



UNAM
POSGRADO



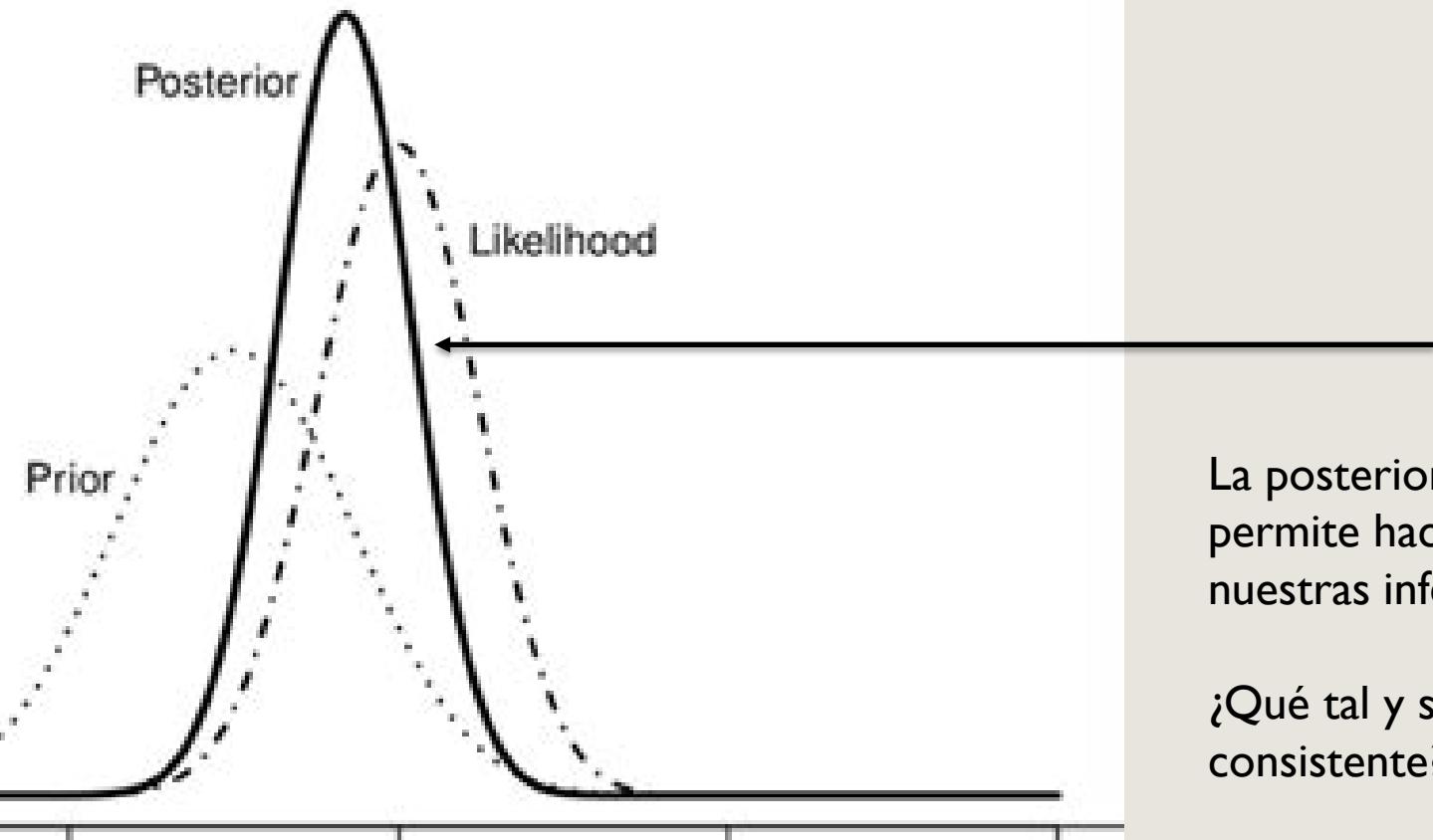
Programa
Universitario
de Estudios
del Desarrollo
UNAM

Algoritmos de muestreo y Monte Carlo

Dr. Héctor Nájera
Dr. Curtis Huffman



¿La distribución posterior es confiable?



La posterior nos
permite hacer
nuestras inferencias

¿Qué tal y si no es
consistente?

Propiedades

- ¡OK! Si me interesa entonces encontrar la masa y la densidad de una distribución (posterior), eso significa que tengo que calcular el área bajo la curva
- La integral

Probability Density Function

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

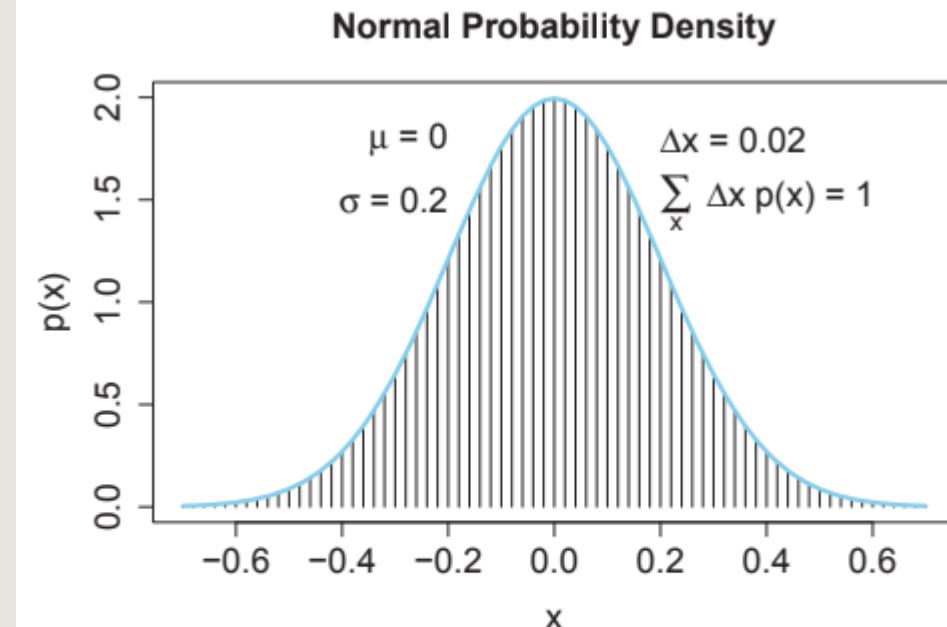


Figure 4.4: A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.



¿Cuál es el problema?

Table 4.1: Proportions of combinations of hair color and eye color. Some rows or columns may not sum exactly to their displayed marginals because of rounding error from the original data. Data adapted from Snee (1974). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.

Eye Color	Hair Color				Marginal (Eye Color)
	Black	Brunette	Red	Blond	
Brown	.11	.20	.04	.01	.37
Blue	.03	.14	.03	.16	.36
Hazel	.03	.09	.02	.02	.16
Green	.01	.05	.02	.03	.11
Marginal (Hair Color)	.18	.48	.12	.21	1.0

Probabilidad condicional

Reubicación de acuerdo al color de cabello

Table 4.2: Example of conditional probability. Of the blue-eyed people in Table 4.1, what proportion have hair color h ? Each cell shows $p(h|\text{blue}) = p(\text{blue}, h)/p(\text{blue})$ rounded to two decimal points. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Eye Color	Hair Color				Marginal (Eye Color)
	Black	Brunette	Red	Blond	
Blue	.03/.36 = .08	.14/.36 = .39	.03/.36 = .08	.16/.36 = .45	.36/.36 = 1.0

Esto es inferencia bayesiana:

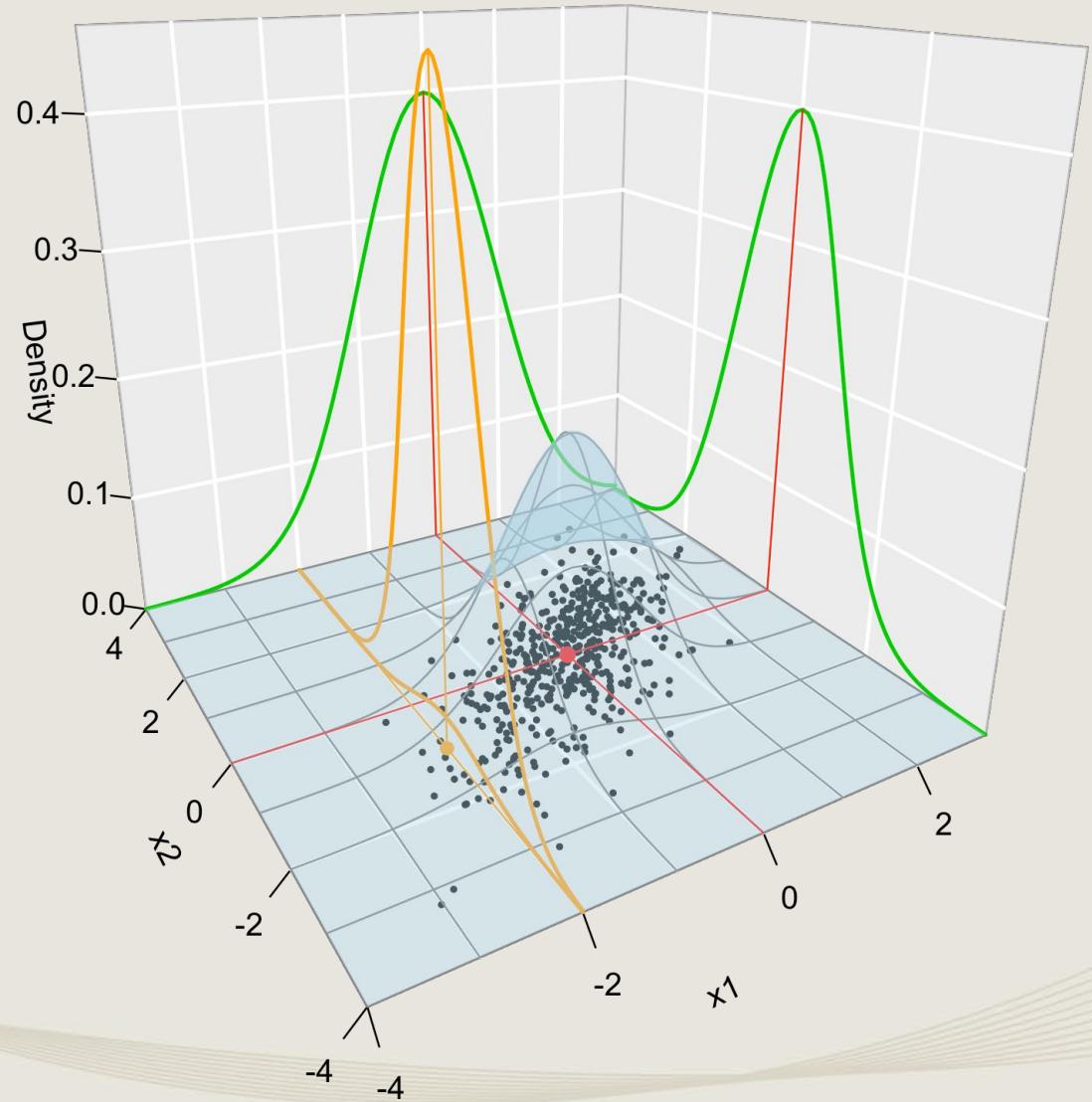
$$P(\beta|D)$$

Probabilidad condicional

Cuando tenemos un modelo de dos variables la inferencia descansa en la distribución conjunta de las posteriores de los parámetros de interés.

El espacio de solución deja de ser obvio

Recuerden que para hacer inferencia (HDI - ROPE) necesitamos conocer las áreas bajo la curva





¿Problema?

- Hay verosimilitudes sencillas

$$h_i \sim \text{Normal}(\mu, \sigma) \rightarrow \text{Likelihood}$$

$$\mu \sim \text{Normal}(178, 20) \rightarrow \text{Prior}$$

$$\sigma \sim \text{Uniform}(0, 50) \rightarrow \text{Prior}$$

- Monstruo pasado por teorema de Bayes

$$Pr(\mu, \sigma | h) \frac{\prod_i \text{Normal}(h_i | \mu, \sigma) \text{Normal}(\mu | 178, 20) \text{Uniform}(\sigma | 0, 50)}{\int \int \prod_i \text{Normal}(h_i | \mu, \sigma) \text{Normal}(\mu | 178, 20) \text{Uniform}(\sigma | 0, 50) d\mu d\sigma}$$



Para
regularizar/normalizar
hay que integrar esto!

PROBLEMA: ALTAS DIMENSIONES

- 6 parámetros
- El espacio de los parámetros es de seis dimensiones –distribución conjunta de todos los parámetros-
- 1,000,000,000,000,000 de valores de los parámetros

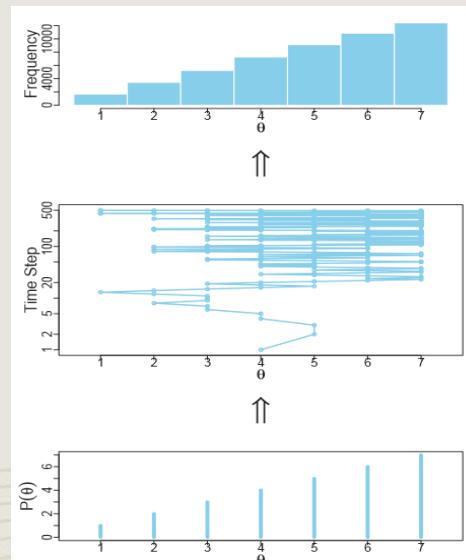
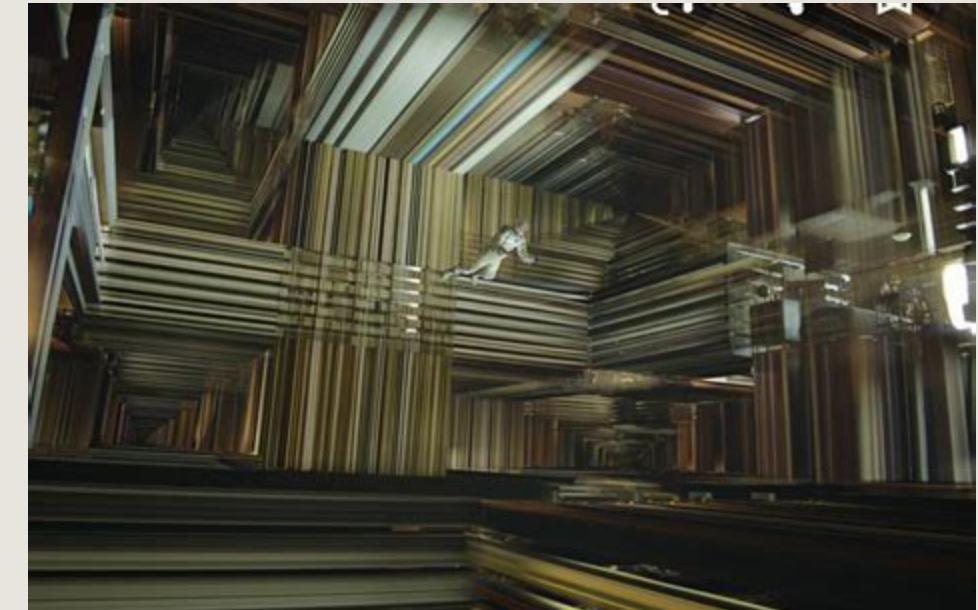


Figure 7.2: Illustration of a simple Metropolis algorithm. The bottom panel shows the values of the target distribution. The middle panel shows one random walk, at each time step proposing to move either one unit right or one unit left, and accepting the proposed move according the heuristic described in the main text. The top panel shows the frequency distribution of the positions in the walk. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.





