Simulación Estocástica

Métodos de Monte Carlo vía Cadenas de Markov

Vanda Inácio de Carvalho

Primer Semestre 2015

- En estadística clásica (o frequentista) dado un modelo paramétrico, asumimos que los parámetros del modelo son fijos, típicamente, constantes desconocidas.
- Dados los datos, se puede hacer inferencia sobre los parámetros, frecuentemente a través de la función de verosimilitud. Es decir, determinamos las estimaciones de máxima verosimilitud de los parámetros. Podemos también calcular errores estándar e intervalos de confianza.
- En estadística Bayesiana también asumimos un modelo paramétrico para los datos.
- Sin embargo, en lugar de asumir que los parámetros del modelo son constantes fijas, asumimos que estos son variables aleatorias.
- Además, podemos hacer uso de algún tipo de conocimiento que tengamos *a priori* sobre los valores de los parámetros.
- Esta es una característica importante de la estadística Bayesiana, la presencia de una distribución a priori.

- La distribución a posteriori de los parámetros del modelo depende del modelo paramétrico elegido, de los datos, y de la distribución a priori.
- La distribución a posteriori contiene toda la información que necesitamos sobre los parámetros.
- De esta distribución podemos extraer la média de los parámetros, la varianza e intervalos de probabilidad/credibilidad (el equivalente en el contexto Bayesiano a los intervalos de confianza).

- Sea $x_1, \ldots, x_n$ una muestra aleatoria y sea θ (puede ser un vector) un parámetro continuo.

- Los 'ingredientes' principales para llevar a cabo inferencia Bayesiana sobre θ son la función de verosimilitud
$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta),$$
y la distribución a priori de θ, $p(\theta)$.

- El teorema de Bayes da una 'receta' para la inferencia a posteriori
$$p(\theta \mid \mathbf{x}) = \frac{L(\theta; \mathbf{x})p(\theta)}{\int_{\Theta} L(\theta; \mathbf{x})p(\theta)\mathrm{d}\theta}. \tag{1}$$

- La interpretación de (1) es la siguiente: cuando la información a priori sobre el parámero θ, expresa por $p(\theta)$, se combina con los datos observados, la información a priori sobre θ se actualiza y se expresa por $p(\theta \mid \mathbf{x})$, la distribución a posteriori.

- El denominador en (1) asegura que $p(\theta \mid \mathbf{x})$ es de hecho una densidad, es decir, que integra a uno (constante de normalización).

- Ya que el denominador en (1) sólo depende de los datos observados, que se suponen fijos en un contexto Bayesiano, podemos escribir

$$p(\theta \mid \mathbf{x}) \propto L(\theta; \mathbf{x}) p(\theta).$$

- Por palabras,

$$\text{posteriori} \propto \text{verosimilitud} \times \text{priori}$$

- Supongamos que la distribución a priori $p(\theta)$ pertenece a una clase de distribuciones paramétricas $\mathcal{F}$.
- Entonces, se dice que la distribución a priori $p(\theta)$ es conjugada con respecto a la verosimilitud $L(\theta; \mathbf{x})$, si la distribución a posteriori $p(\theta \mid \mathbf{x})$ también pertenece a $\mathcal{F}$, es decir

$$p(\theta) \in \mathcal{F} \Rightarrow p(\theta \mid \mathbf{x}) \in \mathcal{F}.$$

- A continuación vamos a ver vários ejemplos de modelos conjugados.

Likelihood	Prior	Posterior
Binomial	Beta	Beta
Negative Binomial	Beta	Beta
Poisson	Gamma	Gamma
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Normal (mean unknown)	Normal	Normal
Normal (variance unknown)	Inverse Gamma	Inverse Gamma
Normal (mean and variance unknown)	Normal/Gamma	Normal/Gamma
Multinomial	Dirichlet	Dirichlet

- Vamos a considerar un ejemplo ficticio.
- Supongamos que n trabajadores de una determinada empresa son examinados para detectar el consumo de drogas.
- Además, supongamos que x (de los n) de estos trabajadores presentan pruebas positivas para el consumo de drogas y que un trabajador presenta una prueba positiva con probabilidad θ.
- El objetivo es estimar θ y evaluar la incertidumbre acerca de θ.
- La verosimilitud de este tipo de datos es obviamente binomial

$$L(\theta; x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

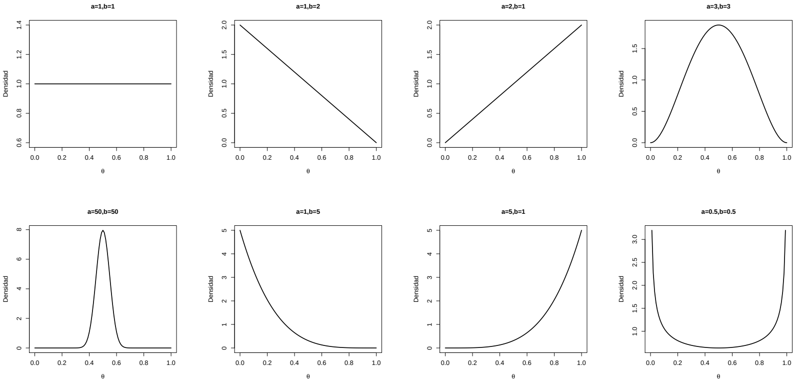
- Necesitamos de elegir una distribución para θ.
- Podemos usar la distribución beta como distribución a priori para θ, dado que tiene soporte en $[0, 1]$ (y más importante, es conjugada de la dist. binomial).
- Si $\theta \sim \text{Beta}(a, b)$, entonces

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1.$$

- La media, varianza y moda de la dist. beta son, respectivamente,

$$\frac{a}{a+b}, \quad \frac{ab}{(a+b)^2(a+b+1)}, \quad \frac{a-1}{a+b-2}.$$

- Los parámetros a y b son llamados hiperparámetros y pueden ser fijos o no (modelos jerárquicos).


a=1,b=1
a=1,b=2
a=2,b=1
a=3,b=3
a=50,b=50
a=1,b=5
a=5,b=1
a=0.5,b=0.5
Densidad
θ

- La dist. a posteriori es

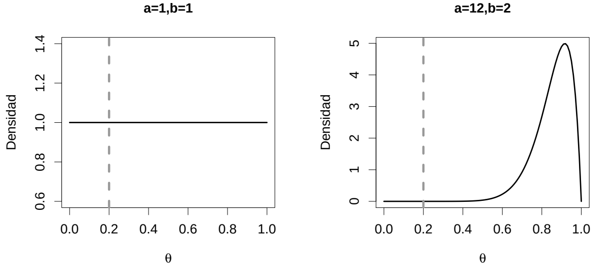
$$
\begin{aligned}
p(\theta \mid x) &\propto L(\theta; x)p(\theta) \\
&= \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{a+x-1}(1-\theta)^{b+n-x-1}.
\end{aligned}
$$

- La última expresión la reconocemos como el núcleo de una dist. beta con parámetros $a+x$ y $b+n-x$. Así

$$\theta \mid x \sim \text{Beta}(a+x, b+n-x).$$

- Es este modelo, a puede ser visto como el número de sucesos a priori y b como el númuero de insucesos.

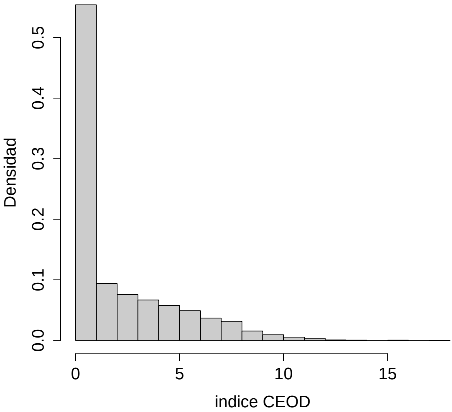
- Supongamos que $n = 200$ y $x = 40$.
- EMV: $\hat{\theta} = \frac{x}{n} = \frac{40}{200} = 0.2$.
- Además consideremos los siguientes conjuntos de hiperparámetros: $a = b = 1$, y $a = 12, b = 2$.



- Gris: EMV. Negro: Dist. a priori.

- Se observa que cuando $a = 12, b = 2$, la información a priori contradice la información de los datos.
- Cuando $a = b = 1$, obtenemos $\mathbb{E}(\theta \mid x) = 41/202 \approx 0.2$, $\text{Var}(\theta \mid x) \approx 0.00080$.
- Cuando $a = 12$ y $b = 2$, obtenemos $\mathbb{E}(\theta \mid x) = 52/214 \approx 0.24$, $\text{Var}(\theta \mid x) \approx 0.00086$.
- Consideremos ahora $n = 50, x = 10$. La proporción de sucesos aún es 0.2 $(10/50)$.
- Cuando $a = b = 1$, obtenemos $\mathbb{E}(\theta \mid x) = 11/52 \approx 0.21$, $\text{Var}(\theta \mid x) \approx 0.0034$.
- Sin embargo, cuando $a = 12, b = 2$, obtenemos que $\mathbb{E}(\theta \mid x) = 22/64 \approx 0.34$, $\text{Var}(\theta \mid x) \approx 0.0035$.
- **Mensaje**: la información a priori es importante, especialmente para tamaños de muestra pequeños.

