Antes de poder continuar con la construcción y especificación del modelo, se debe hacer una pausa para estudiar los fundamentos de la escuela bayesiana de la estadística. Esto pues el algoritmo asociado al modelo recae en un método fundamental de la disciplina: el muestreador de Gibbs, para lo cual falta definir algunos detalles más del modelo.

0.1. Fundamentos de la estadística bayesiana

Dado el problema de describir fenómenos bajo incertidumbre, existen dos escuelas dominantes de la estadística: la frecuentista y la bayesiana. La escuela bayesiana, nombrada así en honor a Thomas Bayes (1702 - 1761), enfatiza el componente *probabilista* del proceso inferencial, desarrollando un paradigma completo para la inferencia y la toma de decisiones bajo incertidumbre. Asimismo, la estadistica bayesiana está axiomatizada bajo la *teoría de la decisión*. Esta teoría formaliza conceptos como la coherencia entre preferencias y utilidad, sobre los que desarrolla un marco metodológic, **bernardo2001bayesian** y **mendoza2011estadistica**.

Esta metodología, además de proveer técnicas concretas para resolver problemas, también formaliza en una forma de pensar sobre la probabilidad como una *medida racional para cuantificar la incertidumbre*. Este paradigma es el que más corresponde con el sentido que usualmente se le da a la palabra. La inferencia o predicción sobre eventos, se realiza mediante una *actualización* de la información que se tiene bajo la luz de nueva evidencia, modificando así la medida de incertidumbre. El teorema

de Bayes es el mecanismo que permite realizar este proceso de actualización. De manera informal el teorema (1) explica que dado un evento E bajo condiciones C, la probabilidad *posterior* de ocurrencia del evento, será proporcional a la probabilidad *previa* que se tiene sobre este, ponderado por la probabilidad de ocurrencia de las condiciones presentes.

Teorema 0.1. *El teorema de Bayes (informal):*

$$P(E|C) \propto P(C|E)P(E) \tag{1}$$

Donde, el término central $P(C|E)$ es una medida descriptiva de las condiciones (usualmente datos), $P(E)$ es la probabilidad previa (*a priori*) que se tiene del evento E y $P(E|C)$ es la probabilidad posterior del evento (actualizada).

En un contexto de estadistica paramétrica más formal, los eventos E se abstraen en una serie de parámetros θ que usualmente son desconocidos. Asimismo las condiciones C quedan resumidas en datos observados $\mathbf{X}$ que son interpretados como *evidencia*. Bajo este paradigma antes de poder hacer cualquier intento de inferencia sobre θ, se debe especificar el *modelo probablistico* que se asume describe el fenómeno observado, pues es a través de este modelo que se da una medida concreta para cuantificar la incertidumbre. Primero, se tienen ciertas creencias, hipótesis u conocimiento previo, *a priori*, sobre los parámetros θ, los cuales se representan por una medida de probabilidad $\pi(\theta)$. Segundo, se tienen datos $\mathbf{X}$ a los que se asigna un modelo de probabilidad dependiente de los parámetros $\pi(\mathbf{X}|\theta)$, a la que se le conoce como *verosimilitud* (**bernardo2003bayesian**).

Teorema 0.2. *El teorema de Bayes:*

$$\pi(\theta|\mathbf{X}) \propto \pi(\mathbf{X}|\theta)\,\pi(\theta) \tag{2}$$

Habiendo especificado el modelo, el teorema de Bayes (2) describe el proceso de actualización de conocimiento sobre los parámetros θ. La idea es que este proceso de actualización sea, de la misma forma, un *proceso de aprendizaje*, en el cual los parámetros capturen la información contenida en los datos.

Bajo el paradigma frecuentista, se adopta un enfoque diferente para el aprendizaje. Se asume que no hay incertidumbre inherente en los parámetros dado los datos por lo que simplemente son desconocidos y se deben de estimar. El mecanismo que permite su estimación, usualmente consiste en plantear una función objetivo y optimizarla. Por ejemplo, si se escoge la verosimilitud $\pi(\mathbf{X}|\theta)$, se busca dar un estimador que la maximice, pues equivaldría a encontrar los parámetros que hagan más *posibles* los datos bajo el modelo planteado. Si por el contrario, es escoge una función como la suma de residuales cuadrados (RSS por sus siglas en inglés) de los modelos ANOVA, se busca la $\hat{\theta}$ que minimice los residuales, así, el modelo logra capturar toda la variabilidad posible de los datos.

