Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

Document de Treball 2020/07 1/39 pág.
Working Paper 2020/07 1/39 pág.

"Retos para el análisis y la estimación de la distribución de probabilidad en Big-data"

Catalina Bolancé

UBIREA | Institut de Recerca en Economia Aplicada Regional i Pública

UBIREA | Institut de Recerca en Economia Aplicada Regional i Pública
UNIVERSITAT DE BARCELONA

WEBSITE: www.ub.edu/irea/ • CONTACT: irea@ub.edu

The Research Institute of Applied Economics (IREA) in Barcelona was founded in 2005, as a research institute in applied economics. Three consolidated research groups make up the institute: AQR, RISK and GiM, and a large number of members are involved in the Institute. IREA focuses on four priority lines of investigation: (i) the quantitative study of regional and urban economic activity and analysis of regional and local economic policies, (ii) study of public economic activity in markets, particularly in the fields of empirical evaluation of privatization, the regulation and competition in the markets of public services using state of industrial economy, (iii) risk analysis in finance and insurance, and (iv) the development of micro and macro econometrics applied for the analysis of economic activity, particularly for quantitative evaluation of public policies.

Abstract

En este documento se describen los principales conceptos relacionados con el ajuste no paramétrico de la distribución de probabilidades cuando se dispone de datos masivos y estos poseen fuerte asimetría a la derecha. En concreto, se estudian datos que representan perdidas positivas, que son muy heterogéneos y, por tanto, que pueden ser muy reducidos, cercanos a cero, o muy elevados y, además, pueden proceder de distintas distribuciones de probabilidad. Además, se mostrará cómo, aún disponiendo de una gran cantidad de datos, el efecto de la censura y el truncamiento sigue siendo un problema de falta de información que provoca grandes sesgos en los valores estimados. También, se describirán algunos resultados relacionados con la estimación paramétrica desde la perspectiva del uso de datos masivos. Finalmente, se presentarán algunos estimadores tipo núcleo, que ya han sido propuestos en la literatura, y que abordan algunas dificultades de los estimadores núcleos más clásicos cuando en los datos existen valores muy extremos los cuales es necesario modelizar para la cuantificación del riesgo.

JEL classification: E30, E39, Y10

Keywords: Análisis univariante, Estimación paramétrica, Estimación no paramétrica, Censura, Truncamiento, Quantificación del riesgo.

Catalina Bolancé: Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain. Email: bolance@ub.edu

Acknowledgements

La autora agradece el apoyo de la Fundación BBVA.

1. Introducción

En este documento, además de analizar los algoritmos para el ajuste paramétrico de distribuciones, el principal reto es poder estudiar los precedentes que permitan plantear algoritmos que ayuden a mejorar la efiencia computacional de los métodos de estimación no paramétricos. En concreto, la estimación núcleo de diferentes funciones que dibujan el comportamiento de la distribución. Todo ello con el objetivo general de predecir el riesgo y, particularmente, cuando la distribución es de cola pesada.

El resultado de cualquier proceso de observación, sea este científico, financiero, económico o social, consiste en unos valores que pueden representarse mediante una variable aleatoria que puede ser continua o discreta. En las mejores condiciones, dicho proceso de observación debería garantizar obtener información que permita realizar el análisis de forma insesgada y eficiente. Teniendo en cuenta esta idea, a priori, cuando se hace referencia a datos masivos (en adelante big-data) se asume que la disponibilidad de muchos datos debería garantizar que las estimaciones cumplirán ambas propiedades. Sin embargo, en la práctica más clásica del análisis estadístico y, sobre todo, en un contexto del big-data, donde la información regogida no está sujeta a un diseño muestral, dicha situación ideal puede no producirse. Además, se dan situaciones de información incompleta y/o sesgada, algunas de las cuales se describirán en este documento. Abordar dichas situaciones en un contexto del big-data es fundamental y pone de manifiesto la necesidad del preprocesamiento de los datos, que permita explorar sus principales características y facilite la decisión sobre el uso de modelos más o menos complejos para la cuantificación del riesgo.

En general, los algoritmos en big-data consisten en guardar la información en forma de estadísticos suficientes que permitan la estimación sencilla de los parámetros del modelo. Sin embargo, existen muchas situacines, particularmente en el ámbito de la cuantificación del riesgo, en las cuales la estimación no puede obtenerse a partir de estadísticos suficientes, como puede ser el caso de algunas distribuciones como la log-logística generalizada, la distribución de Pareto generalizada o algunas otras distribuciones de cola pesada (see Bahraoui et al., 2015; Buch-Larsen et al., 2005). Además, no siempre se estará interesados en los valores centrales de la distribución, por el contrario, en el contexto del análisis del riesgo el interés radica en los cuantiles más extremos.

Desde una perspectiva univariante, se analiza cómo se distribuyen los datos observados dentro de un dominio. Como alternativa a la estimación paramétrica a la que se hizo referencia en el párrafo anterior, en un contexto de big-data, el análisis desde una perspectiva totalmente no-paramétrica, de modo que la información muestral sea la que dibuje la forma de la función a estimar, cada vez es más aplicable. Ante ello, la estimación núcleo es un método no-paramétrico, lo suficientemente flexible y que puede adaptarse de forma muy sencilla al ajuste de distintas funciones (véase Bowman y Azzalini, 1997, donde se presentan las aplicaciones más clásicas de

la estimación núcleo). Partircularmente, el estimador núcleo de la función de densidad (véase Silverman, 1986, para una revisión) ha sido ampliamente utilizado como herramienta de análisis (see Tsay, 2016).

La estimación núcleo de la función de distribución ha sido menos estudiada que la de la función de densidad. Sin embargo, se mostrará cómo ésta puede ser una herramienta válida para la cuantificación del riesgo. Fue Azzalini (1981) quien analizó las propiedades estadísticas del estimador núcleo de la función de distribución y, además, planteó la estimación del cuantil a partir de su inversa y también, paralelamente, dedujo sus propiedades (véase también Reiss, 1981). En referencia a la estimación de los cuantiles de la distribución, alternativamente al estimador de Azzalini (1981), se han propuesto diversos estimadores del cuantil, basados en el estimador núcleo de la regresión, que mejoran las propiedades en muestra finita del primero. Sheather y Marron (1990) realizan una amplia revisión de dichos estimadores. Sin embargo, en algunos ámbitos de la cuantificación del riesgo, estos estimadores no son eficientes dado que fallan cuando el objetivo es estimar un cuantil extremo de una distribución de cola pesada. Ante ello, se han propuesto alternativas que mejoran la eficiencia de los estimadores existentes, las cuales se basan en la transformación de los datos. Además, esta estrategia basada en la transformación también implica una mejora computacional en la cuantificación del riesgo basada en la estimación del cuantil extremo.

El uso combinado de un estimador paramétrico y no-paramétrico en el contexto del big-data permite la aproximación de las funciones de densidad, de distribución y los cuantiles, garantizando la consistencia de los estimadores cuando el modelo paramétrico no coincide con el que teóricamente es el generador de los datos y una menor varianza que los métodos no-paramétricos puros.

A continuación, se presentan las herramientas básicas para el análisis empírico de la distribución de una variable aleatoria a partir de unos datos, los cuales, además, pueden tener problemas de falta de información, como son la censura y el truncamiento. También, se muestran algunos resultados relacionados con las distribuciones de valor extremo que permiten clasificar las distintas formas de decrecimiento de la cola derecha de la distribución, lo cual es fundamental para el análisis de las propiedades de los estimadores que combinan métodos paramétricos y no paramétricos. En este contexto, también se aborda la estimación paramétrica de la distribución univariante en big-data, que junto a la estimación núcleo, permiten obtener un estimador que combina ambas técnicas y que, además, plantea retos de cara al ajuste de distribuciones y a la predicción de cuantiles en big-data, los cuáles se describen al final de este documento.

2. Conceptos básicos en el análisis univariante

Formalmente, desde el punto de vista de la estadística teórica, una variable aleatoria es una función definida en el espacio de probabilidad: $(\Omega, \mathcal{A}, \mathcal{P})$, donde Ω es el espacio muestral, también conocico como el conjunto de sucesos elementales; $\mathcal{A}$ es el conjunto de todos sucesos aleatorios y $\mathcal{P}$ es la función de probabilidad. Por lo tanto, se define variable aleatoria X como la función: $X : \Omega \longrightarrow \Re$.

Toda variable aleatoria X tiene asociada una distributión de probabilidad con función de densidad (pdf-probability distribution function) $f_X(x)$ y función de distribución (cdf-cumulative distribution function) $F_X(x)$, a partir de las cuales podremos calcular todas las probabilidades en $\mathcal{P}$ asociadas al conjunto de sucesos aleatorios en el espacio de probabilidad. En la práctica, desde una perspectiva tradicional se debería garantizar que los datos observados permitan la estimación de estas funciones de forma consistente, eficiente e insesgada. Es decir, cuando el número de observaciones tiende a infinito se debería ser capaz de reproducir exactamente el comportamiento de la distribución de probabilidades.

La forma más directa y sencilla de estimar las funciones de densidad y de distribución son, respectivamente, el histograma y la distribución empírica. Ambos son herramientas que se basan únicamente en la información muestral y, por tanto, cuanto mayor es el número de datos mejor recogeran la forma de f_X y F_X. Sin embargo, dependiendo de la forma de estas funciones, esto puede no ser así. A continuación, se describe brevemente en qué consisten el histograma y la distribución empírica e se ilustrarán sus deficiencias con un ejemplo típico en el contexto de la cuantificación de riesgo.

El histograma proporciona una estimación de la función de densidad para variable aleatorias continuas, que consiste en definir unos intervalos de valores a lo largo del dominio de la variable, los cuales pueden ser de la misma amplitud o no. Sea $X_1, ..., X_n$ una muestra de observaciones de la variable aleatoria X y sean I_j, $j = 1, ..., m$ una serie de intervalos $I_j = [l_j, u_j)$ consecutivos, de modo que $l_{j+1} = u_j$, definidos dentro del dominio de la variable, si el número de observaciones dentro de cada intervalo es n_j, con $\sum_{j=1}^{m} n_j = n$, la densidad se aproxima dentro de cada intervalo como $\hat{f}_j = \frac{n_j}{u_j - l_j}$. Existen distintos criterios para seleccionar el número de intervalos m, el más común es $m = \sqrt{n}$.

En la Figura 1 se muestra un ejemplo de histograma estimado a partir de una muestra de $1,000,000$ de datos procedente de una variable aleatoria con fuerte asimetría a la derecha, cuya forma es muy común en el contexto de la cuantificación del riesgo. Para poder ver la forma del histograma en las distintas partes del dominio de la variable, se ha dividido el gráfico en cuatro partes y se ha adaptado la amplitud de los intervalos a la disponibilidad de muestra en cada parte del dominio analizada. La Figura 1 muestra como, aunque el tamaño muestral es de 1 millón de observaciones, en los dos gráficos inferiores hay intervalos de valores del dominio

de la variable donde no hay datos y, además, dichos intervalos se sitúan en los valores más elevados de la variable, que suele ser la zona de la distribución a la que está asociada el cálculo del riesgo.

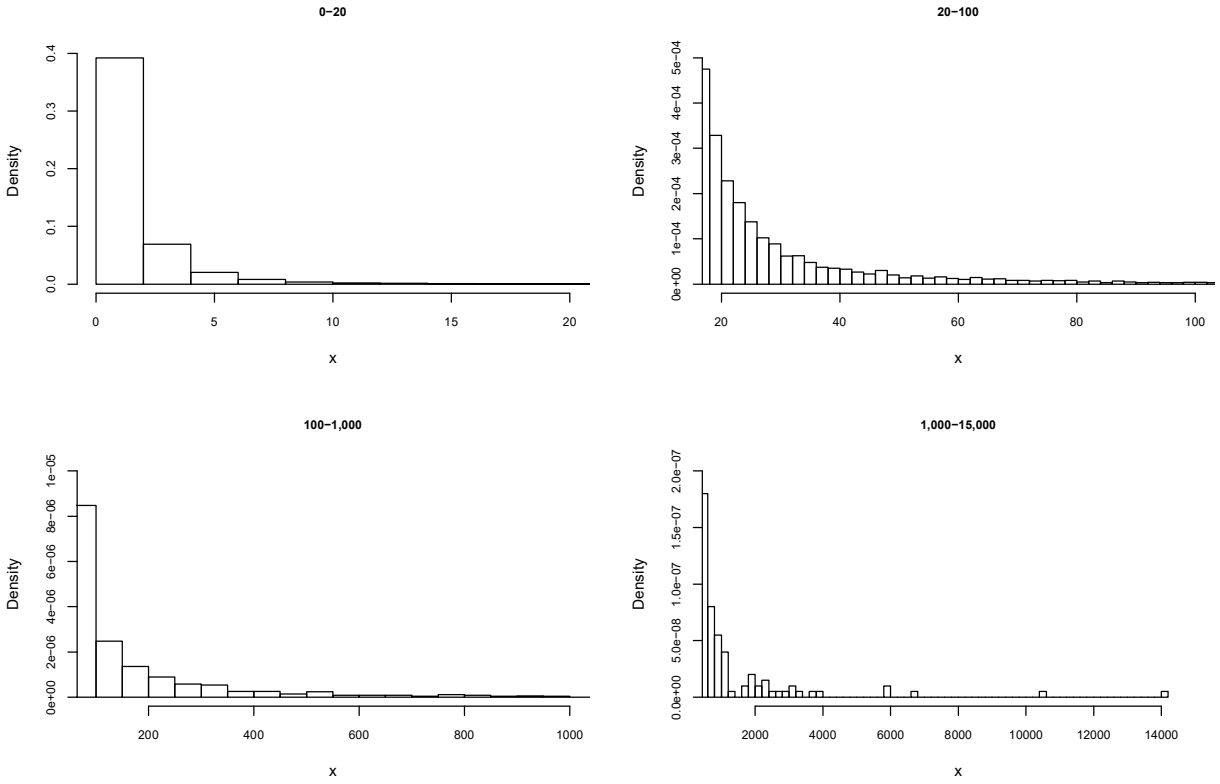

Figura 1: Ejemplo de histograma de una variable aleatoria asimétrica a la derecha, en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

