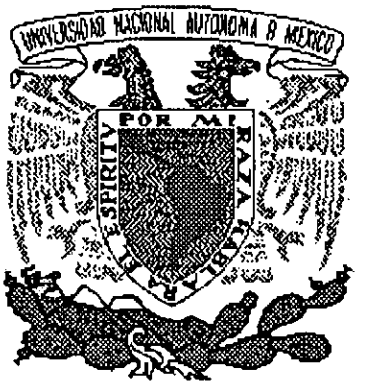
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

U.A.C.P. y P. I.I.M.A.S.



INFERENCIA BAYESIANA A PARTIR DE DISTRIBUCIONES FINALES MULTIMODALES

MAESTRIA EN ESTADISTICA E INESTIGACION DE OPERACIONES

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA

LUIS ENRIQUE NIETO BARAJAS

MEXICO, D.F., 1998

PAG. DESCONTINUADA

A mi niña Lyn.

Reconocimientos

◇ A mi esposa Lyn, por el amor, apoyo y comprensión que me brindó durante la realización de este trabajo.

◇ A mis padres Jorge A. y Ma. Teresa, por el cariño que me brindaron para realizar mis estudios de maestría.

◇ A mi asesor Eduardo Gutiérrez Peña, por el tiempo y conocimiento que me compartió para realizar este trabajo de investigación.

◇ Un especial reconocimiento y agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) que ayudó a sustentar mis estudios de maestría y finalmente poder concluir con este trabajo de investigación.

◇ De igual manera, agradezco al Instituto tecnológico Autónomo de México (ITAM) y en especial al departamento de Estadística por las facilidades brindadas para la realización de este trabajo.

INFERENCIA BAYESIANA A PARTIR DE DISTRIBUCIONES FINALES MULTIMODALES.

En el caso de distribuciones finales unimodales, la aproximación normal asintótica ha jugado tradicionalmente un papel muy importante para poder realizar inferencias, especialmente en casos donde la distribución final de que se trate no es fácil de manejar. Otra forma de realizar inferencias aproximadas es usando el método de Laplace, que se basa en un desarrollo similar al de la aproximación normal. A pesar de que existen otras aproximaciones numéricas quizá más precisas para obtener momentos (por ejemplo vía simulación), el método de Laplace ha tenido un fuerte impacto en el desarrollo de la teoría Bayesiana debido a su utilidad para encontrar aproximaciones analíticas a densidades marginales.

En algunas aplicaciones Bayesianas surgen, como resultado del análisis, distribuciones finales multimodales. En este trabajo se tratará, por una parte, de generalizar la aproximación normal a una densidad final multimodal mediante una mezcla de densidades normales, y por otra parte, de aplicar el método de Laplace o una extensión de él a densidades finales multimodales.

Contenido

Capítulo 1

Introducción.

1.1 Inferencia Bayesiana.

Sea $\mathcal{F} = \{p(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ un modelo paramétrico indexado por $\boldsymbol{\theta}$, donde Θ es un espacio paramétrico de dimensión k. Sea $\mathbf{X}' = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria de $p(x|\boldsymbol{\theta})$ y suponga que el conocimiento inicial sobre $\boldsymbol{\theta}$ se puede representar a través de una distribución de probabilidad $p(\boldsymbol{\theta})$.

El mecanismo de la inferencia Bayesiana consiste en combinar la información inicial que se tiene sobre el parámetro desconocido $\boldsymbol{\theta}$ con la información obtenida a partir de los datos observados, cuyo modelo de probabilidad está indexado por $\boldsymbol{\theta}$, para obtener una descripción del conocimiento final acerca del valor del parámetro de interés $\boldsymbol{\theta}$.

La forma de obtener la descripción conjunta de la incertidumbre sobre los valores del parámetro desconocido es mediante el Teorema de Bayes.

Teorema de Bayes.

La distribución final de $\boldsymbol{\theta}$, después de observar la muestra $\mathbf{X} = \mathbf{x}$, está dada por

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{x})}.$$

Otra forma de escribir el resultado anterior es

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}),$$

dado que el factor $p(\mathbf{x})$ no depende de $\boldsymbol{\theta}$.

El teorema anterior permite obtener una descripción actualizada del conocimiento acerca del valor del parámetro desconocido $\boldsymbol{\theta}$. La distribución final contiene toda la información disponible sobre el valor desconocido de $\boldsymbol{\theta}$. De esta manera, cualquier tipo de inferencia acerca del parámetro $\boldsymbol{\theta}$ se obtiene a partir de su distribución final.

Los problemas básicos de inferencia, como la estimación puntual, la estimación por intervalo y las pruebas de hipótesis, se pueden atacar directamente resumiendo algunas características de la distribución final. Para el problema de estimación puntual, por ejemplo, se puede usar como estimador de $\boldsymbol{\theta}$ una medida de localización de la distribución final tal como la media, la mediana, o la moda . Si la estimación deseada es mediante un rango de valores posibles en donde se pueda encontrar el verdadero valor del parámetro, se puede obtener un intervalo de máxima densidad directamente de la distribución final. Finalmente, para realizar una prueba de hipótesis una posibilidad es calcular las probabilidades finales de cada hipótesis y compararlas entre sí de acuerdo con el contexto del problema.

Si $\boldsymbol{\theta}$ es un vector de dimensión k, digamos $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_k)$ y solamente θ_1 es de interés, entonces se puede obtener la distribución marginal final de θ_1 a partir de la distribución final del vector $\boldsymbol{\theta}$, integrando sobre la región correspondiente los parámetros que no son de interés, es decir,

$$p(\theta_1 | \mathbf{x}) = \int \cdots \int p(\boldsymbol{\theta} | \mathbf{x}) \, \partial\theta_2 \cdots \partial\theta_k.$$

Para hacer inferencias acerca de una observación futura x_F de la variable aleatoria X, se puede obtener la distribución predictiva para X_F de la siguiente forma:

$$p(x_F | \mathbf{x}) = \int p(x_F | \boldsymbol{\theta}) \, p(\boldsymbol{\theta} | \mathbf{x}) \, \partial\boldsymbol{\theta}.$$

A partir de la distribución predictiva, se puede realizar cualquier tipo de inferencia acerca del valor que observará la variable aleatoria X_F.

La mayoría de los problemas de inferencia Bayesiana se concretan en calcular ciertas características (que llamaremos *resúmenes inferenciales*) que resumen el comportamiento de la distribución final del parámetro de interés. En la mayoría de los casos los resúmenes inferenciales básicos se reducen a integrales de la forma

$$S_J[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) \, p(\boldsymbol{\theta}) \, p(\mathbf{x} | \boldsymbol{\theta}) \, \partial\boldsymbol{\theta}_{J^c},$$

donde $g : \Re^k \to \Re$, $J \subseteq \{1, \ldots, k\}$, $J^c = \{1, \ldots, k\} \setminus J$ y $\boldsymbol{\theta}_{J^c} = \{\theta_j : j \in J^c\}$. Por ejemplo, tenemos que

$$\begin{aligned} p(\mathbf{x}) &= S_\varnothing[1]; \\ E(\theta_i \theta_j | \mathbf{x}) &= \frac{S_\varnothing[\theta_i \theta_j]}{S_\varnothing[1]}; \\ p(\theta_j | \mathbf{x}) &= \frac{S_{\{j\}}[1]}{S_\varnothing[1]}; \end{aligned}$$

$$\Pr(\boldsymbol{\theta} \in A) = \frac{S_{\varnothing}[I_A(\boldsymbol{\theta})]}{S_{\varnothing}[1]}; \quad \text{y}$$
$$p(x_F|\mathbf{x}) = \frac{S_{\varnothing}[p(x_F|\boldsymbol{\theta})]}{S_{\varnothing}[1]},$$

donde $I_A(\boldsymbol{\theta})$ denota la función indicadora del conjunto A.

La característica esencial de los métodos Bayesianos es el uso explícito de una medida de probabilidad para cuantificar la incertidumbre que se tiene sobre cantidades desconocidas. Esta cuantificación mediante probabilidades se basa en una interpretación subjetiva de la probabilidad.

En la práctica, generalmente no es fácil cuantificar la incertidumbre o conocimiento inicial acerca del parámetro de interés mediante un modelo de probabilidad. Sin embargo, se han propuesto diversos métodos para atacar este problema (ver, por ejemplo, Diaconis e Ylvisaker, 1985 y Kadane, 1980).

Una característica deseable del proceso de actualización de la información es que la distribución inicial dé lugar a una distribución final que sea tratable, en el sentido de que se pueda integrar analíticamente. Además de la tratabilidad, se requiere que la familia de distribuciones iniciales $p(\boldsymbol{\theta})$ sea suficientemente flexible, de tal forma que el conocimiento inicial del investigador pueda ser bien representado por un miembro de esa familia. Esto ha motivado el uso de las llamadas *familias conjugadas*. El lector interesado puede consultar Bernardo y Smith (1994) y DeGroot (1970).

En ocasiones la información inicial sobre el verdadero valor del parámetro es demasiado vaga o, por diversas razones, se desea hacer las inferencias únicamente a partir de los datos. Para estos casos se han propuesto cierto tipo de distribuciones iniciales que reflejaran dicha ausencia de información inicial. Tales ditribuciones son

llamadas *distribuciones iniciales no informativas.* Jeffreys (1946) propuso una familia de distribuciones iniciales no informativas invariantes ante reparametrizaciones del modelo. Por su parte, Bernardo (1979) propuso un método más general para obtener distribuciones no informativas, basadas en una definición formal del concepto de información. Este tipo de distribuciones son llamadas *distribuciones iniciales de referencia.*

Muchos de los modelos paramétricos utilizados en la práctica dan como resultado distribuciones finales relativamente fáciles de manejar, con un comportamiento suave y la mayoría de las veces unimodal. En este caso, un estimador puntual obtenido a partir de la distribución final, podría ser la moda. La estimación puntual mediante la moda es muy útil, ya que en el caso de que la distribución final no sea fácil de manejar analíticamente, se pueden utilizar métodos aproximados, para obtener características de la distribucion final, que se basan en estimaciones modales.

1.2 Multimodalidad.

En algunas ocasiones, como resultado del análisis Bayesiano, se obtienen distribuciones finales multimodales. Esta multimodalidad de la distribución final se puede deber, por un lado, a que se proponga porque así lo requiere el problema, una distribución inicial multimodal. Por otro lado, la multimodalidad de la distribución final se puede deber a la verosimilitud del problema específico.

Para ilustrar el primer caso, se presenta el siguiente ejemplo (modificación del ejemplo presentado en Diaconis e Ylvisaker, 1985).

Ejemplo 1.2.1.

Se tiene un experimento que consiste en girar una moneda n veces sobre una mesa y registrar el número de soles. Se desea hacer inferencia sobre el parámetro θ, la probabilidad de observar un sol al girar una moneda sobre una mesa.

Sea $X_i = \left\{ \begin{array}{l} 1, \text{ si sale sol} \\ \\ 0, \text{ e.o.c.} \end{array} \right\}$ una variable aleatoria tal que,

$$P[X_i = 1] = \theta = 1 - P[X_i = 0],$$

donde $\theta \in (0,1)$, entonces, $p(x_i|\theta) = Ber(x_i|\theta)$.

Se tienen $X_1, X_2, \ldots, X_n$ v.a. independientes, por lo tanto,

$$p(\mathbf{x}|\theta) = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}.$$

Para este problema se sabe, por experiencia, que la proporción de soles al girar una moneda sobre una mesa un número grande de veces se encuentra alrededor de $1/3$ o bien de $2/3$. Para reflejar este conocimiento inicial sobre θ se propone una mezcla de dos densidades Beta con distintos parámetros para obtener un comportamiento bimodal. Entonces, la distribución inicial de θ es

$$p(\theta) = 0.5\beta(\theta|10,20) + 0.5\beta(\theta|20,10),$$

donde, $\beta(\theta|a,b)$ es la función de densidad Beta con parámetros a y b.

Al realizar el experimento, se observaron 4 soles en 10 giros de la moneda. Combinando la información inicial con la verosimilitud, se obtiene que la distribución final de θ es también una mezcla de dos densidades Beta con los siguientes parámetros:

$$p(\theta|\mathbf{x}) = 0.74\beta(\theta|14,26) + 0.26\beta(\theta|24,16).$$

El comportamiento de las densidades se pueden ver en la Figura 1.1. ■

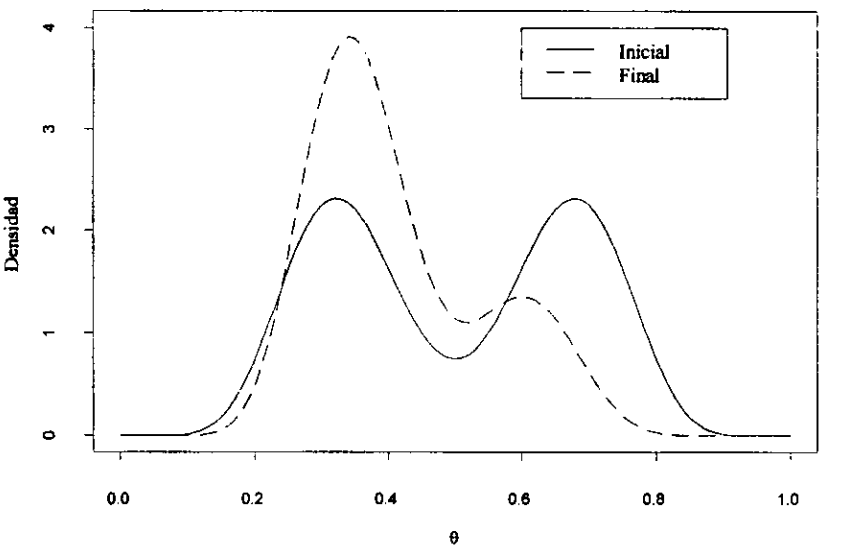


Figura 1.1: Densidades inicial y final de θ.

El segundo caso donde la multimodalidad de la distribución final se debe a la verosimilitud se ilustra con el siguiente ejemplo:

Ejemplo 1.2.2.

Sean $X_1, X_2, \ldots, X_n$ una muestra de variables aleatorias independientes con distribución $Cauchy(\theta, 1)$. El objetivo es obtener inferencias sobre el parámetro de localización θ. En este caso, si no se tiene ningún conocimiento inicial sobre θ, es común utilizar la distribución inicial no informativa

$$p(\theta) \propto 1,$$

la cual es impropia.

Entonces, la densidad final para θ es simplemente una función proporcional a la verosimilitud, por lo tanto,

$$p(\theta|\mathbf{x}) \propto \prod_{i=1}^{n} \frac{1}{\pi\left[1+(x-\theta)^2\right]}.$$

En el caso de que se tenga una muestra de tamaño 2 y se cumpla que $|x_1 - x_2| > 2$ entonces, la función de verosimilitud para θ es bimodal. Como en este caso la densidad final de θ es porporcional a la verosimilitud, la densidad final de θ también es bimodal.

El comportamiento de la distribución final se puede observar en la Figura 1.2, donde se muestran gráficas para distintos valores de x_1 y x_2. La constante de normalización de la densidad final se encontró numéricamente usando la regla trapezoidal.

■

En el caso de que la distribución final del parámetro de interés tenga un comportamiento unimodal, generalmente bastan unos cuantos momentos de esta distribución para dar una idea adecuada del comportamiento del parámetro. Estos momentos se pueden aproximar fácilmente mediante técnicas numéricas en el caso de que no se puedan obtener analíticamente. Además, esta información se puede complementar con los cuantiles de la distribución para tener una mejor idea de su comportamiento. Por otro lado, cuando el parámetro de interés es un vector, se pueden obtener y graficar las distribuciones marginales para cada componente del parámetro.

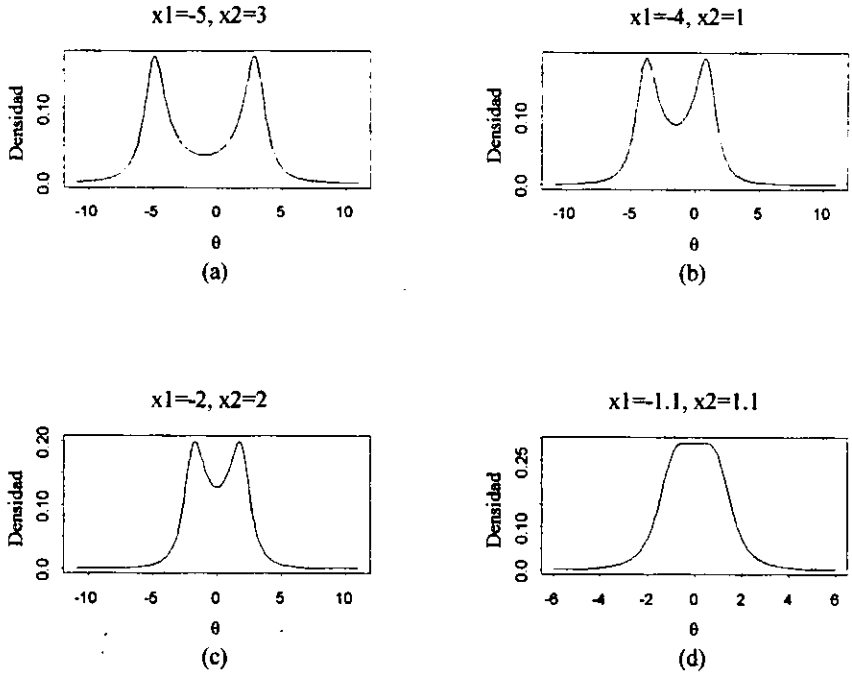


Figura 1.2: Densidad final de θ para distintos valores de x_1 y x_2.

Cuando la distribución final del parámetro de interés tiene un comportamiento más complicado, por ejemplo, si tiene dos o más modas, el esfuerzo computacional para encontrar una aproximación a las medidas de localización y de dispersión es mucho mayor. Más aún, incluso si se pudieran calcular de una manera relativamente sencilla, estas medidas no serían de mucha utilidad para describir el comportamiento final del parámetro. En ese caso, los cuantiles darían una mejor idea de la distribución de la probabilidad sobre el espacio parametral y se tendría una idea de las regiones de mayor densidad. Es necesario señalar que el cálculo de los cuantiles requeriría de un esfuerzo computacional mayor.

1.3 Discusión.

Como resultado del análisis Bayesiano y dada la complejidad de muchos de los modelos probabilísticos utilizados en la práctica, en algunos casos surgen distribuciones finales muy complejas. En ocasiones, ni siquiera es posible calcular analíticamente la constante de normalización, y si se desea, además, describir el comportamiento de la densidad mediante resúmenes inferenciales, es necesario recurrir a aproximaciones analíticas o incluso a aproximaciones numéricas.

Existe mucho trabajo realizado en el campo de aproximaciones analíticas de las densidades finales en el caso de que la distribución final tenga un comportamiento unimodal o al menos tenga una sola moda dominante. Una de las aproximaciones analíticas más usada es la aproximación normal asintótica, que realmente es una aproximación modal; es decir, alrededor de la moda esta aproximación es adecuada, pero lejos de ella la aproximación puede no serlo.

Otra aproximación analítica importante es la aproximación de Laplace, que es una regla de integración basada, al igual que la aproximación normal asintótica, en una expansión en la serie de Taylor del logaritmo de la densidad final. Este tipo de aproximación con el método de Laplace ha tenido mucha importancia por la forma en que se aplica al cálculo de densidades marginales, dando como resultado una aproximación analítica a la densidad marginal.

En este campo de las aproximaciones analíticas a densidades finales, existe muy poco trabajo realizado en el caso de que la densidad final sea multimodal. En O'Hagan (1994) se presenta una generalización a la aproximación normal asintótica medi-

ante una mezcla de densidades normales. Por otra parte, en la literatura de análisis numérico existen diversos métodos para calcular el valor de una integral a través de las llamadas *reglas de cuadratura*. Para una revisión de estos métodos se puede consultar Burden (1993). En aplicaciones estadísticas Naylor y Smith (1982) discuten un procedimiento iterativo basado en la regla de Gauss-Hermite. Este tipo de aproximaciones son muy precisas, pero carecen de la flexibilidad necesaria para muchas de las aplicaciones Bayesianas.

Otro tipo de aproximaciones numéricas son las que utilizan un método basado en simulación. Esta clase de aproximaciones son llamadas aproximaciones por *Monte Carlo*. Dentro de este tipo de aproximaciones por Monte Carlo existen diversas formas, como el método de muestreo por importancia (Wolpert, 1991), el método de muestreo-remuestreo (Rubin, 1988) y unas más sofisticadas que hacen uso de Cadenas de Markov como el Muestreador de Gibbs (Gelfand y Smith, 1990) y el método de Metropolis-Hastings (Metropolis *et al.*, 1953 y Hastings, 1970). Estos métodos de aproximación, a diferencia de las reglas de cuadratura, son muy usados en aplicaciones Bayesianas por su flexibilidad para calcular diversos resúmenes inferenciales con base en una sola muestra. Es necesario tomar en cuenta que a medida que la dimensión del parámetro aumenta, el cómputo necesario para generar una muestra es mucho mayor. Además, en el caso de multimodalidad, algunos métodos de este tipo pueden conducir a resultados completamente erróneos.

Ni las aproximaciones numéricas por cuadratura ni las aproximaciones por Monte Carlo, hacen distinción entre el tipo de comportamiento de la distribución final, es decir, son aplicables tanto a densidades finales unimodales como multimodales.

Finalmente, se puede decir que existe muy poca investigación realizada en cuanto a las aproximaciones analíticas en el caso de que la densidad final sea multimodal. El objetivo de este trabajo es ampliar la investigación realizada en ese sentido, evaluando la aproximación normal generalizada para el caso multimodal (propuesta en O'Hagan, 1994) y usándola como base de una generalización de la aproximación de Laplace.

Para realizar la investigación sobre las posibles aproximaciones analíticas en el caso multimodal y para poder evaluar de una manera objetiva la aproximación propuesta en O'Hagan (1994), es necesario revisar las aproximaciones analíticas ya existentes para el caso unimodal, como son la aproximación normal asintótica y la aproximación de Laplace.

Capítulo 2

Aproximaciones Analíticas (Caso unimodal).

2.1 Aproximación Normal asintótica.

Como se mencionó en el capítulo anterior, para poder encontrar resúmenes inferenciales de la distribución final muchas veces es necesario recurrir a aproximaciones, ya sea analíticas o numéricas, debido a la complejidad de la distribución final. La aproximación analítica más usada, y la más fácil de implementar, es la aproximación normal asintótica.

Existen resultados importantes acerca del comportamiento de la distribución final cuando el tamaño de muestra tiende a infinito. Es necesario hacer la distinción entre el caso de que el espacio parametral de $\boldsymbol{\theta}$ es numerable ($\boldsymbol{\theta}$ es una variable aleatoria discreta) y el caso de que es no numerable ($\boldsymbol{\theta}$ es una variable aleatoria continua). Dicha distinción es importante debido a que el comportamiento asintótico es diferente para

cada caso. El interés en este trabajo se centra esencialmente en espacios parametrales no numerables, es decir, cuando $\boldsymbol{\theta}$ es una variable aleatoria continua.

En el primer caso, el espacio parametral Θ es un conjunto numerable de valores (posiblemente finito), es decir, $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \}$, de manera que el modelo paramétrico correspondiente al verdadero parámetro $\boldsymbol{\theta}_t$, es "distinguible" de los otros valores posibles del parámetro, en el sentido de que las discrepancias

$$\int p(\mathbf{x}\,|\boldsymbol{\theta}_t) \log \left[\frac{p(\mathbf{x}\,|\boldsymbol{\theta}_t)}{p(\mathbf{x}\,|\boldsymbol{\theta}_i)}\right] \partial \mathbf{x}$$

son distintas de cero (i.e, >0), para toda $i \neq t$. Entonces, la distribución límite del parámetro $\boldsymbol{\theta}$ condicional al valor de la muestra $\mathbf{x}$, es una densidad degenerada en el verdadero valor del parámetro $\boldsymbol{\theta}_t$, es decir,

$$\lim_{n\to\infty} p(\boldsymbol{\theta}_t\,|\mathbf{x}) = 1, \qquad \lim_{n\to\infty} p(\boldsymbol{\theta}_i\,|\mathbf{x}) = 0, \qquad i \neq t.$$

La demostración de este resultado se puede encontrar en Bernardo y Smith (1994).

Sin importar la probabilidad inicial asignada al verdadero valor del parámetro, siempre y cuando ésta sea mayor a cero, el resultado anterior plantea que si se obtiene una muestra suficientemente grande de observaciones muestrales $\mathbf{x}$, entonces la probabilidad final del verdadero valor del parámetro tenderá a uno. Esto demuestra que la inferencia Bayesiana es inherentemente "consistente".

Consideremos ahora el caso en el que el espacio parametral Θ es un conjunto no numerable, es decir, que $\boldsymbol{\theta}$ sea una variable aleatoria continua. Supongamos además que $\Theta \subseteq \Re^k$. Uno de los resultados más importantes de la teoría asintótica en la estadística Bayesiana es el siguiente teorema:

Teorema 2.1.1.

Suponga que $X_1, \ldots X_n$ son variables aleatorias independientes de la densidad $p(x|\boldsymbol{\theta})$, de manera que $p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$. Suponga además que la densidad inicial $p(\boldsymbol{\theta})$ y la función de densidad conjunta $p(\mathbf{x}|\boldsymbol{\theta})$ son dos veces diferenciables alrededor de $\widehat{\boldsymbol{\theta}}$, la moda de la densidad final $p(\boldsymbol{\theta}|\mathbf{x})$. Entonces, bajo ciertas condiciones de regularidad y conforme $n \to \infty$

$$p(\boldsymbol{\theta}|\mathbf{x}) = N_k\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)\right)\left\{1 + O\left(n^{-1/2}\right)\right\}, \tag{2.1}$$

donde

$$\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right) = -\left[\frac{\partial^2 \log p\left(\widehat{\boldsymbol{\theta}}|\mathbf{x}\right)}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right]^{-1}.$$

Las condiciones de regularidad necesarias para que se dé la convergencia del Teorema 2.1.1 se pueden verificar en Bernardo y Smith (1994). Una demostración más formal y rigurosa de la normalidad asintótica se puede encontrar en Johnson (1970) y en Walker (1969).

Una justificación heurística del resultado anterior se puede lograr haciendo la expansión en serie de Taylor del logaritmo de la densidad final de $\boldsymbol{\theta}$ alrededor de su moda, como se muestra a continuación.

Sea

$$p_x(\boldsymbol{\theta}) = p(\boldsymbol{\theta})\, p(\mathbf{x}|\boldsymbol{\theta})$$

el kernel de la densidad final; por lo tanto,

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p_x(\boldsymbol{\theta}).$$

Entonces, la expansión en serie de Taylor de $\log p_x(\boldsymbol{\theta})$ alrededor de $\widehat{\boldsymbol{\theta}}$ es

$$\log p_x(\boldsymbol{\theta}) = \log p_x\left(\widehat{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) + R_n$$

donde

$$\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right) = -\left[\frac{\partial^2 \log p\left(\widehat{\boldsymbol{\theta}}|\mathbf{x}\right)}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right]^{-1} = -\left[\frac{\partial^2 \log p_x\left(\widehat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right]^{-1}.$$

Suponiendo que se cumplen las condiciones de regularidad que aseguran que R_n es pequeño para n grande, e ignorando las constantes de proporcionalidad, se tiene que

$$p_x(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right\}.$$

De aquí se observa que la densidad final de $\boldsymbol{\theta}$ es proporcional al kernel de una densidad Normal k-variada. Una vez calculada la constante de normalización necesaria se obtiene el resultado del Teorema 2.1.1. ■

Existen otras formas de la aproximación normal asintótica. Por ejemplo, para tamaños de muestra grandes, el efecto de la densidad inicial en la moda y en la matriz de dispersión $\boldsymbol{\Sigma}$ tenderá a ser pequeño comparado con el efecto de los datos, por lo que la densidad inicial se puede ignorar. Este cambio reemplazaría la moda posterior con el estimador máximo verosímil de $\boldsymbol{\theta}$ y la matriz de dispersión $\boldsymbol{\Sigma}$ con una nueva matriz que se puede expresar como menos la inversa de la matriz de información de Fisher observada.

Alternativamente, se podría reemplazar la moda posterior por la media posterior y la matriz de dispersión por menos la inversa de la matriz de información de Fisher esperada, evaluada en la media posterior.

Las dos formas alternativas de la aproximación normal asintótica descritas anteriormente son asintóticamente equivalentes a la del Teorema 2.1.1.

Si la aproximación (2.1) es adecuada, entonces prácticamente cualquier resumen inferencial de interés (como distribuciones marginales o momentos de funciones lineales de $\boldsymbol{\theta}$) se puede aproximar fácilmente.

En particular,

$$E(\boldsymbol{\theta}|\mathbf{x}) = \widehat{\boldsymbol{\theta}}\left\{1 + O\left(n^{-1}\right)\right\} \quad \text{y} \quad \text{Var}(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)\left\{1 + O\left(n^{-1}\right)\right\}.$$

Desafortunadamente, en aplicaciones específicas no siempre es fácil determinar si la aproximación normal es adecuada para el tamaño de muestra dado. Para tamaños de muestra pequeños, la aproximación normal no es más que una aproximación modal, es decir, alrededor de la moda la aproximación será razonable. Sin embargo, en este caso no hay garantía de que la aproximación sea adecuada en regiones alejadas de la moda.

Es conveniente mencionar que la aproximación normal es razonable en el caso de que la distribución final para algún tamaño de muestra determinado tenga una sola moda dominante, debido a que el teorema está basado en el hecho de que asintóticamente la distribución final tendrá una sola moda que tenderá al verdadero valor del parámetro. Hay que tomar en cuenta que para tamaños de muestra pequeños se pueden tener comportamientos bimodales o incluso multimodales, y en estos casos la

aproximación normal puede carecer de sentido.

En ocasiones es posible mejorar la precisión de la aproximación si se trabaja en términos de alguna transformación del parámetro de interés. El objeto de dicha transformación es que la distribución final del nuevo parámetro sea "más simétrica" que la del parámetro original y que su soporte no esté acotado.

Ejemplo 2.1.1.

