Informe 4: Inferencia Bayesiana con métodos MonteCarlo

B.M. Celiz[1]

[1] *Facultad de Matemática, Astronomía y Física, UNC, Argentina*

Contacto / bruno.celiz@mi.unc.edu.ar

Resumen / En este informe realizamos. vía inferencia Bayesiana, un ajuste de los parámetros de la función de Schechter para los datos de la función de luminosidad de galaxias obtenidos en la banda r del Sloan Digital Sky Survey. Luego, utilizando el método de Metrópolis-Hasting para Markov Chains MonteCarlo en Python, exploramos el espacio de parámetros definido y buscamos maximizar la función *Likelihood* mediante la técnica de Gradiente Descendiente, encontrando así la mejor combinación de parámetros para nuestros datos.

Abstract / In this essay we will do, through Bayesian inference, a fit of the parameters from the Schechter's function for the luminosity function of galaxies obtained in the r band from the Sloan Digital Sky Survey. Then, using Metrópolis-Hasting method for Markov Chains MonteCarlo in Python, we will explore the parameter space and search to maximize the *Likelihood* via Gradient Descent, therefore finding the best combination of parameters for our data.

Keywords / galaxies: luminosity function, mass function — methods: data analysis — methods: numerical — methods: statistical — chaos

1. Introducción

La inferencia Bayesiana es un tipo de inferencia estadística que se basa en el Teorema de Bayes. Si tenemos un conjunto de datos d que queremos describir con un modelo m que depende de los parámetros θ_i, pedir la mejor elección de parámetros posibles es lo mismo que maximizar la probabilidad de los parámetros dado los datos y el modelo, según el Teorema de Bayes (Ec. (1)).

$$P(\theta_i|d,m) = \frac{P(d|m,\theta_i)P(\theta_i|m)}{P(d|m)} \tag{1}$$

Cada factor de Ec. (1) recibe un nombre particular, y es que $P(\theta_i|d,m)$, o como es llamada, *Probabilidad Posterior* (PP), está definida por:

- $P(d|m)$ es la *Evidencia*. En el caso de usarse un modelo fijo, como es nuestro caso, este término es una constante.
- $P(\theta_i|m)$ es el *Prior*, que viene de *Probabilidad Anterior*. Consideramos esta probabilidad como constante en un intervalo definido y nula afuera del mismo, dado a que no conocemos el valor óptimo de cada parámetro pero si podemos acotar sus rangos posibles. Esto es, decimos que los parámetros siguen una distribución uniforme, y por ende el *Prior* es constante.
- $P(d|m,\theta_i)$ es el *Likelihood* ($\mathcal{L}$), el término al que más importancia le daremos en este trabajo. Dado a que es el único no constante, el valor de la PP depende exclusivamente de sus variaciones. Luego, siempre intentamos encontrar la combinación de parámetros que maximice este término (dado a que el modelo es fijo). De esta forma, maximizamos la PP.

Pero ¿Cómo maximizamos el $\mathcal{L}$? Nosotros trabajamos con dos métodos distintos: *Markov Chains Monte-Carlo* (MCMC) vía *Metrópolis-Hasting* (MH), y *Gradiente Descendiente* (GD).

MCMC-MH es un tipo de cadena de Markov que explora el espacio de parámetros realizando saltos aleatorios en el mismo cada vez que se cumple una *Prueba de Aceptación*. Esta prueba consiste en calcular la PP del punto en el espacio de parámetros anterior (PP_{i-1}) y la actual (PP_i), tirar un número aleatorio $x \in [0,1)$ y compararlos:

$$x < min(1, \frac{PP_i}{PP_{i-1}}) \Longrightarrow \text{acepto salto} \tag{2}$$

$$x \geq min(1, \frac{PP_i}{PP_{i-1}}) \Longrightarrow \text{sorteo nuevo x} \tag{3}$$

La forma de los saltos que se realizan en el espacio de parámetros es muy parecida a la de un *random walk* y depende de qué valor mínimo de salto en cada eje le asignemos. Este método no encuentra el valor que maximiza $\mathcal{L}$, por lo que queremos ver alrededor de qué valores converge la cadena que implementamos y luego calcular el promedio para conocer estos valores.

GD, en cambio, explota las propiedades del gradiente y, de manera iterativa (de a saltos), se va acercando a la zona con menor valor de χ^2 (que es lo mismo que maximizar $\mathcal{L}$). χ^2 (Ec. (4)) es un estimador de qué tan alejado está mi modelo ($f(\vec{\theta}, x_i)$) de los datos (y_i).