La distribución empírica es un estimador de F_X que puede obtenerse de forma muy sencilla para cualquier valor x de la variable aleatoria y se define como:

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \tag{1}$$

donde $I(\cdot)$ es la función indicador que toma valor 1 si la condición entre paréntesis es verdadera y 0 en caso contrario. El sesgo de $\widehat{F}_n$ es cero y su varianza es:

$$V\left[\widehat{F}_n(x)\right] = F_X(x)\left[1 - F_X(x)\right]/n. \tag{2}$$

En la Figura 2 se muestra la función de distribución empírica representada en los mismos cuatro intervalos de valores en los que se ha representado el histograma para la Figura 1. La distribución empírica únicamene cambia de valor en los valores observados en la muestra, de ahí su forma escalonada. En el ejemplo de la Figura 2 se observa cómo la forma escalonada de la distribución empírica se hace más evidente a medida que aumentan los valores de la variable, es decir, en la cola derecha de la distribución.

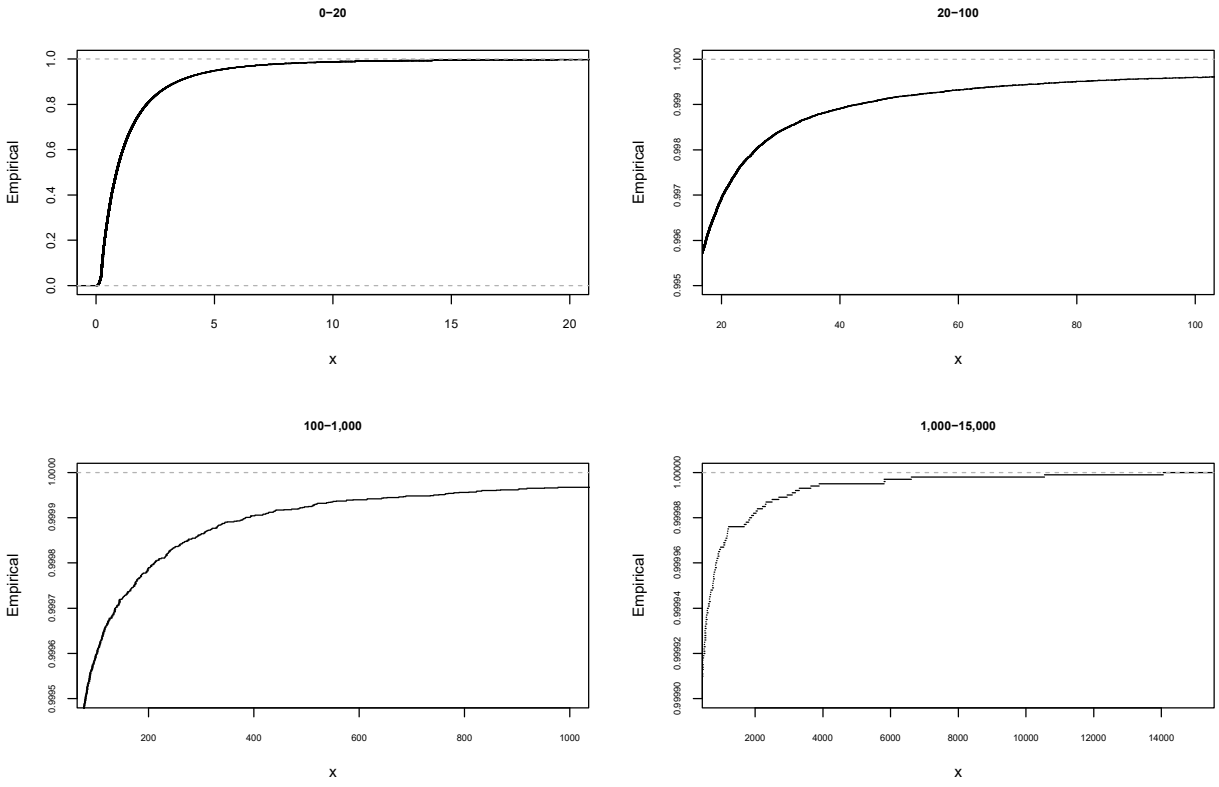

Figura 2: Ejemplo de distribución empírica de una variable aleatoria asimétrica a la derecha, definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

2.1. El dominio de una variable aleatoria

Dado que una variable aleatoria X es una función definida en el espacio muestral Ω, dicha función $X(\omega)$ tiene asociado un dominio en Ω que, en definitiva, contiene los valores que puede tomar dicha variable y que puede ser acotado o no.

En el contexto de la cuantificación de riesgo, en muchos casos la variable aleatoria objeto de análisis, la cual está asociada al riesgo, es una pérdida que toma valores positivos, esto es:[1]

$$X : \Omega \longrightarrow \Re^{+}.$$

En la cuantificación de riesgo el interés se centrará en la cola derecha de la distribución, concretamente en los cuantiles más extremos próximos al máximo. El Teorema de los Valores Extremos, que es similar al Teorema Central del Límite, pero se centra en el comportamiento asintótico de la distribución cuando los valores de la variable tienen al máximo, lo que se conoce como Distribución de Valor Extremo (EVD-*Extreme Value Distribution*). Todo ello se estudia dentro de la Teoría del Valor Extremo (EVT-*Extreme Value Theory*) (véase, por ejemplo, Reiss y Thomas, 2007; Coles, 2001). Dentro de la EVT, el **dominio de atracción del máximo** (MDA-*maximum domain of attraction*) está relacionado con la forma de la EVD (see Kotz y Nadarajah, 2000).

La expresión de la cdf de la EVD generalizada es:

$$\begin{array}{ll} G_{\xi}(x,\mu,\sigma) = \exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} & \text{if } \xi \neq 0 \\ G_{\xi}(x,\mu,\sigma) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \text{if } \xi = 0 \end{array} \tag{3}$$

y su pdf es:

$$\begin{array}{ll} g_{\xi}(x,\mu,\sigma) = \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{(-1/\xi)-1}\exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} & \text{if } \xi \neq 0 \\ g_{\xi}(x,\mu,\sigma) = \exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \text{if } \xi = 0, \end{array} \tag{4}$$

donde ξ es el parámetro de forma, μ es el parámetro de posición y σ el parámetro de escala. La esperanza de la variable aleatoria X con EVD generalizada depende del parámetro de forma ξ como sigue:

$$E(X) = \begin{cases} \mu + \sigma\frac{\Gamma(1-\xi)-1}{\xi} & \text{if} \quad \xi \neq 0, \xi < 1 \\ \mu + \sigma\gamma & \text{if} \quad \xi = 0 \\ \infty & \text{if} \quad \xi \geq 1 \end{cases}, \tag{5}$$

donde $\Gamma(\cdot)$ es la función gamma de Euler y γ es la constante de Euler. El MDA de G_{ξ} también depende del parámetro de forma ξ. En la expresión (4) cuando $\xi = 0$ tendremos una EVD

[1] En el ámbito financiero, cuando se analizan los rendimientos (beneficios), la variable tomará valores positivos o negativos y el riesgo estará asociado a los valores más negativos.

tipo Gumbel (MDA-Gumbel) and cuando $\xi > 0$ el resultado es una EVD tipo Fréchet (MDA-Fréchet). También existe un tercer caso cuando $\xi < 0$ que se corresponde con el tipo Weibull pero que no se suele dar en el contexto de la cuantificación de riesgo dado que supone asimetría a la izquierda (MDA-Weibull).

Algunos ejemplos concretos de distribuciones con MDA Gumbel son la Normal, la Lognormal y la Weibull. Distribuciones que tiene MDA Fréchet son la t de Student, la Pareto, la Cauchy y la Log-Logística (o distribución de Champernowne). El nombre de esta distribución se debe al matemático y economista David Gawen Champernowne que inicialmente la propuso para el ajuste del logaritmo de la renta. Posteriormente fue redefinida como la distribución log-logística, conocida en economía como la distribución de Fisk, de modo que el logaritmo de una variable con distribución de Fisk posee distribución logística, relación similar a la que posee la log-normal con la normal.

Otra características de las EVD es lo que se conoce como el punto final derecho (right endpoint). Definimos el punto final derecho de G como $r(G) = \sup(x|G(x) < 1)$. Además, sabemos que si dos distribuciones G_1 y G_2 son tal que $r(G_1) = r(G_2)$, entonces:

$$\lim_{x \uparrow r(G_1)} \frac{\bar{G}_1(x)}{\bar{G}_2(x)} = c,$$

para alguna constante $0 < c < \infty$, donde $\bar{G}_1(x) = 1 - G_1(x)$ y $\bar{G}_2(x) = 1 - G_2(x)$. En este caso G_1 y G_2 tienen el mismo MDA, además, se dice que G_1 y G_2 son de cola equivalente si:

$$\lim_{x \uparrow r(G_1)} \frac{\bar{G}_1(x)}{\bar{G}_2(x)} = 1.$$

En la Figura 3 se representan las funciones de densidad y de distribución de la EVD generalizada para los tres casos que se contemplan en la expresión (5) y para $\xi > 0$, y se muestra como este parámetro influye en la forma de la distribución de la variable aleatoria.

2.2. Censura y truncamiento

La censura y el truncamiento son situaciones de información incompleta asociados a los valores observados en la muestra. Existen notables diferencias entre ambas situaciones. La censura se da cuando lo que se mide no es el valor de la variable si no parte de ese valor o un valor que lo sustituye, sin embargo, el truncamiento es un filtro sobre la muestra observada, es decir, aquellos casos que no pasen el filtro directamente no se observan.

La censura clásicamente se ha asociado al análisis de supervivencia, donde la variable objeto de análisis es el tiempo hasta la ocurrencia de un suceso. El caso de censura más común es el

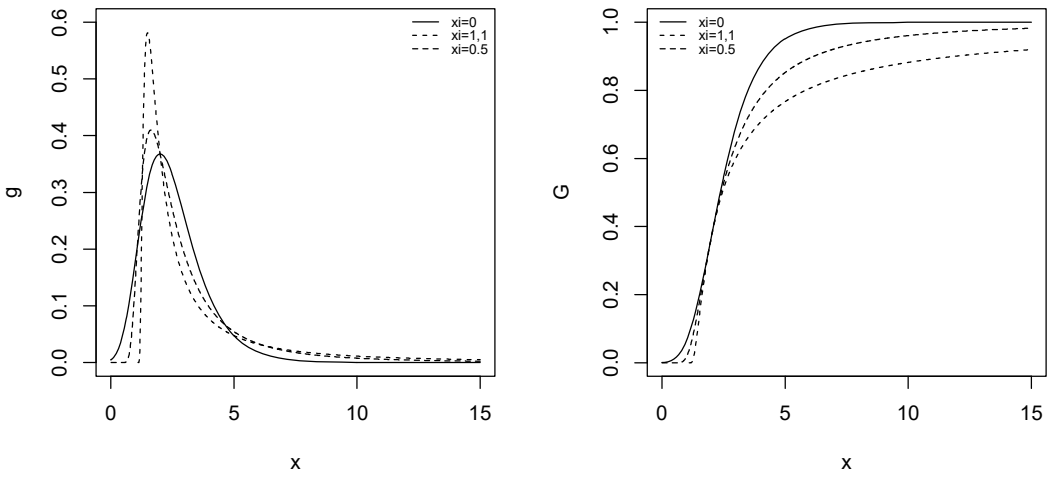

Figura 3: Distribución de Valor Extremo Generalizado para diferentes valores de ξ, pdfs a la izquierda y cdfs a la derecha.

de censura por la derecha, lo que sucede cuando el tiempo del estudio a finalizado -tiempo de censura- y existen casos en los que el suceso analizado no ha ocurrido y por tanto el dato observado es igual al tiempo de censura. A este tipo de censura por la derecha se le denomina fija y puede generalizarse a cualquier tipo de variable que represente una cuantía económica, una velocidad, etc..

