SAPIENZA
UNIVERSITÀ DI ROMA

Catene di Markov ed applicazioni ai metodi Monte Carlo

Facoltà di Ingegneria dell'informazione, Informatica e Statistica

Corso di Laurea in Statistica Economia e Società

Candidato

Gian Mario Sangiovanni
Matricola 1889445

Relatore

Prof. Costantino Ricciuti

Correlatore

Prof.ssa Luisa Beghin

Anno Accademico 2021/2022

Tesi discussa il 11 Luglio 2022
di fronte a una commissione esaminatrice composta da:

Prof. Costantino Ricciuti (presidente)
Prof.ssa Luisa Beghin

Catene di Markov ed applicazioni ai metodi Monte Carlo
Tesi di Laurea. Sapienza – Università di Roma

Questa tesi è stata composta con LaTeX e la classe Sapthesis.

Email dell'autore: sangiovanni.1889445@studenti.uniroma1.it

A ciò che è stato, a ciò che è, a ciò che sarà

Sommario

Nel seguente elaborato verrà trattato in maniera generale uno degli argomenti cardine dei processi stocastici, ossia le **Catene di Markov**, ed una loro applicazione alla classe di metodi **Monte Carlo** (McMc). Analizzeremo il caso particolare in cui le variabili aleatorie nel processo siano discrete come anche il tempo.

L'obiettivo principale consiste nel fornire al lettore una presentazione sintetica del suddetto argomento sia dal punto di vista probabilistico sia computazionale. Saranno indagate le leggi probabilistiche alla base di tali processi e il modo in cui vengono usate per creare metodi in grado di modellizzare la realtà osservata.

Studieremo il funzionamento dell'algoritmo di **Metropolis-Hastings**, il quale genera realizzazioni *correlate* a partite da una Catena di Markov, quando la densità di interesse f presenta problemi di natura computazionale. Riusciremo a comprendere la sua grande portata nella generazione di campioni, dando più rilievo ad una visione probabilistica piuttosto che informatica. Questa branca della statistica cade sotto il nome di **Computational Statistics**.

Sicuramente questo elaborato non possiede alcuna finalità in termini di esaustività, piuttosto si è cercato di creare una sintesi di un argomento complesso con l'approccio di uno studente.

Indice

Elenco delle figure

Introduzione

L'argomento principale della tesi riguarda la descrizione delle Catene di Markov discrete ed una loro applicazione alla classe di metodi Monte Carlo. Nel Capitolo 1 vengono introdotte tali Catene, con l'obiettivo di spiegare le nozioni di ergodicità, reversibilità e stazionarietà, utili per il prosieguo della trattazione. Ovviamente, è impossibile esaurire la dissertazione di un argomento così complesso in poche pagine, per questo é stata evitata tutta una parte relativa alle catene di Markov continue.

Nel Capitolo 2 presentiamo l'applicazione delle catene di Markov alla classe di metodi Monte Carlo (McMc), discutendone i principali aspetti metodologici e probabilistici. Nella prima parte viene spiegato (a grandi linee) il funzionamento di tale metodo e il collegamento con le catene, mentre successivamente vengono introdotte le estensioni di due importanti teoremi nel caso di variabili correlate (Legge dei Grandi Numeri e Teorema Del Limite Centrale).

Nel Capitolo 3 viene descritto il funzionamento di uno degli algoritmi basato sulle McMc, ossia il Metropolis-Hastings, e le sue principali criticitá. Esso si sviluppa a partire dalla seconda metá del secolo scorso, diventando in poco tempo molto *popolare* tra gli statistici. Per poterne comprendere al meglio tutte le proprietà, utilizziamo due esempi teorici.

Un algoritmo, come anche un modello, é solo un modo per semplificare la realtà, un modo di imporre una struttura logica ed armonica ad un qualcosa governato dal caos.

Citando il filosofo Friedrich Nietzsche [16]:

> "C'è qualcosa nell'arte, come nella natura del resto, che ci rassicura, e qualcosa che invece, ci tormenta, ci turba.
>
> Due sentimenti eterni in perenne lotta: la ricerca dell'ordine e il fascino del caos.
>
> Dentro questa lotta abita l'uomo, e ci siamo noi, tutti, Ordine e Disordine.
>
> Cerchiamo regole, forme, canoni, ma non cogliamo mai il reale funzionamento del mondo. É per gli uomini un eterno mistero. L'incapacità di risolvere questo mistero ci terrorizza, ci costringe a oscillare tra la ricerca di un'armonia impossibile e l'abbandono dal caos"

Armonia e **Caos**: non è possibile l'esistenza di uno senza l'altro.

Capitolo 1

Catene di Markov

1.1 Introduzione

Prima di iniziare l'analisi delle catene di Markov, risulta necessario inquadrare il contesto entro cui ci stiamo muovendo, dando una definizione, seppur non formale, di un processo stocastico [1] e spiegando la principale differenza rispetto ad un processo deterministico.

Un **processo stocastico** può essere definito come una famiglia di variabili aleatorie $\{X_t,\ t \in T\}$, dove la generica X_t è definita su uno spazio di probabilità $(\mathbf{\Omega}, \mathbf{A}, \mathbf{P})$. Tramite un processo possiamo quindi modellizzare un fenomeno evolutivo.
Ricordando che una v.a [2] è una funzione $X : \Omega \to R$, allora un processo stocastico é scrivibile come un insieme di funzioni $\{X_{(t,\omega)},\ t \in T\ e\ \omega \in \Omega\}$. Fissato un certo $\omega_0 \in \Omega$, la funzione $X_{(t,\omega_0)}$ è una **Traiettoria** (realizzazione) del processo stocastico $\{X_t,\ t \in T\}$[5]. Se ad esempio studiassimo l'evoluzione di un fenomeno all'interno di una popolazione, fissare ω sarebbe come scegliere una generica unità e seguirla nel tempo. Al contrario, fissare t implicherebbe lo studio della popolazione in un definito istante.
È possibile classificare un processo stocastico in base all'immagine della generica X_t e allo spazio dei tempi T. In particolare, se le X_t sono discrete allora il processo

[1] *stochastikos* (greco) = che tende al bersaglio, approssima
[2] nel prosieguo del testo si alterna l'utilizzo di v.a e variabile aleatoria

si definisce *discreto*, altrimenti è detto *continuo*. Se l'insieme T [3] contiene al più un'infinità numerabile di elementi, allora il processo si definisce a *parametro discreto*, mentre se T [4] contiene un'infinità non numerabile di elementi, allora il processo è a *parametro continuo*.
Secondo il **Teorema di Consistenza** (Estensione) di Kolmogorov (vedi [6] [20]), per poter *conoscere* un processo stocastico bisogna essere in grado di calcolarne la distribuzione congiunta a più tempi, cioè per ogni collezione di tempi $t_1 < t_2 < t_3 < \ldots$ é possibile scrivere la distribuzione congiunta. *"Conoscere"* un processo stocastico vuol dire quindi essere in grado di rappresentare le relazioni tra le variabili. All'interno di un processo possono sussistere o meno relazioni di **dipendenza stocastica** tra le diverse n-uple di variabili.

A differenza del caso precedente, un **processo deterministico** considera solamente l'evoluzione nel tempo di una quantità x_t, non più aleatoria. Scegliamo un valore di input x_0 fisso/costante attraverso il quale modellizzare la realtà, senza l'utilizzo di alcuna legge probabilistica. È un chiaro riferimento alla concezione deterministica: essa assume l'esistenza di una spiegazione per tutti i fenomeni naturali e la presenza di relazioni causa-effetto tra essi.
Citando Laplace: [13]

> "Gli eventi attuali hanno un legame con quelli che li precedono, il quale legame è fondato sul principio evidente che una cosa non possa cominciare ad essere senza una causa che la produca. ... Dobbiamo dunque considerare lo stato presente dell'universo come effetto del suo stato anteriore, e come causa di quello che seguirà"

Lo stesso filosofo, seppur comprendendo la casualità del mondo stesso e delle sue leggi, si dovette accontentare di una spiegazione provvisoria e limitata, espressione della nostra ignoranza (la probabilità!)[11]. Filosofia e scienza sono indissolubilmente legate tra loro, per questo è stato doveroso fare una piccola digressione.

[3] $T = \{0, 1, 2, \ldots\}$
[4] $T \in [0, \infty) = R^+$

Per poter inferire sul futuro è sufficiente osservare il processo nel presente, tralasciando ciò che è accaduto nel passato. Al fine di cogliere questo particolare aspetto della matematica, **Andrey Andreyevich Markov** introdusse lo studio di modelli che ora portano il suo nome.

1.2 Generalitá sulle Catene di Markov

Sia data una successione di variabili aleatorie $\{X_t,\ t \in T\}$, dove la generica X_t assume valori nell'insieme $S = \{s_0, \ldots, s_k\}$, definito come lo **spazio degli stati** associato alla catena. D'ora in avanti $T = \{0, 1, 2, \ldots\}$ e S è fissato con ordine $k+1$, cioè lavoriamo con un processo discreto a parametro discreto. Spesso si è soliti identificare un istante iniziale ($t = 0$) su cui si fanno particolari ipotesi o lo stato si suppone essere noto.

Per comodità occorre pensare alla variabile X_t come uno dei possibili stati del sistema al tempo t. Si può quindi affermare che il sistema è nello stato j al tempo t se $X_t = j$.

Come logica conseguenza del Teorema di Consistenza, risulta necessario concentrarsi sulla distribuzione congiunta associata al processo in quanto è la principale fonte d'informazione che possediamo. Ricapitolando, una catena risulta essere definita da:

1. distribuzione congiunta a più tempi;
2. insieme degli stati **S**;
3. insieme dei tempi **T**.

Ricordando la regola della catena [5], é possibile scrivere la distribuzione congiunta come:

$$
\begin{aligned}
&P(X_0 = s_0 \cap \ldots \cap X_t = s_t) \qquad (1.1)\\
&= P(X_t = s_t | X_0 = s_0 \cap \ldots \cap X_{t-1} = s_{t-1}) \cdot P(X_0 = s_0 \cap \ldots \cap X_{t-1} = s_{t-1})\\
&= \ldots = P(X_0 = s_0) \cdot P(X_1 = s_1 | X_0 = s_0) \cdot \ldots \cdot P(X_t = s_t | X_0 = s_0 \cap \ldots X_{t-1} = s_{t-1})
\end{aligned}
$$

Uno dei principali obiettivi è di trovare un modo per scrivere la distribuzione congiunta in un processo Markoviano e determinare gli elementi da cui dipende.

[5] $P(A \cap B \cap C \ldots \cap D) = P(D|A \cap B \ldots \cap C) \cdot \ldots \cdot P(C|A \cap B) \cdot P(B|A) \cdot P(A)$

Definizione 1.2.1 (Markov)**.** *Un processo stocastico* $\mathbb{X}$ *é una* ***Catena di Markov*** *se soddisfa la seguente proprietá:*

$$P(X_t = s_t | X_0 = s_0 \cap \ldots \cap X_{t-1} = s_{t-1}) = P(X_t = s_t | X_{t-1} = s_{t-1}) \tag{1.2}$$

Stiamo affermando un qualcosa di molto *forte*, in quanto risulta importante solo ciò che accade nell'istante di tempo immediatamente precedente e non tutto il resto. Si dice che il processo goda di *mancanza di memoria*. Tale proprietà non riguarda effettivamente il modello in sè, ma è solo un'ipotesi semplificativa.
Abbiamo aggiunto un tassello davvero importante in quanto un processo aleatorio, se Markoviano, perde memoria del passato in senso condizionato (passato e futuro sono indipendenti se si conosce il presente).
Tornando alla distribuzione congiunta (1.1), essa diventa:

$$\begin{aligned} &P(X_0 = s_0 \cap \ldots \cap X_t = s_t) \\ &= P(X_0 = s_0) \cdot P(X_1 = s_1 | X_0 = s_0) \cdot \ldots \cdot P(X_t = s_t | X_{t-1} = s_{t-1}) \end{aligned}$$

Definizione 1.2.2 (Omogeneità temporale)**.** *Una* ***Catena di Markov*** $\mathbb{X}$ *é detta omogenea nel tempo se:*

$$P(X_t = j | X_{t-1} = i) = P(X_1 = j | X_0 = i) \quad t \in T \quad (i,j) \in S \tag{1.3}$$

Alternativamente, un processo Markoviano è omogeneo nel tempo se le probabilità condizionate, espresse nella Definizione 1.2.1, non variano aggiungendo o sottraendo una certa costante al parametro temporale. Tali processi sono quindi invarianti rispetto a traslazioni temporali.[9]
Si definisce la *matrice di transizione ad un solo passo* **H** di ordine $(k+1) \times (k+1)$ [6], il cui generico elemento $h_{i,j}$ è espresso dalla formula (1.3). Esso rappresenta la probabilità di passare dallo stato i allo stato j in un tempo t uguale ad uno.

