Metodi Monte Carlo

Vers. 1.0.1

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Numeri Pseudo Casuali

Numeri pseudo casuali

Prima di introdurre i vari metodi, chiariamo cosa funziona la simulazione di variabili aleatorie. Il punto fondamentale è che un computer non può simulare qualcosa che sia "random", ma solo valori **pseudo-random**, quindi deterministici (**P.S.:** esistono siti che vendono numeri "veramente" casuali, per esempio `https://www.random.org`)
Partiamo dal caso uniforme, possiamo dire che

Definizione - uniform Pseudo Random generator

Un generatore di variabili pseudo-casuali $U(0, 1)$, è un algoritmo che partendo da un punto iniziale u_0 (seed) e una trasformazione $D()$, produce una sequenza di valori $(u_1, \ldots, u_n)$, con $u_i = D(u_{i-1}) \in [0, 1]$. Il vettore $(u_1, \ldots, u_n)$ riproduce le "caratteristiche" di un vero vettore di realizzazioni di variabili aleatoria $(V_1, \ldots, V_n)$ iid da $U(0, 1)$.

Alcune osservazioni:

- L'algoritmo è deterministico, e conoscendo il seme (seed) e la funzione $D()$, possiamo ricostruire il campione pseudo-casuale;
- Con "riproduce le caratteristiche" si intende che nessun test del tipo

$$H_0 : U_1, \dots, U_n \overset{iid}{\sim} U(0,1)$$

  venga rifiutato.
- Le sequenze sono periodiche.

L'idea è quindi che la sequenza di numeri pseudo-casuali sia indistinguibile da una sequenza di numeri realmente casuali. In R possiamo fare quest'esempio.

```
set.seed(100)
runif(10,0,1)

set.seed(100)
runif(10,0,1)

runif(10,0,1)
```

Integrazione Monte Carlo

Nel codice sopra, i primi 2 runif daranno lo stesso risultato (perchè risettate il seed), nel terzo i risultati sono diversi. Un semplice esempio di generatore pseudo casuale $U(0,1)$ si può ottenere creando le variabili

$$d_i = (ad_{i-1} + c) \text{ mod } m$$

dove $0 \leq c \leq m$ (incremento) e $0 < a < m$ (moltiplicatore), e dopo prendendo

$$u_i = d_i/m.$$

I valori m, a e c, vanno scelti in modo da non creare loop.

Integrazione Monte Carlo

Integrazione

In molti studi Monte Carlo, lo scopo è studiare delle caratteristiche delle variabili aleatorie $X \in \mathcal{X}$ nella forma

$$E(h(X)) = \int_{\mathcal{X}} h(x)f(x)dx$$

se i dati sono continui, oppure

$$E(h(X)) = \sum_{x \in \mathcal{X}} h(x)P(X = x)$$

se discreti e

$$E(h(X)) = \sum_{x \in \mathcal{X}_D} h(x)P(X = x) + \int_{\mathcal{X}_c} h(x)f(x)dx$$

se la variabile è mista con $\mathcal{X}_D$ parte del dominio discreto e $\mathcal{X}_C$ continuo con $\mathcal{X}_D \cup \mathcal{X}_C = \mathcal{X}$. Oppure, per scriverli in una forma unificata

$$E(h(X)) = \int_{\mathcal{X}} h(x)f(x)d\lambda(x)$$

con $d\lambda(x)$

- misura di Lebesgue: $d\lambda(x) = dx$ nel caso continuo
- misura di conteggio nel caso discreto.
- misura ibrida in caso di variabili miste, ovvero

$$d\lambda(x) = d\lambda_D(x) + d\lambda_C(x)$$

dove $d\lambda_D(x)$ è la misura di conteggio su $\mathcal{X}_D$ e $d\lambda_C(x)$ è la misura di Lebesgue su $\mathcal{X}_C$. In questo caso,

$$E(h(X)) = \sum_{x \in \mathcal{X}_D} h(x)P(X = x) + \int_{\mathcal{X}_C} h(x)f(x)dx$$

Esempi

- $h(X) = X$ (media)
- $h(X) = X^2$ (secondo momento)
- ...

Consideriamo un campione di variabili identicamente distribuite, non necessariamente indipendenti (cioè con la stessa marginale) $(X_1, \ldots, X_n)$. Sappiamo che

$$\bar{h}_n = \frac{\sum_{i=1}^n h(X_i)}{n}$$

converge quasi certamente a $E(h(X))$ per la legge dei grandi numeri. Questo risultato vale anche per funzioni h diverse dall'identità. Se poniamo $Y = h(X)$ e indichiamo con $f_Y(y)$ la densità di Y, allora

$$\int_{\mathcal{X}} h(x)f(x)d\lambda(x) = \int_{\mathcal{Y}} y f_Y(y)d\lambda(y) \approx \bar{Y}$$