También existe la censura por la derecha progresiva, en la cual los valores de censuras son diferentes para diferentes observaciones. Esta situación ocurre cuando, debido a causas ajenas al estudio, algunas observaciones dejan de formar parte del análisis. Por ejemplo, en el contexto del seguro podríamos estar interesados en cuantificar el coste total de los siniestros durante un año; sin embargo, existen asegurados que cancelan su póliza antes de un año, estos asegurados estarán en la muestra pero la información recogida será que el coste total será igual o mayor al cuantificado hasta la fecha.

En general, la censura por la izquierda es menos común, pero en el contexto de la cuantificación de riesgo existen situaciones en las que se da. Este tipo de censura por la izquierda ocurre cuando el suceso analizado se da antes del inicio del estudio o, por ciertas razones controladas por el analista, aunque la observación está dentro del estudio el dato no empieza a contabilizarse hasta que este supera una determinado cuantía. Por ejemplo, en el ámbito de la medicina y los seguros de salud, no se sabe desde cuando un individuo está enfermo y el tiempo

se empieza a medir desde en el momento que se detecta la enfermedad; también, en el ámbito de los seguros de automóvil, algunos sensores de control de velocidad no se activan hasta pasado un límite. También existe la censura en un intervalo, la cual se da cuando se sabe que el valor de la variable se situa en un cierto intervalo que puede ser aleatorio aleatorio o no. Por ejemplo, de nuevo en el ámbito médico, cuando se sabe que la enfermedad se ha desarrollado durante el tiempo entre dos pruebas. En una misma muestran pueden existir datos con distintos tipos de censura: por la derecha, por la izquierda o en un intervalo.

Cuando existen datos censurados el supuesto que suele realizarse es que la censura, sea del tipo que sea, es no informativa. Esto implica independencia entre las observaciones censuradas y no censuradas y significa que los valores de censura son totalmente aleatorios y no influyen en la distribución de la variable. Dicho de otro modo, la distribución de los datos censurados y no censurados es la misma.

El truncamiento es un problema de falta de información más simple que la censura. Este se da cuando por algunas razones ajenas o no al analista sólo se observan los datos ligados a una parte del dominio de la variable aleatoria, es decir, sólo se dispone de muestra para representar una parte de la distribución. Al igual que la censura, el truncamiento puede ser por la derecha o por la izquierda. El primero -por la derecha- se da cuando el filtro es del tipo $X \leq x_t$ y el segundo -por la izquierda- cuando el filtro es $X \geq x_t$, siendo x_t un valor de truncamiento.

Dadas f_X y F_X para una observación X_i de la cual se puede obtener el dato completo podremos calcular el valor de $f_X(X_i)$; sin embargo, para las observaciones censuradas esta información la conoceremos parcialmente. Concretamente para los distintos tipos de censura:

- Censura por la derecha $1 - F_X(R_i)$,
- Censura por la izquierda $F_X(L_i)$,
- Censura en un intervalo $F_X(R_i) - F_X(L_i)$,

donde R_i y L_i son valores de censura por la derecha y por la izquierda, respectivamente. Todo ello se tendrá que tener en cuenta en el cálculo de la verosimilitud de la muestra. En definitiva, suponiendo que en la muestra existen los tres tipos de censura, la función de verosimilitud será:

$$L\left(X_1, ..., X_n\right) = \prod_{i \in \mathcal{D}} f_X(X_i) \prod_{i \in \mathcal{R}} \left(1 - F_X(R_i)\right) \prod_{i \in \mathcal{L}} F_X(L_i) \prod_{i \in \mathcal{I}} \left(F_X(R_i) - F_X(L_i)\right), \tag{6}$$

donde $\mathcal{D}$, $\mathcal{R}$, $\mathcal{L}$ y $\mathcal{I}$ son los conjuntos de datos completos, censurados por la derecha, por la izquierda y en un intervalo, respectivamente.

Si f_X y F_X tienen una forma ligada a una distribución de probabilidades paramétrica, el método ampliamente utilizado para la estimación de los parámetros es el que se basa en la maximización de la función de verosimilitud. Aunque en el contexto del big-data la elección de dicho

método de estimación dependerá de la eficiencia computacional, más adelante comentaremos este aspecto.

Cuando hay censura por la derecha, el estimador de Kaplan-Meier (ver Kaplan y Meier, 1958), que se define a continuación, es la aproximación empírica de la función de distribución. Sean $R_1, ...R_n$, respectivamente, los valores de censura para cada observación X_i, cuya función de distribución se representa como F_R. Se supone que las variables aleatorias X y R son independientes. A partir de los n pares de observaciones (X_i, R_i) se definen dos nuevas variables que equivalen a:

$$U_i = \text{mín}(X_i, R_i) \quad \forall i = 1, ..., n \tag{7}$$

y

$$I_i = \begin{cases} 1 & \text{si } X_i \leq R_i \\ 0 & \text{si } X_i > R_i. \end{cases}$$

Las observaciones $U_1, ..., U_n$ son los valores muestrales de los que se dispone para realizar la estimación; su función de distribución es F_U. Además, también se dispone de la variable I_i, que indica si una observación i está o no censurada, tomando valor 1 cuando no existe censura y 0 en caso contrario. Por tanto, en lugar de los pares (X_i, R_i), la información muestral de la que se dispondrá para obtener la estimación de Kaplan-Meier se corresponde con los pares (U_i, I_i). Con estos, el estimador de Kaplan-Meier de $F_X(x)$ es:

$$\hat{F}_X^{K-M}(x) = \begin{cases} 0 & \text{si } 0 \leq x \leq U_{(1)}, \\ 1 - \prod_{i=1}^{j-1} \left(\frac{n-i}{n-i+1}\right)^{I_{(i)}} & \text{si } U_{(j-1)} < x \leq U_{(j)} \ \forall j = 2, ..., n, \\ 1 & \text{si } x > U_{(n)} \quad \text{si } U_{(n)} \text{ no está censurado}, \end{cases} \tag{8}$$

donde los subíndices entre paréntesis indican que los pares $(U_{(i)}, I_{(i)})$ están ordenados de menor a mayor según los valores que toman los U_i. [2]

En la expresión (8) se observa que las observaciones censuradas no modifican el estimador. También es fácil demostrar que si no hay datos censurados el estimador de Kaplan-Meier

[2]Cuando hay empates la expresión se define en forma de datos agrupados:

$$\hat{F}_X^{K-M}(x) = \begin{cases} 0 & \text{si } 0 \leq x \leq U_{(1)}, \\ 1 - \prod_{X_{(i)} \leq x} \left(1 - \frac{d_i}{n_i}\right)^{I_{(i)}} & \text{si } U_{(j-1)} < x \leq U_{(j)} \ \forall j = 2, ..., n, \\ 1 & \text{si } x > U_{(n)} \quad \text{si } U_{(n)} \text{ no está censurado}, \end{cases} \tag{9}$$

donde los subíndices entre paréntesis indican orden de menor a mayor, n_i es el número de datos observados hasta el valor $X_{(i)}$, d_i es el número de casos no censurados con valor $X_{(i)}$.

coincide con la distribución empírica. Una dificultad que se añade al estimador definido en (8) es que cuando el valor máximo observado coincide con un dato censurado la distribución no estará definida a partir de este valor. Por tanto, a partir de este valor será necesario extrapolar los valores de la distribución. Para ello, el estimador núcleo que definiremos más adelante será de utilidad.

En la Figura 4 se representa la función de distribución de Kaplan-Meier censurando un 10 % de la muestra original, cuya distribución empírica se representaba en la Figura 2. Se supone censura por la derecha no fija o progresiva, es decir, cada valor de la variable puede tener un valor de censura distinto. Junto al estimador de Kaplan-Meier se representa la distribución empírica que no tiene en cuenta la censura. Dicha figura evidencia el sesgo en el que se incurre si no se tiene en cuenta que hay datos censurados por la derecha. Se puede observar como con la distribución empírica (línea fina) se subestimaría el riesgo, dado que la curva tiende a 1 más rápidamente.

Como se ha descrito anteriormente, el estimador de Kaplan-Meier está definido para datos censurados por la derecha, si lo que se tiene son datos censurados por la izquierda una opción es cambiar el signo a las observaciones para, posteriormente, una vez obtenido el estimador, volver al signo original.

Cuando en los datos hay truncamiento, simplemente se tiene que corregir la función densidad para que esta integre 1 en el dominio de los datos observados. Para los distintos tipos de truncamiento y dado un valor de truncamiento x_t la función de densidad truncada se obtiene del siguiente modo:

- Truncamiento por la derecha $\frac{f(X_i)}{F_X(x_t)}$,
- Truncamiento por la izquierda $\frac{f(X_i)}{1-F_X(x_t)}$.

La función de distribución se obtiene del mismo modo que la función de densidad.

Así como en el caso de la censura, el caso más común es cuando esta se produce por la derecha, en el caso del truncamiento comúnmente se da por la izquierda. Es decir, viene dado por un filtro en la muestra del tipo $X >= x_t$. En general, esta situación es muy común en los modelos de utilidad, donde dicha utilidad se supone normal y se asocia a un beneficio económico, sin embargo, únicamente se observa una parte de la población, por ejemplo, los asalariados, en este caso solo se medirá la utilidad positiva y la variable se distribuirá como una normal truncada.

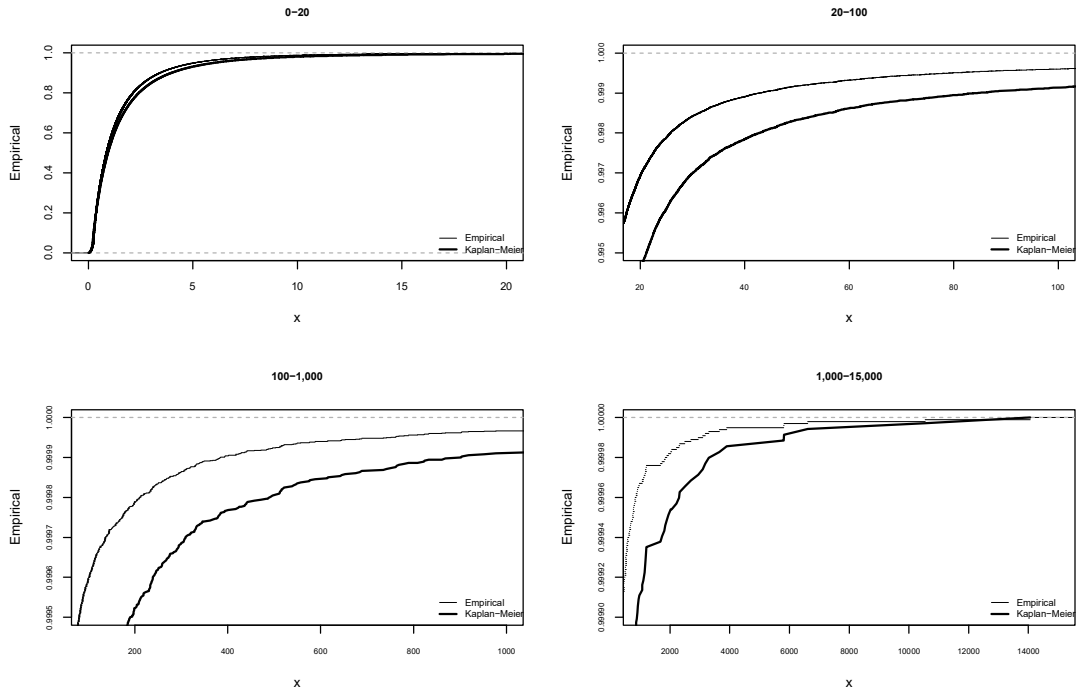

Figura 4: Estimador de Kaplan-Meier suponiendo un 10 % de datos censurados (línea gruesa) y distribución empírica sin tener en cuenta la censura (línea fina) de una variable aleatoria asimétrica a la derecha, definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

2.3. Una variable aleatoria positiva: consecuencias de la censura por la izquierda y el truncamiento por la derecha

Cuando se analiza el riesgos surgen dificultades de información incompleta que no son tan comunes en otros ámbitos como los del análisis de la supervivencia, donde la censura por la derecha y el truncamiento por la izquierda son los más frecuentes. Específicamente, se trata de la censura por la izquierda y el truncamiento por la derecha. En el primer caso, el dato se mide a partir de que supera una determinada cuantía y, en caso contrario, simplemente toma un valor máximo. En el segundo caso, el truncamiento por la derecha, sólo se observarán datos por debajo de un determinado valor.

