



Métodos de Monte Carlo
Desarrollo Matemático
Marco Villacañas
Andrés Moguel
Inferencia Estadística I
Dr. Nelson Muriel
Licenciatura en Actuaría
Semestre Otoño 2021
12 de noviembre de 2021

CONTENIDO:

Generación de Variables Aleatorias	3
Integración de Monte Carlo	7
Controlando Varianza en Monte Carlo	14
Optimización de Monte Carlo	16

ABSTRACT:

En esta entrega se desarrolla la teoría que se considera esencial y necesaria para poder entender y aplicar los Métodos de Monte Carlo. Esta teoría será aplicada en la tercer entrega de este proyecto. En caso de que en la aplicación practica se utilice algo de teoría que no se incluye en estos temas, esta será desarrollada en el último entregable. Los temas fundamentales de Monte Carlo son los problemas de optimización e integración. Las secciones de generación de variables aleatorias y control de varianza son secciones auxiliares que consideramos importantes para una correcta explicación teórica.

Generación de Variables Aleatorias

Los métodos cubiertos en este desarrollo necesitan como base la posibilidad de producir, normalmente con ayuda de herramientas de cómputo, un supuesto flujo infinito de variables aleatorias para distribuciones conocidas. Esta simulación se basa en la generación de variables aleatorias uniformes. Se tomará un enfoque centrado en las estadísticas de producir estas variables aleatorias uniformes y de otros tipos.

Los métodos de simulación se basan en la generación de variables aleatorias, inicialmente independientes, distribuidas de acuerdo con una distribución f , que no necesariamente es explícitamente conocida. Comencemos con la generación de variables aleatorias uniformes en el intervalo $[0,1]$. Todas las demás distribuciones requieren de una secuencia de variables uniformes para poder ser simuladas.

Simulación Uniforme

Definición: Generador uniforme de números pseudo aleatorios

Un generador uniforme de números pseudo aleatorios es un algoritmo que dado un valor inicial u_0 y una transformación D , genera una secuencia $u_i = (D^i(u_0))$ de valores en el intervalo $[0,1]$. Para todo n , los valores (u_1, \dots, u_n) reproducen el comportamiento de una muestra independiente e idénticamente distribuida de variables aleatorias uniformes (V_1, V_n) al compararlas a través de unos ciertos experimentos.

Esta definición se restringe a los aspectos comprobables de la generación de variables aleatorias las cuales están relacionadas o conectadas por la transformación o relación de recurrencia:

$$u_i = D(u_{i-1}).$$

La validez del algoritmo consiste en verificar que la secuencia U_1, \dots, U_n lleva a la aceptación de la hipótesis:

$$H_0: U_1, \dots, U_n \text{ son iid } U_{[0,1]}$$

Se pueden utilizar distintas pruebas o experimentos para ver que los generadores sean adecuados. En otras palabras, lo que esta definición explica es que un algoritmo que genera valores uniformes es aceptable si no es rechazado por una serie de pruebas o experimentos.

La Transformación Inversa

Definición: inversa generalizada

Para una función no decreciente F sobre \mathbb{R} , la inversa generalizada de F , F^- , es la función definida como:

$$F^-(u) = \inf \{x: F(x) \geq u\}$$

De esta definición sale el siguiente lema, que establece una representación de cualquier variable aleatoria como una transformación de la variable aleatoria uniforme.

Lema:

Si $U \sim U_{[0,1]}$, entonces la variable aleatoria $F^-(U)$ se distribuye o tiene la distribución F .

Demostración:

$\forall u \in [0,1]$ y $\forall x \in F^-([0,1])$ la inversa generalizada satisface lo siguiente:

$$F(F^-(u)) \geq u \text{ y } F^-(F(x)) \leq x.$$

Por lo tanto, se tiene que:

$$\{(u, x): F^-(u) \leq x\} = \{(u, x): F(x) \geq u\}$$

y que;

$$P(F^-(U) \leq x) = P(U \leq F(x)) = F(x)$$

En otras palabras, lo que se necesita para poder generar una variable aleatoria $X \sim F$, es suficiente generar U de acuerdo con $\mathcal{U}_{[0,1]}$ y luego aplicar a estas variables uniformes la transformación

$$x = F^-(u).$$

El lema anterior también implica que una mala elección de generador para las variables aleatorias uniformes podría invalidar la simulación precedente.

Para aclarar este concepto de transformación inversa veamos un ejemplo con la distribución exponencial:

Ejemplo 1: Generación de Variables Aleatorias Exponenciales

Consideremos $X \sim \text{Exp}(1)$. Tiene la función de distribución: $F(x) = 1 - e^{-x}$. Entonces se plantea:

$$u = 1 - e^{-x}$$

y se procede a resolver para x :

$$x = -\log(1 - u)$$

Por lo tanto, si $U \sim \mathcal{U}_{[0,1]}$ y dado que tanto U y $U-1$ son variables aleatorias uniformes se tiene que: $X = -\log U$ se distribuye exponencial.

Este ejemplo ilustra que la generación de variables aleatorias uniformes es un elemento clave para determinar el comportamiento de los métodos de simulación para las demás distribuciones. Esto se debe a que todas las demás distribuciones F se pueden representar a través de una transformación de variables aleatorias uniformes.

Este método para generar variables aleatorias solamente puede ser utilizado si la función de distribución es explícitamente conocida o disponibles; es decir que existe un algoritmo que permite el cálculo de $F^-(u)$ en un espacio de tiempo aceptable. Para estos casos hay alternativas.

Métodos de Aceptación y Rechazo: *Accept-Reject Methods*

Existen distintas distribuciones para las cuales es muy difícil o incluso imposible simular mediante la transformación inversa. En otras ocasiones o casos la distribución dada no se puede representar de una manera útil. En ambos casos no es posible utilizar propiedades de probabilidad para encontrar o derivar un método adecuado de simulación. A continuación, serán desarrollados otros métodos que solamente requieren conocer la forma funcional de la densidad f de interés hasta una constante multiplicativa, no se necesita estudiar o conocer f a fondo. La base de este método es utilizar una densidad más simple, g para la cuál se simula. A esta densidad g se le conoce como la densidad instrumental.

Definición: densidad instrumental y densidades objetivo

Dada una función g , la función instrumental, existen muchas densidades f , conocidas como densidades objetivo, que se pueden simular a través de la función g . El algoritmo correspondiente para este proceso, El Algoritmo de Aceptación y Rechazo se basa en una conexión con la distribución uniforme.

El Teorema Fundamental de Simulación:

Sea f la densidad de interés, en un espacio arbitrario, entonces se puede escribir como:

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

Por lo tanto; f aparece como la densidad marginal (de X) en la distribución conjunta:

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

Dado que, U no se relaciona directamente con el problema, esta se define como una variable auxiliar.

Al introducir una variable auxiliar uniforme en la expresión de f , tenemos un nuevo planteamiento. Dado que (X, U) es la densidad conjunta de X y de U , se puede generar a partir de ella una al generar variables aleatorias uniformes sobre el conjunto restringido:

$$\{(x, u): 0 < u < f(x)\}.$$

Además, ya que la distribución marginal de X es la distribución original f , al generar una variable aleatoria uniforme sobre el conjunto restringido mencionado anteriormente, se ha generado una variable aleatoria a partir de f . Esta variable fue generada sin utilizar f sino a través de el cálculo de $f(x)$. Está equivalencia es importante dado el siguiente teorema:

Teorema: (Teorema fundamental de la simulación)

Simular:

$$X \sim f(x)$$

Es equivalente a simular:

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

Hay casos en los cuales la simulación de (X, U) no es tan sencilla. Tomemos por ejemplo que se podría simular $X \sim f(x)$ y $U|X = x \sim \mathcal{U}(0, f(x))$ pero al hacer esto, la representación dada por el teorema anterior es básicamente inutilizable. La solución a este problema es simular la pareja (X, U) en un conjunto más grande, en el que se pueda simular con mayor facilidad, y luego tomar la pareja si la restricción se cumple.

Por ejemplo, tomemos un espacio de una dimensión y supongamos lo siguiente:

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

y que f está acotada por m . Entonces se puede simular la pareja aleatoria $(Y, U) \sim \mathcal{U}(0 < u < m)$ al simular $Y \sim \mathcal{U}(a, b)$ y $U|Y = y \sim \mathcal{U}(0, m)$ y tomando la pareja solamente si la restricción $0 < u < f(y)$ se satisface. Esto resulta en la distribución correcta de el valor aceptado de Y que podemos llamarlo X dado que:

$$P(X \leq x) = P(Y \leq x | U < f(Y))$$

$$= \frac{\int_a^x \int_0^{f(y)} du dy}{\int_a^b \int_0^{f(y)} du dy} = \int_a^x f(y) dy.$$

Esto es lo mismo que decir que si $A \subset B$ y se genera una muestra uniforme sobre B , manteniendo solamente los términos de esta muestra que también se encuentran en A , resulta en una muestra aleatoria sobre A . Es importante notar que tiene un tamaño aleatorio que es independiente de los valores de la muestra.