En inferencia bayesiana no se optimiza... se muestrea!

3.2. SAMPLING TO SUMMARIZE

53

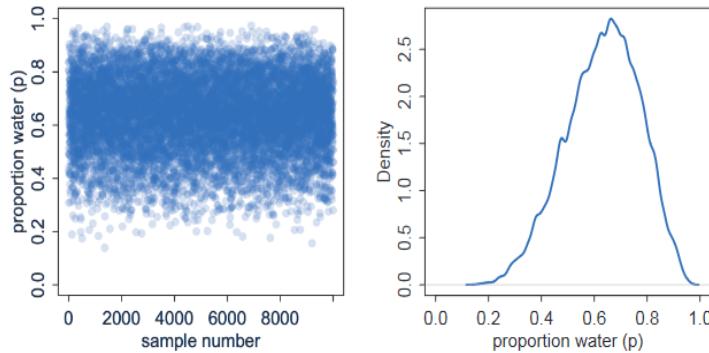
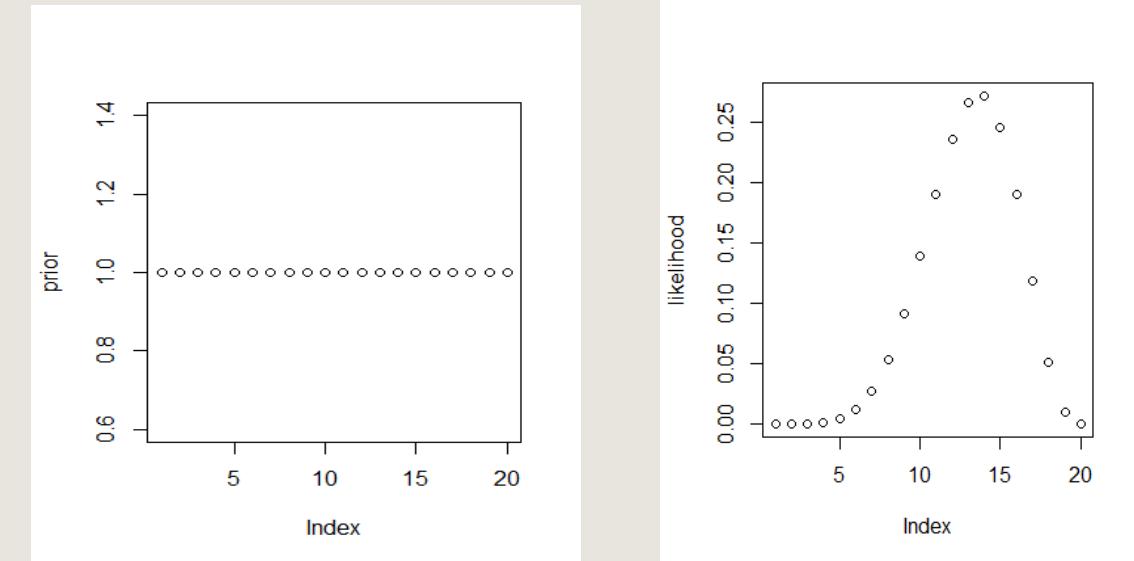
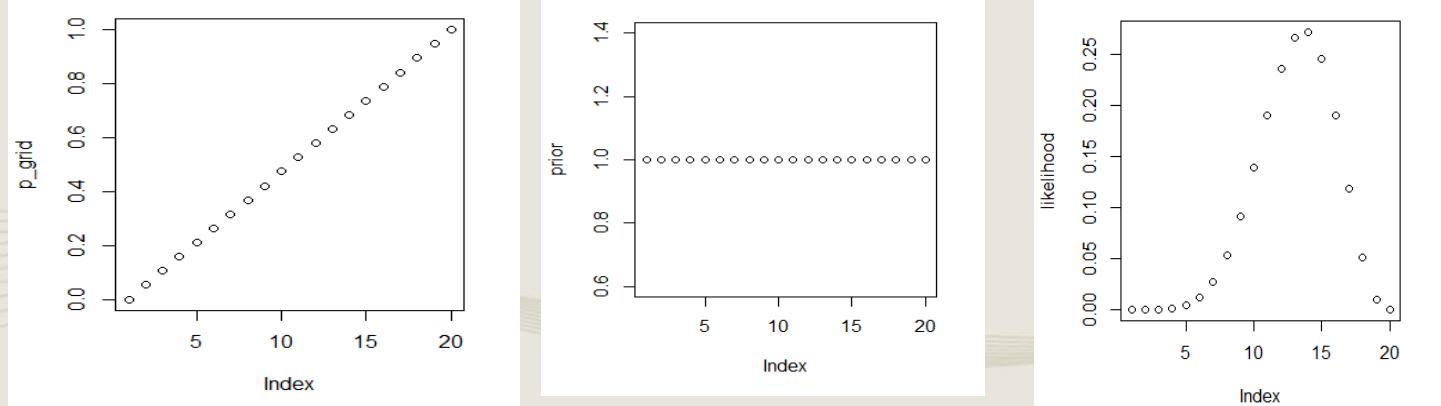


FIGURE 3.1. Sampling parameter values from the posterior distribution. Left: 10,000 samples from the posterior implied by the globe tossing data and model. Right: The density of samples (vertical) at each parameter value (horizontal).



uh? Muestras
de aquí?
Cómo por
qué?

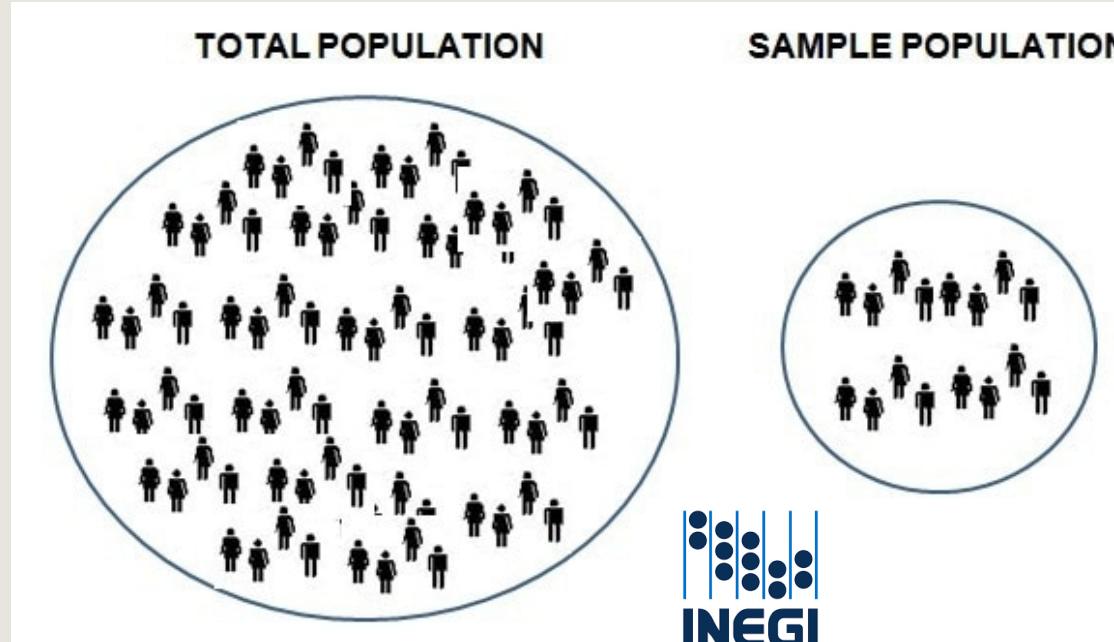




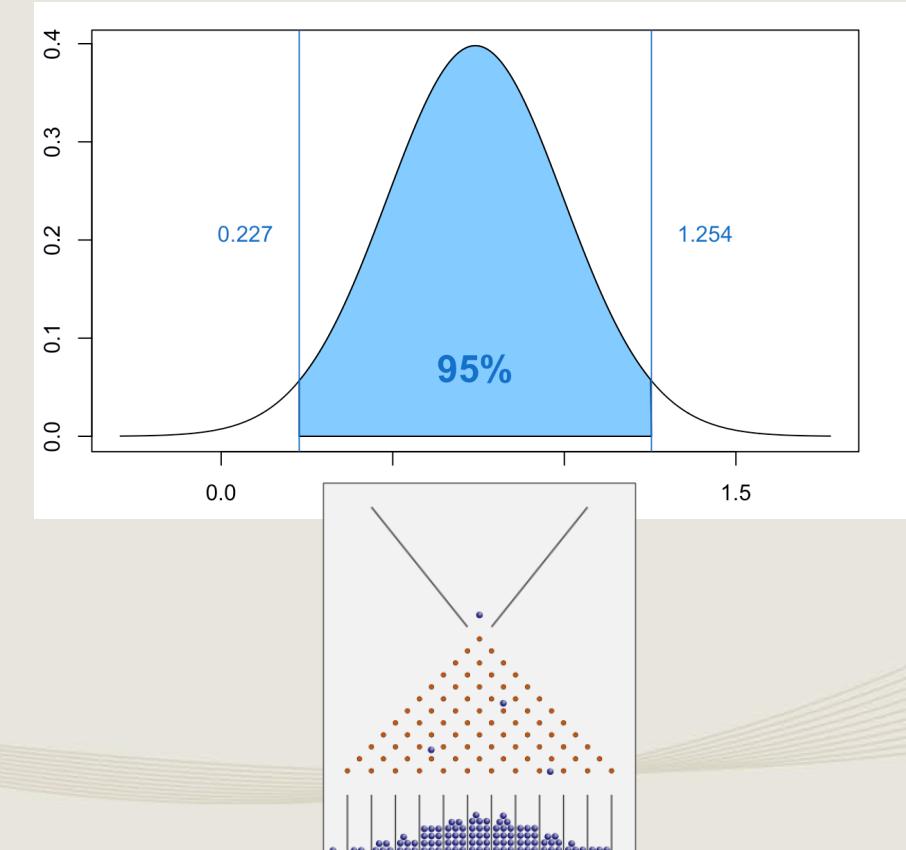
¿Dijo usted muestreo?

Tenemos dos tipos principales de muestreo en estadística

Muestreo de encuestas (survey sampling)



Muestreo a partir de una distribución de probabilidad





Problema de las dos monedas

- Tengo dos monedas: Una normal y otra con dos caras
 - ¿Cuál es la probabilidad de sacar la justa?
 - ¿Cuál es la probabilidad de elegir la justa dado que salió cara?

$$P(F|H) = P(F) \cdot P(H|F) / [P(F) \cdot P(H|F) + P(U) \cdot P(H|U)]$$



¿Cómo se muestrea a partir de una distribución de probabilidad?

