Universidad Nacional Autónoma de México

Facultad de Ciencias

Método de Propagación de Esperanzas.

TESIS

que para obtener el título de

Actuario

presenta

Pablo Ulises Hernández Garcés

Director de tesis:

Dr. Ramsés Humberto Mena Chávez

Ciudad Universitaria, CDMX., 2018

Dedicado a:
Mis padres y
hermano.
Mi esposa y
a la pequeña sofi.

Agradecimientos

En primer lugar quiero agradecer a mi asesor, el Dr. Ramsés, por el apoyo brindado para la realización del presente trabajo.

También agradezco a los profesores que fueron parte importante de mi formación. Especialmente a la Dra. Ruth, que con sus enseñanzas me transmitió el gusto y pasión por la estadística.

Para la UNAM, mi alma máter, quiero expresar mi gratitud y cariño. Agradezco la gran calidad que presenta en sus académicos y en la vida estudiantil. Gracias a esta fue posible compartir excelentes momentos con mis amigos y, además, me dió la oportunidad de conocer a la mujer que hoy acompaña mi camino.

Más aún, quiero agradecer a mis padres y a mi hermano que han creído en mi a cada momento y que gracias a ellos esto ha sido posible.

Resumen

La estadística bayesiana representa una herramienta importante en inferencia estadística, sin embargo su correcta aplicación requiere de técnicas de cómputo y aproximación eficientes. En este trabajo revisamos una técnica de aproximación que permite realizar inferencia bayesiana de una manera relativamente veloz cuando se le compara con otras en la literatura. El método se denomina propagación de esperanzas (*Expectation Propagation*) y fue propuesto por Minka (2001).

En el presente trabajo se introducen al lector los conceptos básicos de la estadística bayesiana y se aborda uno de los temas principales con los que cuenta, la estimación de la distribución posterior. Este se analiza desde una perspectiva computacional, donde se busca tener una forma de realizar inferencia bayesiana con mayor precisión y a un bajo costo computacional.

Dentro del desarrollo de esta tesis, se analizan de forma general algunos de los métodos más usados para la aproximación en la inferencia bayesiana. Comenzando por métodos de Monte Carlo, muestreo por importancia (*Importance Sampling*) y el muestreo de Gibbs (*Gibbs sampling*). También se abordan los métodos de Laplace y variación de Bayes (*Variational Bayes*), hasta llegar a los métodos de densidad de emisión (ADF) y el de propagación de esperanzas (EP). Para cada uno de los métodos expuestos en el presente trabajo se desarrolla la idea principal y se ejemplifica. Además se presentan los algoritmos de cada uno de ellos.

Por último se comparan los métodos en cuestión de precisión y costo computacional en el problema de inferencia bayesiana en el que se cuenta con una verosimilitud dada por una mezcla de dos distribuciones gausianas univariadas. En este experimento se encuentra que el último método expuesto, propagación de esperanzas, es más eficiente que los anteriores mencionados.

Introducción

Uno de los principales problemas de la estadística bayesiana ha sido el costo computacional. Esta tesis presenta una técnica de aproximación que permite realizar inferencia bayesiana de una forma más rápida y precisa que los métodos propuestos con anterioridad. Este método, propagación de esperanzas, *Expectation Propagation (EP) (Minka, 2001)*, es una generalización del método de Densidad de Emisión, *Assumed Density Filtering (ADF)*. Esta mejora muestra una alternativa a la aproximación de la distribución posterior con una distribución más simple, la cual es muy cercana en el sentido de la divergencia de Kullback-Leibler (*KL-divergence*).

En el primer capítulo se presentan los elementos básicos de la estadística bayesiana desde el punto de vista de la teoría de las decisiones, la cual provee una justificación axiomática de esta corriente de la estadística. En este apartado se exponen los elementos básicos de teoría de las decisiones; se define el conjunto de acciones, las funciones de decisión, funciones de pérdida y cómo a partir de estas se puede llegar a la justificación del enfoque bayesiano.

En el segundo capítulo, el lector encontrará los métodos de aproximación más usuales en la literatura; métodos de muestreo como Monte Carlo, muestreo por importancia y muestreo de Gibbs, también métodos de aproximación local y aproximación por límites inferiores como el método de LAPLACE y el método de variación de Bayes, respectivamente. Por último, un método iterativo de igualación de momentos, el método de densidad de emisión. Para cada uno de estos métodos se desarrollan los elementos básicos, se ejemplifican y se proporciona el algoritmo correspondiente.

En el tercer capítulo, el cual es el objetivo principal de esta tesis, se presenta el trabajo propuesto por Minka (2001), el algoritmo de propagación de esperanzas. Este algoritmo, el cual es una extensión del método de densidad de emisión, se desarrolla en su forma general y se ejemplifica en un problema de inferencia bayesiana sobre una mezcla de dos distribuciones gausianas.

Por último, en el capítulo cuatro todos los algoritmos expuestos son desarrollados e implementados para resolver el problema de la mezcla de dos distribuciones normales. A su vez, son comparados entre sí considerando la precisión y el costo computacional que cada uno de los algoritmos tiene para este problema en particular.

Índice general

Agradecimientos **III**

Resumen **V**

Introducción **VII**

1. Perspectiva bayesiana **1**

1.1. Idea principal 1

1.2. Introducción a la teoría de decisión estadística 2

1.3. Teorema de Bayes 5

1.4. Inferencia estadística 5

2. Métodos de integración **9**

2.1. Métodos de integración Monte Carlo 10

2.1.1. Muestreo por importancia 13

2.1.2. Muestreo de Gibbs 16

2.2. Método de *LAPLACE* 19

2.3. Variación de Bayes 22

2.3.1. Variación de Bayes vía el algoritmo EM 24

2.4. Densidad de emisión 27

3. Método de propagación de esperanzas **33**

3.1. Mezcla de dos gausianas vía propagación de esperanzas 37

4. Comparación y resultados **41**

4.1. Métodos aplicados a una mezcla de dos distribuciones gausianas 41
4.1.1. Muestreo por importancia. 41
4.1.2. Muestreo de Gibbs. 42
4.1.3. Método de Laplace. 43
4.1.4. Variación de Bayes. 46
4.2. Comparación 47

5. Conclusiones **51**

A. Esperanza-Maximización **53**

B. Familias exponenciales **55**

B.1. Forma exponencial de la distribución Normal multivariada 56

Bibliografía **57**

Capítulo 1

Perspectiva bayesiana

En este capítulo se abordan conceptos básicos de estadística bayesiana con la finalidad de poner en contexto los temas desarrollados en este trabajo.

Existen varias formas de introducir el enfoque bayesiano al problema de inducción estadística, en este capítulo se introduce dicho enfoque a partir de la teoría de las decisiones. Para el desarrollo de esta línea de pensamiento se consultaron Jayanta et al. (2006) y Berger (1985).

Como punto de partida de este capítulo se plantea una idea básica sobre la estadística bayesiana de una manera simple, en las siguientes secciones se plantea de manera más formal la justificación de este enfoque.

En la segunda sección se encuentra la introducción a la teoría de decisiones. En ella, se plantean los conceptos básicos de la teoría frecuentista y cómo se modifican bajo la perspectiva bayesiana, esto se ilustra a través de algunos ejemplos básicos. Con todo lo expuesto en esta sección se tiene la justifiación de la estadística bayesiana.

Para la tercera sección, se cuenta con una revisión del teorema de Bayes y cómo este cambia la perspectiva de la estadística. Por último, al final de este capítulo se estudia de manera breve la inferencia estadística desde una perspectiva bayesiana.

1.1. Idea principal

Bajo el enfoque bayesiano, los parámetros son modelados como variables aleatorias. Se debe tener claro que esta aleatoriedad no describe su variabilidad, sino que es una descripción de la incertidumbre que se tiene sobre sus verdaderos valores.

Esta modelación requiere de la construcción de una distribución de probabilidad para el parámetro θ, llamada distribución *a priori* o inicial, la cual refleja la incertidumbre que se tiene sobre los parámetros. De esta forma la distribución *a priori*, $\pi(\theta)$, junto con la distribución condicional de los datos dado un valor del parámetro, $f(x|\theta)$, da lugar a una distribución conjunta sobre el espacio χ x Ω, el cual es el espacio subyacente al fenómeno en estudio. Donde χ y Ω son los espacios donde $\mathbf{X}$ y θ toman valores, respectivamente.

Dada la construcción de la distribución conjunta sobre el espacio del fenómeno en estudio, $f(x, \theta)$, es posible obtener información relevante del parámetro teniendo en cuenta la información observada. De esta forma, una vez observados los datos es posible actualizar la incertidumbre inicial y dar lugar a una nueva distribución que refleje de mejor manera el comportamiento de los parámetros, llamada distribución *a posteriori* o distribución posterior.

Esta actualización se realiza a través del teorema de Bayes

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}, \tag{1.1}$$

donde $f(\theta|x)$ es la distribución posterior, y a la integral que se encuentra en el denominador de (1.1) se le conoce como constante de normalización, ya que no depende de θ, o evidencia del modelo.

En la siguiente sección se aborda una de las formas de justificar lo antes mencionado, esto a través de teoría de decisión.

1.2. Introducción a la teoría de decisión estadística

La teoria de decisión estadística está enfocada en la toma de decisiones en presencia de conocimiento estadístico, i.e. en presencia de los datos, lo cual ayuda a despejar la incertidumbre involucrada en el problema de decisión. Esta incertidumbre se considera una cantidad desconocida y se representa por θ, el cual es un elemento del espacio Θ, llamado espacio parametral.

El elemento básico de esta teoría son las decisiones, comunmente llamadas acciones, denotadas por α, las cuales son elementos del espacio de todas las posibles acciones, denotado por $\mathscr{A}$. Una vez seleccionada una acción, α, se considera uno de los elementos principales de la teoría de decisión, la función de pérdida. Si se toma la acción α_1 y parámetro θ_1, entonces se incurre en un pérdida, denotada por $L(\theta_1, \alpha_1)$. Se supone que la función de pérdida, $L(\theta, \alpha)$, está definida para todo $(\theta, \alpha) \in \Theta \times \mathscr{A}$. Se considera que las funciones de pérdida son funciones acotadas, de tal forma que las integrales sobre ellas estén bien definidas.

Por otro lado, la teoría de decisiones cuenta con funciones de decisión o reglas de decisión, $\delta(\mathbf{x})$, que toman valores en $\mathscr{A}$. Suponga que $\delta(\mathbf{x}) = \alpha$ para un conjunto de datos $\mathbf{x}$, entonces el investigador que tome esta regla de decisión elegirá la acción α dado un conjunto particular de datos e incurrirá en una pérdida $L(\theta, \alpha)$.

Los problemas de inferencia estadística tienen su representación en la teoría de las decisiones. Por ejemplo, en un problema en el que se desea estimar $\tau(\theta)$, una función de θ que toma valores reales, se tiene que $\mathscr{A} = \mathbb{R}$, la función de pérdida cuadrática está dada por

$$L(\theta, \alpha) = (\alpha - \tau(\theta))^2$$

y la función de decisión $\delta(\mathbf{x})$ es un estimador de $\tau(\theta)$.

Si el problema es una prueba de hipótesis como $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, entonces $\mathscr{A} = \{\alpha_0, \alpha_1\}$, donde α_j se refiere a la acción de aceptar H_j, $L(\theta, \alpha) = 0$ si θ cumple H_j y $L(\theta, \alpha) = 1$ en otro caso. Ahora, si $I(\mathbf{x})$ es una función indicadora de la región de rechazo de H_0, entonces la correspondiente regla de decisión, $\delta(\mathbf{x})$, es igual a α_j si $I(\mathbf{x}) = j$; $j = 0, 1$.

Como se ha mencionado, existe incertidumbre que involucra al problema de decisión. Debido a esto, la pérdida en la que se incurre no es conocida con certeza, para mitigar esto se considera la pérdida esperada dada una decisión y posteriormente se elige un óptimo con respecto a esta pérdida esperada.

Definición 1.1 *Se define la pérdida esperada o función de riesgo de una regla de decisión* $\delta(\boldsymbol{x})$ *como*

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(\boldsymbol{X}))] = \int_\chi L(\theta, \delta(\boldsymbol{x})) f(x|\theta) dx, \tag{1.2}$$

donde χ *es el espacio en el que toma valores la variable aleatoria* $\boldsymbol{X}$.

En el ejemplo de estimación de $\tau(\theta)$, se tiene que la función de riesgo de la regla de decisión asociada, $\delta(\mathbf{x})$, está dada por

$$R(\theta, \delta) = \mathbb{E}_\theta[(\tau(\theta) - \delta(\mathbf{x}))^2],$$

también conocido como error cuadrático medio. Por otro lado, si $\delta(\mathbf{x})$ es la función indicadora de la región de rechazo de H_0, entonces $R(\theta, \delta)$ es la probabilidad del error tipo I si $\theta \in \Theta_0$ y la probabilidad del error tipo II si $\theta \in \Theta_1$.

La Definición (1.1) pertenece a un punto de vista frecuentista, ya que esta promedia la pérdida sobre todos los posibles conjuntos de datos. Sin embargo, a partir de la Definición (1.1) se puede definir la equivalente bajo una perspectiva bayesiana.

Desde un punto de vista bayesiano la idea frecuentista no es convincente, debido a que este no se adhiere a la idea de ponderar sobre todos los posibles conjuntos de datos, visto en la ecuación (1.2). Para los bayesianos la forma más natural de considerar la pérdida esperada es aquella que considera la incertidumbre propia de θ, ya que θ es desconocida al momento de tomar la decisión y a fin de representar esta incertidumbre es considerada como una variable aleatoria, con densidad $\pi(\theta)$.

A partir de esto se definen los conceptos de la teoría de decisión bayesiana, la cual construye las bases de la estadística bayesiana.

Definición 1.2 *Si* $\pi(\theta|\boldsymbol{x})$ *es la función de distribución posterior, considerando a* $\pi(\theta)$ *como la distribución inicial de* θ*, se define la pérdida esperada bayesiana o riesgo posterior de una regla de decisión* $\delta(x)$ *como*

$$\rho(\pi^*, \delta(x)) = \mathbb{E}^{\pi^*}[L(\theta, \delta(x))] = \int_\Theta L(\theta, \delta(x)) \pi(\theta|\boldsymbol{x}) d\theta,$$

donde

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)\pi(\theta)$$

A la función $f(x|\theta)$ se le conoce como verosimilutd, y corresponde con la distribución de los datos dado un valor del parámetro. Por otro lado, la función $\pi(\theta|\mathbf{x})$ es la distribución posterior de θ, la cual incluye la actualización de la creencia sobre el parámetro una vez observados los datos.

A partir de la Definición (1.2) se define la correspondiente acción de Bayes.

Definición 1.3 *Se define la acción de Bayes $\delta^*(x)$ como aquel valor de $\delta(x)$ que minimice el riesgo posterior. i.e.*

$$\delta^*(x) = \arg\min_{\delta(x)} \mathbb{E}^{\pi^*}[L(\theta, \delta(x))].$$

Considerando el ejemplo de estimación en el que $L(\theta, \delta(x)) = (\theta - \delta(x))^2$ se tiene que el estimador bayesiano es la media posterior; i.e. $\delta^*(x) = \mathbb{E}[\theta|\mathbf{x}]$.

Una de las diferencias más importantes entre la función de riesgo y el riesgo posterior es que éste último, al promediar sobre todos los posibles valores de θ y para un conjunto de datos dado, es un número, mientras que la función de riesgo promedia sobre todos los posibles conjuntos de datos dado un valor fijo de θ. Por lo que esta función puede llevar consigo problemas de ordenamiento, en el sentido de no saber claramente qué regla de decisión es mejor[1].

A pesar de las distintas definiciones de cada uno de los enfoques, la teoría de decisión bayesiana toma algunos conceptos de la clásica para definir otros nuevos.

Definición 1.4 *La regla de Bayes, δ_π, es una función que minimiza*

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)d\theta, \tag{1.3}$$

donde $R(\theta, \delta)$ es el riesgo frecuentista, Definición (1.1).

Definición 1.5 *Se define el riesgo de Bayes como*

$$r(\pi) = r(\pi, \delta_\pi).$$

La regla de Bayes, Definición (1.4), promedia el riesgo frecuentista sobre una distribución a priori de θ. Por otro lado, el riesgo de Bayes, Definición (1.5), es igual a la ecuación (1.3) evaluada en la regla de Bayes. Mientas que el riesgo de Bayes es un concepto frecuentista, ya que promedia sobre x, la expresión $r(\pi, \delta)$ puede ser interpretada de una forma distinta.

$$\begin{aligned} r(\pi, \delta) &= \int \left[\int L(\theta, \delta(x)) f(x|\theta) dx \right] \pi(\theta) d\theta \\ &= \int \left[\int L(\theta, \delta(x)) \pi(\theta|x) d\theta \right] f(x) dx \\ &= \int \rho(\pi, \delta(x)) f(x) dx \end{aligned} \tag{1.4}$$

[1]Se dice que una regla de decisión δ_1 es R-mejor que otra δ_2 si $R(\theta, \delta_1) \leq R(\theta, \delta_2) \forall \theta \in \Theta$, con la desigualdad estricta para algún $\theta \in \Theta$. Derivado de esta definición se dice que δ es una regla de decisión admisible si ninguna otra regla de decisión es R-mejor.

Es importante notar que la ecuación (1.4) es el riesgo posterior promediado sobre la distribución marginal de los datos, $f(x)$. Esto implica que la regla de Bayes se puede obtener tomando la acción de Bayes para cada x.

Por lo tanto, con lo mostrado en esta sección se encuentra una justificación del enfoque bayesiano, ya que éste se puede estructurar como un problema de decisión.

1.3. Teorema de Bayes

Para encontrar la distribución posterior se hace uso del Teorema de Bayes, el cual parte del concepto de probabilidad condicional y da una buena interpretación del modelo bayesiano, el cual debe su nombre a este teorema.

Teorema 1.1 *Sean A y B dos eventos tales que $P(B) > 0$. Suponga que $P[B|A]$ existe, entonces la $P[A|B]$ está dada de la siguiente forma*

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B|A]P[A]}{P[B]}. \tag{1.5}$$

Una generalización de (1.5) nos permite calcular la distribución del parámetro dado los datos. Considerando las funciones de distribución *a priori* y de verosimilitud tenemos, para el caso continuo:

$$f_{\Theta|\mathbf{x}}(\theta|x) = \frac{f_{\mathbf{x}|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int_{\Omega} f_{\mathbf{x}|\Theta}(x|t)f_{\Theta}(t)dt}, \tag{1.6}$$

con $f_{\mathbf{x}|\Theta}(x|\theta)$ la función de verosimilitud, y $f_{\Theta}(\theta)$ la distribución *a priori* de Θ. El denominador de la ecuación (1.6) es llamado la densidad predictiva a priori de los datos $\mathbf{X}$, denotada como $f_{\mathbf{X}}(x)$. El caso discreto es análogo.

La expresión en (1.6) se interpreta de manera natural como la probabilidad de la cantidad de interés, θ, dado el valor del experimento, x. La cual está en función del modelo que rige el experimento, verosimilitud, y la incertidumbre que se tiene sobre el valor del parámetro de la verosimilitud. El denominador, al ser una función que no depende de la cantidad de interés, θ, es considerada una constante de normalización.

1.4. Inferencia estadística

Cualquier inferencia estadística que se desee realizar sobre el parámetro θ se encuentra basada en la distribución posterior. A esto se le conoce como el paradigma bayesiano. Por ejemplo, muchas veces se tiene interés en conocer la media y varianza posterior del

parámetro θ, las cuales son integrales sobre la distribución (1.6):

$$\mathbb{E}[\theta|\mathbf{x}] = \int_{\Omega} \theta f_{\theta|\mathbf{x}}(\theta|\mathbf{x}) d\theta, \tag{1.7}$$

$$\mathbb{E}[\theta^2|\mathbf{x}] = \int_{\Omega} \theta^2 f_{\theta|\mathbf{x}}(\theta|\mathbf{x}) d\theta, \tag{1.8}$$

$$Var[\theta|\mathbf{x}] = \mathbb{E}[\theta^2|\mathbf{x}] - \mathbb{E}[\theta|\mathbf{x}]^2. \tag{1.9}$$

Estas expresiones son fáciles de encontrar cuando la distribución posterior tiene una forma paramétrica conocida. Lo cual es común cuando la distribución inicial es conjugada con la distribución condicional de los datos dado un valor del parámetro.

Definición 1.6 *Una familia de distribuciones sobre el espacio parametral Ω, definida como $\mathscr{G} := \{q_\phi(\theta); \phi \in \Phi\}$, se dice conjugada (cerrada bajo muestreo) para la familia paramétrica $\mathscr{P}_\Omega$ si la distribución posterior $q \in \mathscr{G}$.*

En otras palabras, si la distribución inicial, $q_\Theta(\theta)$, pertenece a una familia paramétrica $\mathscr{G}$, se dice que $q_\Theta(\theta)$ es conjugada para la verosimilitud, $f_{\mathbf{x}|\Theta}(\mathbf{x}|\theta)$, si la distribución posterior, $f_{\Theta|\mathbf{x}}(\theta|\mathbf{x})$, pertenece también a la familia paramétrica $\mathscr{G}$.

Bajo el esquema de familias conjugadas, la inferencia sobre los parámetros resulta sumamente sencilla, ya que todo está determinado por las características de la familia. Sin embargo, esta característica del modelo no se encuentra con facilidad en casos prácticos.

Para verosimilitudes que siguen una cierta familia paramétrica se han encontradado distribuciones iniciales que permiten obtener un modelo conjugado. Las más conocidas se muestran en la siguiente tabla.