El teorema anterior se puede generalizar para contemplar conjuntos de mayor tamaño. Esto puede permitir simular para casos donde el soporte de f y/o el valor máximo de f no se encuentren acotados.

Si el conjunto de mayor tamaño es de la forma:

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

entonces las restricciones son de que $m(x) \geq f(x)$ y que la simulación de una uniforme sobre \mathcal{L} es posible. La eficiencia dicta que m debe ser tan próxima a f como sea posible, para evitar gastar simulaciones. Una nota importante es que dada la restricción $m(x) \geq f(x)$, m no puede ser una densidad de probabilidades. Por eso definimos:

$$m(x) = Mg(x) \text{ donde } \int_x m(x) dx = \int_x Mg(x) dx = M,$$

dado que m es necesariamente integrable. Podemos utilizar el teorema fundamental de simulación al revés para simular la uniforme sobre \mathcal{L} , es decir simular $Y \sim g$ y luego $U|Y = y \sim \mathcal{U}(0, Mg(y))$. Si solamente se aceptan las y 's tales que se satisface la restricción $u < f(y)$ se tiene que:

$$\begin{aligned} P(X \in \mathcal{A}) &= P(Y \in \mathcal{A} | U < f(Y)) \\ &= \frac{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy}{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy} = \int_{\mathcal{A}} f(y) dy \end{aligned}$$

Para cualquier conjunto medible \mathcal{A} y las X 's aceptadas se distribuyen de f . Dado este desarrollo se ha obtenido una generalización del teorema fundamental dada por el siguiente corolario:

Corolario:

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

$$Y \sim g, U|Y = y \sim \mathcal{U}(0, Mg(y)) \text{ hasta } 0 < u < f(y).$$

Este corolario tiene dos consecuencias importantes. Provee un método genérico para simular cualquier densidad f que es conocida hasta un factor multiplicativo. Es decir que la constante que normaliza f no tiene que ser conocida ya que el método solamente requiere la entrada de f/M , el cual no depende de la constante de normalización. La segunda consecuencia es que la probabilidad de aceptación es exactamente $1/M$ y el número esperado de intentos hasta que una variable sea aceptada es M .

El Algoritmo de Aceptación y rechazo:

La implementación del corolario anterior es conocido como el método de Aceptación y rechazo. El algoritmo de Aceptación y Rechazo consta de los siguientes pasos:

1. Generar $X \sim g, U \sim \mathcal{U}_{[0,1]}$;
2. Aceptar $Y = X$ si $U \leq f(x)/Mg(X)$;
3. Si no se cumple la condición en (2) regresar a (1)

Algunas observaciones importantes de este algoritmo:

1. En los casos donde f y g están normalizadas, para que ambas sean densidades de probabilidades, la constante M es necesariamente mayor que 1.
2. El tamaño de M y la eficiencia del algoritmo se vuelven una función de que tanto puede imitar o aproximarse g a f .
3. Para que f/g se mantenga acotado es necesario que g tenga colas más gruesas que las de f .

Con esto se termina el desarrollo acerca de generación de variables aleatorias y se comenzará a desarrollar la teoría que es propiamente de los Métodos de Monte Carlo. Se desarrollarán los temas de integración e optimización.

Integración de Monte Carlo

Introducción

Los problemas y aplicaciones numéricas de la estadística inferencial son aquellos de optimización e integración. A través de métodos y aproximaciones como la máxima verosimilitud, el Bayesiano, método de momentos o incluso los métodos de bootstrap, se puede llegar a calcular analíticamente un estimador. Sin embargo, hay algunos problemas en los que es muy difícil o imposible llegar a calcular estos estimadores asociados a un paradigma dado.

Muchos cálculos de inferencia Bayesiana requieren integración. En general, la estimación de Bayes bajo la función de pérdida $L(\theta, \delta)$ y el π anterior es la solución del programa de minimización:

$$\min_{\delta} \int_{\theta} L(\theta, \delta) \pi(\theta) f(x|\theta) d\theta$$

El estimador de Bayes será una esperanza posterior solamente cuando la función de pérdida es la función cuadrática $\|\theta - \delta\|^2$. Hay casos en donde una configuración específica de la función de pérdidas que se construye mediante una toma de decisiones, excluye la solución analítica de la integración del problema de minimización. Ya que no es posible integrar directamente, se necesita llegar a una solución aproximada del programa de minimización a través de métodos numéricos o mediante simulación.

Sin importar con que tipo de inferencia estadística se esté trabajando, esta nos llevará a considerar soluciones numéricas. El tema anterior de generación de variables aleatorias provee una base para la construcción de soluciones de distintos problemas en estadística. La estadística inferencial sobre modelos complejos resultará muchas veces en la necesidad de usar técnicas de simulación. Veamos un ejemplo de esto para luego desarrollar métodos de integración basados en simulación.

Ejemplo 2: Funciones de Pérdida Lineales y Cuadráticas por partes

Consideremos la función de pérdida cuadrática por partes:

$$L(\theta, \delta) = w_i(\theta - \delta)^2 \text{ cuando } \theta - \delta \in [a_i, a_{i+1}), w_i > 0$$

Al diferenciar la esperanza posterior, la función de pérdida muestra que el estimador de Bayes asociado satisface:

$$\sum_i w_i \int_{a_i}^{a_{i+1}} (\theta - \delta^\pi(x)) \pi(\theta|x) d\theta = 0.$$

es decir,

$$\delta^\pi(x) = \frac{\sum_i w_i \int_{a_i}^{a_{i+1}} \theta \pi(\theta) f(x|\theta) d\theta}{\sum_i w_i \int_{a_i}^{a_{i+1}} \pi(\theta) f(x|\theta) d\theta}$$

Formalmente, la computación de $\delta^\pi(x)$ requiere la computación de las medias posteriores restringidas a los intervalos $[a_i, a_{i+1})$ y de las probabilidades posteriores de los mismos.

Una solución general para distintos problemas computacionales como el del ejemplo anterior es usar simulación. Se pueden simular las distribuciones dadas o aproximadas para calcular las cantidades deseadas.

Integración Clásica de Monte Carlo

Antes de poder aplicar técnicas de simulación a problemas más prácticos es necesario desarrollar algunas de sus propiedades. Se puede empezar con el problema pertinente a la siguiente integral:

$$\mathbb{E}_f[h(X)] = \int_X h(x) f(x) dx$$

Se puede proponer usar una muestra (X_1, \dots, X_m) generada utilizando la densidad f para aproximar la ecuación anterior a través del promedio empírico. A este método se le conoce como el método de Monte Carlo. Definimos el promedio empírico:

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^m h(x_j),$$

Dado que \overline{h}_m converge casi-seguro a $\mathbb{E}_f[h(X)]$ por la ley fuerte de los grandes números. Además, cuando h^2 tiene esperanza finita bajo f , la velocidad con la que converge \overline{h}_m puede ser medida dado que la varianza:

$$\text{var}(\overline{h}_m) = \frac{1}{m} \int_X (h(x) - \mathbb{E}_f[h(X)])^2 f(x) dx$$

También puede ser estimada con la muestra (X_1, \dots, X_m) mediante:

$$v_m = \frac{1}{m^2} \sum_{j=1}^m [h(x_j) - \overline{h}_m]^2.$$

Cuando m es “grande” se tiene:

$$\frac{\overline{h}_m - \mathbb{E}_f[h(X)]}{\sqrt{v_m}}$$

y por lo tanto se puede observar que se distribuye como una variable normal estándar, $N(0,1)$. Esto lleva a la construcción de una prueba de convergencia y de intervalos de confianza sobre la aproximación de $\mathbb{E}_f[h(X)]$. A continuación veamos un ejemplo de una primera idea de cómo usar la integración de Monte Carlo.

Ejemplo 3: Una primera integración de Monte Carlo

Consideremos la función: $h(x) = [\cos 50x + \sin 20x]^2$. Como un primer ejemplo consideremos integrar la función en $[0,1]$. Esto se muestra en la primer imagen de la figura 1. Aunque es posible integrar la función dada de forma analítica, es un buen ejemplo para una primer aproximación e introducción a Monte Carlo. Para calcular la integral con este método, primero se generan U_1, U_2, \dots, U_n variables aleatorias i.i.d que se distribuyen $\mathcal{U}_{[0,1]}$ y se aproxima $\int h(x)dx$ utilizando $\sum h(U_i)/n$. La imagen del centro de la figura 1 muestra un histograma con los valores de $h(U_i)$. La última imagen de esta figura muestra medios de ejecución y errores estándar. De esta imagen se puede observar que el promedio de Monte Carlo converge con el valor de 0.963 tras 10,000 iteraciones. Si lo comparamos con el valor exacto de 0.965, se puede concluir que la estimación con Monte Carlo se compara favorablemente.

