



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas

Computación Estadística

Alan Riva Palacio Cohen

Agosto 31, 2021

Índice

1. Monte Carlo vía Cadenas de Markov	1
1.1. Introducción y problema a resolver	1
1.2. Cadenas de Markov	2
1.3. Algoritmo de Metropolis-Hastings	6
1.3.1. Elección de q	6
1.3.2. Metropolis-Hastings adaptativo	8
1.3.3. Mezcla de cadenas	8
1.4. Muestreo de Gibbs	8
1.5. Monte Carlo Hamiltoniano (HMC)	12
1.6. Diagnóstico de convergencia	18
1.7. Tamaño efectivo de muestra	24
1.8. Slice Sampler	27
2. Código	33
3. Bibliografía	35

1. Monte Carlo vía Cadenas de Markov

1.1. Introducción y problema a resolver

Sea f_k el kernel de una función de densidad de probabilidad f en \mathbb{R}^d , es decir f positiva y $0 < \int_{\mathbb{R}^d} f_k(x)dx < \infty$. Tal densidad induce la medida de probabilidad:

$$P[X \in A] = \frac{\int_A f_k(x)dx}{\int_{\mathbb{R}^d} f_k(x)dx}$$

para $A \in \mathbb{R}^d$. Si se quiere estimar la esperanza

$$\mathbb{E}_f[h(X)] = \frac{\int_{\mathbb{R}^d} h(x)f_k(x)dx}{\int_{\mathbb{R}^d} f_k(x)dx}$$

para $h : \mathbb{R}^d \rightarrow \mathbb{R}$. Si f_k es complicada puede resultar difícil simular variables aleatorias independientes e idénticamente distribuidas con la ley de probabilidad correspondiente.

La alternativa provista por los métodos de MCMC consiste en construir una cadena de Markov que tome valores en \mathbb{R}^d que sea fácil de simular en una computadora y que tenga a f como distribución estacionaria. Es decir, se busca construir cadenas de Markov sencillas de programar y que tengan kernel de transición de probabilidad $P(x, (dy))$ tal que (para $A \in \mathcal{B}(\mathbb{R}^d)$)

$$\int_{\mathbb{R}^d} \int_{y \in A} f(x)P(x, y)dydx = \int_A f(x)dx,$$

alternativamente se utiliza la notación

$$\int_{\mathbb{R}^d} \int_{y \in A} f(dx)P(x, dy) = \int_A f(dx)$$

o

$$\int_{\mathbb{R}^d} f(dx)P(x, A) = f(A).$$

Así si se generan $X_0, X_1, \dots, X_{n-1}, X_n$ para $n \in N$ grande, entonces la ley de probabilidad de X_n será aproximadamente la de la distribución estacionaria

$$\mathcal{L}(X_n) \approx \mathcal{L}(f).$$

Con lo que se puede tomar $Z_1 = X_n$, volver al inicio de la cadena y simularla nuevamente para generar Z_2, Z_3, \dots y estimar $\mathbb{E}_f[h(X)]$ con

$$\frac{1}{m} \sum_{i=1}^m h(Z_i).$$

En práctica para los métodos MCMC no se reinicia la simulación de la cadena para generar cada Z_i , en vez se utiliza la cola de una sola cadena dando el estimador

$$\frac{1}{m-b} \sum_{i=b+1}^m h(X_i),$$

donde a b se le llama el periodo de "quemado" (o "burn-in"). Si b es suficientemente grande se espera que $\mathcal{L}(X_b) \approx \mathcal{L}(f)$. En este caso sin embargo $h(X_{b+1}), h(X_{b+2}), \dots, h(X_m)$ no son independientes pero el estimador puede ser calculado con mayor eficiencia.

Puede aparentar ser más complicado construir tal cadena de Markov en vez de estimar $\mathbb{E}_f[h(X)]$ directamente; sin embargo estos métodos han sido muy bien estudiados e implementados para un uso eficiente en comparación de otras técnicas.

1.2. Cadenas de Markov

Definición 1.1 (Kernel de transición).

Un kernel de transición de probabilidad P definido en $\mathcal{X} \times \mathcal{B}(\mathcal{X})$, con \mathcal{X} un subconjunto de \mathbb{R} o \mathbb{R}^d arbitrario, es una función tal que

1. $\forall x \in \mathcal{X}, P(x, \cdot)$ es una medida de probabilidad.
2. $\forall A \in \mathcal{B}(\mathcal{X}), P(\cdot, A)$ es medible.

Si \mathcal{X} es discreto entonces el kernel de transición de probabilidad puede ser visto como una matriz de transición de probabilidad P con entradas dadas por

$$P[x, y] = \mathbb{P}[X_n = y | X_{n-1} = x].$$

En el caso continuo, el kernel también denota una probabilidad condicional. En ese caso dado por una densidad $P(x, x')$ de la transición condicionada a estar en x en el paso anterior, es decir

$$\mathbb{P}[X_n \in A | X_{n-1} = x] = \int_A P(x, x') dx'$$

Definición 1.2 (Cadena de Markov).

Una secuencia de variables aleatorias $X_0, X_1, \dots, X_{n-1}, X_n, \dots$ es una cadena de Markov si para cualquier $n \in \mathbb{N} \setminus \{0\}$ la distribución condicional de X_n dado $X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_0 = x_0$ es igual a la distribución de X_n dado $X_{n-1} = x_{n-1}$, es decir

$$\begin{aligned} \mathbb{P}[X_n \in A | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_0 = x_0] &= \mathbb{P}[X_n \in A | X_{n-1} = x_{n-1}] \\ &\stackrel{\text{def}}{=} \int_A P(x_{n-1}, x') dx' = \int_A P(x_{n-1}, dx') = P(x_{n-1}, A) \end{aligned}$$

donde P es un kernel de transición de probabilidad.

Se dice que una cadena de Markov es homogénea en el tiempo o simplemente homogénea, si la distribución de $(X_{t_1}, \dots, X_{t_k})$ dado X_{t_0} es igual a la distribución de $(X_{t_1-t_0}, X_{t_2-t_0}, \dots, X_{t_k-t_0})$ dado x_0 para todo k y para toda malla $t_0 \leq t_1 \leq \dots \leq t_k$.

Se hará uso de la notación

$$\mathbb{P}_x[A] = \mathbb{P}[A | X_0 = x]$$

con lo que

$$\begin{aligned} \mathbb{P}_x[X_1 \in A_1] &= P(x, A_1) \\ \mathbb{P}_x[(X_1, X_2) \in A_1 \times A_2] &= \int_{A_1} P(y_1, A_2) P(x, dy_1) \\ \mathbb{P}_x[(X_1, \dots, X_n) \in A_1 \times \dots \times A_n] &= \int_{A_1} \dots \int_{A_{n-1}} P(y_{n-1}, A_n) P(x, dy_1) P(y_1, dy_2) \dots P(y_{n-2}, dy_{n-1}) \end{aligned}$$

Si se denota $P^1(x, A) = P(x, A)$, para $n > 1$ se puede definir el kernel de n transiciones de probabilidad dado por

$$P^n(x, A) = \int_{\mathcal{X}} p^{n-1}(y, A) P(x, dy)$$

Definición 1.3 (Ecuaciones de Chapman-Kolmogorov).

Para cualesquiera $m, n \in \mathbb{N}, x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$

$$p^{m+n}(x, A) = \int_{\mathcal{X}} P^n(y, A) P^m(x, dy).$$

Para el caso de estados discretos las ecuaciones de Chapman-Kolmogorov pueden ser interpretadas como productos matriz de transición. Sin embargo en el caso general se debe considerar a P como un operador

$$Ph(x) = \int_{\mathcal{X}} h(y)P(x, dy)$$

con $h \in L_1(\lambda)$ para λ una medida tal que $P(x, \cdot) \sim \lambda$.

Definición 1.4 (Propiedad débil de Markov).

Sea π una distribución inicial, para el estado inicial X_0 , entonces

$$\mathbb{E}_\pi[h(X_{n+1}, X_{n+2}, \dots) | X_0 = x_0, X_1 = x_1, \dots, X_n = x_n] = \mathbb{E}_{x_n}[h(X_1, X_2, \dots)]$$

dado que las esperanzas existan.

Definición 1.5 (Tiempo de paro).

Sea $A \in \mathcal{B}(\mathcal{X})$. El primer tiempo " n " para el cual una cadena de Markov entra al conjunto A es llamado el tiempo de paro en A y es denotado por

$$\tau_A = \inf\{n \geq 1 : X_n \in A\}.$$

Definición 1.6 (Propiedad fuerte de Markov). Sea π una distribución inicial, para el estado inicial X_0 , y ζ un tiempo de paro casi seguramente finito

$$\mathbb{E}_\pi = [h(X_{\zeta+1}, X_{\zeta+2}, \dots) | X_0 = x_0, X_1 = x_1, \dots, X_\zeta = x_\zeta] = \mathbb{E}_{x_\zeta}[h(X_1, X_2, \dots)]$$

dado que las esperanzas existan.

Definición 1.7 (Medida invariante y estacionariedad).

Se dice que una medida $\pi(\sigma\text{-finita})$ es invariante para un kernel de transición de probabilidad P si

$$\pi(B) = \int_{\mathcal{X}} P(x, B)\pi(dx), B \in \mathcal{B}(\mathcal{X}).$$

Si una medida invariante π es una medida de probabilidad (integra 1), entonces se dice que π es la distribución de probabilidad invariante.

Ejemplo 1.1 (Auto-regresivo AR(1)).

Sea $\{\epsilon_i\}_{i=1}^{\infty} \stackrel{i.i.d.}{\sim} \text{Normal}(0, \sigma^2)$, $\theta, x_0 \in \mathbb{R}$. Se define la cadena auto-regresiva de orden 1 AR(1), como $X = \{X_n\}_{n=1}^{\infty}$ tales que $X_0 = x_0$ y para $n \geq 1$

$$X_{n+1} = \theta X_n + \epsilon_n.$$

Para $\theta < 1$ la cadena es ergódica y tiene distribución estacionaria $\text{Normal}\left(0, \frac{\sigma^2}{1-\theta^2}\right)$.

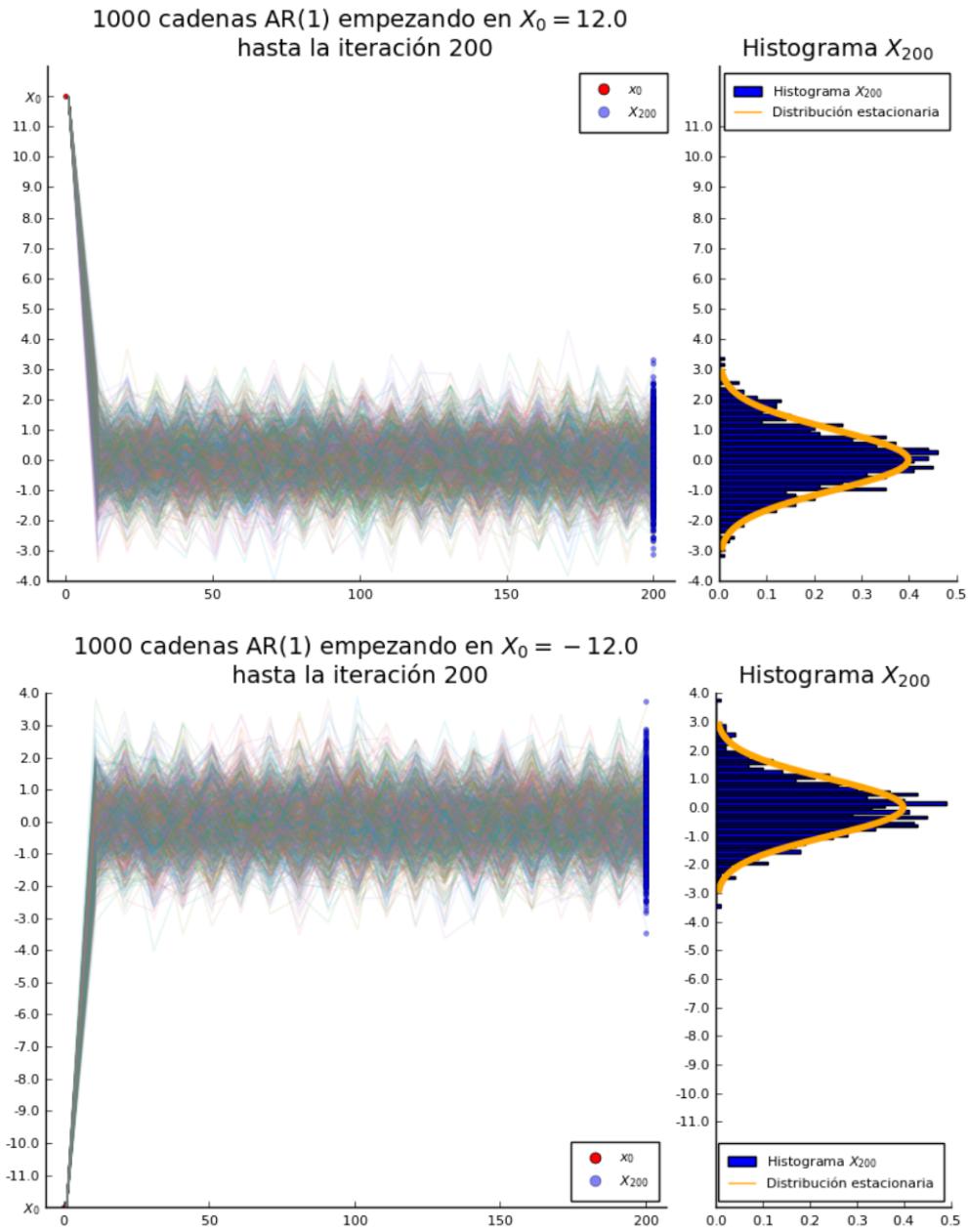


Figura 1: Convergencia con periodo de quemado y distribuciones de 1000 series $AR(1)$ para distintos puntos iniciales x_0 ; notese como ambas simulaciones convergen a la misma distribución.

