UNIVERSITÀ DI PISA

DIPARTIMENTO DI MATEMATICA

Corso di Laurea Triennale in Matematica

Un metodo Markov Chain Monte Carlo: il campionamento di Gibbs

Relatore:
Prof. Dario Trevisan

Candidato:
Cecilia Marchi

ANNO ACCADEMICO 2020/2021

Indice

Introduzione

La tesi ha l'obiettivo di presentare il campionamento di Gibbs, un algoritmo appartenente alla classe dei metodi Markov Chain Monte Carlo (MCMC), introdotto nel 1984 dai fratelli Geman.
Il primo capitolo è un'introduzione alla teoria delle catene di Markov realizzato seguendo il libro di G. Modica e L. Poggiolini ([2]). Viene trattato in particolar modo il concetto di distribuzione invariante, e ne viene data una caratterizzazione grazie ad un teorema di convergenza. Nel libro si dimostra la convergenza delle potenze di una matrice di transizione finita e regolare utilizzando il teorema di Perron-Frobenius, ma del quale è omessa la dimostrazione. Invece, alla fine del primo capitolo di questo lavoro, si arriverà allo stesso risultato, ma evitando il teorema.
Nel secondo capitolo, basato sugli articoli di Roberts e Smith ([3]) e di Gelfand ([1]), viene illustrato il campionamento di Gibbs, ed inoltre è descritto l'algoritmo di Metropolis-Hastings, uno dei più noti metodi MCMC. Nell'articolo di Roberts e Smith sono discusse le condizioni sufficienti per la convergenza degli algoritmi. In questo elaborato viene riportato il tutto semplificato al caso in cui lo spazio degli stati del processo è discreto. Inoltre è presente un'applicazione del metodo di Gibbs, incluso il codice in MATLAB usato per l'implementazione e alcuni grafici che ne mostrano la convergenza.
L'ultima parte è dedicata al metodo di Gibbs applicato a densità gaussiane multivariate. Nel caso bidimensionale si possono trovare delle formule esplicite per gli errori di approssimazione dell'algoritmo, utili a stimare la velocità di convergenza. Nel caso d-dimensionale, con $d > 2$, sono riportate alcune simulazioni a supporto di una congettura che estende il risultato rigoroso di convergenza bidimensionale.

Capitolo 1

Catene di Markov a tempi discreti

1.1 Processi stocastici

Definizione 1.1. Un *processo stocastico* è una famiglia di variabili aleatorie $(X_t)_{t\in T}$ sullo spazio di probabilità $(\Omega, \mathcal{E}, \mathbb{P})$ e tutte a valori in uno stesso spazio misurabile $(S, \mathcal{S})$. $T \subset \mathbb{R}$ è detto *insieme dei tempi*, e S è l'*insieme degli stati*.

Fissato $\omega \in \Omega$, la *traiettoria* o *cammino* di ω è la funzione $t \mapsto X_t(\omega)$.
Un processo stocastico si può interpretare anche come una variabile aleatoria a valori nelle traiettorie, cioè $X : \Omega \to S^T$, dove S^T è l'insieme delle funzioni da T in S, munito della σ-algebra prodotto (eventualmente infinito) $\mathcal{S}^{\otimes T}$.
Infine, si può pensare al processo come una mappa $X : T \times \Omega \to S$ tale che $X(t, \omega) = X_t(\omega)$ per ogni $t \in T$ e $\omega \in \Omega$.

Il processo si dice *a tempi discreti* se T è discreto, mentre si dice a *tempi continui* se T è un intervallo (anche illimitato). Inoltre se S è finito o numerabile, allora il processo si dice *discreto* o *a stati discreti*, altrimenti è detto *a stati continui*.

In questo lavoro verranno considerati soltanto processi stocastici a tempi discreti, e in questo capitolo viene posta l'attenzione sui processi a stati discreti.
Si può assumere, senza perdita di generalità, $T = \mathbb{N}$ o $T = \{1, \dots, d\}$, e $S = \mathbb{N}$ o $S = \mathbb{Z}$, oppure, nel caso finito, $S = \{1, \dots, N\}$.

Sia $\{X_n\}_{n\in\mathbb{N}}$ un processo stocastico discreto con S l'insieme degli stati.

Definizione 1.2. Per ogni $n \geq 1$ la matrice $\mathbf{P}(n) = (\mathbf{P}(n)^i_j)_{i,j\in S}$ definita da

$$\mathbf{P}(n)^i_j := \begin{cases} \mathbb{P}(X_n = j \mid X_{n-1} = i) & \text{se } \mathbb{P}(X_{n-1} = i) > 0 \\ \delta^i_j & \text{se } \mathbb{P}(X_{n-1} = i) = 0 \end{cases}$$

è detta *matrice di transizione* del processo $\{X_n\}$ al passo n.

$\mathbf{P}(n)$ è una matrice stocastica, ovvero una matrice quadrata (eventualmente con infinite righe e colonne [1]) con tutte le entrate maggiori o uguali a zero e tale per cui la somma degli elementi in ogni riga è uguale a uno.
Infatti, è chiaro che $\mathbf{P}(n) \geq 0$. Inoltre, se $\mathbb{P}(X_{n-1} = i) = 0$, allora

$$\sum_{j\in S} \mathbf{P}(n)^i_j = \delta^i_i = 1,$$

[1] In questo caso le regole per le operazioni tra matrici rimangono le stesse del caso finito, l'unica differenza è che le somme diventano serie.

invece se $\mathbb{P}(X_{n-1} = i) > 0$, usando la legge delle probabilità totali, si ha

$$\sum_{j\in S} \mathbf{P}(n)^i_j = \sum_{j\in S} \mathbb{P}(X_n = j \mid X_{n-1} = i) = \sum_{j\in S} \frac{\mathbb{P}(X_n = j, X_{n-1} = i)}{\mathbb{P}(X_{n-1} = i)} = \frac{\mathbb{P}(X_{n-1} = i)}{\mathbb{P}(X_{n-1} = i)} = 1.$$

Rappresentiamo le densità marginali del processo con dei vettori riga $\pi(n) = (\pi(n)_i)_{i\in S}$ tali che $\pi(n)_i = \mathbb{P}(X_n = i)$.

La matrice $\mathbf{P}(n+1)$ mette in relazione le leggi (marginali) di X_n e di X_{n+1}. Infatti,

$$\pi(n+1)_j = \sum_{i\in S} \mathbb{P}(X_{n+1} = j \mid X_n = i)\mathbb{P}(X_n = i) = \sum_{i\in S} \pi(n)_i \mathbf{P}(n+1)^i_j$$

ovvero

$$\pi(n+1) = \pi(n)\mathbf{P}(n+1). \tag{1.1}$$

Iterando la (1.1) si ottiene

$$\pi(n+k) = \pi(k)\mathbf{P}(k+1)\mathbf{P}(k+2)\cdots\mathbf{P}(k+n) \qquad \forall n, k \geq 0. \tag{1.2}$$

Definizione 1.3. Un processo stocastico $\{X_n\}$ discreto a valori in S si dice *omogeneo* se esiste una matrice stocastica $\mathbf{P} = (\mathbf{P}^i_j)_{i,j\in S}$ tale che per ogni $n \in \mathbb{N}$ e per ogni $i, j \in S$, se $\mathbb{P}(X_n = i) > 0$, allora

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbf{P}^i_j.$$

In altre parole, un processo stocastico è omogeneo se le sue matrici di transizione non dipendono da n.

Se $\{X_n\}$ è omogeneo, la (1.2) diventa

$$\pi(n+k) = \pi(k)\mathbf{P}^n \qquad \forall n, k \geq 0. \tag{1.3}$$

1.2 Catene di Markov

Notazione. Data una funzione tra spazi misurabili $X : (E, \mathcal{E}) \to (F, \mathcal{F})$, indichiamo con $\sigma(X)$ la σ-algebra generata da X, cioè la più piccola σ-algebra in E che rende X misurabile.

Definizione 1.4. Sia $(X_t)_{t\in T}$ un processo a valori in S, $T \subset \mathbb{R}$. Indicando con $(\mathcal{F}_t)_{t\in T} = (\sigma(X_i \mid i \leq t))_{t\in T}$ la sua *filtrazione naturale*, si dice che (X_t) è un *processo di Markov* se per ogni $\phi : S \to \mathbb{R}$ misurabile limitata e per ogni $t, t' \in T$, $t' \geq t$,vale

$$\mathbb{E}[\phi(X_{t'}) \mid \mathcal{F}_t] = \mathbb{E}[\phi(X_{t'}) \mid X_t] \tag{1.4}$$

La relazione (1.4) si chiama *proprietà di Markov*.

Se il processo è a tempi e stati discreti, allora la proprietà di Markov si può riformulare in questo modo: per ogni $n \geq k \geq 0$ e per ogni $i_k, \ldots, i_{n+1} \in S$ vale

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_{k+1} = i_{k+1}, X_k = i_k) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) \tag{1.5}$$

sotto l'ipotesi $\mathbb{P}(X_n = i_n, \ldots, X_{k+1} = i_{k+1}, X_k = i_k) > 0$. In questo caso il processo si chiama *catena di Markov*.

Dalla (1.5) segue la proprietà più generale data dal teorema:

Teorema 1.1. *Sia $\{X_n\}$ un processo stocastico discreto a valori in S. $\{X_n\}$ è una catena di Markov se e solo se per ogni $r, n, k \in \mathbb{N}$ con $0 \leq r < n < n+k$, e per ogni $G \in \sigma((X_0, \ldots, X_r)), F \in \sigma((X_{r+1}, \ldots, X_n)), E \in \sigma((X_{n+1}, \ldots, X_{n+k}))$, se $\mathbb{P}(F \cap G) > 0$, si ha*

$$\mathbb{P}(E \mid F \cap G) = \mathbb{P}(E \mid F). \tag{1.6}$$

Inoltre,

$$\mathbb{P}(E \cap F \mid G) = \mathbb{P}(E \mid F)\mathbb{P}(F \mid G). \tag{1.7}$$

Dimostrazione. Che la (1.6) implichi la proprietà di Markov (1.5) è ovvio, scegliendo opportuni E, F, G. Dimostriamo il viceversa. Per ogni $n \geq 0$, dato $i_n \in S$ sia

$$A_n := \{\omega \in \Omega \mid X_n(\omega) = i_n\}.$$

Preso $B \in \sigma(X_0, \ldots, X_{n-1})$ tale che $\mathbb{P}(A_n \cap B) > 0$, ricordando che B si può scrivere come unione disgiunta di eventi della forma $\{X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\}$, dalla legge delle probabilità totali e dalla proprietà di Markov, si ottiene

$$\mathbb{P}(A_{n+1} \mid A_n \cap B) = \mathbb{P}(A_{n+1} \mid A_n). \tag{1.8}$$

Dal fatto che $\mathbb{P}(E \cap F \mid G) = \mathbb{P}(E \mid F \cap G)\mathbb{P}(F \mid G)$ e applicando la (1.8) si può scrivere

$$\begin{aligned}
&\mathbb{P}(A_{n+1+k} \cap \cdots \cap A_{n+1} \mid A_n \cap B) \\
=&\mathbb{P}(A_{n+1+k} \mid A_{n+k} \cap \cdots \cap A_n \cap B)\mathbb{P}(A_{n+k} \cap \cdots \cap A_{n+1} \mid A_n \cap B) \\
=&\mathbb{P}(A_{n+1+k} \mid A_{n+k})\mathbb{P}(A_{n+k} \cap \cdots \cap A_{n+1} \mid A_n \cap B).
\end{aligned}$$

Con un semplice argomento induttivo si può concludere che vale

$$\mathbb{P}(A_{n+1+k} \cap \cdots \cap A_{n+1} \mid A_n \cap B) = \prod_{j=n}^{k+n} \mathbb{P}(A_{j+1} \mid A_j). \tag{1.9}$$

Da quest'ultima uguaglianza e dalla legge delle probabilità totali, segue l'uguaglianza (1.6) della tesi. Infine la (1.7) è conseguenza di (1.6), infatti

$$\begin{aligned}
\mathbb{P}(E \cap F \mid G) &= \mathbb{P}(E \mid F \cap G)\mathbb{P}(F \cap G \mid G) \\
&= \mathbb{P}(E \mid F \cap G)\mathbb{P}(F \mid G) = \mathbb{P}(E \mid F)\mathbb{P}(F \mid G).
\end{aligned}$$

□

La proprietà di Markov, spesso chiamata anche "assenza di memoria", esprime il fatto che la previsione statistica dello stato futuro della catena dipende solo dallo stato presente. Questo semplifica notevolmente lo studio del processo stocastico. Infatti, in generale, per calcolare la probabilità di eventi individuati dalle $\{X_n\}$ serve conoscere le probabilità congiunte $\mathbb{P}(X_n = i_n, \ldots, X_0 = i_0)$ al variare di n e degli stati $i_0, \ldots, i_n$, e questo è un problema difficile, se non si hanno altre ipotesi, come l'indipendenza delle variabili. Invece, per una catena di Markov basta conoscere le diverse probabilità di transizione $\mathbb{P}(X_{n+1} = j \mid X_n = i)$. Infatti per qualsiasi $i_0, \ldots, i_k \in S$, ogni volta che $\{X_{n+k-1} = i_{k-1}, \ldots, X_n = i_0\}$ è non trascurabile, vale

$$\begin{aligned}
&\mathbb{P}(X_{n+k} = i_k, \ldots, X_{n+1} = i_1, X_n = i_0) \\
=&\mathbb{P}(X_{n+k} = i_k \mid X_{n+k-1} = i_{k-1}, \ldots, X_n = i_0)\mathbb{P}(X_{n+k-1} = i_{k-1}, \ldots, X_n = i_0) \\
=&\mathbb{P}(X_{n+k} = i_k \mid X_{n+k-1} = i_{k-1})\mathbb{P}(X_{n+k-1} = i_{k-1}, \ldots, X_n = i_0)
\end{aligned}$$

e, iterando, si ottiene

$$\begin{aligned}
&\mathbb{P}(X_{n+k} = i_k, \ldots, X_{n+1} = i_1, X_n = i_0) \\
=&\prod_{j=0}^{k-1} \mathbb{P}(X_{n+j+1} = i_{j+1} \mid X_{n+j} = i_j)\mathbb{P}(X_n = i_0) \\
=&\pi(n)_{i_0}\mathbf{P}(n+1)_{i_1}^{i_0} \cdots \mathbf{P}(n+k)_{i_k}^{i_{k-1}}
\end{aligned} \tag{1.10}$$

Se la catena è omogenea, è chiaro che la complessità si riduce ulteriormente, infatti la (1.10) diventa

$$\mathbb{P}(X_{n+k} = i_k, \ldots, X_{n+1} = i_1, X_n = i_0) = \pi(n)_{i_0}\mathbf{P}_{i_1}^{i_0} \cdots \mathbf{P}_{i_k}^{i_{k-1}}. \tag{1.11}$$

Inoltre, sempre nell'ipotesi di omogeneità, se $\mathbb{P}(X_n = i_0) = \mathbb{P}(X_0 = i_0)$, allora

$$\begin{aligned}
&\mathbb{P}(X_{n+k} = i_k, \ldots, X_{n+1} = i_1 \mid X_n = i_0) \\
=&\mathbb{P}(X_k = i_k, \ldots, X_1 = i_1 \mid X_0 = i_0).
\end{aligned} \tag{1.12}$$

La (1.12) si chiama *proprietà di rinnovo* della catena.

La seguente proposizione, che si dimostra facilmente per induzione, mostra la relazione tra le probabilità condizionate $\mathbb{P}(X_{n+k} = j \mid X_k = i)$ e matrici di transizione.

Proposizione 1.1. *Sia $\{\mathbf{P}(n)\}$ la sequenza delle matrici di transizione di una catena di Markov discreta $\{X_n\}$ a valori in S. Per ogni $i, j \in S$ e $n, k \geq 0$ vale*

$$\mathbb{P}(X_{n+k} = j \mid X_k = i) = (\mathbf{P}(k+1) \cdots \mathbf{P}(k+n))^i_j. \tag{1.13}$$

Se la catena è omogenea, cioè $\mathbf{P}(n) = \mathbf{P} \ \ \forall n \in \mathbb{N}$, allora

$$\mathbb{P}(X_{n+k} = j \mid X_k = i) = (\mathbf{P}^n)^i_j \tag{1.14}$$

Dimostriamo ora che data una variabile aleatoria X_0 e una matrice stocastica $\mathbf{P}$, si può costruire una catena di Markov che ha X_0 come variabile iniziale e $\mathbf{P}$ come matrice di transizione. Questo garantisce l'esistenza della probabilità condizionata rispetto a X_0, infatti condizionare per $\{X_0 = i\}$ (quando $\mathbb{P}(X_0 = i) > 0$) equivale a considerare la catena che al tempo zero si trova nello stato i con probabilità 1.

