UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Selección de umbral para modelar paramétricamente la distribución de excesos.

Matias Ezequiel Zylbersztejn

Dirección: Dra. Mariela Sued

Agosto 2024

All models are wrong, but some are useful.
[Todos los modelos están mal, pero
algunos son útiles.]

George Box

Agradecimientos

Esta tesis pude terminarla gracias al apoyo de mucha gente, no solo durante la confección de la misma, sino también durante el camino recorrido en la vida, y particularmente en la facultad.

Quiero agradecer a mi mama y a mis hermanos apoyándome y ayudando para poder llegar hasta acá, y aunque ya no puede estar acá, también a mi papa que se que estaría muy contento en este momento.

A mis amigos de siempre con los que compartimos numerosas etapas de nuestro crecimiento.

También a todos mis compañeros y amigos que me hice acá en la facultad, a los geos con los que a pesar de cambiarme de carrera continuamos compartiendo montones de cosas. Pero también a los biólogos, los químicos, los físicos, los matemáticos. Es verdad que no conocí a todos al mismo tiempo, sino que en distintos momentos y etapas, con algunos compartimos muchos momentos de estudio, de "estrés académico", de trabajos prácticos, viajes de campo. Con otros compartimos partidos de futbol, noches de juegos, y hasta encuentros literarios.

A mi compañera.

Agradezco también a Mariela, mi directora, que además de la paciencia que me tuvo, me marcaba límites para no irme por la ramas y poder cerrar esta etapa antes de abrir nuevas.

Y finalmente a todos los trabajadores de la educación de todos los niveles educativos tanto docentes como no docentes, ya que sin ellos no hubiese tenido educación pública y de calidad, y la culminación de esta etapa es también posible gracias a ellas y ellos.

Índice general

1. Introducción **3**

2. EMV y divergencia-KL **5**
2.1. Introducción 5
2.2. Estimador de Máxima Verosimilitud 6
2.2.1. Consistencia del EMV 6
2.3. Divergencia de Kullback - Leibler 9
2.4. EMV y divergencia KL 11

3. El problema de los extremos **13**
3.1. Introducción 13
3.2. Modelando la distribución del máximo 14
3.3. Excesos por encima de un umbral 15
3.4. Aplicación 17

4. Detección de umbrales **19**
4.1. Un modelo semiparamétrico 19
4.1.1. Definiciones 19
4.1.2. El modelo 20
4.2. Estimando el umbral 21

5. Simulación **23**
5.1. El modelo uniforme-exponencial 23
5.2. Procedimiento 24
5.2.1. Generación de la muestra 24
5.2.2. Procesamiento 24
5.3. Análisis 24
5.3.1. Calidad general del estimador del u_0 25
5.3.2. Calidad del modelo estimado a partir del umbral 26

6. Conclusiones **33**

A. R scripts **35**
A.1. Funciones utilizadas para generar las muestras 35
A.2. Funciones utilizadas para estimar u_0 35

Capítulo 1

Introducción

En este trabajo se propone un método automático de detección de umbral para valores extremos de una distribución a partir de una muestra aleatoria de la misma, utilizando un modelo semiparamétrico en donde existe un umbral a partir del cual la distribución de excesos se distribuye como $\mathcal{E}(\lambda)$, que pertenece a la *familia de Pareto generalizada (FPG)*. Probaremos dicha prupuesta numéricamente con el software R.

La teoría de valores extremos es un área de interés en la estadística con numerosas utilidades en diversas aplicaciones en las ciencias naturales y sociales, en áreas como la hidrología, climatología, entre otras. Su importancia radica en que brinda herramientas para analizar la ocurrencia de eventos extremos a partir de los datos disponibles.

En ella se desarrolla un enfoque que permite estudiar, entre otras cosas, dos tipos de fenómenos. Por un lado, el comportamiento asintótico de la distribución del máximo de n variables aleatorias *independientes e idénticamente distribuidas (iid)*. Y por otra parte, el comportamiento de las colas de la distribución de una variable a través de la distribución de $X - u \mid X > u$, llamada distribución de los excesos, a medida que u tiende a infinito. Un resultado importante, desarrollado por [Pickands, 1975], es que en caso de existir el límite, este pertenece a la *FPG*.

La pregunta que buscamos resolver es a partir de que valor *umbral* es válido considerar que se está dentro del régimen donde se puede modelar la distribución de excesos mediante una distribución de la *FPG*. Es sabido que al encontrar un valor adecuado, cualquier valor mayor lo es también. Pero hay un interés en encontrar el mínimo; hacerlo implica poder aprovechar mejor los datos disponibles y además tener un modelo que permita realizar estimaciones en un rango mayor de valores. Además se busca hacerlo de una manera automática, que no dependa de la experiencia o de la subjetividad del analista.

En la bibliografía se pueden encontrar diversos métodos que abordan este problema. Existen métodos gráficos que se basan en la media de los residuos respecto del candidato a umbral, o en QQ-plots, ver [Coles, 2001, ch. Threshold Models]. Más reciente en el tiempo se abordó el tema también de manera paramétrica, donde se asume que los valores extremos son outliers de un modelo paramétrico y proponen un método de detección del *umbral* a partir de la detección de estos outliers, ver [Cabras and Morales, 2007]. También existen propuestas paramétricas en donde particionan la recta real, y asumen una distribución para los valores previos al umbral, y una distribución de la *FPG* para los valores posteriores, ver [Wong and Li, 2010]. Esta idea de partir la recta en dos también se puede utilizar sin asumir ninguna distribución para los valores previos al umbral y en su lugar estimar la densidad de la distribución previa a través de núcleos no paramétricos, ver por ejemplo [Gonzalez et al., 2013] o [MacDonald et al., 2011].

En este trabajo se utiliza un enfoque similar a estos últimos, pero sin estimar ninguna distribución para los datos previos al umbral. Se espera que minimizando la distancia entre el modelo y

las distribuciones de excesos empírica se lo pueda identificar.

La presente tesis se estructura de la siguiente manera: en primer lugar veremos elementos teóricos que permiten fundamentar el origen de la propuesta, en el capítulo 2 describiremos la relación entre el *estimador de máxima verosimilitud (EMV)* y la *divergencia de Kullback-Leibler (divergencia KL)*. Definiremos cada uno de estos conceptos, y daremos condiciones suficientes para garantizar consistencia del *EMV* y a que converge en ese caso, aún cuando el modelo esté equivocado. Veremos que ese límite representa una especie de proyección de la distribución real en el espacio de distribuciones dado por el modelo considerando como pseudo-distancia a la *divergencia KL*, el término pseudo-distancia se debe a que la *divergencia KL* no es una métrica matemáticamente hablando pues no es simétrica.

En el capítulo 3 introduciremos matemáticamente la teoría de valores extremos. En primer lugar se estudiará la distribución del *máximo* de muestras aleatorias y se darán condiciones suficientes bajo las cuales se puede garantizar que el *máximo* se distribuye asintóticamente según la distribución *Gumbel, Fréchet y Weibull*, y que estas tres son las únicas posibles. Mencionaremos las limitaciones de este enfoque y luego describiremos una alternativa partiendo de estudiar la distribución de excesos, el cual tiene como ventaja la posibilidad de utilizar un mayor número de datos, y daremos condiciones suficientes bajo las cuales, asintóticamente, se distribuyen según *FPG*. Debido a que en este trabajo se restringió el alcance solamente a cuando los excesos se distribuyen de manera exponencial, se demostrará que la distribución exponencial pertenece a la *FPG*.

En el capítulo 4 describiremos nuestra propuesta para estimar el umbral y el argumento en el cual se basa. La única hipótesis que se asume es que las variables aleatorias se distribuyen de manera tal que existe un umbral, a partir del cual la distribución de excesos se distribuye como una distribución exponencial.

En el capítulo 5 se presentarán los resultados de una simulación, donde exploramos de manera empírica el comportamiento de nuestra propuesta tanto para estimar el umbral como para valernos del mismo para estimar de manera más eficiente percentiles extremos de la distribución.

Finalmente, en el capítulo 6 haremos una recopilación de lo estudiado en la presente tesis y presentaremos los principales resultados obtenidos.

Capítulo 2

Método de máxima verosimilitud y minimización de la divergencia de Kullback - Leibler

En este capítulo estudiaremos la relación entre la estimación del parámetro de una distribución de probabilidad por el método de máxima verosimilitud y el parámetro que minimiza la divergencia de Kullback-Leibler entre una distribución de probabilidad y una familia paramétrica de distribuciones.

2.1. Introducción

Un problema clásico de la estadística es inferir información acerca de un proceso a partir de los datos que éste genera. Es decir, trabajamos con X_i, $i \geq 1$ iid distribuídos como X, con $X \sim F$ y se procura estimar F o alguna característica de la misma. Para ello, se propone un modelo y luego se utilizan los datos para, mediante alguna regla de decisión, *ajustarlo*.

El modelo propuesto puede estar compuesto por una familia de distribuciones indexadas por un conjunto finito de parámetros. Cuando éste es el caso se habla de *modelo paramétrico*. Por ejemplo, al proponer que una serie de datos siguen una distribución normal $\mathcal{N}(\mu, \sigma^2)$ estamos primero proponiendo como modelo a $\mathcal{M} = \left\{F : F \sim \mathcal{N}(\mu, \sigma^2)\right\}$, dejando lugar para ajustar con los datos a los parámetros (μ, σ^2). Otro ejemplo es el que vamos a desarrollar en los siguientes capítulos de este trabajo, el cual consiste en modelar paramétricamente la distribución de excesos $F_u \sim X - u \mid X > u$ mediante una función de distribución que pertenece a la *familia de Pareto Generalizada*.

En un extremo opuesto está el caso cuando el modelo está compuesto por una familia que no están indexadas por un conjunto finito de parámetros. Por ejemplo, dadas una serie de observaciones independientes $X_1, X_2, \ldots, X_n$ con $X_i \sim F$, y se busca estimar F sin ninguna otra hipótesis.

Cuando proponemos un modelo paramétrico, algo que puede suceder es que el modelo propuesto no sea lo suficientemente *bueno* o flexible para explicar el fenómeno; más precisamente que la distribución de probabilidad del fenómeno que genera datos no pertenezca al modelo $\mathcal{M}$ propuesto. En este caso es de interés saber que tan lejos está el modelo del fenómenos real. O sea es útil tener una medida de cuan distintas son dos distribuciones de probabilidad o de alguna forma de cuantificar la información que se pierde cuando se utiliza un modelo incorrecto.

En este capítulo abordaremos el caso en que se plantea un modelo *paramétrico* de un parámetros, y como estimarlo mediante la *estimación de máxima verosimilitud (EMV)*. Y para medir su

distancia con el fenómeno real, estudiaremos la *divergencia de Kullback-Leibler (divergencia-KL).*

Por otra parte, la estimación no paramétrica de la distribución F está dada por $\widehat{F}$, la denominada distribución empírica: $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{X_i \leq t}$.

También asumiremos que tanto la distribución verdadera F como las propuestas por diferentes modelos tienen función de densidad; que tienen *esperanza*. Y finalmente consideraremos un modelo paramético $\mathcal{M} = \{F(\cdot;\theta), \theta \in \Theta\}, \Theta \subset \mathbb{R}$; es decir, un conjunto de distribuciones indexadas con el parámetro θ . Para referirnos a los elementos de $\mathcal{M}$ a veces usaremos la siguiente notación: F_θ para referirnos al elemento $F(\cdot;\theta)$.

2.2. Estimador de Máxima Verosimilitud

Intuitivamente este método propone asumir que los datos provienen del modelo paramétrico $\mathcal{M}$ elegido. Esto significa que los datos son observaciones (o realizaciones) de una muestra que corresponden a un elemento $F(\cdot;\theta_0) \in \mathcal{M}$. Lo que resta para terminar de describir el proceso que genera los datos es encontrar θ_0. Para ello se considera a la función de densidad conjunta evaluada en las observaciones como una función del parámentro θ, y la estrategia consiste en estimar θ_0 con aquel $\tilde{\theta}$ que la maximice. Esta función es conocida como *función de verosimilitud*. Más precisamente:

Definición 2.2.1 (Función de verosimilitud)**.** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria *iid*, con $X_i \sim F(\cdot;\theta) \in \mathcal{M}$, y $\mathbf{x} = (x_1, x_2, ..., x_n)$ una realización de esta muestra. Definimos:

$$L(\theta|\mathbf{x}) := \prod_{i=1}^{n} f(x_i;\theta), \tag{2.1}$$

donde $f(\cdot;\theta)$ es la función de densidad correspodniente a $F(\cdot;\theta)$.

La notación $L(\theta|\mathbf{x})$ enfatiza que el parámetro θ varía mientras que el vector $\mathbf{x}$ de las realizaciones de la muestra aleatoria está fijo.

