Inferencia Bayesiana con métodos MonteCarlo: Cadenas de Markov

A.R. Callen[1]

[1] *Observatorio Astronómico de Córdoba, UNC, Argentina*

Contacto / ailen.callen@mi.unc.edu.ar

Resumen / En este trabajo se implementan distintos conceptos y técnicas estudiadas, relacionados con la inferencia estadística, ajuste de funciones, selección de modelos, cuadrados mínimos, interpolación y minimización.

Abstract / In this work, various concepts and techniques studied in the context of statistical inference, function fitting, model selection, least squares, interpolation, and minimization are implemented.

Keywords / Sun: MonteCarlo — inferencia Bayesiana — Gradiente descendente — Metrópolis-Hastings

1. Introducción

La inferencia estadística se puede llevar a cabo como una aplicación del teorema de Bayes. Si tenemos un conjunto de datos d que se puede describir por un modelo m con parámetros ϕ, queremos calcular el mejor modelo que puede dar lugar a esos datos, es decir, maximizar la probabilidad posterior de los parámetros dados los datos para un modelo m, $p(\phi|d,\ m)$. Esta probabilidad es proporcional al *Likelihood* $p(d|\ \phi,\ m)$ por la función distribución de la probabilidad anterior *(prior,* $p(\phi,\ m)$*)*.

$$p(\phi|d,m) = \frac{p(d|\phi,m)p(\phi|m)}{p(d|m)} \tag{1}$$

y está normalizada por la evidencia, es decir, la probabilidad marginal del *Likelihood* para el modelo m:

$$p(d|m) = \int_{\Omega} p(d|\phi,m)p(\phi|m) \cdot d\phi \tag{2}$$

donde Ω denota el espacio de parámetros. Cuando se ajusta un modelo a un conjunto de datos, se quiere conocer la función de *Likelihood,* $p(d|\ \phi,\ m)$, que depende de los parámetros ϕ. Existen varios métodos para llevar esto a cabo, entre ellos las Cadenas de Markov Monte Carlo (MCMC).

En este trabajo se usa el algoritmo de Metrópolis-Hastings, que se utiliza para simular distribuciones multivariadas, y el gradiente descendente para obtener los parámetros de la función de luminosidad de galaxias y compararlos con los valores obtenidos en Blanton et al (2001). Para ello se utilizará como modelo la función de Schechter dada por la ecuación 3.

$$\Phi(M)\,dM = 0.4 \ln 10 \Phi^* 10^{-0.4(M-M*)(a+1)} e^{10^{-0.4(M-M*)}} dM \tag{3}$$

Donde

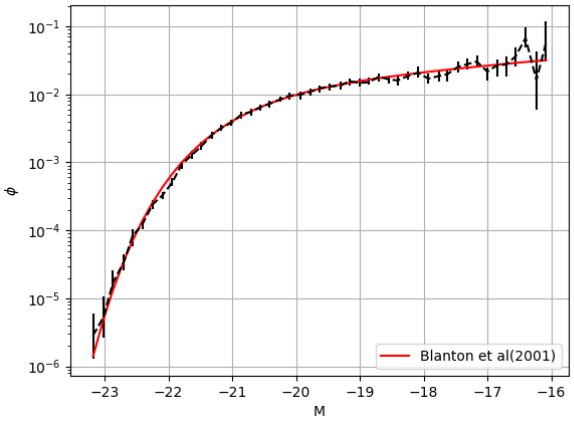

Figura 1: Ajuste de la función de luminosidad de galaxias realizado por Blanton et al (2001).

a = (-1.20 ± 0.03)
ϕ = (0.0146 ± 0.0012)
M = (-20.83 ± 0.03)

2. Procedimiento

Lo primero que se realizó, como se muestra en la Figura 1, fue graficar los datos pertenecientes a la función de luminosidad de galaxias obtenida por Blanton et al (2001).Luego, se procedió a cambiar los valores de los parámetros obtenidos por los autores para poder elegir lo limites del *prior*. La función de *priors* determina el espacio de parámetros en donde realizaremos la búsqueda de estos, por lo tanto, se seleccionó una región acotada con el fin de disminuir el tiempo de calculo del algoritmo. En este caso, se usó un *prior* plano,por lo que esta función devuelve un valor constante si se encuentra dentro de la región que se va a muestrear, mientras que es nula fuera de la región. De la Figura 2 a la Figura 4 se

