CIMAT Centro de Investigación en Matemáticas, A.C.

DISTRIBUCIONES INFERENCIALES PARA ESTIMACIÓN DE DENSIDADES PREDICTIVAS Y PONDERACIÓN DE HIPÓTESIS

T E S I S

Que para obtener el grado de

Maestro en Ciencias

con Especialidad en

Probabilidad y Estadística

Presenta

Vidal Alí González Cucurachi

Director de Tesis:

Dr. Rolando José Biscay Lirio

Autorización de la versión final

Guanajuato, Gto., Febrero de 2018

Centro de Investigación en Matemáticas, A.C.

Acta de Examen de Grado

Acta No.: 127

Libro No.: 002

Foja No.: 127

En la Ciudad de Guanajuato, Gto., siendo las 15:00 horas del día 2 de febrero del año 2018, se reunieron los miembros del jurado integrado por los señores:

DRA. ELOÍSA DÍAZ-FRANCÉS MURGUÍA (CIMAT)
DR. EDUARDO ARTURO GUTIÉRREZ PEÑA (IIMAS-UNAM)
DR. ROLANDO JOSÉ BISCAY LIRIO (CIMAT)

bajo la presidencia del primero y con carácter de secretario el segundo, para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

Sustenta

VIDAL ALÍ GONZÁLEZ CUCURACHI

en cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

"DISTRIBUCIONES INFERENCIALES PARA ESTIMACIÓN DE DENSIDADES PREDICTIVAS Y PONDERACIÓN DE HIPÓTESIS "

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a):

APROBADO

DRA. ELOÍSA DÍAZ-FRANCÉS MURGUÍA
Presidente

DR. EDUARDO ARTURO GUTIÉRREZ PEÑA
Secretario

DR. ROLANDO JOSÉ BISCAY LIRIO
Vocal

CIMAT
DIRECCIÓN GENERAL

DR. VÍCTOR MANUEL RIVERO MERCADO
Director General

Agradecimientos

Este trabajo se lo dedico a mis padres, Vidal González y Ana Lilia Cucurachi. Muchas gracias papás por todo lo que me han dado, por construir con lecciones, consejos y a veces regaños la persona que ahora soy, los amo. Agradezco a mi hermana Chuy porque a pesar de los años y la distancia, siempre me has demostrado tu apoyo y amor de hermanita mayor.

Natalia, mi Nat, muchas gracias por darme los ánimos en el proceso paso a paso de esta tesis, en esas noches de desvelo y esas tardes de frustación frente a la compu. Gracias por siempre alentarme a ver hacia adelante y preferir el camino hacia crecer, gracias por hacer de nuestra relación el mejor equipo, te amo.

A la Sra. Herminia y al Sr. Ramón les agradezco por acogerme en su hogar, por darme un consejo cuando es necesario, por hacerme sentir como en casa estando lejos de mi tierra, han sido partícipes del logro que representa este trabajo.

Agradezco a mi asesor, el Dr. Rolando Biscay, por confiar en mí e instruirme de forma estupenda para este trabajo. También le agradezco me haya compartido de su gran conocimiento y experiencia estadística a través de las horas de discusión en su oficina.

Gracias a mis sinodales, la Dra. Eloísa Días-Francés y el Dr. Eduardo Gutierrez por sus valiosas opiniones y sugerencias a este trabajo. Así también les agradezco su paciencia y dedicación para la revisión del mismo. En especial agradezco a la Dra. Eloísa, a usted le debo el despertar mi interés por la probabilidad y estadística, gracias a sus cursos y a sus ayudantías me adentré en este mundo y ahora le dedico mi día a día.

Agradezco al CIMAT y al CONACYT por los apoyos económicos que me otorgaron a lo largo

de la maestría. Al CIMAT también le agradezco su hospitalidad y la oportunidad de formar parte de una de las mejores comunidades científicas del país.

Resumen

Para modelos estadísticos paramétricos, la construcción de distribuciones de probabilidad sobre el espacio de parámetros que son dependientes de los datos disponibles, teniendo la forma específica de distribuciones posteriores, constituye un instrumento básico de los métodos inferenciales Bayesianos. Más generalmente, Hirotugu Akaike introdujo el término distribución inferencial (DI) para referirse a cualquier distribución sobre el espacio de parámetros dependiente de datos. Cualquier DI puede utilizarse como una distribución mezcladora o de ponderación probabilística aplicada sobre la distribución paramétrica del modelo, para obtener una estimación H de la distribución de m datos futuros que sigan la misma distribución F que los n datos de la muestra original disponible. Sobre esta base, Akaike propuso la esperanza de la divergencia de Kullback-Leibler de H con respecto a F (que denotamos $KL(F, H)$) como un criterio de calidad de una DI. Sin embargo, Akaike se limitó a considerar DI que son distribuciones posteriores con respecto a distribuciones previas especificadas. En la presente tesis se extiende este enfoque de Akaike en dos direcciones principales. 1) Se proponen métodos para construir DI sobre la base de la minimización de versiones empíricas del criterio $KL(F, H)$. Esto provee un enfoque no Bayesiano (o "frecuentista") para construir DI con clara interpretación estadística. 2) Además, se consideran espacios de parámetros estructurados, en el sentido de consistir en una familia finita de espacios de parámetros regulares. Esto incluye en especial el caso de hipótesis múltiples (simples o compuestas), lo que permite ofrecer un enfoque no Bayesiano para la ponderación probabilística de hipótesis. Ambas líneas de trabajo se integran en las propuestas elaboradas en esta tesis. Estas se formulan matemáticamente, se comentan sus fundamentos estadísticos y se ilustran sus comportamientos a través de simulaciones computacionales en diversos ejemplos simples.

Contenido

Capítulo 1

Introducción

Sea una muestra $x = (x_1, .., x_n)$ consistente en variables aleatorias x_i independientes e idénticamente distribuidas (i.i.d) con valores en un espacio $\mathcal{X}$. Supóngase que se especifica un modelo estadístico paramétrico para la densidad $f(x)$ de la muestra x, a través de una familia de densidades $\{f(\cdot;\theta) : \theta \in \Theta\}$. Aquí, Θ es el espacio de valores posibles del parametro θ, y $f(\cdot;\theta)$ es una función de densidad sobre $\mathcal{X}$ para cada $\theta \in \Theta$ (con respecto a cierta medida sobre $\mathcal{X}$; típicamente, la medida de Lebesgue). En caso de modelos estadísticos bayesianos (Bernardo y Smith (2000)), se especifica además una densidad previa (o *a priori*) $f(\theta)$ sobre el espacio de parámetros Θ. Esto permite definir la densidad posterior

$$f(\theta \mid x) = \frac{f(x;\theta)\, f(\theta)}{\int f(x;\theta)\, f(\theta)\, d\theta}.$$

Tal densidad posterior $f(\theta \mid x)$ es una densidad sobre el espacio de parámetros que depende de los datos, y constituye el ingrediente básico de los métodos de inferencia bayesianos (Bernardo y Smith (2000)). En particular, supóngase que interesa estimar la densidad $f(z)$ de una muestra futura $z = (z_1, .., z_m)$ consistente en una muestra i.i.d. de variables z_i con la misma densidad que los datos x_i. La estadística bayesiana ofrece una solución inmediata a este problema mediante la llamada densidad predictiva, definida como

$$h(z \mid x) = \int f(z;\theta)\, f(\theta \mid x)\, d\theta.$$

Notar que en esta fórmula la densidad predictiva $h(z \mid x)$ (es decir, la estimación de la densidad $f(z)$ de una muestra futura) se obtiene integrando cada densidad $f(z;\theta)$ del modelo multiplica-

da por una ponderacion probabilística o distribución mezcladora dada por la densidad posterior $f(\theta \mid x)$.

Más generalmente, H. Akaike (ver Akaike (1977), Akaike (1978)) llamó una *densidad inferencial* $q(\theta; x)$ a cualquier densidad sobre el espacio de parámetros dependiente de la muestra x. La densidad posterior asociada a un modelo bayesiano es pues un caso particular de densidad inferencial. Akaike señala que cualquier distribución inferencial $q(\theta; x)$ puede ser utilizada como densidad mezcladora para obtener una estimación $h(z; x)$ de la densidad $f(z)$ de una muestra futura z, mediante

$$h(z; x) = \int f(z; \theta)\, q(\theta; x)\, d\theta.$$

Además, Akaike propuso utilizar como criterio de calidad de una distribución inferencial o mezcladora $q(\theta; x)$ la esperanza de la divergencia de Kullback-Leibler de la mezcla $h(z; x)$ con respecto a la verdadera distribución $f(z)$ de la muestra futura:

$$\mathbb{E}\left(KL\left(f, h(\cdot; x)\right)\right) = \mathbb{E}\left(\int f(z) \ln\left(\frac{f(z)}{h(z; x)}\right) dz\right),$$

donde la esperanza es con respecto a la densidad $f(x)$ de la muestra x.

Es importante resaltar que este criterio permite darle un claro sentido probabilístico, tanto en modelos bayesianos como no bayesianos, a cualquier densidad inferencial o mezcladora: una mezcladora $q(\theta; x)$ se considera una ponderanción del elemento $f(\cdot; \theta)$ del modelo que resulta tanto más adecuada cuanto menor sea la divergencia esperada $\mathbb{E}\left(KL\left(f, h(\cdot; x)\right)\right)$.

La utilidad de distribuciones inferenciales o mezcladoras generales, no necesariamente bayesianas (densidades posteriores), para la construcción de densidades predictivas mediante mezclas, ha tenido fundamento teórico adicional por diversos resultados de optimalidad (ver Brown et al. (2008)). Sin embargo, Akaike sólo aplicó este criterio al caso particular de densidades inferenciales dadas por densidades posteriores (es decir, en caso de modelos bayesianos), y no propuso métodos para la construcción de densidades inferenciales no bayesianas. El uso de densidades inferenciales no bayesianas, sea para la obtención de densidades predictivas u otras aplicaciones, ha permanecido prácticamente no desarrollado en la literatura estadística. La construcción de densidades predictivas no bayesianas suele hacerse mediante enfoques que no involucran mezclas (Bjornstad

(1990)). Una excepción es la propuesta de Harris (1989) de utilizar una estimación de la densidad del estimador máximo verosímil (EMV) como distribución inferencial para la obtención de densidades predictivas. Pero esta propuesta tiene la limitación de que requiere existencia del estimador máximo verosímil y conocimiento de la forma de su densidad, además no se deriva de un criterio de optimalidad que la fundamente.

Es propósito subyacente de este trabajo de tesis elaborar y evaluar métodos para la construcción de densidades inferenciales no bayesianas sobre la base de un claro criterio de calidad. Para ésto, la idea clave del enfoque que se propone en esta tesis consiste en minimizar, con respecto a la distribución inferencial $q(\theta; x)$, una version empírica del criterio $\mathbb{E}(KL(f, h(\cdot; x)))$, donde $h(z; x)$ es la mezcla resultante de integrar el modelo con respecto a la mezcladora $q(\theta; x)$.

Se centrará la atencion en el caso de modelos estructurados, en el sentido de que el espacio de parámetros Θ está particionado en una familia finita de subespacios o hipótesis $H_0, \ldots, H_{k-1}$. Para ejemplificar tales modelos estructurados, una de las situaciones mas sencillas es la de evaluar una hipótesis simple *versus* una alternativa simple $H_0{:}\theta = \theta_0$ *vs.* $H_0{:}\theta = \theta_1$, para cierto parámetro de interes θ. Otra clase de ejemplos lo constituye el problema de dos hipótesis $H_0{:}\theta \in \Theta_0$ *vs.* $H_0{:}\theta \in \Theta_1$, quizas alguna de ellas compuesta, para un parámetro de interes θ. Para este tipo de situación, se construirán distribuciones inferenciales $q(\cdot; x)$ sobre el conjunto $\{0, 1, .., k-1\}$; es decir, la distribución inferencial puede interpetarse como una ponderación probabilística $q(j; x)$ de cada hipótesis H_j, $j = 0, 1, .., k-1$. El método de construcción que se propone en este documento se basará en minimizar, con respecto a la distribución inferencial $q(\cdot; x)$, una versión empírica del criterio $\mathbb{E}(KL(f, h(\cdot; x)))$, donde $h(z; x)$ es la mezcla resultante de integrar el modelo con respecto a la mezcladora $q(; x)$. De este modo, se ofrece un nuevo enfoque para la ponderación de hipótesis, que tiene claro sentido probabilístico de acuerdo al criterio de optimalidad $\mathbb{E}(KL(f, h(\cdot; x)))$, tanto en modelos no bayesianos como bayesianos

Desde la perspectiva bayesiana, esta problemática se aborda en general mediante probabilidades posteriores de hipótesis. Por ejemplo, en Berger y Sellke (1987) se describe el procedimiento de calcular $\mathbb{P}(H_0 \mid x)$, y dependiendo del valor de esta probabilidad (dependiente de una muestra observada) se toma la decisión de aceptar o rechazar H_0. No obstante, se mencionan algunos obstáculos de este procedimiento, que recaen principalmente en el de escoger una distribución

previa adecuada, lo cual en general no es del todo sencillo. En Bernardo y Rueda (2002), para superar la dificultad de especificar una distribución previa subjetiva, se propone el Criterio de Referencia Bayesiano (BRC por sus siglas en Inglés), el cual considera una distribucion previa objetiva o no informativa. Sobre la base del valor que tome la denominada discrepancia intrínseca, respecto a cierto umbral, se toma la decisión de rechazar la hipótesis a prueba. Si bien, el BRC soluciona la problemática de elegir una distribución previa subjetiva, la escala en la que se mide la discrepancia intrínseca no tiene una interpretación general expedita.

Desde un punto de vista no bayesiano, algunos han hecho uso del p-valor para la ponderación de hipótesis. Pero este enfoque ha estado en discusión desde que fue propuesto, ha sido ampliamente criticado, y es consenso actual considerarlo no aceptable. Uno de los mas recientes artículos sobre el tema es Wasserstein y Lazar (2016), donde por primera vez la ASA (siglas en Inglés de Asociación Estadística Americana) expone su postura sobre el uso del p-valor. En general, todas las críticas fundamentan que el p-valor por sí solo no representa un buen instrumento para la ponderación de hipótesis.

Desde un punto de vista también no bayesiano, como alternativa se ha propuesto el uso de la verosimilitud relativa como medida de plausibilidad de una hipótesis, como se expone el Capítulo 9 de Edwards (1992) y en Pawitan (2001), por ejemplo. Sin embargo, las unidades en las que se mide la verosimilitud relativa no tienen una interpretación sencilla, lo que difculta su uso para ponderar una hipótesis de un modo que tenga claro sentido probabilístico.

