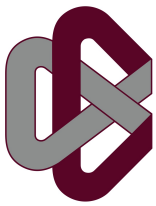

CIMAT

Centro de Investigación en Matemáticas, A.C.

Gibbs Direccional Óptimo: Aproximación Normal

Tesina

Que para obtener el Grado de:

Maestro en Ciencias con Especialidad en Probabilidad y Estadística

P R E S E N T A:

Mario Santana Cibrian

Director:

Dr. José Andrés Christen Gracia

Guanajuato, Gto., 5 de Agosto de 2011

Integrantes del Jurado

Presidente: Dr. Rogelio Ramos Quiroga (CIMAT)
Secretario: Dr. Marcos Aurelio Capistrán Ocampo (CIMAT)
Vocal y director de la tesina: Dr. José Andrés Christen Gracia (CIMAT)

Asesor:

Dr. José Andrés Christen Gracia

Sustentante:

Mario Santana Cibrian

Agradecimientos

A Roxana, por todo su amor y sus cuidados.

A mi madre, por todos los sacrificios que hizo para darme la oportunidad de estudiar.

A mi padre, por todas sus valiosas enseñanzas.

A David, porque siempre estuvo conmigo en los momentos más difíciles.

A Saturnino y Lupita, porque nunca dudaron de mí.

A mis amigos y compañeros, por todos sus consejos y su apoyo incondicional.

A mi asesor, el Dr. José Andrés Christen Gracia, por creer en mi y permitirme trabajar a su lado.

Al Dr. Rogelio Ramos Quiroga, porque siempre estuvo dispuesto a escucharme.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Centro de Investigación en Matemáticas (CIMAT), por haberme brindado el apoyo económico necesario para realizar mis estudios de posgrado.

Prefacio

Los métodos Markov Chain Monte Carlo (MCMC) son algoritmos que permiten obtener una muestra de una distribución de probabilidad f sin necesidad de simular directamente de ella. Para ello, estos métodos se basan en la construcción de una cadena de Markov ergódica cuya distribución estacionaria es precisamente f. Esta clase de algoritmos ha resultado ser de gran utilidad en diversas áreas, particularmente en la Estadística Bayesiana.

Christen y Fox (2011) exploran un criterio de optimalidad para el algoritmo MCMC Gibbs direccional. Dicho criterio consiste en minimizar la información mutua entre dos pasos consecutivos de la cadena de Markov generada por el algoritmo. Además, proponen una distribución de direcciones óptimas para el caso en el que la distribución objetivo es una distribución Normal multivariada.

Este trabajo retoma las ideas anteriores y las utiliza para crear un algoritmo Gibbs direccional óptimo que permite simular de distribuciones objetivo más generales. La principal característica del algoritmo es que utiliza una aproximación Normal local a la distribución de interés. Para ello es necesario que se pueda calcular de forma analítica el gradiente y el Hessiano de $-\log f(\mathbf{x})$, donde $f(\mathbf{x})$ es la función de densidad de la distribución objetivo.

Más aún, el trabajo propone distribuciones óptimas distintas a la que aparece en Christen y Fox (2011). Para evaluar el desempeño del algoritmo con cada una de las distribuciones de las direcciones, se utiliza una variante de la llamada distribución Normal sesgada.

Para exponer lo anterior, la tesina se dividió en cuatro capítulos. El Capítulo 1 presenta breves introducciones a los temas que sustentan el trabajo, como lo son la Estadística Bayesiana y los métodos Markov Chain Monte Carlo.

A lo largo del Capítulo 2 se desarrolla el algoritmo MCMC Gibbs direccional óptimo. En el Capítulo 3 se realizan los experimentos para evaluar el desempeño del algoritmo propuesto. Aquí se muestran los resultados obtenidos y las dificultades que aparecieron en la implementación computacional.

Por último, el Capítulo 4 presenta las conclusiones que se obtuvieron, mencionando los alcances del trabajo y posibles proyectos a futuro.

Índice general

Índice de cuadros

Índice de figuras

CAPÍTULO 1

Introducción

En este primer capítulo se expondrán de forma breve los temas que dan sustento al presente trabajo de tesina. Comenzaremos dando una introducción a la Estadística Bayesiana, enfocándonos principalmente en la metodología utilizada para hacer inferencia. Continuaremos con la descripción de los métodos Markov Chain Monte Carlo, los cuales son de gran utilidad precisamente en el área de inferencia Bayesiana.

Por último se presentará una breve introducción al área de problemas inversos. Aquí se explicará el por qué esta clase de problemas proporcionó parte de la motivación para el desarrollo de este trabajo.

1.1. Estadística Bayesiana

La Estadística Bayesiana es un enfoque muy particular de la Estadística. El nombre "Bayesiana" hace referencia a Thomas Bayes (1702 - 1761), un matemático británico que demostró un caso particular del teorema que ahora lleva su nombre. El teorema de Bayes es la base de la inferencia Bayesiana.

Es importante señalar dos de las características que distinguen el enfoque Bayesiano del resto de la Estadística:

- La primera característica está asociada a la parte filosófica. En Estadística Bayesiana se considera que todas las formas de incertidumbre se expresan en términos de una

medida de probabilidad. Más aún, se piensa que la probabilidad es una medida de lo que se sabe acerca de un evento. En este sentido, toda probabilidad depende de una serie de consideraciones y supuestos que envuelven al evento de interés. Por esta razón, en la literatura se suele decir que la probabilidad en la Estadística Bayesiana es *subjetiva*. Debido a la connotación negativa que puede tener esta palabra, optamos por referirnos a este tipo de probabilidad como *condicional*.

- La segunda característica tiene que ver con la parte operativa. En la metodología Bayesiana se considera que las probabilidades se modifican con evidencia. El teorema de Bayes es la herramienta que permite hacer posible esta idea.

  Teorema 1.1.1. *Teorema de Bayes: Sean A y B dos eventos cualesquiera tales que $P[B] > 0$. Entonces*

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B]}. \tag{1.1}$$

  $P[A]$ es llamada probabilidad a priori y representa el conocimiento inicial que se tiene del evento A; $P[B \mid A]$ es el modelo observacional e indica cómo sería la probabilidad del evento B si se conociera A. $P[A \mid B]$ es llamada probabilidad *a posteriori* (o posterior) y muestra cuál es la probabilidad de A después de haber observado B. $P[B]$ es sólo una constante de normalización que hace que $P[A \mid B]$ sea en efecto una probabilidad.

La posibilidad de incorporar el conocimiento previo a las probabilidades a través de la probabilidad a priori es sin duda el aspecto más llamativo y también el más criticado de este esquema.

Hasta ahora únicamente hemos dado un panorama general de lo que es la Estadística Bayesiana. Si el lector desea profundizar más en la parte filosófica y los fundamentos teóricos de este enfoque, recomendamos referirse a Bernardo y Smith (1994)[2], Berger (1985)[3] y Robert (2001)[14]. A continuación nos centraremos solamente en la parte operativa de la Inferencia Bayesiana.

1.1.1. Inferencia Bayesiana

A diferencia de la inferencia frecuentista, la inferencia Bayesiana supone que las cantidades desconocidas (parámetros) son variables aleatorias en vez de constantes, y que los datos, una vez observados, son fijos en vez de aleatorios. Por esta razón, la estimación en Bayesiana no consiste en encontrar estimadores puntuales de los parámetros de interés, sino en encontrar una distribución de probabilidad completa para dichos parámetros.

Suponga que se tiene una muestra $\mathbf{d} = (d_1, d_2, d_3, \ldots, d_m)$ de una variable aleatoria Y con distribución $f_{Y|\mathbf{X}}$, y se desea hacer inferencia sobre el vector de parámetros $\mathbf{X}$. La solución Bayesiana a este problema es la siguiente. Primero se propone una distribución a priori $f(\mathbf{x})$ para el vector de parámetros $\mathbf{X}$, la cual debe capturar el conocimiento que se tiene sobre estos parámetros. Después se establece un modelo observacional $f(\mathbf{d} \mid \mathbf{x})$ que representa cómo es la probabilidad de los datos suponiendo que conocemos $\mathbf{X}$. Este modelo no es más que la función de verosimilitud. Por último, se utiliza la regla de Bayes para resumir el conocimiento previo y la evidencia que se tiene sobre $\mathbf{X}$ en una distribución de probabilidad dada por

$$f(\mathbf{x} \mid \mathbf{d}) = \frac{f(\mathbf{d} \mid \mathbf{x}) f(\mathbf{x})}{\int f(\mathbf{d} \mid \mathbf{x}) f(\mathbf{x}) d\mathbf{x}}. \tag{1.2}$$

El término $\int f(\mathbf{d} \mid \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$ que aparece en la expresión (1.2) es sólo una constante de normalización que no depende de $\mathbf{x}$. La distribución $f(\mathbf{x} \mid \mathbf{d})$ recibe el nombre de *distribución posterior* y es el objeto de estudio de la inferencia Bayesiana.

Tener una distribución completa para $\mathbf{X}$ proporciona una gran ventaja sobre otros esquemas de inferencia, ya que no sólo provee una gran cantidad de información por si misma, sino que cuenta con todos los resultados de la teoría de probabilidad a su disposición. Además, algunos procedimientos estadísticos como las pruebas de hipótesis se vuelven sencillos e intuitivos con este esquema.

Si bien las ventajas de la inferencia Bayesiana son vastas, también existen dificultades importantes. Dos de las más comunes son: a) seleccionar la distribución a priori adecuada, y b) realizar los cálculos asociados a la distribución posterior.

Durante mucho tiempo el uso de la inferencia Bayesiana se vio limitado debido a que la distribución posterior llegaba a ser muy compleja y la constante de normalización era incalculable con técnicas convencionales. Sin embargo, a partir de 1990 hubo un incremento importante en la investigación y aplicaciones de los métodos Bayesianos, lo cual se debió principalmente al descubrimiento (por parte de los estadísticos) de los métodos Markov Chain Monte Carlo.

1.2. Markov Chain Monte Carlo

Los métodos Markov Chain Monte Carlo (MCMC) son algoritmos que permiten obtener una muestra de una distribución de probabilidad f sin necesidad de simular directamente de dicha distribución. Para ello estos métodos se basan en la construcción de una cadena de Markov ergódica cuya distribución estacionaria es precisamente f.

El principio general sobre el cual trabajan estos algoritmos es bastante sencillo. Dado un punto inicial arbitrario $x^{(0)}$, se construye una cadena de Markov ergódica $(X^{(t)})_{t\in\mathbb{N}}$ cuya distribución estacionaria es la distribución de interés f. Esto garantiza que, para un valor $l \in \mathbb{N}$ suficientemente grande, se cumple $X^{(l)}, X^{(l+1)}, X^{(l+2)}, \ldots \sim f$. Obviamente $X^{(l)}$ y $X^{(l+1)}$ no son independientes; sin embargo, para un cierto valor $m \in \mathbb{N}$ se puede considerar que $X^{(l)}$ y $X^{(l+m)}$ son aproximadamente independientes. Por lo tanto, si se simula de dicha cadena y se toma $Z^{(t)} = X^{(l+mt)}$, entonces se obtiene una muestra aproximadamente independiente de f.

Como la cadena es ergódica, desde el punto de vista teórico el valor inicial $x^{(0)}$ no tiene mayor relevancia. En la práctica cualquier punto inicial dentro del soporte de la distribución hará que la cadena converja; sin embargo, para ciertos valores iniciales el tiempo de convergencia podría crecer demasiado.

En la terminología de los métodos MCMC, l recibe el nombre de *long run* o *burn in* e indica el tiempo que se necesita para que la cadena converja a la distribución estacionaria. Por otro lado, m se suele llamar *thinning* y determina cada cuántas simulaciones se debe hacer el submuestreo. La calidad de las simulaciones depende fuertemente de estas dos cantidades. Contrario a lo que se esperaría, suele ser más complicado determi-

nar los valores de l y m que construir una cadena de Markov con las propiedades deseadas.

Una de las aplicaciones más comunes de estos algoritmos es el cálculo numérico de integrales multidimensionales. Por esta razón suelen ser de gran utilidad en áreas como la Física y la Estadística.

Algunos de los algoritmos MCMC más comunes son: Metropolis-Hastings, Gibbs sampler y Slice sampler. A continuación describiremos los primeros dos ya que son la base del algoritmo que se desarrolla en esta tesina.

1.2.1. Metropolis-Hastings

El algoritmo Metropolis-Hastings (M-H) es probablemente el método MCMC por excelencia. Fue nombrado en honor a los trabajos realizados por Nicholas Metropolis y W. Keith Hastings. Metropolis propuso por primera vez el algoritmo y lo usó para el caso específico de la distribución Boltzmann (ver Metropolis et al. 1953[12]). Posteriormente Hastings presentó una versión para casos más generales (ver Hastings, 1970[8]).

Para presentar el funcionamiento del algoritmo introduciremos un poco de notación. Sea π una distribución de probabilidad multivariada con soporte $\mathcal{X} \subseteq \mathbb{R}^n$ y con función de densidad $\pi(\mathbf{x})$. Supondremos que se tiene una expresión explícita para $\pi(\mathbf{x})$, salvo tal vez por una constante multiplicativa independiente de $\mathbf{x}$. En adelante π será la distribución de la cual deseamos simular y nos referiremos a ella como *distribución objetivo* o *distribución de interés*, que en el caso de la inferencia Bayesiana es la distribución posterior no normalizada, por ejemplo $\pi(\mathbf{x}) = f(\mathbf{x} \mid \mathbf{d}) \propto f(\mathbf{d} \mid \mathbf{x}) f(\mathbf{x})$.

Sea $q(\mathbf{y} \mid \mathbf{x})$ una densidad de probabilidad condicional tal que $q(\cdot \mid \mathbf{x})$ está bien definida para todo $\mathbf{x} \in \mathcal{X}$. Llamaremos a q *distribución instrumental* o *distribución propuesta*. En general se busca una distribución q de la cual es sencillo simular. Además, es necesario que $q(\mathbf{y} \mid \mathbf{x})$ tenga una expresión explícita o que sea simétrica, es decir, $q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{x} \mid \mathbf{y})$.

