UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Departamento de Matemática

Simulación Estocástica

Francisco Cuevas Pacheco
Gabriel Riffo Jara

Departamento de Matemática
Universidad Técnica Federico Santa María
Valparaíso - Chile

https://fcocuevas87.github.io/
https://github.com/cab0512

Versión Septiembre 2023

Para la confección de este texto, se ha utilizado como base el template LaTeX desarrollado por Mathias Legrand, disponible en http://www.latextemplates.com/template/the-legrand-orange-book.

Índice general

Prefacio

La simulación se puede entender como la acción de representar algo mediante la imitación. En el momento que este documento se está redactando, la simulación se ha vuelto una herramienta relevante para enfrentar diversos problemas de las ciencias básicas, tales como neurociencia, ecología, estudios climáticos, agricultura, astronomía, entre otras.

Estos apuntes se basan en una recopilación de resultados y algoritmos que permitirán introducir al estudiante al mundo de la simulación estocástica. Este documento nace con el objetivo de estandarizar las clases para el curso *simulación estocástica MAT 448* del programa de Magíster en Ciencias mención matemática de la Universidad Técnica Federico Santa María y se inician con el trabajo de, en ese entonces estudiante del programa de magister, Gabriel Riffo (de quien solo esperaba un par de hojas y terminó en un apunte).

En estas notas introductorias se presentan los algoritmos clásicos para simular variables aleatorias y aplicaciones estadísticas de estos mismos. En general, se espera que el estudiante que lea estas notas sea capaz de comprender los elementos básicos de los métodos *MCMC* y sus consecuencias.

Las aplicaciones de estos elementos a las diferentes áreas del conocimiento se dejan como inquietud para el lector.

Introducción

Con el pasar del tiempo, los modelos matemáticos han ido evolucionando a tal punto que las diversas áreas de la ciencia recurren a estos para poder describir fenómenos de interés. Debemos tener presente que los modelos son simplificaciones del mundo real basados en diversos criterios científicos (Ver Box (1976) y *all models are wrong*) y que, como tales, permiten describir o representar ciertas características de un determinado fenómeno.

La modelación matemática se suele clasificar según varias naturalezas: según la técnica utilizada, según el área de procedencia, según la complejidad, etc. Una clasificación de interés para este curso es la basada en la aleatoriedad: si el modelo no incorpora aleatoriedad se llamará *modelo determinista*, donde las mismas entradas o condiciones iniciales producirán invariablemente las mismas salidas o resultados; en el caso opuesto se llamará *modelo estocástico*, y el objetivo será entender la distribución de probabilidad subyacente a las diversas entidades del modelo.

En ambos casos la simulación juega un rol trascendental, ya que para los modelos deterministas permiten experimentar, sin error, situaciones que podrían ser difíciles o imposibles de realizar en un laboratorio. Por otra parte, los modelos estocásticos nos permiten generar predicciones y, además, comprender las fuentes de error y como estas impactan en los distintos elementos involucrados en el modelo. En estas notas nos enfocaremos en técnicas para simular variables aleatorias y sus aplicaciones a la los modelos estocásticos.

La piedra angular de los métodos de simulación estocástica es la secuencia aleatoria $\{u_i\}_{i=1}^{\infty}$ proveniente de una distribución uniforme. Sin embargo, esta secuencia requiere ser tratada con cuidado, ya que si consideramos que un computador es una máquina que se dedica a seguir instrucciones propuestas por un usuario, entonces cualquier secuencia generada es determinista... y en efecto es así. Es por esto, que la generación de la secuencia $\{u_i\}_{i=1}^{\infty}$ no es aleatoria, si no que pseudo aleatoria y debe ser comprendida.

Estas notas están divididas en dos grandes secciones: la primera sección es una introducción a los elementos básicos de probabilidades y a los métodos tradicionales de la simulación estocástica; la

segunda sección muestra aplicaciones de la simulación estocástica a diversos problemas de la estadística, como tópicos de inferencia y análisis de datos.

Preliminares

0.1 Nociones básicas de probabilidad

En esta sección introduciremos los elementos básicos de la teoría de probabilidad para poder comprender el apunte y, además, estandarizar la notación. Muchos de estos resultados se encuentran en los libros de Durrett (2019) o Ross (2014). Comenzaremos con los elementos básicos de la teoría de la medida.

Definición 0.1.1 — σ-álgebra. Sea $\Omega \neq \emptyset$. Una colección $\mathscr{F}$ de subconjuntos de Ω se llama una σ-álgebra en Ω si y solamente si:

1. $\Omega \in \mathscr{F}$,
2. Si $A \in \mathscr{F}$, entonces $A^c \in \mathscr{F}$,
3. Si $A_1, A_2, \ldots \in \mathscr{F}$, entonces $\bigcup_{i=1}^{n} A_i \in \mathscr{F}$.

Recordamos como un caso particular a la La σ-álgebra de Borel, la cual corresponde a la σ-álgebra generada por los conjuntos abiertos. Los elementos de σ-álgebra de Borel se conocen como Borelianos.

Definición 0.1.2 — Espacio medible. *Sea $\Omega \neq \emptyset$ y $\mathscr{F}$ una σ-álgebra sobre Ω. La pareja $(\Omega, \mathscr{F})$ se llama espacio medible.*

Definición 0.1.3 — Medida de probabilidad. Se dice que $\mathbb{P} : \mathscr{F} \to [0,1]$ es una medida de probabilidad si cumple que

1. $\mathbb{P}(A) \geq \mathbb{P}(\emptyset)$, para todo $A \in \mathscr{F}$,
2. Sea $\{A_i\}_{i \in \mathscr{I}}$, una sucesión de conjuntos disjuntos con $\mathscr{I}$ siendo un conjunto numerable de índices. Entonces
$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_{i \in \mathscr{I}} \mathbb{P}(A_i),$$
3. $\mathbb{P}(\Omega) = 1$.

Definición 0.1.4 — Espacio de probabilidad. *Se define un espacio de probabilidad como la tripleta $(\Omega, \mathscr{F}, \mathbb{P})$, donde Ω es un espacio (espacio muestral), $\mathscr{F}$ es una σ-álgebra (conjunto de eventos), y $\mathbb{P}$*

es una medida de probabilidad.

Definición 0.1.5 — Variable aleatoria. *Sea $\mathbb{P}$ una medida de probabilidad y $X(\cdot)$ una función a valores reales definida en el espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$ tal que*

$$\begin{aligned} f :\Omega &\longrightarrow \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

Se dice que $X(\cdot)$ es una variable aleatoria si para cada conjunto de Borel B se tiene que

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathscr{F}$$

Notamos que una variable aleatoria induce una medida de probabilidad en $\mathbb{R}$. De hecho, basta con considerar a un conjunto $A \in \mathbb{R}$, notar que $X^{-1}(A) \in \mathscr{F}$ y luego medir el conjunto utilizando $\mathbb{P}(X^{-1}(A))$. En lo que viene del texto, utilizamos la notación clásica $\mathbb{P}(X \in A)$. Esto nos lleva a la siguiente definición

Definición 0.1.6 — Función de distribución acumulada. *La medida inducida es la llamada distribución de probabilidad y se caracteriza mediante la siguiente fórmula:*

$$F(x) = \mathbb{P}(X \leq x).$$

Definición 0.1.7 Una función $F(x)$ es una distribución de probabilidad si y solamente si

1. $F(x)$ es monótona no decreciente,
2. $\lim_{x \longrightarrow \infty} F(x) = 1$,
3. $\lim_{x \longrightarrow -\infty} F(x) = 0$,
4. $F(x)$ es continua por la derecha.

Demostración: Durret

Es decir, existe una caracterización para las funciones de distribución de una variable aleatoria. Es natural ver que toda variable aleatoria tiene asociada una función de distribución y viceversa. Para introducir el siguiente resultado, es necesario recordar algunos elementos de la teoría de la medida.

Definición 0.1.8 — Medida absolutamente contínua. *Una medida $\mathbb{P}$ es absolutamente contínua respecto a la medida de Lebesgue en $\mathbb{R}$, digamos ν, si para cada conjunto ν-medible A se tiene que $\nu(A) \implies \mathbb{P}(A)$. Denotamos esta relación como $\mathbb{P} \ll \nu$.*

El siguiente resultado es una consecuencia del teorema de Radon-Nykodim y nos muestra como expresar una distribución en términos de una función de densidad,

Teorema 0.1.1 — Función de densidad. *Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad y sea $\nu(\cdot)$ la medida de Lebesgue en $\mathbb{R}$. Si asumimos que $\mathbb{P} \ll \nu$, entonces se tiene que existe una función de densidad de probabilidad f tal que*

$$\mathbb{P}(A) = \int_A f(x)\mathrm{d}x.$$

Definición 0.1.9 — Momentos y momentos centrales. *Se define el k-ésimo momento y el k-ésimo momento centrado de una variable aleatoria X mediante*

$$\begin{aligned} \mathbb{E}[X^k] &= \int_{-\infty}^{\infty} x f(x)\mathrm{d}x, \\ M_X^k &= \mathbb{E}[(X - \mathbb{E}[X])^k], \end{aligned}$$

respectivamente y cuando estas integrales sean finitas.

El primer momento se conoce como *esperanza*, mientras que el segundo momento central se conoce como la *varianza*. Se puede observar que si el momento p_0 no existe, entonces todos los momentos $p > p_0$ no existen.

Las definiciones anteriores se pueden extender al caso en que se tengan multiples variables aleatorias. La primera extensión tiene que ver con un conjunto finito numerable de variables aleatorias descrito a continuación

Definición 0.1.10 — Distribución y densidad multivariada. *Consideremos un vector aleatorio $X = [X_1, \ldots, X_n]$ y un vector determinista $x = [x_1, \ldots, x_n]$. Entonces, se define la función de densidad conjunta del vector X por*

$$F_X(x) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

más aún, si la medida de probabilidad $\mathbb{P}$ es absolutamente contínua, se define la función de densidad conjunta $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ por

$$F_X(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n,$$

o equivalentemente

$$\frac{\partial^n F_X(x)}{\partial x_1 \cdots \partial x_n} = f_{X_1,\ldots,X_n}(x)$$

Definición 0.1.11 — Distribución marginal y densidad marginal. *Sea X un vector aleatorio particionado $X = [X^{(1)}, X^{(2)}]^\top$, donde $X^{(1)} = [X_1, \ldots, X_p]$ y $X^{(2)} = [X_{p+1}, \ldots, X_n]$. Se define la distribución marginal de $X^{(1)}$ mediante*

$$F_{X^{(1)}}(x_1, \ldots, x_p) = \mathbb{P}(X_1 \leq x_1, \ldots, X_p \leq x_p, X_{p+1} \leq \infty, \ldots, X_n \leq \infty),$$

mientras que la densidad marginal está dada por

$$f_{X^{(1)}}(x_1, \ldots, x_p) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x) \mathrm{d}x_{p+1} \cdots \mathrm{d}x_n.$$

Definición 0.1.12 — Distribución condicional. *Sea X un vector aleatorio particionado $X = [X^{(1)}, X^{(2)}]^\top$, donde $X^{(1)} = [X_1, \ldots, X_p]$ y $X^{(2)} = [X_{p+1}, \ldots, X_n]$ y sea $x^{(2)}$ es una realización de $X^{(2)}$. Se define la densidad condicional de $X^{(1)}$ dado $X^{(2)} = x^{(2)}$, denotado por $X^{(1)}|X^{(2)} = x^{(2)}$, mediante*

$$f_{X^{(1)}|X^{(2)}=x^{(2)}}(x^{(1)}) = \frac{f_X(x)}{f_{X^{(2)}}(x^{(2)})}$$

Definición 0.1.13 — Independencia estadística. *Sea X un vector aleatorio particionado de la siguiente forma $X = [X^{(1)}, X^{(2)}]^\top$, donde $X^{(1)} = [X_1, \ldots, X_p]$ y $X^{(2)} = [X_{p+1}, \ldots, X_n]$. Se dice que $X^{(1)}$ es independiente de $X^{(2)}$ si*

$$F_X(x) = F_{X^{(1)}}(x^{(1)}) F_{X^{(2)}}(x^{(2)}),$$

o, en términos de su función de densidad

$$f_X(x) = f_{X^{(1)}}(x^{(1)}) f_{X^{(2)}}(x^{(2)}).$$

La noción de independencia es fundamental para desarrollar algoritmos de simulación y, usualmente, es una propiedad que se suele buscar. Análogamente al caso univariado, podemos calcular el valor esperado y los momentos para los vectores aleatorios. En particular, nos centraremos en el valor esperado y la varianza.

Definición 0.1.14 — Valor esperado y varianza. *Sea $X = [X_1, \dots, X_n]$ un vector aleatorio. Se define el valor esperado de un vector aleatorio por*

$$\mathbb{E}[X] = [\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]]^\top,$$

Además, si $\mathbb{E}[X_i^2] < \infty$ para todo $i \in \{1, \dots, n\}$, se define la matriz de covarianza de un vector aleatorio por

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top].$$

Doob (1990). A continuación introduciremos una de las definiciones fundamentales para el desarrollo de este curso.

Definición 0.1.15 — Proceso estocástico. *Considere un espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$ y un espacio medible $(\mathscr{S}, \mathscr{U})$. Sea $D \subset \mathscr{S}$ un conjunto de índices y X_s una variable aleatoria donde $s \in D$. Un proceso estocástico X es una colección de de variables aleatorias, es decir*

$$X = \{X_s : s \in \mathscr{S}\}.$$

La secuencia $X_s(\omega)$, con $\omega \in \Omega$ se conoce como trayectoria o realización de un proceso estocástico.

Los procesos estocásticos son utilizados como modelos para describir fenómenos de distintas áreas de la ciencia, como la ingeniería, física, biología, entre otros. En la literatura se suele encontrar una extensa discusión para distintos dominios. A modo de ejemplo, si $\mathscr{S} = \mathbb{Z}$ y $D = \mathbb{Z}$, entonces el proceso estocástico es conocido como *serie temporal*, mientras que si $\mathscr{S} = \mathscr{R}^2$ y $D = [0,1]^2$, entonces estamos frente a un *proceso estocástico espacial*. Los detalles y tratamiento de cada uno de estos procesos suele requerir bastante atención, basado en lo que se suele modelar.

Como consecuencia del teorema de consistencia de Kolmogorov (Ver Durrett (2019)), un proceso estocástico puede ser caracterizado por su distribución finito dimensional. Esto requiere estudiar subconjuntos finitos de D, lo cual requiere introducir más notación.

Algoritmos de simulación

Generación de objetos Aleatorios

La generación de números aleatorios ha sido un tópico de investigación en si mismo durante mucho tiempo. Esto debido a que se genera una paradoja: *¿sé pueden generar elementos totalmente aleatorios mediante pasos deterministas?*.

En rigor, un computador genera *números pseudo aleatorios*, ya que son secuencias reproducibles que se generan con algoritmos. Estos algoritmos, a pesar de ser deterministas, emulan el comportamiento de una variable aleatoria.

Un buen generador de números aleatorios debe satisfacer las siguientes condiciones:

1. Aleatoriedad: las secuencias generadas deben pasar una serie de tests estadísticos de aleatoriedad.
2. Periodo largo: Debido a que la cantidad de estados que un computador puede tomar es finita, las secuencias, eventualmente, se repetirán cada cierto periodo (principio del palomar).
3. Eficiencia: En general, los procedimientos basados en simulación requieren un número grande de simulaciones.
4. Reproducible: Si se escoge el mismo estado inicial (o semilla), el algoritmo debería de generar la misma secuencia.

En la literatura temprana donde se generaban secuencias de números que aparentaban ser pseudo aleatorias. Algunos ejemplos son RANDU y otros

0.2 Generación de variables aleatorias uniformes

La piedra angular de cualquier algoritmo de simulación estocástica es la generación de una secuencia de números $u_1, u_2, \ldots$ provenientes de una secuencia de variables aleatorias $U_1, U_2, \ldots$ cuya densidad es uniforme en el intervalo $[0, 1]$. La importancia de la simulación como herramienta de modelado es tal, que muchos softwares poseen comandos para generar muestras provenientes de una variable aleatoria uniforme (Como el comando `runif` de R o `np.random.uniform` de Python). Cualquier algoritmo propuesto para simular variables aleatorias uniformes debe pasar los llamados test intransigentes (o

diehard test). A modo de ejemplo, veremos el generador de números aleatorios más básico de la literatura estadística.

■ **Ejemplo 0.2.1 — Generador congruencial mixto.** Comenzando con la semilla z_0, constantes a, c y M, la secuencia

$$z_k = (az_{k-1} + c) \text{mod} M$$

se conoce como generador congruencial mixto. Notar que $\{z_k\}_{i=1}^{N_{max}}$ es un secuencia de números entre 0 y $M-1$. Para generar números en el intervalo $[0,1)$ se considera un reescalamiento de la secuencia $u_k = z_k/M$.

En lo que sigue de estas notas, nos referiremos a la secuencia de números pseudo aleatorios simplemente como números aleatorios. Además, tenemos la suerte de que en estos tiempos el generador de números aleatorios uniforme es de muy buena calidad. La generación de variables aleatorias uniformes da pie a una serie de algoritmos que se explorarán a continuación.

0.3 Generación de variables aleatorias no uniformes

En esta sección revisaremos diversos métodos para simular, de manera exacta, variables aleatorias no uniformes.

0.3.1 Método de la inversa

La simulación de variables aleatorias no uniformes se basa en el siguiente teorema

Teorema 0.3.1 *Definamos la inversa generalizada de F mediante*

$$F^-(u) := \text{ínf}\{x : F(x) \geq u\}, \quad 0 < u < 1.$$

Si X es una variable aleatoria con función de distribución F, es decir $X \sim F$, y $U \sim U[0,1]$, entonces

$$F^-(U) \sim F$$

Proof: Tarea

■ **Ejemplo 0.3.2** Supongamos que se desea simular la variable aleatoria $X \sim Exp(\lambda)$, cuya distribución de densidad está dada por

$$F(x) = 1 - \exp\{-\lambda x\}, \quad x \geq 0.$$

El método de la inversa requiere calcular $F^{-1}(u)$. Mediante inspeción directa notamos que

$$F^{-1}(u) = -\frac{\log(1-u)}{\lambda}.$$

Finalmente, por el Teorema 0.3.1 se tiene que si $U \sim U[0,1]$, se concluye que $F^{-1}(U) \sim Exp(\lambda)$

■ **Ejemplo 0.3.3** Supongamos que se desea simular de variable aleatoria discreta $X \sim Geo(p)$, cuya densidad está dada por

$$F(k) = 1 - (1-p)^k \quad k \in \mathbb{N}.$$

Notamos que U es una variable aleatoria continua, mientras que los valores de $F(k)$ se pueden mapear $\mathbb{N}$, por lo que calcularemos la inversa generalizada. Para esto, notamos que para $u \in [0,1]$ fijo, el entero k mas pequeño que satisface $F(k) \geq u$ está dado por

$$k \geq \frac{\log(1-u)}{\log(1-p)}.$$

En consecuencia, se tiene que

$$F^{-}(u) = \left\lceil \frac{\log(1-u)}{\log(1-p)} \right\rceil,$$

donde $\lceil x \rceil$ es la función techo.

0.4 Teorema de transformación

Dado que se sabe simular desde una distribución uniforme, es posible obtener otras distribuciones desde el teorema de transformación, el cual se recuerda a continuación:

Teorema 0.4.1 — Teorema de transformación. *Sea $X = (X_1, \ldots, X_p)$ un vector aleatorio p-dimensional con función de densidad f_X y considere el vector aleatorio $Y = [Y_1, \ldots, Y_p]$ dado por las biyecciones $Y_i = g_i(X_1, \ldots, X_p)$. Luego, la densidad del vector Y está dada por*

$$f_Y(y) = f_X(g_1^{-1}(y_1, \ldots, y_p), \ldots, g_p^{-1}(y_1, \ldots, y_p)) \left| J\left(\frac{x_1, \ldots, x_p}{y_1, \ldots, y_p}\right) \right|,$$

donde

$$J\left(\frac{x_1, \ldots, x_p}{y_1, \cdots, y_p}\right) = \begin{bmatrix} \frac{\partial x_1}{\partial y 1} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_p}{\partial y 1} & \cdots & \frac{\partial x_p}{\partial y_p} \end{bmatrix}$$

es la matriz Jacobiana.

■ **Ejemplo 0.4.2** Sea $U \sim U[0,1]$, entonces se tiene que $(b-a)U + a \sim U[a,b]$.

■ **Ejemplo 0.4.3** Sea $U \sim U[0,n]$, entonces $\lceil x \rceil$ es la distribución uniforme discreta.

■ **Ejemplo 0.4.4** (Box-Muller) Sean $U, V \sim U[0,1]$ y considere la siguiente transformación:

$$X_1(U,V) = \sqrt{-2\log(U)}\cos(2\pi V),$$
$$X_2(U,V) = \sqrt{-2\log(V)}\sin(2\pi U).$$

Entonces, el vector aleatorio $X = [X_1(U,V), X_2(U,V)]^\top \sim N(0, I_{2\times 2})$, donde $I_{2\times 2}$ es la matriz identidad de dimensión 2.

■ **Ejemplo 0.4.5** Sea $U \sim U[0,1]$ entonces se tiene que $X = -\log(U_i)/\beta \sim Exp(\beta)$

■ **Ejemplo 0.4.6** Sea $X \sim N_d(0, I_{d\times d})$ entonces se tiene que $U = X/\|X\| \sim U(\mathbb{S}^{d-1})$

El teorema de transformación nos da permite utilizar la representación estocástica para simular distribuciones más complejas. En efecto:

■ **Ejemplo 0.4.7** Sea $U \sim U(\mathbb{S}^d)$ un vector aleatorio y R una variable aleatoria estrictamente positiva. Considere además un vector d dimensional μ una matriz definida positiva Σ tal que $\Sigma = A'A$. Luego la representación

$$X = \mu + RUA,$$

posee una distribución elíptica (Ver Cambanis, Huang y Simons (1981)).

■ **Ejemplo 0.4.8** Sean $Y_1, Y_2 \sim N(0,1)$. Luego la transformación

$$X = \frac{\theta|Y_1| + Y_2}{\sqrt{1+\theta^2}}$$

posee una distribución normal asimétrica (Ver Henze (1986))

■ **Ejemplo 0.4.9** Sean $X|\theta \sim f(x|\theta)$ una distribución con parámetro θ. Asumiendo que $\theta \sim g(\theta)$, entonces se tiene que $X \sim \int_\Theta f(x|\theta)g(\theta)\mathrm{d}\theta$ es una distribución de mixtura de escala.

0.5 Métodos basados en la función característica

En algunos contextos es común no disponer de una forma cerrada para la función de densidad, sin embargo se conoce la función característica. Recordar que la función característica de una variable aleatoria X está dada por

$$\psi(\omega) := \mathbb{E}[e^{i\omega X}] = \int_{\mathbb{R}} f(x)e^{i\omega x}\mathrm{d}x,$$

mientras que la inversa se obtiene como consecuencia directa del teorema de inversión de Fourier:

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \psi(\omega)e^{-i\omega x}\mathrm{d}\omega, \tag{1}$$

de aquí, se podría aproximar $f(x)$ en un reticulado fino utilizando técnicas de integración numérica. Considerando esto, un algoritmo para simular X sería

1. En un reticulado, digamos $\{x_1\}_{i=1}^n$, evaluar f utilizando (1),
2. Calcular los pesos $p_k = f(x_k)/\sum_{i=1}^n f(x_i)$,
3. Simular X desde la distribución discreta p_k.

Se puede generar una variante continua de este método si se considera la interpolación $p_{k+\delta} = p_k + \tau p_{k+1}$, con τ un valor determinista a escoger.

0.6 Algoritmo Aceptación-Rechazo

Hay muchas distribuciones de las que es difícil, o imposible, simular directamente mediante el método de la transformada inversa. Más aún, en algunos casos ni siquiera somos capaces de representar la distribución de forma explicita. Ni tampoco podemos usar la distribución para obtener simulaciones directas.

En estos casos tenemos que recurrir a otra clase de métodos, que sólo requieren que conozcamos la forma funcional de f salvo alguna constante multiplicativa.

La clave de este método consiste en utilizar una densidad g más fácil de simular. Para luego filtrar, de una forma apropiada, las simulaciones obtenidas. nos basaremos en el siguiente teorema.

Sea f la densidad de interés, que desde ahora denominaremos densidad objetivo, notemos que f se puede escribir como:

$$f(x) = \int_0^{f(x)} du$$

Entonces tenemos que f es la densidad marginal (de X) en la distribución conjunta:

$$(X,U) \sim U\{(x,u) : 0 < u < f(x)\}$$

A la variable U, que no esta relacionada al problema inicial, la denominaremos variable auxiliar. La idea para simular es la siguiente: generamos de la distribución conjunta variables aleatorias uniformes en el conjunto $\{(x,u) : 0 < u < f(x)\}$. Como la distribución marginal de X es nuestra densidad objetivo f al generar una variable uniforme en $\{(x,u) : 0 < u < f(x)\}$ habremos generado una variable aleatoria con densidad f. Notar que, salvo para calcular $f(x)$, en el procedimiento no hemos utilizado la densidad f. Lo cual nos lleva al siguiente teorema:

Teorema 0.6.1 — Teorema Fundamental de la simulación. *Simular:*

$$X \sim f(x)$$

es equivalente a simular:

$$(X,U) \sim U\{(x,u) : 0 < u < f(x)\}$$

Este teorema nos da el fundamento teórico para simular, sin embargo, presenta el problema que la simulación del par uniforme (X,U) no siempre es sencilla. Por ejemplo simulando $X \sim f(x)$ y $U|X = x \sim U(0, f(x))$ pero entonces esto hace que toda la representación sea inútil. Y el enfoque simétrico, que incluye simular U de su distribución marginal y X de la distribución condicional $X|U = u$ no siempre es factible. Entonces la solución entonces consiste en simular (X,U) en un conjunto mas grande, donde la simulación sea mas simple y luego tomar los pares en donde si se cumple la restricción.

Por ejemplo, en el caso 1-dimensional, supongamos que:

$$\int_a^b f(x)dx = 1,$$

donde f esta acotada por m. Luego, podemos simular del par $(Y,U) \sim U(0 < u < m)$ mediante simular $Y \sim U(a,b)$ y $U|Y = y \sim U(0,m)$ con esto, aceptamos el par (y,u) si $0 < u < f(y)$. Notemos que los valores aceptados tienen la distribución deseada. En efecto:

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(Y \leq x | U < f(Y)) \\ &= \frac{\int_a^x \int_0^{f(y)} dudy}{\int_a^b \int_0^{f(y)} dudy} = \int_a^x f(y)dy\end{aligned}$$

Entonces, si nuestra distribución esta en un conjunto A, que llamaremos conjunto objetivo, basta tomar un conjunto B tal que $A \subseteq B$ y generar una uniforme en B que aceptamos si el resultado esta también en A.

La idea anterior se puede generalizar a la situación en donde el conjunto no es una caja, sino que un conjunto simulable. Esta generalización permite simular en los casos en donde el soporte de f es un conjunto no acotado.

Si el conjunto grande es de la forma:

$$\mathscr{L} = \{(y,u) : 0 < u < m(y)\}$$

con la condición $m(x) \geq f(X)$ y que la simulación de una uniforme en L sea posible. Para lograr la mayor eficiencia se espera que m este lo mas cerca posible de f para evitar desperdiciar muchas simulaciones.

Claramente m no puede ser una función de densidad, por lo cual escribiremos $m(x) = Mg(x)$ con $g(x)$ otra función de densidad, que llamaremos densidad instrumental.

Para simular $\mathscr{L}$ simulamos primero $Y \sim g$ y luego $U|Y \sim U(0, Mg(y))$. Solo aceptaremos y si la condición $u < f(y)$ se satisface, entonces tendremos:

$$\begin{aligned}
\mathbb{P}(X \in A) &= \mathbb{P}(Y \in A | U < f(Y)) \\
&= \frac{\int_A \int_0^{f(y)} \frac{du}{M(g(y))} g(y) dy}{\int \int_0^{f(y)} \frac{du}{M(g(y))} g(y) dy} \\
&= \int_A f(y) dy
\end{aligned}$$

Esto nos lleva al siguiente colorario.

Corolario 0.6.2 *Sea $X \sim f(X)$ y sea $g(x)$ una función de densidad que satisface que $f(x) \leq Mg(x)$ para alguna constante $M \geq 1$. Entonces, simular $X \sim f$ es equivalente a generar:*

$$Y \sim g \quad \text{y} \quad U|Y = y \sim U(0, Mg(y))$$

hasta que $0 < u < f(y)$.

El corolario tiene dos consecuencias. En primer lugar, podemos simular la densidad f aunque no conozcamos la constante normalizadora de f. Pues el método solo necesita el radio f/M, el cual no depende de la constante normalizadora.

La segunda consecuencia, es que la probabilidad de aceptar es exactamente $1/M$, y por lo tanto, la cantidad de intentos esperados hasta que una variable es aceptada es M. Por lo cual una forma de definir entre varias densidades g, ie $g_1, g_2, g_3, \ldots$ es a través de comparar las constantes $M_1, M_2, \ldots$ y escoger la densidad que tenga la cota M_i mas pequeña.

0.6.1 Algoritmo Aceptación-Rechazo

La implementación del Corolario 0.6.2 es conocida como el método de Aceptación-Rechazo (o rejection sampling), el cual se presenta a continuación:

A. 1: Algoritmo Aceptación Rechazo

Dado f, g, M:

1. Generar $X \sim g$, $U \sim U(0,1)$.
2. Se acepta $Y = X$ si $U \leq f(X)/Mg(X)$.
3. En caso de no aceptar se retorna al paso 1.

En los casos en donde f y g están normalizados, y por tanto son funciones de densidad, la constante M debe ser mayor que 1. Notemos que para que (f/g) sea acotado, es necesario que las colas de g sean mas pesadas que las de f. A modo de ejemplo, esto significa que es imposible simular una distribución

de Cauchy usando una distribución normal, sin embargo si se puede generar una normal usando una Cauchy.

Para optimizar el algoritmo, se puede elegir la densidad instrumental g de una familia paramétrica, y determinar el valor de los parámetros que minimice la cota M, como en el siguiente ejemplo:

■ **Ejemplo 0.6.3 — Gamma Accept-Reject.** Si $F \sim Ga(\alpha,\beta)$ sabemos que si $\alpha \in \mathbb{N}$ entonces F se puede escribir como suma de α variables aleatorias exponenciales con $\varepsilon_i \sim Exp(\beta)$ las cuales son sencillas de simular. Sin embargo si $\alpha \notin \mathbb{N}$ simular $Ga(\alpha,\beta)$ por este método no es posible.

Una opción posible es usar el algoritmo Aceptación-Rechazo con la distribución instrumental $Ga(a,b)$, con $a = [\alpha]$ $(\alpha \geq 1)$. (Sin perdida de generalidad supongamos que $\beta = 1$). El radio f/g es $b^{-a}x^{\alpha-a}\exp(-(1-b)x)$, salvo la constante normalizadora, teniendo así la cota:

$$M = b^{-a}\left(\frac{\alpha-a}{(1-b)e}\right)^{\alpha-a}$$

para $b < 1$. El máximo de $b^{-a}(1-b)^{\alpha-a}$ se obtiene a $b = a/\alpha$. Por ende la elección óptima de b para simular $Ga(\alpha,1)$ es $b = a/\alpha$.

Sin embargo la estrategia anterior no siempre garantiza un algoritmo eficiente, como se ilustra a continuación:

■ **Ejemplo 0.6.4 — Simular una Normal a través de una Doble exponencial.** Considere simular $N(0,1)$ a través de la densidad instrumental $g(x,\alpha) = (\alpha/2)\exp(-\alpha|x|)$. Entonces a través de un calculo simple se puede mostrar que:

$$\frac{f(x)}{g(x,\alpha)} \leq \sqrt{2/\pi}\alpha^{-1}\exp\left(\alpha^2/2\right),$$

donde la cota inferior se alcanza cuando $\lambda = 1$. Notar que en ese caso la probabilidad de aceptar es $\sqrt{\pi/2e} = 0{,}76$, por lo cual, para producir una variable aleatoria $N(0,1)$ necesitaríamos generar, en promedio $1/0{,}76 = 1{,}3$ variables uniformes, lo cual es mucho menos eficiente que el algoritmo Box-Muller, en el cual solo necesitábamos una variable uniforme.

0.6.2 Algoritmo Aceptación-Rechazo con envolturas

En muchos casos, la distribución asociada a la densidad f es difícil de simular debido a la complejidad de la propia función f, que puede requerir un tiempo de cálculo considerable en cada evaluación. Para estos casos, podemos usar una forma de acelerar el algoritmo Accept-Reject a través de una función g_l mas simple de calcular y que acota por abajo de f. Este algoritmo se conoce como Algoritmo Aceptación-Rechazo con envolturas, y se basa en el siguiente lema, el cual es una extensión del Corolario 0.6.2:

Lema 0.6.5 *Si existe una densidad g_m y una función g_l y una constante M tal que:*

$$g_l(x) \leq f(x) \leq Mg_m(x)$$

entonces el algoritmo

A. 2: Algoritmo Envelope Accept-Reject

Dado f, g_m, g_l, M:

1. Generar $X \sim g_m$, $U \sim U(0,1)$.
2. Se acepta $Y = X$ si $U \leq g_l(X)/Mg_m(X)$.
3. En otro caso, aceptar si $U \leq f(X)/Mg_m(X)$.
4. Si no se acepta X, repetir 1

produce variables aleatorias con distribución f.

Este algoritmo permite disminuir potencialmente la evaluaciones que hacemos de f en un factor:

$$\frac{1}{M}\int g_l(x)dx$$

que es donde la probabilidad de f no esta evaluada. Este método sigue por el principio de compresión propuesto por Marsaglia (1977). Una forma de obtener las cotas g_l es a través de usar la serie de Taylor de $f(x)$, lo cual se ilustra en el siguiente ejemplo:

■ **Ejemplo 0.6.6 — Cota para una distribución normal.** Por la expresión en serie de Taylor de $\exp(-x^2/2)$ tal que $\exp(-x^2/2) \geq 1-(x^2/2)$ y entonces:

$$\left(1-\frac{x^2}{2}\right) \leq \exp(-x^2/2)$$

teniendo así una cota inferior para simular $N(0,1)$. Notar que la cota solo es útil cuando $|X| \leq \sqrt{2}$, evento que ocurre con probabilidad 0,61.

■ **Ejemplo 0.6.7 — Variables Poisson proveniente de variables logísticas.** Este algoritmo fue propuesto por Atkinson (1979), como una alternativa para simular una distribución $Poiss(\lambda)$ usando su relación con la distribución logística. Recordar que la distribución de densidad y distribución de una V.A logística está dada por:

$$f(x) = \frac{1}{\beta}\frac{\exp(-(x-\alpha)/\beta)}{[1+\exp(-(x-\alpha)/\beta))]^2} \quad F(x) = \frac{1}{1+\exp(-(x-\alpha)/\beta)}$$

Para relacionar la distribución logística con la distribución Poisson, vamos a considerar la variable $N(x) = \lfloor x+0{,}5 \rfloor$, donde$\cdot$ es la función parte entera. Como la distribución logística tiene soporte $(-\infty,\infty)$, nos restringiremos al dominio $[-1/2,\infty)$. Entonces la variable aleatoria N tiene función de distribución:

$$P(N=n) = \frac{1}{1+\exp(-(n+0{,}5-\alpha)/\beta)} - \frac{1}{1+\exp(-(n-0{,}5-\alpha)/\beta)},$$

si $x > 1/2$, y

$$P(N=n) = \left(\frac{1}{1+\exp(-(n+0{,}5-\alpha)/\beta)} - \frac{1}{1+\exp(-(n-0{,}5-\alpha)/\beta)}\right)\frac{1+\exp(-(0{,}5+\alpha)/\beta)}{\exp(-(0{,}5+\alpha)/\beta)}$$

si $-1/2 < x \leq 1/2$ y el ratio de densidades es:

$$\lambda^n / P(N=n)e^{\lambda} n!$$

Para optimizar (α, β) Atkinson (1979) propusó la elección $\alpha = \lambda$ y $\beta = \pi/\sqrt{3\lambda}$. Para esta elección de α y β, una optimización analítica de la cota de la expresión anterior es imposible, pero una maximización numérica y una interpolación arroja la cota $c = 0{,}767 - 3{,}36/\lambda$.

El algoritmo obtenido es:

A. 3: Simulación Poisson de Atkinson

1. Defina $\beta = \pi/\sqrt{3\lambda}$, $\alpha = \lambda\beta$ y $k = \log(c) - \lambda - \log(\beta)$
2. Generamos $U_1 \sim U[0,1]$ y calcular:

$$x = [\alpha - \log((1-u)/u)]/\beta$$

   hasta que $X > -0{,}5$
3. Definimos $N = \lfloor X + 0{,}5 \rfloor$ y generamos $U_2 \sim U[0,1]$-
4. Aceptar $N \sim P(\lambda)$ si

$$\alpha - \beta x + \log[u_2/(1+\exp(\alpha - \beta x))^2] \leq k + N\log(\lambda) - \log(N!)$$

Aunque la simulación es exacta, este algoritmo se basa en una serie de de aproximaciones, tanto en la elección de los parámetros (α, β) como en el cálculo de los límites de mayorización y los coeficientes de densidad. A pesar de esto el algoritmo tiene una eficacia razonable, sin embargo algoritmos mas complejos, como el de Devroye (1986) pueden ser preferibles.

II

Métodos de Montecarlo basados en cadenas de Markov

Simulación basadas en cadenas de Markov

En esta sección se introducirán los conceptos básicos sobre cadenas de Markov y sus consecuencias. La naturaleza y las propiedades de las cadenas de Markov permiten generar mecanismos para poder generar, de alguna manera, realizaciones cuya distribución será la que se desea muestrear.

0.7 Introducción a las cadenas de Markov

Definición 0.7.1 — Cadena de Markov. *Un proceso estocástico a tiempo discreto* $\{X_n; n \in \mathbb{N}\}$ *con espacio de estados* Ω *es una cadena de Markov, si para todo* $n \in \mathbb{N}$*, para todo* $A \in \Omega$ *y para todos* $x_1, \dots x_n \in \Omega$*, se tiene que:*

$$\mathbb{P}(X_{n+1} \in A | X_0 = x_0, X_1 = x_1, \dots X_n = x_n) = \mathbb{P}(X_{n+1} \in A | X_n = x_n) := P(X_n = x_n, A).$$

La medida de probabilidad P se denomina como kernel de transición. Informalmente si al tiempo $n+1$ se le considera como un tiempo futuro, n como el presente y $0, 1, \dots, n-1$ como el pasado entonces la distribución de probabilidad del estado del proceso depende únicamente del estado presente n y no del pasado $0, 1, \dots, n-1$.

■ **Ejemplo 0.7.1 — Cadenas de Markov con espacio de estados discreto .** Si Ω un subconjunto finito de $\mathbb{N}$ entonces podemos definir las probabilidades de transición del estado i al estado j en el tiempo $n+1$ de la siguiente manera:

$$P(i, j) = \mathbb{P}(X_{n+1} = j | X_n = i).$$

Estas probabilidades se conocen como probabilidades de transición a un paso.

■ **Ejemplo 0.7.2 — Paseo aleatorio.** Un paseo aleatorio X_n es un proceso estocástico a tiempo discreto donde, en el n-ésimo instante el proceso puede aumentar en 1 unidad con probabilidad p, o

disminuir en 1 unidad con probabilidad $1-p$. Entonces:

$$\mathbb{P}(X_{n+1}=j|X_n=i)=\begin{cases} p & j=i+1, \\ 1-p & j=i-1, \\ 0 & \text{en otro caso.} \end{cases}$$

Notemos que dado una trayectoria del proceso hasta el instante n, el proceso en el instante $n+1$ solo dependerá del estado mas reciente (X_n) y no de los anteriores.

■ **Ejemplo 0.7.3 — Autorregresivo de orden 1.** Una proceso autorregresivo de orden 1 (o $AR(1)$) X_n es un proceso estocástico a tiempo discreto que toma valores en $\mathbb{R}$, y que está dado por

$$X_n=\phi X_{n-1}+\sigma\varepsilon_n$$

donde $\varepsilon_n \sim N(0,1)$

Si la medida de probabilidad P no depende del instante n, entonces se dice que la cadena de Markov es homogénea. Para suavizar la notación denotamos $P(X_n=x,A)=P(x,A)$.

Para el caso homogéneo se denota $p_{ij}=P(X_{n+1}=j|X_n=i)$. Variando los índices i y j sobre el conjunto de estados, podemos obtener una matriz P en donde la entrada (i,j) es la probabilidad de pasar del estado i al estado j.

A pesar de que nos centraremos en el caso de cadenas de Markov en donde el espacio de estados es un subespacio continuo de $\mathbb{R}^n$, varias de las propiedades de las Cadenas de Markov con espacio de estados discretos se pueden extrapolar para el caso continuo, por lo cual se recomienda al lector tener siempre este caso en cuenta.

En general, se puede considerar que una cadena de Markov inicia su evolución partiendo de un estado x_0 cualquiera, o más generalmente podemos asumir que el punto de partida es aleatorio, por lo cual tendrá su propia distribución de probabilidad Π. Tal que $\Pi(A)$ es la probabilidad de empezar en un punto dentro del conjunto A. En el caso que exista π tal que para todo A:

$$\Pi(A)=\int_A \pi(x)dx, \quad A\subseteq\Omega$$

Se dice que π es la densidad de probabilidad inicial.