El tema desarrollado en este trabajo tiene conexión con algunos otros trabajos previos. En Lindsay (1983) se aborda el enfoque de mezcla de modelos paramétricos, dándole el mayor énfasis en la estimación por máxima verosimilitud de dicha mezcla y en la demostración de algunas de las propiedades del estimador. Este artículo es un precursor de la utilización de mezclas, no obstante, por sí solo dicho trabajo no presenta un objetivo particular en la estimación del modelo mezclado. Por otro lado, en Walker et al. (2001) se presenta un enfoque de teoría de decisión para mezcla de modelos. En este artículo se propone un criterio bayesiano, basado en la maximización de una función de utilidad, para la obtención de la mezcla que mejor estime la densidad predictiva. Por su parte, en Gutiérrez-Peña y Walker (2005) se describe un procedimiento vía Teoría de decisión bayesiana y estimación no-paramétrica de la densidad predictiva para evitar incoherencia en la

verificación de modelos bayesianos. Dicho criterio tiene como elemento escencial la maximización de una función de utilidad de una mezcla de modelos no-paramétricos. La distinción principal del trabajo realizado en esta tesis respecto a los mencionados en este párrafo es la de presentar un método con interpretación probabilística clara y sencilla para la estimación de la densidad predictiva mediante una mezcla de modelos, no restringido a modelos bayesianos.

Resumiendo, a partir de lo dicho anteriormente, este trabajo de tesis se plantea los **objetivos** siguientes:

1.Elaborar un enfoque general para la construcción de densidades inferenciales o mezcladoras para modelos estructurados, sobre la base de la minimización de versiones empíricas de la esperanza de la divergencia de Kullback-Leibler de la mezcla (densidad predictiva) correspondiente con respecto a la verdadera densidad de una muestra futura.

2. Estudiar mediante simulaciones el comportamiento de los métodos propuestos, y compararlos con enfoques alternativos.

El documento de tesis está estructurado del modo siguiente. En el Capítulo 2 se considera el caso de hipótesis simple *versus* alternativa simple. Primeramente se da una breve descripción del problema predictivo. A continuación, se exponen dos algoritmos propuestos para la construcción de distribuciones inferenciales en esta situación, conduciendo a correspondientes estimaciones de la densidad de una muestra futura. Se presentan resultados de simulaciones realizadas para la evaluación de estos métodos en el caso Normal, y se explora su relación con la distribución del estimador por máxima verosimilitud. En el Capítulo 3 se elabora un enfoque para construcción de distribuciones inferenciales para el caso de dos hipótesis, enfocado principalmente al caso de una hipótesis simple *versus* alternativa compuesta. Además se expone la aplicación de este enfoque a la inferencia sobre el parámetro de forma de la distribución de valores extremos generalizada (DVEG). Por último, se presenta una breve discusión sobre las ventajas y desventajas del criterio propuesto, y se dan sugerencias sobre posibles líneas de investigación a seguir a partir de los resultados obtenidos. Finalmente, se enuncian conclusiones de este trabajo de tesis.

Capítulo 2

Modelos paramétricos con dos componentes simples

Se tiene una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$ con distribución $f(\bullet; \theta_0) \in \{f(\bullet; \theta) : \theta \in \Theta\}$. Se considera una distribución inferencial sobre el espacio de parámetros $q(\theta; x)$, $\theta \in \Theta$. Esta distribución inferencial induce una mezcla de modelos, $h(z; x) = \int f(z; \theta) q(\theta; x) d\theta$. Como se ha mencionado, con el enfoque bayesiano resulta natural la estimación de la densidad predictiva, la distribución inferencial comunmente es obtenida como una distribución posterior, para así obtener una estimación de $f(z \mid x)$ con la mezcla h, para mayor detalle de este enfoque se recomienda al lector consultar Bernardo y Smith (2000). Esta notación se seguirá de manera estándar a lo largo de todo el escrito. Para facilitarla, $f(z; \theta_0)$ la denotaremos simplemente como $f(z)$, teniendo siempre en mente que es un modelo paramétrico con valor real θ_0 del parámetro θ.

Obviamente, es estadísticamente deseable que la mezcla $h(z; x)$ sea una buena aproximación a la densidad verdadera $f(z)$ de datos futuros z. Sobre esta base, en Akaike (1977) y Akaike (1978) se hace uso de la esperanza de la *Divergencia de Kullback-Leibler* (KL) de la mezcla h con respecto a f como criterio de optimalidad y utilizan este criterio para medir el ajuste de la estimación bayesiana de densidades predictivas, es decir, de mezclas mediante distribuciones inferenciales que son distribuciones posteriores. En Brown et al. (2008), bajo este mismo criterio, se determina la mejor distribución inferencial en el caso en el que el modelo de los datos observados es normal multivariado. Bajo estos precedentes, se propone en esta tesis utilizar la divergencia de Kullback-

Leibler, para definir un criterio general de calidad de la mezcla obtenida mediante una distribución inferencial especificada.

La divergencia de Kullback-Leibler, como se define en Cover y Thomas (2006), de una densidad $g(x)$ respecto a otra $f(x)$, se define como

$$\begin{aligned} KL(f,g) &= \int f(x) \ln\left[\frac{f(x)}{g(x)}\right] dx \\ &= \int f(x) \ln\left[f(x)\right] dx - \int f(x) \ln\left[g(x)\right] dx. \end{aligned}$$

Por tanto, el criterio de calidad de una mezcla mencionado arriba, es decir, la esperanza de la divergencia de Kullback-Leibler de la mezcla h con respecto a la densidad f, dada por

$$\mathbb{E}\left[KL(f, h(\bullet; x))\right] = \int f(x) \left\{ \int f(z) \ln\left[\frac{f(z)}{h(z;x)}\right] dz \right\} dx. \tag{2.1}$$

Akaike utiliza este criterio en el caso particular en el que la mezcla $h(z;x)$ resulta de una distribución posterior, es decir, la distribución posterior toma el papel de mezcladora. No obstante, la propuesta de esta tesis es construir una distribución inferencial general de tal forma que al ser utilizada como mezcladora, la mezcla h minimice el valor medio de KL respecto a la verdadera densidad $f(z)$. Esta divergencia no es una distancia entre densidades (no es simétrica), pero es una medida de desviación muy utilizada en Estadística (Kullback y Leibler (1951), Cover y Thomas (2006), Lovric (2011)). A continuación se enuncian algunas propiedades de la divergencia de Kullback-Leibler.

- **Premétrica**. $KL(f,g) \geq 0$; además $KL(f,g) = 0$ si y sólo si $f = g$ casi seguramente respecto a la medida generadora de f.

- **Convexidad**. Para densidades f_1, f_2, g_1, g_2 y $\alpha \in [0,1]$, se tiene que

$$KL[\alpha f_1 + (1-\alpha) f_2, \alpha g_1 + (1-\alpha) g_2] \leq \alpha KL(f_1, g_1) + (1-\alpha) KL(f_2, g_2).$$

- **Relación con la logverosimilitud**. Dada una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$ en donde cada x_i se distribuye independientemente de acuerdo a una densidad $f(\bullet; \theta_0)$, la logversimilitud está dada por

$$l(\theta; x) = \sum_{i=1}^{n} \ln\left[f(x_i; \theta)\right].$$

Por la Ley de Grandes Números, se tiene que

$$\frac{1}{n}l(\theta;x) = \frac{1}{n}\sum_{i=1}^{n}\ln\left[f(x_i;\theta)\right] \to \mathbb{E}_{\theta_0}\{\ln\left[f(x;\theta)\right]\}.$$

Por otra parte, se tiene que

$$\begin{aligned} KL\left[f(\bullet;\theta_0), f(\bullet;\theta)\right] &= \int f(x;\theta_0)\ln\left[\frac{f(x;\theta_0)}{f(x;\theta)}\right]dx \\ &= \int f(x;\theta_0)\ln\left[f(x;\theta_0)\right]dx - \int f(x;\theta_0)\ln\left[f(x;\theta)\right]dx \\ &= c - \mathbb{E}_{\theta_0}\{\ln\left[f(x;\theta)\right]\}, \end{aligned}$$

donde c es una constante no dependiente de θ. De aquí que minimizar $KL\left[f(\bullet;\theta_0), f(\bullet;\theta)\right]$ con respecto a θ es aproximadamente equivalente a maximizar la logverosimilitud $l(\theta;x) \approx n\mathbb{E}_{\theta_0}\{\ln\left[f(x;\theta)\right]\}$. O sea, la logverosimilitud puede considerarse una versión empírica de

$$-nKL\left[f(\bullet;\theta_0), f(\bullet;\theta)\right],$$

salvo una constante aditiva.

Desde el punto de vista de estimación puntual por sustitución, dado un estimador $\widehat{\theta}$ de θ, se tiene una estimación de $f(z;\theta)$ dada por $f(z;\hat{\theta})$. Aquí, $f(z;\hat{\theta})$ puede verse como una mezcla particular, en donde la mezcladora es una medida delta de Dirac centrada en $\hat{\theta}$. Por tanto, en este caso el criterio de calidad (2.1) se reduce a

$$\mathbb{E}_{\hat{\theta}}\left\{KL\left[f(\bullet), f(\bullet;\widehat{\theta})\right]\right\} = k - \int g(\theta)\left[\int f(z)\ln\left(f(z;\theta)\right)dz\right]d\theta,$$

donde $f(\bullet)$ es la densidad verdadera de z, $k = \int g(\theta)\{\int f(z)\ln\left[f(z)\right]dz\}d\theta$ es una constante y $g(\bullet)$ es la densidad de $\hat{\theta}$. Por tanto, minimizar $\mathbb{E}_{\hat{\theta}}\left\{KL\left[f(z), f(z;\widehat{\theta})\right]\right\}$ es equivalente a maximizar $\int g(\theta)\left[\int f(z)\ln\left(f(z;\theta)\right)dz\right]d\theta$. Además, por la desigualdad de Jensen

$$\begin{aligned} \mathbb{E}_{\hat{\theta}}\left\{KL\left[f(z), f(z;\widehat{\theta})\right]\right\} &= k - \int g(\hat{\theta})\left[\int f(z)\log\left(f(z;\widehat{\theta})\right)dz\right]d\hat{\theta} \\ &= k - \int f(z)\left[\int g(\hat{\theta})\log\left(f(z;\widehat{\theta})\right)d\hat{\theta}\right]dz \\ &= k - \mathbb{E}_Z\{\mathbb{E}_{\hat{\theta}}\left[log\left(f(\bullet;\hat{\theta})\right)\right]\} \\ &\geq k - \mathbb{E}_Z\left[\log\left(\mathbb{E}_{\hat{\theta}}f(z;\widehat{\theta})\right)\right] \\ &= \mathbb{E}_Z\left\{KL\left[f(z), \mathbb{E}_{\hat{\theta}}f(z;\widehat{\theta})\right]\right\}, \end{aligned}$$

donde,

$$\mathbb{E}_{\hat{\theta}} f(z;\widehat{\theta}) = \int_{\Theta} f(z;\theta) g(\theta) d\theta,$$

es una densidad para z. De lo que se deduce que es mejor considerar una mezcla de las distribuciones $f(z;\theta)$ mediante la densidad g de $\hat{\theta}$, en lugar de tomar una estimación puntual para estimar la densidad $f(z)$ de datos futuros z.

Esto sugiere la idea, propuesta por Harris (1989), de usar la distribución del EMV como una mezcladora. Pero en la práctica, tal distribución g es desconocida, pues depende del verdadero valor θ_0 del parámetro. Esto nos motiva a considerar una distribución inferencial general, la cual es una distribución dependiente de datos sobre el espacio de parámetros, que funja como mezcladora del modelo estadístico para obtener una mezcla óptima, es decir, en nuestro contexto, que minimicie el valor esperado de la divergencia de Kullback-Leibler de la mezcla respecto a la verdadera densidad $f(z)$.

Más específicamente, se busca $q(\theta;x)$ de tal manera que $h(z;x) = \int_{\Theta} f(z;\theta) q(\theta;x) d\theta$ sea, respecto al valor esperado de la divergencia de Kullback-Leibler, la mejor estimación de $f(z)$. Es decir, estamos interesados en obtener $\widehat{q}$ que minimice el valor medio de KL, que como ya mencionamos, es lo mismo que

$$\widehat{q} = \arg\max_{q} \mathbb{E}_Z \left[\ln \left(\int_{\Theta} f(z;\theta) q(\theta;x) d\theta \right) \right]. \tag{2.2}$$

Debido a que en general la distribución verdadera $f(z)$ es desconocida, se considerará la versión bootstrap de la expresión anterior,

$$\widehat{q} = \arg\max_{q} \mathbb{E}^* \left[\ln \left(\int_{\Theta} f(x^*;\theta) q(\theta;x) d\theta \right) \right]. \tag{2.3}$$

En el presente capítulo se elaborará esta propuesta para el caso en el que el espacio de parámetros tiene sólo dos valores $\Theta = \{\theta_0, \theta_1\}$. Pero los algoritmos desarrollados se pueden extender fácilmente al caso en el que el espacio parametral tiene un número finito de elementos. A continuación se describirán los distintos algoritmos para la obtención de $\widehat{q}$ en el caso específico en el que $\Theta = \{\theta_0, \theta_1\}$.

2.1. Métodos

Supóngase que se tiene una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$, de la variable $X \sim f(\bullet; \theta)$, $\theta \in \Theta = \{\theta_0, \theta_1\}$. Se busca $\widehat{q} = \arg \max_q \mathbb{E}^* \left[\ln \left(\int_\Theta f(x^*; \theta) q(\theta; x) d\theta \right) \right]$, lo cual en este caso se traduce a obtener π_{opt} tal que

$$\begin{aligned} \pi_{opt} &= \arg \max_{\pi} \mathbb{E}^* \left[\ln \left(\pi f(x^*; \theta_0) + (1 - \pi) f(x^*; \theta_1) \right) \right] \\ &= \arg \max_{\pi} \varphi(\pi). \end{aligned}$$

donde $\varphi(\pi) = \mathbb{E}^* \left[\ln \left(\pi f(x^*; \theta_0) + (1 - \pi) f(x^*; \theta_1) \right) \right]$.