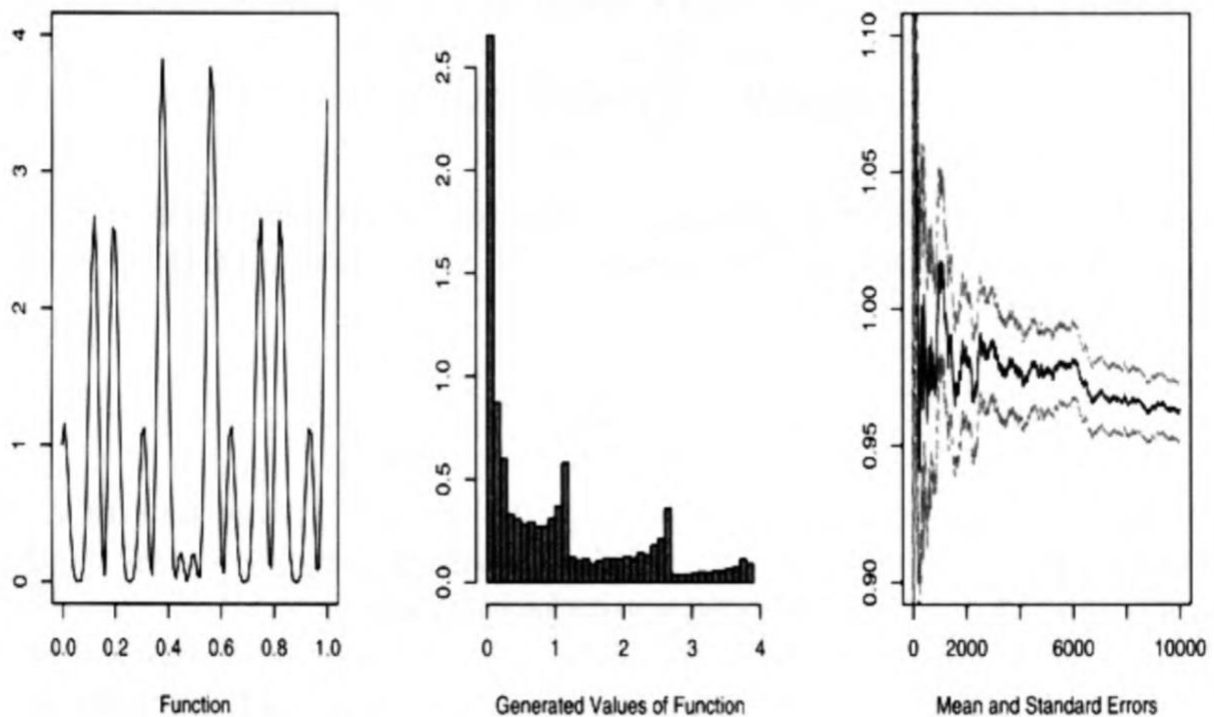


Figura 1 Cálculo de la integral de $h(x)$: izq.: función, centro: histograma de $h(U)$ con generación uniforme, der: media y error estándar.

Parece ser que el método propuesto anteriormente es suficiente para evaluar integrales como la que fue planteada al principio de la sección, que define a $\mathbb{E}_f[h(X)]$, de manera controlada. Sin embargo, mientras que el método de Monte Carlo directo provee una buena aproximación de la integral ya mencionada en varios casos y ejemplos, existen algunas alternativas más eficientes

que no sólo evitan una simulación directa de f pero que también pueden ser utilizadas repetidamente para varias integrales que definen a $\mathbb{E}_f[h(X)]$. Este uso repetido puede ser para una familia de funciones h o una familia de densidades f .

Muestreo de Importancia

Principios:

El método desarrollado a continuación se conoce como muestreo de importancia ya que se basa en las llamadas funciones de importancia. La evaluación de la integral para $\mathbb{E}_f[h(X)]$ basada en la simulación de f no necesariamente es siempre óptima y hasta se puede decir que es sub óptima. Es importante considerar que dicha integral se puede representar en una infinidad de maneras con la tripleta (X, h, f) . Por lo tanto la búsqueda de un estimador óptimo debe considerar todas estas representaciones posibles. Será importante considerar que, en el primer teorema establecido en esta sección, el método que óptimo que propone depende de la función h para la que se define en $\mathbb{E}_f[h(X)]$. Por lo tanto, no se puede considerar como óptimo si se evalúan distintas integrales relacionadas a f simultáneamente. La principal alternativa a hacer un muestreo directamente de f para luego evaluar la integral de $\mathbb{E}_f[h(X)]$, es utilizar el muestreo de importancia definido a continuación.

Definición: Muestreo de Importancia

El método del muestreo de importancia es una evaluación de $\mathbb{E}_f[h(X)] = \int_X h(x) f(x) dx$ basada en generar una muestra X_1, \dots, X_m de una distribución g dada y aproximar:

$$\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j).$$

Este método se basa en que podemos representar a $\mathbb{E}_f[h(X)] = \int_X h(x) f(x) dx$ como:

$$\mathbb{E}_f[h(X)] = \int_X h(x) \frac{f(x)}{g(x)} g(x) dx$$

a la cual se le conoce como la identidad fundamental del muestreo de importancia. El estimador de la aproximación de $\mathbb{E}_f[h(X)]$ converge a $\mathbb{E}_f[h(X)] = \int_X h(x) f(x) dx$ por la misma razón por la cual el estimador regular de Monte Carlo, $\overline{h_m}$, converge. Esto es sin importar la elección de la distribución de g mientras se cumpla que:

$$\text{supp}(g) \supset \text{supp}(f)$$

Es importante hacer la nota de que la identidad fundamental del muestreo de importancia es una representación muy general que expresa el hecho de que una integral dada no está intrínsecamente asociada con la distribución dada. El muestreo de importancia es de particular interés ya que impone pocas restricciones sobre la elección de la distribución g . Además, la muestra generada a partir de g puede ser utilizada repetidamente para distintas funciones h , así como distintas densidades f .

Estimadores de Varianza Finita

Aunque la distribución g puede ser casi cualquier densidad para que el estimador $\mathbb{E}_f[h(X)] \approx$

$\frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$ converja, hay algunas elecciones que son mejores que las demás. Por lo tanto,

es natural intentar comparar distintas distribuciones de g para la evaluación de $\mathbb{E}_f[h(X)] = \int_X h(x) f(x) dx$. Es importante notar que, aunque $\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$ si converge en casi-seguramente a $\mathbb{E}_f[h(X)] = \int_X h(x) f(x) dx$ su varianza solamente es finita cuando la esperanza:

$$\mathbb{E}_g \left[h^2(X) \frac{f^2(X)}{g^2(X)} \right] = \mathbb{E}_f \left[h^2(X) \frac{f(X)}{g(X)} \right] = \int_X h^2(X) \frac{f^2(X)}{g(X)} dx < \infty.$$

Por lo tanto, distribuciones con colas más ligeras que las de f , no son apropiadas para aplicar muestreo de importancia. Además, en estos casos las varianzas de los estimadores para $\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$, tienen varianza infinita para alguna funciones de h . Una alternativa a esto y que resuelve el problema de varianzas infinitas y genera un estimador más estable es utilizar:

$$\frac{\sum_{j=1}^m h(x_j) f(x_j) / g(x_j)}{\sum_{j=1}^m f(x_j) / g(x_j)}$$

Aunque este estimador sea sesgado, es sesgo es pequeño y la mejora en la varianza lo hace preferible en vez de $\frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$. A continuación enunciamos un teorema que relaciona g , la varianza y este último estimador.

Teorema:

La elección de g que minimiza la varianza del estimador $\frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$ es:

$$g^*(x) = \frac{|h(x)|f(x)}{\int_X |h(z)|f(z)dz}$$

Demostración:

Primero nótese que:

$$\text{var} \left[\frac{h(X)f(X)}{g(X)} \right] = \mathbb{E}_g \left[\frac{h^2(X)f^2(X)}{g^2(X)} \right] - \left(\mathbb{E}_g \left[\frac{h(X)f(X)}{g(X)} \right] \right)^2$$

y el segundo término no depende de g . Entonces para minimizar la varianza solamente es necesario minimizar el primer término. De la desigualdad de Jensen tenemos que:

$$\mathbb{E}_g \left[\frac{h^2(X)f^2(X)}{g^2(X)} \right] \geq \left(\mathbb{E}_g \left[\frac{|h(X)|f(X)}{g(X)} \right] \right)^2 = \left(\int |h(x)|f(x)dx \right)^2$$

Esto provee una cota inferior que es independiente de la elección de g . Con esto podemos decir que esta cota inferior se obtiene al elegir $g = g^*$.