Sea $x\in\Omega$ el estado de la cadena en un tiempo arbitrario y sea A un conjunto en el espacio de estado, sabemos que la probabilidad de pasar del punto x a A esta denotada por $P(x,A)$. Sera de relevancia el caso en donde para cada $x\in\Omega$ existe un $p(x,\dot{})$ tal que para todo $A\subseteq\Omega$ tenemos que:

$$\int_A p(x,y)dy=\mathbb{P}(x,A)$$

Y llamaremos a $p(\cdot,\cdot)$ densidad de transición a 1 paso. Estos dos conceptos se pueden generalizar para un numero arbitrario de pasos. Sea $n\in\mathbb{N}$, $x\in\Omega$ y $A\subseteq\Omega$, denotamos por:

$$P^n(x,A)=\mathbb{P}(X_n\in A|X_0=x_0)$$

la distribución de X_n dado $X_0=x$, lo cual se conoce por kernel de transición a $n-$pasos. Igualmente es de interés el caso en el cual $\forall x$ en Ω existe $p^n(x,\cdot)$ tal que para todo Ω tenemos que:

$$\int_A p^n(x,A)=P^n(x,A)$$

y llamaremos a $p^n(\cdot,\cdot)$ la densidad de transición a n pasos. Nos gustaría tener una forma sencilla de calcular tanto P^n como p^n, una idea para lograrlo es pensar en las transiciones a n pasos como trayectorias que parten en el punto x y terminan en el punto y, las cuales se pueden descomponer en n pasos.

Motivados por esta idea es que surge la siguiente ecuación:

Proposición 0.7.4 — Ecuación de Chapman-Kolmogorov. *Para cualquier par de números enteros r y n tales que $0 \leq r \leq n$, para cualquier par de puntos $x, y \in \Omega$ y para cualquier partición B_i del espacio Ω se cumplen:*

$$p^n(x,y) = \int_\Omega p^r(x,k)p^{n-r}(k,y)dk$$

Para el caso discreto la ecuacion de Chapman-Kolmogorov se escribe de la siguiente manera:

$$p_{ij}(n) = \sum_k p_{ik}(r)p_{kj}(n-r)$$

Notar entonces que podemos calcular $p^n(x,\cdot)$ si conocemos $p_n(x,\cdot)$. Además si conocemos la distribución inicial de la cadena entonces podemos calcular la probabilidad de que en el tiem po n la cadena este en el conjunto A de la siguiente manera:

$$\mathbb{P}(X_n \in A) = \int_A \int_\Omega p^n(x,A)\pi(x)dx$$

Como p^n se puede obtener si conocemos p entonces la medida de probabilidad de P solo depende de p y de π, por lo cual estas dos funciones nos determinan el comportamiento de la cadena de Markov.

Notar que de esta manera podemos estudiar las trayectorias que puede seguir una cadena de Markov, sin embargo un tipo de trayectoria que nos interesa es la trayectoria constante, ie cuando existe un punto fijo en el cual cuando la cadena llega a este punto se mantiene constante. Como las trayectorias son estocásticas no es buena idea pensar en encontrar un punto fijo, pero si podemos extrapolar esta idea buscando un estado en donde las distribuciones iniciales no cambien al momento de evaluar la distribución algún paso hacia el futuro. Esto nos motiva definir las distribuciones invariantes.

0.7.1 Distribución invariante

Definición 0.7.2 — Distribución invariante. *Una distribución de probabilidad inicial π para un cadena de Markov $\{X_n : n \in \mathbb{N}\}$ con kernel de transición es una distribución invariante, si y solo si para todo $A \subseteq \Omega$:*

$$\int \pi(x)P(x,A)dx = \int_A \pi(x)dx := \Pi(A).$$

Dado que como nos interesa el caso cuando existe densidad de transición entonces tendremos:

$$\int \int_A \pi(x)p(x,y)dydx = \int_A \pi(x)$$
$$\int_A \int \pi(x)p(x,y)dxdy = \int_A \pi(x)dx \quad \text{(Por Fubini-Tonelli)}$$

Por lo cual, teniendo nuestra densidad de transición nos bastaría resolver la ecuación para encontrar la distribución invariante. Debido a la complejidad de resolver una ecuación diferencial parcial de forma

explicita, nos interesan los casos en donde $p(\cdot,\cdot)$ cumpla alguna propiedad que nos haga mas fácil el calculo.

Para el primer, veamos el comportamiento asintótico de la cadena, en especial no centraremos en el caso de que la función p^n converge a alguna función cuando n tiende a infinito, lo cual es el símil de converger a un punto fijo, en contextos no estocásticos. Esto nos lleva a la siguiente proposición.

Proposición 0.7.5 *Sea Ω espacio de medida finita. Supongamos que para algún $x \in \Omega$ existe π tal que:*

$$p^n(x,y) \to \pi(y) \quad \text{cuando} \quad n \to \infty \,, \forall y \in \Omega$$

Entonces π es una distribución invariante.

Demostración: Tenemos que:

$$\int_\Omega \lim_{n\to\infty} p^n(x,y)dy = \int_\Omega \pi(y)dy = 1 = \lim_{n\to\infty}\int_\Omega p^n(x,y)dy$$

y entonces:

$$\pi(y) = \lim_{n\to\infty} p^n(x,y) = \lim_{n\to\infty}\int_\Omega p^n(x,z)p(z,y)dz = \int \lim_{n\to\infty} p^n(x,z)p(z,y)dz = \int \pi(z)p(z,y)dz$$

Otra propiedad importante sucede cuando la densidad de partir del punto x y llegar al punto y es la misma que la de partir en el punto y y llegar al punto x, lo cual nos motiva a la siguiente definición:

Definición 0.7.3 — DBC: Condición de detalle de balance. *Sea π densidad inicial y sea p la densidad de transición, diremos que la cadena satisface la condicion de detalles de balance o DBC por sus siglas en ingles, si para todos los estado $x,y \in \Omega$, tenemos que:*

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

Si para una cadena de markov, la distribución π satisface el *DBC* entonces se puede demostrar que es una distribución invariante. Las cadenas de Markov que cumplen el *DBC* también cumplen otra propiedad, la reversibilidad, que estudiaremos a continuación.

0.7.2 Reversibilidad

Sea $\{X_n : n \geq 0\}$ una cadena de Markov con kernel de transición P, m un entero fijo, y sea $Y_n = X_{m-n}$ un nuevo proceso para $n = 0,\ldots,m$, ie $\{Y_n : n = 0,\ldots,m\}$ es la cadena original invertida en el tiempo, ahora del tiempo m al tiempo 0. Veamos si este proceso es una cadena de Markov.

Sea $A \in \Omega$ y $y_1, y_2 \ldots y_n \in \Omega$ arbitrario, además asumamos que el kernel de transición P se le asocia una densidad de transición p, definida por $p(X_n = x_n)$ la densidad de probabilidad de la variable X_n, entonces tenemos que:

$$p(Y_n = y_n | Y_{n-1} = y_{n-1}, \ldots Y_1 = y_1, Y_0 = y_0) = p(X_{m-n} = x_{m-n} | X_{m-n+1} = x_{m-n+1} \ldots X_m = x_m).$$

Desarrollando el termino de la derecha tenemos que:

$$
\begin{aligned}
&= p(X_{m-n} = x_{m-n} | X_{m-n+1} = x_{m-n+1} \dots X_m = x_m) \\
&= \frac{p(X_{m-n} = x_{m-n}, X_{m-n+1} = x_{m-n+1} \dots X_m = x_m)}{p(X_{m-n+1} = x_{m-n+1} \dots X_m = x_m)} \\
&= \frac{p(X_m = x_m | X_{m-1} = x_{m-1}) \cdots p(X_{m-n+2} = x_{m-n+2} | X_{m-n+1} = x_{m-n+1})}{p(X_m = x_m | X_{m-1} = x_{m-1}) \cdots p(X_{m-n+2} = x_{m-n+2} | X_{m-n+1} = x_{m-n+1})} \\
&\cdot \frac{p(X_{m-n+1} = x_{m-n+1} | X_{m-n} = x_{m-n}) p(X_{m-n} = x_{m-n})}{p(X_{m-n+1} = x_{m-n+1})} \\
&= \frac{p(X_{m-n+1} = x_{m-n+1} | X_{m-n} = x_{m-n}) p(X_{m-n} = x_{m-n})}{p(X_{m-n+1} = x_{m-n+1})} \\
&= p(X_{m-n} = x_{m-n} | X_{m-n+1} = x_{m-n+1}) \\
&= p(Y_n = y_n | Y_{n-1} = y_{n-1})
\end{aligned}
$$

Demostrando así que el proceso cumple la propiedad de Markov. Notar que no es necesaria la homogeneidad de la cadena original $\{X_n : n \in \mathbb{N}\}$ para que la nueva cadena sea efectivamente una cadena de Markov. Si ademas se asume que la cadena original es homogénea sería de interés el ver si la nueva cadena es también homogénea, sin embargo, esto no siempre es cierto, para esto sean $z_1, z_2 \in \Omega$ arbitrarios tenemos:

$$
\begin{aligned}
p(Y_{n+1} = z_2 | Y_n = z_1) &= \frac{p(X_{m-n} = z_1, X_{m-n+1} = z_2)}{p(X_{m-n} = z_1)} \\
&= p(X_{m-n} = z_1 | X_{m-n} = z_2) \frac{p(X_{m-n-1} = z_2)}{p(X_{m-n} = z_1)} \\
&= p_X(z_1, z_2) \frac{p(Y_{n+1} = z_2)}{p(Y_{n+1} = z_1)},
\end{aligned}
$$

con $p_X(\cdot, \cdot)$ la probabilidad de transición de la cadena original. Por lo tanto, tenemos que la probabilidad de transición depende del tiempo a través del cociente $p(Y_{n+1} = z_2)/p(Y_{n+1} = z_1)$. Para solucionar este problema tomaremos como hipótesis la existencia de una distribución inicial invariante π. Por lo tanto tendremos que,

$$
p(Y_{n+1} = z_2 | Y_n = z_1) := p_Y(z_1, z_2) = p_X(z_2, z_1) \frac{\pi(z_2)}{\pi(z_1)}
$$

Teniendo así la homogeneidad. Por ultimo una pregunta importante es el caso en donde la cadena de Markov original tiene la misma densidad de transición que la cadena de Markov invertida, para eso necesitamos que $p_X = p_Y$, para eso sean $z_1, z_2 \in \Omega$ arbitrarios, buscamos que:

$$
\begin{aligned}
p_Y(z_1, z_2) &= p_X(z_1, z_2) \\
p_X(z_2, z_1) \frac{\pi(z_2)}{\pi(z_1)} &= p_X(z_1, z_2) \\
p_X(z_2, z_1) \pi(z_2) &= \pi(z_1) p_X(z_1, z_2)
\end{aligned}
$$

Lo cual es condicion DBC. Luego se debe notar que si tenemos esta propiedad obtenemos la invariante de π. Esto nos motiva a la siguiente definición.

Definición 0.7.4 — Reversibilidad. *Una cadena de Markov $\{X_n : n \in \mathbb{N}\}$ homogénea con densidad de transición $p(\cdot,\cdot)$ y distribución inicial π, es reversible si para todo $x, y \in \Omega$ tenemos que:*

$$\pi(x)p(x,y) = \pi(y)p(y,x).$$

Además la distribución π es una distribución invariante de π.

0.7.3 Irreducibilidad

En algunos casos es posible descomponer nuestra cadena de Markov en ciertos grupos o clases incomunicados entre si, ie, a veces existen conjuntos B_1, B_2 con $\Pi(B_1), \Pi(B_2) \geq 0$ del espacio Ω tales que $P^n(X_1 \in B_1 | X_2 \in B_2) = 0$ para todo $n \in \mathbb{N}$ en cuyo caso tenemos que de empezar en un conjunto es imposible llegar al otro. Sin embargo estaremos interesados en el caso en donde, independientemente del punto inicial, podamos llegar siempre a cualquier conjunto del espacio de estados en un numero finito de pasos. Esta propiedad se conoce como irreducibilidad.

Definición 0.7.5 — Irreducibilidad. *Sea una cadena de Markov con distribución invariante Π. La cadena es irreducible si para todo $x \in \Omega$ y $A \subseteq \Omega$ con $\Pi(A) > 0$ existe un n tal que $P^n(x,A)) > 0$. Es decir, independiente del punto inicial, la cadena puede llegar, en un numero finito de pasos, a cualquier región A con $\Pi(A)) > 0$. Para enfatizar el rol de la distribución Π (o la densidad π) decimos que la cadena es Π- irreducible. Además, decimos que la cadena es Harris recurrente, si para todo A:*

$$\mathbb{P}(X_n \in A \text{ para un numero infinito de n} | X_0 = x) = 1$$

Se puede demostrar que la irreductibilidad implica la unicidad de la distribución invariante.

El concepto de la recurrencia Harris implica la irreductibilidad, sin embargo el reciproco no siempre es verdadero.

Notemos que a pesar de tener una cadena $\Pi-$irreducible aun existe la posibilidad de que dos conjuntos B_1, B_2 con $\Pi(B_1), \Pi(B_2) > 0$ tales que estos dos conjuntos queden incomunicados, ie $P(X_2 \in B_2 | X_1 \in B_1) = 0$, pues puede ser que la cadena oscile entre varios conjuntos.

Definición 0.7.6 *Una cadena de Markov $\Pi-$irreducible es periódica si existe una partición del espacio $\Omega = \bigcup_{i=0}^{n} A_i$ con $n \geq 2$ y los A_i conjuntos disjuntos tal que $\Pi(A_n) = 0$ y*

$$x \in A_0 \Rightarrow \mathbb{P}(x,A_1) = 1, x \in A_1 \Rightarrow \mathbb{P}(x,A_2) = 1,$$
$$\ldots, x \in A_{n-2} \Rightarrow \mathbb{P}(x,A_{n-2}) = 1,\ x \in A_{n-1} \Rightarrow \mathbb{P}(x,A_0) = 1$$

En otro caso la cadena se dice que es aperiódica.

Ya teniendo todas estas propiedades podemos introducir el siguiente teorema:

Teorema 0.7.6 — Teorema de Convergencia para cadenas de Markov. *Para una cadena de Markov, $\Pi-$irreducible y aperiódica, donde Π es una distribución invariante, entonces existe un conjunto $C \subseteq \Omega$ tales que $\Pi(C) = 1$ y para todo $x \in C$ y $A \subseteq \Omega$.*

$$\mathbb{P}(X_n \in A | X_0 = x) \to \Pi(A), \quad n \to \infty$$

Gracias a este teorema, no nos importara la distribución inicial de la cadena de Markov, pues sabemos que cumpliendo esas propiedades convergeremos a la distribución deseada.

0.7.4 Teorema ergódico

Definición 0.7.7 *Una función medible h es armonica para la cadena X_n si:*

$$\mathbb{E}[h(X_{n+1})|x_n] = h(x_n)$$

Estas funciones son invariantes para el kernel de transición y caracterizan la recurrencia Harris:

Proposición 0.7.7 *Sea una cadena de Markov X_n invariante, con una distribución ϕ-irreducible. Entonces la cadena X_n es Harris recurrente si y solo si las unicas funciones harmonicas de X_n son las funciones constantes.*

Demostración: Ver Meyn y Tweedie (2012)

Lo cual nos lleva a el teorema ergodico para cadenas de Markov:

Teorema 0.7.8 *Si X_n tiene una distribución invariante π entonces los dos siguientes teoremas son equivalentes:*

1. *Si $f, g \in L^1$ con $\int g(x)\mathrm{d}\pi(\mathrm{x}) \neq 0$, entonces:*

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} f(X_i)}{\frac{1}{n}\sum_{i=1}^{n} g(X_i)} = \frac{\int f(x)\mathrm{d}\pi(\mathrm{x})}{\int g(x)\mathrm{d}\pi(\mathrm{x})}$$

2. *La cadena de Markov X_n es Harris recurrente.*

Ya teniendo estas propiedades, podemos pasar al teorema que da fundamento a los algoritmos que veremos en esta sección.

0.8 Algoritmo de Metropolis-Hastings

Gracias a la sección 0.7, sabemos que si tenemos una cadena de Markov con distribución estacionaria f, entonces podemos usar esta cadena de Markov para generar una muestra $X_1, X_2, \cdots \sim f(x)$, a través del siguiente mecanismo: Generamos un valor inicial arbitrario y luego se genera una cadena de Markov, $\{X_t, t \in \mathbb{N}\}$ con un kernel de transición con densidad estacionaria f, lo cual garantiza la convergencia en distribución de X_t a f. Motivando la siguiente definición:

Definición 0.8.1 — MCMC: Monte Carlo basado en cadenas de Markov. Un método de Monte Carlo basado en cadena de Markov o *MCMC* por las siglas de Markov chain Monte Carlo, es un método que se basa en generar una cadena de Markov $\{X_t : t \in \mathbb{N}\}$ cuya distribución estacionaria f sea nuestra densidad objetivo.

Con este principio, si debemos simular de una densidad objetivo f solo es necesario encontrar un kernel de transición adecuado y generar una cadena de Markov. Cabe destacar, que este kernel de transición dependerá de la distribución objetivo. La elección de los distintos kernels de transición posibles nos llevará a algoritmos distintos. Sin embargo, nos enfocaremos en el algoritmo de Metropolis-Hastings y sus variantes.

0.8.1 Algoritmo Metropolis-Hastings

Para introducir el algoritmo Metropolis-Hastings asumamos que la densidad objetivo f, que deseamos simular es conocida. A continuación, se elige una densidad de transición $g(x,y)$, tal que $g(x,\cdot)$ sea fácil de simular y conozcamos su forma explicita. La densidad objetivo f debe ser tal que el radio:

$$f(y)/g(x,y)$$

sea posible de calcular salvo constantes multiplicativas que no dependan de x.

Dado entonces $f(\cdot)$ y $q(\cdot,\cdot)$ con las propiedades ya mencionadas definiremos la probabilidad de aceptación de la siguiente manera:

$$a(x,y) = \text{mín}\left\{1, \frac{f(y)g(y,x)}{f(x)g(x,y)}\right\}, \tag{2}$$

donde por convención $a(x,y) = 1$ cuando $f(x)q(x,y) = 0$. Lo cual nos permite definir el algoritmo Metropolis-Hastings usando el siguiente procedimiento :

A. 4: Algoritmo Metropolis-Hastings

Dado $X_t = x_t$

1. Generar $Y_t \sim g(x_t, y_t)$
2. Generar $U_t \sim U(0,1)$
3. Definir:

$$X_{t+1} = \begin{cases} y_t & \text{Si } u < a(x_t, y_t) \\ x_t & \text{Si } u \geq a(x_t, y_t) \end{cases}$$

La distribución g se le denomina distribución instrumental y a la probabilidad $a(x,y)$ como probabilidad de aceptación del algoritmo Metropolis-Hastings (o simplemente probabilidad de aceptación).

Notar que el algoritmo siempre aceptará valores y_t tales que el radio $f(y_t)/g(x_t,y_t)$ incrementa con respecto al valor $f(x_t)/g(y_t,x_t)$. Solo en el caso simétrico, i.e. $g(x,y) = g(y,x)$, el radio de aceptación depende del radio objetivo $f(y_t)/f(x_t)$, notamos ademas que el algoritmo también aceptará valores y_t tales que el radio anterior decrezca.

Igual que en el caso de el algoritmo Aceptación-Rechazo el algoritmo Metropolis-Hastings depende solo de los ratios:

$$f(y_t)/f(x_t) \quad \text{y} \quad g(y_t,x_t)/g(x_t,y_t)$$

y por lo tanto, es independiente de constantes normalizadoras, cuando estas no dependen de y o x.

El algoritmo Metropolis-Hasting es un algoritmo genérico, definido para todo f y g, por lo que es necesario imponer condiciones para asegurar que la cadena de Markov producida por el algoritmo converja a una secuencia cuya distribución sea f, y a la distribución instrumental g para que f sea la distribución límite de la cadena de Markov producida por el algoritmo.

En primera instancia, impondremos condiciones sobre el soporte de la función f, digamos $supp(f)$, en particular impondremos que todas las componentes conexas de $supp(f)$ están conectadas con todas las demás a través del kernel de transición. Diremos que el $supp(f)$ está truncado por g si existe $A \subset supp(f)$ tal que:

$$\int_A f(y)dy > 0 \quad \text{y} \quad \int_A g(x,y)dy = 0, \quad \forall x \in supp(f).$$

En tal caso tendremos que el algoritmo no tendrá a f como su distribución límite, pues para $x_0 \notin A$, la cadena X_t no visitará A. De esto se deriva que, una condición necesaria para que f sea una densidad invariante de la cadena de Markov, es que:

$$supp(f) \subset \bigcup_{x \in supp(f)} supp(g(x,\cdot))$$

Otra condición que se suele imponer es la de DBC (Definición 0.7.3). Esto debido a que nos asegura que la convergencia de la cadena de Markov. En la sección 0.7.2 vimos que si la cadena satisface la condición DBC entonces es una densidad invariante, por lo cual procederemos a verificar las condiciones sobre g y f.

Teorema 0.8.1 *Sea $\{X_t : t \in \mathbb{N}\}$ la cadena generada por el algoritmo Metropolis-Hastings. Entonces para toda distribución instrumental g cuyo soporte incluye ε se cumple que:*

1. *El kernel de transición satisface el DBC para la densidad f*
2. *f es la distribución estacionaria de la cadena.*

Demostración: Para la propiedad 1. para demostrar la primera consecuencia, notamos que el kernel de transición asociado al algoritmo de Metropolis-Hastings es:

$$p(x,y) = a(x,y)g(x,y) + (1 - r(x))\delta_x(y)$$

En donde $r(x) = \int a(x,y)g(x,y)dy$ y δ_x denota la función delta de Dirac en x. Lo que nos permite verificar que:

$$a(x,y)g(x,y)f(x) = a(y,x)g(,y,x)f(y)$$

y

$$(1 - r(x))\delta_x(y)f(x) = (1 - r(y))\delta_y(x)f(y)$$

Demostrando así que la cadena cumple la condición de DBC.

La propiedad 2. es consecuencia directa de la propiedad 1.

0.8.2 Convergencia

Por construcción se sabe que el algoritmo Metropolis-Hastings tiene una distribución invariante f. Ahora nos falta demostrar que, independiente del punto inicial, la cadena converge a la densidad objetivo. Notar que si la cadena de Markov es aperiódica y Harris recurrente entonces la convergencia la garantiza el teorema de Convergencia para las cadenas de Markov (teorema 0.7.6.

Para que la cadena sea aperiódica es suficiente que el algoritmo permita eventos de la forma $\{X_{t+1} = X_t\}$, es decir:

$$\mathbb{P}(f(X_t)g(X_t,Y_t) \leq f(Y_t)g(Y_t,X_t)) < 1 \tag{3}$$

Por otro lado, para obtener la propiedad de la irreducibilidad de la cadena generada por el Metropolis-Hastings $\{X_t : t \in \mathbb{N}\}$ se deben asumir condiciones adicionales.

Primeramente la densidad instrumental g debe ser positiva, ie,

$$g(x,y) > 0 \quad \text{para todo } (x,y) \in \varepsilon \times \varepsilon, \tag{4}$$

es decir, que desde cualquier punto $x \in \varepsilon$ podemos llegar con 1 paso a cualquier conjunto $A \subseteq \varepsilon$, si $\nu(A) > 0$ (con ν la medida de Lebesgue). Luego, como f es la distribución invariante, la cadena es positiva. Por lo cual solo nos faltaría demostrar la recurrencia Harris, la cual podemos obtener del siguiente resultado.

Lema 0.8.2 *Si la cadena generada por el algoritmo Metropolis-Hastings $\{X_t : t \in \mathbb{N}\}$ es $f-$irreducible, entonces es Harris recurrente.*

Demostración: Utilizaremos el hecho de que si las únicas funciones armónicas acotadas son las funciones constantes, entonces la cadena es Harris recurrente.

Si h es una función harmónica, entonces satisface:

$$h(x_0) = E[h(X_1)|x_0] = E[h(X_t)|x_0].$$

Como la cadena de Metropolis-Hastings es positiva y aperiódica, entonces podemos usar el teorema (citar Proposición), y podemos concluir que h es f constante en casi todas partes, e igual a $E_f[h(X)]$. Para mostrar que h es constante en casi todas partes, veamos que:

$$E[h(X_1)|x_0] = \int a(x_0,x_1)g(x_0,x_1)h(x_1)dx + (1-r(x_0))h(x_0)$$

y sustituyendo $h(x_1)$ por $E[h(X)]$ en la integral. Teniendo que:

$$E_f[h(X)]r(x_0) + (1-r(x_0))h(x_0) = h(x_0)$$

teniendo que $(h(x_0) - E[h(X)])r(x_0) = 0$ para todo $x_0 \in \varepsilon$. Como $r(x_0) > 0$ para todo $x_0 \in \varepsilon$, luego por la $f-$irreductibilidad tenemos que h es constante y entonces la cadena es Harris recurrente.

Por lo tanto, tenemos el siguiente resultado de convergencia para cadenas de Markov generadas por Metropolis-Hastings.

Teorema 0.8.3 *Si la cadena de Markov generada por el Metropolis-Hastings ($\{X_t : t \in \mathbb{N}\}$) es f-irreducible, entonces:*

1. *Si $h \in L^1(f)$ entonces:*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X_t) = \int h(x)f(x)dx, \quad \text{casi seguramente.}$$

2. *Si además $\{X_t : t \in \mathbb{N}\}$ es aperiódica, entonces:*

$$\lim_{n\to\infty} \left|\left| \int K^n(x,\cdot)\mu(dx) - f \right|\right|_{TV} = 0$$

para toda distribución inicial μ, donde $K^n(x,\cdot)$ denota el Kernel de transición de n y $||\cdot||_{TV}$ denota la norma de variación total definida por $||\mu||_{TV} = \sup_A ||\mu(A)||$.

Notar que tanto para la $f-$irreductibilidad, y la positividad de la densidad instrumental g pueden obtenerse de las condiciones 3 y 4. Obteniendo así el siguiente corolario.

Corolario 0.8.4 *Las conclusiones del teorema 0.8.3 se mantienen si la cadena de Markov generada por el Metropolis-Hastings $\{X_t : t \in \mathbb{N}\}$ tiene una densidad de transición g que satisface 3 y 4*

0.8.3 Metropolis-Hastings Independiente

Este submétodo surge de considerar la distribución instrumental g independiente de X_t, es decir para todo $x_1, x_2, y \in \mathbb{R}$ $g(x_1,y) = g(x_2,y)$, por lo cual ahora definiremos $g(y) = g(x,y)$. Teniendo así el siguiente algoritmo:

A. 5: Algoritmo Metropolis-Hastings Independiente

Dado $X_t = x_t$

1. Generar $Y_t \sim g(y)$
2. Generar $U_t \sim U(0,1)$
3. Definir:

$$X_{t+1} = \begin{cases} Y_t & \text{Si } u < a(x_t,Y_t) \\ x_t & \text{Si } u \geq a(x_t,Y_t) \end{cases}$$

con

$$a(x,y) = \text{mín}\left\{\frac{f(Y_t)g(x_t)}{f(x_t)g(Y_t)}, 1\right\}$$

Notemos que a pesar de que la v.a Y_t es independiente de X_t la probabilidad de aceptar la variable Y_t depende de X_t.

La convergencia de la cadena $\{X_t : t \in \mathbb{N}\}$ vienen otorgadas por la densidad g, pues si g es positiva c.t.p en el soporte de f entonces la cadena es irreducible y aperiódica, teniendo así la convergencia por el Corolario 0.8.4. Sin embargo este submetodo posee otra propiedad mas fuerte de convergencia, que sigue del siguiente resultado de Mengersen y Tweedie (1996):

Teorema 0.8.5 *Si existe una constante M tal que:*

$$f(x) \leq Mg(x), \quad \forall x \in supp(f),$$

entonces:

$$||K^n(x,\cdot) - f||_{TV} \leq 2\left(1 - \frac{1}{M}\right)^n.$$

Si las funciones f, g cumplen las hipótesis del lema anterior, entonces también se puede usar un algoritmo de Aceptación-Rechazo para simular f, por lo cual es natural comparar estos dos algoritmos. Una ventaja que presenta el algoritmo Metropolis-Hastings es que la probabilidad de aceptación esperada es mayor que en el algoritmo Aceptación-Rechazo.

Proposición 0.8.6 *Si la condiciones del teorema 0.8.5 se cumplen, entonces la probabilidad de aceptación del algoritmo Metropolis-Hastings es al menos* $1/M$ *cuando la cadena es estacionaria.*

Demostración: Si la distribución dada por $f(X)/g(X)$ es absolutamente continua, entonces la probabilidad de aceptación es:

$$\begin{aligned}
E\left[\text{mín}\left\{\frac{f(Y_t)g(X_t)}{f(X_t)g(Y_t)}, 1\right\}\right] &= \int \mathbb{I}_{\frac{f(y)g(x)}{g(y)f(x)}>1} f(x)g(y)dxdy + \int \frac{f(y)g(x)}{g(y)f(x)} \mathbb{I}_{\frac{f(y)g(x)}{g(y)f(x)}\leq 1} f(x)g(y)dxdy \\
&= 2\int \mathbb{I}_{\frac{f(y)g(x)}{g(y)f(x)}\geq 1} f(x)g(y)dxdy \\
&\geq 2\int \mathbb{I}_{\frac{f(y)}{g(y)}\geq\frac{f(x)}{g(x)}} f(x)\frac{f(y)}{M}dxdy \\
&= \frac{2}{M}P\left(\frac{f(X_1)}{g(X_1)} \geq \frac{f(X_2)}{g(X_2)}\right) = \frac{1}{M}.
\end{aligned}$$

Como X_1 y X_2 son independientes y distribuyen con respecto a f, teniendo así que la ultima probabilidad es igual a $1/2$.

Tenemos así que el algoritmo de Metropolis-Hastings independiente es mas eficiente que el algoritmo Aceptación-Rechazo, pues para la misma densidad instrumental g se aceptan , en promedio, mas propuestas.

0.8.4 Metropolis-Hastings con Caminatas Aleatorias

Un enfoque para la construcción de un algoritmo Metropolis-Hastings es utilizar el valor simulado anteriormente para generar el valor siguiente, es decir, se explota una vecindad del valor actual de la cadena de Markov.

Como el candidato a g en el algoritmo Metropolis-Hastings, puede depender del valor actual X_t entonces una opción es definir Y_t de la siguiente manera:

$$Y_t = X_t + \varepsilon_t$$

donde los ε_t se considera un paso aleatorio con distribución g e independiente de X_t. Entonces tendremos que la densidad instrumental $g(x,y)$ es ahora de la forma $g(y-x)$. La cadena de Markov asociada a q es entonces la caminata aleatoria.

Notar que podemos usar los mismos resultados de convergencia de la sección 0.8.2, por lo cual nos basta que g sea positiva para obtener que la cadena es $f-$irreducible y aperiódica, teniendo así la convergencia. La elección mas común para la distribución g son distribuciones uniformes centradas en un intervalo simétrico en el 0, o distribuciones estándar como la normal o la T de Student. Notar que en todas las opciones anteriores solo consideramos distribuciones simétricas (ie $g(t) = g(-t)$), lo cual sigue de la expresión original del algoritmo propuesto por Metropolis et al. (1953).

A. 6: Algoritmo Metropolis-Hastings con paseos aleatorios

Dado $X_t = x_t$

1. Generar $Y_t \sim g(|y - x_t|)$
2. Generar $U_t \sim U(0,1)$
3. Definir:

$$X_{t+1} = \begin{cases} Y_t & \text{Si } u < a(x_t, Y_t) \\ x_t & \text{Si } u \geq a(x_t, Y_t) \end{cases}$$

con

$$a(x,y) = \text{mín}\left\{\frac{f(Y_t)}{f(x_t)}, 1\right\}$$

0.8.5 Ley Fuerte de los Grandes Números

Teorema 0.8.7 — Ley fuerte de los grandes números para Cadenas de Markov. *Sea* $\{X_n : n \in \mathbb{N}\}$ *sea una cadena de Markov* $\Pi-$*irreducible, y sea* $h : \Omega \to \mathbb{R}$ *sea una función tal que la media* $\theta = \int h(x)\pi(x)dx$ *existe. Para un entero arbitrario* $m \geq 0$ *se define el promedio ergódico como:*

$$\hat{\theta}_n = \frac{1}{n+1}\sum_{i=m}^{m+n} h(X_i) \tag{5}$$

Entonces existe un conjunto $C \subseteq \Omega$ *tal que* $\Pi(C) = 1$ *y para todo* $x \in C$,

$$P(\hat{\theta}_n \to \theta \text{ cuando } n \to \infty | X_0 = x) = 1 \tag{6}$$

Además, si la cadena es Harris recurrente, podemos tomar $C = \Omega$.

La recurrencia Harris nos da la consistencia para todos los estados iniciales posibles $x \in \Omega$, por lo cual no hay que preocuparse de las condiciones iniciales. Sin embargo, dado que la recurrencia Harris es mas difícil de probar, nos centraremos en la irreductibilidad en la cual, gracias al teorema de convergencia, el estimador $\hat{\theta}_n$ es consistente para todos los estados iniciales $x \in C$.

Una pregunta importante es la elección de m en el teorema anterior, para eso escogeremos el m conocido como el "burn in", que es el momento cuando la cadena esta en equilibrio, ie m es lo suficientemente grande para que la distribución X_m es bastante cercana a π.

0.9 Aproximación de Langevine ajustada por Metrópolis

La clase de algoritmos Metropolis-Hastings es muy amplia, pues, dada una densidad objetivo f tenemos infinitas distribuciones g que podemos usar como distribuciones instrumentales para la

simulación. Si descartamos las densidades g que no son simulables, todavía nos quedan muchas densidades que podemos utilizar. Teniendo así un problema, pues no tenemos una forma de escoger una densidad instrumental g de un conjunto de distribuciones instrumentales posibles.

Una distribución clásica es la normal cuya media depende de (x,y) y varianza constante. La elección de la varianza es de suma importancia para la convergencia del algoritmo, pues una varianza pequeña provocará que se den pasos muy pequeños, y por lo tanto el algoritmo se demorará en converger a la densidad objetivo, pues le cuesta variar mucho de su posición actual. Por otro lado una varianza muy grande provocará que se rechacen muchas simulaciones, por lo cual se requerirá de un gran número de simulaciones para poder generar la muestra.

Por otra parte, en la sección 0.8.2 vimos que al imponer condiciones sobre la densidad objetivo, obtuvimos una ventaja considerable: logramos demostrar que el algoritmo es geométricamente ergódico. Esto nos da una convergencia mucho más rápida a la densidad objetivo. Lo cual nos motiva a buscar condiciones sobre g que nos garanticen una convergencia mas rápida a la densidad objetivo f.

Por lo dicho anteriormente estudiaremos una variación de los algoritmos Metropolis-Hastings, el Metropolis-adjusted Langevine algorithm (MALA) y el Metropolis-adjusted Langevine truncated algorithm (MALTA), propuestos por Roberts y Tweedie (1996), los cuales utilizan información sobre la densidad objetivo (en la forma del gradiente de $\log f$) para construir una distribución de propuesta específica para el problema.

0.9.1 Ecuación de difusión de Langevine

Para introducir el algoritmo MALA, es necesario estudiar la Ecuación de difusión de Langevine, dada por la siguiente ecuación diferencial estocástica a tiempo continuo:

$$dL_t = \frac{1}{2}\nabla \log f(L_t)dt + dW_t, \tag{7}$$

en donde W_t es una movimiento Browniano estándar. Se puede demostrar que, bajo condiciones apropiadas de f, la distribución de L_t converge a f cuando t tiende a infinito. Por lo tanto, generar un algoritmo para resolver (7), nos puede servir para obtener un algoritmo que simule de la distribución f.

Una forma de solucionar esto es mediante la aproximación de primer orden de Euler, la cual se basa en realizar una discretización de (7), que se obtiene a través de aproximar:

$$\begin{aligned} dL_t &\approx L_{t+h} - L_t \\ dt &\approx h \\ dW_t &\approx W_{t+h} - W_t, \end{aligned}$$

obteniendo así lo siguiente:

$$L_{t+1} \sim N\left(L_t + h\frac{1}{2}\nabla \log f(L_t), hI_k\right). \tag{8}$$

Notamos que la Ecuación (8) soluciona varios problemas, puesto que logramos obtener una cadena de Markov en donde la distribución muestral es una normal, la cual es facil de simular, cuya varianza viene determinada por el paso h de la discretización, sin embargo, este método requiere ser estudiado con cautela puesto que tanto la convergencia de (7) como la convergencia de (8) dependeran de la funcion f, además la convergencia de (7) no necesariamente implicará la convergencia de (8). Esto será discutido a continuación.

0.9.2 Convergencia

Asumiremos que la distribución f es distinta de 0 en todo su dominio y, además que es diferenciable tal que $\nabla \log(f(x))$ esta siempre bien definido.

Como mencionamos anteriormente, esta ecuación converge a la distribución objetivo f gracias al siguiente teorema:

Teorema 0.9.1 *Si $\nabla \log f(x)$ es continuamente diferenciable y existen $N, a, b < \infty$, tales que:*

$$\nabla \log f(x) \cdot x \leq a|x|^2 + b, \quad |x| > N$$

Entonces la difusión L_t en (7), es aperiódica, f es una distribución invariante de L_t y además, para todo x tenemos que:

$$||\mathbb{P}_l^t(x,\cdot) - f|| \to 0 \tag{9}$$

La ecuación (9) justifica utilizar la ecuación de difusión de Langevine. La ventaja del método se basa en la rápida convergencia detallada a continuación:

Definición 0.9.1 — Ergodicidad Exponencial. *Sea X_t un proceso estocástico, diremos que el proceso es exponencialmente ergódico si existen una distribución π invariante, $\rho > 0$ y además para todo x existe un $M_x < \infty$ tal que:*

$$||\mathbb{P}_L^t(x,\cdot) - \pi|| \leq R_x \rho^t$$

A pesar de que bajo las condiciones del Teorema 0.9.1 el proceso estocástico siempre converge, no siempre tendremos la ergodicidad exponencial, como por ejemplo:

Teorema 0.9.2 *Si $|\nabla \log f(x)| \to 0$ entonces L_t no es exponencialmente ergódico.*

Sin embargo el siguiente teorema nos da condiciones suficientes (pero no necesarias) que nos aseguran la ergodicidad exponencial:

Teorema 0.9.3 *Supongamos que existe $S > 0$ tal que $|f(x)|$ es acotada para todo $|x| \geq S$. Entonces si existe un $0 < d < 1$ tal que:*

$$\liminf_{|x|\to\infty} (1-d)|\nabla \log f(x)|^2 + \nabla^2 \log f(x) > 0$$

entonces L_t es exponencialmente ergódico.

En la práctica, se implementa una aproximación discreta a la difusión L_t, por lo cual es necesario verificar si los resultados anteriores pueden extenderse para estas aproximaciones discretas.

0.9.3 Algoritmo desajustado de Langevine

El algoritmo desajustado de Langevine (o ULA, del inglés *unadjusted Langevine algorithm*) es una cadena de Markov a tiempo discreto que se obtiene a través de la discretización de la difusión de Langevine L_t. Para crear la cadena de Markov consideraremos un paso $h > 0$ fijo, y asumiremos que $f > 0$ y además que $|\nabla \log(f)| = \infty$ en un conjunto numerable de $\text{supp} f$. Con esto, se define la cadena de Markov ULA mediante la recursión $X_{t+1} \sim N(X_t + h\frac{1}{2}\nabla \log f(x_{t-1}), h_t I_k)$.

A. 7: Algoritmo desajustado de Langevine

Dado $X_t = x_t$ y $h > 0$ fijo

1. Generar $X_{t+1} \sim N(X_t + h\frac{1}{2}\nabla \log f(x_{t-1}), h_t I_k)$

Notemos que la cadena ULA no tiene necesariamente a f como su distribución invariante. Por ejemplo, si $f \sim N(0,1)$, y tomando $h = 2$ tendremos que $X_n \sim N(0,2)$, teniendo así que la cadena es

estacionaria, pero convergerá a una densidad distinta de la densidad objetivo. Esto ocurre pues ULA es sólo una aproximación, sin embargo al ser la discretización simple y natural de la difusión Langevine mencionaremos alguna de sus propiedades.

Primeramente definiremos los siguientes limites, supongamos que para algún d fijo los siguientes limites existen:

$$\lim_{x\to\infty} \frac{1}{2}h\nabla \log(f(x))x^{-d} := S_d^+, \qquad \text{y} \qquad \lim_{x\to-\infty} \frac{1}{2}h\nabla \log(f(x))|x|^{-d} := S_d^-.$$

Lo cual nos permite introducir el siguiente teorema:

Teorema 0.9.4 *La cadena ULA $\{X_n : n \in \mathbb{N}\}$ es geométricamente ergódica si se cumple una de las siguientes condiciones:*

a) *Para algún $d \in [0,1)$ tanto $S_d^+ < 0$ y $S_d^- > 0$ existen.*

b) *Para $d = 1$ tanto $S_d^+ < 0$ y $S_d^- > 0$ existen y*

$$(1+S_d^+)(1-S_d^-) < 1.$$

Sin embargo, existen determinados casos donde la cadena ULA no es geométricamente ergódica, es más puede que ni siquiera sea ergódica.

Teorema 0.9.5 a) *La cadena ULA es ergódica en $\mathbb{R}$, pero no geométricamente ergódica si, para algún $d \in (-1,0)$ tanto $S_d^+ < 0$ y $S_d^- > 0$ existen.*

b) *La cadena ULA en $\mathbb{R}$ nos es ergódica, si para algún $d > 1$ tanto $S_d^+ < 0$ y $S_d^- > 0$ existen, o si, para $d = 1$ tanto $S_d^+ < -2$ y $S_d^- > 2$ existen.*

Esto nos da la idea de realizar ajustes a nuestro algoritmo con el fin de obtener la convergencia deseada.

