UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Selección de Modelos en Teoría de Valores Extremos

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A:

CARMEN JOSEFINA AYALA MACÍAS

DIRECTOR DE TESIS:
DR. RAÚL RUEDA DÍAZ DEL CAMPO

Ciudad Universitaria, Cd. de México.

2017

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

FACULTAD DE CIENCIAS
Secretaría General
División de Estudios Profesionales

Votos Aprobatorios

DR. ISIDRO ÁVILA MARTÍNEZ
Director General
Dirección General de Administración Escolar
P r e s e n t e

Por este medio hacemos de su conocimiento que hemos revisado el trabajo escrito titulado:

Selección de Modelos en Teoría de Valores Extremos

realizado por **AYALA MACÍAS CARMEN JOSEFINA** con número de cuenta **0-9653196-1** quien ha decidido titularse mediante la opción de **tesis** en la licenciatura en **Actuaría**. Dicho trabajo cuenta con nuestro voto aprobatorio.

Propietario	Dr. Ramsés Humberto Mena Chávez	
Propietario	Mat. Margarita Elvira Chávez Cano	
Propietario Tutor	Dr. Raúl Rueda Díaz del Campo	
Suplente	Act. Jaime Vázquez Alamilla	
Suplente	Dr. Eduardo Arturo Gutiérrez Peña	

Atentamente
"POR MI RAZA HABLARÁ EL ESPÍRITU "
Ciudad Universitaria, D. F., a 26 de junio de 2014
EL JEFE DE LA DIVISIÓN DE ESTUDIOS PROFESIONALES

ACT. MAURICIO AGUILAR GONZÁLEZ

Señor sinodal: antes de firmar este documento, solicite al estudiante que le muestre la versión digital de su trabajo y verifique que la misma incluya todas las observaciones y correcciones que usted hizo sobre el mismo.
MAG/mdm

Dedicatoria y Agradecimientos

A mi papá, con todo mi cariño y añoranza.

A mis hijos, Emilio y Patricio, inagotable fuente de amor y alegría.

A David, gracias por siempre estar, por darme la mano y por compartir sueños, alegrías y también tristezas, que siempre son más llevaderas a tu lado.

A mi mamá, con mucho cariño y agradecimiento por lo que sigues haciendo por mi.

A mi hermana Mirtha, por todos los momentos compartidos y el amor que me das.

A mi asesor, el Dr. Raúl Rueda, por todo el tiempo, paciencia y trabajo que dedicó para que concluyera este trabajo. No tengo palabras para expresar todo mi agradecimiento y aprecio.

A mis Sinodales: Dr. Ramsés H. Mena, Mat. Margarita Chávez, Act. Jaime Vázquez y Dr. Eduardo A. Gutiérrez, por su disposición para ayudarme en la revisión de este trabajo, por los valiosos comentarios que me hicieron para mejorarlo, y por el tiempo que le dedicaron.

Índice general

Abreviaturas y símbolos principales

M_n	máximo de $\{X_1, \ldots, X_n\}$, $n \geq 2$
c.s.	casi seguramente
$\xrightarrow{\mathcal{D}}$	convergencia en distribución
$\stackrel{d}{=}$	equivalencia en distribución
$\mathbb{E}$	esperanza
$\mathbb{E}(X\|G)$	esperanza condicional
F	función de distribución
$\overline{F}$	cola de F
$F^{-1}(y)$	función inversa generalizada
$f(t)$	función de densidad
i.i.d.	independiente e idénticamente distribuidas
$\xrightarrow{P}$	convergencia en probabilidad
$\xrightarrow{c.s.}$	converge casi seguro
$\mathbb{P}$	medida de probabilidad
$PD(\alpha, F)$	proceso Dirichlet con parámetros α, F
$\mathbb{N}$	el conjunto de números naturales
$\mathbb{R}$	el conjunto de números reales
$X_{(n)}$	n-ésima estadística de orden
x_F	punto final a la derecha de F
v.a.	variable aleatoria

Introducción

La importancia de los eventos o valores extremos radica en que, si bien se presentan con una probabilidad muy baja, su ocurrencia tiene un gran impacto en el fenómeno de estudio. La teoría tiene aplicaciones en muchas áreas, como ingeniería, ciencias ambientales, y en los últimos años en finanzas y seguros. Parte de la teoría fue desarrollada para resolver problemas relacionados con el diseño de estructuras que deben resistir algún fenómeno ambiental; si el fenómeno es de gran intensidad, la estructura fallará, por lo tanto es necesario diseñarla de modo que resista fenómenos meteorológicos extremos y que la probabilidad de falla sea pequeña.

La teoría de valores extremos busca extrapolar la información que proveen los datos para estimar la probabilidad de ocurrencia e intensidad de un evento extremo, con la problemática adicional de que existen pocas observaciones disponibles de eventos extremos. Con frecuencia queremos estimar valores que van más allá del máximo valor de la muestra, y las técnicas estándar de estimación de densidades ajustan bien donde los datos tienen mayor densidad, pero pueden tener sesgos importantes al estimar las colas.

Todas las conclusiones o resultados estadísticos que se obtienen están condicionadas por el modelo probabilístico y por el o los parámetros que se hayan especificado. La selección del modelo que mejor explique el fenómeno de valores extremos es, por lo tanto, fundamental en el análisis estadístico que de él se haga.

En este trabajo se aborda el problema de valores extremos desde la perspectiva Bayesiana, lo que permite, entre otras cosas, contrarrestar la escasez de datos extremos con el conocimiento que se tenga *a priori* del fenómeno en estudio, para finalmente encontrar la distribución predictiva que nos permita hacer inferencia sobre el comportamiento futuro de dicho fenómeno. A este enfoque se incorpora el uso de herramientas de selección de modelos, con

lo que se busca elegir al modelo que represente de mejor manera un cierto proceso, partiendo de un conjunto de modelos contendientes.

Capítulo 1

Teoría de valores extremos

Existen dos cuestiones fundamentales a las que se busca dar respuesta al estudiar el comportamiento de los eventos extremos. La primera de ellas se refiere a la frecuencia con la que ocurren dichos eventos; la segunda, al tamaño o severidad del evento en cuestión. Al contar con esta información, se pueden predecir y prevenir muchos eventos catastróficos; de ahí la creciente importancia que se le ha dado a la teoría de valores extremos desarrollada en las últimas décadas.

Al estudiar el comportamiento de los eventos extremos la primera pregunta que surge es ¿cómo ocurren estos eventos? Esta pregunta ha llevado a muchos estadísticos a buscar los métodos matemáticos apropiados para explicar eventos que ocurren con una probabilidad relativamente pequeña, pero que tienen una influencia significativa en el modelo que describe el comportamiento global de los datos.

En el campo de los eventos extremos, los modelos de movimiento browniano, los procesos Poisson homogéneos y el proceso de caminata aleatoria, por ejemplo, forman parte fundamental de la teoría probabilística y de la teoría de valores extremos clásica, de los cuales se derivan muchos otros modelos.

Los eventos extremos que ocurren en estos problemas son descritos a través de algunas distribuciones y procesos estocásticos. En este capítulo veremos algunas de las distribuciones más importantes, como la de valores extremos generalizada, la Pareto generalizada y la clase de distribuciones subexponenciales.

Muchos de los resultados que se presentarán se basan en leyes límite y

métodos asintóticos, por lo que serán tratados también en este capítulo. Para profundizar más en el tema, véase Embrechts *et al* (1997).

§1.1. El rol de la teoría de valores extremos y sus áreas de aplicación

Como la gran mayoría de las áreas de la estadística, la teoría de valores extremos trata de hacer extrapolación a partir de los datos disponibles. Reduciendo el problema a su forma más simple, se cuenta con una serie de observaciones independientes $X_1, \ldots, X_n$ de una función de distribución F desconocida, de la cual se debe estimar la cola de la manera más precisa posible. La dificultad de este problema radica en que la mayor parte de los datos se concentran al centro de la distribución. Por definición, los datos extremos no son fáciles de observar, lo que dificulta su estimación.

Los principales aspectos a considerar en este problema son:

(a) Existen muy pocas observaciones en la cola de la distribución;

(b) Muchas veces se requieren estimaciones más allá de X_{max} o X_{min}, es decir, la observación más grande o más pequeña de los datos observados;

(c) Los métodos comunes para estimar la densidad se ajustan bien donde los datos tienen mayor densidad, pero pueden presentar grandes sesgos al estimar las probabilidades de la cola.

El papel de la teoría de valores extremos es desarrollar técnicas científica y estadísticamente sustentadas que permitan estimar el comportamiento extremo de procesos o variables aleatorias, por lo que esta teoría tiene aplicaciones en varios campos para modelar fenómenos y datos de la vida real. En seguros, por ejemplo, se pueden encontrar modelos para el tamaño de demandas o de pérdidas; en finanzas se pueden modelar las colas pesadas de las distribuciones del tipo de cambio o del rendimiento de ciertos activos en el mercado; en las ciencias ambientales las aplicaciones varían desde la hidrología en el modelo de inundaciones, hasta la geología en el estudio de depósitos minerales.

Otra aplicación estándar de los modelos de valores extremos es en teoría de la confiabilidad. Conceptualmente, los componentes de un sistema están hechos de componentes más pequeños, por lo que la falla de uno de los componentes pequeños ocasiona la falla del sistema global. Este principio es conocido como "el componente más débil", y su importancia radica en que la probabilidad de falla del sistema depende fundamentalmente de la probabilidad de falla del componente más débil, por lo que es necesario un modelo preciso de la cola de la distribución de los tiempos de falla de cada componente.

Más adelante veremos que la forma en que se desarrolla la teoría de valores extremos, al menos en su forma más simple, es mediante un argumento análogo al del teorema central del límite, pero aplicado a máximos muestrales en lugar de sumas.

El análisis de valores extremos se realiza generalmente bajo alguno de los siguientes enfoques:

1. La teoría clásica de valores extremos, cuyos modelos describen el comportamiento estadístico de M_n=max$\{X_1, \ldots, X_n\}$, y

2. Los modelos de picos sobre el umbral, que hacen uso de todos los datos o valores extremos disponibles, entendiendo por valores extremos a aquéllos que exceden un umbral alto.

En las siguientes secciones se explica con mayor detalle estos dos enfoques.

§1.2. Teoría clásica de valores extremos

Supongamos que $X_1, \ldots, X_n$ es una sucesión de v.a.i.i.d.'s con función de distribución F. Una forma simple de describir el comportamiento de los valores extremos, es considerar el comportamiento de las estadísticas de orden máximas

$$M_n = \max\{X_1, \ldots, X_n\}, \quad n \geq 2. \tag{1.1}$$

En la práctica, las X_i's generalmente representan observaciones de algún proceso medido en una escala de tiempo, como el nivel semanal del agua en una presa, la temperatura diaria en una cierta región, etc., de manera

que M_n representa el máximo del proceso sobre n unidades de tiempo de observación. Por ejemplo, si n es el número de observaciones en un año, M_n corresponde al máximo anual.

La mayoría de los resultados de valores extremos se obtienen al considerar este máximo; sin embargo se pueden obtener resultados equivalentes para el mínimo a partir de la siguiente igualdad

$$\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n). \tag{1.2}$$

Una vez definido M_n y conociendo la función de distribución de X, resulta sencillo encontrar la distribución del máximo

$$\begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, \dots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_n \leq x) \\ &= F^n(x), \quad x \in \mathbb{R}, \quad n \in \mathbb{N}. \end{aligned}$$

En la práctica, sin embargo, la dificultad radica en que la función de distribución F es desconocida. Una posible solución sería estimar a la función de distribución F a partir de los datos observados para después sustituirla en la ecuación anterior; sin embargo, dada la naturaleza del problema, las pequeñas discrepancias en la estimación de F pueden significar discrepancias importantes en la estimación de F^n. Un método alternativo es aceptar que F^n es desconocido y buscar familias de modelos para F^n que puedan ser estimadas a partir de los datos extremos solamente.

Si bien es cierto que existen algunas cotas para el comportamiento de M_n, éstas son muy grandes cuando se trata de aplicaciones prácticas, lo que da origen a un método basado en argumentos asintóticos. De manera más específica, veremos qué distribuciones límite son factibles para M_n conforme $n \to \infty$, y utilizaremos esta familia de distribuciones como una aproximación a la distribución de M_n para n finita pero grande.

Comenzaremos por decir que los valores extremos se presentan en el extremo derecho de la cola de la distribución, lo que intuitivamente nos indica que el comportamiento de M_n está relacionado con el *punto final a la derecha* de la cola de la función de distribución F.

Definición 1.1. Punto final a la derecha de F

Se define a

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}, \tag{1.3}$$

como el punto final a la derecha de F.

De la definición anterior se tiene que para todo $x < x_F$,

$$\mathbb{P}(M_n \leqslant x) = F^n(x) \to 0, \quad n \to \infty. \tag{1.4}$$

Por otra parte, cuando $x_F < \infty$, y $x \geqslant x_F$

$$\mathbb{P}(M_n \leqslant x) = F^n(x) = 1. \tag{1.5}$$

Por lo tanto $M_n \xrightarrow{P} x_F$ conforme $n \to \infty$, donde $x_F \leqslant \infty$. Como la sucesión (M_n) es no decreciente en n, converge casi seguramente, y por lo tanto podemos concluir que

$$M_n \xrightarrow{c.s.} x_F, \quad n \to \infty. \tag{1.6}$$

Lo anterior, sin embargo, no proporciona la información suficiente que nos permita determinar las posibles leyes límite para el máximo M_n de la sucesión (X_n). Este problema de valores extremos puede considerarse análogo al problema del teorema central del límite, como se verá a continuación.

§1.2.1. Valores extremos vs. sumas

Una de las razones por las que resulta interesante comparar el teorema central del límite con algunos de los resultados importantes de la teoría de valores extremos es el uso de herramientas matemáticas similares, además de que los resultados derivados del teorema central del límite nos llevan a encontrar nuevas formas de resolver problemas relacionados con eventos extremos.

Comencemos por la formulación general del teorema central del límite.

Supongamos que $X_1, \ldots, X_n$ es una sucesión de v.a.i.i.d.'s con función de distribución F, y definamos la suma parcial $S_n = X_1 + \ldots + X_n$. El teorema central del límite da solución a los siguientes problemas:

- Dada F, encontrar las constantes $a_n > 0$ y $b_n \in \mathbb{R}$ tales que

$$\frac{S_n - b_n}{a_n} \xrightarrow{d} Y, \quad n \to \infty, \tag{1.7}$$

  donde Y es una variable aleatoria no degenerada con función de distribución G.

- Caracterizar la función de distribución G de Y en (1.7).

- Dada una posible función de distribución G en (1.7), encontrar todas las funciones de distribución F que satisfacen (1.7) (problema del dominio de atracción), caracterizando las sucesiones (a_n) y (b_n).

Los problemas anteriores pueden resolverse y dar origen a nuevos resultados en el campo de las sumas. Por ejemplo, bajo la condición general de momentos $\mathbb{E}[X^2] < \infty$ el teorema central del límite nos lleva a la función de distribución Normal estándar ($G = N(0,1)$), con $a_n = \sqrt{n}\sigma$, $b_n = n\mu$ y todas las funciones de distribución con segundo momento finito son atraídas a la N(0,1). Cuando la condición anterior no se cumple ($\mathbb{E}[X^2] = \infty$), se trabaja con una clase de distribuciones límite relativamente pequeña conocida como α-estable. Únicamente en ese caso de colas pesadas las condiciones en la cola de la distribución garantizan la existencia de una distribución límite.

La solución general a 1.7 está dada por la clase de distribuciones estables, como se verá a continuación.

Definición 1.2. Distribuciones y variables aleatorias estables.