Se $h^2(X)$ ha attesa finita, possiamo calcolarne anche la varianza

$$v_n = \mathrm{var}(\bar{h}_n) = \mathrm{var}\left(\frac{\sum_{i=1}^n h(X_i)}{n}\right) =$$

$$\frac{1}{n^2}\left(\sum_{i=1}^n var\left(h(X_i)\right) + \sum_{i=1}^n \sum_{j=1}^n cov\left(h(X_i), h(X_j)\right)\right)$$

Integrazione Monte Carlo

Nel caso in cui le variabili siano indipendenti abbiamo che

$$\mathrm{var}(\bar{h}_n) = \frac{\mathrm{var}\,(h(X_i))}{n} = \frac{1}{n}\int_{\mathcal{X}} (h(x) - E(h(X)))^2 f(x) d\lambda(x)$$

che si può approssimare con

$$v_n = \frac{1}{n^2}\sum_{i=1}^{n} (h(x_i) - \bar{h}_n)^2$$

Attenzione:

i) Lo stimatore della varianza non è uno stimatore Monte Carlo.

ii) Si può dimostrare che è distorto

iii) Una versione non distorta dello stimatore della varianza si trova sostituendo$(n-1)$ a n nel denominatore

Naturalmente più piccola è $\mathrm{var}(\bar{h}_n)$ e migliore sarà la stima di $E(h(X))$.

Da ora in poi, a meno che non sia detto esplicitamente, assumeremo che le variabili siano anche indipendenti

Integrazione Monte Carlo

Monte Carlo vale anche per funzioni di più variabile:

$$E(h(X,Y)) = \int_{\mathcal{X}} \int_{\mathcal{Y}} h(x,y) f(x,y) d\lambda(y) d\lambda(x) \approx \frac{\sum_{i=1}^{n} h(x_i, y_i)}{n}$$

e anche nel caso in cui $h(x,y) = h^*(x)$:

$$E(h^*(X)) = \int_{\mathcal{X}} h^*(x) f(x) d\lambda(x) =$$

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} h(x,y) f(x,y) d\lambda(y) d\lambda(x) \approx \frac{\sum_{i=1}^{n} h(x_i, y_i)}{n} = \frac{\sum_{i=1}^{n} h^*(x_i)}{n}$$

Funzione di ripartizione

La funzione di ripartizione per una variabile può essere scritta come

$$F(a) = \int_{-\infty}^{a} f(x)d\lambda(x)$$

che non è nella forma che ci permette di usare un metodo Monte Carlo. Possiamo però introdurre la funzione $\mathbf{1}_a(x)$ che assume valore 1 se $x \leq a$, altrimenti zero, e usare

$$\int_{-\infty}^{a} f(x)d\lambda(x) = \int_{-\infty}^{\infty} \mathbf{1}_a(x)f(x)d\lambda(x) = E(\mathbf{1}_a(X)) \approx \frac{\sum_{i=1}^{n} \mathbf{1}_a(X_i)}{n}$$

dove naturalmente $X_i \sim F$ e in questo caso $h(X) = \mathbf{1}_a(X)$.

Possiamo quindi stimare l'intera funzione di ripartizione con

$$\hat{F}(a) = \frac{\sum_{i=1}^{n} \mathbf{1}_a(X_i)}{n} \approx E(\mathbf{1}_a(X)) = \int_{-\infty}^{\infty} \mathbf{1}_a(x)f(x)dx = \int_{-\infty}^{a} f(x)dx = F(a)$$

$\hat{F}(a)$ è una stima della funzione di ripartizione

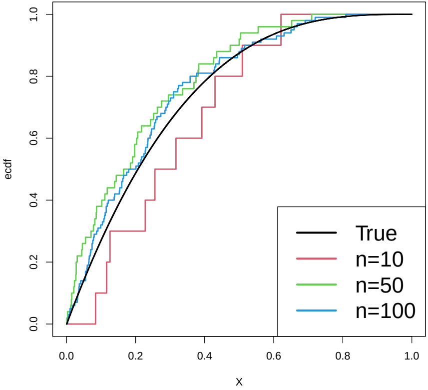

Figure: ecdf di una $B(1,3)$ per diversi valori di n

Code: Codice della figura

```
# simuliamo dalla beta
x1 = rbeta(10,1,3)
x2 = rbeta(50,1,3)
x3 = rbeta(100,1,3)

# cacoliamo le funzioni di ripartizioni empiriche su xseq
xseq= seq(0,1, by=0.001)
ec1 = ecdf(x1)(xseq)
ec2 = ecdf(x2)(xseq)
ec3 = ecdf(x3)(xseq)

plot(xseq,ec1, type="s", lwd=2, col=2, xlab="X", ylab="ecdf")
lines(xseq,ec2, type="s", col=3, lwd=2)
lines(xseq,ec3, type="s", col=4, lwd=2)
lines(xseq, pbeta(xseq,1,3), type="s", col=1, lwd=2)
legend("bottomright", c("True","n=10","n=50","n=100")
        ,col=1:4, lwd=3, cex=2
        )
```