$f_{\mathbf{x}\|\theta}$	f_θ (con parámetros conocidos)	$f_{\theta\|\mathbf{x}}$
$N(\theta, \sigma^2)$ σ^2 conocida	$N(\mu, \tau^2)$	$N(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
$\mathsf{Poiss}(\theta)$	$\mathsf{Ga}(\alpha, \beta)$	$\mathsf{Ga}(\alpha + x, \beta + 1)$
$\mathsf{Ga}(\nu, \theta)$	$\mathsf{Ga}(\alpha, \beta)$	$\mathsf{Ga}(\alpha + \nu, \beta + x)$
$\mathsf{Bin}(n, \theta)$ n conocida	$\mathsf{Be}(\alpha, \beta)$	$\mathsf{Be}(\alpha + x, \beta + n - x)$
$\mathsf{BinNeg}(m, \theta)$ m conocida	$\mathsf{Be}(\alpha, \beta)$	$\mathsf{Be}(\alpha + m, \beta + x)$
$\mathsf{Multinom}(\theta_1, \ldots, \theta_n)$	$\mathsf{Dirichlet}(\alpha_1, \ldots, \alpha_n)$	$\mathsf{Dirichlet}(\alpha_1 + x, \ldots, \alpha_n + x)$

Tabla 1.1: Tabla de Familias conjugadas.

Se puede observar que para distribuciones como las de la Tabla (1.1) los cálculos relacionados con la inferencia del parámetro es sencilla y pueden ser realizados de manera analítica.

En general, una forma de facilitar los cálculos para la obtención de la distribución posterior ó *a posteriori*, es observando que el denominador es una expresión que no depende del valor de θ, por lo que es considerado como una constante. De esta forma podemos ver la ecuación (1.6) del teorema de Bayes como:

$$f_{\Theta|\mathbf{x}}(\theta|x) \propto f_{\mathbf{x}|\Theta}(x|\theta) f_\Theta(\theta). \tag{1.10}$$

Así, es posible trabajar con los *kernels* de la distribución inicial y de verosimilitud, de esta manera se encuentra el *kernel* de la distribución posterior, la cual queda completamente especificada una vez que se encuentra la constante de normalización.

Para el caso de familias conjugadas, una forma de verificar que las distribuciones de la Tabla (1.1) cumplen esa propiedad es utilizando (1.10).

EJEMPLO **1.1** *Considere una distribución inicial* $N(\theta|\mu,\tau^2)$ *y verosimilitud* $N(x|\theta,\sigma^2)$, *con* σ^2 *conocida. Partiendo de la ecuación* (1.6) , *se tiene:*

$$\begin{aligned}
f(\theta|\boldsymbol{x}) &= \frac{N(x|\theta,\sigma^2)N(\theta|\mu,\tau^2)}{\int_{-\infty}^{\infty} N(x|t,\sigma^2)N(t|\mu,\tau^2)dt} \\
&\propto N(x|\theta,\sigma^2)N(\theta|\mu,\tau^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \\
&\propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}\exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \\
&= \exp\left\{-\frac{x^2-2x\theta+\theta^2}{2\sigma^2}-\frac{\theta^2-2\mu\theta+\mu^2}{2\tau^2}\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left[\theta^2(\tau^2+\sigma^2)-2\theta(x\tau^2+\mu\sigma^2)+(x^2\tau^2+\mu^2\sigma^2)\right]\right\} \\
&\propto \exp\left\{-\frac{(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left[\theta^2-2\theta\left(\frac{x\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)+\left(\frac{x\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)^2\right]\right\} \\
&= \exp\left\{-\frac{(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left[\theta-\left(\frac{x\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)\right]^2\right\}
\end{aligned}$$

La última igualdad es el kernel de una distribución normal con parámetros:

$$\begin{aligned}
media = \mu^* &= \left(\frac{x\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right) \\
varianza = \sigma^* &= \frac{\sigma^2\tau^2}{(\tau^2+\sigma^2)}
\end{aligned}$$

De este modo, es posible observar que el kernel de la distribución posterior es el mismo que el de la a priori, salvo por los parámetros, por lo cual no es necesario realizar el cálculo de la constante de normalización pues el modelo está totalmente especificado. Ya que, al ser $f(\theta|\boldsymbol{x})$ *una función de densidad de probabilidad, la constante que porporciona la igualdad en:*

$$f(\theta|\boldsymbol{x}) \propto \exp\left\{-\frac{(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left[\theta-\left(\frac{x\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)\right]^2\right\} \tag{1.11}$$

Debe ser la constante de normalización de una distribución $N(\mu^*,\sigma^*)$, *esto es:*

$$f(\theta|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^*}}\exp\left\{-\frac{1}{2\sigma^*}(\theta-\mu^*)^2\right\} \tag{1.12}$$

De esta forma se pueden obtener todas las distribuciones de la Tabla (1.1)

De manera gráfica se puede observar la actualización *de la distribución inicial, la cual representa la creencia sobre el parámetro, una vez considerados los datos, mediante la verosimilitud, dando lugar a la distribución posterior.*

La siguiente Figura presenta gráficamente el comportamiento de la familia conjugada Normal, para una distribución inicial $N(\theta|3,4)$ *y una verosimilitud* $N(x|\theta,1.25)$*, donde* θ *fue generado a partir de la distribución inicial. Por lo cual, la distribución posterior resultante es una distribución* $N(\theta|2.06,0.976)$*.*

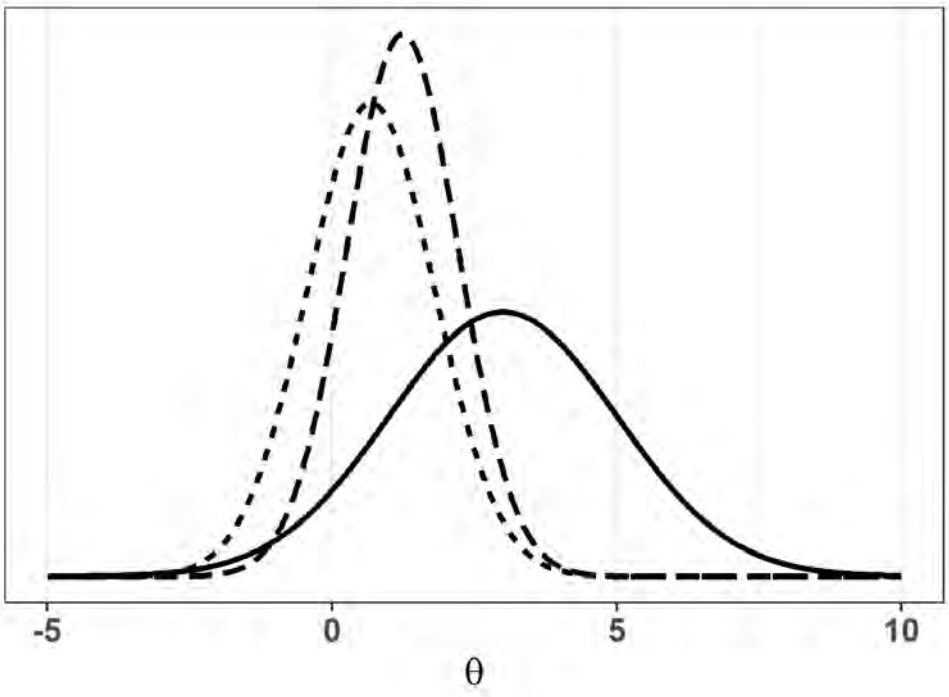

Figura 1.1: Familia conjugada Normal. La línea sólida representa la distribución a priori. la línea punteada representa la verosimilitud y la línea discontinua representa la distribución posterior.

De esta forma, la estadística bayesiana puede incorporar conocimiento previo sobre el fenómeno de estudio. Cómo este afecta a la inferencia se puede observar en la Figura (1.1). Por lo que se puede obtener mejores resultados en algunos problemas en comparación con el enfoque clásico.

Sin embargo, en un contexto distinto al de familias conjugadas, el cálculo de la constante de normalización no siempre es una tarea sencilla. La forma del producto de la verosimilitud con la distribución inicial puede ser intratable al momento de integrarla de manera analítica o inclusive puede llegar a ser muy costosa computacionalmente para una aproximación numérica convencional, como la de cuadratura.

Por ello se han desarrollado distintas técnicas de aproximación que han atacado este problema, las cuales son objetivo del presente trabajo.

Capítulo 2

Métodos de integración

En este capítulo se describe el trabajo realizado previamente al método de propagación de esperanzas (EP), (Minka, 2001), para aproximaciones en la inferencia bayesiana.

La aproximación clásica para la integración numérica es el método determinístico de cuadratura, (Davis and Rabinowitz, 1984). En esta aproximación la integral es evaluada en una serie de puntos y se construye una función que interpole esos valores. La función interpoladora se elige dentro de una familia de funciones que puedan ser integradas analíticamente, como los polinomios. La integral de la función interpolante aproxima la integral deseada, y con un número suficiente de puntos la aproximación puede ser arbitrariamente precisa.

Tomando en cuenta un conjunto de puntos $x_1, \ldots, x_n$, la interpolación y la integración se reduce a una suma ponderada de los valores de la función

$$I = \int_A f(x)dx,$$
$$\hat{I} = \sum_{i=1}^{n} w_i f(x_i).$$

Donde los pesos w_i son conocidos. Este método es excelente con integrandos que son simples al momento de interpolar. En una dimensión se obtienen resultados muy precisos y es difícil conseguir una mejor aproximación con otras técnicas. Sin embargo, en varias dimensiones es intratable, debido al vasto número de nodos que se requieren para realizar una interpolación de una función compleja.

En problemas de inferencia bayesiana, el método de cuadratura no es una buena opción debido a que los integrandos son distintos de cero en regiones pequeñas, i.e. son *escasos*, lo cual implica que muchos de los puntos serán en vano. Existen otras técnicas deterministas que tratan de evitar estos problemas explotando más propiedades de la integral y no sólo sus valores en puntos dados. Una alternativa para la aproximación son los métodos no deterministicos.

Para los métodos no deterministicos, como el método Monte Carlo (Robert and Casella, 2004), se hace uso de la ley de los grandes números en lugar de intentar interpolar o aproximar la integral de alguna otra forma. El método Monte Carlo, a pesar de ser ineficiente en bajas dimensiones, debido a que requiere demasiados nodos comparado con

otros métodos como la cuadratura, es regularmente el único método factible en grandes dimensiones.

Considerando los problemas de *escasez* de la inferencia bayesiana, éstos pueden ser tratados mediante la técnica de muestreo por importancia (Robert and Casella, 2004), la cual muestrea a partir de una distribución propuesta que empata la forma del integrando tan bien como sea posible. El muestreo por importancia tiene además la ventaja de trabajar para regiones infinitas.

Con muestreo por importancia intercambiamos las dificultades de integración numérica por dificultades en el muestreo a partir de funciones complejas, ya que una buena propuesta de distribución, i.e. que ajuste de mejor manera la forma del integrando, puede ser difícil de muestrear. Por ello se proponen métodos de Monte Carlo vía cadenas de Markov como el muestreo de Gibbs (Casella and George, 1992), en el cual se genera una muestra que aproxima a una generada por el integrando, posteriormente se puede utilizar el método de Monte Carlo para realizar la aproximación con la muestra generada.

Una forma de aprender sobre las características de la función es calcular un gran número de derivadas en un sólo punto. Esta idea es cubierta por la expansión en serie de Taylor, en la que para cierto número de puntos se obtiene un número equivalente de derivadas. Esto resulta de utilidad para los problemas de inferencia bayesiana, al tener *escazes* tiene sentido enfocarse en el área donde se concentran los datos. La aplicación más popular de este razonamiento es el método de Laplace, (Kass and Raftery, 1993), donde se realiza la expansión del $\log(f)$ sobre su moda.

Por otra parte, una forma más general para abordar el problema de la inferencia bayesiana es introduciendo una distribución arbitaria $q(x)$ que sea tratable y que a través de la desigualdad de Jensen forme un límite inferior para la integral. De esta forma la integración se traduce en un problema de optimización de límites. La elección de $q(x)$ está sujeta a la obtención del límite más estrecho, lo cual es equivalente a minimizar la divergencia KL, $KL(q||f)$. Este es un método de variación de límites conocido como variación de Bayes, (Beal, 2003).

Por último, la técnica que precede a la propagación de esperanzas es la densidad de emisión, (Ramakrishnan et al., 2011), la cual es un método de aproximación recursivo. La idea principal es elegir una distribución con la cual sea sencillo trabajar y proyectarla sobre la distribución posterior en cada iteración, restringida a la familia de distribuciones de la función propuesta. De esta forma, los momentos de la distribución de aproximación son encontrados minimizando la KL divergencia entre la distribución posterior y la distribución propuesta, haciendo así una igualación de momentos.

Estos métodos; Monte Carlo, muestro por importancia, muestreo de Gibbs, Laplace, variación de Bayes y densidad de emisión, son presentados con mayor detalle en este capítulo.

2.1. Métodos de integración Monte Carlo

El origen del método de Monte Carlo se remonta a 1944, pero fue hasta la segunda guerra mundia cuando el método se utilizó como una herramienta de investigación para el desarrollo de la bomba atómica, (Hammersle and Handscomb, 1964). El trabajo consistía

en la simulación de la difusión de un neutrón. Por su parte, el desarrollo sistemático de las ideas utilizadas fueron desarrolladas por Harris y Herman Kahn en 1948. Sin embargo, el crédito del desarrollo de las técnicas del método Monte Carlo se atribuye a Ulam, von Neumann y Fermi.

El método de integración Monte Carlo aborda la solución de integrales de la forma

$$\mathbb{E}_f[h(x)] = \int_\chi h(x)f(x)dx. \tag{2.1}$$

El principio detrás del método es generar una muestra $(x_1, \ldots, x_n)$ de la función f y proponer como aproximación la media empírica

$$\overline{h}_n = \frac{1}{n}\sum_{j=1}^{n} h(x_j).$$

Este estimador, por la ley fuerte de los grandes números, converge casi seguramente al valor de la integral, i.e.

$$P(\overline{h}_n \to \mathbb{E}_f[h(x)], \text{cuando} n \to \infty) = 1.$$

Además, cuando $h^2(x)$ tiene esperanza finita bajo f, i.e.

$$\mathbb{E}_f[h^2(x)] = \int_\chi h^2(x)f(x)dx < \infty,$$

la velocidad de convergencia de $\overline{h}_n$ puede ser evaluada, ya que la varianza

$$\begin{aligned}
\mathsf{Var}(\overline{h}_n) &= \mathsf{Var}\left(\frac{1}{n}\sum_{j=1}^{n} h(x_j)\right) \\
&= \frac{1}{n}\mathsf{Var}(h(x)) \\
&= \frac{1}{n}\int_\chi \left(h(x) - \mathbb{E}_f[h(X)]\right)^2 f(x)dx
\end{aligned}$$

puede ser estimada partiendo de la muestra $(x_1, \ldots, x_n)$ a través de

$$\nu_n = \frac{1}{n^2}\sum_{j=1}^{n}\left[h(x_j) - \overline{h}_n\right]^2.$$

Por lo que para una n lo suficientemente grande,

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

tiene aproximadamente una distribución $N(0,1)$, y esto permite la construcción de una prueba de convergencia y de intervalos de confianza sobre la aproximación de $\mathbb{E}_f[h(X)]$.

EJEMPLO **2.1** *Para la siguiente función evaluar su integral sobre* $[0,1]$,

$$h(x) = (1-x^2)^{\frac{3}{2}}.$$

Se puede escribir la integral de $h(x)$ *en la forma de* (2.1) *utilizando la función de densidad de una variable aleatoria uniforme en el* $[0,1]$. *i.e.* $f(x) = 1\ \forall x \in [0,1]$

$$\int_0^1 h(x)dx = \int_0^1 h(x)f(x)dx = \int_0^1 h(x)1dx = \int_0^1 (1-x^2)^{\frac{3}{2}}dx.$$

Así, para aproximar h *se genera* $(x_1,\ldots,x_n)$, *de manera independiente, de una variable aleatoria* $X \sim U(0,1)$ *y se evalúa* $h(x)$ *en cada punto de la muestra para calcular el estimador* $\overline{h}_n$.

Para ello, se utilizó el lenguaje de programación R y se calculó el estimador, obteniendo $\overline{h}_n = 0.5889909$.[1] *Por otro lado, la integral de la función* $h(x)$ *en el intervalo* $[0,1]$ *puede ser resuelta de manera analítica y tiene un valor de* $\frac{3\pi}{16} \approx 0.5890486225$, *lo cual indica que la aproximación realizada es buena. En este caso, se tiene un error aproximado de* 5.768389×10^{-5}

De manera alternativa, si los límites de la integral, a *y* b *con* $a < b$, *son finitos, se puede optar por el cambio de variable* $y = \frac{x-a}{b-a}$ *y así la muestra podrá ser generada de una variable aleatoria* $U(0,1)$ *y la integral resultante tiene la forma*

$$I = \int_a^b f(x)dx = \int_0^1 h(y)dy$$

Con $h(y) = f(y(b-a)+a)(b-a)$, *y* $f(x)$ *la función cuya integral se desea calcular.*

La Figura 2.1 muestra la gráfica de $h(x)$ *en el* $[0,1]$ *y la convergencia de la media,* $\overline{h}_n$, *y un intervalo de confianza para cada observación generada.*

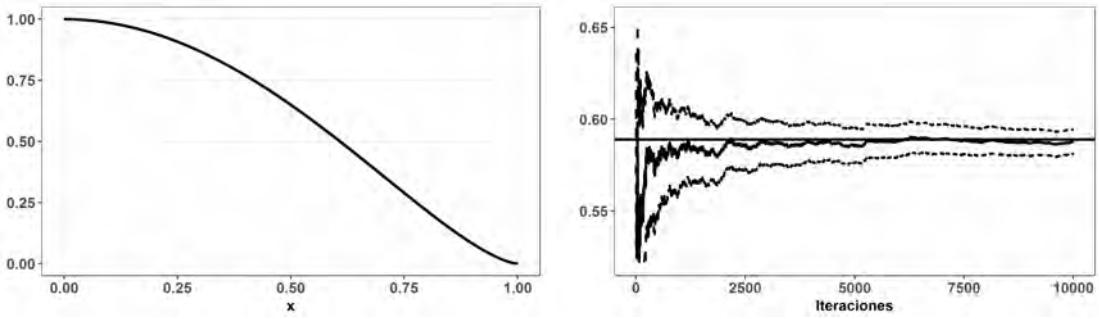

Figura 2.1: Izquierda: Gráfica de la función $h(x)$. Derecha: Gráfica de convergencia del estimador $\overline{h}_n$, línea sólida. Las líneas punteadas representan los intervalos de confianza de cada observación y la línea sólida horizontal es el valor real de la integral

Así, bajo este esquema, pareciera que el método Monte Carlo es suficiente para aproximar integrales, ya que bastaría con escribir la función que deseamos integrar en la forma

[1] En el Anexo se encuentran los códigos empleados para todos los ejercicios aqui expuestos.

de (2.1). Sin embargo, satisfacer la necesidad de generar números que provengan de la distribución f puede llegar a ser una tarea complicada, debido a que la forma de f en ocasiones es demasiado compleja. Mientras que el método provee buenas aproximaciones en la mayoría de los casos, existen alternativas más eficientes que no solo evitan una simulación directa de la función f sino que además pueden ser usadas de manera iterativa para varias integrales.

Por último, una forma de implementar el método de integración Monte Carlo es presentada en el Algoritmo 1.

Algoritmo 1 Método de Integración Monte Carlo

1: Escribir la integral en la forma de (2.1).
2: Generar n observaciones independientes de $f(x)$.
3: Calcular el estimador $\overline{h}_n$ como:

$$\overline{h}_n = \frac{1}{n}\sum_{j=1}^{n} h(x_j).$$

2.1.1. Muestreo por importancia

Este método es llamado *muestreo por importancia* debido a que depende de funciones *importantes* en lugar de la distribución original. (Robert and Casella, 2004)

El *Muestreo por importancia* se basa en una representación alternativa de la expresión con la que trabaja Monte Carlo,

$$\mathbb{E}_f[h(x)] = \int_\chi h(x)f(x)dx = \int_\chi h(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}_g\left[\frac{h(x)f(x)}{g(x)}\right]. \tag{2.2}$$

Donde la función g es una densidad arbitraria tal que es estrictamente positiva cuando $h(x)f(x) \neq 0$. χ es el conjunto donde X toma valores y por lo tanto puede ser más pequeño que el soporte de la densidad g.

La relación que se muestra en (2.2) justifica el uso del estimador para el valor de la integral

$$\overline{h}_n = \frac{1}{n}\sum_{j=1}^{n}\frac{f(x_j)h(x_j)}{g(x_j)} \to \mathbb{E}_f[h(x)]. \tag{2.3}$$

Para este estimador la muestra $(x_1, \ldots, x_n)$ es generada a partir de la función g y no de la función f como se había visto en el método Monte Carlo.

Se puede observar que el estimador (2.3) es análogo al presentado con el método de Monte Carlo. Para su construcción se utiliza la media empírica y con base en la ley de los grandes números el estimador converge casi seguramente al valor de la integral.

La escencia del muestro por importancia es resolver el problema presente en el método de Monte Carlo, ya que no es necesario muestrear de la función f, lo cual en ocasiones

podría ser muy complicado, sino de una función de densidad propuesta g. De esta forma, generar la muestra para calcular el estimador no tiene complicación, pues g es una función de densidad previamente conocida.

A pesar de que la distribución g puede ser casi cualquier densidad, de tal forma que (2.3) converge, existen opciones que son mejores que otras, y es natural comparar distintas distribuciones de g para la evaluación de (2.1). Para ello, se debe notar que mientras (2.3) converge casi seguramente a (2.1) su varianza es finita cuando

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

Entonces para aquellas distribuciones g que tengan las colas más ligeras que f, i.e. aquellas cuyo cociente $f(x)/g(x)$ no es acotado, no son apropiadas para un muestreo por importancia. De hecho, en estos casos, la varianza del estimador (2.3) puede ser infinita para algunas funciones h y al tener una variación tan amplia en los pesos $f(x_i)/g(x_i)$ se está dando más importancia a valores pequeños de x_i.