Considere el Ejemplo 1.1, donde se quiere estimar la probabilidad de observar un sol al girar una moneda sobre una mesa. Supóngase que no existe razón suficiente para creer *a priori* que tal probabilidad está cargada hacia 1/3 ó 2/3, y que ahora se propone una distribución inicial no informativa uniforme sobre el intervalo $(0, 1)$, i.e.,

$$p(\theta) = I_{(0,1)}(\theta).$$

Al combinar esta información inicial con la verosimilitud de los datos y sabiendo que se observaron 4 soles en 10 eventos, se obtiene que la distribución final de θ es

$$p(\theta|\mathbf{x}) = Beta(\theta|5, 7).$$

Aplicando la aproximación normal (2.1) a la densidad final de θ se tiene que

$$p(\theta|\mathbf{x}) \approx N(\theta|0.4, 0.024).$$

Como la densidad normal toma valores en toda la recta real, se puede truncar esta aproximación para restringirla únicamente al rango $(0, 1)$ y renormalizarla dividiendo entre la probabilidad acumulada en el intervalo, que aproximadamente es 0.9950. Gráficas de la densidad final de θ y de su aproximación normal truncada se encuentran

en la Figura 2.1(a). Se puede observar que ambas distribuciones tienen modas muy similares, pero la aproximación normal no es tan precisa para valores de θ cercanos a cero. En particular, cuando $\theta = 0$ la aproximación normal no llega a cero.

Estas fallas en la aproximación se deben a que la densidad final de θ tiene un ligero sesgo hacia la izquierda y además tiene un rango de valores acotado. Para tratar de que la aproximación normal se parezca más a la densidad final original, es necesario realizar una transformación sobre θ que elimine el sesgo y el rango acotado. La transformación que se propone es la logística, es decir,

$$\eta = \log\left(\frac{\theta}{1-\theta}\right).$$

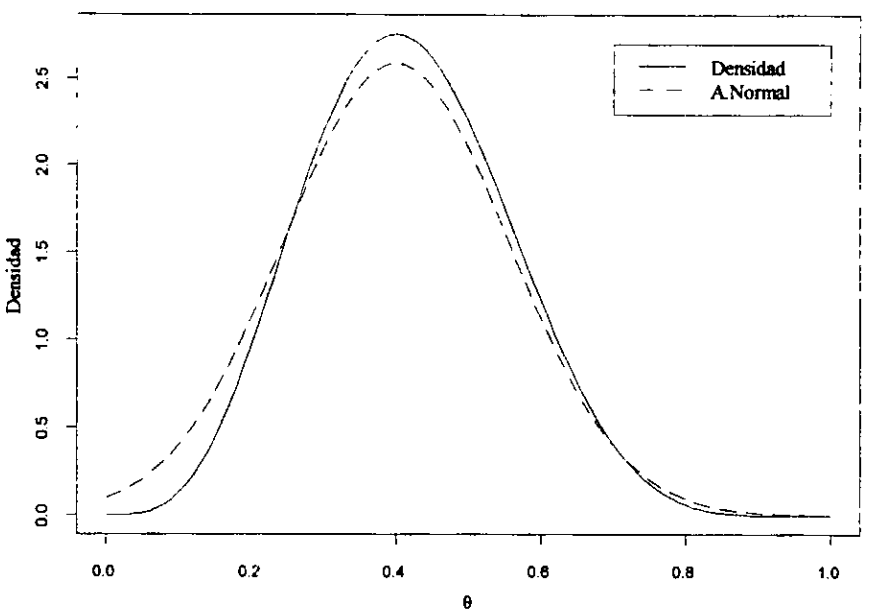
Realizando la transformación anterior se obtiene que la densidad de η es

$$p(\eta|\mathbf{x}) = \frac{1}{B(5,7)} e^{5\eta} (1+e\eta)^{-12} I_{(-\infty,\infty)}(\eta).$$

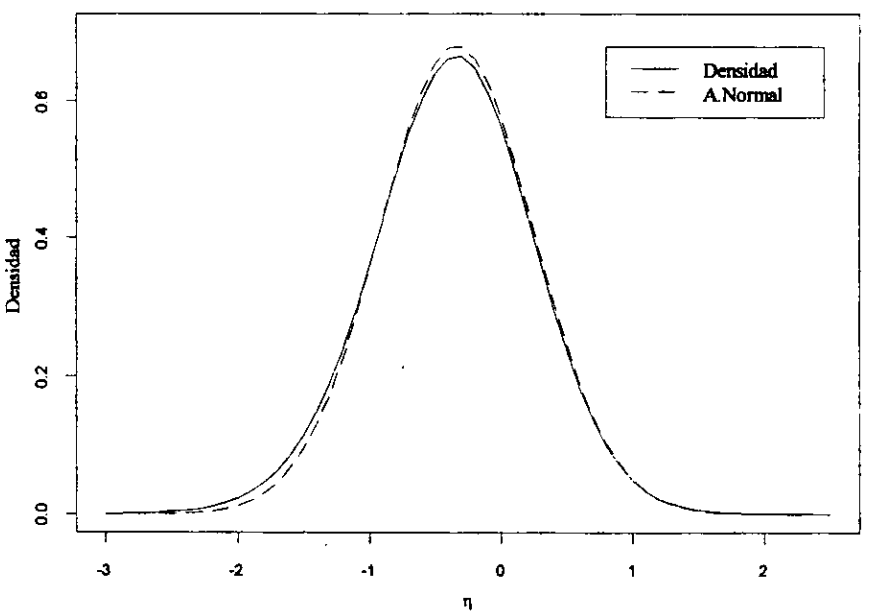
Aplicando la ecuación de la aproximación normal (2.1) al parámetro transformado η, se tiene que

$$p(\eta|\mathbf{x}) \approx N(\eta|-0.3364, 0.3428).$$

El comportamiento de la transformación y su correspondiente aproximación normal se pueden observar en la Figura 2.1(b). Dado que el sesgo se corrigió casi del todo y el rango abarca toda la recta real, la aproximación normal a este parámetro transformado mejora considerablemente con respecto a la aproximación del parámetro sin transformar. ■



(a)



(b)

Figura 2.1: Densidad final y aproximación normal. (a) θ y (b) η.

2.2 Aproximación de Laplace.

2.2.1 Ideas básicas.

La aproximación de Laplace es un método asintótico para calcular integrales y ha tenido un uso muy amplio en las matemáticas aplicadas.

Para iniciar la exposición de la aproximación de Laplace, recordemos que en el contexto Bayesiano el problema de encontrar resúmenes inferenciales de la distribución final se reduce a evaluar integrales de la forma

$$E\left[g\left(\boldsymbol{\theta}\right)|\mathbf{x}\right]=\int g\left(\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\,|\mathbf{x}\right)\partial\boldsymbol{\theta},$$

donde $p\left(\boldsymbol{\theta}\,|\mathbf{x}\right)$ es la densidad final de $\boldsymbol{\theta}$ y $g\left(\boldsymbol{\theta}\right)$ es una función real de interés. Sabemos que $p\left(\boldsymbol{\theta}\,|\mathbf{x}\right)$ se puede escribir como

$$p\left(\boldsymbol{\theta}\,|\mathbf{x}\right)=\frac{p\left(\mathbf{x}\,|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)}{\int p\left(\mathbf{x}\,|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)\partial\boldsymbol{\theta}},$$

donde, $p\left(\mathbf{x}\,|\boldsymbol{\theta}\right)$ es la función de verosimilitud y $p\left(\boldsymbol{\theta}\right)$ es la densidad inicial. Entonces, se puede ver que $E\left[g\left(\boldsymbol{\theta}\right)|\mathbf{x}\right]$ puede expresarse como el cociente de dos integrales de la forma

$$E\left[g\left(\boldsymbol{\theta}\right)|\mathbf{x}\right]=\frac{\int g\left(\boldsymbol{\theta}\right)p\left(\mathbf{x}\,|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)\partial\boldsymbol{\theta}}{\int p\left(\mathbf{x}\,|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)\partial\boldsymbol{\theta}}. \tag{2.2}$$

Los integrandos del cociente (2.2) se pueden reexpresar, de manera que la esperanza tome la forma

$$E\left[g\left(\boldsymbol{\theta}\right)|\mathbf{x}\right]=\frac{\int b_N\left(\boldsymbol{\theta}\right)\exp\left\{-nh_N\left(\boldsymbol{\theta}\right)\right\}\partial\boldsymbol{\theta}}{\int b_D\left(\boldsymbol{\theta}\right)\exp\left\{-nh_D\left(\boldsymbol{\theta}\right)\right\}\partial\boldsymbol{\theta}} \tag{2.3}$$

donde,

$$\begin{aligned} b_N(\boldsymbol{\theta}) \exp\{-n h_N(\boldsymbol{\theta})\} &= g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \quad \text{y} \\ b_D(\boldsymbol{\theta}) \exp\{-n h_D(\boldsymbol{\theta})\} &= p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}). \end{aligned}$$

Supongamos que $b_N(\cdot)$ y $b_D(\cdot)$ son funciones que no dependen de n, y que $h_N(\cdot)$ y $h_D(\cdot)$ son funciones de orden constante en n, cuando $n \to \infty$.

Cuando $h_N(\boldsymbol{\theta}) = h_D(\boldsymbol{\theta})$ se dice que (2.3) está en *forma estándar* (Lindley, 1980), mientras que si $b_N(\boldsymbol{\theta}) = b_D(\boldsymbol{\theta})$ [en cuyo caso $h_N(\boldsymbol{\theta}) \neq h_D(\boldsymbol{\theta})$] se dice que (2.3) está en *forma exponencial* (Tierney y Kadane, 1986).

La elección de las funciones $b(\cdot)$ y $h(\cdot)$ en la forma estándar depende del problema específico, pero la elección más común es

$$\begin{aligned} h_N(\boldsymbol{\theta}) &= h_D(\boldsymbol{\theta}) = -\left(\frac{1}{n}\right) \log\{p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\}, \\ b_N(\boldsymbol{\theta}) &= g(\boldsymbol{\theta}) \quad \text{y} \\ b_D(\boldsymbol{\theta}) &= 1 \end{aligned} \tag{2.4}$$

En cambio, la elección más común de las funciones $b(\cdot)$ y $h(\cdot)$ en la forma exponencial es seleccionarlas de tal manera que

$$\begin{aligned} b_N(\boldsymbol{\theta}) &= b_D(\boldsymbol{\theta}) = 1, \\ h_N(\boldsymbol{\theta}) &= -\left(\frac{1}{n}\right) \log\{g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\} \quad \text{y} \\ h_D(\boldsymbol{\theta}) &= -\left(\frac{1}{n}\right) \log\{p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\} \end{aligned} \tag{2.5}$$

Teorema 2.2.1.

Supongamos que se desea calcular una integral de la forma

$$I = \int b(\boldsymbol{\theta}) \exp\{-nh(\boldsymbol{\theta})\} \partial\boldsymbol{\theta}$$

donde, $b : \Re^k \to \Re$ no depende de n, y $h : \Re^k \to \Re$ es una función doblemente diferenciable de orden constante en n, cuando $n \to \infty$. Supongamos además que $h(\cdot)$ tiene un mínimo en $\widehat{\boldsymbol{\theta}}$. La aproximación de Laplace básica a la integral I es

$$I = \left(\frac{2\pi}{n}\right)^{\frac{k}{2}} \left|\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)\right|^{\frac{1}{2}} b\left(\widehat{\boldsymbol{\theta}}\right) \exp\left\{-nh\left(\widehat{\boldsymbol{\theta}}\right)\right\} \left[1 + O\left(n^{-1}\right)\right] \qquad (2.6)$$

donde

$$\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\frac{\partial^2 h\left(\widehat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right]^{-1}.$$

La demostración de este resultado se puede encontrar en Tierney, Kass y Kadane (1989).

Una justificación heurística se puede encontrar fácilmente haciendo la expansión en serie de Taylor tanto de $h(\boldsymbol{\theta})$, como de $b(\boldsymbol{\theta})$ alrededor de $\widehat{\boldsymbol{\theta}}$. Para facilitar la exposición, supongamos que $\boldsymbol{\theta}$ es de dimensión uno.

Haciendo la expansión de $nh(\theta)$ alrededor de $\widehat{\theta}$ y despreciando los términos de orden mayor a dos, tenemos

$$nh(\theta) \approx nh\left(\widehat{\theta}\right) + nh'\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right) + \frac{n}{2\Sigma\left(\widehat{\theta}\right)}\left(\theta - \widehat{\theta}\right)^2$$

donde $h'\left(\widehat{\theta}\right)$ es la primera derivada de $h(\theta)$ evaluada en $\widehat{\theta}$, y

$$\Sigma\left(\widehat{\theta}\right) = \left[\frac{\partial^2 h\left(\widehat{\theta}\right)}{\partial\theta^2}\right]^{-1} \quad \text{y} \quad h'(\theta) = \frac{\partial h(\theta)}{\partial\theta}.$$

Ahora, dado que $h'\left(\widehat{\theta}\right) = 0$, obtenemos

$$\exp\{-nh(\theta)\} \approx \exp\left\{-nh\left(\widehat{\theta}\right)\right\} \exp\left\{-\frac{n}{2\Sigma\left(\widehat{\theta}\right)}\left(\theta-\widehat{\theta}\right)^2\right\}.$$

De manera similar, desarrollando $b(\theta)$ alrededor de $\widehat{\theta}$ tenemos

$$b(\theta) \approx b\left(\widehat{\theta}\right) + b'\left(\widehat{\theta}\right)\left(\theta-\widehat{\theta}\right) + \frac{b''\left(\widehat{\theta}\right)}{2}\left(\theta-\widehat{\theta}\right)^2,$$

donde $b'\left(\widehat{\theta}\right)$ y $b''\left(\widehat{\theta}\right)$ son la primera y segunda derivada de $b(\theta)$ evaluadas en $\widehat{\theta}$, respectivamente.

Entonces, el integrando de I puede escribirse como

$$\begin{aligned} b(\theta)\exp\{-nh(\theta)\} \;\approx\; & \left\{b\left(\widehat{\theta}\right) + b'\left(\widehat{\theta}\right)\left(\theta-\widehat{\theta}\right) + \frac{b''\left(\widehat{\theta}\right)}{2}\left(\theta-\widehat{\theta}\right)^2\right\} \\ & \times \exp\left\{-nh\left(\widehat{\theta}\right)\right\} \exp\left\{-\frac{n}{2\Sigma\left(\widehat{\theta}\right)}\left(\theta-\widehat{\theta}\right)^2\right\}. \end{aligned}$$

Se puede observar que el último factor de esta expresión es proporcional a una densidad $N\left(\theta\left|\widehat{\theta}, \frac{\Sigma\left(\widehat{\theta}\right)}{n}\right.\right)$. Al integrar ambos lados, se obtiene que el segundo término de la primera llave desaparece debido a que $E\left(\theta-\widehat{\theta}\right) = 0$ (tomando el valor esperado con respecto a la densidad normal antes mencionada).

Finalmente,

$$I \approx \left(\frac{2\pi}{n}\right)^{\frac{1}{2}} \left[\Sigma\left(\widehat{\theta}\right)\right]^{\frac{1}{2}} b\left(\widehat{\theta}\right) \exp\left\{-nh\left(\widehat{\theta}\right)\right\} \left\{1 + \frac{b''\left(\widehat{\theta}\right)}{2}\mathrm{Var}(\theta)\right\}$$

donde

$$\text{Var}(\theta) = \frac{\Sigma\left(\widehat{\theta}\right)}{n} = O\left(n^{-1}\right).$$

■

El resultado del Teorema 2.2.1 proporciona una regla general de integración basada en una aproximación normal. La aproximación de Laplace es particularmente útil en el contexto de la estadística Bayesiana debido a que proporciona una forma fácil de aproximar resúmenes inferenciales de la distribución final del parámetro, simplemente con el cálculo de primeras y segundas derivadas.

2.2.2 Valores esperados.

Retomando el objetivo inicial de calcular valores esperados de funciones del parámetro con respecto a la densidad final, es necesario aplicar el resultado del Teorema 2.2.1 tanto al numerador como al denominador del cociente (2.3).

Dependiendo de la factorización que se haga tanto del numerador como del denominador del cociente (2.2) se tendrán aproximaciones distintas. En particular, para las factorizaciones *estándar* y *exponencial* se tienen resultados importantes. Estos dos resultados se enuncian en los siguientes dos teoremas.

Teorema 2.2.2. Forma estándar.

Sea $g(\boldsymbol{\theta})$ una función suave definida sobre el espacio parametral Θ de dimensión k. El valor esperado de $g(\boldsymbol{\theta})$ con respecto a la densidad final de $\boldsymbol{\theta}$ se puede escribir como

$$E\left[g\left(\boldsymbol{\theta}\right)|x\right]=\frac{\int g\left(\boldsymbol{\theta}\right)p\left(\mathbf{x}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)\partial\boldsymbol{\theta}}{\int p\left(\mathbf{x}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)\partial\boldsymbol{\theta}}=\frac{\int b_N\left(\boldsymbol{\theta}\right)\exp\left\{-nh_N\left(\boldsymbol{\theta}\right)\right\}\partial\boldsymbol{\theta}}{\int b_D\left(\boldsymbol{\theta}\right)\exp\left\{-nh_D\left(\boldsymbol{\theta}\right)\right\}\partial\boldsymbol{\theta}}$$

donde las funciones $b_J\left(\cdot\right)$ y $h_J\left(\cdot\right)$, $J=N,D$ se eligen de la forma estándar (2.4), es decir, $h_N\left(\boldsymbol{\theta}\right)=h_D\left(\boldsymbol{\theta}\right)=h\left(\boldsymbol{\theta}\right)$ y $b_N\left(\boldsymbol{\theta}\right)=g\left(\boldsymbol{\theta}\right)b_D\left(\boldsymbol{\theta}\right)$. Entonces, aplicando (2.6) tanto al numerador como al denominador se obtiene la aproximación de Laplace en forma estándar

$$E\left[g\left(\boldsymbol{\theta}\right)|\mathbf{x}\right]=g\left(\widehat{\boldsymbol{\theta}}\right)\left\{1+O\left(n^{-1}\right)\right\} \tag{2.7}$$

donde $\widehat{\boldsymbol{\theta}}$ es el mínimo de $h\left(\boldsymbol{\theta}\right)$.

La demostración de este resultado se puede encontrar en Tierney *et al.* (1989).

La aproximación de Laplace en forma estándar para el cálculo de valores esperados es una aproximación de primer orden y además se puede observar que coincide con la aproximación modal al valor esperado final. Para aplicar la aproximación de Laplace en forma estándar únicamente es necesario el cálculo de primeras y segundas derivadas, con la ventaja de que existen algoritmos numéricos eficientes para aproximarlas en el caso de que no se puedan encontrar analíticamente.

Para obtener una aproximación más precisa al valor esperado final, Tierney y Kadane (1986) propusieron una ligera modificación en la factorización de los integrandos. Este cambio se enuncia en el siguiente teorema:

Teorema 2.2.3. Forma exponencial.

Sea $g(\boldsymbol{\theta})$ una función suave y positiva definida sobre el espacio parametral Θ de dimensión k. El valor esperado de $g(\boldsymbol{\theta})$ con respecto a la densidad final de $\boldsymbol{\theta}$ se puede escribir como

$$E[g(\boldsymbol{\theta})|x] = \frac{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \partial\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \partial\boldsymbol{\theta}} = \frac{\int b_N(\boldsymbol{\theta}) \exp\{-nh_N(\boldsymbol{\theta})\} \partial\boldsymbol{\theta}}{\int b_D(\boldsymbol{\theta}) \exp\{-nh_D(\boldsymbol{\theta})\} \partial\boldsymbol{\theta}}$$

donde las funciones $b_J(\cdot)$ y $h_J(\cdot)$, $J = N, D$ son elegidas de la forma (2.5), es decir, $h_N(\boldsymbol{\theta}) = h_D(\boldsymbol{\theta}) - \left(\frac{1}{n}\right) \log g(\boldsymbol{\theta})$ y $b_N(\boldsymbol{\theta}) = b_D(\boldsymbol{\theta}) = 1$. Entonces, aplicando (2.6) tanto al numerador como al denominador se obtiene la aproximación de Laplace en forma exponencial

$$E[g(\boldsymbol{\theta})|x] = \frac{\left|\boldsymbol{\Sigma}_N\left(\widehat{\boldsymbol{\theta}}_N\right)\right|^{\frac{1}{2}} b_N\left(\widehat{\boldsymbol{\theta}}_N\right) \exp\left\{-nh_N\left(\widehat{\boldsymbol{\theta}}_N\right)\right\}}{\left|\boldsymbol{\Sigma}_D\left(\widehat{\boldsymbol{\theta}}_D\right)\right|^{\frac{1}{2}} b_D\left(\widehat{\boldsymbol{\theta}}_D\right) \exp\left\{-nh_D\left(\widehat{\boldsymbol{\theta}}_D\right)\right\}} \left\{1 + O\left(n^{-2}\right)\right\}$$

donde $\widehat{\boldsymbol{\theta}}_J$ es el mínimo de $h_J(\boldsymbol{\theta})$, $J = N, D$.

Expresando este resultado en términos de la densidad inicial y la verosimilitud tenemos que

$$E[g(\boldsymbol{\theta})|x] = \frac{\left|\boldsymbol{\Sigma}_N\left(\widehat{\boldsymbol{\theta}}_N\right)\right|^{\frac{1}{2}} g\left(\widehat{\boldsymbol{\theta}}_N\right) p\left(\mathbf{x}\middle|\widehat{\boldsymbol{\theta}}_N\right) p\left(\widehat{\boldsymbol{\theta}}_N\right)}{\left|\boldsymbol{\Sigma}_D\left(\widehat{\boldsymbol{\theta}}_D\right)\right|^{\frac{1}{2}} p\left(\mathbf{x}\middle|\widehat{\boldsymbol{\theta}}_D\right) p\left(\widehat{\boldsymbol{\theta}}_D\right)} \left\{1 + O\left(n^{-2}\right)\right\} \tag{2.8}$$

La demostración de este resultado se puede encontrar en Tierney *et al.* (1989).

De los Teoremas 2.2.2 y 2.2.3 se puede observar que el orden del error de la aproximación de Laplace cambia al factorizar de una manera diferente el integrando del numerador y del denominador del cociente (2.2), obteniéndose una aproximación

de segundo orden al realizar la factorización en forma exponencial. A diferencia de la aproximación de Laplace en forma estándar, en la forma exponencial es necesario calcular primeras y segundas derivadas tanto del numerador como del denominador, por lo que generalmente se requieren más cálculos en este último caso.

La diferencia en el orden del error de la aproximación de Laplace en forma exponencial (2.8) se debe a que los primeros términos de los errores, al aplicar (2.6) al numerador y al denominador, son ambos de orden $O(n^{-1})$ y además idénticos, por lo que al tomar el cociente se cancelan, obteniéndose la aproximación de orden $O(n^{-2})$. Se pueden verificar los detalles de este argumento en Tierney y Kadane (1986).

La aproximación de Laplace en forma exponencial tiene la desventaja de que para ser aplicada, la función $g(\cdot)$ debe de ser una función positiva. Sin embargo, Tierney *et* al. (1989) obtuvieron una aproximación de segundo orden sin ninguna restricción sobre la función $g(\cdot)$, basada en la aproximación de Laplace a la función generadora de momentos. Otra manera de aplicar la aproximación de Laplace en forma exponencial en el caso de que la función $g(\cdot)$ no sea positiva es sumando una constante suficientemente grande para poder aplicar la aproximación y al resultado final restarle la constante, i.e,

$$\widehat{E}[g(\boldsymbol{\theta})|\mathbf{x}] = \widehat{E}[g(\boldsymbol{\theta}) + a|\mathbf{x}] - a$$

Ejemplo 2.2.1. (Modificación del ejemplo presentado en Press, 1989).

Sean $X_1, \ldots, X_n$ un conjunto de variables aleatorias independientes de la densidad

$$p(x|\theta) = e^{-\theta}\frac{\theta^x}{x!}I_{\{0,1,\ldots,\}}, \qquad \theta > 0.$$

Entonces, la función de densidad conjunta de los datos es

$$p(\mathbf{x}|\theta) = e^{-n\theta}\frac{\theta^{\sum_1^n x_i}}{\prod_1^n x_i!}.$$

Supongamos que se propone una densidad inicial conjugada para θ,

$$p(\theta) = Gamma(\theta|\alpha,\beta).$$

Por lo tanto, la densidad final para θ resulta ser

$$p(\theta|\mathbf{x}) \propto \theta^{\alpha+\sum x_i-1}e^{-(\beta+n)\theta}.$$

Si $\alpha_1 = \alpha + \sum x_i$ y $\beta_1 = \beta + n$, entonces $\theta|\mathbf{x} \sim Gamma(\alpha_1,\beta_1)$.

Por las características de la densidad Gamma, sabemos que el valor esperado final de θ está dado por

$$E(\theta|\mathbf{x}) = \frac{\alpha_1}{\beta_1}.$$

- *Aproximación de Laplace (forma estándar).*

  Factorizando el integrando de la forma (2.4) para encontrar el valor esperado final de θ, tenemos que

$$\begin{aligned} h_N(\theta) &= h_D(\theta) = -\frac{1}{n}\log\left\{\theta^{\alpha_1-1}e^{-\beta_1\theta}\right\} \\ b_N(\theta) &= \theta \\ b_D(\theta) &= 1, \end{aligned}$$

  aplicando la aproximación (2.7), observamos que la aproximación de Lapace en forma estándar para este caso es la moda de la densidad final de θ. Por lo tanto,

$$\widetilde{E}(\theta|\mathbf{x}) = \frac{(\alpha_1-1)}{\beta_1}.$$

- *Aproximación de Laplace (forma exponencial).*

  Ahora, factorizando el integrando de la forma (2.5) para encontrar el valor esperado final de θ, tenemos que

$$\begin{aligned} h_N(\theta) &= -\frac{1}{n}\log\left\{\theta^{\alpha_1}e^{-\beta_1\theta}\right\} \\ h_D(\theta) &= -\frac{1}{n}\log\left\{\theta^{\alpha_1-1}e^{-\beta_1\theta}\right\} \\ b_N(\theta) &= b_D(\theta) = 1. \end{aligned}$$

  Es inmediato comprobar que $\widehat{\theta}_N = \frac{\alpha_1}{\beta_1}$, $\widehat{\Sigma}_N = \frac{n\alpha_1}{\beta_1^2}$, $\widehat{\theta}_D = \frac{\alpha_1-1}{\beta_1}$, $\widehat{\Sigma}_N = \frac{n(\alpha_1-1)}{\beta_1^2}$. Aplicando la aproximación (2.8), obtenemos que

$$\widehat{E}(\theta\,|\mathbf{x}) = \frac{\alpha_1}{\beta_1}\left(\frac{\alpha_1}{\alpha_1-1}\right)^{\alpha_1-\frac{1}{2}}e^{-1}$$

  es la aproximación de Laplace al valor esperado final de θ en forma exponencial.

Notemos que tanto en la aproximación de Laplace en la forma estándar como en la forma exponencial, se debe cumplir que $\alpha_1 > 1$ para que sean aplicables.

Para evaluar la efectividad de ambas aproximaciones en este ejemplo, se presenta la siguiente tabla que muestra los errores relativos, como función de α_1.

α_1	2	3	4	6	8	10
Exponencial	4.05%	1.38%	0.69%	0.28%	0.15%	0.09%
Estándar	100%	50%	33%	20%	14.2%	11.1%

De la Tabla anterior se observa que conforme α_1 crece, los errores relativos de ambas aproximaciones decrecen, notándose además que los errores relativos para la aproximación de Laplace en forma exponencial son mucho menores que los correspondientes a la aproximación estándar.

2.2.3 Densidades predictivas.

El método de Laplace también se puede utilizar como un método de aproximación a densidades predictivas, debido a que la densidad predictiva se puede expresar como un valor esperado con respecto a la densidad final.

Sea x_F una observación futura de la variable aleatoria X, entonces la distribución predictiva para X_F está dada por

$$p(x_F|\mathbf{x}) = \int p(x_F|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x})\, \partial\boldsymbol{\theta}$$

o equivalentemente, como un valor esperado se puede expresar como

$$p(x_F|\mathbf{x}) = E_{\boldsymbol{\theta}|\mathbf{x}}\left[p(x_F|\boldsymbol{\theta})\right].$$

La aproximación de Laplace en la forma estándar (2.7) a la densidad predictiva $\widetilde{p}(x_F|\mathbf{x}) = p\left(x_F\left|\widehat{\boldsymbol{\theta}}\right.\right)$ es una aproximación de primer orden, donde $\widehat{\boldsymbol{\theta}}$ es la moda de la distribución final.

Como la densidad de $X_F|\boldsymbol{\theta}$ es una función no negativa, se puede aplicar la aproximación de Laplace en la forma exponencial (2.8), obteniéndose que $\widehat{p}(x_F|\mathbf{x}) = \widehat{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[p(x_F|\boldsymbol{\theta})\right]$ es una aproximación de segundo orden a la densidad predictiva de X_F.

Más generalmente, la aproximación de Laplace permite aproximar cualquier densidad marginal de interés. Esto se discute en la siguiente sección.

2.2.4 Densidades marginales.

Sea $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, con $\boldsymbol{\theta}_1 \in \Re^{k_1}$ y $\boldsymbol{\theta}_2 \in \Re^{k-k_1}$. Supongamos que la distribución de $\boldsymbol{\theta}$ se puede escribir como

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \propto b(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \exp\{-nh(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\},$$

donde $b(\cdot)$ no depende de n y $h(\cdot)$ es una función de orden constante en n, cuando $n \to \infty$, y nos interesa calcular la densidad marginal de $\boldsymbol{\theta}_1$, es decir,

$$p(\boldsymbol{\theta}_1) \propto \int b(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \exp\{-nh(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\} \partial\boldsymbol{\theta}_2. \tag{2.9}$$

Para cada valor de $\boldsymbol{\theta}_1$, se definen

$$b_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = b(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \quad \text{y} \quad h_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

de manera que $b_{\boldsymbol{\theta}_1}(\cdot)$ y $h_{\boldsymbol{\theta}_1}(\cdot)$ son respectivamente $b(\cdot)$ y $h(\cdot)$ vistas únicamente como funciones de $\boldsymbol{\theta}_2$. Finalmente, supongamos que $h_{\boldsymbol{\theta}_1}(\cdot)$ tiene un mínimo en $\widehat{\boldsymbol{\theta}}_2 = \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$.

Aplicando el resultado (2.6) del Teorema 2.2.1 a la ecuación (2.9), se obtiene que la aproximación de Laplace a la densidad marginal de $\boldsymbol{\theta}_1$ se reduce a

$$\widetilde{p}(\boldsymbol{\theta}_1) \propto \left|\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_1)\right|^{\frac{1}{2}} p\left(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\right),$$

donde $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_1) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}\left(\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\right)$, con

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = \left[\frac{\partial^2 h_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2' \partial\boldsymbol{\theta}_2}\right]^{-1}.$$

Dependiendo de la factorización de las funciones $b_{\boldsymbol{\theta}_1}(\cdot)$ y $h_{\boldsymbol{\theta}_1}(\cdot)$, se tendrá la aproximación de Laplace en forma estándar si se utiliza (2.4), y la aproximación de Laplace forma exponencial si se utiliza (2.5).