- Vamos a considerar el estudio de caries presentado en Lesaffre y Lawson (Bayesian Biostatistics, 2012).
- El conjunto de datos contiene información sobre la experiencia de caries en los dientes primarios (o de leche) de 4351 niños (de los Flanders) de 7 años de edad. El estudio fue realizado en 1996.
- La experiencia de caries en los dientes primarios se mide clínicamente por el indice CEOD.
- El indice CEOD es la sumatoria de dientes primarios cariados, con indicación de extracción o obturados. Varia de 0 (sin experiencia de caries) a 20 (todos los dientes primarios afectados).


Densidad
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
indice CEOD

- La dist. de Poisson es una opción natural para modelar los indices CEOD observados.
- La verosimilitud es

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \left\{ \frac{\theta^{x_i}}{x_i!} e^{-\theta} \right\},$$

  donde x_i representa el indice CEOD para el niño i.
- Aunque antes del estudio la información sobre la higiene oral de los niños Flanders era limitada, los siguientes hechos eran conocidos:
  1. El articulo de revisión de Van Obbergen et al. (2001) reportaba un indice CEOD promedio de 4.1, obtenido en un estudio de 109 niños Flanders de 7 años de edad. El estudio fue realizado en Liege en 1983.
  2. Un promedio para el indice CEOD de 1.39 se obtuvo alrededor de Ghent en 200 niños de 5 años de edad examinados en 1994.
  3. Se sabe que la higiene oral he mejorado considerablemente en los Flanders en los últimos años.

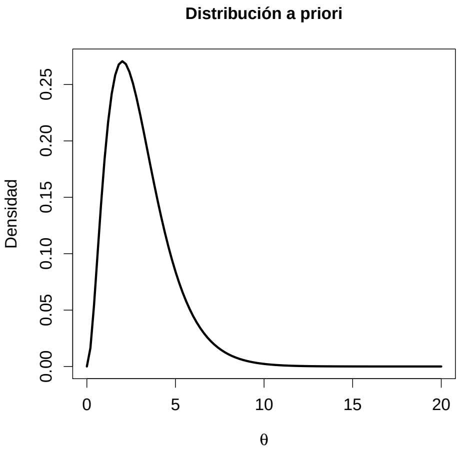
- Una distribución a priori adecuada para θ debe reflejar estos 3 hechos.
- La dist. Gama es conjugada con respecto a la verosimilitud Poisson.
- Por lo tanto, asumiremos que $\theta \sim \text{Gama}(a, b)$, donde a es el parámetro de forma y b es el inverso del parámetro de escala (también denominado de tasa), implicando que

$$\mathbb{E}(\theta) = \frac{a}{b}, \quad \text{Var}(\theta) = \frac{a}{b^2}.$$

- La densidad de la dist. Gama(a, b) es

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-\theta b}, \quad \theta > 0, \quad a, b > 0.$$

- Los autores consideraron que cuando $a = 3$ y $b = 1$, el conocimiento a priori parecía adecuadamente representado.


Distribución a priori
Densidad
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
θ

- Derivemos entonces la dist. a posteriori

$$\begin{aligned} p(\theta \mid \mathbf{x}) &\propto L(\theta; \mathbf{x}) p(\theta) \\ &= \prod_{i=1}^{n} \left\{ \frac{\theta^{x_i}}{x_i!} e^{-\theta} \right\} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \\ &\propto \theta^{a+\sum_{i=1}^{n} x_i - 1} e^{-\theta(b+n)}. \end{aligned}$$

- Así,

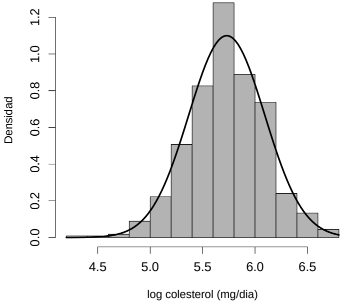
$$\theta \mid \mathbf{x} \sim \text{Gamma}\left(a + \sum_{i=1}^{n} x_i, b + n \right).$$

- En el ejemplo tenemos que $a = 3$, $b = 1$, $n = 4351$ y $\sum_{i=1}^{4351} x_i = 9758$. Luego, $\theta \mid \mathbf{x} \sim \text{Gamma}(4761, 4352)$.

- Claramente el efecto de la información a priori en la dist. a posteriori es minimo. Tenemos que

$$\mathbb{E}(\theta \mid \mathbf{x}) = \frac{9761}{4352} \approx 2.24, \quad \text{Var}(\theta \mid \mathbf{x}) = \frac{9761}{4352^2} \approx 0.0005.$$

- En este ejemplo usaremos los níveles de (log) colesterol de 563 empleados bancarios en Belgica. Los datos fueran recogidos en 1990 (Lesaffre & Lawson, Bayesian Biostatistics, 2012).



- Histograma de los niveles de log colesterol juntamente con la aproximación dada por la dist. normal.

- Sean $x_1, \ldots, x_n$ los niveles de log colesterol de los $n = 563$ empleados bancarios.
- Por simplicidad vamos a suponer que los log niveles de colesterol siguen una dist. normal

$$x_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2), \quad i = 1, \ldots, n.$$

- Por ahora, asumiremos que σ^2 es fijo.
- La verosimilitud es

$$\begin{aligned} L(\mu; \sigma^2, \mathbf{x}) &= \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\} \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}. \end{aligned}$$

- Vamos a considerar

$$\mu \sim \mathrm{N}(\mu_0, \sigma_0^2),$$

con μ_0 y σ_0^2 fijos.

- La dist. a posteriori es

$$
\begin{aligned}
p(\mu \mid \sigma^2, \mathbf{x}) &\propto L(\mu; \sigma^2, \mathbf{x}) p(\mu) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(-2\mu n\bar{x}+n\mu^2) - \frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{-2\mu n\bar{x}\sigma_0^2 + n\mu^2\sigma_0^2 + \mu^2\sigma^2 - 2\mu\mu_0\sigma^2}{\sigma^2\sigma_0^2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2(n\sigma_0^2+\sigma^2) - 2\mu(n\bar{x}\sigma_0^2+\mu_0\sigma^2)}{\sigma^2\sigma_0^2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu(n\bar{x}\sigma_0^2+\mu_0\sigma^2)/(n\sigma_0^2+\sigma^2)}{\sigma^2\sigma_0^2/(n\sigma_0^2+\sigma^2)}\right)\right\}
\end{aligned}
$$

- Sea
$$m = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$
- La distribución a posteriori queda entonces

$$\begin{aligned}
p(\mu \mid \sigma^2, \mathbf{x}) &\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu m}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu m + m^2 - m^2}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu m + m^2}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{(\mu - m)^2}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\}
\end{aligned}$$

- Reconocemos la expresión anterior como el núcleo de una dist. normal con média m y varianza $\sigma^2\sigma_0^2/(n\sigma_0^2+\sigma^2)$.

- Luego (dividindo todo por $\sigma^2\sigma_0^2$)

$$\mu \mid \sigma^2, \mathbf{x} \sim \mathrm{N}\left(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\right).$$

- Volviendo a los datos de níveles de (log) colesterol.
- Consideremos $\sigma^2 = s^2 = 0.13$.
- Además, consideremos $\mu_0 = 0$ y $\sigma_0^2 = 100$ (reflejando la ausencia de información a priori).
- Obtenemos así,

$$\mathbb{E}(\mu \mid \sigma^2, \mathbf{x}) = 5.73, \quad \text{Var}(\mu \mid \sigma^2, \mathbf{x}) = 0.00023.$$

Cadenas de Markov

Nociones fundamentales

- Informalmente, podemos definir un proceso estocástico como una colección de variables aleatorias $X_0, X_1, \ldots, X_n, \ldots$
- La mayoría de las veces asumimos que las variables aleatorias son independientes y identicamente distribuidas.
- Sin embargo, en la práctica, para modelizar fenómenos reales, el supuesto de independencia puede ser algo restrictivo.
- En el otro extremo, permitir interacciones arbitrarias entre los X_i, hace con que sea muy dificil hacer cálculos, mismo que elementales.
- Una cadena de Markov es una secuencia de variables aleatorias que presenta la dependencia en un paso, presentando un equilibrio entre independencia total y dependencia total.

Definición (Cadena de Markov)

Una secuencia de variables aleatorias $X_0, X_1, X_2, \ldots$ definida en un espacio de estados discreto $\{1, 2, \ldots, N\}$ es una cadena de Markov si

$$\Pr(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \Pr(X_{n+1} = j \mid X_n = i), \quad \forall n \geq 0.$$

Interpretación: el futuro es independiente del pasado dado el presente.

