Inferencia Bayesiana con m etodos MonteCarlo: Cadenas de Markov

Camila Sofia Beatriz Hormaeche

[1] *Facultad de Matematica,Astronomia y Fisica (FAMAF)-(UNC)*

Received: ... / Accepted: ...

Resumen / En este trabajo se aplican herramientas estadísticas para ajustar un modelo con ciertos parámetros a datos observados, se utiliza métodos como Cadena de Markov-Monte Carlo (MCMC), el metodo de Metropolis-Hasting, El gradiente Descendente. Se aplica un ajuste Bayesiano al modelo de Schechter y se explora el espacio de parámetros para optimizar los parámetros del modelo. Ademas, se aplican enfoques frecuentistas y bayesianos para determinar un dado parámetro y comparar entre ambos métodos. Ademas, se analizan prior y cual es conveniente aplicar.

Keywords /

1. Introducción

En este trabajo práctico implementaremos distintos conceptos y técnicas estudiados, relacionados con la inferencia estadística, ajuste de funciones, selección de modelos, cuadrados mínimos, interpolación y minimización. Las actividades a realizar son:

- Inferencia Bayesiana, para realizar el ajuste de un modelo paramétrico a un conjunto de datos.
- Exploración del espacio de parámetros para estimar la función de Likelihood usando Cadenas de Markov Monte Carlo (MCMC).
- Implementación del algoritmo de Metrópolis-Hastings para llevar a cabo realizaciones de MCMC.
- Interpolación de datos para construir una función continua y derivable que pase por un conjunto de puntos.
- Minimización de funciones, mediante la técnica del gradiente descendente.
- Implementación de funciones en Python.

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, $p(\phi,m)$):

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

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

$$p(d|m) = \int_\Omega p(d|\phi,m)\,p(\phi|m)\,d\phi,$$

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 particular, el algoritmo de Metrópolis-Hastings es un método de MCMC que se utiliza para simular distribuciones multivariadas.

2. Función de Luminosidad de Galaxias y el Modelo de Schechter

La función de Schechter está dada por:
$\phi(M) = 0.4\ln(10)\,\phi_0\,10^{-0.4(M-M_0)(a_0+1)}\exp(-10^{-0.4(M-M_0)})$
Donde los parámetros son:

- ϕ_0: Densidad normalizada de galaxias.
- a_0: Índice de la pendiente en la región de magnitudes altas.
- M_0: Magnitud característica, donde la distribución cambia de exponencial a potencia. La función de luminosidad describe la distribución de galaxias según su brillo o magnitud absoluta. El modelo de Schechter se utiliza para describir esta distribución de manera empírica.

En la figura 1 se gráfica el modelo (función de Schechter) con los parámetros propuestos por Blanton, se observa que los datos se ajustan correctamente al modelo y se reproduce correctamente la figura obtenida por Blanton.

Se observa que los parámetros que afectan el modelo son ϕ_0,M_0,α_0, por lo que se puede estudiar como estos afectan el modelo dado que el principal objetivo de este trabajo es realizar un análisis profundo del **espacio de parámetros**.

En la figura 3 se muestra este análisis y se observa lo siguiente:

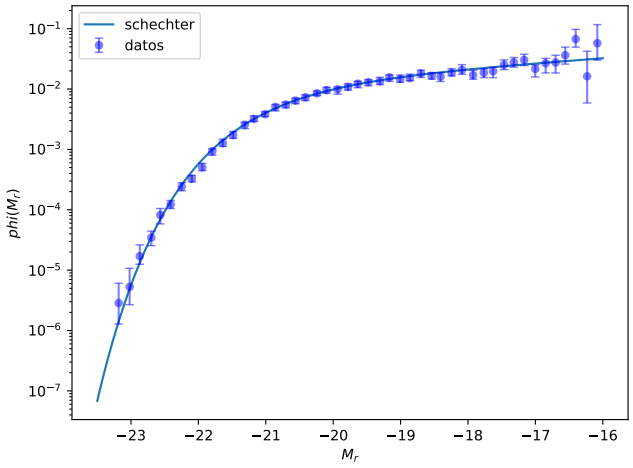

Fig. 1. Simulación de los datos observados usando la función de schechter y ajuste de modelo de Blanton

- ϕ_0: Cambiar este parámetro afecta la altura de la curva, es decir, cómo de densa es la distribución de galaxias en las magnitudes más brillantes
- α_0:Este parámetro controla la pendiente de la función, determinando cómo decrece la densidad de galaxias con magnitudes más brillantes.
- M_0: Al modificar este parámetro se desplaza la curva horizontalmente. Define el punto de transición entre galaxias luminosas y poco luminosas.

