Estadística Española • Volumen 56, número 184 / 2014, pp. 127-145

Estimadores de Truncamiento

Juan Fco. Ortega Dato
Facultad de CC. Económicas y Empresariales de Albacete
Universidad de Castilla-La Mancha

Resumen

Con objeto de proteger las estimaciones de posibles contaminaciones por observaciones atípicas (outliers), se han desarrollado estimadores que no son influenciados por estas observaciones; los llamados estimadores robustos.

En este trabajo se proponen unos nuevos estimadores robustos de posición y escala, tanto para el caso univariante como multivariante, denominados *estimadores de truncamiento*. De ellos se estudian sus principales propiedades y cómo construirlos para que sean utilizados en diferentes escenarios reales. Concretamente, se proponen procedimientos para su perfecta definición sobre muestras que procedan de distribuciones normales o de distribuciones t-Student.

Palabras clave: Estimadores robustos, Observaciones atípicas, Outliers.

Clasificación: AMS: 62F35

Trimmed Estimators

Abstract

Outliers are contaminate observations that severely modify usual estimators. To overcome this problem, robust estimators have been introduced in the literature.

In this paper we propose new families of location and scale robust estimators for univariate and multivariate distributions. We call these estimators *trimmed estimators*. We study their main properties and propose procedures to define all their elements in different situations. More concisely, we propose procedures to be used with real data in normal and t-Student distributions.

Keywords: Robust estimators, Outliers.

AMS Classification: 62F35

1. Introducción

Un primer problema que se nos puede plantear al estimar los parámetros de un modelo a partir de una muestra, es que ésta puede contener observaciones contaminantes que enturbian y nos proporcionan información errónea. A estas observaciones contaminates se les denomina observaciones atípicas u outliers (Barnett y Lewis, 1994).

Para proteger las estimaciones de las posibles observaciones atípicas, se han definido y estudiado en la literatura diferentes tipos de estimadores resistentes a ellas, conocidos como estimadores robustos (Huber, 1964; Andrews y otros, 1972; Hampel, 1974; Rousseeuw, 1984; Maronna, Martin y Yohai, 2006).

Los estimadores robustos son poco usados en la práctica, quizá por desconocimiento o porque en algunos casos son difíciles de calcular e implementar en el ordenador. En los casos en los que el investigador decide usar estos estimadores, los elige de entre los más conocidos, incluso en algunos casos el definido por él mismo, o su elección se basa en las propiedades de robustez del estimador, es decir, su resistencia ante contaminaciones. En Rousseeuw y Leroy (1987) se realiza una revisión de los diferentes estimadores robustos de la literatura y de las propiedades de robustez definidas para ellos.

La media truncada (traducción de trimmed mean) es un estimador robusto de posición univariante dentro de la familia de los L-Estimadores (Bickel, 1973). Ésta se define como la media de las observaciones de la muestra tras haber eliminado un porcentaje de observaciones. Usando la estructura de los estimadores clásicos basados en la media (varianza, covarianza y matriz de varianzas-covarianzas) y sustituyendo ésta por la media truncada, es posible construir unas familias de estimadores de posición y escala con buenas propiedades de robustez que llamaremos *estimadores de truncamiento*.

En este trabajo se proponen las familias de estimadores de truncamiento, tanto para el caso univariante como multivariante, estudiando sus principales propiedades y determinando sus elementos para su uso en diferentes escenarios. Nos centraremos, particularmente, en el desarrollo de estos estimadores para el caso de muestras que proceden de distribuciones normales o distribuciones t-Student.

El estudio realizado se organiza de la siguiente manera. En la Sección 2 se presentan los estimadores de truncamiento, tanto en el caso univariante como multivariante. En la siguiente sección, Sección 3, se proponen unos métodos para determinar, en la práctica y de manera general para cualquier distribución de la que procedan los datos, todos los elementos necesarios para el uso de estos estimadores. En la Sección 4 se presenta un estudio que proporciona cómo determinar dichos elementos en muestras de distribuciones normales y en muestras t-Student. Por último, en la Sección 5 realizaremos unos comentarios finales a modo de conclusiones.

2. Estimadores de truncamiento

La media truncada es un estimador de posición bien conocido de la literatura estadística (ver, por ejemplo, Maronna, Martin y Yohai, 2006). Este estimador es una alternativa robusta a la media muestral que usa parte de la muestra en estudio para inferir información

sobre los parámetros poblacionales. Partiendo de la estructura de los estimadores clásicos de posición y escala basados en la media muestral, y sustituyendo ésta por medias truncadas, es posible definir unas familias de estimadores con buenas propiedades y comportamientos ante observaciones atípicas.

A continuación presentamos y estudiamos los citados estimadores, basados en truncamientos, tanto para el caso univariante como multivariante.

2.1 Caso univariante

La media truncada es realmente un elemento de una familia de estimadores de posición, donde el resto de elementos se definen utilizando un número real α que denominaremos *nivel de truncamiento*. Así, la media truncada a nivel de truncamiento α (α_*Truncada*) se construye como la media de las observaciones de la muestra eliminando el *α%* de las observaciones mayores y el *α%* de las menores.

De manera más rigurosa, dada una muestra $x=\{x_1,x_2,\ldots,x_n\}$ y un valor $\alpha\in[0,50)$, α_*Truncada* sobre *x* se denota y define por:

$$T_\alpha(x)=\frac{1}{n-2a}\sum_{i=a+1}^{n-a}x_{[i]} \qquad [1]$$

con $a=Int(\alpha n/100)$, donde *Int* denota parte entera, y $x_{[i]}$ la observación en la posición i-ésima de una ordenación de menor a mayor de los elementos de *x*. Así, $T_0(x)$ es la media muestral de *x* y si α tiende a 50, entonces *α_Truncada* tiende a la mediana de *x*, por lo que, tomando por convenio que $T_{50}(x)$ es la mediana muestral de *x,* podemos definir las *α_Truncadas* para todo α en el intervalo cerrado [0,50].

Estos estimadores tienen buenas propiedades de robustez, como: el ser, para cualquier α, estimadores insesgados de la media en variables con distribuciones simétricas; ser equivariantes, es decir, cumplir que $T_\alpha(sx+t)=sT_\alpha(x)+t \quad \forall s,t\in R$; para tamaños muestrales *n*, tener punto de ruptura de valor *(a+1)/n*, y punto de ruptura asintótico de valor α/100; y el tener curva de sensibilidad acotada, siempre que $a\geq 1$. Por otra parte, su eficiencia bajo normalidad (considerando eficiencia relativa con respecto a la media) puede ser aproximada mediante la forma lineal (100−0,7230α)% para $\alpha\in[0,50]$, por lo que los elementos de esta familia tienen una eficiencia entre el 64% de la mediana y el máximo (el 100%) de la media.

Para el caso de estimadores de escala, en Ortega (2004) se propone una familia de estimadores usando truncamientos, al sustituir medias por medias truncadas en la conocida varianza. De esta forma, dada la muestra *x* como antes, que suponemos procede de una variable *X*, los niveles de truncamiento $\alpha, \beta\in[0,50]$ y las medias truncadas definidas en [1], se denota y define *αβ_Truncada* sobre *x* como:

$$T^2_{\alpha\beta}(x) = C_{\alpha\beta}(X)\, T_\beta\left(\left\{\left(x_i - T_\alpha(x)\right)^2\right\}_i\right) \qquad [2]$$

donde $C_{\alpha\beta}(X)$ es un coeficiente de consistencia, positivo, dependiente de los niveles α y β y de la distribución de *X*, del que supondremos por convenio que es invariante ante cambios de posición y escala, y con $C_{0,0}(X)=1$.