Come si vede dalla figura la funzione è a scalini, con scalini di altezza $1/n$ e per $X_{i-1} < a \leq X_i$ è costante.
Dalla funzione di ripartizione stimata (chiamata anche **funzione di ripartizione empirica**), si possono calcolare i quantili empirici. Per esempio il quantile empirico di livello p è

$$min\{x_i : i = 1, \ldots, n \text{ and } \hat{F}(x) \geq p\}$$

Stima di densità

Un caso interessante si presenta quando non conosciamo la forma esplicita di $f(x)$, ma conosciamo $f(x|y)$, dove $Y \sim G$ è definita su $\mathcal{Y}$, e sappiamo campionare da G.
Possiamo calcolare la densità $f(x)$ in un punto specifico x^*:

$$f(x^*) = \int_{\mathcal{Y}} f(x^*|y)g(y)d\lambda(y)$$

e se indichiamo con $Y_1, \ldots, Y_n$ campioni iid da G, e assumiamo $h(X) = f(x^*|y)$, allora

$$\frac{\sum_{i=1}^{n} f(x^*|Y_i)}{n} \approx \int_{\mathcal{Y}} f(x^*|y)g(y)d\lambda(y)$$

Densità marginale

Esempio

Ipotizziamo che $Y \sim \text{Exp}(\lambda)$, e $X|Y = y \sim G(\exp y, 1)$, come è fatta la densità marginale di X?.

Questo è un semplice caso in cui non siamo in grado di descrivere analiticamente una distribuzione e possiamo usare la stima Monte Carlo.

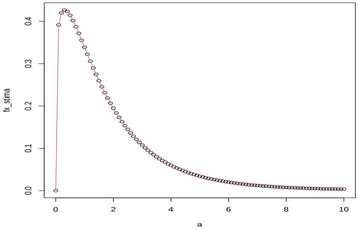

Figure: Esempio MC

Code: Codice della figura

```
n = 1000

# valori di x su cui calcolare la densita'
a = seq(0,10,by=0.1)
lambda = 2


fx_stima = c()
y = rexp(n,lambda)
for(i in 1:length(a))
{
    # densita' condizionata
  fzgiveny = dgamma(a[i],exp(y),1)
    # stima di fx nel punto a[i]
  fx_stima[i] = sum(fzgiveny)/n
}
plot(a,fx_stima)
lines(a,fx_stima, col=2)
```

Naturalmente non sappiamo determinare la forma funzionale, ma possiamo trovarla utilizzando qualche metodo interpolante, o usando funzioni a tratti.

Rao-Blackwellization

Prendiamo la seguente relazione

$$E(h(X)) = \int_{\mathcal{X}} h(x) f_X(x) d\lambda(x) =$$

$$\int_{\mathcal{Y}} \int_{\mathcal{X}} h(x) f(x|y) f(y) d\lambda(x)\, d\lambda(y) = \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} h(x) f(x|y) d\lambda(x) \right) f(y) d\lambda(y)$$

e quindi

$$E_x(h(X)) = \int_{\mathcal{Y}} E_x(h(X)|y) f(y) d\lambda(y)$$

quindi questi valori attesi potremmo calcolarli usando i campioni di X, con

$$\frac{\sum_{i=1}^n h(X_i)}{n} \approx E(h(X))$$

o con quelli di Y, con

$$\frac{\sum_{i=1}^n E_x(h(X)|y_i)}{n} \approx \int_{\mathcal{Y}} E_x(h(X)|y) f(y) d\lambda(y)$$

se siamo in grado di calcolare $E(h(X)|Y_i)$. Questa eguaglianza tra i valori attesi è la classica legge

$$E(X) = E_y(E_x(X|Y))$$

Assumendo indipendenza, abbiamo che

$$var_x\left(\frac{\sum_{i=1}^n h(X_i)}{n}\right) = \frac{\sum_{i=1}^n var_x(h(X_i))}{n^2}$$

e

$$var_y\left(\frac{\sum_{i=1}^n E_x(h(X)|Y_i)}{n}\right) = \frac{\sum_{i=1}^n var_y(E_x(h(X)|Y_i))}{n^2}$$

Sebbene i due valori attesi stimano/calcolano la stessa cosa, abbiamo che

$$var_x(h(X)) = var_y(E_x(h(X)|Y)) + E_y(var_x(h(X)|Y))) \geq var_y(E_x(h(X)|Y))$$

(legge della varianza totale)

Abbiamo quindi che

$$var_x\left(\frac{\sum_{i=1}^{n} h(X_i)}{n}\right) \geq var_y\left(\frac{\sum_{i=1}^{n} E_x(h(X)|Y_i)}{n}\right)$$

e quindi la stima

$$\frac{\sum_{i=1}^{n} E_x(h(X)|Y_i)}{n}$$

è migliore in termine di varianza.

