Regiones de tolerancia robustas multivariadas

Andrés Farall

Directora: Dra. Graciela Boente

Tesis para acceder al título de

Magister en Estadística Matemática

Universidad de Buenos Aires

Resumen

Esta tesis aborda el problema de definir regiones de tolerancia para datos multivariados y estudiar sus propiedades de robustez. El objetivo de este trabajo incluye realizar una revisión bibliográfica sobre los resultados existentes relacionados con las regiones de tolerancia multivariadas, proponer una versión robusta de las mismas y calcular los factores necesarios para la determinación de dichas regiones. Por otra parte, un estudio de la función de influencia permite analizar la sensibilidad a datos atípicos de las propuestas consideradas.

En primer lugar, se definen regiones de tolerancia multivariadas para distribuciónes normales y se describen algunas aproximaciones para el cálculo de las mismas. Mediante un estudio de simulación se realiza un análisis del comportamiento de estas regiones de tolerancia, a las que llamaremos clásicas, ante la presencia de datos atípicos externos (outliers) e internos (inliers). Dicho estudio permite mostrar la falta de robustez del procedimiento tradicionalmente utilizado.

Con la idea de solucionar esta falta de robustez se proponen regiones de tolerancia robustas mediante un procedimiento "*plug–in*", es decir, mediante el reemplazo de los estimadores de posición y escala clásicos con los cuales se construyen las regiones, por estimadores robustos. Para el cálculo de las constantes de tolerancia y el estudio de simulación, se trabajó con los estimadores definidos por Stahel (1981) y Donoho (1982). Los factores de tolerancia para estas regiones robustas en el caso de la distribución normal se calcularon mediante un estudio de Monte Carlo.

Una deficiencia común de los métodos aproximados de cálculo de los factores de tolerancia (Monte Carlo) es la carencia de elementos que permitan calcular el error cometido en dicha aproximación. Es así que conjuntamente con los factores estimados definimos y calculamos valores conservativos para estos factores, que dan una idea de la medida del error.

Por último, un estudio de simulación permite comparar la cobertura real de las regiones clásicas y robustas cuando las observaciónes provienen de una distribución normal y de distribuciónes alternativas.

Índice general

Capítulo 1

Introducción

Dentro del campo de la inferencia estadística existe la intención de definir regiones que den idea de los valores posibles que toman las variables en estudio.

Se han definido tres nociones distintas de regiones aleatorias: la región de confianza, de predicción y de tolerancia. En particular, cuando el interés está puesto en un parámetro de centralidad como la media poblaciónal ($\boldsymbol{\mu}$), y se supone que las observaciónes tienen distribución normal, las regiones de confianza están basadas en la media muestral ($\overline{\mathbf{x}}$) y la matriz de covarianza muestral ($\mathbf{S}$). Es decir, cuando $\mathbf{x}_i$, $1 \leq i \leq n$, son independientes $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, una región de confianza para $\boldsymbol{\mu}$ de nivel q verificará

$$P_{\boldsymbol{\mu}}\left((\boldsymbol{\mu} - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\boldsymbol{\mu} - \overline{\mathbf{x}}) \leq \mathcal{T}^2_{d,n-1,1-q}\right) = q \quad \forall \boldsymbol{\mu} \, .$$

donde $T^2_{d,n-1,1-q}$ es el punto que deja una probabilidad $1-q$ a la derecha de la distribución de Hotelling, y es equivalente a $\frac{d(n-1)}{n-d}\mathcal{F}_{d,n-d,1-q}$.

Cuando el interés es la predicción de nuevos valores, como ser la media de las futuras r observaciónes $\mathbf{y}_r = (\mathbf{x}_{n+1} + \cdots + \mathbf{x}_{n+r})/r$, se busca definir una región aleatoria $\mathcal{R}$ que con alta probabilidad contenga a $\mathbf{y}_r$. En el caso de observaciónes normales, el problema se reduce a buscar el valor de la constante K tal que

$$P\left((\mathbf{y}_r - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{y}_r - \overline{\mathbf{x}}) \leq K\right) = q \, .$$

La ley bajo la cual se miden las probabilidades anteriores, es la correspondiente a los vectores aleatorios $\overline{\mathbf{x}}$ e $\mathbf{y}_r$, así como a la matriz aleatoria $\mathbf{S}$. Cuando las observaciónes provienen de una normal multivariada la determinación de esta región es trivial ya que es fácil ver que la región $\mathcal{R}$ está dada por

$$\mathcal{R} = \left\{ \mathbf{y}_r : (\mathbf{y}_r - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{y}_r - \overline{\mathbf{x}}) \leq \frac{(n+r)}{nr} \, \frac{(n-1)\, d}{(n-d)} \, \mathcal{F}_{d,n-d,1-q} \right\} \, ,$$

donde $\mathcal{F}_{m,r,\alpha}$ es el punto que deja una probabilidad α a la derecha de la distribución $\mathcal{F}$ con m y r grados de libertad.

