Cómputo Científico
Tarea VII - Comodín

Joel Chacón Castillo

Guanajuato, México

---

1. Punto 1

Sean $x_i \sim Gamma(\alpha, \beta)$, $i = 1, 2, ..., n$. Simular datos x_i con $alpha = 3$ y $\beta = 100$ considerando los casos $n = 3$ y $n = 30$ con $alpha \sim U(1, 4)$, $\beta \sim exp(1)$ distribuciones a priori, se tiene la posterior

$$f(\alpha, \beta|\hat{x}) \propto \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} r_1^{\alpha-1} e^{-\beta(r_2+1)} \mathbb{1}(1 \leq \alpha \leq 4)\mathbb{1}(\beta > 0) \tag{1}$$

En ambos casos, grafica los contornos para visualizar dónde está concentrada la posterior. Utilizar la propuesta:

$$q\left((\alpha_p \quad \beta_p)^T|(\alpha \quad \beta)^T\right) = (\alpha \quad \beta)^T + (\epsilon_1 \quad \epsilon_2) \tag{2}$$

donde $(\epsilon_1 \quad \epsilon_2)^T \sim N_2((0 \quad 0)^T, \delta_i\sigma_j^2)$ $(\delta_i = 1 \text{ si } i = j)$

Comentarios

Inicialmente se observa que la función posterior está compuesta por el producto de la función de verosimilitud de la distribución Gamma y una distribución exponencial, esto es:

$$\begin{aligned} f(\alpha, \beta|x) &\propto \left(\prod_{i=1}^{n} x_i^{\alpha-1} e^{-x_i\beta} \mathbb{1}_{(1,4)}\right)\left(e^{-\beta}\mathbb{1}_{(0,\infty)}\right) \\ &\propto \left(\frac{\beta}{\Gamma(\alpha)}\right)^n r_2^{\alpha-1} e^{-\beta(r_1+1)} \mathbb{1}_{(1,4)}\mathbb{1}_{(0,\infty)} \end{aligned} \tag{3}$$

donde $r_1 = \prod_{i=1}^{n} x_i$ y $r_2 = \sum_{i=1}^{n} x_i$.

Considerando como propuesta una distribución normal bivariada $N((\alpha \quad \beta)^T, \delta_i\sigma_j^2) = q(\alpha, \beta|\alpha_p, \beta_p)$. En consecuencia se tiene que el criterio de aceptación $\rho = min\left(1, (\frac{f(\alpha_p,\beta_p|x)}{f(\alpha,\beta|x)})(\frac{q(\alpha_p,\beta_p|\alpha,\beta)}{q(\alpha,\beta|\alpha_p,\beta_p)})\right) = min\left(1, (\frac{f(\alpha_p,\beta_p|x)}{f(\alpha,\beta|x)})\right)$, esto se debe a que la propuesta es una distribución simétrica y en consecuencia $\left(\frac{q(\alpha_p,\beta_p|\alpha,\beta)}{q(\alpha,\beta|\alpha_p,\beta_p)}\right) = 1$.

Al final el criterio de aceptación resulta en los siguiente:

$$\rho = min\left\{\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha_p)}\right)^n\left(\frac{\beta_p^{\alpha_p}}{\beta^{\alpha}}\right)\left(\frac{r_2^{\alpha_p-1}}{r_2^{\alpha-1}}\right)\left(e^{-\beta_p(r_2+1)+\beta(r_2+1)}\mathbb{1}_{(1,4)}\mathbb{1}_{(0,\infty)}\right)\right\} \quad (4)$$

Es importante aclarar que en las notas el soporte de está definido en $\mathbb{1}_{(1,4)}\mathbb{1}_{(1,\infty)}$, pero en realidad es $\mathbb{1}_{(1,4)}\mathbb{1}_{(0,\infty)}$.

Como distribuciones iniciales se consideraron $alpha \sim U(1,4)$ y $\beta \sim U(0,4)$.

En las figuras 1 y 2 (n=3 y n=30 respectivamente) se consideraron 1000 iteraciones con un burn-in de 100 iteraciones, con un parámetro de $\sigma_1 =_2=$ 0,1. En la parte izquierda de cada figura se muestra la caminata sin el burn-in y en la parte derecha se muestra la caminata del burn-in. En el primer caso ($n = 3$) el sesgo es menor y por lo tanto se observa un sub-ajuste para el cálculo de la verosimilitud, caso contrario con $n = 30$ (figura 2) en resultado el soporte de la función es mayor con $n = 3$ que con $n = 30$.