Después de obtener la constante de normalización, Kass *et al.* (1988) muestran que el error relativo para esta aproximación, al usar una factorización exponencial, es de orden $O\left(n^{-\frac{3}{2}}\right)$ en vecindades alrededor de la moda que se compactan a una tasa de $n^{-\frac{1}{2}}$. Este resultado puede ser comparado con la aproximación normal a densidades marginales, la cual tiene un error de orden $O\left(n^{-\frac{1}{2}}\right)$ en las mismas vecindades.

La constante de normalización de la densidad conjunta de $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ puede ser aproximada por (2.6), con un error de orden $O\left(n^{-1}\right)$.

Ejemplo 2.2.2.

Sea

$$p(\mu, \tau) = N\left(\mu \middle| \mu_0, \frac{1}{\tau}\right) Ga\left(\tau \middle| \frac{\alpha}{2}, \frac{\beta}{2}\right)$$

la densidad Normal-Gamma, es decir,

$$p(\mu, \tau) \propto \tau^{\frac{\alpha+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[\beta + (\mu - \mu_0)^2\right]\right\}.$$

La densidad marginal de μ está dada por

$$p(\mu) \propto \left[1 + \frac{1}{\beta}(\mu - \mu_0)^2\right]^{-\frac{\alpha+1}{2}},$$

por lo tanto,

$$p(\mu) = St\left(\mu \middle| \mu_0, \frac{\beta}{\alpha}, \alpha\right).$$

- *Aproximación de Laplace (forma estándar).*

  Sean

$$b\mu(\tau) = \tau^{\frac{1}{2}} \exp\left\{-\frac{\tau}{2}(\mu - \mu_0)^2\right\} \quad \text{y} \quad h\mu(\tau) = \frac{1}{2n}\left[\beta\tau - (\alpha - 2)\log\tau\right].$$

Entonces,

$$\widehat{\tau}(\mu) = \widehat{\tau} = \frac{\alpha - 2}{\beta} \quad \text{y} \quad \Sigma\mu(\widehat{\tau}) = \frac{2n\widehat{\tau}^2}{\alpha - 2}.$$

Al observar que estas dos últimas expresiones no dependen de μ, se obtiene que la aproximación de Laplace estándar a la densidad marginal de μ es

$$\widetilde{p}(\mu) \propto \exp\left\{-\frac{\widehat{\tau}}{2}(\mu - \mu_0)^2\right\}$$

o equivalentemente,

$$\widetilde{p}(\mu) = N\left(\mu \middle| \mu_0, \frac{\beta}{\alpha - 2}\right).$$

- *Aproximación de Laplace (forma exponencial).*

Sean

$$b\mu(\tau) = 1 \quad \text{y} \quad h\mu(\tau) = \frac{1}{2n}\left\{\tau\left[\beta + (\mu - \mu_0)^2 - (\alpha - 1)\log\tau\right]\right\}.$$

En este caso,

$$\begin{aligned} \widehat{\tau}(\mu) &= (\alpha - 1)\left[\beta + (\mu - \mu_0)^2\right]^{-1} \quad \text{y} \\ \Sigma\mu(\widehat{\tau}(\mu)) &= 2n(\alpha - 1)\left[\beta + (\mu - \mu_0)^2\right]^{-2}. \end{aligned}$$

Por lo tanto, la aproximación de Laplace exponencial a la densidad marginal de μ es

$$\widehat{p}(\mu) \propto \left[1 + \frac{1}{\beta}(\mu - \mu_0)^2\right]^{\frac{-(\alpha+1)}{2}}.$$

En otras palabras,

$$\widehat{p}(\mu) = St\left(\mu \middle| \mu_0, \frac{\beta}{\alpha}, \alpha\right) = p(\mu).$$

La aproximación obtenida en la forma estándar es la mejor aproximación normal a la verdadera densidad marginal. Sin embargo, en el caso de la aproximación en forma exponencial se reprodujo la verdadera densidad marginal. ■

Vale la pena hacer notar que en algunos casos como Normal-Gamma, Dirichlet y Logística bivariada, las aproximaciones a la densidad marginal por el método de Laplace resultan ser exactas al normalizarlas para que integren uno (Efstathiou *et al.*, 1998).

Una ventaja considerable del método de Laplace para obtener densidades marginales es que dan lugar a aproximaciones analíticas. Por ejemplo, si cambian los hiperparámetros de la distribución inicial o se aumenta el tamaño de muestra, se pueden encontrar rápidamente aproximaciones a las densidades marginales correspondientes si se cuenta con una aproximación analítica. En cambio, en el caso de las aproximaciones numéricas no siempre se tendrá esta ventaja.

Tierney *et al.* (1989 b) trabajaron en las aproximaciones a densidades marginales para generalizar la aproximación de Laplace al caso de que se quiera aproximar la distribución marginal de una función general del parámetro.

2.2.5 Discusión.

Existen varias ventajas del método de Laplace. Es un procedimiento computacionalmente rápido, sustituye la integración numérica con diferenciación numérica (en el caso de que no se puedan obtener analíticamente las derivadas requeridas), por lo que es más fácil de obtener. Además, es un algoritmo determinístico en el sentido de que

no depende de números aleatorios como los métodos de Monte Carlo.

Por otra parte, el método de Laplace tiene también varias limitaciones. Para que la aproximación sea válida o aplicable, la función $-nh(\cdot)$ debe ser unimodal o al menos dominada por una sola moda. La precisión de la aproximación depende de la parametrización utilizada (por ejemplo θ vs. $\log(\theta)$), y la parametrización adecuada puede ser difícil de obtener. Al igual que en la aproximación normal, en aplicaciones específicas generalmente no es posible determinar si la aproximación de Laplace es adecuada para un tamaño de muestra dado. Finalmente, en el caso en que la dimensión del parámetro sea grande, el método de Laplace raramente tendrá mucha precisión, además de que el trabajo computacional y/o analítico, necesario para el cálculo de la moda y las matrices de dispersión se complica.

Capítulo 3

Aproximaciones analíticas (Caso multimodal).

3.1 Aproximación con mezclas de normales.

Como se mencionó en el Capítulo 1, en ciertas aplicaciones Bayesianas surgen, como resultado del análisis, distribuciones finales multimodales. En estos casos, para poder obtener resúmenes inferenciales de la distribución final generalmente se tiene que recurrir a aproximaciones, ya que en algunos casos ni siquiera la constante de normalización es fácil de obtener.

Al igual que en el caso unimodal, se tratará de aproximar a la distribución final mediante una distribución fácil de manejar. Una familia fácil de manipular que permite modelar casi cualquier comportamiento es la familia de las mezclas. Las mezclas han sido utilizadas para reproducir el comportamiento de ciertas distribuciones (Diaconis e Ylvisaker, 1985) y se han obtenido muy buenos resultados debido a su flexibilidad.

En particular, las mezclas de densidades normales han sido ampliamente utilizadas en el contexto Bayesiano para aproximar distribuciones finales (West, 1993).

Existen otras familias de distribuciones que representan comportamientos multimodales (Cobb *et al.*, 1983), pero no tienen la ventaja de ser flexibles, como las mezlcas, para modelar cualquier comportamiento.

3.1.1 Aproximación básica.

Una generalización de la aproximación normal asintótica, basada en una mezcla de densidades normales, se encuentra en O'Hagan (1994). Esta aproximación consiste en lo siguiente:

Sea $p(\boldsymbol{\theta}|\mathbf{x}) \propto p_x(\boldsymbol{\theta})$ una densidad multimodal, y sean $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_d$ las modas, con $\boldsymbol{\Sigma}_1\left(\widehat{\boldsymbol{\theta}}_1\right), \ldots, \boldsymbol{\Sigma}_d\left(\widehat{\boldsymbol{\theta}}_d\right)$ las correspondientes matrices de dispersión dadas por

$$\boldsymbol{\Sigma}_i\left(\widehat{\boldsymbol{\theta}}_i\right) = \widehat{\boldsymbol{\Sigma}}_i = -\left[\frac{\partial^2 \log p_x\left(\widehat{\boldsymbol{\theta}}_i\right)}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]^{-1}.$$

La idea es entonces aproximar $p_x(\boldsymbol{\theta})$ a través de una mezcla de densidades normales,

$$p_x(\boldsymbol{\theta}) \approx \sum_{i=1}^{d} p_x\left(\widehat{\boldsymbol{\theta}}_i\right) \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)' \widehat{\boldsymbol{\Sigma}}_i^{-1}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)\right\},$$

donde los parámetros de cada una de las componentes normales se calculan a partir de la moda y la matriz de dispersión correspondiente como en el caso unimodal.

Por lo tanto,

$$p(\boldsymbol{\theta}|\mathbf{x}) \approx \sum_{i=1}^{d} w_i N\left(\boldsymbol{\theta}\left|\widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\Sigma}}_i\right.\right), \tag{3.1}$$

donde $w_i = \frac{p_x\left(\widehat{\boldsymbol{\theta}}_i\right)\left|\widehat{\boldsymbol{\Sigma}}_i\right|^{\frac{1}{2}}}{\sum_{j=1}^{d} p_x\left(\widehat{\boldsymbol{\theta}}_j\right)\left|\widehat{\boldsymbol{\Sigma}}_j\right|^{\frac{1}{2}}}$.

Esta aproximación será mejor si las modas están suficientemente separadas, en el sentido de que la distancia $\left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_i\right)' \widehat{\boldsymbol{\Sigma}}_i^{-1} \left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_i\right)$ sea "grande" para toda $i \neq j$ (ver O'Hagan, 1994, Secc. 8.9). En ese caso, características de $p(\boldsymbol{\theta}|\mathbf{x})$ se aproximan por las correspondientes características de la aproximación básica. En particular, el valor esperado final y la varianza final se pueden aproximar por

$$\widehat{E}[\boldsymbol{\theta}|\mathbf{x}] = \sum_{i=1}^{d} w_i \widehat{\boldsymbol{\theta}}_i \quad \text{y}$$

$$\begin{aligned} \widehat{\mathrm{Var}}[\boldsymbol{\theta}|\mathbf{x}] &= \sum_{i=1}^{d} w_i \left(\widehat{\boldsymbol{\Sigma}}_i + \widehat{\boldsymbol{\theta}}_i \widehat{\boldsymbol{\theta}}_i'\right) - \widehat{E}[\boldsymbol{\theta}|\mathbf{x}]\, \widehat{E}[\boldsymbol{\theta}|\mathbf{x}]' \\ &= \sum_{i=1}^{d} w_i \widehat{\boldsymbol{\Sigma}}_i + \sum_{i=1}^{d} w_i \left(\widehat{\boldsymbol{\theta}}_i - \widehat{E}[\boldsymbol{\theta}|\mathbf{x}]\right)\left(\widehat{\boldsymbol{\theta}}_i - \widehat{E}[\boldsymbol{\theta}|\mathbf{x}]\right)'. \end{aligned}$$

Las distribuciones marginales para cada θ_j, $j = 1, \ldots, k$ se pueden aproximar fácilmente mediante

$$\widehat{p}(\theta_j|\mathbf{x}) = \sum_{i=1}^{d} w_i N\left(\theta_j \left| \widehat{\boldsymbol{\theta}}_{i,j}, \widehat{\boldsymbol{\Sigma}}_{i,jj} \right.\right),$$

donde $\widehat{\boldsymbol{\theta}}_{i,j}$ es la j-ésima coordenada del vector $\widehat{\boldsymbol{\theta}}_i$ y $\widehat{\boldsymbol{\Sigma}}_{i,jj}$ es la entrada (j,j)de la matriz $\widehat{\boldsymbol{\Sigma}}_i$.

Esta aproximación básica, a diferencia de la aproximación normal usada comúnmente en el caso unimodal, no tiene ninguna justificación asintótica. En general, no hay garantía de que al aumentar el tamaño de muestra la aproximación sea mejor, ni de que la distribución final converja a una mezcla de densidades normales. Es conveniente aclarar que en el caso de que se cumplan las condiciones del Teorema 2.1.1 la distribución final será unimodal para un tamaño de muestra suficientemente grande y, dado que la aproximación normal asintótica se puede ver como un caso particular

de la aproximación básica, en este caso la aproximación básica será cada vez mejor a medida que aumente el tamaño de muestra y sí habrá convergencia.

Una característica importante de esta aproximación es el hecho de que en el caso de que la función de densidad sea unimodal, la aproximación básica se reduce a la aproximación normal asintótica. Por otro lado, una desventaja de la aproximación básica es que si la función de densidad es una mezcla de densidades normales, entonces la aproximación no reproduce a la verdadera función de densidad. (En contraste, en el caso unimodal la aproximación normal asintótica reproduce a la densidad verdadera si ésta es normal).

Ejemplo 3.1.1.

Considere el Ejemplo 1.2.2, donde para un tamaño de muestra $n = 2$ y para puntos x_1 y x_2 tales que $|x_1 - x_2| > 2$ se tiene que la distribución final de θ es bimodal. Gráficas de la aproximación (3.1) aplicada a cada una de las cuatro distribuciones de la Figura 1.2 se presentan en la Figura 3.1.

En la Figura 3.1 se observa que la aproximación básica no resulta adecuada en ninguno de los cuatro casos considerados. Para las distribuciones (a), (b) y (c) la aproximación básica captura bien la localización de las modas pero la dispersión es menor a la que tiene la distribución verdadera. Para la distribución (d) la dispersión de la aproximación básica es mayor a la de la verdadera distribución y ni siquiera refleja la bimodalidad. Este mal comportamiento de la aproximación básica a las cuatro densidades puede atribuirse a que las modas no están suficientemente separadas.

Para contar con una evaluación numérica de la precisión de la aproximación, y para tener una medida de comparación con aproximaciones posteriores, se calculará

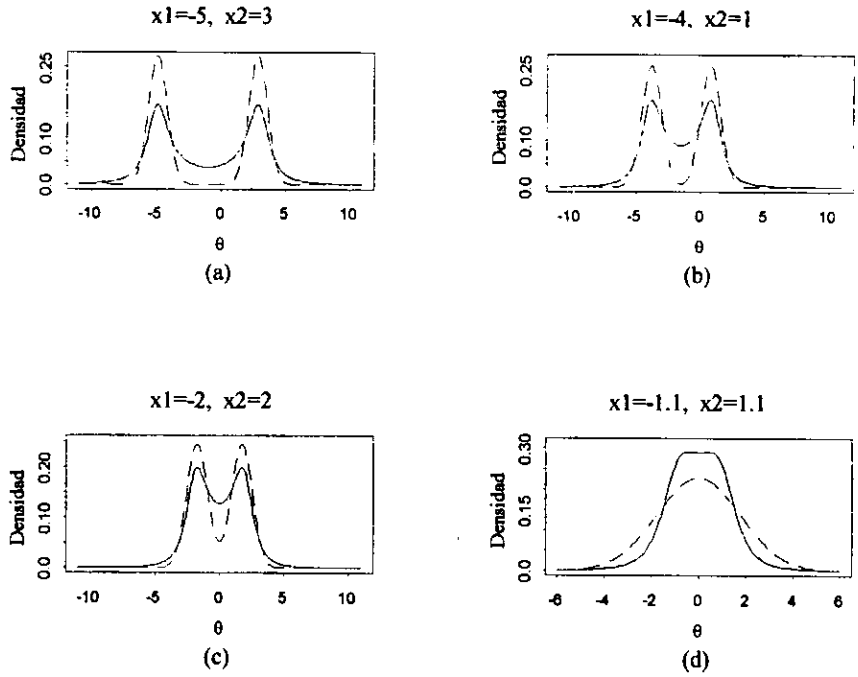


Figura 3.1: Densidad final de θ y Aproximación básica.

la siguiente

$$\sup_{\boldsymbol{\theta}} \left| p\left(\boldsymbol{\theta} \,|\mathbf{x}\right) - \widehat{p}\left(\boldsymbol{\theta} \,|\mathbf{x}\right) \right|. \tag{3.2}$$

La aproximación será mejor si la medida (3.2) es cercana a cero.

La medida (3.2) para cada una de las cuatro densidades de la Figura 3.1 es (a) 0.1035, (b) 0.0816, (c) 0.0759 y (d) 0.0766, notándose que la mejor aproximación es para la densidad (c). ■

En algunas aplicaciones Bayesianas, como en el análisis de modelos lineales, en el análisis de modelos mixtos y en el análisis de modelos dinámicos, frecuentemente surge un tipo particular de distribución llamada *poly-t* (ver Broemeling, 1985). Esta distribución tiene una función de densidad de la forma

$$p\left(\boldsymbol{\theta}\left|\left\{\boldsymbol{\mu}_j, \mathrm{M}_j, \nu_j\right\}, r, k\right.\right) \propto \prod_{j=1}^{r}\left[1+\frac{1}{\nu_j}\left(\boldsymbol{\theta}-\boldsymbol{\mu}_j\right)' \mathrm{M}_j\left(\boldsymbol{\theta}-\boldsymbol{\mu}_j\right)\right]^{-\frac{1}{2}(\nu_j+k)},$$

donde, $\nu_j > 0$, $\sum_{j=1}^{r} \nu_j > k$, $\mathrm{M}_j \geq 0$ (semidefinida positiva), $\sum_{j=1}^{r} \mathrm{M}_j > 0$ (definida positiva). Vale la pena comentar que Dréze (1977) presenta dos tipos de densidades poly-t, en forma de producto y en forma de cociente. La densidad anterior corresponde a la densidad poly-t en forma de producto.

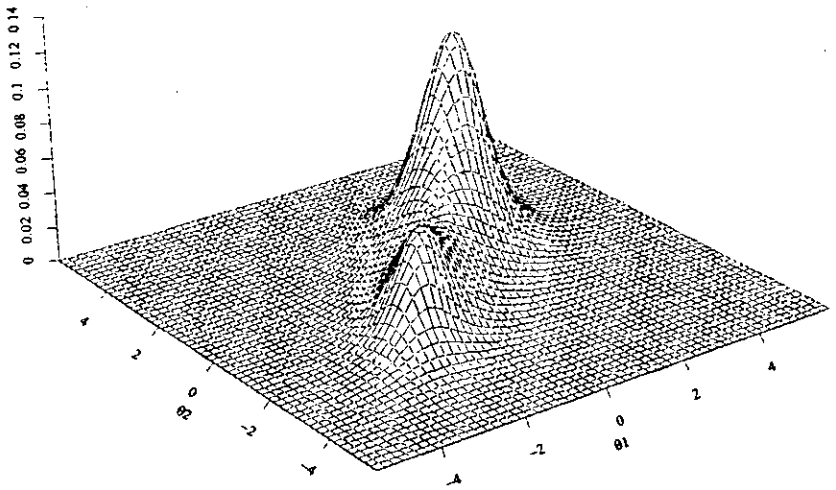
Si $r = 1$, θ tiene una distribución t multivariada, pero cuando $r \geq 2$ se conoce muy poco acerca de la distribución de θ. La constante de normalización y los momentos son desconocidos y es muy difícil trabajar con ella debido a que en general es multimodal y asimétrica.

Para ilustrar el comportamiento de la distribución poly-t y de la aproximación básica a ésta se presenta el siguiente ejemplo.

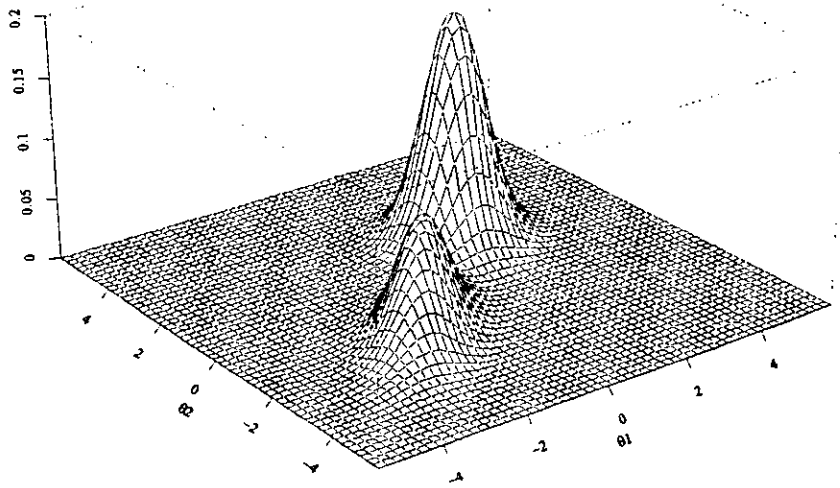
Ejemplo 3.1.2.

Sea $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ un vector aleatorio con distribución poly-t con parámetros $k = 2$, $r = 2$, $\boldsymbol{\mu}_1 = \begin{pmatrix} -2 \\ -2 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\mathrm{M}_1 = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}$, $\mathrm{M}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ y $\nu_1 = \nu_2 = 2$. La gráfica en perspectiva de la distribución conjunta de $\boldsymbol{\theta}$ y las correspondientes curvas de nivel se presentan en las Figuras 3.2(a) y 3.3(a). La constante de normalización se obtuvo numéricamente utilizando la regla trapezoidal.

En las Figuras 3.2(a) y 3.3(a) se puede observar claramente el comportamiento bimodal de la distribución de $\boldsymbol{\theta}$. Aplicando (3.1) a esta distribución se tiene que la



(a)



(b)

Figura 3.2: (a) Densidad Poly-t bivariada y (b) Aproximación básica (perspectiva).

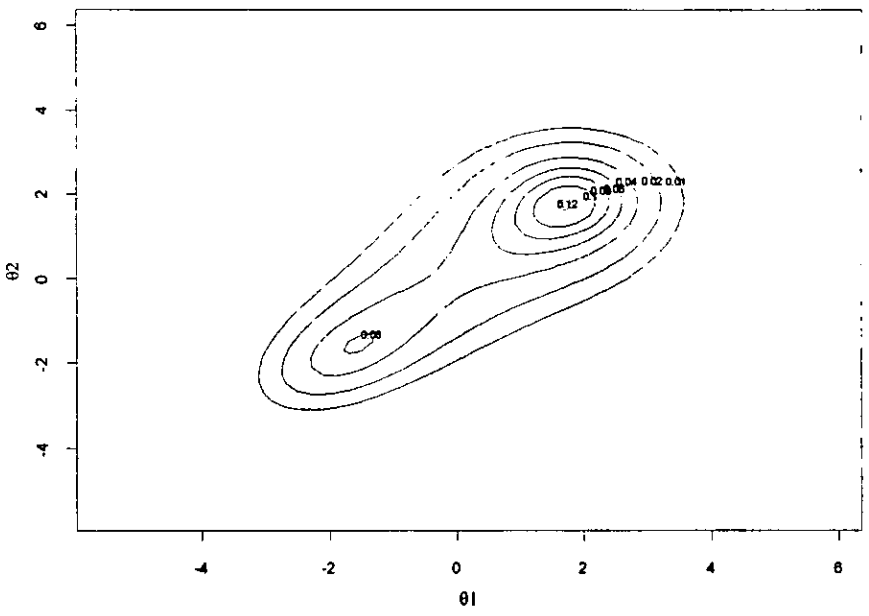


Figura 3.3: (a) Densidad Poly-t bivariada y (b) Aproximación básica (curvas de nivel).

aproximación básica está dada por

$$\widehat{p}(\boldsymbol{\theta}) = 0.36\ N\left(\boldsymbol{\theta}\left|\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\Sigma}}_1\right.\right) + 0.64\ N\left(\boldsymbol{\theta}\left|\widehat{\boldsymbol{\theta}}_2, \widehat{\boldsymbol{\Sigma}}_2\right.\right),$$

donde

$$\widehat{\boldsymbol{\theta}}_1 = \begin{pmatrix} -1.56 \\ -1.56 \end{pmatrix}, \widehat{\boldsymbol{\theta}}_2 = \begin{pmatrix} 1.74 \\ 1.74 \end{pmatrix},$$

$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.77 & 0.50 \\ 0.50 & 0.77 \end{pmatrix} \text{ y } \widehat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.55 & 0.10 \\ 0.10 & 0.55 \end{pmatrix}.$$

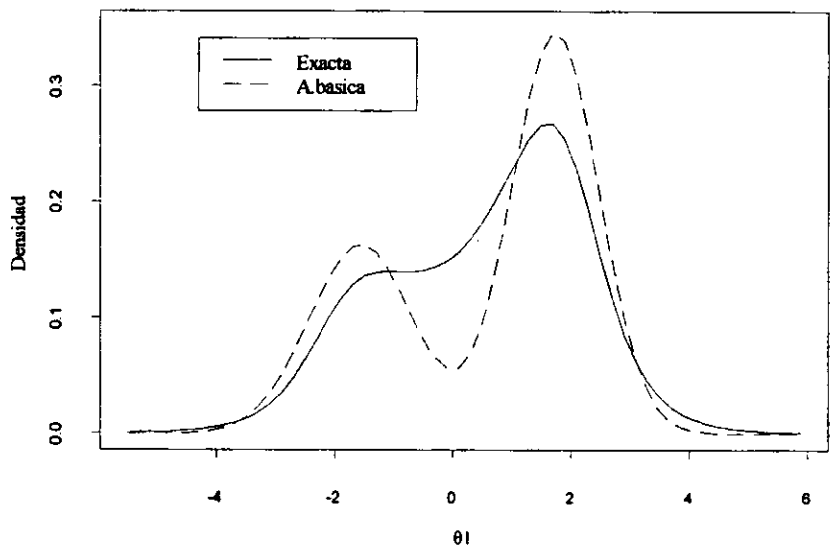


Figura 3.4: Densidad marginal de θ_1 y Aproximación básica.

La gráfica en perspectiva y las curvas de nivel de esta aproximación se encuentran en las Figuras 3.2(b) y 3.3(b). La aproximación básica captura muy bien el comportamiento bimodal de la distribución verdadera, pero con una mayor concentración

respecto a las modas. En la Figura 3.3(b), donde se presentan las curvas de nivel para los mismos niveles que la verdadera, se aprecia mejor que la aproximación básica no es adecuada en la región que se encuentra entre las modas. La medida (3.2) de la precisión de la aproximación a la densidad poly-t es 0.0671.

En las aplicaciones, es frecuente que el interés se centre en el cálculo de las distribuciones marginales, que en el caso de las distribuciones poly-t no se conoce una forma explícita. La distribución marginal para θ_1 obtenida mediante integración numérica y su correspondiente aproximación básica se presentan en la Figura 3.4.

Como era de esperarse, la aproximación básica correspondiente a la distribución marginal tampoco resulta ser adecuada. En este caso, la medida (3.2) de la precisión de la aproximación a la densidad marginal es 0.1002. ■

3.1.2 Nueva aproximación.

Motivación.

Las mezclas de densidades son esencialmente densas en el espacio de densidades (Diaconis e Ylvisaker, 1985), en el sentido de que cualquier densidad puede ser aproximada por una mezcla suficientemente grande; esto sugiere que una mezcla de densidades normales puede dar lugar a una aproximación adecuada de cualquier densidad con soporte no acotado, si el número de componentes de la mezcla es suficientemente grande. En este caso, la elección de las correspondientes medias y varianzas no es tan importante. Por ejemplo, en la estimación de densidades por el método de kernel (Silverman, 1986; Capítulos 3 y 4) la localización está dada por los puntos muestrales

y la dispersión es un factor que únicamente influye en la suavidad de la aproximación.

A continuación se presenta el siguiente resultado.

Resultado 3.1.1.

Sea

$$p\left(\boldsymbol{\theta}\,|\mathbf{x}\right) = cp_x\left(\boldsymbol{\theta}\right),$$

donde c es la constante de normalización.

Sea $\left\{\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2, \ldots, \widetilde{\boldsymbol{\theta}}_m\right\}$ un conjunto de puntos cualesquiera en la región de mayor densidad de $p_x\left(\boldsymbol{\theta}\right)$.

La densidad final $p\left(\boldsymbol{\theta}\,|\mathbf{x}\right)$ se puede aproximar mediante

$$\widetilde{p}\left(\boldsymbol{\theta}\,|\mathbf{x}\right) = \sum_{i=1}^{m} w_i N_k\left(\boldsymbol{\theta}\,\middle|\widetilde{\boldsymbol{\theta}}_i, h\mathbf{I}_k\right) \tag{3.3}$$

donde, $h > 0$ es un parámetro de suavizamiento y los pesos w_i, $i = 1, \ldots, m$ son la solución al sistema de ecuaciones lineales

$$\begin{pmatrix} p_x\left(\widetilde{\boldsymbol{\theta}}_1\right) & N_{1,m} - N_{1,1} & \cdots & N_{1,m} - N_{1,m-1} \\ p_x\left(\widetilde{\boldsymbol{\theta}}_2\right) & N_{2,m} - N_{2,1} & \cdots & N_{2,m} - N_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_x\left(\widetilde{\boldsymbol{\theta}}_m\right) & N_{m,m} - N_{m,1} & \cdots & N_{m,m} - N_{m,m-1} \end{pmatrix} \begin{pmatrix} c \\ w_1 \\ \vdots \\ w_{m-1} \end{pmatrix} = \begin{pmatrix} N_{1,m} \\ N_{2,m} \\ \vdots \\ N_{m,m} \end{pmatrix} \tag{3.4}$$

con $N_{i,j} = N_k\left(\widetilde{\boldsymbol{\theta}}_i\,\middle|\widetilde{\boldsymbol{\theta}}_j, h\mathbf{I}_k\right)$.

Justificación.

Esta nueva aproximación está motivada por el hecho de que los pesos de la aproximación básica no son muy flexibles en el sentido de que solamente dependen de la altura de las modas y de la curvatura de la densidad en las modas. La idea es tratar de determinar los pesos de una mezcla de densidades normales, suponiendo que se conocen las medias y las varianzas de cada uno de los componentes, así como el número de componentes.

Sea

$$p(\boldsymbol{\theta}|\mathbf{x}) = cp_x(\boldsymbol{\theta}) = \sum_{i=1}^{m} w_i N_k(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

una mezcla de densidades normales con las medias y varianzas $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, m$ conocidas y los pesos w_i, $i = 1, \ldots, m$ desconocidos. Para poder determinar los pesos que definen a la mezcla, basta con evaluar la distribución final en cada una de las medias, obteniéndose las siguientes ecuaciones

$$cp_x(\boldsymbol{\mu}_j) = \sum_{i=1}^{m} w_i N_k(\boldsymbol{\mu}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad j = 1, \ldots, m.$$

Usando el hecho de que $\sum_{i=1}^{m} w_i = 1$ y resolviendo el sistema de m ecuaciones con m incógnitas se obtienen los pesos que definen a la mezcla.