Entonces el algoritmo M-H funciona de la siguiente manera:

Dado $\mathbf{X}^{(t)} = \mathbf{x} \in \mathcal{X}$:

1. Se simula $\mathbf{y}$ de $q(\cdot \mid \mathbf{x})$.

2. Se *propone* que la cadena salte a $\mathbf{X}^{(t+1)} = \mathbf{y}$.

3. Se acepta la propuesta con probabilidad Metropolis-Hastings, dada por

$$\rho(\mathbf{x}, \mathbf{y}) = \text{mín}\left\{1, \frac{\pi(\mathbf{y})\, q(\mathbf{x} \mid \mathbf{y})}{\pi(\mathbf{x})\, q(\mathbf{y} \mid \mathbf{x})}\right\}. \tag{1.3}$$

   Es decir, se genera $u \sim U(0,1)$ y se toma

$$\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{y}, & \text{si } u \leq \rho(\mathbf{x}, \mathbf{y}), \\ \mathbf{x}, & \text{si } u > \rho(\mathbf{x}, \mathbf{y}). \end{cases} \tag{1.4}$$

Este algoritmo claramente genera una cadena de Markov ya que $\mathbf{X}^{(t+1)}$ sólo depende de $\mathbf{X}^{(t)}$. El kernel de transición de la cadena M-H está dado por:

$$K(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x}, \mathbf{y}) q(\mathbf{y} \mid \mathbf{x}) + \delta_{\mathbf{x}}(\mathbf{y})(1 - r(\mathbf{x})), \tag{1.5}$$

donde $r(\mathbf{x}) = \int \rho(\mathbf{x}, \mathbf{y}) q(\mathbf{y} \mid \mathbf{x}) d\mathbf{y}$, y $\delta_{\mathbf{x}}(\mathbf{y})$ es la función delta de Dirac.

Hay que notar que, para obtener muestras de π, no es necesario contar con una expresión completa de la densidad de nuestra distribución objetivo, basta con tener una función proporcional a dicha densidad. Esto se debe a que el algoritmo sólo requiere calcular cocientes de la forma

$$\frac{f(\mathbf{y})}{f(\mathbf{x})} \quad \text{y} \quad \frac{q(\mathbf{x} \mid \mathbf{y})}{q(\mathbf{y} \mid \mathbf{x})}. \tag{1.6}$$

Por esta razón el algoritmo M-H ha resultado de gran utilidad para la inferencia Bayesiana ya que, como se mencionó anteriormente, la constante de normalización de la densidad posterior puede ser muy difícil de calcular.

Teoremas de Convergencia

Una vez definido el algoritmo, es necesario establecer las condiciones bajo las cuales se cumple que π es la distribución estacionaria y la cadena es ergódica. Para llegar a esto primero considérense las siguientes definiciones.

Definición 1.2.1. *Una cadena de Markov estacionaria* $(\mathbf{X}^{(t)})$ *es reversible si la distribución de* $\mathbf{X}^{(t+1)} \mid \mathbf{X}^{(t+2)} = \mathbf{x}$ *es la misma que la distribución de* $\mathbf{X}^{(t+1)} \mid \mathbf{X}^{(t)} = \mathbf{x}$.

Definición 1.2.2. *Una cadena de Markov con kernel de transición K satisface la condición de balance detallado si existe una función f que cumple*

$$K(\mathbf{y}, \mathbf{x}) f(\mathbf{y}) = K(\mathbf{x}, \mathbf{y}) f(\mathbf{x}). \tag{1.7}$$

Luego

Teorema 1.2.1. *Suponga que una cadena de Markov con kernel de transición K satisface la condición de balance detallado con una función de densidad de probabilidad f. Entonces*

(a) La densidad f es una densidad invariante de la cadena.

(b) La cadena es reversible

Entonces el siguiente teorema nos proporciona el primer resultado que necesitamos.

Teorema 1.2.2. *Sea* $\left(X^{(t)}\right)$ *la cadena generada por el algoritmo M-H. Entonces para cualquier distribución condicional q cuyo soporte incluya a* $\mathcal{X}$*:*

(a) el kernel de la cadena satisface la condición de balance detallado con π*;*

(b) π *es una distribución estacionaria de la cadena.*

La demostración de a) se sigue de la definición de balance detallado. El inciso b) es consecuencia del Teorema 1.2.1.

Para establecer la ergodicidad es necesario que la cadena sea aperiódica y Harris recurrente. Esto garantizará que la distribución límite es π y que la cadena converge sin importar cuál sea el punto inicial $\mathbf{x}^{(0)}$. Definamos primero los conceptos anteriores.

Definición 1.2.3. *Dada una medida ψ, la cadena de Markov $\left(X^{(t)}\right)$ con kernel de transición $K(\mathbf{x}, \mathbf{y})$ es ψ-irreducible si, para cada $A \in \mathcal{B}(\mathcal{X})$ con $\psi(A) > 0$, existe n tal que $K^n(\mathbf{x}, A) > 0$ para todo $\mathbf{x} \in \mathcal{X}$.*

Definición 1.2.4. *Un conjunto A es Harris recurrente si $P_x(\eta_A = \infty) = 1$ para todo $\mathbf{x} \in A$. La cadena $\left(X^{(t)}\right)$ es Harris recurrente si existe una medida ψ tal que $\left(X^{(t)}\right)$ es ψ-irreducible y para cada conjunto A con $\psi(A) > 0$, A es Harris recurrente.*

Una condición suficiente para que la cadena sea aperiódica es que $K(\mathbf{x}, \mathbf{x}) > 0$ para todo $\mathbf{x} \in \mathcal{X}$. Esto se cumple a su vez siempre que

$$P\left[\pi(\mathbf{x})\, q(\mathbf{y} \mid \mathbf{x}) \leq \pi(\mathbf{y})\, q(\mathbf{x} \mid \mathbf{y})\right] < 1. \tag{1.8}$$

Consideremos ahora el siguiente teorema:

Teorema 1.2.3. *Si la cadena M-H $\left(\mathbf{X}^{(t)}\right)$ es π-irreducible, entonces es Harris recurrente.*

Se puede probar que, si $q(\mathbf{y} \mid \mathbf{x}) > 0$ para todo $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, entonces la cadena M-H es π-irreducible. En consecuencia del teorema anterior, la cadena también es Harris recurrente.

El lector puede encontrar los detalles de estos teoremas y sus demostraciones en Robert y Casella (2004)[15], así como una introducción completa a cadenas de Markov en Meyn y Tweedie (1993)[13].

1.2.2. Gibbs Sampler

El algoritmo Gibbs sampler puede verse como un caso particular del algoritmo M-H. A pesar de esto, sus fundamentos metodológicos y motivación histórica son completamente distintos.

Supongamos nuevamente que π es nuestra distribución objetivo. Antes de mostrar el algoritmo Gibbs sampler introduciremos un poco más de notación. Supongamos que se tiene un vector aleatorio n-dimensional $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ con soporte $\mathcal{X} \subseteq \mathbb{R}^n$ y función de densidad $\pi(\mathbf{x})$. Sea $\mathbf{X}_{-i} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ el vector de dimensión $n-1$ que se obtiene al eliminar la i-ésima componente del vector $\mathbf{X}$. Entonces la función

$$\pi_{X_i|\mathbf{X}_{-i}}\left(x_i \mid \mathbf{x}_{-i}\right) \propto \pi(\mathbf{x}), \tag{1.9}$$

representa la función de densidad condicional de la variable X_i dado el vector $\mathbf{X}_{-i}$, para $i = 1, 2, \ldots, n$. Estas funciones de densidad univariadas se mueven sobre un único eje de la base elegida para representar a $\mathbf{X}$ y reciben el nombre de densidades condicionales totales (full conditionals).

Entonces el algoritmo Gibbs sampler genera las transiciones de $\mathbf{X}^{(t)}$ a $\mathbf{X}^{(t+1)}$ como sigue:

Dado $\mathbf{x}^{(t)} = \left(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}\right)$, se generan:

1. $x_1^{(t+1)} \sim \pi_{X_1|\mathbf{X}_{-1}}\left(x_1 \mid x_2^{(t)}, \ldots, x_n^{(t)}\right)$;

2. $x_2^{(t+1)} \sim \pi_{X_2|\mathbf{X}_{-2}}\left(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_n^{(t)}\right)$;

   $\vdots$

n. $x_n^{(t+1)} \sim \pi_{X_n|\mathbf{X}_{-n}}\left(x_n \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{n-1}^{(t+1)}\right)$.

Una de las ventajas más importantes de este algoritmo es que todas las simulaciones son aceptadas, y en cada transición se obtiene un punto diferente de la cadena. Esto se debe a que la probabilidad de aceptación M-H es 1 en todo momento. Además, las simulaciones sólo se realizan a través de las densidades condicionales totales. El hecho de que éstas sean densidades unidimensionales representa una ventaja computacional.

Por otro lado, para poder implementar el algoritmo es necesario que se pueda simular de forma sencilla de cada una de las densidades condicionales totales. En muchos casos esto no es posible ni siquiera para una sola de estas densidades, lo cual limita la aplicación de este método.

1.2.3. Random Scan Gibbs Sampler

Una variante del algoritmo anterior es el llamado random scan Gibbs sampler, el cual no simula de forma sistemática de las densidades condicionales, sino que modifica una componente del vector $\mathbf{X}^{(t)}$ escogida de forma aleatoria.

Supongamos que $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ es un vector n-dimensional tal que $0 < \alpha_i < 1$ para todo i, y $\sum_i \alpha_i = 1$. Entonces el algoritmo funciona de la siguiente manera:

Dado $\mathbf{x}^{(t)} = \left(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}\right)$:

1. Se elige $i \in 1, 2, \ldots, n$ con probabilidad α_i.
2. Se genera $x_i^{(t+1)} \sim \pi_{X_i|\mathbf{X}_{-i}}\left(x_i \mid \mathbf{x}_{-i}^{(t)}\right)$.
3. Se toma $\mathbf{X}^{(t+1)} = \left(x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_i^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)}\right)$.

Este algoritmo mantiene casi todas las ventajas y desventajas del Gibbs sampler original. A pesar de esto es preferible usar random scan ya que la correlación entre las simulaciones disminuye.

1.2.4. Gibbs Direccional

Si bien el algoritmo Gibbs direccional es una generalización del Gibbs sampler, continúa siendo un caso particular del algoritmo M-H. La idea es escoger una dirección arbitraria $\mathbf{e} \in \mathbb{R}^n$ tal que $\|\mathbf{e}\| = 1$, y muestrear de la distribución condicional total de esa dirección. Esto puede ser descrito como

$$\mathbf{X}^{(t+1)} = \mathbf{x}^{(t)} + r\mathbf{e}, \tag{1.10}$$

donde $\mathbf{e}$ indica la dirección y r la magnitud de la transición. El algoritmo entonces trabaja como sigue:

Dado $\mathbf{x}^{(t)}$:

1. Se genera $\mathbf{e} \sim h\left(\mathbf{e} \mid \mathbf{x}^{(t)}\right)$.

2. Se genera $r \sim g\left(r \mid \mathbf{e}, \mathbf{x}^{(t)}\right)$.

3. Se propone $\mathbf{X}^{(t+1)} = \mathbf{x}^{(t)} + r\mathbf{e}$.

Es importante señalar que, para este caso particular, la densidad propuesta puede escribirse como

$$q(\mathbf{x} + r\mathbf{e} \mid \mathbf{x}) = g(r \mid \mathbf{e}, \mathbf{x})h(\mathbf{e} \mid \mathbf{x}). \tag{1.11}$$

Es sencillo demostrar que la distribución condicional de $r \mid \mathbf{e}, \mathbf{x}$ debe ser proporcional a $\pi\left(\mathbf{x} + r\mathbf{e}\right)$, es decir

$$g(r \mid \mathbf{e}, \mathbf{x}) \propto \pi\left(\mathbf{x} + r\mathbf{e}\right). \tag{1.12}$$

Si la distribución de $\mathbf{e}$ tiene como soporte la esfera S^n, entonces la cadena de Markov generada por el algoritmo anterior será ergódica y su distribución estacionaria será π. No obstante, el desempeño del algoritmo dependerá de la correlación entre $\mathbf{X}^{(t+1)}$ y $\mathbf{X}^{(t)}$.

La pregunta natural que surge es: ¿Cómo elegir una distribución para la dirección $\mathbf{e}$ que optimice (en algún sentido) el desempeño de este algoritmo? En el Capítulo 2 se presentará la respuesta que dan Christen y Fox (2011)[5] a esta pregunta.

1.3. Problemas Inversos

En esta sección mencionaremos algunas de las ideas principales que caracterizan a los problemas inversos. Lo único que se pretende aquí es argumentar cómo estos problemas proporcionaron parte de la motivación para desarrollar esta tesina. Si el lector desea adentrarse más en el área de problemas inversos, recomendamos revisar Vogel (2002)[18] y Calvetti (2007)[4].

Los problemas inversos aparecen constantemente en aplicaciones de las ciencias y de la industria, por lo que han recibido mucha atención de parte de los matemáticos aplicados, físicos, estadísticos e ingenieros. Este campo ha experimentado un crecimiento explosivo en las últimas décadas. Esto se debe a la importancia de las aplicaciones y, principalmente, al reciente desarrollo de computadoras potentes y de rápidos y confiables métodos numéricos.

En estas aplicaciones el objetivo es estimar algunos atributos desconocidos de interés mediante ciertas mediciones que están indirectamente relacionadas con dichos atributos. Para entender mejor esta idea, considere los siguientes ejemplos:

a) La exploración sísmica realiza mediciones de las vibraciones que ocurren en la superficie de la tierra. Éstas sólo están relacionadas de forma indirecta a las formaciones geológicas del subsuelo, las cuales son las que se desea determinar.

b) La tomografía médica computarizada tiene como objetivo producir imágenes de las estructuras dentro del cuerpo a partir de mediciones de rayos X que han pasado a través del cuerpo.

Los problemas inversos suelen involucrar modelos muy complejos como lo son las ecuaciones diferenciales ordinarias no lineales y ecuaciones diferenciales parciales, por mencionar algunos. En estos modelos una pequeña cantidad de ruido en los datos puede llevar a enormes errores en las estimaciones. Debido a ésto, los matemáticos y físicos consideran que este proceso de estimación está "mal planteado".