Independientemente del paradigma estadístico que se escoja, siempre es importante la validación del modelo y de sus supuestos. No obstante, tanto la teoría bayesiana como la frecuentista han resultado de infinita utilidad en la práctica y el avance de la estadística y ciencia en general.

Una de las dificultades que surgen en la estadística bayesiana, es que la obtención de resultados analíticos cerrados es difícil o muy tedioso una vez que los modelos se empiezan a complicar. Por ejemplo, en las ecuaciones anteriores, se ha usado el argumento de proporcionalidad $\propto$. Esto pues, para que se de la igualdad, el lado derecho de la ecuación (2) se debe dividir entre

$$\pi(\mathbf{X}) = \int \pi(X|\tilde{\theta})\,\pi(\tilde{\theta})\,d\tilde{\theta}$$

, el cual usualmente es difícil, sino imposible, de calcular. A este término se le conoce como *constante de proporcionalidad* y su función es la de reescalar la expresión del lado derecho para que en realidad se tenga una distribución en el izquierdo. Usualmente, para evitar estas complicaciones, se escogen *distribuciones conjugadas*, para que tanto la distribución a priori como la posterior pertenezcan a la misma familia.[1] Sin embargo, con los avances en el poder computacional disponible y técnicas numéricas para resolver integrales (**robert2004monte**), se han desarrollado muchos métodos para aplicar el proceso de aprendizaje independientemente de que tan complejo sea el modelo. Muchos de estos métodos recaen en la teoría de las *cadenas de Markov*, como lo es, el muestreador Gibbs a presentarse en la sección 0.2.

1. En el apéndice **??** se detallan las distribuciones conjugadas y se realiza más a fondo la derivación de los resultados de este trabajo.

Estimadores Bayesianos

Una vez realizado el proceso de actualización, se cuenta con una distribución posterior de probabilidad para los parámetros de interés.[2] No obstante, por practicaldiad y utilidad, en ocasiones se busca dar un *estimador puntual* de los parámetros. La teoría de la decisión dicta que para medir la deseabilidad de escoger cierto parámetro en particular, se debe definir una función de perdida (L) o utilidad que optimice esta elección. Particularmente, las funciones de perdida logran medir las consecuencias incurridas, al tomar $\hat{\theta}$ como el valor puntual del parámetro. Lo hacen, penalizando la distancia entre el valor real θ y su estimador puntual $\hat{\theta}$. Por lo tanto y sin entrar mucho en los detalles técnicos, para dar un estimador puntual se resuelve el problema de optimización:

$$\hat{\theta} = \min_{\theta \in \Theta} \mathbb{E}[L(\hat{\theta}, \theta)] \tag{3}$$

con Θ el espacio de todas las posibles valores de θ. Sin embargo, se demuestra que para funciones de perdida sencillas, pero intuitivas, se tiene que el estimador puntual posterior es alguna medida de centralidad de la distribución posterior. Por ejemplo:

> Función de pérdida cuadrática: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, deriva en la media posterior, es decir: $\hat{\theta} = \mathbb{E}[\theta|\mathbf{X}]$
>
> Función de perdida valor absoluto: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, deriva en la mediana de la distribución posterior.

2. Es común tener, no es la distribución analítica, sino una muestra de ella.

Función de pérdida 0-1: $L(\hat{\theta}, \theta) = I(\hat{\theta} \neq \theta)$, deriva en la moda de la distribución posterior.

En la práctica, estas cantidades son fáciles de calcular cuando se tiene una muestra simulada de θ proveniente de la distribución posterior.[3]. En el paquete, se implementa una forma sencilla de obtener estimadores puntuales con cualquiera de las dos primeras funciones de pérdida (cuadrática y valor absoluto). Para la aplicación de este modelo, sin embargo, dado el uso de familias conjugadas, las distribuciones posteriores resultantes tienen la característica que la media, la mediana y la moda coinciden facilitando la elección por parte del analista.