En la figura 3 se muestra todo el recorrido realizado dada la configuración $n = 30$ $\sigma_1 = \sigma = 2 = 0{,}01$ y un máximo de iteraciones de $10,000$, es decir se realizan desplazamientos menores, en resultado la caminata requiere un mayor número de iteraciones para converger a la distribución objetivo. Entonce se puede decir que la matriz de

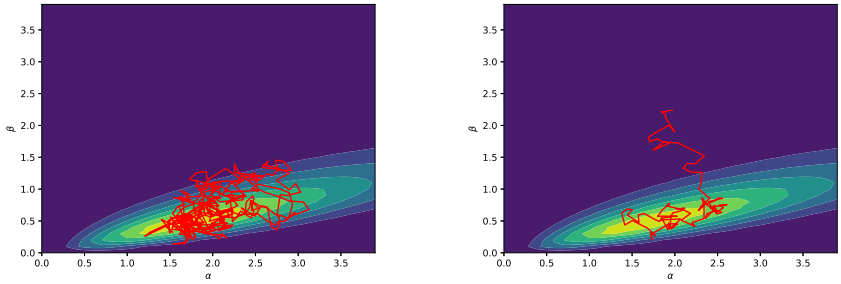

Figura 1: Figura de contorno de la distribución posteriori considerando $n = 3$, 1000 iteraciones (100 de burn-in) y una propuesta normal con matriz de covarianza diagonal con entradas 0,1. En la parte izquierda se muestra la caminata sin el burn-in y en la derecha s emuestra el recorrido en el burn-in.

.

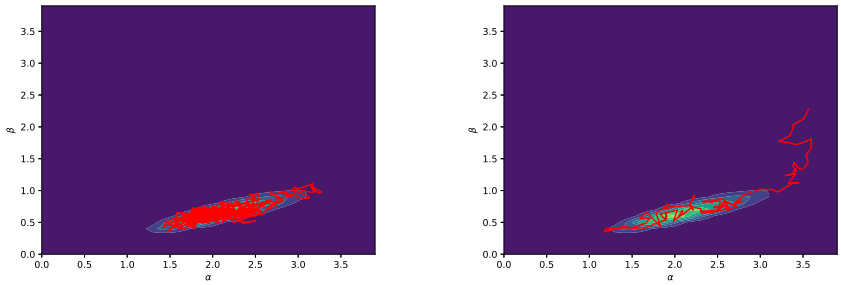

Figura 2: Figura de contorno de la distribución posteriori considerando $n = 30$, 1000 iteraciones (100 de burn-in) y una propuesta normal con matriz de covarianza diagonal con entradas 0,1. En la parte izquierda se muestra la caminata sin el burn-in y en la derecha s emuestra el recorrido en el burn-in.

.

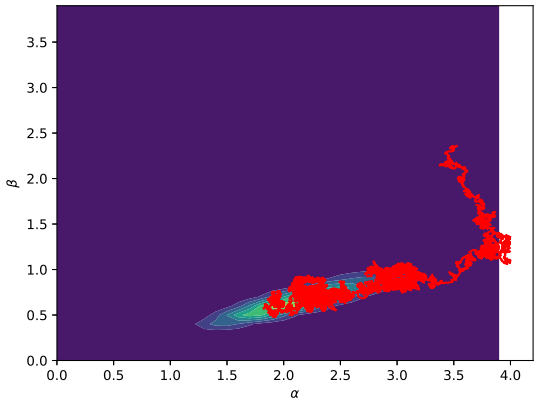

Figura 3: Figura de contorno de la distribución posteriori considerando $n = 30$, 10000 iteraciones (0 de burn-in) y una propuesta normal con matriz de covarianza diagonal con entradas 0,01.

.