Para lidiar con estas dificultades se desarrollaron técnicas matemáticas conocidas como *regularización*. Si bien son útiles, atacan los problemas en un marco determinista con poca o ninguna atención a la incertidumbre inherente a los mecanismos de modelación y de medición de los datos. Este podría no ser el mejor enfoque ya que las mediciones de los datos son inexactas por su naturaleza. Por esta razón, los modelos estadísticos son una alternativa ya que proporcionan un método eficaz y riguroso para hacer frente a errores de medición.

La teoría estadística de inversión reformula los problemas inversos como problemas de inferencia por medio de la estadística Bayesiana. Desde esta perspectiva, la solución a un problema inverso es la distribución de probabilidad posterior de la cantidad de interés, la

cual se obtiene después de incorporar toda la información disponible al modelo.

En muchos casos, los problemas inversos tienen como solución una distribución posterior f muy compleja por lo que se requiere el uso de métodos MCMC para su análisis. Diseñar un algoritmo MCMC que funcioné bien para esta clase de distribuciones no es una tarea sencilla, sobre todo si la dimensionalidad de f es alta. No obstante, a pesar de la complejidad de f, en algunos casos se puede obtener de forma explicita el gradiente y el Hessiano de $-logf$. Es aquí donde surgió la idea de crear un algoritmo MCMC que pueda simular de este tipo de distribución posterior utilizando "únicamente" la información del gradiente y el Hessiano. Un algoritmo con tales características sería de gran utilidad ya que no se requeriría saber nada más de la distribución f.

En el siguiente capítulo se propondrá un algoritmo que tiene las características antes mencionadas y que creemos será de utilidad en las aplicaciones relacionadas a problemas inversos.

CAPÍTULO 2

Gibbs Direccional Óptimo

En el capítulo anterior se planteó la pregunta: ¿Cómo escoger la distribución de las direcciones $\mathbf{e}$ de forma que se optimize (en algún sentido) el desempeño de algoritmo Gibbs direccional? Para responder a esta pregunta, es necesario primero definir un criterio de "optimalidad". Una vez hecho esto, entonces se debe seleccionar una distribución de probabilidad h para $\mathbf{e}$ que satisfaga dicho criterio. Aquí es importante recalcar que el algoritmo será *óptimo* sólo en el sentido de ese criterio particular.

En este capítulo se mostrará el criterio de optimalidad que Christen y Fox (2011)[5] proponen en su reporte *Optimal Direction Gibbs for Sampling from Very High Dimension Normal Distributions*. Además se muestra cuál es una posible distribución óptima para las direcciones $\mathbf{e}$, para el caso en el que la distribución objetivo es una distribución Normal multivariada.

Comenzaremos recordando que el algoritmo Gibbs direccional genera una transición de $\mathbf{X}^{(t)} = \mathbf{x}^{(t)}$ a $\mathbf{X}^{(t+1)} = \mathbf{x}^{(t)} + r\mathbf{e}$, donde r es el tamaño del salto y $\mathbf{e}$ la dirección del mismo. Ya se estableció en al capítulo anterior que el desempeño de este algoritmo depende de que tan correlacionados estén $\mathbf{X}^{(t)}$ y $\mathbf{X}^{(t+1)}$. Entonces una aproximación natural al problema de optimización es buscar una medida de dependencia entre dos variables aleatorias y después minimizarla.

Christen y Fox (2011)[5] proponen como medida de dependencia la *información mutua*[1] entre las variables aleatorias $\mathbf{X}$ y $\mathbf{Y}$, la cual mide la divergencia de Kullback-Leibler entre el modelo conjunto $f_{\mathbf{X},\mathbf{Y}}$ y el modelo independiente $f_{\mathbf{X}} f_{\mathbf{Y}}$. Para calcular dicha información se utiliza la siguiente expresión:

$$I(\mathbf{Y},\mathbf{X}) = \int\int f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x}) \log \frac{f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y}) f_{\mathbf{X}}(\mathbf{x})} d\mathbf{x} d\mathbf{y}. \tag{2.1}$$

Nos interesa calcular $I\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right)$. En adelante supondremos que $\mathbf{X}^{(t+1)} = \mathbf{Y}$ y $\mathbf{X}^{(t)} = \mathbf{X}$, esto es sólo para simplificar la notación.

Al igual que antes nuestra distribución objetivo es π. Si suponemos que $\mathbf{X} \sim \pi$, entonces la fórmula de densidad condicional nos permite ver que

$$\begin{aligned} f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x}) &= f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \\ &= K(\mathbf{x},\mathbf{y}) \pi(\mathbf{x}). \end{aligned} \tag{2.2}$$

Sustituyendo el resultado de la ecuación (2.2) en la ecuación (2.1) obtenemos

$$\begin{aligned} I(\mathbf{Y},\mathbf{X}) &= \int\int \pi(\mathbf{x}) K(\mathbf{x},\mathbf{y}) \log \frac{\pi(\mathbf{x}) K(\mathbf{x},\mathbf{y})}{\pi(\mathbf{x}) \pi(\mathbf{y})} d\mathbf{x} d\mathbf{y} \\ &= \int\int \pi(\mathbf{x}) K(\mathbf{x},\mathbf{y}) \log \frac{K(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})} d\mathbf{x} d\mathbf{y}. \end{aligned} \tag{2.3}$$

Lo que se desea ahora es escoger direcciones $\mathbf{e}$ para las cuales $I(\mathbf{Y},\mathbf{X})$ sea pequeña. Como I es una divergencia Kullback-Liebler, entonces I está bien definida ya que $I \geq 0$ e $I = 0$ si y sólo si $\mathbf{Y}$ y $\mathbf{X}$ son independientes, es decir $f_{\mathbf{X},\mathbf{Y}} = f_{\mathbf{X}} f_{\mathbf{Y}}$ (sin embargo, I no es una métrica ya que no es simétrica). Entonces la idea de minimizar I en términos de $\mathbf{e}$ tiene sentido.

[1] Ver Cover y Thomas (1991)[6].

Dedicaremos la siguiente sección al estudio del caso en el que π es una distribución Normal multivariada. En este caso particular es posible obtener una expresión analítica para $I(\mathbf{Y}, \mathbf{X})$.

2.1. Caso Normal

Supongamos que la distribución objetivo π es una distribución Normal n-variada con matriz de precisión[2] $\mathbf{A}_{n\times n}$ y vector de medias $\boldsymbol{\mu}$. Si $\mathbf{X} \sim \pi$ entonces su función de densidad está dada por

$$\pi(\mathbf{x}) = \left(\frac{|\mathbf{A}|}{(2\pi)^n}\right)^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{2.4}$$

Como se explicó anteriormente, dada la dirección $\mathbf{e} \in \mathbb{R}^n$, $\|\mathbf{e}\| = 1$, la cadena se mueve de $\mathbf{X}^{(t)} = \mathbf{x}$ a $\mathbf{X}^{(t+1)} = \mathbf{x}^{(t)} + r\mathbf{e}$, donde la longitud $r \in \mathbb{R}$ tiene distribución g proporcional a $\pi\left(\mathbf{x}^{(t)} + r\mathbf{e}\right)$. De esto se sigue que

$$\begin{aligned} g(r \mid \mathbf{e}, \mathbf{x}) &\propto \exp\left\{-\frac{1}{2}(\mathbf{x}+r\mathbf{e}-\boldsymbol{\mu})^T \mathbf{A}(\mathbf{x}+r\mathbf{e}-\boldsymbol{\mu})\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{v}+r\mathbf{e})^T \mathbf{A}(\mathbf{v}+r\mathbf{e})\right\}, \end{aligned} \tag{2.5}$$

donde $\mathbf{v} = \mathbf{x} - \boldsymbol{\mu}$. Haciendo un poco de álgebra sobre la ecuación (2.5) se puede determinar la distribución de r. Así

[2]La matriz de precisión es la inversa de la matriz de varianzas y covarianzas. En Estadística Bayesiana se usa con mayor frecuencia ya que hace los cálculos más sencillos.

$$
\begin{aligned}
g(r \mid \mathbf{e}, \mathbf{x}) &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{v}+r\mathbf{e}\right)^T \mathbf{A}\left(\mathbf{v}+r\mathbf{e}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mathbf{v}^T+r\mathbf{e}^T\right)\mathbf{A}\left(\mathbf{v}+r\mathbf{e}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mathbf{v}^T\mathbf{A}+r\mathbf{e}^T\mathbf{A}\right)\left(\mathbf{v}+r\mathbf{e}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mathbf{v}^T\mathbf{A}\mathbf{v}+r\mathbf{e}^T\mathbf{A}\mathbf{v}+\mathbf{v}^T\mathbf{A}r\mathbf{e}+r\mathbf{e}^T\mathbf{A}r\mathbf{e}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mathbf{v}^T\mathbf{A}\mathbf{v}+2r\mathbf{e}^T\mathbf{A}\mathbf{v}+r^2\mathbf{e}^T\mathbf{A}\mathbf{e}\right)\right\} \\
&= \exp\left\{-\frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{2}\left(r^2+2r\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}+\frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)\right\} \\
&\propto \exp\left\{-\frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{2}\left[r^2-2r\left(-\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{2}\left[r^2-2r\left(-\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)+\left(-\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2\right]\right\} \\
&\propto \exp\left\{-\frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{2}\left[r-\left(-\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)\right]^2\right\}. \qquad (2.6)
\end{aligned}
$$

La ecuación (2.6) implica que r sigue una distribución Normal con media $-\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}$ y precisión $\mathbf{e}^T\mathbf{A}\mathbf{e}$.

Ahora hay que notar que, dada la dirección $\mathbf{e}$, el Kernel de transición correspondiente es

$$
\begin{aligned}
K_{\mathbf{e}}(\mathbf{x},\mathbf{y}) &= g\left(\mathbf{e}^T\left(\mathbf{y}-\mathbf{x}\right) \mid \mathbf{e}, \mathbf{x}\right) 1\left(\mathbf{y}=\mathbf{x}+\mathbf{e}^T\left(\mathbf{y}-\mathbf{x}\right)\mathbf{e}\right) \\
&= \left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{2\pi}\right)^{1/2} \exp\left\{-\frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{2}\left(\mathbf{e}^T\left(\mathbf{y}-\mathbf{x}\right)+\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2\right\} \\
&\quad \cdot 1\left(\mathbf{y}=\mathbf{x}+\mathbf{e}^T\left(\mathbf{y}-\mathbf{x}\right)\mathbf{e}\right). \qquad (2.7)
\end{aligned}
$$

Esto se debe a que, para ir de $\mathbf{x}$ a $\mathbf{y}$ sobre la dirección $\mathbf{e}$ es necesario que $r = \mathbf{e}^T\left(\mathbf{y}-\mathbf{x}\right)$, lo cual se sigue de $\mathbf{y}-\mathbf{x} = r\mathbf{e}$ y $\mathbf{e}^T\mathbf{e} = 1$. Así, condicionalmente a $\mathbf{e}$ y $\mathbf{x}$, $\mathbf{y}$ está restringida

a la linea $\mathbf{y} = \mathbf{x} + \mathbf{e}^T (\mathbf{y} - \mathbf{x}) \mathbf{e}$.

Con esto ya podemos calcular la información mutua entre $\mathbf{Y} = \mathbf{X}^{(t+1)}$ y $\mathbf{X} = \mathbf{X}^{(t)}$ dada la dirección $\mathbf{e}$. Utilizando la ecuación (2.3) se obtiene:

$$I(\mathbf{Y}, \mathbf{X}) = \int \int \pi(\mathbf{x}) K_{\mathbf{e}}(\mathbf{x}, \mathbf{y}) \log \frac{K_{\mathbf{e}}(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{y})} d\mathbf{x} d\mathbf{y}. \tag{2.8}$$

Para desarrollar la ecuación anterior trabajaremos primero con el término que involucra el logaritmo. Entonces

$$\begin{aligned}
\log \frac{K_{\mathbf{e}}(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{y})} &= \log K_{\mathbf{e}}(\mathbf{x}, \mathbf{y}) - \log \pi(\mathbf{y}) \\
&= log \left(\left(\frac{\mathbf{e}^T \mathbf{A} \mathbf{e}}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\mathbf{e}^T \mathbf{A} \mathbf{e}}{2} \left(\mathbf{e}^T (\mathbf{y} - \mathbf{x}) + \frac{\mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \right)^2 \right\} \right) \\
&\quad - \log \left(\left(\frac{|\mathbf{A}|}{(2\pi)^n} \right)^{1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{y} - \boldsymbol{\mu}) \right\} \right) \\
&= \frac{1}{2} \left[\log \left(\mathbf{e}^T \mathbf{A} \mathbf{e} \right) - log(2\pi) \right] - \frac{\mathbf{e}^T \mathbf{A} \mathbf{e}}{2} \left(\mathbf{e}^T (\mathbf{y} - \mathbf{x}) + \frac{\mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \right)^2 \\
&\quad - \frac{1}{2} \left[\log |\mathbf{A}| - n \log(2\pi) \right] + \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{y} - \boldsymbol{\mu}) \\
&= -\frac{1}{2} \log |\mathbf{A}| + \frac{n}{2} \log(2\pi) - \frac{1}{2} \log(2\pi) + \frac{1}{2} \log \left(\mathbf{e}^T \mathbf{A} \mathbf{e} \right) \\
&\quad - \frac{1}{2} \left[\mathbf{e}^T \mathbf{A} \mathbf{e} \left(\mathbf{e}^T (\mathbf{y} - \mathbf{x}) - \frac{\mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \right)^2 - (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{y} - \boldsymbol{\mu}) \right] \\
&= c + \frac{1}{2} log \left(\mathbf{e}^T \mathbf{A} \mathbf{e} \right) - \frac{1}{2} \left[Q_1(\mathbf{e}, \mathbf{x}, \mathbf{y}) - Q_2(\mathbf{y}) \right],
\end{aligned} \tag{2.9}$$

donde

$$\begin{aligned}
c &= \frac{n-1}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{A}|, \\
Q_1(\mathbf{e}, \mathbf{x}, \mathbf{y}) &= \mathbf{e}^T \mathbf{A} \mathbf{e} \left(\mathbf{e}^T (\mathbf{y} - \mathbf{x}) + \frac{\mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \right)^2, \\
Q_2(\mathbf{y}) &= (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{y} - \boldsymbol{\mu}).
\end{aligned}$$

Desarrollando la primera integral de la ecuación (2.8) obtenemos

$$\begin{aligned}\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})\log\frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})}d\mathbf{y} &= \int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})\left[c+\frac{1}{2}log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)-\frac{1}{2}\left[Q_1(\mathbf{e},\mathbf{x},\mathbf{y})-Q_2(\mathbf{y})\right]\right]d\mathbf{y}\\ &= c\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}+\frac{1}{2}log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}\\ &\quad-\frac{1}{2}\int Q_1(\mathbf{e},\mathbf{x},\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}+\frac{1}{2}\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}\\ &= c+\frac{1}{2}log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)-\frac{1}{2}\int Q_1(\mathbf{e},\mathbf{x},\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}\\ &\quad+\frac{1}{2}\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}, \qquad (2.10)\end{aligned}$$

lo cual se debe a que $\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}=1$ por definición de Kernel de transición.