Proposizione 1. *La matrice* **H** *è una matrice stocastica i cui elementi soddisfano le seguenti proprietà:*

- $h_{i,j} \geq 0 \rightarrow$ *gli elementi sono positivi*
  $\forall (i,j) \in S$

[6] l'ordine della matrice dipende dalla cardinalità di S

- $\sum_{j=s_0}^{s_k} h_{i,j} = 1 \rightarrow$ *la somma sulle righe é unitaria*
  $\forall i \in S$

Osservazione 1.2.1. 2 *considerazioni:*

- *All'interno di una catena di Markov, la transizione tra uno stato ed un altro avviene in maniera totalmente probabilistica e non deterministica;*
- *tramite H stiamo studiando l'evolversi della catena nel* breve periodo *(tempo unitario).*

Possiamo ora scrivere in forma estesa la generica matrice H:

$$H = \begin{bmatrix} h_{0,0} & h_{0,1} & \dots & h_{0,k} \\ h_{1,0} & h_{1,1} & \dots & h_{1,k} \\ \vdots & & & \\ h_{k,0} & h_{k,1} & \dots & h_{k,k} \end{bmatrix} \tag{1.4}$$

Osservazione 1.2.2. *Prendendo in considerazione una situazione reale, esiste un modo per conoscere le diverse probabilità di transizione?*
Si possono trovare almeno due soluzioni:

- *vengono stabilite direttamente dal ricercatore in accordo a qualche modello;*
- *le* inferiamo *a partire da dati reali, cioè osservando il comportamento nel passato di un sistema simile a quello in esame.*

Affinchè si possa studiare una catena di Markov, è necessario essere in grado di rappresentarla tramite la sua matrice di transizione o tramite grafi connessi e orientati.

Esempio 1.2.1 (Previsioni meteo)**.** *[19] Ipotizziamo che il fatto di piovere o meno nei giorni successivi dipenda esclusivamente dalle condizioni metereologiche del giorno corrente e non sia legato in alcun modo a ciò che è successo nei giorni precedenti. Facciamo le seguenti assunzioni:*

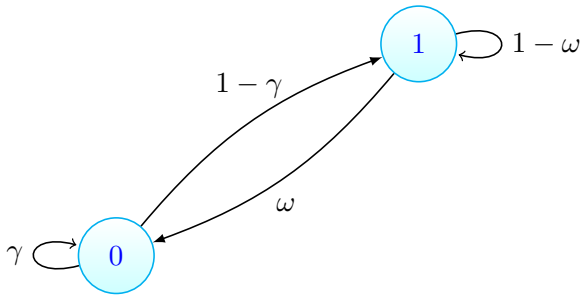
- *se oggi piove, domani pioverà con probabilità pari ad γ.*
  Questo implica che non pioverá con probabilità $1 - \gamma$.

- *se oggi non piove, domani pioverà con probabilità pari a* ω.
  Questo implica che non pioverá con probabilità $1-\omega$.

Il sistema è nello stato 0 *quando piove ed* 1 *quando non piove. Quindi, stiamo definendo una catena di Markov discreta con parametro discreto avente la seguente matrice di transizione* (1.4)*:*

$$H = \begin{bmatrix} \gamma & 1-\gamma \\ \omega & 1-\omega \end{bmatrix}$$

Figura 1.1. grafo rappresentativo (1)



Fino ad ora ci siamo focalizzati solamente sul passaggio da uno stato ad un altro quando t é uguale ad 1. Come si calcola la probabilità di transizione tra uno stato i e uno stato j in m passi?

Definizione 1.2.3. *Dati due generici stati* i *e* j *appartenenti ad* S*, si definisce la probabilità di transizione tra* i *due stati in* m *passi come:*

$$p_{i,j}^{(m)} = P(X_{m+n} = j | X_n = i) = P(X_m = j | X_0 = i) \tag{1.5}$$

Osservazione 1.2.3. *Il secondo passaggio nella* (1.5) *dipende dalla proprietà di omogeneità temporale* (1.3)*. Se* j *è uguale ad* i*, stiamo calcolando la probabilità di tornare nello stesso identico stato in un tempo esattamente uguale ad* m.

Definizione 1.2.4. *Con considerazioni analoghe per la matrice* H*, possiamo definire la* ***Matrice di transizione in m passi*** *il cui generico elemento* $p_{i,j}^{(m)}$ *è descritto*

dalla (1.5)*:*

$$P^{(m)} = \begin{bmatrix} P_{0,0}^{(m)} & P_{0,1}^{(m)} & \dots & P_{0,k}^{(m)} \\ P_{1,0}^{(m)} & P_{1,1}^{(m)} & \dots & P_{1,k}^{(m)} \\ \vdots & & & \\ P_{k,0}^{(m)} & P_{k,1}^{(m)} & \dots & P_{k,k}^{(m)} \end{bmatrix} \tag{1.6}$$

Osservazione 1.2.4. *Sono valide le stesse proprietà relative alla matrice H (Proposizione 1) e anch'essa presenta ordine $k+1$. Inoltre, tramite la matrice di transizione ad m passi* (1.6)*, si può studiare l'evoluzione della catena nel lungo periodo.*

Osservazione 1.2.5. 2 *considerazioni:*

- $P^{(1)} = H$*;*
- $P^{(0)} = I$*, dove I è la matrice identità.*

1.2.1 Legame tra H e P

Teorema 1.2.1.

$$P^{(m)} = H^m \tag{1.7}$$

Dimostrazione.

$$\begin{aligned} p_{i,j}^{(m)} = P(X_m = j | X_0 = i) &= \sum_s [P(X_m = j | X_{m-1} = s \cap X_0 = i) \cdot P(X_{m-1} = s | X_0 = i)] \\ &= \sum_s [P(X_m = j | X_{m-1} = s) \cdot P(X_{m-1} = s | X_0 = i)] \\ &= \sum_s (H_{s,j} \times P_{i,s}^{(m-1)}) = \sum_s (P_{i,s}^{(m-1)} \times H_{s,j}) \end{aligned}$$

Sfruttando la legge delle alternative [7] e la proprietà di markovianità, abbiamo trovato un collegamento tra le due probabilità di transizione. In particolar modo:

[7] $P(A) = \sum_v P(A|B_v) \times P(B_v) \rightarrow B_v$ sono le "*cause*" dell'evento A. Le B_v formano una partizione dello spazio campionario

$$P^{(m)} = P^{(m-1)} \times H.$$
$$\vdots$$
$$P^{(1)} = H$$
$$P^{(0)} = I$$

Iterando questo procedimento per $m-1$ volte si giunge al risultato. ■

Osservazione 1.2.6. *la matrice di transizione* (1.4) *ad un solo passo è sufficiente per studiare l'evoluzione del processo (sia nel breve sia nel lungo periodo).*

Teorema 1.2.2. *(Equazioni Chapman-Kolmogorov)*

$$P^{(m_1+m_2)} = P^{(m_1)} \cdot P^{(m_2)}$$
$$p_{i,j}^{(m_1+m_2)} = P(X_{n+m_1+m_2} = j | X_n = i) = \textstyle\sum_s (p_{i,s}^{(m_1)} \cdot p_{s,j}^{(m_2)})$$

Dimostrazione. [8]

$$\begin{aligned} p_{i,j}^{(m_1+m_2)} &= P(X_{n+m_1+m_2} = j | X_n = i) = \sum_s [P(X_{n+m_1+m_2} = j \cap X_{n+m_1} = s | X_n = i] \\ &= \sum_s [P(X_{n+m_1+m_2} = j | X_{n+m_1} = s \cap X_n = i) \cdot P(X_{n+m_1} = s | X_n = i)] \\ &= \sum_s [P(X_{n+m_1+m_2} = j | X_{n+m_1} = s) \cdot P(X_{n+m_1} = s | X_n = i] \\ &= \sum_s [P(X_{m_2+m_1} = j | X_{m_1} = s) \cdot P(X_{m_1} = s | X_0 = i)] \\ &= \sum_s (p_{s,j}^{(m_2)} \cdot p_{i,s}^{(m_1)}) \end{aligned}$$

In questa operazione si riconosce il prodotto tra due matrici [2]:

$$P^{(m_1+m_2)} = P^{(m_1)} \cdot P^{(m_2)}$$

■

Osservazione 1.2.7. *Dato che le variabili di partenza sono aleatorie, lo è anche il tempo di attesa per passare da uno stato i ad uno stato j (ovvero il tempo di permanenza nello stato i).*

[8] $P(A \cap B|C) = P(A|B \cap C) \cdot P(B|C)$

1.3 Tempo d'attesa

Ci troviamo di fronte ad un nuovo quesito:
Qual è la probabilità di permanere nello stato i un tempo t uguale ad $l-1$? [9]

Teorema 1.3.1. *(Distribuzione intertempi) Dato un generico stato i, definisco T_i come il tempo di attesa prima di spostarsi dallo stato i-esimo.*
Allora:

$$T_i \sim Geom(1-h_{i,i}) \tag{1.8}$$

Ovvero:

$$P(T_i = l) = \begin{cases} (h_{i,i})^{l-1} \cdot (1-h_{i,i}) & l = 1,2,3,\dots \\ 0 & altrimenti \end{cases} \tag{1.9}$$

Dimostrazione. Per questa dimostrazione useremo un semplice espediente. Assumendo di trovarci nello stato i, si resta nello stesso stato con probabilità $h_{i,i}$ mentre ci si muove verso un qualsiasi altro stato con probabilità $1-h_{i,i}$. Per ogni singolo tempo t siamo di fronte ad una scelta binaria, modellizzabile tramite la distribuzione bernoulliana. È una successione di l prove "indipendenti". [10]

$$\begin{aligned} P(T_i = l) &= \sum_{j\neq i}[P(X_l = j \cap X_{l-1} = i \cap \dots \cap X_1 = i | X_0 = i)] \\ &= \sum_{j\neq i}[P(X_l = j | X_{l-1} = i \cap \dots \cap X_0 = i) \cdot P(X_{l-1} = i \cap \dots \cap X_1 = i | X_0 = i)] \\ &= \dots = \sum_{j\neq i}[h_{i,j}] \cdot h_{i,i}^{l-1} = (1-h_{i,i}) \cdot h_{i,i}^{l-1} \end{aligned}$$

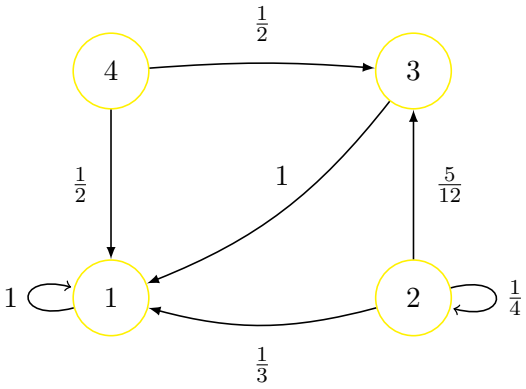
■

Osservazione 1.3.1. *Non è un caso che si possa modellizzare il tempo di attesa tramite una distribuzione geometrica, in quanto essa gode di mancanza di memoria. Tale distribuzione è usata per studiare il tempo di attesa prima di un successo nel caso di esperimenti bernoulliani.*

[9] conseguentemente di uscire dallo stato i in un tempo t uguale ad l
[10] si sfrutta ancora che $P(A \cap B|C) = P(A|B \cap C) \cdot P(B|C)$

Esempio 1.3.1. *Supponiamo di osservare una catena di Markov identificata da* $\mathbb{X} = \{X_t,\ t \in T\}$, *in cui la generica variabile* X_t *assume valori in* $S = \{1, 2, 3, 4\}$. *La matrice di transizione è:*

Figura 1.2. grafo rappresentativo (2)



$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{4} & \frac{5}{12} & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Possiamo porci diverse domande:

- *Qual è la probabilità di uscire dallo stato* 2 *dopo un tempo* l*?*
  $P(T_2 = l) = (\frac{1}{4})^{l-1} \cdot (\frac{3}{4})$
- *Qualè la probabilità di non uscire mai dallo stato* 1*?*
  $P(T_1 = \infty) = 1 \to 1$ *è uno stato* ***assorbente***[11]
- *Qual è la probabilità di uscire dallo stato* 4 *dopo un tempo* t *uguale ad* 1*?*
  $P(T_4 = 1) = 1 \to 4$ *è uno stato* ***repellente***[12].