Los gráficos muestran cómo cada parámetro afecta al modelo, ayudando a comprender su importancia.

Continuando, mediante un ajuste Bayesiano, se pretende ajustar el modelo (función schechter) a los datos. Se busca estudiar el espacio de parámetros y encontrar los mejores parámetros ϕ_0 α_0 y M_0. Esto implica que:

- Función Likelihood: Mide qué tan bien el modelo explica los datos observados.
- Definición prior: Se utilizan prior planos, que verifica si los parámetros están dentro de un rango, 1 o 0.
- Función Probabilidad Posterior: Es la combinación de la verosimilitud y el prior, según el teorema de Bayes.

Las funciones utilizadas son las siguientes:

Función de verosimilitud o Likelihood:

$L = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(y-y_{model})^2}{2\sigma^2})$

Función de log-verosimilitud (log-Likelihood):

$\log(L) = A - \sum \frac{(y-y_{model})^2}{2\sigma^2}$

Función de probabilidad posterior (primer termino corresponde a la función likelihood y el segundo a los prior:

$\log(p(\phi \mid d, m)) = \log(L) + \log(P(\phi, m)) - \log(p(d \mid m))$ Este término se usa para identificar los valores óptimos de los parámetros que maximizan la probabilidad posterior. Se puede ver claramente que la función posterior maximiza la combinación de verosimilitud y prior. Los gráficos de probabilidad posterior muestran cómo cambia esta probabilidad con respecto a cada parámetro. Las curvas típicamente presentan un máximo que indica el mejor ajuste.

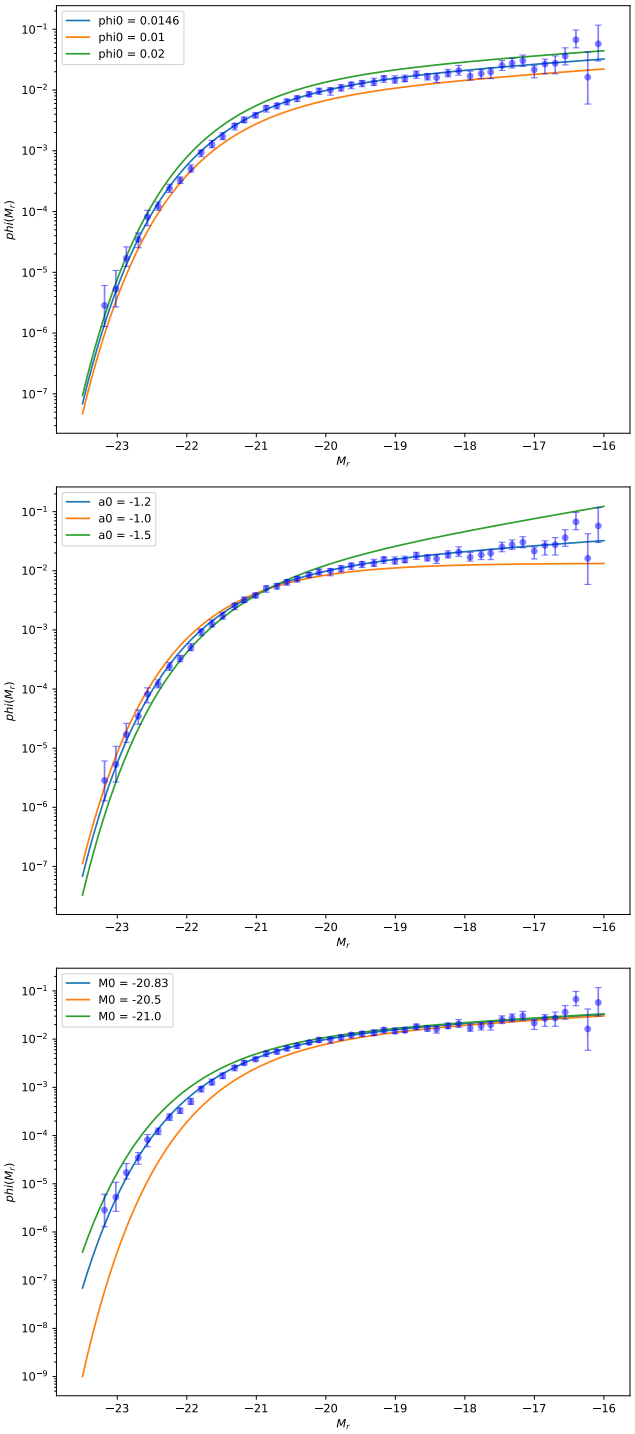

Fig. 2. Análisis de los parámetros que afectan el modelo

El logaritmo de la verosimilitud confirma que los valores de referencia propuestos por Blanton et.al proporcionan un buen ajuste.