En estas condiciones, los extremos de la familia *αβ_Truncadas* son: la varianza, para el caso de $\alpha=\beta=0$; y para $\alpha=\beta=50$ la *mediana de las distancias al cuadrado a la mediana*, concepto denotado y definido por:

$$MDM(x) = T^2_{50,50}(x) \qquad [3]$$

Es fácil comprobar que la raíz cuadrada de *MDM* es muy similar, salvo por la paridad del tamaño muestral, al estimador conocido por *MAD* (desviación absoluta a la mediana), definido por $MAD(x)=C_m\, med_i\, \{|x_i - med(x)|\}$, donde C_m es también un coeficiente de consistencia y *med* denota mediana (ver Rousseeuw y Croux, 1993 o Hampel y otros, 1986).

De las propiedades de robustez de los elementos de la familia αβ_*Truncadas* se pueden destacar que: son equivariantes, es decir, $T_\alpha\ (sx+t)=sT_\alpha(x)+t\ \ \forall s,t\in R$; que su punto de ruptura es de valor $Min\{(a+1),(b+1)\}$, con $a=Int(\alpha n/100)$ y $b=Int(\beta n/100)$ en muestras de tamaño *n*, y su punto de ruptura asintótico es $Min\{\alpha/100,\beta/100\}$; y que tienen curva de sensibilidad acotada siempre que $Min\{a,b\}\geq 1$. Por otro lado, su eficiencia bajo normalidad (considerando de nuevo eficiencia relativa, ahora con respecto a la varianza) está, dependiendo de los niveles de truncamiento considerados, entre el 38,35% de *MDM* y el máximo del 100% de la varianza.

Las definiciones [1] y [2] sobre muestras pueden ser extendidas al caso de variables, igual que la media y varianza muestrales a las de esperanza y varianza poblacionales. Para ello, siendo *X* una variable con función de densidad *f*(*x*) y función de distribución *F*(*x*), α_*Truncada* sobre *X* para $\alpha\in[0,50]$ se define de la forma:

$$T_\alpha(X) = \frac{100}{100-2\alpha}\int_{a1}^{a2} x f(x)\,dx \qquad [4]$$

donde $F(a_1)=\alpha/100$ y $F(a_2)=1-(\alpha/100)$. De esta forma, $T_0(X)$ es la esperanza de *X* y $T^2_{0,0}(X)$ su varianza, y si α tiende a 50 entonces α_*Truncada* tiende a la mediana poblacional, por lo que, tomando por convenio que $T_{50}(X)=F^{-1}(1/2)$ mediana poblacional, podemos definir las α_*Truncadas* y αβ_*Truncadas* sobre *X* para todo $\alpha,\beta\in[0,50]$.

En definitiva, para una muestra *x* las medidas $T_{\alpha}(x)$ y $T^2_{\alpha\beta}(x)$ para $\alpha\beta \in [0,50]$ pueden ser consideradas como estimadores robustos de su media y varianza, respectivamente, de manera que la presencia de observaciones atípicas en *x* no influirá de manera significativa en estas medidas, para niveles de truncamiento adecuados.

2.2 Caso multivariante

El estimador clásico de posición multivariante es el vector de medias. Como alternativa a éste, proponemos definir un vector basado en truncamientos usando su misma estructura donde en cada componente se utilice el estimador α_*Truncada*.

De esta forma, dada la muestra *p*-dimensional *z* de tamaño *n*, donde denotaremos por z_j a la muestra marginal de la variable *j*-ésima en *z* para *j=1,2,...,p,* se define el vector α_*Truncada* sobre *z* para nivel de truncamiento $\bar{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)$ como:

$$T_{\bar{\alpha}}(z) = \left(T_{\alpha_1}(z_1),\ T_{\alpha_2}(z_2), ..., T_{\alpha_p}(z_p) \right) \qquad [5]$$

Nótese que en esta definición se ha utilizado un nivel de truncamiento $\alpha_j \in [0,50]$ para cada marginal z_j, siendo entonces $T_0(z)$ el vector de medias muestrales de *z*, para 0 el vector de ceros de IR^p, y donde se cumple que $T_{\alpha}(sz+t) = sT_{\alpha}(z)+t$, para cualquier $s \in IR$ y $t \in IR^n$.

Para otro lado, el estimador clásico de escala multivariante es la matriz de varianzas-covarianzas, basada en la media muestral y definida mediante la varianza y la covarianza dos a dos de las muestras marginales. Como alternativa a esta medida, proponemos sustituir la varianza en cada marginal por las *αβ_Truncadas* y las covarianzas entre marginales por unas nuevas medidas, sustituyendo medias por medias truncadas en la estructura de la covarianza, las cuales definimos a continuación.

Dada las muestras $x = \{x_1, x_2, ..., x_n\}$ e $y = \{y_1, y_2, ..., y_n\}$ de las variables *X* e *Y*, respectivamente, denotamos y definimos *αβ_CoTruncada* sobre *x* e *y* de niveles de truncamiento $\bar{\alpha} = (\alpha_x, \alpha_y)$ y β como:

$$CT_{\bar{\alpha}\beta}(x,y) = C_{\bar{\alpha}\beta}(X,Y)\, T_{\beta}\left(\left\{ \left(x_i - T_{\alpha_x}(x) \right)\left(y_i - T_{\alpha_y}(y) \right) \right\}_i \right) \qquad [6]$$

donde $\alpha_x, \alpha_y, \beta \in [0,50]$ y $C_{\bar{\alpha}\beta}(X,Y)$ será un nuevo coeficiente de consistencia, dependiente de los niveles α y β y de la distribución de *X* e *Y*, del cual supondremos que es invariante ante cambios de posición y escala, y con $C_{\bar{0},0}(X,Y) = 1$. Estas medidas tienen la misma estructura que la covarianza, donde se han sustituido medias por medias

truncadas, y propiedades similares, cumpliendo: $CT_{\bar{\alpha}\beta}(x,y) = CT_{\bar{\alpha}'\beta}(y,x)$ donde $\bar{\alpha}' = (\alpha_x, \alpha_y)$; $CT_{\bar{\alpha}\beta}$ $(sx+t, s'y+t') = ss'$ $CT_{\bar{\alpha}\beta}(x,y)$ para $s,s',t,t' \in$ IR cualesquiera; y $CT_{\bar{\alpha}\beta}(x,x) = T^2_{\alpha\beta}(x)$ para $\bar{\alpha}' = (\alpha,\alpha)$. Además, sus extremos son la covarianza, para el caso de $\bar{\alpha}$ = (0,0) y β=0, y para $\bar{\alpha}$ = (50,50) y β=50 una medida que llamamos *co-mediana de las distancias a las medianas* entre marginales, denotado y definido por:

$$CMDM(x,y) = CT_{(50,50),50}(x,y) \quad [7]$$

Retomando la construcción de una alternativa mediante truncamientos de la matriz de varianzas-covarianzas, proponemos construir esta nueva matriz sustituyendo los estimadores clásicos por los estimadores de truncamiento definidos. Así, dada la muestra *p*-dimensional *z* como antes, llamamos matriz de truncamientos sobre *z* de niveles de truncamiento $\hat{\alpha} = (\bar{\alpha}_{ij})_{i,j=1,2,...,p}$ y $\hat{\beta} = (\beta_{ij})_{i,j=1,2,...,p}$ a la matriz:

$$MT_{\hat{\alpha}\hat{\beta}}(z) = \left(CT_{\bar{\alpha}_{i,j}\beta_{i,j}}(z_i, z_j) \right)_{i,j=1,2,...,p} \quad [8]$$

donde: $\bar{\alpha}_{ij} = (\alpha_i, \alpha_j)$ y β_{ij} son los niveles de truncamiento para cada par de muestras marginales z_i y z_j; $\beta_{ij} = \beta_{ji}$; y con $\alpha_i, \beta_{ij} \in [0,50]$ para todo *i* y *j*. De esta forma, $MT_{\hat{\alpha}\hat{\beta}}(z)$ es una matriz cuadrada de orden *p*, simétrica, con $MT_{\hat{0},\hat{0}}(z)$ (todos los niveles ceros) la matriz de varianzas-covarianzas en *z*, y definida positiva en general (cuestión que ha sido comprobada en la gran mayoría de los casos donde se ha calculado para su uso en muestras reales, ejemplos y simulaciones).