Nótese que π caracteriza a la distribución mezcladora dada por $q(\theta; x) = \pi \delta_{\theta_0}(\theta) + (1 - \pi) \delta_{\theta_1}(\theta)$. Como se mencionó en la introducción, en general se puede considerar predecir m datos futuros, donde $m \geq 1$. Se distinguirán a partir de aquí dos casos: cuando $m = 1$ y cuando $m > 1$. Para ambos casos en esta tesis se han propuesto dos tipos de métodos, que se diferencían entre ellos por el uso de una regularización sobre π. A continuación se expondrán los algoritmos de ambos métodos, para después describir los escenarios de simulación que se trabajaron para probar el desempeño de dichos algoritmos, así como mostrar los resultados de las simulaciones. Todo el trabajo de simulación se realizó en el software Matlab.

2.1.1. Método sin regularización

Primero es necesario hacer un comentario acerca de la notación que se utilizará. Para este método denotaremos el estimador de π_{opt}, como $\widetilde{\pi}_k$, donde el subíndice indica el número de observaciones que se están prediciendo. Es decir, $\widetilde{\pi}_1$ se refiere a la estimación de π_{opt}, bajo este método, para la densidad predictiva de un dato; en cambio, $\widetilde{\pi}_n$ denota la estimación para la densidad predictiva de n datos. Hacemos mención sólo de $\widetilde{\pi}_1$ y $\widetilde{\pi}_n$ debido a que las simulaciones están principalmente situadas en estos dos casos, sin embargo, algunas pruebas fueron realizadas para $\widetilde{\pi}_{n/2}$.

El algoritmo del método sin regularización para obtener $\widetilde{\pi}_1$ es el siguiente:

Algoritmo: Sin regularización para 1 dato futuro, caso de modelos paramétricos con dos componentes simples

1. Se obtienen B muestras bootstrap $x_1^*, x_2^*, \ldots, x_B^*$; donde $x_b^* = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$.

2. Para cada $b = 1, 2, \ldots, B$ se obtiene

$$\pi_b = \arg\max_{\pi} \sum_{j=1}^{n} \ln\left(\pi f(x_j^{*b}; \theta_0) + (1-\pi) f(x_j^{*b}; \theta_1)\right).$$

3. Obtenemos el estimador final de π_{opt} mediante

$$\widetilde{\pi}_1 = \frac{1}{B} \sum_{b=1}^{B} \pi_b.$$

Análogamente, para obtener $\widetilde{\pi}_m$, $m \in \{2, 3, \ldots, n\}$, se tiene que lo que se desea estimar es la mejor mezcla $\pi f_m(z; \theta_0) + (1-\pi) f_m(z; \theta_1)$ que aproxime la densidad $h(z_1, z_2, \ldots, z_m \mid x)$. El siguiente algoritmo obtiene la estimación $\widetilde{\pi}_m$.

Algoritmo: Sin regularización para m datos futuros, caso de modelos paramétricos con dos componentes simples

1. Se obtienen B muestras bootstrap $x_1^*, x_2^*, \ldots, x_B^*$; donde $x_b^* = (x_1^{*b}, x_2^{*b}, \ldots, x_m^{*b})$.

2. Y se obtiene,

$$\widetilde{\pi}_m = \arg\max_{\pi} \sum_{b=1}^{B} \ln\left(\pi f_m(x_b^*; \theta_0) + (1-\pi) f_m(x_b^*; \theta_1)\right).$$

En donde $f_m(x; \theta) = \prod_{i=1}^{m} f(x_i; \theta)$ si $x = (x_1, x_2, \ldots, x_m)$.

Estos dos algoritmos se implementaron en Matlab y se realizaron simulaciones para estudiar su comportamiento bajo distintos escenarios de simulación y para compararlos entre ellos. Los resultados de estas simulaciones se mostrarán más adelante. En la sección en la que se muestran dichos resultados, se puede corroborar que la estimación de π_{opt} es muy variable cuando se estima

mediante el método sin regularización. Con la intención de reducir la posibilidad de que la solución π_{opt} no sea única, se consideró un método alternativo mediante regularización sobre λ. A continuación se describen dos algoritmos que se distinguen de los recién expuestos por la consideración de una regularización sobre π.

2.1.2. Método con regularización

La entropía de Shannon para una variable aleatoria X que toma valores en un conjunto finito $\{x_1, x_2, \ldots, x_k\}$ se define como

$$\begin{aligned} \mathbb{H}(X) &= \mathbb{E}[I(X)] \\ &= \mathbb{E}[-\ln(\mathbb{P}(X))] \\ &= -\sum_{i=1}^{k} p_i \ln(p_i), \end{aligned}$$

donde $p_i = \mathbb{P}(X = x_i)$ e $I(X) := -\ln(\mathbb{P}(X))$ se puede ver como el contenido de información de X. Se tiene que $\mathbb{H}(X)$ se maximiza, respecto a los p_i, en $p_1 = p_2 = \ldots = p_k = 1/k$, es decir, cuando ningún valor x_i tiene mayor probabilidad de ocurrir que algún otro, por lo que la entropía de Shannon representa una medida de impredicibilidad de la variable X.

Recordemos que a π se le asocia la distribución mezcladora Bernoulli: $\pi = \mathbb{P}(\theta = \theta_0) = 1 - \mathbb{P}(\theta = \theta_1)$; osea, la mezcladora $\theta \sim Bern(\pi)$, donde $\theta \in \{\theta_0, \theta_1\}$. De esta caracterización surge la idea de considerar una nueva función objetivo $\varphi(\pi)$,

$$\varphi(\pi) = \sum_{j=1}^{n} \ln\left(\pi f(x_j; \theta_0) + (1-\pi) f(x_j; \theta_1)\right) + \lambda \mathbb{H}(Bern(\pi)),$$

donde $\mathbb{H}(Bern(\pi)) = -\pi \ln(\pi) - (1-\pi)\ln(1-\pi)$ es la entropía de una variable aleatoria con distribución Bernoulli con probabilidad π y $\lambda > 0$ es el parámetro de regularización. Nótese que $\mathbb{H}(Bern(\pi))$ funge como regularizador para $\sum_{j=1}^{n} \ln\left(\pi f(x_j; \theta_0) + (1-\pi) f(x_j; \theta_1)\right)$. Nótese también que el máximo de $\mathbb{H}(Bern(\pi))$ se alcanza en $\pi = 1/2$, como se comentó anteriormente.

A continuación se describe el algoritmo con regularización, tanto en el caso para predecir sólo 1 dato futuro, como $m > 1$.

Algoritmo: Con regularización para 1 dato futuro, caso de modelos paramétricos con dos componentes simples

1. Para una rejilla de valores de λ obtenemos, para $i = 1, 2, \ldots, n$

$$\widehat{\pi}_{\lambda}^{(i)} = \arg\max_{\pi} \sum_{j \neq i} \ln\left(\pi f(x_j; \theta_0) + (1-\pi) f(x_j; \theta_1)\right) + \lambda \mathbb{H}(Bern(\pi)).$$

2. Obtenemos $\widehat{\lambda}$,

$$\widehat{\lambda} = \arg\max_{\lambda} \sum_{j=1}^{n} \ln\left(\widehat{\pi}_{\lambda}^{(i)} f(x_j; \theta_0) + (1-\widehat{\pi}_{\lambda}^{(i)}) f(x_j; \theta_1)\right).$$

3. Obtenemos el estimador final de π_{opt},

$$\widehat{\pi}_1 = \arg\max_{\pi} \sum_{j=1}^{n} \ln\left(\pi f(x_j; \theta_0) + (1-\pi) f(x_j; \theta_1)\right) + \widehat{\lambda} \mathbb{H}(Bern(\pi)).$$

Algoritmo: Con regularización para m datos futuros, caso de modelos paramétricos con dos componentes simples

1. Se obtienen B muestras bootstrap $x_1^*, x_2^*, \ldots, x_B^*$; donde $x_b^* = (x_1^{*b}, x_2^{*b}, \ldots, x_m^{*b})$.

2. Para una rejilla de valores de λ obtenemos, para $b = 1, 2, \ldots, B$

$$\widehat{\pi}_{\lambda}^{(b)} = \arg\max_{\pi} \sum_{j \neq b} \ln\left(\pi f_m(x_j^*; \theta_0) + (1-\pi) f_m(x_j^*; \theta_1)\right) + \lambda \mathbb{H}(Bern(\pi)).$$

3. Obtenemos $\widehat{\lambda}$,

$$\widehat{\lambda} = \arg\max_{\lambda} \sum_{b=1}^{B} \ln\left(\widehat{\pi}_{\lambda}^{(b)} f_n(x_b^*; \theta_0) + (1-\widehat{\pi}_{\lambda}^{(b)}) f_n(x_b^*; \theta_1)\right).$$

4. Obtenemos el estimador final de π_{opt},

$$\widehat{\pi}_m = \arg\max_{\pi} \sum_{b=1}^{B} \ln\left(\pi f_m(x_b^*; \theta_0) + (1-\pi) f_m(x_b^*; \theta_1)\right) + \widehat{\lambda} \mathbb{H}(Bern(\pi)).$$

Como parte del trabajo que incluye esta tesis, se implementaron los algoritmos sin regularización, tanto para predecir un dato como para predecir m, además se implementó el algoritmo con regularización para la predicción de un dato. En particular, se realizaron simulaciones para probar los algoritmos en el caso de 1 y n datos, a su vez que algunas simulaciones se realizaron para el caso de predicción de $n/2$ datos. A continuación se describirán los resultados de simulaciones para ilustrar el comportamiento de los algoritmos y para hacer una comparación entre ellos.

2.2. Simulaciones

2.2.1. Comportamiento de cada estimador y comparación

Para estudiar los estimadores se realizaron simulaciones bajo el siguiente escenario. Recordemos que se tiene un espacio de parámetros $\Theta = \{\theta_0, \theta_1\}$. Considérese $f(\bullet; \theta_j) = N(\theta_j; \sigma)$ para $j = 0, 1$, donde $\theta_0 = 0$ y $\sigma = 1$. Se simuló una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$, donde $x_i \sim N(\theta_0, \sigma)$ con $\theta_0 = 0$, $\sigma = 1$ y $n = 70$. Un punto importante es que el tamaño de muestra $n = 70$ que se eligió fue con la intención de evaluar las estimaciones con tamaño de muestra pequeño, pero suficiente para obtener buenos resultados en el procedimiento bootstrap. En la Figura 1 (a)-(c) se muestran los histogramas de $\widetilde{\pi}_1$ bajo estas condiciones, considerando varios valores de θ_1. Además se muestra la estimación de $\widetilde{\pi}_1$ haciendo variar el valor de

$$\Delta_n = \sqrt{n}\frac{|\theta_0 - \theta_1|}{\sigma},$$

esto manteniendo fijos $n = 70$, $\theta_0 = 0$ y $\sigma = 1$, variando θ_1, equivalentemente, Δ_n. Como se puede apreciar en la Figura 1, para un valor de Δ_n pequeño, la estimación de $\tilde{\pi}_1$ se encuentra concentrada principalmente en 0 y 1, lo que sugiere ponderar uniformente entre $f(\bullet; \theta_0)$ y $f(\bullet; \theta_1)$. Sin embargo, conforme Δ_n crece, la estimación de $\tilde{\pi}_1$ se acerca al verdadero valor $\pi = 1$. Esto útimo se corrobora con la gráfica en la Figura 1 (d), haciendo crecer Δ_n, $\tilde{\pi}_1$ también crece aproximándose cada vez más a 1.

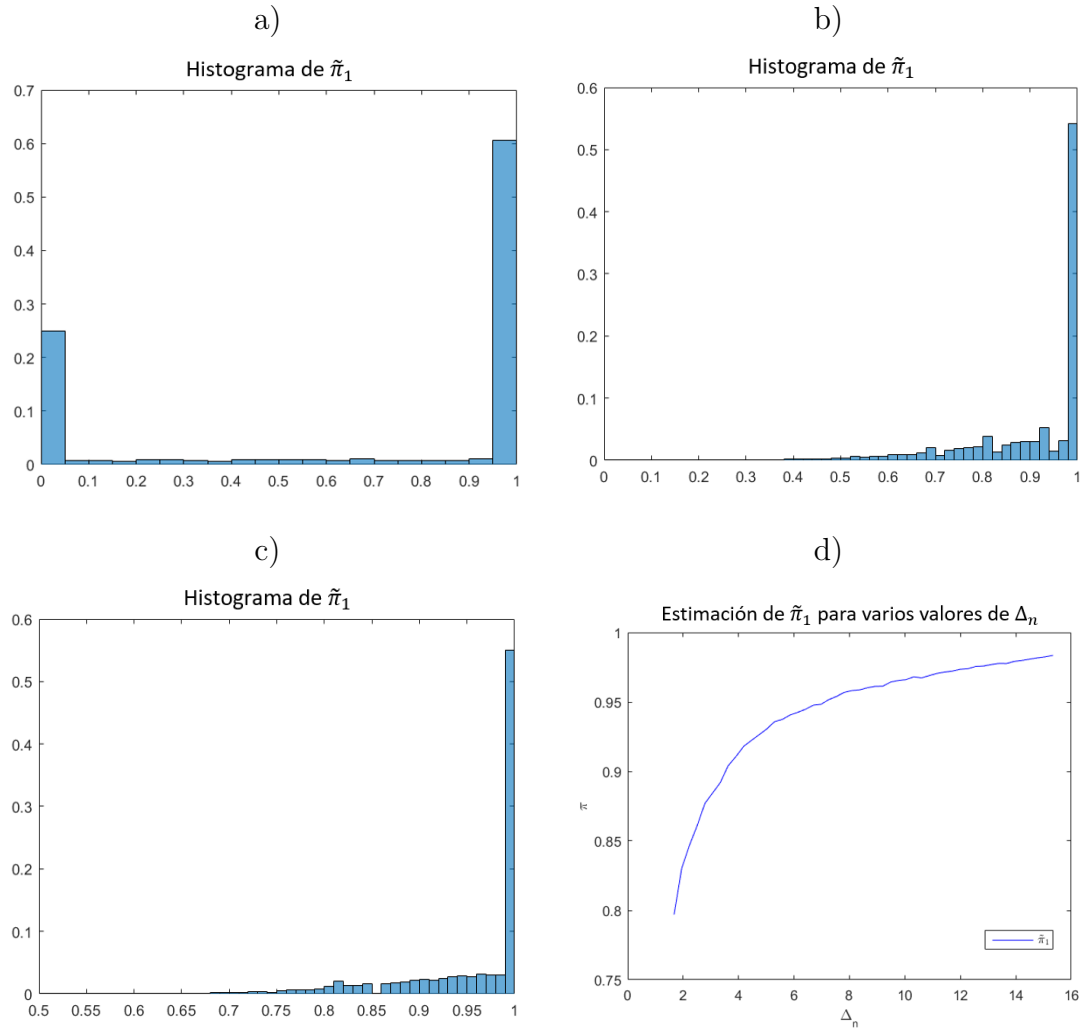

Figura 1. a) Histograma de $\widetilde{\pi}_1$ con $\Delta_n = 0.4183$ ($\sqrt{70}(0.05)$). b) Histograma de $\widetilde{\pi}_1$ con $\Delta_n = 8.3666$ ($\sqrt{70}(1)$). c) Histograma de $\widetilde{\pi}_1$ con $\Delta_n = 20.9165$ ($\sqrt{70}(2.5)$). d) Estimación de $\widetilde{\pi}_1$ para cierto rango de valores de Δ_n.