[11] una volta arrivato nello stato 1 risulta impossibile uscirne
[12] il sistema permane nello stato 4 al massimo $t = 1$

1.4 Distribuzione congiunta

Possiamo definire la notazione per la distribuzione marginale. Essa rappresenta la probabilità che in un certo istante di tempo la catena assuma un determinato valore (stato).
In formule [13]:

$$u_d^{(n)} = P(X_n = d) \quad d \in S,\ n \in T$$

Generalizzando: $u^{(n)} = \left(u_0^{(n)}, \ldots, u_k^{(n)}\right) = \Big(P(X_n = 0), \ldots, P(X_n = k)\Big)$

Questa è la distribuzione marginale della v.a X_n.
$u^{(n)}$ prende anche il nome di *vettore delle probabilità di stato* ed è un vettore riga.

Teorema 1.4.1. *(regola iterativa) Data una Catena di Markov $\mathbb{X}$ con matrice di transizione H, è possibile affermare che:*

$$u^{(n+1)} = u^{(n)} \cdot H \tag{1.10}$$

Dimostrazione. Sappiamo che:

$$\begin{aligned} u_j^{(n+1)} &= P(X_{n+1} = j) = \sum_i [P(X_{n+1} = j | X_n = i) \cdot P(X_n = i)] \\ &= \sum_i (p_{i,j}^{(1)} \cdot u_i^{(n)}) = \sum_i (u_i^{(n)} \cdot h_{i,j}) = (u^{(n)} \cdot H)_j \quad \forall j \end{aligned}$$

■

Cosa accadrebbe se iterassimo tale formula?

Lemma 1.4.1. *È valida la seguente relazione:*

$$u^{(n+1)} = u^{(0)} \cdot H^{n+1} = u^{(0)} \cdot P^{(n+1)}$$

Dimostrazione.

$$u^{(n+1)} = u^{(n)} \cdot H = u^{(n-1)} \cdot H^2 = \ldots = u^{(0)} \cdot H^{n+1}$$

■

[13] probabilità che il processo al tempo n sia nello stato d

Per ricavare la distribuzione di probabilità congiunta di una catena di Markov è necessario il possesso di 2 elementi:

- $u^{(0)}$, ovvero la distribuzione marginale iniziale;
- H, ossia la matrice di transizione ad un solo passo.

Molti problemi legati alle catene di Markov possono essere espressi in termini di queste quantità, facendo ridurre il tutto ad uno studio della matrice H.

Proposizione 2. *[17] È possibile scrivere la distribuzione congiunta* (1.1) *nel seguente modo:*
$P(X_0 = s_0 \cap \ldots \cap X_t = s_t) = u^{(0)}_{s_0} \cdot H_{s_0,s_1} \cdot \ldots \cdot H_{s_{t-1},s_t}$

Come riportato nell'introduzione, lo studio della distribuzione congiunta risulta essere di fondamentale importanza per l'analisi della **dipendenza stocastica**.

Esempio 1.4.1. *(Semafori) Avete mai avuto l'impressione che il semaforo rosso per i pedoni duri un'eternità? No? Evidentemente sono io sempre in ritardo!*
Supponiamo di avere una Catena di Markov $\mathbb{X}$*, tale per cui la generica* X_t *assuma valori in* $S = \{1, 2, 3\}$*, dove lo stato* 1 *equivale al rosso, lo stato* 2 *al giallo, mentre lo stato* 3 *al verde.*
Supponiamo che al tempo 0 *nessuno stato sia favorito rispetto ad un altro, quindi* $u^{(0)} \sim Uniforme\{1, 2, 3\}$
La matrice di transizione è tale per cui:

$$H = \begin{pmatrix} \frac{5}{12} & \frac{7}{12} & 0 \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

Possiamo porci diverse domande:

1. *qual è la probabilità di aspettare piú di 5 secondi per attraversare?*
   $P(T_1 > 5) = 1 - P(1 \leq T_1 \leq 5) = 1 - \sum_{g=1}^{5}[(\frac{5}{12})^{g-1} \cdot \frac{7}{12}]$

2. *qual è la probabilità che al tempo $t = 40s$ il semaforo sia rosso?*
   $u_1^{(40)} = (u^{(0)} \cdot H^{40})_1 \simeq 0.2352521$

3. *qual è la probabilità della seguente intersezione di eventi?*
   $P(X_0 = 3 \cap X_1 = 2 \cap X_2 = 1 \cap X_3 = 1) = u_3^{(0)} \cdot h_{3,2} \cdot h_{2,1} \cdot h_{1,1} \simeq 0.0185$

Ipotiziamo di simulare il percorso di tale catena di Markov. Osservando i valori assunti da essa in un tempo t uguale a cento secondi, come si comporta allo scorrere del tempo? che traiettoria assume? Nel seguente grafico è riportato il risultato di tale simulazione.

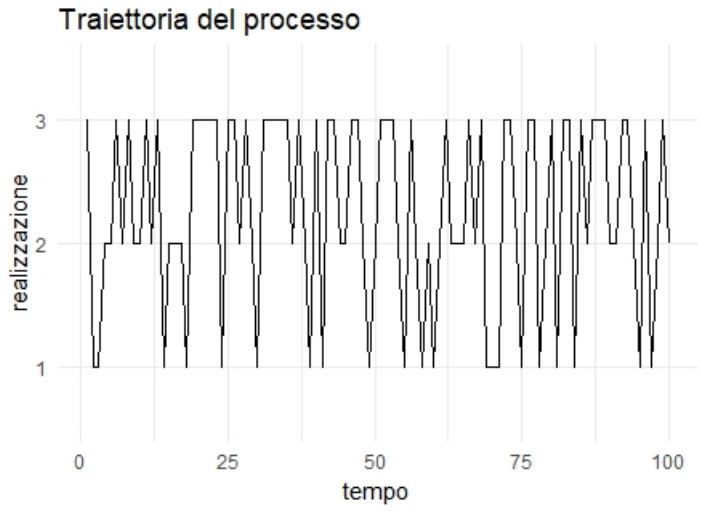


Figura 1.3. Traiettoria del processo con $T \in [0, 100]$

1.4.1 Classificazione degli stati

In maniera del tutto intuitiva, possiamo pensare allo sviluppo di una catena come al moto di particelle che *saltano* da uno stato all'altro. Per lo studio di una Catena di Markov è fondamentale esaminare nel dettaglio le caratteristiche degli stati, giungendo ad una *Classificazione degli stati* stessi. È possibile classificare uno stato in base alla probabilità di passare ad un altro o di tornare su se stesso dopo un certo tempo t.

Prima di entrare nel vivo della questione, è necessario introdurre la seguente v.a:

$$F_i = min\{t \geq 1 : \ X_t = i\}$$

Essa è definita come il tempo della prima *visita* allo stato i. Convenzionalmente, $F_i = \infty$ se tale visita non avviene mai.

Definizione 1.4.1. *(Tempo di primo ritorno) Dato uno stato iniziale i, si definisce* ***Tempo di primo ritorno*** *il tempo trascorso prima di ritornare in tale stato. La sua distribuzione è data da:*

$$P(F_i = t | X_0 = i) \tag{1.11}$$

In aggiunta, si definisce il ***Tempo medio di primo ritorno*** *come il valore atteso di tale distribuzione.*
Formalmente:

$$\mu_i = E[F_i | X_0 = i] \tag{1.12}$$

La (1.12) rappresenta la ricorrenza media di un determinato stato, cioè quanto tempo in media la catena impiega per ritornare in quello stato. Analogamente, $\frac{1}{\mu_i}$ rappresenta la frazione di tempo trascorso nello stato i; tale valore può essere interpretato come una sorta di peso associato al medesimo stato.

Definizione 1.4.2. *Sono valide le seguenti definizioni [12]:*

- *Uno stato j si definisce* ***Accessibile*** *tramite uno stato i se per un qualche intero $t \geq 0 \rightarrow p_{i,j}^{(t)} > 0$;* [14]
- *Se due stati i e j sono entrambi accessibili tramite l'altro, allora diremo che* ***Comunicano****: $i \iff j$;*
- *Se due stati non comunicano allora:*
  $p_{i,j}^{(t)} = 0$ *o* $p_{j,i}^{(t)} = 0 \ \forall t \in T$;
- *Uno stato g si definisce* ***Transitorio*** *se:*
  $P(F_g = \infty | X_0 = g) > 0$ [15] *e* $\mu_g = \infty$;
- *Uno stato a si definisce* ***Ricorrente (persistente)*** *se:*
  $P(F_a < \infty | X_0 = a) = 1 \iff P_{a,a}^{(t)} = 1$ *per un qualche t;*

[14] esiste una probabilità positiva di raggiungere j a partire da i in un numero finito di transizioni.
[15] questo deriva dal fatto che $P(F_g = \infty) > 0 \rightarrow$ v.a impropria

- *Dato uno stato a ricorrente, esso è definito* ***Nullo*** *se* $\mu_a = \infty$.
  Se invece $\mu_a < \infty$ [16], *allora a é* ***Non Nullo***.

Osservazione 1.4.1. *Si può dimostrare che se due stati comunicano, allora devono essere dello stesso tipo (transitori o ricorrenti). Un insieme di stati C è definito insieme chiuso se tutti gli stati all'interno comunicano ma non comunicano con gli stati esterni a C, quindi* $h_{j,i} = 0$ *se* $j \in C$ *e* $i \notin C$.

Ovviamente non è sempre possibile individuare delle sottocatene. Un caso molto interessente si ha quando esiste solo la catena principale e non è scomponibile; questo avviene quando tutti gli stati comunicano tra loro.