Teorema 1.2. *Sia X_0 variabile aleatoria sullo spazio $(\Omega, \mathcal{E}, \mathbb{P})$ a valori nell'insieme discreto S. Sia $\{\xi_n\}$ una sequenza di variabili $\xi_n : \Omega \to \mathbb{R}^N$ indipendenti e identicamente distribuite (i.i.d.), tutte indipendenti da X_0. Se $f : S \times \mathbb{R}^N \to S$ è una funzione misurabile, allora la sequenza di variabili $\{X_n\}$ tale che $X_n : \Omega \to S$ e*

$$X_{n+1}(\omega) = f(X_n(\omega), \xi_n(\omega)) \quad \forall \omega \in \Omega \quad \forall n \geq 0$$

è una catena di Markov omogenea con matrice di trasizione

$$\mathbf{P}^i_j = \mathbb{P}(f(i, \xi_n) = j) \quad \forall i, j \in S.$$

Dimostrazione. Per ogni $n > 0$ la variabile ξ_n è indipendente da $(X_0, \ldots, X_n)$, in quanto ogni X_k è funzione di $X_0, \xi_1, \ldots, \xi_{k-1}$, che sono tutte indipendenti da ξ_n. Allora, dalla definizione di X_{n+1}, per ogni $j, i, i_{n-1}, \ldots, i_0 \in S$, si ottiene

$$\begin{aligned} &\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ =&\mathbb{P}(f(i, \xi_n) = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ =&\mathbb{P}(f(i, \xi_n) = j), \end{aligned} \tag{1.15}$$

dove l'ultima uguaglianza è dovuta al fatto che anche $f(i, \xi_n)$ è indipendente da $(X_0, \ldots, X_n)$. Inoltre, dalla (1.15) si può notare che la probabilità condizionata non dipende dai valori assunti da $X_{n-1}, \ldots, X_0$, quindi

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i),$$

cioè $\{X_n\}$ è una catena di Markov. Dato che le variabili ξ_n sono identicamente distribuite, la probabilità di transizione $\mathbb{P}(f(i, \xi_n) = j)$ non dipende da n, quindi la catena è omogenea. □

Teorema 1.3. *Siano S un insieme finito o numerabile, $\mathbf{P}$ una matrice stocastica e $X_0 : \Omega \to S$ variabile aleatoria sullo spazio di probabilità $(\Omega, \mathcal{E}, \mathbb{P})$. Allora esiste una catena di Markov omogenea $\{X_n\}_{n \geq 0}$ con insieme degli stati S e matrice di transizione $\mathbf{P}$.*

Dimostrazione. Sia $\{\xi_n\}$, $\xi_n : \Omega \to [0, 1]$, una successione di variabili i.i.d. con distribuzione uniforme, indipendenti da X_0 (la cui esistenza è garantita dall'esistenza della misura prodotto infinito numerabile). Sia $f : S \times \mathbb{R} \to S$ tale che

$$f(i, s) := \min \Big\{ j \Big| \sum_{h=1}^{j} \mathbf{P}^i_h \geq s \Big\}.$$

Sia $\{X_n\}$, con $X_n : \Omega \to S$, la sequenza definita da

$$X_{n+1}(\omega) = f(X_n(\omega), \xi_n(\omega)) \quad \forall \omega \in \Omega \quad \forall n \geq 0.$$

Per il teorema 1.2, $\{X_n\}$ è una catena di Markov omogenea con probabilità di transizione dallo stato i allo stato j uguale a $\mathbb{P}(f(i, \xi_n) = j)$. Inoltre, $f(i, \xi_n(\omega)) = j$ se e solo se

$$\sum_{h=1}^{j-1} \mathbf{P}_h^i < \xi_n(\omega) \leq \sum_{h=1}^{j} \mathbf{P}_h^i,$$

quindi, poichè le ξ_n sono distribuite uniformemente, si ha

$$\begin{aligned} &\mathbb{P}(f(i, \xi_n(\omega)) = j) \\ =&\mathbb{P}\Big(\sum_{h=1}^{j-1} \mathbf{P}_h^i < \xi_n \leq \sum_{h=1}^{j} \mathbf{P}_h^i\Big) \\ =&\sum_{h=1}^{j} \mathbf{P}_h^i - \sum_{h=1}^{j-1} \mathbf{P}_h^i = \mathbf{P}_j^i. \end{aligned}$$

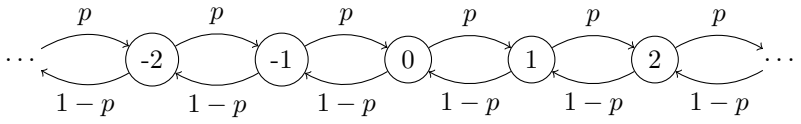
□

Un modo utile per rappresentare una catena di Markov (ma in generale un processo stocastico a tempi e stati discreti) è usare i grafi orientati pesati. Se $\mathbf{P}$ è la matrice di transizione al passo k della catena $\{X_n\}$ con insieme degli stati S, si può considerare il grafo pesato che ha S come insieme dei nodi, e per ogni $i, j \in S$ un arco diretto da i verso j con peso $\mathbf{P}_j^i$, ogni volta che $\mathbf{P}_j^i > 0$.

Esempio 1.1 (Random walk)**.** Ad ogni istante $n \in \mathbb{N}$ una persona si trova in una posizione $k \in \mathbb{Z}$. All'istante successivo si può muovere in posizione $k+1$ con probabilità p, o in posizione $k-1$ con probabilità $1-p$. Sia X_n la variabile aleatoria che indica la posizione all'istante n. Allora $\{X_n\}$ è una catena di Markov omogenea con matrice di transizione

$$\mathbf{P} = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & p & 0 & \dots \\ \dots & 1-p & 0 & p & \dots \\ \dots & 0 & 1-p & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Il grafo che rappresenta $\mathbf{P}$ è



1.3 Classificazione degli stati e parametri caratteristici

Consideriamo una catena di Markov omogenea sullo spazio $(\Omega, \mathcal{E}, \mathbb{P})$ con insieme degli stati discreto S e matrice di transizione $\mathbf{P} = (p_{ij})$. Per ogni $n \geq 1$ indichiamo con $p_{ij}^{(n)}$ le entrate della matrice $\mathbf{P}^n$.
Dato $C \subset S$, si dice che $\omega \in \Omega$ *visita* C al passo k se $X_k(\omega) \in C$, e più in generale $\omega \in \Omega$ *visita* C se esiste $n \geq 1$ tale che $X_n(\omega) \in C$. Inoltre, se $i, j \in S$, si dice che i *visita* j, e si scrive $i \to j$, se esiste $n \geq 1$ tale che $p_{ij}^{(n)} > 0$. Se $i \to j$ e $j \to i$, si scrive $i \leftrightarrow j$.
Introduciamo una variabile aleatoria t_C a valori in $\{1, 2, \dots\} \cup \{+\infty\}$ tale che per ogni $\omega \in \Omega$

$$t_C(\omega) := \begin{cases} +\infty & \text{se } X_n(\omega) \notin C \ \ \forall n \geq 1 \\ \min\{n \geq 1 \mid X_n(\omega) \in C\} & \text{altrimenti.} \end{cases}$$

$t_C(\omega)$ indica il passo in cui ω visita per la prima volta C.

Per ogni $i \in S$, purchè $\mathbb{P}(X_0 = i) > 0$, si può introdurre il parametro $f_{iC}^{(k)}$, che indica la probabilità di entrare in C al passo k partendo dallo stato i, definito da

$$\begin{aligned} f_{iC}^{(k)} :=& \mathbb{P}(t_C = k \mid X_0 = i) \\ =& \mathbb{P}(X_k \in C, X_{k-1} \notin C, \dots, X_1 \notin C \mid X_0 = i). \end{aligned}$$

Inoltre, chiamiamo f_{iC} la probabilità di visitare C partendo da i. Questa, dato che gli eventi $\{\omega \in \Omega \mid t_C(\omega) = k\}$ sono disgiunti, vale

$$f_{iC} := \mathbb{P}(t_C < +\infty \mid X_0 = i) = \sum_{k=1}^{+\infty} \mathbb{P}(t_C = k \mid X_0 = i) = \sum_{k=1}^{+\infty} f_{iC}^{(k)}.$$

Se $C = \{j\}$, scriviamo t_j, $f_{ij}^{(k)}$ e f_{ij}. Se $i = j$, $f_{jj}^{(k)}$ e f_{jj} si dicono *probabilità di ritorno in j al passo k* e *probabilità di ritorno in j*.

É utile calcolare il numero medio di passi che i cammini che partono da i impiegano per visitare C, ovvero il valore atteso della variabile t_C rispetto alla misura di probabilità condizionata $\mathbb{P}_i(\cdot) := \mathbb{P}(\cdot \mid X_0 = i)$:

$$\mathbb{E}_i[t_C] := \int_\Omega t_C(\omega)\, d\mathbb{P}_i(\omega) = \frac{1}{\mathbb{P}(X_0 = i)} \int_{\{X_0 = i\}} t_C(\omega)\, d\mathbb{P}(\omega)$$

$\mathbb{E}_i[t_C]$ si chiama anche *tempo medio di attesa* per entrare in C partendo da i. Quando $C = \{j\}$ e $i = j$, $\mathbb{E}_i[t_C]$ si chiama più propriamente *tempo medio di ritorno* allo stato j e si denota con $\overline{T}_{jj}$.

Dal teorema di Beppo Levi segue che

$$\mathbb{E}_i[t_C] = \begin{cases} +\infty & \text{se } \mathbb{P}(t_C = +\infty \mid X_0 = i) > 0 \\ \sum_{k=1}^{+\infty} k\, \mathbb{P}(t_C = k \mid X_0 = i) & \text{altrimenti.} \end{cases}$$

Dalla definizione di f_{iC} si ha che $\mathbb{P}(t_C = +\infty \mid X_0 = i) > 0$ se e solo se $f_{iC} < 1$, quindi

$$\mathbb{E}_i[t_C] = \begin{cases} +\infty & \text{se } f_{iC} < 1 \\ \sum_{k=1}^{+\infty} k f_{iC}^{(k)} & \text{se } f_{iC} = 1 \end{cases} \tag{1.16}$$

In particolare,

$$\overline{T}_{jj} = \begin{cases} +\infty & \text{se } f_{jj} < 1 \\ \sum_{k=1}^{+\infty} k f_{jj}^{(k)} & \text{se } f_{jj} = 1 \end{cases} \tag{1.17}$$

Infine, introduciamo le variabili $V_C^{(n)}$ e V_C che indicano rispettivamente il *numero di visite in C nei primi n passi* e il *numero totale di visite in C*. Considerati gli eventi $E_n = \{X_n \in C\}$, tali variabili sono definite da

$$V_C^{(n)} = \sum_{k=1}^{n} \mathbb{1}_{E_k} \quad \text{e} \quad V_C = \sum_{k=1}^{\infty} \mathbb{1}_{E_k}.$$

Se $C = \{j\}$, si scrive più brevemente $V_j^{(n)}$ e V_j.
Il *numero medio di visite* in j partendo da i è definito come il valore atteso di V_j rispetto alla misura di probabilità $\mathbb{P}_i(\cdot)$:

$$\mathbb{E}_i[V_j] := \frac{1}{\mathbb{P}(X_0 = i)} \int_{\{X_0 = i\}} V_j(\omega)\, d\mathbb{P}(\omega)$$

Proposizione 1.2. *Valgono le seguenti formule:*

$$\mathbb{E}_i[V_j] = \sum_{n=0}^{\infty} p_{ij}^{(n)} \tag{1.18}$$

$$\mathbb{E}_i[V_j] = \begin{cases} +\infty & \text{se } f_{jj} = 1 \\ \frac{f_{ij}}{1 - f_{jj}} & \text{se } f_{jj} < 1 \end{cases} \tag{1.19}$$

Dimostrazione. Dal teorema di Beppo Levi si ha

$$\mathbb{E}_i\Big[\sum_{n=0}^{\infty} \mathbb{1}_{E_n}\Big] = \sum_{n=0}^{\infty} \mathbb{E}_i[\mathbb{1}_{E_n}],$$

allora, usando la proprietà (1.14), si dimostra la prima uguaglianza:

$$\mathbb{E}_i[V_j] = \sum_{n=0}^{\infty} \mathbb{E}_i[\mathbb{1}_{E_n}] = \sum_{n=0}^{\infty} \mathbb{P}(X_n = j \mid X_0 = i) = \sum_{n=0}^{\infty} p_{ij}^{(n)}.$$

Per dimostrare la seconda uguaglianza, usiamo la formula di Cavalieri e scriviamo

$$\mathbb{E}_i[V_j] = \sum_{k=0}^{\infty} \mathbb{P}(V_j > k \mid X_0 = i).$$

Osserviamo che

$$\mathbb{P}(V_j \geq k \mid X_0 = i) = f_{ij} f_{jj}^{k-1}. \tag{1.20}$$

Infatti, se $k = 1$, dal fatto che $V_j(\omega) \geq 1$ se e solo se $t_j(\omega) < \infty$,

$$\mathbb{P}(V_j \geq 1 \mid X_0 = i) = \mathbb{P}(t_j < \infty \mid X_0 = i) = f_{ij}.$$

Assumiamo $k \geq 2$, e per ogni $h \geq 1$ consideriamo $R_h = \sum_{k=h+1}^{\infty} \mathbb{1}_{E_k}$ la variabile che conta il numero di visite in j dopo il passo h. In questo modo

$$\{V_j > k\} = \bigcup_{h=1}^{\infty} \{R_h > k-1, t_j = h\}.$$

Allora, grazie alla proprietà di rinnovo (1.12) e all'omogeneità della catena, si ottiene

$$\begin{aligned}
&\mathbb{P}(R_h \geq k-1,\, t_j = h \mid X_0 = i) \\
=&\mathbb{P}(R_h \geq k-1 \mid t_j = h,\, X_0 = i)\mathbb{P}(t_j = h \mid X_0 = i) \\
=&\mathbb{P}(R_h \geq k-1 \mid X_h = j)\mathbb{P}(t_j = h \mid X_0 = i) \\
=&\mathbb{P}(V_j \geq k-1 \mid X_0 = j) f_{ij}^{(h)}.
\end{aligned}$$

Quindi,

$$\begin{aligned}
\mathbb{P}(V_j \geq k \mid X_0 = i) &= \sum_{h=1}^{\infty} \mathbb{P}(R_h \geq k-1,\, t_j = h \mid X_0 = i) \\
&= \Big(\sum_{h=1}^{\infty} f_{ij}^{h}\Big)\mathbb{P}(V_j \geq k-1 \mid X_0 = j) \\
&= f_{ij}\mathbb{P}(V_j \geq k-1 \mid X_0 = j).
\end{aligned}$$

Induttivamente si conclude che

$$\mathbb{P}(V_j \geq k \mid X_0 = i) = f_{ij} f_{jj}^{k-1}.$$

A questo punto, si ottiene la tesi:

$$\begin{aligned}
\mathbb{E}_i[V_j] &= \sum_{k=0}^{\infty} \mathbb{P}(V_j > k \mid X_0 = i) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(V_j \geq k \mid X_0 = i) \\
&= \sum_{k=1}^{\infty} f_{ij} f_{jj}^{k-1} = \begin{cases} +\infty & \text{se } f_{jj} = 1 \\ \frac{f_{ij}}{1-f_{jj}} & \text{se } f_{jj} < 1. \end{cases}
\end{aligned}$$

□

Le formule precedenti servono a dare una caratterizzazione degli stati di una catena di Markov.

Definizione 1.5. Uno stato $j \in S$ si dice *ricorrente* se $\sum_{k=1}^{+\infty} p_{jj}^{(k)} = +\infty$, si dice *transiente* se $\sum_{k=1}^{+\infty} p_{jj}^{(k)} < +\infty$

Proposizione 1.3. *Sono equivalenti:*

1. *j è ricorrente;*
2. *i cammini che partono da j tornano in j con probabilità 1, cioè $f_{jj} = 1$;*
3. *i cammini che partono da j visitano j infinite volte con probabilità 1.*

Inoltre, la probabilità che un cammino che parte da i passi infinite volte per j è f_{ij}. Infine, se j è ricorrente e $j \leftrightarrow i$, allora i cammini che partono da i visitano j con probabilità 1, cioè $f_{ij} = 1$.

Dimostrazione. L'equivalenza tra *1* e *2* segue immediatamente dalla proposizione 1.2. É ovvio che da *3* segua *2*. Infine, supponendo $f_{jj} = 1$ e applicando la (1.20)

$$\mathbb{P}(V_j = +\infty \mid X_0 = i) = \lim_{k \to +\infty} \mathbb{P}(V_j > k \mid X_0 = i) = \lim_{k \to +\infty} f_{ij} f_{jj}^{k-1} = f_{ij},$$

cioè *3*, se $i = j$. Infine, supponiamo $j \to i$ e sia $k > 0$ tale che $p_{ji}^{(k)} > 0$. Il prodotto $p_{ji}^{(k)}(1 - f_{ij})$ indica la probabilità che un cammino che parte da j passi per i al passo k e non torni più in j (si può dimostare usando la formula (1.7)), e $1 - f_{jj}$ indica la probabilità di non ritorno in j. Allora $p_{ji}^{(k)}(1 - f_{ij}) \leq 1 - f_{jj}$, e se $f_{jj} = 1$, anche $f_{ij} = 1$. □

Proposizione 1.4. *Sono equivalenti:*

1. *j è transiente;*
2. *i cammini che partono da j tornano in j con probabilità minore di 1, cioè $f_{jj} < 1$;*
3. *la probabilità che i cammini che partono da j visitino j infinite volte è 0;*
4. *i cammini che partono da j visitano j un numero finito di volte con probabilità 1.*

Dimostrazione. *1* e *2* sono equivalenti per la proposizione 1.2. Inoltre,

$$\mathbb{P}(V_j = +\infty \mid X_0 = j) = \lim_{k \to +\infty} \mathbb{P}(V_j > k \mid X_0 = j) = \lim_{k \to +\infty} f_{jj}^k.$$

Perciò *2* è equivalente a *3* perchè $\mathbb{P}(V_j = +\infty \mid X_0 = j) = 0$ se e solo se $f_{jj} < 1$. Se j è transiente, si ha $\mathbb{P}(V_j = +\infty \mid X_0 = j) = 0$, altrimenti $\sum_{k=1}^{\infty} p_{jj}^{(k)} = \mathbb{E}_j[V_j] = +\infty$. Infine, l'equivalenza tra *3* e *4* è ovvia. □

Da questa caratterizzazione degli stati e da 1.17, si osserva che il tempo medio di ritorno allo stato j può essere finito solo se j è ricorrente. Più precisamente, $\overline{T}_{jj}$ è finito se e solo se j è ricorrente e la serie $\sum_{k=1}^{+\infty} k f_{jj}^{(k)}$ converge.