Asimismo, bajo las mismas condiciones que se describieron en el párrafo anterior, en la Figura 2 se muestran gráficas de la función objetivo $\varphi(\pi)$ del método con regularización para el caso de predicción de 1 dato, es decir, para la estimación de $\widehat{\pi}_1$. Además se muestra la estimación de $\widehat{\pi}_1$ (Figura 2 d) haciendo variar el valor de Δ_n, esto manteniendo fijos $n = 70$, $\theta_0 = 0$ y $\sigma = 1$ y variando θ_1. Como se puede observar en la Figura 2 (a) se tiene que para valores pequeños de Δ_n, el algoritmo estima una mezcla uniforme de $f(\bullet;\theta_0)$ y $f(\bullet;\theta_1)$, esto es, $\hat{\pi}_1 = 1/2$. A su vez, si se aumenta el valor de Δ_n, se obtiene una función objetivo $\varphi(\pi)$ con un óptimo cada vez más cercano

a 1. Esto se puede ver en la Figura 2 (b) y (c), además el comportamiento de $\hat{\pi}_1$ se observa en la Figura 2 (d), en donde se puede apreciar que conforme Δ_n crece, $\hat{\pi}_1$ tiende a al verdadero valor $\pi = 1$.

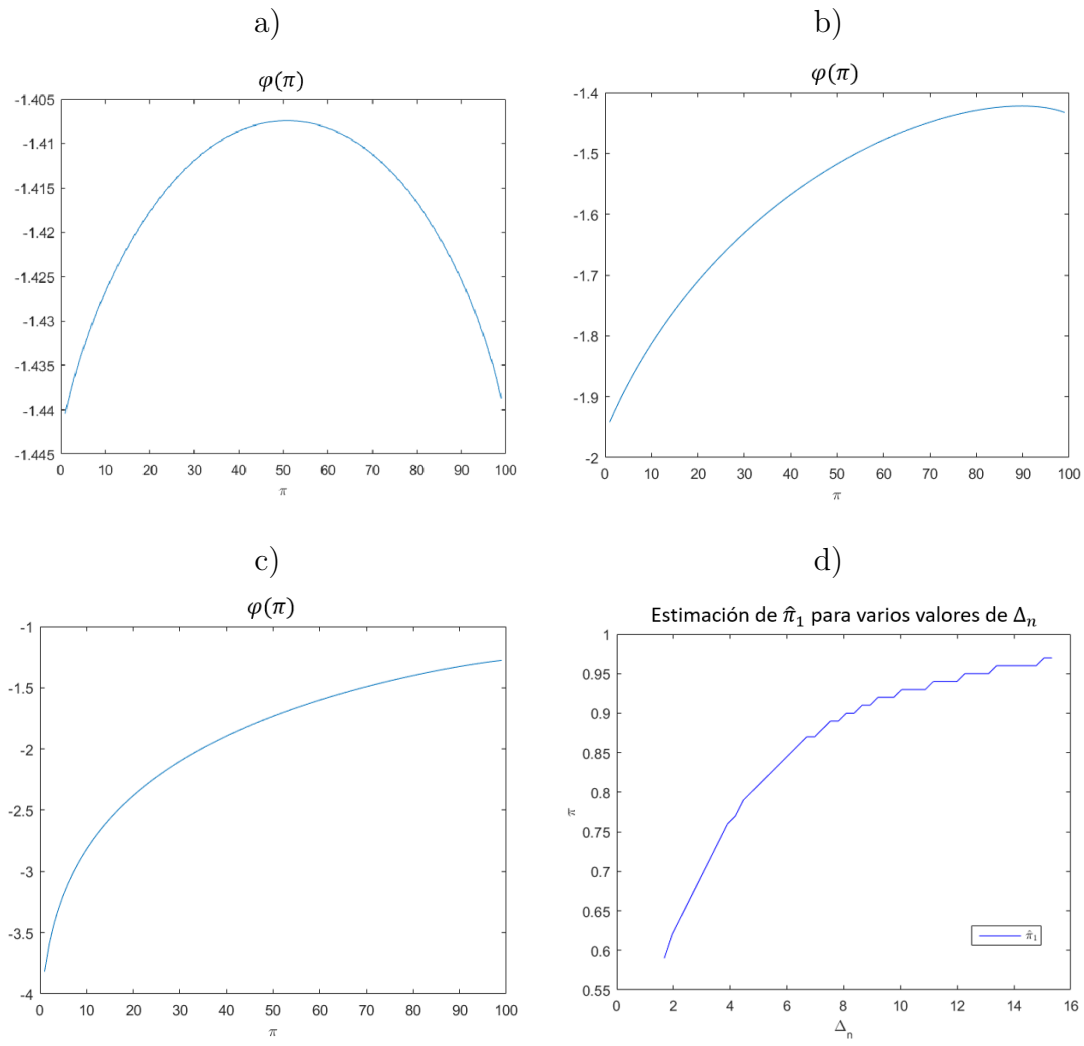

Figura 2. Método con regularización para estimar $\widehat{\pi}_1$. a) $\varphi(\pi)$ con $\Delta_n = 0.4183$ $(\sqrt{70}(0.05))$. b) $\varphi(\pi)$ con $\Delta_n = 8.3666$ $(\sqrt{70}(1))$. c) $\varphi(\pi)$ con $\Delta_n = 20.9165$ $(\sqrt{70}(2.5))$. d) Estimación de $\widehat{\pi}_1$ para cierto rango de valores de Δ_n.

En la Figura 3 se muestran las gráficas de $\varphi(\pi)$ para la estimación de $\widetilde{\pi}_n$. Dichas funciones $\varphi(\pi)$ para varios valores de θ_1. Como se puede ver en todas las gráficas de la Figura 3, incluso para valores pequeños de Δ_n, se tiene una estimación de $\tilde{\pi}_n = 1$.

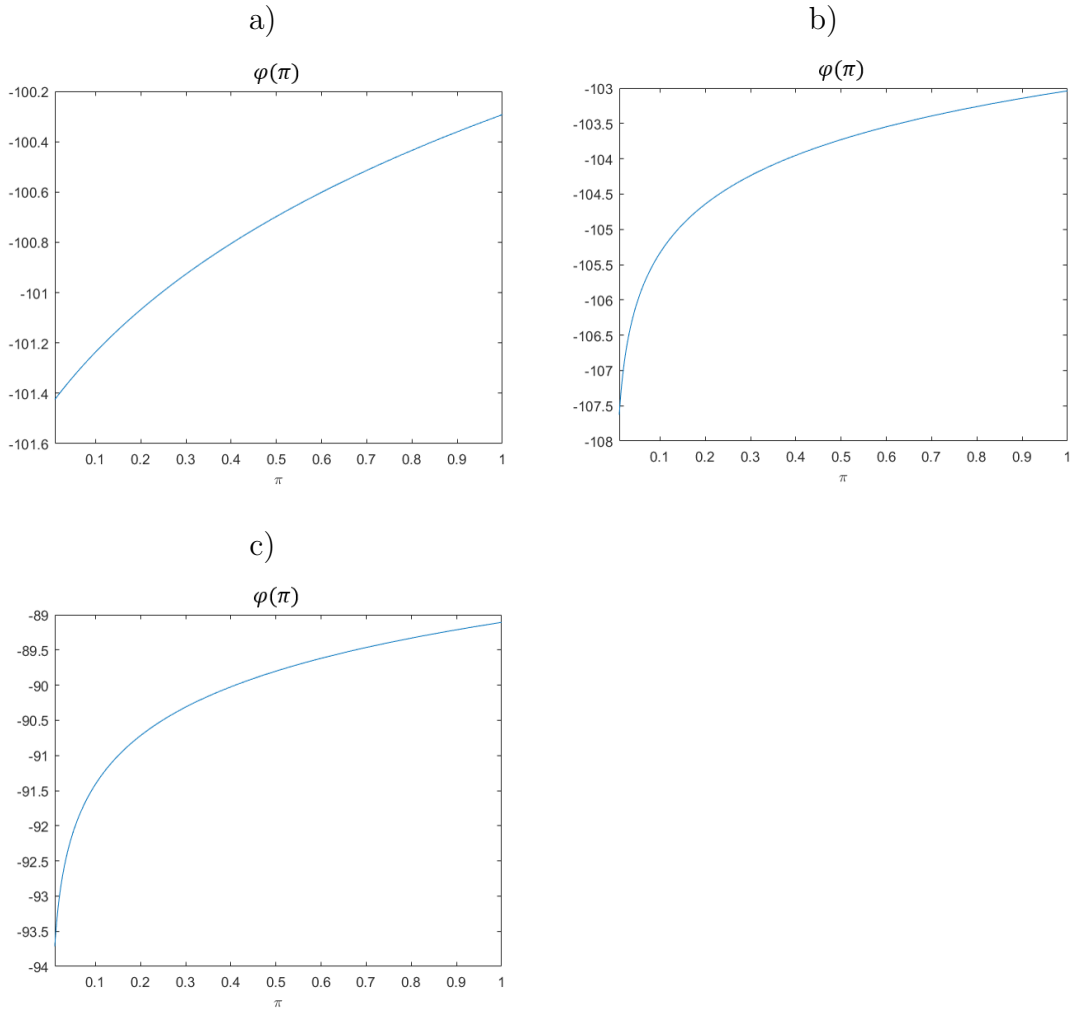

Figura 3. Método sin regularización para estimar $\widetilde{\pi}_n$. a) $\varphi(\pi)$ con $\Delta_n = 0.4183$ $(\sqrt{70}(0.05))$. b) $\varphi(\pi)$ con $\Delta_n = 8.3666$ $(\sqrt{70}(1))$. c) $\varphi(\pi)$ con $\Delta_n = 20.9165$ $(\sqrt{70}(2.5))$.

En la Tabla 1 se muestran concentradas las estimaciones puntuales de $\widetilde{\pi}_1$, $\widetilde{\pi}_n$ y $\widehat{\pi}_1$ en cada uno de los casos mostrados en las Figuras 1, 2 y 3.

θ_1	Δ_n	$\widetilde{\pi}_1$	$\widehat{\pi}_1$	$\widetilde{\pi}_n$
0.05	0.4183	0.5548	0.51	1
1	8.3666	0.9461	0.9	1
2.5	20.9165	1	0.99	1

Tabla 1. Estimaciones puntuales de $\widetilde{\pi}_1$, $\widehat{\pi}_1$ y $\widetilde{\pi}_n$

Dentro de las preguntas que surgieron en el desarrollo de la tesis, una de mucho interés fue si la estimación de π_{opt} arrojaba información similar tanto en el caso de predicción de 1 dato como en el caso de predicción de $m > 1$ datos, en particular de $m = n$. En la Figura 4 se muestra una gráfica comparativa de las estimaciones de $\widetilde{\pi}_1$, $\widehat{\pi}_1$, $\widetilde{\pi}_n$ y $\widetilde{\pi}_{n/2}$, ésto con la intención de evaluarlas en distintos escenarios. Como se puede observar, en el caso en el que se predicen $n/2$ o n datos la convergencia es más rapida al verdadero valor $\pi = 1$. Además, entre el algoritmo sin regularización y con regularización, el primero muestra una mejor estimación en el sentido que siempre se tiene que $\tilde{\pi}_1 > \hat{\pi}_1$ sin importar el valor de Δ_n.

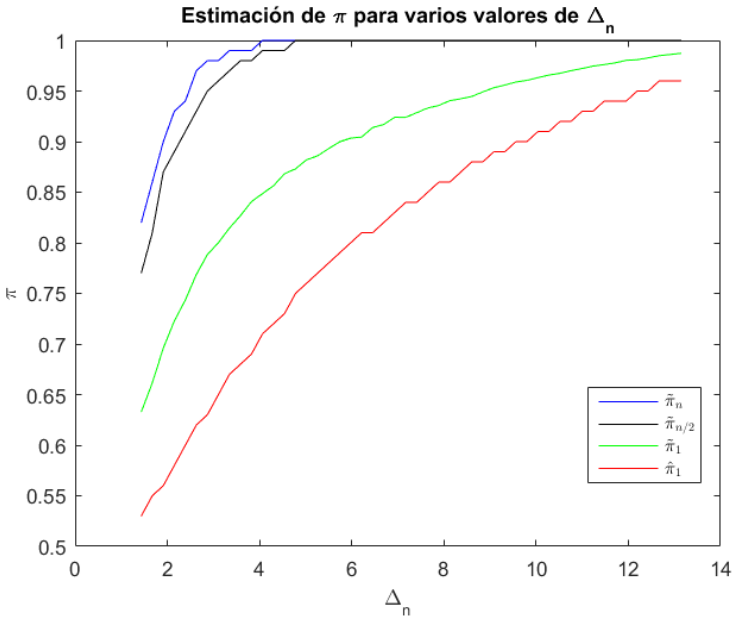

Figura 4. Comparación de las estimaciones de $\widetilde{\pi}_1$, $\widehat{\pi}_1$, $\widetilde{\pi}_n$ y $\widetilde{\pi}_{n/2}$ haciendo variar Δ_n

Como se mencionó anteriormente, en Harris (1989) se propone a la distribución del estimador por máxima verosimilitud $\widehat{\theta}$ de θ como mezcladora para obtener una buena estimación de la densidad predictiva. En el caso particular donde $\Theta = \{\theta_0, \theta_1\}$, se tiene que dicha distribución está dada por $\mathbb{P}(\widehat{\theta} = \theta_0) = 1 - \mathbb{P}(\widehat{\theta} = \theta_1)$. Como notación, sea

$$h_{\widetilde{\pi}}(z; x) = \widetilde{\pi} f(z; \theta_0) + (1 - \widetilde{\pi}) f(z; \theta_1),$$

el estimador de $h(z; x)$ respecto a $\widetilde{\pi}_1$. Además, más adelante se mostrará que en el caso Normal con un espacio para la media $\theta \in \{\theta_0, \theta_1\}$, se puede calcular analíticamente $\mathbb{P}(\widehat{\theta} = \theta_0)$. Sea

$$h_{MV}(z; x) = \mathbb{P}(\widehat{\theta} = \theta_0) f(z; \theta_0) + [1 - \mathbb{P}(\widehat{\theta} = \theta_0)] f(z; \theta_1),$$

el estimador de $h(z;x)$ respecto a la probabilidad teórica $\mathbb{P}(\widehat{\theta} = \theta_0)$. Por último, sea

$$h_{freq}(z;x) = \pi_{freq} f(z;\theta_0) + (1 - \pi_{freq}) f(z;\theta_1),$$

la estimación de $h(z;x)$ respecto a la frecuencia con la que ocurre el evento $A_0 = \{\widehat{\theta} = \theta_0\}$, donde π_{freq} es la frecuencia relativa de A_0 en un procedimiento bootstrap. En los escenarios de simulación se tiene que $h(z;x)$ es la verdadera densidad de los datos $f(z) = f(z;\theta_0)$. En la Figura 5 se muestra la estimación, bajo simulaciones Monte Carlo, de

$$\mathbb{E}[KL(f, \widehat{h})],$$

donde $\widehat{h}$ representa las estimaciones que acabamos de enlistar.

