

ANALIZĂ BAYESIANĂ

Fundamente Teoretice și Exemple

Constantin Manuel Bosancianu

Școala Doctorală de Științe Politice, Politici Publice și Relații Internaționale
Central European University

Metode Aplicate de Cercetare Socială
Cluj-Napoca, România: 24 Iulie, 2014

Întrebări și teme

- ▶ Ce este nou în analiza Bayesiană?
- ▶ Concepte de bază
- ▶ Principiile analizei Bayesiene
- ▶ Metode moderne: algoritmul Gibbs, algoritmul Metropolis–Hastings
- ▶ Exemple, exemple, exemple...

Scurt istoric

Ramuri statistice:

1. Paradigma frecventistă: J. Neyman (CIs), E. Pearson (NHST), A. Wald (experimente);
2. Paradigma verosimilității (*likelihood*): R. A. Fisher – inferență bazată pe un eșantion unic;
3. Paradigma Bayesiană: T. Bayes, P.-S. Laplace, B. de Finetti.

Scurt istoric

Paradigma Bayesiană are cele mai vechi rădăcini – secolul XVIII (T. Bayes și P.-S. Laplace).

Nu a devenit fezabilă pentru inferențe statistice în situații practice decât recent.

Motive pentru adoptarea târzie

Ostilitate a practicienilor celorlalte ramuri statistice, e.g. R. A. Fisher.

Dificultăți în a accepta caracterul ei *subiectiv*.¹

Cerințele computaționale foarte ridicate în majoritatea aplicațiilor practice (algoritmi Gibbs sau Metropolis–Hastings).

¹Mai multe despre aceasta peste câteva slide-uri.

Exemplele practice

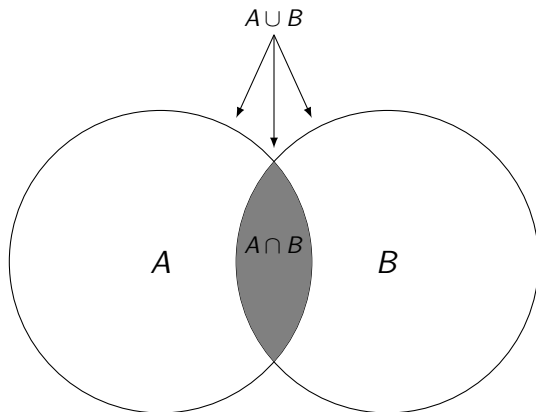
Prezentarea: <https://manuelbosancianu.github.io/workshops/2014-07-Cluj>

Am evitat să ofer și codul R, deoarece scopurile noastre sunt mai teoretice de data aceasta.

Pe lângă R, mai este nevoie de un program care se ocupă de analiza propriu-zisă (**JAGS**, BUGS, STAN).

Probabilități

Probabilitate



Probabilitate

$$p(A \cup B) = p(A) + p(B) - p(A, B).$$

Formula poate fi citită ca “probabilitatea ca A sau B să se întâmple”.

Probabilitate

$p(A \cap B) = p(A, B) = p(A) * p(B)$, **dacă și numai dacă** A și B sunt independente.

Formula poate fi citită ca “probabilitatea ca A și B să se întâmple”.

Dacă A și B nu sunt independente, $p(A \cap B) = p(A|B) * p(B)$

Probabilitate condițională

$$p(A|B) = \frac{p(A, B)}{p(B)} \quad (1)$$

Poate fi citită ca “probabilitatea ca A să se întâmple ținând cont că B deja s-a întâmplat”

Probabilitate condițională

	Interes politic				
	1	2	3	4	Total rând
Ciclu secundar	19	37	39	14	4%
Liceu	70	375	387	155	34%
Facultate	49	397	617	224	44%
MA și peste	5	95	274	180	19%
Total coloană	5%	31%	45%	20%	$\approx 100\%$

Distribuțiile marginale sunt marginile, iar distribuțiile condiționale sunt coloanele sau rândurile (condiționate de valoarea celeilalte variabile).

Legea

$$\begin{cases} p(A|B) = \frac{p(A,B)}{p(B)} \\ p(B|A) = \frac{p(B,A)}{p(A)} \end{cases} \quad (2)$$

Înmulțirea cu numitorul transformă ecuația 2 în

$$\begin{cases} p(A, B) = p(A|B) * p(B) \\ p(B, A) = p(B|A) * p(A) \end{cases} \quad (3)$$

Legea

Deoarece $p(A, B) = p(B, A)$

$$p(A|B) * p(B) = p(B|A) * p(A) \Rightarrow p(A|B) = \frac{p(B|A) * p(A)}{p(B)} \quad (4)$$

Legea lui Bayes!

Semnificația

În termeni simpli, oferă posibilitatea de a inversa probabilități condiționale.

Posibilități imense apar dacă înlocuim B cu $date$, iar A cu un parametru θ .

$$\underbrace{p(\theta|date)}_{\text{posterioară}} = \frac{\overbrace{p(date|\theta)}^{\text{verisimilitate}} * \overbrace{p(\theta)}^{\text{anterioară}}}{\underbrace{p(date)}_{=1}} \quad (5)$$

Semnificația

Combinăm probabilitatea de a observa datele din eșantion condiționată de modelul nostru statistic, $p(\text{date}|\theta)$, cu gradul nostru de încredere în acel model, $p(\theta)$, pentru a obține probabilitatea modelului condiționată de observarea datelor din eșantion.