Definizione 1.4.3. *Se tutti gli stati di una catena comunicano tra loro, allora essa è detta* ***Irriducibile*** [17]. *In formule:*

$$p_{i,j}^{(t)} > 0 \text{ per qualche } t \in T \text{ e } \forall (i,j) \in S$$

Definizione 1.4.4. *(Periodo) Il* ***periodo*** $d(j)$ *di uno stato* j *è definito come il massimo comun divisore tra i tempi possibili per ritornare allo stato* j. *In formule:*

$$d(j) = mcd\{t : p_{j,j}^{(t)} > 0\}$$

se:

- $d(j) > 1 \rightarrow$ j *è periodico;*
- $d(j) = 1 \rightarrow$ j *è aperiodico*

Osservazione 1.4.2. *Se tutti gli stati di una catena sono aperiodici (oppure periodici), allora la catena è detta aperiodica di periodo 1 (oppure periodica di periodo d)*

Definizione 1.4.5. *Uno stato si definisce* ***Ergodico*** *se è persistente, non nullo e aperiodico.*

Osservazione 1.4.3. *Se tutti gli stati di una catena sono ergodici, allora la catena è detta Ergodica*

[16] il tempo medio di ricorrenza è finito
[17] il nome stesso dà l'idea che non si possano riscontrare sottocatene

Esempio 1.4.2. *(Previsioni meteo a Roma) Con riferimento all'esempio 1.2.1*
È facile vedere che la catena è aperiodica poiché, considerando tutti i tempi per ritornare allo stato j, ci sono almeno due numeri primi .
Inoltre, i due stati comunicano e sono ricorrenti, ergo la catena è irriducibile.
Potremmo essere interessati a sapere quanti giorni di pioggia (oppure di non pioggia) consecutivi ci sono stati in questo mese. Siamo interessati alla probabilità che il primo giorno di non pioggia arrivi dopo k giorni di pioggia. [18]

$$P(X_{k+1} = 1 \cap X_k = 0 \cap \ldots \cap X_{k-5} = 0 \cap \ldots \cap X_1 = 0 | X_0 = 0)$$
$$= P(X_{k+1} = 1 | X_0 = 0 \cap \ldots \cap X_k = 0) \cdot P(X_k = 0 \cap \ldots \cap X_1 = 0 | X_0 = 0)$$
$$= \ldots = h_{0,1} \cdot h_{0,0}^k = (1 - \gamma) \cdot \gamma^k$$

Potremmo anche essere interessati a sapere quale sia la probabilità che piova dopo k giorni, sapendo che ha piovuto al tempo zero. (1.11)

$$f_{0,0} = P(X_k = 0 \cap \ldots \cap X_1 = 1 | X_0 = 0)$$
$$= P(X_k = 0 | X_{k-1} = 1 \cap \ldots \cap X_0 = 0) \cdot P(X_{k-1} = 1 \cap \ldots \cap X_1 = 1 | X_0 = 0)$$
$$= \ldots = h_{1,0}^2 \cdot h_{1,1}^{k-2} = \omega^2 \cdot \omega^{k-2}$$

1.5 Distribuzione stazionaria

Un concetto importante riguarda lo studio di un sistema in equilibro, cioè quando le sue condizioni non variano nel tempo.
Poniamoci ora due nuovi quesiti:

1. Esiste una distribuzione che non cambia con lo scorrere del tempo?
2. Come si comporta la distribuzione della catena X_n al crescere di n?

Definizione 1.5.1. *Il vettore $\pi = (\pi_0, \ldots, \pi_k)$ è definito **distribuzione invariante** (stazionaria) se i suoi elementi π_j sono tali che:*

- $\pi_j \geq 0 \quad \forall j \in S \quad e \quad \sum_j \pi_j = 1$;

[18] abbiamo anche dimostrato empiricamente che gli intertempi seguono una distribuzione geometrica

- $\pi = \pi H \rightarrow$ π *è autovettore sinistro con autovalore unitario.*

Osservazione 1.5.1. *Essa assume importanza in virtù di questo semplice fatto:*

$$\pi P^{(m)} = (\pi H)H^{m-1} = \ldots = \pi H = \pi \text{ [19]}$$

Se quindi $u^{(0)} = \pi$, allora la distribuzione di probabilità marginale rimane la stessa ad ogni passo della catena. Detto in maniera intuitiva, la distribuzione di X_n non dipende dal tempo n. Il fatto che una catena possieda una distribuzione stazionaria non implica necessariamente la stazionarietà della catena.

Definizione 1.5.2. *(stazionarietà) Una catena di Markov è definita* ***stazionaria*** *se tutte le distribuzioni marginali sono uguali tra loro.*
In formule:

$$X_t \overset{d}{=} X_0 \quad \forall t$$

Prendendo in considerazione $u^{(0)}$, dal lemma 1.4.1 è valido che:

$$u^{(0)} \cdot P^{(m)} = u^{(m)}$$

Se la catena è stazionaria, allora:

$$u^{(0)} \cdot P^{(m)} = u^{(0)}$$

Questo equivale a dire che $u^{(0)}$ é la distribuzione stazionaria. In generale, non esiste sempre una distribuzione stazionaria e, anche se esistesse, nulla garantirebbe la sua unicità.

Teorema 1.5.1. *(esistenza ed unicità) Sia data una catena di Markov irriducibile* $\mathbb{X}$ *con stati ricorrenti non nulli, allora esiste ed è unica una distribuzione invariante* $\pi = (\pi_0, \ldots, \pi_k)$.
Inoltre:

$$\pi = (\pi_0, \ldots, \pi_k) = \left(\frac{1}{\mu_0}, \ldots, \frac{1}{\mu_k}\right) \tag{1.13}$$

dove il generico μ_k *rappresenta il tempo medio di primo ritorno per lo stato k, come enunciato nella definizione* (1.12).

[19] questa cosa è valida in virtù del Teorema 1.2.1

Osservazione 1.5.2. *Dato che gli stati sono tutti persistenti non nulli,* $\frac{1}{\mu_i} > 0$. *Questo significa che tutti gli stati vengono visitati con una probabilità diversa da* 0, *inversamente proporzionale al tempo medio di primo ritorno in ciascuno stato.*

Per rispondere al secondo quesito, posto all'inizio del paragrafo, dovremmo studiare l'andamento di $u^{(n)}$ per $n \to \infty$. Stiamo quindi ragionando sulla **Distribuzione Limite**. È ovvio che nel caso in cui $u^{(0)} = \pi$, la distribuzione stazionaria coincide con quella limite. In generale, se la catena ammette distribuzione limite essa è anche stazionaria

Infatti, per il Teorema (1.4.1):

$u^{(n+1)} = u^{(n)} \cdot H$

se n $\to \infty$, allora $u^{(\infty)} = u^{(\infty)} \cdot H$ soddisfa l'equazione invariante (Definizione 1.5.1)

Non vale il viceversa a meno di particolari condizioni, più rigorose rispetto a quelle del Teorema (1.5.1). Infatti, è necessario aggiungere l'ipotesi di aperiodicità, cioè andando a considerare una catena ergodica.

Teorema 1.5.2. *(Ergodico) Sia data una catena di Markov ergodica* $\mathbb{X}$, *allora esiste una distribuzione stazionaria* $\pi = (\pi_0, \dots, \pi_k)$ *e concide con quella limite per ogni scelta di* $u^{(0)}$.

Se $n \to \infty$:

$$p_{i,j}^{(n)} \to \pi_j = \frac{1}{\mu_j}$$

$$P^{(n)} \to \Pi = \begin{bmatrix} \pi_0 & \dots & \pi_k \\ \vdots & & \vdots \\ \pi_0 & \dots & \pi_k \end{bmatrix}$$

Osservazione 1.5.3. *Per un n sufficientemente grande,* π_j *rappresenta la probabilità di essere nello stato* j *al tempo n, senza tener conto dello stato di partenza. Quindi, gli stati* j *che presentano* π_j *maggiore sono quelli che vengono visitati più volte.*

Proposizione 3. *Supponendo valido il precedente teorema, vale anche che se n*$\to \infty$:

$$u^{(n)} \to \pi = \left(\frac{1}{\mu_0}, \dots, \frac{1}{\mu_k}\right)$$

Dimostrazione. Per un certo stato $j \in S$:

$$u_j^{(n)} = \sum_i u_i^{(0)} \cdot p_{i,j}^{(n)} \to \sum_i u_i^{(0)} \cdot \pi_j = \pi_j \tag{1.14}$$

■

Esempio 1.5.1. *Data una una catena di Markov* $\mathbb{X}$ *in sui* $S = \{s_0, s_1, s_2, s_3\}$ *e* $u^{(0)} \sim Uniforme\{1, 2, 3, 4\}$. *La matrice di transizione* H *è tale per cui:*

$$H = \begin{bmatrix} 0.2 & 0.3 & 0 & 0.5 \\ 0.7 & 0.2 & 0.1 & 0 \\ 0 & 0.4 & 0.4 & 0.2 \\ 0.9 & 0 & 0 & 0.1 \end{bmatrix}$$

Soddisfa le condizioni del teorema ergodico, in quanto la catena è irriducibile ed aperiodica. Allora:

$u^{(n)} \simeq (0.48828606, 0.19975339, 0.03329223, 0.27866831)$ $(n \to \infty)$

Essa sarà anche la distribuzione invariante. Graficamente:

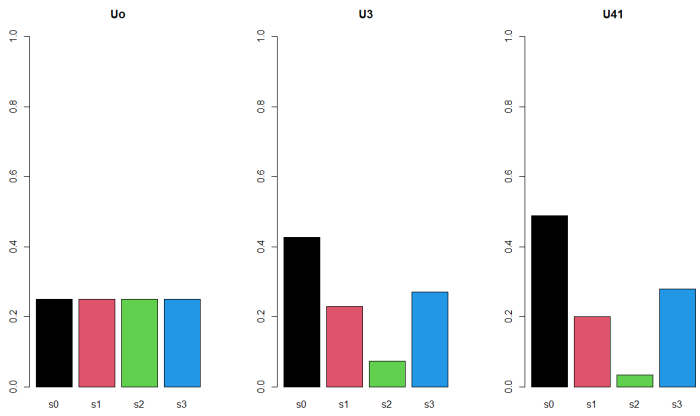


Figura 1.4. Distribuzione marginale allo scorrere del tempo

1.5.1 Catene reversibili

Il concetto di reversibilità temporale è tanto affascinante quanto complesso da capire ed applicare a situazioni reali. Non è nostro compito interrogarci sulla sua veridicità,

piuttosto introdurre un meccanismo utile per lo studio del *bilancio globale* di una catena.

Sia data una catena ergodica $\mathbb{X} = \{X_t,\ t \in [0, N]\}$, in cui H è la matrice di transizione e π è la distribuzione invariante. Si definisce la *catena inversa* $\mathbb{Y}$ tale per cui il generico $Y_t = X_{N-t}$. Si può dimostrare che essa è ancora una catena di Markov [9].

Definizione 1.5.3. *Sia data una catena di Markov ergodica $\mathbb{X}$ con distribuzione invariante π. Allora, essa è reversibile se la matrice di transizione di $\mathbb{X}$ è uguale a quella di $\mathbb{Y}$, ossia:*

$$\pi_s \cdot h_{s,s^*} = \pi_{s^*} \cdot h_{s^*,s} \quad \forall (s, s^*) \tag{1.15}$$

Osservazione 1.5.4. *La suddetta equazione prende il nome di **bilancio dettagliato**, in quanto afferma che la quantità uscente da un nodo è uguale a quella entrante (bilancio globale). È immediato il non saper distinguere, osservando la traiettoria di una catena, se si muove nel futuro o nel passato.*

Teorema 1.5.3. *(Unicità) Sia data una catena di Markov ergodica e reversibile $\mathbb{X}$ tale per cui π soddisfi (1.15), allora essa è la distribuzione invariante ed è unica.*

Dimostrazione. Supponiamo che π soddisfi la (1.15), allora (fissando uno stato s):

$$\sum_s (\pi_s \cdot h_{s,s^*}) = \sum_s (\pi_{s*} \cdot h_{s^*,s}) \rightarrow \sum_s (\pi_s \cdot h_{s,s^*}) = \pi_{s^*}$$

Essendo valida $\forall s \in S$, allora:

$$\pi H = \pi$$

■

Osservazione 1.5.5. *Sono valide le seguenti affermazioni:*

- *è immediato il fatto che la medesima proprietà di simmetria (1.5.3) sia valida con $P^{(m)}$ [1];*
- *questa considerazione ci consente di ricavare più facilmente la distribuzione stazionaria, in quanto non abbiamo dubbi di unicità.*

Capitolo 2

MCMC

2.1 Introduzione

Prima dell'avvento di metodi computazionali potenti ed accessibili, sia costruire un grande esperimento sia descrivere un fenomeno con un modello accurato richiedeva un processo lungo e complesso [18]. Per ovviare a questo problema si era soliti usare dei *modelli standard*, che permettevano di effettuare il tutto in maniera più agevole, andando però a distaccarsi da modelli piú corretti.

