

Funciones Iniciales de Referencia para Predicción Bayesiana.

Lizbeth Román Padilla

Tesis Doctoral



Facultad de Matemáticas
Departamento de Estadística e Investigación Operativa.
Universitat de València

Diciembre 2010

Don José Miguel Bernardo Herranz, Catedrático del Departamento de Estadística e Investigación Operativa de la Universitat de València,

CERTIFICA que la presente tesis

“Funciones iniciales de referencia para predicción Bayesiana”

ha sido realizada bajo mi dirección por Lizbeth Román Padilla y constituye su tesis para optar al grado de correspondiente al programa del Departamento de Estadística e Investigación Operativa.

Y para que así conste, firmo el presente certificado.

Valencia, Diciembre de 2010.

José Miguel Bernardo Herranz

A mi madre

Agradecimientos

Este trabajo fue escrito gracias al infinito amor que mi mamá me tiene, lo dedico a Blanca Guadalupe Padilla Robles, al amor de mis hermanas Bety y Abby, y de Fernando Román Nava, mi padre. Gracias.

Agradezco a mi familia en México, a mi abuelita Esperanza, a mis tios y primos, y a la memoria de los que no me logré despedir, a tío Fernando.

A las amigas que hicieron de esta experiencia en común una gran aventura de crecimiento profesional y humano, a Daniela Saucedo Reyes.

Al personal del departamento de Estadística e Investigación Operativa por acogerme y ofrecerme un cálido hogar académico donde he podido concebir esta investigación. Agradezco sinceramente su apoyo y comprensión en los más duros momentos.

A Antonio Félix Agusti por su apoyo, por sus consejos y por su compañía.

Al P. José Antonio Galindo Rodrigo y al grupo de catequesis de la iglesia de Ntra. Sra. de la Esperanza por su compañía. Por ayudarme a encontrar el Camino de la Fe.

Agradezco la paciencia y tiempo dedicados a la lectura de esta memoria a los profesores Miguel Ángel Gómez Villegas, Francisco Javier Girón González-Torré y a Elias Moreno Bas.

A mi tutor, director y maestro, José Miguel Bernardo Herranz por su paciencia, por sus ánimos y elogios que hicieron que no perdiera la *fe* en la investigación, y por su contagioso optimismo.

Finalmente, agradezco a México por la confianza que ha puesto en mi.

Índice general

Introducción	1
Notación	4
1 Métodos de obtención de distribuciones iniciales objetivas	9
1.1 Laplace y el <i>principio de razón insuficiente</i>	9
1.1.1 Principio de razón insuficiente	9
1.1.2 El postulado de Bayes-Laplace	10
1.2 Invarianza: las reglas de Jeffreys.	11
1.3 Verosimilitud de datos trasladados.	17
1.4 Máxima entropía	19
1.5 Geometría	20
1.6 Iniciales de cobertura consistente	21
1.7 Método de Zellner	22
1.8 Modelos de teoría de la decisión	22
1.9 Método de Rissanen	24
2 Análisis de referencia	27
2.1 Introducción	27
2.2 Discrepancia intrínseca e información esperada	29
2.3 Distribuciones de referencia	32
2.3.1 Modelos uniparamétricos	32
2.3.2 Un parámetro de ruido	40
2.3.3 Varios parámetros	44
2.3.4 Modelos no regulares con varios parámetros	46
2.3.5 Parámetros discretos	48
2.4 Algunas dificultades de usar iniciales objetivas	49
2.4.1 Inconsistencia fuerte	49
2.4.2 El efecto dominante de la inicial	50
2.4.3 Inadmisibilidad	50

2.4.4	Paradojas de marginalización	50
2.4.5	Densidades finales impropias	51
2.4.6	Dependencia del espacio muestral	52
3	Predicción con distribuciones iniciales objetivas.	55
3.1	Planteamiento del problema.	56
3.1.1	Probabilidad predictiva de que una nueva observación se encuentre contenida en el intervalo formado por dos observaciones previas. .	56
3.1.2	Probabilidad predictiva de que una nueva observación x_{n+1} pertenezca al intervalo formado por las observaciones mínima $x_{(1)}$ y máxima $x_{(n)}$ de una muestra dada.	58
3.2	Métodos de decisión de referencia para predicción	61
3.3	Inconsistencia fuerte	62
3.4	Distribuciones iniciales con validez frecuentista aproximada	64
3.5	Funciones de pérdida	67
3.6	La densidad predictiva máximo verosímil.	68
4	Funciones iniciales para predicción.	69
4.1	Predicción con datos Exponenciales	70
4.2	Predicción con datos Cauchy	72
4.3	Predicción con datos Normales	74
4.4	Predicción con datos Log-Normales	76
4.4.1	La media $E(x \mu, \sigma)$ como cantidad de interés	79
4.4.2	La moda $Mo(x \mu, \sigma)$ como cantidad de interés	81
4.4.3	La mediana $Me[x \mu, \sigma]$ como cantidad de interés	82
4.5	Modelos continuos no regulares	85
4.5.1	Modelo uniforme en $(0, \theta)$	86
4.5.2	Modelo uniforme en $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$	88
4.5.3	Modelo Exponencial con dos parámetros y parámetro de escala conocido	91
4.6	Modelo Exponencial con dos parámetros	96
4.6.1	Obtención de la densidad predictiva	100
5	Funciones iniciales para predicción en el modelo Normal bivariado.	107
5.1	Densidad Normal bivariada: estudio frecuentista	108
5.2	Densidad Normal bivariada: estudio Bayesiano	111
5.2.1	Caso ρ desconocido.	111
5.2.2	Caso ρ y σ desconocidos.	118

5.2.3	Caso todos los parámetros desconocidos	127
5.3	Predicción de una función de observaciones futuras	128
Catálogo de iniciales para predicción		134
Conclusiones y líneas de investigación		137
Apéndices		141
A Modelo Log-Normal		141
B Modelo Exponencial con dos parámetros.		145
B.1	Aproximación numérica de la función inicial de referencia $\pi_\phi(\phi, \varphi)$ cuando ϕ (mediana) es la cantidad de interés.	146
B.1.1	Aproximación numérica de la función inicial de referencia $\pi(\varphi \phi)$.	146
B.1.2	Obtención del modelo integrado con un solo parámetro $p(x \phi)$. .	146
B.1.3	Aproximación numérica de la función inicial $\pi(\phi)$	148
B.2	Obtención de la función de densidad predictiva resultante de la inicial $\pi_a(\alpha, \varphi)$. Expresión de la probabilidad requerida $\Pr(x_{(1)} < x < x_{(n)} \mathbf{x})$, para $n \geq 2$ observaciones aleatorias.	148
B.2.1	Dos observaciones	149
B.2.2	Más de dos observaciones	150
B.3	Obtención de la inicial de referencia $\pi_\varphi(\alpha, \varphi)$ cuando φ (parámetro de escala) es la cantidad de interés.	151
B.3.1	Obtención de $\pi(\alpha \varphi)$	152
B.3.2	Modelo integrado con un solo parámetro $p_i(x \varphi)$	152
B.3.3	Obtención de la función inicial de referencia $\pi(\varphi)$ del modelo integrado $p_i(\cdot \varphi)$	154
C Modelo normal Bivariado.		157
C.1	Densidades final y predictiva de la inicial de referencia π_ρ	157
C.2	Densidad final de la inicial de referencia π_σ	158
C.3	Demostración de $ g > 1$ y $ k > 1$	160
D Coeficiente de correlación desconocido		165
Referencias		169

Índice de figuras

2.1	Función inicial de referencia numérica para el modelo Exponencial. La línea continua corresponde a la inicial de referencia teórica $\pi(\theta) \propto \theta^{-1}$, los puntos corresponden a la aproximación numérica de la inicial de referencia	40
4.1	Comportamiento de la cobertura de la densidad predictiva resultante de la familia de distribuciones iniciales $\pi(\mu, \sigma) \propto \sigma^{-\alpha}$ ($\alpha > 0$), en el modelo $N(\cdot \mu, \sigma)$.	75
4.2	Función inicial de referencia numérica de $\pi(\alpha)$ para veinte valores de α uniformemente espaciados en $(-e^3, e^3)$ y re-escalados para que $\pi(1) = 1$.	94
4.3	Aproximación numérica de la inicial $\pi(\varphi \phi)$ con 20 valores de φ en un rango de e^{-3} a e^3 uniformemente log-espaciados re-escalados para que $\varphi(1 \phi) = 1$	101
4.4	Aproximación numérica de la inicial de referencia $\pi(\phi)$ del modelo integrado $p_i(x \phi)$	101
4.5	Probabilidad de $\Pr(x_1 < x < x_2 a)$ donde la inicial es $\pi(\alpha, \varphi) \propto \varphi^{-a}$, $a > 0$, con $a = 1$ se tiene $\pi_\phi(\phi, \varphi)$ y con $a = 2$ se tiene $\pi_G(\alpha, \varphi)$	102
5.1	Gráfica de la frecuencia con que una nueva observación se sitúa en el rectángulo formado por dos observaciones provenientes de $N_2(\cdot \mu, \Sigma)$ junto con la función teórica $\frac{1}{9\sqrt{1-\rho^2}}$.	110
B.1	Histograma y función de densidad teórica del modelo $p_i(x \phi)$, correspondiente de $n = 5000$ muestras aleatorias con $i = 3$ y $\phi = 2$.	148
B.2	Histograma correspondiente de $n = 10000$ observaciones simuladas del modelo integrado con un parámetro $p_i(x \varphi)$ versus la función de densidad teórica.	155
B.3	Aproximación numérica de la función inicial de referencia $\pi(\varphi)$	155

Índice de tablas

1	Distribuciones continuas univariadas	5
2	Distribuciones continuas univariadas (cont.)	6
3	Distribuciones continuas multivariadas	6
4.1	Valor medio y desviación estándar de la cobertura de probabilidad $\Pr(x_{(1)} < x < x_{(n)} x_1, \dots, x_n)$ provenientes del modelos $\text{LgN}(x \mu, \sigma)$ e inicial $\pi(\mu, \sigma) \propto \sigma^{-1}$	78
4.2	Valor medio (desviación estándar) de las probabilidades $\Pr(x_{(1)} < x < x_{(n)} \mathbf{x}) = \int_{x_1}^{x_2} p(x \mathbf{x}) dx$ provenientes de $\text{LgN}(x \mu, \sigma)$ para diferentes valores de μ y σ y con iniciales dadas por π_{ϕ_1} , π_{ϕ_2} y π_{ϕ_3}	84
4.3	Frecuencia con que una nueva observación pertenece al intervalo formado por dos observaciones dadas (x_1, x_2) ($x_1 < x_2$) provenientes de una distribución $\text{Un}(0, \theta)$	87
4.4	Valor medio y desviación estándar de la cobertura de probabilidad de la densidad predictiva de $\text{Un}(0, \theta)$ con previa de referencia dada por $\pi(\theta) \propto \theta^{-1}$	87
4.5	Frecuencia con que una tercera observación se encuentra contenida en el intervalo formado por dos observaciones previas provenientes de $\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$	89
4.6	Valor medio y desviación estándar de la cobertura de probabilidad de la densidad predictiva del modelo $\text{Un}(0, \theta)$ y la inicial de referencia $\pi(\theta) \propto \theta^{-1}$	90
4.7	Frecuencias del evento $x \in (x_1, x_2)$ correspondientes de $n = 10000$ muestras simuladas de tamaño dos provenientes de $\text{Ex}(\cdot \alpha)$	92
4.8	Media muestral (desviación estándar) de la cobertura de probabilidad de la densidad predictiva resultante de la inicial de referencia $\pi(\alpha) \propto 1$ y el modelo no regular $\text{Ex}(x \alpha, 1)$	95
4.9	Frecuencia de que una nueva observación x con distribución $\text{Ex}(x \alpha, \beta)$ pertenezca al intervalo $(x_{(1)}, x_{(n)})$ formado por las observaciones máxima y mínima de la muestra $\mathbf{x} = \{x_1, \dots, x_n\}$	98

4.10	Media muestral de la cobertura de probabilidad $\Pr(x_1 < x < x_2 x_1, x_2)$ proveniente de $m = 100$ muestras simuladas de tamaño dos de $\text{Ex}(x \alpha, \varphi)$, con diferentes valores de α y φ	103
4.11	Valor medio de la cobertura de probabilidad $\Pr(x_{(1)} < x < x_{(n)} \mathbf{x})$ de $\text{Ex}(x \alpha, \theta)$, para diferentes valores de α y φ con función inicial de referencia $\pi_\phi(\alpha, \varphi) \propto \varphi^{-1}$	104
5.1	Frecuencia con que una tercera observación cae en el rectángulo formado por dos observaciones dadas provenientes de $N_2(\mathbf{x} \boldsymbol{\mu}, \Sigma)$	110
5.2	Convergencia de las densidades finales $\pi_{R\rho}$, π_{DM} y π_φ	116
5.3	Valor medio y desviación estándar de la cobertura de probabilidad de la densidad predictiva resultante de las iniciales $\pi_{R\rho}$, π_{DM} y $\pi_{R\rho}$	117
5.4	Valor medio y desviación estándar de la cobertura de probabilidad de la densidades predictivas resultantes de las iniciales π_ρ y π_α	122
5.5	Valor medio (desviación estándar) de la cobertura de probabilidad de las densidades predictivas provenientes de las iniciales π_ρ y π_σ	126
5.6	Frecuencia con que una nueva observación pertenece al intervalo formado por dos observaciones previas.	130
5.7	Valor medio (desviación estándar) de la cobertura de probabilidad de la densidad predictiva resultante de la función inicial $\pi(\lambda) \propto \lambda^{-1}$	132
5.8	Valor medio (desviación estándar) de la cobertura de probabilidad de la densidad predictiva $p(\cdot y_1, y_2)$ correspondiente de $n = 100$ muestras aleatorias de tamaño dos.	134
	Funciones iniciales de referencia para predicción bajo hipótesis de cobertura	135
	Iniciales objetivas para predicción para el modelo $N_2(\mathbf{x} \mathbf{0}, \Sigma)$	135

Introducción

Motivación

Según la Real Academia Española predecir es el acto de declarar o indicar por adelantado, anticiparse a los hechos, *anunciar por revelación, ciencia o conjetura algo que ha de suceder*.

La predicción siempre ha jugado un papel relevante en el medio científico puesto que se desea saber de modo preciso lo que ocurrirá bajo determinadas condiciones, y un método de predicción tendría éxito en la medida que sean acertadas sus predicciones.

Predecir fue la forma inicial de la inferencia estadística, cuyo objetivo original fue *inferir acerca de los posibles valores que no han sido observados en base a los valores que ya han sido observados*(Geisser, 1993).

Casi todas las áreas científicas ha abordado el problema de la predicción y no ha sido la excepción la estadística matemática y en particular la estadística Bayesiana.

La estadística Bayesiana debe su nombre al uso frecuente del teorema de Bayes, que es un procedimiento que revisa y actualiza la probabilidad de un suceso a la luz de los datos observados. Ésta es la esencia de la *inferencia Bayesiana*.

Para hacer inferencia Bayesiana se necesita especificar una función de *verosimilitud* basada en los datos observados, suponer un modelo de probabilidad con parámetros desconocidos y una distribución *inicial* o *a priori* que describa lo que se sabe de los parámetros desconocidos antes de que los datos hayan sido observados. El uso del teorema de Bayes por tanto proporciona la distribución *final* o *posterior* de los parámetros desconocidos, que nos explica *lo que se conoce acerca de los parámetros [desconocidos] después de que los datos han sido observados*. La distribución final permite hacer resúmenes inferenciales tales como la estimación puntual o la estimación por intervalos.

La distribución inicial representa el conocimiento previo (o falta de él) que tiene el investigador antes de observar los datos. Pero, ¿cómo especificar esta distribución inicial? Varias son las propuestas para elegir esta distribución inicial *no informativa* u *objetiva*, y variados son los soportes teóricos que sustentan cada uno de éstos métodos. Bayes (1763) y Laplace (1814) propusieron basarse en la especificación de una distribución que reflejara la completa *ignorancia* acerca de los parámetros. Sus propuestas permanecieron

vigentes hasta la llegada de los trabajos de Jeffreys (1946, 1961) quién revolucionó esta área dándo luz a la estadística Bayesiana moderna. A partir de ese momento, varios trabajos más han surgido compartiendo el mismo objetivo aunque los enfoques filosóficos han ido cambiando. Por ejemplo, se intenta que sean *los datos los que hablen por sí mismos* en el momento de elegir una distribución inicial.

Actualmente, ésta área de la estadística Bayesiana sigue siendo campo fértil para la investigación y se la conoce como análisis Bayesiano objetivo.

Por otro lado, predecir según el análisis Bayesiano objetivo también fue uno de los temas que Jeffreys abordó; de hecho se cree que fue el primero en hacerlo. Su trabajo se centró en los modelos de localización y escala univariados continuos. Desde entonces la predicción Bayesiana objetiva ha sido abordada muy frecuentemente.

Los métodos Bayesianos objetivos necesitan iniciales que únicamente dependan del modelo supuesto y de la cantidad de interés. El *análisis de referencia* (Bernardo 1979, Berger y Bernardo 1992a, Bernardo 2005, Berger, Bernardo y Sun, 2009a) proporciona una solución general cuando la cantidad de interés es alguna función de los parámetros.

Una posibilidad es, abordar el problema de predecir con el análisis de referencia, pero no está claro cómo proceder cuando la cantidad de interés es alguna función de las observaciones futuras. En este sentido, Berger y Bernardo (1992a) proponen el uso de alguna función del modelo como cantidad de interés. Es decir, elegir como cantidad de interés una función $\phi(\boldsymbol{\theta}) = g[p_x(\cdot|\boldsymbol{\theta})]$ del vector de parámetros que únicamente depende del modelo supuesto $\{p_x(\cdot|\boldsymbol{\theta}), x \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$, donde g puede ser cualquier estimador de las observaciones futuras tales como la media $E(x|\boldsymbol{\theta})$, la moda $Mo(x|\boldsymbol{\theta})$ o la mediana $Me(x|\boldsymbol{\theta})$.

Esta propuesta puede ser útil para predecir en la estadística Bayesiana, pero surge un nuevo interrogante, pues dentro de todas las posibles iniciales provenientes de dichas cantidades de interés hay que averiguar cuál es la que debe elegirse.

Por otro lado, una forma de *discriminar* entre las posibles iniciales alternativas es usar las propiedades de cobertura de la densidad final resultante. En efecto, el uso de las propiedades de cobertura de probabilidad para discriminar entre esas iniciales tiene una historia larga (Welch y Peers 1963, Berger y Bernardo 1989, Datta y Sweeting, 2005). En este trabajo, usaremos esta técnica para elegir entre las iniciales resultantes de las distintas funciones g que pueden definir la cantidad de interés.

Nuestro objetivo es la elección de la inicial de referencia *adecuada* para predecir de forma que la densidad predictiva resultante tenga buenas propiedades de cobertura.

Estructura

Varias son las propuestas para trabajar el problema de la predicción dentro del análisis Bayesiano objetivo, algunas de las más recientes son las dadas por Kuboki (1998), que sigue los mismos pasos de obtención dados en Bernardo (1979); Eaton y Sudderth (1998), intentando evitar la propiedad de inconsistencia fuerte; Komaki (1996) y Ghosh y Sweeting (2000) quienes usan la validez frecuentista aproximada; Smith (1999) quien asegura que la función de pérdida juega un papel importante, y Hartigan (1998) que propone la distribución inicial de máxima verosimilitud para el problema de predicción. Ninguna de estas propuestas proporciona soluciones generales con buenas propiedades de cobertura frecuentista. Un objetivo adicional de este trabajo es el de proponer un procedimiento general de obtención de distribuciones iniciales objetivas para la predicción, basadas en las propiedades de cobertura de la densidad predictiva.

La estructura de este trabajo es como sigue: El primer capítulo proporciona un resumen general de los métodos de construcción de iniciales objetivas con ejemplos y explicación de sus posibles dificultades. El segundo capítulo se centra sólo en el análisis de referencia pues es el antecedente de nuestra propuesta de solución para el problema de predicción. El tercer capítulo contiene un resumen de las propuestas actuales de iniciales objetivas para la predicción. El cuarto capítulo desarrolla el objetivo de este trabajo, y es donde se estudia el problema de predicción en modelos paramétricos univariados continuos regulares y no regulares. Asimismo, es donde encontramos que si se define la mediana como cantidad de interés la inicial de referencia produce una densidad predictiva que posee adecuadas propiedades de cobertura. También encontramos que dentro de las iniciales objetivas propuestas, las resultantes del análisis de referencia en general y sus respectivas densidades, poseen adecuadas propiedades de cobertura. El quinto capítulo estudia el modelo Normal bivariado con vector de medias conocido y elementos de la matriz de dispersión desconocidos, donde se trabaja con varios supuestos sobre sus parámetros. Cuando se trabaja con modelos multivariados no está muy claro cual es el valor de la mediana, debido a ello, hacemos uso de las propiedades de cobertura de probabilidad de la predictiva para discriminar entre las posibles iniciales objetivas alternativas. Nos encontramos entonces, que no hay una única cantidad de interés para el problema de predicción en este modelo, pues más de una de ellas proporciona una adecuada cobertura. Finalizamos con un catálogo de iniciales para predicción de los modelos estudiados, así como con una sección de conclusiones y futuras líneas de investigación. Las referencias bibliográficas, tablas y apéndices completan esta Tesis.

Notación

A lo largo de este trabajo se supone que las distribuciones de probabilidad son descritas por sus respectivas funciones de densidad, en caso de ambigüedad, se hace una distinción entre las cantidades aleatorias de los valores que éstos pudieran tomar. Los vectores aleatorios se representan en negritas, si éstos son observables se escribirán en el alfabeto latino y cuando no son observables, en alfabeto griego. La densidad de probabilidad condicional de los datos $\mathbf{x} \in \mathcal{X}$ dado $\boldsymbol{\theta}$ se representa como $p(\mathbf{x}|\boldsymbol{\theta})$ o $p_{\mathbf{x}|\boldsymbol{\theta}}$, con $p(\mathbf{x}|\boldsymbol{\theta}) \geq 0$ y $\int_{\mathcal{X}} p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = 1$; la distribución final o *posterior* de $\boldsymbol{\theta} \in \Theta$ dado \mathbf{x} se representará como $p(\boldsymbol{\theta}|\mathbf{x})$ o $p_{\boldsymbol{\theta}|\mathbf{x}}$, con $p(\boldsymbol{\theta}|\mathbf{x}) \geq 0$ y $\int_{\Theta} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = 1$. En este trabajo, las funciones de densidad de distribuciones específicas se denotan por su correspondiente nombre abreviado y la Tabla 1 muestra las funciones de densidad usadas junto con su espacio paramétrico, así como el espacio muestral y algunas de sus características, tales como la media y la varianza.

Tabla 1: Distribuciones continuas univariadas.

Uniforme: $\text{Un}(x a, b)$	
$b > a$	$a < x < b$
$p(x) = c$	$c = \frac{1}{b-a}$
$E[x] = \frac{a+b}{2}$	$\text{Var}[x] = \frac{(b-a)^2}{12}$
Gama: $\text{Ga}(x \alpha, \beta)$	
$\alpha > 0, \beta > 0$	$x > 0$
$p(x) = cx^{\alpha-1}e^{-\beta x}$	$c = \frac{\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \frac{\alpha}{\beta}$	$\text{Var}[x] = \frac{\alpha}{\beta^2}$
Exponencial: $\text{Ex}(x \theta)$	
$\theta > 0$	$x > 0$
$p(x) = ce^{-\theta x}$	$c = \theta$
$E[x] = \frac{1}{\theta}$	$\text{Var}[x] = \frac{1}{\theta^2}$
Raíz cuadrada de la Gama inversa: $\text{Ga}^{-1/2}(x \alpha, \beta)$	
$\alpha > 0, \beta > 0$	$x > 0$
$p(x) = cx^{-(2\alpha+1)}e^{-\beta/x^2}$	$c = \frac{2\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \frac{\sqrt{\beta}\Gamma(\alpha-1/2)}{\Gamma(\alpha)}$	$\text{Var}[x] = \frac{\beta}{\alpha-1} - E[x]^2$
Gama inversa: $\text{IGa}(x \alpha, \beta)$	
$\alpha > 0, \beta > 0$	$x > 0$
$p(x) = cx^{-(\alpha+1)}e^{-\beta/x}$	$c = \frac{\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \frac{\beta}{\alpha-1}$	$\text{Var}[x] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
Normal: $\text{N}(x \mu, \sigma)$	
$-\infty < \mu < \infty, \sigma > 0$	$-\infty < x < \infty$
$p(x) = c \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$	$c = \frac{1}{\sqrt{\pi\sigma^2}}$
$E[x] = \mu$	$\text{Var}[x] = \sigma^2$
t de Student: $\text{St}(x \mu, \sigma, \alpha)$	
$-\infty < \mu < \infty, \sigma > 0, \alpha > 0$	$-\infty < x < \infty$
$p(x) = c \left(1 + \frac{1}{\alpha\sigma^2}(x - \mu)^2\right)^{-(\alpha+1)/2}$	$c = \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha}{2})} \left(\frac{1}{\sigma^2\alpha\pi}\right)^{1/2}$
$E[x] = \mu$	$\text{Var}[x] = \alpha\sigma^2(\alpha - 2)^{-1}$
Gamma-Gamma: $\text{GaGa}(x \alpha, \beta, n)$	
$\alpha > 0, \beta > 0, n > 0$	$x > 0$
$p(x) = c \frac{x^{n-1}}{(\beta+x)^{\alpha+n}}$	$c = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+n)}{\Gamma(n)}$
$E[x] = n \frac{\beta}{\alpha-1}$	$\text{Var}[x] = \frac{\beta^2(n^2+n(\alpha-1))}{(\alpha-1)^2(\alpha-2)}$
Log-Normal: $\text{LN}(x \mu, \sigma)$	
$-\infty < \mu < \infty, \sigma > 0$	$x > 0$
$p(x) = c \frac{1}{x} \exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\}$	$c = \frac{1}{\sqrt{2\pi\sigma^2}}$
$E[x] = e^{\mu+\sigma^2/2}$	$\text{Var}[x] = \sigma^2$
$\text{Med}[x] = e^\mu, \text{Mo}[x] = e^{\mu-\sigma^2}$	
Log-Student: $\text{LSt}(x \mu_1, \nu, \alpha)$	
$\mu_1 > 0, \nu > 0, \alpha > 0$	$x > 0$
$p(x) = c \frac{1}{x} \left(1 + \frac{1}{\alpha\nu^2} \left(\log \frac{x}{\mu_1}\right)^2\right)^{-\frac{\alpha+1}{2}}$	$c = \frac{\Gamma((\alpha+1)/2)}{\nu\sqrt{\alpha\pi}\Gamma(\alpha/2)}$
(Véase Good y Crook, 1987)	
Cauchy: $\text{Ca}(x \mu, \sigma) = \text{St}(x \mu, \sigma, 1)$	

Tabla 2: Distribuciones continuas univariadas (cont.).

Ji cuadrada: $\chi^2(x \nu) = \chi_\nu^2$	
$\nu > 0$	$x > 0$
$p(x) = cx^{\nu/2-1}e^{-x/2}$	$c = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)}$
$E(x) = \nu$	$\text{Var}(x) = 2\nu$
Ji cuadrada descentralizada: $\chi^2(x \nu, \lambda)$	
$\nu > 0, \lambda > 0$	$x > 0$
$p(x) = \sum_{i=0}^{\infty} \text{Po}(i \lambda/2)\chi^2(x \nu + 2i)$	
$E(x) = \nu + \lambda$	$\text{Var}(x) = 2(\nu + 2\lambda)$

Tabla 3: Distribuciones continuas multivariadas.

Normal multivariada: $N_k(\mathbf{x} \boldsymbol{\mu}, \Sigma)$	
$\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)' \in \mathcal{R}^k$	$\mathbf{x} = (x_1, \dots, x_k)'$
Σ simétrica, positiva definida	
$p(\mathbf{x}) = c \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$	$c = (2\pi)^{-k/2} \Sigma ^{-1/2}$
$E(\mathbf{x}) = \boldsymbol{\mu}$	$\text{Var}(\mathbf{x}) = \Sigma$

El análisis de referencia siempre se define en términos de modelos paramétricos con la siguiente forma general:

$$\mathcal{M} \equiv \{p(\mathbf{x}|\boldsymbol{\omega}), \mathbf{x} \in \mathcal{X}, \boldsymbol{\omega} \in \Omega\}$$

con $p(\mathbf{x}|\boldsymbol{\omega}) \geq 0$ y $\int_{\mathcal{X}} p(\mathbf{x}|\boldsymbol{\omega})d\mathbf{x} = 1$. A lo largo de este trabajo siempre se supone que el modelo describe las condiciones bajo las cuales los datos fueron generados. Algunas veces, pero no siempre, los datos consistirán en una muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_n\}$ de tamaño fijo n proveniente de alguna distribución de probabilidad. Si $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ es el vector de interés, el modelo puede reparametrizarse, sin pérdida de generalidad, en la forma $\mathcal{M} \equiv \{p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\lambda}), \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda} \in \Lambda\}$ donde $\boldsymbol{\lambda}$ es un vector de parámetros de ruido y $\Omega = \Theta \times \Lambda$ el espacio paramétrico. Además, $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ denota una distribución inicial objetiva conjunta para $\{\boldsymbol{\theta}, \boldsymbol{\lambda}\}$ cuando $\boldsymbol{\theta}$ es la cantidad de interés.

Los momentos y otras medidas de dispersión de las distribuciones de probabilidad se describen con adecuados símbolos, por ejemplo,

$$E(\mathbf{x}|\boldsymbol{\theta}) = \int_{\mathcal{X}} \mathbf{x}p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}$$

$$\text{Var}(\mathbf{x}|\boldsymbol{\theta}) = \int_{\mathcal{X}} (\mathbf{x} - E[\mathbf{x}|\boldsymbol{\theta}])^2 p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}$$

son la esperanza o media, y la varianza de \mathbf{x} bajo el modelo $\{p_{\mathbf{x}|\boldsymbol{\theta}}, \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$. De

manera similar, $\text{Med}[\mathbf{x}|\boldsymbol{\theta}]$ y $\text{Mo}[\mathbf{x}|\boldsymbol{\theta}]$ denotan respectivamente la mediana y la moda. Los estadísticos comúnmente usados en modelos de una dimensión son la media y varianza muestral, denotadas por

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Asímismo, $x_{(i)}$ denota el i -ésimo estadístico de orden, en particular $x_{(1)}$ (x_{\min}) y $x_{(n)}$ (x_{\max}) denotan el mínimo y máximo de la muestra, respectivamente.

Capítulo 1

Métodos de obtención de distribuciones iniciales objetivas

Este capítulo contiene un breve resumen de algunos métodos de obtención de distribuciones iniciales objetivas. Presentamos una descripción de cada método y una breve discusión de la filosofía que lo sustenta; también incluimos algunos ejemplos. Asimismo, mencionamos algunas de las interpretaciones de las distribuciones iniciales que se han hecho a lo largo de los años, desde los trabajos de Bayes, Laplace y Jeffreys hasta este momento. Citaremos también algunas de las propiedades que esperamos cumplirán las distribuciones iniciales objetivas.

1.1. Laplace y el *principio de razón insuficiente*

Los primeros intentos para definir una distribución inicial no informativa con la que se pueda representar la *ignorancia* se encuentran registradas en los trabajos de Bayes (1763) y posteriormente en Laplace (1814).

El método de construcción de funciones iniciales no informativas que ellos proponen es conocido como *principio de razón insuficiente* o también como *el postulado de Bayes-Laplace*.

A continuación haremos una breve descripción de este principio e incluiremos algunas de las críticas surgidas a lo largo de los años respecto del mismo.

1.1.1. Principio de razón insuficiente

El principio de razón insuficiente especifica que *las probabilidades de los eventos serán iguales si no existen razones para pensar que deberían ser diferentes*.

Bayes (1763) fue el primero en utilizar este principio y lo ejemplificó con el parámetro θ de la distribución Binomial, al cual le asignó la función inicial uniforme argumentando que tenemos un *total desconocimiento de su valor*.

Años más tarde, y de acuerdo al historiador Stigler (1986b), cuando Laplace estableció su *principio* de probabilidad inversa no lo reconoció como un equivalente al teorema de Bayes, pues asumió que todas las causas *a priori* son igualmente posibles o *indiferentemente indistinguibles*. A su principio también se le conoce como *principio de indiferencia de Laplace*.

De manera implícita, Laplace aplicó su principio al muestreo binomial, asignándole en sorprendentemente coincidencia la misma distribución uniforme al parámetro θ . También lo aplicó a tasas de decrecimiento expresando de manera *elegante* una curva de error (Stigler, 1986a, p. 111).

Por otro lado, desde el punto de vista de la inferencia inductiva de Jeffreys, el principio de Laplace corresponde a considerar todas las probabilidades igualmente posibles y así expresar de manera formal la ignorancia (Jeffreys, 1961, p. 33-34), o *distribución de la ignorancia*,

“...to say that the probabilities are equal is a precise way of saying that we have no ground for choosing between the alternatives.”

Posteriormente, Hartigan (1983, p. 2) señala que el principio de razón insuficiente no puede ser aceptado debido a que demasiadas probabilidades son igualmente posibles lo que daría como resultado algunas paradojas de partición. Entonces es inconsistente asignar simultáneamente esta regla a todas las *vastedades y refinamientos* del espacio paramétrico (véase los ilustrativos ejemplos discutidos en Shafer (1976, p. 23-24) y Kass y Wasserman (1996, secc. 3.1)).

Debe notarse que Laplace supuso que el problema sería especificado de tal manera que todos los resultados fueran igualmente probables, y en el caso de un resultado con una alta probabilidad de ocurrencia, éste debería subdividirse hasta lograr que todos los sub-resultados fuesen igualmente probables. Lo que demuestra que Laplace era consciente de que el supuesto “igualmente probables” solo tendría sentido en una parametrización en particular.

1.1.2. El postulado de Bayes-Laplace

Aplicar el principio de razón insuficiente cuando se tiene un espacio paramétrico continuo, significa usar una distribución inicial plana, más específicamente, usar una constante positiva como distribución inicial. Pero el uso de una inicial plana implica problemas de invarianza, como por ejemplo, si θ es uniforme, la función $\phi = e^\theta$ no lo es, e inversamente, si ϕ es uniforme, $\theta = \log \phi$ no lo será. Por tanto, en el caso continuo y dependiendo de la parametrización podría ser apropiado usar la distribución uniforme como inicial.

Según Kass y Wasserman (1996), la primera (y más famosa) aplicación registrada de la distribución Uniforme como distribución inicial es la hecha por Bayes (1763), pues la

atribuye como inicial al parámetro de una distribución Binomial.

Según Stigler (1982), el artículo inicial de Bayes ha sido malinterpretado, debido principalmente a su difícil lectura y comprensión. E incluso, Stigler asegura que relevantes personalidades como Fisher, Pearson y Jeffreys también cayeron en el error de no entender la idea principal de Bayes.

De acuerdo a Stigler, la idea original de Bayes fue suponer una variable aleatoria X proveniente de una distribución binomial basada en n intentos con probabilidad de éxito θ . Entonces, antes de observar X nosotros *no sabemos nada*. Bayes definió esta ignorancia como $\Pr(X = r)$ lo cual equivale a ser constante y esto requiere que θ sea uniforme. Específicamente,

$$\begin{aligned}\Pr(X = r) &= \int_0^1 \text{Bi}(x|\theta, r)\pi(\theta)d\theta \\ &= \frac{1}{n+1} \quad r = 0, 1, 2, \dots, n\end{aligned}$$

con $\pi(\theta) = 1$.

Así tenemos el postulado de *Bayes-Laplace*, que es considerado como casi equivalente al principio de razón insuficiente.

Sin embargo, Fisher (Stigler, 1982, Sec. 6) objeta que este argumento es restrictivo pues está basado en mecanismos físicos. Además, la distribución uniforme de X está sujeta a paradojas de refinamiento (Kass y Wasserman, 1996).

1.2. Invarianza: las reglas de Jeffreys.

Desde el inicio de la historia de las distribuciones iniciales objetivas, la propiedad de invarianza ha sido una de las propiedades que se suele esperar que cumplan las distribuciones iniciales. Como se ha mencionado anteriormente, era una práctica común utilizar la distribución Uniforme para representar la ignorancia, pero esta primera práctica Bayesiana deriva en inconsistencias. Por ejemplo, si $\phi = \sin^{-1} \sqrt{\theta}$ tiene una función inicial $\pi(\phi) = 1$ entonces θ tiene distribución $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$, y si además uno no sabe nada del valor de ϕ , tampoco sabrá nada acerca del valor de θ .

Jeffreys (1946) fue el primero en comprender completamente este hecho y dedicó todos sus esfuerzos para conseguir “reglas” de obtención de distribuciones iniciales objetivas, las cuales deberían ser invariantes en el sentido explicado.

Las reglas de Jeffreys

La primera dificultad a la que Jeffreys se enfrentó fue la de encontrar la forma de decir que la magnitud del problema es desconocida cuando no se exige especial atención con respecto a ningún valor en particular. Él consideró varios escenarios para proponer sus

reglas de acuerdo al posible rango de valores del parámetro.

Si el rango del parámetro es:

- finito o la recta real $(-\infty, \infty)$, entonces la distribución inicial debería ser la distribución Uniforme; o
- si el rango son los reales positivos $(0, \infty)$, entonces la distribución inicial del logaritmo del parámetro debería ser también la distribución Uniforme.

Jeffreys aspiraba a encontrar una regla general de obtención de iniciales que produjera los mismos resultados, y propuso que la propiedad de invarianza fuese la que la que hiciera de esa regla de aplicación general. Además, la esencia de esas reglas fue la de expresar en lenguaje formal la *ignorancia* acerca del valor del parámetro en un rango de valores permitido, es decir, las reglas formales deberían ser tan impersonales hasta donde fuese posible.

La regla general de Jeffreys se encuentra postulada en Jeffreys (1946) donde menciona que para poder definir la regla de obtención es necesario hacer una distinción entre la estimación y los problemas de pruebas de hipótesis.

La regla de Jeffreys nos dice que la distribución inicial será la raíz cuadrada del determinante de $I(\boldsymbol{\theta})$, la matriz de información de Fisher,

$$\pi_{\theta}(\theta) \propto \det(I(\boldsymbol{\theta})_{ij})^{1/2},$$

donde

$$I(\boldsymbol{\theta})_{ij} = \mathbb{E} \left(-\frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right)$$

con $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ y l es la log-verosimilitud de los datos.

Esta regla es aplicable si $I(\boldsymbol{\theta})$ es definida positiva y además es invariante ante posibles parametrizaciones, *i.e.*,

$$\pi_{\theta}(\theta) = \pi_{\gamma}(\gamma(\theta)) \left| \det \left(\frac{\partial \gamma}{\partial \theta} \right) \right|.$$

Es decir, la distribución inicial resultante de la parametrización $\gamma = \gamma(\theta)$ es probabilísticamente consistente con la que resulte de usar θ .

En este sentido su regla es general y Jeffreys la describió como “auto-consistente”. Sin embargo, Jeffreys (1946) encontró ciertas limitaciones en su aplicación, pues no siempre es verdad que todos los parámetros de un modelo de probabilidad son diferenciables. Por

ejemplo, Jeffreys discutió el caso cuando los parámetros sólo toman valores discretos, como sucede en el caso de la distribución Rectangular.

Más aún, sus dos reglas entran en conflicto entre ellas mismas lo cual puede verse muy fácilmente con el modelo normal: De acuerdo a la regla basada en el rango de valores del parámetro, la función inicial es $\pi_{\mu,\sigma}(\mu, \sigma) \propto \frac{1}{\sigma}$. Mientras que la regla general resulta en $\pi_{\mu,\sigma}(\mu, \sigma) \propto \sigma^{-2}$. Es más, si la regla general se aplica a más parámetros, deberá añadirse un factor de corrección σ^{-1} a cada uno de ellos, lo cual es *inacceptable* según Jeffreys (1961, p. 182); porque en el caso de varias medias $\{\mu_i, i = 1, \dots, k\}$ y desviación estándar única σ , los grados de libertad de una distribución t de Student dependerán del número de observaciones de la muestra para cualesquier k parámetros que pudiese haber. Véase Kass y Wasserman (1996, p. 1345) y Bernardo y Smith (1994, Ej. 5.25).

Para evitar el problema anterior, la sugerencia de Jeffreys fue la de suponer independencia entre μ y σ y así tratarlos separadamente. Entonces cuando σ fuese desconocida y μ un valor fijo, la función inicial para σ debería ser $\pi_\sigma(\sigma) \propto \sigma^{-1}$, y cuando σ fuese conocida y μ desconocida, la función inicial para μ debería ser la uniforme.

Jeffreys fue un poco más lejos al proponer una generalización para el caso de varios parámetros. En particular, abordó los problemas que envolvían parámetros de escala y de localización sugiriendo separar los parámetros en dos vectores, es decir, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$ para todos los parámetros de localización, y los restantes en el vector $\boldsymbol{\theta}$. Para el vector $\boldsymbol{\mu}$ asigna la distribución Uniforme y para $\boldsymbol{\theta}$ propone utilizar su regla general (Jeffreys, 1961, p. 182-183). En este caso la propuesta es,

$$\pi(\mu_1, \dots, \mu_k; \boldsymbol{\theta}) \propto (\det(I(\boldsymbol{\theta})))^{-1/2}$$

donde $I(\boldsymbol{\theta})$ se obtiene suponiendo que $\boldsymbol{\mu}$ es conocida. En caso de haber parámetros de escala, éstos deberán incluirse en el vector $\boldsymbol{\theta}$.

Jeffreys además consideró los siguientes casos: para el caso de tener un modelo de probabilidad que envuelva parámetros de localización, Jeffreys sugirió una versión modificada de su regla general (*ib.*, p. 186). Cuando hay parámetros mezclados Jeffreys sugiere trabajar éstos parámetros de manera independiente (*ib.*, p. 185).

Finalmente, han sido los capítulos 3 y 4 de *Theory of Probability* de Jeffreys los que han dado luz a la teoría moderna Bayesiana. Se recomienda ver una discusión reciente de su contenido en Robert, Chopin y Rousseau (2009).

Invarianza

Como se ha discutido anteriormente, una de las propiedades deseables que la distribuciones iniciales deberían de cumplir es la propiedad de la invarianza, y en cierta manera esto ha definido las distintas metodologías de obtención de distribuciones iniciales. Recorde-

mos brevemente: Si con algún método se obtiene una inicial no-informativa $p(\phi)$ y con la misma metodología se obtiene $p(\xi)$ como una distribución inicial para $\xi = \xi(\phi)$, que es una transformación uno-a-uno de ϕ , debe verificarse que $p(\xi)d\xi = p(\phi)d\phi$. De esta manera, el procedimiento de obtención será invariante ante transformaciones uno-a-uno del parámetro.

En Dawid (1983) puede encontrarse un resumen de algunas de las propiedades que tales metodologías deben tener y que mencionamos a continuación:

(i) *Invarianza frente a transformaciones del parámetro (PI).*

Sea $\phi = \phi(\theta)$ una función invertible de $\theta \in \Theta$. Si $\pi_\theta(\theta)$ es una función inicial objetiva para θ y $\pi_\phi(\phi)$ es una función inicial objetiva para ϕ , entonces

$$\pi_\phi(\phi) = \pi_\theta(\theta)|J(\theta)|^{-1}|_{\theta=\theta(\phi)}$$

donde $|J(\theta)| = \det\left(\frac{\partial\phi(\theta)}{\partial\theta}\right)$ es el determinante del Jacobiano de la transformación.

(ii) *Invarianza frente a transformaciones de los datos (DI)*

Considérese el modelo $\{p_X(\cdot|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ y sea $Y = y(X)$ una función invertible de X . Entonces, las funciones iniciales de θ inducidas por los modelos $p_X(\cdot|\theta)$ y $p_Y(\cdot|\theta)$ deberán ser iguales.

De acuerdo a Dawid (1983) estos dos criterios de invarianza (PI y DI) no están relacionados con la *ignorancia* o *el no saber nada*, puesto que también pueden ser usados para distribuciones iniciales subjetivas. Sin embargo, la siguiente característica es la *clave* de la invarianza puesto que formaliza la *ignorancia* como la irrelevancia del contexto.

3. *Invarianza del contexto (CI).*

Para definir una distribución inicial, el contexto del modelo no debería de considerarse. Expresado en términos formales, si $\mathcal{M} = (\mathcal{X}, \Theta, \mathcal{P})$ es un modelo con espacio muestral \mathcal{X} , espacio paramétrico Θ y una familia de distribuciones parametrizadas \mathcal{P} y consideramos otro modelo $\mathcal{M}' = (\mathcal{X}', \Theta', \mathcal{P})$ con la misma familia de distribuciones paramétricas \mathcal{P} , entonces

$$\pi(\theta) = \pi(\theta').$$

Si se aplican los criterios (PI), (DI) y (CI) simultáneamente se imponen importantes restricciones a la obtención de distribuciones iniciales.

Por otro lado, nótese que la regla de Jeffreys $\pi(\theta) \propto |I(\theta)|^{1/2}$, cuando existe, satisface todos estos criterios.

En un sentido más general sobre la propiedad de invarianza, el trabajo de Hartigan (1964) considera las *inversiones*. Una inversión g es una función que asigna una densidad final $g_f(\theta|x)$ a cada familia \mathcal{F} de densidades de $f(x|\theta)$, esto es,

$$g_{\mathcal{F}}(\theta|x) \propto f(x|\theta)h_{\mathcal{F}}(\theta).$$

Si $h_{\mathcal{F}}(\theta)$ es la densidad inicial de cada \mathcal{F} , de acuerdo a Hartigan (1964) $h_{\mathcal{F}}$ debe cumplir, bajo la *inversion*, las siguientes propiedades: (i) invarianza frente a transformaciones de los datos, (ii) invarianza frente a transformaciones de los parámetros, (iii) invarianza de restricción del espacio paramétrico Θ (Θ -restriction invariance), (iv) suficiencia, (v) producto directo y (vi) producto repetido.

Sea la familia de densidades $\mathcal{F} = \{f(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$ y sea z una transformación uno-a-uno diferenciable sobre X y Θ , donde $z : \mathcal{X} \rightarrow \mathcal{X}$, $z : \Theta_1 \rightarrow \Theta_2$, $\Theta_1, \Theta_2 \subseteq \Theta$ tal que

$$f(z(x)|z(\theta)) \left(\frac{dz(x)}{dx} \right) = f(x|\theta)$$

entonces las propiedades (i), (ii) y (iii) implican que para alguna constante c

$$h_{\mathcal{F}}(z(\theta))|J(z^{-1}(\theta))| = ch_{\mathcal{F}}(\theta).$$

Cualquier densidad que satisfaga esta relación es una densidad inicial *relativamente invariante*. De hecho, la clase de densidades iniciales *relativamente invariantes* es más grande que la clase de densidades iniciales invariantes (Hartigan, 1964).

Nótese que las propiedades (i), (ii), (iii) son las mismas que Dawid (1983) pide para las distribuciones iniciales invariantes.

Hartigan (1964) discute las densidades invariantes por la derecha y por la izquierda, señalando que ambas densidades serán miembros diferentes de la familia de densidades invariantes. Además, define las densidades iniciales *localmente invariantes asintóticamente* (ALI, *asymptotically locally invariant*) $\pi_{ALI}(\theta)$, en el caso de un solo parámetro, que son las que satisfacen

$$\frac{d}{d\theta} \log \pi_{ALI}(\theta) = \frac{E_{x|\theta}(f_1 f_2)}{E_{x|\theta}(f_2)}$$

donde $f_1 = \frac{\partial}{\partial \theta} \log f(x|\theta)$ and $f_2 = \frac{\partial^2}{\partial \theta^2} \log f(x|\theta)$. Nótese que $\frac{\partial}{\partial \theta} \log \pi(\theta)$ depende de las propiedades de f_1 y de f_2 . Hartigan discute una generalización para varios parámetros lo que implica la solución de un sistema de ecuaciones diferenciales, que no podría tener solución. En las familias exponenciales, Hartigan (1964) encuentra una forma simple de la densidad inicial relativamente invariante. Además, demuestra que la distribución

inicial de Jeffreys satisface todos los criterios antes mencionados al igual que la densidad inicial ALI.

Resulta interesante la sugerencia que hace Hartigan (1964) para obtener distribuciones iniciales, al combinar las densidades iniciales ALI y la de Jeffreys en la forma de $\pi(\theta) \propto \pi_J^\alpha(\theta)\pi_{ALI}^\beta(\theta)$ con $\alpha + \beta = 1$, donde $\pi_J^\alpha(\theta)$ y $\pi_{ALI}^\alpha(\theta)$ son las iniciales de Jeffreys y ALI respectivamente.

En un trabajo más reciente, Hartigan (1998) propone obtener una distribución inicial (ALI) que maximice la distancia entre la densidad final y la inicial bajo la distancia truncada de Kullback-Leibler, es decir, elegir una inicial ALI será igual a encontrar una estimación máximo verosímil bajo el supuesto de pérdida truncada de Kullback-Leibler. Hartigan (1998) también aborda el problema de predicción y encuentra la densidad predictiva que es asintóticamente cercana, en términos de la distancia de Kullback-Leibler, a la verdadera densidad. Esta propuesta será estudiada ampliamente en el Capítulo 3 de esta Tesis.

Dawid (1983) propone una regla para obtener una inicial invariante en familias exponenciales regulares de la forma

$$f(x|\theta) = \exp\{a(x) + b(\theta) + \phi(\theta)^T t(x)\}, \quad \theta \in \tilde{\Theta}$$

donde $\tilde{\Theta} = \{\theta : \int \exp\{a(x) + \phi(\theta)^T t(x)\} dx < \infty\}$, $\tilde{\Phi} = \phi(\tilde{\Theta})$ es un subconjunto convexo en \mathcal{R}^k y ϕ es el parámetro canónico. Dawid (1983) sugiere asignar una inicial adecuada a ϕ y entonces usar la propiedad de invarianza del parámetro (PI) para obtener la inicial objetiva en cualquier parametrización de θ . Por ejemplo, la distribución inicial sobre $\tilde{\Phi}$ se asigna dependiendo de su dominio $\tilde{\Phi}$ y no de otra característica del modelo. Si $\tilde{\Phi} = \mathcal{R}^k$, la inicial deberá ser relativamente invariante bajo transformaciones que preservan $\tilde{\Phi}$, y la distribución uniforme $\pi(\phi) \propto 1$ tiene esta propiedad, la cual coincide con la inicial ALI de Hartigan.

Por otro lado, un *grupo de transformaciones* que actúa sobre el espacio muestral puede actuar de manera diferente en el espacio paramétrico. Consecuentemente, deberán considerarse dos grupos de transformaciones, uno que actúe sobre el espacio muestral y el otro sobre el espacio paramétrico. En muchos de los casos, es mejor suponer que es un solo grupo el que actúa de manera diferente sobre los espacios paramétrico y muestral.

Si un modelo es invariante bajo la acción de algún grupo G , la medida de Haar por la izquierda y por la derecha sobre G proporcionan distribuciones iniciales sobre Θ (espacio paramétrico), y son invariantes (o relativamente invariantes) sobre la acción del grupo G . En general, existen varias razones para elegir la medida de Haar por la derecha más que la medida de Haar por la izquierda. Un ejemplo es que la medida de Haar por

la derecha arroja la mejor decisión invariante en problemas de decisión (Berger, 1985). Pueden encontrarse algunas razones más en Kass y Wasserman (1996).

Sin embargo, los requisitos de invarianza no son suficientes para garantizar un buen comportamiento de las metodologías de obtención de distribuciones iniciales objetivas. Por esa razón se agregan dos importantes requisitos que son:

- La no existencia de la *inconsistencia fuerte*

La distribución final deberá concentrarse alrededor del verdadero valor del parámetro conforme los datos se van acumulando. Si esto no ocurre se tiene una inconsistencia (fuerte). En Stone (1976) pueden encontrarse dos ejemplos de inconsistencia fuerte que son resultado de usar la inicial uniforme “no informativa”. Posteriormente, en la Sección 2.4 se discutirá esta propiedad.

- La no existencia de *paradojas de marginalización*

Estas paradojas pueden aparecer cuando se utilizan iniciales impropias y el espacio paramétrico es multidimensional. Brevemente, supóngase el modelo $p(x|\theta, \lambda)$ y la distribución inicial conjunta $\pi(\theta, \lambda)$. También supóngase que la densidad final de θ satisface $\pi(\theta|x) = \pi(\theta|z(x))$ para alguna función $z(x)$ de los datos, y que la densidad en el muestreo de z es tal que $f(z|\theta, \lambda) = f(z|\theta)$, por tanto depende únicamente de θ . Debería ser entonces posible obtener la densidad final $\pi(\theta|x)$ de $f(z|\theta)$ y de alguna inicial $\pi(\theta)$; esto siempre sucede si $\pi(\theta, \lambda)$ es propia, pero no siempre es así cuando $\pi(\theta, \lambda)$ es impropia. Véase la subsección 2.4.4 de esta Tesis donde se ejemplifica lo anterior con el coeficiente de variación del modelo Normal. Véase también Dawid, Stone y Zidek (1973) y los ejemplos ahí mencionados.

Además de todas las referencias bibliográficas mencionadas, una discusión más reciente de las propiedades de invarianza de algunas distribuciones iniciales puede encontrarse en Datta y Ghosh (1996).

1.3. Verosimilitud de datos trasladados.

La función de verosimilitud es de *datos trasladados* en términos de $\phi(\theta)$, que es una transformación uno-a-uno de θ , si se puede expresar como

$$l(\theta|\mathbf{y}) = g[\phi(\theta) - f(\mathbf{y})] \tag{1.1}$$

donde $g(\cdot)$ es una función independiente de los datos \mathbf{y} y $f(\mathbf{y})$ es una función de \mathbf{y} . Si esto es posible, es decir, la verosimilitud puede escribirse en la forma (1.1), entonces puede considerarse a la distribución inicial uniforme sobre $\phi(\theta)$ como no-informativa.

La idea de las verosimilitudes de datos trasladados fue propuesta por Box y Tiao (1973). Ellos supusieron que es posible expresar que *sabemos muy poco* acerca de θ en términos de $\phi(\theta)$, por lo que la verosimilitud es determinada por su localización.

Dado que para muestras de tamaño moderado no siempre existe una transformación exacta que exprese la verosimilitud en la forma (1.1), Box y Tiao (1973) proponen buscar una transformación sobre θ donde la verosimilitud sea aproximadamente de datos trasladados, y por lo tanto la verosimilitud será *casi* independiente de los datos excepto por su localización.

Es interesante mencionar que este método proporciona otra justificación para el uso de la regla general de Jeffreys. En efecto, suponiendo una muestra de n observaciones aleatorias, la aproximación normal de la verosimilitud es

$$l(\theta|\mathbf{y}) \simeq N(\theta|\hat{\theta}, \hat{\sigma}_y^2)$$

donde $\hat{\theta}$ es el estimador máximo verosímil $\hat{\sigma}_y^2 = (nI(\theta))^{-1}$ y $I(\theta)$ es la matriz de información de Fisher. Para estabilizar la varianza es necesario una reparametrización tal que para alguna constante c se tenga

$$l(\phi|\mathbf{y}) \simeq N(\phi|\hat{\phi}, \frac{c}{n}),$$

pero la varianza de $\phi(\theta)$ es $\text{Var}(\theta)[\phi'(\theta)]^2 = I(\hat{\theta})^{-1}[\phi'(\theta)]^2$ y será constante si sucede $\phi'(\theta) = I(\theta)^{1/2}$.

Una distribución inicial uniforme $\pi(\phi) = 1$ para ϕ implicará la inicial $\pi(\theta) = \pi(\phi)\phi'(\theta) = I(\theta)^{1/2}$ para θ , que es la regla de Jeffreys.

Kass (1990) propone una modificación, que usa una función $h_a(\cdot)$ que depende de un estadístico auxiliar (pivote), es decir, $h_a(\cdot)$ dependerá de los datos a través de a . Esta extensión de la definición es conocida como *grupo teórico* y se cumple en modelos de transformación general. Kass (1990) analiza las propiedades de la verosimilitud de datos trasladados, que son:

- *Aproximación local.* Para muestras independientes e idénticamente distribuidas, las funciones de verosimilitud de datos trasladados son localmente aproximadas hasta un orden $O(n^{-1})$.
- *Densidades posteriores aproximadas.* Si se usa una distribución inicial uniforme, la densidad final puede considerarse como asintóticamente local de datos trasladados hasta un orden $O(n^{-1})$.
- *Aproximación local multiparamétrica.* Si se usan métricas de información geodésicas,

la verosimilitud puede considerarse como de datos trasladados hasta un orden $O(n^{-1})$.

La extensión multiparamétrica de Kass (1990) es poco sólida (Kass y Wasserman, 1996) puesto que no es posible encontrar una verosimilitud de datos trasladados conjunta.

Finalmente, para más detalles sobre verosimilitud de datos trasladados véase Box y Tiao (1973), Kass (1990) y el resumen de Kass y Wasserman (1996).