Este resultado es algo formal dado que cuando $h(x) > 0$, la elección óptima de $g^*(x)$ requiere que se conozca $\int h(x)f(x)dx$. Una alternativa, que nos permite el teorema anterior es utilizar el estimador $\frac{\sum_{j=1}^m h(x_j)f(x_j)/g(x_j)}{\sum_{j=1}^m f(x_j)/g(x_j)}$ como:

$$\frac{\sum_{j=1}^m h(x_j) f(x_j) / g(x_j)}{\sum_{j=1}^m f(x_j) / g(x_j)} = \frac{\sum_{j=1}^m h(x_j) |h(x_j)|^{-1}}{\sum_{j=1}^m |h(x_j)|^{-1}}$$

Dónde $x_j \sim g \propto |h|/f$. Es importante resaltar que el numerador es el número de veces que $h(x_j)$ es positiva menos las veces que es negativa. En particular cuando h es positiva la equivalencia anterior es la media armónica. También hay que notar que este nuevo estimador es sesgado y puede ser inestable.

Si se toma una visión más práctica, el teorema anterior sugiere buscar distribuciones g para las cuales $|h|f/g$ es casi constante y tiene varianza finita. Es importante notar que, aunque no es necesario que la varianza sea finita para que $\frac{\sum_{j=1}^m h(x_j) |h(x_j)|^{-1}}{\sum_{j=1}^m |h(x_j)|^{-1}}$ y $\frac{1}{m} \sum_{j=1}^m \frac{f(x_j)}{g(x_j)} h(x_j)$ converjan, el muestreo de importancia obtiene resultados muy pobres cuando se tiene que:

$$\int \frac{f^2(x)}{g(x)} dx = +\infty$$

ya sea en términos del comportamiento del estimador (brincos de amplitud, inestabilidad en el promedio, convergencia lenta) o de comparaciones con Métodos de Monte Carlo directos. No se recomienda utilizar distribuciones g que cumplen con el resultado anterior.

Muestreo de Importancia y Aceptación-Rechazo

El teorema enunciado en la sección anterior soluciona formalmente el problema de comparar el método de Aceptación y Rechazo con el muestreo de importancia. Esto es dado que a excepción de funciones de h que son constantes, la densidad óptima g^* siempre es diferente de f . Una comparación más útil y realista sería tomar en cuenta que el teorema ya mencionado tiene aplicaciones limitadas en un aspecto práctico ya que prescribe una densidad instrumental que depende de la función de interés h . Esto no permite que se pueda reutilizar la muestra generada para estimar distintas cantidades deseadas. Cuando los métodos de Aceptación y Rechazo se implementan con una densidad g tal que satisface $f(x) \leq M g(x)$ para una constante $1 < M < \infty$, la densidad g puede fungir como la densidad instrumental del muestreo de importancia. Algo que es buena noticia es que f/g está acotado, asegurando la finites de la varianza para los estimadores del método de muestreo de importancia correspondientes. Es importante recordar que, en el método de Aceptación y Rechazo, la muestra resultante, X_1, \dots, X_n es un subconjunto de la muestra Y_1, \dots, Y_t , donde las Y_i 's se simulan a partir de g y dónde t es el número aleatorio de simulaciones de g necesarias para producir las n variables de f .

Para poder hacer una comparación de estimación utilizando el método de Aceptación y Rechazo y estimación usando muestreo de importancia es entonces razonable comenzar con los dos estimadores ya conocidos:

$$\delta_1 = \frac{1}{n} \sum_{i=1}^n h(X_i) \quad y \quad \delta_2 = \frac{1}{t} \sum_{j=1}^t \frac{f(Y_j)}{g(Y_j)} h(Y_j)$$

Estos estimadores corresponden a la aplicación directa a una muestra producida por el método de Aceptación y Rechazo y a una estimación de muestreo de importancia derivado de la muestra, es decir a un reciclado de las variables rechazadas por el algoritmo de Aceptación y Rechazo. Si

el cociente de f/g solamente es conocido hasta una cierta constante, entonces δ_2 puede ser remplazado por:

$$\delta_3 = \frac{\sum_{j=1}^t h(Y_j) f(Y_j) / g(Y_j)}{\sum_{j=1}^t f(Y_j) / g(Y_j)}$$

y si escribimos δ_2 de una forma más explícita:

$$\delta_2 = \frac{n}{t} \left\{ \frac{1}{n} \sum_{i=1}^n h(X_i) \frac{f(X_i)}{g(X_i)} + \frac{t-n}{n} \frac{1}{t-n} \sum_{i=1}^{t-n} h(Z_i) \frac{f(Z_i)}{g(Z_i)} \right\},$$

donde $\{Y_1, \dots, Y_t\} = \{X_1, \dots, X_n\} \cup \{Z_1, \dots, Z_{t-n}\}$; las Z_i siendo las variables rechazadas por el algoritmo de Aceptación y Rechazo. Basándonos en el tamaño de muestra la varianza de δ_2 es mas pequeña que la del estimador:

$$\frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i).$$

Si se pudiese aplicar el teorema enunciado en la sección anterior, se podría entonces concluir que este último estimado domina δ_1 (para una elección apropiada de g) y, por lo tanto, es mejor reciclar las Z_i en lugar de descartarlas. Sin embargo, este razonamiento no es correcto dado que t es una variable aleatoria y es la regla de detención del algoritmo de Aceptación y Rechazo. La distribución de t es una binomial negativa con parámetros n , $1/M$ y por lo tanto $\{Y_1, \dots, Y_t\}$ no es una muestra independiente e idénticamente distribuida de g .

Comparar δ_1 y δ_2 se puede reducir a comparar $\delta_1 = f(y_t)$ y δ_2 para t que se distribuye geométrica de parámetro $1/M$ y con $n = 1$. Con todo y dicha simplificación, poder comparar δ_1 y δ_2 es complicado y por lo tanto resulta difícil comparar el sesgo y la varianza de δ_2 con $var_f(h(X))$.

Mientras que el estimador δ_2 se basa en una representación errónea de la distribución de $\{Y_1, \dots, Y_t\}$, una alternativa razonable basada en la distribución apropiada para la muestra es:

$$\delta_4 = \frac{n}{t} \delta_1 + \frac{1}{t} \sum_{j=1}^{t-n} h(Z_j) \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)}$$

Donde las Z_j son los elementos de $\{Y_1, \dots, Y_t\}$ que han sido rechazados. Este estimador es insesgado y se puede comparar con δ_1 cuando $n = 1$. Esto es a través de ña comparación de las varianzas de $h(X_1)$ y de δ_4 que ahora lo podemos escribir como:

$$\delta_4 = \frac{1}{t} h(X_1) + (1 - \rho) \frac{1}{t} \sum_{j=1}^{t-1} h(Z_j) \left(\frac{g(Z_j)}{f(Z_j)} - \rho \right)^{-1}$$

Ahora asumiendo que $\mathbb{E}_f[h(X)] = 0$, la varianza de δ_4 es:

$$var(\delta_4) = \mathbb{E} \left[\frac{t-1}{t^2} \int h^2(x) \frac{(M-1)f^2(x)}{Mg(x) - f(x)} dx + \frac{1}{t^2} \mathbb{E}_f[h^2(X)] \right]$$

Este desarrollo ya es demasiado específico para ciertos casos, es decir que depende demasiado de f , g , y h para poder hacer una comparación general.

La distribución marginal de las Z_i 's del algoritmo de Aceptación y Rechazo es $\frac{Mg-f}{M-1}$, y el estimador de muestreo de importancia asociado a esta distribución es:

$$\delta_5 = \frac{1}{t-n} \sum_{j=1}^{t-n} h(Z_j) \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)}$$

y este estimador por lo tanto nos permite representar a δ_4 como:

$$\delta_4 = \frac{n}{t} \delta_1 + \frac{t-n}{t} \delta_5,$$

el cual es un promedio ponderado del estimador usual de Monte Carlo y de δ_5 .

De acuerdo con el teorema enunciado en la sección de estimadores de varianza finita, la distribución instrumental puede ser elegida tal que la varianza de δ_5 es menor o más baja que la varianza de δ_1 . Dado que dicho estimador es insesgado, δ_4 dominará a δ_1 para una elección apropiada de g . Este resultado quiere decir que para una función g fija, existen funciones h tales que δ_4 mejora sobre δ_1 .

