

ESTADÍSTICA BAYESIANA Y COSMOLOGÍA

XX
XX
XX
XX
XX
XX
XX
XX

ESTADÍSTICA BAYESIANA VS FRECUENTISTA

Antes de comenzar nuestra incursión a través del mundo de un estadístico Bayesiano es importante conocer cuales son las principales diferencias que distinguen un pensamiento Bayesiano de uno Frecuentista. Así pues, en esta sección introdujeron estas discrepancias, analizando las consecuencias desprendidas de estos conceptos.

Comenzaremos definiendo el concepto más importante para cualquier procedimiento estadístico, esto es, el concepto de probabilidad. Si consideramos que x es una variable aleatoria relacionada con un evento en particular y $P(x)$ su probabilidad correspondiente, entonces para ambos casos se debe cumplir que

$$P(x) \geq 0, \quad (1a)$$

$$\int_{-\infty}^{\infty} dx P(x) = 1. \quad (1b)$$

Para eventos mutuamente excluyentes,

$$P(x_1 \cup x_2) = P(x_1) + P(x_2), \quad \text{si } x_1 \cap x_2 = \emptyset \quad (1c)$$

En general

$$P(x_1, x_2) = P(x_1)P(x_1|x_2) \quad (1d)$$

En palabras, la última regla nos dice que la probabilidad de que x_1 y x_2 pasen estará dada por el producto entre la probabilidad de que x_1 pase con la probabilidad condicional de que x_2 pase tal que x_1 ha pasado.

Y entonces, ¿en qué consiste la diferencia entre ambas disciplinas? Básicamente, la diferencia principal entre la estadística Bayesiana y la Frecuentista consiste en su definición de probabilidad. En la estadística frecuentista, **probabilidad está fundamentalmente relacionada con una frecuencia de eventos**, i.e. $p = n/N$, donde n es el numero de sucesos de un evento, mientras que N es el número total de intentos. Por otro lado, en la estadística Bayesiana, **probabilidad es fundamentalmente relacionada con nuestro propio conocimiento sobre algún evento**.

Ahora bien, cualquier procedimiento estadístico conciste de 3 ingredientes básicos que deben ser entendidos dependiendo el tipo de estadística que estemos empleando: los datos, el modelo y un método de estimación.

Los datos son una medida de nuestras observaciones, los cuales denotaremos como D . Para un estadístico Frecuentista, dichos datos son el producto de un evento repetible, mientras que para uno Bayesiano los datos se observan a partir de una muestra realizada, no necesariamente repetible, por lo que estos deben tomarse fijos para el análisis estadístico.

Luego tenemos el modelo. Este concepto difiere dependiendo de que interpretación estadística estemos utilizando. En general, un *modelo*, Q , es una colección de mediciones de probabilidades P . Las distribuciones P_θ son llamadas distribuciones del modelo, donde θ son aquellos parámetros que Q contiene. La diferencia entre ambas clases de estadísticas recae en el valor que los parámetros θ pueden tener. Mientras que para un pensamiento frecuentista los parámetros θ son fijos y existe sólo un valor del parámetro θ_0 al que nuestros datos pertenecen, para un pensamiento Bayesiano los parámetros θ no poseen un único valor. De hecho, en la estadística Bayesiana, se considera que no existe diferencia matemática formal entre parámetros y datos, por lo que para ambos casos se debe asociar una distribución de probabilidad.

Finalmente nos encontramos con el estimador de P_0 . Un estimador para P_0 es la representación de nuestra “mejor creencia” de P dados los datos D , i.e. $P_0 = P_{mejor}$.

De esta manera en la tabla ?? podemos resumir las principales diferencias entre ambas estadísticas.

Frecuentista	Bayesiano
Los datos son una muestra aleatoria repetible. Existe una frecuencia de eventos.	Los datos son observados a partir de la muestra realizada.
Los parámetros del modelo se mantienen constantes durante este proceso repetible.	Los parámetros son desconocidos y deben ser descritos probabilísticamente.
Los parámetros se toman fijos.	Los datos se toman fijos.

TABLE I: Principales diferencias entre las interpretacion Frecuentista y la Bayesiana.