Dimostriamo una formula che tornerà utile in seguito.

Proposizione 1.5. *Per ogni $n \geq 1$, per ogni coppia di stati i, j, vale*

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}; \tag{1.21}$$

Dimostrazione. Fissato $n \geq 1$, per ogni $k \in \{1, \ldots, n\}$ si consideri l'evento

$$F_{k,n} := \{X_1 \neq j, \ldots, X_{k-1} \neq j, X_k = j, X_n = j\}.$$

Siano, inoltre,

$$\begin{aligned} E &:= \{X_n = j\}, \\ F &:= \{X_k = j, X_{k-1} \neq j, \dots X_1 \neq j\}, \\ G &:= \{X_0 = i\}. \end{aligned}$$

Osservando che $F_{k,n} = E \cap F$, applicando la (1.7) e usando l'ipotesi di omogeneità, si ottiene

$$\begin{aligned} &\mathbb{P}(F_{k,n} \mid X_0 = i) \\ =&\mathbb{P}(E \mid F)\mathbb{P}(F \mid G) \\ =&\mathbb{P}(X_n = j \mid X_k = j)\mathbb{P}(X_k = j, X_{k-1} \neq j, \dots, X_1 \neq j \mid X_0 = i) \\ =&p_{jj}^{(n-k)} f_{ij}^{(k)}. \end{aligned}$$

Gli eventi $\{F_{k,n}\}_{k=1}^n$ partizionano l'insieme $\{\omega \in \Omega \mid X_n(\omega) = j\}$, quindi si ha la tesi:

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i) = \sum_{k=1}^{n} \mathbb{P}(F_{k,n} \mid X_0 = i) = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}.$$

□

1.4 Distribuzioni invarianti

Definizione 1.6. Data una catena di Markov omogenea a valori in S con matrice di transizione $\mathbf{P}$, un vettore stocastico (o vettore di probabilità) $\pi = (\pi_j)_{j \in S}$, cioè tale che $\pi_j \geq 0$ per ogni $j \in S$ e $\sum_{j \in S} \pi_j = 1$, si chiama *distribuzione invariante* per la catena se vale $\pi\mathbf{P} = \pi$.

Osservazione 1.1. Dalla relazione $\pi(n+1) = \pi(n)\mathbf{P}$ si ottengono due importanti osservazioni:

1. π è una distribuzione invariante se e solo se ogni volta che la catena ha densità marginale π al tempo iniziale, allora anche le densità marginali agli istanti successivi sono uguali a π;
2. se la successione $\{\pi(n)\}_{n \in \mathbb{N}}$ converge ad un vettore stocastico π, allora π è un distribuzione invariante.

Teorema 1.4. *Sia $\{X_n\}$ una catena di Markov omogenea. Se l'insieme degli stati S è finito, allora esiste una distribuzione invariante π.*

Dimostrazione. Siano $\mathbf{P}$ la matrice di transizione del processo e $\pi(0) = (\pi(0)_i)_{i \in S}$ il vettore che rappresenta la densità marginale della catena al tempo $n = 0$. Per ogni $n > 0$ sia

$$\tilde{\pi}(n) = \frac{1}{n} \sum_{k=1}^{n} \pi(0)\mathbf{P}^k.$$

Ogni $\tilde{\pi}(n)$ è un vettore stocastico, in quanto media aritmetica di vettori stocastici (cfr. (1.3)). Per il teorema di Bolzano-Weierstrass nel caso vettoriale, esiste una sottosuccessione $\{\tilde{\pi}(n_k)\}_k$ che converge ad un certo $\tilde{\pi}_\infty$. Poichè l'insieme dei vettori stocastici

$$\left\{ x \in \mathbb{R}^{|S|} \,\middle|\, x_i \geq 0 \text{ e } \sum_{i=1}^{|S|} x_i = 1 \right\}$$

è chiuso in $\mathbb{R}^{|S|}$, grazie al fatto che S è finito, allora anche $\tilde{\pi}_\infty$ è un vettore di probabilità. Tale $\tilde{\pi}_\infty$ è la distribuzione invariante cercata. Infatti, per ogni $n > 0$ vale

$$\begin{aligned}\tilde{\pi}(n)\mathbf{P} &= \left(\frac{1}{n}\sum_{k=1}^{n}\pi(0)\mathbf{P}^k\right)\mathbf{P} \\ &= \frac{1}{n}\sum_{k=1}^{n}\pi(0)\mathbf{P}^{k+1} \\ &= \frac{1}{n}\sum_{k=1}^{n}\pi(0)\mathbf{P}^k + \frac{1}{n}(\pi(0)\mathbf{P}^{n+1} - \pi(0)\mathbf{P}) \\ &= \tilde{\pi}(n) + \frac{1}{n}(\pi(0)\mathbf{P}^{n+1} - \pi(0)\mathbf{P}).\end{aligned}$$

Il termine $\frac{1}{n}(\pi(0)\mathbf{P}^{n+1} - \pi(0)\mathbf{P})$ tende a 0 per $n \to +\infty$, poichè le componenti del vettore $\pi(0)\mathbf{P}^{n+1}$ sono limitate per ogni n. Passando al limite per $n_k \to +\infty$, l'identità di sopra, con n_k al posto di n, permette di concludere che $\tilde{\pi}_\infty\mathbf{P} = \tilde{\pi}_\infty$. □

Osservazione 1.2. Nel teorema di esistenza l'ipotesi di S finito è fondamentale. Si consideri il caso della passeggiata aleatoria nell'esempio 1.1 con $p = 1$. Poichè la matrice di transizione è

$$\mathbf{P} = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & 1 & 0 & 0 & \dots \\ \dots & 0 & 0 & 1 & 0 & \dots \\ \dots & 0 & 0 & 0 & 1 & \dots \\ \dots & 0 & 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix},$$

una distribuzione invariante deve essere tale che $\pi_{i+1} = \pi_i$ per ogni $i \in \mathbb{Z}$, cioè deve essere una distribuzione uniforme. Dato che gli stati sono infiniti, tale distribuzione non esiste.

Definizione 1.7. Una matrice di transizione $\mathbf{P} = (p_{ij})$ su un insieme di stati S si dice *irriducibile* se per ogni $i, j \in S$ esiste $n \geq 1$ tale che $p_{ij}^{(n)} > 0$; si dice *regolare* se esiste $n_0 \geq 1$ tale che $p_{ij}^{(n_0)} > 0$ per ogni $i, j \in S$.

Lemma 1.1. *Se* $\mathbf{P}$ *è una matrice di transizione irriducibile sull'insieme di stati* S *finito, ed esiste* $h \in S$ *tale che* $p_{hh} > 0$*, allora* $\mathbf{P}$ *è regolare.*

Dimostrazione. Per ogni $i, j \in S$ sia $n(i,j) \geq 1$ per cui $p_{ij}^{n(i,j)} > 0$. Sia $n_0 := 2\max_{i,j} n(i,j) + 1$. Allora, per ogni $i, j \in S$, preso $q := k_0 - n(i,h) - n(h,i)$, vale

$$p_{ij}^{(n_0)} = \sum_{k\in S} p_{ik}^{(n(i,k))} p_{kk}^{(q)} p_{kj}^{(n(k,j))} \geq p_{ih}^{(n(i,h))} p_{hh}^{(q)} p_{hj}^{(n(h,j))} \geq p_{ih}^{(n(i,h))} (p_{hh})^q p_{hj}^{(n(h,j))} > 0.$$

□

Teorema 1.5. *Sia* $\mathbf{P}$ *matrice di transizione regolare sull'insieme finito di stati* S*. Allora esiste un'unica distribuzione invariante.*

Dimostrazione. Sia $n_0 \geq 0$ per cui tutte le entrate di $\mathbf{P}^{n_0}$ sono non nulle. Siano π e $\tilde{\pi}$ due distribuzioni invarianti per $\mathbf{P}$ (ne esiste almeno una per il teorema 1.4). Chiaramente π e $\tilde{\pi}$ sono distribuzioni invarianti anche per le potenze di $\mathbf{P}$, che sono ancora matrici stocastiche, quindi si

può scrivere:

$$\begin{aligned}\sum_{i\in S}|\pi_i-\tilde{\pi}_i| &= \sum_{i\in S}\Big|\sum_{j\in S}\pi_j p_{ji}^{(n_0)}-\tilde{\pi}_j p_{ji}^{(n_0)}\Big| \\ &= \sum_{i\in S}\Big|\sum_{j\in S}(\pi_j-\tilde{\pi}_j)p_{ji}^{(n_0)}\Big| \\ &\leq \sum_{i\in S}\sum_{j\in S}|\pi_j-\tilde{\pi}_j|p_{ji}^{(n_0)} \\ &= \sum_{j\in S}|\pi_j-\tilde{\pi}_j|\sum_{i\in S}p_{ji}^{(n_0)} \\ &= \sum_{j\in S}|\pi_j-\tilde{\pi}_j|.\end{aligned}$$

Ma allora vale l'uguaglianza

$$\sum_{i\in S}\Big|\sum_{j\in S}(\pi_j-\tilde{\pi}_j)p_{ji}^{(n_0)}\Big| = \sum_{i\in S}\sum_{j\in S}|\pi_j-\tilde{\pi}_j|p_{ji}^{(n_0)},$$

ed in particolare

$$\Big|\sum_{j\in S}(\pi_j-\tilde{\pi}_j)p_{ji}^{(n_0)}\Big| = \sum_{j\in S}|\pi_j-\tilde{\pi}_j|p_{ji}^{(n_0)}.$$

É un fatto noto che la disuguaglianza triangolare tra numeri reali è un'uguaglianza se e solo se tutti i numeri sono concordi, quindi si può supporre

$$(\pi_j-\tilde{\pi}_j)p_{ji}^{(n_0)}\geq 0 \quad \forall j\in S.$$

Dividendo per $p_{ji}^{(n_0)}$, che è positivo, si ottiene

$$\pi_j\geq\tilde{\pi}_j \quad \forall j\in S,$$

e ricordando che si tratta di vettori stocastici, si ha

$$\sum_{j\in S}\pi_j = 1 = \sum_{j\in S}\tilde{\pi}_j,$$

perciò deve valere

$$\pi_j=\tilde{\pi}_j \quad \forall j\in S.$$

□

Osservazione 1.3. La distribuzione invariante è unica anche se la matrice di transizione $\mathbf{P}$ è irriducibile. Nella dimostrazione del teorema di unicità basta sostituire $\mathbf{P}^{n_0}$ con la matrice $\frac{1}{2}\sum_{n=0}^{\infty}2^{-n}\mathbf{P}^n$, che è anch'essa di transizione e con tutte le entrate non nulle.

1.5 Teorema di convergenza

Sia $\{X_n\}$ una catena di Markov omogenea con insieme degli stati finito o numerabile S e matrice di transizione $\mathbf{P}=(p_{ij})$. Per ogni $i,j\in S$ siano $\overline{T}_{jj}\in\mathbb{R}\cup\{+\infty\}$ il tempo medio di ritorno allo stato j e f_{ij} la probabilità di visitare j partendo da i.

Teorema 1.6. *Se $p_{jj}^{(n)}\to w_j$, allora*

$$w_j=\frac{1}{\overline{T}_{jj}} \quad e \quad p_{ij}^{(n)}\to f_{ij}w_j \tag{1.22}$$

(assumiamo $w_j=0$ quando $\overline{T}_{jj}=+\infty$).

Dimostrazione. Per ogni $n, k \geq 1$, sia $\varphi_n(k) := \mathbb{1}_{\{1,\dots,n\}}(k) f_{ij}^{(k)} p_{jj}^{(n-k)}$. Allora la (1.21) si può scrivere come

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} \varphi_n(k).$$

Nello spazio misurabile $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ dotato della "misura che conta i punti", si può applicare il teorema di convergenza dominata alla successione $\{\varphi_n\}$, che è dominata dalla funzione $k \mapsto f_{ij}^{(k)}$. Poichè, fissato k, $\lim_{n\to\infty} \varphi_n(k) = f_{ij}^{(k)} w_j$, si ottiene

$$\lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{k=1}^{\infty} \varphi_n(k) = \sum_{k=1}^{\infty} \lim_{n\to\infty} \varphi_n(k) = \Big(\sum_{k=1}^{\infty} f_{ij}^{(k)} \Big) w_j = f_{ij} w_j.$$

Per dimostrare che $w_j = \frac{1}{\overline{T}_{jj}}$ distinguiamo due casi:

- se j è transiente, allora $\sum_{n=1}^{\infty} p_{jj}^{(n)} < +\infty$, quindi $p_{jj}^{(n)} \to 0 = w_j$. Inoltre $f_{jj} < 1$, allora dalla (1.17) si ha $\overline{T}_{jj} = +\infty$;

- se j è ricorrente, allora $f_{jj} = 1$ e da (1.17) $\overline{T}_{jj} = \sum_{k=1}^{+\infty} k f_{jj}^{(k)}$.
  Usiamo il seguente lemma riguardante le successioni di numeri reali non negativi:

  Lemma. *Siano $\{a_n\}$ e $\{b_n\}$ sequenze di numeri non negativi tali che $a_0 = 1$, $\sum_{i=1}^{\infty} b_i = 1$ e $a_n = \sum_{j=1}^{n} b_j a_{n-j}$ per ogni $n \geq 1$. Se $a_n \to \tilde{a} \in \mathbb{R}_+ \cup \{+\infty\}$, allora*

$$\tilde{a} = \frac{1}{B}, \quad \textit{dove} \quad B := \sum_{k=1}^{\infty} k\, b_k.$$

  Dal fatto che $p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}$ per ogni $n \geq 1$ e che $\sum_{k=1}^{\infty} f_{jj}^{(k)} = f_{jj} = 1$, si può applicare il lemma con $a_n := p_{jj}^{(n)}$ e $b_k := f_{jj}^{(k)}$, ottenendo la tesi.

□

Osservazione 1.4. Se j è transiente, allora $w_j = 0 = f_{ij} w_j$. Se j è ricorrente e $j \leftrightarrow i$ allora $f_{ij} = 1$ (cfr. Proposizione 1.3) e $f_{ij} w_j = w_j$. Quindi il teorema 1.6 mostra che il calcolo dei tempi medi di ritorno permette di trovare l'unica distribuzione invariante w, nel caso in cui S sia finito e $\mathbf{P}$ regolare (o irriducibile). Infatti, in queste ipotesi e quelle del teorema,

$$\pi(n)_j = \sum_{i\in S} \pi(0)_i p_{ij}^{(n)} \to \sum_{i\in S} \pi(0)_i f_{ij} w_j = \sum_{i\in S} \pi(0)_i w_j = w_j.$$

Invece, se S è infinito, per quanto visto nella sezione precedente, non è garantito che il limite sia una distribuzione di probabilità.

Dimostriamo che se $\mathbf{P}$ è regolare e S finito, $\mathbf{P}^n$ converge per $n \to \infty$, quindi vale il teorema sulla convergenza, e il limite di $\mathbf{P}^n$ è la matrice che ha in ogni riga l'unica distribuzione invariante.

