CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS

MAESTRÍA EN PROBABILIDAD Y ESTADÍSTICA

Gibbs Direccional Óptimo: Distribución Normal Truncada

T E S I N A

PARA OBTENER EL TÍTULO DE:

Maestro en Ciencias
con Especialidad en Probabilidad y Estadística

PRESENTA:

Diego Andrés Pérez Ruiz

ASESOR:

Dr. José Andrés Christen Gracia

Mayo del 2011

JURADO ASIGNADO:

Presidente:	Dr. José Héctor Morales Bárcenas.
Secretario:	Dra. Angélica Hernández Quintero.
Vocal:	Dr. José Andrés Christen Gracia.

Lugar donde se realizó la tesina:

Centro de Investigación en Matemáticas A.C , CIMAT.

ASESOR:

Dr. José Andrés Christen Gracia

SUSTENTANTE:

Diego Andrés Pérez Ruiz

Resumen

Los métodos de simulación de cadenas de Markov (MCMC) son algoritmos usados para producir simulaciones de distribuciones complejas. Consisten en diseñar una cadena de Markov cuya distribución estacionaria es la distribución de la cual se quiere simular. La necesidad de simular la distribución normal truncada es frecuente en problemas inversos, por lo que en ocasiones se recurre al muestreo de Gibbs, un algoritmo de MCMC que simula de manera sistemática o de manera aleatoria de las distribuciones condicionales. El Gibbs generalizado toma una dirección arbitraria en el espacio sobre el cual se simula. La pregunta es: ¿Qué distribución de direcciones sería la óptima? Éste trabajo da una respuesta a esta pregunta para el caso de la distribución normal truncada.

Palabras clave: MCMC, Gibbs, Metropolis Hastings, Normal Truncada, Problemas Inversos.

Abstract

Markov Chain Monte Carlo (MCMC) algorithms are used to produce simulations of complex distribution. It consists in designing a Markov Chain whose stationary distribution is the distribution of interest. The need to simulate from the truncated normal distribution is common in Bayesian inference and in inverse problems, the Gibbs algorithm, MCMC algorithm that simulates systematically or randomly from full conditional distributions, is an useful strategy to simulating from the Truncated Normal distribution. The general Gibbs sampler takes an arbitrary direction in space along which to sample. The natural question is: What distribution of directions would be optimal? In this work we give an answer to this question for the truncated Normal distribution.

Keywords: Markov Chain Monte Carlo, Gibbs, Metropolis Hastings, Truncated Normal, Inverse Problems.

"You can never know everything". Lan said quietly, "and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyway"

Robert Jordan, *Winter's Heart, Book IX of the Wheel of Time*.

Dedicatorias

Por un conjunto no numerable de razones:

a mis padres.

Agradecimientos

Gracias al Consejo Nacional de Ciencia y Tecnología por la beca de estudios de posgrado durante estos 2 años en el programa.

Gracias al Centro de Investigación en Matemáticas A.C. por permitirme estudiar en sus aulas, por el uso de la biblioteca, del cubículo y de las instalaciones en general y, sobre todo, por brindarme una educación de calidad.

Gracias a mis padres, por todo su apoyo incondicional en cada instante de mi vida, gracias por su comprensión y amor ofrecidos en todo momento, ustedes son mi inspiración para alcanzar mis metas.

Gracias a mis hermanos Dorian Andrea y Daniel Alberto, son parte de mi ser y siempre los llevo conmigo, por compartir emociones, sentimientos y momentos tan especiales.

Gracias a mi asesor, Dr. Andrés Christen, por su apoyo constante y su paciencia en la elaboración de este trabajo, por sus correciones y por la manera de hacerme ver el mundo, gracias por hacerme crecer en muchos aspectos de mi vida.

Gracias al Dr. Víctor Rivero Mercado, coordinador de la Maestría en Probabilidad y Estadística del CIMAT, por su apoyo recibido durante el posgrado.

Gracias al Dr. Rogelio Ramos Quiroga, por sus consejos y haberme transmitido sus experiencias profesionales en todo momento. Gracias por ser como es.

Gracias a mis sinodales, Dra. Angélica Hernández Quintero y Dr. Héctor Morales por las revisiones y a portaciones a este trabajo.

Índice general

Índice de figuras

Índice de Algoritmos

Introducción

> *Mathematics, rightly viewed, posseses not only truth, but supreme beauty -a beauty cold and austere-, like that of sculpture.*
>
> **Bertrand Russell**

El enfoque Bayesiano de la inferencia estadística está basado en axiomas que proporcionan una estructura lógica y garantizan la consistencia de los métodos propuestos, constituyen un paradigma completo a la inferencia y una revolución científica en el sentido de Kuhn (1962). La Estadística Bayesiana intenta crear una teoría para hacer inferencia. Sin embargo, para lograrlo hace uso de herramientas matemáticas que describan, por medio de las distribuciones de probabilidad, la incertidumbre en el problema.

El rol de los métodos de Monte Carlo y la simulación en las ciencias han incrementado su importancia durante los últimos 10 años, debido a que juegan un rol central en el desarrollo de disciplinas como la física, las ciencias sociales y la biología. Además, los cálculos que generalmente aparecen en el análisis Bayesiano se han vuelto viables debido a dichos métodos, dando pauta a un sinnúmero de aplicaciones en la estadística Bayesiana, permitiendo la creación de nuevos métodos y el mejoramiento de los ya existentes.

Las técnicas que comprende MCMC son generales, su uso ha revolucionado por completo la estadística bayesiana. Hoy se manejan modelos muy complejos en donde las distribuciones no son convencionales y es necesario utilizar métodos eficientes para simular de ellos. A esto se debe la necesidad de desarrollar técnicas para simular dichas distribuciones.

Los métodos de simulación de cadenas de Markov (MCMC) son algoritmos usados para producir simulaciones de distribuciones complejas, típicamente en dimensiones altas. Consisten en diseñar una cadena de Markov cuya distribución estacionaria es la distribución objetivo o

de interés.

La necesidad de simular de la distribución normal truncada es frecuente en la inferencia bayesiana (donde surgen problemas con un espacio de parámetros acotados) y en problemas inversos. Es raro el caso en el que es posible hacer el cálculo analítico, y la integración numérica puede ser muy complicada para dimensiones altas. Es por ello que en ocasiones se recurre al muestreo de Gibbs, un algoritmo de MCMC que simula de manera sistemática o de manera aleatoria de las distribuciones condicionales.

El Gibbs generalizado toma una dirección arbitraria en el espacio sobre el cual se simula (en lugar de solamente utilizar los ejes canónicos, como en Gibbs tradicional). Una serie de direcciones puede ser usada. La pregunta natural que surge es:

¿Qué distribución de direcciones sería la óptima?

El presente trabajo da una respuesta a la pregunta anterior.

Para ello, en el Capítulo 1, se da una introducción a la Inferencia Bayesiana, se revisan las técnicas más usuales basadas en la simulación de cadenas de Markov y el algoritmo de Metropolis-Hastings, así como el del muestreo de Gibbs. Al final del primer capítulo se da una introducción al análisis bayesiano de problemas inversos lineales, en los cuales de manera natural surge la distribución normal truncada como distribución posterior.

El Capítulo 2 trata de responder la pregunta inicial, para el caso de la distribución normal truncada. Construimos una cadena de Markov mediante el algoritmo Gibbs y generalizamos esta idea mediante el algoritmo de Gibbs Direccional Óptimo, que muestrea a lo largo de una dirección. Utilizamos la divergencia de Kullback-Leibler para encontrar mejores direcciones en el algoritmo que minimicen dicha divergencia.

En el Capítulo 3, a través de un modelo de metabolismo del hígado en un estado estacionario, utilizamos la inferencia bayesiana en problemas inversos lineales, donde vemos que surge la distribución normal truncada y es de interés simular de ella. Al final de este capítulo se incluyen las comparaciones entre el Gibbs Direccional Óptimo, Gibbs sistemático o tradiccional y el Gibbs aleatorio para el caso de la distribución normal truncada.

En el Capítulo 4, se presentan las conclusiones más relevantes del trabajo. Se incluye, al final, un Anexo sobre el Integrated Autocorrelation Time (IAT), herramienta importante utilizada en el análisis de simulaciones de MCMC, pues indica el número de simulaciones que debemos descartar de nuestra muestra para obtener una muestra pseudo independiente.

El código del programa se pone a disposición del lector en la página de internet del Dr. Andrés Christen (http://www.cimat.mx/∼jac/software/) dicho programa fue desarrollado en el lenguaje de programación Pyhton utilizando las librerías de Numpy y Scipy.

Guanajuato, Mayo del 2011.

Capítulo 1

Inferencia Bayesiana

En este capítulo se abordan los temas que sustentan el trabajo. Primero se dará una introducción a la Inferencia Bayesiana con el propósito de mostrar los elementos para la construcción de un modelo estadístico bayesiano. En segundo lugar examinaremos las técnicas más usuales basadas en la simulación de cadenas de Markov y revisaremos el algoritmo de Metropolis-Hastings, así como el del muestreo de Gibbs.

Finalmente, daremos una introducción al análisis de problemas inversos lineales, cuyo propósito es estimar parámetros de interés dado los datos que están relacionados de manera indirecta a los parámetros. Explicaremos el enfoque bayesiano para problemas inversos y discutiremos la interpretación que surja de este enfoque.

Sin embargo, esto corresponde a una breve introducción al tema, por lo que referimos al lector a Robert (2001), que presenta los fundamentos del análisis bayesiano; Gamerman y Migon (1993), que ofrece una introducción a la estadística bayesiana, en especial la relación entre la estadística bayesiana y frequentista; O'Hagan (1994), como una buena referencia para la Inferencia Bayesiana; Berger (1985), para revisar la teoría de decisión bayesiana, Bernardo y Smith (1994), libro que ejemplifica los conceptos clave y los resultados teóricos en estadística bayesiana, y finalmente Kaipio y Somersalo (2004) libro clásico en el área de problemas inversos.

1.1. Teorema de Bayes

El enfoque Bayesiano se desarrolla en la presencia de observaciones x, que se pueden describir a través de una distribución de probabilidad con densidad o función de probabilidad $f(x|\theta)$, donde θ representa el vector de parámetros de interés .

Asumiremos que las observaciones $x_1,\ldots,x_n$ se generan a partir de una distribución de probabilidad paramétrica. Esto es x_i para $(1 \leq i \leq n)$ tiene distribución con densidad $f_i(x_i|\theta_i)$ en $\mathbb{R}^p$ tal que la función f_i es conocida y el parámetro θ_i es desconocido. Este modelo se puede representar de manera más sencilla de la siguiente forma:

$$x_i \sim f(x_i|\theta), \tag{1.1}$$

donde x es el vector de observaciones y θ representa al vector de parámetros $\theta_1,\ldots,\theta_n$. Usamos la notación x se distribuye de acuerdo a f o $x \sim f$ en lugar de x es una observación de la distribución con densidad f.

Consideremos n observaciones independientes idénticamente distribuidas (i.i.d) $x_1,\ldots,x_n$ de una distribución común, $f(x_1|\theta)$, de tal forma que podemos escribir 1.1 como:

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{1.2}$$

que se conoce como la función de verosimilitud para una muestra de tamaño n, y denotamos $l(\theta;x) = f(x|\theta)$.

Una vez definido el modelo, el propósito principal de la estadística paramétrica es hacer inferencia sobre el parámetro θ. Sin embargo, es posible que el investigador tenga información sobre el valor del parámetro y es posible incorporar dicha información al análisis.

Bayes y Laplace consideraron que la incertidumbre del parámetro θ podría ser modelada por medio de una distribución de probabilidad π, a la cual la llamaron *distribución a priori* (o inicial). El enfoque Bayesiano incorpora dicha información al análisis a través de la distribución *a priori* $\pi(\theta)$.

Una vez que el problema cuenta con estos dos elementos - la distribución a priori y la verosimilitud -, la inferencia se basa en la distribución de θ condicionada en x, $\pi(\theta|x)$, a la cual

llamamos *distribución a posteriori* (o posterior) y, utilizando el teorema de Bayes, esto es:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}. \tag{1.3}$$

De lo anterior, es posible observar que el término $\int f(x|\theta)\pi(\theta)d\theta$ es simplemente una constante de normalización, por lo que el teorema de Bayes puede ser escrito de manera más compacta de la siguiente manera:

$$\pi(\theta|x) \propto l(\theta;x)\pi(\theta), \tag{1.4}$$

es decir, $\pi(\theta|x)$ es proporcional a la verosimilitud multiplicada por la distribución a priori de θ.