INTRODUCCIÓN A LA ESTADÍSTICA BAYESIANA

En esta sección introduciremos brevemente los conceptos matemáticos básicos que son necesarios para poder emplear el formalismo Bayesiano a la cosmología. Posteriormente, en la siguiente sección nos concentraremos en las herramientas computacionales que nos ayudarán a simplificar la vida cuando un análisis analítico no sea posible (que será prácticamente en todos los casos).

Teorema de Bayes, priors, posteriores y esas cosas

El tratamiento Bayesiano a un problema estadístico siempre estará centrado en lo que es conocido como el teorema de Bayes. Este teorema es una consecuencia directa de los axiomas de la probabilidad (??). Podemos observar de (??), sin pérdida de generalidad, que siempre es posible reescribir $P(x_2, x_1) = P(x_2)P(x_2|x_1)$. Luego, como es de esperarse que la relación $P(x_1, x_2) = P(x_2, x_1)$ se cumpla, obtenemos que

$$P(x_2|x_1) = \frac{P(x_2)P(x_1|x_2)}{P(x_1)}. \quad (2)$$

Este último resultado es conocido como el *teorema de Bayes*. Por otro lado, ya que habíamos mencionado anteriormente que en la estadística Bayesiana no existe distinción formal entre datos y parámetros, podemos reescribir la ecuación anterior haciendo los cambios $x_1 \rightarrow D$ y $x_2 \rightarrow H$, como

$$P(\theta, H|D) = \frac{P(\theta, H)P(D|\theta, H)}{P(D)} \quad (3)$$

En esta expresión hemos agregado el término extra θ para especificar que H depende de dichos parámetros. Por otro lado, notemos que hemos agregado una nueva cantidad H , que llamaremos nuestra “hipótesis”. Esta última corresponde al modelo Q que según nuestra creencia mejor se ajustaría a los datos, i.e. $H = Q_{mejor}$.

Notemos que ahora, de la expresión anterior, contamos con 4 nuevos entes que deben ser entendidos a la perfección antes de poder continuar con este capítulo. En primera instancia $P(\theta, H|D)$ es conocida como la distribución posterior de probabilidad (o simplemente posterior). Esta cantidad es básicamente el resultado principal a obtener ya que nos habla sobre la probabilidad de nuestro modelo (o parámetros del modelo), tal que hemos obtenido cierto conjunto de datos D . Usualmente, dicha cantidad es utilizada para acotar los parámetros del modelo. Luego tenemos los denominados “priors”, $P(\theta, H)$, los cuales son una distribución de probabilidad para nuestros parámetros, definidos a partir de nuestro propio conocimiento acerca del modelo. Generalmente, en el límite donde tenemos muchos datos, estos priors no son estadísticamente importantes, por lo que un prior típico a tomar para cada parámetro del modelo es uno plano (una distribución uniforme). A continuación tenemos lo que es conocido como el Likelihood $L(D; \theta) \equiv P(D|\theta, H)$ y es básicamente el elemento más importante al momento de hacer un estudio estadístico para la inferencia de los parámetros de nuestro modelo. Nos concentraremos en este elemento más adelante. Finalmente nos encontramos con la evidencia Bayesiana (o simplemente evidencia). Podemos notar que ésta actúa como un factor de normalización

$$P(D) = \int d\theta P(\theta, H)P(D|\theta, H). \quad (4)$$

De esta manera, esta cantidad es usualmente ignorada cuando el espacio de parámetros de un único modelo es probado. Por otro lado, esta cantidad resulta ser fundamental si lo que se quiere es hacer una comparación entre modelos. Sin embargo, para los propósitos de este capítulo, no nos interesaremos en calcular $P(D)$.

Podemos ver que el teorema de Bayes tiene una implicación enorme respecto a una inferencia (de parámetros, por ejemplo) desde el punto de vista Bayesiano. En un escenario típico podemos tomar un conjunto de datos y esperar interpretarlos en términos de algún modelo. Sin embargo, lo que usualmente podemos hacer es lo opuesto, es decir, podemos tener un conjunto de datos y posteriormente confrontar un modelo tomando en cuenta la probabilidad que nuestro modelo se ajuste a los datos. De esta manera, como se puede ver de (??), el teorema de Bayes nos permite relacionar ambos escenarios, dándonos la posibilidad de conocer cual es el modelo (o parámetros del modelo) que mejor se ajusta a los datos.