Lo stimatore che utilizza la distribuzione condizionata per calcolare il valore atteso marginale di $h(X)$, si chiama **Rao-blackwell estimator**.

Importance sampling

Ritorniamo al problema del calcolo di attese

$$E(h(X)) = \int_{\mathcal{X}} h(x)f(x)d\lambda(x)$$

Questo integrale può essere scritto come

$$E(h(X)) = \int_{\mathcal{X}} h(x)f(x)d\lambda(x) = \int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)} g(x)d\lambda(x)$$

dove $g(x)$ è una qualsiasi funzione per cui $g(x) > 0$ per ogni $x \in \mathcal{X}$ tale che $f(x) > 0$.

Nel caso in cui $g(x)$ è una densità abbiamo che

$$\int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)} g(x)d\lambda(x) = E_g\left(\frac{h(X)f(X)}{g(X)}\right)$$

Come per l'esempio precedente,

$$\int_{\mathcal{X}} h(x)f(x)d\lambda(x)$$

Importance sampling

e

$$\int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)} g(x) d\lambda(x)$$

calcolano la stessa cosa,

$$E_g\left(\frac{h(X)f(X)}{g(X)}\right) = E_f(h(X))$$

Abbiamo quindi due possibili stimatori

- con $X_i \sim f$

$$\frac{\sum_{i=1}^{n} h(x_i)}{n} \approx E_f(h(X))$$

- con $X_i \sim g$

$$\frac{\sum_{i=1}^{n} \frac{h(x_i)f(x_i)}{g(x_i)}}{n} \approx E_g\left(\frac{h(X)f(X)}{g(X)}\right)$$

Importance sampling

Possiamo vedere quale stimatore ha varianza minore, se assumiamo indipendenza, e calcoliamo

$$var_f(h(X)) = E(h(X)^2) - E^2(h(X)) = \int_{\mathcal{X}} h^2(x) f(x) d\lambda(x) - E_f^2(h(X))$$

e

$$var_g\left(\frac{h(X)f(X)}{g(X)}\right) = \int_{\mathcal{X}} \frac{h^2(x)f^2(x)}{g^2(x)} g(x) d\lambda(x) - E_f^2(h(X))$$

dove, nella seconda varianza, abbiamo usato la relazione

$$E_g^2\left(\frac{h(X)f(X)}{g(X)}\right) = E_f^2(h(X))$$

Abbiamo allora che l'importance sampling è migliore se

$$var_f(h(X)) - var_g\left(\frac{h(X)f(X)}{g(X)}\right) > 0$$

Abbiamo che

$$var_f(h(X)) - var_g\left(\frac{h(X)f(X)}{g(X)}\right) = \int_{\mathcal{X}} \left(1 - \frac{f(x)}{g(x)}\right) h^2(x) f(x) d\lambda(x)$$

Importance sampling

visto che $f(x)$ e $g(x)$ sono densità, il rapporto $f(x)/g(x)$ non può sempre essere $<$1 o $>$1. Quindi, per avere una differenza di varianze positiva (i.e., importance sampling migliore del vanilla), dobbiamo avere che

- $f(x)/g(x) > 1$ se $h^2(x)f(x)$ è piccolo; $1 - \frac{f(x)}{g(x)}$ negativo;
- $f(x)/g(x) < 1$ se $h^2(x)f(x)$ è grande; $1 - \frac{f(x)}{g(x)}$ positivo.

quindi $g(x)$ deve essere elevata (i.e., importante), per i punti dove $h^2(x)f(x)$ è elevata. Si può dimostrare che la scelta migliore per $g(x)$ è

$$g(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(x)|f(x)d\lambda(x)}$$

Simulazioni

Simulazione

Se sappiamo generare da una distribuzione, con trasformazioni di variabili possiamo ottenere campioni da altre distribuzioni.

Esempio

Se $U \sim U(0,1)$, allora $X = -\log U \sim Exp(1)$ e $\frac{X}{\lambda} \sim Exp(\lambda)$

Soluzione:

possiamo usare la regola della trasformazione di variabili:

$$f_X(x) = f_u(u) \left| \frac{du}{dx} \right| = |-\exp(-x)| = \exp(-x)$$

e

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \exp(-\lambda y) \, |\lambda| = \lambda \exp(-\lambda y)$$

$\square$

Esempio - Normale Multivariata

Se $U_1 \sim U(0,1)$ e $U_2 \sim U(0,1)$, allora

$$X = \sqrt{-2\log U_1}\cos(2\pi U_2) \quad Y = \sqrt{-2\log U_1}\sin(2\pi U_2)$$

sono normali standard indipendenti (si chiama trasformazione di Box–Muller) e

$$\mathbf{Z} = \boldsymbol{\mu} + \mathbf{A}\left(\begin{array}{c} X \\ Y \end{array}\right)$$

dove $\boldsymbol{\mu} \in \mathbb{R}^2$ e $\mathbf{A}$ è una matrice 2×2 con elementi in $\mathbb{R}$, allora $\mathbf{Z} \sim N_2(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^T)$ (quest'ultima parte è facilmente dimostrabile).

e quindi da una coppia di uniformi possiamo avere qualsiasi tipo di normale bivariata. Le relazioni tra variabili aleatorie sono molte e spesso da una si può andare all'altra, con opportune trasformazioni.