Sean X, X_1 y X_2 v.a.i.i.d.'s. Una variable aleatoria o función de distribución es conocida como estable si satisface la siguiente identidad

$$c_1 X_1 + c_2 X_2 \stackrel{d}{=} b(c_1, c_2)X + a(c_1, c_2), \tag{1.8}$$

para todos los números c_1 y c_2 no negativos y los números reales apropiados $b(c_1, c_2) > 0$ y $a(c_1, c_2)$.

Consideremos ahora la suma de variables aleatorias estables. De (1.8) tenemos que, para algunas constantes a_n y $b_n > 0$ y $X = X_1$,

$$S_n = X_1 + \ldots + X_n \stackrel{d}{=} b_n X + a_n, \quad n \geq 1,$$

que puede escribirse como

$$b_n^{-1}(S_n - a_n) \stackrel{d}{=} X,$$

De lo anterior se concluye que, si una distribución es estable, entonces es la distribución límite para las sumas de v.a.i.i.d.'s; pero ¿existen otras distribuciones límite? La respuesta la da el siguiente teorema.

Teorema 1.1. *Propiedad límite de las leyes estables.*
La clase de las distribuciones estables coincide con la clase de todas las leyes límite posibles para la suma de v.a.i.i.d.'s.

Los elementos de las distribuciones estables están caracterizados principalmente por un parámetro $\alpha \in (0, 2]$. El caso $\alpha = 2$ corresponde a la distribución Normal, $\alpha = 1$ a la distribución Cauchy, y $\alpha = 1/2$ a la distribución Levy.

Existen varias razones para utilizar distribuciones estables en la descripción de un proceso. La primera de ellas es que existen razones teóricas sólidas para esperar un modelo estable no Gaussiano. La segunda razón es el teorema central del límite generalizado, que establece que el único límite no trivial posible para las sumas de términos independientes e idénticamente distribuidos es estable, por lo que se debería usar un modelo estable en los casos en que las cantidades observadas correspondan a la suma de otros términos (como el precio de una acción, el ruido en un sistema de comunicación, etc.). El tercer argumento para modelar con distribuciones estables es empírico: además de lo que establece el teorema central del límite, existe evidencia empírica utilizada por algunos autores para justificar el uso de modelos estables al trabajar con conjuntos grandes de datos que tienen sesgos y colas pesadas (Nolan, 2005). Algunos ejemplos del uso de las distribuciones estables con este tipo de datos en las áreas de economía y finanzas se pueden encontrar en Mandelbrot (1963), Fama (1965), Embrechts et al. (1997), Rachev, Mittnik y Paolella (2000), entre muchos otros. Estos datos son vagamente descritos por modelos Gaussianos, pero pueden describirse de manera precisa por una distribución estable. El uso de estas distribuciones al modelar datos extremos y distribuciones de colas pesadas puede consultarse en Embrechts et al. (1997), Adler et al. (1998), y en Reiss & Thomas (2001).

En contraste con las sumas, al trabajar con máximos es necesario que se cumplan condiciones más específicas en la cola de la distribución, $\overline{F} = 1 - F$, para asegurar que $\mathbb{P}(M_n \leqslant u_n)$ converja a un límite no trivial, es decir, un número en $(0, 1)$. Pero ¿cuáles son estas condiciones? En primer lugar se observa que son necesarias ciertas condiciones de continuidad hacia el punto final a la derecha de F, lo que deja fuera a varias distribuciones importantes. Por ejemplo, si F tiene una distribución Poisson, $\mathbb{P}(M_n \leq u_n)$ no tiene límite en $(0, 1)$ sin importar cuál sea la sucesión (u_n), lo que significa que los máximos normalizados de v.a.i.i.d.'s Poisson no tienen una distribución límite no degenerada.

De manera similar a lo que se ha visto hasta ahora en el campo de las sumas, en la siguiente sección se buscará caracterizar a las posibles leyes límite para el máximo M_n bajo transformaciones afines positivas del tipo $(c_n)^{-1}(M_n - d_n)$, es decir, mediante la solución de un problema análogo al teorema central del límite.

§1.2.2. Convergencia débil de máximos bajo transformaciones afines

Considerando la analogía existente entre el problema de valores extremos para caracterizar las leyes límite del máximo M_n y el problema del teorema central del límite, plantearemos el problema sobre máximos bajo el mismo esquema que se siguió al encontrar leyes límite para las sumas en el problema del teorema central del límite. Así, se busca determinar las distribuciones que satisfacen, para todo $n \geqslant 2$, la identidad

$$\text{máx}(X_1, \ldots, X_n) \stackrel{d}{=} c_n X + d_n, \tag{1.9}$$

para constantes apropiadas, $c_n > 0$ y $d_n \in \mathbb{R}$.

Esto no es más que buscar la clase de distribuciones que se ajustan a los máximos bajo transformaciones afines. Como se vio anteriormente, en el caso de las sumas centradas y estandarizadas las distribuciones estables son las únicas leyes límite posible. De igual forma, existe una clase de distribuciones que cumplen con esta propiedad para máximos centrados y estandarizados.

Definición 1.3. *Distribuciones max-estables.*
Una variable aleatoria X no degenerada (así como su función de distribución) es llamada max-estable si satisface (1.9) para v.a.i.i.d.'s $X, X_1, \ldots, X_n$, constantes apropiadas $c_n > 0$, $d_n \in \mathbb{R}$ y todo $n \geq 2$.

Supongamos que (X_n) es una sucesión de v.a.i.i.d. max-estables, entonces podemos escribir (1.9) como

$$c_n^{-1}(M_n - d_n) \stackrel{d}{=} X. \tag{1.10}$$

de donde se concluye que toda distribución max-estable es una distribución límite para el máximo de v.a.i.i.d.'s. Más aún, las distribuciones max-estables son las únicas leyes límite para máximos estandarizados.

Teorema 1.2. *Probabilidad límite de las leyes max-estables.*
La clase de distribuciones max-estables coincide con la clase de todas las leyes límite posibles (no degeneradas) para máximos apropiadamente estandarizados de v.a.i.i.d.'s.

El siguiente resultado es la base de la teoría clásica de valores extremos y en donde se resumen los resultados más importantes.

Teorema 1.3. *Teorema de Fisher-Tippet (Fisher & Tippet, 1928).*

Sea $\{X_n : n \in \mathbb{N}\}$ *una sucesión de v.a.i.i.d.'s y* $M_n = max\{x_1, \ldots, x_n\}$ $\forall\ n \in \mathbb{N}$. *Si existen constantes normalizadoras* $c_n > 0$, $d_n \in \mathbb{R}$ *y alguna función de distribución no degenerada* H *tal que*

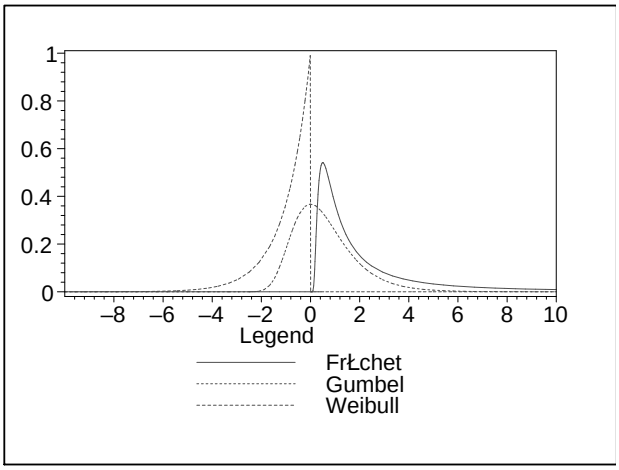
$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H \tag{1.11}$$

entonces H *pertenece a uno de los siguientes tres tipos de funciones de distribución*

$$\textit{Fréchet:}\quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp\left\{-x^{-\alpha}\right\}, & x > 0 \end{cases} ;\ \alpha > 0$$

$$\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\}, & x \leq 0 \\ \\ 0, & x > 0 \end{cases} \quad ; \ \alpha > 0$$

$$\textit{Gumbel:} \quad \Lambda(x) = \exp\{-e^{-x}\}, \quad x \in \mathbb{R}.$$



La demostración de este teorema no es trivial y está fuera de los objetivos de este trabajo. Un esbozo de la demostración se puede encontrar en Embrechts, Klüppelberg, Mikosch, *et al.* (1997).

En términos menos formales, el Teorema 1.3 establece que el máximo muestral normalizado $(M_n - d_n)/c_n$ converge en distribución a una variable aleatoria cuya distribución pertenece a las familias Gumbel, Weibull o Fréchet. Lo anterior implica que, una vez que se estabiliza a M_n con las sucesiones adecuadas $\{c_n\}$ y $\{d_n\}$, la distribución límite de la variable normalizada M_n debe ser alguna de las distribuciones de valores extremos, como se denomina al conjunto de distribuciones Gumbel, Weibull y Fréchet.

Cada una de estas familias tiene parámetros de localización y escala, d_n y c_n, respectivamente, y las distribuciones Fréchet y Weibull tienen un parámetro de forma adicional.

Lo más importante de este teorema es que los tres tipos de distribuciones de valores extremos son los únicos límites posibles para las distribuciones de

M_n, sin importar cuál sea la distribución F de la población. En este sentido, el teorema nos brinda una representación para valores extremos análoga al teorema central del límite.

§1.2.3. La distribución de valores extremos generalizada

La forma que adoptan las distribuciones de valores extremos depende del comportamiento de la cola de la distribución de F (la función de distribución de las $X_i's$), para lo cual resulta útil analizar el comportamiento de la distribución F en x_F. En el caso de la distribución Weibull x_F es finito, contrario a lo que sucede en las distribuciones Gumbel y Fréchet. Por otra parte, la densidad de H decae exponencialmente cuando se trata de la distribución Gumbel y polinomialmente cuando se trata de la Fréchet. Esto significa que, en la práctica, las tres familias nos llevan a representaciones muy diferentes del comportamiento de los valores extremos.

En las primeras aplicaciones de la teoría de valores extremos, era usual elegir una de las tres familias y posteriormente estimar los parámetros relevantes de dicha distribución. Esto daba origen a dos problemas: en primer lugar, se necesita de una técnica para elegir cuál de las tres familias es la más apropiada considerando los datos; en segundo lugar, en cualquier inferencia que se hace una vez que se eligió una familia, se supone que se hizo la elección correcta, por lo que no permite incorporar ningún tipo de incertidumbre respecto a la elección que se hizo, aún cuando dicha incertidumbre puede ser sustancial.

Un mejor análisis puede obtenerse de una reformulación de los modelos anteriores, ya que se puede verificar que las familias Gumbel, Weibull y Fréchet pueden representarse en una sola familia con función de distribución

$$H(x) = \exp\Big\{ -\Big[1 + \xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}, \tag{1.12}$$

definida en el conjunto $\{x : 1 + \xi(x-\mu)/\sigma > 0\}$, donde $-\infty < \mu < \infty$, $\sigma > 0$ y $-\infty < \xi < \infty$. Esta familia es conocida como la distribución de valores extremos generalizada y está compuesta por tres parámetros: el de localización, μ; el de escala, σ; y el de forma, ξ. Las distribuciones Fréchet

y Weibull corresponden a los casos $\xi > 0$ y $\xi < 0$, respectivamente. El caso $\xi = 0$ es interpretado como el límite de la distribución (1.12) cuando $\xi \to 0$, lo que nos lleva a la distribución Gumbel.

§1.2.4. Dominio de atracción al máximo y constantes normalizadoras

Hasta ahora hemos identificado a las distribuciones de valores extremos como las leyes límite para máximos normalizados de v.a.i.i.d.'s. El siguiente paso es examinar las condiciones que debe cumplir la función de distribución F de manera que los máximos normalizados converjan en distribución a H, la distribución de valores extremos. Lo anterior guarda estrecha relación con la forma de elegir las constantes normalizadoras $c_n > 0$ y $d_n \in \mathbb{R}$ tales que

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H. \tag{1.13}$$

Es importante notar que el teorema de convergencia a tipos (ver Embrechts, P., Klüppelberg, C. y Mikosch, T. (1997), pg. 554) asegura que, salvo por transformaciones afines de la forma $G(x) = G(ax + b)$, las leyes límite están determinadas de manera única, por lo que un cambio en las constantes normalizadoras no derivará en la convergencia del máximo a un límite diferente. Más aún, se puede agrupar a todas las funciones de distribución F que compartan la misma distribución límite estable en una clase, llamada dominio de atracción al máximo.

Definición 1.4. Dominio de atracción al máximo.

Decimos que la v.a. X (la función de distribución de X, o la distribución de X) pertenece al dominio de atracción al máximo de la distribución de valores extremos H, si existen constantes $c_n > 0$, $d_n \in \mathbb{R}$ tales que

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H \tag{1.14}$$

o, de manera equivalente,

$$\lim_{n\to\infty} \Pr(M_n \leq c_n x + d_n) = \lim_{n\to\infty} F^n(c_n x + d_n) = H(x) \tag{1.15}$$

Notación: $X \in DAM(H)$ o $F \in DAM(H)$.

Adicionalmente, se establece una relación de equivalencia de colas para las funciones de distribución que, al cumplirse, permite observar ciertas propiedades relacionadas con el dominio de atracción al máximo.

Definición 1.5. Equivalencia de colas.

Dos funciones de distribución F y G son llamadas de colas equivalentes si su punto final es el mismo, es decir, $x_F = x_G$, y

$$\lim_{x\to x_F} \overline{F}(x)/\overline{G}(x) = c \tag{1.16}$$

para alguna constante $0 < c < \infty$.

En colas equivalentes cada dominio de atracción al máximo es cerrado, es decir, para dos funciones de distribución F y G de colas equivalentes, $F \in DAM(H)$ si y sólo si $G \in DAM(H)$.

Otra propiedad importante, y de gran ayuda al momento de calcular las constantes normalizadoras, es el hecho de que dos funciones de distribución de colas equivalentes pueden utilizar las mismas constantes normalizadoras.

§1.2.5. Algunos resultados sobre el dominio de atracción al máximo de la distribución de valores extremos

Supongamos que se tiene una sucesión de v.a.i.i.d.'s $X_1, X_2, \ldots$ de una función de distribución desconocida F, y que además podemos encontrar sucesiones de números reales $a_n > 0$ y b_n tales que

$$\Pr[(M_n - b_n)/a_n \leq x] = F^n(a_n x + b_n) \xrightarrow{d} H(x), n \to \infty, \tag{1.17}$$

para alguna función no degenerada H. Por el Teorema de Fisher-Tippet y los resultados del dominio de atracción al máximo, sabemos que si se cumple la condición anterior entonces $F \in DAM(H)$.

La clase de distribuciones F para la que se cumple la condición (1.17) es grande, y se pueden encontrar algunas condiciones equivalentes. Una de estas condiciones se relaciona con las funciones de distribución F que están en el dominio de atracción al máximo de la distribución Fréchet (H_ξ donde $\xi > 0$). Esta condición resulta importante si se considera que la Fréchet es la única distribución de valores extremos con colas pesadas, y los datos con que se trabaja en la teoría de valores extremos generalmente provienen de distribuciones de este tipo.

Gnedenko y Kolmogorov (1954) mostraron que para $\xi > 0$, $F \in \text{DAM}(H_\xi)$ si y sólo si $1 - F(x) = x^{-1/\xi}L(x)$, donde $L(x)$ es una función de variación lenta. Este resultado dice esencialmente que si la cola de la función de distribución F decae como una función potencia, entonces la distribución se encuentra en el dominio de atracción al máximo de la Fréchet. La clase de distribuciones en las que esto sucede es muy grande, e incluye a la Pareto, Burr, Log-gamma, t y Cauchy, así como a varios modelos mezclados. A las distribuciones que se encuentran en esta clase se les llama *distribuciones de colas pesadas*.