Teorema 1.7. *Se $\mathbf{P}$ è regolare, l'autovalore 1 ha molteplicità geometrica uguale a 1.*

Dimostrazione. Sia $n_0 \geq 1$ tale che $\mathbf{P}^{n_0}$ abbia tutte le entrate non nulle. Sia π l'unica distribuzione invariante per $\mathbf{P}$. Per ogni $i \in S$ $\pi_i = \sum_{j\in S} \pi_j p_{ji}^{(n_0)}$, e poichè esiste $j \in S$ con $\pi_j > 0$, anche $\pi_i > 0$. Sia $v \in \mathbb{R}^{|S|}$ un altro autovettore relativo a 1. Visto che le entrate di π sono tutte strettamente positive, esiste $\epsilon > 0$ per cui $\overline{v} := \frac{\pi + \epsilon v}{\langle \pi + \epsilon v, 1 \rangle}$ è un vettore di probabilità. Dall'unicità della distribuzione invariante segue che $\overline{v} = \pi$, quindi $v \in Span\{\pi\}$. □

Teorema 1.8. *Se $\mathbf{P}$ è regolare, ogni autovalore $\lambda \neq 1$ è tale che $|\lambda| < 1$.*

Dimostrazione. Consideriamo $k \geq 1$ per cui $p_{ii}^{(k)} > 0$ per ogni $i \in S$. Applicando il primo teorema di Gershgorin si può affermare che lo spettro di $\mathbf{P}^k$ è contenuto in $\bigcup_{i \in S} K_i$, con $K_i = \{z \in \mathbb{C} \mid |z - p_{ii}^{(k)}| \leq 1 - p_{ii}^{(k)}\}$. Da questo si ottiene che ogni autovalore di $\mathbf{P}^k$ diverso da 1 ha modulo strettamente minore di 1. É noto che, data una matrice A, $Sp(A) = \{\lambda_1, \ldots, \lambda_n\}$ se e solo se $Sp(A^k) = \{\lambda_1^k, \ldots, \lambda_n^k\}$. Allora per ogni $\lambda \in Sp(\mathbf{P})$, $\lambda \neq 1$, da $|\lambda^k| < 1$ si ha $|\lambda| < 1$. □

Dai teoremi precedenti si ottiene che la forma canonica di Jordan di una matrice di transizione regolare $\mathbf{P}$ è della forma

$$\mathbf{J} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{J}_1 & 0 & \ldots & 0 \\ 0 & 0 & \mathbf{J}_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \mathbf{J}_p \end{pmatrix},$$

dove $\mathbf{J}_1, \ldots, \mathbf{J}_p$ sono blocchi di Jordan relativi agli autovalori diversi da 1. Poichè

$$\mathbf{J}^n = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{J}_1^n & 0 & \ldots & 0 \\ 0 & 0 & \mathbf{J}_2^n & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \mathbf{J}_p^n \end{pmatrix},$$

si ha che $\mathbf{J}^n$ converge a

$$\mathbf{K} := \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}$$

Sia $\mathbf{S}$ una matrice tale che $\mathbf{S}^{-1}\mathbf{P}\mathbf{S} = \mathbf{J}$, allora da $\mathbf{P}^n = \mathbf{S}\mathbf{J}^n\mathbf{S}^{-1}$ si ottiene la convergenza di $\mathbf{P}^n$ a $\mathbf{S}\mathbf{K}\mathbf{S}^{-1}$.

Capitolo 2

Metodi Markov Chain Monte Carlo

I metodi Markov Chain Monte Carlo (MCMC) rappresentano una classe di algoritmi usati per il campionamento di variabili aleatorie, ovvero, data una densità di probabilità (discreta o continua) π su $\mathbb{R}^d$, i metodi permettono di costruire una sequenza $\{X_i\}_{i=1}^n$ di variabili i.i.d. con densità π. Lo scopo è calcolare l'integrale $\int_{\mathbb{R}^d} f(x)\pi(x)dx$, che corrisponde al valore atteso di $f(X_i)$, con $f : \mathbb{R}^d \to \mathbb{R}$ boreliana limitata, e che quindi può essere approssimato con $\frac{1}{n}\sum_{i=1}^n f(X_i)$ (per la legge forte dei grandi numeri). L'idea alla base di questi algoritmi è costruire una catena di Markov che ha π come distribuzione invariante. In questo modo, dopo un numero sufficiente di passi, cioè dopo un periodo di "burn-in" abbastanza lungo, si otterrà un campione da una densità che approssima π.
In questo capitolo vengono presentati due algoritmi MCMC: l'algoritmo di Metropolis-Hastings e il campionamento di Gibbs.
Consideriamo il caso in cui $S := \{x \in \mathbb{R}^d \mid \pi(x) > 0\}$ è discreto.

2.1 Algoritmo di Metropolis-Hastings

Sia $\pi = (\pi_x)_{x\in S}$ la densità dalla quale si deve estrarre il campione, chiamata anche densità target. Siano $\mathbf{Q} = (q_{xy})_{x,y\in S}$ una matrice stocastica, e $x^{(0)}$ un qualsiasi elemento in S. Se $x^{(t)} = x$, $x^{(t+1)}$ viene generato come segue. Si estrae $y \in S$ dalla densità $(q_{xz})_{z\in S}$, che viene considerato come candidato per $x^{(t+1)}$. Si costruisce un'altra funzione di probabilità $\alpha : S \times S \to [0, 1]$

$$\alpha(x, z) = \begin{cases} \min\left\{\frac{\pi_z q_{zx}}{\pi_x q_{xz}}, 1\right\} & \text{se } \pi_x q_{xz} > 0 \\ 1 & \text{se } \pi_x q_{xz} = 0 \end{cases}.$$

Infine si accetta y, ponendo $x^{(t+1)} = y$, con probabilità $\alpha(x, y)$, altrimenti si rimane nello stato x, ponendo $x^{(t+1)} = x$ con probabilità $1 - \alpha(x, y)$. Notiamo che $\alpha(x, x) = 1$, quindi se viene estratto x come candidato, lo stato corrente non cambia, pur avendo accettato.
La catena di Markov omogenea $\{X^{(t)}\}_{t\geq 0}$ che viene creata in questo modo ha matrice di transizione $\mathbf{P} = (p_{xy})$ tale che

$$p_{xy} = \begin{cases} q_{xy}\alpha(x, y) & \text{se } x \neq y \\ 1 - \sum_{z\neq x} q_{xz}\alpha(x, z) & \text{se } x = y \end{cases}.$$

Questo algoritmo è una generalizzazione dell'algoritmo di Metropolis, che richiede che $\mathbf{Q}$ sia simmetrica.

Proposizione 2.1. *La distribuzione target π è invariante per $\{X^{(t)}\}$.*

Dimostrazione. Per verificare che $\pi\mathbf{P} = \pi$, è sufficiente dimostare che per ogni $x, y \in S$ vale $p_{xy}\pi_x = p_{yx}\pi_y$. Infatti, in tal caso, si avrebbe

$$(\pi\mathbf{P})_x = \sum_{y\in S} \pi_y p_{yx} = \sum_{y\in S} \pi_x p_{xy} = \pi_x \sum_{y\in S} p_{xy} = \pi_x.$$

Nel caso $\pi_x q_{xy} > 0$, $\alpha(x,y)$ si può riscrivere come

$$\alpha(x,y) = \frac{1}{\pi_x q_{xy}} \min\{\pi_y q_{yx}, \pi_x q_{xy}\}.$$

Supponendo $\min\{\pi_y q_{yx}, \pi_x q_{xy}\} > 0$, e $x \neq y$, si ha

$$\pi_x p_{xy} = \pi_x q_{xy} \alpha(x,y) = \min\{\pi_y q_{yx}, \pi_x q_{xy}\} = \pi_y q_{yx} \alpha(y,x) = \pi_y p_{yx}$$

Nel caso $\min\{\pi_y q_{yx}, \pi_x q_{xy}\} = 0$, senza perdita di generalità si può supporre $\pi_y q_{yx} = 0$. Se $\pi_x q_{xy} = 0$, l'uguaglianza è ovvia, se $\pi_x q_{xy} > 0$, allora $\alpha(x,y) = 0$, quindi $p_{xy}\pi_x = p_{yx}\pi_y = 0$. □

Osservazione 2.1. Supponiamo $\{X^{(t)}\}$ definita sullo spazio $(\Omega, \mathcal{E}, \mathbb{P})$. Per ogni $x \in S$, definiamo $r(x) := \sum_{z \neq x} q_{xz}\alpha(x,z)$ la probabilità di cambiare stato, quando lo stato corrente è x. Sia $K = (K_{xy})_{x,y \in S}$ la matrice substocastica (cioè con tutte le entrate positive o nulle e tale che la somma sulle righe sia minore o uguale a 1) definita da:

$$K_{xy} := \begin{cases} \alpha(x,y) q_{xy} & \text{se } x \neq y \\ 0 & \text{se } x = y. \end{cases}$$

Quando $x \neq y$, K_{xy} rappresenta la probabilità di passare dallo stato x allo stato y. Notiamo che $K_{xy}^{(t)}$ è la probabilità di passare da x a y in t passi, cambiando almeno una volta stato. Nel caso di Metropolis-Hastings, con queste definizioni si ha, per $t \geq 1$,

$$K_{xy}^{(t)} = \begin{cases} p_{xy}^{(t)} & \text{se } x \neq y \\ p_{xx}^{(t)} - (1 - r(x))^t & \text{se } x = y. \end{cases}$$

Dati $x, y \in S$, sia $A(x,y,t) = \{\omega \in \Omega \mid X^{(t)}(\omega) = y,\, X^{(0)}(\omega) = x,\, \exists s\, \exists z \,:\, 1 \leq s \leq t,\, z \neq x,\, X^{(s)}(\omega) = z\}$ l'insieme dei punti le cui traiettorie sono non costanti e partono da x e al tempo t si trovano in y. Per definizione,

$$\mathbb{P}(A(x,y,t) \mid X_0 = x) = K_{xy}^{(t)}.$$

$A(x,y,t)$ si può scrivere come unione disgiunta degli insiemi $B_1, B_2 \subset \Omega$, dove B_1 è l'insieme delle traiettorie da x a y passanti per $z \in S \setminus \{x\}$ al tempo $t-1$, B_2 è l'insieme della traiettorie non costanti che passano in x al tempo $t-1$ e al tempo t si spostano in y. Con queste definizioni si ha

$$\begin{aligned}
\mathbb{P}(B_1 \mid X^{(0)} = x) &= \sum_{z \neq x} \mathbb{P}(X^{(t)} = y \mid X^{(t-1)} = z) \mathbb{P}(X^{(t-1)} = z \mid X^{(0)} = x) \\
&= \sum_{z \neq x} p_{zy} K_{xz}^{(t-1)} \\
&= \begin{cases} \sum_{\substack{z \neq x \\ z \neq y}} K_{zy} K_{xz}^{(t-1)} + (1 - r(y)) K_{xy}^{(t-1)} & \text{se } x \neq y \\ \sum_{z \neq x} K_{zy} K_{xz}^{(t-1)} & \text{se } x = y \end{cases} \\
\mathbb{P}(B_2 \mid X^{(0)} = x) &= \begin{cases} p_{xy} p_{xx}^{(t-1)} & \text{se } x \neq y \\ p_{xx} \mathbb{P}(A(x,x,t-1) \mid X^{(0)} = x) & \text{se } x = y \end{cases} \\
&= \begin{cases} p_{xy} K_{xx}^{(t-1)} + p_{xy}(1 - r(x))^{t-1} & \text{se } x \neq y \\ p_{xx} K_{xx}^{(t-1)} & \text{se } x = y \end{cases} \\
&= \begin{cases} K_{xy} K_{xx}^{(t-1)} + K_{xy}(1 - r(x))^{t-1} & \text{se } x \neq y \\ (1 - r(x)) K_{xx}^{(t-1)} & \text{se } x = y \end{cases}
\end{aligned}$$

Poichè l'unione è disgiunta,

$$\mathbb{P}(A(x,y,t)) \mid X^{(0)} = x) = \sum_{i=1}^{2} \mathbb{P}(B_i \mid X^{(0)} = x).$$

Allora, se $x \neq y$, usando la definizione di K e il fatto che $K_{yy} = 0$ per ogni $y \in S$, si può scrivere

$$\begin{aligned} K_{xy}^{(t)} &= \sum_{\substack{z \neq x \\ z \neq y}} K_{zy} K_{xz}^{(t-1)} + (1 - r(y)) K_{xy}^{(t-1)} + K_{xy} K_{xx}^{(t-1)} + K_{xy}(1 - r(x))^{t-1} \\ &= \sum_{z} K_{zy} K_{xz}^{(t-1)} + (1 - r(y)) K_{xy}^{(t-1)} + K_{xy}(1 - r(x))^{t-1} \end{aligned}$$

Se $x = y$,

$$\begin{aligned} K_{xx}^{(t)} &= \sum_{z \neq x} K_{zy} K_{xz}^{(t-1)} + (1 - r(x)) K_{xx}^{(t-1)} \\ &= \sum_{z} K_{zy} K_{xz}^{(t-1)} + (1 - r(x)) K_{xx}^{(t-1)} + K_{xx}(1 - r(x))^{t-1} \end{aligned}$$

Allora si ottiene la seguente formula iterativa per le potenze di K:

$$K_{xy}^{(t)} = \sum_{z} K_{zy} K_{xz}^{(t-1)} + (1 - r(y)) K_{xy}^{(t-1)} + K_{xy}(1 - r(x))^{t-1}.$$

Teorema 2.1. *Supponiamo $q_{xy} = 0$ se e solo se $q_{yx} = 0$. Se $\mathbf{Q}$ è irriducibile, allora $\mathbf{P}$ è irriducibile; se $\mathbf{Q}$ è regolare, allora $\mathbf{P}$ è regolare.*

Dimostrazione. L'ipotesi $q_{xy} = 0$ se e solo se $q_{yx} = 0$ implica che $\alpha(x, y) > 0$ per ogni $x, y \in S$. Definiamo, per ogni $t \geq 1$, $U_x^{(t)} = \{y \in S \mid p_{xy}^{(t)} > 0\}$ e $V_x^{(t)} = \{y \in S \mid q_{xy}^{(t)} > 0\}$. Si può dimostrare per induzione che $V_x^{(t)} \subset U_x^{(t)}$, e infine, da questo fatto si può dedurre che l'irriducibilità di $\mathbf{Q}$ implica l'irriducibilità di $\mathbf{P}$, e la regolarità di $\mathbf{Q}$ implica la regolarità di $\mathbf{P}$. Se $y \in V_x^{(1)}$, $y \neq x$, allora $p_{xy} = q_{xy}\alpha(x, y) > 0$. Se $x \in V_x^{(1)}$, allora

$$p_{xx} = 1 - \sum_{z \neq x} p_{xz} = \sum_{z \neq x} q_{xz}(1 - \alpha(x, z)) + q_{xx} > 0.$$

Supponiamo $V_x^{(t)} \subset U_x^{(t)}$ e consideriamo $y \in V_x^{(t+1)}$. Poichè $q_{xy}^{(t+1)} = \sum_{z \in S} q_{xz}^{(t)} q_{zy} > 0$, esiste $z \in S$ tale che $q_{xz}^{(t)} q_{zy} > 0$, cioè $z \in V_x^{(t)}$ e $y \in V_z^{(1)} \subset U_z^{(1)}$. Allora per ipotesi $z \in U_x^{(t)}$, quindi $p_{xy}^{(t+1)} = \sum_{z \in S} p_{xz}^{(t)} p_{zy} > 0$, ovvero $y \in U_x^{(t+1)}$. □

2.2 Campionamento di Gibbs

Consideriamo ogni elemento di $\mathbb{R}^d$ partizionato in r blocchi, cioè $x = (x_1, \ldots, x_r)$, dove per ogni $i \in \{1, \ldots, r\}$ $x_i = (x_{i1}, \ldots, x_{in(i)})$, $n(i) \geq 1$ e $n(1) + \cdots + n(r) = d$, con $x_{ij} \in \mathbb{R}$. Data π la densità da approssimare, denotiamo con

$$\pi(x_i \mid \bar{x}_1, \ldots, \bar{x}_{i-1}, \bar{x}_{i+1}, \ldots, \bar{x}_r) = \frac{\pi(\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_r)}{\sum_{y \in \mathbb{R}^{n(i)}} \pi(\bar{x}_1, \ldots, \bar{x}_{i-1}, y, \bar{x}_{i+1}, \ldots, \bar{x}_r)}$$

la densità condizionale di x_i dati i valori $\bar{x}_j \in \mathbb{R}^{n(j)}$ delle altre componenti.
L'algoritmo parte da un generico $x^{(0)} = (x_1^{(0)}, \ldots, x_r^{(0)}) \in S$. Alla $(t+1)$-esima iterazione viene generato il campione $x^{(t+1)}$ come segue:
si estrae $x_1^{(t+1)}$ da $\pi(x_1 \mid x_2^{(t)}, \ldots, x_r^{(t)})$;
si estrae $x_2^{(t+1)}$ da $\pi(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_r^{(t)})$;
$\vdots$
si estrae $x_j^{(t+1)}$ da $\pi(x_j \mid x_1^{(t+1)}, \ldots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \ldots, x_r^{(t)})$;
$\vdots$
si estrae $x_r^{(t+1)}$ da $\pi(x_r \mid x_1^{(t+1)}, \ldots, x_{r-1}^{(t+1)})$.