EJEMPLO **2.2** *Considere la siguiente función:*

$$f(x) = \frac{1}{2}e^{-|x|}.$$

La cual corresponde a la función de densidad de una variable aleatoria con distribución doble exponencial. La función de distribución correspondiente a esta familia paramétrica tiene la forma

$$F(x) = \frac{1}{2}e^x I_{(x\leq 0)} + \left(1 - \frac{e^{-x}}{2}\right) I_{(x>0)}.$$

La cual es difícil de invertir y por ello es complicado generar una muestra[2]*. Suponga que se quiere calcular $E[X^2]$ para esta distribución, la cual tiene soporte en $\mathbb{R}$. Esto es, se quiere calcular la integral*

$$\int_{-\infty}^{\infty} x^2 \frac{1}{2}e^{-|x|}dx.$$

Para ello, reescribiremos la expresion de la siguiente forma,

$$\int_{-\infty}^{\infty} x^2 \frac{\frac{1}{2}e^{-|x|}}{\frac{1}{\sqrt{8\pi}}e^{-x^2/8}} \frac{1}{\sqrt{8\pi}}e^{\frac{-x^2}{8}}dx,$$

donde

$$\frac{1}{\sqrt{8\pi}}e^{\frac{-x^2}{8}}$$

es la función propuesta g, de (2.2). La cual corresponde a la función de densidad de una variable aleatoria $N(0,4)$.

[2]Un método de simulación sencillo es el método de la función inversa, el cual consiste en invertir la función de acumulación de una variable aleatoria y evaluarla en una simulación de una variable uniforme.

Es claro que generar una muestra de g es más fácil que generarla a partir de f. Una vez generada la muestra $(x_1, \ldots, x_n)$ de una $N(0, 4)$ podemos estimar

$$\mathbb{E}\left[x^2 \frac{\frac{1}{2}e^{-|x|}}{\frac{1}{\sqrt{8\pi}}e^{-x^2/8}}\right]$$

utilizando la media empírica, i.e. con $\overline{h}_n$. Con ayuda de R se obtiene que el valor del estimador es 1.998898*, mientras que el valor real de la integral es 2. Esto quiere decir que se realizó una buena aproximación.*

De manera análoga se puede calcular cualquier momento de la variable aleatoria doble exponencial.

La Figura (2.2)*, muestra la función de densidad doble exponencial, $f(x)$, y la convergencia del estimador* (2.3) *junto con el intervalo de confianza para cada observación.*

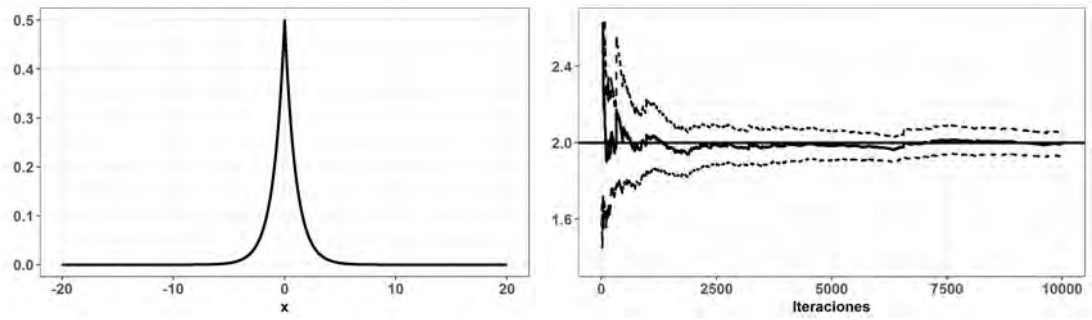

Figura 2.2: Izquierda: Gráfica de la función doble exponencial, $f(x)$. Derecha: Gráfica de convergencia del estimador $\overline{h}_n$, línea sólida. Las líneas punteadas representan los intervalos de confianza de cada observación y la línea sólida horizontal es el valor real de la integral

La forma en que el muestreo por importancia resuelve el problema de muestreo del método de Monte Carlo permite al usuario la elección de la función a muestrear. Sin embargo, es importante notar que una mejor elección de la función g corresponderá a una varianza menor para el cálculo de la integral. Por lo que puede llegar a ser un problema, debido a que la forma de encontrar la función g que optimice la aproximación no es muy clara.

La forma en que se puede implementar el método de muestreo por importancia se muestra en el Algoritmo 2.

Algoritmo 2 Método de *muestreo por importancia*

1: Proponer una función g y escribir la integral de la forma de (2.2)
2: Generar n observaciones independientes de $g(x)$.
3: Calcular el estimador $\overline{h}_n$ como

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

2.1.2. Muestreo de Gibbs

El muestreo de Gibbs (*Gibbs sampler*) es una técnica para generar variables aleatorias de una distribución marginal, sin tener que calcular la densidad. Este método se basa en propiedades elementales de las cadenas de Markov (Casella and George, 1992).

Suponga que se tiene la densidad conjunta $f(x, y_1, \ldots, y_m)$, y se tiene interés en obtener características de la densidad marginal

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_m) dy_1 \ldots dy_m. \tag{2.4}$$

Quizás la forma más natural para obtener cualquier característica sobre $f(x)$ sería calcular la íntegral (2.4). Sin embargo, la integración de (2.4) puede llegar a ser extremadamente difícil de realizar analítica o numéricamente. En estos casos, el muestreo de Gibbs provee una alternativa para obtener $f(x)$.

El muestreo de Gibbs permite generar una muestra $X_1, \ldots, X_m \sim f(x)$ sin la necesidad de contar con $f(x)$. Si además se simula una muestra lo suficientemente grande, las características de f que se deseen calcular se pueden obtener con el grado deseado de precisión.

Para entender mejor el funcionamiento del muestreo de Gibbs se analizará el caso de dos variables (Casella and George, 1992). Comenzando con la pareja de variables aleatorias (X, Y), el muestreo de Gibbs genera una muestra de $f(x)$ muestreando a partir de las distribuciones condicionales $f(x|y)$ y $f(y|x)$, distribuciones que regularmente son conocidas en los modelos estadísticos. Este proceso es realizado para generar una sucesión de variables aleatorias

$$Y_0', X_0', Y_1', X_1', \ldots, Y_k', X_k'. \tag{2.5}$$

Donde el valor inicial $Y_0' = y_0'$ es especificado, y los demás términos de (2.5) se obtienen de manera iterativa generando valores alternativamente de

$$X_j' \sim f(x|Y_j' = y_j'),$$
$$Y_{j+1}' \sim f(y|X_j' = x_j').$$

Nos referimos a la generación de (2.5) como una muestra de Gibbs.

Esto resulta, bajo condiciones generales, en que la distribución de X_k' converge a $f(x)$, la verdadera distribución marginal de X, cuando $k \to \infty$. Así, para una k lo suficientemente grande, la observación final en (2.5), $X_k' = x_k'$, es efectivamente un punto muestral de $f(x)$. De esta forma, si generamos m muestras de Gibbs independientes de longitud k, para una k lo suficientemente grande, y tomamos los valores finales, X_k', de cada secuencia tendremos una aproximación a una muestra independiente e idénticamente distribuida de $f(x)$.

EJEMPLO **2.3** *Binomial-Beta. Considere* $X|\theta \sim Binomial(n, \theta)$ *y* $\pi(\theta) = Beta(a, b)$. *Por el teorema de bayes obtenemos la densidad condicional*

$$\pi(\theta|x) \propto \theta^x (1-\theta)^{n-x} \theta^{a-1} (1-\theta)^{b-1}$$

La cual es nuevamente una distribución Beta con parámetros actualizados $a + x, b + n - x$*, debido a que la familia Binomial y la familia Beta son conjugadas. A partir de esta expresión se obtiene que la densidad conjunta de* (X, θ) *es*

$$f(x,\theta) = f(x|\theta)\pi(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

Por lo que la distribución marginal de X *tiene la forma*

$$f(x) = \binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)},\ x = 0, 1, \ldots, n.$$

La cual es conocida como la distribución Binomial-Beta. Generar números aleatorios de esta distribución es sumamente complicado. Así que se generarán a partir del muestreo de Gibbs.

Para ello se asignan los valores iniciales $(x^{(0)}, \theta^{(0)})$*. Para una* M *lo suficientemente grande, generaremos la secuencia de Gibbs* (2.5) *como sigue:*

$$x^{(i)} \sim f(x|\theta^{(i-1)}) = Binomial(n, \theta^{(i-1)}),$$
$$\theta^{(i)} \sim \pi(\theta|x^{(i)}) = Beta(a + x^{(i)}, b + n - x^{(i)}).$$

Este proceso genera una muestra de (x, θ)*. Esto implica que se puede aproximar la densidad de* X*,* $f(x)$*, con los valores* $x^{(i)}$*. Por ejemplo:*

$$f(x) = P(X = x) \approx (\#\ de\ \ X^{(i)} = x)/M,\ \ x = 0, 1, \ldots, n$$

Utilizando R, con valores iniciales $\theta^{(0)} = .5$*,* $x^{(0)} = 1$ *y parámetros de la distribución Beta* $a = 2, b = 4$ *y* $n = 16$ *se generó una muestra de la distribución marginal de* X*.*

A continuación se muestran los histogramas de los valores de θ *y* x *los cuales son una aproximación a las densidades* $\pi(\theta)$ *y* $f(x)$ *respectivamente.*

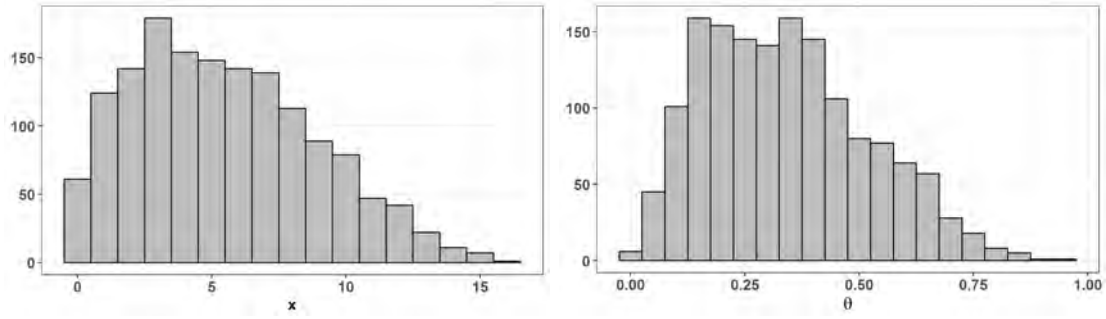

Figura 2.3: Histogramas de la muestra generada. Izquierda: Histograma de la aproximación de la densidad $f(x)$. Derecha: Histograma de la aproximación de la densidad $\pi(\theta)$

De esta forma si se desea obtener alguna característica de la distribución Binomial-Beta, como sus momentos, estas se pueden obtener mediante métodos que requieren de una muestra de la función objetivo, como el método de Monte Carlo.

El muestreo de Gibbs es un ejemplo del método de Monte Carlo vía cadenas de Markov ya que el punto $\theta^{(s+1)}$ depende de $\theta^{(s)}$. Y el límite, o distribución estacionaria, es $\pi(\theta|x)$

Algoritmo 3 Método de Muestreo de Gibbs

1: Iniciar $\theta^{(0)}$
2: Para la verosimilitud $f(x|\theta)$ y la distribución apriori $\pi(\theta)$
3: **for** iteración $i = 1, 2, \ldots, M$ **do**
4: $\quad x^{(i)} \sim f(x|\theta^{(i-1)})$
5: $\quad \theta^{(i)} \sim \pi(\theta|x^{(i)})$
6: **end for**

la distribución posterior. La cual es la base de la estadística bayesiana. El Algortimo 3 muestra la implementación del muestreo de Gibbs.

Es importante notar que θ puede ser un vector k-dimensional. Para este caso se supone que para todo $i = 1, \ldots, k$ se puede simular de la distribución posterior condicional $\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$. El Algoritmo 4 muestra la dinámica del muestreo para un vector k-dimensional.

Algoritmo 4 Método de Muestreo de Gibbs, con $\theta = (\theta_1, \ldots, \theta_k)$

1: Iniciar $(\theta_1^{(0)}, \ldots, \theta_k^{(0)})$ y $x^{(0)}$
2: **for** iteración $i = 1, 2, \ldots, M$ **do**
3: $\quad \theta_1^{(i)} \sim \pi(\theta_1|\theta_2^{i-1}, \ldots, \theta_k^{i-1}, x^{i-1})$
4: $\quad \theta_2^{(i)} \sim \pi(\theta_2|\theta_1^{i-1}, \theta_3^{i-1}, \ldots, \theta_k^{i-1}, x^{i-1})$
5: $\quad \theta_3^{(i)} \sim \pi(\theta_3|\theta_1^{i-1}, \theta_2^{i-1}, \theta_4^{i-1}, \ldots, \theta_k^{i-1}, x^{i-1})$
6: $\quad \ldots$
7: $\quad \theta_k^{(i)} \sim \pi(\theta_k|\theta_1^{i-1}, \theta_2^{i-1}, \ldots, \theta_{k-1}^{i-1}, x^{i-1})$
8: $\quad x^{(i)} \sim f(x|\theta_1^{(i)}, \ldots, \theta_k^{(i)})$
9: **end for**

2.2. Método de *LAPLACE*

El método de Laplace, a diferencia de los métodos presentados anteriormente, realiza una aproximación gaussiana alrededor de un parámetro, (Kass and Raftery, 1993). Este método permite evaluar integrales de la forma

$$I(\lambda) = \int_a^b f(t)e^{-\lambda g(t)}dt \tag{2.6}$$

En la cual se debe tener que λ es lo suficientemente grande, $g(t)$ es una función suave [3] con un mínimo local en $y^* \in (a, b)$ y $f(t)$ es una función suave. Esta integral puede ser la función generadora de momentos de la distribución de $g(T)$ cuando T tiene densidad $f(t)$, también se puede tratar, en el contexto de la estadística bayesiana, de la esperanza posterior de $f(t)$ o simplemente ser una integral que se desee calcular.

Cuando λ es lo suficientemente grande, la contribución al valor de la integral se encuentra esencialmente en una vecindad alrededor de y^*, ya que es ahí cuando $\exp\{-\lambda g(t)\}$ tiene un valor máximo. Por ello, es de especial interés estudiar el comportamiento de $g(t)$ alrededor de y^*, lo cual se lleva a cabo mediante la aproximación por series de Taylor

$$g(y) = g(y^*) + g'(y^*)(y - y^*) + g''(y^*)\frac{(y - y^*)^2}{2} + \ldots = \sum_{n=0}^{\infty} \frac{g^{(n)}(y^*)}{n!}(y - y^*)^n$$

Ya que y^* es un mínimo local, se tiene que $g'(y^*) = 0$ y $g''(y^*) > 0$, por lo que se cumple

$$g(y) - g(y^*) = \frac{g''(y^*)}{2}(y - y^*)^2 + \ldots$$

Ahora, reescribiendo la integral (2.6) y para una vecindad alrededor de y^* tenemos

$$\begin{aligned}
I(\lambda) &\approx e^{-\lambda g(y^*)} \int_{y^*-\epsilon}^{y^*+\epsilon} f(t)e^{-\lambda(g(t)-g(y^*))}dt \\
&\approx e^{-\lambda g(y^*)} f(y^*) \int_{y^*-\epsilon}^{y^*+\epsilon} e^{-\lambda(g(t)-g(y^*))}dt + e^{-\lambda g(y^*)} f'(y^*) \int_{y^*-\epsilon}^{y^*+\epsilon} (t - y^*)e^{-\lambda(g(t)-g(y^*))}dt \\
&\approx e^{-\lambda g(y^*)} f(y^*) \int_{y^*-\epsilon}^{y^*+\epsilon} e^{-\lambda(g'(y^*)(t-y^*)+\frac{1}{2}g''(y^*)(t-y^*)^2)}dt \\
&\approx e^{-\lambda g(y^*)} f(y^*) \int_{-\infty}^{\infty} e^{-\frac{\lambda}{2}g''(y^*)(t-y^*)^2}dt
\end{aligned}$$

Para el cálculo anterior, se realizó una aproximación de la función f alrededor de y^*. Además, se puede identificar el kernel de una distribución normal con media y^* y varianza $(\lambda g''(y^*))^{-1}$ lo cual simplifica el resultado y se obtiene una forma cerrada

$$I(\lambda) \approx e^{-\lambda g(y^*)} f(y^*) \sqrt{\frac{2\pi}{\lambda g''(y^*)}}$$

[3] Una función f se dice suave o de clase C^∞ si existen todas sus derivadas de cualquier orden.

En la literatura se muestra que para este tipo de aproximaciones se tiene un término de error que es $O(\frac{1}{\lambda})$, por lo que el resto de la aproximación de Taylor,

$$\sum_{n=3}^{\infty} \frac{g^{(n)}(y^*)}{n!}(y-y^*)^n$$

es menor o igual que c/λ. Lo cual significa que para λ grande el error es muy pequeño.

Dada la naturaleza de este estimador, la extensión a una mayor dimensión, cuando $\overline{y}^* \in \mathbb{R}^m$, es sencilla de realizar y es completamente análoga a la presentada. Para este caso, la integración es realizada sobre el dominio $m-$dimensional, obteniendo

$$I(\lambda) \approx e^{-\lambda g(\overline{y}^*)} f(\overline{y}^*)(\frac{2\pi}{\lambda})^{\frac{m}{2}}(det\Sigma_{\overline{y}^*})^{\frac{1}{2}}$$

EJEMPLO **2.4** *Considere la función gamma, definida como*

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt.$$

Es útil recordar que para $\lambda \in \mathbb{N}$ *se tiene que* $\Gamma(\lambda+1) = \lambda!$. *Utilizando el método de Laplace tenemos*

$$\begin{aligned}
\Gamma(\lambda+1) &= \int_0^{\infty} t^{\lambda}e^{-t}dt \\
&= \int_0^{\infty} e^{\lambda \ln(t)}e^{-t}dt \\
&= \int_0^{\infty} e^{-\lambda\left(\frac{t}{\lambda}-\ln(t)\right)}dt \quad (Haciendo\ \ t=\lambda z) \\
&= \lambda\int_0^{\infty} e^{-\lambda\left(z-\ln(\lambda z)\right)}dz \\
&= \lambda e^{\lambda\ln(\lambda)}\int_0^{\infty} e^{-\lambda\left(z-\ln(z)\right)}dz \\
&= \lambda^{\lambda+1}\int_0^{\infty} e^{-\lambda\left(z-\ln(z)\right)}dz.
\end{aligned}$$

Para este caso se tiene que $f(z) \equiv 1$ *y* $g(z) = z - \ln(z)$. *Se puede observar que* g *tiene un valor mínimo sobre* $(0,\infty)$ *en* $z = 1$, *ya que* $g'(1) = 0$, $g''(1) = 1 > 0$. *Por último, el método de Laplace genera el siguiente resultado*

$$\begin{aligned}
\Gamma(\lambda+1) &= \lambda^{\lambda+1}e^{-\lambda g(t^*)}\sqrt{\frac{2\pi}{\lambda g''(t^*)}}\{1+O(\frac{1}{\lambda})\} \\
&= \lambda^{\lambda+\frac{1}{2}}e^{-\lambda}\sqrt{2\pi}\{1+O(\frac{1}{\lambda})\} \quad (con\ \ t^*=1).
\end{aligned}$$

Este resultado es conocido como la fórmula de Stirling. *La precisión de ésta aproximación depende del grado del polinomio de Taylor considerado. Este cálculo se vuelve más sencillo*

cuando $f(x)$ es constante, obteniendo

$$I = \int_a^b e^{-\lambda g(y)} dy$$
$$= e^{-\lambda g(t^*)} \sqrt{\frac{2\pi}{\lambda g''(t^*)}} \{1 + \frac{5\rho_3^* - 3\rho_4^*}{24\lambda} + O(\frac{1}{\lambda^2}).\}$$

Donde $\rho_3^* = \frac{g^{(3)}(t^*)}{\{g''(t^*)\}^{\frac{3}{2}}}$ *y* $\rho_4^* = \frac{g^{(4)}(t^*)}{\{g''(t^*)\}^2}$. *Así, la fórmula de* Stirling *tiene la forma*

$$\Gamma(\lambda + 1) = \lambda^{\lambda + \frac{1}{2}} e^{-\lambda} \sqrt{2\pi} \{1 + \frac{1}{12\lambda} + O(\frac{1}{\lambda^2})\} -$$

La cual es más precisa, incluso para valores pequeños de λ como muestra la siguiente tabla del $\log(\Gamma(\lambda + 1))$.

λ	Valor Exacto	Fórmula de Stirling	Stirling mejorada
2	0.6931472	0.6518048	0.6926268
4	3.1780538	3.1572615	3.1778807
8	10.6046029	10.5941899	10.6045527
16	30.6718601	30.6666508	30.6718456
32	205.1681995	205.1668957	205.1681970

Tabla 2.1: Estimación de la función gamma vía método de Laplace.

Se puede observar que la precisión del modelo depende en gran medida de los términos utilizados provenientes del polinomio de Taylor, como se muestra en la tabla (2.1). De esta forma, las características con las que cuenten las funciones $f(x)$ y $g(x)$ determinarán el tamaño de la precisión que será posible alcanzar, ya que para funciones cuya descomposición por el polinomio de Taylor tengan una forma complicada será muy costoso realizar el cálculo, en términos computacionales.