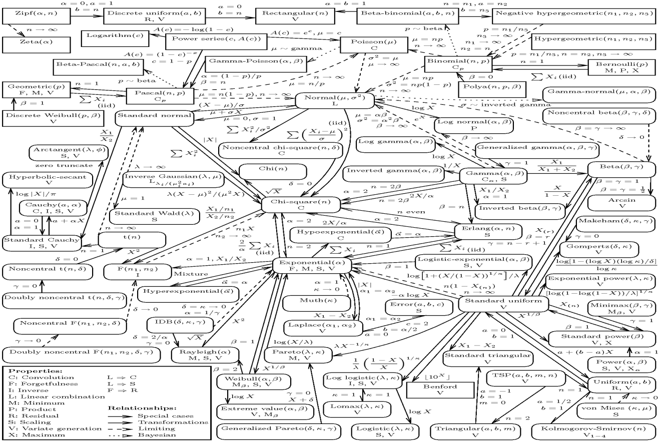

Figure: Trasformazione di variabili

Definiamo l'inversa generalizzata:

Definizione - Inversa Generalizzata o funzione quantilica

Per una funzione non decrescente F su $\mathbb{R}$, l'inversa generalizzata di F, indicata con F^{-1} è definita come

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}$$

nei casi a cui siamo interessati, F è sempre la funzione di ripartizione. L'utilità delle variabili aleatorie uniformi è chiara del teorema

Teorema

Se $U \sim U(0,1)$, allora la variabile aleatoria $F^{-1}(U)$ proviene da F.

Dimostrazione:

Caso continuo:

$P(F^{-1}(U) \leq x) = P(F(F^{-1}(U)) \leq F(x)) = P(U \leq F(x)) = F(x)$ Per l'ultimo passaggio usiamo il fatto che

$$P(U \leq y) = y$$

se $U \sim U(0,1)$, e $y \in [0,1]$.

Simulazione

Caso discreto:

- Assumiamo che $x \in \mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \dots\}$ con $\mathcal{X}_j > \mathcal{X}_{j-1}$
- indichiamo con $\mathcal{U}_j = (a_{j-1}, a_j]$ il set tale che se $u \in \mathcal{U}_j$ allora $F^{-1}(u) = \mathcal{X}_j$

Se $\mathcal{X}_j \leq x < \mathcal{X}_{j+1}$ abbiamo che

$$F(x) = F(\mathcal{X}_j)$$

e

$$a_j - a_{j-1} = F(\mathcal{X}_j) - F(\mathcal{X}_{j-1}) = P(X = \mathcal{X}_j)$$

quindi

$$P(F^{-1}(U) \leq x) = \sum_{h=1}^{j} P(U \in \mathcal{U}_h) = \sum_{h=1}^{j} (a_j - a_{j-1}) = \sum_{h=1}^{j} P(X = \mathcal{X}_h) = F(\mathcal{X}_j)$$

$\square$

Simulazione

Si può dimostrare che l'esempio precedente con le esponenziali è conseguenza di questo teorema. Ricordiamo che l'esempio era

Esempio

Se $U \sim U(0,1)$, allora $X = -\log U \sim Exp(1)$ e $\frac{X}{\lambda} \sim Exp(\lambda)$

i.e. $X = \frac{-\log U}{\lambda} \sim Exp(\lambda)$.
Possiamo verificare che

$$F(x) = u = 1 - e^{-\lambda x} \Rightarrow \log(1-u) = -\lambda x \Rightarrow \frac{-\log(1-u)}{\lambda} = x$$

dove u è un campione da un'uniforme. Sappiamo anche che se $U \sim U(0,1)$ allora $U^* = 1 - U \sim U(0,1)$ e quindi

$$\frac{-\log(u^*)}{\lambda} = x$$

Simulazione

Prima di procedere con altri risultati, mostriamo una proprietà dei campioni casuali multivariati. Per semplicità prendiamo un variabile bivariata (X_1, X_2) (ma si può generalizzare) e sappiamo che la sua densità (o pmf) è

$$f(x_1, x_2) = f(x_1)f(x_2|x_1).$$

Allora, un campione di (X_1, X_2) può essere ottenuto simulando prima $X_1 = x_1$ dalla marginale, e poi X_2 dalla condizionata dato $X_1 = x_1$. Questo si dimostra facilmente se assumiamo che (X_1, X_2) siano discrete, altrimenti, con continue possiamo prendere un intorno delle variabili.