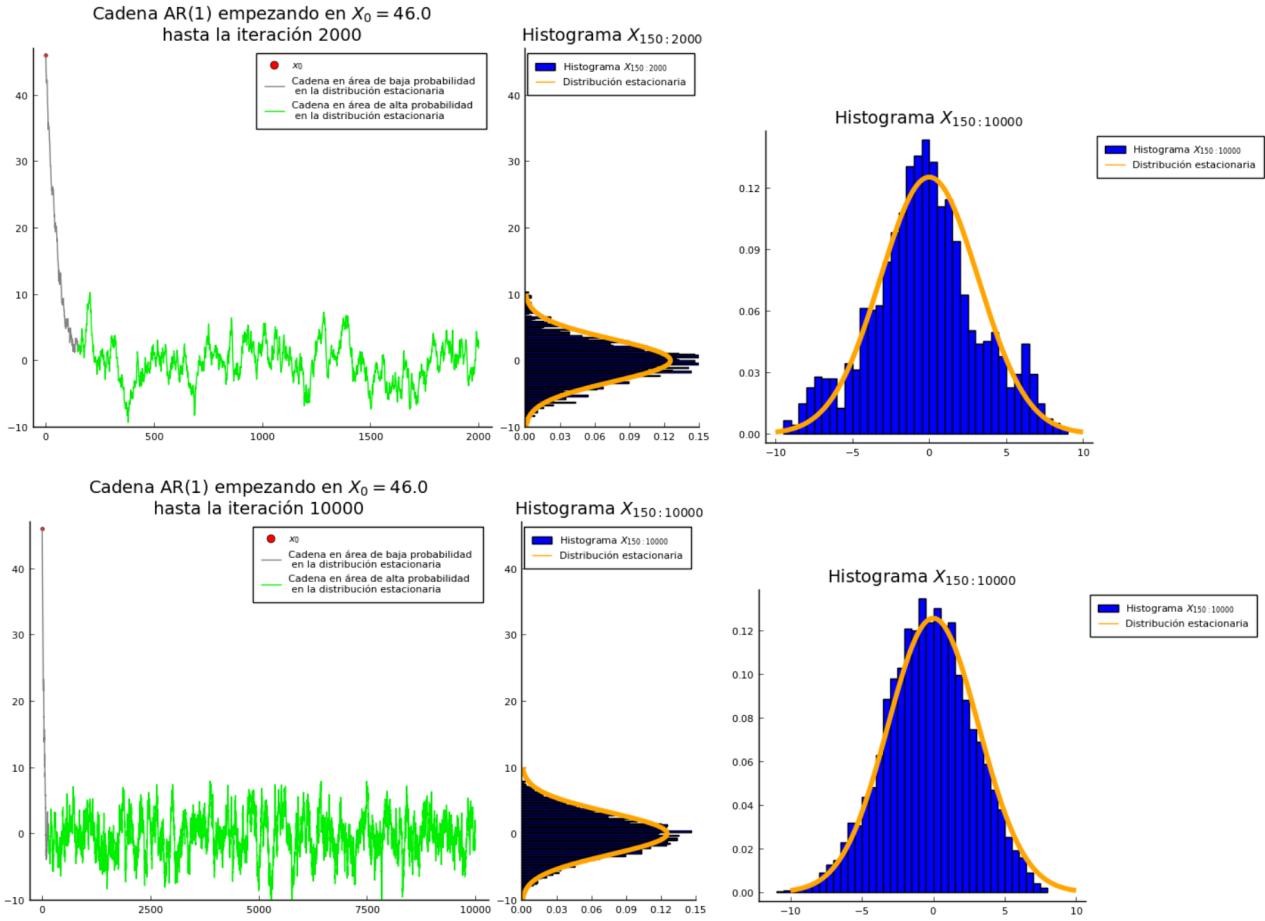


Figura 2: Convergencia con periodo de quemado con distintas iteraciones de la serie $AR(1)$; note la afinación en la distribución conforme se incrementa el número de iteraciones.

Definición 1.8 (Reversibilidad).

Se dice que una cadena de Márkov $\{X_n\}_{n=0}^{\infty}$ es reversible si la distribución de X_{n+1} condicionada en $X_{n+2} = x$ es igual a la distribución de X_{n+1} condicionada en $X_n = x$.

Definición 1.9 (Balance detallado).

Dado un kernel de transición de probabilidad $P(x, dy) = p(x, y)dy$, se dice que la condición de balance detallado se cumple si existe una función π tal que

$$\pi(x)p(x, y) = \pi(y)p(y, x).$$

Si no hay problemas con la definición de probabilidades condicionales dadas por el kernel de transición de probabilidad P , cuando π es una función de densidad de probabilidad reversible utilizando a π como distribución inicial, es equivalente a la condición de balance detallado.

Teorema 1.1. Si un kernel de transición de probabilidad P satisface balance detallado con π una densidad de probabilidad, entonces π es la densidad invariante de la cadena de Markov asociada a P .

Gracias al teorema anterior, una manera de crear cadenas markovianas estacionarias puede consistir de la construcción de cadenas reversibles. El caso más popular es el algoritmo de Metropolis-Hastings.

1.3. Algoritmo de Metropolis-Hastings

Definición 1.10 (Algoritmo de Metropolis-Hastings).

Sean f_k el kernel de una función de densidad de probabilidad y $Q(x, dy) \propto q(x, y)dy$ un kernel de transición de probabilidad, dado X_{n-1}

1. Simular $Y_n \sim \mathcal{L}(Q(X_{n-1}, \cdot))$ y $U \sim Unif(0, 1)$.
2. Aceptar $X_n = Y_n$ si se cumple que

$$U \leq \alpha(X_{n-1}, Y_n) = \min \left\{ 1, \frac{f_k(Y_n)q(Y_n, X_{n-1})}{f_k(X_{n-1})q(X_{n-1}, Y_n)} \right\};$$

en caso contrario $X_n = X_{n-1}$.

Si en α el denominador $f(x)q(x, y) = 0$ entonces por convención, para evitar ambigüedades, se define $\alpha = 1$.

Proposición 1.1. El algoritmo de Metropolis-Hastings genera una cadena de Markov reversible respecto a f .

Obsérvese que se debe de poder generar cadenas con kernel de transición de probabilidad Q aunque para la evaluación del paso de aceptación no es necesario conocer la constante de normalización de la correspondiente distribución condicional. Más aún solo la evaluación de f_k es necesaria para el paso de aceptación por lo que la constante de normalización de f tampoco debe de ser conocida.

En ocasiones resulta numéricamente conveniente utilizar el paso de aceptación equivalente:

$$\begin{aligned} \log(U) &\leq \min\{0, \log(f_k(Y_n)) + \log(q(Y_n, X_{n-1})) - \log(f_k(X_{n-1})) - \log(q(X_{n-1}, Y_n))\} \\ &\Leftrightarrow U \leq \min \left\{ 1, \frac{f_k(Y_n)q(Y_n, X_{n-1})}{f_k(X_{n-1})q(X_{n-1}, Y_n)} \right\} \end{aligned}$$

(Código 1).

1.3.1. Elección de q

- **Metropolis-Hastings simétrico:** Si $q(x, y) = q(y, x)$ entonces la probabilidad de aceptación se reduce a

$$\alpha(x, y) = \min \left\{ 1, \frac{f_k(y)}{f_k(x)} \right\}.$$

- **Metropolis-Hastings con caminata aleatoria:** Se denomina así cuando $q(x, y) = q(y - x)$. Por ejemplo con $Q(x, \cdot) = Unif(x - 1, x + 1)$.

- **Algoritmo de Langevin:** Se denomina cuando para $\delta > 0$ pequeño se tiene que

$$Y_{n+1} \sim Normal\left(X_n + \frac{\delta}{2} \frac{d}{dx}(\log(f(x)))|_{x=X_n}, \delta\right).$$

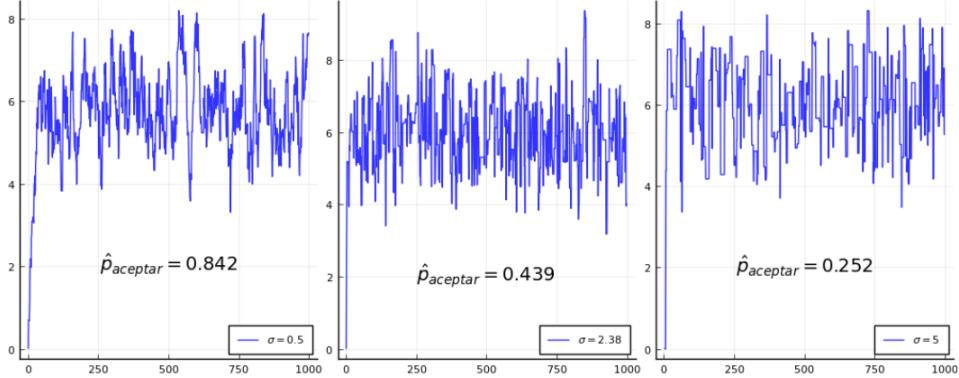


Figura 3: Cadena de Metropolis-Hastings con caminatas aleatorias $Normal(\mu, \sigma)$ para converger a una estacionaria $Normal(6, 1)$.

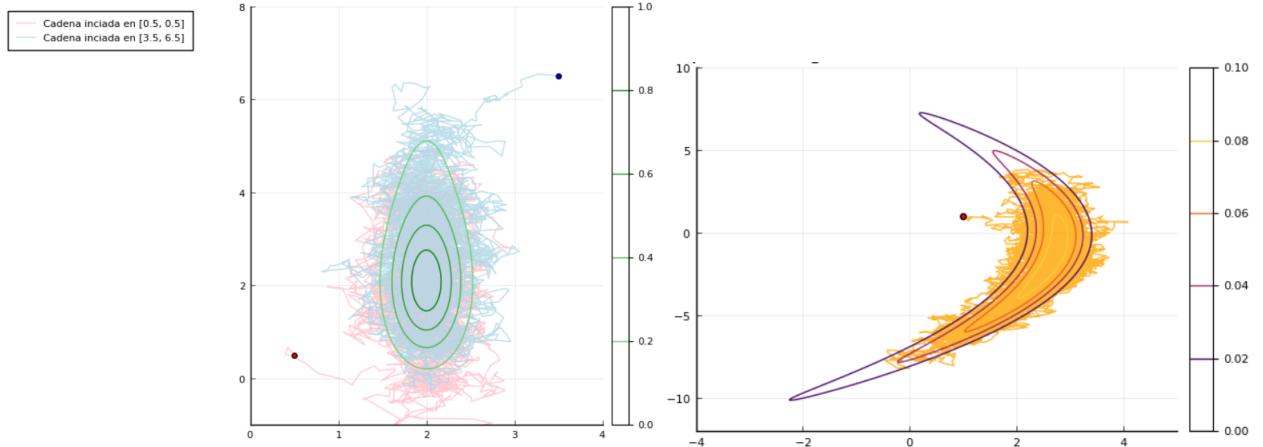


Figura 4: Cadenas de Metropolis-Hastings que convergen a distribuciones con forma de aguacate y de banana.

Una manera para construir algoritmos de Metropolis-Hastings en cierto subconjunto $I \subset \mathbb{R}$ consiste en considerar el algoritmo de Metropolis-Hastings en \mathbb{R} y una transformación $g : \mathbb{R} \rightarrow I$ invertible.

Ejemplo 1.2. Considérese el caso de Metropolis-Hastings con caminata aleatoria gaussiana

$$\tilde{q}_k(z, w) = e^{-(z-w)^2/(2\sigma^2)}.$$

Dado x el estado de la cadena se propone

$$y = g(g^{-1}(x) + \epsilon), \epsilon \sim Normal(0, \sigma^2)$$

con lo que

$$q_k(x, y) = \tilde{q}_k(g^{-1}(x), g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

Obsérvese que explotando la simetría del kernel \tilde{q}

$$\frac{q_k(y, x)}{q_k(x, y)} = \frac{\frac{d}{dx} g^{-1}(x)}{\frac{d}{dy} g^{-1}(y)}$$

la siguiente tabla facilita las equivalencias necesarias para distribuciones exponenciales restringidas a intervalos:

I	$g(x)$	$g^{-1}(x)$	$dg^{-1}(x)$	$\frac{q_k(y,x)}{q_k(x,y)}$
$(0, \infty)$	$\exp(x)$	$\log(x)$	x^{-1}	$\frac{y}{x}$
(a, ∞)	$\exp(x) + a$	$\log(x - a)$	$\frac{1}{x-a}$	$\frac{y-a}{x-a}$
$(-\infty, a)$	$a - \exp(x)$	$\log(a - x)$	$-\frac{1}{a-x}$	$\frac{y-x}{a-x}$
$[0, 1]$	$\frac{\exp(x)}{\exp(x)+1}$	$\log\left(\frac{x}{1-x}\right)$	$\frac{1}{x(1-x)}$	$\frac{y(1-y)}{x(1-x)}$
$[a, b]$	$\frac{b\exp(x)+a}{\exp(x)+1}$	$\log\left(\frac{x-a}{b-x}\right)$	$\frac{b-a}{(b-x)(x-a)}$	$\frac{(b-y)(y-a)}{(b-x)(x-a)}$

Figura 5: Inversas, derivadas y cocientes para exponenciales restringidas a intervalos.

1.3.2. Metropolis-Hastings adaptativo

Definición 1.11 (Metropolis-Hastings adaptativo).

Dados X_{n-1}, μ_{n-1} y Γ_{n-1}

1. Simular $Y_n \sim N(\mu_{n-1}, \Gamma_{n-1})$ y $U \sim Unif(0, 1)$.
2. Aceptar $X_n = Y_n$ si se cumple que

$$U \leq \alpha(X_{n-1}, Y_n) = \min \left\{ 1, \frac{f_k(Y_n)}{f_k(X_{n-1})} \right\};$$

en caso contrario $X_n = X_{n-1}$.

3. Actualizar

$$\mu_n = \mu_{n-1} + \frac{1}{n}(X_n - \mu_{n-1})$$

y

$$\Gamma_n = \Gamma_{n-1} + \frac{1}{n}((X_n - \mu_{n-1})(X_n - \mu_{n-1})' - \Gamma_{n-1})$$

1.3.3. Mezcla de cadenas

Si P_1 y P_2 son kernels de transición de probabilidad, ambos con distribución estacionaria f entonces la cadena dada por el kernel de transición de probabilidad P_1P_2 también tiene distribución estacionaria f ; por lo que en la práctica es posible mezclar cadenas para generar algoritmos de MCMC.

Obsérvese que aunque cada P_1 y P_2 sea reversible para f , la cadena de Márkov dada por P_1P_2 no será reversible en general. Por esa razón es importante no restringirse al caso reversible en el estudio de las cadenas de Márkov.

1.4. Muestreo de Gibbs

Definición 1.12 (Muestreo de Gibbs).

Sea $\mathcal{X} \subset \mathbb{R}^d$ abierto. El muestreador de Gibbs en la entrada $i, 1 \leq i \leq d$, para un vector $x \in \mathcal{X}$ está dado por un kernel de transición de probabilidad que sólo modifica el valor de la entrada i y deja al resto de las entradas de x sin modificación. El valor x_i es generado con la distribución condicional $f_{X_i|X_1=x_1, \dots, X_{i-1}=x_{i-1}, X_{i+1}=x_{i+1}, \dots, X_n=x_n}$, es decir

$$P_i(x_{-1}, S_{x_{-i}, i, a, b}) = \frac{\int_a^b f_k(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n) dz}{\int_{-\infty}^{\infty} f_k(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n) dz}$$

con $X_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ y $S_{x_{-i}, i, a, b} = \left\{ y \in \mathcal{X} : y_j = x_j \text{ para } j \neq i, a \leq y_i \leq b \right\}$.

Se construyen así los muestradores de Gibbs al mezclar los Kernels de P_i :

- **Muestrador de Gibbs con escaneo determinístico:** $P = P_1 P_2 \cdots P_d$.
- **Muestrador de Gibbs con escaneo aleatorio:** $P = \frac{1}{d} \sum_{i=1}^d P_i$.