La Figura (2.4) muestra las funciones gamma, la función gamma estimada mediante el método de Laplace y la función estimada mediante el método mejorado. Como se puede observar la última estima de buena manera la función gamma, sin embargo, ambas estimaciones son pobres para valores cercanos a uno.

Por otro lado, basados en el enfoque bayesiano, una forma de ver la densidad posterior, para un conjunto de variables aleatorias, $X_1, \ldots, X_n$, condicionalmente independientes sobre el parámetro θ con distribución $f(x|\theta)$ y distribución a priori $\pi(\theta)$, es como sigue

$$\pi(\theta|x) \propto e^{l(\theta)} \pi(\theta)$$

donde $l(\theta) = \log(L(\theta))$ es la función de log-verosimilitud. Si se considera

$$\bar{l}_n(\theta) = \frac{l(\theta)}{n} = \frac{1}{n} \sum_{i=1}^{n} \log(f(X_i|\theta))$$

Por la ley de los grandes números tenemos que cuando $n \to \infty$

$$\bar{l}_n(\theta) \to \mathbb{E}_\theta[\log(f(X|\theta))] = -H(\theta)$$

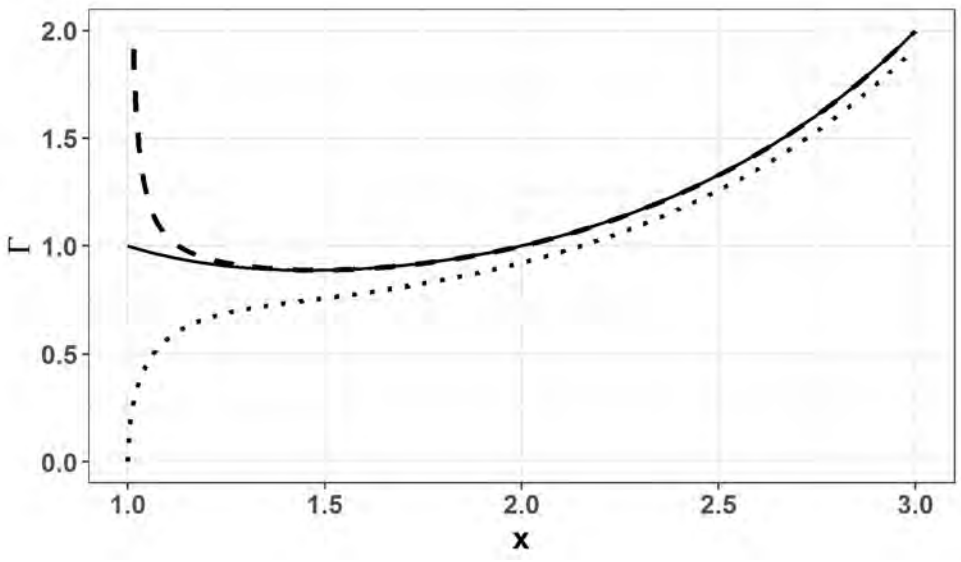

Figura 2.4: Funciones gamma estimadas. La línea sólida corresponde al valor exacto de la función gamma. La línea punteada corresponde al método de Laplace y la línea discontinua corresponde al método de Laplace mejorado.

donde $H(\theta)$ es la entropía[4] de la densidad $f(x|\theta)$.

De esta forma, la variación en la densidad posterior, para una n lo suficientemente grande, será dominada por la función de log-verosimilitud. Así, se puede implementar el método de Laplace a la función posterior

$$\pi(\theta|x) \propto e^{n\bar{l}_n(\theta)}\pi(\theta) \tag{2.7}$$

Para la implementación de este método se puede seguir lo descrito en el Algoritmo 5.

Algoritmo 5 Método de Integración de Laplace

1: Escribir la integral en la forma de (2.7).
2: Identificar las funciones $f(x)$ y $g(x)$.
3: Utilizar el resultado para un error de orden $O(\frac{1}{\lambda})$
4: Si se desea una aproximación más precisa, aproximar $f(x)$ y $g(x)$ considerando más términos del polinomio de Laplace

2.3. Variación de Bayes

La inferencia *variacional* bayesiana generaliza la idea del método de Laplace. En esta metodología se desea encontrar una densidad que sea lo más similar posible a la verdadera distribución posterior. Esta similitud se basa en la divergencia de *Kullback-Liebler (KL)*

Definición 2.1 *La divergencia de* Kullback-Liebler *(KL) para las funciones de densidad p y q, de variables aleatorias continuas, se define como*

$$D_{KL}(p||q) := \int_{-\infty}^{\infty} p(x)\ln(\frac{p(x)}{q(x)})dx. \tag{2.8}$$

[4]Es una medida de incertidumbre de la distribución de probabilidad de una variable aleatoria.

La divergencia KL es una medida no simétrica de la similitud o diferencia entre dos funciones de densidad de probabilidad p y q.

En este método, la distribución de aproximación, $\hat{f}$, no tiene una forma dada, sino que se restringe funcionalmente utilizando la suposición de independencia condicional

$$f(\theta|x) \approx \hat{f}(\theta|x) = \hat{f}(\theta_1|x)\hat{f}(\theta_2|x)\dots\hat{f}(\theta_q|x)$$

Donde $\hat{f}(\theta|x)$ es la variante funcional usada en el proceso de optimización y que da lugar a la distribución de aproximación de θ dados los datos.

Para hacer más tratable el proceso, el método no minimiza la divergencia KL en su forma *original*, (2.8), de $f(\theta|x)$ a $\hat{f}(\theta|x)$ sino la divergencia KL *inversa*, $KL(\hat{f}(\theta|x)||f(\theta|x))$, de $\hat{f}(\theta|x)$ a $f(\theta|x)$.

El siguiente teorema proporciona una expresión para $\hat{f}_i(\theta_i)$ que minimiza la divergencia de KL y cuya demostración se puede consultar en (mídl, 2004).

Teorema 2.1 ***(Variational Bayes)*** *Sea $f(\theta|x)$ la función de densidad posterior del parámetro multivariado θ, tal que $\theta = [\theta'_1, \dots, \theta'_q]'$. Sea $\hat{f}(\theta|x)$ una función de densidad restringida al conjunto de distribuciones condicionalmente independientes sobre $\theta'_1, \dots, \theta'_q$*

$$\hat{f}(\theta|x) = \hat{f}(\theta_1, \dots, \theta_q|x) = \prod_{i=1}^{q} \hat{f}_i(\theta_i|x)$$

Entonces, el mínimo de la divergencia KL

$$\hat{f}(\theta|x) = \arg\min_{\hat{f}} KL(\hat{f}(\theta|x)||f(\theta|x))$$

esta dado por

$$\hat{f}_i(\theta_i|x) \propto \exp\left(E_{\hat{f}_{/i}(\theta_{/i}|x)}\left[\ln f(\theta, x)\right]\right), \quad i = 1, \dots, q \tag{2.9}$$

Donde $\theta_{/i}$ denota el complemento de θ_i en θ, y $\hat{f}_{/i}(\theta_{/i}|x) = \prod_{\substack{i=1 \\ j\neq i}}^{q} \hat{f}_j(\theta_j|x)$

Estos supuestos tienen consecuencias que pueden ser vistas como desventajas dentro del método; la independencia condicional trae consigo que el método sólo pueda ser usado para modelos con más de un parámetro. Por otro lado, el uso de la divergencia KL *inversa* es menos óptima que la divergencia *original*, además un mínimo en la divergencia *inversa* puede no ser único.

Sin embargo, estas desventajas pueden ser sobrepasadas por las ventajas computacionales que presenta el método de VB; optimización funcional, ya que tiene una forma libre en $\hat{f}$, que cuenta con una solución analítica, y los parámetros de la distribución posterior aproximada óptima pueden ser evaluados usando algorítmos alternativos como el EM (*Expectation maximization*)[5].

[5] Ver apéndice (A)

2.3.1. Variación de Bayes vía el algoritmo EM

La clave del método variacional es aproximar la integral con una forma más simple de tratar, formando un límite superior o inferior. Por lo que la integración se traduce en un problema de optimización de límites; haciendo el límite tan estrecho como sea posible hacia el verdadero valor.

Suponga que las variables observadas son $\mathbf{x} = (x_1, \ldots, x_n)$, y existen variables ocultas denotadas por $\mathbf{y} = (y_1, \ldots, y_n)$ y $\theta = (\theta_1, \ldots, \theta_q)$ denota los parámetros. Se supone una distribución a priori para θ, $f(\theta)$. Entonces, la verosimilitud marginal $f(\mathbf{x})$ se puede acotar inferiormente introduciendo una distribución sobre la variable latente[6], $\mathbf{y}$, y el parámetro θ la cual tiene el mismo soporte. Esto es, la integral de interés

$$f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\theta) f(\theta) d\theta$$

se puede acotar inferiormente de la siguiente forma

$$\begin{aligned}
\ln f(\mathbf{x}) &= \ln \int_{\Theta} f(\mathbf{x}|\theta) f(\theta) d\theta \\
&= \ln \int_{\Theta} \int_{\chi} f(\mathbf{x}, \mathbf{y}|\theta) f(\theta) d\mathbf{y} d\theta \\
&= \ln \int_{\Theta} \int_{\chi} f(\mathbf{x}, \mathbf{y}, \theta) d\mathbf{y} d\theta \\
&= \ln \int_{\Theta} \int_{\chi} \frac{f(\mathbf{x}, \mathbf{y}, \theta)}{q(\mathbf{y}, \theta)} q(\mathbf{y}, \theta) d\mathbf{y} d\theta \\
&\geq \int_{\Theta} \int_{\chi} \ln \left(\frac{f(\mathbf{x}, \mathbf{y}, \theta)}{q(\mathbf{y}, \theta)} \right) q(\mathbf{y}, \theta) d\mathbf{y} d\theta, \qquad (2.10)
\end{aligned}$$

donde (2.10) se debe a la desigualdad de Jensen[7]. Maximizando el límite inferior con respecto a la distribución libre $q(\mathbf{y}, \theta)$, se tiene que:

$$q(\mathbf{y}, \theta) = f(\mathbf{y}, \theta|\mathbf{x}),$$

lo cual implica la igualdad en (2.10). Sin embargo esto no simplifica el problema ya que evaluar la distribución posterior exacta, $f(\mathbf{y}, \theta|\mathbf{x})$, requiere conocer la constante de normalización, i.e. $f(\mathbf{x})$, la integral de interés. Por lo que el método se limita a una distribución posterior más simple, una aproximación factorizada:

$$q(\mathbf{y}, \theta) \approx q_{\mathbf{y}}(\mathbf{y}) q_{\theta}(\theta)$$

[6]Una variable latente, o variable oculta, es aquella que no es observada directamente sino que es inferida a partir de las demás variables que se observan.

[7]Sea $f(x)$ una función cóncava, entonces para la variable aleatoria X se tiene que $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$

con la cual partiendo de (2.10) se tiene:

$$\begin{aligned}
\ln f(\mathbf{x}) &\geq \int_{\Theta}\int_{\chi} \ln\left(\frac{f(\mathbf{x},\mathbf{y},\theta)}{q(\mathbf{y},\theta)}\right) q(\mathbf{y},\theta)d\mathbf{y}d\theta \\
&= \int_{\Theta}\int_{\chi} \ln\left(\frac{f(\mathbf{x},\mathbf{y},\theta)}{q_{\mathbf{y}}(\mathbf{y})q_{\theta}(\theta)}\right) q_{\mathbf{y}}(\mathbf{y})q_{\theta}(\theta)d\mathbf{y}d\theta \\
&= \int_{\Theta} q_{\theta}(\theta)\int_{\chi} q_{\mathbf{y}}(\mathbf{y}) \ln\left(\frac{f(\mathbf{x},\mathbf{y},\theta)}{q_{\mathbf{y}}(\mathbf{y})q_{\theta}(\theta)}\right) d\mathbf{y}d\theta \\
&= \int_{\Theta} q_{\theta}(\theta)\int_{\chi} q_{\mathbf{y}}(\mathbf{y}) \ln\left(\frac{f(\mathbf{x},\mathbf{y}|\theta)f(\theta)}{q_{\mathbf{y}}(\mathbf{y})q_{\theta}(\theta)}\right) d\mathbf{y}d\theta \\
&= \int_{\Theta} q_{\theta}(\theta)\int_{\chi} q_{\mathbf{y}}(\mathbf{y}) \left[\ln\left(\frac{f(\mathbf{x},\mathbf{y}|\theta)}{q_{\mathbf{y}}(\mathbf{y})}\right) + \ln\left(\frac{f(\theta)}{q_{\theta}(\theta)}\right)\right] d\mathbf{y}d\theta \\
&= \mathscr{F}(q_{\mathbf{y}}(\mathbf{y}), q_{\theta}(\theta)) \\
&= \mathscr{F}(q_{\mathbf{y}_1}(y_1), \ldots, q_{\mathbf{y}_n}(y_n), q_{\theta}(\theta))
\end{aligned} \tag{2.11}$$

donde la última igualdad es una consecuencia de la independencia e idéntica distribución de los datos $\mathbf{x}$.

El algoritmo de variación de Bayes maximiza iterativamente $\mathscr{F}$ en (2.11) con respecto a las distribuciones libres $q_{\mathbf{y}}(\mathbf{y})$ y $q_{\theta}(\theta)$. El siguiente teorema provee las ecuaciones de actualización para el aprendizaje variacional bayesiano.

Teorema 2.2 ***Variación de Bayes vía maximización de esperanzas (VBEM)*** *Considera un conjunto* $\boldsymbol{x} = (x_1, \ldots, x_n$ *con parámetros* $\theta = (\theta_1, \ldots, \theta_q$ *y variables latentes* $\boldsymbol{y} = y_1, \ldots, y_n$. *Una cota inferior sobre la log-verosimilitud marginal es*

$$\mathscr{F}(q_{\boldsymbol{y}}(\boldsymbol{y}), q_{\theta}(\theta)) = \int_{\Theta} q_{\theta}(\theta)\int_{\chi} q_{\boldsymbol{y}}(\boldsymbol{y}) \left[\ln\left(\frac{f(\boldsymbol{x},\boldsymbol{y}|\theta)}{q_{\boldsymbol{y}}(\boldsymbol{y})}\right) + \ln\left(\frac{f(\theta)}{q_{\theta}(\theta)}\right)\right] d\boldsymbol{y}d\theta$$

y ésta puede ser iterativamente optimizada realizando actualizaciones para cada iteración i, *en dos pasos,*

$$\begin{aligned}
\textbf{\textit{Paso E}}&: q_{y_j}^{(i+1)}(y_j) = \frac{1}{Z_{y_j}} \exp \int q_{\theta}^{(i)}(\theta) \ln f(x_j, y_j|\theta)d\theta \;\; \forall j \in \{1, \ldots, n\}, \\
\textbf{\textit{Paso M}}&: q_{\theta}^{(i+1)}(\theta) = \frac{1}{Z_{\theta}} p(\theta) \exp \int q_{y}^{(i+1)}(y) \ln f(x, y|\theta)dy,
\end{aligned}$$

donde Z *es la constante de normalización y*

$$q_{y}^{(i+1)}(y) = \prod_{i=1}^{n} q_{y_j}^{(i+1)}(yi).$$

Más aún, las reglas de actualización convergen a un máximo local de $\mathscr{F}(q_{\boldsymbol{y}}(\boldsymbol{y}), q_{\theta}(\theta))$.

La prueba a este teorema se puede consultar en (Beal, 2003). Se puede observar que las expresiones provistas por el teorema (2.2) proveen del algoritmo necesario para (2.9).

Para comprender el proceso de la aproximación se implementa el método para el caso de una distribución normal univariada utilizando el resultado de (2.9).

EJEMPLO **2.5** ***Normal Univariada.*** *Suponga que tenemos un conjunto de datos* $x = (x_1, \ldots x_n)$ *de una distribución con media* μ *y precisión* τ*. Dado el conjunto de datos la función de verosimilitud tiene la forma*

$$p(x|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\}.$$

Para los parámetros μ *y* τ *se puede tener cierta idea sobre su distribución, para facilitar el análisis, se introduce las siguientes distribuciones a priori conjugadas*

$$p(\mu|\tau) = N(\mu|\mu_0, (\lambda_0\tau)^{-1}),$$
$$p(\tau) = Gamma(\tau|\alpha_0, \beta_0).$$

Recordando que se tiene interés en estimar la distribución posterior $q(\mu,\tau)$ *y asumiendo, de acuerdo con las hipótesis del método, que ésta distribución se puede factorizar de la siguiente forma*

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau).$$

La cual no es, necesariamente, como la verdadera distribución posterior se factoriza. Para encontrar el valor óptimo para el factor $q_\mu(\mu)$ *se aplica la expresión* (2.9)

$$\begin{aligned}
\log\, q^*_\mu(\mu) &= \mathbb{E}[\log\, p(x,\mu,\sigma)] = \mathbb{E}_\tau[\log[\, p(x|\mu,\tau)p(\mu|\tau)p(\tau)]] \\
&= \mathbb{E}_\tau[\log\, p(x|\mu,\tau) + \log\, p(\mu|\tau) + \log\, p(\tau)] \\
&= \mathbb{E}_\tau[\log\, p(x|\mu,\tau) + \log\, p(\mu|\tau)] + C \\
&= -\frac{\mathbb{E}[\tau]}{2}\left(\lambda_0(\mu-\mu_o)^2 + \sum_{j=1}^{n}(x_i-\mu)^2\right) + C
\end{aligned}$$

El siguiente paso es completar el cuadrado sobre μ *para obtener la forma de una distribución gausiana,* $N(\mu|\mu_n, \lambda_n^{-1})$*, para* $q_\mu(\mu)$ *donde*

$$\mu_n = \frac{\lambda_0\mu_0 + n\overline{x}}{\lambda_0 + n} \quad , \quad \lambda_n = (\lambda_0 + n)\mathbb{E}[\tau]$$

Un análisis similar para $q_\tau(\tau)$ *muestra que sigue una distribución* $Gamma(\tau|\alpha_n, \beta_n)$ *donde*

$$\begin{aligned}
\alpha_n &= \alpha_0 + \frac{n+1}{2} \\
\beta_n &= \beta_0 + \frac{1}{2}\mathbb{E}\left[\sum_{j=i}^{n}(x_j-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right].
\end{aligned}$$

Es importante notar que se puede usar las ecuaciones anteriormente derivadas, (2.9)*, para iterar computacionalmente estimaciones más refinadas para los parámetros del modelo. Se puede observar que los parámetros de* $q_\mu(\mu)$ *dependen del valor de la media de* τ *y viceversa.*

2.4. Densidad de emisión

La Densidad de emisión, (ADF), es una técnica general para aproximar distribuciones posteriores en modelos estadísticos. Este método aplica cuando se tiene una distribución conjunta $p(x, \theta)$, donde x ha sido observado y θ es una variable latente, y se requiere conocer la distribución posterior sobre θ, $p(\theta|x)$, así como la probabilidad de los datos observados, también llamada evidencia del modelo, $p(x)$. Esto es útil para la estimación y para la selección del modelo, respectivamente.

Considerando la distribución conjunta de θ y n observaciones independientes $x = (x_1, \ldots, x_n)$

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

la cual al momento de aplicar ADF se escribe como un producto de términos que realizan la aproximación

$$p(x, \theta) \approx \prod_{i=0}^{n} t_i(\theta),$$

donde $t_0(\theta) = p(\theta)$ y $t_i(\theta) = p(x_i|\theta)$. Por otro lado, para realizar la aproximación se debe eligir una familia de distribuciones y partiendo del hecho de que se desea una aproximación a la posterior que sea simple de tratar, la distribución de aproximación será aquella que tenga un vector de momentos finito. Por ello, una muy buena opción para la distribución de aproximación es una perteneciente a la familia exponencial, cuya estructura reduce el método a una igualación de momentos.

El siguiente paso es incorporar iterativamente los términos t_i dentro de la distribución posterior aproximada. En cada iteración, i, se pasa de una *vieja* $q^{/i}(\theta)$[8] a una *nueva* $q^i(\theta)$. Se inicializa con $q^0(\theta) = 1$. Incorporar el término de la distribución a priori no requiere de una aproximación, por otro lado, para incorporar un término más complicado, $t_i \;\; i > 0$, se debe tomar la distribución posterior exacta, la cual está dada por

$$\hat{p}(\theta) = \frac{t_i(\theta) q^{/i}(\theta)}{\int_\theta t_i(x) q^{/i}(x) dx}. \tag{2.12}$$

Además se debe minimizar la divergencia, $KL(\hat{p}(\theta)||q(\theta))$, sujeto a la condición que $q(\theta)$ pertenece a la familia de aproximación.

Para entender mejor la metodología considere el caso de una mezcla de dos distribuciones normales

$$\begin{aligned} p(x|\theta) &= (1-\omega) N(x|\theta, \mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I}) \qquad , \omega \in (0, 1). \\ N(x|\mathbf{m}, \mathbf{V}) &= \frac{1}{|2\pi\mathbf{V}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x-\mathbf{m})^T \mathbf{V}^{-1} (x-\mathbf{m}) \right\}, \end{aligned} \tag{2.13}$$

con $\mathbf{I}$ la matríz idéntidad.