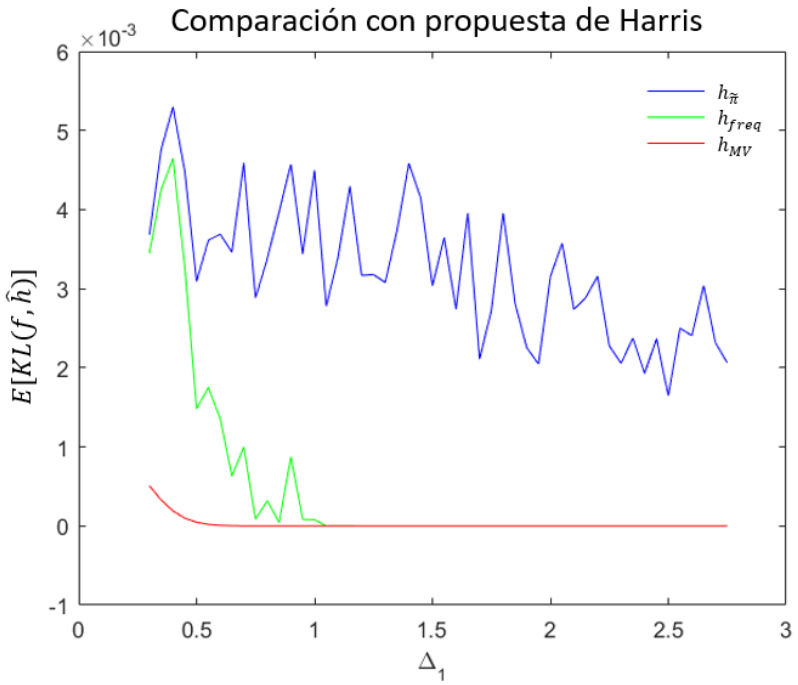

Figura 5. Estimación por Monte Carlo de $\mathbb{E}[KL(f, h_{\widetilde{\pi}})]$, $\mathbb{E}[KL(f, h_{freq})]$ y $\mathbb{E}[KL(f, h_{MV})]$.

Antes de hacer los comentarios finales (en la última sección de este capítulo) sobre estos resultados obtenidos en las simulaciones, se discutirá a continuación una conjetura que se hizo respecto a $\widetilde{\pi}_1$ y la distribución del estimador por máxima verosimilitud $\widehat{\theta}$ de θ, dada por $\mathbb{P}(\widehat{\theta} = \theta_0)$.

2.2.2. Relación con la distribución del EMV

Se planteó la conjetura siguiente: cuando $n \to \infty$, se cumple que

$$\widetilde{\pi}_1 \approx \mathbb{P}\left(\widehat{\theta} = \theta_0\right),$$

donde $\widehat{\theta}$ es el estimador máximo verosimil de θ. Por un lado, tenemos que, si $t_1 = \sum_i x_i^2$ y $t_2 = \sum_i x_i$, entonces

$$\begin{aligned}
\mathbb{P}\left(\widehat{\theta} = \theta_0\right) &= \mathbb{P}\left[\ell(\theta_0) \geq \ell(\theta_1)\right] \\
&= \mathbb{P}\left[-n\ln(\sigma) - \frac{t_1 - 2\theta_0 t_2 + n\theta_0^2}{2\sigma^2} \geq -n\ln(\sigma) - \frac{t_1 - 2\theta_1 t_2 + n\theta_1^2}{2\sigma^2}\right] \\
&= \mathbb{P}\left[-2\theta_0 t_2 + n\theta_0^2 \leq -2\theta_1 t_2 + n\theta_1^2\right] \\
&= \mathbb{P}\left[n\left(\theta_0^2 - \theta_1^2\right) \leq 2t_2\left(\theta_0 - \theta_1\right)\right] \\
&= \mathbb{P}\left[\frac{n\left(\theta_0^2 - \theta_1^2\right)}{2\left(\theta_0 - \theta_1\right)} \geq t_2\right] \text{ si } \theta_1 > \theta_0 \\
&= \mathbb{P}\left[t_2 \leq \frac{n}{2}\left(\theta_0 + \theta_1\right)\right].
\end{aligned}$$

Como $t_2 = \sum_i x_i \sim N\left(n\theta_0, n\sigma^2\right)$, entonces

$$\begin{aligned}
\mathbb{P}\left(\widehat{\theta} = \theta_0\right) &= \mathbb{P}\left[\frac{t_2 - n\theta_0}{\sqrt{n}\sigma} \leq \frac{n\left(\theta_0 + \theta_1\right) - 2n\theta_0}{2\sqrt{n}\sigma}\right] \\
&= \mathbb{P}\left[Z \leq \frac{\sqrt{n}\left(\theta_1 - \theta_0\right)}{2\sigma}\right],
\end{aligned}$$

donde $Z \sim N\left(0, 1\right)$.

Se planea explorar esta conjetura mediante simulaciones haciendo crecer el tamaño de muestra n y obteniendo mediante bootstrap el estimador $\widetilde{\pi}_1$ y la frecuencia relativa con la que ocurre el evento $\widehat{\theta} = \theta_0$ para aproximar $\mathbb{P}\left(\widehat{\theta} = \theta_0\right)$. En la Figura 6 se muestran los resultados de las simulaciones bajo distintos escenarios, cambiando el valor de Δ_n.

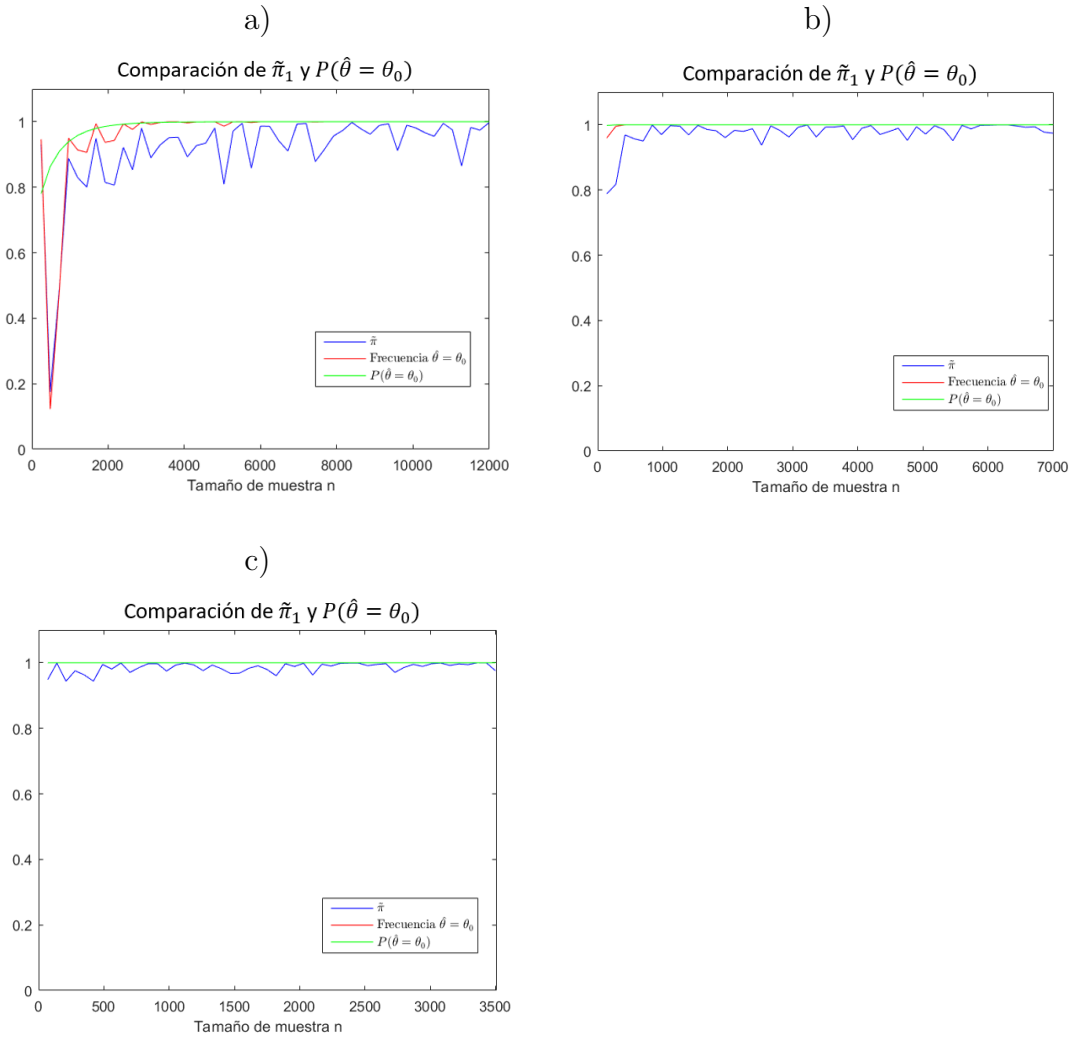

Figura 6. Comparación de $\widetilde{\pi}_1$, la frecuencia con la que ocurre $A_0 = \{\widehat{\theta} = \theta_0\}$ y la probabilidad teórica $\mathbb{P}(\widehat{\theta} = \theta_0)$, ésto para distintos valores de n. a) $\Delta_n = 0.4183$ $(\sqrt{70}(0.05))$. b) $\Delta_n = 4.183$ $(\sqrt{70}(0.5))$. c) $\Delta_n = 8.3666$ $(\sqrt{70}(1))$.

Como se puede observar en la Figura 6, a pesar de que Δ_n sea pequeño, conforme n crece, efectivamente $\mathbb{P}(\hat{\theta} = \theta_0) \approx \tilde{\pi}_1$. Además, para valores de Δ_n no tan pequeños, la convergencia es más rápida, aunque no tanto como la convergencia de la frecuencia realtiva observada en que ocurre el evento $A_0 = \{\hat{\theta} = \theta_0\}$.

2.3. Comentarios finales

En distintas simulaciones se mostraron varios resultados que versan sobre el comportamiento individual de los estimadores, la comparación entre ellos bajo diferentes escenarios de estimación, la comparación con la propuesta en Harris (1989) y sobre la conjetura de que $\widetilde{\pi}_1 \approx \mathbb{P}(\widehat{\theta} = \theta_0)$ cuando $n \to \infty$. De estos resultados se pueden concluir diversas aseveraciones. Así también motivan diferentes direcciones en las cuales se puede encaminar el trabajo realizado.

Antes que todo, cabe mencionar que en todas las simulaciones la verdadera densidad predictiva es $f(z) = f(z; \theta_0)$, es decir, idealmente $\pi_{opt} = 1$. De las Figuras 1 y 2 y de la Tabla 1 es posible aseverar que ambos estimadores para la predicción de un dato, tanto con regularización como sin ella, se aproximan a 1 cuando Δ_n se hace más grande, esto era de esperarse. También, era de esperarse que conforme Δ_n se hace pequeño, debido a la "cercanía" de los modelos, que el estimador de π_{opt} se acerque a 0.5, representando una elección aleatoria del parámetro θ. Cuando se habla de cercanía, se refiere justo a la diferencia relativa entre sus medias $\Delta_1 = \Delta_n/\sqrt{n}$, como se ilustra en la Figura 7. Por último, de la Figura 4, se puede concluir que el estimador sin regularización, $\widetilde{\pi}_1$, resultó un poco mejor que el estimador por el método con regularización, en el sentido de estar más cerca de 1.

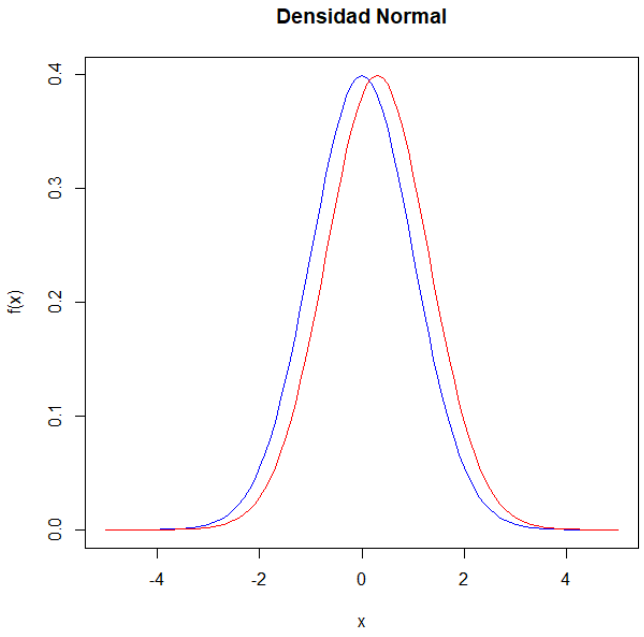

Figura 7. $\Delta_1 = 0.2$

Como se mencionó anteriormente, una pregunta importante fue si la estimación de π_{opt} resultaba parecida en los casos de predicción de 1 dato como de más datos futuros. Para tratar de contestar esta pregunta, se realizó una comparación de $\widetilde{\pi}_1$, $\widehat{\pi}_1$, $\widetilde{\pi}_n$ y $\widetilde{\pi}_{n/2}$, haciendo variar el valor de Δ_n, como se mostró en la Figura 4. De estos resultados es muy claro que $\widetilde{\pi}_n$ y $\widetilde{\pi}_{n/2}$ resultan mucho mejor, pues son más cercanos a $\pi_{opt} = 1$, que los estimadores en el caso de predicción de sólo 1 dato. Esto se debe a que la divergencia entre los modelos $f_n(x;\theta_0)$ y $f_n(x;\theta_1)$ (entre sus densidades) es de Δ_n mientras que la de $f_1(x;\theta_0)$ y $f_1(x;\theta_1)$ es de $\Delta_1 = \Delta_n/\sqrt{n}$. Como en la Figura 4 se muestran las estimaciones en escala de Δ_n, esto hace que los modelos en el caso de 1 dato sean muy cercanos, en el sentido que ya describimos, incluso con valores de Δ_n no tan cercanos a 0. Por lo que en un mismo escenario, las estimaciones de π_{opt} serán mejores en el caso de predicción de $m > 1$ datos que en el caso de sólo 1 dato.