Ahora si f solamente es conocida hasta una constante de integración, se puede reemplazar a δ_4 con:

$$\delta_6 = \frac{n}{t} \delta_1 + \frac{t-n}{t} \sum_{j=1}^{t-n} \frac{h(Z_j)f(Z_j)}{Mg(Z_j) - f(Z_j)} / \sum_{j=1}^{t-n} \frac{f(Z_j)}{Mg(Z_j) - f(Z_j)}$$

Aunque el dominio de δ_1 mediante δ_4 nos se extiende para δ_6 , δ_6 estima correctamente funciones constantes y es asintóticamente equivalente a δ_4 .

En la próxima sección se comentará más a fondo sobre las varianzas y el control de las mismas en las estimaciones.

Controlando Varianza en Monte Carlo

Monitoreo de la variación con el Teorema del Limite Central (TLC)

Teorema de limite central evaluado la convergencia al estimador Monte Carlo,

$$\bar{h}_{.m} = \frac{1}{m} \sum_{j=1}^m h(X_j) \quad X_j \sim f(x)$$

Integral de interés

$$J = \int h(x)f(x)dx$$

Monitoreo Univariado

Cuando se considera evaluar la integral de $h(x) = \cos(50x) + \sin(20x)]^2$ con parámetros 0,1, resulta en convergencia con un error estándar, moviéndose de una manera no racional, es decir, las orillas exhiben unas ondulaciones desiguales que la estimación puntual. Si, en cambio, producimos secuencias paralelas de estimaciones, obtenemos el resultado de la integral J , el rango y la banda empírica del 90% (derivados del conjunto de estimaciones en cada iteración

tomando los cuantiles empíricos del 5% y el 95%) son mucho más amplios que el intervalo de confianza del 95% predicho por el TLC, donde se calculó la varianza.

Este simple ejemplo advierte aún más contra el uso de una aproximación normal cuando se usa repetidamente sobre iteraciones con estimadores dependientes, simplemente porque la aproximación de confianza normal solo tiene una validación marginal y estática. El uso de una banda de estimadores en paralelo es obviamente más costoso, pero proporciona la evaluación correcta de la variación de estos estimadores.

Ejemplo 4: Cauchy

Para estimar una media normal, se puede conseguir cierto grado de robustez con un Cauchy previo, por lo que tenemos el modelo.

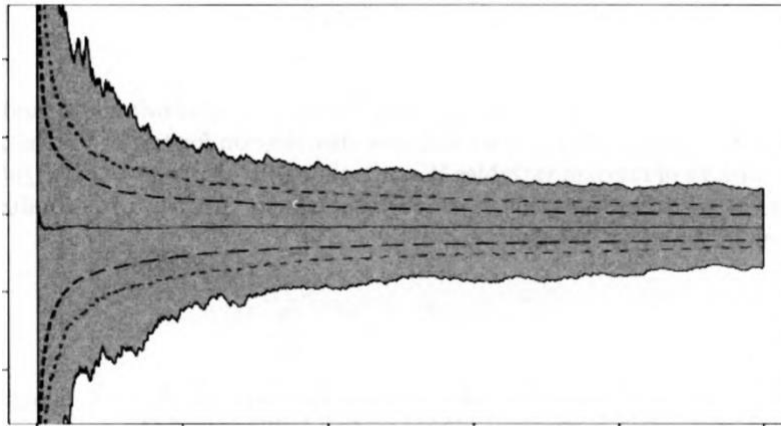


Figura 2:

Convergencia de 1000 iteraciones de la integral $h(x) = \text{Cos}(50x) + \sin(20x)]^2$ en donde observamos la línea recta es el promedio de las 1000 iteraciones, la línea con puntos es el 90% empírico de la banda y la línea punteada es el 95% del intervalo de confianza. Y el color grs representa todo el rango de las iteraciones.

Monitoreo Multivariado

Como se mencionó en la introducción de este capítulo, un método válido para adjuntar varianzas a una gráfica, y para tener un teorema del límite central válido, es derivar los límites usando un enfoque multivariado. Aunque el cálculo completo no es difícil.

Suponga que X_1, X_2, \dots es una secuencia de variables aleatorias independientes (iid) que se simulan con el objetivo de estimar $\mu = Ef(x_1)$ trabajamos con X_i , cuando normalmente estamos interesados en $h(X_i)$ para algunas funciones de h . Se define $\bar{X}_m = \left(\frac{1}{m}\right) \sum_{i=1}^m x_i$, para $m=1, 2, \dots, n$, donde n es el número de variables aleatorias que simularemos. Una gráfica de media continua. Asumimos $x_i \sim N(0, \sigma^2)$ independientes con una distribución $\bar{X}_m = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n)$ Dado que cada elemento de esta variable aleatoria tiene media μ , un intervalo de confianza simultáneo, basado en la distribución normal multivariante, será una evaluación válida de la varianza.

Sea $\mathbf{1}$ el vector $n \times 1$ de unidades. Entonces $E[X] = \mathbf{1}\mu$.

$$\text{cov}(X_m, X_m) = \frac{\sigma^2}{\max(k, k')}$$

Optimización de Monte Carlo

Introducción

Similar al problema de la integración, las diferencias entre los valores numéricos enfoque y el enfoque de simulación del problema

$$\max h(\theta)$$

Tratando en la función h . Al abordar un problema de optimización utilizando métodos numéricos deterministas, las propiedades analíticas de La función objetivo (convexidad, delimitación, suavidad) suele ser primordial. Para el enfoque de la simulación, estamos más interesados en h desde un método probabilístico (más que analítico) punto de vista. Obviamente, esta dicotomía es algo artificial, ya que existen enfoques de simulación donde la probabilidad no se utiliza la interpretación de h . No obstante, el uso de las propiedades analíticas de h juega un papel menor en el enfoque de simulación.

Ejemplo 5: Procesamiento de Señales:

Ó Ruanaidh y Fitzgerald (1996) estudiaron datos para el procesamiento de señales con un modelo simple ($i = 1, 2, \dots, N$)

$$x_i = \alpha_1 \cos(\omega t_i) + \alpha_2 \sin(\omega t_i) + \varepsilon_i. \quad \varepsilon_i \sim N(0, \sigma^2)$$

Con parámetros desconocidos $\alpha = (\alpha_1, \alpha_2)$, ω , y σ y las observaciones de t_1, \dots, t_N La función de verosimilitud es de la forma

$$\sigma^{-N} \exp \left(-\frac{(x - G\alpha)^t (x - G\alpha)}{2\sigma^2} \right),$$

Con $X = (x_1, \dots, x_N)$ y

$$G = \begin{pmatrix} \cos(\omega t_1) & \sin(\omega t_1) \\ \cos(\omega t_N) & \sin(\omega t_N) \end{pmatrix}.$$

La prioridad $\pi(\alpha, \omega, \sigma) = \sigma^{-1}$ Distribución marginal

$$\pi(\omega|x) \propto (x^t x - x^t G (G^t G)^{-1} G^t x)^{\frac{2-N}{2}} (\det G^t G)^{-1/2}$$

Explícitamente en ω no es fácil de calcular Esta configuración también es ilustrativo de funciones con muchos modos, como lo muestra O Ruanaidh y Fitzgerald (1996).

Siguiendo a Geyer (1996), queremos distinguir entre dos enfoques a la optimización de Monte Carlo. El primero es un enfoque exploratorio, en el que el objetivo es optimizar la función h describiendo su rango completo. Las propiedades reales de la función juegan un papel menor aquí, con el Monte Carlo.

Aspectos mejor vinculados a la exploración de todo el espacio θ , aunque, por ejemplo, la pendiente de h se puede utilizar para acelerar la exploración. (Tal una técnica puede ser útil para describir funciones con múltiples modos, para ejemplo.) El segundo enfoque se basa en una aproximación *Likelihood* de la función objetivo h y es algo así como un paso preliminar a la optimización real. Aquí, el aspecto de Montecarlo explota la Likelihood propiedades de la función h para llegar a una aproximación aceptable y casi omitiendo el 0. Veremos que este enfoque

puede vincularse a métodos de datos faltantes, como el algoritmo EM. Observamos también que Geyer (1996) solo considera que el segundo enfoque es la "optimización de Monte Carlo". Obviamente, aunque estemos considerando estos dos enfoques diferentes por separado, pueden combinarse en un problema dado. De hecho, métodos como el algoritmo EM o el algoritmo de Robbins-Monro aproveche la aproximación de Monte Carlo para mejorar su técnica de optimización particular.

