Algoritmi Markov chain Monte Carlo (MCMC)

Vers. 1.1.1

Gianluca Mastrantonio

email: gianluca.mastrantonio@polito.it

Markov chain Monte Carlo

Nei casi più complessi, è difficile lavorare con la a posteriori

- Multidimensionale
- Costante di normalizzazione generalmente non disponibile
- ....

e difficilmente è possibile trovare la distribuzione a posteriori in forma chiusa.

Abbiamo che

- **Monte Carlo**: possiamo studiare le caratteristiche di una distribuzione attraverso campioni dalla distribuzione di interesse
- **Markov Chain**: Data una distribuzione π, si può definire una Markov Chain (sotto opportune condizioni) che abbia π come distribuzione stazionaria

L'idea è di construire in Markov chain che abbia come distribuzione stazionaria la a-posteriori di interesse, quindi, indipendentemente da dove iniziamo l'algoritmo, prima o poi arriviamo alla distribuzione limite che è la stazionaria

Markov chain Monte Carlo

Per capire come fare, introdurre il concetto di **full-conditional**

Full conditional

Dato un vettore di variabili aleatorie $\boldsymbol{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, con congiunta $f(\mathbf{X}_1, \ldots, \mathbf{X_p})$, la full conditional di $\mathbf{X}_j$ è la distribuzione di

$$\mathbf{X}_j | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1} \ldots, \mathbf{X}_p$$

che è quindi la distribuzione di una variabile, dato tutte le altre. Per semplicità spesso si scrive

$$\mathbf{X}_j | \ldots$$

per indicare la full conditional di $\mathbf{X}_j$ e con

$$f(\mathbf{x}_j | \ldots)$$

la sua densità

Quello che ci servirà sono le full conditional della a posteriori, e queste sono connesse ai DAG come vedremo poi

Gibbs Sampler

Mettiamoci in un caso particolare in cui siamo interessati alla distribuzione (che può essere una posteriori o no)

$$f(\boldsymbol{\theta}|\mathbf{y})$$

con $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$ e immaginiamo di voler ottenere campioni dalla distributione

L'idea dell'algoritmo è di partizionare il set $\boldsymbol{\theta}$ in p subset $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$

- Possono essere anche multivariati
- $\boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2 \cup \ldots \boldsymbol{\theta}_p = \boldsymbol{\theta}$
- generalmente di richiede che $\boldsymbol{\theta}_i \cap \boldsymbol{\theta}_{i'} = \emptyset$, ma fintanto che $\boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2 \cup \ldots \boldsymbol{\theta}_p = \boldsymbol{\theta}$ potete anche avere intersezione non vuota

e campionare una variabile alla volta tramite una probabilità di transizione. In questo caso, il **tempo** delle catena di Markov è in realtà l'iterazione e verra indicata con un indice, generalmente b, in alto.

Gibbs Sampler

Definiamo

- $\boldsymbol{\theta}^b$ campioni b-esimi
- $\boldsymbol{\theta}_{1:h} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_h)$: il vettore delle prima h variabili
- $\boldsymbol{\theta}_{(h+1):p} = (\boldsymbol{\theta}_{h+1}, \ldots, \boldsymbol{\theta}_p)$: le restanti variabili
- $\boldsymbol{\theta}_{1:0} = \boldsymbol{\theta}_{(p+1):p} = \emptyset$
- transizione $T_h((\boldsymbol{\theta}_{1:(h-1)}, \boldsymbol{\theta}_{h:p}), (\boldsymbol{\theta}_{1:h-1}, \boldsymbol{\theta}^*_h, \boldsymbol{\theta}_{(h+1):p}))$: una transizione che utilizziamo per campionare solamente il valore del' $h-$esima componente.

Gibbs Sampler

Partendo da $\boldsymbol{\theta}^{b-1}$, l'algoritmo prevede di

- simuliamo $\boldsymbol{\theta}_1^b$ utilizzando $T_1(.,.)$
- simuliamo $\boldsymbol{\theta}_2^b$ utilizzando $T_2(.,.)$
- ...
- simuliamo $\boldsymbol{\theta}_p^b$ utilizzando $T_p(.,.)$

ad ogni passo, le variabili che entrano nei calcoli/simulazioni devono sempre essere le ultime generate.
Possiamo facilmente verificare che la a posteriori è la stazionaria. Per farlo assumiamo di avere solo due variabili. Allora, dobbiamo dimostrare che

$$f(\theta_1^{b+1}, \theta_2^{b+1}|\mathbf{y}) = \int_{\Theta_1}\int_{\Theta_2} f(\theta_1^b, \theta_2^b|\mathbf{y}) f(\theta_1^{b+1}|\theta_2^b, \mathbf{y}) f(\theta_2^{b+1}|\theta_1^{b+1}, \mathbf{y}) d\lambda(\theta_2^b) d\lambda(\theta_1^b) =$$

Calcoliamo l'integrale

$$\int_{\Theta_1}\int_{\Theta_2} f(\theta_1^b, \theta_2^b|\mathbf{y}) f(\theta_1^{b+1}|\theta_2^b, \mathbf{y}) f(\theta_2^{b+1}|\theta_1^{b+1}, \mathbf{y}) d\lambda(\theta_2^b) d\lambda(\theta_1^b) =$$

Gibbs Sampler

$$f(\theta_2^{b+1}|\theta_1^{b+1},\mathbf{y})\int_{\Theta_2} f(\theta_1^{b+1}|\theta_2^{b},\mathbf{y})\left(\int_{\Theta_1} f(\theta_1^{b},\theta_2^{b}|\mathbf{y})d\lambda(\theta_1^{b})\right)d\lambda(\theta_2^{b}) =$$

$$f(\theta_2^{b+1}|\theta_1^{b+1},\mathbf{y})\int_{\Theta_2} f(\theta_1^{b+1}|\theta_2^{b},\mathbf{y})f(\theta_2^{b}|\mathbf{y})d\lambda(\theta_2^{b}) =$$

$$f(\theta_2^{b+1}|\theta_1^{b+1},\mathbf{y})\int_{\Theta_2} f(\theta_1^{b+1},\theta_2^{b}|\mathbf{y})d\lambda(\theta_2^{b}) = f(\theta_2^{b+1}|\theta_1^{b+1},\mathbf{y})f(\theta_1^{b+1}|\mathbf{y}) =$$

$$f(\theta_2^{b+1},\theta_1^{b+1}|\mathbf{y})$$

CVD

Gibbs Sampler

Algoritmo Gibbs sampler

Vogliamo ottenere campioni da $f(\boldsymbol{\theta}|\mathbf{y})$

Inizializzazione: imponi $b = 0$ e inizializza i valori di $\boldsymbol{\theta}_1^0, \ldots, \boldsymbol{\theta}_p^0$ con dei valori "possibili".

1: **repeat**
2: incrementa b di 1
3: campiona $\boldsymbol{\theta}_1^b$ da $\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{b-1}, \boldsymbol{\theta}_3^{b-1}, \ldots, \boldsymbol{\theta}_{p-1}^{b-1}, \boldsymbol{\theta}_p^{b-1}, \mathbf{y}$
4: campiona $\boldsymbol{\theta}_2^b$ da $\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{b}, \boldsymbol{\theta}_3^{b-1}, \ldots, \boldsymbol{\theta}_{p-1}^{b-1}, \boldsymbol{\theta}_p^{b-1}, \mathbf{y}$
5: campiona $\boldsymbol{\theta}_3^b$ da $\boldsymbol{\theta}_3|\boldsymbol{\theta}_1^{b}, \boldsymbol{\theta}_2^{b}, \ldots, \boldsymbol{\theta}_{p-1}^{b-1}, \boldsymbol{\theta}_p^{b-1}, \mathbf{y}$
6: $\ldots$
7: campiona $\boldsymbol{\theta}_p^b$ da $\boldsymbol{\theta}_p|\boldsymbol{\theta}_1^{b}, \boldsymbol{\theta}_2^{b}, \ldots, \boldsymbol{\theta}_{p-2}^{b}, \boldsymbol{\theta}_{p-1}^{b}, \mathbf{y}$
8: definisci $\boldsymbol{\theta}_p^b = (\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b \ldots, \boldsymbol{\theta}_p^b)^T$
9: **until** $b < B$
10: $\boldsymbol{\theta}_p^b$ sono campioni dipendenti da $f(\boldsymbol{\theta}|\mathbf{y})$.