Teste pentru boală

Să presupunem că un virus afectează 1% din populație.

Avem un test, cu 95% rată de succes: categorizează corect 95% din cazurile de infecție, și 95% din cazurile lipsite de infecție.

Dacă testul îmi spune că sunt infectat, care e probabilitatea să fiu infectat?

Teste pentru boală

Probabilitatea de depistare în cazul infecției: $p(D|I) = 95\%$.

Probabilitatea infecției: $p(I) = 1\%$.

Probabilitatea depistării:

$$\begin{aligned} p(D) &= p(D|I) * p(I) + [1 - p(D| \sim I)] * p(\sim I) = \\ &= 0.95 * 0.01 + 0.05 * 0.99 = 0.059 \end{aligned}$$

Teste pentru boală

Punând toate cele 3 elemente împreună, avem:

$$\begin{aligned} p(I|D) &= \frac{p(D|I) * p(I)}{p(D)} = \\ &= \frac{0.95 * 0.01}{0.059} \approx 0.161 \end{aligned}$$

Legea lui Bayes poate fi aplicată încă o dată, și încă o dată...

Întrebări?

Analiza Bayesiană

Interpretarea probabilității²

NON-BAYESIENI

Proporția de “succese” dintr-o serie lungă (infinită) de probe (experimente) derulate în condiții identice.

BAYESIENI

Gradul de încredere (**subiectivă!**) a cercetătorului în valoarea unui parametru înainte de a observa datele.

²Am folosit textul lui Jeff Gill (2008) pentru această secțiune.

Fix vs. aleatoriu

NON-BAYESIENI

Eșantionul este aleatoriu, însă parametrii care generează datele sunt ficși (există în natură).

BAYESIENI

Eșantionul este fix (observat de noi), însă parametrii care au generat datele sunt aleatorii (fiecare este descris printr-o distribuție a cărei varianță denotă gradul nostru de încredere în valoarea parametrului).

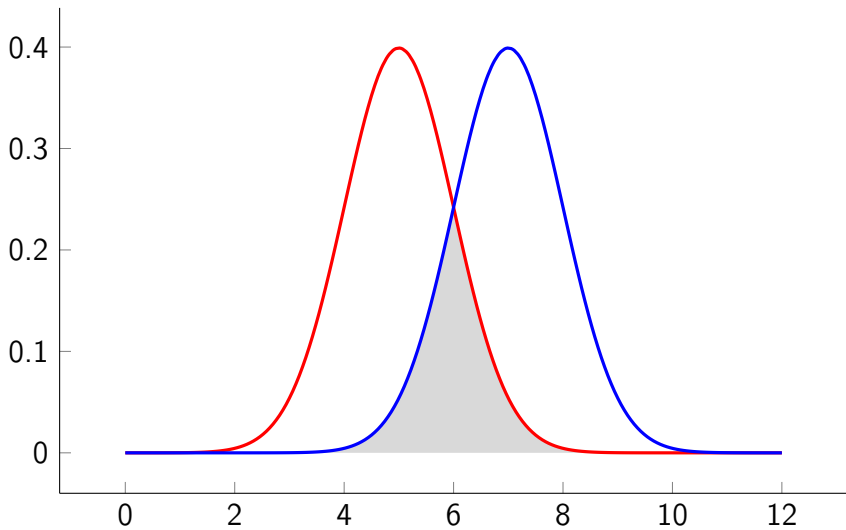
Sumarizarea rezultatelor

NON-BAYESIENI

Coeficienți și erori standard. Intervale de încredere care indică faptul că în 19/20 de încercări intervalul include parametrul.

BAYESIENI

Descrierea distribuției posterioare: medie, IQR etc. Indică faptul că suntem 95% siguri că parametrul se găsește în intervalul HPD (*highest posterior density*).



Ce concluzie ar lua un non-Bayesian și un Bayesian?

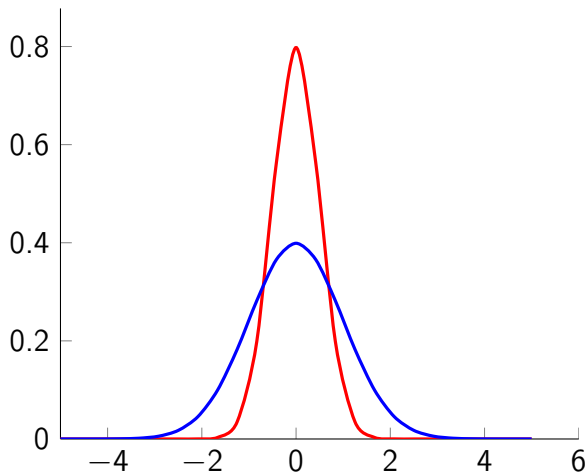
Distribuție anterioară

Am trecut de la probabilități clare (ca în exemplul cu infecția), la probabilități exprimate în distribuții.