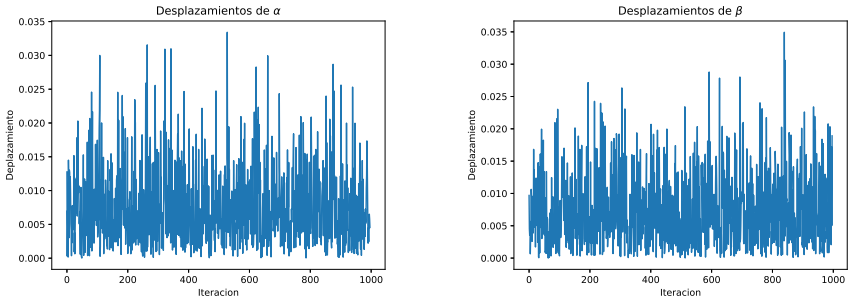

Figura 4: Desplazamiento de α y β considerando considerando $n = 30$, 1000 iteraciones y una propuesta normal con matriz de covarianza diagonal con entradas 0,01.

2. Punto 2

Simular de la distribución $Gamma(\alpha, 1)$ con la propuesta $Gamma([\alpha], 1)$, donde $[\alpha]$ denota la parte entre de α. Además,realizar el siguiente experimento: poner como punto inicial $x_0 = 1000$ y graficar la evolución de la cadena, es decir, $f(x_t)$ vs. t.

Comentarios

En este caso se considera que la propuesta es indemendiente, es decir $q(y|x) = q(y)$, por lo tanto el criterio de aceptación es el siguiente:

$$\rho(x, y) = min\left(1, \left(\frac{f(y)}{f(x)}\right)\left(\frac{q(x)}{q(y)}\right)\right) \tag{5}$$

Principalmente porque si $q(x) \approx f(x)$, entonces $\rho(x, y) \approx 1$, esto quiere decir que la propuesta $q(.)$ es buena si se parece a $f(.)$.

La distribución inicial que se consideró es $x \sim U(0{,}0, 5{,}0)$. Los parámetros que se consideraron fueron los siguientes. $\alpha = 1{,}9$, $\beta = 10$, iteraciones $= 1000$. En la figura 5 se muestran los resultados obtenidos, en la parte superior izquierda se muestra una estimación de la función objetivo dada la simulación y su recorrido. En la parte superior derecha se pone la evaluación de la distribución objetivo de cada estado visitado en el recorrido. Principalmente, las distribución simulada y la distribución objetivo tienen formas visualmente similares. Aedmás, en la parte inferir izquierda se presentan los estados de cada iteracion que corresponden a la cadena, por lo tanto se obsera convergencia a la distribución objetivo, a pesar de que la distribución propuesta tenga una forma distinta que se puede observar en la figura 6.

Como se indica en el ejercicio, se probaron dos puntos iniciales alejados de la distribución inicial, $x_0 = 1000$ con $\alpha = 200{,}5$ y $\beta = 1$ y $x_0 = 100$ con $\alpha = 20{,}9$ y $\beta = 1$, el comportamiento de la cadena depende fuertemente de los parametros utilizados en la distribución. En resultado puede suceder que si el punto inicial está extremadamente alejado de la media de la función de distribución, entonces sus valores pueden ser extremadamente bajos causando problemas numéricos, en consecuencia se tendrían problemas de convergencia. Se probaron varios parametros de la distribución para que existiera convergencia dado el punto inicial, en la figura 8 se muestra la evolución de la cadena de cada configuración y sus respectivos PDF's. Principalmente se puede obsevar que al aumentar el parámetro α la distribución objetivo y la distribución propuestas tienen a ser muy similares, esto se puede observa en la evolución de la cadena. En los dos casos la convergencia fue demasiado rápido.