I metodi sotto la denominazione di **Monte Carlo** sono coloro i quali generano un insieme di realizzazioni a partire da una distribuzione nota o una nuova distribuzione, in cui le singole variabili sono i.i.d [20]. Più in generale, tali metodi fanno parte di una classe di algoritmi che, da un insieme di numeri generati in maniera casuale, permette di studiare il comportamento teorico come approssimazione del valore medio. In linea teorica, stiamo approssimando un modello complesso tramite un campione casuale. All'interno di tale classe, esistono alcuni modelli basati sulla generazione di numeri casuali originati da variabili aleatorie uniformi nell'intervallo $(0, 1)$. Per puntualizzare, tramite la distribuzione uniforme non generiamo numeri randomici ma pseudo-randomici, in quanto i computer sono delle macchine deterministiche.
Perchè utilizzare la distribuzione uniforme? [3]

[20] indipendenti ed identicamente distribuite

Teorema 2.1.1. *(Inversione) Presa una qualsiasi v.a continua $X \sim f(x)$, assumiamo che la sua funzione di ripartizione F ammetta inversa. Allora $F(x)$è distribuita come una uniforme nell'intervallo $(0, 1)$.*
In formule:

$$Y = F(X) \sim Unif(0, 1)$$

Dimostrazione.

$$\begin{aligned} P(Y \leq y) &= P(F(X) \leq y) = P(F^{-1}(F(X)) \leq F^{-1}(y)) \\ &= P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y \\ &\text{Quindi, se } Y \sim Unif(0, 1) \text{ allora } X \sim F^{-1}(Y) \end{aligned}$$

■

Questo teorema non è valido solo per variabili continue, ma è possibile trovarne un'estensione anche per variabili discrete.

Osservazione 2.1.1. *Si possono fare due considerazioni:*

- *La generazione di campioni casuali a partire da v.a uniformi è fondamentale per lo studio del comportamento dei metodi di simulazione per altre distribuzioni di probabalità, in quanto esse possono essere rappresentate come una trasformazione deterministica delle variabili uniformi di partenza;*
- *Se volessi generare un campione da una certa variabile aleatoria $X \sim f(x)$, basterebbe generare u da $U \sim Unif(0, 1)$ e applicare la trasformazione $x = F^{-1}(u)$.*

Come vedremo in seguito, non siamo interessati a tale meccanismo di generazione poichè, come accennato nel Capitolo 1, i processi Markoviani sono affetti da dipendenza stocastica. È comunque doveroso farne un breve cenno per giustificare il loro futuro utilizzo.
Entrando nello specifico, data una certa funzione di densità $f : L \to \Re$, una certa funzione $T : L \to \Re$ e un vettore n-dimensionale di v.a i.i.d $X = (X_1, \ldots, X_n)$. Il

nostro interesse è rivolto verso la valutazione di una *quantità* associata a tali variabili:

$$E_f[T(X)] = \int_{\mathfrak{X}} T(x)f(x)\ dx \quad \text{se } X_t \text{ v.a continua} \tag{2.1}$$
$$= \sum_{\mathfrak{X}} T(x)f(x) \quad \text{se } X_t \text{ v.a discreta}$$

Senza addentrarci eccessivamente nella teoria alla base del metodo, esso approssima il (2.1) mediante la generazione di un campione casuale $(X_1, \ldots, X_n)$ (dove $n >> 1$), estratto dalla densità f e proponendo uno stimatore corretto:

$$\widehat{t}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i) \qquad V[\widehat{t}_n] = E[(\widehat{t}_n - E[\widehat{t}_n])^2]$$

Per quanto riguarda la convergenza dello stimatore, sono soddisfatte le ipotesi della *Legge dei Grandi Numeri*:

$$\widehat{t}_n \to E_f[T(X)] \quad \text{q.c}$$

In merito alla distribuzione approssimata dello stimatore, sono soddisfatte le ipotesi del *Teorema del Limite Centrale:*

$$\frac{\widehat{t}_n - E_f[T(X)]}{\sqrt{V(\widehat{t}_n)}} \sim \mathcal{N}(0,1)$$

2.2 Metodi Monte Carlo per le catene di Markov

I metodi precedentemente specificati si basano sulla generazione di un campione i.i.d direttamente dalla funzione di densità f o indirettamente (come nel cosiddetto **importance sampling**) tramite l'ausilio di altre densità. Uno dei problemi principali della famiglia di metodi Monte Carlo riguarda la simulazione di quantità multivariate. Nel nostro caso specifico, andremo a generare campioni identicamente distribuiti e correlati a partire da una catena di Markov discreta, dove la correlazione del campione è dovuta alla proprietà di Markovianità. Per semplicità, cambieremo notazione per non confonderci con la situazione precedente.
Supponiamo di avere un certo vettore di stati $\theta = (\theta_1, \ldots, \theta_n)$, una certa funzione $g : \theta \to \Re$ e una funzione di densità π [21]. Per quanto il discorso possa essere

[21] si potrebbe denominare massa di probabilità in quanto lavoriamo nel campo discreto

generalizzato, supponiamo di lavorare nel discreto, tale per cui θ può assumere al più un'infinità numerabile di valori.
In molti problemi inferenziali è spesso richiesto di calcolare quantità che possono assumere la seguente forma:

$$\sum_{\theta} g(\theta)\pi(\theta) \quad \text{nel caso discreto} \tag{2.2}$$

$$\int_{\theta} g(\theta)\pi(\theta)\ d\theta \quad \text{nel caso continuo} \tag{2.3}$$

Quando la funzione π risulta essere complicata oppure θ assume un numero di valori molto elevato, allora queste quantità sono molto difficili da calcolare. Per questo, dovremo introdurre dei metodi di stima in grado di fornire un'approssimazione quanto più possibile corretta e ragionevole.
Una tecnica alternativa al classico metodo Monte Carlo riguarda un approccio basato sulle **Monte Carlo Markov Chains (McMc)**. Una prima traccia di tale metodo si trova in una pubblicazione del 1953 redatta dall'*American Institute of Physics* [14], in merito alla creazione di una *fast computing machine* [22] per investigare le proprietà relative alle equazioni di stato per interazioni molecolari.

Tale approccio richiede la costruzione di una catena di Markov con le seguenti proprietà:

- la catena presenta una distribuzione invariante π unica;
- la distribuzione invariante π coincide con quella limite;
- la matrice di transizione H ha una forma semplice.

Al fine di soddisfare la prima condizione, servirebbe una catena irriducibile con stati persistenti non nulli. Questo però non garantirebbe la coincidenza tra distribuzione limite ed invariante, per questo richiediamo la catena ergodica. Per semplificare il problema, viene preteso il soddisfacimento della proprietà di reversibilità (per via della condizione di unicità espressa nel Teorema 1.5.3) e di stazionarietà.
Prima di addentrarci nel funzionamento di tali metodi, occorre discuterne alcuni

[22] Calcolatore ad alta potenza.

aspetti probabilistici. Il principale problema riguarda la correlazione, o per meglio dire l'autocorrelazione tra le variabili, legata alle catene di Markov, la quale non permette il naturale utilizzo di 2 teoremi importanti: *Teorema del Limite Centrale* e *Legge dei Grandi Numeri*

Definizione 2.2.1. *(Autocovarianza) Data una catena di Markov* $\mathbb{X}$*, si definisce la funzione di* ***Autocovarianza*** *come:*

$$C(i,j) = Cov(X_i, X_j) \quad (i,j) \in \theta$$

Si definisce la ***Autocorrelazione*** *di una catena:*

$$\rho_{i,j} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} \in [-1,1] \tag{2.4}$$

Avendo a disposizione una catena di Markov ergodica e stazionaria, è possibile scrivere la relazione (2.4) come segue:

$$C(i, i+t) = Cov(X_i, X_{i+t}) = Cov(X_0, X_t) \ (i,t) \in \Theta$$

Tale funzione dipende esclusivamente da t, il quale prende il nome di lag [23].
In particolare, si definisce l'autocovarianza al lag t come:

$$\gamma_t = Cov(X_0, X_t) \tag{2.5}$$

Introdurre questa funzione è importante in quanto la varianza del nostro stimatore non dipende solo dalle varianze. Infatti, la varianza di una somma di variabili è uguale alla somma delle varianze solo nel caso di incorrelazione.
Prima di ragionare su questo aspetto, conviene estendere la nozione di convergenza per lo stimatore media.

Teorema 2.2.1. *(Ergodico per catene finite) [8] Sia data una catena di Markov ergodica* $\mathbb{X}$ *con* π *distribuzione invariante e una funzione* $g : \theta \to \Re$*.*
Definisco:

$$\frac{N_n(g)}{n} = \frac{1}{n}\sum_{i=1}^{n}[g(X_i)] = \textit{media dei valori assunti nel campione}$$

[23] ritardo

Allora:

$$\frac{N_n(g)}{n} \to \pi \cdot g \quad q.c \quad \textit{Oppure} \quad \frac{\sum_{i=1}^{n}[g(X_i)]}{n} \to \sum_{i\in\theta} g(i)\pi_i \quad q.c$$

Dove $\pi = (\pi_1, \ldots, \pi_n)$, $g = (g_1, \ldots, g_n)$, $\pi \cdot g$ *è il prodotto scalare tra i due vettori ed* $n >> 1$.

Dimostrazione. (Euristica) Dividiamo la dimostrazione in due casi per comprendere al meglio il funzionamento di tale teorema.

1. Supponiamo $g(X_j) = I_{(X_j=i)} \to \ g(X_j) = 1 \ \text{ se } \ X_j = i$

   Come conseguenza, $N_n(g) = n_i$ ossia il tempo trascorso nello stato i

   $\frac{N_n(g)}{n} = \frac{n_i}{n}$ non è altro che la frequenza relativa di tempo trascorso nell'unità i.

   Per via del Teorema (1.5.1), esso è approssimabile a $\frac{1}{\mu_i}$.

   Allora:

   $$\frac{N_n(g)}{n} \to \frac{1}{\mu_i} = \pi_i \quad \text{q.c}$$

2. Nel caso completo, proviamo a generalizzare il risultato. Per via del ragionamento precedente possiamo affermare:

   $$\begin{aligned}\frac{N_n(g)}{n} &= \frac{n_0 g(0) + \ldots + n_n g(n)}{n} \qquad n_i = \text{numero di passaggi per unità i} \\ &\approx \frac{1}{\mu_0} g(0) + \ldots + \frac{1}{\mu_n} g(n) \\ &= \sum_{i\in\theta} g(i)\pi_i \quad q.c\end{aligned}$$

■

Abbiamo quindi definito uno stimatore per la quantità (2.2) di interesse.
Nel nostro caso:

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to E_\pi(g) \ \text{ q.c} \tag{2.6}$$

Tale somma è approssimabile attraverso il tempo di soggiorno in un determinato stato e la funzione calcolata in quello stato. Sembra essere una diretta conseguenza del teorema ergodico, in quanto π_i può essere interpretata come la probabilità che lo stato i venga visitato. In aggiunta, tale probabilità non dipende dallo stato di partenza.

Molto più complicata è l'estensione del TLC, in quanto richiede condizioni molto più specifiche. Per spiegazioni più approfondite è possibile consultare il [15].

Teorema 2.2.2. *(TLC) Data una catena di Markov $\mathbb{X}$ geometricamente ergodica* [24]*, reversibile con $E_\pi(g^2) < \infty$, con distribuzione invariante π e una funzione $g : \Theta \to \Re$ ($n >> 1$):*

$$\sqrt{n}(\bar{g}_n - E_\pi(g)) \sim N(0, \sigma_g^2) \tag{2.7}$$

Osservazione 2.2.1. *Per ragioni pratiche, la trattazione di σ_g^2 verrà attuata più avanti.*

Teoricamente, per via del Teorema Ergodico 1.5.2 le due relazioni dovrebbero essere vere a prescindere dalla distribuzione iniziale $u^{(0)}$.