Para este problema es razonable proponer una distribución inicial para $\theta = (\theta_1, \ldots, \theta_d)$ Normal $d-$variada:

$$p(\theta) \sim N(\mathbf{0}, 100\mathbf{I}_d).$$

[8] $q^{/i}(\theta) = \frac{q(\theta)}{t_i(\theta)} = \prod_{\substack{j=1 \\ j \neq i}}^{n} t_j(\theta)$

Dado que se está trabajando con dstribuciones normales, las funciones de aproximación con distribución gaussiana esférica[9] proporcionan la mejor opción. Por lo que la distribución posterior aproximada es

$$q(\theta) \sim N(\mathbf{m}_\theta, \nu_\theta \mathbf{I}).$$

El siguiente paso es incorporar cada uno de los términos $t_i(\theta)$ tomando la distribución posterior exacta como en (2.12) y minimizar la divergencia $KL(\hat{p}(\theta)||q^{'}(\theta))$ sujeto a la condición de que $q^{'}(\theta)$ sigue una distribución gaussiana. Para ello, es necesario igualar a cero el gradiente de la divergencia KL con respecto a $(\mathbf{m}_\theta, \nu_\theta)$ de la siguiente forma:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \nu_q} KL(\hat{p}(\theta)||q^{'}(\theta)) \\ 0 &= \frac{\partial}{\partial \mathbf{m}_q} KL(\hat{p}(\theta)||q^{'}(\theta)) \end{aligned}$$

utilizando la forma exponencial de la distribución normal[10] para $\hat{p}(\theta)$ y $q^{'}(\theta)$,

$$\begin{aligned} \hat{p}(\theta) &\propto \exp\left\{ (\mathbf{m}_p^{'}\nu_p^{-1}, \nu_p^{-1})(\theta, -\frac{\theta^{'}\theta}{2})^{'} - \frac{\mathbf{m}_p^{'}\nu_p^{-1}\mathbf{m}_p}{2} + \ln\left(\frac{1}{|\nu_p \mathbf{I}|^{\frac{1}{2}}}\right)\right\} \\ q^{'}(\theta) &\propto \exp\left\{ (\mathbf{m}_q^{'}\nu_q^{-1}, \nu_q^{-1})(\theta, -\frac{\theta^{'}\theta}{2})^{'} - \frac{\mathbf{m}_q^{'}\nu_q^{-1}\mathbf{m}_q}{2} + \ln\left(\frac{1}{|\nu_q \mathbf{I}|^{\frac{1}{2}}}\right)\right\} \end{aligned}$$

se tiene

$$\begin{aligned} \frac{\partial}{\partial \mathbf{m}_q} KL(\hat{p}(\theta)||q^{'}(\theta)) &= \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\left((\mathbf{m}_p^{'}\nu_p^{-1} - \mathbf{m}_q^{'}\nu_q^{-1}, \nu_p^{-1} - \nu_q^{-1})(\theta, -\frac{\theta^{'}\theta}{2})^{'}\right)d\theta \\ &+ \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\frac{\mathbf{m}_q^{'}\nu_q^{-1}\mathbf{m}_q}{2}d\theta - \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\frac{\mathbf{m}_p^{'}\nu_p^{-1}\mathbf{m}_p}{2}d\theta \\ &+ \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\ln\left((\nu_p)^{-\frac{n}{2}}\right)d\theta - \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\ln\left((\nu_q)^{-\frac{n}{2}}\right)d\theta \\ &= \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\left((\mathbf{m}_p^{'}\nu_p^{-1} - \mathbf{m}_q^{'}\nu_q^{-1}, \nu_p^{-1} - \nu_q^{-1})(\theta, -\frac{\theta^{'}\theta}{2})^{'}\right)d\theta \\ &+ \frac{\partial}{\partial \mathbf{m}_q}\left(\frac{\mathbf{m}_q^{'}\nu_q^{-1}\mathbf{m}_q}{2}\right) \\ &= \frac{\partial}{\partial \mathbf{m}_q}\left(\frac{\mathbf{m}_q^{'}\nu_q^{-1}\mathbf{m}_q}{2}\right) - \frac{\partial}{\partial \mathbf{m}_q}\int \hat{p}(\theta)\theta\mathbf{m}_q^{'}\nu_q^{-1}d\theta \\ &= \nu_q^{-1}\mathbf{m}_q - \nu_q^{-1}\int \theta\hat{p}(\theta)d\theta = 0 \end{aligned}$$

Por lo que se debe cumplir que

$$\mathbf{m}_q = \int \theta\hat{p}(\theta)d\theta = \mathbb{E}_{\hat{p}}[\theta]$$

[9]Un vector $\mathbf{x} = (x_1, \ldots, x_n)$ se dice que tiene distribución gaussiana esférica si $\mathbf{x} \sim N(\mu, \sigma\mathbf{I}_n)$, con $\mathbf{I}_n$ la matríz idéntidad.

[10]ver (B).

De igual forma para ν_q

$$\begin{aligned}\frac{\partial}{\partial \nu_q} KL(\hat{p}(\theta)||q^{'}(\theta)) &= \frac{\partial}{\partial \nu_q}\int \hat{p}(\theta)\frac{\theta^{'}\theta}{2}\nu_q^{-1}d\theta - \frac{\partial}{\partial \nu_q}\int \hat{p}(\theta)\theta \mathbf{m}_q^{'}\nu_q^{-1}d\theta \\ &+ \frac{\partial}{\partial \nu_q}\int \hat{p}(\theta)\frac{\mathbf{m}_q^{'}\nu_q^{-1}\mathbf{m}_q}{2}d\theta - \frac{\partial}{\partial \nu_q}\int \hat{q}(\theta)\ln\left((\nu_p)^{-\frac{n}{2}}\right)d\theta \\ &= \int \hat{p}(\theta)\theta\mathbf{m}_q^{'}\nu_q^{-2}d\theta - \int \hat{p}(\theta)\frac{\theta^{'}\theta}{2}\nu_q^{-2}d\theta + \frac{n}{2\nu_q} - \frac{\mathbf{m}_q^{'}\nu_q^{-2}\mathbf{m}_q}{2} \\ &= \mathbf{m}_q^{'}\nu_q^{-2}\mathbb{E}_{\hat{p}}[\theta] - \frac{\nu_q^{-2}}{2}\mathbb{E}_{\hat{p}}[\theta^{'}\theta] + \frac{n}{2\nu_q} - \frac{\mathbf{m}_q^{'}\nu_q^{-2}\mathbf{m}_q}{2} \\ &= 0\end{aligned}$$

Por lo que se sigue

$$\mathbb{E}_{\hat{p}}[\theta^{'}\theta] = \int \hat{p}(\theta)\theta^{'}\theta d\theta = 2\mathbf{m}_q^{'}\mathbb{E}_{\hat{p}}[\theta] - \mathbf{m}_q^{'}\mathbf{m}_q + n\nu_q = n\upsilon_q + \mathbf{m}_q^{'}\mathbf{m}_q.$$

Debido a que

$$\mathbb{E}_{\hat{p}}[\theta] = \mathbf{m}_q.$$

Por lo tanto, para minimizar la divergencia $KL(\hat{p}(\theta)||q(\theta))$ se tienen las siguientes restricciones sobre los momentos

$$\begin{aligned}\mathbb{E}_{\hat{p}}[\theta] &= \mathbf{m}_q = \mathbb{E}_{q'}[\theta], \\ \mathbb{E}_{\hat{p}}[\theta^{'}\theta] &= n\nu_q + \mathbf{m}_q^{'}\mathbf{m}_q = \mathbb{E}_q^{'}[\theta^{'}\theta].\end{aligned}$$

Como se acaba de mostrar la distribución gaussiana esférica está caracterizada por las esperanzas $(\mathbb{E}[\theta], \mathbb{E}[\theta^{'}\theta])$. Para otros miembros de la familia exponencial se encuentran diferentes condiciones sobre las esperanzas, para calcularlas es útil tener presente las siguientes ecuaciones, (Minka, 2001):

$$\begin{aligned}Z(\mathbf{m}_q, \nu_q) &= \int_\Theta t(\theta)q(\theta)d\theta \\ &= \int_\Theta \frac{t(\theta)}{(2\pi\nu_q)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\nu_q}(\theta-\mathbf{m}_q)^{'}(\theta-\mathbf{m}_q)\right)d\theta \\ \nabla_m \ln Z(\mathbf{m}_q, \nu_q) &= \frac{1}{Z}\int_\Theta \frac{(\theta-\mathbf{m}_q)}{\nu_q}\frac{t(\theta)}{(2\pi\nu_q)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\nu_q}(\theta-\mathbf{m}_q)^{'}(\theta-\mathbf{m}_q)\right)d\theta \\ &= \frac{\mathbb{E}_{\hat{p}}[\theta]}{\nu_q} - \frac{\mathbf{m}_q}{\nu_q} \\ \mathbb{E}_{\hat{p}}[\theta] &= \mathbf{m}_q + \nu_q\nabla_m \ln Z(\mathbf{m}_q, \nu_q) \qquad (2.14) \\ \mathbb{E}_{\hat{p}}[\theta^{'}\theta] - \mathbb{E}_{\hat{p}}[\theta]^{'}\mathbb{E}_{\hat{p}}[\theta] &= n\nu_q - \nu_q^2\left(\nabla_m^{'}\nabla_m - 2\nabla_\nu \ln Z(\mathbf{m}_q, \nu_q)\right) \qquad (2.15)\end{aligned}$$

Estas relaciones son propiedades de una distribución $q(\theta)$ gaussiana esférica y se sostienen para toda $t(\theta)$. En el ejemplo de la mezcla de dos distribuciones normales se tiene lo

siguiente

$$
\begin{aligned}
Z(\mathbf{m}_q, \nu_q) &= \int_\Theta \left[(1-\omega)N(x|\theta, \mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I})\right] q(\theta) d\theta \\
&= (1-\omega)\int_\Theta N(x|\theta, \mathbf{I}) q(\theta) d\theta + \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= (1-\omega)\int_\Theta \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\theta)^{'}(x-\theta)\right) q(\theta) d\theta + \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= \frac{(1-\omega)}{(2\pi)^n \nu_q^{\frac{n}{2}}} \int_\Theta \exp\left(-\frac{(x-\theta)^{'}(x-\theta)}{2}\right) \exp\left(-\frac{(\theta-\mathbf{m}_q)^{'}(\theta-\mathbf{m}_q)}{2\nu_q}\right) d\theta \\
&+ \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= \frac{(1-\omega)}{(2\pi)^n \nu_q^{\frac{n}{2}}} \int_\Theta \exp\left(-\frac{1}{2\nu_q}\left[\nu_q (x-\theta)^{'}(x-\theta) - (\theta-\mathbf{m}_q)^{'}(\theta-\mathbf{m}_q)\right]\right) \\
&+ \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= \frac{(1-\omega)}{(2\pi)^n \nu_q^{\frac{n}{2}}} \int_\Theta \exp\left(-\frac{1}{2\nu_q}\left[\nu_q x^{'} x - 2\nu x^{'}\theta - 2\theta^{'}\mathbf{m}_q + \mathbf{m}^{'}\mathbf{m}\right]\right) \\
&+ \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= \frac{(1-\omega)}{(2\pi)^n \nu_q^{\frac{n}{2}}} \int_\Theta \exp\left(-\frac{(\nu_q+1)}{2\nu_q}\left[\left(\theta + \frac{\nu_q x + \mathbf{m}_q}{\nu_q+1}\right)^{'}\left(\theta + \frac{\nu_q x + \mathbf{m}_q}{\nu_q+1}\right)\right] + C\right) \\
&+ \omega N(x|\mathbf{0}, 10\mathbf{I})
\end{aligned}
$$

Con

$$
C = -\frac{1}{2\nu_q(\nu_q+1)}\left[\nu_q(\nu_q+1)x^{'}x + (\nu_q+1)\mathbf{m}_q^{'}\mathbf{m}_q - [\nu_q x^{'} + \mathbf{m}_q^{'}][\nu_q x + \mathbf{m}_q]\right]
$$

Por lo que

$$
\begin{aligned}
Z(\mathbf{m}_q, \nu_q) &= \frac{(1-\omega)}{(2\pi)^{\frac{n}{2}}(\nu_q+1)^{\frac{n}{2}}} \exp(C) + \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= \frac{(1-\omega)}{(2\pi)^{\frac{n}{2}}(\nu_q+1)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\nu_q(\nu_q+1)}\left[\nu_q x^{'}x - 2\nu_q x^{'}\mathbf{m}_q + \nu_q \mathbf{m}_q^{'}\mathbf{m}_q\right]\right) \\
&+ \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= \frac{(1-\omega)}{(2\pi)^{\frac{n}{2}}(\nu_q+1)^{\frac{n}{2}}} \exp\left(-\frac{1}{2(\nu_q+1)}\left[x^{'}x - 2x^{'}\mathbf{m}_q + \mathbf{m}_q^{'}\mathbf{m}_q\right]\right) \\
&+ \omega N(x|\mathbf{0}, 10\mathbf{I}) \\
&= (1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I})
\end{aligned}
$$

Ahora considera el gradiente de $N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I})$ con respecto a $\mathbf{m}_q$

$$
\begin{aligned}
\nabla_m N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I}) &= \nabla_m \frac{\exp\left(-\frac{1}{2(\nu_q+1)}(x-\mathbf{m}_q)^{'}(x-\mathbf{m}_q)\right)}{(2\pi(\nu_q+1))^{\frac{n}{2}}} \\
&= \frac{1}{(2\pi(\nu_q+1))^{\frac{n}{2}}} \frac{(x-\mathbf{m}_q)}{(\nu_q+1)} \exp\left(-\frac{(x-\mathbf{m}_q)^{'}(x-\mathbf{m}_q)}{2(\nu_q+1)}\right) \qquad (2.16)
\end{aligned}
$$

Entonces, tomando en cuenta la expresión en (2.16), se tiene

$$\begin{aligned}\nabla_m \ln Z(\mathbf{m}_q, \nu_q) &= \frac{\nabla_m (1-\omega) N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I})}{(1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I})} \\ &= \frac{(1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I})}{(1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I})} \frac{(x-\mathbf{m}_q)}{(\nu_q+1)} \\ &= r\frac{(x-\mathbf{m}_q)}{(\nu_q+1)}\end{aligned}$$

Con

$$r = \frac{(1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I})}{(1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I})}$$

De manera análoga para el gradiente del $\ln Z(\mathbf{m}_q, \nu_q)$ con respecto de ν_q

$$\begin{aligned}\nabla_\nu N(x|\mathbf{m}_q, \nu_q) &= -\frac{2\pi\frac{n}{2}(2\pi(\nu_q+1))^{\frac{n}{2}-1}}{(2\pi(\nu_q+1))^n} \exp\left(-\frac{1}{2(\nu_q+1)}(x-\mathbf{m}_q)'(x-\mathbf{m}_q)\right) \\ &+ \frac{\exp\left(-\frac{1}{2\pi(\nu_q+1)(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}\right)}{(2\pi(\nu_q+1))^{\frac{n}{2}}} \frac{(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}{2(\nu_q+1)^2} \\ &= -\frac{n}{2(2\pi(\nu_q+1))^{\frac{n}{2}}(\nu_q+1)} \exp\left(-\frac{1}{2(\nu_q+1)}(x-\mathbf{m}_q)'(x-\mathbf{m}_q)\right) \\ &+ \frac{\exp\left(-\frac{1}{2\pi(\nu_q+1)(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}\right)}{(2\pi(\nu_q+1))^{\frac{n}{2}}} \frac{(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}{2(\nu_q+1)^2}\end{aligned}$$

Así se tiene que

$$\begin{aligned}\nabla_\nu \ln Z(\mathbf{m}_q, \nu_q) &= \frac{\nabla_\nu (1-\omega) N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I})}{(1-\omega)N(x|\mathbf{m}_q, (\nu_q+1)\mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I})} \\ &= -r\frac{n}{2(\nu_q+1)} + r\frac{(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}{2(\nu_q+1)^2}\end{aligned}$$

Por su parte, para el cálculo de la varianza se requiere conocer

$$\begin{aligned}\nabla_m' \nabla_m - 2\nabla_\nu \ln Z(\mathbf{m}_q, \nu_q) &= \nabla_m' \ln Z(\mathbf{m}_q, \nu_q) \nabla_m \ln Z(\mathbf{m}_q, \nu_q) - 2\nabla_\nu \ln Z(\mathbf{m}_q, \nu_q) \\ &= \left[r\frac{(x-\mathbf{m}_q)}{(\nu_q+1)}\right]' \left[r\frac{(x-\mathbf{m}_q)}{(\nu_q+1)}\right] + r\frac{n}{(\nu_q+1)} - r\frac{(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}{(\nu_q+1)^2} \\ &= r^2\frac{(x-\mathbf{m}_1)'(x-\mathbf{m}_1)}{(\nu_q+1)^2} + r\frac{n}{(\nu_q+1)} - r\frac{(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}{(\nu_q+1)^2} \\ &= r\frac{n}{(\nu_q+1)} - r(1-r)\frac{(x-\mathbf{m}_q)'(x-\mathbf{m}_q)}{(\nu_q+1)^2}\end{aligned}$$

Por último, para estimar el factor de escala de la distribución posterior, $p(X)$, basta con multiplicar los factores de normalización $Z_i(\mathbf{m}_q, \nu_q)$ producidos por cada iteración.

Una vez teniendo las expresiones (2.14) y (2.15) para la mezcla de dos distribuciones normales se puede obtener el algoritmo de densidad de emisión para una mezcla de dos distribuciones gaussianas, el cual se muestra en el Algoritmo 6.

De esta forma se puede observar que para cada punto se calcula la probabilidad de estar en la zona de *confusión*, esto es si forma parte de $N(x|\theta, \mathbf{I})$ o de $N(x|\mathbf{0}, 10\mathbf{I})$, al igual que se actualiza el estimador de θ, $\mathbf{m}_q$, y de la confianza dentro de la estimación, ν_q.

Algoritmo 6 Método de densidad de emisión

1: Iniciar $\mathbf{m}_q = 0, \nu_q = 100$, los parámetros de la distribución a priori. Iniciar $s = 1$, el cual es el factor de escala, $p(X)$.

2: Para cada punto $x_i = (x_{1i}, \ldots, x_{mi})$ actualizar los parámetros $(\mathbf{m}_q, \nu_q, s)$ siguiendo las expresiones (2.14) y (2.15):

$$\mathbf{m}_q^{nueva} = \mathbf{m}_q + \nu_q r_i \frac{(x_i - \mathbf{m}_q)}{\nu_q + 1}$$

$$\nu_q^{nueva} = \nu_q - r_i \frac{\nu_q^2}{\nu_q + 1} + r_i(1 - r_i) \frac{\nu_q^2 (x_i - \mathbf{m}_q)'(x_i - \mathbf{m}_q)}{n(\nu_q + 1)^2}$$

$$s^{nueva} = s \cdot Z_i(\mathbf{m}_q, \nu_q)$$

Sin embargo, el algoritmo depende del orden en el que los datos son analizados, ya que la probabilidad de estar en la zona de confusión depende del estimado de θ. Lo que puede implicar una inferencia poco confiable.

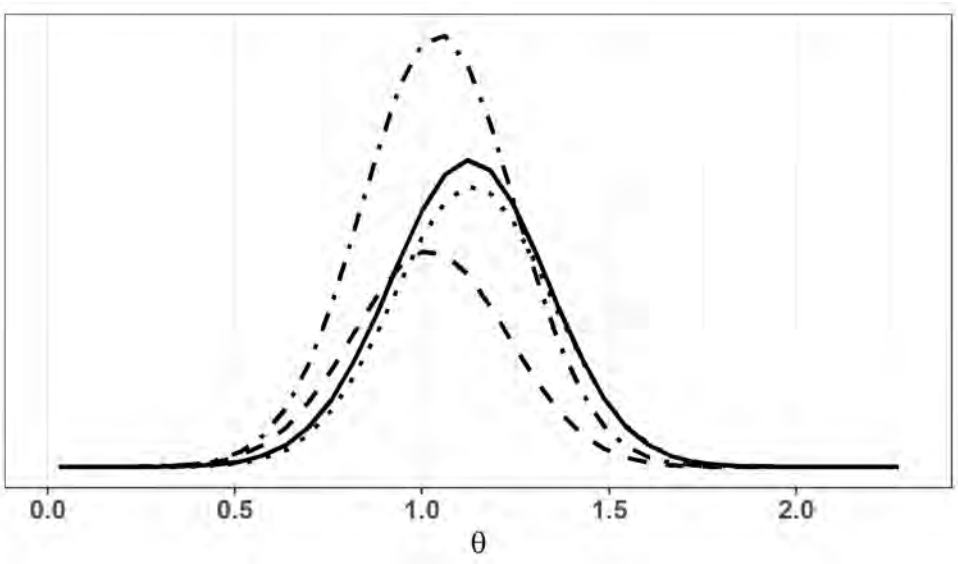

Figura 2.5: Función de distribución conjunta, línea sólida, vs. las distribuciones aproximadas vía densidad de emisión para tres ordenaciones de los datos

Capítulo 3

Método de propagación de esperanzas

En este capítulo se describe la metodología propuesta por Minka (2001), propagación de esperanzas, el cual se basa en una nueva interpretación del método densidad de emisión.

Como se observó en la sección anterior, en el método de densidad de emisión se trata cada observación en el término t_i para que se aproxime la distribución posterior que incluye a t_i, lo cual implica un problema en la ordenación de los datos. Por otra parte, esto se puede realizar aproximando primeramente la función t_i, con otra función $\hat{t}_i$, para así obtener la distribución posterior utilizando la nueva función de aproximación $\hat{t}_i$.

Esta interpretación es siempre posible ya que se puede definir la función de aproximación $\hat{t}_i$ como el cociente entre la nueva posterior y la antigua multiplicada por una constante

$$\hat{t}(\theta) = Z\frac{q^{nueva}(\theta)}{q(\theta)}. \tag{3.1}$$

Por lo que al multiplicar la ecuación (3.1) por $q(\theta)$ se obtiene la función deseada $q(\theta)^{nueva}$.

Una de las propiedades importantes de este algoritmo es que si la distribución posterior aproximada es gaussiana entonces el término de aproximación, $\hat{t}_i$, también tendrá la misma distribución ya que es el cociente de las posteriores. De igual forma si la distribución posterior pertenece a la familia exponencial, entonces el término de aproximación tendrá la misma forma funcional de esa distribución.