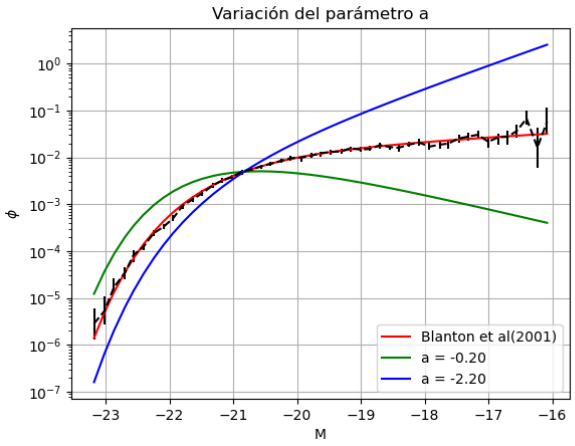

Figura 2: Variación del parámetro a entre a=-2.20 y a=-0.20 dejando M y ϕ fijos.

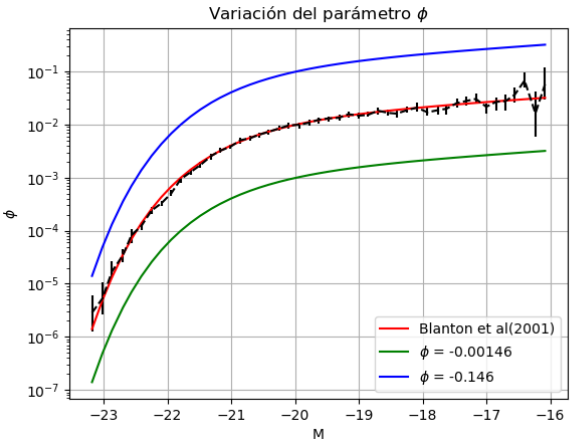

Figura 4: Variación del parámetro ϕ entre ϕ=0.00146 y ϕ=0.146 dejando M y a fijos.

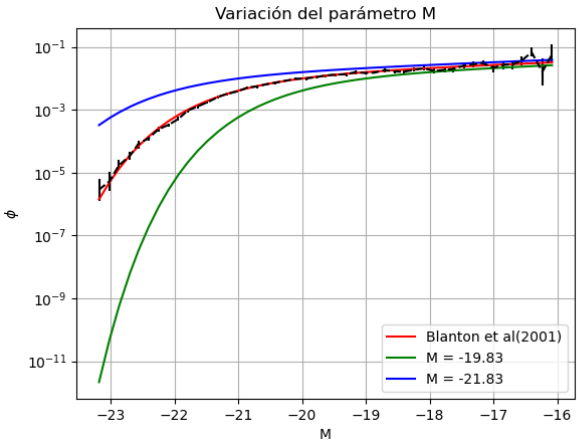

Figura 3: Variación del parámetro M entre M=-19.83 y M=-21.83 dejando a y ϕ fijos.

pueden ver las variaciones de cada uno de los parámetros marcando los límites del espacio de los parámetros.

Una vez determinado el espacio de los parámetros se calcula el *Likelihood* y la probabilidad posterior de los parámetros dada por la ecuación 1. Además, como en este caso se usó solo un modelo para ajustar los datos se toma $p(d \mid m) = 1$.

Teniendo esto en cuenta y que además se eligió un *prior* plano se ve que en el caso en que los parámetros caigan dentro de los limites dados por este, la probabilidad posterior de los parámetros va a ser igual a el *Likelihood* , mientras que será cero cuando los parámetros caigan fuera de los límites seleccionados.

Finalmente, con todas las funciones determinadas se procede a realizar primero el algoritmo de Metrópolis-Hastings y luego el del gradiente descendente.