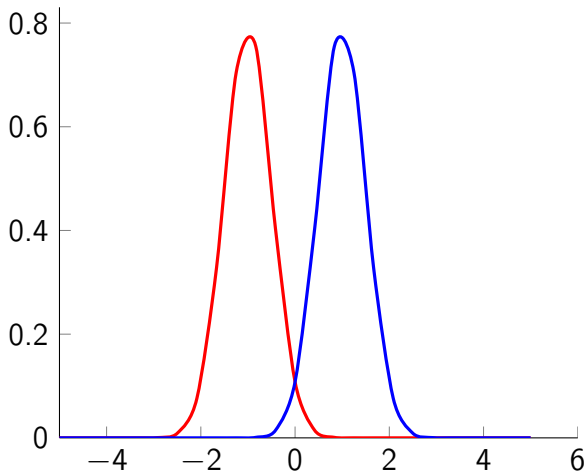
Din punct de vedere teoretic, saltul nu e chiar atât de mare.

O distribuție exprimă gradul nostru de încredere privind parametrul estimat.

Grad de încredere diferit...



Parametri diferiți



Exemplul cu infecția

$p(I) = 1\%$. Însă, știm că există variabilitate în populație, e.g. cei tineri au o probabilitate mai scăzută de a fi infectați decât cei mai în vârstă.

Putem exprima aceasta ca $p \sim \mathcal{N}(1, 0.25)$.

Un grad mai mare de incertitudine ar putea fi exprimat ca $p \sim \mathcal{N}(1, 0.5)$.

De unde vin distribuțiile anterioare?

Un aspect foarte controversal al Analizei Bayesiene, deoarece aduce subiectivism în procedură.

Pe de altă parte, există suficient subiectivism ascuns în practica paradigmei frecventiste.

Bayesienii trebuie să depună eforturi considerabile pentru a arăta că rezultatele obținute nu sunt datorate alegerii “foarte atente” a distribuției anterioare.

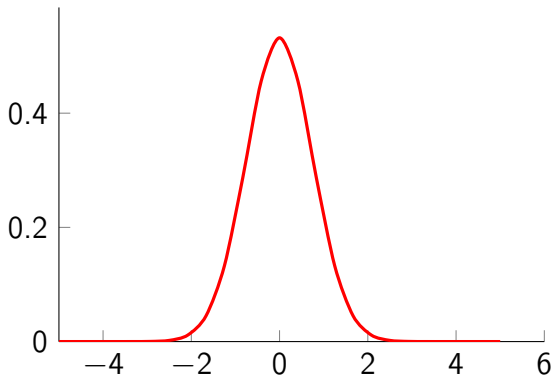
De unde vin distribuțiile anterioare?

1. Experiența cercetătorului;
2. Experți externi (procedura se numește *elicitare*);
3. Analize existente.

Tipuri de anterioare

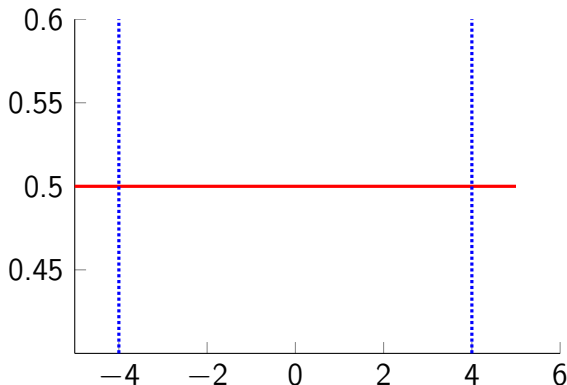
1. Informative;
2. Non-informative;
3. Conjugate (*conjugate*).

Anterioare informative



Exprimă cunoștințele noastre despre valoarea parametrului și gradul nostru de încredere în aceste cunoștințe

Anterioare non-informative



Exprimă lipsa oricărei cunoștințe despre parametru, ceea ce transformă analiza într-una bazată integral pe ML

Anterioare conjugate

În cele mai multe cazuri, distribuția posterioară e foarte dificil de obținut analitic.

Caracterul conjugat e o proprietate a distribuției anterioare și a funcției verosimilității (*likelihood function*), care **asigură** că posterioara va aparține aceleiași familii de distribuții ca anterioara.

Anterioare conjugate

Foarte importante din punct de vedere istoric, pentru ca altfel doar un număr minuscule de analize Bayesiene ar fi putut fi rezolvate.

Recent, cu metodele MCMC (*Markov chain Monte Carlo*), care abordează problema prin simulare iar nu analitic, anterioarele conjugate nu mai sunt atât de importante.

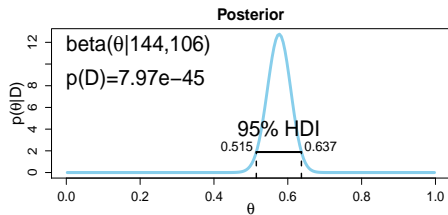
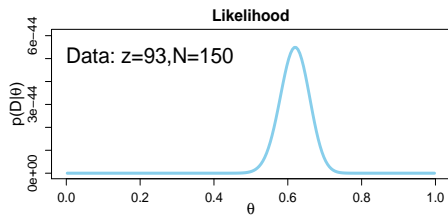
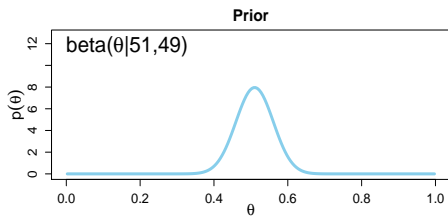
Întrebări?