Abbiamo adesso il problema di come si determinano le full conditional. Per capirlo vediamo la relazione che ha il modello Bayesiano con i DAG. Immaginiamo di prendere i dati di temperatura massima di una locazioni spaziale dei dati ecologici. Indichiamo queste misurazioni con y_t, con $t = 1, \ldots, T$, e di avere il seguente gerarchico

$$y_t = \beta_0 + \beta_1 t + w_t + \epsilon_t$$
$$\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$
$$w_t | w_{t-1} \sim N(\alpha w_{t-1}, \tau^2), t = 2, \ldots, T$$

e che la stazione al tempo 2 abbia avuto un problema e non ci sia la registrazione di y_2. Fate attenzione che la prima riga è equivalente a scrivere

$$y_t | w_t \sim N(\beta_0 + \beta_1 t + w_t, \sigma^2)$$

La trattazione dei dati mancanti è uno dei vantaggi dell'approccio Bayesiano, visto che si possono trattare in una maniera molto naturale, cos a non possibile nel frequentista (a

meno di fare Expectation Maximization). I parametri del modello sono tutte le cose che non conosciamo. In questo caso quindi anche il dato mancante è un parametro

In questo caso i parametri (tutti ciò che non conosciamo) sono $(\beta_0, \beta_1, \alpha, \sigma^2, \tau^2, w_1, y_2)$ e la a posteriori di interesse è

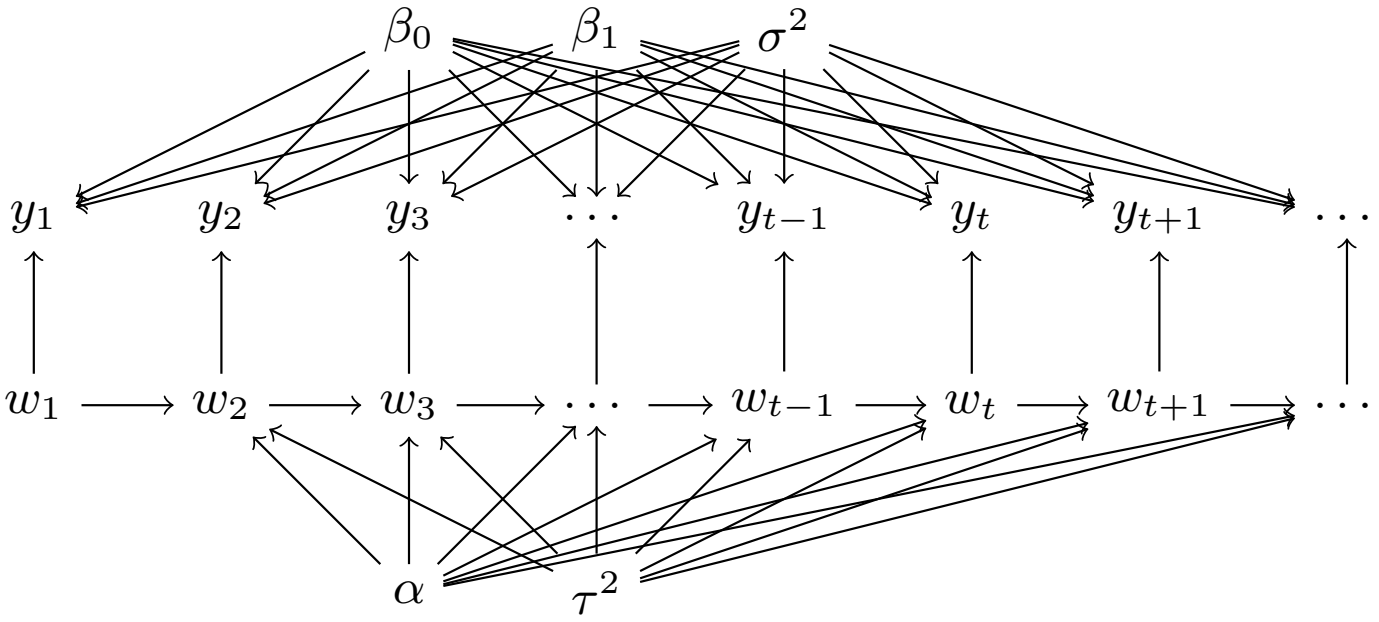
$$f(\beta_0, \beta_1, \alpha, \sigma^2, \tau^2, \mathbf{w}, y_2|\mathbf{y}_{-2}) =$$

$$\frac{f(\mathbf{y}|\beta_0, \beta_1, \sigma^2, \mathbf{w}) f(\mathbf{w}_{-1}|w_1, \tau^2, \alpha) f(w_1, \tau^2, \alpha, \beta_0, \beta_1, \sigma^2)}{f(\mathbf{y}_{-2})}$$

e assumiamo che $f(w_1, \tau^2, \alpha, \beta_0, \beta_1, \sigma^2) = f(w_1) f(\tau^2) f(\alpha) f(\beta_0) f(\beta_1) f(\sigma^2)$.
Attenzione! non potete decidere la distribuzione su y_2 perché è già definita dal modello.

Possiamo rappresentare le relazioni tra le variabili aleatorie (osservate, latenti e parametri) tramite il DAG

Modelli grafici, Variabili Latenti e Missing data



Immaginiamo di voler trovare la full conditional di w_2. Abbiamo che

$$f(w_2|...) = \frac{f(\beta_0, \beta_1, \alpha, \sigma^2, \tau^2, \mathbf{w}, \mathbf{y})}{f(\beta_0, \beta_1, \alpha, \sigma^2, \tau^2, \mathbf{w}_{-2}, \mathbf{y})} \propto f(\beta_0, \beta_1, \alpha, \sigma^2, \tau^2, \mathbf{w}, \mathbf{y})$$

e $f(\beta_0, \beta_1, \alpha, \sigma^2, \tau^2, \mathbf{w}, \mathbf{y})$ è il numeratore della a posteriori e la congiunta rappresentata dal DAG. In questo caso abbiamo quindi che

$$f(w_2|...) = f(w_1)f(\tau^2)f(\alpha)f(\beta_0)f(\beta_1)f(\sigma^2)f(y_2|w_2, \beta_0, \beta_1, \sigma^2)\times$$

$$\prod_{t=2}^{T} f(w_t|w_{t-1}, \tau^2, \alpha)f(y_t|w_t, \beta_0, \beta_1, \sigma^2) \propto$$

$$f(w_2|w_1, \tau^2, \alpha)f(w_3|w_2, \tau^2, \alpha)f(y_2|w_2, \beta_0, \beta_1, \sigma^2)$$

Quindi la full conditional di una variabile dipende solo dalla sua distribuzione e da quella di tutti i nodi a cui punta nel DAG.
Un altro modo di vederlo è che la distribuzione di un elemento del DAG dipenda da tutti i nodi entranti e uscenti da quale parametro (l'insieme di questi nodi è spesso chiamato **Markov Blanket**)