Otro aspecto importante es que no seguiremos la convención usual de que las variables aleatorias están representadas por letras mayúsculas, X, y la realización por x. Esto se debe a que desde el punto de vista Bayesiano, siempre condicionamos con respecto al valor observado x, y consideramos al parámetro θ como la variable aleatoria.

En estadística bayesiana se suele hacer uso indiscriminado de abusos de notación, que, sin embargo, resulta en textos más compactos. En términos más precisos, deberíamos escribir

$$\pi_{\theta|x}(t|X) = \frac{f_{x|\theta}(X|t)\pi_\theta(t)}{\int_{t\in M_\theta} f_{x|\theta}(X|t)\pi_\theta(t)dt}. \tag{1.5}$$

en lugar de $\pi(\theta|x) \propto l(\theta;x)\pi(\theta)$. Donde M_θ denota el espacio parametral.

La aportación principal de un modelo estadístico bayesiano es considerar una distribución de probabilidad en los parámetros. En términos estadísticos, el Teorema de Bayes actualiza la información de θ extrayendo información de θ contenida en las observaciones x.

Mientras que las dificultades relacionadas con los métodos de máxima verosimilitud son principalmente problemas de optimización (múltiples modas, solución a la ecuación de verosimilitud, entre otros) el enfoque bayesiano presenta en su mayoría con problemas de integración.

Los recursos de cómputo con los que se contaba hace 20 años permitían que se conociera cierto tipo de cálculos explícitos, llamadas las *conjugadas a priori*. Éstas son distribuciones *a priori* para las cuales las correspondientes distribuciones *a posteriori* son miembros de la

misma familia *a priori* original. La principal motivación para usar *a prioris conjugadas* es su fácil manejo, ver Bernardo y Smith (1994), Capítulo 1 con tablas de conjugadas.

Cuando no se tienen distribuciones conjugadas, sólo se tienen $l(\theta;x)\pi(\theta)$, y uno tiene como obstáculo la integración de la posterior para encontrar la constante de normalización. Desde hace mucho años se ha buscado una solución a dicho problema; la solución que se ha implementado desde 1990 (aún cuando en la física estadística ya se conocía desde los años setentas) es la simulación de cadenas de Markov o Markov Chain Monte Carlo. En la siguiente sección se explican las técnicas de simulación de Cadenas de Markov más conocidas.

1.2. Métodos de Simulación de Cadenas de Markov

Los métodos que se consideran en esta sección se basan en la simulación de cadenas de Markov, cuya distribución límite corresponde a la distribución de interés, de la que se desea simular en este caso, de la posterior $\pi(\theta|x)$. En el caso de Inferencia Bayesiana, el objetivo es simular una cadena que converja a una distribución límite igual a la distribución posterior de los parámetros de interés. Sin embargo, un problema común es que no se obtienen muestras independientes, debido a la estructura markoviana de simulación.

1.2.1. El algoritmo de Metropolis - Hastings

El algoritmo de Metropolis - Hastings fue inicialmente desarrollado por Metropolis para el cálculo de las propiedades de sustancias químicas en el año de 1950. La formulación matemática en el ámbito estadístico se debe a Hastings casi veinte años después. Dicho método tiene una variedad de aplicaciones en la física matemática y en la reconstrucción de imágenes. El método fue publicado por Metropolis en su artículo *Equations of state calculations by fast computing machines* publicado en el Journal of Chemical Physics, Metropolis (1953).

Antes de introducir de manera formal el algoritmo, definiremos lo que es un Método Monte Carlo de Cadenas de Markov.

Definición 1 *Un Método Monte Carlo de Cadenas de Markov o MCMC, para la simulación de una distribución f es cualquier método que produce una Cadena de Markov ergódica $(X^{(t)})$ cuya distribución estacionaria es f, y para la cual existe un algoritmo (eficiente) de simulación.*

El principio general de los algoritmos MCMC es que dado un valor inicial arbitrario $x^{(0)}$, la cadena $(X^{(t)})$ se genera utilizando un kernel de transición con distribución estacionaria, digamos f, la cual garantiza la convergencia en distribución de $(X^{(t)})$ a f. Dado que la cadena es ergódica, el valor $x^{(0)}$ en principio no es importante.

El algoritmo de Metropolis-Hastings comienza con una distribución objetivo (o de interés), $f(\theta) = \pi(\theta|x) \propto l(\theta;x)\pi(\theta)$ y una distribución instrumental $q(\cdot|x)$.

La distribución objetivo debe de satisfacer que el cociente:

$$\frac{f(\theta_1)}{f(\theta_2)},$$

es conocido para todo θ_1, θ_2. El algoritmo de Metropolis - Hastings asociado con la distribución objetivo y la densidad condicional q produce una Cadena de Markov $(\theta^{(t)})$ a través de la siguiente secuencia:

Algoritmo 1 Metropolis - Hastings.

1: Se inicializa con un valor inicial $\theta^{(0)}$.

2: Se considera una densidad $q(\cdot|\theta^{(j-1)})$ de la cual se genera ϕ_t.

3: Se evalúa la probabilidad:

$$\alpha(\theta^{(j-1)}, \phi_t) = \text{mín}\left\{1, \frac{\pi(\phi_t)q(\phi_t|\theta^{(j-1)})}{\pi(\theta^{(j-1)})q(\theta^{(j-1)}|\phi_t)}\right\}.$$

4: Se acepta ϕ_t con probabilidad α. Si ϕ_t es aceptado entonces $\theta^{(j)} = \phi_t$ de otra forma $\theta^{(j)} = \theta^{(j-1)}$.

5: Continuar con 2.

La distribución q se conoce como la distribución instrumental (o propuesta) y la probabilidad $\alpha(\theta^{(j-1)}, \phi_t)$ es la probabilidad de aceptación de Metropolis-Hastings. Un caso particular importante, es cuando $q(\theta, \phi_t) = q(\phi_t, \theta)$, obteniéndose:

$$\alpha(\theta^{(j-1)}, \phi_t) = \text{mín}\left\{1, \frac{\pi(\phi_t)}{\pi(\theta^{(j-1)})}\right\}. \tag{1.6}$$

El caso anterior se presenta cuando ϕ_t se genera a partir de una distribución simétrica en $\theta^{(j-1)}$. Este algoritmo siempre acepta valores ϕ_t tal que el cociente $\pi(\phi_t)/q(\phi_t|\theta^{(j-1)})$ se incrementa, comparado con el valor previo $\pi(\theta^{(j-1)})/q(\theta^{(j-1)}|\phi_t)$.

La probabilidad $\alpha(\theta^{(j-1)}, \phi_t)$ está definida sólo cuando $\pi(\theta^{(j-1)}) > 0$, sin embargo la cadena comienza con un valor inicial $\theta^{(0)}$ tal que $\pi(\theta^{(0)}) > 0$, y $\pi(\theta^{(j-1)}) > 0$ para toda $t \in \mathbb{N}$ puesto que valores de ϕ_t, tal que $\pi(\phi_t) = 0$ conducen a $\alpha(\theta^{(j-1)}, \phi_t) = 0$ y son rechazados por el algoritmo.

Observamos que el algoritmo anterior se encuentra definido para toda π y q, y no hemos puesto restricción alguna a éstas. Sin embargo, es necesario imponer condiciones mínimas para π y la distribución condicional q, con el objeto de que π sea la distribución límite de la cadena $(\theta^{(t)})$.

Asumamos que el soporte de π[1], $\mathscr{E}$, es un soporte conexo, puesto que uno no conexo invalida el algoritmo; se procede a encontrar un componente conexo y mostrar que los diferentes componentes conexos de $\mathscr{E}$ están ligados por el kernel del algoritmo de Metropolis-Hastings.

El soporte de π es un subconjunto de la σ-álgebra Σ, es:

$$supp\ (\pi) = \overline{\{\ A \in \Sigma | \pi(x) > 0,\ x \in A\}}. \tag{1.7}$$

Si el soporte $\mathscr{E}$ está truncado por q, esto es que existe un $A \subset \mathscr{E}$ tal que

$$\int_A \pi(\theta) d\theta > 0 \quad y \quad \int_A q(\phi|\theta) d\phi = 0, \quad \forall \theta \in \mathscr{E},$$

entonces el algoritmo no tiene a π como una distribución límite, puesto que $\theta^{(0)} \notin A$, y la cadena nunca visitará A. De manera que la condición mínima necesaria es que:

$$\bigcup_{\theta\ \in\ supp\ \ \pi} supp\ \ q(\cdot|\theta) \supset\ \ supp\ \ \pi. \tag{1.8}$$

Para ver que π es la distribución estacionaria de la cadena de Metropolis - Hastings, examinaremos el kernel de manera detallada y mostraremos que satisface las ecuaciones que balance detallado:

$$\pi(\theta) K(\phi, \theta) = \pi(\phi) K(\theta, \phi). \tag{1.9}$$

Teorema 1 Robert y Casella (2004). *Sea $(\theta^{(t)})$ la cadena producida por el algoritmo de Metropolis - Hastings, para cada distrubución condicional q cuyo suporte incluye $\mathscr{E}$,*

[1]Se dice que una función tiene soporte conexo si la adherencia del conjunto donde no es nula conforma un conjunto cerrado y acotado.

- *El kernel de transición de la cadena satisface las ecuaciones de balance detallado con π y la cadena es reversible.*
- *π es la distribución estacionaria de la cadena.*

Demostración. El kernel de transición asociado al algoritmo de Metropolis - Hastings es:

$$K(\theta,\phi) = \alpha(\theta,\phi)q(\phi|\theta) + (1-r(\theta))\delta_\theta(\phi), \tag{1.10}$$

donde $r(\theta) = \int \alpha(\theta,\phi)q(\phi|\theta)d\phi$, $\alpha(\theta,\phi)$ es la probabilidad de aceptación y $\delta_\theta(\phi)$ denota la Delta de Dirac en θ, de manera que se verifica que:

$$\alpha(\theta,\phi)q(\phi|\theta)\pi(\theta) = \alpha(\phi,\theta)q(\theta|\phi)\pi(\phi) \tag{1.11}$$

$$(1-r(\theta))\delta_\theta(\phi)\pi(\theta) = (1-r(\phi))\delta_\phi(\theta)\pi(\phi) \tag{1.12}$$

que establecen el balance detallado para la cadena de Metropolis - Hastings. ∎

La demostración de que π es la distribución estacionaria de la cadena se sigue del Teorema 2.

Hemos visto que el diseño de la cadena depende de la distribución instrumental que se utilice. La distribución instrumental define entonces un kernel de transición $K(\theta,\phi)$, esto es, la probabilidad (o densidad) de pasar de θ a ϕ y $K(\theta,A)$ es la medida de probabilidad de pasar de θ a un medible A.

Teorema 2 Robert y Casella (2004). *Si K cumple con balance detallado con respecto a π, entonces π es una densidad invariante de la cadena (y la cadena es reversible).*

Demostración.

$$\begin{aligned} K(\phi,B)\pi(\phi)d\phi &= \int\int_B K(\phi,\theta)\pi(\phi)d\theta d\phi \\ &= \int\int_B K(\theta,\phi)\pi(\theta)d\theta d\phi \\ &= \int_B \int K(\theta,\phi)d\phi\,\pi(\theta)d\theta \\ &= \int_B \pi(\theta)d\theta \\ &= \pi(B) \end{aligned}$$

∎

1.2.2. El Muestreador de Gibbs

El nombre del Muestreador de Gibbs o *Gibbs Sampling* proviene de un artículo histórico de Geman y Geman (1984), quienes por primera vez aplicaron el Muestreador de Gibbs a un campo aleatorio de Gibbs. El muestreador de Gibbs es un caso particular del algoritmo de Metropolis - Hastings.

El trabajo de Geman y Geman (1984) persuadió a Gelfand y Smith (1990) a escribir un artículo que provocó interés en los métodos bayesianos, en especial en el cómputo estadístico bayesiano. Cabe mencionar que Tanner y Wong (1987), y Besag y Clifford (1989), habían propuesto soluciones similares, pero no recibieron la misma respuesta de la comunidad estadística.

El muestreador de Gibbs Sistemático

Supóngase $\theta = (\theta_1, \ldots, \theta_p)$, un vector aleatorio, con $p \geq 1$, las θ's son componentes unidimensionales. Además, supongamos que podemos simular de las correspondientes densidades condicionales $\pi_1, \ldots, \pi_p$, esto es, podemos simular de

$$\theta_i \sim \pi_i(\theta_i | \theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p), \tag{1.13}$$

para $i = 1, 2, \ldots, p$. El algoritmo de Gibbs asociado a la transición de $\theta^{(t)}$ a $\theta^{(t+1)}$ es el siguiente:

Algoritmo 2 Muestreador de Gibbs Sistemático.