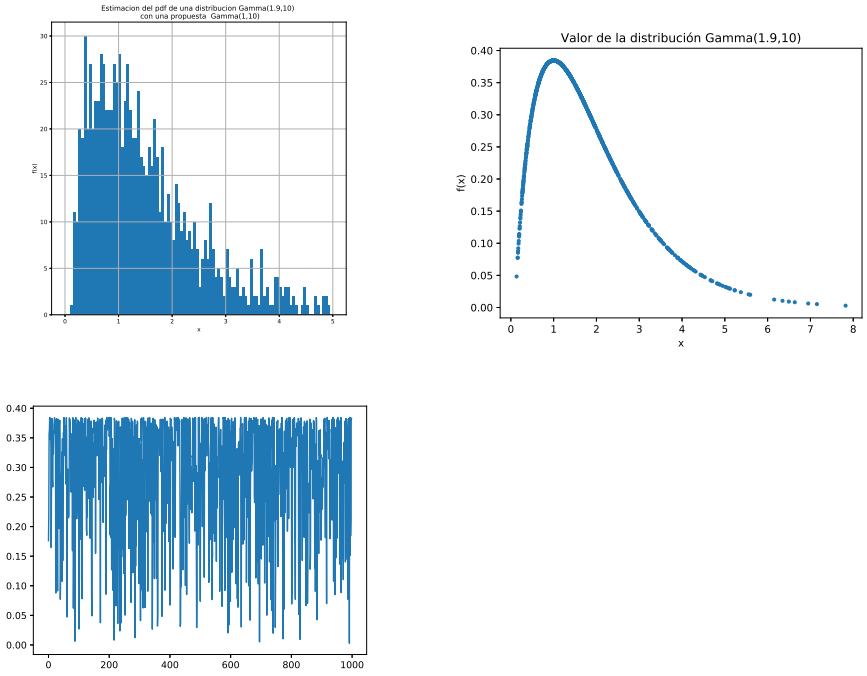

Figura 5: En la parte superior izquierda está la distribución simulada, en la parte superior derecha la distribución objetivo y en la parte inferior izquierda la desplazamientos de la cadena.

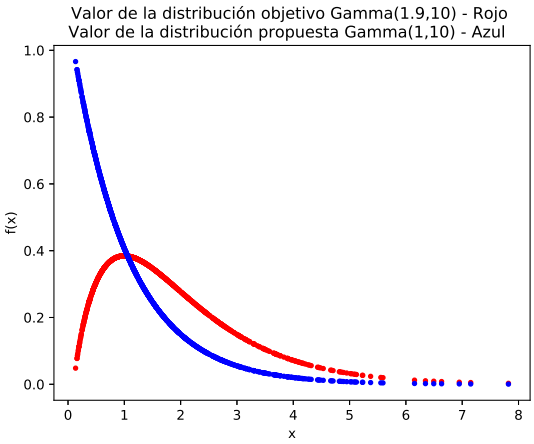

Figura 6: Evaluación de la distribución objetivo y propuesta dados los estados visitados.

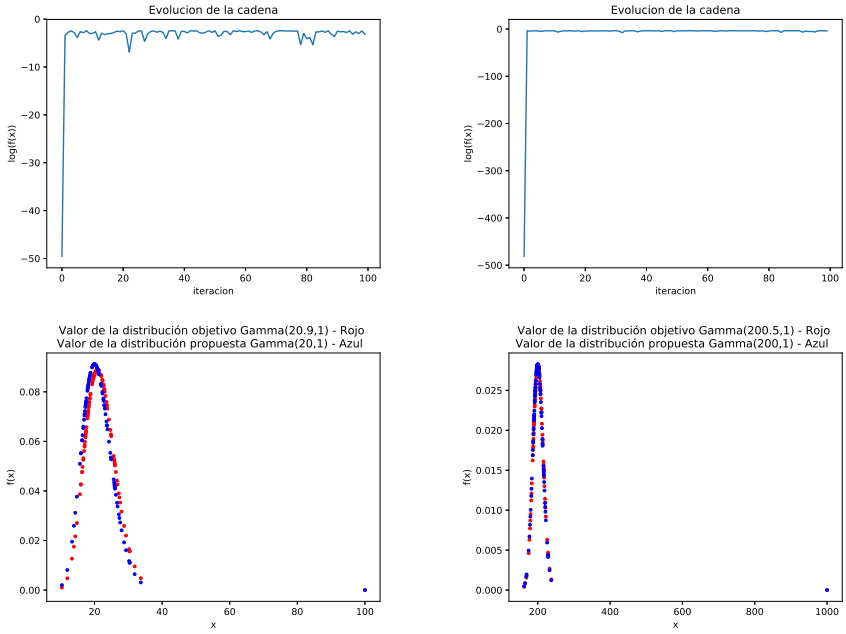

Figura 7: En la parte superior izquierda se muestran los resultados con la configuración $x_0 = 1000$ con $\alpha = 200{,}5$ y $\beta = 1$ y en la parte derecha con y $x_0 = 100$ con $\alpha = 20{,}9$ y $\beta = 1$. La parte superior indica la evolución de la cadena y la parte inferior indica la forma de la distribución objetivo (rojo) y propuesta (azul)