- Induciendo ignorabilidad (a partir de números pseudo-aleatorios) en una estructura apropiada de forma iterada

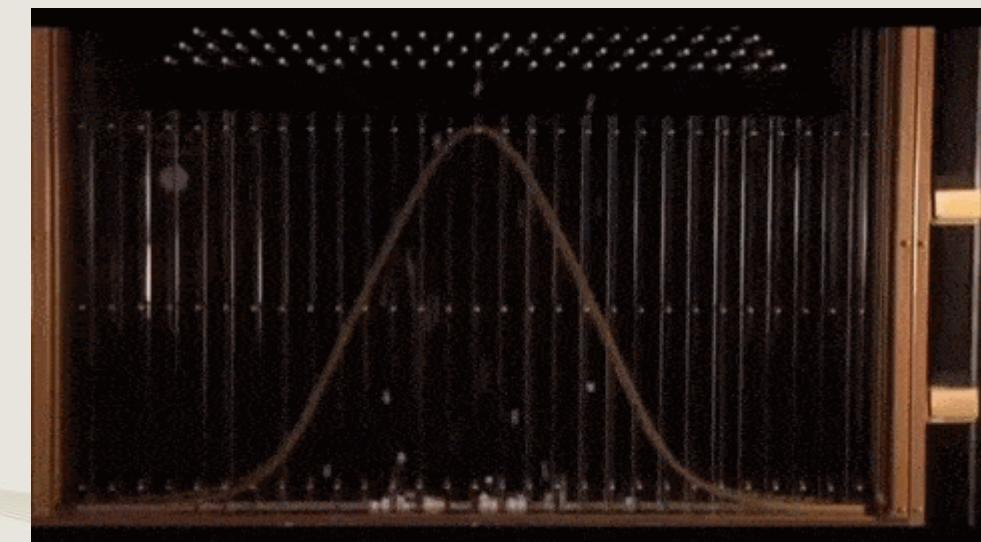
D10004	A	B	C	D
1	Coin	Flip	Head	Head Fair
2	U	H	1	
3	U	H	1	
4	F	H	1	1
5	U	H	1	
6	F	H	1	1
7	U	H	1	
8	F	T		
9995	F	H	1	1
9996	F	T		
9997	F	T		
9998	U	H	1	
9999	U	H	1	
10000	F	T		
10001	F	H	1	1
10002	5019		7471	2490
10003				
10004	P(F)= 0,5019		P(F1 H1)= 0,33329	

A2=SI(ALEATORIO()<0.5,"F","U")

B2=SI(A2="U","H",SI(ALEATORIO()<0.5,"H","T"))

C2=SI(B2="H",I,"")

D2=SI(Y(A2="F",B2="H"),I,"")

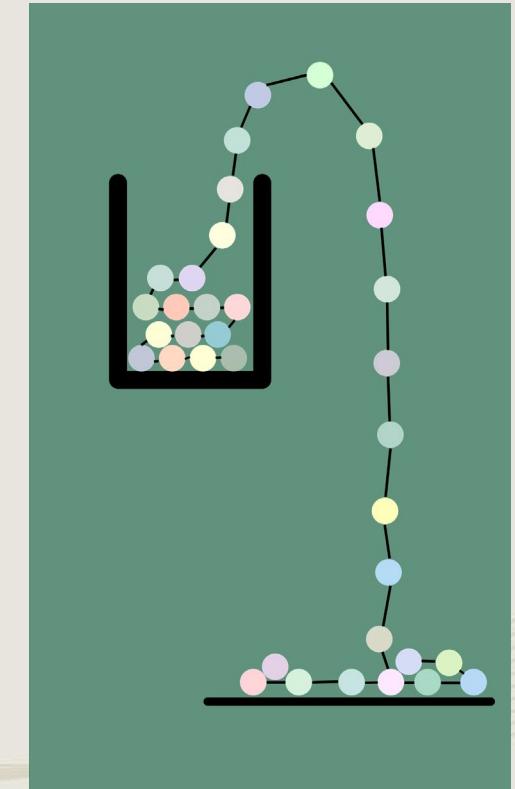


¿Cómo se muestrea a partir de una distribución de probabilidad?

- Usualmente, en las aplicaciones al “mundo real”, no es nada fácil muestrear a partir de distribuciones (CDFs a posteriori) complicadas.

Los Métodos Montecarlo aplicados a Cadenas de Markov permiten (tarde o temprano) recuperar las distribuciones objetivo.

- Metropolis (1953)
- Metropolis-Hastings (1970)
- Gibbs (1984)
- Hamiltoniano (o híbrido; 1987)

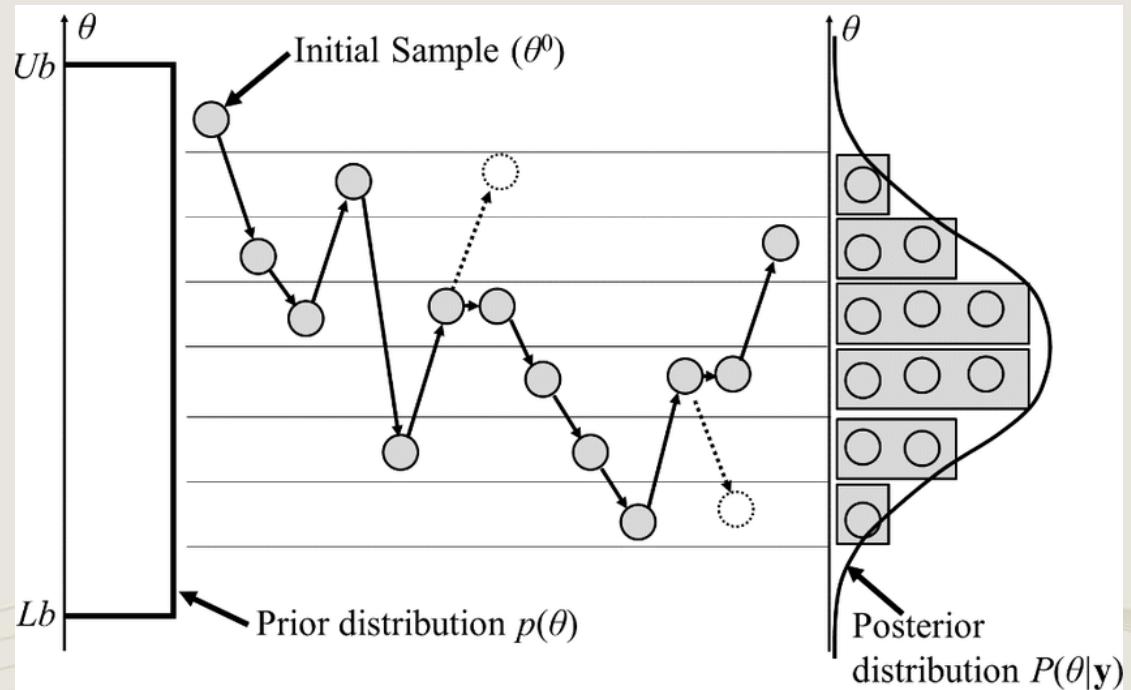


Algoritmo Metropolis



Diagrama de flujo del algoritmo Metropolis

¡Tan sencillo que cabe en un tweet!



Richard McElreath
@rlmcelreath

y=sum(rpois(20,2))
n=1e4
p=rep(1,n)
for(i in 2:n){
r=p[i-1]
q=exp(log(r)+rnorm(1)/9)
p[i]=ifelse(runif(1)<q^y*r^(-y)*exp(-20*(q-r)),q,r)
}

9:53 AM · May 18, 2016

34 2 Share this Tweet

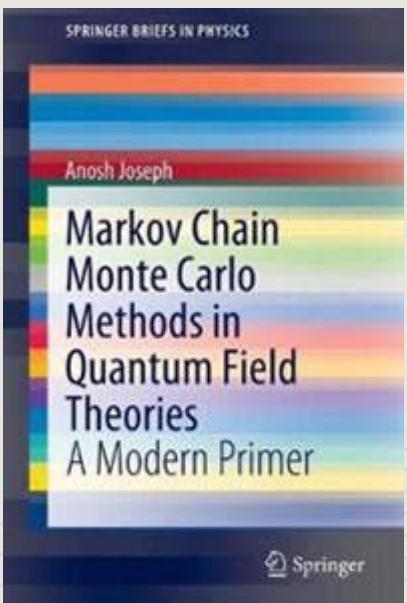


El camino a la posterior - MH

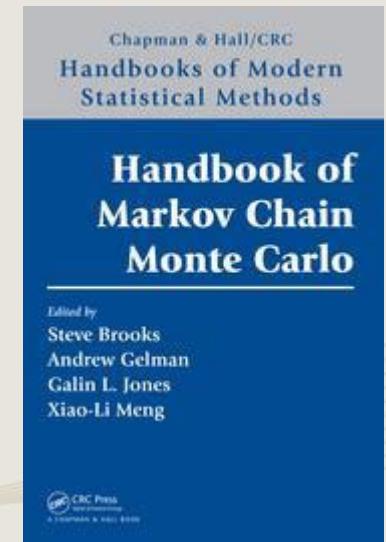
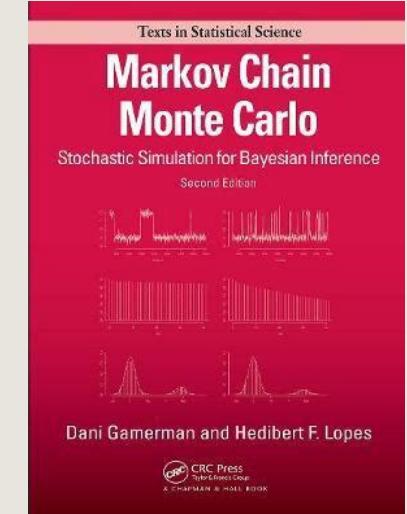
Monte Carlo

Markov Chains

MCMC



Algoritmo de Metrópolis





Monte Carlo

Monte Carlo

$$\theta_t \sim N(\mu, \sigma)$$

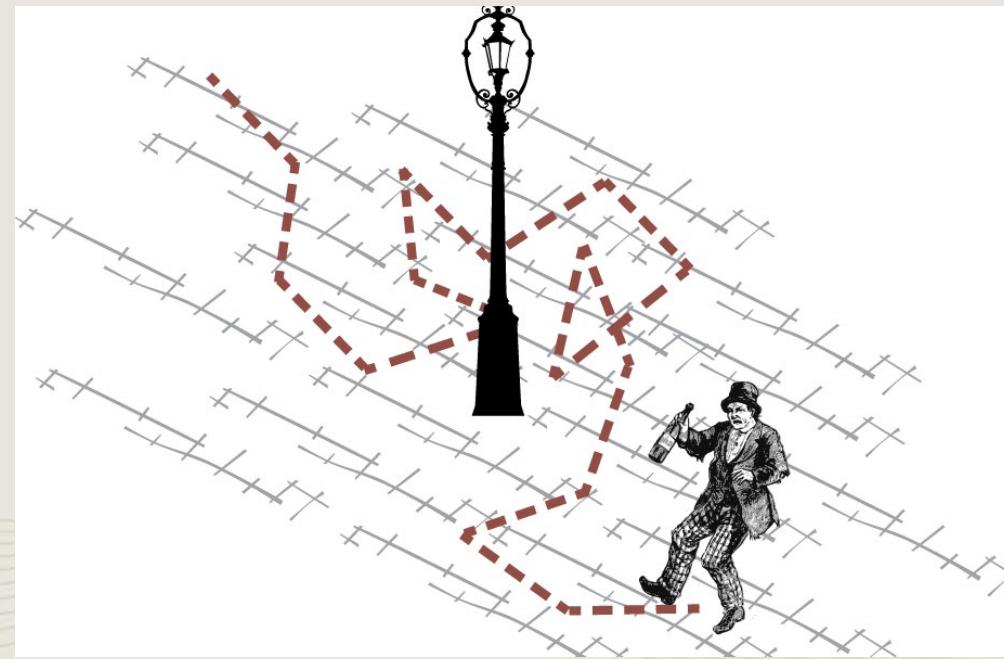


Pausa (k)

¿Dijo usted cadenas de Markov?

- Una cadena de Markov es un modelo (gráfico) de probabilidad que tiene la propiedad de Markov.
- La propiedad de Markov (de orden 1) consiste en suponer que en una serie de variables aleatorias X_1, X_2, \dots, X_n , cada variable aleatoria sólo depende de la más reciente y no del resto.

$$P(X_i | X_{i-1}, X_{i-2}, \dots, X_2, X_1) = P(X_i | X_{i-1})$$



https://miro.medium.com/max/934/1*bohRgb802KJvL8YVu9ynHg.png



Markov Chain Monte Carlo

$$\theta_t \sim N(\theta_{t-1}, \sigma)$$



Algoritmo Metropolis

- Este algoritmo muestrea de una función objetivo complicada (la posteriori) usando una distribución simple como propuesta de transición a partir de la última posición.
 - Se arranca de una posición aleatoria cualquiera (válida) en el espacio de parámetros.
 - Con base en la ubicación actual, se extrae de manera aleatoria, de una función de distribución simple y simétrica, una propuesta del siguiente paso (en la vecindad) a dar.
 - Si el paso conduce a un punto con mayor densidad de probabilidad (evaluando la posteriori en el punto), se acepta la propuesta y se avanza en la dirección planteada.
 - Si el paso conduce a un punto con menor densidad, se acepta la propuesta con probabilidad igual a la proporción de densidad que la propuesta representa de la posición actual. Si la propuesta se rechaza la posición actual se cuenta de nuevo.

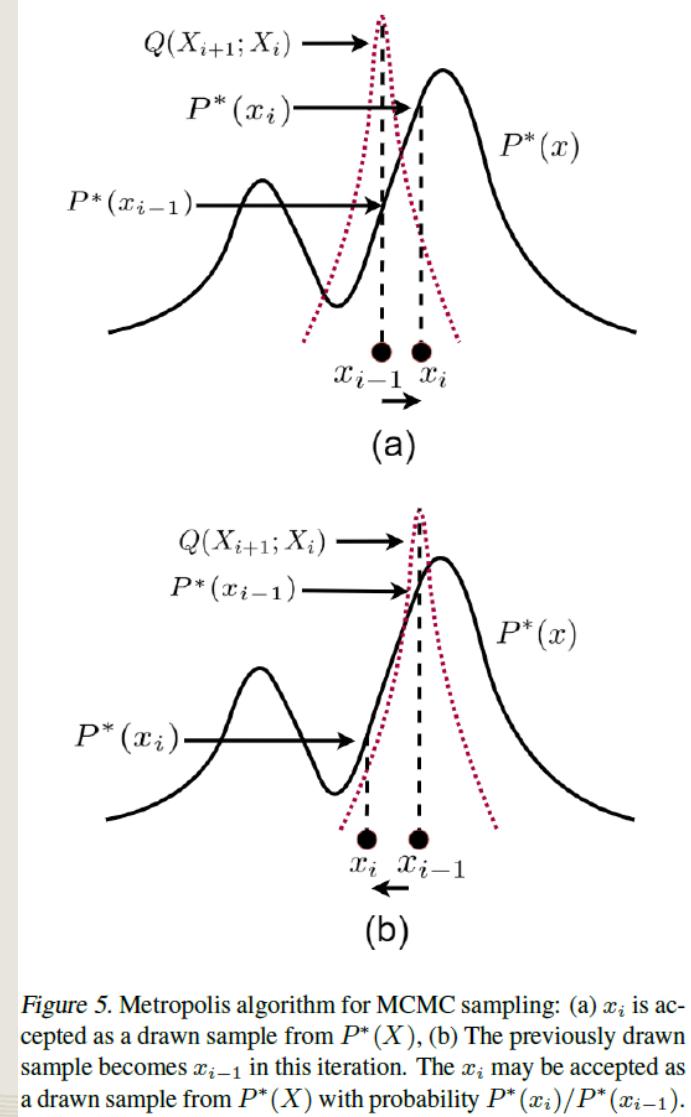
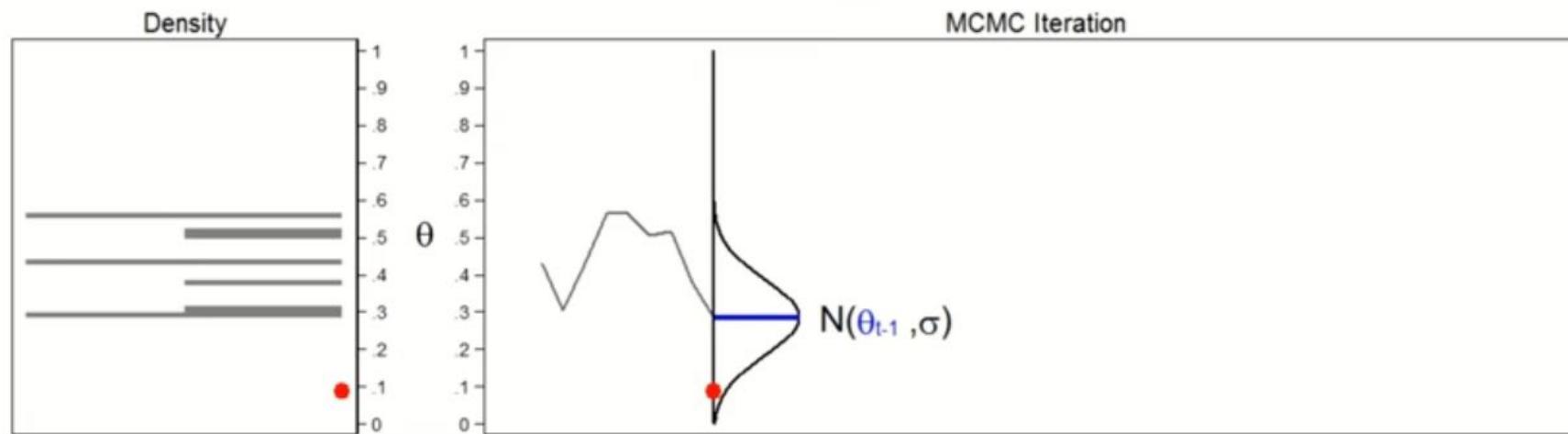