Para simplificar la ecuación (2.10) desarrollamos la primera integral. Los cálculos son sencillos recordando que $r=\mathbf{e}^T(\mathbf{y}-\mathbf{x})$. Así

$$\begin{aligned}\int Q_1(\mathbf{e},\mathbf{x},\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y} &= \int \mathbf{e}^T\mathbf{A}\mathbf{e}\left(\mathbf{e}^T(\mathbf{y}-\mathbf{x})+\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2 K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}\\ &= \mathbf{e}^T\mathbf{A}\mathbf{e}\int\left(r+\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2 g(r\mid\mathbf{e},\mathbf{x})dr\\ &= \mathbf{e}^T\mathbf{A}\mathbf{e}\int\left[r^2+2r\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}+\left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2\right]g(r\mid\mathbf{e},\mathbf{x})dr\\ &= \mathbf{e}^T\mathbf{A}\mathbf{e}\left[\int r^2 g(r\mid\mathbf{e},\mathbf{x})dr+2\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\int r g(r\mid\mathbf{e},\mathbf{x})dr\right.\\ &\quad\left.+\left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2\int g(r\mid\mathbf{e},\mathbf{x})dr\right]\\ &= \mathbf{e}^T\mathbf{A}\mathbf{e}\left[\mathbb{E}\left[r^2\right]+2\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\mathbb{E}\left[r\right]+\left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2(1)\right]\\ &= \mathbf{e}^T\mathbf{A}\mathbf{e}\left[\frac{1}{\mathbf{e}^T\mathbf{A}\mathbf{e}}+\left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2-2\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)\right.\\ &\quad\left.+\left(\frac{\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\right)^2\right]\\ &= 1. \qquad (2.11)\end{aligned}$$

Es posible calcular la segunda integral de la ecuación (2.10) sustituyendo $\mathbf{y} = \mathbf{x} + r\mathbf{e}$ y $\mathbf{v} = \mathbf{x} - \boldsymbol{\mu}$. Así

$$\begin{aligned}
\int Q_2(\mathbf{y}) K_{\mathbf{e}}(\mathbf{x},\mathbf{y}) d\mathbf{y} &= \int (\mathbf{y}-\boldsymbol{\mu})^T \mathbf{A} (\mathbf{y}-\boldsymbol{\mu}) K_{\mathbf{e}}(\mathbf{x},\mathbf{y}) d\mathbf{y} \\
&= \int (\mathbf{x}+r\mathbf{e}-\boldsymbol{\mu})^T \mathbf{A} (\mathbf{x}+r\mathbf{e}-\boldsymbol{\mu}) K_{\mathbf{e}}(\mathbf{x},\mathbf{x}+r\mathbf{e}) dr \\
&= \int (r\mathbf{e}+\mathbf{v})^T \mathbf{A} (r\mathbf{e}+\mathbf{v}) g_{\mathbf{e}}(r) dr \\
&= \int \left(r^2 \mathbf{e}^T \mathbf{A} \mathbf{e} + 2r\mathbf{e}^T \mathbf{A} \mathbf{v} + \mathbf{v}^T \mathbf{A} \mathbf{v} \right) g_{\mathbf{e}}(r) dr \\
&= \int r^2 \mathbf{e}^T \mathbf{A} \mathbf{e} g_{\mathbf{e}}(r) dr + 2 \int r \mathbf{e}^T \mathbf{A} \mathbf{v} g_{\mathbf{e}}(r) dr + \int \mathbf{v}^T \mathbf{A} \mathbf{v} g_{\mathbf{e}}(r) dr \\
&= \mathbf{e}^T \mathbf{A} \mathbf{e} \mathbb{E}\left[r^2\right] + 2\mathbf{v}^T \mathbf{A} \mathbf{e} \mathbb{E}\left[r\right] + \mathbf{v}^T \mathbf{A} \mathbf{v} \\
&= \mathbf{e}^T \mathbf{A} \mathbf{e} \left[\left(\frac{\mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \right)^2 + \frac{1}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \right] - 2 \frac{\mathbf{v}^T \mathbf{A} \mathbf{e} \mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} + \mathbf{v}^T \mathbf{A} \mathbf{v} \\
&= \frac{\left(\mathbf{e}^T \mathbf{A} \mathbf{v}\right)^2}{\mathbf{e}^T \mathbf{A} \mathbf{e}} + 1 - 2 \frac{\left(\mathbf{e}^T \mathbf{A} \mathbf{v}\right)^2}{\mathbf{e}^T \mathbf{A} \mathbf{e}} + \mathbf{v}^T \mathbf{A} \mathbf{v} \\
&= 1 - \frac{\left(\mathbf{e}^T \mathbf{A} \mathbf{v}\right)^2}{\mathbf{e}^T \mathbf{A} \mathbf{e}} + \mathbf{v}^T \mathbf{A} \mathbf{v}. \qquad (2.12)
\end{aligned}$$

Por lo tanto, sustituyendo (2.11) y (2.12) en (2.10) se tiene

$$\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y}) \log \frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})} d\mathbf{y} = c + \frac{1}{2} log\left(\mathbf{e}^T \mathbf{A} \mathbf{e}\right) - \frac{1}{2} \frac{\mathbf{v}^T \mathbf{A} \mathbf{e} \mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} + \mathbf{v}^T \mathbf{A} \mathbf{v}. \quad (2.13)$$

Sustituimos la expresión anterior en (2.8) y calculamos la integral respecto a $\mathbf{x}$. Así

$$\begin{aligned}
I\left(\mathbf{Y},\mathbf{X}\right) &= \int \pi(\mathbf{x}) \left[c + \frac{1}{2} log\left(\mathbf{e}^T \mathbf{A} \mathbf{e}\right) - \frac{1}{2} \frac{\mathbf{v}^T \mathbf{A} \mathbf{e} \mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} + \mathbf{v}^T \mathbf{A} \mathbf{v} \right] d\mathbf{x} \\
&= c \int \pi(\mathbf{x}) d\mathbf{x} + \frac{1}{2} \log\left(\mathbf{e}^T \mathbf{A} \mathbf{e}\right) \int \pi(\mathbf{x}) d\mathbf{x} - \frac{1}{2} \int \frac{\mathbf{v}^T \mathbf{A} \mathbf{e} \mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \pi(\mathbf{x}) d\mathbf{x} \\
&\quad + \int \pi(\mathbf{x}) \mathbf{v}^T \mathbf{A} \mathbf{v} d\mathbf{x} \\
&= c + \frac{1}{2} \log\left(\mathbf{e}^T \mathbf{A} \mathbf{e}\right) - \frac{1}{2} \int \frac{\mathbf{v}^T \mathbf{A} \mathbf{e} \mathbf{e}^T \mathbf{A} \mathbf{v}}{\mathbf{e}^T \mathbf{A} \mathbf{e}} \pi(\mathbf{x}) d\mathbf{x} + \int \pi(\mathbf{x}) \mathbf{v}^T \mathbf{A} \mathbf{v} d\mathbf{x}. \qquad (2.14)
\end{aligned}$$

Para el cálculo de la primera integral de la ecuación anterior definimos $\mathbf{R} = \frac{\mathbf{A}\mathbf{e}\mathbf{e}^T\mathbf{A}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}$. Recordemos que la esperanza de una forma cuadrática está dada por

$$\mathbf{E}[\mathbf{z}\mathbf{R}\mathbf{z}] = tr(\mathbf{R}\Sigma_{\mathbf{Z}}) + \boldsymbol{\mu}_{\mathbf{Z}}^T\mathbf{R}\boldsymbol{\mu}_{\mathbf{Z}}, \tag{2.15}$$

donde $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}[\mathbf{Z}]$ y $\Sigma_{\mathbf{Z}} = Var[\mathbf{Z}]$. Además se puede mostrar que $\mathbf{E}[\mathbf{v}] = \mathbf{0}$ y $Var[\mathbf{v}] = \mathbf{A}^{-1}$. Luego

$$\begin{aligned}
\int \frac{\mathbf{v}^T\mathbf{A}\mathbf{e}\mathbf{e}^T\mathbf{A}\mathbf{v}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\pi(\mathbf{x})d\mathbf{x} &= \int \mathbf{v}^T\mathbf{R}\mathbf{v}\pi(\mathbf{x})d\mathbf{x} \\
&= \mathbb{E}\left[\mathbf{v}^T\mathbf{R}\mathbf{v}\right] \\
&= tr\left(\mathbf{R}\mathbf{A}^{-1}\right) + \mathbf{0}^T\mathbf{R}\mathbf{0} \\
&= tr\left(\frac{\mathbf{A}\mathbf{e}\mathbf{e}^T\mathbf{A}}{\mathbf{e}^T\mathbf{A}\mathbf{e}}\mathbf{A}^{-1}\right) \\
&= \frac{1}{\mathbf{e}^T\mathbf{A}\mathbf{e}} tr\left(\mathbf{A}\mathbf{e}\mathbf{e}^T\right) \\
&= \frac{1}{\mathbf{e}^T\mathbf{A}\mathbf{e}} tr\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right) \\
&= \frac{\mathbf{e}^T\mathbf{A}\mathbf{e}}{\mathbf{e}^T\mathbf{A}\mathbf{e}} \\
&= 1.
\end{aligned} \tag{2.16}$$

La segunda integral de la ecuación (2.14) se calcula como sigue:

$$\begin{aligned}
\int \mathbf{v}^T\mathbf{A}\mathbf{v}\pi(\mathbf{x})d\mathbf{x} &= \int (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\pi(\mathbf{x})d\mathbf{x} \\
&= \int \mathbf{x}^T\mathbf{A}\mathbf{x}\pi(\mathbf{x})d\mathbf{x} - 2\boldsymbol{\mu}^T\mathbf{A}\int \mathbf{x}\pi(\mathbf{x})d\mathbf{x} + \int \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\pi(\mathbf{x})d\mathbf{x} \\
&= \mathbb{E}\left[\mathbf{x}^T\mathbf{A}\mathbf{x}\right] - 2\boldsymbol{\mu}^T\mathbf{A}\mathbf{E}[\mathbf{X}] + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}(1) \\
&= tr(\mathbf{A}\mathbf{A}^{-1}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \\
&= tr\left(\mathbf{I}_{n\times n}\right) \\
&= n.
\end{aligned} \tag{2.17}$$

Sustituyendo (2.16) y (2.17) en (2.14) podemos concluir que la información mutua entre $\mathbf{X}^{(t+1)}$ y $\mathbf{X}^{(t)}$ es

$$I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right) = c + n - \frac{1}{2} + \frac{1}{2}\log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right) = C + \frac{1}{2}\log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right), \tag{2.18}$$

donde $C = c + n - \frac{1}{2}$. Es importante notar que $I_{\mathbf{e}}$ no depende de dónde se encuentra la cadena, es decir, no depende de $\mathbf{X}^{(t)}$.

2.1.1. Eligiendo un Conjunto de Direcciones

De acuerdo a la ecuación (2.18), la mejor dirección es aquella que minimiza $C + \frac{1}{2}\log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)$. Como este algoritmo es un caso particular del algoritmo M-H, entonces los teoremas de convergencia expuestos anteriormente son válidos. Sabemos que la cadena debe ser π-irreducible para que su distribución ergódica sea π. Por esta razón no podemos tomar únicamente la mejor dirección ya que la cadena sólo recorrería una línea en el espacio $\mathcal{X} \subseteq \mathbb{R}^n$.

Necesitamos entonces obtener una distribución completa para $\mathbf{e}$. Si las direcciones tienen función de densidad $h(\mathbf{e})$, y el soporte de h es la esfera S^n, entonces se garantiza que la cadena de Markov con kernel de transición $\int K_{\mathbf{e}}(\mathbf{x}, \mathbf{y})h(\mathbf{e})d\mathbf{e}$ es π-irreducible.

Kaufman y Smith (1998)[17] argumentaron que una distribución óptima para las direcciones es aquella cuya función de densidad satisface

$$h(\mathbf{e}) \propto \sup_{\mathbf{x}\in\mathcal{X}, r\in\mathbb{R}} \left\{ \int \pi(\mathbf{x} + \tau\mathbf{e})d\tau \frac{|r|^{n-1}}{\pi(\mathbf{x} + r\mathbf{e})} \right\}, \tag{2.19}$$

y optimiza la tasa de convergencia geométrica a la distribución estacionaria del Gibbs direccional resultante. Sin embargo, el resultado anterior sólo aplica para distribuciones con soporte $\mathcal{X}$ acotado. Para soporte no acotado no es posible controlar el término $\frac{|r|^{n-1}}{\pi(\mathbf{x}+r\mathbf{e})}$. Además de esto, muy poco se ha dicho en la literatura respecto a las direcciones óptimas del algoritmo Gibbs direccional.

Christen y Fox (2011)[5] argumentan que la distribución h^*, cuya densidad satisface

$$h^*(\mathbf{e}) \propto \left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)^{-1/2}, \tag{2.20}$$

permite minimizar la ecuación (2.18). Para ello hacen notar que minimizar $I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right)$ es equivalente a maximizar

$$\begin{aligned} \exp\left\{-I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right)\right\} &= \exp\left\{-C - \frac{1}{2}\log\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)\right\} \\ &= C_0\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)^{-1/2}, \end{aligned} \tag{2.21}$$

donde $C_0 = \exp\{-C\}$. Si se toman simulaciones de h^*, entonces se generan direcciones $\mathbf{e}$ con valores grandes de $\left(\mathbf{e}^T\mathbf{A}\mathbf{e}\right)^{-1/2}$, con lo cual se maximiza la ecuación (2.21) y, en consecuencia, se minimiza $I_{\mathbf{e}}$.