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(\vec{\theta}, x_i))^2}{\sigma_i^2} \tag{4}$$

Donde el par (x_i, y_i) son los datos obtenidos y σ_i el error de cada uno.

Luego, el método de GD se puede escribir, de manera iterativa, como

$$\vec{\theta_i} = \theta_{i-1} - \eta \nabla \chi^2|_{\theta_{i-1}} \tag{5}$$

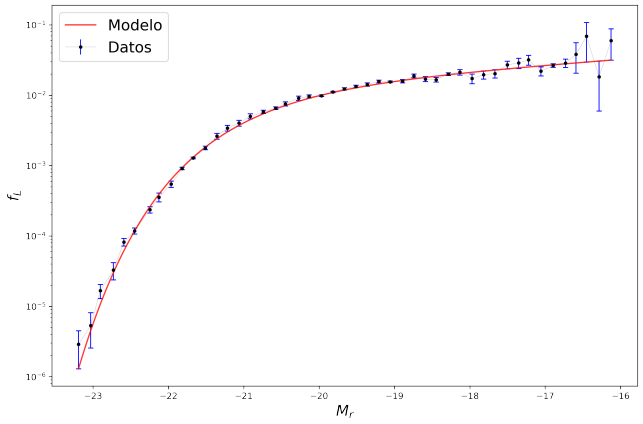

Figura 1: Modelo de Blanton et al. (2001) para la función de luminosidad $f_L(M_c)$ (en línea continua) para sus datos y sus errores correspondientes (puntos oscuros).

Ahora, θ_i son los parámetros nuevos y θ_{i-1} los viejos. Se agrega la *tasa de aprendizaje* η (que puede ser constante o no), y se deben probar varios valores hasta encontrar uno que devuelva resultados convincentes.

2. Datos y Modelo

Los datos con los que trabajamos corresponden a la función de luminosidad de galaxias vs. magnitud absoluta en la banda r ($f_L(M_r)$). A este conjunto de datos lo representamos con el modelo de Schechter (Ec. (6)).

$$f(M_r) = 0.4ln(10)\phi 10^{-0.4(M-M_c)(\alpha+1)} \times exp(10^{-0.4(M-M_c)}) \quad (6)$$

Donde ϕ, α y M_c son los parámetros que ajustamos para maximizar $\mathcal{L}$. Los pares de datos a analizar fueron los obtenidos por Blanton et al. (2001), cuyo modelo ajustado corresponde a la elección de parámetros $\phi = (0.0146 \pm 0.0012)$, $M_c = (-20.83 \pm 0.03)$ y $\alpha = (-1.20 \pm 0.03)$, que llamaremos de ahora en adelante "parámetros de Blanton"(Fig. 1).

Supusimos que todos los datos son independientes entre sí, entonces la probabilidad de un par de datos dado el modelo y los parámetros (es decir, $\mathcal{L}(\phi, \alpha, M_c)$) viene dado por Ec. (7).

$$\mathcal{L} = \prod_{i=1}^{n} f(x_i, \phi, \alpha, M_c) \quad (7)$$

Donde x_i son las mediciones de M_r de Blanton et al. (2001) y n es la cantidad de estas mediciones (cantidad de datos).

2.1. Espacio de parámetros

Para comenzar a implementar los distintos métodos de maximización de $\mathcal{L}$ variamos los parámetros involucrados en Ec. (6) (partiendo desde los de Blanton) para encontrar cotas inferiores y superiores y así delimitar el dominio del espacio de parámetros. Este espacio es el que define dónde el *Prior* es distinto de 0, y por ende, dónde voy a maximizar mi $\mathcal{L}$.

Entonces, encontramos los rangos a los que pertenece cada uno variando la forma de posible envolventes a todos los datos. Obtuvimos: $\phi \in (0.001; 0.044)$, $\alpha \in (-2.0; -0.8)$ y $M_c \in (-23.0; -19.5)$. Por lo que nuestro dominio del espacio de parámetros resulta una región de $\mathbb{R}^3$ (prisma) de lados $\Delta\phi = 0.043$, $\Delta\alpha = 1.2$ y $\Delta M_c = 3.5$.

3. MCMC-MH

Una vez definidos el modelo, el espacio de parámetros y los métodos para encontrar la mejor combinación posible para los datos originales, comencemos implementando MCMC-MH sobre la región definida.