1.4. Máxima entropía

Si $\Theta = \{\theta_1, \dots, \theta_k\}$ es finito y π es una medida de probabilidad sobre Θ , entonces la entropía de π , que puede considerarse como una medida de la incertidumbre sobre θ implicada por π , se define como

$$H(\pi) = - \sum_{i=1}^k \pi(\theta_i) \log \pi(\theta_i).$$

La función inicial de mayor entropía es la menos informativa, y el *método de máxima entropía* consiste en elegir la inicial que maximice la entropía. Cuando no se tienen restricciones, la inicial que maximiza es la distribución inicial uniforme. Pero cuando se cuenta con información especificada por las esperanzas de un conjunto de funciones de las variables aleatorias $\{E[f_1], \dots, E[f_r]\}$ con $r < k$, entonces el método sugiere la inicial que maximiza la entropía sujeta a las restricciones impuestas por estos momentos, lo que da como resultado

$$\pi(\theta_i) \propto \exp \left\{ \sum_{j=1}^r \lambda_j f_j(\theta_i) \right\}.$$

Jaynes (1968) fue el mayor impulsor de los métodos de máxima entropía para la obtención de distribuciones iniciales objetivas. Él asegura que el método de máxima entropía es una herramienta formal que garantiza que no se han considerado supuestos erróneos, que podrían permanecer ocultos, y que se consideran todas las posibilidades que la información disponible aporta en ese momento. Demuestra que la unicidad de la expresión $-p \log p$ es consecuencia de su correspondencia con las frecuencias, es decir, que la inicial resultante de maximizar la entropía ha de concordar con la distribución de frecuencias proveniente de resultados experimentales realizados en un gran número de veces. Otras funciones convexas pueden usarse para definir la incertidumbre, pero Jaynes asegura que carecerían de una correspondencia frecuentista como la descrita anteriormente.

La versión continua de la máxima entropía carece de la propiedad de invarianza y Jaynes (1968) sugiere usar la forma invariante $-\int p(\theta) \log \frac{p(\theta)}{\pi(\theta)} d\theta$ que requiere la especi-

ficación de una inicial “no informativa” $\pi(\theta)$, de tal suerte que el argumento se vuelve circular.

Finalmente, Seidenfeld (1987) además de aportar un resumen sobre los métodos de máxima entropía, describe dos de los problemas asociados a ellos:

- *Conflictos con la actualización Bayesiana.* Supóngase que existe la inicial P_0 tal que maximiza la entropía dada una restricción, y se dé la ocurrencia del suceso A . Existen dos formas para incluir esta nueva información, (i) condicionando P_0 para tener $P_0(\cdot|A)$, o (ii) incluyendo el suceso como una restricción, esto es, $E[I_A]$ donde I_A es la función indicadora para el suceso A , pero la nueva probabilidad resultante Q que maximiza la entropía difiere de $P_0(\cdot|A)$.
- *Paradojas de partición.* La máxima entropía puede incurrir en paradojas de partición al igual que el *principio de razón insuficiente*. Las inconsistencias aparecen cuando el espacio paramétrico es subdividido en presencia y ausencia de restricciones empíricas sobre las cantidades de interés.

Para más detalles sobre este método de obtención de distribuciones iniciales objetivas véase Seidenfeld (1987) y el resumen dado por Kass y Wasserman (1996) junto con las referencias ahí citadas.

1.5. Geometría

Jeffreys (1961) introduce el cuadrado de las distancias como una aproximación a las distancias entre distribuciones. De hecho, encuentra que la divergencia de Kullback-Leibler se comporta localmente como el cuadrado de la función distancia determinada por la métrica Riemanniana, que es la información de Fisher $|I(\theta)|^{1/2}$. Su idea se basó principalmente en la propiedad de invarianza dado que los volúmenes provenientes de métricas Riemannianas son invariantes ante reparametrizaciones.

En esta misma línea de investigación, Kass (1989) proporciona una interpretación geométrica de una medida local de la distancia entre elementos de una familia de distribuciones, la cual es la divergencia Kullback-Leibler o equivalentemente, la información de Fisher. Kass define la *distancia de información* como la distancia entre dos elementos de distribuciones. Demuestra también que esta distancia no depende de las parametrizaciones y que existe una relación entre esta distancia de información y la divergencia Kullback-Leibler. Encuentra que la divergencia Kullback-Liebler tiene un comportamiento local como el del cuadrado de la función distancia, coincidiendo con el razonamiento inicial de Jeffreys.

Nótese que la regla general de Jeffreys es generada por los elementos naturales del volumen de la métrica de información, además es invariante e uniforme respecto de al-

gunos espacios paramétricos, como la medida de Lebesgue, puesto que asigna igual masa a regiones que tienen igual volumen. Esta propiedad la hace atractiva en el momento de elegir una inicial objetiva.

1.6. Iniciales de cobertura consistente

Desde finales de los años cincuenta, uno de los aspectos buscados de la obtención de distribuciones iniciales es que sus respectivas distribuciones posteriores concuerden con los resultados frecuentistas. Expresado en términos formales, considérese un parámetro escalar θ y las funciones cotas superior e inferior de los datos $l(x)$ y $u(x)$ que satisfacen $\Pr[l(x) < \theta < u(x)] = 1 - \alpha$, así que $A_x = (l(x), u(x))$ es un conjunto creíble (de probabilidad) final $1 - \alpha$, con $\int_{A_x} \pi(\theta|x)d\theta = 1 - \alpha$. Ahora bien, pueden entonces considerarse las propiedades frecuentistas de A_x bajo el muestreo repetido dado θ . En general, para algún tipo de distribuciones iniciales, la proporción de muestras que producirán conjuntos A_x que contengan al verdadero valor de θ (la cobertura frecuentista) no será exactamente $1 - \alpha$, pero puede encontrarse una aproximación a la misma.

Mencionaremos a continuación algunos de los principales puntos sobre esta línea de investigación:

Lindley (1958) encuentra un acuerdo exacto entre las distribuciones finales y los intervalos de confianza provenientes de distribuciones fiduciales para un parámetro escalar, cuando el modelo de probabilidad pertenece a la familia de distribuciones de localización o puede transformarse en un modelo de localización. Más aún, si existe un estadístico suficiente para el parámetro y se preserva la forma de localización, también hay un acuerdo exacto. Lindley (1958) incluye resultados sobre la familia exponencial, en particular las distribuciones Normal y Gama, que son distribuciones pertenecientes a familias de localización.

Welch y Peers (1963) proporcionan una generalización de este resultado para el caso uniparamétrico, donde los intervalos Bayes asociados de un solo lado se aproximan asintóticamente a intervalos de confianza, y encuentran que la aproximación es mejor cuando se usa la inicial de Jeffreys.

Finalmente, un método de obtención de iniciales objetivas que proporcionen cierto acuerdo con los resultados frecuentistas son las iniciales de cobertura consistente o *matching*. Una inicial de cobertura consistente es aquella inicial bajo la cual las probabilidades finales de ciertas regiones coinciden con sus coberturas de probabilidad, exacta o aproximadamente.

Nótese que una inicial de cobertura consistente no es necesariamente la mejor opción, como demuestran importantes ejemplos. Uno de ellos es el problema de Fieller-Creasy, donde los intervalos de confianza frecuentista no funcionan y en donde una inicial objetiva

Bayesiana no debería reproducirlos.

Un resumen completo sobre estas iniciales objetivas puede verse en Datta y Sweeting (2005).

Por otro parte, las propiedades frecuentistas pueden ser usadas como reglas de selección. En esta línea Berger y Bernardo (1989) proponen el uso de las propiedades de cobertura frecuentista de los intervalos creíbles para discriminar entre posibles distribuciones iniciales objetivas alternativas.

1.7. Método de Zellner

Zellner (1977) propone la *inicial de máxima información de datos* (Maximal Data Information Prior) basada en medidas de información. Zellner elige la inicial usando una función \mathcal{G} como función criterio u objetivo para ser maximizada, sujeta no únicamente a la condición $\int \pi(\theta)d\theta = 1$ sino también a otras posibles condiciones relevantes.

Expresado en términos formales, considérese una observación y proveniente de $p(y|\theta)$, con parámetro escalar θ tal que $a \leq \theta \leq b$, con a, b finitos. Zellner define la *información de los datos* como $\mathcal{I} = \int p(y|\theta) \log p(\theta) dy$ y el *valor promedio de la información a priori* de los datos como

$$\bar{\mathcal{I}}(\theta) = \int \mathcal{I}(\theta)\pi(\theta)d\theta.$$

La inicial *MDIP* $\pi(\theta)$ es la inicial que maximiza

$$\mathcal{G} = \bar{\mathcal{I}}(\theta) - \int \pi(\theta) \log \pi(\theta) d\theta,$$

cuya solución es $\pi(\theta) \propto \exp\{\mathcal{I}(\theta)\}$. Ésta será la inicial que maximiza la entropía sujeta a la restricción de que la entropía en los datos deberá ser constante.

Desafortunadamente, el método propuesto por Zellner no es invariante ante parametrizaciones y por lo tanto no puede proponerse como un procedimiento general de obtención de iniciales. Sin embargo puede forzarse y lograr una invarianza parcial en cierta forma *ad hoc*. Zellner (1977) incluye un catálogo de iniciales *MDIP* para diferentes modelos de probabilidad.

Además de la referencia aquí citada, puede encontrarse más información sobre este método en Kass y Wasserman (1996, Sec. 3.8) y en las referencias ahí mencionadas.

1.8. Modelos de teoría de la decisión

Otra forma de obtención de distribuciones iniciales no informativas consiste en usar teoría de la decisión. En esta sección haremos una revisión de tres de los principales precursores de este método, Hartigan (1965), Good (1969) y Kashyap (1971), en ese orden.

La propuesta de Hartigan se basa en un problema de decisión con los elementos: Una función de densidad $p(x|\theta)$ y una función de pérdida $L(d, \theta)$ con $d \in D$ un espacio de decisiones.

Para cada densidad inicial $h(\theta)$ existe una decisión de Bayes d , dada la observación x , tal que minimiza la pérdida promedio L con respecto a la densidad final. Hartigan dice que la decisión $d(x)$ es *insesgada* para la función de pérdida L si

$$E[L(d(x), \theta_0)|\theta] \geq E[L(d(x), \theta)|\theta] \quad \forall \theta, \theta_0 \in \Theta$$

La función de pérdida $L(\cdot)$ es interpretada como una medida de incompatibilidad entre d y θ , cuando θ es verdadera d es en promedio más compatible con θ que cualquier otro valor del parámetro.

Si θ es unidimensional, la distribución inicial $h(\cdot)$ es asintóticamente insesgada si

$$h(\theta) = \frac{E \left[\left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right)^2 | \theta \right]}{\sqrt{\frac{\partial^2}{\partial d^2} L(d(x), \theta) |_{d=\theta}}}$$

En particular, para la función de pérdida $L(d, \theta) = \alpha(\theta)(d - \theta)^2$, la inicial insesgada es $\pi_U(\theta) = \frac{I(\theta)}{\alpha(\theta)}$ donde I es la información de Fisher. Si se usa la función de pérdida logarítmica,

$$L(d, \theta) = \int \log \left(\frac{p(x|d)}{p(x|\theta)} \right) p(x|\theta) dx,$$

la inicial insesgada es $\pi_U(\theta) = \sqrt{I(\theta)}$, que coincide con la inicial de Jeffreys. Hartigan también proporciona una extensión a varios parámetros; estudia el caso de la familia exponencial multidimensional y el modelo de regresión.

El trabajo de Good (1969) difiere un poco del de Hartigan, pues Good propone utilizar la función $U(G|F)$ como “la utilidad de afirmar la distribución G , cuando en realidad es F ”. Good demuestra que con respecto a un caso especial de función de utilidad U , la distribución inicial *menos favorable*, es proporcional a $|I(\theta)|^{1/2}$, donde $I(\theta)$ es la matriz de información de Fisher. Para más detalles véase Good (1969).

Otro trabajo interesante es el hecho por Kashyap (1971), que encuentra una expresión de una inicial objetiva al considerar un juego de suma-cero entre dos jugadores (“N” la naturaleza *versus* “S” nosotros) donde nosotros elegimos la inicial que minimiza la divergencia entre la densidad de los datos y la densidad predictiva. Kashyap demuestra que la mejor estrategia del problema minimax es elegir la inicial $\pi(\theta)$ que minimice $E[\log \frac{f(y|\theta)}{\pi(\theta)}]$, donde el valor esperado es evaluado sobre la densidad conjunta de θ y los

datos y . Cuando el tamaño de la muestra crece, la solución minimax se aproxima a la inicial de Jeffreys.

1.9. Método de Rissanen

Al igual que en el caso continuo, también se ha intentado definir una inicial objetiva universal sobre parámetros que solo toman valores enteros $\Theta = \{1, 2, \dots\}$, y que no dependan del modelo asumido.

Jeffreys (1961, p. 238) sugirió la inicial $Q(n) \propto n^{-1} + O(n^{-2})$ cuando se desea estimar un entero positivo, dicha inicial la deduce con el ejemplo de un viajero en la estación de trenes. Un hombre que viaja dentro de un país extranjero tiene que cambiar de tren en cierta estación. El hombre desconoce el número total de vagones y lo único que logra ver es el vagón número 100. Con esta información ¿el hombre que puede inferir acerca del total de vagones? Sólo se tiene el supuesto de que los vagones están enumerados consecutivamente a partir de 1. Aparentemente, la cantidad que se desea estimar (entero positivo) no tiene un límite superior sobre los posibles valores que podría tener. Este ejemplo usó Jeffreys para sugerir la función $Q(n)$ para la estimación de un entero positivo, pero no proporcionó ningún tipo de soporte teórico.

Otro punto de vista completamente diferente es el dado por Rissanen (1983) basado en la teoría de códigos. Explicaremos brevemente su propuesta: Supóngase que se ha construido un código que asigna una cadena binaria a cada entero. Cada código necesita un único prefijo que identificará el principio y final de dicho código. El prefijo ayudará al decodificador a identificar la proporción de código que corresponde a cada uno de los códigos (no existe el símbolo coma). Mientras tanto, alguien elige un entero M (cota superior del entero) de una distribución de probabilidad P , donde P refleja el conocimiento previo (o casi ignorancia). La primera tarea es definir el *conocimiento inicial*, que será un conjunto de distribuciones Ω de P , donde M es el único parámetro conocido, y en vez de usar el código se usarán las longitudes de cada código $L = (L(1), L(2), \dots)$, así sólo se tiene un código para todo el conjunto de distribuciones. La segunda tarea es asignar los códigos de tal suerte que el código con la información de las longitudes sea lo más pequeño posible, es decir, minimizar el inverso del *código eficiencia*, el cual es

$$\min_L \sup_P \lim_{N \rightarrow \infty} \frac{\sum_{i=1}^N P(i)L(i)}{-\sum_{i=1}^N P(i) \log P(i)}$$

con algunas condiciones de regularidad.

Rissanen (1983) demostró que existe un código $L_0(n) = \log_2^*(n) + \log_2 c$ donde $\log_2^*(n) =$

$\log_2 n + \log_2 \log_2 n + \dots$, que satisface la condición de minimización y sugiere usar

$$\begin{aligned} Q(n) &\propto 2^{L_0(n)}, \\ &\propto \frac{1}{n} \times \frac{1}{\log_2 n} \times \dots \times \frac{1}{\log_2 n \dots \log_2 n} \times \frac{1}{c}, \quad \text{for } n > 0; \end{aligned}$$

como inicial universal para todos los enteros. Esta inicial también puede extenderse a los enteros negativos.

Sin embargo, la aplicación de esta inicial general frecuentemente da como resultado soluciones insatisfactorias a ciertos ejemplos y uno de ellos es el de datos binomiales $\text{Bi}(r|n, \theta)$ donde n es desconocido.

Capítulo 2

Análisis de referencia

2.1. Introducción

Desde los trabajos de Bayes (1763) y de Laplace (1814/1825) hasta nuestros días, ha adquirido gran relevancia el problema de encontrar un método de obtención de iniciales *no informativas*, tarea a la que le han dedicado gran parte de sus esfuerzos hombres y mujeres de ciencia.

El trabajo de Jeffreys (1946) fue un parteaguas en este tema y es considerado como el precursor de la estadística Bayesiana moderna. El trabajo inicial cristalizó en el libro pionero de Jeffreys (1961) a partir del cual han surgido continuas propuestas tales como las de Zellner (1971a) y Box y Tiao (1973) por mencionar algunas. En el Capítulo 1 se ha proporcionado un breve resumen de algunas de ellos.

Actualmente la investigación en esta área frecuentemente se basa en el entorno del *análisis de referencia* donde una *función inicial de referencia* representa la completa *ignorancia* sobre una cantidad de interés bien especificada. La inicial de referencia produce una distribución final de referencia para dicha cantidad, que siempre será dominada por la información que proveen los datos.

Las funciones iniciales de referencia no representan las creencias personales del investigador, por tanto pueden usarse como un estándar de comunicación en el medio científico. Su elección incorpora dos importantes *desiderata*: La función inicial no debe influir en la información que proporcionan los datos, y su elección se hace de tal forma que las creencias personales del investigador tampoco influyen en la información que los datos proporcionan.

La densidad final de referencia se obtiene mediante el uso formal del teorema de Bayes, y el papel de la función inicial de referencia es el de una simple *herramienta de obtención* pues generalmente es una distribución impropia. Como consecuencia directa de su modo de obtención, la distribución final de referencia es objetiva dado que se basa únicamente en el modelo supuesto y los datos observados. Es una representación

matemática de la *ignorancia* acerca de la cantidad de interés relativa a la información que podrían proporcionar los datos bajo un modelo dado.

En general, se espera que las iniciales objetivas cumplan las propiedades siguientes:

(i) *Generalidad.*

El procedimiento de obtención debe ser general, es decir, debe ser aplicable a cualquier problema de inferencia y sus resultados deben ser apropiados. En particular, la densidad final $\pi(\theta|\mathbf{x})$ debe ser propia para cualquier conjunto de datos \mathbf{x} .

(ii) *Invarianza.*

El método deberá ser invariante bajo cualquier transformación que no modifique el problema. Véase la sección 1.2 donde se explica detalladamente esta propiedad.

(iii) *Marginalización consistente.*

Las densidades finales deben ser las mismas si se han obtenido del modelo completo o de la distribución marginal de un estadístico suficiente. Es particular, si la densidad final $\pi_1(\theta|\mathbf{x})$ se obtuvo del modelo completo $p(\mathbf{x}|\theta, \lambda)$ y tiene la forma $\pi(\theta|\mathbf{x}) = \pi_1(\theta|t)$ donde $t = t(\mathbf{x})$ es algún estadístico de los datos; si t tiene distribución en el muestreo $p(t|\theta)$ que depende únicamente de θ y la densidad final resultante de usar $p(t|\theta)$ es $\pi_2(\theta|t)$, entonces la densidad final $\pi_2(\cdot|t)$ deberá ser igual a $\pi_1(\cdot|\mathbf{x})$.

(iv) *Propiedades del muestro consistente.*

El muestreo repetido de la distribución final debe ser consistente con el modelo, es decir, la distribución final deberá concentrarse alrededor del verdadero valor del parámetro cuando el tamaño de la muestra se incrementa. Se dice que hay una *inconsistencia fuerte* si la densidad final carece de esta propiedad.

Las funciones iniciales de referencia (Bernardo 1979, Berger y Bernardo, 1989, 1992a, 1992b, 1992c, Bernardo, 2005, Berger, Bernardo y Sun, 2009a) parecen ser las únicas que cumplen con todas estas propiedades.

La idea central de las funciones iniciales de referencia es la de maximizar la cantidad de información que los datos proporcionan acerca de lo uno podría esperar sobre una cantidad de interés. Las iniciales de referencia generalmente son impropias y por lo tanto no pueden normalizarse en una función de densidad. Recuérdese que las iniciales de referencia son una herramienta técnica de obtención de finales de referencia a través del uso formal del teorema de Bayes.

La siguiente sección se centrará en los conceptos de discrepancia intrínseca e información esperada, centrales en la teoría de la información, y esenciales en el análisis de referencia.

2.2. Discrepancia intrínseca e información esperada

En teoría de la probabilidad, es frecuente el uso de medidas de divergencia entre las distribuciones de probabilidad, basadas en teoría de la información. Cada una de éstas medidas define un tipo de convergencia y se ha encontrado que describen mejor algunos de los procesos límites existentes en la teoría de la probabilidad y en la inferencia estadística. Una de ellas es la *discrepancia intrínseca* (Bernardo y Rueda, 2002).

Definición 1 (Discrepancia intrínseca). *La discrepancia intrínseca $\delta\{p_1, p_2\}$ entre dos densidades de probabilidad $p_1(\mathbf{x})$ y $p_2(\mathbf{x})$ para la cantidad aleatoria $\mathbf{x} \in \mathcal{X}$ es*

$$\delta\{p_1(\mathbf{x}), p_2(\mathbf{x})\} = \min \left\{ \int_{\mathcal{X}_1} p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x}, \int_{\mathcal{X}_2} p_2(\mathbf{x}) \log \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} d\mathbf{x} \right\}.$$

La discrepancia intrínseca entre dos familias de densidades de probabilidad

$$\mathcal{M}_1 \equiv \{p_1(\mathbf{x}|\phi), \phi \in \Phi\} \quad \text{y} \quad \mathcal{M}_2 \equiv \{p_2(\mathbf{x}|\psi), \psi \in \Psi\}$$

para la cantidad aleatoria $\mathbf{x} \in \mathcal{X}$, se define como

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} = \min_{\phi \in \Phi, \psi \in \Psi} \delta\{p_1(\mathbf{x}|\phi), p_2(\mathbf{x}|\psi)\}.$$

La discrepancia intrínseca es un ejemplo de función de pérdida intrínseca (Robert, 1996) y puede expresarse en términos de la divergencia de Kullback-Leibler como

$$\delta\{p_2, p_1\} = \min \{ \kappa\{p_1|p_2\}, \kappa\{p_2|p_1\} \}$$

donde $\kappa\{p_j|p_i\} = \int_{\mathcal{X}_i} p_i(\mathbf{x}) \log \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} d\mathbf{x}$, con $\mathcal{X}_i \subseteq \mathcal{X}_j$. Consecuentemente, la discrepancia intrínseca hereda algunas de las propiedades de la divergencia de Kullback-Leibler. En particular, es no negativa, está bien definida, y es cero sí y sólo si $p_1(\mathbf{x}) = p_2(\mathbf{x})$ casi por todas partes. Nótese que $\kappa(p_1|p_2)$ es el valor esperado del logaritmo de la densidad (o probabilidad) de la razón de p_1 en contra de p_2 . De este modo, la discrepancia intrínseca entre dos modelos, $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ es la *razón de log-verosimilitudes mínima esperada en favor del modelo (verdadero) del que provienen los datos*. Además, la discrepancia intrínseca entre dos modelos no depende de la parametrización usada para describirlos. Si p_1 y p_2 tienen dominios estrictamente anidados y alguna de las divergencias no es

infinita, la discrepancia intrínseca sigue estando definida. En efecto, si $\mathcal{X}_i \subset \mathcal{X}_j$, entonces $\delta\{p_i, p_j\} = \delta\{p_j, p_i\} = \kappa\{p_j|p_i\}$.

La discrepancia intrínseca también tiene la propiedad de la aditividad de la información, que es consecuencia de que la densidad conjunta de cantidades aleatorias independientes es el producto de sus marginales.

Nótese que la discrepancia intrínseca es una medida de la “distancia” entre dos modelos diferentes para $\mathbf{x} \in \mathcal{X}$. Entonces $\delta\{p_1(\mathbf{x}|\phi), p_2(\mathbf{x}|\psi)\}$ es una medida simétrica (en unidades naturales de información) de lo diferentes que son las densidades de probabilidad $p_1(\mathbf{x}|\phi)$ y $p_2(\mathbf{x}|\psi)$, como función de los parámetros ϕ y ψ , de forma que

$$\delta\{p_1(\mathbf{x}|\phi), p_2(\mathbf{x}|\psi)\} = \delta(\phi, \psi).$$

La discrepancia intrínseca de un punto a un conjunto puede definirse como la distancia de ese punto al elemento más cercano del conjunto. Como cualquier función de pérdida intrínseca es invariante bajo transformaciones uno-a-uno de los parámetros usados para describir los modelos de probabilidad.

Una ventaja de la discrepancia intrínseca es que puede usarse como una función general de pérdida en estadística, porque $\delta(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ mide la pérdida de usar el modelo estimado $p(\mathbf{x}|\hat{\boldsymbol{\theta}})$ como si este fuese el modelo verdadero $p(\mathbf{x}|\boldsymbol{\theta})$. Esto puede usarse en un contexto de pruebas de hipótesis como se propone en Bernardo (1999), Bernardo y Rueda (2002) y Bernardo y Juárez (2003). Específicamente, sea el modelo $p(\mathbf{x}|\boldsymbol{\theta}_0)$ (modelo nulo); uno podría preguntarse lo lejos que se encuentra del verdadero modelo $p(\mathbf{x}|\boldsymbol{\theta})$ en función de $\boldsymbol{\theta}$, que es precisamente lo que mide $\delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}$. El modelo nulo será rechazado si el *estadístico intrínseco* $d(\boldsymbol{\theta}_0|\mathbf{x}) = \int_{\Theta} \delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$ es muy grande. De manera similar, el *estimador intrínseco* $\boldsymbol{\theta}^*$ se define como el que minimiza $d(\tilde{\boldsymbol{\theta}}|\mathbf{x})$ entre todos los $\tilde{\boldsymbol{\theta}} \in \Theta$. La función de pérdida esperada $d(\tilde{\boldsymbol{\theta}}|\mathbf{x})$ también puede usarse para definir las *regiones intrínsecas p-creíbles* $R_p = \{\tilde{\boldsymbol{\theta}} : d(\tilde{\boldsymbol{\theta}}|\mathbf{x}) < d_p^*\}$ donde d_p^* es elegido tal que cumpla con $\int_{\boldsymbol{\theta} \in R_p} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = p$.

Asímismo, la discrepancia intrínseca define un nuevo tipo de convergencia para las distribuciones de probabilidad, la *convergencia intrínseca*, que será usada posteriormente para definir las funciones iniciales de referencia.

Definición 2 (Convergencia intrínseca). *Una sucesión de distribuciones de probabilidad especificada por sus funciones de densidad $\{p_i(\mathbf{x})\}_{i=1}^{\infty}$ se dice que converge intrínsecamente a una distribución de probabilidad con densidad $p(\mathbf{x})$ siempre que la sucesión de sus discrepancias intrínsecas $\{\delta(p_i, p)\}_{i=1}^{\infty}$ converja a cero.*

Supóngase el modelo $\mathcal{M} \equiv \{p(\mathbf{x}|\boldsymbol{\theta}), \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ y una función inicial estrictamente positiva $\pi(\boldsymbol{\theta})$. Si $\pi(\boldsymbol{\theta})$ es impropia, es decir, $\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \infty$, el teorema de Bayes no

puede usarse formalmente para obtener la densidad final $\pi(\boldsymbol{\theta}|\mathbf{x})$,

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

a menos que suceda que $\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, puesto que sólo entonces $\pi(\boldsymbol{\theta}|\mathbf{x})$ será propia. Para garantizar lo anterior, se debe encontrar un adecuado limite de subconjuntos del espacio paramétrico de tal forma las iniciales sean propias en ellos, y por tanto inducirán finales propias.

Considérese nuevamente el modelo y una función estrictamente positiva $\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, tal que, para toda $\mathbf{x} \in \mathcal{X}$, $\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. Una *sucesión de compactos aproximada del espacio paramétrico* es una sucesión creciente de subconjuntos compactos de Θ , $\{\Theta_i\}_{i=1}^{\infty}$, que converge a Θ . Sea $\{\pi_i(\boldsymbol{\theta})\}_{i=1}^{\infty}$ la sucesión de iniciales propias obtenidas de renormalizar $\pi(\boldsymbol{\theta})$ a las Θ_i 's. La correspondiente sucesión de finales $\{\pi_i(\boldsymbol{\theta}|\mathbf{x})\}_{i=1}^{\infty}$, con $\pi_i(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})$, se le llama la *sucesión de finales aproximada* a la final $\pi(\boldsymbol{\theta}|\mathbf{x})$.

Puesto que las iniciales normalizadas $\pi_i(\boldsymbol{\theta})$ son propias, puede esperarse que las finales resultantes convergerán en el sentido de *convergencia intrínseca* a la final $\pi(\boldsymbol{\theta}|\mathbf{x})$. Esto lo asegura el siguiente teorema:

Teorema 1 *Considérese el modelo $\mathcal{M} \equiv \{p(\mathbf{x}|\boldsymbol{\theta}); \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ y una función continua estrictamente positiva $\pi(\boldsymbol{\theta})$ tal que $\int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, para toda $\mathbf{x} \in \mathcal{X}$. Para cualquier sucesión de compactos aproximada del espacio paramétrico, la correspondiente sucesión de finales aproximada converge intrínsecamente a la final $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.*

Demostración. Véase la demostración del Teorema 1 en Berger, Bernardo y Sun (2009a).

□

Sin embargo, para evitar patologías se necesita de una forma más fuerte de convergencia; la correspondiente sucesión de finales $\{\pi_i(\boldsymbol{\theta}|\mathbf{x})\}_{i=1}^{\infty}$ se dice que es *intrínsecamente convergente en sentido fuerte* a la final $\pi(\boldsymbol{\theta}|\mathbf{x})$ si

$$\lim_{i \rightarrow \infty} \int_{\mathcal{X}} \delta\{\pi_i(\boldsymbol{\theta}|\mathbf{x}), \pi(\boldsymbol{\theta}|\mathbf{x})\} p_i(\mathbf{x}) = 0$$

donde $p_i(\mathbf{x}) = \int_{\Theta_i} p(\mathbf{x}|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})d\boldsymbol{\theta}$.

Así, una sucesión de distribuciones finales converge intrínsecamente en sentido fuerte a algún límite, si la discrepancia intrínseca fuerte *esperada* de este límite se aproxima a cero. Por lo tanto, una sucesión de finales convergentes intrínsecamente en sentido fuerte se espera proporcione una buena aproximación a la final límite para cualquier conjunto de datos. Las iniciales objetivas deberían cumplir esta propiedad.

Definición 3 (Función inicial permisible). Una función continua estrictamente positiva $\pi(\boldsymbol{\theta})$ es una función inicial permisible para el modelo

$$\mathcal{M} = \{p(\mathbf{x}|\boldsymbol{\theta}), \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$$

si

(i) para toda $\mathbf{x} \in \mathcal{X}$, $\pi(\boldsymbol{\theta}|\mathbf{x})$ es propia, es decir, $\int_{\Theta} \pi(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} < \infty$;

(ii) para alguna sucesión de compactos aproximada la correspondiente sucesión de finales es convergente intrínsecamente en sentido fuerte a

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

Por definición, las funciones iniciales permisibles siempre producen finales propias.

2.3. Distribuciones de referencia

2.3.1. Modelos uniparamétricos

Un concepto importante en el análisis de referencia es la *información esperada de Shannon*.

Definición 4 (Información esperada). La información que se espera proporcione una observación del modelo $\mathcal{M} \equiv \{p(\mathbf{x}|\theta), \mathbf{x} \in \mathcal{X}, \theta \in \Theta\}$, cuando la inicial para θ es $p(\theta)$ es

$$I\{p(\theta)|\mathcal{M}\} = \int \int_{\mathcal{X} \times \Theta} p(\mathbf{x}|\theta)p(\theta) \log \frac{p(\theta|\mathbf{x})}{p(\theta)} d\mathbf{x}d\theta$$

donde $p(\theta|\mathbf{x}) = p(\mathbf{x}|\theta)p(\theta)/p(\mathbf{x})$ y $p(\mathbf{x}) = \int_{\Theta} p(\mathbf{x}|\theta)p(\theta)d\theta$.

La cantidad de información esperada de $I\{p(\theta)|\mathcal{M}\}$ que proporciona una observación proveniente del modelo \mathcal{M} depende de la inicial $p(\theta)$. Si la inicial contiene una gran cantidad de información acerca de θ poca información proveerán los datos. Considérense k observaciones independientes de \mathcal{M} , para k grande $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, $I\{p(\theta)|\mathcal{M}^k\}$ proporcionará gran cantidad de la información faltante acerca del parámetro de interés θ . Por lo tanto, si $k \rightarrow \infty$, $I\{p(\theta)|\mathcal{M}^k\}$ es la información faltante acerca de θ con $p(\theta)$. Intuitivamente, una función inicial de referencia $\pi(\theta|\mathcal{M}, \mathcal{P})$ será una *función permisible* que maximiza la información faltante acerca de θ entre la clase \mathcal{P} de iniciales compatibles con conocimiento previo de que se disponga sobre el valor de θ .

Existen dos importantes dificultades con la formulación intuitiva de las funciones iniciales de referencia que son: Si el espacio paramétrico es continuo o infinitamente numerable $I\{p(\theta)|\mathcal{M}^k\}$ diverge cuando $k \rightarrow \infty$, puesto que se necesita una cantidad

infinita de información para saber todo acerca de θ . Por otra parte, la información esperada está siempre definida en un conjunto no acotado. Estas dificultades se resuelven con la siguiente definición (Bernardo, 1979, Berger y Bernardo, 1992a, Berger *et al.* 2009a).

Definición 5 (Función inicial de referencia).

Sea el modelo $\mathcal{M} \equiv \{p(\mathbf{x}|\theta), \mathbf{x} \in \mathcal{X}, \theta \in \Theta \subset \mathcal{R}\}$ con parámetro escalar continuo y sea \mathcal{P} una clase de funciones iniciales para θ para las cuales se cumple que $\int_{\Theta} p(\mathbf{x}|\theta)p(\theta)d\theta < \infty$. La función $\pi(\theta) = \pi(\theta|\mathcal{M}, \mathcal{P})$ es una inicial de referencia para el modelo \mathcal{M} dado \mathcal{P} , si existe una función permisible tal que, para cualquier conjunto compacto $\Theta_0 \subset \Theta$,

$$\lim_{k \rightarrow \infty} \{I\{\pi_0|\mathcal{M}^k\} - I\{p_0|\mathcal{M}^k\}\} \geq 0, \quad \text{para toda } p \in \mathcal{P},$$

donde $\pi_0(\theta)$ y $p_0(\theta)$ son, respectivamente, las restricciones renormalizadas de $\pi(\theta)$ y $p(\theta)$ en Θ_0

Las funciones iniciales de referencia tienen las siguientes propiedades:

- *Independencia del tamaño de muestra.*

Si los datos $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ consisten en una muestra aleatoria de tamaño n provenientes del modelo $\mathcal{M} \equiv \{p(\mathbf{y}|\theta), \mathbf{y} \in \mathcal{Y}, \theta \in \Theta\}$, y con inicial de referencia $\pi(\theta|\mathcal{M}, \mathcal{P})$. Entonces $\pi(\theta|\mathcal{M}^n, \mathcal{P}) = \pi(\theta|\mathcal{M}, \mathcal{P})$, para cualquier tamaño de muestra fijo n (Berger *et al.* 2009a, Teo. 4).

- *Compatibilidad con estadísticos suficientes.*

Considérese el modelo $\mathcal{M} = \{p(\mathbf{x}|\theta), \mathbf{x} \in \mathcal{X}, \theta \in \Theta\}$ con estadístico suficiente $\mathbf{t} = \mathbf{t}(\mathbf{x}) \in \mathcal{T}$, y sea $\mathcal{M}_{\mathbf{t}} = \{p(\mathbf{t}|\theta), \mathbf{t} \in \mathcal{T}, \theta \in \Theta\}$ el modelo correspondiente en términos de \mathbf{t} . Entonces $\pi(\theta|\mathcal{M}, \mathcal{P}) = \pi(\theta|\mathcal{M}_{\mathbf{t}}, \mathcal{P})$ (Berger *et al.* 2009a, Teo. 5).

- *Consistencia bajo parametrización.*

Considérese el modelo $\mathcal{M} = \{p(\mathbf{x}|\theta), \mathbf{x} \in \mathcal{X}, \theta \in \Theta\}$ y sea $\phi = g(\theta)$, donde $g : \Theta \rightarrow \Phi$ es una transformación invertible (a trozos) de θ . Entonces $\pi_{\phi}(\phi|\mathcal{M}_{\phi}, \mathcal{P})$ es la densidad inicial inducida por la transformación de $\pi_{\theta}(\theta|\mathcal{M}_{\theta}, \mathcal{P})$ (Berger *et al.* 2009a, Teo. 6). Específicamente, para alguna $c > 0$ y para toda $\phi \in \Phi$,

- (i) $\pi_{\phi}(\phi) = c \pi_{\theta}(g^{-1}(\phi))$, si Θ es discreta; y
- (ii) $\pi_{\phi}(\phi) = c \pi_{\theta}(g^{-1}(\phi))|J_{\phi}|$, si $J_{\phi} = \frac{\partial g^{-1}(\phi)}{\partial \phi}$ existe.

Esta propiedad es conocida también como *invarianza* bajo transformaciones.

Puede encontrarse una detallada explicación de las propiedades antes citadas en las Secciones 5.4 y 5.6 de Bernardo y Smith (1994).

Retomando la definición de la inicial de referencia, ésta tiene sentido si $I\{\pi_0|\mathcal{M}^k\}$ y $\{p_0|\mathcal{M}^k\}$ son finitos, por tanto es necesario definir algunas condiciones que la inicial deberá cumplir: Si $p \in \mathcal{P}_s$, donde \mathcal{P}_s es la clase de funciones continuas estrictamente positivas en Θ , entonces

$$(i) \quad \forall \theta \in \Theta, p(\theta) > 0,$$

y

$$(ii) \quad \forall \mathbf{x} \in \mathcal{X}, \int_{\Theta} p(\mathbf{x}|\theta)p(\theta)d\theta < \infty.$$

Las iniciales que cumplan con estas dos condiciones se llamarán *funciones iniciales estándar*. También, será necesario definir los *modelos estándar*, que básicamente son aquellos que dan lugar a una información esperada finita proveniente de cualquier muestra finita.

Expresado en términos formales, el modelo $\mathcal{M} \equiv \{p(\mathbf{x}|\theta); \mathbf{x} \in \mathcal{X}, \theta \in \Theta\}$ es un *modelo estándar* si para cualquier función inicial $p(\theta) \in \mathcal{P}$ y cualquier conjunto compacto de Θ_0 ,

$$I\{p_0|\mathcal{M}^k\} < \infty \tag{2.1}$$

donde p_0 es la inicial resultante de restringir $p(\theta)$ a Θ_0 . Es importante verificar las condiciones bajo las que (2.1) se satisface.

- Cuando todas $p(\mathbf{t}_k|\theta)$, $\theta \in \Theta_0$ tienen el mismo dominio, la Ec. (2.1) se satisface si, para cualquier $\theta, \theta' \in \Theta_0$,

$$\int_{\mathcal{T}_k} p(\mathbf{t}_k|\theta) \log \frac{p(\mathbf{t}_k|\theta)}{p(\mathbf{t}_k|\theta')} d\mathbf{t}_k < \infty,$$

Berger *et al.* (2009a, Lema 1).

- Cuando todas $p(\mathbf{t}_k|\theta)$, $\theta \in \Theta_0$ tienen dominios diferentes, la Ec. (2.1) se satisface si

$$(i) \quad H[p(\mathbf{t}_k|\theta)] \equiv - \int_{\mathcal{T}_k} p(\mathbf{t}_k|\theta) \log(p(\mathbf{t}_k|\theta)) d\mathbf{t}_k \text{ es acotada inferiormente para } \theta \in \Theta_0, \text{ y}$$

$$(ii) \quad \int_{\mathcal{T}_k} p_0(\mathbf{t}_k) \log(p_0(\mathbf{t}_k)) d\mathbf{t}_k > -\infty, \text{ donde } p_0(\mathbf{t}_k) \text{ es la verosimilitud marginal proveniente de la inicial uniforme, es decir,}$$

$$p_0(\mathbf{t}_k) = L(\Theta)^{-1} \int_{\Theta_0} p(\mathbf{t}_k|\theta) d\theta,$$

donde $L(\Theta_0)$ es la medida de Lebesgue de Θ_0 , véase Berger *et al.* (2009a, Lema 2).

El siguiente teorema proporciona una expresión explícita de la función inicial de referencia.

Teorema 2 (Forma explícita de la función inicial de referencia). *Supóngase el modelo estándar $\mathcal{M} \equiv \{p(\mathbf{x}|\theta), \mathbf{x} \in \mathcal{X}, \theta \in \Theta \subset \mathcal{R}\}$, la clase estándar \mathcal{P}_s de iniciales candidatas, y sea $\mathbf{t}_k = \mathbf{t}_k(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathcal{T}_k$ cualquier estadístico suficiente (que puede ser la muestra completa). Sea $\pi^*(\theta)$ una función continua estrictamente positiva tal que la final (formal) correspondiente*

$$\pi^*(\theta|\mathbf{x}) = \frac{p(\mathbf{t}_k|\theta)\pi^*(\theta)}{\int_{\Theta} p(\mathbf{t}_k|\theta)\pi^*(\theta)d\theta}$$

es propia y asintóticamente consistente, y defínase, para cualquier punto interior $\theta_0 \in \Theta$,

$$f_k(\theta) = \exp \left\{ \int_{\mathcal{T}_k} p(\mathbf{t}_k|\theta) \log \pi^*(\theta|\mathbf{t}_k) d\mathbf{t}_k \right\}, \text{ y} \quad (2.2)$$

$$f(\theta) = \lim_{k \rightarrow \infty} \frac{f_k(\theta)}{f_k(\theta_0)}. \quad (2.3)$$

Si, (i) cada $f_k(\theta)$ es continua y, para cualquier θ fija y k suficientemente grande, $\left\{ \frac{f_k^0(\theta)}{f_k^0(\theta_0)} \right\}$ es monótonica en k o es acotada superiormente por alguna función $h(\theta)$ integrable sobre cualquier conjunto compacto; y (ii) $f(\theta)$ es una función inicial permisible, entonces $\pi(\theta|\mathcal{M}, \mathcal{P}_s) = f(\theta)$ es una inicial de referencia para el modelo \mathcal{M} y la clase de iniciales \mathcal{P}_s .

Demostración. Véase la demostración del Teorema 7 de Berger *et al.* 2009a. □

Nótese que π_0 y θ_0 pueden ser elegidas arbitrariamente, además no se especificaron conjuntos compactos, por tanto las iniciales de referencia funcionan, simultáneamente, para todos los subconjuntos compactos de Θ .

Una función inicial permisible para cualquier modelo (estándar) de localización $\mathcal{M}_1 \equiv \{f_1(\mathbf{x} - \mu); \mathbf{x} \in \mathcal{X}, \mu \in \mathcal{R}\}$ es la inicial uniforme $\pi(\mu) \propto 1$. Similarmente, si es permisible, la inicial de referencia para un modelo de escala

$$\mathcal{M}_2 = \{\sigma^{-1}f_2(\mathbf{x}/\sigma); \mathbf{x} \in \mathcal{X}, \sigma \in \mathcal{R}^+\}$$

es $\pi(\sigma) \propto \sigma^{-1}$.

Un ejemplo más sofisticado es el que provee el modelo general uniforme, que es discutido ampliamente en Berger *et al.* (2009a).

Considérese el modelo

$$\mathcal{M} = \left\{ \text{Un}(x|a_1(\theta), a_2(\theta)); \frac{1}{a_2(\theta) - a_1(\theta)}, a_1(\theta) < x < a_2(\theta) \right\} \quad (2.4)$$

donde $\theta > \theta_0$ y $0 < a_1(\theta) < a_2(\theta)$, $\Theta = (\theta_0, \infty)$. Defínase

$$b_k \equiv b_k(\theta) = \frac{a_2'(\theta) - a_1'(\theta)}{a_k'(\theta)}, \quad k = 1, 2.$$

La inicial de referencia para el modelo (2.4) resulta ser

$$\pi(\theta) = \frac{a_2'(\theta) - a_1'(\theta)}{a_2(\theta) - a_1(\theta)} \exp \left\{ b_1 + \frac{1}{b_1 - b_2} (b_1 \psi(b_1^{-1}) - b_2 \psi(b_2^{-1})) \right\}$$

donde $\psi(z)$ es la función digama definida por $\psi(z) = \frac{d}{dz} \log \Gamma(z)$ para $z > 0$, y las derivadas $a_1'(\theta)$ y $a_2'(\theta)$ satisfacen $0 < a_1'(\theta) < a_2'(\theta)$.

Funciones iniciales de referencia numéricas

En algunos modelos complejos las funciones iniciales de referencia pueden ser de difícil obtención, pero puede aproximarse la Ec. (2.2) a través de evaluaciones numéricas mediante el siguiente pseudo-código:

(i) Valores iniciales:

elíjase un valor moderado de k ;

elíjase una función positiva arbitraria $\pi^*(\theta)$, por ejemplo $\pi^*(\theta) = 1$;

elíjase el número m de muestras a simular.

(ii) Para cualquier θ dado, **repítase**, para $j = 1, \dots, m$:

simúlese una muestra aleatoria $\{\mathbf{x}_{1j}, \dots, \mathbf{x}_{kj}\}$ de tamaño k proveniente de $p(\mathbf{x}|\theta)$;

calcúlese (numérica o analíticamente) la integral

$$c_j = \int_{\Theta} \prod_{i=1}^k p(\mathbf{x}_{ij}|\theta) \pi^*(\theta) d\theta;$$

evalúese $r_j(\theta) = \log \left(\prod_{i=1}^k \frac{p(\mathbf{x}_{ij}|\theta) \pi^*(\theta)}{c_j} \right)$.

(iii) Calcúlese $\pi(\theta) = \exp\{m^{-1} \sum_{j=1}^m r_j(\theta)\}$ y **almacénese** el par $(\theta, \pi(\theta))$.

(iv) **Repítanse** las rutinas (ii) y (iii) para todos los θ valores para los cuales se requiera el par $(\theta, \pi(\theta))$.

En caso de necesitar una aproximación analítica a $\pi(\theta)$, esta puede obtenerse por medio de alguna técnica de interpolación.

Al final de esta subsección reproducimos un ejemplo de Bernardo (2005a) donde se implementa esta aproximación numérica.

Funciones iniciales de referencia bajo condiciones de regularidad

Sea la muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_k\}$ proveniente del modelo uniparamétrico $\mathcal{M} \equiv \{p(x|\theta); x \in \mathcal{X}, \theta \in \Theta \subset \mathcal{R}\}$ y supóngase que existe un estimador suficiente asintóticamente consistente $\tilde{\theta}_k$ de θ . La inicial de referencia de θ puede obtenerse entonces en términos de una aproximación asintótica $\pi(\theta|\tilde{\theta})_n$ a la distribución final de θ . Específicamente,

$$\pi(\theta) \propto \pi(\theta|\tilde{\theta}_n)|_{\tilde{\theta}_n=\theta}.$$

Funciones iniciales de referencia bajo normalidad asintótica

Supóngase que la distribución final de θ , la cantidad de interés, es asintóticamente normal. Específicamente, sea $\{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ una muestra aleatoria proveniente del modelo $\mathcal{M} = \{p(\mathbf{y}|\theta); \mathbf{y} \in \mathcal{Y}, \theta \in \Theta\}$ y supóngase que la densidad final de θ , $\pi(\theta|\mathbf{y}_1, \dots, \mathbf{y}_k)$, es asintóticamente normal con desviación estándar $s(\tilde{\theta}_k)/\sqrt{k}$, donde $\tilde{\theta}_k$ es un estimador consistente de θ . Entonces, la inicial de referencia tiene la forma

$$\pi(\theta|\mathcal{M}, \mathcal{P}_0) \propto s(\theta)^{-1}.$$

Bajo condiciones de regularidad (Bernardo y Smith, 1994, p.314) la distribución final de referencia de θ es asintóticamente normal con varianza $i(\hat{\theta}_k)^{-1}/k$ donde $\hat{\theta}_k$ es el estimador máximo verosímil de θ y

$$i(\theta) = - \int_{\mathcal{Y}} p(\mathbf{y}|\theta) \frac{\partial^2}{\partial \theta^2} \log p(\mathbf{y}|\theta) d\mathbf{y}$$

es la información de Fisher. Si este es el caso, y $i(\theta)^{1/2}$ es una función inicial permisible, la inicial de referencia coincide con la inicial de Jeffreys $\pi(\theta) \propto i(\theta)^{1/2}$.

Las funciones iniciales de referencia y el principio de verosimilitud

Nótese que las funciones iniciales de referencia dependen del modelo asumido. Por consiguiente, si el modelo cambia, cambiará la inicial de referencia. Esto genera una incompatibilidad con el *principio de verosimilitud* el cual dice que toda evidencia acerca de θ proporcionada por los datos, se encuentra contenida en la función de verosimilitud.

Sin embargo, nótese que el principio de verosimilitud se aplica *después* de haber observado los datos mientras que las iniciales de referencia se obtienen *antes* de que

los datos se han observado (véase el Ej. 12, Bernardo, 2005a.)

Como se ha mencionado anteriormente, las iniciales de referencia son simples herramientas de obtención de distribuciones finales y no tienen una interpretación probabilística. Las distribuciones finales de referencia resultantes resumen lo que se podría decir acerca de la cantidad de interés si la información *a priori* fuese mínima relativa a la información que el muestreo repetido de los datos pudiera proveer, datos provenientes de un modelo *específico* \mathcal{M} .

Funciones iniciales de referencia restringidas

En caso de que la inicial deba cumplir algunas restricciones será necesario también restringir la clase de iniciales de referencia propias \mathcal{P}_0 , con dominio Θ , a una clase más pequeña \mathcal{P} .

Considérese la clase restringida de iniciales continuas propias con dominio Θ definidas por

$$\mathcal{P} = \left\{ p_\theta : \int_{\Theta} p(\theta) d\theta = 1, \int_{\Theta} g_i(\theta) p(\theta) d\theta = \beta_i, i = 1, \dots, m \right\}$$

donde $E(g_i(\theta)) = \beta_i, i = 1, \dots, m$ son las restricciones impuestas. Si existe, la inicial de referencia restringida se obtiene en términos de la inicial de referencia no restringida $\pi(\theta|\mathcal{M}, \mathcal{P}_0)$ como

$$\pi(\theta|\mathcal{M}, \mathcal{P}) \propto \pi(\theta|\mathcal{M}, \mathcal{P}_0) \exp \left\{ \sum_{i=1}^m \lambda_i g_i(\theta) \right\}$$

donde $\lambda_i (i = 1, \dots, m)$ son constantes determinadas por las condiciones que definen la clase \mathcal{P} .

Datos exponenciales

Se reproduce a continuación el ejemplo del modelo Exponencial $\text{Ex}(x|\theta)$ dado en Bernardo (2005a) para ilustrar los diferentes procedimientos de obtención de la inicial de referencia.

Ejemplo 1 (Datos exponenciales). Sea $\mathbf{x} = \{x_1, \dots, x_n\}$ una muestra aleatoria proveniente del modelo $\mathcal{M} = \{f(x|\theta); x > 0, \theta > 0\}$, donde $f(x|\theta) = \theta e^{-\theta x}$, y $t_n = \sum_{i=1}^n x_i$ es un estadístico suficiente.

- La distribución en el muestreo del estadístico suficiente t_n es una densidad Gama $p(t_n|\theta) = \text{Ga}(t_n|n, \theta)$, puesto que $\text{Ex}(y|\theta) = \text{Ga}(y|1, \theta)$ y $s_y = \sum_{i=1}^n y_i \sim \text{Ga}(s_y|n\alpha, \beta)$. Siguiendo el Teorema 2, con función $h(\theta)$ constante, la Ec.

(2.2) ahora es

$$\begin{aligned}
f_n(\theta) &= \exp \left\{ \int_0^\infty \text{Ga}(t_n|n, \theta) \log \frac{\text{Ga}(t_n|n, \theta)h(\theta)}{\int_0^\infty \text{Ga}(t_n|n, \theta)h(\theta)d\theta} dt_n \right\} \\
&= \exp \left\{ \int_0^\infty \text{Ga}(t_n|n, \theta) \log \text{Ga}(\theta|n+1, t_n) \right\} \\
&= \exp \{-\log \Gamma(n+1) - n + (n+1)\psi(n) - \log \theta\} = c_n \frac{1}{\theta}
\end{aligned}$$

donde $\psi(z)$ es la función digama, $\psi(z) = \frac{d}{dz} \log \Gamma(z)$ para $z > 0$, y c_n es una función que no depende de θ . Usando la Ec. (2.3)

$$f(\theta) = \lim_{n \rightarrow \infty} \frac{f_n(\theta)}{f_n(\theta_0)} = \frac{\theta_0}{\theta}$$

así, entonces $\pi(\theta) \propto \theta^{-1}$, la cual puede demostrarse que es una función inicial permisible.

- Alternativamente, uno puede considerar $\hat{\theta}_n = \frac{1}{\bar{x}}$, $\bar{x} = \sum_{i=1}^n x_i/n$, que es el estimador máximo verosímil de θ , que es consistente. La función de verosimilitud es $l(\theta) = \theta^n e^{-n\theta/\hat{\theta}_n}$, y dada una función inicial continua positiva $h(\theta)$, la aproximación asintótica a la final tiene la forma $\pi(\theta|\mathbf{x}) \propto \theta^n e^{-n\theta/\hat{\theta}_n} h(\theta)$. Si $h(\theta) = 1$, $\pi(\theta|\mathbf{x}) \propto \text{Ga}(\theta|n+1, n/\hat{\theta}_n) \propto \pi(\theta|\hat{\theta}_n)$, bajo muestreo repetido el valor $\hat{\theta}_n$ se concentrará en el verdadero valor θ , por tanto

$$\pi(\theta) \propto \pi(\theta|\hat{\theta}_n)|_{\hat{\theta}_n=\theta} = \frac{1}{\theta}.$$

- Función inicial de referencia numérica.

Usando el algoritmo numérico de obtención de iniciales de referencia descrita anteriormente, obtendremos una aproximación a la inicial de referencia para este modelo. Con $m = 500$ muestras de $k = 100$ observaciones aleatorias, para 10 valores de θ uniformemente log-espaciados en un rango de e^{-3} a e^3 , y reescalados para que $\pi(1) = 1$. La Figura 2.1 muestra la inicial de referencia exacta (línea continua) contra los diez valores aproximados (puntos) de la inicial de referencia numérica.

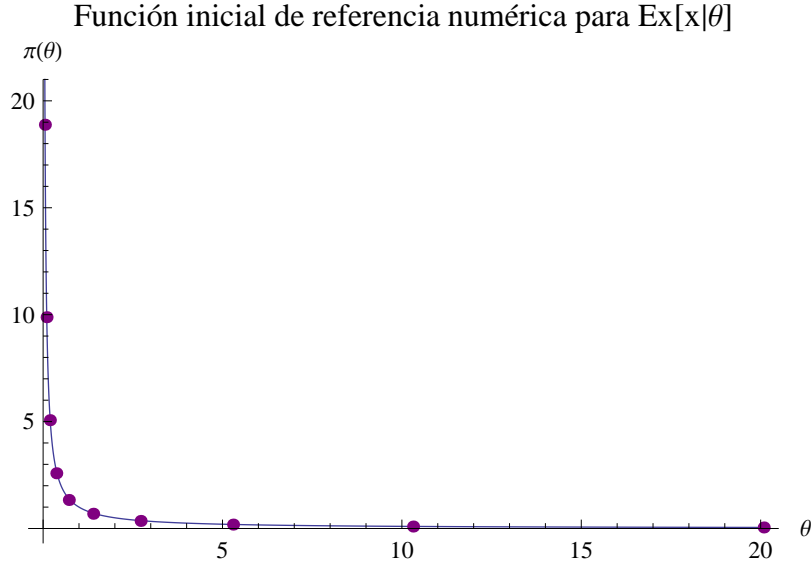


Figura 2.1: Función inicial de referencia numérica para el modelo Exponencial. La línea continua corresponde a la inicial de referencia teórica $\pi(\theta) \propto \theta^{-1}$, los puntos corresponden a la aproximación numérica de la inicial de referencia.

2.3.2. Un parámetro de ruido

Considérese ahora el modelo $\mathcal{M} \equiv \{p(x|\theta, \lambda), x \in \mathcal{X}, \theta \in \Theta, \lambda \in \Lambda\}$ donde θ es el parámetro de interés y λ un parámetro de ruido. Es obvio que para la obtención de la inicial de referencia $\pi(\theta|\mathbf{x})$, dados los datos $\mathbf{x} = \{x_1, \dots, x_n\}$, es necesario obtener la inicial de referencia conjunta $\pi_\theta(\theta, \lambda)$, e integrar con respecto al parámetro de ruido λ , es decir

$$\pi(\theta|\mathbf{x}) = \int_{\Lambda} \pi_\theta(\theta, \lambda|\mathbf{x}) d\lambda \propto \int_{\Lambda} p(\mathbf{x}|\theta, \lambda) \pi_\theta(\theta, \lambda) d\lambda.$$

La obtención de la inicial de referencia puede hacerse aplicando recursivamente el algoritmo para un problema de dos parámetros a dos problemas con un solo parámetro como se describe a continuación:

1. Condicionando sobre θ , $p(\mathbf{x}|\theta, \lambda)$ ahora sólo depende de λ dado θ . Entonces uno puede obtener $\pi(\lambda|\theta)$ con el algoritmo de iniciales de referencia para un parámetro.
2. Dependiendo de la naturaleza de $\pi(\lambda|\theta)$, existen dos formas para obtener $\pi(\theta)$:
 - Si $\pi(\lambda|\theta)$ es propia de forma que $\int_{\Lambda} \pi(\lambda|\theta) d\lambda = 1$, $\pi(\lambda|\theta)$ se usará para definir el modelo integrado de un solo parámetro

$$p(\mathbf{x}|\theta) = \int_{\Lambda} p(\mathbf{x}|\theta, \lambda) \pi(\lambda|\theta) d\lambda \tag{2.5}$$

al que el algoritmo puede aplicarse nuevamente para obtener $\pi(\theta)$.