Il vantaggio è che, anche quando d è molto grande, si campiona da densità in dimensioni minori. La sequenza $\{x^{(0)}, x^{(1)}, \dots, x^{(t)}, \dots\}$ è una traiettoria di una catena di Markov omogenea $\{X^{(t)}\}_{t\geq 0}$ (l'estrazione di $x^{(t+1)}$ non dipende da $x^{(t-1)}, \dots, x^{(0)}$) definita in un certo spazio di probabilità $(\Omega, \mathcal{E}, \mathbb{P})$ e con matrice di transizione $\mathbf{P} = (p_{xy})_{x,y\in S}$ definita da

$$p_{xy} := \prod_{l=1}^{r} \pi(y_l \mid y_i,\, i < l,\, x_j,\, j > l)$$

Infatti,

$$\begin{aligned} p_{xy} &= \mathbb{P}(X^{(1)} = y \mid X^{(0)} = x) \\ &= \mathbb{P}(X_r^{(1)} = y_r \mid X_1^{(1)} = y_1, \dots X_{r-1}^{(1)} = y_{r-1}, X^{(0)} = x) \cdots \mathbb{P}(X_2^{(1)} = y_2 \mid X_1^{(1)} = y_1, X^{(0)} = x) \\ &= \pi(y_r \mid y_1, \dots, y_{r-1}, x_r) \cdots \pi(y_2 \mid y_1, x_3, \dots, x_r)\pi(y_1 \mid x_2, \dots x_r). \end{aligned}$$

Proposizione 2.2. *La distribuzione π è invariante per $\{X^{(t)}\}$.*

Dimostrazione. Consideriamo la catena di Markov (non necessariamente omogenea) $\{Z^{(k)}\}_{k\geq 0}$ definita in questo modo:

$$\begin{aligned} &Z^{(0)} := X^{(0)}; \\ &Z^{(1)} := (X_1^{(1)}, X_2^{(0)}, \dots, X_r^{(0)}); \\ &Z^{(2)} := (X_1^{(1)}, X_2^{(1)}, X_3^{(0)}, \dots, X_r^{(0)}); \\ &\vdots \\ &Z^{(r)} := (X_1^{(1)}, \dots, X_r^{(1)}) = X^{(1)}; \\ &Z^{(r+1)} := (X_1^{(2)}, X_2^{(1)}, \dots, X_r^{(1)}); \\ &\vdots \end{aligned}$$

Supponiamo che $Z^{(k)} = (X_1^{(t+1)}, \dots, X_{l-1}^{(t+1)}, X_l^{(t)}, \dots, X_r^{(t)})$ abbia densità π. Siano $z = (z_1, \dots, z_r) \in \mathbb{R}^d$ e $z_{x,l} := (z_1, \dots, z_{l-1}, x, z_{l+1}, \dots, z_r) \in \mathbb{R}^d$, con $l \in \{1, \dots, r\}$ e $x \in \mathbb{R}^{n(l)}$.

$$\begin{aligned} \mathbb{P}(Z^{(k+1)} = z) &= \sum_x \mathbb{P}(Z^{(k+1)} = z \mid Z^{(k)} = z_{x,l})\mathbb{P}(Z^{(k)} = z_{x,l}) \\ &= \sum_x \mathbb{P}(X_l^{(t+1)} = z_l \mid X_i^{(t+1)} = z_i,\, i < l,\, X_j^{(t)} = z_j,\, j > l)\pi(z_{x,l}) \\ &= \sum_x \pi(z_l \mid z_j,\, j \neq l)\pi(z_{x,l}) \\ &= \sum_x \frac{\pi(z)}{\sum_y \pi(z_{y,l})}\pi(z_{x,l}) = \pi(z), \end{aligned}$$

ovvero anche $Z^{(k+1)}$ ha densità π. Quindi, se $X^{(0)}$ è distribuita secondo π, anche $Z^{(k)}$ lo è per ogni k, ed in particolare $X^{(t)}$ per ogni t. Di conseguenza, π è una distribuzione invariante per $\{X^{(t)}\}$. □

Lemma 2.1. *Se S è finito e $\mathbf{P}$ è irriducibile, allora è anche regolare.*

Dimostrazione. Per ogni $x \in S$ e per ogni $l \in \{1, \dots, r\}$, $\pi(x_l \mid x_i, i \neq l) > 0$ per definizione di S, quindi $p_{xx} = \prod_l \pi(x_l \mid x_i, i \neq l) > 0$. Allora dal lemma 1.1 segue la tesi. □

Da questo lemma e da quanto dimostrato nel capitolo 1, segue che nel caso in cui S è finito e $\mathbf{P}$ irriducibile, le densità marginali della catena convergono a π, quindi il metodo è convergente. Nella pratica spesso si considera S con una struttura a reticolo. In questo modo, la matrice di transizione non ha entrate nulle, quindi è regolare.

Per produrre un campione da π di taglia n, cioè n campioni indipendenti, si può far partire l'algoritmo n volte, generando n catene di Markov con variabili iniziali indipendenti. Un altro modo è fissare un k opportuno e, dopo il periodo di "burn-in", considerare un campione ogni k iterazioni del metodo.

2.3 Esempio di applicazione del metodo di Gibbs

Consideriamo un'urna contentente N palline rosse e blu. Sia X la variabile a valori in $\{\frac{j}{N}\}_{j=0,\dots,N}$ che rappresenta la frazione di palline rosse nell'urna. Fissati due parametri $\alpha, \beta > 0$, scegliamo X con densità discreta π_X tale che per ogni $x \in \{\frac{j}{N}\}_{j=0,\dots,N}$

$$\pi_X(x) = c_{\alpha,\beta}\, x^{\alpha-1}(1-x)^{\beta-1},$$

dove

$$c_{\alpha,\beta} = \left[\sum_{j=0}^{N} \left(\frac{j}{N}\right)^{\alpha-1}\left(1-\frac{j}{N}\right)^{\beta-1}\right)\right]^{-1}$$

è la costante di normalizzazione. In questo modo π_X è una discretizzazione della distribuzione continua $Beta(\alpha, \beta)$.
Fissato $n \in \mathbb{N}$, sia K la variabile che conta il numero di palline rosse estratte in n estrazioni con rimpiazzo.
Usiamo il metodo di Gibbs per campionare (X, K). La densità condizionata di K rispetto a $\{X = x\}$ è una binomiale di parametri n e x, cioè

$$\pi_{K|X}(k \mid x) = \binom{n}{k} x^k (1-x)^{n-k}.$$

Invece, usando la formula di Bayes, la densità di X sapendo $\{K = k\}$ è definita da

$$\begin{aligned}
\pi_{X|K}(x \mid k) &= \frac{\pi_{K|X}(k \mid x)\pi_X(x)}{\sum_{j=0}^{N} \pi_{K|X}(k \mid \frac{j}{N})\pi_X(\frac{j}{N})} \\
&= \frac{\binom{n}{k}x^k(1-x)^{n-k}c_{\alpha,\beta}\, x^{\alpha-1}(1-x)^{\beta-1}}{\sum_{j=0}^{N}\binom{n}{k}(\frac{j}{N})^k(1-\frac{j}{N})^{n-k}c_{\alpha,\beta}(\frac{j}{N})^{\alpha-1}(1-\frac{j}{N})^{\beta-1}} \\
&= c_{\alpha+k,\beta+n-k}\, x^{\alpha+k-1}(1-x)^{\beta+n-k-1},
\end{aligned}$$

che è una discretizzazione della distribuzione $Beta(\alpha + k, \beta + n - k)$.
Si può calcolare anche la densità discreta di K, π_K, in questo modo:

$$\begin{aligned}
\pi_K(k) &= \sum_x \pi_{K|X}(k \mid x)\pi_X(x) \\
&= \sum_x \binom{n}{k} x^k(1-x)^{n-k}c_{\alpha,\beta}x^{\alpha-1}(1-x)^{\beta-1} \\
&= \binom{n}{k} c_{\alpha,\beta}c^{-1}_{\alpha+k,\beta+n-k}.
\end{aligned}$$

Di seguito il codice in MATLAB usato per implementare il metodo. La funzione prende in input i parametri N, n, α e β, il numero di campioni da estrarre *npoints*, il numero di iterazioni *it* per il periodo di "burn-in" e un numero m. Se $m \leq 0$ la funzione genera i campioni da catene di Markov diverse, facendo ripartire l'algoritmo tante volte quante il numero di punti da generare. Se $m > 0$, dopo il periodo di "burn-in", i campioni vengono selezionati dalla stessa catena di Markov ogni m passi. La funzione restituisce una matrice che in ogni riga ha un campione.

```
function [z] = betabin(N,n,alpha,beta,npoints,it,m)
S = [0:N]./N;
if(m<=0)
    for np = 1:npoints
        x = alpha/(alpha+beta);
        k = floor(n*alpha/(alpha+beta));
        for i = 1:it
            for j=0:N
                W(j+1) = (j/N)^(alpha+k-1)*(1-j/N)^(beta+n-k-1);
            end
            W = W./sum(W);
            x = randsample(S,1,true,W);
            k = binornd(n,x);
        end
        z(np,:) = [x,k];
    end
else
    x = alpha/(alpha+beta);
    k = floor(n*alpha/(alpha+beta));
    for i = 1:it
        for j=0:N
            W(j+1) = (j/N)^(alpha+k-1)*(1-j/N)^(beta+n-k-1);
        end
        W = W./sum(W);
        x = randsample(S,1,true,W);
        k = binornd(n,x);
    end
    for np = 1:npoints
        z(np,:) = [x,k];
        i = 0;
        for i = 1:m
            for j=0:N
                W(j+1) = (j/N)^(alpha+k-1)*(1-j/N)^(beta+n-k-1);
            end
            W = W./sum(W);
            x = randsample(S,1,true,W);
            k = binornd(n,x);
        end
    end
end
end
```

Listing 2.1: Metodo di Gibbs per la densità di (X, K)

Riportiamo i grafici a barre delle densità marginali a confronto con le approssimazioni di queste date dall'algoritmo, con periodo di "burn-in" di 10 iterazioni. Sono stati generati 10000 campioni considerando $N = 15$ e $n = 10$, chiamando la funzione *betabin* con $m = 0$ e $m = 10$. Per le figure 2.1 e 2.2 sono stati scelti $\alpha = 2$ e $\beta = 5$, per le figure 2.3 e 2.4 $\alpha = 50$ e $\beta = 100$.

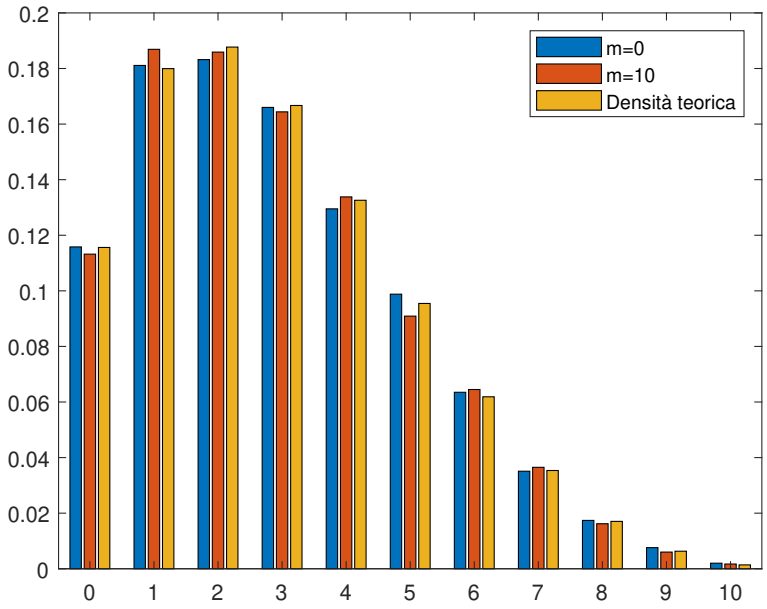


Figura 2.1: Grafici a barre della densità (marginale) di K e delle sue approssimazioni ottenute con il metodo di Gibbs, con $N = 15$, $n = 10$, $\alpha = 2$, $\beta = 5$.

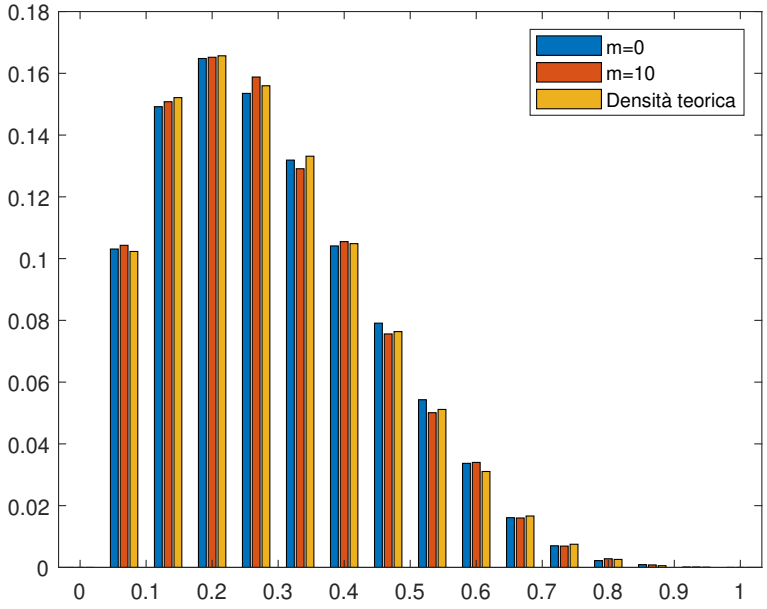


Figura 2.2: Grafici a barre della densità (marginale) di X e delle sue approssimazioni ottenute con il metodo di Gibbs, con $N = 15$, $n = 10$, $\alpha = 2$, $\beta = 5$.

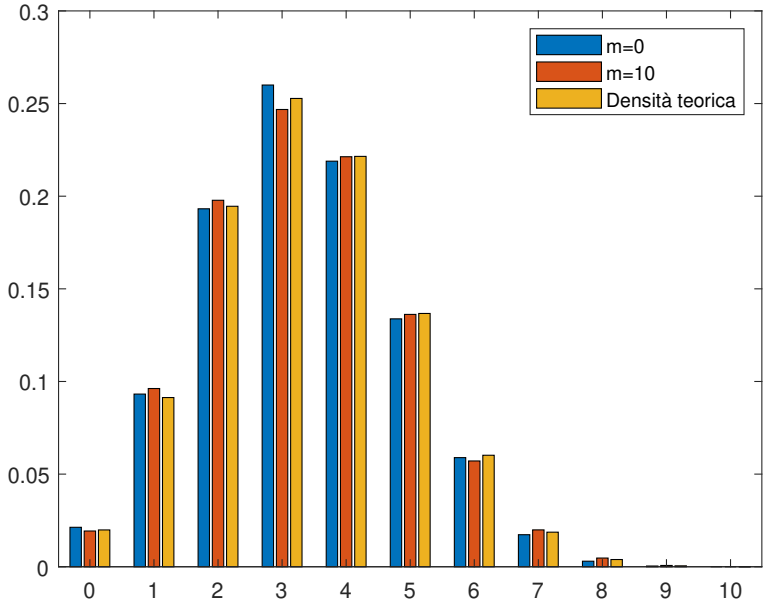


Figura 2.3: Grafici a barre della densità (marginale) di K e delle sue approssimazioni ottenute con il metodo di Gibbs, con $N = 15$, $n = 10$, $\alpha = 50$, $\beta = 100$.

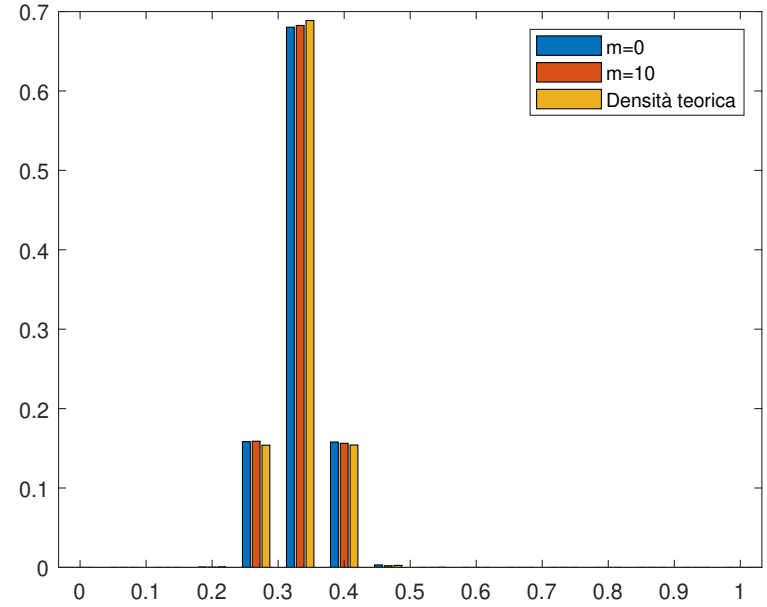


Figura 2.4: Grafici a barre della densità (marginale) di X e delle sue approssimazioni ottenute con il metodo di Gibbs, con $N = 15$, $n = 10$, $\alpha = 50$, $\beta = 100$.

Capitolo 3

Il metodo di Gibbs per densità gaussiane multivariate

I concetti esposti nel capitolo precedente si possono generalizzare al caso di densità continue ([3]). In questo capitolo vediamo come si applica il campionamento di Gibbs a densità gaussiane multivariate.

Definizione 3.1. Un vettore aleatorio $X = (X_1, \dots, X_d)$ a valori in $\mathbb{R}^d$ è un *vettore gaussiano* se per ogni $u = (u_1, ..., u_d) \in \mathbb{R}^d$ si ha che $\langle X, u\rangle = \sum_{i=1}^{d} u_i X_i$ è una variabile gaussiana.
Il vettore delle medie $m = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]) \in \mathbb{R}^d$, e la matrice delle covarianze $Q \in \mathbb{R}^{d\times d}$, con $(Q)_{ij} = Cov(X_i, X_j)$, identificano la legge di X, e si scrive $X \sim \mathcal{N}(m, Q)$.

Notazione. Dati due vettori aleatori X, Y a valori in $\mathbb{R}^m$ e $\mathbb{R}^n$ rispettivamente, indichiamo con $Cov(X, Y) \in \mathbb{R}^{m\times n}$ la matrice tale che $(Cov(X, Y))_{ij} = Cov(X_i, Y_j)$. In accordo alla notazione usata per variabili unidimensionali, indichiamo $Cov(X, X)$ con $Var(X)$.