Podemos ahora sí definir el *estimador de máxima verosimilitud* $\tilde{\theta}_n^{EMV}$:

Definición 2.2.2 (Estimador de máxima verosimilitud)**.** Dada $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria *iid*, con $X_i \sim F(\cdot,\theta) \in \mathcal{M}$, y $\mathbf{x} = (x_1, x_2, ..., x_n)$ una realización de esta muestra. Consideremos $\tilde{\theta}_n^{EMV}(\mathbf{x}) = \underset{\theta\in\Theta}{argmax}\, L(\theta|\mathbf{x})$. Entonces definimos al *estimador de máxima verosimilitud (EMV)* como $\tilde{\theta}_n^{EMV}(\mathbf{X})$

Debido a las dificultades comunes al trabajar con productorias y que las probabilidades son valores entre 0 y 1, cuando es posible, se suele trabajar con logaritmos, el cual convierten el producto en sumas, y que al ser una función monótona creciente preservan el *argmax*, obteniendo la siguiente formulación:

$$\begin{aligned}\tilde{\theta}_n^{EMV}(\mathbf{x}) &= \underset{\theta\in\Theta}{argmax}\,(\log\,(L(\theta|\mathbf{x}))) \\ &= \underset{\theta\in\Theta}{argmax}\left(\sum_{i=1}^{n}\log\,(f(x_i;\theta))\right)\end{aligned} \tag{2.2}$$

2.2.1. Consistencia del EMV

Una propiedad razonable que se espera del estimador utilizado es que converja al parámetro verdadero al tener más datos. Esta propiedad es la consistencia.

Definición 2.2.3 (Consistencia de un estimador)**.** Sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria *iid*, con $X_i \sim F(\cdot, \theta) \in \mathcal{M}$. Diremos que $\tilde{\theta}_n(\mathbf{X})$ es un estimador consistente para $\theta \in \Theta$ si, en algún sentido,

$$\tilde{\theta}_n(\mathbf{X}) \xrightarrow[n\to\infty]{} \theta \ , \quad \text{para todo } \theta \in \Theta$$

En el caso en que la convergencia sea *casi segura*, se dice que es fuertemente consistente. Y si la convergencia es en *probabilidad* se dice que es débilmente consistente.

A continuación, valiéndonos del enfoque de los *M-estimadores* daremos algunas condiciones que aseguran la consistencia del *EMV*. Se utiliza este enfoque ya que permite generalizar los resultados a una familia de estimadores más general que solo a los estimadores de máxima verosimilitud, y también nos va a permitir definir un marco de trabajo en el cual consideremos que el modelo es incorrecto. Además se espera que pueda ser un aporte a quien lea el presente trabajo ya que es un enfoque poco conocido en los estudios de grado.

M-estimadores

En este apartado nos basaremos en el capítulo 5 del libro *Asymptotic Statistics* de A.W. van der Vaart, [Vaart, 1998, ch. M-and Z-Estimators]. Este no es el enfoque original, Fisher ya había estado estudiando algunas propiedades de los *EMV* en la década de 1920, ver por ejemplo [Fisher, 1925], mientras que los *M-estimadores* se comenzaron a sistematizar y estudiar recién en la década de 1960, cuando se hicieron fundamentales para el estudio de estimadores robustos ([Vaart, 1998, p. 61]).

Sea X una variable aleatoria con distribución F: $X \sim F$. Dada una función $m(x;\theta)$, definimos $M(\theta) = M(\theta; F) = \mathbb{E}_F(m(X;\theta))$. Supongamos que existe un único valor $\theta_0 = \theta_0(F)$ que maximiza la función $M(\theta; F)$. Estamos interesados en estimar $\theta_0(F)$ utilizando una muestra $X_1, \ldots, X_n$ iid, $X_i \sim F$. Siendo $M(\theta)$ una esperanza, la *Ley de los Grande Números (LGN)* sugiere que puede ser estimada con un promedio; podemos entonces estimar el parámetro de interés maximizando el promedio. Más específicamente, consideremos la siguiente definición:

Definición 2.2.4. $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria, con $X_i \sim F$ y $\mathbf{x} = (x_1, x_2, ..., x_n)$ una realización de esta muestra. Consideremos la función:

$$M_n(\theta|\mathbf{x}) = \frac{\sum_{i=1}^{n} m(x_i;\theta)}{n}. \tag{2.3}$$

Definimos

$$\hat{\theta}_n(\mathbf{x}) := \underset{\theta\in\Theta}{argmax}\ M_n(\theta|\mathbf{x}).$$

Decimos que $\hat{\theta}_n(\mathbf{X})$ es un M-estimador de $\theta_0(F)$. Cabe mencionar que, puede suceder que $\hat{\theta}_n(\mathbf{X})$ no exista o que haya múltiples maximizadores; en este trabajo asumiremos el caso habitual en el que $\hat{\theta}_n(\mathbf{X})$ existe y es único.

Para estos casos más generales es posible utilizar la siguiente definición, basada en el libro *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory* de Brown, D, [Brown, 1986, ch. Maximum Likelihood Estimation]. En el cual primero define el conjunto de que maximizan las *estimaciones*, el cual puede ser el conjunto vacío, poseer un único elemento, o poseer más. Y luego un M-estimador sería una función que seleccione algún elemento de ese conjunto. Más precisamente:

Definición 2.2.5. Sean $\mathbf{X}, \mathbf{x}, M_n(\theta|\mathbf{x}), m(x;\theta)$ como en la definición anterior (2.2.4). Consideremos la siguiente función:

$$l(\Theta|\mathbf{x}) = \sup\{M_n(\theta|\mathbf{x}) : \theta \in \Theta\}$$

Y sea

$$\widehat{\theta}_\Theta(\mathbf{x}) = \{\theta \in \Theta : M_n(\theta|\mathbf{x}) = l(\Theta|\mathbf{x})\}$$

Notemos que $\widehat{\theta}_\Theta$ es un subconjunto de Θ compuesto por aquellos θ que maximizan las M_n. Si no existen dichos θ este conjunto es vacío.
Luego, sea una función $\delta : \mathbb{R}^n \to \Theta$, tal que $\delta(\mathbf{x}) \in \widehat{\theta}_\Theta(\mathbf{x})$, entonces esta función δ diremos que es un M-estimador.

Observación 2.2.6. El *estimador de máxima verosimilitud* es un *M-estimador*. Notar que si

$$m(x, \theta) = \log(f(x, \theta))$$

se tiene que

$$\underset{\theta\in\Theta}{argmax}\ M_n(\theta|\mathbf{x}) = \underset{\theta\in\Theta}{argmax} \sum_{i=1}^{n} \log(f(\theta|\mathbf{x}))$$

O sea $\widehat{\theta}_n(\mathbf{X}) = \tilde{\theta}_n^{EMV}(\mathbf{X})$

La observación 2.2.6 justifica porque el enfoque de M-estimadores, pues el EMV es un caso particular de M-estimador.

Como decíamos al introducir a los *M-estimadores*, el parámetro de interés es un $\theta_0(F)$, que maximiza la función $M(\theta; F) = E_F(m(X; \theta))$, por lo tanto, $\hat{\theta}_n$ es consistente si $\hat{\theta}_n \xrightarrow{p} \theta_0$. Por la *LGN* es razonable que se cumpla, sin embargo para demostrar que los maximizadores de las M_n converjan a un θ_0 que maximice $M(\theta; F)$ es necesaria una condición que de una convergencia en un sentido más "funcional" [Vaart, 1998, p. 45], una posibilidad es que las funciones M_n converjan en un sentido uniforme a un $M(\theta; F)$. Asumiento esta hipótesis, vamos a demostrar consistencia del M-estimador vamos a utilizar la demostración del libro ya mencionado [Vaart, 1998], asumiendo también que θ_0 sea el único máximo de $M(\theta)$, y que solo valores de θ cercanos a θ_0 cumplan que $M(\theta)$ esté cerca de $M(\theta_0)$. Esta propiedad se la conoce como que θ_0 sea un punto máximo de M bien separado. Más precisamente:

Teorema 2.2.7 (Consistencia de M-estimadores)**.** *Sea M_n, $n \geq 1$ una sucesión de funciones aleatorias que convergen uniformemente a M, tal que para $\forall \epsilon > 0$*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\quad p \quad} 0 \tag{2.4}$$

$$\sup_{\theta:\, d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0) \tag{2.5}$$

Si $\hat{\theta}_n$ es una sucesión de estimadores que cumplen con $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - Y_n$ con Y_n V.A. que convergen en probabilidad a 0. Entonces la sucesión de estimadores $\hat{\theta}_n \xrightarrow{\quad p \quad} \theta_0$ Notar que la última condición dada sobre los estimadores incluye al estimador de máxima verosimilitud y a otros que se aproximen razonablemente bien al máximo.

Demostración. Por hipótesis tenemos que los $\hat{\theta}_n$ cumplen $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - Y_n$ con $Y_n \xrightarrow{p} 0$. Como M_n converge uniformemente a M:

$$\begin{aligned} 0 \leq M(\theta_0) - M(\hat{\theta}_n) &\leq M(\theta_0) - M_n(\theta_0) + M_n(\theta_0) - M_n(\hat{\theta}_n) + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \\ &\leq 2 \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + Y_n \xrightarrow{p} 0 \end{aligned}$$

Donde en el último paso utilizamos la conidición 2.4 y que por hipótesis $Y_n \xrightarrow{p} 0$. Luego, por la condición 2.5 se tiene que para todo $\epsilon > 0$, existe un número $\nu > 0$ tal que $M(\theta) < M(\theta_0) - \nu$ para

todo θ con $d(\theta, \theta_0) \geq \epsilon$. Por lo que el conjunto $d(\hat{\theta}_n, \theta_0) \geq \epsilon$ está incluido en $M(\theta_0) - M(\hat{\theta}_n) > \nu$, el cual tiene probabilidad convergente a 0, por lo que $P(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \to 0$ que es lo que queríamos probar. □

La convergencia uniforme es una hipótesis fuerte, sin embargo *van der Vaart* afirma que la condición es equivalente a que el conjunto $\{m_\theta : \theta \in \Theta\}$ sea *Glivenko - Cantelli* [Vaart, 1998, p. 46], y que las familias paramétricas $\{f_\theta : \theta \in \Theta,\ \Theta \text{ acotado}\}$ lo son [Vaart, 1998, pp. 271,272].

Entonces, recapitulando, asumimos un modelo paramétrico y que si las funciones densidades que lo describen cumplen con ciertas condiciones, como las descriptas en él último teorema, podemos hallar un método consistente para estimar los parámetros. Incluso de la formulación se desprende algo de utilidad para aplicaciones prácticas. El M-estimador se construyen maximizando cierta función, pero en realidad lo que importa es que estén "cerca" de maximizar. Este hecho es de utilidad, pues a la hora de realizar simulaciones en la computadora se introducen siempre problemas de redondeo, sin embargo estos teoremas nos garantizan igual el comportamiento asintótico de los estimadores.

La pregunta que sigue es, ¿qué sucede si el modelo es incorrecto? El estimador sigue siendo consistente, pero ¿a qué converge entonces? Para responder esa pregunta vamos introducir otro concepto. La *divergencia de Kullback-Leibler*.

2.3. Divergencia de Kullback - Leibler

Como se mencionó al final de la sección anterior, ¿qué sucede cuando nos equivocamos en la propuesta del modelo? Es decir, ¿qué pasa cuando asumimos por verdadero un modelo incorrecto y estimamos el *parámetro del supuesto modelo* por máxima verosimilitud? La divergencia de Kullback-Leibler es una poderosa herramienta para dar respuesta a esta inquietud. Se basa en la esperanza matemática de la diferencia logarítmica de las funciones de densidad. Más precisamente:

Definición 2.3.1 (Divergencia de Kullback-Leibler)**.** Sean P, Q dos funciones de distribución de probabilidad, supongamos que $P \ll Q$ y $Q \ll P$, o sea que cada una es absolutamente continua con respecto a la otra, y que tienen densidades p, q:

$$D_{KL}(P||Q) := \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{2.6}$$

Que no es otra cosa que:

$$D_{KL}(P||Q) = \mathbb{E}_P\left(\log\left(\frac{p(X)}{q(X)}\right)\right),$$

donde la notación $\mathbb{E}_P$ explicita que la esperanza es con respecto a la distribución P: $X \sim P$.

Se pide $P \ll Q$ y $Q \ll P$, para evitar el caso en que exista algún conjunto E tal que $\int_E p(x)dx = 0$ y $\int_E q(x)dx \neq 0$ y viceversa. Cuando se está en ese caso las distribuciones son perfectamente identificables una respecto de otra [Kullback and Leibler, 1951]. Sin embargo se puede relajar un poco la condición pidiendo solamente $P \ll Q$ (ver [Kullback, 1968, ch. Properties of Information, section Information and Sufficiency]). También esta definición es valida si p y q son densidades respecto a una medida de referencia arbitraria ν ([Kullback and Leibler, 1951]), permitiendo así contemplar el caso discreto.