1: Se considera el iterado inicial $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$

2: Se actualiza el j-ésimo iterado generando:

$$\theta_1^{(j)} \sim \pi(\theta_1 | \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}) \tag{1.14}$$

$$\theta_2^{(j)} \sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)} \ldots, \theta_p^{(j-1)}) \tag{1.15}$$

$$\vdots \tag{1.16}$$

$$\theta_p^{(j)} \sim \pi(\theta_p | \theta_1^{(j)}, \ldots, \theta_{(p-1)}^{(j)}) \tag{1.17}$$

3: Ir al paso 2 e iterar hasta convergencia.

---

donde las densidades $\pi(\theta_i|\theta_{-i})$ son llamadas las condicionales totales, y una característica particular de este algoritmo es que éstas son las únicas densidades utilizadas para la simulación.

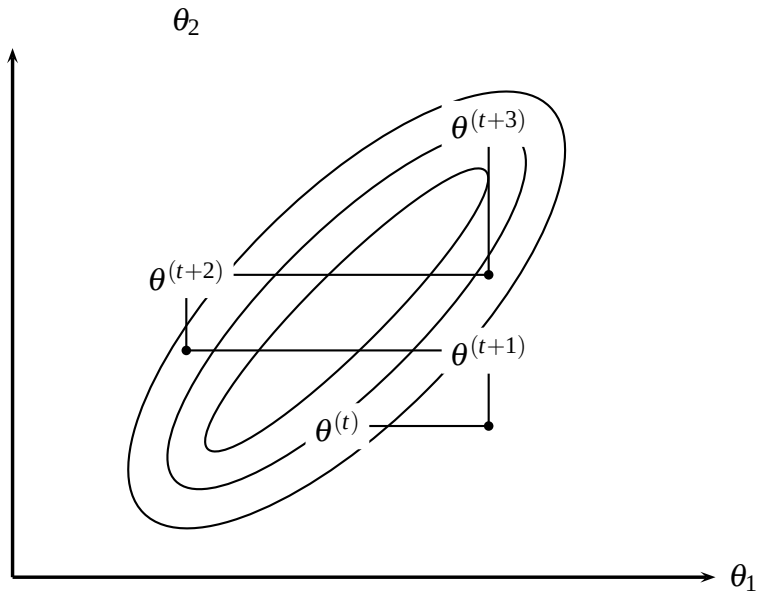

Figura 1.1: Representación gráfica del Gibbs Sistemático para el caso bidimensional

El muestreador de Gibbs Aleatorio

El muestreador de Gibbs aleatorio elige de manera aleatoria en cada paso la secuencia en la cual la p-ésima componente de $\theta_1, \ldots, \theta_p$, (el vector aleatorio p dimensional) es visitada.

Sea $\Gamma = \{\alpha_1, \ldots, \alpha_p\}$ el conjunto de probabilidades de visitar cierto componente, y asúmase que $0 < \alpha_i < 1$ para toda $i = 1, 2, \ldots, p$ y $\sum \alpha_i = 1$. El algoritmo del muestreador de Gibbs Aleatorio se resume de la siguiente manera:

Algoritmo 3 Muestreador de Gibbs Aleatorio.

1: Se considera un valor inicial $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.
2: En la i-ésima iteración:

- Aleatoriamente elija $j \in \{1, \ldots, p\}$ con probabilidad α_j
- Genere $\theta_j^{(t)} \sim \pi(\theta_j | \theta_{-j}^{(t-1)})$

3: Ir al paso 2 hasta convergencia.

1.2.3. Gibbs Direccional

El algoritmo Gibbs Direccional es un caso especial del Algoritmo de Metropolis-Hastings, fue propuesto de manera independiente por Boneh y Golan (1979) y Smith (1980) para generar puntos uniformemente distribuidos sobre regiones acotadas. Smith lo llama como un algoritmo de mezcla, y prueba la convergencia en Smith (1984).

En el contexto de estadística Bayesiana, Berger y Chen (1993) utilizan dicho algoritmo para simular de una distribución multinomial con parámetros acotados. Sin embargo, puede encontrarse en la literatura un sinfin de aplicaciones de dicho algoritmo, desde la estimación de la matriz de covarianzas hasta el análisis de coeficientes en modelos de regresión.

Algoritmo 4 Gibbs Direccional

1: Supóngase que el estado inicial $x^{(t)}$.
2: Seleccione una dirección $e^{(t)}$ con $||e^{(t)}|| = 1$.
3: Simular una v.a $r^{(t)}$ de:

$$f(r) \propto \pi(x^{(t)} + r^{(t)} e^{(t)}).$$

4: Actualizar $x^{(t+1)} = x^t + r^{(t)} e^{(t)}$.

Kauffman y Smith (1998) desarrollaron una dirección óptima para el algoritmo, y demostraron que existe una única distribución óptima de direcciones (en el sentido de que la tasa de convergencia es geométrica con la norma del supremo) para $e^{(t)}$ y π con soporte compacto.

En casos más generales, uno puede usar un grupo de transformaciones para los posibles mo-

vimientos, siempre y cuando los movimientos preserven una probabilidad invariante.

El algoritmo de Kauffman y Smith (1998) se comporta como el Muestreador de Gibbs Aleatorio y permite explorar la dirección escogida. Tiende a ser útil cuando la distribución objetivo es multimodal. Con respecto a aplicaciones de este algoritmo en estadística Bayesiana, Berger (1985) considera que este método es particularmente útil cuando θ tiene un espacio paramétrico fuertemente acotado.

1.2.4. Propiedades de Convergencia

Uno de los problemas fundamentales de los algoritmos de MCMC es determinar la convergencia de los mismos, así como seleccionar un número de iteraciones idóneas. Teóricamente, el valor inicial es irrelevante, ya que tanto el Gibbs Sampling como el algoritmo de Metropolis - Hastings convergen con cualquier valor inicial. Sin embargo, existen problemas en los que es importante partir de un valor inicial adecuado para garantizar una convergencia rápida.

En el caso del algoritmo de Metropolis - Hastings, la velocidad de convergencia depende fuertemente de la elección de la distribución instrumental.

En el caso de Gibbs Sampling, la velocidad de convergencia depende del problema. Una regla general que funciona, en muchos casos, es escribir a los parámetros usando bloques de dimensión alta para permitir que la cadena recorra eficientemente el espacio muestral.

Para poder determinar la convergencia, la prueba más fácil es la exploración gráfica de las componentes del vector de parámetros. Se considera que la cadena ha alcanzado el equilibrio cuando se estabiliza alrededor de un valor. Es importante explorar varias trayectorias con diferentes valores iniciales.

1.3. Análisis de Problemas Inversos Lineales

Los problemas inversos pueden aplicarse de diferentes maneras, desde problemas relacionados con la biomedicina hasta dinámica de fluidos. En este tipo de problemas el objetivo principal es estimar parámetros de interés, dado los datos que están relacionados de manera indirecta a los parámetros. Supongamos que los datos observados d depende de una x (desconocida), a través de un proceso de medición y queremos recuperar a x de d. En términos matemáticos, representamos el proceso de medición por una familia de modelos parametrizados por x, donde todos los parámetros necesarios están contenidos en x, incluyendo los parámetros de ruido.

Un ejemplo de problema inverso es la tomografía axial computarizada: uno desea reconstruir imágenes del cuerpo a través de mediciones realizadas por rayos X. Sin embargo, pequeñas cantidades de *ruido* en los datos nos llevan a estimaciones erróneas de los parámetros. Este fenómeno se conoce como *ill-posedness*, y muchas técnicas matemáticas conocidas como *regularización* se han desarrollado con el objetivo de tratar de entender el fenómeno.

1.3.1. Teoría de Reguralización

Para explicar la idea básica de la regularización, consideremos un problema lineal inverso de la forma $Ax = y$, los valores de algunos parámetros del modelo pueden ser obtenidos de los datos observados. Sea $\mathscr{H}_1$ y $\mathscr{H}_2$ espacios de Hilbert[2] separables de dimensión finita o infinita y sea $A : \mathscr{H}_1 \to \mathscr{H}_2$ un operador compacto. Estamos interesados en encontrar $x \in \mathscr{H}_1$ que satisface la siguiente ecuación:

$$Ax = y, \tag{1.18}$$

donde $y \in \mathscr{H}_2$ es conocido. Esta ecuación se conoce como la ecuación de *Fredholm* de primer orden. Para esta ecuación la solución

[2]Decimos que X es un espacio de Hilbert si es un espacio vectorial con producto interno completo. Es decir, si d es la métrica inducida por la norma en X, inducida a su vez por el producto interno, entonces (X,d) es completo. Si X es un espacio con producto interno, escribiremos el producto interno de $x,y \in X$ como (x,y), y la norma inducida como $\|x\|$. Entre las propiedades que satisface la norma inducida se encuentra la identidad del paralelogramo. Sirven para clarificar y para generalizar el concepto de series de Fourier, ciertas transformaciones lineales tales como la transformación de Fourier.

- existe $\Leftrightarrow y \in Rango(A)$,
- es única $\Leftrightarrow Ker(A) = \{0\}$.

Ambas condiciones deben satisfacerse para garantizar una única solución. Desde el punto de vista práctico, existe un obstáculo para resolver 1.18. Cuando se tiene un problema inverso lineal discreto que describa a un sistema lineal, y y x son vectores, donde el vector y representa datos medidos, generalmente contaminados por un error (de medición), y en lugar de tener $Ax = y$, tenemos:

$$Ax \approx y. \tag{1.19}$$

Se sabe que aunque la inversa de A exista, es decir si se tiene un operador compacto que manda conjuntos acotados de X en conjuntos precompactos de Y, para X, Y espacios normados, ésta no puede ser continua a menos que los espacios $\mathscr{H}_1$ y $\mathscr{H}_2$ sean finito dimensionales, de manera que pequeños errores en y causan errores en el tamaño arbitrario de x.

La idea básica de los métodos de regularización es que en lugar de resolver de manera exacta la ecuación $Ax = y$, uno busca encontrar un problema parecido que tiene solución única y que es robusto en el sentido que pequeños errores en los datos no afectan en la solución, Kaipo y Somersalo (2004).

1.3.2. Teoría Estadística de Inversión Bayesiana

La esencia de los métodos de inversión estadística es el replanteamiento del problema como un problema de inferencia. Contamos de manera directa con cantidades observables y no observables. El objetivo de la teoría de inversión estadística es extraer información y evaluar la incertidumbre basada tanto en el conocimiento sobre el proceso de medición como en la información y modelos de variables desconocidas que están disponibles antes de la medición.

El enfoque de Inversión Bayesiano se basa en los siguientes principios:

- Las variables incluidas en el modelo se modelan como variables aleatorias.
- La aleatoriedad describe nuestro grado de información relativa a sus realizaciones.
- El grado de información relativa a estos valores se cuantifica con una distribución de probabilidad.

- La solución del problema inverso es la distribución de probabilidad a posteriori.

El último punto lo hace de interés particular, puesto que el enfoque estadístico Bayesiano es diferente al tradicional discutido en la literatura, usual de problemas inversos. Puede consultarse más en Vogel (2002).

Los métodos de regularización producen estimaciones puntuales de las incógnitas, mientras que los métodos bayesianos, producen una distribución de la cual se pueden obtener estimaciones.

En el enfoque tradicional de problemas inversos, típicamente escribimos el modelo de la forma:

$$y = f(x,e), \tag{1.20}$$

donde $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ es la función del modelo y $e \in \mathbb{R}^k$ es el vector que contiene los parámetros con *ruido*. En problemas inversos, los parámetros son vistos como variables aleatorias, de manera que la ecuación 1.20 se escribe como:

$$Y = f(X,E). \tag{1.21}$$

A este se le conoce como *forward mapping* exacto. Sin embargo en ocasiones se tiene una aproximación a la ecuación anterior por lo que en lugar de tener $Y = f(X,E)$ se tiene

$$Y \approx f(X,E). \tag{1.22}$$

Y a este se le conoce como *forward mapping* aproximado.

1.3.3. Ejemplos

En muchas situaciones, las cantidades que se desean determinar son diferentes a las que medimos, si los datos medidos dependen, de alguna manera, de las cantidades que queremos, entonces los datos contienen información sobre dichas cantidades. A partir de los datos que hemos medido, el problema inverso consiste en tratar de reconstruir dichas cantidades.

A continuación se muestran algunos ejemplos de los problemas inversos más frecuentes en la literatura:

Red Eléctrica.

Consideremos el problema de recuperar el valor para cada resistencia en una red cuadrada de resistores $(N \times N)$ a partir de mediciones con ruido realizadas en la frontera de la red. En Christen y Fox (2005) los autores modelan el voltaje y calculan el forward map exacto y aproximado.