2.1. Algoritmo Metrópolis-Hastings

Es un algoritmo para generar cadenas de Markov, diseñado para explorar el espacio de parámetros y estimar distribuciones de probabilidad posterior en el marco

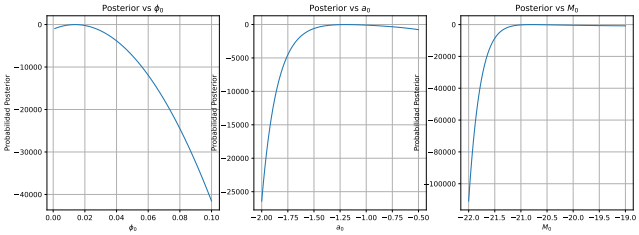

Fig. 3. La función probabilidad posterior maximiza

de la inferencia bayesiana. En el Apéndice A: Resultados del Algoritmo Metrópolis-Hastings, se observan los resultados obtenidos por este algoritmo.

- **Propuesta de un nuevo estado**: Se propone un nuevo estado en el espacio de parámetros desde el estado actual utilizando una distribución propuesta.
- **Aceptación o rechazo del nuevo estado**: Se acepta o rechaza el nuevo estado basado en una probabilidad de aceptación que evalúa qué tan probable es el nuevo estado en comparación con el estado actual.
- **Generación de una cadena**: Se genera una cadena que converge hacia la distribución objetivo.

Este algoritmo se usa cuando:

- La distribución posterior es compleja y no puede ser normalizada fácilmente.
- Se desea estimar parámetros en un modelo de probabilidad.

El código implementa el algoritmo M-H con los parámetros iniciales definidos en un rango razonable basado en datos y la literatura.

- Las cadenas muestran convergencia hacia una región específica del espacio de parámetros, lo que es consistente con un único máximo de la función posterior.
- Se simulan 15 cadenas comenzando desde diferentes regiones del espacio de parámetros. Las cadenas convergen al mismo máximo, evidenciando que el modelo tiene un único máximo global y un buen funcionamiento del algoritmo. Muestran buena convergencia hacia un único máximo, indicando que la exploración del espacio de parámetros es efectiva
- En el histograma y marginalizacion de los parámetros ϕ_0, α_0 y M_0 se observa que convergen cerca de los valores esperados, los valores determinados por Blanton. Los histogramas sugieren que el algoritmo explora eficientemente el espacio de parámetros también.

A continuación se presentan los promedios y las desviaciones estándar de los parámetros estimados:

Parámetro	Promedio	Desviación Estándar
a_0	-1.1673	0.0286
M_0	-20.7579	0.0503
ϕ_0	0.0154	0.0009

Los valores obtenidos están en concordancia con los parámetros de referencia.

2.2. Algoritmo del Gradiente Descendente

El gradiente descendente es un método iterativo utilizado para encontrar el mínimo (o máximo) de una función, en este caso, se emplea para maximizar la verosimilitud marginalizada respecto a cada parámetro. Calculo la derivada y me muevo en una dirección. El objetivo es encontrar los parámetros del modelo que mejor se ajusten a los datos observados y estudiar el espacio de parámetros de manera similar al Algoritmo de M-H.

El gradiente hace referencia a la derivada de una función en cada coordenada, lo que también indica el cambio en cada dirección. La derivada parcial también muestra la dirección de cambio en una dirección específica.

El algoritmo de gradiente descendente sigue estos pasos en cada iteración:

- **Cálculo del Gradiente**: Se utiliza el gradiente descendente para maximizar la verosimilitud de un modelo de Schechter ajustado a datos de galaxias. El gradiente indica la dirección de mayor aumento de la función.
- **Actualización de Parámetros**: Los parámetros se actualizan moviéndose en la dirección negativa del gradiente. La magnitud del paso depende de la tasa de aprendizaje (η), un escalar. Si η es muy grande, podemos sobrepasar el máximo (o mínimo), y si es muy pequeño, puede demorar en converger.
- **Convergencia**: El algoritmo continúa iterando hasta que el cambio en los parámetros entre iteraciones es suficientemente pequeño o hasta que se alcanza un número máximo de iteraciones.