La elección de h^* resulta muy conveniente ya que es sencillo simular de ella. Si se genera $\mathbf{e}$ de una Normal multivariada centrada en el origen y con matriz de precisión $\mathbf{A}$, y se calcula $\mathbf{e}^* = \dfrac{\mathbf{e}}{\|\mathbf{e}\|}$, entonces $\mathbf{e}^* \sim h^*$.

Hay dos cosas que remarcar hasta el momento. La primera es que para simular las direcciones sólo se necesita conocer la matriz de precisión $\mathbf{A}$ y es posible prescindir de la media $\boldsymbol{\mu}$. Por otro lado, para simular de la Normal multivariada es necesario simular prácticamente de la misma Normal. Para este caso particular el resultado no es muy útil.

No obstante, este resultado nos permitirá trabajar con distribuciones objetivo más generales mediante el uso de una aproximación local Normal a la distribución objetivo. Desarrollaremos esta idea en la siguiente sección.

2.2. Aproximación Local Normal

Sea π la distribución objetivo con soporte $\mathcal{X} \subseteq \mathbb{R}^n$ como antes. Supongamos que $\mathbf{X} \in \mathbf{R}^n$ es un vector aleatorio con función de densidad $\pi(\mathbf{x})$. Supongamos además que para cada

$\mathbf{x} \in \mathcal{X}$ podemos calcular el gradiente $\nabla(\mathbf{x})$ y el Hessiano $\mathbf{H}(\mathbf{x})$ de $-\log \pi(\cdot)$. Si esto es cierto, entonces podemos obtener una aproximación Normal a $\pi(\mathbf{x})$ en cada punto de $\mathcal{X}$, es decir, una aproximación Normal local.

Para hacer esto consideramos la aproximación de Taylor de segundo orden de $-\log \pi(\mathbf{y})$

$$-\log \pi(\mathbf{y}) \approx -\log \pi(\mathbf{x}) + \nabla(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \mathbf{H}(\mathbf{x})(\mathbf{y}-\mathbf{x}). \tag{2.22}$$

Completando cuadrados en la expresión anterior obtenemos

$$\begin{aligned}
-\log \pi(\mathbf{y}) &\approx -\log \pi(\mathbf{x}) + \nabla(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}\left(\mathbf{y}^T - \mathbf{x}^T\right)\mathbf{H}(\mathbf{x})(\mathbf{y}-\mathbf{x}) \\
&= -\log \pi(\mathbf{x}) + \frac{2}{2}\nabla(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x}) - \mathbf{x}^T\mathbf{H}(\mathbf{x})\right](\mathbf{y}-\mathbf{x}) \\
&= -\log \pi(\mathbf{x}) + \frac{1}{2}\left[2\nabla(\mathbf{x})\mathbf{y} - 2\nabla(\mathbf{x})\mathbf{x}\right] \\
&\quad + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{x} + \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{x}\right] \\
&= -\log \pi(\mathbf{x}) + \frac{1}{2}\left[\nabla(\mathbf{x})\mathbf{y} + \nabla(\mathbf{x})\mathbf{y} - \nabla(\mathbf{x})\mathbf{x} - \nabla(\mathbf{x})\mathbf{x}\right] \\
&\quad + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{x} + \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{x}\right] \\
&= -\log \pi(\mathbf{x}) + \frac{1}{2}\left[\nabla(\mathbf{x})\mathbf{y} + \nabla(\mathbf{x})\mathbf{y} - \nabla(\mathbf{x})\mathbf{x} - \nabla(\mathbf{x})\mathbf{x}\right] \\
&\quad + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{x} + \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{x}\right] \\
&\quad + \frac{1}{2}\nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x}) - \frac{1}{2}\nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x}) \\
&= \left[-\log \pi(\mathbf{x}) - \frac{1}{2}\nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})\right] + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{y} + \nabla(\mathbf{x})\mathbf{y}\right] \\
&\quad - \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{x} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{x} + \nabla(\mathbf{x})\mathbf{x}\right] \\
&\quad + \frac{1}{2}\left[\mathbf{y}^T\nabla^T(\mathbf{x}) - \mathbf{x}^T\nabla^T(\mathbf{x}) + \nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})\right] \\
&= c(\mathbf{x}) + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{y} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{y} + \nabla(\mathbf{x})\mathbf{y}\right] \\
&\quad - \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{x} - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{x} + \nabla(\mathbf{x})\mathbf{x}\right] \\
&\quad + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})\mathbf{H}^{-1}(\mathbf{x})\nabla^T(\mathbf{x}) - \mathbf{x}^T\mathbf{H}(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x}) + \nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})\right] \\
&= c(\mathbf{x}) + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x}) - \mathbf{x}^T\mathbf{H}(\mathbf{x}) + \nabla(\mathbf{x})\right]\mathbf{y} \\
&\quad + \frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x}) - \mathbf{x}^T\mathbf{H}(\mathbf{x}) + \nabla(\mathbf{x})\right](-\mathbf{x})
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})-\mathbf{x}^T\mathbf{H}(\mathbf{x})+\nabla(\mathbf{x})\right]\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})\\
&=c(\mathbf{x})+\frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})-\mathbf{x}^T\mathbf{H}(\mathbf{x})+\nabla(\mathbf{x})\right]\left[\mathbf{y}-\mathbf{x}+\mathbf{H}^{-1}(\mathbf{x})\nabla^T(\mathbf{x})\right]\\
&=c(\mathbf{x})+\frac{1}{2}\left[\mathbf{y}^T\mathbf{H}(\mathbf{x})-\mathbf{x}^T\mathbf{H}(\mathbf{x})+\nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\mathbf{H}(\mathbf{x})\right]\left[\mathbf{y}-\mathbf{x}+\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})\right]\\
&=c(\mathbf{x})+\frac{1}{2}\left[\mathbf{y}^T-\mathbf{x}^T+\nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\right]\mathbf{H}(\mathbf{x})\left[\mathbf{y}-\mathbf{x}+\mathbf{H}^{-1}(\mathbf{x})\nabla^T(\mathbf{x})\right]\\
&=c(\mathbf{x})+\frac{1}{2}\left[\mathbf{y}-\mathbf{x}+\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})\right]^T\mathbf{H}(\mathbf{x})\left[\mathbf{y}-\mathbf{x}+\mathbf{H}^{-1}(\mathbf{x})\nabla^T(\mathbf{x})\right]\\
&=c(\mathbf{x})+\frac{1}{2}\left[\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\right]^T\mathbf{H}(\mathbf{x})\left[\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\right], \qquad (2.23)
\end{aligned}
$$

donde $c(\mathbf{x})=-\log\pi(\mathbf{x})-\frac{1}{2}\nabla(\mathbf{x})\mathbf{H}(\mathbf{x})\nabla^T(\mathbf{x})$ y $\boldsymbol{\mu}(\mathbf{x})=\mathbf{x}-\mathbf{H}^{-1}(\mathbf{x})\nabla^T(\mathbf{x})$. Luego

$$
\begin{aligned}
\pi(\mathbf{y}) &= \exp-[-\log\pi(\mathbf{y})]\\
&\approx \exp\left\{-c(\mathbf{x})-\frac{1}{2}\left[\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\right]^T\mathbf{H}(\mathbf{x})\left[\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\right]\right\}\\
&\approx \exp\left\{-\frac{1}{2}\left[\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\right]^T\mathbf{H}(\mathbf{x})\left[\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\right]\right\}. \qquad (2.24)
\end{aligned}
$$

Entonces tenemos una aproximación Normal local a π con vector de medias $\boldsymbol{\mu}(\mathbf{x})$ y matriz de precisión $\mathbf{H}(\mathbf{x})$.

Utilizando la ecuación (2.24) y los resultados de la sección anterior podemos proponer un algoritmo Gibbs direccional óptimo que nos permita simular de una distribución objetivo más general que la distribución Normal. Este algoritmo funciona como sigue:

Dado $\mathbf{X}^{(t)}=\mathbf{x}$:

1. Se genera $\mathbf{e}=\dfrac{\mathbf{e}^*}{\|\mathbf{e}^*\|}$, donde $\mathbf{e}^*\sim N(\mathbf{0},\mathbf{H}(\mathbf{x}))$.

2. Se genera $r\sim N(\mu,\tau)$, donde $\mu=-\dfrac{\mathbf{e}^T\nabla(\mathbf{x})}{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}$ y $\tau=\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}$.

3. Se propone $\mathbf{X}^{(t+1)}=\mathbf{x}+r\mathbf{e}$.

4. Se acepta la propuesta con probabilidad Metropolis-Hastings.

A continuación analizaremos la probabilidad de aceptación. Recordamos que el cociente M-H está dado por $R = \frac{\pi(\mathbf{y})q(\mathbf{x}|\mathbf{y})}{\pi(\mathbf{x})q(\mathbf{y}|\mathbf{x})}$. En nuestro contexto $\mathbf{y} = \mathbf{x} + r\mathbf{e}$,

$$q(\mathbf{x} \mid \mathbf{y}) = g(r \mid -\mathbf{e}, \mathbf{y})h^*(-\mathbf{e} \mid \mathbf{y}), \tag{2.25}$$

y

$$q(\mathbf{y} \mid \mathbf{x}) = g(r \mid \mathbf{e}, \mathbf{x})h^*(\mathbf{e} \mid \mathbf{x}). \tag{2.26}$$

Si hacemos $f(\mathbf{x}) = \log \pi(\mathbf{x})$, entonces

$$\log R = f(\mathbf{y}) + \log q(\mathbf{x} \mid \mathbf{y}) - f(\mathbf{x}) + \log q(\mathbf{y} \mid \mathbf{x}). \tag{2.27}$$

Definimos ahora

$$\begin{aligned} \psi_{\mathbf{x}}(\mathbf{e}, r) &= f(\mathbf{x}) + \log q(\mathbf{y} \mid \mathbf{x}) \\ &= f(\mathbf{x}) + \log q(\mathbf{x} + r\mathbf{e} \mid \mathbf{x}). \end{aligned} \tag{2.28}$$

Como $q(\mathbf{x} + r\mathbf{e} \mid \mathbf{x}) = g(r \mid \mathbf{x}, \mathbf{e})h^*(\mathbf{e})$, entonces

$$\begin{aligned} q(\mathbf{x} + r\mathbf{e} \mid \mathbf{x}) &= K_{\mathbf{H}(\mathbf{x})}(\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e})^{-\frac{1}{2}} \frac{(\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e})^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left\{-\frac{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}{2}\left(r + \frac{\mathbf{e}^T\nabla(\mathbf{x})}{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}\right)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}} K_{\mathbf{H}(\mathbf{x})} \exp\left\{-\frac{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}{2}\left(r + \frac{\mathbf{e}^T\nabla(\mathbf{x})}{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}\right)^2\right\}. \end{aligned} \tag{2.29}$$

Por lo tanto

$$\psi_{\mathbf{x}}(\mathbf{e}, r) = f(\mathbf{x}) + \log K_{\mathbf{H}(\mathbf{x})} - \frac{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}{2}\left(r + \frac{\mathbf{e}^T\nabla(\mathbf{x})}{\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}}\right)^2, \tag{2.30}$$

donde

$$K_{\mathbf{H}(\mathbf{x})}^{-1} = \int_{\|\mathbf{e}\|=1} \left(\mathbf{e}^T \mathbf{H}(\mathbf{x})\mathbf{e}\right)^{-\frac{1}{2}} d\mathbf{e} \tag{2.31}$$

es la constante de normalización de la función de densidad $h^*(\mathbf{x})$, la cual depende de $\mathbf{x}$. Podemos ver entonces que

$$R(\mathbf{x}, \mathbf{e}, r) = \exp\left\{\psi_{\mathbf{x}+r\mathbf{e}}(-\mathbf{e}, r) - \psi_{\mathbf{x}}(\mathbf{e}, r)\right\}. \tag{2.32}$$

Igual que antes la probabilidad de aceptar un salto de $\mathbf{x}$ a $\mathbf{y} = \mathbf{x}+r\mathbf{e}$ es mín$\{1, R(\mathbf{x}, \mathbf{e}, r)\}$. Se utilizó la expresión (2.27) para redefinir R y esto nos permite tener mayor estabilidad numérica al simular.

El algoritmo es bastante general ya que para simular de π sólo es necesario conocer el gradiente y el Hessiano de $-\log \pi(\mathbf{x})$ para cada punto $\mathbf{x} \in \mathcal{X}$. Debido a su construcción, se esperaría que el desempeño sea muy bueno.

Sin embargo, existe una problema importante con la implementación. La constante de normalización de la función de densidad h^*, definida en la ecuación (2.31), no se puede calcular de forma analítica. Esto hace que tengamos que recurrir a una aproximación numérica, lo cual sólo es factible para dimensiones pequeñas.

2.3. Otras Direcciones Óptimas

Debido a la problemática señalada anteriormente, consideraremos una distribución para las direcciones $\mathbf{e}$ distinta de la distribución propuesta por Christen y Fox (2011)[5].

Una idea simple es hacer que la cadena se mueva sólo en un número finito de direcciones. Para ello, haremos que las direcciones $\mathbf{e}$ sean los eigenvectores de la matriz $\mathbf{H}(\mathbf{x})$, así $\mathbf{e} \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. La i-ésima dirección se seleccionará con probabilidad proporcional a λ_i^{-1}, donde λ_i es el eigenvalor correspondiente al i-ésimo eigenvector , $i = 1, 2, \ldots, n$. Luego

$$h_1(\mathbf{e_i}) = (k\lambda_i)^{-1}, \tag{2.33}$$

donde $k = \sum_{i=1}^{n} \lambda_i^{-1}$. Haciendo esto la distribución de $\mathbf{e}$ se vuelve discreta. Además se elimina por completo el problema que teníamos con la constante de normalización de la distribución h^* y la dimensión de π ya no es un problema.