Abbiamo anche che

$$f(x_1, x_2) = f(x_1)f(x_2|x_1) = f(x_2)f(x_1|x_2) = f(x_1, x_2)$$

Questo mi dice che il campione (x_2) può essere visto sia come un campione dalla condizionata, che dalla marginale. Quindi, se vogliamo un set n di campioni dalla marginale di X_2, possiamo simulare dalla congiunta e tenere solo i campioni di X_2. Visto che

$$f(x_2) = \int f(x_1, x_2)d\mu(x_1) = \int f(x_2|x_1)f(x_1)d\mu(x_1)$$

questa procedura può essere anche vista come una **marginalizzazione.**

Metodi Accept-Reject

Algoritmi Accept-Reject

Teorema

Simulare

$$X \sim f(x)$$

dove X è definita in un dominio arbitrario, è equivalente a simulare

$$(X, U) \sim U\{(x, u) : 0 < u < f(x)\}$$

i.e, (X, U) è uniforme in $\{(x, u) : 0 < u < f(x)\}$

Dimostrazione:

Prima di tutto notiamo che

i) il dominio $\{(x, u) : 0 < u < f(x)\}$ ha aria 1;

ii) siccome $f(x, u)$ è uniforme il suo valore è costante, i.e., $f(x, u) = c \, \forall (x, u) \in \{(x, u) : 0 < u < f(x)\}$,

iii) visto che $f(x, u)$ è una densità, per integrare a 1 dobbiamo avere $c = 1$.

La dimostrazione discende dal fatto che

$$f(x) = \int_0^{f(x)} 1 d\lambda(u) = \int_0^{f(x)} f(x, u) d\lambda(u) = \int_0^{f(x)} f(u|x) f(x) d\lambda(u)$$

L'ultimo passaggio è una marginalizzazione rispetto a u di $f(x, u)$.

$\square$

La variabile U è un caso particolare di variabile **ausiliaria** o **latente**.

Potremmo quindi simulare prima X e poi $U|X$, ma questo non ci da nessun vantaggio. Il vantaggio nasce quando non siamo in grado di simulare da X, ma siamo in grado di simulare dall'uniforme (X, U), il che non è sempre facile.
Possiamo usare il teorema e una sua estensione per riuscire a simulare X in casi più generali.

Accept-Reject sampling

Per semplicità ipotizziamo che

$$\int_a^b f_X(x) d\lambda(x) = 1$$

e che $f_X(x)$ sia limitata da m, i.e., sup $f_X(x) \leq m$.
Ipotizziamo di simulare un punto in una "scatola" di dimensioni $[a, b] \times [0, m]$, simulando

- $Y \sim U(a, b)$ (possibile solo se X è limitato)
- $U|Y = y \sim U(0, m)$ (il condizionamento a Y=y è superfluo)

Notate che

- Y e U sono in realtà indipendenti;
- $f(y, u) = \frac{1}{(b-a)m}$

L'idea è di accettare la coppia (y, u) iff $0 < u < f_X(y)$ è soddisfatta.

Accept-Reject sampling

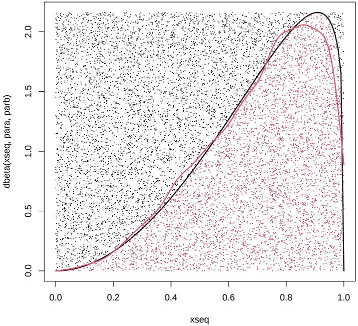

Figure: Simulazione di un $B(3, 1.2)$

Code: Codice della figura

```
para = 3 # parametri
parb = 1.2
# calcoliamo il massimo
moda =  (para-1)/(para+parb-2)
m = dbeta(moda,para,parb)

## campioni U e Y
n = 10000   # numero di campioni
Y = runif(n,0,1)
U = runif(n,0,m)
X = Y[U < dbeta(Y,para,parb)]
U_X = U[U < dbeta(Y,para,parb)]
# plot densita'
xseq = seq(0,1,by=0.01)
plot(
  xseq, dbeta(xseq,para,parb), ylim=c(0,m),
  type="l", lwd=2,
)
points(Y,U, pch=20, cex = 0.1)
points(X,U_X, pch=20, cex = 0.1, col=2)
lines(density(X, from=0, to = 1), col=2, lwd=2)
```

Accept-Reject sampling

Possiamo dimostrare che le y che teniamo sono un campione da X. Indichiamo con x i campioni y che soddisfano $0 < u < f_X(y)$, abbiamo che

$$P(X \leq x) = P(Y \leq x | U < f_X(Y)) = \frac{\int_a^x \int_0^{f_X(y)} (b-a)^{-1} m^{-1} du \, d\lambda(y)}{\int_a^b \int_0^{f_X(y)} (b-a)^{-1} m^{-1} du \, d\lambda(y)} =$$

$$\frac{\int_a^x \int_0^{f_X(y)} du \, d\lambda(y)}{\int_a^b \int_0^{f_X(y)} du \, d\lambda(y)},$$

che è uguale a

$$\int_a^x f_X(y) d\lambda(y)$$

visto che il denominatore

$$\int_a^b \int_0^{f_X(y)} d\lambda(u) \, d\lambda(y) = \int_a^b f_X(y) d\lambda(y) = 1$$

Accept-Reject sampling

Notate come abbiamo un campione da $f_X()$ senza mai simulare da questa distribuzioni, ma solo da uniformi (Importante! Dobbiamo solo essere in grado di calcolare $f_X()$, ma non di simulare da essa.).