¿ Y qué hay acerca del Likelihood?

Nuevamente tomando el teorema de Bayes (??), si ignoramos la evidencia Bayesiana y consideramos un prior uniforme, el cálculo del posterior estará únicamente determinado por maximizar el Likelihood. Por supuesto, el ignorar la evidencia ocasiona que no sea posible el poder dar una probabilidad absoluta de los parámetros de nuestro modelo, sin embargo, lo que sí podemos hacer es dar una probabilidad relativa definida como el cociente entre probabilidades. De esta manera, el Likelihood en un punto particular en el espacio de parámetros puede ser comparado con el que mejor se ajuste a las observaciones, L_{max} . Podemos decir entonces que un modelo es aceptable si el cociente de los Likelihoods

$$\Lambda = -2 \ln \left[\frac{L(D; \theta, H)}{L_{max}} \right] \quad (5)$$

es mayor que algún número dado.

Por otra parte, notemos que si nuestro posterior posee un único máximo global en θ_0 , siempre es posible hacer una expansión en serie de la forma

$$\ln L(D; H) = \ln L(D; H_0) + \frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) \frac{\partial^2 \ln L}{\partial \theta_\alpha \partial \theta_\beta} (\theta_\beta - \theta_{0\beta}) + \dots \quad (6)$$

donde H_0 corresponde al modelo con los parámetros que mejor se ajustan a los datos y $\theta_{0\alpha}$ son los componentes del vector de parámetros θ_0 .

De esta manera, para estos casos podemos reescribir nuestro Likelihood como

$$L(D; H) = L(D; H_0) \exp \left[-\frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) H_{\alpha\beta} (\theta_\beta - \theta_{0\beta}) \right] \quad (7)$$

donde

$$H_{\alpha\beta} = \frac{\partial^2 \ln L}{\partial \theta_\alpha \partial \theta_\beta} \quad (8)$$

es llamada la matriz Hessiana y controla si la estimación de θ_α y θ_β se encuentran correlacionadas. Si esta es diagonal, se dice que las estimaciones no están correlacionadas.

Aquí es preciso remarcar que esta aproximación es buena siempre que contemos con un único máximo global en nuestro posterior. Si por el contrario éste poseiera múltiples máximos locales, entonces la aproximación no serviría.

Actualizando los posteriores

Una pregunta natural que puede surgir al momento de hacer cualquier análisis Bayesiano es: ¿qué pasa cuando obtengo nuevos datos? es decir, ¿puedo hacer uso de la información que ya conocía o es preciso volver a calcular todo tomando como mi conjunto de datos los datos viejos más los nuevos? Como [referencias] explica, siempre que los datos sean consistentes entre ellos, no importa cual sea la manera que se elija para analizar los datos, i.e. da lo mismo si primero se utilizan los datos viejos para obtener un posterior y luego dicho posterior se utiliza como prior para analizar los datos nuevos, que tomar el conjunto completo de datos y analizarlos con un prior inicial, digamos uno no informativo (plano).

Chi-cuadrada

Como se comentó anteriormente, la forma precisa para encontrar el posterior de algún modelo en particular estará dado por maximizar el Likelihood. Notemos que en la aproximación gaussiana, dicha tarea es equivalente a minimizar la expresión

$$\chi^2 \equiv (\theta_\alpha - \theta_{0\alpha}) H_{\alpha\beta} (\theta_\beta - \theta_{0\beta}). \quad (9)$$

La cantidad χ^2 es usualmente llamada *chi-cuadrada* y se relaciona con el Likelihood Gaussiano vía $L = L_0^{-\chi^2/2}$. Así pues, podemos decir que el proceso de maximizar un Likelihood Gaussiano con el de minimizar un chi-cuadrada son equivalentes. Sin embargo, en las circunstancias donde un Likelihood no pueda ser bien especificado por una distribución gaussiana, entonces ambos procesos serán totalmente distintos.

Curvas de contorno y regiones de confianza

Una vez que se han obtenido los parámetros que mejor se ajustan a los datos, lo que nos gustaría saber es si hay regiones de confianza donde otros parámetros pueden ser considerados como buenos candidatos para nuestro modelo. Una elección natural sería considerar regiones donde χ^2 es menor que algún cierto valor dado. Para el caso de distribuciones Gaussianas multi-variadas, como es el caso de (??), dichas regiones corresponden con elipsoides.