La divergencia-KL cumple con la siguiente propiedad:

Proposición 2.3.2. *$D_{KL}(P||Q) \geq 0$ y cumple la igualdad sii $P = Q$*

Demostración. Si p y q son densidades:

$$\begin{aligned} D_{KL}(P||Q) &= \mathbb{E}_P\left(\log\left(\frac{p}{q}\right)\right) \\ &= -\mathbb{E}_P\left(\log\left(\frac{q}{p}\right)\right) \\ &\underbrace{\geq}_{\text{desigualdad de Jensen}} -\log\left(\mathbb{E}_P\left(\frac{q}{p}\right)\right) \qquad (2.7) \\ &= -\log\left(\int_{-\infty}^{\infty} p(x)\frac{q(x)}{p(x)}dx\right) \\ &= -\log\left(\int_{-\infty}^{\infty} q(x)dx\right) \\ &\underbrace{=}_{\text{q es densidad}} 0 \end{aligned} \qquad (2.8)$$

Y la igualdad en 2.7 vale solamente cuando P = Q. □

Para una demostración con una medida de referencia arbitraria de esta propiedad, ver [Kullback and Leibler, 1951].

Las familias paramétricas que contemplamos son identificables, es decir, cada parámetro está asociado a una única distribución en el modelo. Tenemos entonces el siguiente resultado.

Corolario 2.3.3. *Si F_{θ_1} y F_{θ_2} son elementos de $\mathcal{M}$, o sea $F_{\theta_1} = F(\cdot, \theta_1)$ y $F_{\theta_2} = F(\cdot, \theta_2)$. Se tiene que $D_{KL}(F_{\theta_1}||F_{\theta_2}) = 0 \iff \theta_1 = \theta_2$*

Si bien la divergencia de Kullback Leibler no es una *métrica* se la utiliza igualmente como una medida de similitud entre dos distribuciones de probabilidad, donde en el caso de que sean iguales, la *divergencia KL* es 0.

Observación 2.3.4. De la definición de la divergencia-KL, (con $P \ll Q$ y $Q \ll P$), tenemos:

$$\begin{aligned} D_{KL}(P||Q) &= \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx = \int_{-\infty}^{\infty} p(x)\log(p(x))dx - \int_{-\infty}^{\infty} p(x)\log(q(x))dx \\ &= \mathbb{E}_P(\log(p(X))) - \mathbb{E}_P(\log(q(X))) \qquad (2.9) \end{aligned}$$

En algunos textos se puede encontrar referencias a la divergencia de Kullback-Leibler como entropía relativa en el contexto de teoría de la información. Para ver más propiedades se puede consultar [Kullback, 1968], y también [Brown, 1986, ch. The Dual to the Maximum Likelihood Estimator] en el cual profundiza en las propiedades que tiene esta divergencia dentro del conjunto de las familias exponenciales. Consideremos ahora un modelo $\mathcal{M} = \{F(\cdot, \theta) : \theta \in \Theta\}$ y una distribución F que no necesariamente pertenece al modelo $\mathcal{M}$. Vamos a asumir que existe $\theta \in \Theta$ tal que F es absolutamente continua respecto de F_θ. En tal caso, reemplazando en la ecuación (2.9), tenemos que:

$$D_{KL}(F||F(\cdot,\theta)) = \mathbb{E}_F(\log(f(X)) - \mathbb{E}_F(\log(f(X,\theta))$$

Esta expresión sugiere entonces una forma de elegir $\theta_0 = \theta_0(F)$, un parámetro que identifica a una distribución perteneciente a $\mathcal{M}$ más cercana a la distribución de interés: simplemente se trata de minimizar, en el caso que sea posible, la *divergencia KL*. En lo que sigue, asumiremos que el valor que minimiza es único. Notemos que, si $F \in \mathcal{M}$ y $F = F_{\theta^*}$, entonces $\theta_0(F) = \theta^*$, debido al Corolario 2.3.3.

2.4. EMV y divergencia KL

Ya tenemos todos las piezas para entender que pasa con el estimador de máxima verosimilitud, incluso cuando el modelo propuesto para la distribución F no la contiene. Aunque el modelo no sea el adecuado, el *EMV* tiende en cierta forma al parámetro que representa la distribución más "cercana" dentro del modelo a la distribución de interés. En caso de que el modelo sea correcto, efectivamente tiende al parámetro adecuado que también minimiza la *divergencia KL*. Esta es una manera de justificar la idea intuitiva que está detrás del *EMV*.

Sea $\mathcal{M} = \{F_\theta : \theta \in \Theta\}$ un modelo *regular*, o sea en condiciones donde el *EMV* funcione razonablemente bien, bien identificado y F una función de distribución. Definimos $m(x;\theta) = \log(f(x;\theta))$ donde $f(\cdot;\theta)$ denota a la función de densidad asociada a la distribución $F(\cdot;\theta)$, también definimos $M(\theta;F) = E_F(m(X;\theta))$

Consideremos los siguientes supuestos.

C1. $F \ll F_\theta$, $F_\theta \in \mathcal{M}$

C2. Existe un único $\theta_0 = \theta_0(F)$ tal que

$$D_{KL}(F||F_{\theta_0}) < D_{KL}(F||F_\theta)\ , \quad \text{para todo } \theta \in \Theta,\ \theta \neq \theta_0. \tag{2.10}$$

C3.

$$D_{KL}(F||F_{\theta_0}) \ < \inf_{\theta:\, d(\theta,\theta_0)\geq\epsilon} D_{KL}(F||F_\theta)\ . \tag{2.11}$$

El siguiente resultado establece que pasa con el estimador de máxima verosimilitud en modelos mal especificados.

Teorema 2.4.1. *Sea $(X_i)_{i\geq 1}$ una sucesión de variables aleatorias iid con distribución F y $\mathcal{M} =$ un modelo paramétrico regular identificable. Asumimos que valen las condiciones C.1, C.2, y C.3 y asumimos también que $\tilde{\theta}_n^{EMV}(\mathbf{X})$, estimador de máxima verosimilitud, esta bien definido: existe un valor que maximiza la función de verosimilitud presentada en* (2.1) *y es único. Se tiene que*

$$\tilde{\theta}_n^{EMV}(\mathbf{X}) \xrightarrow[n\to\infty]{} \theta_0(F) = \underset{\theta}{argmin}\ D_{KL}(F||F_\theta).$$

Es decir, el estimador de máxima verosimilitud converge al valor θ_0 cuya distribucón asociada F_{θ_0} está más cerca de *F, en el sentido de minimizar $D_{KL}(F||F_\theta)$.*

Demostración. Las condiciones pedidas permiten invocar los resultados presentados para los *M-estimadores*. Tenemos entonces que

$$\begin{aligned}
\tilde{\theta}_n^{EMV}(\mathbf{X}) &= \underset{\theta}{argmax}\left(\sum_{i=1}^{n}\log(f(x_i,\theta))\right)\\
&= \underset{\theta}{argmax}\left(\frac{\sum_{i=1}^{n}\log(f(x_i,\theta))}{n}\right)\\
&\xrightarrow[n\to\infty]{} \underset{\theta}{argmax}\,(\mathbb{E}_F(\log(f_\theta))\\
&= \underset{\theta}{argmin}\,(-\mathbb{E}_F(\log(f_\theta))\\
&= \underset{\theta}{argmin}\,(\mathbb{E}_F(\log(f) - \mathbb{E}_F(\log(f_\theta))\\
&= \underset{\theta}{argmin}\,(D_{KL}(F||F_\theta))
\end{aligned}$$

□

Observación 2.4.2. En particular, este resultado se puede verificar de manera directa para las familias de distribuciones exponenciales, donde $f_\theta(x,\theta) = \theta e^{-\theta x}$, para $x > 0$, y F una distribución con primer momento finito y positivo, y tal que $F \ll F_\theta$ y función de densidad f:

Tenemos por un lado que:

$$\begin{aligned} g(\theta) = D_{KL}(F||F_\theta) &= \int_{-\infty}^{\infty} f(x) \log\left(\frac{f(x)}{\theta e^{-\theta x}}\right) dx \\ &= \underbrace{\int_{-\infty}^{\infty} \log(f(x)) f(x) dx}_{C(F)} - \int_{-\infty}^{\infty} (\log(\theta) - \theta x) f(x) dx \\ &= C(F) - \log(\theta) \underbrace{\int_{-\infty}^{\infty} f(x) dx}_{=1} + \theta \underbrace{\int_{-\infty}^{\infty} x f(x) dx}_{=E_F(X)} \\ &= C(F) - \log(\theta) + \theta E_F(X) \end{aligned}$$

Derivando g:

$$g'(\theta) = -\frac{1}{\theta} + E_F(X)$$

Se tiene que $g'(\theta) = 0 \iff \theta = \frac{1}{E_F(x)}$.

Derivando una vez más:

$$g''(\theta) = \frac{1}{\theta^2} > 0 \quad \forall \theta$$

Luego resulta que $\theta = \frac{1}{E_F(x)}$ es un mínimo absoluto y único de $D_{KL}(F||F_\theta)$, por lo que $\theta_0(F) = E_F(X)^{-1}$. Por otro lado, cuando se proponen a las distribuciones exponenciales como modelo se tiene que

$$\tilde{\theta}_n^{EMV}(\mathbf{X}) = \frac{1}{\overline{X}_n}$$

y, por *LGN*, se verifica que converge a $\theta_0(F)$.

Vale la pena mencionar que esta relación entre el *EMV* y la *divergencia KL* se puede demostrar para más casos de las familias exponenciales, ver [Brown, 1986, ch. The Dual to the Maximum Likelihood Estimator].

Capítulo 3

El problema de los extremos

> There is always going to be an element of doubt as one is extrapolating into areas one doesn't know about. But what EVT is doing is making the best use of whatever data you have about extreme phenomena. [Siempre habrá un elemento de duda, ya que se extrapola a áreas que se desconocen. Pero lo que hace la teoría de valores extremos es aprovechar al máximo los datos de que se dispone sobre fenómenos extremos.]
>
> Richard Smith

3.1. Introducción

La teoría de valores extremos es una de las disciplinas de la estadística más importantes de los últimos 50 años [Coles, 2001, p. 50]. Con aplicaciones en rubros muy variados como la industria de seguros, la hidrología o la predicción de tráfico en telecomunicaciones.

En general busca responder preguntas y hacer inferencias en torno a procesos inusualmente grandes o pequeños. Más en concreto, en modelar probabilidades de eventos más extremos que los observados hasta el momento.

Consideremos una muestra aleatoria $X_1, X_2, \ldots, X_n$ y a su máximo $M_n = \text{máx}(X_1, X_2, \ldots, X_n)$. Uno quisiera conocer la distribución de probabilidades de M_n y poder tomar decisiones contando con esa base. Si se conociera las distribuciones de las X_i, el problema sería relativamente sencillo. Pero en general éstas no se conocen. Acá es donde cobran relevancia las ideas de la teoría de valores extremos.

Resulta que bajo condiciones razonables M_n tiene un comportamiento asintótico que se encuentra dentro de una familia paramétrica. Al llevar adelante estas ideas, se está asumiendo como válido un modelo que extrapola información como límite de una serie finita de eventos, ésto no tiene porque ser verdad, pero sin embargo hasta la fecha no hubo propuestas superadoras [Coles, 2001, ch. Introduction]. Esto es conocido como el *paradigma de valores extremos*.

Para entender mejor lo dicho hasta aquí desarrollemos un poco más con un ejemplo. Supongamos que X_i representa el nivel del mar en alguna ciudad costera. Luego extraigamos por cada año el valor máximo. De esta forma tenemos una nueva variable aleatoria M_n que representa el máximo

anual. Una pregunta de interés para planificar el desarrollo urbano consiste en estimar cual puede llegar a ser el máximo nivel del mar en los próximos 100 o 1000 años. Es en este tipo de preguntas que la *teoría de valores extremos* resulta de utilidad dando un marco de trabajo para aprovechar la información lo mejor posible, sin embargo asume que las condiciones que generaron los datos obtenidos se mantienen estables a lo largo del tiempo. Por ejemplo, debido al cambio climático, ¿es razonable suponer que el nivel del mar no se va a ver influenciado por éste? Es de esperar que sí sea influenciado, sin embargo, las conclusiones obtenidas a partir de los datos de nivel del mar recogidos hasta el momento, y luego utilizando la *teoría de valores extremos para variables iid* no van a dar cuenta de este cambio en las condiciones en las que suceden los hechos. Sin embargo, si se considera un período de tiempo más corto, sí es razonable esperar que las condiciones se mantienen constantes, y por lo tanto este marco de trabajo resulta de utilidad. Otra posibilidad es contemplar modelos alternativos, que contemplen la no estacionaridad del proceso.