Al tratar de aproximar cualquier distribución final mediante una mezcla de densidades normales y basándonos en el hecho de que las mezclas son densas, se propone utilizar un número considerablemente grande de puntos arbitrarios como localización de las componentes con una varianza común y finalmente los pesos se encuentran mediante el sistema de ecuaciones anterior, obteniéndose así la aproximación propuesta en el Resultado 3.1.1. ■

La precisión de la aproximación (3.3) depende de varios aspectos importantes. En primer lugar, es conveniente que la distribución que se quiera aproximar tenga un soporte no acotado, y no es necesario que tenga un comportamiento simétrico. Por otro lado, el número de puntos necesarios para lograr una buena aproximación depende del comportamiento de la distribución, pero en general la nueva aproximación será mejor al aumentar en número de puntos m. En particular, el grado de suavidad de la función $p_x(\boldsymbol{\theta})$ determina el valor del parámetro de suavizamiento h, que depende principalmente del número de puntos así como de la separación entre ellos.

Los pesos obtenidos mediante el sistema de ecuaciones (3.4) siempre suman uno, pero en algunos casos (para ciertos valores de h), los pesos pueden tomar valores negativos o mayores que uno.

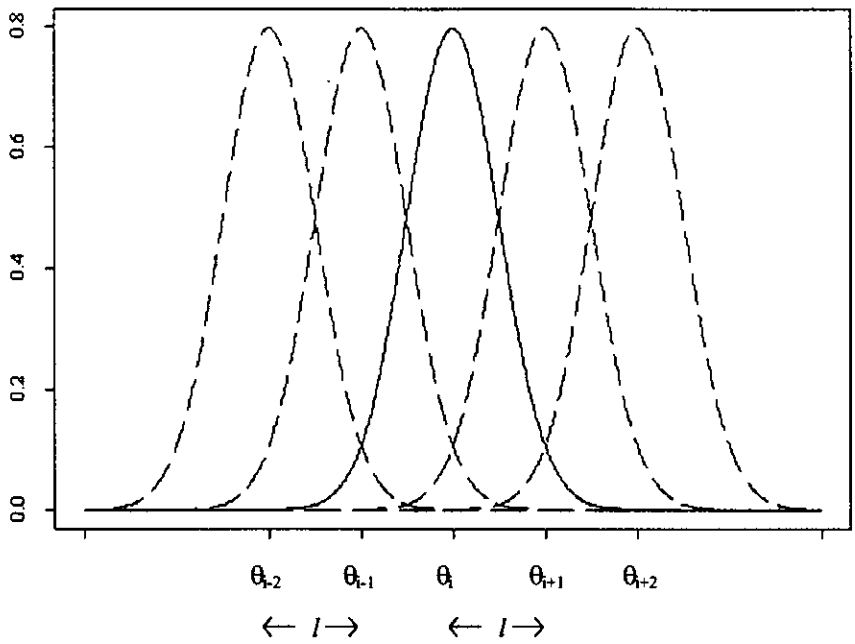
Es conveniente que los puntos elegidos sean equidistantes para que el parámetro h sea fácil de encontrar, en ese caso, una posible selección inicial de h sería

$$h^* = \left(\frac{3}{2}\right)^{2k} \frac{l^2}{9}, \tag{3.5}$$

donde l es la distancia ortogonal entre las coordenadas de cada punto contiguo. Por ejemplo, si $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, entonces $l = \widetilde{\theta}_{1_i} - \widetilde{\theta}_{1_{i-1}} = \widetilde{\theta}_{2_i} - \widetilde{\theta}_{2_{i-1}}$.

El punto inicial (3.5) tiene una justificación empírica basada en el hecho de que para valores de h cercanos a h^*, la aproximación resultante es suave. En el caso univariado se observó que la suavidad de la aproximación depende de la cantidad de densidad, de cada componente de la mezcla, que se intersecta para cada punto $\widetilde{\boldsymbol{\theta}}_i$. En varios ejemplos se obtuvo un comportamiento suave cuando únicamente las densidades (componentes) localizadas en los puntos contiguos $\widetilde{\boldsymbol{\theta}}_{i-2}$, $\widetilde{\boldsymbol{\theta}}_{i-1}$, $\widetilde{\boldsymbol{\theta}}_{i+1}$ y $\widetilde{\boldsymbol{\theta}}_{i+2}$

se combinan con la densidad centrada en el punto $\widetilde{\boldsymbol{\theta}}_i$, para lo cual se requería que la varianza cumpliera con la restricción $\frac{3}{2}l = 3\sqrt{h}$ (ver siguiente figura).



Finalmente, despejando el valor de h se obtiene la elección (3.5). Para distintas distribuciones se probó esta selección inicial de h, en una y dos dimensiones, obteniéndose buenos resultados para alguna h cercana a la obtenida por (3.5).

Muchas veces, al imponer la restricción de que los pesos obtenidos como solución del sistema de ecuaciones (3.4) sean todos positivos (para algun h conveniente), la aproximación resultante puede no tener un comportamiento suave, en el sentido de que puede tener muchos picos. Es posible lograr mejores aproximaciones (al menos en la región de mayor densidad bajo $p(\boldsymbol{\theta}|\mathbf{x})$) permitiendo que algunos de los pesos sean negativos, pero en este caso no se podría interpretar a la aproximación como una densidad. Una solución a ésto sería asignar el valor de cero a los pesos negativos

y renormalizar los pesos restantes para que sumen uno (siempre y cuando el valor de los pesos negativos sea relativamente pequeño en valor absoluto).

Para el caso de $m = 2, 3$ se estudió el comportamiento de la matriz (3.4) y se demostró que para cualquier distribución final tal que $p_x\left(\widetilde{\boldsymbol{\theta}}_i\right) > 0$, $i = 1, \ldots, m$ y $h > 0$ el determinante de la matriz es estrictamente positivo, por lo que el sistema siempre tiene solución. Lo anterior y la experiencia empírica al trabajar con matrices con $m \geq 4$ sugieren que la matriz siempre tiene solución también en este último caso.

Igual que en el caso de la aproximación básica, si la aproximación (3.3) es adecuada, entonces cualquier resumen inferencial de la distribución final $p(\boldsymbol{\theta}|\mathbf{x})$ se puede aproximar por el correspondiente resumen inferencial de la aproximación pro-puesta. En particular, el valor esperado final y la varianza final se pueden aproximar respectivamente por

$$\widetilde{E}[\boldsymbol{\theta}|\mathbf{x}] = \sum_{i=1}^{m} w_i \widetilde{\boldsymbol{\theta}}_i \quad \text{y}$$

$$\begin{aligned} \widetilde{\text{Var}}[\boldsymbol{\theta}|\mathbf{x}] &= \sum_{i=1}^{m} w_i \left(h\mathbf{I}_k + \widetilde{\boldsymbol{\theta}}_i \widetilde{\boldsymbol{\theta}}_i'\right) - \widetilde{E}[\boldsymbol{\theta}|\mathbf{x}]\, \widetilde{E}[\boldsymbol{\theta}|\mathbf{x}]' \\ &= h\sum_{i=1}^{m} w_i \mathbf{I}_k + \sum_{i=1}^{m} w_i \left(\widetilde{\boldsymbol{\theta}}_i - \widetilde{E}[\boldsymbol{\theta}|\mathbf{x}]\right)\left(\widetilde{\boldsymbol{\theta}}_i - \widetilde{E}[\boldsymbol{\theta}|\mathbf{x}]\right)'. \end{aligned}$$

Por otro lado, las distribuciones marginales para cada θ_j, $j = 1, \ldots, k$ se pueden aproximar mediante

$$\widetilde{p}(\theta_j|\mathbf{x}) = \sum_{i=1}^{m} w_i N\left(\theta_j \left| \widetilde{\boldsymbol{\theta}}_{i,j}, h\right.\right),$$

donde $\widehat{\boldsymbol{\theta}}_{i,j}$ es la j-ésima coordenada del vector $\widehat{\boldsymbol{\theta}}_i$.

Ejemplo 3.1.3.

Con referencia al Ejemplo 1.2.2, para un tamaño de muestra $n = 2$, y con $x_1 = -5$ y $x_2 = 3$, se cumple que la distribución final de θ es bimodal (Figura 1.2(a)).

Sean $\widetilde{\theta}_1 = -10$, y $\widetilde{\theta}_m = 8$ el primer y el último punto sobre la región de mayor densidad de $p_x(\theta) = \frac{1}{\pi^2\left[1+(-5-\theta)^2\right]\left[1+(3-\theta)^2\right]}$. (Se puede observar que $p_x(\theta)$ es el kernel de una distribución poly-t con parámetros $k = 1$, $r = 2$, $\mu_1 = -5$, $\mu_2 = 3$, $M_1 = M_2 = 1$ y $\nu_1 = \nu_2 = 1$).

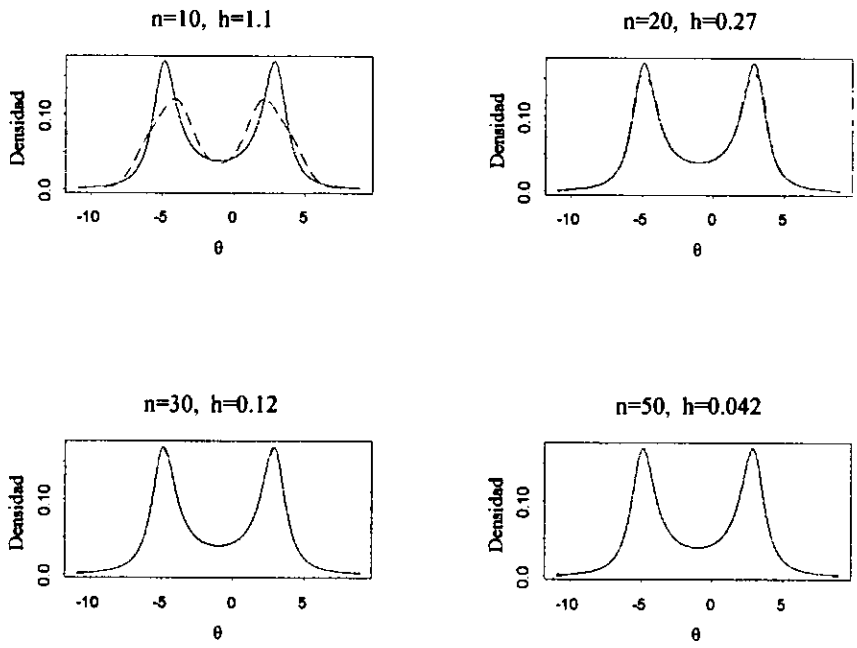


Figura 3.5: Densidad final de θ y Nueva aproximación.

La Figura 3.5 muestra la aproximación (3.3) para $m = 10, 20, 30, 50$ puntos equidistantes en el intervalo $(-10, 8)$. El parámetro h con el que se realizaron las aproximaciones fueron encontrados por ensayo y error, tomando como punto de partida la elección (3.5). Por ejemplo, para $m = 50$, $l = 0.36$, y $k = 1$ se obtiene

$h^* = 0.032$, y el valor de h que mejor funcionó fue 0.042. En todos los casos, el valor de h elegido finalmente es cercano al correspondiente valor de h^*.

En la Figura 3.5 se observa que la aproximación propuesta mejora considerablemente al aumentar el número de puntos. Con $m = 10$ puntos se ve un comportamiento bimodal, pero la aproximación no se parece en nada a la verdadera. Con $m = 20$ puntos se captura perfectamente la bimodalidad y la dispersión, fallando un poco en la altura de las modas. Al aumentar a 30 y a 50 el número de puntos, las gráficas de las respectivas aproximaciones no se alcanzan a apreciar a simple vista, debido a que reproducen casi perfectamente a la verdadera distribución.

Otra forma de ver que la aproximación mejora al aumentar en número de puntos m, es mediante la medida de la precisión (3.2). Con $m = 10, 20, 30$ y 50 las correspondintes medidas son 0.0644, 0.0170, 0.0040 y 0.0018 respectivamente, lo que confirma que al aumentar el número de puntos mejora la aproximación. Comparando esta aproximación con la aproximación básica (Figura 3.1(a)), aún con $m = 10$ puntos la nueva aproximación tiene una medida de precisión menor que la correspondiente a la aproximación básica que es 0.1035.

Como la aproximación a la distribución poly-t es bastante buena, se calcularán algunos resúmenes inferenciales de interés, tales como la constante de normalización y los tres primeros momentos, y se compararán con los encontrados mediante una regla trapezoidal. Los errores relativos para las cuatro distribuciones poly-t, con los mismos puntos muestrales $\mathbf{x}$ del Ejemplo 1.2.2 para los cuales se obtienen distintas separaciones entre las modas (Figura 1.2), se presentan en la Tabla 3.1.1.

Tabla 3.1.1

x	c^{-1}	$E(\theta)$	$E(\theta^2)$	$E(\theta^3)$	$\widetilde{\theta}_1, \widetilde{\theta}_m$
$(-5,3)$	2.2%	0.0%	5.2%	5.4%	-10,8
$(-4,1)$	2.0%	0.2%	5.4%	5.4%	-9,6
$(-2,2)$	2.1%	0.0%	8.8%	0.0%	-6,6
$(-1.1,1.1)$	2.4%	0.0%	14.1%	0.0%	-4,4

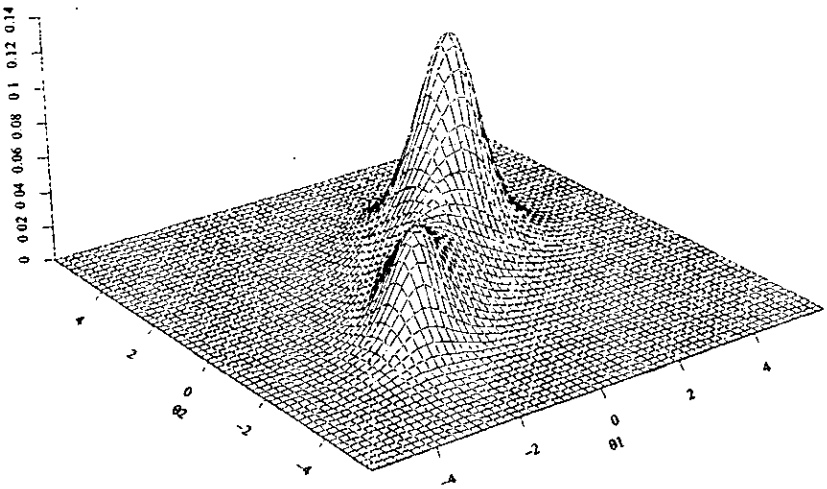
Para calcular las aproximaciones a los momentos se utilizaron $m = 30$ puntos equidistantes sobre el intervalo $\left(\widetilde{\theta}_1, \widetilde{\theta}_m\right)$ y con $h = h^*$ calculado a partir de (3.5).

En la tabla anterior se observa que las aproximaciones tienen errores menores a 10% (exceptuando la aproximación al segundo momento de la cuarta distribución), a pesar de que sólo se utilizaron 30 puntos. Se puede lograr una mayor precisión si se utiliza un mayor número de puntos. ■

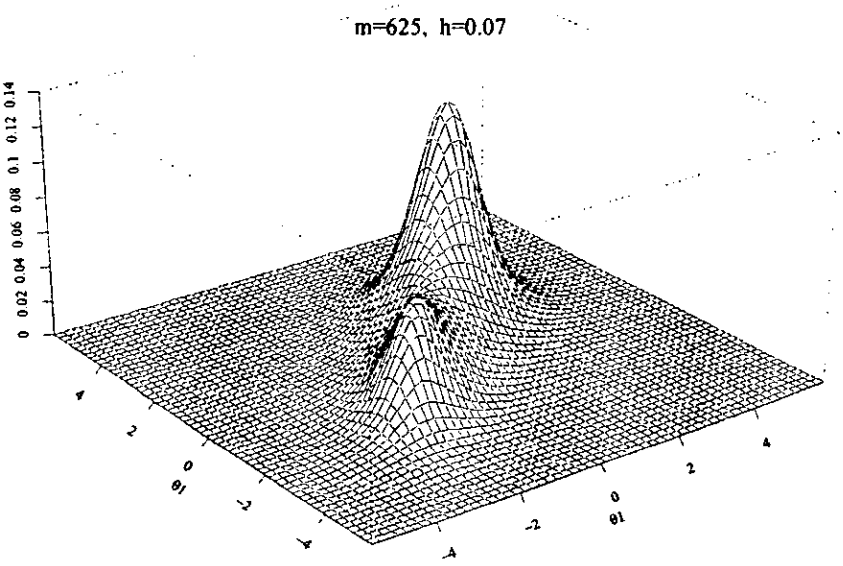
Ejemplo 3.1.4.

Considere el Ejemplo 3.1.2. Para los mismos parámetros, $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ tiene una distribución poly-t bimodal. Aplicando la aproximación (3.3) a esta distribución, con base en $m = 25^2 = 625$ puntos en una retícula regular sobre el cuadrado $(-4, 4.5) \times (-4, 4.5)$ y un parámetro de suavizamiento $h = 0.07$ se obtienen las Figuras 3.6(b) y 3.7(b).

La gráfica en perspectiva de la aproximación (Figura 3.6(b)) parece ser idéntica a la de la Figura 3.6(a). Aunque la gráfica en perspectiva pudiera ser engañosa, al observar las curvas de nivel en la Figura 3.7(b) y compararlas con las verdaderas curvas de nivel (Figura 3.7(a)), se observa que también son casi idénticas, reproduciéndose perfectamente la estructura de correlación entre las variables θ_1 y θ_2, así como la lo-

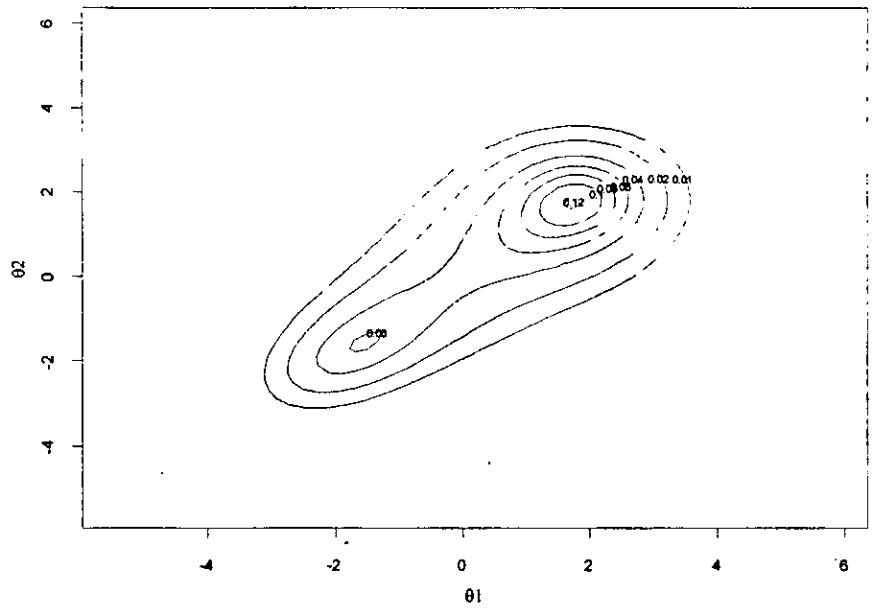


(a)



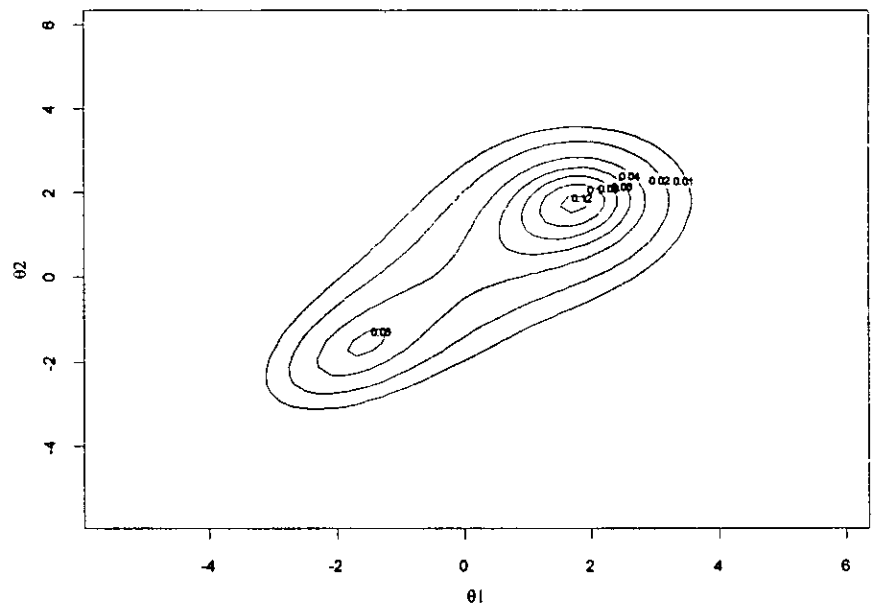
(b)

Figura 3.6: (a) Densidad Poly-t bivariada y (b) Nueva aproximación (perspectiva).



(a)

m=625, h=0.07



(b)

Figura 3.7: (a) Densidad Poly-t bivariada y (b) Nueva aproximación (curvas de nivel).

calización y la altura de las modas. La medida (3.2) de la precisión de la aproximación es 0.0023, que es un número muy cercano a cero, lo que indica que la aproximación es muy buena.

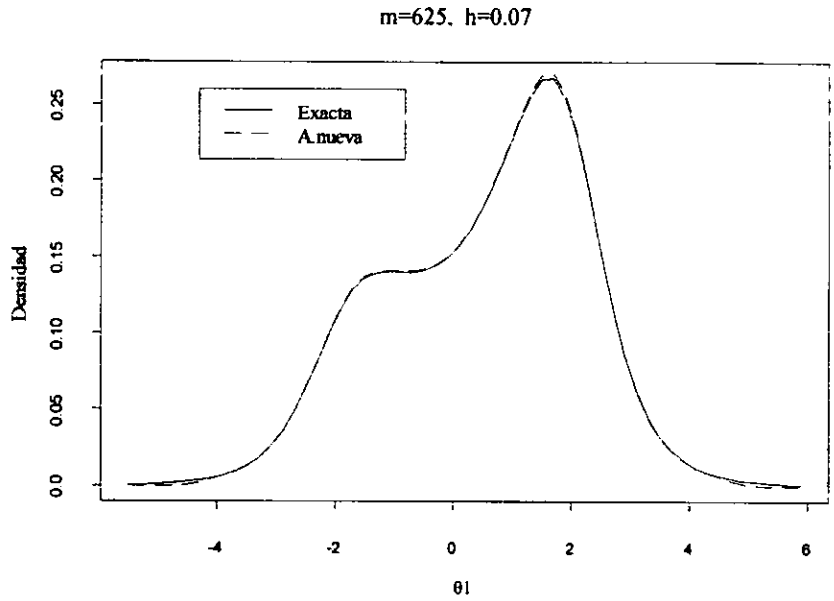


Figura 3.8: Densidad marginal de θ_1 y Nueva aproximación.

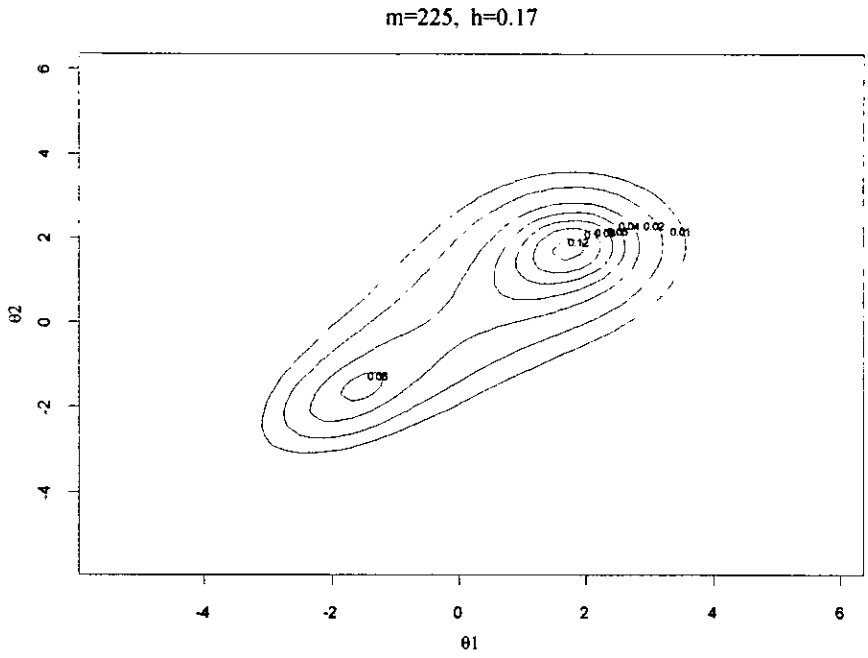
De la misma manera, la Figura 3.8 presenta la distribución marginal de θ_1 y su correspondiente aproximación obtenida a través del método propuesto. En esta figura se observa que la aproximación es muy precisa, aunque con una ligera diferencia en los extremos de las colas. La medida de la precisión en este caso es 0.0028, que comparado con la medida de la precisión de la aproximación básica (Sección 3.1.1) a esta distribución marginal (0.1002), es mucho menor.

Con el objeto de determinar el mínimo número de puntos necesarios para obtener

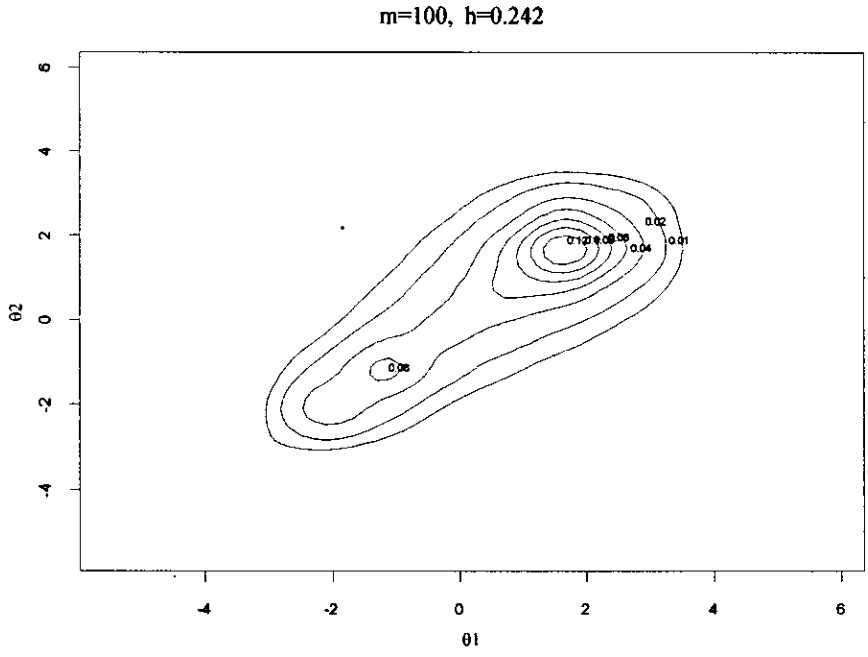
una aproximación razonable, a continuación se presentan las curvas de nivel para la aproximación de la distribución conjunta de $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ con base en $m = 15^2 = 225$ y $m = 10^2 = 100$ puntos en una retícula regular sobre la misma región $(-4, 4.5) \times (-4, 4.5)$.

En la Figura 3.9(a) se presentan las curvas de nivel de la aproximación con 225 puntos. Se puede observar que la aproximación es bastante buena, localiza muy bien las modas y refleja la estructura de correlación perfectamente bien, aunque la altura en las modas es un poco menor. Su medida de precisión es 0.0056 un poco mayor que cuando se utilizaron 625 puntos, pero sigue siendo un valor bastante bajo.

Con una rejilla de 100 puntos, las curvas de nivel de la nueva aproximación (Figura 3.9(b)) son menos suaves que las de la verdadera densidad, pero aún capturan adecuadamente tanto la localización como la estructura de correlación de ésta. Obsérvese que aún con 100 puntos la nueva aproximación es mejor que la aproximación básica a está densidad (Figuras 3.2(b) y 3.3(b)). Al comparar las medidas de la precisión con 100 puntos (0.0181) con la correspondiente medida de la aproximación básica (0.0671) se corrobora que la aproximación básica es mucho mejor. ■



(a)



(b)

Figura 3.9: Nueva aproximación con (a) $m = 225$ puntos y (b) $m = 100$ puntos (curvas de nivel).

Ejemplo 3.1.5.