La forma en la que estas regiones de confianza pueden ser calculadas es un poco técnica para los fines de este capítulo, sin embargo, para los lectores interesados se les recomienda revisar las referencias [ref].

Marginalización

Se puede dar el caso en el que al realizar un experimento sólo estemos interesados en analizar ciertos parámetros de nuestro modelo o existan parámetros extras considerados como ruido. Un ejemplo de estos parámetros de ruido podría ser aquellos correspondientes a efectos de calibración del experimento. Si este es el caso, lo que usualmente se hace es marginalizar sobre los parámetros que no son de interés mediante la expresión

$$P(\theta_1, \dots, \theta_j, H|D) = \int d\theta_{j+1} \dots d\theta_m P(\theta, H|D) \quad (10)$$

donde m es el número total de parámetros en nuestro modelo y $\theta_1, \dots, \theta_j$ corresponden a los parámetros en los que estamos interesados.

MÉTODOS NUMÉRICOS

En un escenario típico casi nunca es posible hacer el cálculo del posterior de manera analítica. Para estas circunstancias existen toda una caja de herramientas computacionales que pueden ayudarnos a realizar esta tarea de forma numérica. En esta sección nos enfocaremos en revisar las herramientas más simples (pero no por eso menos eficientes) que están a nuestra disposición.

Técnicas de MCMC para la inferencia de parámetros

Una secuencia X_1, X_2, \dots de elementos aleatorios se dice que es una cadena de Markov si la distribución condicional de X_{n+1} dados X_1, \dots, X_n depende únicamente de X_n . Lo que es importante de este proceso es que se puede demostrar que éstas convergen a un estado estacionario donde varios elementos sucesivos de la cadena recaen. De esta manera, es posible estimar todas las cantidades de interés con éstas (promedios, varianza, etc.).

El uso de este método en nuestro análisis Bayesiano es el de tratar de muestrear nuestro posterior con ayuda de estas cadenas. Para esto $P(\theta, H|D)$ es aproximada por un conjunto de funciones delta

$$p(\theta, H|D) \simeq \frac{1}{N} \sum_{i=1}^N \delta(\theta - \theta_i) \quad (11)$$

Así, el promedio de nuestro posterior puede ser calculado como

$$\langle \theta \rangle = \int d\theta \theta P(\theta, H|D) \simeq \frac{1}{N} \sum_{i=1}^N \theta_i \quad (12)$$

De esta manera, podemos promediar sobre funciones de nuestros parámetros mediante la expresión

$$\langle f(\theta) \rangle \simeq \frac{1}{N} \sum_{i=1}^N f(\theta_i) \quad (13)$$

Algoritmo de “Metropolis Hasting”

Como se explicó anteriormente, en un proceso MCMC es necesario proponer un nuevo paso en nuestra cadena tomando sólo como información el paso presente. Sin embargo, como es de esperarse, necesitamos un criterio de aceptación (o no) de este nuevo paso, dependiendo de si éste resulta ser mejor para nuestro modelo o no. Por otro lado, si resulta que el nuevo paso es mejor, nos gustaría no siempre aceptarlo, ya que de hacerlo, se podría dar el caso de que no estemos mapeando completamente el espacio de parámetros e ir convergiendo a lo que podría ser un máximo local de probabilidad para nuestro posterior. El algoritmo más simple que contiene toda esta información en su metodología es conocido como el *algoritmo de Metropolis Hasting* (MHA, por sus siglas en inglés). Éste básicamente consiste en lo siguiente: Comenzando con un punto inicial aleatorio θ_i , con probabilidad posterior asociada $p_i = P(\theta_i, H|D)$, se necesita proponer un candidato θ_c tomándolo de alguna *distribución propuesta* $q(\theta_i, \theta_c)$ que sea simétrica, i.e. $q(\theta_i, \theta_c) = q(\theta_c, \theta_i)$. De esta manera, la probabilidad de aceptar un nuevo punto estaría dada por

$$p(\text{aceptar}) = \min \left[1, \frac{p_c}{p_i} \right] \quad (14)$$

De esta manera, el algoritmo completo puede ser descrito siguiendo los siguientes pasos:

1. Elegir un valor inicial aleatorio θ_i en el espacio de parámetros y calcular su correspondiente posterior.
2. Gerar un nuevo candidato en el espacio de parámetros conciderando alguna distribución propuesta y calcular su correspondiente posterior.
3. Aceptar (o no) el nuevo punto con ayuda del MHA.
4. Si el punto no es aceptado, repeterir el punto anterior en la cadena.
5. Repetir los pasos del 2 al 4 hasta tener una cadena lo suficientemente larga.

Pruebas de convergencia

Es claro que es necesario conocer algún criterio que nos permita saber cuando nuestras cadenas han convergido al estado estacionario, de lo contrario, uno podría tener una gran pérdida de tiempo computacionalmente hablando al sobre muestrear las cadenas. La forma más simple de hacer dicha prueba consiste en simplemente correr varias cadenas y ver, a ojo de buen cubero, que todas converjan a la misma región de parámetros. Sin embargo, existen métodos más formales que nos permiten hacer dicha verificación. La prueba clásica de convergencia utilizada en Cosmología es el llamado *criterio de convergencia de Gelman-Rubin* (1992) [ref]. En éste se considera que una cadena ha convergido siempre que la cantidad

$$\hat{R} = \frac{\frac{N-1}{N}W + B(1 + \frac{1}{M})}{W}, \quad (15)$$

se aproxime a la unidad. Un criterio típico de convergencia es cuando $R < 1.03$. En esta expresión M es el número total de cadenas, N el número de puntos por cadena,

$$B = \frac{1}{M-1} \sum_{j=1}^M (\langle \theta^j \rangle - \langle \theta \rangle)^2 \quad (16)$$

es la varianza relacionada entre cadenas y

$$W = \frac{1}{M(N-1)} \sum_{i=1}^N \sum_{j=1}^M (\theta_i^j - \langle \theta^j \rangle)^2 \quad (17)$$

es la varianza de cada cadena.

Algunos detalles importantes

Sobre la distribución propuesta.- La elección de una función propuesta para generar el nuevo paso en nuestra cadena es de vital importancia para la eficiencia al momento de hacer la inferencia en cuestión. Se puede dar el caso de que si el paso propuesto en la función es muy pequeño, entonces nuestro código muestrearía muy lentamente el espacio de parámetros, lo que ocasionaría que nuestro método se hiciera muy ineficiente. Por otro lado, si el paso de nuestra función propuesta es muy grande ocasionaría que este pudiera mapear de forma muy ineficiente el espacio de parámetros, haciendo saltos grandes y dejando el valor de nuestros parámetros que mejor se ajustan a los datos entre las regiones intermedias entre pasos.

Sobre el “burn-in”.- Dado que al generar nuestras cadenas solemos comenzar en un punto arbitrario del espacio de parámetros, existirán algunos puntos al inicio que no tendrán nada que ver con la región de parámetros que son de nuestro interés. A estos puntos iniciales es a lo que se le denomina el “burn-in” de la cadena y debe ser ignorada durante el análisis estadístico.

Pruebas de autocorrelación.- Una forma complementaria de ver por convergencia en una estimación utilizando MCMCs es viendo la autocorrelación entre las muestras. La autocorrelación log k es definida como la correlación entre todas las muestras y la muestra k . Ésta puede ser cuantificada por calcular [ref]

$$\rho_k = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)Var(X_{t+k})}} = \frac{E[(X_t - X)(X_{t+k} - X)]}{\sqrt{E[(X_t - X)^2]E[(X_{t+k} - X)^2]}} \quad (18)$$

donde X_i es la i -ésima muestra y X es el promedio de las muestras. Esta autocorrelación debe hacerse pequeña conforme k crece, lo que implicaría que las muestras comienzan a ser independientes.