Exploración Estocástica

Una solución básica

Hay varios casos en los que el método de exploración es particularmente bueno y adecuado. Primero, si θ está acotado, lo que a veces puede lograrse mediante una reparación de parametrización, una primera aproximación a la resolución es simular desde una distribución uniforme en θ , $u_1, \dots, u_m \sim U\theta$ y usar la aproximación $h_m = \max(h(u_1), \dots, h(u_m))$. Este método converge (de m a ∞), pero puede ser muy lento ya que no tiene en cuenta ninguna característica específica de h . Distribuciones distintas del uniforme, que posiblemente puedan estar relacionadas con h , entonces puede hacerlo mejor. En particular, en configuraciones donde la función de Likelihood es extremadamente costosa de calcular, el número de evaluaciones de la función h eses mejor mantenerlo al mínimo.

Una primera maximización de Monte Carlo

En la función $h(x) = [\cos(50x) + \sin(20x)]^2$ la función se define en un intervalo acotado, intentamos la estrategia de simular $u_1, \dots, u_m \sim U(0,1)$ y la aproximación $h_m = \max(h(u_1), \dots, h(u_m))$. El resultado Ahí vemos que la búsqueda aleatoria ha hecho un buen trabajo al imitar la función. El máximo de Monte Carlo es 3.832, lo que concuerda perfectamente con el "verdadero" máximo, obtenido mediante una evaluación exhaustiva. por supuesto, este es un pequeño ejemplo, y como se mencionó anteriormente, este ingenuo El método puede ser costoso en muchas situaciones. Sin embargo, el ejemplo ilustra el hecho de que, en problemas de baja dimensión, si la evaluación de la función es rápida, esto el método es una opción razonable.

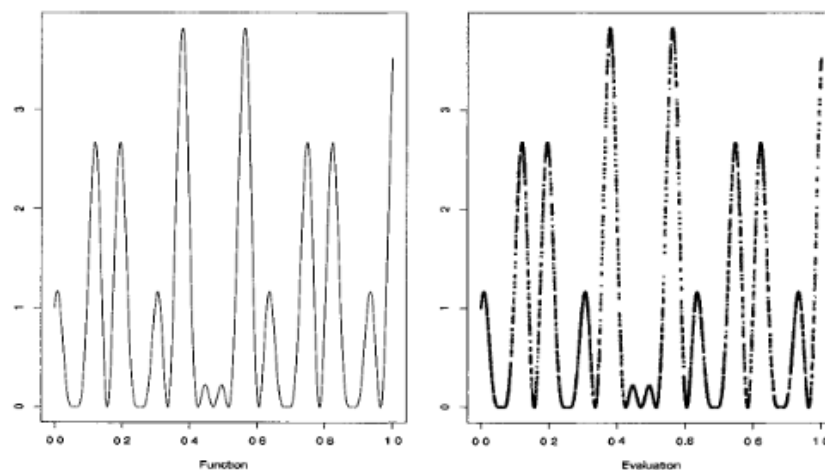


Figura 3: maximización de $h(x) = [\cos(50x) + \sin(20x)]^2$ a la izquierda es una grafica de la función, a la derecha es una simulación de 5000 variables aleatorias uniformes $[0,1]$.

Esto nos lleva a otra dirección en la que se relaciona a h con una distribución de probabilidad. Si h es positiva y:

$$\int_{\theta} h(\theta) d\theta < +\infty$$

la resolución de equivale a encontrar los modos de la densidad h . Más

En general, si no se cumplen estas condiciones, es posible que podamos transformar la función $h(\theta)$ en otra función $H(\theta)$ que satisfaga lo siguiente:

- (i) La función H no es negativa y satisface $\int H < \infty$.
- (ii) Las soluciones son aquellas que maximizan $H(\theta)$ en 0.

Un ejemplo, podemos tomar:

$$H(\theta) = \exp\left(\frac{h(\theta)}{T}\right) \text{ ó } H(\theta) = \exp\left\{\frac{h(\theta)}{T}\right\} / (1 + \exp\left\{\frac{h(\theta)}{T}\right\})$$

Al elegir T para acelerar la convergencia o para evitar los máximos locales (como en recocido simulado) Cuando el problema se expresa en términos estadísticos, resulta natural generar una muestra $(\theta_1, \dots, \theta_m)$ de h o H y aplicar un método de estimación en modo estándar (o simplemente compare los $h(\theta_i)$ en algunos casos ves más útil descomponer $h(\theta)$ en $h(\theta) = h_1(\theta)h_2(\theta)$ y simular de h_1

Ejemplo 6: Minimización de una función compleja

Se considera minimizar la función en función de \mathbb{R}^2

$$h(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(\sin(10x) x) + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y) y)$$

cuyo mínimo global es 0, alcanzado en $(x, y) = (0, 0)$. (Este es el hermano mayor de la función Dado que esta función tiene muchos mínimos locales, no satisface las condiciones bajo las cuales los métodos de minimización están garantizados para proporcionar el mínimo global. Sobre el. ¿Por otro lado, la distribución en M ? con densidad proporcional $\exp(-h(x, y))$ puede ser simulada, aunque esta no es una distribución estándar, utilizando, las técnicas de Monte Carlo en cadena de Markov, y una aproximación del mínimo de $h(x, y)$ y puede ser derivado del mínimo del resultado $h(x_i, y_i)$ y simular a partir de la densidad proporcional de

$$h_1(x, y) = \exp\{-(x \sin(20y) + y \sin(20x))^2 - (x \cos(10y) - y \sin(10x))^2\}$$

Elimina el cálculo de \cosh y \sinh en la simulación paso.

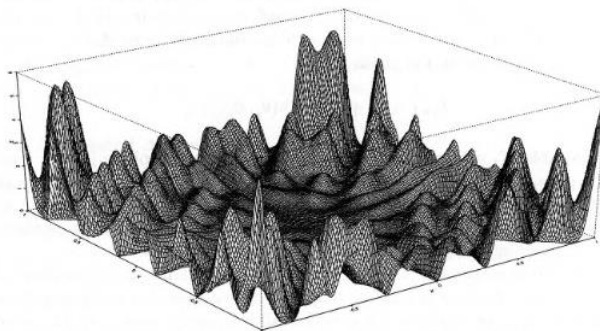


Figura 4: Grafica de la función $h(x, y)$ del ejemplo 6 en $[-1, 1]^2$

En particular si $H(x, \theta)$ es una densidad y si es posible simular a partir de la densidad, como solución es la distribución marginal de θ , se verán varios métodos para encontrar máximos que se puedan clasificar como métodos exploratorios.

Método del Gradiente

Es un método numérico determinista numérico, aprovechando el problema (5.1) que produce una secuencia (θ_j) que converge a la solución exacta de (5.1), θ cuando el dominio $\theta \in \mathbb{R}^2$ y la función $-h$ y ambos son convexos, la secuencia (θ_j) Es construido de manera recursiva de manera.

$$\theta_{j+1} = \theta_j + \alpha_j \nabla h(\theta_j), \quad \alpha_j > 0,$$

Donde ∇h es el gradiente de h . para varias opciones de la secuencia de (α_j) el algoritmo converge al máximo (único).

En configuraciones más generales (es decir, cuando la función o el espacio es menor regular), la ecuación puede modificarse mediante perturbaciones estocásticas para volver a lograr la convergencia, como se describe en detalle en Rubinstein (1981) o Duflo (1996) una de estas modificaciones estocásticas es escoger una segunda secuencia (β_j) para definir cadena (θ_j)

$$\theta_{j+1} = \theta_j + \frac{\alpha_j}{2\beta_j} \Delta h(\theta_j, \beta_j C_j) C_j$$

Las variables C_j se distribuyen uniformemente en la esfera unitaria $\|C\| = 1$ y $\Delta h(x, y) = h(x + y) - h(x - y)$ aproxima $2 \|y\| \nabla h(x)$ En contraste con el enfoque determinista, este método no procede necesariamente a lo largo del pendiente más pronunciado en (θ_j) pero esta propiedad a veces es una ventaja en el sentido de que puede evitar quedar atrapado en máximos locales o en puntos de h .

La convergencia de (θ_j) a la solución θ nuevamente depende de la elección de (α_j) y (β_j) observamos que (θ_j) puede verse como un no homogéneo.

Cadena de Márkov que casi con seguridad converge a un determinado valor. El estudio de estas cadenas es bastante complicado dado su siempre cambiante kernel de transición (ver Winkler 1995 para algunos resultados en esta dirección). Sin embargo, condiciones suficientemente fuertes como la disminución de α_j hacia 0 y de (α_j/β_j) a una constante distinta de cero son suficientes para garantizar la convergencia de la secuencia (θ_j) .

Reconocido Simulado

El algoritmo de recocido simulado fue introducido por Metrópolis et al. (1953) para minimizar una función de criterio en un conjunto finito con un tamaño muy grande, pero también se aplica a la optimización en un conjunto continuo y a la simulación (ver Kirkpatrick y col. 1983, Ackley et al 1985 y Neal 1993, 1995).