La censura por la izquierda se da cuando solo se puede medir la variable a partir de un determinado valor. Por ejemplo, en el contexto de los datos telemáticos procedentes de sensores que miden la velocidad del vehículo, dicho sensor suele activarse a partir de una determinada

velocidad y en el caso de no activarse la velocidad se asume igual a un mínimo. Por tanto, si se quisie analizar la distribución de la velocidad, tendríamos que tener en cuenta la censura por la izquierda. Formalmente, se tendrían que definir los valores $L_1, ..., L_n$ de censura para cada observación X_i (puede asumirse que cada sensor tiene distinta sensibilidad). A partir de los n pares de observaciones (X_i, L_i) se definen dos nuevas variables que equivalen a:

$$V_i = \text{máx}(X_i, L_i) \quad \forall i = 1, ..., n$$

y

$$I_i = \begin{cases} 1 & \text{si } X_i \geq L_i \\ 0 & \text{si } X_i < L_i. \end{cases}$$

En la práctica se dispondrá de los pares de valores (V_i, I_i). Como se ha mostrado con anterioridad en la expresión (9), el estimador de Kaplan−Meier puede utilizarse cuando hay censura por la derecha. En el caso de la censura por la izquierda un truco es cambiar el signo de los valores de V_i para, una vez estimada la función de distribución, volver al signo original. En la Figura 5 se representa la función de distribución de Kaplan-Meier con un 10 % de la muestra original con censura por la izquierda. Se supone censura por la iqquierda progresiva, es decir, para cada observación hay un valor de censura distinto. Junto al estimador de Kaplan−Meier se representa la distribución empírica que no tiene en cuenta la censura. Dicha figura evidencia el sesgo en el que incurrimos si no se tiene en cuenta que hay datos censurados por la izquierda. Se puede observar como con la distribución empírica (línea fina) se sobreestima el riesgo, dado que la curva tiende a 1 más rápidamente.

Como se observa en la Figura 5 la distribución empírica tiende a 1 más lentamente que el estimador de Kaplan−Meier, por lo tanto, si no se tiene en cuenta la censura por la izquierda, al contrario de lo que sucedía con la censura por la derecha, el analista sobreestimará el riesgo.

Cuando se trata de una variable que toma valores positivos, el truncamiento por la derecha se da si el analista decide eliminar aquellos casos extremos (valores muy elevados) que distorsionan el valor de los estimadores asociados al centro y a la dispersión de la distribución. Sin embargo, en el contexto del análisis del riesgo, son estos valores extremos los que influyen en su cuantificación.

3. Ajuste de distribuciones en un entorno de big-data

En el contexto del big-data, el ajuste de distribuciones con métodos paramétricos o no paramétricos es similar al que se plantea ante una muestra clásica, pero con una dificultad añadida ligada a que se tienen que gestionar una gran cantidad de datos. En esta sección se

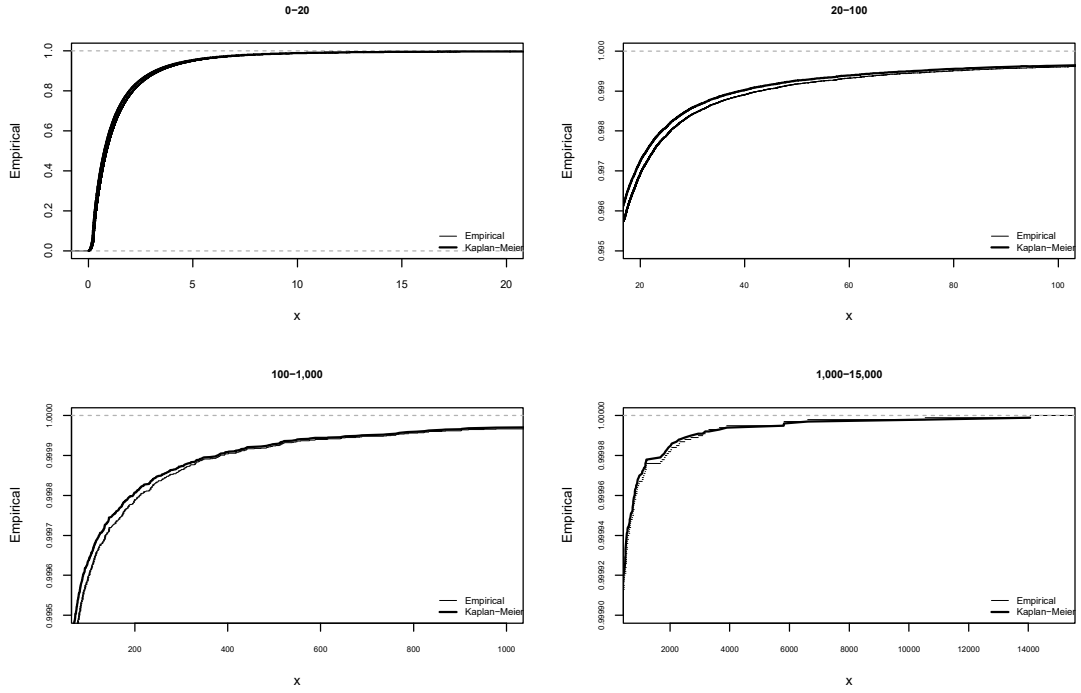

Figura 5: Estimador de Kaplan-Meier suponiendo un 10 % de datos censurados por la izquierda (línea gruesa) y distribución empírica sin tener en cuenta la censura (línea fina) de una variable aleatoria asimétrica a la derecha, definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

presentan los estimadores más clásicos y se plantean las principales retos cuando se dispone de datos masivos.

3.1. Distribuciones paramétrica

Sea X una variable aleatoria cuya distribución tiene asociada unas funciones de densidad $f_X(\cdot|\Theta)$ y de distribución $F_X(\cdot|\Theta)$ cuya forma funcional está asociada a un vector de parámetros $\Theta = (\theta_1, ..., \theta_k)'$ ($'$ indica traspuesta) que se tiene que estimar a partir de una muestra. Para ello puede utilizarse el método basado en la maximización de la verosimilitud, el método de los momentos y el método de los momentos generalizados.

El método de máxima verosimilitud se basa en la maximización del logaritmo de la función de verosimilitud de la muestra. Suponiendo que no existe censura ni truncamiento, dicho logaritmo

de la función de verosimilitud es:

$$l(\hat{\Theta}) = \sum_{i=1}^{n} log\left[f_X(X_i|\hat{\Theta})\right], \tag{10}$$

donde $\hat{\Theta}$ son los parámetros estimados.

El método de los momentos o los momentos generalizados consiste en encontrar la solución de un sistema de ecuaciones que relaciona k momentos de la distribución, $m_j = E(X^j),\ j = 1, ..., k$, o una versión ponderada de los mismos con los k parámetros a estimar.

Asumiendo que se dispone de un elevado número de datos y que la distribución paramétrica coincide con el modelo generador de dichos datos, las propiedades estadísticas de los distintos estimadores están garantizadas y, por tanto, su consistencia. Por este motivo, el utilizar uno u otro estimador dependerá de su mayor o menor dificultad computacional. Es decir, dado el elevadísimo número de datos, tenemos que maximizar la eficicencia tanto en lo que se refiere a las necesidades de almacenamiento de la información como a su procesamiento. Por este motivo los algoritmos para la estimación de los parámetros suelen ser definidos en función de estadísticos suficientes. Algunos ejemplos sencillos de estimadores máximo verosímiles se describen a continuación en la Tabla 1.

Tabla 1: Ejemplos de estimadores máximo verosímiles y estadísticos suficientes.

Distribución	$f_X(x\|\Theta)$	Estadísticos Suficientes	Estimadores
Normal(μ_N, σ_N)	$\frac{1}{\sigma_N\sqrt{2\pi}} exp\left(\frac{(x-\mu_N)^2}{2\sigma_N^2}\right)$	$S_1 = \sum_{i=1}^n X_i$ y $S_2 = \sum_{i=1}^n X_i^2$	$\hat{\mu}_N = \frac{S_1}{n}$ y $\hat{\sigma}_N = \sqrt{\frac{S_2}{n} - \hat{\mu}^2}$
Lognormal(μ_L, σ_L)	$\frac{1}{x\sigma_L\sqrt{2\pi}} exp\left(\frac{(log(x)-\mu_L)^2}{2\sigma_L^2}\right)$	$S_1 = \sum_{i=1}^n log(X_i)$ y $S_2 = \sum_{i=1}^n log(X_i)^2$	$\hat{\mu}_L = \frac{S_1}{n}$ y $\hat{\sigma}_L = \sqrt{\frac{S_2}{n} - \hat{\mu}^2}$
Exponencial(λ)	$\lambda exp\,(-\lambda x)$	$S_1 = \sum_{i=1}^n X_i$	$\hat{\lambda} = \frac{n}{S_1}$

En la Tabla 1 los estimadores se corresponden con distribuciones de cola con decrecimiento exponencial o tipo Gumbel. Sin embargo, cuando se analizan pérdidas positiva la distribución podría tener cola con decrecimiento potencial o tipo Fréchet, en estos casosZhang y Nadarajah (2017) obtienen los estimadores de los parámetros de las distribuciones de Pareto tipo I y II, que mostramos en la Tabla 2. Estos autores también describen el algoritmo para la estimación de la distribución de Pareto generalizada.

3.2. Bondad de ajuste, prueba de adecuación para n grande

Cuando se analizan muestras muy grandes, la inferencia clásica tal y como la conocemos no puede aplicarse. Los niveles de significación (p-values) serán prácticamente cero, por tanto,

Tabla 2: Ejemplos de estimadores máximo verosímiles de la Pareto tipo I y II.

Distribución	$f_X(x\|\Theta)$	Estimadores
Pareto Tipo I	$\frac{\alpha\lambda^\alpha}{(x^{\alpha+1}}$	$\hat{\lambda} = min(X_1, ..., X_n)$ y $\hat{\alpha} = n\left[\sum_{i=1}^n log(X_i) - nlog\hat{\lambda}\right]^{-1}$
Pareto Tipo II	$\frac{\alpha\lambda^\alpha}{(x+\lambda)^{\alpha+1}}$	$\hat{\lambda} = min(X_1, ..., X_n)$ y $\hat{\alpha} = n\left[\sum_{i=1}^n log(X_i + \hat{\lambda}) - nlog\hat{\lambda}\right]^{-1}$

tenderemos a rechazar todas las hipótesis nulas. En marzo de 2016 la ASA (*American Statistical Association*) ya planteaba seis principios sobre la interpretación del p-value, los cuáles invalidaban los criterios basados en la comparación de este valor con un límite que en general ha venido siendo igual a 0,05. Los criterios de bondad de ajuste y adecuación de los modelos tienen que basarse en estadísticos sencillos y comparables fácilmente, como lo son los criterios de información estadística como el de Akaike y el criterio bayesiano de Schwarz, que son conocidos, respectivamente, como AIC y BIC y se calculan a partir del valor del logaritmo de la función de verosimilitud definido en (10) como:

$$AIC = 2k - 2l(\hat{\Theta}), \tag{11}$$

y

$$BIC = log(n)k - 2l(\hat{\Theta}). \tag{12}$$

En ambos casos, cuando menor sea el valor del criterio mejor es el ajuste del modelo.

3.3. Estimación no paramétrica

Al contrario que la estimación paramétrica, donde el reto es estimar los parámetros a partir de los cuáles tendremos definidos las funciones asociadas a la forma de la distribución, como son la función de densidad y función de distribución, la estimación no paramétrica plantea retos distintos, dado que su objetivo se centra en una única función. Es decir, los estimadores no paramétricos son distintos para la función de densidad y la función de distribución. En esta sección se describe la estimación núcleo de ambas funciones, junto a los retos que los estimadores plantean en el contexto del big-data.

En la práctica, la estimación no paramétrica en general y la estimación núcleo en particular nos ayudan a tener una visualización previa de la distribución de los datos, pero también pueden utilizarse como herramienta para cuantificar el riesgo sin asumir ningún modelo paramétrico. Ante ello, el gran reto será diseñar algoritmos que nos permitan minimizar el coste computacional de obtención de los resultados, garantizando las propiedades asintóticas de los estimadores.

Analizando los estimadores núcleo desde una perspectiva computacional observamos que en su definición, en cada punto de la función de densidad necesitamos realizar el sumatorio de una función núcleo $k(\cdot)$ de todos los datos, lo que en el contexto del big-data supone una gran

necesidad de cálculo computacional, sobretodo si el dominio de la variable es infinito. Además, en el caso de la estimación núcleo transformada, la obtención de la transformación óptima en ocasiones no es sencilla. Sin embargo, el uso de transformaciones también puede facilitar la obtención de la estimación núcleo cuando tenemos un gran número de observaciones.

3.3.1. Estimación núcleo de la función de densidad

Sea $X_1, ..., X_n$ los n datos observados que se suponen independientes e igualmente distribuidos, procedentes de una variable aleatoria continua X con función de densidad f_X, definida en el conjunto de los números reales $(-\infty < x < +\infty)$. El estimador núcleo de $f_X(x)$, para cualquier punto x del dominio se define del siguiente modo:

$$\hat{f}_X(x) = \frac{1}{nb} \sum_{i=1}^{n} k\left(\frac{x - X_i}{b}\right). \tag{13}$$

La función $k(\cdot)$ se denomina núcleo de la estimación y $b > 0$ es el parámetro de alisamiento o ventana. Tradicionalmente, el núcleo $k(\cdot)$ es una función continua, simétrica respecto al cero y acotada o asintóticamente acotada, que se elige convenientemente para que cumpla las siguientes propiedades:

$$\int_{-\infty}^{+\infty} k(t)\, dt = 1, \int_{-\infty}^{+\infty} tk(t)\, dt = 0, \int_{-\infty}^{+\infty} t^2 k(t)\, dt = \sigma_k^2 > 0$$

$$\text{y } k(t) \geq 0 \ \forall t.$$

Por tanto, $k(\cdot)$, aunque no es estrictamente necesario, suele ser una función de densidad. Algunas de las funciones que pueden utilizarse como núcleo de la estimación se presentan en la Tabla 3.