Como en el caso univariante, es posible extender las definiciones [5] y [8] al caso de variables usando la ecuación [4].

En conclusión, para una muestra multivariante *z* las medidas $T_{\bar{\alpha}}(z)$ y $MT_{\hat{\alpha}\hat{\beta}}(z)$ pueden ser consideradas como estimadores robustos del vector de medias y matriz de varianzas-covarianzas, respectivamente, donde de nuevo, para niveles de truncamiento adecuados, la presencia de observaciones atípicas en *z* no influirá de manera significativa en sus cálculos.

3. Elementos de los estimadores de truncamiento

Con el fin de tener perfectamente definidos los estimadores de truncamiento para su uso sobre datos reales, debemos determinar los coeficientes de consistencia y los niveles de truncamiento en cada caso. Estos elementos dependerán en última instancia de la distribución de la población de donde se extrae la muestra, distribución que supondremos es simétrica, ya que es en estos casos donde los estimadores de truncamiento son especialmente apropiados.

3.1 Coeficientes de consistencia

Los coeficientes de consistencia, tanto para el estimador *αβ_Truncada* en [2] como para el estimador *αβ_CoTruncada* en [6], pretenden servir para que los estimadores respeten las características poblacionales de la variable de la que se supone procede la muestra. Para ellos supondremos que no son afectados por cambios de posición y/o escala en la muestra, y que dependen únicamente de los niveles de truncamiento y de la distribución de la variable que genera la muestra.

Para X una variable unidimensional con distribución simétrica, media μ_x y varianza σ^2_x no nula, el coeficiente $C_{\alpha\beta}(X)$ será definido de manera que $T^2_{\alpha\beta}(X)$ sea consistente en media cuadrática con σ^2_x. Bajo este supuesto, recordando que $T_\alpha(X)$ son estimadores insesgados de μ_x y que son equivariantes para todo $\alpha \in [0,50]$, se cumple que:

$$T^2_{\alpha\beta}(X) = C_{\alpha\beta}(X)\sigma^2_x T_\beta(X_0)$$

donde $X_0=((X-\mu_x)/\sigma_x)^2$. Imponiendo ahora que $T^2_{\alpha\beta}(X)=\sigma^2_x$, se obtiene:

$$C_{\alpha\beta}(X) = \left[T_\beta(X_0)\right]^{-1} \qquad [9]$$

para cualquier α. Así, el coeficiente de consistencia para *αβ_Truncada* en [2] depende del nivel de truncamiento β y de la distribución de la variable X, siendo su cálculo, en general, de fácil implementación en el ordenador.

Para el caso del coeficiente de consistencia $C_{\alpha\beta}(X,Y)$ usado en la ecuación [6], consideremos ahora dos variable X e Y con distribuciones simétricas, esperanzas μ_x y μ_y, desviaciones típicas σ_x y σ_y no nulas y covarianza σ_{xy}. Entonces $T_\alpha(X)=\mu_x$ y $T_\alpha(Y)=\mu_y$ para todo $\alpha \in [0,50]$, de manera que tomando las estandarizaciones $X'=(X-\mu_x)/\sigma_x$ e $Y'=(Y-\mu_y)/\sigma_y$ y recordando que T_α es una estimador equivariante, se obtiene:

$$CT_{\bar{\alpha}\beta}(X,Y) = C_{\bar{\alpha}\beta}(X,Y)\sigma_x\sigma_y T_\beta(X'Y')$$

Imponiendo ahora que $CT_{\bar{\alpha}\beta}(X,Y)=\sigma_{xy}$, si $T_\beta(X'Y')\neq 0$ se obtiene:

$$C_{\bar{\alpha}\beta}(X,Y) = r_{xy}\left[T_\beta(X'Y')\right]^{-1} \qquad [10]$$

para cualquier $\bar{\alpha}$, donde $r_{xy}=\sigma_{xy}/(\sigma_x\sigma_y)$ es el coeficiente de correlación entre X e Y.

En este caso, el coeficientes de consistencia para αβ_*CoTruncada* en [6] también depende del nivel β y de la distribución de la variables X e Y, donde la correlación entre ellas será de especial interés.

En resumen, los coeficientes de consistencia $C_{\alpha\beta}$ y $C_{\overline{\alpha}\beta}$ dependen de los niveles de truncamiento y de la distribución de donde se extrae la muestra. Estos coeficientes serán determinados para el caso particular de distribuciones normales y t-Student en la Sección 4.

3.2 Niveles de truncamiento

Los niveles de truncamiento condicionan el comportamiento de los estimadores de truncamiento ante ciertas propiedades deseables. Más concretamente, su valor determina, por una parte, el comportamiento del estimador ante las propiedades aconsejables para estudios realizados en presencia de observaciones atípicas (robustez), mientras que por otra, también determina la eficiencia de dicho estimador.

El objeto de los niveles de truncamiento es determinar cuántas observaciones pueden ser consideradas como observaciones atípicas en la muestras. De esta forma, si la muestra no contiene observaciones atípicas estos niveles deberían ser ceros, de manera que los estimadores de truncamiento coincidan con los estimadores clásicos. Por otro lado, el hecho de que algún nivel de truncamiento sea cero, no significa que no existan observaciones atípicas en la muestra, sino que éstas no afectan de manera significativa en el cálculo de las estimaciones. Así, por ejemplo, si en una muestra univariante tenemos sólo dos valores muy extremos que podrían ser considerados como observaciones atípicas (uno a cada lado del grueso de la muestra, pero igualmente muy distantes de éste), el nivel de truncamiento α en la medida α_*Truncadas* debería ser cero, ya que sus influencias se contrarrestarían.

Centraremos nuestra atención en el nivel de truncamiento α para las α_*Truncadas* y β para las αβ_*Truncadas* y las αβ_*CoTruncadas*, ya que el resto de niveles se construirán (como veremos) a partir de estos. Así, para empezar, debemos determinar qué valores son posibles para estos niveles y, posteriormente, cuáles serían los aconsejables en cada situación particular.

Es fácil comprobar que los posibles valores que puede tomar α, para que las α_*Truncadas* proporcionen diferentes valores entre sí en una muestra de tamaño n, se incluyen en el conjunto denotado y definido por:

$$P\nu\ (n) = \{0, s, 2s, ..., ks\} \qquad [11]$$

donde s=100/n y k=*Int*((*n*−1)/2). De igual manera ocurre para el nivel β en las medidas αβ_*Truncadas* y αβ_*CoTruncadas*.

Veamos ahora cómo determinar unos niveles de truncamiento apropiados, según unos criterios, tanto para el caso univariante como el multivariante.

La elección de un nivel de truncamiento para la familia α_*Truncada* se estudia en diferentes trabajos (ver, por ejemplo, Dodge y Jurecková, 2000). Esta elección puede depender de diversos criterios, como: las características particulares de la muestra donde

se aplica la medida; criterios basados sobre la idea de asegurar una determinada eficiencia o robustez, dependiendo de las necesidades del estudio; o incluso de las preferencias particulares del investigador.

En este trabajo proponemos un nuevo criterio de elección, basado en: elegir el menor nivel de truncamiento de manera que las diferencias entre las α_*Truncadas* para valores mayores que éste sean "pequeñas", entendiendo por esto que sean menores que una determinada cota. Es decir, elegir el nivel a partir del cual las diferencias entre las estimaciones proporcionadas por las α_*Truncadas* estén controladas, o sean, en algún sentido, estables. Este control o estabilidad se fija mediante la citada cota, a la cual llamaremos *cota de estabilidad*.