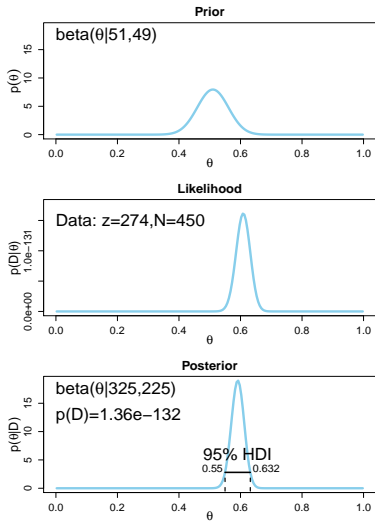
Estimarea unei proporții

Un politician face un sondaj cu 2 săptămâni înainte de alegeri ($N=100$); 51% din respondenți ar vota cu ea.

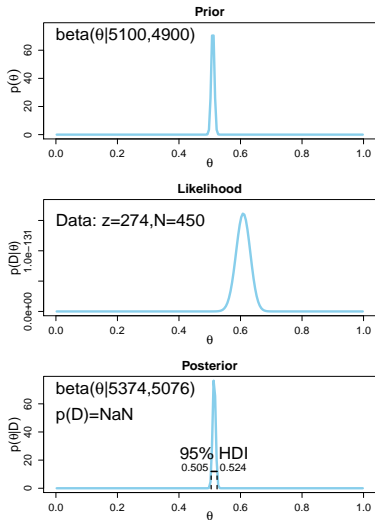
Cu o săptămână înainte, un nou sondaj ($N=150$) arată că 60% din alegători ar vota cu ea.

Ce șanse de câștig are politicianul nostru?





Dar dacă al doilea sondaj e cu $N=450$?



Dar dacă al primul sondaj e cu $N=10000$?

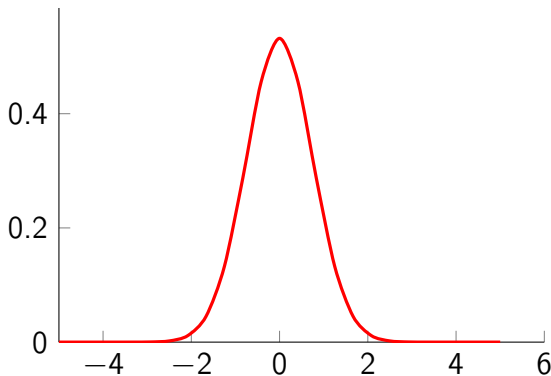
Exemplu

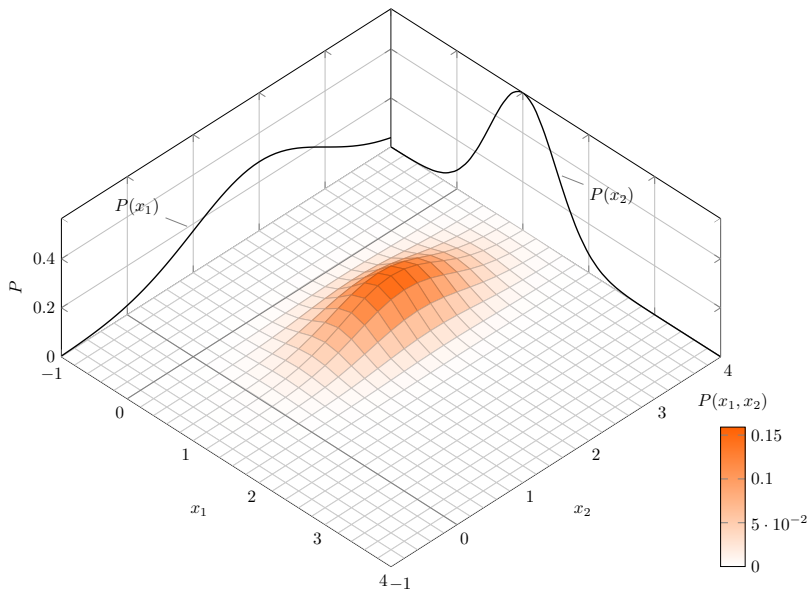
Am putea continua cu exemple de alte distribuții.

Matematica devine puțin mai complexă, însă raționamentul e același.

Iar teorie...

Realitatea e rareori așa





Alte cazuri problematice⁴

Distribuții univariate care nu pot fi descrise printr-un set de parametri.³

În aceste cazuri, nu mai putem obține o soluție analitică pentru a descrie distribuția posterioară.

³Spre exemplu, cei doi parametri care descriu o distribuție Gaussiană specifică: $\mathcal{N}(1, 0.25)$.