- Si $\pi(\lambda|\theta)$ tiene una integral infinita sobre Λ , *i.e.*, $\int_{\Lambda} \pi(\lambda|\theta)d\lambda = \infty$, la integración deberá hacerse en una sucesión creciente de subconjuntos $\{\Lambda_i\}_{i=1}^{\infty}$ de Λ , de forma que la restricción de $\pi(\lambda|\theta)$ a Λ_i $\pi_i(\lambda|\theta) = \frac{\pi(\lambda|\theta)}{\int_{\Lambda_i} \pi(\lambda|\theta)d\lambda}$ sea propia. Por lo tanto, la Ec. (2.5) es ahora

$$p_i(x|\theta) = \int_{\Lambda_i} p(\mathbf{x}|\theta, \lambda)\pi_i(\lambda|\theta)d\lambda$$

y da lugar a una sucesión de modelos integrados $\{p_i(x|\theta)\}_{i=1}^{\infty}$, que es proporcional a una sucesión de funciones $\{\pi_i(\theta)\}_{i=1}^{\infty}$.

3. Con $\pi(\lambda|\theta)$ y $\pi(\theta)$, la inicial conjunta es $\pi_{\theta}(\theta, \lambda) = \pi(\lambda|\theta)\pi(\theta)$. El uso formal del teorema de Bayes arroja la final requerida

$$\pi(\theta|\mathbf{x}) \propto \int_{\Lambda} p(\mathbf{x}|\theta, \lambda)\pi_{\theta}(\theta, \lambda)d\lambda = p(\mathbf{x}|\theta)\pi(\theta),$$

cuando $\pi(\lambda|\theta)$ es propia, y una sucesión de θ -finales de referencia $\{\pi_i(\theta|\mathbf{x})\}_{i=1}^{\infty}$, cuando $\pi(\lambda|\theta)$ es impropia en cuyo caso $\pi(\theta|\mathbf{x})$ se obtiene como el límite intrínseco correspondiente

$$\pi(\theta|\mathbf{x}) = \lim_{i \rightarrow \infty} \pi_i(\theta|\mathbf{x}).$$

Invarianza

Las funciones iniciales de referencia, en caso de tener un parámetro de ruido, también preservan la propiedad de invarianza.

Dado el parámetro de interés θ y un parámetro de ruido λ , uno esperaría que la correspondiente inicial de referencia para θ no dependa de cómo es elegido el parámetro de ruido λ .

Teorema 3 (Invarianza con respecto a la elección del parámetro de ruido). *Supóngase los datos \mathbf{x} que han sido obtenidos ya sea de*

$$p(\mathbf{x}|\theta, \lambda), (\theta, \lambda) \in \Theta \times \Lambda \subset \mathcal{R} \times \mathcal{R}, \text{ o de } p(\mathbf{x}|\theta, \phi), (\theta, \phi) \in \Theta \times \Phi \subseteq \mathcal{R} \times \mathcal{R},$$

donde $(\theta, \lambda) \rightarrow (\theta, \phi)$ es una transformación uno-a-uno, con $\phi = g_{\theta}(\lambda)$. Entonces, las finales de referencia para θ dada \mathbf{x} son idénticas.

Demostración. Véase la demostración de la Proposición 5.22 en Bernardo y Smith (1994).

□

Alternativamente, si necesitamos aplicar una transformación al parámetro de interés

$\gamma = g(\theta)$, uno espera que la final del vector (θ, λ) debe ser consistente, con un apropiado elemento Jacobiano, con la final de referencia del vector (γ, λ) .

Teorema 4 (Invarianza bajo transformaciones uno-a-uno). *Supóngase que los datos pueden proceder de $p(\mathbf{x}|\theta, \lambda), (\theta, \lambda) \in \Theta \times \Lambda \subset \mathcal{R} \times \mathcal{R}$, o de $p(\mathbf{x}|\gamma, \phi), (\gamma, \lambda) \in \Gamma \times \Lambda \subseteq \mathcal{R} \times \mathcal{R}$ donde $\gamma = g(\theta)$. Entonces, dados cualesquier datos \mathbf{x} , las finales de referencia para θ y γ están relacionadas por:*

$$(i) \pi_\gamma(\gamma|\mathbf{x}) = \pi_\theta(g^{-1}(\gamma)|\mathbf{x}), \text{ si } \Theta \text{ es discreta; y}$$

$$(ii) \pi_\gamma(\gamma|\mathbf{x}) = \pi_\theta(g^{-1}(\gamma)|\mathbf{x})|J_{g^{-1}}(\gamma)|, \text{ si } J_{g^{-1}}(\gamma) = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \text{ existe.}$$

Demostración. Véase la demostración de la Proposición 5.28 de Bernardo y Smith (1994).

□

Funciones iniciales de referencia bajo normalidad asintótica

Bajo el supuesto de binormalidad asintótica para la distribución final $\pi(\theta, \lambda|\mathbf{x})$, las iniciales de referencia pueden obtenerse en términos de la matriz de información de Fisher:

Teorema 5 (Funciones iniciales de referencia bajo binormalidad asintótica). *Sea la muestra aleatoria $\mathbf{x} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$, consistente de n observaciones independientes provenientes del modelo $\mathcal{M} \equiv \{p(\mathbf{y}|\theta, \lambda), \mathbf{y} \in \mathcal{Y}, \theta \in \Theta, \lambda \in \Lambda\}$*

y sea \mathcal{P}_0 la clase de todas las iniciales conjuntas con dominio $\Theta \times \Lambda$. Si la distribución final de $\{\theta, \lambda\}$ es asintóticamente binormal con matriz de dispersión $V(\hat{\theta}_n, \hat{\lambda}_n)/n$, donde $\{\hat{\theta}_n, \hat{\lambda}_n\}$ son los estimadores consistentes de $\{\theta, \lambda\}$, defínase

$$V(\theta, \lambda) = \begin{pmatrix} v_{\theta\theta}(\theta\lambda) & v_{\theta\lambda}(\theta\lambda) \\ v_{\lambda\theta}(\theta\lambda) & v_{\lambda\lambda}(\theta\lambda) \end{pmatrix}, \quad H(\theta, \lambda) = V^{-1}(\theta, \lambda).$$

Entonces

$$\pi(\lambda|\theta) \propto \sqrt{h_{\lambda\lambda}(\theta, \lambda)}, \quad \lambda \in \Lambda,$$

y, si $\pi(\lambda|\theta)$ es propia,

$$\pi(\theta) \propto \exp \left\{ \int_{\Lambda} \pi(\lambda|\theta) \log \frac{1}{\sqrt{v_{\theta\theta}(\theta, \lambda)}} d\lambda \right\}, \quad \theta \in \Theta. \quad (2.6)$$

Si $\pi(\theta, \lambda)\pi(\theta)$ es una función inicial permisible, la θ -inicial de referencia es entonces

$$\pi(\theta|\mathcal{M}^n, \mathcal{P}_0) \propto \pi(\lambda|\theta)\pi(\theta).$$

Si $\pi(\lambda|\theta)$ no es propia, la integración de la Ec. (2.6) se realiza sobre los elementos de una sucesión creciente de subconjuntos $\{\Lambda_i\}_{i=1}^{\infty}$ tal que $\int_{\Lambda_i} \pi(\lambda|\theta)d\lambda < \infty$, para obtener la sucesión $\{\pi_i(\lambda|\theta)\pi_i(\theta)\}_{i=1}^{\infty}$, donde $\pi_i(\lambda|\theta)$ es la renormalización de $\pi(\lambda|\theta)$ a Λ_i , y la inicial de referencia $\pi_{\theta}(\theta, \lambda)$ se define como su correspondiente límite intrínseco en sentido fuerte.

Demostración. Véase una justificación heurística en Bernardo (2005a, Teo. 11). □

Bajo condiciones de regularidad (Bernardo y Smith, 1994, Sect. 5.3), si la distribución final conjunta es asintóticamente binormal con matriz de dispersión $nI(\hat{\theta}_n, \hat{\lambda}_n)$, donde $I(\theta, \lambda)$ es la matriz de información de Fisher, entonces $V(\theta, \lambda) = I^{-1}(\theta, \lambda)/n$ en el Teorema 5.

Teorema 6 (Funciones iniciales de referencia bajo factorización). *Con las mismas condiciones del Teorema 5, (i) si θ y λ son de variación independiente, es decir, Λ no depende de θ , y (ii) ambas $h_{\lambda\lambda}(\theta, \lambda)$ y $v_{\theta\theta}(\theta, \lambda)$ factorizan, de forma que*

$$(v_{\theta\theta}(\theta, \lambda))^{-1/2} \propto f_{\theta}(\theta)g_{\theta}(\lambda), \quad \sqrt{h_{\lambda\lambda}(\theta, \lambda)} \propto f_{\lambda}(\theta)g_{\lambda}(\lambda),$$

entonces la función θ -inicial de referencia es simplemente

$$\pi_{\theta}(\theta, \lambda) = f_{\theta}(\theta)g_{\lambda}(\lambda),$$

incluso cuando la función inicial de referencia condicional $\pi(\lambda|\theta) = \pi(\lambda) \propto g_{\lambda}(\lambda)$ sea impropia.

Demostración. Véase la demostración del Teorema 12 de Bernardo (2005a). □

En el caso de un modelo de localización y escala y bajos ciertas condiciones de regularidad, existe una función inicial de referencia (Fernández y Steel, 1999).

Teorema 7 (Modelos de localización y escala). *Considere un modelo de localización y escala*

$$\mathcal{M} \equiv \{\sigma^{-1}f(\sigma^{-1}(\mathbf{x} - \mu)); \mathbf{x} \in \mathcal{X}, \mu \in \mathcal{R}, \sigma > 0\},$$

regular, de forma que la distribución final conjunta de μ y σ es asintóticamente normal, y sea \mathcal{P}_0 la clase de todas las iniciales conjuntas continuas estrictamente positivas de (μ, σ) . Entonces la inicial de referencia ya sea para μ o para σ , es de la forma

$$\pi_{\mu}(\mu, \sigma|\mathcal{M}, \mathcal{P}_0) = \pi_{\sigma}(\mu, \sigma|\mathcal{M}, \mathcal{P}_0) = \sigma^{-1}.$$

Demostración. Véase Fernández and Steel (1999). □

2.3.3. Varios parámetros

Ahora extendemos el concepto de función inicial de referencia para el caso de más de dos parámetros.

Sea $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$, $\boldsymbol{\theta} \in \Theta$. Sucesivos condicionamientos dan como resultado la inicial de referencia conjunta para $\boldsymbol{\theta}$ de la forma siguiente

$$\pi(\boldsymbol{\theta}) = \pi(\theta_m | \theta_1, \dots, \theta_{m-1}) \times \pi(\theta_{m-1} | \theta_1, \dots, \theta_{m-2}) \times \dots \times \pi(\theta_2 | \theta_1) \pi(\theta_1).$$

Aplicando secuencialmente el algoritmo de obtención de iniciales de referencia uno puede obtener las condicionales $\pi(\theta_m | \theta_1, \dots, \theta_{m-1}), \dots, \pi(\theta_2 | \theta_1), \pi(\theta_1)$, en ese orden y así obtener la inicial de referencia que corresponde a la parametrización ordenada $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$.

Para describir este algoritmo es necesario especificar la siguiente notación: Supóngase el modelo $p(\mathbf{x} | \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ con matriz de información de Fisher de rango completo

$$I(\boldsymbol{\theta}) = -E_{\mathbf{x} | \boldsymbol{\theta}} \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\mathbf{x} | \boldsymbol{\theta}) \right\},$$

sea $S(\boldsymbol{\theta}) = I^{-1}(\boldsymbol{\theta})$ la matriz inversa de $I(\boldsymbol{\theta})$ y defínase los subvectores del vector $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{[j]} = (\theta_1, \dots, \theta_j)$, y $\boldsymbol{\theta}_{[j]} = (\theta_{j+1}, \dots, \theta_m)$; denótese a $S_j(\boldsymbol{\theta})$ como la submatriz $j \times j$ superior izquierda de $S(\boldsymbol{\theta})$, y sea $h_j(\boldsymbol{\theta})$ el elemento inferior derecho de $S_j^{-1}(\boldsymbol{\theta})$. Considérese $\Theta = \Theta_1 \times \dots \times \Theta_m$ con $\theta_i \in \Theta_i$, y sea $\{\Theta_i^l\}_{l=1}^\infty$ una sucesión creciente de subconjuntos compactos de Θ_i ; y defínase $\Theta_{[j]}^l = \Theta_{j+1}^l \times \dots \times \Theta_m^l$.

Teorema 8 (Funciones iniciales de referencia ordenadas bajo normalidad asintótica).

Con la notación mencionada anteriormente y bajo ciertas condiciones de regularidad (Proposición 5.14, Bernardo y Smith, 1994) que aseguran normalidad asintótica de la final conjunta, la función inicial de referencia $\pi(\boldsymbol{\theta})$ relativa a la ordenación de los parámetros $\{\theta_1, \dots, \theta_m\}$ está dada por

$$\pi(\boldsymbol{\theta}) = \lim_{l \rightarrow \infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)},$$

para algún punto interior $\boldsymbol{\theta}^* \in \Theta$, y donde $\pi^l(\boldsymbol{\theta})$ está definida por la siguiente recursión:

(i) Para $j = m$ y $\theta_m \in \Theta_m^l$

$$\pi_m^l(\theta_{[m-1]} | \boldsymbol{\theta}^{[m-1]}) = \pi_m^l(\theta_m | \theta_1, \dots, \theta_{m-1}) = \frac{\sqrt{h_m(\boldsymbol{\theta})}}{\int_{\Theta_m^l} \sqrt{h_m(\boldsymbol{\theta})} d\theta_m}.$$

(ii) Para $j = m - 1, m - 2, \dots, 2$ y $\theta_j \in \Theta_j^l$

$$\pi_j^l(\boldsymbol{\theta}_{[j-1]}|\boldsymbol{\theta}^{[j-1]}) = \pi_{j+1}^l(\boldsymbol{\theta}_{[j]}|\boldsymbol{\theta}^{[j]}) \frac{\exp\{E_j^l[\log \sqrt{h_j(\boldsymbol{\theta})}]\}}{\int_{\Theta_j^l} E_j^l[\log \sqrt{h_j(\boldsymbol{\theta})}] d\theta_j}$$

donde

$$E_j^l[\log \sqrt{h_j(\boldsymbol{\theta})}] = \int_{\Theta_{[j]}^l} \log \sqrt{h_j(\boldsymbol{\theta})} \pi_{j+1}^l(\boldsymbol{\theta}_{[j]}|\boldsymbol{\theta}^{[j]}) d\boldsymbol{\theta}_{[j]}.$$

(iii) Para $j = 1$, $\boldsymbol{\theta}_{[0]} = \boldsymbol{\theta}$, con $\boldsymbol{\theta}^{[0]}$ vacío, y

$$\pi^l(\boldsymbol{\theta}) = \pi_1^l(\boldsymbol{\theta}_{[0]}|\boldsymbol{\theta}^{[0]}).$$

Demostración. Véase la demostración de la Proposición 5.30 de Bernardo y Smith (1994).

□

Pueden encontrarse más detalles sobre la iniciales de referencia para más de dos parámetros en Berger and Bernardo (1992b, 1992c, 1992a).

La inicial de referencia puede obtenerse fácilmente si $\{h_j(\boldsymbol{\theta})\}$ depende únicamente de $\boldsymbol{\theta}^{[j]}$, e incluso se simplifica más si $H(\boldsymbol{\theta})$ es diagonal por bloques, de forma que el j -ésimo término puede factorizarse en un producto de una función de θ_j y una función que no depende de θ_j . El siguiente corolario formaliza esta idea.

Corolario 1 Si $h_j(\boldsymbol{\theta})$ depende únicamente de $\boldsymbol{\theta}^{[j]}$, $j = 1, \dots, m$, entonces

$$\pi^l(\boldsymbol{\theta}) = \prod_{j=1}^m \frac{\sqrt{h_j(\boldsymbol{\theta})}}{\int_{\Theta_j^l} \sqrt{h_j(\boldsymbol{\theta})} d\theta_j}, \quad \boldsymbol{\theta} \in \Theta^l.$$

Si $H(\boldsymbol{\theta})$ es diagonal por bloques, es decir, $\theta_1, \dots, \theta_m$ son mutuamente ortogonales, con

$$H(\boldsymbol{\theta}) = \begin{pmatrix} h_{11}(\boldsymbol{\theta}) & 0 & \dots & 0 \\ 0 & h_{22}(\boldsymbol{\theta}) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & h_{mm}(\boldsymbol{\theta}) \end{pmatrix}$$

entonces $h_j(\boldsymbol{\theta}) = h_{jj}(\boldsymbol{\theta})$, $j = 1, \dots, m$. Además, si, en el caso anterior

$$\sqrt{h_{jj}(\boldsymbol{\theta})} = f_j(\theta_j)g_j(\boldsymbol{\theta}),$$

donde $g_j(\boldsymbol{\theta})$ no depende de θ_j , y si las Θ_j^l 's no dependen de $\boldsymbol{\theta}$, entonces

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^m f_j(\theta_j).$$

Demostración. Véase la demostración del Corolario de la Proposición 5.30 en Bernardo y Smith (1994). □

En un problema específico, es natural preguntarse cuál es la mejor parametrización ordenada. Lo recomendable es elegir la ordenación de acuerdo al interés inferencial, con los parámetros en un ordenamiento decreciente de mayor a menor importancia. Es recomendable *no agrupar a menos que exista una razón específica para hacerlo* (Berger y Bernardo, 1992b,1992a).

A lo largo de esta Tesis, algunas funciones iniciales se obtendrán considerando diferentes ordenaciones en los parámetros.

2.3.4. Modelos no regulares con varios parámetros

Esta subsección se centra en los modelos no regulares y en el trabajo propuesto por Ghosal (1997).

Ghosal (1997) propuso la obtención de la función inicial de referencia para modelos multiparamétricos y no regulares. Siguiendo una notación similar a la dada por Ghosal, considérese un modelo $f(\cdot|\alpha, \varphi)$ que es *no regular* porque el espacio muestral depende de alguna función de uno de sus parámetros del modelo, presentando discontinuidades alrededor de un parámetro escalar α , mientras que en los parámetros restantes φ es un modelo regular si se mantiene fijo α . Un ejemplo de estos modelos es el Exponencial con dos parámetros,

$$\text{Ex}(\cdot|\alpha, \varphi) = \frac{1}{\varphi} e^{-\frac{x-\alpha}{\varphi}}, \quad x > \alpha \geq 0, \varphi > 0.$$

Ghosal (1997) obtiene la función inicial de referencia a través de la expansión asintótica de la divergencia esperada de Kullback-Leibler entre la densidad final y la función inicial, a la que Ghosal llamó como la *medida de información de Lindley*.

Ghosal trabaja bajo las condiciones de regularidad que brevemente mencionamos a continuación:

G.1 Uniformidad sobre subconjuntos compactos de $\Lambda \times \Phi$.

G.2 La densidad $f(\cdot|\alpha, \varphi)$ tiene dominio en un intervalo $S(\alpha) := [a_1(\alpha), a_2(\alpha)]$.

G.3 El dominio $S(\alpha)$ puede ser creciente o decreciente.

G.4 Para cada $\alpha \in \Lambda$ y $\varphi \in \Phi$, los siguiente límites existen:

$$p(\alpha, \varphi) = \lim_{x \downarrow a_1(\alpha)} f(x|\alpha, \varphi) \quad (2.7)$$

$$q(\alpha, \varphi) = \lim_{x \uparrow a_2(\alpha)} f(x|\alpha, \varphi) \quad (2.8)$$

G.5 Las funciones definidas por

$$\begin{aligned} J(\alpha, \varphi) &= \int (g_\alpha(x|\alpha, \varphi)) dx, \quad j = 1, \dots, d \\ J_j(\alpha, \varphi) &= \int g_\alpha(x|\alpha, \varphi) g_{\varphi_j}(x|\alpha, \varphi) dx \\ J_{k,j}(\alpha, \varphi) &= \int g_{\varphi_j}(x|\alpha, \varphi) g_{\varphi_k}(x|\alpha, \varphi) dx, \quad j, k = 1, \dots, d \end{aligned} \quad (2.9)$$

son finitas y continuas en (α, φ) , donde $g_\alpha = \frac{\partial g}{\partial \alpha}$, $g_{\varphi_j} = \frac{\partial g}{\partial \varphi_j}$ y $g = \sqrt{f}$, $\varphi = \{\varphi_1, \dots, \varphi_d\}$. La matriz $((J_{k,j}(\alpha, \varphi)))$ es positiva definida.

G.6 Las funciones $a_1(\alpha)$, $a_2(\alpha)$, $a'_1(\alpha)$, $a'_2(\alpha)$ (donde “ ’ ” denota la derivada), $J(\alpha, \varphi)$, $J_j(\alpha, \varphi)$, $J_{jk}(\alpha, \varphi)$, $j, k = 1, \dots, d$, tiene un *majorant* que es el producto de una función polinomial en $|\alpha|$ y una función exponencial en $\|\varphi\|$.

G.7 Para cualesquiera (α, φ) fijos, $\max\{\frac{\partial}{\partial \varphi_j} \log f(x|\alpha, \varphi) : j = 1, \dots, d\}$ permanece acotado cuando $x \downarrow a_1(\alpha)$ y $x \uparrow a_2(\alpha)$.

G.8 Para alguna $C, \gamma > 0$,

$$\int f^{1/2}(x|\alpha, \varphi) f^{1/2}(x|\alpha + u, \varphi + \mathbf{v}) dx \leq C|u|^{-\gamma} \exp(-\gamma\|\mathbf{v}\|).$$

G.9 Para $x \in (a_1(\alpha), a_2(\alpha))$, $\log f(x|\alpha, \varphi)$ es tres veces diferenciable (y continua) en (α, φ) . Para cualquier (α, φ) , existe una vecindad $N_{\alpha, \varphi}$ y una función $P_{\alpha, \varphi}$ -integrable $H_{\alpha, \varphi}$. Además, $E_{\alpha, \varphi} H_{\alpha, \varphi}(X_1)$ es continua en α, φ . Iguale $c(\alpha, \varphi) = E_{\alpha, \varphi} (\frac{\partial}{\partial \alpha} \log f(x|\alpha, \varphi))$, es decir,

$$c(\alpha, \varphi) = a'_1(\alpha)p(\alpha, \varphi) - a'_2(\alpha)q(\alpha, \varphi) > 0. \quad (2.10)$$

G.10 Para cualquier subconjunto compacto $K \subset \Lambda \times \Phi$, se tiene que

$$\sup_{(\alpha, \varphi) \in K} \sup_{n \geq 1} E_{\alpha, \varphi} \sigma_n(\alpha) < \infty,$$

donde $\sigma_n(\alpha) = n(W_n - \alpha)$, $W_n = \min\{a_1^{-1}(X_{(1)}), a_2^{-1}(X_{(n)})\}$.

Una vez que las condiciones (A.1)-(A.10) se han cumplido, la inicial de referencia π^* (que es continua y definida en cualquier subconjunto compacto K) después de normalización, es la que maximiza la funcional

$$\int_K \pi(\alpha, \varphi) \log \left(\frac{c(\alpha, \varphi)(\det \mathbf{F}(\alpha, \varphi))^{1/2}}{\pi(\alpha, \varphi)} \right) d\alpha d\varphi.$$

Se sigue que la inicial e referencia es de la forma

$$\pi^*(\alpha, \varphi) = c(\alpha, \varphi)(\det \mathbf{F}(\alpha, \varphi))^{1/2} \quad (2.11)$$

$$\text{con } \mathbf{F}(\alpha, \varphi) = ((4J_{jk}(\alpha, \varphi))) \quad (2.12)$$

que preserva la propiedad de invarianza, y donde $\mathbf{F}(\alpha, \varphi)$ es una matriz definida positiva (véase la condición G.5).

En el caso de tener un parámetro de interés α y los restantes parámetros φ son de ruido, Ghosal (1997) sugiere dos *procedimientos*: Procedimiento **1** que es básicamente una adaptación del método dado por Berger y Bernardo (1989) al fijar un parámetro, digamos α , encontrar la densidad condicional $\pi(\varphi|\alpha)$, seleccionar una sucesión de subconjuntos compactos de Φ y finalmente obtener la inicial de referencia $\pi(\alpha)$, por lo tanto $\pi_\alpha(\alpha, \varphi) \propto \pi(\varphi|\alpha)\pi(\alpha)$. Cuando factorizan tanto (2.10) como la raíz cuadrada de (2.12), la inicial de referencia puede obtenerse muy fácilmente de manera similar al algoritmo de Berger y Bernardo. El procedimiento **2** se basa en una maximización funcional, la cual resulta en la elección de $\pi(\alpha, \varphi) = c(\alpha, \varphi)$.

Cuando el parámetro de interés es φ y α es el parámetro de ruido, los procedimientos anteriores se modifican de forma *obvia*, es decir, intercambiando los roles de $c(\alpha, \varphi)$ y $\mathbf{F}(\alpha, \varphi)$.

Para más detalles véase Ghosal (1997).

2.3.5. Parámetros discretos

En esta subsección se describen las funciones iniciales de referencia para el caso de parámetros discretos.

Recordemos que el problema de definir una inicial objetiva cuando el espacio paramétrico es discreto ha sido estudiado por varios autores. Jeffreys (1961) propone el uso de $\pi(\theta) \propto \theta^{-1}$ cuando $\theta \in \mathcal{N}^+$ y Rissanen (1983) propone una inicial propia para los enteros positivos usando teoría de códigos; ambas propuestas han sido mencionadas en la Sección 1.9.

Berger, Bernardo y Sun (2009b) sugieren procedimientos generales para encontrar las funciones iniciales de referencia en espacios paramétricos discretos, presentan algunos procedimientos que convierten un problema discreto en un problema continuo y así usar la teoría convencional de iniciales de referencia. A continuación incluimos un breve resumen de su trabajo.

Sea θ un parámetro discreto de interés y la variable aleatoria x ,

A.1 *Supóngase que el parámetro es continuo.*

Si θ fuese tratado como si fuese continuo sería necesario introducir un parámetro de normalización. Por otro lado, algunas veces los datos también puede considerarse como continuos, lo cual haría necesario introducir un segundo parámetro de normalización. Consecuentemente, se corre el riesgo que el modelo continuo resultante sea diferente del modelo discreto original.

A.2 *Introducción de un hiperparámetro jerárquico continuo.*

Añadir un nivel jerárquico con hiperparámetro continuo y resolver el problema. Aunque no es frecuente tener una única estructura jerárquica definida.

A.3 *Aplicar la teoría de iniciales de referencia con un estimador consistente.*

Encontrar un estimador consistente, posiblemente lineal, basado en las repeticiones del modelo considerado y encontrar su distribución asintótica en el muestreo cuando el tamaño de la muestra crece. Pretenda que el parámetro es continuo en esta nueva distribución y encontrar la correspondiente inicial de referencia.

Berger *et al.* (2009b) discuten cinco problemas discretos que son: El modelo de crecimiento poblacional, las distribuciones Hipergeométrica, Beta Binomial y Binomial, y selección de modelos.

2.4. Algunas dificultades de usar iniciales objetivas

Finalizaremos este capítulo con un resumen de algunas de las dificultades encontradas al usar iniciales objetivas.

2.4.1. Inconsistencia fuerte

La inconsistencia fuerte ya se ha mencionado previamente en la Sección 1.2 pero ahora discutiremos principalmente el trabajo de Stone (1976).

El trabajo de Stone (1976) proporciona interesantes ejemplos que muestran claramente posibles inconsistencias de usar iniciales impropias. Stone ilustra la inconsistencia entre las coberturas de probabilidad y las probabilidades finales resultantes de usar iniciales uniformes en casos específicos. Por ejemplo, si uno considera una región C , uno no esperaría diferencias entre la probabilidad final $p(C|\mathbf{x})$ y la cobertura de la probabilidad

en el muestreo $p(C|\theta)$ para todos los valores de θ . La clave está en la elección de la inicial y la distinción entre los parámetros de interés de los de ruido. Lo cual remarca la importancia que tiene definir claramente los parámetros o la cantidad de interés, véase Bernardo (1979) quien discute el primer ejemplo de Stone (1976).

2.4.2. El efecto dominante de la inicial

Algunas veces la inicial domina los datos, incluso cuando la densidad final es consistente. La *paradoja de Stein*, descrita posteriormente, es un buen ejemplo de dominancia de la inicial sobre los datos. También es ilustrativo el ejemplo dado en Kass and Wasserman (1996, Sección 4.2.2.)

2.4.3. Inadmisibilidad

Un problema latente con el uso de iniciales objetivas es que pueden producir estimadores Bayes inadmisibles. Un estimador Bayes es *inadmissible* si existe otro estimador con menor o igual riesgo para cada θ , pero existe al menos una θ donde el riesgo es estrictamente menor. Siguiendo a Stein (1956), considérese la estimación de varias medias normales: Sean $x_i \sim N(x_i|\theta_i, 1)$ ($i = 1, \dots, n$) observaciones aleatorias independientes, y supóngase que se requiere estimar el vector de medias $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$. Si se supone $\pi(\boldsymbol{\theta}) \propto c$, una constante, la distribución final es una normal multivariada $N_m(\boldsymbol{\theta}|\mathbf{x}, I)$, donde $\mathbf{x} = (x_1, \dots, x_n)'$ y matriz identidad I como matriz de covarianza. Stein (1956) demostró que bajo la función de pérdida cuadrática cuando $n \geq 3$, la media final de $\boldsymbol{\theta}$ es inadmissible. Por otro lado, Bernardo y Rueda (2002) proponen como función de pérdida la discrepancia intrínseca y usar la inicial de referencia del parámetro de interés, bajo estos supuestos, la media final tiene un riesgo apropiado.

2.4.4. Paradojas de marginalización

Esta dificultad se ha mencionado anteriormente en la Sección 1.2. Aquí discutiremos con detalle el conocido ejemplo de inferencias acerca del coeficiente de variación de una variable aleatoria Normal (Bernardo y Smith, 1994, Ej.5.25 y Stone y Dawid, 1972, Ej. 2). Sea $\mathbf{x} = \{x_1, \dots, x_n\}$ una muestra aleatoria de $N(x|\mu, \sigma^2)$ y supóngase que se requiere hacer inferencias acerca del coeficiente de variación $\phi = \frac{\mu}{\sigma}$. Supóngase la inicial impropia $\pi(\mu, \sigma) \propto \sigma^{-1}$, por tanto la final conjunta para (μ, σ) es

$$\pi(\mu, \sigma|\mathbf{x}) \propto \sigma^{-(n+1)} \exp \left\{ -\frac{n}{2\sigma^2} (s^2 + (\bar{x} - \mu)^2) \right\},$$

con $s^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$. La final en la nueva parametrización (ϕ, σ) es $\pi(\phi, \sigma|\mathbf{x}) \propto \frac{1}{\sigma^n} \exp \left\{ -\frac{1}{2}n\phi^2 + \frac{n\phi\bar{x}}{\sigma} - \frac{R^2}{2\sigma^2} \right\}$ donde $R = \sqrt{\sum_{i=1}^n x_i^2}$. Integrandolo sobre σ , la densidad

marginal posterior final para ϕ es

$$\pi(\phi|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}n\phi^2\right\} \int_0^\infty \omega^{n-2} \exp\left\{-\frac{1}{2}\omega^2 + r\phi\omega\right\} d\omega$$

con $r = n\bar{x}/R$. Nótese que $\pi(\phi|r, n)$ es una función de r . Por otro lado, la distribución de r dados μ y σ es

$$p(r|\mu, \sigma) \propto \exp\left\{-\frac{1}{2}n\phi^2\right\} (1 - r^2/n)^{(n-3)/2} \int_0^\infty \omega^{n-1} \exp\left\{-\frac{1}{2}\omega^2 + r\phi\omega\right\} d\omega,$$

la cual es una función sólo de ϕ . Entonces, con $p(r|\mu, \sigma)$ y una distribución inicial apropiada para ϕ uno esperaría obtener la misma densidad final que al usar $p(x|\mu, \sigma)$. En realidad no existe tal inicial satisfaciendo lo anterior. Es decir, usando la inicial conjunta $\pi(\mu, \sigma) \propto \sigma^{-1}$, las inferencias sobre el modelo completo son necesariamente diferentes de aquellas obtenidas a través del modelo marginal, y por tanto se tiene una paradoja de marginalización.

2.4.5. Densidades finales impropias

Algunas veces las iniciales impropias producen finales impropias.

Consideremos el ejemplo dado en Kass and Wasserman (1996) de un modelo jerárquico

$$y_i \sim N(y_i|\mu_i, \sigma) \quad y \quad \mu_i \sim N(\mu_i|\mu, \tau)$$

para $i = 1, \dots, n$, donde σ es conocida. Si se elige la inicial $\pi(\tau) \propto \tau^{-1}$ se obtiene una final impropia. Este ejemplo también se discute en el Teorema 2 de Berger (1985, Sección 4.6.2, p. 183-187).

Cuando se utiliza una inicial impropia hay que garantizar que se obtiene una final propia. A pesar de que el poder de cómputo facilita la verificación numérica de que las finales son propias, los modelos son cada vez más complejos y resulta una tarea muy difícil demostrar analíticamente que la final sea propia. Aparentemente, una forma de evitar finales impropias es usar iniciales propias difusas o *vagas*; pero en la práctica esto no resuelve el problema, véase en Bernardo y Smith (1994, p. 365) un ejemplo del peligro del uso ingenuo de las iniciales difusas.

Ejemplo 2 (Paradoja de Stein). Sea $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ una muestra aleatoria proveniente de $N_k(\mathbf{x}_i|\boldsymbol{\mu}, I_k)$. Sea $t = \sum_{i=1}^n \bar{x}_i^2$ donde \bar{x}_i es la media de las n observaciones aleatorias de la coordenada i . Puesto que $\boldsymbol{\mu}$ es un parámetro de localización, la inicial objetiva para la medias es $\pi(\mu_1, \dots, \mu_k) \propto 1$, la cual puede aproximarse con la densidad inicial propia

$$\pi(\mu_1, \dots, \mu_k) \propto \prod_{i=1}^k N(\mu_i|0, \sigma)$$

con desviación estándar σ pequeña.

Supóngase que se requiere hacer inferencias acerca de $\phi = \sum_{i=1}^k \mu_i^2$. Se verá que con la inicial anterior, para k grande, ésta dominará a los datos. En efecto, la distribución final de $n\phi$ es una distribución no centralizada *Ji cuadrada* con k grados de libertad y parámetro de descentralización nt ,

$$n\phi \sim \chi_k^2(nt)$$

de forma que la media y varianzas finales de ϕ son, respectivamente,

$$E(\phi|\mathbf{x}_1, \dots, \mathbf{x}_n) = t + \frac{k}{n}, \quad \text{Var}(\phi|\mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{2}{n} \left(2t + \frac{k}{n} \right).$$

Por otro lado, la distribución en el muestreo de t es una *Ji cuadrada* con k grados de libertad y parámetro de centralización $n\phi$, así que

$$nt \sim \chi_k^2(n\theta) \quad \text{con} \quad E(t|k, n\phi) = \phi + \frac{k}{n}.$$

Entonces, si $k = 100$, $n = 1$ y $t = 200$, la media posterior $E(t|\mathbf{x}) \approx 300$, $\text{Var}(\phi|\mathbf{x}) \approx 32^2$, mientras que el estimador insesgado $\hat{\phi} = E(\phi|k, n\phi)$ bajo su distribución en el muestreo es

$$\hat{\phi} = t - \frac{k}{n} = 100$$

que es completamente diferente a la media, y esta diferencia se incrementa conforme $k \rightarrow \infty$. Entonces la inicial objetiva elegida domina a los datos.

Por otro lado, si se usa la inicial de referencia relativa a la ordenación de los parámetros $\{\phi, \omega_1, \dots, \omega_{k-1}\}$, donde ω_i ($i = 1, \dots, k-1$) son los parámetros de ruido, con inicial marginal dada por $\pi(\phi) \propto \phi^{-1/2}$, la distribución final es

$$\pi(\phi|\mathbf{x}) \propto \phi^{-1/2} \chi_k^2(nt|t, n\phi)$$

cuya moda es cercana a $\hat{\phi}$ (véase Bernardo, 1979), y tiene regiones creíbles con adecuadas propiedades de cobertura.

De esta paradoja se pueden subrayar dos hechos relevantes: La inicial objetiva puede dominar a los datos y el uso de iniciales difusas no evita las dificultades que podrían tener las iniciales impropias.

2.4.6. Dependencia del espacio muestral

Algunas veces las iniciales objetivas dependen del espacio muestral; a estas iniciales se les conoce como “dependientes del diseño” y pueden resultar en la violación del *principio de verosimilitud*, tal como se discutió en la Sección 2.3.1. Veamos el siguiente ejemplo:

Supóngase x_1 es el número de éxitos en n lanzamientos de una moneda con probabilidad de éxito p , así $x_1 \sim \text{Bi}(x_1|n, p)$ con correspondiente inicial de Jeffreys $\pi_1(p) \propto \frac{1}{\sqrt{p(1-p)}}$.

Ahora considérese z el número de fracasos antes de obtener r éxitos, es decir la distribución es la Binomial negativa $z \sim \text{NBi}(z|r, p)$, con inicial de Jeffreys $\pi_2(p) \propto \frac{1}{p\sqrt{1-p}}$. Entonces, dependiendo del espacio muestral han resultado dos iniciales diferentes, incluso cuando la función de verosimilitud es proporcional.

Recuérdese, las iniciales de referencia típicamente dependen del modelo *supuesto*, no de la función de verosimilitud *observada*.

Capítulo 3

Predicción con distribuciones iniciales objetivas.

En este capítulo presentamos una revisión de algunas investigaciones recientes sobre distribuciones predictivas usando distribuciones iniciales objetivas e incluimos interesantes resultados frecuentistas.

Según Robert *et al.* (2008), Jeffreys (1961, Sec. 3.8) fue el primero en usar la predicción Bayesiana en modelos de localización y escala. La idea que Jeffreys introduce es la siguiente: En un modelo de localización y escala, se quiere saber la probabilidad de que una tercera observación, se encuentre contenida en el intervalo (x_1, x_2) formado por dos observaciones previas. Puede demostrarse que ésta probabilidad es $\frac{1}{3}$ para todo valor de x_1 y x_2 y Jeffreys demuestra que con la inicial objetiva que propone se obtiene una distribución predictiva compatible con este resultado. En este capítulo generalizamos a más de dos observaciones éste último resultado.

Analizaremos además varias de las propuestas de obtención de iniciales objetivas para predicción y que mencionamos brevemente a continuación: Kuboki (1998) es un razonamiento parecido al de Bernardo (1979) para problemas de inferencia paramétrica, Eaton y Sudderth (1998) tratan de evitar la inconsistencia fuerte. La idea dada por Komaki (1996) es la clave para el trabajo de Ghosh y Sweeting (2000) quienes usan aproximación de la validez frecuentista. Por otro lado, Smith (1999) propone el uso de la función de pérdida, mientras que Hartigan (1998) propone las distribuciones basadas en las funciones de máxima verosimilitud para predicción.

3.1. Planteamiento del problema.

3.1.1. Probabilidad predictiva de que una nueva observación se encuentre contenida en el intervalo formado por dos observaciones previas.

Comenzaremos con una interesante propiedad frecuentista de los modelos univariados continuos.

Teorema 9 Sean x_1, x_2 dos observaciones provenientes de una distribución univariada continua perteneciente a una familia regular de distribuciones de la forma $\mathcal{M} \equiv \{f(x|\theta), \theta \in \Theta\}$. La probabilidad de que una tercera observación x se encuentra contenida en el intervalo (x_1, x_2) ($x_1 < x_2$) formado por dos observaciones previas es $\frac{1}{3}$.

Demostración. Este resultado es un caso particular del Teorema 10. □

Por otro lado, Jeffreys (1961, Sec. 3.8) consideró la probabilidad final predictiva de que una nueva observación x se encuentre contenida entre las dos observaciones x_1 y x_2 dadas, suponiendo un modelo de localización y escala de la forma

$$\mathcal{M} = \{\sigma^{-1}f(\sigma^{-1}(x - \mu)), x \in \mathcal{X}, \mu \in \mathcal{R}, \sigma > 0\}$$

donde μ es el parámetro de localización y σ el parámetro de escala, ambos desconocidos; con la inicial $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$ y donde \mathcal{M} cumple

- $f(\cdot)$ es una función continua, y
- $F(z) = \int_{-\infty}^z f(t)dt$ con $F(\infty) = 1$.

A continuación reproducimos la obtención de Jeffreys: Para un modelo de localización y escala Jeffreys propone usar la inicial conjunta $\pi(\mu, \sigma) \propto \sigma^{-1}$, de forma que la correspondiente densidad final conjunta es proporcional a

$$\pi(\mu, \sigma|x_1, x_2) \propto \frac{1}{\sigma^3} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right).$$

La probabilidad de tener una nueva observación x contenida en el intervalo (x_1, x_2) formado por dos observaciones previas con $x_1 < x_2$ es

$$\begin{aligned} \Pr(x_1 < x < x_2|x_1, x_2) &= \frac{\int_0^\infty \int_{-\infty}^\infty \int_{x_1}^{x_2} \frac{1}{\sigma^4} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right) f\left(\frac{x_3 - \mu}{\sigma}\right) dx d\mu d\sigma}{\int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\sigma^4} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right) f\left(\frac{x_3 - \mu}{\sigma}\right) dx d\mu d\sigma} \\ &= \frac{I_3}{I_4}. \end{aligned}$$

Evaluemos por separado las integrales I_3 y I_4 . Puede verse claramente que

$$\begin{aligned} I_3 &= \int_0^\infty \int_{-\infty}^\infty \int_{x_1}^{x_2} \frac{1}{\sigma^4} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right) f\left(\frac{x - \mu}{\sigma}\right) dx d\mu d\sigma \\ &= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma^3} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right) \\ &\quad \times \left[F\left(\frac{x_2 - \mu}{\sigma}\right) - F\left(\frac{x_1 - \mu}{\sigma}\right) \right] d\mu d\sigma. \end{aligned}$$

Haciendo el siguiente cambio de variables,

$$\theta = \frac{x_1 - \mu}{\sigma}, \quad \phi = \frac{x_2 - \mu}{\sigma} \quad (3.1)$$

donde $-\infty < \theta < \phi < \infty$, ($\Psi(\theta, \phi) = w(\mu, \sigma)$), uno tiene que

$$J_{w^{-1}}(\Psi) = (\phi - \theta)^{-2} \begin{pmatrix} \phi(x_1 - x_2) & \theta(x_2 - x_1) \\ x_2 - x_1 & -(x_2 - x_1) \end{pmatrix}.$$

Por lo tanto

$$\begin{aligned} I_3 &= \frac{1}{x_2 - x_1} \int_{-\infty}^\infty \int_\theta^\infty f(\phi) f(\theta) [F(\phi) - F(\theta)] d\phi d\theta \\ &= \frac{1}{2(x_2 - x_1)} \int_{-\infty}^\infty (1 - F(\theta))^2 f(\theta) d\theta = \frac{1}{6(x_2 - x_1)}. \end{aligned}$$

De manera análoga y usando el mismo cambio de variables dado en (3.1), la integral I_4 queda como

$$\begin{aligned} I_4 &= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma^3} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right) \int_{-\infty}^\infty \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) dx d\mu d\sigma \\ &= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma^3} f\left(\frac{x_1 - \mu}{\sigma}\right) f\left(\frac{x_2 - \mu}{\sigma}\right) d\mu d\sigma \\ &= \frac{1}{x_2 - x_1} \int_{-\infty}^\infty \int_\theta^\infty f(\theta) f(\phi) d\phi d\theta = \frac{1}{x_2 - x_1} \int_{-\infty}^\infty (1 - F(\theta)) f(\theta) d\theta \\ &= \frac{1}{2(x_2 - x_1)}. \end{aligned}$$

Consecuentemente, la probabilidad requerida es

$$\Pr(x_1 < x < x_2 | x_1, x_2) = \frac{I_3}{I_4} = \frac{1}{3}.$$

De este modo, para problemas de localización y escala, Jeffreys estableció una perfecta correspondencia entre la cobertura observada y la probabilidad predictiva resultante de usar la inicial que él recomendó para este caso.

Ahora procederemos a generalizar el Teorema 9 y el resultado de Jeffreys.

3.1.2. Probabilidad predictiva de que una nueva observación x_{n+1} pertenezca al intervalo formado por las observaciones mínima $x_{(1)}$ y máxima $x_{(n)}$ de una muestra dada.

Teorema 10 Sea $\{x_1, \dots, x_n\}$ una muestra aleatoria de tamaño n proveniente de una distribución continua univariada que pertenece a la familia regular $\mathcal{M} = \{f(x|\theta), \theta \in \Theta\}$. La probabilidad de que una nueva observación x se encuentra contenido en el intervalo $(x_{(1)}, x_{(n)})$ formado por las observaciones mínima y máxima de la muestra, es $\frac{n-1}{n+1}$.

Demostración. Calcúlese la probabilidad de tener una nueva observación x contenida en el intervalo $(x_{(1)}, x_{(n)})$ dados los datos $\mathbf{x} = \{x_1, \dots, x_n\}$, donde $x_{(1)} = \min\{x_1, \dots, x_n\}$ y $x_{(n)} = \max\{x_1, \dots, x_n\}$. Esta probabilidad es

$$p(x_{(1)} < x < x_{(n)}|\theta) = \int_{x_{(1)}}^{x_{(n)}} f(x|\theta)dx = F(x_{(n)}|\theta) - F(x_{(1)}|\theta), \quad (3.2)$$

que es una función del mínimo y máximo de \mathbf{x} . Entonces necesitamos calcular el valor esperado de (3.2), evaluando sobre la densidad conjunta del mínimo y máximo, puesto que ahora ambas observaciones son estadísticos suficientes. La densidad conjunta del mínimo y máximo es

$$g(x_{(1)}, x_{(n)}|\theta) = n(n-1) (F(x_{(n)}|\theta) - F(x_{(1)}|\theta))^{n-2} f(x_{(n)}|\theta)f(x_{(1)}|\theta) \\ -\infty < x_{(1)} < x_{(n)} < \infty.$$

donde $f = \frac{d}{dx}F$.

Entonces el valor esperado requerido es

$$E_{x_{(1)}, x_{(n)}|\theta} [F(x_{(n)}|\theta) - F(x_{(1)}|\theta)] = n(n-1) \int_{-\infty}^{\infty} \int_{-\infty}^{x_{(n)}} (F(x_{(n)}|\theta) - F(x_{(1)}|\theta)) \\ *(F(x_{(n)}|\theta) - F(x_{(1)}|\theta))^{n-2} f(x_{(1)}|\theta)f(x_{(n)}|\theta)dx_{(1)}dx_{(n)} \\ = n(n-1) \int_{-\infty}^{\infty} \int_{-\infty}^{x_{(n)}} (F(x_{(n)}|\theta) - F(x_{(1)}|\theta))^{n-1} f(x_{(1)}|\theta)f(x_{(n)}|\theta)dx_{(1)}dx_{(n)}.$$

que será

$$\begin{aligned}
& n(n-1) \int_{-\infty}^{\infty} \int_{-\infty}^{x_{(n)}} (F(x_{(n)}|\theta) - F(x_{(1)}|\theta))^{n-1} f(x_{(n)}|\theta) f(x_{(1)}|\theta) dx_{(1)} dx_{(n)} \\
&= n(n-1) \int_{-\infty}^{\infty} \frac{1}{n} (F(x_{(n)}|\theta) - F(x_{(1)}|\theta))^n \Big|_{-\infty}^{x_{(n)}} f(x_{(n)}|\theta) dx_{(n)} \\
&= (n-1) \int_{-\infty}^{\infty} (F(x_{(n)}|\theta))^n f(x_{(n)}|\theta) dx_{(n)} \\
&= (n-1) \frac{1}{n+1} (F(x_{(n)}|\theta))^{n+1} \Big|_{-\infty}^{\infty} = \frac{n-1}{n+1}.
\end{aligned}$$

□

Ahora generalizamos el resultado de Jeffreys para el caso de tener $n > 2$ observaciones. La probabilidad requerida es

$$\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x}) = \int_{x_{(1)}}^{x_{(n)}} p(x | \mathbf{x}) dx \quad (3.3)$$

donde $\mathbf{x} = \{x_1, \dots, x_n\}$ es un conjunto de n observaciones idénticamente distribuidas provenientes de un modelo de localización y escala,

$$x_{(1)} = \min\{x_1, \dots, x_n\}, \quad x_{(n)} = \max\{x_1, \dots, x_n\},$$

con modelo de localización y escala \mathcal{M} de la forma

$$\mathcal{M} \equiv \{\sigma^{-1} f(\sigma^{-1}(x - \mu)), x \in \mathcal{X}, \mu \in \mathcal{R}, \sigma > 0\},$$

donde $f(\cdot | \mu, \sigma)$ es una función continua que cumple con las mismas propiedades mencionadas anteriormente. Sea $\mathbf{y} = \{x_{(1)}, \dots, x_{(n)}\}$ el vector de estadísticos de orden, nótese que \mathbf{y} es un estimador suficiente para el vector $\boldsymbol{\theta} = \{\mu, \sigma\}$ cuya densidad conjunta es

$$\prod_{i=1}^n f(x_i | \mu, \sigma) = \prod_{i=1}^n f(x_{(i)} | \mu, \sigma) = f(\mathbf{y} | \mu, \sigma).$$

Por tanto, la probabilidad (3.3) puede reescribirse como

$$\begin{aligned}
\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x}) &= \Pr(x_{(1)} < x < x_{(n)} | \mathbf{y}) \\
&= \Pr(x_{(1)} < x < x_{(n)} | x_{(1)}, x_{(n)}),
\end{aligned}$$

nótese que depende únicamente de $x_{(1)}$ y $x_{(n)}$, ahora ambos son estadísticos suficientes. Por lo tanto, es necesario especificar la densidad conjunta de $(x_{(1)}, x_{(n)})$

$$g(x_{(1)}, x_{(n)}|\mu, \sigma) = n(n-1)[F(\sigma^{-1}(x_{(n)} - \mu)) - F(\sigma^{-1}(x_{(1)} - \mu))]^{n-2} \\ \times \frac{1}{\sigma^2} f(\sigma^{-1}(x_{(n)} - \mu)) f(\sigma^{-1}(x_{(1)} - \mu)).$$

con $f = \frac{d}{dx}F$.

Con la inicial $\pi(\mu, \sigma) \propto \sigma^{-1}$ la densidad final es proporcional a

$$\pi(\mu, \sigma|x_{(1)}, x_{(n)}) \propto g(x_{(1)}, x_{(n)}|\mu, \sigma)\pi(\mu, \sigma).$$

Luego entonces la probabilidad requerida es

$$\Pr(x_{(1)} < x < x_{(n)}|x_{(1)}, x_{(n)}) = \frac{\int_0^\infty \int_{-\infty}^\infty \int_{x_1}^{x_2} \frac{1}{\sigma^2} g(x_{(1)}, x_{(n)}|\mu, \sigma) f\left(\frac{x-\mu}{\sigma}\right) dx d\mu d\sigma}{\int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\sigma^2} g(x_{(1)}, x_{(n)}|\mu, \sigma) f\left(\frac{x-\mu}{\sigma}\right) dx d\mu d\sigma} \\ = \frac{I_5}{I_6}.$$

Trabajando de manera separada las dos integrales, y usando el siguiente cambio de variables

$$\theta = \frac{x_{(1)} - \mu}{\sigma}, \quad \phi = \frac{x_{(n)} - \mu}{\sigma} \tag{3.4}$$

donde $-\infty < \theta < \phi < \infty$, $(\Psi(\theta, \phi) = w(\mu, \sigma))$, uno tiene

$$J_{w^{-1}}(\Psi) = (\phi - \theta)^{-2} \begin{pmatrix} \phi(x_{(1)} - x_{(n)}) & \theta(x_{(n)} - x_{(1)}) \\ x_{(n)} - x_{(1)} & -(x_{(n)} - x_{(1)}) \end{pmatrix} \\ |\det J_{w^{-1}}(\Psi)| = \frac{(x_{(n)} - x_{(1)})^2}{(\phi - \theta)^3}.$$

Primero procederemos a trabajar con la parte más interna de la integral I_5 , expresándola en términos de θ y ϕ

$$\int_{x_{(1)}}^{x_{(n)}} \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) dx = \int_\theta^\phi f(\tau) d\tau = F(\phi) - F(\theta)$$

con $\tau = \frac{x-\mu}{\sigma}$, $\theta = \frac{x_{(1)}-\mu}{\sigma} < \tau < \phi = \frac{x_{(n)}-\mu}{\sigma}$. Usando el cambio de variables dado en (3.4), entonces se sigue que

$$\begin{aligned} I_5 &= \frac{n(n-1)}{x_{(n)} - x_{(1)}} \int_{-\infty}^{\infty} \int_{\theta}^{\infty} f(\phi)f(\theta)[F(\phi) - F(\theta)]^{n-1} d\phi d\theta \\ &= \frac{n(n-1)}{x_{(n)} - x_{(1)}} \int_{-\infty}^{\infty} f(\phi)[1 - F(\theta)]^n d\theta = \frac{n-1}{n+1} \frac{1}{x_{(n)} - x_{(1)}} \end{aligned}$$

y de manera similar,

$$\begin{aligned} I_6 &= \int_0^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sigma^2} g(x_{(1)}, x_{(n)}) f\left(\frac{x-\mu}{\sigma}\right) dx d\mu d\sigma \\ &= \int_0^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sigma} g(x_{(1)}, x_{(n)}) d\mu d\sigma \\ &= \frac{n(n-1)}{x_{(n)} - x_{(1)}} \int_{-\infty}^{\infty} \int_{\theta}^{\infty} f(\phi)f(\theta)[F(\phi) - F(\theta)]^{n-2} d\phi d\theta \\ &= \frac{n}{x_{(n)} - x_{(1)}} \int_{-\infty}^{\infty} f(\theta)[1 - F(\theta)]^{n-1} d\theta = \frac{1}{x_{(n)} - x_{(1)}}. \end{aligned}$$

Finalmente, tenemos

$$\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x}) = \frac{I_5}{I_6} = \frac{\frac{n-1}{n+1} \frac{1}{x_{(n)} - x_{(1)}}}{\frac{1}{x_{(n)} - x_{(1)}}} = \frac{n-1}{n+1}.$$

Una comparación con el Teorema 10 demuestra nuevamente que en modelos de localización y escala, con la inicial objetiva convencional $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$, existe una perfecta concordancia entre la cobertura de observación y la probabilidad predictiva objetiva.

Uno de los objetivos de esta Tesis es analizar si este tipo de resultados son generalizables a modelos generales, o si sólo resultan válidos en modelos de localización y escala.

3.2. Métodos de decisión de referencia para predicción

En el trabajo *Reference priors for prediction* de Kuboki (1998) se encuentra una propuesta de obtención de iniciales objetivas para predicción. El autor trata de extender el argumento dado en Bernardo (1979) y Berger y Bernardo (1989, 1992a, 1992b) para obtener iniciales objetivas que ayuden a hacer inferencias sobre los parámetros. Kuboki caracteriza una inicial no informativa para predicción basada en la información predictiva esperada que proveen las observaciones actuales. Kuboki pretende encontrar la inicial que maximiza la información predictiva faltante, que es típicamente infinita; y evita esta dificultad siguiendo el procedimiento de Bernardo (1979). La inicial de Kuboki para

predicción satisfice las propiedades de simetría e invarianza requeridas por Jeffreys.

Kuboki encuentra una expresión explícita para la inicial si el modelo es un caso regular, es decir,

- (i) el modelo $f(x|\theta)$ es doblemente diferenciable y continuo en θ para casi cualquier x , y el valor esperado de $\log \theta$ es finito y continuo como función de θ ;
- (ii) la matriz de información de Fisher $I(\theta)$ es definida positiva; y
- (iii) una parametrización del modelo resulta en un modelo diferente.