Ejemplo 1.3. Considérese el modelo

$$\begin{aligned}\frac{1}{\sigma^2} &\sim \text{Gamma}(\alpha, \beta) \\ \mu | \sigma^2 &\sim \text{Normal}(\mu_0, \frac{\sigma^2}{v}) \\ X_i &\sim \text{Normal}(\mu, \sigma^2), i \in \{1, \dots, n\}\end{aligned}$$

con

$$\begin{aligned}m\mu_n &= \frac{n\bar{x} + v\mu_0}{v + n} \\ v &= v + n \\ \alpha &= \alpha + \frac{n}{2} \\ \beta &= \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{vn(\bar{x} - \mu_0)^2}{2(v + n)}\end{aligned}$$

se tiene que

$$\mu | \sigma^2 \sim \text{Normal}\left(\mu_n, \frac{\sigma^2}{v_n}\right)$$

y

$$\sigma^2 | \mu, X \sim \Gamma^{-1}\left(\alpha_n + 0.5, \frac{2\beta n + v_n(\mu - \mu_n)^2}{2}\right)$$

por lo que se puede implementar un muestreador de Gibbs;

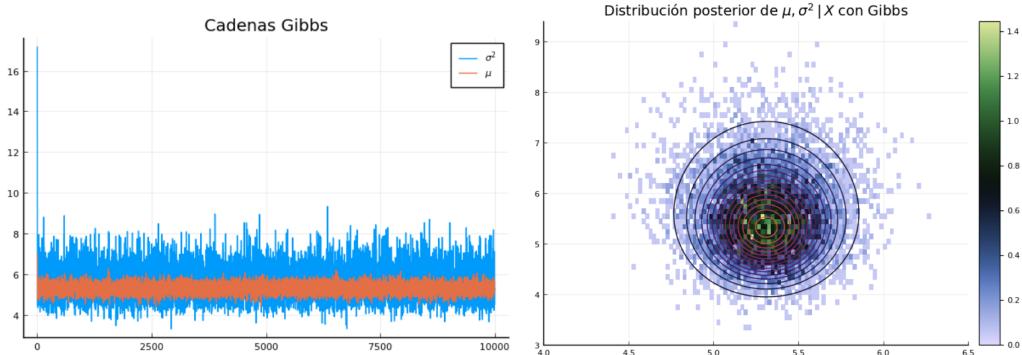


Figura 6: Muestreo de Gibbs para la distribución *Normal* descrita previamente (Código 2).

En el modelo considerado se pueden simular directamente los parámetros dado que se conocen las posteriores previas y las respectivas marginales

$$\mu | X \sim t_{2\alpha_n} \left(\cdot | \mu_n, \frac{\beta_n}{\alpha_n v_n} \right)$$

y

$$\sigma^2 | X \sim \Gamma^{-1}(\alpha_n, \beta_n).$$

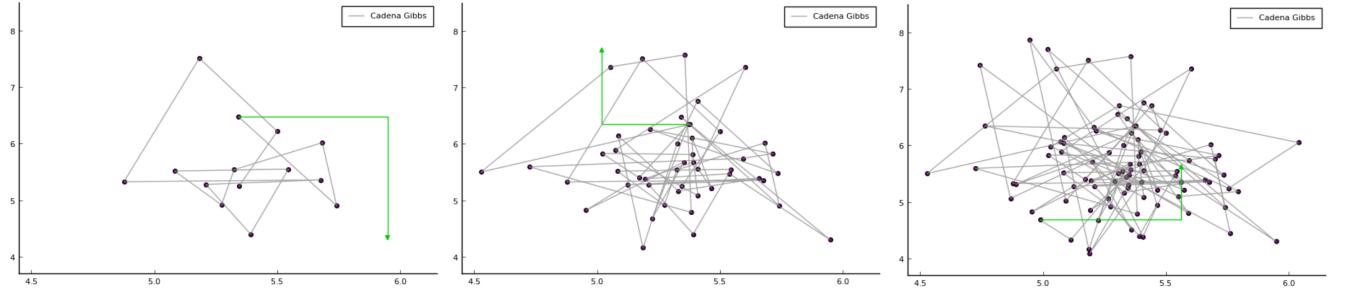


Figura 7: Secuencia de simulaciones de los parámetros μ, σ^2 mediante la cadena de Gibbs.

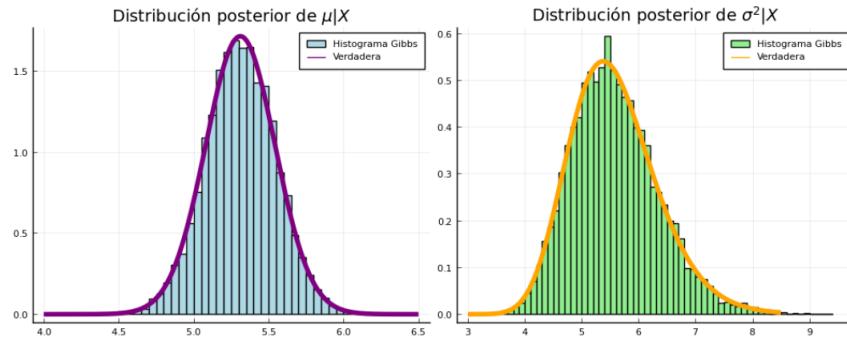
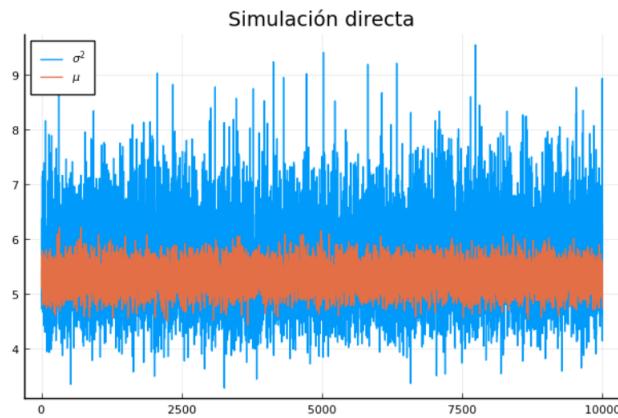


Figura 8: Distribuciones posteriores de los parámetros $\mu, \sigma^2|X$ mediante la cadena de Gibbs.



Ejemplo 1.4 (Mixturas Gaussianas).

Considérese el modelo

$$\frac{1}{\sigma_j^2} \sim \text{Gamma}(\alpha, \beta); \mu_j | \sigma_j^2 \sim \text{Normal}\left(\mu_0, \frac{\sigma_j^2}{v}\right), j \in \{1, \dots, d\}$$

$$\begin{aligned} \Pi &\sim \text{Dirichlet}(\alpha_1, \dots, \alpha_d) \\ Z_i &\sim \text{Categorico}(\Pi), i \in \{1, \dots, n\} \\ X_i | Z_i = j, \mu, \sigma^2 &\sim \text{Normal}\left(\mu_j, \sigma_j^2\right), i \in \{1, \dots, n\} \end{aligned}$$

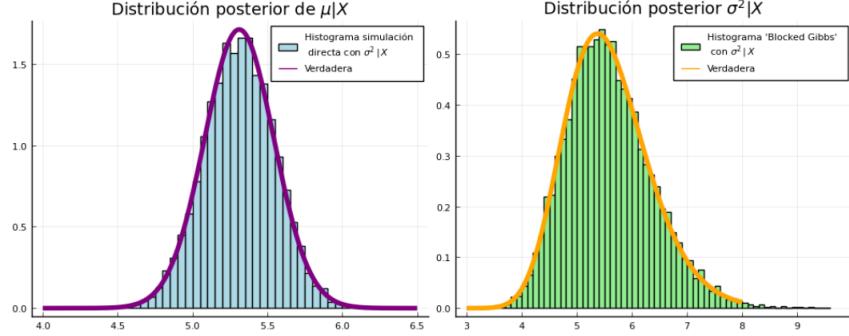


Figura 10: Secuencia de simulaciones de los parámetros μ, σ^2 y distribuciones posteriores mediante la cadena de Gibbs (Código 3).

Se tiene que

$$\mathbb{P}[X_i = x | \pi, \mu, \sigma^2] = \sum_{j=1}^d \pi_j f_{Normal(\mu_j, \sigma_j^2)}(x)$$

es una mixtura de gaussianas. Obsérvese que

$$\begin{aligned} f(\pi | X = x, z, \mu, \sigma^2) &= f(\pi | z) \propto f_{Dir(\alpha)}(\pi) \prod_{i=1}^n f_\pi(z_i) \\ &\propto \prod_{j=1}^d \pi_j^{\alpha_j - 1} \prod_{i=1}^n f_\pi(z_i) \stackrel{\hat{n}_j = \#\{i : z_i = j\}}{=} \prod_{j=1}^d \pi_j^{\alpha_j + \hat{n}_j - 1} \propto f_{Dir(\alpha + \hat{n})}(\pi) \\ f(z_i | X = x, z_{-i}, \pi, \mu, \sigma^2) &= f(z_i | X_i = x_i, \pi, \mu, \sigma^2) \\ &\propto f_\pi(z_i) f_{normal(\mu_{z_i}, \sigma_{z_i}^2)}(x_i) = \pi_{z_i} f_{normal(\mu_{z_i}, \sigma_{z_i}^2)}(x_i) \\ f(\mu_j, \sigma_j^2 | X = x, \pi, \mu, \sigma^2) &= f(\mu_i, \sigma_i^2 | X'_i s : z_i = j) \end{aligned}$$

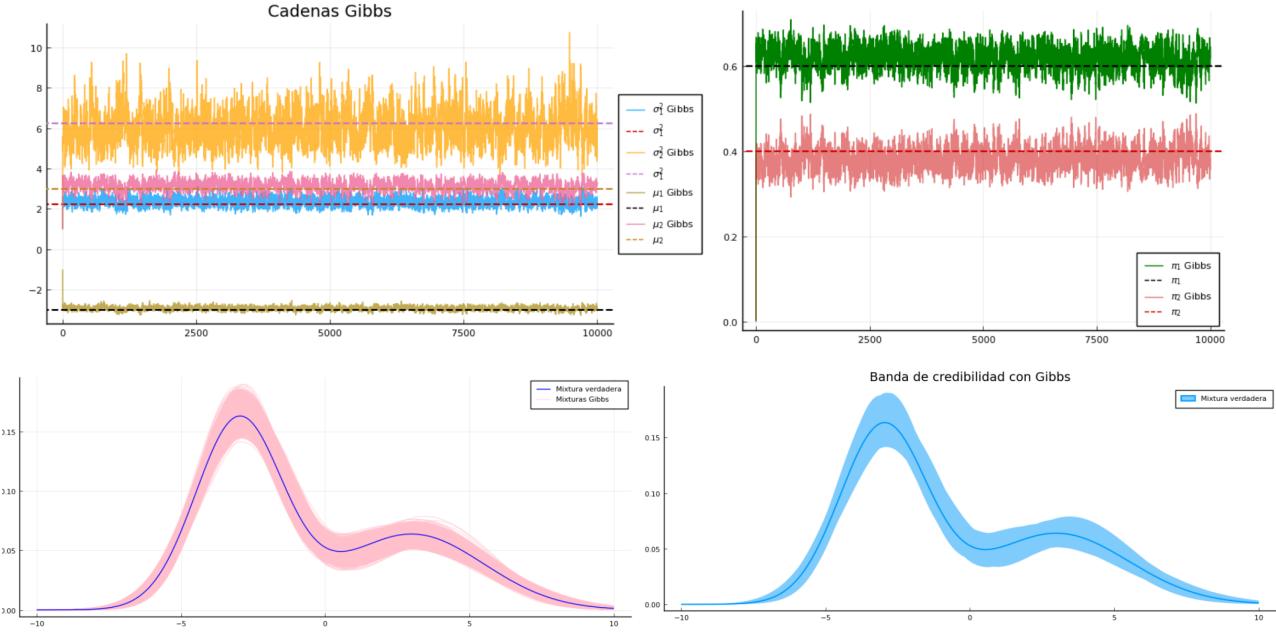


Figura 11: Convergencias de los parámetros de la mixtura y ajuste de la distribución generada a la mixtura real.

Ejemplo 1.5 (Regresión Poisson).

Supóngase que se tienen muestras $(Y_1, X_1), \dots, (Y_n, X_n)$ tales que

$$Y_i | a, b, X_i \sim \text{Poisson}(e^{aX_i + b})$$

$$a \sim \text{Normal}(0, \sigma^2)$$

$$b \sim \text{Normal}(0, \tau^2)$$

Se tiene entonces una verosimilitud dada por

$$\pi(a, b | y, x) \propto \pi(y | x, a, b) \pi(a) \pi(b) \propto \exp\left(a \sum_{i=1}^n y_i + b \sum_{i=1}^n y_i x_i - e^a \sum_{i=1}^n e^{x_i b} - \frac{a^2}{2\sigma^2} - \frac{b^2}{2\tau^2}\right)$$

Entonces

$$\log(\pi(a | y, x, b)) = a \sum_{i=1}^n y_i - e^a \sum_{i=1}^n e^{x_i b} - \frac{a^2}{2\sigma^2}$$

$$\log(\pi(b | y, x, a)) = b \sum_{i=1}^n y_i x_i - e^a \sum_{i=1}^n e^{x_i b} - \frac{b^2}{2\tau^2}$$

y se puede utilizar aceptación rechazo adaptativo para implementar el muestreador de Gibbs.

1.5. Monte Carlo Hamiltoniano (HMC)

La idea principal es construir una distribución propuesta en el paso de Metropolis-Hastings de tal forma que

1. Se conserven las condiciones de balance detallado.
2. Se utilice la geometría de la distribución a muestrear para construir la distribución de propuesta.

3. La distribución de propuesta explore flexiblemente el soporte de la distribución a muestrear.
4. La distribución de propuesta sea simétrica.

Para tales motivos se hará uso de una variable auxiliar $p \in \mathbb{R}^d$ con d la dimensión de la distribución a samplear x y una dinámica determinista llamada hamiltoniana la cual será utilizada dentro de un algoritmo de Metropolis-Hastings.

Sea $H : \mathbb{R}^{2d} \rightarrow \mathbb{R}$ (que se denominará Hamiltoniano), se define el sistema de ecuaciones diferenciales

$$\begin{aligned}\frac{dx_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial x_i}\end{aligned}$$

para $i \in \{1, 2, \dots, d\}$. Estos sistemas resultan importantes ya que

$$\frac{dH}{dt} = \sum_{i=1}^d \left(\frac{\partial H}{\partial x_i} \frac{dx_i}{dt} + \frac{\partial H}{\partial p_i} \frac{dp_i}{dt} \right) = \sum_{i=1}^d \left(\frac{\partial H}{\partial x_i} \frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial x_i} \right) = 0$$

Ejemplo 1.6.

Considérese

$$H(x, p) = \frac{x^2}{2} + \frac{p^2}{2}$$

con lo que se tiene la dinámica

$$\begin{aligned}\frac{dx}{dt} &= \frac{\partial H}{\partial p} = p \\ \frac{dp}{dt} &= -\frac{\partial H}{\partial x} = -x\end{aligned}$$

cuya solución está dada por

$$x(t) = r\cos(a + t), p(t) = -r\sin(a + t)$$

$$\text{Hamiltoniano } H(x, p) = 0.5(x^2 + p^2)$$

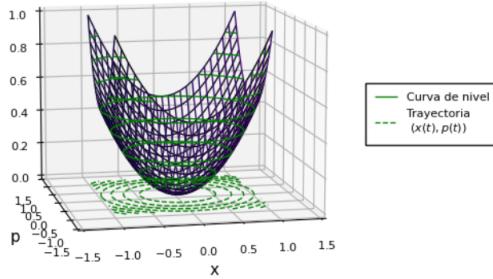


Figura 12: Hamiltoniano con respectivas curvas de nivel.

Para el algoritmo HMC se hará uso de hamiltonianos con la forma

$$H(x, p) = U(q) + K(p).$$

A la función U se le llamará la energía potencial y a K la energía cinética, con lo que H puede ser pensada como energía total y definir la densidad de probabilidad

$$P(x, p) = \frac{1}{M} e^{-H(x, p)} = \frac{1}{M} e^{-U(x) - K(p)}$$

de tal forma que $x \perp\!\!\!\perp p$. Si se quiere que $x \sim \mathcal{L}(f)$ entonces se toma $U(x) = -\log(f(x))$ y para la energía cinética usualmente se utiliza $K(p) = p' M^{-1} p / 2$, aún más se tomará $M = D$ diagonal, $D_{i,i} = d_i$, con lo que $K(p) = \sum_{i=1}^n \frac{p_i^2}{2d_i}$.