Proviamo a calcolare la full conditional utilizzando il trick del kernel. Abbiamo che tutto ciò che dipende da w_2 è

$$
\begin{aligned}
f(w_2|\dots) \propto & \exp\left(-\frac{1}{2}\left(\frac{w_2^2 - 2w_2\alpha w_1}{\tau^2}\right)\right) \times \\
& \exp\left(-\frac{1}{2}\left(\frac{w_2^2\alpha^2 - 2w_2\alpha w_3}{\tau^2}\right)\right) \times \\
& \exp\left(-\frac{1}{2}\left(\frac{w_2^2 - 2w_2(y_2 - \beta_0 - 2\beta_1)}{\sigma^2}\right)\right) \times = \\
& \exp\left(-\frac{1}{2}\left(w_2^2\left(\frac{1+\alpha^2}{\tau^2} + \frac{1}{\sigma^2}\right) - 2w_2\left(\frac{\alpha w_1 + \alpha w_3}{\tau^2} + \frac{y_2 - \beta_0 - 2\beta_1}{\sigma^2}\right)\right)\right)
\end{aligned}
$$

e se definisco

$$
v = \left(\frac{1+\alpha^2}{\tau^2} + \frac{1}{\sigma^2}\right)^{-1} \qquad m = v\left(\frac{\alpha w_1 + \alpha w_3}{\tau^2} + \frac{y_2 - \beta_0 - 2\beta_1}{\sigma^2}\right)
$$

posso vedere che

$$\exp\left(-\frac{1}{2}\left(\frac{w_2^2}{v}-2\frac{w_2 m}{v}\right)\right)\propto\exp\left(-\frac{(w_2-m)^2}{2v}\right)$$

e quindi

$$w_2|\cdots\sim N(m,v)$$

Se volesismo trovare la full conditional di y_2, che essendo missing va campionata, è facile vedere che la sua full conditional dipende solo da

$$f(y_2|w_2,\beta_0,\beta_1,\sigma^2)$$

e quindi

$$y_2|\cdots\sim N(\beta_0+2\beta_1+w_2,\sigma^2)$$

Gibbs Sampler

Vedremo tra poco cosa fare quando non siamo in grado di trovare le full conditional, o campionare, e concentriamoci sulla correlazione tra i valori e la dipendenza dall'inizializzazione.
Ipotizziamo di aver usato il Gibbs sampler e di aver ottenuto campioni $\boldsymbol{\theta}^b = (\boldsymbol{\theta}_1^b, \ldots, \boldsymbol{\theta}_p^b)$, con $b = 1, \ldots, B$, da

$$f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p | \mathbf{y})$$

visto che siamo, in genere, interessati a statistiche della distribuzione

$$f(\boldsymbol{\theta}_p | \mathbf{y})$$

ci possiamo focalizare sui campioni $(\boldsymbol{\theta}_p^1, \ldots, \boldsymbol{\theta}_p^B)$, che può essere vista come una serie temporale. Possiamo allora plottarla e vedere il suo comportamento, come nell'esempio seguente:

Gibbs Sampler

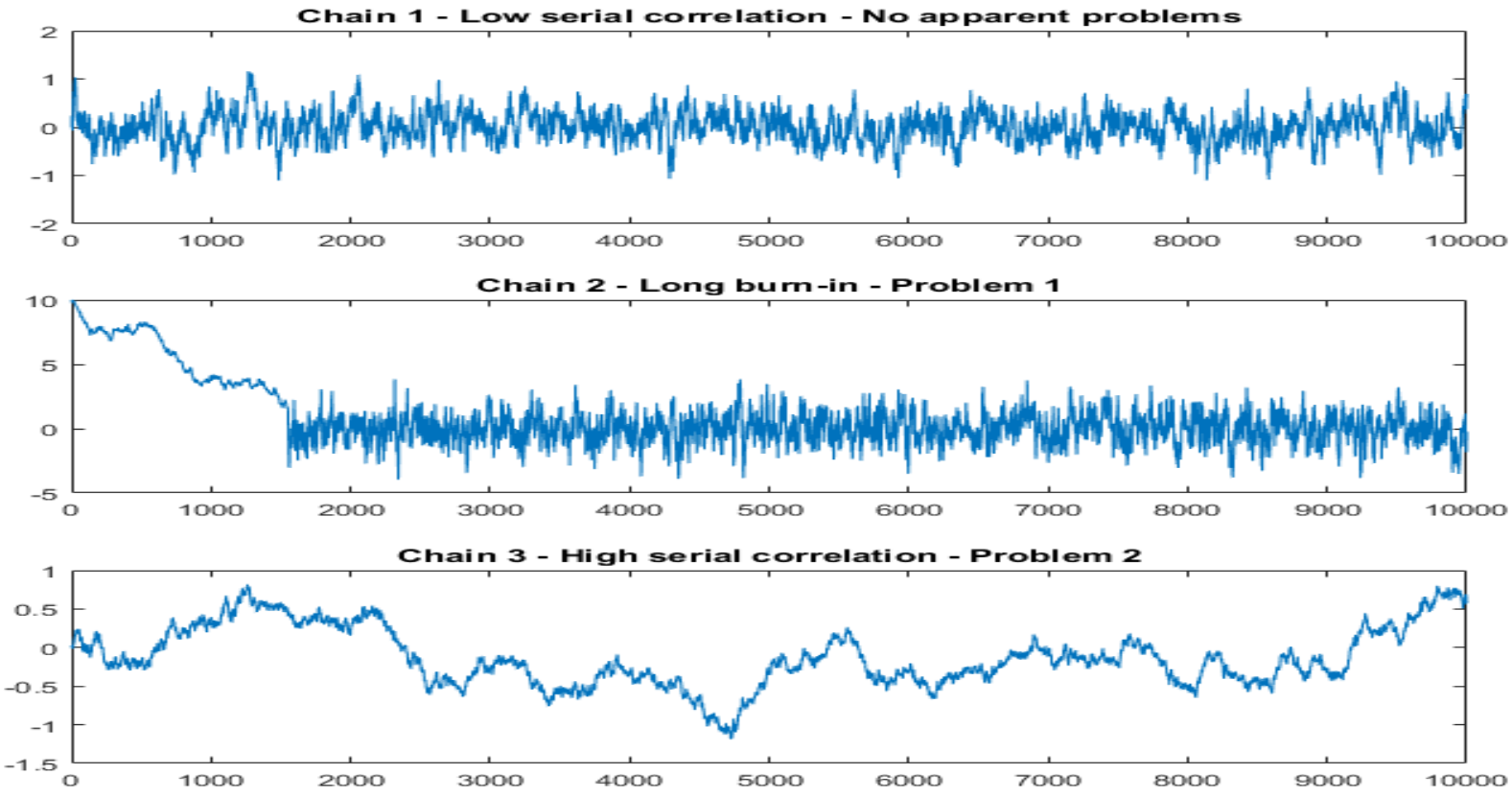


Figure: Esempi di catene - presa da https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo-diagnostics

Questi plot si chiamano **traceplot**

Gibbs Sampler

- La prima catena è una catena che è arrivata a convergenza, visto che i campioni sembrano indipendenti e stazionari, e sembrerebbero essere campioni da una a posteriori con media zero e varianza bassa
- Nella seconda catena, per i primi 1500 campioni, si vede un forte effetto dell'inizializzazione $=10$, e ci sono voluti paracchi campioni per arrivare a convergenza (per assomigliare alla prima)
- Nella terza catena i campioni sembrano essere arrivati a convergenza, visto che ruotano intorno alla media 0 e anche il range di variabilità assomiglia alla prima catena, ma c'è una forte correlazione tra i campioni

Gibbs Sampler

Per risolvere i problemi della catena 2 e 3 si usa il burnin e il thin

- **burnin** il numero di campioni all'inizio della catena che vengono buttati. Nella catena 2, se facessimo un burnin di 2000, avremmo che i rimanenti campioni assomigliano alla catena 1
- **thin** prendere un campione ogni thin. Per esempio, con thin=2, invece di $(\boldsymbol{\theta}_j^1, \boldsymbol{\theta}_j^2, \boldsymbol{\theta}_j^3, \boldsymbol{\theta}_j^4, \ldots, \boldsymbol{\theta}_p^B)$, utilizzeremo $(\boldsymbol{\theta}_j^1, \boldsymbol{\theta}_j^3, \boldsymbol{\theta}_j^5, \boldsymbol{\theta}_j^7, \ldots, \boldsymbol{\theta}_j^B)$. Visto che la dipendenza tra i campioni dipende dalla distanza (se vediamo i campioni come serie temporale), se aumento la distanza diminuisco la correlazione/dipendenza

Un modo per valutare la correlazione tra i campioni è l'autocorrelation function (ACF) al log h:

$$r(h) = \frac{\sum_{b=1}^{N-h}(\boldsymbol{\theta}_j^b - \bar{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_{j+h}^b - \bar{\boldsymbol{\theta}}_j)}{\sum_{b=1}^{N}(\boldsymbol{\theta}_j^b - \bar{\boldsymbol{\theta}}_j)^2}$$

che valuta la correlazione tra i campioni "originali", e i campioni shiftati di un tempo/lag h. Notate che più aumenta h e meno sono gli elementi che si usano per calcolare il numeratore di $r(h)$.