Supongamos que la distribución final de $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ consiste en una mezcla de densidades normales, por ejemplo,

$$p(\boldsymbol{\theta}|\mathbf{x}) = 0.3N(\boldsymbol{\theta}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.2N(\boldsymbol{\theta}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.5N(\boldsymbol{\theta}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3),$$

donde

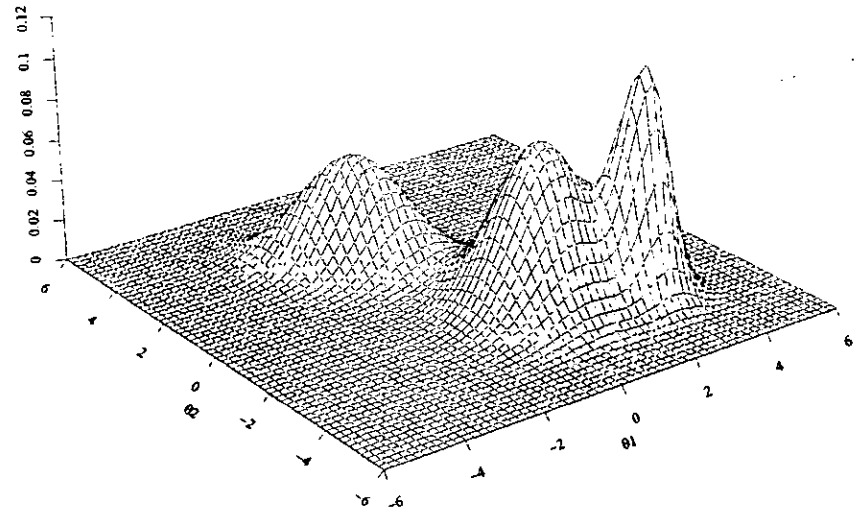
$$\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ -3 \end{pmatrix}, \ \boldsymbol{\mu}_3 = \begin{pmatrix} 1 \\ -2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix} \text{ y } \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

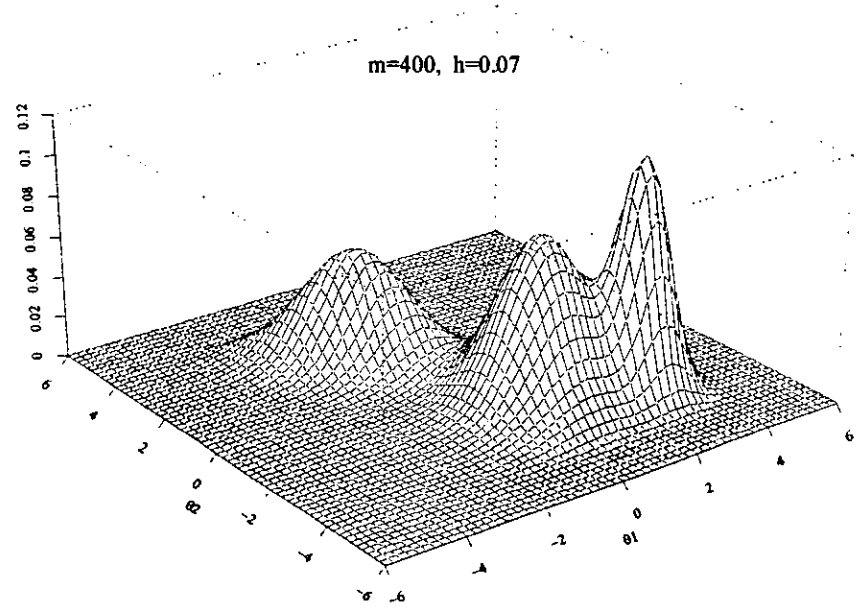
Las Figuras 3.10(a) y 3.11(a) presentan la gráfica en perpectiva y las curvas de nivel de esta distribución. En ellas se puede observar que el comportamiento de la distribución es trimodal, con dos modas relativamente juntas y la otra un poco más alejada.

Al aplicar la aproximación (3.3) con base en $m = 20^2 = 400$ puntos distribuidos en una retícula regular sobre la región $(-4, 4) \times (-4, 4)$, y con $h = 0.07$ se obtienen las Figuras 3.10(b) y 3.11(b).

A simple vista, la gráfica de la aproximación en perspectiva (Figura 3.10(b)) parece ser idéntica a la verdadera distribución, pero si se observa cuidadosamente la Figura 3.11(b) se puede notar que en la aproximación el pico correspondiente a la moda central es un poco más pronunciado. A pesar de esto, se puede decir que la aproximación propuesta reproduce casi perfectamente a la verdadera distribución. El valor de la medida de precisión (3.2) es 0.0033, lo que nos indica que la aproximación es bastante

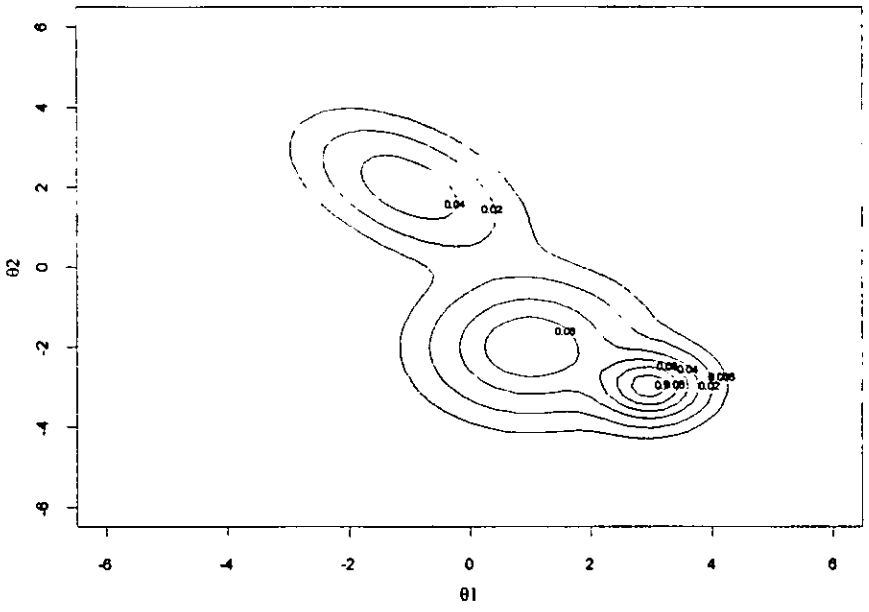


(a)

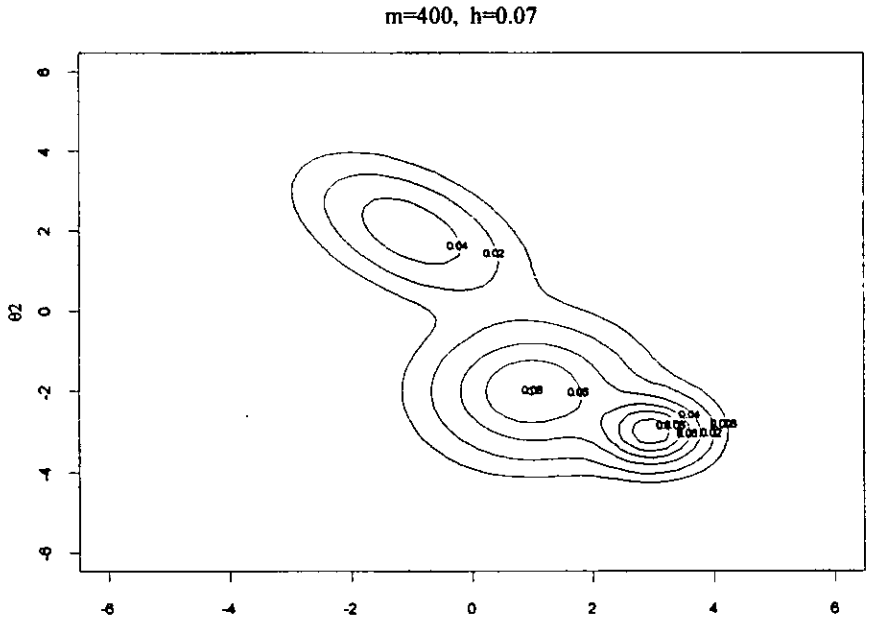


(b)

Figura 3.10: (a) Mezcla de tres densidades Normales y (b) Nueva aproximación (perspectiva).



(a)



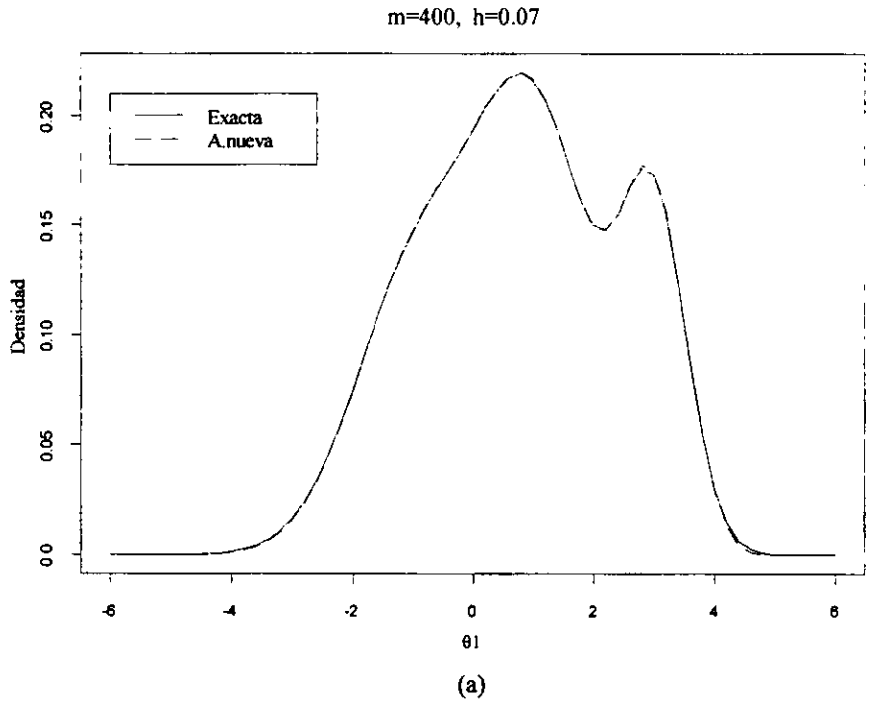
(b)

Figura 3.11: (a) Mezcla de tres densidades Normales y (b) Nueva aproximación (curvas de nivel).

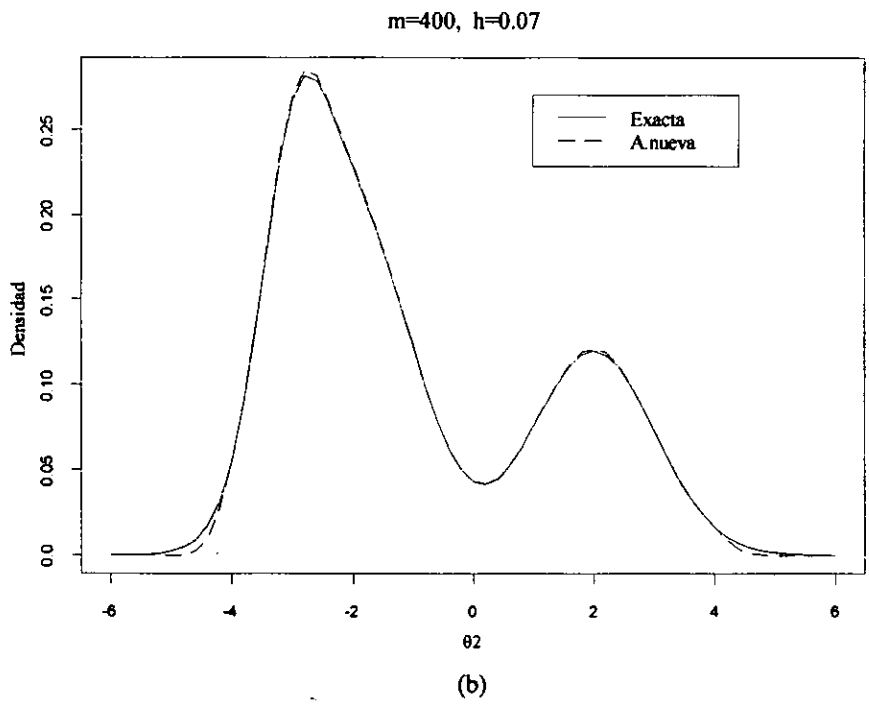
precisa.

Finalmente, una vez que se observó que la aproximación a la distribución conjunta es buena, se pueden calcular las correspondientes aproximaciones a las distribuciones marginales de θ_1 y θ_2. En la Figura 3.12(a) se puede observar que la distribución marginal de θ_1 es bimodal, a pesar de que la distribución conjunta es trimodal. La aproximación a la distribución marginal de θ_1 es bastante buena, reproduciéndose muy bien el comportamiento bimodal y la caída de las colas. Así mismo, la aproximación a la distribución marginal de θ_2 (Figura 3.12(b)) es también bastante razonable, aunque con una ligera diferencia en las colas. Las correspondientes medidas de la precisión dela aproximación a cada una de las densidades marginales es 0.0023 para la densidad marginal de θ_1 y 0.0068 para la densidad marginal de θ_2. Estos valores indican que las aproximaciones son bastante buenas.

Vale la pena comentar que como la densidad $p(\boldsymbol{\theta}|\mathbf{x})$ es una mezcla de densidades normales, la aproximación básica (vista en la Sección 3.1.1) a esta distribución es razonable. La parte de la moda que está más separada es reproducida de manera adecuada por la aproximación básica, pero la parte de las otras dos modas que están más juntas no es razonablemente representada por la aproximación básica. Este comportamiento es de esperarse tomando como referencia a los ejemplos presentados en la Sección 3.1.1. ■



(a)



(b)

Figura 3.12: Densidades marginales y Nueva aproximación (a) θ_1 y (b) θ_2.

3.1.3 Discusión.

La aproximación propuesta en esta tesis es fácil de implementar porque no se tiene que utilizar un algoritmo para encontrar las modas de la función $p_x(\boldsymbol{\theta})$, lo que sí es necesario para implementar la aproximación básica.

Para que la nueva aproximación sea aplicable en el caso de que la dimensión de $\boldsymbol{\theta}$ sea mayor a uno hay que considerar dos puntos importantes. En primer lugar, es conveniente que el rango de valores donde se encuentra la mayor densidad de $p_x(\boldsymbol{\theta})$ sea el mismo para cada componente θ_i. Esto se debe a que la nueva aproximación se basa en una mezcla de densidades normales circulares (lo que simplifica la aproximación determinando únicamente un solo valor para la varianza común (h)) y en otro caso, se tendría que encontrar un valor h_i para cada componente θ_i. En segundo lugar, es conveniente tratar de disminuir lo más posible la correlación entre los parámetros, para que al aplicar la nueva aproximación no se tengan problemas numéricos. El problema se debe a que como los puntos arbitrarios se situan sobre una retícula regular que cubre la mayor densidad de $p_x(\boldsymbol{\theta})$ y al tener correlación alta entre los componentes del parámetro, exitirán muchos puntos $\widetilde{\boldsymbol{\theta}}_i$ para los cuales el valor de $p_x\left(\widetilde{\boldsymbol{\theta}}_i\right)$ sea cero o prácticamente cero.

Además de ser una aproximación a densidades, la aproximación propuesta se puede ver también como una regla de integración, ya que a partir del mismo sistema de ecuaciones (3.4) se obtiene el valor de la integral, c. Si el objetivo es encontrar el valor de la integral, c, entonces es suficiente con resolver el sistema de ecuaciones utilizando h^*.

Por otra parte, la nueva aproximación no distingue si la densidad es unimodal o multimodal; sin embargo, el número de puntos necesarios para tener una buena aproximación aumenta exponencialmente con la dimensión del parámetro. Por ejemplo, si en una dimensión se necesitaron m puntos para tener una buena aproximación, en cambio, en dos dimensiones se necesitarían m^2 puntos para cubrir el plano de mayor densidad y lograr una buena aproximación.

La nueva aproximación tiene ciertas similitudes con la estimación de densidades por el método de kernel (Silverman, 1986), en el sentido de que es una mezcla de densidades normales que utiliza una varianza común como parámetro de suavizamiento. Una diferencia esencial es el hecho de que el método de kernel se basa en una muestra aleatoria de la distribución cuya densidad se quiere estimar y utiliza un peso común para cada componente de la mezcla. En cambio, la aproximación propuesta en esta tesis se basa en puntos arbitrarios (no aleatorios) y los pesos están dados por la solución de un sistema de ecuaciones lineales.

3.2 Aproximación de Laplace multimodal.

Una generalización de la aproximación de Laplace, en el caso de que la función a integrar sea multimodal, se puede lograr utilizando mezclas de densidades normales.

3.2.1 Laplace multimodal básica.

Valores esperados.

Sea $p(\boldsymbol{\theta}|\mathbf{x})$ la distribución final (multimodal) de $\boldsymbol{\theta}$. Para encontrar cualquier resumen inferencial de interés es necesario resolver integrales en donde algunas veces el integrando es una función multimodal. En particular, supóngase que se quiere encontrar el valor esperado de una función real $g(\boldsymbol{\theta})$ con respecto a la distribución final de $\boldsymbol{\theta}$, es decir,

$$E[g(\boldsymbol{\theta})|\mathbf{x}] = \frac{\int g(\boldsymbol{\theta}) p_x(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}}{\int p_x(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}},$$

donde $p_x(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$.

Se puede reexpresar el integrando de cada una de las integrales del cociente anterior de manera que el valor esperado tome la forma

$$E[g(\boldsymbol{\theta})|\mathbf{x}] = \frac{\int b_N(\boldsymbol{\theta})\, h_N(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}}{\int b_D(\boldsymbol{\theta})\, h_D(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}}, \tag{3.6}$$

para algunas funciones $b_J(\cdot)$ y $h_J(\cdot)$, $J = N, D$ escogidas convenientemente.

Supongamos que las funciones $h_J(\cdot)$, $J = N, D$ son multimodales y no negativas. Entonces la aproximación de Laplace multimodal, basada en la aproximación normal básica, para cada una de las integrales anteriores es de la forma

$$\int b(\boldsymbol{\theta})\, h(\boldsymbol{\theta})\, \partial\boldsymbol{\theta} \approx \sum_{i=1}^{d} (2\pi)^{\frac{k}{2}} \left|\widehat{\boldsymbol{\Sigma}}_i\right|^{\frac{1}{2}} b\left(\widehat{\boldsymbol{\theta}}_i\right) h\left(\widehat{\boldsymbol{\theta}}_i\right) \tag{3.7}$$

donde $\widehat{\boldsymbol{\theta}}_i$ es la i-ésima moda de $h(\boldsymbol{\theta})$, y $\widehat{\boldsymbol{\Sigma}}_i = -\left[\frac{\partial^2 \log h\left(\widehat{\boldsymbol{\theta}}_i\right)}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right]^{-1}$, $i = 1, \ldots, d$.

Una justificación heurística para la aproximación (3.7) es la siguiente.

Sea $h(\boldsymbol{\theta})$ una función multimodal no negativa con d modas. Utilizando una expansión en serie de Taylor (e ignorando los términos de orden mayor que dos) alrededor de cada una de las modas de la función $h(\boldsymbol{\theta})$ se tiene que

$$h(\boldsymbol{\theta}) \approx h\left(\widehat{\boldsymbol{\theta}}_i\right) \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)' \widehat{\boldsymbol{\Sigma}}_i^{-1}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)\right\}, \quad i=1,\ldots,d.$$

donde $\widehat{\boldsymbol{\theta}}_i$ es la i-ésima moda y $\widehat{\boldsymbol{\Sigma}}_i = -\left[\frac{\partial^2 \log h\left(\widehat{\boldsymbol{\theta}}_i\right)}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]^{-1}$.

Por otro lado, realizando la expansión en serie de Taylor (e ignorando los términos de orden mayor que uno) de la función $b(\boldsymbol{\theta})$ alrededor de cada una de las modas de $h(\boldsymbol{\theta})$ obtenemos

$$b(\boldsymbol{\theta}) \approx b\left(\widehat{\boldsymbol{\theta}}_i\right) + \frac{\partial}{\partial \boldsymbol{\theta}} b\left(\widehat{\boldsymbol{\theta}}_i\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right), \quad i=1,\ldots,d.$$

Combinando las aproximaciones alrededor de cada moda entonces, el integrando $b(\boldsymbol{\theta})\, h(\boldsymbol{\theta})$ se puede aproximar mediante

$$\sum_{i=1}^{d}\left[b\left(\widehat{\boldsymbol{\theta}}_i\right) + \frac{\partial}{\partial \boldsymbol{\theta}} b\left(\widehat{\boldsymbol{\theta}}_i\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)\right]\left[h\left(\widehat{\boldsymbol{\theta}}_i\right) \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)' \widehat{\boldsymbol{\Sigma}}_i^{-1}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_i\right)\right\}\right].$$

Por lo tanto, al realizar la integral se cancela el segundo término del primer corchete, obtiéndose la aproximación de Laplace multimodal básica (3.7). ■

Continuando con la aproximación al valor esperado final, aplicando (3.7) a cada una de las integrales del cociente (3.6) se obtiene

$$\widehat{E}\left[g(\boldsymbol{\theta})\,|\mathbf{x}\right] = \frac{\sum_{i=1}^{d_N}\left|\widehat{\boldsymbol{\Sigma}}_{N_i}\right|^{\frac{k}{2}} b_N\left(\widehat{\boldsymbol{\theta}}_{N_i}\right) h_N\left(\widehat{\boldsymbol{\theta}}_{N_i}\right)}{\sum_{i=1}^{d_D}\left|\widehat{\boldsymbol{\Sigma}}_{D_i}\right|^{\frac{k}{2}} b_D\left(\widehat{\boldsymbol{\theta}}_{D_i}\right) h_D\left(\widehat{\boldsymbol{\theta}}_{D_i}\right)}, \tag{3.8}$$

donde $\widehat{\boldsymbol{\theta}}_{J_i}$ es la i-ésima moda de $h_J(\boldsymbol{\theta})$, y $\widehat{\boldsymbol{\Sigma}}_{J_i} = -\left[\frac{\partial^2 \log h_J\left(\widehat{\theta}_i\right)}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]^{-1}$, $i=1,\ldots,d_J$ y $J=N,D$.

En particular, si se utiliza una factorización en forma "estándar", es decir, $h_N(\boldsymbol{\theta}) = h_D(\boldsymbol{\theta}) = p_x(\boldsymbol{\theta})$, $b_N(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ y $b_D(\boldsymbol{\theta}) = 1$, entonces la aproximación de Laplace multimodal se reduce a

$$\widehat{E}\left[g(\boldsymbol{\theta})\,|\mathbf{x}\right] = \sum_{i=1}^{d} w_i g(\widehat{\boldsymbol{\theta}}_i),$$

donde $\widehat{\boldsymbol{\theta}}_i$, $i = 1, \ldots, d$ son las modas de $p_x(\boldsymbol{\theta})$ y $w_i = \frac{p_x(\widehat{\boldsymbol{\theta}}_i)|\widehat{\boldsymbol{\Sigma}}_i|^{\frac{1}{2}}}{\sum_{j=1}^{d} p_x(\widehat{\boldsymbol{\theta}}_j)|\widehat{\boldsymbol{\Sigma}}_j|^{\frac{1}{2}}}$, $i = 1, \ldots, d$ (pesos de la aproximación básica).

Al igual que la aproximación básica a densidades, esta generalización de la aproximación de Laplace no tiene ninguna propiedad asintótica. Una característica importante de esta generalización es que, en el caso de que la función $h(\boldsymbol{\theta})$ en (3.7) sea unimodal, la aproximación propuesta en esta tesis se reduce a la aproximación de Laplace (2.6) vista en la Sección 2.2.1.

Ejemplo 3.2.1.

Considere el Ejemplo 1.2.2, donde la distribución final para θ es una poly-t univariada. Suponga que es de interés calcular los primeros 3 momentos de esta distribución, es decir, $E\left[\boldsymbol{\theta}^j\,|\mathbf{x}\right] = \frac{\int \theta^j p_x(\theta)\partial\theta}{\int p_x(\theta)\partial}$, $j = 1, 2, 3$. Tomemos el caso en el que para $n = 2$ la distribución final es bimodal y se utilizan los mismos puntos x_1 y x_2 del Ejemplo 1.2.2.

Al aplicar (3.8) para aproximar los primeros 3 momentos de cada una de las cuatro distribuciones del Ejemplo 1.2.2, se tiene lo siguiente.

Al realizar una factorización estándar, donde $b_N(\theta) = \theta^j$, $h_N(\theta) = p_x(\theta)$, $b_D(\theta) = 1$ y $h_N(\theta) = p_x(\theta)$ para $j = 1, 2, 3$, ó utilizando una factorización exponencial, donde $b_N(\theta) = 1$, $h_N(\theta) = \theta^j p_x(\theta)$, $b_D(\theta) = 1$ y $h_N(\theta) = p_x(\theta)$ para $j = 2$, se

obtienen los siguientes errores relativos para cada una de las aproximaciones. (La factorización exponencial no se realizó para $j = 1, 3$ porque la función $h_N(\theta)$ toma valores negativos).

Tabla 3.2.1

x	Factorización	c^{-1}	$E(\theta)$	$E(\theta^2)$	$E(\theta^3)$
$(-5, 3)$	estándar	37.9%	0.0%	2.5%	2.6%
	exponencial			0.4%	
$(-4, 1)$	estándar	28.5%	0.1%	14.5%	16.6%
	exponencial			10.9%	
$(-2, 2)$	estándar	18.4%	0.0%	31.2%	0.0%
	exponencial			20.6%	
$(-1.1, 1.1)$	estándar	148.0%	0.0%	88.4%	0.0%
	exponencial			72.2%	

El valor "exacto" utilizado para obtener los errores relativos, se calculó numéricamente utilizando la regla trapezoidal.

En la tabla anterior se observa que los errores relativos de la aproximación de Laplace multimodal son, en algunos casos, bastante grandes. La aproximación al primer y tercer momentos, al utilizar la factorización estándar, es buena para las cuatro distribuciones debido a que tanto las ditribuciones originales como las respectivas aproximaciones básicas son simétricas. En el caso de la constante de normalización y el segundo momento, para ninguna de las cuatro distribuciones la aproximación es adecuada si se utiliza la factorización estándar. Es conveniente notar que al usar la

factorización exponencial para aproximar el segundo momento, el error relativo disminuye en los cuatro casos comparado con el error correspondiente de la aproximación estándar, aunque sigue siendo un error considerable.

En general, se podría decir que esta aproximación no da buenos resultados, debido a que la aproximación básica utilizada para aproximar la función $h(\theta)$ no es muy buena.

Densidades marginales.

Una de las principales aplicaciones de la aproximación de Laplace (vista en la Sección 2.2) es en el cálculo de distribuciones marginales, ya que permite obtener una aproximación analítica. La aproximación de Laplace multimodal también puede ser utilizada para obtener distribuciones marginales.

Sea $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, con $\boldsymbol{\theta}_1 \in \Re^{k_1}$ y $\boldsymbol{\theta}_2 \in \Re^{k-k_1}$. Supongamos que la distribución de $\boldsymbol{\theta}$ es multimodal y que se puede escribir como

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \propto b(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

donde $h(\cdot)$ es una función multimodal. Nos interesa calcular la densidad marginal de $\boldsymbol{\theta}_1$, es decir,

$$p(\boldsymbol{\theta}_1) \propto \int b(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, \partial\boldsymbol{\theta}_2. \tag{3.9}$$

Para cada valor de $\boldsymbol{\theta}_1$, se definen

$$b_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = b(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \quad \text{y} \quad h_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

de manera que $b_{\boldsymbol{\theta}_1}(\cdot)$ y $h_{\boldsymbol{\theta}_1}(\cdot)$ son respectivamente $b(\cdot)$ y $h(\cdot)$ vistas únicamente

como funciones de $\boldsymbol{\theta}_2$. Finalmente, supongamos que $h_{\boldsymbol{\theta}_1}(\cdot)$ tiene máximos relativos en $\widehat{\boldsymbol{\theta}}_{2_i} = \widehat{\boldsymbol{\theta}}_{2_i}(\boldsymbol{\theta}_1)$, $i = 1, \ldots, d$.

Aplicando el resultado (3.7) a la ecuación (3.9), se obtiene que la aproximación de Laplace multimodal básica a la densidad marginal de $\boldsymbol{\theta}_1$ se reduce a

$$\widehat{p}(\boldsymbol{\theta}_1) \propto \sum_{i=1}^{d} \left|\widehat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta}_1)\right|^{\frac{1}{2}} p\left(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_{2_i}(\boldsymbol{\theta}_1)\right),$$

donde $\widehat{\boldsymbol{\Sigma}}_i(\boldsymbol{\theta}_1) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}\left(\widehat{\boldsymbol{\theta}}_{2_i}(\boldsymbol{\theta}_1)\right)$, con

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2) = -\left[\frac{\partial^2 \log h_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2' \partial \boldsymbol{\theta}_2}\right]^{-1}.$$

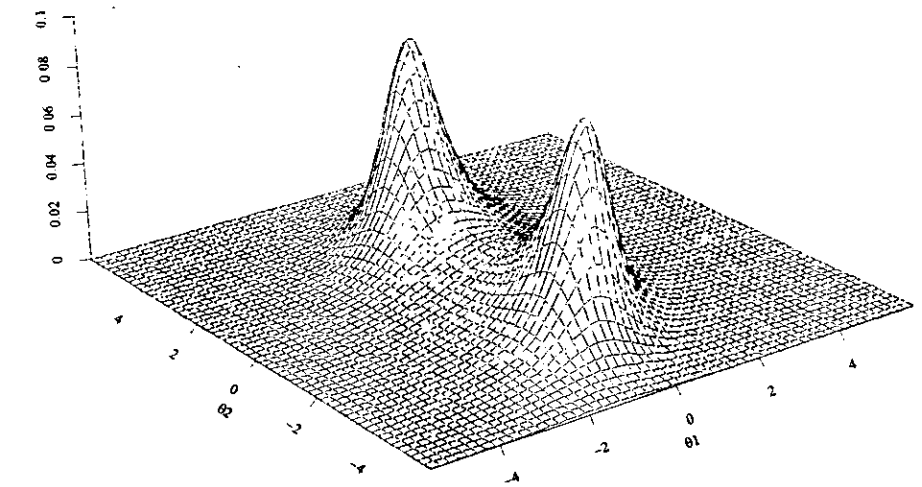
Ejemplo 3.2.2.

Sea $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ un vector aleatorio con distribución poly-t con parámetros $k = 2$, $r = 2$, $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$, $\mathbf{M}_1 = \mathbf{M}_2 = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}$ y $\nu_1 = \nu_2 = 2$.

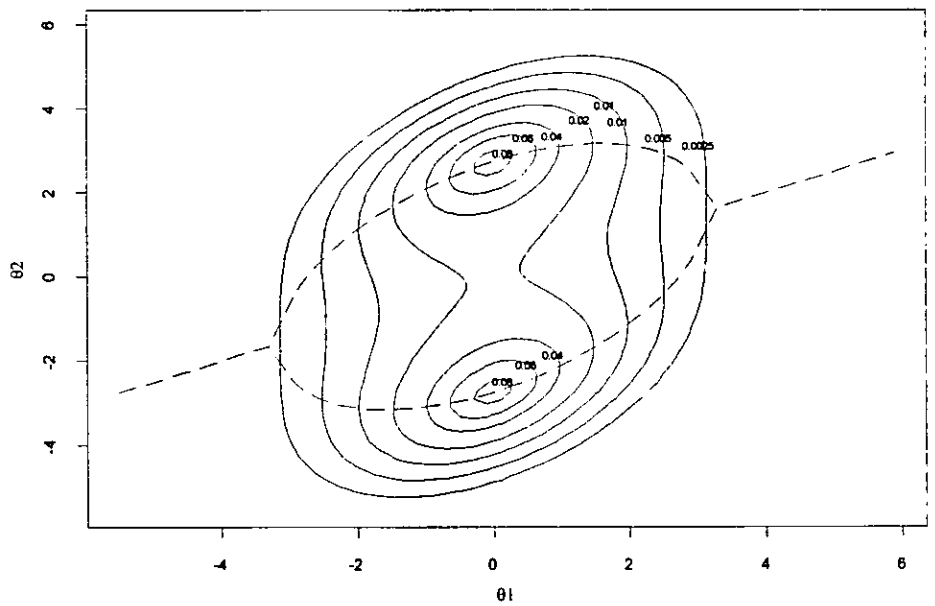
La gráfica en perspectiva de la distribución conjunta de $\boldsymbol{\theta}$ y las correspondientes curvas de nivel se presentan en las Figura 3.13. La constante de normalización se obtuvo numéricamente utilizando la regla trapezoidal.