Así, el algoritmo de propagación de esperanzas se puede interpretar como una secuencia de aproximaciones gaussianas $\hat{t}_i$ para obtener una distribución posterior gaussiana en θ. Por lo que bajo esta perspectiva, el orden en que son procesados los datos no importa ya que éste sólo determina cómo se hace la aproximación.

Para hacer más claro cómo ésta nueva interpretación funciona considere la situación planteada en la sección anterior para la densidad de emisión en la que se tenian dos distribuciones gausianas esféricas

$$\begin{aligned} q(\theta) &= N_n(\mathbf{m}, \nu\mathbf{I}), \\ q(\theta)^{nueva} &= N_n(\mathbf{m}^*, \nu^*\mathbf{I}). \end{aligned}$$

Para estas distribuciones el término de aproximación $\hat{t}_i$ está dado de la siguiente forma:

$$
\begin{aligned}
Z_i(\mathbf{m},\nu) &= \int_\theta t_i q(\theta) d\theta \\
\hat{t}_i(\theta) &= Z_i(\mathbf{m},\nu)\frac{q^{nueva}(\theta)}{q(\theta)} \\
&= Z_i(\mathbf{m},\nu)\left(\frac{\nu}{\nu^*}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\nu^*}(\theta-\mathbf{m}^*)'(\theta-\mathbf{m}^*)\right)\exp\left(\frac{1}{2\nu}(\theta-\mathbf{m})'(\theta-\mathbf{m})\right) \\
&= Z_i(\mathbf{m},\nu)\left(\frac{\nu}{\nu^*}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\nu^*}(\theta'\theta - 2\theta'\mathbf{m}^* + \mathbf{m}^{*'}\mathbf{m}^*)\right) \\
&\quad * \exp\left(\frac{1}{2\nu}(\theta'\theta - 2\theta'\mathbf{m} + \mathbf{m}'\mathbf{m})\right) \\
&= Z_i(\mathbf{m},\nu)\left(\frac{\nu}{\nu^*}\right)^{\frac{n}{2}} \exp\left(-\frac{(\nu-\nu^*)}{2\nu^*\nu}\left[\left(\theta - \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)'\left(\theta - \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)\right]\right) \\
&\quad * \exp\left(-\frac{(\nu-\nu^*)}{2\nu^*\nu}\left[\frac{\mathbf{m}^{*'}\mathbf{m}^*\nu - \mathbf{m}'\mathbf{m}\nu^*}{\nu-\nu^*} - \left(\frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)'\left(\frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)\right]\right) \\
&= Z_i(\mathbf{m},\nu)\left(\frac{\nu}{\nu^*}\right)^{\frac{n}{2}} \exp\left(-\frac{(\nu-\nu^*)}{2\nu^*\nu}\left[\left(\theta - \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)'\left(\theta - \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)\right]\right), \\
&\quad * \exp(C)
\end{aligned}
$$

observando el argumento de la segunda exponencial

$$
\begin{aligned}
C &= -\frac{(\nu-\nu^*)}{2\nu^*\nu}\left[\frac{\mathbf{m}^{*'}\mathbf{m}^*\nu - \mathbf{m}'\mathbf{m}\nu^*}{\nu-\nu^*} - \left(\frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)'\left(\frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)\right] \\
&= -\frac{1}{2\nu^*\nu(\nu-\nu^*)}\left[(\nu-\nu^*)(\mathbf{m}^{*'}\mathbf{m}^*\nu - \mathbf{m}'\mathbf{m}\nu^*) - (\mathbf{m}^*\nu - \mathbf{m}\nu^*)'(\mathbf{m}^*\nu - \mathbf{m}\nu^*)\right] \\
&= \frac{1}{2(\nu-\nu^*)}\left[\mathbf{m}^{*'}\mathbf{m}^* + \mathbf{m}'\mathbf{m} - 2\mathbf{m}^{*'}\mathbf{m}\right] \\
&= \frac{1}{2(\nu-\nu^*)}(\mathbf{m}^*-\mathbf{m})'(\mathbf{m}^*-\mathbf{m}),
\end{aligned}
$$

con lo cual se tiene

$$
\begin{aligned}
\hat{t}_i(\theta) = Z_i(\mathbf{m},\nu)\left(\frac{\nu}{\nu^*}\right)^{\frac{n}{2}} &\exp\left(-\frac{(\nu-\nu^*)}{2\nu^*\nu}\left[\left(\theta - \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)'\left(\theta - \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu-\nu^*}\right)\right]\right) \\
&* \exp\left(\frac{1}{2(\nu-\nu^*)}(\mathbf{m}^*-\mathbf{m})'(\mathbf{m}^*-\mathbf{m})\right). \qquad (3.2)
\end{aligned}
$$

Ahora se definen las siguientes variables para la distribución de aproximación $\hat{t}_i$ con base en las expresiones del argumento de la primera exponencial

$$\nu_i = \frac{\nu^*\nu}{(\nu-\nu^*)},$$

lo cual implica que

$$
\begin{aligned}
\nu_i^{-1} &= \frac{(\nu-\nu^*)}{\nu^*\nu} \\
&= \nu^{*-1} - \nu^{-1}
\end{aligned}
$$

de igual manera para el parámetro de localización

$$\begin{aligned}\mathbf{m}_i &= \frac{\mathbf{m}^*\nu - \mathbf{m}\nu^*}{\nu - \nu^*} \\ &= \mathbf{m}^*\nu_i\nu^{*-1} - \mathbf{m}\nu_i\nu^{-1},\end{aligned}$$

sustituyendo ν^{*-1}

$$\begin{aligned}\mathbf{m}_i &= \mathbf{m}^*\nu_i(\nu_i^{-1} + \nu^{-1}) - \mathbf{m}\nu_i\nu^{-1} \\ &= \mathbf{m}^* + \nu_i\nu^{-1}(\mathbf{m}^* - \mathbf{m}) + \mathbf{m} - \mathbf{m} \\ &= \mathbf{m} + (\mathbf{m}^* - \mathbf{m}) + \nu_i\nu^{-1}\mathbf{m}^* \\ &= \mathbf{m} + \nu^{-1}(\nu_i + \nu)(\mathbf{m}^* - \mathbf{m}).\end{aligned}$$

Finalmente se tienen las expresiones que permiten reducir e interpretar de mejor manera la forma de $\hat{t}_i$

$$\nu_i^{-1} = \nu^{*-1} - \nu^{-1} \tag{3.3}$$

$$\mathbf{m}_i = \mathbf{m} + \nu^{-1}(\nu_i + \nu)(\mathbf{m}^* - \mathbf{m}) \tag{3.4}$$

Por último la expresión (3.2) se puede reescribir usando las expresiones (3.3) y (3.4) obteniendo

$$\begin{aligned}\hat{t}_i(\theta) &= Z_i(\mathbf{m}, \nu)\left(\frac{\nu_i + \nu}{\nu_i}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\nu_i\nu}(\theta - \mathbf{m}_i)'(\theta - \mathbf{m}_i)\right) \\ &\quad * \exp\left(\frac{1}{2(\nu_i + \nu)}(\mathbf{m}_i - \mathbf{m})'(\mathbf{m}_i - \mathbf{m})\right) \\ &= \frac{Z_i(\mathbf{m}, \nu)}{N(\mathbf{m}_i|\mathbf{m}, (\nu_i + \nu)\mathbf{I}}N(\theta|\mathbf{m}_i, \nu_i\mathbf{I}).\end{aligned} \tag{3.5}$$

Por cómo fue construida la función de aproximación, se tiene que $\hat{t}_i q(\theta)$ produce la actualización del algoritmo de densidad de emisión con el respectivo factor de escala, (2.12). Se debe señalar que la notación $N(\cdot|\mu, \sigma)$ en (3.5) es usada de manera simbólica, y no como una densidad propia implícita en el modelo. La función $\hat{t}_i$ tiene una forma exponencial cuadrática en θ, la cual tiene forma de una distribución gausiana con media $\mathbf{m}_i$ y varianza $\nu_i\mathbf{I}$ multiplicada por un factor de escala. Sin embargo, lo que hace distinta a $\hat{t}_i$ de una distribución gausiana es que la varianza, ν_i, puede ser negativa o infinita, lo cual implica que $\hat{t}_i$ es una función con una curva hacia arriba o bien una constante, respectivamente. Esto caracteriza la forma en que se aproxima cada término t_i.

La Figura (3.1) muestra la función $\hat{t}_i(\theta)$ y la función $t_i(\theta)$ para el problema de la mezcla de dos gausianas.

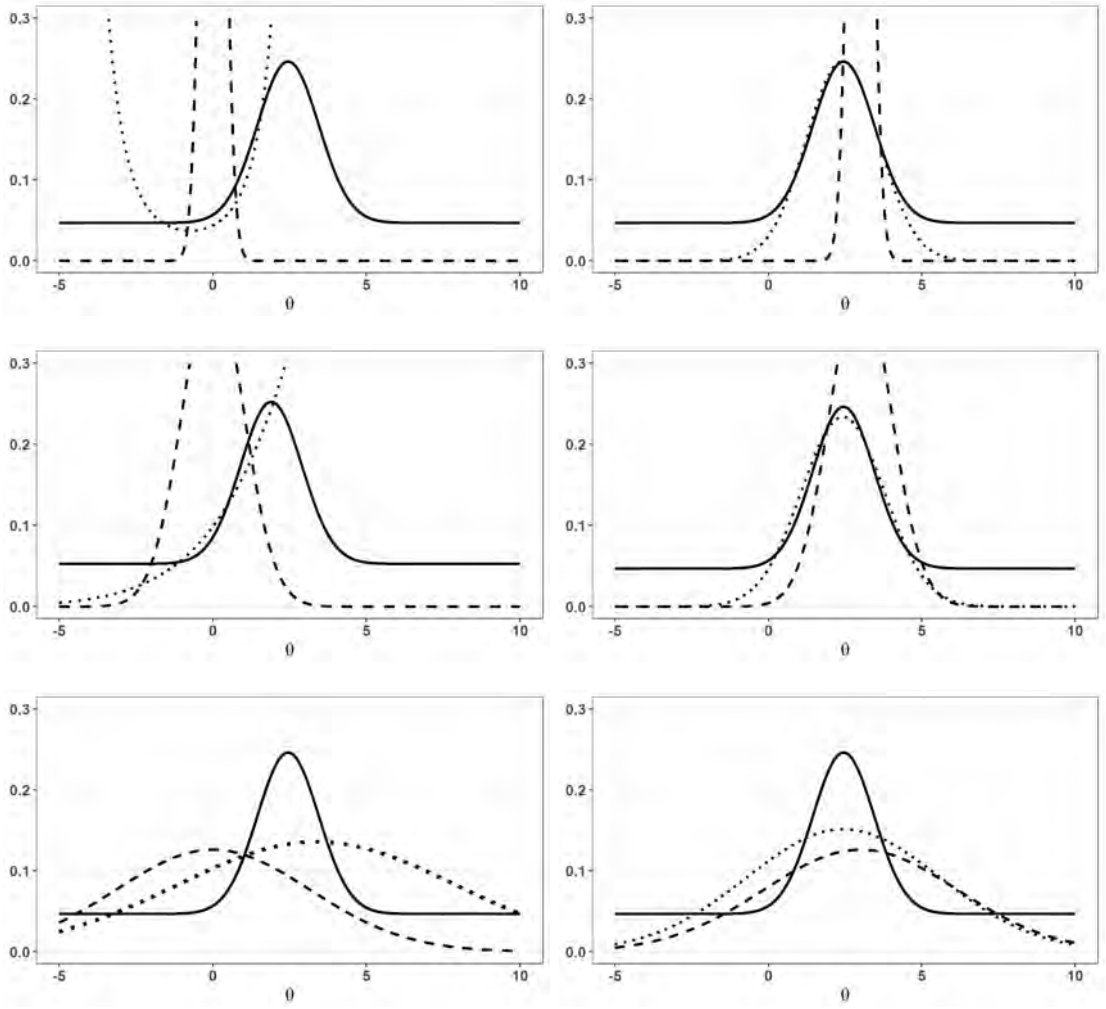

Figura 3.1: Término de aproximación $\hat{t}_i(\theta)$, representado por líneas punteadas, como una función de θ vs. el término exacto $t_i(\theta)$, representado por líneas sólidas, y la distribución posterior $q(\theta)$, representada por líneas discontinuas, en la iteración i en el problema de mezcla de dos gausianas. En cada gráfica se consideran distintos valores de m y ν, la columna de la izquierda corresponde a valores de $m = 0$ y la derecha a $m = 3$, mientras que para los renglones, el primero corresponde a $\nu = 0.1$, el segundo a $\nu = 1$ y el último a $\nu = 10$.

En resumen, el término $t_i = p(x_i|\theta)$ en la densidad de emisión es una distribución gausiana en θ, el término $\hat{t}_i$ es una exponencial cuadrática en θ. La distribución $q(\theta)$ determina en dónde la función de aproximación $\hat{t}_i$ es precisa. Por ejemplo, si $q(\theta)$ es estrecha, i.e. ν es pequeña, entonces $\hat{t}_i$ es precisa sólo en rangos estrechos de valores de θ, determinado por $\mathbf{m}$, por otro lado, si ν es grande entonces $\hat{t}_i$ trata de aproximar t_i más ampliamente y $\mathbf{m}$ es menos importante.

Hasta este punto, sólo se tiene una expresión diferente de las actualizaciones realizadas por el algoritmo de densidad de emisión. Para obtener el método de propagación de esperanzas se refinan las variables $(\mathbf{m}_i, \nu_i)$ considerando todas las demás aproximaciones. El algoritmo general del método de propagación de esperanzas, (Minka, 2001), se describe en el Algoritmo 7.

Algoritmo 7 Método de propagación de esperanzas

1: Iniciar el término de aproximación $\hat{t}_i$.
2: Calcular la distribución posterior para θ a partir del producto de $\hat{t}_i$:

$$q^{nueva}(\theta) = \frac{\prod_i \hat{t}_i}{\int \prod_i \hat{t}_i d\theta}. \tag{3.6}$$

3: Hasta la convergencia de todas las $\hat{t}_i$:

- Elegir una $\hat{t}_i$ para refinar.
- Remover el término $\hat{t}_i$ de la posterior para obterner una distribución posterior *vieja* $q(\theta)^{/i}$ dividiendo y normalizando:

$$q(\theta)^{/i} \propto \frac{q^{nueva}(\theta)}{\hat{t}_i}. \tag{3.7}$$

- Calcular la distribución posterior $q^{nueva}(\theta)$ y el factor de normalización Z_i de $q^{/i}(\theta)$ y t_i mediante el algoritmo de filtración de desidad asumida.
- Actualizar $\hat{t}_i$ como:

$$\hat{t}_i = Z_i \frac{q^{nueva}(\theta)}{q(\theta)}.$$

4: Utilizar la constante de normalización de $q^{nueva}(\theta)$, en (3.6), como una aproximación a $p(x)$:

$$p(x) \approx \int \prod_i \hat{t}_i d\theta.$$

En la convergencia, el resultado no dependerá del orden de procesamiento. Durante este proceso, en (3.7) se utiliza la división para remover el término $\hat{t}_i$, este paso puede ser realizado de igual forma acumulando los términos $\hat{t}$ excepto el $i-$ésimo:

$$q(\theta) \propto \prod_{j \neq i} \hat{t}_j(\theta). \tag{3.8}$$

Se puede notar que debido a que el término de aproximación se inicializa en uno, el resultado de pasar por primera vez a través de los datos es exactamente igual que realizar el método de densidad de emisión.

3.1. Mezcla de dos gausianas vía propagación de esperanzas

Ahora, para entender mejor el procedimiento del algoritmo considere el ejemplo de la mezcla de dos distribuciones gausianas planteado en el método de densidad de emisión, el algoritmo de propagación de esperanzas para este caso, considerando el Algoritmo 7, se describe en el Algoritmo 8.

Algoritmo 8 Método de propagación de esperanzas para mezcla de dos gausianas

1: El término de aproximación tiene la forma:

$$\hat{t}_i = s_i \exp\left(-\frac{1}{2\nu_i}(\theta - \mathbf{m}_i)'(\theta - \mathbf{m}_i)\right)$$

Iniciar el término de aproximación $\hat{t}_i$, el término $\hat{t}_0$ se inicia con la distribución a priori y los términos restantes en 1.

$$\begin{aligned} \nu_0 &= 100 & \nu_i &= \infty \\ \mathbf{m}_0 &= 0 & \mathbf{m}_i &= 0 \\ s_0 &= (2\pi\nu_0)^{-\frac{n}{2}} & s_i &= 1 \end{aligned}$$

2: Calcular la distribución posterior a partir de las $\hat{t}_i$:

$$\mathbf{m}^{nueva} = \mathbf{m}_0\,, \quad \nu^{nueva} = \nu_0$$

3: Hasta que $(\mathbf{m}_i, \nu_i, s_i)$ convergan, considerando cambios menores a 10^{-4}, para $i = 1, \ldots, n$ hacer:

- Remover $\hat{t}_i$ de la distribución posterior para obtener una *vieja* posterior, $q(\theta)$, utilizando el resultado de (3.3) y (3.4):

$$\begin{aligned} \nu^{-1} &= (\nu^{nueva})^{-1} - \nu_i^{-1} \\ \mathbf{m} &= \nu(\nu^{nueva})^{-1}\mathbf{m}^{nueva} - \nu\nu_i^{-1}\mathbf{m}_i = \mathbf{m}^{nueva} + \nu\nu^{-1}(\mathbf{m}^{nueva} - \mathbf{m}_i) \end{aligned} \tag{3.9}$$

- Actualizar $(\mathbf{m}^{nueva}, \nu^{nueva}, Z_i)$ a partir de $(\mathbf{m}, \nu)$ utilizando la densidad de emisión, Algoritmo (6):

$$\begin{aligned} \mathbf{m}^{nueva} &= \mathbf{m} + \nu_i \frac{(x_i - \mathbf{m})}{\nu + 1} \\ \nu^{nueva} &= \nu - r_i \frac{\nu^2}{\nu + 1} + r_i(1 - r_i)\frac{\nu^2 (x_i - \mathbf{m})'(x_i - \mathbf{m})}{n(\nu + 1)^2} \\ Z_i &= (1 - \omega)N(x_i|\mathbf{m}, (\nu + 1)\mathbf{I}) + \omega N(x_i|\mathbf{0}, 10\mathbf{I}) \end{aligned} \tag{3.10}$$

- Actualizar $\hat{t}_i$ a partir de (3.9) y (3.10):

$$\nu_i = \left(\frac{r_i}{\nu + 1} - r_i(1 - r_i)\frac{(x_i - \mathbf{m})'(x_i - \mathbf{m})}{n(\nu + 1)^2}\right)^{-1} - \nu \tag{3.11}$$

$$\begin{aligned} \mathbf{m}_i &= \mathbf{m} - (\nu_i + \nu)\nu^{-1}(\mathbf{m}^{nueva} - \mathbf{m}) \\ &= \mathbf{m} - (\nu_i + \nu) r_i \frac{x_i - \mathbf{m}}{\nu + 1} \end{aligned} \tag{3.12}$$

$$s_i = \frac{Z_i}{(2\pi\nu_i)^{\frac{n}{2}} N(\mathbf{m}_i|\mathbf{m}, (\nu_i + \nu)\mathbf{I})} \tag{3.13}$$

4: Calcular la constante de normalización

$$B = \frac{\nu^{nueva}}{2}\left(\sum_i \frac{\mathbf{m}_i}{\nu_i}\right)'\left(\sum_i \frac{\mathbf{m}_i}{\nu_i}\right) - \sum_i \frac{\mathbf{m}_i'\mathbf{m}_i}{\nu_i}$$

$$p(x) \approx \int \prod_i \hat{t}_i(\theta) d\theta = (2\pi\nu^{nueva})^{\frac{n}{2}} \exp\left(\frac{B}{2}\right) \prod_i^n s_i \tag{3.14}$$

Para obtener las expresiones (3.11), (3.12), (3.13) y (3.14) del algoritmo se hace uso de lo siguiente

$$\begin{aligned}
\nu_i &= \frac{\nu^{nueva}\nu}{\nu - \nu^{nueva}} \\
&= \frac{\nu^2 - r_i\frac{\nu^3}{\nu+1} + r_i(1-r_i)\frac{\nu^3(x_i-\mathbf{m})'(x_i-\mathbf{m})}{n(\nu+1)^2}}{r_i\frac{\nu^2}{\nu+1} - r_i(1-r_i)\frac{\nu^2(x_i-\mathbf{m})'(x_i-\mathbf{m})}{n(\nu+1)^2}} \\
&= \frac{1 - r_i\frac{\nu}{\nu+1} + r_i(1-r_i)\frac{\nu(x_i-\mathbf{m})'(x_i-\mathbf{m})}{n(\nu+1)^2}}{\frac{r_i}{\nu+1} - r_i(1-r_i)\frac{(x_i-\mathbf{m})'(x_i-\mathbf{m})}{n(\nu+1)^2}} \\
&= \left(\frac{r_i}{\nu+1} - r_i(1-r_i)\frac{(x_i-\mathbf{m})'(x_i-\mathbf{m})}{n(\nu+1)^2}\right)^{-1} - \nu \\
\mathbf{m}_i &= \nu_i(\nu^{nueva})^{-1}\mathbf{m}^{nueva} - \mathbf{m}(\nu^{-1}\nu_i) \\
&= \mathbf{m} - \mathbf{m}(\frac{\nu^{nueva}}{\nu - \nu^{nueva}} + 1) + \mathbf{m}^{nueva}\frac{\nu}{\nu - \nu^{nueva}} \\
&= \mathbf{m} - (\nu_i + \nu)\nu^{-1}(\mathbf{m}^{nueva} - \mathbf{m}) \\
&= \mathbf{m} - (\nu_i + \nu)r_i\frac{x_i - \mathbf{m}}{\nu + 1}
\end{aligned}$$

Para el término s_i se parte de la siguiente expresión

$$\begin{aligned}
\hat{t}_i(\theta) &= s_i \exp\left(-\frac{1}{2\nu_i}(\theta_i - \mathbf{m}_i)'(\theta_i - \mathbf{m}_i)\right) \\
&= \frac{Z_i(\mathbf{m}, \nu)}{N(\mathbf{m}_i|\mathbf{m}, (\nu_i + \nu)\mathbf{I})}N(\theta|\mathbf{m}_i, \nu_i\mathbf{I})
\end{aligned}$$

La última igualdad se debe a (3.5), para obtener la expresion (3.13) basta con completar la distribución normal en la primer igualdad y despejar s_i.