- La probabilidad $\Pr(X_{n+1} = j \mid X_n = i)$ se llama probabilidad de transición del estado i para el estado j.
- Asumiremos homogeneidad en el tiempo, que significa decir que la probabilidad de transición $\Pr(X_{n+1} = j \mid X_n = i)$ es la misma para todo el n. En particular,

$$\Pr(X_{n+1} = j \mid X_n = i) = \Pr(X_1 = j \mid X_0 = i).$$

Definición (Matriz de transición)

Sea $X_0, X_1, X_2, \ldots$ una cadena de Markov con espacio de estados $\{1, 2, \ldots, N\}$ y sea $p_{ij} = \Pr(X_{n+1} = j \mid X_n = i) = \Pr(X_1 = j \mid X_0 = i)$ la probabilidad de transición del estado i para el estado j. La matriz $P = (p_{ij})$, de dimensión $N \times N$, se llama matriz de transición de la cadena.

Nota: P es una matriz no negativa y la suma de cada una de sus filas es 1.

Definición (Probabilidad de transición a n pasos)

La probabilidad de transición a n pasos del estado i para el estado j, la cual vamos a denotar por $p_{ij}^{(n)}$, es la probabilidad de la cadena estar en el estado j exactamente n pasos después de haber estado en i

$$p_{ij}^{(n)} = \Pr(X_n = j \mid X_0 = i).$$

Se prueba que $p_{ij}^{(n)}$ es la (i, j)-ésima entrada de $P^{(n)}$.

- Una cadena de Markov queda completamente caracterizada por
  1. La distribución inicial $\Pr(X_0 = i)$, $i = \{1, \ldots, N\}$.
  2. La matriz de transción P.

- Una cadena de Markov es
  - **Irreducible**: si para todos los estados i y j, existe n tal que $\Pr(X_n = j \mid X_0 = i) > 0$.
  - **Recurrente**: si existe n tal que $\Pr(X_n = i \mid X_0 = i) = 1$, $\forall i$. La cadena es positivo-recurrente si el tempo esperado de retorno a i es finito.
  - **Aperiódica**: El período de un estado i de una cadena de Markov es el máximo comun divisor del número posible de pasos para volver a i partindo de i. Es decir, el período de i es el máximo comun divisor de los números n tales que $P^n(i,i) > 0 \Leftrightarrow \Pr(X_n = i \mid X_0 = i) > 0$. Un estado es llamado de aperiódico si su período es 1. Consecuentemente, la cadena es aperiódica si todos los estados son aperiódicos.

Definición (Cadena de Markov ergódica)

Si la cadena de Markov es irreducible, positivo-recurrente y aperiódica, entonces la cadena se denomina ergódica.

Definición (Distribución estacionaria)

Una distribución estacionaria de una cadena de Markov con matriz de transición P se define como un vector $\pi = (\pi_1, \ldots, \pi_N)$, cuyo j-ésimo componente ($j = 1, \ldots, N$) está dado por

$$\pi_j = \lim_{n \to \infty} \Pr(X_n = j \mid X_0 = i).$$

Si la cadena de Markov es ergodica, se demuestra que las probabilidades π_j son la única solución del sistema de ecuaciones

$$\pi_j = \sum_{i=1}^{N} \pi_i p_{ij}, \quad j = 1, \ldots, N,$$

o equivalentemente

$$\pi = \pi P.$$

- El objetivo de los métodos de Monte Carlo vía Cadenas de Markov (MCMC, del ingles *Markov chain Monte Carlo*) es construir una cadena de Markov cuya distribución estacionaria es la distribución de interés (la distribución en la cual estamos interesados en simular).
- Así, una muestra de la cadena de Markov es una muestra de la distribución de interés.

Teorema (Teorema Ergódico)

Si la secuencia de vectores simulados $\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$ *es una realización de una cadena de Markov ergódica cuya distribución estacionaria es* π*, entonces*

$$\frac{1}{M}\sum_{j=0}^{M} g(\mathbf{x}^{(j)}) \longrightarrow \mathbb{E}_{\pi}\{g(\mathbf{X})\} = \int g(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}.$$

El teorema ergódico es una generalización de la ley fuerte de los grandes números para cadenas de Markov.

- El muestreador de Gibbs (o *Gibbs sampler*) fue originalmente propuesto por Geman & Geman (1984).
- En un contexto general, suponga que estamos interesados en obtener una muestra de una distribución multivariada $\pi(\mathbf{x}) = \pi(x_1, \ldots, x_d)$.
- Si es facil determinar las distribuciones condicionales de

$$\begin{gathered}\pi(x_1 \mid x_2, x_3 \ldots, x_d) \\ \pi(x_2 \mid x_1, x_3, , \ldots x_d) \\ \vdots \\ \pi(x_d \mid x_1, x_2, \ldots, x_{d-1})\end{gathered}$$

y si es fácil muestrear desde estas distribuciones entonces podemos ocupar el siguiente algoritmo.

Algoritmo

1. *Especificar valores iniciales* $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_d^{(0)})$. *Hacer i=1.*
2. *Para* $i = 2, \ldots, M$
   - *Generar* $x_1^{(i)}$ *de la dist. condicional*
     $$\pi(x_1 \mid x_2^{(i-1)}, x_3^{(i-1)}, \ldots, x_d^{(i-1)})$$
   - *Generar* $x_2^{(i)}$ *de la dist. condicional*
     $$\pi(x_2 \mid x_1^{(i)}, x_3^{(i-1)}, \ldots, x_d^{(i-1)})$$
   - ...
   - *Generar* $x_d^{(i)}$ *de la dist. condicional*
     $$\pi(x_d \mid x_1^{(i)}, x_2^{(i)}, \ldots, x_{d-1}^{(i)})$$

Algoritmo

Si π corresponde a una dist. a posteriori entonces se aplica el mismo algoritmo pero reemplazando

$$\pi(x_j \mid x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_d^{(i-1)})$$

por

$$\pi(\theta_j \mid \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_d^{(i-1)}, \mathbf{x})$$

Ejemplo

- *Generar una muestra de una dist. normal bivariada estándar con correlación ρ, cuya función de densidad de probabilidad está dada por*

$$\pi(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\right\},$$

usando el muestreador de Gibbs.

- *Para implementar el muestreador de Gibbs, en este caso, necesitamos de conocer las distribuciones condicionales de $x_1 \mid x_2$ y de $x_2 \mid x_1$.*

Ejemplo

- *Determinemos entonces* $x_1 \mid x_2$

$$\begin{aligned}
\pi(x_1 \mid x_2) &\propto \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2)\right\} \\
&= \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + \rho^2 x_2^2 - \rho^2 x_2^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + \rho^2 x_2^2)\right\} \\
&= \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1 - \rho x_2)^2\right\},
\end{aligned}$$

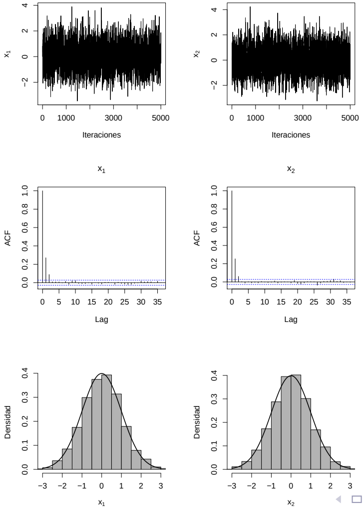
el cual reconocemos como el núcleo de una dist. normal con media ρx_2 *y varianza* $1-\rho^2$.

- *Así,* $x_1 \mid x_2 \sim N(\rho x_2, 1-\rho^2)$.
- *De manera similar,* $x_2 \mid x_1 \sim N(\rho x_1, 1-\rho^2)$.

Ejemplo

1. *Inicializar* $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)})$. *Hacer* $i = 1$.
2. *Para* $i = 1, \ldots, M$
   - $x_1^{(i)} \sim N\left(\rho x_2^{(i-1)}, 1 - \rho^2\right)$,
   - $x_2^{(i)} \sim N\left(\rho x_1^{(i)}, 1 - \rho^2\right)$.

```
set.seed(123); M=5000
x1=x2=numeric(M)
x1[1]=x2[1]=0
rho=0.5
for(i in 2:M){
x1[i]=rnorm(1,rho*x2[i-1],sqrt(1-rho**2))
x2[i]=rnorm(1,rho*x1[i],sqrt(1-rho**2))
}

plot(1:M,x1,type="l"); plot(1:M,x2,type="l")
acf(x1); acf(x2)
```


x_1
Iteraciones
x_2
Iteraciones
x_1
ACF
Lag
x_2
ACF
Lag
Densidad
x_1
Densidad
x_2

Ejemplo

- *Consideremos que* $x_1 \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$*, con ambos* μ *y* σ^2 *desconocidos.*
- *En este ejemplo el objetivo es hacer inferencia sobre* (μ, σ^2).
- *Para eso necesitamos de obtener una muestra de la dist. a posteriori* $p(\mu, \sigma^2 \mid \mathbf{x})$.
- *La verosimilitud es*

$$L(\mu, \sigma^2; \mathbf{x}) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right] \right\}$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