Podemos ver que la dirección $\mathbf{e}_n$ asociada al eigenvalor más pequeño λ_n de $\mathbf{H}(\mathbf{x})$ es óptima. Para ello hay que notar que, si deseamos minimizar $I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right)$, es decir, minimizar la ecuación (2.18), entonces

$$\begin{aligned} \min_{\|\mathbf{e}\|=1} I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right) &= \min_{\|\mathbf{e}\|=1} \left\{ C + \frac{1}{2}\log\left(\mathbf{e}^T A \mathbf{e}\right) \right\} \\ &= C + \frac{1}{2}\log\left(\min_{\|\mathbf{e}\|=1} \left\{ \mathbf{e}^T A \mathbf{e} \right\} \right) \\ &= C + \frac{1}{2}\log \lambda_n. \end{aligned} \tag{2.34}$$

Además, el mínimo se alcanza cuando $\mathbf{e}$ es el eigenvector asociado a λ_n, es decir, $\mathbf{e} = \mathbf{e}_n$.

Como mencionamos antes, no es posible tomar sólo la dirección que minimiza $I_{\mathbf{e}}$, por lo cual debemos tomar el resto de los eigenvectores para garantizar que la cadena sea π-irreducible.

En el Capítulo 3 realizaremos pruebas para evaluar nuestro algoritmo utilizando ambas distribuciones, h^* y h_1.

CAPÍTULO 3

Experimentos

Para probar el desempeño de nuestro algoritmo utilizaremos como distribución objetivo una variante de la llamada distribución Normal sesgada. Nuestra elección de esta distribución se debe a su sencillez y versatilidad, además del hecho de que nos permite calcular de forma analítica el gradiente y el Hessiano de $-\log \pi$.

A lo largo del capítulo probaremos el algoritmo con varios casos particulares de la distribución objetivo y señalaremos las ventajas y desventajas de nuestro enfoque.

3.1. Distribución Normal Sesgada Mediante una Distribución Logística

Antes de definir nuestra distribución objetivo es indispensable hablar de la distribución Normal sesgada (Skew-normal). El siguiente Lema nos dará las bases para definirla en el caso univariado.

Lema 3.1.1. *Sea f_0 una función de densidad de probabilidad unidimensional simétrica alrededor de 0, y sea G una función de distribución unidimensional tal que G' existe y es una función de densidad de probabilidad simétrica alrededor de 0. Entonces*

$$f(z) = 2f_0(z)G(w(z)), \tag{3.1}$$

para $-\infty < z < \infty$, es una función de densidad de probabilidad para cualquier función impar $w(\cdot)$.

Este resultado nos permite modificar una función de densidad simétrica "base" f_0 mediante una función de "perturbación" $G(w(x))$ y obtener una nueva función de densidad f. Las condiciones que se piden sobre G y w son mínimas y esto abre la posibilidad de crear una gran variedad de distribuciones con una misma base f_0. Es sencillo ver que el conjunto de densidades "perturbadas" incluye a la densidad f_0, y se obtiene cuando $w(x) \equiv 0$.

Un resultado muy útil asociado al Lema anterior es el siguiente.

Teorema 3.1.1. *Si $X \sim G'$ y $Y \sim f_0$ son variables independientes, entonces*

$$Z = \begin{cases} Y, & \text{si } X < w(Y), \\ -Y, & \text{en otro caso,} \end{cases} \tag{3.2}$$

sigue la función de densidad en (3.1).

La expresión anterior da un método muy sencillo para obtener simulaciones de la densidad (3.1).

Si utilizamos el Lema 3.1.1 con $f_0 = \phi$ y $G = \Phi$, la función de densidad y la función de distribución de una variable $N(0,1)$, respectivamente, y $w(x) = \alpha x$, $\alpha \in \mathbb{R}$, obtenemos la densidad

$$\phi(z;\alpha) = 2\phi(z)\Phi(\alpha z), \tag{3.3}$$

para $-\infty < z < \infty$, la cual es llamada densidad "Normal sesgada" con parámetro de forma α.

Para definir la versión multivariada ocuparemos la extensión del Lema 3.1.1 al caso multidimensional.

Lema 3.1.2. *Si f_0 es una función de densidad de probabilidad n-dimensional tal que $f_0(\mathbf{x}) = f_0(-\mathbf{x})$ para todo $\mathbf{x} \in \mathbb{R}^n$, G es una función de distribución unidimensional diferenciable tal que G' es una función de densidad simétrica alrededor de 0, y w es una función real tal que $w(-\mathbf{x}) = -w(\mathbf{x})$ para todo $\mathbf{x} \in \mathbb{R}^n$, entonces*

$$f(\mathbf{z}) = 2f_0(\mathbf{z})G(w(\mathbf{z})), \tag{3.4}$$

$z \in \mathbb{R}$, es una función de densidad sobre $\mathbb{R}^n$.

La condición $f_0(\mathbf{x}) = f_0(-\mathbf{x})$ es llamada "simetría central". El teorema anterior nos permite manipular distribuciones multivariadas de forma similar al caso univariado.

Un resultado análogo al que se presenta en el Teorema 3.1.1 nos permitirá simular de este tipo de distribuciones de forma simple.

Teorema 3.1.2. *Si $X \sim G'$ y $\mathbf{Y} \sim f_0$ son variables independientes, entonces*

$$\mathbf{Z} = \begin{cases} \mathbf{Y}, & si\ X < w(\mathbf{Y}), \\ -\mathbf{Y}, & en\ otro\ caso, \end{cases} \tag{3.5}$$

sigue la función de densidad en (3.4).

Este resultado es muy útil ya que podremos comparar las simulaciones obtenidas con nuestro algoritmo con una muestra independiente de la distribución objetivo.

Supongamos ahora que $f_0(\mathbf{x}) = \phi_n(\mathbf{x}; \mathbf{\Sigma})$ es la función de densidad de una variable $N_n(\mathbf{0}, \mathbf{\Sigma})$, $G = \Phi$ y que $w(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{x}$ es una función lineal. Si además permitimos la presencia de un parámetro n-dimensional de localización $\boldsymbol{\xi}$, entonces utilizando la ecuación (3.4) obtenemos la densidad

$$f(\mathbf{y}) = 2\phi_n(\mathbf{y} - \boldsymbol{\xi}; \mathbf{\Sigma})\Phi(\boldsymbol{\alpha}^T(\mathbf{y} - \boldsymbol{\xi})), \tag{3.6}$$

$\mathbf{y} \in \mathbb{R}^n$, donde $\boldsymbol{\alpha}$ es un parámetro de forma. Si una variable aleatoria n-dimensional $\mathbf{Y}$ tiene función de densidad (3.6), entonces se dice que $\mathbf{Y}$ sigue una distribución Normal sesgada n-variada.

Hasta aquí sólo hemos definido la distribución Normal sesgada. Si el lector desea conocer más acerca de las propiedades de esta distribución puede encontrar una revisión detallada en Azzalini (2005)[1].

Nuestra función objetivo será una variante de la distribución Normal sesgada ya que, en lugar de tomar la función de perturbación $G = \Phi$, tomaremos

$$G(x) = \frac{1}{1 + \exp\{-(x-\mu)/s\}}. \tag{3.7}$$

$G(x)$ es la función de distribución Logística con parámetro de localización μ y parámetro de escala s. Por esta razón decimos que nuestra distribución objetivo π es una distribución Normal sesgada mediante una distribución Logística.

Para implementar nuestro algoritmo, supondremos que $f_0(\mathbf{x})$ es la densidad de una variable $N_n(\mathbf{0}, \mathbf{A})$, donde $\mathbf{A}$ es la matriz de precisión, esto es

$$f_0(\mathbf{x}) = \left(\frac{|\mathbf{A}|}{(2\pi)^n}\right)^{1/2} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\right\}. \tag{3.8}$$

Supondremos además que el parámetro de localización de la distribución logística es 0 y que el parámetro de escala es $s = \frac{\sqrt{3}}{\pi}$, lo cual hace que la varianza de G sea 1; así

$$G(x) = \frac{1}{1 + \exp\left\{-\pi x/\sqrt{3}\right\}}. \tag{3.9}$$

Por último, consideraremos que el parámetro de localización $\boldsymbol{\xi}$ es el vector n-dimensional $\mathbf{0}$. Luego, utilizando la expresión (3.4) obtenemos

$$\begin{aligned}\pi(\mathbf{x}) &= 2 f_0(\mathbf{x}) G\left(\boldsymbol{\alpha}^T\mathbf{x}\right) \\ &= 2\left(\frac{|\mathbf{A}|}{(2\pi)^n}\right)^{1/2} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\right\} \frac{1}{1 + \exp\left\{-\pi\boldsymbol{\alpha}^T\mathbf{x}/\sqrt{3}\right\}}.\end{aligned}$$

La Figura 3.1 muestra los contornos de π para diferentes valores del parámetro $\boldsymbol{\alpha}$ y la matriz de precisión $\mathbf{A}$, en el caso bidimensional. Además, la gráfica muestra $10,000$ simulaciones que se obtuvieron utilizando el Teorema 3.1.2. Esto nos permite observar el comportamiento de una muestra independiente.

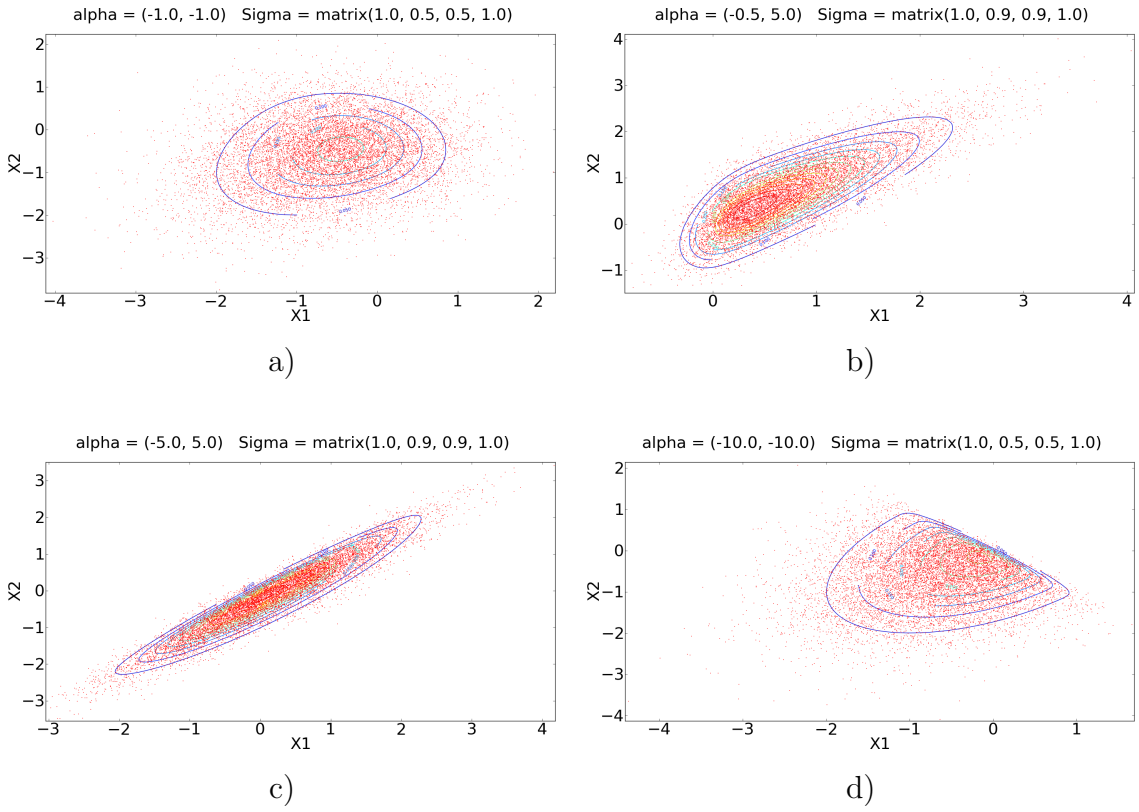

Figura 3.1: Contornos y muestra independiente de la distribución Normal sesgada mediante una distribución Logística.

Para implementar el algoritmo necesitamos calcular el gradiente y el Hessiano de $-\log \pi(\mathbf{x})$. Para ello comenzamos notando que

$$-\log \pi\left(\mathbf{x}\right) = -\log 2 - \log f_0(\mathbf{x}) - \log G\left(\boldsymbol{\alpha}^T \mathbf{x}\right). \tag{3.10}$$

.

Luego

$$\nabla(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}\left(-\log f_0(\mathbf{x})\right) - \frac{\partial}{\partial \mathbf{x}}\left(\log G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right). \tag{3.11}$$

Calculamos por separado las parciales de la expresión anterior. Así

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} \left(-\log f_0(\mathbf{x})\right) &= \frac{\partial}{\partial \mathbf{x}} \left(\frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x}\right) \\ &= \mathbf{A}\mathbf{x}, \end{aligned} \tag{3.12}$$

y

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} \left(\log G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right) &= \frac{1}{G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)} G'\left(\boldsymbol{\alpha}^T \mathbf{x}\right) \boldsymbol{\alpha} \\ &= \frac{1}{G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)} \left[G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\left(1 - G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right)\right] \boldsymbol{\alpha} \\ &= \left(1 - G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right) \boldsymbol{\alpha}, \end{aligned} \tag{3.13}$$

lo cual se debe a que $G'\left(\boldsymbol{\alpha}^T \mathbf{x}\right) = G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\left(1 - G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right)$. Sustituyendo (3.12) y (3.13) en (3.11) obtenemos

$$\nabla(\mathbf{x}) = \mathbf{A}\mathbf{x} - \left(1 - G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right) \boldsymbol{\alpha}. \tag{3.14}$$

Por otro lado

$$\begin{aligned} \mathbf{H}(\mathbf{x}) &= \frac{\partial}{\partial \mathbf{x}} \left(\mathbf{A}\mathbf{x} - \left[1 - G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right] \boldsymbol{\alpha}\right) \\ &= \frac{\partial}{\partial \mathbf{x}} \left(\mathbf{A}\mathbf{x}\right) - \frac{\partial}{\partial \mathbf{x}} \left(\left[1 - G\left(\boldsymbol{\alpha}^T \mathbf{x}\right)\right] \boldsymbol{\alpha}\right) \\ &= \mathbf{A} + G'\left(\boldsymbol{\alpha}^T \mathbf{x}\right) \boldsymbol{\alpha}\boldsymbol{\alpha}^T \\ &= \mathbf{A} + g\left(\boldsymbol{\alpha}^T \mathbf{x}\right) \boldsymbol{\alpha}\boldsymbol{\alpha}^T, \end{aligned} \tag{3.15}$$

donde g es la función de densidad logística.