Sin embargo, muchas veces se cuenta con una serie de observaciónes previas de una cierta población, llamada muestra de referencia, y en base a ésta se quiere establecer una región que contenga a una alta proporción (q) de dicha población y de modo tal que el procedimiento tenga una alta confianza (δ). Es decir, dada una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$, con distribución $P_{\boldsymbol{\theta}}$, las regiones de tolerancia de Tipo 1, se definen como las regiones aleatorias $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ tales que

$$P_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right) \geq q\right] \geq \delta \quad \forall \boldsymbol{\theta} \ ,$$

donde $P_{1,\boldsymbol{\theta}}$ indica la probabilidad asociada a $\mathbf{x}$ (condicional a los valores $\mathbf{x}_1, \ldots, \mathbf{x}_n$) y $P_{n,\boldsymbol{\theta}}$ la asociada a $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Para una tal región, el usuario tiene una confianza δ de que la probabilidad de cobertura de la región $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ sea por lo menos q, independientemente de la distribución de la muestra elegida, siempre que ésta pertenezca a la familia dada.

Otro tipo de regiones de tolerancia que ha recibido menos atención son la regiones de tolerancia de Tipo 2. Una región $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ se dice una región de tolerancia $\delta-$esperada si

$$E_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right)\right] \geq \delta \quad \forall \boldsymbol{\theta} \ .$$

Para tales regiones la probabilidad de cobertura (condicional) media es como mínimo δ. Como antes, $E_{n,\boldsymbol{\theta}}$ es el valor esperado asociado a la distribución de $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Es interesante notar que estas regiones de tipo 2 coinciden con las regiones de predicción para una nueva observación ($r = 1$). En efecto, teniendo en cuenta que

$$\begin{aligned} E_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right)\right] &= E_{n,\boldsymbol{\theta}}\left[E_{1,\boldsymbol{\theta}}[I\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right) | \mathbf{x}_1, \ldots, \mathbf{x}_n]\right] \\ &= E[I(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n))] \\ &= P[\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)) \ , \end{aligned}$$

se obtiene que $E_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right)\right] \geq \delta$ equivale a $P[\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)) \geq \delta$.

Por lo tanto, en el caso de datos normales, si $n > d$, la región de tolerancia de Tipo 2, tiene una expresión explícita ya que es un caso particular de región de predicción normal multivariada con $r = 1$, y está dada por $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \{\mathbf{y} : (\mathbf{y} - \overline{\mathbf{x}})\, \mathbf{S}^{-1}\, (\mathbf{y} - \overline{\mathbf{x}}) \leq K\}$, donde $K = \dfrac{(n+1)}{n}\, \dfrac{(n-1)d}{(n-d)} \mathcal{F}_{d,n-d,1-\delta}$.

Si llamamos C a la variable aleatoria $P_{1,\boldsymbol{\theta}}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right)$, que mide la cobertura (condicional), la equivalencia anterior muestra que mientras las regiones de predicción sólo acotan el valor promedio de C ($E(C) \geq \delta$), las regiones de tolerancia establecen como cuantil $1 - \delta$ de C al valor q ($P(C \geq q) \geq \delta$). Esto produce que factores de predicción que garantizan un nivel δ_0 definan regiones de tolerancia con

nivel $\delta = \delta_0$ y cobertura menor a δ_0 ($q \leq \delta_0$) o regiones de tolerancia con cobertura $q = \delta_0$ que poseen una confianza $\delta \leq \delta_0$.

Propiedades que relacionan las regiones de tolerancia de Tipo 2 con un problema de test de hipótesis y las regiones obtenidas en el caso de la distribución normal pueden hallarse en Fraser y Guttman (1956).

La región de tolerancia se dice no–paramétrica si la probabilidad de cobertura condicional $P_{1,\boldsymbol{\theta}}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right)$ es una variable aleatoria cuya distribución no depende de $\boldsymbol{\theta}$. Varios autores han considerado regiones de tolerancia no–paramétricas, entre otros podemos mencionar Fraser (1951), Fraser y Guttman (1956), Krzanowski y Radley (1989), Mee (1970), Quesenberry y Gessaman (1968).