Tabla 3: Funciones núcleo.

Núcleo	$k(t) =$
Uniforme	$1/2$, si $\lvert t\rvert \leq 1$
Triangular	$(1 - \lvert t\rvert)$, si $\lvert t\rvert \leq 1$
Epanechnikov	$(3/4)(1 - t^2)$, si $\lvert t\rvert \leq 1$
Doble-Ponderado	$(15/16)(1 - t^2)^2$, si $\lvert t\rvert \leq 1$
Triple-Ponderado	$(35/32)(1 - t^2)^3$, si $\lvert t\rvert \leq 1$
Gaussiano	$\left(1/\sqrt{2\pi}\right) e^{-(1/2)t^2}$, $\forall\ -\infty \leq t \leq +\infty$

El hecho de utilizar una función núcleo u otra distinta en el estimador de f_X afecta a sus propiedades matemáticas, en el sentido de que dicho estimador adopta las mismas propiedades de continuidad y derivabilidad que $k(\cdot)$. Por ejemplo, si se utilizara un núcleo Gaussiano, se obtendrá una estimación de la función de densidad continua en todo su dominio y que posee infinitas derivadas. Por otro lado, la selección de una función núcleo entre las que se apuntan en la Tabla 3 afecta muy levemente a las propiedades de sesgo y varianza del estimador núcleo. En función de si el dominio de la función de densidad estimada es acotado o no se recomienda utilizar un núcleo acotado (el de Epanechnikov es el más frecuente) o no (el núcleo Gaussiano es el más frecuente).

Si $f_X(x)$ posee sus dos primeras derivadas continuas, la expresión del sesgo de su estimación núcleo es:

$$E\left[\hat{f}_X(x)\right] - f_X(x) = \frac{1}{2}\sigma_k^2 b^2 f_X''(x) + o(b^2) \tag{14}$$

y su varianza es:

$$V\left[\hat{f}_X(x)\right] = \frac{1}{nb} f_X(x) \int_{-\infty}^{+\infty} k(t)^2\, dt + o\left(\frac{1}{nb}\right), \tag{15}$$

donde $o(\cdot)$ es un término que tiende a cero con el mismo orden que su argumento (véase Silverman, 1986, donde se incluyen las demostraciones correspondientes).

Las expresiones del sesgo y la varianza anteriores permiten deducir que, si se cumple la siguiente relación entre la ventana y el tamaño muestral:

$$b \to 0 \text{ y } nb \to \infty \text{ cuando } n \to \infty,$$

tanto el sesgo como la varianza tienden a cero cuando la muestra tiende a infinito, por lo que el estimador es consistente. Por tanto, cuando se dispone de muchos datos el estimador núcleo de la función de densidad proporcionará una buena aproximación de la realidad. Sin embargo, esta afirmación no es del todo cierta, ya que dependerá de como están distribuidos los datos.

Uno de los principales retos de la estimación núcleo es calcular el valor del parámetro de alisamiento b en (13) que, como ya se había dicho, sirve para controlar el grado de alisamiento de la estimación y/o para delimitar la influencia de las observaciones muestrales en la estimación de la densidad en un punto x. Es decir, cuanto menor sea el valor de b, menos suave es $\hat{f}_X$ (menor es el grado de alisamiento) y mayor la influencia de las observaciones muestrales cercanas al punto x en la estimación de $f_X(x)$. Cuanto mayor es el valor de b, ocurre lo contrario.

Se han propuestos diversos métodos para la obtención del parámetro de alisamiento, todos ellos se basan en la minimización del error entre el resultado de la estimación y la función de desidad teórica. Los criterios utilizados basados en la distancia L^2 son el error al cuadrado integrado (*integrated square error*, ISE) y el error al cuadrado integrado medio (*mean integrated*

square error, MISE) o error cuadrático medio integrado (*integrated mean square error*, IMSE), los cuales se definen del siguiente modo:

$$ISE\left(\hat{f}_X\right) = \int_{-\infty}^{+\infty} \left[\hat{f}_X(x) - f_X(x)\right]^2 dx \tag{16}$$

y

$$\begin{aligned} IMSE\left(\hat{f}_X\right) &= \int_{-\infty}^{+\infty} E\left[\hat{f}_X(x) - f_X(x)\right]^2 dx \\ &= E\left\{\int_{-\infty}^{+\infty} \left[\hat{f}_X(x) - f_X(x)\right]^2 dx\right\} = MISE\left(\hat{f}_X\right). \end{aligned} \tag{17}$$

En general, los métodos de cálculo del parámetro de alisamiento son de dos tipos, los que utilizan validación cruzada para estimar ISE o MISE (ver, por ejemplo, Rudemo, 1982; Bowman, 1984; Jones et al., 1991; Hall et al., 1992) y los métodos *plug-in*, que consisten en aproximar el valor de b que asimptóticamente minimiza MISE (ver, por ejemplo, Silverman, 1986; Park y Marron, 1990; Sheather y Jones, 1991).

Los métodos de cálculo de b basados en validación cruzada suponen un gran esfuerzo computacional, dado que requieren del cálculo de una función de error eliminando cada una de las observaciones. Los métodos *plug-in*, por el contrario, que pueden ser más eficientes computacionalmente, se basan en la optimización numérica de la aproximación de MISE, tal y como se describe a continuación.

Partiendo de que el error al cuadrado medio es la suma de la varianza más el sego al cuadrado del estimador, en este caso se obtiene:

$$MSE\left[\hat{f}_X(x)\right] = \frac{1}{nb} f_X(x) \int_{-\infty}^{+\infty} k(t)^2\, dt + \frac{1}{4}\sigma_k^2 b^4 f_X''(x)^2 + o\left(\frac{1}{nb}\right) + o(b^4), \tag{18}$$

integrando la expresión anterior, la expresión del MISE es:

$$MISE\left(\hat{f}_X\right) = \frac{1}{nb} \int_{-\infty}^{+\infty} k(t)^2\, dt + \frac{1}{4}\sigma_k^2 b^4 \int_{-\infty}^{+\infty} f_X''(x)^2\, dx + o\left(\frac{1}{nb}\right) + o(b^4). \tag{19}$$

Eliminando los términos de orden $o(\cdot)$ de la expresión anterior, se obtiene la aproximación de asintótica de MISE, a la que se denomina A-MISE:

$$A - MISE\left(\hat{f}_X\right) = \frac{1}{nb} \int_{-\infty}^{+\infty} k(t)^2\, dt + \frac{1}{4}\sigma_k^2 b^4 \int_{-\infty}^{+\infty} f_X''(x)^2\, dx. \tag{20}$$

Minimizando la expresión anterior con respecto al parámetro de alisamiento, se obtiene que el valor de b que minimiza asintóticamente MISE y que depende del número de datos n es:

$$b^* = \left(\sigma_k^2\right)^{-\frac{2}{5}} \left[\frac{\int_{-\infty}^{+\infty} k(t)^2\,dt}{\int_{-\infty}^{+\infty} f_X''(x)^2\,dx}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}. \tag{21}$$

Si sustituimos el valor de b^* en la expresión (20) de $A - MISE$ se obtiene que el mínimo asintótico de MISE es:

$$\frac{5}{4} k_2^{\frac{2}{5}} \left[\int_{-\infty}^{+\infty} k(t)^2\,dt\right]^{\frac{4}{5}} \left[\int_{-\infty}^{+\infty} f_X''(x)^2\,dx\right]^{\frac{1}{5}} n^{-\frac{4}{5}}. \tag{22}$$

Por tanto, se deduce que cuanto menor sea el funcional $\int_{-\infty}^{+\infty} f_X''(x)^2\,dx$ menor será el valor asintótico de MISE, por tanto, cuanto más alisada sea f_X menor será este error.

Partiendo de la expresión (21), en general, los métodos *plug-in* consisten en aproximar esta expresión a partir de la estimación directa o no del funcional $\int_{-\infty}^{+\infty} f_X''(x)^2\,dx$. Silverman (1986) propone aproximar el valor de este funcional suponiendo que es igual al que se obtiene con una función de densidad normal con varianza σ^2, el valor del parámetro resultante suponiendo un núcleo gausiano es $b = 1{,}059\hat{\sigma}n^{-1/5}$, donde $\hat{\sigma}$ es un estimador consistente de σ; a esta estrategia se le denomina *rule-of-thumb* y se podría generalizar a cualquier otra densidad distinta a la normal. Este es un modo sencillo y rápido de calcular el parámetro de alisamiento, pero puede llevar a errores si el grado de alisamiento de la función de densidad que se desea estimar no se asemeja al de la normal.

Otro modo de aproximar (21) es mediante la estimación núcleo de $\int_{-\infty}^{+\infty} f_X''(x)^2\,dx$, lo que puede volver a ser computacionalmente largo. Entre estos métodos, el que se ha comprobado que es computacionalmente más eficiente y se aproxima más a la solución óptima b^* en la mayoría de distribuciones es el propuesto por Sheather y Jones (1991). Sin embargo, como se muestra en la Figura 6, que se ha obtenido con el mismo millón de datos utilizados en figuras anteriores de este documento, cuando la función de densidad posee una larga cola derecha, el método de Sheather-Jones no funciona, proporcionando una función excesivamente subalisada en todo su dominio ($b = b = 2{,}3e-05$). Alternativamente, en la Figura7 se muestra el resultado utilizando el criterio *rule-of-thumb* basado en la densidad de la normal. En este caso, dada la forma de la función, este parámetro de alisamiento sobrealisa la función de densidad real en su moda, sin embargo, aún tomando un valor mucho más elevado ($b = 1{,}6396$) que el obtenido con el método de Sheather-Jones, sigue proporcionando un resultado muy sub-alisado en los valores de la cola derecha de la función de densidad.

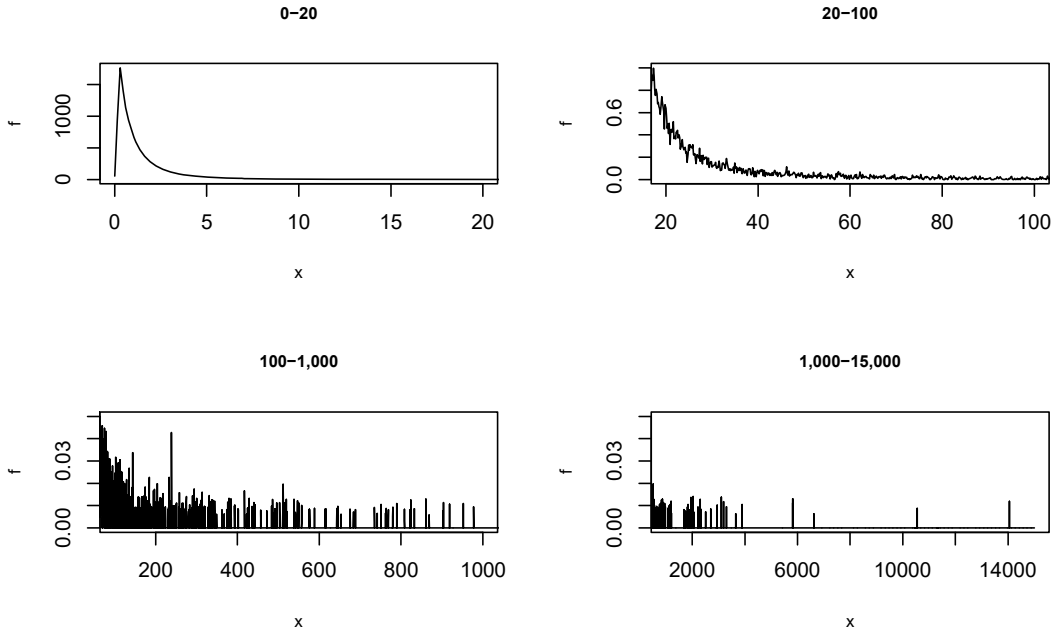

Figura 6: Estimación núcleo de la función de densidad con el parámetro de alisamiento calculado con el método de Sheather-Jones, definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

Un modo de abordar el problema que se plantea en las Figuras 6 y 7, una estrategia es la de transformar los datos de modo que los nuevos datos tengan asociados una función de densidad similar a un modelo fácil de aproximar a partir de la estimación núcleo, por ejemplo, como la función de densidad de la normal. En esta línea, la transformación logarítmica ha sido ampliamente utilizada en economía y finanzas con el objetivo de eliminar asimetrías en los datos. Por ejemplo, si los datos proceden de una distribución log-normal, su logaritmo tendrá distribución normal; otro ejemplo es el de la distribución log-logística, en este caso el logaritmo de la variable tendrá distribución logística. A partir de la estimación núcleo de la función de densidad de la variable transformada, con un cambio de variable podremos deducir la función de densidad de la variable original. A esta estrategia se la denomina estimación núcleo transformada, que se obtiene como se describe a continuación. Sea T una función de transformación con al menos dos derivadas continuas, se define:

$$\widehat{f}_X^T(x) = \frac{T'(x)}{nb} \sum_{i=1}^{n} k\left[\frac{T(x) - T(X_i)}{b}\right] = T'(x)\, \hat{f}_{T(X)}\left[T(x)\right]. \tag{23}$$