Las distribuciones en el dominio de atracción al máximo de la Gumbel (H_0) incluyen a la Normal, Exponencial, Gamma y Log-normal. A estas distribuciones se les conoce como *de colas medianas* y encuentran una gran variedad de aplicaciones en campos como el de seguros.

Las distribuciones en el dominio de atracción al máximo de la Weibull (H_ξ para $\xi < 0$) son distribuciones *de colas pequeñas*, como la uniforme y la beta. Esta clase es la de menor interés en la teoría de valores extremos.

Denotemos ahora

$$U(t) = F^{-1}(1 - t^{-1}) \tag{1.18}$$

donde F^{-1} es la función cuantil.

El siguiente teorema es uno de los resultados fundamentales de la teoría de valores extremos.

Teorema 1.4. *Caracterización del dominio de atracción al máximo.*
Para $\xi \in \mathbb{R}$, las siguientes aseveraciones son equivalentes:
a) $F \in DAM(H_\xi)$

b) Existe una función positiva y medible $a(\cdot)$ tal que, para $1+\xi x > 0$,

$$\lim_{u\to x_F} \frac{\overline{F}(u+xa(u))}{\overline{F}(u)} = \begin{cases} (1+\xi x)^{-1/\xi} & \text{si } \xi \neq 0 \\ e^{-x} & \text{si } \xi = 0 \end{cases} \tag{1.19}$$

c) Para $x, y > 0, y \neq 1$,

$$\lim_{s\to\infty} \frac{U(sx)-U(s)}{U(sy)-U(s)} = \begin{cases} \frac{x^\xi - 1}{y^\xi - 1} & \text{si } \xi \neq 0 \\ \\ \frac{\text{In}(x)}{\text{In}(y)} & \text{si } \xi = 0 \end{cases} \tag{1.20}$$

El teorema anterior resume la información esencial relativa al dominio de atracción al máximo, y por lo tanto constituye la base de muchas de las técnicas estadísticas para eventos extremos. En particular, la condición (1.19) tiene una interpretación probabilística importante dentro de los objetivos de este trabajo.
Sea X una v.a. con f.d. $F \in DAM(H_\xi)$; desarrollando el lado izquierdo de la igualdad (1.19) se tiene

$$\begin{aligned} \lim_{u\to x_F} \frac{\overline{F}(u+xa(u))}{\overline{F}(u)} &= \lim_{u\to x_F} \frac{\mathbb{P}(X > u + xa(u))}{\mathbb{P}(X > u)} \\ &= \lim_{u\to x_F} \mathbb{P}(X > u + Xa(u) | X > u) \\ &= \lim_{u\to x_F} \mathbb{P}\Big(\frac{X-u}{a(u)} > x | X > u\Big), \end{aligned}$$

y por lo tanto, de (1.19) se tiene que

$$\lim_{u\to x_F} \mathbb{P}\Big(\frac{X-u}{a(u)} > x | X > u\Big) = \begin{cases} (1+\xi x)^{-1/\xi} & \text{si } \xi \neq 0 \\ e^{-x} & \text{si } \xi = 0. \end{cases} \tag{1.21}$$

La ecuación (1.21) nos da una aproximación de la distribución para excesos sobre un umbral alto y forma parte de un conjunto de resultados equivalentes a los que se han mostrado para máximos, pero que describen el comportamiento de grandes observaciones que exceden un umbral alto. Bajo este enfoque, la pregunta a responder es: dado que una observación es extrema, ¿qué tan grande puede ser?

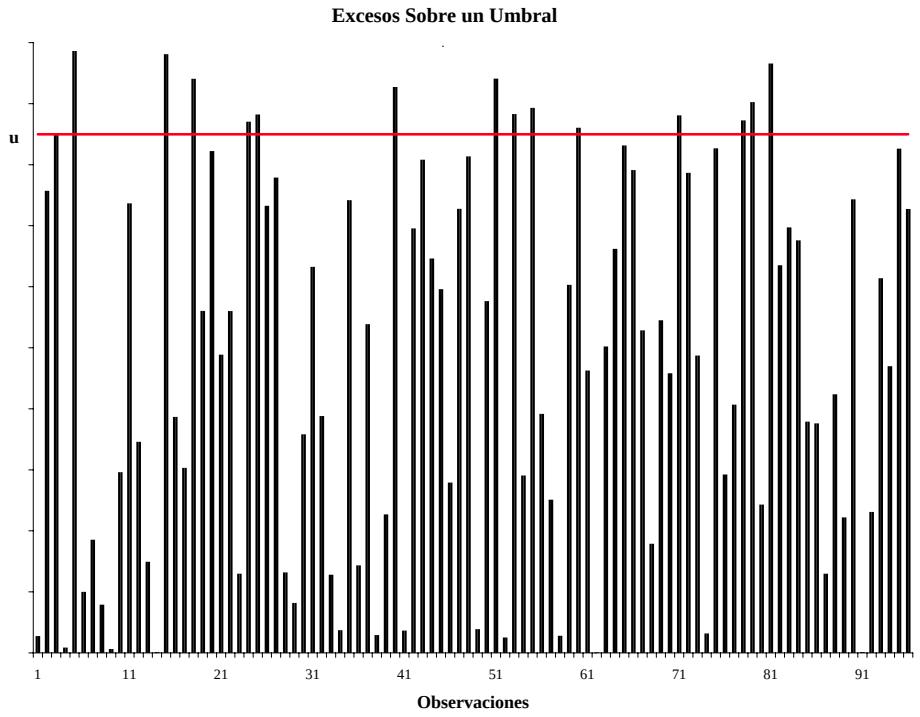
En la siguiente sección se mostrará cuál es la función de distribución que se ajusta mejor a este tipo de datos.

§1.3. Modelo de picos sobre el umbral

Como se puede observar a partir de lo expuesto en la sección anterior, considerar únicamente el máximo de un bloque implica perder información muy valiosa de valores extremos si se cuenta con una mayor cantidad de datos extremos disponibles, sobre todo si un mismo bloque contiene más datos extremos que otro.

Un método alternativo de modelar los datos extremos consiste en analizar el comportamiento de los datos que exceden un cierto umbral, sin tomar en cuenta si están o no dentro de un mismo bloque.

Sea $X_1, X_2, \ldots$ una sucesión de v.a.i.i.d.'s, con función de distribución F. Consideraremos entonces como eventos extremos a aquellas X_i que exceden el umbral u.



La descripción del comportamiento estocástico de estos eventos extremos está dada por la probabilidad condicional

$$\mathbb{P}(X > u + x | X > u) = \frac{1 - F(x+u)}{1 - F(u)}. \tag{1.22}$$

Si la función de distribución F fuera conocida, la distribución de los excesos sobre el umbral también lo sería. Sin embargo, como en el caso de la fundón de distribución para máximos de bloques, la distribución F es desconocida, por lo que se debe buscar una aproximación para la distribución de valores que exceden un umbral, de manera semejante a lo que se hizo para el caso de máximos.

§1.3.1. La distribución Pareto generalizada.

El resultado más importante respecto a la distribución de los excesos sobre un umbral lo da el teorema Pickands-Balkema-de Haan, que consiste en la caracterización asintótica del modelo de este tipo de datos extremos.

Teorema Pickands-Balkema-de Haan

Definimos la función de distribución de los excesos sobre el umbral u como

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}$$

para $0 \leq x < x_F - u$.

Este teorema (Balkema & de Haan (1974), Pickands (1975)), muestra que cuando las condiciones del dominio de atracción al máximo se cumplen, es decir, cuando la sucesión de máximos normalizados converge en distribución a una de las distribuciones de valores extremos, la distribución Pareto generalizada es la distribución límite para la distribución de los excesos, conforme el umbral u tiende al punto final a la derecha x_F. Esto significa que podemos encontrar una función positiva y medible $\sigma(u)$ tal que

$$\lim_{u \to x_F} \sup_{0 \leq x_F - u} |F_u(x) - G_{\xi, \sigma(u)}(x)| = 0,$$

sí y sólo si $F \in DAM(H_\xi)$.

Este teorema sugiere que, para umbrales u suficientemente grandes, la función de distribución de los excesos puede aproximarse por $G_{\xi,\sigma}(x)$ para algunos valores de ξ y σ. De forma equivalente, para $x - u \geq 0$, la función de distribución de los excesos más u, puede aproximarse por $G_{\xi,\sigma}(x-u) = G_{\xi,u,\sigma}(x)$.

La relevancia estadística de este resultado es que nos lleva a ajustar la distribución Pareto generalizada a los datos que exceden umbrales altos. Asimismo, este teorema nos da las bases teóricas para esperar que, si elegimos un umbral suficientemente alto, los datos que lo sobrepasen tendrán un comportamiento Pareto generalizado. Sin embargo, la principal dificultad práctica se encuentra en elegir el umbral adecuado, ya que la teoría no nos brinda ayuda alguna al respecto, por lo que el analista de los datos debe tomar la decisión con base en su experiencia.

A continuación revisaremos algunos de los aspectos más importantes de la distribución Pareto generalizada y la conexión que existe con los resultados para el máximo.

Definición 1.6. La distribución Pareto generalizada.

Definimos a $G_{\xi,\beta}$ como

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi\frac{x}{\beta})^{-1/\xi} & \text{si} \quad \xi \neq 0, \\ \\ 1 - \exp^{-x/\beta} & \text{si} \quad \xi = 0, \end{cases} \quad ; x \in \mathcal{D}(\xi, \beta)$$

donde

$$\mathcal{D}(\xi, \beta) = \begin{cases} [0, \infty) & \text{si} \quad \xi \geq 0, \\ [0, \frac{-\beta}{\xi}] & \text{si} \quad \xi < 0. \end{cases}$$

Se puede extender esta familia agregando un parámetro de localización μ. Así, la distribución Pareto generalizada $G_{\xi,\mu,\beta}$ se define como $G_{\xi,\beta}(x-\mu)$.

Una relación importante entre la distribución Pareto generalizada y la distribución de valores extremos se puede notar a partir de la siguiente igualdad

$$\begin{aligned}\overline{G}_{\xi,\beta}(x) &= 1 - G_{\xi,\beta} \\ &= \begin{cases} (1+\xi\frac{x}{\beta})^{-1/\xi} & \text{si} \quad \xi \neq 0, \\ \\ \exp^{x/\beta} & \text{si} \quad \xi = 0, \end{cases} \quad ; x \in \mathcal{D}(\xi,\beta) \\ &= \lim_{u \to x_F} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)}\end{aligned}$$

donde

$$\mathcal{D}(\xi,\beta) = \begin{cases} [0,\infty) & \text{si} \quad \xi \geq 0, \\ [0, \frac{-\beta}{\xi}] & \text{si} \quad \xi < 0. \end{cases}$$

es decir, la distribución límite que se encontró en (1.19) al trabajar con máximos de valores extremos.

El teorema anterior implica que si los máximos de bloques tienen distribución H, entonces los excesos sobre un umbral tienen su distribución correspondiente dentro de la familia Pareto generalizada. Más aún, los parámetros de la distribución Pareto generalizada de excesos sobre un umbral están determinados de manera única por aquéllos de la distribución de valores extremos generalizada de los máximos de bloques.
Las semejanzas entre las dos distribuciones se reflejan también en el parámetro de forma ξ, pues al igual que en la distribución de valores extremos, es dominante al determinar el comportamiento cualitativo de la distribución Pareto generalizada. Si $\xi > 0$ se tiene una versión reparametrizada de la Pareto usual, sin límite superior; si $\xi < 0$ se llega a una distribución tipo Pareto II, que está acotada por la derecha; y $\xi = 0$ representa una distribución no acotada, la exponencial.

Finalmente, es importante señalar que la propiedad de modelar los excesos sobre un umbral alto es sólo una de varias propiedades de la distribución Pareto generalizada relativas a observaciones que exceden un umbral alto. Otros resultados interesantes son los siguientes:

- El número de excesos sobre un umbral alto sigue un proceso Poisson.
- Se puede encontrar el valor del umbral más adecuado graficando la función media de exceso, como se verá más adelante.

- La distribución del máximo de un número Poisson de excesos sobre un umbral alto es la distribución de valores extremos.

Con lo que se ha revisado hasta ahora, se distingue claramente la aplicación más importante que se da a las distribuciones anteriores:

- La distribución de valores extremos H_ξ describe las distribuciones límite de máximos normalizados.

- Por su parte, la distribución Pareto generalizada $G_{\xi,\beta}$ es la distribución límite de los excesos sobre un umbral alto.

§1.4. Análisis exploratorio de datos extremos

Uno de los análisis de datos más importantes es el exploratorio, es decir, aquél en el que se observa y analiza los datos tratando de entender qué nos dicen, antes de comenzar un análisis estadístico más específico. Con este propósito, el uso de diferentes tipos de gráficas resulta de gran utilidad, como se verá a continuación.

§1.4.1. Gráficas de cuantiles (QQ-Plots)

Las gráficas de cuantiles son procedimientos gráficos de bondad de ajuste desarrollados tras observar que los cuantiles $Q(p)$ de las distribuciones, se encuentran relacionados de manera lineal con los cuantiles correspondientes de un conjunto de datos que provienen de dichas distribuciones. Como la linealidad en una gráfica puede ser verificada fácilmente a simple vista o cuantificada por medio de un coeficiente de correlación, esta herramienta se utiliza para responder a la siguiente pregunta de bondad de ajuste:

¿Cuál es el modelo que mejor se ajusta a los datos $X_1, \ldots, X_n$ *i.i.d.?*

La distribución normal es la primera clase de modelos para los cuales las gráficas de cuantiles constituyeron una herramienta poderosa para responder esta pregunta. Sin embargo, este procedimiento ha adquirido mayor importancia y su uso se ha extendido rápidamente en la búsqueda de otro tipo de distribuciones.

Supongamos que $F(.;\theta)$ es un modelo paramétrico que se ajustará a $X_1, \ldots, X_n$, resultando un estimador $\hat{\theta}$ y de ahí un modelo ajustado $\hat{F} = F(.;\hat{\theta})$. Definamos la muestra ordenada $X_{1,n} \leqslant \cdots \leqslant X_{n,n}$. La gráfica de los puntos

$$\left\{\left(X_{k,n}, \hat{F}^{-1}(p_{k,n})\right) : k = 1, \ldots, n\right\} \tag{1.23}$$

para una sucesión apropiada de puntos $(p_{k,n})$ es la gráfica de las estadísticas de orden de los datos $X_1, \ldots, X_n$ cuando se ajusta a la familia paramétrica $F(.;\theta)$. Generalmente se toma $p_{k,n} = k/(n+1)$.

Al realizar estas gráficas debe considerarse que generalmente los datos se contrastan contra una familia con parámetros de localización y escala $F((\cdot - \mu)/\psi)$, donde en muchos casos μ y ψ representan a la media y la desviación estándar de X. Una gráfica de cuantiles para esta F seguirá siendo lineal, pero con ordenada al origen en μ y pendiente ψ, por lo que estos parámetros pueden ser estimados usando regresión lineal. Sin embargo, en el caso de las distribuciones de valores extremos, como se vio en las secciones anteriores, además de los parámetros de localización y escala existe un parámetro de forma $\xi \in \mathbb{R}$, lo que hace más delicada la interpretación de estas gráficas. Una forma de afrontar este problema es obteniendo un estimador $\hat{\xi}$ para ξ y después hacer la gráfica de cuantiles para la distribución de valores extremos $H_{\hat{\xi};0,1}$, donde μ y ψ pueden ser estimadas ya sea por inspección visual o mediante una regresión lineal.