Supóngase la densidad conjunta $h(x, y|\theta)$ de las observaciones actuales x y futuras y , $\theta = \{\theta_1, \dots, \theta_d\}$, $\theta \in \mathcal{R}^d$, con densidades marginales $g(x|\theta)$ y $f(y|\theta)$. La inicial de Kuboki es

$$\pi(\theta) \propto \sqrt{\frac{|I_f(\theta)||I_g(\theta)|}{|I_h(\theta)|}}$$

donde I_f , I_g y I_h son, respectivamente, la matrices de información de Fisher para las densidades $f(\cdot|\cdot)$, $g(\cdot|\cdot)$ and $h(\cdot, \cdot|\cdot)$.

Entre los ejemplos que Kuboki estudia, encuentra que la inicial para predicción en el modelo de localización y escala es $\pi(\mu, \sigma) \propto \sigma^{-2}$ cuya predictiva carece de adecuadas propiedades de cobertura, mientras que la inicial predictiva $\pi(\mu, \sigma) \propto \sigma^{-1}$ si las cumple como se demostró anteriormente.

En general, la inicial propuesta por Kuboki se reduce a la inicial de la regla general de Jeffreys, y como se demostrará posteriormente, las densidades predictivas resultantes de esta inicial no cuentan con adecuadas propiedades de cobertura.

3.3. Inconsistencia fuerte

El trabajo de Eaton y Sudderth (1998, 1999) ilustra claramente que el uso *no crítico* de iniciales objetivas para predicción podría producir respuestas inaceptables.

En Eaton and Sudderth (1998), los autores analizan las distribuciones predictivas para la modelos Multinormales lineales, obtenidos de iniciales impropias relativamente invariantes.

Sea $Y : n \times r$ una matriz de datos basada en los datos observados Y , queremos obtener una distribución predictiva para un nuevo conjunto de datos observables $Z : m \times r$. Sean $X : n \times p$ y $W : n \times p$, ambas matrices de rango p conocido, $\beta : p \times r$ es una matriz de parámetros desconocidos, Σ es la matriz de covarianzas. El par (X, Y) es condicionalmente independiente y tiene distribución conjunta Multinormal $\begin{pmatrix} X \\ Y \end{pmatrix} \sim$

$N\left(\left(\begin{array}{c} X \\ W \end{array}\right) \beta, I_{m+n} \otimes \theta\theta'\right)$. Cada renglón de $\left(\begin{array}{c} X \\ Y \end{array}\right)$ tiene matriz de covarianzas $\Sigma = \theta\theta'$, donde $\theta \in G_T^+$ son matrices triangulares inferiores con elementos positivos G_T^+ , I_{m+n} denota la matriz identidad con rango $m+n$ y \otimes denota el producto Kronecker. Con el uso de iniciales impropias de la forma $\pi(\beta, \Sigma) \propto |\Sigma|^\alpha$ los autores encuentran la expresión general de las distribuciones predictivas para el problema de predicción MANOVA.

Eaton y Sudderth mencionan que muchas de las distribuciones predictivas recomendadas en la literatura son *fuertemente inconsistentes* en el sentido de Stone (1976). En particular, discuten los casos de las medidas invariantes por la derecha y por la izquierda, donde la primera resulta en una distribución predictiva consistente mientras que la segunda, que es la inicial de Jeffreys para el modelo MANOVA, produce una distribución predictiva fuertemente inconsistente. También, introducen una nueva distribución predictiva, que sí es consistente (Eaton and Sudderth, 1998, Teo. 4.3) y que tiene la siguiente forma

$$f(z|y) = \alpha(0)|A|^{-r/2}|S|^{-m/2}|U(z)|^{-(m+n-p)/2}$$

donde

$$\begin{aligned} A &= I_m + W(X'X)^{-1}W', \\ S &= y'(I_n - X(X'X)^{-1}X')y, \\ U(z) &= I_r + (\tau(S))^{-1}V(z)(\tau(S)')^{-1} \\ \alpha(0) &= \frac{\int_{G_T^+} |\theta|^{-n-m-p} \exp\{-\frac{1}{2}tr(\theta\theta')^{-1}\}d\theta}{\int_{G_T^+} |\theta|^{-n-p} \exp\{-\frac{1}{2}tr(\theta\theta')^{-1}\}d\theta}, \\ V(z) &= (z - W\hat{\beta})'(I_m + W(X'X)^{-1}W')^{-1}(z - W\hat{\beta}), \\ \beta &= (X'X)^{-1}X'y. \end{aligned}$$

Esta distribución predictiva es el resultado de usar la medida invariante por la *derecha* sobre G_0 . Para más detalles véase Eaton y Sudderth (1998).

En un trabajo posterior, Eaton and Sudderth (1999) demuestran que bajo ciertos supuestos, el pivoteo Fisheriano y la medida de Haar por la derecha usada como inicial, producen la misma distribución predictiva invariante, que es consistente. Más aún, demuestran que cualquier otra distribución predictiva invariante es inconsistente.

3.4. Distribuciones iniciales con validez frecuentista aproximada

En esta sección presentamos un resumen de la obtención de iniciales objetivas para predicción que cumplan con que la cobertura predictiva Bayesiana asegure una concordancia frecuentista (*match*) aproximada o exacta.

Komaki (1996) investigó la asintoticidad de segundo orden de las distribuciones predictivas considerando modelos de familias exponenciales curvadas multidimensionales.

Es una práctica común construir distribuciones predictivas *sustituyendo* (*shifting*) el estimador asintóticamente eficiente en el modelo. Usando esta práctica, Komaki encuentra que el valor medio de la divergencia de Kullback-Leibler entre la verdadera predictiva y la predictiva *sustituída* es de orden $o(n^{-1})$, donde n es el tamaño muestral. También encuentra una expresión asintótica para las distribuciones Bayesianas y discute la inicial de Jeffreys cuya predictiva resultante tiene una coincidencia de orden $o(n^{-1})$.

En esta misma línea de investigación, Datta, Mukerjee, Ghosh and Sweeting (2000) proponen obtener iniciales para predicción usando como criterio de selección la cobertura consistente asintótica y lo llaman el *criterio de consistencia de probabilidad*. Para poder caracterizar la inicial para predicción usan dos formas comunes de cobertura consistente: Cuantiles posteriores predictivos y regiones de máxima densidad predictiva.

Intervalos predictivos de un solo lado

Sea $\mathbf{x} = \{x_1, \dots, x_n\}$, n observaciones aleatorias de $f(\cdot|\boldsymbol{\theta})$ donde $\boldsymbol{\theta}$ pertenece a cualquier dimensión finita. Sea $0 < \alpha < 1$ y una densidad inicial $\tau(\boldsymbol{\theta})$, y sea x una futura observación aleatoria de $f(\cdot|\boldsymbol{\theta})$. Sea $p_\tau(x|\mathbf{x})$ la densidad posterior predictiva de x resultante de usar $\tau(\boldsymbol{\theta})$.

Estamos interesados en la inicial τ tal que

$$\Pr_\tau[x > h(\tau, \alpha)|\mathbf{x}] = \int_{h(\tau, \alpha)}^{\infty} p_\tau(x|\mathbf{x})dx = \alpha$$

donde $h(\tau, \alpha)$ es el α -cuantil superior de la densidad predictiva $p_\tau(\cdot|\mathbf{x})$ y se esperaría que se cumpliera, exacta o aproximadamente, que

$$\Pr_{\boldsymbol{\theta}}[x > h(\tau, \alpha)|\boldsymbol{\theta}] = \int_{h(\tau, \alpha)}^{\infty} f(x|\boldsymbol{\theta})dx \simeq \alpha.$$

Datta *et al.* (2000) encontraron la siguiente aproximación

$$\Pr_{\boldsymbol{\theta}}[x > h(\tau, \alpha)|\boldsymbol{\theta}] = \alpha - \frac{1}{n\tau(\boldsymbol{\theta})} D_s \{ \kappa^{st}(\boldsymbol{\theta}) \mu_t(\boldsymbol{\theta}, \alpha) \tau(\boldsymbol{\theta}) \} + o(n^{-1}) \quad (3.5)$$

donde $\mu_t(\boldsymbol{\theta}, \alpha) = \int_{q(\boldsymbol{\theta}, \alpha)}^{\infty} \frac{\partial}{\partial \theta_j} f(x|\boldsymbol{\theta})dx$, $\kappa^{ij}(\boldsymbol{\theta})$ es el i, j -ésimo elemento de $I(\boldsymbol{\theta})^{-1}$, $I(\boldsymbol{\theta})$ la

matriz de información de Fisher, y $q(\boldsymbol{\theta}, \alpha)$ satisface

$$\int_{q(\boldsymbol{\theta}, \alpha)}^{\infty} f(x|\boldsymbol{\theta})dx = \alpha.$$

Además, el lado izquierdo de la Ec. (3.5) es igual a $\alpha + o(n^{-1})$ sí y sólo si

$$\frac{\partial}{\partial \theta_s} \{ \kappa^{st}(\boldsymbol{\theta}) \mu_t(\boldsymbol{\theta}, \alpha) \tau(\boldsymbol{\theta}) \} = 0. \quad (3.6)$$

Entonces la inicial $\tau(\cdot)$ que es solución de (3.6) asegura una cobertura frecuentista aproximada hasta un orden $o(n^{-1})$. Esta inicial es invariante bajo transformaciones del parámetro, pero, y en general depende de α .

En el caso uniparamétrico, la inicial de Jeffreys, cuando existe, satisface (3.6) para todos los valores de α , sucediendo lo mismo con los modelos de localización y escala. Los autores añaden que la condición (3.6) ayuda a la comparación del comportamiento de las coberturas de probabilidad de cualesquiera iniciales disponibles.

Regiones de máxima densidad predictiva

El caso multivariado, cuando $\boldsymbol{x}^n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ son vectores aleatorios provenientes de la densidad $f(\boldsymbol{x}|\boldsymbol{\theta})$. Datta *et al.* (2000) consideran la región de máxima densidad predictiva para predecir una nueva observación \boldsymbol{x} , y deducen las condiciones bajo las cuales dicha predicción tiene una validez frecuentista adecuada.

Teniendo una región de densidad predictiva para una nueva observación \boldsymbol{x} , $H(\tau, \boldsymbol{x}^n, \alpha)$, con cobertura de probabilidad posterior predictiva α , para cada $\boldsymbol{\theta}$ y $\alpha \in (0, 1)$, existe un único cuantil $m(\boldsymbol{\theta}, \alpha)$ tal que $\Pr_{\boldsymbol{\theta}}(\boldsymbol{x} \in A(\boldsymbol{\theta}, \alpha)) = \alpha$ donde $A(\boldsymbol{\theta}, \alpha) = \{\boldsymbol{x} : f(\boldsymbol{x}|\boldsymbol{\theta}) \geq m(\boldsymbol{\theta}, \alpha)\}$.

Demuestran que

$$\Pr_{\boldsymbol{\theta}}(\boldsymbol{x} \in H(\tau, \boldsymbol{x}^n, \alpha)) = \alpha - \frac{1}{\tau(\boldsymbol{\theta})} D_s \{ \kappa^{st}(\boldsymbol{\theta}) \xi_t(\boldsymbol{\theta}, \alpha) \tau(\boldsymbol{\theta}) \} + o(n^{-1}),$$

donde

$$\xi_t(\boldsymbol{\theta}, \alpha) = \int_{A(\boldsymbol{\theta}, \alpha)} D_t f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}.$$

Puede notarse que $\Pr_{\boldsymbol{\theta}}(\boldsymbol{x} \in H(\tau, \boldsymbol{x}^n, \alpha)) = \alpha + o(n^{-1})$ sí y sólo si τ satisface la ecuación diferencial parcial

$$D_s \{ \kappa^{st}(\boldsymbol{\theta}) \xi_t(\boldsymbol{\theta}, \alpha) \tau(\boldsymbol{\theta}) \} = 0,$$

y las soluciones generales también dependen del nivel α .

Datta *et al.* (2000) proporcionan interesantes ejemplos de los modelos Normal bivariado, multivariado de escala, multivariado de localización y el de localización y escala.

Una aplicación a los modelos de regresión puede encontrarse en el trabajo de Datta y Mukerjee (2003), donde discuten los problemas de caracterizar las iniciales de cobertura consistente para predicción de variables dependientes dada una variable independiente y observaciones pasadas de ambas.

Severini, Mukerjee y Ghosh (2002) proporcionan una exploración la propiedad de la consistencia exacta de algunas de las regiones Bayesianas predictivas basada en iniciales invariantes de Haar por la derecha. En particular, su investigación da soporte a la recomendación dada por Jeffreys de usar la inicial $\pi(\mu, \sigma) \propto \sigma^{-1}$ en modelos de localización y escala $f(\cdot|\mu, \sigma)$, que es la inicial invariante de Haar por la derecha. Su resultado principal está basado en un artículo publicado por Hora y Buehler (1966) quienes también trabajaron con el problema de predicción (Hora y Buehler, 1967), específicamente la predicción puntual. Hora y Buehler (1967) se centran en las funciones invariantes predictivas (funciones de las observaciones futuras con la propiedad de ser invariante) y aseguran que los límites obtenidos de las respectivas distribuciones fiduciales son análogos a los límites de confianza en el caso frecuentista.

Mayores detalles sobre predicción y cobertura consistente puede encontrarse en Datta y Mukerjee (2004, Sec. 6).

Iniciales de cobertura consistente para efectos aleatorios

Motivados por los efectos aleatorios de los modelos ANOVA, Chang, Kim y Mukerjee (2003) caracterizan las iniciales de cobertura consistente para predicción de los efectos aleatorios no observados. Debe observarse que su propuesta de *criterio de consistencia para predicción* de los efectos aleatorios no observados arroja resultados que pudieran diferir a los dados cuando se desea hacer estimación sobre los parámetros usando iniciales de cobertura consistente.

Iniciales de cobertura consistente uniforme para predicción

En un trabajo relativamente reciente, Sweeting (2007) caracteriza a las *iniciales predictivas de cobertura consistente uniformes (UPMP)*, que son las iniciales predictivas de cobertura de probabilidad consistente para todos los valores de α .

Para cuantiles con cobertura consistente de una sola dimensión, la inicial de Jeffreys es UPMP y es única; para más de una dimensión, si es que existe una inicial UPMP esta será única. En un modelo de localización y escala $f(x|\boldsymbol{\theta}) = \frac{1}{\theta_2} f^*\left(\frac{x-\theta_1}{\theta_2}\right)$ la inicial UPMP es $\pi(\boldsymbol{\theta}) \propto \frac{1}{\theta_2}$, que es la medida de Haar por la derecha y que a su vez Jeffreys recomendó en estos modelos. Además, coincide con la obtenida por Kuboki (1998).

El caso multiparamétrico es más complicado ya que la inicial debe ser la solución de

un sistema de ecuaciones diferenciales, cuya solución puede ser única o una infinidad, dependiendo de la estructura del problema.

Sweeting (2007) discute dos modelos multiparamétricos: El primer ejemplo es el modelo Normal bivariado con vector de medias cero y elementos de la matriz de covarianza σ_1, σ_2 y ρ . Trabaja con la parametrización ortogonal dada por

$$T^{-1} = \begin{pmatrix} \theta_1 & 0 \\ \theta_2\theta_3 & \theta_2 \end{pmatrix}$$

donde $\Sigma = TT'$. Bajo esta parametrización la solución general es

$$\pi(\theta) \propto \theta_1^{-2} h(\theta_2^{-1}\theta_1, \theta_3),$$

la cual, en términos de $(\sigma_1, \sigma_2, \rho)$, $\left(\theta_1 = \frac{1}{\sigma_1}, \theta_2 = \frac{1}{\sigma_2}(1 - \rho^2)^{-1/2}, \theta_3 = -\rho\frac{\sigma_2}{\sigma_1}\right)$, es

$$\pi(\sigma_1, \sigma_2, \rho) \propto \pi_J(\sigma_1, \sigma_2, \rho) h\left(\frac{\sigma_2}{\sigma_1}, \sqrt{1 - \rho^2}\right)$$

donde $\pi_J(\sigma_1, \sigma_2, \rho) \propto \frac{1}{\sigma_1\sigma_2(1-\rho^2)^{3/2}}$ es la inicial resultante de la regla general multiparamétrica de Jeffreys y h es una función positiva arbitraria. En particular, tomando $h(x, y) = x^a y^b$, todas las iniciales que cumplen con esta forma son UPMP. Si $a = 1, b = 1/2$ resulta en la inicial conjunta $\pi(\sigma_1, \sigma_2, \rho) \propto \sigma_1^{-2}(1 - \rho^2)^{-1}$, la cual es la inicial de Haar por la derecha proveniente del grupo de transformaciones $T^{-1}X$.

El segundo ejemplo es el modelo de localización multiparamétrico en donde encuentra que la inicial Haar por la derecha es una inicial exacta UPMP, dado que el modelo es invariante a transformaciones de traslación. En este caso, la inicial de Haar por la derecha y la inicial de Jeffreys coinciden, y además es única.

Finalmente, Sweeting (2007) sugiere que en caso de tener varias iniciales predictivas de cobertura consistente se debe seleccionar aquellas que son simultáneamente de cobertura consistente predictiva y posterior.

3.5. Funciones de pérdida

Con el objetivo de encontrar una forma de obtención de iniciales objetivas para predicción con adecuadas propiedades de cobertura frecuentista, Smith (1999) propone usar teoría de la decisión y expansiones asintóticas, sugiriendo que la función de pérdida juega un papel importante bajo este esquema.

Con la ayuda de varios ejemplos, Smith compara el comportamiento de la distribución predictiva Bayesiana *versus* la distribución *plug in*, o *sustituida* o *shifting*, que es la distribución resultante de sustituir el parámetro máximo verosímil en ella y suponer que

esta es la distribución predictiva. Evalúa sus propiedades frecuentistas usando diferentes funciones de pérdida y el error cuadrático medio de la probabilidad predictiva. Concluye que los desempeños predictivos Bayesianos son mejores si se tiene un especial cuidado en la elección de la función de pérdida.

En Smith (1999) pueden encontrarse ejemplos de la distribución general de Pareto (GPD), una estimación de varias medias normales, el modelo Poisson-Gama, y una interesante discusión de su trabajo.

3.6. La densidad predictiva máximo verosímil.

Hartigan (1998) propone usar la inicial de máxima verosimilitud para estimación que ya ha sido discutida en la Sección 1.2. Pero, además señala que la densidad predictiva correspondiente minimiza el valor esperado de la distancia truncada de Kullback-Leibler, dados los datos, a la verdadera densidad. Para más detalles véase Hartigan (1998, Sec. 6).

Capítulo 4

Funciones iniciales para predicción.

En este capítulo se estudian problemas de predicción con algunos modelos univariados continuos con uno o con dos parámetros.

Recordemos el problema de predicción, en el que no es obvio cómo proceder desde el punto de vista del análisis de referencia puesto que la cantidad de interés es una función t de las observaciones futuras y no una función de los parámetros. Una primera solución podría ser utilizar la esperanza condicionada del modelo como cantidad de interés para predicción (Berger y Bernardo, 1992), o más formalmente alguna función $\phi(\theta) = g[p_x(\cdot|\theta)]$ del modelo asumido, donde g puede ser algún estimador de la función a predecir tal como la esperanza $E(t|\theta)$, la mediana $\text{Med}(t|\theta)$ o la moda $\text{Mo}(t|\theta)$.

Por otro lado, un método para poder *discriminar* entre funciones iniciales alternativas es considerar la cobertura de probabilidad de su respectiva función final (Welch y Peers, 1963, Berger y Bernardo, 1989, Datta y Sweeting, 2005). Esta técnica se utiliza en el marco de predicción como se describe en el presente capítulo.

Para cada uno de los modelos estudiados en este capítulo, se analizan una o varias de sus funciones iniciales de referencia y su respectiva densidad predictiva. De acuerdo al comportamiento de la cobertura de probabilidad de la densidad predictiva se determina la función inicial apropiada para predicción entre las posibles funciones iniciales alternativas.

Los modelos estudiados incluyen modelos regulares y modelos no regulares. Los modelos regulares estudiados son los modelos Exponencial, Cauchy, Normal y Log-Normal. Los modelos no regulares son aquellos cuyo espacio muestral depende de alguna función del parámetro del modelo. Entre estos modelos se estudian el modelo Uniforme y el Exponencial trasladado o *shifted Exponential*.

En cada uno de estos modelos, se encontrará que la inicial correspondiente a la mediana del modelo como cantidad de interés, produce una densidad predictiva con propiedades de cobertura adecuadas.

4.1. Predicción con datos Exponenciales

Supóngase que se tienen n observaciones aleatorias independientes $\{x_1, \dots, x_n\}$ provenientes de una distribución Exponencial $\text{Ex}(\cdot|\theta)$ con parámetro θ y función de densidad dada por

$$\text{Ex}(x|\theta) = \theta e^{-x\theta}, \quad \theta > 0, x > 0.$$

Puesto que solo hay un parámetro, la inicial de referencia es esencialmente única. En particular, la inicial de referencia es

$$\pi_\phi = \pi_\theta | J_\phi |$$

para cualquier transformación uno-a-uno $\phi(\theta)$ de θ . En términos de θ , la función inicial de referencia es $\pi(\theta) \propto \theta^{-1}$. Supóngase que se tiene una muestra $\mathbf{x} = \{x_1, \dots, x_n\}$ de n observaciones aleatorias provenientes de $\text{Ex}(\cdot|\theta)$, entonces la densidad final correspondiente será igual a

$$\pi(\theta|\mathbf{x}) \sim \text{Ga}(\theta|n, t)$$

donde $\text{Ga}(x|\alpha, \beta)$ es una densidad de probabilidad *Gama* con parámetros α y β (véase pág. 4), n es el tamaño de muestra y $t = \sum_{i=1}^n x_i$.

La densidad predictiva puede obtenerse fácilmente y es igual a

$$p(x|\mathbf{x}) = \frac{nt^n}{(t+x)^{n+1}} = \text{GaGa}(x|n, t, 1) \quad (4.1)$$

donde $\text{GaGa}(x|\alpha, \beta, m)$ es la densidad *Gama-Gamma* con parámetros (α, β, m) (véase pág. 4).

Una vez obtenida la densidad predictiva, nuestro interés se centra en determinar la probabilidad de que una nueva observación x pertenezca al intervalo $(x_{(1)}, x_{(2)})$ formado por las observaciones mínima y máxima de la muestra \mathbf{x} . En particular, dadas dos observaciones x_1 y x_2 , con $x_1 < x_2$, queremos saber si la cobertura de probabilidad de la densidad predictiva en el intervalo (x_1, x_2) se aproxima al valor teórico de $\frac{1}{3}$ dado en el Teorema 9, esto es, queremos el valor medio en el muestreo de

$$\begin{aligned} \Pr(x_1 < x < x_2 | x_1, x_2) &= \int_{x_1}^{x_2} \text{GaGa}(x|n, t, 1) dx \\ &= t^n \left(\frac{1}{(x_1 + t)^n} - \frac{1}{(x_2 + t)^n} \right). \end{aligned}$$

donde $n = 2$ y $t = x_1 + x_2$, que se reduce a

$$\Pr(x_1 < x < x_2 | x_1, x_2) = \frac{3(x_2 - x_1)(x_1 + x_2)^3}{(2x_1 + x_2)^2(x_1 + 2x_2)^2}. \quad (4.2)$$

En este caso la probabilidad buscada depende de las observaciones x_1 y x_2 , de forma que debe calcularse su valor esperado sobre todos los valores posibles que una muestra de tamaño dos pudiese tener, *i.e.*, el valor esperado de (4.2) evaluado con la densidad condicional conjunta de (x_1, x_2) , $p(x_1, x_2 | \theta)$.

Recordemos como es la densidad conjunta de las observaciones máxima y mínima de una muestra aleatoria. Supóngase que una muestra aleatoria $\{x_1, \dots, x_n\}$ de tamaño n con función de densidad f y función de distribución F , considere las observaciones $x_{(1)} = \min\{x_1, \dots, x_n\}$ y $x_{(n)} = \max\{x_1, \dots, x_n\}$. La función de densidad conjunta de $x_{(1)}$ y $x_{(n)}$ es

$$g(x_{(1)}, x_{(n)}) = n(n-1)(F(x_{(n)}) - F(x_{(1)}))^{n-2} f(x_{(1)})f(x_{(n)}), \\ -\infty < x_{(1)} < x_{(n)} < \infty.$$

Si la muestra es de tamaño dos $\{x_1, x_2\}$, donde $x_{(1)} = \min\{x_1, x_2\}$ y $x_{(2)} = \max\{x_1, x_2\}$, la función de densidad conjunta de $(x_{(1)}, x_{(2)})$ es

$$g(x_{(1)}, x_{(2)}) = 2f(x_{(1)})f(x_{(2)}), \quad (4.3) \\ -\infty < x_{(1)} < x_{(2)} < \infty.$$

La densidad conjunta bajo el supuesto $x_1 < x_2$ es la misma que la densidad conjunta de las observaciones mínima y máxima de una muestra de tamaño dos.

Sean $x_{(1)} = \min\{x_1, x_2\} = x_1$ y $x_{(2)} = \max\{x_1, x_2\} = x_2$, la densidad conjunta de x_1 y x_2 bajo la condición $x_1 < x_2$ está dada por

$$p(x_1, x_2 | \theta) = 2\text{Ex}(x_1 | \theta)\text{Ex}(x_2 | \theta), \quad 0 < x_1 < x_2 < \infty, \theta > 0.$$

Por lo tanto, el valor esperado de (4.2) es

$$\mathbb{E}_{x_1, x_2 | \theta} \left[\frac{3(x_2 - x_1)(x_1 + x_2)^3}{(2x_1 + x_2)^2(x_1 + 2x_2)^2} \right] = 2 \int_0^\infty \int_0^{x_2} \frac{3(x_2 - x_1)(x_1 + x_2)^3}{(2x_1 + x_2)^2(x_1 + 2x_2)^2} \theta^2 e^{-\theta(x_1 + x_2)} dx_1 dx_2 \\ = \frac{1}{3}.$$

Esto demuestra que el valor esperado de la cobertura definida por (4.2) *no depende* de θ y es consistente con el resultado del Teorema 9. Consecuentemente, la inicial de referencia $\pi(\theta) \propto \theta^{-1}$ produce una densidad predictiva con adecuadas propiedades de

cobertura.

4.2. Predicción con datos Cauchy

Un modelo interesante es el modelo Cauchy, que pertenece a la familia de modelos de localización y escala, pero con la particularidad de que carece de media. Sin embargo, la mediana y la moda están bien definidas y son iguales al parámetro de localización.

Supóngase que se tiene una muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_n\}$ de tamaño n proveniente de una distribución Cauchy con parámetros de localización y escala μ y σ , respectivamente.

La función de densidad está dada por

$$\text{Ca}(x|\mu, \sigma) = \frac{1}{\pi\sigma} \left(1 + \frac{(x - \mu)^2}{\sigma^2}\right)^{-1}, \quad x \in \mathcal{R}, \sigma > 0, \mu \in \mathcal{R}.$$

Puesto que es un modelo regular de localización-escala, la inicial de referencia conjunta cuando μ o σ es el parámetro de interés es

$$\pi(\mu, \sigma) \propto \sigma^{-1}$$

(Fernández y Steel, 1999).

Con esta inicial de referencia la densidad final es

$$\pi(\mu, \sigma|\mathbf{x}) \propto \sigma^{-1} (\pi\sigma)^{-n} \prod_{i=1}^n \left(1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right)^{-1}.$$

Consecuentemente, la densidad predictiva de referencia es

$$p(x|\mathbf{x}) = \pi^{-1} \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma^n} \left(1 + \left(\frac{x - \mu}{\sigma}\right)^2\right)^{-1} \pi(\mu, \sigma|\mathbf{x}) d\mu d\sigma.$$

En el caso particular de $n = 2$, $\mathbf{x} = \{x_1, x_2\}$, la densidad final y predictiva tienen respectivamente las siguientes expresiones analíticas,

$$\pi(\mu, \sigma|x_1, x_2) = \frac{2\sigma^2|x_1 - x_2|}{\pi^3 \left(((x_1 - \mu)^2 + \sigma^2)((x_2 - \mu)^2 + \sigma^2)((x - \mu)^2 + \sigma^2) \right)},$$

$$\begin{aligned}
p(x|x_1, x_2) &= \\
&= \frac{(x_1 - x_2) \log((x_1 - x_2)^2) + (x - x_1) \log((x_1 - x)^2) + (x_2 - x) \log((x_2 - x)^2)}{2\pi^2(x_1 - x)(x - x_2)\text{Sign}(x_1 - x_2)} \\
&= \frac{1}{2\pi\text{Sign}(x_1 - x_2)} \left\{ \frac{1}{x_1 - x} \log \left[\left(\frac{x_1 - x_2}{x_2 - x} \right)^2 \right] + \frac{1}{x_2 - x} \log \left[\left(\frac{x_1 - x}{x_1 - x_2} \right)^2 \right] \right\} \\
&\quad (x_1 \neq x_2, x \neq x_1, x \neq x_2)
\end{aligned}$$

donde $\text{Sign}[x]$ toma en los valores $-1, 0$ ó 1 , dependiendo de si x es negativo, cero o positivo.

Finalmente, la probabilidad de que una nueva observación x pertenezca al intervalo formado por (x_1, x_2) , $x_1 < x_2$, es

$$\begin{aligned}
\Pr(x_1 < x < x_2|x_1, x_2) &= \int_{x_1}^{x_2} p(x|x_1, x_2)dx \\
&= \frac{1}{2\pi^2\text{Sign}(x_1 - x_2)} \times \left\{ \begin{array}{l} 2 \log \left[\frac{x_1 - x}{x_1 - x_2} \right] - \log \left[\left(\frac{x_1 - x}{x_1 - x_2} \right)^2 \right] \log(x - x_2) \\ - \log(x - x_1) \left(\log \left[\left(\frac{x_1 - x_2}{x_2 - x} \right)^2 \right] + 2 \log \left[\frac{x - x_2}{x_1 - x_2} \right] \right) \\ - 2\text{PolyLog} \left(2, \frac{x_1 - x}{x_1 - x_2} \right) + 2\text{PolyLog} \left(\frac{x - x_2}{x_1 - x_2} \right) \end{array} \right\} \\
&= -\frac{2}{2\pi^2} \left[\text{PolyLog} \left(2, \frac{x - x_2}{x_1 - x_2} \right) - \text{PolyLog} \left(2, \frac{x_1 - x}{x_1 - x_2} \right) \right]_{x_1}^{x_2} \\
&= -\frac{1}{\pi^2} (\text{PolyLog}(2, 0) - \text{PolyLog}(2, 1) - \text{PolyLog}(2, 1) + \text{PolyLog}(2, 0)),
\end{aligned}$$

donde $\text{Sign}(x_1 - x_2) = -1$ pues $x_1 < x_2$ y $\text{PolyLog}(n, z) \equiv \text{Li}_n(z)$ es la función polilogaritmo, que es igual a

$$\text{Li}_n(z) = \frac{(-1)^{n-1}}{(n-2)!} \int_0^1 \frac{\log^{n-2}(t) \log(1-zt)}{t} dt,$$

entonces

$$\text{Li}_2(0) = - \int_0^1 \frac{\log(1)}{t} dt = 0 \quad \text{y} \quad \text{Li}_2(1) = - \int_0^1 \frac{\log(1-zt)}{t} dt = \frac{\pi^2}{6},$$

de acuerdo a la Ec. (4.291.2) de Gradshteyn y Ryzhik (1994) la ecuación $\text{Li}_2(1) = \frac{\pi^2}{6}$.

Sustituyendo, resulta

$$\Pr(x_1 < x < x_2|x_1, x_2) = -\frac{1}{\pi^2} \left(-\frac{\pi^2}{3} \right) = \frac{1}{3}.$$

Por tanto, la cobertura $\Pr(x_1 < x < x_2 | x_1, x_2)$ no depende de las observaciones (x_1, x_2) y es igual a $\frac{1}{3}$. Un resultado similar se obtiene en el caso $x_1 > x_2$.

A la luz de este ejemplo, uno esperaría que para cualquier modelo de localización (μ) y escala (σ), la inicial de referencia dada por σ^{-1} produzca una densidad predictiva con adecuadas propiedades de cobertura.

Nótese además, que en el modelo Cauchy la mediana y la moda coinciden con el valor del parámetro de localización, surgiendo de manera natural la elección de la mediana o la moda como cantidades de interés en el problema de predicción.

El ejemplo siguiente ilustra la influencia de la distribución inicial en las propiedades de cobertura de la distribución predictiva.

4.3. Predicción con datos Normales

Considérese ahora una muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_n\}$ de tamaño n proveniente de una distribución Normal $N(x|\mu, \sigma)$. Sea la familia de funciones iniciales conjuntas de $\{\mu, \sigma\}$ dadas por

$$\pi(\mu, \sigma) \propto \sigma^{-\alpha}, \quad \alpha > 0. \quad (4.4)$$

La densidad final es proporcional a

$$\pi(\mu, \sigma | \mathbf{x}) \sim N\left(\mu | \bar{x}, \frac{\sigma^2}{n}\right) \times \text{Ga}^{-1/2}\left(\sigma \left| \frac{n + \alpha - 1}{2}, \frac{ns^2}{2}\right.\right),$$

donde $\text{Ga}^{-1/2}(\cdot | \gamma, \beta)$ denota la *raíz cuadrada* de una densidad *Gama inversa* (squared-root inverted-gamma) con parámetros γ y β (véase pág.4), y donde s y \bar{x} son los estimadores máximo verosímiles de σ y μ , respectivamente.

La densidad predictiva es igual a

$$p(x | \mathbf{x}) = \left(1 + \frac{(x - \bar{x})^2}{(n + 1)s^2}\right)^{-\frac{n + \alpha - 1}{2}} \sim \text{St}\left(x | \bar{x}, s\sqrt{\frac{n + 1}{n + \alpha - 2}}, n + \alpha - 2\right),$$

donde $\text{St}(\cdot | \theta, \lambda, \nu)$ denota una densidad *t-Student* con parámetros de localización θ y escala λ , y con ν grados de libertad.

Consecuentemente, dadas dos observaciones aleatorias x_1 y x_2 , la probabilidad de que una nueva observación x se encuentre contenida en el intervalo (x_1, x_2) , $x_1 < x_2$, es

$$\Pr(x_1 < x < x_2 | x_1, x_2) = c \int_{x_1}^{x_2} \left(1 + \frac{(x - \bar{x})^2}{s^2(n + 1)}\right)^{-\frac{n + \alpha - 1}{2}} dx \quad (4.5)$$

$$\text{con } c = \frac{\Gamma(\frac{n+\alpha-1}{2})}{\Gamma(\frac{n+\alpha-2}{2})\Gamma(1/2)} \frac{1}{\sqrt{s^2(n+1)}}.$$

Puesto que $n = 2$, y haciendo el cambio de variable

$$t = \frac{x - \bar{x}}{\sqrt{3}s}, \quad dt = \frac{1}{\sqrt{3}s},$$

el dominio se transforma en

$$0 < t < \frac{|x_2 - x_1|}{2\sqrt{3}s} = \frac{1}{\sqrt{3}} = t_1 \quad \text{ya que } s = \frac{|x_2 - x_1|}{2}.$$

Por tanto, la integral (4.5) se transforma en

$$\Pr(x_1 < x < x_2 | x_1, x_2) = 2 \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha}{2})\sqrt{\pi}} \int_0^{\frac{1}{\sqrt{3}}} (1+t^2)^{-(\alpha+1)/2} dt,$$

que depende únicamente de α .

El comportamiento de $\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x})$ puede verse gráficamente en la Figura (4.1). Obsérvese que la probabilidad de $x \in (x_1, x_2)$ tiende a uno conforme α crece.

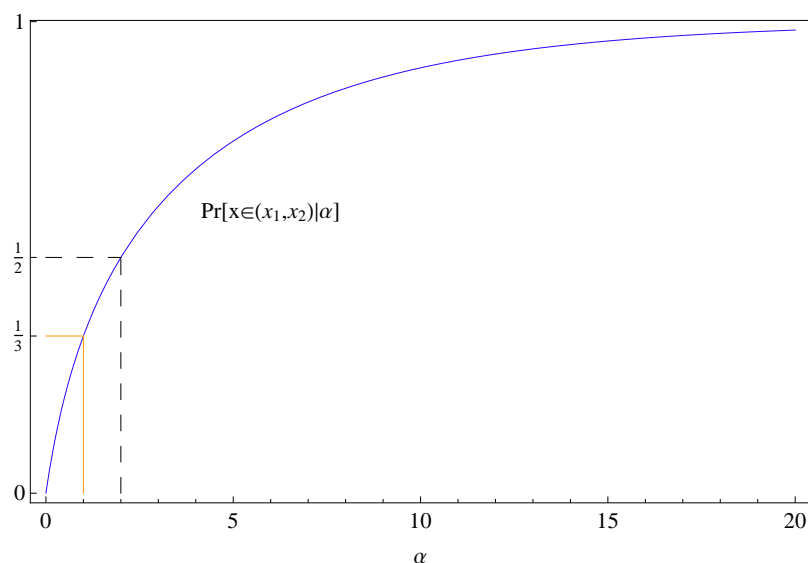


Figura 4.1: Comportamiento de la cobertura de la densidad predictiva resultante de la familia de distribuciones iniciales $\pi(\mu, \sigma) \propto \sigma^{-\alpha}$ ($\alpha > 0$), en el modelo $N(\cdot | \mu, \sigma)$.

Supóngase que la cantidad de interés es el vector $\boldsymbol{\theta} = \{\mu, \sigma\}$; se sabe que la inicial de referencia es entonces la función inicial multiparamétrica de Jeffreys $\pi(\mu, \sigma) \propto \sigma^{-2}$, que es también la medida de Haar por la izquierda. En este caso $\alpha = 2$ dando como resultado que la probabilidad requerida resulta $\Pr(x_1 < x_3 < x_2 | \alpha = 2) = \frac{1}{2}$.

Supóngase ahora que la cantidad de interés es el parámetro de localización μ , que es

simultáneamente la media, la moda y la mediana de x dados μ y σ , por tanto la inicial de referencia correspondiente es $\pi(\mu, \sigma) \propto \sigma^{-1}$, que es la medida de Haar por la derecha. Consecuentemente $\alpha = 1$ y la probabilidad requerida es exacta $\Pr(x_1 < x < x_2 | \alpha = 1) = \frac{1}{3}$.

Debe observarse que dentro de la familia de iniciales relativamente invariantes $\pi(\mu, \sigma) \propto \sigma^{-\alpha}$, la única inicial que produce una densidad predictiva cuya cobertura coincide con el resultado del Teorema 9 es la que corresponde al valor de $\alpha = 1$. Que a su vez, es igual a la inicial de referencia cuando la cantidad de interés es el parámetro de localización del modelo.

La expresión analítica de la densidad predictiva dadas dos observaciones $x_1 \neq x_2$ está dada por

$$p(x|x_1, x_2) = \left(1 + \frac{(x - \bar{x})^2}{3s^2}\right)^{-\frac{\alpha+1}{2}}$$

con $\bar{x} = \frac{x_1+x_2}{2}$ y $s = \frac{|x_2-x_1|}{2}$, y por tanto para $\alpha = 1$,

$$p(x|x_1, x_2) = \left(1 + \frac{1}{3} \left(\frac{2x - (x_1 + x_2)}{x_2 - x_1}\right)^2\right)^{-1}.$$

4.4. Predicción con datos Log-Normales

Considérese ahora el modelo de localización y escala Log-Normal con función de densidad dada por

$$\begin{aligned} \text{LgN}(x|\mu, \sigma) &= \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\} \\ x &> 0, \mu \in \mathcal{R}, \sigma > 0, \end{aligned}$$

de forma que $y = \log x$ tiene una densidad Normal $N(y|\mu, \sigma)$, y la matriz de información de Fisher es

$$I(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \quad (4.6)$$

Como demuestran Fernández y Steel (1999), si la cantidad de interés es μ o es σ , la inicial de referencia será $\pi(\mu, \sigma) \propto \sigma^{-1}$, que es también con la inicial de Haar por la derecha. La densidad final correspondiente es

$$\pi(\mu, \sigma|\mathbf{x}) = \frac{2^{1-n/2} s^{n-1} n^{n/2}}{\Gamma(\frac{n-1}{2})\sqrt{\pi}} \sigma^{-(n+1)} \exp\left\{-\frac{ns^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right\}$$

donde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n \log x_i = \log(x_1 \dots x_n)^{1/n} \quad \text{y} \quad s^2 = \frac{1}{n} \sum_{i=1}^n (\log x_i - \bar{x})^2.$$

La densidad predictiva para una nueva observación dados los datos \mathbf{x} es

$$\begin{aligned} p(x|\mathbf{x}) &= \frac{s^{n-1} \Gamma(n/2)}{\sqrt{\pi(n+1)} x \Gamma(\frac{n-1}{2}) \left(s^2 + \frac{(\bar{x} - \log x)^2}{n+1} \right)^{n/2}} \\ &\sim \text{LSt}(x | \exp(\bar{x}), s \sqrt{\frac{n+1}{n-1}}, n-1) \end{aligned}$$

donde $\text{LSt}(x|\mu, \nu, \alpha)$ denota la función de densidad *Log-Student generalizada* definida por

$$\begin{aligned} \text{LSt}(x|\mu, \nu, \alpha) &= \frac{\nu^\alpha}{x \alpha^{1/2} \text{Beta}(1/2, \alpha/2)} \left\{ \nu^2 + \frac{1}{\alpha} \left(\log \frac{x}{\mu} \right)^2 \right\}^{-\frac{\alpha+1}{2}} \\ &x > 0, \mu > 0, \nu > 0, \alpha > 0 \end{aligned}$$

con α grados de libertad (Good y Crook, 1987).

A continuación calculemos la probabilidad del suceso $B = \{x : x_{(1)} < x < x_{(n)}\}$

$$\Pr(B|\mathbf{x}) = \int_{x_{(1)}}^{x_{(n)}} \text{LSt}(x|\mu, \nu, \alpha) dx = F_{ls}(x_{(n)}|\mu, \nu, \alpha) - F_{ls}(x_{(1)}|\mu, \nu, \alpha) \quad (4.7)$$

donde $F_{ls}(\cdot|\mu, \nu, \alpha)$ es la función de distribución acumulada de *Log-Student* anteriormente descrita, $\mu = e^{\bar{x}}$, $\nu = s \sqrt{\frac{n+1}{n-1}}$ y $\alpha = n - 1$ grados de libertad. Obsérvese que en general, se trata de una función de las observaciones y no únicamente del tamaño muestral.

Cuando solo se tienen dos observaciones el resultado es exacto como se verá a continuación. Dadas x_1 y x_2 ($x_1 < x_2$), la densidad predictiva es

$$p(x|x_1, x_2) = \frac{\sqrt{3} \left| \log \frac{x_1}{x_2} \right|}{\pi x (3(\log^2 x + \log^2 x_1 + \log^2 x_2) - (\log x x_1 x_2)^2)}$$

y la probabilidad predictiva de que una nueva observación x pertenezca al intervalo

(x_1, x_2) es (omitiendo algunos cálculos)

$$\begin{aligned} \Pr[x_1 < x < x_2 | x_1, x_2] &= \int_{x_1}^{x_2} p(x | x_1, x_2) dx \\ &= \frac{|\log x_1/x_2|}{\pi \sqrt{-\log^2 x_1/x_2}} \operatorname{arctanh} \left(\frac{\log x_1/x_2}{\sqrt{-3 \log^2 x_1/x_2}} \right) \Big|_{x_1}^{x_2} \\ &= \frac{2}{\pi i} \operatorname{arctanh} \left(\frac{\log x_1/x_2}{\sqrt{3i} |\log x_1/x_2|} \right) = \frac{2}{\pi i} \operatorname{arctanh} \left(\frac{\operatorname{Sign}[\log x_1/x_2]}{\sqrt{3i}} \right) \end{aligned}$$

donde $\operatorname{Sign}[x] = 0, -1, 1$ dependiendo de que x sea cero, negativo o positivo.

Puesto que $x_1 < x_2$, si $\log \frac{x_1}{x_2} < 0 \in \mathcal{R}$, se tiene que

$$\Pr[x_1 < x < x_2 | x_1, x_2] = \frac{2}{\pi i} \operatorname{arctanh} \left(-\frac{1}{i\sqrt{3}} \right) = \frac{2}{\pi i} \frac{i\pi}{6} = \frac{1}{3} \quad (4.8)$$

para todos los posibles valores (x_1, x_2) . El caso $x_1 > x_2$ arroja el mismo resultado.

Con más de dos observaciones y de acuerdo al Teorema 10, cabría esperar que el valor medio en el muestreo de la función (4.7) evaluado sobre todas las muestras de tamaño n *debiera* ser cercano a $\frac{n-1}{n+1}$ para todos los valores de μ y σ , y la Tabla 4.1 confirma que esto es así.

n	μ	E[Pr($x_{(1)} < x < x_{(n)} \mathbf{x}$)]				$\frac{n-1}{n+1}$
		$\sigma = 1$		$\sigma = 2$		
		Media	DvSt	Media	DvSt	
2	0	0.3333	-	0.3333	-	1/3
	1	0.3333	-	0.3333	-	
3	0	0.5002	0.0004	0.5003	0.0004	1/2
	1	0.4950	0.0004	0.5004	0.0004	
4	0	0.5998	0.0008	0.6006	0.0008	0.6
	1	0.6007	0.0008	0.6005	0.0009	
7	0	0.7496	0.0014	0.7501	0.0014	0.75
	1	0.7500	0.0015	0.7507	0.0014	
9	0	0.7997	0.0015	0.8003	0.0015	0.8
	1	0.7994	0.0015	0.8004	0.0015	
15	0	0.8748	0.0011	0.8748	0.0012	0.875
	1	0.8751	0.0012	0.8751	0.0012	

Tabla 4.1: Valor medio y desviación estándar de las probabilidades $\Pr[x_{(1)} < x < x_{(n)} | x_1, \dots, x_n]$ correspondientes a $m = 5000$ muestras simuladas de tamaño $n \in \{2, 3, 4, 7, 9, 15\}$ provenientes de $\operatorname{LgN}(x | \mu, \sigma)$, $\mu \in \{0, 1\}$, $\sigma \in \{1, 2\}$ con inicial dada por $\pi(\mu, \sigma) \propto \sigma^{-1}$

Los valores en la Tabla 4.1 fueron obtenidos de la siguiente manera:

- Genérese una muestra aleatoria $\{x_1, \dots, x_n\}$ de tamaño n proveniente de una distribución $\text{LgN}(x|\mu, \sigma)$.
- Obténgase la densidad final $\pi(\mu, \sigma|x_1, \dots, x_n)$ y la correspondiente densidad predictiva $p(x|x_1, \dots, x_n)$.
- Calcúlese $\Pr[x_{(1)} < x < x_{(n)}|x_1, \dots, x_n] = \int_{x_{(1)}}^{x_{(n)}} p(x|x_1, \dots, x_n)$ donde $x_{(1)} = \min\{x_1, \dots, x_n\}$ y $x_{(n)} = \max\{x_1, \dots, x_n\}$.
- Repítanse los tres pasos anteriores m veces. Calcúlese el valor promedio y la desviación estándar de las m coberturas de probabilidad.

Para m grande el valor medio de la cobertura de probabilidad de la densidad predictiva se aproximará a su valor esperado. Se observa que éste se aproxima al valor teórico dado en el Teorema 10.

Ahora nos planteamos para que medida de tendencia central (media, moda o mediana), definida como cantidad de interés, resultaría en una densidad predictiva con propiedades de cobertura similares.

En la siguiente subsección se desarrollará con detalle la obtención de la inicial de referencia cuando la media es la cantidad de interés. Los casos de la mediana y la moda se omiten por ser similares.

4.4.1. La media $E(x|\mu, \sigma)$ como cantidad de interés

Se sabe que en el modelo Log normal la media es igual a $E(x|\mu, \sigma) = e^{\mu + \sigma^2/2}$. Considérese pues la función

$$\phi_1 = E(x|\mu, \sigma) = \exp\{\mu + \sigma^2/2\}.$$

Asúmase la parametrización $\psi = (\phi_1, \sigma) = g(\mu, \sigma)$, donde ahora ϕ_1 es nuestro parámetro de interés y σ el parámetro de ruido. El Jacobiano de la transformación inversa $g^{-1}(\phi_1, \sigma)$ es

$$J_{g^{-1}}(\psi) = \begin{pmatrix} \frac{1}{\phi_1} & -\sigma \\ 0 & 1 \end{pmatrix}.$$

Usando el supuesto de normalidad asintótica bajo transformaciones (Corolario 1 de la Proposición 5.17, Bernardo y Smith, 1994), se tiene que la matriz de información de

Fisher en términos de (ϕ_1, σ) está dada por

$$H_\psi(\psi) = J_{g^{-1}}^t(\psi) I_{g^{-1}}(\psi) J_{g^{-1}} = \begin{pmatrix} \frac{1}{\phi_1^2 \sigma^2} & -\frac{1}{\sigma \phi_1} \\ -\frac{1}{\sigma \phi_1} & 1 + \frac{2}{\sigma^2} \end{pmatrix},$$

y la matriz de dispersión $S(\varphi) = H_\psi(\psi)^{-1}$ es

$$S(\varphi) = \frac{1}{2} \begin{pmatrix} \phi_1^2 \sigma^2 (2 + \sigma^2) & \phi_1 \sigma^3 \\ \phi_1 \sigma^3 & \sigma^2 \end{pmatrix}.$$

Defínase una sucesión creciente de intervalos cerrados de la forma

$$\Lambda_i = \{\sigma : e^{-i} \leq \sigma \leq e^i\}, \quad i = 1, 2, \dots$$

independientes de ϕ_1 .

De H_ψ pueden obtenerse las iniciales de referencia al identificar productos de funciones independiente de ϕ_1 y σ (Corolario de la Proposición 5.29, Bernardo y Smith, 1994). Específicamente,

$$\begin{aligned} \pi(\sigma|\phi_1) \propto \sqrt{h_{22}} &= \sqrt{1 + \frac{2}{\sigma^2}} = f_2(\phi) \times g_2(\sigma) \\ \sqrt{h_\phi} &= \sqrt{h_{11} - h_{12} h_{22}^{-1} h_{21}} \\ &= \frac{1}{\sigma \phi_1} (1 - (1 + \frac{2}{\sigma^2}))^{-1/2} = f_1(\phi) \times g_1(\sigma). \end{aligned}$$

Entonces, la inicial de referencia para la parametrización ordenada (ϕ_1, σ) es

$$\pi(\phi_1, \sigma) \propto f_1(\phi) g_2(\sigma) = \phi_1^{-1} \sqrt{1 + \frac{2}{\sigma^2}}.$$

En términos de los parámetros originales, la inicial de referencia es

$$\pi_{\phi_1}(\mu, \sigma) \propto \sqrt{1 + \frac{2}{\sigma^2}},$$

que difiere de la inicial de referencia $\pi(\mu, \sigma) \propto \sigma^{-1}$ cuando μ es el parámetro de interés.

Dadas dos observaciones x_1 y x_2 , la densidad final será proporcional a

$$\pi_{\phi_1}(\mu, \sigma | x_1, x_2) \propto \frac{1}{2\pi\sigma^2 x_1 x_2} \sqrt{1 + \frac{2}{\sigma^2}} \exp \left\{ -\frac{((\log x_1 - \mu)^2 + (\log x_2 - \mu)^2)}{2\sigma^2} \right\},$$

que es impropia (véase el Apéndice A), lo que implica que con esta distribución inicial no pueden hacerse inferencias con solo dos observaciones.

Considérese ahora una muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_n\}$ de tamaño n con $n > 2$; la densidad final será proporcional a

$$\pi_{\phi_1}(\mu, \sigma | \mathbf{x}) \propto \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n \frac{1}{\prod_{i=1}^n x_i} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\log x_i - \mu)^2\right\} \sqrt{1 + \frac{2}{\sigma^2}}$$

y la densidad predictiva para una nueva observación será

$$p_{\phi_1}(x | \mathbf{x}) = \int_0^\infty \int_{-\infty}^\infty \text{LN}(x | \mu, \sigma) \pi_{\phi_1}(\mu, \sigma | \mathbf{x}) d\mu d\sigma.$$

que no cuenta con una expresión analítica explícita.

Como se ha procedido anteriormente, deseamos saber la probabilidad de una nueva observación x se encuentra contenida en el intervalo $(x_{(1)}, x_{(n)})$ formado por las observaciones mínima y máxima de la muestra \mathbf{x} . En particular, deseamos saber si la cobertura de probabilidad la densidad predictiva $p_{\phi_1}(x | \mathbf{x})$ en $(x_{(1)}, x_{(n)})$, *i. e.*,

$$\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x}) = \int_{x_{(1)}}^{x_{(n)}} p_{\phi_1}(x | \mathbf{x}) dx, \quad (4.9)$$

se aproxima al valor teórico $\frac{n-1}{n+1}$ del Teorema 10.

Dado que no se cuenta con una expresión analítica de la densidad predictiva $p_{\phi_1}(x | \mathbf{x})$, el valor esperado de la integral (4.9) será aproximado numéricamente y la Tabla 4.2 muestra el resultado de dichas aproximaciones (véanse las columnas correspondientes a π_{ϕ_1}), que ponen de manifiesto una cobertura inapropiada.

4.4.2. La moda $\text{Mo}(x | \mu, \sigma)$ como cantidad de interés

Considérese ahora la moda del modelo $\text{Mo}(x | \mu, \sigma) = \exp\{\mu - \sigma^2\}$ y sea

$$\phi_2 = \text{Mo}(x | \mu, \sigma) = \exp\{\mu - \sigma^2\}$$

con la que se tiene la parametrización (ϕ_2, σ) . Análogo a la obtención de la inicial de referencia π_{ϕ_1} , la inicial de referencia para el orden (ϕ_2, σ) es

$$\pi_{\phi_2}(\phi_2, \sigma) \propto \phi_2^{-1} \sqrt{4 + 2/\sigma^2}$$

la cual expresada en los parámetros originales es

$$\pi_{\phi_2}(\mu, \sigma) \propto \sqrt{4 + 2/\sigma^2},$$

que es diferente a la inicial de referencia cuando μ es la cantidad de interés.

Dadas dos observaciones, la densidad final será proporcional a

$$\pi_{\phi_2}(\mu, \sigma | x_1, x_2) \propto \frac{\sqrt{4 + \frac{2}{\sigma^2}}}{2\pi\sigma^2 x_1 x_2} \exp \left\{ -\frac{((\log x_1 - \mu)^2 + (\log x_2 - \mu)^2)}{2\sigma^2} \right\}$$

que nuevamente es impropia (véase el Apéndice A).

Supóngase nuevamente que se tiene una muestra de observaciones aleatorias $\mathbf{x} = \{x_1, \dots, x_n\}$ con $n > 2$; la densidad final será proporcional a

$$\pi_{\phi_2}(\mu, \sigma | \mathbf{x}) \propto \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \frac{1}{\prod_{i=1}^n x_i} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (\log x_i - \mu)^2 \right\} \sqrt{4 + \frac{2}{\sigma^2}}$$

y la densidad predictiva es

$$p_{\phi_2}(x | \mathbf{x}) = \int_0^\infty \int_{-\infty}^\infty \text{LN}(x | \mu, \sigma) \pi_{\phi_2}(\mu, \sigma | \mathbf{x}) d\mu d\sigma$$

para la que no se tiene una expresión analítica.

El objetivo es verificar las propiedades de cobertura de $p_{\phi_2}(x | \mathbf{x})$ en el intervalo $(x_{(1)}, x_{(n)})$, *i.e.*,

$$\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x}) = \int_{x_{(1)}}^{x_{(n)}} p_{\phi_2}(x | \mathbf{x}) dx,$$

que se esperararía concuerden con el valor teórico $\frac{n-1}{n+1}$ dado en el Teorema 10. Como no se cuenta con una expresión analítica de $p_{\phi_2}(x | \mathbf{x})$, el valor esperado de la integral anterior será aproximado numéricamente donde las columnas π_{ϕ_2} de la Tabla 4.2 muestran estos valores. Obsérvese, que como en el caso de la media, no se obtienen coberturas apropiadas.

4.4.3. La mediana $\text{Me}[x | \mu, \sigma]$ como cantidad de interés

Finalmente, considérese ahora la mediana del modelo $\text{Me}[x | \mu, \sigma] = e^\mu$ y sea la transformación

$$\phi_3 = \text{Me}[x | \mu, \sigma] = e^\mu.$$

Puesto que ϕ_3 es una transformación uno-a-uno de μ , la inicial de referencia cuando ϕ_3 es la cantidad de interés es igual a la inicial de referencia cuando el parámetro de interés es μ , $\pi_{\phi_3}(\mu, \sigma) \propto \sigma^{-1}$.

Lo anterior puede comprobarse fácilmente al obtener directamente la inicial de refe-

rencia en términos de la parametrización (ϕ_3, σ) resulta ser

$$\pi_{\phi_3}(\phi_3, \sigma) \propto \frac{1}{\phi_3 \sigma}$$

y en términos de los parámetros originales (μ, σ) es

$$\pi_{\phi_3}(\mu, \sigma) \propto \sigma^{-1}$$

Cuando la función inicial de referencia es proporcional a σ^{-1} ya hemos establecido en (4.8) que su respectiva densidad predictiva dadas dos observaciones x_1 y x_2 ($x_1 < x_2$), tiene una cobertura exactamente igual a $\frac{1}{3}$ coincidiendo con el valor teórico del Teorema 9.