UNA PRIMERA SESIÓN DE TRABAJO COMPLETO: AJUSTANDO UNA LÍNEA RECTA

En esta sección pondremos en uso las herramientas aprendidas anteriormente con uno de los ejemplos más simples que hay: ajustar una línea recta. Para esto, primero generaremos un conjunto de 25 datos a lo largo de la recta $y = 1 + x$. Digamos que estos datos fueron obtenidos a partir de una teoría en que la línea recta es el modelo que puede describir al sistema. Para dificultar un poco más las cosas, vamos a agregar a cada dato un ruido gaussiano con desviación estandar $\sigma = 0.3$, que pueden deberse a problemas en la medición. Con esto, nuestros datos lucirán como se pueden ver en la figura ??.

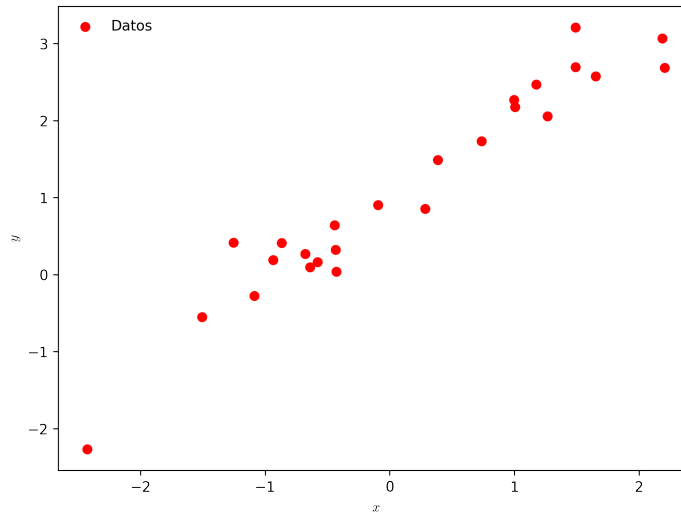


FIG. 1: Datos a lo largo de la recta $y = 1 + x$ con un ruido gaussiano con desviación estandar $\sigma = 0.3$.

Ahora bien, supongamos que estamos interesados en conocer el valor de los parámetros a y m de nuestra recta $y = a + mx$, por lo que nos gustaría estimar dichos parámetros mediante un proceso de inferencia Bayesiana. También

es posible estimar cual es la desviación estandar de nuestro conjunto de datos a lo largo de la recta, lo que nos agregaría un parámetro σ extra por estimar. Si consideramos que en principio no conocemos nada sobre el valor de nuestros parámetros libres, salvo las cotas límite en que éstos deberían estar, un buen prior para comenzar nuestra inferencia Bayesiana sería el considerar distribuciones planas. Para esto comenzaremos considerando los priors

$$a \propto U[0, 1.5], \quad m \propto U[0, 1.5] \quad \text{y} \quad \sigma \propto U[0, 0.5] \quad (19)$$

Considerando que existen valores de a , m y σ para los cuales los datos se fijan mejor (tal que el posterior es máximo global), entonces de (??) podemos escribir el Likelihood de nuestro sistema como

$$L(D; \text{recta}) \propto \exp \left[-\frac{(y_d - y)^2}{2\sigma} \right]. \quad (20)$$

Con esto ya podemos generar nuestras MCMC utilizando el MHA. Para ésto hicimos uso del módulo PyMC3 [ref] ya implementado en Python que nos permite emplear este método de forma más sencilla. Para el lector interesado, el código puede encontrarse en [ref]. En nuestro análisis corrimos un total de 6 cadenas con 10,000 pasos cada una. Nuestro resultado obtenido puede verse en la tabla ?? y la figura ?. Como podemos ver en el lado izquierdo de la figura, existen regiones para las cuales la frecuencia de eventos en nuestro muestreo se ve incrementada. De esta manera, podemos decir que dichas regiones poseen un posterior más probable para ajustarse a nuestros datos. De hecho, de la tabla ?? podemos observar que los valores reales para a y b parecen encontrarse dentro de una desviación estandar del valor promedio estimado para estos parámetros, mientras que el valor real de σ parece encontrarse ligeramente por arriba. Adicionalmente se obtuvo el valor del criterio de convergencia de Gelman-Rubin para cada variable con la intención de verificar que nuestros resultados obtenidos han convergido. Como podemos ver este número es muy cercano a la unidad, por lo que nuestro criterio de convergencia se cumple.