Por último, para calcular la constante de normalización se tiene

$$\begin{aligned}
\prod_i \hat{t}_i &= \prod_i s_i \exp\left(-\frac{1}{2\nu_i}(\theta - \mathbf{m}_i)'(\theta - \mathbf{m}_i))\right) \\
&= \exp\left(-\sum_i \frac{(\theta - \mathbf{m}_i)'(\theta - \mathbf{m}_i)}{2\nu_i}\right)\prod_i s_i \\
&= \exp\left(-\sum_i \frac{(\theta'\theta - 2\theta'\mathbf{m}_i + \mathbf{m}_i'\mathbf{m}_i)}{2\nu_i}\right)\prod_i s_i \\
&= \exp\left(-\theta'\theta\sum_i \frac{1}{2\nu_i} + \theta'\sum_i \frac{\mathbf{m}_i}{\nu_i} - \sum_i \frac{\mathbf{m}_i'\mathbf{m}_i}{2\nu_i}\right)\prod_i s_i \\
&= \exp\left(-C\left[\theta'\theta - \theta' A\right] - \sum_i \frac{\mathbf{m}_i'\mathbf{m}_i}{2\nu_i}\right)\prod_i s_i \\
&= \exp\left(-C\left(\theta - \frac{A}{2}\right)'\left(\theta - \frac{A}{2}\right) + C\frac{A'A}{4} - \sum_i \frac{\mathbf{m}_i'\mathbf{m}_i}{2\nu_i}\right)\prod_i s_i
\end{aligned} \tag{3.15}$$

donde

$$C = \sum_i \frac{1}{2\nu_i} = \frac{1}{2\nu^{nueva}}$$
$$A = \sum_i \frac{\mathbf{m}_i}{\nu_i C} = \sum_i \frac{\nu^{nueva}\mathbf{m}_i}{2\nu_i}$$

Integrando la expresión (3.15) con respecto a θ y completando la distribución $N(\frac{A}{2}, \nu^{nueva}\mathbf{I})$ se obtiene (3.14).

Siguiendo el algoritmo (8), y para una muestra simulada de una mezcla de distribuciones normales como en (2.13) con $\omega = 0.5$ se obtiene el ajuste mostrado en la Figura (3.2).

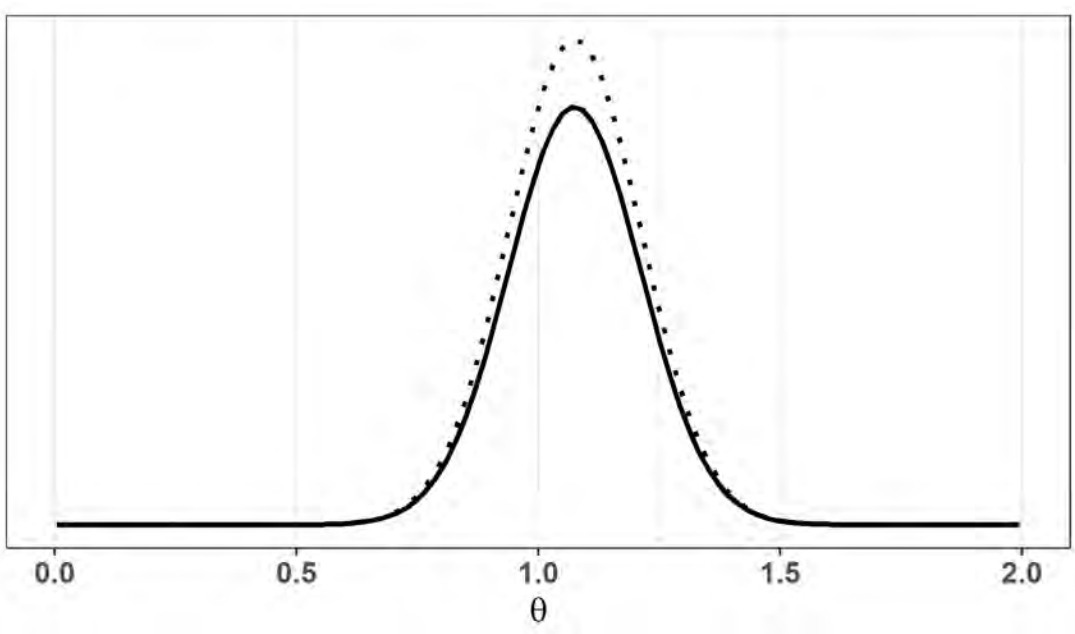

Figura 3.2: Distribución conjunta de la mezcla gausiana, línea sólida, vs. la distribución conjunta aproximada vía propagación de esperanzas, línea punteada.

Capítulo 4

Comparación y resultados

En este capítulo se evalúa el algoritmo de densidad de emisión y propagación de esperanzas en el problema de mezcla de dos distribuciones gausianas y además se compara con los otros cuatro algoritmos de inferencia aproximada: muestreo por importancia, muestreo de Gibbs, método de Laplace y variación de Bayes vía el algoritmo EM.

4.1. Métodos aplicados a una mezcla de dos distribuciones gausianas

En esta sección se describe la implementación de las metodologías para el caso de la mezcla de dos distribuciones gausianas.

4.1.1. Muestreo por importancia.

Para implementar la metodología de muestreo por importancia se utiliza la expresión (2.2) utilizando la forma de la mezcla de dos distribuciones gausianas

$$p(x|\theta) = (1-\omega)N(x|\theta, \mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I}) \qquad , \omega \in (0,1).$$

Considerando la distribución a priori $p(\theta)$, se puede calcular la constante de normalización de la distribución posterior integrando sobre el dominio de θ

$$\begin{aligned} p(x) &= \int_{\Theta} p(x|\theta)p(\theta)d\theta \\ &= \int_{\Theta} \prod_i p(x_i|\theta)p(\theta)d\theta \\ &= \mathbb{E}_\theta \left[\prod_i p(x_i|\theta) \right] \end{aligned}$$

Por lo que si se genera una muestra $(\theta_1, \ldots, \theta_S)$ de $p(\theta)$ y utilizando el estimador (2.3) se tiene

$$p(x) \approx \frac{1}{S} \sum_{i=1}^{S} p(x|\theta_i).$$

Por otro lado, es de interés conocer la esperanza posterior, la cual se puede estimar como sigue

$$\mathbb{E}[\theta|x] \approx \frac{\sum_{i=1}^{S} \theta_i p(x|\theta_i)}{\sum_{i=1}^{S} p(x|\theta_i)},$$

debido a que

$$\begin{aligned} \mathbb{E}[\theta|x] &= \int_{\Theta} \theta p(\theta|x) d\theta \\ &= \int_{\Theta} \theta \frac{p(\theta, x)}{p(x)} d\theta \\ &= \int_{\Theta} \theta \frac{p(x|\theta)p(\theta)}{p(x)} d\theta. \end{aligned}$$

Sin embargo, éste método sólo estima integrales específicas, por lo que para compararlo con los demás algoritmos de manera cuantitativa se calcula el valor absoluto de la diferencia entre la media exacta y la estimada, al igual que el valor exacto de $p(x)$ y el estimado.

4.1.2. Muestreo de Gibbs.

En el muestro de Gibbs, se introducen variables ocultas c_i las cuales indican a qué categoría pertenecen las x_i, donde cada categoría es una distribución normal. Dado un valor de θ, se genera un valor de c_i a partir de una distribución Bernoulli con media la probabilidad de pertenencia al primer grupo. Por lo que dado c_i, se obtiene una distribución posterior exacta normal para θ y se genera una nueva observación de θ. El promedio de las θ's generadas por este proceso es la estimación de la media posterior.

De manera más formal, para el modelo de mezcla de dos distribuciones normales

$$p(x|\theta) = (1-\omega)N(x|\theta, \mathbf{I}) + \omega N(x|\mathbf{0}, 10\mathbf{I}) \qquad \omega \in (0, 1)$$

el muestreo de Gibbs se desarrolla considerando lo siguiente, (Zhihui, 2010)

$$\begin{aligned} \theta &\sim Normal(\mu_0, \sigma_0^2) \\ c_i|\theta, x_i &\sim Bernoulli(r_i) \\ \theta_i|c_i, \mathbf{x} &\sim Normal\left(\frac{\sigma_1^2 s_1^x + \sigma_0^2 \mu_0}{\sigma_1^2 n_1 + \sigma_0^2}, \frac{1}{\sigma_1^2 n_1 + \sigma_0^2}\right) \end{aligned}$$

donde

$$r_i = \frac{(1-\omega)N(x_i|\theta_i, \mathbf{I})}{(1-\omega)N(x_i|\theta_i, \mathbf{I}) + \omega N(x_i|\mathbf{0}, 10\mathbf{I})},$$
$$n_1 = \sum_{i=1}^{n} \mathbb{I}_{(c_i=1)},$$
$$s_1^x = \sum_{i=1}^{n} \mathbb{I}_{(c_i=1)} x_i.$$

Siguiendo las expresiones anteriores para cada $i \in \{1, \ldots, n\}$ y realizando m simulaciones de este proceso se puede generar la muestra de Gibbs con la que se estima el valor de θ mediante la media de la muestra. El Algoritmo 9 muestra el proceso a seguir para estimar la media de la mezcla de distribuciones normales.

Algoritmo 9 Método de muestreo de Gibbs

1: Iniciar θ_0 muestreando de la distribución a priori.

$$\theta_0 \sim Normal(\mu_0, \sigma_0^2)$$

2: Para cada $t \in 1, \ldots, m$:
   a: Generar $c_i^{(t)}$, $(i = 1, \ldots, n)$, a partir de una distribución Bernoulli

$$c_i^{(t)} \sim Bernoulli(r_i)$$

   b: Calcular

$$n_1^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{(c_i=1)},$$
$$(s_1^x)^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{(c_i=1)} x_i.$$

   c: Generar $\theta^{(t)}$ a partir de

$$Normal\left(\frac{\sigma_1^2 (s_1^x)^{(t)} + \sigma_0^2 \mu_0}{\sigma_1^2 n_1^{(t)} + \sigma_0^2}, \frac{1}{\sigma_1^2 n_1^{(t)} + \sigma_0^2}\right).$$

Al igual que el muestreo por importancia, éste método sólo estima integrales específicas, por lo que para compararlo con los demás algoritmos se calcula el valor absoluto de la diferencia entre la media exacta y la estimada.

4.1.3. Método de Laplace.

En la implementación del método de Laplace, se parte de la expansión de Taylor alrededor de un máximo, $\hat{\theta}$. De manera general se tiene que el polinomio de Taylor de segundo orden para la distribución conjunta de los datos y el parámetro es

$$\ln p(x, \theta) \approx \ln p(x, \hat{\theta}) + (\theta - \hat{\theta}) \left(\frac{\partial \ln p(x, \theta)}{\partial \theta}\right)\Big|_{\theta=\hat{\theta}} + \frac{(\theta - \hat{\theta})'}{2} \left(\frac{\partial^2 \ln p(x, \theta)}{\partial \theta \partial \theta'}\right)\Big|_{\theta=\hat{\theta}} (\theta - \hat{\theta})$$

Para obtener $\hat{\theta}$ se utiliza el algoritmo EM para maximizar la función de distribución conjunta, o equivalentemente la distribución posterior. Por lo que el segundo término de la ecuación anterior es igual a cero y se tiene

$$
\begin{aligned}
p(x,\theta) &\approx p(x,\hat{\theta}) \exp\left(\frac{1}{2}(\theta-\hat{\theta})'\mathbf{H}(\theta-\hat{\theta})\right), \\
p(x) &\approx \int_{\Theta} p(x,\hat{\theta}) \exp\left(\frac{1}{2}(\theta-\hat{\theta})'\mathbf{H}(\theta-\hat{\theta})\right) d\theta = p(x,\hat{\theta})(2\pi)^{\frac{n}{2}}|-\mathbf{H}|^{\frac{1}{2}}, \\
\mathbf{H} &= \left(\frac{\partial^2 \ln p(x,\theta)}{\partial\theta\partial\theta'}\right)\Big|_{\theta=\hat{\theta}}.
\end{aligned}
$$

Para el problema de la mezcla de dos distribuciones gausianas se construye una aproximación gausiana para la distribución posterior utilizando la curvatura de ésta en $\hat{\theta}$. Para ello se consideran las expresiones anteriores, por lo que la curvatura tiene la forma, (Minka, 2000)

$$
\mathbf{H} = \nabla_{\theta\theta'} \log p(\theta|x) = -\frac{1}{100}\mathbf{I} - \sum_i r_i \mathbf{I} + \sum_i (1-r_i)(x_i-\hat{\theta})(x_i-\hat{\theta})', \tag{4.1}
$$

$$
r_i = \frac{(1-\omega)N(x_i|\hat{\theta},\mathbf{I})}{(1-\omega)N(x_i|\hat{\theta},\mathbf{I}) + \omega N(x_i|\mathbf{0},10\mathbf{I})}.
$$

La matriz Hessiana $\mathbf{H}$, es una matriz semidefinida negativa si $\hat{\theta}$ es un máximo. Por lo tanto, la aproximación gausiana y la constante de normalización son

$$
p(\theta|x) \approx N(\hat{\theta}, -\mathbf{H}^{-1}), \tag{4.2}
$$

$$
p(x) \approx p(x,\hat{\theta})(2\pi)^{\frac{n}{2}}|-\mathbf{H}|^{-\frac{1}{2}}. \tag{4.3}
$$

La expresión de $\mathbf{H}$ para el problema de mezclas se obtiene de la siguiente forma

$$
\begin{aligned}
\mathbf{H} &= \nabla_{\theta\theta'} \log \frac{p(x|\theta)p(\theta)}{p(x)} \\
&= \nabla_{\theta\theta'} \log p(x|\theta) + \nabla_{\theta\theta'} \log p(\theta)
\end{aligned}
$$

Donde para el problema de la mezcla la distribución a priori es una distribución normal y la verosimilitud es la mezcla de dos distribuciones gausianas

$$
\begin{aligned}
p(\theta) &= N(\theta|\mathbf{0},100\mathbf{I}) \\
p(x|\theta) &= \prod_i p(x_i|\theta) = \prod_i \left[(1-\omega)N(x_i|\theta,\mathbf{I}) + \omega N(x_i|\mathbf{0},10\mathbf{I})\right]
\end{aligned}
$$

donde sus respectivas derivadas se obtienen como sigue

$$
\begin{aligned}
\nabla_{\theta\theta'} \log p(\theta) &= \nabla_{\theta\theta'}\left(-\frac{1}{2(100)}\theta'\mathbf{I}\theta\right) \\
&= \nabla_{\theta} - \frac{1}{200}[2\theta\mathbf{I}] = -\frac{1}{100}\mathbf{I}
\end{aligned}
$$

Para el gradiente de la verosimilitud

$$\begin{aligned}\nabla_{\theta\theta'}\log p(x|\theta) &= \sum_i \nabla_{\theta\theta'}\log\left[(1-\omega)N(x_i|\theta,\mathbf{I})+\omega N(x_i|\mathbf{0},10\mathbf{I})\right]\\ &= \sum_i\left[\nabla_\theta\frac{\nabla_{\theta'}(1-\omega)N(x_i|\theta,\mathbf{I})}{(1-\omega)N(x_i|\theta,\mathbf{I})+\omega N(x_i|\mathbf{0},10\mathbf{I})}\right]\\ &= \sum_i\left[\nabla_\theta\frac{(1-\omega)N(x_i|\theta,\mathbf{I})(x_i-\theta)}{(1-\omega)N(x_i|\theta,\mathbf{I})+\omega N(x_i|\mathbf{0},10\mathbf{I})}\right].\end{aligned}$$

Si se considera

$$s_i(\theta) = (1-\omega)N(x_i|\theta,\mathbf{I})+\omega N(x_i|\mathbf{0},10\mathbf{I}).$$

Entonces

$$\begin{aligned}\nabla_{\theta\theta'}\log p(x|\theta) = (1-\omega)\sum_i\left[\frac{N(x_i|\theta,\mathbf{I})x_i(x_i-\theta)' - N(x_i|\theta,\mathbf{I}) - N(x_i|\theta,\mathbf{I})\theta(x_i-\theta)'}{s_i(\theta)}\right]\\ -(1-\omega)\sum_i\left[\frac{(N(x_i|\theta,\mathbf{I})(x_i-\theta))\nabla_\theta s_i(\theta)^2}{s_i(\theta)}\right].\end{aligned}$$

Donde la expresion del gradiente de la función $s_i(\theta)$ es

$$\nabla_\theta s_i(\theta) = \nabla_\theta(1-\omega)N(x_i|\theta,\mathbf{I}) = (1-\omega)N(x_i|\theta,\mathbf{I})(x_i-\theta)'.$$

Por lo cual se tiene lo siguiente

$$\begin{aligned}\nabla_{\theta\theta'}\log p(x|\theta) &= (1-\omega)\sum_i\left[\frac{N(x_i|\theta,\mathbf{I})(x_i-\theta)'(x_i-\theta)}{s_i(\theta)}\right] - (1-\omega)\sum_i\frac{N(x_i|\theta,\mathbf{I})}{s_i(\theta)}\\ &\quad -\sum_i(1-\omega)^2\frac{N(x_i|\theta,\mathbf{I})N(x_i|\theta,\mathbf{I})(x_i-\theta)'(x_i-\theta)}{s_i(\theta)^2}\\ &= \sum_i\left[\frac{(1-\omega)N(x_i|\theta,\mathbf{I})}{s_i(\theta)} - (1-\omega)^2\frac{N(x_i|\theta,\mathbf{I})N(x_i|\theta,\mathbf{I})}{s_i(\theta)^{2b}}\right](x_i-\theta)'(x_i-\theta)\\ &\quad -\sum_i r_i\\ &= -\sum_i r_i + \sum_i(1-r_i)r_i(x_i-\theta)'(x_i-\theta).\end{aligned}$$

De esta forma se obtiene la expresión de $\mathbf{H}$ en la ecuación (4.1).

Por lo tanto, para calcular la esperanza posterior y $p(x)$ basta con seguir las expresiones (4.2) y (4.3) respectivamente.

4.1.4. Variación de Bayes.

Para la variación de Bayes, se acota inferiormente el logaritmo de la verosimilitud mediante el acotamiento de cada término de los datos introduciendo una variable oculta c_i

$$\begin{aligned}
\ln p(x_i|\theta) &= \ln \sum_j^k p(x_i|c_i = j, \theta) p(c_i = j|\theta) \\
&= \ln \left(\mathbb{E}_{c|\theta}[p(x_i|c_i = j, \theta)] \right) \\
&\geq \mathbb{E}_{c|\theta}[\ln p(x_i|c_i = j, \theta)] \\
&= \sum_j^k \ln \left(\frac{p(x_i|c_i = j, \theta))}{p(c_i = j|\theta)} \right) p(c_i = j|\theta) \\
&= \ln \left(\prod_j^k \left[\frac{p(x_i|c_i = j, \theta)}{q_{ij}} \right]^{q_{ij}} \right).
\end{aligned}$$

Donde $\sum_j q_{ij} = 1$.

Lo anterior se deriva de la desigualdad de Jensen y de la inclusión de una variable oculta c_i que indica si la observación pertenece a la $j-$ésima distribución en la mezcla, $j \in \{1, 2, \ldots, k\}$. En el caso de la mezcla aquí presentada se considera $k = 2$. Por lo tanto, bajo estas consideraciones se tiene que

$$p(x_i|\theta) \geq \left(\frac{(1-\omega) N(x|\theta, \mathbf{I})}{q_{i1}} \right)^{q_{i1}} \left(\frac{\omega N(x|0, 10\mathbf{I})}{q_{i2}} \right)^{q_{i2}}$$

Los parámetros de variación q_{ij} son optimizados de tal forma que se maximiza la cota inferior. Dados los límites, la distribución gausiana posterior para θ es, (Minka, 2001)

$$\begin{aligned}
K_j &= \sum_i q_{ij} \\
\overline{x}_j &= \frac{1}{K_j} \sum_i q_{ij} x_i \\
S_j &= \sum_i q_{ij} (x_i - \overline{x}_j)(x_i - \overline{x}_j)' \\
\mathbf{V} &= \left(K_1 \mathbf{I} + \frac{\mathbf{I}}{100} \right)^{-1} \\
\mathbf{m} &= \mathbf{V} K_1 \overline{x}_1 \\
p(\theta|D) &\approx N(\mathbf{m}, \mathbf{V}) \\
p(D) &\geq \frac{1}{(2\pi)^{\frac{n(K_1-1)}{2}} K_1^{\frac{n}{2}}} N(\overline{x}_1, \mathbf{0}, \frac{\mathbf{I}}{K_1} + 100) \exp -\frac{1}{2} tr(S_1) \\
&\quad \frac{1}{(2\pi 10)^{\frac{n(K_2-1)}{2}} K_2^{\frac{n}{2}}} N(\overline{x}_2, \mathbf{0}, \frac{10\mathbf{I}}{K_2}) \exp -\frac{1}{2} tr(\frac{S_2}{10}) \\
&\quad \prod_i (\frac{\omega}{q_{ij}})^{q_{i1}} (\frac{1-\omega}{q_{i2}})^{q_{i2}}
\end{aligned}$$

Una vez desarrollados los métodos para el problema de la mezcla de dos gausianas es posible realizar la comparación de los métodos presentados, en un nivel de precisión y costo computacional.