Se esperaría que para más de dos observaciones aleatorias la cobertura de probabilidad de la densidad predictiva en el intervalo $(x_{(1)}, x_{(n)})$ fuese aproximado al valor $\frac{n-1}{n+1}$. La Tabla 4.2 (columnas π_{ϕ_3}) muestra las aproximaciones al valor esperado de $\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x})$, que demuestran una concordancia con el Teorema 10.

Los valores en la Tabla 4.2 fueron calculados así:

- (i) Genérese una muestra aleatoria de tamaño n , $\mathbf{x} = \{x_1, \dots, x_n\}$ de $\text{LgN}(\cdot | \mu, \sigma)$.
- (ii) Obténgase $\pi(\mu, \sigma | \mathbf{x})$ y $p(x | \mathbf{x})$.
- (iii) Calcúlese $\Pr(x_{(1)} < x < x_{(n)} | \mathbf{x}) = \int_{x_{(1)}}^{x_{(n)}} p(x | \mathbf{x}) dx$ donde $x_{(1)} = \min\{x_1, \dots, x_n\}$ y $x_{(n)} = \max\{x_1, \dots, x_n\}$. Guardar en memoria.
- (iv) Repítanse los pasos (i) a (iii) m veces, promédiese y calcúlese la desviación estándar de las coberturas.

Este procedimiento se aplica a cada inicial de referencia π_{ϕ_1} , π_{ϕ_2} y π_{ϕ_3} .

n	μ	$\sigma = 1$			$\sigma = 2$			$\frac{n-1}{n+1}$ $\frac{1}{3} = 0.3333$
		π_{ϕ_1}	π_{ϕ_2}	π_{ϕ_3}	π_{ϕ_1}	π_{ϕ_2}	π_{ϕ_3}	
2	0	-	-	0.3333 (0×10^{-4})	-	-	0.3333 (0×10^{-4})	$\frac{1}{3} = 0.3333$
	1	-	-	0.3333 0	-	-	0.3333 (0×10^{-4})	
3	0	0.6753 (0.015)	0.6659 (0.016)	0.5000 (0×10^{-3})	0.6634 (0.102)	0.6452 (0.105)	0.5001 (0×10^{-3})	$\frac{1}{2} = 0.5$
	1	0.6744 (0.015)	0.6655 (0.016)	0.5001 (0×10^{-3})	0.6606 (0.102)	0.6474 (0.105)	0.5000 (0×10^{-3})	
4	0	0.7518 (0.01)	0.7459 (0.010)	0.6000 (0×10^{-3})	0.8769 (0.025)	0.8721 (0.027)	0.6001 (0×10^{-3})	$\frac{3}{5} = 0.6$
	1	0.7523 (0.01)	0.7464 (0.010)	0.6001 (0×10^{-3})	0.8767 (0.025)	0.8720 (0.029)	0.6002 (0×10^{-3})	
7	0	0.8509 (0.004)	0.8481 (0.005)	0.7503 (0.001)	0.9641 (0.0022)	0.9640 (0.002)	0.7500 (0.001)	$\frac{3}{4} = 0.75$
	1	0.8517 (0.004)	0.8473 (0.005)	0.7500 (0.001)	0.9639 (0.002)	0.9669 (0.0018)	0.7502 (0.001)	
9	0	0.8822 (0.003)	0.8793 (0.003)	0.7999 (0.001)	0.9797 (0.001)	0.9794 (0.001)	0.7998 (0.001)	$\frac{4}{5} = 0.8$
	1	0.8812 (0.003)	0.8798 (0.003)	0.7999 (0.001)	0.9793 (0.001)	0.9792 (0.0005)	0.8004 (0.001)	
15	0	0.9258 (0.001)	0.9251 (0.002)	0.8752 (0.001)	0.9951 (0×10^{-3})	0.9951 (0×10^{-3})	0.8750 (0.001)	$\frac{7}{8} = 0.875$
	1	0.9262 (0.001)	0.9245 (0.002)	0.8752 (0.001)	0.9951 (0×10^{-3})	0.9952 (0×10^{-3})	0.8747 (0.001)	
200	0	0.9933 (0×10^{-3})	0.9933 (0×10^{-3})	0.9908 (0×10^{-3})	0.9999 (0×10^{-3})	0.9999 (0×10^{-3})	0.9903 (0×10^{-3})	$\frac{1999}{2001} \approx 0.999$
	1	0.9933 (0×10^{-3})	0.9932 (0×10^{-3})	0.9907 (0×10^{-3})	0.9999 (0×10^{-3})	0.9999 (0×10^{-3})	0.9903 (0×10^{-3})	

Tabla 4.2: Valor medio (desviación estándar) de las probabilidades $\Pr[x_{(1)} < x < x_{(n)} | \mathbf{x}] = \int_{x_1}^{x_2} p(x|\mathbf{x}) dx$ correspondiente a $m = 15000$ muestras simuladas de tamaño n provenientes de $\text{LgN}(x|\mu, \sigma)$ para diferentes valores de μ y σ . La columnas π_{ϕ_1} , π_{ϕ_2} y π_{ϕ_3} muestran la cobertura promedio de la densidad predictiva cuando la cantidad de interés es la media, la moda y la mediana, respectivamente.

Observaciones finales sobre el modelo Log-Normal

Jeffreys (1961, pp. 172) mencionó que en modelos de localización y escala (ambos parámetros desconocidos), y la función inicial objetiva dada por

$$\pi(\mu, \sigma) = \sigma^{-1},$$

la densidad predictiva resultante cumple con la propiedad

$$\Pr(x_1 < x < x_2 | x_1, x_2) = 1/3.$$

Los resultados en la Tabla 4.2 demuestran que se extiende a

$$\Pr[x_{(1)} < x < x_{(n)}] = \frac{n-1}{n+1}$$

cuando se consideran muestras de tamaño n .

En el caso del modelo Log-Normal, se ha visto que cuando la media o la moda son la cantidad de interés, la densidad final correspondiente no existe si solo se tienen dos observaciones aleatorias. Sin embargo, cuando la cantidad de interés es la mediana no existe esta restricción.

Por tanto, la mediana (pero no la media o la moda) puede considerarse como una elección general apropiada de la cantidad de interés para problemas de predicción.

4.5. Modelos continuos no regulares

Dentro de esta sección se estudiarán los modelos no regulares de la forma $\mathcal{M} \equiv \{p(x|\theta) : x \in S(\theta), \theta \in \Theta \in \mathcal{R}\}$ donde $S(\theta)$ es el espacio muestral que depende del parámetro del modelo.

Dentro del *análisis de referencia*, los trabajos de Ghosal (1997) y Ghosal y Samanta (1997) definen las condiciones de regularidad bajo las cuales la función inicial que maximiza la información esperada del muestreo repetido de modelos de la forma \mathcal{M} , y esa es la función inicial de referencia.

Sin embargo, como se verá más adelante, el método propuesto por Ghosal (1997) no es muy claro cuando se desea obtener la inicial de referencia de una determinada cantidad de interés. En cambio, el método sugerido por Bernardo (1979), Berger y Bernardo (1992a), Bernardo (2005a) y Berger *et al.* (2009a) es claro y arroja una inicial cuya densidad predictiva cuenta con adecuadas propiedades de cobertura.

La presentación de los modelos de esta sección será de acuerdo a su grado de complejidad. Empezamos con dos modelos Uniformes y concluimos con dos casos del modelo Exponencial trasladado o *shifted Exponential*.

4.5.1. Modelo uniforme en $(0, \theta)$

Considérese el Ejemplo 11 dado en Bernardo (2005a) correspondiente al modelo $\text{Un}(0, \theta)$.

Sean dos observaciones aleatorias x_1 y x_2 provenientes de la distribución $\text{Un}(0, \theta)$ con $\theta > 0$. Este modelo también cumple con la propiedad frecuentista de que dadas dos observaciones con $x_1 < x_2$, la probabilidad de que una nueva observación pertenezca al intervalo (x_1, x_2) es un tercio, es decir, $\Pr(x_1 < x < x_2 | \theta) = \frac{1}{3}$.

Teorema 11 *Sean dos observaciones aleatorias e independientes x_1 y x_2 con densidad $\text{Un}(0, \theta)$ y $x_1 < x_2$, la probabilidad de que una tercera observación x se encuentre contenida en (x_1, x_2) es un tercio.*

Demostración. La probabilidad requerida es

$$\Pr(x_1 < x < x_2 | \theta) = \int_{x_1}^{x_2} \text{Un}(0, \theta) dx = \frac{x_2 - x_1}{\theta}$$

que es una función que depende de las observaciones, entonces debemos encontrar el valor esperado de ésta sobre todos los posibles valores que una muestra de tamaño dos pudiera tener. Bajo el supuesto de $x_1 < x_2$, la función de densidad conjunta de (x_1, x_2) es

$$f(x_1, x_2 | \theta) = \begin{cases} \frac{2}{\theta^2} & \text{si } 0 < x_1 < x < x_2 < \theta, \\ 0 & \text{en otro caso.} \end{cases}$$

Entonces el valor esperado será

$$\begin{aligned} E_{x_1 x_2} \left(\frac{x_2 - x_1}{\theta} \right) &= \frac{2}{\theta^3} \int_0^\theta \int_0^{x_2} (x_2 - x_1) dx_1 dx_2 \\ &= \frac{2}{\theta^3} \int_0^\theta (x_1 x_2 - \frac{x_1^2}{2}) \Big|_0^{x_2} dx_2 \\ &= \frac{1}{\theta^3} \int_0^\theta x_2^2 dx_2 = \frac{1}{3}. \end{aligned}$$

□

El resultado del Teorema 11 anterior puede comprobarse numéricamente y la Tabla 4.3 muestra las frecuencias de $x \in (x_1, x_2)$ para diferentes valores de θ . Las frecuencias se aproximan al valor teórico de $\frac{1}{3}$ como se esperaría.

Con la inicial de referencia $\pi(\theta) \propto \theta^{-1}$ (Bernardo, 2005a, Ejemplo 11), dadas dos observaciones aleatorias $\{x_1, x_2\}$, la densidad final es igual a

$$\pi(\theta | x_1, x_2) = c^{-1} \frac{1}{\theta^3}, \quad \theta > x_2$$

Frecuencias

θ	1	2	4	6	8	10
	0.3344 (0.0035)	0.3337 (0.0057)	0.33223 (0.0055)	0.3348 (0.0028)	0.335 (0.0051)	0.3333 (0.0043)

Tabla 4.3: Frecuencia con que una nueva observación pertenece al intervalo formado por dos observaciones dadas (x_1, x_2) ($x_1 < x_2$) provenientes de una distribución $\text{Un}(0, \theta)$. Valor promedio (desviación estándar) comprendida de $m = 10$ frecuencias de $n = 10000$ muestras aleatorias de tamaño dos.

con $c = (2x_2^2)^{-1}$. La densidad predictiva resulta ser

$$p(x|x_1, x_2) = c^{-1} \int_{\max\{x, x_2\}}^{\infty} \frac{1}{\theta^4} d\theta = \frac{2}{3} \begin{cases} \frac{1}{x_2} & x < x_2 \\ \frac{x_2^2}{x^3} & x > x_2 \end{cases}.$$

Puede entonces calcularse la probabilidad requerida,

$$\begin{aligned} \Pr(x_1 < x < x_2 | x_1, x_2) &= \int_{x_1}^{x_2} p(x|x_1, x_2) dx \\ &= \int_{x_1}^{x_2} \frac{2}{3x_2} dx = \frac{2(x_2 - x_1)}{3x_2} = g(x_1, x_2). \end{aligned}$$

que depende de las observaciones obtenidas. Se desea saber el valor esperado en el muestreo de dicha función

$$\begin{aligned} E_{x_1, x_2 | \theta}(g(x_1, x_2)) &= 2 \int_0^{\theta} \int_0^{x_2} \frac{2(x_2 - x_1)}{3x_2} \frac{1}{\theta^2} dx_1 dx_2 \\ &= \frac{4}{3\theta^2} \int_0^{\theta} \frac{x_2}{2} dx_2 = \frac{1}{3} \end{aligned}$$

que es exacto e igual a $\frac{1}{3}$.

La comprobación numérica de lo anterior puede verse en la Tabla 4.4 que muestra el valor promedio (desviación estándar) de la cobertura de probabilidad de la densidad predictiva correspondiente de $n = 10000$ muestras aleatorias de tamaño dos y diferentes valores de θ . Nótese que el valor medio se aproxima a $\frac{1}{3}$ independientemente del valor de θ .

θ	1/2	1	2	4	6	8	10
	0.3271 (0.0366)	0.3203 (0.0379)	0.3282 (0.0381)	0.3408 (0.0370)	0.3339 (0.0370)	0.3340 (0.0371)	0.3327 (0.0372)

Tabla 4.4: Valor medio y desviación estándar de la cobertura de probabilidad de la densidad predictiva de $\text{Un}(0, \theta)$ con previa de referencia dada por $\pi(\theta) \propto \theta^{-1}$, los valores se aproximan $1/3$.

Los valores en la Tabla 4.4 se calcularon así:

- Generar dos observaciones aleatorias (x_1, x_2) de $\text{Un}(0, \theta)$.
- Obtener las densidades final $\pi(\theta|x_1, x_2)$ y predictiva $p(x|x_1, x_2)$.
- Calcular $\Pr(x_1 < x < x_2|x_1, x_2)$ y almacenarlo.
- Repetir todos los pasos anteriores n veces, promediar el resultado y calcular la desviación estándar.

La conclusión final de $\text{Un}(0, \theta)$ es que la inicial de referencia $\pi(\theta) \propto \theta^{-1}$ produce una densidad predictiva con adecuadas propiedades de cobertura.

4.5.2. Modelo uniforme en $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$

Continuando con los modelos no regulares, otro modelo interesante es el Uniforme sobre el intervalo $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, para el cual su inicial de referencia es uniforme en $(x_{\max} - \frac{1}{2}, x_{\min} + \frac{1}{2})$ (Bernardo, 1979, Sec. 3.5), con x_{\max} y x_{\min} las observaciones mínima y máxima de la muestra, respectivamente.

Este modelo, también cumple con la propiedad frecuentista de que dadas dos observaciones aleatorias $x_1 < x_2$ la probabilidad del evento $x_1 < x < x_2$ es igual a un tercio, es decir, $\Pr(x_1 < x < x_2|\theta) = \frac{1}{3}$.

Teorema 12 Sean x_1 y x_2 dos observaciones aleatorias e independientes con densidad $\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ con $x_1 < x_2$, la probabilidad de que una tercera observación x pertenezca a (x_1, x_2) es un tercio.

Demostración. La probabilidad requerida es

$$\Pr(x_1 < x < x_2|\theta) = \int_{x_1}^{x_2} dx = x_2 - x_1,$$

que es una función de los datos entonces debemos encontrar el valor esperado de ésta sobre todos los posibles valores que pudieran tener (x_1, x_2) . Puesto que $x_1 < x_2$, la función de densidad conjunta es

$$f(x_1, x_2|\theta) = \begin{cases} 2 & \text{si } \theta - \frac{1}{2} < x_1 < x < x_2 < \theta + \frac{1}{2}, \\ 0 & \text{en otro caso.} \end{cases}$$

Omitiendo por brevedad algunos resultados intermedios, el valor esperado resulta

$$\begin{aligned}
 E_{x_1 x_2 | \theta} (x_2 - x_1) &= 2 \int_{\theta-1/2}^{\theta+1/2} \int_{\theta-1/2}^{x_2} (x_2 - x_1) dx_1 dx_2 \\
 &= 2 \int_{\theta-1/2}^{\theta+1/2} \left(\frac{1}{8} + \frac{x_2 + x_2^2 - \theta + \theta^2}{2} - x_2 \theta \right) dx_2 \\
 &= 2 \left(\frac{1}{6} \right) = \frac{1}{3}.
 \end{aligned}$$

□

La Tabla 4.5 muestra las frecuencias de este evento y puede verse que los resultados son congruentes con el Teorema 12 (la obtención de las frecuencias fueron similares a los de la Tabla 4.3).

Frecuencias						
θ	1	2	4	6	8	10
	0.3335	0.3335	0.3333	0.3330	0.3351	0.3338
	(0.0049)	(0.0035)	(0.0041)	(0.0041)	(0.0032)	(0.0034)

Tabla 4.5: Frecuencia con que una tercera observación se encuentra contenida en el intervalo formado por dos observaciones previas. Valor medio (desviación estándar) de $m = 10$ frecuencias con que ocurre el suceso en $n = 10000$ muestras aleatorias de tamaño dos de $\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

A continuación obtengamos las densidades final y predictiva. Sean dos observaciones aleatorias x_1 y x_2 ($x_1 < x_2$) con distribución $\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. La correspondiente densidad final de referencia está dada por $\text{Un}(x_2 - \frac{1}{2}, x_1 + \frac{1}{2})$ y la densidad de predictiva será

$$p(x|x_1, x_2) = \int_{\Theta} \frac{1}{1 + x_1 - x_2} d\theta$$

donde $\Theta = \{\theta : \theta \in (x - \frac{1}{2}, x + \frac{1}{2}) \cap (x_2 - \frac{1}{2}, x_1 + \frac{1}{2})\}$. Debe notarse que el intervalo $(x - \frac{1}{2}, x + \frac{1}{2})$ corresponde a la restricción del modelo $\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ en términos de θ .

La intersección de los intervalos se reduce al siguiente espacio paramétrico,

$$\Theta = \left\{ \theta : \theta \in \left(\max\{x, x_2\} - \frac{1}{2}, \min\{x, x_1\} + \frac{1}{2} \right) \right\}.$$

Por tanto, en la densidad predictiva deben considerarse todos los casos posibles,

$$p(x|x_1, x_2) = \frac{\min\{x, x_1\} - \max\{x, x_2\} + 1}{1 + x_1 - x_2}$$

$$= \frac{1}{1 + x_1 - x_2} \times \begin{cases} x - x_2 + 1 & \text{si } x_2 - 1 < x < x_1, \\ x_1 - x + 1 & \text{si } x_2 < x < x_1 + 1, \\ x_1 - x_2 + 1 & \text{si } x_1 < x < x_2, \\ 0 & \text{en otro caso.} \end{cases}$$

Finalmente, la probabilidad requerida es

$$\Pr(x_1 < x < x_2|x_1, x_2) = \int_{x_1}^{x_2} p(x|x_1, x_2) = \int_{x_1}^{x_2} dx = x_2 - x_1 = g(x_1, x_2)$$

que es una función de los datos. Por lo tanto, determinaremos el valor esperado de $g(x_1, x_2)$ evaluado sobre todos los posibles valores que una muestra de tamaño dos pudiera tener. Bajo el supuesto de $x_1 < x_2$, se tiene que

$$E_{x_1, x_2|\theta}(g(x_1, x_2)) = 2 \int_{\theta - \frac{1}{2}}^{\theta + \frac{1}{2}} \int_{\theta - \frac{1}{2}}^{x_2} (x_2 - x_1) dx_1 dx_2 = \frac{1}{3}$$

el cual es exacto coincidiendo con el resultado del Teorema 9. El caso $x_1 > x_2$ produce el mismo resultado.

La Tabla 4.6 muestra la media muestral de la cobertura de probabilidad de la densidad predictiva en el intervalo formado por dos observaciones dadas, proveniente de n muestras simuladas de tamaño dos. Nótese que los valores se aproximan al valor teórico de $\frac{1}{3}$. Los valores de la Tabla 4.6 se calcularon de manera similar a los de la Tabla 4.4.

θ	1/2	1	2	4	6	8	10
	0.3302	0.3291	0.3356	0.3340	0.3325	0.3338	0.3344
	(0.0551)	(0.0551)	(0.0554)	(0.0545)	(0.0557)	(0.0563)	(0.0558)

Tabla 4.6: Valor medio y desviación estándar de la cobertura de probabilidad de la densidad predictiva del modelo $\text{Un}(0, \theta)$ y la inicial de referencia $\pi(\theta) \propto \theta^{-1}$ provenientes de $n = 10000$ muestras aleatorias de tamaño dos.

Por lo tanto, la distribución uniforme como inicial de referencia para $\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, produce una densidad predictiva con adecuadas propiedades de cobertura.

4.5.3. Modelo Exponencial con dos parámetros y parámetro de escala conocido

Consideremos ahora el modelo no regular Exponencial trasladado $\text{Ex}(\cdot|\alpha, \varphi)$, también conocido como Exponencial trasladado, *shifted Exponential* o *Exponencial con dos parámetros*. La función de densidad está dada por

$$\text{Ex}(x|\alpha, \varphi) = \frac{1}{\varphi} e^{-\frac{x-\alpha}{\varphi}}, \quad x > \alpha, \alpha \in \mathcal{R}, \varphi > 0.$$

Obsérvese que es un modelo no regular pues su espacio muestral depende del parámetro α .

En esta subsección nos concentraremos en el supuesto de que φ es conocido e igual a uno. Por tanto, el modelo se reduce a uno de un solo parámetro con función de densidad dada por

$$\text{Ex}(x|\alpha) = e^{-(x-\alpha)}, \quad x > \alpha, \alpha \in \mathcal{R},$$

y función de distribución acumulada dada por

$$F(x|\alpha) = 1 - e^{-(x-\alpha)}.$$

Nótese que sigue siendo un modelo no regular.

Las medidas de tendencia central son las siguientes: la media es el valor $\alpha + 1$, la mediana es igual a $\alpha + \log 2$ y la moda es el valor más pequeño de α en su rango de valores. Hemos simplificado la notación de $\text{Ex}(\cdot|\alpha, \varphi)$ por $\text{Ex}(\cdot|\alpha)$ y sólo en este apartado hemos abusado de la notación de $\text{Ex}(\cdot|\cdot)$ dada en la pág. 4.

El modelo $\text{Ex}(\cdot|\alpha)$ cumple igualmente con la propiedad frecuentista de que una nueva observación se encuentre contenida en el intervalo formado por dos observaciones previas.

Teorema 13 Sean x_1 y x_2 dos observaciones aleatorias e independientes con función de densidad $\text{Ex}(\cdot|\alpha)$, con $x_1 < x_2$. La probabilidad de que una nueva observación x se encuentre en (x_1, x_2) es un tercio.

Demostración. La probabilidad requerida es

$$\Pr(x_1 < x < x_2|\alpha) = \int_{x_1}^{x_2} e^{-(x-\alpha)} dx = e^{-(x_1-\alpha)} - e^{-(x_2-\alpha)}.$$

Puede verse que es una función que depende de las observaciones, entonces evaluaremos ésta sobre todos los posibles valores que pudieran tener las observaciones (x_1, x_2) . La función de densidad conjunta de (x_1, x_2) con $x_1 < x_2$ es

$$f(x_1, x_2|\alpha) = 2e^{-(x_1+x_2-2\alpha)}, \quad x_2 > x_1 > \alpha.$$

Entonces, omitiendo por brevedad algunos cálculos intermedios, se tiene que

$$\begin{aligned}
& E_{x_1 x_2} (e^{-(x_1-\alpha)} - e^{-(x_2-\alpha)}) \\
&= 2 \int_0^\infty \int_\alpha^{x_2} (e^{-(x_1-\alpha)} - e^{-(x_2-\alpha)}) e^{-(x_1+x_2-2\alpha)} dx_1 dx_2 \\
&= \int_0^\infty e^{-(3x_2-\alpha)} (e^{x_2} - e^\alpha)^2 dx_2 \\
&= \frac{1}{3}.
\end{aligned}$$

□

También puede comprobarse en la Tabla 4.7 que las frecuencias provenientes de $n = 10000$ muestras simuladas de tamaño dos de $\text{Ex}(\cdot|\alpha)$ se aproximan al valor teórico del teorema anterior, independientemente del valor que α pudiera tener.

Frecuencias							
α	0	2	4	6	8	10	16
	0.3315	0.3323	0.3324	0.3325	0.3296	0.33322	0.3308
	(0.0042)	(0.0061)	(0.0045)	(0.0034)	(0.0058)	(0.0041)	(0.0048)

Tabla 4.7: Frecuencias del evento $x \in (x_1, x_2)$ correspondientes de $n = 10000$ muestras simuladas de tamaño dos provenientes de $\text{Ex}(\cdot|\alpha)$. Media muestral y desviación estándar de $m = 10$ frecuencias.

Función inicial de referencia

El modelo $\text{Ex}(\cdot|\alpha)$ es no regular y se esperaría poder aplicar el método de obtención de iniciales de referencia sugerido por Ghosal y Samanta (1997) para esta clase de modelos, pero $\text{Ex}(\cdot|\alpha)$ no cumple con las condiciones de regularidad citadas por los autores.

Sin embargo, aplicando el algoritmo explícito de obtención de iniciales de referencia dada por Bernardo (1979), Berger y Bernardo (1992a), Bernardo (2005a) y Berger *et al.* (2009a) (Teorema 2 en esta Tesis) resulta en la función inicial uniforme como se detalla a continuación.

Sean n observaciones aleatorias $\mathbf{x} = \{x_1, \dots, x_n\}$, todas con función de distribución de probabilidad dada por $F(x|\alpha)$. La función de verosimilitud estará dada por

$$\begin{aligned}
l(\alpha|\mathbf{x}) &= \prod_{i=1}^n \text{Ex}(x_i|\alpha) = \exp\{-(t - n\alpha)\}, \quad x_i < \alpha, \quad i = 1, \dots, n \\
&= \exp\{-(t - n\alpha)\} \mathbf{1}_{\{\alpha < x_{(1)}\}}(\alpha) \\
&\propto e^{n\alpha} \mathbf{1}_{\{\alpha < x_{(1)}\}}(\alpha)
\end{aligned}$$

con $t = \sum_{i=1}^n x_i$, $x_{(1)} = \min\{x_1, \dots, x_n\}$ y $\mathbf{1}$ es la función indicadora tal que

$$\mathbf{1}_{\{\alpha < x_{(1)}\}}(\alpha) = \begin{cases} 1 & \text{si } \alpha < x_{(1)}, \\ 0 & \text{en otro caso.} \end{cases}$$

Nótese que $x_{(1)}$ es el estadístico suficiente y tiene función de densidad dada por

$$g(x_{(1)}|\alpha) = ne^{-n(x_{(1)}-\alpha)}, \quad \alpha < x_{(1)} < \infty.$$

Sea $h(\alpha) = 1$, para la cual se cumple que

$$\begin{aligned} \int_{-\infty}^{x_{(1)}} g(x_{(1)}|\alpha)h(\alpha)d\alpha &= ne^{-nx_{(1)}} \int_{-\infty}^{x_{(1)}} e^{nx_{(1)}} d\alpha \\ &= 1 < \infty. \end{aligned}$$

y

$$\begin{aligned} f_n(\alpha) &= \exp \left\{ \int_{\alpha}^{\infty} g(x_{(1)}|\alpha) \log g(x_{(1)}|\alpha) dx_{(1)} \right\} \\ &= \exp \left\{ \int_{\alpha}^{\infty} ne^{-n(x_{(1)}-\alpha)} (\log n - n(x_{(1)} - \alpha)) dx_{(1)} \right\} \\ &= n \exp \left\{ -n \left(\int_{\alpha}^{\infty} x_{(1)} ne^{-n(x_{(1)}-\alpha)} dx_{(1)} - \alpha \right) \right\} \\ &= n \exp \left\{ -n \left(\frac{1}{n} + \alpha - \alpha \right) \right\} \\ &= ne^{-1} \end{aligned}$$

que no depende de α . Por lo tanto, la función inicial de referencia es

$$\pi(\alpha) \propto \lim_{n \rightarrow \infty} \frac{f_n(\alpha)}{f_n(\alpha_0)} = \frac{ne^{-1}}{ne^{-1}} = 1.$$

Como verificación adicional, hemos calculado la aproximación numérica de $\pi(\alpha)$ que puede verse en la Figura 4.2 y que es efectivamente aproximadamente constante.

La Figura 4.2 muestra la inicial de referencia numérica $\pi(\alpha)$ aproximada numéricamente para veinte valores de α en un rango de $-e^3$ a e^3 uniformemente espaciados y re-escalados de tal forma que $\pi(1) = 1$. Para la aproximación fueron simuladas $m = 10000$ muestras aleatorias de tamaño $k = 500$ de $\text{Ex}(\cdot|\alpha, 1)$.

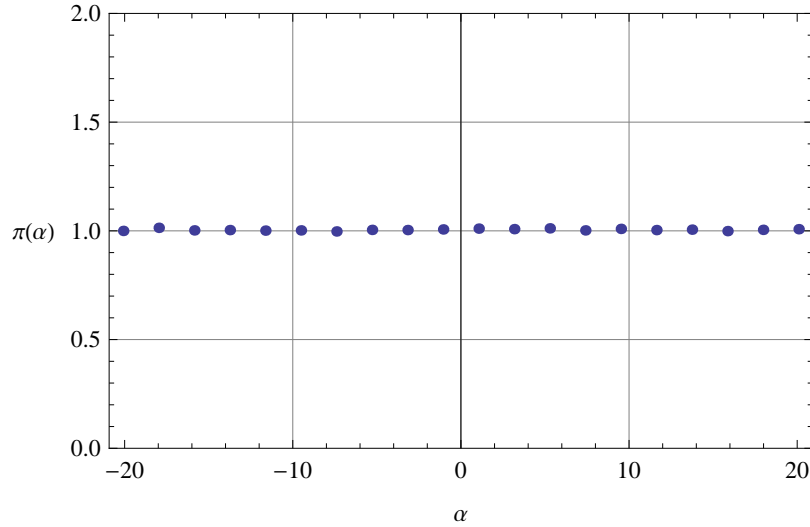


Figura 4.2: *Función inicial de referencia numérica de $\pi(\alpha)$ para veinte valores de α uniformemente espaciados en $(-e^3, e^3)$ y re-escalados para que $\pi(1) = 1$. Simulación de $m = 10000$ muestras de tamaño $k = 500$ provenientes de $\text{Ex}(\cdot|\alpha)$.*

Densidad predictiva y su cobertura de probabilidad

Obtengamos ahora la densidad final con $\pi(\alpha) \propto 1$. Dadas dos observaciones x_1 y x_2 ($x_1 < x_2$), se tiene que

$$\pi(\alpha|x_1, x_2) = c^{-1}e^{-(x_1+x_2-2\alpha)}, \quad -\infty < \alpha < x_1,$$

donde la constante de marginalización es

$$c = \int_{-\infty}^{x_1} \exp\{-(x_1 + x_2 - 2\alpha)\} = \frac{1}{2}e^{x_1-x_2}.$$

La densidad predictiva puede obtenerse fácilmente. Nótese que para una nueva observación x , el rango de valores de α estará acotado por el valor mínimo entre las variables x_1 y x , es decir $\alpha \in (-\infty, \min\{x, x_1\})$,

$$\begin{aligned} p(x|x_1, x_2) &= \frac{1}{c} \int_{-\infty}^{\min\{x, x_1\}} e^{-(x_1+x_2+x-3\alpha)} d\alpha = \frac{1}{c} e^{-(x_1+x_2+x)} \int_{-\infty}^{\min\{x, x_1\}} e^{3\alpha} d\alpha \\ &= \frac{1}{3c} e^{-(x_1+x_2+x)} e^{3\alpha} \Big|_{-\infty}^{\min\{x, x_1\}} \\ &= \frac{1}{3c} e^{-(x_1+x_2+x)} \begin{cases} e^{3x} & x < x_1 \\ e^{3x_1} & x > x_1 \end{cases} \\ &= \frac{2}{3} \begin{cases} e^{-2(x_1-x)} & x < x_1 \\ e^{x_1-x} & x > x_1 \end{cases}. \end{aligned}$$

Puede comprobarse que $p(x|x_1, x_2)$ integra a uno en todo su espacio muestral:

$$\frac{2}{3} \left(\int_{-\infty}^{x_1} e^{-2(x_1-x)} dx + \int_{x_1}^{\infty} e^{-(x_1-x)} dx \right) = \frac{2}{3} \left(\frac{e^{-2(x_1-x)}}{2} \Big|_{-\infty}^{x_1} + e^{-(x_1-x)} \Big|_{x_1}^{\infty} \right) = 1.$$

La probabilidad requerida es

$$\begin{aligned} \Pr(x_1 < x < x_2 | x_1, x_2) &= \int_{x_1}^{x_2} p(x|x_1, x_2) dx = \frac{2}{3} e^{x_1} \int_{x_1}^{x_2} e^{-x} dx \\ &= \frac{2}{3} e^{x_1} (e^{-x_1} - e^{-x_2}) = \frac{2}{3} (1 - e^{x_1-x_2}) = h(x_1, x_2) \end{aligned}$$

que es una función de las observaciones.

Como se ha procedido anteriormente, calculemos el valor esperado de $h(x_1, x_2)$ sobre todos los posibles valores que una muestra de tamaño dos pudiera tener. Recordemos el supuesto inicial $x_1 < x_2$ para el cual se tiene que

$$\begin{aligned} E_{x_1, x_2 | \alpha}(h(x_1, x_2)) &= 2 \int_{\alpha}^{\infty} \int_{\alpha}^{x_2} h(x_1, x_2) \text{Ex}(x_1 | \alpha) \text{Ex}(x_2 | \alpha) dx_1 dx_2 \\ &= \frac{4}{3} \int_{\alpha}^{\infty} \int_{\alpha}^{x_2} (1 - e^{x_1-x_2}) e^{-(x_1+x_2-2\alpha)} dx_1 dx_2 \\ &= \frac{4e^{2\alpha}}{3} \int_{\alpha}^{\infty} e^{-x_2} \int_{\alpha}^{x_2} e^{-x_1} (1 - e^{x_1-x_2}) dx_1 dx_2 \\ &= \frac{4e^{2\alpha}}{3} \int_{\alpha}^{\infty} (e^{-(x_2+\alpha)} - (x_2 - \alpha + 1) e^{-2x_2}) dx_2 \\ &= \frac{4e^{2\alpha}}{3} \left(e^{-2\alpha} - \frac{3}{4} e^{-2\alpha} \right) = \frac{1}{3} \end{aligned}$$

que es exacto e igual a $1/3$. Este resultado confirma la concordancia con el Teorema 13. El caso $x_1 > x_2$ es similar y por tanto se omite.

La Tabla 4.8 muestra el valor promedio y desviación estándar de la cobertura de probabilidad de $p(x|x_1, x_2)$ en el intervalo (x_1, x_2) correspondiente de n muestras aleatorias de tamaño dos. Las aproximaciones numéricas confirman el resultado teórico.

α	-2	1/2	0	4	16
	0.3358 (0.0372)	0.3320 (0.03670)	0.3343 (0.0374)	0.3356 (0.0377)	0.3351 (0.03750)

Tabla 4.8: Media muestral (desviación estándar) de la cobertura de probabilidad de la densidad predictiva resultante de la inicial de referencia $\pi(\alpha) \propto 1$ y el modelo no regular $\text{Ex}(x|\alpha, 1)$. Simulación de $n = 10000$ muestras aleatorias de tamaños dos con varios valores de α .

Observaciones finales sobre el modelo no regular $\text{Ex}(x|\alpha)$

Puesto que es un modelo no regular con un solo parámetro, la función inicial de referencia para α es esencialmente única. En particular la mediana y la media son dos transformaciones uno-a-uno de α , por tanto sus funciones iniciales de referencia serán iguales a la función inicial de referencia de α .

Dado que $\pi(\alpha) = 1$ arroja una densidad predictiva con adecuadas propiedades de cobertura entonces puede deducirse que $\pi(\alpha) = 1$ es la inicial adecuada para predicción en este modelo.

4.6. Modelo Exponencial con dos parámetros

Estudiaremos a continuación el modelo Exponencial con dos parámetros $\text{Ex}(x|\alpha, \varphi)$, ambos desconocidos. La función de densidad está dada por

$$\text{Ex}(x|\alpha, \varphi) = \varphi^{-1} e^{-\varphi^{-1}(x-\alpha)} \quad x > \alpha, \varphi > 0.$$

y la función de distribución acumulada es $F(x|\alpha, \varphi) = 1 - e^{-(x-\alpha)/\varphi}$. La media y varianzas son, respectivamente,

$$\text{Me}(x|\alpha, \varphi) = \alpha + \varphi \quad \text{y} \quad \text{Var}(x|\alpha, \varphi) = \varphi^2.$$

La mediana es

$$\text{Med}(x|\alpha, \varphi) = \alpha + \varphi \log 2$$

y la moda se encuentra en el valor α más pequeño de su rango de variación.

Como se ha procedido anteriormente, con el modelo $\text{Ex}(\cdot|\alpha, \varphi)$ es fácil determinar la frecuencia con la que una nueva observación pertenece al intervalo formado por dos observaciones previas y la Tabla 4.9 muestra que dichas frecuencias se aproximan al valor de $\frac{1}{3}$ cuando se cuenta con dos observaciones, y en caso de tener una muestra aleatoria $\{x_1, \dots, x_n\}$, la probabilidad de una nueva observación se encuentre contenida en $(x_{(1)}, x_{(n)})$ es $\frac{n-1}{n+1}$, donde $x_{(1)} = \min\{x_1, \dots, x_n\}$ y $x_{(n)} = \max\{x_1, \dots, x_n\}$. Véase el siguiente teorema.

Teorema 14 *Sea la muestra aleatoria $\{x_1, \dots, x_n\}$ con densidad de probabilidad dada por $\text{Ex}(\cdot|\alpha, \varphi)$ y sean las observaciones mínima y máxima de la muestra $x_{(1)}$ y $x_{(n)}$, respectivamente. La probabilidad de que una nueva observación x se encuentre contenida en $(x_{(1)}, x_{(n)})$ es $\frac{n-1}{n+1}$.*

Demostración. La probabilidad del suceso $x_{(1)} < x < x_{(n)}$ es

$$\begin{aligned}\Pr(x_{(1)} < x < x_{(n)}|\alpha, \varphi) &= \int_{x_{(1)}}^{x_{(n)}} \text{Ex}(x|\alpha, \varphi)dx \\ &= e^{-\frac{x_{(1)}-\alpha}{\varphi}} - e^{-\frac{x_{(n)}-\alpha}{\varphi}}.\end{aligned}$$

Como puede verse, es una función que depende de $x_{(1)}$ y $x_{(n)}$ y de los parámetros, entonces evaluemos el valor esperado de ésta sobre todos los posibles valores que $(x_{(1)}, x_{(n)})$ pudiesen tener. La función de densidad conjunta de $(x_{(1)}, x_{(n)})$ es

$$\begin{aligned}f(x_{(1)}, x_{(n)}|\alpha, \varphi) &= \frac{n(n-1)}{\varphi^2} \left(e^{-\frac{x_{(1)}-\alpha}{\varphi}} - e^{-\frac{x_{(n)}-\alpha}{\varphi}} \right)^{n-2} e^{-\frac{x_{(1)}+x_{(n)}-2\alpha}{\varphi}} \\ &\alpha < x_{(1)} < x_{(n)}, \alpha \in \mathcal{R}, \varphi > 0.\end{aligned}$$

El valor esperado será igual a (se han omitido algunos cálculos)

$$\begin{aligned}\mathbb{E}_{x_{(1)}x_{(n)}} \left(e^{-\frac{x_{(1)}-\alpha}{\varphi}} - e^{-\frac{x_{(n)}-\alpha}{\varphi}} \right) &= \frac{n(n-1)}{\varphi^2} \int_{\alpha}^{\infty} \int_{\alpha}^{x_{(n)}} \left(e^{-\frac{x_{(1)}-\alpha}{\varphi}} - e^{-\frac{x_{(n)}-\alpha}{\varphi}} \right)^{n-1} e^{-\frac{x_{(1)}+x_{(n)}-2\alpha}{\varphi}} dx_{(1)} dx_{(n)} \\ &= \frac{n-1}{\varphi} \int_{\alpha}^{\infty} \left(e^{-\alpha/\varphi} - e^{-x_{(n)}/\varphi} \right)^n e^{-\frac{x_{(n)}-(n+1)\alpha}{\varphi}} dx_{(n)} \\ &= \frac{n-1}{n+1}\end{aligned}$$

□

Nótese que al ser un modelo no regular, la inicial de referencia no puede obtenerse a partir de la matriz de información de Fisher. Sin embargo, pertenece a los modelos no regulares analizados por Ghosal (1997) quien obtuvo una función inicial de referencia en este tipo de modelos.

Sigamos el método propuesto por Ghosal (1997) descrito en la Subsección 2.3.4 de esta Tesis. Puesto que obtener las iniciales de referencia cuando α o φ es la cantidad de interés es relativamente fácil se omite su desarrollo. Sin embargo, desarrollaremos la obtención de la inicial de la mediana cuando ésta es la cantidad de interés.

Entonces,

- Si α fuese la cantidad de interés y φ el parámetro de ruido la inicial de referencia es $\pi_{G\alpha}(\alpha, \varphi) \propto \varphi^{-1}$.
- Inversamente, si la cantidad de interés fuese φ , la inicial de referencia es $\pi_{G\varphi}(\alpha, \varphi) \propto \varphi^{-2}$

Frecuencia del evento $x_{(1)} < x < x_{(n)}$					
n	$\alpha \backslash \varphi$	1	1/2	1/5	$\frac{n-1}{n+1}$
2	1/2	0.3356	0.3339	0.3326	1/3 = 0.3333
	1	0.3334	0.3344	0.3314	
	4	0.3320	0.3337	0.3335	
3	1/2	0.5016	0.5005	0.5011	1/2 = 0.5
	1	0.4992	0.5001	0.4989	
	4	0.50136	0.4979	0.5013	
4	1/2	0.6	0.5993	0.6048	3/5 = 0.6
	1	0.5998	0.6028	0.6024	
	4	0.6001	0.6029	0.5988	
7	1/2	0.7527	0.7503	0.7488	3/4 = 0.75
	1	0.7508	0.7507	0.7480	
	4	0.7502	0.7487	0.7504	
9	1/2	0.7999	0.7975	0.8022	4/5 = 0.8
	1	0.8011	0.7998	0.8004	
	4	0.8007	0.8001	0.7992	
15	1/2	0.8756	0.8743	0.8743	7/8 = 0.875
	1	0.8756	0.8745	0.8745	
	4	0.8758	0.8766	0.8757	

Tabla 4.9: Frecuencia de que una nueva observación x con distribución $\text{Ex}(x|\alpha, \beta)$ pertenece al intervalo $(x_{(1)}, x_{(n)})$ formado por las observaciones máxima y mínima de la muestra $\mathbf{x} = \{x_1, \dots, x_n\}$. Frecuencias correspondientes de $m = 100000$ muestras aleatorias de tamaño n .

la cual coincide con $\pi_G(\alpha, \varphi)$.

Puesto que estamos interesados en la mediana como cantidad de interés, definámosla como

$$\phi \equiv \phi(\alpha, \varphi) = \alpha + t\varphi$$

con $t = \log 2$. Puede verse que es una transformación lineal de α y de φ , y si $t = 1$ estaríamos trabajando con la media como cantidad de interés.

El modelo expresado en términos de (α, ϕ) es

$$\text{Ex}(x|\alpha, \phi) = \frac{t}{\phi - \alpha} e^{-\frac{t}{\phi - \alpha}(x - \alpha)} \quad x > \alpha, \alpha > 0, \phi > \alpha.$$

Con la notación de Ghosal (1997) se tiene que:

$$S(\alpha) := [a_1(\alpha), a_2(\alpha)] = (\alpha, \infty),$$

por lo tanto $a_1(\alpha) = \alpha$ y $a_2(\alpha) = \infty$. Los límites correspondientes a las ecuaciones (2.7) y (2.8) serán

$$p(\alpha, \phi) = \frac{t}{\phi - \alpha} \quad \text{y} \quad q(\alpha, \phi) = 0.$$

Entonces la ecuación (2.10) es

$$c(\alpha, \phi) = \frac{t}{\phi - \alpha},$$

y la ecuación (2.12) es

$$\mathbf{F}(\alpha, \phi) = 4 \int_{\alpha}^{\infty} (g_{\phi}(x|\alpha, \phi))^2 dx = (\alpha - \phi)^{-2},$$

$$\text{con } g_{\phi}(x|\alpha, \phi) = \frac{t(x-\alpha)+\alpha-\phi}{2(\phi-\alpha)^{3/2}} \exp \left\{ -\frac{t(x-\alpha)}{\phi-\alpha} \right\}.$$

De acuerdo con la Ec. (2.11), la función inicial de referencia para (α, ϕ) , cuando el parámetro de interés es el par (α, ϕ) , $\pi_G(\alpha, \phi)$ es proporcional a

$$\pi_G(\alpha, \phi) = \pi_{G\phi}(\alpha, \phi) = \pi_{G\alpha}(\alpha, \phi) = c(\alpha, \phi) \sqrt{\mathbf{F}(\alpha, \phi)} \propto \frac{1}{(\alpha - \phi)^2}.$$

Puede obtenerse fácilmente la correspondiente inicial en términos de los parámetros originales, la cual es

$$\begin{aligned} \pi_{G\phi}(\alpha, \varphi) &= \pi(\alpha, \phi(\alpha, \varphi)) |J| = \varphi^{-2} \\ \text{con } J &= \begin{pmatrix} \frac{\partial \alpha}{\partial \alpha} & \frac{\partial \alpha}{\partial \varphi} \\ \frac{\partial \phi}{\partial \alpha} & \frac{\partial \phi}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & t \end{pmatrix}. \end{aligned}$$

Por lo tanto, la inicial cuando el parámetro de interés es el vector (α, ϕ) en términos de los parámetros originales (α, φ) es

$$\pi_{G\phi}(\alpha, \varphi) \propto \varphi^{-2}.$$

En los términos que hemos trabajado, nuestro interés *debería* centrarse en obtener la inicial de referencia cuando la mediana es la cantidad de interés más. El método de Ghosal (1997) propone dos procedimientos en estos casos, de los cuales al aplicar el primer procedimiento resulta en una función inicial $\pi_{Proc1}(\alpha, \phi)$ indeterminada. Mientras que el segundo procedimiento arroja la función inicial $\pi_{Proc2}(\alpha, \phi) \propto (\phi - \alpha)^{-1}$, que en los parámetros originales corresponde a $\pi_{Proc2}(\alpha, \varphi) \propto \varphi^{-1}$. Es evidente que *no hay una certeza* en sobre cual procedimiento debe usarse.

Apliquemos ahora el método de obtención de iniciales de referencia propuesto por Bernardo (1979, 2005a), Berger and Bernardo (1992a) y Berger *et al.* (2009a). Para esto, el modelo se expresará en términos de la mediana ϕ y del parámetro φ ,

$$\text{Ex}(x|\phi, \varphi) = \frac{e^{-t}}{\varphi} \exp\left\{-\frac{x-\phi}{\varphi}\right\},$$

$$\varphi > 0, \phi - t\varphi < x.$$

donde $t = \log 2$.

Puesto que ϕ es la cantidad de interés, se obtendrá la inicial de referencia $\pi_\phi(\phi, \varphi)$ aplicando recursivamente el algoritmo de obtención de la siguiente manera:

- (i) Fijando ϕ , obtener $\pi(\varphi|\phi)$.
- (ii) Obtener el modelo integrado $p(x|\phi) = \int_{\max\{\frac{\phi-x}{t}, 0\}}^{\infty} \text{Ex}(x|\phi, \varphi)\pi(\varphi|\phi)d\varphi$.
- (iii) Aplicar nuevamente el algoritmo de obtención al modelo integrado de un solo parámetro $p(x|\phi)$ para obtener $\pi(\phi)$.

Finalmente, $\pi_\phi(\phi, \varphi) \propto \pi(\varphi|\phi)\pi(\phi)$.

Para el paso (i) hemos encontrado que la función teórica φ^{-1} ajusta a las aproximaciones numéricas de $\pi(\varphi|\phi)$ para un valor fijo ϕ lo que puede verse en la Figura 4.3. Paso (ii), suponiendo que $\pi(\varphi|\phi) \propto \varphi^{-1}$ y dado que es impropia se define una secuencia creciente de intervalos cerrados donde $\pi_i(\varphi|\phi)$ es propia. Y así obtener el modelo integrado de un solo parámetro $p_i(x|\phi)$ (véase el Apéndice B sección B.1). Finalmente, aplicar nuevamente el algoritmo numérico de obtención de iniciales al modelo $p_i(x|\phi)$ para obtener $\pi_i(\phi)$. Puede verse en la Figura 4.4 que $\pi(\phi)$ es constante para un valor i fijo.

De la aproximación numérica, podemos concluir que la inicial de referencia cuando la mediana es la cantidad de interés puede aproximarse por $1/\varphi$.

En la siguiente subsección se estudiarán las propiedades de cobertura de las densidades de probabilidad predictivas resultante de las funciones iniciales obtenidas en esta sección.

Obsérvese que, las iniciales objetivas obtenidas en la sección anterior pertenecen a la familia de funciones iniciales dadas por

$$\pi_a(\alpha, \varphi) \propto \varphi^{-a}, \quad a > 0. \quad (4.10)$$

Recuérdese que si $a = 1$ corresponde a la inicial $\pi_\phi(\phi, \varphi)$ con la mediana como cantidad de interés, obtenida de aplicar el método de Bernardo (1979, 2005a), Berger y Bernardo (1989a), Berger *et al.* (2009a) y la de Ghosal (1997, Procedimiento 2). Mientras que si

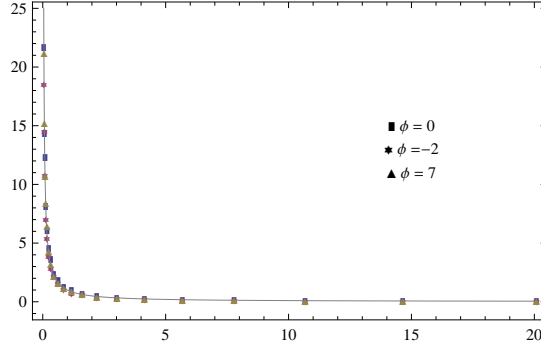


Figura 4.3: Aproximación numérica de la inicial $\pi(\varphi|\phi)$ con 20 valores de φ en un rango de e^{-3} a e^3 uniformemente log-espaciados re-escalados para que $\varphi(1|\phi) = 1$. Con tres valores de ϕ y cada valor aproximado corresponde de $n = 500$ muestras aleatorias de tamaño $k = 50$. La línea continua es la inicial teórica φ^{-1} .

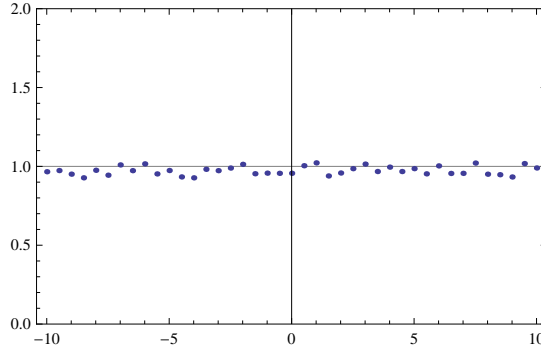


Figura 4.4: Aproximación numérica de la inicial de referencia $\pi(\phi)$ del modelo integrado $p_i(x|\phi)$. Aproximación numérica de $\pi(\phi)$ proveniente de cuarenta valores de ϕ en un rango de -10 a 10 espaciados uniformemente, cada valor fue estimado con $n = 1000$ muestras aleatorias de tamaño $k = 50$ y valor fijo $i = 3$. Los valores se centran alrededor de uno.

$a = 2$ se tiene la inicial de referencia para (α, ϕ) obtenida al aplicar el método de Ghosal (1997) cuando α o ϕ o el par (α, ϕ) son la cantidad de interés.

Sean dos observaciones aleatorias x_1 y x_2 de $\text{Ex}(x|\alpha, \varphi)$ con $x_1 < x_2$. Utilizando la función inicial π_a , la densidad final tendrá la siguiente expresión:

$$\pi_a(\alpha, \varphi|x_1, x_2) = c_a^{-1} \varphi^{-(a+2)} \exp \left\{ -\frac{x_1 + x_2 - 2\alpha}{\varphi} \right\}, \quad \alpha < x_1, \varphi > 0,$$

con constante de marginalización dada por $c_a = \frac{\Gamma(a+1)}{2a(x_2-x_1)^a}$. La correspondiente densidad

predictiva puede obtenerse muy fácilmente y será igual a

$$p(x|x_1, x_2) = \frac{2}{3}a(x_2 - x_1)^a \begin{cases} (x_1 + x_2 - 2x)^{-(a+1)} & x < x_1 \\ (x_2 + x - 2x_1)^{-(a+1)} & x > x_1. \end{cases}$$

Ahora podemos calcular la probabilidad requerida $\Pr(x_1 < x < x_2|x_1, x_2)$ que será una función que depende de a

$$\Pr(x_1 < x < x_2|x_1, x_2) = \int_{x_1}^{x_2} p(x|x_1, x_2)dx = \frac{2}{3}(1 - 2^{-a}).$$

- Si $a = 1$, que es la inicial de referencia $\pi_\phi(\alpha, \varphi)$ y la inicial $\pi_{Proc2}(\alpha, \varphi)$ entonces

$$\Pr(x_1 < x < x_2|x_1, x_2) = \frac{1}{3}.$$

- Si $a = 2$, que corresponde a la inicial $\pi_G(\alpha, \varphi)$ para la que se obtiene

$$\Pr(x_1 < x < x_2|x_1, x_2) = \frac{1}{2}.$$

Claramente, la inicial $\pi_\phi(\alpha, \varphi) \propto \frac{1}{\varphi}$ produce una densidad predictiva con adecuadas propiedades de cobertura cuando se tienen dos observaciones.

La Figura 4.5 muestra el comportamiento de la probabilidad requerida $\Pr(x_1 < x < x_2|a)$ como función de a . Obsérvese que conforme a crece el límite la probabilidad tiende al valor de $2/3$.

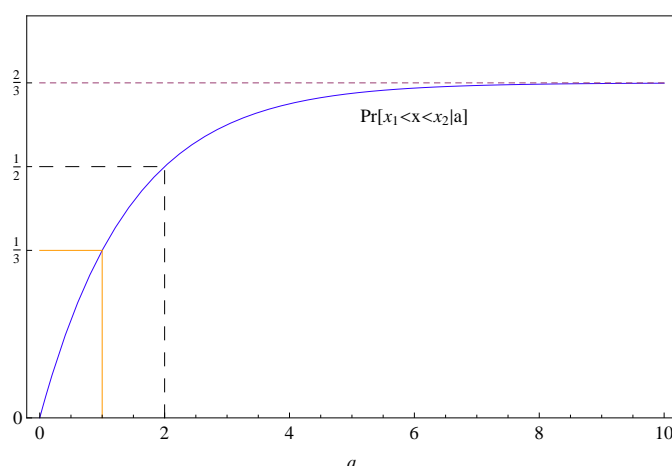


Figura 4.5: Probabilidad de $\Pr(x_1 < x < x_2|a)$ donde la inicial es $\pi(\alpha, \varphi) \propto \varphi^{-a}$, $a > 0$, con $a = 1$ se tiene $\pi_\phi(\phi, \varphi)$ y con $a = 2$ se tiene $\pi_G(\alpha, \varphi)$.

Además, la Tabla 4.10 muestra la cobertura de probabilidad de la densidad predictiva para $a = 1, 2$. Puesto que el valor es exacto, la desviación estándar es insignificante por lo que fue necesario solo simular $m = 100$ muestras aleatorias de tamaño dos.

a	α	$\varphi = 1$	$\varphi = 1/2$	$\varphi = 1/5$
		Media	Media	Media
1	-2	0.3333	0.3333	0.3333
	1/2	0.3333	0.3333	0.3333
	4	0.3333	0.3333	0.3333
2	-2	0.5	0.5	0.5
	1/2	0.5	0.5	0.5
	4	0.5	0.5	0.5

Tabla 4.10: *Media muestral de la cobertura de probabilidad $\Pr[x_1 < x < x_2|x_1, x_2]$ proveniente de $m = 100$ muestras simuladas de tamaño dos de $\text{Ex}(x|\alpha, \varphi)$, con diferentes valores de α y φ .*

Es natural preguntarse si con más de dos variables aleatorias la densidad predictiva resultante de la inicial $\pi_\phi(\phi, \varphi) \propto \varphi^{-1}$ cuenta con adecuadas propiedades de cobertura. En particular, si se acerca al valor $\frac{n-1}{n+1}$ (Teorema 10).

Supóngase que se tiene una muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_n\}$ con $n > 2$, la densidad predictiva tiene la siguiente expresión

$$p(x|\mathbf{x}) = \frac{n(n+a-2)(t-nx_{(1)})^{n+a-2}}{n+1} \times \begin{cases} (t-nx)^{-(n+a-1)} & x < x_{(1)}, \\ (t+x-(n+1)x_{(1)})^{-(n+a-1)} & x > x_{(1)}. \end{cases}$$

Por tanto, la probabilidad requerida $\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x})$ es la función de los datos

$$\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x}) = \frac{n}{n+1} \left(1 - \left(\frac{t-nx_{(1)}}{t+x_{(n)}-(n+1)x_{(1)}} \right)^{n+a-2} \right)$$

cuyo valor esperado en el muestreo será aproximado numéricamente.