3.1. Explorando el espacio de parámetros

Sorteamos un *guess inicial* tal que su *Likelihood* no sea muy alto. Es decir, comenzamos lejos de los valores óptimos. Implementamos el algoritmo y de comparar las posiciones en cada *step* de la cadena en el espacio de parámetros, obtuvimos la Fig. 2.

A primera vista notamos que los pasos finales de la cadena oscilan entre valores ligeramente mayores a los de Blanton. Veamos mejor esto en la Fig. 3, donde graficamos el valor de cada parámetro paso a paso.

Acá comprobamos nuestra hipótesis, y es que casi todos los θ_i caen por encima de los valores de Blanton.

Por completitud decidimos realizar otra corrida del algoritmo, pero forzando las condiciones iniciales para que comencemos desde el otro lado del espacio de parámetros. Es decir, en vez de comenzar con valores menores, probamos con mayores. El resultado fue similar, por lo que asumimos que los resultados no estaban sesgados por la posición inicial, y que las cadenas convergen cerca del paso 2000. Esto significa que puedo calcular el promedio de los valores pertenecientes a la cadena desde el paso 2000 y obtener resultados preliminares de nuestros parámetros óptimos.

Para este primer intento obtuvimos que $\phi = (0.0154 \pm 0.0004)$, $M_c = (-20.80 \pm 0.02)$ y $\alpha = (-1.18 \pm 0.02)$, que pertenecen al intervalo propuesto por Blanton et al. (2001).

3.2. Múltiples MCMC-MH

Una vez probado el algoritmo, procedimos a realizar de manera simultánea 5 *runs* con un *guess inicial* aleatorio (diferentes y alejados del máximo de *Likelihood*), y así estudiar el comportamiento de las mismas de manera más general, sin importar la condición inicial determinada. Los resultados de estas cadenas se muestran en la Fig. 4. Donde cada panel muestra la frecuencia de valores de cada parámetro.

Luego, así como mencionamos en la Sec. 3.1., nos quedamos sólo con los valores de la cadena después del paso 2000 y calculamos el promedio de los promedios de cada parámetro. Para ver si se comparan con los valores de Blanton, colocamos nuestros resultados y los de Blanton et al. (2001) en la Tabla 1.

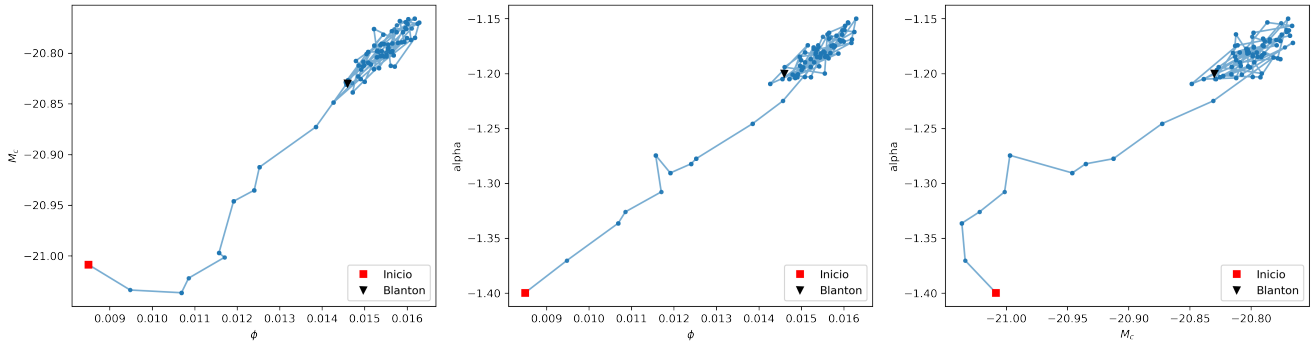

Figura 2: Recorrido de nuestra primera *run* del algoritmo MCMC-MH en las 3 proyecciones posibles de nuestro espacio de parámetros. En línea continua se muestra la evolución de nuestro *guess* paso a paso. Marcado con un triángulo negro se muestran los valores de Blanton y con un cuadrado rojo nuyestro *guess inicial*. *Panel izquierdo*: M_c vs ϕ. *Panel central*: α vs ϕ. *Panel derecho*: α vs M_c.