El código se basa en la aplicación del método de gradiente descendente para ajustar los parámetros de la función de Schechter mediante la maximización de la log-verosimilitud. El gradiente total de la log-verosimilitud es simplemente la combinación de las derivadas parciales de la log-verosimilitud con respecto a los tres parámetros ϕ_0, α_0 y M_0.

En el Apéndice B: Resultados del Algoritmo Gradiente Descendente, se observa la evolución de los parámetros durante el proceso de optimización. Muestra cómo cambian los valores de los parámetros a lo largo de las iteraciones del algoritmo de gradiente descendente. Se observa el inicio de la cadena y el valor de referencia, que serían los valores obtenidos por Blanton, a lo que se espera llegar.

Se generan gráficos de contornos para cada par de parámetros, donde el color representa la log-verosimilitud. Los contornos ayudan a identificar cómo

varía la log-verosimilitud y dónde es máxima. Además, se dibuja la trayectoria del gradiente descendente y se marca el valor de referencia de Blanton.

Por último, se simulan varias cadenas de Markov, cada una comenzando desde un punto inicial aleatorio. Para cada cadena, se muestra cómo los parámetros evolucionan a lo largo de las iteraciones. Se marca el inicio de cada cadena con una estrella verde y el punto de convergencia con una estrella roja. Esto ayuda a identificar si las cadenas están convergiendo hacia un valor común, lo cual parece ser lo que sucede. Además, se gráfica el valor de Blanton para cada parámetro, lo que permite comparar de manera correcta los valores obtenidos.

Los valores obtenidos a partir del Algoritmo de Gradiente Descendente son los siguientes:

Parámetro	Promedio	Desviación Estándar
a_0	-1.1731	8.8978e-05
ϕ_0	0.0150	9.3435e-05
M_0	-20.7709	0.0001

Table 1. Promedio y desviación estándar de los parámetros a_0, ϕ_0, y M_0.

2.3. Conclusión de la Cadena de Markov

Entonces, se analizo y estudio dos métodos para explorar un modelo y obtener ciertos parámetros:

- **Algoritmo Metrópolis-Hasting**: Es más adecuado para problemas en los que se pretende explorar el espacio de parámetros, especialmente en contextos bayesianos donde se necesitan las distribuciones posteriores. Es útil cuando no se tiene un buen conocimiento previo sobre la distribución de parámetros. Es mejor para escapar de los mínimos locales y explorar toda la distribución de parámetros, lo que lo hace útil en problemas de alta dimensionalidad, aunque puede ser lento para converger. Depende de la elección del tamaño de salto. Para obtener una buena exploración del espacio de parámetros, el número de iteraciones necesarias puede ser muy grande.
- **Algoritmo Gradiente Descendente**: Es más adecuado para problemas de optimización rápida cuando se busca un mínimo local y la función o modelo es diferenciable. Cuando el espacio de parámetros es relativamente simple y la función de verosimilitud es suave, el gradiente descendente puede converger mucho más rápido que Metrópolis-Hastings. Es relativamente simple de implementar. Para problemas donde se busca optimizar un único punto, el gradiente descendente es mejor. Sin embargo, es susceptible a quedarse atrapado en mínimos locales. La elección del tamaño del paso de aprendizaje es crucial. El gradiente descendente no proporciona información sobre la incertidumbre o la distribución de los parámetros, solo da un punto óptimo de parámetros. En este caso, se obtuvo incertidumbre debido a la generación de varias cadenas. Además, requiere que las funciones sean diferenciables, lo que es un problema si la función que se ajusta a un modelo no es diferenciable o es discontinua.

En conclusión, la utilización de uno u otro algoritmo dependerá de lo que se esté buscando obtener o del objetivo a analizar.

3. Teorema de Bayes

Se busca aplicar el teorema de Bayes para estimar el parámetro p (la probabilidad de obtener cara al lanzar una moneda cargada), usando datos observados en un experimento donde se arroja una moneda 100 veces y se obtiene 60 caras y 40 secas. La obtención del parámetro p se realiza utilizando dos tipos de prior: uno uniforme y otro gaussiano. La idea es comparar los resultados obtenidos con ambos prior. En principio se debe definir la función de verosimilitud o likelihood, aquí se aplica una función binomial.