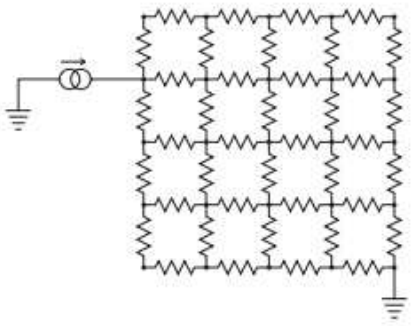

Figura 1.2: Red de resistores para el caso $N = 4$. Se puede observar el nodo de entrada y la dirección del flujo. Las mediciones son hechas en la frontera de dicha red.

Tomografia Axial Computarizada.

Dado un paciente, queremos obtener cortes transversales del cuerpo y obtener imágenes de dichos cortes. Es sabido que los rayos X cruzan de manera parcial el cuerpo y que la opacidad de diversas estructuras internas (huesos, órganos, etc.) varían, de modo que una imagen del coeficiente de variación en el cuerpo nos dará una imagen de dichas estructuras. Sin embargo, las únicas mediciones que uno puede realizar es medir la absorción total a lo largo de las líneas a través del cuerpo. Dada una colección de las integrales de línea (los datos), deseamos reconstruir la absorción en función de la posición en el cuerpo (la imagen).

Imágenes Radio Astronómicas.

Cuando los astrónomos hacen uso de un interferómetro[3] en un radiotelescopio, resulta que las mediciones no corresponde a la distribución de las fuentes en el cielo (llamada la función de brillo), sino de la transformada de Fourier del brillo del cielo. No es posible medir toda la transformada de Fourier, por lo que se toma una muestra de la transformada en un conjunto de curvas irregulares en el espacio de Fourier. De estos datos, la pregunta natural es ¿cómo es posible reconstruir la distribución deseada?

Es posible encontrar más ejemplos de problemas inversos en la literatura, estos solo constituyen una pequeña muestra de un sinfin de ejemplos, se sugiere consultar las notas del curso de ELEC 404: Imaging and Inference, que pueden consultarse en la siguiente dirección: http://elec.physics.otago.ac.nz/w/images/c/cb/ELEC404_chapter1.pdf, impartido por Colin Fox en la universidad de Otago en Nueva Zelanda.

[3]Instrumento que emplea la interferencia de las ondas de luz para medir con gran precisión longitudes de onda de la luz misma.

Capítulo 2

Gibbs Direccional Óptimo

En este segundo capítulo se implementa el Gibbs Direccional Óptimo. Primero comenzaremos construyendo una cadena de Markov mediante el algoritmo Gibbs. Posteriormente mostraremos cómo el Gibbs Direccional Óptimo generaliza esta idea muestreando a lo largo de una dirección, y utilizaremos la divergencia de Kullback-Leibler para encontrar mejores direcciones en el algoritmo de Gibbs que minimicen dicha divergencia. Finalmente, daremos los detalles de la implementación en Python de dicho algoritmo.

Este capítulo está basado en el criterio de optimalidad propuesto en el reporte de Christen y Fox (2011).

2.1. Divergencia Kullback-Leibler

Sea $\pi(\mathbf{x})$ la distribución objetivo de interés. Sea $\mathbf{X} \in \mathbb{R}^n$ una variable aleatoria con densidad $\pi(\mathbf{x})$. El Gibbs sampling es un algoritmo MCMC que simula a partir de la distribución condicional, como se mencionó en el Capítulo 1.

$$f_{X_i|\mathbf{X}_{-i}}(x_i|\mathbf{x}_{-i}) \propto \pi(\mathbf{x}), \tag{2.1}$$

donde la notación $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$ representa el vector $(n-1)$ dimensional creador a partir de eliminar la entrada i del n-ésimo vector v. La ecuación anterior, representa las distribuciones condicionales totales.

Usando 2.1, es posible crear una cadena de Markov $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \ldots$, considerando el siguiente kernel de transición

$$K_i\left(\mathbf{x}^{(t)},\mathbf{x}^{(t+1)}\right)=f_{X_i|\mathbf{X}_{-i}}\left(x_i^{(t+1)}|\mathbf{x}_{-i}^{(t)}\right)\mathbb{I}\left(\mathbf{x}_{-i}^{(t+1)}=\mathbf{x}_{-i}^{(t)}\right). \tag{2.2}$$

Donde $\mathbb{I}$ representa a la función indicadora. Esto es, el i-ésimo kernel cambia en la i-ésima coordenada simulando de las distribuciones condicionales totales $f_{X_i|\mathbf{X}_{-i}}\left(\cdot|x_{-i}^{(t)}\right)$. En el muestreador de Gibbs aleatorio, un kernel de transición completo se define como

$$K(\mathbf{x},\mathbf{y})=\sum_{i=1}^{n}\omega_i K_i(\mathbf{x},\mathbf{y}), \tag{2.3}$$

para $\omega_i \geq 0$ y $\sum_{i=1}^{n}\omega_i = 1$.

El Gibbs Sampling direccional generaliza esta idea escogiendo una dirección $\mathbf{e}\in\mathbb{R}^n$ con $\|\mathbf{e}\|=1$, donde $\|\mathbf{e}\|$ representa la norma euclideana de $\mathbf{e}=(e_1,\ldots,e_n)$ dada por $\|\mathbf{e}\|=\left(\sum_i^n |e_i|^2\right)^{1/2}$ y muestreando de la distribución condicional a lo largo de dicha dirección. Esto es

$$\mathbf{X}^{(t+1)}=\mathbf{x}^{(t)}+r\mathbf{e}, \tag{2.4}$$

donde $r\in\mathbb{R}$ y tiene distribución proporcional a $\pi(\mathbf{x}^{(t)}+r\mathbf{e})$. El kernel de transición satisface las ecuaciones de balance detallado con respecto a π. Asumiendo π-irreducibilidad la cadena de Markov tiene a π como distribución ergódica.

La pregunta natural hasta este punto es ¿cómo elegir $\mathbf{e}$ tal que optimice la convergencia de la Cadena de Markov? Una vez que se tenga irreducibilidad, cualquier cadena tendrá como distribución ergódica a π, pero el rendimiento dependerá de qué tan dependientes son $\mathbf{X}^{(t)}$ y $\mathbf{X}^{(t+1)}$. Una medida de dependencia es utilizar la información mutua entre X e Y $(I(X,Y))$, que mide la divergencia de Kullback-Leibler entre el modelo conjunto $f_{X,Y}$ y el alterno (independencia) $f_X f_Y$. Por las propiedades de la divergencia de Kullback-Leibler $I=0\Leftrightarrow \mathbf{X}^{(t+1)}\perp\mathbf{X}^{(t)}$, además es siempre positiva (puede probarse usando la desigualdad de Jensen) y no es simétrica (por lo que no se trata de una distancia).

Ahora, si nos fijamos en la información mutua entre $\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}$:

$$I\left(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}\right)=\int\int f_{\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}}\left(\mathbf{x}^{(t+1)},\mathbf{x}^{(t)}\right)\log\frac{f_{\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}}\left(\mathbf{x}^{(t+1)},\mathbf{x}^{(t)}\right)}{f_{\mathbf{X}^{(t+1)}}\left(\mathbf{x}^{(t+1)}\right)f_{\mathbf{X}^{(t)}}\left(\mathbf{x}^{(t)}\right)}d\mathbf{x}^{(t)}d\mathbf{x}^{(t+1)}. \tag{2.5}$$

Asumiendo que $\mathbf{X}^{(t)}\sim\pi$ observamos que

$$f_{\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}}\left(\mathbf{x}^{(t+1)},\mathbf{x}^{(t)}\right)=\pi\left(\mathbf{x}^{(t)}\right)K\left(\mathbf{x}^{(t)},\mathbf{x}^{(t+1)}\right), \tag{2.6}$$

y que $f_{x^{(t+1)}}\left(x^{(t+1)}\right)=\pi\left(x^{(t)}\right)$, de manera que la información mutua se puede escribir de la siguiente manera

$$I\left(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}\right)=\int\int\pi(\mathbf{x})K(\mathbf{x},\mathbf{y})\log\frac{\pi(\mathbf{x})K(\mathbf{x},\mathbf{y})}{\pi(\mathbf{x})\pi(\mathbf{y})}d\mathbf{y}d\mathbf{x} \tag{2.7}$$

$$=\int\int\pi(\mathbf{x})K(\mathbf{x},\mathbf{y})\log\frac{K(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})}d\mathbf{y}d\mathbf{x}. \tag{2.8}$$

Lo ideal será elegir las direcciones para las cuales $I\left(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}\right)$ se minimiza ya que $I(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)})\geq 0$ y $I=0\Leftrightarrow\mathbf{X}^{(t+1)}\perp\mathbf{X}^{(t)}$. Encontrar direcciones $\mathbf{e}$ para las cuales se minimice I nos proporcionará un criterio de optimización para encontrar mejores direcciones en Gibbs Sampling.

2.2. Caso Normal

Asumiremos π normal multivariada, y nos permitirá calcular $I(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)})$; consideremos la dimensión *baja*, de manera que tenemos la descomposición espectral de la matriz de precisión (inversa de la matriz de varianzas y covarianzas). $\pi\sim MNV(\mu,A)$. Dada una dirección $\mathbf{e}$, la cadena se mueve de $\mathbf{X}^{(t)}=\mathbf{x}$ a

$$\mathbf{Y}=\mathbf{X}^{(t+1)}=\mathbf{x}^{(t)}+r\mathbf{e}, \tag{2.9}$$

donde la longitud $r\in\mathbb{R}$ tiene distribución g proporcional a $\pi(\mathbf{x}^{(t)}+r\mathbf{e})$, esto es $g(r)\propto\exp\{-\frac{1}{2}(\mathbf{v}+r\mathbf{e})'A(\mathbf{v}+r\mathbf{e})\}$ con $\mathbf{v}=\mathbf{x}-\mu$.

Observamos que:

$$\begin{aligned} g(r) &\propto \exp\{-\frac{1}{2}(\mathbf{v}+r\mathbf{e})'A(\mathbf{v}+r\mathbf{e})\} \\ &= \exp\{-\frac{1}{2}(\mathbf{v}'+\mathbf{e}'r)A(\mathbf{v}+r\mathbf{e})\} \\ &\propto \exp\{-\frac{1}{2}(r^2\mathbf{e}'A\mathbf{e}+2r\mathbf{e}'A\mathbf{v})\} \\ &= \exp\{-\frac{1}{2}\mathbf{e}'A\mathbf{e}(r^2+2r\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})\} \\ &\propto \exp\{-\frac{1}{2}\mathbf{e}'A\mathbf{e}(r+\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2\}. \end{aligned}$$

Lo anterior corresponde al kernel de una distribución normal, y consecuentemente $r \sim N(-\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}, \mathbf{e}'A\mathbf{e})$ y $\mathbf{e}'A\mathbf{e}$ es la matriz de precisión. Haciendo $\mathbf{e} = \mathbf{e}_i$ el i-ésimo vector base estándar, se obtiene $Y_i \sim N(\mu_i - a_{ii}^{-1}(\mathbf{v}'\mathbf{v} - (x_i - \mu_i)^2), a_{ii})$, que corresponden a las distribuciones condicionales totales de una distribución Normal Multivariada, con entrada *i*, las cuales son precisamente las que usaremos para simular en el algoritmo de Gibbs tradicional.

A partir de lo anterior, el kernel de transición correspondiente a la dirección **e** se escribe como:

$$K_{\mathbf{e}}(\mathbf{x},\mathbf{y}) = \left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(\mathbf{e}'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2\right\} \mathbb{I}\left(\mathbf{y} = \mathbf{x} + \mathbf{e}'(\mathbf{y}-\mathbf{x})\,\mathbf{e}\right). \quad (2.10)$$

Note que $\mathbf{y}-\mathbf{x} = r\mathbf{e}$ y $\mathbf{e}\mathbf{e}' = 1$, $r = \mathbf{e}'(\mathbf{y}-\mathbf{x})$ por lo que $\mathbf{y}$ está restringida por la línea $\mathbf{y} = \mathbf{x} + \mathbf{e}'(\mathbf{y}-\mathbf{x})\mathbf{e}$, y de ahí la función indicadora en (2.10).