Un buen ajuste produce la gráfica de una línea recta, mientras que una desviación del modelo (como la presencia de colas pesadas, sesgo, *outliers*, etc.) pueden ser diagnosticados fácilmente. En resumen, las cualidades más importantes de las gráficas de cuantiles según Embrechts et al. (1997) se deben a las siguientes propiedades:

- Comparación de distribuciones: Si los datos provienen de una muestra aleatoria de la distribución de referencia o de una transformación lineal de la misma, la gráfica de cuantiles debe ser casi lineal.

- *Outliers*: Pueden ser identificados fácilmente en la gráfica cuando uno o algunos de los datos presentan un error grande o son visiblemente

diferentes del resto de los datos, si éstos se distribuyen igual que la distribución de referencia.

- Localización y escala: Estos parámetros pueden estimarse gráficamente en una muestra de datos mediante la ordenada al origen y la pendiente, bajo el supuesto de que los datos provienen de la distribución de referencia.
- Forma: Algunas diferencias en la forma de la distribución se pueden deducir de la gráfica. En el caso de colas pesadas, por ejemplo, se observarán curvaturas en la parte inferior izquierda y/o superior derecha de la gráfica.

Finalmente, podemos decir que para aceptar un modelo propuesto como un modelo posible de la población es necesario:

1. Comenzar por una caracterización de la relación lineal entre una función creciente de los cuantiles teóricos $Q(p)$ de la distribución propuesta y los cuantiles calculados de una distribución específica;
2. Reemplazar los cuantiles teóricos $Q(p)$ por los cuantiles empíricos correspondientes $\hat{Q}_n(p)$;
3. Graficar la función creciente de los cuantiles empíricos contra el cuantil específico correspondiente.

§1.4.2. Función media de excesos muestral

Como se mencionó anteriormente, la principal dificultad práctica al momento de ajustar una distribución Pareto generalizada a la muestra que excede un umbral, se encuentra precisamente en elegir el umbral adecuado. La gráfica de la función media de excesos muestral es una herramienta muy útil al momento de elegir el tamaño de dicho umbral, además de que permite observar aspectos importantes del comportamiento de los datos en la cola de la distribución.

Definimos a la función media de excesos muestral como

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)^+}{\sum_{i=1}^{n} 1_{\{X_i > u\}}}$$

donde

$$(X_i - u)^+ = \begin{cases} X_i - u, & X_i - u > 0 \\ \\ 0, & X_i - u \leq 0 \end{cases}$$

es decir, la función media de excesos muestral es igual a la suma de los excesos sobre el umbral u, dividida entre el número de datos que exceden el umbral u.

La función media de excesos muestral $e_n(u)$ es un estimador empírico de la función media de excesos, que se define como $e(u) = \mathbb{E}[X - u | X > u]$. Esta función describe la cantidad en que se espera sobrepasar un umbral, dado que el exceso ocurre.

En el gráfico se contrasta el valor del umbral u contra la función media de exceso muestral

$$\{(u, e_n(u)), X_{1,n} < u < X_{n,n}\}$$

donde $X_{1,n}$ y $X_{n,n}$ son la primera y la n-ésima estadísticas de orden y $e_n(u)$ es la función media de excesos muestral.

Una de las propiedades más importantes de las gráficas $\{(u, e_n(u)) : u \geq 0\}$, es que permiten distinguir fácilmente a las distribuciones de colas pesadas de las que no lo son. Una función media de excesos con pendiente positiva significa que la distribución tiene colas pesadas; en particular, una línea recta con pendiente positiva sobre un umbral dado es signo de un comportamiento tipo Pareto en la cola de la distribución. Una tendencia negativa muestra que la distribución es de colas ligeras, mientras que una línea con pendiente cero es signo de una cola exponencial.

Cuando consideramos la forma de la función media de excesos la distribución exponencial resulta particularmente importante, pues al no tener memoria no importa si $X > a$, el resultado para $\mathbb{E}(X - u)$ es el mismo que se obtendría si $a = 0$ y se calculara $\mathbb{E}(X)$, de donde se verifica que la función media de excesos es constante con el mismo valor de la esperanza de una

exponencial

$$e(u) = \frac{1}{\lambda} \quad \text{para toda } u > 0. \tag{1.24}$$

Para ilustrar lo anterior, en el Cuadro 1.1 se presentan las funciones medias de exceso teóricas de algunas de las distribuciones más importantes.

Cuadro 1.1: *Función media de exceso para algunas distribuciones*

Función	**Cola** $\overline{F} = 1 - F(c)$	$e(u)$
Exponencial	$\exp(-\lambda c)$	$1/\lambda$
Log-normal	$\int_c^\infty \frac{1}{\sqrt{2\pi}\sigma x} exp\Big(- \frac{1}{2\sigma^2}(\log(x) - \mu)^2\Big) dx$	$\frac{a}{\log(a)}(1 + o(1))$
Pareto	$\mathrm{c}^{-\alpha}$	$\frac{a}{\alpha - 1} \quad (\alpha > 1)$
Uniforme	$1 - c$	$\frac{1}{2}(1 - a) \quad (a < 1)$
Weibull	$\exp(-\lambda c^\tau)$	$\frac{a^{1-\tau}}{\lambda\tau}(1 + o(1))$

Capítulo 2

Inferencia bayesiana

Los resultados científicos o experimentales generalmente consisten en datos de la forma $x = \{x_1, \ldots, x_n\}$, donde las x_i son observaciones "homogeneas". Los métodos estadísticos se utilizan para llegar a conclusiones sobre la naturaleza y el comportamiento futuro del proceso que ha generado las observaciones.

Un elemento central de todo análisis estadístico es la especificación de un modelo probabilístico que, se supone, describe el mecanismo que genera los datos como función de un parámetro θ (posiblemente multidimensional), y sobre este valor se tiene, en el mejor de los casos, información limitada. Es obvio, entonces, que todas las conclusiones estadísticas que se obtienen están condicionadas por el modelo probabilístico y por el valor de el o los parámetros que se hayan especificado.

En la inferencia estadística se estima el valor del parámetro poblacional θ a partir de los datos observados x, por lo que se puede inferir que los valores de θ que dan una probabilidad alta a los valores observados x son más deseables que aquéllos que asignan una probabilidad baja. Esto es lo que se conoce como el principio de máxima verosimilitud.

A diferencia de este enfoque clásico, el razonamiento bayesiano está basado fundamentalmente en la teoría de probabilidad. No es casualidad que algunos de los libros más importantes de la estadística bayesiana, como los de Laplace (1812), de Finetti (1970), o Jeffreys (1939), se titulen "Teoría de la probabilidad". La consecuencia más importante de esta fundamentación es el hecho de que toda la incertidumbre presente en el problema se debe

expresar a través de distribuciones de probabilidad.

El contexto en el que se desarrolla la inferencia bayesiana es el siguiente: existe un parámetro poblacional θ sobre el que queremos hacer inferencia, y un modelo probabilístico $p(x|\theta)$ que determina la probabilidad de observar diferentes valores de x, según el valor del parámetro θ. La principal diferencia con la inferencia estadística clásica es que, en el enfoque bayesiano, el valor de θ, por ser desconocido, es tratado como una variable aleatoria.

En esencia, el interés se centra en $p(\theta|x)$, conocida como *distribución final*, en lugar de $p(x|\theta)$, la distribución de los datos dado el parámetro. Para encontrar dicha distribución, los parámetros desconocidos y los modelos probabilísticos deben tener una distribución de probabilidad conjunta que describa toda la información disponible sobre su valor. Este principio es visto como uno de los elementos más importantes del enfoque bayesiano, y en muchos sentidos nos lleva a inferencias más naturales, pero para ello es necesario especificar una *distribución inicial* $p(\theta)$ que representa las ideas o la información que se tiene sobre θ antes de tener alguna información de los datos.

§2.1. Características del método bayesiano

Según O'Hagan (1994), podemos indentificar cuatro aspectos fundamentales que caracterizan el método bayesiano de inferencia estadística:

- Información inicial. Todos los problemas son únicos y tienen su propio contexto. De este contexto se obtiene la información inicial, y es la formulación y explotación de ese conocimiento inicial que mantiene a la inferencia bayesiana aparte de la estadística clásica.

- Probabilidad subjetiva. A diferencia del enfoque frecuentista de la estadística clásica, la estadística bayesiana hace explícita y formaliza la noción de que la probabilidad es subjetiva, dependiendo del conocimiento o expectativas de un individuo. Es por esto que se dice que el análisis bayesiano es personalista, ya que es único bajo las especificaciones de las ideas y el conocimiento inicial de cada individuo. Por otra parte, la inferencia se basa en la distribución final $p(\theta|x)$, cuya forma

depende de las especificaciones particulares de la distribución inicial $p(\theta)$.

- Consistencia. Al tratar al parámetro θ como aleatorio, el desarrollo de la inferencia bayesiana surge naturalmente únicamente de la teoría probabilística, lo que significa que cualquier inferencia puede ser planteada en términos de probabilidad sobre θ.
- No uso de métodos *ad hoc*. Como en la estadística clásica no se estima probabilidad alguna para los diferentes valores que puede tomar θ, se han desarrollado varios criterios para evaluar qué tan bueno es un estimador en uno u otro sentido. La inferencia bayesiana deja de lado esta tendencia de juzgar y comparar estimadores al dejar que la distribución posterior exprese en términos probabilísticos toda la inferencia sobre el parámetro desconocido θ.

§2.2. El paradigma bayesiano

El análisis estadístico de los datos observados generalmente comienza con una evaluación informal descriptiva, que se utiliza para proponer un modelo probabilístico formal $\{p(x|\theta), \theta \in \Theta\}$ que se supone representa, para algún valor (desconocido) de θ, el mecanismo probabilístico que ha generado los datos observados. En la inferencia bayesiana se considera una necesidad lógica el asignar una probabilidad inicial $p(\theta)$ sobre el espacio de parámetros Θ, que describa el conocimiento disponible sobre θ antes de que los datos hayan sido observados. Con esta información, y a partir de cálculos estándar de la teoría de probabilidad, si el modelo probabilístico es correcto, toda la información disponible sobre el valor de θ una vez que los datos han sido observados está contenida en la distribución final, que se obtiene de manera inmediata a partir del teorema de Bayes,

$$p(\theta|x, A) = \frac{p(x|\theta)p(\theta|A)}{\int p(x|\theta)p(\theta|A)d\theta}, \tag{2.1}$$

donde A representa la hipótesis hecha sobre el modelo de probabilidad, y para simplificar la representación suprimimos la referencia a la hipótesis aceptada.

Este uso sistemático del teorema de Bayes para incorporar la información proporcionada por los datos es lo que justifica el adjetivo *bayesiana* por el cual es conocido el paradigma.

§2.3. Distribución inicial

Como se ha mencionado a lo largo de este capítulo, uno de los elementos más importantes del paradigma bayesiano es la especificación de una distribución inicial que capture la información inicial que se tiene sobre todos los parámetros desconocidos, de manera que dicha información pueda ser incorporada al modelo junto con los datos obtenidos en la muestra. Esta información puede venir, por ejemplo, de la opinión de un experto o de experimentos anteriores de naturaleza similar.

La distribución inicial describe lo que sabemos acerca del parámetro θ y no cómo varía, pues θ es un parámero fijo pero desconocido.

§2.3.1. Análisis de referencia

Un caso particularmente importante surge cuando no se quiere incorporar información sobre el parámetro θ, es decir, se busca que las inferencias sólo dependan de los datos observados y del modelo $p(x|\theta)$ supuesto. En estos casos se utiliza el análisis de referencia, que se basa en conceptos de teoría de la información para determinar distribuciones iniciales llamadas *de referencia*, de manera que la inferencia sobre los parámetros se base únicamente en el modelo supuesto y en los datos observados.
El problema de caracterizar a una distribución inicial de este tipo es mucho más complejo de lo que intuitivamente se puede pensar, ya que en el modelo bayesiano los datos no pueden hablar completamente por sí mismos, pues cualquier especificación inicial tiene alguna consecuencia en la distribución final.

La distribución inicial de referencia para θ, denotada por $\pi(\theta)$, se define como aquella distribución que maximiza la información funcional faltante.

Dados los datos x, la distribución final de referencia se deriva simplemente

del teorema de Bayes como

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta)$$

La regla de Jeffreys

La regla de Jeffreys es el procedimiento más comúnmente utilizado, en el caso uniparametral, para obtener distribuciones iniciales que representen la falta de información que se tiene respecto al parámetro de interés. Una cualidad importante de esta regla es que es invariante bajo transformaciones uno a uno del parámetro, lo que significa que la misma falta de conocimiento que se tiene respecto al valor que puede tomar θ se tiene sobre el valor que puede tomar cualquier transformación uno a uno sobre θ.

Con base en estas consideraciones de invarianza, y en estrecha relación con un resultado muy importante del análisis de referencia (en el cual se demuestra que si la distribución posterior asintótica de θ, bajo ciertas condiciones de regularidad, es Normal con cierta precisión, entonces la distribución inicial de referencia tiene la forma $\pi(\theta) \propto \{h(\theta)\}^{1/2}$), Jeffreys (1946) usa como distribución inicial a la densidad $\pi(\theta)$ (generalmente impropia, es decir, el área bajo la densidad inicial no es la unidad)

$$\pi(\theta) \propto h(\theta)^{1/2}$$

donde

$$h(\theta) = \int p(x|\theta)\Big(- \frac{\partial^2}{\partial\theta^2} \log\ p(x|\theta)\Big)dx,$$

Esta densidad se deriva del hecho de que Jeffreys notó que la divergencia logarítmica se comporta localmente como el cuadrado de una distancia, determinada por una métrica de Riemman, cuya medida natural de distancia es $h(\theta)^{1/2}$.

En el caso multiparamétrico, la generalización de la regla de Jeffreys es

$$\pi(\theta) \propto |H(\theta)|^{1/2},$$

donde

$$H(\theta)_{ij} = - \int p(x|\theta) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\ p(x|\theta)dx,$$

es decir, H es la matriz de información de Fisher.

En su trabajo, Jeffreys exploró las implicaciones de la distribución inicial propuesta para un gran número de problemas de inferencia, y encontró que su regla (restringida por definición a un parámetro continuo) funciona bien en el caso real, pero puede llevar a resultados poco deseables cuando se trata de extender a casos multiparamétricos, en los que además se dificulta considerablemente el cálculo.

En general, si θ es un parámetro de localización, $\pi(\theta) \propto 1$, y si θ es un parámetro de escala, entonces $\pi(\theta) \propto \theta^{-1}$.

Cabe señalar que frecuentemente las distribuciones iniciales así obtenidas son impropias. Como ejemplo tenemos a $\pi(\mu) = 1$ y $\pi(\sigma) = 1/\sigma$, cuyas integrales son infinitas. La mayoría de las veces las distribuciones iniciales impropias pueden usarse sin problemas en el análisis bayesiano; sin embargo, en algunos modelos el uso de distribuciones iniciales impropias pueden llevar a distribuciones finales impropias. En la selección de modelos, que es el caso que nos interesa, el uso de distribuciones iniciales impropias dificulta el proceso de selección.

Distribuciones iniciales "vagas"

Las distribuciones iniciales vagas son esencialmente densidades con una gran dispersión (como una distribución Normal con varianza muy grande), con lo que se obtiene un valor inicial de los parámetros muy parecido sobre un intervalo muy grande. Por ejemplo, la distribución inicial $\pi(\sigma) = 1/\sigma$ puede aproximarse por una densidad Gamma con parámetros de forma y escala muy pequeños.

§2.4. Distribución predictiva

Una vez que se obtuvieron algunos datos sobre un fenómeno de interés, generalmente se quiere conocer más sobre la población de donde provienen. Tradicionalmente, esto se hace suponiendo que dicha población sigue cierta distribución y se estiman los parámetros de esa distribución.