Siguiendo esta idea, dada una muestra *x* de tamaño *n*, se define el nivel de truncamiento para α_*Truncada* de cota ε_x en *x* como el valor:

$$\alpha_x = Min\left\{\alpha \in Pv(n) / \left|T_a(x) - T_b(x)\right| < \varepsilon_x \forall a,b \geq \alpha\right\} \quad [12]$$

donde se propone que la cota ε_x dependa de la muestra *x* y sea determinada como una medida de la diferencia máxima entre las posibles estimaciones de α_*Truncada* para muestras de la distribución de la que se extrae *x*. Ya que la medida α_*Truncada* es equivariante, es posible fijar este valor usando muestras estandarizadas de la distribución. En definitiva, ε_x es construida por simulación mediante:

$$\varepsilon_x = \hat{\sigma}_x \; Percentil_{99}\left\{Max_{a>b}\left\{\left|T_a(m_x) - T_b(m_x)\right|\right\}\right\} \quad [13]$$

donde $\hat{\sigma}_x$ es determinado mediante un estimador de escala robusto sobre *x, y* m_x son muestras estandarizadas generadas por ordenador de la distribución de la que se supone procede *x*.

Para la elección de los niveles de truncamiento en αβ_*Truncada* seguiremos la misma idea anterior, adaptándola a este nuevo caso. De esta forma, dada una muestra *x* de tamaño *n* usaremos como α_x el determinado mediante [12] y tomaremos como nivel de truncamiento β el valor:

$$\beta_x = Min\left\{\beta \in Pv(n) / \left|T^2{}_{\alpha_x \alpha}(x) - T^2{}_{\alpha_x b}(x)\right| < \varepsilon'_x \forall a,b \geq \beta\right\} \quad [14]$$

donde ε'_x es una nueva cota de estabilidad, que proporciona una información similar a la que generaba ε_x, determinada de nuevo por simulación mediante:

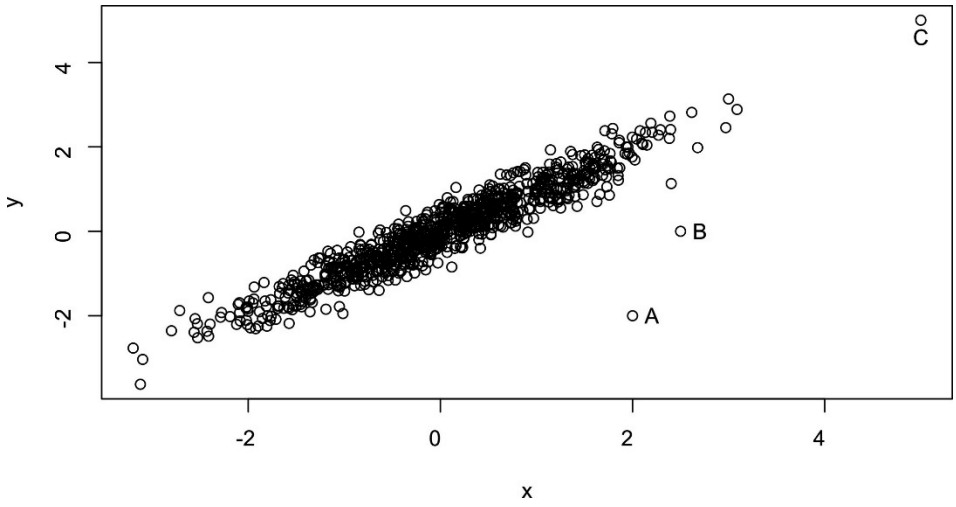
$$\varepsilon'_x = \hat{\sigma}_x^2 \; Percentil_{99}\left\{Max_{a>b}\left\{\left|T^2{}_{\alpha_x a}(m_x) - T^2{}_{\alpha_x b}(m_x)\right|\right\}\right\} \quad [15]$$

donde $\hat{\sigma}_x$ es también determinado mediante un estimador de escala robusto sobre *x, y* m_x son muestras estandarizadas generadas por ordenador de la distribución de la que se supone procede *x*.

Para determinar los niveles de truncamiento en el caso multivariante, comenzamos haciendo un estudio para la medida αβ_*CoTruncada* definida en [6]. Para ilustrar el procedimiento de elección usaremos, sin pérdida de generalidad, una muestra bivariante estandarizada representada en la Figura 1.

Figura 1

Muestra bivariante estandarizada, donde las observaciones *A* y *C* influyen significativamente en el cálculo de *αβ CoTruncada*, mientras que *B* no



Las observaciones que pueden contaminar αβ_*CoTruncada* pueden ser de dos tipos: observaciones que no siguen la estructura del grueso de los datos, llamadas *outlier de regresión* (como las observaciones del tipo *A* y *B* en la Figura 1), donde unas (las de tipo *A*) producen una influencia contaminante mientras que otras (las observaciones del tipo *B*) no producen influencias significativas en el cálculo; y observaciones que sí siguen la estructura del grueso de los datos, incluso fortaleciendo esa relación, pero que son extremadamente alejadas, que son llamadas observaciones *good leverage* (como las observaciones del tipo *C* en la Figura 1). En consecuencia, para determinar los niveles α y β en αβ_*CoTruncada* debemos tener en cuenta, por un lado, las observaciones extremas en cada muestra marginal y, por otro, las observaciones del tipo outliers de regresión con influencia contaminante (observaciones del tipo *C* y *A* en la Figura 1, respectivamente).

Para el nivel α en la medida αβ_*CoTruncada* proponemos que sea el vector formado por los niveles de truncamiento en las α_*Truncadas* marginales. Esto está justificado con

objeto de proteger los resultados ante contaminaciones en ambas marginales (observaciones del tipo *C* en la Figura 1), ya que para esta medida los outliers de regresión no influyen. Por otro lado, para la elección del nivel β en la medida αβ_*CoTruncada*, tendremos en cuenta la influencia contaminante tanto de los outliers de regresión como de las observaciones good leverage. Es inmediato que al multiplicar las componentes marginales de cada observación de la muestra bivariante, los outliers de regresión aparecen como valores extremos negativos, por lo que su presencia hace que la covarianza (es decir, la media de estos valores) sea menor. Además, las observaciones good leverage serán eliminadas en cada marginal usando los niveles de truncamiento β en cada αβ_*Truncada*. En definitiva, proponemos como nivel de truncamiento β en αβ_*CoTruncada* aquél que maximice la media truncada de los productos de las componentes de las muestras una vez estandarizadas, teniendo en cuenta el signo de su correlación, siempre que este β sea mayor que los niveles β en las αβ_*Truncadas* para cada muestra marginal.

En resumen, dadas las muestras $x=\{x_1,x_2,\ldots,x_n\}$ e $y=\{y_1,y_2,\ldots,y_n\}$ y siendo α_x, α_y, β_x y β_y los niveles de truncamiento para αβ_*Truncada* en *x* e *y*, calculados mediante [12] y [14] respectivamente, proponemos tomar como nivel de truncamiento α en la medida αβ_*CoTruncada* para *x* e *y* el vector:

$$\bar{\alpha}_{xy} = \left(\alpha_x, \alpha_y\right) \qquad [16]$$

y, construyendo $x'=\{x'_i\}_i$ e $y'=\{y'_i\}_i$ donde $x'_i=\left(x_i - T_{\alpha x}(x)\right)/\sqrt{T^2{}_{\alpha x \beta x}(x)}$ e $y'_i=\left(y_i - T_{\alpha y}(y)\right)/\sqrt{T^2{}_{\alpha y \beta y}(y)}$ proponemos usar como nivel β el de valor:

$$\beta_{xy} = Arg\ \max_\beta \left\{T_\beta\left(ax'y'\right) / \beta \geq \max\left\{\beta_x, \beta_y\right\}\right\} \qquad [17]$$

donde, construyendo $s=med_i(sig(x'_i y'_i))$, si $s\in\{-1,1\}$ tomamos $a=s$ y si $s=0$ tomaremos por convenio $a=1$. Notar que *a* proporciona una estimación del signo de la correlación entre las muestras *x* e *y*.