Calculamos la información mutua en la ecuación (2.8), dada la dirección **e**. Note que:

$$\begin{aligned} \frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})} &= \frac{\left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(\mathbf{e}'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2\right\}}{\frac{1}{(2\pi)^{n/2}|A|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)'A(\mathbf{y}-\mu)\right\}} \\ &= \left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{\frac{1}{2}} (2\pi)^{n/2}|A|^{1/2} \exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(\mathbf{e}'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2 + \frac{1}{2}(\mathbf{y}-\mu)'A(\mathbf{y}-\mu)\right\} \\ &= \left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{\frac{n}{2}-\frac{1}{2}} |A|^{1/2} \exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(e'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2 + \frac{1}{2}(\mathbf{y}-\mu)'A(\mathbf{y}-\mu)\right\} \end{aligned}$$

Tomando el logaritmo tenemos

$$\log\frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})} = \frac{n-1}{2}\log(2\pi) - \frac{1}{2}\log(|A|) + \frac{1}{2}(\mathbf{e}'A\mathbf{e}) - \frac{\mathbf{e}'A\mathbf{e}}{2}\left(\mathbf{e}'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2 + \frac{1}{2}(\mathbf{y}-\mu)'A(\mathbf{y}-\mu).$$

Haciendo $C_1 = \frac{n-1}{2}\log(2\pi) - \frac{1}{2}\log(|A|)$, $Q_1(\mathbf{e},\mathbf{x},\mathbf{y}) = \mathbf{e}'A\mathbf{e}(\mathbf{e}'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2$ y $Q_2(\mathbf{y}) = (\mathbf{y}-\mu)'A(\mathbf{y}-\mu)$, la información mutua se escribe como:

$$I_{\mathbf{e}}(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}) = C_1 + \frac{1}{2}\log(\mathbf{e}'A\mathbf{e}) - \frac{1}{2}(Q_1(\mathbf{e},\mathbf{x},\mathbf{y}) - Q_2(\mathbf{y})). \tag{2.11}$$

De lo anterior

$$\begin{aligned}\int \log\frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})}K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y} &= \int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})C_1 + \int \frac{1}{2}K_{\mathbf{e}}(\mathbf{x},\mathbf{y})\log(\mathbf{e}'A\mathbf{e}) \\ &\quad - \int \frac{1}{2}K_{\mathbf{e}}(\mathbf{x},\mathbf{y})(Q_1(\mathbf{e},\mathbf{x},\mathbf{y})) + \int \frac{1}{2}K_{\mathbf{e}}(\mathbf{x},\mathbf{y})Q_2(\mathbf{y})) \\ &= C_1 + \frac{1}{2}\log(\mathbf{e}'A\mathbf{e}) - \frac{1}{2} + \frac{1}{2}\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}.\end{aligned}$$

Puesto que $\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y} = 1$ y $\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})(Q_1(\mathbf{e},\mathbf{x},\mathbf{y})) = 1$, esto se observa regresando a la transformación en r

$$\begin{aligned}\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})(Q_1(\mathbf{e},\mathbf{x},\mathbf{y})) &= \int K_e(\mathbf{x},\mathbf{y})\mathbf{e}'A\mathbf{e}(\mathbf{e}'(\mathbf{y}-\mathbf{x}) + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2 \\ &= \int \mathbf{e}'A\mathbf{e}(r + \frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2 g_{\mathbf{e}}(r)dr \\ &= \int \mathbf{e}'A\mathbf{e}(r^2 + 2r\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}} + \left(\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2)g_{\mathbf{e}}(r)dr \\ &= (\mathbf{e}'A\mathbf{e})\int r^2 g_{\mathbf{e}}(r)dr + 2(\mathbf{e}'A\mathbf{v})\int r g_{\mathbf{e}}(r)dr + (\mathbf{e}'A\mathbf{e})\left(\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2\int g_{\mathbf{e}}(r)dr \\ &= (\mathbf{e}'A\mathbf{e})(\frac{1}{\mathbf{e}'A\mathbf{e}} + (\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2) - 2(\mathbf{e}'A\mathbf{v})(\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}) + (\mathbf{e}'A\mathbf{e})(\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2 \\ &= 1 + \frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}} - 2\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}} + \frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}} \\ &= 1 + 2\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}} - 2\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}} \\ &= 1\end{aligned}$$

por lo que $\int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})(Q_1(\mathbf{e},\mathbf{x},\mathbf{y}))=1$ y (2.12) se reescribe como:

$$\int \log\frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})}K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}=C_1+\frac{1}{2}\log(\mathbf{e}'A\mathbf{e})-\frac{1}{2}+\frac{1}{2}\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}.$$

Sin embargo, $\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}$ se calcula regresando a la transformación en r ya que $\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}=\int(r\mathbf{e}-\mathbf{v})'A(r\mathbf{e}-\mathbf{v})g_{\mathbf{e}}(r)dr$, donde

$$\begin{aligned}g_{\mathbf{e}}(r)&=\left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}(\mathbf{e}'(\mathbf{y}-\mathbf{x})+\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2\right\}\\&=\left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}(r+\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}})^2\right\}.\end{aligned}$$

Y calculamos

$$\begin{aligned}\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}&=\int(r\mathbf{e}-\mathbf{v})'A(r\mathbf{e}-\mathbf{v})g_{\mathbf{e}}(r)dr\\&=\int(\mathbf{e}'r-\mathbf{v}')A(r\mathbf{e}-\mathbf{v})g_{\mathbf{e}}(r)dr\\&=\int(r^2\mathbf{e}'A\mathbf{e}-2(\mathbf{v}'A\mathbf{e})r+\mathbf{v}'A\mathbf{v})g_{\mathbf{e}}(r)dr\\&=\mathbf{e}'A\mathbf{e}\int r^2g_{\mathbf{e}}(r)dr-2(\mathbf{v}'A\mathbf{e})\int rg_{\mathbf{e}}(r)dr+(\mathbf{v}'A\mathbf{v})\int g_{\mathbf{e}}(r)dr\\&=\mathbf{e}'A\mathbf{e}\left\{\frac{1}{\mathbf{e}'A\mathbf{e}}+\frac{(\mathbf{v}'A\mathbf{e})\mathbf{e}'A\mathbf{v}}{(\mathbf{e}'A\mathbf{e})^2}\right\}-2\frac{(\mathbf{v}'A\mathbf{e})\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}+\mathbf{v}'A\mathbf{v}\\&=1+\frac{(\mathbf{e}'A\mathbf{v})^2}{(\mathbf{e}'A\mathbf{e})}-2\frac{(\mathbf{v}'A\mathbf{e})\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}+\mathbf{v}'A\mathbf{v}\\&=1-\frac{\mathbf{v}'A\mathbf{e}\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}+\mathbf{v}'A\mathbf{v}.\end{aligned}$$

Finalmente, observamos

$$\int Q_2(\mathbf{y})K_{\mathbf{e}}(\mathbf{x},\mathbf{y})d\mathbf{y}=1-\frac{\mathbf{v}'A\mathbf{e}\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}+\mathbf{v}'A\mathbf{v}, \tag{2.12}$$

por lo que

$$\int \log \frac{K_{\mathbf{e}}(\mathbf{x},\mathbf{y})}{\pi(\mathbf{y})} K_{\mathbf{e}}(\mathbf{x},\mathbf{y}) d\mathbf{y} = C + \frac{1}{2}\log(\mathbf{e}'A\mathbf{e}) - \frac{1}{2}\frac{\mathbf{v}'A\mathbf{e}\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}} + \frac{1}{2}\mathbf{v}'A\mathbf{v}. \tag{2.13}$$

Ahora es necesario integrar con respecto a $\pi(d\mathbf{x})$. Observamos

$$\begin{aligned}
\int \mathbf{v}'A\mathbf{v}\pi(\mathbf{x})d\mathbf{x} &= \int \left\{(\mathbf{x}-\mu)'A(\mathbf{x}-\mu)\right\}\pi(\mathbf{x})d\mathbf{x} \\
&= \int \left\{(\mathbf{x}'-\mu')A(\mathbf{x}-\mu)\right\}\pi(\mathbf{x})d\mathbf{x} \\
&= \int \left\{(\mathbf{x}'A-\mu'A)(\mathbf{x}-\mu)\right\}\pi(\mathbf{x})d\mathbf{x} \\
&= \int \left\{(\mathbf{x}'A\mathbf{x}-2\mu'A\mathbf{x}+\mu'A\mu)\right\}\pi(\mathbf{x})d\mathbf{x} \\
&= \int \left\{(\mathbf{x}'A\mathbf{x}-2\mu'A\mathbf{x}+\mu'A\mu)\right\}\pi(\mathbf{x})d\mathbf{x} \\
&= \int (\mathbf{x}'A\mathbf{x})\pi(\mathbf{x})d\mathbf{x} - 2\mu'A\int \mathbf{x}\pi(\mathbf{x})d\mathbf{x} + \mu'A\mu \\
&= tr(AA^{-1}) + \mu'A\mu - 2\mu'A\mu + \mu'A\mu \\
&= tr(I_n) \\
&= n.
\end{aligned}$$

Más aún, el valor esperado de una forma cuadrática es $E(z'Rz) = tr(R\Sigma) + \mu'A\mu$ donde μ y Σ es el vector de medias y matriz de varianzas y covarianzas de la variable aleatoria z, respectivamente.

Haciendo $R = \frac{A\mathbf{e}\mathbf{e}'A}{\mathbf{e}'A\mathbf{e}}$ y puesto que el valor de $E(\mathbf{v}) = 0$, obtenemos

$$\begin{aligned}
E\left(\mathbf{v}'\frac{A\mathbf{e}\mathbf{e}'A}{\mathbf{e}'Ae}\mathbf{v}\right) &= \frac{1}{\mathbf{e}'A\mathbf{e}} tr\left(A\mathbf{e}\mathbf{e}'AA^{-1}\right) \\
&= \frac{1}{\mathbf{e}'A\mathbf{e}} tr\left(A\mathbf{e}\mathbf{e}'\right) \\
&= \frac{1}{\mathbf{e}'A\mathbf{e}} tr\left(\mathbf{e}A\mathbf{e}'\right) \\
&= 1.
\end{aligned}$$

De manera que:

$$\begin{aligned}
I_{\mathbf{e}}(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}) = I_{\mathbf{e}} &= C_1 + n - \frac{1}{2} + \frac{1}{2}\log(\mathbf{e}'A\mathbf{e}) \\
&= C_2 + \frac{1}{2}\log(\mathbf{e}'A\mathbf{e}).
\end{aligned}$$

Con $C_2 = C_1 + \frac{1}{2}(2n-1)$. Note que la información mutua no depende de $X^{(t)}$ (ni, desde luego, de $X^{(t+1)}$), sino sólo de la dirección $\mathbf{e}$.

2.2.1. Eligiendo un conjunto de direcciones

Necesitamos una distribución para h que genere un Gibbs Sampling irreducible. De acuerdo a lo anterior, la mejor dirección es la que minimiza $I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}\right)$. Sin embargo, no podemos únicamente elegir dicha dirección, porque la cadena resultante no será irreducible y no estaremos muestreando de π (circularíamos sólo en una recta dentro del espacio de estados). Si las direcciones tienen distribución $h(\mathbf{e})$, y ésta tiene como soporte a la n- esfera, $\mathbb{S}^n$, la cadena resultante

$$K(\mathbf{x},\mathbf{y}) = \int K_{\mathbf{e}}(\mathbf{x},\mathbf{y})h(\mathbf{e})d\mathbf{e}, \tag{2.14}$$

es irreducible. Kaufman y Smith (1994) argumentan que la distribución de la dirección óptima, en un sentido de convergencia geométrica es

$$h(\mathbf{e}) \propto \sup_{x\in\mathscr{X}, r\in\mathbb{R}} \left\{ \int \pi(\mathbf{x}+r\mathbf{e})dr \frac{|r|^{n-1}}{\pi(\mathbf{x}+r\mathbf{e})} \right\}. \tag{2.15}$$

Sin embargo, esto sólo se satisface para un soporte acotado $\mathscr{X}$ de π. No podemos controlar el término $\frac{|r|^{n-1}}{\pi(\mathbf{x}+r\mathbf{e})}$ para soportes no acotados. Por lo tanto, esto sugiere utilizar una dirección

$$h^*(e) \propto \sup_{x\in\mathscr{R}^n, r\in\mathbb{R}} \left\{ \int \pi(\mathbf{x}+r\mathbf{e})dr \right\}, \tag{2.16}$$

Primero calculamos la $\int \pi(\mathbf{x}+r\mathbf{e})dr$, de lo cual se observa

$$\begin{aligned} \int \pi(\mathbf{x}+r\mathbf{e})dr &= \int \left(\frac{|A|}{(2\pi)^n}\right)^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}+r\mathbf{e}-\mu)'A(\mathbf{x}+r\mathbf{e}-\mu)\right\} dr \\ &= \left(\frac{|A|}{(2\pi)^n}\right)^{1/2} \int \exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left[r^2 + 2r\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}} + \frac{\mathbf{v}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right]\right\} dr \end{aligned} \tag{2.17}$$