	Promedio	Desv. Est.	Gelman-Rubin
a	0.957023	0.075513	0.99999
b	1.059392	0.062649	1.00029
σ	0.374450	0.060069	1.00027

TABLE II: Promedios obtenidos tras nuestra inferencia Bayesiana. Se calcula también el criterio de Gelman-Rubin para la convergencia de las cadenas.

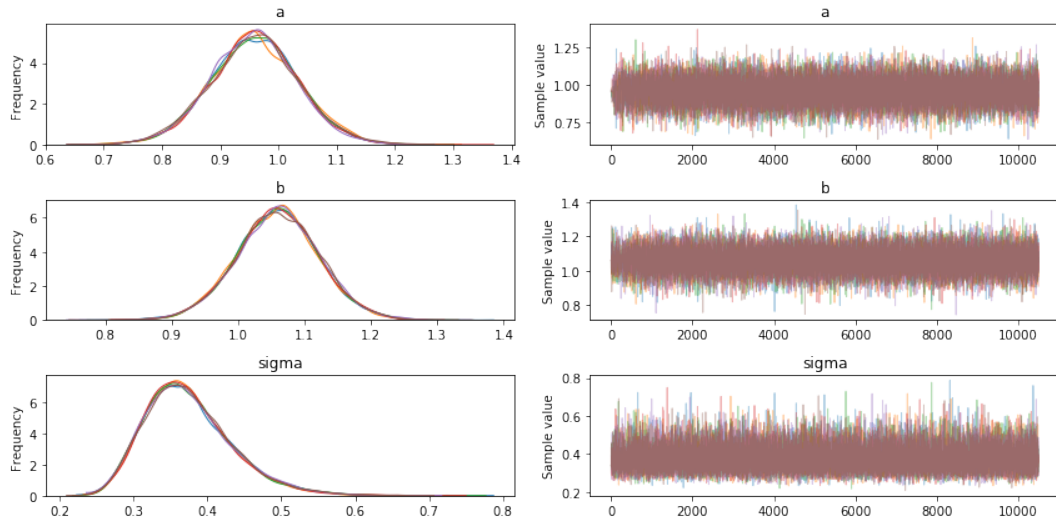


FIG. 2: Resultados obtenidos de nuestro muestreo de pasos en nuestras cadenas de Markov.

Como se mencionó, también es importante checar si no existe autocorrelación entre las cadenas mediante las pruebas de autocorrelación. Como se puede observar en la figura ??, conforme $lag\ k$ crece, la correlación tiende a cero, lo que nos dice que en nuestras muestras son independientes, con lo que podemos considerar que nuestro análisis va bien y ha convergido completamente.