La forma de operar del teorema de Bayes nos permite obtener de forma

natural la distribución predictiva para nuevos datos no observados y a partir de los datos disponibles x, al pasar de la distribución

$$p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta$$

a la distribución

$$p(x_{n+1}, \ldots, x_{n+m}|x_1, \ldots, x_n) = \int p(x_{n+1}, \ldots, x_{n+m}|\theta)p(\theta|x_1, \ldots, x_n)d\theta$$

a través de

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}.$$

Si usamos $y = \{y_1, \ldots, y_m\} = \{x_{n+1}, \ldots, x_{n+m}\}$ para denotar observaciones futuras o no observadas, y $x = \{x_1, \ldots, x_n\}$ a las cantidades observadas, la relación anterior puede expresarse en términos más simples como

$$\begin{aligned} p(x) &= \int p(x|\theta)p(\theta)d\theta; \\ p(y|x) &= \int p(y|\theta)p(\theta|x)d\theta; \\ p(\theta|x) &= p(x|\theta)p(\theta)/p(x). \end{aligned}$$

Capítulo 3

Selección de modelos

Seleccionar el modelo que explique de mejor manera un fenómeno determinado es uno de los problemas fundamentales de la ciencia, particularmente en la rama estadística, ya que el conocimiento que se tiene respecto a casi cualquier fenómeno de la naturaleza es limitado y, por lo tanto, la traducción de dicho fenómeno en un modelo estadístico es imperfecto. Es por esta razón que una de las etapas más importantes del análisis estadístico corresponde a la selección del modelo que represente de mejor manera el fenómeno bajo estudio, de manera que se pueda hacer inferencia, e incluso predicción, sobre dicho fenómeno.

La determinación de un modelo tiene dos vertientes: la selección de modelos y la adecuación o ajuste de un modelo. En el primer caso se parte de un conjunto de modelos posibles y se pretende elegir aquél que represente de mejor manera un cierto proceso; en el segundo caso se parte de un modelo determinado y el interés se centra en la capacidad de ajuste de dicho modelo a los datos disponibles. En la mayoría de los casos, la determinación de un modelo deberá apoyarse en las dos vertientes anteriores: por una parte, existen razones importantes para considerar un rango de modelos posibles, ya sea desde la perspectiva de un individuo o de un grupo de individuos, lo que nos lleva naturalmente a un problema de selección de modelos; por otra parte, una vez que ha sido seleccionado el modelo, es necesario evaluar su comportamiento y capacidad predictiva a partir de los datos ya observados, con el objeto de determinar su validez.

Recientemente se ha observado un creciente interés en los métodos baye-

sianos de selección de modelos, lo cual es avalado por la basta cantidad de literatura escrita al respecto y por el reconocimiento de algunas ventajas que surgen de manera natural al usar estos métodos en lugar de los clásicos.

Tradicionalmente, los procesos de selección se han basado en los factores de Bayes (ver por ejemplo Jeffreys, 1939) para los cuales Smith y Spiegehalter (1980) mostraron que, con cierta distribución inicial, proveen una generalización del criterio bayesiano de información (Schwartz, 1978) y del criterio de información de Akaike (1973). Surgieron después los factores de Bayes intrínsecos aritméticos y geométricos (Berger y Pericchi, 1996) y los factores de Bayes fraccionales (O'Hagan, 1995), entre muchos otros. En secciones posteriores se presenta una breve descripción de algunos de estos métodos de selección de modelos, desarrollados en el marco de la teoría de decisiones.

Un enfoque diferente a aquél en el que se busca elegir un modelo único a partir de un conjunto de modelos contendientes es el de combinación de modelos, particularmente el que se refiere a la combinación de modelos predictivos (ver Clemen, 1990). Este enfoque se basa en la premisa de que la precisión de una predicción puede mejorarse considerablemente al combinar las predicciones de varios individuos. En el enfoque bayesiano el problema de mezcla de modelos se conoce como promedio bayesiano de modelos, donde la combinación de modelos se representa como una mezcla de distribuciones finales o predictivas, según sea el caso.

§3.1. Perspectivas de la comparación de modelos

Denotemos por $\{M_i, i \in I\}$ a cada uno de los modelos posibles, donde I es un conjunto índice (finito o numerable), y por $\mathcal{M} = \{M_i : i \in I\}$ a la clase que comprende estos modelos. Bajo el enfoque bayesiano, el modelo M_i está definido como

$$M_i = \{p_i(x|\theta_i), \pi_i(\theta_i)\},$$

donde $p_i(x|\theta_i)$ denota la distribución condicional de la variable aleatoria x dado el parámetro θ_i y el modelo M_i, y π_i es la distribución inicial de los parámetros θ_i condicional al modelo M_i para cada $i \in I$.

En el contexto del problema de decisión relativo a la selección de modelos entre un conjunto de modelos propuestos $\{M_i, i \in I\}$, existen tres diferentes perspectivas en que se puede considerar a un rango de modelos posibles (ver Bernardo y Smith, 1994):

1) $\mathcal{M}$-cerrado

Desde esta perspectiva se supone que existe un modelo dentro de la clase $\mathcal{M}$ que es el modelo verdadero, en el sentido de que representa al proceso que genera los datos observados. Claramente, esta perspectiva está detrás de muchos de los criterios más comunes de selección de modelos, ya que se limita a elegir a uno de los modelos bajo consideración que puede reflejar tanto la incertidumbre de un individuo ante un conjunto de modelos propuestos, como un rango de modelos propuestos por diferentes individuos.

El modelo especifica la distribución de x de la forma

$$p(x) = \sum_{i \in I} p(M_i) p(x|M_i),$$

donde $p(M_i)$ denota las ponderaciones iniciales del conjunto de modelos $\{M_i, i \in I\}$.

Un cuestionamiento lógico a esta perspectiva es qué tan sensato resulta hablar de un modelo "verdadero", considerando la información o conocimiento limitado que generalmente se tiene sobre la naturaleza de un fenómeno. Dado que esta limitante nos obliga a tener que definir modelos que representen una aproximación al verdadero proceso subyacente, es difícil aceptar esta perspectiva más allá de situaciones controladas, por ejemplo, cuando se sabe que la muestra ha sido generada utilizando un conjunto de programas de simulación, o cuando se presenta una aplicación en la que debe reconsiderarse el continuar utilizando un modelo con parámetros específicos o incorporar la incertidumbre sobre el valor apropiado de los mismos.

2) $\mathcal{M}$-abierto

Una perspectiva más realista que la descrita anteriormente supone que ninguno de los modelos de la clase $\mathcal{M}$ es el modelo verdadero, por lo

que $\{M_i, i \in I\}$ es simplemente un conjunto de modelos que sirven como una aproximación al proceso subyacente del fenómeno en estudio.
La ausencia de un modelo supuesto requiere que la comparación se realice sobre una serie de modelos (como pueden ser los modelos de regresión con diferentes elecciones de regresores, modelos lineales con diferentes selecciones de covarianzas, etc.) con base en un criterio de optimización o discrepancia respecto al modelo "verdadero".

3) $\mathcal{M}$-completo

Esta perspectiva se adopta generalmente cuando la implementación del modelo que se considera más realista resulta difícil o muy costoso, por lo que el conjunto de modelos $\{M_i, i \in I\}$ es considerado una aproximación al modelo más adecuado, constituido por un conjunto de modelos con los que se cuenta para ser comparados.
Así, los modelos alternativos son contemplados como una aproximación al modelo real M_t. Sin embargo, los modelos alternativos deberán ser evaluados y comparados a la luz de lo que realmente se cree, de manera que el modelo predictivo para x se puede representar como

$$p(x) = p_t(x) = p(x|M_t). \tag{3.1}$$

Generalmente los modelos alternativos que se consideran son aquellos que resultan atractivos desde el punto de vista de la implementación y comunicación de resultados en comparación con el modelo "real" M_t.

§3.2. Selección de modelos como un problema de decisión

El problema de selección de modelos es un problema de decisión, y por lo tanto es necesario incluir una función de utilidad y una distribución inicial en el espacio de los "estados de la naturaleza" relativos al problema. La función de utilidad debe cuantificar las consecuencias de una acción particular en el espacio de decisiones (en este caso, elegir un modelo de un número finito de

ellos), dado un "estado de la naturaleza" particular.

En el enfoque bayesiano se han propuesto varios métodos de selección de modelos que han sido ampliamente aceptados y utilizados en diferentes contextos. El planteamiento del problema en cada uno de estos criterios de selección depende en gran medida de la forma en que se definen los elementos del problema de decisión, en donde se considera implícitamente la perspectiva de comparación de modelos que se haya adoptado, a través del espacio de acciones.

A continuación se presentan algunos de los criterios de selección bajo la perspectiva de un problema de decisión.

§3.2.1. Factores de Bayes

Mediante el uso de los factores de Bayes se busca comparar dos modelos entre sí y elegir uno de ellos, sin alguna decisión subsecuente y restringidos a la perspectiva $\mathcal{M}$-cerrada. El "estado de la naturaleza" de interés se define como el verdadero modelo M_t, por lo que suponiendo una muestra futura $y = (y_1, \ldots, y_m)$, $\Pr(M_t|y) \to 1$ cuando $m \to \infty$. Desde esta perspectiva, el problema consiste en elegir al modelo "verdadero", por lo que una forma natural de la función de utilidad es

$$u(M_i, w) = \begin{cases} 1 & \text{si} \quad w = M_i \\ 0 & \text{si} \quad w \neq M_i, \end{cases} \tag{3.2}$$

donde w es un "estado de la naturaleza"particular.

En estas condiciones, el grado de preferencia entre un rango de modelos puede determinarse a partir de sus probabilidades finales, $P(M_i|x)$, $i \in I$. La decisión óptima es, por lo tanto, elegir el modelo que tenga la probabilidad final más alta mediante el cociente de sus probabilidades finales

$$\frac{P(M_i|x)}{P(M_j|x)} = \frac{p(x|M_i)}{p(x|M_j)} \times \frac{P(M_i)}{P(M_j)}, \tag{3.3}$$

donde

$$p(x|M_i) = \int p_i(x|\theta_i)p_i(\theta_i)d\theta_i.$$

La igualdad (3.3) indica que el cociente de probabilidades finales se obtiene de actualizar el cociente de las probabilidades iniciales a través del cociente de las verosimilitudes integradas, y es precisamente este último cociente al que se conoce como factor de Bayes del modelo M_i con respecto a M_j.

Definición 3.1. Factor de Bayes.

Dadas dos hipótesis H_i, H_j correspondientes a dos modelos alternativos M_i, M_j, con los datos x, el factor de Bayes a favor de H_i (contra H_j) está dado por el cociente

$$B_{ij} = \frac{p(x|M_i)}{p(x|M_j)} = \left\{\frac{P(M_i|x)}{P(M_j|x)}\right\} \Big/ \left\{\frac{P(M_i)}{P(M_j)}\right\}. \tag{3.4}$$

Intuitivamente, el factor de Bayes cuantifica la evidencia que proveen los datos en favor de alguno de los modelos. Por lo tanto, si $B_{ij} > 1$ significa que H_i es más creíble que H_j a la luz de x, mientras que $B_{ij} < 1$ significa que H_j es más creíble que H_i.

Good (1950) sugirió el uso del logaritmo de los cocientes anteriores, a los que nombró *pesos de evidencia*, de manera que $\log B_{ij}$ corresponde al peso de la evidencia de la verosimilitud integrada en favor de M_i. Por otra parte, el uso de ciertas distribuciones iniciales en los factores de Bayes puede derivar, por ejemplo, en el criterio de información de Akaike o en el criterio bayesiano de información.

Al emplear los factores de Bayes como criterio de selección de modelos se deben considerar algunos aspectos importantes. Gelfand y Ghosh (1998) señalan que desde la perspectiva de teoría de decisiones, los factores de Bayes únicamente son válidos cuando se trabaja con una función de utilidad 0-1, sin embargo, en la práctica se prefiere usar funciones cuantitativas que cualitativas. Por otra parte, Gelfand y Ghosh (1998) indican que las distribuciones predictivas son comparables entre modelos y las distribuciones finales no lo

son, por lo que, en la mayoría de las situaciones, el enfoque predictivo es el más adecuado.

Otro problema común, y que ha atraído la atención de muchos bayesianos, tiene que ver con la elección de distribuciones iniciales impropias (ver Spiegelhalter y Smith, 1982; Aitkin, 1991; Berger y Pericchi, 1996; O'Hagan, 1995). El problema surge cuando, debido a la falta de información inicial sobre los parámetros o a la complejidad del problema en cuestión, resulta difícil asignar una distribución inicial. En estos casos generalmente se recurre a distribuciones iniciales no informativas que muchas veces llevan a distribuciones finales impropias, y como consecuencia el factor de Bayes resulta indeterminado.

Algunas variantes del factor de Bayes han sido propuestas para corregir parcialmente el problema de indeterminación, entre las que se encuentra el factor de Bayes parcial (Aitkin, 1991).

§3.3. Selección de modelos predictivos

En muchas situaciones la elección de un modelo tiene como principal objetivo predecir valores futuros de una variable aleatoria de interés. En estos casos el criterio de selección debe dar prioridad a la capacidad predictiva de cada modelo de acuerdo con algún criterio de optimalidad.

La comparación entre modelos sugiere el uso de distribuciones predictivas incluso cuando el fin último no sea el de predicción, pues a diferencia de las distribuciones finales, las distribuciones predictivas son comparables entre modelos y permiten considerar implícitamente el uso que se dará al modelo seleccionado, además de que la selección de un modelo con buena capacidad predictiva es una forma de garantizar que dicho modelo representa el proceso subyacente que genera los datos. En este sentido, Box (1980) señala que las distribuciones predictivas y las distribuciones finales tienen roles complementarios en el análisis de datos: las distribuciones finales proveen una base para la estimación de parámetros condicional al ajuste del modelo considerado, mientras que las distribuciones predictivas permiten verificar la validez del modelo considerado a la luz de los datos disponibles.

El uso de la aproximación predictiva para seleccionar modelos ha sido

propuesta por Geisser y Eddy (1979), San Martini y Spezzaferri (1984), Gelfand et al. (1992), Gelfand (1995), Gelfand y Ghosh (1998), Laud e Ibrahim (1995), entre otros. San Martini y Spezzaferri (1984) propusieron un criterio de selección que se basa en el comportamiento de las distribuciones predictivas finales (que deben ser propias, sin importar si las distribuciones iniciales lo son) y en la probabilidad final de cada uno de los modelos, que puede no estar bien definida si alguna de las distribuciones iniciales son impropias. Asimismo, uno de los supuestos implícitos es que el verdadero modelo está contenido dentro del conjunto de modelos bajo consideración, es decir, se adopta la perspectiva del $\mathcal{M}$-cerrado.

Gutiérrez-Peña y Walker (2001) plantean un criterio de selección de modelos para el caso de v.a.i.i.d.'s en el cual las perspectivas $\mathcal{M}$-cerrada y $\mathcal{M}$-abierta pueden ser vistas como casos especiales, y en el que se permite además incorporar el conocimiento que se tenga respecto a los parámetros del modelo "verdadero". Dado que este modelo puede verse como una combinación ponderada de las perspectivas $\mathcal{M}$-abierto y $\mathcal{M}$-cerrado, este criterio se conoce como enfoque $\mathcal{M}$-mixto. Una de las ventajas de este método es que se puede especificar una probabilidad que refleje la hipótesis que se tenga respecto a que el modelo verdadero se encuentre o no en el conjunto de modelos $\mathcal{M}$.

En este trabajo implementaremos el criterio predictivo de selección de modelos planteado por Gutiérrez-Peña y Walker (2001) para el análisis de datos de valores extremos, de manera que la elección del modelo de valores extremos se haga con base en un criterio eficiente y no se determine únicamente en función de la estimación que se tenga del parámetro de forma de la distribución de valores extremos o en la experiencia del investigador, como se hace generalmente.