2.1. Metrópolis-Hastings

Este método consiste en realizar cadenas, diez en este trabajo, que se pueden pensar como saltos (en 3 dimensiones), cuyo nuevo paso está dado por la función posterior evaluada en el paso anterior. Se calcula la probabilidad de aceptación como el mínimo entre 1 y la razón (función posterior en el paso nuevo)/(función posterior en el paso anterior). Se genera un número aleatorio u entre 0 y 1, y si u es menor a la probabilidad de aceptación se acepta el paso.

Para elegir los parámetros iniciales de la cadena se sortean inicialmente en el rango que se definió el *prior*, y se le pide que elija aquellos valores que no anulen la función posterior.

Los factores que intervinieron principalmente en las cadenas son: el ancho de los intervalos de los parámetros (*priors*) y el número de pasos. Estos valores cambian mucho la forma de las cadenas por lo tanto es necesario llegar a un balance entre el tamaño del salto y el ancho del intervalo, recordando que no debe ser demasiado grande ya que a grandes distancias la función posterior es nula (por el *likelihood y prior)* y además podría saltar el mínimo sin caer nunca sobre el, ni tampoco demasiado chico, ya que se necesitaría más tiempo de cómputo o podría nunca llega al valor esperado.

Para comprobar si las cadenas muestrean correctamente y para comprobar su convergencia, se realiza un gráfico del parametro a vs ϕ, el cual se muestra en la Figura 5. Además, se ve la tendencia de cada parámetro con el número de pasos en cada cadena en las Figuras 6,7 y 8.

Analizando las imágenes se puede ver que a partir de los N=4000 pasos los parámetros se estabilizan, indicando la convergencia de la cadena. En esta cantidad de paso se realiza el "quemado"de la cadena, es decir, solo se tienen en cuenta los valores de los parámetros a partir de N=4000 para luego calcular el promedio y sacar los valores finales que van a ser comparados.

Con este método los valores obtenidos son:

a = -1.173

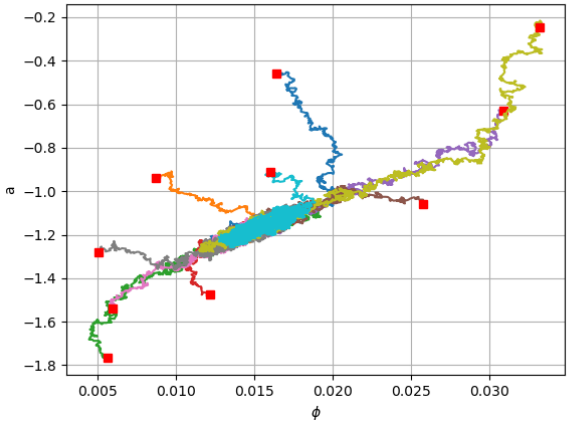

Figura 5: Convergencia de diez cadenas de Markov en el espacio de parámetros a vs ϕ. Los cuadrados rojos simbolizan el comienzo de cada cadena.

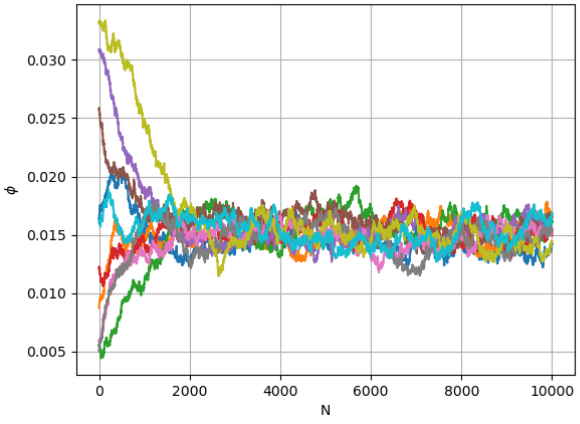

Figura 6: Convergencia del parámetro ϕ en función de la cantidad de pasos.

$\phi = 0.0150$
M = -20.772

Que caen dentro del intervalo de error de los valores obtenidos por Blanton et al(2001) por lo que son valores comparables.