Ejemplo

- *La dist. a posteriori es*
$$p(\mu, \sigma^2 \mid \mathbf{x}) \propto L(\mu, \sigma^2; \mathbf{x}) p(\mu, \sigma^2).$$
- *Hay varias posibilidades para* $p(\mu, \sigma^2)$.
- *Consideremos que a priori* μ *y* σ^2 *son independientes, es decir,* $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$.
- *Consideremos además que* $\mu \sim N(\mu_0, \sigma_0^2)$ *y que* $\sigma^2 IG(a, b)$ *donde IG denota una dist. Gamma-inversa. Así,*

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\}, \quad p(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}.$$

Ejemplo

- *Luego,*

$$p(\mu, \sigma^2 \mid \mathbf{x}) \propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$
$$\times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$
$$\times \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}.$$

- *Ya hemos visto que*

$$\mu \mid \sigma^2, \mathbf{x} \sim N\left(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\right).$$

Ejemplo

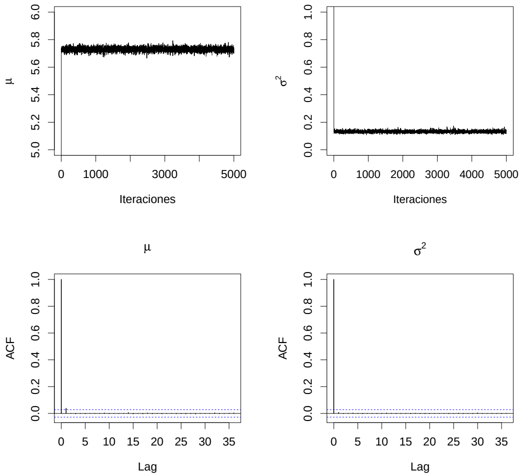
- *Determinemos ahora $p(\sigma^2 \mid \mu, \mathbf{x})$. Tenemos que*

$$
\begin{aligned}
p(\sigma^2 \mid \mu, \mathbf{x}) &\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \times \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\} \\
&\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}(\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\} \\
&= (\sigma^2)^{-(a+n/2+1)} \exp\left\{-\frac{1}{\sigma^2}\left(b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right\}.
\end{aligned}
$$

- *Así,*

$$
\sigma^2 \mid \mu, \mathbf{x} \sim IG\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right).
$$

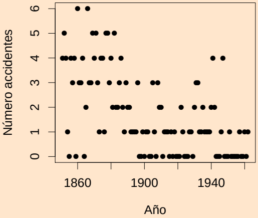
```
data=read.table(file="cholesterol.txt",dec=",",header=TRUE)
chol=data[,1]; lchol=log(chol); n=length(lchol)

require(pscl) #to generate random numbers from an inverse gamma
M=5000
mu=sigma2=numeric(M)
mu0=0; sigma02=100
a=b=0.1
mu[1]=rnorm(1); sigma2[1]=rigamma(1,a,b)
for(i in 2:M){
meanmu=((mu0/sigma02)+((n*mean(lchol))/(sigma2[i-1])))/((1/sigma02)+(n/sigma2[i-1]))
varmu=1/((1/sigma02)+(n/sigma2[i-1]))
mu[i]=rnorm(1,meanmu,sqrt(varmu))

a1=a+(n/2)
b1=b+0.5*sum((lchol-mu[i])**2)
sigma2[i]=rigamma(1,a1,b1)
}

mean(mu)
5.72913
quantile(mu,c(0.025,0.975))
5.699633 5.759370

mean(sigma2)
0.1335819
quantile(sigma2,c(0.025,0.975))
0.1175706 0.1481610
```



μ
6.0
5.8
5.6
5.4
5.2
5.0
0
1000
3000
5000
Iteraciones
σ2
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
Iteraciones
μ
ACF
0
5
10
15
20
25
30
35
Lag
σ2
ACF
Lag

Ejemplo

- *Se realizó un estudio sobre el número de accidentes en minas de carbón durante 112 años (1851-1962) en Reino Unido (datos analisados en Carlin, Gelfand & Smith, 1992).*
- *Se observa que hay un número de accidentes relativamente elevado en los primeros años y un número relativamente bajo en los últimos años.*
- *La pregunta que se plantea es cuando las mejoras en tecnología y en los procedimientos de seguridad tuvieran un efecto positivo en el número de accidentes.*



Ejemplo

- *El siguiente modelo fue propuesto por Carlin, Gelfand & Smith (1992)*

$$\begin{cases} x_i \mid \lambda \sim Po(\lambda), & i = 1, \ldots, k, \\ x_i \mid \phi \sim Po(\phi), & i = k+1, \ldots, n. \end{cases}$$

- *El número de accidentes por año sigue una dist. de Poisson. El número medio de accidentes en los primeros k años es λ, mientras que en los restantes $n-k$ años la media es ϕ.*
- *El objetivo desde el punto de vista estadístico es estimar k, el momento de cambio en la tendencia (changepoint), y también λ y ϕ.*
- *Las distribuciones a priori (independientes) que se consideran son*

$$\begin{aligned} &\lambda \sim Gamma(a, b), \\ &\phi \sim Gamma(c, d), \\ &k \sim UD\{1, \ldots, n\} \quad (UD{=}Uniforme\ Discreta). \end{aligned}$$

Ejemplo

- *Luego,*

$$
\begin{aligned}
p(\lambda, \phi, k \mid \mathbf{x}) &\propto L(\lambda, \phi, k; \mathbf{x}) p(\lambda) p(\phi) p(k) \\
&\propto \left\{ \prod_{i=1}^{k} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right\} \left\{ \prod_{i=k+1}^{n} \frac{e^{-\phi} \phi^{x_i}}{x_i!} \right\} \lambda^{a-1} e^{-b\lambda} \phi^{c-1} e^{-d\phi} \frac{1}{n} \\
&\propto e^{-\lambda k} \lambda^{\sum_{i=1}^{k} x_i} e^{-\phi(n-k)} \phi^{\sum_{i=k+1}^{n} x_i} \lambda^{a-1} e^{-b\lambda} \phi^{c-1} e^{-d\phi}.
\end{aligned}
$$

- *Así,*

$$
\begin{aligned}
p(\lambda \mid \phi, k, \mathbf{x}) &\propto e^{-\lambda k} \lambda^{\sum_{i=1}^{k} x_i} \lambda^{a-1} e^{-b\lambda} \\
&= \lambda^{a + \sum_{i=1}^{k} x_i - 1} e^{-\lambda(b+k)} \\
&\propto Gamma \left(a + \sum_{i=1}^{k} x_i, b + k \right)
\end{aligned}
$$

Ejemplo

- *De manera similar,*

$$\phi \mid \lambda, k, \mathbf{x} \sim Gamma\left(c + \sum_{i=k+1}^{n} x_i, d + n - k\right).$$

- *La dist. condicional de k resulta ser la más complicada*

$$\begin{aligned} p(k \mid \lambda, \phi, \mathbf{x}) &\propto e^{-\lambda k} \lambda^{\sum_{i=1}^{k} x_i} e^{\phi k} \phi^{\sum_{i=k+1}^{n} x_i} \\ &= e^{-k(\lambda-\phi)} \lambda^{\sum_{i=1}^{k} x_i} \phi^{\sum_{i=1}^{n} x_i - \sum_{i=1}^{k} x_i} \\ &\propto e^{-k(\lambda-\phi)} \left(\frac{\lambda}{\phi}\right)^{\sum_{i=1}^{k} x_i}. \end{aligned}$$

- *Luego,*

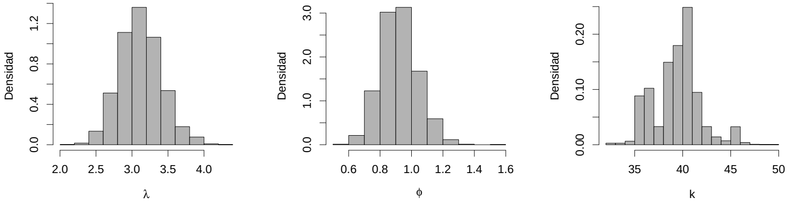
$$p(k \mid \lambda, \phi, \mathbf{x}) = \frac{e^{-k(\lambda-\phi)} \left(\frac{\lambda}{\phi}\right)^{\sum_{i=1}^{k} x_i}}{\sum_{j=1}^{n} e^{-j(\lambda-\phi)} \left(\frac{\lambda}{\phi}\right)^{\sum_{i=1}^{j} x_i}}.$$

```
k=20; a=0.1; b=0.1; c=0.1; d=0.1
M=5000; output=matrix(0,nrow=M,ncol=3); set.seed(123)

for(i in 1:M){
lambda=rgamma(1,a+sum(x[1:k]),b+k)
phi=rgamma(1,c+sum(x[(k+1):n]),d+n-k)
post.dist=rep(0,n)
for(j in 1:n){
post.dist[j]=exp(-j*(lambda-phi))*((lambda/phi)**sum(x[1:j]))
}
post.dist=post.dist/sum(post.dist)
k=sample(1:n,size=1,prob=post.dist)
output[i,]=c(lambda,phi,k)
}

plot(1:M,output[,1],type="l"); plot(1:M,output[,2],type="l")
plot(1:M,output[,3],type="l")

acf(output[,1]); acf(output[,2]); acf(output[,3])
```

```
mean(output[,1])
3.118172

mean(output[,2])
0.9214093

mean(output[,3])
39.9218

quantile(output[,1],c(0.025,0.975))
2.581782 3.722834

quantile(output[,2],c(0.025,0.975))
0.7060668 1.1704523

quantile(output[,3],c(0.025,0.975))
36 46
```