Veamos ahora cómo determinar los niveles de truncamiento en el caso multivariante en general. Para la elección de los niveles de truncamiento en el vector de truncamientos definido en [5] y la matriz de truncamientos definida en [8], proponemos que estos dependan de los elementos antes definidos. Es decir, sobre una muestra *p*-dimensional *z*, donde z_j es la muestra marginal de la variable *j*-ésima en *z* para cada *j=1,2,...,p*, siendo α_i y β_i los niveles de truncamiento para αβ_*Truncada* en cada z_i univariante, usaremos para la construcción de $T_\alpha(z)$ el vector:

$$\bar{\alpha}_z = \left(\alpha_1, \alpha_2, \ldots, \alpha_p\right) \qquad [18]$$

y para la construcción de $MT_{\alpha\beta}(z)$ los niveles de truncamiento:

$$\hat{\alpha}_z = \left(\bar{\alpha}_{ij}\right)_{i,j=1,2,\ldots,p} \qquad \hat{\beta}_z = \left(\beta_{ij}\right)_{i,j=1,2,\ldots,p} \qquad [19]$$

donde $\bar{\alpha}_{ij} = (\alpha_i, \alpha_j)$ y $\beta_{ij} = \beta_{ji}$ es el nivel de truncamiento para αβ_*CoTruncada* en las muestras z_i y z_j definido mediante [17] con $\beta_{ii} = \beta_i$ para todo i.

En definitiva, la elección de los niveles de truncamiento que determinan los estimadores de truncamiento sobre una muestra, se basa en dos cotas de estabilidad que dependen, en última instancia, de la muestra y de la distribución de donde se extrae ésta. En la Sección 4 desarrollaremos estos estudios para algunos casos particulares.

4. Estimadores de truncamiento en la práctica

Algunos estimadores robustos son poco usados en estudios con datos reales, quizá por no estar perfectamente definidos todos sus elementos para su implementación en el ordenador. Además, en la mayoría de estudios se supone que los datos proceden de una distribución normal, en algunos casos por ser la más conocida y estudiada (Hoaglin y otros, 1982).

En el caso de los estimadores de truncamiento, los elementos necesarios para su uso en la práctica son los coeficientes de consistencia y los niveles de truncamiento, donde estos últimos se basan en unos ciertos valores llamados cotas de estabilidad.

Con objeto de que los estimadores de truncamiento sí sean usados en la práctica, en esta sección se estudian y determinan estos elementos, tanto para el caso de normalidad como para distribuciones t-Student, donde todos los cálculos y simulaciones se han realizado con el paquete estadístico R (R-program version 3.0.1. Copyright (c) 2013. The R Foundation for Statistical Computing Platform: i386-w64-mingw32/i386, 32-bit).

4.1 Los estimadores en distribuciones normales

Denotando por $N_p(\mu,\Sigma)$ a una distribución normal p-dimensional con esperanza μ y matriz de varianzas-covarianzas Σ, vamos a determinar los coeficientes de consistencia y los niveles de truncamiento para el uso de los estimadores de truncamiento sobre muestras de esta distribución.

Coeficientes de consistencia. Siendo $X \sim N(\mu_x, \sigma_x)$, normal univariante, el cuadrado de su estandarización será $((X-\mu_x)/\sigma_x)^2 \sim {\chi_1}^2$, por lo que el coeficiente de consistencia $C_{\alpha\beta}(X)$ usando la ecuación [9] queda definido por:

$$C_{\alpha\beta}(X) = \left[T_\beta\left(\chi^2_1\right)\right]^{-1} \qquad [20]$$

para cualquier α. En el Cuadro 1 se presentan estos valores para ciertos β, donde: $C_{\alpha,0}(X)$=1, de manera que $T^2_{0,0}(x)$ es la varianza de una muestra x de la variable; y $C_{\alpha,50}(X)$=2,1981, es decir, el coeficiente de consistencia para *MDM* coincide con el cuadrado del considerado para el estimador *MAD*, de valor C_m=1,4826 (Hampel y otros, 1986).

Cuadro 1

Coeficientes de consistencia $C_{\alpha\beta}(X)$ para $X \sim N(\mu,\sigma)$ y algunos β.

	$\beta = 0$	5	10	15	20	25	30	35	40	45	50
$C\alpha\beta$	1	1,2485	1,4280	1,5884	1,7344	1,8653	1,9788	2,0717	2,1409	2,1836	2,1981

Para el caso del coeficiente de consistencia $C_{\tilde{\alpha}\beta}(X,Y)$ en la ecuación [10], donde $(X,Y)\sim N_2(\mu,\Sigma)$ normal bivariante, éste depende del coeficiente de correlación r_{xy} y de la distribución del producto de estandarizaciones $X'Y'$. Para determinar esta dependencia, hemos realizado un estudio de simulación sobre muestras normales estandarizadas con diferentes r_{xy}. En el Cuadro 2 se presentan los valores medios de los coeficientes obtenidos para algunos casos en la simulación realizada, con sus varianzas entre paréntesis.

Cuadro 2

Valores medios (varianzas) del coeficiente de consistencia $C_{\tilde{\alpha}\beta}(X,Y)$ obtenidos mediante simulación de muestras de $(X,Y) \sim N_2(\mu,\Sigma)$ de tamaño n = 10.000 para diferentes r_{xy} y β.

$\lvert r_{xy}\rvert$	$\beta = 0$	5	10	15	20	25	30	35	40	45	50
≃0	0,25	-0,13	0,37	0,02	0,7369	-0,25	0,10	0,06	-0,82	0,01	-0 , 28
	(28,4)	(10,9)	(25,8)	(17,5)	(58,4)	(21,2)	(13,2)	(8,2)	(50,5)	(12,0)	(99,9)
0,1	1,01	1,28	1,50	1,74	2,00	2,28	2,62	3,04	3,59	4,44	5 , 22
	(0,10)	(0,14)	(0,17)	(0,21)	(0,26)	(0,31)	(0,37)	(0,46)	(0,58)	(0,76)	(1,13)
0,25	1,00	1,27	1,49	1,71	1,96	2,23	2,55	2,92	3,37	3,78	3 , 93
	(0,04)	(0,05)	(0,07)	(0,08)	(0,10)	(0,12)	(0,14)	(0,18)	(0,22)	(0,29)	(0,33)
0,5	1,00	1,26	1,48	1,69	1,92	2,16	2,42	2,67	2,88	3,01	3 , 06
	(0,02)	(0,03)	(0,03)	(0,04)	(0,05)	(0,06)	(0,08)	(0,09)	(0,11)	(0,12)	(0,13)
0,75	1,00	1,26	1,46	1,66	1,85	2,04	2,21	2,35	2,46	2,53	2 , 56
	(0,01)	(0,02)	(0,02)	(0,03)	(0,04)	(0,04)	(0,05)	(0,06)	(0,07)	(0,07)	(0,07)
0,9	1,00	1,25	1,44	1,62	1,79	1,94	2,07	2,18	2,26	2.31	2 , 33
	(0,01)	(0,02)	(0,02)	(0,03)	(0,03)	(0,03)	(0,04)	(0,05)	(0,05)	(0,05)	(0,06)
≃ 1	1,00	1,24	1,42	1,58	1,73	1,86	1,97	2,07	2,14	2,18	2 , 19
	(0,01)	(0,01)	(0,02)	(0,02)	(0,03)	(0,03)	(0,03)	(0,04)	(0,04)	(0,04)	(0,05)

Del estudio de simulación observamos que: para valores de r_{xy} cercanos a cero, las desviaciones típicas son grandes, por lo que el estudio no proporciona unos valores fiables para el coeficiente de consistencia; por otro lado, para valores mayores de $|r_{xy}|$ las

desviaciones típicas son pequeñas, por lo que sí que los resultados pueden ser más fiable, con valores muy parecidos al coeficiente de consistencia $C_{\alpha\beta}(X)$ cuando $|r_{xy}|$ es mayor de 0,5, convergiendo finalmente a los valores de este coeficiente conforme $|r_{xy}|$ se acerca a 1. En definitiva, el coeficiente de consistencia depende fuertemente del parámetro r_{xy}, no pudiendo ser determinado de manera independiente de éste.