Gibbs Sampler

Se l'ACF è uguale a 1 al lag 0 e 0 per tutti gli altri lag, allora i campioni sono indipendenti. Se l'ACF si annulla per lag $\geq c$, allora basta fare un thin di c per avere campioni indipendenti.

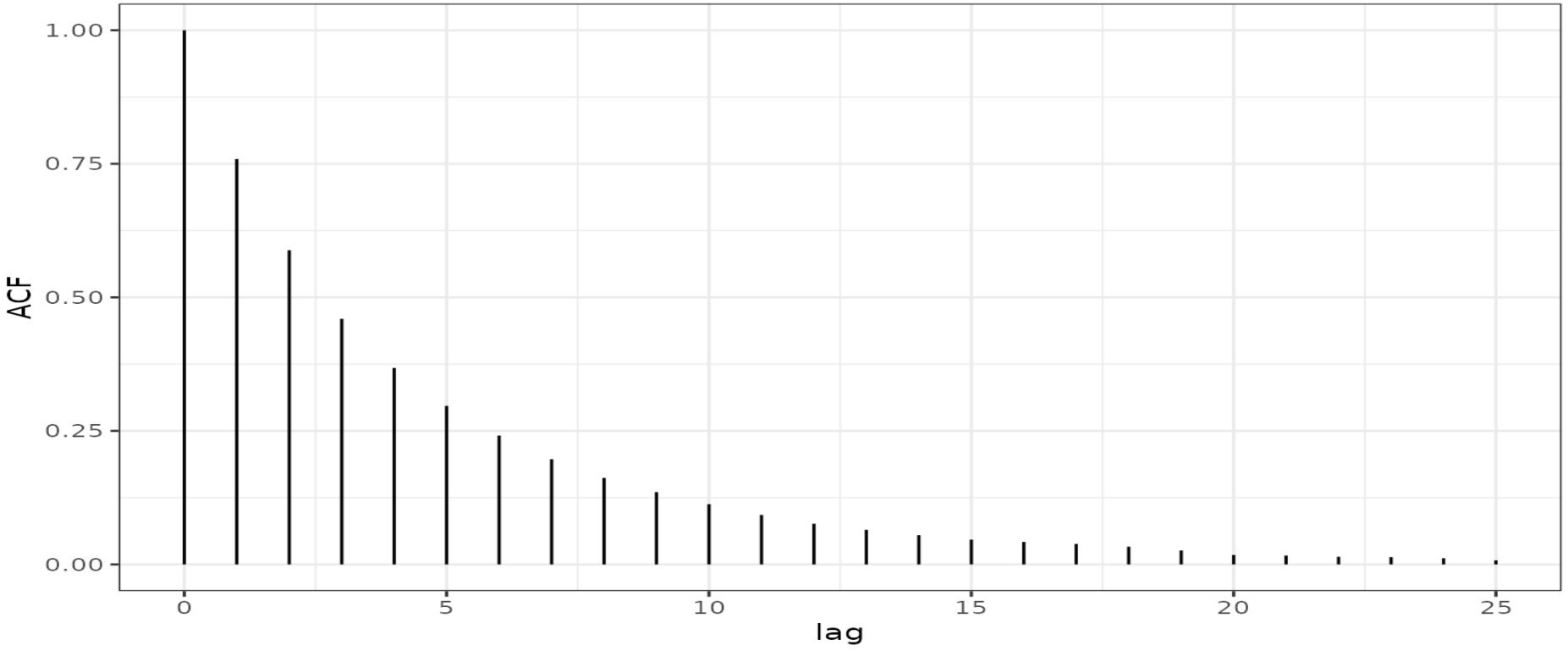


Figure: Esempio di Acf - presa da https://michael-stevens-27.github.io/silverblaze/reference/plot_acf.html

Metropolis-Hastings

Spesso, non siamo in grado di trovare un forma chiusa la forma della full conditional: per esempio, in una regressione potete vedere che se $\sigma^2 \sim G(a, b)$, la full conditional non ha una forma nota, mentre se $\sigma^2 \sim IG(a, b)$ la full conditional è ancora $IG()$.
Continuiamo a lavorare nel setting del Gibbs Sampler, ma all'iterazione b, step h, non campioniamo direttamente dalla full conditional, ma usiamo un algoritmo accept-reject per ottenere un campione. L'idea è

- decidiamo una distribuzione $Q()$ che chiamiamo proposta, con densità $q()$, che può dipendere dallo stato corrente della catena (quindi da $(\boldsymbol{\theta}^b_{1:(h-1)}, \boldsymbol{\theta}^{b-1}_{h:p}, \mathbf{y})$ se continuiamo con l'esempio precedente)
- proponiamo un nuovo valore per la catena, indicato con $\boldsymbol{\theta}^*_h$
- con un meccanismo probabilistico, decidiamo se il nuovo valore della catena per il parametro $h-$esimo debba essere $\boldsymbol{\theta}^*_h$ o teniamo il vecchio valore.

Metropolis-Hastings

La distribuzione proposal è molto importante. Generalmente richiede che la densità $q()$ sia maggiore di 0 dove la full conditional è maggiore di 0, ma, basta che

- utilizzando la proposta $Q()$, con un numero finito di passi, si possa raggiungere tutti i punti dove la densità $q()$ è maggiore di 0
- in generale si definisce la proposta come dipendente solo dal parametro h all'interazione precedente: $q(.|\boldsymbol{\theta}_h^{b-1})$
- si preferisce lavorare con variabili in $\mathbb{R}$ e come proposta si usa una $N(\boldsymbol{\theta}_h^{b-1}, \eta^2)$ (se la variabile originale non è su $\mathbb{R}$ bisogna trasformare le variabili, e calcolare la nuova a posteriori)

Metropolis-Hastings

Metropolis-Hasting algorithm

Scopo: Ottenere un campione da $\boldsymbol{\theta}_h^b$ da $\boldsymbol{\theta}_h|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b, \ldots, \boldsymbol{\theta}_{h-1}^b, \boldsymbol{\theta}_{h+1}^b, \ldots, \boldsymbol{\theta}_{p-1}^{b-1}, \boldsymbol{\theta}_p^{b-1}, \mathbf{y}$

Inizializza: Scegli una distribuzione $Q(\boldsymbol{\theta}_{1:(h-1)}^b, \boldsymbol{\theta}_{h:p}^{b-1}, \mathbf{y})$, chiamata distribuzione **proposta** (**proposal distribution**) con densità $q(\boldsymbol{\theta}_h^*|\boldsymbol{\theta}_{1:(h-1)}^b, \boldsymbol{\theta}_{h:p}^{b-1}, \mathbf{y})$.