La probabilidad de obtener k caras en n lanzamientos, dada una probabilidad p, está dada por la distribución binomial:

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Según el teorema de Bayes, la probabilidad posterior de p dados los datos observados D es proporcional a la verosimilitud multiplicada por el prior:

$$P(p \mid D) = \frac{L(p)P(p)}{P(D)}$$

Donde $P(p)$ es el prior y $P(D)$ es la probabilidad marginal de los datos, que se usa para normalizar la distribución posterior. En este caso, la normalización se realiza dividiendo por la integral de $L(p)P(p)$ sobre todo el espacio de p.

El prior uniforme es la distribución más simple, donde todas las probabilidades de p son igualmente probables. Se define como:

$$P(p) = 1 \quad \text{para} \quad p \in [0,1]$$

Se asume, que todas las probabilidades de obtener una cara son igualmente probables en el rango $[0,1]$.

El prior gaussiano es una distribución normal centrada en $\mu = 0.5$ con una desviación estándar de $\sigma = 0.1$. Esto refleja que, la probabilidad de obtener cara es probablemente cerca de 0.5, con una cierta incertidumbre.

$$P(p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right)$$

En este caso, $\mu = 0.5$ y $\sigma = 0.1$.

Se usa la verosimilitud y el prior para calcular la probabilidad posterior.

$$P(p \mid D) \propto P(D \mid p)P(p)$$

Donde:

- $P(p)$ es el **prior** (probabilidad previa de los valores de p).
- $P(D \mid p)$ es la **verosimilitud** de los datos dados los valores de p.

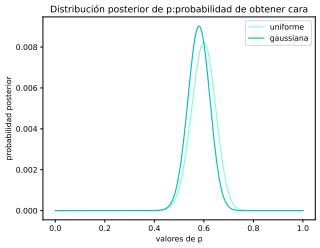

Fig. 4. Comparación entre probabilidad posterior con prior plano y prior gaussiano.

Luego, la normalizamos para que la distribución posterior sume a 1. Esto se realiza tanto para el prior uniforme como para el prior gaussiano.

- **Probabilidad posterior con prior Uniforme**: La distribución posterior resulta bastante "plana", lo que sugiere que, aunque los datos observados favorecen una probabilidad de p cercana a 0.6, la distribución no presenta una inclinación significativa hacia un valor específico de pp.
- **Probabilidad posterior con prior Gaussiano**: la distribución posterior se centra alrededor de p=0.5, pero sigue siendo influenciada por los datos.

En la figura 4, se compara ambas distribuciones y se observa cómo la elección del prior afecta los resultados de la estimación bayesiana.

4. Frecuentistas vs Bayesianos

Suponiendo, el tiempo de decaimiento de una partícula, modelado como una función exponencial con una constante de decaimiento λ, se busca comparar la estimación de esta constante mediante dos enfoques: el método frecuentista (Maximum Likelihood Estimation, MLE) y la inferencia bayesiana. El tiempo de decaimiento de una partícula sigue una distribución exponencial. La Función de Densidad de Probabilidad (PDF) de esta distribución está dada por:

$$f(t, \lambda) = \lambda e^{-\lambda t}$$

donde:

- t es el tiempo de decaimiento,
- λ es la constante de decaimiento (el parámetro que estamos tratando de estimar).

La Función de Distribución Acumulada (CDF) es:

$$F(t, \lambda) = 1 - e^{-\lambda t}$$

Si se genera un número aleatorio uniforme y entre 0 y 1,a su vez se puede generar un valor de t (tiempo de decaimiento) a partir de la función inversa de la CDF:

$$t = -\frac{\ln(1-y)}{\lambda}$$

Con lo cual, permite obtener datos de t a partir de valores aleatorios uniformemente distribuidos.