Con objeto de tener unos valores concretos para el coeficiente $C_{\bar{\alpha}\beta}(X, Y)$ en muestras normales, proponemos usar como aproximación el valor del coeficiente de consistencia para una de las marginales con el mismo nivel de truncamiento, es decir, proponemos:

$$C_{\bar{\alpha}\beta}(X,Y) = \left[T_\beta\left(\chi^2{}_1\right)\right]^{-1} \qquad [21]$$

para cualquier $\bar{\alpha}$, donde esta aproximación será más apropiada cuanto mayor sea el valor absoluto de la correlación entre X e Y.

Niveles de truncamiento. Si x es una muestra de una variable normal, para determinar los niveles de truncamiento α_x y β_x usando las ecuaciones [12] y [14], respectivamente, es necesario fijar las cotas de estabilidad ε_x y ε'_x.

La cota de estabilidad ε_x para el caso de normalidad será calculada usando la ecuación [13] donde se propone como estimador de escala robusto $\sqrt{MDM}$, es decir:

$$\varepsilon_x = \sqrt{MDM(x)}\, Percentil_{99}\left\{Max_{a>b}\left\{\left|T_a(m_x) - T_b(m_x)\right|\right\}\right\}$$

siendo m_x muestras de $N(0,1)$. Con el fin de obtener unos valores concretos de esta cota para cada situación, proponemos usar una aproximación que será construida generando muestras m_x de diferentes tamaños n y realizando un ajuste mínimo cuadrático de los valores obtenidos sobre este parámetro. Como resultado, la cota de estabilidad para una muestra x de tamaño n extraida de una distribución normal $N(\mu_x,\sigma_x)$ se fija de la forma:

$$\varepsilon_N(x;n) = 1,7350 n^{-0,4746}\sqrt{MDM(x)} \qquad [22]$$

Para el caso de la cota ε'_x seguiremos el mismo procedimiento. Para la misma muestra x y teniendo en cuenta que αβ_*Truncada* es equivariante, la cota de estabilidad ε'_x será calculada usando la ecuación [15] donde el estimador de escala robusto es ahora *MDM*, es decir:

$$\varepsilon'_x = MDM(x)\, Percentil_{99}\left\{Max_{a>b}\left\{\left|T^2{}_{0,a}(m_x) - T^2{}_{0,b}(m_x)\right|\right\}\right\}$$

donde, de nuevo, m_x son muestras $N(0,1)$. Generando muestras m_x de diferentes tamaños n, proponemos volver a construir la cota de estabilidad mediante un ajuste mínimo cuadrático de los valores obtenidos. Como resultado, la cota de estabilidad ε'_x para una muestra x de tamaño n que se supone procede de una distribución normal $N(\mu_x, \sigma_x)$ se fija de la forma:

$$\varepsilon'_N(x;n) = 4{,}3940 n^{-0{,}4691} MDM(x) \qquad [23]$$

4.2 Los estimadores en distribuciones t-Student

Denotemos por $Z=(Z_1, Z_2,...,Z_p) \sim t_p(\nu,\mu,\Sigma)$ a una variable t-Student de dimensión p, con ν grados de libertad, parámetro de posición μ y parámetro de escala Σ. En estas condiciones, $E[Z]=\mu$ *(para* $\nu>1$*),* $Var[Z]=\nu/(\nu-2)\,\Sigma$ *(para* $\nu>2$*)* y si ν tiende a infinito (habitualmente $\nu>30$) entonces Z sigue una distribución normal $N_p(\mu,\Sigma)$. Además, cada marginal $Z_i \sim t(\nu,\mu_i,\sigma_i)$, t-Student univariante, donde $Z'_i=(Z_i-\mu_i)/\sigma_i \sim t(\nu,0,1)=t_\nu$ para cada $i=1,2,...,p$.

Dado que las distribuciones t-Student son simétricas, los estimadores de truncamiento son adecuados en muestras sobre estas variables, ya que los recortes se consideran por igual en ambas colas de la muestra. Por otro lado, en algunos trabajos ν se supone dado, considerándolo no como un parámetro de la distribución sino como un elemento conocido. Notar que este valor es especialmente importante en el caso de modelos contaminados, ya que una observación que puede ser considerada como atípica para un valor de ν grande, puede ser perfectamente genuina para un modelo con un valor de éste pequeño. En este trabajo supondremos ν conocido, con $\nu \in (2,100]$ donde para $\nu>100$ la distribución t-Student será considerada como una normal.

Para el uso de los estimadores de truncamiento sobre muestras t-Student necesitamos determinar, como en normalidad, los coeficientes de consistencia y los niveles de truncamiento para cada caso.

Coeficientes de consistencia. Suponiendo que $X \sim t(\nu,\mu_x,\sigma_x)$, sabemos que $E[X]=\mu_x$, $Var[x] = \nu/(\nu-2)\sigma_x^2$ y entonces $X_0 = (\nu-2)/\nu\left((X-\mu_x)/\sigma_x\right)^2 \sim (\nu-2)/\nu F_{1,\nu}$

donde $F_{1,\nu}$ es la distribución F-Snedecor de grados de libertad *1* y ν. Por lo tanto, el coeficiente de consistencia usando [9] depende de β y ν de la forma:

$$C_{\alpha\beta}(X) = \frac{\nu}{\nu-2}\left[T_\beta\left(F_{1,\nu}\right)\right]^{-1} \qquad [24]$$

que puede ser fácilmente implementado en el ordenador. En el Cuadro 3 se presentan algunos valores como muestra, observándose que: $C_{\alpha,\,0}(X)=1$, ya que para $\nu>2$ se tiene

que $T_0\ (F_{1,\nu})=\nu/(\nu-2)$; para un nivel de truncamiento β fijo, cuando ν disminuye los valores para $C_{\alpha\beta}(X)$ crecen; y que cuando los grados de libertad crecen el coeficiente se acerca al dado para el caso de normalidad (Cuadro 1), para cada nivel de truncamiento.

Cuadro 3

Coeficientes de consistencia $C_{\alpha\beta}(X)$ con $X \sim t(\nu,\mu,\sigma)$ para diferentes ν y β.