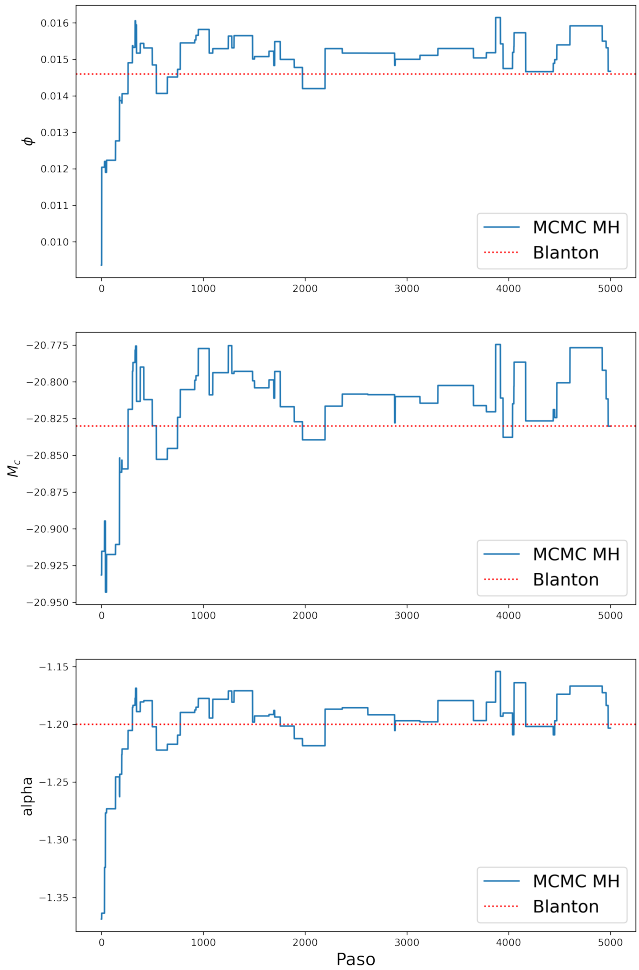

Figura 3: Valor de cada parámetro en función del *step* de nuestra cadena. En línea continua se grafican los valores en función del paso, y en línea punteada los valores de Blanton para cada uno. *Panel superior*: ϕ. *Panel central*: M_c. *Panel inferior*: α.

Tabla 1: Comparación de los valores de parámetros encontrados vs. los de Blanton. No se agregaron las incertidumbres de nuestros resultados dado a que eran despreciables.

Parámetro	Blanton et al.(2001)	MCMC-MH
ϕ	(0.0146 ± 0.0012)	0.0153
M_c	(-20.83 ± 0.03)	-20.80
α	(-1.20 ± 0.03)	-1.18

3.3. Mal Mixing

Algo importante aclarado en Sec. 3.1. fue que una condición necesaria para el *guess inicial* era que el valor de la función *Likelihood* en ese punto no debía ser alta, pero no especificamos la ubicación de este punto en el espacio de parámetros. Y es que es muy importante que estas distintas realizaciones de cadenas comiencen en puntos del espacio distantes entre sí, para evitar que alguna de ellas caiga en un máximo local (si es que hay) en vez de en el máximo global de $\mathcal{L}$, que es lo que buscamos.

No tener en consideración esto puede causar un *Mal mixing*, lo que significa que la cadena va a converger a valores no óptimos, diferentes a los obtenidos por Blanton et al. (2001).

No lo mostramos en este trabajo, pero se puede ver que para este caso en particular, si graficamos $\mathcal{L}(\phi)$, $\mathcal{L}(M_c)$ y $\mathcal{L}(\alpha)$ podemos identificar sólo 1 máximo en cada caso. Por lo que nos despreocupamos de un posible *Mal Mixing*.

4. Gradiente Descendiente

El método alternativo a MCMC-MH para optimizar la elección de parámetros que me maximizan la PP es el de *Gradiente Descendiente*. En este caso no nos interesa sortear randoms ni evaluar la PP para distintos *steps*, sino que calculamos las derivadas de χ^2 respecto de cada parámetro $(\frac{\partial \chi^2}{\partial \theta_i})$ para conocer el valor de cada componente del gradiente de Ec. (5) en cada *step*, y así acercarnos al máximo de $\mathcal{L}$.