La idea fundamental de los métodos de recocido simulados es que un cambio de escala, llamada temperatura permite movimientos más rápidos en la superficie de la función h para maximizar, cuyo negativo se llama energía. Por lo tanto, rasarla evita parcialmente la atracción de trampas

de los máximos locales. Dada una temperatura parámetro $T > 0$, se genera una muestra $\theta_1^T, \theta_2^T, \dots$ a partir de la distribución: (primer expresión de la pg. Siguiende)

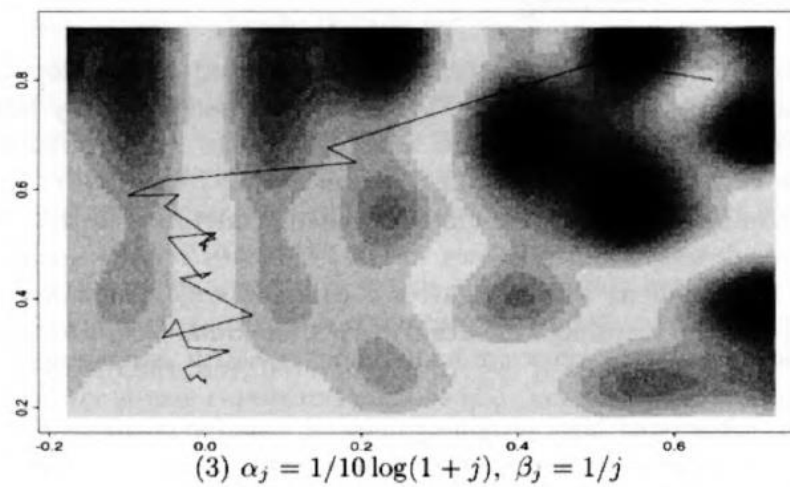
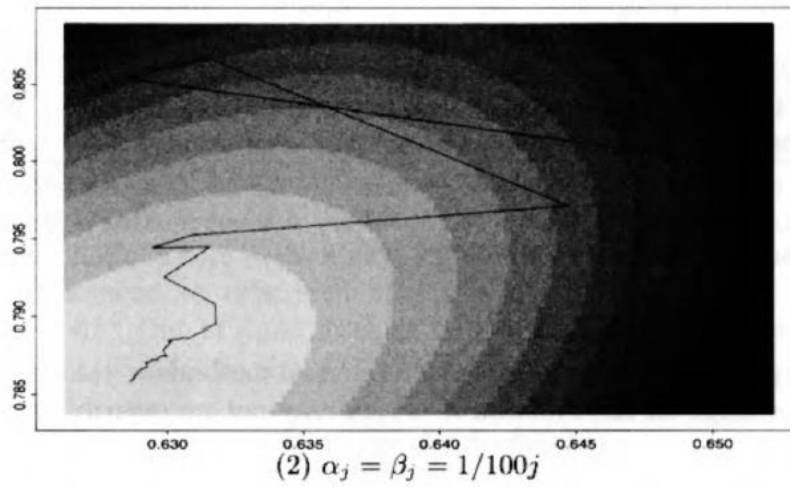
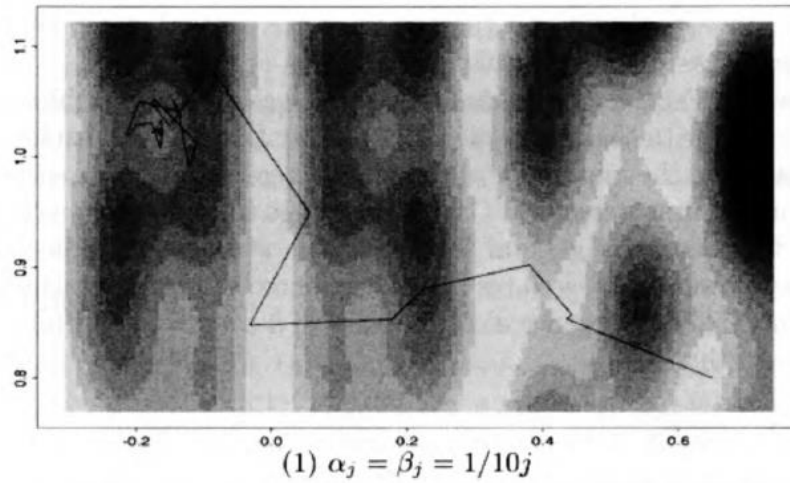


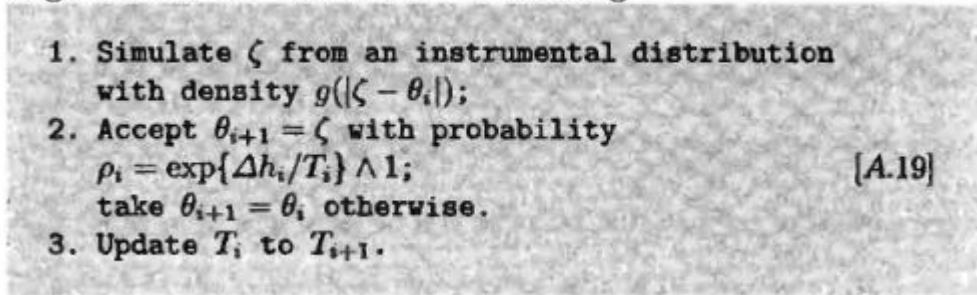
Figura 5: El método del gradiente estocástico parte de 3 opciones diferentes (α_i) y (β_j) empiezan en el punto $(.65, .8)$ para la misma secuencia (C_j) los niveles de gris son tales que los tonos más oscuros significan elevaciones más altas. La función h para ser minimizado.

$$\pi(\theta) \propto \exp\left(\frac{h(\theta)}{T}\right)$$

y se puede utilizar como en la Sección 5.2.1 para llegar a un máximo aproximado de h . A medida que T disminuye hacia 0, los valores simulados de esta distribución concentrarse en un vecindario cada vez más estrecho del local máximos de h (ver Teorema 5.10 y Winkler 1995). El hecho de que este enfoque tenga un efecto moderador sobre la atracción de los máximos locales de h se hace más evidentes cuando consideramos la simulación método propuesto por Metrópolis et al. (1953), empezando de θ_0 , se genera ζ de una distribución uniforme con $\nu = (\theta_0)$ of (θ_0) o más en general para la distribución de la densidad $g(|\zeta - \theta_0|)$

$$\begin{aligned}\theta_1 &= \left\{ \zeta \text{ con probabilidad } \rho = \exp\left(\frac{\Delta h}{T}\right) \wedge 1 \right\} \\ \theta_1 &= \{\theta_0 \text{ con probabilidad } 1 - \rho\}\end{aligned}$$

Algorithm A.19 –Simulated Annealing–



```

1. Simulate  $\zeta$  from an instrumental distribution
   with density  $g(|\zeta - \theta_i|)$ ;
2. Accept  $\theta_{i+1} = \zeta$  with probability
    $\rho_i = \exp\{\Delta h_i / T_i\} \wedge 1$ ;
   take  $\theta_{i+1} = \theta_i$  otherwise.
3. Update  $T_i$  to  $T_{i+1}$ .
[A.19]
```

Figura 6: Algoritmo de Reconocido Simulado

Teorema:

Considere un sistema en el que es posible vincular dos estados por una secuencia finita de estados. Si para cada $h > 0$ y cada par $((e_i, e_j))$ e_i puede llegar a una altitud h de (e_i) si y solo si se puede llegar a una altura de h de (e_i) y si (T_i) decrece a 0 la secuencia (θ_i) satisface

$$\lim_{i \rightarrow \infty} P(\theta_i \in \Omega) = 1$$

Si y solo si

$$\sum_{i=1}^{\infty} \exp\left(-\frac{D}{T_i}\right) = +\infty,$$

Este teorema, por lo tanto, da una condición necesaria y suficiente, en la tasa de disminución de la temperatura, de modo que el algoritmo de recocido simulado converge al conjunto de máximos globales. Esto sigue siendo un resultado ya que D es, en la práctica, desconocido. Por ejemplo, si $T_i = \text{gamma} / \log i$ hay convergencia a un máximo global si y solo si $F > D$. Numerosos artículos

y los libros han considerado la determinación práctica de la secuencia (T_n) (ver Geman y Geman 1984, Mitra et al. 1986, Van Laarhoven y Aarts 1987, Aarts y Kors 1989, Winkler 1995 y referencias allí). En lugar de la tasa logarítmica anterior, una tasa geométrica $T_i = \alpha^i T_0$ ($0 < \alpha < 1$) también es adoptada a menudo en la práctica, con la constante α calibrada al comienzo del algoritmo para que la tasa de aceptación sea lo suficientemente alta en el algoritmo Metrópolis.