ν	*β = 0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*50*
≃ 2	2,55	11,30	15,28	18,75	21,90	24,74	27,22	29,28	30,81	31,77	32 , 09
3	1	2,1667	2,7401	3,2379	3,6878	4,0917	4,4432	4,7323	4,9484	5,0823	5,1276
5	1	1,5309	1,8443	2,1186	2,3666	2,5889	2,7819	2,9402	3,0583	3,1313	3,1561
8	1	1,3839	1,6308	1,8483	2,0455	2,2222	2,3754	2,5010	2,5946	2,6525	2,6721
10	1	1,3487	1,5788	1,7821	1,9665	2,1317	2,2751	2,3925	2,4800	2,5340	2,5523
15	1	1,3091	1,5198	1,7066	1,8762	2,0283	2,1602	2,2682	2,3486	2,3983	2,4151
20	1	1,2919	1,4940	1,6735	1,8365	1,9827	2,1094	2,2133	2,2906	2,3383	2,3545
30	1	1,2762	1,4702	1,6429	1,7998	1,9406	2,0626	2,1625	2,2369	2,2828	2,2984
100	1	1,2563	1,4400	1,6039	1,7530	1,8867	2,0027	2,0976	2,1682	2,2119	2,2266

Para el coeficiente de consistencia $C_{\bar{\alpha}\beta}(X,Y)$ siguiendo la ecuación [10], donde $(X,Y)\sim t_2(\nu,\mu,\Sigma)$, ocurre lo mismo que en el caso normal, este coeficiente depende del parámetro de correlación r_{xy} y de la distribución del producto de las marginales estandarizadas X' e Y', donde cada una de ellas siguen distribuciones t_ν marcadas por los grados de libertad ν. Como en el caso de normalidad, para determinar esta dependencia hemos realizado un estudio de simulación para diferentes r_{xy} y ν. En el Cuadro 4 se presentan algunos casos de los resultados obtenidos para los valores medios de los coeficientes, siendo las varianzas omitidas por ser similares a las del caso normal.

Cuadro 4

Valores medios del coeficiente de consistencia $C_{\bar{\alpha}\beta}(x,y)$ obtenidos mediante simulación de muestras (X,Y) ~ $t_2(\nu,\mu,\Sigma)$ de tamaño n = 10.000 y diferentes r_{xy}, ν y β. (Continúa)

r_{xy}	ν	β=0	5	10	15	20	25	30	35	40	45	50
≃0	≃ 2	0,00	0,01	4,38	-3,22	20,37	5,64	-44,44	3,93	-3,50	0,05	12,75
	3	0,06	-0,19	-0,78	-0,35	0,28	0,32	-0,21	-1,11	0,16	-9,68	1,56
	5	0,08	0,09	-0,00	-0,20	-0,06	0,35	0,08	-1,11	0,26	-3,89	-3,09
	8	0,02	-0,28	-0,38	-0,04	1,84	-0,33	1,30	0,06	-0,40	-0,06	1,74
	10	0,01	0,90	1,12	0,36	-0,91	1,21	0,37	-3,47	-0,40	-0,24	-0,44
	15	0,04	-0,05	-0,12	0,16	3,00	0,23	0,07	83,61	-8,03	-2,52	-0,71
	20	0,76	-0,21	-0,03	-0,52	-0,03	-0,06	-1,38	-0,00	0,15	-0,71	-7,37
	30	0,01	0,09	-0,19	0,13	-0,26	0,01	-0,17	-0,16	-9,57	-0,52	-1,10
	100	0,12	-0,25	0,06	0,21	-0,15	-0,39	0,35	0,28	-0,77	-0,21	-0,79
0,1	≃ 2	8,07	12,39	17,66	22,69	28,29	34,81	42,38	51,25	63,42	81,47	98,71
	3	1,91	2,30	3,04	3,77	4,55	5,46	6,51	7,78	9,63	12,22	14,51
	5	1,03	1,60	1,99	2,38	2,85	3,34	3,92	4,61	5,58	6,98	8,24
	8	1,00	1,43	1,75	2,07	2,42	2,80	3,23	3,82	4,57	5,73	6,81
	10	1,01	1,40	1,68	1,98	2,33	2,66	3,11	3,61	4,34	5,39	6,50
	15	1,00	1,35	1,62	1,89	2,19	2,52	2,91	3,41	4,05	5,02	5,93
	20	1,01	1,34	1,59	1,84	2,14	2,45	2,82	3,29	3,92	4,84	5,78
	30	1,01	1,32	1,56	1,82	2,07	2,41	2,74	3,20	3,83	4,77	5,48

	100	1,01	1,28	1,52	1,75	2,01	2,31	2,65	3,07	3,65	4,53	5,35

Cuadro 4

Valores medios del coeficiente de consistencia $C_{\bar{\alpha}\beta}(x,y)$ obtenidos mediante simulación de muestras (X,Y) ~ $t_2(\nu,\mu,\Sigma)$ de tamaño n = 10.000 y diferentes r_{xy}, ν y β. (Conclusión)

r_{xy}	ν	$\beta=0$	5	10	15	20	25	30	35	40	45	50
0,25	≃2	5,15	12,22	17,26	22,26	27,66	33,69	40,53	48,68	58,50	67,65	71,14
	3	1,00	2,27	2,99	3,71	4,46	5,32	6,29	7,42	8,88	10,12	10,60
	5	1,00	1,58	1,97	2,36	2,77	3,25	3,78	4,43	5,17	5,88	6,17
	8	1,00	1,42	1,72	2,03	2,36	2,72	3,16	3,66	4,27	4,81	5,06
	10	1,00	1,38	1,66	1,95	2,26	2,61	3,01	3,47	4,06	4,56	4,73
	15	1,00	1,34	1,60	1,86	2,14	2,46	2,83	3,26	3,77	4,25	4,42
	20	1,00	1,31	1,56	1,82	2,09	2,39	2,75	3,16	3,67	4,11	4,30
	30	0,99	1,30	1,54	1,78	2,05	2,34	2,68	3,08	3,57	4,00	4,17
	100	1,00	1,28	1,50	1,73	1,98	2,26	2,58	2,98	3,43	3,84	4,00
0,50	≃ 2	2,87	12,06	16,96	21,68	26,66	31,86	37,56	43,04	47,89	50,93	52,09
	3	1,02	2,26	2,95	3,63	4,34	5,08	5,87	6,67	7,32	7,74	7,92
	5	1,00	1,57	1,95	2,32	2,71	3,12	3,56	3,98	4,33	4,57	4,66
	8	0,99	1,41	1,71	2,00	2,31	2,64	2,99	3,32	3,60	3,80	3,86
	10	0,99	1,37	1,65	1,92	2,21	2,51	2,84	3,15	3,42	3,60	3,65
	15	1,00	1,33	1,58	1,83	2,10	2,38	2,68	2,97	3,20	3,36	3,42
	20	1,00	1,31	1,55	1,80	2,05	2,32	2,61	2,89	3,11	3,26	3,33
	30	1,00	1,29	1,53	1,76	2,00	2,26	2,54	2,81	3,03	3,17	3,23
	100	0,99	1,27	1,49	1,71	1,94	2,19	2,46	2,71	2,93	3,05	3,10
0,75	≃ 2	2,60	11,83	16,44	20,70	24,95	29,01	32,81	36,10	38,73	40,40	40,87
	3	1,02	2,23	2,88	3,50	4,10	4,68	5,22	5,67	6,04	6,25	6,33
	5	0,99	1,56	1,91	2,25	2,58	2,90	3,19	3,44	3,63	3,76	3,80
	8	1,00	1,40	1,68	1,95	2,22	2,47	2,70	2,90	3,04	3,14	3,18
	10	1,00	1,36	1,63	1,88	2,12	2,36	2,57	2,76	2,89	2,98	3,02
	15	1,00	1,32	1,56	1,79	2,02	2,24	2,43	2,60	2,73	2,81	2,84
	20	1,00	1,30	1,53	1,75	1,97	2,18	2,37	2,53	2,66	2,74	2,76
	30	1,00	1,29	1,51	1,72	1,93	2,13	2,31	2,47	2,59	2,66	2,68
	100	1,00	1,27	1,47	1,68	1,88	2,07	2,24	2,39	2,50	2,57	2,59
0,90	≃ 2	2,62	11,57	15,89	19,73	23,32	26,70	29,68	32,14	34,07	35,22	35,62
	3	1,01	2,20	2,81	3,37	3,87	4,35	4,77	5,11	5,38	5,54	5,59
	5	1,00	1,54	1,88	2,18	2,47	2,73	2,95	3,14	3,28	3,37	3,40
	8	0,99	1,39	1,66	1,90	2,12	2,33	2,50	2,65	2,77	2,84	2,86
	10	1,00	1,36	1,60	1,83	2,04	2,23	2,39	2,53	2,64	2,71	2,73
	15	0,99	1,31	1,54	1,75	1,94	2,11	2,27	2,40	2,49	2,55	2,57
	20	1,00	1,30	1,51	1,71	1,90	2,06	2,21	2,33	2,43	2,48	2,51
	30	1,00	1,28	1,49	1,68	1,86	2,02	2,16	2,28	2,37	2,42	2,44
	100	1,00	1,26	1,46	1,64	1,81	1,96	2,10	2,21	2,29	2,34	2,36
≃1	≃ 2	2,60	11,31	15,29	18,76	21,91	24,78	27,24	29,28	30,84	31,82	32,10
	3	1,01	2,16	2,74	3,23	3,68	4,09	4,44	4,73	4,96	5,09	5,13
	5	1,00	1,53	1,84	2,11	2,36	2,58	2,78	2,94	3,06	3,13	3,16
	8	1,00	1,38	1,63	1,84	2,04	2,22	2,37	2,50	2,59	2,65	2,67
	10	1,00	1,34	1,57	1,78	1,96	2,13	2,27	2,39	2,48	2,53	2,55
	15	1,00	1,30	1,52	1,70	1,87	2,02	2,16	2,26	2,35	2,39	2,41
	20	0,99	1,29	1,49	1,67	1,83	1,98	2,11	2,21	2,29	2,33	2,35
	30	0,99	1,27	1,47	1,64	1,80	1,94	2,06	2,16	2,24	2,28	2,29
	100	0,99	1,25	1,44	1,60	1,75	1,88	2,00	2,10	2,17	2,21	2,22