3.2. Modelando la distribución del máximo

Según [Coles, 2001] entender el comportamiento del máximo de las muestras es fundamental en la teoría de valores extremos. Consideremos $X_1, X_2, \ldots, X_n$ una muestra *iid* con $X_i \sim F$. Luego $M_n = \text{máx}(X_1, X_2, \ldots, X_n)$ es también una variable aleatoria. Cuando F es conocida, se sabe que $P(M_n \leq z) = (F(z))^n$. Pero cuando F no es conocida, es donde se hace necesario modelar.

Una posibilidad es calcular empíricamente a F, sin embargo, no resulta estable al querer calcular F^n. Es por ello que se busca alguna familia paramétrica para modelar. Resulta que esto es posible:

Teorema 3.2.1. *Si existen sucesiones de constantes $a_n > 0$ y b_n tal que*

$$P\left(\frac{(M_n - b_n)}{a_n} < z\right) \xrightarrow[n\to\infty]{} G(z) \tag{3.1}$$

y G no es una distribución degenerada, entonces G pertenece a una de las siguientes familias:

1. *Extremos Tipo I: Gumbel*

$$G(z) = \exp\left(-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right) \qquad -\infty < z < \infty$$

2. *Extremos tipo II: Fréchet*

$$G(z) = \begin{cases} 0 & z \leq b \\ \exp\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right] & z > b \end{cases}$$

3. *Extremos tipo III: Weibull*

$$G(z) = \begin{cases} \exp\left(-\left[-\left(\frac{z-b}{a}\right)\right]^{\alpha}\right); & z < b \\ 1 \qquad z \geq b \end{cases}$$

Con $a > 0$, y $\alpha > 0$ en los tipos II y III.

Este teorema implica que cuando al máximo muestral se lo puede estabilizar mediante un cambio de escala (o sea que posee una distribución límite no degenerada), entonces su distribución límite es solo uno de estos tres casos. Éste resulta ser un resultado análogo al más conocido Teorema Central del Límite. Para una demostración del teorema consultar [Leadbetter et al., 1983].

A pesar de ser un resultado muy útil tiene el inconveniente de que para su uso práctico primero hay que elegir una de las tres familias y luego con la muestra estimar los parámetros correspondientes. Una consecuencia es que la elección de la familia puede dar lugar a resultados muy distintos. Es por eso que existe un enfoque distinto que consiste en agregar un parámetro más permitiendo así unificar las tres familias en una una única dando origen a la *familia generalizada de valores extremos*.

$$G(z) = \exp\left\{-\left[1+\epsilon\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\epsilon}}\right\} \tag{3.2}$$

Donde el caso $\epsilon = 0$ se interpreta como $\epsilon \to 0$. O sea, cuando $\epsilon < 0$ volvemos a una distribución de tipo Weibull, con $\epsilon > 0$ es una distribución de tipo Fréchet, y el caso $\epsilon = 0$ es el caso de una distribución de tipo Gumbel. Luego, con este enfoque no es necesario adoptar a priori el tipo de distribución que modela el máximo, sino también estimarlo mediante las muestras.

Sin embargo, este enfoque general tiene un punto negativo importante. Es verdad que permite modelar valores extremos de una distribución desconocida, pero lo hace a partir de estimar una distribución para el comportamiento del máximo muestral. O sea, requiere que dada una muestra de datos, primero dividirla en bloques, y luego quedarse con un único valor de cada bloque (el máximo), esto tiene la consecuencia de ser un método poco robusto, ya que depende de la partición de los datos seleccionados, y además de cada partición solo considera un único valor, perdiendo un montón de información. A continuación veremos un enfoque alternativo a la estimación de valores extremos que parcialmente da cuenta de esta problemática, éste se basa en caracterizar los excesos que ocurren por encima de un umbral.

3.3. Excesos por encima de un umbral

La teoría de valores extremos no se encarga solo de estudiar la distribución del máximo, otro tema de interés es la distribución de *excesos por encima de un umbral*, de hecho ambos temas están relacionados como se puede ver en el Teorema 3.3.2 (para ver con más detalle se recomienda [Reiss and Thomas, 2007]). Su estudio no tiene solo interés matemático, sino que puede ser útil al modelar fenómenos cuyos datos solo importen o se registren a partir de cierto valor. Incluso a la hora de hacer inferencias permite realizar un mayor aprovechamiento de los datos ya que utiliza una mayor proporción de datos de la muestra a la hora de ajustar modelos.

En primer lugar, definamos a los *excesos* respecto de un umbral. Consideremos $X_1, X_2, \ldots, X_n$ una muestra aleatoria *iid* con $X_i \sim F$, los excesos respecto a un umbral u son los $X_i - u$ con $X_i > u$. Se define la distribución de excesos respecto de un umbral de la siguiente manera:

Definición 3.3.1 (Distribución de excesos respecto a un umbral)**.** Sea X variable aleatoria tal que $X \sim F$, la distribución de excesos respecto el umbral u se define como:

$$F_u(y) = P(X - u \leq y | X > u),\ y > 0$$

Si F fuera conocida, se sabe que:

$$P(X > u + y | X > u) = \frac{1 - F(u+y)}{1 - F(u)}, \qquad y > 0 \tag{3.3}$$

Sin embargo, F no suele ser conocida, pero existe un resultado análogo al visto en la sección anterior.

Teorema 3.3.2 (Pickands, 1975)**.** *Sea $X_1, X_2, \dots X_n$ una muestra aleatoria iid con $X_i \sim F$. Consideremos $M_n = \max(X_1, X_2, \dots, X_n)$. Supongamos que F verifique* (3.1)*. Es decir, satisface las condiciones requeridas para que la familia de distribuciones de valores extremos generalizada aproxima a M_n. O sea*

$$P(M_n < z) \approx G(z)$$

con

$$G(z) = \exp\left\{ -\left[1 + \epsilon\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\epsilon}} \right\}$$

Entonces para u suficientemente grande, la distribución de $(X-u)$, condicionado a $X > u$, es aproximadamente

$$H_{\sigma,\epsilon}(y) = 1 - \left(1 + \frac{\epsilon y}{\sigma}\right)^{-\frac{1}{\epsilon}} \tag{3.4}$$

definida en $\{y : y > 0 \ \& \ (1 + \frac{\epsilon y}{\sigma}) > 0\}$ Esta familia de distribuciones se conoce como familia de Pareto Generalizada (FPG)

Más precisamente, se tiene que

$$\lim_{u \to u^*} \sup_{0 \le y < \infty} |F_u(y) - H_{\sigma,\epsilon}(y)| = 0,$$

donde u^* denota su borde derecho, o sea: $u^* := sup\{u : F(u) < 1\}$

Una demostración de este teorema se puede ver en [Pickands, 1975].

Notar que los parámetros de la familiade Pareto Generalizada están determinados unívocamente por los de la *familia generalizada de valores extremos*, en particular el parámetro ϵ es el mismo para ambos.

Observación 3.3.3. La distribución exponencial pertenece a la familia de Pareto Generalizada, y la distribución de excesos asociada con cualquier umbral también.

Demostración 3.3.4. En primer lugar tenemos que

$$X \sim \mathcal{E}(\lambda) \iff P(X \le x) = F(x) = 1 - e^{-\lambda x}$$

Consideremos ahora

$$F_u(y) = P(X - u \le y | X > u)$$

Luego

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

Finalmente

$$F_u(y) = 1 - \frac{1 - F(u+y)}{1 - F(u)} = 1 - \frac{e^{-\lambda(u+y)}}{e^{-\lambda u}} = 1 - e^{-\lambda y}, \qquad y > 0$$

para cualquier umbral u

Observación 3.3.5. Notar que en el caso de la observación anterior no importa el umbral, la distribución de excesos sigue siendo exponencial con el mismo parámetro λ.

Una propiedad interesante y que le confiere a este enfoque utilidad, es un cierto tipo de estabilidad en el siguiente sentido:

Proposición 3.3.6. *Si $Y \sim F \in$ Familia de Pareto Generalizada y $u > 0$ entonces*

$$(Y - u | Y > u) \sim F^* \in \textit{Familia de Pareto Generalizada}$$

Y además vale que si para algún u_0 que $F_{u_0}(y) = H_{\sigma_0,\epsilon}(y)$, entonces para $u > u_0$ se tiene que $F_u(y) = H_{\tilde{\sigma},\epsilon}(y)$ con $\tilde{\sigma} = \sigma(u) = \sigma_0 + \epsilon(u - u_o)$

3.4. Aplicación

El Teorema 3.3.2 sugiere entonces una manera de modelar los excesos de una distribución. Teniendo una serie de datos $x_1, x_2, \ldots, x_n$ a los cuales identificar con realizaciones de una muestra aleatoria *iid*, se puede considerar como eventos extremos por encima de un umbral u a los $x_i : x_i > u$. Renombrando a los elementos de este conjunto de excesos como $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$, podemos definir $y_j = x_{(j)} - u$, con $j = 1, \ldots, k$ como realizaciones independientes de una variable aleatoria cuya distribución se puede aproximar por un elemento de la *Familia de Pareto Generalizada*. A partir de esto es posible entonces con los métodos tradicionales de inferencia estadística, encontrar los parámetros que mejor se ajusten a la muestra. Sin embargo, este método es sensible al valor del umbral elegido. En la bibliografía se pueden encontrar distintas técnicas para determinarlo, entre ellas técnicas gráficas que requieren de la habilidad del investigador, o técnicas que requieren asumir hipótesis sobre los datos por debajo del umbral (ver [Coles, 2001, ch. Threshold Models], [Cabras and Morales, 2007], [Wong and Li, 2010], [Gonzalez et al., 2013] o [MacDonald et al., 2011]). Este problema es el tema central de esta tesis y se propone explorar un algoritmo para estimarlo sin que asuma hipótesis extra de los datos, ni que requiera de la experiencia o habilidad del investigador. En el siguiente capítulo abordaremos esta propuesta.

Capítulo 4

Distribución de los excesos. Una propuesta de detección de umbrales

> Pero el papel de la estadística no es tanto resumir lo que ya ha ocurrido, sino inferir las características de aleatoriedad en el proceso que generó los datos.
>
> Coles

Al final del capítulo anterior mencionamos el enfoque de umbrales y el hecho de que aprovecha mejor los datos de la muestra frente a solo buscar modelar la distribución del máximo, sin embargo, la calidad del ajuste depende de la selección del umbral. Por ejemplo, en aplicaciones prácticas a la hora de estimar los parámetros correspondientes, si se elige un umbral chico, el modelo es malo, pero si el umbral es grande quedan pocos datos para ajustarlo.

Es por lo tanto útil el estudio de propuestas para abordar este problema. En la presente tesis proponemos explorar una manera ,partiendo de un modelo semiparamétrico, de estimar este umbral.

4.1. Un modelo semiparamétrico

4.1.1. Definiciones

Si bien los resultados del capítulo anterior proponen un modelo paramétrico (la FPG) para los excesos respecto un umbral, para valores de u suficientemente grandes, en este trabajo vamos a asumir que la distribución de excesos pertenece a la FPG a partir de cierto umbral. Para ser más especificos, recordemos la definición 3.3.1, en la cual definimos la distribución de excesos como:

$$F_u(y) = P(X - u \leq y | X > u),\ y > 0$$

Consideremos ahora el conjunto de las distribuciones donde existe un valor u al partir del cual la distribución de excesos es un elemento de la FPG, o sea:

$$\mathcal{M} = \{F \text{ distribución } : \exists u \in \mathbb{R} \text{ tal que } F_u = G,\ \ G \in FPG\}$$

Notemos que la propiedad 3.3.6 implica que si $F_{u_0} \in FPG$, entonces $F_u \in FPG\ \forall u > u_0$. De acá se puede definir de manera precisa el umbral de interés:

Definición 4.1.1. Sea $F \in \mathcal{M}$, consideremos

$$\Gamma_F = \{u \ : F_u = G,\ G \in \ FPG\}$$

Definimos como el *umbral* para F a

$$u(F) = \text{ínf}\, \Gamma_F$$

En [Gonzalez et al., 2013], se demuestra que en realidad este ínfimo es un mínimo. No vamos a realizar ninguna hipótesis sobre la distribución por debajo del umbral, es en este sentido que hablamos que nuestra propuesta se basa en un modelo semiparamétrico.

4.1.2. El modelo

En el presente trabajo vamos a considerar solo el subconjunto de $\mathcal{M}$ compuesto por las distribuciones para las cuales existe un *umbral* al partir del cual su distribución de excesos es exponencial. O sea:

$$\mathcal{M}_{\mathcal{E}} = \{F \text{ distribución } : \exists u \in \mathbb{R} \text{ tal que } F_u \sim \mathcal{E}(\lambda)\}$$

Recordemos que según la observación 3.3.3, la distribución exponencial pertenece a la *FPG* y también mantiene el criterio de estabilidad, en el sentido de que si vale para un valor u también vale para los mayores.

Caracterización alternativa del umbral

A continuación vamos a dar una caracterización del *umbral* para poder determinarlo con herramientas provenientes de la estadística.