Proposizione 3.1. *Sia $(X, Y) \sim \mathcal{N}(m, Q)$ vettore gaussiano a valori in $\mathbb{R}^d$, con X a valori in $\mathbb{R}^m$ e Y a valori in $\mathbb{R}^n$, e $Q = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, $A \in \mathbb{R}^{m\times m}, B \in \mathbb{R}^{m\times n}, C \in \mathbb{R}^{n\times n}$. Allora la legge di X condizionata a $\{Y = y\}$ è una distribuzione gaussiana a valori in $\mathbb{R}^m$ di media $\mathbb{E}[X] + BC^{-1}(y - \mathbb{E}[Y])$ e varianza $A - BC^{-1}B^T$. Notiamo che la varianza non dipende da y.*

Dimostrazione. Cerchiamo $K \in \mathbb{R}^{m\times n}$ tale che $Z := X + KY$ sia indipendente da Y. Poichè (Z, Y) è un vettore gaussiano, Z e Y sono indipendenti se e solo se $Cov(Z, Y) = 0 \in \mathbb{R}^{m\times n}$. Sfruttando le proprietà della matrice delle covarianze, si può scrivere

$$Cov(Z, Y) = Cov(X, Y) + Cov(KY, Y) = B + KVar(Y) = B + KC,$$

da cui,

$$Cov(Z, Y) = 0 \iff K = -BC^{-1}.$$

Poichè $\mathbb{E}[X \mid Y]$ è $\sigma(Y)$-misurabile, per il lemma di Doob esiste $g : \mathbb{R}^n \to \mathbb{R}^m$ misurabile tale che $\mathbb{E}[X \mid Y] = g(Y)$. Indichiamo $\mathbb{E}[X \mid Y = y] := g(y)$. Per la varianza condizionale usiamo una notazione analoga. Ricordando che se $X = \phi(Y, Z)$, con Y e Z variabili indipendenti e ϕ boreliana limitata, vale $\mathbb{E}[X \mid Y = y] = \mathbb{E}[\phi(y, Z)]$, si ha

$$\begin{aligned}\mathbb{E}[X \mid Y = y] &= \mathbb{E}[Z - KY \mid Y = y] = \mathbb{E}[Z - Ky] = \mathbb{E}[Z] - Ky \\ &= \mathbb{E}[X] + K\mathbb{E}[Y] - Ky = \mathbb{E}[X] + BC^{-1}(y - \mathbb{E}[Y]),\end{aligned}$$

$$\begin{aligned}Var(X \mid Y = y) &= Var(Z - KY \mid Y = y) = Var(Z - ky) \\ &= Var(Z) = Var(X + KY) = Var(X) + Var(KY) + Cov(X, KY) + Cov(KY, X) \\ &= Var(X) + KVar(Y)K^T + Cov(X, KY)^T + Cov(KY, X) \\ &= Var(X) + KVar(Y)K^T + 2KCov(Y, X) \\ &= A + BC^{-1}CC^{-1}B^T - 2BC^{-1}B^T = A - BC^{-1}B^T.\end{aligned}$$

□

Osservazione 3.1. Se $Q = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, allora $A - BC^{-1}B^T$ è il complemento di Schur di C in Q. Dalla relazione di congruenza tra matrici, si ha che Q è definita positiva se e solo se $A - BC^{-1}B^T$ e C sono definite positive, infatti,

$$\begin{pmatrix} A - BC^{-1}B^T & 0 \\ 0 & C \end{pmatrix} = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix}^T.$$

3.1 Caso d=2

Sia $X = (X_1, X_2) \sim \mathcal{N}(m, Q)$ gaussiana bidimensionale, con $m = (m_1, m_2) \in \mathbb{R}^2$, $Q = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathbb{R}^{2\times 2}$. Per ogni $k \geq 0$ sia $X^{(k)} = (X_1^{(k)}, X_2^{(k)})$ la varibile ottenuta alla fine della k-esima iterazione del metodo di Gibbs, con media $m^{(k)} = (m_1^{(k)}, m_2^{(k)})$ e varianza $Q^{(k)} = \begin{pmatrix} v_1^{(k)} & v_{12}^{(k)} \\ v_{12}^{(k)} & v_2^{(k)} \end{pmatrix}$. Supponiamo che $X_1^{(0)}$ e $X_2^{(0)}$ siano gaussiane (anche costanti) indipendenti. La matrice Q è semidefinita positiva, quindi in particolare $\det Q = ac - b^2 \geq 0$, da cui $\frac{b^2}{ac} \leq 1$. Sia $\delta := \frac{b^2}{ac}$.

Teorema 3.1. *Se Q è invertibile, gli errori di approssimazione della media, varianza e covarianza dovuti all'algoritmo convergono a zero esponenzialmente. Più precisamente,*

$$\begin{aligned} |m_1^{(k)} - m_1| &= \delta^k |m_1^{(0)} - m_1| \\ |m_2^{(k)} - m_2| &= \delta^k |m_2^{(0)} - m_2| \\ |v_1^{(k)} - a| &= \delta^{2k} |v_1^{(0)} - a| \\ |v_2^{(k)} - c| &= \delta^{2k} |v_2^{(0)} - c| \\ |v_{12}^{(k)} - b| &= \delta^{2k} \frac{|b|}{a} |v_1^{(0)} - a| \end{aligned} \tag{3.1}$$

Inoltre, anche nel caso in cui $\det Q = 0$, se le densità marginali di X sono note, considerando $X^{(0)} \sim \mathcal{N}\Big((m_1, m_2), \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}\Big)$, l'algoritmo converge in un passo.

Per dimostare il teorema, vediamo nel dettaglio come agisce il metodo per campionare X.
Sia $(x_1^{(k)}, x_2^{(k)})$ il valore assunto da $X^{(k)}$. Dalla proposizione 3.1 si deduce che

$$\begin{aligned} X_1^{(k+1)} &\sim \mathcal{N}\Big(m_1 + \frac{b}{c}(x_2^{(k)} - m_2), a - \frac{b^2}{c}\Big) \\ X_2^{(k+1)} &\sim \mathcal{N}\Big(m_2 + \frac{b}{a}(x_1^{(k+1)} - m_1), c - \frac{b^2}{a}\Big). \end{aligned}$$

Si dimostrano le seguenti proposizioni.

Proposizione 3.2. *Per ogni $k \geq 0$ valgono le formule iterative:*

$$\begin{cases} m_1^{(k+1)} = m_1 + \frac{b}{c}(m_2^{(k)} - m_2) \\ m_2^{(k+1)} = m_2 + \frac{b}{a}(m_1^{(k+1)} - m_1) \end{cases} \tag{3.2}$$

$$\begin{cases} v_1^{(k+1)} = (\frac{b}{c})^2 v_2^{(k)} + a - \frac{b^2}{c} \\ v_2^{(k+1)} = (\frac{b}{a})^2 v_1^{(k+1)} + c - \frac{b^2}{a} \end{cases} \tag{3.3}$$

$$v_{12}^{(k)} = \frac{b}{a} v_1^{(k)}. \tag{3.4}$$

Inoltre, se $i \in \{1, 2\}$, $X_i^{(k)}$ ha densità gaussiana $\mathcal{N}(m_i^{(k)}, v_i^{(k)})$.

Dimostrazione. Definiamo $r := \frac{b}{c}$, $s := \frac{b}{a}$, $v_1 := a - \frac{b^2}{c}$ e $v_2 := c - \frac{b^2}{a}$. Per individuare la legge di $X_i^{(k)}$ basta calcolare la funzione caratteristica $\varphi_{X_i^{(k)}}$, ricordando che la funzione caratteristica di una gaussiana di media μ e varianza σ^2 è $\varphi(t) = \exp(it\mu - \frac{1}{2}t^2\sigma^2)$.
Procediamo per induzione. Dato che $X_1^{(1)} \sim \mathcal{N}(m_1 + r(x_2^{(0)} - m_2), v_1)$, si ha

$$\mathbb{E}[e^{itX_1^{(1)}} \mid X_2^{(0)} = x_2^{(0)}] = \exp(it(m_1 + r(x_2^{(0)} - m_2)) - \frac{1}{2}t^2 v_1),$$

allora

$$\mathbb{E}[e^{itX_1^{(1)}} \mid X_2^{(0)}] = \exp(it(m_1 + r(X_2^{(0)} - m_2)) - \frac{1}{2}t^2 v_1).$$

Quindi si ottiene

$$\begin{aligned}
\varphi_{X_1^{(1)}}(t) &= \mathbb{E}[e^{itX_1^{(1)}}] = \mathbb{E}[\mathbb{E}[e^{itX_1^{(1)}} \mid X_2^{(0)}]] \\
&= \mathbb{E}[\exp(it(m_1 + r(X_2^{(0)} - m_2)) - \frac{1}{2}t^2 v_1)] \\
&= \exp(it(m_1 - rm_2) - \frac{1}{2}t^2 v_1)\mathbb{E}[\exp(itrX_2^{(0)})] \\
&= \exp(it(m_1 - rm_2) - \frac{1}{2}t^2 v_1)\exp(itrm_2^{(0)} - \frac{1}{2}t^2 r^2 v_2^{(0)}) \\
&= \exp(it(m_1 - rm_2 + rm_2^{(0)}) - \frac{1}{2}t^2(r^2 v_2^{(0)} + v_1)),
\end{aligned}$$

cioè $X_1^{(1)} \sim \mathcal{N}(m_1 - rm_2 + rm_2^{(0)}, r^2 v_2^{(0)} + v_1)$.
Invece, $X_2^{(1)} \sim \mathcal{N}(m_2 + s(x_1^{(1)} - m_1), v_2)$, quindi

$$\mathbb{E}[e^{itX_2^{(1)}} \mid X_1^{(1)}] = \exp(it(m_2 + s(X_1^{(1)} - m_1)) - \frac{1}{2}t^2 v_2).$$

Allora si ha

$$\begin{aligned}
\varphi_{X_2^{(1)}}(t) &= \mathbb{E}[e^{itX_2^{(1)}}] = \mathbb{E}[\mathbb{E}[e^{itX_2^{(1)}} \mid X_1^{(1)}]] \\
&= \mathbb{E}[\exp(it(m_2 + s(X_1^{(1)} - m_1)) - \frac{1}{2}t^2 v_2)] \\
&= \exp(it(m_2 - sm_1) - \frac{1}{2}t^2 v_2)\mathbb{E}[\exp(itsX_1^{(1)})] \\
&= \exp(it(m_2 - sm_1) - \frac{1}{2}t^2 v_2)\exp(itsm_1^{(1)} - \frac{1}{2}t^2 s^2 v_1^{(1)}) \\
&= \exp(it(m_2 - sm_1 + sm_1^{(1)}) - \frac{1}{2}t^2(s^2 v_1^{(1)} + v_2)),
\end{aligned}$$

ovvero, $X_2^{(1)} \sim \mathcal{N}(m_2 - sm_1 + sm_1^{(1)}, s^2 v_1^{(1)} + v_2)$.
Infine calcoliamo $v_{12}^{(1)}$:

$$\begin{aligned}
v_{12}^{(1)} &= Cov(X_1^{(1)}, X_2^{(1)}) = \mathbb{E}[X_1^{(1)} X_2^{(1)}] - m_1^{(1)} m_2^{(1)} \\
&= \mathbb{E}[\mathbb{E}[X_1^{(1)} X_2^{(1)} \mid X_1^{(1)}]] - m_1^{(1)} m_2^{(1)} \\
&= \mathbb{E}[X_1^{(1)} \mathbb{E}[X_2^{(1)} \mid X_1^{(1)}]] - m_1^{(1)} m_2^{(1)} \\
&= \mathbb{E}[X_1^{(1)}(m_2 + s(X_1^{(1)} - m_1))] - m_1^{(1)} m_2^{(1)} \\
&= m_1^{(1)} m_2 + s\mathbb{E}[(X_1^{(1)})^2] - sm_1 m_1^{(1)} - m_1^{(1)} m_2^{(1)} \\
&= m_1^{(1)} m_2 + s(v_1^{(1)} + (m_1^{(1)})^2) - sm_1 m_1^{(1)} - m_1^{(1)} m_2^{(1)} \\
&= sv_1^{(1)} + m_1^{(1)}(m_2 - sm_1 + sm_1^{(1)} - m_2^{(1)}) = sv_1^{(1)}
\end{aligned}$$

Nel passo induttivo, supponiamo $X_1^{(k)} \sim \mathcal{N}(m_1^{(k)}, v_1^{(k)})$ e $X_2^{(k)} \sim \mathcal{N}(m_2^{(k)}, v_2^{(k)})$. Si procede esattamente come nel passo base:

$$\begin{aligned}
\varphi_{X_1^{(k+1)}}(t) &= \mathbb{E}[e^{itX_1^{(k+1)}}] = \mathbb{E}[\mathbb{E}[e^{itX_1^{(k+1)}} \mid X_2^{(k)}]] \\
&= \mathbb{E}[\exp(it(m_1 + r(X_2^{(k)} - m_2)) - \frac{1}{2}t^2 v_1)] \\
&= \exp(it(m_1 - rm_2) - \frac{1}{2}t^2 v_1)\mathbb{E}[\exp(itrX_2^{(k)})] \\
&= \exp(it(m_1 - rm_2) - \frac{1}{2}t^2 v_1)\exp(itrm_2^{(k)} - \frac{1}{2}t^2 r^2 v_2^{(k)}) \\
&= \exp(it(m_1 - rm_2 + rm_2^{(k)}) - \frac{1}{2}t^2(r^2 v_2^{(k)} + v_1)),
\end{aligned}$$

$$\begin{aligned}
\varphi_{X_2^{(k+1)}}(t) &= \mathbb{E}[e^{itX_2^{(k+1)}}] = \mathbb{E}[\mathbb{E}[e^{itX_2^{(k+1)}} \mid X_1^{(k+1)}]] \\
&= \mathbb{E}[\exp(it(m_2 + s(X_1^{(k+1)} - m_1)) - \frac{1}{2}t^2 v_2)] \\
&= \exp(it(m_2 - sm_1) - \frac{1}{2}t^2 v_2)\mathbb{E}[\exp(itsX_1^{(k+1)})] \\
&= \exp(it(m_2 - sm_1) - \frac{1}{2}t^2 v_2)\exp(itsm_1^{(k+1)} - \frac{1}{2}t^2 s^2 v_1^{(k+1)}) \\
&= \exp(it(m_2 - sm_1 + sm_1^{(k+1)}) - \frac{1}{2}t^2(s^2 v_1^{(k+1)} + v_2)),
\end{aligned}$$

ovvero $X_1^{(k+1)} \sim \mathcal{N}(m_1 - rm_2 + rm_2^{(k)}, r^2 v_2^{(k)} + v_1)$ e $X_2^{(k+1)} \sim \mathcal{N}(m_2 - sm_1 + sm_1^{(k+1)}, s^2 v_1^{(k+1)} + v_2)$. Inoltre,

$$\begin{aligned}
v_{12}^{(k+1)} &= Cov(X_1^{(k+1)}, X_2^{(k+1)}) = \mathbb{E}[X_1^{(k+1)} X_2^{(k+1)}] - m_1^{(k+1)} m_2^{(k+1)} \\
&= \mathbb{E}[\mathbb{E}[X_1^{(k+1)} X_2^{(k+1)} \mid X_1^{(k+1)}]] - m_1^{(k+1)} m_2^{(k+1)} \\
&= \mathbb{E}[X_1^{(k+1)} \mathbb{E}[X_2^{(k+1)} \mid X_1^{(k+1)}]] - m_1^{(k+1)} m_2^{(k+1)} \\
&= \mathbb{E}[X_1^{(k+1)}(m_2 + s(X_1^{(k+1)} - m_1))] - m_1^{(k+1)} m_2^{(k+1)} \\
&= m_1^{(k+1)} m_2 + s\mathbb{E}[(X_1^{(k+1)})^2] - sm_1 m_1^{(k+1)} - m_1^{(k+1)} m_2^{(k+1)} \\
&= m_1^{(k+1)} m_2 + s(v_1^{(k+1)} + (m_1^{(k+1)})^2) - sm_1 m_1^{(k+1)} - m_1^{(k+1)} m_2^{(k+1)} \\
&= sv_1^{(k+1)} + m_1^{(k+1)}(m_2 - sm_1 + sm_1^{(k+1)} - m_2^{(k+1)}) = sv_1^{(k+1)}
\end{aligned}$$

□

Proposizione 3.3. *Per ogni $k \geq 0$ $X^{(k)} = (X_1^{(k)}, X_2^{(k)})$ è un vettore gaussiano.*

Dimostrazione. Sia $t = (t_1, t_2) \in \mathbb{R}^2$. Calcoliamo la funzione caratteristica del vettore aleatorio $X^{(k)}$.

$$\begin{aligned}
\varphi_{X^{(k)}}(t) &= \mathbb{E}[e^{i\langle t, X^{(k)}\rangle}] \\
&= \mathbb{E}[e^{i(t_1 X_1^{(k)} + t_2 X_2^{(k)})}] \\
&= \mathbb{E}[e^{it_1 X_1^{(k)}} \mathbb{E}[e^{it_2 X_2^{(k)}} \mid X_1^{(k)}]] \\
&= \mathbb{E}[\exp(it_1 X_1^{(k)} + it_2(m_2 + s(X_1^{(k)} - m_1)) - \frac{1}{2}t_2^2 v_2)] \\
&= \mathbb{E}[\exp(i(t_1 + st_2)X_1^{(k)})]\exp(it_2(m_2 - sm_1) - \frac{1}{2}t_2^2 v_2) \\
&= \exp(i(t_1 + st_2)m_1^{(k)} - \frac{1}{2}(t_1 + st_2)^2 v_1^{(k)} + it_2(m_2 - sm_1) - \frac{1}{2}t_2^2 v_2) \\
&= \exp(i(t_1 m_1^{(k)} + t_2(m_2 + sm_1^{(k)} - sm_1)) - \frac{1}{2}(t_1^2 v_1^{(k)} + 2t_1 t_2 s v_1^{(k)} + t_2^2(s^2 v_1^{(k)} + v_2))) \\
&= \exp(i(t_1 m_1^{(k)} + t_2 m_2^{(k)}) - \frac{1}{2}(t_1^2 v_1^{(k)} + 2t_1 t_2 v_{12}^{(k)} + t_2^2 v_2^{(k)})) = \exp(i\langle t, m^{(k)}\rangle - \frac{1}{2}\langle tQ^{(k)}, t\rangle).
\end{aligned}$$