Además, se hizo la comparación de la estimación de la densidad predictiva haciendo uso de $\widetilde{\pi}_1$, $h_{\widetilde{\pi}}$, con h_{freq} y h_{MV}. Como ya se comentó, la propuesta de Harris es optimal en diferentes condiciones, y bajo distintos métodos de estimación de la distribución del estimador de máxima verosimilitud. Los resultados mostrados en la Figura 5 ratifican la buena estimación que se obtiene de la densidad predictiva con h_{freq} y h_{MV}. No obstante, las estimaciones de $\mathbb{E}[KL(f, h_{\widetilde{\pi}})]$ resultaron en el orden de 0.003, lo que es un punto a favor de la estimación de $f(z)$ mediante el enfoque propuesto.

Respecto a la conjetura que se planteó de que $\widetilde{\pi}_1 \approx \mathbb{P}(\widehat{\theta} = \theta_0)$ cuando $n \to \infty$, es posible dar algunos comentarios. Como se puede observar en la Figura 6, la conjetura parece ser cierta en el caso Normal con espacio para la media con dos componentes, que es el caso en el que se llevaron a cabo las simulaciones. Como se mostró, $\mathbb{P}(\widehat{\theta} = \theta_0)$ depende del valor de Δ_n, por lo que, manteniendo fijo $|\theta_1 - \theta_0|$, para apreciar la convergencia, es necesario un número menor de n conforme $|\theta_1 - \theta_0|$ es mayor. Con esto, las simulaciones sugieren que es posible concluir que efectivamente $\widetilde{\pi}_1$ se acerca a $\mathbb{P}(\widehat{\theta} = \theta_0)$ cuando n crece.

A partir del trabajo realizado en este capítulo, salen a relucir varias direcciones en las que se pueden enfocar esfuerzos futuros. Por un lado, los algoritmos descritos son para el caso particular en el que se tiene un espacio de parámetros $\Theta = \{\theta_0, \theta_1\}$. A pesar de que se implementó para el caso normal univariado y multivariado con matriz de covarianzas la identidad, estos algoritmos

pueden ser generalizados facilmente en el caso en donde Θ es un subconjunto finito de $\mathbb{R}^k$. Por otro lado, las implementaciones y las simulaciones que se llevaron a cabo fueron situadas en el caso del parámetro de localización de la distribución normal, sin embargo, es interesante plantear el enfoque en distribuciones diferentes, como por ejemplo las pertenecientes a la familia de localización y escala, o bien, motivados por los casos de estudio tratados en Harris (1989), la distribución Poisson o Binomial. Se hacen estos comentarios con la intención de motivar trabajos futuros en este enfoque.

Bajo el enfoque que se describió a lo largo de este capítulo, el problema de estimar la densidad predictiva $h(z;x)$ se puede atacar considerando el modelo de mezcla óptimo, a la luz de los datos x, $\pi_{opt}f(z;\theta_0) + (1-\pi_{opt})f(z;\theta_1)$. No obstante, si suponemos un modelo $f(z;\theta)$ para $h(z;x)$, y consideramos la prueba de hipótesis $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, se tiene que π_{opt} representa el peso estimado desde una perspectiva de ponderación de hipótesis en el caso específico de dos hipótesis simples. En el capítulo 3 se abordará el caso de una familia estructurada de distribuciones involucrando componentes compuestas, en donde el espacio de parámetros Θ puede no ser finito y lo que es de interés es considerar una familia de subconjuntos del mismo, situando el problema en un contexto de ponderación de hipótesis compuestas.

Capítulo 3

Modelos paramétricos con componentes compuestas

En muchas áreas de la ciencia, como medicina, biología y genómica, surge la necesidad de poner a prueba la veracidad de una familia de hipótesis $H_0, H_1, \ldots, H_{k-1}$. Para mostrar la diversidad de aplicación de esta problemática, mencionemos un par de ejemplos. En genética, es de común interés identificar las diferencias en la expresión de genes a partir de microarreglos de ADN, contexto en el que Dudoit (2003) hace una discusión de distintos enfoques. En el área de neurología, Narayan et al. (2015) propone un método para contrastar si dos modelos gráficos gaussianos son "iguales", haciendo uso del enfoque por pruebas de hipótesis múltiples sobre las probabilidades de existencia de una arista en el modelo. Además, en Austin et al. (2014) se hace una revisión de las aportaciones más relevantes en el campo de pruebas de hipótesis múltiples. Concretando en un caso particular, supóngase que se tiene una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$, donde cada x_i se distribuye de acuerdo a una densidad $f(\bullet; \theta)$; además, se tiene una familia de hipótesis $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1, \ldots, H_{k-1} : \theta \in \Theta_{k-1}$ concernientes al parámetro θ.

Adaptado a esta situación, el enfoque expuesto en las Capítulos 1 y 2, es posible contruir una distribución inferencial basada en la muestra x. Específicamente, se busca $q(\theta; x)$ de tal forma que su masa esté distribuida en la familia de subconjuntos $\Theta_0, \Theta_1, \ldots, \Theta_{k-1}$ como $\pi_i = \mathbb{P}(\theta \in \Theta_i)$, es decir, q está definida por los pesos $(\pi_0, \pi_1, \ldots, \pi_{k-1})$. Así, $\hat{q}$ es

$$\hat{q} = \arg\max_{\pi} \mathbb{E}\left(KL(f, h(\bullet; x))\right) \tag{3.1}$$

donde $\pi = (\pi_0, \pi_1, \ldots . \pi_{k-1})$, $h(z;x) = \sum_{i=0}^{k-1} \pi_i f(z;\hat{\theta}_i)$, $\hat{\theta}_i$ es el estimador por máxima verosimilitud restringido a Θ_i, $z = (z_1, z_2, \ldots, z_m)$ es una muestra futura y f es la densidad de z. De aquí que $\hat{q}$ es una ponderación para la familia de hipótesis, óptima según dicho criterio. Claro está que dicho criterio depende de una esperanza desconocida, que habrá que sustituir por alguna versión empírica de ella. A continuación se detalla el enfoque propuesto en el contexto de ponderación de hipótesis.

3.1. Ponderación de hipótesis

Supóngase que se tiene una muestra aleatoria x que se distribuye de acuerdo a una densidad $f(\bullet;\theta)$, donde $\theta \in \Theta = \sqcup_{i=0}^{k-1}\Theta_i$. Ahora, el objetivo es evaluar el conjunto de hipótesis $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1, \ldots, H_{k-1} : \theta \in \Theta_{k-1}$. Para lograr dicho objetivo, es posible adaptar el enfoque planteado en los capítulos 1 y 2, para construir una densidad inferencial $q(\bullet;x)$ que pondere las hipótesis propuestas. Retomando el enfoque, se busca $\hat{q}$ tal que

$$\hat{q} = \arg\max_{q} \mathbb{E}\left[\ln\left(\int_{\Theta} f(z;\theta)q(\theta;x)d\theta\right)\right], \tag{3.2}$$

donde el valor esperado es respecto a x y z. Se considerará la siguiente forma específica de $q(\theta;x)$:

$$q(\theta;x) = \sum_{i=0}^{k-1} \pi_i \delta_{\hat{\theta}_i}(\theta),$$

donde $\delta_{\hat{\theta}_i}(\bullet)$ es la delta de Dirac centrada en el estimador por máxima verosimilitud $\hat{\theta}_i$ restringido a Θ_i y $\pi_0, \pi_1, \ldots, \pi_{k-1} > 0$ satisfacen que $\sum_{i=0}^{k-1} \pi_i = 1$. Así, (3.2) es equivalente a

$$\hat{q} = \arg\max_{q} \mathbb{E}\left[\ln\left(\sum_{i=0}^{k-1} f(z;\hat{\theta}_i)q(\hat{\theta}_i;x)\right)\right].$$

Luego, la búsqueda de $\hat{q}$ se reduce a determinar $\pi = (\pi_0, \pi_1, \ldots, \pi_{k-1})$ mediante

$$\hat{\pi} = \arg\max_{\pi} \mathbb{E}\left[\ln\left(\sum_{i=0}^{k-1} \pi_i f(z;\hat{\theta}_i)\right)\right]. \tag{3.3}$$

Nótese que en (3.3) en general se desconoce la densidad verdadera, por lo que calcular explícitamente la esperanza no es posible. Por esta razón, para aproximar el término a maximizar en (3.3), se hará uso de una versión bootstrap, como se detallará más adelante en los algoritmos.

Para los alcances de esta tesis considérese el caso particular en el que se tienen dos hipótesis $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, en donde $\Theta_0 = \{\theta_0\}$ y $\Theta_1 = \Theta_0^c$, lo cual se traduce en el caso $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

3.1.1. Hipótesis simple vs. hipótesis compuesta

Concentrando la atención en un caso simple en particular, supóngase que se ponen a prueba las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

Supóngase que se tiene una muestra aleatoria x que se distribuye conforme a una densidad $f(\bullet;\theta)$. Sobre la base de esta muestra aleatoria, desde el enfoque propuesto, lo que se busca es obtener $\hat{\pi}$ tal que maximice una versión empírica del criterio (3.3). Para esto, se propone el siguiente método de ponderación de hipótesis:

Algoritmo: Para ponderación de hipótesis simple vs. compuesta

1. Se obtienen B muestras bootstrap $x_1^*, x_2^*, \ldots, x_B^*$; donde $x_b^* = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$.

2. Para cada $b = 1, 2, \ldots, B$ se obtiene

$$\pi_b = \arg\max_{\pi} \sum_{j=1}^{n} \ln\left(\pi f(x_j^{*b};\theta_0) + (1-\pi) f(x_j^{*b};\hat{\theta}_b^{*(j)})\right),$$

   donde $\hat{\theta}_b^{*(j)}$ es el estimador por máxima verosimilitud basado en $x_b^* \setminus \{x_j^{*b}\}$

3. Obtenemos el estimador final de π_{opt},

$$\widetilde{\pi}_1 = \frac{1}{B}\sum_{b=1}^{B} \pi_b.$$

Se realizaron simulaciones con la intención de comparar este enfoque con algunos otros que forman parte de los procedimientos usuales en el contexto de pruebas de hipótesis. A continuación se exponen dichos enfoques para después explicar los escenarios de simulación.

3.1.2. Comparación con otros enfoques

En la literatura se pueden encontrar varias formas de tratar esta prueba de hipótesis. A pesar de ser muy criticado desde distintos frentes, el p-valor es posiblemente el punto de referencia más utilizado para el rechazo de una hipótesis. Por otro lado, los intervalos de verosimilitud también han tomado un papel importante para muchos estadísticos, siendo utilizados como regiones de plausibilidad para una hipótesis como $H_0 : \theta = \theta_0$. No obstante, como se discute en Wasserstein y Lazar (2016), no se recomienda al p-valor como un referente único para la toma de decisiones. A su vez, la escala en la que se mide la verosimilitud relativa, punto de partida para los intervalos de verosimilitud, no tiene una interpretación sencilla en general. Por otro lado, desde el enfoque bayesiano, en Berger y Sellke (1987) se discuten varias propuestas para tratar bayesianamente esta prueba de hipótesis. En este capítulo se comparará el enfoque propuesto con el procedimiento clásico de pruebas de significancia, con el enfoque de intervalos de verosimilitud y con el enfoque bayesiano basado en una distribución previa tipo Jeffreys.

Sea $X = (X_1, X_2, \ldots, X_n)$, donde las X_i son iid con densidad $f(\bullet; \theta)$. Se desea probar la hipótesis nula $H_0 : \theta = \theta_0$ contra la hipótesis alternativa $H_1 : \theta \neq \theta_0$. El procedimiento clásico de pruebas de significancia es el de definir una estadística de prueba $T(X)$ y tomar la decisión de rechazar o no H_0 dependiendo de cuán pequeño es el p-valor, definido como

$$p = \mathbb{P}\left[T(X) \geq T(x)\right],$$

donde $T(x)$ es la estadística de prueba evaluada en la muestra aleatoria observada. Existe actualmente una gran discusión sobre qué valor considerar como umbral para identificar a un p-valor como suficientemente pequeño. Existen convenciones en áreas de la ciencia en particular, en las que un umbral de 0.05 o 0.01 representa suficiente evidencia en contra de la hipótesis nula. No obstante, como se menciona en Wasserstein y Lazar (2016), estos valores no deben ser puestos como una medida universal para el rechazo de una hipótesis.

Particularmente, supóngase que se tiene que $X = (X_1, X_2, \ldots, X_n)$, donde las X_i son iid con distribución $N(\theta, \sigma^2)$. La estadística de prueba usual para la prueba que se está tratando ($H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$) es

$$T(X) = \sqrt{n}\frac{|\bar{X} - \theta_0|}{\sigma},$$

donde $\bar{X}$ es el promedio. Así, se tiene que

$$t = T(x) = \sqrt{n}\frac{|\bar{x} - \theta_0|}{\sigma},$$

y el p-valor en este caso es

$$p = 2(1 - \Phi(t)),$$

donde $\Phi(\bullet)$ es la función de distribución de una normal estándar.

Considérese ahora un enfoque bayesiano en el que se elige como distribución previa de θ, aquella que le da $1/2$ de probabilidad tanto a H_0 como a H_1, y que además distribuye la masa en H_1 de acuerdo a una distribución $N(\theta_0, \sigma^2)$. Esta distribución previa es similar a la propuesta por Jeffreys (1961), quien propuso los mismos pesos para las hipótesis, pero con una distribución de masa sobre H_1 de acuerdo a un distribución Cauchy. Ahora bien, con esto se puede calcular $\mathbb{P}(H_0|x)$,

$$\mathbb{P}(H_0|x) = \frac{\mathbb{P}(x|H_0)}{m(x)}$$

donde $m(x) = 1/2 f(x;\theta_0) + 1/2 \int f(x;\theta) g(\theta) d\theta$ es la densidad marginal de x y $g(\theta)$ es la densidad de $N(\theta_0, \sigma^2)$. Es fácil comprobar que (ver Berger y Sellke (1987) para detalles)

$$P(H_0|x) = \left(1 + (1+n)^{-1/2} exp\left\{\frac{t^2}{[2(1+1/n)]}\right\}\right)^{-1}.$$

Como se verá más adelante, para una n fija, no necesariamente pequeña, $P(H_0|x)$ es grande para valores grandes de t en casos en que otros enfoques proveen fuerte evidencia en contra de H_0.