Haciendo $k = \exp\left\{-\frac{\mathbf{v}^T A\mathbf{e}}{2}\right\}$. Y multiplicando y diviendo por k para completar los cuadrados se obtiene

$$\begin{aligned}
\int \pi(\mathbf{x}+r\mathbf{e})dr &= \left(\frac{|A|}{(2\pi)^n}\right)^{1/2} \int k\exp\left\{\frac{1}{2}\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}}\right\}\exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(r+\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2\right\}dr \\
&= k\exp\left\{\frac{1}{2}\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}}\right\}\left(\frac{|A|}{(2\pi)^n}\right)^{1/2}\int \exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(r+\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2\right\}dr \\
&= k\exp\left\{\frac{1}{2}\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}}\right\}\left(\frac{2\pi}{\mathbf{e}'A\mathbf{e}}\right)^{1/2}\left(\frac{|A|}{(2\pi)^n}\right)^{1/2} \\
&\int \left(\frac{\mathbf{e}'A\mathbf{e}}{2\pi}\right)^{1/2}\exp\left\{-\frac{\mathbf{e}'A\mathbf{e}}{2}\left(r+\frac{\mathbf{e}'A\mathbf{v}}{\mathbf{e}'A\mathbf{e}}\right)^2\right\}dr \\
&= k\exp\left\{\frac{1}{2}\frac{\left(\mathbf{e}^TA\mathbf{v}\right)^2}{\mathbf{e}^TA\mathbf{e}}\right\}\left(\frac{2\pi}{\mathbf{e}^TA\mathbf{e}}\right)^{1/2}\left(\frac{|A|}{(2\pi)^n}\right)^{1/2} \\
&\leq \exp\left\{\frac{1}{2}\frac{(\mathbf{e}'A\mathbf{v})^2}{\mathbf{e}'A\mathbf{e}}\right\}\frac{(2\pi)^{-(n-1)/2}|A|^{1/2}}{\sqrt{\mathbf{e}'A\mathbf{e}}},
\end{aligned}$$

El valor de $k \in (0,1]$ y al eliminarlo de la expresión original, obtenemos la desigualdad. De acuerdo a Christen y Fox (2010), lo anterior sugiere tomar

$$h^*(\mathbf{e}) \propto \left(\mathbf{e}'A\mathbf{e}\right)^{-1/2}. \tag{2.18}$$

Se puede minimizar $I_{\mathbf{e}}\left(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)}\right)$ maximizando $\exp\{-I_{\mathbf{e}}(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)})\}$, por lo que de elegir $h^*(\mathbf{e}) \propto (\mathbf{e}'\mathbf{A}\mathbf{e})^{-1/2}$ elegirá direcciones con $I_{\mathbf{e}}(\mathbf{X}^{(t+1)},\mathbf{X}^{(t)})$ mínima.

Para muestrear de $h^*(\mathbf{e})$ basta observar que $h^*(\mathbf{e}) \propto \int \pi(\mu+r\mathbf{e})d\tau \propto (\mathbf{e}'\mathbf{A}\mathbf{e})^{-1/2}$ y el máximo se alcanza cuando $\mathbf{x}=\mu$. Si simulamos $\mathbf{e}_u$ de una $MNV(\mathbf{0},A) \sim \pi_o$ y tomamos $\mathbf{e} = \frac{\mathbf{e}_u}{||\mathbf{e}_u||}$, entonces $\mathbf{e}$ tiene densidad

$$\propto \int \pi_o(r\mathbf{e})dr \propto (\mathbf{e}'\mathbf{A}\mathbf{e})^{-1/2} \tag{2.19}$$

i.e., $\mathbf{e} \sim \mathbf{h}^*$.

2.3. Normal Truncada

Para simular de una distribución normal multivariada se necesita prácticamente muestrear de la misma normal multivariada.

Sea π una distribución Normal Mutivariada con matriz de precisión $\mathbf{A}$ y vector de medias $\mu = \left(\sqrt{\frac{1}{n}}, \ldots, \sqrt{\frac{1}{n}}\right)$ y soporte $x_i \geq 0$ todas las entradas son no negativas y la función objetivo es $\pi(x)$.

Consideremos un muestreo aleatorio de Gibbs con distribución de direcciones h^*, con un punto inicial $X^{(0)} = \mu$, no es necesario un *burn in time*. Esto debido a que el valor inicial se encuentra dentro del soporte de la función objetivo. Generalmente, el *burn in time* se refiere a tener que desechar algunas iteraciones antes de que la cadena alcance el equilibrio en un algoritmo de MCMC.

Consideremos primero el problema de simular de una distribución Normal Truncada univariada, y soporte $[a, \infty)$.

$$p(x) = \frac{\phi_{\mu,\sigma^2}(x)}{\Phi(-a)} \mathbb{I}_{x \geq 0} \tag{2.20}$$

donde $\phi(x)$ denota la función de densidad de la función normal con media μ y varianza σ^2. En nuestro caso $a = 0$. Un método conveniente para poder simular una v.a X de 2.20 es utilizando el método de la transformada inversa. Esto es

$$x = -\Phi^{-1}(\Phi(-a)u), \tag{2.21}$$

donde $u \sim U[0,1]$ es una variable uniforme. Haciendo un poco de álgebra, podemos observar que la fórmula anterior se puede escribir de la siguiente manera:

$$x = -\Phi^{-1}(\Phi(a) + (1 - \Phi(a))u). \tag{2.22}$$

Sin embargo, la expresión anterior resulta ser menos inestable numéricamente, es decir los errores debidos a las aproximaciones se atenúan a medida en que el cómputo procede. El método de la transformada inversa por sí mismo produce una simulación rápida, es por eso que se implementa dicho método de simulación. Puede consultarse a Glasserman (2003) para conocer más sobre el orden de convergencia de dicho algoritmo.

Para el caso multivariado, la generalización de la distribución normal multivariada puede darse de la siguiente forma:

Sea $\phi(x_{\mu,\Sigma})$ una densidad normal *n*-variada, con vector de medias μ y matriz de varianzas y covarianzas Σ. Esto es:

$$\phi(x,\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}. \tag{2.23}$$

Decimos que X tiene distribución normal truncada multivariada si

$$\phi(x,\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}\mathbb{I}_R(x), \tag{2.24}$$

donde $\mathbb{I}_R(\cdot)$ es la función indicadora de R. En el siguiente capítulo veremos la necesidad de simular de una normal truncada que surge de manera natural en problemas inversos lineales.

2.4. Implementación en Python

A principios de los años 90 Guido van Rossum creó Python un lenguaje de programación multiparadigma. Esto es, más que forzar a los programadores a adoptar un estilo particular de programación, permite varios estilos: programación orientada a objetos, programación estructurada, programación funcional y programación orientada a aspectos. Es un lenguaje interpretado - está diseñado para ser ejecutado por medio de un intérprete - y es multiplataforma.

El paradigma de programación orientado a objetos es un paradigma para la implementación, de una estructura que puede contener a la vez variables y métodos propios, y que dispone de ciertas propiedades como la herencia[1] y el polimorfismo. En Python los objetos se definen a partir de clases. Una clase crea un tipo de dato. Cuando se asigna una clase a una variable esta no contiene un objeto sino una instancia de la clase.

Pyhton es además una poderosa herramienta para llevar a cabo tareas de cómputo científico, tales como análisis de datos, graficación e implementación de algoritmos de MCMC, entre otros. La mayoría de los programas que involucran cálculo numérico son intensivos, y no podemos contar con un programa lento a la hora de realizar un bucle a través de cálculos intensos.

[1] Quiere decir que una clase es de un tipo o subtipo de otra clase.

Python evita este problema mediante el uso de dos librerías llamadas NumPy y Scipy. NumPy ofrece funciones eficientes para el almacenamiento y manipulación de matrices numéricas. Mientras que Scipy proporciona un mayor nivel de funcionalidad, como la integración numérica.

Python, junto con Scipy y NumPy, proporciona un sistema con capacidad numérica similar a Matlab. La librería Matplotlib permite al programador generar gráficos similares a los que se pueden hacer en Matlab. Puede consultarse la página oficial de Pyhton y su creador: http://www.python.org/∼guido/.

Capítulo 3

Problemas Inversos Lineales y Gibbs Direccional Óptimo

En este tercer capítulo veremos la motivación de utilizar la inferencia bayesiana en problemas inversos lineales, la motivación de trabajar en problemas inversos lineales surge a partir de un modelo de metabolismo del hígado en un estado estacionario. Para ello nos basaremos en el artículo de Calvetti, Kuceyesky y Somersalo (2008), donde se plantea un análisis de dicho modelo utilizando algoritmos de MCMC.

Finalmente, resolveremos este problema que surge en análisis de problemas inversos lineales, mediante un enfoque bayesiano. Una vez resuelto el problema, consideramos la simulación de una distribución normal truncada, donde simulamos mediante Gibbs Direccional Óptimo, con direcciones simuladas de h^*. Daremos una comparación en el caso de varias dimensiones $n = 2, 3, 5, 10, 15, 20$, y en cada caso compararemos con el algoritmo de Gibbs Sampling, Random Scan Gibbs y direcciones de h^*.

3.1. Modelo para el metabolismo hepático en estado estacionario

A fin de que los modelos matemáticos del metabolismo del hígado puedan utilizarse como experimentos *in vivo*, es necesario que capturen ciertas características esenciales del sistema biológico, para que sean computacionalmente implementables. Dichas características dependen del sistema biológico que se este modelando. El desafío computacional de los modelos

espacialmente distribuidos[1] es que involucran un número muy grande de parámetros, por ello, se requiere de técnicas de cómputo intensivo.

El hígado, como órgano regulador en funciones importantes de metabolismo, juega un rol central en todo el cuerpo. Su capacidad para utilizar o producir glucosa de acuerdo a las necesidades lo distinguen de otros órganos como el corazón o el cerebro. El interés de entender el metabolismo hepático se debe al aumento reciente de casos de diabetes tipo II[2] en el mundo, relacionada a los cambios alimenticios de la población.

El enfoque que presentan Calvetti y Somersalo (2009) es el llamado análisis bayesiano de flujo de balance (BFBA), el cual estima parámetros tomando en cuenta su dependencia mutua. Dicho análisis estima los parámetros desconocidos modelándolos como variables aleatorias, donde la aleatoriedad expresa nuestra incertidumbre sobre los valores.

Consideremos tener que estimar un vector x de parámetros desconocidos. De manera que el modelo lineal, para el modelo estacionario, se puede expresar de manera compacta en:

$$Ax = 0 \tag{3.1}$$

donde A es conocida. Debido a la indeterminación del sistema, el espacio nulo de A no es trivial. En lugar de forzar la solución a x en estar en el espacio nulo, asumimos que las ecuaciones del modelo son conocidas y remplazamos (3.1) por

$$Ax = y \quad \text{con} \quad x \geq 0. \tag{3.2}$$

donde y es un vector aleatorio que modela las observaciones con error.

3.2. Posterior Normal Truncada

Partiendo de (3.2), ahora supongamos que $A_{mxn}x_{nx1} = y_{mx1}$, puede ser visto como un problema lineal inverso, donde A es una matriz de dimensión $m \times n$ conocida.

[1] Si el modelo tiene en cuenta de forma explícita el espacio, estamos ante un modelo espacialmente distribuido.

[2] La diabetes tipo II es una enfermedad metabólica caracterizada por altos niveles de glucosa en la sangre, debido a una resistencia celular a las acciones de la insulina, combinada con una deficiente secreción de insulina por el páncreas.

Supóngase que $y|x \sim MVN(Ax,\tau)$ donde $\tau = diag(\frac{1}{\sigma_1^2},\ldots,\frac{1}{\sigma_m^2})$ es conocida, y suponga a priori $x \sim MVNT(\mu,B)$ (normal multivariada truncada a $x_i > 0$) , donde μ y B al igual que A son conocidas. Lo que necesitamos es calcular la distribución posterior de x.