Allora $X^{(k)} \sim \mathcal{N}(m^{(k)}, Q^{(k)})$. □

Applicando una sostituzione, le formule (3.2) e (3.4) si possono riscrivere come

$$\begin{cases} m_1^{(k+1)} = (1-\delta)m_1 + \delta m_1^{(k)} \\ m_2^{(k+1)} = (1-\delta)m_2 + \delta m_2^{(k)}, \end{cases}$$

$$\begin{cases} v_1^{(k+1)} = (1-\delta^2)a + \delta^2 v_1^{(k)} \\ v_2^{(k+1)} = (1-\delta^2)c + \delta^2 v_2^{(k)}. \end{cases}$$

Da queste si ricavano induttivamente le formule del teorema 3.1 per gli errori alla k-esima iterazione:

$$\begin{aligned} |m_1^{(k+1)} - m_1| &= \delta |m_1^{(k)} - m_1| = \cdots = \delta^{k+1} |m_1^{(0)} - m_1| \\ |m_2^{(k+1)} - m_2| &= \delta |m_2^{(k)} - m_2| = \cdots = \delta^{k+1} |m_2^{(0)} - m_2| \\ |v_1^{(k+1)} - a| &= \delta^2 |v_1^{(k)} - a| = \cdots = \delta^{2(k+1)} |v_1^{(0)} - a| \\ |v_2^{(k+1)} - c| &= \delta^2 |v_1^{(k)} - c| = \cdots = \delta^{2(k+1)} |v_2^{(0)} - c| \\ |v_{12}^{(k)} - b| &= \left| \frac{b}{a} v_1^{(k)} - b \right| = \frac{|b|}{a} |v_1^{(k)} - a| = \frac{|b|}{a} \delta^{2k} |v_1^{(0)} - a| \end{aligned}$$

Supponiamo di aver estratto i campioni $\{x^{(1)}, \ldots, x^{(n)}\}$ usando l'algoritmo. Per verificare la convergenza del metodo, si possono considerare la media campionaria $\widehat{m} \in \mathbb{R}^2$ definita da

$$\widehat{m} := \frac{1}{n} \sum_{k=1}^{n} x^{(k)},$$

e la matrice delle covarianze empirica $\widehat{Q} \in \mathbb{R}^{2\times 2}$ definita da

$$(\widehat{Q})_{ij} := \frac{1}{n-1} \sum_{k=1}^{n} (x_i^{(k)} - \widehat{m}_i)(x_j^{(k)} - \widehat{m}_j).$$

Se i campioni sono distribuiti approssimativamente secondo la distribuzione normale $\mathcal{N}(m, Q)$, le quantità $||\widehat{m} - m||_2$ e $||\widehat{Q} - Q||_\infty$ tendono a zero al crescere di n (per la legge forte dei grandi numeri).

Le figure 3.1 e 3.2 mostrano la relazione tra δ e $||\widehat{m} - m||_2$ e $||\widehat{Q} - Q||_\infty$. Il metodo compie 3 iterazioni, partendo dal vettore nullo $x^{(0)} = (0, 0)$, ed estraendo 500 campioni da diverse catene di Markov. L'algoritmo è stato eseguito 500 volte per campionare le densità target $\mathcal{N}(m, Q)$, con m e Q scelti in modo casuale. In particolare, m è stato scelto con coefficienti presi da una distribuzione $\mathcal{N}(5, 1)$, e $Q = AA^T$, con A matrice con entrate scelte da una distribuzione normale standard[1] (in questo modo, A è invertibile quasi certamente e AA^T è definita positiva).

Dalle figure si deduce che al crescere di δ, più precisamente se $\delta > 0.3$ per la media, e $\delta > 0.6$ per la matrice delle covarianze, diventano necessarie più di 3 iterazioni per ottenere una buona approssimazione di m e Q.

[1] In MATLAB si può usare il comando $normrnd(\mu, \sigma, [n, m])$ che genera una matrice $n \times m$ con coefficienti presi dalla distribuzione $\mathcal{N}(\mu, \sigma^2)$.

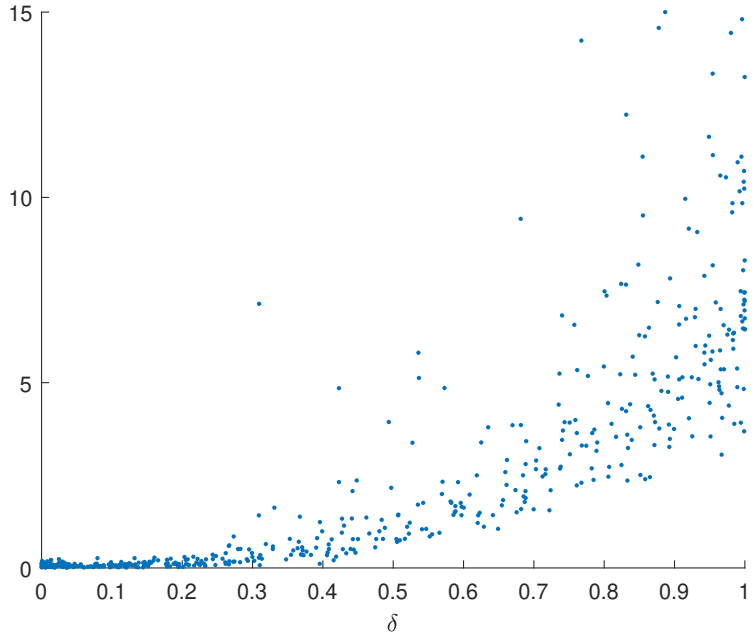


Figura 3.1: Relazione tra δ e $||\widehat{m} - m||_2$ in scala logaritmica

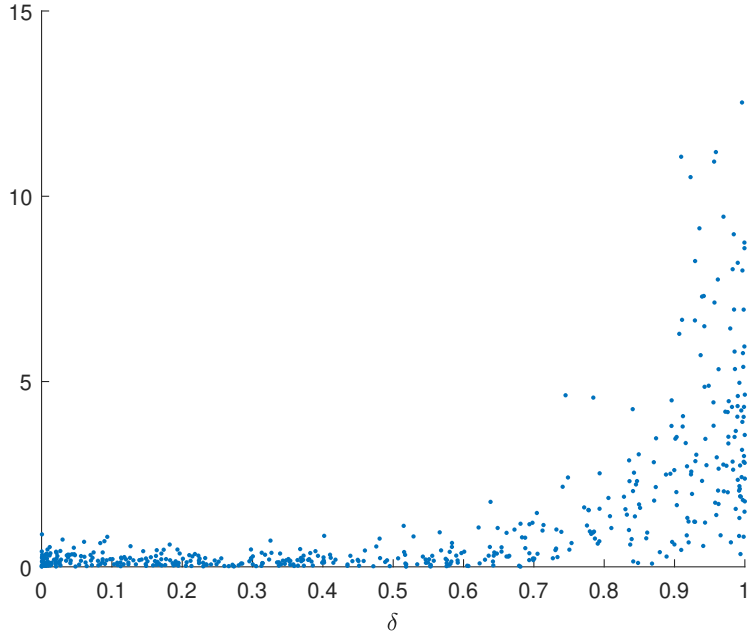


Figura 3.2: Relazione tra δ e $||\widehat{Q} - Q||_\infty$ in scala logaritmica

Infine, riportiamo dei grafici che mostrano l'andamento degli errori $||m^{(k)} - m||_2$ e $||Q^{(k)} - Q||_\infty$ al variare di k. Per approssimare $m^{(k)}$ e $Q^{(k)}$ sono state considerate la media campionaria e la matrice di covarianza empirica di 500 punti estratti con k iterazioni. m e Q sono casuali con coefficienti scelti dalla distribuzione $\mathcal{N}(5, 1)$.
Dal teorema 3.1, si ottiene $||m^{(k)} - m||_2 = \delta^k ||m^{(0)} - m||_2$ e $||Q^{(k)} - Q||_\infty = \delta^{2k} M$, dove $M = \max\{(1 + \frac{|b|}{a})|v_1^{(0)} - a|, |v_2^{(0)} - c| + \frac{|b|}{a}|v_1^{(0)} - a|\}$. In questo caso, dato che $x^{(0)}$ è sempre il vettore

nullo, $m^{(0)} = (0,0)$ e $v_1^{(0)} = v_2^{(0)} = 0$, quindi $M = ||Q||_\infty$. L'andamento di $||m^{(k)} - m||_2$ si può quindi confrontare con il grafico della funzione $f(k) = \delta^k ||m||_2$ (figura 3.3), e $||Q^{(k)} - Q||_\infty$ può essere confrontato con la funzione $g(k) = \delta^{2k} ||Q||_\infty$ (figura 3.4). Le piccole oscillazioni dei punti sono dovute all'errore generato dall'approssimazione di $m^{(k)}$ e $Q^{(k)}$ con la media e la covarianza empiriche.

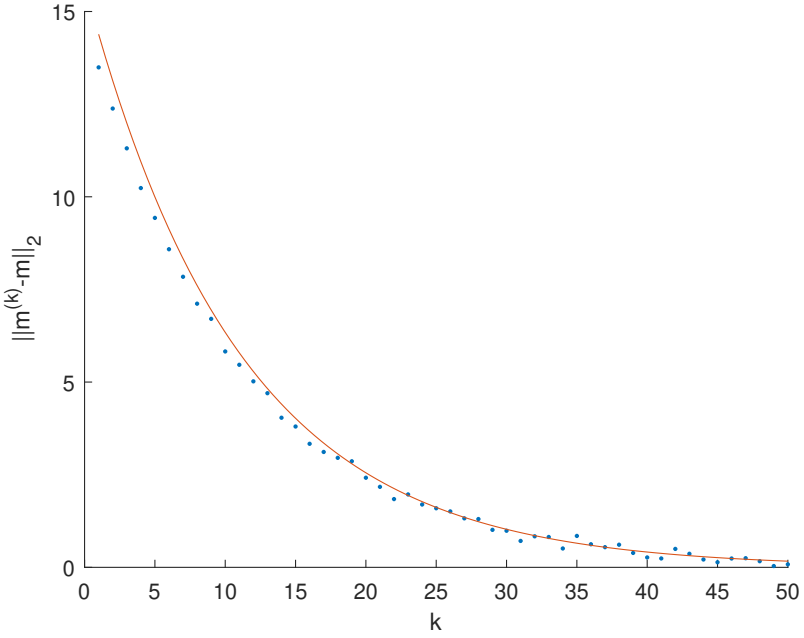


Figura 3.3: Andamento di $||m^{(k)} - m||_2$ in funzione di $k \in \{1, \dots, 50\}$ confrontato con la funzione $f(k) = \delta^k ||m||_2$.

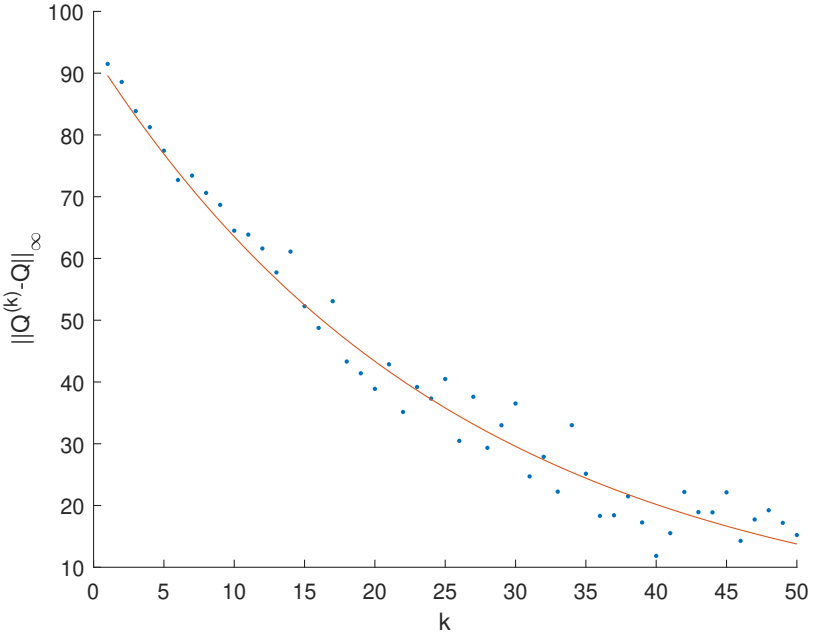


Figura 3.4: Andamento di $||Q^{(k)} - Q||_\infty$ in funzione di $k \in \{1, \dots, 50\}$ confrontato con la funzione $g(k) = \delta^{2k} ||Q||_\infty$.

3.2 Caso $d>2$

Sia $X \sim \mathcal{N}(m, Q)$ la variabile da campionare, con $m \in \mathbb{R}^d$, $Q = (Q_{ij}) \in \mathbb{R}^{d\times d}$ definita positiva. Supponiamo $d > 2$. Se $X^{(k)}$ è la variabile ottenuta alla fine della k-esima iterazione dell'algoritmo di Gibbs, si può dimostrare, come nella sezione precedente, che ha densità gaussiana. Per ogni $i \in \{1, \dots, d\}$ indichiamo con Q_i la sottomatrice di Q di taglia $d-1$ ottenuta eliminando l'i-esima riga e l'i-esima colonna, e con B_i l'i-esima riga di Q senza l'elemento Q_{ii}. Consideriamo

$$\delta_i := \frac{1}{Q_{ii}} B_i Q_i^{-1} B_i^T$$

e

$$\delta := \max_i \delta_i.$$

Osservazione 3.2. Poichè $0 < \delta_i < 1$ per ogni $i \in \{1, \dots, d\}$, si ha $0 < \delta < 1$. Infatti, sia $v_i = Q_{ii} - B_i Q_i^{-1} B_i^T$ il complemento di Schur di Q_i in Q. Per ipotesi Q è definita positiva, quindi dall'osservazione 3.1 si ottiene $v_i > 0$, ovvero $\delta_i < 1$. Inoltre, essendo Q_i definita positiva e quindi anche Q_i^{-1}, si ha $B_i Q_i^{-1} B_i^T > 0$, da cui segue $\delta_i > 0$.

Ripetiamo le simulazioni fatte nel caso bidimensionale, per verificare se esiste una correlazione tra la velocità di convergenza del metodo e δ. Applichiamo il metodo di Gibbs considerando le densità condizionali unidimensionali.
Come nella sezione precedente indichiamo con $\widehat{m}$ la media campionaria e $\widehat{Q}$ la matrice delle covarianze empirica. Le seguenti figure mostrano la relazione tra δ e $||\widehat{m} - m||_2$ e $||\widehat{Q} - Q||_\infty$. Il metodo parte dal vettore nullo, ed estrae 500 campioni da diverse catene di Markov. L'algoritmo è stato eseguito 500 volte per campionare le densità target $\mathcal{N}(m, Q)$, scegliendo in modo casuale m e Q.
Nelle figure 3.5 e 3.6 si ha $d = 3$ e l'algoritmo compie 5 iterazioni. Nelle figure 3.7 e 3.8 si ha $d = 5$ con 10 iterazioni del metodo.
Si può osservare che in entrambi i casi al crescere di δ crescono anche gli errori.

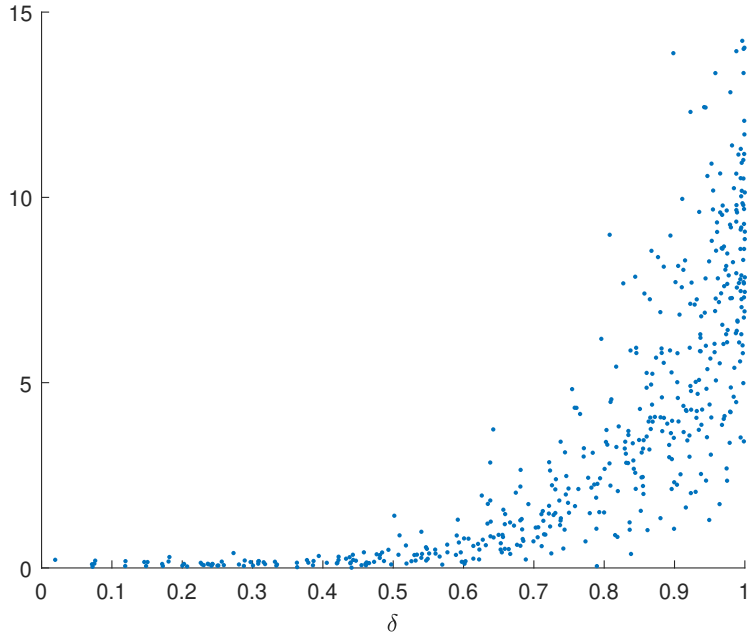


Figura 3.5: Relazione tra δ e $||\widehat{m} - m||_2$ in scala logaritmica, con $d = 3$.