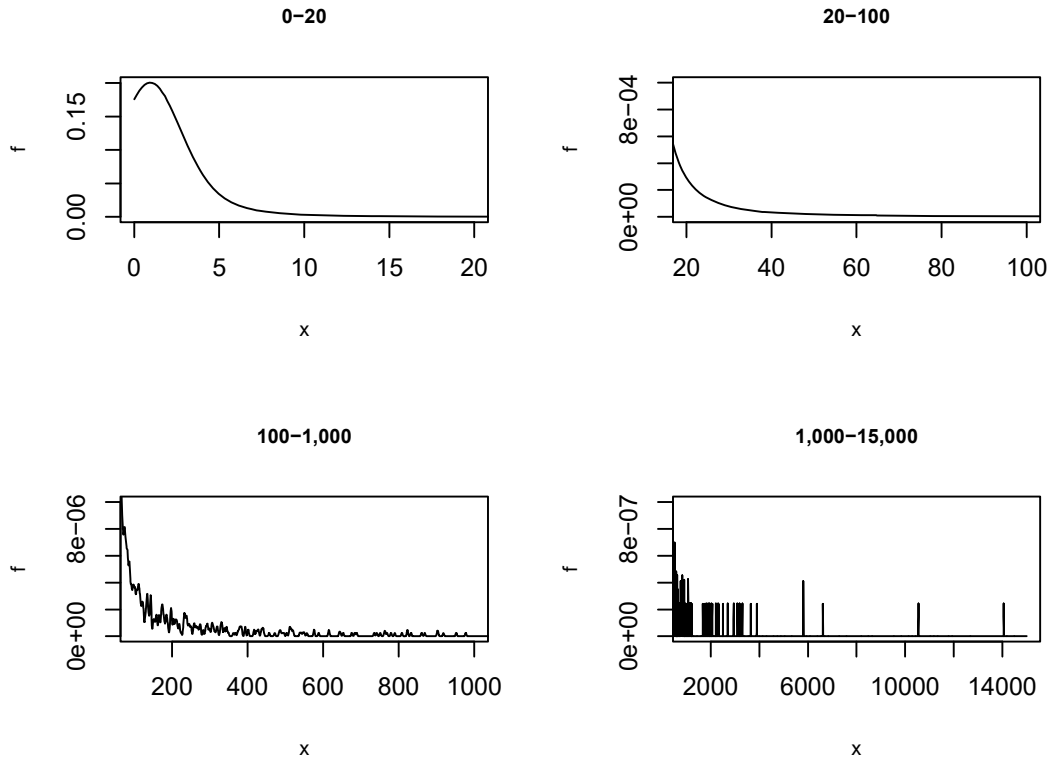

Figura 7: Estimación núcleo de la función de densidad con el parámetro de alisamiento calculado con el método *rule-of-thumb* basado en la densidad de la normal, definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

El estimador núcleo transformado en x equivale al producto de la derivada de la función de transformación y el estimador núcleo de la variable transformada en $T(x)$.

En la Figura 8 se muestra el resultado de la estimación núcleo transformada a partir del logaritmo de los datos utilizados anteriormente en las Figuras 6 y 7, donde el parámetro de alisamiento es el *rule-of-thumb* basado en la densidad de la normal con varianza σ^2, cuyo valor ya ha sido calculado anteriormente para obtener la Figura 7 y es $b = 1{,}6396$. En la Figura 8 se observa una mejora respecto a la anterior, sin embargo, seguimos observando subalisamiento de la curva en los valores más extremos. Esto se debe a que la cola de la distribución asociada a nuestros datos es mucho larga y pesada, de modo que con la transformació logarítmica no se consigue corregir toda la asimetría.

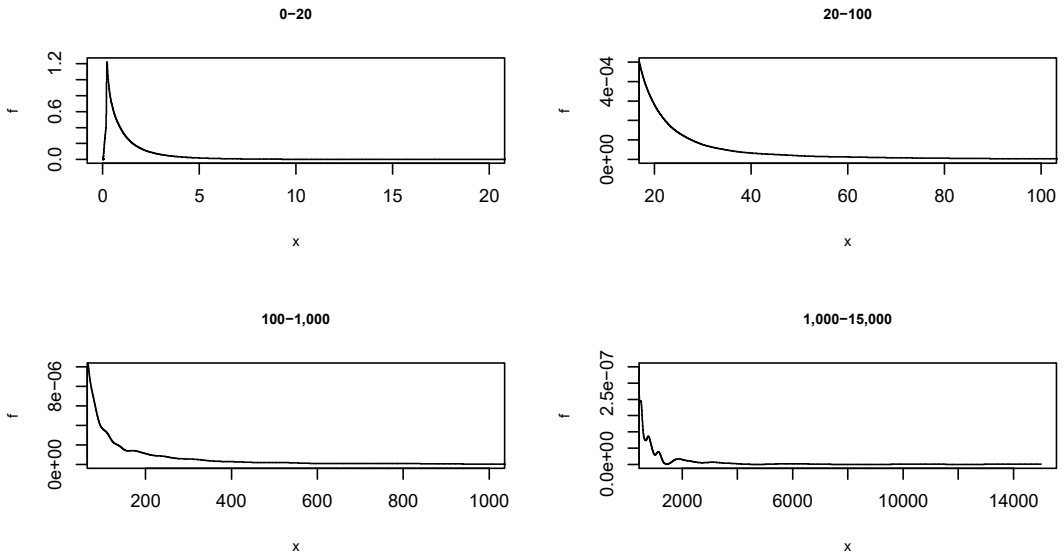

Figura 8: Estimación núcleo transformada de la función de densidad con el parámetro de alisamiento *rule-of-thumb*, definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

Es fundamental encontrar la transformación óptima que permita obtener la estimación núcleo de la función de densidad que minimice el asintóticamente MISE. Con este objetivo existen diversos trabajos que estudian diferentes familias de transformaciones a partir de las cuales se han propuesto estimadores núcleo transformados que han permitido mejorar notablemente los resultados del estimador núcleo cuando la variable posee fuerte asimetría positiva, como ocurre en muchas ocasiones cuando medimos un pérdida económica (véase, por ejemplo, Bolancé et al., 2003; Buch-Larsen et al., 2005; Bolancé et al., 2012).

La forma propuesta para abordar las dificultades computacionales de la estimación núcleo transformada es buscar una estrategia de transformación cuyo resultado sea una variable acotada y fácil de estimar mediante el estimador núcleo. En base a los trabajos de Terrell y Scott (1985) y Terrell (1990), que demuestran que la densidad que minimiza asintóticamente el MISE el estimador núcleo es la de una Beta simétrica cuyos parámetros dependenrán de su dominio acotado en $[0, 1]$ o en $[-1, 1]$, Bolancé et al. (2008); Bolancé (2010) proponen una doble transformación que consiste en aplicar primero una función de distribución paramétrica, la cual puede ser fácil de estimar en el contexto del big-data con un estimador máximo verosímil o de momentos, como por ejemplo la lognormal, y posteriormete aplicar la inversa de la función de distribución de la Beta asintóticamente óptima. Con esto se consigue fijar el valor del parámetro

de alisamiento, dado que supondremos que la función densidad de la variable doble transformada es conocida y coincide con la de la Beta utilizada en la segunta transformación. Además, el hecho de que el dominio de la variable transformada esté acotada nos permite utilizar una función núcleo acotada y a partir del valor conocido del parámetro de alisamiento podremos determinar cuantos datos necesitamos para obtener la estimación en cada punto del dominio. Esta estrategia la desarrollaremos más detalladamente en la siguiente sección dedicada al estimador núcleo de la función de distribución y los cuantiles. Además, diseñar un algoritmo computacionalmente eficiente para implementarla es uno de los retos en los que se está trabajando en la actualidad en este proyecto.

La estimación núcleo también se puede corregir cuando tenermos datos censurados o truncados, la corrección del truncamiento es directa, y es igual que en el caso paramétrico. En el caso de disponer de datos censurados por la derecha o por la izquierda (tras cambiar el signo tendremos censura por la derecha) la corrección se realiza a partir del estimador de Kaplan-Meier definido anteriormete en (8). En concreto, cada elemento dentro del sumatorio que define cualquier estimador tipo núcleo de los que hemos definido quedaría multiplicado por:

$$t_1 = \hat{F}_X^{K-M}(U_{(2)}) \tag{24}$$

$$t_i = \hat{F}_X^{K-M}(U_{(i+1)}) - \hat{F}_X^{K-M}(U_{(i)}) \tag{25}$$

$$t_n = 1 - \hat{F}_X^{K-M}(U_{(n)}), \tag{26}$$

donde las observaciones $U_{(1)}, ..., U_{(n)}$ fueron definidas en (7), teniendo en cuenta que el subíndice entre parétesis indica orden (véase Padgett y McNichols, 1984; Mielniczuk, 1986, para una revisión de las propiedades). Esta misma idea también es aplicable a los estimadores tipo núcleo de la función de distribución que describimos a continuación.

3.3.2. Estimación núcleo de la función de distribución

La estimación de la función de distribución es fundamental en la cuantificación del riesgo, dado que nos permite evaluar la probabilidad de que una pérdida se sitúe por debajo (o, a partir de su contrario, por encima) de un valor muy elevado. Además, a partir de la inversa de esta función obtenemos el valor del cuantil con un determinado nivel de confianza. Este cuantil, con un nivel de confianza cercano a 1 ha sido y sigue siendo ampliamente utilizado como medida de riesgo y se conoce como el valor en riesgo (VaR-*value-at-risk*).

De forma natural, el estimador núcleo de la función de distribución F_X se obtiene integrando el estimador núcleo de la función de densidad. La expresión del estimador se deduce fácilmente mediante el siguiente cambio de variable $t = \frac{u-X_i}{b}$ y equivale a (véase Azzalini, 1981; Reiss,

1981):

$$\begin{aligned}\widehat{F}_X(x) &= \int_{-\infty}^{x} \widehat{f}_X(u)du = \int_{-\infty}^{x} \frac{1}{nb}\sum_{i=1}^{n} k\left(\frac{u-X_i}{b}\right) du \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\frac{x-X_i}{b}} k\left(t\right) dt = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{b}\right),\end{aligned} \tag{27}$$

donde $K(\cdot)$ es la función de distribución de $k(\cdot)$. El parámetro de alisamiento b posee un efecto similar al que tiene en el estimador núcleo de la función de densidad, aunque en este caso el efecto puede ser algo menos evidente a simple vista. En la práctica, al igual que sucedía con el estimador núcleo de la función de densidad, el parámetro de alisamiento depende de la distribución teórica y, por tanto, éste se tendrá que aproximar. En la literatura se han propuesto métodos de obtención de la ventana óptima similares a los utilizados en la estimación núcleo de la función de densidad. Sarda (1993) y Bowman et al. (1998) analizaron el cálculo de b por el método de validación cruzada. Altman y Léger (1995) adaptaron el método de Sheather-Jones al estimador núcleo de la función de distribución. Por otro lado, también puede utilizarse el valor del parámetro de alisamiento con referencia a una distribución conocida (*rule-of-thumb*).

En la Figura 9 se muestra el estimador núcleo de la función de distribución obtenido con los mismos datos utilizados en el apartado dedicado a la función de densidad y utilizando, en la línea continua, el parámetro de alisamiento basado en el método de Sheather-Jones adaptado a los criterios de alisamiento del estimador númcleo de la función de distribución, cuyo valor es $b = 5{,}9e-05$, y utilizando, en la línea discontinua, el parámetro de alisamiento calculado con el criterio *rule-of-thumb* que proporciona un valor del parámetro de alisamiento $b = 0{,}0674$. Si comparamos ambas curvas observamos que las diferencias son apenas perceptibles a simple vista, sin embargo, los parámetros de alisamiento utilizados en ambos casos son bastante distintos. En ambas figuras se puede observar que el estimador núcleo de la función de distribución plantea una problemática similar al estimador de la función de densidad cuando la distribución es asimétrica. El estimador en la cola derecha de la distribución, es bastanta similar a la distribución empírica y, por tanto, tiene una varianza similar.

El estimador núcleo de la función de distribución que se define en (27) tiene muchas similitudes con la expresión de la distribución empírica definida en (1). Simplemente si en (27) remplazamos $K\left(\frac{x-X_i}{b}\right)$ por $I(X_i \leq x)$ se obtiene (1). La principal diferencia entre (1) y (27) es que la distribución empírica solo usa datos con valor inferior o igual a x para obtener la estimación de $F_X(x)$, mientras que la estimación núcleo se obtiene utilizando todos los datos alrededor de x, pero dando más peso a las observaciones que son inferiores a x que a las observaciones que son superiores. Estas diferencias implican que el estimador núcleo esté definido en todos los puntos de dominio de la función de distribución y no únicamente en los observados en la muestra.