Osservazione 2.2.2. *Sorgono spontanee due domande:*

- *cosa garantisce che lo starting point scelto sia generato a partire dalla distribuzione stazionaria?;*
- *nella pratica, cosa garantisce che $u^{(0)} = \pi$?*

Da un punto di vista teorico, come accennato precedentemente, imponiamo che la catena possegga la proprietà di stazionarietà. Come visto nella definizione 1.5.2, tale proprietà garantisce che $u^{(0)}$ coincida con la distribuzione stazionaria. Da un punto di vista pratico, si potrebbero applicare diverse tecniche volte a risolvere questo problema. Una delle più semplici ed intuitive è la cosiddetta **Burn-in** [25].

[24] la convergenza verso la distribuzione stazionaria avviene ad un tasso geometrico

[25] autocombustione della catena

Essa prevede l'eliminazione delle prime $p-1$ iterazioni della catena in modo che ci si avvicini sufficientemente verso la distribuzione stazionaria al tempo p. Per via della mancanza di memoria della catena di Markov, scartare i primi $p-1$ passi è giustificabile se reinizializzassimo la simulazione a partire dal p-esimo passo.
Intuitivamente, questa tecnica ci permette di entrare con maggiore probabilità in una regione dove gli stati sono più rappresentativi della densità che vorremmo approssimare. Inoltre, permette di superare il problema della scelta dello starting point. Nella maggior parte dei casi, il p-esimo è sicuramente uno stato che verrà visitato più spesso rispetto allo stato scelto inizialmente. Tale constatazione è importante: infatti il TLC e la LGN dovrebbero valere a prescindere dallo stato di partenza. Tuttavia, da un punto di vista teorico, nulla assicura che tale intuizione si verifichi. Per questo si è soliti dire che tale metodo è solo uno dei possibili da utilizzare (neanche particolarmente buono).
Esistono almeno due problemi: il primo è legato al numero di iterazioni da scartare, poiché non esiste una regola universale. Si è soliti dire che bisogna scartare un numero *sufficientemente* grande di stati. In aggiunta, se la numerosità campionaria non è troppo elevata, non si possono eliminare troppe iterazioni senza avere un contrappasso in termini di bontà dell'approssimazione.

2.2.1 Stimatore della varianza

La varianza dello stimatore $\bar{g}_n$ (2.6) è utile per capirne la precisione e la velocità di convergenza verso il valore desiderato. Rispetto al caso di variabili i.i.d, bisogna tener conto del fattore di covarianza tra le variabili.

Teorema 2.2.3. *Data una catena di Markov ergodica $\mathbb{X}$, si definisce la varianza dello stimatore $\bar{g}_n$ come:*

$$var(\bar{g}_n) = \gamma_0 + 2\sum_{k=1}^{n-1} \frac{n-k}{k}\gamma_k \tag{2.8}$$

Inoltre, se la catena è geometricamente ergodica e reversibile con $E_\pi(g^2) < \infty$ per $n \to \infty$:

$$n \cdot var(\bar{g}_n) \approx \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \tag{2.9}$$

dove γ_0 *e* γ_k *sono le funzioni di autocovarianza (definita nella formula* (2.5)*) rispettivamente al lag* 0 *e al lag* k*.*

Dimostrazione.

$$var(\bar{g}_n) = var\left(\frac{1}{n}\sum_{i=1^n} g(X_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\Big(g(X_i), g(X_j)\Big)$$
$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} var\Big(g(X_i)\Big) + 2\sum_{i=i}^{n-1}\sum_{j=i+1}^{n} Cov\Big(g(X_i), g(X_j)\Big)\right]$$

Dato che la catena è ergodica, essa ammette distribuzione invariante e limite π. Questo implica che nè varianza nè covarianza dipendono più da n poiché restano costanti al passare del tempo.

$$var(\bar{g}_n) = \frac{1}{n^2}\left[n \cdot var\Big(g(X_j)\Big) + 2\sum_{k=1}^{n-1}(n-k)\cdot Cov\Big(g(X_j), g(X_{j+k})\Big)\right]$$
$$= \frac{1}{n^2}\left[n\cdot\gamma_0 + 2\sum_{k=1}^{n-1}(n-k)\cdot\gamma_k\right]$$

Il precedente passaggio sfrutta la (2.5)

$$n \cdot var(\bar{g}_n) = \gamma_0 + 2\sum_{k=1}^{n-1}\Big(\frac{n-k}{n}\Big)\gamma_k$$
$$\text{se n} \to \infty \quad n\cdot var(\bar{g}_n) \approx \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k = \sigma_g^2$$

■

Osservazione 2.2.3. *Seguono due semplici osservazioni:*

- *al crescere del lag, il fattore di autocovarianza tende ad annullarsi;*
- *Essendo impossibile l'osservazione di tale quantità, bisogna trovare un modo per stimarla. Studiamo quindi la* ***diagnostica*** *del metodo.*

Esistono diversi modi per stimare σ_g^2, uno dei più semplici ed intuitivi è conosciuto come **Batch Means method** [26] [4]. Consiste nel dividere la catena in segmenti consecutivi di lunghezza m, chiamati grappoli, mediante i quali dare una stima della

[26] media dei grappoli

varianza.

In formule:

$$Z_1 = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} \quad Z_2 = \begin{pmatrix} X_{m+1} \\ \vdots \\ X_{2m} \end{pmatrix} \dots \quad Z_j = \begin{pmatrix} X_{m(j-1)+1} \\ \vdots \\ X_{mj} \end{pmatrix} \dots \quad Z_n = \begin{pmatrix} X_{m(n-1)+1} \\ \vdots \\ X_{mn} \end{pmatrix}$$

Intuitivamente, è come se la popolazione composta da N unità fosse ripartita in n gruppi composti da m unità.

Ciascuna Z_j con $j = 1, \dots, n$ forma una catena di Markov, tale per cui si può definire la media del grappolo come:

$$B_j = g(Z_j) = \frac{1}{m} \sum_{i=1}^{m} g(X_{m(j-1)+i})$$

Per il Teorema 2.2.1, sono valide le seguenti convergenze:

$$B_j \to E_\pi(g_j) = \mu \quad \text{q.c} \quad j = 1, \dots, m$$
$$\bar{B}_n = \frac{1}{n} \sum_{j=1}^{n} B_j \to E_\pi(g) = \mu \quad \text{q.c}$$

Inoltre, per quanto riguarda la varianza, dalla definizione 2.2.3 segue che:

$$s^2_{batch} = \frac{1}{n} \sum_{j=1}^{n} (B_j - \bar{B}_n)^2 \to var(B_j) \approx \frac{\sigma^2_m}{m}$$

Per m sufficientemente grande:

$$\frac{\sigma^2_m}{m} \approx \frac{\sigma^2_g}{m}$$

Allora, per via del Teorema 2.2.2:

$$\bar{g}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2_g}{n}\right) \approx \mathcal{N}\left(\mu, \frac{m}{n} s^2_{batch}\right)$$

Osservazione 2.2.4. *Lo stimatore media del grappolo B_j ha approssimativamente la stessa varianza dello stimatore media generale, a meno di un fattore correttivo che dipende dalla diversa lunghezza tra i grappoli e la catena. La varianza dei singoli grappoli è stimata tramite la varianza del campione estratto.*

Esponiamo il principale problema pratico: Quanti grappoli bisogna considerare? Quale deve essere la loro numerosità?

Paradossalmente, le due domande sono in conflitto. Euristicamente, vorremmo una numerosità per grappolo elevata al fine di riuscire a soddisfare la relazione $\sigma_m^2 \approx \sigma_g^2$. Quanto elevata? È difficile rispondere a priori a questa domanda, in quanto molto spesso non conosciamo al meglio il problema sulla catena di Markov. D'altro canto, vorremmo un numero di grappoli elevato, al fine di approssimare al meglio la varianza dello stimatore. Teoricamente, sarebbero necessari almeno 20 grappoli. Questo genera un conflitto. A meno che la numerosità campionaria non sia veramente elevata, è piuttosto difficile soddisfarli entrambi.

In generale, si tende a non superare i 30 grappoli che permettono una stima più che soddisfacente. Non bisogna esagerare con il numero di grappoli poichè potremmo incorrere in un problema: generiamo in maniera indipendente. Per questo, una delle possibili diagnostiche riguarda il controllo del fattore di autocovarianza al lag k tra gli stimatori media dei grappoli (2.5). Se tale fattore assume un valore basso per lag piccoli vuol dire che vi è un problema nel numero di grappoli selezionati.

Capitolo 3

Algoritmo di Metropolis-Hastings

3.1 Funzionamento del metodo

L'algoritmo di Metropolis-Hastings è uno dei più popolari metodi di simulazione nella statistica. Venne introdotto nel 1953 da Robert Metropolis [27] [14], il quale, insieme ad altri scienziati, si trovava presso i laboratori di Los Alamos. Solo in seguito, più precisamente nel 1970, Hastings ne diede una generalizzazione, permettendone quindi una maggiore applicabilità. Tale metodo può essere definito come il *cavallo di battaglia* degli McMc, per via della sua estrema semplicità e versatilità. Per usare un'analogia, esso permette di esplorare il quadro generale in maniera locale e graduale piuttosto che darne una rappresentazione globale ma sfocata.

Entrando nel dettaglio, sia data una certa distribuzione $f(\theta)$ ed una distribuzione π tale che $f(\theta) \propto \pi$. Esse differiscono solamente per una costante moltiplicativa, in quanto il nucleo della distribuzione è lo stesso. Nello specifico, π prende il nome di *Distribuzione Target* mentre $\theta = (\theta_1, \ldots, \theta_n) \in \Theta$, dove Θ può assumere al limite un'infinitá numerabile di valori. L'intuizione proposta da Metropolis-Hastings riguarda la costruzione di una catena di Markov Ergodica $\mathbb{X}$ sullo spazio degli stati Θ con π distribuzione invariante per le motivazioni espresse nella sezione 2.2.

[27] collaborarono anche i coniugi Rosenbluth e i coniugi Teller

In aggiunta, è richiesto il soddisfacimento della condizione di reversibilità (Teorema (1.5.3)) e di stazionarietà (Definizione 1.5.2). Essendo lo spazio degli stati discreto, anche la catena di Markov è discreta.
Sono quindi validi i teoremi 2.2.1 e 2.2.2, i quali, come visto precedentemente, non dipendono dallo starting point. Per tutte le considerazioni fatte nel Capitolo 2, simulare una catena di Markov è intrinsecamente simile ad una simulazione i.i.d dalla distribuzione target, tenendo conto della perdita di efficienza (come visto nella definizione 2.2.3). Per diminuire tale perdita è richiesta la simulazione di un numero maggiore di elementi interni al campione, al fine di poter disporre di un'approssimazione migliore della suddetta distribuzione.
Tornando all'esempio di prima, la costruzione di una catena di Markov ergodica permette un'esplorazione locale di tutti gli stati per poter coprire completamente la regione di interesse. Alla catena è associata una certa matrice di transizione H, la quale è difficile da ricavare per una qualche motivazione. H non può essere scelta in maniera casuale, poiché la sua distribuzione limite deve essere proprio π. Per il Teorema 1.5.2, la distribuzione limite coincide con quella invariante. H può essere decomposta in 2 parti:

- una matrice di transizione Q, definita *Distribuzione proposta* (proposal distribution);
- una matrice A, la quale contiene le probabilità di accettazione-rifiuto.[28]

Q deve avere una forma semplice tale per cui è possibile generare dei campioni a partire da tale distribuzione. Tale matrice è associata ad un'altra catena di Markov definita sullo stesso spazio degli stati, sulla quale non valgono le stesse proprietà di prima. Infatti, se la catena associata a tale matrice fosse ergodica, allora useremmo direttamente Q per approssimare π, senza dover necessariamente generare H.
Una volta proposto un campione, tramite la matrice A siamo in grado di accettare o rifiutare con una certa probabilità di *movimento* $a_{i,j}$. Essa permette di delineare il percorso della catena in quanto è la probabilità di accettare il campione proposto tramite $q_{i,j}$.