Ejemplo 1.7.

Si se considera $U(x) = x^2/2$ y $K(p) = p^2/2$ entonces

$$P(x, p) \propto e^{-x^2/2} e^{-p^2/2}$$

con lo que se definen variables aleatorias $X \sim \text{Normal}(0, 1)$ y $\mathcal{P} \sim \text{Normal}(0, 1)$.

Definición 1.13 (Algoritmo Monte Carlo Hamiltoniano).

Dados valores previos X_{n-1} y \mathcal{P}_{n-1}

1. Simular $\tilde{\mathcal{P}}_n \sim \mathcal{L}\left(\frac{1}{M_k e^{-K(p)}}\right)$ (usualmente Gaussiana multivariada).
2. Proponer Y_n, G_n aproximando numéricamente la dinámica del hamiltoniano asociando empezando en $(X_{n-1}, \tilde{\mathcal{P}}_n)$.
3. Simular $V \sim \text{Unif}(0, 1)$ y aceptar $X_n = Y_n, \mathcal{P}_n = G_n$ si se cumple que

$$V < \min \left\{ 1, e^{-H(Y_n, G_n) + H(X_{n-1}, \tilde{\mathcal{P}}_n)} \right\}$$

en caso contrario $X_n = X_{n-1}$.

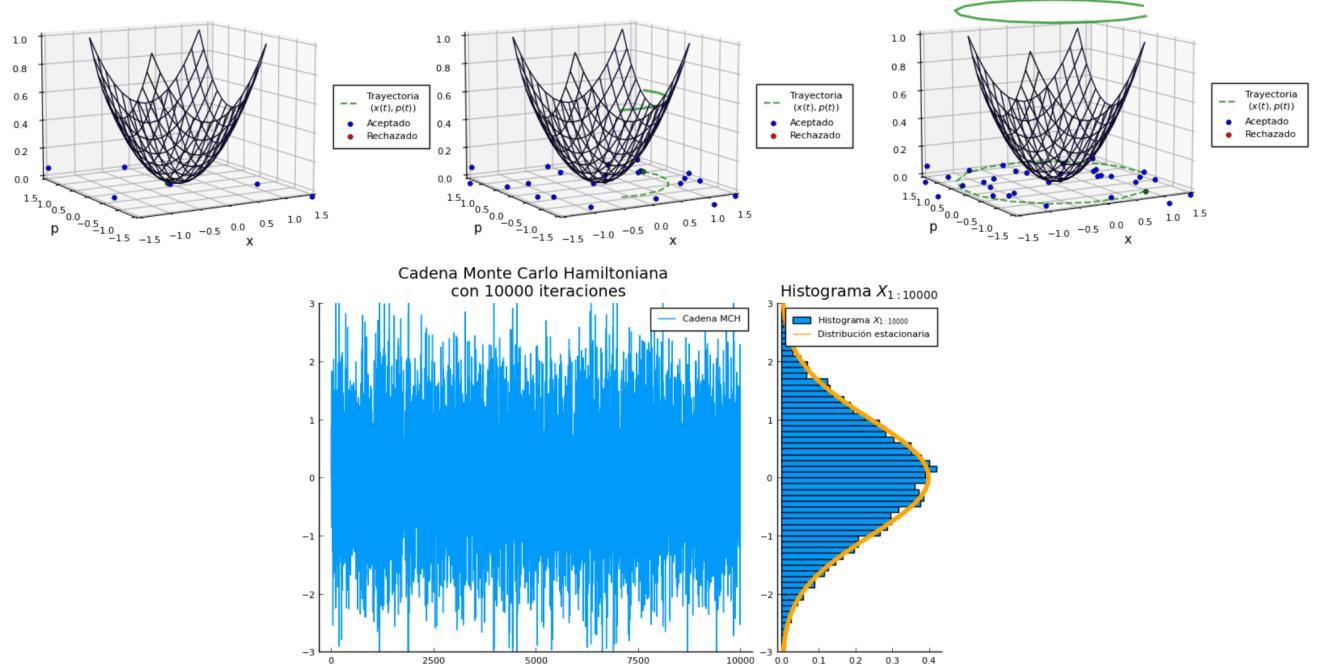


Figura 13: Muestreo de datos con Monte Carlo Hamiltoniano con esquema de Euler.

Definición 1.14 (Esquema de Euler).

Para $p_0, x_0 \in \mathbb{R}^d$ al tiempo t_0 y $\epsilon > 0$ se genera una malla de tamaño m dada por

$$1. p_{i,t_j} = p_{i,t_{j-1}} - \epsilon \frac{\partial U}{\partial x_i}(x(t_{j-1})) = \left(p_{i,t_{j-1}} = +\epsilon \frac{dp_i}{dt}(t_{j-1}) \right)$$

$$2. x_{i,t_j} = x_{i,t_{j-1}} - \epsilon \frac{\partial K}{\partial p_i}(p(t_{j-1})) = \left(x_{i,t_{j-1}} = +\epsilon \frac{dx_i}{dt}(t_{j-1}) \right)$$

con $t_j = t_{j-1} + \epsilon, j = 1, \dots, m, i = 1, \dots, d$.

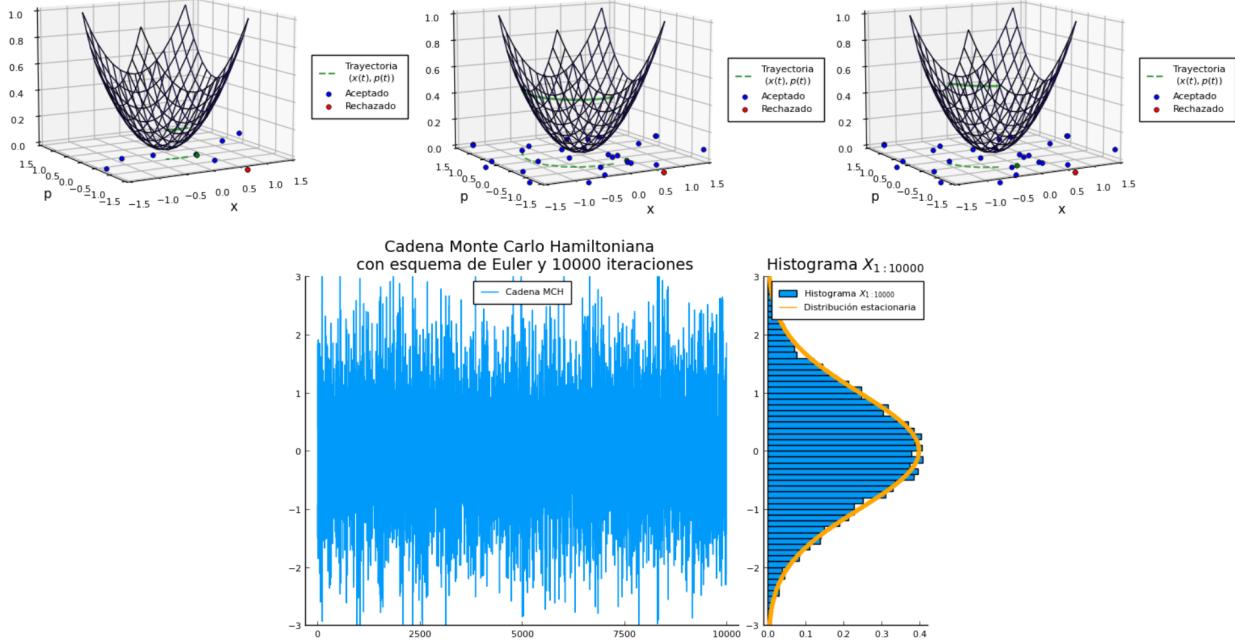


Figura 14: Muestreo de datos con Monte Carlo Hamiltoniano con esquema de Euler.

El método de Euler no suele preservar el volumen de un sistema dinámico por lo que no será usado para el Monte Carlo Hamiltoniano. Para esto se presenta el método de Leapfrog (o salto de rana) que está construido de forma que el volumen se preserve exactamente.

Definición 1.15 (Método de Leapfrog).

Para $p_0, x_0 \in \mathbb{R}^d$ al tiempo t_0 y $\epsilon > 0$ se genera una malla de tamaño m dada por

$$1. p_{i,t_j,\epsilon/2} = p_{i,t_{j-1}} - \frac{\epsilon}{2} \frac{\partial U}{\partial x_i}(x(t_{j-1}))$$

$$2. x_{i,t_j} = x_{i,t_{j-1}} - \frac{\epsilon}{2} \frac{\partial K}{\partial p_i}(p(t_{j-1} + \frac{\epsilon}{2}))$$

$$3. p_{i,t_j} = p_{i,t_{j-1},\epsilon/2} - \frac{\epsilon}{2} \frac{\partial U}{\partial x_i}(x(t_{j-1}))$$

con $t_j = t_{j-1} + \epsilon, j = 1, \dots, m, i = 1, \dots, d$.

Definición 1.16 (MCH con NUTS (no u-turns)).

Se debe de poner cuidado en la elección del tiempo que se recorre la curva de nivel del hamiltoniano en el algoritmo MCH. Para los esquemas de Euler y Leapfrog este tiempo de recorrido está dado por $t = \epsilon L$ donde L es el número de elementos en la malla y ϵ es el tamaño del paso asociados. Para evitar que al estar recorriendo la curva de nivel del hamiltoniano se comience a regresar hacia el punto inicial se hará uso de la siguiente cantidad:

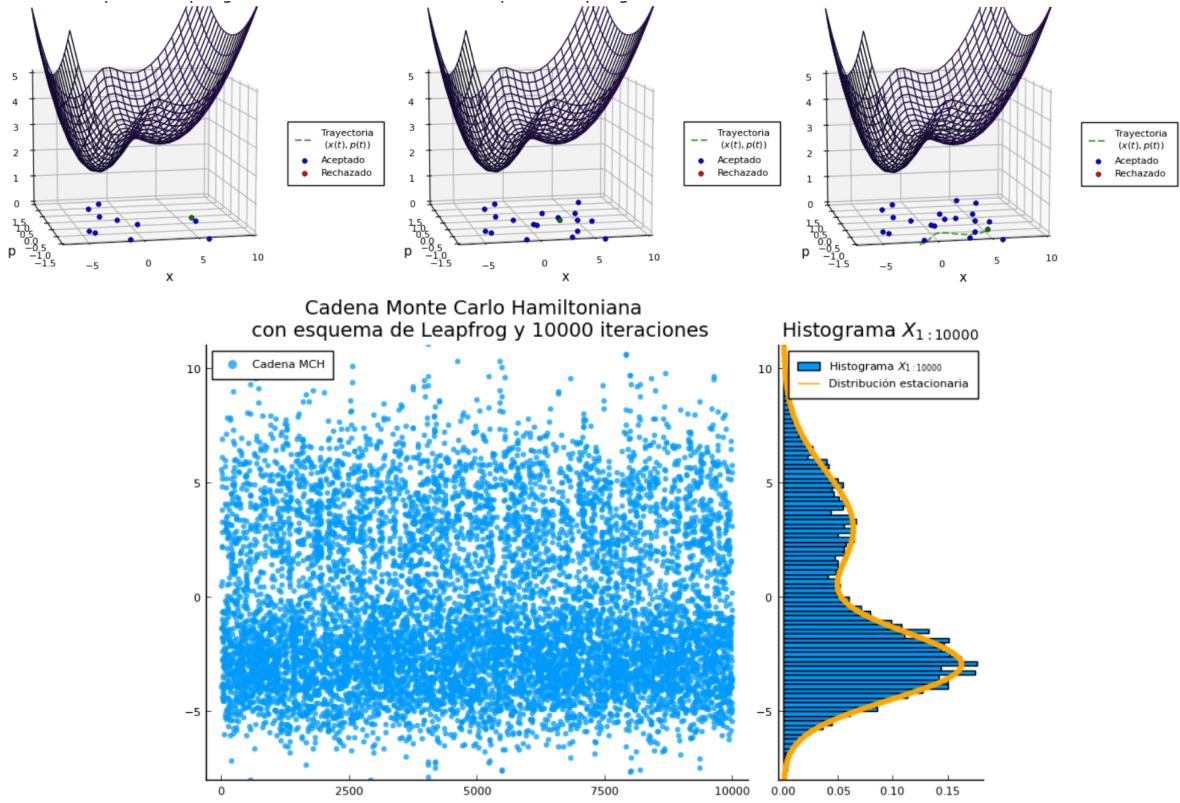


Figura 15: Muestreo de datos con Monte Carlo Hamiltoniano con esquema de Leapfrog.

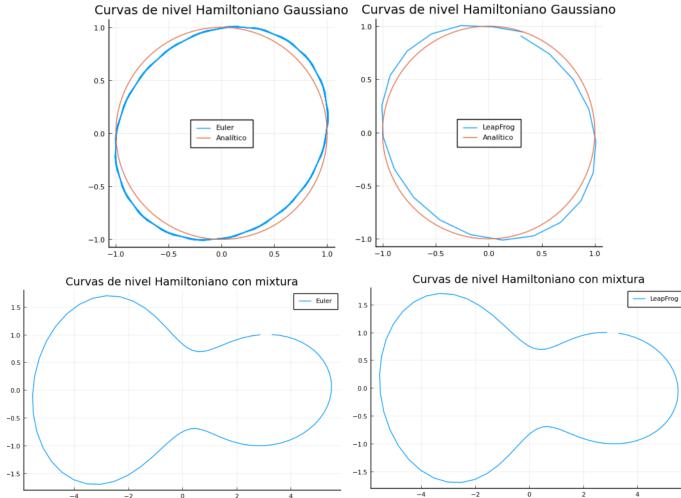


Figura 16: Comparación de curvas de nivel para esquemas de Euler y Leapfrog.

$$\frac{d}{dt} \left(\frac{(x(t) - x_0) \cdot (x(t) - x_0)}{2} \right) = (x(t) - x_0) \cdot \frac{d}{dt}((x(t) - x_0)) = (x(t) - x_0) \cdot p(t)$$

la cual es proporcional al cambio infinitesimal de la distancia entre el punto final de nuestra trayectoria y el punto inicial por lo que para evitar "vueltas en U" nos hay que detenerse cuando ésta es negativa.

La primera manera en que se implementó esta idea en MCH fue mediante un muestreo por rebanadas con doubling en el cual se detiene el doblamiento de la trayectoria cuando el producto interior anterior es negativo. Posteriormente se simplificó esto al usar muestreo multinomial en vez de la construcción de una rebanada para simular un estado dentro de la trayectoria construida con un método de doubling.

Definición 1.17 (MCH con elección adaptativa de ϵ).

Para elegir ϵ de manera adaptativa en el algoritmo de MCH se hace uso de un esquema de aproximación estocástica. Uno de los algoritmos más usados de adaptación estocástica es el algoritmo de Robbins-Monro donde se quiere encontrar la raíz de una función h (es decir encontrar θ^* tal que $h(\theta^*) = 0$) para lo cual sólo se tiene la evaluación de una sucesión aleatoria H_n tal que

$$h(\theta) = \lim_{m \rightarrow \infty} \sum_{n=1}^m \frac{1}{m} \mathbb{E}[H_n | \theta]$$

si h es no-decreciente, bajo condiciones de regularidad se tiene que

$$\theta_{n+1} = \theta_n - \eta_n H_n$$

satisface $h(\theta_n) \rightarrow 0$ cuando

$$\sum_{n=1}^{\infty} \eta_n = \infty, \sum_{n=1}^{\infty} \eta_n^2 < \infty$$

(usualmente se toma $\eta_n = n^{-\alpha}$ con $\alpha \in (0.5, 1]$).