Prior Feedback

Otro enfoque del problema de maximización (5.1) se basa en el resultado de Hwang (1980) de la convergencia (en T) de la llamada propuestas de Gibbs $\exp(\frac{h(\theta)}{T})$ a la distribución uniforme en el conjunto de máximos de h . Este enfoque, llamado integración recursiva o retroalimentación previa en Robert (1993) (ver también Robert y Soubiran 1993), se basa en el siguiente resultado de la convergencia.

Teorema

Considera h una función de valor real definida cerrada y acotada, $\theta \in \mathbb{R}^p$ si existe una única solución θ^* satisfice

$$\theta^* = \arg \max_{\theta \in \Theta} h(\theta),$$

Entonces

$$\lim_{\lambda \rightarrow \infty} \left(\frac{\int_{\Theta} \theta e^{\lambda h(\theta)} d\theta}{\int_{\Theta} e^{\lambda h(\theta)} d\theta} \right) = \theta^*$$

Dado que h es continua de θ^*

Corolario

π es una densidad positiva de θ , y si existe un único estimador de verosimilitud θ^* satisfice que:

$$\lim_{\lambda \rightarrow \infty} \frac{\int \theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}{\int e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta} = \theta^*$$

Este resultado utiliza la misma técnica que en el teorema anterior, a saber, la aproximación de Laplace de las integrales del numerador y del denominador. Expresa principalmente el hecho de que el máximo El estimador de verosimilitud se puede escribir como un límite de los estimadores de Bayes asociados con una distribución arbitraria π y con observaciones virtuales correspondientes a las

$\lambda \ell(\theta|x)$, $\exp\{\lambda \ell(\theta|x)\}$ cuando $\lambda \in \mathbb{N}$

$$\delta_{\lambda}^{\pi}(x) = \frac{\int \theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}{\int e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}$$

es simplemente el estimador de Bayes asociado con la distribución anterior π y una muestra correspondiente que consta de A réplicas de la muestra inicial x . La intuición detrás de estos

resultados es que a medida que el tamaño de la muestra aumenta infinito, la influencia de la distribución anterior se desvanece y la distribución asociada con $e^{\lambda \ell(\theta|x)} \pi(\theta)$ se concentra cada vez más al máximo global $\ell(\theta|x)$ cuando λ decrece.

Ejemplo 7 Estimación de la Gamma

Considerar la estimación de la distribución de la forma α de $\mathcal{G}(\alpha, \beta)$ con β conocido. Sin perder la generalidad se toma $\beta = 1$ para una constante (impropia) en α la distribución posterior satisface

$$\pi_{\lambda}(\alpha|x) \propto x^{\lambda(\alpha-1)} e^{-\lambda x} \Gamma(\alpha)^{-\lambda}$$

Para encontrar λ , el cálculo $\mathbb{E}[\alpha|x, \lambda]$ podemos obtener la simulación del algoritmo Metropolis-Hastings basado en la distribución $\exp(\frac{1}{\alpha(n-1)})$ donde $\alpha^{(n-1)}$ se denota el valor previo asociado cadena Markov.

El atractivo de la integración recursiva también es claro en el caso de restricciones de estimación de parámetros.

Aproximación estocástica

A continuación, pasamos a los métodos que trabajan más directamente con la función objetivo. En lugar de preocuparse por exploraciones rápidas del espacio. Informalmente hablando, estos métodos son algo preliminares a la verdadera optimización paso, en el sentido de que utilizan aproximaciones de la función objetivo h . Observamos que estas aproximaciones tienen un propósito diferente a las que previamente se observaron (por ejemplo, Laplace y saddlepoint aproximaciones, los métodos descritos aquí a veces puede resultar en un nivel adicional de error al mirar el máximo de una aproximación a h).

Dado que la mayoría de estos métodos de aproximación solo funcionan en los llamados modelos de datos faltantes, comenzamos esta sección con una breve introducción a estos modelos. Volvemos al supuesto de que la función objetivo h satisface $h(x) = \mathbb{E}[H(x, Z)]$ muestran que esta suposición surge en muchas configuraciones realistas. Además, tenga en cuenta que las extensiones artificiales que utilizan esta representación son sólo dispositivos computacionales y no invalida la inferencia general.

Modelos de datos faltantes

Los modelos de datos faltantes se consideran mejor como modelos en los que la likelihood. se puede expresar como

$$g(x|\theta) = \int_Z f(x, z|\theta) dz$$

Mas generalizado, donde la función $h(x)$ deberá ser optimizada para ser expresada

$$h(x) = \mathbb{E}[H(x, Z)]$$

Verosimilitud de datos censurados

Suponer que las observaciones Y_1, \dots, Y_n i.i.d. de $f(y - \theta)$ y debemos ordenar las observaciones entonces si $Y = (y_1, \dots, y_m)$ censurado y (y_{m+1}, \dots, y_n) Para la función que es:

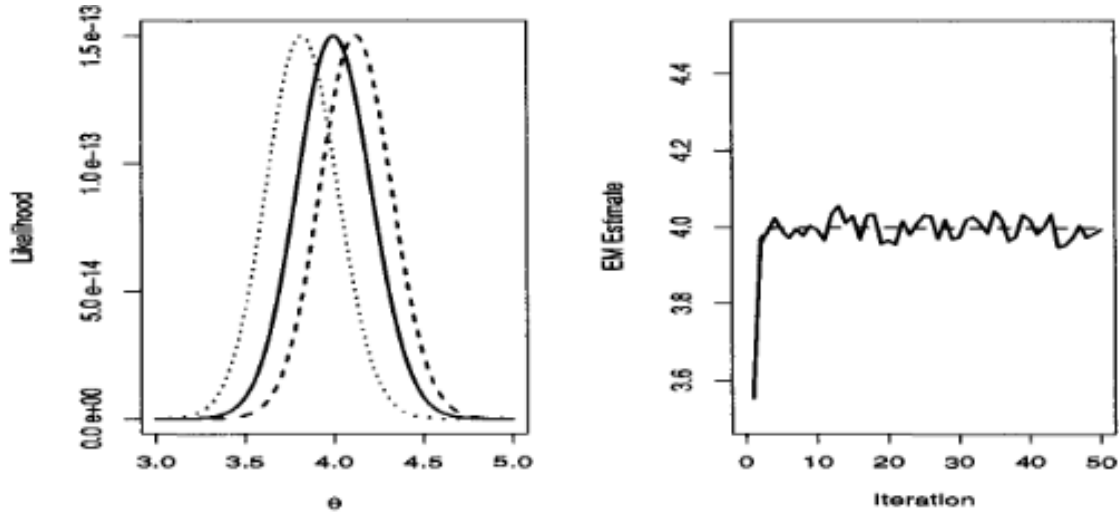


Figura 7: La grafica de la izquierda muestra 3 verosimilitudes de tamaño 25 $N(4,1)$ El más a la izquierda es la verosimilitud de la muestra donde los valores superiores a 4.5 se reemplazan por el valor 4.5 (punteado), el centro (sólido) son los datos observados probabilidad (5.9), y la más a la derecha (punteada) es la probabilidad usando los datos reales.

El panel de la derecha muestra estimaciones EM (discontinuas) y MCEM (sólidas).

$$L(\theta|y) = [1 - F(\alpha - \theta)]^{n-m} \prod_{i=m+1}^m f(y_i - \theta)$$

Donde F es la cdf asociada con f . si nosotros observamos los últimos $n-m$ valores, decimos $Z = (Z_{m+1}, \dots, Z_n)$ con $z_i > \alpha$ ($i = m+1, \dots, n$) deberamos construir la verosimilitud

$$L^c(\theta|y, z) = \prod_{i=1}^m f(y_i - \theta) \prod_{i=m+1}^n f(z_i - \theta)$$

Con el cual es fácil obtener

$$L(\theta|y) = \mathbb{E}[L^c(\theta|y, Z)] = \int_Z L^c(\theta|y, z) f(z|y, \theta) dz$$

Donde $f(z|y, \theta)$ se sostiene el vector Z implemente sirve para simplificar los cálculos, y la forma en que se selecciona Z para satisfacer, no debería afectar el valor del estimador. Este es un modelo de datos que faltan y nos referimos a la función $L^c(\theta|y, z) = f(x|z|\theta)$ como la likelihood de "modelo completo" o "datos completos", que corresponde a la observación de los datos completos (x, z) .

Referencias

- Pease, C. (2018, September 11). *An overview of Monte Carlo methods*. Retrieved October 15, 2021.
- Robert, C. P., & Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer.
- Robert, C., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. Springer.
- Shonkwiler, R. W., & Mendivil, F. (2009). Introduction to monte carlo methods. *Undergraduate Texts in Mathematics*, 1–49.