0.2. Herramientas de simulación

Una vez establecida el proceso de actualización, se estudian las técnicas para simular de la distribución posterior $\pi(\theta|\mathbf{X})$. Desde principios de los años noventa, se han desarrollado algoritmos y paquetería estadística que permiten plantear modelo de una forma sencilla y obtener una muestra arbitrariamente grande de θ. Sin embargo, la gran mayoría de estos algoritmos recaen en los *métodos Monte Carlo de cadenas de Markov* (MCMC). Estos métodos, como su nombre lo indica, hacen alusión a principios de aleatoriedad, como se daría en un casino. Usando ideas intuitivas de probabilidad y números pseudoaleatorios, se pueden generar muestras prácticamente de cualquier distribución, incluso si su forma funcional es desconocida. La simula-

3. Excepto la moda muestra para los casos continuos.

ción, como tal es un tema que merece un estudio más profundo, no obstante, sus aplicaciones prácticas son muy intuitivas (**robert2004monte**). Las técnicas de simulación, permiten que los estadísticos y experimentadores puedan hacer el menor número de supuestos posibles sobre los modelos, puesto que ya no se buscan resultados analíticos sino más bien, describir el fenómeno de la forma más precisa posible y dejar los cálculos a una computadora.

Breve introducción a las cadenas de Markov

Definición 0.3. Una cadena de Markov, es una secuencia de variables aleatorias: $X^{(1)}$, $X^{(2)}$, ... que cumplen la *propiedad Markoviana*:

$$P(X^{(t+1)}|X^{(t)} = x^{(t)}, X^{(t-1)} = x^{(t-1)}, \ldots, X^{(2)} = x^{(2)}, X^{(1)} = x^{(1)})$$
$$= P(X^{(t+1)}|X^{(t)} = x^{(t)}) \quad \forall t$$

con t interpretado como *tiempo* y $x^{(t)}$ el estado en el que se encuentra la variable aleatoria $X^{(t)}$.

Esta definición, implica que la siguiente variable de la cadena, $X^{(t+1)}$, únicamente depende de el estado actual $X^{(t)}$ y no de los anteriores. Usualmente esta propiead es explicada como: el futuro, condicionando al presente, es independiente del pasado. El ejemplo canónico que se presenta es la caminata aleatoria: $X^{(t+1)} = X^{(t)} + e^{(t)}$, con $e^{(t)}$ error aleatorio generado de forma independiente. De esta idea se desarrolla toda una rica teoría revisada en cursos de procesos estocásicos **ross2009introduction**.

Una de las ideas más relevantes para lo que concierne este trabajo, es la de *matrices de transición*. Dada una cadena con n posibles estados ($X^{(t)}$ únicamente puede tomar valores de un subconjunto de cardinalidad n) se puede construir una matriz cuadrada $P \in \mathbb{R}^{n \times n}$ donde cada entrada $0 \leq p_{i,j} \leq 1$ representa la probabilidad de transicionar del estado i al estado j. Se demuestra, que si una cadena es *ergodica*,[4] entonces existe una *distribución límite* que es igual a la *distribución estacionaria*: $\exists\, \pi$, un vector de estados, tal que $\pi P = \pi$. Sin entrar en los detalles técnicos, la ergodicidad es la propiedad que asegura que eventualmente se alcanza la convergencia de la cadena sin importar el estado inicial tras repetidas aplicaciones de la matriz de transición P.[5] Esta idea se puede extender a casos más complejos donde se relajan o se cambian algunos de los supuestos. Incluso, se extiende a casos donde el número de estados es no finito, pero el concepto fundamental es el mismo. En el contexto de este trabajo, la idea es poder simular *secuencialmente* cadenas de parámetros θ que eventualmente converjan a la distribución estacionaria.

0.2.1. Muestreador de Gibbs

El el muestreador de Gibbs (*Gibbs sampler*) es método, para simular variables aleatorias de una *distribución conjunta* sin tener que calcularla directamente, (**gelfand1990sampling**) Usualmente, el muestreo de Gibbs se usa dentro de un contexto bayesiano, aunque

4. Aperiódica, irreducible y recurrente positiva. Para efectos dse simplicidad en la exposición, la ergodicidad es tratada como una propiedad en si misma. Las definiciones formales, puede ser consultadas en cualquier texto de procesos estocásticos.

5. Esta convergencia es una convergencia estocástica aplicable al paradigma bayesiano. El paradigma frecuentista, presenta resultados de convergencia que recaen en el análisis funcional (**stone1985additive**)

también funciona para otras aplicaciones. A primera vista, pareciera complejo, pero en realidad, se basa únicamente en las propiedades revisadas (relativamente sencillas) de las cadenas de Markov.