Figure 5. Metropolis algorithm for MCMC sampling: (a) x_i is accepted as a drawn sample from $P^*(X)$, (b) The previously drawn sample becomes x_{i-1} in this iteration. The x_i may be accepted as a drawn sample from $P^*(X)$ with probability $P^*(x_i)/P^*(x_{i-1})$.

Metropolis-Hastings MCMC



Step 1: $r(\theta_{\text{new}}, \theta_{t-1}) = \frac{\text{Posterior}(\theta_{\text{new}})}{\text{Posterior}(\theta_{t-1})} = \frac{\text{Beta}(1,1, 0.088) \times \text{Binomial}(10,4, 0.088)}{\text{Beta}(1,1, 0.286) \times \text{Binomial}(10,4, 0.286)} = 0.039$



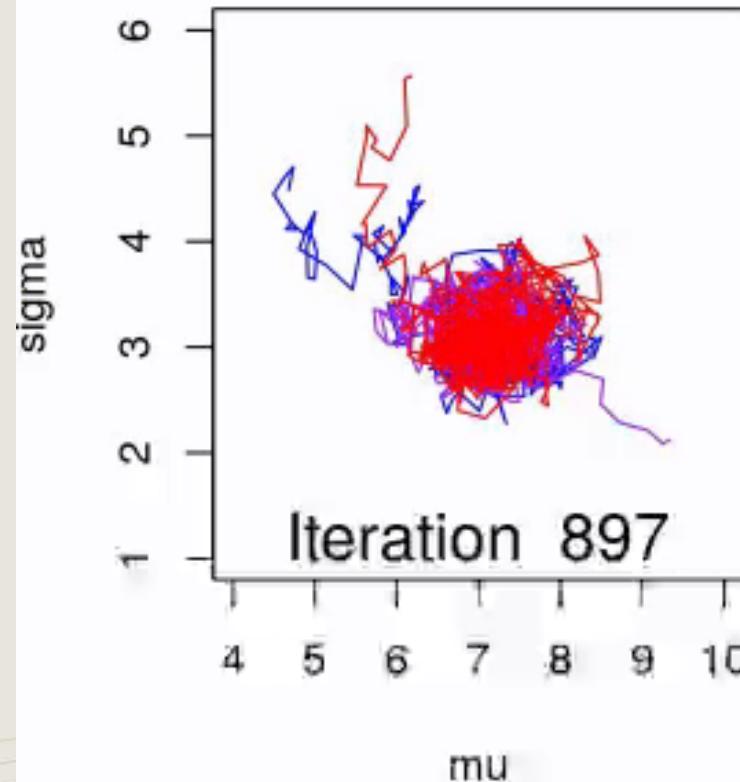
Step 2: Acceptance probability $\alpha(\theta_{\text{new}}, \theta_{t-1}) = \min\{r(\theta_{\text{new}}, \theta_{t-1}), 1\} = \min\{0.039, 1\} = 0.039$

Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.247$

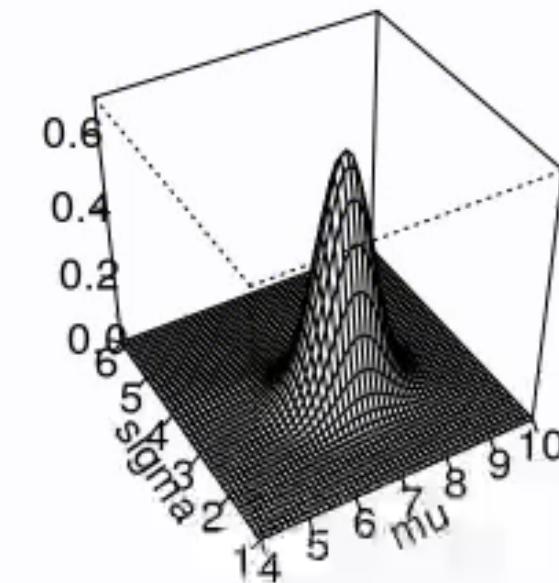
Reproducir (k) If $u < \alpha(\theta_{\text{new}}, \theta_{t-1}) \rightarrow$ If $0.247 < 0.039$ Then $\theta_t = \theta_{\text{new}} = 0.088$

M-H Dos variables

Markov chains

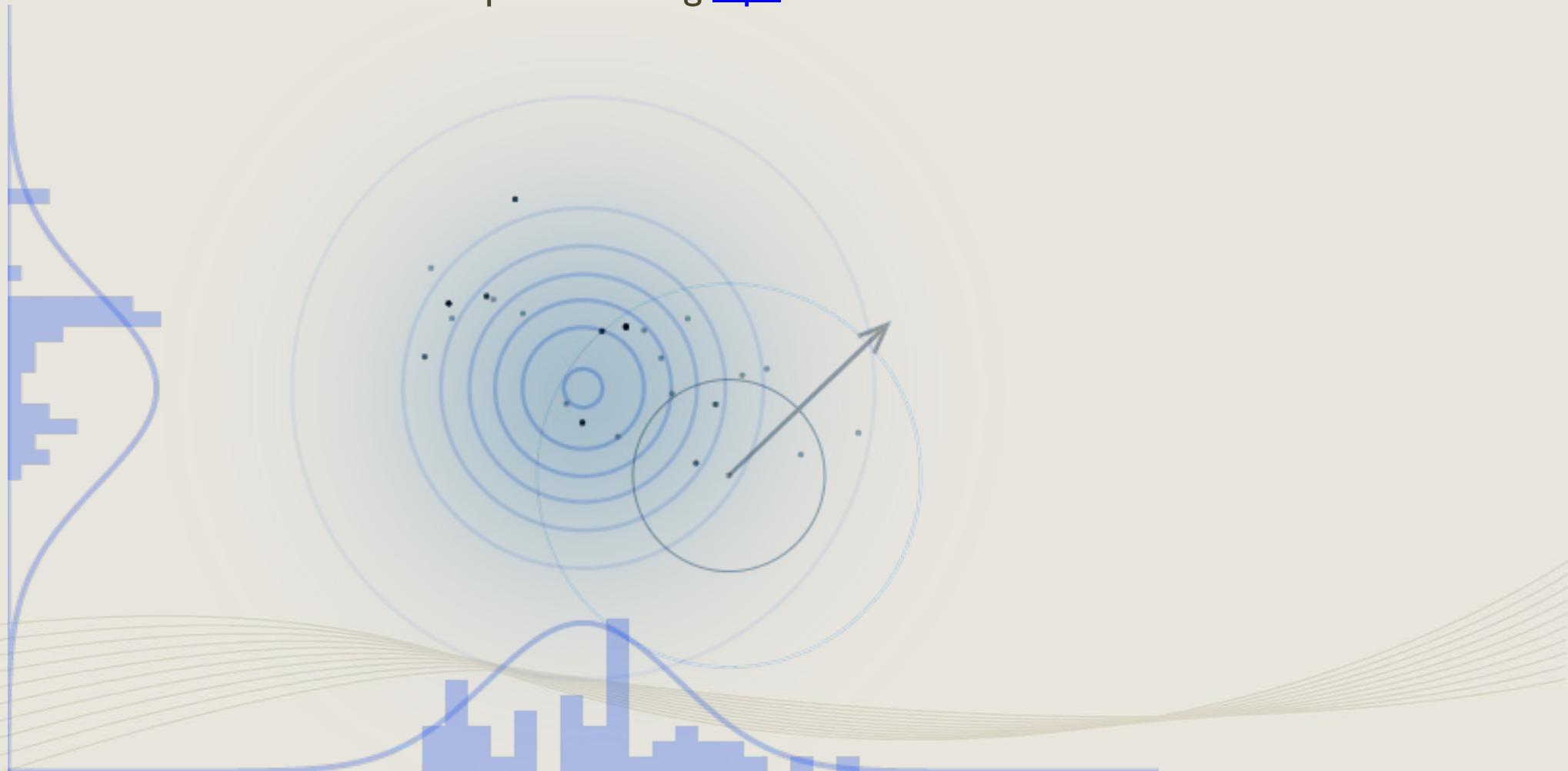


Posterior density



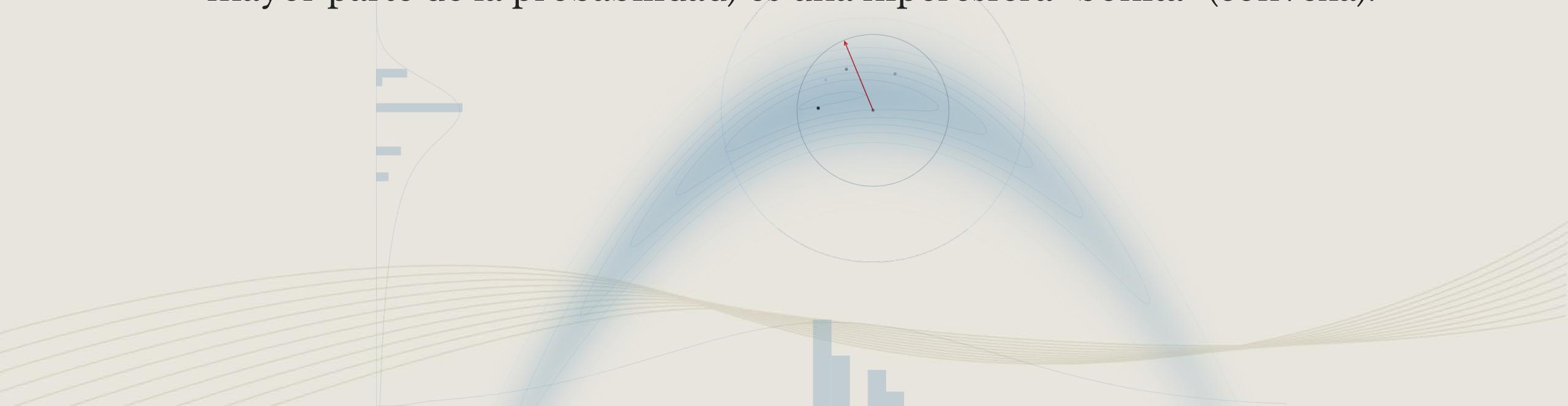
Simulaciones

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).



Algoritmo Metropolis

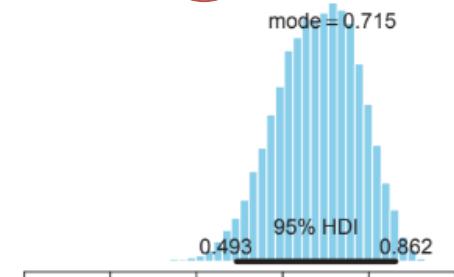
- A pesar de ir “dando tumbos” en una *senda aleatoria*, el algoritmo de Marshall y Arianna Rosenbluth (y su generalización Metropolis-Hastings) funciona, pero ese es justo su problema: **es demasiado aleatorio.**
 - Gasta muchísimo tiempo reexplorando las mismas partes de la distribución objetivo desperdiциando tiempo de cómputo valioso.
 - Prácticamente sólo es eficiente en el caso normal/gausiano de bajas dimensiones donde el *conjunto típico* (la parte del domino que concentra la mayor parte de la probabilidad) es una hiperesfera “bonita” (convexa).