1: proponi un valore $\boldsymbol{\theta}_h^*$ dalla proposta

2: calcola

$$\alpha(\boldsymbol{\theta}_h^{b-1}, \boldsymbol{\theta}_h^*) = \alpha = \min\left\{ \frac{f(\boldsymbol{\theta}_h^*|\boldsymbol{\theta}_{1:(h-1)}^b, \boldsymbol{\theta}_{(h+1):p}^{b-1}, \mathbf{y}) q(\boldsymbol{\theta}_h^{b-1}|\boldsymbol{\theta}_{1:(h-1)}^b, \boldsymbol{\theta}_h^*, \boldsymbol{\theta}_{(h+1):p}^{b-1}, \mathbf{y})}{f(\boldsymbol{\theta}_h^{b-1}|\boldsymbol{\theta}_{1:(h-1)}^b, \boldsymbol{\theta}_{(h+1):p}^{b-1}, \mathbf{y}) q(\boldsymbol{\theta}_h^*|\boldsymbol{\theta}_{1:(h-1)}^b, \boldsymbol{\theta}_h^{b-1}, \boldsymbol{\theta}_{(h+1):p}^{b-1}, \mathbf{y})}, 1 \right\}$$

3: $\boldsymbol{\theta}_h^b = \boldsymbol{\theta}_h^*$ con probabilità α, e con probabilità $1-\alpha$ è uguale a $\boldsymbol{\theta}_h^b = \boldsymbol{\theta}_h^{b-1}$

Metropolis-Hastings

Notate che

- L'agoritmo non richiede di campionare dalla full conditional, ma solo di saperne calcolare la densità
- nel calcolo di α, la costante di normalizzazione della full conditional si semplifica.
- se la proposal è una distribuzione simmetrica con media data dal valore precedente della variabili, scompare dal rapporto α.
- l'ultimo step si fa generalmente simulando $U \sim U(0,1)$, e se $u < \alpha$ allora $\boldsymbol{\theta}_h^b = \boldsymbol{\theta}_h^*$, altrimenti $\boldsymbol{\theta}_h^b = \boldsymbol{\theta}_h^{b-1}$
- se la proposal è la full conditional, α è sempre uguale a 1 $\Rightarrow$ Il campionamento diretto si può vedere come un Metropolis in cui la proposta è al full conditional.

Metropolis-Hastings

Possiamo dimostrare che anche il MH lascia invariata la distribuzione stazionaria. In questo caso possiamo dimostrare che un passo Metropolis rispetta il detailed balanced. Prendiamo il caso in cui

- abbiamo solo $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
- facciamo l'update di $\boldsymbol{\theta}_1$

In questo caso dobbiamo dimostrare che

$$f(\boldsymbol{\theta}_1^b | \boldsymbol{\theta}_2^b, \mathbf{y}) T_1((\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b), (\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b)) =$$

$$f(\boldsymbol{\theta}_1^{b+1} | \boldsymbol{\theta}_2^b, \mathbf{y}) T_1((\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b), (\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b))$$

nel caso in cui α sia definito come nell'algoritmo Metropolis. Il caso in cui non si accetta la proposta e quindi $\boldsymbol{\theta}_1^b = \boldsymbol{\theta}_1^{b+1}$, è banalmente vero. Nel caso in cui la proposta viene accettata abbiamo che le transizioni sono

$$T_1((\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b), (\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b)) = q(\boldsymbol{\theta}_1^{b+1} | \boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b, \mathbf{y}) \alpha(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^{b+1})$$

Metropolis-Hastings

e

$$T_1((\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b), (\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b)) = q(\boldsymbol{\theta}_1^b | \boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b, \mathbf{y}) \alpha(\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_1^b)$$

Uno tra

- $\alpha(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^{b+1})$ e
- $\alpha(\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_1^b)$

deve essere 1.

Assumiamo che $\alpha(\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_1^b) = 1$.

Abbiamo quindi che

- $f(\boldsymbol{\theta}_1^b | \boldsymbol{\theta}_2^b, \mathbf{y}) T_1((\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b), (\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b)) = f(\boldsymbol{\theta}_1^b | \boldsymbol{\theta}_2^b, \mathbf{y}) q(\boldsymbol{\theta}_1^{b+1} | \boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b, \mathbf{y}) \alpha(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^{b+1})$
- $f(\boldsymbol{\theta}_1^{b+1} | \boldsymbol{\theta}_2^b, \mathbf{y}) T_1((\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b), (\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b)) = f(\boldsymbol{\theta}_1^{b+1} | \boldsymbol{\theta}_2^b, \mathbf{y}) q(\boldsymbol{\theta}_1^b | \boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b, \mathbf{y})$

Metropolis-Hastings

Se c'è **detailed balance**, dobbiamo avere che queste due quantità siano uguali e quindi segue che

$$\alpha(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^{b+1}) = \frac{f(\boldsymbol{\theta}_1^{b+1}|\boldsymbol{\theta}_2^b, \mathbf{y})q(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b, \mathbf{y})}{f(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_2^b, \mathbf{y})q(\boldsymbol{\theta}_1^{b+1}|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b, \mathbf{y})}$$

che è il valore di $\alpha(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^{b+1})$ del Metropolis quando è diverso da 1.

Se invece è $\alpha(\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_1^{b+1})$ a essere uguale a 1, con calcoli simili arriviamo a

$$\alpha(\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_1^b) = \frac{f(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_2^b, \mathbf{y})q(\boldsymbol{\theta}_1^{b+1}|\boldsymbol{\theta}_1^b, \boldsymbol{\theta}_2^b, \mathbf{y})}{f(\boldsymbol{\theta}_1^{b+1}|\boldsymbol{\theta}_2^b, \mathbf{y})q(\boldsymbol{\theta}_1^b|\boldsymbol{\theta}_1^{b+1}, \boldsymbol{\theta}_2^b, \mathbf{y})}$$

Esempio regressivo

Riprendiamo l'esempio precedente, non assumendo σ^2 noto

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

Noi vogliamo campionare da $f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$, assumendo $f(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta})f(\sigma^2)$ con

$$\boldsymbol{\beta} \sim N_p(\mathbf{M}, \mathbf{V})$$
$$\sigma^2 \sim IG(a, b)$$

che hanno densità

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

$$f(\boldsymbol{\beta}) = (2\pi)^{-\frac{p}{2}}|\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{M})^T\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{M})\right)$$

$$f(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right)$$

Utilizziamo il Gibbs e quindi, all'iterazione b-esima dobbiamo campionare

Esempio regressivo

- $\boldsymbol{\beta}^b$ dalla condizionata $\boldsymbol{\beta}|(\sigma^2)^{b-1}, \mathbf{y}$;
- $(\sigma^2)^b$ dalla condizionata $\sigma^2|\boldsymbol{\beta}^b, \mathbf{y}$

Utilizzando gli stessi passaggi di prima, abbiamo che

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^2 \sim N_p(\mathbf{M}_p, \mathbf{V}_p)$$

con

$$\mathbf{V}_p = (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1})^{-1}$$

$$\mathbf{M}_p = \mathbf{V}_p(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{M})$$

questo perchè quando al condizionante c'è σ^2 questo non fa parte della distribuzione full conditional di $\boldsymbol{\beta}$ ed è considerato come se fosse noto.

Esempio regressivo

Usando il "trucco" del kernel, la full conditional di σ^2 si trova prendendo tutte le componenti della a-posteriori che dipendono da σ^2:

$$(\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)(\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right) =$$

$$(\sigma^2)^{-\left(\frac{n}{2}+a+1\right)} \exp\left(-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2+b}{\sigma^2}\right)$$

che è il kernel di una Gamma inversa. Quindi

$$\sigma^2|\boldsymbol{\beta},\mathbf{y} \sim IG\left(\frac{n}{2}+a, \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+b\right)$$

Possiamo quindi costruire il nostro algoritmo MCMC

Esempio regressivo

Immaginiamo adesso che, nelle stesse condizioni, non riuscissimo a vedere che il kernel della condizionata di σ^2 è di una Gamma inversa. In questo caso, il campione da $\sigma^2|\boldsymbol{\beta}^b, \mathbf{y}$ potrebbe essere ottenuto utilizzando un Metropolis.
Dobbiamo proporre un valore $(\sigma^2)^*$, calcolare l'acceptance ratio

$$\alpha = \min\left\{\frac{f((\sigma^2)^*|\boldsymbol{\beta}^b, \mathbf{y})q((\sigma^2)^{b-1}|(\sigma^2)^*)}{f((\sigma^2)^{b-1}|\boldsymbol{\beta}^j, \mathbf{y})q((\sigma^2)^*|(\sigma^2)^{b-1})}, 1\right\}$$

e vedere, simulando dall'uniforme, se accettare o meno.