[28] essendo una probabilità $a_{i,j} \in [0, 1]$

In formule:

$$h_{i,j} = \begin{cases} q_{i,j} \cdot a_{i,j} & i \neq j \\ 1 - \sum_n (q_{i,n} \cdot a_{i,n}) & i = j \end{cases} \tag{3.1}$$

$$X_{t+1} = \begin{cases} j & \text{con probabilità } a_{i,j} \\ X_t & \text{con probabilità } 1 - a_{i,j} \end{cases} \tag{3.2}$$

dove X_t rappresenta lo stato che la catena ha visitato l'istante precedente.
Dati 2 stati $(i, j) \in \Theta$, con $i \neq j$, come conseguenza del Teorema 1.5.3:

$$\pi_i \cdot h_{i,j} = \pi_j \cdot h_{j,i} \rightarrow \pi_i \cdot q_{i,j} \cdot a_{i,j} = \pi_j \cdot q_{j,i} \cdot a_{j,i}$$
$$a_{i,j} = a_{j,i} \cdot \frac{\pi_j}{\pi_i} \cdot \frac{q_{j,i}}{q_{i,j}}$$

per soddisfare l'equazione di bilancio dettagliato:

$$a_{i,j} = min\left(1, \frac{\pi_j}{\pi_i} \cdot \frac{q_{j,i}}{q_{i,j}}\right)$$

stiamo quindi affermando che $a_{i,j} = 1$ se $\pi_j \cdot q_{j,i} \geq \pi_i \cdot q_{i,j}$. In particolar modo, il rapporto $\frac{q_{j,i}}{q_{i,j}}$ prende il nome di *Hastings-ratio*, il quale venne introdotto per generalizzare tale metodo per ogni tipo di distribuzione Q [10]. Infatti, se la distribuzione Q è simmetrica allora $\frac{q_{j,i}}{q_{i,j}} = 1$ e quindi $a_{i,j} = 1$ se $\pi_j \geq \pi_i$. Se la catena associata a Q passa ad uno stato che presenta un tempo di soggiorno maggiore allora l'algoritmo accetta sempre il candidato proposto.

Le probailità di accettazione-rifiuto non solo permettono di preservare la distribuzione invariante, ma possono essere anche usate per valutare la performance dell'algoritmo. Si può definire il **tasso di accettazione**, ossia un resoconto del numero di volte in cui lo stato proposto viene accettato. L'obiettivo è di trovare quella distribuzione proposta che mi permette di avere un tasso di accettazione tra un quarto ed un mezzo [7] (non è una regola universale). Da una parte, un alto tasso di accettazione potrebbe essere un segnale di una pessima convergenza verso la distribuzione stazionaria in quanto la funzione da approssimare apparirebbe "piatta". Dall'altra parte, un basso tasso di accettazione potrebbe non permettere la visita di tutti gli stati; come conseguenza bisognerebbe simulare un campione con alta

numerosità per poter esplorare completamente la catena.
Per quanto la scelta di Q sia casuale, dovremmo almeno essere in grado di poter visitare (almeno in teoria) tutti gli stati della catena, scegliendone una irriducibile.

Esempio 3.1.1. *Sia π la distribuzione stazionaria e $S = \{1, 2, 3, 4\}$ lo spazio degli stati, tale per cui $\pi = \left(\frac{1}{12}, \frac{4}{12}, \frac{5}{12}, \frac{2}{12}\right)$. Nonostante il metodo si applichi per distribuzioni complesse, ne utilizziamo una facile per spiegarne il funzionamento. Ci appoggiamo quindi ad una catena di Markov a tempo discreto con stessa S, ma con una matrice di transizione Q nota e semplice da cui proporre campioni (simmetrica).*

$$Q = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{2}{5} \end{bmatrix}$$

Supponiamo che lo starting point sia lo stato 3. Al passo 1 la catena propone lo stato 1 con probabilità di accettazione pari ad $\frac{1}{5}$. Si estrae $u = \frac{1}{7}$ da un'uniforme in $(0, 1)$ e quindi la proposta viene accettata poiché u è minore della probabilità di accettazione. Al passo 2 la catena propone lo stato 4 con $a_{i,j} = 1$. In questo caso sicuramente il candidato proposto verrà accettato.

3.1.1 Definizione Algoritmo

Una volta compreso il funzionamento di tale metodo da un punto di vista probabilistico, è necessario formalizzare l'algoritmo. Nella pratica, come funziona il meccanismo di generazione basato sulle McMc?
Algoritmo:

1. Al tempo t uguale ad 1 si sceglie randomicamente uno stato di partenza θ_0 [29];
2. Viene proposto un candidato θ_t in accordo con q_{θ_0,θ_t};
3. Si calcola la probabilità di accettare a_{θ_0,θ_t};
4. Si estrae u da una Uniforme in $(0, 1)$. Se $u \leq a_{\theta_0,\theta_t}$ allora si accetta il campione proposto θ_t, altrimenti viene mantenuto θ_0;

[29] operativamente applico la tecnica del burn-in

5. Si ricomincia dal punto 2 fino a che non si estrae un campione n-dimensionale.

Una volta finito il ciclo, possediamo un campione n-dimensionale in grado di approssimare la densità di partenza $f(\theta)$. Nel punto 4 si utilizza una distribuzione uniforme per le ragioni spiegate nel Teorema 2.1.1.

Esempio 3.1.2. *(Normale Univariata) Sia data una distribuzione normale univariata con media uguale* 30 *e varianza uguale* 4*. Il nostro obiettivo è riuscire ad approssimare tale densità. Essendo il supporto continuo, andrà discretizzato.*

```
funzione_target=function(x){
  (1/sqrt(2*pi*2^2))*exp(-(x-30)^2/(2*2^2))
}
```

Come funzione proposta utilizziamo una normale univariata, anch'essa con supporto discretizzato, di varianza unitaria e media variabile (funzione simmetrica). La media di tale distribuzione verrà assunta dall'ultimo stato accettato per approssimare la densità. Quest'ultima considerazione è la riprova empirica che tale algoritmo si muove in maniera locale per ricostruire il quadro generale.

```
funzione_proposta=function(x){
  rnorm(1,mean = x,sd=1)
}
```

La varianza è definita lo ***step size*** *(passo di campionamento) in quanto definisce un insieme di stati che presentano maggiore probabilità di essere la vera media della distribuzione target.*

$$P(\mu - \sigma < x < \mu + \sigma) = 0.68 \quad P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.96$$

Scegliendo uno step size troppo elevato si rischia che la traiettoria della catena passi per un numero elevato di stati, influenzando negativamente l'approssimazione (bisogna estrarre un campione più grande). Viceversa, con uno step size piccolo si rischia di visitare sempre gli stessi stati. Il problema più grande è quello di proporre stati che si trovano vicini allo starting point, senza riuscire mai ad allontanarsi.
Continuando con il procedimento, estraiamo casualmente 4 *stati e simuliamo una*

catena di Markov prendendo come stato iniziale uno dei 4 *a turno. Come ultimo passo, sceglieremo l'ultimo valore assunto da ciascuna simulazione come starting point per la successiva generazione degli elementi che compongono i campioni (Burn-In).*

```
l=c(-30,0,25,50)
k=c()
burn_in=function(x){
  for (j in 1:length(x)) {
    for (i in 1:1000) {
      xprimo=funzione_proposta(x[j])
      ratio=min(1,funzione_target(xprimo)/funzione_target(x[j]))
      prob= runif(1)< ratio
      montecarlo[i]=ifelse(prob,xprimo,x[j])
      x[j]=montecarlo[i]
    }
    k[j]=montecarlo[1000]
  }
  return(k)
}
starting=round(burn_in(l),1)
```

Nel precedente codice è spiegato il metodo utilizzato per generare un campione da una distribuzione proposta.

Come si è comportato l'algoritmo?

Tutto sommato, con la sola generazione di un campione di 3000 *elementi, in tutti e* 4 *i casi riusciamo ad approssimare la densità desiderata (con qualche piccolo errore). Ovviamente, se avessimo generato un campione con una dimensione maggiore, l'approssimazione sarebbe stata più accurata, a discapito però del calcolo computazionale. Anche in un caso semplificato come questo, si capisce l'enorme influenza dello starting point.*

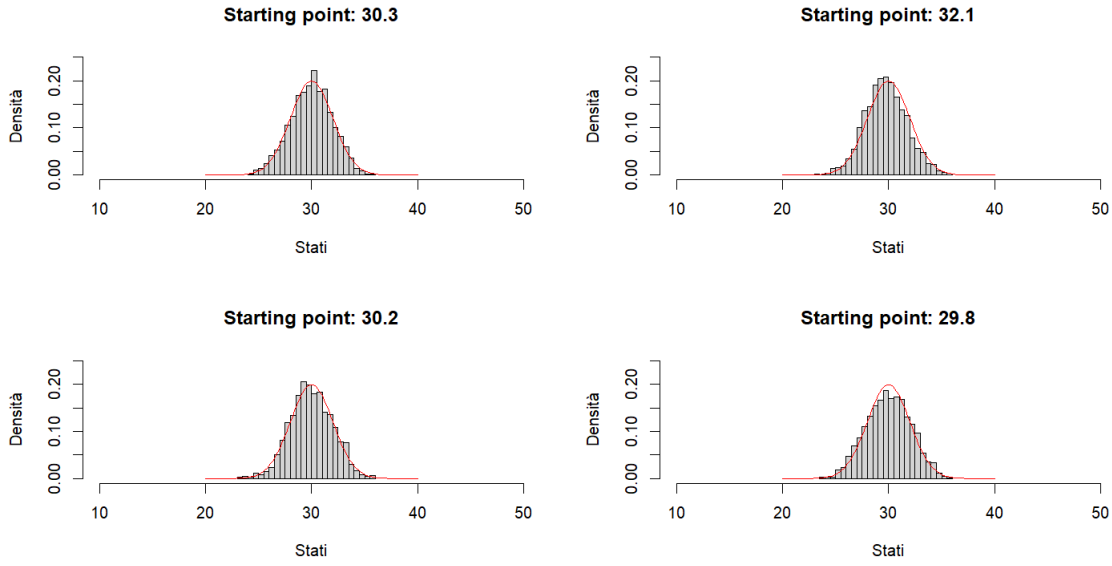


Figura 3.1. Istogramma delle simulazioni per diversi starting points

Qual è la traiettoria della catena? È facile notare che la catena assume, indifferentemente dallo starting point, valori compresi tra 24 *e* 36*. Questo è l'intervallo in cui si troverebbero il* 96% *delle realizzazioni se generassimo dalla distribuzione target. A parità di numerosità campionaria, estrarre dalla distribuzione target in maniera i.i.d permette di avere un'accuratezza maggiore.*

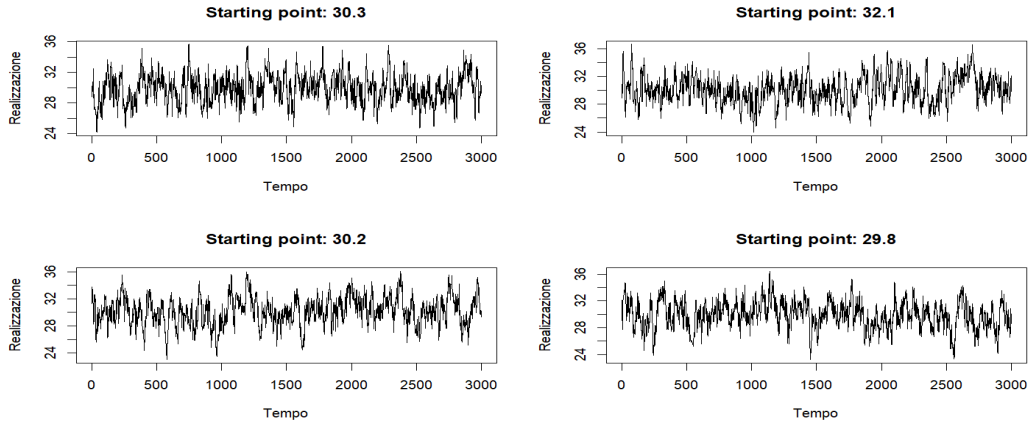


Figura 3.2. Traiettoria della catena per diversi starting points

La differenza principale tra i due metodi di generazione riguarda la dipendenza stocastica che esiste tra un elemento ed un altro all'interno del campione. Prendendo in considerazione solo una delle 4 *simulazioni, come si comporta l'autocorrelazione al crescere del lag? Poiché tutte le variabili sono correlate, il processo si definisce ad* ***oscillazioni lente****. Questo implica che l'algoritmo tende ad esplorare lo spazio degli stati in maniera graduale.*