Con esto tenemos todos los ingredientes listos para experimentar con el algoritmo de la sección 2.2. Para analizar los datos de forma gráfica nos concentraremos en el caso $n = 2$. La Figura 3.2 muestra las simulaciones obtenidas para los mismos ejemplos de la

Figura 3.1, después de $10,000$ iteraciones del programa. El punto inicial de la cadena es $\mathbf{X}^{(0)} = (0, 0)$, lo cual hace innecesario el burn in.

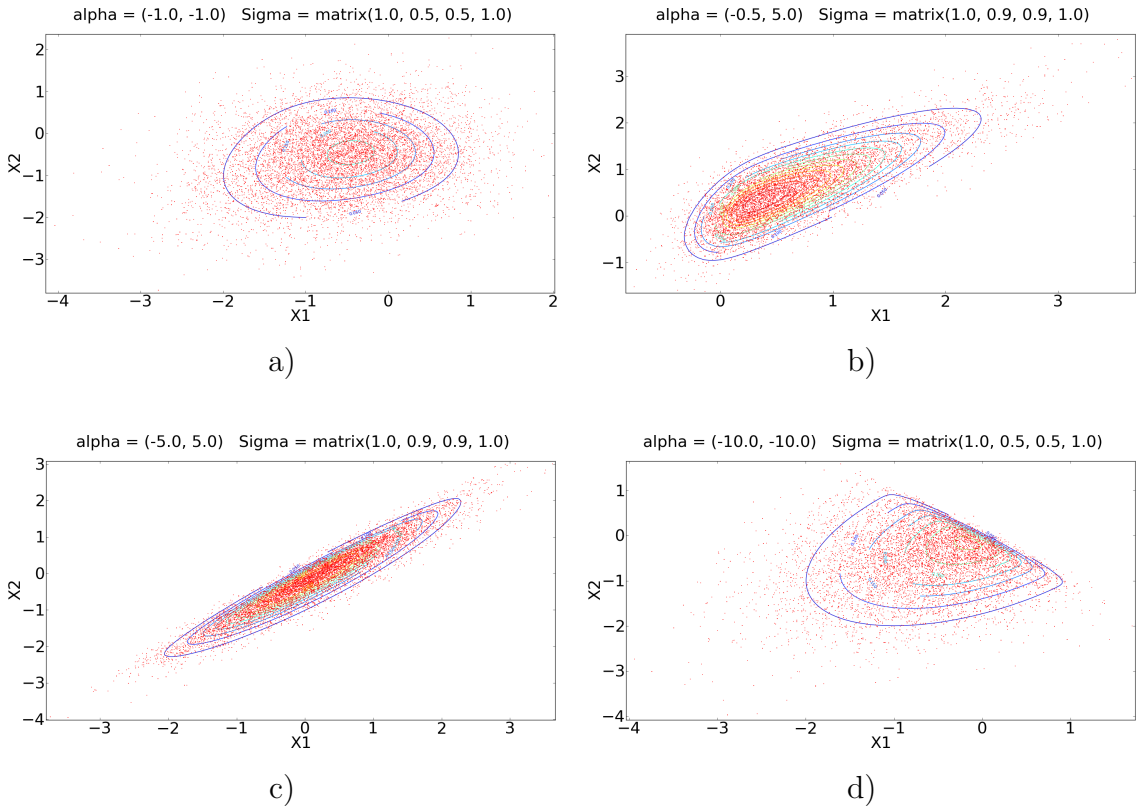

Figura 3.2: Simulaciones obtenidas con el algoritmo Gibbs direccional óptimo utilizando la distribución h^* .

Podemos ver que la distribución de las simulaciones generadas con el algoritmo Gibbs direccional óptimo es muy similar a la que se observa en la muestra independiente. Si bien esto es solamente una impresión visual, nos da idea del funcionamiento del algoritmo.

La Tabla 3.1 muestra la tasa de aceptación y el valor del Integrated Autocorrelation Time[1] (IAT) para cada uno de los ejemplos anteriores. El IAT da una idea de la calidad y correlación de la cadena resultante.

Hay que notar que el valor del IAT es pequeño en todos los casos. Eso es bueno ya

[1]Ver Apéndice.

	$\boldsymbol{\alpha}$	$\boldsymbol{\Sigma}$	IAT	Tasa de aceptación
a)	$(-1.0, -1.0)$	$(1.0, 0.5, 0.5, 1.0)$	4.039870	0.8712
b)	$(-0.5, 5.0)$	$(1.0, 0.9, 0.9, 1.0)$	6.944900	0.7693
c)	$(-5.0, 5.0)$	$(1.0, 0.9, 0.9, 1.0)$	3.909896	0.8485
d)	$(-10.0, -10.0)$	$(1.0, 0.5, 0.5, 1.0)$	7.629253	0.6892

Cuadro 3.1: IAT y tasa de aceptación de las simulaciones obtenidas con el algoritmo Gibbs direccional óptimo utilizando la distribución h^* .

que si queremos generar una muestra independiente de tamaño N, entonces el número de observaciones que debemos generar con nuestro algoritmo Gibbs direccional óptimo es aproximadamente $M = N * IAT$. Por otro lado, la tasa de aceptación es grande en todos los casos. Esto es muy relevante ya que nos indica que la cadena se mueve constantemente. Ambas medidas nos indican que el desempeño del algoritmo es bastante bueno, y confirman la impresión que tuvimos al observar la Figura 3.2.

Es importante mencionar que no es posible utilizar el algoritmo Gibbs sampler convencional en estos ejemplos particulares, ya que las distribuciones condicionales totales no son distribuciones conocidas de las cuales se pueda simular de forma sencilla. Esto es un punto a favor del esquema que estamos manejando ya que, incluso para ejemplos relativamente sencillos, el algoritmo Gibbs clásico no es viable.

Sin embargo, el algoritmo no es perfecto y presenta varios problemas importantes. Una restricción muy fuerte es que la matriz Hessiana $\mathbf{H}(\mathbf{x})$ tiene que ser definida positiva para todo $\mathbf{x} \in \mathbb{R}^n$, ya que de otra forma no existiría la aproximación Normal.

El problema más grande es que no podemos normalizar de forma analítica la función de densidad $h^*(\mathbf{e}) \propto \left(\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}\right)^{-1/2}$. Esto nos obliga a calcular la constante de normalización $K_{\mathbf{H}(\mathbf{x})} = \int_{\|\mathbf{e}\|} \left(\mathbf{e}^T\mathbf{H}(\mathbf{x})\mathbf{e}\right)^{-1/2} d\mathbf{e}$ numéricamente. La implementación del algoritmo se complica desde $n = 3$. En la actualidad muchas aplicaciones trabajan con distribuciones con dimensiones muy altas. Por esta razón el algoritmo que proponemos se vuelve inviable.

3.2. Direcciones de Eigenvectores

En la Sección 2.3 propusimos una distribución alternativa a h^* con la que se intenta eliminar las dificultades anteriormente mencionadas. La idea es tomar las direcciones $\mathbf{e}$ como los eigenvectores de la matriz $\mathbf{H}(\mathbf{x})$.

Veamos ahora el comportamiento del algoritmo con esta variante. La Figura 3.3 muestra el desempeño sobre los mismos ejemplos de la Figura 3.2. En ella podemos observar algo muy interesante. En los ejemplos a) y b) las simulaciones se encuentran bien ditribuidas sobre toda la región de interés. Sin embargo, en los ejemplos c) y d) las simulaciones están fuertemente concentradas en una región muy particular. Esto se debe a que en dicha región los eigenvalores de $\mathbf{H}(\mathbf{x})$ son muy contrastantes, uno es mucho más grande que el otro, y por consecuencia se obtienen simulaciones en direcciones similares por mucho tiempo.

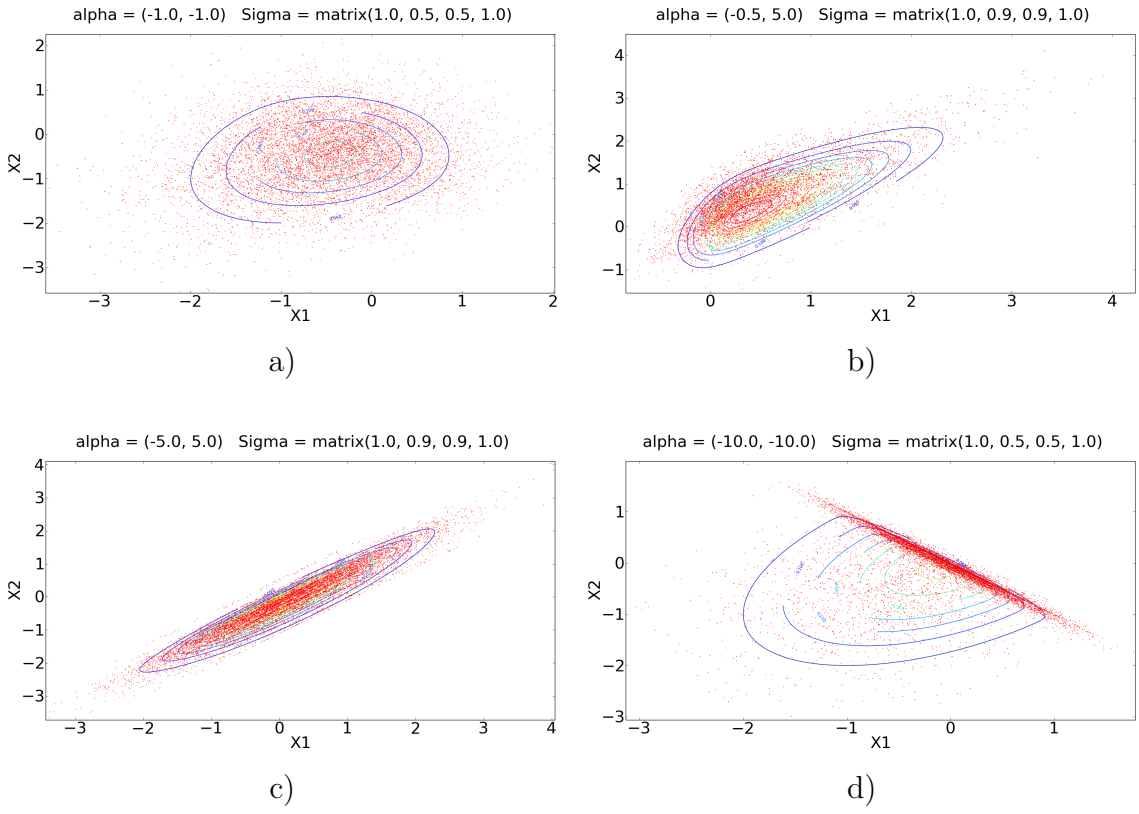

Figura 3.3: Simulaciones obtenidas con el algoritmo Gibbs direccional óptimo utilizando la distribución h_1 .

Esto parece indicar que este esquema funciona bien en casos sencillos, pero se vuelve ineficiente ante distribuciones raras. No obstante, es posible modificar la probabilidad con las que se escogen las direcciones para evitar en cierta medida estos problemas.

Nuevamente tomaremos las direcciones $\mathbf{e}$ como los eigenvectores de $\mathbf{H}(\mathbf{x})$, pero la probabilidad de escoger $\mathbf{e}_i$ ahora será proporcional a $(\lambda_i)^{-b}$, donde $b \sim Beta(1, 9)$, esto es $h_2(\mathbf{e}_i) \propto (\lambda_i)^{-b}$. Esto permite que las probabilidades puedan hacerse más balanceadas en regiones en las que hay un eigenvalor muy dominante.

La Figura 3.4 muestra el comportamiento del algoritmo resultante de esta modificación. Seguimos usando los mismos ejemplos para poder hacer comparaciones directas.

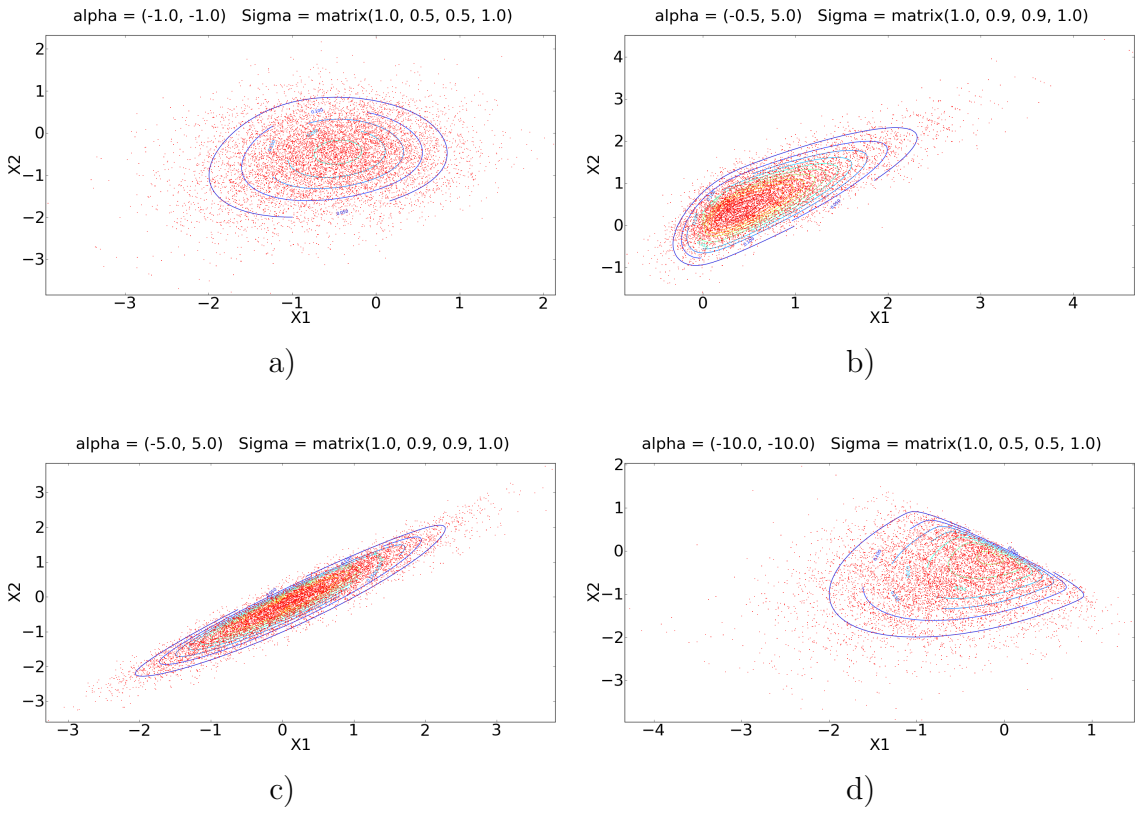

Figura 3.4: Simulaciones obtenidas con el algoritmo Gibbs direccional óptimo utilizando la distribución h_2 .