0.9.4 Metropolis-adjusted Langevine Algorithm

Para lograr la convergencia y ergodicidad del algoritmo, se realiza una modificación al algoritmo ULA que consiste en añadir un paso de aceptación rechazo para construir un algoritmo tipo Metropolis-Hastings. Igualmente que en ULA, consideramos $h > 0$ fijo, teniendo así el siguiente algoritmo:

A. 8: Algoritmo de Langevine ajustado con Metropolis

Dado $X_t = x_t$ y $h > 0$,

1. Generar $Y_t \sim g(x_t, y)$, donde $g(x,\cdot) \sim N(x + \frac{1}{2}h\nabla \log(f(x)), h)$
2. Generar $U_t \sim U(0,1)$
3. Definir:

$$X_{t+1} = \begin{cases} Y_t & \text{Si } u < a(x_t, y_t) \\ x_t & \text{Si } u \geq a(x_t, y_t) \end{cases}$$

en donde,

$$a(x,y) = \text{mín}\left\{\frac{f(y_t)g(y_t,x_t)}{f(x_t)g(x_t,y_t)}, 1\right\}$$

.

Notamos que, por lo demostrado en la sección 0.8, la cadena converge a f, en el sentido que

$$||P^n(x,\cdot) - f||_{TV} \to 0$$

donde $P^n(x,\cdot)$ denota la probabilidad de transición a n pasos.

Para asegurar la convergencia geométrica del algoritmo MALA, necesitamos imponer algunas restricciones sobre la forma en que se aceptan los movimientos propuestos.

Denotaremos por $A(x)$ la región de aceptación de MALA desde el punto x, es decir, $A(x)$ sera la región en donde siempre se acepta la propuesta. Formalmente,

$$A(x) = \{y : f(x)g(x,y) \leq f(y)g(y,x)\}$$

en donde $g(x,y)$ es el kernel de transición de ULA, dado por A.7 . Denotando por $R(x) := A^c(x)$, el conjunto de puntos potencialmente rechazados, el cual contiene los puntos propuestos cuya probabilidad de ser rechazados es mayor a 0, y por

$$I(x) = \{y; |y| \leq |x|\}.$$

Con todos los elementos anteriores podemos introducir la siguiente definición:

Definición 0.9.2 — Convergencia hacia el interior. *Diremos que $A(\cdot)$ converge hacia el interior de q si:*

$$\lim_{|x|\to\infty} \int_{A(x)\Delta I(x)} g(x,y)dy = 0,$$

en donde $A\Delta B = (A \cup B)/(A \cap B)$.

Para las densidades que convergen hacia el interior tenemos una condición que garantiza la convergencia geométrica.

Teorema 0.9.6 *Supongamos que $c(x) = x + \frac{1}{2}h\nabla \log f(x)$ es la media de "la posición del nuevo candidatoz que:*

$$\liminf_{|x|\to\infty}(|x| - |c(x)|) > 0$$

Asumamos que $A(\cdot)$ converge hacia el interior de g. Entonces la cadena MALA es geométricamente ergódica.

Demostración: ver Roberts y Tweedie (1996).

La propiedad de la convergencia hacia el interior es, generalmente, posible de evaluar, pues se puede reescribir $A(x)$ como:

$$A(x) = \left\{ y : \int_y^x \nabla \log(f(x))dx \leq \frac{1}{2}(x-y)(\nabla \log(x) + \nabla \log f(y)) + \frac{h}{8}\left(|\nabla \log(f(x))|^2 - |\nabla f(y)|^2\right) \right\}$$

Sin embargo, al igual que para el algoritmo ULA, existen casos en donde no se logra la ergodicidad geométrica. Por ejemplo, en el caso en donde las colas de la distribución son mas livianas que la Gaussiana, no tenemos la ergodicidad geométrica, como lo expone el siguiente teorema.

Teorema 0.9.7 Si f es acotado y,

$$\liminf_{|x|\to\infty} \frac{|\nabla \log(f(x))|}{|x|} > \frac{4}{h}$$

Entonces la cadena MALA no es exponencialmente ergódica.

Por otra parte, si las colas de f son muy pesadas, entonces tampoco tendremos la ergodicidad geométrica, como ejemplifica el siguiente teorema:

Teorema 0.9.8 Si $\nabla \log f(x) \to 0$ entonces MALA no es geométricamente ergódico.

0.9.5 Metropolis adjusted Langevine truncated algorithm

Es posible ajustar el algoritmo mala para intentar capturar las mejores propiedades tanto del algoritmo de paseo aleatorio de Metrópolis como del candidato de Langevine ULA. Este algoritmo se denomina MALTA (Metropolis adjusted Langevine Truncated Algorithm).

El algoritmo consiste en truncar el paso del algoritmo ULA, obteniendo así el siguiente algoritmo

A. 9: Algoritmo truncado de Langevine ajustado con Metropolis

Dado $X_t = x_t$ y $h, D > 0$,

1. Generar $Y_t \sim g(x_t, y)$, donde $g(x,\cdot) \sim N(x + hR(x), h)$ y

$$R(x) = \frac{D\nabla \log(f(x))}{2\,\text{máx}\{D, |\nabla \log(f(x))|\}}$$

2. Generar $U_t \sim U(0,1)$
3. Definir:

$$X_{t+1} = \begin{cases} Y_t & \text{Si } u < a(x_t, y_t) \\ x_t & \text{Si } u \geq a(x_t, y_t) \end{cases}$$

en donde,

$$a(x,y) = \text{mín}\left\{\frac{f(y_t)g(y_t,x_t)}{f(x_t)g(x_t,y_t)}, 1\right\}$$

0.10 Metropolis-Hastings Adaptativo

En esta sección discutiremos otra alternativa de distribución instrumental que mejora la velocidad de convergencia, esta propuesta consiste en modificar la distribución muestral en cada paso de la iteración con el fin de optimizar la convergencia. Esta clase de algoritmos se conocen como Metropolis-Hastings adaptativo.

0.10.1 Metropolis-Hastings usando paseos aleatorios

En esta sección nos enfocaremos en distribuciones muestrales de la forma: $Y_{n+1} = X_n + Z_{n+1}$ con $X_i, Y_i \in \mathbb{R}^d$ y en donde el paso $Z_i \sim N(0, \sigma^2 I_d)$. Además asumiremos que la distribución objetivo f es continua y positiva. El objetivo es encontrar el σ^2 que optimice la velocidad de la convergencia.

Como se discutió en las secciones 0.8 y 0.9 la selección de la varianza σ^2 es crucial, puesto que si es muy pequeña entonces la cadena recorrerá el recorrido de la variable aleatoria objetivo muy lentamente y, por ende, la cadena se demorará en converger; mientras que si la la varianza σ^2 es muy grande, el recorrido de la variable aleatoria objetivo se recorre de manera muy errática. En ambos casos, se presentan problemas: la cantidad de observaciones aceptadas y la calidad de la simulación dependerá de la selección de σ^2. Esto nos motiva a buscar criterios de optimalidad.

Uno de los criterios de optimalidad más utilizados para encontrar la varianza consiste en fijar la tasa de de aceptación. La razón del porque esto es un punto de atención nace por los casos extremos: una tasa de aceptación de 1 implica que somos capaces de simular f directamente, lo cual sugiere que $\sigma^2 \to 0$; mientras que una tasa de aceptación cercana a 0 implica que nos es difícil simular y sugiere que $\sigma^2 \to \infty$. La regla empírica sugiere una tasa de aceptación *lejos de 0 y lejos de 1* (lo que sea que esto signifique), dejando demasiadas alternativas, en las cuales indagaremos a continuación.

0.10.2 Ratio de Aceptación Óptimo

El problema de selección de σ^2 en función del ratio de aceptación se ha estudiado anteriormente, ver Lee y Jung (2016)

Informalmente, se busca encontrar un ratio de aceptación $a(x,y)$ que optimice la convergencia del algoritmo. Para eso consideraremos un paseo aleatorio de Metropolis-Hastings y una distribución objetivo f de la forma:

$$f(x_1,x_2,\ldots,x_d) = \pi(x_1)\pi(x_2)\ldots\pi(x_d),$$

para alguna densidad π suave. Es decir, la densidad objetivo consiste en componentes independientes idénticamente distribuidos (i.i.d). Notemos que este supuesto es bastante poco realista pues, si sabemos simular de π entonces podemos simular de f. Bajo este supuesto, si tomamos pasos de la forma $N(0,\sigma^2 I_d)$, se puede probar que cuando $d \to \infty$, el radio óptimo es exactamente 0,234. Mas precisamente si asumimos que $\sigma^2 = \ell^2/d$, entonces, si el tiempo se acelera por un factor de d, y el espacio se reduce por un factor de $\sqrt{d}$, cada componente de la cadena de Markov convergerá a una difusión U_t definida por la solución de la siguiente ecuación diferencial estocástica:

$$dU_t = ((h(\ell))^{1/2} dB_t + h(\ell)\frac{f'(U_t)}{2f(U_t)}dt$$

donde h. a menudo llamada función de velocidad, esta dada por $h(\ell) = 2\ell^2\Phi(-\sqrt{I}\ell/2)$ en donde Φ es la distribución acumulada de una normal estándar, y I una constante que depende de π, de hecho:

$$I = \int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx.$$

Entonces la difusión es optimizada cuando $\ell = 2{,}38/\sqrt{I}$. Para mas detalles ver Gelman, Gilks y Roberts (1997)

Este resultado se pude extrapolar para el caso de densidades de la forma:

$$f(x) = \prod_{i=1}^{d} C_i f(C_i x_i)$$

en donde $\{C_i\}$ son i.i.d con distribución de varianza finita. Bajo estas hipótesis se puede demostrar que el ratio de aceptación óptimo 0.234 se mantiene y la eficiencia asintótica es proporcional a d^{-1}. Sin embargo, el algoritmo óptimo para una propuesta no homogénea tendrá eficiencia relativa asintótica mayor que el algoritmo óptimo para la propuesta homogénea, por un factor de bC_1^2 en donde $b = E(C_i^2)/E(C_i)^2 \geq 1$. Es decir, una menor homogéneidad de la distribución objetivo (es decir hay alta variabilidad de los C_i), resultará en un algoritmo mas lento.

Un caso de especial es cuando la distribución objetivo es $N(0,\Sigma)$ con Σ matriz de covarianza $d \times d$. Y los pasos son de la forma $N(0,\Sigma_p)$, lo cual es equivalente a tener un paso que distribuye $N(0,I)$ y una distribución objetivo de la forma $N(0,\Sigma\Sigma_p^{-1})$. Si definimos como $C_i = \sqrt{\lambda_i}$ con $\{\lambda_i\}_{i=1}^d$ los valores propios de la matriz $\Sigma\Sigma_p^{-1}$. Entonces para d grande, esto corresponde al caso en donde el $\{C_i\}$ es aleatorio con $E(C_i) = \frac{1}{d}\sum_{j=1}^d \sqrt{\lambda_j}$ y $E(C_i^2) = \frac{1}{d}\sum_{j=1}^d \lambda_j$. Por ende el factor b toma la forma:

$$b = \frac{E(C_i^2)}{(E(C_i))^2} \approx \frac{\frac{1}{d}\sum_{j=1}^d \lambda_j}{(\frac{1}{d}\sum_{j=1}^d \sqrt{\lambda_j})^2} = d\frac{\sum_{j=1}^d \lambda_j}{\left(\frac{1}{d}\sum_{j=1}^d \sqrt{\lambda_j}\right)^2}.$$

Notar que esta expresión se maximiza cuando los $\{\lambda_j\}$ son constantes, es decir, cuando $\Sigma\Sigma_p^{-1}$ es un múltiplo de la identidad.

Por lo tanto concluimos que para los pasos de la forma $N(0,\Sigma_p)$ y distribución objetivo de la forma $N(0,\Sigma)$ entonces el Σ_p óptimo es proporcional a Σ, es decir $\Sigma_p \approx k\Sigma$ para algún $k > 0$. Si fijamos $\Sigma_p = k\Sigma$ entonces podemos aplicar el resultado del ratio óptimo para el caso homogéneo, es decir fijamos $k = (2{,}38)^2/d$, por lo cual :

$$\Sigma_p = \left[\frac{(2{,}38)}{d}\right]\Sigma.$$

Notemos que a pesar de lo potente del resultado para este caso particular, no es muy útil si lo utilizamos directamente, pues estamos asumiendo que la densidad objetivo es una normal y el paso es también normal, lo que implica que podemos simular directamente de la distribución objetivo. Sin embargo este resultado es relevante pues nos motiva a definir un nuevo tipo de algoritmo Metropolis-Hastings, en el cual intentaremos modificar la varianza para así alcanzar el ratio de estimación óptimo.

0.10.3 Algoritmo de Metropolis-Hastings Adaptativo

Asumamos que tenemos un criterio bajo el cual la distribución muestral q es óptima. Por ejemplo, asumamos que una distribución muestral es óptima si alcanza el anterior ratio de aceptación óptimo 0,234. En ese caso necesitamos una forma de obtener una distribución muestral que nos garantice llegar a este radio.

Una forma de lograrlo, es tratar de modificar la densidad muestral para que el algoritmo alcance siempre el ratio de aceptación óptimo, en particular para las primeras iteraciones del algoritmo, pues es donde necesitamos alcanzar lo conocido como "burn in", que es el momento en donde la cadena converge a la distribución objetivo. A diferencia de los pasos posteriores al "burn in", pues ahí nos interesará mantener la densidad instrumental propuesta pues ya habremos alcanzado la distribución objetivo y querremos mantener la estacionaridad. Lo anterior motiva a definir el concepto de cadenas de Markov adaptativas.

Definición 0.10.1 — Cadenas de Markov adaptativas. *Sea $\{\mathbb{P}_\gamma\}_{\gamma\in\mathcal{Y}}$ una familia de kernels de transición para una cadena de Markov, cada una con la misma distribución estacionaria f. Sea Γ_n el kernel escogido en la $n-$ésima iteración, si el proceso estocástico $\{X_n\}_{n\in N}$ cumple con la siguiente propiedad:*

$$\mathbb{P}(X_{n+1}\in A|(X_n=x_n,\Gamma_n=\gamma_n),(X_{n-1}=x_{n-1},\Gamma_{n-1}=\gamma_{n-1})\ldots,(X_0=x_0,\Gamma_0=\gamma_0)=\mathbb{P}_\gamma(x,A)$$

Entonces $\{X_n : n\in\mathbb{N}\}$ es una cadena de Markov adaptativa.

Notemos que en la definición anterior tenemos directamente que el proceso $\{(X-n,\Gamma_n) : n\in\mathbb{N}\}$ es una cadena de Markov. En la practica la elección de Γ_n depende de el historial de la cadena, es decir, dependerá de $X_{n-1},\ldots,X_0,\Gamma_{n-1},\ldots,\Gamma_0$.

A pesar de que cada kernel de transición P_γ tiene como distribución estacionaria a f, no necesariamente la cadena de Markov adaptativa $\{X_n : n\in\mathbb{N}\}$ converge a la distribución estacionaria f. Por lo cual es importante imponer condiciones suficientes sobre las cuales la cadena $\{X_n : n\in\mathbb{N}\}$ converja en distribución a f. En particular tenemos el siguiente teorema que nos garantiza la convergencia

Teorema 0.10.1 *Asumamos que se cumplen las siguientes condiciones*

- *Condición de adaptación decreciente:*

$$\lim_{n\to\infty}\sup_x ||\mathbb{P}_{\Gamma_{n+1}}(x,\cdot)-\mathbb{P}_{\Gamma_n}(x,\cdot)||=0 \quad \textit{en probabilidad}$$

- *Condición de contención :*

$$\{M_\varepsilon(X_n,\Gamma_n)\}_{n=0}^{\infty} \quad \textit{es acotado en probabilidad, } \varepsilon > 0$$

donde $M_\varepsilon(x,y) = \text{ínf}\{n \geq 1 : ||P_\gamma^n(x,\cdot) - f(\cdot)|| \leq \varepsilon\}$ *es la convergencia del Kernel* P_γ *para la condición inicial* $x \in \Omega$.

Entonces se cumple que:

$$\lim_{n\to\infty} \sup_A ||\mathbb{P}(X_n \in A) - f(A)|| = 0$$

y

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = f(g)$$

para toda función $g : \Omega \to \mathbb{R}$ *acotada.*

Teninedo asi una condición suficiente que nos garantiza la convergencia de la cadena de Markov.

0.10.4 Metropolis-Hastings adaptativo

Definidas las cadenas de Markov adaptativas, notemos que podemos utilizarlas para modificar el algoritmo Metropolis-Hastings. Para lograr esto asumiremos que la densidad muéstrala g_t varia a través de cada iteración, más aun, asumiremos que la densidad muestral g_t depende solamente de la cadena $x_1,\cdots,x_t$ generada hasta el momento t, es decir, existe una función Γ tal que $g_{t+1} = \Gamma((x_1,x_2,\ldots,x_t))$. Bajo este supuesto podemos definir el algoritmo de Metropolis-Hastings adaptativo.

A. 10: Algoritmo Metropolis-Hastings adaptativo

Dado $X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0$ y $\Gamma(\cdot)$

1. Definir $g_t = \Gamma((x_1,x_2,\cdot,x_n))$
2. Generar $Y_t \sim g_t(x,y)$
3. Generar $U_t \sim U(0,1)$
4. Definir:

$$X_{t+1} = \begin{cases} Y_t & \text{Si } u < a(x_t,Y_t) \\ x_t & \text{Si } u \geq a(x_t,Y_t) \end{cases}$$

donde

$$a(x,y) = \text{mín}\left\{\frac{f(y_t)g_t(x,y)}{f(x_t)g_t(y,x)}, 1\right\}$$

La convergencia del método dependerá de la función Γ, la cual para cada muestra de tamaño $(x_1,x_2,\cdots,x_t)$ generada anteriormente le asigna una distribución g_t. En particular buscamos propuestas tal que se cumpla la condición de adaptación decreciente, lo cual nos asegura la convergencia. Si además asumimos que utilizaremos un paseo aleatorio con paso Z normal , entonces debemos hacer propuestas en las cuales la covarianza de g_t dependa de la muestra generada hasta el momento t.

Basándose en esta idea Haario, Saksman y Tamminen (2001) propusieron un algoritmo de Metropolis Hastings adaptativo, en el cual la distribución g_t es una distribución normal con media en el punto actual x_{t-1} y covarianza $\Sigma_t = Cov(x_0,x_1,\ldots,x_{t-1})$ y $\Sigma_0 = C_0$ con C_0 una matriz de covarianza arbitraria, definida positiva.

Esto se resume en el siguiente algoritmo:

A. 11: Algoritmo Metropolis-Hastings adaptativo de paseo aleatorio

Dado $X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0$

1. Definir $\Sigma_t = Cov(X_0, \ldots, X_n)$ y $g_t(x,y) \sim N\left(x, \left[\frac{(2{,}38)^2}{d}\right]\Sigma_t\right)$
2. Generar $Y_t \sim g(x, \cdot)$
3. Generar $U_t \sim U(0,1)$
4. Definir:

$$X_{t+1} = \begin{cases} Y_t & \text{Si } u < a(x_t, Y_t) \\ x_t & \text{Si } u \geq a(x_t, Y_t) \end{cases}$$

donde

$$a(x,y) = \text{mín}\left\{\frac{f(y_t)g_t(x,y)}{f(x_t)g_t(y,x)}, 1\right\}$$

Una problemática del algoritmo anterior es que las covarianzas pueden colapsar numéricamente a 0. Por lo cual Haario, Saksman y Tamminen (2001) propuso la siguiente forma de la covarianza: $C_t = s_d(cov(X_0, \ldots, X_{t-1} + \varepsilon I_d)$ con s_d un factor para lograr el ratio de aceptación deseado, en nuestro caso $s_d = (2{,}38)^2/d$ y $\varepsilon > 0$ una constante escogida para evitar que las covarianzas se vayan a 0. Otra posibilidad es asumir una distribución instrumental basada en la mixtura de distribuciones Gaussianas:

$$f(x;\beta) = (1-\beta)N\left(X_t, \left[\frac{(2{,}38)^2}{d}\right]\Sigma_t\right) + \beta N(X_t, \Sigma_0)$$

para algún $0 < \beta < 1$ y para alguna matriz de covarianza Σ_0, propuestos. Para mas detalles ver Roberts y Jeffrey S Rosenthal (2009).

Para los tres algoritmos anteriores, tendremos que para la $n-$ésima iteración, la variación de Σ_t con respecto de Σ_{t-1} sera sintéticamente de tamaño $1/t$. Por ejemplo si escogemos el primer algoritmo, A.10, y definimos $s_d = \frac{(2{,}38)^2}{d}$, podemos usar la formula recursiva para la covarianza por lo cual:

$$\lim_{t\to\infty} \|\Sigma_{t+1} - \Sigma_t\|_\infty = \lim_{t\to\infty} \left\| \frac{t-1}{t} C_t + \frac{s_d}{t}(t\bar{X}_{t-1}\bar{X}_{t-1}^\top - (t+1)\bar{X}_t\bar{X}_t^\top + X_t X_t^\top) - C_t \right\|_{\infty=0}.$$

Lo anterior permite demostrar que se cumple la condición de adaptación creciente. Además se puede demostrar (ver Haario, Saksman y Tamminen (2001) y Roberts y Jeffrey S Rosenthal (2009)) que si uno se restringe a regiones compactas entonces se cumplira también la condición de contención.

Es más la condición de contención se cumple tambien en los casos en donde la densidad objetivo decae, al menos, polinomialmente en cada coordenada (ver Bai, Roberts y Jeffrey Seth Rosenthal (2009)). Lo anterior junto con el Teorema 0.10.1 nos garantiza la convergencia de los algoritmos anteriores.

0.11 Gibbs Sampler

En muchos contextos aplicados, tenemos que tratar con distribuciones de probabilidad complejas definidas en espacios multidimensionales, en donde no es posible simular directamente de la distribución conjunta, pero si somos capaces de simular de las distribuciones condicionales. Además, los espacios multidimensionales son difíciles de visualizar, por lo cual es complicado calcular las regiones con mayor probabilidad de capturar un punto. Lo anterior produce que la simulación sea un reto en este contexto. Para solucionar este problema, introduciremos el algoritmo de Gibbs-Sampler, el cual nos permite simular distribuciones conjuntas a través de las distribuciones condicionales. A pesar de

que la motivación viene dada por el caso cuando la dimensión es grande, veremos primero el caso 2-dimensional para comprender mejor el algoritmo.

0.11.1 Gibbs Sampler de dos etapas

Sean X, Y variables aleatorias con densidad conjunta $f(x,y)$, además asumamos que podemos simular de las distribuciones condicionales, es decir, podemos simular de las distribuciones condicionales. Entonces se define el algoritmo Gibbs-Sampler de la siguiente manera:

A. 12: Gibbs Sampler

Dado $X_t = x_t$,
1. Generar $Y_{t+1} \sim f_{Y|X}(\cdot|x_t)$.
2. Generar $X_{t+1} \sim f_{X|Y}(\cdot|y_t)$

En donde por simplicidad denotamos $f_{X|Y}(\cdot|y)$ la densidad condicional de X dado $Y = y$. En el algoritmo anterior, la sucesión (X_t, Y_t), generada por el algoritmo es una cadena de Markov, más aún, las subsucesiones X_t e Y_t son también cadenas de Markov.

Notemos que la cadena X_t tiene kernel de transición:

$$K(x,x') = \int f_{Y|X}(y|x) f_{X|Y}(x'|y)\mathrm{d}y$$

la cual solo depende del ultimo valor de la cadena $\{X_t : n \in \mathbb{N}\}$. Además, podemos demostrar que la cadena de Markov tiene como distribución estacionaria a la densidad marginal f_X, pues:

$$\begin{aligned} f_X(x') &= \int f_{X|Y}(x'|y) f_Y(y)\mathrm{d}y \\ &= \int f_{X|Y}(x'|y) \int f_{Y|X} f_X(x)\mathrm{d}x \\ &= \int \left[\int f_{X|Y}(x'|y) f_{Y|X}(y|x) \right] f_X(x)\mathrm{d}x \\ &= \int K(x,x') f_X(x)\mathrm{d}x. \end{aligned}$$

Por lo cual este método tambien nos sirve para simular de una distribución univariada.

■ **Ejemplo 0.11.1 — Normal bivariada.** Consideremos la V.A dada por:

$$(X,Y) \sim N_2\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

recordemos las condicionales de la normal bivariada (MAT-041).

Dado y_t generar:

$$\begin{aligned} X_{t+1}|y_t &\sim N(\rho y_t, 1-\rho^2) \\ Y_{t+1}|x_{t+1} &\sim N(\rho x_{t+1}, 1-\rho^2) \end{aligned}$$

Notemos que la correspondiente cadena de Markov de X_t (resp. Y_t) está definida por la relación:

$$X_{t+1} = \rho^2 X_t + \sigma\varepsilon_t \quad \varepsilon_t \sim N(0,1)$$

con $\sigma^2 = 1-\rho^2+\rho^2(1-\rho^2) = 1-\rho^4$. Y además X tiene distribución estacionaria $N(0,1)$

■ **Ejemplo 0.11.2 — Mixtura de normales.** Consideremos la función de densidad de una mixtura de Gaussianas:

$$X|p \sim pN(\mu_1,\sigma^2)+(1-p)N(\mu_2,\sigma^2), \qquad p \sim ber(p),$$

es decir, X es puede ser generado con probabilidad p por la distribución $N(\mu_1,\sigma^2)$ o puede ser generado con probabilidad $1-p$ por la distribución $N(\mu_2,\sigma^2)$. Para simplificar la notación, denotemos la variable no observada Z con:

$$\mathbb{P}(Z_i=1)=1-\mathbb{P}(Z_i=2)=p \quad \text{y} \quad X|Z=i \sim N(\mu_k,\sigma^2)$$

Dada una muestra $(x_1,\ldots,x_n)$ i.i.d asumamos que queremos hacer inferencia sobre el parámetro $\mu=(\mu_1,\mu_2)$, el cual asumiremos aleatorio. Además asumamos que, para lograr lo anterior, necesitamos simular de la distribución $f_{\mu|x}$. Para la simulación podemos hacer un Gibbs Sampler incluyendo la variable no observada Z por lo cual necesitamos realizar los siguientes pasos:

- Simular μ de $\mu|x,z$
- Simular de $z|x,\mu$

Primeramente asumiremos que μ tiene distribución $N(0,10\sigma^2 I_{2\times 2})$. Para eso notemos que la densidad de X dado μ y dado Z es proporcional a:

$$f_{X|Z=z,\mu=(\mu_1,\mu_2)} \propto \exp(-(\mu_1^2+\mu_2^2)^2/20\sigma)\left(\prod_{z_i=1} p\exp(-(x_i-\mu_1^2)/2\sigma^2)\right)\left(\prod_{z_i=2}(1-p)\exp(-(x_i-\mu_2)^2/2\sigma^2)\right).$$

Como μ_1,μ_2 son independientes, dado (z,x) la distribución condicional para μ_i (con $i\in\{1,2\}$) es:

$$\mu_i|Z=z,X=x \sim N\left(\sum_{z_j=i} x_i/(0{,}1+n_i),\sigma^2/(0{,}1+n_i)\right),$$

con $n_i=\sum_{j=1}^n 1_{z_j=i}$, es decir la cantidad de z_j que son iguales a i. De igual forma la distribución condicional de Z dado μ es un producto de binomiales (por la independencia entre los z_j), con:

$$\mathbb{P}(Z_j=1|X_i=x_i,\mu=(\mu_1,\mu_2))=\frac{p\exp-(x_i-\mu_1)^2/2\sigma^2}{p\exp(-(x_i-\mu_1)^2/2\sigma^2+(1-p)\exp(x_i-\mu_2)^2/2\sigma^2},$$

teniendo así los pasos para simular la distribución objetivo.

0.11.2 Propiedades fundamentales de Gibbs Sampler de dos etapas

Una característica importante del Gibbs-Sampler es que las distribuciones condicionales tienen información suficiente para producir una muestra a partir de la distribución conjunta. Esta propiedad está fundamentada en el siguiente resultado importante que nos permite reconstruir la densidad conjunta a través de las densidades condicionales

Teorema 0.11.3 — Teorema de Hammersley-Clifford. *La distribución conjunta asociada a las densidades condicionales $f_{Y|X}$ y $f_{X|Y}(x|y)$ tiene densidad conjunta:*

$$f(x,y)=\frac{f_{Y|X}(y|x)}{\int[f_{Y|X}(y|x)/f_{X|Y}(x|y)]dy}.$$

Demostración: Como $f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$ entonces:

$$\int \frac{f_{Y|X}(y|x)}{f_{X|Y}(x|y)}\mathrm{d}y = \int \frac{f_Y(y)}{f_X(x)}\mathrm{d}y = \frac{1}{f_X(x)},$$

obteniendo el resultado.

Anteriormente demostramos que las subcadenas X_t e Y_t de la cadena de Markov $\{(X_t,Y_t) : n \in \mathbb{N}\}$ tienen como distribución estacionaria a sus distribuciones marginales, f_x y f_y respectivamente. Ahora nos enfocaremos en demostrar la irreductibilidad y la convergencia de las subcadenas y de la cadena conjunta.

Una condición suficiente para la irreductibilidad de la cadena (X_t,Y_t) es la llamada condición de positividad introducida por Besag (1974):

Definición 0.11.1 — Condición de positividad. *Sea $[Y_1,Y_2,\ldots,Y_p]$ un vector aleatorio con densidad conjunta $g(y_1,y_2,\ldots y_p)$, y sean $g_i(y_i)$ la distribución marginal de Y_i. g satisface la condición de positividad si y solo si $g_i(y_i) > 0$ para todo $i = 1,\ldots,p$ implica que $g(y_1,y_2,\ldots y_p) > 0$.*

Es decir, el soporte de g es el producto cartesiano del soporte de las marginales g_i. Esta definición nos permite introducir el siguiente teorema:

Teorema 0.11.4 *Cada una de las subcadenas X_t e Y_t son cadenas de Markov con distribuciones estacionarias f_X e f_Y respectivamente. Además, si f cumple la condición de positividad, entonces las dos cadenas son irreducibles.*

Demostración: La estacionaridad ya fue demostrada en la sección 4.6.1. Bajo la condición de positividad, si f es positiva entonces $f_{X|Y}(x|y)$ es positiva en el soporte proyectado de f y para cada conjunto de Borel en Ω puede ser visitado en 1 paso por la cadena, estableciendo así la irreductibilidad. El mismo razonamiento aplica para la cadena $\{Y_t : t \in \mathbb{N}\}$.

La convergencia de la cadena de Markov conjunta $\{(X_t,Y_t) : t \in \mathbb{N}\}$, es un caso especial del algoritmo Gibbs-Sampler de múltiples etapas. Sin embargo, estableceremos aquí el siguiente resultado de convergencia, cuya demostración es similar a la del lema 4.3.1 y la del teorema 4.3.2.

Teorema 0.11.5 *Bajo la condición de positividad, si el kernel de transición:*

$$K((x,y),(x',y')) = f_X(x')f_{Y|X}(y'|x')$$

es absolutamente continuo con respecto a la medida de Lebesgue, entonces la cadena (X_t,Y_t) es Harris recurrente y ergódica con distribución estacionaria f.

0.11.3 Gibbs Sampler de múltiples etapas

El algoritmo de Gibbs Sampler de múltiples etapas es una extensión del algoritmo Gibbs Sampler de dos etapas. Asumamos que X es una variable aleatoria p dimensional con $p > 2$. Además supongamos que podemos simular de las distribuciones condicionales $f_1,\ldots,f_p$, en donde por simplicidad denotamos por $f_i(x_i|x_1,x_2,\ldots,x_{i-1},x_{i+1},\ldots,x_p)$ la densidad de X_i condicionada a $X_1 = x_1, X_2 = x_2, \ldots X_p = x_P$. Entonces el algoritmo Gibbs Sampler de p etapas se define de la siguiente manera:

A. 13: Algoritmo Gibbs Sampler de múltiples etapas

Dado $X^t = (x_1^{(t)}, \ldots, x_p^{(t)})$

1. Generar $X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \ldots, x_p^{(t)})$.
2. Generar $X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)})$.

   $\vdots$

p. Generar $X_p^{(t+1)} \sim f_p(x_p | x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$

Las densidades $f_1, \cdots, f_p$ se llaman full condicionales. Una particular característica del Gibbs Sampler es que estas son las únicas densidades utilizadas para la simulación. Así, incluso para casos de dimensión grande, todas las simulaciones se hacen de densidades univariadas, lo que suele ser una ventaja, como lo ilustra el siguiente ejemplo:

■ **Ejemplo 0.11.6 — Modelo Auto-exponencial.** El modelo auto-exponencial Besag (1974) se ha encontrado útil en algunos aspectos bajo el contexto de estadística espacial. Cuando $y \in \mathbb{R}_+^3$, la densidad correspondiente es:

$$f(y_1, y_2, y_3) \propto \exp(-(y_1 + y_2 + y_3 + \theta_{12} y_1 y_2 + \theta_{23} y_2 y_3 + \theta_{31} y_3 y_1))$$

para $\theta > 0$ conocido. Notamos que las densidades condicionales son exponenciales. Por ejemplo:

$$Y_3 | y_1, y_2 \sim \exp(1 + \theta_{23} y_2 + \theta_{31} y_1),$$

las cuales son muy fáciles de simular. En contraposición las otras densidades condicionales y las densidades marginales tienen la siguiente forma:

$$f(y_2 | y_1) \propto \frac{\exp(-(y_1 + y_2 + \theta_{12} y_1 y_2))}{1 + \theta_{23} y_2 + \theta_{31} y_1},$$
$$f(y_1) \propto e^{-y_1} \int_0^{+\infty} \frac{\exp(-y_2 - \theta_{12} y_1 y_2)}{1 + \theta_{23} y_2 + \theta_{31} y_1} \mathrm{d}y_2,$$

las cuales no pueden ser simuladas de forma fácil

El algoritmo de Gibbs Sampler presenta las siguientes diferencias con el algoritmo de Metrópolis-Hastings:

- El radio de aceptación de Gibbs Sampler es uniformemente igual a 1. Por lo tanto, cada valor simulado es aceptado y las sugerencias de la Sección 4.5 sobre el radio de aceptación óptimo no se aplican en este caso. Esto significa que la evaluación de la convergencia para este algoritmo debe tratarse de forma distinta que lo utilizado para el Metrópolis-Hastings.
- El uso del Gibbs Sampler implica limitaciones en la elección de las distribuciones instrumentales y requiere un conocimiento previo de algunas propiedades de la función objetivo f.
- El Gibbs Sampler es, por construcción, multidimensional. Aunque algunos componentes del vector simulado pueden ser artificiales y/o innecesarios para la para el problema de interés, la construcción es al menos bidimensional.

0.11.4 Teorema general de Hammersley-Clifford

Ya vimos el teorema de Hammersley-Clifford en el caso especial de un Gibbs Sampler dos dimensional. Este teorema puede ser generalizado para el caso en donde la dimensión d es mayor que

2. Para lo anteriori necesitamos la condicion de positividad (Definición 0.11.1), con la cual podemos obtener el siguiente resultado:

Teorema 0.11.7 — Teorema de Hammersley Clifford. *Bajo la condición de positividad, la distribución conjunta f satisface:*

$$f(x_1,\ldots,x_p) \propto \prod_{j=1}^{p} \frac{f_{\ell_j}(x_{\ell_j}|x_{\ell_1},\ldots,x_{\ell_{j-1}},x'_{\ell_{j+1}},\ldots,x'_{\ell_p})}{f_{\ell_j}(x'_{\ell_j}|x_{\ell_1},\ldots,x_{\ell_{j-1}},x'_{\ell_{j+1}},\ldots,x'_{\ell_p})}$$

para toda permutación ℓ y para toda $x' \in \Omega$

Demostración:

$$\begin{aligned} f(x_1,\ldots,x_p) &= \frac{f_p(x_p|x_1,\ldots,x_{p-1})}{f_p(x'_p|x_1,\ldots,x_{p-1})} f(x_1,\ldots,x_{p-1},x'_p) \\ &= \frac{f_p(x_p|x_1,\ldots,x_{p-1})}{f_p(x'_p|x_1,\ldots,x_{p-1})} \frac{f_{p-1}(x_{p-1}|x_1,\ldots,x'_p)}{f_{p-1}(x'_{p-1}|x_1,\ldots,x'_p)} f(x_1,\ldots,x'_{p-1},x'_p). \end{aligned}$$

Usando un argumento recursivo tenemos que:

$$f(x_1,\ldots,x_p) = \prod_{j=1}^{p} \frac{f_{\ell_j}(x_{\ell_j}|x_{\ell_1},\ldots,x_{\ell_{j-1}},x'_{\ell_{j+1}},\ldots,x'_{\ell_p})}{f_{\ell_j}(x'_{\ell_j}|x_{\ell_1},\ldots,x_{\ell_{j-1}},x'_{\ell_{j+1}},\ldots,x'_{\ell_p})} f(x'_1,\ldots,x'_p).$$

0.11.5 Propiedades del Gibbs Sampler de múltiples etapas

Una propiedad importante que debemos asegurar para cualquier algoritmo es la convergencia. Para esto, se demostrará la convergencia de cada una de las subcadenas de la forma $(X^{(t)}_{\ell_1}, X^{(t)}_{\ell_2}, \ldots X^{(t)}_{\ell_n})$ que pueda presentar nuestra cadena principal. Para eso denotemos por Y_t una subcadena arbitraria de la cadena de Markov principal, tal que $Y_t = (X^{(t)}_{\ell_1}, X^{(t)}_{\ell_2} \ldots X^{(t)}_{\ell_n})$ con $\ell \subset \{1,2,\ldots,p\}$.

Primero, demostraremos que, si la cadena principal es ergódica, toda subcadena es también ergódica.

Teorema 0.11.8 *Para el algoritmo Gibbs Sampler de múltiples etapas, si la cadena principal $\{X_t : t \in \mathbb{N}\}$ es ergódica entonces la distribución f es una distribución estacionaria de la cadena $\{X_t : t \in \mathbb{N}\}$, y toda subcadena $\{Y_t : t \in \mathbb{N}\}$ es ergódica con distribución limite f_ℓ en donde f_ℓ denota la densidad marginal en las variables dadas por el conjunto de índices ℓ.*

Demostración: El kernel de transición de la cadena $\{X_t : t \in \mathbb{N}\}$ es el producto:

$$K(x,x') = f_1(x'_1|x_2,\ldots,x_p) f_2(x'_2|x'_1,x_3\ldots,x_p) f_p(x'_p|x'_1,\ldots,x'_{p-1})$$

Para el vector $x = (x_1,x_2,\ldots,x_p)$, sea $f^i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_p)$ la densidad marginal del vector x, obtenida integrando la componente x_i. Si $X \sim f$ y A es medible con respecto a la medida de Lebesgue, entonces:

$$\begin{aligned} \mathbb{P}(X' \in A) &= \int 1_A(x') K(x,x') f(x) \mathrm{d}x' \mathrm{d}x \\ &= \int 1_A(x') [f_1(x_1|x_2,\ldots,x_p)\ldots f_p(x'_p,x'_1,\ldots,x'_{p-1})][f_1(x_1|x_2,\ldots,x_p) f^1(x_2,\ldots,x_p)] \mathrm{d}x_1\ldots x_p \mathrm{d}_{x'_1} \mathrm{d}_{x'_p} \\ &= \int 1_A(x') [f_2(x_1|x_2,\ldots,x_p)\ldots f_p(x'_p,x'_1,\ldots,x'_{p-1}] f(x'_1,x_2,\ldots,x_p) \mathrm{d}x_1\ldots x_p \mathrm{d}_{x'_1} \mathrm{d}_{x'_p}, \end{aligned}$$

en donde integramos la componente x_1 y combinamos $f(x'_1|x_2,\ldots,x_p) f^1(x_2,\ldots,x_p) = f(x'_1,x_2,\ldots,x_p)$. Entonces escribimos $f(x'_1,x_2,\ldots,x_p) = f(x_2|x'_1,x_3,\ldots x_p) f^2(x'_1,x_3,\ldots,x_p)$ e integramos la variable x_2

obteniendo:

$$\mathbb{P}(X' \in A) = \int 1_A(x') f_3(x'_3|x'_1, x'_2, \dots x_p) \dots f_p(x'_p|x'_1, \dots, x'_{p-1}) f(x'_1, x'_2, x_3, \dots, x_p) \mathrm{d}x_3 \dots \mathrm{d}x_p \mathrm{d}x_{1'} \dots \mathrm{d}x_{x'_p}$$

Si seguimos integrando de esta manera los x_i, la probabilidad resultante es:

$$\mathbb{P}(Y' \in A) = \int_A f(x'_1, \dots, x'_p) \mathrm{d}x'$$

Demostrando que f es la distribución estacionaria de la cadena. Además, por el teorema (to do), la cadena X tiene como distribución límite a f y por lo tanto la subcadena Y_t tiene distribución límite a la marginal de f, ie f_ℓ.

Al igual que en el caso del Gibbs Sampler de dos etapas, para obtener la irreductibilidad de la cadena necesitamos la condición de positividad 0.11.1:

Teorema 0.11.9 *Para el algoritmo Gibbs Sampler de múltiples etapas, si la densidad f satisface la condición de positividad, entonces es irreducible.*

CITA. La demostración de este teorema es similar a la demostración del Teorema 0.11.4.