En las secciones siguientes se presenta una descripción general del criterio de selección de Gutiérrez-Peña y Walker (2001). Dado que la función de utilidad del modelo predictivo juega un papel fundamental en este criterio de selección, dedicaremos la siguiente sección a presentar el criterio bayesiano de la máxima utilidad esperada.

§3.4. Criterio bayesiano de máxima utilidad esperada

Algunos autores (Box y Hill, 1967; San Martini y Spezzaferri, 1984; Poskitt, 1987) plantean reformular el problema de selección de modelos como un problema de decisión cuya solución es maximizar la utilidad esperada. La idea crucial de este método consiste en introducir una función de utilidad que capture la utilidad de un modelo dados los datos. Las funciones de utilidad que contengan distribuciones finales como argumento no podrán ser utilizadas en la mayoría de los criterios de selección de modelos, ya que el vector de parámetros puede tener diferentes interpretaciones de un modelo a otro, por lo que se utilizarán distribuciones predictivas para evitar este problema. San Martini y Spezzaferri (1984) consideran la función de utilidad $U(f(Y_0|Y), y_0)$ de la distribución predictiva del valor futuro Y_0, donde Y_0 es el verdadero valor futuro desconocido. Por un argumento de Bernardo (1979), los autores indican que la única función de utilidad local propia tiene la forma $b_0 \log f(y_0|Y) + b_1(y_0)$.

El criterio bayesiano de maximizar la función de utilidad esperada final tiene una equivalencia en términos de la divergencia de Kullback-Leibler:

Elegimos el modelo M_1 en lugar de M_2 si

$$w_1 K(f(Y_0|y, M_1), f(Y_0|y, M_2)) > w_2 K(f(Y_0|y, M_2), f(Y_0|y, M_1))$$

donde $K(f_1, f_2) = \int f_1 \log(f_1/f_2)$ denota la divergencia Kullback-Leibler de la densidad f_1 con la densidad f_2.

Con lo visto hasta ahora, se tienen los elementos más importantes para revisar el criterio predictivo de selección de modelos planteado por Gutiérrez-Peña y Walker (2001), en el cual se basa este trabajo.

§3.5. Enfoque $\mathcal{M}$-mixto

§3.5.1. Espacio de "estados de la naturaleza" y espacio de acciones

Denotemos por $Y_f \in \mathcal{Y}_f \subset \mathbb{R}$ a la variable aleatoria futura sobre la que deseamos inferir. En el contexto de la teoría de decisiones $\mathcal{Y}_f$ es el espacio de

"estados de la naturaleza". En el caso de la selección predictiva de modelos se tiene un espacio de acciones consecuentes: en la primera etapa de las acciones debemos elegir un modelo dentro de la clase $\mathcal{M}$, es decir, $\mathcal{A}_1 = \mathcal{M}$, y en la segunda etapa debemos inferir respecto a la variable aleatoria $\mathbf{Y}_f$. El problema se simplifica si definimos como espacio de acciones a la clase

$$\mathbf{P} = \{p_k(\cdot|x^n) : k \in K\}, \tag{3.5}$$

formada por todas las funciones predictivas finales generadas por los modelos de la clase $\mathcal{M}$, las cuales están definidas sobre el espacio $\mathcal{Y}_f \subset \mathbb{R}$, y $x^n = (x_1, \ldots, x_n)$.

§3.5.2. La función de utilidad

Bernardo (1979) plantea el uso de una función logarítmica como una función de utilidad en los problemas de decisión en los cuales el espacio de decisiones consiste en densidades de probabilidad. Esta idea fue retomada por San Martini y Spezzaferri (1984) y Bernardo y Smith (1994), quienes se enfocaron en la construcción de una función de utilidad que describiera la utilidad de la distribución predictiva de cada uno de los modelos en el espacio de decisiones. Específicamente, en situaciones donde la incertidumbre se encuentra en el valor de una observación futura $y = x_{n+1}$, los autores se inclinan por el uso de la función logarítmica $\log\{p_i(y|x^n)\}$, donde $p_i(y|x^n)$ denota la densidad predictiva final bajo el modelo M_i, y $x^n = (x_1, \ldots, x_n)$. En estos casos, la utilidad esperada final está dada por

$$\int \log\{p_i(y|x^n)\}\, p(y|x^n) dy, \tag{3.6}$$

donde $p(y|x^n)$ denota la densidad predictiva final de y.

Como en la práctica generalmente se desconoce cuál es el verdadero modelo, $p(y|x^n)$ es también desconocida; sin embargo, desde la perspectiva del $\mathcal{M}$-cerrado se tiene que

$$p(y|x^n) = \sum_{i=1}^{k} \mathbb{P}(M_i|x^n) p_i(y|x^n), \tag{3.7}$$

donde $\mathbb{P}(M_i|x^n)$ es la ponderación final ligada a cada uno de los modelos en $\mathcal{M}$, $p_i(y|x^n)$ es la densidad predictiva final bajo el modelo M_i, y k el número de modelos considerados. Esto da origen al criterio de San Martini y Spezzaferri (1984).

Para la perspectiva del $\mathcal{M}$-abierto, Bernardo y Smith (1994) plantean una aproximación de validación cruzada para la utilidad posterior esperada, sin embargo, dicha aproximación puede no ser apropiada incluso para tamaños de muestra moderados.

Las ideas anteriores sugieren de manera natural el uso de una función de utilidad en $\mathcal{D} \times \mathcal{F}$ dada por

$$U(M_i, F) = \int \log \{p_i(y|x^n)\}\, dF(y). \tag{3.8}$$

Por lo tanto, el objetivo es maximizar la utilidad esperada final

$$\begin{aligned} \bar{U}(M_i) &= \mathbb{E}_{F|x^n}\left[\int \log \{p_i(y|x^n)\}\, dF(y)\right] \\ &= \int \log \{p_i(y|x^n)\}\, d\mathbb{E}_{F|x^n}\{F(y)\}. \end{aligned}$$

La elección de F debe reflejar la percepción que se tiene respecto a la distribución muestral.

Sea $PD(\alpha_0; F_0)$ un proceso Dirichlet con parámetro de localización F_0 y parámetro de escala α_0. Ferguson (1973) introduce el proceso Dirichlet como una forma de asignar probabilidades en el espacio de las funciones de distribución tales que $\mathbb{E}(F) = F_0$. El resultado de mayor interés sobre el proceso Dirichlet es el siguiente: dado x^n, una muestra de tamaño n de F, entonces $[F|x^n] \sim PD(\alpha_n, F_n)$ y

$$F_n = \frac{\alpha_0 F_0 + nG_n}{\alpha_0 + n}$$

donde G_n es la distribución empírica de la muestra, y la noción general es que α_0 representa el tamaño de muestra inicial y $\alpha_n = \alpha_0 + n$.

Consideremos en primer lugar el proceso Dirichlet desde la perspectiva del $\mathcal{M}$-cerrado, es decir, suponemos que existe una i y un θ_i tal que $f_i(.|\theta_i)$ es la función de densidad muestral verdadera. Posteriormente tomamos $F_0 = f_i(.|\theta_i)$ y hacemos $[F|i, \theta_i] \sim PD(\alpha_0, F_0)$. El hecho de que el proceso esté centrado en F_0 nos permitirá alejarnos de la perspectiva del $\mathcal{M}$-cerrado: cuanto más cerca esté α_0 de $+\infty$, más cerca se estará de la perspectiva del $\mathcal{M}$-cerrado; por el contrario, cuanto más cerca esté α_0 a 0, la perspectiva del $\mathcal{M}$-cerrado se hace más débil. Lo anterior implica que mientras más débil sea la perspectiva de un $\mathcal{M}$-cerrado, más fuerte se hace la de un $\mathcal{M}$-abierto, y viceversa.

Concentrémonos ahora en la perspectiva del $\mathcal{M}$-cerrado, y sea $w_i = \mathbb{P}(M_i)$. Para facilitar la notación, se introduce la variable $\phi = (\phi_1, \ldots, \phi_k)$, donde $\phi_i = 1$ y $\phi_j = 0$, $j \neq i$, si el i-ésimo modelo es verdadero. La distribución inicial para (ϕ, θ), donde $\theta = (\theta_1, \ldots, \theta_k)$, puede escribirse como

$$p(\theta, \phi) = p(\theta|\phi)p(\phi), \tag{3.9}$$

donde

$$p(\theta|\phi) = \sum_{i=1}^{k} \phi_i \pi_i(\theta_i) \tag{3.10}$$

y $p(\phi)$ es la distribución Bernoulli multivariada con pesos $(\phi_1, \ldots, \phi_k)$. Tenemos entonces que:

$$[F|\phi, \theta] \sim PD\Big(\alpha_0, \sum_{i=1}^{k} \phi_i f_i(.|\theta_i)\Big), \tag{3.11}$$

lo que nos lleva a una mezcla del proceso Dirichlet inicial. La distribución inicial para F de San Martini y Spezzaferri (1984) es completamente paramétrica, y corresponde al caso $\alpha_0 = +\infty$. Está dada por

$$[F|\phi,\theta] = \sum_{i=1}^{k} \phi_i f_i(\cdot|\theta_i), \tag{3.12}$$

lo que supone implícitamente que el modelo del mundo real es un miembro del espacio de decisiones: la perspectiva del $\mathcal{M}$-cerrado.
Si no existen empates, Antoniak (1974) muestra que la mezcla final de procesos Dirichlet está dada por

$$F_{\phi,\theta,x^n} \sim PD(\alpha_n, F_n)$$

y

$$p(\theta,\phi|x^n) = p(\theta|\phi,x^n)p(\phi|x^n),$$

con

$$p(\theta|\phi,x^n) = \sum_{i=1}^{k} \phi_i \pi_i(\theta_i|x^n),$$

donde $\pi_i(\theta_i|x^n)$ es la distribución final para θ_i bajo el i-ésimo modelo, y $p(\phi|x^n)$ es la distribución Bernoulli multivariada con ponderaciones dadas por $(w_1^*,\ldots,w_k^*)$, donde $w_i^* = \Pr(M_i|x^n) \propto w_i \int f_i(x^n|\theta_i)\pi_i(\theta_i)d\theta_i$.

§3.5.3. La perspectiva $\mathcal{M}$-mixta.

Con lo que se ha visto hasta ahora, es posible encontrar una solución al problema de decisión para seleccionar el modelo adecuado.

La utilidad esperada final está dada por

$$\overline{U}(M_i) = \mathbb{E}_{F|x^n}\left[\int log\left\{p_i(y|x^n)\right\} dF(y)\right],$$

por lo que es necesario obtener $\mathbb{E}_{F|x^n}\{F(y)\}$, que está dada por

$$\mathbb{E}_{\phi|x^n}(\mathbb{E}_{\theta|\phi,x^n}[\mathbb{E}_{F|\theta,\phi,x^n}\{F(y)\}]) = \frac{\alpha_0 \sum_i w_i^* p_i(y|x^n) + nG_n(y)}{\alpha_0 + n} \tag{3.13}$$

Por lo tanto

$$\overline{U}(M_i) = q_n \overline{U}_C(M_i) + (1 - q_n)\overline{U}_A(M_i), \tag{3.14}$$

donde $q_n = \alpha_0/(\alpha_0 + n)$,

$$\overline{U}_C(M_i) = \int \log\{p_i(y|x^n)\} \sum_{j=1}^{k} w_j^* p_j(y|x^n) dy$$

(la utilidad esperada final obtenida por San Martini y Spezzaferri (1984) para la perspectiva del $\mathcal{M}$-cerrado), y

$$\overline{U}_A(M_i) = \int \log\{p_i(y|x^n)\}\, dG_n(y).$$

Es fácil notar que $\overline{U}(M_i) = \overline{U}_A(M_i)$ cuando $\alpha_0 = 0$. De la ecuación (3.13) se puede ver que cuando $\alpha_0 = 0$ la utilidad final depende únicamente de los datos, lo que sugiere que $\overline{U}_A(M_i)$ corresponde a la perspectiva del $\mathcal{M}$-abierto. El caso de $\alpha_0 = +\infty$ corresponde obviamente a la perspectiva del $\mathcal{M}$-cerrado, y $0 < \alpha_0 < +\infty$ corresponde a la mezcla de los dos modelos. En la perspectiva del $\mathcal{M}$-abierto, se selecciona a M_i en lugar de M_j si

$$\frac{1}{n} \sum_{l=1}^{n} \log\left(\frac{p_i(x_l|x^n)}{p_j(x_l|x^n)}\right) > 0,$$

es decir, se prefiere a M_i en lugar de M_j si

$$\Delta_{ij} \stackrel{def}{=} \prod_{l=1}^{n} \left\{\frac{p_i(x_l|x^n)}{p_j(x_l|x^n)}\right\}^{1/n} > 1.$$

§3.5.4. Ejemplo.

Consideremos el siguiente problema de selección de modelos, desde la perspectiva del $\mathcal{M}$-abierto (ver Gutiérrez-Peña y Walker, 2001).
Sean

$$M_1 = \{N(\centerdot|\theta_1, 1), \pi_1(\theta_1) \propto 1\},$$

y

$$M_2 = \left\{ N(\cdot|0, \theta_2), \pi_2(\theta_2) \propto \theta_2^{-1}) \right\}.$$

Para resolver el problema de decisión y seleccionar al modelo más apropiado, debemos calcular la distribución predictiva final y posteriormente Δ_{12}.

a) La distribución predictiva final está dada por:

$$p_1(y|x^n) = \int p(y|\theta_1) p(\theta_1|x^n) d\theta_1$$

donde

$$\begin{aligned} p(\theta_1|x^n) &= p(\theta_1) p(x^n|\theta_1) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x_i - \theta_1)^2 \right\} \end{aligned}$$

y

$$p(y|\theta_1) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(y - \theta_1)^2 \right\}.$$

Entonces

$$\begin{aligned} &\mathrm{p}_1(y|x^n) = \textstyle\int (2\pi)^{-(\frac{n+1}{2})} \exp\left\{ -\frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \theta_1)^2 + (y - \theta_1)^2 \right] \right\} d\theta_1 \\ &\propto \textstyle\int \exp\left\{ -\frac{1}{2}\left[n\theta_1^2 - 2\theta_1 n\bar{x} + y^2 - 2y\theta_1 + \theta_1^2 \right] \right\} d\theta_1 \\ &\propto \textstyle\int \exp\left\{ -\frac{1}{2}\left[(n+1)\theta_1^2 - 2\theta_1(y + n\bar{x}) + y^2 \right] \right\} d\theta_1 \\ &\propto \textstyle\int \exp\left\{ -\frac{1}{2}(n+1)\left[\theta_1^2 - 2\theta_1\left(\frac{y+n\bar{x}}{n+1}\right) + \left(\frac{y+n\bar{x}}{n+1}\right)^2 \right] \right\} \\ &\qquad \exp\left\{ -\frac{1}{2}y^2 - \frac{(y+n\bar{x})^2}{n+1} \right] \right\} \end{aligned}$$

$$\begin{aligned}
&\propto \exp\left\{-\tfrac{1}{2}\Big(y^2 - \tfrac{y^2+2yn\bar{x}+n\bar{x}^2}{n+1}\Big\}\int \exp\left\{-\tfrac{1}{2}(n+1)\Big(\theta_1 - \tfrac{y+n\bar{x}}{n+1}\Big)^2\right\} d\theta_1\right. \\
&\propto \exp\left\{-\tfrac{1}{2}\Big(\tfrac{n}{n+1}\Big)\Big(y^2 - 2\bar{x}y + \bar{x}^2\Big)\right\} \\
&\propto \exp\left\{-\tfrac{1}{2(1-1/n)}(y-\bar{x})^2\right\} \\
&\propto N(y|\bar{x}, 1-1/n)
\end{aligned}$$

b) Distribución predictiva final para el modelo M_2

$$p_2(y|x^n) = \int_0^\infty p(y|\theta_2)p(\theta_2|x^n)d\theta_2.$$

En primer lugar tenemos que

$$p_2(y|\theta_2) \propto \frac{1}{\theta_2}\exp\left\{-\frac{1}{2\theta_2^2}y^2\right\};$$