Es deseable utilizar el anterior esquema durante el periodo de quemado de un algoritmo MCMC para encontrar el valor de algún parámetro θ de tal forma que pueda ser fijado en la cadena después del periodo de quemado.

Por ejemplo si a_n denota el evento de aceptar en la iteración n e un algoritmo de Metropolis-Hastings y se considera $H_n = -\delta - a_n$ se estarían encontrando valores de θ que ayudan a tener probabilidad δ de aceptación.

Sin embargo en el contexto de MCMC con Robbins-Monro la forma de la sucesión η_n hace que iteraciones iniciales tengan un peso considerable lo cual puede ser problemático al querer una valor de θ relacionado con la dinámica estacionaria y no inicial transitoria de la cadena.

Por lo anterior se usa el siguiente esquema de adaptación estocástica para elegir parámetros en esquemas de MCMC.

Definición 1.18 (Promedio dual de Nésterov).

$$\theta_{n+1} = \mu - \frac{\sqrt{n}}{\gamma(n + n_0)} \sum_{i=1}^n H_i$$

$$\tilde{\theta}_{n+1} = \eta_n \theta_{n+1} + (1 - \eta_n) \tilde{\theta}_n$$

es tal que $\tilde{\theta}_n \rightarrow 0$. Con $\mu \in \mathbb{R}$ y $t_0, \gamma \in (0, \infty)$ parámetros libres a elegir.

1.6. Diagnóstico de convergencia

La distancia de variación total entre 2 medidas de probabilidad v_1 y v_2 está dada por

$$\|v_1(\cdot) - v_2(\cdot)\| = \sup_A |v_1(A) - v_2(A)|$$

Con eso surgen las preguntas

1. ¿ $\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - f(\cdot)\| = 0$?
2. Para $\epsilon > 0$, ¿qué tan grande debe de ser n para que $\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - f(\cdot)\| < \epsilon$?

Proposición 1.2.

1. $\|v_1(\cdot) - v_2(\cdot)\| = \sup_{g: \mathcal{X} \rightarrow [0,1]} |\int g dv_1 - \int g dv_2|$
2. Si f es la distribución estacionaria para un kernel de transición de probabilidad P entonces $\forall n \in \mathbb{N}$

$$\|P^n(x, \cdot) - f(\cdot)\| \leq \|P^{n-1}(x, \cdot) - f(\cdot)\|$$

Una cadena de Márkov con distribución estacionaria f puede no converger a f . Por ejemplo con $\mathcal{X} = \{1, 2, 3\}$, $f = (1/3, 1/3, 1/3)$ y

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

se tiene $fP = f$ sin embargo si $X_0 = 1$ entonces $X_n \in \{1, 2\} \forall n \geq 1$ con lo que $\mathbb{P}[X_n = 3] = 0 \neq \frac{1}{3} = f(\{3\})$. De hecho en este caso la distribución estacionaria no es única y se está convergiendo a la estacionaria $f_2 = (0.5, 0.5, 0)$.

Definición 1.19 (Cadena ϕ -irreducible).

Una cadena de Márkov es ϕ -irreducible si existe una medida ϕ distinta de 0, σ -finita en \mathcal{X} tal que para todo $A \subset \mathcal{X}$ tal que $\phi(A) > 0$ y $x \in \mathcal{X}$ se tiene que existe $n = n(x, A) \in \mathbb{N} \setminus \{0\}$ que satisface

$$P^n(x, A) > 0.$$

Ejemplo 1.8. Para el algoritmo de Metropolis-Hastings con distribución estacionaria dada por una densidad de probabilidad f finita en \mathbb{R}^d , y kernel de transición Q auxiliar con una densidad $q(\cdot, \cdot)$ positiva y continua en $\mathbb{R}^d \times \mathbb{R}^d$ se tiene que la cadena de Márkov asociada al algoritmo es f -irreducible.

Sea A tal que $f(A) > 0$ entonces $\exists r > 0$ tal que $A_r = A \cap B_r(0)$, con $B_r(0)$ la bola con radio r centrada en 0, satisface $f(A_r) > 0$. Por continuidad de q para cualquier $x \in \mathbb{R}^d$

$$\inf_{y \in A_r} \min\{q(x, y), q(y, x)\} \geq \epsilon$$

para algún $\epsilon > 0$. En el caso $f(x) = 0$ se tiene por convención que $\alpha = 1$ por lo que siempre se acepta y se tiene $P(x, A) = Q(x, A) > 0$. Entonces hay que enfocarse en x tal que $f(x) > 0$ con lo que

$$\begin{aligned} P(x, A) &\geq P(x, A_r) \geq \int_{A_r} q(x, y) \min\left\{1, \frac{f(y)q(y, x)}{f(x)q(x, y)}\right\} dy \\ &= \int_{A_r} \min\left\{\frac{f(x)q(x, y)}{f(x)}, \frac{f(y)q(y, x)}{f(x)}\right\} dy \geq \epsilon |\{y \in A_r : f(y) \geq f(x)\}| + \frac{\epsilon}{f(x)} f(\{y \in A_r : f(y) < f(x)\}) \end{aligned}$$

Donde $|C|$ denota el área o medida de Lebesgue del conjunto C . Si ambos términos en la última igualdad son 0 entonces $f \equiv 0$ lo cual es una contradicción. Se concluye que $P(x, A) > 0$ y la cadena de Metropolis-Hastings es f -irreducible.

Incluso cadenas ϕ -irreducibles pueden no converger en distribución debido a problemas de periodicidad. Considérese $\mathcal{X} + \{1, 2, 3\}$, $f = (1/3, 1/3, 1/3)$ y

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

entonces la cadena es ϕ -irreducible para cualquier medida ϕ y el límite del $\mathbb{P}[X_n = 1]$ conforme $n \rightarrow \infty$ no existe por lo que no puede haber estacionariedad.

Definición 1.20 (Cadena aperiódica).

Una cadena de Márkov con distribución estacionaria f es aperiódica si no existe $d \geq 2$ y subconjuntos disjuntos $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n \subset \mathcal{X}$ con $P(x, \mathcal{X}_{i+1}) = 1 \forall \mathcal{X}_i, i \in \{1, 2, \dots, d-1\}$, y $P(x, \mathcal{X}_1) = 1 \forall x \in \mathcal{X}_d$ tal que $f(\mathcal{X}_1) > 0$, y en consecuencia $f(\mathcal{X}_i) > 0$ para $i \in \{2, \dots, d\}$.

Ejemplo 1.9. La cadena de Metropolis-Hastings 1.8 es aperiódica. para probarlo supóngase lo contrario, Si $\mathcal{X}_1, \mathcal{X}_2$ son disjuntos y tales que $P(x, \mathcal{X}_2) = 1 \forall x \in \mathcal{X}_1$ entonces para $x \in \mathcal{X}_1$ fijo

$$p(x, \mathcal{X}_1) \geq \int_{y \in \mathcal{X}_1} q(x, y) \alpha(x, y) dy > 0$$

lo cual da una contradicción (recuérdese que si $f(x) = 0$ entonces $\alpha(x, y) = 1$ por convención).

Teorema 1.2 (Convergencia en variación total).

Una cadena de Márkov (con un espacio de estados cuya σ -álgebra es generada con una familia numerable de conjuntos) aperiódica y ϕ -irreducible con distribución estacionaria f satisface

$$\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - f(\cdot)\| = 0$$

(para f -casi-todo $x \in \mathcal{X}$).

En particular $\lim_{n \rightarrow \infty} P^n(x, A) = f(A) \forall A \subset \mathcal{B}(\mathcal{X})$.

Teorema 1.3 (Teorema ergódico).

Si $h : \mathcal{X} \rightarrow \mathbb{R}$ es tal que $\int_{\mathcal{X}} |h(x)| f(x) dx = f(|h|) < \infty$ entonces se tiene una ley fuerte de los grandes números para la cadena de Márkov $\{X_i\}_{i=0}^{\infty}$:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n h(X_i) \stackrel{c.s.}{=} \int_{|\mathcal{X}|} h(x) f(x) dx = \mathbb{E}_f[X].$$

Definición 1.21 (Ergodicidad uniforme).

Una cadena de Márkov con distribución estacionaria f es uniformemente ergódica si

$$\|P^n(x, \cdot) - f(\cdot)\| \leq M\rho^n, n \in \{1, 2, \dots\}$$

para $\rho < 1$ y $M < \infty$.

Para dar condiciones que aseguren ergodicidad uniforme se utiliza el siguiente concepto:

Definición 1.22 (Conjunto pequeño).

Se dice que un conjunto C es pequeño, o (n_0, ϵ, v) -pequeño, si \exists un entero positivo n_0 , real positivo, $\epsilon > 0$ y una medida de probabilidad v en \mathcal{X} tales que

$$P^{n_0}(x, A) \geq \epsilon v(A), \forall x \in C, A \in \mathcal{B}(\mathcal{X}).$$

Ejemplo 1.10. Cadena de Metropolis-Hastings 1.8 en los conjuntos compactos donde f es acotada son pequeños. Sea C un conjunto compacto tal que $F|_C \leq k \leq \infty$. Sea $x \in C, A$ un conjunto y A_r subconjunto de A compacto tal que $f(A_r) > 0$; se define

$$\epsilon = \inf_{x \in C, y \in A_r} \min\{q(x, y), q(y, x)\} > 0.$$

entonces

$$P(x, dy) \geq q(x, y) \min \left\{ 1, \frac{q(y, x)f(x)}{q(x, y)f(y)} \right\} dy = \min \left\{ q(x, y), \frac{f(y)q(y, x)}{f(x)} \right\} dy > \epsilon,$$

$$\min \left\{ 1, \frac{f(y)}{k} \right\} dy > 0$$

por lo que C es pequeño. En particular si \mathcal{X} es compacto entonces \mathcal{X} es pequeño.

Teorema 1.4 (Ergodicidad uniforme para espacio de estados pequeño).

Dada una cadena de Márkov con distribución de probabilidad estacionaria f y espacio de estados \mathcal{X} pequeño respecto a algunos $n_0, \epsilon > 0$ y medida de probabilidad v , entonces la cadena es uniformemente ergódica, de hecho

$$\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - f(\cdot)\| \leq (1 - \epsilon)^{\lfloor n/n_0 \rfloor}.$$

Definición 1.23 (Cadena geométricamente ergódica).

Una cadena de Márkov con distribución estacionaria f es geométricamente ergódica si

$$\|P^n(x, \cdot) - f(\cdot)\| \leq M(x)\rho^n, n \in \{1, 2, \dots\}$$

para $\rho < 1$ y $M(x) < \infty$ para (f -casi-todo) $x \in \mathcal{X}$.

Definición 1.24 (Pruebas de convergencia usando construcciones de couplings(acoplamientos)).

Sean X y Y variables aleatorias en algún espacio \mathcal{X} , entonces

$$\begin{aligned} \|\mathcal{L}(X) - \mathcal{L}(Y)\| &= \sup_A |\mathbb{P}[X \in A] - \mathbb{P}[Y \in A]| \\ &= \sup_A |\mathbb{P}[X \in A, X = Y] + \mathbb{P}[X \in A, X \neq Y] - \mathbb{P}[Y \in A, X = Y] - \mathbb{P}[Y \in A, X \neq Y]| \\ &= \sup_A |\mathbb{P}[X \in A, X \neq Y] - \mathbb{P}[Y \in A, X \neq Y]| \leq \mathbb{P}[X \neq Y] \end{aligned}$$

Se le llama a

$$\|\mathcal{L}(X) - \mathcal{L}(Y)\| \leq \mathbb{P}[X \neq Y]$$

La desigualdad de coupling (o acoplamiento).

Definición 1.25 (Algoritmo para construcción de un coupling).

Sea $X_0 = x, X'_0 \sim \mathcal{L}(f)$ y P un kernel de transición de probabilidad que define una cadena de Márkov con C un conjunto (n_0, ϵ, v) -pequeño.

Para $n \in \{1, 2, \dots\}$ realizar

1. Si $X_n = X'_n$ se elige $X_{n+1} = X'_{n+1} \sim P(X_n, \cdot)$
2. En caso contrario, si $(X_n, X'_n) \in C \times C$ entonces con probabilidad ϵ se toma

$$X_{n+n_0} = X'_{n+n_0} \sim v(\cdot)$$

y con probabilidad $1 - \epsilon$ se toma

$$X_{n+n_0} \sim \frac{1}{1 - \epsilon} \left(P^{n_0}(X_n, \cdot) - \epsilon v(\cdot) \right)$$

$$X'_{n+n_0} \sim \frac{1}{1 - \epsilon} \left(P^{n_0}(X'_n, \cdot) - \epsilon v(\cdot) \right)$$

si $n_0 > 1$ se toma $X_{n+1}, \dots, X_{n+n_0-1}|X_n, X_{n+n_0}$ y, condicionalmente independientes, $X'_{n+1}, \dots, X'_{n+n_0-1}|X'_n, X'_{n+n_0}$.

3. En caso contrario, de forma condicionalmente independiente, se toman

$$\begin{aligned} X_{n+1} &\sim P(X_n, \cdot) \\ X'_{n+1} &\sim P(X'_n, \cdot). \end{aligned}$$

Obsérvese que si $X_n \neq X'_n$ (para lo cual se dice que el coupling no se ha acoplado), entonces la ley de probabilidad $X_{n+1}|X_n$ es $P(X_n, \cdot)$ pues en caso de que $(X_n, X'_n) \in C \times C$

$$\epsilon v(\cdot) + (1 - \epsilon) \frac{1}{1 - \epsilon} (P^{n_0}(X_n, \cdot) - \epsilon v(\cdot)) = P^{n_0}(X_n, \cdot)$$

y $X_{n+1}, \dots, X_{n+n_0-1}$ son generados por las distribuciones condicionales adecuadas. Análogamente $X'_{n+1}|X'_n$ tiene distribución dada por $P(X'_n, \cdot)$ y al tenerse $X'_n \sim \mathcal{L}(f)$ se sigue que $X'_n \sim \mathcal{L}(f) \forall n \geq 1$.

Demostración 1.1 (Teorema de ergodicidad uniforme para espacio de estados pequeño).

Sea una cadena de Márkov con distribución de probabilidad estacionaria f y espacio de estados \mathcal{X} pequeño respecto a algunos $n_0, \epsilon > 0$ y medida de probabilidad v . Entonces para la construcción del coupling asociado hay probabilidad ϵ cada n_0 iteraciones del coupling para que $X_n = X'_n$, en particular si $n = mn_0$, por la desigualdad de coupling se tiene

$$\lim_{n \rightarrow \infty} \|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| = \lim_{n \rightarrow \infty} \|P_n(x, \cdot) - f(\cdot)\| \leq \mathbb{P}[X_n \neq X'_n] \leq (1 - \epsilon)^m = (1 - \epsilon)^{n/n_0}.$$

El teorema de convergencia en variación total no hace supuestos sobre la existencia de un conjunto pequeño para poder construir un coupling por lo que para su demostración se hace uso del siguiente teorema técnico:

Teorema 1.5 (Existencia de un conjunto pequeño).