En los resultados de este estudio hemos observado que: fijado ν, los valores son similares entre sí para $|r_{xy}|$ mayores de 0,5; y, para estos mismos $|r_{xy}|$ fijados, al disminuir ν los valores obtenidos también disminuyen siendo muy similares al coeficiente univariante dado en la ecuación [24]. Por lo tanto, de nuevo el coeficiente de consistencia depende del parámetro r_{xy}, no pudiendo ser determinado su valor de manera independiente de éste parámetro. En definitiva, con objeto de tener unos valores fáciles de calcular en cada caso, de nuevo proponemos usar como aproximación para este coeficiente el valor del coeficiente de consistencia para una de las marginales, es decir:

$$C_{\bar{\alpha}\beta}(X,Y) = \frac{v}{v-2}\left[T_\beta\left(F_{1,v}\right)\right]^{-1} \qquad [25]$$

para cualquier $\bar{\alpha}$, donde este valor será más apropiado cuanto mayor sea el valor absoluto de la correlación entre *X* e *Y*.

Niveles de truncamiento. Para determinar los niveles de truncamientos para *x* una muestra de una variable t-Student, igual que en el caso normal, necesitamos fijar las cotas de estabilidad ε_x y ε'_x. De nuevo, para determinar estas cotas mediante las ecuaciones [13] y [15], hemos usado el estimador *MDM* y realizado unos estudios de simulación sobre muestras t-Student estandarizadas de diferentes tamaños muestrales *n* y parámetros ν. Después, para poder ser usadas en la práctica, las cotas de estabilidad han sido aproximadas mediante unos ajustes mínimo cuadráticos dependiendo de *n* y ν.

Como resultado de estos estudios, para una muestra *x* de tamaño *n* que procede de una variable $t(\nu,\mu_x,\sigma_x)$, las cotas de estabilidad ε_x y ε'_x aproximadas son de la forma:

$$\varepsilon_t(x;n,v) = 3{,}1189 n^{-0{,}4753} v^{-0{,}1257} \sqrt{MDM(x)} \qquad [26]$$

$$\varepsilon'_t(x;n,v) = 60{,}7580 n^{-0{,}5162} v^{-0{,}4965} MDM(x) \qquad [27]$$

5. Conclusiones

Los estimadores clásicos, como la media y la varianza, pueden ser muy influenciados por la presencia de observaciones atípicas (outliers) en la muestra en estudio. Para resolver este problema, se proponen unos estimadores resistentes a estas observaciones; los estimadores robustos. Como un caso de estos estimadores, la media truncada (de la familia de los L-estimadores) es un estimador de posición robusto bien conocido, muy intuitivo, con buenas propiedades en robustez y fácil de implementar en cualquier paquete estadístico.

Utilizando la estructura de los estimadores clásicos y la media truncada, en este trabajo se proponen los llamados *estimadores de truncamiento*. De ellos se estudian sus principales características, comprobándose que cumplen propiedades adecuadas en presencia de observaciones atípicas, por lo que proporcionarán estimaciones no influenciadas por contaminaciones. También se proponen procedimientos para determinar todos sus elementos y puedan ser utilizados en la práctica.

Como caso particular, se estudian los estimadores de truncamiento para muestras que procedan de distribuciones normales y de distribuciones t-Student. Así, para muestras de estas distribuciones, se propone cómo construir los coeficientes de consistencia y se proporciona una aproximación de las cotas de estabilidad que determinan la elección de unos niveles de truncamiento adecuados. Las rutinas para el cálculo de los estimadores de truncamiento han sido desarrolladas para el paquete estadístico R, y pueden ser proporcionadas bajo petición al autor.

Actualmente estamos realizando diferentes estudios para aplicar los estimadores propuestos. En uno de ellos se propone construir una alternativa robusta de la distancia de Mahalanobis, para detectar observaciones atípicas en muestras contaminadas, sustituyendo los estimadores de posición y escala clásicos por los estimadores de truncamiento.

Referencias

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H. Y TUCKEY, J.W. (1972). «Robust Estimates of Location». *Ed. Princeton University Press*.

BARNETT, V. Y LEWIS, T. (1994). «Outliers in Statistical Data». *Ed. Wiley and Sons*.

BICKEL, P.J. (1973). «On some analogues to linear combination of order statistics in the linear model». *Annals of Statistics*, 1.

DODGE, Y. Y JURECKOVÁ, J. (2000). «Adaptive regression». *Ed. Springer-Verlag N.Y.*

HAMPEL, F.R. (1974): «The Influence Curve and its role in robust estimation» *Journal of the American Statistical Association*, 69.

HAMPEL, F.R., RONCHETTI, E.M. ROUSSEEUW, P.J. Y STAHEL, W.A. (1986). «Robust Statistics: The approach based on influence functions». *Ed. Wiley and Sons*.

HOAGLIN, D.C., MOSTELLER, F. Y TUCKEY, J.W. (1982). «Understanding Robust and Exploratory Data Analysis». *Ed. Wiley and Sons*.

HUBER, P.J. (1964). «Robust estimation of a location parameter». *Annals of Mathematical Statistics*, 35.

MARONNA, R.A., MARTIN, R.D. Y YOHAI, V.J. (2006). «Robust statistics. Theory and methods». *Ed. Wiley and Sons*.

ORTEGA, J.FCO. (2004). «A family of scale estimators by means of trimming». In *Theory and Applications of Recent Robust Methods*. pp. 259-269. Ed. Birkhauser Verlag.

ROUSSEEUW, P.J. (1984): «Least median of squares regression». *Journal of American Statistical Association*, 388.

ROUSSEEUW, P.J. Y LEROY, A.M. (1987). «Robust Regression and Outlier Detection». Ed. *Wiley and Sons.*

ROUSSEEUW, P.J. Y CROUX C. (1993). «Alternatives to the Median Absolute Deviation». *Journal of the American Statistical Association*, 424.