Los ejemplos a) y b) muestran un buen comportamiento de las simulaciones igual que

antes. Por otro lado, los ejemplos c) y d), los cuales identificamos como difíciles anteriormente, parecen haberse corregido pues ya no se aprecia la concentración de las simulaciones que aparecía en la Figura 3.3.

La Tabla 3.2 muestra los valores del IAT y la tasa de aceptación. En ella podemos observar un comportamiento similar al de la Tabla 3.1. Las tasas de aceptación se mantienen altas y no representan un problema. El IAT por su parte es ligéramente más alto que en la tabla anterior, salvo por el ejemplo c). No obstante, la diferencia no es muy grande, sobre todo si consideramos que el IAT varía entre muestra y muestra hasta por una unidad.

	$\boldsymbol{\alpha}$	$\boldsymbol{\Sigma}$	IAT	Tasa de aceptación
a)	$(-1.0, -1.0)$	$(1.0, 0.5, 0.5, 1.0)$	4.497301	0.8793
b)	$(-0.5, 5.0)$	$(1.0, 0.9, 0.9, 1.0)$	7.844630	0.7534
c)	$(-5.0, 5.0)$	$(1.0, 0.9, 0.9, 1.0)$	2.705416	0.8809
d)	$(-10.0, -10.0)$	$(1.0, 0.5, 0.5, 1.0)$	8.821470	0.7327

Cuadro 3.2: IAT y tasa de aceptación de las simulaciones obtenidas con el algoritmo Gibbs direccional óptimo utilizando la distribución h_2 .

Todo parece indicar que las correcciones sobre las probabilidades tuvieron un efecto positivo. Nuestra elección de la distribución $Beta(1, 9)$ fue meramente intuitiva. De hecho, los parámetros de la distribución Beta son una perilla del algoritmo que podemos usar a nuestro favor.

Entre los tres esquemas presentados, este último parece el más viable ya que muestra un buen desempeño y los valores del IAT son muy similares a los obtenidos con nuestro esquema original.

En el siguiente capítulo resumiremos los resultados que hemos obtenido y daremos nuestras conclusiones.

CAPÍTULO 4

Discusión y Conclusiones

En este trabajo de tesina se estudió la propuesta de Christen y Fox (2011) sobre Gibbs direccional óptimo. Como consecuencia se desarrollo un algoritmo Gibbs direccional que permite simular de diversas distribuciones objetivo mediante el uso de una aproximación Normal local.

El algoritmo propuesto requiere tres elementos indispensables para su implementación:

1. Conocer el gradiente de $-log\pi(\mathbf{x})$, $\nabla(\mathbf{x})$, para todo $\mathbf{x}$ en el soporte de la distribución objetivo.

2. Conocer el Hessiano de $-log\pi(\mathbf{x})$, $\mathbf{H}(\mathbf{x})$, para todo $\mathbf{x}$ en el soporte de la distribución objetivo.

3. Que $\mathbf{H}(\mathbf{x})$ sea una matriz positiva definida para todo $\mathbf{x}$ en el soporte de la distribución objetivo.

Es cierto que estas tres condiciones no se tienen para cualquier distribución objetivo. Sin embargo, para los casos en los que se cuenta con esta información, el algoritmo propuesto es una excelente opción ya que no requiere tener mayor conocimiento de π. En aplicaciones complejas donde la dimensionalidad es alta este enfoque resulta muy conveniente.

Después de las pruebas realizadas en el Capítulo 3 consideramos que la distribución de direcciones óptimas que genera mejores resultados es h^*. La mejor característica de esta

distribución es que tiene como soporte la esfera S^n. Sin embargo, es necesario calcular la constante de normalización de dicha distribución numéricamente. Hacer esto no es sencillo para aplicaciones con dimensionalidad alta.

Debido a los problemas que presenta h^* optamos por elegir la distribución h_2 como la distribución de direcciones óptimas. Esta elección nos permite trabajar de forma sencilla con distribuciones objetivo en dimensiones altas. La distribución tiene la gran ventaja de ser discreta. Además, la posibilidad de variar los parámetros de la distribución Beta la hacen más versátil.

En la tesina sólo consideramos tres distribuciones para las direcciones $\mathbf{e}$. Por supuesto éstas no son las únicas posibles. Aún queda mucho por investigar sobre las direcciones óptimas.

El algoritmo únicamente fue probado con ejemplos sencillos para los cuales se tiene una gran cantidad de información. En estos casos se obtuvieron buenos resultados. El siguiente paso es probar el algoritmo en una distribución más compleja como las que aparecen en el área de problemas inversos. Esto es un posible proyecto a futuro.

Por último, es importante mencionar que el algoritmo fue implementado en el lenguaje de programación orientado a objetos Python. Durante la programación del algoritmo pudimos apreciar que este lenguaje es muy superior al lenguaje R. Una de sus principales ventajas es que no está limitado al cómputo estadístico ya que ofrece una extensa gama de herramientas para computo científico en general. Si bien es cierto que es un lenguaje más completo, y por ende más complejo, su aprendizaje es sólo un poco más complicado que el aprendizaje de R. Por estas razones consideramos que Phyton es una excelente opción tanto para el desarrollo de aplicaciones académicas sencillas como para el desarrollo de aplicaciones de investigación complejas.

APÉNDICE A

Integrated Autocorrelation Time

Sea $\mathbf{X}^{(0)} = \mathbf{x}^{(0)}, \mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \ldots, \mathbf{X}^{(m)} = \mathbf{x}^{(m)} \in \mathcal{X}$ una realización de una cadena de Markov con distribución ergódica π. Suponga que dicha realización es obtenida a través de un algoritmo MCMC y que la cadena se encuentra en equilibrio desde $\mathbf{X}^{(0)}$.

Sea g una función medible tal que $g : \mathcal{X} \to \mathbb{R}$. Nos interesa calcular $\mu_g = \mathbb{E}_\pi[g(\mathbf{X})]$. Como $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ es una muestra correlacionada de π, entonces podemos utilizarla para estimar μ_g mediante la siguiente expresión:

$$\hat{\mu}_g = \frac{1}{m} \sum_{t=1}^{m} g\left(\mathbf{x}^{(t)}\right). \tag{A.1}$$

Para evaluar que tan buen estimador es $\hat{\mu}_g$ calculamos $Var[\hat{\mu}_g]$. Si la muestra fuese independiente, entonces

$$Var[\hat{\mu}_g] = \frac{Var[g]}{m}. \tag{A.2}$$

Sea

$$\mathcal{C}_g(t) = Cov\left[g(\mathbf{X}^{(i)}), g(\mathbf{X}^{(i+t)})\right] \tag{A.3}$$

la autocovarianza entre $\mathbf{X}^{(i)}$ y $\mathbf{X}^{(i+t)}$, y sea

$$c_g(t) = \frac{\mathcal{C}_g(t)}{\mathcal{C}_g(0)} \tag{A.4}$$

la autocorrelación a un corrimiento (lag) t, que no depende de i ya que la cadena es homogenea y ergódica. Es importante notar que la varianza de g es un acaso particular de la autocovarianza, esto es $Var[g] = \mathcal{C}_g(0)$. Luego

$$\begin{aligned}
Var\left[\hat{\mu}_g\right] &= Var\left[\frac{1}{m}\sum_{t=1}^{m} g\left(\mathbf{x}^{(t)}\right)\right] \\
&= \frac{1}{m}\sum_{t=1}^{m}\sum_{s=1}^{m} Cov\left[g\left(\mathbf{x}^{(t)}\right), g\left(\mathbf{x}^{(s)}\right)\right] \\
&= \mathcal{C}_g(0) + \mathcal{C}_g(1) + \ldots + \mathcal{C}_g(m) \\
&\quad + \mathcal{C}_g(1) + \mathcal{C}_g(0) + \mathcal{C}_g(1) + \mathcal{C}_g(2) + \ldots + \mathcal{C}_g(m-1) \\
&\quad + \mathcal{C}_g(2) + \mathcal{C}_g(1) + \mathcal{C}_g(0) + \mathcal{C}_g(1) + \ldots + \mathcal{C}_g(m-2) \\
&\quad \vdots \\
&\quad + \mathcal{C}_g(m) + \mathcal{C}_g(m-1) + \mathcal{C}_g(m-2) + \ldots + \mathcal{C}_g(0) \\
&= \frac{1}{m^2}\left[m\mathcal{C}_g(0) + 2(m-1)\mathcal{C}_g(1) + 2(m-2)\mathcal{C}_g(2) + \ldots + 2\mathcal{C}_g(m)\right] \\
&= \frac{m\mathcal{C}_g(0)}{m^2}\left[1 + 2\sum_{t=1}^{m}\left(1 - \frac{t}{m}\right)\frac{\mathcal{C}_g(t)}{\mathcal{C}_g(0)}\right] \\
&= \frac{Var[g]}{m}\left[1 + 2\sum_{t=1}^{m}\left(1 - \frac{t}{m}\right)c_g(t)\right] \\
&= \frac{Var[g]}{m}\tau_g(m),
\end{aligned} \tag{A.5}$$

donde

$$\tau_g(m) = 1 + 2\sum_{t=1}^{m}\left(1 - \frac{t}{m}\right)c_g(t). \tag{A.6}$$

Entonces, las ecuaciones (A.2) y (A.5) muestran que la varianza del estimador obtenida con la muestra MCMC es más grande que la varianza del estimador obtenida con la

muestra independiente por el factor $\tau_g(m)$.

En la ecuación (A.2) $Var[\hat{\mu}_g]$ decrece como $1/m$, donde m es el número de muestras independientes. Por otro lado, en la ecuación (A.5) $Var[\hat{\mu}_g]$ decrece como $\tau_g(m)/m$. Entonces $\tau_g(m)$ es el número de muestras correlacionadas con el mismo poder de reducción de la varianza que el de una muestra independiente.

Sea

$$\tau_g = \tau_g(\infty) = \lim_{m\to\infty} \tau_g(m). \tag{A.7}$$

Si suponemos que τ_g converge, entonces

$$\tau_g = 1 + 2\sum_{t=1}^{\infty} c_g(t). \tag{A.8}$$

El valor τ_g recibe el nombre de *Integrated Autocorrelation Time* (IAT). En general se desea desarrollar algoritmos MCMC con un IAT pequeño pues permite estimar de forma más precisa $Var[\hat{\mu}_g]$, evitando generar muestras demasiado grandes.

En la práctica el IAT nos indica el número de simulaciones que debemos descartar de nuestra muestra MCMC para obtener una muestra casi independiente, es decir, nos indica el valor del *thinning*[1]

Estas ideas son sólo una breve introducción al IAT. Para mayores referencia el lector puede consultar Roberts and Rosenthal (2001)[16]. Para la estimación de τ_g ver Geyer (1992)[7].

[1] Ver Sección 1.2

Bibliografía

[1] Azzalini, Adelchi (2005). *The Skew-normal Distribution and Related Multivariate Families*. Scandinavian Journal of Statistics, Vol. 32, Pág. 159-188.

[2] Bernardo, José M. y Smith, Adrian F. M. (1994). *Bayesin Theory*. Jhon Wiley. Chichester.

[3] Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag. New York.

[4] Calvetti, Daniela y Somersalo, Erkki (2007). *Introduction to Bayesian Scientific Computing*. Springer Verlag. New York.

[5] Christen, J. Andrés y Fox, Colin (2011). *Optimal Direction Gibbs for Sampling from Very High Dimension Normal Distributions*. http://www.cimat.mx/~jac/ChristenFox2011.pdf

[6] Cover, Thomas M. y Thomas, Joy A. (1991). *Elements of Information Theory*. Jhon Wiley. New York.

[7] Geyer, Charles J. (1992). *Practical Markov Chain Monte Carlo*. Statistical Science, Vol. 7, Pág. 473-511.

[8] Hastings, W. Keith (1970). *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. Biometrika, Volumen 57, Número 1, Pag. 97-109.

[9] Kaipio, Jari y Somersalo, Erkki (2004). *Statistical and Computational Inverse Problems*. Springer Verlag. New York.

[10] Karlin, Samuel y Taylor, Howard M. (1975). *A First Course in Stochastic Processes*. Segunda Edición. Academic Press. San Diego.

[11] MacKay, David J.C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. New York.

[12] Metropolis, Nicholas; Rosenbluth, Arianna; Rosenbluth, Marshall; Teller, Augusta H. y Teller, Edward (1953). *Equation of State Calculations by Fast Computing Machines*. The Journal of Chemical Physics, Volumen 21, Número 6.

[13] Meyn, Sean y Tweedie, Richard L. (1993). *Markov Chains and Stochastic Stability*. Cambridge University Press. New York.

[14] Robert, Christian P. (2001). *The Bayesian Choice*. Segunda Edición. Springer-Verlag. New York.

[15] Robert, Christian P. y Casella, George (2004). *Monte Carlo Statistical Methods*. Segunda Edición. Springer Verlag. New York.

[16] Roberts, Gareth O. y Rosenthal, Jeffrey S. (2001). *Optimal Scaling for Various Metropolis-Hastings algorithms*. Statistical Science, Vol. 16, Pág. 351-367.

[17] Smith, Robert L. y Kaufman, David E. (1998). *Direction Choice for Accelerated Convergence in Hit-ans-Run Sampling*. Operations Research, Volumen 46, Pág. 84-95.

[18] Vogel, Curtis R. (2002). *Computational Methods for Inverse Problems*. SIAM. Philadelphia.