Como se puede observar en las Figura 3.13, la distribución de $\boldsymbol{\theta}$ presenta un comportamiento bimodal. Aplicando la aproximación de Laplace multimodal básica (3.7) para el cálculo de la densidad marginal de θ_1 se tiene lo siguiente:

Sean $b_{x_1}(x_2) = 1$ y $h_{x_1}(x_2) = p(x_1, x_2)$. En la Figura 3.13(b) se puede observar que la densidad conjunta de $\boldsymbol{\theta}$ vista únicamente como función de θ_2 (para cada valor de θ_1) es bimodal para valores de θ_1 cercanos a cero y deja de ser bimodal para valores de θ_1 alejados del cero. Como no es posible encontrar analíticamente una función $\widehat{\theta}_2 = \widehat{\theta}_2(\theta_1)$ que maximice $h_{\theta_1}(\theta_2)$ para cada valor de θ_1, se aproximará



(a)



(b)

Figura 3.13: Densidad Poly-t bivariada (a) Perspectiva y (b) Curvas de nivel.

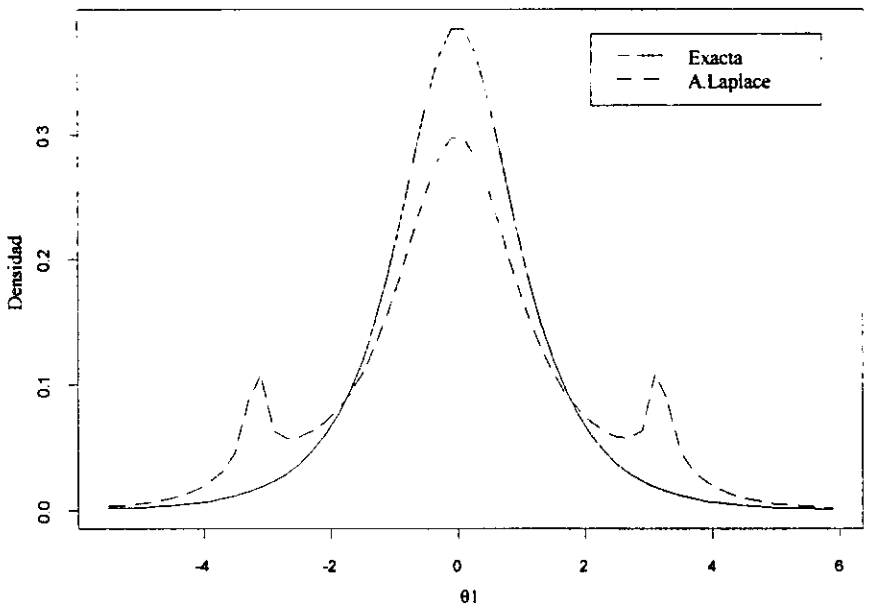
numéricamente para valores de θ_1 entre -5.5 y 5.9. De esta manera se podrá observar qué tan buena es la aproximación por el método de Laplace multimodal básica a la densidad marginal. Las líneas puntedas de la Figura 3.13(b) representan las modas de $h_{\theta_1}(\theta_2)$ para cada valor de θ_1.

En la Figura 3.14(a) se puede observar que la aproximación de Laplace multimodal básica a la distribución marginal de θ_1 no es adecuada debido a que presenta modas espurias (picos) en valores de θ_1 cercanos a los puntos donde la función $h_{\theta_1}(\theta_2)$ se hace unimodal. Esto se debe a que hay un aumento brusco en la varianza $\widehat{\Sigma}(\theta_1)$ para esos valores de θ_1.

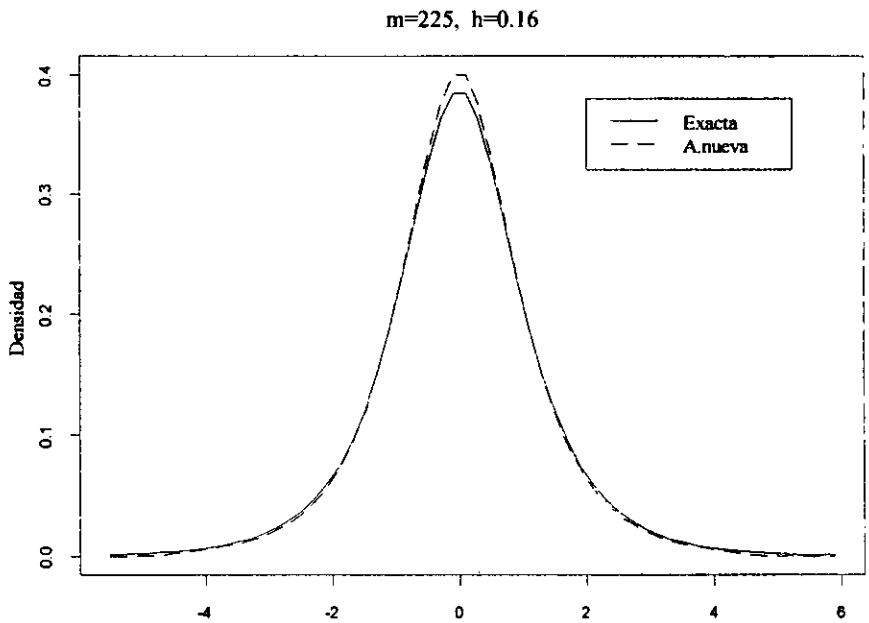
Para mostrar la capacidad de la aproximación propuesta en la Sección 3.1.2, la Figura 3.14(b) presenta tal aproximación a la distribución marginal de θ_1 con $m = 225$ puntos dstribuidos en una retícula regular sobre la región $(-4,-4) \times (4,4)$ y un valor $h = 0.16$. En esa figura se observa que la aproximación es bastante buena, captando muy bien la localización de la moda pero ligeramente con una mayor altura.

Las medidas de precisión (3.2) de las aproximaciónes Laplace multimodal básica y aproximación nueva a la densidad marginal de θ_1 son 0.0902 y 0.0163 respectivamente indicando que la nueva aproximación es más precisa. ■

En el ejemplo anterior se puede observar que la aproximación de Laplace multimodal básica para obtener distribuciones marginales no es muy buena en el caso de que la función $h_{\theta_1}(\theta_2)$ cambie en el número de modas al variar θ_1, debido a que la varianza aumenta bruscamente en la región donde ocurre el cambio. El mismo comportamiento se ha observado en otros ejemplos sin embargo, esto no siempre es así. A continuación se presenta un ejemplo en el cual el cambio en el número de modas



(a)



(b)

Figura 3.14: Densidad marginal de θ_1, (a) Aproximación de Laplace multimodal y (b) Nueva aproximación.

no afecta a la aproximación.

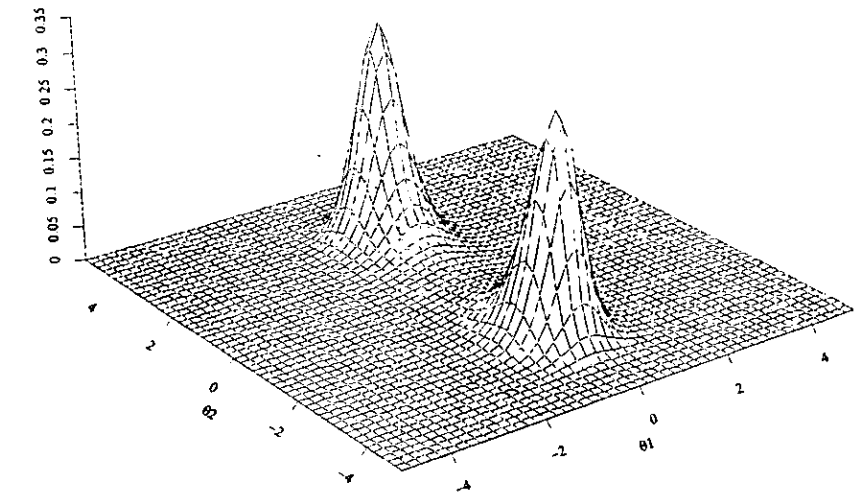
Ejemplo 3.2.3.

Sea $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ un vector aleatorio con distribución poly-t con parámetros $k = 2$, $r = 2$, $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$, $\mathrm{M}_1 = \mathrm{M}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$ y $\nu_1 = \nu_2 = 10$.

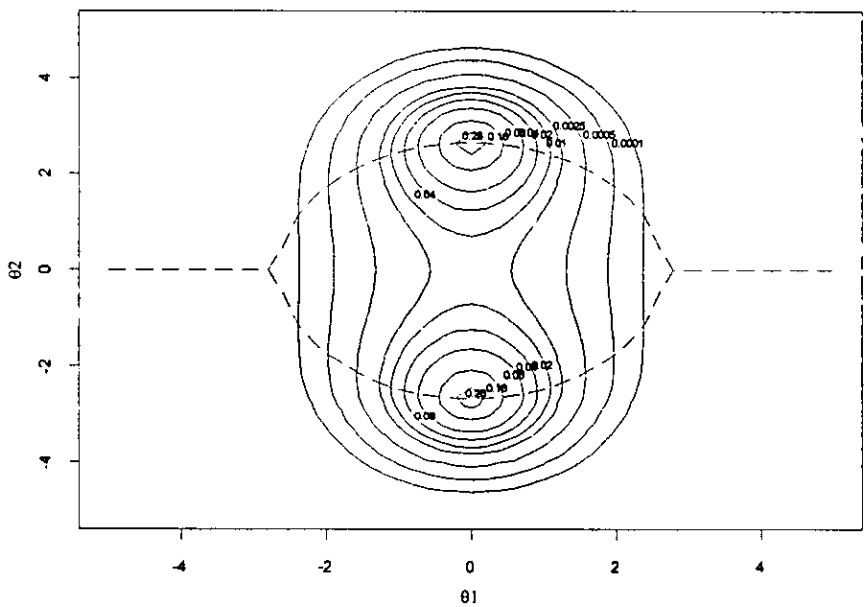
Como se puede observar en la Figura 3.15, la distribución de $\boldsymbol{\theta}$ presenta un comportamiento bimodal, con una varianza muy pequeña en cada moda lo que hace que las modas estén separadas. En la Figura 3.15(b) se presentan las curvas de nivel y la localización (obtenida numéricamente) de las modas $\widehat{\theta}_2 = \widehat{\theta}_2(\theta_1)$ para cada valor de θ_1.

Tomando $b_{\theta_1}(\theta_2) = 1$ y $h_{\theta_1}(\theta_2) = p(\theta_1, \theta_2)$ y aplicando la aproximación de Laplace multimodal básica para obtener la distribución marginal de θ_1, se obtiene la aproximación presentada en la Figura 3.16(a). En este caso la aproximación es muy buena, incluso en la gráfica no se aprecia ninguna diferencia con la original. Con este ejemplo podemos ver que el efecto de los picos alrededor de los puntos donde la función se hace unimodal no es tan pronunciado. Realmente tal efecto sigue existiendo, pero no se alcanza a apreciar gráficamente debido a que el cambio de bimodal a unimodal ocurre en áreas de muy baja densidad y no influye en la aproximación.

En este caso, la nueva aproximación con $m = 225$ puntos distribuidos en una retícula regular sobre la región $(-3, 3) \times (-3, 3)$ y un valor $h = 0.05$, para la distribución marginal de θ_1 se presenta en la Figura 3.16(b). En esa figura se observa que la nueva aproximación también es bastante buena, fallando ligeramente en la altura de la moda.

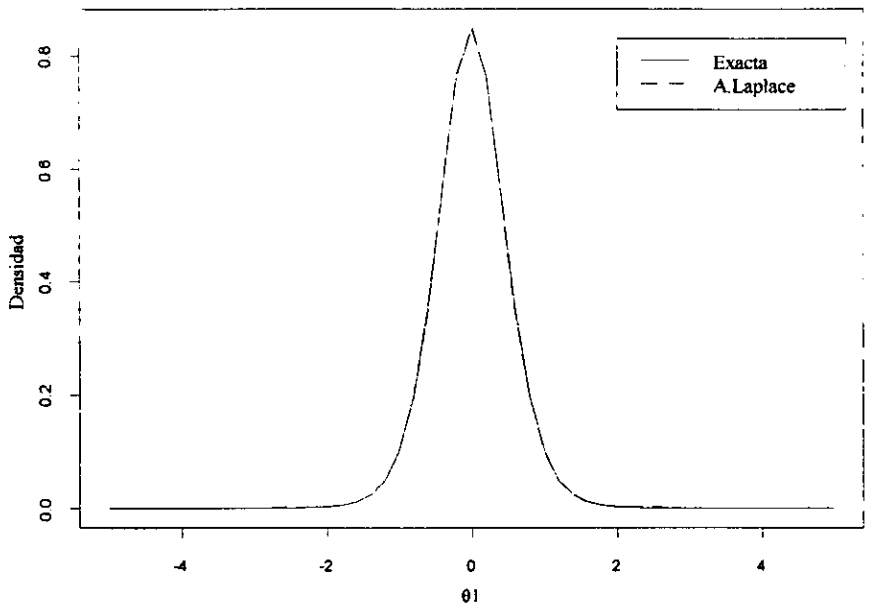


(a)

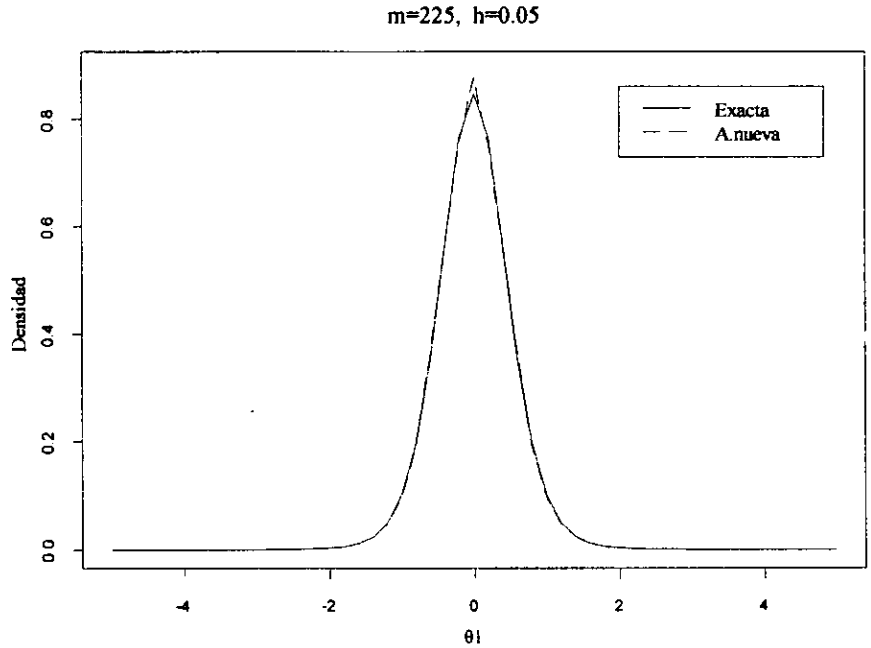


(b)

Figura 3.15: Densidad Poly-t bivariada (a) Perspectiva y (b) Curvas de nivel.



(a)



(b)

Figura 3.16: Densidad marginal de θ_1, (a) Aproximación de Laplace multimodal y (b) Nueva aproximación.

Las medidas de precisión (3.2) de las aproximaciónes Laplace multimodal básica y aproximación nueva a la densidad marginal de θ_1 son 0.0044 y 0.0304 respectivamente lo que indica que en esta caso la aproximación de Laplace multimodal es más precisa.

■

3.2.2 Laplace multimodal nueva.

Como se mencionó en la Sección 3.1.3, la nueva aproximación puede ser usada como una regla de integración, dejando a un lado el contexto estadístico de aproximación a densidades.

Supongamos que se desean calcular resúmenes inferenciales de la distribución final $p(\boldsymbol{\theta}|\mathbf{x})$ de $\boldsymbol{\theta}$, en particular supóngase que se desea calcular el valor esperado final de la función $g(\boldsymbol{\theta})$ mediante

$$E[g(\boldsymbol{\theta})|\mathbf{x}] = \frac{\int g(\boldsymbol{\theta}) p_x(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}}{\int p_x(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}},$$

donde $p_x(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$.

Supongamos que existen funciones $b_J(\cdot)$ y $h_J(\cdot)$, $J = N, D$ escogidas convenientemente, de manera que el valor esperado tome la forma

$$E[g(\boldsymbol{\theta})|\mathbf{x}] = \frac{\int b_N(\boldsymbol{\theta})\, h_N(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}}{\int b_D(\boldsymbol{\theta})\, h_D(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}}. \tag{3.10}$$

Por un lado, si se realiza una factorización en forma estándar, es decir, $h_N(\boldsymbol{\theta}) = h_D(\boldsymbol{\theta}) = p_x(\boldsymbol{\theta})$, $b_N(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ y $b_D(\boldsymbol{\theta}) = 1$, se obtiene que la aproximación de Laplace multimodal nueva al valor esperado final de $g(\boldsymbol{\theta})$, basada en la nueva aprox-

imación a densidades (3.3), toma la forma

$$\widetilde{E}\left[g(\boldsymbol{\theta})\,|\mathbf{x}\right] = \sum_{i=1}^{m} w_i g\left(\widetilde{\boldsymbol{\theta}}_i\right),$$

donde $\widetilde{\boldsymbol{\theta}}_i$, $i = 1, \ldots, m$ son puntos arbitrarios sobre la región de mayor densidad de $p_x(\boldsymbol{\theta})$, y los w_i, $i = 1, \ldots m$ son la solución al sistema de ecuaciones (3.4) utilizando como parámetro de suavizamiento el valor h^* de la ecuación (3.5).

Una justificación heurística de este resultado es la siguiente.

Aplicando la nueva aproximación con mezcla de densidades normales (3.3) a la función $p_x(\boldsymbol{\theta})$ se tiene que

$$h_N(\boldsymbol{\theta}) = h_N(\boldsymbol{\theta}) = p_x(\boldsymbol{\theta}) \approx \frac{1}{c}\sum_{i=1}^{m} w_i N_k\left(\boldsymbol{\theta}\,\middle|\,\widetilde{\boldsymbol{\theta}}_i, h\mathbf{I}_k\right),$$

y realizando la expansión en serie de Taylor (e ignorando los términos de orden mayor que uno) de la función $b_N(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ alrededor de cada uno de los puntos $\widetilde{\boldsymbol{\theta}}_i$ se obtiene

$$b_N(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \approx g\left(\widetilde{\boldsymbol{\theta}}_i\right) + \frac{\partial}{\partial \boldsymbol{\theta}} g\left(\widetilde{\boldsymbol{\theta}}_i\right)\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_i\right), \quad i = 1, \ldots, m$$

Como cada una de las componentes $w_i N_k\left(\boldsymbol{\theta}\,\middle|\,\widetilde{\boldsymbol{\theta}}_i, h\mathbf{I}_k\right)$ se puede ver como una aproximación a $p_x(\boldsymbol{\theta})$ en el punto $\widetilde{\boldsymbol{\theta}}_i$, se puede combinar cada una de las componentes con la correspondiente aproximación de la función $g(\boldsymbol{\theta})$ en cada punto $\widetilde{\boldsymbol{\theta}}_i$ obtiéndose que

$$b_N(\boldsymbol{\theta})\, h_N(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) p_x(\boldsymbol{\theta}) \approx \frac{1}{c}\sum_{i=1}^{m}\left[g\left(\widetilde{\boldsymbol{\theta}}_i\right) + \frac{\partial}{\partial \boldsymbol{\theta}} g\left(\widetilde{\boldsymbol{\theta}}_i\right)\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_i\right)\right] w_i N_k\left(\boldsymbol{\theta}\,\middle|\,\widetilde{\boldsymbol{\theta}}_i, h\mathbf{I}_k\right).$$

Integrando la expresión anterior se cancela el segundo término del corchete para obtener la siguiente aproximación para el numerador de (3.10)

$$\int b_N(\boldsymbol{\theta})\, h_N(\boldsymbol{\theta})\, \partial\boldsymbol{\theta} \approx \frac{1}{c}\sum_{i=1}^{m} g\left(\widetilde{\boldsymbol{\theta}}_i\right).$$

A su vez, el denominador de (3.10) se puede aproximar mediante

$$\int b_D(\boldsymbol{\theta})\, h_D(\boldsymbol{\theta})\, \partial\boldsymbol{\theta} \approx \int \frac{1}{c} \sum_{i=1}^{m} w_i N_k\left(\boldsymbol{\theta} \left| \widetilde{\boldsymbol{\theta}}_i, h\mathbf{I}_k \right.\right) \partial\boldsymbol{\theta} = \frac{1}{c}.$$

Finalmente, al realizar el cociente de ambas aproximaciones se obtiene la aproximación de Lapace multimodal nueva en forma estándar. ■

Por otra parte, si se utiliza una factorización en forma exponencial, es decir, $h_N(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) p_x(\boldsymbol{\theta})$, $h_D(\boldsymbol{\theta}) = p_x(\boldsymbol{\theta})$ y $b_N(\boldsymbol{\theta}) = b_D(\boldsymbol{\theta}) = 1$, entonces, aplicando la nueva aproximación a cada una de las funciones $h_J(\cdot)$, $J = N, D$ se obtiene que la aproximación al valor esperado final es

$$\widehat{E}\left[g(\boldsymbol{\theta})\right] = \frac{c_D}{c_N},$$

donde c_J, $J = N, D$ son las constantes de normalización obtenidas al resolver el sistema de ecuaciones lineales (3.4) al utilizar m puntos arbitrarios sobre la región de mayor densidad de $h_J(\cdot)$ y usando como parámetro de suavizamiento h_J^*, $J = N, D$ de la ecuación (3.5). Para poder realizar la aproximación en forma exponencial, es necesario que la función $g(\boldsymbol{\theta})$ sea no negativa para todo $\boldsymbol{\theta}$.

La justificación de este resultado se sigue directamente del desarrollo de la aproximación de Laplace multimodal nueva en forma estándar.

Ejemplo 3.2.4.

Considere el Ejemplo 3.1.6, donde para cuatro distribuciones poly-t univariadas se quiere aproximar la constante de normalización y los primeros 3 momentos.

Usando la matriz (3.4) para aproximar el valor de los momentos, con $n = 30$ y el valor h^* de la ecuación (3.5) se obtienen los sigiuentes errores relativos.

Tabla 3.2.2

x	Factorización	c^{-1}	$E(\theta)$	$E(\theta^2)$	$E(\theta^3)$	$\widetilde{\theta}_1, \widetilde{\theta}_m$
$(-5, 3)$	estándar	2.2%	0.0%	5.7%	6.0%	-10,8
	exponencial			0.8%		-15,12
$(-4, 1)$	estándar	2.0%	0.2%	6.2%	6.4%	-9,6
	exponencial			0.9%		-13,8
$(-2, 2)$	estándar	2.1%	0.0%	9.8%	0.0%	-6,6
	exponencial			5.4%		-14,14
$(-1.1, 1.1)$	estándar	2.4%	0.0%	15.2%	0.0%	-4,4
	exponencial			14.0%		-50,50

De la tabla anterior se puede observar que los errores relativos al usar la factorización en forma exponencial son menores que los correspondientes errores relativos al usar la aproximación estándar.

Comparando estos errores con los obtenidos por la aproximación de Laplace básica (ver Tabla 3.2.1) se observa que en su mayoría son menores, notándose en particular una gran dismi-nución en el error relativo de la aproximación a la constante de normalización. Ahora bien, al comparar estos mismos errores con los obtenidos por la nueva aproximación a la densidad correspondiente (ver Tabla 3.1.1), se observa que los errores relativos de la nueva aproximación son todos menores o iguales a los errores de la aproximación de Laplace nueva al usar la factorización estándar, mientras que los errores de la aproximación de Laplace nueva al usar la factorización exponencial siguen siendo menores que los respectivos errores obtenidos al usar la nueva

aproximación a la función de densidad.

3.2.3 Discusión.

La generalización directa de la aproximación de Laplace al caso en donde la función a integrar sea multimodal es la aproximación de Laplace básica (llamada así en este trabajo), teniendo como caso particular a la aproximación de Laplace (Sección 2.2) cuando la función a integrar sea unimodal.

La ventaja de la aproximación de Laplace (unimodal) para aproximar densidades marginales de forma analítica se pierde al generalizarla al caso multimodal debido a que se tienen funciones más complicadas que difícilmente se pueden trabajar analíticamente para obtener la "función moda" y lograr una aproximación analítica.

La aproximación de Laplace nueva, basada en la nueva aproximación con mezclas de densidades normales, en la mayoría de los ejemplos presentados en este trabajo logra mejores aproximaciones a la constante de normalización y a resúmenes inferenciales tales como los tres primeros momentos.

La aproximación de Laplace nueva también podría usarse para aproximar densidades marginales, pero la aproximación resultante sería únicamente numérica. En cambio, la nueva aproximación (Sección 3.1.2) se puede utilizar para obtener analíticamente una muy buena aproximación a densidades marginales utilizando mezclas de las correspondientes densidades normales marginales.

Capítulo 4

Aplicaciones.

Como se mencionó en la Sección 3.1.1, una de las aplicaciones estadísticas en donde surgen las distribuciones poly-t es en el análisis Bayesiano de modelos lineales. Zellner (1971) y Dréze (1977) presentan una descripción detallada de cómo surgen las distribuciones poly-t en el análisis Bayesiano de Regresión en el caso de heterocedasticidad. Una breve explicación, basada en Zellner (1971), es la siguiente:

Sea

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \\ \mathbf{Y}_2 &= \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \end{aligned} \tag{4.1}$$

un modelo de regresión lineal, donde

$\mathbf{Y}_i$ es un vector de observaciones de la variable dependiente de dimensión $n_i \times 1$,

$\mathbf{X}_i$ es una matriz de observaciones de k variables independientes de dimensión $n_i \times k$ y con rango k,

$\boldsymbol{\beta}$ es un vector de coeficientes de dimensión $k \times 1$ y

$\boldsymbol{\epsilon}_i$ un vector de errores aleatorios de dimensión $n_i \times 1$, $i = 1, 2$.

Supongamos que los errores $\boldsymbol{\epsilon}_1$ y $\boldsymbol{\epsilon}_2$ son independientes y tienen una distribución Normal con media cero. Los elementos de $\boldsymbol{\epsilon}_1$ tienen varianza común σ_1^2 y los elementos de $\boldsymbol{\epsilon}_2$ tienen varianza común σ_2^2, i.e, $\boldsymbol{\epsilon}_i \sim N_n\left(\mathbf{0}, \sigma_i^2 \mathbf{I}_n\right)$, $i = 1, 2$. Bajo estos supuestos, la verosimilitud está dada por

$$\begin{aligned} p\left(\mathbf{y} \,|\mathbf{X}, \boldsymbol{\beta}, \sigma\right) \;\propto\; & \sigma_1^{-n_1} \exp\left\{-\frac{1}{2\sigma_1^2}\left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)'\left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)\right\} \\ & * \sigma_2^{-n_2} \exp\left\{-\frac{1}{2\sigma_2^2}\left(\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\right)'\left(\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\right)\right\}. \end{aligned}$$

Considere la distribución inicial no informativa para $\boldsymbol{\beta}$, σ_1 y σ_2 representada por

$$p\left(\boldsymbol{\beta}, \sigma_1, \sigma_2\right) \propto \frac{1}{\sigma_1 \sigma_2},$$

con $0 < \sigma_1 < \infty$, $0 < \sigma_2 < \infty$ y $-\infty < \beta_i < \infty$, $i = 1, \ldots, k$.

Combinando la distribución inicial con la verosimilitud se obtiene que la distribución final es

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma_1, \sigma_2 \,|\mathbf{X}, \mathbf{y}\right) \;\propto\; & \left(\sigma_1^{n_1+1}\right)^{-1} \exp\left\{-\frac{1}{2\sigma_1^2}\left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)'\left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)\right\} \\ & * \left(\sigma_2^{n_2+1}\right)^{-1} \exp\left\{-\frac{1}{2\sigma_2^2}\left(\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\right)'\left(\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\right)\right\}. \end{aligned}$$

Integrando sobre σ_1 y σ_2 se obtiene la distribución marginal para $\boldsymbol{\beta}$:

$$p\left(\boldsymbol{\beta} \,|\mathbf{X}, \mathbf{y}\right) \propto \left[\left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)'\left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)\right]^{-\frac{n_1}{2}} \left[\left(\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\right)'\left(\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\right)\right]^{-\frac{n_2}{2}}$$

en otras palabras, esta distribución se puede escribir como:

$$\begin{aligned} p\left(\boldsymbol{\beta} \,|\mathbf{X}, \mathbf{y}\right) \;\propto\; & \left[1 + \frac{1}{\nu_1 S_1^2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1\right)' \mathbf{Z}_1 \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1\right)\right]^{-\frac{1}{2}(\nu_1 + k)} \\ & * \left[1 + \frac{1}{\nu_2 S_2^2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_2\right)' \mathbf{Z}_2 \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_2\right)\right]^{-\frac{1}{2}(\nu_2 + k)} \end{aligned} \tag{4.2}$$

donde $\nu_i = n_i - k$, $\mathbf{Z}_i = \frac{\mathbf{X}_i'\mathbf{X}_i}{S_i^2}$, $\widehat{\boldsymbol{\beta}}_i = \mathbf{Z}_i^{-1}\mathbf{X}_i'\mathbf{y}_i$ y $\nu_i S_i^2 = \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i\right)'\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i\right)$, $i = 1, 2$.

Se puede observar que la distribución (4.2) es una distribución poly-t con $r = 2$.

Para ejemplificar el uso de las distribuciones poly-t en el análisis Bayesiano de regresión se presentan a continuación dos ejemplos, uno con datos simulados y otro con datos reales.