Reiss (1981) y Azzalini (1981) demuestran que el sesgo de $\widehat{F}_X(x)$ se aproxima como:

$$E\left[\hat{F}_X(x)\right] - F_X(x) = b^2\frac{\sigma_k^2}{2} f'_X\left(x\right) + o(b^2) \tag{28}$$

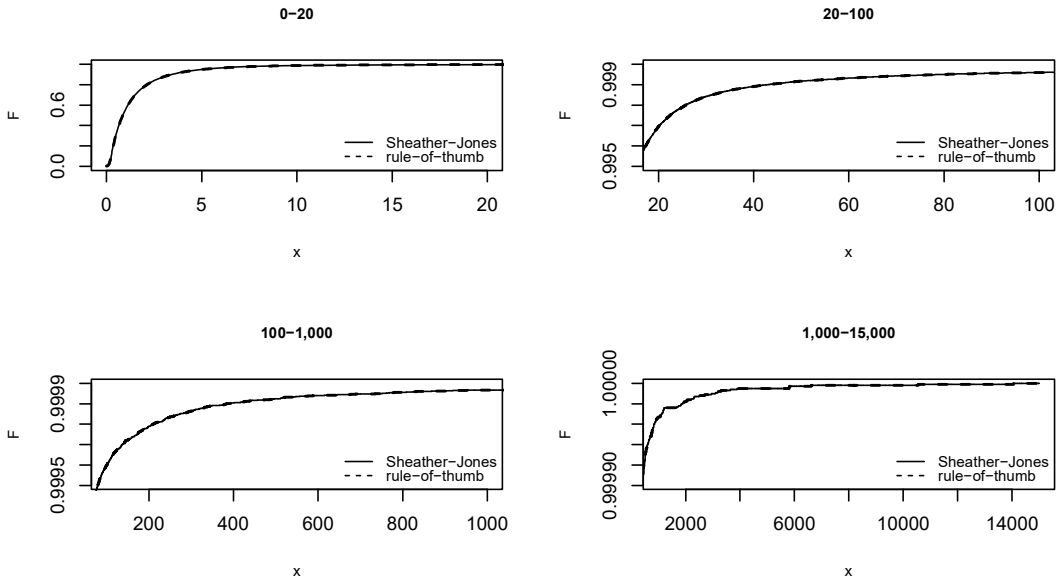

Figura 9: Estimación núcleo de la función de distribución con parámetros de alisamiento calculado con el método de Sheather-Jones (línea continua) y con el *rule-of-thumb* (línea discontinua), definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

y la varianza es:

$$V\left[\widehat{F}_X(x)\right\} = \frac{F_X(x)\left[1 - F_X(x)\right]}{n} - f_X(x)\frac{b}{n}\left[1 - \int K^2(t)\,dt\right] + o\left(\frac{b}{n}\right). \tag{29}$$

Las expresiones (28) y (29) indican que si $b \to 0$ con $n \to \infty$, el estimador núcleo de la función de distribución es consistente.

Comparando la expresión (29) con la varianza de la aproximación empírica definida en (2), el estimador núcleo de la función de distribución tiene menor varianza. Sin embargo, añade cierto sesgo a la estimación que, tal y como se puede observar en (28), tiende a cero si el tamaño de la muestra es grande.

Al igual que para el estimador núcleo de la función de densidad, puede obtenerse un valor del parámetro de alisamiento que minimice una aproximación asintótica de:

$$MISE\left[\widehat{F}_X(x)\right] = E\left\{\int \left[F_X(x) - \widehat{F}_X(x)\right]^2 dx\right\}.$$

El valor asintótico de MISE conocido como A-MISE equivale a la suma de la aproximación asintótica de la varianza integrada más el sesgo asintótico al cuadrado integrado, esto es:

$$\begin{aligned} A-MISE\left[\widehat{F}_X\right] &= \frac{1}{n}\int F_X(x)\left[1-F_X(x)\right]dx \\ &- \frac{1}{n}b\int K\left(t\right)\left[1-K\left(t\right)\right]dt+\frac{\sigma_k^4}{4}b^4\int\left[f'_X\left(x\right)\right]^2dx. \end{aligned} \tag{30}$$

La minimización de (30) respecto a b da como resultado el parámetro de alisamiento que minimiza asintóticamente el valor de MISE, que resulta ser:

$$b^*=\left(\frac{\int K\left(t\right)\left[1-K\left(t\right)\right]dt}{\sigma_k^4\int\left[f'_X\left(x\right)\right]^2dx}\right)^{\frac{1}{3}}n^{-\frac{1}{3}}. \tag{31}$$

Los errores en la estimación núcleo de la función de distribución son similares a los que se analizaban en el caso de la función de densidad. De nuevo, el estimador núcleo es similar a la distribución empírica en las zonas donde la información muestral es escasa. Ante ello, se han adaptado las estrategias de transformación utilizadas en la estimación núcleo de la función de densidad a la estimación núcleo de la función de distribución. La idea parte de un resultado similar al obtenido en la expresión (22) para la función de densidad. Concretamente, en este caso, substituyendo el parámetro de alisamiento definido en (21) en la expresión (30) de A-MISE se obtiene el valor asintóticamente mínimo de MISE:

$$\begin{aligned} &\frac{1}{n}\int F_X(x)\left[1-F_X(x)\right]dx \\ -\ &\frac{\left\{\int K\left(t\right)\left[1-K\left(t\right)\right]dt\right\}^{\frac{1}{3}}-\frac{1}{4}\left\{\int K\left(t\right)\left[1-K\left(t\right)\right]dt\right\}^{\frac{4}{3}}}{\sigma_k^{\frac{4}{3}}\left\{\int\left[f'_X\left(x\right)\right]^2dx\right\}^{\frac{1}{3}}}n^{-\frac{4}{3}}. \end{aligned} \tag{32}$$

De la expresión anterior se deduce que, en el caso del estimador núcleo de la función de distribución, el valor asintóticamente óptimo de MISE depende directamente del funcional $\int_{-\infty}^{+\infty}f'_X(x)^2\,dx$. Por tanto, se puede plantear la misma estrategia de doble transformación ampliamente estudiada para el estimador núncleo de la función de densidad (KIBMCE), adaptada a la minimización del nuevo funcional, que en este caso depende de la primera derivada de la función de densidad, lo que equivale a la segunda derivada de la función de distribución. A continuación, se describen algunas propiedades que se han analizado en el contexto de la estimación núcleo transformada de la función de distribución. Por conveniencia, a partir de este momento se supondrá que la variable aleatoria analizada toma valores positivos y es asimétrica a la derecha.

Sea $T(\cdot)$ una función de transformación con al menos una derivada continua, dado que por definición $F_X(x) = F_{T(X)}(T(x))$, el estimador núcleo transformado de la variable original es igual al estimador núcleo de la variable transformada, es decir:

$$\widehat{F}_X^T(x) = \frac{1}{n}\sum_{i=1}^{n} K\left[\frac{T(x) - T(X_i)}{b}\right] = \widehat{F}_{T(X)}\left[T(x)\right]. \tag{33}$$

Como se describía anteriormente, son muchos los trabajos que han estudiado el uso de transformaciones en el contextos del estimador núcleo de la función de densidad (Wand et al., 1991; Ruppert y Cline, 1994; Bolancé et al., 2003; Buch-Larsen et al., 2005; Bolancé et al., 2008; Bolancé, 2010). Sin embargo, son pocos los estudios que analizan el uso de transformaciones en el estimador núcleo de la función de distribución (Alemany et al., 2013; Swanepoel y Van Graan, 2005). En general, el tipo de transformaciones propuestas son similares a las utilizadas en el estimador núcleo de la función de densidad: paramétricas o no paramétricas y que son función de distribución o no.

En Alemany et al. (2013) se demostró que el sesgo y la varianza del estimador núcleo transformado se aproximan como:

$$Sesgo\left[\widehat{F}_X^T(x)\right] \sim \frac{1}{T'(x)}\left[1 - \frac{\frac{f_X(x)}{f'_X(x)}}{\frac{T'(x)}{T''(x)}}\right]\frac{1}{2}f'_X(x)\,\sigma_k^2 b^2 \tag{34}$$

y

$$V\left[\hat{F}_X^T(x)\right] \sim \frac{F_X(x)\left[1 - F_X(x)\right]}{n} - \frac{f_X(x)}{T'(x)}\frac{b}{n}\left[1 - \int K^2(t)\,dt\right]. \tag{35}$$

De la expresión del sesgo se deduce que si $T = F_X$ entonces $E\left(\widehat{F}_X^T(x)\right) = F_X(x), \forall x \in \mathbb{R}^+$. Sin embargo, en la práctica utilizar una transformación que es función de distribución implica que la variable transformada es $Uniforme(0,1)$, que es una distribución que no cumple las propiedades de derivabilidad requeridas por el estimador núcleo, su función de densidad no posee al menos una primera derivada distinta de cero, y por tanto no se obtendrán los resultados esperados.

De nuevo, de las expresiones del sesgo y la varianza del estimador núcleo se deduce que puede existir una transformación que mejore las propiedades del estimador que pasa por la estrategia de una doble transformación. Para ello hay que encontrar aquella densidad que minimice el funcional $\int_{-\infty}^{+\infty} f'_X(x)^2\,dx$.

Terrell (1990) demostró que la densidad de la $Beta(3,3)$ con dominio $[-1,1]$ minimiza $\int_{-\infty}^{+\infty} f'_X(x)^2\,dx$, siendo sus funciones de densidad y de distribución:

$$h(x) = \frac{15}{16}\left(1 - x^2\right)^2,\ -1 \leq x \leq 1 \tag{36}$$

y

$$H(x) = \frac{3}{16}x^5 - \frac{5}{8}x^3 + \frac{15}{16}x + \frac{1}{2},\ -1 \leq x \leq 1. \tag{37}$$

El estimador núcleo doble transformado se define como:

$$\widehat{F}_X^{DT}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left\{\frac{H^{-1}\left[T(x)\right] - H^{-1}\left[T(X_i)\right]}{b}\right\}, \tag{38}$$

donde T es la función de distribución asimétrica a la derecha y que tiene un comportamiento similar al de la distribución de los datos. Por ejemplo, podría utilizarse una log-normal o una log-logística.

El parámetro de alisamiento que se utiliza en (38) puede calcularse directamente sustituyendo en el valor de b que minimiza A-MISE definido en (31) la función de densidad $h(\cdot)$ definida en (36). El núcleo que se propone utilizar es el de Epanechnikov, dado que la $Beta(3,3)$ es una densidad definida en $[-1,1]$, el resultado de este parámetro de alisamiento es $n = 3^{1/3}n^{-1/3}$.

Se han desarrollado las expresiones del sesgo y la varianza del estimador núcleo transformado teniendo en cuenta la doble transformación. Siendo $\widehat{F}_X^{DT}(x)$ el estimador núcleo doble transformado se han obtenido las siguientes expresiones para el sesgo y la varianza, respectivamente:

$$\begin{aligned} & E\left[\widehat{F}_X^{DT}(x)\right] - F_X(x) \\ \approx\ & \frac{\sigma_k^2}{2}b^2\left[\frac{m'\left\{M^{-1}\left[T(x)\right]\right\}}{m\left\{M^{-1}\left[T(x)\right]\right\}}f_X(x) + \frac{f_X'(x)}{T'(x)}m\left\{M^{-1}\left[T(x)\right]\right\}\left(1 - \frac{\frac{f_X(x)}{f_X'(x)}}{\frac{T'(x)}{T''(x)}}\right)\right] \end{aligned} \tag{39}$$

y

$$\begin{aligned} & V\left[\widehat{F}_X^{DT}(x)\right] \\ \approx\ & \frac{F_X(x)\left[1 - F_X(x)\right]}{n} - \frac{f_X(x)}{T'(x)}m\left\{M^{-1}\left[T(x)\right]\right\}\frac{b}{n}\left(1 - \int K^2(t)\,dt\right). \end{aligned} \tag{40}$$

De la expresión del sesgo en (39) se deduce que éste, además de depender de la diferencia entre las funciones de densidad y de su primera derivada asociada a la distribución teórica y a la utilizada como primera transformación, respectivamente, también depende de la función de densidad y de su primera derivada asociada a la Beta utilizada en la segunda transformación. El primer término de la suma en la expresión (39) será positivo o negativo en función de la derivada de $m(\cdot)$ en (36); sin embargo, el segundo término dependerá del signo de la derivada

de la densidad teórica y del término:

$$\left(1 - \frac{\frac{f_X(x)}{f'_X(x)}}{\frac{T'(x)}{T''(x)}}\right).$$

Por tanto, según se seleccione la primera transformación $T(\cdot)$, se podrá añadir más sesgo o, por el contrario, ambos términos podrán compensarse resultando en una reducción del sesgo.