MCMC

Muestreador de Gibbs



Ejemplo

- *Datos referentes a animales distribuidos multinomialmente en cuatro categorias.*
- *Sea x_i el número total de animales en la categoria i ($i = 1, \ldots, 4$) y sea p_i la probabilidade de un animal pertencer a la categoria i.*
- *Así,*

$$\mathbf{x} = (x_1, x_2, x_3, x_4),$$

con probabilidades

$$(p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right), \quad 0 < \theta < 1.$$

- *El objetivo es estimar θ desde una perspectiva Bayesiana.*

Ejemplo

- *Los datos claramente siguen una dist. binomial*

$$
\begin{aligned}
L(\theta; \mathbf{x}) &= \frac{(x_1 + x_2 + x_3 + x_4)!}{x_1!x_2!x_3!x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \\
&\propto \left(\frac{1}{2} + \frac{\theta}{4}\right)^{x_1} \left(\frac{1}{4}(1-\theta)\right)^{x_2+x_3} \left(\frac{\theta}{4}\right)^{x_4} \\
&\propto (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4}.
\end{aligned}
$$

- *Consideremos* $\theta \sim$ *Beta*(α, β)*.*
- *Así,*

$$
\begin{aligned}
p(\theta \mid \mathbf{x}) &\propto (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\
&= (2+\theta)^{x_1} \theta^{\alpha+x_4-1} (1-\theta)^{\beta+x_2+x_3-1}
\end{aligned}
$$

- *Como se puede observar,* $p(\theta \mid \mathbf{x})$ *no corresponde a ninguna dist. conocida.*

Ejemplo

- *Una manera de enfrentar este problema es la siguiente:*
    - *Supongamos que la primera categoría se puede dividir en dos subcategorias A y B y suponga además que* $p_A = 1/2$*(prob. de un animal pertenecer a la subcategoria a) y* $p_B = \theta/4$*(prob. de un animal pertenecer a la subcategoria b).*
    - *Sea z el número (desconocido) total de animales que pertenecen a la categoria A. Como consecuencia el número total de animales en la categoría B es* $x_1 - z$*.*
- *Al vector*

$$(z, x_1 - z, x_2, x_3, x_4),$$

se llama conjunto de datos aumentados.

Ejemplo

- *La verosimilitud de los datos aumentados es*

$$\begin{aligned} L(\theta \mid \mathbf{x}; z) &\propto \left(\frac{1}{2}\right)^{z} \left(\frac{\theta}{4}\right)^{x_1 - z} \left(\frac{1}{4}(1-\theta)\right)^{x_2+x_3} \left(\frac{\theta}{4}\right)^{x_4} \\ &\propto \theta^{x_1 - z + x_4}(1-\theta)^{x_2+x_3}. \end{aligned}$$

- *La dist. a posteriori es*

$$\begin{aligned} p(\theta \mid \mathbf{x}, z) &\propto \theta^{x_1 - z + x_4}(1-\theta)^{x_2+x_3}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha + x_1 - z + x_4 - 1}(1-\theta)^{\beta + x_2 + x_3 - 1}, \end{aligned}$$

donde concluimos que

$$\theta \mid \mathbf{x}, z \sim Beta(\alpha + x_1 - z + x_4, \beta + x_2 + x_3).$$

Ejemplo

- *Si conoceriamos z, podriamos simular desde la dist. condicional de θ.*
- *Pero, aun que no conozcamos z, conocemos su dist. condicional:*
  - *Hay x_1 animales en las categorías A y B.*
  - *La prob. condicional de que un animal este en la categoria A dado que esta en la categoria 1 es*

$$\frac{\Pr(\textit{pertenecer categoria A})}{\Pr(\textit{pertenecer categoria 1})} = \frac{1/2}{1/2 + \theta/4} = \frac{2}{2+\theta}.$$

  - *Así,*

$$z \mid \theta, \mathbf{x} \sim Bin\left(y_1, \frac{2}{2+\theta}\right).$$

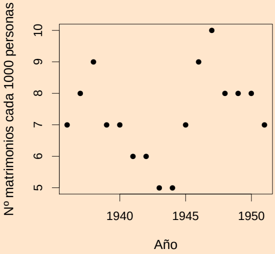
Ejemplo

1. *Especificar* $\theta^{(0)}$. *Hacer* $i = 1$.
2. *Para* $i = 1, \ldots, M$
   - $z \sim Binomial\left(x_1, \frac{2}{2+\theta^{(i-1)}}\right)$.
   - $\theta^{(i)} \sim Beta(\alpha + x_1 - z + x_4, \beta + x_2 + x_3)$.

```
x=c(125,18,20,34)
set.seed(123); M=5000
theta=numeric(M); theta[1]=0.1
alpha=beta=1
for(i in 2:M){
z=rbinom(1,x[1],2/(2+theta[i-1]))
theta[i]=rbeta(1,alpha+x[1]-z+x[4],beta+x[2]+x[3])
}

plot(1:M,theta,type="l"); acf(theta)

mean(theta[101:M])
0.6225588

quantile(theta[101:M],c(0.025,0.975))
0.5213662 0.7185131
```

Ejemplo (modelo jerárquico)

- *Se consideran el número de matrimonios cada 1000 personas en Italia desde el año 1936 hasta el año 1951.*
- *La pregunta que se plantea es si es correcto modelizar las tasas de matrimonio durante la guerra mundial (1939–1945) del mismo modo que antes y después de ella.*



Ejemplo (modelo jerárquico)

- *Sea x_i el número el número de matrimonios en el año i.*
- *Consideremos*

$$\begin{aligned} x_i \mid \lambda_i &\sim Po(\lambda_i), \quad i = 1, \ldots, n \\ \lambda_i \mid \beta &\sim Exp(\beta), \quad i = 1, \ldots, n \\ \beta &\sim Exp(1). \end{aligned}$$

- *Asumimos que $\lambda_1, \ldots, \lambda_n$ son independientes, es decir*

$$p(\lambda_1, \ldots, \lambda_n \mid \beta) = \prod_{i=1}^{n} p(\lambda_i \mid \beta).$$

- *La dist. a posteriori es*

$$\begin{aligned} p(\boldsymbol{\lambda}, \beta \mid \mathbf{x}) &\propto L(\boldsymbol{\lambda}, \beta; \mathbf{x}) p(\boldsymbol{\lambda}, \beta) \\ &= L(\boldsymbol{\lambda}; \mathbf{x}) p(\boldsymbol{\lambda} \mid \beta) p(\beta) \end{aligned}$$

Ejemplo (modelo jerárquico)

- *Así,*

$$p(\boldsymbol{\lambda}, \beta \mid \mathbf{x}) \propto \prod_{i=1}^{n} \left\{ \frac{e^{-\lambda_i} \lambda_i^{x_i}}{x_i!} \right\} \prod_{i=1}^{n} \left\{ \beta e^{-\beta \lambda_i} \right\} e^{-\beta}.$$

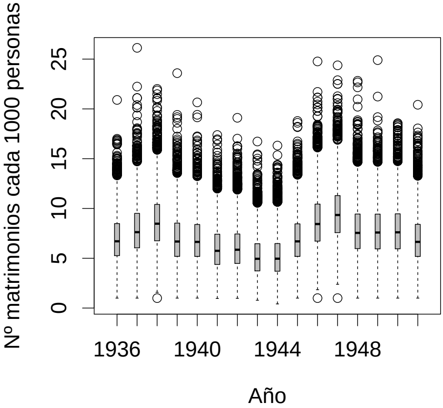
- *Sea* $\boldsymbol{\lambda}_{(-i)} = (\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_n)$.
- *La dist. condicional a posteriori para* λ_i *es*

$$\begin{aligned} p(\lambda_i \mid \boldsymbol{\lambda}_{(-i)}, \beta, \mathbf{x}) &\propto e^{-\lambda_i} \lambda_i^{x_i} e^{-\beta \lambda_i} \\ &= \lambda_i^{x_i + 1 - 1} e^{-\lambda_i (1+\beta)} \\ &\propto Gamma(x_i + 1, 1 + \beta), \quad i = 1, \ldots, n \end{aligned}$$