Por otra parte

$$p(\theta_2|x^n) \propto \pi(\theta_2)p(x^n|\theta_2),$$

$$\begin{aligned}
p_2(\theta_2|x^n) &\propto \pi(\theta_2)p(x^n|\theta_2) \\
&\propto \frac{1}{\theta_2}\prod_{i=1}^n \frac{1}{\theta_2}\exp\left\{-\frac{1}{2\theta_2^2}x_i^2\right\} \\
&\propto \theta_2^{-(n+1)-1}\exp\left\{-\frac{1}{2\theta_2^2}\sum_{i=1}^n x_i^2\right\}.
\end{aligned}$$

Por lo tanto,

$$p_2(y|x^n) = \int_0^\infty \theta_2^{-(n+1)-1}\exp\left\{-\frac{1}{2\theta_2^2}(nS^2+y^2)\right\}d\theta_2,$$

donde

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Sea $h = 1/\theta_2^2$, entonces

$$\theta_2 = \frac{1}{h^{1/2}} \quad \text{y} \quad d\theta_2 = -\frac{1}{2\theta_2^{3/2}}dh$$

y por lo tanto

$$\begin{aligned}
p_2(y|x^n) &\propto \int_0^\infty h^{\frac{n+1+1}{2}-\frac{3}{2}} \exp\Big\{ -\frac{h}{2}(nS^2+y^2)\Big\}\mathrm{dh} \\
&\propto \int_0^\infty h^{\frac{n+1}{2}-1} \exp\Big\{ -\frac{h}{2}(nS^2+y^2)\Big\}\mathrm{dh} \\
&\propto \Big(nS^2+y^2\Big)^{-\frac{n+1}{2}} \int_0^\infty \mathrm{Gamma}\Big(h, \frac{n+1}{2}, nS^2+y^2\Big)\ \mathrm{dh} \\
&\propto \Big(nS^2+y^2\Big)^{-\frac{n+1}{2}} \\
&\propto \left(1+\frac{y^2}{nS^2}\right)^{-\frac{n+1}{2}} \\
&\propto \mathrm{St}\Big(y|n,0,S^{-2}\Big) \\
&\propto \mathrm{St}\left(y|n,0,\frac{n}{\sum_{i=1}^n x_i^2}\right).
\end{aligned}$$

Una vez encontradas las distribuciones predictivas para los dos modelos, es posible calcular de manera acumulada (para n=100, 200, ... , 10000) la función

$$\Delta_{ij} = \prod_{l=1}^{n}\left\{\frac{p_i(x_l|z_n)}{p_j(x_l|z_n)}\right\}^{1/n}.$$

Observaremos ahora el resultado que se obtuvo de Δ_{12} bajo diferentes escenarios de los parámetros θ_1 y θ_2:

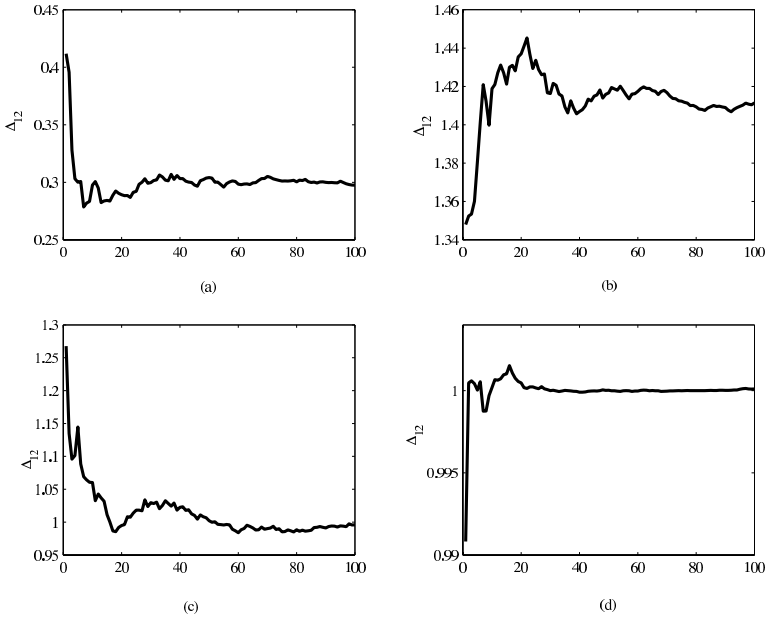
a) $\theta_1 = 0$ y θ_2 =5.

b) $\theta_1 = 1$ y $\theta_2 = 1$.

c) $\theta_1 = 2$ y $\theta_2 = 3$.

d) $\theta_1 = 0$ y $\theta_2 = 1$.

Para cada uno de los casos anteriores, se simularon muestras de tamaño 100, 200, ... , 10000 de la distribución Normal (θ_1, θ_2), y se calculó la función Δ_{ij} como función de n. El programa se realizó en Mathematica, y los resultados obtenidos para los cuatro casos se graficaron en la siguiente figura.



Como se puede observar en la gráfica anterior, los resultados obtenidos para el criterio Δ_{12} son congruentes en cuanto a la muestra que se toma y el modelo que resulta elegido. En el caso (a), $\Delta_{12} < 1$, indicando que se debe elegir el modelo 2. En el caso (b), $\Delta_{12} > 1$, lo que significa que se prefiere al modelo 1. En el caso (c), sin embargo, ninguno de los modelos es el correcto, y por lo tanto Δ_{12} fluctúa alrededor de 1, indicando que Δ_{12} no discrimina entre los dos modelos. Finalmente, en el caso (d) los dos modelos

son adecuados, y nuevamente se tiene que Δ_{12} fluctúa alrededor de 1.
Asimismo, se puede notar que conforme aumenta el tamaño de muestra (n), el criterio Δ_{12} converge a un número que permite decidir cuál de los dos modelos en consideración se debe elegir.

Capítulo 4

Aplicaciones

En los capítulos anteriores se ha revisado un conjunto de herramientas y modelos que permiten formular, de una forma matemáticamente precisa, cuestiones fundamentales sobre valores extremos en el caso unidimensional, enfocándonos particularmente en el procedimiento de selección de modelos planteado por Gutiérrez-Peña y Walker (2001).

En el mundo real, los eventos extremos se presentan a través de un conjunto de datos, como niveles de agua de una presa, reclamaciones a compañías de seguros, velocidad del viento en ciertos lugares, rendimientos de ciertas acciones en el mercado, entre otros. Si bien en todos estos casos el problema subyacente se refiere al comportamiento de valores extremos, el encargado de modelar dicho comportamiento deberá hacer uso de las herramientas que más se adecúen al tipo de problema en cuestión para poder llegar a conclusiones científicamente sustentables a partir de los datos. Asimismo, al realizar un análisis de datos es importante saber reportar: los datos deben presentarse de manera clara y objetiva, se deben formular preguntas precisas y responder con base en los resultados obtenidos en los modelos. Todo este proceso constituye un arte: la teoría estadística juega un papel relativamente pequeño, pero crucial.

En este capítulo se presenta una aplicación de la teoría de selección de modelos en un problema de valores extremos, con el propósito de ejemplificar y poner a prueba las herramientas desarrolladas en los capítulos anteriores. Para ello se trabajó con datos que reflejan el comportamiento de la serie de rendimientos de la Bolsa Mexicana de Valores (BMV).

§4.0.5. Análisis de eventos extremos aplicado a la serie de rendimientos de la BMV

Supongamos que un inversionista analiza la posibilidad de invertir en el índice de la BMV o en acciones de alguna sociedad de inversión que reflejan el comportamiento de dicho índice. Obviamente, un factor importante para tomar la decisión sobre invertir o no, y en su caso, cuánto invertir, será el nivel de rentabilidad y de riesgo esperados en el índice. Una forma de conocer el comportamiento de estos factores es a través de la distribución de pérdidas y ganancias, conocida en inglés como *P & L Distribution*.

Entre las distribuciones utilizadas para describir el comportamiento de este tipo de series de rendimiento se encuentran la Normal y la Lognormal. Sin embargo, cuando se trata de datos que se refieren a rendimientos de algún activo, las colas de estas distribuciones no se ajustan adecuadamente a los datos, lo que generalmente se debe a que las colas de la "verdadera" distribución de los datos son más pesadas, por lo que usar estas distribuciones implica que se subestime la probabilidad de ocurrencia de pérdidas o ganancias mayores.

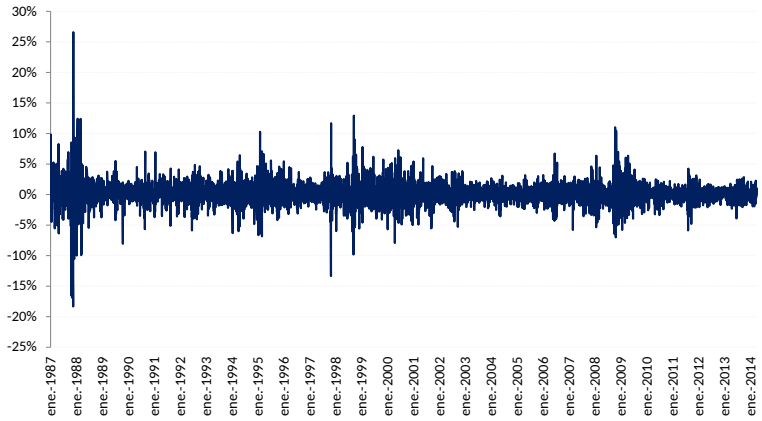
Por otra parte, debemos tomar en cuenta que para cualquier inversionista potencial, los datos de mayor interés no serán únicamente aquéllos que se encuentran alrededor de la media de la distribución, sino también, y de manera muy relevante, los más alejados de ella.

Considerando lo anterior, nuestro interés se centrará en buscar una distribución que describa de la manera más precisa posible el comportamiento de dichos valores, es decir, buscaremos ajustar la mejor distribución a los valores extremos. Utilizaremos el método de selección de modelos desarrollado en el Capítulo 3 para elegir el mejor modelo, considerando los resultados del teorema de Fisher-Tippet, que establece que bajo ciertos criterios de convergencia la distribución solamente puede pertenecer a una de tres familias: Gumbel, Weibull o Fréchet.

Comenzaremos por hacer un análisis exploratorio de los datos, lo que nos permitirá tener una idea mucho más clara de su comportamiento antes de iniciar un análisis estadístico más profundo.

El conjunto de datos que se estudiará consiste en los rendimientos logarítmicos diarios de la serie de precios del Indice de Precios y Cotizaciones (IPC), del 2 de enero de 1987 al 1o de abril de 2014, cuya serie de tiempo e histograma se muestran en las Figuras 4.1 y 4.2.

Figura 4.1
Rendimientos diarios de la BMV



La gráfica de la serie nos ayuda a identificar las pérdidas y ganancias extremas y el momento aproximado en el que ocurrieron. En esta gráfica observamos, por ejemplo, el efecto que tuvo en la BMV la crisis de 1988, la caída que tuvo la bolsa mexicana en octubre de 1997 como consecuencia de la crisis asiática que impactó a la mayoría de las bolsas del mundo, así como la crisis global que se vivió alrededor de 2009, desencadenada en gran parte por el mercado hipotecario *subprime*.

Por otra parte, en el histograma de rendimientos (Figura 4.2) podemos ver el rango de los datos, en dónde se acumula la mayor parte de ellos, y nos da una idea del comportamiento de las colas de la distribución.

Figura 4.2
Histograma de rendimientos de la BMV

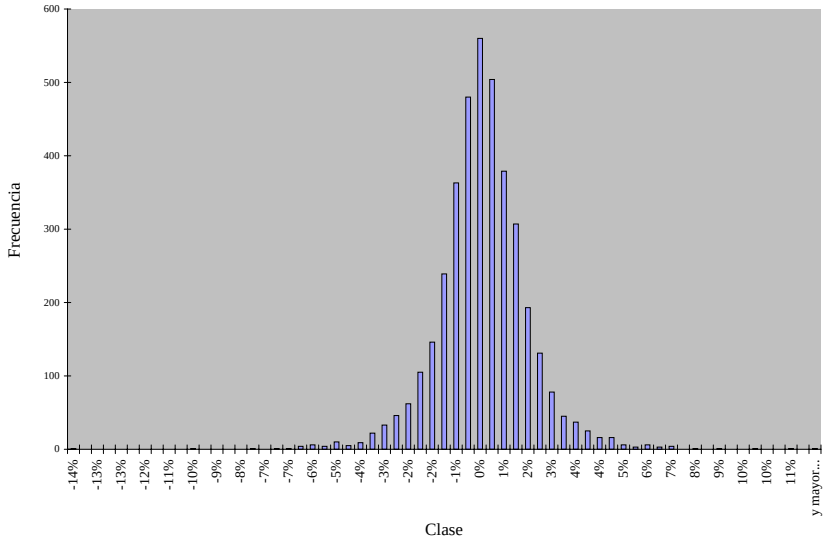
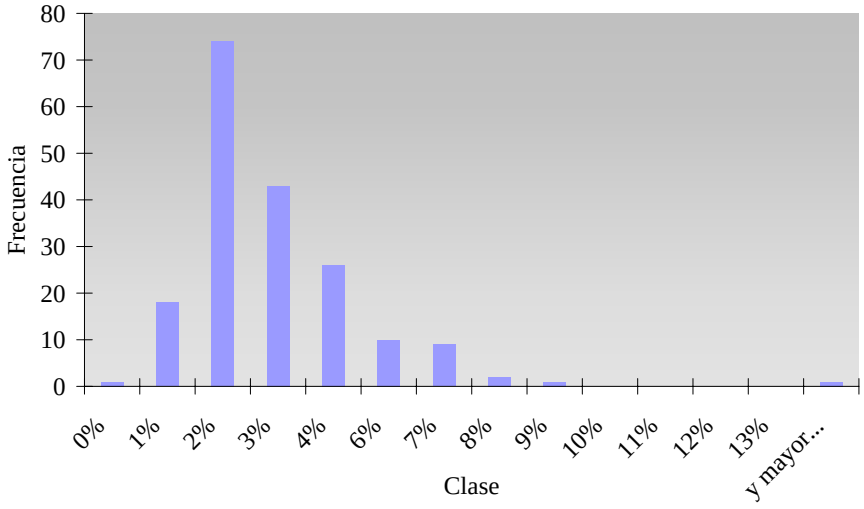


Figura 4.3
Histograma de Valores Extremos de la serie de rendimientos del IPC



Finalmente, para conocer el comportamiento de los datos de valores extremos de la muestra, en la Figura 4.3 se presenta el histograma de estos datos. En él se puede apreciar que los datos presentan un ligero sesgo hacia la derecha; además, existen datos superiores al 13 %, que es un valor muy alto si consideramos que se trata de rendimientos diarios.

Tras este breve análisis centraremos nuestra atención en el problema de selección de modelos.

§4.0.6. Método de selección de modelos

La selección del modelo la trataremos como un problema de decisión que nos permita encontrar la "mejor" distribución predictiva de los datos, entendida como aquella distribución que maximiza la utilidad esperada. Con ese propósito, y bajo las premisas anteriores, implementaremos el procedimiento de Gutiérrez-Peña y Walker (2001) descrito en el Capítulo 3.