Sin perdida de generalidad, se busca simular una muestra de los parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\lambda)^t$ que provienen de la distribución conjunta $\pi(\boldsymbol{\theta})$. Esta distribución usualmente no es conocida analíticamente, sin embargo el muestreador de Gibbs permite simular una muestra arbitrariamente grande de la distribución con la que se puede aproximar empíricamente $\hat{\pi}(\boldsymbol{\theta}) \approx \pi(\boldsymbol{\theta})$. Posteriormente esta muestra se estudia con medidas de centralidad y dispersión, gráficos, cuantiles, etcétera.

Para llevar a cabo el muestreo, se intercambia el difícil cálculo de la distribución conjunta al cálculo de las distribuciones condicionales que usualmente son más fáciles de derivar. Las distribuciones condicionales están dadas por:

$$\begin{aligned} \theta_1 &\sim \pi(\theta_1 | \theta_2, \ldots, \theta_\lambda) \\ \theta_2 &\sim \pi(\theta_2 | \theta_1, \theta_3, \ldots, \theta_\lambda) \\ &\vdots \\ \theta_\lambda &\sim \pi(\theta_\lambda | \theta_1, \ldots, \theta_{\lambda-1}) \end{aligned} \tag{4}$$

Se comienza con un valor inicial arbitrario $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_\lambda^{(0)})^t$, donde el superíndice $^{(k)}$ corresponde a la iteración k. Se comienza a simular de las correspondientes distribuciones condicionales, las cuales quedan especificadas para los valores inicia-

les. En este caso, para $k = 1, 2, 3, \ldots$ se tiene:

$$\begin{aligned}
\theta_1^{(k)} &\sim \pi(\theta_1|\theta_2^{(k-1)}, \ldots, \theta_\lambda^{(k-1)}) \\
\theta_2^{(k)} &\sim \pi(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_\lambda^{(k-1)}) \\
&\vdots \\
\theta_\lambda^{(k)} &\sim \pi(\theta_\lambda|\theta_1^{(k)}, \ldots, \theta_{\lambda-1}^{(k)})
\end{aligned} \tag{5}$$

Este proceso se itera hasta tener una muestra de tamaño arbitrario, que haya alcanzado la región de probabilidad donde se encuentra la distribución estacionaria, en este caso la distribución posterior $\pi(\boldsymbol{\theta})$.

La convergencia no es intuitiva, es decir, no es trivial derivar que al muestrear de las distribuciones condicionales, se obtenga eventualmente una muestra de la distribución conjunta Sin embargo, la prueba formal recae en las mismas ideas de las cadenas de Markov. Definido el problema, se puede formar una kernel de transición, generalización de las matrices de transición, derivado de las distribuciones condicionales de $\theta_i \; \forall i = 1, \ldots, \lambda$. A la larga ($k \to \infty$) y dadas las propiedades de ergodicidad, los valores de la cadena corresponden a valores muestreados de la distribución conjunta. En **casella1992explaining** y **tierney1994markov** se presentan versiones más rigurosas de el porqué las cadenas Markov de un muestreador de Gibbs convergen.

En la práctica, una vez obtenida la cadena $\left\{\theta^{(k)}\right\}_{k=0}^{N_{\text{sim}}}$, donde N_{sim} es el número total de elementos simulados, es importante revisar si esta ya ha alcanzado la distribución posterior. Para ello, es usual revisar la media ergódica (media acumulada) de cada

parámetro, de donde se esperaría ver que la variación hacia el final de la cadena ser mínima. Asimismo, se suele analizar la traza de la cadena en sí y los histogramas de ella. Para ejemplificar, en la figura 1 se tienen tres imágenes de las cadenas simuladas por el mustreador Gibbs implementado en este trabajo,[6] en particular, las cadenas de la realización 1 del ejemplo 1 en la página la página **??**. Para esta realización en praticular, se escogen los parámetros $M = 2$, $J = 2$ y $K = 1$, implicando que se tienen rectas continuas en tres nodos ($N^* = 2$), derivando en un total de $\lambda = 5$ parámetros por estimar ($\boldsymbol{\beta} \in \mathbb{R}^5$). La imagen 1a presenta la media ergódica de todos los parámetros que se empiezan a estabilizar conforme avanzan el número de iteraciones del algoritmo. En 1b, se grafican las trazas de los primeros 3 parámetros (β_0, β_1 y β_2) y en 1c sus correspondientes histogramas.[7] Se observa como los primeros valores de los parámetros aún no se estabilizan del todo y sus medias fluctúan, asimismo, se puede observar claramente, como los histogramas tienen formas similares a la de una distribución normal; este hecho se esclarecerá en la sección **??**.