Toda cadena de Márkov ϕ -irreducible (en un espacio de estados con σ -álgebra generada contablemente), tiene un conjunto pequeño $C \subset \mathcal{X}$ tal que $\phi(C) > 0$ y $v(C) > 0$.

Lema 1.1 (técnico).

Sea una cadena de Márkov ϕ -irreducible y aperiódica en \mathcal{X} y $A \subset \mathcal{X}$ (medible), si $\mathbb{P}_x[\tau_A < \infty] > 0 \forall x \in \mathcal{X}$, entonces $\mathbb{P}_x[\tau_A < \infty] = 1$ para f -casi-todo $x \in \mathcal{X}$, con

$$\tau_A = \inf\{n \geq 1 : X_n \in A\}.$$

Demostración 1.2 (Teorema de convergencia en variación total).

Por el teorema de existencia de un conjunto pequeño sea C pequeño tal que $\phi(C), v(C) > 0$. Se construye el coupling asociado para C y se considera la cadena de Márkov $\{X_i, X'_i\}_{i=1}^{\infty}$. Sea $G = \{(x, x') : \mathbb{P}_{(x, x')}[\exists n \geq 1 : X_n = X'_n] = 1\} \subset \mathcal{X} \times \mathcal{X}$. Si $(X_0, X'_0) = (x, X'_0) \in G$ entonces usando la desigualdad de coupling como la prueba del teorema de ergodicidad uniforme se obtiene

$$\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - f(\cdot)\| = 0$$

por lo que vanta ver que para f -casi-todo $x \in \mathcal{X} \mathbb{P}[(x, X'_0) \in G] = 1$.

Se tiene que $\phi(C) > 0$ por o que de la definición de ϕ -irreducibilidad $\forall x \in \mathcal{X}$ hay un τ tal que $P(x, C) > 0$, por otro lado $f(C) > 0$ por lo que desde cualquier $(x, x') \in \mathcal{X} \times \mathcal{X}$ la cadena tiene probabilidad positiva de alcanzar $C \times C$ ya que por la construcción del coupling $\mathbb{P}[(X, X') \in C \times C | (X_0 = x, X'_0)] \geq P(x, C)f(C)$, del lema técnico se sigue que para f -casi-todo $x \in \mathcal{X}$ la cadena del coupling visita $C \times C$ con probabilidad 1.

En ese momento la cadena se acopla con probabilidad ϵ y en caso contrario toma valores en un conjunto donde f integra positivo (al ser f invariante para P ϕ -irreducible $P(x, \cdot)'f \forall x \in \mathcal{X}$ de lo cual se sigue de la definición de conjunto pequeño que $v'f$) por lo que se puede aplicar el lema técnico nuevamente para ver que con probabilidad 1 se regresa nuevamente a $C \times C$ con probabilidad ϵ de acoplarse. Se sigue que $\mathbb{P}_{(x, X'_0)}[\exists n \geq 1 : X_n = X'_n] = 1$ por lo que $(x, X'_0) \in G$.

Para la demostración del teorema ergódico se utilizará la siguiente definición:

Definición 1.26 (Tiempo de regeneración).

Un tiempo de regeneración es un tiempo de paro τ con la propiedad de que $(X_\tau, X_{\tau+1}, \dots) \perp\!\!\!\perp (X_{\tau-1}, X_{\tau-2})$.

Para la demostración del teorema ergódico también es necesario comprender el siguiente algoritmo:

Definición 1.27 (Algoritmo para splitting (partimiento) de una cadena de Márkov).

Sea $X_0 = X$ y P un kernel de transición de probabilidad que define una cadena de Márkov con C un conjunto ($n_0 = 1, \epsilon v$ -pequeño). Para $n \in \{1, 2, \dots\}$ se realiza:

- Si $X_n \in C$ entonces con probabilidad ϵ se toma

$$X_{n+1} \sim v(\cdot)$$

y con probabilidad $1 - \epsilon$ se toma

$$X_{n+1} \sim \frac{1}{1-\epsilon} \left(P(X_n, \cdot) - \epsilon v(\cdot) \right)$$

En esta construcción se dice que la cadena se parte en X_n o al tiempo n si $X_n \in C$ y X_{n+1} es generada con la distribución de probabilidad $v(\cdot)$. Obsérvese que los tiempos de partimiento son tiempos de regeneración por construcción y que para cualquier $n \geq 1$ la distribución marginal de X_n está dada por

$$\epsilon v(\cdot) + (1 - \epsilon) \frac{1}{1-\epsilon} \left(P(X_n, \cdot) - \epsilon v(\cdot) \right) = P(X_n, \cdot).$$

Sea $\tau_0 = 0$ y

$$\tau_j = \inf\{n > \tau_{j-1} : X_{n-1} \in C, X_n \sim v(\cdot)\}.$$

Demostración 1.3 (Teorema ergódico).

Se demostrará algo más fuerte. Sean h, g integrables en valor absoluto respecto a f . Se demostrará que

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n h(X_i)}{\sum_{i=1}^n g(X_i)} \xrightarrow{c.s.} \frac{\int h(x)f(x)dx}{\int g(x)f(x)dx}$$

Sea $K_n + 1$ el número de veces que el splitting se ha partido hasta el tiempo n . Obsérvese que procediendo como en la prueba del teorema para la variación total $K_n \xrightarrow{c.s.} \infty$, se tienen las desigualdades

$$\sum_{j=0}^{K_n} \sum_{i=r_j+1}^{\tau_{j+1}} f(x_i) \leq \sum_{i=1}^n f(X_i) \leq \sum_{j=0}^{K_n+1} \sum_{i=r_j+1}^{\tau_{j+1}} f(X_i)$$

y las variables aleatorias

$$S_j(f) = \sum_{i=r_j+1}^{\tau_{j+1}} f(X_i)$$

son independientes e idénticamente distribuidas.

Se sigue que

$$\frac{K_n + 1}{K_n + 2} \frac{\frac{1}{K_n+1} \sum_{j=0}^{K_n} S_j(h)}{\frac{1}{K_n+2} \sum_{j=0}^{K_n+1} S_j(g)} \leq \frac{\sum_{i=1}^n h(X_i)}{\sum_{i=1}^n g(X_i)} \leq \frac{K_n + 2}{K_n + 1} \frac{\frac{1}{K_n+2} \sum_{j=0}^{K_n+1} S_j(h)}{\frac{1}{K_n+1} \sum_{j=0}^{K_n} S_j(g)}$$

por lo que utilizando la Ley de los Grandes números usual para los bloques S_j y el hecho de que $\frac{K_n+2}{K_n+1} \xrightarrow{c.s.} 1$ y el teorema de Slutsky (o del mapeo continuo) se sigue que

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n h(X_i)}{\sum_{i=1}^n g(X_i)} \xrightarrow{c.s.} \frac{\int h(x)f(x)dx}{\int g(x)f(x)dx}.$$

Definición 1.28 (Átomo de una cadena).

Una cadena de Márkov tiene un átomo $\alpha \in \mathcal{B}(\mathcal{X})$ si

$$P(x, A) = v(A), \forall x \in \alpha, \forall A \in \mathcal{B}(\mathcal{X})$$

Sea α un átomo de una cadena de Márkov y

$$\tilde{\pi}(A) = \mathbb{E}_\alpha \left[\sum_{i=0}^{\tau_\alpha - 1} \mathbb{1}\{X_i \in A\} \right]$$

Se tiene que

$$\begin{aligned} \tilde{\pi}(A) &= \mathbb{E}[\mathbb{1}\{X_i \in \alpha, X_0 \in A\}] + \sum_{i=1}^{\infty} \mathbb{E}_\alpha[\mathbb{1}\{X_i \in A, \tau_\alpha > i\}] \\ &= \mathbb{E}[\mathbb{1}\{X_i \in \alpha, X_0 \in A\}] + \sum_{i=1}^{\infty} \mathbb{E}_\alpha[\mathbb{1}\{X_i \in A, \tau_\alpha > i-1\}] - \sum_{i=1}^{\infty} \mathbb{E}_\alpha[\mathbb{1}\{X_i \in A, \tau_\alpha = i\}] \\ &= \mathbb{E}[\mathbb{1}\{X_i \in \alpha, X_0 \in A\}] + \mathbb{E}_\alpha \left[\sum_{i=1}^{\infty} \mathbb{1}\{X_i \in A, \tau_\alpha > i-1\} \right] - \sum_{i=1}^{\infty} \mathbb{E}_\alpha[\mathbb{1}\{X_i \in A, \tau_\alpha = i\}] \\ &= \mathbb{E}[\mathbb{1}\{X_i \in \alpha, X_0 \in A\}] + \mathbb{E}_\alpha \left[\sum_{i=1}^{\infty} \mathbb{1}\{X_i \in A, \tau_\alpha > i-1\} \right] - \mathbb{E}_\alpha[\mathbb{1}\{X_{\tau_\alpha} \in A\}] \\ &= \mathbb{E}[\mathbb{1}\{X_i \in \alpha, X_0 \in A\}] + \mathbb{E}_\alpha \left[\sum_{i=1}^{\infty} \mathbb{1}\{X_i \in A, \tau_\alpha > i-1\} \right] - \mathbb{E}_\alpha[\mathbb{1}\{X_0 \in \alpha, X_0 \in A\}] \\ &= \mathbb{E}_\alpha \left[\sum_{i=1}^{\infty} \mathbb{1}\{X_i \in A, \tau_\alpha > i-1\} \right] = \mathbb{E}_\alpha \left[\sum_{i=1}^{\tau_\alpha} P(X_{i-1}, A) \right] \\ &= \mathbb{E}_\alpha \left[\sum_{i=1}^{\tau_\alpha - 1} P(X_i, A) \right] = \int \tilde{\pi}(dx) P(x, A). \end{aligned}$$

Definición 1.29 (Teorema de límite central para cadenas de Márkov).

Para una cadena de Márkov ϕ -irreducible y aperiódica con distribución estacionaria f , se dice que $h : \mathcal{X} \rightarrow \mathbb{R}$ satisface un teorema del límite central (TLC) si al ser iniciada la cadena en estacionariedad, es decir $X_0 \sim \mathcal{L}(f)$ entonces

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (h(X_i) - f(h)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (h(X_i) - \mathbb{E}_f[h(X)]) \xrightarrow{d} N(0, \sigma^2)$$

con

$$\sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[\left(\sum_{i=1}^n (h(X_i) - \mathbb{E}_f[h(X)]) \right)^2 \right]$$

Teorema 1.6. Sea una cadena de Márkov reversible iniciada en estacionariedad y $r(x) = \mathbb{P}[X_{n+1} = X_n | X_n = x]$. Si

$$\lim_{n \rightarrow \infty} n f((h - f(h))^2 r^n) = \infty$$

entonces no se satisface un TLC para h .

Demostración 1.4.

$$\begin{aligned}\sigma^2 &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_f \left[\left(\sum_{i=1}^n (h(X_i) - \mathbb{E}_f[h(X)]) \right)^2 \right] \geq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_f \left[\left(\sum_{i=1}^n (h(X_i) - \mathbb{E}_f[h(X)]) \right)^2 \mathbb{1}\{X_0 = X_1 = \dots = X_n\} \right] \\ &= \frac{1}{n} \mathbb{E}_f [h(h(X_0) - f(h))^2 r(X_0)^n] = \lim_{n \rightarrow \infty} n f(h - f(h))^2 r^n = \infty.\end{aligned}$$

Teorema 1.7. Si una cadena de Márkov con distribución estacionaria f es uniformemente ergódica entonces se satisface un TLC para h tal que $f(h^2) < \infty$.

Teorema 1.8. Si una cadena de Márkov con distribución estacionaria f es geométricamente ergódica entonces se satisface un TLC para h tal que $f(|h|^{2+\delta}) < \infty$ para algún $\delta > 0$.

Teorema 1.9. Si una cadena de Márkov con distribución estacionaria f es geométricamente ergódica y reversible entonces se satisface un TLC para h tal que $f(h^2) < \infty$.

Teorema 1.10. Si una cadena de Márkov es ϕ -irreducible, aperiódica y reversible con distribución estacionaria f entonces se satisface un TLC cuando $\sigma^2 < \infty$.

1.7. Tamaño efectivo de muestra

Si se consideran variables aleatorias X_1, \dots, X_n entonces

$$\begin{aligned}Var\left(\frac{X_1 + \dots + X_n}{n}\right) &= Cov\left(\frac{X_1}{n} + \dots + \frac{X_n}{n}, \frac{X_1}{n} + \dots + \frac{X_n}{n}\right) \\ &= \frac{1}{n^2} \sum_{1 \leq j \leq n} Var(X_j) + \frac{1}{n^2} \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n; j \neq i} Cov(X_i, X_j)\end{aligned}$$

En el caso más sencillo de que X_1, \dots, X_n son i.i.d. y $\sigma^2 = Var(X_1)$ se tiene que

$$Var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \sum_{1 \leq i \leq n} \sigma^2 + \frac{1}{n^2} \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n; j \neq i} Cov(X_i, X_j) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Si para todo $1 \leq i \leq n$, $Var(X_i) = \sigma^2$ y para todo $1 \leq k \leq n-i$ se tiene que $Cov(X_i, X_{i+k}) = \rho_k \sigma^2$ entonces

$$\begin{aligned}Var\left(\frac{X_1 + \dots + X_n}{n}\right) &= \frac{1}{n^2} \sum_{i=1}^n Var(X_1) + \frac{2}{n^2} \sum_{k=1}^{n-1} Cov(X_1, X_{1+k})(n-k) = \frac{n\sigma^2}{n^2} + \frac{2}{n} \sum_{k=1}^{n-1} \sigma^2 \rho_k \left(1 - \frac{k}{n}\right) \\ &= \frac{\sigma^2}{n} \left(1 + 2 \sum_{k=1}^{n-1} \rho_k \left(1 - \frac{k}{n}\right)\right)\end{aligned}$$

en este caso si se toma $m = \lfloor \frac{n}{1+2\sum_{k=1}^{n-1} \rho_k \left(1 - \frac{k}{n}\right)} \rfloor$ observaciones i.i.d. entonces la varianza de su promedio empírico coincide con la varianza del promedio empírico de n variables aleatorias con la estructura de correlación anterior.

Considérese $\{X_i\}_{i=1}^\infty$ tales que $Var(X_i) = \sigma^2$ para $1 \leq i$ y $Cov(X_i, X_j) = \rho \sigma^2$, $1 \leq i, j, i \neq j$. Entonces

$$m = \lfloor \frac{n}{1+2\rho \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right)} \rfloor = \lfloor \frac{n}{1+2\rho \left(n-1 - \frac{(n-1)n}{2n}\right)} \rfloor = \lfloor \frac{n}{1+\rho(n-1)} \rfloor$$

(Código de ejemplo:4).

Ejemplo 1.11.

Considérese una cadena $AR(1)$ iniciada en su distribución estacionaria, es decir $X_0 \sim N\left(0, \frac{\sigma^2}{1-\theta^2}\right)$ entonces

$$\rho_k = \frac{Cov(X_1, X_{1+k})}{\sqrt{Var(X_1)Var(X_{1+k})}} = \frac{\sigma^2 \theta^k}{\sigma^2} = \theta^k$$

y

$$m = \lfloor \frac{n}{1 + 2 \sum_{k=1}^{n-1} \theta^k \left(1 - \frac{k}{n}\right)} \rfloor$$

Motivado por lo anterior se define $n_{ef} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho_k \left(1 - \frac{k}{n}\right)}$ el tamaño de muestra efectiva (Código: 5).