3. Punto 3

Implementar Random Walk Metropolis Hasting (RWMH) donde la distribución objetivo es $N_2(\mu., \Sigma)$ con $\mu = (3, 5)^T$

$$\Sigma = \begin{pmatrix} 1 & 0{,}9 \\ 0{,}9 & 1 \end{pmatrix}$$

Utilizar una propuesta $\epsilon_t \sim N_2(0, \sigma I)$

- ¿Cómo elegir σ para que la cadena sea eficiente?
- ¿Qué consecuencias tiene la elección de *sigma*?
- Como experimento, elige como punto inicial $X_0 = (1000, 1)^T$ y comenta los resultados.

Comentarios

Para evitar problemas numéricos se utiliza el logaritmo de la función objetivo y de la propuesta. En particular el método utilizado de scipt es "multivariate.log.pdfÇomo parte de este análisis se consideran dos escenarios, en el primero se toma una disitribución inicial por medio de una distribución uniforme $x_i \sim U(0{,}0, 5{,}0)$ (figura 8). En el segundo escenario se fija le punto inicial en $x_0 = (1000, 1)^t$ (figura 9). Además en cada escenario se prueban las configuraciones con $\sigma = 5{,}0$ y $\sigma = 0{,}1$ En el primer escenario (distribución uniforme inicial –ver figura 8–), se considera un máximo de iteraciones distinto, en el caso de $\sigma = 5{,}0$ se consideran 1000 iteraciones y en $\sigma = 0{,}5$ se consideran 10000 iteraciones. Además se toma la eficiencia de cada ejecución de las cuatro como $Eficiencia = \frac{Aceptadas}{Aceptadas + Rechazadas}$ En la tabla 1 se presenta la eficiencia obtenida en cada configuración (sin burn-in). En general (considerando todos los experimentos) se puede establecer los siguiente:

- En este caso existe una mejor eficiencia al considerar un $\sigma = 0{,}1$, esto tiene sentido, ya que después de que la distribución instrumental converge a la distribución objetivo, la cantidad de rechazos es significativamente menor.
- Tener un $\sigma = 0{,}1$ tiene implicaciones en la convergencia –dependiendo del punto inicial–.
- Valores elevados $\sigma = 5{,}0$ ayuda a una convergencia rápida si el punto inicial está alejado de la distribución objetivo, no obstante una vez que converge pueden exisitr un número de rechazos mayores.

3.1. Propuesta !!!

Dados los resultados anteriores, la propuesta consiste en un kernel híbrido con pesos adaptativos, es decir $\epsilon_t = w_1 N_2(0, (0{,}1 * I)) + w_2 N_2(0, (5{,}0 * I))$, donde el peso de cada distribución es ajustado en base a su éxito en una ventana de tiempo (batch). Una alternativa que suena muy interesante, y no se utiliza en esta tarea, es de utilizar descenso de gradiente para optimizar los pesos y el sigma (varianza), donde lo que se desea es minimizar la divergencia de Kullback entre la mezcla de distribuciones y la distribución objetivo.

Aunque el procedimiento de adaptar los pesos parece ser contradictorio ya que de cierta forma la cadena depende de los movimientos anteriores en el trabajo de Gareth and Rosenthal [1] se aclara que esto en la práctica

tiene propiedades de convergencia ya que fácilmente se puede demostrar la condición de *Diminishing Adaptation*, y en ciertas situaciones (que se cumple en este caso) la condición de *Boundary Convergence*.

	$x_0 \sim U(0,5,0)$	$x_0 = (1000,1)^t$
$\sigma = 0{,}1$	0.68	0.58
$\sigma = 5{,}0$	0.11	0.26

Cuadro 1: Eficencia obtenida en cada configuración.