En la Figura 9 se puede ver que el ajuste en este trabajo pasa dentro de las barras de error de todos los puntos y coincide en la mayoría de los puntos con el ajuste obtenido por Blanton et al (2001), excepto en la parte de luminosidades altas.

2.2. Gradiente descendente

Cuando se quiere realizar un ajuste de un modelo que depende no-linealmente de un conjunto de p $a_k, k = 1...p$ parámetros el procedimiento se basa en buscar los parámetros que minimizan la función χ^2.

La idea general es dar valores iniciales y desarrollar un procedimiento iterativo para mejorarlos. Uno de los procedimientos para mejorar un conjunto inicial de parámetros es seguir la dirección de máximo crecimiento (o decrecimiento) indicada por el gradiente de χ^2.

El método que usa el gradiente como indicador para

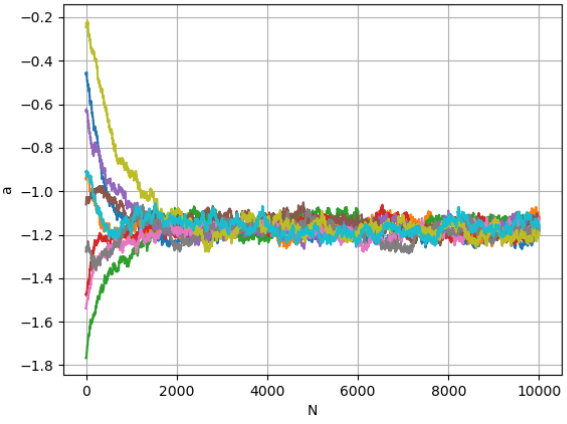

Figura 7: Convergencia del parámetro a en función de la cantidad de pasos.

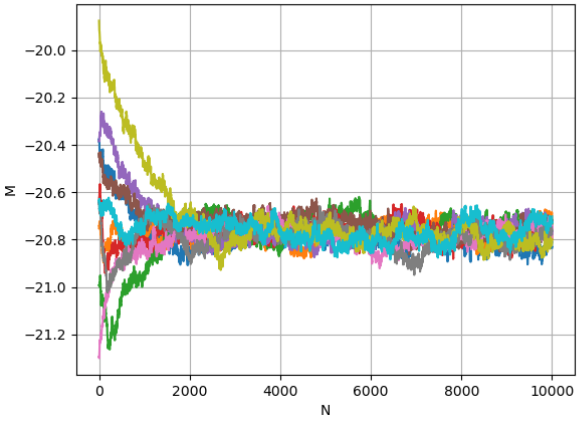

Figura 8: Convergencia del parámetro M en función de la cantidad de pasos.

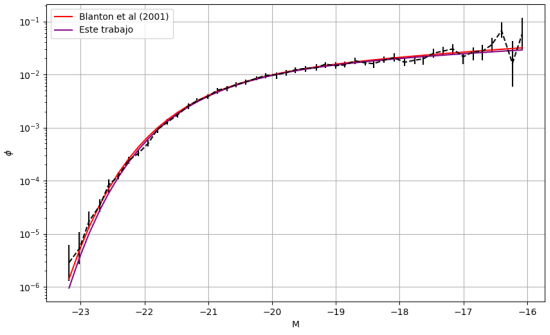

Figura 9: Linea a trazos: datos, Linea violeta: ajuste realizado en este trabajo, Linea roja: ajuste obtenido por Blanton et al(2001).

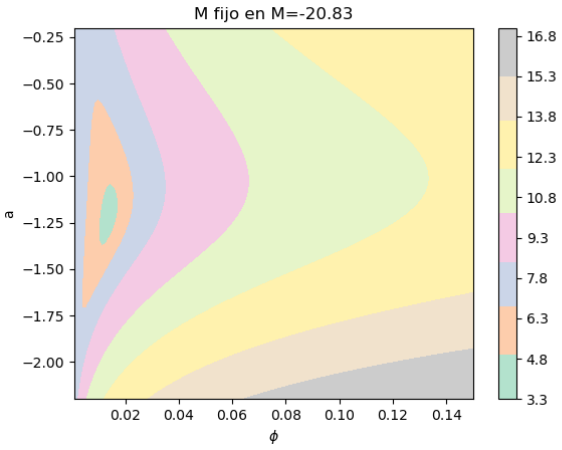

Figura 10: Comportamiento del χ^2 mostrando que solo tiene un mínimo, dejando el parámetro M fijo.