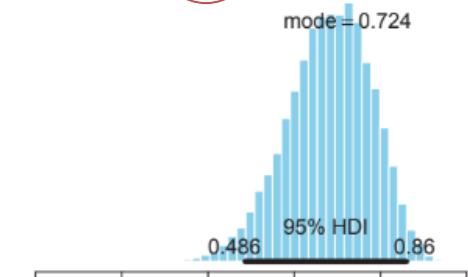
Buenas y malas muestras

Tamaño del
salto

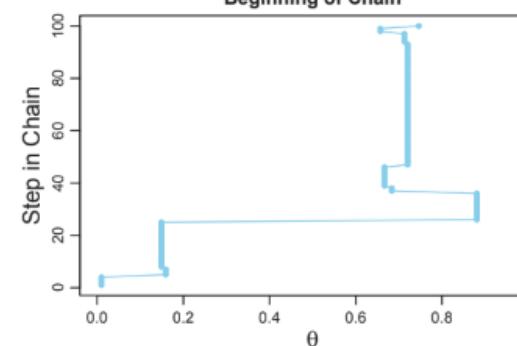
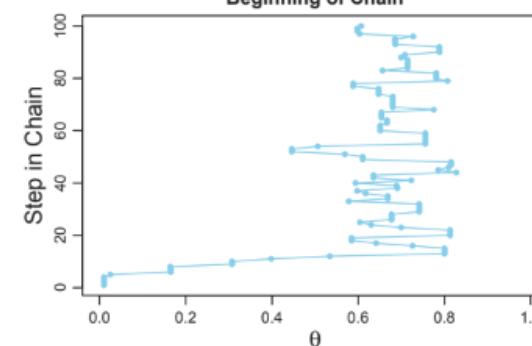
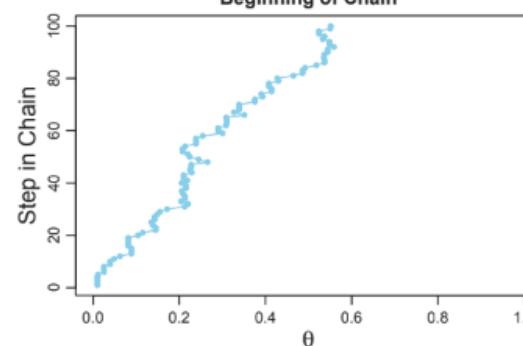
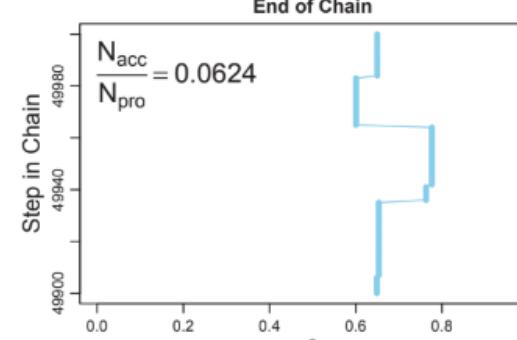
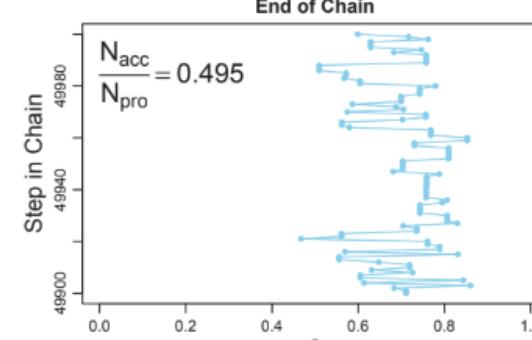
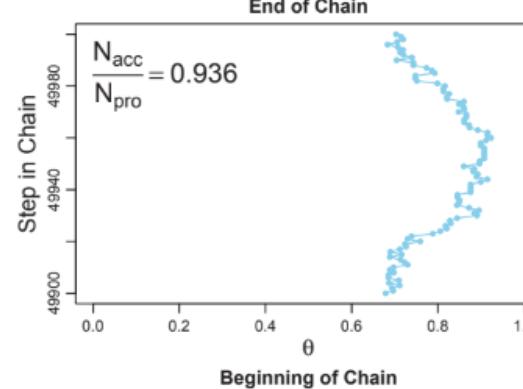
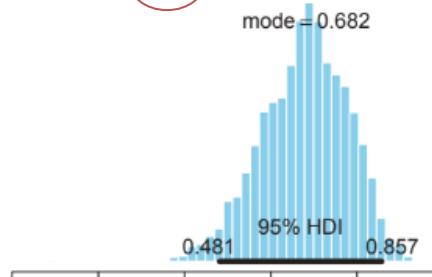
Prpsl.SD = 0.02, Eff.Sz. = 468.9



Prpsl.SD = 0.2, Eff.Sz. = 11723.9



Prpsl.SD = 2, Eff.Sz. = 2113.4



Los problemas en la exploración con MCMC

Por qué es un problema?

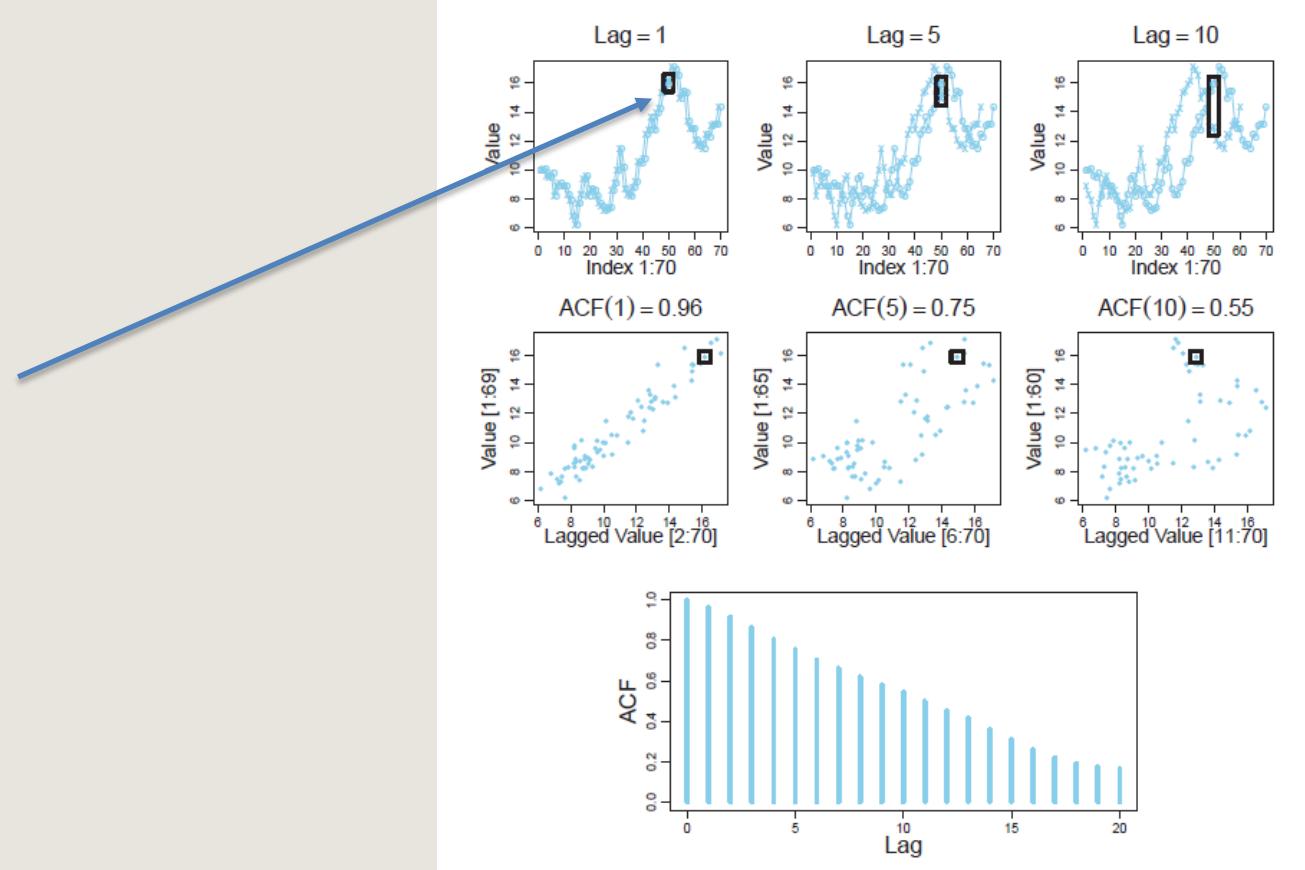


Figure 7.12: Autocorrelation of a chain. Upper panels show examples of lagged chains. Middle panels show scatter plots of chain values against lagged chain values, with their correlation annotated. Lowest panel shows the autocorrelation function (ACF). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.



Algoritmo Gibbs (MH)

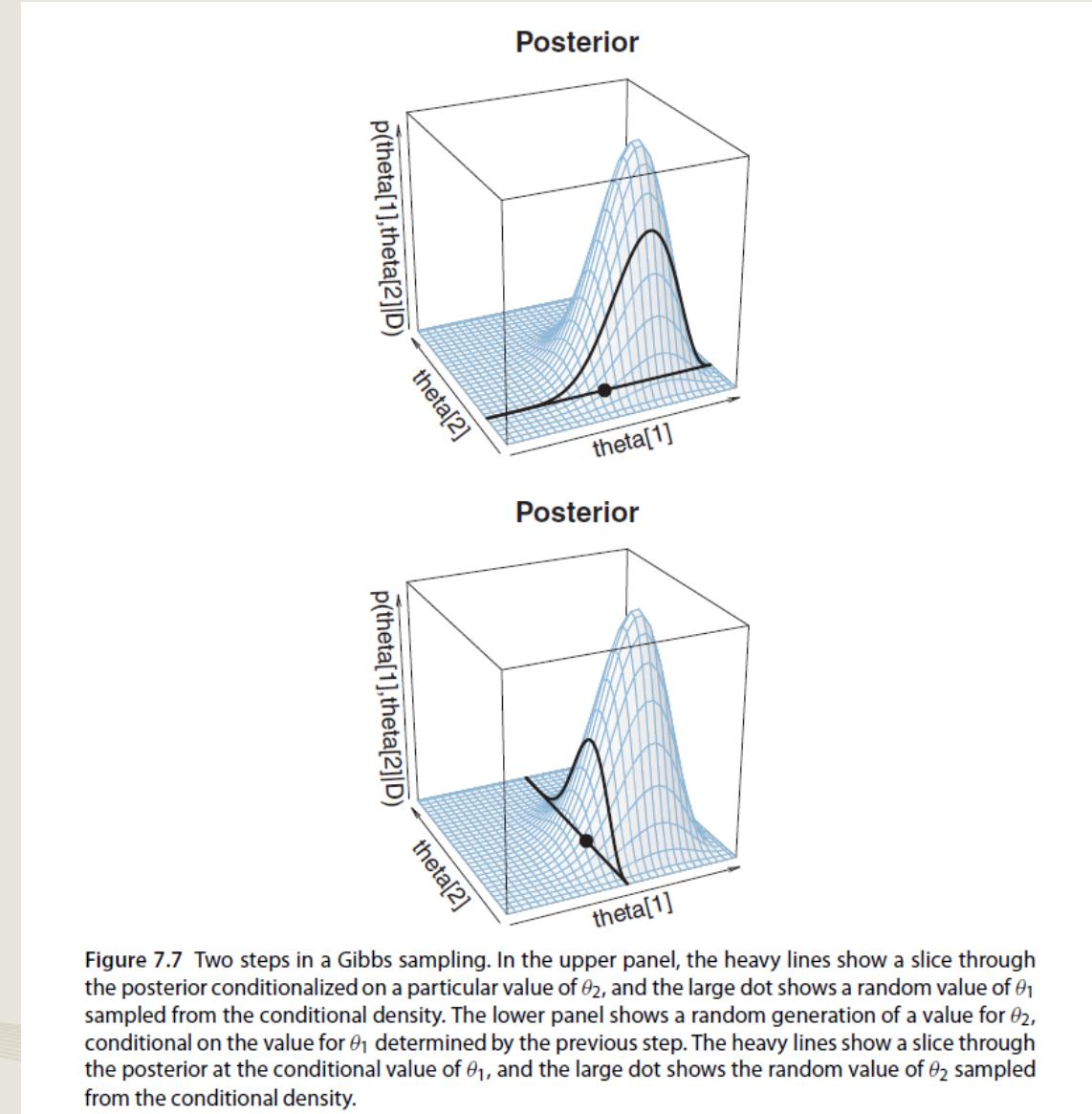


De izquierda a derecha, Stuart y Donald Geman, París 1983

Algoritmo Gibbs (MH)