Por el teorema de Bayes, la distribución posterior de $x|y$ es proporcional al producto de la verosimilitud por la distribución a priori,

$$f(x|y) \propto l(y|x)\, f(x).$$

Para una muestra de tamaño m, la verosimilitud se escribe como

$$l(y|x) \propto \exp\left\{ -\frac{1}{2}\sum_i^m (y_i - Ax)'\tau(y_i - Ax) \right\}$$

y la distribución a priori de x satisface

$$f(x) \propto \exp\left\{ -\frac{1}{2}(x-\mu)'B^{-1}(x-\mu) \right\}$$

Calculando la distribución posterior de $x|y$,

$$f(x|y) \propto \exp\left\{ -\frac{1}{2}\sum_i^m (y_i - Ax)'\tau(y_i - Ax) \right\} \exp\left\{ -\frac{1}{2}(x-\mu)'B^{-1}(x-\mu) \right\} \tag{3.3}$$

$$\propto \exp\left\{ -\frac{1}{2}(\sum_i^m (y_i - Ax)'\tau(y_i - Ax) + (x-\mu)'B^{-1}(x-\mu)) \right\} \tag{3.4}$$

$$\propto \exp\left\{ -\frac{1}{2}(tr(\tau\sum_i^m (y_i - Ax)(y_i - Ax)') + (x-\mu)'B^{-1}(x-\mu)) \right\} \tag{3.5}$$

Observese que el término $\sum_i^m (y_i - Ax)(y_i - Ax)'$ se escribe como:

$$\begin{aligned}\sum_i^m (y_i - Ax)(y_i - Ax)' &= \sum_i^m (y_i - \bar{y} + \bar{y} - Ax)(y_i - \bar{y} + \bar{y} - Ax)' \\ &= \sum_i^m (y_i - \bar{y})(y_i - \bar{y})' + m(\bar{y} - Ax)(\bar{y} - Ax)'.\end{aligned}$$

Sustituyendo en (3.5), obtenemos

$$
\begin{aligned}
f(x|y) &\propto \exp\left\{-\frac{1}{2}(tr(\tau(\sum_{i}^{m}(y_i-\bar{y})(y_i-\bar{y})'+m(\bar{y}-Ax)(\bar{y}-Ax)'))+(x-\mu)'B^{-1}(x-\mu))\right\}\mathbb{I}(x_i>0)\\
&\propto \exp\left\{-\frac{1}{2}(tr(\tau m(\bar{y}-Ax)(\bar{y}-Ax)')+(x-\mu)'B^{-1}(x-\mu))\right\}\mathbb{I}(x_i>0)\\
&\propto \exp\left\{-\frac{1}{2}(\bar{y}-Ax)'(\tau m)(\bar{y}-Ax)+(x-\mu)'B^{-1}(x-\mu)\right\}\mathbb{I}(x_i>0)\\
&\propto \exp\left\{-\frac{1}{2}(\vec{y}'-x'A')(\tau m)(\bar{y}-Ax)+(x'-\mu')B^{-1}(x-\mu)\right\}\mathbb{I}(x_i>0)\\
&\propto \exp\left\{-\frac{1}{2}(x'(B^{-1}+A'(m\tau)A)x-2(\mu'B^{-1}+\vec{y}'(m\tau)A)x+\vec{y}'(m\tau)\bar{y}+\mu'B^{-1}\mu\right\}\mathbb{I}(x_i>0)\\
&\propto \exp\left\{-\frac{1}{2}((x-\mu^*)'(B^{-1}+A'(m\tau)A)(x-\mu^*))\right\}\mathbb{I}(x_i>0),
\end{aligned}
$$

el cual corresponde al kernel de una distribución normal multivariada truncada con matriz de varianzas y covarianzas Σ, dada de la siguiente forma

$$\Sigma^{-1}=(B^{-1}+A'(m\tau)A) \tag{3.6}$$

y vector de medias μ^* dado por

$$\mu^*=(B^{-1}+A'(m\tau)A)^{-1}(\mu'B^{-1}+\vec{y}'(m\tau)A) \tag{3.7}$$

3.3. Direcciones Óptimas y Experimentos

Consideramos diferentes dimensiones $n=2,3,5,10,15,20$, en cada caso se compara cada una con el Gibbs sistemático, con una dirección aleatoria de Gibbs (Gibbs Aleatorio) y con direcciones simuladas del Gibbs Direccional Óptimo.

Usando la descomposición QR de una matriz $n\times n$ con entradas uniformes, uno obtiene una matriz ortogonal aleatoria **P** que representa una base ortogonal de eigenvalores. Los eigenvectores de **A**, $\lambda_1,\ldots,\lambda_n$ representan la precisión en cada dirección del eigenvector y son $\lambda_i=\sigma_i^{-2}$.

La desviación estandar en cada dirección principal es

$$\sigma_i=i^{-\frac{\alpha}{n}} \tag{3.8}$$

con $\alpha \geq 0$ y $\mathbf{A} = \mathbf{P}'\Lambda\mathbf{P}$ donde $\Lambda = diag(\lambda_i)$. Éstas representan desviaciones estandard decrecientes e incrementan inversamente conforme los α's se incrementan. Consideremos valores de $\alpha = 0, 5, 10, 20$ que hacen las desviaciones estándares progresivamente más contrastantes. Ver figura 3.2.

Para cada combinación de n y α calculamos el IAT (ver apéndice 1) de la cadena resultante, que nos da un índice de la eficiencia de la cadena simulada.

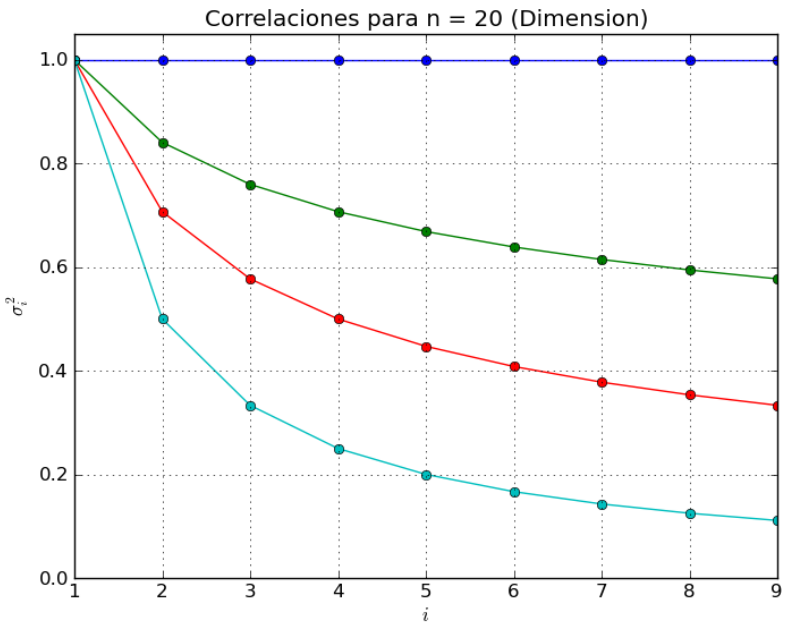

Figura 3.1: Desviaciones estándares progresivamente más contrastantes.

La gráfica (a) muestra el IAT (Integrated Autocorrelation Time) dividido entre la dimensión de la distribución normal multivariada de interés. La línea en azul corresponde al IAT para el Gibbs Sistemático, la línea en verde corresponde al Gibbs Aleatorio y la linea en rojo corresponde al Gibbs Direccional Óptimo.

Las desviaciones estandar para la distribución normal multivariada truncada se eligen con $\alpha = 0$ es decir, normales truncadas independientes. De la gráfica anterior se observa que el IAT dividido entre la dimensión para direcciones de acuerdo a h^* parece comportarse igual que Gibbs tradicional. Mientras que el IAT para el Gibbs Direccional Aleatorio se comporta

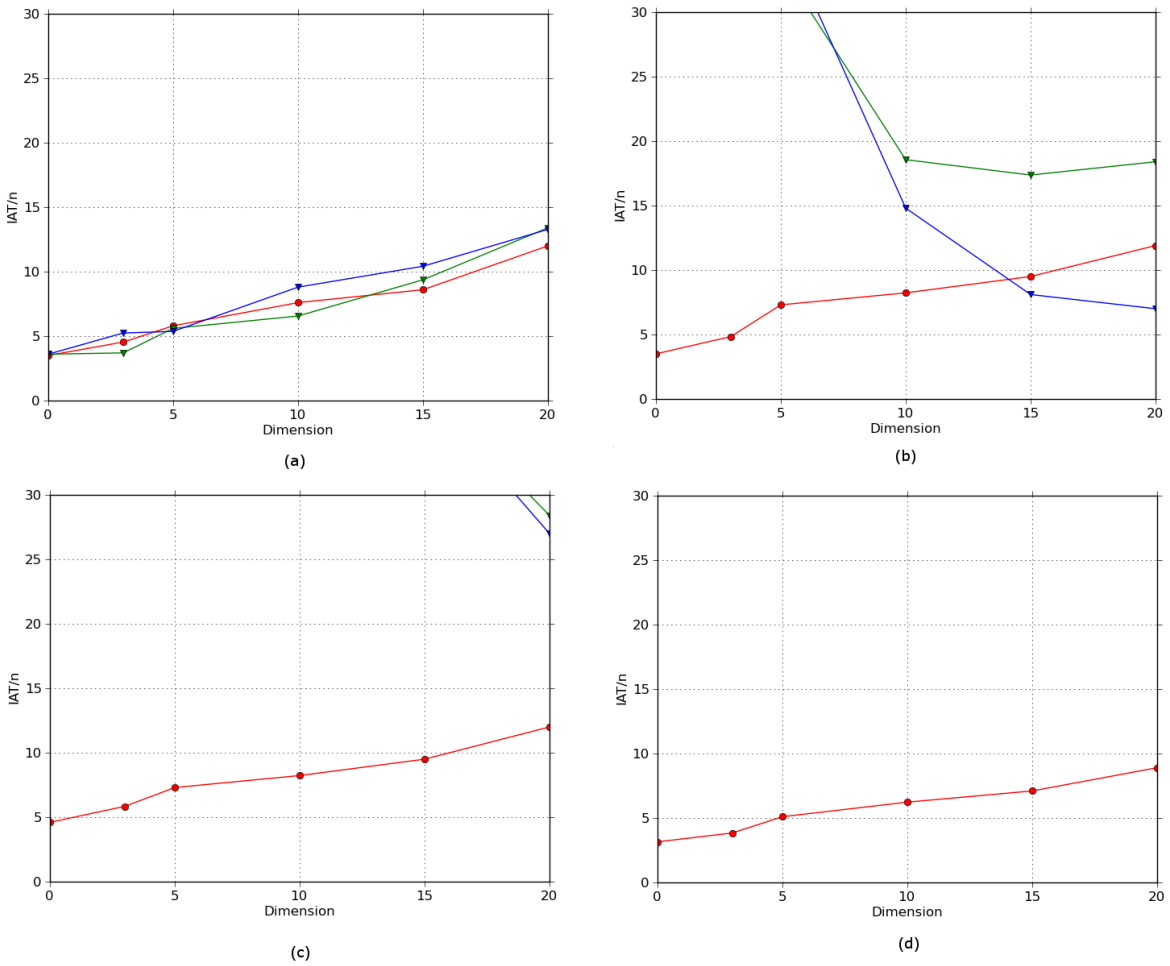

Figura 3.2: IAT dividido entre la dimensión de la distribución normal multivariada truncada objetivo, las desviaciones para la MN se eligen de acuerdo (a) $\alpha = 0$, (b) $\alpha = 5$, (c) $\alpha = 10$ y (d) $\alpha = 20$ en (3.16).

mejor que estos dos últimos.

La gráfica (b) muestra el IAT dividido entre la dimensión de la distribución normal multivariada truncada, en este caso las desviaciones estandar para la distribución normal multivariada se eligen con $\alpha = 10$ es decir,

$$\sigma_i = i^{-\frac{10}{n}}. \tag{3.9}$$

Es posible observar que en la gráfica (b) que el IAT/n para direcciones de h^* se comporta mejor que para Gibbs Aleatorio y el Gibbs Sistemático. Sin embargo, al principio el IAT para el Gibbs Aleatorio y el Gibbs Sistemático tienden a ser muy grande, y decrece conforme la dimensión de la función objetivo de interés (normal truncada) aumenta, llegando en última instancia, a ser el IAT del Gibbs Aleatorio a ser menor que el IAT del Gibbs Direccional Óptimo. Durante toda la gráfica se muestra que el IAT del Gibbs Direccional Óptimo permanece con $IAT/n \leq 12$, contrario con lo que sucede con el IAT/n del Gibbs Aleatorio y el Gibbs Sistemático.

La gráfica (c) muestra el IAT dividido entre la dimensión de la distribución de interés. Las desviaciones estandar para la distribución normal multivariada truncada se eligen para esta gráfica con $\alpha = 20$ es decir,

$$\sigma_i = i^{-\frac{20}{n}} \tag{3.10}$$

Se observa que el IAT/n para direcciones de h^* presenta un mejor comportamiento que el Gibbs Aleatorio y el Gibbs Sistemático. Éstos dos últimos presentan problemas cuando la dimensión de la función objetivo es alta, en este caso la dimensión es mayor al orden de 20.