- *La dist. condicional a posteriori para* β *es*

$$\begin{aligned} p(\beta \mid \boldsymbol{\lambda}, \mathbf{x}) &\propto \beta^n e^{-\beta \sum_{i=1}^{n} \lambda_i} e^{-\beta} \\ &= \beta^{n+1-1} e^{-\beta (1 + \sum_{i=1}^{n} \lambda_i)} \\ &\propto Gamma(n + 1, 1 + \sum \lambda_i) \end{aligned}$$

```
x=c(7,8,9,7,7,6,6,5,5,7,9,10,8,8,8,7); n=length(x)
set.seed(123); M=5000
lambda=matrix(0,nrow=M,ncol=n); beta=numeric(M)
beta[1]=1; lambda[1,]=rep(1,n)
for(j in 2:M){
for(i in 1:n){
lambda[j,i]=rgamma(1,1+x[i],1+beta[j-1])
}
beta[j]=rgamma(1,n+1,1+sum(lambda[j,]))
}
```



Nº matrimonios cada 1000 personas
0
5
10
15
20
25
1936
1940
1944
1948
Año

Algoritmo de Metropolis–Hastings

- El algoritmo de Metropolis–Hastings es una generalización del algoritmo de Metropolis.
- El algoritmo de Metropolis (Metropolis et al.) fue introducido en el año de 1953 y ha sido utilizado por muchos años en la comunidad física.
- El artículo de Hastings en 1970 generaliza el algoritmo de Metropolis para un contexto estadístico.
- El algoritmo quedó conocido como el algoritmo de Metropolis–Hastings.
- El algoritmo es especialmente útil cuando no se puede determinar alguna dist. condicional de interés, no siendo entonces posible utilizar el muestreador de Gibbs.

- La idea principal del algoritmo es generar una cadena de Markov $\{x^{(i)} : i = 0, 1, 2, \ldots\}$ cuya dist. estacionaria sea la dist. de interés, la dist. de la cual queremos simular, llamésmole π.
- Para un determinado estado de la cadena, $x^{(i-1)}$, el algoritmo debe especificar como generar el próximo estado $x^{(i)}$.
- Esto se hace a través de la distribución de propuesta (*proposal distribution*), $q(\cdot \mid x^{(i-1)})$, de la cual simulamos un valor candidato x^*.
- Si el valor candidato x^* es aceptado, entonces la cadena se mueve y hacemos $x^{(i)} = x^*$.
- Por otra parte si x^* es rechazado, la cadena permanece en el estado actual y $x^{(i)} = x^{(i-1)}$.
- Por ejemplo, si la dist. de propuesta es la dist. normal, entonces una opción natural para $q(\cdot \mid x^{(i-1)})$ podría ser $\mathrm{N}(\mu = x^{(i-1)}, \sigma^2)$, para algún valor de σ^2 fijo.
- La dist. de propuesta depende del tipo de problema y debe obedecer a ciertas condiciones de regularidad de manera que se tenga la convergencia de la cadena para la dist. estacionaria. Si, por ejemplo, la dist. de propuesta tiene el mismo soporte que la dist. de la cual queremos simular π, entonces, en general, las condiciones de regularidad son satisfechas.

Algoritmo

1. *Especificar* $x^{(0)}$.
2. *Para* $i = 1, \ldots, M$
   - *Generar* $x^* \sim q(\cdot \mid x^{(i-1)})$.
   - *Calcular la probabilidade de aceptación*

$$\alpha(x^{(i-1)}, x^*) = \min\left\{1, \frac{\pi(x^*)q(x^{(i-1)} \mid x^*)}{\pi(x^{(i-1)})q(x^* \mid x^{(i-1)})}\right\}.$$

   - *Generar* $u \sim U(0,1)$.
   - *Si* $u \leq \alpha(x^{(i-1)}, x^*)$ *aceptar* x^* *y hacer* $x^{(i)} = x^*$*; en otro caso, rechazar* x^* *y hacer* $x^{(i)} = x^{(i-1)}$

- En el contexto Bayesiano, en que el interés es simular de una dist. a posteriori, llamésmole $p(\theta \mid \mathbf{x})$ el algoritmo es identico. Notemos que

$$\begin{aligned}\alpha(\theta^{(i-1)}, \theta^*) &= \min\left\{1, \frac{p(\theta^* \mid \mathbf{x}) q(\theta^{(i-1)} \mid \theta^*)}{p(\theta^{(i-1)} \mid \mathbf{x}) q(\theta^* \mid \theta^{(i-1)})}\right\} \\ &= \min\left\{1, \frac{L(\theta^*; \mathbf{x}) q(\theta^{(i-1)} \mid \theta^*)}{L(\theta^{(i-1)}; \mathbf{x}) q(\theta^* \mid \theta^{(i-1)})}\right\}.\end{aligned}$$

- Como se puede observar no necesitamos de la constante de normalización de la dist. a posteriori.

- El el algoritmo de Metropolis original (Metropolis et al. 1953) simplemente distribuciones de propuesta simétricas son consideradas, es decir,

$$q(x^* \mid x^{(i-1)}) = q(x^{(i-1)} \mid x^*).$$

- Por ejemplo, una dist. de propuesta simétrica es $q(x^* \mid x^{(i-1)}) = \mathrm{N}(\mu = x^{(i-1)}, \sigma^2)$, σ^2 fijo.

- Veamos

$$\begin{aligned} q(x^* \mid x^{(i-1)}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x^* - x^{(i-1)})^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x^{(i-1)} - x^*)^2\right\} \\ &= q(x^{(i-1)} \mid x^*). \end{aligned}$$

- En el caso del algoritmo de Metropolis la probabilidad de aceptación queda simplificada

$$\alpha(x^{(i-1)}, x^*) = \min\left\{1, \frac{\pi(x^*)}{\pi(x^{(i-1)})}\right\}.$$

- En este caso, si $\pi(x^*) \geq \pi(x^{(i-1)})$, entonces la cadena si mueve para x^* una vez que $\alpha(x^{(i-1)}, x^*) = 1$.

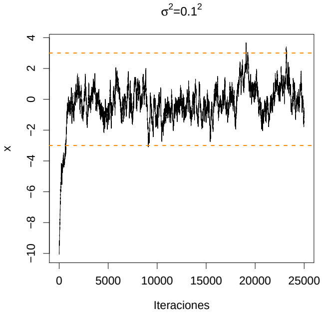
- Por otra parte, si $\pi(x^*) \leq \pi(x^{(i-1)})$, entonces la cadena si mueve para x^* con probabilidad α.

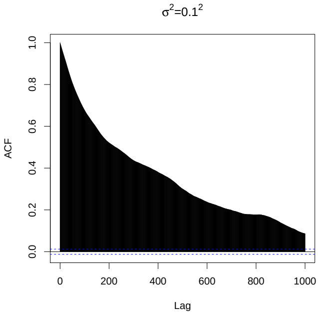
Ejemplo

- *En cualquier de los dos algoritmos (Metropolis o Metropolis–Hastings) es importante entender como la escala de la dist. de propuesta (cuando la dist. de propuesta depende de un parámero de escala) afecta la eficiencia del algoritmo.*
- *Consideremos el ejemplo de simular desde una dist. normal.*
- *Obviamente que no necesitamos de un algoritmo del tipo de Metropolis–Hastings para simular desde una dist. normal, pero el ejemplo sirve para illustrar la importancia de la elección del parámetro de escala (cuando este esté presente).*
- *Así, consideremos que la dist. de la cual queremos simular,* π*, es la dist.* $N(0, 1)$.
- *Consideremos además* $q(\mid x^{(i-1)}) = N(\mu = x^{(i-1)}, \sigma^2)$.
- *Iremos considerar* $\sigma^2 = 0.1^2, 2.5^2, 50^2$.

```
set.seed(123); M=25000; x=numeric(M)
x[1]=-10; sigma=2.5; t=0
for(i in 2:M){
xstar=rnorm(1,mean=x[i-1],sd=sigma)
num=dnorm(x=xstar,0,1)
den=dnorm(x=x[i-1],0,1)
alpha=min(1,num/den)
u=runif(1)
if(u<=alpha){x[i]=xstar;t=t+1}
else{x[i]=x[i-1]}
}
```

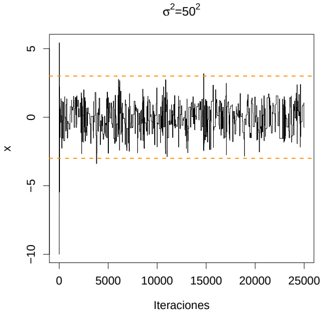
- $\sigma^2 = 0.1^2$; tasa de aceptación: 96.732%.

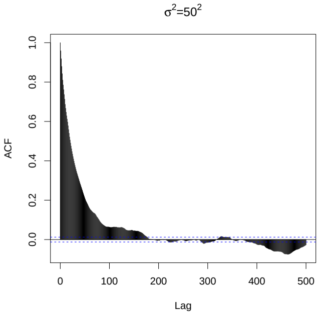




- Para $\sigma^2 = 0.1^2$, la tasa de aceptación es muy elevada, cerca de 97%.
- La cadena, aún que se mueva mucho, no explora bien todo el espacio de posible valores.
- También se puede observar que la cadena tarda cerca de 1000 iteraciones a llegar a la dist. estacionaria.
- Como resultado de la elevada tasa de aceptación, la autocorrelación de la cadena es muy elevada. En la práctica tendríamos de aplicar un desfasamento entre observaciones superior a 1000.