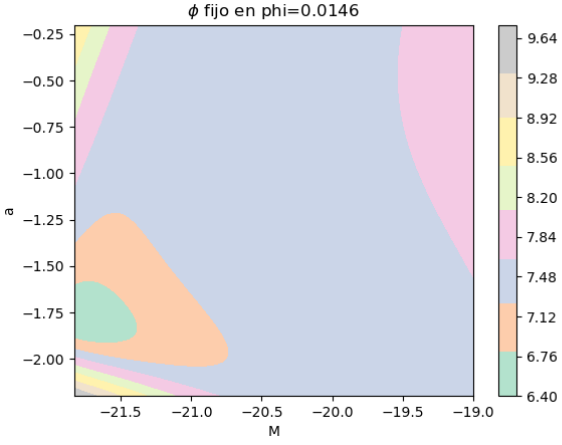

Figura 11: Comportamiento del χ^2 mostrando que solo tiene un mínimo, dejando el parámetro ϕ fijo.

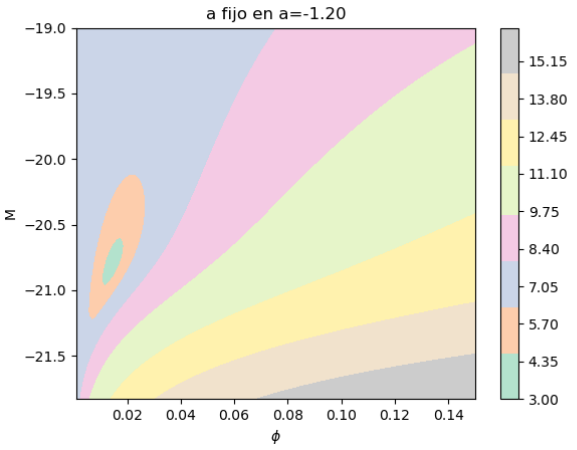

Figura 12: Comportamiento del χ^2 mostrando que solo tiene un mínimo,dejando el parámetro a fijo.

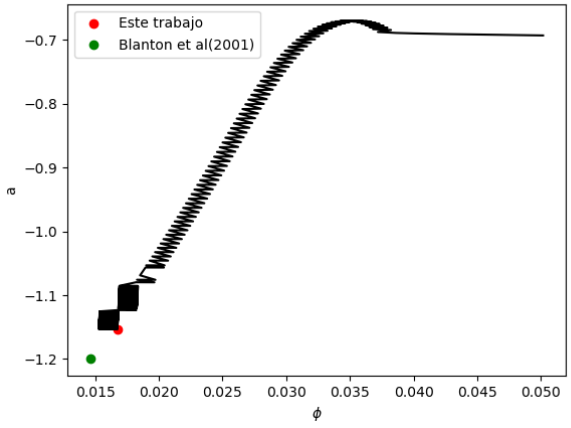

Figura 13: Parámetro a vs parámetro ϕ para una caminata de N=500 pasos. Punto rojo= valores finales obtenidos en este trabajo. Punto verde=Blanton et al(2001)

optimizar los parámetros es el de "Gradiente Descendiente", en el cual:

$a_(new) = a_(old) - \eta \nabla \chi^2$

donde a son los parámetros y η es el tamaño del paso (fijo).

Una vez hecho esto se analiza como se comporta χ^2. Para ello, se determina un rango para los parámetros y se observa una *tajada* en cada plano posible entre los parámetros dejando el tercer parámetro fijo. Además se indica con un código de colores el valor de la función (Figura 10,11 y 12) para ver gráficamente donde se encuentra el mínimo de la misma y corroborar que sea único, ya que de otro modo el método no funcionaria.