Sin embargo, las condiciones de este teorema muchas veces son difíciles de verificar, por lo cual Tierney (1994) dio una condición que es mucho más usada en la práctica.

Teorema 0.11.10 *Si el kernel de transición asociado al algoritmo de Gibbs Sampler es absolutamente continuo con respecto a la medida de Lebesgue entonces la cadena resultante es Harris recurrente.*

Cuya demostración es similar al lema 0.8.2.

Gracias a la recurrencia Harris podemos obtener el siguiente resultado, el cual es bastante similar al Teorema 0.8.3

Teorema 0.11.11 *Si el kernel de transición de la cadena $\{X_t : t \in \mathbb{N}\}$ es absolutamente continuo con respecto a la medida μ.*

1. *Si $h_1, h_2 \in L^1(f)$ con $\int h_2(x)\mathrm{d}(x) \neq 0$ entonces:*

$$\lim_{n\to\infty} \frac{\sum_{t=1}^T h_1(X^{(t)})}{\sum_{t=1}^T h_2(X^{(t)})} = \frac{\int h_1(x)\mathrm{d}f(x)}{\int h_2(x)\mathrm{d}f(x)}, \text{ casi seguramente.}$$

2. *Si, además $(X^{(t)})$ es aperiódica, entonces, para toda distribución inicial μ.*

$$\lim_{n\to\infty} \left|\left| \int K^n(x, \cdot)\mu(\mathrm{d}x) - f \right|\right| = 0.$$

0.11.6 Algoritmo de Gibbs Sampler como un algoritmo de Metropolis-Hastings

Si bien es cierto, la construcción del algoritmo de Gibbs Sampler es diferente de un algoritmo de Metropolis-Hastings, existen relaciones entre ambos. Intuitivamente, un algoritmo de Gibbs Sampler es la composición de p kernels Markovianos.

Teorema 0.11.12 *El algoritmo de Gibbs Sampler es equivalente a la composición de p algoritmos de Metropolis-Hastings cada una con probabilidad de aceptación uniformemente igual a 1*

Demostración: Si escribimos el algoritmo de Gibbs Sampler en p algoritmos, cada uno correspondiente al $p-$ésimo paso de simulación de la distribución condicional, es suficiente mostrar que cada uno de estos algoritmos tiene probabilidad de aceptación igual a 1. Para $1 \leq i \leq p$, la distribución instrumental en el paso i está dada por:

$$q_i(x'|x) = \delta_{(x_1,\dots,x_{i-1},x_{i+1},\dots x_p)}(x'_1, \dots, x'_{i-1}, x'_{i+1}, x'_p) f(x'_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p),$$

y el ratio de aceptación está dado por:

$$\begin{aligned}\frac{f(x')f_i(x|x')}{f(x)f_i(x'|x)} &= \frac{f(x')f_i(x_i|x_1,\ldots,x_{i-1},x_{i+1},x_p)}{f(x)f_i(x_i'|x_1,\ldots,x_{i-1},x_{i+1},x_p)}\\ &= \frac{f_i(x_i'|x_1,\ldots,x_{i-1},x_{i+1}\ldots x_p)f_i(x_i|x_1,\ldots,x_{i-1},x_{i+1},x_p)}{f_i(x_i'|x_1,\ldots,x_{i-1},x_{i+1},x_p)f_i(x_i|x_1,\ldots,x_{i-1},x_{i+1}m\ldots x_p)}\\ &= 1.\end{aligned}$$

■ **Ejemplo 0.11.13 — Continuación ejemplo 0.11.6.** Consideremos el modelo auto-exponencial dos-dimensional:

$$f(x_1,x_2) \propto \exp(-x_1 - x_2 - \theta_{12}x_1x_2),$$

con las siguientes densidades condicionales:

$$\begin{aligned}f_1(x_1|x_2) &= (1+\theta_{12}x_2)\exp(-(1+\theta_{12}x_2)x_1)\\ f_2(x_2|x_1) &= (1+\theta_{12}x_1)\exp(-(1+\theta_{12}x_1)x_2),\end{aligned}$$

entonces el ratio:

$$\frac{f(x_1',x_2)f(x_1|x_2)}{f(x_1,x_2)f(x_1'|x_2)} = \frac{\exp(-x_1' - x_2 - \theta_{12}x_1'x_2)(1+\theta_{12}x_2)\exp(-(1+\theta_{12}x_2)x_1)}{\exp(-x_1 - x_2 - \theta_{12}x_1x_2)(1+\theta_{12}x_2)\exp(-(1+\theta_{12}x_2)x_1')} = 1.$$

A pesar de lo anterior, el algoritmo Gibbs Sampler no es un caso particular del algoritmo Metropolis-Hastings, pues el ratio (global) de aceptación del vector $(x_1,\ldots,x_p)$ usualmente no es igual a 1, por lo cual un algoritmo tipo Metropolis-Hastings generaría una probabilidad de rechazo del vector.

Lo anterior se ilustra en el siguiente ejemplo:

■ **Ejemplo 0.11.14 — Continuación ejemplo 0.11.13.** Notemos que para el modelo auto-exponencial dos-dimensional el kernel de transición esta dado por:

$$K((x_1,x_2),(x_1',x_2'))) = f_1(x_1'|x_2)f_2(x_2'|x_1'),$$

entonces el ratio:

$$\frac{f(x_1',x_2')}{f(x_1,x_2)}\frac{K((x_1',x_2'),(x_1,x_2)}{K((x_1,x_2),(x_1',x_2')} = \frac{(1+\theta_{12}x_2')(1+\theta_{12}x_1)}{(1+\theta_{12}x_2)(1+\theta_{12}x_1')}\exp(\theta_{12}(x_2'x_1 - x_1'x_2))$$

lo cual es diferente de 1 para casi todo vector (x_1,x_2,x_1,x_2').

0.11.7 Aplicaciones del Gibbs Sampler a estructuras jerárquicas

Existen diversos casos donde proponer un algoritmo tipo Gibbs Sampler resulta ser más natural que proponer otro tipo de algoritmo de simulación. Este es el caso de los llamados *modelos con estructuras jerárquicas*, los cuales se basan son estructuras en donde la distribución conjunta f puede ser descompuesta de la siguiente manera:

$$f(x) = \int f_1(x|z_1)f_2(z_1|z_2)\ldots f_{n-1}(z_{n-1}|z_n)f_n(z_n)\mathrm{d}z_1\ldots\mathrm{d}z_n,$$

Tales modelos aparecen naturalmente en el análisis bayesiano (ver capitulo 0.21.8) de modelos complejos, donde la diversidad de distribuciones a priori o la variabilidad de las observaciones puede requerir la introducción de varios niveles de distribuciones a priori (ver Wakefield et al. (1994)).

En estos modelos, las distribuciones pueden tener varios niveles de jerarquía. Los siguientes ejemplos ilustran situaciones donde los modelos jerárquicos son particularmente útiles.

■ **Ejemplo 0.11.15 — Epidemiología veterinaria.** La investigación en epidemiología veterinaria a veces utiliza datos de grupos de animales, como camadas o rebaños. Tales datos pueden no seguir algunas de las suposiciones usuales de independencia, y, como resultado, las varianzas de las estimaciones de los parámetros tienden a ser mayores (este fenómeno es conocido como "sobre-dispersión"). Schukken, Casella y Van den Broek (1991) obtuvieron conteos del número de casos de mastitis clínica en rebaños de ganado lechero durante el periodo de un año.

Si asumimos que, en cada rebaño, la aparición de una mastitis es una variable Bernoulli, y sea $X_i, i = 1,\ldots,m$ el número de casos de mastitis en el rebaño i, entonces es razonable asumir que $X_i \sim Poisson(\lambda_i)$ donde λ_i es la tasa de infección del rebaño i. Sin embargo, no hay independencia entre las X_i pues la mastitis es infecciosa, lo cual podría provocar una sobre-dispersión. Por esto, Schukken, Casella y Van den Broek (1991) modelaron λ_i con una distribución gamma que depende de un parámetro aleatorio β. Teniendo entonces el siguiente modelo jerárquico:

$$\begin{aligned} X_i &\sim Poisson(\lambda_i) \\ \lambda_i &\sim Ga(\alpha, \beta_i) \\ \beta_i &\sim Ga(a, c) \end{aligned}$$

en donde α, a, c son conocidos. Entonces las densidades condicionales de λ_i y β_i son:

$$\begin{aligned} \lambda_i | X_i = x_i, \beta_i = b &\sim Ga(x_i + \alpha, 1 + \beta_i) \\ \beta_i | \lambda_i = l_i, X_i = x_i &\sim Ga(\alpha + a, l_i + b) \end{aligned}$$

■ **Ejemplo 0.11.16 — Modelos Médicos.** En la Farmacocinética, uno de los tópicos más frecuentes es modelar la relación entre la dosis de un fármaco y la concentración resultante en la sangre. (En términos más generales, la farmacocinética estudia las diferentes interacciones de un fármaco y el organismo.)

Gilks et al. (1993) introdujo un nuevo enfoque para estimar los parámetros farmacocinéticos, utilizando un modelo tradicional de efectos mixtos y una estructura no lineal, pero que también es robusto a los valores atípicos comunes en los ensayos clínicos. Para una dosis d_i administrada en el momento 0 al paciente i, la concentración logarítmica X_{ij} medida en la sangre en el tiempo t_{ij}, sigue una distribución t de Student. Es decir,

$$\frac{X_{ij} - \log(g_{ij}(\lambda_i)}{\sigma\sqrt{n/(n-2)}} \sim t(n)$$

donde $\lambda_i = (\log C_i, \log V_i)$ es un vector de parámetros para el $i-$ésimo individuo, σ^2 es la varianza del error, y g_{ij} está dado por:

$$g_{ij}(\lambda_i) = \frac{d_i}{V_i} \exp\left(-\frac{C_i}{V_i} y_i j\right).$$

Gilks et al. (1993) completo el modelo, asumiendo una distribución para σ tal que $\pi(\sigma) = 1/\sigma$ y además propuso:

$$\begin{aligned} \lambda_i &\sim N(\theta, \sigma) \\ \theta &\sim N(\tau_1, T_1) \\ \Sigma^{-1} &\sim W_2(\tau_2, T_2) \end{aligned}$$

donde los valores τ_1, T_1, τ_2 y T_2 son conocidos. Se pueden usar estructuras conjugadas para la mayoría de los parámetros usando la descomposición de Dickey para la distribución t de Student, la cual consiste en asociar a cada variable X_{ij}, una variable (artificial) W_{ij} tal que:

$$X_{ij}|W_{ij} = w_{ij} \sim N\left(\log(g_{ij}(\lambda_i)), \sigma^2 \frac{w_{ij} n}{n-2}\right)$$

Usando esta variable aleatoria, las distribuciones full condicionales de C_i y θ son distribuciones normales, mientras que las distribuciones full condicionales de σ^2 y Σ son una gamma inversa y una inversa Wishart, respectivamente. El caso de V_i es más complicado, pues la full condicional es proporcional a:

$$\exp\left(-\frac{1}{2}\left\{\sum_j \left(\log V_i + \frac{C_i t_{ij}}{V_i} - \mu_i\right)^2 / \left(\varepsilon + (\log V_i - \gamma_i)^2/\zeta\right)\right\}\right)$$

donde los parámetros $\mu_i, \varepsilon, \gamma_i$ y ζ depende de los otros parámetros x_{ij}. Gilks et al (1993) sugirió usar un algoritmo aceptación rechazo para simular la distribución anterior. Otra posibilidad es usar un paso de Metrópolis-Hastings (ver sección 4.6.8).

■ **Ejemplo 0.11.17 — Falla de estaciones de bombeo.** Gaver y O'Muircheartaigh (1987) (ver Gaver y O'Muircheartaigh (1987)) introdujeron un modelo que es muy usado en la literatura de Gibbs Sampler.

Supongamos que tenemos 10 estaciones de bombeo , y sea y_i el número de fallas de la $i-$ésima estación en un periodo de tiempo $t_i > 0$. Asumiremos que el vector $t = (t_1, \ldots, t_{10})$ es conocido. Y además denotaremos $Y = (y_1, \ldots, y_{10})$.

Entonces asumiremos que el modelo tiene la siguiente estructura.

1. La distribución de cada Y_i depende de la realización de una variable aleatoria positiva Λ_i, que interpretaremos como una tasa de fallas no observadas. Además denotaremos $\Lambda = (\Lambda_1, \ldots, \Lambda_{10})$. Particularmente, la distribución de Y_i dado $\Lambda_i = \lambda_i$ es una distribución Poisson con media $t_i\lambda_i$.
2. Asumiremos que las variables Y_i condicionadas a Λ_i son independientes.
3. La distribución de Λ depende de una variable aleatoria β, el cual interpretaremos como el nivel medio de la tasa de fallas no observadas. Particularmente, asumiremos que los Λ_i condicionados a β son independientes y además cada Λ_i distribuyen exponencial con parámetro β.
4. Asumiremos que β distribuye exponencial con parámetro $b = 40$ (este valor se obtuvo mediante algunos experimentos).
5. Finalmente, asumiremos que la distribución condicional de Y dado (Λ, β) no depende de β sino solo de Λ.

Todo lo anterior se resume en la siguiente imagen: .

Solo necesitamos calcular las distribuciones full condicionales:

Primeramente notemos que por el diagrama anterior $Y|\Lambda = \lambda, \beta = b \sim Y|\Lambda = \lambda$ entonces tendremos que:

$$Y|\Lambda = (\lambda_1, \ldots, \lambda_{10}), \beta = b \sim \prod_{i=1}^{10} \frac{(t_i\lambda_i)^{y_i}}{y_i!} e^{-t_i\lambda_i}, \quad y = (y_1, \ldots, y_{10}) \in \mathbb{N}^{10}$$

Por otro lado como (Y, Λ, β) siguen un modelo jerárquico entonces para la distribución conjunta $f_{Y,\Lambda,\beta}(y, \lambda, b)$, tenemos la siguiente igualdad:

$$f_{Y,\Lambda,\beta}(y, \lambda, b) = f_{Y|\Lambda}(y|\lambda, b) f_{\Lambda,\beta}(\lambda, \beta) = f_{Y|\Lambda}(y|\lambda) f_{\Lambda}(\lambda|b) f_{\beta}(b)$$

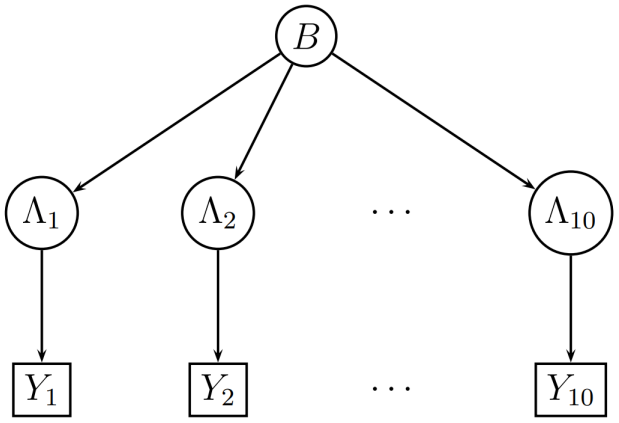

Figura 1: Estructura jerárquica para las fallas en estaciones de bombeo

entonces tendremos que:

$$f_{\Lambda,\beta|Y}(\lambda,b|y) \propto \beta^{10} e^{-40\beta} \prod_{i=1}^{10} \lambda_i^{y_i} e^{-(t_i+\beta)\lambda_i}$$

Además sabemos que los $\Lambda_i|\beta = b$ son independientes, por lo tanto podemos obtener la distribución de $\Lambda_i|Y_i = y_i, \beta = b$ usando el teorema de Bayes llegando al siguiente resultado:

$$f_{\Lambda_i|Y,\beta}(\lambda_i|y,b) \propto \lambda_i^{y_i} e^{-(t_i+b)\lambda_i},$$

lo cual es una densidad gamma con parámetros $y_i + 1$ y $t_i + b$.

Por otro lado, igualmente por el teorema de Bayes podemos obtener que:

$$f_{\beta|\Lambda,Y}(b|\lambda,b) \propto \beta^{10} e^{-(40+\sum_{i=1}^{n}\lambda_i)b}$$

lo cual es una densidad gamma con parámetros 11 y $40 + \sum_{i=1}^{n} \lambda_i$.

Todo esto resulta en el siguiente algoritmo:

I Escoger los parámetros iniciales $Y_0 = y^0, \Lambda_0 = \lambda^0, \beta_0$

II For $n = 0, 1, \ldots$:

a) Generar β_{n+1} de una distribución gamma de parámetros 11 y $\sum_{i=1}^{n} \lambda_i^n + 40$

b) For $i = 1, \ldots, 10$ generar Λ_i^{n+1} de una distribución gamma de parámetros $y_i + 1$ y $t_i + \beta_n$.

c) For $i = 1, \ldots, 10$ generar Y_i de una distribución Poisson con media $\lambda_i t_i$.

0.11.8 Algoritmo de Gibbs Sampler con Metrópolis-Hastings

Un algoritmo MCMC híbrido es un algoritmo que combina Metrópolis-Hastings con un Gibbs Sampler. Son generalmente usados cuando las densidades de $x_i|X_{-i}$ condicionales no pueden ser fácilmente simuladas. En tales casos para cualquier paso i del Gibbs Sampler sustituimos la simulación de $f_i(x_i|x_j)$ por la simulación de una distribución instrumental q_i. Esta modificación fue propuesta por Muller (1991) (ver Miiller (1991)) dando paso al siguiente algoritmo:

A. 14: MCMC Híbrido

For $i = 1, \ldots, p$ dado $X^t = (x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_i^{(t)}, \ldots, x_p^{(t)})$

1. Simular

$$y_i \sim q_i(y_i | x_1^{(t+1)}, \ldots, x_i^{(t)}, x_{i+1}^{(t)}, \ldots, y_p^{(t)})$$

2. Tomar

$$x_i^{(t+1)} = \begin{cases} x_i^{(t)} & \text{con probabilidad} \quad 1-a \\ \\ y_i & \text{con probabilidad} \quad a \end{cases}$$

donde

$$a = \text{mín} \left\{ 1, \frac{f_i(y_i | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots x_p^{(t)})}{q_i(y_i | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_i^{(t)}, x_{i+1}^{(t)}, \ldots y_p^{(t)})} \right.$$
$$\left. \cdot \frac{q_i(x_i^{(t)} | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_i^{(t)}, x_{i+1}^{(t)}, \ldots x_p^{(t)})}{f_i(x_i^{(t)} | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, y_i, x_{i+1}^{(t)}, \ldots x_p^{(t)})} \right\}$$

Un punto importante acerca de esta modificación, es que el paso anterior del Metropolis-Hastings solo se usa una vez en cada iteración del algoritmo. El paso modificado produce así una sola simulación de X_i, en lugar de tratar de aproximar $f_i(x_i | x_j, j \neq i)$ con mayor precisión mediante la producción de N simulaciones de q_i. Las razones para esta elección son dos: primero, el algoritmo híbrido resultante es válido ya que f es su distribución estacionaria. En segundo lugar, el Gibbs Sampler también conduce a una aproximación de f, proporcionar una aproximación más .exacta"de $f_i(x_i | x_j, j \neq i)$ en el MCMC Híbrido no nos conduce necesariamente a una mejor aproximación de f y la sustitución de f_i por q_i puede ser incluso beneficiosa para la velocidad de excursión de la cadena en la superficie de f (Chen y Schmeiser (1998)).

III Métodos estadísticos basados en simulación

Integracion Monte Carlo

Integrar funciones es un elemento fundamental en diversas áreas de la matemática. En probabilidad y estadística es importante debido a que estamos enfocados en calcular el valor esperado de una variable aleatoria X o su varianza, por ejemplo. Usualmente, el cálculo de estas cantidades puede ser complicado y técnicas numéricas son requeridas para tener una estimación de dicha integral.

En este capítulo nos enfocaremos en realizar estimaciones de integrales mediante la simulación de números aleatorios o de proceso estocásticos, según sea el caso. Para esto, nos enfocaremos en estimar la integral

$$I_D = \mathbb{E}_f[h(X)] = \int_D h(x)f(x)\mathrm{dx}, \tag{10}$$

donde $x \in D \subset \mathbb{R}^d$.

0.12 Integracion Monte Carlo

La integración Monte Carlo se basa en estimar (III) utilizando la ley fuerte de los grandes números.

Teorema 0.12.1 — Ley fuerte de los grandes números. Sea $X_1, X_2, \ldots$ una secuencia de variables aleatorias independientes e identicamente distribuidas con $\mathbb{E}(X_1) = \mu$ y $\mathbb{E}(X_1^4) < \infty$. Entonces

$$\frac{X_1 + \ldots + X_n}{n} \longrightarrow \mu \quad \text{casi seguramente.}$$

Luego, del Teorema 0.12 la cantidad

$$\hat{I}_D = \frac{1}{N}\sum_{i=1}^{N} h(X_i).$$

es un estimador insesgado para I. Para poder obtener una estimación del error de integración y una noción sobre la velocidad de convergencia, recordamos el teorema del límite central.

Teorema 0.12.2 — Teorema del límite central. Sea $X_1, X_2, \ldots$ una secuencia de variables aleatorias independientes e identicamente distribuidas con $\mathbb{E}(X_1) = \mu$ y $\mathbb{V}(X_1) = \sigma^2 < \infty$. Entonces se tiene que

$$\sqrt{N}(\hat{I}_D - \mu) \xrightarrow{d} N\left(0, \frac{\sigma^2}{n}\right).$$

En este contexto, el Teorema 0.12 tiene diversas consecuencias: se obtiene que la varianza de la estimación es σ^2/n, mostrando que la estimación se vuelve certera a medida que n crece; la tasa de convergencia de la estimación es del tipo $\mathcal{O}(n^{-1/2})$, la cual es independiente de la dimensión del problema.

Para estimar el error, o controlar el número total de puntos requeridos, se suele proceder mediante la creación de intervalos de confianza asintóticos. Con esto, es posible tener una aproximación de el total de puntos.

La velocidad de cómputo del método dependerá de tres elementos fundamentales

1. La velocidad para simular la secuencia $\{X_i\}_{i=1}^N$.
2. La el costo computacional requerido para evaluar la función.
3. La tolerancia requerida por el usuario.

Una manera de reducir el número total de simulaciones, es mediante la reducción de la varianza del estimador. Las metodologías que se presentan a continuación tratan de lidiar con este problema.

0.13 Muestreo importado

El Muestreo importado (una traducción del inglés *Importance Sampling*) se basa en el principio de reescribir I mediante

$$I_D = \mathbb{E}_f[h(X)] = \int_D h(x)f(x)\mathrm{d}x = \int_D \frac{h(x)f(x)}{g(x)}g(x)\mathrm{d}x = \mathbb{E}_g\left[\frac{h(x)f(x)}{g(x)}\right], \tag{11}$$

donde g es una función de densidad, llamada función de importación (o *importance function*). Esta nueva forma de ver el problema genera el siguiente estimador insesgado

$$\hat{I}_{D;g} = \frac{1}{N}\sum_{i=1}^{N}\frac{h(X_i)f(X_i)}{g(X_i)} = \frac{1}{N}\sum_{i=1}^{N}h(X_i)w(X_i),$$

donde $w(X_i) = f(X_i)/g(X_i)$. Notamos que $\hat{I}_{D;g}$ es un estimador insesgado para I_D, mientras que la varianza está dada por

$$\mathrm{Var}[\hat{I}_{D;g}] = \int_D \frac{(h(x)f(x) - \mu g(x))^2}{g(x)}\mathrm{d}x. \tag{12}$$

La Ecuación (12) nos permite encontrar características sobre g para minimizar la varianza del estimador. La propuesta trivial sería considerar $g(x) = h(x)f(x)/\mu$ para obtener una varianza de 0, sin embargo, esta elección no es útil en la práctica ya que requiere conocer I_D. Más aún, si h es una función que toma valores negativos, g no sería una función de densidad.

Considerando esta idea, se puede ver que la elección que ayuda a reducir la varianza debe cumplir que $q(x) = c|h(x)|f(x)$, donde c es una constante normalizadora. En efecto, esto se resume en el siguiente teorema

Teorema 0.13.1 Sea $q(x) = c|h(x)|f(x)$ una función de densidad, donde c es la constante normalizadora. Luego, para cualquier densidad de probabilidad φ se cumple que $\text{Var}[\hat{I}_{D;q}] \leq \text{Var}[\hat{I}_{D;\varphi}]$.

Demostración:

$$\begin{aligned}
\text{Var}[\hat{I}_{D;q}] - \mu^2 &= \int_D \frac{(h(x)f(x))^2}{c|h(x)|f(x)} \mathrm{d}x \\
&= \int_D \frac{|h(x)|f(x)}{c} \mathrm{d}x \\
&= \left(\int_D |h(x)|f(x) \mathrm{d}x \right)^2 \\
&= \left(\int_D \frac{|h(x)|f(x)}{\varphi(x)} \varphi(x) \mathrm{d}x \right)^2 \\
&\leq \int_D \left(\frac{|h(x)|f(x)}{\varphi(x)} \right)^2 \varphi(x) \mathrm{d}x \\
&= \int_D \frac{h(x)^2 f(x)^2}{\varphi(x)} \mathrm{d}x \\
&= \text{Var}[\hat{I}_{IS;\varphi}] - \mu^2
\end{aligned}$$

■ **Ejemplo 0.13.2** Aproxime y compare con el valor real la siguiente integral:

$$I = \int_0^\infty x^{-\alpha} e^{-x} \mathrm{d}x$$

■ **Ejemplo 0.13.3** Sea $D = [0,1]^2$. Calcular

$$I_{D,\phi} = \int_D \int_D e^{5 * e^{\|x-y\|/0{,}05}} \mathbb{I}_{\{\|x-y\| \leq \phi\}} \mathrm{d}x\mathrm{d}y$$

para distintos valores de ϕ

0.14 Método de estratificación recursiva

Los métodos de estratificación se basan en la lógica de que las zonas de mayor variabilidad requieren mayor cantidad de observaciones. Esta idea fue propuesta por Press y Farrar (1990) y se detalla a continuación.

Notamos que la integral I_D se puede descomponer en $I_D = I_{D_1} + I_{D_2}$, donde D_1 y D_2 son dos dominios disjuntos, es decir, $D_1 \cap D_2 = \emptyset$ y $D = D_1 \cup D_2$. Luego, si realizamos integración Monte Carlo en cada una de las regiones con $N/2$ puntos, se tiene que un estimador insesgado I_D es

$$\begin{aligned}
\hat{I}_D &= \hat{I}_{D_1} + \hat{I}_{D_2} \\
&= \frac{1}{N/2} \sum_{i=1}^{N/2} h(X_{iD_2}) + \frac{1}{N/2} \sum_{i=1}^{N/2} h(X_{iD_2})
\end{aligned} \tag{13}$$

donde X_{iD} es la i-ésima variable aleatoria en D. De aquí se desprende que la varianza de (13) es

$$\text{Var}[I_D] = (\text{Var}[\hat{I}_{D_1}] + \text{Var}[\hat{I}_{D_2}]) = \frac{2}{N}(\text{Var}_{D_1}[h(X)] + \text{Var}_{D_2}[h(X)])$$

donde $\text{Var}_D[h(X)] = \text{Var}[h(X)\mathbb{I}_D(X)]$. Análogamente, si ahora se realiza una estimacíon de Monte Carlo con N puntos, donde se usarán N_1 en D_1 y N_2 en D_2, entonces se obtiene que

$$\hat{I}_D = \frac{1}{N_1}\sum_{i=1}^{N_1} h(X_{iD_2}) + \frac{1}{N_2}\sum_{i=1}^{N_2} h(X_{iD_2}),$$

y su respectiva varianza,

$$\text{Var}[\hat{I}_D] = \left(\frac{\text{Var}_{D_1}[h(X)]}{N_1} + \frac{\text{Var}_{D_2}[h(X)]}{N - N_1}\right). \tag{14}$$

Luego, nos interesa saber como se deben repartir los puntos N_1 y N_2 de manera tal que la Ecuación (14) alcanze su mínimo. Es posible demostrar que el mínimo se alcanza cuando se cumple la relación

$$\frac{N_1}{N} = \frac{\text{Var}_{D_1}[h(X)]}{\text{Var}_{D_1}[h(X)] + \text{Var}_{D_2}[h(X)]}, \tag{15}$$

por lo cual, N_1/N representa el porcentaje del error que es explicado por los puntos N_1. Entonces, este ratio se nos indica en qué lugar se deberían añadir más puntos, de manera tal que se minimice la varianza del estimador. Dicho esto, estamos listos para presentar el algoritmo:

Entrada: Función $h(x)$ a integrar, una densidad $f(x)$, el dominio de integración, tolerancia ε, y un total de puntos N
Salida: Una estimación de I_D con su error estandar
begin
  $N^* = pN$;
  Dividir el dominio D en dos subdominios disjuntos D_1, D_2;
  Generar $\{X_i\}_{i=1}^{N^*}$ desde una densidad f, sobre el dominio D.;
  Calcular $\hat{\sigma}_{D_1}$ y $\hat{\sigma}_{D_2}$, la varianza del estimador en los dominios;
  Encontrar N_1 y N_2 tales que $N_1 + N_2 = (1-p)N$ y que $N_1/N_2 \approx \hat{\sigma}_{D_2}/\hat{\sigma}_{D_2}$;
  $(\hat{I}_{D1}, \hat{\sigma}_{D_1}) = MISER(D_1, N_1)$;
  $(\hat{I}_{D2}, \hat{\sigma}_{D_2}) = MISER(D_2, N_2)$;
  Resultado: $(\hat{I}_{D1} + \hat{I}_{D2}, \sigma_{D_1} + \sigma_{D_2})$
end

Algorithm 1: Algoritmo de estratificación recursiva (MISER)

La cantidad $\hat{\sigma}_D$ representa un estimador de la desviación estandar $\hat{I}_D$. La estimación clásica se obtiene mediante calcular la raiz cuadrada de $\text{Var}_{D_1}[h(X)]$, sin embargo, esta estimación no es del todo correcta, ya que es a posteriori de la selección de la región. Para corregir este efecto, Press y Farrar (1990) propone modificar la Ecuación (14) mediante una potencia de los tamaños muestrales, es decir

$$\text{Var}[\hat{I}_D] = \left(\frac{\text{Var}_{D_1}[h(X)]}{N_1^\alpha} + \frac{\text{Var}_{D_2}[h(X)]}{N_2^\alpha}\right). \tag{16}$$

Esta modificación trata de compensar el sesgo de estimación de la varianza. El mínimo respecto a N_1 se alcanza cuando

$$N_1 = N\frac{\mathrm{Var}_{D_1}[h(X)]^{\frac{1}{1+\alpha}}}{\mathrm{Var}_{D_1}[h(X)]^{\frac{1}{1+\alpha}} + \mathrm{Var}_{D_2}[h(X)]^{\frac{1}{1+\alpha}}} \tag{17}$$

Por lo tanto, se sugiere modificar las líneas X e Y del algoritmo 2 por las estimaciones provenientes de (16) y (17).

0.15 Integración con árboles de regresión

La ventaja del algoritmo *MISER* se basa en la subdivisión del dominio D en dos dominios que, en principio, minimizarían la varianza. Este procedimiento suele ser caro en la medida que el dominio D sea muy complejo. Para solucionar este problema se propone el uso de árboles de regresión (Ver Breiman (2017)).

Recordar que un árbol de regresión se basa en aproximar una función $h(x)$ mediante

$$\hat{h}(x) = \sum_{i=1}^{m} c_i \mathbb{I}_{R_i}(x), \tag{18}$$

donde la secuencia $c_1, \ldots, c_m$ son números reales y es una secuencia de conjuntos disjuntos tales que $\cup_{i=1}^{m} R_i = D$. es decir, funciones localmente constantes. Si los dominios R_i son conocidos para todo $i = 1, \ldots, m$, entonces el problema es trivial (de hecho, c_i es el promedio de $h(x)$ en R_i). Sin embargo, encontrar estos conjuntos es poco factible y requiere ser tratado con cuidado. Asumiendo que se dispone de una serie de observaciones $\{(x_i, h(x_i))\}_{i=1}^{N}$, y que los conjuntos R_i son hyperplanos que segmentan la coordenada j en dos partes, es decir $R_1(j,s) = \{X|X_j \leq s\}$ y $R_2(j,s) = \{X|X_j > s\}$, entonces se plantea el siguiente problema de minimización

$$\min_{j,s}\left(\min_{c_1}\sum_{x_i \in R_1(j,s)}(h(x)-c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(h(x)-c_2)^2\right). \tag{19}$$

Notamos que el Problema (19) se puede resolver mediante la exploración total del conjunto solución. El cálculo de esto es muy veloz ya que los subproblemas tienen solución explícita (Ver Hastie et al. (2009)).

Volviendo al problema de estimar I_D, donde D es un dominio compacto de $\mathbb{R}^d$, los árboles de regresión permiten mezclar de manera heurística las ideas de *importance sampler* y MISER, es decir, nos acercaremos a una densidad $q(x) = c|h(x)|f(x)$ mediante dividir el dominio D de manera tal que se reduzca la varianza de la estimación. Una manera de desarrollar estas ideas es mediante la aproximación de la función $q(x)$ mediante árboles de regresión normalizados, es decir

$$\hat{q}(x) = \frac{\sum_{i=1}^{m} c_i \mathbb{I}_{R_i}(x)}{\int_D \sum_{i=1}^{m} c_i \mathbb{I}_{R_i}(x)\mathrm{d}x} = \frac{\sum_{i=1}^{m} c_i \mathbb{I}_{R_i}(x)}{\sum_{i=1}^{m} c_i |R_i|} = \sum_{i=1}^{m} c_i w_i \mathbb{I}_{R_i}(x).$$

por lo tanto, es posible realizar una estimación Monte Carlo $\hat{I}_{D;\hat{q}}$ siempre y cuando la estimación de $\hat{q}(x)$ sea razonable. Notamos que se puede simular fácilmente desde la distribución (0.15). De hecho, basta con simular desde una distribución discreta proporcional a c_i, para luego simular una distribución uniforme en el rectángulo $R_i \subset \mathbb{R}^d$.

El único inconveniente que tenemos es sobre las muestras de entrenamiento para comenzar el algoritmo de árbol de regresión. Recordando los resultados de MISER, el dominio de integración es determinante al momento de realizar una estimación Monte Carlo, ya que de este dependerá la varianza. Para solventar dicho problema, se propone el siguiente algoritmo:

Entrada: Función $h(x)$ a integrar, una densidad $f(x)$, un total de iteraciones K, un total de muestras M por iteracion, y el dominio de integración
Salida: Una estimación de I_D con su error estandar
begin
Simular $x_1,\ldots,x_M \sim \text{Unif}(D)$;
Formar el conjunto de datos $D_{train} = \{(x_i, h(x_i))\}_{i=1}^{M}$;
Fijar $D_{train.new} = \emptyset$;
Entrenar un árbol de regresión usando D_{train} ;
Obtener $\hat{q}^0$ de la Ecuación (0.15) ;
for $j = 0$ **to** K **do**
Generar la muestra $x_i^j \sim \hat{q}^j$;
Formar el conjunto de datos $D_{train.new} = \{(x_i^j, h(x_i^j))\}_{i=1}^{M}$;
Crear $D_{train} = D_{train} \cup D_{train.new}$.;
Entrenar un árbol de regresión usando D_{train} ;
Obtener $\hat{q}^{j+1}$ de la Ecuación (0.15) ;
end
Generar $x_1,\ldots,x_{M'} \sim \hat{q}^{j+1}$ Calcular $\hat{I}_{D;\hat{q}^{j+1}}$ **Resultado:** $(\hat{I}_{D;\hat{q}^{j+1}}, \sigma_D)$
end

Algorithm 2: Algoritmo de integración via árboles de regresión iterativos.

■ **Ejemplo 0.15.1** Compare los métodos de integración para $\mathbb{E}(X)$ cuando solo se dispone de la función característica. Para esto se proponen cuatro alternativas:

1. Mediante la fórmula integral de Cauchy.
2. Mediante el algoritmo de Risch.
3. Mediante un método de cuadratura determinista.
4. Mediante integración Monte Carlo.

Considere que la función característica está dada por $\psi(\omega) = 1/(1+(b\omega)^2)$

Algoritmo Esperanza-Maximización

En estadistica, uno de los procesos mas frecuentes es la estimación de un parametro de una distribución de probabilidad. En muchos contextos, no es posible acceder directamente a lo datos necesarios para la estimación, ya sea porque nos encontramos con datos faltantes o incompletos, o porque el modelo produce que no sea posible observar una variable que es necesario para la estimación. Estas dificultades surgen cuando un resultado depende de otros resultados subyacentes, o cuando hay agrupamientos de los datos bajo algun criterio, por ejemplo con datos censurados

En este contexto surgue el algoritmo EM (esperanza-maximización) el cual produce estimaciones a través de un procedimiento iterativo en el cual se estiman los datos perdidos (o no observados) para luego hacer una estimación a traves de máxima verosimilitud. Este algoritmo es muy usado en modelos de mixturas, agrupamiento de datos, psicometría, ingenieria estructural (STRIDE), en reconstrucción de imagenes, etc. Esto pues nos da una forma sencilla y eficaz de realizar estimaciones en problemas que, por naturaleza, contienen datos que no pueden observados.

0.16 Algoritmo EM

Asumamos que tenemos un vector aleatorio X con función de densidad $f(x,\theta)$, donde $\theta = (\theta_1, \theta_2, \ldots, \theta_d)^T$ es un vector de parámetros perteneciente al espacio paramétrico Θ.

Supongamos que tenemos un vector de observaciones x, en donde cierta parte de los datos (que denotaremos como x_{mis}) están perdidos (o no es posible observarlos). Dividiremos entonces nuestro vector de datos completos $x_{\text{com}} = (x_{\text{obs}}^T, x_{\text{mis}}^T)^T$ en donde x_{obs} denota los datos observados y x_{mis} los datos perdidos.

Entonces dado el vector de datos observados x_{obs}, la verosimilitud adopta la forma:

$$\ell_0(\theta, x_{\text{obs}}) = \log f(x_{\text{obs}}; \theta) \tag{20}$$

$$= \log \int f(x_{\text{com}}; \theta) \text{dx}_{\text{mis}} \tag{21}$$

$$= \log \int f(x_{\text{com}} | x_{\text{mis}}; \theta) f(x_{\text{mis}}; \theta) \text{dx}_{\text{mis}}, \tag{22}$$

notemos que la expresión anterior no siempre es fácil de maximizar. Sin embargo tenemos la siguiente igualdad:

$$f(x_{\text{mis}} | x_{\text{obs}}; \theta) = \frac{f(x_{\text{com}}; \theta)}{f(x_{\text{obs}}; \theta)}$$

por lo tanto la función de verosimilitud de los datos observados ℓ_0 se puede escribir de la siguiente manera:

$$\ell_0(x_{\text{obs}}; \theta) = \log f(x_{com}; \theta) - \log f(x_{\text{mis}} | x_{\text{obs}}; \theta). \tag{23}$$

Motivados por la Ecuación (23), es común utilizar la siguiente notación:

$$\ell_c(x_{\text{com}}; \theta) := \log f(x_{\text{com}}; \theta) \quad \text{y} \quad \ell_m(x_{\text{com}}; \theta) := \log f(x_{\text{mis}} | x_{\text{obs}}; \theta),$$

la cual hace mención a la verosimilitud de los datos completos ℓ_c, y a la verosimilitud de los datos perdidos ℓ_m respectivamente.

Multiplicando ambos lados de la Ecuación 23 por $f(x_{\text{mis}} | x_{\text{obs}}; \theta')$ e integrando, se obtiene que:

$$\begin{aligned} \int \ell_0(x_{\text{obs}}; \theta) f(x_{\text{mis}} | x_{\text{obs}}; \theta') \text{dx}_{\text{mis}} &= \int \ell_c(x_{\text{com}}; \theta) f(x_{\text{mis}} | x_{\text{obs}}; \theta') \text{dx}_{\text{mis}} \\ &\quad - \int \ell_m(x_{\text{com}}; \theta) f(x_{\text{mis}} | x_{\text{obs}}; \theta') \text{dx}_{\text{mis}} \\ \Rightarrow \ell_0(x_{\text{obs}}; \theta) &= \mathbb{E}[\ell_c(x_{\text{com}}; \theta) | x_{\text{obs}}; \theta')] - \mathbb{E}[\ell_c(x_{\text{mis}}; \theta) | x_{\text{obs}}; \theta')] \end{aligned}$$

Definiendo las cantidades

$$\begin{aligned} Q(\theta, \theta') &= \mathbb{E}[\ell_c(y_{\text{com}}; \theta) | x_{\text{obs}}; \theta')], \\ H(\theta, \theta') &= \mathbb{E}[\ell_c(y_{\text{mis}}; \theta) | x_{\text{obs}}; \theta')], \end{aligned}$$

tendremos que:

$$\ell_0(x_{\text{obs}}, \theta) = Q(\theta, \theta') - H(\theta, \theta'), \tag{24}$$

para todo θ'.