En Pawitan (2001) se describe la forma de utilizar intervalos de verosimilitud como base informativa para rechazar o no una hipótesis. Un intervalo de verosimilitud de nivel c se define como

$$\begin{aligned} IV(c) &= \left\{\theta : \frac{L(\theta;x)}{L(\hat{\theta};x)} \geq c\right\} \\ &= \{\theta : R(\theta;x) \geq c\}, \end{aligned}$$

donde $L(\bullet;x)$ y $R(\bullet;x)$ son la verosimilitud y la verosimilitud relativa de θ respectivamente. Sólo con la definición de un intervalo de verosimilitud, no es claro por qué algún valor de c sería adecuado

para considerar relevante a $IV(c)$. Sin embargo, para el caso de la distribución $N(\theta, \sigma^2)$ se tiene que

$$\begin{aligned} R(\theta; x) &= exp\left[-\frac{n}{2\sigma^2}\left(\bar{x}-\theta\right)^2\right] \\ \ln\left[R(\theta; x)\right] &= -\frac{n}{2\sigma^2}\left(\bar{x}-\theta\right)^2. \end{aligned}$$

Es conocido que $\bar{x} \sim N(\theta; \sigma^2/n)$, así,

$$\frac{n}{\sigma^2}\left(\bar{x}-\theta\right)^2 \sim \chi^2_1,$$

esto es,

$$W := -2\ln\left[R(\theta; x)\right] \sim \chi^2_1.$$

W es conocido como el estadístico de Wilk. Su distribución χ^2_1 es exacta para el caso normal y, como se demuestra en el capítulo 9 de Pawitan (2001), en el caso general W tiene distribución asintótica χ^2_1 bajo condiciones de regularidad sobre la verosimilitud $L(\theta; x)$. Así, es posible calcular la probabilidad de cobertura de $IV(c)$.

Entonces,

$$\begin{aligned} \mathbb{P}\left[IV(c) \ni \theta\right] &= \mathbb{P}\left[R(\theta; x) \geq c\right] \\ &= \mathbb{P}\left[-2\ln\left[R(\theta; x)\right] \leq -2\ln(c)\right] \\ &= \mathbb{P}\left[\chi^2_1 \leq -2\ln(c)\right]. \end{aligned}$$

De aquí, que para obtener $\mathbb{P}\left[IV(c) \ni \theta\right] = 1-\alpha$, entonces

$$\begin{aligned} -2\ln(c) &= \chi^2_{1,1-\alpha} \\ c &= exp\left[\frac{\chi^2_{1,1-\alpha}}{2}\right], \end{aligned}$$

donde $\chi^2_{1,1-\alpha}$ es el cuantil de probabilidad $1-\alpha$ de una distribución χ^2_1. En particular, para $\alpha = 0.05$ se tiene que $c \approx 0.1465$. Con esto, es posible interpretar un intervalo $IV(c)$ en términos de probabilidad de cubrimiento del verdadero valor del parámetro.

A continuación se describen los escenarios de simulación en los cuales se comparan los enfoques recién expuestos, así como también se muestran los resultados de dichas simulaciones.

3.2. Simulaciones

Se realizaron varias simulaciones para mostrar el desempeño del estimador respecto al enfoque bayesiano con distribución previa tipo Jeffreys, con la verosimilitud relativa y con el enfoque clásico por pruebas de significancia. Se consideraron dos comparaciones. Primero, para comparar el enfoque propuesto en esta tesis con el bayesiano con previa tipo Jeffreys y el procedimiento clásico por pruebas de significancia. Por otro lado, se comparó el enfoque propuesto con la verosimilitud relativa, el p-valor y el valor esperado del p-valor. A continuación se explican los escenarios de cada simulación

Para la primera comparación, en todos los casos se consideró una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$, donde cada x_i se distribuye como una normal $N(\theta_1, \sigma^2)$, $\sigma = 1$ y θ_1 variable dependiendo del escenario (el objetivo es evaluar $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$). Primero, para comparar con el enfoque bayesiano y el p-valor, se retomaron los escenarios de simulación de Berger y Sellke (1987), en los cuales se generaron muestras de tamaño $n = 50, 100$ y 1000 y la intención es comparar para un p-valor fijo correspondientes a una t fija también. Los dos casos son $p-valor = 0.05$ y $p-valor = 0.01$, correspondientes a $t = 1.96$ y $t = 2.576$ respectivamente. Para obtener $\tilde{\pi}$ en cada caso, se obtuvo el valor de θ_0 correspondiente a cada valor de t.

Por simplicidad, se consideraron distintos valores de θ_0 positivos. Observe la Figura 8 (comparación con la VR); para que la verosimilitud relativa tome el valor de 0.1465 o 0.035, correspondientes a $t = 1.96$ y $t = 2.576$ respectivamente, el valor de θ_0 debe ser aproximadamente 0.2 y 0.3. Es necesario aclarar que las gráficas mostradas en la Figura 8 corresponden a un tamaño de muestra $n = 70$; sin embargo, se siguió esta metodología para escoger el valor de θ_0 correspondientes a tamaños de muestra $n = 50, 100$ y 1000, y con esto calcular $\tilde{\pi}$, considerando que la muestra observada se distribuye como $N(0, 1)$. En las tablas 2 y 3 se muestran los resultados de estas simulaciones, en donde el valor de la verosimilitud relativa se muestra sólo como referencia de la metodología utilizada para calcular $\tilde{\pi}$.

p-valor=0.05, t=1.96	**Tamaño de muestra**		
Enfoque	**50**	**100**	**1000**
Previa tipo Jeffreys	0.52	0.6	0.82
Verosimilitud relativa	0.1465		
Propuesta de tesis	0.3012	0.3647	0.4424

Tabla 2. En el renglón de Previa tipo Jeffreys se muestra $\mathbb{P}(H_0|x)$, en el de Verosimilitud relativa se muestra el valor que toma ésta para el correspondiente valor de t, en el renglón de la Propuesta de tesis se muestra el valor de $\tilde{\pi}$.

p-valor=0.01, t=2.576	**Tamaño de muestra**		
Enfoque	**50**	**100**	**1000**
Previa tipo Jeffreys	0.22	0.27	0.53
Verosimilitud relativa	0.036		
Propuesta de tesis	0.1784	0.1349	0.2844

Tabla 3. En el renglón de Previa tipo Jeffreys se muestra $\mathbb{P}(H_0|x)$, en el de Verosimilitud relativa se muestra el valor que toma ésta para el correspondiente valor de t, en el renglón de la Propuesta de tesis se muestra el valor de $\tilde{\pi}$.

Con una $t = 1.96$ o $t = 2.576$, con un procedimiento de pruebas de significancia, la muestra presenta suficiente evidencia en contra de H_0, con un nivel del 0.05 ó 0.01 según sea el caso. No obstante, en ambos casos $\mathbb{P}(H_0|x)$ crece conforme el tamaño de muestra n crece. Para $t = 1.96$, en todos los escenarios, $\mathbb{P}(H_0|x) > 0.5$, sugiriendo alta plausibilidad para H_0 sobre la base de la muestra x. El enfoque propuesto resulta ser mejor que el enfoque bayesiano descrito, pues en todos los escenarios se obtiene un $\tilde{\pi}$ menor al de la previa tipo Jeffreys, como sería lógico esperar. Por otra parte, con un enfoque de prueba de significancia, se hubiese rechazado fuertemente la hipótesis H_0. El crecimiento de $\tilde{\pi}$ conforme n crece se debe a que la verosimilitud relativa se vuelve más angosta al tener mayor información, es decir, los intervalos de verosimilitud tienden a ser más pequeños haciendo que θ_0 se aproxime al EMV, por lo que $\tilde{\pi}$ tiende a 0.5.

Por otro lado, para la comparación con el p-valor, su valor medio y la verosimilitud relativa, se realizó el siguiente experimento. Se simuló una muestra aleatoria $x = (x_1, x_2, \ldots, x_n)$, donde $n = 70$ y cada x_i se distribuye como una normal $N(0, 1)$. Para valores de θ_0 entre 0 y 3 se obtuvo la correspondiente estimación de $\tilde{\pi}$. Además, para cada valor de θ_0 también se obtuvo el correspondiente valor $R(\theta_0; x)$ y el p-valor $2\left[1 - \Phi\left(\sqrt{n}\frac{|\bar{x}-\theta_1|}{\sigma}\right)\right]$, donde Φ es la función de distribución de una normal estándar, así como también se obtuvo mediante bootstrap la estimación del valor esperado de dicho p-valor. En la Figura 8 se muestra la comparación de estos resultados.

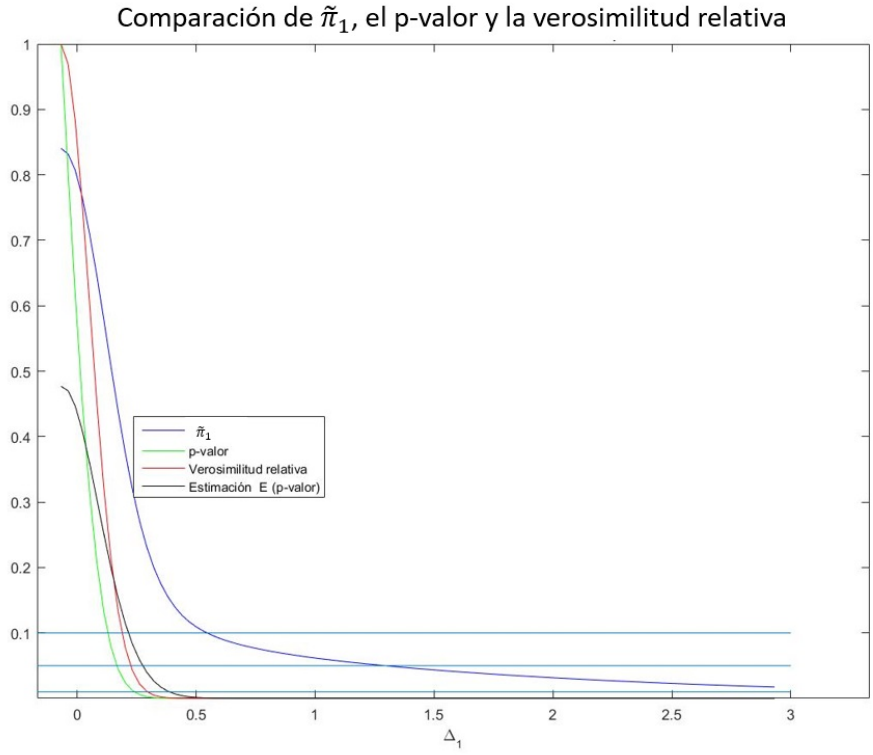

Figura 8. Comparación de $\tilde{\pi}$, el p-valor, la Verosimilitud relativa y el valor esperado del p-valor.

Como se puede observar en la Figura 8, el valor de $\tilde{\pi}$ tiende a 0 conforme θ_0 crece, es decir, cuando θ_0 se aleja del verdadero valor de $\theta = 0$. No obstante, comparado con los demás enfoques, la convergencia de $\tilde{\pi}$ es menos acelerada. Al final de este capítulo se darán algunos comentarios finales complementarios a los que se mencionan en esta sección.

A continuación se expone la forma en que la maquinaria propuesta en esta tesis funciona para la estimación de la densidad predictiva en el caso de la distribución generalizada de valores

extremos (DGVE), donde los parámetros de localización y escala son conocidos, pero se desconoce el parámetro de forma. Se tiene que para ciertos valores de dicho parámetro de forma, la DGVE toma nombres conocidos: Gumbel, Fréchet y Weibull. Se comenzará la sección describiendo el contexto de este problema.

3.3. Aplicación al modelo de distribución de valores extremos generalizada

Como se describe en Coles (2001), la teoría de valores extremos es una disciplina muy particular dentro de la estadística, debido a que provee de técnicas y modelos para describir la ocurrencia de eventos inusuales. No es coincidencia que las primeras aplicaciones de los modelos de valores extremos hayan sido en ingeniería civil: para los ingenieros es esencial diseñar sus estructuras con la capacidad de resistir la cantidad de fuerza que razonablemente se espera que las impacte. La teoría de valores extremos proporciona un marco de trabajo en el cual es posible tener una estimación anticipada de dichas fuerzas en función de datos históricos. Sin embargo, en los últimos años, esta teoría ha tomado relevancia en otras áreas, como meteorología, oceanografía, geología, finanzas, entre otras.

La DGVE es comunmente utilizada para modelar distintos fenómenos tratados en la teoría de valores extremos. La densidad de esta distribución tiene tres parámetros, uno de localización, uno de escala y uno de forma. Es de particular interés la estimación del parámetro de forma, ya que dependiendo de si este es positivo, negativo o nulo, la DGVE resulta en distribuciones conocidas: Fréchet, Weibull y Gumbel, respectivamente. Esto es, la DGVE engloba estas tres últimas distribuciones, las cuales tomaron importancia tiempo atrás en el modelado de distintos fenómenos presentes en el estudio de la teoría de valores extremos.

Supóngase que se tiene una muestra x con distribución de valores extremos generalizada y se desea inferir la densidad de datos futuros z, $h(z;x)$, en donde z se distribuye igual que x. De manera usual, existen procedimientos clásicos para la estimación de los parámetros de la DGVE, tales como el método de máxima verosimilitud y el método de momentos ponderado. Con estos métodos es posible obtener una estimación puntual de los parámetros y con ello una estimación

por sustitución de la densidad de datos futuros $h(z;x)$. No obstante la propuesta de esta tesis permite estimar la densidad predictiva mediante una mezcla de las densidades Gumbel, Fréchet y Weibull. A continuación se explica com mayor detalle la propuesta en este contexto.