Definición 4.1.2. Sea X v.a. con $X \sim F$, $F \in \mathcal{M}_{\mathcal{E}}$ y $u \in \mathbb{R}_+$:

$$\lambda_F(u) := \frac{1}{\mathbb{E}_F(X - u | X > u)}$$

Notar que si $F_u \sim \mathcal{E}(\lambda)$, $\lambda_F(u)$ es el parámetro de la exponencial según la cual se distribuye F_u.

Proposición 4.1.3. *Sea u_0 el* umbral *de F, o sea $u(F) = u_0$, y λ_0 tal que $F_{u_0} \sim \mathcal{E}(\lambda_0)$ Entonces:*

$$\lambda_F(u) = \lambda_0 \quad \forall u \geq u_0$$

Demostración 4.1.4. Es una consecuencia directa de las observaciones 3.3.3 y 3.3.5.

Definición 4.1.5. Definimos la siguiente función de pérdida

$$l_F(u) = \sup_{y \geq 0} |F_u(y) - F_{\lambda_F(u)}(y)|,$$

Esta función no es otra cosa que la distancia en norma supremo que hay entre la distribución de excesos por encima del valor u y el modelo, en el sentido de elegir la distribución exponencial que mejor la aproxima en términos de divergencia de Kullback Leibler, a la distribución de excesos F_u.

De la definición anterior se desprende directamente que:

$$l_F(u) \begin{cases} = 0 & u \geq u_0 \\ > 0 & u < u_0 \end{cases}$$

Finalmente obtenemos que

$$u(F) = \text{mín}(\{u : l_F(u) = 0\}) \tag{4.1}$$

O sea el primer u que minimice la función de pérdida.

4.2. Estimando el umbral

Con esta caracterización es posible plantear un camino para estimar el umbral a partir de muestras aleatorias considerando las versiones empíricas de las funciones y parámetros mencionados al final de la sección anterior. Consideremos una muestra aleatoria $X_1, X_2, \ldots, X_n$ *iid* con $X_i \sim F \in \mathcal{M}_{\mathcal{E}}$

Definición 4.2.1. Dada una muestra aleatoria $X_1, X_2, \ldots, X_n$ *iid* con $X_i \sim F \in \mathcal{M}_{\mathcal{E}}$, consideremos:

$$\widehat{\lambda}(u) = \frac{\sum_{i=1}^{n} \mathbf{I}_{X_i > u}}{\sum_{i=1}^{n} (X_i - u)\mathbf{I}_{X_i > u}}$$

$$\widehat{F}_u(y) = \frac{\sum_{i=1}^{n} \mathbf{I}_{X_i - u \leq y} \mathbf{I}_{X_i > u}}{\sum_{i=1}^{n} \mathbf{I}_{X_i > u}}$$

$$\widehat{l}(u) = \sup_{y \geq 0} |\widehat{F}_u(y) - F_{\widehat{\lambda}(u)}|$$

Si bien el siguiente paso sería minimizar la versión empírica de la función de pérdida, está el problema de que según el modelo teórico, todo valor u mayor al *umbral* verdadero la va a minimizar, es por lo tanto necesario penalizar los valores más grandes. Definimos por ello:

Definición 4.2.2.

$$PN(u) = \hat{l}(u) + c_n u$$

y a la estimación del *umbral*:

$$\hat{u} = argmin\, PN(u)$$

Esta función con penalización, está basada en [Boente et al., 2023].

Finalmente, encontrando $\hat{u}$, es posible también modelar la distribución de excesos:

Definición 4.2.3.

$$\tilde{F}_u := F_{\widehat{\lambda}(\widehat{u})}$$

Capítulo 5

Simulación

> The key message is that extreme value theory cannot do magic but it can do a whole lot better than empirical curve fitting and guesswork. [El mensaje principal es que si bien la teoría de valores extremos no puede hacer magia, sí puede hacerlo mejor que el ajuste empírico y las conjeturas.]
>
> Jonathan Tawn

En este capítulo describiremos el estudio de simulación realizado para explorar empíricamente el desempeño de nuestra propuesta de estimación.

5.1. El modelo uniforme-exponencial

Se realizaron una serie de simulaciones en donde los datos tenían una función de densidad dada por:

$$f(x) = p_{umbral}\mathbf{I}_{[0\leq x\leq 1]} + (1 - p_{umbral})\lambda e^{-\lambda(x-1)}\mathbf{I}_{[x>1]} \tag{5.1}$$

donde p_{umbral} y λ son parámetros que se fueron variando para probar en distintos casos. Esta densidad se entiende como datos menores a 1 siguen una distribución uniforme $[0, 1]$ y datos mayores una distribución exponencial de parámetros λ desplazada hacia la derecha en 1. p_{umbral} representa la proporción de datos que siguen la distribución uniforme. En este caso el *umbral* poblacional u_0 es igual a 1.

De esta manera para $x = 1$ tenemos un cambio de régimen, en donde la distribución de los excesos por encima siguen una distribución exponencial, la cual como vimos anteriormente pertenece a la *Familia de Pareto Generalizada*, y si la proporción de datos por encima de este umbral es adecuada, estamos ante un caso en donde es posible modelar a estos últimos según la teoría excesos descripta en el capítulo anterior. Nuestro objetivo entonces es poder detectar el umbral a partir del cual el modelo se ajusta bien.

Observación 5.1.1. Notar que, en general, la función de densidad $f(\cdot)$ presentada en (5.1) no es una función continua. Sin embargo, cuando

$$\lambda = \frac{p_{umbral}}{1 - p_{umbral}}$$

sí lo es.

5.2. Procedimiento

La simulación se realizó enteramente en el software R, los scripts utilizados para generar las muestras y calcular la estimación de u_0 se encuentran en el apéndice A.

5.2.1. Generación de la muestra

Una variable aleatoria con la densidad descripta por 5.1 se puede generar mediante el siguiente procedimiento: simular una variable aleatoria $Y \sim Be(p_{umbral})$, si $Y = 1$ entonces tomo un elemento generado por una uniforme $(0, 1)$, en cambio, si $Y = 0$ tomo un elemento generado por una exponencial de parámetro λ trasladada sumando el valor 1. Repitiendo este procedimiento n veces obtenemos una realización de una muestra de tamaño n. En adelante, utilizaremos muestra o datos de manera indistinta.

5.2.2. Procesamiento

A cada muestra se le realizaron los siguientes pasos:

1. Proponer un intervalo candidato que contenga al *umbral* y que garantice que para cualquier candidato a *umbral* dentro de ese intervalo, exista un dato por encima de él.
2. Repetir para cada valor u de la muestra dentro del intervalo definido en el punto anterior:
   - *a*) Quedarse con los excesos respecto al u correspondientes de esa muestra.
   - *b*) Proponer que la distribución de esos excesos es exponencial.
   - *c*) Estimar el parámetro λ mediante máxima verosimilitud.
   - *d*) Calcular distribución acumulada empírica de los excesos.
   - *e*) Calcular la distancia entre la exponencial estimada con el parámetro del punto 3 y la distribución acumulada empírica de los excesos del punto 4. Utilizar distancia del supremo.
3. Quedarse con el valor u que minimice dicha distancia, considerando una penalización ya que se quiere el valor umbral más chico a partir del cual el modelo ajuste bien, esto implica que la distancia se puede minimizar con un valor u mayor al umbral verdadero y por eso la necesidad de la penalización. El u hallado será la estimación de u_0 y será llamado $\hat{u}$.

5.3. Análisis

Se probó este algoritmo en diferentes escenarios considerando los siguientes parámetros:

- Tamaño de la muestra n: 500, 1000, 1500, 2000, 5000, 10000
- Proporción de datos por debajo del umbral p_{umbral}: 0.8, 0.95
- Parámetro λ de la exponencial: 0.5, 1, 10, 30, 50, y el λ que hace continua la densidad (5.1) ($\lambda = 4$ y $\lambda = 19$ cuando $p_{umbral} = 0{,}8$ y $p_{umbral} = 0{,}95$ respectivamente).
- Replicaciones $Nrep$: 1000
- Penalización: $0{,}8n^{-0,4}u$

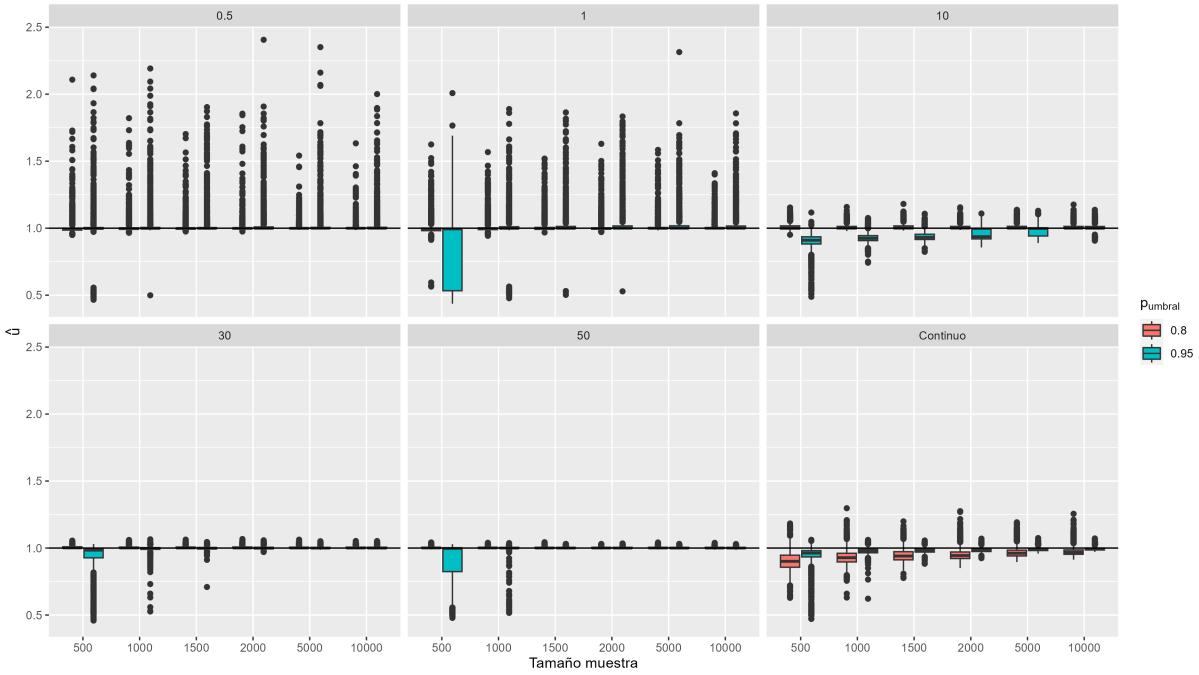

Figura 5.1: Estimación del umbral para diferentes n y p_{umbral} y λ. Cada cuadro representa un λ distinto. La línea negra sólida representa u_0.

La simulación se realizó repitiendo $Nrep = 1000$ veces la experiencia de tomar una muestra de tamaño n, definidos previamente, siguiendo la distribución dada por la ecuación (5.1). Por cada *tamaño de muestra* n y cada *proporción de datos por debajo del umbral* p_{umbral} se varió también el parámetro λ. Combinando los valores contemplados para λ y para p_{umbral}, quedan determinados 6×2 distribuciones para generar datos.

5.3.1. Calidad general del estimador del u_0

En la figura 5.1 se pueden apreciar los resultados obtenidos al estimar u_0. En todos los casos se observan la presencia de valores atípicos. Estos valores atípicos en los casos en que $\lambda < 10$ se ubican principalmente por encima del umbral verdadero. Pero en λ mayores se los ve más cercanos al umbral verdadero y en los casos en que $n = 500$ o $n = 1000$ mayormente se ubican por debajo del valor del umbral. También se observa que para el caso en λ continuo, las estimaciones realizadas con $p_{umbral} = 0{,}95$ se encuentran más cerca y con menos dispersión respecto de u_0.

Para evaluar entonces el rendimiento del estimador se calculó la mediana empírica del valor absoluto del error (EMAD, por sus siglas en inglés *empirical median absolute deviation*) definido como:

$$EMAD = mediana\,|\hat{u} - u_0|\,,$$

donde $\hat{u}$ es el estimador del *umbral* y u_0 es el *umbral* poblacional del modelo.

Observando la figura 5.2 se observa como el EMAD disminuye al aumentar el n, se observa un comportamiento anómalo con $p_{umbral} = 0{,}95$ y $\lambda = 10$ para tamaños de muestra $n = 500, 1000, 1500, 2000$ el cual requiere más estudio para analizar el motivo. Distinto es el caso con el caso en que λ hace continua a la densidad del modelo, es esperable que al ser continuo el cambio de régimen entre la parte uniforme y la exponencial cueste más identificar un salto en la función de pérdida.