En este trabajo, nos interesarán regiones de tolerancia para la distribución normal, que conserven sus propiedades en un entorno de la misma. En el caso particular de la distribución normal, las regiones de tolerancia son elipsoides aleatorios centrados en la media muestral, es decir

$$\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left\{\mathbf{y} : (\mathbf{y} - \overline{\mathbf{x}})\, \mathbf{S}^{-1}\, (\mathbf{y} - \overline{\mathbf{x}}) \leq K\right\}$$

y la constante K o factor de tolerancia se elige de modo tal que se verifique

$$P_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}\left((\mathbf{x} - \overline{\mathbf{x}})\, \mathbf{S}^{-1}\, (\mathbf{x} - \overline{\mathbf{x}}) \leq K | \overline{\mathbf{x}}, \mathbf{S}\right) \geq q\right] \geq \delta \quad \forall \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ .$$

Es importante destacar que en ambos tipos de regiones la probabilidad de cobertura ($P_{1,\boldsymbol{\theta}}$) es condicional a las variables $\overline{\mathbf{x}}$ y $\mathbf{S}$ calculadas con la muestra de referencia, en tanto que la probabilidad externa ($P_{n,\boldsymbol{\theta}}$) o la esperanza ($E_{n,\boldsymbol{\theta}}$) se toman sobre estas últimas.

En el Capítulo 2, se presenta una revisión de los principales resultados relacionados con las regiones de tolerancia univariadas y multivariadas y se propone un procedimiento de un paso para mejorar la cobertura de dichas regiones. Se estudia además la bondad de las aproximaciones dadas por Guttman (1970) para calcular el factor de tolerancia.

Como es previsible, por estar centrados en la media y la matriz de covarianza muestral, la probabilidad de cobertura y el tamaño de estas regiones se verá fuertemente afectado por observaciónes atípicas. En el Capítulo 3 se analiza el comportamiento de la región de tolerancia clásica ante datos atípicos y se propone una alternativa robusta que consiste en reemplazar los estimadores clásicos de posición y escala multivariados por estimadores robustos Fisher–consistentes y afín equivariantes.

El valor de los factores de tolerancia para las regiones robustas obtenidas utilizando los estimadores de Donoho–Stahel se obtiene mediante un estudio de Monte Carlo. El algoritmo utilizado y los valores obtenidos se encuentran en el Capítulo 3 donde además se realiza un estudio de simulación que permite comparar el comportamiento de la nueva propuesta y de la propuesta clásica para muestras de referencia con

distribución normal y con distintos tipos de contaminaciones. En el Capítulo 4, se calcula la función de influencia de la probabilidad de cobertura. La función de influencia permite explicar la sensibilidad del procedimiento clásico ante datos atípicos externos. Por otra parte, resulta ser acotada si el estimador multivariado de escala utilizado tiene influencia acotada.

En el Capítulo 5, se describe brevemente una generalización al caso multivariado del procedimiento para corregir la probabilidad de cobertura dado por Fernholz y Gillespie (2001) que será objeto de estudio en el futuro.

Finalmente, en el Capítulo 6, se presentan las conclusiones de esta tesis. Las Tablas, Figuras y Programas se presentan en el Apéndice.

Capítulo 2

Regiones de Tolerancia

2.1. Introducción

Las regiones de tolerancia son de amplio uso en la industria, siendo quizá la principal aplicación el Control de Calidad, en donde se busca garantizar, para una multiplicidad de variables en estudio, el cumplimiento de ciertos estándares.

Otra importante aplicación consiste en la utilización de las regiones de tolerancia como elementos de decisión sobre la pertenencia de ciertas muestras a determinadas poblaciónes.

Un ejemplo de la primera aplicación lo constituye el estudio realizado por Fuchs y Kenett (1988). En dicho estudio, se emplean regiones de tolerancia en el control de calidad de la producción de obleas cerámicas utilizadas en la industria electrónica. En ese trabajo se definen regiones de tolerancia basadas en una muestra de referencia de 13 observaciónes, que se supone cumplen con los estándares requeridos por la industria, con la finalidad de utilizarlas en la toma de decisión de la aceptación o rechazo de nuevos lotes de obleas.