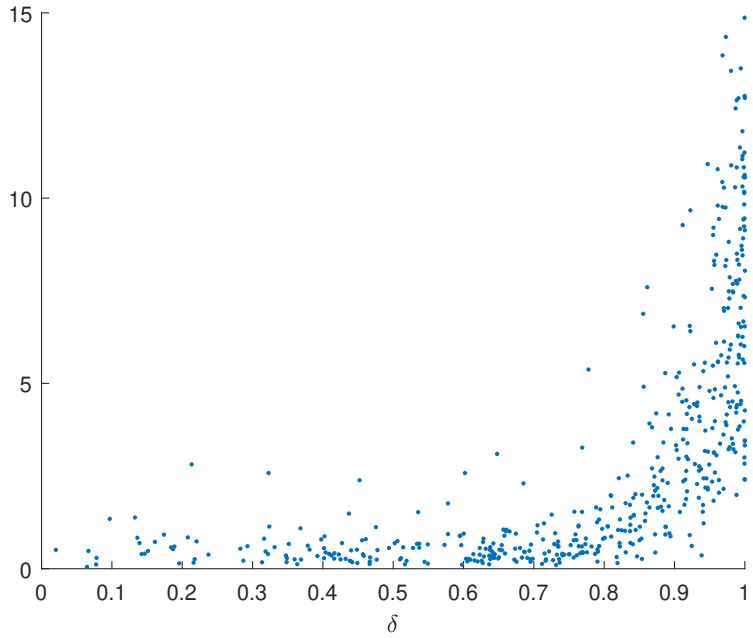


Figura 3.6: Relazione tra δ e $||\widehat{Q} - Q||_\infty$ in scala logaritmica, con $d = 3$.

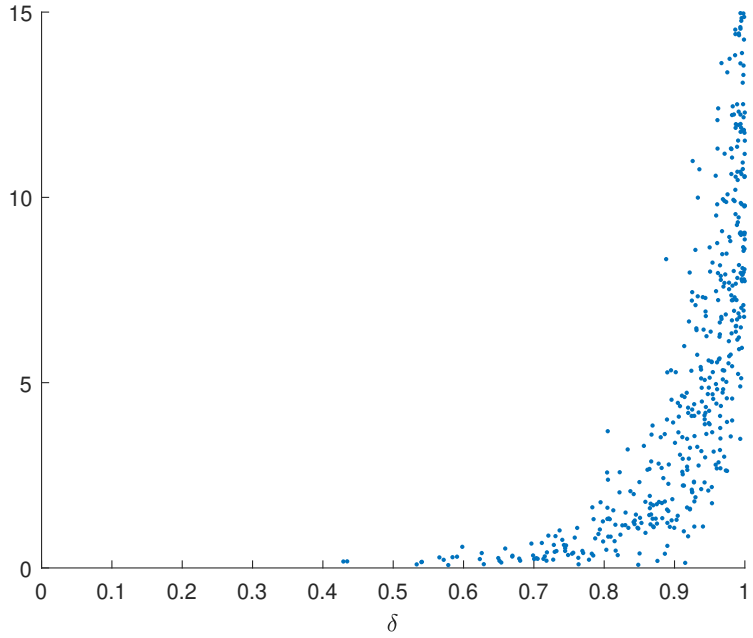


Figura 3.7: Relazione tra δ e $||\widehat{m} - m||_2$ in scala logaritmica, con $d = 5$.

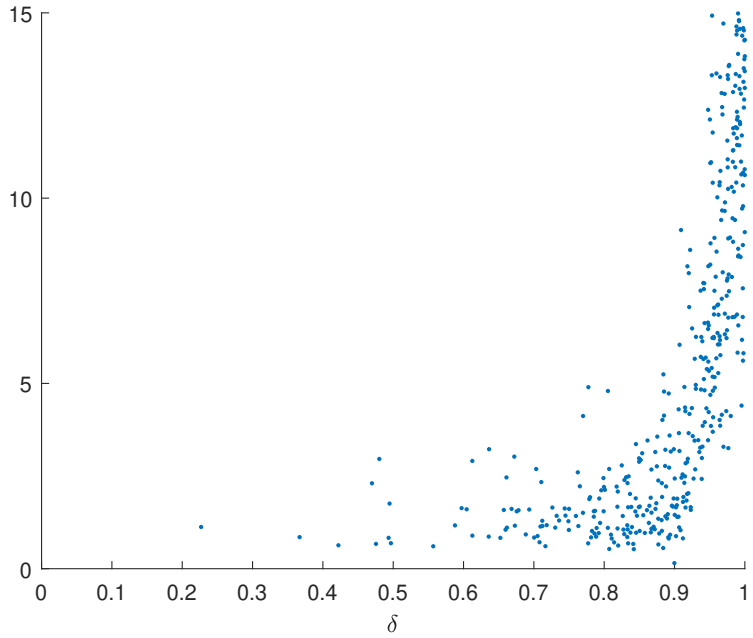


Figura 3.8: Relazione tra δ e $||\widehat{Q} - Q||_\infty$ in scala logaritmica, con $d = 5$.

Infine, confrontiamo l'errore di approssimazione della media al variare del numero di iterazioni con la funzione $f(k) = \delta^k ||m^{(0)} - m||_2$, e l'errore di approssimazione della matrice delle covarianze con la funzione $g(k) = \delta^{2k} ||Q^{(0)} - Q||_\infty$. Come in precedenza, l'algoritmo genera 500 campioni, partendo da $x^{(0)} = 0 \in \mathbb{R}^d$, quindi si può supporre $m^{(0)} = 0 \in \mathbb{R}^d$ e $Q^{(0)} = 0 \in \mathbb{R}^{d \times d}$.
Dalle figure 3.10 e 3.11 sembra che l'errore converga a zero più velocemente di $f(k) = \delta^k ||m||_2$. Considerando anche l'errore dovuto all'approssimazione di $Q^{(k)}$, le figure 3.12, 3.13 e 3.14 sembrano suggerire che l'errore sulle covarianze segua l'andamento di $g(k)$. Lo studio teorico di questo andamento è lasciato aperto ad indagini future.

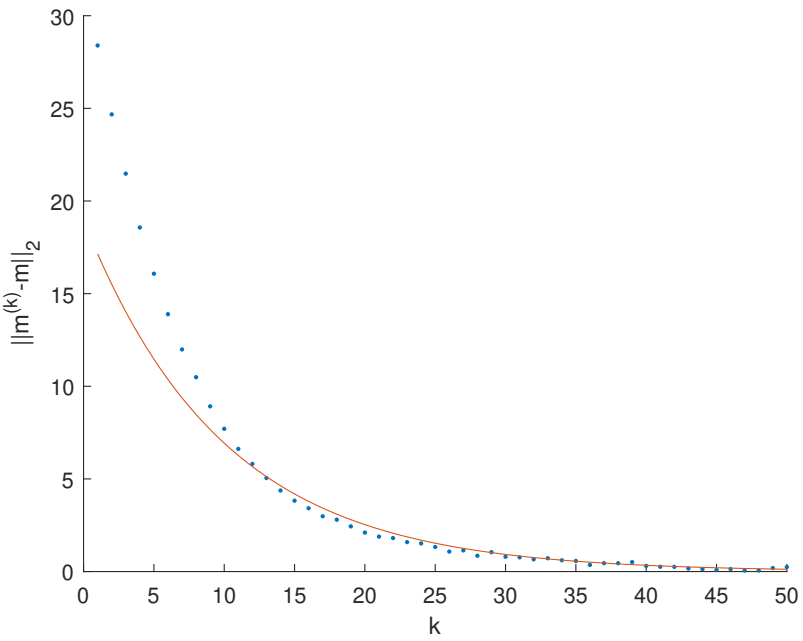


Figura 3.9: Andamento di $||m^{(k)} - m||_2$ in funzione di $k \in \{1, \dots, 50\}$ confrontato con la funzione $f(k) = \delta^k ||m||_2$, con $d = 3$.

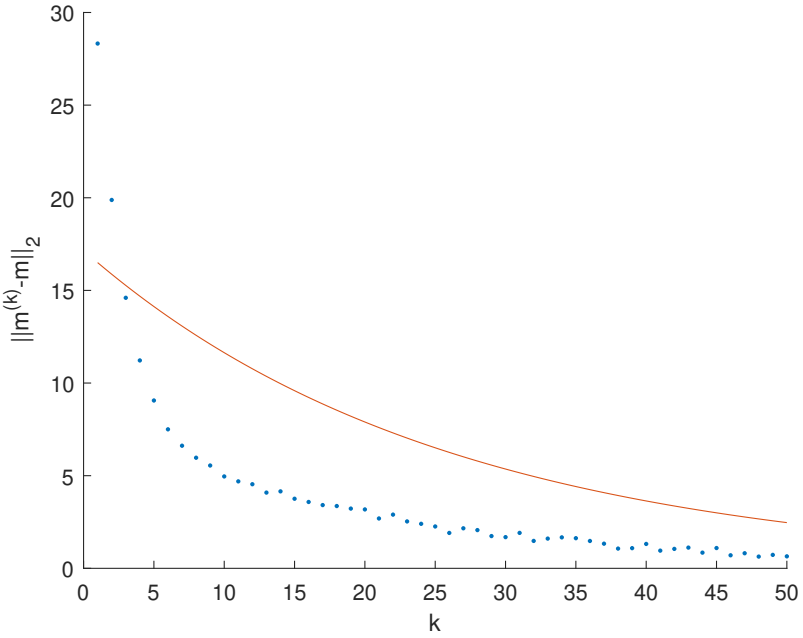


Figura 3.10: Andamento di $||m^{(k)} - m||_2$ in funzione di $k \in \{1, \dots, 50\}$ confrontato con la funzione $f(k) = \delta^k ||m||_2$, con $d = 3$.

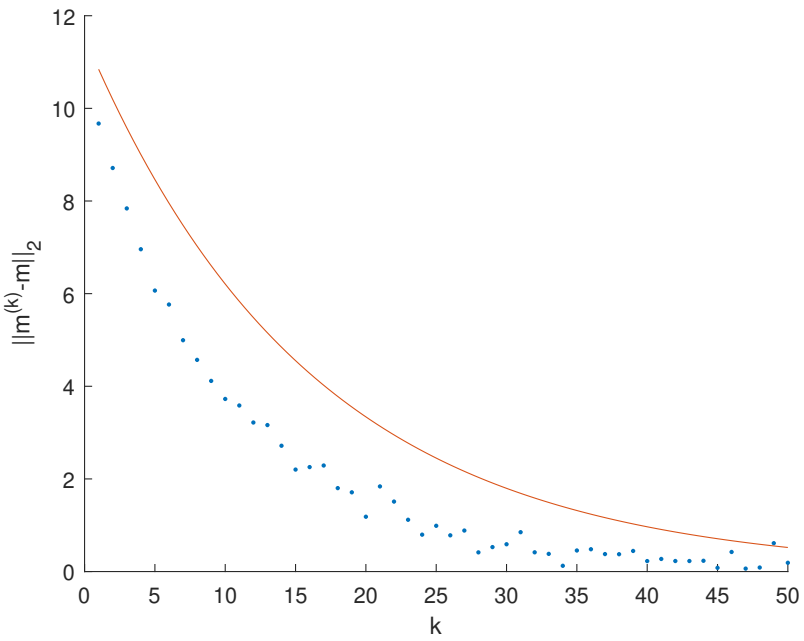


Figura 3.11: Andamento di $||m^{(k)} - m||_2$ in funzione di $k \in \{1, \ldots, 50\}$ confrontato con la funzione $f(k) = \delta^k ||m||_2$, con $d = 5$.

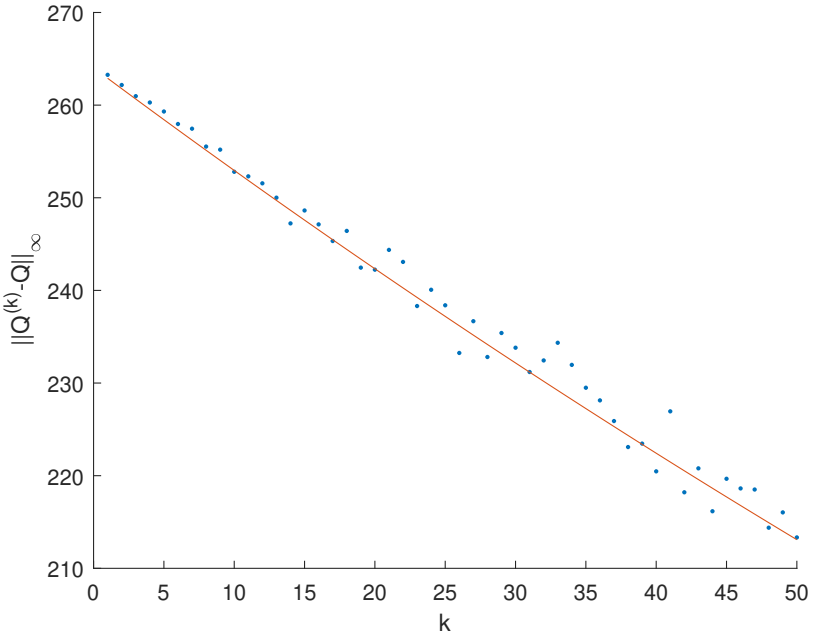


Figura 3.12: Andamento di $||Q^{(k)} - Q||_\infty$ in funzione di $k \in \{1, \ldots, 50\}$ confrontato con la funzione $g(k) = \delta^{2k} ||Q||_\infty$, con $d = 3$.

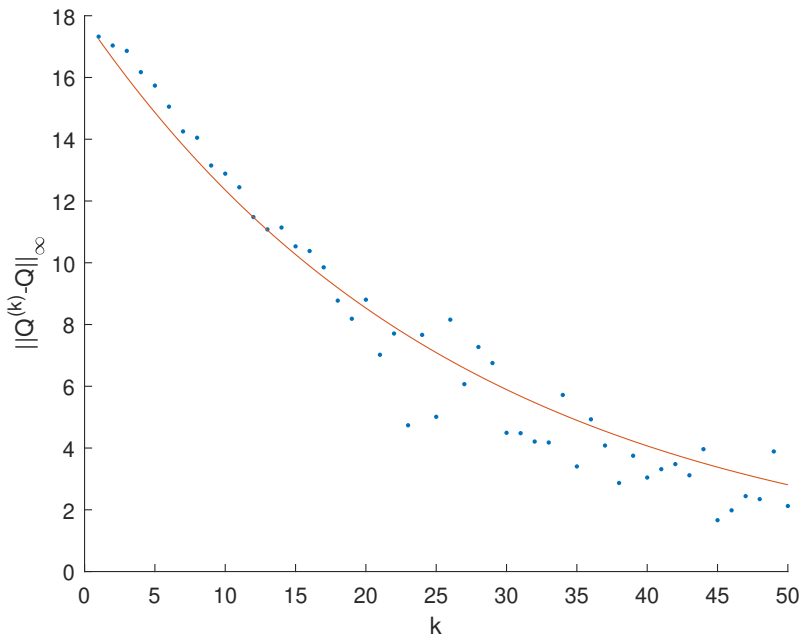


Figura 3.13: Andamento di $||Q^{(k)} - Q||_\infty$ in funzione di $k \in \{1, \dots, 50\}$ confrontato con la funzione $g(k) = \delta^{2k}||Q||_\infty$, con $d = 3$.

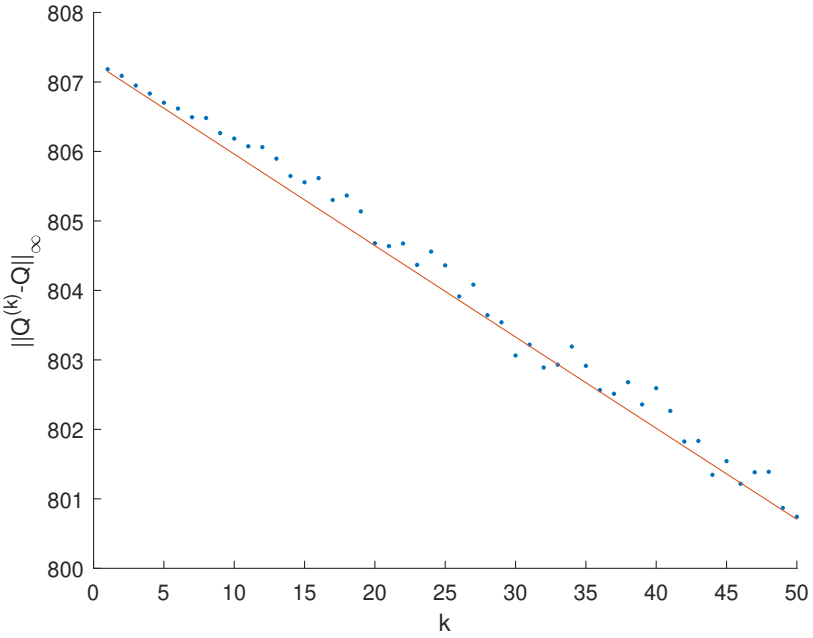


Figura 3.14: Andamento di $||Q^{(k)} - Q||_\infty$ in funzione di $k \in \{1, \dots, 50\}$ confrontato con la funzione $g(k) = \delta^{2k}||Q||_\infty$, con $d = 5$.

Bibliografia

[1] Alan E Gelfand. Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304, 2000.

[2] G. Modica and L. Poggiolini. *A First Course in Probability and Markov Chains*. Wiley, 2012.

[3] Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.