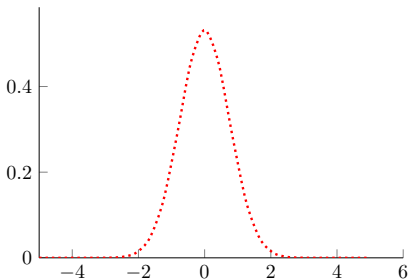
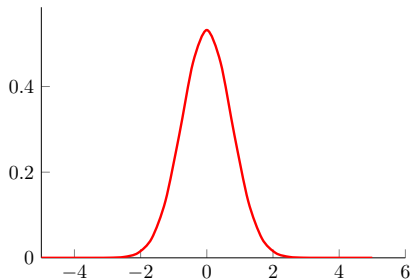
⁴Puteți găsi figura de mai sus la: <http://tex.stackexchange.com/questions/31708/draw-a-bivariate-normal-distribution-in-tikz>.

Metode Monte Carlo

Soluția este de a obține o “image” aproximativă a acestei distribuții, și de a folosi această “copie” pentru a obține parametrii de care suntem interesați.

Acesta este rolul metodelor Monte Carlo (numite astfel deoarece implică procese aleatorii).

Rolul “imaginii”



Algorithmul Gibbs

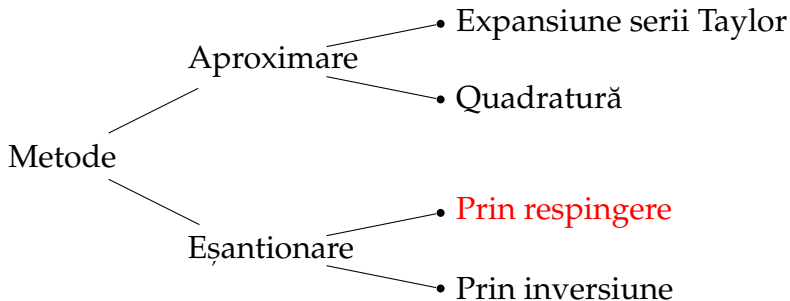
Simplificare

Putem aborda problema încercând să simplificăm: o distribuție multivariată poate fi secționată într-o serie de distribuții de ordin mai mic, e.g. univariate.

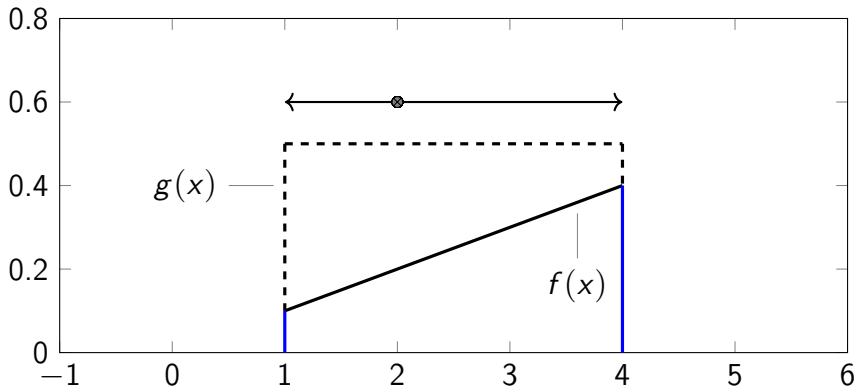
În asta constă esența algoritmului Gibbs.

Însă cum putem analiza aceste distribuții de ordin mai mic dacă nu analitic?

Metode

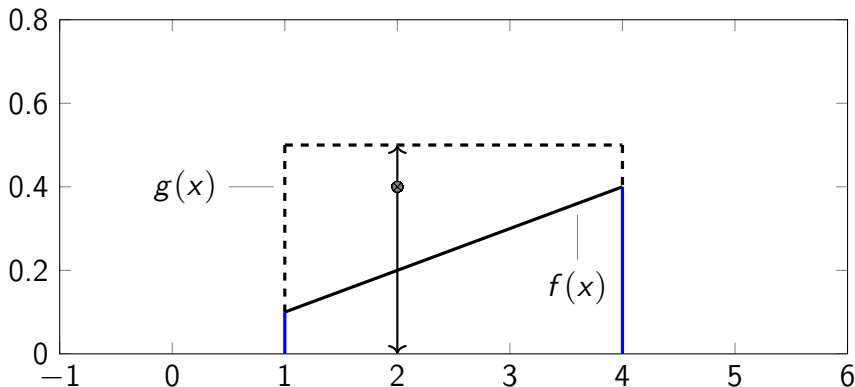


Eșantionare prin respingere



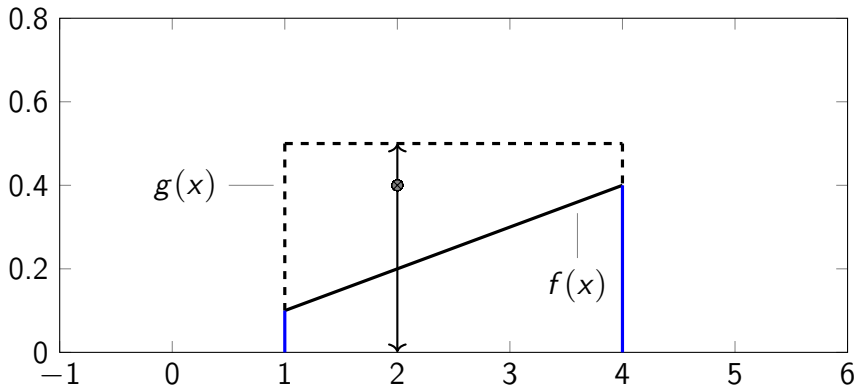
Alegem un punct, z , pe funcția “plic” $g(x)$. Suntem interesați de funcția $f(x)$