La Ecuación (24) nos permite ver el problema de maximizar la función de verosimilitud como un problema aumentado y, por lo tanto, diversas técnicas de optimización pueden ser utilizadas. En particular, nos interesa encontrar un algoritmo iterativo que en cada paso permita incrementar la verosimilitud, es decir, si $\theta^{(k)}$ es el vector de parámetros en la iteración k-ésima entonces buscamos un algoritmo tal que:

$$\ell_0(x_{\text{obs}}; \theta^{(k+1)}) \geq \ell_0(x_{\text{obs}}; \theta^{(k)}) \tag{25}$$

para todo $k \in \mathbb{N}$, es decir, (25) es monótona en θ. Evaluando $\theta' = \theta^{(k)}$ en la igualdad (24) tendremos que

$$\ell_0(x_{\text{obs}};\theta^{(k+1)}) - \ell_0(x_{\text{obs}};\theta^{(k)}) = Q(\theta^{(k+1)};\theta^{(k)}) - Q(\theta^{(k)};\theta^{(k)}) - (H(\theta^{(k+1)};\theta^{(k)}) - H(\theta^{(k)};\theta^{(k)})).$$

Notando que:

$$H(\theta^{(k+1)},\theta^{(k)}) - H(\theta^{(k)},\theta^{(k)}) = \mathbb{E}\left[\log\left(\frac{f(x_{\text{mis}}|x_{\text{obs}};\theta^{(k+1)})}{f(x_{\text{mis}}|x_{\text{obs}};\theta^{(k)})}\right)\middle| x_{\text{obs}};\theta^{(k)}\right] \tag{26}$$

$$\leq \log\left(E\left[\frac{f(x_{\text{mis}}|x_{\text{obs}};\theta^{(k+1)})}{f(x_{\text{mis}}|x_{\text{obs}};\theta^{(k)})}\middle| x_{\text{obs}};\theta^{(k)}\right]\right) \tag{27}$$

$$= \log\left(\int f(x_{\text{miss}}|x_{\text{obs}};\theta^{(k+1)})\text{dx}_{\text{miss}}\right) = 0, \tag{28}$$

entonces se puede concluir que, para generar un algoritmo de ascenso que satisfaga:

$$\ell_0(x_{\text{obs}};\theta^{(k+1)}) \geq \ell_0(x_{\text{obs}};\theta^{(k)}),$$

basta con encontrar $\theta^{(k+1)}$ tal que:

$$Q(\theta^{(k+1)};\theta^{(k)}) \geq Q(\theta^{(k)};\theta^{(k)}),$$

en particular podemos tomar:

$$\theta^{k+1} = \arg\max_{\theta} Q(\theta,\theta^{(k)}).$$

Lo anterior, se conoce como *algoritmo de esperanza maximización* (o algoritmo EM) introducido por Dempster, Laird y Rubin (1977) y se resume en el siguiente algoritmo.

A. 15: Algoritmo Esperanza-Maximización

Dado $\Theta^{(k)}$, X_{obs}:

1. **Paso E:** Calcular:

$$Q(\Theta,\Theta^{(k)}) = \mathbb{E}[\ell_c(\Theta;X_{\text{com}})|X_{\text{obs}} = x_{\text{obs}};\Theta^{(k)}]$$

2. **Paso M:** Actualizar $\Theta^{(k+1)}$ como:

$$\Theta^{(k+1)} = \arg\max_{\Theta} Q(\Theta,\Theta^{(k)})$$

Ya conocido el algoritmo, pasaremos a ilustrar su uso con algunos ejemplos:

■ **Ejemplo 0.16.1 — Distribución t multivariada.** La distribución t multivariada tiene muchas aplicaciones en estadística aplicada. Se dice que una variable aleatoria p-dimensional X sigue una distribución t multivariada $t_p(\mu,\Sigma,\nu)$ con posición μ, escala Σ, y con ν grados de libertad, si dado el peso u entonces tenemos:

$$X \mid u \sim N(\mu,\Sigma/u) \tag{29}$$

donde la variable aleatoria latente U correspondiente al peso u tiene una distribución:

$$U \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{30}$$

la función de densidad gamma(α, β) se define de la siguiente manera:

$$f(u;\alpha,\beta) = \left(\beta^{\alpha} u^{\alpha-1}/\Gamma(\alpha)\right) \exp(-\beta u) I_{(0,\infty)}(u), \tag{31}$$

con $\alpha, \beta > 0$. Notar que la función de densidad de X, integrando u en la función de densidad conjunta de X y U que se puede formar a partir de las Ecuaciones (30) y (31), obteniendo:

$$f_p(x;\mu,\Sigma,\nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\Sigma|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma\left(\frac{\nu}{2}\right)\left\{1+\delta(x,\mu;\Sigma)/\nu\right\}^{-\frac{1}{2}(\nu+p)}}, \tag{32}$$

donde:

$$\delta(x,\mu;\Sigma) = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

denota la distancia de Mahalanobis al cuadrado entre x y μ (con Σ matriz de covarianza). A medida que ν tiende a infinito, U converge a uno con probabilidad uno, y por lo tanto, X se converge marginalmente en una distribución normal multivariada con media μ y matriz de covarianza Σ.

Supongamos que queremos estimar con máxima verosimilitud los parametros μ y Σ en la densidad de t (ver la Ecuación 32), en el caso en donde los grados de libertad ν son conocidos.

Por lo tanto asumamos que $x = (x_1, \ldots, x_n)^T$ son observaciones de la distribución $t_p(\mu, \Sigma, \nu)$. Buscamos encontrar el estimador máximo verosímil de θ en base a x, donde $\theta = (\mu^T, \text{veech}(\Sigma)^T)^T$ contiene los elementos de μ y los elementos distintos de Σ. A partir de (32), la función de logverosimilitud para θ con las observaciones x es:

$$\begin{aligned}
\ell_0(\theta, x_{\text{obs}}) &= \sum_{j=1}^{n} \log f_p\left(x_j;\mu,\Sigma,\nu\right) \\
&= -\frac{1}{2} np \log(\pi\nu) + n\left\{\log\Gamma\left(\frac{\nu+p}{2}\right) - \log\Gamma\left(\frac{1}{2}\nu\right)\right\} \\
&- \frac{1}{2} n \log|\Sigma| + \frac{1}{2} n(\nu+p)\log\nu - \frac{1}{2}(\nu+p)\sum_{j=1}^{n} \log\left\{\nu + \delta\left(x_j,\mu;\Sigma\right)\right\}
\end{aligned}$$

la que no admite una solución de forma cerrada.

Dada la Ecuación (29), es conveniente considerar a los datos provenientes de la variable latente U como datos perdidos. El vector de datos completos y se define como:

$$y = \left(x^T, u^T\right)^T$$

donde

$$u = \left(u_1, \ldots, u_n\right)^T.$$

Las variables perdidas $u_1, \ldots, u_n$ se definen tal que:

$$X_j \mid u_j \sim N\left(\mu, \Sigma/u_j\right)$$

sean independientes para $j = 1, \ldots, n$, y además se define $U_1, \ldots U_n$ i.i.d tal que:

$$U_1, \ldots, U_n \sim \text{gamma}\left(\frac{1}{2}\nu, \frac{1}{2}\nu\right).$$

Recalcar que, en este ejemplo, el vector de datos faltantes u consta de observaciones de variables que nunca serían observables como datos en el sentido habitual.

Debido a lo obtenido en las ecuaciones (29) y (30), la función de verosimilitud de los datos completos se puede descomponer en el producto de la densidad condicional de X dado $U = u$ y la densidad marginal de U. Por ende, la logverosimilitud de los datos completos se puede escribir como:

$$\ell_c(\theta) = \ell_{1c}(\theta) + a(z)$$

en donde

$$\begin{aligned} \ell L_{1c}(\Theta) = &-\frac{1}{2}np\log(2\pi) - \frac{1}{2}n\log|\Sigma| + \frac{1}{2}p\sum_{j=1}^{n}\log u_j \\ &-\frac{1}{2}\sum_{j=1}^{n} u_j\,(x_j-\mu)^T\,\Sigma^{-1}\,(x_j-\mu) \end{aligned} \tag{33}$$

y

$$\begin{aligned} a(u) = &-n\log\Gamma\left(\frac{1}{2}\nu\right) + \frac{1}{2}n\nu\log\left(\frac{1}{2}\nu\right) \\ &+\frac{1}{2}\nu\sum_{j=1}^{n}(\log u_j - u_j) - \sum_{j=1}^{n}\log u_j. \end{aligned} \tag{34}$$

Teniendo ya calculada la verosimilitud de los datos compuestos podemos pasar a calcular los pasos E y M del algoritmo.

Para el paso E en la $(k+1)$-ésima iteración del algoritmo EM, necesitamos calcular: $Q\left(\theta;\theta^{(k)}\right)$. Como ν es conocido, solo necesitamos enfocarnos en el primer término $\ell_{1c}(\theta)$ en la expresión (33) para $\ell_c(\theta)$ (ya que los otros términos no tienen parámetros desconocidos). Dado que este término es lineal en los datos no observables u_j, el paso E se efectúa simplemente reemplazando u_j con su esperanza condicional con respecto a $x_j, \theta^{(k)}$.

Se puede demostrar que la distribución de U dado $X = x$ es:

$$U|X = x \sim \text{Gamma}(m_1, m_2) \tag{35}$$

en donde

$$m_1 = \frac{1}{2}(\nu + p)$$

y

$$m_2 = \frac{1}{2}\{\nu + \delta(w,\mu;\Sigma)\} \tag{36}$$

entonces tendremos que:

$$E(U \mid X = x) = \frac{\nu + p}{\nu + \delta(x,\mu;\Sigma)} \tag{37}$$

y por lo tanto:

$$E\left[U_j \mid x_j;\Theta^{(k)}\right] = u_j^{(k)}$$

en donde:

$$u_j^{(k)} = \frac{\nu + p}{\nu + \delta\left(x_j, \mu^{(k)}; \Sigma^{(k)}\right)}.$$

Para obtener el paso M, notemos que $ell_{1c}(\theta)$ corresponde a la función de verosimilitud formada por n observaciones independientes $x_1,\ldots,x_n$ con media común μ y matrices de covarianza $\Sigma/u_1,\ldots,\Sigma/u_n$, respectivamente. Después de realizar el paso E, cada u_j se reemplaza por $u_j^{(k)}$. Por lo tanto el paso M es equivalente a calcular la media ponderada y la matriz de covarianza muestral de $x_1,\ldots,x_n$ con pesos $u_1^{(k)},\ldots,u_n^{(k)}$. Teniendo entonces que

$$\mu^{(k+1)} = \frac{\sum_{j=1}^n u_j^{(k)} x_j}{\sum_{j=1}^n u_j^{(k)}}$$

y

$$\Sigma^{(k+1)} = \frac{1}{n}\sum_{j=1}^n u_j^{(k)}\left(x_j - \mu^{(k+1)}\right)\left(x_j - \mu^{(k+1)}\right)^T.$$

En este caso de ν conocido, se puede observar que el algoritmo EM es equivalente a mínimos cuadrados ponderados iterativos. El paso E actualiza los pesos $u_j^{(k)}$, mientras que el paso M elige de manera efectiva $\mu^{(k+1)}$ y $\Sigma^{(k+1)}$ mediante una estimación de mínimos cuadrados ponderados.

■ **Ejemplo 0.16.2 — Modelo de Mixturas.** Una de las aplicaciones del algoritmo EM es la estimación de máxima verosimilitud de modelos de mixturas finitas, ya trabajamos con un modelo de mixturas en 0.11.2 para una mixtura de 2 normales, sin embargo en este ejemplo consideraremos modelos de mixtura con una cantidad arbitraria finita de variables aleatorias. Un modelo de mixtura finita está compuesto por k modelos componentes cuyos parámetros son $\theta_1,\ldots,\theta_K$, y una distribución sobre estos componentes, denotada por p. La función de densidad de probabilidad de este modelo de mixtura finita se expresa como:

$$f(x \mid \theta) = \sum_{i=1}^M p_i f_i(x \mid \theta_i)$$

Donde $\theta = (p,\theta_1,\ldots,\theta_i)$ denota todos los parámetros del modelo de mixtura y f_i es la función de densidad de un modelo componente, el cual tiene parámetro θ_i. Cuando cada modelo componente es un modelo gaussiano, se denomina modelo de mixturas gaussiana.

La simulación de una muestra x a partir de un modelo de mixtura finita se puede realizar en dos pasos:

- Simulamos una variable indicatriz Z de la distribución p es decir $Z \sim p(p_1,p_2,\ldots,p_M)$, donde Z puede tomar valores en $\{1,2,3\ldots,M\}$. Este valor indica de que componente generaremos la variable aleatoria X.
- Dado $Z = z$ simular X de la z-ésima componente.

Dado un conjunto de observaciones $x = \left\{x^1, \ldots, x^n\right\}$, podemos tratar el conjunto asociado de indicatrices $z = \left\{z^1, \ldots, z^n\right\}$ como variables latentes. Para estimar los parámetros del modelo, podemos utilizar el algoritmo EM.

Para obtener los pasos E y M, notemos que la logverosimilitud de datos completos es el logaritmo de la función de densidad conjunta de las variables observadas y latentes, dados los parámetros del modelo. En un modelo de mixtura finita, se expresa como:

$$\log(f(x,z;\theta)) = \log\left(\prod_{j=1}^{n} f\left(x^j|z^j;\theta\right) f\left(z^j;\theta\right)\right)$$

donde, $f\left(x^j|z^j;\theta\right) = f_{z^j}\left(x^j;\theta_{z^j}\right)$ y $f\left(z^j;\theta\right) = p_{z^j}$. Por lo tanto, la sustitución de estos valores en la fórmula anterior resulta en:

$$\log f(x,z \mid \theta) = \sum_{j=1}^{n} \left(\log(p_{z^j}) + \log\left(f_{z^j}\left(x^j \mid \theta_{z^j}\right)\right)\right)$$

esta forma nos dificultara obtener el paso M, por lo tanto se introduce un función indicatriz δ_i:

$$\delta_i(z) = \begin{cases} 1 & (z = i) \\ 0 & (z \neq i) \end{cases}$$

con esta notación podemos reescribir la función de verosimilitud como:

$$\log f(x,z|\theta) = \sum_{j=1}^{n}\sum_{i=1}^{M} \delta_j\left(z^j\right)\left(\log(p_i) + \log\left(f_i\left(x^j \mid \theta_i\right)\right)\right).$$

Con esta función de verosimilitud podemos obtener los pasos E y M.

Primeramente para el paso E, dado θ y x^j, notemos que la indicatriz z^j es independiente de las otras observaciones, por lo tanto, tendremos que:

$$f(z|x;\theta) = \prod_{j=1}^{n} f(z^j|x^j;\theta)$$

Usando el teorema de Bayes podemos obtener que:

$$f(z^j = i|x^j;\theta) = \frac{p_i f_i(x^j|\theta_i)}{\sum_{i=l}^{M} p_l f_l(x^j|\theta_l)}$$

Entonces:

$$f(z|x;\theta^{(k)}) = \prod_{j=1}^{n} f(z^j|x^j;\theta^{(k)})$$

Entonces tenemos que Q toma la siguiente forma:

$$\begin{aligned} Q(\theta,\theta^{(k)}) &= \mathbb{E}\left[\sum_{j=1}^{n}\sum_{i=1}^{M} \delta_i\left(z^j\right)\left(\log p_i + \log f_i\left(x^j \mid \theta_i\right)\right)|x;\theta^{(k)}\right] \\ &= \sum_{i=1}^{n}\sum_{i=1}^{M} \mathbb{E}\left[\delta_i\left(z^j\right)\left(\log p_i + \log f_i\left(x^j \mid \theta_i\right)\right)|x;\theta^{(k)}\right]. \end{aligned}$$

Donde en cada termino

$$\mathbb{E}\left[\delta_i\left(z^j\right)\left(\log p_i+\log f_i\left(x^j|\theta_i\right)\right)|x;\theta^{(k)}\right]=\mathbb{E}\left[\delta_i\left(z^j\right)|x;\theta^{(i)}\right]\cdot\left(\log p_i+\log f_i\left(x^j\mid\theta_i\right)\right)$$

y el valor de $\delta_i\left(z^j\right)$ puede ser solo 0 o 1 , entonces tenemos:

$$\mathbb{E}\left[\delta_i\left(z^j\right)|X=x;\theta^{(k)}\right]=f(i|x^i;\theta^{(k)})$$

Y por lo tanto tendremos que:

$$Q(\theta,\theta^{(k)})=\sum_{j=1}^{n}\sum_{i=1}^{M}f(z^j|x^j;\theta^{(k)})(\log p_i+\log f_i(x^j|\theta_i))$$

Notemos que para el caso M podemos calcular $\theta_i^{(k)}$ de la siguiente manera:

$$\theta_i^{(k+1)}=\arg\max_{\theta}\sum_{j=1}^{n}f(i|x^j;\theta^{(k)})\log(f_i(x^j|\theta_i))$$

En cambio para estimar los p_i tendremos que resolver el siguiente problema:

$$p_i^{(k+1)}=\arg\max_{p}\sum_{j=1}^{n}f(i|x^j;\theta^{(k)})\log(p_i)$$
$$\text{s.a}\sum_{i=1}^{M}p_i=1,\ \ p_i\geq 0\text{ para todo }i$$

Teniendo entonces que la solución esta dada por:

$$p_i^{(k+1)}=\frac{1}{n}\sum_{j=1}^{n}f(i|x^j;\theta^{(k)}))$$

0.16.1 Algoritmo EM para la familia exponencial

Existe un caso especial para el algoritmo EM en donde es posible obtener una forma simplificada de los pasos E y M, este caso ocurre si la función de densidad pertenece a la familia exponencial.

La función de densidad de probabilidad de los datos completos $f_c(x;\theta)$ pertenece a la familia exponencial si se cumple:

$$f_c(x;\theta)=b(x)\exp\left\{c^T(\theta)t(x)\right\}/a(\theta),\tag{38}$$

donde el estadístico suficiente $t(x)$ es un vector de tamaño $k\times 1$ ($k\geq d$) y $c(\theta)$ es una función vectorial de tamaño $k\times 1$ que depende del vector de parámetros θ de tamaño $d\times 1$. Además, $a(\theta)$ y $b(x)$ son funciones escalares. El espacio parámetrico Θ es un conjunto convexo de dimensión d tal que la ecuación (38) define una función de densidad de probabilidad para todos los θ en Θ:

$$\Theta=\left\{\theta:\int b(x)\exp\left\{c^T(\theta)t(x)\right\}dx<\infty\right\}.\tag{39}$$

Si $k=d$ y el Jacobiano de $c(\theta)$ tiene rango completo, se dice que $f_c(x;\theta)$ pertenece a una familia exponencial regular. El coeficiente $c(\theta)$ del estadístico suficiente $t(x)$ en la Ecuación (38) se conoce

como parámetro natural o canónico. Por lo tanto, si la función de densidad de los datos completos $f_c(x;\theta)$ pertenece a una familia exponencial regular en su forma canónica, entonces:

$$f_c(x;\theta) = b(x)\exp\left\{\theta^T t(x)\right\}/a(\theta). \tag{40}$$

El parámetro θ en (40) es único salvo transformaciones lineal no singulares de tamaño $d \times d$, al igual que la elección correspondiente de $t(x)$. La esperanza del estadístico suficiente $t(X)$ en (40) se calcula como:

$$\mathbb{E}\left[t(X);\theta\right] = \partial \log a(\theta)/\partial\theta. \tag{41}$$

Al tomar la esperanza condicional de $\ell_c(\theta)$ dado y, podemos expresar $Q\left(\theta;\theta^{(k)}\right)$ como:

$$Q\left(\theta;\theta^{(k)}\right) = \theta^T t^{(k)} - \log a(\theta) \tag{42}$$

donde $t^{(k)} = \mathbb{E}\left[t(X) \mid Y = y;\theta^{(k)}\right]$ y $\theta^{(k)}$ representa el valor actual de θ.

Al diferenciar 42 con respecto a θ y usando (41), se sigue que el paso M requiere escoger $\theta^{(k+1)}$ resolviendo la ecuación:

$$E\left[t(X);\theta\right] = t^{(k)} \tag{43}$$

Si la ecuación (43) se puede resolver para $\theta^{(k+1)}$ dentro del espacio parámetrico Θ, entonces la solución es única debido a la convexidad del logaritmo negativo de la verosimilitud en la familia exponencial regular. En casos donde la ecuación no tiene solución, el máximo para $\theta^{(k+1)}$ de θ se encuentra en la frontera de Θ.

■ **Ejemplo 0.16.3 — Normal bivariada con datos faltantes.** Sea $X = (X_1, X_2)^T$ un vector bivariado con distribución normal:

$$X \sim N(\mu, \Sigma)$$

con media $\mu = (\mu_1, \mu_2)^T$ y matriz de covarianzas Σ dada por:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}. \tag{44}$$

La densidad de una normal bivariada esta dada por:

$$\phi(x;\theta) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\right\}.$$

donde el vector de parámetros θ esta dado por:

$$\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$$

Supongamos que deseamos encontrar el estimador máximo verosímil de θ, basandonos en una muestra aleatoria de tamaño n tomada de X, donde los datos de la i-ésima variable X_i están perdidos en m_i observaciones $i \in \{1,2\}$. Denotaremos los datos completamente observados de la siguiente manera: $x_j = (x_{1j}, x_{2j})^T$ con $j \in \{1,\ldots,m\}$, donde $m = n - m_1 - m_2$, $x_{:2j}$, con $j \in \{m+1,\ldots,m+m_1\}$ representa las m_1 observaciones con los valores de la primera variable x_{1j} perdidos, y x_{1j} con $j = \{m+m_1+1,\ldots,n\}$ representa las m_2 observaciones con los valores de la segunda variable x_{2j} perdidos.

Se supone la perdida de datos puede considerarse como completamente aleatoria, de modo que los datos observados pueden ser considerados como una muestra aleatoria de tamaño m de la distribución normal bivariada, y un par de muestras aleatorias independientes de tamaño m_i de las distribuciones normales univariadas:

$$X_i \sim N(\mu_i, \sigma_{ii})$$

para $i \in \{1,2\}$. Por ende, los datos observados están dados por:

$$y = \left(x_1^T, \ldots, x_m^T, v^T\right)^T$$

donde el vector v esta dado por:

$$v = \left(x_{2,m+1}, \ldots, x_{2,m+m_1}, x_{1,m+m_1+1}, \ldots, x_{1,n}\right)^T.$$

La función de logverosimilitud para Θ con respecto a los datos observados y es:

$$\log L(\Theta) = -n\log(2\pi) - \frac{1}{2}m\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{m}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) - \frac{1}{2}\sum_{i=1}^{2}m_i\log\sigma_{ii}$$
$$-\frac{1}{2}\left\{\sigma_{11}^{-1}\sum_{j=m+m_1+1}^{n}(x_{1j}-\mu_1)^2 + \sigma_{22}^{-1}\sum_{j=m+1}^{m+m_1}(x_{2j}-\mu_2)^2\right\}.$$

Una elección para los datos completos aquí es tomar un vector de n observaciones bivariadas. El vector de datos completos x entonces estará dado por:

$$x = \left(x_1^T, \ldots, x_n^T\right)^T$$

en donde el vector de datos perdidos z es:

$$z = \left(x_{1,m+1}, \ldots, x_{1,m+m_1}, x_{2,m+m_1+1}, \ldots, x_{2,n}\right)^T.$$

La función de log verosimilitud para los datos completos θ es:

$$\begin{aligned}\ell_c(\theta) &= -n\log(2\pi) - \frac{1}{2}n\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\\ &= -n\log(2\pi) - \frac{1}{2}n\log\xi\\ &\quad -\frac{1}{2}\xi^{-1}\left[\sigma_{22}T_{11} + \sigma_{11}T_{22} - 2\sigma_{12}T_{12}\right.\\ &\quad -2\left\{T_1(\mu_1\sigma_{22} - \mu_2\sigma_{12}) + T_2(\mu_2\sigma_{11} - \mu_1\sigma_{12})\right\}\\ &\quad \left.+n\left(\mu_1^2\sigma_{22} + \mu_2^2\sigma_{11} - 2\mu_1\mu_2\sigma_{12}\right)\right] \qquad (45)\end{aligned}$$

en donde

$$T_i = \sum_{j=1}^{n}x_{ij} \quad (i=1,2),$$

$$T_{hi} = \sum_{j=1}^{n}x_{hj}x_{ij} \quad (h,i=1,2),$$

y donde

$$\xi = \sigma_{11}\sigma_{22}\left(1-\rho^2\right)$$

y, además:

$$\rho = \sigma_{12}/\left(\sigma_{11}\sigma_{22}\right)^{\frac{1}{2}}$$

es la correlación entre X_1 y X_2.

Se puede observar que $L_c(\theta)$ pertenece a la familia exponencial con estadístico suficiente:

$$T = \left(T_1, T_2, T_{11}, T_{12}, T_{22}\right)^T.$$

Si el vector de datos completos x estuviera disponible, entonces el estimador máximo verosímil de los datos completos de $\theta, \hat{\theta}$, puede calcularse de forma directa. Según los resultados para la estimación usando el método de máxima verosimilitud (datos completos) para la distribución normal bivariada, $\hat{\theta}$ se obtiene que:

$$\begin{aligned}
&\hat{\mu}_i = T_i/n \quad (i = 1,2),\\
&\hat{\sigma}_{hi} = \left(T_{hi} - n^{-1}T_hT_i\right)/n \quad h \in \{1,2\}.
\end{aligned}$$

Ahora consideramos el paso E en la $(k+1)$-ésima iteración del algoritmo EM, donde $\theta^{(k)}$ denota el valor de θ después de la k-ésima iteración del algoritmo EM. Se puede ver a partir de (45) que, para calcular la esperanza condicional actual de la log-verosimilitud de los datos completos:

$$Q\left(\theta;\theta^k\right) = E\left[\ell_c(\theta) \mid Y = y;\theta^{(k)}\right]$$

necesitamos las esperanzas condicionales actuales de los estadísticos suficientes T_i y $T_{hi}(h,i = 1,2)$. Por lo tanto, necesitamos:

$$E\left[X_{1j} \mid X_{2j} = x_{2j};\theta^{(k)}\right]$$

y

$$E\left[X_{1j}^2 \mid X_{2j} = x_{2j};\theta^{(k)}\right]$$

para $j = m+1,\ldots,m+m_1$. Y, además necesitamos:

$$E\left[X_{2j} \mid X_{1j} = x_{1j};\theta^{(k)}\right]$$

y

$$E\left[X_{2j}^2 \mid X_{1j} = x_{1j};\theta^{(k)}\right]$$

para $j = m+1,\ldots,m+m_1$.

A partir de las propiedades de la distribución normal bivariada, la distribución condicional de X_2 dado $X_1 = x_1$ es una normal con media

$$\mu_2 + \sigma_{12}\sigma_{11}^{-1}\left(x_1 - \mu_1\right)$$

y varianza

$$\sigma_{22,1} = \sigma_{22}\left(1-\rho^2\right)$$

Por lo tanto,

$$E\left[X_{2j} \mid X_{1j} = x_{1j}; \theta^{(k)}\right] = x_{2j}^{(k)} \tag{46}$$

donde

$$x_{2j}^{(k)} = \mu_2^{(k)} + \left(\sigma_{12}^{(k)}/\sigma_{11}^{(k)}\right)\left(x_{1j} - \mu_1^{(k)}\right), \tag{47}$$

y

$$E\left[X_{2j}^2 \mid X_1 j = x_{1j}; \theta^{(k)}\right] = \left(x_{2j}^{(k)}\right)^2 + \sigma_{22,1}^{(k)} \tag{48}$$

para $j = m + m_1 + 1, \ldots, n$. De manera similar, $E\left[X_{1j} \mid X_{2j} = x_{2j}; \theta^{(k)}\right]$ y $E\left[X_{1j}^2 \mid X_{2j} = x_{2j}; \theta^{(k)}\right]$ se obtienen intercambiando los subíndices 1 y 2 en las ecuaciones (47) y (48).

Cabe destacar que si simplemente imputáramos los $x_{2j}^{(k)}$ para los x_{2j} faltantes en la verosimilitud de los datos completos (y de manera similar para los x_{1j} faltantes), no obtendríamos la misma expresión que la Q-función generada por el paso E, debido a la omisión del término $\sigma_{22,1}^{(k)}$ en el lado derecho de 48.

El paso M en la $(k+1)$-ésima iteración se implementa simplemente reemplazando T_i y T_{hi} por $T_i^{(k)}$ y $T_{hi}^{(k)}$, respectivamente, donde estos últimos están definidos por la sustitución de los valores x_{ij} y x_{ij}^2 con sus experanzas condicionales actuales dadas por (47) y (48) para $i = 2$ y por sus formas correspondientes para $i = 1$. En consecuencia, $\theta^{(k+1)}$ esta dado por:

$$\begin{aligned} \mu_i^{(k+1)} &= T_i^{(k)}/n & i &\in \{1,2\} \\ \sigma_{hi}^{(k+1)} &= (T_{hi}^{(k)} - n^{-1} T_h^{(k)} T_i^{(k)})/n & h &\in \{1,2\} \end{aligned}$$

0.16.2 Algoritmo GEM

A menudo, la solución del paso M existe y es posible calcularla (verMcLachlan y Krishnan (2007):página 28) . Sin embargo, en los casos que no exista o en donde no sea factible encontrar el valor de θ que maximice la función $Q(\theta, \theta^{(k)})$, basta ver que se cumpla la propiedad de la monotonía 25 y solo es necesario lograr que: $Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)})$. Dado lo anterior Dempster, Laird y Rubin (1977) definió un algoritmo EM generalizado (algoritmo GEM) para estos casos, el cual se ilustra a continuación:

A. 16: Algoritmo Esperanza-Maximización generalizado

Dado $\theta^{(k)}$, x_{obs}:

1. **Paso E:** Calcular:

$$Q(\theta, \theta^{(k)}) = \mathbb{E}[\ell_c(\Theta; x_{\text{com}}) | X_{\text{obs}} = x_{\text{obs}}; \theta^{(k)}]$$

2. **Paso M:** Seleccionar θ^{k+1} satisfaciendo:

$$Q(\theta^{(k+1)}, \theta^{(k)}) > Q(\theta^{(k)}, \theta^{(k)})$$

0.17 Convergencia algoritmo EM

Como lo demostramos en la sección 0.16 para toda iteración $k+1$-esima del algoritmo EM tendremos que:

$$\ell_0(x_{\text{obs}};\theta^{k+1}) \geq \ell_0(x_{\text{obs}};\theta^k) \tag{49}$$

Por lo tanto podemos tomar al algoritmo EM como un algoritmo iterativo que toma una dirección de ascenso de la función ℓ_0. Por lo tanto si la función ℓ_0 esta acotada, entonces la función $\{\ell_0(x_{\text{obs}};\theta^k)\}$ converge a algún valor ℓ_0^*. La mayoria de las veces, ℓ_0^* sera un punto fijo del algoritmo. Es decir, $\ell_0^* = L(\theta^*)$ para algún θ^* tal que:

$$\left.\frac{\partial \ell_0(x_{\text{obs}};\theta)}{\partial \theta}\right|_{\theta=\theta^*} = 0$$

Es más, en la practica, la mayoría de las veces ℓ_0^* sera un máximo local. En los casos en que la sucesión $\{\theta^{(k)}\}$ se mantiene constante para algún punto fijo θ^* que no es un máximo global ni local de ℓ_0 (por ejemplo un punto silla), basta tomar una pequeña perturbación de θ con respecto al punto fijo θ^k para que esto cause que el algoritmo EM diverja de este punto fijo. En general, si ℓ_0 tiene muchos puntos fijos, la convergencia del algoritmo EM dependerá de la elección del punto inicial $\theta^{(0)}$.

A pesar de lo anterior, el algoritmo EM puede converger teóricamente a un punto fijo que no sea un máximo local ni global del algoritmo, más específicamente, puede converger a un punto silla. Para eso sea θ^* un punto silla del algoritmo. Notemos que diferenciando la ecuación 24 tendremos:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\partial Q(\theta;\theta^k)}{\partial \theta} - \frac{H(\theta;\theta^k)}{\partial \theta}$$

De la desigualdad 26:

$$H(\theta;\theta^k) \leq H(\theta^k;\theta^k) \tag{50}$$

Para todo $\theta \in \Theta$, entonces:

$$\left.\frac{\partial H(\theta;\theta^k)}{\partial \theta}\right|_{\theta=\theta^k} = 0 \tag{51}$$

Sea $\theta_0 \in \Theta$ arbitrario de θ. Entonces si $\theta^k = \theta_0$ entonces 51:

$$\frac{\partial \ell(\theta_0)}{\partial \theta} = \left.\frac{\partial Q(\theta,\theta_0)}{\partial \theta}\right|_{\theta=\theta_0} \tag{52}$$

donde $\dfrac{\partial \ell(\theta_0)}{\partial \theta}$ denota $\dfrac{\partial l(\theta)}{\partial \theta}$ evaluado en θ_0. Supongamos que $\theta = \theta^*$ donde θ^* es un punto fijo de $\ell(\theta)$. Entonces tenemos que:

$$\frac{\partial \ell(\theta^*)}{\partial \theta} = \left.\frac{\partial Q(\theta,\theta^*)}{\partial \theta}\right|_{\theta=\theta^*}$$

Supongamos que $\theta = \theta^*$, donde θ^* es un punto fijo de ℓ. Entonces tendremos que:

$$\begin{aligned}\frac{\partial \ell(\theta^*)}{\partial \theta} &= \left.\frac{\partial Q(\theta,\theta^*)}{\partial \theta}\right|_{\theta=\theta^*} \\ &= 0\end{aligned}$$

Por lo tanto, podemos ver que el algoritmo EM puede converger a un punto silla θ^* si $Q(\theta,\theta^*)$ es maximizado globalmente sobre θ en θ^*.

Para lograr la convergencia a un máximo, Wu (1983) propuso la condición:

$$\sup_{\theta \in \Theta} Q(\theta, \theta^*) > Q(\theta^*; \theta^*) \tag{53}$$

para todo punto fijo θ^* que no es un máximo local de de $\ell_0(x_{\text{obs}}; \theta)$.

Esta condición en conjunto con la condiciones de regularidad que se darán en la próxima subsección asegurará que todos los puntos límite de cualquier instancia del algoritmo EM son máximos locales de $L(\theta)$ y que $L(\theta^k)$ converge monótonamente a $L^* = L(\theta^*)$ para algún máximo local θ^*. Sin embargo la condición (53) suele ser difícil de verificar.

0.17.1 Condiciones de Regularidad de Wu

En el algoritmo EM (A.15), el paso M involucra el mapeo de un punto a un conjunto, pues, el paso M requiere obtener:

$$\mathscr{M}(\theta^{(k)}) = \arg\max_{\theta} Q(\theta; \theta^{(k)})$$

en donde $\mathscr{M}(\theta^{(k)})$ es un conjunto con los valores de θ que maximizan $Q(\theta, \theta^{(k)})$.

Para el algoritmo GEM, el mapeo $\mathscr{M}(\theta^{(k)})$ esta definido por la elección de un $\theta^{(k+1)}$ tal que:

$$Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)})$$

A través de los resultados existentes en la literatura de optimización para mapeos de un punto a un conjunto, Wu (1983) estableció condiciones que aseguran la convergencia de la sucesión $\{\ell_0(\theta^{(k)})\}$ a un valor estacionario $L(\theta)$.

Lo primero que necesitamos es que se cumpla la propiedad que definiremos a continuación:

Definición 0.17.1 — Mapeo Cerrado. Un mapeo $\mathscr{M}$ es cerrado en θ_0 si para toda sucesión $\{\theta_m\}_{m \in \mathbb{N}} \subset \Theta$ tal que $\theta_m \to \theta_0$ y para toda sucesión $\{\varphi_m\}_{n \in \mathbb{N}}$ con $\varphi_m \in \mathscr{M}(\theta_m)$ φ_m converge a φ_0 entonces esto implica que $\varphi_0 \in \mathscr{M}(\theta_0)$.

Además de estas condiciones, Wu (1983) supuso las siguientes condiciones:

$$\Theta \text{ es un subconjunto de } \mathbb{R}^d. \tag{54}$$

$$\Theta_{\theta_0} = \{\theta \in \Theta : \ell(\theta) \geq \ell(\theta_0)\} \text{ es compacto para todo } L(\theta_0) > -\infty. \tag{55}$$

$$L(\theta) \text{ es continua en } \Theta \text{ y diferenciable en el interior de } \Theta. \tag{56}$$

Desde ahora denominaremos a las condiciones (54), (55) y (56), como condiciones de regularidad de Wu, o simplemente condiciones de Wu.

Una consecuencia de las condiciones de Wu es que toda sucesión $\{\ell(\theta^{(k)})\}$ esta acotada superiormente para todo punto inicial $\theta^{(0)} \in \Theta$, donde, se asume que el punto inicial satisface que $L(\theta^{(0)}) > -\infty$.

Como señala Wu (1983), suponer la compacidad en (55) puede ser restrictiva cuando el espacio paramétrico Θ no es compacto.

Por otro lado, en lo que resta de esta sección, asumiremos también que para todo $\theta^{(k)}$ está en el interior de Θ. Es decir, $\theta^{(k+1)}$ es una solución de la ecuación:

$$\partial Q\left(\theta; \theta^{(k)}\right) / \partial \theta = 0.$$

Wu (1983) señala que esta condición puede ser una implicancia de la siguiente condición:

$$\Theta_{\theta_o} \text{ está en el interior de } \Theta \text{ para cualquier } \theta_o \in \Theta. .$$

0.17.2 Teorema de Convergencia para el algoritmo GEM

Ahora se presentara, sin demostración, el principal teorema de convergencia dado por Wu (1983) para el algoritmo GEM, el cual también se aplica al algoritmo EM, ya que este último es un caso especial de un algoritmo GEM.

Teorema 0.17.1 *Sea* $\left\{\Theta^{(k)}\right\}$ *una instancia de un algoritmo GEM generada por* $\theta^{(k+1)} \in \mathscr{M}\left(\theta^{(k)}\right)$. *Supongamos que se cumplen las condiciones de reguralidad, y además se cumple que:*

i) $\mathscr{M}\left(\theta^{(k)}\right)$ *es cerrado sobre el complemento de S, donde S es el conjunto de puntos estacionarios en el interior de* Θ.

ii) *Se cumple que:*

$$L\left(\theta^{(k+1)}\right) > L\left(\theta^{(k)}\right) \text{ para todo } \theta^{(k)} \notin \mathscr{S}$$

Entonces, todos los puntos límite de $\left\{\theta^{(k)}\right\}$ *son puntos estacionarios y* $L\left(\theta^{(k)}\right)$ *converge monótonamente a* $L^* = L(\theta^*)$ *para algún punto estacionario* $\theta^* \in S$.

La condición ii) se cumple para una sucesión EM. Consideremos un $\theta^{(k)} \notin S$. Entonces, gracias a la ecuación (52), tendremos:

$$\begin{aligned}\left[\partial Q\left(\theta;\theta^{(k)}\right)/\partial\theta\right]_{\theta=\theta^{(k)}} &= \partial \log L\left(\theta^{(k)}\right)/\partial\theta \\ &\neq \mathbf{0},\end{aligned}$$

ya que $\theta^{(k)} \notin S$. Por lo tanto, $Q\left(\theta;\theta^{(k)}\right)$ no se maximiza en $\theta = \theta^{(k)}$, y entonces, por la definición del paso M del algoritmo EM (A.15) y por la condición en (53) tendremos:

$$Q\left(\theta^{(k+1)};\theta^{(k)}\right) > Q\left(\theta^{(k)};\theta^{(k)}\right)$$

Esto implica que:

$$L\left(\theta^{(k+1)}\right) > L\left(\theta^{(k)}\right)$$

mostrando que se cumple la condición ii) del Teorema 0.17.1.

Para obtener que el mapeo $\mathscr{M}(\theta)$ es cerrado, Wu (1983) dió la siguiente condición suficiente:

$$Q(\theta;\Psi) \text{ sea continua tanto en } \theta \text{ como en } \Psi. \tag{57}$$

La cual se denomina condición de continuidad.

Esta condición es muy débil y debería cumplirse en la mayoría de las situaciones prácticas. Por ejemplo, se ha demostrado que se cumple para el caso de una familia exponencial curvada:

$$g_c(x;\theta) = b(x)\exp\left\{\theta^T t(x)\right\}/a(\theta),$$

donde θ se encuentra en una subvariedad compacta Θ_0 de la región Θ de dimensión d , la cual se define en (39).

Esto conduce al siguiente teorema de Wu (1983) para una sucesión generada por el algoritmo EM.

0.17.3 Convergencia del algoritmo EM

Teorema 0.17.2 *Supongamos que $Q(\theta;\Psi)$ cumple la condición de continuidad* (57) *y las condiciones de regularidad de Wu. Entonces, todos los puntos límite de cualquier instancia $\{\theta^{(k)}\}$ del algoritmo EM son puntos estacionarios de $L(\theta)$, y $L(\theta^{(k)})$ converge monótonamente a algún valor $L^* = L(\theta^*)$ para algún punto estacionario θ^*.*

El Teorema anterior es un consecuencia del Teorema 0.17.1 ya que la condición (i) se deduce de la suposición de continuidad (57), mientras que la condición (ii) se cumple siempre para una sucesión generada por el algoritmo EM.

Según Wu (1983) , el Teorema 0.17.1 es el resultado más general para algoritmos EM y GEM. Sin embargo, desde el punto de vista del usuario, el Teorema 0.17.2 proporciona el resultado mas útil, pues solo requiere condiciones que son fáciles de verificar.