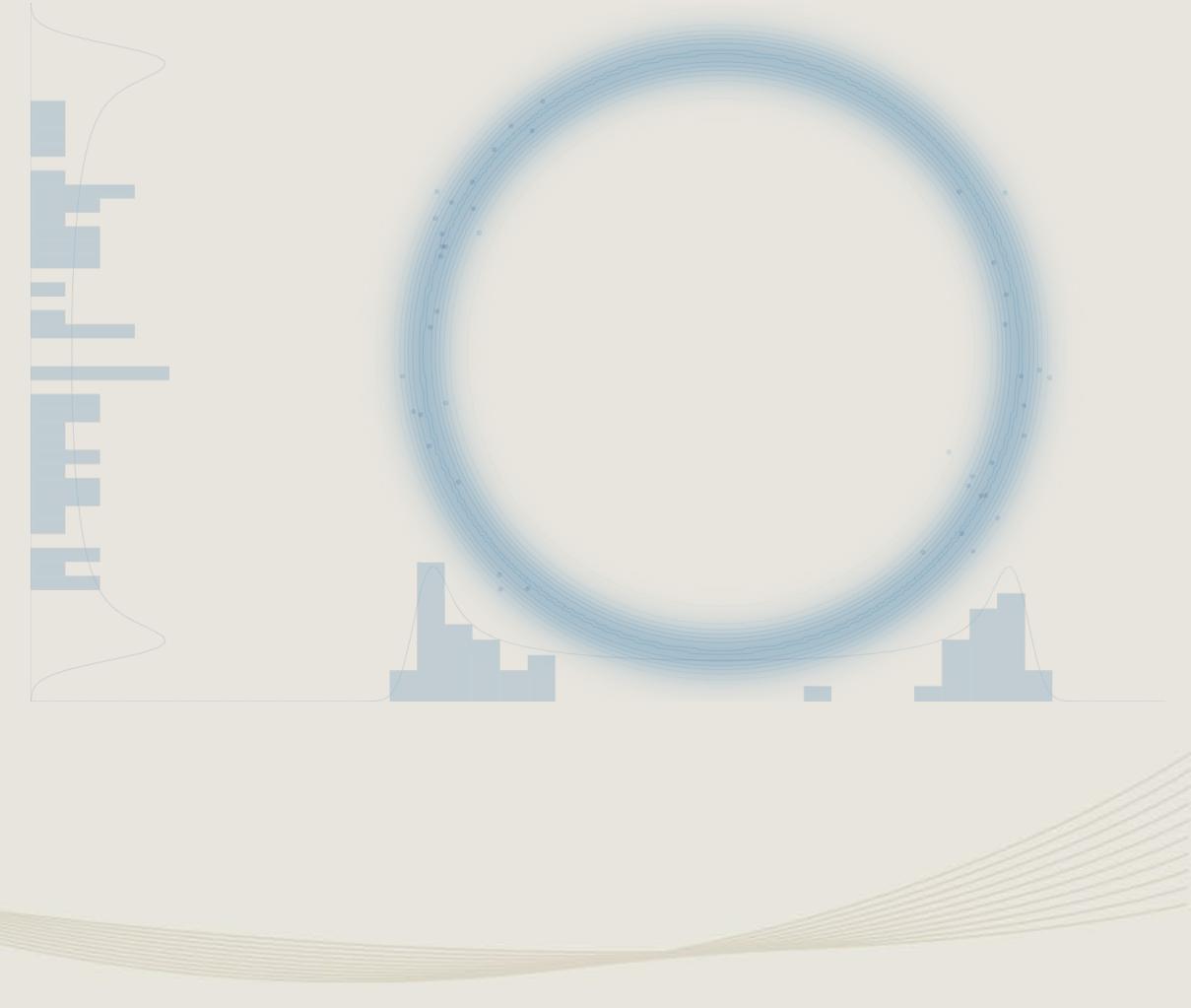
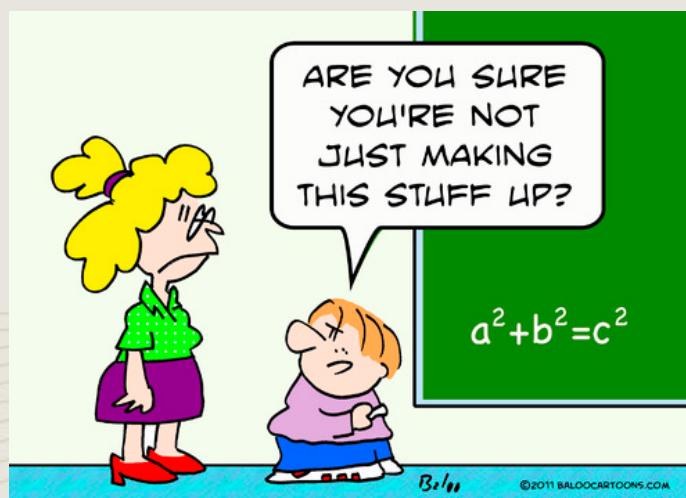
- El muestreo Gibbs es un caso particular de MH que modifica la distribución de donde se extraen las propuestas de siguiente paso en la senda aleatoria.
 - A cada paso de la senda, sólo una de las dimensiones del espacio de parámetros es seleccionada, y la propuesta de paso siguiente se extrae de manera aleatoria de la **distribución posterior condicional** (univariada) de ese parámetro, $p(\theta_1 | \{\theta_{j \neq 1}\}, D)$.

Noten que este algoritmo requiere que se pueda muestrear **directamente** de todas las distribuciones condicionales.



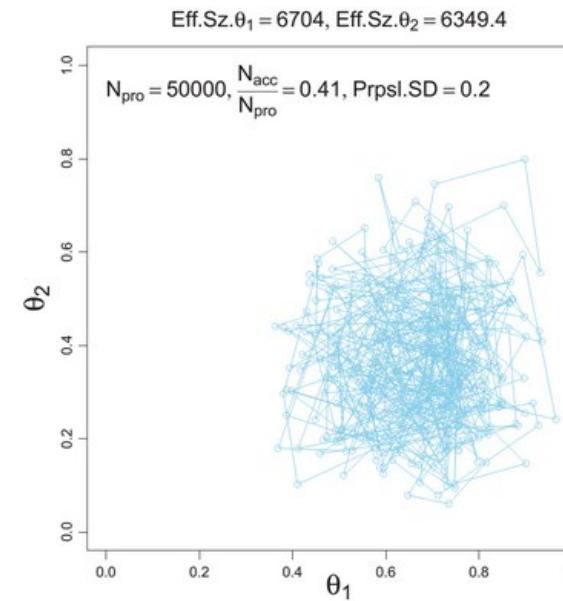
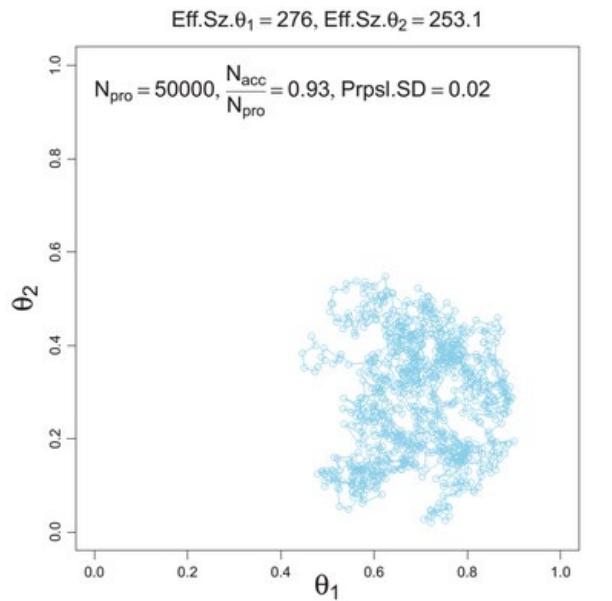
Algoritmo Gibbs (MH)

- Debido a que el muestreo à la Gibbs sólo cambia un parámetro a la vez, éste puede “atascarse” (dar pasos muy pequeños) con parámetros altamente correlacionados (conjuntos típicos esbeltos).



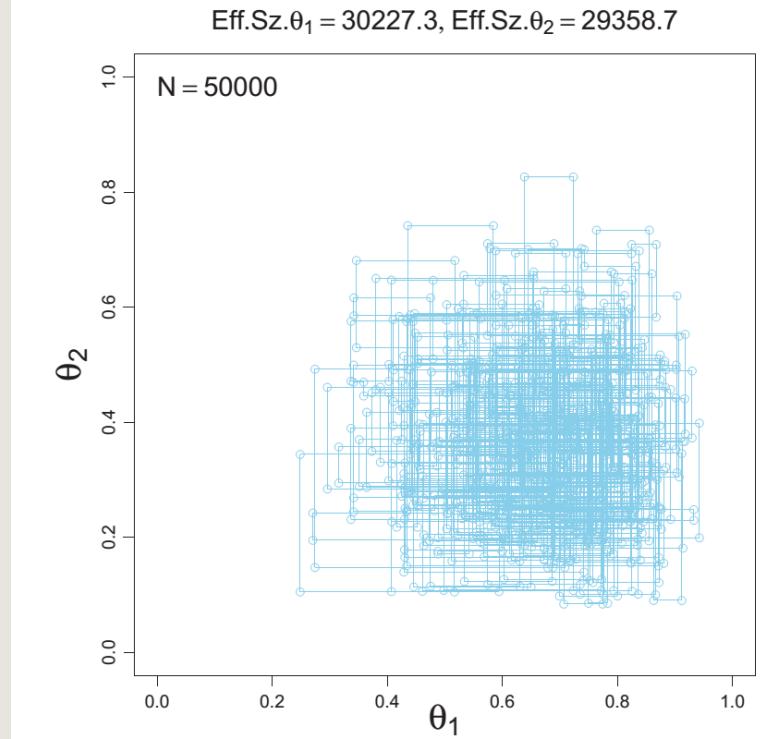
MH vs Gibbs

Metropolis-Hastings



Metropolis-Hastings is great, simple, and general. BUT ... sensitive to step size. AND ... can be too slow, because it can end up rejecting a great many steps.

Gibbs sampling



Lo que ustedes quieren son más muestras de calidad (efectivas) en menos tiempo!

De qué depende que sean más eficientes?



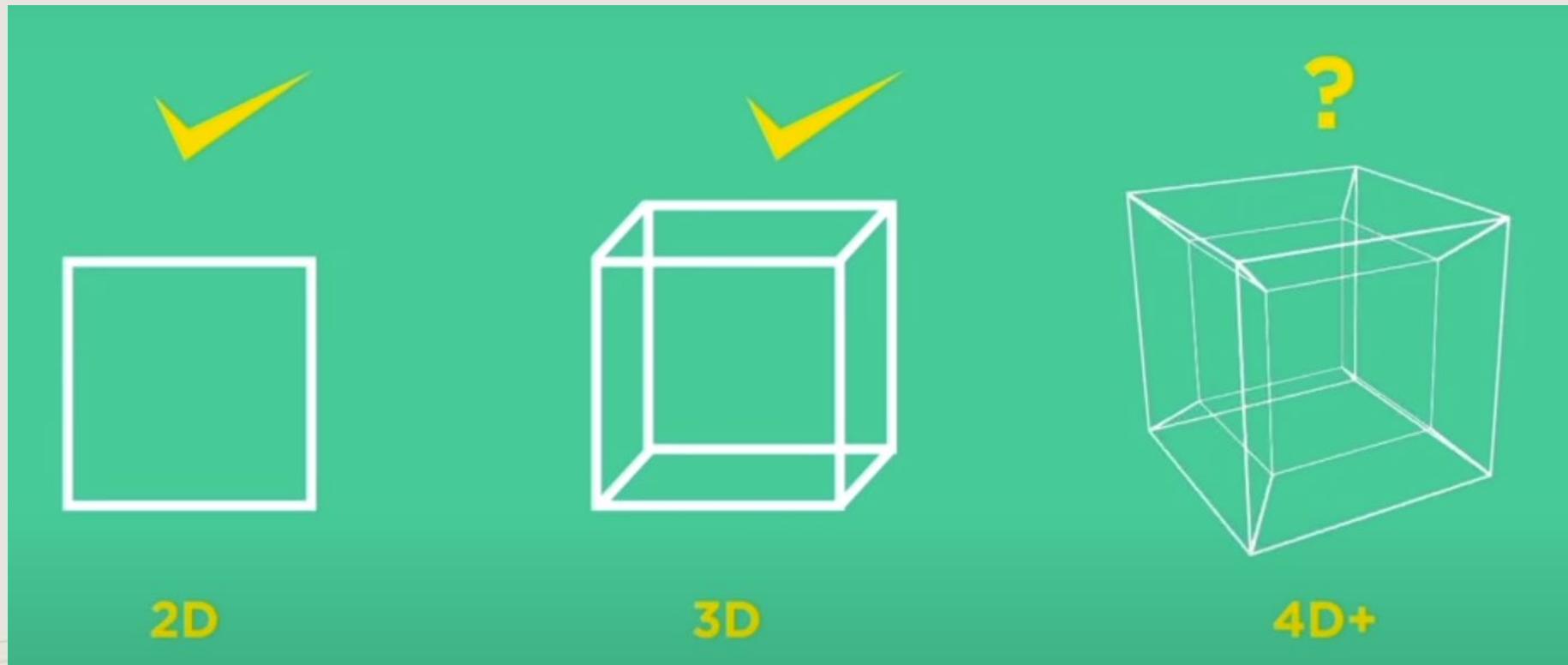
Mirando el Gibbs

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).



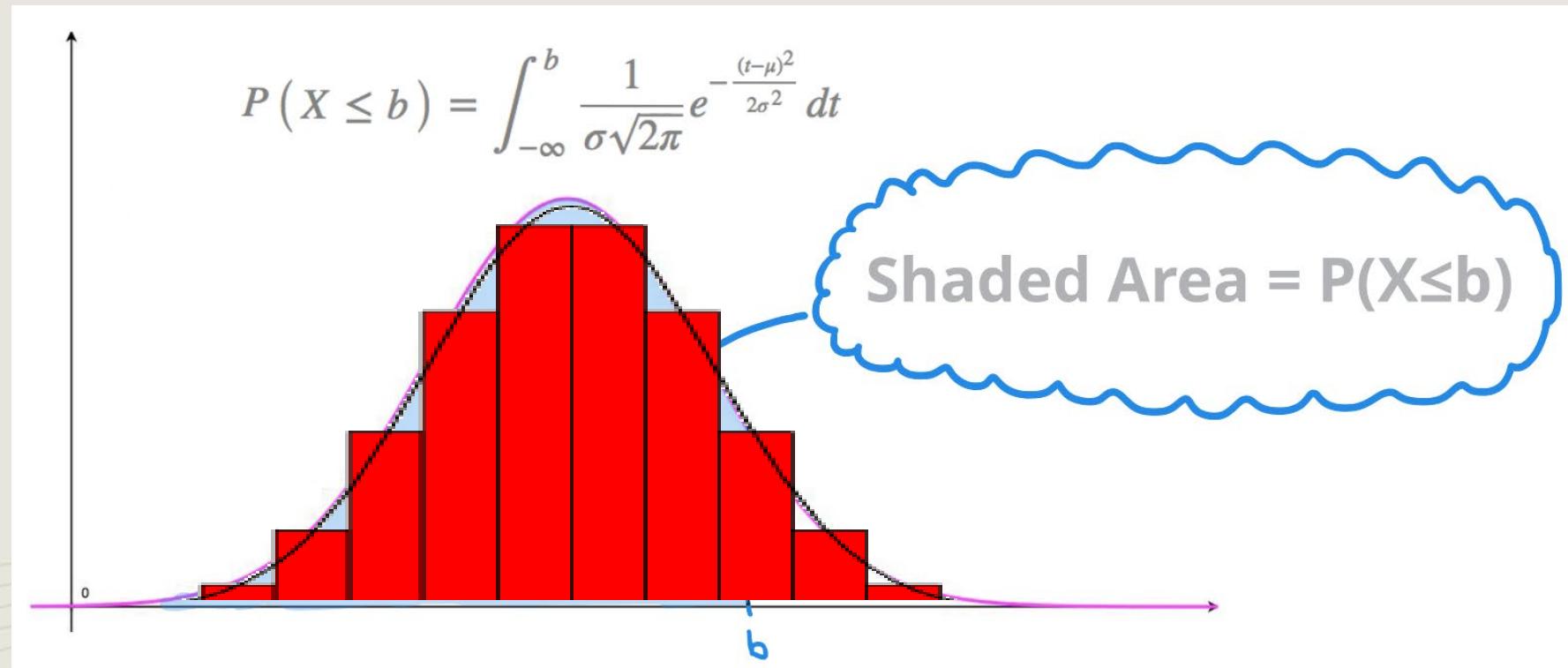
¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales



¿Conjuntos típicos esbeltos?