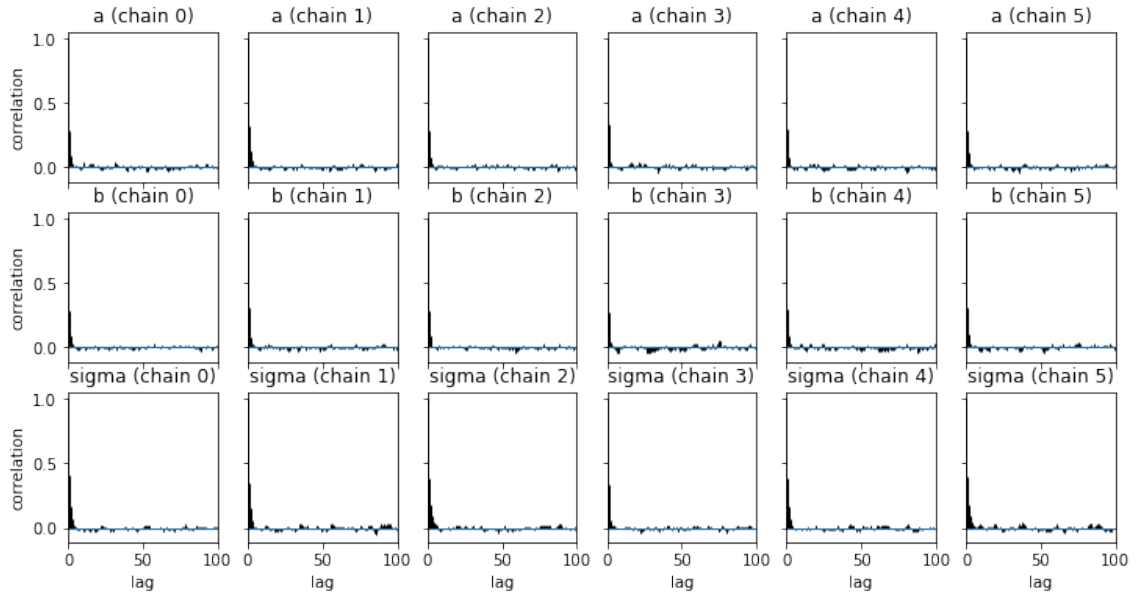
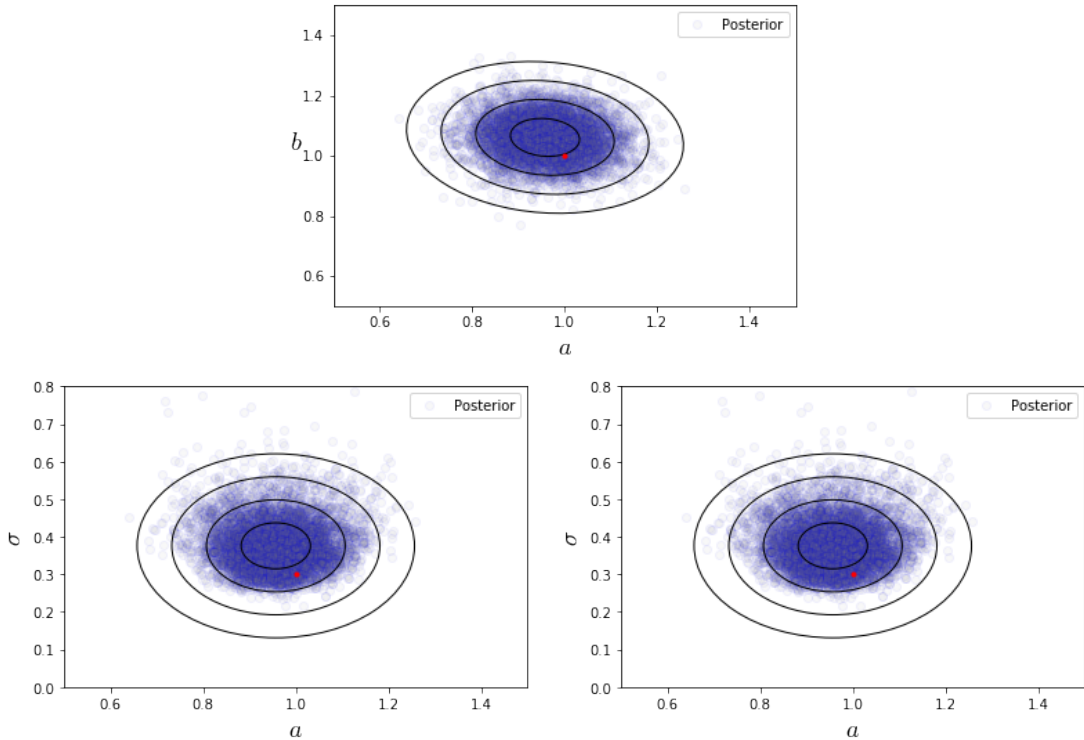


FIG. 3: Pruebas de autocorrelación.

Finalmente sólo resta mostrar las típicas regiones de confianza para nuestros parámetros. Las regiones que son usuales mostrar son aquellas que corresponden a un múltiplo de desviaciones estandar de los parámetros de interés. En la figura ?? mostramos las típicas regiones de confianza a $1-4\sigma$. Además hemos graficado en color rojo los valores reales de nuestros parámetros. Lo que podemos observar rápidamente es que ninguno de estos se encuentra alejado a más de 2σ del máximo estimado. Por supuesto, la razón de esta discrepancia es debido a que sólo contamos con 25 datos para hacer nuestra inferencia Bayesiana; de contar con más, nuestros resultados deberían ser más precisos y podríamos ser capaces de acotar en una región muy pequeña los posibles valores que nuestros parámetros podrían tener.

FIG. 4: Regiones de confianza en 2D para nuestros parámetros a , b y σ . Graficamos los contornos a $1-4\sigma$.