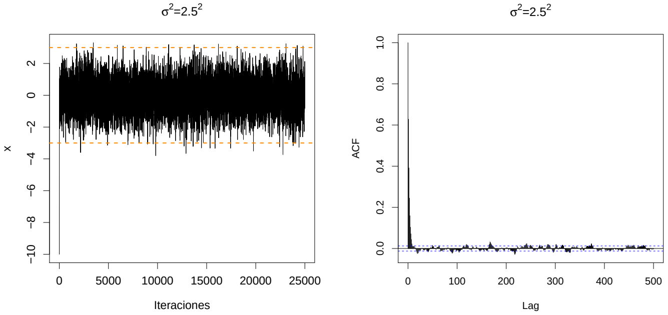
- $\sigma^2 = 50^2$; tasa de aceptación: 2.596%





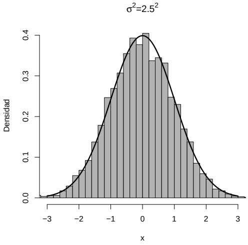
- En el otro extremo, cuando $\sigma^2 = 50^2$, la tasa de aceptación es muy baja, cerca de 2.6%.
- Como consecuencia de la baja tasa de aceptación, la cadena permanece por varias iteraciones en el mismo valor.
- La autocorrelación de la cadena también es elevada, aún que no tanto como en el caso anterior. Un desfasamento entre observaciones de cerca de 200 tendría de ser aplicado.

- $\sigma^2 = 2.5^2$; tasa de aceptación: 42.84%



MCMC
Algoritmo de Metropolis–Hastings

- Para $\sigma^2 = 2.5^2$, la tasa de aceptación es de cerca de 43%.
- Se observa que la cadena explora bastante bien todos los valores posibles y que la cadena rapidamente llega a la dist. estacionaria.
- La función de autocorrelación también se ve bien, aun que un desfasamento de cerca de 100 observaciones seria conveniente.
- En general, las tasas de aceptación deben estar entre 20% y 50%.
- Veamos como queda el ajuste para este caso.



Ejemplo

- *Simular desde la distribución lognormal estándar, cuya función densidad de probabilidad está dada por*

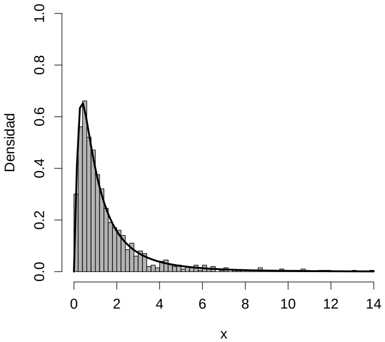
$$\pi(x) = \frac{1}{x\sqrt{2\pi}} \exp\left\{ -\frac{(\log x)^2}{2} \right\}, \quad x \geq 0.$$

- *Para dist. de propuesta consideremos una dist. normal truncada en cero y centrada en el estado anterior de la cadenas, es decir*

$$q(\cdot \mid x^{(i-1)}) = NT_{(0,\infty)}(\mu = x^{(i-1)}, \sigma^2),$$

con σ^2 fijo.

```
require(msm)
M=5000; x=numeric(M)
x[1]=rtnorm(1,0,1,lower=0); sigma2=2; t=0
for(i in 2:M){
xstar=rtnorm(1,mean=x[i-1],sd=sqrt(sigma2),lower=0)
num=dlnorm(x=xstar,0,1)*dtnorm(x=x[i-1],mean=xstar,sd=sqrt(sigma2),lower=0)
den=dlnorm(x=x[i-1],0,1)*dtnorm(x=star,mean=x[i-1],sd=sqrt(sigma2),lower=0)
alpha=min(1,num/den)
u=runif(1)
if(u<=alpha){x[i]=xstar;t=t+1}
else{x[i]=x[i-1]}
}

plot(1:M,x,type="l"); acf(x,lag.max=100)
x1=x[seq(101,M,50)]; acf(x1)
```



Densidad
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
x

Ejemplo

- *Simular desde una dist. de Rayleigh con parámetro de escala σ^2, cuya función de densidad de probabilidad está dada por*

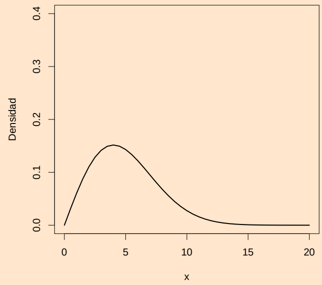
$$\pi(x \mid \sigma^2) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, \quad x \geq 0, \quad \sigma > 0.$$

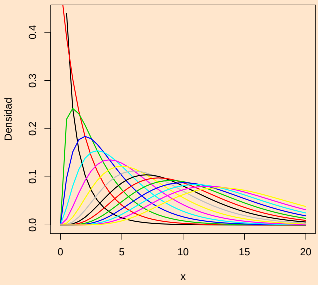
- *Consideremos $\sigma = 4$ (p.e.) y como distribución de propuesta utilizaremos una dist. chi cuadrado con grados de libertad iguales al estado anterior de la cadena, o sea*

$$q(\cdot \mid x^{(i-1)}) = \chi^2_{x^{(i-1)}}.$$

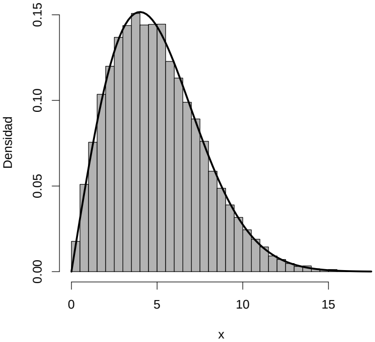
Ejemplo

- *Izquierda: Densidad de la dist. Rayleigh con $\sigma = 4$. Derecha: Densidad de la dist. chi cuadrado con diferentes grados de libertad.*





```
dray=function(x,sigma){
ifelse(x>=0,(x/(sigma**2))*exp((-x**2)/(2*sigma**2)),0)
}

set.seed(123); M=50000; x=numeric(M)
x[1]=rchisq(1,df=1); t=0; sigma=4
for(i in 2:M){
xstar=rchisq(1,df=x[i-1])
num=dray(xstar,sigma=sigma)*dchisq(x=x[i-1],df=xstar)
den=dray(x[i-1],sigma=sigma)*dchisq(x=xstar,df=x[i-1])
alpha=min(1,num/den)
u=runif(1)
if(u<=alpha){x[i]=xstar; t=t+1}
else{x[i]=x[i-1]}
}

t/M; plot(1:M,x,type="l"); acf(x,lag.max=500)
```



0.00
0.05
0.10
0.15
Densidad
0
5
10
15
x

Ejemplo

- *Vamos a volver al ejemplo de datos multinomiales presentado anteriormente.*
- *Tenemos que*
$$p(\theta \mid \mathbf{x}) \propto (2+\theta)^{x_1}\theta^{a+x_4-1}(1-\theta)^{b+x_2+x_3-1}.$$
- *Para dist. de propuesta consideremos una dist. normal truncada en el intervalo* $(0,1)$ *y centrada en el estado anterior de la cadena, es decir*
$$q(\cdot \mid \theta^{(i-1)}) = NT_{(0,1)}(\mu = \theta^{(i-1)}, \sigma^2),$$
con σ^2 *fijo.*

```
post=function(theta,a,b,x){
((2+theta)**x[1])*(theta**(a+x[4]-1))*((1-theta)**(b+x[2]+x[3]-1))
}
x=c(125,18,20,34)
require(msm)
M=10000; theta=numeric(M); theta[1]=runif(1); t=0; set.seed(123)
sigma2=0.1
for(i in 2:M){ thetastar=rtnorm(1,mean=theta[i-1],sd=sqrt(sigma2),lower=0,upper=1)
num=post(thetastar,a=1,b=1,x=x)*dtnorm(x=theta[i-1],mean=thetastar,sd=sqrt(sigma2),lower=0,upper=1
den=post(theta[i-1],a=1,b=1,x=x)*dtnorm(x=thetastar,mean=theta[i-1],sd=sqrt(sigma2),lower=0,upper=
alpha=min(1,num/den)
u=runif(1)
if(u<=alpha){theta[i]=thetastar; t=t+1}
else{theta[i]=theta[i-1]}
}
t/M
plot(1:M,theta,type="l"); acf(theta)
mean(theta)
0.6199123
quantile(theta,c(0.025,0.975))
0.5156620 0.7178733
```

- El muestreador independiente (*independence sampler*) es un caso particular del algoritmo de Metropolis–Hastings y fue propuesto por Tierney (1994).
- En el muestreador independiente, la dist. de propuesta no depende del estado anterior de la cadena, es decir,
$$q(x^* \mid x^{(i-1)}) = q(x^*).$$
- La probabilidad de aceptación queda así,
$$\alpha(x^{(i-1)}, x^*) = \min\left\{1, \frac{\pi(x^*)q(x^{(i-1)})}{\pi(x^{(i-1)})q(x^*)}\right\}.$$
- De notar que, aún que generemos valores candidatos que no dependen del estado anterior de la cadena, la cadena resultante no es independiente, pues la probabilidad de aceptación aún depende de $x^{(i-1)}$.
- En general, q debe ser una buena aproximación de π, y Gilks et al. (1996) demostraran que es preferible que q tenga colas más pesadas que π.
- Para el muestreador independiente, dado que la dist. de propuesta no depende de $x^{(i-1)}$, cuanto mayor sea la tasa de aceptación, mejor.