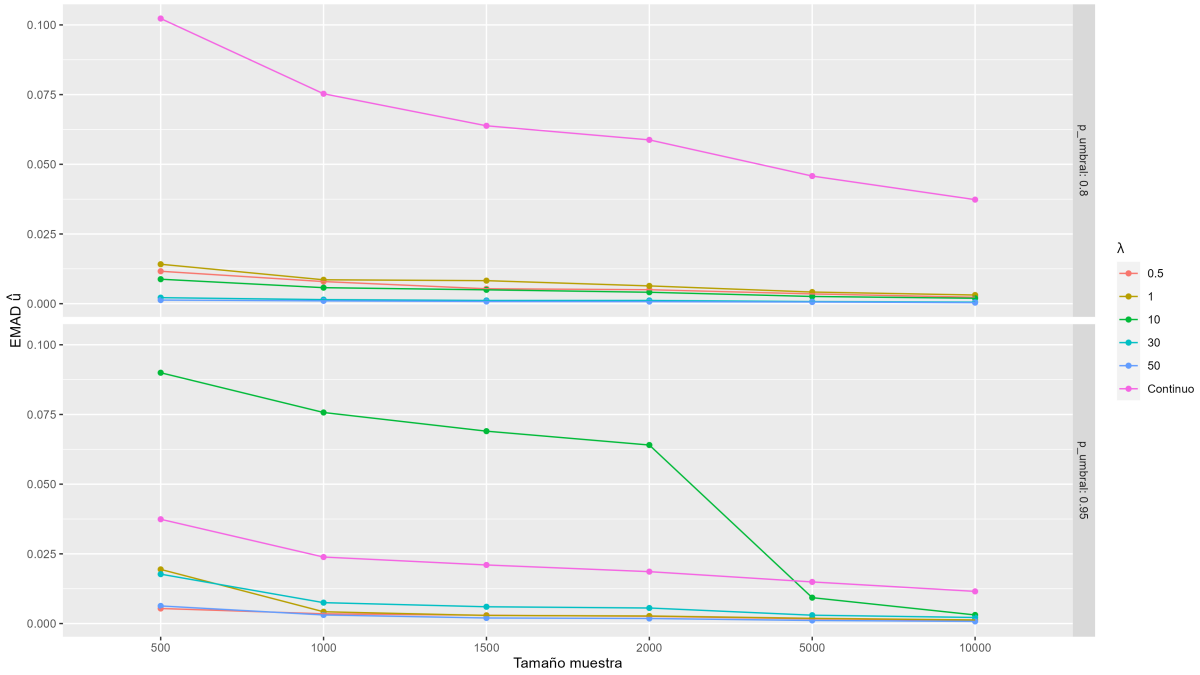

Figura 5.2: Comportamiento del EMAD en función del n para diferentes λ y p_{umbral}. Las líneas se representan para mayor claridad al seguir la evolución de la medida del error según cada escenario, pero solo se tomaron medidas en los $n = 500, 1000, 1500, 2000, 5000, 10000$.

5.3.2. Calidad del modelo estimado a partir del umbral

Es importante recordar que el objetivo es modelar la distribución de excesos, por lo que además de encontrar el umbral adecuado, es necesario ver si se puede tener una "buena" estimación del modelo en general.

Estimación de λ

En el gráfico 5.3 se muestra el cociente entre el lambda estimado a partir de los excesos resultantes luego de estimar u_0 y el λ poblacional que dio origen a los datos, en función del n, para los distintos p_{umbral}. Se observa que cuando $p_{umbral} = 0.95$ la dispersión es mayor, y en el caso de un tamaño de muestra de 500 esa dispersión es visiblemente mayor. Sin embargo en general la estimación resultante pareciera ser buena y mejora con el n. También se observa un mayor sesgo cuando el λ poblacional es 10 y $p_{umbral} = 0{,}95$, donde la mayoría de los $\hat{\lambda}$ estimados es mayor al λ poblacional.

Para aportar más claridad a este punto se puede utilizar el gráfico 5.4, que compara las medianas de $\frac{\hat{\lambda}}{\lambda}$. En donde se aprecia como mejora la estimación con el tamaño de muestra. Resulta curioso observar lo que sucede con $\lambda = 10$, donde se muestra un comportamiento distinto según si p_{umbral} 0,8 o 0,95. Y a la vez distinto al resto. En casi todos los casos con $p_{umbral} = 0{,}95$, la estimación mejora abruptamente al cambiar el tamaño de muestra de 500 a 1000.

Estimación de percentiles

Otro aspecto que da información sobre la calidad de las estimaciones, es la comparación que surge al calcular los cuantiles a partir de la muestra bajo el supuesto del modelo propuesto vs los

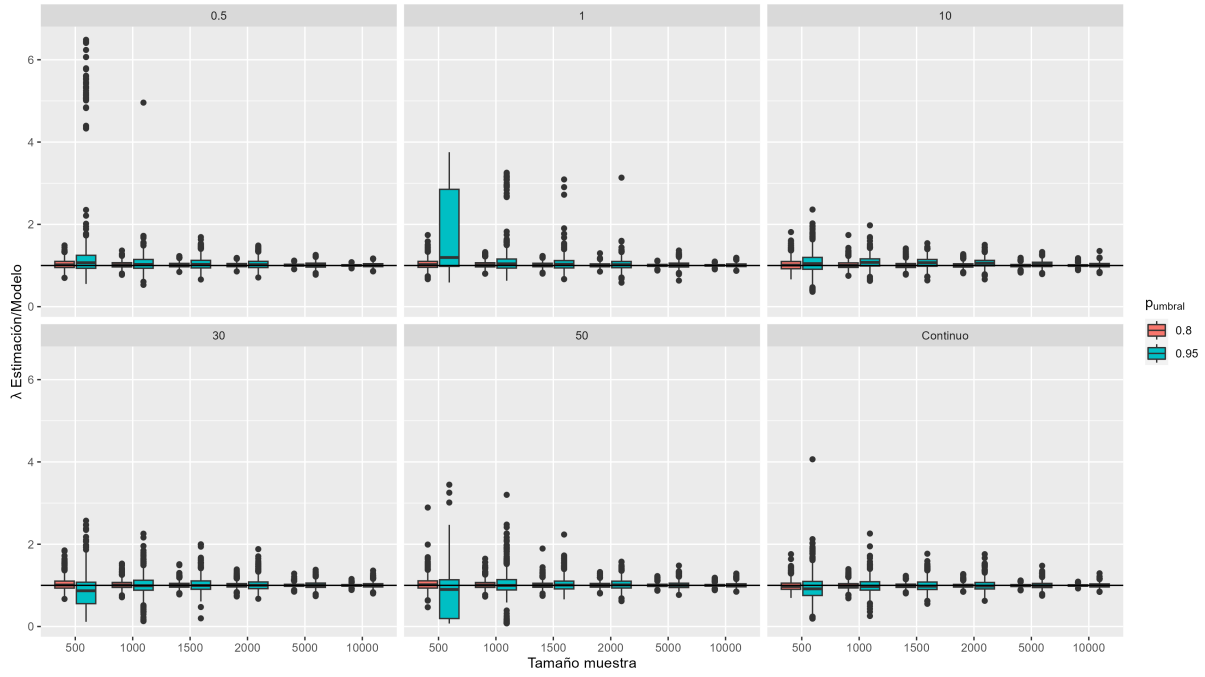

Figura 5.3: Relación $\frac{\hat{\lambda}}{\lambda}$ para diferentes n y p_{umbral} y λ. Cada cuadro representa un λ distinto. La línea horizontal representa el valor 1 que es cuando la estimación coincide con el parámetro poblacional.

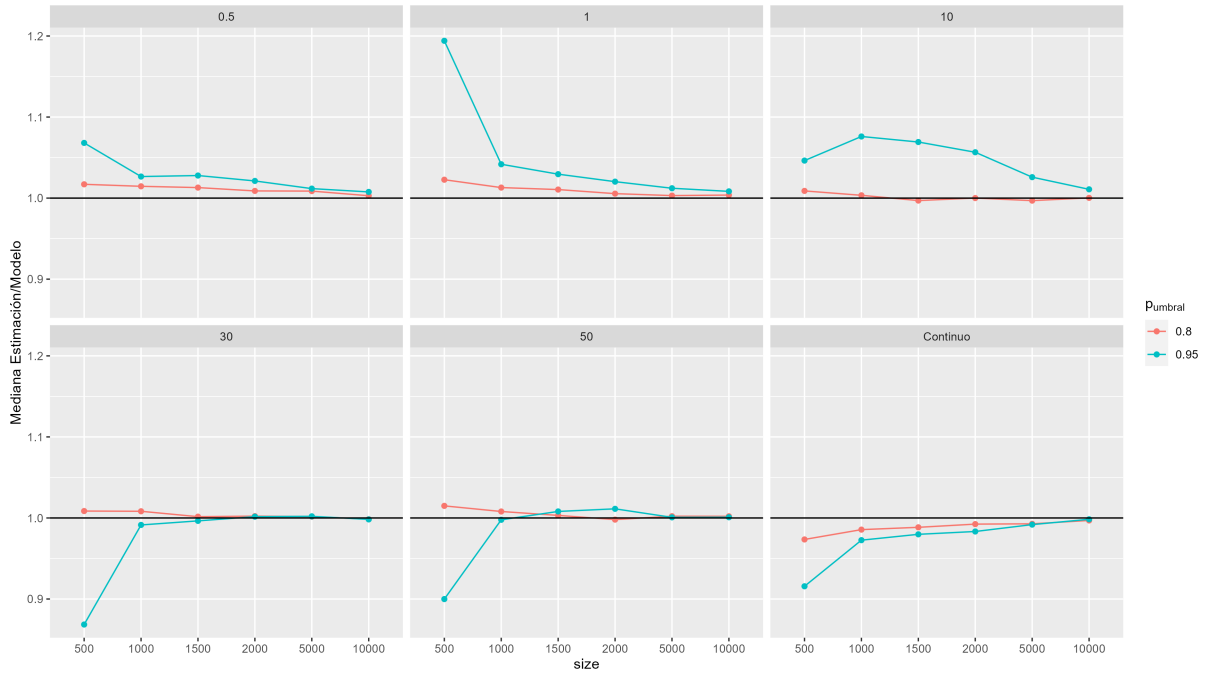

Figura 5.4: Mediana de la razón $\frac{\hat{\lambda}}{\lambda}$ vs el tamaño de muestra, discriminado según la proporción de datos debajo del umbral verdadero. Cada cuadro muestra el gráfico para un λ distinto. La línea horizontal representa el valor 1 que es cuando la estimación coincide con el parámetro poblacional.

obtenidos de manera no paramétrica, y ver cuales estiman mejor a los cuantiles poblacionales.

Notemos que bajo el modelo propuesto, la función de distribución acumulada está dada por

$$F(x) = \begin{cases} p_{umbral} F_1(x) & x \leq u_0 \\ p_{umbral} + (1 - p_{umbral})(1 - e^{-\lambda(x-u_0)}) & x > u_0 \end{cases}$$

En tal caso, para todo valor $p > p_{umbral}$ tenemos que el p-ésimo cuantil $F^{-1}(p)$ está dado por

$$F^{-1}(p) = \frac{-1}{\lambda} \log\left(1 - \frac{p - p_{umbral}}{1 - p_{umbral}}\right) + u_0 = \frac{-1}{\lambda} \log\left(\frac{1-p}{1 - p_{umbral}}\right) + u_0$$

Esto sugiere un procedimiento de tipo plug-in para su estimación, en la medida que $\hat{p}_{umbral}$ sea menor a p. Es decir, podemos estimar el $p_{cuantil}$ reemplazando p_{umbral}, u_0 y λ por sus estimadores.

Se consideraron los cuantiles 0.95, 0.99, 0.9999, 0.999999 y debido a las diferencias entre las magnitudes de los mismos, se compararon los cocientes $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$. Cuando el cuantil estimado coincide con el poblacional el cociente es 1, mientras que si da mayor, quiere decir que el cuantil está sobreestimado y si da menor está subestimado. Como ya se vio en los análisis anteriores la presencia de valores atípicos influyen en el promedio aritmético, para resumir los valores se utilizó la mediana.

En las figuras 5.5, 5.6, 5.7, 5.8, 5.9, 5.10 se compararon las estimaciones de los cuantiles paramétricos en función del tamaño de muestra y p_{umbral}. En todos los casos se observa que para los $p_{cuantil} = 0{,}95$ y $0{,}99$ los resultados son muy parecidos, sin embargo en los cuantiles más extremos se aprecia una mejor estimación mediante el modelo. Esto es importante, pues mejorar la estimación de los cuantiles más extremos es justamente un objetivo de interés.

Mientras que de manera no paramétrica no se pueden estimar probabilidades para valores mayores al máximo de la muestras, tener un modelo que se ajuste bien sí lo permite. En otras palabras, estos modelos de valores extremos permite asignar probabilidades a valores que estén por fuera del rango de las muestras observadas, permiten extrapolar.