Possiamo anche calcolare la probabilità che un campione sia accettato

$$P(\text{Acc.}) = P(U \leq f_X(Y)) \equiv P(U \leq f_X(Y), Y \in [a,b]) =$$

$$\int_a^b \int_0^{f_X(y)} \frac{1}{(b-a)m} d\lambda(u)\, d\lambda(y) = \frac{\int_a^b f_X(y) d\lambda(y)}{(b-a)m} = \frac{1}{(b-a)m}$$

che può diventare molto piccola molto velocemente. Nell'esempio precedente la probabilità di accettazione è 0.46.

Accept-Reject sampling

Possiamo estende l'esempio precedente per ottenere un qualcosa più efficiente.
Ipotizziamo che Y ha densità $g(y)$, e che esista un m tale per cui

- $1 < M < \infty$
- $f_X(y) \leq Mg(y) = m(y)$

In questo caso il dominio di X e Y ($\mathcal{X}$) può essere anche $\mathbb{R}$.
Supponiamo inoltre che simulare Y sia facile e possibile, e che sia possibile calcolare $g(y)$ per ogni valore di y. Vogliamo simulare nello spazio

$$\ell = \{(y, u) : 0 < u < m(y)\}$$

nel seguente modo

- Simulare $Y \sim G$ (La sua distribuzione);
- Simulare $U^*|Y = y \sim U(0, g(y))$
- Definire $u = u^* M$, che è una campione da $U|Y = y \sim (0, m(y))$ (Possiamo anche campionare direttamente $U|Y = y \sim (0, m(y))$ senza passare per U^*.)

Accept-Reject sampling

Se accettiamo il campione solo se $u < f_X(y)$, e indichiamo con x solo i campioni accettati, le x sono da F_x visto che

$$P(X \in \mathcal{A}) = P(Y \in \mathcal{A}|U < f_X(Y)) = \frac{\int_{\mathcal{A}} \int_0^{f_X(y)} \frac{g(y)}{Mg(y)} du\, d\lambda(y)}{\int_{\mathcal{X}} \int_0^{f_X(y)} \frac{g(y)}{Mg(y)} du\, d\lambda(y)} = \int_{\mathcal{A}} f_X(y) d\lambda(y)$$

dove

$$f(y,u) = f(y)f(u|y) = g(y)\frac{1}{Mg(y)}$$

La probabilità di accettare è

$$P(\text{Acc.}) = P(U < f_X(Y)) = \int_{\mathcal{X}} \int_0^{f_X(y)} \frac{g(y)}{Mg(y)} du\, d\lambda(y) = \frac{1}{M} \int_{\mathcal{X}} f_X(y) d\lambda(y) = \frac{1}{M}$$

quindi più $\lambda(y)$ è vicina a $f_X()$ e più accetto, con il caso limite di probabilità 1 se $g(y) \equiv f_X(y)$ e $m = 1$. Rispetto al caso precedente non abbiamo bisogno di restrizioni sul dominio.

Accept-Reject sampling

Formalizziamo il tutto in un corollario

Corollario

Assumiamo che $X \sim f(x)$ e definiamo $g(x)$ come una funzione di densità che soddisfa

$$f(x) \leq Mg(x)$$

per qualche costante $M \geq 1$, allora per simulare una X da f basta generare

$$Y \sim g \qquad U|Y = y \sim U(0, Mg(y))$$

finchè $0 < u < f(y)$

Kernel e costante di normalizzazione

C'è un aspetto di questi approcci che non è immediatamente visibile. Ogni densià di un vettore di variabili aleatorie $\mathbf{x}$, dipendente da dai parametri θ, si può dividere in due parti

$$f(\mathbf{x}|\theta) = \frac{k(\mathbf{x}|\theta)}{C(\theta)}$$

dove $k(\mathbf{x}|\theta)$, chiamato kernel, dipende da i dati e dai parametri, e una costante di normalizzazione $C(\theta)$ che non dipende dai dati. Una distribuzione è totalmente descritta dal solo kernel visto che

$$1 = \int_{\mathcal{X}} f(\mathbf{x}|\theta) d\lambda(\mathbf{x}) = \int_{\mathcal{X}} \frac{k(\mathbf{x}|\theta)}{C(\theta)} d\lambda(\mathbf{x}) = \frac{1}{C(\theta)} \int_{\mathcal{X}} k(\mathbf{x}|\theta) d\lambda(\mathbf{x})$$

e quindi deve essere

$$\int_{\mathcal{X}} k(\mathbf{x}|\theta) d\lambda(\mathbf{x}) = C(\theta)$$

e

$$f(\mathbf{x}|\theta) \propto k(\mathbf{x}|\theta)$$

o, in altre parole,

$$f(\mathbf{x}|\theta) = \frac{k(\mathbf{x}|\theta)}{C(\theta)} = \frac{k(\mathbf{x}|\theta)}{\int_{\mathcal{X}} k(\mathbf{x}|\theta) d\lambda(\mathbf{x})}$$