Se invece di σ^2 lavorassimo con $\tau^2 = log(\sigma^2)$, possiamo assumere una proposta normale, con media data dal condizionante, e calcolare l'acceptance ratio come

$$\alpha = \min\left\{\frac{f((\tau^2)^*|\boldsymbol{\beta}^b, \mathbf{y})}{f((\tau^2)^{b-1}|\boldsymbol{\beta}^b, \mathbf{y})}, 1\right\}$$

dove

$$f((\tau^2)^*|\boldsymbol{\beta}^b, \mathbf{y})$$

Esempio regressivo

si può calcolare facilmente con le regole di trasformazione di variabili aleatorie, da

$$f((\sigma^2)^*|\boldsymbol{\beta}^b, \mathbf{y}).$$

Esempio regressivo

Facciamo adesso vedere che al posto di $f(\sigma^2|\boldsymbol{\beta}, \mathbf{y})$ potremmo mettere

$$f(\mathbf{y}|\sigma^2, \boldsymbol{\beta}) f(\sigma^2, \boldsymbol{\beta})$$

i.e., il numeratore della a posteriori, e il risultato non cambia. Questo perchè possiamo scrivere la formula precedente come

$$f(\mathbf{y}|\sigma^2, \boldsymbol{\beta}) f(\sigma^2, \boldsymbol{\beta}) = f(\mathbf{y}|\sigma^2, \boldsymbol{\beta}) f(\sigma^2|\boldsymbol{\beta}) f(\boldsymbol{\beta})$$

e abbiamo anche che

$$f(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) = \frac{f(\mathbf{y}|\sigma^2, \boldsymbol{\beta}) f(\sigma^2|\boldsymbol{\beta})}{f(\mathbf{y}|\boldsymbol{\beta})}$$

quindi

$$f(\mathbf{y}|\sigma^2, \boldsymbol{\beta}) f(\sigma^2, \boldsymbol{\beta}) = f(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) f(\mathbf{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})$$

Vediamo cosa succede a α quando mettiamo il numeratore della a posteriori:

$$\alpha = \min \left\{ \frac{f(\mathbf{y}|(\sigma^2)^*, \boldsymbol{\beta}^b) f((\sigma^2)^*, \boldsymbol{\beta}^b) q((\sigma^2)^{b-1}|(\sigma^2)^*)}{f(\mathbf{y}|(\sigma^2)^{b-1}, \boldsymbol{\beta}^b) f((\sigma^2)^{b-1}, \boldsymbol{\beta}^b) q((\sigma^2)^*|(\sigma^2)^{b-1})}, 1 \right\} =$$

Esempio regressivo

$$\min\left\{\frac{f((\sigma^2)^*|\boldsymbol{\beta}^b,\mathbf{y})f(\mathbf{y}|\boldsymbol{\beta}^b)f(\boldsymbol{\beta}^b)q((\sigma^2)^{b-1}|(\sigma^2)^*)}{f((\sigma^2)^{b-1}|\boldsymbol{\beta}^b,\mathbf{y})f(\mathbf{y}|\boldsymbol{\beta}^b)f(\boldsymbol{\beta}^b)q((\sigma^2)^*|(\sigma^2)^{b-1})},1\right\}=$$

$$\min\left\{\frac{f((\sigma^2)^*|\boldsymbol{\beta}^b,\mathbf{y})q((\sigma^2)^{b-1}|(\sigma^2)^*)}{f((\sigma^2)^{b-1}|\boldsymbol{\beta}^b,\mathbf{y})q((\sigma^2)^*|(\sigma^2)^{b-1})},1\right\}$$

che è l'acceptance ratio che si basa sulla full conditional

$$f(\sigma^2|\boldsymbol{\beta},\mathbf{y})$$

Adaptive Proposal

Il problema del MH è sulla scelta della proposal. Anche nel più semplice caso in cui si possa utilizzare una $N(\theta_p^{b-1}, \eta)$, nasce il problema di come decidere η

- un η troppo piccolo mi fa accettare molto, ma le catene saranno molto dipendenti
- un η troppo grande mi fa accettare poco, rendendo i campioni dipendenti

euristicamente, si dovrebbe accettare il 25% delle volte

Esistono dei metodi adattivi che permettono di "trovare" un valore accettabile di η, nel caso univariato e se se può usare una proposta normale $N(\theta_p^{b-1}, \eta)$. Definiamo

- b è l'iterazione
- α^* come un rate di accettazione desiderato
- A, D come delle costanti
- c specifica ogni quante iterazioni bisogna aggiornare η (c=50 è un valore classico)
- $\bar{\alpha}$ la media campionaria degli α del Metropolis nelle ultima c iterazioni.

Adaptive Proposal

definiamo un valore iniziale per η, e poi ogni c iterazioni, e dopo aver fatto il passo Metropolis, modifichiamo la varianza come

$$\eta \leftarrow\!\shortmid \exp\left(log(\eta) + \gamma_b(\bar{\alpha} - \alpha^*)\right)$$

dove γ_b è una funzione

- che $\to 0$ con $b \to \infty$ (condizione di **diminish adaptation**)
- $\gamma_{b+1} \leq \gamma_b$
- $0 \leq \gamma_b < 1$

Classiche scelte sono

$$\gamma_b = \frac{A}{D+b} \qquad \gamma_b = \frac{1}{b^A}$$

Notate che in questo caso l'algoritmo MCMC non è più Markoviano, e per questa ragione serve la condizione di diminish adaptation per avere un algoritmo con la giusta distribuzione stazionaria.

Adaptive Proposal

Nel caso multivariato, possiamo fare qualcosa di simile se la proposta è una normale multivariata $N(\boldsymbol{\theta}_p^{b-1}, \lambda\boldsymbol{\Sigma})$. In questo caso, ad ogni iterazione, dopo aver fatto il passo Metropolis, potete aggiornare

$$\lambda \leftarrow\!\shortmid \exp\left(log(\lambda) + \gamma_b(\alpha - \alpha^*)\right)$$

$$\boldsymbol{\mu} \leftarrow\!\shortmid \boldsymbol{\mu} + \gamma_b(\boldsymbol{\theta}_p^b - \boldsymbol{\mu})$$

$$\boldsymbol{\Sigma} \leftarrow\!\shortmid \boldsymbol{\Sigma} + \gamma_b\left(\left(\boldsymbol{\theta}_p^b - \boldsymbol{\mu}\right)\left(\boldsymbol{\theta}_p^b - \boldsymbol{\mu}\right)' - \boldsymbol{\Sigma}\right)$$

dove α è il rapporto metropolis, e λ, $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ vanno inizializzate.

Capita spesso, soprattutto nelle prime iterazioni, che $\boldsymbol{\Sigma}$ sia numericamente singolare, e quindi, si tenda a utilizzare

$$N(\boldsymbol{\theta}_p^{b-1}, \lambda\boldsymbol{\Sigma} + \epsilon\mathbf{I})$$

dove ϵ è un valore piccolo rispetto a $\boldsymbol{\Sigma}$, invece di

$$N(\boldsymbol{\theta}_p^{b-1}, \lambda\boldsymbol{\Sigma})$$

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo

Il problema del Gibbs Samples, è che se anche teoricamente le componenti delle variabili aleatorie possono essere multivariati, quando la dimensione è troppo grande, si tende a accettare poco. Esiste un metodo, che può essere utilizzato come alternativa al Gibbs Sampler, o come un singolo passo del Gibbs. Questo è l'Hamiltonian Monte Carlo.
Da un punto di vista molto generale, L'Hamiltoniana è solo un modo per fare una **proposta in un Metropolis** in maniera "furba".