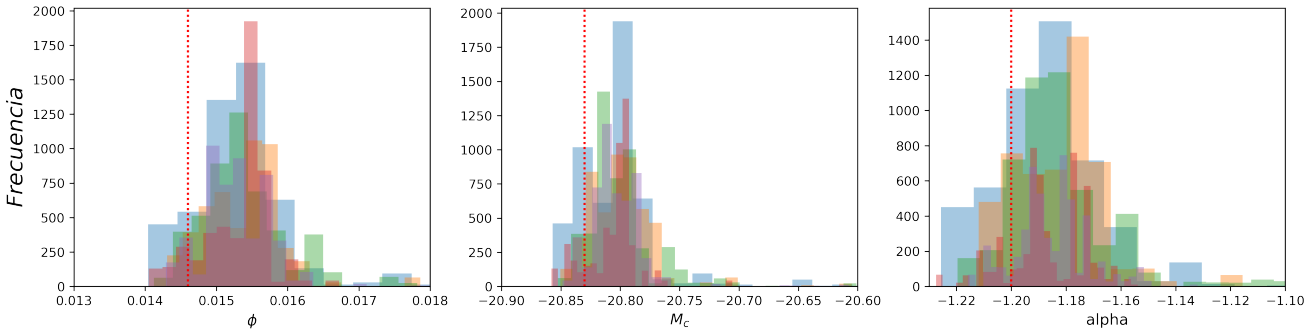

Figura 4: Histogramas superpuestos de 5 realizaciones de MCMC-MH. En barras de colores (bines = 30) las frecuencias de cada valor del parámetro y en línea punteada el valor de Blanton. *Panel izquierdo*: ϕ. *Panel central*: M_c. *Panel derecho*: α.

4.1. Evaluando χ^2 sobre todo el espacio de parámetros

El paso preliminar para implementar este método es dividir al espacio de parámetros en una grilla y calcular el valor de χ^2 en cada una de estas celdas. Por la naturaleza del estimador χ^2 podemos decir que mientras éste sea menor, mayor será $\mathcal{L}$. Un mapeo de estos valores se muestra en la Fig. 5.

4.2. Descenso

Una vez evaluado χ^2 en el espacio y conocida la forma funcional de cada componente del gradiente pedimos un *guess inicial* aleatorio (nuevamente, cuyo valor de $\mathcal{L}$ debe ser bajo, así comenzamos lejos de los valores óptimos) y corrimos el método iterativo durante 2000 pasos, para una tasa de aprendizaje $\eta = 0.5^\star$ (Fig. 5).

Para esta primera *run*, las coordenadas del punto final en el espacio de parámetros resultaron $\phi = 0.0153$, $M_c = -20.80$ y $\alpha = -1.18$. Valores muy similares** a los de la Tabla 1, obtenidos de correr el algoritmo MCMC-MH.

Repetimos este proceso (al igual que en la Sec. 3.2.) para varios *guess inicial* distantes entre sí en el espacio de parámetros, y obtuvimos en todos los intentos resultados parecidos al mostrado en Fig. 5. Esto puede demostrarse por el mapeo de valores de χ^2 (colores) de la figura. ¡Hay sólo un mínimo en todo el espacio! tanto las cadenas MCMC-MH como el GD van a tender, sean cualesquiera los puntos de partida, a los mismos valores óptimos, porque sólo hay 1 región mejor que el resto. Esto coincide con lo dicho en la Sec. 2.

5. Conclusión

El Teorema de Bayes (Ec. (1)) nos permitió encontrar una forma de maximizar la probabilidad de los parámetros dados los datos y un modelo. Esto es, realizar un ajuste de un modelo a un conjunto de datos medidos, vía inferencia Bayesiana.

Exploramos el espacio de parámetros (subconjunto de $\mathbb{R}^3$, definido por el rango posible de valores de cada parámetro) con cadenas de Markov (método MCMC-MH) para así encontrar alrededor de qué valores oscilaban. Vimos el comportamiento de las mismas para regiones exteriores (condiciones iniciales distintas) y alrededor de los valores óptimos.

Si en vez de buscar el máximo de $\mathcal{L}$ (Ec. (7)) buscamos el mínimo de χ^2 (Ec. (4)) utilizamos el método del Gradiente Descendiente. Para ello dividimos al espacio de parámetros en una grilla y calculamos χ^2 en cada una de estas, para luego implementar el método (de manera iterativa) de GD.

Los resultados de ambos métodos resultaron idénticos para las primeras cifras significativas: $\phi = 0.0153$, $M_c = -20.80$ y $\alpha = -1.18$. Estos valores encontrados pertenecen al intervalo propuesto por Blanton et al. (2001).