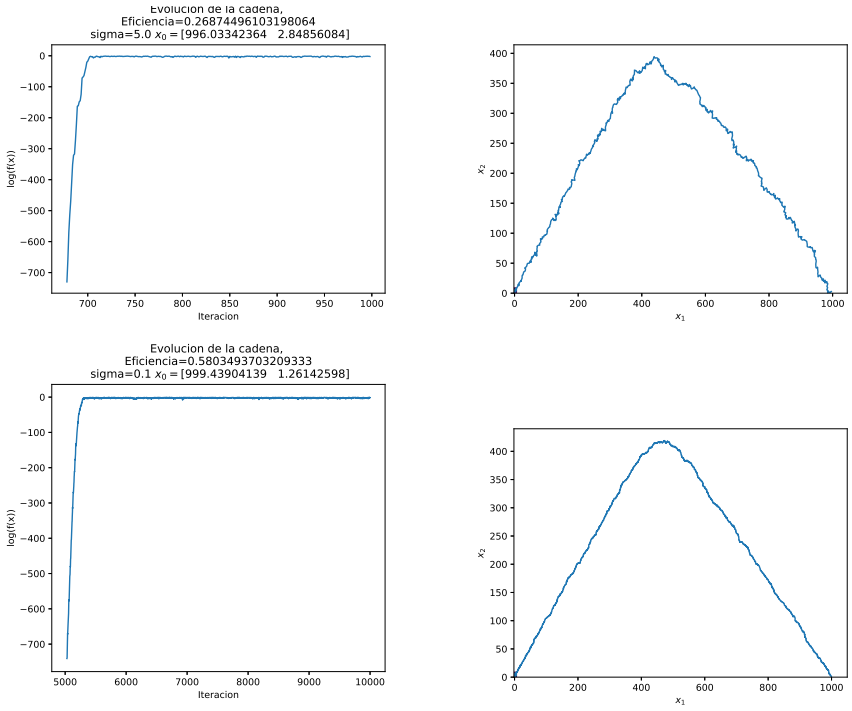

Figura 8: Evolución de la cadena (izquierda) y recorrido (derecha) considerando un punto inicial como $x_0 = (1000,1)^t$ (en el título se reporta el segundo punto inicial).

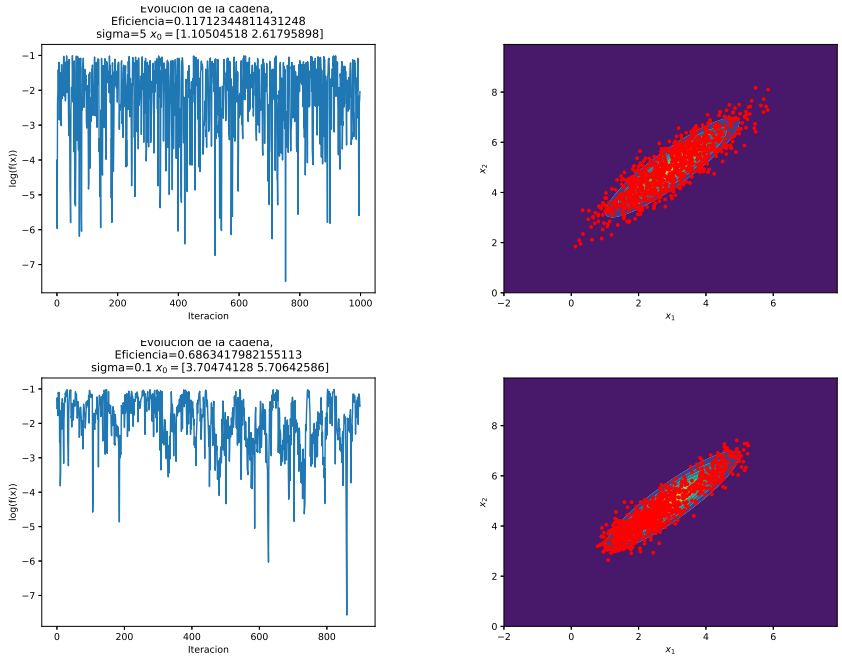

Figura 9: Evolución de la cadena (izquierda) y recorrido (derecha) considerando un punto inicial $x_0 \sim U(0,5,0)$.