0.18 Versiones MonteCarlo del Algoritmo EM

En el algoritmo EM, el paso E puede ser difícil de implementar debido a la dificultad para calcular la esperanza de la log-verosimilitud. Para solucionar esto Wei y Tanner (1990a) y Wei y Tanner (1990b) propusieron un enfoque de Monte Carlo mediante la simulación de los datos faltantes z a partir de la distribución condicional $f(x_{\text{mis}}|x_{\text{obs}};\theta^{(k)})$ en el paso E de la $k+1$-ésima iteración, y luego maximizar la aproximación de la esperanza condicional de la log-verosimilitud de los datos completos, es decir, definir una estimación para Q de tal manera:

$$\hat{Q}(\theta,\theta^{(k)}) = \frac{1}{m}\sum_{j=1}^{m} \ell_c(x_{\text{obs}}, x_{\text{mis}};\theta) \tag{58}$$

Tomando el límite cuando $m \to \infty$ de $\hat{Q}(\theta;\theta^{(k)})$ tendremos que $\hat{Q}(\cdot,\cdot)$ converge a $\hat{Q}(\cdot,\cdot)$. El algoritmo resultante de remplazar Q por $\hat{Q}$ se conoce como algoritmo MCEM y se resume en lo siguiente:

A. 17: Algoritmo MCEM

Dado $\theta^{(k)}$, x_{obs}:

1. **Simulación:** Generar $z_1,\ldots,z_M \sim f(x_{\text{mis}}|x_{\text{obs}};\theta^{(k)})$ y formar el conjunto de datos completos $\{x_1^{(k+1)},\ldots,x_m^{(k+1)}\}$ en donde
$$x_i^{(k+1)} = (x_{\text{obs}}, z_i).$$
2. **Esperanza:** Definir
$$\hat{Q}(\theta;\theta^{(k)}) = \frac{1}{m}\sum_{j=1}^{m} \ell_c(x_j^{(k+1)};\theta^{(k-1)})$$
3. **Maximización** Actualizar $\theta^{(k+1)}$:
$$\theta^{(k+1)} = \arg\max_{\theta} \hat{Q}(\theta,\theta^{(k)}).$$

Aunque la maximización de (58) a menudo puede ser difícil, existen situaciones, como en el caso de la familia exponencial, donde hay soluciones en forma cerrada para el problema de maximización.

Para reducir los costos de la simulación de los datos faltantes, se pueden reutilizar simulaciones de pasos anteriores (ver Levine y George Casella (2001)). Cuando es difícil simular el paso E, se puede usar el algoritmo Gibbs Sampler (ver Chan y Ledolter (1995)), lo cual ilustraremos con un ejemplo en esta sección.

El Algoritmo A.17 , introduce un error de Monte Carlo en el paso E, además de que pierde la propiedad de la monotonía (25). Sin embargo, en ciertos casos, el algoritmo se acerca a un maximizador con una alta probabilidad, Booth y Hobert (1999). Los problemas de especificar m y monitorear la convergencia son de importancia central en el uso del algoritmo. Wei y Tanner (1990a) recomiendan que se utilicen valores pequeños de m en las etapas iniciales y que los valores se incrementen a medida que el algoritmo se acerca a la convergencia. En cuanto al monitoreo de la convergencia, recomiendan que se grafiquen los valores de $\theta^{(k)}$ en función de k y si la convergencia al valor $\hat{\theta}$ se mantiene, el proceso de estimación puede terminarse; de lo contrario, se debe continuar el proceso con un valor mayor de m. Booth y Hobert (1999) y McCulloch (1997) propusieron métodos alternativos para especificar m y obtener reglas de parada.

A continuación, presentaremos un ejemplo del algoritmo MCEM para una normal con datos censurados.

■ **Ejemplo 0.18.1 — Normal con datos censurados.** Supongamos que $x_1,\ldots,x_n$ es una muestra aleatoria de la distribución $N(\mu,1)$. Además asumamos que los datos $x_{m+1},\ldots,x_n$ están censurados por la derecha en c, es decir para $i \in \{m+1,\ldots,n\}$ solo sabemos que $x_i \geq c$ pero no conocemos sus valores reales. Denotemos por $x = (x_1,\ldots,x_m,x_{m+1},\ldots,x_n)^T$ al vector de datos completos y por

$$z = (x_{m+1},\ldots,x_n)^T$$

al vector que contiene los datos faltantes. Además, sea $\bar{x}_{\text{obs}}$ la media de las m observaciones no censuradas. La función de verosimilitud de los datos completos para $\theta = \mu$ (salvo constantes) es

$$\ell_c(\mu) = -\sum_{j=1}^{m} \frac{(x_j-\mu)^2}{2} - \sum_{j=m+1}^{n} \frac{(x_j-\mu)^2}{2}.$$

La densidad de los datos faltantes z es un producto de normales truncadas y, por lo tanto,

$$\ell_m(x;\mu) \propto \sum_{j=m+1}^{n} \frac{(x_j-\mu)^2}{2}.$$

En el paso E, se calcula la esperanza de la verosimilitud condicional de los datos completos para obtener (salvo constantes aditivas):

$$Q\left(\mu;\mu^{(k)}\right) = -\frac{1}{2}\left\{\sum_{j=1}^{m}(x_j-\mu)^2 + \sum_{j=m+1}^{n} E\left[(X_j-\mu)^2 \mid X_j > c;\mu^{(k)}\right]\right\}$$

lo que conduce en el paso M a la estimación actualizada de μ,

$$\mu^{(k+1)} = \frac{m\bar{x}_{\text{obs}} + (n-m)E[X \mid X > c;\mu^{(k)}]}{n} \tag{59}$$

al sustituir la esperanza condicional actual de X en (59), obtenemos

$$\mu^{(k+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\mu^{(k)} + \frac{1}{n}\frac{\phi\left(c-\mu^{(k)}\right)}{1-\Phi\left(c-\mu^{(k)}\right)}$$

donde ϕ es la función de densidad de una distribución normal estándar y Φ es la función de distribución de una normal estándar.

La solución del EM con Monte Carlo en este ejemplo es reemplazar $E[X \mid X > c; \mu^{(k)}]$ por

$$\frac{1}{m}\sum_{j=1}^{m} x_j$$

donde x_j se genera a partir de la densidad normal truncada (en c) con media $\mu^{(k)}$ y varianza unitaria.

0.18.1 Cadenas de Makov para el algoritmo EM con Montecarlo

En situaciones en las que no es posible realizar una simulación directa de la distribución objetivo, podemos recurrir a las técnicas de simulación Monte Carlo con cadenas de Markov que se presentaron en el capítulo 4. Para ilustrar cómo utilizar estos métodos previamente vistos, presentaremos dos ejemplos. En el primero utilizaremos el algoritmo Metropolis-Hastings, mientras que en el segundo emplearemos el algoritmo Gibbs Sampler.

■ **Ejemplo 0.18.2** En el modelo lineal mixto generalizado (GLMM), el algoritmo EM presenta dificultades debido a que la etapa E es intratable incluso bajo suposiciones gaussianas de los efectos aleatorios. Para solucionar este problema Vaida y Meng (2005) propusieron implementar un paso E con Monte Carlo a través de una técnica de Monte Carlo con cadenas de Markov.

El modelo GLMM se resume en lo siguiente, asumamos que tenemos un vector Y de variables aleatorias a predecir, además de un vector U de variables de efectos aleatorios, tal que los Y_i condicionados a los U_i son independientes y pertenecen a la familia exponencial:

$$Y_j|u \sim \text{indep} f_j(y_j|u) \quad (j = 1,\ldots,n)$$

es decir, los Y_j son condicionalmente independientes dados los u_j, donde:

$$f_j(y_j|u_j) = \exp[\kappa^{-1}(\theta_j y_j - b(\theta_j)) + c(y_j;\kappa)]$$

Denotaremos por $\mu_j = E[y_j|u]$. Asumamos que tenemos un conjunto de variables regresoras X tal que tenemos el siguiente predictor lineal:

$$\nu_j = x_j^T\beta + z_j^T\mu_j$$

además tenemos una función $h(\cdot)$ que permite enlazar nuestro predictor lineal con μ_j de la siguiente manera:

$$h(\mu_j) = \nu_j = x_j^T\beta + z_j^T u_j$$

y por ultimo tenemos que la función de distribución de U tiene a D como parámetros del modelo, es decir:

$$U \sim f(u|D)$$

Resumiendo, tenemos el siguiente modelo:

$$\begin{aligned}
Y_j|u &\sim \text{indep} f_j(y_j|u) \quad (j = 1,\ldots,n)\\
f_j(y_j \mid u_j) &= \exp\left[\kappa^{-1}\left\{\theta_j y_j - b(\theta_j)\right\} + c(y_j;\kappa)\right],\\
E(y_j \mid u) &= \mu_j,\\
h(\mu_j) &= x_j^T\beta + z_j^T u_j,\\
u &\sim f(u \mid D),
\end{aligned}$$

Nótese que z_j es un vector de diseño y no denota datos faltantes.

El vector de datos completos estará dado por $\left(y^T, u^T\right)^T$, donde $y = (y_1, \ldots, y_n)^T$ y $u = (u_1, \ldots, u_n)^T$. A pesar que $\theta = (\theta_1, \ldots, \theta_n)$ forma parte de los parámetros del modelo y son desconocidos, nuestro objetivo sera estimar los parametros $\Psi = (\beta, \kappa, D)$ sin necesidad de estimar θ. La distribución de los datos completos se puede escribir a través de la densidad condicional de la siguiente manera: $f(y,u) = f(u)f(y \mid u)$, de modo que la log-verosimilitud de los datos completos es:

$$\begin{aligned}
\ell_c(\Psi) &= \log f(y \mid u; \beta) + \log f(u; D) \\
&= \sum_{j=1}^{n} \log f\left(y_j \mid u; \beta\right) + \log f(u; D) \\
&= \kappa^{-1} \left\{ \sum_{j=1}^{n} \theta_j y_j - b(\theta_j) \right\} + \sum_{j=1}^{n} \log c(y_j; \kappa) + \log f(u; D).
\end{aligned}$$

Esto se debe a que, los y_j, condicionados a u, son independientes. Luego, β y κ aparecen en la primera parte (la parte del modelo lineal generalizado) y D aparece en la segunda parte. La maximización se puede realizar por separado, la primera parte con respecto a una ecuación que es bastante común en problemas de GLM y la segunda parte similar a la máxima verosimilitud con respecto a $f(u; D)$, lo cual puede ser simple si f pertenece a la familia exponencial. Así, una iteración del algoritmo EM se reduce a:

1. Encontrar $\beta^{(k+1)}$ y $\kappa^{(k+1)}$ para maximizar $E[\log f(y \mid u; \beta, \kappa) \mid y; \Psi^{(k)}]$,
2. Encontrar $D^{(k+1)}$ para maximizar $E\left[\log f\left(u \mid D^{(k)}\right) \mid y; \Psi^{(k)}\right]$.

La ventaja de este procedimiento es evitar el cálculo de la verosimilitud (f_Y) y conformarse con la distribución condicional. Aun así, no siempre es fácil derivar expresiones analíticas para las dos expectativas involucradas. En tales casos, podemos recurrir a métodos de Monte Carlo tanto para estimar estas integrales (expectativas) como para maximizar. Por lo tanto, el algoritmo consiste en obtener una muestra aleatoria $u^{(1)}, \ldots, u^{(M)}$ de la distribución condicional y actualizar los parámetros β, κ, D mediante:

1. Calcular los valores actualizados $\beta^{(k+1)}$ y $\kappa^{(k+1)}$ para maximizar la estimación de Monte Carlo

$$\frac{1}{M} \sum_{m=1}^{M} \log f\left(y \mid u^{(m)}; \beta, \kappa\right)$$

de

$$E[\log f(y \mid u; \beta, \kappa) \mid y; \Psi^{(k)}]$$

2. Calcular el valor $D^{(k+1)}$ actualizado que maximice:

$$\frac{1}{m} \sum_{j=1}^{m} \log f(u^{(m)} | D)$$

El algoritmo Metropolis-Hastings A.4 se utiliza para muestrear de la distribución condicional de U. En el algoritmo Metropolis-Hastings, si elegimos la distribución instrumental como la densidad marginal $f(u)$ de U, entonces el ratio de aceptación sera de la forma:

$$\frac{f(u^* \mid y,\beta,\kappa,D)\,f(u)}{f(u \mid y,\beta,\kappa,D)f(u^*)} = \frac{\prod_{j=1}^{n} f(y_j \mid u^*,\beta,\kappa)\,f(u^* \mid D)\,f(u \mid D)}{\prod_{j=1}^{n} f(y_j \mid u,\beta,\kappa)\,f(u \mid D)f(u^* \mid D)}$$
$$= \frac{\prod_{j=1}^{n} f(y_j \mid u^*,\beta,\kappa)}{\prod_{j=1}^{n} f(y_j \mid u,\beta,\kappa)}$$

Notese que este cálculo solo involucra la distribución condicional de Y dada u.

Cabe señalar que existen varias formas de utilizar MCMC en estos contextos. Por ejemplo McCulloch (1994) y McCulloch (1997) utilizan el Algoritmo Gibbs Sampler (A.12) para modelos probit y el Algoritmo Metropolis-Hastings (A.4) para GLMM, respectivamente.

■ **Ejemplo 0.18.3 — Gibbs Sampler para Mixturas Gausianas.** El algoritmo Gibbs Sampler (A.12) se utiliza ampliamente en muchos problemas bayesianos donde la distribución conjunta es complicada y difícil de manejar, pero las distribuciones condicionales suelen ser lo suficientemente fáciles de simular. Como es en el caso del modelo de mixturas finitas bayesiano que presentaremos a continuación:

Consideremos una versión bayesiana del problema de mixtura finita que discutimos en el Ejemplo 0.16.2 aquí ψ_i denotara el vector de parámetros de la densidad del componente $f_i(x;\psi_i)$ en una mixtura de la forma:

$$f(x;\theta) = \sum_{i=1}^{g} \pi_i f_i(x;\psi_i),$$

entonces la totalidad de los parámetros para el problema es:

$$\theta = \left(\psi^T, \pi^T\right)^T$$

donde $\psi = \left(\psi_1^T, \ldots, \psi_g^T\right)^T$ y $\pi = (\pi_1, \ldots, \pi_{g-1})^T$. Si las densidades de los componentes tienen parámetros comunes, entonces ψ es el vector de esos parámetros conocidos a priori como distintos. Dado un conjunto de datos observados $x_{\text{obs}} = \left(x_1^T, \ldots, x_n^T\right)^T$ de la distribución de mixtura y una densidad a priori $p(\theta)$ para θ, la densidad a posteriori para θ,

$$p(\theta \mid x_{\text{obs}}) \propto \prod_{j=1}^{n} \left(\sum_{i=1}^{g} \pi_i f_i(x_j;\psi_i) \right) p(\theta),$$

la cual es intratable (ver Titterington et al. (1985)). Sin embargo, supongamos que formulamos un problema de datos completos como en el ejemplo 0.16.2 al introducir el vector de datos faltantes $z = \left(z_1^T, \ldots, z_n^T\right)^T$, donde z_j es el vector que contiene las variables indicatrices binarias (cero-uno) que definen la pertenencia del componente de cada x_j. Luego, los componentes del modelo condicional en un marco bayesiano son

$$p(x_j \mid z_j,\theta) = \prod_{i=1}^{g} p_i^{(z_j)_i}(x_j;\psi_i)\, p(z_j \mid \pi)$$

donde

$$p(z_j \mid \pi) = \prod_{i=1}^{g} \pi_i^{(z_j)_i}$$

Si se toma $p(\theta)$ como el producto de una distribución a priori Dirichlet para π y una distribución a priori independiente para θ, entonces las distribuciones condicionales de $(\pi \mid \pi, z), (\psi \mid \pi, z)$, y $(z \mid \psi, \pi)$ están definidas y son conocidas, lo que permite utilizar un procedimiento de Gibbs Sampler para obtener estimaciones de una distribución a posteriori que de otra manera sería intratable.

0.18.2 EM Monte Carlo con Newton Rapson

En los casos donde no se pueda calcular $\arg\max_\theta \hat{Q}_m(\theta, \theta^{(k)})$ de forma explicita, podemos usar el algoritmo Newton Rapson para calcular este valor de manera iterativa. Para ello notemos que:

$$\frac{\partial \hat{Q}(\theta;\theta^{(k)})}{\partial \theta} = \frac{1}{m}\sum_{j=1}^{M} \frac{\partial}{\partial \theta}\ell_c(x_{\text{obs}}, x_{\text{mis}};\theta)$$
$$\frac{\partial^2 \hat{Q}(\theta;\theta^{(k)})}{\partial \theta \partial \theta^T} = \frac{1}{m}\sum_{j=1}^{m} \frac{\partial^2}{\partial \theta \partial \theta^T}\ell_c(x_{\text{obs}}, x_{\text{mis}};\theta)$$

Entonces, en cada iteración del algoritmo Newton-Rapson tendremos que:

$$(\theta^{(k)})^{(r+1)} = (\theta^{(k)})^{(r)} + \left(-\frac{\partial^2 \hat{Q}((\theta^{(k)})^{(r)}, \theta^{(k)})}{,}\partial\theta\partial\theta^T\right)^{-1} \frac{\partial \hat{Q}((\theta^{(k)})^{(r)}, \theta^{(k))}}{\partial \theta}$$

Al momento de que la iteración en (0.18.2) converja a un valor θ^* definimos $\theta^{(k+1)} = \theta^*$.

A pesar de la eficacia de este método para calcular $\arg\max_\theta \hat{Q}_m(\theta, \theta^{(k)})$ aumentamos la complejidad del algoritmo al calcular el arg max con un algoritmo iterativo.

0.19 Otras variantes Estocasticas

El algoritmo EM con Monte Carlo presenta desafíos relacionados con la complejidad computacional asociada a la simulación en el paso E, así como la dificultad para realizar la maximización en el paso M la cual no siempre se puede hacer de manera directa. Estos factores contribuyen a un aumento en la complejidad computacional del algoritmo. Para abordar estas complicaciones se han propuesto varias variantes del algoritmo EM con Monte Carlo. A continuación, nos centraremos en presentar dos de estas variantes.

0.19.1 Algoritmo MH-RM

Este algoritmo esta fuertemente motivado por la Identidad de Fisher, pues en el caso del algoritmo EM toma una forma particular. Para ilustrar lo anterior denotemos el gradiente de la log-verosimilitud de los datos completos de la siguiente manera:

$$s(\theta; x_{\text{com}}) = \nabla_\theta \ell_c(x_{\text{com}};\theta).$$

Entonces por la identidad de Fisher tendremos la siguiente igualdad:

$$\nabla_\theta \ell_0(x_{\text{obs}};\theta) = \int s(x_{\text{com}},\theta) f(x_{\text{mis}}|x_{\text{obs}};\theta)\text{dx}_{\text{mis}} \tag{60}$$

Notemos que la Ecuación (60) sugiere que, se puede optimizar $\ell_0(x_{\text{obs}};\theta)$ sin evaluar directamente su gradiente. En su lugar, las direcciones de ascenso están dadas por la esperanza condicional del gradiente de los datos completos $\mathbf{s}(x_{\text{com}};\theta)$. Una solución que sea igual a cero para el lado derecho de la Ecuación (60) también satisface las ecuaciones de verosimilitud y maximiza la función $\ell_0(x_{\text{obs}};\theta)$. La conexión central radica en tomar la esperanza de $\mathbf{s}(x_{\text{com}};\theta)$ con respecto a la distribución condicional de x_{mis} dado x_{obs} . Dado que θ es desconocido y $f(x_{\text{mis}}|x_{\text{obs}};\theta)$ depende de θ, la solución solo se puede obtener de forma iterativa. El algoritmo MH-RM no es más que una formalización de esta idea.

Para obtener la solución de la ecuación anterior usamos el algoritmo de Robbins y Monro (1951). El algoritmo Robbins y Monro (1951) es un algoritmo para encontrar raíces de funciones de regresión afectadas por un ruido aleatorio. En el caso más simple, sea $g(\cdot)$ una función de θ a valores reales. Si $g(\cdot)$ es conocida y es continua entonces podríamos usar el algoritmo de Newton-Rapson:

$$\theta_{k+1} = \theta_k + \left[-\nabla_\theta g\left(\theta_k\right)\right]^{-1} g\left(\theta_k\right)$$

para encontrar su raíz. Alternativamente, si no se puede asumir diferenciabilidad, se puede utilizar la siguiente aproximación:

$$\theta_{k+1} = \theta_k + \gamma g\left(\theta_k\right)$$

en una vecindad de la raíz si γ es suficientemente pequeño. Ahora supongamos que $g(\theta)$ solo se puede medir de manera imprecisa como $g(\theta)+\zeta$, donde ζ es una variable aleatoria de media cero que representa el ruido aleatorio. Esta es la situación original con la que Robbins y Monro (1951) estaban tratando. El método de Robbins-Monro actualiza de forma iterativa la aproximación a la raíz de acuerdo al siguiente esquema recursivo:

$$\theta_{k+1} = \theta_k + \gamma_k R_{k+1} \tag{61}$$

donde $R_{k+1} = g\left(\theta_k\right) + \zeta_{k+1}$ es una estimación de $g\left(\theta_k\right)$ y $\{\gamma_k; k \geq 1\}$ es una sucesión de constantes de ganancia tales que:

$$\gamma_k \in (0,1], \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \quad \text{y} \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

Si juntamos las tres condiciones obtendremos que las constantes de ganancia disminuiran lentamente a cero. El atractivo de este algoritmo es que R_{k+1} no tiene que ser muy preciso. Esto se puede entender de la siguiente manera: si θ_k todavía está lejos de la raíz, tomar un gran número de observaciones para calcular una buena estimación de $g\left(\theta_k\right)$ es ineficiente porque R_{k+1} es útil en la medida en que proporciona la dirección correcta para el siguiente movimiento. Las constantes de ganancia decrecientes eventualmente eliminaran el efecto del ruido para que la sucesión de estimaciones converja a la raíz.

En el algoritmo MH-RM, se extiende el algoritmo básico de la Ecuación (61) a problemas de múltiples parámetros que involucran la ampliación estocástica de datos faltantes. Sea

$$\mathbf{H}(\theta \mid x_{\text{com}}) = -\frac{\partial^2 \ell_c(x_{\text{com}}; \theta)}{\partial \theta \partial \theta'}$$

la matriz de información de los datos completos, y sea $P(\cdot, A \mid x_{\text{obs}}; \theta)$ un kernel de transición de una cadena de Markov tal que, para todo θ y cualquier conjunto medible A, genera una cadena uniformemente ergódica que tiene a $f(x_{\text{mis}} \mid x_{\text{obs}}; \theta)$ como su distribución invariante, de modo que:

$$\int_A f(x_{\text{mis}}|x_{\text{obs}}; \theta) = \int f(x_{\text{mis}}|x_{\text{obs}}; \theta) P(x_{\text{mis}}, A \mid x_{\text{obs}}; \theta.$$

Lo anterior permite definir el algoritmo MH-RM de la siguiente manera:

A. 18: Algoritmo MH-RM

Dado $\theta^{(k)}$, x_{obs} y Γ_k

1. **Simulación:** Generar $z_1, \ldots, z_M \sim f(x_{\text{mis}}|x_{\text{obs}}; \theta^{(k)})$ y formar el conjunto de datos completos $\{x_1^{(k+1)}, \ldots, (x_m^{(k+1)}\}$ en donde

$$x_i^{(k+1)} = (x_{\text{obs}}, z_i).$$

2. **Aproximación:** Aproximar $\nabla_\theta \ell_0(x_{\text{obs}}; \theta^{(k)})$ y la matriz de datos completos, de la siguiente manera:

$$s_{k+1} = \frac{1}{m} \sum_{j=1}^{m} s(x_j^{(k+1)}; \theta^{(k)})$$

$$\Gamma_{k+1} = \Gamma_k + \gamma_k \left(\frac{1}{m} \sum_{j=1}^{m} H_0 \left(x_j^{(k)}; \theta^{(k)} \right) \right)$$

3. **Actualización de Robbins-Monro:** Actualizar θ:

$$\theta^{(k+1)} = \theta^{(k)} + \gamma_k \Gamma_{k+1}^{-1} s_{k+1}.$$

En la práctica, γ_k se puede tomar como $1/k$, en cuyo caso la elección de Γ_0 es arbitraria. Se puede demostrar que, bajo ciertas condiciones de regularidad, el algoritmo MH-RM converge a un máximo local de $\ell_0(x_{\text{obs}}; \theta)$ con probabilidad uno (ver Cai (2010)). Aunque el tamaño de la simulación m_k puede depender del número de iteración k, no es en absoluto necesario. El resultado de convergencia muestra que el algoritmo converge con un tamaño de simulación fijo y relativamente pequeño, es decir, $m_k \equiv m$, para todo k.

0.19.2 Algoritmo SAEM

Notemos que en los algoritmos anteriores, en cada iteración se debe simular todo un conjunto de datos perdidos y se descartan todos los datos perdidos simulados durante las iteraciones anteriores, lo cual provoca que nuestro algoritmo no aproveche las simulaciones realizadas anteriormente. Para solucionar este problema, Delyon, Lavielle y Moulines (1999) idearon un algoritmo usando aproximaciones estocásticas para los valores de $Q(\theta; \theta^{(k)})$ las cuales dependieran siempre de todos los datos simulados anteriormente, definiendo así el siguiente algoritmo:

A. 19: Algoritmo SAEM

Dado $\theta^{(k-1)}$, $\theta^{(k-1)}$, x_{obs}, $\hat{Q}(\theta, \theta^{(k-1)})$

1. **Simulacion:** Generar $z_1, \ldots, z_M \sim f(x_{\text{mis}}|x_{\text{obs}}; \theta^{(k)})$ y formar el conjunto de datos completos $\{x_1^{(k+1)}, \ldots, (x_m^{(k+1)}\}$ en donde

$$x_i^{(k+1)} = (x_{\text{obs}}, z_i).$$

2. **Aproximación:** Actualizar $\hat{Q}(\theta; \theta^{(k)})$ por

$$\hat{Q}(\theta; \theta^{(k)}) = \hat{Q}(\theta; \theta^{(k-1)}) + \gamma_k \left(\frac{1}{m} \sum_{j=1}^{m} \ell_c(x_j^{(k+1)}; \theta^{(k-1)}) - \hat{Q}(\theta; \theta^{(k-1)} \right)$$

3. **Actualización** Actualizar θ:

$$\theta^{(k+1)} = \arg\max_{\theta} \hat{Q}(\theta, \theta^{(k)})$$

Tecnicas de Remuestreo

Pueden existir múltiples dificultades a las que podemos enfrentarnos al utilizar metodos estadísticos paramétricos y no paramétricos. Por ejemplo, el tamaño de nuestra muestra puede ser bastante pequeño por lo cual no podemos usar el teorema del limite central para hacer intervalos de confianza asintoticos. Es posible que estemos interesados en un estadistico del cual no conocemos su distribución por lo cual no podemos calcular los valores como, intervalos de confianza, valores p, valores críticos, etc.

Para solucionar lo anterior se utilizan métodos de remuestreo, que consisten en en generar nuevas muestras llamadas submuestras utilizando nuestra muestra obtenida. Alguna de las ventajas de estos métodos es que son sencillos de comprender e implementar, a menudo se pueden utilizar aunque no se conozca la distribución, y muchas veces conducen a estimadores mas robustos.

En esta sección nos centraremos en dos métodos de remuestreos muy comunes en la bibliografia: Jacknife y Bootstrap. Primeramente expondremos el metodo de Jacknife el cual genera submuestras de una población eliminando un elemento de la muestra. Este método se creo para obtener la varianza asintótica de un estimador, pero actualmente se utiliza para calcular el valor de otras cantidades de interés, además de generar un nuevo estimador que es mas robusto que el usual. En segundo lugar explicaremos Bootstrap el cual es una técnica que expande la idea de Jacknife al obtener submuestras de tamaño arbitrario de nuestra muestral principal, y al igual que Jacknife, es bastante útil para calcular cantidades asociadas a nuestro estimador.

0.20 Jacknife

Si $\widehat{\theta}$ es un estimador basado en una muestra i.i.d $y_1,\ldots,y_n$, entonces denotaremos por $\widehat{\theta}_{[i]}$ el estimador que obtenemos al estimar θ eliminando y_i de la muestra. Denotaremos el promedio de las estimaciónes que se obtienen al eliminar un dato por $\bar{\theta}_1 = n^{-1}\sum_{i=1}^{n}\widehat{\theta}_{[i]}$ y definimos los pseudo-valores por:

$$\widehat{\theta}_{ps,i} = n\widehat{\theta} - (n-1)\widehat{\theta}_{[i]} \tag{62}$$

La intuición de esta idea es que el pseudo-valor es la parte de $\widehat{\theta}$ que depende de y_i, es decir, se resta $(n-1)\widehat{\theta}_{[i]}$, que no depende de y_i, de $n\widehat{\theta}$, dejando solo la parte de $\widehat{\theta}$ que depende de y_i. El promedio de estos pseudo-valores es el estimador de Jackknife ajustado del sesgo, es decir:

$$\begin{aligned}\widehat{\theta}_{\text{J}} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{ps,i} = n\widehat{\theta} - (n-1)\bar{\theta}_1 \\ &= \widehat{\theta} - (n-1)\left(\bar{\theta}_1 - \widehat{\theta}\right).\end{aligned}$$

Estimación del Sesgo

El estimador de Jackknife del sesgo es $(n-1)\left(\bar{\theta}_1 - \widehat{\theta}\right)$, y el estimador $\widehat{\theta}_{\text{J}}$ es el estimador de Jackknife ajustado del sesgo. Asuamamos que se cumple la siguiente igualdad para $\widehat{\theta}$:

$$\text{E}[\widehat{\theta}] = \theta + \frac{\beta_1}{n} + \frac{\beta_2}{n^2} + O\left(n^{-3}\right) \tag{63}$$

Ahora, escribamos $\widehat{\theta}$ y $\bar{\theta}_1$ como variables de respuesta en un modelo de regresión lineal simple, obtenido después de agrupar los términos de orden superior en un termino de error:

$$\widehat{\theta} = \theta + \frac{\beta_1}{n} + e_1, \quad \bar{\theta}_1 = \theta + \frac{\beta_1}{n-1} + e_2.$$

En este modelo de regresión, el intercepto es $= \theta$, la pendiente es $= \beta_1$ y el predictor es el inverso del tamaño de la muestra, n^{-1} y $(n-1)^{-1}$. Usando los dos puntos $(x_1, y_1) = (1/n, \widehat{\theta})$ y $(x_2, y_2) = \left(1/(n-1), \bar{\theta}_1\right)$, el segmento de línea que los conecta tiene una pendiente

$$\widehat{\beta}_1 = \frac{y_2 - y_1}{x_2 - x_1} = n(n-1)\left(\bar{\theta}_1 - \widehat{\theta}\right)$$

y el intercepto es

$$\frac{y_1 x_2 - y_2 x_1}{x_2 - x_1} = \widehat{\theta} - (n-1)\left(\bar{\theta}_1 - \widehat{\theta}\right) = \widehat{\theta} - \frac{\widehat{\beta}_1}{n} = \widehat{\theta}_{\text{J}}$$

como vimos anteriormente, también es el promedio de los pseudo-valores. Aquí podemos ver que $\widehat{\theta}_{\text{J}} = \widehat{\theta} - \frac{\widehat{\beta}_1}{n}$ es un estimador corregido del sesgo, donde $\frac{\widehat{\beta}_1}{n} = (n-1)\left(\bar{\theta}_1 - \widehat{\theta}\right)$ es el sesgo estimado restado de $\widehat{\theta}$.

No es difícil de mostrar que, usando (63), tenemos:

$$E[\widehat{\theta}_J] = \theta - \frac{\beta_2}{n^2} + O(n^{-2})$$

y así el sesgo de θ se ha reducido de $O\left(n^{-1}\right)$ a $O\left(n^{-2}\right)$. Es posible ampliar este enfoque dejando de lado dos o más observaciones a la vez, para mas detalles ver Schucany, Gray y Owen (1971), pero aparentemente esto no se usa con frecuencia en la práctica, Bradley Efron (1982). De hecho, incluso el estimador corregido del sesgo de primer orden $\widehat{\theta}_{\text{J}}$ puede no ser una mejora sobre $\widehat{\theta}$ en términos del error cuadrático medio Kim y Singh (1998), por lo cual suele no usarse en la práctica. Por otro lado, el estimador de varianza de Jackknife es un método práctico importante utilizado ampliamente en el muestreo de encuestas y en otros contextos. Por lo tanto, ahora nos enfocaremos en este método.

Estimación de Varianza de Jackknife

El estimador de varianza de Jackknife para $\widehat{\theta}$ se puede definir en términos de las varianzas muestrales de los $\widehat{\theta}_{[i]}$ o de los pseudo-valores (62) de la siguiente manera:

$$\widehat{V}_{\mathrm{J}} = \frac{(n-1)^2}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{\theta}_{[i]} - \bar{\theta}_{1} \right)^2 = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{\theta}_{ps,i} - \widehat{\theta}_{\mathrm{J}} \right)^2 .$$

Para entender $\widehat{V}_{\mathrm{J}}$, examinemos los pseudo-valores. En el caso de que $\widehat{\theta} = \bar{Y}$, los pseudo-valores son simplemente los valores de la muestra en sí, y el estimador de varianza de Jackknife es exactamente el mismo que la estimación estándar de la varianza s^2_{n-1}/n. (Con s^2_{n-1} la varianza muestral con $n-1$ en el divisor). Más generalmente, tendremos:

$$\widehat{\theta}_{ps,i} - \widehat{\theta} = (n-1) \left(\widehat{\theta} - \widehat{\theta}_{[i]} \right) = \frac{\widehat{\theta}_{[i]} - \widehat{\theta}}{-\frac{1}{n-1}}.$$

A modo de ejemplo, vamos a considerar la media muestral, y la varianza muestral. Notar que la media muestral es un estadístico lineal de los datos, mientras que la varianza muestral es un estadístico no lineal de los datos.

■ **Ejemplo 0.20.1 — Media muestral.** Para la media muestral $\widehat{\theta} = \widehat{\mu} = \bar{y}$, los estimadores dejando la muestra i-ésima afuera, son $\widehat{\theta}_{[i]} = (n\bar{y} - y_i)/(n-1)$, y los pseudo-valores son $n\bar{y} - (n\bar{y} - y_i) = y_i$, las propias observaciones. Por lo tanto, $\widehat{V}_{\mathrm{J}} = s^2_{n-1}/n$.

■ **Ejemplo 0.20.2 — Varianza Muestral.** Para la varianza muestral $\widehat{\theta} = s^2_n = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$, los estimadores eliminando el i-esimo dato son:

$$\begin{aligned}
\widehat{\theta}_{[i]} &= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} (Y_j - \bar{y}_{n-1,i})^2 \\
&= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left\{ y_j - \bar{y} - (\bar{y}_{n-1,i} - \bar{y}) \right\}^2 \\
&= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} (y_j - \bar{y})^2 - (\bar{y}_{n-1,i} - \bar{y})^2 \\
&= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} (y_j - \bar{y})^2 - \frac{1}{(n-1)^2} (\bar{y} - y_i)^2 ,
\end{aligned}$$

donde hemos usado que $\bar{y}_{[i]} - \bar{y} = (n-1)^{-1} (\bar{y} - y_i)$ en el último paso. Luego, los pseudovalores son:

$$\begin{aligned}
\widehat{\theta}_{ps,i} = n\widehat{\theta} - (n-1)\widehat{\theta}_{[i]} &= \sum_{j=1}^{n} (y_j - \bar{y})^2 - \sum_{\substack{j=1 \\ j \neq i}}^{n} (y_j - \bar{y})^2 + \frac{1}{n-1} (\bar{y} - y_i)^2 \\
&= (y_i - \bar{y})^2 + \frac{1}{n-1} (\bar{y} - y_i)^2 \\
&= \left(\frac{n}{n-1} \right) (y_i - \bar{y})^2 .
\end{aligned}$$

Al tomar el promedio de los pseudovalores, obtenemos:

$$\widehat{\theta}_{\text{J}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{ps,i} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

lo cual es la versión usual de la varianza muestral (insesgada). Esto ilustra cómo $\widehat{\theta}_{\text{J}}$ tiene menos sesgo que $\widehat{\theta}$, donde en este caso $\text{E}_F[\widehat{\theta}] = \{(n-1)/n\}\sigma^2$. De hecho, Bradley Efron (1982) afirma que la verdadera justificación de $\widehat{\theta}_{\text{J}}$ es que elimina por completo el sesgo en estimadores cuadráticos.

Notese que tenemos que:

$$\widehat{\theta}_{ps,i} - \widehat{\theta}_{\text{J}} = \left(\frac{n}{n-1}\right)\left\{(Y_i - \bar{Y})^2 - s_n^2\right\}.$$

Finalmente, el promedio muestral de $\widehat{\theta}_{ps,i}$ es $\widehat{\theta}_{\text{J}}$ y al tomar la varianza muestral (con $n-1$) de $\widehat{\theta}_{ps,i}$ dividida por n, obtenemos:

$$\widehat{V}_{\text{J}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{n}{n-1}\right)^2\left\{(Y_i - \bar{Y})^2 - s_n^2\right\}^2 = \left(\frac{n}{n-1}\right)^3\frac{\left(m_4 - s_n^4\right)}{n}$$

donde recordemos que $m_k = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^k$.

Como los momentos y las funciones de los momentos tienen un rol muy importante en estadistica, vamos a enunciar un teorema de Shao y Tu (2012), para justificar el uso de $\hat{V}_j$ con este estadistico.

Teorema 0.20.3 *Supongamos que $Y_1, \ldots Y_n$ son vectores i.i.d con media finita $E[Y_1] = \mu$ y covarianza Σ. Si g es una función a valores reales (escalar) con con $g'(\mu) \neq 0$ y $g'(y)$ es continua en μ, entonces para $T = g(\bar{Y})$, $n\hat{V}_j \to \sigma^2$ con $n \to \infty$, donde $\sigma^2 = g'(\mu)\Sigma g'(\mu)^T$.*

0.21 Bootstrap

Bootstrap es una técnica general para estimar cantidades desconocidas asociadas con modelos estadísticos. A menudo, Bootstrap se utiliza para encontrar:

1. Errores estándar para estimadores.
2. Intervalos de confianza para parámetros desconocidos.
3. Valores p para estadísticos de prueba bajo una hipótesis nula.

Por lo tanto, el método Bootstrap se utiliza típicamente para estimar cantidades asociadas con la distribución muestral de estimadores y estadísticos de prueba. Para introducir la idea detras de esta técnica nos centraremos en describir el primer uso de los nombrados anteriormente: Errores estandarizados para estimadores.

0.21.1 Errores Estandarizados con Bootstrap

El método de jackknife se centra en estimar la varianza asintótica de un estimador. En cambio, el método Bootstrap intenta estimar directamente la varianza de un estimador. La idea básica es

1. Escribir la varianza del estimador en términos de la función de distribución desconocida F de los datos: $\text{Var}_F(\widehat{\theta})$
2. A continuación, sustituir F por la estimación $\hat{F}$ en la expresión de la varianza, dando como resultado $\text{Var}_{\hat{F}}(\widehat{\theta})$.

En este sentido, el estimador Bootstrap se denomina estimador "plug-in", porque $\hat{F}$ se introduce ("plug") en la expresión de la varianza.

Una visión alternativa del método Bootstrap se basa en muestrear de una pseudo-población ficticia que llamaremos "mundo Bootstrap". Es esta visión la que hace que el Bootstrap sea fácil de entender y de implementar. La idea básica es:

1. Primero, creamos una pseudo-población a partir de los valores de la muestra. En una situación sencilla, consideramos el conjunto de valores de la muestra $\{y_1,\ldots,y_n\}$ como una población.
2. Luego, concebimos la idea de extraer una muestra aleatoria (a menudo llamada remuestreo) de esta pseudo-población, imitando el proceso de muestreo real lo más fielmente posible. Esto es un muestreo aleatorio en el mundo Bootstrap. Para la población simple anterior y el muestreo independiente e idénticamente distribuido, este remuestreo significa extraer muestras con reposición de $\{y_1,\ldots,y_n\}$.
3. Debido a que conocemos todo sobre la pseudo-población, teóricamente podemos calcular la varianza de nuestro estimador cuando la muestra se extrae de la pseudo-población. Esta varianza es el estimador de varianza Bootstrap.

En la práctica, el cálculo del último paso es demasiado difícil, excepto para casos muy simples, por lo que aproximamos el cálculo de la varianza mediante el muestreo repetido de la pseudo-población, calculando el estimador para cada remuestreo Bootstrap y luego calculando la varianza muestral de estos estimadores. En otras palabras, estimamos la varianza teórica en el mundo Bootstrap mediante métodos de Monte Carlo. Recuerde que la varianza teórica de nuestro estimador en el mundo Bootstrap es en realidad una estimación de la varianza cuando se ve desde el mundo real. Por lo tanto, un enfoque de Monte Carlo nos da una estimación de la estimación bootstrap de la varianza.