Immaginiamo di voler campionare da

$$f(\boldsymbol{\theta}|\mathbf{y})$$

dove $\boldsymbol{\theta}$ è $d-$dimensionale, utilizzando un metodo del tipo Metropolis. Possiamo introdurre un vettore di variabile latenti, indicate con $\mathbf{p}$, della stessa dimensione di $\boldsymbol{\theta}$ e, invece di $f(\boldsymbol{\theta}|\mathbf{y})$, proviamo a campionare da

$$f(\mathbf{p})f(\boldsymbol{\theta}|\mathbf{y})$$

Hamiltonian Monte Carlo

con la classica assunzione che

$$f(\mathbf{p}) \propto \exp\left(-\frac{\mathbf{p}'\mathbf{p}}{2}\right)$$

Introduciamo la funzione Hamiltoniana

$$H(\mathbf{p}, \boldsymbol{\theta}) = -log(f(\mathbf{p})) - log(f(\boldsymbol{\theta}|\mathbf{y})) = K(\mathbf{p}) + V(\boldsymbol{\theta})$$

dove possiamo interpretare $K(\mathbf{p})$ come l'energia cinetica e $V(\mathbf{q})$ quella potenziale. Naturalmente la distribuzione di interesse è

$$f(\mathbf{p})f(\boldsymbol{\theta}|\mathbf{y}) \propto \exp(-H(\mathbf{p}, \boldsymbol{\theta}))$$

Se io voglio campionare dalla distribuzione di interesse, posso usare il sistema Hamiltoniamo per muovermi nello spazio, e proporre un valore **decidendo poi se accettarlo o meno**, visto che

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial K(\mathbf{p})}{\partial \mathbf{p}}$$

$$\frac{d\mathbf{p}}{dt} = -\frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Hamiltonian Monte Carlo

Possiamo usare il **leap-frog** per ottenere dei valori proposti:

$$\mathbf{p} \leftarrow\!\shortmid \mathbf{p} - \frac{\epsilon}{2}\frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\boldsymbol{\theta} \leftarrow\!\shortmid \boldsymbol{\theta} + \epsilon \mathbf{p}$$

$$\mathbf{p} \leftarrow\!\shortmid \mathbf{p} - \frac{\epsilon}{2}\frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Quando vogliamo fare un passo per campionare una variabile $\boldsymbol{\theta}$ dato che abbiamo il $b-$esimo valore simualato $\boldsymbol{\theta}^b$

- decidiamo un ϵ
- generiamo un $\mathbf{p}^b$ dalla sua marginale (in questo caso $N(\mathbf{0}, \mathbf{I})$)
- utilizziamo il leap-frog per L iterezioni e chiamiamo gli ultimi valori come $\boldsymbol{\theta}^*$ e $\mathbf{p}^*$
- calcoliamo il rapporto metropolis

$$\alpha((\mathbf{p}^b, \boldsymbol{\theta}^b), (\mathbf{p}^*, \boldsymbol{\theta}^*)) = min\{1, \exp(H(\mathbf{p}^b, \boldsymbol{\theta}^b) - H(\mathbf{p}^*, \boldsymbol{\theta}^*))\}$$

  accettare il nuovo valore con probabilità α

Hamiltonian Monte Carlo

- Se invece del leap-frog, si potessero usare direttamente le equazioni Hamiltoniane, accettarei sempre perchè il valore dell'Hamiltoniano non cambia.
- Il leap-frog introduce un piccolo "errore" e per questo il rapporto potrebbe essere diverso, ma è generalmente vicino a 1. Anceh se si accetta molto i campioni, se L è grande, i campioni sono quasi indipendenti
- l'HMC funzione per qualsiasi dimensione, anche se d è elevato. Per questo motivo si può usare per proporre/accettare l'intero set delle variabili della a posteriori
- **Problema:** vale solo per le variabili continue

Attenzione: questa è una proposta deterministica. In generale, le proposte deterministiche si possono fare, ma se ne deve tener conto nel rate Metropolis visto che invece che una proposta vera e propria, può essere vista come una trasformazione di variabili.
Nel Rapporto Metropolis bisognerebbe mettere lo Jacobiano della trasformazione $\Rightarrow$ Questo è uguale a 1 per trasformazioni che non cambiano "volumi", e il leap-frog non cambia il volume dello spazio

Inferenza

Statistiche a posteriori

Un aspetto interessante e utile dei metodi Monte Carlo è che se vogliamo un campione da

$$f(\mathbf{z}|\boldsymbol{\psi})$$

dove $\mathbf{z} = (z_1, \ldots, z_g)$ è multivariata, visto che

$$f(\mathbf{z}|\boldsymbol{\psi}) = f(z_1|\boldsymbol{\psi}) f(z_2|z_1, \boldsymbol{\psi}) f(z_3|z_2, z_1, \boldsymbol{\psi}) \ldots f(z_g|z_{g-1}, z_{g-2}, \ldots, z_1, \boldsymbol{\psi})$$

possiamo ottenerlo simulando in sequenza

- il b-esimo campione z_1^b di z_1 da $f(z_1|\boldsymbol{\psi})$
- il b-esimo campione z_2^b di z_2 da $f(z_2|z_1^b, \boldsymbol{\psi})$
- il b-esimo campione z_3^b di z_3 da $f(z_3|z_2^b, z_1^b, \boldsymbol{\psi})$
- ...
- il b-esimo campione z_g^b di z_g da $f(z_g|z_{g-1}^b, z_{g-2}^b, \ldots, z_1^b, \boldsymbol{\psi})$

Statistiche a posteriori

Il vettore $\mathbf{z}^b = (z_1^b, z_2^b, \ldots, z_g^b)$ è il b-esimo campione da $f(\mathbf{z}|\boldsymbol{\psi})$.

E' facile vedere come il set di campioni $(z_1^1, z_1^2, \ldots, z_1^B)$, se presi singolarmente, sono stati tutti simulati dalla marginale $f(z_1|\boldsymbol{\psi})$, e quindi sono un set di campioni dalla marginale. Ma visto che l'ordine dato al campionamento delle variabili è invariante, cioè

- il b-esimo campione z_2^b di z_2 da $f(z_2|\boldsymbol{\psi})$
- il b-esimo campione z_1^b di z_1 da $f(z_1|z_2^b, \boldsymbol{\psi})$
- il b-esimo campione z_3^b di z_3 da $f(z_3|z_2^b, z_1^b, \boldsymbol{\psi})$
- ...
- il b-esimo campione z_g^b di z_g da $f(z_g|z_{g-1}^b, z_{g-2}^b, \ldots, z_1^b, \boldsymbol{\psi})$

Statistiche a posteriori

produce sempre un campione da $f(\mathbf{z}|\boldsymbol{\psi})$, è vero anche che i valori $(z_2^1, z_2^2, \ldots, z_2^B)$, presi singolarmente, sono campioni dalla marginale $f(z_2|\boldsymbol{\psi})$. Per simmetria questo è quindi vero per tutte le variabili.
Per vederlo meglio, assumiamo che $g = 2$, scriviamo la congiunta come

$$f(z_1, z_2) = f(z_1)f(z_2|z_1) = f(z_2)f(z_1|z_2)$$

quindi se abbiamo un campione da $f(z_1, z_2)$ possiamo averlo ottenuto simulando z_1 dalla marginale $(f(z_1)f(z_2|z_1))$ oppure simulando z_2 dalla marginale $(f(z_2)f(z_1|z_2))$.

Quindi, preso un campione a posteriori $\mathbf{z}^b = (z_1^b, z_2^b, \ldots, z_g^b)$,

- preso tutto insieme, proviene dalla distribuzione $f(\mathbf{z}|\boldsymbol{\psi})$
- z_1^b, preso singolarmente, proviene da $f(z_1|\boldsymbol{\psi})$
- z_2^b, preso singolarmente, proviene da $f(z_2|\boldsymbol{\psi})$
- z_3^b, preso singolarmente, proviene da $f(z_3|\boldsymbol{\psi})$
- ...