La gráfica (d) muestra el IAT dividido entre la dimensión de la distribución normal multivariada truncada. Observamos que el IAT/n para direcciones de h^* presenta un mejor comportamiento que el Gibbs Aleatorio y el Gibbs Sistemático, al igual que la gráfica anterior.

De las gráficas anteriores es posible observar que el Gibbs Aleatorio y Gibbs Sistemático tienen un peor desempeño conforme los α's incrementan, mientras que el Gibbs Direccional Óptimo satisface que $IAT/n \leq 12$.

Observese que para este caso, el Gibbs Direccional Óptimo se puede ver como una caminata aleatoria, los IAT resultantes se acercan al óptimo teórico para una caminata aleatoria escalada óptima como se explica en Roberts y Rosenthal(2001).

Discusión y Conclusiones

Las técnicas que comprende MCMC son técnicas generales; su uso ha revolucionado por completo a la estadística bayesiana. Actualmente en esta área se manejan modelos muy complejos en donde las distribuciones no son convencionales y es necesario utilizar métodos eficientes para simular de ellos.

Por ello existe la necesidad de desarrollar técnicas eficientes para simular de dichas distribuciones. El Gibbs generalizado toma una dirección arbitraria en el espacio sobre el cual se simula y tratamos de responder a la pregunta de qué distribución de direcciones sería la óptima.

Hemos observado que la necesidad de simular de la distribución normal truncada es frecuente en la inferencia bayesiana, y es común encontrar este tipo de distribuciones en el área de problemas inversos.

Una vez realizada la experimentación, que se explica en el capítulo 3, y consideramos diferentes dimensiones $n = 2, 3, 5, 10, 15, 20$, en cada caso se compara cada una con el Gibbs Aleatorio, el Gibbs Sistemático y con direcciones simuladas del Gibbs Direccional Óptimo.

De los resultado obtenidos es posible observar que el Gibbs Aleatorio y el Gibbs Sistemático tienen un peor desempeño conforme la distribución normal truncada se vuelve más correlacionada y satisface que $IAT/n \leq 12$.

Observe que para este caso, el Gibbs Direccional Óptimo se puede ver como una caminata aleatoria, los IAT resultantes se acercan al óptimo teórico para una caminata aleatoria escalada óptima como lo explican en Roberts y Rosenthal(2001).

En cuanto a la implementación del algoritmo de simulación de una variable normal truncada

univariada, como se mencionó en el trabajo corresponde a un método de la transformada inversa, como trabajo a futuro se puede implementar un método de simulación mediante un algoritmo de aceptación y rechazo, y bien se pueden combinar dichos métodos para su optimización. Como se mencionó durante el trabajo, estos programas fueron implementados en Python, el cual representa una excelente opción para el cómputo estadístico intensivo.

Anexo 1

En el presente anexo se definirán y explicarán de manera muy breve el Integrated Autocorrelation Time (IAT).

IAT

Sea $X_0 = x^{(0)}, X_1 = x^{(1)}, \ldots, X_N = x^{(N)}$ una realización de una cadena de Markov homogénea, como resultado de algún algoritmo de MCMC que muestrea de una distribución objetivo $\pi(x)$. Recordemos que la varianza de una cantidad es una medida de variabilidad alrededor de su media. De manera que si $x \sim \pi(x)$ y $\mu_f = E[f(X)]$ entonces,

$$Var(f) = E[f(X)^2] - \mu_f^2$$

la cual mide la varianza de f en muestras x distribuidas como $\pi(x)$. Considere la cantidad $\hat{f}_N$ (un estimador para $E[f(X)]$) y estamos interesados en conocer la variabilidad de nuestro estimador, esto es:

$$Var(\hat{f}_N)$$

Es sabido que para N suficientemente grande (por el Teorema del Límite Central),

$$\hat{f}_N \approx N(E(f), Var(\hat{f}_N)). \tag{3.11}$$

Usualmente la ecuación anterior se escribe como:

$$\lim_{N\to\infty} \sqrt{N}(\hat{f}_N - E[f]) \xrightarrow{D} N(0, C)$$

con C una constante positiva independiente de N. Si la cadena de Markov $(X^{(t)})$ fuera producida por muestras totalmente independientes, y $\hat{f}_N = \frac{1}{N}\sum_{n=1}^{N} f(x^{(n)})$ entonces la varianza del estimador fuera

$$Var(\hat{f}_N) = \frac{Var(f)}{N}. \tag{3.12}$$

Sin embargo, por la naturaleza de la simulación la cadena está correlacionada. Y veremos que

$$Var(\hat{f}_N) = \frac{\tau_f Var(f)}{N}, \tag{3.13}$$

donde τ_f representa un número utilizado para generar una cadena de Markov. De la ecuación 3.12 basta observar que la varianza decrece conforme a $1/N$ donde N es el número de muestras independientes. Mientras que para la ecuación 3.13 la varianza decrece conforme a τ_f/N. De manera que τ_f representa el número de muestras correlacionadas que tiene el mismo efecto en la reducción de la varianza que una muestra independiente.

La covarianza $(cov(f,g))$ de dos cantidades f y g es una medida de su correlación. Consideremos $\{X^{(n)}\}_{n=0}^{N-1}$ una secuencia de N variables aleatorias de una cadena de Markov con distribución ergódica π, y escribimos la función de covarianza (con retraso s) en términos de $f(X)$ de la siguiente manera:

$$\begin{aligned} C_{ff}(s) &= cov(f(X_n), f(X_{n+s})) \\ &= E(f(X_n)f(X_{n+s})) - \mu_f^2 \end{aligned}$$

puesto que la cadena es homogénea y tiene distribución estacionaria, $C_{ff}(s)$ depende sólo de s y no de n. Definimos una función de autocovarianza normalizada de la siguiente manera:

$$\begin{aligned} \rho_{ff}(s) &= C_{ff}(s)/C_{ff}(0) \\ &= C_{ff}(s)/Var(f) \end{aligned}$$

observamos que $\rho_{ff}(0) = 1$ puesto que la cadena está correlacionada consigo mismo, esperamos que $\rho_{ff}(s) \to 0$ conforme $s \to \infty$. Asumiremos que para una secuencia de v.a suficientemente grande $\rho_{ff}(s) \approx 0$ cuando $s > M$. También asumiremos que $N >> M$, de manera que la primera, $x^{(0)}$, y la última muestra $x^{(N)}$ prácticamente no están correlacionadas. Bajo el supuesto de $Var(\hat{f}_N) = \frac{\tau_f Var(f)}{N}$, observamos

$$
\begin{aligned}
Var(\hat{f}_N) &= E\left(\hat{f}_N^2\right) - \left(E\left(\hat{f}_N\right)\right)^2 \\
&= E\left[\left(\frac{1}{N}\sum_{n=1}^{N} f(X^{(n)})\frac{1}{N}\sum_{m=1}^{N} f(X^{(n)})\right)\right) - \left(E\left(\frac{1}{N}\sum_{n=1}^{N} f(X^{(n)})\right)\right)^2 \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} E\left(f(X_m)f(X_n)\right) - E\left(f^2\right)
\end{aligned}
$$

puesto que el valor de $E(f(X_n)) = E(f)$, independiente de n, si la cadena es homogénea y estacionaria. Para el caso en que $M << N$

$$
\begin{aligned}
\sum_{n=1}^{N}\sum_{m=1}^{N} E(f(X_m)f(X_n)) &\approx \sum_{n=1}^{N}\left[E(f(X_n)f(X_n)) + 2\sum_{s=1}^{N-M} E(f(X_n)f(x^{(n+s)}))\right] \\
&= \sum_{n=1}^{N}\left[C_{ff}(0) + (E(f))^2 + 2\sum_{s=1}^{N-M} C_{ff}(s) + (E(f))^2\right]
\end{aligned}
$$

de manera que:

$$
\begin{aligned}
Var(\hat{f}_N) &= \frac{1}{N^2}\sum_{n=1}^{N} Var(f) + 2\sum_{s=1}^{N-M} C_{ff}(s) \\
&\approx \frac{Var(f)}{N^2}\sum_{n=1}^{N}\left[1 + 2\sum_{s=1}^{M}\rho_{ff}(s)\right] \\
&\approx \frac{Var(f)\tau_f}{N}
\end{aligned}
$$

Sea

$$
\tau_g = \tau_g(\infty) = \lim_{N\to\infty} \tau_f. \tag{3.14}
$$

Si suponemos que τ_g converge, entonces

$$
\tau_g = 1 + 2\sum_{t=1}^{\infty}\rho_{ff}(s). \tag{3.15}
$$

El valor τ_g recibe el nombre de **Integrated Autocorrelation Time** (IAT). Es de interés desarrollar algoritmos MCMC con un IAT pequeño, pues permite estimar de forma más precisa

la varianza, evitando generar muestras demasiado grandes.

El IAT nos indica el número de simulaciones que debemos descartar de nuestra muestra MCMC para obtener una muestra pseudo independiente. Lo anterior sólo constituye una breve introducción al IAT, por lo que referimos al lector a consultar Geyer (1992) para la estimación.

Bibliografía

[1] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Verlag.

[2] Berger, J. and Chen, M. H. (1993). *Predicting retirement patterns: prediction for a multinomial distribution with constrained parameter space*. The Statistician, 42, 427-443.

[3] Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, Wiley: Chichester, UK.

[4] Besag, J. y Clifford, P. (1989). *Generalized Monte Carlo significant test*, Biometrika 76:633 642.

[5] Boneh, A. y Golan, A. (1979) *Constraints redundancy and feasible region boundedness by Random Feasible Point Generator (RFPG)*. Proceedings of the EURO III Conference, Amsterdam, April 9-11, 1979.

[6] Calvetti, Kuceyeski and Somersalo, *Sampling-Based Analysis of a Spatially Distributed Model for Liver Metabolism at Steady State*, Multiscale Model Simul. Vol. 7 2008, No. 1 pp. 407 431.

[7] Christen J A and Fox C (2005), *MCMC using an Approximation*, Journal of Computational and Graphical Statistics, Vol. 14, No. 4, December 2005 pp.795-810.

[8] Christen J A and Fox C (2010), *A general purpose sampling algorithm for continuous distributions (the t-walk)* Bayesian Analysis 5 (2) 1-20. doi:10.1214/10 BA603.

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

[10] DeGroot, M. H. (1970), *Optimal statistical decisions*, McGraw Hill: NY.

[11] Geman, S. and Geman, D. (1984). *Stochastic relaxation, gibbs distributions, and the bayesian restoration of images*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 6(6):721-741.

[12] Gamerman, D. and H. S. Migon (1993). *Inferencia Estatística: Uma Abordagem Integrada*. Textos de Métodos Matemáticos. Instituto de Matemática, UFRJ.

[13] Gelfand, A.E. (with A.F.M. Smith) 1990. *Sampling Based Approaches to Calculating Marginal Densities*, Journal Amer. Stat. Assoc., 85, 398-409.

[14] Geyer (1992). Practical MCMC, Statistical Science, Vol. 7, No. 4., pp. 473-483.

[15] Glasserman, Paul (2003). *Monte Carlo methods in financial engineering*. Springer-Verlag.

[16] Hastings, W. (1970). *Monte Carlo sampling methods using Markov chains and their application*. Biometrika, 57:97 109.

[17] Kaipio, Jari y Somersalo E. (2004) *Statistical and Computational Inverse Problems*.

[18] Kaufman, DE, and Smith, RL (1998) *Direction choice for accelerated convergence in hit-and-run sampling*, Oper. Res. 46, no. 1 , 84-95.

[19] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University Press.

[20] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). *Equations of state calculations by fast computing machines*. Journal of Chemical Physics, 21(6):1087 1092.

[21] Robert, C. (2001). *The Bayesian Choice*, second edition. Springer Verlag, New York.

[22] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second edition. Springer Verlag, New York.

[23] Roberts, G. O. and Rosenthal, J. S. (2001). *Optimal scaling for various Metropolis Hastings algorithms*. Statist. Sci. 16 351–367.

[24] Smith,R. L. (1980). *A Monte Carlo procedure for generating random feasible solutions to mathematical programs*. A Bulletin of the 0RSA.TIMS Joint National Meeting, Washington, D.C., 101.

[25] Smith,R. L. (1984). *Efficient Monte Carlo procedure for generating points uniformly distributed over bounded regions*. Oper. Res. 32 1297 1308.

[26] Tanner, Martin A. y Wong, Hung Wing (1987) *The Calculation of Posterior Distributions by Data Augmentation*, Journal of the American Statistical Association, Vol. 82, No. 398. (Jun., 1987), pp. 528-540.

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