0.21.2 Interpretacion "Plug-in"para la estimación de la varianza con Bootstrap

La varianza de $\widehat{\theta}$ bajo la distribución F se define como:

$$\mathrm{Var}_F(\widehat{\theta}) = \int \left\{\widehat{\theta}(y_1,\ldots,y_n) - \mathrm{E}_F[\widehat{\theta}]\right\}^2 \mathrm{dF(y_1)}\cdots\mathrm{dF(y_n)}$$

donde

$$\mathrm{E}_F[\widehat{\theta}] = \int \widehat{\theta}(y_1,\ldots,y_n)\mathrm{dF(y_1)}\cdots\mathrm{dF(y_n)}$$

La estimación Bootstrap no paramétrica de $\mathrm{Var}(\hat{\theta})$ consiste en reemplazar F por la función de distribución empírica $F_n(y) = \frac{1}{n}\sum_{i=1}^n I(Y_i \le y)$:

$$\mathrm{Var}_{F_n}(\hat{\theta}) = \int \left\{\hat{\theta}(y_1,\ldots,y_n) - \mathrm{E}_{F_n}[\hat{\theta}]\right\}^2 \mathrm{dF_n(y_1)}\cdots\mathrm{dF_n(y_n)}.$$

Ahora, la expresión general para estos cálculos de esperanza es complicada. Por ejemplo,

$$\mathrm{E}_{F_n}(\hat{\theta}) = \int \hat{\theta}(y_1,\ldots,y_n)\mathrm{dF_n(y_1)}\cdots\mathrm{dF_n(y_n)} = \frac{1}{\mathrm{n^n}}\sum_{\mathrm{i_1}=1}^{\mathrm{n}}\ldots\sum_{\mathrm{i_n}=1}^{\mathrm{n}}\hat{\theta}(\mathrm{Y_{i_1}},\ldots,\mathrm{Y_{i_n}})$$

Recordar que, al integrar una función $g(y)$ en una dimensión con respecto a dF, se tiene que $\int g(y)\mathrm{dF(y)} = \int \mathrm{g(y)f(y)dy}$ para y_1 continuo, donde f es la derivada de F, y $\int g(y)dF(y) = \sum g(y_i)P(Y_1 = y_i)$ para Y_1 discreta con posibles valores $y_1, y_2, \ldots$. Por lo tanto, al reemplazar $F(y)$ por la función de distribución empírica $F_n(y) = \frac{1}{n}\sum_{i=1}^n I(y_i \le y)$, se tiene que $\int g(y)\mathrm{dF_n(y)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{g(y_i)}$, debido a que la función de distribución empírica corresponde a una variable aleatoria discreta con probabilidad $1/n$ en los

valores observados de los datos. Por supuesto, una integral de n dimensiones conduce a una suma de n términos.

Para estimadores muy simples, podemos realizar los cálculos anteriores de manera exacta. Por ejemplo, si $\widehat{\theta} = \bar{Y}$, sabemos que $\mathrm{E}_F[\bar{Y}] = \mathrm{E}_F[Y_1]$ y, por lo tanto,

$$\mathrm{E}_{F_n}[\bar{Y}] = \mathrm{E}_{F_n}[Y_1] = \int y \, \mathrm{dF_n(y)} = \bar{\mathrm{y}}$$

De manera similar, sabemos que $\mathrm{Var}_F(\bar{Y}) = \mathrm{Var}_F(Y_1)/n = [\mathrm{E}_F[Y_1^2] - \{\mathrm{E}_F[Y_1]\}^2]/n$, y luego

$$\mathrm{Var}_{F_n}(\bar{Y}) = \frac{\mathrm{E}_{F_n}[Y_1^2] - \{\mathrm{E}_{F_n}[Y_1]\}^2}{n} = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2\right) = \frac{s_n^2}{n}$$

Por lo tanto, s_n^2/n es el estimador Bootstrap no paramétrico de la varianza de la media muestral.

Para ejemplificar, asumamos que $F(y) = F(y;\sigma) = 1 - \exp(-y/\sigma)$, es decir, una distribución exponencial con media σ. Entonces, el estimador de máxima verosimilitud de F es $\widehat{F} = F(y;\bar{Y}) = 1 - \exp(-y/\bar{Y})$, y el estimador Bootstrap paramétrico de la varianza de la media muestral es simplemente $\mathrm{Var}_{\widehat{F}}\bar{Y} = \mathrm{Var}_{\widehat{F}}Y_1/n = \bar{Y}^2/n$, ya que el cuadrado de la media es la varianza para una distribución exponencial.

0.21.3 Propiedades Asintóticas

Singh (1981) y Bickel y Freedman (1981) iniciaron el estudio de las propiedades asintóticas de el método Bootstrap. El resultado básico es que, bajo condiciones bastante débiles, la estimación Bootstrap de la distribución estandarizada de un estimador converge casi seguramente a la misma distribución asintótica que el estimador. El caso más común es cuando el estimador o estadístico es asintóticamente normal, y en ese caso, un método típico de demostración es utilizar aproximaciones por promedios. Demostrar que la estimación Bootstrap de la varianza es consistente requiere de condiciones adicionales. Se recomienda consultar a Shao y Tu (2012) para más detalles.

Para ilustrar la convergencia casi segura en distribución de la estimación Bootstrap de la distribución, mencionaremos el Teorema 3.1 de Bickel y Freedman (1981). Este resultado se centra el estadístico V de grado $m = 2$:

$$V_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \nu(y_i, y_j)$$

donde ν es un kernel simétrico, es decir, $\nu(x,y) = \nu(y,x)$. Por ejemplo sea $\nu(x,y) = (x-y)^2/2$, entonces tendremos que $V_n = s_n^2$. La varianza asintótica multiplicada por n de V_n es:

$$\sigma_\nu^2 = 4\left[\int\left\{\int \nu(x,y)\mathrm{dF(y)}\right\}^2 \mathrm{dF(y)} - \theta_\nu^2\right],$$

donde $\theta_\nu = \mathrm{E}[\nu(Y_1, Y_2)]$ es su límite en probabilidad. En el siguiente resultado, utilizaremos muestras Bootstrap $y_1^*, \ldots, y_n^*$ que son i.i.d con una función de distribución igual a F_n, la función de distribución empírica de la muestra dado $Y_1, \ldots, Y_n$, y V_n^* es el estadístico basado en una muestra Bootstrap. En este mundo Bootstrap, el verdadero parámetro es $\theta_{\nu,n} = \mathrm{E}_{F_n}[\nu(Y_1^*, Y_2^*)] = V_n$.

Teorema 0.21.1 — Bickel y Freedman (1981). *Supongamos que $Y_1, \ldots, Y_n$ son variables aleatorias independientes e idénticamente distribuidas con función de distribución F, $E[\nu(Y_1,Y_2)]^2 < \infty$, $E[\nu(Y_1,Y_1)]^2 < \infty$, y $\sigma_\nu^2 > 0$. Entonces, para casi todos los valores de $Y_1, Y_2, \ldots$, dado $(Y_1 = y_1, \ldots, Y_n = y_n)$, cuando $n \to \infty$:*

$$\sqrt{n}\,(V_n^* - \theta_{v,n}) \text{ converge débilmente (i.e en distribución) a } N\left(0, \sigma_v^2\right)$$

Este resultado significa que:

$$P^*(\sqrt{n}\,(V_n^* - \theta_{v,n}) \leq x) \longrightarrow \Phi\left(\frac{x}{\sigma_v}\right) \quad \text{a medida que } n \to \infty$$

donde escribimos la probabilidad anterior como P^* para enfatizar que la probabilidad es condicional a la muestra dada, o una probabilidad en el mundo Bootstrap. Por lo tanto, la convergencia asintótica Bootstrap de primer orden implica verificar que el estadístico en el mundo Bootstrap (en el caso anterior, V_n^*) tiene la misma normalidad asintótica que el estadístico V_n en el mundo real. La demostración generalmente utiliza una aproximación por promedios y luego un Teorema del Límite Central que tiene en cuenta el hecho de que la verdadera distribución subyacente F_n en el mundo Bootstrap está cambiando con cada n.

Estos resultados asintóticos de primer orden, sugieren que dado que la distribución Bootstrap converge, también convergen cantidades relacionadas utilizadas en la inferencia, como los cuantiles o las varianzas de la distribución Bootstrap. Sin embargo, a menudo, estos resultados no son suficientes para analizar las variaciones en el método Bootstrap o para compararlo con otros tipos de métodos. Por lo tanto, se utilizan expansiones de Edgeworth de las funciones de distribución Bootstrap, como $P^*\left(\sqrt{n}\,(V_n^* - \theta_{v,n}) \leq x\right)$, para realizar tales análisis.

Para ilustrar, supongamos que $\widehat{\theta}$ sigue una distribución asintóticamente normal AN$(\theta, \sigma/n)$. Una expansión de Edgeworth de un término para $\widehat{\theta}$ daría

$$P(\sqrt{n}(\widehat{\theta} - \theta) \leq x) = \Phi(x/\sigma) + \frac{c}{\sqrt{n}} + o\left(n^{-1/2}\right)$$

a medida que $n \to \infty$ para cada x, donde Φ es la función de distribución normal estándar. Análogamente, en el mundo Bootstrap, deberíamos tener el resultado relacionado

$$P^*\left(\sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}\right) \leq x\right) = \Phi\left(x/\sigma_n^*\right) + \frac{c_n}{\sqrt{n}} + o\left(n^{-1/2}\right) \tag{64}$$

casi seguramente a medida que $n \to \infty$ para cada x, donde $\sigma_n^* \xrightarrow{p} \sigma$ y $c_n \xrightarrow{p} c$. En donde θ^* denota el estimador de θ en el mundo Bootstrap y σ^* su desviación estandar. Entonces restando (64) de (0.21.3), tendremos:

$$\begin{aligned} P\left(\sqrt{n}(\widehat{\theta} - \theta) \leq x\right) - P^*\left(\sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}\right) \leq x\right) &= \Phi(x/\sigma) - \Phi\left(x/\sigma_n^*\right) + o_p\left(n^{-1/2}\right) \\ &= O_p\left(n^{-1/2}\right) \end{aligned}$$

donde en el último paso hemos utilizado la expansión de Taylor de Φ y el supuesto que $\sigma_n^* - \sigma = O_p\left(n^{-1/2}\right)$. Por lo tanto, la distribución Bootstrap de $\sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}\right)$ está dentro de $O_p\left(n^{-1/2}\right)$ de la distribución de $\sqrt{n}(\widehat{\theta} - \theta)$.

Para comparar, consideremos ahora las expansiones análogas para las cantidades tipo t, definimos entonces $t = (\widehat{\theta} - \theta)/\widehat{\sigma}$ y $t^* = \left(\widehat{\theta}^* - \hat{\theta}\right)/\hat{\sigma}^*$, donde $\hat{\sigma}$ es un estimador de σ y asumimos convergencia

a la normal estándar:

$$P(t \leq x) = \Phi(x) + \frac{d}{\sqrt{n}} + o\left(n^{-1/2}\right) \tag{65}$$

$$P^*(t^* \leq x) = \Phi(x) + \frac{d_n}{\sqrt{n}} + o_p\left(n^{-1/2}\right). \tag{66}$$

Ahora, asumiendo que $d_n - d = O_p\left(n^{-1/2}\right)$, tenemos que $P(t \leq x) - P(t^* \leq x) = O_p\left(n^{-1}\right)$, una tasa de convergencia más rápida que para las cantidades sin estandarizar. Estas expansiones han proporcionado una forma importante de comparar intervalos de confianza. Una buena fuente para comprender las expansiones de Edgeworth es Hall (2013).

0.21.4 Intervalos de confianza con Bootstrap

Los intervalos de confianza con Bootstrap han sido el foco de una gran parte de la literatura de investigación sobre Bootstrap. Básicamente ha habido tres líneas principales de desarrollo: El método original del percentil de Efron (1979) y sus mejoras que dieron lugar al intervalo acelerado con corrección de sesgo (BCa), el intervalo t bootstrap introducido en Bradley Efron (1982) y analizado en Hall (1988), y el intervalo doble bootstrap introducido en Hall (1986). Nosotros nos centraremos en el primero de los tres anteriores.

0.21.5 Intervalo del Percentiles

Efron (1979) propuso un intervalo de percentiles bootstrap de $100(1-2\alpha)\,\%$, que consiste en tomar los percentiles empíricos 100α y $100(1-\alpha)$ de los valores bootstrap $\theta_1^*, \ldots, \theta_B^*$ como los límites izquierdo y derecho, respectivamente. Si $\widehat{K}_B$ es la función de distribución empírica de los valores bootstrap, entonces el intervalo de percentiles $100(1-2\alpha)\,\%$ es:

$$\left(\widehat{K}_B^{-1}(\alpha), \widehat{K}_B^{-1}(1-\alpha)\right)$$

0.21.6 Justificación Heuristica del Intervalo del Percentiles

La motivación de Efron para el intervalo de percentiles se basa en suponer la existencia de una transformación creciente g tal que

$$P\left(g(\widehat{\theta}) - g(\theta) \leq x\right) = H(x) \tag{67}$$

donde H es la función de distribución de una variable aleatoria simétrica alrededor de 0, es decir, $H^{-1}(\alpha) = -H^{-1}(1-\alpha)$ para $0 < \alpha < 1$. (Típicamente, Efron describe $H(x)$ como una función de distribución normal con media 0, $\Phi(x/\sigma)$, pero solo se utiliza la simetría en la derivación). De manera similar, en el mundo bootstrap, supongamos que (67) se cumple aproximadamente,

$$P^*\left(g\left(\widehat{\theta}^*\right) - g(\widehat{\theta}) \leq x\right) \approx H(x), \tag{68}$$

recordar que el superindice $*$ se utiliza para denotar cálculos en el mundo bootstrap donde $\widehat{\theta}$ es el valor verdadero. Supongamos por un momento que g es conocida. Entonces, (68) conduce a $P\left(g^{-1}\{g(\widehat{\theta}) - x\} \leq \theta\right) = H(x)$. Al igualar esta probabilidad a $1-\alpha$ obtenemos $x = H^{-1}(1-\alpha)$ y sustituyendo x se obtiene

$$P\left(g^{-1}\left\{g(\widehat{\theta})-H^{-1}(1-\alpha)\right\}\leq\theta\right)=1-\alpha$$

Bajo la suposición de (67) y el conocimiento de g, $\left(g^{-1}\left\{g(\widehat{\theta})-H^{-1}(1-\alpha)\right\},\infty\right)$ es un intervalo unilateral exacto con una probabilidad de cobertura de $1-\alpha$. Dado que el límite superior $1-\alpha$ se deriva de manera similar, nos concentraremos en este límite inferior. Ahora, el objetivo es mostrar que $g^{-1}\left\{g(\widehat{\theta})-H^{-1}(1-\alpha)\right\}$ es estimado por el extremo izquierdo del intervalo de percentiles. Para hacerlo, recordemos que denotamos la función de distribución empírica de los valores bootstrap $\theta_1^*,\ldots,\theta_B^*$ como $\widehat{K}_B$, por lo que el extremo izquierdo del intervalo de percentiles es simplemente $L_\alpha=\widehat{K}_B^{-1}(\alpha)$. Entonces,

$$\begin{aligned}\alpha&=P^*\left(\widehat{\theta}^*\leq L_\alpha\right)\\&=P^*\left\{g\left(\widehat{\theta}^*\right)\leq g\left(L_\alpha\right)\right\}\\&=P^*\left\{g\left(\widehat{\theta}^*\right)-g(\widehat{\theta})\leq g\left(L_\alpha\right)-g(\widehat{\theta})\right\}\\&\approx H\left\{g\left(L_\alpha\right)-g(\widehat{\theta})\right\},\end{aligned}$$

donde la aproximación en el último paso se deriva de (68). Finalmente, resolviendo $\alpha\approx H\left\{g\left(L_\alpha\right)-g(\widehat{\theta})\right\}$ para L_α obtenemos

$$L_\alpha\approx g^{-1}\left\{H^{-1}(\alpha)+g(\widehat{\theta})\right\}=g^{-1}\left\{g(\widehat{\theta})-H^{-1}(1-\alpha)\right)\right\}$$

porque $H^{-1}(\alpha)=-H^{-1}(1-\alpha)$. Esta expresión es la misma que el límite inferior exacto dado anteriormente. Aunque se asumió la existencia de g que satisface (67) y (68), g no se utiliza en la definición del intervalo de percentiles. La justificación teórica rigurosa del intervalo sigue de la convergencia asintótica adecuada de $\widehat{K}_B$.

0.21.7 Precisión asintótica de los intervalos de confianza

Consideremos un intervalo de confianza unilateral nominal $1-\alpha$ $(L_n(\alpha),\infty)$ para θ, es decir, $L_n(\alpha)$ es un límite inferior $1-\alpha$. Si la probabilidad de no cobertura

$$P\left(\theta<L_n(\alpha)\right)=\alpha+O\left(n^{-k/2}\right),$$

se dice que el intervalo es de orden k. Por lo tanto, $P\{\theta<L_n(\alpha)\}=\alpha+c/\sqrt{n}$ se llama de primer orden preciso, y $P\{\theta<L_n(\alpha)\}=\alpha+d/n$ se llama de segundo orden preciso. Una definición similar se aplica a los límites superiores e intervalos de dos lados. No es difícil mostrar que los intervalos unilaterales estándar de Wald y el intervalo de percentiles son de primer orden preciso. Las versiones de dos lados son de segundo orden preciso para estadísticas que son asintóticamente normales. En secciones futuras veremos intervalos bootstrap que tienen una precisión asintótica mejor que el intervalo de percentiles. Estos resultados de orden superior requieren la inversión de las expansiones de Edgeworth discutidas brevemente al final de la Sección 0.21.3. Hall (1986) y Hall (1988) inició el estudio de la precisión de orden superior de los intervalos bootstrap.

0.21.8 Enfoques Monte Carlo para Boostrap

Sea $f(x;\theta)$ distribución de densidad que depende solo de θ y sea $\widehat{\theta}$ la estimación de θ utilizando los datos $x_1,\ldots,x_n$ provenientes de una muestra i.i.d de $f(x;\theta)$. Además sea $\mathscr{T}(\cdot)$, función tal que, dada una muestra (como por ejemplo $x_1,\ldots,x_n$) nos entrega la estimación de θ para dicha muestra (por ejemplo $\widehat{\theta}$).

Dado que conocemos la distribución de probabilidad de la cual provienen los datos, entonces podemos realizar simulaciones para obtener estimaciones para la varianza de nuestro estimador usando bootstrap. Para aquello simularemos B muestras aleatorias de tamaño m de la densidad $f(x;\widehat{\theta})$. Existen ventajas en considerar $m \neq n$ en algunos casos Shao y Tu (2012), pero para ilustrar como funciona el método usaremos $m = n$. Denominaremos $x^{*1}, x^{*2}, \ldots, x^{*B}$ cada una de estas muestras aleatorias de tamaño n, es decir, $x^{*b} = (x_1^{*b}, \ldots x_n^{*b})$. Luego calculamos el valor estimado para θ usando cada una de las muestras x^{*b}, obtiendo asi la estimación θ_b^* de θ. Teniendo así los siguientes estimadores bootstrap para el sesgo b, la varianza s^2 y los percentiles H :

$$b^{(B)} = \frac{1}{B}\sum_{b=1}^{B} \theta_b^* - \widehat{\theta} \tag{69}$$

$$v^{(B)} = \frac{1}{B}\sum_{b=1}^{B} \left(\theta_b^* - \frac{1}{B}\sum_{b=1}^{B} \theta_b^*\right) \tag{70}$$

y además

$$H^{(B)}(x) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\theta_b^* \leq x\right\}. \tag{71}$$

Teniendo así nuestra estimación Bootstrap con Montecarlo para estas cantidades.

Es necesario decidir qué tan grande debe ser el tamaño de muestra de Monte Carlo, B. Este problema, que a menudo preocupa a los estadísticos aplicados, es un poco similar al problema de determinar el tamaño de la muestra original $X_1,\ldots,X_n$. Sin embargo, hay dos diferencias importantes entre ellos. En primer lugar, en la mayoría de los casos podemos permitirnos un B mucho mayor que n. En segundo lugar, desperdiciamos tiempo y recursos si vamos demasiado lejos, ya que realizar más cálculos no ayuda a reducir el error del estimador bootstrap original. Por lo tanto, debemos seleccionar un B de manera que el error de la aproximación de Monte Carlo sea insignificante en comparación con el error del estimador bootstrap original.

Efron (1987) consideró el coeficiente de variación (CV) como una medida de la variabilidad de la aproximación de Monte Carlo utilizando un B fijo. Dado que la estimación de la desviación estándar y la construcción de intervalos de confianza son las dos principales aplicaciones del bootstrap, Efron (1987) solo estudió la determinación de B para estos dos problemas.

La aproximación de Monte Carlo para el estimador de desviación estándar bootstrap de un estimador $\widehat{\theta}$ es:

$$s^{(B)} = \sqrt{v^{(B)}}$$

donde $v^{(B)}$ es la varianza bootstrap definida en (70). Las fórmulas estándar para estimadores de momentos Serfling (2009) conducen a

$$E\left[s^{(B)}; \widehat{\theta}\right] \approx \sqrt{v_{\text{Boot}}}$$

y

$$\text{var}\left[s^{(B)};\widehat{\theta}\right] \approx \frac{\rho_{\text{Boot}} - v_{\text{Boot}}^2}{4Bv_{\text{Boot}}}$$

donde v_{Boot} es:

$$v_{\text{Boot}} = \text{var}_*[\theta^*] := \text{var}[\theta^*;\widehat{\theta}]$$

y

$$\rho_{\text{Boot}} = E\left[\left(\hat{\theta}^* - E[\theta^*;\widehat{\theta}_n]\right)^4\right].$$

Por lo tanto, el coeficiente de variación de $s^{(B)}$, condicional a $X_1,\ldots,X_n$, es igual a

$$\text{cv}_*\left(s^{(B)}\right) = \frac{\sqrt{\text{var}_*\left[s_{\text{Boot}}^{(B)}\right]}}{E_*[s_{\text{Boot}}^{(B)}]} \approx \sqrt{\frac{\hat{\delta}_n + 2}{4B}} \tag{72}$$

En donde:

$$\begin{aligned} E_*[g(X)] &= E[g(X);\widehat{\theta}] \\ \text{cv}_*(g(X)) &= cv(g(X);\widehat{\theta}) \\ \widehat{\delta}_n &= \rho_{\text{Boot}}\,/v_{\text{Boot}}^2 - 3 \end{aligned}$$

Hay al menos dos formas de determinar el tamaño de muestra de Monte Carlo, B, utilizando la ecuación (72).

El primer método, sugerido por Bradley Efron (1987), es determinar B estableciendo que:

$$\text{cv}_*\left(s^{(B)}\right) = \varepsilon_0 \tag{73}$$

donde ε_0 es un nivel deseado dado. En muchos casos, $\hat{\delta}_n \approx 0$ $\left(\hat{\delta}_n \to_p 0\right.$ a medida que $\left.n \to \infty\right)$. Entonces, la ecuación (73) se reduce a $B = \frac{1}{2}\varepsilon_0^{-2}$. Por ejemplo, si $\varepsilon_0 = 0{,}05$, entonces $B = 200$; si $\varepsilon_0 = 0{,}1$, entonces $B = 50$.

Sea $s_{\text{Boot}} = \sqrt{v_{\text{Boot}}}$ el estimador de desviación estándar bootstrap y $\text{cv}\left(s_{\text{Boot}}\right)$ su coeficiente de variación. Cabe destacar que no vale la pena intentar hacer que $\text{cv}_*\left(s^{(B)}\right)$ sea demasiado pequeño en comparación con $\text{cv}\left(s_{\text{Boot}}\right)$, pues, en la mayoría de los casos, es suficiente tener $\text{cv}_*\left(s^{(B)}\right)/\text{cv}\left(s_{\text{Boot}}\right) \to 0$. En algunos casos, incluso $\text{cv}_*\left(s_{\text{Boot}}^{(B)}\right) = \text{cv}\left(s_{\text{Boot}}\right)$ es suficiente.

El segundo método para determinar B es establecer

$$\text{cv}_*\left(s^{(B)}\right) = \text{cv}\left(s_{\text{Boot}}\right) \text{ o } o\left(\text{cv}\left(s_{\text{Boot}}\right)\right)$$

lo cual conduce a

$$B = \frac{a_n\left(\hat{\delta}_n + 2\right)}{4\left[\text{cv}\left(s_{\text{Boot}}\right)\right]^2} \tag{74}$$

donde $a_n \equiv 1$ o $\{a_n\}$ es cualquier secuencia de números positivos que divergen hacia el infinito (por ejemplo, $a_n = \log\log n$). El tamaño de muestra de Monte Carlo seleccionado según (74) depende de n y de la precisión del estimador bootstrap original.

Estadistica Bayesiana

La estadística Bayesiana es un enfoque de la estadística que se basa en reinterpretar el concepto de probabilidad. En estadística frecuentista (estadística clásica) la probabilidad de un evento se interpreta como la frecuencia con la que ocurre dicho evento en una serie larga de ensayos repetidos. En contraposición, la estadística Bayesiana interpreta la probabilidad como una medida de confianza o credibilidad que un individuo pueda tener sobre un evento en concreto. Podemos tener una creencia previa sobre un suceso, pero es probable que nuestras creencias cambien al momento de que obtengamos nuevas pruebas.

Bajo estos dos diferentes paradigmas, obtendremos dos enfoques distintos al momento de hacer inferencia. En la estadística frecuentista intentamos eliminar la incertidumbre de un suceso realizando estimaciones. En cambio, la estadística bayesiana trata de mantener la incertidumbre, pero la actualiza ajustando las creencias personales con las nuevas pruebas. Este proceso se puede realizar gracias al Teorema de Bayes que expondremos a continuación.

0.22 Teorema de Bayes

En esta sección usaremos la notación usual en estadística bayesiana. En la cual, si X, Y son variables aleatorias denotamos la densidad de X por $\pi(x)$, la densidad conjunta por $\pi(x,y)$ y la densidad condicional de X dado $Y = y$ por $\pi(x|y)$.

Como mencionamos anteriormente la estadística Bayesiana se basa en el teorema de Bayes, el cual enunciaremos a continuación.

Teorema 0.22.1 — Teorema de Bayes. *Sean X, Y variables aleatorias, y sea y, x tal que $\pi(x) > 0$ entonces:*

$$\pi(y|x) = \frac{\pi(x|y)\pi(y)}{\pi(x)}$$

Cuya demostración se obtiene usando la definición de densidad condicional:

$$\pi(y|x) = \frac{\pi(x,y)}{\pi(x)} = \frac{\pi(x|y)\pi(y)}{\pi(x)}$$

Notar que podemos usar el teorema de probabilidades tenemos que, $\pi(x) = \int \pi(x|y)\pi(y)\mathrm{dy}$. Y por lo tanto si conocemos $\pi(x|y), \pi(y)$ y además x esta fijo, podemos calcular $\pi(x)$ pues este no depende de y. Por lo anterior tendremos que $\pi(y|x)$ es proporcional a $\pi(x|y)\pi(y)$ o, como se denota en estadistica bayesiana:

$$\pi(y|x) = \pi(x|y)\pi(y)$$

0.23 Teorema de Bayes para inferencia paramétrica

Sea X una variable aleatoria que proviene de un modelo estadística que depende de un parámetro θ, en particular si θ es conocido, usando la notación de estadística bayesiana, $X|\theta \sim \pi(X|\theta)$. Consideremos que tenemos una muestra x y queremos hacer inferencia sobre θ. En el análisis Bayesiano se considera que el parámetro θ desconocido es una variable aleatoria, por lo cual posee una función de densidad que denotaremos por $\pi(\theta)$. Entonces usando el teorema de Bayes obtenemos lo siguiente:

$$\begin{aligned}\pi(\theta|x) &= \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)} \\ &\propto \pi(x|\theta)\pi(\theta).\end{aligned}$$

Lo cual se conoce como:

$$\text{Posteriori} \propto \text{Verosimilitud} \times \text{Priori}$$

Dado lo anterior la estadística bayesiana se basa en los siguientes componentes:

1. Una distribución a priori $\pi(\theta)$ la cual representa nuestro conocimiento previo sobre θ antes de observar los datos.
2. Una función de verosimilitud $\pi(x|\theta)$, que es la distribución de x en el caso de que θ fuera conocido.
3. Una función de distribución a posteriori $\pi(\theta|x)$, la cual representa el conocimiento de θ después de observar los datos.

Si ya tenemos una muestra x entonces para obtener la distribución a posteriori solo necesitamos usar el teorema de Bayes, como se ilustra en el siguiente ejemplo.

■ **Ejemplo 0.23.1** Supongamos que $X|\theta \sim Bin(n,\theta)$. Podemos especificar una distribución a priori para θ, por ejemplo, consideremos $\theta \sim Beta(\alpha,\beta)$ con $\alpha,\beta > 0$ conocidos. Entonces para $0 \leq \theta \leq 1$ tenemos:

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

donde $B(\alpha,\beta) = \dfrac{\gamma(\alpha)\gamma(\beta)}{\gamma(\alpha+\beta)}$ es la funcion beta y $E[\theta] = \dfrac{\alpha}{\alpha+\beta}$. Como:

$$\int_0^1 \pi(\theta)\mathrm{d}\theta = 1$$

entonces:

$$B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta \tag{75}$$

Usando el teorema de Bayes, obtenemos que la densidad a posteriori es:

$$\begin{aligned}\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-1} \times \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.\end{aligned}$$

Entonces, $\pi(\theta|x) = c\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$ para alguna constante c que no depende de θ. Ahora tenemos que:

$$\int_0^1 \pi(\theta|x)\mathrm{d}\theta = 1 \;\Rightarrow\; c^{-1} = \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\mathrm{d}\theta.$$

Notemos que gracias a (75) podemos evaluar la integral y por lo tanto tenemos que:

$$c^{-1} = \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\mathrm{d}\theta = \mathrm{B}(\alpha+\mathrm{x},\beta+\mathrm{n}-\mathrm{x})$$

De donde:

$$\pi(\theta|x) = \frac{1}{B(\alpha+x,\beta+n-x)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$$

es decir, $\theta|x \sim Beta(\alpha+x,\beta+n-x)$

Notemos lo simple que es realizar esta actualización:el número de aciertos observados, x, se suma a α mientras que el número de fallas observadas, $n-x$, se suma a β. Esto ocurre porque la distribución a priori y posteriori son ambas de la misma familia de distribuciones, en este caso la familia Beta. Cuando lo anterior ocurre se dice que las distribuciones son conjugadas.

0.23.1 Distribuciones conjugadas y la familia exponencial

Definición 0.23.1 — Familia de conjugación. *Una familia* Π *de distribuciones a priori forman una familia de conjugación con respecto a la verosimilitud* $\pi(x|\theta)$ *si la densidad a posteriori esta en la familia* Π *de distribución para todo x independiente de la distribución a priori en* Π.

En el ejemplo 0.23.1 mostramos que, con respecto a la verosimilitud Binomial, la distribución Beta es una familia de conjugación.

Una de las mayores ventajas mostradas en el ejemplo 0.23.1, la cual se cumple en general para familias de conjugación, es que es sencillo calcular la constante de proporcionalidad. En el análisis Bayesiano tenemos que $\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta)$ tal que $\pi(\theta|x) = c\pi(x|\theta)\pi(\theta)$ donde c es una constante que no depende de θ. Como $\pi(\theta|x)$ es una función de densidad e integra 1 tenemos que:

$$c^{-1} = \int_\theta \pi(x|\theta)\pi(\theta)\mathrm{d}\theta.$$

Recordemos que la distribución binomial pertenece a la familia exponencial de un parámetro, con estadístico suficiente $t(x) = x$, lo cual motiva a buscar distribuciones a priori conjugadas para verosimilitudes de esta familia. Lo anterior fue exactamente lo realizado por, Diaconis e Ylvisaker (1979) quienes estudiaron distribuciones a priori conjugadas para la familia exponencial, y en particular distribuciones a priori de la forma $\pi(\phi \mid n_0,t_0) = \kappa(n_0,t_0)\,c(\phi)^{n_0}e^{n_0t_0\phi}$. Combinar esta información a priori con la información de $X_1,\ldots,X_n \sim$ i.i.d. $\pi(x \mid \phi) = b(x)\exp(\phi t(x))/a(\phi)$ nos da la siguiente distribución a posteriori:

$$\begin{aligned}\pi(\phi \mid x_1,\ldots,x_n) &\propto \pi(\theta)\pi(x_1,\ldots,x_n \mid \theta)\\ &\propto a(\phi)^{-n_0-n}\exp\left\{\phi \times \left[n_0 t_0 + \sum_{i=1}^{n} t(x_i)\right]\right\}\\ &\propto \pi(\phi \mid n_0+n, n_0 t_0 + n\bar{t}(x)),\end{aligned}$$

donde $\bar{t}(x) = \sum t(x_i)/n$. La similitud entre las distribuciones a posteriori y a priori sugiere que n_0 se puede interpretar como un "tamaño de muestra previoz t_0 como una "suposición previa"de $t(X)$. Esta interpretación puede ser mas precisa: Diaconis e Ylvisaker (1979) demostraron que:

$$\begin{aligned}\mathrm{E}[t(X)] &= \mathrm{E}[\mathrm{E}[t(X) \mid \phi]]\\ &= \mathrm{E}\left[c'(\phi)/c(\phi)\right] = t_0,\end{aligned}$$

por lo que t_0 representa el valor esperado previo de $t(X)$. El parámetro n_0 es una medida de cuán informativa es la distribución a priori. Hay una variedad de formas de cuantificar lo anterior, pero quizás la más simple es observar que, como función de $\phi, \pi(\phi \mid n_0, t_0)$ tiene la misma forma que una verosimilitud $\pi(\tilde{x}_1,\ldots,\tilde{x}_{n_0} \mid \phi)$ basada en n_0 .ºbservaciones previas"$\tilde{x}_1,\ldots,\tilde{x}_{n_0}$ para las cuales $\sum t(\tilde{x}_i)/n_0 = t_0$. En este sentido, la distribución a priori $\pi(\phi \mid n_0, t_0)$ contiene la misma cantidad de información que se obtendría a partir de n_0 muestras independientes de la población.

■ **Ejemplo 0.23.2 — Distribución Binomial.** La representación de la familia exponencial de la distribución binomial (θ) se puede obtener a partir de la función de densidad de una sola variable aleatoria binaria:

$$\begin{aligned}\pi(x \mid \theta) &= \theta^x(1-\theta)^{1-x}\\ &= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)\\ &= e^{\phi y}\left(1+e^{\phi}\right)^{-1},\end{aligned}$$

donde $\phi = \log[\theta/(1-\theta)]$. La distribución conjugada a priori para ϕ está dada por $p(\phi \mid n_0, t_0) \propto \left(1+e^{\phi}\right)^{-n_0} e^{n_0 t_0 \phi}$, donde t_0 representa la esperanza a priori de $t(x) = x$, o equivalentemente, t_0 representa nuestra probabilidad a priori de que $X = 1$. Utilizando el teorema de transformación , esto se traduce en una distribución a priori para θ tal que $\pi(\theta \mid n_0, t_0) \propto \theta^{n_0 t_0 - 1}(1-\theta)^{n_0(1-t_0)-1}$, que es una distribución Beta $(n_0 t_0, n_0(1-t_0))$. Se puede obtener una distribución a priori débilmente informativa estableciendo t_0 igual a nuestra esperanza a priori y $n_0 = 1$. Si nuestra esperanza a priori es $1/2$, el resultado es una distribución Beta $(1/2, 1/2)$. Bajo la distribución a priori beta débilmente informativa $(t_0, (1-t_0))$, la distribución posteriori sería $\{\theta \mid x_1,\ldots,x_n\} \sim Beta(t_0 + \sum x_i, (1-t_0) + \sum(1-x_i))$.

■ **Ejemplo 0.23.3 — Distribución Poisson.** El modelo Poisson(θ) se puede mostrar como un modelo de familia exponencial con las siguientes características:

- $t(x) = x$
- $\phi = \log\theta$;
- $c(\phi) = \exp\left(-e^{-\phi}\right)$.

La distribución a priori conjugada para ϕ es entonces $\pi(\phi \mid n_0, t_0) = \exp\left(n_0 e^{-\phi}\right) e^{n_0 t_0 x}$, donde t_0 es el valor esperado a priori de la media poblacional de X. Esto se traduce en una densidad a priori para θ de la forma $\pi(\theta \mid n_0, t_0) \propto \theta^{n_0 t_0 - 1} e^{-n_0 \theta}$, que es una densidad gamma $(n_0 t_0, n_0)$. Se puede obtener una distribución a priori débilmente informativa estableciendo t_0 como la esperanza a priori de X y $n_0 = 1$, lo que resulta en una distribución a priori Gamma $(t_0, 1)$. La distribución a posteriori bajo tal priori sería $\{\theta \mid x_1, \ldots, x_n\} \sim Gamma(t_0 + \sum x_i, 1 + n)$.

0.23.2 Intervalos de credibilidad

A menudo es deseable identificar regiones del espacio parametrico las cuales tengan una alta probabilidad de contener al valor del verdadero parametro. Para hacer esto, después de observar los datos $X = x$, podemos construir un intervalo $[l(x), u(x)]$ de manera que la probabilidad de que $l(x) < \theta < u(x)$ sea alta.

Definición 0.23.2 — Intervalo de credibilidad. Un intervalo $[l(y), u(y)]$, basado en los datos observados $Y = y$, tiene una credibilidad bayesiana (probabilidad de cobertura) del $(100 - \alpha)\,\%$ para θ si

$$\mathbb{P}(l(y) < \theta < u(y) \mid Y = y) = (100 - \alpha)\,\%$$

La forma de obtener $l(y)$ y $u(y)$ dependen del metodo que ocupemos para construir el intervalo de credibilidad. Un metodo es el intervalo basado en cuartiles, el cual expondremos a continuación,

Intervalos basados en cuantiles

Una forma simple de obtener un intervalo de credibilidad es utilizar cuantiles a posteriori. Para crear un intervalo de credibilidad basado en cuantiles del $100 \times (1 - \alpha)\,\%$, hay que encuentrar los números $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ de manera que se cumplan las siguientes condiciones:

1. $\mathbb{P}(\theta < \theta_\alpha \mid Y = y) = \alpha/2$
2. $\mathbb{P}\left(\theta > \theta_{1-\alpha/2} \mid Y = y\right) = \alpha/2$.

Los números $\theta_{\alpha/2}$ y $\theta_{1-\alpha/2}$ son los cuantiles a posteriori del $\alpha/2$ y $1 - \alpha/2$ de θ, por lo que

$$\begin{aligned}
\mathbb{P}\left(\theta \in \left[\theta_{\alpha/2}, \theta_{1-\alpha/2}\right] \mid X = x\right) &= 1 - \mathbb{P}\left(\theta \notin \left[\theta_{\alpha/2}, \theta_{1-\alpha/2}\right] \mid X = x\right) \\
&= 1 - \left[\mathbb{P}\left(\theta < \theta_{\alpha/2} \mid X = x\right) + \mathbb{P}\left(\theta > \theta_{1-\alpha/2} \mid X = x\right)\right] \\
&= 1 - \alpha.
\end{aligned}$$

■ **Ejemplo 0.23.4 — Binomial con a priori uniforme.** Supongamos que, de $n = 10$ observaciones condicionalmente independientes de una variable aleatoria binaria, observamos $X = 2$ unos. Usando una distribución uniforme como distribución a priori para θ, la distribución a posteriori es $\theta \mid \{X = 2\} \sim \text{beta}(1 + 2, 1 + 8)$. Un intervalo de credibilidad a posteriori del 95 % se puede obtener a partir de los cuantiles 0,025 y 0,975 de esta distribución beta. Estos cuantiles son 0,06 y 0,52 respectivamente, por lo que la probabilidad a posteriori de que $\theta \in [0{,}06, 0{,}52]$ es del 95 %.

Intervalo de mayor densidad a posteriori

La imagen anterior muestra que la distribución a posteriori y un intervalo de credibilidad del 95 % para θ del ejemplo anterior. Notese que hay valores de θ fuera del intervalo basado en cuantiles que tienen una densidad más alta que algunos puntos dentro del intervalo. Esto sugiere un tipo de intervalo más restrictivo:

Definición 0.23.3 — Region región de mayor densidad a posteriori . Una región de mayor densidad a posteriori (o High Density Posterior por sus siglas en ingles), es un subconjunto del espacio parametrico $s(y)$ tal que:

1. $\mathbb{P}(\theta \in s(y)|Y = y) = 1 - \alpha$
2. Si $\theta_a \in s(y)$ y $\theta_b \notin s(y)$ entonces, $\pi(\theta_a|Y = y) > \pi(\theta_b|Y = y)$

Todos los puntos en una región HPD tienen una mayor densidad a posteriori que los puntos fuera de la región. Sin embargo, una región HPD puede no ser un intervalo si la densidad a posteriori es multimodal (tiene múltiples peaks). Para el ejemplo binomial anterior, la región 95 %HPD es $[0{,}04, 0{,}048]$, que es más estrecha (más precisa) que el intervalo basado en cuantiles, pero ambos contienen el 95 % de la probabilidad a posteriori.

Relación intervalos de credibilidad e intervalos de confianza

Recordando la definición de intervalos de credibilidad (Def0.23.2) la interpretación de este intervalo es que describe la información que se tiene sobre la ubicación del valor verdadero de θ después de haber observado $X = x$. Esto es diferente de la interpretación frecuentista de la probabilidad de cobertura, que describe la probabilidad de que el intervalo cubra el valor verdadero antes de observar los datos:

Definición 0.23.4 — Intervalo de confianza. Un intervalo aleatorio $[l(Y), u(Y)]$ tiene una confianza (probabilidad de cobertura) del $(100 - \alpha)\,\%$ para θ si, antes de recolectar los datos,

$$\mathbb{P}(l(X) < \theta < u(X) \mid \theta) = (100 - \alpha)\,\%$$

En cierto sentido, las nociones frecuentista y bayesiana de cobertura describen la cobertura pre y post experimental, respectivamente.

Bajo el contexto frecuentista, una vez que se observa $Y = y$, y se sustituyan estos datos en la fórmula del intervalo de confianza $[l(y), u(y)]$, entonces

$$\mathbb{P}(l(y) < \theta < u(y) \mid \theta) = \begin{cases} 0 & \text{si } \theta \notin [l(y), u(y)] \\ 1 & \text{si } \theta \in [l(y), u(y)] \end{cases}$$

Esto destaca la falta de una interpretación post-experimental de los intervalos de confianza. Aunque esto puede hacer que la interpretación frecuentista parezca algo insuficiente, aún es útil en muchas situaciones. Supongamos que se esta realizando un gran número de experimentos no relacionados y se estan creando un intervalo de confianza para cada uno de ellos. Si los intervalos tienen una probabilidad de cobertura frecuentista del 95 %, se puede esperar que el 95 % de los intervalos contengan el valor correcto del parámetro.