Definición 1.30 (Error estándar Monte-Carlo).

Sea θ una cantidad aleatoria de interés en una cadena para MCMC $\theta_1, \dots, \theta_n$, sea

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(\theta^{(i)} - \frac{1}{n-1} \sum_{j=1}^n \theta^{(j)} \right)^2}$$

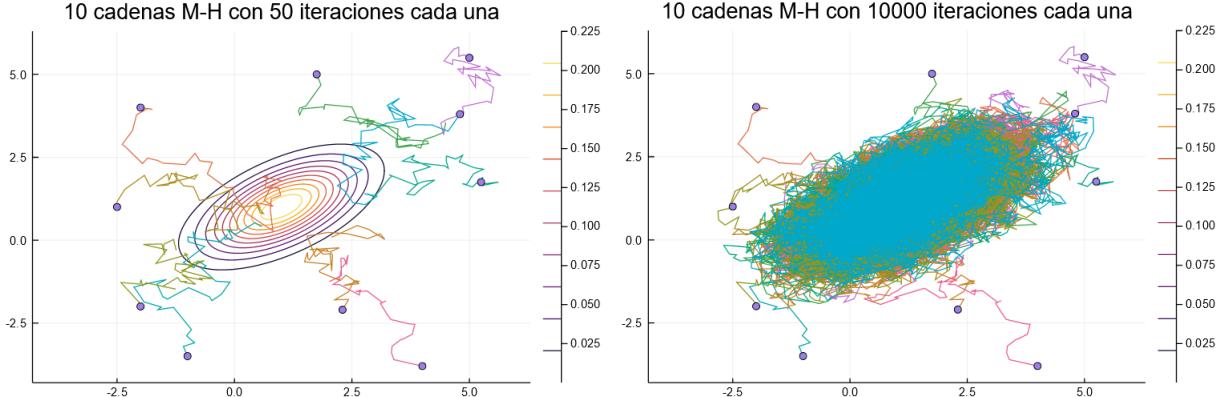
un estimador de la desviación estándar de θ sobre la cadena. Se define el error estándar de Monte-Carlo como

$$MCSE = \frac{\hat{\sigma}}{\sqrt{n_{ef}}}.$$

Está cantidad sirve para estimar σ en el teorema del límite central para MCMC y puede ser utilizada como diagnóstico para la precisión de la estimación MCMC.

Definición 1.31 (Estadístico \hat{R} de Gelman y Rubin (1992)).

El estadístico \hat{R} se define en términos de M cadenas de Márkov para una cantidad de interés θ .



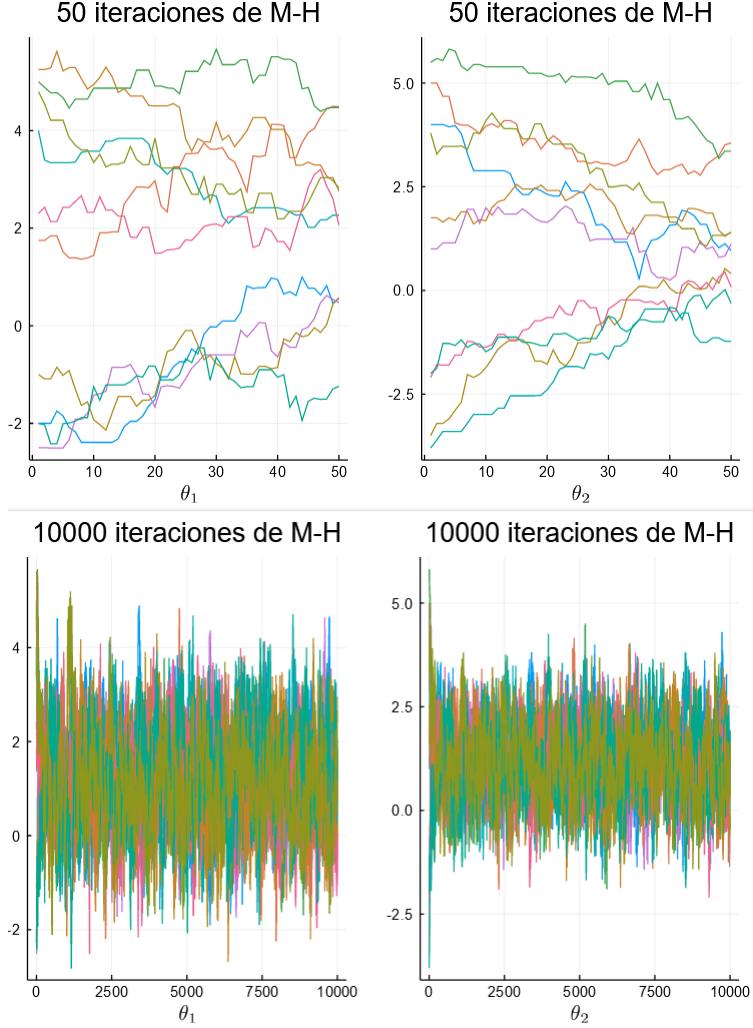


Figura 18: Cadenas de Metropolis-Hastings.

Para $\theta \in \mathbb{R}$ una cantidad de interés en las M cadenas de tamaño N , $\{\theta_i^{(m)}\}_{i=1}^N$ con $m \in \{1, \dots, M\}$, se definen las medias y varianzas por cadena

$$\bar{\theta}^{(m)} = \frac{1}{N} \sum_{i=1}^n \theta_i^{(m)}$$

$$s^{2(m)} = \frac{1}{N-1} \sum_{i=1}^N (\theta_i^{(m)} - \bar{\theta}^{(m)})^2.$$

La varianza "dentro de las cadenas" se define como

$$W = \frac{1}{M} \sum_{m=1}^M s^{2(m)}.$$

Sea

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^M \bar{\theta}^{(m)}.$$

La varianza entre las cadenas "se define como

$$B = \frac{N}{M-1} \sum_{j=1}^M (\bar{\theta}^{(m)} - \bar{\theta})^2.$$

Se define el siguiente estimador para la varianza de θ

$$\widehat{var}^+ = \frac{N-1}{N} W + \frac{1}{N} B$$

Finalmente se define

$$\hat{R} = \sqrt{\frac{\widehat{var}^+}{W}}$$

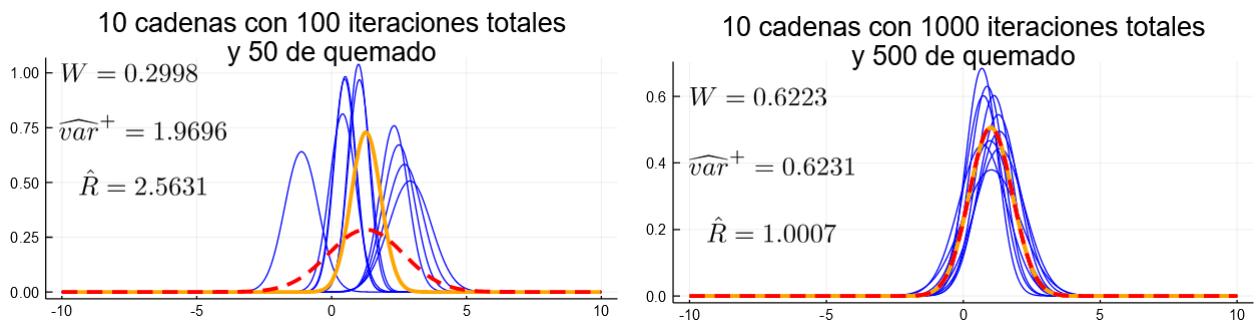


Figura 19: Estimadores de varianza y estadístico de Gelman y Rubin para cadenas MH.

1.8. Slice Sampler

Recuérdese que el teorema fundamental de la simulación dice que

$$X \sim \mathcal{L}(f(x)) \Leftrightarrow (X, U) \sim \text{Unif}\{(x, u) : 0 < u < f(x)\}$$

con lo que se tienen las condicionales

$$U|X \sim \text{Unif}(0, f(X))$$

y

$$X|U \sim \text{Unif}\{x : U \leq f(x)\}$$

por lo que se puede implementar un muestreador de Gibbs llamado muestreador por rebanadas (o slice).

Definición 1.32 (Algoritmo de muestreo por rebanadas).

Sea U_{n-1} y X_{n-1}

1. Simular $U_n \sim \text{Unif}(0, f(X_{n-1}))$
2. Simular $X_n \sim \text{Unif}\{x : U_n \leq f(x)\}$

Ejemplo 1.12.

Considérese la función de densidad de probabilidad

$$f(x) = \frac{1}{2} e^{-\sqrt{x}} \mathbb{1}_{\{x \geq 0\}}$$

entonces para $X|U = u$ las rebanadas de las cuales se simula uniformemente están dadas por

$$\left\{0 \leq x : u \leq \frac{1}{2}e^{\sqrt{x}}\right\} = \{0 \leq x : -\log(2u) \geq \sqrt{x}\} = \{0 \leq x : (\log(2u))^2 \geq x\} = [0, (\log(2u))^2]$$

entonces

$$X|U \sim \text{Unif}(0, (\log(2u))^2)$$

(Código: 6).

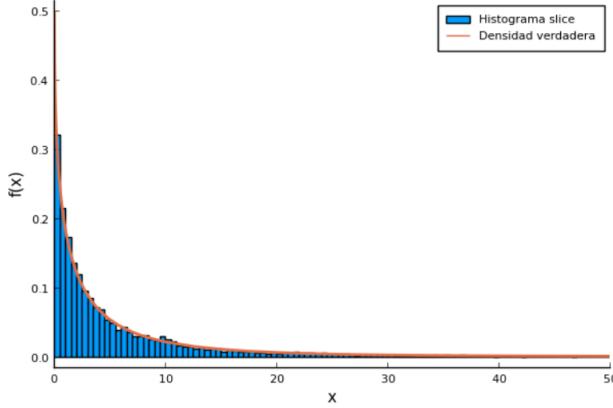


Figura 20: Simulación de densidad con Slice sampler.

Si se considera una función de densidad de probabilidad

$$f(x) \propto \prod_{i=1}^k f_i(x)$$

se puede expresar

$$f_i(x) = \int \mathbb{1}_{\{0 \leq w_i \leq f_i(x)\}} dw_i$$

y tomar $(x, w_1, \dots, w_k) \sim \mathcal{L}(p)$ con

$$p(x, w_1, \dots, w_k) \propto \prod_{i=1}^k \mathbb{1}_{\{0 \leq w_i \leq f_i(x)\}}$$

Definición 1.33 (Algoritmo general de muestreo por rebanadas).

Sean $W_{1,n-1}, \dots, W_{k,n-1}$ y X_{n-1}

1. Simular $W_{1,n-1} \sim \text{Unif}(0, f_1(X_{n-1}))$

⋮

- k.** Simular $W_{k,n-1} \sim \text{Unif}(0, f_k(X_{n-1}))$

- k + 1.** Simular $X_n \sim \text{Unif}\{x : W_{i,n} \leq f_i(x), 1 \leq i \leq k\}$.

Ejemplo 1.13. Considérese

$$f(x) \propto (1 + \sin^2(3x))e^{-x^2/2}$$

entonces se puede tomar $f_1(x) = (1 + \sin^2(3x))$ y $f_2(x) = e^{-x^2/2}$. Se tiene

$$\begin{aligned} & \{x : w_1 \leq 1 + \sin^2(3x)\} \\ &= \begin{cases} \cup_{k \in \mathbb{N}} \left(\frac{\arcsin(\sqrt{w_1 - 1})}{3} + \frac{k}{3}\pi, \frac{\pi - \arcsin(\sqrt{w_1 - 1})}{3} + \frac{k}{3}\pi \right) & \text{si } w_1 \geq 1 \\ \mathbb{R} & \text{si } w_1 < 1 \end{cases} \end{aligned}$$

y

$$\{x : w_2 \leq e^{-x^2/2}\} = [-\sqrt{-2\log(w_2)}, \sqrt{-2\log(w_2)}]$$

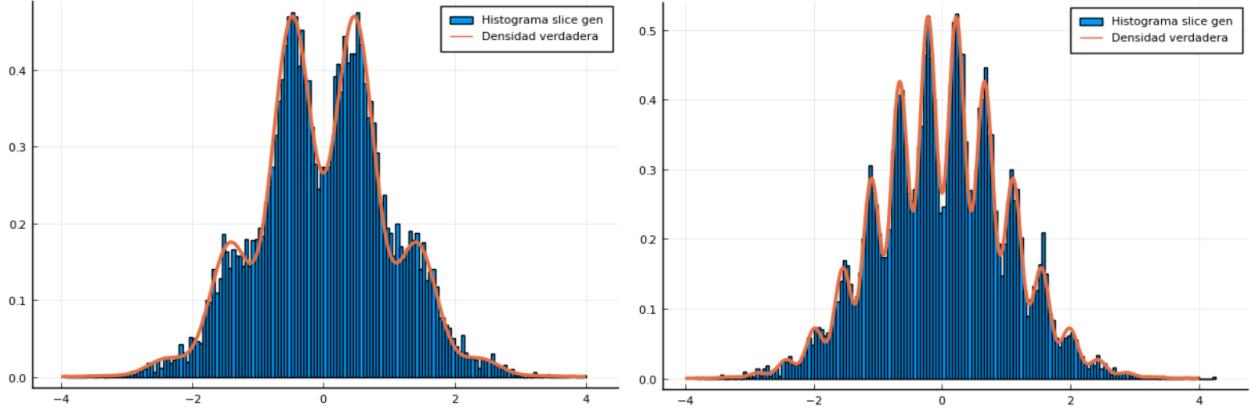


Figura 21: Simulación de densidades con algoritmo general de Slice sampling.

Cuando una densidad no puede ser factorizada ni invertida para obtener de manera relativamente sencilla las rebanadas horizontales, se pueden construir rebanadas aleatorias $[L, R]$ que contengan casi seguramente a la rebanada horizontal desconocida. De esta manera se puede simular uniformemente de $[L, R]$ y aceptar cuando el punto generado pertenezca a la rebanada horizontal deseada. Para justificar que tal procedimiento conforma un MCMC se debe tener cuidado en no romper "la reversibilidad del muestreo por rebanadas (caso particular del muestreo de Gibbs).

Definición 1.34 (Algoritmo Slice Stepout).

1. Elegir un punto inicial x_0 , w una longitud supuesta para la rebanada, m un entero para limitar el grosor de la rebanada y y el nivel vertical definiendo la rebanada.
2. Sean $U, V \sim Unif(0, 1)$ independientes.
3. Tomar $L = x_0 - wU$, $R = x_0 + w(1 - U)$, $J = \lfloor mV \rfloor$ y $K = m - 1 - J$
4. Mientras $J > 0$ y $y < f(L)$
 - a) Actualizar $L = L - w$ y $J = J - 1$
5. Mientras $K > 0$ y $y < f(R)$
 - a) Actualizar $R = R + w$ y $K = K - 1$.

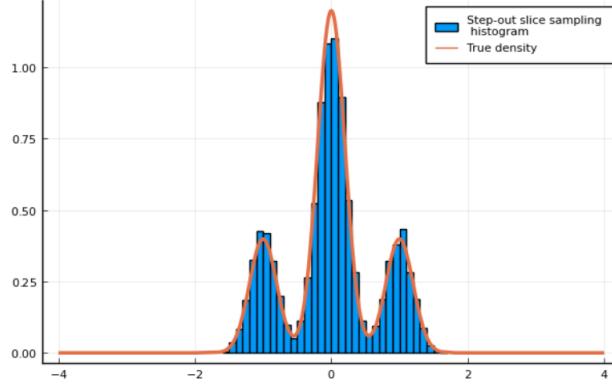


Figura 22: Simulación de densidad con Algoritmo Slice Stepout.