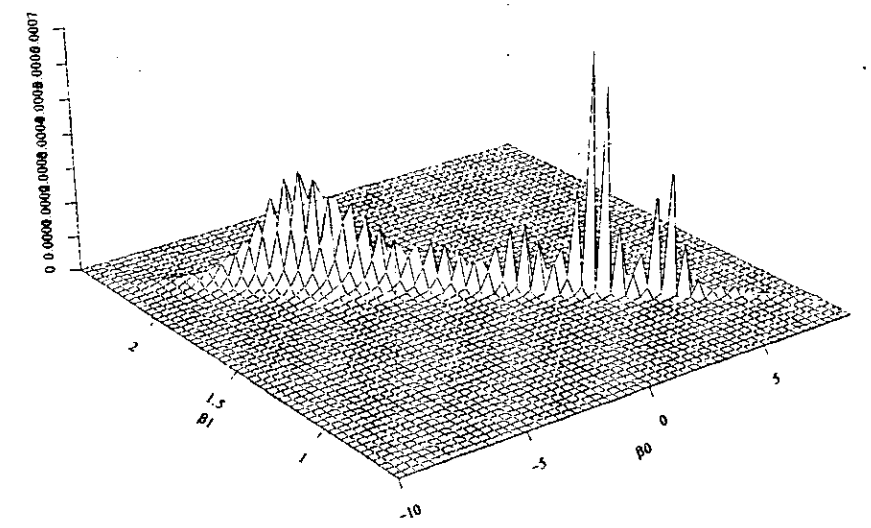
4.1 Ejemplo con datos simulados.

Sean $\mathbf{Y}_1$ y $\mathbf{Y}_2$ dos muestras aleatorias que cumplen con los supuestos del modelo (4.1).

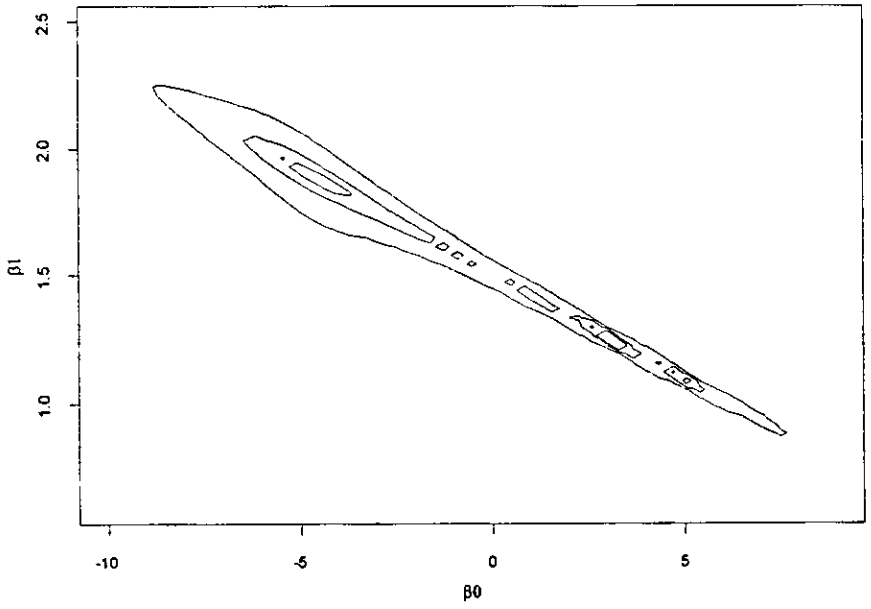
Supongamos que $\mathbf{X}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$, $\mathbf{X}_2 = \begin{pmatrix} 1 & 11 \\ 1 & 12 \\ 1 & 13 \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} -5 \\ 2 \end{pmatrix}$, $\sigma_1^2 = 5$ y $\sigma_2^2 = 15$.

Para $n_1 = n_2 = 3$ se obtuvieron los siguientes valores simulados de $\mathbf{y}_1$ y $\mathbf{y}_2$: $\mathbf{y}_1 = (-2.091, -0.153, -1.592)'$ y $\mathbf{y}_2 = (17.002, 17.999, 19.357)'$.

Por el análisis presentado al inicio de este capítulo, sabemos que la distribución final marginal del vector $\boldsymbol{\beta}$ al usar una distribución inicial no informativa es una distribución poly-t de la forma (4.2) con $\nu_1 = \nu_2 = 1$, $\widehat{\boldsymbol{\beta}}_1 = \begin{pmatrix} -1.77 \\ 0.24 \end{pmatrix}$, $\widehat{\boldsymbol{\beta}}_2 = \begin{pmatrix} 3.99 \\ 1.17 \end{pmatrix}$, $\mathbf{Z}_1 = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}$, $\mathbf{Z}_2 = \begin{pmatrix} 3 & 36 \\ 36 & 434 \end{pmatrix}$, $\nu_1 S_1^2 = 1.9011$, $\nu_2 S_2^2 = 0.0218$ y $k = 2$.



(a)



(b)

Figura 4.1: Kernel de la distribución final de $\boldsymbol{\beta}$. (a) Perspectiva y (b) Curvas de nivel.

La forma de la distribución final de $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ se puede observar en la Figura 4.1 (los picos se deben a que los puntos de la rejilla con la que se realizó la gráfica no están lo suficientemente cercanos entre sí). De la gráfica en perspectiva y las curvas de nivel es claro el comportamiento bimodal de esta distribución. Para obtener cualquier resumen inferencial de esta distribución final y apreciar el uso de la distribución poly-t como forma de atacar problemas de heterocedasticidad, es necesario recurrir a aproximaciones. Zellner (1971) utiliza una aproximación normal para obtener resúmenes inferenciales de la distribución poly-t, pero por lo observado en la Figura 4.1, la aproximación normal no sería una aproximación adecuada. Apliquemos entonces la aproximación (3.3) propuesta en esta tesis.

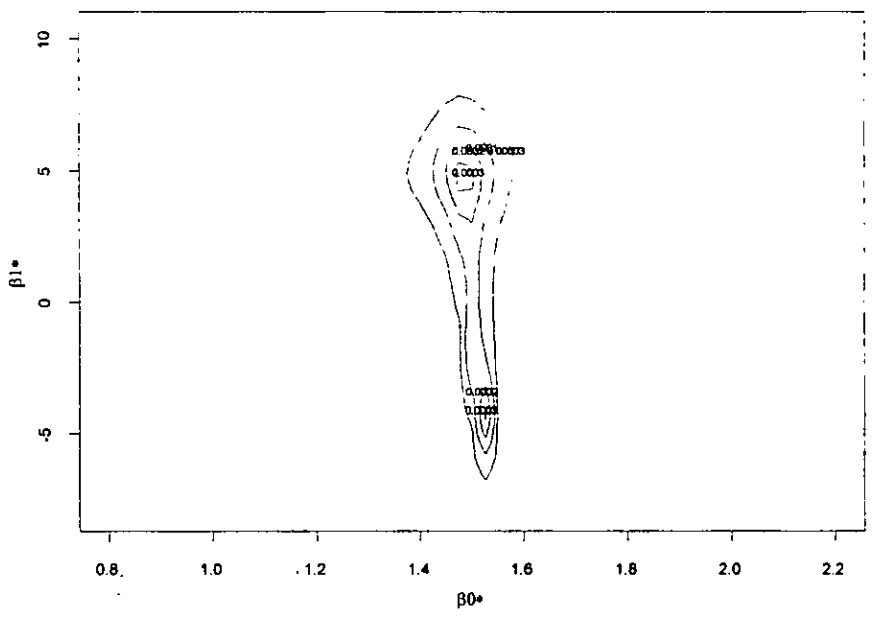
En la Figura 4.1 se puede observar que las escalas para β_0 y para β_1 son distintas. Como se mencionó en la Sección 3.1.3, es conveniente tener la misma escala para los dos parámetros. Si se realizara una transformación lineal sobre el parámetro β_1 para tener la misma escala que β_0, se tendría una correlación muy grande entre los parámetros y esto también dificultaría la aplicación de la aproximación (3.3).

Primeramente, para que la nueva aproximación se pueda aplicar eficazmente es necesario realizar una rotación que corrija el problema de la correlación entre β_0 y β_1.

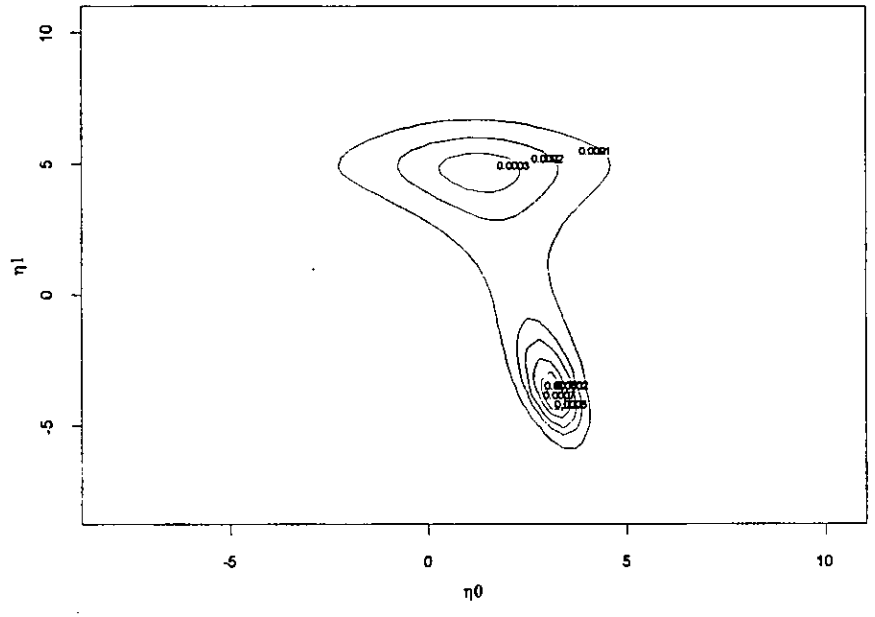
Sea $\gamma = 1.4846$ el ángulo de rotación para la transformación lineal

$$\begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} = \begin{pmatrix} \cos\gamma & \text{sen}\gamma \\ -\text{sen}\gamma & \cos\gamma \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

En la Figura 4.2(a) se presenta las curvas de nivel para la distribución de $\boldsymbol{\beta}^*$. En ella se puede observar que la correlación entre β_0^* y β_1^* es mucho menor que en el caso anterior. Hasta este momento tampoco es posible aplicar el método propuesto debido



(a)



(b)

Figura 4.2: Curvas de nivel de las densidades transformadas. (a) $\boldsymbol{\beta}^*$ y (b) $\boldsymbol{\eta}$.

a que la escala es muy distinta para los dos parámetros aún transformados.

En segundo lugar, únicamente es necesario realizar un cambio de escala en el parámetro β_1^*, de la siguiente manera: sea

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 60 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} - \begin{pmatrix} 87.8 \\ 0 \end{pmatrix}.$$

Esta transformación lineal permite tener a los dos parámetros en la misma escala. La Figura 4.2(b), que presenta las curvas de nivel del nuevo vector de parámetros $\boldsymbol{\eta}$, muestra que la escala para η_0 y η_1 es la misma, de -5 a 7.6.

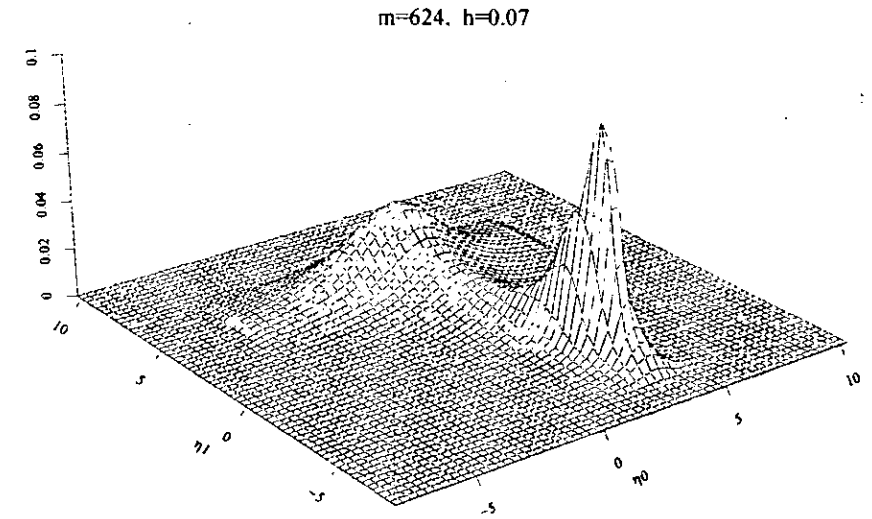
Es conveniente aclarar que, por ser lineales, las transformaciones que se han realizado tienen un Jacobiano constante, de manera que la gráfica de la Figura 4.2(b) representa el kernel de la distribución de $\boldsymbol{\eta}$.

Finalmente, aplicando la aproximación (3.3) al kernel de la distribución de $\boldsymbol{\eta}$ con base en $m = 624$ puntos en una retícula regular sobre el rectángulo $(-4.8, 6.8) \times (-5.9, 6.8)$ y un parámetro de suavizamiento $h = 0.07$ se obtienen las gráficas de la Figura 4.3. En esta figura se observa perfectamente el comportamiento bimodal de la distribución de $\boldsymbol{\eta}$. Más aún, al comparar las curvas de nivel de la aproximación (Figura 4.3(b)) con las de la Figura 4.2(b) se observa que se reproduce adecuadamente la estructura de correlación y la localización de las modas.

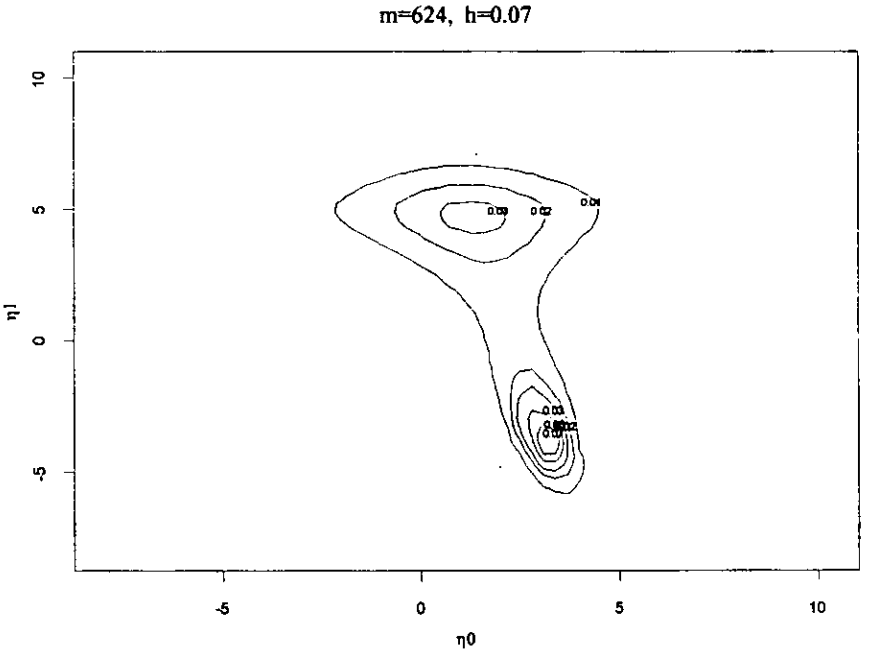
Por lo tanto la distribución final para $\boldsymbol{\eta}$ se puede aproximar mediante

$$\widehat{p}(\boldsymbol{\eta} \,|\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{m} w_i N_2 (\boldsymbol{\eta} \,|\widetilde{\boldsymbol{\eta}}_i, h\mathbf{I}_2),$$

donde w_i, $i = 1, \ldots, m$ son la solución al sistema de ecuaciones (3.4).



(a)



(b)

Figura 4.3: Nueva aproximación a la distribución de $\boldsymbol{\eta}$. (a) Perspectiva y (b) Curvas de nivel.

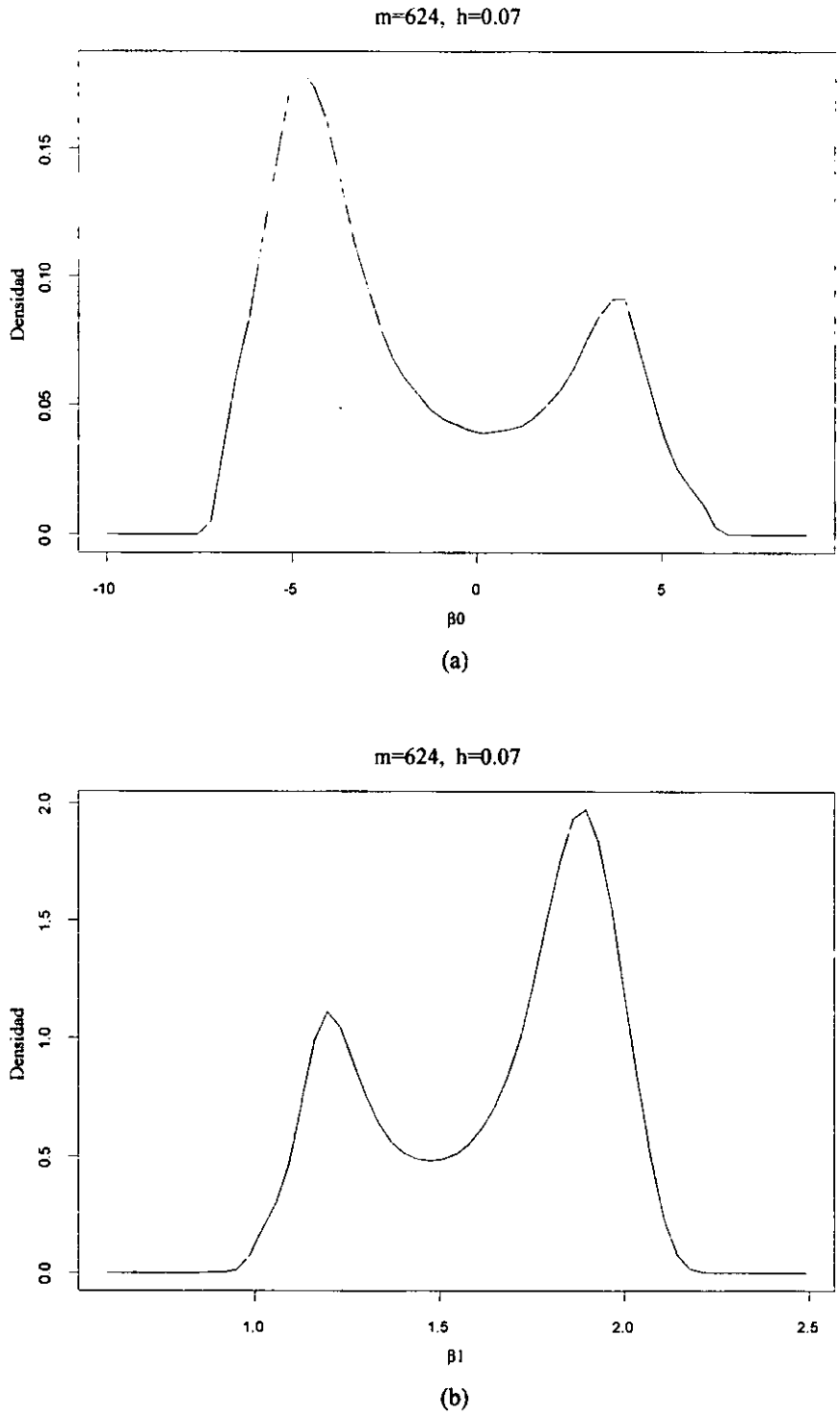


Figura 4.4: Densidades marginales. (a) β_0 y (b) β_1.

Al combinar las dos transformaciones lineales en una sola se tiene que

$$\boldsymbol{\eta} = \mathbf{D}\boldsymbol{\beta} - \mathbf{d},$$

donde $\mathbf{D} = \begin{pmatrix} 5.162 & 59.77 \\ 0.996 & 0.086 \end{pmatrix}$ y $\mathbf{d} = \begin{pmatrix} 87.8 \\ 0 \end{pmatrix}$.

Invirtiendo la tranformación se tiene que $\boldsymbol{\beta} = \mathbf{D}^{-1}\boldsymbol{\eta} + \mathbf{D}^{-1}\mathbf{d}$ y el Jacobiano de la transformación es $|\mathbf{J}| = |\det(\mathbf{D})| = 60$. Finalmente, realizando un poco de álgebra se obtiene que la distribución final para $\boldsymbol{\beta}$ se puede aproximar adecuadamente mediante

$$\widehat{p}(\boldsymbol{\beta}\,|\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{m} w_i N_2\left(\boldsymbol{\beta}\,\middle|\,\mathbf{D}^{-1}(\widetilde{\boldsymbol{\eta}}_i + \mathbf{d}), h(\mathbf{C}'\mathbf{C})^{-1}\right). \tag{4.3}$$

Las gráficas de la aproximación a las correspodientes distribuciones marginales de β_0 y β_1 se presentan en la Figura 4.4. En esta figura se puede observar que ambas distribuciones marginales son bimodales, como podría esperarse al observar el kernel de la distribución conjunta (Figura 4.1).

En ambos casos, el verdadero valor del parámetro se localiza en una zona de alta densidad y es cercano a la mayor de las modas.

4.2 Ejemplo con datos reales.

Este ejemplo fue tomado de Neter, Wasserman y Kutner (1989; Sección 11.8).

Un investigador de salud está interesado en estudiar la relación entre la Presión diastólica arterial (Y) y la Edad (X_1) en mujeres adultas que gozan de buena salud, entre los 20 y 40 años de edad. Para ello, recolectó datos de 54 mujeres con las características requeridas, los datos se presentan en la Tabla 4.2.1.

De acuerdo con Neter *et al.* (1989), el diagrama de dispersión de estos datos (Figura 4.5(a)) sugiere una relación lineal entre la presión diastólica y la edad. Por lo tanto, un modelo que podría representar el comportamiento de los datos es:

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i, \quad i = 1, \ldots n.$$

En forma matricial el modelo se puede escribir como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

donde $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ es un vector de observaciones, $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1)$ es una matriz de dimensión $n \times k$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ es un vector de coeficientes de dimensión $k \times 1$ y $\boldsymbol{\varepsilon} \sim N_2(\mathbf{0}, \sigma^2\mathbf{I})$ es un vector de errores de dimensión $n \times 1$, con $n = 54$ y $k = 2$.

En la Figura 4.5(a) también se puede observar un incremento en la varianza al aumentar la edad. Para verificar el comportamiento de la varianza, se llevó a cabo un análisis preliminar con el objeto de observar el comportamiento de los residuos. Se realizó un ajuste por mínimos cuadrados, que en el contexto Bayesiano coincide con el análisis de regresión con errores Normales tomando una distribución inicial no informativa y considerando una función de pérdida cuadrática (Zellner, 1971).

Tabla 4.2.1. Datos de Presión diastólica arterial.

Sujeto	Edad	Presión	Sujeto	Edad	Presión	Sujeto	Edad	Presión
i	X_{1_i}	Y_i	i	X_{1_i}	Y_i	i	X_{1_i}	Y_i
1	27	73	19	37	78	37	42	85
2	21	66	20	38	87	38	44	71
3	22	63	21	33	76	39	46	80
4	26	79	22	35	79	40	47	96
5	25	68	23	30	73	41	45	92
6	28	67	24	37	68	42	55	76
7	24	75	25	31	80	43	54	71
8	25	71	26	39	75	44	57	99
9	23	70	27	46	89	45	52	86
10	20	65	28	49	101	46	53	79
11	29	79	29	40	70	47	56	92
12	24	72	30	42	72	48	52	85
13	20	70	31	43	80	49	57	109
14	38	91	32	46	83	50	50	71
15	32	76	33	43	75	51	59	90
16	33	69	34	49	80	52	50	91
17	31	66	35	40	90	53	52	100
18	34	73	36	48	70	54	58	80

El diagrama de dispersión de los residuos *versus* edad (X_1) se presenta en la Figura 4.5(b), la cual confirma el hecho de que la varianza es no constante.

Con el propósito de explorar si la varianza del error tiene una relación simple con la edad, Neter *et al.* (1989) dividen los casos en cuatro grupos de aproximadamente el mismo tamaño de acuerdo a la edad. Entonces, basado en los residuos del análisis preliminar estiman la varianza dentro de cada grupo. Los cuatro grupos de edad junto con su varianza estimada se presentan en la Tabla 4.2.2.

Tabla 4.2.2.

Grupo (i)	Edad (X_i)	n_i	$\widehat{\sigma}_i^2$	Ponderación (γ_i)
1	$[20, 30)$	13	17.74	0.056
2	$[30, 40)$	13	42.13	0.023
3	$[40, 50)$	15	87.93	0.011
4	$[50, 60)$	13	124.14	0.008

Se consideró que las varianzas estimadas $(\widehat{\sigma}_i^2)$ tenían una diferencia importante, por lo que se decidió realizar un análisis con mínimos cuadrados ponderados (MCP) tomando como ponderación para cada grupo de edad el inverso de la varianza estimada (γ_i). Al final de esta sección se presenta una comparación entre las estimaciones obtenidas por mínimos cuadrados ordinarios (MCO) y por MCP.

Por otro lado, como se mencionó al inicio de este capítulo, las distribuciones poly-t permiten atacar el problema de heterocedasticidad en el análisis Bayesiano de regresión. Generalizando el modelo (4.1) al caso en el que se tienen cuatro grupos de datos con varianza distinta en cada grupo, y suponiendo una distribución inicial no

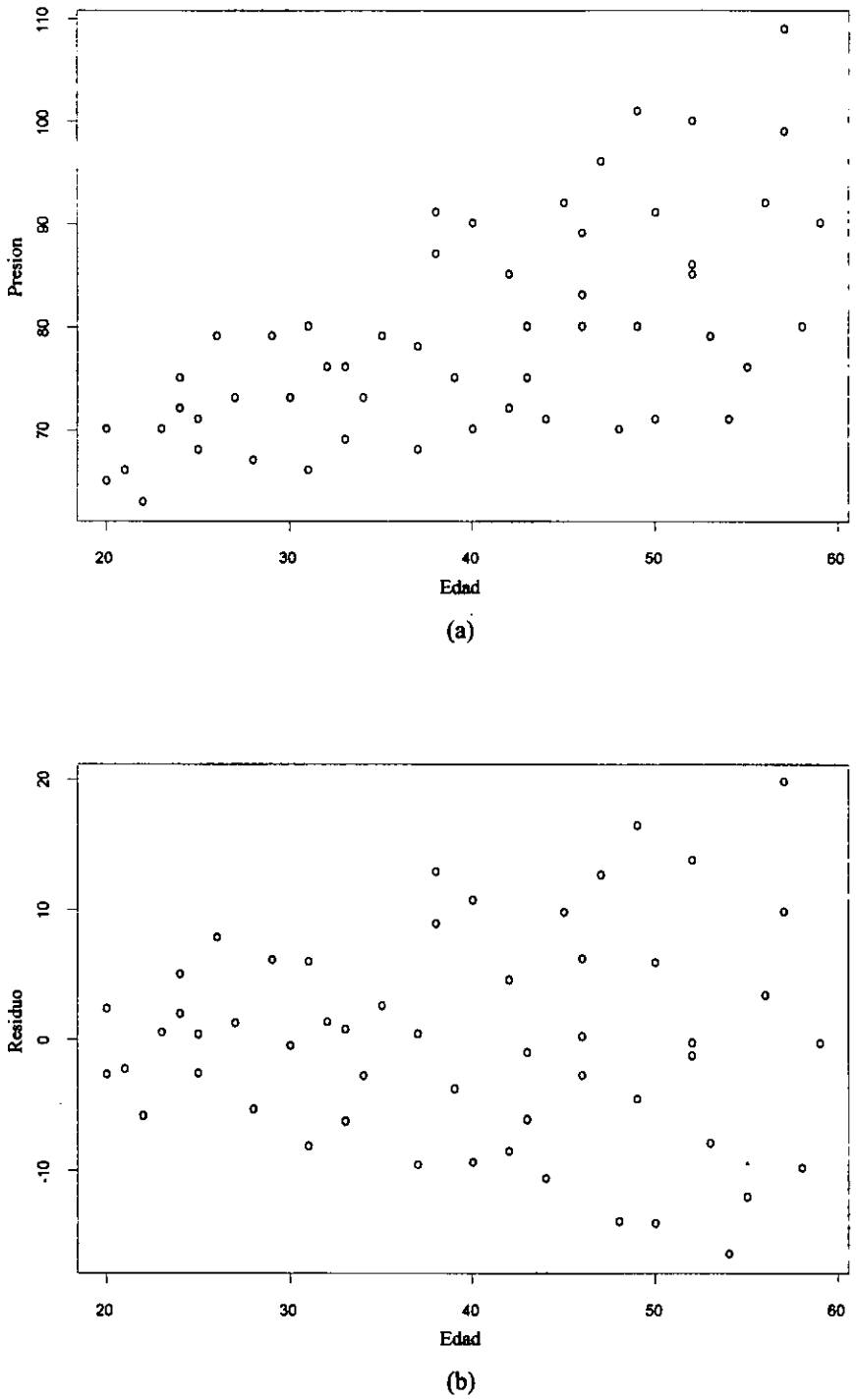


Figura 4.5: Diagramas de dispersión. (a) Presión *vs.* Edad y (b) Residuos *vs.* Edad.

informativa, se obtiene que la distribución final marginal para $\boldsymbol{\beta}$ está dada por

$$p(\boldsymbol{\beta}\,|\mathbf{X},\mathbf{y}) \propto \prod_{i=1}^{4}\left[1+\frac{1}{\nu_i S_i^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_i\right)'\mathbf{Z}_i\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_i\right)\right]^{-\frac{1}{2}(\nu_i+k)}, \qquad (4.4)$$

donde ν_i, $\mathbf{Z}_i$, $\widehat{\boldsymbol{\beta}}_i$ y $\nu_i S_i^2$, $i=1,\ldots,4$ están definidos como en (4.2).