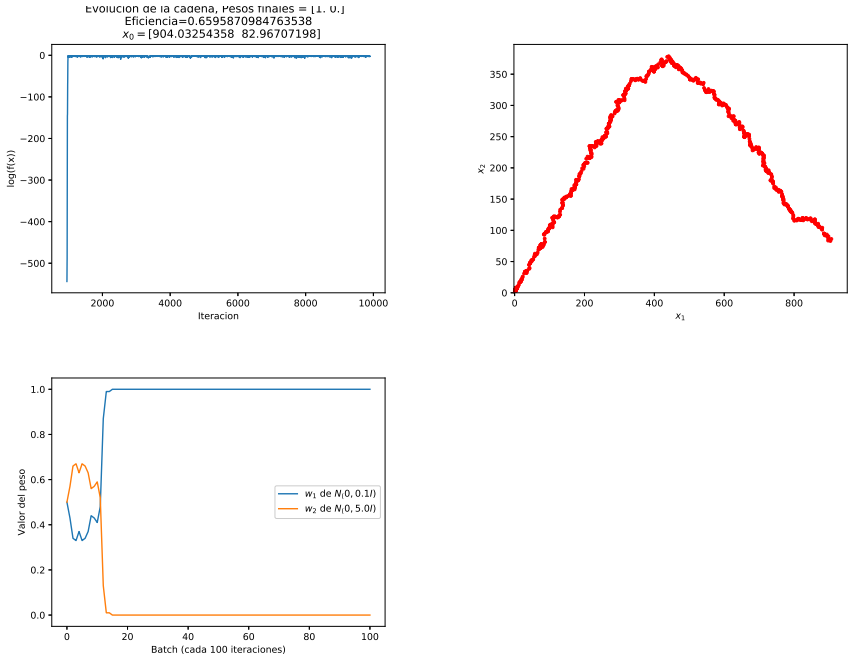

Figura 10: Evolución de la cadena (izquierda) y recorrido (derecha) considerando un punto inicial como $x_0 = (1000, 1)^t$ considerando como propuesta una mezcla de distribuciones y pesos aadptativos. En la parte inverior se reporta la evolución de los pesos.

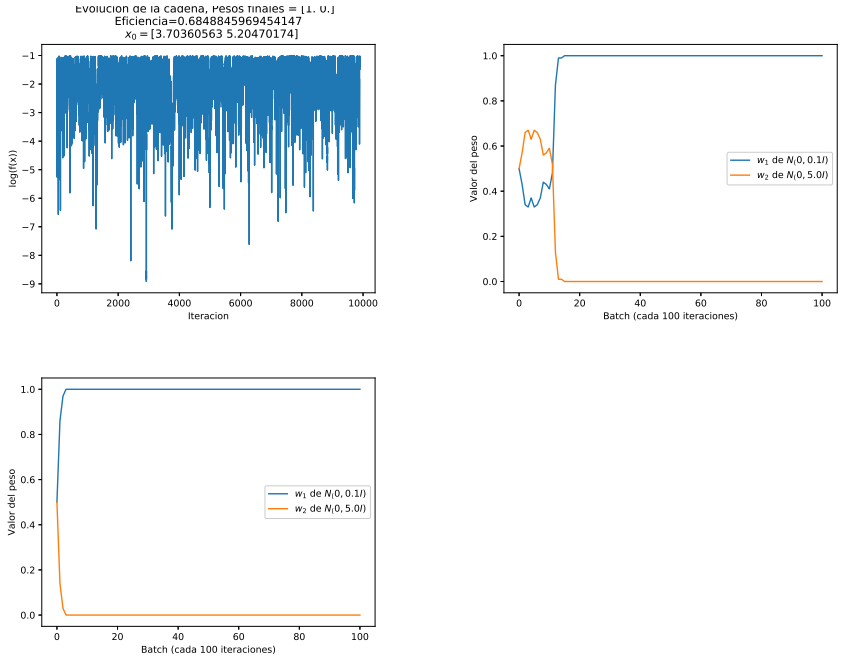

Figura 11: Evolución de la cadena (izquierda) y recorrido (derecha) considerando un punto inicial $x_0 \sim U(0,5,0)$ considerando la propuesta como una mezcla de distribuciones con pesos adaptativos. En la parte inferior se reporta la evolución de los pesos

Referencias

[1] G. O. Roberts, J. S. Rosenthal, Examples of adaptive mcmc, Journal of Computational and Graphical Statistics 18 (2009) 349–367.