Ejemplo

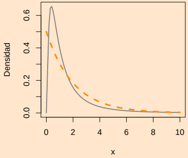
- *Simular desde una dist. lognormal estándar, cuya función de densidad de probabilidad está dada por*

$$\pi(x) = \frac{1}{x\sqrt{2\pi}} \exp\left\{-\frac{(\log x)^2}{2}\right\}, \quad x \geq 0,$$

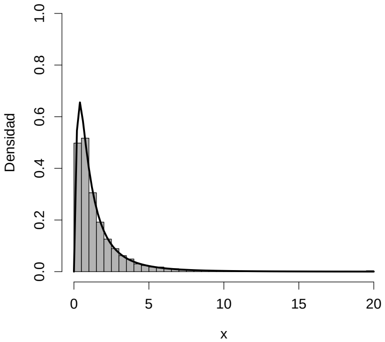
  usando el muestreador independiente y la dist. Gamma como dist. de propuesta.

- *Debemos buscar una configuración de parámetros tal que la dist. Gamma tenga colas más pesadas que la dist. lognormal.*

- *Para $\alpha = 1$ (forma) y $s = 2$ (escala), la dist. Gamma tiene colas más pesadas que la dist. lognormal.*



```
set.seed(123); M=20000
x=numeric(M); x[1]=rgamma(1,2,1); t=0
for(i in 2:M){
xstar=rgamma(1,1,scale=2)
num=dlnorm(xstar,0,1)*dgamma(x[i-1],1,scale=2)
den=dlnorm(x[i-1],0,1)*dgamma(xstar,1,scale=2)
alpha=min(1,num/den)
u=runif(1)
if(u<=alpha){x[i]=xstar; t=t+1}
else{x[i]=x[i-1]}
}

plot(1:M,x,type="l");acf(x)
```


Densidad
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
x

- Hasta ahora hemos utilizado el algoritmo de Metropolis–Hastings (y sus variantes) para simular desde distribuciones univariadas. Sin embargo, es en el caso multivariado que estos algoritmos son más útiles.
- Supongamos, entonces, que queremos simular desde una dist. multivariada $\pi(\mathbf{x}) = \pi(x_1, \ldots, x_d)$ utilizando el algoritmo de Metropolis–Hastings.
- La aceptación/rechazo de los valores generados se puede hacer en bloco o componente a componente.

Algoritmo (Actualización en bloco)

1. *Especificar un valor inicial* $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$.
2. *Para* $i = 1, \ldots, M$
   - *Generar* $\mathbf{x}^* \sim q(\cdot \mid \mathbf{x}^{(i-1)})$, *donde* q *es una dist.* d*-dimensional.*
   - *Probabilidad de aceptación*

$$\alpha(\mathbf{x}^{(i-1)}, \mathbf{x}^*) = \min\left\{1, \frac{\pi(\mathbf{x}^*)q(\mathbf{x}^{(i-1)} \mid \mathbf{x}^*)}{\pi(\mathbf{x}^{(i-1)})q(\mathbf{x}^* \mid \mathbf{x}^{(i-1)})}\right\}.$$

   - *Generar* $u \sim U(0,1)$.
   - *Si* $u \leq \alpha(\mathbf{x}^{(i-1)}, \mathbf{x}^*)$ *aceptar* $\mathbf{x}^*$ *y hacer* $\mathbf{x}^{(i)} = \mathbf{x}^*$; *en otro caso, rechazar* $\mathbf{x}^*$ *y hacer* $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)}$.

Algoritmo (Actualización componente a componente)

1. *Especificar un valor inicial* $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$.
2. *Para* $i = 1, \ldots, M$
   - *Generar* $x_1^* \sim q(\cdot \mid x_1^{(i-1)})$.
   - *Calcular la probabilidad de aceptación de* x_1^*

$$\alpha_1 = \min\left\{1, \frac{\pi(x_1^*, x_2^{(i-1)}, \ldots, x_d^{(i-1)})q(x_1^{(i-1)} \mid x_1^*)}{\pi(x_1^{(i-1)}, x_2^{(i-1)}, \ldots, x_d^{(i-1)})q(x_1^* \mid x_1^{(i-1)})}\right\}.$$

   - *Generar* $u_1 \sim U(0,1)$. *Si* $u_1 \leq \alpha_1$ *hacer* $x_1^{(i)} = x_1^*$*; en otro caso, hacer* $x_1^{(i)} = x_1^{(i-1)}$.
   - ...
   - *Generar* $x_d^* \sim q(\cdot \mid x_d^{(i-1)})$.
   - *Calcular la probabilidad de aceptación de* x_d^*

$$\alpha_d = \min\left\{1, \frac{\pi(x_1^{(i)}, x_2^{(i)}, \ldots, x_d^*)q(x_d^{(i-1)} \mid x_d^*)}{\pi(x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i-1)})q(x_d^* \mid x_d^{(i-1)})}\right\}.$$

   - *Generar* $u_d \sim U(0,1)$. *Si* $u_d \leq \alpha_d$ *hacer* $x_d^{(i)} = x_d^*$*; en otro caso, hacer* $x_d^{(i)} = x_d^{(i-1)}$.

Ejemplo

- *Simular una muestra de la dist. normal estándar bivariada, cuya función de densidad de probabilidad está dada por*

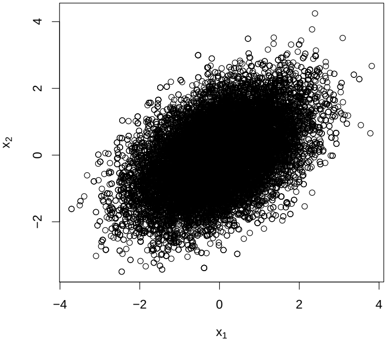
$$\pi(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\right\},$$

- *Utilizaremos un algoritmo de Metropolis–Hastings con actualización en bloco y otro con actualización componente a componente.*

```
dbinom=function(x1,x2,rho){
(1/(2*pi*sqrt(1-rho**2)))*exp(-(1/(2*(1-rho**2)))*(x1**2-2*rho*x1*x2+x2**2))
}

set.seed(123); M=30000; x1=x2=numeric(M); x1[1]=x2[1]=0
Sigma=matrix(c(1.75,0,0,1.75),nrow=2,byrow=T); t=0
require(MASS); rho=0.5
for(i in 2:M){
xstar=mvrnorm(1,c(x1[i-1],x2[i-1]),Sigma)
num=dbinom(xstar[1],xstar[2],rho=rho)
den=dbinom(x1[i-1],x2[i-1],rho=rho)
alpha=min(1,num/den)
u=runif(1)
if(u<=alpha){x1[i]=xstar[1];x2[i]=xstar[2];t=t+1}
else{x1[i]=x1[i-1];x2[i]=x2[i-1] }
}

plot(1:M,x1,type="l"); plot(1:M,x2,type="l")

acf(x1,lag.max=500); acf(x2,lag.max=500)

cor(x1,x2)
```

MCMC

Algoritmo de Metropolis–Hastings



```
set.seed(123); M=30000; x1=numeric(M); x2=numeric(M)
x1[1]=rnorm(1); x2[1]=rnorm(1); sigma12=3; sigma22=3
t1=0; t2=0; rho=0.5

for(i in 2:M){
x1star=rnorm(1,x1[i-1],sqrt(sigma12))
num1=dbinom(x1star,x2[i-1],rho=rho)
den1=dbinom(x1[i-1],x2[i-1],rho=rho)
alpha1=min(1,num1/den1)
u1=runif(1)
if(u1<=alpha1){x1[i]=x1star; t1=t1+1} else{x1[i]=x1[i-1]}

x2star=rnorm(1,x2[i-1],sqrt(sigma22))
num2=dbinom(x1[i],x2star,rho=rho)
den2=dbinom(x1[i],x2[i-1],rho=rho)
alpha2=min(1,num2/den2)
u2=runif(1)
if(u2<=alpha2){x2[i]=x2star; t2=t2+1}
else{x2[i]=x2[i-1]}
}

plot(1:M,x1,type="l"); plot(1:M,x2,type="l")

acf(x1,lag.max=100); acf(x2,lag.max=100)
```