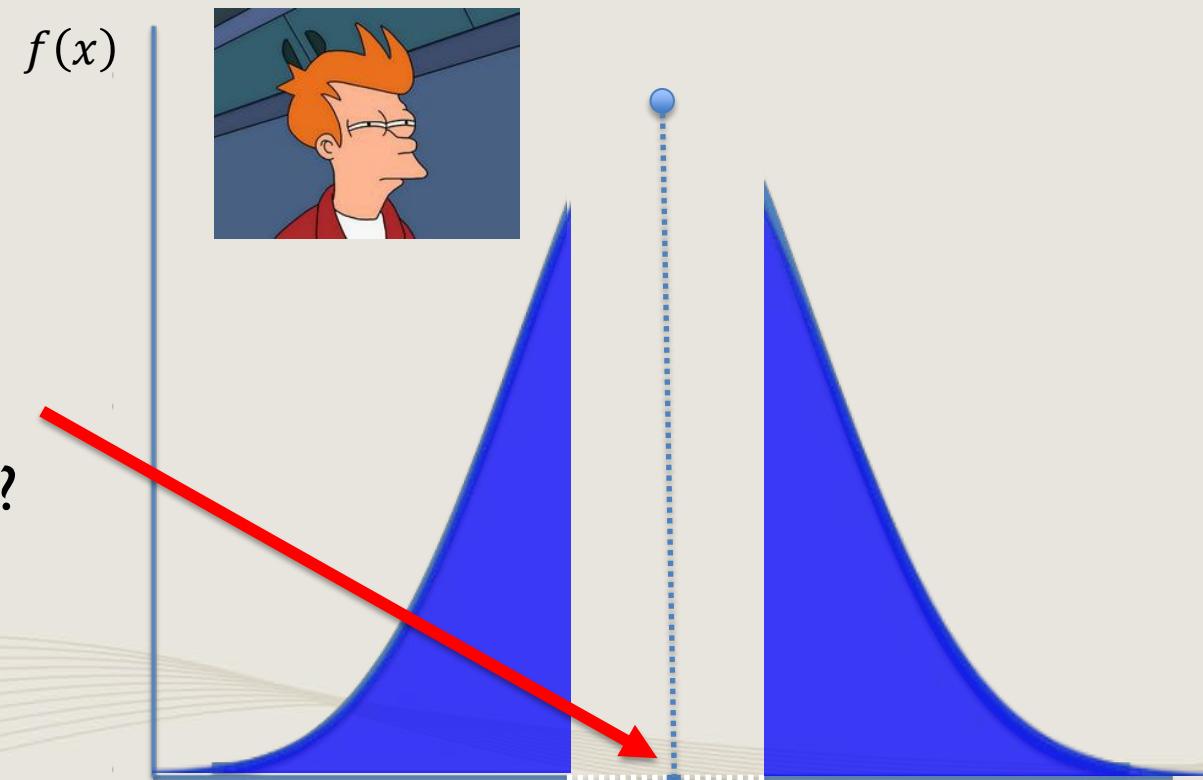
- Recuerden que, en el caso continuo, la masa de probabilidad está dada no meramente por la densidad (la altura de la función), sino por el **área** (base x altura) bajo la curva: el límite del área de los rectángulos rojos cuando la base tiende a cero (la integral)



¿Conjuntos típicos esbeltos?

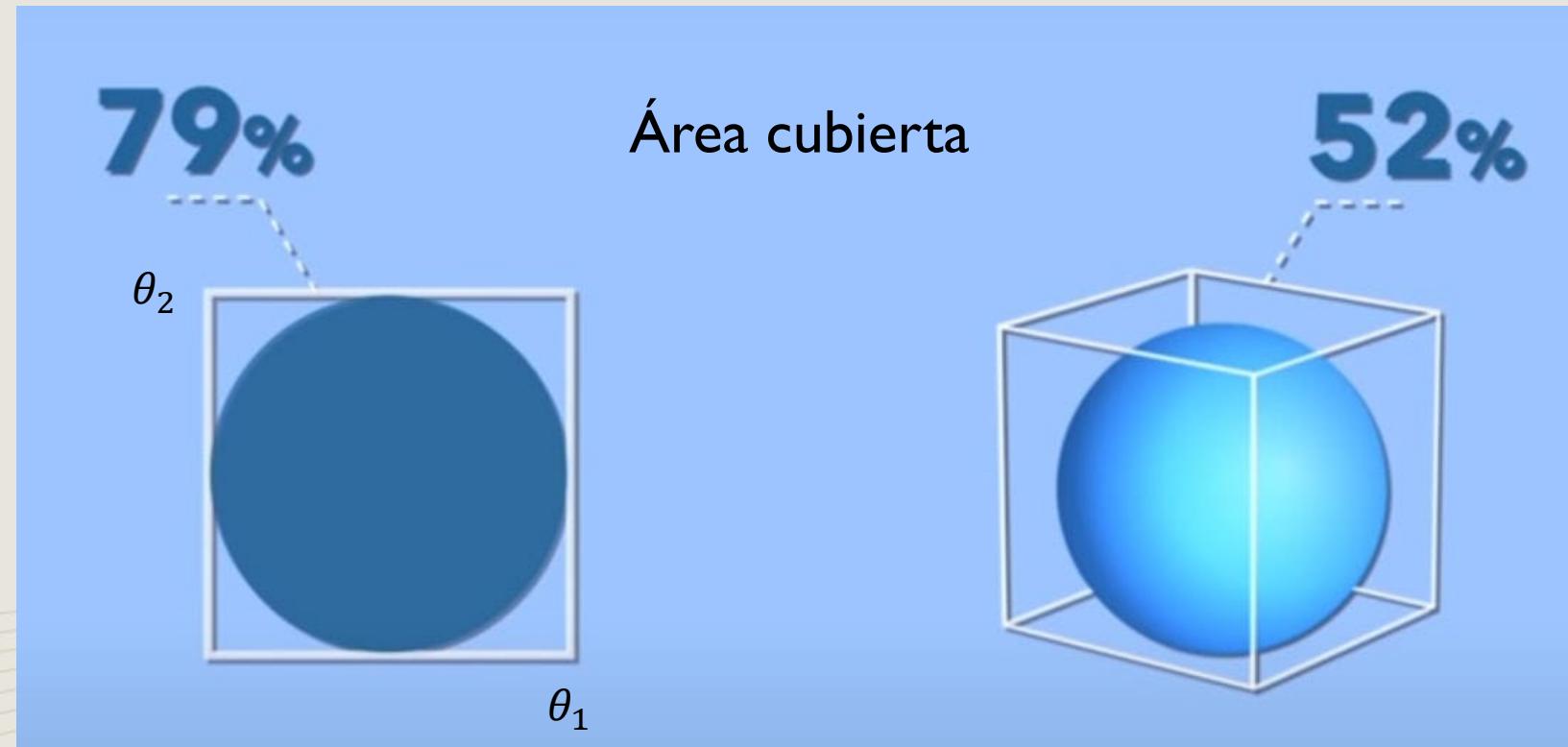
- La densidad (la altura), $f(x)$, puede ser mucha (la moda), pero si el dominio (la base) recorre poca distancia (área/volumen, dt), la masa de probabilidad que acumula (la integral) es despreciable.

¿Cómo puede recorrer
poca distancia
(área/volumen) el dominio?



¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).





¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).





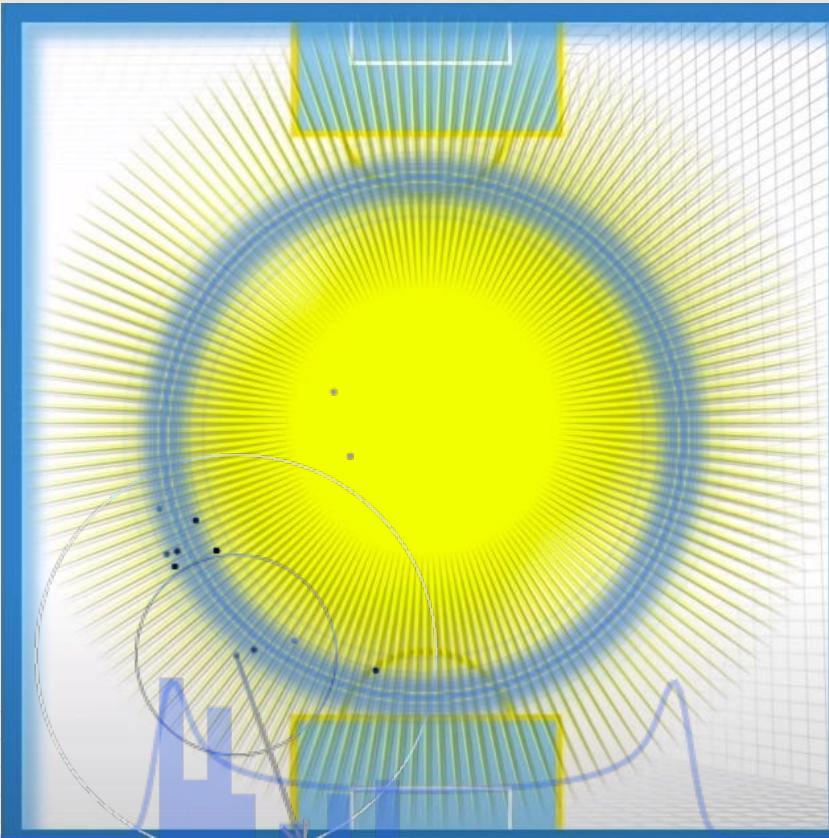
¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).



¿Conjuntos típicos esbeltos?

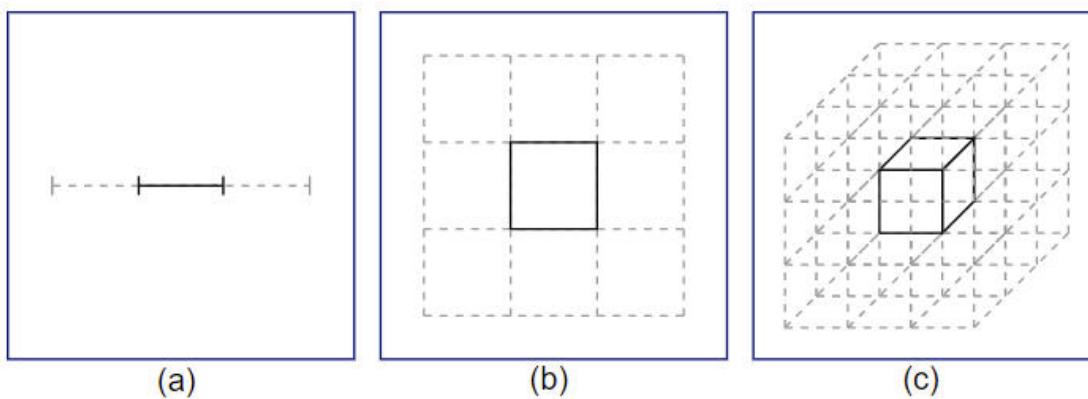
- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).



- Con sólo 30 dimensiones (parámetros), el recorrido/área/volumen cubierto por una hiperesfera circunscrita en ese espacio es proporcional a la de un grano de arroz en una cancha de futbol.
- La densidad puede ser mucha cerca de la moda (el centro de la cancha), pero el recorrido del dominio es despreciable.
- Hay muchísima más área lejos de la moda.

¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).

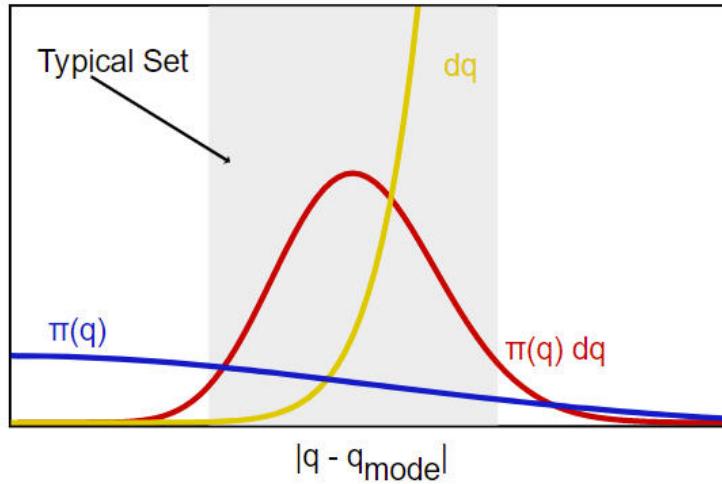


To illustrate the distribution of volume in increasing dimensions, consider a rectangular partitioning centered around a distinguished point such as the mode. The relative weight of the center partition is (a) $1/3$ in one dimension, (b) $1/9$ in two dimensions, (c) and only $1/27$ in three dimensions. In d dimensions there are 3^{d-1} neighboring partitions. Very quickly the volume in the center partition becomes negligible compared to the neighboring volume. This effect only amplifies if we consider larger regions around the mode, i.e. partitions beyond the nearest neighbors. For instance, for next-to-nearest neighbors, the base would be 5^{d-1} .

- Con sólo 30 dimensiones (parámetros), el recorrido/área/volumen cubierto por una hiperesfera circunscrita en ese espacio es proporcional a la de un grano de arroz en una cancha de futbol.
- La densidad puede ser mucha cerca de la moda (el centro de la cancha), pero el recorrido del dominio es despreciable.
- Hay muchísima más área lejos de la moda.

¿Conjuntos típicos esbeltos?

- Cosas inesperadas y contraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).