Eșantionare prin respingere



Alege aleatoriu $u \in (0, 1)$ și alege punctul $u * g(z)$

Eșantionare prin respingere



Dacă $u * g(z) > f(z)$, respinge și selectează un nou z . Dacă nu, păstrează-l ca aparținând de $f(x)$. În cazul de față, trebuie respins.

Avantaje

Metoda funcționează pentru orice formă a funcției $f(x)$.

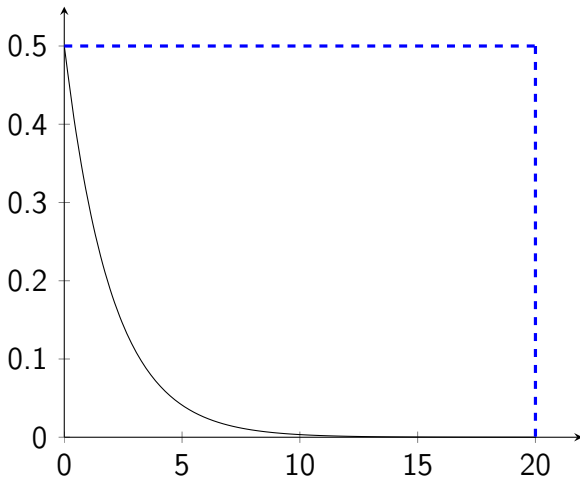
Într-o formă puțin mai complexă funcționează pentru distribuții multivariate.

Limite

Uneori e dificil de a găsi o funcție “plic” potrivită.

Alteori, procesul poate deveni foarte ineficient dacă funcția “plic” e mult mai înaltă decât funcția de care suntem interesați.

Ineficiență



Distribuție gamma dificil de “acoperit” eficient

Cum o folosim?

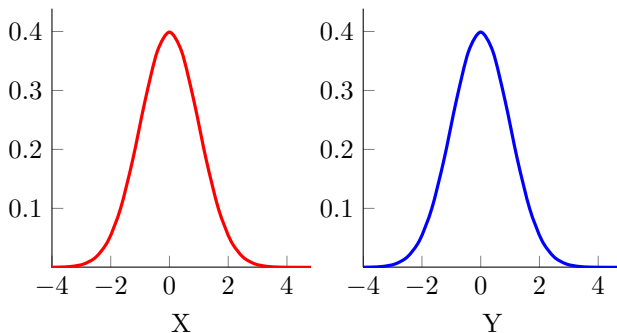
Algoritmul Gibbs poate folosi eşantionarea prin respingere pentru a ajunge la o “imagine” a distribuţiei multivariate de interes.

Să presupunem că avem un vector de parametri, $\theta_1, \theta_2, \dots, \theta_k$, cărora le alocăm nişte valori iniţiale aleatorii, S .

Algoritmul Gibbs

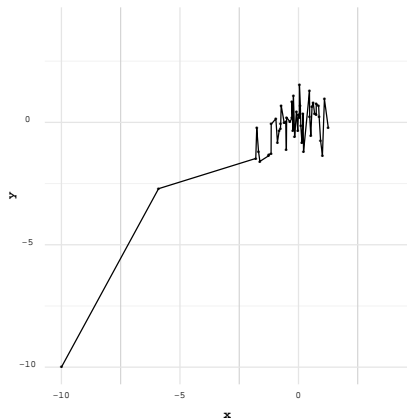
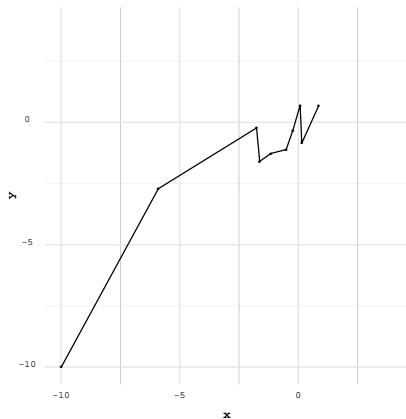
1. $\theta^{j=0} = S$
2. j devine $j + 1$
3. Eșantionăm $(\theta_1^j \mid \theta_2^{j-1}, \theta_3^{j-1}, \dots, \theta_k^{j-1})$
4. Eșantionăm $(\theta_2^j \mid \theta_1^j, \theta_3^{j-1}, \dots, \theta_k^{j-1})$
5. ...
6. Eșantionăm $(\theta_k^j \mid \theta_1^j, \theta_2^j, \dots, \theta_{k-1}^j)$
7. Ne întoarcem la pasul 1.