La Tabla 4.11 muestra el valor medio (desviación estándar) de $\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x})$ para más de dos observaciones y con diferentes valores de α y φ . Puede verse que el valor medio se acerca al valor teórico $\frac{n-1}{n+1}$ mostrando una concordancia con el Teorema 10.

La obtención detallada de las densidades final y predictiva para $n > 2$ puede encontrarse en el Apéndice B.2.

n	α	$\varphi = 1$		$\varphi = 1/2$		$\varphi = 1/5$	
		Med	DvSt	Med	DvSt	Med	DvSt
3	-2	0.4998	0.0018	0.4990	0.0018	0.4990	0.0019
	1/2	0.4990	0.0017	0.4972	0.0018	0.4991	0.0017
	4	0.5029	0.0017	0.4988	0.0017	0.5001	0.0017
4	-2	0.6006	0.0024	0.6019	0.0024	0.6005	0.0026
	1/2	0.60019	0.0024	0.5985	0.0026	0.6000	0.0026
	4	0.6025	0.0025	0.5995	0.0026	0.5990	0.0026
7	-2	0.7511	0.0027	0.7458	0.0026	0.7479	0.0026
	1/2	0.7487	0.0026	0.7532	0.0025	0.7469	0.0025
	4	0.7568	0.0023	0.7501	0.0025	0.7508	0.0026
9	-2	0.7998	0.0028	0.8036	0.0022	0.8009	0.0022
	1/2	0.8009	0.0022	0.7983	0.0028	0.7975	0.0022
	4	0.7998	0.0026	0.8000	0.0021	0.7997	0.0021
15	-2	0.8746	0.0012	0.8776	0.0011	0.8728	0.0014
	1/2	0.8751	0.0012	0.8753	0.0012	0.8755	0.0012
	4	0.8744	0.0013	0.8775	0.0012	0.8764	0.0012

Tabla 4.11: Valor medio de la cobertura de probabilidad $\Pr[x_{(1)} < x < x_{(n)} | \mathbf{x}]$ proveniente de $m = 100$ muestras aleatorias de tamaño n de $\text{Ex}(x|\alpha, \theta)$, para diferentes valores de α y φ con función inicial de referencia $\pi_\phi(\alpha, \varphi) \propto \varphi^{-1}$.

Consideraciones finales sobre el modelo $\text{Ex}(\cdot|\alpha, \varphi)$.

Supongamos ahora que nuestro parámetro de interés es φ . En la Sección 4.5.3 hemos obtenido la inicial $\pi(\alpha)$ cuando se tiene el modelo $\text{Ex}(\cdot|\alpha, 1)$, nótese que la obtención de $\pi(\alpha|\varphi)$ es esencialmente la misma, por tanto $\pi(\alpha|\varphi) \propto 1$.

Al ser una inicial impropia, deberá definirse una sucesión creciente de intervalos donde $\pi_i(\alpha|\varphi)$ sea propia. De esa manera, se podrá obtener el modelo integrado de un solo parámetro $p_i(x|\varphi)$ (véase el Apéndice B.3.2).

Aplicando nuevamente el algoritmo numérico de obtención de iniciales al modelo integrado $p_i(x|\varphi)$ y así obtener $\pi_i(\varphi)$, resulta que la función φ^{-1} se ajusta a las aproximaciones de $\pi_i(\varphi)$. Véase el Apéndice B.3.3.

Por tanto, puede concluirse que la inicial de referencia cuando φ es el parámetro de interés es $\pi_\varphi(\alpha, \varphi) \propto \varphi^{-1}$.

Por otro lado, partamos de $\pi_\varphi(\alpha, \varphi) \propto \varphi^{-1}$, si definimos la transformación lineal de $\phi = \alpha + \varphi \log 2$ (mediana) y φ , es muy fácil verificar que

$$\pi_\varphi(\phi, \varphi) = \pi_\varphi(\phi(\alpha, \varphi), \varphi) |J| = \varphi^{-1}$$

donde

$$J = \begin{pmatrix} \frac{\partial \alpha}{\partial \phi} & \frac{\partial \alpha}{\partial \varphi} \\ \frac{\partial \varphi}{\partial \phi} & \frac{\partial \varphi}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} 1 & -\log 2 \\ 0 & 1 \end{pmatrix}.$$

Es importante subrayar que para las iniciales de esta forma la densidad predictiva resultante cuenta con adecuadas propiedades de cobertura. De lo que puede concluirse que *no hay una única* cantidad de interés para el problema de predicción en el sentido de que distintas cantidades de interés pueden producir la misma inicial de referencia. Sin embargo, la elección de la mediana del modelo como cantidad de interés *siempre* produce una inicial apropiada.

Una consideración final: el método de obtención de funciones iniciales propuesto por Bernardo (1979, 2005a), Berger y Bernardo (1992a) y Berger *et al.* (2009a) no presenta problemas pues sólo es necesario definir una cantidad de interés y sobre ésta aplicar el algoritmo de obtención (Teorema 2 de esta Tesis). Esta situación contrasta con la metodología de Ghosal (1997) con la que no es obvio como proceder cuando existen parámetros de ruido.

Capítulo 5

Funciones iniciales para predicción en el modelo Normal bivariado.

Como extensión natural de haber estudiado modelos univariados en el capítulo anterior, en este capítulo se pretende definir la inicial de referencia para predicción en el modelo Normal bivalente.

Como primer punto, comprobaremos numéricamente si se cumple en el modelo Normal bivariado la proposición del Teorema 9: Dados dos puntos \mathbf{x}_1 y \mathbf{x}_2 provenientes de una distribución $N_2(\cdot|\boldsymbol{\mu}, \Sigma)$, encontrar la probabilidad de que un tercer punto \mathbf{x} esté contenido en el rectángulo formado por las dos observaciones previas. La probabilidad de que esto ocurra se aproxima a 3^{-2} cuando las variables son independientes, y a $h(\rho)3^{-2}$ cuando existe dependencia.

Partiendo de este resultado frecuentista, nos centraremos en el caso particular bivariado $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$ con vector de medias conocido e igual al vector de ceros y $\{\sigma_1, \sigma_2, \rho\}$ los elementos de la matriz de dispersión Σ . De esta manera solo trabajaremos con un modelo de tres parámetros en sus distintos casos.

El objetivo de este capítulo se centrará en encontrar la función inicial cuya densidad predictiva cuente con adecuadas propiedades de cobertura, es decir, que tenga una concordancia con los resultados frecuentistas. Recordemos que esta ha sido la forma de discriminación entre las funciones iniciales alternativas y dado que en los modelos multivariados no está inequívocamente definida la mediana del modelo, su aplicación será esencial en la elección de la cantidad de interés que *sugerirá* la función inicial adecuada para predicción.

Debido a la complejidad computacional de los cálculos presentados en este capítulo, la mayoría de las aproximaciones numéricas fueron hechas en Mathematica v.6.0 para Linux x86 (64 bit), puesto que son aproximaciones que requieren gran cantidad de memoria virtual y ser de larga duración. Mientras que las más simples fueron hechas en Mathematica

v.7.0 en un ordenador portátil con procesador T7500.

5.1. Densidad Normal bivariada: estudio frecuentista

A continuación verificamos numéricamente si el modelo Normal bivariado cumple con la propiedad frecuentista similar al caso univariado expuesto en el Teorema 9.

Frecuencia con que una nueva observación se encuentra contenido en el rectángulo formado por dos observaciones previas.

Considérese el vector aleatorio $\mathbf{x} = (x_1, x_2)$ proveniente del modelo Normal bivariado $N_2(\cdot|\boldsymbol{\mu}, \Sigma)$, cuya función de densidad es dada por

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) \right\}$$

con vector de medias $\boldsymbol{\mu} = (\mu_1, \mu_2)$ y matriz de varianzas covarianzas

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}.$$

Sean $\mathbf{x}_1 = (x_{11}, x_{12})$ y $\mathbf{x}_2 = (x_{21}, x_{22})$ dos vectores aleatorios independientes con función de densidad de probabilidad $N_2(\cdot|\boldsymbol{\mu}, \Sigma)$, y se requiere saber la probabilidad de que un nuevo punto \mathbf{x} esté contenido en el rectángulo formado por \mathbf{x}_1 y \mathbf{x}_2 , es decir,

$$\Pr(\mathbf{x} \in C|\boldsymbol{\mu}, \Sigma) = \int_C N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x}$$

donde $C = \{(x_1, x_2) : (x_1, x_2) \in (x_{11}, x_{12}) \times (x_{21}, x_{22})\}$.

Nótese que existen 2^2 casos de estudio, todos ellos similares al siguiente caso

$$x_{11} < x_{21} \quad \text{y} \quad x_{12} < x_{22}, \tag{5.1}$$

dado lo anterior centraremos nuestro desarrollo en éste caso.

Defínase el área $C_{\mathbf{x}_1\mathbf{x}_2} = \{(x_1, x_2) : x_{11} < x_1 < x_{21} \text{ y } x_{12} < x_2 < x_{22}\}$. La probabilidad es ahora

$$\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2}|\boldsymbol{\mu}, \Sigma) = \int_{C_{\mathbf{x}_1\mathbf{x}_2}} N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x}.$$

Trabajemos con la función de densidad $N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ y el siguiente cambio de variables con

su respectivo Jacobiano de la transformación

$$w = \frac{x_1 - \mu_1}{\sigma_1}, \quad z = \frac{x_2 - \mu_2}{\sigma_2}, \quad \text{y} \quad J = \begin{pmatrix} \frac{\partial x_1}{\partial w} & \frac{\partial x_1}{\partial z} \\ \frac{\partial x_2}{\partial w} & \frac{\partial x_2}{\partial z} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}.$$

Entonces,

$$\phi(w, z; \rho) = N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma)|J| = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{w^2 - 2\rho wz + z^2}{2(1-\rho^2)}\right\}$$

donde ϕ es la Normal bivariada estandarizada y

$$\Phi(w, z; \rho) \equiv \int_{-\infty}^w \int_{-\infty}^z \phi(t, s; \rho) dt ds$$

Apliquemos estas transformaciones para obtener la probabilidad que nos interesa,

$$\Pr(\mathbf{x} \in C_{\mathbf{x}_1 \mathbf{x}_2} | \boldsymbol{\mu}, \Sigma) = \int_{\frac{x_{11}-\mu_1}{\sigma_1}}^{\frac{x_{21}-\mu_1}{\sigma_1}} \int_{\frac{x_{12}-\mu_2}{\sigma_2}}^{\frac{x_{22}-\mu_2}{\sigma_2}} \phi(w, z; \rho) dw dz$$

donde w y z se han definido anteriormente. Claramente,

$$\begin{aligned} \Pr(\mathbf{x} \in C_{\mathbf{x}_1 \mathbf{x}_2} | \boldsymbol{\mu}, \Sigma) &= \begin{pmatrix} \Phi\left(\frac{x_{21}-\mu_1}{\sigma_1}, \frac{x_{22}-\mu_2}{\sigma_2}; \rho\right) - \Phi\left(\frac{x_{11}-\mu_1}{\sigma_1}, \frac{x_{22}-\mu_2}{\sigma_2}; \rho\right) \\ -\Phi\left(\frac{x_{21}-\mu_1}{\sigma_1}, \frac{x_{12}-\mu_2}{\sigma_2}; \rho\right) + \Phi\left(\frac{x_{11}-\mu_1}{\sigma_1}, \frac{x_{12}-\mu_2}{\sigma_2}; \rho\right) \end{pmatrix} \\ &= g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma). \end{aligned}$$

Será necesario calcular el valor esperado de $g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma)$ sobre todos los posibles valores que pudiesen tener \mathbf{x}_1 y \mathbf{x}_2 bajo el supuesto (5.1).

$$\begin{aligned} E_{\mathbf{x}_1 \mathbf{x}_2} (g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma) | x_{11} < x_{21}, x_{12} < x_{22}) &= \\ \int_{C_{\mathbf{x}_1 \mathbf{x}_2}} g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma) N_2(\mathbf{x}_1 | \boldsymbol{\mu}, \Sigma) N_2(\mathbf{x}_2 | \boldsymbol{\mu}, \Sigma) d\mathbf{x}_1 d\mathbf{x}_2. & \end{aligned} \quad (5.2)$$

Puesto que existen 2^2 casos similares, el valor esperado total será

$$E_{\mathbf{x}_1 \mathbf{x}_2} (g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma)) = 4E (g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma) | x_{11} < x_{21}, x_{12} < x_{22}).$$

No hemos logrado obtener una expresión analítica de la esperanza (5.2) pero hemos

encontrado una aproximación numérica de la esperanza total $E_{\mathbf{x}_1, \mathbf{x}_2} (g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma))$

$$E_{\mathbf{x}_1, \mathbf{x}_2} (g(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}, \Sigma)) \approx \begin{cases} \frac{1}{9} & \text{si } \rho = 0 \\ \frac{h(\rho)}{9} & \text{si } \rho \neq 0 \end{cases}$$

donde $h(\rho)$ puede ser aproximada por la función $\frac{1}{\sqrt{1-\rho^2}}$.

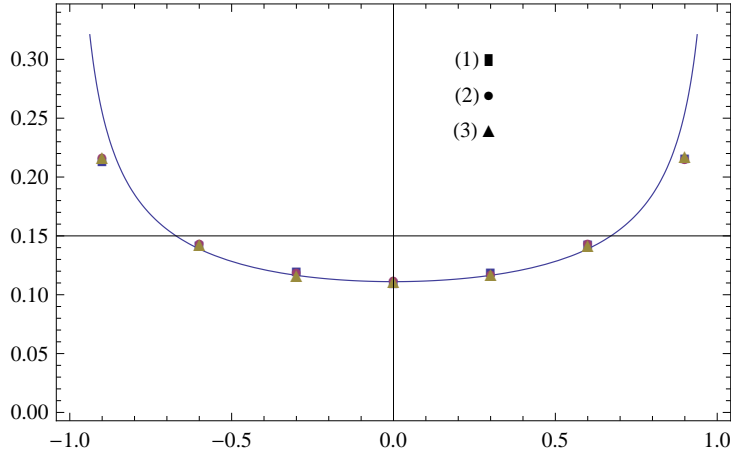


Figura 5.1: Gráfica de las frecuencias con que una nueva observación se sitúa en el rectángulo formado por dos observaciones provenientes de $N_2(\cdot | \boldsymbol{\mu}, \Sigma)$ junto con la función teórica $\frac{1}{9\sqrt{1-\rho^2}}$ (línea continua). Las frecuencias fueron generadas con parámetros $\{\mu_1, \mu_2, \sigma_1, \sigma_2\}$ donde (1) corresponde a $\{0, 0, 1, 1\}$, (2) a $\{-2, 1, 2, \frac{1}{2}\}$ y (3) a $\{-7, 4, 4, \frac{1}{4}\}$.

La Tabla 5.1 muestra las frecuencias con que ocurre el evento descrito anteriormente para diferentes valores de $\{\mu_1, \mu_2, \sigma_1, \sigma_2\}$ y ρ , y la Figura 5.1 muestra el comportamiento de éstas con respecto a la función $9^{-1}(1-\rho^2)^{-1/2}$. Los valores graficados corresponden a los parámetros $\{\mu_1, \mu_2, \sigma_1, \sigma_2\}$, por ejemplo, los puntos correspondientes a (1) fueron generados de una $N_2(\cdot | \boldsymbol{\mu}, \Sigma)$ con valores dados por $\{0, 0, 1, 1\}$ y $\rho = 0, \pm 0.3, \pm 0.6, \pm 0.9$.

Frecuencias

(μ_1, μ_2)	$\{\sigma_1, \sigma_2\}$	ρ						
		-0.9	-0.6	-0.3	0	0.3	0.6	0.9
(0, 0)	{1, 1}	0.2155	0.1425	0.1180	0.1118	0.1155	0.1435	0.2165
(-2, 1)	{2, $\frac{1}{2}$ }	0.2171	0.1422	0.1180	0.1122	0.1188	0.1425	0.2167
(-7, 4)	{4, $\frac{1}{4}$ }	0.2167	0.1444	0.1184	0.1113	0.1177	0.1446	0.2161
$\frac{1}{9} h(\rho)$		0.2549	0.1389	0.1165	0.1111	0.1165	0.1389	0.2549

Tabla 5.1: Frecuencia con que una tercera observación cae en el rectángulo formado por dos observaciones dadas provenientes de $N_2(\mathbf{x} | \boldsymbol{\mu}, \Sigma)$. Frecuencias de $m = 100000$ muestras simuladas de tamaño dos con varios valores de $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\{\sigma_1, \sigma_2\}$ y ρ .

5.2. Densidad Normal bivariada: estudio Bayesiano

Esta sección se centrará en el modelo normal bivariado $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$ con vector de medias conocido e igual a $\mathbf{0}$ y matriz de dispersión Σ con elementos σ_1 , σ_2 y ρ . Dado que el modelo se reduce a tres parámetros, trabajaremos con algunos de los posibles casos, en particular,

- Varianzas conocidas e iguales, parámetro de correlación desconocido.
- Varianzas desconocidas e iguales y parámetro de correlación desconocido.

También discutimos el caso general, cuando todos los parámetros son desconocidos.

5.2.1. Caso ρ desconocido.

Analicemos las iniciales alternativas para el modelo

$$N_2\left(\mathbf{x}|\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad (5.3)$$

donde ρ es desconocido y se asume que los parámetros de escala son conocidos e iguales $\sigma_1 = \sigma_2 = 1$. Entonces la función de densidad es

$$f(x_1, x_2|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1 x_2)\right\}.$$

En este caso particular, aparentemente existen únicamente dos iniciales no informativas sugeridas en la literatura como posibles candidatas:

- La inicial de referencia:

$$\pi_{R_p}(\rho) \propto \sqrt{i(\rho)} = \frac{\sqrt{1+\rho^2}}{1-\rho^2};$$

- y la inicial HPD *matching* sugerida por Datta y Mukerjee (2004) para este modelo:

$$\pi_{DM}(\rho) \propto \frac{(1-\rho^2)^2}{1+\rho^2}.$$

Complementamos nuestro análisis al considerar también la transformación $\varphi = \operatorname{arctanh}\rho$ (Jeffreys, 1961, Sec. 8.2) junto con el supuesto

$$\pi_\varphi(\varphi) \propto 1$$

que expresada en términos de ρ es la inicial $\pi_\varphi(\rho) \propto \frac{1}{1-\rho^2}$.

Utilizaremos la siguiente expresión general de las iniciales anteriores

$$\pi_{ab}(\pi) \propto (1 - \rho^2)^a (1 + \rho^2)^b. \quad (5.4)$$

A continuación detallamos el desarrollo de cada una de estas iniciales. Al final de esta subsección pueden encontrarse los resultados numéricos.

Inicial de referencia π_{R_ρ} .

Bajo el supuesto de normalidad asintótica de la densidad final de ρ , la inicial de referencia puede obtenerse fácilmente en modelos de un solo parámetro. En este caso la función inicial es

$$\pi_{R_\rho}(\rho) \propto \sqrt{i(\rho)}$$

donde

$$i(\rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(-\frac{\partial^2}{\partial \rho^2} \log[f(\mathbf{x}|\rho)] \right) f(\mathbf{x}|\rho) d\mathbf{x} = \frac{1 + \rho^2}{(1 - \rho^2)^2},$$

por lo que

$$\pi_{R_\rho}(\rho) \propto \frac{\sqrt{1 + \rho^2}}{1 - \rho^2}.$$

Dadas dos observaciones \mathbf{x}_1 y \mathbf{x}_2 , la densidad final es proporcional a

$$\pi_{R_\rho}(\rho|\mathbf{x}_1, \mathbf{x}_2) \propto f(\mathbf{x}_1|\rho) f(\mathbf{x}_2|\rho) \pi_{R_\rho}(\rho)$$

con constante de marginalización $c = \int_{-1}^1 f(\mathbf{x}_1|\rho) f(\mathbf{x}_2|\rho) \pi_{R_\rho}(\rho) d\rho$.

Véase el Apéndice D donde se demuestra que la densidad final es propia para $n \geq 2$.

Para ver la cobertura de probabilidad de la densidad final sobre el valor de ρ , la Tabla 5.2 (primer renglón), muestra la frecuencia con que el verdadero valor de ρ se encuentra contenido en el intervalo $(-1, q)$, con q un cuantil de la densidad final con probabilidad acumulada

$$\Pr(\rho < q|\mathbf{x}_1, \mathbf{x}_2) = \alpha.$$

La densidad predictiva es expresada en términos de $\pi_{R_\rho}(\rho|\mathbf{x}_1, \mathbf{x}_2)$ y de la función de densidad de una nueva observación puesto que no se tiene una expresión analítica

explícita,

$$p_{R_\rho}(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \int_{-1}^1 \pi_R(\rho|\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}|\rho) d\rho.$$

Recuérdese que estamos interesados en la probabilidad del suceso $\{\mathbf{x} \in \mathbf{x}_1 \times \mathbf{x}_2\}$ para lo que definimos la región

$$C_{\mathbf{x}_1\mathbf{x}_2} = \{(x_1, x_2) : (x_1, x_2) \in (x_{11}, x_{12}) \times (x_{21}, x_{22})\}$$

entonces

$$\Pr(\mathbf{x} \in \mathbf{x}_1 \times \mathbf{x}_2) = \int_{C_{\mathbf{x}_1\mathbf{x}_2}} p_{R_\rho}(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}.$$

Ésta probabilidad será aproximada numéricamente y en la Tabla 5.3 (primer renglón) pueden verse el valor medio (y desviación estándar) de la cobertura de probabilidad $\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2}|\mathbf{x}_1, \mathbf{x}_2)$. La obtención de los resultados se discute posteriormente.

Función inicial *HPD matching* de Datta y Mukerjee (2004).

La función inicial *HPD matching* sugerida por Datta and Mukerjee (2004, p. 71) para el modelo 5.3 es

$$\pi_{DM}(\rho) \propto \frac{(1 - \rho^2)^2}{1 + \rho^2}.$$

Recordemos que las regiones con mayor densidad final o HPD (*highest posterior density*) son regiones de la forma

$$\{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\mathbf{X}) \geq K\}$$

donde $\boldsymbol{\theta}$ puede ser un escalar o un vector, $\pi(\boldsymbol{\theta}|\mathbf{x})$ es la densidad final resultante de la inicial $\pi(\cdot)$ y los datos \mathbf{X} , K es un nivel dado de credibilidad sobre la densidad final.

Las regiones HPD contienen el mínimo volumen de densidad posible dados los datos \mathbf{X} para un nivel de credibilidad de la densidad final. Las funciones iniciales que aseguren una cobertura frecuentista de estas regiones son conocidas como *HPD matching*.

Datta y Mukerjee (2004) demuestran que la inicial de referencia $\pi(\rho) \propto \sqrt{i(\rho)}$ no es *HPD matching* mientras que $\pi_{DM} \propto i(\rho)^{-1}$ si cumple con esta propiedad. Para más detalles, véase Datta y Sweeting (2005, Secc.5.6) y referencias ahí incluidas.

Retornando a nuestro objetivo, dadas dos observaciones $\{\mathbf{x}_1, \mathbf{x}_2\}$ la densidad final es

proporcional a

$$\pi_{DM}(\rho|\mathbf{x}_1, \mathbf{x}_2) \propto f(\mathbf{x}_1|\rho)f(\mathbf{x}_2|\rho)\pi_{DM}(\rho)$$

con constante de marginalización $c = \int_{-1}^1 f(\mathbf{x}_1|\rho)f(\mathbf{x}_2|\rho)\pi_{DM}(\rho)d\rho < \infty$, por ser π_{DM} propia.

Véase el comportamiento de la convergencia de la densidad final π_{DM} que se muestra en el segundo renglón de la Tabla 5.2. En dicha tabla se presenta la frecuencia con que la densidad final contiene al verdadero valor ρ , es decir, dadas varias muestras aleatorias de tamaño n , contar el número de veces en que la densidad final de ρ incluye al verdadero valor ρ , al final de dicha tabla se encuentra una discusión más detallada de sus resultados.

La densidad predictiva tiene la siguiente expresión

$$p_{DM}(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \int_{-1}^1 \pi_{DM}(\rho|\mathbf{x}_1, \mathbf{x}_2)f(\mathbf{x}|\rho)d\rho.$$

Abordando el problema de nuestro estudio, calculemos la cobertura de probabilidad de la densidad predictiva en el área del rectángulo formado por los puntos \mathbf{x}_1 y \mathbf{x}_2 , es decir, la región $C_{\mathbf{x}_1\mathbf{x}_2}$ definida anteriormente,

$$\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2}|\mathbf{x}_1, \mathbf{x}_2) = \int_{C_{\mathbf{x}_1\mathbf{x}_2}} p_{R_\rho}(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x},$$

que será aproximada numéricamente.

La Tabla 5.3 (segundo renglón) muestra el valor medio (y desviación estándar) de la cobertura de probabilidad $\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2}|\mathbf{x}_1, \mathbf{x}_2)$ para diferentes valores de ρ , al final de la tabla se discuten los resultados.

Función inicial de $\varphi = \operatorname{arctanh}\rho$ con $\pi_\varphi(\varphi) \propto 1$

Finalmente, usando la transformación

$$\varphi = \operatorname{arctanh}(\rho)$$

de ρ (Jeffreys, 1961, Sec.8.2) y suponiendo que la densidad inicial es uniforme

$$\pi_\varphi(\varphi) \propto 1,$$

que es equivalente a la función inicial $\pi_\varphi(\rho) \propto \frac{1}{1-\rho^2}$ en términos de ρ .

La función de densidad para una observación \mathbf{x} y $\varphi = \operatorname{arctanh}\rho$ tendrá la expresión

siguiente:

$$f(\mathbf{x}|\varphi) = \frac{1}{2\pi\operatorname{sech}\varphi} \exp\left\{-\frac{1}{2} \cosh^2 \varphi (x_1^2 + x_2^2 - 2x_1x_2 \tanh \varphi)\right\}.$$

Dadas dos observaciones $\mathbf{x}_1 = (x_{11}, x_{12})$ and $\mathbf{x}_2 = (x_{21}, x_{22})$, la función de verosimilitud es proporcional a la densidad final ya que $\pi_\varphi(\varphi)$ es constante

$$\begin{aligned} L(\varphi|\mathbf{x}_1, \mathbf{x}_2) &\propto \frac{1}{4\pi^2\operatorname{sech}^2\varphi} \exp\left\{-\frac{1}{2} \cosh^2 \varphi (s_{.1} + s_{.2} - 2s_{12} \tanh \varphi)\right\} \\ \pi_\varphi(\varphi|\mathbf{x}_1, \mathbf{x}_2) &\propto L(\varphi|\mathbf{x}_1, \mathbf{x}_2)\pi_\varphi(\varphi) \end{aligned}$$

donde $s_{.1} = x_{11}^2 + x_{21}^2$, $s_{.2} = x_{12}^2 + x_{22}^2$ y $s_{12} = x_{11}x_{12} + x_{21}x_{22}$.

No hay una expresión analítica de la densidad final, pero puede demostrarse fácilmente que, en términos de ρ , la constante de marginalización es finita

$$c = \int_{-1}^1 \pi(\rho|\mathbf{x}_1, \mathbf{x}_2) d\rho < \infty$$

cuyos detalles pueden verse en el Apéndice D.

Análogo a las dos funciones iniciales antes estudiadas, presentamos en la Tabla 5.2 (tercer renglón) la convergencia de la cobertura de π_φ , la discusión de los resultados pueden verse al pie de la tabla.

Obtengamos ahora la expresión de la densidad predictiva

$$p_\varphi(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \int_{-\infty}^{\infty} \pi_\varphi(\varphi|\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}|\varphi) d\varphi.$$

de la que tampoco se cuenta con una expresión analítica explícita.

Puesto que nos interesa la probabilidad

$$\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2}|\mathbf{x}_1, \mathbf{x}_2) = \int_{C_{\mathbf{x}_1\mathbf{x}_2}} p_\varphi(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}$$

del que aproximaremos su valor esperado numéricamente. La Tabla 5.3 (tercer renglón) muestra el valor medio (y la desviación estándar) de la cobertura de $\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2}|\mathbf{x}_1, \mathbf{x}_2)$ para diferentes valores de ρ .

Inicial	Forma (5.4) (a,b)	ρ		
		0	0.5	0.9
$\pi_{R\rho}(\rho) \propto \frac{\sqrt{1+\rho^2}}{1-\rho^2}$	$(-1, \frac{1}{2})$	0.8888	0.8862	0.8955
$\pi_{DM}(\rho) \propto \frac{(1-\rho^2)^2}{1+\rho^2}$	$(2, -1)$	0.9738	0.6398	0.0694
$\pi_\varphi(\varphi) \propto 1$		0.8989	0.901	0.9990

Tabla 5.2: Convergencia de las densidades finales $\pi_{R\rho}$, π_{DM} y π_φ . Frecuencias con que el verdadero valor ρ pertenece al intervalo $(-1, q)$ dado que $\Pr(\rho < q | \mathbf{x}_1, \mathbf{x}_2) = \alpha$. Valores provenientes de $m = 10000$ muestras simuladas de tamaño $n = 5$ con $\alpha = 0.9$.

El contenido de la Tabla 5.2 fue calculado como se describe a continuación:

- (i) Fíjese una probabilidad α y un valor ρ .
- (ii) Simúlese n vectores aleatorios de $N_2(\mathbf{x} | \mathbf{0}, \Sigma)$ con $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
- (iii) Dada una inicial $\pi(\rho)$, obténgase la densidad final $\pi(\rho | \mathbf{x}_1, \dots, \mathbf{x}_n)$ y el cuantil q para el cual se tiene una probabilidad final acumulada igual a α , es decir,

$$\Pr(\rho < q | \mathbf{x}_1, \mathbf{x}_2) = \int_{-1}^q \pi(\tau | \mathbf{x}, \dots, \mathbf{x}_n) d\tau = \alpha.$$

- (iv) Verifíquese si el verdadero valor de ρ esté incluido en el intervalo $(-1, q)$ y almacene el caso.
- (v) Repítanse los pasos (ii)-(iv) m -veces.

Conclusiones de los resultados de la Tabla 5.2

La Tabla 5.2 muestra la convergencia de la densidad final de ρ para una función inicial dada y los datos. Para estudiar la convergencia de dichas funciones finales, se presentan las frecuencias con que el valor ρ pertenece al intervalo $(-1, q)$, con q tal que $\alpha = \Pr(\rho < q | \mathbf{x}) = \int_{-1}^q \pi(\tau | \mathbf{x}) d\tau$, para cada función inicial. Se espera que el 100% de las veces, el valor verdadero de ρ esté contenido en el intervalo $(-1, q)$ para un α dado. Las frecuencias de la Tabla 5.2 fueron obtenidas con los valores $\alpha = 0.90$, $n = 5$, $m = 10000$ y tres valores para ρ .

De la Tabla 5.2 puede concluirse que las densidades finales provenientes de usar $\pi_{R\rho}$ y π_φ tienen una mejor cobertura frecuentista que π_{DM} . Además para altas correlaciones, la función inicial de referencia es claramente preferible.

$\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2} \mathbf{x}_1, \mathbf{x}_2) = \int_{C_{\mathbf{x}_1\mathbf{x}_2}} p(\mathbf{x} \mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}$							
Inicial	Forma (5.4)	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
		Prom	DvSt	Prom	DvSt	Prom	DvSt
$\pi_{R\rho}(\rho)$	$(-1, \frac{1}{2})$	0.1296	0.0146	0.1552	0.0159	0.2161	0.0237
$\pi_{DM}(\rho)$	$(2, -1)$	0.1157	0.0144	0.1245	0.0162	0.1619	0.0210
$\pi_\varphi(\varphi)$		0.1317	0.0134	0.1481	0.0179	0.2152	0.0226
Frecuencias		0.1147		0.1296		0.2056	

Tabla 5.3: Valor medio y desviación estándar de la cobertura de probabilidad de la densidad predictiva resultante de las iniciales $\pi_{R\rho}$, π_{DM} y $\pi_{R\rho}$. Valores correspondientes de $l = 50$ frecuencias de $m = 100$ muestras de $N_2(\mathbf{x} | \boldsymbol{\mu}, \Sigma)$. Frecuencias (en negro) de que un nuevo punto aleatorio \mathbf{x} esté contenido en el rectángulo formado por los puntos $\mathbf{x}_1 \times \mathbf{x}_2$ correspondientes de $m = 10000$ muestras de $N_2(\mathbf{x} | \mathbf{0}, \Sigma)$. Las coberturas de probabilidad deberían de aproximarse a las frecuencias, es decir $1/3^2 \simeq 0.1111$ cuando $\rho = 0$ y $h(\rho)/9$ cuando $\rho \neq 0$.

Descripción del contenido de la Tabla 5.3:

- (i) Dado un valor de ρ , genérese dos puntos aleatorios \mathbf{x}_1 y \mathbf{x}_2 de $N_2(\mathbf{x} | \mathbf{0}, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
- (ii) Dada una función inicial $\pi(\rho)$, calcúlese la densidad final $\pi(\rho | \mathbf{x}_1, \mathbf{x}_2)$ y la correspondiente densidad predictiva $p(\mathbf{x} | \mathbf{x}_1, \mathbf{x}_2)$.
- (iii) Calcúlese $\int_{C_{\mathbf{x}_1\mathbf{x}_2}} p(\mathbf{x} | \mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}$ y almacénese en memoria el resultado.
- (iv) Repítanse los pasos (i)-(iii) m veces y promédiense las m probabilidades.
- (v) Repítanse todos los pasos l -veces, promédiense y calcúlese la desviación estándar.

El objetivo de la Tabla 5.3 es contrastar el valor frecuentista y el valor medio de la cobertura de probabilidad de la predictiva de las tres iniciales bajo estudio. Puede apreciarse, a la luz de las aproximaciones numéricas, que la inicial de referencia $\pi_{R\rho}$ y la inicial π_φ producen densidades predictivas con mejores propiedades de cobertura. No sucediendo lo mismo con la inicial HPD *matching* π_{DM} sugerida por Datta y Mukerjee (2004) pues puede verse que se tiene una pobre aproximación de la cobertura de probabilidad para el valor extremo de $\rho = 0.9$.

El comportamiento de la cobertura de la densidad predictiva resultante de usar π_{DM} es consecuencia natural de la convergencia de su densidad final π_{DM} , pues como puede verse en la Tabla 5.2 (segundo renglón) el número de veces en que el valor verdadero de ρ es considerado por la densidad final para el valor $\rho = 0.9$ es cercano a cero.

Por otro lado, nótese que la cobertura de las predictivas provenientes de $\pi_{R\rho}$ y de π_φ se mantienen similares en los tres diferentes los valores de ρ en el modelo 5.3. De este hecho,

podemos sugerir estas dos funciones iniciales como adecuadas para predicción dado que producen densidades predictivas con buenas propiedades de cobertura frecuentista.

En la Tabla se utilizaron los valores $m = 100$ y $l = 50$ para calcular el valor promedio de la cobertura de probabilidad de la densidad predictiva. Los valores en negro corresponden a la frecuencia con que una nueva observación se encuentra contenida en el rectángulo formado por dos observaciones previas. Las frecuencias corresponden a $m = 10000$ muestras simuladas de $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$, con

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

para cada valor de ρ .

5.2.2. Caso ρ y σ desconocidos.

En esta subsección se estudia el caso cuando σ_1, σ_2 y ρ son desconocidos, pero las varianzas son iguales $\sigma = \sigma_1 = \sigma_2$, es decir,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Para los supuestos anteriores, se proponen las siguientes funciones iniciales:

- Las funciones iniciales de referencia cuando σ o ρ es el parámetro de interés.
- La función inicial de Jeffreys sugerida de su regla general.
- Finalmente, se incluyen y discuten algunas transformaciones de los elementos de Σ para las cuales se obtienen sus funciones iniciales de referencia.

Recuérdese que se asume siempre que el vector de medias es conocido e igual a $\boldsymbol{\mu} = (0, 0)'$.

Como primer paso, se especifica la notación usada en esta subsección:

La función de densidad de $\mathbf{x} = (x_1, x_2)$ denotada por $N_2(\mathbf{x}|\mathbf{0}, \sigma^2 I_2)$ tiene la forma siguiente

$$f(\mathbf{x}|\sigma, \rho) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma^2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1 x_2) \right\}.$$

Dadas dos observaciones $\mathbf{x}_1 = (x_{11}, x_{12})$ y $\mathbf{x}_2 = (x_{21}, x_{22})$, la función de verosimilitud es expresada en términos de sus elementos o en términos de sus sumas, como puede verse a

continuación

$$\begin{aligned} L(\sigma, \rho) &= \frac{1}{4\pi^2\sigma^4(1-\rho^2)} \exp \left\{ -\frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}{2\sigma^2(1-\rho^2)} \right\} \\ &= \frac{1}{4\pi^2\sigma^4(1-\rho^2)} \exp \left\{ -\frac{1}{2\sigma^2(1-\rho^2)}(s_{.1} + s_{.2} - 2\rho s_{12}) \right\} \end{aligned}$$

donde $s_{.1} = x_{11}^2 + x_{21}^2$, $s_{.2} = x_{12}^2 + x_{22}^2$ and $s_{12} = x_{11}x_{12} + x_{21}x_{22}$.

Todas las funciones iniciales estudiadas pertenecen a la familia paramétrica:

$$\pi_{abc}(\sigma, \rho) \propto \frac{(1+\rho^2)^c}{\sigma^a(1-\rho^2)^b}. \quad (5.5)$$

Funciones iniciales de referencia

Bajo el supuesto de normalidad asintótica, es fácil identificar la iniciales de referencia de la matriz de información de Fisher,

$$I(\sigma, \rho) = \begin{pmatrix} \frac{4}{\sigma^2} & -\frac{2\rho}{\sigma(1-\rho^2)} \\ -\frac{2\rho}{\sigma(1-\rho^2)} & \frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix}, \quad (5.6)$$

con matriz inversa dada por $H = I^{-1}(\sigma, \rho)$

$$H = \begin{pmatrix} \frac{\sigma^2(1+\rho^2)}{4} & \frac{\sigma\rho(1-\rho^2)}{2} \\ \frac{\sigma\rho(1-\rho^2)}{2} & (1-\rho^2)^2 \end{pmatrix}.$$

Se sigue de Berger y Bernardo (1992a) que las iniciales de referencia conjunta cuando ρ o σ es la cantidad de interés son, respectivamente,

$$\pi_\rho(\rho, \sigma) \propto \frac{1}{\sigma(1-\rho^2)} \quad \text{y} \quad \pi_\sigma(\sigma, \rho) \propto \frac{\sqrt{1+\rho^2}}{\sigma(1-\rho^2)}.$$

Densidades final y predictiva de la inicial π_ρ .

Considérese la inicial de referencia

$$\pi_\rho(\rho, \sigma) \propto \frac{1}{\sigma(1-\rho^2)}$$

donde ρ es la cantidad de interés.

Dadas dos observaciones \mathbf{x}_1 y \mathbf{x}_2 , la densidad final es igual a

$$\pi_\rho(\sigma, \rho | \mathbf{x}_1, \mathbf{x}_2) = \frac{1}{c_\rho} \frac{\sigma^{-5}(1-\rho^2)^{-2}}{(2\pi)^2} \exp \left\{ -\frac{1}{2\sigma^2(1-\rho^2)}(s_{.1} + s_{.2} - 2\rho s_{12}) \right\}$$

donde $c_\rho = \frac{1}{(2\pi)^2} \frac{\Gamma(2)}{2} \frac{1}{s_{12}^2} \frac{2}{g^2-1}$ y $g \equiv g(\mathbf{x}_1, \mathbf{x}_2) = \frac{s_{.1}+s_{.2}}{2s_{12}}$, la cual está definida cuando $|g| > 1$. En el Apéndice C.1 se detalla la obtención de las densidades posterior y predictiva, la función g se detalla en el Apéndice C.3.

La densidad predictiva para una nueva observación $\mathbf{x} = (x_1, x_2)$ es

$$p_\rho(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \frac{\Gamma(3)}{2^4 c_\rho \pi^3} \frac{1}{(s_{12} + x_1 x_2)^3} \frac{\pi|k|}{2k(k^2 - 1)^{3/2}}.$$

donde $k \equiv k(\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2) = \frac{s_{.1}+s_{.2}+x_1^2+x_2^2}{2(s_{12}+x_1 x_2)}$ siempre y cuando $|k| > 1$ (véase el Apéndice C.3).

La probabilidad requerida es

$$\Pr(\mathbf{x} \in C_{\mathbf{x}_1 \mathbf{x}_2} | \mathbf{x}_1, \mathbf{x}_2) = \int_{C_{\mathbf{x}_1 \mathbf{x}_2}} p_\rho(\mathbf{x} | \mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}$$

donde la región de interés la definimos como

$$C_{\mathbf{x}_1 \mathbf{x}_2} = \left\{ \begin{array}{l} (x_1, x_2) : \min\{x_{11}, x_{21}\} < x_1 < \max\{x_{11}, x_{21}\}, \\ \text{y} \quad \min\{x_{21}, x_{22}\} < x_2 < \max\{x_{21}, x_{22}\} \end{array} \right\}.$$

El valor esperado la cobertura de probabilidad anterior se aproximará numéricamente. La Tabla 5.4 (primer renglón) muestra la cobertura de probabilidad de $p_\rho(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2)$ cuando hay independencia entre variables. Adicionalmente, la Tabla 5.5 (primer renglón) muestra la cobertura de probabilidad para diferentes valores de ρ .

Cada una de las tablas contiene una descripción del cálculo de las aproximaciones presentadas junto con una conclusión de los resultados.

Densidades final y predictiva de la inicial π_σ .

Ahora considérese la inicial de referencia siguiente

$$\pi_\sigma(\sigma, \rho) \propto \frac{\sqrt{1 + \rho^2}}{\sigma(1 - \rho^2)}$$

donde σ es la cantidad de interés.

Dadas dos observaciones \mathbf{x}_1 y \mathbf{x}_2 , la densidad posterior resultante es

$$\pi_\sigma(\sigma, \rho | \mathbf{x}_1, \mathbf{x}_2) = \frac{\sigma^{-5} (1 - \rho^2)^{-2} (1 + \rho^2)^{1/2}}{c_\sigma (2\pi)^2} \exp \left\{ -\frac{(s_{.1} + s_{.2} - 2\rho s_{12})}{2\sigma^2 (1 - \rho^2)} \right\}$$

donde la constante de marginalización es (véase el Apéndice C.2)

$$c_\sigma = \frac{2\Gamma(2)}{(2\pi * 2s_{12})^2} \frac{1}{(g^2 - 1) \sqrt{1 + g^2}} \left(\begin{array}{l} 2\sqrt{1 + g^2} (\sqrt{2} + (g^2 - 1) \operatorname{arcsinh}(1)) \\ + g (g^2 - 1) \log \left(\frac{g-1}{1+g} * \frac{1-g+\sqrt{2(1+g^2)}}{1+g+\sqrt{2(1+g^2)}} \right) \end{array} \right),$$

con $g \equiv g(\mathbf{x}_1, \mathbf{x}_2) = \frac{s_{.1} + s_{.2}}{2s_{12}}$ y $|g| > 1$.

La densidad predictiva para una nueva observación $\mathbf{x} = (x_1, x_2)'$ está definida por

$$p_\sigma(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \frac{4\Gamma(3)}{c_\sigma(2\pi)^3} \frac{(5k^5\sqrt{2\pi})^{-1}}{(2(s_{12} + x_1x_2))^3} \\ \times \left(\begin{array}{c} 32\Gamma(\frac{3}{4})\Gamma(\frac{7}{4})\text{Hypergeometric2F1}(\frac{7}{4}, 2, \frac{9}{4}, \frac{1}{k^4}) \\ -5k^2\Gamma(-\frac{3}{4})\Gamma(\frac{5}{4})\text{HypergeometricPFQ}(\{1, \frac{5}{4}, \frac{3}{2}\}, \{\frac{1}{2}, \frac{7}{4}\}, \frac{1}{k^4}) \end{array} \right)$$

con $k \equiv k(\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2) = \frac{s_{.1} + s_{.2} + x_1^2 + x_2^2}{2(s_{12} + x_1x_2)}$ y $|k| > 1$; la función $\text{Hypergeometric2F1}(a, b, c, z)$ tiene expansión en series de sumas

$${}_2F_1(a, b; c; z) = \sum_{k=0}^{\infty} (a)_k (b)_k / (c)_k z^k / k!,$$

y la función $\text{HypergeometricPFQ}(\{a_1, \dots, a_p\}, \{b_1, \dots, b_q\}, z)$ tiene la expansión en series de sumas

$${}_pF_q(a; b; z) = \sum_{k=0}^{\infty} (a_1)_k \dots (a_p)_k / (b_1)_k \dots (b_q)_k z^k / k!.$$

La expresión usual de la probabilidad requerida

$$\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2} | \mathbf{x}_1, \mathbf{x}_2) = \int_{C_{\mathbf{x}_1\mathbf{x}_2}} p_\sigma(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}.$$

El valor esperado de esta probabilidad se aproximará numéricamente y puede verse en la Tabla 5.4 (segundo renglón) el comportamiento de la cobertura de $\Pr(\mathbf{x} \in C_{\mathbf{x}_1\mathbf{x}_2} | \mathbf{x}_1, \mathbf{x}_2)$ cuando las variables son independientes. Para el caso de dependencia entre variables, la Tabla 5.5 (segundo renglón) muestra el valor esperado de la cobertura para diferentes valores de ρ .

Inicial	Forma(5.5)	Prom	DvSt
$\pi_\rho(\rho, \sigma) \propto \frac{1}{\sigma(1-\rho^2)}$	(1, 1, 1/2)	0.1103	0.0117
$\pi_\sigma(\sigma, \rho) \propto \frac{\sqrt{1+\rho^2}}{\sigma(1-\rho^2)}$	(1, 1, 0)	0.1084	0.0114

Tabla 5.4: Valor medio y desviación estándar de la cobertura de probabilidad de la densidades predictivas resultantes de las iniciales π_ρ y π_σ . Los valores deben aproximarse a 3^{-2} pues se asume independencia de las variables.

Los resultados mostrados en la Tabla 5.4 fueron calculados como se describe a continuación:

- (i) Simúlense dos vectores aleatorios $\{\mathbf{x}_1, \mathbf{x}_2\}$ de $N_2(x|\mathbf{0}, \mathbf{I})$.
- (ii) Dada una función inicial $\pi(\sigma, \rho)$, obténgase las densidades final $\pi(\sigma, \rho|\mathbf{x}_1, \mathbf{x}_2)$ y predictiva $p(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2)$ correspondientes.
- (iii) Calcúlese la probabilidad predictiva del área $C_{\mathbf{x}_1\mathbf{x}_2}$, es decir,

$$\int_{C_{\mathbf{x}_1\mathbf{x}_2}} p(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}$$

y almacenarla en memoria.

- (iv) Repítanse los pasos (i)-(iii) n -veces. Obténgase la media muestral y la desviación estándar de las n probabilidades predictivas calculadas.

Nótese que la cobertura de la densidad predictiva resultante de la inicial de referencia π_ρ tiene un mejor comportamiento comparado con la inicial de referencia π_σ , pero la diferencia no es importante.

El tamaño de muestra usado en las Tabla 5.4 es $n = 10^5$. Las integrales en el paso (iii) fueron aproximadas mediante integración Monte Carlo que es una opción de la integración numérica que proporciona Mathematica y solo se muestran cuatro dígitos de precisión.

Función inicial de la regla general de Jeffreys

Siguiendo la regla general Jeffreys,

$$\pi(\sigma, \rho) \propto |\det I_{\sigma, \rho}|^{1/2}$$

donde $I_{\sigma, \rho}$ es la matriz de información de Fisher dada en ec. (5.6)

$$I(\sigma, \rho) = \begin{pmatrix} \frac{4}{\sigma^2} & -\frac{2\rho}{\sigma(1-\rho^2)} \\ -\frac{2\rho}{\sigma(1-\rho^2)} & \frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix},$$

Entonces la inicial es

$$\pi_{JR}(\sigma, \rho) \propto \frac{1}{\sigma(1-\rho^2)}$$

que coincide con la función inicial $\pi_\rho(\sigma, \rho)$ donde ρ es la cantidad de interés. Nótese que para esta función inicial su respectiva densidad predictiva cumple con tener buenas propiedades de cobertura, véase nuevamente los resultados mostrados en las Tablas 5.4 y 5.5, primer renglón.

Otras parametrizaciones

Nuestro objetivo es saber cual es la función inicial cuya correspondiente densidad predictiva tiene adecuadas propiedades de cobertura. Trabajemos ahora con dos grupos de parametrizaciones de los elementos de la matriz de dispersión Σ . Los cuales se analizan en grupos de dos, obteniendo de cada uno de ellos la función inicial de referencia y su respectiva densidad predictiva.

(i) Los valores propios de Σ .

Los valores propios de $\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}$ son

$$\lambda_1 = \sigma^2(1 - \rho) \quad \text{y} \quad \lambda_2 = \sigma^2(1 + \rho).$$

(ii) El conjunto de parametrizaciones $\{\theta_1, \theta_2, \theta_3\}$, donde

$$\theta_1 = \rho, \quad \theta_2 = \sigma^2(1 - \rho^2) \quad \text{y} \quad \theta_3 \equiv |\Sigma| = \sigma^4(1 - \rho^2).$$

Nótese que las parametrizaciones (ii) son transformaciones de uso común en la literatura.

Trabajemos con detalle la inicial de referencia cuando los valores propios son las cantidades de interés, omitiendo el desarrollo de las parametrizaciones restantes por ser

similares.

Dada la matriz de información de Fisher $I(\sigma, \rho)$ en (5.6),

$$I(\sigma, \rho) = \begin{pmatrix} \frac{4}{\sigma^2} & -\frac{2\rho}{\sigma(1-\rho^2)} \\ -\frac{2\rho}{\sigma(1-\rho^2)} & \frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix},$$

y la transformación $\boldsymbol{\psi} = (\lambda_1, \lambda_2) = g(\sigma, \rho)$ con Jacobiano de la transformación inversa dada por

$$J_{g^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} \frac{\partial \sigma}{\partial \lambda_1} & \frac{\partial \sigma}{\partial \lambda_2} \\ \frac{\partial \rho}{\partial \lambda_1} & \frac{\partial \rho}{\partial \lambda_2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2\sqrt{2(\lambda_1+\lambda_2)}} & \frac{1}{2\sqrt{2(\lambda_1+\lambda_2)}} \\ -\frac{2\lambda_2}{(\lambda_1+\lambda_2)^2} & \frac{2\lambda_1}{(\lambda_1+\lambda_2)^2} \end{pmatrix}.$$

Bajo el supuesto de normalidad asintótica bajo transformaciones (Corolario 1 de la Prop. 5.17, Bernardo y Smith, 1994, p. 295) se tiene que la matriz de información de Fisher de la transformación es

$$I_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = J_{g^{-1}}^t I(g^{-1}(\boldsymbol{\psi})) J_{g^{-1}} = \begin{pmatrix} \frac{1}{2\lambda_1^2} & 0 \\ 0 & \frac{1}{2\lambda_2^2} \end{pmatrix}.$$

Puesto que $I_{\boldsymbol{\psi}}$ es una matriz diagonal, las iniciales de referencia para la ordenación (λ_1, λ_2) y (λ_2, λ_1) son iguales a $\frac{1}{\lambda_1 \lambda_2}$, y expresada en términos de los parámetros originales es

$$\pi(\sigma, \rho) = \pi_{\boldsymbol{\lambda}}(\sigma^2(1-\rho), \sigma^2(1+\rho)) |J_g| = \frac{4\sigma^3}{\sigma^4(1-\rho^2)} \propto \frac{1}{\sigma(1-\rho^2)}$$

donde J_g es el Jacobiano de la transformación,

$$J_g = \begin{pmatrix} \frac{\partial \lambda_1}{\partial \sigma} & \frac{\partial \lambda_1}{\partial \rho} \\ \frac{\partial \lambda_2}{\partial \sigma} & \frac{\partial \lambda_2}{\partial \rho} \end{pmatrix} = \begin{pmatrix} 2\sigma(1-\rho) & -\sigma^2 \\ 2\sigma(1+\rho) & \sigma^2 \end{pmatrix}, \quad |J_g| = 4\sigma^3.$$

Nótese que $\pi(\sigma, \rho)$ es igual a la inicial $\pi_{\rho}(\sigma, \rho)$, inicial de referencia cuando ρ es la cantidad de interés, y para la cual su cobertura de probabilidad ha sido estimada (Tablas 5.4 y 5.5).

Finalmente, discutamos las parametrizaciones restantes:

Parametrizaciones (ii) : Si se trabaja con los pares (θ_1, θ_2) y (θ_1, θ_3) . La inicial resultante de cada par coincide con la inicial de referencia cuando ρ es el parámetro de interés en la inicial de referencia, es decir, $\pi_{\rho} \propto \frac{1}{\sigma(1-\rho^2)}$ (en términos de los parámetros originales).

En este caso se sabe que la correspondiente densidad predictiva cuenta con una

buena cobertura de probabilidad como se muestra en las Tablas 5.4 y 5.5.

Adviértase que las iniciales resultantes de los casos (i) y (ii) confirman que no hay una *única* cantidad de interés cuya densidad predictiva tenga una buena cobertura frecuentista.

Cobertura de probabilidad de la densidad predictiva							
Inicial	Forma (5.5)	Prom (Dv.St)			Prom (Dv.St)		
		$\sigma = 1$			$\sigma = 2$		
		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
π_ρ	(1,1,0)	0.1100 (3×10^{-4})	0.1248 (0.0003)	0.1910 (0.0003)	0.1100 (0.0003)	0.1247 (0.0002)	0.1913 (0.0006)
π_σ	(1, 1, $\frac{1}{2}$)	0.1086 (3×10^{-4})	0.1232 (0.0003)	0.1901 (0.0004)	0.1084 (0.0003)	0.1233 (0.0002)	0.1900 (0.0004)
Frecuencias		0.1147	0.1296	0.2056	0.1094	0.1404	0.2176

Tabla 5.5: Valor medio (desviación estándar) de la cobertura de probabilidad de las densidades predictivas provenientes de las iniciales π_ρ y π_σ . Aproximaciones correspondientes de $n = 10$ valores de $m_1 = 100000$ muestras aleatorias de $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$. Frecuencias (en negro) con que una nueva observación se encuentra contenida en el rectángulo formado por dos observaciones previas correspondientes de $m_2 = 10000$ muestras de $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$. La cobertura debe ser cercana a las frecuencias.

Los valores contenidos en la Tabla 5.5 fueron calculados como se describe a continuación:

- (i) Genérese dos puntos aleatorios $\mathbf{x}_{i1}, \mathbf{x}_{i2}$ de $N(\mathbf{x}|\mathbf{0}, \Sigma)$, donde

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

- (ii) Dada una inicial $\pi(\sigma, \rho)$, obténgase la densidad final $\pi(\cdot, \cdot | \mathbf{x}_{i1}, \mathbf{x}_{i2})$ y predictiva $p(\mathbf{x} | \mathbf{x}_{i1}, \mathbf{x}_{i2})$.
- (iii) Calcúlese

$$\Pr(\mathbf{x} \in \mathbf{x}_{i1} \times \mathbf{x}_{i2} | \mathbf{x}_{i1}, \mathbf{x}_{i2}) = \int_{x_{\min 1}}^{x_{\max 1}} \int_{x_{\min 2}}^{x_{\max 2}} p(\mathbf{x} | \mathbf{x}_{i1}, \mathbf{x}_{i2}) d\mathbf{x}$$

$$\text{con } x_{\min 1} = \min\{x_{11}, x_{12}\}, \quad x_{\max 1} = \max\{x_{11}, x_{12}\}, \\ x_{\min 2} = \min\{x_{21}, x_{22}\} \text{ y } x_{\max 2} = \max\{x_{21}, x_{22}\}.$$

Guardar en memoria el resultado.

- (iv) Repítase los pasos (i)-(iii) n -veces y promedie la cobertura de probabilidad con tal de obtener $E_j = \frac{1}{n} \sum_{i=1}^n \Pr(\mathbf{x} \in C_{\mathbf{x}_{i1}\mathbf{x}_{i2}} | \mathbf{x}_{i1}, \mathbf{x}_{i2})$.
- (v) Repítase el paso (iv) m_1 -veces y obtenga la lista $\{E_1, \dots, E_{m_1}\}$. Calcúlese nuevamente la media y desviación estándar de esta muestra.

El objetivo de la Tabla 5.5 es contrastar el valor medio de las coberturas de probabilidad de las densidades predictivas y la frecuencia con que una nueva observación se encuentra contenida en el rectángulo formado por dos observaciones previas. De acuerdo a las aproximaciones numéricas, puede verse que ambas iniciales π_σ y π_ρ producen densidades predictivas con cobertura frecuentista adecuada.

Los valores utilizados para aproximar las coberturas de las densidades predictivas son $n = 10$ valores de $m_1 = 100000$ muestras simuladas de $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$ con

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Las frecuencias (negro) corresponden a $m_2 = 10000$ muestras simuladas de $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$, con Σ definida anteriormente.