Esempio1:

Nel caso in cui $X \sim N(\mu, \sigma^2)$, abbiamo che

$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-0.5} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-0.5} \exp\left(-\frac{x^2+\mu^2-2x\mu}{2\sigma^2}\right)$$

dove

$$k(x|\mu, \sigma^2) = \exp\left(-\frac{x^2-2x\mu}{2\sigma^2}\right)$$

$$C(\mu, \sigma^2) = \frac{1}{(2\pi)^{-0.5} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)}$$

e quindi

$$\int_{\mathbb{R}} \exp\left(-\frac{x^2-2x\mu}{2\sigma^2}\right) d\lambda(x) = (2\pi)^{-0.5} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

Esempio2:

Nel caso in cui $X \sim B(a, b)$, abbiamo che

$$f(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}$$

Accept-Reject sampling

dove

$$k(x|a,b) = x^{a-1}(1-x)^{b-1}$$

$$C(a,b) = B(a,b)$$

e quindi

$$\int_0^1 x^{a-1}(1-x)^{b-1} d\lambda(x) = B(a,b)$$

Accept-Reject sampling

Torniamo alla simulazione e vediamo che la richiesta $f(x) \leq Mg(x)$ può anche essere scritta come

$$f(x) \leq Mg(x) \rightarrow \frac{k(\mathbf{x}|\theta)}{C(\theta)} \leq Mg(x) \rightarrow k(\mathbf{x}|\theta) \leq C(\theta)Mg(x) = M^{*}g(x)$$

quindi non abbiamo bisogno di conoscere la costante di normalizzazione per usare il metodo. Per esempio, assumiamo che

$$f(x) \propto \exp(-x^2/2)(\sin^2(6x) + 3\cos^2(x)\sin^2(4x) + 1) = k(\mathbf{x})$$

con $x \in [-\pi, \pi)$. Possiamo vedere che $k(\mathbf{x}) < 12g(y)$, dove $g(y)$ è la densità di una normal standard. Possiamo quindi richiedere che il kernel sia dominato da una $M^{*}g(x)$

Accept-Reject sampling

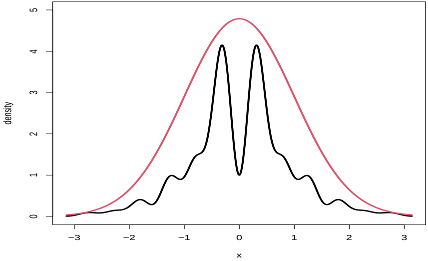

Figure: Kernel e densità normale ($\times M$)

Accept-Reject sampling

Code: Codice della figura

```
xseq = seq(-pi, pi, by=0.01)
dens = exp( -xseq^2/2 )*
      (sin(6*xseq)^2+3*cos(xseq)^2*sin(4*xseq)^2+1)
plot(
    xseq, dens, type="l", ylim=c(0,5),
    lwd=3, xlab="x", ylab="density"
)
lines(xseq, 12*dnorm(xseq), col=2, lwd=3)
```

Accept-Reject sampling

Un altro modo per scrivere il corollario precedente, che è spesso usato, è l'algoritmo **Accept-Reject**

Algoritmo Accept-Reject

Scopo: Generare un campione x da $f_x()$

1: **repeat**
2: Generare $Y \sim G$, e $U \sim U(0,1)$
3: **until** $u < \frac{f(y)}{Mg(y)}$
4: $x \leftarrow y$

Soluzione:

Se simuliamo $U \sim U(0,1)$ e accettiamo se $u < \frac{f(y)}{Mg(y)}$, allora
$U^* = U \times Mg(y) \sim U(0, Mg(y))$ e $u \times Mg(y) < \frac{f(y)}{Mg(y)} Mg(y) \Rightarrow u^* < f(y)$ $\square$

Dall'algoritmo è chiaro come per un intorno di x, la probabilità di accettare dipende dalla distanza tra $f(x)$ e $Mg(x)$, ci sono quindi punti più facili da campionare e punti più complicati.

Accept-Reject sampling

Fate attenzione che

- simulare $U \sim U(0,1)$ e accettare iff $u < \frac{f(y)}{Mg(y)}$;
- simulare $U \sim U(0, Mg(y))$ e accettare iff $u < f(y)$

sono la stessa cosa.