5.1. Discusión

Debido a las distintas implementaciones de cada método (y que en ambos llegamos a los mismos resultados), nos vemos en condiciones para debatir acerca de las ventajas y desventajas de cada uno, por lo menos aplicado a nuestro caso en particular.

Aplicar MCMC-MH es más sencillo que GD, por el código fuente necesario, por la no necesidad de realizar cálculo analítico (calcular χ^2 y sus derivadas) y porque los métodos que involucran sortear radnoms (en este caso, para la prueba de aceptación de un salto) suelen ser robustos, necesitando meramente realizar un promedio de los valores una vez converge la cadena.

Por otro lado, GD es mucho más *estético*. Una vez calculado el valor de χ^2 en todo el espacio y la forma funcional de sus derivadas respecto a cada parámetro, el paso iterativo que nos acerca al mínimo del espacio produce un camino suave desde la posición inicial hasta el punto buscado. La contraparte es que para otra situación de modelo/datos pueden ocurrir desviaciones que nos lleven a valores de parámetros erróneos. Esto lo podemos ver en la Fig. 5, donde ocurren cambios de dirección repentinos y no se realiza un camino directo hasta el mínimo.

Ambos métodos dependen de los valores mínimos posibles de los saltos, pero GD es mucho más susceptible a introducir un valor erróneo (dependemos de prueba y error, no sólo del salto sino de la tasa de aprendizaje).

*Estos valores se eligieron luego de prueba y error.
**No son idénticos porque difieren en un $\sim 0.01\,\%$

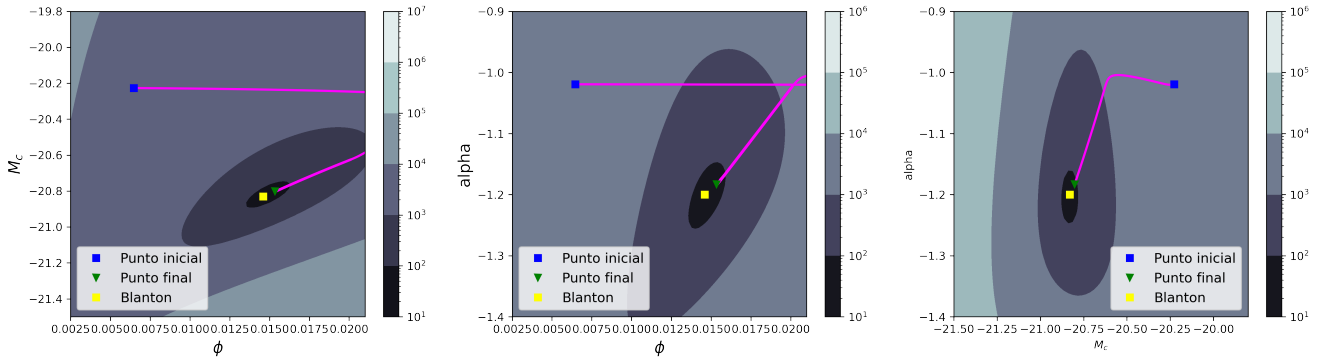

Figura 5: Valor de χ^2 en el espacio de parámetros (acotado). En escala de color logarítimica, el color oscuro corresponde a mínimos de χ^2, y el cuadrado amarillo la combinación de parámetros de Blanton. El cuadrado azul es nuestro *guess inicial*, el triángulo verde el punto final de la iteración y la línea continua magenta es el recorrido paso a paso. Cada proyección deja fijo al tercer parámetro en el valor de Blanton. *Panel izquierdo*: $\alpha = -1.2$. *Panel central*: $M_c = -20.83$. *Panel derecho*: $\phi = 0.0146$.

Por último comentamos que mapear en colores al espacio de parámetros según el valor de χ^2 nos evitó tener que calcular $\mathcal{L}(\theta_i)$, como comentamos en la Sec. 3.3. Dado a que en este caso que el *Prior* y la *Evidencia* son constantes, y los datos son independientes, pedir el mínimo de χ^2 es lo mismo que pedir el máximo de $\mathcal{L}$. De esto encontramos que existe un solo máximo en el dominio definido por lo que no hace falta preocuparse por un *Mal mixing*, y podemos utilizar cualquier punto del espacio de parámetros definido en Sec. 3.1. como condición inicial.

Agradecimientos: En agradecimiento a Marítima y Solenoide, que sin ellas este trabajo se hubiera terminado muchísimo antes.

Referencias

Blanton M.R., et al., 2001, AJ, 121, 2358