5.2.3. Caso todos los parámetros desconocidos

Como consecuencia natural de los casos anteriores estudiados, consideremos ahora los siguientes casos:

- (i) $\{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$ desconocidos. Puesto que la función inicial objetiva de los parámetros de localización es constante, el problema se reduce a estudiar sólo $\{\sigma_1, \sigma_2, \rho\}$.
- (ii) $\{\sigma_1, \sigma_2, \rho\}$ desconocidos.

Para ambos casos, puede utilizarse la familia de iniciales objetivas de la forma

$$\pi_{abc} \propto \frac{1}{\sigma_1^a \sigma_2^a (1 - \rho^2)^b}.$$

Se sabe que en los modelos multivariados no es obvia la definición de la mediana. Parecería natural definir la mediana del modelo como el vector de medianas, análogo al vector de medias. Debido a esta dificultad, optaremos por elegir la densidad inicial objetiva en función de las propiedades de cobertura de la densidad predictiva que implica.

$\{\sigma_1, \sigma_2, \rho\}$ desconocidos.

La elección de la función (o funciones) inicial objetiva es similar a la hecha en las dos subsecciones anteriores:

Dados dos vectores aleatorios \mathbf{x}_1 y \mathbf{x}_2 de $N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, encontrar los valores a, b y c tales que la cobertura de la densidad predictiva sea

$$\Pr_{abc}(\mathbf{x} \in C_{\mathbf{x}_1 \mathbf{x}_2} | \mathbf{x}_1, \mathbf{x}_2) \approx \frac{1}{3},$$

donde $C_{\mathbf{x}_1 \mathbf{x}_2}$ es el rectángulo cuyos vértices son los elementos de \mathbf{x}_1 y \mathbf{x}_2 .

$\{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$ desconocidos

Cuando las varianzas son distintas, se requieren de al menos tres observaciones aleatorias para que la densidad final sea propia. Por tanto, la elección de la función (o funciones) inicial objetiva adecuada para predicción se propone trabajar en la línea siguiente:

- (i) **Resultado frecuentista.** Dado que se requieren tres observaciones para asegurar que la densidad final sea propia, deberá calcularse la probabilidad de que una nueva observación se encuentre contenida en el menor rectángulo formado que contiene a las tres observaciones obtenidas. Este resultado ayudará a discriminar entre las posibles funciones iniciales la que es adecuada para predicción.
- (ii) **Estudio Bayesiano.** Utilizando la familia de funciones iniciales objetivas π_{abc} . Encontrar los valores de a, b y c para los cuales la densidad inicial produce una densidad predictiva con adecuadas propiedades de cobertura de probabilidad, es decir, la cobertura de la probabilidad predictiva en el rectángulo formado que contiene a las tres observaciones dadas deberá acercarse al valor frecuentista obtenido en (i).

Estas líneas de investigación se encuentran abiertas y son discutidas ampliamente en la Sección de conclusiones y líneas futuras de investigación (pág. 137).

5.3. Predicción de una función de observaciones futuras

Vamos a estudiar la predicción de una función univariada de dos observaciones binormales suponiendo que todos los parámetros de Σ ($\{\sigma_1, \sigma_2, \rho\}$) son desconocidos con σ_1 y σ_2 no necesariamente iguales.

Empezaremos mencionando algunas propiedades del modelo normal bivariado:

Dados dos vectores aleatorios \mathbf{x}_1 y \mathbf{x}_2 de $N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ con

$$\boldsymbol{\mu} = (\mu_1, \mu_2)' \quad \text{y} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}.$$

El vector aleatorio $\mathbf{x}_1 - \mathbf{x}_2 = (x_{11} - x_{21}, x_{12} - x_{22})$ tiene distribución de probabilidad $N_2(\mathbf{0}, 2\Sigma)$. Supóngase que nos interesa predecir el valor del cociente

$$y = \frac{x_{11} - x_{12}}{x_{12} - x_{22}}.$$

Hinkley (1969) desarrolló la expresión de la función de densidad de esta nueva variable aleatoria unidimensional cuando el vector de medias es $\boldsymbol{\mu} = (0, 0)$. El problema también lo abordan Pham-Gia, Turkkan y Marchand (2006) con los resultados que se resumen a continuación.

A . Cociente de dos variables normales independientes.

Sean $x_1 \sim N(\mu_2, \sigma_1)$ y $x_2 \sim N(\mu_2, \sigma_2)$ dos variables aleatorias independientes cuyo cociente $y = x_1/x_2$ tiene función de densidad dado por

$$f(y|\mu_1, \mu_2, \sigma_1, \sigma_2) = \frac{K_1}{\sigma_2^2 y^2 + \sigma_1^2} {}_1F_1(1; 1/2; \theta_1(y)), \quad -\infty < y < \infty$$

donde

$$\theta_1(y) = \frac{1}{2\sigma_1^2\sigma_2^2} \frac{(\sigma_2^2\mu_1 y + \mu_2\sigma_1^2)^2}{\sigma_2^2 y^2 + \sigma_1^2} \geq 0, \text{ y } K_1 = \frac{\sigma_1\sigma_2}{\pi} \exp \left\{ -\frac{1}{2} \left(\frac{\mu_1^2}{\sigma_1^2} + \frac{\mu_2^2}{\sigma_2^2} \right) \right\}.$$

Si $\mu_1 = \mu_2 = 0$ la función de densidad se reduce a una densidad Cauchy

$f(y|\sigma_1, \sigma_2) = \frac{\sigma_1\sigma_2}{\pi(\sigma_2^2 y^2 + \sigma_1^2)}$, $-\infty < y < \infty$, y si $\sigma_1 = \sigma_2$ corresponde a la densidad Cauchy estándar $f(y) = 1/\pi(y^2 + 1)$.

B . Cociente de dos variables normales dependientes.

Sea el vector aleatorio $\mathbf{x} = (x_1, x_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, con

$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$, y $\rho \neq 0$. El cociente $y = x_1/x_2$ tiene función de densidad dada por

$$f(y|\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = K_2 \frac{2(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_2^2 y^2 - 2\rho\sigma_1\sigma_2 y + \sigma_1^2} {}_1F_1(1; 1/2; \theta_2(y))$$

donde $\theta_2(y) = \frac{(-\sigma_2^2\mu_1 y + \rho\sigma_1\sigma_2(\mu_2 y + \mu_1) - \mu_2\sigma_1^2)^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)(\sigma_2^2 y^2 - 2\rho\sigma_1\sigma_2 y + \sigma_1^2)} \geq 0$ y

$$K_2 = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{\sigma_2^2\mu_1^2 - 2\rho\sigma_1\sigma_2\mu_1\mu_2 + \mu_2^2\sigma_1^2}{2(1-\rho^2)\sigma_1^2\sigma_2^2} \right\},$$

y ${}_1F_1$ definida por ${}_1F_1(\alpha, \gamma; z) = \sum_{k=0}^{\infty} \frac{(\alpha, k)}{(\gamma, k)} \cdot \frac{z^k}{k!}$, $\gamma \neq 0, -1, -2, \dots$ con los factoriales ascendentes (α, k) , o coeficientes de Pochhammer, definidos por $(\alpha, k) = \alpha(\alpha+1)\dots(\alpha+k-1) = \Gamma(\alpha+k)/\Gamma(\alpha)$ con $(\alpha, 0) = 1$.

Si $\mu_1 = \mu_2 = 0$ y $\sigma_1, \sigma_2 \neq 1$, se reduce a la distribución Cauchy $\text{Ca}(y|\xi, \nu)$, con $\xi = \rho\frac{\sigma_1}{\sigma_2}$ y $\nu = \frac{\sigma_1}{\sigma_2}\sqrt{1-\rho^2}$; y si $\sigma_1 = \sigma_2 = 1$ (o iguales) la densidad de probabilidad

se reduce a $f(y|\rho) = \frac{\sqrt{1-\rho^2}}{\pi(1-2\rho y+y^2)}$.

Debido a que suponemos $\boldsymbol{\mu} = \mathbf{0}$, la función de densidad de la variable aleatoria y es

$$f(y|\lambda, \rho) = \frac{1}{\pi} \frac{\lambda\sqrt{1-\rho^2}}{\lambda^2 - 2y\rho\lambda + y^2}, \quad (5.7)$$

$$-\infty < y < \infty, \lambda > 0, \rho \in (-1, 1)$$

donde $\lambda = \frac{\sigma_1}{\sigma_2}$. La Ec.(5.7) es una densidad Cauchy, $\text{Ca}(y|\lambda\rho, \lambda\sqrt{1-\rho^2})$, perteneciente a la familia de densidades de localización y escala.

A lo largo de esta subsección estudiamos los siguientes casos para $\lambda = \frac{\sigma_1}{\sigma_2}$ y ρ :

- (i) Caso $\lambda \neq 1$ es desconocida y ρ conocido, el modelo se reduce a uno de un solo parámetro.
- (ii) Caso $\sigma_1 \neq \sigma_2$ y además desconocidos (λ es desconocida), entonces el modelo mantiene ambos parámetros $f(y|\lambda, \rho)$.

Una vez definidos los posibles casos, se procede a obtener las respectivas iniciales de referencias y sus densidades predictivas. Recuérdese que para el problema de predicción, elegiremos la función inicial cuya densidad predictiva cuente con adecuadas propiedades de cobertura.

Primero comprobemos que el modelo cumple con la propiedad frecuentista de que una nueva observación pertenece al intervalo formado por dos observaciones previas. La Tabla 5.6 muestra $m = 10$ frecuencias correspondientes de $n = 100000$ muestras aleatorias de tamaño dos de $\text{Ca}(y|\lambda\rho, \lambda\sqrt{1-\rho^2})$ para diferentes valores de λ y ρ . Nótese que los resultados concuerdan con el Teorema 9 pues se aproximan al valor teórico de $\frac{1}{3}$.

λ	$\rho = -0.9$	$\rho = -0.5$	$\rho = -0.1$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
0.5	0.3331 (0.0013)	0.3332 (0.0014)	0.3332 (0.0013)	0.3331 (0.0013)	0.3332 (0.0020)	0.3332 (0.0011)
1	0.3337 (0.0008)	0.3331 (0.0017)	0.3335 (0.0009)	0.3340 (0.0017)	0.3334 (0.0013)	0.3330 (0.0014)
5	0.3337 (0.0012)	0.3323 (0.0012)	0.3338 (0.0020)	0.3343 (0.0010)	0.3327 (0.0010)	0.3331 (0.0013)

Tabla 5.6: Frecuencia con que una nueva observación pertenece al intervalo formado por dos observaciones previas. Valor medio (desviación estándar) de $m = 10$ frecuencias correspondientes de $n = 100000$ muestras de tamaño dos de $\text{Ca}(\cdot|\lambda\rho, \lambda\sqrt{1-\rho^2})$. Nótese que el valor promedio se aproxima al valor teórico $\frac{1}{3}$.

Caso λ desconocido y ρ conocido.

Este caso es un modelo de un solo parámetro y suponiendo normalidad asintótica de la densidad final de λ , la función inicial de referencia es

$$\pi(\lambda) \propto \sqrt{i(\lambda)} = \sqrt{\frac{1}{2\lambda^2(1-\rho^2)}} \propto \frac{1}{\lambda}.$$

donde $i(\lambda) = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \lambda} \log f(y|\lambda) dy$.

Sean dos observaciones y_1 y y_2 , la densidad final será proporcional a

$$\pi(\lambda|y_1, y_2) \propto \frac{\lambda}{\pi^2(y_1^2 + \lambda^2 - 2y_1\lambda\rho)(y_2^2 + \lambda^2 - 2y_2\lambda\rho)}$$

y la constante de marginalización dada por

$$\begin{aligned} c &= \int_0^{\infty} \frac{\lambda}{\pi^2(y_1^2 + \lambda^2 - 2y_1\lambda\rho)(y_2^2 + \lambda^2 - 2y_2\lambda\rho)} d\lambda \\ &= \frac{(y_1 + y_2)\sqrt{1-\rho^2} \log \left[\frac{y_1^2}{y_2^2} \right] - 2(y_2 - y_1)\rho(\pi + 2 \arcsin \rho)}{2\pi^2(y_1 - y_2)\sqrt{1-\rho^2}((y_1 + y_2)^2 - 4y_1y_2\rho^2)}. \end{aligned}$$

La densidad predictiva es entonces

$$p(y|y_1, y_2) = \int_0^{\infty} f(y|\lambda)\pi(\lambda|y_1, y_2)d\lambda.$$

La probabilidad requerida, bajo el supuesto de $y_1 < y_2$, es

$$\Pr(y_1 < y < y_2|y_1, y_2) = \int_{y_1}^{y_2} p(y|y_1, y_2)dy.$$

El valor esperado en el muestreo de esta probabilidad será aproximado numéricamente y se espera que se aproxime al valor teórico de $\frac{1}{3}$, como se comprueba en la Tabla 5.7.

A la luz de los resultados numéricos puede concluirse que la función inicial λ^{-1} es la adecuada para predicción en este caso.

$\lambda \backslash \rho$	$\frac{1}{10}$	$\frac{1}{2}$	$\frac{9}{10}$
1/2	0.3331 (0.0017)	0.3323 (0.0011)	0.3336 (0.0013)
1	0.3340 (0.0016)	0.3318 (0.0011)	0.3335 (0.0015)
5	0.3322 (0.0010)	0.3322 (0.0013)	0.3322 (0.0012)

Tabla 5.7: Valor medio (desviación estándar) de la cobertura de probabilidad de la densidad predictiva resultante de la función inicial $\pi(\lambda) \propto \lambda^{-1}$. Aproximaciones correspondientes de $n = 10000$ muestras aleatorias de tamaño dos para diferentes valores de λ y tres valores fijos de ρ . Nótese que la media muestral se aproxima a $\frac{1}{3}$.

Parámetros de escala desconocidos y distintos (λ desconocida)

Si ambos parámetros de escala son iguales pero desconocidos entonces $\lambda = 1$, este caso ha sido estudiado en la subsección anterior. Veamos entonces el caso cuando los parámetros de escala son distintos y por tanto λ es desconocida.

El valor de la mediana en este modelo es $\lambda\rho$ y puede comprobarse muy fácilmente que la probabilidad acumulada hasta este punto es un medio,

$$\int_{-\infty}^{\lambda\rho} \frac{1}{\pi} \frac{\lambda\sqrt{1-\rho^2}}{\lambda^2 - 2y\rho\lambda + y^2} dy = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y - \lambda\rho}{\lambda\sqrt{1-\rho^2}}\right) \Big|_{-\infty}^{\lambda\rho} = \frac{1}{2}.$$

Puesto que nos interesa la mediana, usemos la siguiente parametrización

$$\theta = \lambda\rho \quad y \quad \nu = \lambda\sqrt{1-\rho^2},$$

de esta manera se tiene una densidad de probabilidad $\text{Ca}(\cdot|\theta, \nu)$ (modelo de localización-escala).

Se sabe que la función inicial de referencia en este tipo de modelos, cuando la escala o la localización es la cantidad de interés, es ν^{-1} .

El modelo Cauchy ya ha sido estudiado en la Sección 4.2 de esta Tesis y se ha demostrado que la cobertura de probabilidad de la densidad predictiva resultante de la inicial anterior es igual a un tercio. Dado que tenemos un caso similar, es de esperarse que la densidad predictiva cuente con adecuadas propiedades de cobertura.

La inicial $\pi_\theta(\theta, \nu) \propto \nu^{-1}$ expresada en los parámetros originales es

$$\begin{aligned} \pi_\theta(\lambda, \rho) &= \pi_\theta(\theta^{-1}(\lambda, \rho), \nu^{-1}(\lambda, \rho)) |J| \\ &= \frac{1}{1-\rho^2} \end{aligned}$$

con

$$J = \begin{pmatrix} \frac{\partial \theta}{\partial \lambda} & \frac{\partial \theta}{\partial \rho} \\ \frac{\partial \nu}{\partial \lambda} & \frac{\partial \nu}{\partial \rho} \end{pmatrix} = \begin{pmatrix} \rho & \lambda \\ \sqrt{1-\rho^2} & -\frac{\lambda}{\sqrt{1-\rho^2}} \end{pmatrix} \quad \text{y} \quad |J| = \frac{\lambda}{\sqrt{1-\rho^2}}.$$

Dada $\pi_\theta(\alpha, \rho)$, la densidad final dadas dos observaciones tiene la siguiente forma

$$\pi_\theta(\alpha, \rho | y_1, y_2) = \frac{1}{c} \frac{\lambda^2}{(y_1^2 + \lambda^2 - 2y_1\lambda\rho)(y_2^2 + \lambda^2 - 2y_2\lambda\rho)}, \quad y_1 \neq y_2$$

con constante de marginalización igual a

$$c = \int_0^\infty \frac{\lambda}{2\pi^2(y_1 - y_2)(y_1y_2 - \lambda^2)} \log \left(\frac{(y_1 - \lambda)^2(y_2 + \lambda)^2}{(y_2 - \lambda)^2(y_1 + \lambda)^2} \right) d\lambda$$

con $\lambda \neq |y_1|, |y_2|$, y $\lambda^2 \neq y_1y_2$ (ρ ya ha sido integrado).

Consecuentemente, la densidad predictiva es,

$$p(y | y_1, y_2) = \frac{1}{c} \int_0^\infty \int_{-1}^1 \frac{\lambda^3 \sqrt{1-\rho^2} d\rho d\lambda}{\pi^3(y^2 + \lambda^2 - 2y\lambda\rho)(y_1^2 + \lambda^2 - 2y_1\lambda\rho)(y_2^2 + \lambda^2 - 2y_2\lambda\rho)}$$

de la que no se tiene una expresión analítica explícita.

La probabilidad requerida, suponiendo $y_1 < y_2$, es

$$\Pr(y_1 < y < y_2 | y_1, y_2) = \int_{y_1}^{y_2} p(y | y_1, y_2) dy.$$

Como se ha procedido anteriormente, el valor esperado en el muestreo de esta cobertura será aproximado numéricamente,

$$E_{y_1, y_2 | \lambda \rho} (\Pr(y_1 < y < y_2 | y_1, y_2) | y_1, y_2) \approx \frac{1}{n} \sum_{i=1}^n \Pr(y_{1i} < y < y_{2i} | y_{1i}, y_{2i}).$$

La Tabla 5.8 muestra las aproximaciones correspondientes de $n = 100$ muestras aleatorias de tamaño dos para diferentes valores de λ y ρ . Nótese que la desviación estándar es cercana a cero, lo que nos dice que el valor promedio debiera ser exacto a $1/3$ pero no se alcanza por los redondeos inherentes de la precisión numérica del ordenador.

Puede concluirse que una función inicial para predicción en el modelo $f(y | \lambda, \rho)$ es

$$\pi(\lambda, \rho) \propto \frac{1}{1 - \rho^2}$$

que es la correspondiente de usar a la mediana como cantidad de interés.

$\lambda \backslash \rho$	$\frac{1}{10}$	$\frac{1}{2}$	$\frac{9}{10}$
1/2	0.3330 (≈ 0)	0.3330 (≈ 0)	0.3328 (3×10^{-6})
1	0.3330 (≈ 0)	0.3328 (4×10^{-6})	0.3325 (1×10^{-5})
5	0.3329 (1×10^{-6})	0.3329 (≈ 0)	0.3323 (9×10^{-6})

Tabla 5.8: Valor medio (desviación estándar) de la cobertura de probabilidad de la densidad predictiva $p(\cdot|y_1, y_2)$ correspondiente de $n = 100$ muestras aleatorias de tamaño dos. El valor medio se aproxima a un tercio.

La siguiente subsección presenta el caso cuando $\lambda \neq 1$ desconocida y ρ conocido.

Conclusión del modelo $f(\cdot|\lambda\rho, \lambda\sqrt{1-\rho^2})$

Se analiza el problema de encontrar una inicial de referencia para predicción de una función de dos observaciones binormales. La densidad de esta nueva función una densidad Cauchy de la que se sabe que la inicial de referencia de la mediana (o la moda) produce una densidad predictiva con cobertura frecuentista exacta (Sección 4.2).

Catálogo de iniciales para predicción

Iniciales objetivas para predicción bajo hipótesis de cobertura		
Modelo	Parámetro/cantidad de interés	Función inicial para predicción
$\text{Ex}(x \theta)$	θ	$\pi(\theta) \propto \frac{1}{\theta}$
$\text{Ca}(x \mu, \sigma)$	μ Mediana y moda	$\pi(\mu, \sigma) \propto \frac{1}{\sigma}$
$\text{N}(x \mu, \sigma)$	μ Media, mediana y moda	$\pi(\mu, \sigma) \propto \frac{1}{\sigma}$
$\text{LgN}(x \mu, \sigma)$	$\phi_3 = e^\mu$ Mediana	$\pi_{\phi_3}(\mu, \sigma) \propto \frac{1}{\sigma}$
$\text{Un}(0, \theta)$	θ	$\pi(\theta) \propto \frac{1}{\theta}$
$\text{Un}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$	θ	$\pi(\theta) \propto \text{Un}(x_{\max} - \frac{1}{2}, x_{\min} + \frac{1}{2})$
$\text{Ex}(x \alpha, \varphi)$		
φ conocido	α	$\pi(\alpha) \propto 1$
α, φ desconocidos	$\phi = \alpha + \varphi \log 2$	$\pi_\phi(\alpha, \varphi) \propto \frac{1}{\varphi}$

Tabla 5.9: Funciones iniciales de referencia para predicción bajo hipótesis de cobertura

Iniciales objetivas para predicción para $N_2(\mathbf{x} \mathbf{0}, \Sigma)$			Desempeño
Supuestos	Parámetro/cantidad de interés	Función inicial para predicción	
ρ desconocido $\sigma_1, \sigma_2 = 1$ conocidos	ρ	$\pi_\rho(\rho) \propto \frac{\sqrt{1+\rho^2}}{1-\rho^2}$	Bueno
	$\varphi = \text{arctanh}\rho$	$\pi_\varphi(\rho) \propto \frac{1}{1-\rho^2}$	Bueno
ρ, σ desconocidos $\sigma_1 = \sigma_2 = \sigma$	ρ	$\pi_\rho(\rho, \sigma) \propto \frac{1}{\sigma(1-\rho^2)}$ Inicial de la regla general de Jeffreys	Bueno
	σ	$\pi_\sigma(\rho, \sigma) \propto \frac{\sqrt{1+\rho^2}}{\sigma(1-\rho^2)}$	Bueno
σ_1, σ_2 y ρ desconocidos			
Cociente de normales $\text{Ca}(y \lambda\rho, \lambda\sqrt{1-\rho^2})$, con $\lambda = \sigma_1/\sigma_2$ y ρ			
ρ conocido	λ	$\pi(\lambda) \propto \frac{1}{\lambda}$	Excelente (casi exacto)
ρ y σ desconocidos	$\theta = \lambda\rho$ (con $\nu = \lambda\sqrt{1-\rho^2}$)	$\pi_\theta(\lambda, \rho) \propto \frac{1}{1-\rho^2}$	Excelente (Casi exacto)

Tabla 5.10: Iniciales objetivas para predicción para el modelo $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$, con $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$, desempeño de la cobertura de probabilidad predictiva resultantes de diferentes funciones iniciales objetivas.

Conclusiones y líneas de investigación futuras

Conclusiones

1. Se ha visto que no existe una propuesta general de obtención de iniciales objetivas para predicción. Nuestra propuesta es encontrar una inicial objetiva para predicción mediante el análisis de referencia, basada en alguna función de los datos como cantidad de interés. En particular, para los modelos continuos univariados sugerimos elegir la mediana de los datos, cuando exista o una aproximación analítica a ella, como cantidad de interés, puesto que a la luz de los modelos estudiados, la densidad predictiva correspondiente tiene propiedades de cobertura adecuadas.
2. Cuando se trabaja con modelos no regulares de la forma descrita en las Secciones 4.5 y 4.6, el método de obtención de iniciales de referencia propuesto por Bernardo (1979, 2005a), Berger y Bernardo (1992a) y Berger *et al.* (2009a) produce una inicial de referencia cuya densidad predictiva cuenta con adecuadas propiedades de cobertura. Recuérdese que en el modelo Exponencial con dos parámetros, fue la aproximación numérica de la inicial de referencia de la mediana, como cantidad de interés, la que produjo una densidad predictiva con cobertura de probabilidad exacta.
3. Al inicio del Capítulo 1 se generaliza el resultado de Jeffreys para predicción en el modelo de localización y escala. Jeffreys propuso la inicial para predicción en este modelo y demostró que la probabilidad de tener una nueva observación contenida en el intervalo formado por dos observaciones previas es un tercio. Nosotros hemos extendido este caso a más de dos observaciones encontrando que la probabilidad de que una nueva observación se encuentra contenida en el intervalo formado por las observaciones mínima y máxima es $(n - 1)/(n + 1)$ con n el tamaño de muestra.
4. En casos multivariados, no es claro como proceder, pues la mediana no está definida. En esta situación la propuesta es usar las propiedades de cobertura de la densidad predictiva para elegir entre las posibles iniciales alternativas.

Líneas de investigación futuras

1. Una línea de investigación es la predicción cuando se tiene un modelo de probabilidad discreto. Se sabe que los teoremas de cubrimiento sólo se aplican a variables continuas, y en el límite, debería ser aproximadamente cierto con modelos discretos para muestras grandes. Se propone estudiar desde esta perspectiva el modelo binomial.
2. Dentro del capítulo 5 se trabaja con detalle el caso del modelo normal Bivariado con vector de medias conocido y elementos de la matriz de dispersión desconocidos pero varianzas iguales. Falta considerar los casos donde $\{\sigma_1, \sigma_2, \rho\}$ ($\sigma_1 \neq \sigma_2$) son desconocidos y el caso general donde todos los parámetros $\{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$ son desconocidos. La propuesta de obtención de la función inicial objetiva adecuada para predicción en el primer caso ya se ha discutido brevemente en la Sección 5.2.3, por tanto, centraremos nuestra discusión en el segundo caso. Como es sabido, en los modelos multidimensionales no es claro que valor pudiera tener la mediana; teniendo un vector de medias parecería natural definir un vector de medianas. Nuestra propuesta es analizar una colección de funciones iniciales objetivas y discriminar entre ellas de acuerdo a las propiedades de cobertura de sus respectivas densidades predictivas. Sugerimos las siguientes funciones iniciales:

- **Iniciales de referencia** Nótese que la matriz de información de Fisher no depende de $\{\mu_1, \mu_2\}$ y el lugar que tomen en cualquier ordenación no afecta la función inicial resultante. Entonces, sólo se consideran las ordenaciones de los tres parámetros restantes, que son los siguientes casos:

- Orden $\{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$ y $\{\sigma_1, \sigma_2, \rho, \mu_1, \mu_2\}$, la inicial es

$$\pi_{R_\sigma}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \frac{\sqrt{1 + \rho^2}}{\sigma_1 \sigma_2 (1 - \rho^2)}.$$

- Orden $\{\mu_1, \mu_2, \sigma_1, \rho, \sigma_2\}$ y $\{\sigma_1, \rho, \sigma_2, \mu_1, \mu_2\}$, la función inicial correspondiente es

$$\pi_{R_{\sigma^*}}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)} \sqrt{\frac{2}{2 - \rho^2}}.$$

- Orden $\{\mu_1, \mu_2, \rho, \sigma_1, \sigma_2\}$ y $\{\rho, \sigma_1, \sigma_2, \mu_1, \mu_2\}$, la función inicial correspondiente es

$$\pi_{R_\rho}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)}. \quad (5.8)$$

Estas funciones iniciales son obtenidas siguiendo el algoritmo dado en Berger y Bernardo (1992a) y estudiadas extensamente en Berger y Sun (2006, 2008). La función inicial π_{R_p} (5.8) fue propuesta por Lindley (1965) y es ampliamente estudiada en Bayarri (1981).

- Función inicial de la regla general de Jeffreys.

La función inicial obtenida de aplicar la regla general de Jeffreys es

$$\begin{aligned}\pi_J(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &\propto \sqrt{\det I_{\theta}} \\ &\propto \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)^2}.\end{aligned}$$

- Función inicial para predicción propuesta por Geisser (1993).

Geisser (1993, cap. 9) propone una función inicial para predicción en el modelo $N_d(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ que es $\pi_G(\boldsymbol{\mu}, \Sigma^{-1}) \propto |\Sigma|^{-(d+1)/2}$. En nuestro ($d = 2$) caso la función inicial es

$$\pi_G(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto |\Sigma|^{-(d+1)/2} = \frac{1}{\sigma_1^3 \sigma_2^3 (1 - \rho^2)^{3/2}}$$

Las densidades finales basadas en las funciones iniciales anteriores son propias cuando se tienen $n \geq 3$ observaciones. Por tanto encontrar un resultado similar al del Teorema 10 junto con el resultado obtenido en la sección 5.1 pero extenderlo a tres observaciones. Es decir, dadas tres observaciones $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ provenientes de $N_2(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, y una nueva observación $\mathbf{x} = \{x_1, x_2\}$, la probabilidad

$$\Pr(\mathbf{x} \in \mathbf{x}_{\min} \times \mathbf{x}_{\max} | \boldsymbol{\mu}, \Sigma) = \int_{\mathbf{x}_{\min} \times \mathbf{x}_{\max}} f(\mathbf{x} | \boldsymbol{\mu}, \Sigma) d\mathbf{x} \approx \frac{1}{2^2} h(\rho)$$

donde

$$\mathbf{x}_{\min} \times \mathbf{x}_{\max} = \{(x_1, x_2) : x_{\min 1} < x_1 < x_{\max 1}, x_{\min 2} < x_2 < x_{\max 2}\},$$

con $x_{\min} = \min\{x_{11}, x_{21}, x_{31}\}$, $x_{\max} = \max\{x_{12}, x_{22}, x_{32}\}$. En otras palabras, la probabilidad de que una nueva observación \mathbf{x} se encuentre contenido en el rectángulo formado por las coordenadas mínima y máxima de $\mathbf{x}_1 = (x_{11}, x_{12})$, $\mathbf{x}_2 = (x_{21}, x_{22})$, $\mathbf{x}_3 = (x_{31}, x_{32})$, y esto *debería* de ser cercano a $\frac{h(\rho)}{2^2}$. Se espera que las coberturas de las densidades resultantes de cada una de las iniciales alternativas debe aproximarse al resultado frecuentista anterior.

3. Modelo del cociente de variables normales. Como ejemplo de predicción de

una función de dos observaciones binormales se estudia el problema del cociente de sus diferencias que se reduce a un problema de una sola variable cuya densidad es $\text{Ca}(y|\lambda\rho, \lambda\sqrt{1-\rho^2})$.

Bajo el supuesto de que se conoce el parámetro λ el modelo se reduce a uno de un solo parámetro, del cual es fácil obtener su inicial de referencia. Para dos observaciones sus densidades final y predictiva están bien definidas y se debe aproximar el valor esperado de la cobertura $\Pr(y_1 < y < y_2|y_1, y_2)$ con $y_1 < y_2$. Debido a su lenta convergencia hacia el valor teórico $\frac{1}{3}$ es necesario mejorar el algoritmo para encontrar resultados concluyentes.

4. **Regresión lineal simple.** El modelo de regresión lineal simple donde $y_i \sim N(\alpha + \beta x, \sigma)$ con mediana $E(y|x) = \alpha + \beta x$. Si definimos como cantidad de interés a $\phi = \alpha + \beta x$, debemos obtener su inicial correspondiente y dadas dos observaciones, comprobar si la cobertura de probabilidad de la densidad predictiva resultante $\Pr(y_1 < y < y_2|y_1, y_2)$ se aproxima al valor teórico $\frac{1}{3}$ para $y_1 < y_2$ con $y_1, y_2 \sim N(\alpha + \beta x, \sigma)$.
5. **Right Haar.** El trabajo de Severini, Mukerjee y Ghosh (2002) exploran la propiedad de cobertura consistente de regiones Bayesianas invariantes, es decir, regiones Bayesianas resultantes de usar una inicial invariante. Trabajan con la inicial de Haar por la derecha y estudian tres interesantes ejemplos, uno de ellos es el del análisis de varianza. Otra línea de investigación que proponemos es probar las propiedades de cobertura frecuentista de la densidad predictiva resultante de usar esta inicial en el análisis de varianza.

Apéndice A

Modelo Log-Normal

La densidad final no es propia cuando la media (moda) es la cantidad de interés y sólo se tienen $n = 2$ observaciones.

A continuación demostraremos que la densidad final de referencia cuando la media (o la moda) es la cantidad de interés no es propia en el caso de tener solo dos observaciones.

Trabajaremos con detalle el caso de la media y se omitirá el desarrollo de la moda por ser similar. Para simplificar la exposición se usará la notación ϕ en lugar de ϕ_1 .

Considérese dos observaciones aleatorias x y y con distribución Log Normal $\text{LgN}(\cdot|\mu, \sigma)$ y sea la cantidad de interés $\phi(\mu, \sigma) = E(x|\mu, \sigma) = \exp(\mu + \frac{\sigma^2}{2})$ con inicial de referencia dada por $\pi_\phi(\mu, \sigma) \propto \sqrt{1 + \frac{2}{\sigma^2}}$.

La densidad final es proporcional a

$$\pi_\phi(\mu, \sigma|x, y) \propto \frac{1}{xy2\pi} \sqrt{1 + \frac{2}{\sigma^2}} \frac{1}{\sigma^2} \exp \left\{ -\frac{1}{2\sigma^2} ((\log x - \mu)^2 + (\log y - \mu)^2) \right\},$$

y se desea verificar si es propia o no. Sea c la constante de marginalización

$$c = \frac{1}{xy2\pi} \int_{-\infty}^{\infty} \frac{\sqrt{1 + \frac{2}{\sigma^2}}}{\sigma^2} \int_0^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} ((\log x - \mu)^2 + (\log y - \mu)^2) \right\} d\mu d\sigma$$

que debe integrarse sobre los dominios de μ y σ .

Nótese que el término del exponente es de la forma $A(x - \alpha)^2 + B(x - \beta)^2$ al cual se le puede aplicar la siguiente identidad

$$A(x - \alpha)^2 + B(x - \beta)^2 = (A + B)(x - m)^2 + \frac{AB}{A + B}(\alpha - \beta)^2,$$

donde $m = \frac{a\alpha + b\beta}{A + B}$.

Entonces, la constante de marginalización se reduce a

$$c = \frac{1}{xy2\pi} \int_0^\infty \frac{1}{\sigma^2} \sqrt{1 + \frac{2}{\sigma^2}} \exp \left\{ -\frac{1}{4\sigma^2} \left(\log \frac{x}{y} \right)^2 \right\} \int_{-\infty}^\infty e^{-\frac{(\mu-m)^2}{\sigma^2}} d\mu d\sigma$$

con $m = \log \sqrt{xy}$.

Observéese que el integrando del extremo derecho corresponde al *kernel* de una densidad normal para μ , *i.e.*, $\mu \sim N(\mu|m, \sigma/\sqrt{2})$. Por lo que la constante se reduce a

$$\begin{aligned} c &= \frac{1}{xy2\sqrt{\pi}} \int_0^\infty \frac{1}{\sigma} \sqrt{1 + \frac{2}{\sigma^2}} e^{-\frac{1}{4\sigma^2} (\log \frac{x}{y})^2} d\sigma \\ &= \frac{1}{xy2\sqrt{\pi}} \int_0^\infty \frac{\sqrt{2 + \sigma^2}}{\sigma^2} e^{-\frac{1}{4\sigma^2} (\log \frac{x}{y})^2} d\sigma. \end{aligned}$$

Nótese nuevamente, que el integrando es el *kernel* de la *raíz cuadrada* de la densidad *Gama inversa* (squared-root inverted-Gamma) para σ , es decir, $\sigma \sim \text{Ga}^{-1/2} \left(\sigma | \frac{1}{2}, \frac{1}{4} \log^2 \left(\frac{x}{y} \right) \right)$. Entonces, la constante es ahora

$$c = \frac{1}{xy2 \left| \log \frac{x}{y} \right|} E_\sigma [\sqrt{2 + \sigma^2}],$$

donde $E_\sigma[\sqrt{2 + \sigma^2}]$ es el valor esperado de la función $\sqrt{2 + \sigma^2}$ con $\sigma \sim \text{Ga}^{-1/2} \left(\sigma | \frac{1}{2}, \frac{1}{4} \log^2 \left(\frac{x}{y} \right) \right)$.

Recordemos el siguiente hecho: una observación aleatoria continua w tiene función de densidad $\text{Ga}^{-1/2}(w|\alpha, \beta)$ si $v = \frac{1}{w^2}$ tiene una densidad $\text{Ga}(v|\alpha, \beta)$.

Entonces, el valor esperado requerido puede expresarse en términos de una nueva variable $\nu = \frac{1}{\sigma^2}$ como se detalla a continuación,

$$\begin{aligned} E_\sigma[\sqrt{2 + \sigma^2}] &= c_2 \int_0^\infty \sqrt{2 + \sigma^2} \sigma^{-2} \exp \left\{ -\frac{1}{4\sigma^2} \log^2 \frac{x}{y} \right\} d\sigma \\ &= c_2 \int_0^\infty \sqrt{2 + \frac{1}{\nu}} \exp \left\{ -\frac{\nu}{4} \log^2 \frac{x}{y} \right\} \left| -\frac{1}{2\nu^{3/2}} \right| d\nu \\ &= \frac{c_2}{2} \int_0^\infty \sqrt{2 + \frac{1}{\nu}} \sqrt{\frac{1}{\nu}} \exp \left\{ -\frac{\nu}{4} \log^2 \frac{x}{y} \right\} d\nu \\ &= E_\nu \left[\sqrt{2 + \frac{1}{\nu}} \right] \end{aligned}$$

con $\sigma = \frac{1}{\sqrt{\nu}}$, $d\sigma = -\frac{1}{2\nu^{3/2}}$ y constante de marginalización $c_2 = \frac{|\log x/y|}{\sqrt{\pi}}$.

Ahora se tiene que $\nu \sim \text{Ga}(\nu | \frac{1}{2}, \frac{1}{4} \log^2 \frac{x}{y})$ y el siguiente paso es verificar la existencia de la esperanza $E_\nu(\sqrt{2 + 1/\nu})$.

Considere una variable aleatoria z proveniente de una distribución *Gama*, $z \sim \text{Ga}(z|\alpha, \beta)$, y las dos funciones siguientes

$$f(z) = \sqrt{2 + 1/z} \quad \text{y} \quad g(z) = z^{-1/2}.$$

Puede verificarse fácilmente que

$$g(z) = \frac{1}{\sqrt{z}} < \sqrt{2 + \frac{1}{z}} = f(z), \quad \forall z > 0.$$

Obtengamos el valor esperado de $g(z)$, que es igual a

$$E_z[z^{-1/2}|\alpha, \beta] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty z^{-1/2} z^{\alpha-1} \exp\{-\beta z\} dz = \frac{\sqrt{\beta} \Gamma(\alpha - 1/2)}{\Gamma(\alpha)}$$

la cual está definida si y sólo si $\alpha > \frac{1}{2}$.

Usemos ahora la prueba de comparación de convergencia para integrales,

$$E_z[f(z)|\alpha, \beta] = \int_0^\infty \sqrt{2 + \frac{1}{z}} z^{\alpha-1} e^{-\beta z} dz > \int_0^\infty \frac{1}{\sqrt{z}} z^{\alpha-1} e^{-\beta z} dz = E_z[g(z)|\alpha, \beta].$$

Nótese que si $\alpha = 1/2$, $E_z[g(z)|\alpha, \beta]$ diverge por tanto $E_z[f(z)|\alpha, \beta]$ también.

Recordemos que el valor esperado requerido es $E_\nu[\sqrt{2 + \frac{1}{\nu}} | \frac{1}{2}, \frac{1}{4} \log^2 \frac{x}{y}]$ el cual no está definido dado que el valor de parámetro α es $1/2$.

Por tanto la constante c de marginalización de la densidad final $\pi(\mu, \alpha|x, y)$ diverge.

Hemos demostrado que la densidad final $\pi(\mu, \alpha|x, y)$ no existe cuando la media es la cantidad de interés con solo dos observaciones. La demostración es análoga para el caso de la moda como cantidad de interés.

A continuación obtengamos el valor de la constante de marginalización cuando se tiene una muestra aleatoria $\mathbf{x} = \{x_1, \dots, x_n\}$ con tamaño de muestra $n > 2$,

$$c = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \prod_{i=1}^n x_i^{-1} \int_{-\infty}^{\infty} \int_0^{\infty} \sigma^{-n} \sqrt{1 + \frac{2}{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n \log x_i - \bar{x}\right)^2\right\} d\mu d\sigma$$

donde el elemento del exponente puede expresarse como

$$\sum_{i=1}^n \log(x_i - \bar{x})^2 = ns^2 + n(\bar{x} - \mu)^2$$

con $s^2 = \frac{\sum_{i=1}^n (\log x_i - \bar{x})^2}{n}$ y $\bar{x} = n^{-1} \sum_{i=1}^n \log x_i$. Entonces,

$$c = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \prod_{i=1}^n x_i^{-1} \int_{-\infty}^{\infty} \sigma^{-(n+1)} \sqrt{2 + \sigma^2} e^{-\frac{ns^2}{2\sigma^2}} \int_0^{\infty} \exp\left\{-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right\} d\mu d\sigma.$$

Nótese que el integrando del extremo derecho corresponde al *kernel* de una densidad Normal $N(\mu|\bar{x}, \sigma/\sqrt{n})$. Por tanto c se reduce a la siguiente expresión,

$$c = \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \frac{1}{\sqrt{n} \prod_{i=1}^n x_i} \int_0^{\infty} \sqrt{2 + \sigma^2} \sigma^{-n} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\} d\sigma.$$

Nuevamente, véase que el integrando puede identificarse como el valor esperado de $\sqrt{2 + \sigma^2}$ cuando σ proviene de la *raíz cuadrada* de la distribución *Gama inversa* (squared-root inverted-Gamma), *i.e.*, $\sigma \sim \text{Ga}^{-1/2}(\sigma|\frac{n-1}{2}, \frac{ns^2}{2})$ y recuérdese que para que esta esperanza esté definida es necesario que $\frac{n-1}{2} > \frac{1}{2}$, *i.e.*, $n > 2$.

Un resultado similar se obtiene cuando el parámetro de interés es la moda $\text{Mo}(x|\mu, \sigma) = e^{\mu - \sigma^2}$.

Apéndice B

Modelo Exponencial con dos parámetros.

En este apéndice se incluyen los siguientes desarrollos:

- Aproximación numérica de la inicial de referencia $\pi_\phi(\phi, \varphi)$ cuando ϕ (mediana) es la cantidad de interés.
 - Aproximación numérica de $\pi(\varphi|\phi)$.
 - Modelo integrado con un solo parámetro $p(x|\phi)$.
 - Aproximación numérica de la inicial $\pi(\phi)$.
- Función de densidad predictiva $\pi_a(\alpha, \varphi)$ y expresión de la probabilidad requerida $\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x})$, para $n = 2$ y $n > 2$ observaciones aleatorias.
- Obtención de la inicial de referencia $\pi_\varphi(\alpha, \varphi)$ cuando φ (parámetro de escala) es la cantidad de interés.
 - Obtención de $\pi(\alpha|\varphi)$.
 - Modelo integrado con un solo parámetro $p_i(x|\varphi)$.
 - Inicial de referencia $\pi(\varphi)$ del modelo integrado $p_i(\cdot|\varphi)$.

Las aproximaciones numéricas de las funciones iniciales de referencia seguirán el pseudo código descrito en la subsección 2.3.1.

B.1. Aproximación numérica de la función inicial de referencia $\pi_\phi(\phi, \varphi)$ cuando ϕ (mediana) es la cantidad de interés.

El modelo Exponencial con dos parámetros $\text{Ex}(x|\alpha, \varphi)$ expresado en términos de la mediana $\phi \equiv \phi(\alpha, \varphi) = \alpha + t\varphi$ y φ , con $t = \log(2)$ es

$$\text{Ex}(x|\phi, \varphi) = \frac{e^{-t}}{\varphi} e^{-(x-\phi)/\varphi}, \quad \varphi > 0, \quad \phi - t\varphi < x.$$

Nótese que el espacio paramétrico es definido por

$$\varphi > \max\left\{\frac{\phi - x}{t}, 0\right\} \quad \text{y} \quad \phi - t\varphi < x.$$

Apliquemos nuevamente el algoritmo de obtención de iniciales de manera recursiva: para un valor fijo de ϕ obtener $\pi(\varphi|\phi)$, luego obtener el modelo integrado de un solo parámetro $p(x|\phi)$ y aplicar nuevamente el algoritmo para obtener $\pi(\phi)$.

B.1.1. Aproximación numérica de la función inicial de referencia $\pi(\varphi|\phi)$

Aplicando el pseudo código de obtención de iniciales dado en la subsección 2.3.1 la Figura 4.3 (Cap. 4) muestra que las aproximaciones numéricas de $\pi(\varphi|\phi)$ se ajustan a la función φ^{-1} .

La Figura 4.3 muestra 20 valores de φ en un rango de e^{-3} a e^3 uniformemente log-espaciados y reescalados tal que $\pi(1|\phi) \propto 1$. Cada aproximación fue calculada con $n = 500$ muestras aleatorias de tamaño $k = 50$.

B.1.2. Obtención del modelo integrado con un solo parámetro $p(x|\phi)$

Dado que la inicial $\pi(\varphi|\phi) \propto \varphi^{-1}$ es impropia, es necesario definir una secuencia creciente de intervalos cerrados de la forma $\{\varphi \in [i^{-1}, i]\}$ con $i > 0$, de tal forma que

$$\pi(\varphi|\phi) = \frac{1}{2 \log(i) \varphi}, \quad \varphi \in [i^{-1}, i].$$

Por tanto, el modelo integrado tendrá la siguiente forma

$$p_i(x|\phi) = \frac{e^{-t}}{2 \log(i)} \int_S \frac{1}{\varphi^2} \exp\left\{-\frac{x-\phi}{\varphi}\right\} d\varphi$$

con $S = \{\varphi : \varphi \in (\max\{\frac{\phi-x}{\varphi}, i^{-1}\}, i]\}$. Nótese que φ estará acotada inferiormente por $\max\{\frac{\phi-x}{\varphi}, i^{-1}\}$ al supuesto $i^{-1} > 0$, y estará acotada superiormente por i . Entonces la densidad $p_i(x|\phi)$ constará de los siguientes dos casos:

- Si $\frac{\phi-x}{t} < \varphi \leq i$,

$$p_i(x|\phi) = \frac{e^{-t}e^{-(x-\phi)/i} - 1}{2 \log(i)(x - \phi)}, \quad \phi - ti \leq x < \phi - t/i$$

- Si $i^{-1} \leq x \leq i$,

$$p_i(x|\phi) = \frac{e^{-t}(e^{-(x-\phi)/i} - e^{-i(x-\phi)})}{2 \log(i)(x - \phi)}, \quad \phi - t/i \leq x < \infty.$$

Finalmente la densidad queda expresada como

$$p_i(x|\phi) = \frac{1}{2 \log(i)(x - \phi)} \begin{cases} e^{-t}e^{-(x-\phi)/i} - 1 & \phi - ti \leq x < \phi - t/i \\ e^{-t}(e^{-(x-\phi)/i} - e^{-i(x-\phi)}) & \phi - t/i \leq x < \infty. \end{cases}$$

Es fácil ver que $p_i(x|\phi)$ integra a uno en \mathcal{X}_i

$$\begin{aligned} 1 &= \int_{\mathcal{X}_i} p_i(x|\phi) dx = \int_{\phi-ti}^{\phi-t/i} p_i(x|\phi) dx + \int_{\phi-t/i}^{\infty} p_i(x|\phi) dx \\ &= \frac{1}{2 \log i} e^{-t} \{ \text{Ei}[t/i^2] - \text{Ei}[t] + 2e^t \log i + \text{Ei}[t] - \text{Ei}[t/i^2] \} = \frac{2 \log i}{2 \log i} \end{aligned}$$

donde $\text{Ei}[a]$ es la función integral exponencial $\text{Ei}[a] = - \int_{-a}^{\infty} \frac{e^{-t}}{t} dt$.

La función de distribución acumulada está dada por

$$P_i(x < X|\phi) = \begin{cases} \frac{e^{-t}}{2 \log(i)} (\Gamma[0, -t] - \Gamma[0, \frac{x-\phi}{i}] - e^t \log \frac{\phi-x}{it}), & \phi - ti \leq x < \phi - \frac{t}{i} \\ 1 + \frac{e^{-t}(\text{Ei}[\frac{-x+\phi}{i}] - \text{Ei}[i(\phi-x)])}{2 \log(i)}, & \phi - t/i \leq x < \infty. \end{cases}$$

donde $\Gamma[a, b]$ es la función incompleta Gamma $\Gamma[a, b] = \int_b^{\infty} t^{a-1} e^{-t} dt$.

El valor esperado y la varianza son, respectivamente,

$$\begin{aligned} \text{E}(x|\phi) &= \phi - \frac{(-1 + i^2)(-1 + t)}{2i \log(i)}, \\ \text{Var}(x|\phi) &= \frac{\log(i)(2 + t(t-2))(i^4 - 1) - (i^2 - 1)^2(t-1)^2}{4i^2(\log(i))^2}. \end{aligned}$$

Para ver el comportamiento del modelo $p_i(x|\phi)$ se han simulado $n = 5000$ observa-

ciones aleatorias con $\phi = 2$ y $i = 3$ cuyo histograma puede verse en la Figura B.1 junto con la función de densidad teórica $p_i(\cdot|\phi)$ (línea continua). La esperanza y varianza son, respectivamente, 2.37241 y 2.07452.

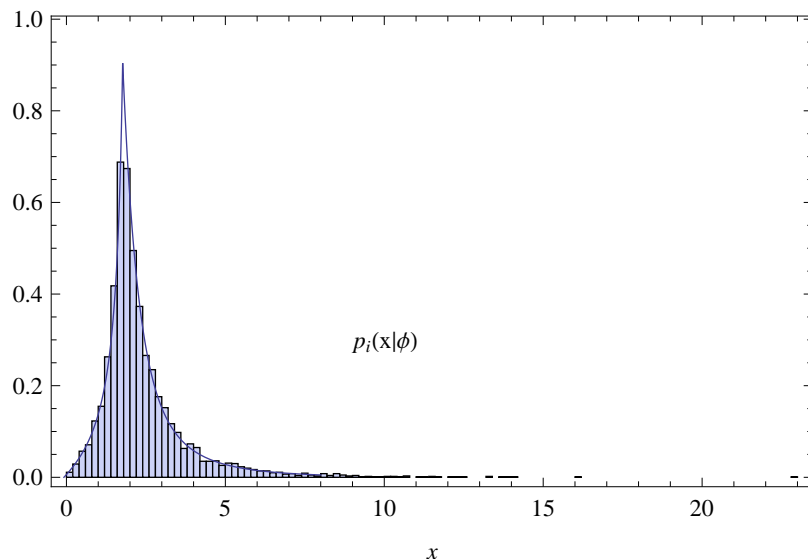


Figura B.1: Histograma y función de densidad teórica del modelo $p_i(x|\phi)$, correspondiente de $n = 5000$ muestras aleatorias con $i = 3$ y $\phi = 2$.

B.1.3. Aproximación numérica de la función inicial $\pi(\phi)$

Una vez obtenido el modelo integrado con un solo parámetro, la inicial de referencia $\pi(\phi)$ será aproximada con el pseudo código de obtención dado en la subsección 2.3.1.

La Figura 4.4 (Cap. 4) muestra que la aproximación numérica de $\pi_i(\phi)$ es constante, por tanto puede suponerse que $\pi_i(\phi) \propto 1$.

La Figura 4.4 muestra las aproximaciones de $\pi(\phi)$ de cuarenta valores de ϕ en un rango de -10 a 10 uniformemente espaciados y re-escalados para que $\pi(1) = 1$. Cada valor corresponde de $n = 1000$ muestras aleatorias de tamaño $k = 50$ con $i = 3$.

B.2. Obtención de la función de densidad predictiva resultante de la inicial $\pi_a(\alpha, \varphi)$. Expresión de la probabilidad requerida $\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x})$, para $n \geq 2$ observaciones aleatorias.

Para obtener las densidades final y predictiva, se utilizará la siguiente expresión general de la función inicial $\pi(\alpha, \varphi) \propto \varphi^{-a}$ con $a > 0$.

B.2.1. Dos observaciones

Sean dos observaciones aleatorias x_1 and x_2 con función de densidad $\text{Ex}(\cdot|\alpha, \varphi)$ y supóngase que $x_1 < x_2$. La densidad final será

$$\pi(\alpha, \varphi|x_1, x_2) = c_a^{-1} \frac{1}{\varphi^2} \exp \left\{ -\frac{x_1 + x_2 - 2\alpha}{\varphi} \right\}, \quad \alpha < x_1, \varphi > 0;$$

con constante de marginalización dada por

$$c_a = \int_{-\infty}^{x_1} \int_0^{\infty} \frac{1}{\varphi^2} e^{-\frac{x_1+x_2-2\alpha}{\varphi}} d\varphi d\alpha.$$

Calculemos el valor de c_a . Nótese que φ tiene una densidad de probabilidad *Gamma inversa*, es decir, $\varphi \sim \text{Ga}^{-1}(\cdot|a+1, x_1+x_2-2\alpha)$.

Entonces, la constante c_a se reduce a

$$\begin{aligned} c_a &= \Gamma(a+1) \int_{-\infty}^{x_1} (x_1 + x_2 - 2\alpha)^{a+1} d\alpha = \Gamma(a+1) \left(\frac{1}{2a(x_1 + x_2 - 2\alpha)^a} \right) \Big|_{-\infty}^{x_1} \\ &= \frac{\Gamma(a+1)}{2a(x_2 - x_1)^a}. \end{aligned}$$

Para la obtención de la densidad predictiva el rango de valores de α debe considerar el valor de la nueva observación, es decir, $-\infty < \alpha < \min\{x, x_1\}$. Entonces la densidad predictiva es la siguiente

$$p(x|x_1, x_2) = c_a^{-1} \int_{-\infty}^{\min\{x, x_1\}} \int_0^{\infty} \frac{1}{\varphi^{a+3}} e^{-\frac{x_1+x_2+x-3\alpha}{\varphi}} d\varphi d\alpha.$$

Nótese que el integrando del extremo derecho corresponde al *kernel* de una densidad $\text{Ga}^{-1}(\cdot|a+2, x_1+x_2+x-3\alpha)$ para φ . Por tanto, la densidad predictiva se reduce a

$$\begin{aligned} p(x|x_1, x_2) &= c_a^{-1} \Gamma(a+2) \int_{-\infty}^{\min\{x, x_1\}} (x_1 + x_2 + x - 3\alpha)^{-(a+2)} d\alpha \\ &= \frac{\Gamma(a+2)}{3c_a(a+1)} (x_1 + x_2 + x - 3 \min\{x, x_1\})^{-(a+1)} \\ &= \frac{\Gamma(a+2)}{3c_a(a+1)} \begin{cases} (x_1 + x_2 - 2x)^{-(a+1)} & x < x_1 \\ (x_2 + x - 2x_1)^{-(a+1)} & x > x_1. \end{cases} \end{aligned}$$

Finalmente, sustituyendo el valor de c_a , resulta

$$p(x|x_1, x_2) = \frac{2}{3}a(x_2 - x_1)^a \begin{cases} (x_1 + x_2 - 2x)^{-(a+1)} & x < x_1 \\ (x_2 + x - 2x_1)^{-(a+1)} & x > x_1. \end{cases}$$

B.2.2. Más de dos observaciones

Sea $\{x_1, \dots, x_n\}$ una muestra aleatoria de $\text{Ex}(\cdot|\alpha, \varphi)$ y utilicemos la expresión general de la previa dada por $\pi(\alpha, \varphi) \propto \varphi^{-a}$. Entonces la densidad final es proporcional a

$$\pi(\alpha, \varphi|\mathbf{x}) \propto \frac{1}{\varphi^{n+a}} \exp\left\{-\frac{t - n\varphi}{\varphi}\right\} \\ -\infty < \alpha < x_{(1)}, \quad \varphi > 0,$$

con $t = \sum_{i=1}^n x_i$ y $x_{(1)} = \min\{x_1, \dots, x_n\}$.