Una vez corroborado la unicidad del mínimo se realiza una caminata siguiendo la dirección del gradiente descendente. Para ello primero se lo normaliza y además se define el parámetro η como una constante multiplicada por el ancho de cada intervalo de cada parámetro definido antes para el *prior*. Esto se hace para poder balancear los diferentes ordenes de magnitud de cada parámetro. Para la caminata se tomo un total de N=500 pasos. Los resultados se muestran en la Figura 13,14 y 15. Además, se puede ver una comparación entre los valores finales obtenidos con este método y los valores obtenidos por Blanton et al(2001).

Con este método los valores obtenidos son:

a = -1.159

ϕ = 0.0166

M = -20.752

Finalmente, se relizó una superposicion de los graficos del comportamiento del χ^2 con las caminatas para cada par de parametros(Figuras 16,17 y 18). Con esto se pudo corroborar que, al menos en dos de los casos, la caminata converge correctamente al minimo del χ^2. Las diferencias observadas en la Figura 18 se debe principalmente a que al dejar un valor fijo para ver el comportamiento del χ^2 no estamos viendo una representacion confiable del mismo.

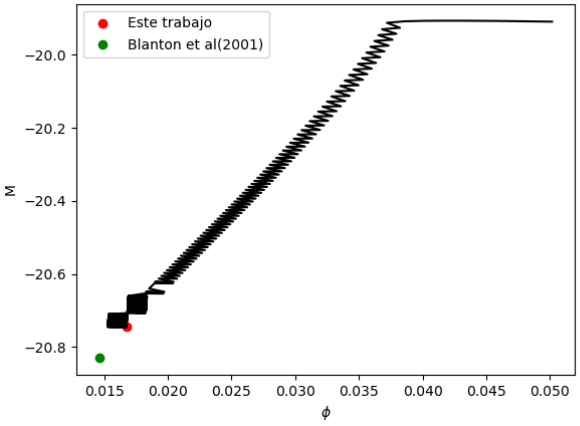

Figura 14: Parámetro M vs parámetro ϕ para una caminata de N=500 pasos. Punto rojo= valores finales obtenidos en este trabajo. Punto verde=Blanton et al(2001)

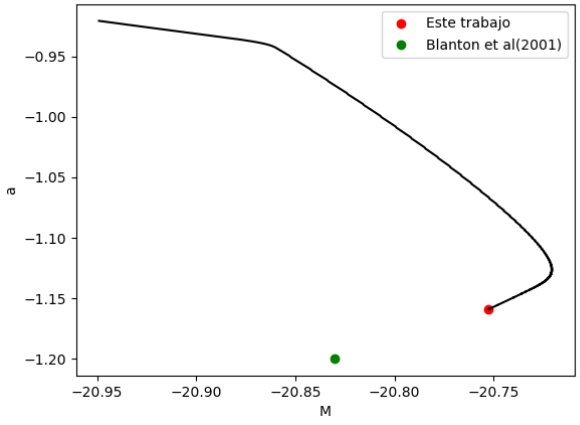

Figura 15: Parámetro a vs parámetro M para una caminata de N=500 pasos. Punto rojo= valores finales obtenidos en este trabajo. Punto verde=Blanton et al(2001)

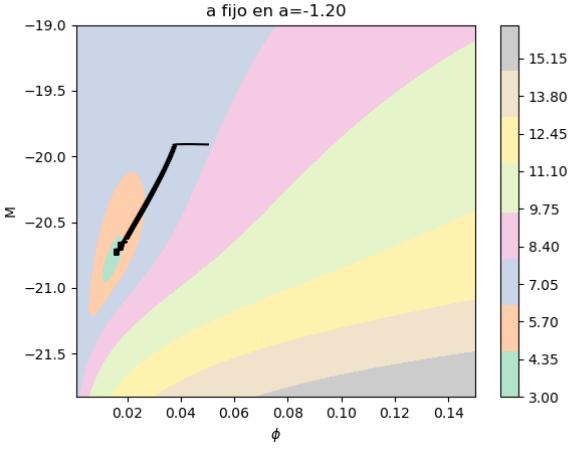

Figura 16: Caminata por el método del gradiente descendente sobre el comportamiento de la función χ