Exemplu



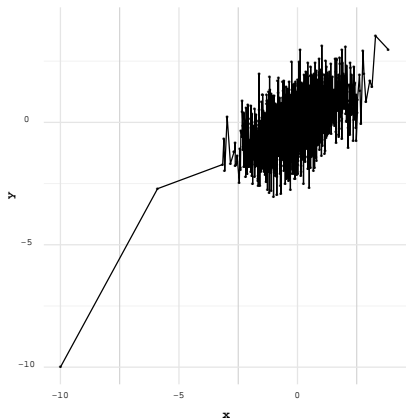
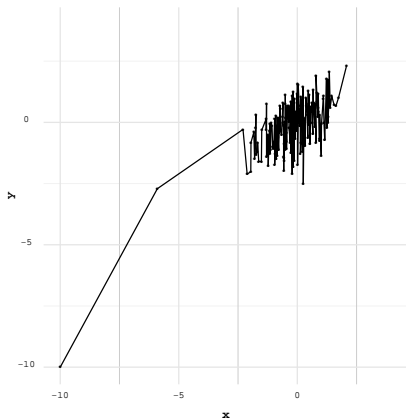
Două variabile distribuite normal, corelate la 0.5

Exemplu



Parcursul algoritmului Gibbs după 10 și după 50 de iterații

Exemplu




Parcursul algoritmului Gibbs după 200 și după 2000 de iterații

Algoritmul Metropolis–Hastings

Probleme cu Gibbs

Uneori, nu putem găsi o funcție “plic”, sau algoritmul e prea ineficient (respinge prea multe valori pentru fiecare valoare acceptată).

Algoritmul Metropolis–Hastings poate fi folosit.⁵

⁵Matematic vorbind, Gibbs este un caz special al Metropolis–Hastings. 

Exemplul politicianului

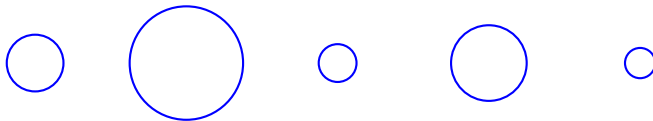
Un politician trebuie să meargă în campanie pe o serie de insule, cu populații diferite.

Pornește de la insula I_x , dar nu poate merge decât pe insulele adiacente, I_{x-1} sau I_{x+1} .

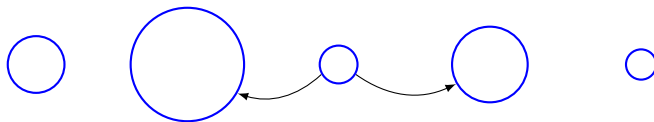
Ea dorește să meargă mai des pe insulele mai populate, pentru a câștiga alegerile.

Ce regulă trebuie adoptată?

Exemplul politicianului

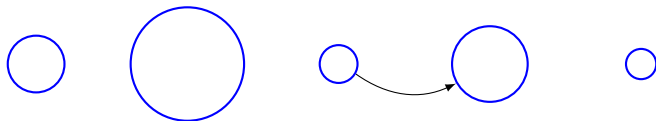


Pasul 1: dreapta sau stânga?

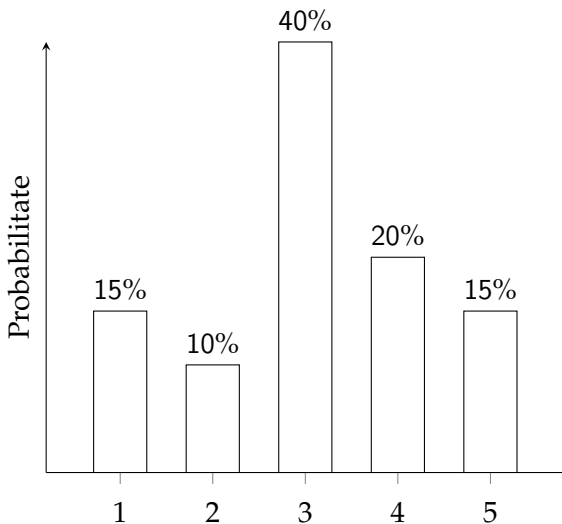


Probabilitatea $p = 0.5$

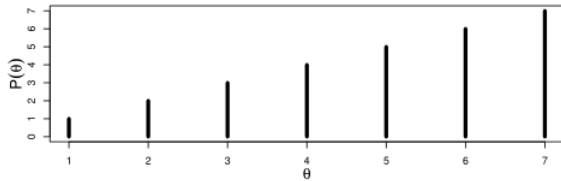
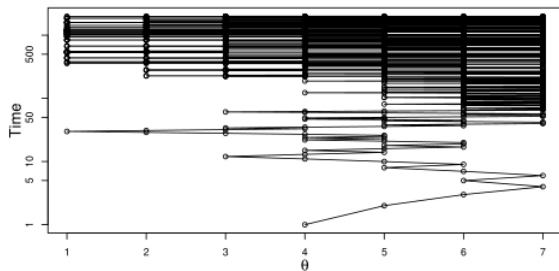
Pasul 2: schimbăm insula sau nu?