4.2. Comparación

Se muestran los resultados de la implementación de cada uno de los algoritmos estudiados en los capitulos anteriores. La comparación se hace con base en los resultados de cada uno en la estimación de la esperanza posterior bajo el modelo de mezcla de dos distribuciones gausianas y el término de normalización.

El modelo de mezclas para el cual se hace la comparación es

$$p(x|\theta) = 0.5N(x|\theta, 1) + 0.5N(x|0, 10)$$

con $\theta = 2$.

Utilizando las expresiones presentadas en este capítulo para cada uno de los algoritmos y el lenguaje de programación R, se realizaron las aproximaciones para dos tamaños de muestra distintos, $n = 200$ y $n = 100$.

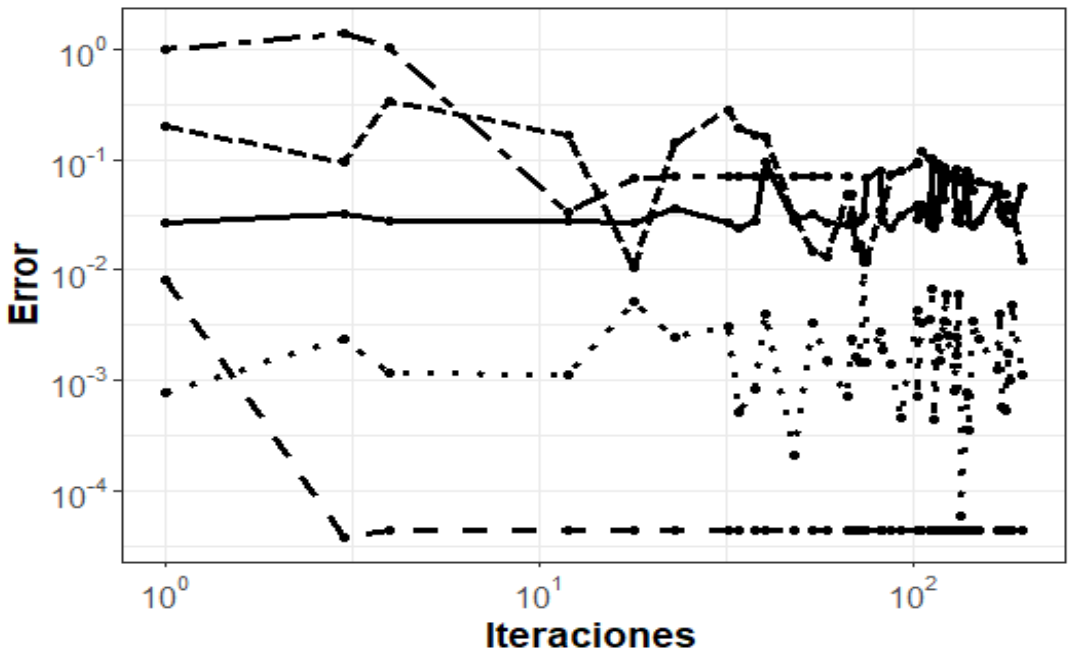

Figura 4.1: Gráfica de error vs. iteraciones en la aproximación de la media posterior para cada metodología, con un tamaño de muestra de $n = 200$. El método EP corresponde a la línea discontinua, el método de Gibbs corresponde a la línea punteada, la línea continua corresponde al muestreo por importancia, el método de variación de Bayes corresponde a la línea de doble discontinuidad y el método de Laplace corresponde a la línea compuesta por puntos y guiones.

En la Figura (4.1), se puede observar que el algorítmo de propagación de esperanzas (EP) tiene el menor error y converge de manera rápida. En la línea que corresponde al método

EP, el primer punto es la aproximación realizada por la densidad de emisión (ADF). También se puede apreciar la estabilidad de la metodología de propagación de esperanzas con respecto a la de los demás algoritmos, ya que estos presentan mayores variaciones con forme incrementan las iteraciones.

Por otro lado, en la Figura (4.2) se puede apreciar la comparación de los algortimos con los que se puede aproximar la evidencia del modelo, $p(x)$, los cuales son: variación de Bayes, muestreo por importancia, Laplace, densidad de emisión y propagación de esperanzas.

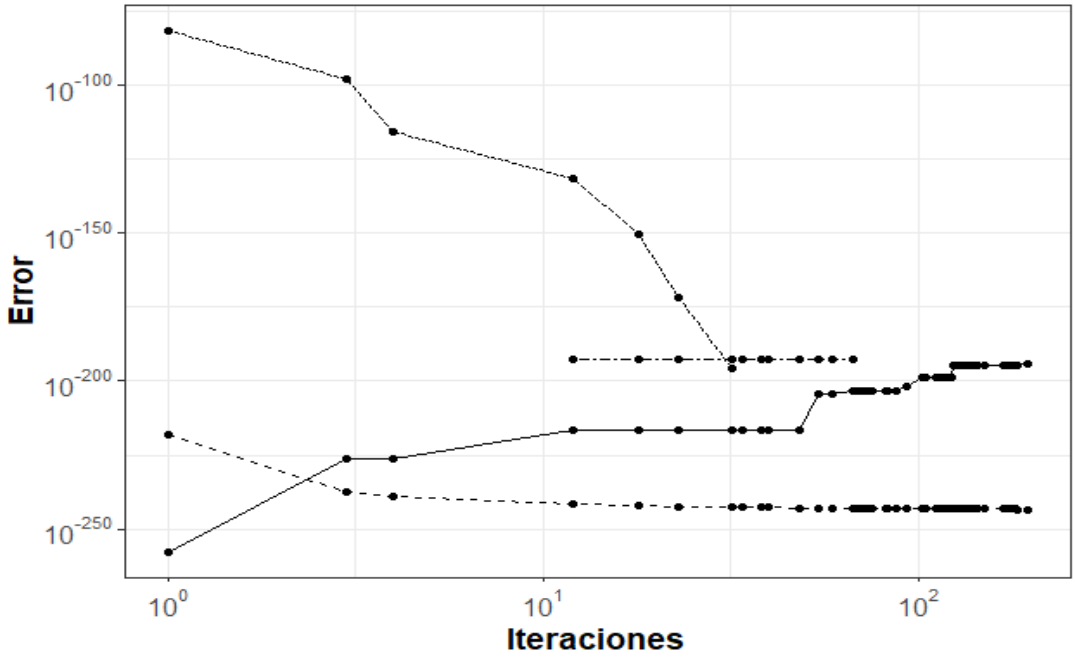

Figura 4.2: Gráfica de error vs. iteraciones en la aproximación de la evidencia del modelo para cada metodología, con un tamaño de muestra de $n = 200$. El método EP corresponde a la línea discontinua, la línea continua corresponde al muestreo por importancia, el método de variación de Bayes corresponde a la línea de doble discontinuidad y método de Laplace corresponde a la línea compuesta por puntos y guiones.

En esta gráfica se puede observar, al igual que con la media, que el método de propagación de esperanzas aproxima de mejor forma y con una convergencia más rápida.

Por otro lado, es interesante observar que las características propias de cada metodología se ven presentes en estas gráficas, ya que métodos como EP, Laplace y variación de Bayes asumen una aproximación gausiana, y su eficiencia se ve condicionada a qué tanto se sostenga ese supuesto. Por ello, en las Figuras (4.1) y (4.2) se puede apreciar que tienen una mayor eficiencia que los métodos de muestreo, los cuales hacen menos supuestos sobre la distribución posterior y por lo cual no pueden explotar esas características. Sin embargo, para distribuciones más complejas como distribuciones multimodales, éstos supuestos se convierten en ventajas para los métodos de muestreo, ya que al no asumir una distribución sobre la posterior no se pierden en máximos locales.

De igual forma los resultados se mantienen para un tamaño de muestra menor, en la Figura (4.3), se puede apreciar que se tiene el mismo comportamiento descrito anteriormente para una muestra de tamaño 100.

Como se pudo observar para el caso aquí expuesto, propagación de esperanzas ofrece una

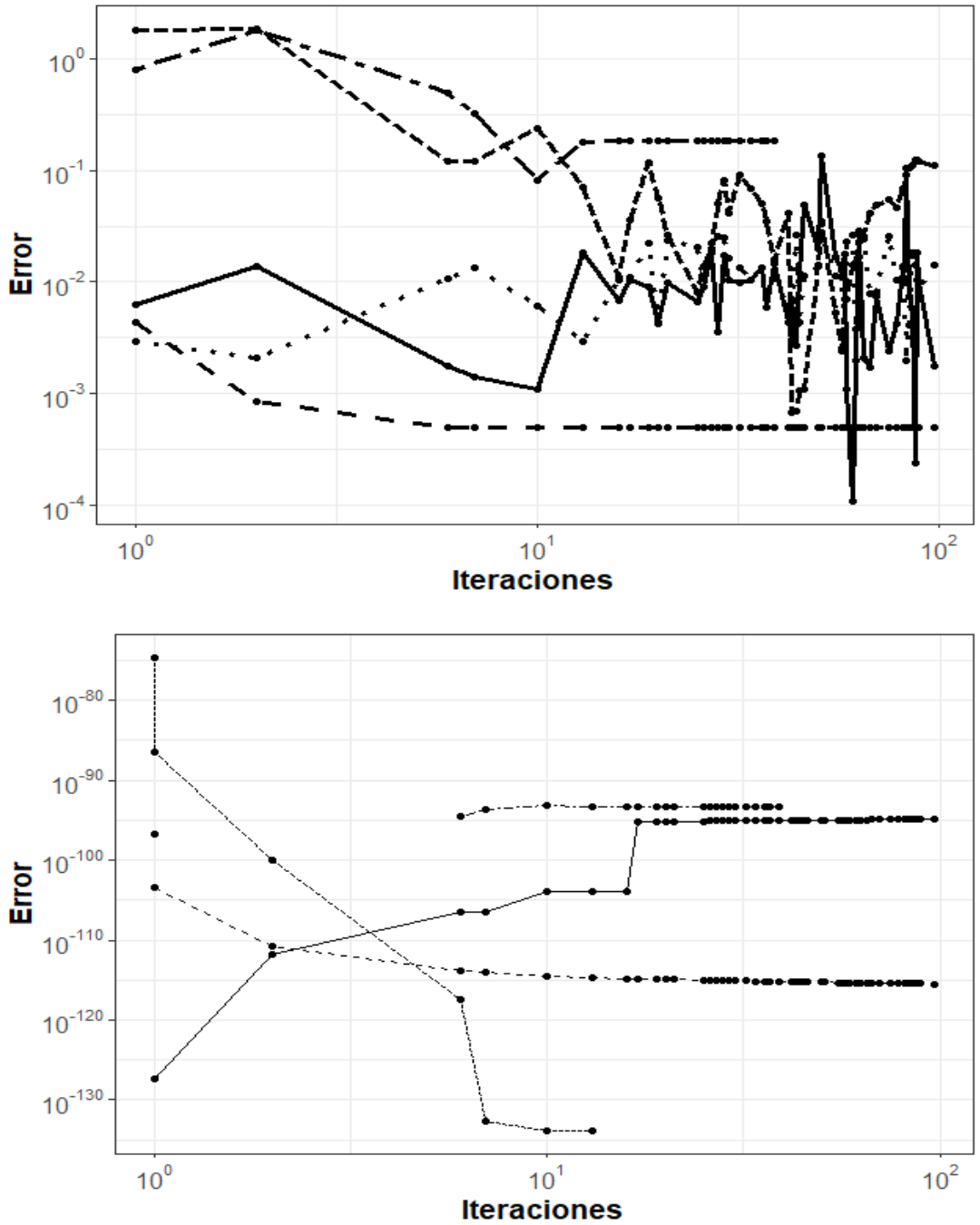

Figura 4.3: Gráficas de error vs. iteraciones en la aproximación de la media posterior y evidencia del modelo para cada metodología, con un tamaño de muestra de $n = 100$. El método EP corresponde a la línea discontinua, el método de Gibbs corresponde a la línea punteada, la línea continua corresponde al muestreo por importancia, el método de variación de Bayes corresponde a la línea de doble discontinuidad y método de Laplace corresponde a la línea compuesta por puntos y guiones.

alternativa con un menor costo computacional y que además proporciona una precisión mayor o igual a los demás métodos. Sin embargo, para problemas más complejos como distribuciones multimodales, puede no llegar a converger. Para estos casos se recomiendan métodos de muestreo que puedan adoptar formas complejas.

Capítulo 5

Conclusiones

En esta tesis se abordaron los principales métodos de aproximación bayesiana como: muestreo por importancia, muestreo de Gibbs, el método de Laplace, variación de Bayes y densidad de emisión. Además se introdujo un algoritmo de aproximación basado en la extensión de la densidad de emisión (ADF), mediante el refinamiento iterativo, propuesto por Minka (2001). Esta extensión corrige la dependencia en el orden de los datos con la que cuenta ADF, además de que incrementa la precisión en la estimación.

Dicha metodología de propagación de esperanzas fue puesta a prueba junto con las otras técnicas de aproximación estudiadas, y se demostró que es superior tanto en precisión como en velocidad de convergencia. Esto tomando el caso de una mezcla gausiana unidimensional de dos componentes.

En general, el algoritmo EP es el más eficiente al momento de estimar distribuciones con formas simples. Sin embargo, para distribuciones con más de una moda tiene problemas de convergencia, para dichas distribuciones una metolodología basada en muestreo es más eficiente.

Finalmente, la propagación de esperanzas como método de inferencia es el más eficiente comparado con los demás algoritmos. Además, para temas de clasificación, se demuestra su eficacia en conjunto con el método de Máquina de punto de Bayes, *Bayes point machine*, comparado con la máquina de soporte vectorial, *support vector machine SVM*, (Minka, 2001). Por lo que el algoritmo EP provee una nueva alternativa en la estimación bayesiana que además de ser muy precisa es de bajo costo, computacionalmente hablando.

Apéndice A

Esperanza-Maximización

El algoritmo EM es un procedimiento iterativo para calcular el estimador máximo verosimil en presencia de datos faltantes o de variables ocultas.

Cada iteración del algoritmo consiste en dos procesos, de los cuales deriva su nombre, el paso E y el paso M. En el paso E, o el paso de la esperanza, los datos faltantes o las variables ocultas son estimadas dados los datos observados y los parámetros del modelo son estimados. Lo cual se logra utilizando la esperanza condicional. En el paso M, o paso de maximización, la función de verosimilitud es maximizada bajo el supuesto que los datos faltantes o las variables ocultas son conocidas. Las estimaciones realizadas en el paso E son utilizadas para realizar la maximización.

En esencia, en el paso E se crea un límite inferior para la función de verosimilitud y posteriormente éste es maximizado en el paso M. Para cada iteración i, sea z_i las variables latentes y Q_i la función de densidad de las mismas, entonces para cada i se tiene que la log verosimilitud de los datos está acotada inferiormente como sigue

$$l(\theta) = \sum_i \ln p(x_i|\theta) \tag{A.1}$$

$$= \sum_i \ln \int_Z p(x_i, z_i|\theta) dz_i \tag{A.2}$$

$$= \sum_i \ln \int_Z Q_i(z_i) \frac{p(x_i, z_i|\theta)}{Q_i(z_i)} dz_i \tag{A.3}$$

$$\geq \sum_i \int_Z Q_i(z_i) \ln \frac{p(x_i, z_i|\theta)}{Q_i(z_i)} dz_i \tag{A.4}$$

donde el último paso se debe a la desigualdad de Jensen.[1]

Ahora, como se puede observar la distribución Q_i da una cota inferior para la verosimilitud y esto se cumple para cualquier distribución Q_i. Sin embargo, a pesar de existir una infinidad de distribuciones con las que es posible trabajar, se debe elegir aquella que alcance la igualdad en la expresión anterior para un valor particular de θ, ya que se desea maximizar el límite inferior.

Para obtener la igualdad en (A.4) se debe analizar la desigualdad de Jensen, la cual se

[1] Sea $f(x)$ una función cóncava, entonces para la variable aleatoria X se tiene que $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$

sostiene como igualdad cuando la esperanza es calculada sobre una constante

$$\frac{p(x_i, z_i|\theta)}{Q_i(z_i)} = c$$

para alguna constante c, que no dependa de z_i, por lo que se tiene

$$Q_i(z_i) \propto p(x_i, z_i|\theta)$$

Además, dado que Q_i es una distribución su integral sobre el dominio de z_i es igual a uno, con lo cual es posible calcular la constante como sigue

$$\begin{aligned} 1 &= \int_Z Q_i(z_i) dz_i \\ &= c \int_Z p(x_i, z_i|\theta) \end{aligned}$$

Entonces la distribución Q_i que debe elegirse es aquella que cumpla

$$\begin{aligned} Q_i(z_i) &= \frac{p(x_i, z_i|\theta)}{\int_Z p(x_i, z_i|\theta) dz_i} \\ &= \frac{p(x_i, z_i|\theta)}{\int_Z p(x_i|\theta) dz_i} \\ &= p(z_i|x_i, \theta) \end{aligned}$$

Por lo tanto la distribución Q_i es la distribución posterior de z_i dado x_i y un valor de θ. Así, la expresión (A.4) se cumple en igualdad.

Una vez realizado el paso E, en el paso M se maximiza la expresión (A.4) con respecto al parámetro para obtener un nuevo ajuste de θ.

Realizar estos pasos iterativamente constituye al algoritmo EM, cuyo algoritmo se puede expresar como sigue.

Algoritmo 10 Maximización de Esperanzas

Paso E: Para cada i, defina:

$$Q_i(z_i) := p(z_i|x_i, \theta)$$

Paso M: Estime θ maximizando:

$$\theta = arg\,\max_{\theta} \sum_i \int_Z Q_i(z_i) \ln \frac{p(x_i, z_i|\theta)}{Q_i(z_i)} = 0$$

Apéndice B

Familias exponenciales

Definición B.1 *Se define a la familia exponencial de distribuciones de probabilidad como aquellas cuya densidad tiene la forma general:*

$$p(x|\theta) = h(x)\exp\Big(\theta^{'}T(x) - A(\theta)\Big) \tag{B.1}$$

Para un vector de paráetros θ, referido como el parámetro canónico, y para funciones dadas T y h.

En la definición de familia exponencial, la estadística $T(x)$ es referida como la estadística suficiente, mientras que la función $A(\theta)$ es conocida como la función cumulante, la cual guarda la siguente relación:

$$A(\theta) = \log \int h(x)\exp\Big(\theta^{'}T(x)\Big)dx$$

que se obtiene integrando la ecuación (B.1). Esto muestra que $A(\theta)$ no es un grado de libertad en la especificación del la familia exponencial, sino que está especificada una vez que $T(x)$ y $h(x)$ son determinadas.

B.1. Forma exponencial de la distribución Normal multivariada

La distribución Normal Multivariada $n-$dimensional puede ser escrita de la siguiente forma

$$\begin{aligned}
p(x|\mu,\Sigma) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^{'}\Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}(x^{'}\Sigma^{-1}-\mu^{'}\Sigma^{-1})(x-\mu)-\frac{1}{2}\ln(|\Sigma|)\right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}(x^{'}\Sigma^{-1}x-2x^{'}\Sigma^{-1}\mu+\mu^{'}\Sigma^{-1}\mu)-\frac{1}{2}\ln(|\Sigma|)\right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}(x^{'}\Sigma^{-1}x-2x^{'}\Sigma^{-1}\mu)-\frac{1}{2}\mu^{'}\Sigma^{-1}\mu-\frac{1}{2}\ln(|\Sigma|)\right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left((\Sigma^{-1}\mu,-\frac{1}{2}\Sigma^{-1})(x,x^{'}x)^{'}-\frac{1}{2}\mu^{'}\Sigma^{-1}\mu-\frac{1}{2}\ln(|\Sigma|)\right)
\end{aligned}$$

Así se tiene que para la distribución Normal multivariada

$$\begin{aligned}
h(x) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \\
\theta &= \begin{pmatrix}\Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1}\end{pmatrix} \\
T(x) &= \begin{pmatrix}x \\ x^{'}x\end{pmatrix} \\
A(\theta) &= \frac{1}{2}\mu^{'}\Sigma^{-1}\mu+\frac{1}{2}\ln(|\Sigma|)
\end{aligned}$$

Bibliografía

Beal, M. (2003). *Variational algorithms for approximate bayesian inference.* PhD thesis, University College London.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer, New York, US.

Casella, G. and George (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Šmídl, V. (2004). *The variational bayes approach in signal processing.* PhD thesis, University of Dublin.

Davis, P. and Rabinowitz, P. (1984). *Methods of numerical integration.* Academic press, CA, US., second edition.

Hammersle, J. and Handscomb, D. (1964). *Monte Carlo methods.* Methuen and Co ltd, GB., first edition.

Jayanta, K. G., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods.* Springer, New York, US., first edition.

Kass, R. and Raftery, A. (1993). Bayes factors and model uncertainty. Technical Report 254, University of Washington.

Minka, T. (2000). Variational bayes for mixtures models: Reversing em.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology.

Ramakrishnan, N., Ertin, E., and Moses, R. L. (2011). Ieee statistical signal processing workshop. In *Assumed density filtering for learning gaussian process models.*

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods.* Springer, New York, US., second edition.

Zhihui, L. (2010). Bayesian mixture models. Master's thesis, McMaster University.