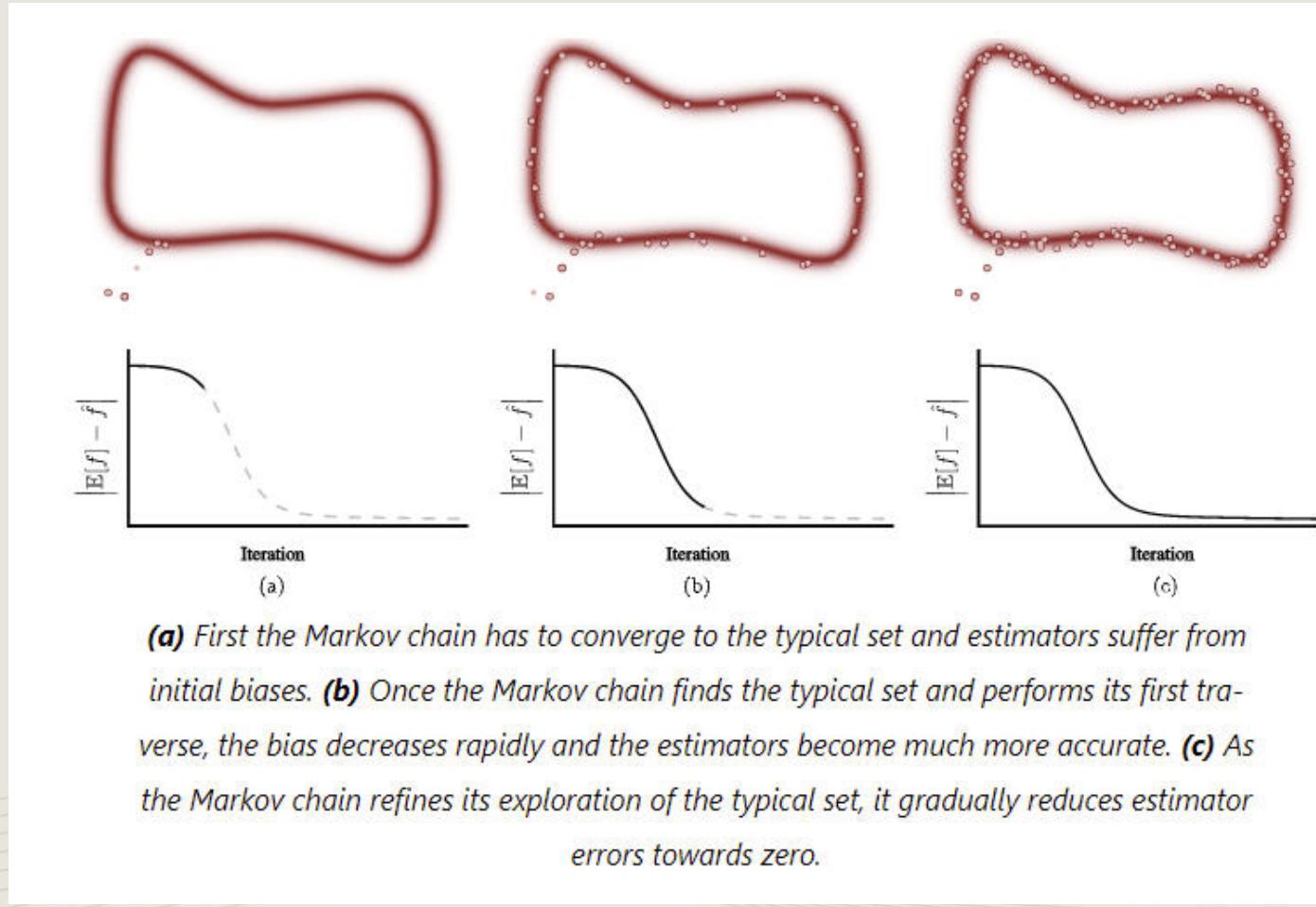
In high dimensions a probability density $\pi(\mathbf{q})$ will concentrate around its mode, but the volume $d\mathbf{q}$ over which we integrate that density is much larger away from the mode. Contributions to expectations are determined by the product $\pi(\mathbf{q}) d\mathbf{q}$. In sufficiently high dimensions, this condition is satisfied only in a nearly-singular region of \mathcal{Q} called the typical set. (This plot only shows a 10-dimensional independent, identically-distributed unit Gaussian. Hence the finite width of the typical set.. \mathbf{q} is the full 10-dimensional vector in parameter space and q in the above figure denotes its radial component.)

- Como aumenta el número de dimensiones (parámetros en el modelo), el conjunto típico (que acumula el grueso de la masa de probabilidad) se vuelve súper esbelto y prácticamente imposible de explorar dando tumbos (MH) o à la Gibbs.



¿Conjuntos típicos esbeltos?

En condiciones ideales, las cadenas de Markov exploran la distribución posterior en tres fases.



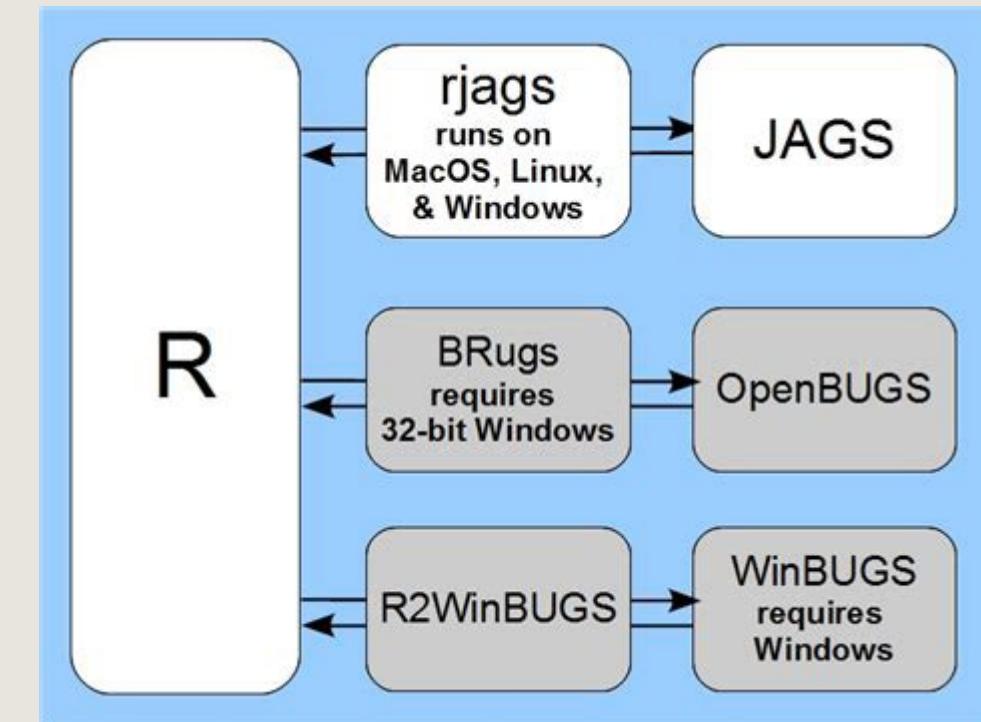
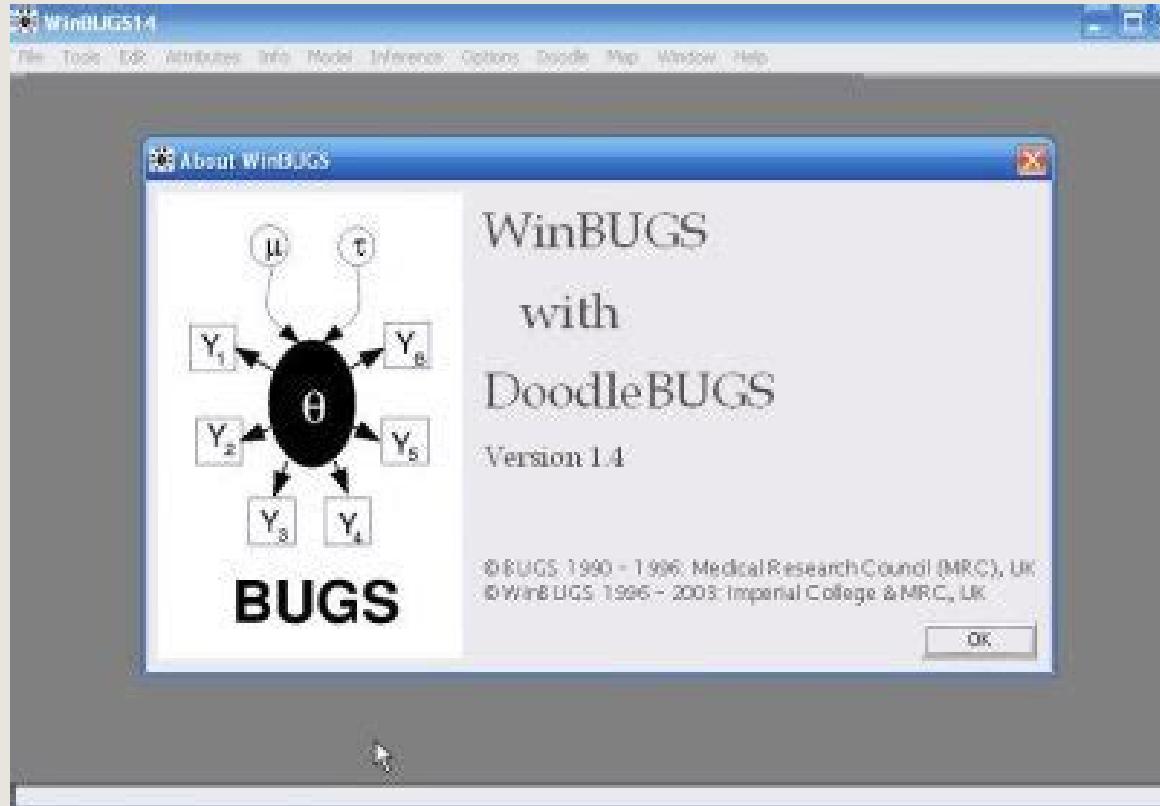


Mirando el Gibbs

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).

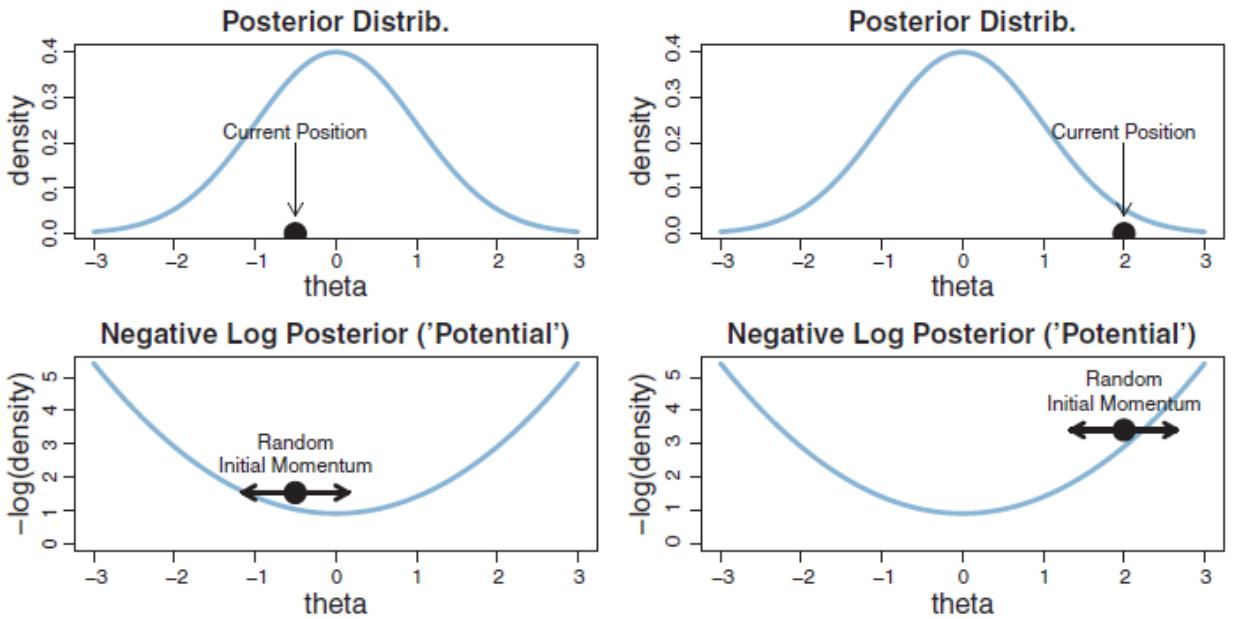


Software



¿Hay otra forma de hacerlo?

Capítulo 14





Referencias

Geman, Stuart and Geman, Donald. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6(6):721–741, 1984.

Ghojogh, B., Nekoei, H., Ghojogh, A., Karray, F., & Crowley, M. (2020). Sampling algorithms, from survey sampling to Monte Carlo methods: Tutorial and literature review. arXiv preprint arXiv:2011.00901.

Hastings, W Keith. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109, 1970.

Metropolis, Nicholas, Rosenbluth, Arianna W, Rosenbluth, Marshall N, Teller, Augusta H, and Teller, Edward. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Sitios de interés

<http://albertolumbreras.net/posts/maldicion-dimensionalidad.html>

<https://janosh.dev/blog/hmc-intro>

<https://chi-feng.github.io/mcmc-demo/>

Series y conferencias de Youtube

<https://youtu.be/ciM6wigZK0w>

<https://youtu.be/U561HGMVWjcw>



CONTACTO

Dr. Héctor Nájera y Dr. Curtis Huffman
Investigadores

Programa Universitario de Estudios del Desarrollo (PUED)
Universidad Nacional Autónoma de México (UNAM)
Antigua Unidad de Posgrado (costado sur de la Torre II de Humanidades), planta baja.
Campus Central, Ciudad Universitaria, Ciudad de México, México.
Tel. (+52) 55 5623 0222, Ext. 82613 y 82616

Tel. (+52) 55 5622 0889
Email: hecatalan@hotmail.com, chuffman@unam.mx