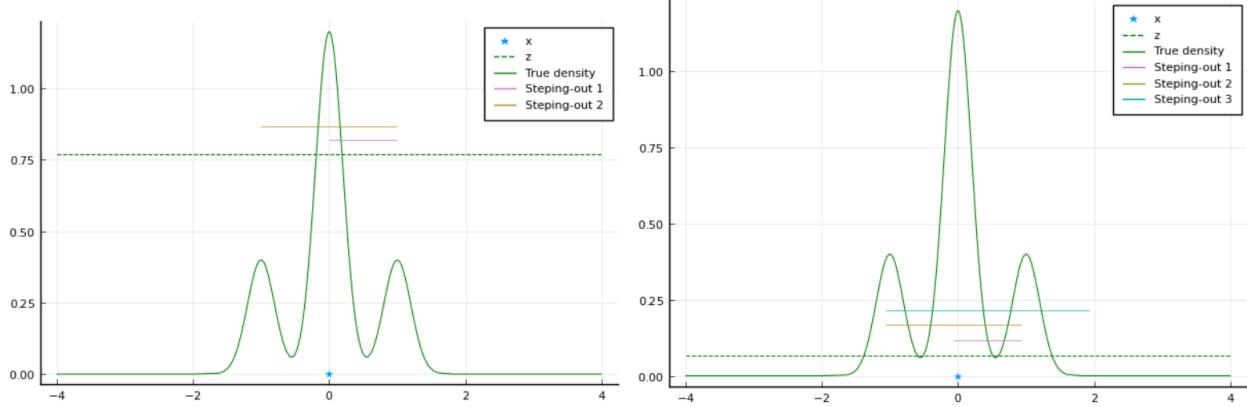


Figura 23: Rebanadas con Algoritmo Slice Stepout.

Definición 1.35 (Algoritmo Slice Doubling).

1. Elegir un punto inicial x_0 , w una longitud supuesta para la rebanada, m un entero para limitar el grosor de la rebanada y y el nivel vertical definiendo la rebanada.
2. Sea $U \sim Unif(0, 1)$.
3. Tomar $L = x_0 - wU$, $R = x_0 + w(1 - U)$ y $K = p$.
4. Mientras $K > 0$ y $\{y < f(L) \text{ ó } y < f(R)\}$.
 - a) $V \sim Unif(0, 1)$, si $V < 0.5$ actualizar $L = L - (R - L)$ y en caso contrario $R = R + (R - L)$. Actualizar $K = K - 1$.

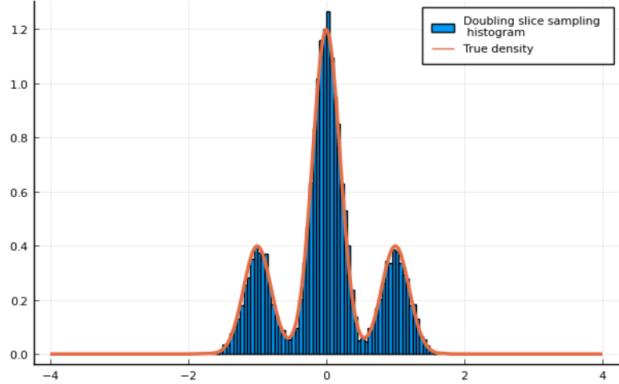


Figura 24: Simulación de densidad con Algoritmo Slice Doubling.

Definición 1.36 (Algoritmo Shrinking).

Sean L y R los extremos izquierdo y derecho, respectivamente, de la rebanada horizontal dada por y y x_0 el valor aceptado de la rebanada anterior.

1. Tomar $U \sim \text{Unif}(0, 1)$ y $x_1 = L + U(R - L)$.
2. Si $y < f(x_1)$, y en caso de doubling, se acepta x_1 entonces se termina el algoritmo. En caso contrario se continua al paso 3.
3. Si $x_1 < x_0$ entonces se toma $L = x_1$ y se vuelve al paso 1, en caso contrario se continua al paso 4.
4. Si $x_1 \geq x_0$ entonces se toma $R = x_1$ y se vuelve al paso 1.

Definición 1.37 (Algoritmo Aceptación para Doubling).

Sea x_0 el punto previo, x_1 el punto a aceptar, w la longitud supuesta para la rebanada, y el nivel vertical definiendo la rebanada y (\hat{L}, \hat{R}) el intervalo generado por el algoritmo doubling.

1. Sean $\hat{L} = L$, $\hat{R} = R$ y $D = \text{False}$. Mientras $\hat{R} - \hat{L} > w$
 - a) Tomar $M = (\hat{R} + \hat{L})/2$.
 - b) Si $\{x_0 < M, x_1 \geq M\}$ ó $\{x_0 \geq M, x_1 < M\}$ entonces se define $D = \text{True}$.
 - c) Si $X_1 < M$ entonces $\hat{R} = M$ en caso contrario $\hat{L} = M$.
 - d) Si D y $y \geq f(\hat{L})$ y $y \geq f(\hat{R})$ entonces x_1 no es aceptable.
2. El punto x_1 es aceptable si nunca es determinado ni aceptable en el paso 2.

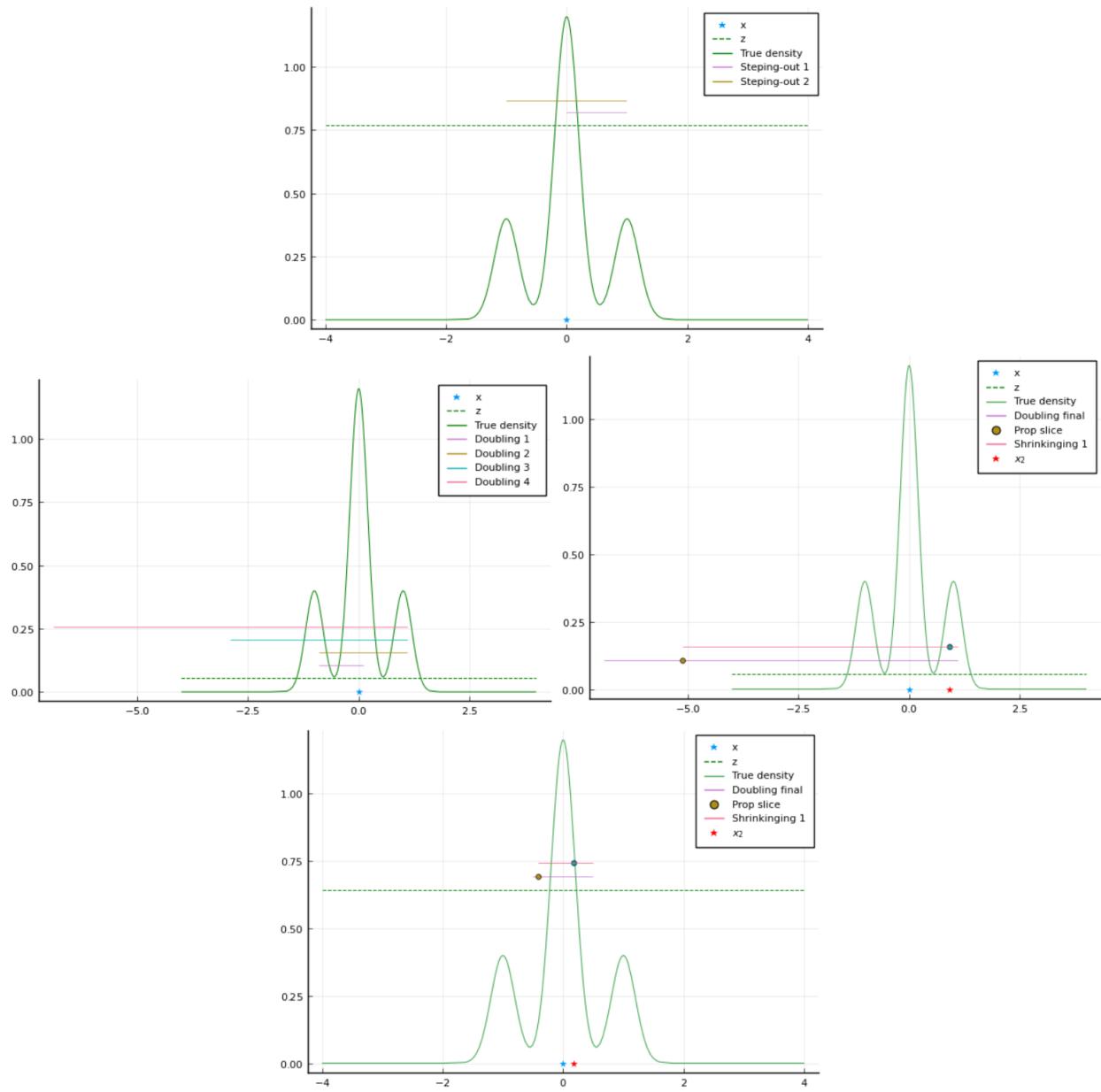


Figura 25: Rebanadas e intervalos con Doubling.

2. Código

Para replicar cualquiera de los resultados mencionados a lo largo del curso solo se requiere ejecutarlos en un entorno de Julia.

Listing 1: Algoritmo Metropolis-Hastings

```
function MetropolisHasting_logscaled_randwalk( 1f :: Function , 1 :: Int64 ,
v_0::Array{Float64,1} , prop_sigma::Float64=0.5)
    v = [ zeros(length(v_0)) for _ in 1:(1+1) ]
    v[1] = v_0
    log_pi_v = 1f( v_0 )
    for i in 2:(1+1)
        v_prop = v[i-1] + rand( MvNormal( zeros(length(v_0)), prop_sigma ) )
        log_pi_v_prop = 1f( v_prop )
        u = rand(Uniform())
        if log(u) < min( log_pi_v_prop - log_pi_v , 0.0 )
            v[i] = v_prop
            log_pi_v = log_pi_v_prop
        else
            v[i] = v[i-1]
        end
    end
    return v
end
MetropolisHasting_logscaled_randwalk( loglik , 3 , [1.0 ,1.0] , 0.5);
```

Listing 2: Ejemplo Muestreo Gibbs

```
# Muestreo de Gibbs
m = 10000 # iteraciones para la cadena
chain_mu = zeros(Float64,m+1) # cadena para mu
chain_sigma_2 = zeros(Float64,m+1) # cadena para sigma_2
chain_mu[1] = rand( Normal(0.0,9.0) ) # inicializacion aleatoria para mu
chain_sigma_2[1] = rand( Normal(0.0,4.0) )^2.0 # inicializacion aleatoria para sigma_2
for i in 2:(m+1)
    chain_mu[i] = rand( Normal(mu_n,sqrt(chain_sigma_2[i-1]/eta_n) ) )
    chain_sigma_2[i]=1.0/rand(Gamma(alpha_n+0.5, 2.0/(2*beta_n+eta_n*(chain_mu[i]*mu_n)^2.0)))
end
```

Listing 3: Ejemplo Muestreo Gibbs directo

```
# Usando directamente la marginal sigma_2 | X y posterior mu | sigma_2 , X
m = 10000 # iteraciones para la cadena
chain_mu_B2 = zeros(Float64,m) # cadena para mu
chain_sigma_2_B2 = zeros(Float64,m) # cadena para sigma_2
for i in 1:m
    chain_sigma_2_B2[i] = rand( InverseGamma(alpha_n,beta_n) )
    chain_mu_B2[i] = rand( Normal(mu_n,sqrt(chain_sigma_2_B2[i]/eta_n)) )
end
```

Listing 4: Ejemplo de tamaño efectivo de muestra

```
sigma_2 = 1.0
rho = 0.49
Sigma = ( zeros(N,N) .+ rho ) + I*(sigma_2/rho)
N = 100
M = floor( Int64 , N/(1+rho*(N-1)) )
N_ = 100000
mu_1 = zeros(N_)
```

```

mu_2 = zeros(N_)
for i in 1:N_
    Y = rand(Normal(0, sqrt(sigma_2)), M) # Obs i.i.d.
    mu_1[i] = mean(Y)
    X = rand(MvNormal(Sigma)) # Obs con matriz de covarianza Sigma
    mu_2[i] = mean(X)
end

# Varianza media muestral i.i.d.
v1 = var(mu_1)
# = 0.4971443145970311

# Varianza teorica de media muestral
sigma_2/M
# = 0.5

M
# = 2

# Varianza media muestral de N variables con correlacion rho
v2 = var(mu_2)
# = 0.5084195971290772

# Varianza teorica de media muestral
sigma_2 * (1+rho*(N-1))/N
# = 0.4951

N
# = 100

```

Listing 5: Ejemplo de tamaño efectivo de muestra AR(1)

```

sigma_2 = 1.0
Theta = 0.5
N = 100
M = floor(Int64, N/(1+2*sum([ (1-k/N)*Theta^k for k in 1:(N-1) ])))
N_ = 100000
mu_1 = zeros(N_)
mu_2 = zeros(N_)
for i in 1:N_
    Y=rand(Normal(0, sqrt(sigma_2/(1-Theta^2.0))), M) # Obs i.i.d. con dist estac AR(1) marginal
    mu_1[i] = mean(Y)
    X = zeros(N)
    X[1] = rand(Normal(0, sqrt(sigma_2/(1-Theta^2.0)))) # Estado inicial
    de cadena AR(1)
    for j in 2:N
        X[j] = Theta*X[j-1] + rand(Normal()) # Construccion iterativa de cadena AR(1)
    end
    mu_2[i] = mean(X)
end

# Varianza media muestral i.i.d.
v1 = var(mu_1)

# = 0.040634910586029756

# Varianza teorica de media muestral
( sigma_2/(1-Theta^2.0))/M

# 0.040404040404040404

```

```

M
# =33

# Varianza media muestral AR(1) con M iteraciones de la cadena
v2 = var(mu_2)

# =0.039518505016971545

# Varianza teorica de media muestral
( sigma_2/(1-theta^2.0)) * (1+2*sum( [ (1-k/N)*theta^k   for k in 1:(N-1) ] ) )/N

# =0.03946666666666664

```

Listing 6: Ejemplo de simulación con Slice Sampler

```

n = 10000
u = zeros(Float64,n)
x = zeros(Float64,n)
u[1] = rand(Uniform(0,f(0)))
x[1] = rand(Uniform(0,log(2*u[1])^2.0))
for i in 2:n
    u[i] = rand(Uniform(0,f(x[i-1])))
    x[i] = rand(Uniform(0,log(2*u[i])^2.0))
end

```

3. Bibliografía

1. Asmussen, S., Glynn, P. (2011). *Stochastic simulation*. New York: Springer.
2. Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference, Algorithms, Evidence and Data Science*. Cambridge University Press. Wiley.
3. Gamerman, D. y López, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall-CRC.
4. Givens, G.H. and Hoeting, J.A. (2013). *Computational Statistics*. (2nd ed.). Wiley.
5. Nelson, B. (2015). *Foundations and methods of stochastic simulation*. [s.l]: Springer.
6. Judd, K., Maliar, L., Maliar, S. (2011). *One-node quadrature beats Monte Carlo*. Cambridge, Mass.: National Bureau of Economic Research.
7. Weihs, C., Mersmann, O. and Ligges, U. (n.d.). *Foundations of statistical algorithms*. United States: CRC Press.