Destacar también que cuando $b \to 0$ con $n \to \infty$ tanto el sesgo como la varianza del estimador núcleo doble transformado definido en (38) tienden a cero, aunque en función de la diferencia entre T y F_X el sesgo tenderá a cero en una proporción más o menos grande, es decir, los términos que multiplican a b^2 y n en las expresiones del sesgo y la varianza definidas en (39) y (40), respectivamente, serán más o menos reducidos. Aunque en un contexto de big-data la magnitud de estos términos no debería ser un problema, en la actualidad se está investigando la mejora que se produce en la estimación núcleo doble transformada cuando utilizamos algunas correcciones del sesgo en el contexto de la estimación núcleo propuestas en la literatura (Kim et al., 2006).

En la Figura 10 se muestra el estimador núcleo doble transformado (línea continua), utilizando como primera transformación la log-logística, dado que tenemos una distribución de cola más pesada que la lognormal, y como segunda transformación la inversa de la función de distribución de la Beta (H) definida en (37), también representamos la estimación núcleo obtenida con el parámetro de alisamiento calculado con el criterio *rule-of-thumb* sin doble transformación (línea discontinua). Comparando ambas curvas observamos el comportamiento más alisado de la estimación núcleo transformada, sobretodo en los valores más extremos de la variable.

De cara al análisis del riesgo, el analista estará interesado en la probabilidad acumulada en la cola derecha de la distribución. En este sentido, dado que la estimación núcleo doble transformada proporciona una variable acotada y con una función de distribución que será similar a la H definida en (37) y un parámetro de alisamiento conocido, podremos acotar el problema teniendo en cuenta las probabilidades acumuladas en esta función de distribución.

3.4. Bases para el diseño de un algoritmo rápido para estimar la función de distribución y el cuantil con una nueva observación

Sea $\hat{F}_X^{(n)}(x)$ un estimador núcleo de la función de distribución en el punto x obtenido con n observaciones. Definiremos $\hat{F}_X^{(n+1)}(x)$ como el estimador obtenido con $n+1$ datos. Si observamos la expresión del estimador núcleo de la función de distribución, por ejemplo en la expresión (27), y tenemos en cuenta que el parámetro de alisamiento depende de n, vemos que esto supone

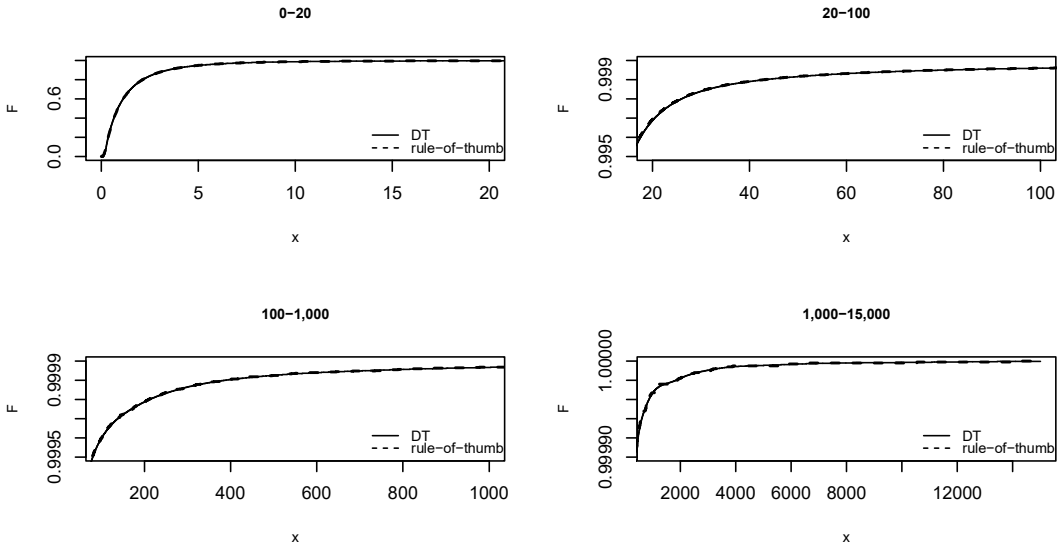

Figura 10: Ejemplo de estimación núcleo doble transformada con T log-logística (línea continua) y estimación núcleo con parámetro de alisamiento *rule-of-thumb* (línea discontinua), definida en $[0, 20)$, $[20, 100)$, $[100, 1000)$ y $[1900, 15000)$.

recalcular $n + 1$ pesos y promerdiarlos, lo que en el contexto del *big-data* puede suponer un elevado coste computacional.

Sin embargo, el problema computacional anterior puede acotarse de dos formas:

- Utilizando núcleos acotados como el de Epanechnikov, lo cual apenas afectará a las propiedades asintóticas del estimador.
- Utilizando parámetros de alisiento en base a una distribución conocida, de modo que dado que dicho parámetro depende de n, su cálculo sea directo $b_n = c \cdot n^{-1/3}$, donde c es una constante conocida.

Teniendo en cuenta los dos puntos anteriores y la expresión de la estimación núcleo, cuando llega una observación nueva, a la que denominamos X_{n+1}, el nuevo estimador de $\hat{F}_X^{(n+1)}(x)$, se obtiene mediante el siguiente proceso:

- Si $\left|\frac{x - X_{n+1}}{b_{n+1}}\right| \geq 1$, entonces poemos aproximar $\hat{F}_X^{(n+1)}(x) \approx \frac{n}{n+1}\hat{F}_X^{(n)}(x)$ asumiendo que con muchos datos $b_n \approx b_{n+1}$, donde b_{n+1} es el parámetro de alisamiento calculado con $n + 1$ datos.

- Si $\left|\frac{x-X_{n+1}}{b_{n+1}}\right| \leq 1$, entonces la corrección se realiza en dos pasos:

  1. Localizar todas las observaciones en el conjunto $M_1 = \left\{X_i \middle| \left|\frac{x-X_i}{b_{n+1}}\right| \leq 1\right\}$.
  2. Calcular el estimador en función de si asumimos $b_n \neq b_{n+1}$ o no:
     - Si modificamos el valor del parámetro de alisamiento tenemos que calcular $F_X^{(n+1)}(x) = \frac{1}{n+1}\sum_{X_i \in M_1} K\left(\frac{x-X_i}{b_{n+1}}\right)$.
     - Si asumimos que $b_n \approx b_{n+1}$ y no modificamos el valor del parámetro de alisamiento el cálculo es directo: $F_X^{(n+1)}(x) = \frac{n}{n+1}\hat{F}_X^{(n)}(x) + \frac{1}{n+1}K\left(\frac{x-X_{n+1}}{b_n}\right)$.

La dificultad fundamental asociada al proceso de estimación anterior es el cálculo del valor de c para obtener el valor del parámetro de alisamiento. El estimador basado en la doble transformación definido en (38), con el cual conseguimos que los datos transformados se comporten como $Beta\,(3,3)$, cuyas funciones de densidad y de distribución se definen, respectivamente, en las expresiones (36) y (37), permite determinar que con el núcleo de Epanechnikov $c = 3^{1/3}$.

Si el objetivo es estimar el cuantil de la distribución $Q_X(p)$, donde p es una probabilidad y cuyo estimador núcleo puede obtenerse directamente a partir de la inversa del estimador núcleo de la función de distribución: $\hat{Q}_X^{(n)}(p) = \hat{F}_X^{(n)^{-1}}(p)$. Dada la expresión del estimador núcleo de la función de distribución, su inversa no puede calcularse de forma exacta y, por tanto, la tendremos que obtener de forma numérica mediante un algoritmo como el de Newton (véase Azzalini, 1981). Esto implica que el procedimiento definido para la estimación de $\hat{F}_X^{(n+1)}(x)$ no es directamente aplicable a la estimación del cuantil cuando disponemos de una nueva observación, dado que en la práctica $F_X^{(n)}(x) \neq F_X^{(n+1)}(x)$ y por tanto hay que volver a aplicar Newton para obtener una nueva inversa.

Para la estimación del cuantil podemos tener en cuenta que en el contexto de la cuantificación de riesgo estamos interesados en cuantiles extremos con un valor de p cercano a 1, podemos acotar el problema a una parte de la distribución, por ejemplo, podemos asumir que en general $p \geq 0{,}95$. Si asumimos un núcleo acotado y un parámetro de alisamiento dado que depende del número de datos, podemos definir el conjunto de observaciones de la cola que necesitamos para obtener nuestro estimador núcleo del cuantil con $p \geq 0{,}95$ como $M_2 = \left\{X_i \middle| \frac{Q_X(0{,}95)-X_i}{b_n} \leq 1\right\}$, de modo que reducimos considerablemente el número de datos que necesitamos para obtener el estimador núcleo del cuantil extremo. La dificultad para definir este conjunto es determinar el valor de $Q_X(0{,}95)$. Podríamos utilizar el cuantil empírico, incluso un estimador núcleo clásico basado en la inversa del estimador núcleo de la distribución o calquier otro estimador del cuantil basado en el estimador núcleo de la regresión (véase Sheather y Marron, 1990, para una revisión). Sin embargo, cuando la cola derecha de la distribución es larga y pesada todos estos

estimadores son similares a la distribución empírica, no tienen una forma totalmente alisada y dependen excesivamente de la información muestral disponible. Como se ha mostrado anteriormente en las Figuras 9, aún disponiendo de datos masivos, si la distribución es fuertemente asimétrica a la derecha, los datos observados en la cola derecha de la distribución no son capaces de dibujar una forma alisada de la misma; por tanto, lo que suele suceder es que la presencia o no de una única observación en la parte de la cola derecha afecta considerablemente al valor del cuantil estimado a partir de la empírica o de un estimador núcleo clásico. El uso de la doble transformación, que nos permite trasladar la distribución original que es desconocida a una $Beta\,(3,3)$, es una forma más eficiente que la distribución empírica y la estimación núcleo clásica de abordar el problema.

Agradecimientos

La autora agradece el apoyo de la Fundación BBVA.

Referencias

Alemany, R., Bolancé, C., y Guillen, M. (2013). A nonparametric approach to calculating value-at-risk. *Insurance: Mathematics and Economics*, 52:255–262.

Altman, N. y Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46:195–214.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68:326–328.

Bahraoui, Z., Bolancé, C., Pelican, E., y Vernic, R. (2015). On the bivariate sarmanov distribution and copula. an application on insurance data using truncated marginal distributions. *SORT-Statistics and Operations Research Transactions*, 39:209–230.

Bolancé, C. (2010). Optimal inverse beta(3,3) transformation in kernel density estimation. *SORT-Statistics and Operations Research Transactions*, 34:223–237.

Bolancé, C., Guillen, M., Gustafsson, J., y Nielsen, J. P. (2012). *Quantitative Operational Risk Models*. Chapman & Hall/CRC Finance Series, London.

Bolancé, C., Guillen, M., y Nielsen, J. P. (2003). Kernel density estimation of actuarial loss functions. *Insurance: Mathematics and Economics*, 32:19–36.

Bolancé, C., Guillen, M., y Nielsen, J. P. (2008). Inverse beta transformation in kernel density estimation. *Statistics& Probability Letters*, 78:1757–1764.

Bowman, A., Hall, P., y Prvan, T. (1998). Bandwidth selection for smoothing of distribution function. *Biometrika*, 85:799–808.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360.

Bowman, A. W. y Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

Buch-Larsen, T., Guillen, M., Nielsen, J. P., y Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, 39:503–518.

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

Hall, P., Marron, J. S., y Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, 92:1–20.

Jones, M. C., Marron, J. S., y Park, B. U. (1991). A simple root n bandwidth selector. *The Annals of Statistics*, 19:1919–1932.

Kaplan, E. L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457 – 481.

Kim, C., Kim, S., Park, M., y Lee, H. (2006). A bias reducing technique in kernel distribution function estimation. *Computational Statistics*, 21:589–601.

Kotz, S. y Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.

Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *Annals of Statistics*, 14:766–773.

Padgett, W. J. y McNichols, D. T. (1984). Nonparametric density estimation from censored data. *Communications in Statistics - Theory and Methods*, 13:1581–1611.

Park, B. U. y Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85:66–72.

Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8:116–119.

Reiss, R.-D. y Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other Fields*. Springer, New York.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9:65–78.

Ruppert, D. R. y Cline, D. B. H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformation. *Annals of Statistics*, 22:185–210.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference*, 35:65–75.

Sheather, S. J. y Jones, M. C. (1991). A reliable data-based bandwidth selection method form kernel density estimation. *Journal of the Royal Statistical Society. Serie B*, 53:683–690.

Sheather, S. J. y Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85:410–416.

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

Swanepoel, J. W. H. y Van Graan, F. C. (2005). A new kernel distribution function estimator based on a nonparametric transformation of the data. *Scandinavian Journal of Statistics*, 32:551–562.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85:270–277.

Terrell, G. R. y Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association*, 80:209–214.

Tsay, R. S. (2016). Some methods for analyzing big dependent data. *Journal of Business & Economic Statistics*, 34:673–688.

Wand, P., Marron, J. S., y Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, 86:343–361.

Zhang, Y. y Nadarajah, S. (2017). Flexible heavy tailed distributions for big data. *Annals of Data Science*, 4:421–432.

UBIREA
Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics
Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona

WEBSITE: www.ub.edu/irea/ • **CONTACT:** irea@ub.edu