Este método de selección de modelos se implementó bajo la perspectiva del $\mathcal{M}$-abierto, por considerarla la alternativa más honesta debido, por una parte, a que no se restringe al conjunto de modelos paramétricos, y por otra, a que únicamente sabemos que los datos extremos convergen en el límite a las distribuciones de valores extremos que estamos considerando. Asimismo es importante notar que, particularmente en este caso, el uso de la distribución predictiva resulta natural, independientemente de que se recomiende su uso en la comparación de modelos, ya que la serie histórica de rendimientos que se analiza se utiliza únicamente como una referencia para medir la incertidumbre del comportamiento futuro del IPC.

Corrida del modelo

Dadas las distribuciones límite Gumbel, Weibull y Fréchet y $x^n = (x_1, ..., x_n)$, tenemos que calcular

$$U_m = \frac{1}{n}\sum_{j=1}^{n} \log\ p_m(x_j|\underline{x}), \qquad m \in G, W, F$$

y elegir el modelo M que maximice U_m.

Las distribuciones finales se obtuvieron usando como distribuciones iniciales las obtenidas con el algoritmo de referencia (Berger y Bernardo, 1992), y que resultaron en:

Distribución	Distribución inicial
Gumbel	$\pi(\mu, \sigma) \propto \sigma^{-1}$
Fréchet	$\pi(\mu, \sigma, \alpha) \propto \sigma^{-1}\alpha^{-1}$
Weibull	$\pi(\mu, \sigma, \alpha) \propto \sigma^{-1}\alpha^{-1}$

Un primer intento fue calcular numéricamente las distribuciones predictivas de cada modelo evaluadas en cada observación que se tenía, es decir, evaluar numéricamente

$$\begin{aligned} p_m(x_j|\underline{x}) &= \int p_m(x_j, \underline{\theta}_m|\underline{x})d\underline{\theta}_m \\ &= \int p_m(x_j|\underline{\theta}_m)p_m(\underline{\theta}_m|\underline{x})d\underline{\theta}_m \\ &= \frac{\int p_m(x_j|\underline{\theta}_m)p_m(\underline{\theta}_m)p(\underline{x}|\theta_m)d\theta_m}{\int p_m(\theta_m)p(\underline{x}|\theta_m)d\theta_m} \\ &\quad \forall\ j \in 1, \ldots, n. \end{aligned}$$

Para probar tanto la integral numérica como el procedimiento, se utilizó Mathematica para resolver la integral numéricamente, utilizando una muestra de 186 datos de mínimos mensuales del IPC correspondientes al período de enero de 1987 a julio de 1990. Se obtuvieron los siguientes resultados:

Distribución	Gumbel	Weibull	Fréchet
Utilidad esperada	-1.18	-0.41	-0.36

En este proceso se tuvieron algunos problemas, como la inestabilidad en el valor de las integrales y el consumo de mucho tiempo y recursos.

Una segunda opción era simular muestras de la distribución final (MCMC), pero no lo hicimos.

Dado que la muestra es grande, se optó por usar la aproximación de Laplace (1774), que permite estimar el cociente de integrales $p_m(x_j|\underline{x})$ como

$$p_m(x_j|\underline{x}) \approx p_m(x_j|\tilde{\theta}_m),$$

donde $\tilde{\theta}_m$ es la moda de la distribución final $p_m(\theta|x)$, y por lo tanto

$$\begin{aligned} U_m &= \frac{1}{n}\sum_{j=1}^{n} \log\ p_m(x_j|\underline{x}) \\ &\approx \frac{1}{n}\sum_{j=1}^{n} \log\ p_m(x_j|\tilde{\theta}_m). \end{aligned}$$

En esta ocasión, para la corrida del modelo se utilizaron los datos de mínimos mensuales de rendimientos diarios del Indice de Precios y Cotizaciones (IPC) de la Bolsa Mexicana de Valores (BMV) de Agosto de 2005 a Mayo de 2014, para trabajar con un total de 327 datos.
A grandes rasgos, lo que se hizo para implementar el planteamiento de Gutiérrez-Peña y Walker (2001) para elegir entre los modelos Gumbel, Fréchet y Weibull, fue lo siguiente:

1. Con los primeros 150 datos resolvimos el problema de selección de modelos:

   - Se obtuvieron las modas de las distribuciones finales:

f.d.	moda_α	moda_μ	moda_σ
Fréchet	3.31473	-0.0146337	0.0329445
Gumbel	-	0.0201029	0.0118115
Weibull	360.254	4.31295	4.29273

- Se encontró la utilidad esperada de cada función de distribución utilizando la moda de la distribución final, conforme a la aproximación de Laplace (1774):

Distribución	Fréchet	Gumbel	Weibull
Utilidad esperada	2.66885	-549.632	2.55099

De acuerdo con este resultado, se elige a la distribución Fréchet por resultar la de mayor función de utilidad.

2. Con los siguientes 100 datos, generamos muestras de la aproximación a la distribución predictiva Fréchet

   - Encontramos $\tilde{\underline{\theta}} = (\tilde{\alpha}, \tilde{\mu}, \tilde{\sigma})$ que maximiza $p(\underline{\theta}|\underline{\mathrm{x}})$
   - Usamos Fréchet$(x|\hat{\underline{\theta}})$ para generar una muestra de tamaño 10,000 $\{x^*_1, \ldots, x^*_{10,000}\} \sim F(x|\hat{\theta})$
   - Encontramos los intervalos de probabilidad al 90 y 95 %:
     $(x^*_{0,05}, x^*_{0,95})$=(0.0104233, 0.0593542)
     $(x^*_{0,025}, x^*_{0,975})$= (0.00940235,0.0790346)

3. Finalmente, con los 77 datos restantes se revisó la cobertura de los intervalos obtenidos vs. los datos reales. En el caso del intervalo de probabilidad al 90 % se observó una cobertura del 87.01 %, y en el del 95 % la cobertura registrada fue del 93.51 %.

   Dada la aproximación, estos intervalos pueden resultar más pequeños que los verdaderos.

El resultado del modelo coincide con lo generalmente establecido en la literatura de valores extremos, en el sentido de que la distribución Fréchet es generalmente la mejor distribución para modelar valores extremos cuando se trata de datos financieros.

Conclusiones

En el marco de la estadística Bayesiana y haciendo uso de herramientas de selección de modelos, en este trabajo se mostró y aplicó una metodología para elegir una densidad y obtener predicciones de valores extremos. A diferencia de las técnicas más comunes utilizadas en estos casos, en este trabajo se considera en el modelo de predicción el desconocimiento que se tiene respecto a la distribución "verdadera", dejando que en su lugar "hablen los datos" y nos permitan elegir una de las tres distribuciones límite a las que, como se ha demostrado, sabemos que converge la distribución verdadera.

La implementación del modelo se hizo de dos maneras: en la primera se utilizó la integración numérica para aproximar la distribución predictiva, basados en la función de utilidad propuesta en la metodología de selección de modelos de Gutiérrez-Peña & Walker (2001); en la segunda, se utilizó la aproximación de Laplace. En ambos casos se eligió la misma distribución, la Fréchet, la misma que se utiliza generalmente en este tipo de datos según la literatura escrita sobre valores extremos. En este caso, sin embargo, tenemos mayor certeza de que esta distribución es la que mejor modela el fenómeno en estudio.

Apéndice

APROXIMACIÓN DE LAPLACE

Nos interesa encontrar

$$\mathbb{E}[g(\theta)] = \frac{\int g(\theta)\pi(\theta)p(x|\theta)d\theta}{\int \pi(\theta)p(x|\theta)d\theta},$$

en particular cuando $g(\theta) = p(y|\theta)$ de manera que $\mathbb{E}[g(\theta)] = p(y|x)$, la distribución predictiva de y dada la muestra observada x de tamaño m.

Haciendo

$$h_N(\theta) = h_D(\theta) = \ -\tfrac{1}{m}\ \log\ \pi(\theta)p(x|\theta),$$

$$b_N(\theta) = g(\theta) \text{ y } b_D(\theta) = 1,$$

tenemos que

$$\mathbb{E}[g(\theta)] = \frac{\int b_N(\theta)\exp\{-mh_N(\theta)\}d\theta}{\int b_D(\theta)\exp\{-mh_D(\theta)\}d\theta}.$$

Laplace (1774) demostró que si b y h son funciones suaves y

$$\tilde{\theta} = \text{argmax}[-h(\underline{\theta})] \quad y \quad \tilde{\sigma}^2 = [h''(\theta)_{\theta=\tilde{\theta}}]^{-1}$$

entonces

$$\int b(\theta)\exp\{-mh(\theta)\}d\theta \approx \sqrt{2\pi\sigma^2}\exp\{-mh(\tilde{\theta})\}\{b(\tilde{\theta}) + \sigma(m^{-2})\}$$

La demostración es similar a la utilizada para demostrar normalidad asintótica: se desarrolla el logaritmo del integrando en series de Taylor alrededor de $\tilde{\theta}$ y se omiten los términos de orden mayor o igual a tres (ver Tierney & Kadane (1986), Tierney et al. (1989) y Kass et al. (1990), en donde se dan las condiciones para que la aproximación sea válida).

Si $g(\theta) = p(y|\theta)$, entonces

$$p(y|x) = \mathbb{E}[p(y|\theta)|x] = \frac{\int p(y|\theta)p(\theta)p(x|\theta)d\theta}{\int p(\theta)p(x|\theta)d\theta}$$

y haciendo

$$h_N(\theta) = h_D(\theta) = \ -\frac{1}{m}\ \log\ p(\theta)p(x|\theta)\ d\theta,$$

$$b_N(\theta) = p(y|\theta) \quad y \quad b_D(\theta) = 1,$$

tenemos que

$$p(y|x) = \frac{\int b_N(\theta)\ \exp\{-mh_N(\theta)\}d\theta}{\int b_D(\theta)\exp\{-mh_D(\theta)\}d\theta}.$$

Si $\tilde{\theta}$ es el máximo de la distribución final y $\tilde{\sigma}^2 = [h''_N(\theta)_{\theta=\tilde{\theta}}]^{-1}$, tenemos que

$$p(y|x) \approx \frac{b_N(\tilde{\theta})\tilde{\sigma}\exp\{-mh_N(\tilde{\theta})\}}{\tilde{\sigma}\exp\{-mh_N(\tilde{\theta})\}} \approx p(y|\tilde{\theta}),$$

es decir, podemos aproximar a la densidad predictiva $p(y|x)$ con la densidad muestral de y evaluada en la moda de la distribución final.

Bibliografía

[1] Adler, R., Feldman, Raya, Taqqu, Murad (1998). *A Practical Guide to Heavy Tails.* Birkhäuser Basel.

[2] Aitkin, M. (1991). Posterior Bayes factors. *J. Roy. Statist. Soc. 53, 111-142.*

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. En: *B. N. Petrov F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267-281).* Budapest: Akademiai Kiado.

[4] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. *Ann. Statist. 2, 1152-1174.*

[5] Balkema, A., y de Haan, L. (1974).Residual life time at great age. *Ann. Prob. 2, 792-804.*

[6] Berger, J.O. y Bernardo, J.M. (1992). On the development of reference priors (con discusión). En: *Bayesian Statistics 4 (J.M. Bernardo, J.O. Berger, A.P. David, A.F.M. Smith, eds). 35-60.*

[7] Berger, J. y Pericchi, L. (1996). The instrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc. 91, 109-122.*

[8] Bernardo, J.M. (1979). Expected information as expected utility. *Ann. Statist. 7, 686-690.*

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

[10] Box, G. E. P. (1980). Sampling and Bayes's inference in scientific modelling and robustness. *J. Roy. Statist. Soc. 143, 383-430.*

[11] Box, G. y Hill, W. (1967). Discrimination among mechanistic models. *Technometrics 9, 57-71.*

[12] Clemen, R. (1996). *Making Hard Decisions: An Introduction to Decision Analysis, 2nd edition.* J Belmont CA: Duxbury Press.

[13] de Finetti, B. (1974). *Theory of Probability.* Chichester: Wiley.

[14] Embrechts, P., Klüppelberg, C. y Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Berlin: Springer-Verlag.

[15] Fama E.F. (1965). *The Behavior of Stock Market Prices.* University of Chicago, *34-105.*

[16] Ferguson, T.S. (1973). A bayesian analysis of some nonparametric problems. Ann. Statist. *2, 209-230.*

[17] Fisher, R.A. y Tippet L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc. 24, 180-190.*

[18] Geisser, S., Eddy, W.F., (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc. 74, 153-160.*

[19] Gelfand, A. E., Dey, D.K., Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *(J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, eds.), Bayesian Statistics 4.* Oxford: University Press.147-167.

[20] Gelfand, A.E. (1995). *Model determination using sampling-based methods. En: Markov Chain Monte Carlo in Practice, (W. Gilks, S. Richardson., D. Spiegelhalter, eds.).* London: Chapman-Hall. 145-161.

[21] Gelfand, A.E., Ghosh, S. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika 85, 1-11.*

[22] Gnedenko B.V. y Kolmogorov A.N. (1954). *Limit Theorems for Sums of Independent Random Variables.* Addison-Wesley, Cambridge, MA: Addison Wesley.

[23] Good, I.J. (1950). *Probability and the Weighing of Evidence.* London: Griffin.

[24] Gutiérrez-Peña, E. y Walker, S.G. A bayesian predictive approach to model selection. *J. Statist. Planning and Inference 93, 2001.*

[25] Jeffreys, H. (1939). Theory of Probability. *Oxford Classic Texts in the Physical Sciences.*

[26] Jeffreys, H. (1946). An invariant form for de prior probability in estimation problems. *Proc. Royal Soc. London A, 453-461.*

[27] Kass, R.E., Tierney, L. y Kadane, J.B. (1990) The validity of posterior expansions based on Laplace's method. *Bayesian and likelihood methods in statistics and econometrics. (S. Geisser, J.S. Hodges, S.J. Press y A. Zellner, eds.)* New York: North-Holland. 473-488.

[28] Laplace, P.S. (1774) Memoir sur la probabilité des causes. *Mémoires de L'Académie Royal de Sciences de Paris 6, 621-656.* Traducido con una nota introductoria por Stigler, S.M. (1986). *Laplace's (1774)* Memoir on inverse probability. *Statistical Science 1, 359-378.*

[29] Laplace, P.S. (1812) *Théorie Analytique des Probabilités.* Paris: Courcier.

[30] Laud, P.W., Ibrahim, J.G. (1995). Predictive model selection. *J. Roy. Statist. Soc. 57, 247-262.*

[31] Mandelbrot, B. (1963). New methods in statistical economics. *J. Political Economy 71, 421-440.*

[32] Nolan, J. (2005). *Stable distributions.* American University.

[33] O'Hagan, A. (1994) *Bayesian Inference. (Kendall's Advanced Theory of Statistics 2B)* London: Harold.

[34] O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. B (Methodological) 57, No. 1(1995), pp. 99-138*

[35] Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist. 3 119-131.*

[36] Poskitt, D.S. (1987). Bayes ARMA model determination: some empirical evidence *Australian Journal of Statistics 29, 3, 334-347.*

[37] Rachev, S.T., Mittnik, S. y Paolella, M.S. (2000). Diagnosing and treating the fat tails in financial returns data. *Journal of Empirical Finance 7, 389-416.*

[38] Reiss, R.D., Thomas, M. (2001). *Statistical Analysis of Extreme Values: from Insurance, Finance, Hydrology, and Other Fields.* Birkhäuser Verlag.

[39] San Martini, A., Spezzaferri, F. (1984). A predictive model selection criterion. *J. Roy. Statist. Soc. 46, 296-303.*

[40] Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. Statist. 6*

[41] Smith, A.F.M. Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B, 213-220.*

[42] Spiegelhalter, D.J., Smith, A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. 44, 377-387.*

[43] Tierney, L. y Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc. 81, 82-86.*

[44] Tierney, L., Kass, R.E. y Kadane, J.B. (1989). Approximate marginal densities of non linear functions. *Biometrika 76, 425-433.*