La estimación de máximo Likelihood (MLE) para el parámetro λ de una distribución exponencial se obtiene a partir de la función de Likelihood. Dada una muestra de n tiempos de decaimiento de esta función es:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i}$$

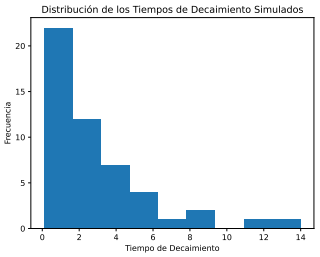

Fig. 5. Distribución de los Tiempos de Decaimiento Simulados

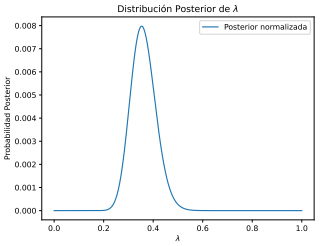

Fig. 6. Distribución Posterior de λ'

Al tomar el logaritmo, obtenemos la log-verosimilitud o log-likelihood (son lo mismo):

$$\log \mathcal{L}(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} t_i$$

Maximizando la log-verosimilitud con respecto a λ, derivamos y igualamos a cero se obtiene la estimación de MLE de λ.

$$\lambda_{\text{MLE}} = \frac{1}{\langle t \rangle}$$

donde $\langle t \rangle$ es el promedio de los tiempos de decaimiento simulados. En la figura 5, se simulan los tiempos de decaimiento, se puede apreciar una caída exponencial, pero se continua analizando esto. El valor de λ que maximiza la log-verosimilitud es el valor de λ que minimiza la suma de los tiempos de decaimiento. En lugar de estimar un único valor para λ, la inferencia bayesiana calcula una distribución posterior de λ dado los datos observados. Asumimos un prior plano (uniforme) para λ, es decir, que todos los valores posibles de λ tienen la misma probabilidad a priori en un rango dado (en este caso, entre 0 y 1).

La probabilidad posterior de λ, dada la observación de los tiempos de decaimiento, es proporcional a la verosimilitud multiplicada por el prior:

$$P(\lambda \mid t) \propto L(\lambda) \cdot P(\lambda)$$

En el código:

- La función `log_likelihood` calcula la log-verosimilitud para un valor dado de λ.
- La función `priors_plano` define el prior uniforme.
- La función `probabilidad_posterior` calcula la probabilidad posterior como el producto de la verosimilitud y el prior.

Se utilizaron prior planos para construir la probabilidad posterior. Para obtener la distribución posterior

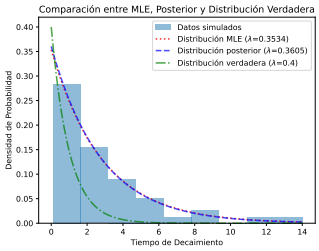

Fig. 7. Comparación entre MLE, Posterior y Distribución Verdadera, utilizando los valores de λ obtenidos para cada caso.

que se observa en la figura 6, se calcula probabilidad posterior para diferentes valores de λ y se normaliza. En la figura 7 se comparan las distintas maneras de estimar lambda con los datos simulados. Es decir,se grafican las distribuciones de probabilidad obtenidas por MLE y la distribución posterior. También se compara con el valor real de λ.

- **MLE:** La estimación de λ por MLE fue:
  $\lambda_{\text{MLE}} = 0.3534$
- **Distribución Posterior:** La media de la distribución posterior es:
  $\lambda_{\text{posterior}} = 0.3605$
- **Desviación estándar (con distribución posterior):**
  Desviación estándar $= 0.0505$
- **Valor Real:** $\lambda_{\text{real}} = 0.4$

Se compararon dos enfoques para determinar λ, el frecuentista y el bayesiano.

- **Enfoque Frecuentista**: Proporciona un valor puntual de λ basado únicamente en los datos observados, sin tener en cuenta ningún prior. El tener un prior da información adicional de los parámetros del modelo antes de observar los datos, ademas que puede guiar el proceso de estimación para evitar resultados poco realistas. Por lo que el enfoque frecuentista sirve cuando se tienen muchos datos y no se incluye un prior.
- **Enfoque Bayesiano**: Genera una distribución posterior, que tiene en cuenta tanto los datos como el prior. Esto puede ser útil cuando los datos son limitados o tienen incertidumbre. Proporciona una probabilidad posterior, por lo que se puede tener en cuenta la incertidumbre del parámetro a estimar y realizar estadística, ademas que incluye un prior, por lo que si se tiene información adicional se puede incorporar a la estimación de la probabilidad posterior y modelar mejor los datos.

Apéndice A: Resultados del Algoritmo Metrópolis-Hastings

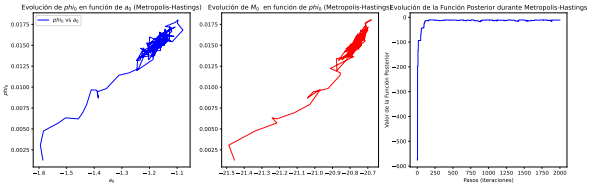

Fig. .1. Caminatas del algoritmo Metrópolis-Hastings.