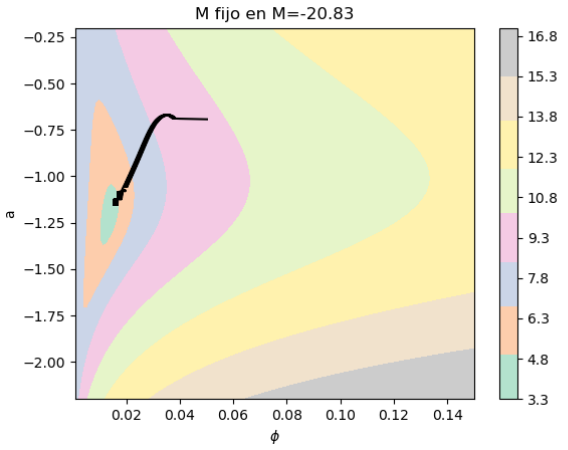

Figura 17: Caminata por el método del gradiente descendente sobre el comportamiento de la función χ

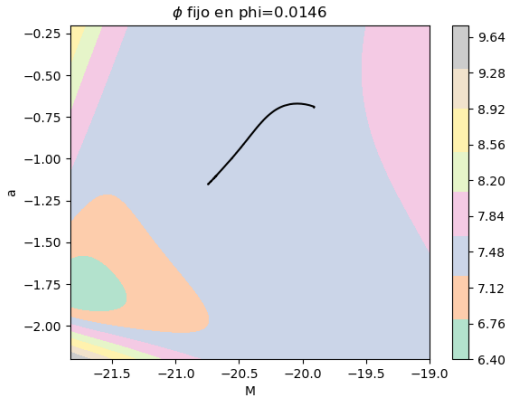

Figura 18: Caminata por el método del gradiente descendente sobre el comportamiento de la función χ

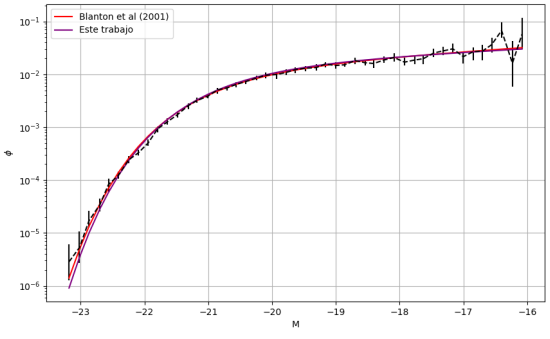

Figura 19: Linea a trazos: datos, Linea violeta: ajuste realizado en este trabajo, Linea roja: ajuste obtenido por Blanton et al(2001).

Por ultimo, se corrobora si los valores obtenidos ajus-

tan los datos medidos y se lo compara con el ajuste obtenido por Blanton et al(2001)(Figura 19). En este ajuste se puede notar que al igual que con el metodo de Metropolis-Hastings los ajustes difieren en las zonas de altas luminosidades, pero este metodo nos brinda un ajuste igual al de Blanton en luminosidades bajas, cosa que no ocurria usando Metropolis-Hastings.

3. Conclusiones

En este trabajo se logró ajustar la función de Schechter a la función de luminosidad de galaxias obtenida por Blanton et al(2001) por medio del método del Gradiente Descendente y Metrópolis-Hastings.

Los métodos Bayesianos, frecuentemente implementados usando las Cadenas de Markov Monte Carlo (MCMC), proveen una manera poderosa de estimar los parámetros de un modelo realizando un muestreo aleatorio del espacio de parámetros.

Los valores encontrados tanto por Metrópolis-Hastings como por Gradiante descendente son comparables con los valores encontrados por Blaton et. al(2001). La diferencia entre algortimos es que el Metrópolis-Hastings nos da mas información acerca del entorno en donde se encuentran los mínimos de la función y además esta tiene la posibilidad de que si se encuentra con un mínimo secundario de la función poder salir de este, lo cual la de Gradiente descendente por construcción no puede hacer.

Referencias

Blanton et al. (2001), AJ, 121, 2358