¿Puede un intervalo de confianza frecuentista tener el mismo grado de credibilidad en el contexto bayesiano? Hartigan (1966) demostró que, para ciertos tipos de intervalos (como los que usaremos en esta sección), un intervalo que tiene una cobertura bayesiana del 95 % adicionalmente tiene la propiedad de que:

$$\mathbb{P}(l(X) < \theta < u(X) \mid \theta) = 0{,}95 + \varepsilon_n$$

donde $|\varepsilon_n| < \frac{a}{n}$ para alguna constante a. Esto significa que un procedimiento de construcción de un intervalo de confianza que proporciona una credibilidad bayesiana del 95 % también tendrá aproximadamente una cobertura frecuentista del 95 %, al menos asintóticamente. Es importante tener en cuenta que la mayoría de los métodos no bayesianos de construcción de intervalos de confianza del 95 % también logran esta tasa de cobertura asintóticamente. Para mas detalles sobre las similitudes entre los intervalos construidos por métodos bayesianos y no bayesianos, consultar Severini (1991) y Sweeting (2001).

0.23.3 Predicción

Una característica importante de la inferencia bayesiana es la existencia de una distribución predictiva para nuevas observaciones. Sean $x_1,\ldots,x_n$ los resultados de una muestra de n variables aleatorias binarias, y sea $\tilde{X} \in \{0,1\}$ un resultado adicional de la misma población que aún no ha sido observado. La distribución predictiva de $\tilde{X}$ es la distribución condicional de $\tilde{X}$ dada $\{X_1 = x_1,\ldots,X_n = x_n\}$. Para variables binarias condicionalmente i.i.d., esta distribución se puede obtener a partir de la distribución de $\tilde{X}$ dada θ y la distribución a posteriori de θ:

$$\begin{aligned}
\mathbb{P}\left(\tilde{X}=1 \mid x_1,\ldots,x_n\right) &= \int \mathbb{P}\left(\tilde{X}=1,\theta \mid x_1,\ldots,x_n\right) d\theta \\
&= \int \mathbb{P}\left(\tilde{X}=1 \mid \theta,x_1,\ldots,x_n\right) \pi\left(\theta \mid x_1,\ldots,x_n\right) d\theta \\
&= \int \theta \pi\left(\theta \mid x_1,\ldots,x_n\right) d\theta \\
&= \mathrm{E}\left[\theta \mid x_1,\ldots,x_n\right] = \frac{a+\sum_{i=1}^{n} x_i}{a+b+n} \\
\mathbb{P}\left(\tilde{X}=0 \mid x_1,\ldots,x_n\right) &= 1-\mathrm{E}\left[\theta \mid x_1,\ldots,x_n\right] = \frac{b+\sum_{i=1}^{n}(1-x_i)}{a+b+n}
\end{aligned}$$

Debemos considerar dos cosas importantes sobre la distribución predictiva:

1. La distribución predictiva no depende de ninguna cantidad desconocida. Si lo hiciera, no podríamos usarla para hacer predicciones.

2. La distribución predictiva depende de nuestros datos observados. En esta distribución, $\tilde{X}$ no es independiente de $X_1,\ldots,X_n$. Esto se debe a que observar $X_1,\ldots,X_n$ nos brinda información sobre θ, que a su vez nos brinda información sobre $\tilde{X}$. Sería malo si $\tilde{X}$ fuera independiente de $Y_X,\ldots,X_n$, ya que eso significaría que nunca podríamos inferir nada sobre la población no muestreada a partir de los casos de muestra.

■ **Ejemplo 0.23.5 — Binomial Beta.** La distribución a priori uniforme, o priori $Beta(1,1)$, se puede considerar equivalente a la información en un conjunto de datos previo que consta de un solo "1τ un solo "0". Bajo esta distribución a priori,

$$\mathbb{P}(\tilde{X}=1 \mid X=x) = \mathrm{E}[\theta \mid X=x] = \frac{2}{2+n}\frac{1}{2} + \frac{n}{2+n}\frac{x}{n},$$

donde $X = \sum_{i=1}^{n} X_i$.

0.23.4 Modelo Normal

Por lo general, una distribución normal se especifica mediante su media, μ, y varianza, σ^2. Al trabajar con una distribución normal en un contexto bayesiano, generalmente es más conveniente especificarla en términos de su media, μ, y su precisión, $\tau := 1/\sigma^2$. Una variable aleatoria X, que distribuye normal con media μ y precisión τ, se denota por $X \sim N(\mu,\tau)$ y tiene función de densidad:

$$\begin{aligned}
\pi(x \mid \mu,\tau) &= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\tau x^2 + \tau\mu x\right)
\end{aligned} \tag{76}$$

Primero, supongamos que tenemos una única observación $x \sim N(\mu, \tau)$ con precisión conocida pero media desconocida. Asumiremos que la distribución a priori de μ es una normal, específicamente $\mu \sim N(\mu_0, \tau_0)$. Entonces, la distribución a posteriori de μ se obtiene mediante:

$$\begin{aligned}\pi(\mu \mid x) &\propto \pi(x \mid \mu)\pi(\mu) \\ &= \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu-\mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}(\tau+\tau_0)\mu^2 + (\tau x + \tau_0\mu_0)\mu\right). \end{aligned} \tag{77}$$

Comparando la Ecuación (77) con la Ecuación (76), podemos observar que esto implica que $\mu \mid x \sim N(\mu_1, \tau_1)$, donde

$$\tau_1 = \tau + \tau_0 \quad \text{y} \quad \mu_1 = \frac{1}{\tau_1}(\tau x + \tau_0\mu_0) = \frac{\tau x + \tau_0\mu_0}{\tau + \tau_0}.$$

Aquí podemos notar que la media a posteriori, μ_1, es un promedio ponderado de x y μ_0 con pesos $\tau/(\tau+\tau_0)$ y $\tau_0/(\tau+\tau_0)$, respectivamente.

En segundo lugar, supongamos que tenemos n observaciones independientes $x_1, \ldots, x_n$, donde $x_i \sim N(\mu, \tau)$. Nuevamente, asumiendo que τ es conocido y que a priori $\mu \sim N(\mu_0, \tau_0)$, se sigue que

$$\mu \mid x_1, \ldots, x_n \sim N(\mu_1, \tau_1)$$

donde

$$\tau_1 = n\tau + \tau_0 \quad \text{y} \quad \mu_1 = \frac{1}{\tau_1}\left(\tau \sum_{i=1}^{n} x_i + \tau_0\mu_0\right) = \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}$$

aquí $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ es el promedio de las n observaciones. Nuevamente, podemos ver que la media a posteriori es un promedio ponderado de la media observada y la media a priori. Ahora consideramos la situación en la que la media es conocida y la precisión es desconocida. Supongamos que la precisión sigue una distribución gamma con parámetro de forma α y parámetro de escala β, es decir,

$$\pi(\tau \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \tau^{\alpha-1} e^{-\tau/\beta}$$

lo cual denotamos como $\tau \sim \text{Gamma}(\alpha, \beta)$. Recordemos que $E(\tau) = \alpha\beta$ y $\text{Var}(\tau) = \alpha\beta^2$. Luego se puede demostrar que la distribución a posteriori de τ también es gamma:

$$\tau \mid x_1, \ldots, x_n \sim \textit{Gamma}\left(\frac{n}{2} + \alpha, \left(\frac{1}{2}\sum_i (x_i - \mu)^2 + \frac{1}{\beta}\right)^{-1}\right)$$

0.24 Aproximaciones Monte Carlo

Gran parte de la inferencia a posteriori en estadística bayesiana se basa en conocer la distribución $\pi(x|\theta)\pi(\theta)/\pi(x)$, como mencionamos anteriormente podemos calcular $\pi(x)$ gracias al teorema de probabilidades totales $\pi(x) = \int \pi(x|\theta)\pi(\theta)\mathrm{d}\theta$. Sin embargo, es posible que la integral anterior no sea posible de calcular, es mas podríamos tener casos en donde la integral anterior diverja ($= \infty$) o tener

casos en donde $\pi(\theta)$ no sea una distribución de probabilidad (ie $\int \pi(\theta)\mathrm{d}\theta = \infty$) lo cual se conoce como priori impropia. Otro problema es que la distribución a posteriori puede no ser una distribución conocida, o una distribución en la cual no podemos calcular parámetros de interés.

Para resolver estos problemas, una solución es usar aproximaciones Monte Carlo, las cuales consisten en realizar simulaciones de $\pi(x|\theta)\pi(\theta)$ y utilizar dichas soluciones como distribución muestral para realizar la inferencia.

Para ilustrar este procedimiento, utilizaremos uno de los ejemplos mas clasicos de la literatura: el modelo normal con media y precisión desconocidas.

0.24.1 Normal con media y precisión desconocidas

Asumamos que tenemos el mismo modelo que en la sección 0.23.4, pero con la media y precisión desconocida. Con respecto a la distribución a priori, asumiremos que μ y τ son independientes y que distribuyen normal y gamma respectivamente es decir:

$$\begin{aligned}\pi(\mu,\tau) &= \pi(\mu)\pi(\tau)\\ \pi(\mu) &\sim N(\mu_0,\tau_0)\\ \pi(\tau) &\sim Gamma(\alpha,\beta)\end{aligned}$$

Por otro lado, asumiremos que la distribución a posteriori tiene la siguiente forma: $\pi(\mu,\tau|x) \propto \pi(x|\mu,\tau)\pi(\mu,\tau) = \pi(x|\mu,\tau)\pi(\mu)\pi(\tau)$.

Bajo estos supuestos, aplicaremos aproximaciones Monte Carlo para generar simulaciones de la distribución a posteriori $\pi(\mu,\tau|x)$. Para lograr lo anterior, notemos que dado que μ y τ son independientes a priori, entonces podemos conocer conocemos las distribuciones full condicionales de μ y τ:

$$\begin{aligned}\pi(\mu|\tau,x) &\sim N\left(\frac{n\tau\bar{x}+\tau_0\mu_0}{n\tau+\tau_0}, n\tau+\tau_0\right)\\ \pi(\tau|\mu,x) &\sim Gamma\left(\frac{n}{2}+\alpha, \left(\frac{1}{2}\sum_i (x_i-\mu)^2+\frac{1}{\beta}\right)^{-1}\right)\end{aligned}$$

Por lo tanto podemos usar un algoritmo Gibbs Sampler para simular de la distribución a posteriori $\pi(\mu,\tau|x)$. En particular dada una muestra x tenemos el siguiente algoritmo:

A. 20: Normal con media y precisión desconocidas

Dado x, μ_i y τ_i

1. Generar μ_{i+1} de:
$$\mu_{i+1}|x,\tau_i \sim N\left(\frac{n\tau_i\bar{x}+\tau_0\mu_0}{n\tau_i+\tau_0}, n\tau+\tau_0\right)$$
2. Generar, τ_{i+1} de:
$$\tau_{i+1}|\mu_i,x \sim Gamma\left(\frac{n}{2}+\alpha, \left(\frac{1}{2}\sum_i (x_i-\mu_{i+1})^2+\frac{1}{\beta}\right)^{-1}\right)$$

Usando el algoritmo anterior podemos obtenemos las siguientes simulaciones.

0.24.2 Modelo de Efectos Aleatorios

Un modelo de efectos aleatorios es un modelo jerarquico que contiene la siguiente estructura:

$$Y_{ij} = \beta + U_i + \varepsilon_{ij} \qquad i = 1,\ldots,I \; j = 1,\ldots J$$

en donde $U_i \sim N(0,\sigma^2)$ y $\varepsilon_{ij} \sim N(0,\tau^2)$. Asumamos la siguiente distribución a priori para los parametros β, σ^2 y τ^2:

$$\pi(\beta,\sigma^2,\tau^2) = \frac{1}{\sigma^2\tau^2}.$$

Notemos que la distribución a priori es impropia si consideramos $\beta \in \mathbb{R}$.

Dado lo anterior, podemos obtener las siguientes distribuciones full condicionales:

$$\begin{aligned}
U_i \mid y,\beta,\sigma^2,\tau^2 &\sim N\left(\frac{J(\bar{y}_i-\beta)}{J+\tau^2\sigma^{-2}}, \left(J\tau^{-2}+\sigma^{-2}\right)^{-1}\right),\\
\beta \mid u,y,\sigma^2,\tau^2 &\sim N\left(\bar{y}-\bar{u}, \tau^2/JI\right)\\
\sigma^2 \mid u,\beta,y,\tau^2 &\sim IG\left(I/2,(1/2)\sum_i u_i^2\right),\\
\tau^2 \mid u,\beta,y,\sigma^2 &\sim IG\left(IJ/2,(1/2)\sum_{i,j}(y_{ij}-u_i-\beta)^2\right),
\end{aligned}$$

están bien definidas y un algoritmo Gibbs Sampler se puede implementar para obtener la distribución a posteriori. Definimos entonces:

A. 21: Modelo de Efectos Aleatorios

Dada una muestra y, y, $\beta^{(n)}$, $(\tau^2)^{(n)}$, $(\sigma^2)^{(n)}$, $u^{(n)}$

1. For $i \in 1,\ldots I$
   - Generar:

$$U_i^{n+1} \mid y,\beta^{(n)},(\sigma^2)^{(n)},(\tau^2)^{(n)} \sim N\left(\frac{J(\bar{y}_i-\beta)}{J+(\tau^2)^{(n)}(\sigma^{-2})^{(n)}}, \left(J(\tau^{-2})^{(n)}+(\sigma^{-2})^{(n)}\right)^{-1}\right)$$

2. Generar:

$$\beta^{(n+1)} \mid u^{(n+1)},y,(\sigma^2)^{(n)},(\tau^2)^{(n)} \sim N\left(\bar{y}-\bar{u}^{(n+1)},(\tau^2)^{(n)}/JI\right)$$

3. Generar :

$$(\sigma^2)^{(n+1)} \mid u^{(n+1)},\beta^{(n+1)},y,(\tau^2)^{(n)} \sim IG\left(I/2,(1/2)\sum_i (u_i^2)^{(n+1)}\right)$$

4. Generar

$$(\tau^2)^{(n+1)} \mid u^{(n+1)},\beta^{(n+1)},y,(\sigma^2)^{(n+1)} \sim IG\left(IJ/2,(1/2)\sum_{i,j}\left(y_{ij}-(u_i)^{(n+1)}-\beta^{(n+1)}\right)^2\right)$$

Bajo ciertas condiciones de regularidad sobre el kernel de transición, Hobert y Casella (1996) han demostrado que si existe una función positiva b, $\varepsilon > 0$ y un conjunto compacto C tal que $b(x) < \varepsilon$ para $x \in C^c$, la cadena $\left(y^{(t)}\right)$ satisface

$$\liminf_{t\to+\infty} \frac{1}{t} \sum_{s=1}^{t} b\left(y^{(s)}\right) = 0.$$

0.24.3 Modelos Jerarquicos: Tumores de ratas

A continuación, consideramos los resultados de un estudio clínico sobre un tipo específico de tumor en ratas. El estudio consistió en 71 experimentos, donde el i-ésimo experimento consistió en contar el número de casos de tumor y_i entre las n_i ratas en ese experimento. Dado que todas las ratas en el conjunto de datos estaban en un grupo de control, no se les expuso a ningún tratamiento especial. Suponemos que el vector de tamaños de los grupos $n = (n_1, \ldots, n_{71})$ es conocido y fijo, y que $y = (y_1, \ldots, y_{71})$ son nuestros datos. Consideramos y como una realización de un vector estocástico $\mathbf{Y} = (Y_1, \ldots, Y_{71})$ donde cada Y_i tiene un espacio de estados $\{0, 1, 2, \ldots, n_i\}$.

Consideremos el siguiente modelo jerárquico:

1. La distribución de cada Y_i depende de una realización de una variable aleatoria $\theta_i \in (0,1)$, que interpretamos como la tasa de mortalidad en el i-ésimo grupo de ratas. Específicamente, asumimos que la distribución condicional de Y_i dado θ_i es binomial con parámetros θ_i y n_i.
2. Condicionado a $\theta = (\theta_1, \ldots, \theta_{71})$, asumimos que $Y_1, \ldots, Y_{71}$ son independientes.
3. La distribución de θ depende de una realización de dos variables aleatorias positivas A y B: dado $A = \alpha$ y $B = \beta$, asumimos que $\theta_1, \ldots, \theta_{71}$ son independientes y θ_i sigue una distribución beta con parámetros α y β.
4. Asumimos que la distribución conjunta de A y B tiene una densidad $\pi(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$.
5. Finalmente, asumimos que la distribución condicional de $\mathbf{Y}$ dado (Θ, A, B) no depende de (A, B).

Obtendremos ahora la distribución a posteriori, para lo cual necesitamos los siguientes componentes

Especificación de la distribución de los datos: Las suposiciones del modelo 1 y 2 implican que la densidad condicional de $\mathbf{Y}$ dado $\theta_1, \ldots, \theta_{71}$ es

$$\pi(y \mid \theta) = \prod_{i=1}^{71} \pi(y_i \mid \theta_i) = \prod_{i=1}^{71} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta)^{n_i - y_i}, \quad y_i \in \{0, 1, \ldots, n_i\}, i = 1, \ldots, 71,$$

donde $y = (y_1, \ldots, y_{71})$ y $\theta = (\theta_1, \ldots, \theta_{71})$. Debido a la suposición 5, $\pi(y \mid \theta, \alpha, \beta) = \pi(y \mid \theta)$ no depende de (α, β).

Especificación de la distribución a priori: Según la suposición 3, condicionado a $A = \alpha$ y $B = \beta$, la densidad condicional de θ es

$$\pi(\theta \mid \alpha, \beta) = \prod_{i=1}^{71} \pi(\theta_i \mid \alpha, \beta) = \prod_{i=1}^{71} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha - 1} (1 - \theta_i)^{\beta - 1}$$

donde $\theta = (\theta_1, \ldots, \theta_{71}) \in (0,1)^{71}$. La densidad a priori de (θ, A, B) se define como $\pi(\theta, \alpha, \beta) = \pi(\theta \mid \alpha, \beta)\pi(\alpha, \beta)$, donde $\pi(\alpha, \beta)$ se especifica en 4.

Especificación de la distribución a posteriori: Se obtiene que la distribución a posteriori de (θ, A, B) dado los datos $\mathbf{Y} = y$ tiene una densidad

$$\begin{aligned}\pi(\theta,\alpha,\beta \mid y) \propto & \pi(\alpha,\beta)\pi(\theta \mid \alpha,\beta)\pi(y \mid \theta) \\ = & \pi(\alpha,\beta)\prod_{i=1}^{71}\pi(\theta_i \mid \alpha,\beta)\prod_{i=1}^{71}\pi(y_i \mid \theta_i) \\ \propto & (\alpha+\beta)^{-5/2}\left\{\prod_{i=1}^{71}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\right\} \\ & \left\{\prod_{i=1}^{71}\binom{n_i}{y_i}\theta_i^{y_i}(1-\theta_i)^{n_i-y_i}\right\} \\ \propto & (\alpha+\beta)^{-5/2}\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{71}\prod_{i=1}^{71}\theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1} \end{aligned} \tag{78}$$

Especificación full condicionales: Podemos demostrar que:

A) Dado $(A,B)=(\alpha,\beta)$ y $\mathbf{Y}=y$, la distribución condicional de θ tiene una densidad

$$\pi(\theta \mid \alpha,\beta,y) \propto \prod_{i=1}^{71}\theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}, \quad \theta \in (0,1)^{71} \tag{79}$$

B) La distribución condicional (79) implica que, condicionado a $(A,B)=(\alpha,\beta)$ y $\mathbf{Y}=y$, tenemos que $\theta_1,\ldots,\theta_{71}$ son independientes y θ_i sigue una distribución beta con parámetros $\alpha+y_i$ y $\beta+n_i-y_i$.

$$f(x) \propto x^{a-1}(1-x)^{b-1} \qquad \text{para } 0<x<1$$

C) La distribución conjunta condicional de (A,B) dado θ y $\mathbf{Y}=y$ tiene una densidad no normalizada

$$\pi(\alpha,\beta \mid \theta,y) \propto (\alpha+\beta)^{-5/2}\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{71}\prod_{i=1}^{71}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}, \quad (\alpha,\beta) \in (0,\infty)^2 \tag{80}$$

Metropolis-Hastings con Gibbs Sampler: Dado que no es inmediato muestrear (A,B) directamente de (80), proponemos utilizar un algoritmo de Metropolis-Hastings dentro de un Gibbs Sampler para muestrear de la densidad a posteriori (78). El algoritmo de Metropolis-Hastings dentro del Gibbs Sampler tiene 72 componentes: $\theta_1,\ldots,\theta_{71}$ y (A,B). Según el punto A. anterior, podemos muestrear fácilmente (mediante un paso de Gibbs Sampler) de la distribución condicional de θ_i dado $(A,B)=(\alpha,\beta)$, pues esta distribución condicional no depende de $\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_{71}$. Para (A,B) proponemos utilizar una actualización de Metropolis-Hastings mediante una paseo aleatorio, donde la propuesta proviene de una distribución normal bivariada como se especifica a continuación.

A. 22: Modelo tumores de ratas

Dado $\theta_n = (\theta_{1,n}, \ldots, \theta_{71,n})$, $\sigma_\alpha^2 > 0$, $\sigma_\beta^2 > 0$, A_n y B_n.

1. Para $i = 1, \ldots, 71$
   - Generar $\theta_{i,n+1} \sim Beta(A_n + y_i, B_n + n_i - y_i)$.
2. Generar $A'_{n+1} \sim \mathrm{N}\left(A_n, \sigma_\alpha^2\right)$
3. Generar $B'_{n+1} \sim \mathrm{N}\left(B_n, \sigma_\beta^2\right)$
4. Generar $U_{n+1} \sim U(0,1)$.
5. Si $U_{n+1} < \pi\left(A'_{n+1}, B'_{n+1} \mid \theta_{n+1}, y\right) / \pi\left(A_n, B_n \mid \theta_{n+1}, y\right)$ entonces $(A_{n+1}, B_{n+1}) = \left(A'_{n+1}, B'_{n+1}\right)$, de lo contrario $(A_{n+1}, B_{n+1}) = (A_n, B_n)$.

Optimización estocástica

En este capítulo nos enfocaremos en estudiar, desde un enfoque basado en simulación, el problema de optimización

$$h_{min} = \min_{x \in \Theta} h(x), \tag{81}$$

donde h es una función a valores reales, donde Θ es un conjunto acotado, y cuyo mínimo se alcanza en $x_{min} \in \Theta$. A continuación, revisaremos heurísticas y modificaciones de los métodos tradicionales de la optimización para resolver (81).

0.25 Optimización Monte Carlo

El método más sencillo para encontrar mínimos se basaría en evaluar la función $h(x)$ en una serie de puntos obtenidos mediante algún mecanismo. Específicamente, si consideramos $u_1, \ldots, u_N \sim U(\Theta)$, entonces bastaría con considerar $\hat{h}_{min} = \min(h(u_1), \ldots, h(u_N))$. La convergencia del método está dada por el siguiente teorema.

Teorema 0.25.1 Sea U una variable aleatoria con distribución uniforme tal que $\text{Rec}\{X\} = \Theta$. Si la distribución de $h(U)$ existe y $\text{Rec}\{h(U)\} = [h_{min}, h_{max}]$, entonces $\hat{h}_{min}$ converge en distribución a h_{min} cuando $N \longrightarrow \infty$.

Demostración: Primero, debemos notar que $\hat{h}^*$ es una variable aleatoria con recorrido $[h_{min}, h_{max}]$ función de distribución

$$F_{\hat{h}^*}(x) = 1 - (1 - F(x))^N.$$

Luego, notamos que la distribución límite está dada por

$$F_{\hat{h}^*}(x) \longrightarrow \begin{cases} 0 & \text{si } x > h_{min} \\ 1 & \text{si } x = h_{min} \end{cases},$$

Finalmente, la demostración está completa si nos damos cuenta que el término de igualdad se obtiene debido a que el recorrido está acotado por h_{min}.

Notamos que este método es muy sencillo y solo requiere dos elementos: simular una distribución en el espacio Θ y ser capaces de evaluar la función h. El costo computacional de este método es proporcional a la dificultad de simular sobre Θ, la evaluación de h y la velocidad de convergencia del algoritmo.

Una idea para acelerar la velocidad del algoritmo, se basa en añadir información de la función h en el proceso de exploración del dominio Θ. Una forma de hacer esto es mediante la creación de una función de densidad que involucre a la función h. Por ejemplo, es posible considerar a la familia de distribuciónes $H(x) = e^{-h(x)}$ digamos c, entonces se podría simular directamente desde H y, en consecuencia, la secuencia $x_1, \ldots, x_N \sim c|H(x)|$ generaría valores cercanos a los extremos de la función con una mayor frecuencia, reduciendo el número de simulaciones requeridas.

0.26 Enjámbre de partículas

El método del enjámbre de partículas (del inglés *particle swarm*, presentado por Kennedy y R. Eberhart (1995)) es un método heurístico inspirado en el pricipio biológico de que *un enjambre que se mueve en conjunto se beneficia de la experiencia de sus otros miembros*. En otras palabras,w el algoritmo entrena un conjunto de valores iniciales (enjambre) mediante la exploración (evaluación) y una etapa de transferencia de información (actualización).

Formalmente, se selecciona el *enjambre* mediante los valores iniciales $u_1^0, \ldots, u_n^0$ (no necesariamente aleatorios) y se *explora* la función mediante la cantidad $\hat{h}_{min} = \text{mín}\{h(u_1), \ldots, h(u_n)\}$. Luego, se *actualizan* los elementos del enjambre mediante la siguiente regla

$$
\begin{aligned}
V_i^k &= \omega V_i^{k-1} + c_1 r_1 (p_i^k - u_i^k) + c_2 r_2 (g^k - u_i^k), \\
u_i^k &= u_i^{k-1} + V_i^k,
\end{aligned}
$$

donde V_i^k es la dirección de cambio, p_i^k es la mejor posición de la i-ésima partícula hasta la iteración k, mientras que g^k es la mejor posición global (Ver Algoritmo **??**). Las constantes ω, c_1, c_2 son hiperparámetros que se interpretan como

1. ω : es el grado de credibilidad de la velocidad anterior,
2. c_1 : es el grado de credibilidad de moverse hacia la mejor posición local,
3. c_2 : es el grado de credibilidad de moverse hacia la mejor posición global.

La investigación en este método se basa en determinar estos hiperparámetros ó en proponer secuencias de parámetros que hagan el algoritmo más automáticos, donde el lector es referido a Bonyadi y Michalewicz (2017); R. C. Eberhart y Shi (2000); Shi y R. Eberhart (1998))

En general, cualquier mejora para un método de optimización se requiere añadir información sobre la función h, lo cual se detalla a continuación.

0.27 Templado simulado (Simulated annealing)

La técnica del templado simulado (del inglés *simulated annealing*) fue introducida por Metropolis et al. (1953) y se basa en construir una cadena de Markov para poder resolver el Problema (81). El rendimiento del método se basa en la correcta selección de dos características:

Entrada: Función a minimizar $h(x)$ a minimizar, un dominio Θ, hiperparámetros $N0, \omega, c_1, c_2$
Salida: Una secuencia de valores $(u_i, h(u_i))$
begin
  Generar $u_1^0, \ldots, u_{N0}^0 \sim U(\Theta)$;
  Calcular $h(u_1^0), \ldots, h(u_{N_0}^0)$;
  Asignar $g^0 = \arg\min h(u_1^0), \ldots, h(u_{N_0}^0)$;
  Asignar $p_i^0 = u_i^0$;
  Fijar $V_i^0 = 0_d$;
  for $i = 0$ **to** K **do**
    Asignar $g^k = \arg\min h(u_1^k), \ldots, h(u_{N_0}^k)$;
    Asignar $p_i^k = \arg\min h(u_i^k), \ldots, h(u_i^0)$;
    Generar $r_1, r_2 \sim U([0,1])$;
    $V_i^k = \omega V_i^{k-1} + c_1 r_1 (p_i^k - u_i^k) + c_2 r_2 (g^k - u_i^k)$;
    $u_i^k = u_i^{k-1} + V_i^k$;
  end
end

Algorithm 3: Algoritmo de enjambre de partículas

- La selección del conjunto de vecinos: una mala selección del conjunto de vecinos podría no explorar el dominio Θ de forma eficiente. Es por esto que se busca un mecanísmo de generación de vecinos que permite explorar Θ de manera razonable.
- La actualización de la temperatura: la regla clave que permite explorar el espacio Θ fuera de los puntos críticos se basa en el parámetro de *temperatura*. A pesar de esto, se espera que la probabilidad de explorar otros estados vaya cambiando con respecto al número de iteraciones.

El pseudo código se describe en **??**.

0.28 Variantes aleatorias del gradiente conjugado

Uno de los métodos clásicos para minimizar una función es el gradiente conjugado. Recordamos que el método del gradiente conjugado, está dado por la secuencia generada por la iteración

$$x^k = x^{k-1} - \eta_k \nabla h(x^{k-1}). \tag{82}$$

Recordar que la iteración (82) se usa para encontrar mínimos locales del Problema (81). La convergencia depende de la elección del punto inicial x^0, la secuencia de constantes η_k y las propiedades de la función h. Debido a que se está utilizando información de primer orden, el método del gradiente conjugado suele converger a los puntos críticos, pero no a los mínimos globales.

0.28.1 Gradiente perturbado

El método del gradiente perturbado modifica las actualizaciones del gradiente conjugado mediante el agregar un ruido en cada iteración. El ruido se escoge de manera tal que uno pueda escapar de los candidatos a puntos silla. El Algoritmo **??** presenta una versión del gradiente perturbado que considera como candidatos a punto silla los valores $\|\nabla h(x)\| \leq g_{tol}$, además de darle un tiempo de espera para

Entrada: Función a minimizar $h(x)$ a minimizar, una regla para generar vecinos f, una regla de actualización de T, llamada τ
Salida: $(x_{max}, h(x_{max}))$
begin
  Generar un conjunto de vecinos iniciales $\mathcal{N} = \{x_1^0, \ldots, x_{N0}^0\}$, donde $x_i^0 \sim f_0$;
  Calcular $x_{min}^0 = \arg\min h(x_1^0), \ldots, h(x_{N0}^0)$;
  Calcular $h_{min}^0 = \min h(x_1^0), \ldots, h(x_{N0}^0)$;
  for $i = i$ **to** K **do**
    Generar, un nuevo conjunto de vecinos $\mathcal{N}_i = \{x_1^k, \ldots, x_{N0}^k\}$, donde $x_i^k \sim f_k$;
    Calcular $x^* = \arg\min_{x \in \mathcal{N}_i} h(x)$;
    Calcular $h^* = \min_{x \in \mathcal{N}_i} h(x)$;
    Calcular $\Delta h = h^* - h_{min}^k$;
    if $\Delta h \leq 0$ **then**
      Asignar $x_{min}^k = x^*$ y $h_{min}^k = h^*$ con probabilidad $\min\{1, e^{-\Delta h/T}\}$;
    end
    Actualizar $T = \tau(T)$
  end
end

Algorithm 4: Templado simulado

verificar la naturaleza del punto crítico. Para ver los detalles de la convergencia, el lector es referido a Jin et al. (2021).

0.28.2 Gradiente estocástico

La técnica del gradiente estocástico data de los años noventa y es introducida por primera vez en Bottou (1998). Este nace para reducir la cantidad de evaluaciones que se pueden tener debido a una cantidad enorme de datos o debido a que el gradiente de h es muy caro de evaluar. En la estructura aditiva, es común que se suele asociar el índice i a el i-ésimo dato. Es por esto que el método del gradiente estocástico propone actualizar cada iteración utilizando una sola observación, la cual será escogida al azar en cada paso iterativo, es decir

$$i \sim U(\{1, \ldots, N\})$$
$$x_k = x^{k-1} - \eta \sum_{i=1}^{N} \nabla h_i(x_i^{k-1}).$$

0.28.3 Gradiente por bloques aleatorios

Un alg

Entrada: Función a minimizar $h(x)$ a minimizar, un dominio Θ, punto inicial x_0, $\eta > 0$, tolerancias g_{tol} y ε, tiempo de reseteo τ
Salida: el par (x_{min}, h_{min})
begin
  $i_{perturb} = 0$;
  for $i = 1$ **to** K **do**
    if $\|\nabla h(x^{k-1})\| \leq g_{tol}$ *y* $i - i_{perturb} > \tau$ **then**
      Simular $\varepsilon_k \sim U(B(0,r))$;
      $x^{k-1} = x^{k-1} + \eta\varepsilon_k$;
    end
    $x^k = x^{k-1} - \eta\nabla h(x^{k-1})$;
  end
end

Algorithm 5: Algoritmo de gradiente perturbado

Bibliografía

Articles

[1] AC Atkinson. "The computer generation of Poisson random variables". En: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28.1 (1979), páginas 29-35.

[2] Yan Bai, Gareth O Roberts y Jeffrey Seth Rosenthal. "On the containment condition for adaptive Markov chain Monte Carlo algorithms". En: (2009).

[3] Julian Besag. "Spatial interaction and the statistical analysis of lattice systems". En: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), páginas 192-225.

[4] Peter J Bickel y David A Freedman. "Some asymptotic theory for the bootstrap". En: *The annals of statistics* 9.6 (1981), páginas 1196-1217.

[5] Mohammad Reza Bonyadi y Zbigniew Michalewicz. "Particle swarm optimization for single objective continuous space problems: a review". En: *Evolutionary computation* 25.1 (2017), páginas 1-54.

[6] James G Booth y James P Hobert. "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm". En: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.1 (1999), páginas 265-285.

[7] Léon Bottou. "Online algorithms and stochastic approximations". En: *Online learning in neural networks* (1998).

[8] George E. P. Box. "Science and Statistics". En: *Journal of the American Statistical Association* 71.356 (1976), páginas 791-799.

[10] Li Cai. "High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm". En: *Psychometrika* 75 (2010), páginas 33-57.

[11] Stamatis Cambanis, Steel Huang y Gordon Simons. “On the theory of elliptically contoured distributions”. En: *Journal of Multivariate Analysis* 11.3 (1981), páginas 368-385.

[12] KS Chan y Johannes Ledolter. “Monte Carlo EM estimation for time series models involving counts”. En: *Journal of the American Statistical Association* 90.429 (1995), páginas 242-252.

[13] Ming-Hui Chen y Bruce Schmeiser. “Toward black-box sampling: A random-direction interior-point Markov chain approach”. En: *Journal of Computational and Graphical Statistics* 7.1 (1998), páginas 1-22.

[14] Bernard Delyon, Marc Lavielle y Eric Moulines. “Convergence of a stochastic approximation version of the EM algorithm”. En: *Annals of statistics* (1999), páginas 94-128.

[15] Arthur P Dempster, Nan M Laird y Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. En: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), páginas 1-22.

[17] Persi Diaconis y Donald Ylvisaker. “Conjugate priors for exponential families”. En: *The Annals of statistics* (1979), páginas 269-281.

[20] B Efron. “The 1977 RIETZ lecture”. En: *The annals of Statistics* 7.1 (1979), páginas 1-26.

[22] Bradley Efron. “Better bootstrap confidence intervals”. En: *Journal of the American statistical Association* 82.397 (1987), páginas 171-185.

[23] Donald P Gaver y Iognaid G O’Muircheartaigh. “Robust empirical Bayes analyses of event rates”. En: *Technometrics* 29.1 (1987), páginas 1-15.

[24] Andrew Gelman, Walter R Gilks y Gareth O Roberts. “Weak convergence and optimal scaling of random walk Metropolis algorithms”. En: *The annals of applied probability* 7.1 (1997), páginas 110-120.

[25] Walter R Gilks et al. “Modelling complexity: applications of Gibbs sampling in medicine”. En: *Journal of the Royal Statistical Society: Series B (Methodological)* 55.1 (1993), páginas 39-52.

[26] Heikki Haario, Eero Saksman y Johanna Tamminen. “An adaptive Metropolis algorithm”. En: *Bernoulli* (2001), páginas 223-242.

[27] Peter Hall. “On the bootstrap and confidence intervals”. En: *The Annals of Statistics* (1986), páginas 1431-1452.

[28] Peter Hall. “Theoretical comparison of bootstrap confidence intervals”. En: *The Annals of Statistics* (1988), páginas 927-953.

[30] JA Hartigan. “Estimation by ranking parameters”. En: *Journal of the Royal Statistical Society: Series B (Methodological)* 28.1 (1966), páginas 32-44.

[32] Norbert Henze. “A Probabilistic Representation of the ’Skew-Normal’ Distribution”. En: *Scandinavian Journal of Statistics* 13.4 (1986), páginas 271-275.

[33] James P Hobert y G Casella. “The effect of improper priors on Gibbs sampling in hierarchical linear mixed models”. En: *Journal of the American Statistical Association* 91.436 (1996), páginas 1461-1473.

[34] Chi Jin et al. “On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points”. En: *Journal of the ACM (JACM)* 68.2 (2021), páginas 1-29.

[36] Yongman Kim y Kesar Singh. “Sharpening estimators using resampling”. En: *Journal of Statistical Planning and Inference* 66.1 (1998), páginas 121-146.

[37] Sungwon Lee y Sungkyu Jung. “Combined analysis of amplitude and phase variations in functional data”. En: *arXiv preprint arXiv:1603.01775* (2016).

[38] Richard A Levine y George Casella. “Implementations of the Monte Carlo EM algorithm”. En: *Journal of Computational and Graphical Statistics* 10.3 (2001), páginas 422-439.

[39] George Marsaglia. “The squeeze method for generating gamma variates”. En: *Computers & Mathematics with Applications* 3.4 (1977), páginas 321-325.

[40] Charles E McCulloch. “Maximum likelihood variance components estimation for binary data”. En: *Journal of the American Statistical Association* 89.425 (1994), páginas 330-335.

[41] Charles E McCulloch. “Maximum likelihood algorithms for generalized linear mixed models”. En: *Journal of the American statistical Association* 92.437 (1997), páginas 162-170.

[43] Kerrie L Mengersen y Richard L Tweedie. “Rates of convergence of the Hastings and Metropolis algorithms”. En: *The annals of Statistics* 24.1 (1996), páginas 101-121.

[44] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. En: *The journal of chemical physics* 21.6 (1953), páginas 1087-1092.

[46] Peter Miiller. “A generic approach to posterior integration and Gibbs sampling”. En: *Tech-nical Report* (1991), páginas 91-09.

[47] William H Press y Glennys R Farrar. “Recursive stratified sampling for multidimensional Monte Carlo integration”. En: *Computers in Physics* 4.2 (1990), páginas 190-195.

[48] Herbert Robbins y Sutton Monro. “A stochastic approximation method”. En: *The annals of mathematical statistics* (1951), páginas 400-407.

[49] Gareth O Roberts y Jeffrey S Rosenthal. “Examples of adaptive MCMC”. En: *Journal of computational and graphical statistics* 18.2 (2009), páginas 349-367.

[50] Gareth O Roberts y Richard L Tweedie. “Exponential convergence of Langevin distributions and their discrete approximations”. En: *Bernoulli* (1996), páginas 341-363.

[52] WR Schucany, HL Gray y DB Owen. “On bias reduction in estimation”. En: *Journal of the American Statistical Association* 66.335 (1971), páginas 524-533.

[53] YH Schukken, G Casella y J Van den Broek. “Overdispersion in clinical mastitis ata from dairy herds: a negative binomial approach”. En: *Preventive Veterinary Medicine* 10.3 (1991), páginas 239-245.

[55] Thomas A Severini. “On the relationship between Bayesian and non-Bayesian interval estimates”. En: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), páginas 611-618.

[58] Kesar Singh. “On the asymptotic accuracy of Efron’s bootstrap”. En: *The annals of statistics* (1981), páginas 1187-1195.

[59] Trevor J Sweeting. “Coverage probability bias, objective Bayes and the likelihood principle”. En: *Biometrika* 88.3 (2001), páginas 657-675.

[60] Luke Tierney. “Markov chains for exploring posterior distributions”. En: *the Annals of Statistics* (1994), páginas 1701-1728.

[62] Florin Vaida y Xiao-Li Meng. “Two slice-EM algorithms for fitting generalized linear mixed models with binary response”. En: *Statistical Modelling* 5.3 (2005), páginas 229-242.

[63] JC Wakefield et al. “Bayesian analysis of linear and non-linear population models by using the Gibbs sampler”. En: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.1 (1994), páginas 201-221.

[64] Greg CG Wei y Martin A Tanner. “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms”. En: *Journal of the American statistical Association* 85.411 (1990), páginas 699-704.

[65] Greg CG Wei y Martin A Tanner. “Posterior computations for censored regression data”. En: *Journal of the American Statistical Association* 85.411 (1990), páginas 829-839.

[66] CF Jeff Wu. “On the convergence properties of the EM algorithm”. En: *The Annals of statistics* (1983), páginas 95-103.

Libros

[9] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[18] Rick Durrett. *Probability: theory and examples*. Volumen 49. Cambridge university press, 2019.

[21] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[29] Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media, 2013.

[31] Trevor Hastie et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Volumen 2. Springer, 2009.

[42] Geoffrey J McLachlan y Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[45] Sean P Meyn y Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[51] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[54] Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

[56] Jun Shao y Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[61] D Michael Titterington et al. *Statistical analysis of finite mixture distributions*. Volumen 198. John Wiley & Sons Incorporated, 1985.

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Departamento de Matemática