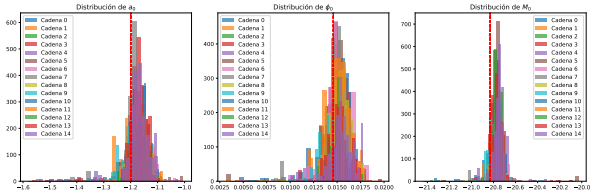

Fig. .2. Marginalización de parámetros en Metrópolis-Hastings.

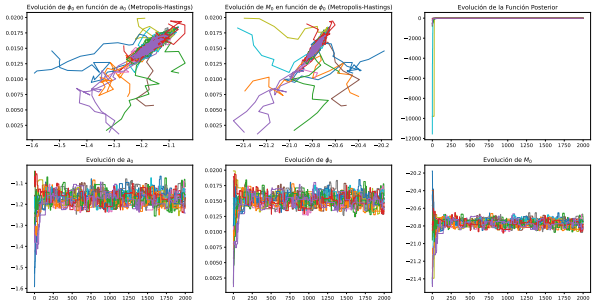

Fig. .3. Evolución de los parámetros en Metrópolis-Hastings.

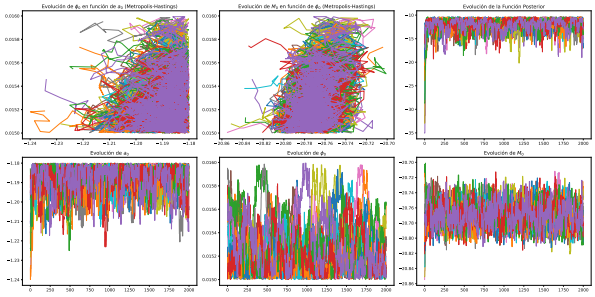

Fig. .4. Evolución de parámetros con malos valores iniciales en Metrópolis-Hastings.

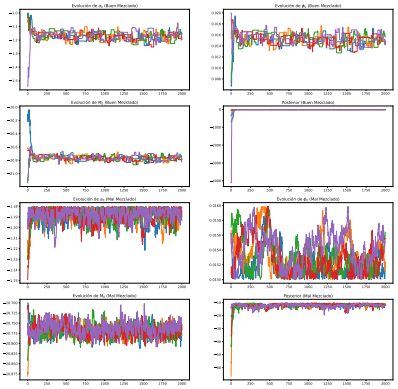

Fig. .5. Comparación entre buenos y malos parámetros en Metrópolis-Hastings.

Apéndice B: Resultados del Algoritmo Gradiente Descendente

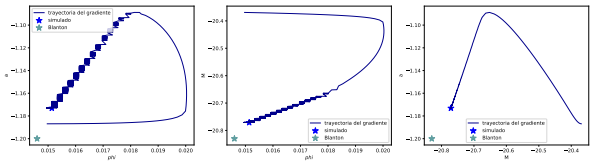

Fig. .6. Caminata y convergencia del algoritmo de gradiente descendente.

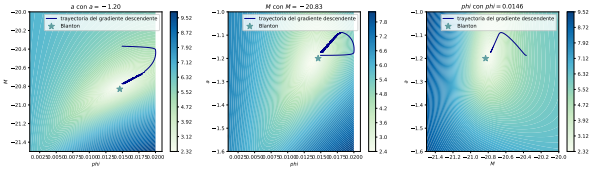

Fig. .7. Espacio de parámetros para el algoritmo de gradiente descendente.

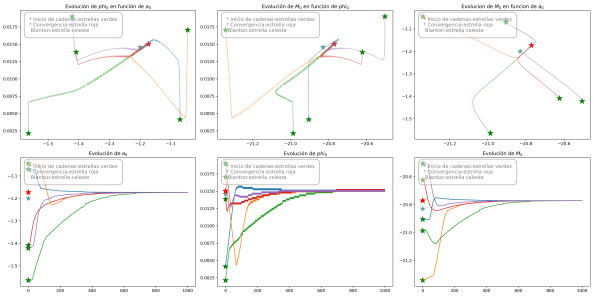

Fig. .8. Evolución de los parámetros en las caminatas del gradiente descendente.