3.3.1. Partición del espacio de valores del parámetro de forma

La densidad de la DVEG está dada por

$$f(x;\mu,\sigma,\kappa) = \begin{cases} \frac{1}{\sigma} exp\Big\{ -\Big[1+\kappa\frac{(x-\mu)}{\sigma}\Big]^{-\frac{1}{\kappa}}\Big\}\Big[1+\kappa\frac{(x-\mu)}{\sigma}\Big]^{-1-\frac{1}{\kappa}} & \kappa > 0 \\ \frac{1}{\sigma} exp\Big\{ -exp\Big[-\frac{(x-\mu)}{\sigma}\Big]\Big\} exp\Big[-\frac{(x-\mu)}{\sigma}\Big] & \kappa = 0 \end{cases}$$

para $1+\kappa\frac{(x-\mu)}{\sigma} > 0$ si $\kappa \neq 0$, y $f(x;\mu,\sigma,\kappa)$ tiene como dominio a todo $\mathbb{R}$ cuando $\kappa = 0$. Aquí $\mu \in \mathbb{R}$ es el parámetro de localización, $\sigma > 0$ es el parámetro de escala y $\kappa \in \mathbb{R}$ es el parámetro de forma. Para ciertos valores de κ, la DGVE tiene nombres conocidos. Para $\kappa = 0$ se tiene la distribución Gumbel, para $\kappa > 0$ se tiene la distribución Fréchet y si $\kappa < 0$ se tiene la distribución Weibull.

Es posible pues dividir el espacio de parámetros $\Theta = \{(\kappa,\sigma,\mu) : \kappa,\mu \in \mathbb{R}, \sigma > 0\}$ en 3 subconjuntos disjuntos,

$$\Theta_1 = \{(0,\sigma,\mu) : \mu \in \mathbb{R}, \sigma > 0\},$$

$$\Theta_2 = \{(\kappa > 0,\sigma,\mu) : \mu \in \mathbb{R}, \sigma > 0\},$$

$$\Theta_3 = \{(\kappa < 0,\sigma,\mu) : \mu \in \mathbb{R}, \sigma > 0\}.$$

Dada una muestra aleatoria $x \sim f(\bullet;\kappa,\sigma,\mu)$, si se quiere estimar $f(z)$, se puede aplicar el enfoque principal de esta tesis, obteniendo la mejor mezcla

$$\pi_1 f(z;\hat{\theta}_1) + \pi_2 f(z;\hat{\theta}_2) + (1-\pi_1-\pi_2) f(z;\hat{\theta}_3).$$

En donde

$$\hat{\theta}_1 = \{(0,\hat{\sigma},\hat{\mu}) : \mu \in \mathbb{R}, \sigma > 0\},$$

$$\hat{\theta}_2 = \{(\hat{\kappa}_+,\hat{\sigma},\hat{\mu}) : \mu \in \mathbb{R}, \sigma > 0\},$$

$$\hat{\theta}_3 = \{(\hat{\kappa}_-,\hat{\sigma},\hat{\mu}) : \mu \in \mathbb{R}, \sigma > 0\},$$

con $\hat{\kappa}_+$ y $\hat{\kappa}_-$ los estimadores por máxima verosimilitud restringidos a $\mathbb{R}^+$ y $\mathbb{R}^-$ respectivamente. Además, supóngase que $0 \leq \pi_1, \pi_2 \leq 1$.

En este contexto, se deriva el siguiente algoritmo para estimar π_{opt}:

Algoritmo: Para el parámetro de forma de la DVEG

1. Se obtienen B muestras bootstrap $x_1^*, x_2^*, \ldots, x_B^*$; donde $x_b^* = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$.

2. Para cada $b = 1, 2, \ldots, B$ se obtiene

$$\pi_b = \arg\max_{\pi} \sum_{j=1}^{n} \log\left(\pi_1 f(x_j^{*b}; \hat{\theta}_1^{*b(j)}) + \pi_2 f(x_j^{*b}; \hat{\theta}_2^{*b(j)}) + (1 - \pi_1 - \pi_2) f(x_j^{*b}; \hat{\theta}_3^{*b(j)})\right).$$

   donde $\hat{\theta}_i^{*b(j)}$ es el estimador por máxima verosimilitud basado en $x_b^* \setminus \{x_j^{*b}\}$, restringido a Θ_i.

3. Obtenemos el estimador final de π_{opt},

$$\widetilde{\pi}_1 = \frac{1}{B} \sum_{b=1}^{B} \pi_b.$$

3.3.2. Simulaciones

Se obtuvieron muestras aleatorias de tamaño $n = 100$ en todas las simulaciones. Cada muestra con distribución de valores extremos generalizada con parámetro de localización $\mu = 3$, de escala $\sigma = 1$ y haciendo variar el parámetro de forma κ. Para analizar la convergencia de los métodos de optimización, se propusieron varios valores para el punto inicial, y así observar el comportamiento del estimador. El estimador $\tilde{\pi}$ se obtuvo con el método sin regularización, es decir, en todos los casos se obtiene $\tilde{\pi}_1$, que por facilidad en la notación sólo se pondrá $\tilde{\pi}$. A continuación se muestran los resultados de las optimizaciones en cada una de las simulaciones.

- Caso $\kappa = 0.05$, por lo que $\pi_{opt} = (0, 1, 0)$. Variando el punto inicial π_0 para la optimización, se obtuvieron los siguientes resultados:

Valor inicial	$\tilde{\pi}$
$\pi_0 = (0.15, 0.7, 0.15)$	$(0, 1, 0)$
$\pi_0 = (0.3, 0.4, 0.3)$	$(0, 1, 0)$
$\pi_0 = (0.05, 0.05, 0.9)$	$(0.0001, 0.821, 0.1789)$

Tabla 4. Valores obtenidos de $\tilde{\pi}$ con diferentes valores iniciales para la optimización.

- Caso $\kappa = -0.05$, por lo que $\pi_{opt} = (0, 0, 1)$. Variando el punto inicial π_0 para la optimización, se obtuvieron los siguientes resultados:

Valor inicial	$\tilde{\pi}$
$\pi_0 = (0.15, 0.15, 0.7)$	$(0, 0, 1)$
$\pi_0 = (0.3, 0.3, 0.4)$	$(0, 0, 1)$

Tabla 5. Valores obtenidos de $\tilde{\pi}$ con diferentes valores iniciales para la optimización.

- Caso $\kappa = 0$, por lo que $\pi_{opt} = (1, 0, 0)$. Variando el punto inicial π_0 para la optimización, se obtuvieron los siguientes resultados:

Valor inicial	$\tilde{\pi}$
$\pi_0 = (0.7, 0.15, 0.15)$	$(0.9999, 0, 0.0001)$
$\pi_0 = (0.4, 0.3, 0.3)$	$(0.9998, 0, 0.0002)$
$\pi_0 = (0.05, 0.05, 0.9)$	$(0.9669, 0, 0.0331)$

Tabla 6. Valores obtenidos de $\tilde{\pi}$ con diferentes valores iniciales para la optimización.

3.4. Comentarios finales

Se realizaron varias simulaciones para estudiar el comportamiento del estimador $\tilde{\pi}$ en el contexto de ponderación de hipótesis, teniendo como fin comparar el enfoque propuesto en esta tesis con algunos procedimientos usuales para este problema. Primero, se comparó con un enfoque bayesiano poniendo como previa una distribución tipo Jeffreys. También, se comparó con el procedimiento

clásico frecuentista por pruebas de significancia (p-valor) y con el enfoque de intervalos de verosimilitud. Por último, se realizaron simulaciones para probar el enfoque aplicado a la distribución de valores extremos generalizada, cuando el interés estadístico recae en el parámetro de forma. De todos estos estudios de simulación salen a relucir comentarios en varios sentidos.

Por una parte, de las simulaciones realizadas para comparar con el enfoque bayesiano, se tiene que a pesar de tener evidencia suficiente para rechazar H_0, utilizando la distribución previa tipo Jeffreys se obtiene $\mathbb{P}(H_0 \mid x)$ muy alta; más aún, esta probabilidad crece cuando el tamaño de muestra n crece también. A pesar de obtener una estimación $\tilde{\pi}$ que también crece conforme n, esta estimación no es mayor a 0.5, desfavoreciendo H_0. Sin embargo, para poder aseverar que $\tilde{\pi}$ no es mayor que este umbral, es necesario hacer un estudio con valores de n más grande.

Ahora bien, de la Figura 8 se puede concluir que tanto el p-valor como la verosimilitud relativa logran detectar la falsedad de la hipótesis H_0 incluso con valores de θ_0 muy cercanos a 0. Como ya se mencionó, estos dos procedimientos presentan algunas dificultades. El p-valor no cuenta con una medida universal para decidir cuándo es suficientemente pequeño o no. Además, la escala en la que se mide la verosimilitud relativa no es intuitiva y depende del valor del EMV y no en sí del verdadero valor de θ. La propuesta de la tesis provee de una interpretación clara a $\tilde{\pi}$, que es la ponderación estimada de $H_0 : \theta = \theta_0$, dependiente de los datos observados x.

De las simulaciones realizadas para analizar el comportamiento del estimador $\tilde{\pi}$ en la aplicación a la DVEG se obtuvo que a pesar de poner un valor inicial π_0 muy alejado del valor óptimo verdadero, el estimador que se obtuvo fue muy cercano a dicho valor óptimo. En la práctica, se desconocen los verdaderos valores de los parámetros, este enfoque permite obtener una estimación mediante una mezcla, librándose de la limitante de tomar una estimación puntual. Además, este enfoque resulta especialmente relevante para el caso en el que el parámetro de forma es cero. Con una buena muestra representativa, un intervalo de confianza contendría tanto valores negativos como positivos. La ventaja de la propuesta de la tesis es que $\tilde{\pi}$ estima la probabilidad de que efectivamente $\theta = 0$, asignando pesos correspondientes a los casos $\theta < 0$ y $\theta > 0$.

No obstante los alentadores resultados obtenidos en este capítulo, se considera que para poder utilizar la propuesta de esta tesis como mecanismo práctico estándar para toma de decisiones con

respecto a una familia de hipótesis, es necesario un estudio más profundo, involucrando mayor variedad de escenarios estadísticos.

Conclusiones

C1 Para modelos estadísticos paramétricos estructurados en múltiples hipótesis, un criterio que permite otorgar un sentido probabilístico preciso y fácilmente interpretable a la ponderación probabilística de hipótesis, válido tanto para modelos no bayesianos como bayesianos, es el siguiente: la esperanza de la divergencia de Kullblack-Leibler de la mezcla de las densidades máximo verosímiles asociadas a las hipótesis según la mezcladora definida por las ponderaciones respecto a la densidad de una muestra futura (de tamaño m definido).

C2 Versiones empíricas de dicho criterio (C1), como la versión bootstrap, ofrecen un enfoque factible para la construcción de ponderaciones de hipótesis (o densidades inferenciales o mezcladoras) para modelos estructurados en múltiples hipótesis.

C3 Las ponderaciones de hipótesis obtenidas mediante (C2) heredan la interpretación probabilística precisa y clara de su contraparte teórica o poblacional (C1). Esto se confirma con aceptable aproximación por resultados de simulación en diversos escenarios presentados en esta tesis.

C4 En el caso de la estructura más sencilla constituida por una hipótesis simple versus una hipótesis alternativa simple, los resultados de simulación obtenidos avalan que, en el sentido del criterio (C1), presenta buen comportamiento la mezcladora dada por las frecuencias relativas de veces que el estimador por máxima verosimilitud coincide con cada hipótesis.

C5 Dada una muestra original de datos de tamaño n, si el interés del investigador radica, no en estimar la densidad de una muestra futura de tamaño especificado m, sino en usar las ponderaciones obtenidas según (C2) como índice de evidencia estadística a favor de las distintas hipótesis, conviene tomar en el método de construcción (C2) un tamaño de muestra futura $m = n$.

C6 Los resultados de simulación en escenarios de hipótesis simple vs. alternativa compuesta para la media de una distribución Normal muestran que el método de ponderación de hipótesis propuesto en esta tesis se comporta ventajosamente en comparación con diversas alternativas existentes, tanto no bayesianas como bayesianas, en el sentido de brindar resultados más acordes con lo que estadísticamente es razonable esperar.

C7 Las simulaciones en escenarios de tres hipótesis para el parámetro de forma de la distribución de valores extremos generalizada muestran un comportamiento del método de ponderación de hipótesis propuesto que resulta también alentadoramente interesante y razonable.

Bibliografía

[1] Akaike, H. (1977). An objective use of bayesian models. *Ann. Inst. Statist. Math*, 29, Part A pp 9-20.

[2] Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, 65, 1, pp. 53-9.

[3] Austin, R. S., Dialsingh, I. y Altman, N. S. (2014). Multiple Hypothesis Testing: A Review. *Journal of the Indian Society of Agricultural Statistics*. 68. 303-314.

[4] Berger, J. O. y Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, Vol. 82, No. 397, pp. 112-122.

[5] Bernardo, J. M. y Rueda, R. (2002). Bayesian Hypothesis Testing: A Reference Approach. *International Statistical Review / Revue Internationale De Statistique*, 70(3), 351-372.

[6] Bernardo, J. M. y Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons.

[7] Bjornstad, J. F. (1990). Predictive Likelihood: A Review. *Statistical Science*, Vol. 5, No. 2, pp. 242-254.

[8] Brown, L. D., George, E. I. y Xu, X. (2008). Admissible predictive density estimation. *The Annals of Statistics*, Vol. 36, No. 3, 1156-1170.

[9] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

[10] Cover, T. M. y Thomas, J. A. (2006). *Elements of Information Theory, 2nd. Edition*. John Wiley & Sons.

[11] Dudoit, S., Shaffer, J. P. y Boldrick, J. C. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18(1), 71-103.

[12] Gutiérrez-Peña, E. y Walker, S.G. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review* 73, 309-330.

[13] Edwards, A. W. F. (1992). *Likelihood, expanded edition*. Cambridge University Press.

[14] Harris, I. N. (1989). Predictive fit for natural exponential families. *Biometrika*, 76, 4, pp. 675-84.

[15] Jeffreys, H. (1961). *Theory of Probability (3rd. ed.)*. Oxford University Press.

[16] Kullback, S. y Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79-86.

[17] Lindsay, B.G. (1983). The geometry of mixture likelihoods: a general theory. *Annals of Statistics* 11, 86-94.

[18] Lovric, M. (2011). *International Encyclopedia of Statistical Science*. Springer.

[19] Narayan, M., Allen, G. I. y Tomson, S. N. (2015). Two sample inference for populations of graphical models with applications to functional connectivity. *arXiv:1502.03853v1*.

[20] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

[21] Walker, S.G., Gutiérrez-Peña, E. y Muliere, P. (2001). A decision theoretic approach to model averaging. *The Statistician* 50, 31-39.

[22] Wasserstein, R. L. y Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, DOI: 10.1080/00031305.2016.1154108