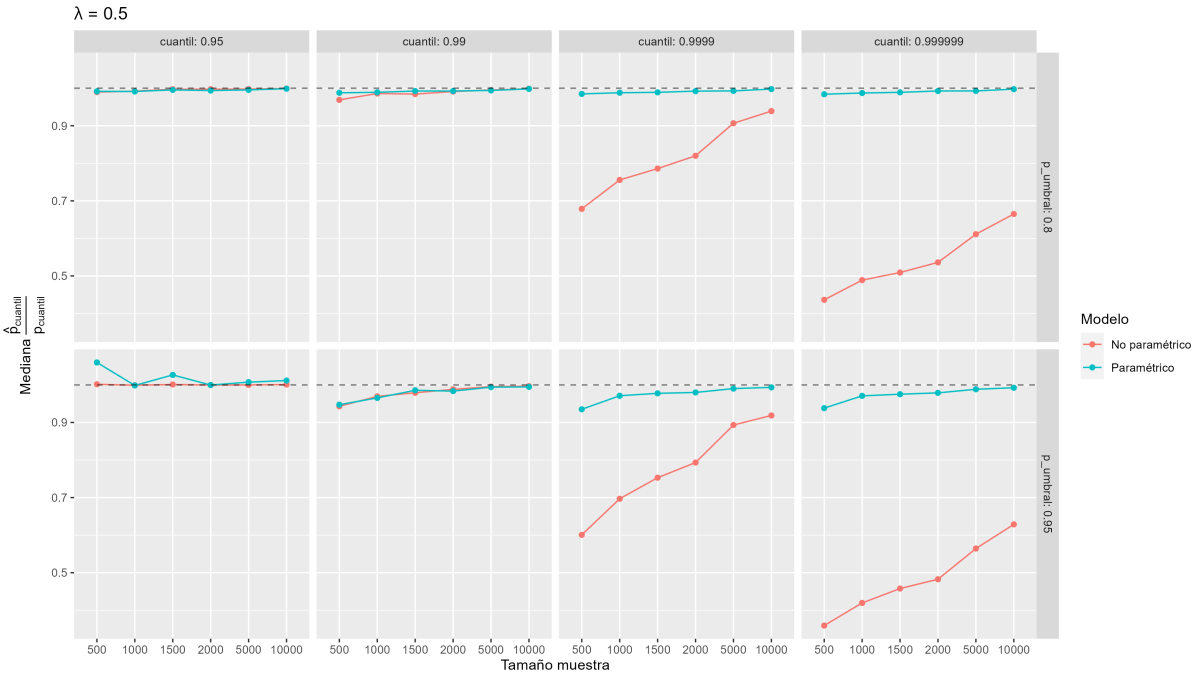

Figura 5.5: Mediana de la razón $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$ vs n con $\lambda = 0{,}5$, discriminado según p_{umbral} y $p_{cuantil}$. La línea horizontal punteada representa el valor 1 que es cuando la estimación coincide con el parámetro verdadero.

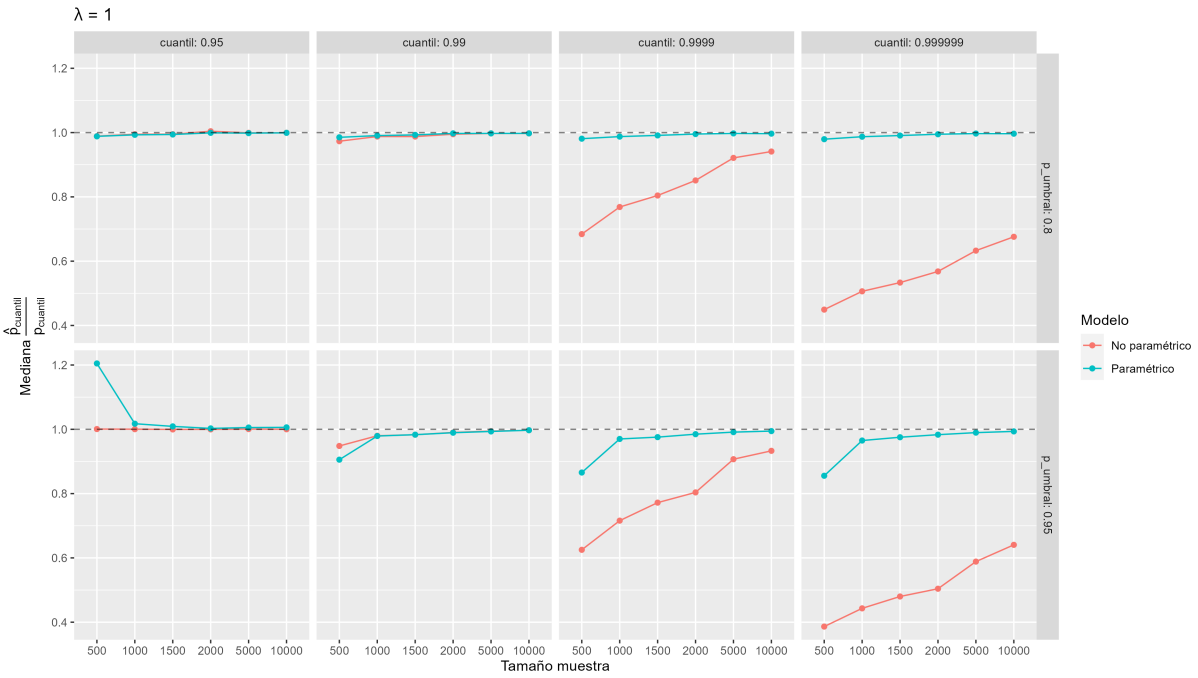

Figura 5.6: Mediana de la razón $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$ vs n con $\lambda = 1$, discriminado según p_{umbral} y $p_{cuantil}$. La línea horizontal punteada representa el valor 1 que es cuando la estimación coincide con el parámetro verdadero.

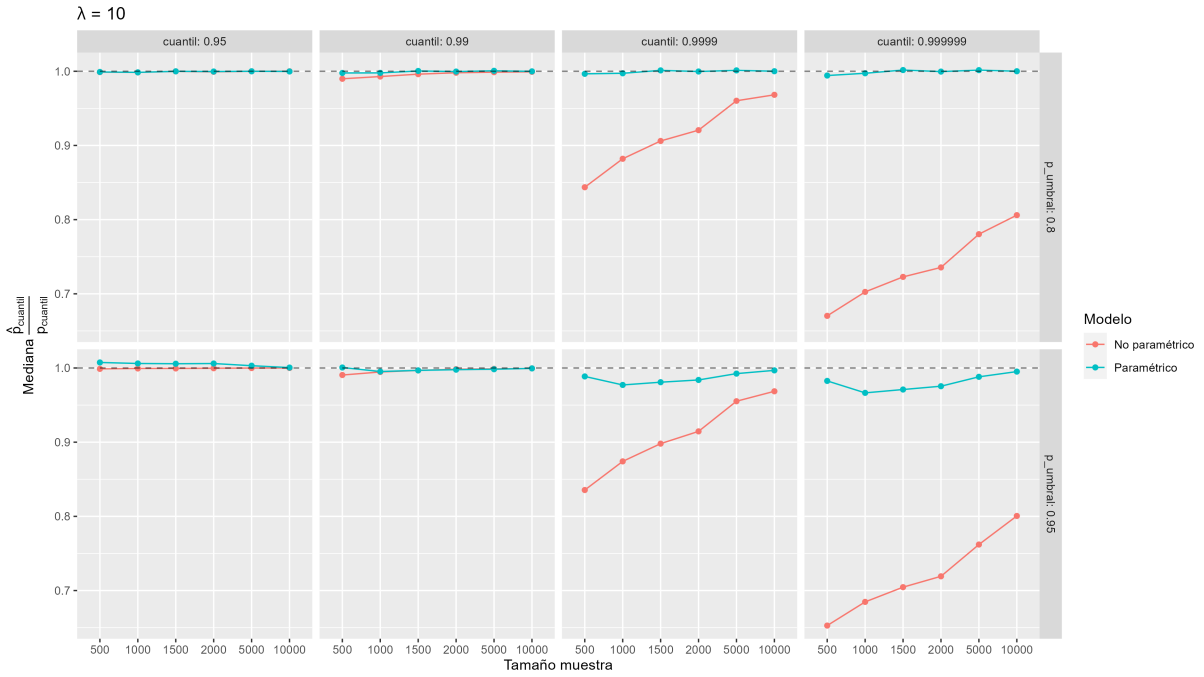

Figura 5.7: Mediana de la razón $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$ vs n con $\lambda = 10$, discriminado según p_{umbral} y $p_{cuantil}$. La línea horizontal punteada representa el valor 1 que es cuando la estimación coincide con el parámetro verdadero.

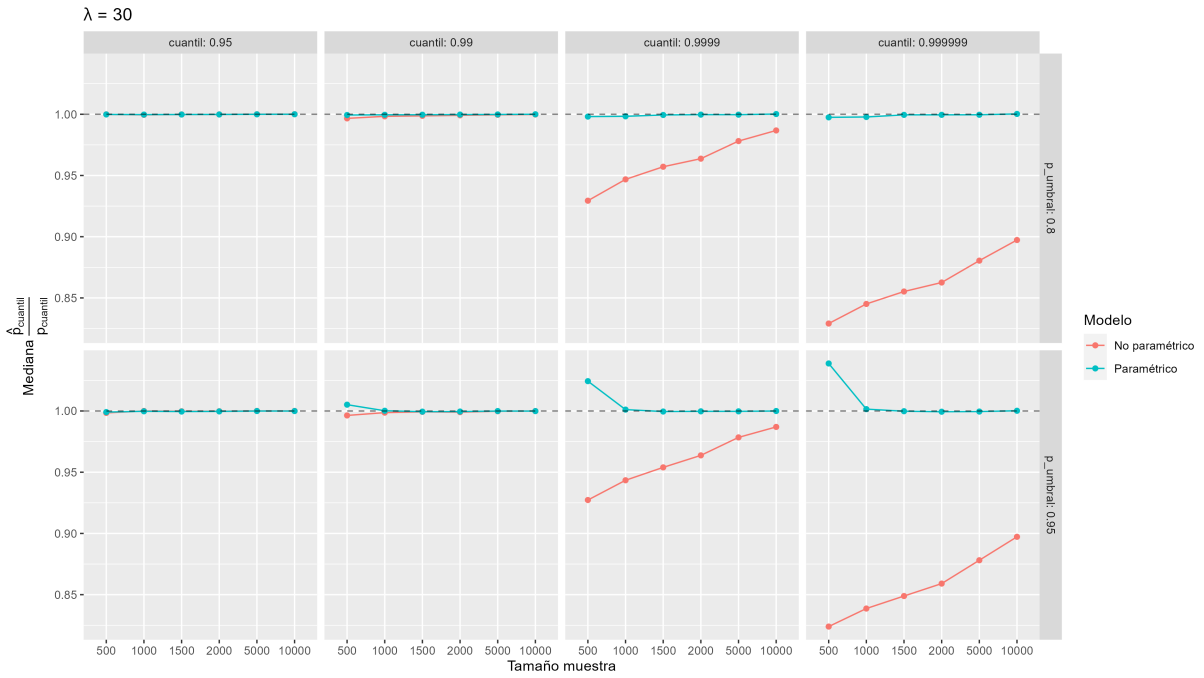

Figura 5.8: Mediana de la razón $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$ vs n con $\lambda = 30$, discriminado según p_{umbral} y $p_{cuantil}$. La línea horizontal punteada representa el valor 1 que es cuando la estimación coincide con el parámetro verdadero.