La constante de marginalización viene dada por

$$c = \int_{-\infty}^{x_{(1)}} \int_0^{\infty} \frac{1}{\varphi^{n+a}} \exp\left\{-\frac{t - n\varphi}{\varphi}\right\} d\varphi d\alpha$$

puede notarse que $\varphi \sim \text{Ga}^{-1}(\varphi|n + a - 1, t - n\alpha)$. Por lo tanto c es

$$c = \Gamma(n + a - 1) \int_{-\infty}^{x_{(1)}} (t - n\alpha)^{-(n+a-1)} d\alpha = \frac{\Gamma(n + a - 1)}{n(n + a - 2)} (t - n\alpha)^{-(n+a-2)} \Big|_{-\infty}^{x_{(1)}} \\ = \frac{\Gamma(n + a - 1)}{n(n + a - 2)} (t - nx_{(1)})^{-(n+a-2)}.$$

A continuación obtengamos la densidad predictiva

$$p(x|\mathbf{x}) = \frac{1}{c} \int_{-\infty}^{\min\{x, x_1\}} \int_0^{\infty} \frac{1}{\varphi^{n+a+1}} \exp\left\{-\frac{t + x - (n+1)\alpha}{\varphi}\right\} d\varphi d\alpha.$$

Nuevamente, nótese que el integrando del extremo derecho corresponde al *kernel* de

una densidad $\text{Ga}^{-1}(\varphi|n+a, t+n-(n+1)\alpha)$ para φ , por lo que se reduce a

$$\begin{aligned}
p(x|\mathbf{x}) &= \frac{\Gamma(n+a)}{c} \int_{-\infty}^{\min\{x, x_1\}} (t+x-(n+1)\alpha)^{-(n+a)} d\alpha \\
&= \frac{\Gamma(n+a)}{c} \frac{1}{(n+1)(n+a-1)} (t+x-(n+1)\alpha)^{-(n+a-1)} \Big|_{-\infty}^{\min\{x, x_1\}} \\
&= \frac{\Gamma(n+a)}{c} \frac{(t+x-(n+1)\min\{x, x_1\})^{-(n+a-1)}}{(n+1)(n+a-1)} \\
&= \frac{\Gamma(n+a)}{c(n+1)(n+a-1)} \begin{cases} (t-nx)^{-(n+a-1)} & x < x_{(1)} \\ (t+x-(n+1)x_{(1)}) & x > x_{(1)} \end{cases}
\end{aligned}$$

Finalmente, sustituyendo el valor de c , la densidad predictiva es igual a

$$\begin{aligned}
p(x|\mathbf{x}) &= \frac{n(n+a-2)(t-nx_{(1)})^{n+a-2}}{n+1} \\
&\quad \times \begin{cases} (t-nx)^{-(n+a-1)} & x < x_{(1)}, \\ (t+x-(n+1)x_{(1)})^{-(n+a-1)} & x > x_{(1)}. \end{cases}
\end{aligned}$$

Nuestro interés se centra en a probabilidad $\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x})$ que es la siguiente

$$\begin{aligned}
\Pr(x_{(1)} < x < x_{(n)}|\mathbf{x}) &= \frac{n(n+a-2)(t-nx_{(1)})^{n+a-2}}{n+1} \\
&\quad \times \int_{x_{(1)}}^{x_{(n)}} (t+x-(n+1)x_{(1)})^{-(n+a-1)} dx \\
&= \frac{n(n+a-2)(t-nx_{(1)})^{n+a-2}}{(n+1)(n+a-2)} \\
&\quad \times \left((t-nx_{(1)})^{-(n+a-2)} - (t+x_{(n)}-(n+1)x_{(1)})^{-(n+a-2)} \right) \\
&= \frac{n}{n+1} \left(1 - \left(\frac{t-nx_{(1)}}{t+x_{(n)}-(n+1)x_{(1)}} \right)^{n+a-2} \right)
\end{aligned}$$

como puede verse es una función que depende de los datos y de a .

B.3. Obtención de la inicial de referencia $\pi_\varphi(\alpha, \varphi)$ cuando φ (parámetro de escala) es la cantidad de interés.

Para obtener la inicial de referencia $\pi_\varphi(\alpha, \varphi)$, aplicaremos recursivamente el algoritmo de obtención de la siguiente manera:

Fijar φ para obtener la previa condicional $\pi(\alpha|\varphi)$, luego obtener el modelo integrado

$$p(x|\varphi) = \int_{\Lambda} \text{Ex}(x|\alpha, \varphi)\pi(\alpha|\varphi)d\alpha$$

y volver a aplicar el algoritmo para obtener $\pi(\varphi)$.

B.3.1. Obtención de $\pi(\alpha|\varphi)$.

Desde que se fija φ , este modelo se convierte en uno que depende solamente de α . La obtención de $\pi(\alpha|\varphi)$ es esencialmente la misma a la dada en la subsubsección 4.5.3 para $\pi(\alpha)$ cuando $\varphi = 1$.

B.3.2. Modelo integrado con un solo parámetro $p_i(x|\varphi)$

Desde que $\pi(\alpha|\varphi) \propto 1$ es impropia, será necesario acotarla en una secuencia creciente de intervalos del espacio paramétrico de α dado por

$$\Lambda_i = \{\alpha : \alpha \in (-i, i), i \neq 0\}$$

tal que $\pi_i(\alpha|\varphi)$ será propia. Entonces $\pi_i(\alpha|\varphi)$ queda expresada como

$$\pi_i(\alpha|\theta) = \begin{cases} \frac{1}{2i} & -i < \alpha < i \\ 0 & \text{en otro caso} \end{cases}.$$

De esta manera el modelo integrado con un solo parámetro $p_i(\cdot|\varphi)$ estará definido como

$$p_i(x|\varphi) = \int_S \text{Ex}(x|\alpha, \varphi)\pi_i(\alpha|\varphi)d\alpha$$

con

$$S = \{\alpha : \alpha \in (-i, i) \cap (-\infty, x)\},$$

que será el espacio paramétrico de α y para el que no solamente las condiciones de $\pi_i(\cdot)$ deben de considerarse, si no también las condiciones del modelo sobre el parámetro, es decir, $-\infty < \alpha < x$.

Así, entonces S es la intersección de ambos intervalos, y $p_i(\cdot|\varphi)$ está definida en los siguiente casos:

- Si $i \leq x$ entonces $-i < \alpha < i$,

$$\begin{aligned}\int_{-i}^i \text{Ex}(x|\alpha, \varphi) \pi_i(\alpha|\varphi) d\alpha &= \frac{1}{2i\varphi} \int_{-i}^i e^{-(x-\alpha)/\varphi} d\alpha \\ &= \frac{1}{2i} (e^{-(x-i)/\varphi} - e^{-(x+i)/\varphi}).\end{aligned}$$

- Si $i > x > -i$ entonces $-i < \alpha < x$,

$$\begin{aligned}\int_{-i}^x \text{Ex}(x|\alpha, \varphi) \pi_i(\alpha|\varphi) d\alpha &= \frac{1}{2i\varphi} \int_{-i}^x e^{-(x-\alpha)/\varphi} d\alpha \\ &= \frac{1}{2i} (1 - e^{-(x+i)/\varphi}).\end{aligned}$$

Nótese que en el caso $x < i$, x se encuentra acotada inferiormente por $-i$, puesto que es la cota inferior de la inicial $\pi_i(\cdot|\alpha)$ cuando $\alpha \in (-i, i)$.

Finalmente, el modelo integrado se resume como:

$$p_i(x|\varphi) = \frac{1}{2i} \begin{cases} e^{-(x-i)/\varphi} - e^{-(x+i)/\varphi} & i \leq x < \infty \\ 1 - e^{-(x+i)/\varphi} & -i < x < i. \end{cases} \quad (\text{B.1})$$

Puede demostrarse muy fácilmente que $p_i(\cdot|\varphi)$ integra a uno en el espacio paramétrico \mathcal{X}_i :

$$\begin{aligned}\int_{\mathcal{X}} p_i(x|\varphi) dx &= \frac{1}{2i} \int_i^\infty (e^{-(x-i)/\varphi} - e^{-(x+i)/\varphi}) dx + 1 - \frac{1}{2i} \int_{-i}^i e^{-(x+i)/\varphi} dx \\ &= \frac{\varphi}{2i} (1 - e^{-2i/\varphi}) + 1 + \frac{\varphi}{2i} (e^{-2i/\varphi} - 1) = 1.\end{aligned}$$

La obtención de la función de distribución acumulada también dividirse en dos casos:

- Si $i \leq x$, entonces

$$\begin{aligned}\Pr(x < X|\varphi) &= \frac{1}{2i} \int_{-i}^i (1 - e^{-(t+i)/\varphi}) dt + \frac{1}{2i} \int_i^x (e^{-(t-i)/\varphi} - e^{-(t+i)/\varphi}) dt \\ &= 1 + \frac{\varphi}{2i} (e^{-2i/\varphi} - 1) + \frac{\varphi}{2i} (1 - e^{-(x-i)/\varphi} + e^{-(x+i)/\varphi} - e^{-2i/\varphi}) \\ &= 1 + \frac{\varphi}{2i} (e^{-(x+i)/\varphi} - e^{-(x-i)/\varphi}).\end{aligned}$$

- Si $-i < x < i$, entonces

$$\Pr(x < X|\varphi) = \frac{1}{2i} \int_{-i}^x (1 - e^{-(t+i)/\varphi}) dt = \frac{1}{2i} (x + i + \varphi(e^{-(x+i)/\varphi} - 1)).$$

Por lo tanto,

$$\Pr(x < X|\varphi) = \begin{cases} 1 + \frac{\varphi}{2i} (e^{-(x+i)/\varphi} - e^{-(x-i)/\varphi}) & i \leq x \\ \frac{1}{2i} (x + i + \varphi(e^{-(x+i)/\varphi} - 1)) & -i < x < i. \end{cases}$$

La media y la varianza también pueden obtenerse. Algunos de los cálculos se han omitido.

$$\begin{aligned} E_{x|\varphi}(x|\varphi) &= \int_i^\infty \frac{x}{2i} (e^{-(x-i)/\varphi} - e^{-(x+i)/\varphi}) dx + \int_{-i}^i \frac{x}{2i} (1 - e^{-(x+i)/\varphi}) dx \\ &= \frac{\varphi(i + \varphi)}{2i} (1 - e^{-2i/\varphi}) + \frac{\varphi}{2i} ((i + \varphi)e^{-2i/\varphi} + i - \varphi) \\ &= \varphi, \end{aligned}$$

y el valor esperado de x^2 , que nos ayudará para calcular la varianza, es

$$\begin{aligned} E_{x|\varphi}(x^2|\varphi) &= \int_i^\infty \frac{x^2}{2i} (e^{-(x-i)/\varphi} - e^{-(x+i)/\varphi}) dx + \int_{-i}^i \frac{x^2}{2i} (1 - e^{-(x+i)/\varphi}) dx \\ &= I_1 + I_2 \end{aligned}$$

evaluando cada una de las integrales I_1 e I_2 se tiene que

$$\begin{aligned} I_1 &= \frac{\varphi}{2i} (1 - e^{-2i/\varphi}) (i^2 + 2i\varphi + 2\varphi^2), \\ I_2 &= \frac{i^3}{3i} - \frac{\varphi}{2i} (e^{-2i/\varphi} (-i^2 - 2\varphi i - 2\varphi^2) + i^2 - 2i\varphi + 2\varphi^2). \end{aligned}$$

Entonces $E_{x|\varphi}(x^2|\varphi) = \frac{i^2}{3} + 2\varphi^2$, y la varianza es

$$\text{Var}(x|\varphi) = E_{x|\varphi}(x^2|\varphi) - \varphi^2 = \frac{i^2}{3} + \varphi^2.$$

Finalmente, simularemos observaciones aleatorias del modelo integrado con un solo parámetro y así para poder ver su comportamiento.

La Figura B.2 muestra el histograma correspondiente de $n = 10000$ observaciones, con $i = 2$ y $\varphi = 1/5$, la línea continua corresponde a la densidad teórica. Adicionalmente, el valor esperado es $\frac{1}{5}$ y la varianza es $\frac{103}{75}$.

B.3.3. Obtención de la función inicial de referencia $\pi(\varphi)$ del modelo integrado $p_i(\cdot|\varphi)$

Habiendo obtenido el modelo integrado con un parámetro $p_i(\cdot|\varphi)$ en la sección B.3.2 en este Apéndice, obtengamos la inicial de referencia para φ .

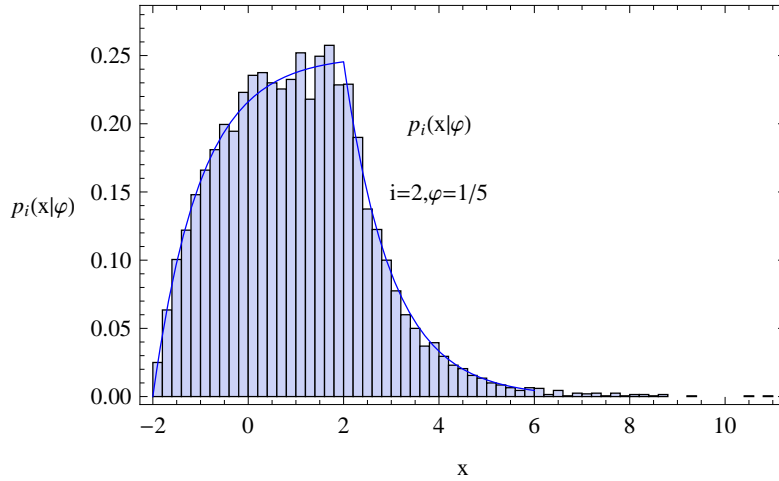


Figura B.2: Histograma correspondiente de $n = 10000$ observaciones simuladas del modelo integrado con un parámetro $p_i(x|\varphi)$, con $i = 2$, $\varphi = 1/5$. La línea continua corresponde a la función de densidad teórica.

La inicial $\pi(\varphi)$ será aproximada usando el método numérico de obtención de previas dado en la subsección 2.3.1. y puede verse en la Figura B.3 las aproximaciones de $\pi(\varphi)$ para veinte valores de φ en un rango de e^{-3} a e^3 , uniformemente log-espaciados y re-escalados para que $\pi(1) = 1$. Por cada valor de φ se simularon $n = 5000$ de muestras de tamaño $k = 300$ con $i = 2$. La línea continua corresponde a φ^{-1} , de esta forma podemos deducir que la inicial buscada es proporcional a φ^{-1} .

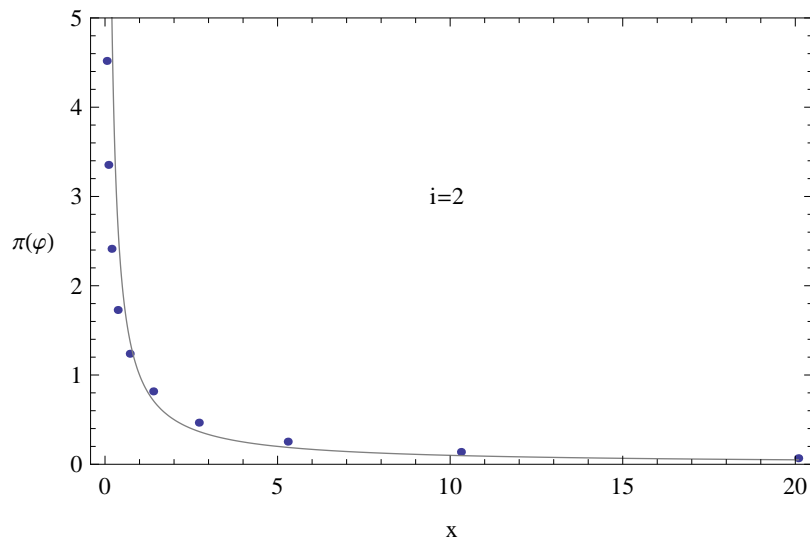


Figura B.3: Aproximación numérica de la función inicial de referencia $\pi(\varphi)$, proveniente de 20 valores de φ en (e^{-3}, e^3) uniformemente log-espaciados y re-escalados para que $\pi(1) = 1$. Cada punto fue calculado con $n = 5000$ muestras aleatorias de tamaño $k = 300$ y valor $i = 2$. La línea continua corresponde a $\pi(\varphi) \propto \varphi^{-1}$.

Apéndice C

Modelo normal Bivariado.

Dentro de este apéndice, se desarrollan la obtenciones de las densidades final y predictiva de cada una de las iniciales sugeridas en la subsección 5.2.2, también se analizan las funciones g y k .

Supóngase el modelo normal bivariado $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$, con $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, σ y ρ desconocidos, vector de medias $\boldsymbol{\mu} = \mathbf{0}$ conocido.

C.1. Densidades final y predictiva de la inicial de referencia π_ρ .

Suponga la inicial

$$\pi_\rho(\rho, \sigma) \propto \frac{1}{\sigma(1 - \rho^2)}.$$

Dadas dos observaciones \mathbf{x}_1 y \mathbf{x}_2 , la densidad final es proporcional a

$$\pi_\rho(\rho, \sigma | \mathbf{x}_1, \mathbf{x}_2) \propto \frac{\sigma^{-5}(1 - \rho^2)^{-2}}{(2\pi)^2} \exp \left\{ -\frac{1}{2\sigma^2(1 - \rho^2)}(s_{.1} + s_{.2} - 2\rho s_{12}) \right\}$$

donde $s_{.1} = x_{11} + x_{21}$, $s_{.2} = x_{12} + x_{22}$ y $s_{12} = x_{11}x_{12} + x_{21}x_{22}$.

El valor de la constante marginalización es

$$c_\rho = \int_{-1}^1 \int_0^\infty \frac{\sigma^{-5}(1 - \rho^2)^{-2}}{(2\pi)^2} \exp \left\{ -\frac{1}{2\sigma^2(1 - \rho^2)}(s_{.1} + s_{.2} - 2\rho s_{12}) \right\} d\sigma d\rho$$

donde puede identificarse en la función exponencial el *kernel* de la raíz cuadrada de la densidad *Gamma inversa* $\text{Ga}^{-1/2}$ para σ , i.e., $\sigma \sim \text{Ga}^{-1/2} \left(\sigma | 2, \frac{s_{.1} + s_{.2} - 2\rho s_{12}}{2(1 - \rho^2)} \right)$. Por lo que c_ρ ahora es

$$c_\rho = \frac{1}{4\pi^2} \frac{\Gamma(2)}{2} \frac{1}{s_{12}^2} \int_{-1}^1 \frac{1}{(g - \rho)^2} d\rho = \frac{1}{4\pi^2} \frac{1}{s_{12}^2} \frac{1}{g^2 - 1}$$

donde

$$g \equiv g(s_{.1}, s_{.2}, s_{12}) = \frac{s_{.1} + s_{.2}}{2s_{12}}$$

es una función de \mathbf{x}_1 y \mathbf{x}_2 .

La constante c_ρ está definida si $|g| > 1$, véase la Sección C.3 de este apéndice.

La densidad predictiva puede obtenerse fácilmente,

$$p_\rho(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{c_\rho 4\pi^2} \int_{-1}^1 \int_0^\infty \frac{1}{\sigma^7(1-\rho^2)^{5/2}} e^{-\frac{s_{.1}+s_{.2}+x_1^2+x_2^2-2\rho(s_{12}+x_1x_2)}{2\sigma^2(1-\rho^2)}} d\sigma d\rho,$$

donde puede identificarse nuevamente el *kernel* de la raíz cuadrada de la densidad *Gamma inversa* para el parámetro σ , i.e.,

$$\sigma \sim \text{Ga}^{-1/2} \left(\sigma | 3, \frac{s_{.1} + s_{.2} + x_1^2 + x_2^2 - 2\rho(s_{12} + x_1x_2)}{2(1-\rho^2)} \right).$$

Entonces la densidad predictiva adquiere la siguiente expresión,

$$\begin{aligned} p_\rho(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) &= \frac{\Gamma(3)}{c_\rho 8\pi^3} \frac{1}{2(s_{12} + x_1x_2)^3} \int_{-1}^1 \frac{\sqrt{1-\rho^2}}{(k-\rho)^3} d\rho \\ &= \frac{\Gamma(3)}{2^4 c_\rho \pi^3} \frac{1}{(s_{12} + x_1x_2)^3} \frac{\pi|k|}{2k(k^2-1)^{3/2}} \end{aligned}$$

donde $k \equiv k(\mathbf{x}_1, \mathbf{x}_2) = \frac{s_{.1}+s_{.2}+x_1^2+x_2^2}{2(s_{12}+x_1x_2)}$, y es válido para $|k| > 1$ (véase la Secc. C.3 de este apéndice).

C.2. Densidad final de la inicial de referencia π_σ .

Dada la inicial de referencia

$$\pi_\rho(\rho, \sigma) \propto \frac{\sqrt{1+\rho^2}}{\sigma(1-\rho^2)},$$

y dos observaciones \mathbf{x}_1 y \mathbf{x}_2 , la densidad final es proporcional a

$$\pi_\sigma(\sigma, \rho|\mathbf{x}_1, \mathbf{x}_2) \propto \frac{(1-\rho^2)^{-2} \sqrt{1+\rho^2}}{4\pi^2 \sigma^5} \exp \left\{ -\frac{1}{2\sigma^2(1-\rho^2)} (s_{.1} + s_{.2} - 2\rho s_{12}) \right\}.$$

La constante de marginalización es

$$c_\sigma = \int_{-1}^1 \int_0^\infty \frac{\sqrt{1-\rho^2}}{\sigma^5(1-\rho^2)^2 4\pi^2} \exp \left\{ -\frac{s_{.1} + s_{.2} - 2\rho s_{12}}{2\sigma^2(1-\rho^2)} \right\} d\sigma d\rho$$

como se ha hecho anteriormente, puede identificarse el *kernel* de la raíz cuadrada de la densidad *Gamma inversa* para σ i.e.,

$$\sigma \sim \text{Ga}^{-1/2} \left(\sigma|2, \frac{s_{\cdot 1} + s_{\cdot 2} - 2\rho s_{12}}{2(1 - \rho^2)} \right).$$

Entonces c_σ es igual a

$$\begin{aligned} c_\sigma &= (2\pi)^{-2} \Gamma(2) \frac{2}{(2s_{12})^2} \int_{-1}^1 \frac{\sqrt{1 + \rho^2}}{(g - \rho)^2} d\rho \\ &= \frac{1}{8\pi^2 s_{12}^2} \frac{1}{(g^2 - 1)\sqrt{1 + g^2}} * \left(\begin{array}{l} 2\sqrt{1 + g^2} (\sqrt{2} + (g^2 - 1) \operatorname{arcsinh}(1)) \\ +g(g^2 - 1) \log \left(\frac{g-1}{1+g} * \frac{1-g+\sqrt{2(1+g^2)}}{1+g+\sqrt{2(1+g^2)}} \right) \end{array} \right) \end{aligned}$$

con $g \equiv g(\mathbf{x}_1, \mathbf{x}_2) = \frac{s_{\cdot 1} + s_{\cdot 2}}{2s_{12}}$ y $|g| > 1$ (Secc. C.3).

La densidad predictiva se puede obtener fácilmente,

$$p_\sigma(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{c_\sigma 8\pi^3} \int_{-1}^1 \int_0^\infty \frac{\sqrt{1 + \rho^2}}{\sigma^7 (1 - \rho^2)^{5/2}} e^{-\frac{s_{\cdot 1} + s_{\cdot 2} + x_1^2 + x_2^2 - 2\rho(s_{12} + x_1 x_2)}{2\sigma^2(1 - \rho^2)}} d\sigma d\rho.$$

Análogamente puede identificarse el *kernel* de la raíz cuadrada de la densidad *Gamma inversa* para el parámetro σ , i.e.,

$$\sigma \sim \text{Ga}^{-1/2} \left(\sigma|3, \frac{s_{\cdot 1} + s_{\cdot 2} + x_1^2 + x_2^2 - 2\rho(s_{12} + x_1 x_2)}{2(1 - \rho^2)} \right)$$

lo cual reduce a p_σ en la siguiente expresión

$$\begin{aligned} p_\sigma(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) &= \frac{\Gamma(3)}{c_\sigma} \frac{4}{(2\pi)^3} \frac{1}{(2(s_{12} + x_1 x_2))^3} \int_{-1}^1 \frac{\sqrt{(1 + \rho^2)(1 - \rho^2)}}{(k - \rho)^3} d\rho \\ &= \frac{\Gamma(3)}{c_\sigma} \frac{4}{(2\pi)^3} \frac{1}{(2(s_{12} + x_1 x_2))^3} \int_{-1}^1 \frac{\sqrt{1 - \rho^4}}{(k - \rho)^3} d\rho \end{aligned}$$

con $k \equiv k(\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2) = \frac{s_{\cdot 1} + s_{\cdot 2} + x_1^2 + x_2^2}{2(s_{12} + x_1 x_2)}$ y la integral converge si $|k| > 1$ (Secc. C.3).

Finalmente, la densidad predictiva es

$$\begin{aligned} p_\sigma(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2) &= \frac{\Gamma(3)}{\mathbf{C}_\sigma} \frac{4}{(2\pi)^3} \frac{1}{(2(s_{12} + x_1 x_2))^3} \frac{1}{5k^5 \sqrt{2\pi}} \\ &\times \left(\begin{array}{l} 32\Gamma\left(\frac{3}{4}\right) \Gamma\left(\frac{7}{4}\right) \operatorname{Hypergeometric2F1}\left(\frac{7}{4}, 2, \frac{9}{4}, \frac{1}{k^4}\right) \\ -5k^2 \Gamma\left(-\frac{3}{4}\right) \Gamma\left(\frac{5}{4}\right) \operatorname{HypergeometricPFQ}\left(\left\{1, \frac{5}{4}, \frac{3}{2}\right\}, \left\{\frac{1}{2}, \frac{7}{4}\right\}, \frac{1}{k^4}\right) \end{array} \right) \end{aligned}$$

donde Hypergeometric2F1(a, b, c, z) tiene la siguiente expansión en series de

$${}_2F_1(a, b; c; z) = \sum_{k=0}^{\infty} (a)_k (b)_k / (c)_k z^k / k!$$

y HypergeometricPFQ($\{a_1, \dots, a_p\}, \{b_1, \dots, b_q\}, z$) con la siguiente expansión en series

$${}_pF_q(a; b; z) = \sum_{k=0}^{\infty} (a_1)_k \dots (a_p)_k / (b_1)_k \dots (b_q)_k z^k / k!$$

C.3. Demostración de $|g| > 1$ y $|k| > 1$.

Los puntos $\mathbf{x}_1 = (x_{11}, x_{12})$ y $\mathbf{x}_2 = (x_{21}, x_{22})$ cumplen la propiedad de

$$(x_{11} - x_{12})^2 = (x_{12} - x_{11})^2 \geq 0 \quad \text{y} \quad (x_{11} + x_{12})^2 \geq 0,$$

entonces

$$(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2 \geq 0, \quad \text{y} \quad \frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2}{2(x_{11}x_{12} + x_{21}x_{22})} \geq 1.$$

Si se tiene el caso de

$$(x_{11} + x_{12})^2 + (x_{21} + x_{22})^2 \geq 0,$$

entonces

$$-\frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2}{2(x_{11}x_{12} + x_{21}x_{22})} \geq 1.$$

Si definimos a $g \equiv g(\mathbf{x}_1, \mathbf{x}_2) = \frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2}{2(x_{11}x_{12} + x_{21}x_{22})}$, luego g deber ser $|g| \geq 1$.

La condición $|k| \geq 1$, tiene un comportamiento similar desde que

$$k \equiv k(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}) = \frac{s_{.1} + s_{.2} + x_1^2 + x_2^2}{2(s_{12} + x_1x_2)}$$

con

$$s_{.1} = x_{11}^2 + x_{21}^2, \quad s_{.2} = x_{12}^2 + x_{22}^2 \quad \text{y} \quad s_{12} = x_{11}x_{12} + x_{21}x_{22}.$$

Demostremos ahora la condición $|g| > 1$ para que la densidad final converja, para eso trabajaremos con la expresión general de la inicial dada por

$$\pi(\sigma, \rho) \propto \sigma^{-a} (1 - \rho^2)^{-b} (1 + \rho^2)^c,$$

entonces, dadas dos observaciones \mathbf{x}_1 y \mathbf{x}_2 la constante de marginalización es

$$c = \int_{-1}^1 \int_0^\infty \frac{(1 + \rho^2)^c}{\sigma^{a+4}(1 - \rho^2)^{b+1}} e^{-\frac{1}{2\sigma^2(1-\rho^2)}(s_{.1}+s_{.2}-2\rho s_{12})} d\sigma d\rho.$$

Comencemos con los valores de $a, b \neq 0, c = 0$ (la inicial π_ρ pertenece a este caso). El valor del constante de la densidad final tiene la siguiente forma

$$c = \int_{-1}^1 \int_0^\infty \frac{(1 - \rho^2)^{-(b+1)}}{\sigma^{a+4}(2\pi)^2} e^{-\frac{1}{2\sigma^2(1-\rho^2)}(s_{.1}+s_{.2}-2\rho s_{12})} d\sigma d\rho. \quad (C.1)$$

Puede identificarse el *kernel* de la raíz cuadrada de la densidad *Gama inversa* para el parámetro σ , *i.e.*,

$$\sigma \sim \text{Ga}^{-1/2} \left(\frac{a+3}{2}, \frac{s_{.1} + s_{.2} - 2\rho s_{12}}{2(1-\rho^2)} \right).$$

Entonces la constante c es

$$c = (2\pi)^{-2} \Gamma \left(\frac{a+3}{2} \right) 2^{-1} \frac{1}{s_{12}^{(a+3)/2}} \underbrace{\int_{-1}^1 \frac{(1 - \rho^2)^{(a+1)/2-b}}{(g - \rho)^{(a+3)/2}} d\rho}_{I_1}.$$

con $g = \frac{s_{.1}+s_{.2}}{2s_{12}}$. Trabajemos únicamente la integral I_1 a la cual le aplicamos la siguiente transformación de ρ :

$$\tau = \log \frac{1 + \rho}{1 - \rho}, \quad \infty < \tau < \infty, \quad d\tau = (1 + \cosh \tau)^{-1}.$$

Entonces,

$$\begin{aligned} I_1 &= \frac{1}{2} \int_{-\infty}^{\infty} (\text{sech}^2 \frac{\tau}{2})^{(3+a-2b)/2} (g - \tanh \frac{\tau}{2})^{-(3+a)/2} (1 + \cosh \tau)^{-1} d\tau \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{(\cosh \frac{\tau}{2})^{(3+a-2b)} (g - \tanh \frac{\tau}{2})^{(3+a)/2}} d\tau. \end{aligned} \quad (C.2)$$

Utilicemos las siguientes definiciones de las funciones hiperbólicas

$$\cosh z = \frac{e^z + e^{-z}}{2}, \quad \sinh z = \frac{e^z - e^{-z}}{2}, \quad \tanh z = \frac{\sinh z}{\cosh z}. \quad (C.3)$$

Entonces C.2 es igual a

$$\frac{1}{2} \int_{-\infty}^{\infty} \left(\frac{e^{\tau/2} + e^{-\tau/2}}{2} \right)^{-(3+a-2b)} \left(g - \frac{e^{\tau/2} - e^{-\tau/2}}{e^{\tau/2} + e^{-\tau/2}} \right)^{-(3+a)/2} d\tau. \quad (C.4)$$

Si $a = b = 1$ y $c = 0$ (inicial π_ρ), la integral (C.4) se reduce a

$$\int_{-\infty}^{\infty} \frac{2e^\tau}{(1+g+e^\tau(g-1))^2} d\tau = \frac{2}{g^2-1}$$

claramente puede verse que $|g| > 1$.

Por otro lado, suponga el caso de $g = 1$ entonces demostremos que la integral (C.4) con los valores $a = b = 1, c = 0$ diverge,

$$\int_{-\infty}^{\infty} \left(\frac{e^{\tau/2} + e^{-\tau/2}}{2} \right)^{-2} \left(1 - \frac{e^{\tau/2} - e^{-\tau/2}}{e^{\tau/2} + e^{-\tau/2}} \right)^{-2} d\tau = \int_{-\infty}^{\infty} e^{-\tau} d\tau = \infty.$$

El caso $g = -1$ arroja un resultado es similar.

Ahora demostremos que para los valores $a, b, c \neq 0$ (la inicial π_σ corresponde a este caso) y $g = 1$, la integral C.1 no converge. Sea la nueva constante,

$$c = \int_{-1}^1 \int_0^\infty \frac{\sigma^{-(a+4)}(1+\rho^2)^c}{(2\pi)^2(1-\rho^2)^{b+1}} e^{-\frac{1}{2\sigma^2(1-\rho^2)}(s_{.1}+s_{.2}-2\rho s_{12})} d\sigma d\rho. \quad (\text{C.5})$$

Nótese que σ tiene una densidad raíz cuadrada de la Gama inversa, i.e.,

$$\sigma \sim \text{Ga}^{-1/2} \left(\sigma \mid \frac{a+3}{2}, \frac{s_{.1}+s_{.2}-2\rho s_{12}}{2(1-\rho^2)} \right).$$

Después de algunas manipulaciones algebraicas la integral C.5 es

$$c = \frac{\Gamma(\frac{a+3}{2})}{8\pi^3 s_{12}^{(a+3)/2}} \int_{-1}^1 (1+\rho^2)^c (1-\rho^2)^{(a+3)/2-(b+1)} (g-\rho)^{-(a+3)/2} d\sigma d\rho.$$

con g determinada previamente.

Aplicando nuevamente la transformación $\tau = \log \frac{1+\rho}{1-\rho}$, y usando las igualdades C.3, la integral tiene la siguiente forma

$$\int_{-\infty}^{\infty} (\text{sech} \frac{\tau}{2})^c (\cosh \tau \sinh^2 \frac{\tau}{2})^{(a+1-2b)/2} (g - \tanh \frac{\tau}{2})^{-(a+3)/2} d\rho.$$

Si sustituímos los valores específicos $a = b = 1, c = 1/2$ (inicial π_σ) con $g = 1$, se tiene que diverge

$$\int_{-\infty}^{\infty} \frac{1}{2} e^\tau \text{sech} \frac{\tau}{2} d\tau = 2(e^{\tau/2} - \text{arctanh } e^{\tau/2})|_{-\infty}^{\infty} = \infty.$$

Si $g = -1$ arroja un resultado similar. Por lo tanto se requiere que $|g| > 1$.

La función k , también debe cumplir con la condición $|k| > 1$ cuya demostración es

similar a la de la función g y por tanto se omite.

Apéndice D

Coeficiente de correlación desconocido

Demostración de la convergencia de las distribuciones finales

En este apéndice se desarrolla la demostración de la convergencia de las funciones finales dadas en la sección 5.2.1.

Sea el modelo Normal bivariado $N_2(\mathbf{x}|\mathbf{0}, \Sigma)$ donde

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

con ρ desconocido.

Se desarrolla la demostración utilizando la expresión general de las funciones iniciales dada por (5.4),

$$\pi_{ab}(\pi) \propto (1 - \rho^2)^a (1 + \rho^2)^b.$$

Recuérdese que

- Si $a = -1$ y $b = 1/2$, tendremos la función inicial de referencia

$$\pi_{R_\rho}(\rho) \propto \sqrt{i(\rho)} = \frac{\sqrt{1 + \rho^2}}{1 - \rho^2};$$

- Si $a = -1$ y $b = 0$ se tiene la función inicial $\pi_\varphi(\rho) \propto \frac{1}{1 - \rho^2}$ correspondiente de la transformación $\varphi = \operatorname{arctanh}\rho$.

Dadas dos observaciones aleatorias $\mathbf{x}_1 = \{x_{11}, x_{12}\}$ y $\mathbf{x}_2 = \{x_{21}, x_{22}\}$, la función final

es proporcional a

$$\begin{aligned} \pi(\rho|\mathbf{x}_1, \mathbf{x}_2) &\propto (1 - \rho^2)^{a-1}(1 + \rho^2)^b \\ &\times \exp \left\{ \frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}{2(1 - \rho^2)} \right\} \end{aligned}$$

donde la constante de marginalización es

$$c = \int_{-1}^1 \frac{(1 + \rho^2)^b}{(1 - \rho^2)^{1-a}} \exp \left\{ \frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}{2(1 - \rho^2)} \right\} d\rho.$$

Nótese que el elemento de la integral anterior $(1 + \rho^2)^b$ es una función continua en el intervalo $(-1, 1)$, cuyo valor siempre está definido, por tanto podemos omitirlo y definir la siguiente función

$$f(\rho) = (1 - \rho^2)^{a-1} \exp \left\{ \frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}{2(1 - \rho^2)} \right\}.$$

entonces nos concentraremos en demostrar que $\int_{-1}^1 f(\rho)d\rho < \infty$. Obsérvese que ambas funciones iniciales coinciden con el valor $a = -1$, por tanto desarrollamos este caso solamente.

Comenzaremos con el elemento del exponente, al cual le sumamos y restamos la cantidad $\rho^2(x_{21}^2 + x_{22}^2)$ para completar los términos cuadráticos,

$$\frac{x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}{1 - \rho^2} = \frac{(x_{11} - \rho x_{12})^2 + (x_{21} - \rho x_{22})^2}{1 - \rho^2} + (x_{12}^2 + x_{22}^2),$$

entonces

$$f(\rho) = \frac{1}{(1 - \rho^2)^2} \exp \left\{ \frac{(x_{11} - \rho x_{12})^2 + (x_{21} - \rho x_{22})^2}{2(1 - \rho^2)} + \frac{x_{12}^2 + x_{22}^2}{2} \right\}. \quad (\text{D.1})$$

Trabajemos ahora el término $(x_{11} - \rho x_{12})^2 + (x_{21} - \rho x_{22})^2$

$$\begin{aligned} &(x_{11} - \rho x_{12})^2 + (x_{21} - \rho x_{22})^2 = \\ &= (x_{12}^2 + x_{22}^2) \times \left\{ \begin{array}{l} \frac{x_{11}^2 + x_{21}^2}{x_{12}^2 + x_{22}^2} - 2\rho \frac{x_{11}x_{12} + x_{21}x_{22}}{x_{12}^2 + x_{22}^2} + \rho^2 \\ - \left(\frac{x_{11}x_{12} + x_{21}x_{22}}{x_{12}^2 + x_{22}^2} \right)^2 + \left(\frac{x_{11}x_{12} + x_{21}x_{22}}{x_{12}^2 + x_{22}^2} \right)^2 \end{array} \right\} \\ &= (x_{12}^2 + x_{22}^2) \left(\rho - \frac{x_{11}x_{12} + x_{21}x_{22}}{x_{12}^2 + x_{22}^2} \right)^2 + (x_{11}^2 + x_{21}^2) - \left(\frac{x_{11}x_{12} + x_{21}x_{22}}{x_{12}^2 + x_{22}^2} \right)^2 \\ &\geq 0. \end{aligned}$$

Nótese que se cumple la siguiente desigualdad

$$(x_{11} - \rho x_{12})^2 + (x_{21} - \rho x_{22})^2 \geq (x_{11}^2 + x_{21}^2) - \left(\frac{x_{11}x_{12} + x_{21}x_{22}}{x_{12}^2 + x_{22}^2} \right)^2 \equiv d$$

Suponiendo que $d > 0$, y puesto que el segundo término del exponente en (D.1) es $\frac{x_{12}^2 + x_{22}^2}{2} \geq 0$, entonces

$$-\frac{(x_{11} - \rho x_{12})^2 + (x_{21} - \rho x_{22})^2}{2(1 - \rho^2)} - \frac{x_{12}^2 + x_{22}^2}{2} \leq -\frac{d}{2(1 - \rho^2)}.$$

Defínase la función $h(\rho)$ como

$$h(\rho) = \frac{1}{(1 - \rho^2)^2} e^{-\frac{d}{2(1 - \rho^2)}},$$

claramente puede verse que $f(\rho) \leq h(\rho)$.

Demostremos ahora que $I = \int_{-1}^1 h(\rho) d\rho < \infty$ usando el hecho de que $h(\rho)$ es simétrica en $(-1, 1)$,

$$I = 2 \int_0^1 h(\rho) d\rho < \infty.$$

Hágase el siguiente cambio de variable

$$u = \frac{d}{1 - \rho^2}, \quad \rho = \sqrt{1 - d/u}, \quad d\rho = \frac{1}{2u^2 \sqrt{1 - d/u}}$$

con el dominio de integración $d < \frac{d}{1 - \rho^2} = u$. Por lo que

$$I = 2 \int_d^\infty (u/d)^2 e^{-d/2} \frac{1}{2u^2 \sqrt{1 - d/u}} du = \frac{1}{d^2} \underbrace{\int_d^\infty \sqrt{\frac{u}{u - d}} e^{-u/2} du}_{I_2}.$$

Por otro lado, sabemos que $u > d > 0$,

- Si $u \geq 1$ entonces $u \geq \sqrt{u}$, y se cumple la desigualdad

$$\sqrt{\frac{u}{u - d}} \leq \frac{u}{\sqrt{u - d}}$$

entonces

$$I_2 \leq \int_d^\infty \frac{u}{\sqrt{u - d}} e^{-u/2} du = (1 + d) \sqrt{2\pi} e^{-d/2} < \infty.$$

- Si $d < u < 1$, entonces $1 > \sqrt{u}$, y

$$I_2 < \int_d^\infty \frac{1}{\sqrt{u-d}} e^{-u/2} du = \sqrt{2\pi} e^{-d/2} < \infty$$

Puesto que $I_2 < \infty$, entonces $I < \infty$ y por tanto $\int_{-1}^1 f(\rho) d\rho < \infty$, las funciones finales son propias.

Referencias

- Abramowitz, M. y Stegun, I. A. (1972). *Handbook of mathematical functions*. New York: Dover Publications, Inc.
- Aitchison, J. (1964). Two papers on the comparison of Bayesian and frequentist approaches to statistical problems for prediction: Bayesian tolerance regions. *Journal of the Royal statistical society B*, Vol. 26, No. 2, p. 161-175.
- Aitchison, J. y Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: University Press.
- Anderson, T. W. (1958) *An introduction to multivariate statistical analysis*. New York: Wiley.
- Ash, R.B. (1972) *Real analysis and probability*. New York: Academic press.
- Balakrishnan, N. y Cohen, A. C. (1991). *Order statistics and inference: estimations methods*. Academic Press: London.
- Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población bivalente. *Trabajos de estadística*, **32**, p. 18-31.
- Bayes, T. R. (1763), "An essay towards solving a problem in the doctrine of chances", *Philosophical Transactions of the Royal Society*, **53**, 370-418. Reprinted in *Biometrika*, **45**, 243-315, 1958.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, New York: Springer-Verlag. 2nd edition.
- Berger, J. O. (2006). The case for objective Bayesian analysis. (with discussion) *Bayesian analysis*, **1**, No.3, p. 385-402 y 457-464.
- Berger J. O. y Bernardo, J. M.(1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of American Statistical Association*, Vol. 84, No. 405, 200-207.
- Berger, J. O. y Bernardo, J. M. (1992a). On the development of the Reference Prior Method. *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) Oxford, U.K.: Clarendon Press, p. 35-60.
- Berger, J. O. y Bernardo, J. M. (1992b). Ordered group reference priors with application to the multinomial model. *Biometrika* 79, p. 25-37.
- Berger, J. O. y Bernardo, J. M. (1992c). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics*, (P.K. Goel and N.S. Iyengar, eds.). Berlin: Springer, p. 323-340.
- Berger, J. O. y Sun, D. (1998). Reference priors with partial information. *Biometrika*, **85**, p. 55-71.

- Berger, J. O. y Sun, D. (2006). Objective priors for a bivariate normal model with multivariate generalizations. Technical Report 07-06, ISDS, Duke Univ.
- Berger, J. O. y Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics* **36**, p. 963-982.
- Berger, J. O., Bernardo, J. M. y Sun, D. (2009a). The formal definition of reference priors. *The Annals of Statistics* **37**.
- Berger, J. O., Bernardo, J. M. y Sun, D. (2009b). Reference priors for discrete parameter spaces. *Technical report*. University of Valencia.
- Berger, J. O., Bernardo, J. M. y Sun, D. (2009c). Natural induction: An objective Bayesian approach. *Rev. Acad. Sci. Madrid*, A 103, p. 125-159 (invited paper with discussion).
- Bernardo, J. M. (1978). Unacceptable implications of the left Haar measure in a standard normal theory inference problem. *Trabajos de Estadística y de Investigación Operativa*, Vol. 29.
- Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 41, p. 113-147.
- Bernardo, J. M. (1999). Nested hypothesis testing: the Bayesian reference criterion. *Bayesian Statistics 6: Proceedings of the Fourth Valencia International Meeting*. (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Eds.) Oxford, U.K.: Clarendon Press, p. 101-130.
- Bernardo, J. M. (2003). Bayesian statistics. *Encyclopaedia of life support systems. Probability and statistics*. (R. Viertl, ed.), Oxford, UK: UNESCO.
- Bernardo, J. M. (2005a). Reference Analysis. *Handbook of Statistics*, Vol. 25. Amsterdam: Elsevier, p. 17-90.
- Bernardo, J. M. (2005b). Intrinsic credible regions: an objective Bayesian approach to interval estimation. *Sociedad de estadística e investigación operativa, TEST*, Vol. 14, No. 2, p. 317-384.
- Bernardo, J. M. (2005c). Intrinsic point estimation of the Normal variance. *Bayesian statistics and its applications* (S.K. Upadhyay, U. Singh and D.K., Dey, eds.) New Delhi: Anamaya Pub. 2005, p. 110-121.
- Bernardo, J. M. y Juárez, M. (2003). Intrinsic estimation, *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds). Oxford: University press, p. 465-476.
- Bernardo, J. M. y Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinormal parameters. *J. Royal Statistical Society D* **47**, p. 101-135.
- Bernardo, J. M. y Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.*, **70**, 351-372.
- Bernardo, J. M. y Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: Wiley

- Box, G.E. P., y Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.
- Chang, I. H., Kim B. H. y Mukerjee, R. (2003) Probability matching priors for predicting unobservable random effects with application to ANOVA models. *Statistics and Probability Letters*, **62**, p. 223-228
- Chernoff, H. (1954) Rational selection of decision functions. *Econometrika*, **22**, p. 422-443
- Clarke, B. y Barron, A. (1994) Jeffreys prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, **41**, p. 36-60.
- Datta, G. S. y Ghosh, M. (1995a) On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 1, p. 37-45.
- Datta, G. S. y Ghosh, M. (1995b). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, p. 1357-1363.
- Datta, G. S. y Ghosh, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24**, p. 141-159.
- Datta, G. S., Mukerjee, R., Ghosh, M. y Sweeting, T. J. (2000) Bayesian prediction with approximate frequentist validity. *The Annals of Statistics*, Vol. 28, No. 5, 1414-1426
- Datta, G. S., Mukerjee, R. (2003) Probability matching priors for predicting a dependent variable with application to regression models. *Annals of the Institute of Statistical Mathematics*, **55**, 1, p. 1-6.
- Datta, G. S., Mukerjee, R. (2004). *Probability matching priors: higher order asymptotics. Lecture notes in Statistics*. New York: Springer.
- Datta, G. S. y Sweeting, T. J. (2005). Probability Matching Priors. *Handbook of Statistics, Vol. 25* Elsevier, p. 91-114
- Dawid, A.P. (1983). "Invariant prior distributions", *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N.L. Jonhson. New York: John Wiley, p. 228-236.
- Dawid, A.P., Stone, M. y Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B*, **35**, p. 189-233 (with discussion).
- DeGroot, M. H. (1973) Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, **68**, Numebr 44, p. 966-969.
- DeGroot, M. H. (1975) *Probability and Statistics*. California: Addison-Wesley.
- Eaton, M. L. y Sudderth, W. D. (1998) A new predictive distribution for normal multivariate linear models. *Sankhyā A : The Indian Journal of Statistics*, Vol. 60, p. 363-382.
- Eaton, M. E. y Sudderth, W. D. (1999) Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli*, **5**, p. 833-854.
- Efron, B. (1970) Comments on Blyth's paper. *Annals of Mathematical Statistics*, **41**, p. 1049-1054.

- Efron, B. (1973) Discussion of “Marginalization paradoxes in Bayesian and structural inference” by A. P. Dawid, M. Stone, and J. V. Zidek (1973). *Journal of the Royal Statistical Society*, Ser. B, **35**, 219.
- Everitt, B.S. (2002). *The Cambridge dictionary of statistics*, 2nd. Ed. Cambridge: University Press.
- Fernández, C. y Steel, F.J. M (1999). Reference priors for the general location-scale model. *Statistical and Probability Letters* 43. Elsevier, p. 377-384.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, Vol. 10, p. 507-521.
- Fogiel, M. (2002). *The advanced calculus problem solver*. New Jersey: Research and education association.
- Geisser, S. (1993). *Predictive inference: an introduction*. London: Chapman and Hall.
- Ghosal, S. (1997). Reference priors in multiparameter nonregular cases. *Test*, Vol. 6, No. 1, p. 159-186.
- Ghosal, S. y Samanta, T. (1997). Expansion of Bayes factor risk for entropy loss and reference prior in nonregular cases, *Statistics and Decisions*, No. 15, p. 129-140.
- Ghosh, J.K. y Mukerjee, R. (1992a). Non-informative priors (with discussion). In Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.). *Bayesian Statistics, vol 4*. Oxford: University Press, p. 195-210.
- Good, I. J. (1969) What is the use of a distribution? *Multivariate Analysis* (Krishnaiah, ed.) New York: Academic Press, p. 183-203.
- Good, I. J. (1969) A subjective evaluation of Bode’s law and an “Objective” test for approximate numerical rationality. *Journal of the American Statistical Association*, **64**, Num. 325, p. 23-49
- Good, I. J. y Crook, J. F. (1987) The robustness and sensitivity of the Mixed-Dirichlet Bayesian test for “independence” in contingency tables. *The Annals of Statistics* **15**, Num. 2, p. 670-693
- Gradshteyn, I.S. y Ryzhik, I.M (1994) *Table of integrals, series and products*, Fifth edition (Alan Jeffrey, ed.) San Diego CA: Academic Press.
- Hartigan, J. A. (1964). Invariant prior distributions. *Annals of Mathematical Statistics*, **35**, p. 836-845.
- Hartigan, J. A. (1965). The asymptotically unbiased prior distributions. *Annals of Mathematical Statistics*, **36**, p. 1137-1152.
- Hartigan, J. A. (1966). Note on the confidence-prior of Welch and Peers. *Journal of the Royal Statistical Society*, **B**, 28, 55-56.
- Hartigan, J. A. (1983). *Bayes theory*. New York: Springer-Verlag.

- Hartigan, J. A. (1998). The maximum likelihood prior. *The annals of statistics*, 26, p. 2083-2103.
- Hinkley, D.V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, 56, 3, p. 635-639.
- Hora, R.B. y Buehler, R.J. (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.* **37**, p. 643-656.
- Hora, R.B. y Buehler, R.J. (1967). Fiducial theory and invariant prediction *Ann. Math. Statist.* **38**, p. 795-801.
- Ilín, V. y Pozniak, E. (1980). *Fundamentos del análisis matemático*. Tomo III. Moscu: Mir.
- Jaynes E.T. (1968), Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, vol. sec-4, no. 3, pp.227-241
- Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London, Ser. A, Vol. 186*, p. 453-561.
- Jeffreys, H. (1961) *Theory of Probability* (Third Edition). Oxford: Clarendon Press.
- Jeffreys, H. (1963) Review of *The foundations of statistical inference* by L.J. Savage, M.S. Bartlett, G.A. Barnard, D.R. Cox, E.S. Pearson and C.A.B. Smith, *Technometrics*, **5**, p. 407-410.
- Jonhson, J. L. y Kotz, S. (1969) *Distributions in statistics: discrete distributions*. USA: Wiley
- Jonhson, J. L. y Kotz, S. (1970) *Continuous univariate distributions*. USA: Wiley
- Jonhson, J. L. y Kotz, S. (1972) *Distributions in statistics: continuous multivariate distributions*. New York: Wiley
- Jonhson, J. L., Kotz, S. y Balakrishnan, N. (1994) *Continuous univariate distributions*. Vol. 1, 2nd Edition. New York: Wiley.
- Jonhson, J. L., Kotz, S. y Balakrishnan, N. (1994) *Continuous univariate distributions*. Vol. 2, 2nd Edition. New York: Wiley.
- Kashyap, R. L. (1971). Prior probability and uncertainty. *IEEE Transactions on information theory*, IT-14, p. 641-650.
- Kass, R. E. (1982). A comment on "Is Jeffreys a 'Necessarist'?", *The American Statistician*, **36**, No. 4, p. 390-391.
- Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical science*, 4, p. 188-234.
- Kass, R. E.(1990). Data-translated likelihood and Jeffreys's rules, *Biometrika*, **77**, 1, p. 107-114.
- Kass, R. E. y Wasserman L. (1996). The Selection of prior distributions by formal rules. *Journal of the American Statistical Association*, Vol. 91, No. 435, p. 1343-1370.

- Komaki, F. (1996). On the asymptotic properties of predictive distributions, *Biometrika*, **83**, p. 299-313.
- Kuboki, H. (1998) Reference priors for prediction. *Journal of Statistical Planning and inference*, **69**, pp. 295-317.
- Laplace, P. S. (1814/1825). *Essai philosophique sur les probabilités*. Paris: Courcier (5th ed.). English translation in 1952 as *Philosophical essay on probabilities*. New York: Dover.
- Levine, R.A. y Casella, G. (2003). Implementing matching priors for frequentist inference. *Biometrika* **30**, pp. 127-137.
- Lindley, D. V. (1958). Fiducial distributions and Bayes's theorem. *Journal of the Royal Statistical Society*, Ser. B, 20, p. 102-107.
- Lindley, D. V. (1965). *Introduction to probability and Statistics from a Bayesian viewpoint, part 2, Inference*. London: Cambridge University Press.
- Mardia, K. V., Kent J. T. y Bibby, J. M. (1979) *Multivariate analysis* London: Academic Press.
- Pérez-Elizalde, S. (2005), *Métodos Bayesianos objetivos de comparación de medias*, PhD. Thesis, Universitat de València, Valencia, Spain.
- Pham-Gia, T., Turkkan, N. y Marchand, E. (2006) Density of the ratio of normal variables and applications. *Communications in statistics: theory and methods.*, 35, p. 1569-1591.
- Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The annals of Statistics*, Vol. 11, No. 11, p. 416-431.
- Robert, C.P. (1996). Intrinsic losses. *Theory and decision*, **40**, p. 191-214.
- Robert, C.P., Chopin, N. y Rousseau, J. (2008). Harold Jeffreys' Theory of probability revisited, *Statistical Science* (to appear, with discussion).
- Savage, L. J. (1962) "Bayesian Statistics", in *Recent developments in information and decision theory*, eds. R.F. Machol and P. Gray, New York: Macmillan, reprinted in the *The writings of Leonard Jimmie Savage - A memorial selection* (1981), Washington DC: American Statistical Association and Institute of Mathematical Statistics.
- Seidenfeld, T. (1987). Entropy and uncertainty. *Foundations of statistical inference*, eds. I.B. MacNeill and G.J. Umphrey, Boston: D. Reidel Publishing company. p. 259-287
- Selby, S. M. (1972) *Standard Mathematical Tables* (twentieth edition). Ohio: The Chemical Rubber.
- Severini, T. A., Mukerjee, R. y Ghosh, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika* **89**, p. 952-957.
- Shafer, G. (1976). *A mathematical theory of evidence*, Princeton, NJ: Princeton University Press.

- Smith, R. L. (1999). Bayesian and frequentist approaches to parametric predictive inference. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, pp.589-612, (with discussion).
- Stigler, S. M. (1982), Thomas Bayes's Bayesian Inference, *J.R.Statist. A*, **145**, p. 250-258.
- Stigler, S. M. (1986a), *The history of Statistics: the measurement of uncertainty before 1900*. Cambridge MA: The Belknap Press of Harvard University.
- Stigler S. M. (1986b), Laplace's 1774 memoir on inverse probability, *Statistical Science*, Vol. 1, **3**, pp. 359-378.
- Stone, M. (1976) Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, Vol. 71, N. 353, p. 114-116.
- Stone, M. y Dawid, A.P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems, *Biometrika*, **59**, No. 2, pp.369-375.
- Sweeting, T.J. (2007). On predictive probability matching priors. *Technical report* No. 278, University College London.
- Thatcher, A.R. (1964). Relationships between Bayesian and confidence limits for predictions. *J. Roy. Statist. Soc., Ser. B* **26**, p. 176-210.
- Tibshirani, R.J. (1989). Noninformative priors for one parameter or many. *Biometrika* **76**, p. 604-608.
- Villegas, C. (1971). On Haar priors. *Foundation of statistical inference* (V.P. Godambe and D.A. Sprott, eds.). Toronto: Holt, Rinehart and Wiston, p. 409-414 (with discussion).
- Villegas, C. (1981). Inner statistical inference II. *Ann. Statist.*, **9**, p. 768-776.
- Welch, B. L. (1939). On the distribution of maximum likelihood estimates. *Biometrika*, **31**, p. 187-190.
- Welch, B.L. y Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc., Ser. B* **35**, p. 318-329.
- Yang, R. y Berger, J.O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Atatistics*, Vol. 22, No. 3, p. 1195-1211.
- Zellner, A. (1971a) Bayesian and non-bayesian analysis of the Log-Normal distribution and Log-Normal regression. *Journal of the American Statistical Association*, **66**, Num. 334, pp. 327-330
- Zellner, A. (1971b). *An introduction to Bayesian inference in econometrics*. New York: John Wiley.
- Zellner, A. (1977). Maximal data information prior distributions. *New Developments in the Applications of Bayesian Methods* (A. Aykac and C. Brumat, eds.). Amsterdam: North-Holland, p. 201-215.

Zellner, A. (1982a) Is Jeffreys a “Necessarist”? *The American Statistician*, **86**, No.1, p. 28-30.

Zellner, A. (1982b), Reply to A comment on “Is Jeffreys a ‘Necessarist’?”, *The American Statistician*, **36**, No. 4, p. 392-393.