Dacă $I_{x+1} \geq I_x$, da cu $p = 1$. Dacă $I_{x+1} < I_x$, da cu $p = Pop_{I_{x+1}} / Pop_{I_x}$.



În loc de insule, ar putea fi o distribuție



Din Kruschke (2010).

Mai formal

Algoritmul Metropolis–Hastings se bazează pe 2 pași:

1. Mai întâi, o decizie **aleatorie** este luată cu privire la direcția de avansare;
2. Apoi, algoritmul avansează spre această direcție cu probabilitatea:

$$\begin{cases} p = 1 & : P(\theta_{propus}) \geq P(\theta_{actual}) \\ p = \frac{P(\theta_{propus})}{P(\theta_{actual})} & : P(\theta_{propus}) < P(\theta_{actual}) \end{cases} \quad (6)$$

“Imaginea”

Algoritmul poartă numele de lanț Markov deoarece are o memorie scurtă: poziția viitoare în distribuție nu este determinată decât de poziția actuală.

Calea parcursă de algoritm reprezintă o imagine suficient de bună a distribuției de care suntem interesați.⁶

⁶Atât timp cât excludem primii pași, care îi dau punctului de pornire o influență prea mare asupra distribuției finale.

Algoritmul

Am prezentat doar un caz special pentru un algoritm mult mai general (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953).

Partea frumoasă a algoritmului este că funcționează și dacă:

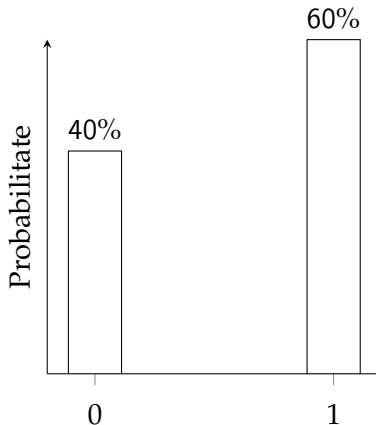
1. Avem o distribuție continuă;
2. Avem mai mult de o dimensiune a distribuției;
3. Salturile dintre poziții sunt mai mari.

Exemplu

Un exemplu foarte simplu de estimare a unei proporții cu ajutorul algoritmului Metropolis–Hastings.

Distribuția anterioară va fi una non-informativă (uniformă).

Datele



O simplă proporție de succese dintr-un eșantion de 20 de încercări

Un fragment important de cod

Partea de cod care specifică distribuția de unde se va face eșantionarea.

```
targetRelProb = function(theta, data) {  
  targetRelProb = likelihood(theta, data) * prior(theta)  
  return(targetRelProb)  
}
```

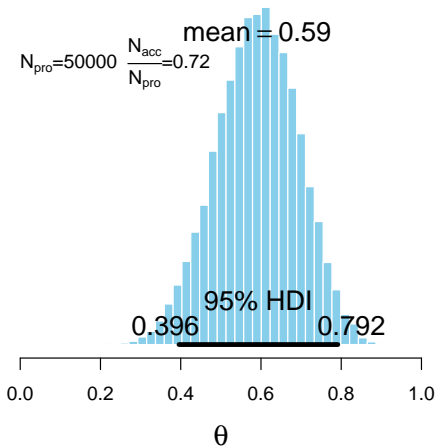
Codul funcționează datorită Legii lui Bayes:

Posterior \propto *Likelihood* \times *Prior*.

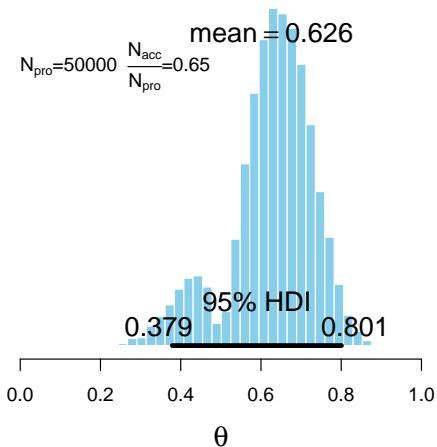
Un fragment important de cod

Partea de cod care decide dacă trecem la o nouă locație în distribuție sau stăm pe loc.

```
probAccept = min( 1,  
targetRelProb( currentPosition + proposedJump , myData )  
/ targetRelProb( currentPosition , myData ) )  
if ( runif(1) < probAccept ) {  
trajectory[ t+1 ] = currentPosition + proposedJump }  
} else {  
trajectory[ t+1 ] = currentPosition  
}
```



Nu suntem foarte departe de proporția inițială



Rezultatul cu o anterioară bimodală

Analiza Bayesiană

Analiza Bayesiană este mult mai flexibilă decât metodele de analiză mai “tradiționale”.

Însă, nu implică doar software sau cod nou, ci o nouă filosofie statistică.

Rezultatele nu mai sunt obținute analitic, ci prin eșantionare, ceea ce implică cerințe computaționale MULT mai înalte.

Muncă mai multă + Putere de calcul \Rightarrow inferențe mai complexe.

Mulțumesc pentru atenție!

Referințe

- Gill, J. (2008). *Bayesian Methods: A Social and Behavioral Sciences Approach* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Burlington, MA: Academic Press.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.