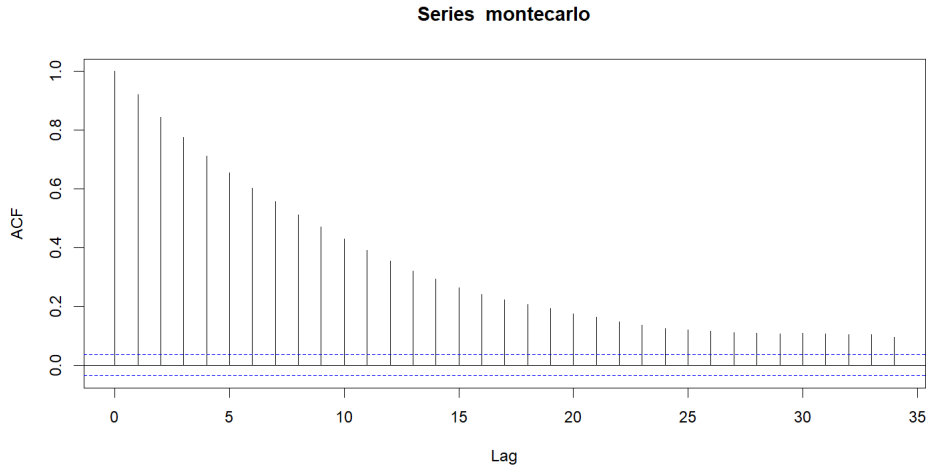


Figura 3.3. Autocorrelazione tra valori assunti dal campione per diversi lag

L'ultimo quesito rimasto in sospeso riguarda la stima della varianza dei valori assunti dalla catena (dal capitolo precedente sappiamo che è possibile utilizzare il ***Batch means method****). Essa assume il valore di* 3.76*, molto simile alla varianza della distribuzione target. Se iterassimo questa procedura un numero sufficiente di volte, otterremmo una stima corretta.*

```
N=length(montecarlo)
n=30
m=N/n
stimatore_varianza=function(a){
  vari=(1/n)*sum((a-mean(a))^2)*m
  out=vari
  return(out)
}
stimatore_varianza(batches)/n #3.76
```

3.2 Problemi algoritmo

Dopo aver introdotto e spiegato nel dettaglio l'algoritmo per approssimare una densità tramite una catena di Markov, non resta altro da fare se non studiarne i possibili problemi. Per semplicità, nella trattazione si eviterà il discorso riguardo la complessità (computazionale) dell'algoritmo e del tasso di convergenza verso la distribuzione stazionaria.
Tra i principali problemi troviamo:

- forma della distribuzione proposta;
- step size;
- scelta dello starting point;
- forma della distribuzione target.

Sebbene abbiano tutti e quattro eguale importanza, l'ultimo è l'unico a non essere mai stato trattato, mentre gli altri hanno trovato un loro spazio all'interno dell'elaborato. In realtà, non possiamo trattarlo singolarmente poichè dipende dai primi tre.
Com'è possibile che la forma della distribuzione target influenzi il funzionamento del metodo? Non dovrebbe funzionare indifferentemente dal tipo di distribuzione da approssimare?
Per dare una risposta completa a queste domande è necessario utilizzare un esempio, poiché consente di capire concretamente il problema.

Esempio 3.2.1. *(Mistura Di Normali) Sia data una distribuzione ottenuta come mistura di due normali con media rispettivamente pari a* 30 *e* 10 *e con varianza pari rispettivamente a* 4 *e* 6*. Il peso associato a ciascuna distribuzione è lo stesso. In teoria, si ottiene una distribuzione bimodale, con media pari alla media delle due medie.*

```
funzione_target=function(x){
  out=((1/2)*(1/sqrt(2*pi*4))*exp(-(1/2)*((x-10)^2)*(1/4)))+
  ((1/2)*(1/sqrt(2*pi*6))*exp(-(1/2)*((x-30)^2)*(1/6)))
}
```

Come funzione proposta utilizziamo una normale con media variabile (vale lo stesso ragionamento del precedente esempio) e varianza pari a sei. Ancora una volta è stata scelta una distribuzione appartenente alla famiglia Normale. Bisogna impostare un passo di campionamento più elevato rispetto al caso precedente per poter esplorare tutti gli stati della catena. In qualche modo, questo è un problema legato alla forma particolare della distribuzione target.

```
funzione_proposta=function(x){
  rnorm(1,mean = x,sd=3)
}
```

Applicando lo stesso procedimento dell'altro esempio, si utilizza la tecnica del Burn-In per ottenere starting points più "ragionevoli". In particolare:

```
starting # 9.707799 10.502720 31.252346 29.822184
```

Rispetto al caso precedente, tali valori non si concentrano intorno ad un solo punto, bensì sono vicini alle due mode della distribuzione target.
Come si comporta l'algoritmo per i diversi starting points (con numerosità fissata pari a mille)? Piuttosto male, infatti solo in un caso su quattro si riesce ad approssimare la densità in maniera corretta. Per quanto lo step size sia elevato, scegliere un punto di partenza troppo vicino ad uno dei due punti di massimo (locale) influenza troppo il percorso della catena. Abbiamo dimostrato il fallimento empirico del metodo Burn-In in caso di distribuzioni target con più punti di massimo (limite applicativo).

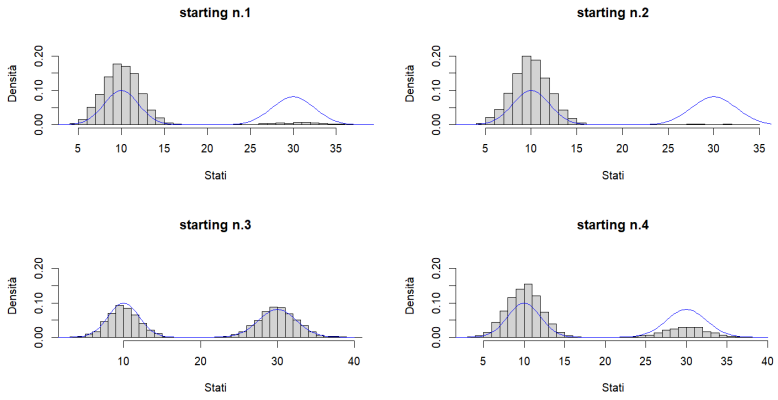


Figura 3.4. Simulazioni per i diversi starting points nel caso di mistura di normali

Qual è il percorso che segue la catena? Tale percorso è influenzato, a parità di step size, dalla numerosità del campione e dal punto di partenza. Appare logico affermare che al crescere della numerosità campionaria l'approssimazione migliora, come è possibile vedere nel grafico 3.5. Tuttavia, la scelta della numerosità adatta è legata allo starting point. Nel nostro caso, è stato scelto pari a venti poiché equidistante dalle due medie. Sicuramente è la scelta più logica, ma nulla assicura che sia la migliore.

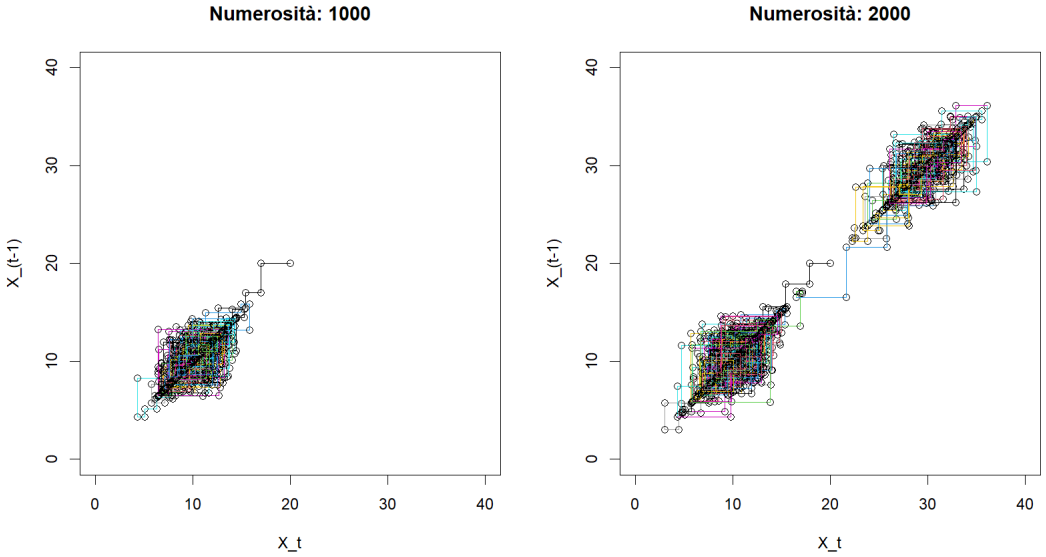


Figura 3.5. Percorso della catena per diverse numerosità campionarie

In conclusione, per distribuzioni target con più mode l'algortimo ha molta difficoltà ad approssimare la densità. In questi casi si predilige l'utilizzo di altri algoritmi, come ad esempio il ***Gibbs sampler****.*
Tutti i problemi, come precedentmente dimostrato, sono legati e risulta difficile trovare il responsabile di un ipotetico fallimento nell'approssimazione.

Bibliografia

[1] David Aldous e James Fill. *Reversible Markov chains and random walks on graphs*. 1995.

[2] Francesco Battaglia. *Metodi di previsione statistica*. Springer Science & Business Media, 2007.

[3] George Casella e Edward I George. «Explaining the Gibbs sampler». In: *The American Statistician* 46.3 (1992), pp. 167–174.

[4] Saptarshi Chakraborty, Suman K Bhattacharya e Kshitij Khare. «Estimating accuracy of the MCMC variance estimator: a central limit theorem for batch means estimators». In: *arXiv preprint arXiv:1911.00915* (2019).

[5] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[6] G Dall'Aglio. «Calcolo delle probabilità. 2003 Terza Edizione». In: *Zanichelli, Bologna* (), pp. 185–186, 280–295.

[7] Andrew Gelman, Walter R Gilks e Gareth O Roberts. «Weak convergence and optimal scaling of random walk Metropolis algorithms». In: *The annals of applied probability* 7.1 (1997), pp. 110–120.

[8] Charles J Geyer. «Markov chain Monte Carlo lecture notes». In: *Course notes, Spring Quarter* 80 (1998).

[9] Geoffrey Grimmett e David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

[10] W Keith Hastings. «Monte Carlo sampling methods using Markov chains and their applications». In: (1970).

[11] GIORGIO Israel. «Il determinismo meccanico e il suo ruolo nelle scienze». In: *Caso, necessità, libertà, Cuen, Napoli* (1998), pp. 45–62.

[12] Samuel Karlin. *A first course in stochastic processes*. Academic press, 2014.

[13] Pierre Simon marquis de Laplace. *Essai philosophique sur les probabilités*. Bachelier, 1840.

[14] Nicholas Metropolis et al. «Equation of state calculations by fast computing machines». In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[15] Sean P Meyn e Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[16] Friedrich Nietzsche. *La nascita della tragedia*. Feltrinelli Editore, 2015.

[17] James R Norris e James Robert Norris. *Markov chains*. 2. Cambridge university press, 1998.

[18] Christian P Robert, George Casella e George Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.

[19] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[20] Paul C Shields. *The ergodic theory of discrete sample paths*. Vol. 13. American Mathematical Soc., 1996.