Utilizando los datos de la Tabla 4.2.1 y los cuatro grupos de la Tabla 4.2.2 se obtiene que

$$\widehat{\boldsymbol{\beta}}_1 = \begin{pmatrix} 45.96 \\ 1.02 \end{pmatrix}, \widehat{\boldsymbol{\beta}}_2 = \begin{pmatrix} 39.82 \\ 1.05 \end{pmatrix}, \widehat{\boldsymbol{\beta}}_3 = \begin{pmatrix} 32.52 \\ 1.11 \end{pmatrix}, \widehat{\boldsymbol{\beta}}_4 = \begin{pmatrix} 24.52 \\ 1.14 \end{pmatrix},$$

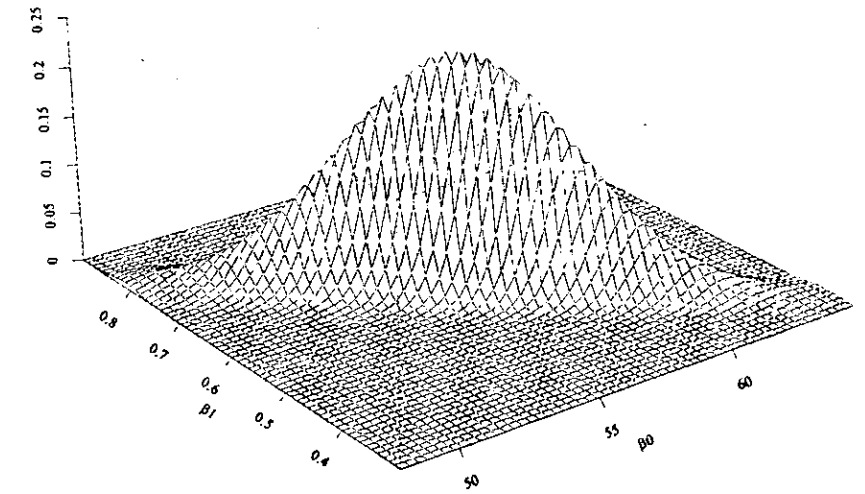
$$\mathbf{Z}_1 = \begin{pmatrix} 13 & 314 \\ 314 & 7686 \end{pmatrix}, \quad \mathbf{Z}_2 = \begin{pmatrix} 13 & 448 \\ 448 & 15552 \end{pmatrix},$$

$$\mathbf{Z}_3 = \begin{pmatrix} 15 & 670 \\ 670 & 30050 \end{pmatrix}, \quad \mathbf{Z}_4 = \begin{pmatrix} 13 & 705 \\ 705 & 38341 \end{pmatrix},$$

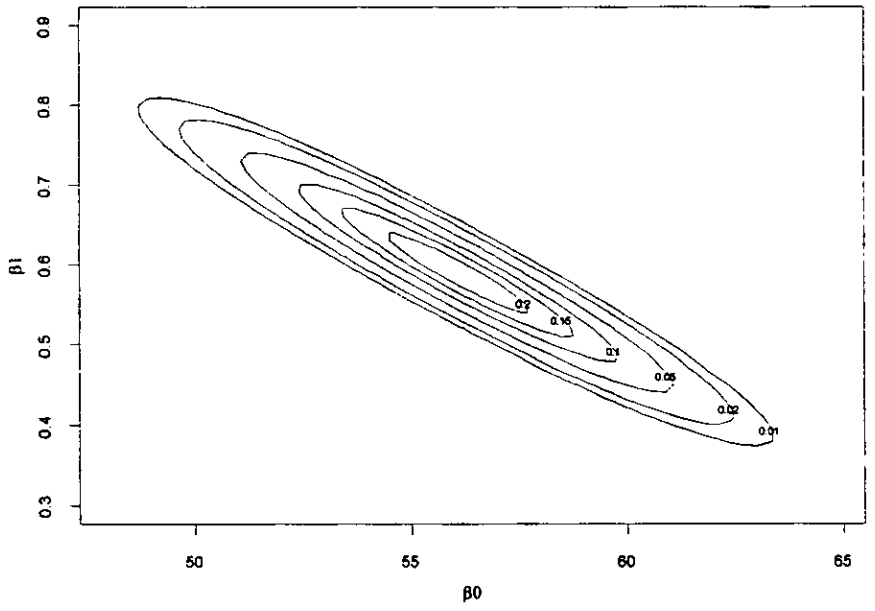
$\nu_1 S_1^2 = 193.18$, $\nu_2 S_2^2 = 479.94$, $\nu_3 S_3^2 = 1196.01$, $\nu_4 S_4^2 = 1454.66$, $k=2$ y $\nu_i = 1$, $i=1,\ldots,4$.

El comportamiento de esta distribución se puede observar en la Figura 4.6. La gráfica en perspectiva y las curvas de nivel indican que la distribución final de $\boldsymbol{\beta}$ tiene únicamente una sola moda.

Como se mencionó en la Sección 3.1.1, la distribución poly-t no es fácil de manejar analíticamente, por lo que es necesario recurrir a aproximaciones para obtener cualquier resumen inferencial de la distribución final de $\boldsymbol{\beta}$. En este caso, para obtener resúmenes inferenciales aproximados es factible aplicar la distribución normal asintótica, Zellner (1971) obtiene que la distribución normal asintótica a la distribución



(a)



(b)

Figura 4.6: Kernel de la distribución final de $\boldsymbol{\beta}$. (a) Prespectiva y (b) Curvas de nivel.

(4.4) es

$$p(\boldsymbol{\beta}\,|\mathbf{X},\mathbf{y}) = N_2\left(\boldsymbol{\beta}\,\middle|\widetilde{\boldsymbol{\beta}},\mathbf{V}\right), \tag{4.5}$$

con

$$\mathbf{V} = (\mathbf{M}_1+\mathbf{M}_2+\mathbf{M}_3+\mathbf{M}_4)^{-1} \text{ y } \widetilde{\boldsymbol{\beta}} = \mathbf{V}\left(\mathbf{M}_1\widehat{\boldsymbol{\beta}}_1+\mathbf{M}_2\widehat{\boldsymbol{\beta}}_2\right),$$

donde $\mathbf{M}_i = \frac{\mathbf{Z}_i}{S_i^2} = \frac{\mathbf{X}_i'\mathbf{X}_i}{S_i^2}$, $i = 1,\ldots,4$.

Para este ejemplo, la aproximación normal asintótica a la distribución final de $\boldsymbol{\beta}$ es de la forma (4.5) con $\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} 56.074 \\ 0.5903 \end{pmatrix}$ y $\mathbf{V} = \begin{pmatrix} 8.2149 & -0.2373 \\ -0.2373 & 0.0075 \end{pmatrix}$. Cabe señalar que, a pesar de que la distribución final de $\boldsymbol{\beta}$ es unimodal y relativamente simétrica, la aproximación normal asintótica no es necesariamente adecuada en este caso ya que la distribución poly-t tiene las colas más pesadas que la normal.

Una nueva forma de aproximar resúmenes inferenciales de la distribución final de $\boldsymbol{\beta}$ es mediante la aproximación propuesta en la Sección 3.1.2 de esta tesis.

Para poder aplicar la aproximación (3.3) es necesario que la escala de β_0 y β_1 sea similar, y para ello se realiza la siguiente transformación lineal:

$$\boldsymbol{\eta} = \mathbf{D}\boldsymbol{\beta} - \mathbf{d},$$

donde $\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 33.3 \end{pmatrix}$ y $\mathbf{d} = \begin{pmatrix} 0 \\ 2.6 \end{pmatrix}$.

Sean $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_m$, con $m = 625$, una colección de puntos en una retícula regular sobre el cuadrado $(48,64)\times(9,25)$. Utilizando un parámetro de suavizamiento $h = 0.15$ se obtiene que la aproximación (3.3) a la distribución de $\boldsymbol{\eta}$ es

$$\widehat{p}(\boldsymbol{\eta}\,|\mathbf{X},\mathbf{y}) = \sum_{i=1}^{m} w_i N_2\left(\boldsymbol{\eta}\,|\widetilde{\boldsymbol{\eta}}_i, h\mathbf{I}_2\right), \tag{4.6}$$

donde w_i, $i = 1, \ldots, m$ son la solución al sistema de ecuaciones (3.4).

Expresando esta aproximación en términos de $\boldsymbol{\beta}$ se tiene que

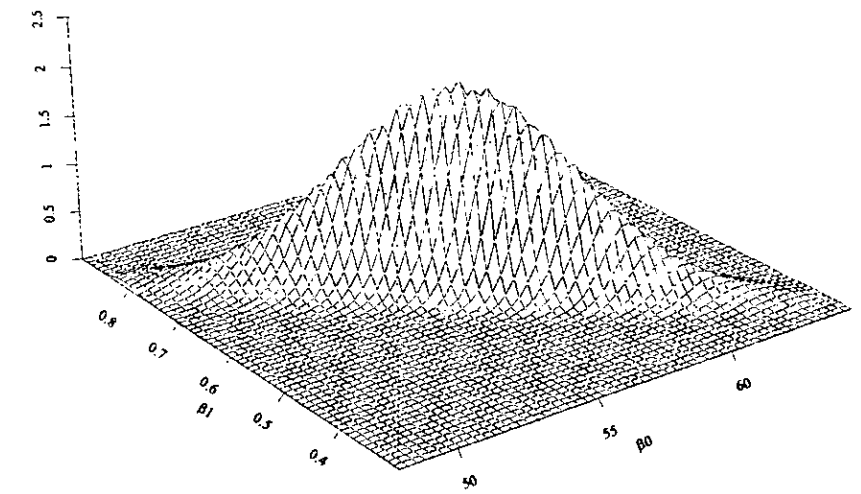
$$\widehat{p}\left(\boldsymbol{\beta}\,|\mathbf{X},\mathbf{y}\right) = \sum_{i=1}^{m} w_i N_2\left(\boldsymbol{\beta}\,\middle|\mathbf{D}^{-1}\left(\widetilde{\boldsymbol{\eta}}_i + \mathbf{d}\right), h\left(\mathbf{C}'\mathbf{C}\right)^{-1}\right).$$

En la Figura 4.7 se puede observar que tanto la gráfica en perspectiva (a) como las curvas de nivel (b) son casi idénticas a las correspondientes gráficas del kernel de la verdadera distribución de $\boldsymbol{\beta}$ (Figura 4.6). Es posible, debido a la flexibilidad de las mezclas de normales obtener aproximaciones a las distribuciones marginales de β_0 y de β_1 a partir de la aproximación conjunta (4.6). En la Figura 4.8 se encuentran estas distribuciones marginales, y en ella se puede observar la unimodalidad de ambas distibuciones.

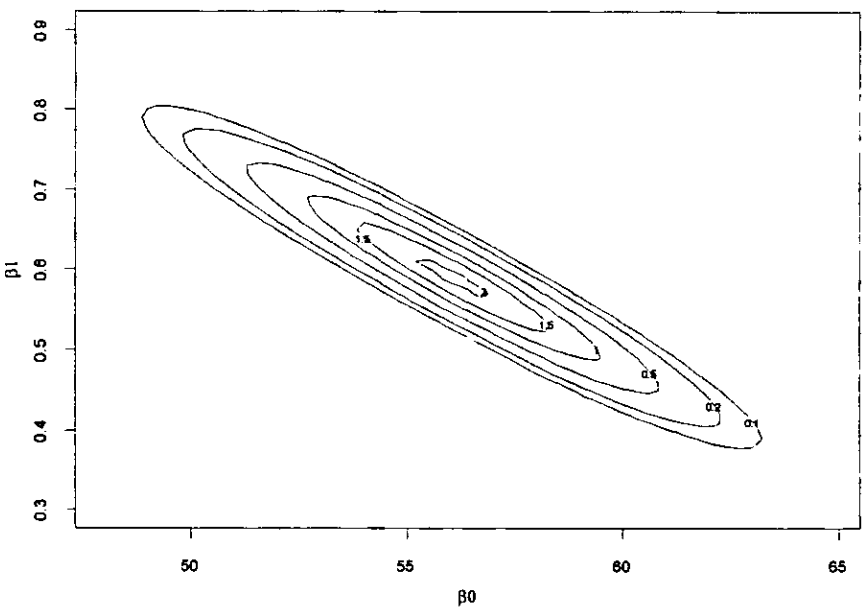
Si se desean hacer estimaciones puntuales sobre los parámetros de la regresión, al suponer una función de pérdida cuadrática se obtiene que el estimador puntual es el valor esperado con respecto a la distribución final de $\boldsymbol{\beta}$ (Bernardo y Smith, 1994; Sección 5.1.5). En este caso, las aproximaciones a los valores esperados finales de β_0 y β_1 usando la aproximación (4.6) están dadas por

$$\widehat{E}\left(\beta_0\,|\mathbf{X},\mathbf{y}\right) = 56.0174 \quad \text{y} \quad \widehat{E}\left(\beta_1\,|\mathbf{X},\mathbf{y}\right) = 0.5915.$$

Finalmente, la Tabla 4.2.3 presenta una comparación entre tres distintos ajustes del modelo: mínimos cuadrados ordinarios (MCO), mínimos cuadrados ponderados (MCP) y el análisis Bayesiano de heterocedasticidad (ABH) discutido al inicio de este capítulo.

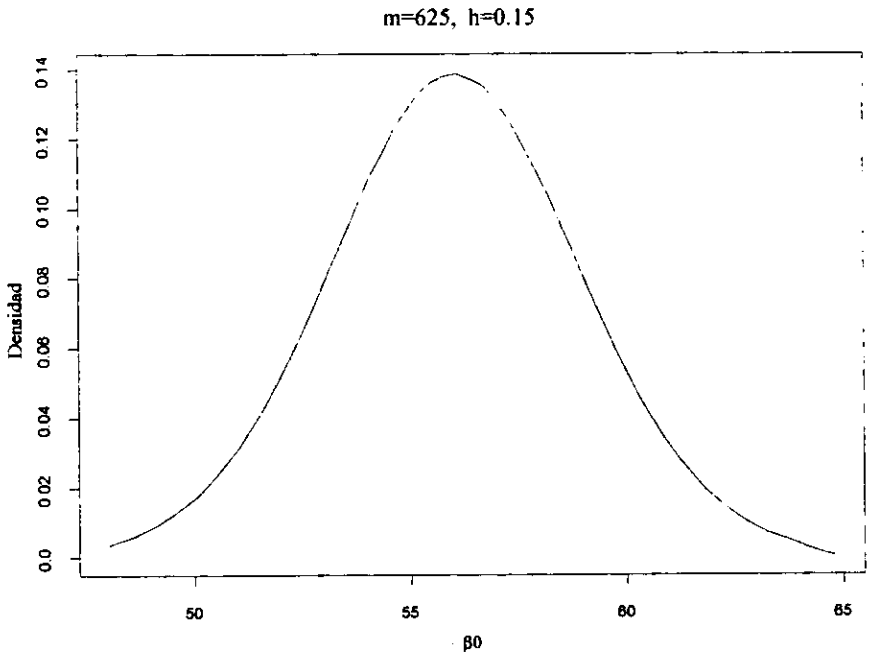


(a)

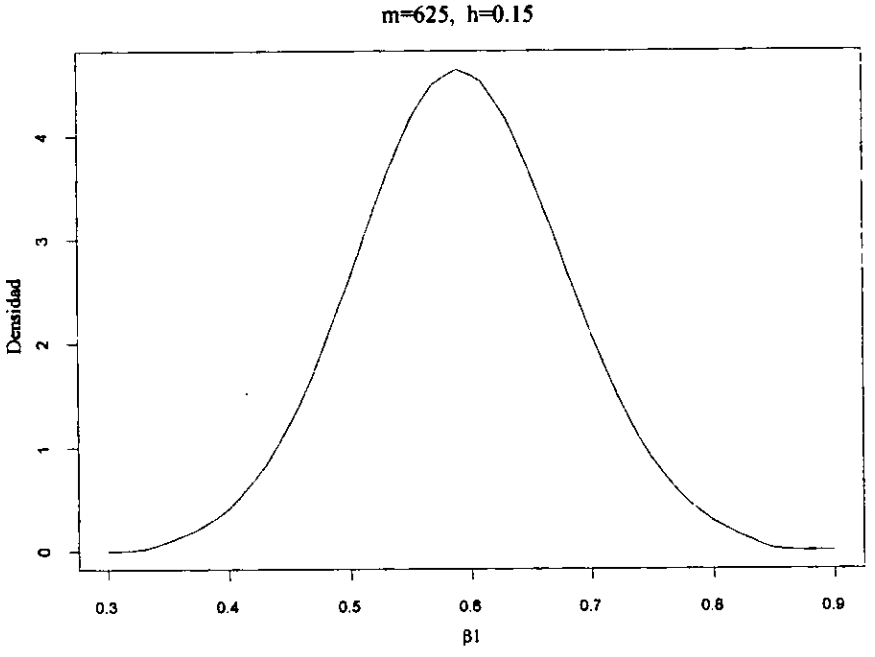


(b)

Figura 4.7: Nueva aproximación a la densidad final de $\boldsymbol{\beta}$. (a) Perspectiva y (b) Curvas de nivel.



(a)



(b)

Figura 4.8: Densidades marginales. (a) β_0 y (b) β_1.

Tabla 4.2.3.

	MCO	MCP	ABH (ana)	ABH (na)
β_0	56.089	56.169	56.074	56.017
$\sqrt{\mathrm{V}(\beta_0)}$	3.9937	2.7908	2.8661	2.8803
β_1	0.5895	0.5800	0.5903	0.5915
$\sqrt{\mathrm{V}(\beta_1)}$	0.0969	0.0840	0.0869	0.0857

(ana) es la aproximación normal asintótica y (na) es la nueva aproximación.

En la Tabla 4.2.3 se observa que las tres estimaciones para los coeficientes de la regresión son muy parecidas entre sí, pero difieren ligeramente en los errores estándar. Es claro que los errores estándar de la estimación por mínimos cuadrados ordinarios son mayores que los correspondientes errores de las otras dos estimaciones. En cambio, entre los errores estándar de la estimación por mínimos cuadrados ponderados y los errores estándar de la estimación Bayesiana casi no hay diferencia. Debe recordarse, sin embargo, que el análisis Bayesiano es más informativo pues produce una distribución de probabilidad conjunta sobre β_0 y β_1, y no sólo estimaciones puntuales y los errores estándar asociados.

Capítulo 5

Conclusiones.

A lo largo de este trabajo se discutieron aspectos importantes de las aproximaciones analíticas más usadas, para obtener resúmenes inferenciales, como son la aproximación normal asintótica y la aproximación de Laplace. Como se mencionó en el Capítulo 2, estas dos aproximaciones tienen su fundamento en argumentos asintóticos, y en muchos casos dan buenos resultados aún cuando se tienen tamaños de muestra relativamente pequeños.

Con la aproximación normal asintótica se pueden aproximar características de la distribución final de una forma rápida, en el caso de que la distribución final sea unimodal o al menos tenga una sola moda dominante. Por su parte, la aproximación de Laplace en forma exponencial produce aproximaciones con un orden de precisión mayor, pero a costa de un ligero incremento en los cálculos. Una de las principales ventajas de la aproximación de Laplace es su aplicación en el cálculo de densidades marginales debido a que se obtiene como resultado una forma analítica en los casos en los que la distribución final sea tratable analíticamente.

El tema central de esta tesis son las distribuciones finales multimodales. Tanto la aproximación normal asintótica como la aproximación de Laplace no son directamente aplicables en estos casos, lo que motivó el estudio de métodos alternativos diseñados especialmente para ellos.

O'Hagan (1994) presenta una generalización de la aproximación normal (asintótica) para el caso de multimodalidad, basada en mezclas de densidades normales (llamada *aproximación básica* en esta tesis). El mismo O'Hagan comenta que esta generalización es adecuada únicamente en el caso de que las modas estén suficientemente separadas. Con los ejemplos presentados en la Sección 3.1.1 de esta tesis se corroboró el hecho de que la aproximación básica no proporciona, en general, buenas aproximaciones debido a que la multimodalidad no siempre se presenta con las modas lo suficientemente separadas. Al estudiar los fundamentos de la aproximación básica se observó que los pesos de la mezcla de densidades normales que la definen, no son muy flexibles debido a que están totalmente determinados por la altura y la curvatura en cada una de las modas. En la búsqueda de una mejor manera de determinar los pesos de la aproximación básica, surgió una nueva idea para aproximar distribuciones mediante mezclas sin necesidad de restringir el número de componentes únicamente al número de modas.

La nueva aproximación presentada por primera vez en esta tesis (Sección 3.1.2), está basada también en mezclas de densidades normales y proporciona una mejor manera de aproximar densidades multimodales debido a que tanto el número de componentes de la mezcla, las localizaciones y las varianzas son hiperparámetros libres, que dependen de la distribución particular que se quiera aproximar. Con los ejem-

plos presentados a lo largo del Capítulo 3, el lector puede darse cuenta de que la nueva aproximación es muy flexible para aproximar casi cualquier comportamiento y se puede lograr cualquier precisión deseada controlando el número de componentes de la mezcla.

El determinar la localización de los componentes de la mezcla es un problema re-lativamente sencilllo ya que únicamente deben de estar en la región de mayor densidad de la distribución que se quiere aproximar. El número de componentes, como ya se mencionó, depende de la precisión deseada. El mayor problema para poder implementar esta nueva aproximación es el encontrar el parámetro de suavizamiento adecuado (h). De forma empírica se logró determinar una posible selección inicial para la búsqueda del valor de h que proporcione la mejor aproximación posible. Dicha selección dió buenos resultados en los ejemplos discutidos en esta tesis.

Una característica importante de esta aproximación es que está basada en mezclas de densidades normales circulares y con base en ellas se logra reproducir multimodalidad y estructuras de correlación complicadas. Es por ello que, para el caso donde la dimensión del parámetro es mayor a uno, la distribución que se desea aproximar debe de tener su área de mayor densidad sobre un rango de valores en la misma escala para cada componente. En el Capítulo 4 se presentaron dos ejemplos en los que hubo la necesidad de realizar un cambio de escala para poder implementar adecuadamente el nuevo método de aproximación a densidades.

Un uso adicional que se le puede dar a la nueva aproximación es como regla de integración, ya que del mismo sistema de ecuaciones de donde se obtienen los pesos se obtiene el valor aproximado de la constante de normalización. Esta idea será

explorada en el futuro.

Por otra parte, una limitación importante de este método es el hecho que el número de puntos necesarios para lograr una buena aproximación crece exponencialmente con la dimensión del parámetro cuya distribución se quiere aproximar, por lo que el esfuezo computacional para implementar el método puede hacerlo prohibitivo.

Tomando en cuenta que la aproximación de Laplace se pudiera ver como una aplicación de la aproximación normal sintótica, en la Sección 3.2 de esta tesis se generalizó la idea de la aproximación de Laplace al caso multimodal tomando como base, por un lado, a la aproximación básica y, por el otro, a la nueva aproximación. Ninguna de estas versiones de la aproximación de Laplace multimodal tiene una justificación analítica formal ya que realizan una aproximación en cada moda por separado y finalmente hacen una mezcla de las aproximaciónes obtenidas. En los ejemplos presentados en la Sección 3.2 se observó que la aproximación de Laplace básica requiere un menor esfuerzo computacional que la aproximación de Laplace nueva, pero la precisión obtenida es menor. La nueva aproximación a través de mezclas de densidades normales presentada en la Sección 3.1.2, utilizada como regla de integración, y la aproximación de Laplace multimodal nueva (tanto en forma estándar como en forma exponencial) tienen errores relativos menores que la aproximación de Laplace básica en los ejemplos presentados en esta tesis.

En los ejemplos del Capítulo 4, se presentó el problema de tener que transformar la escala del soporte de la distibución para poder aplicar de manera adecuada la nueva aproximación. En uno de los ejemplos, además de realizar un cambio de escala hubo la necesidad de hacer una rotación para dismunuir la correlación entre

los parámetros. Los dos ejemplos presentados muestran que la implementación de la nueva aproximación puede tener problemas al aplicarla en problemas prácticos donde se obtienen distribuciones finales no muy sencillas de manejar, pero esos problemas pueden ser resueltos mediante una transformación lineal.

La nueva aproximación, además de proporcionar aproximaciones adecuadas a densidades para un número suficientemente grande de componentes de la mezcla, incluso para un número de componentes de mezcla relativamente pequeño, captura la localización de las modas y, por lo general, también la estructura de correlación. Este tipo de aproximaciones con un número de componentes de mezcla relativamente pequeño, que por sí solas no necesariamente producen una buena aproximación a la verdadera distribución, podrían ser utilizadas como punto de partida para lograr una mejor aproximación usando algunas de las técnicas de Monte Carlo que se han popularizado en los últimos años. En particular, podrían ser utilizadas como "distribución candidato" en el método de muestreo-remuestreo o en el método Metropolis-Hastings. Esta posible aplicación es factible debido a que es muy fácil simular eficazmente de una mezcla de densidades normales.

Más aún, en los casos donde la dimensión del parámetro es grande, para los cuales se requeriría un trabajo computacional enorme para obtener buenos resultados, es posible calcular una aproximación con un número "pequeño" de componentes de mezcla, con la nueva aproximación aquí propuesta, y usar ésta como punto de partida de alguno de los métodos de remuestreo mencionados anteriormente.

Al final de la investigación realizada, y despúes de la experiencia adquirida al aplicar la nueva aproximación en los ejemplos presentados en esta tesis, surgen nuevas

lineas de investigación que se podrían abordar en trabajos posteriores.

El primer punto importante sobre la nueva aproximación a densidades descrita en la Sección 3.1.2, es estudiar y tratar de determinar de una manera más precisa el valor del parámetro de suavizamiento (h) que proporcione la mejor aproximación. Si el uso de la nueva aproximación es como regla de integración entonces determinar el valor de h que tenga el menor error posible en la estimación a la integral. Podría ser que ambos problemas quedaran resueltos al resolver sólo uno de ellos.

El segundo punto, también relativo a la nueva aproximación a densidades, es tratar estudiar el caso en el que las componentes de mezcla no fueran normales esféricas, sino elípticas, es decir, tener un parámetro de suavizamiento distinto para cada componente del vector de parámetros cuya distribución se quiera aproximar. Si $\boldsymbol{\theta} \in \Re^k$ el problema consistiría entonces en determinar $h_1, h_2, \ldots, h_k$, es decir, k parámetros de suavizamiento distintos. El hacer esto eliminaría el problema de realizar cambios de escala para poder aproximar con normales circulares. En todo caso debe evaluarse la conveniencia de buscar k parámetros de suavizamiento en lugar de uno solo junto con una transformación lineal apropiada.

Como tercer punto, la generalización de la aproximación de Laplace al caso multimodal basándose en la nueva aproximación usando una factorización estándar (Sección 3.2.2) tiene la forma de una "regla de integración por cuadratura" (Burden, 1993) que es $\sum_{i=1}^{n} a_i f(x_i)$. Una consideración importante sería el estudiar las propiedades de esta nueva aproximación de Laplace multimodal en forma estándar como regla de integración por cuadratura y compararla con otras reglas ya existentes como la regla trapezoidal y la regla de Gauss-Hermite.

El cuarto y último punto a considerar es la generalización de la nueva aproximación basada en mezclas de densidades normales, a una aproximación basada también en mezclas pero utilizando como componente de mezcla cualquier otra densidad dependiendo de las características del problema. Algunas posibilidades serían utilizar mezclas de densidades t-Student, mezclas de densidades Gamma si la densidad a aproximar tiene un soporte en los reales positivos, o mezclas de densidades Beta si el soporte de la densidad a aproximar está en el intervalo $(0, 1)$.

Referencias

Bernardo, J.M. (1979), Reference posterior distributions for Bayesian inference, *J. Roy. Statist. Soc. B*, **41**, 113-147 (with discussion).

Bernardo, J.M. y Smith A.F.M. (1994), *Bayesian Theory*, Chichester: Wiley.

Broemeling, L.D. (1985), *Bayesian Analysis of Linear Models*, New York: Marcel Dekker.

Burden, R.L. y Faires, J.D. (1993), *Numerical analysis*, Boston: Thomson.

Cobb, L., Koppstein, P. y Chen, N.H. (1983), Estimation and Moment Recursion Relations for Multimodal Distributions of the Exponential Family, *J. Amer. Statist. Assoc.*, **78**, 124-130.

Dréze, J.H. (1977), Bayesian Regression Analysis using Poly-t Densities, *Journal of Econometrics*, **6**, 329-354.

DeGroot, M.H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

Diaconis, P. e Ylvisaker, D. (1985), Quantifying prior opinion, *Bayesian Statistics 2* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley y A.F.M. Smith, eds.), Amsterdam: North-Holland, 133-156 (with discussion).

Efstathiou, M., Gutiérrez-Peña, E. y Smith, A.F.M. (1998), Laplace approximations for natural exponential families with cuts, *Scan. J. Statist.*, **25**, 77-92.

Gelfand, A.E. y Smith, A.F.M. (1990), Sampling based approches to calculating marginal densities, *J. Amer.Statist. Assoc.*, **85**, 398-409.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97-109.

Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, *Proc. Roy. Soc. A*, **186**, 453-461.

Johnson, R.A. (1970), Asymptotic expansions associated with posterior distributions, *Ann. Math. Statist.*, **41**, 851-864.

Kadane, J.D. (1980), Predictive and structural methods for eliciting prior distributions, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner ed.), Amsterdam: North-Holland, 89-109.

Kass, R.E, Tierney, L. y Kadane, J.B. (1988), Asymptotics in Bayesian computation, *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley y A.F.M. Smith, eds.), Oxford: University Press, 261-278 (with discussion).

Lindley, D.V. (1980), Approximate Bayesian methods, *Bayesian Statistics* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley y A.F.M. Smith, eds.), Valencia: University Press, 223-245 (with discussion).

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. y Teller, E. (1953), Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21**, 1087-1092.

Naylor, J.C. y Smith, A.F.M. (1982), Applications of a method for efficient computation of posterior distributions, *Appl. Statist.*, **31**, 214-225.

Neter, J., Wasserman, W. y Kutner M.H., (1989), *Applied Linear Regression Models*, Boston: Irwin.

O'Hagan A. (1994), *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, Cambridge: University Press.

Press, J.S. (1989), *Bayesian Statistics: Principles, Models and Applications*, New York: Wiley.

Rubin, D.B. (1988), Using the SIR algorithm to simulate posterior distributions, *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley y A.F.M. Smith, eds.), Oxford: University Press, 395-402 (with discussion).

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman y Hall.

Tierney L. and Kadane J.B. (1986), Accurate Approximations for Posterior Moments and Marginal Densities, *J. Amer. Statist. Assoc.*, **81**,82-86.

Tierney, L., Kass, R.E., Kadane, J.B. (1989), Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions, *J. Amer. Statist. Assoc.*, **84**, 710-715.

Tierney, L., Kass, R.E., Kadane, J.B. (1989 b), Approximate marginal densities of nonlinear functions, *Biometrika*, **76**, 425-433.

Walker, A.M. (1969), On the asymptotic behaviour of posterior distributions, *J. Roy. Statist. Soc. B.*, **31**, 80-88.

West, M. (1993), Approximating Posterior Distributions by Mixtures, *J. Roy. Statist. Soc. B.*, **55**, 409-422.

Wolpert, R.L. (1991), Monte Carlo importance sampling in Bayesian Statistics, *Statistical Multiple Integration* (N. Flournoy y R.K. Tsutakawa, eds.), Providence: RI: ASA.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.