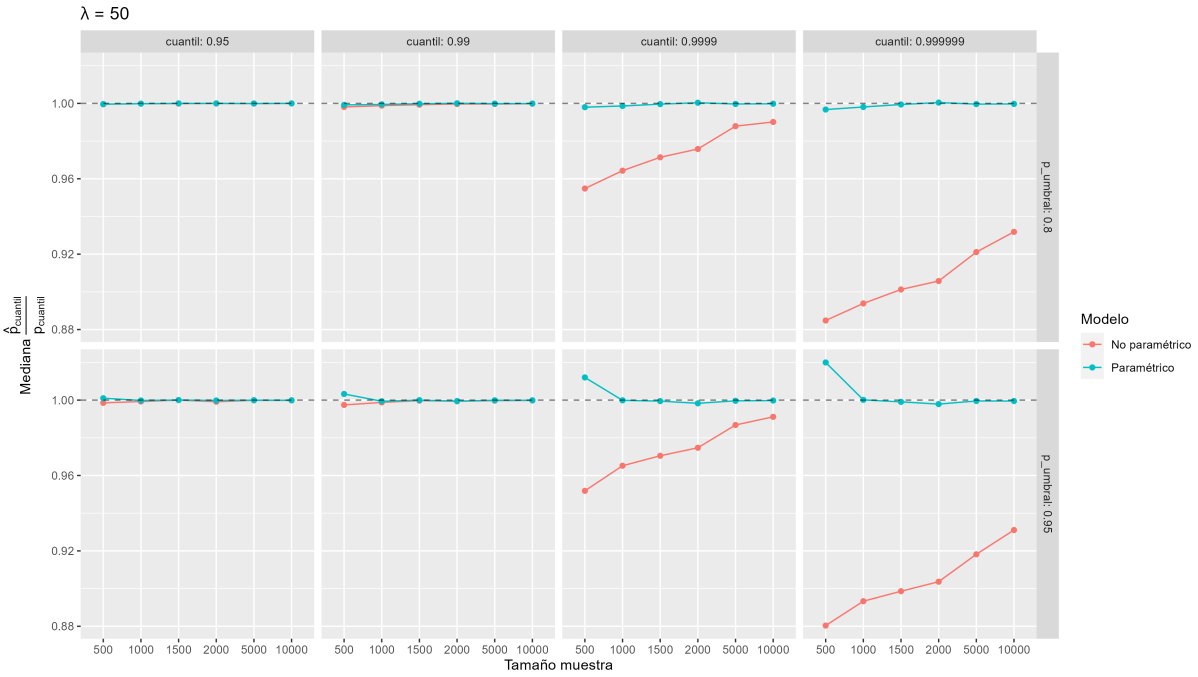

Figura 5.9: Mediana de la razón $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$ vs n con $\lambda = 50$, discriminado según p_{umbral} y $p_{cuantil}$. La línea horizontal punteada representa el valor 1 que es cuando la estimación coincide con el parámetro verdadero.

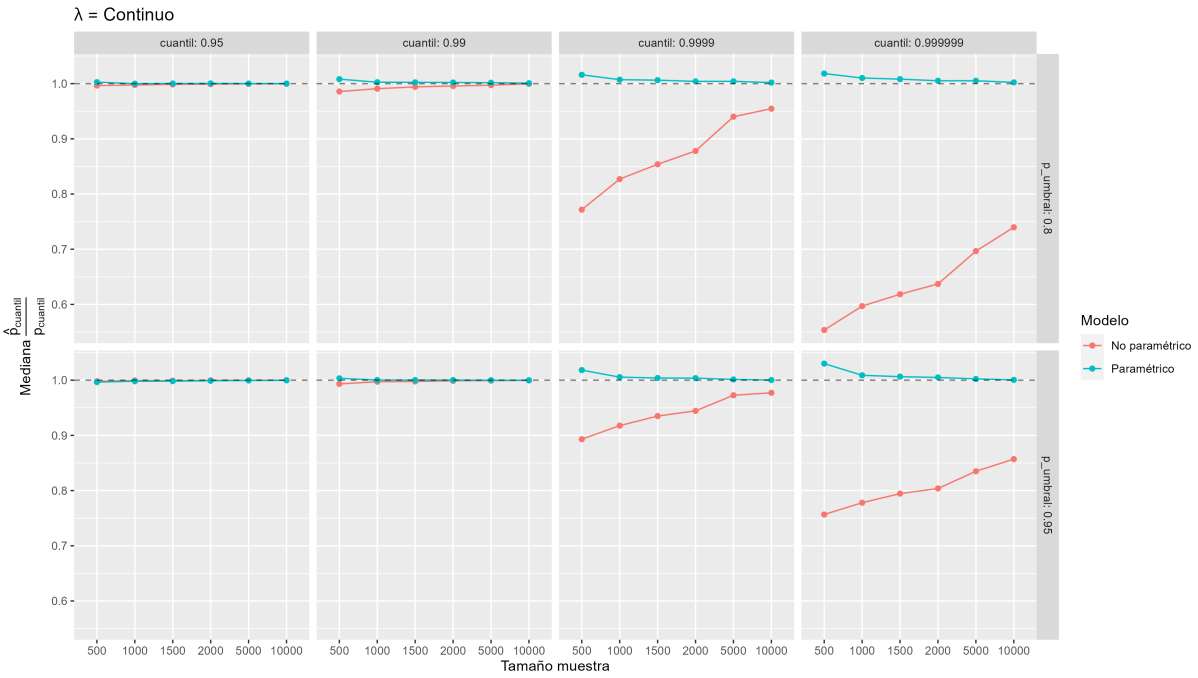

Figura 5.10: Mediana de la razón $\frac{\hat{p}_{cuantil}}{p_{cuantil}}$ vs n con $\lambda = Continuo$, discriminado según p_{umbral} y $p_{cuantil}$. La línea horizontal punteada representa el valor 1 que es cuando la estimación coincide con el parámetro verdadero.

Capítulo 6

Conclusiones

A modo de palabras finales, en este trabajo hemos definido y dado los primeros pasos para estimar el umbral de manera automática al partir del cual poder modelar la distribución de excesos, asumiendo $\mathcal{M}_{\mathcal{E}} = \{F \text{ distribución } : \exists u \in \mathbb{R} \text{ tal que } F_u \sim \mathcal{E}(\lambda)\}$ como modelo semiparamétrico para la distribución de datos.

Dimos fundamentos teóricos a la propuesta, hemos hecho un breve recorrido recapitulando los conceptos de *estimadores de máxima verosimilitud* y de *divergencia de Kullback-Leibler*, describiendo la relación entre ellos, que esencialmente dice que cuando el *EMV* es consistente, las distribuciones del modelo asociadas a dicho parámetro convergen a la distribución cuyo parámetro minimiza la *divergencia KL* con respecto a la distribución poblacional, que es una forma de medir "distancias" entre distribuciones. Esta propiedad se utiliza en la propuesta en el paso donde se mide la distancia en supremos entre la distribución acumulada empírica de los excesos y la propuesta teóricamente. Se utiliza el *EMV* porque aún estando mal el modelo, asintóticamente vamos a encontrar la mejor posible dentro del modelo para la distribución verdadera.

Luego introdujimos el marco de trabajo de la teoría de valores extremos. Se describieron las posibles distribuciones asintóticas de los máximos, pero también los límites que este enfoque tiene en términos prácticos. Se presentó una idea alternativa que consiste en trabajar con las distribuciones de excesos, que además tiene interés por si misma para abordar otro tipo de problemas. Se dieron las condiciones según las cuales con un *umbral* suficientemente grande esas distribuciones se aproximan a algún miembro de la *Familia de Pareto Generalizada*, pero esta idea también presenta un desafío que tiene que ver con encontrar un *umbral* adecuado en donde los excesos estén bien ajustados por algún elemento de esta familia.

A partir de este problema, se definió un modelo que simplifica el problema al caso en que existe un umbral al partir del cual la distribución de excesos es una exponencial de parámetro λ. Y se desarrolló una propuesta en primer lugar para caracterizarlo, y luego también estimarlo junto al parámetro de la exponencial correspondiente. Para evaluar su rendimiento se hicieron una serie de simulaciones obteniendo los siguientes resultados principales:

- A mayor cantidad de datos mayor precisión en la estimación tanto del *umbral* como de λ.
- Cuando el cambio de régimen uniforme al exponencial es continuo, los umbrales estimados son estadísticamente menores al umbral real, teniendo mejores resultados cuando $p_{umbral} = 0{,}95$.
- En cuanto a la estimación de λ se ve una mayor dispersión con $p_{umbral} = 0{,}95$ y con diferencia abrupta entre $n = 500$ y $n = 1000$. Será tarea de otros estudios entender que sucede en el medio, y también el comportamiento anómalo que se observa con $\lambda = 10$ y $p_{umbral} = 0{,}95$.

- Al comparar cuantiles estimados bajo el supuesto del modelo con los parámetros estimados según nuestra propuesta, con los no paramétricos, se observa dentro del rango interpolable una calidad similar en la estimación, con mayor precisión cuando $p_{umbral} = 0{,}8$, mientras que cuando $p_{umbral} = 0{,}95$, $n = 500$ y $\lambda = 0{,}5, 1$ se observa mayor precisión en el método no paramétrico. Sin embargo, para cuantiles mayores, que implican extrapolar la información nuestra propuesta representa una mejora indudable.

Será material para futuros trabajos generalizar el estudio de esta propuesta. Por un lado realizar simulaciones bajo otros escenarios, contemplando distintos casos para la parte no extrema de la densidad, lo que en este trabajo fue considerado con una distribución uniforme. Y también analizar cuando el régimen extremo representa un elemento cualquiera de la *familia de Pareto Generalizada*, en lugar de utilizar solo a la distribuciones exponenciales.

Apéndice A

R scripts

A continuación se encuentran los scripts que permiten generar muestras con el modelo propuesto en el capítulo 5. El resto de los scripts para automatizar la realización de las simulaciones y poder reproducir los resultados se encuentran en el siguiente repositorio: `https://github.com/maty-z/TFL`.

A.1. Funciones utilizadas para generar las muestras

```
# Genero una bernoulli de prob p. Si es 1, entonces tomo una muestra uniforme,
#si no es 1, tomo una muestra de la exponencial.
tomar_muestra.unif_exp <- function(p_umbral,
                                   u_min,
                                   u_max,
                                   exp_lambda) {
  aux <- rbinom(n = 1, size = 1, prob = p_umbral)
  if (aux == 1) {
    z <- runif(1, min = u_min, max = u_max)
  } else {
    z <- rexp(1, exp_lambda) + u_max
  }
  return(z)
}
# Replico el paso para tomar muestras de tamaño size
generar_muestra.unif_exp <- function(size, ...) {
  dist_params <- list(...)
  replicate(size, do.call(tomar_muestra.unif_exp, dist_params))
}
```

A.2. Funciones utilizadas para estimar u_0

```
# Estimo el lambda con maxima verosimilitud
#para los excesos de la muestra con umbral u
lambda_emv <- function(u,muestra) {
  exceso <- muestra[muestra>=u]
  return(length(exceso)/sum(exceso-u)) # lambda_emv = 1/promedio(x-u)|x>u
}
```

```
# Distancia norma supremo entre la empirica de los excesos y la exp
#con parametro dado por el lambda de maxima verosimilitud de: excesos-u
#Aclaracion: Este resultado es el estadistico utilizado por el test
#Kolomogorov-Smirnov
dsup_lambda_u <- function(u,muestra) {
  pexp_rate = lambda_emv(u,muestra)
  exceso <- muestra[muestra>=u]
  return(ks.test(exceso-u,"pexp",rate = pexp_rate)$statistic)
}

# Funcion a optimizar:
# alpha_pen y n_pen son los exponentes del argumento para penalizacion
f_obj <- function(u,muestra,alpha_pen,n_pen,c) {
  return(dsup_lambda_u(u,muestra) + c/length(muestra)^alpha_pen*u^n_pen)
}

# Dada una muestra evalúa la función objetivo en los u candidatos a umbral,
#luego devuelve el que minimiza la f_obj
u_est <- function(muestra, n_pen, alpha_pen, c) {
  u_interval <- c(quantile(muestra, probs = 0.5),
                  quantile(muestra, probs = 0.99)
                 )

  u_candidatos <- muestra[
                          muestra > u_interval[1]
                        & muestra < u_interval[2]
                          ]

  # La función objetivo cambia de comportamiento en los valores muestrales
  #Evalúo la funcion objetivo dentro de los candidatos a umbral,
  #y luego me quedo con el que minimiza la función.
  f_obj.values <- lapply(u_candidatos,
                         f_obj,
                         muestra = muestra,
                         n_pen = n_pen,
                         alpha_pen = alpha_pen,
                         c= c)
  argmin <- which.min(f_obj.values)
  return(u_candidatos[argmin])
}
```

Bibliografía

[Boente et al., 2023] Boente, G., Leonardi, F., Rodriguez, D., and Sued, M. (2023). Threshold detection under a semiparametric regression model.

[Brown, 1986] Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279.

[Cabras and Morales, 2007] Cabras, S. and Morales, J. (2007). Extreme value analysis within a parametric outlier detection framework. *Applied Stochastic Models in Business and Industry*, 23(2):157–164.

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

[Fisher, 1925] Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725.

[Gonzalez et al., 2013] Gonzalez, J., Rodriguez, D., and Sued, M. (2013). Threshold selection for extremes under a semiparametric model. *Statistical Methods & Applications*, 22:481–500.

[Kullback, 1968] Kullback, S. (1968). *Information Theory and Statistics*. Dover.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[Leadbetter et al., 1983] Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Asymptotic Distributions of Extremes*, pages 3–30. Springer New York, New York, NY.

[MacDonald et al., 2011] MacDonald, A., Scarrott, C., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

[Pickands, 1975] Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

[Reiss and Thomas, 2007] Reiss, R.-D. and Thomas, M. (2007). *Parametric Modeling*, pages 3–38. Birkhäuser Basel, Basel.

[Vaart, 1998] Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[Wong and Li, 2010] Wong, T. S. T. and Li, W. K. (2010). A threshold approach for peaks-over-threshold modeling using maximum product of spacings. *Statistica Sinica*, 20(3):1257–1272.