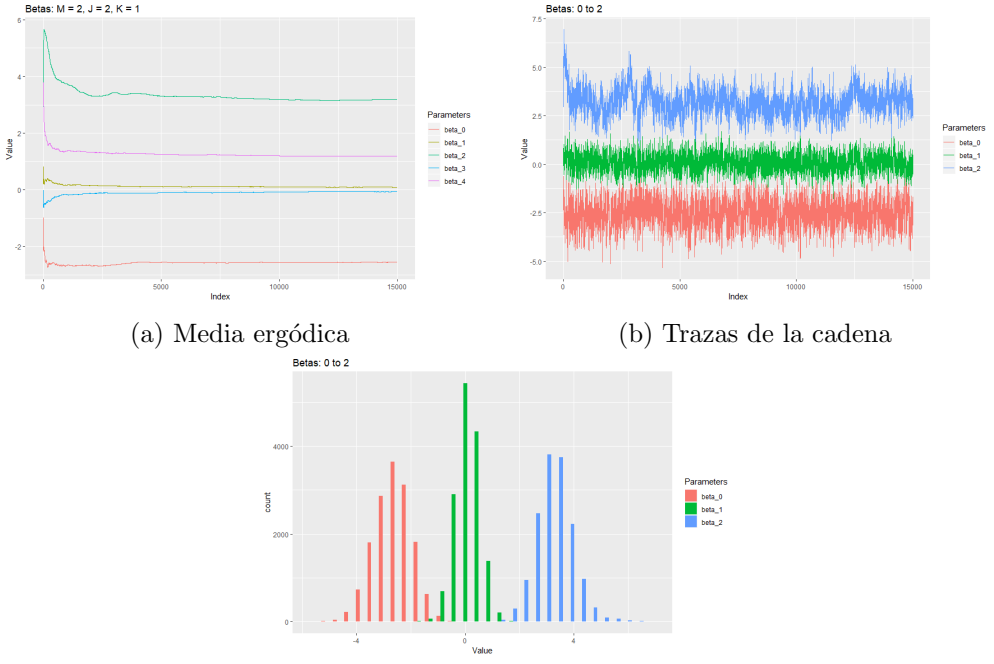
Mejoras a las cadenas

Como se observó en las imágenes previas, el muestreador de Gibbs aunque útil, no es infalible.[8] No obstante, las cadenas pueden ser mejoradas de dos formas sencillas. La primera se conoce como *burn-in* y consiste en eliminar los primeros (k^*)-esimos valores simulados de la cadena. Esto dado que el valor inicial $\theta^{(0)}$ es fijado por el

6. Las imágenes fueron generadas con la librería *ggplot2*, incorporada a las funcionalidades del paquete desarrollado para este trabajo.

7. Solamente se muestran los primeros tres parámetros para evitar tener gráficos muy saturados.

8. En el sentido que no genera una muestra de v.a.i.i.d.



(a) Media ergódica

(b) Trazas de la cadena

(c) Histogramas

Figura 1: Muestro Gibbs para el ejemplo 1 (sección **??**)

estadista, por lo que en ocasiones el algoritmo tiene que explorar una región extensa de posibles valores de θ para converger. Por lo tanto, si se busca una muestra de distribución posterior $\pi(\theta)$ los primeros valores pueden ser descartados. El corte $0 < k^* < N_{\text{sim}}$ es decidido de forma subjetiva una vez que se explora la cadena entera, ya sea por resúmenes numéricos o por representaciones gráficas. El segundo método es conocido como adelgazamiento (*thinning*) y consiste en tomar cada (k_{thin})-ésimo valor de la cadena para reducir (más no desaparecer) la dependencia entre los parámetros. Esto ocurre porque las cadenas de Markov, sobre las que depende el muestreador de Gibbs, son generadas de forma secuencial con base en el

valor actual actual de la cadena (propiedad markoviana). Por lo tanto, los valores simulados están altamente correlacionados. Sin embargo, estos sencillos pasos para mejorar las cadenas logran mejorar las muestras y ya se encuentran implementados en el paquete.