Statistiche a posteriori

- (z_1^b, z_3^b), presi in coppia, provengono da $f(z_1, z_3|\boldsymbol{\psi})$
- ...

Quanto detto sopra è molto utile perchè in generale anche se abbiamo B campioni dalla a posteriori

$$f(\theta_1, \ldots, \theta_p|\mathbf{y})$$

noi siamo generalemte interessate alle marginali a posteriori

$$f(\theta_1|\mathbf{y}) \quad f(\theta_2|\mathbf{y}) \ldots \quad f(\theta_p|\mathbf{y})$$

Le cose che sono generalmente interessanti sono

- Media a posteriori $E(\theta_j|\mathbf{y}) = \sum_{b=1}^{B} \frac{\theta_j^b}{B}$
- Varianza a posteriori $V(\theta_j|\mathbf{y}) = \sum_{b=1}^{B} \frac{(\theta_j^b)^2}{B} - \left(\sum_{b=1}^{B} \frac{\theta_j^b}{B} \right)^2$

Statistiche a posteriori

- Intervallo di credibilità a livello $1-\alpha$: $IC = [q_l, q_U]$ tale che

$$1-\alpha = \int_{q_l}^{q_u} f(\theta_j|\mathbf{y})d\theta_j$$

  si trova utilizzando l'intersa generalizzata e i quantili empirici.

- Previsione di una nuova osservazione y_0 dalla **distribuzione predittiva**:

$$f(y_0|\mathbf{y}) = \int f(y_0|\theta_1, \ldots, \theta_p, \mathbf{y}) f(\theta_1, \ldots, \theta_p|\mathbf{y})$$

- ...

Bontà del Modello

Per valutare la bontà di un modello, abbiamo due scelte:

- informational criteria (criteri informazionali)
- cross validation

I criteri informationali sono indici che tengono in considerazione sia la verosimiglianza del modello, che la sua complessità
Indichiamo con $\hat{\boldsymbol{\theta}}$ una stima Bayesiana dei parametri (per esempio la media) e con $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ la verosimiglianza. Alcuni esempi di informational criteria sono:

- AIC: $-2log(f(\mathbf{y}|\hat{\boldsymbol{\theta}})) + 2\#param$,
- BIC: $-2log(f(\mathbf{y}|\hat{\boldsymbol{\theta}})) + log(n)\#param$

Bontà del Modello

- WAIC che è una versione "Bayesiana" dell'AIC, che si calcola con un integrazione MC

$$\text{WAIC} = -2\sum_{i=1}^{n}\log\left(\frac{1}{B}\sum_{b=1}^{B} f(y_i|\boldsymbol{\theta}^b)\right) + 2\sum_{i=1}^{n}\text{Var}_{\text{posterior}}\left(\log f(y_i|\boldsymbol{\theta})\right)$$

  dove con

$$\text{Var}_{\text{posterior}}\left(\log f(y_i|\boldsymbol{\theta})\right)$$

  intendo una stima della varianza con i campioni della a posteriori. L'indice ha senso però solo se i dati sono indipendenti, dato il processo

- DIC ...
- ...

Bontà del Modello

Se avete diversi modelli, potete calcolare lo stesso indice per ogni modello, e prendere quello con indice più piccolo. Fate attenzione che potete confrontare modelli solo se la $\mathbf{y}$ è la stessa

I problemi maggiori sono da ricercare nella definizione stessa di $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$, per esempio, se riprendiamo l'esempio precedente

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 t + w_t + \epsilon_t \\
\epsilon_t &\overset{iid}{\sim} N(0, \sigma^2) \\
w_t &\sim N(\alpha w_{t-1}, \tau^2), t = 2, \ldots, T
\end{aligned}
$$

le w_t sono parametri? oppure dobbiamo marginalizzare sui w_t? sappiamo poi calcolare la densità dei dati? e i missing values come li trattiamo?

Bontà del Modello

L'altro approccio è più generale e prevede di

- dividere i dati in due set $\mathbf{y}_{obs}$ e $\mathbf{y}_0 = (y_{0,1}, \ldots y_{0,m})$
- utilizzare solo il primo per stimare la a posteriori $f(\boldsymbol{\theta}|\mathbf{y}_{obs})$
- prevedere i valori di $\mathbf{y}_0$.
- confrontare in qualche modo la media a posteriori di $\mathbf{y}_0$ con i valori veri osservati (e.g., differenza al quadrato medie)

Bontà del Modello

Per la previsione di $\mathbf{y}_0$ avete due possibilità:

- Questo si può fare dopo aver stimato il modello, ottenendo campioni dalla predittiva

$$f(\mathbf{y}_0|\mathbf{y}_{obs}) = \int f(\mathbf{y}_0|\boldsymbol{\theta}, \mathbf{y}_{obs}) f(\boldsymbol{\theta}|\mathbf{y}_{obs}) d\lambda(\boldsymbol{\theta}) = \int f(\mathbf{y}_0, \boldsymbol{\theta}|\mathbf{y}_{obs}) d\lambda(\boldsymbol{\theta})$$

- Oppure assumendo che $\mathbf{y}_0$ siano missing e stimare la a posteriori

$$f(\mathbf{y}_0, \boldsymbol{\theta}|\mathbf{y}_{obs})$$

I due approcci danno lo stesso identico risultato, visto che anche se stimo la a posteriori $f(\mathbf{y}_0, \boldsymbol{\theta}|\mathbf{y}_{obs})$, i campioni di $\mathbf{y}_0$ si possono vedere come provenienti dalla marginale $f(\mathbf{y}_0|\mathbf{y}_{obs})$

Bontà del Modello

Per il confronto potete calcolare il valore atteso a posteriori

$$\hat{y}_{0,j}$$

e poi calcolare una differenza quadratica media (MSE)

$$\frac{\sum_{j=1}^{m}(y_{0,j} - \hat{y}_{0,j})^2}{m}$$

o in valore assoluto

$$\frac{\sum_{j=1}^{m}|y_{0,j} - \hat{y}_{0,j}|}{m}$$

Anche se questo è l'approccio più usato, poichè facile da implementare, sta negando tutta l'incertezza presente nella a posteriori. Un metodo più formalmente corretto è il **continuous ranked probability score** (CRPS), che si usa per confrontare una Cumulate F con un valore puntuale

$$CRPS(F, y) = \int (F - \eta(x - y))^2 d\lambda(x)$$

Bontà del Modello

dove $\eta(x-y)$ è la Heaviside step function, che assume valore 0 se è negativa, e 1 se positiva. Si può dimostrare che sotto opportune condizioni (generalmente valide per variabili che discrete o continue univariate, ma per esempio non vale per variabili circolari), si ha che

$$CRPS(F,y) = E_X(|X-y|) - \frac{1}{2}E_{X,X'}(|X-X'|)$$

dove sia X che X' provengono da $F()$
Se F è la cumulata della predittiva $f(y_{0,j}|\mathbf{y}_{obs})$, $F(y_{0,j}|\mathbf{y}_{obs})$, allora possiamo calcolare

$$CRPS(F(y_{0,j}|\mathbf{y}_{obs}), y_{0,j}) = E_X(|X-y_{0,j}|) - \frac{1}{2}E_{X,X'}(|X-X'|)$$

e più questo valore è piccolo e più la previsione è buona.
Visto che sono valori attesi, possiamo usare MC per approssimarli con

$$CRPS(F(y_{0,j}|\mathbf{y}_{obs}), y_{0,j}) = \frac{\sum_{b=1}^{B}|y_{0,j}^b - y_{0,j}|}{B} - \frac{1}{2}\frac{\sum_{b=1}^{B}\sum_{h=1}^{B}|y_{0,j}^b - y_{0,j}^h|}{B^2}$$