Otro ejemplo relacionado con el control de calidad consiste en la utilización de las regiones de tolerancia como herramientas para la detección de cambios en la función de distribución de la cual se muestrean las observaciónes. Más precisamente consideremos un proceso que genera observaciónes independientes $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de vectores aleatorios $d-$dimensionales, como ser la producción de un cierto bien del cual interesa realizar un seguimiento de la calidad de sus características a lo largo del tiempo. A partir de un determinado momento puede interesar saber si el nuevo bien producido ($\mathbf{x}_{n+1}$) se corresponde con los niveles de calidad observados hasta ese momento o si se produjo algún cambio sustancial en el proceso de producción. En este caso, la regla de decisión estaría basada en la pertenencia o no de $\mathbf{x}_{n+1}$ a la región de tolerancia construida con la muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

La teoría de los intervalos de tolerancia (regiones de tolerancia univariadas) se encuentra bastante desarrollada (ver Proschan (1953), por ejemplo). Para el caso normal univariado, o sea, cuando $x_1, \ldots, x_n$ son independientes $x_i \sim N(\mu, \sigma^2)$ y $x \sim N(\mu, \sigma^2)$, si bien no hay solución explícita, la probabilidad de cobertura puede obtenerse como

$$P_1\left(\frac{(x-\overline{x})^2}{s} \leq K \mid \overline{x}, s\right) = F_W\left(Ks, \overline{x}\right)$$

donde $F_W(t, \nu)$ es la función de distribución de una variable aleatoria chi–cuadrado con parámetro de no centralidad ν evaluada en t y donde s^2 es el estimador insesgado de la varianza.

Luego, en este caso, el problema se reduce a hallar la menor constante K que satisface la siguiente ecuación integral

$$\begin{aligned} P_{\overline{x},S}\left[F_W\left(KS, \overline{x}\right) \geq q\right] &\geq \delta \Leftrightarrow \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I\left(F_W\left(K\frac{y}{n-1}, \frac{z}{\sqrt{n}}\right) \geq q\right) f_Y(y)\ f_Z(z)\ dy\ dz &\geq \delta \end{aligned}$$

donde $Y = (n-1)\, s^2 \sim \chi^2_{n-1}$, $Z = \sqrt{n}\,\overline{x} \sim N(0,1)$ e $I(.)$ es la función característica.

Tablas, algoritmos y aproximaciones para éste problema pueden encontrarse en Odeh y Owen (1980).

En el caso multivariado, el desarrollo de las regiones de tolerancia es mucho menor y el cálculo del factor de tolerancia K es más complejo ya que, aún en el caso normal, se debe resolver una ecuación integral $(d+1)-$dimensional. La primer gran clasificación de los distintos enfoques a las regiones de tolerancia multivariadas consiste en enfoques paramétricos por un lado y en enfoques no paramétricos por otro. Entendemos por enfoque paramétrico a aquel en el cual se supone que la función de distribución de la cual provienen las observaciónes es conocida. Dentro de la teoría no paramétrica las regiones de tolerancia se construyen partiendo el espacio muestral de las n observaciónes en $n+1$ bloques y componiendo con estos bloques regiones que satisfagan la definición general (ver Murphy (1948)). Mejoras basadas en esta idea pueden ser halladas en el trabajo de Fraser (1951).

Dentro del enfoque paramétrico resultados como los de John (1962), Chew (1966), Guttman (1970) y Krishnamoorthy y Mathew (1999) se basan en aproximaciones analíticas a la distribución de las medias aritméticas, geométricas y armónicas de los autovalores de una matriz Wishart para poder obtener aproximaciones al factor de tolerancia. Recientemente, Krishnamoorthy y Mathew (1999) dieron aproximaciones para el factor de tolerancia basadas en estudios de Monte Carlo. Aunque computacionalmente costosas, estas evaluaciones son posibles.

En este Capítulo se definen las regiones de tolerancia para el caso de datos normales multivariados y se muestra que es siempre posible reducirse al caso $N(\mathbf{0}, \mathbf{I}_d)$ si los estimadores de posición y escala son afín equivariantes. Se describen, luego, las aproximaciones dadas por varios autores para el cálculo del factor de tolerancia K. Finalmente, se da una propuesta de un paso para mejorar, sin alto costo computacional, la aproximación dada en Guttman (1970).

2.2. Regiones de tolerancia multivariadas para distribuciónes normales.

Dada una muestra aleatoria de referencia $\mathbf{x}_1, \ldots, \mathbf{x}_n$ proveniente de una población con distribución normal $d-$dimensional, esto es $\mathbf{x}_i \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y una nueva observación $\mathbf{x}$ independiente de las anteriores con igual distribucioń, para cada $0 < q < 1$ y $0 < \delta < 1$, una región de tolerancia $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, basada en la muestra de referencia, es la región aleatoria que cubre con probabilidad q a la nueva observación $\mathbf{x}$ con una confianza superior a δ. Como hemos dicho, en el caso normal, la región de tolerancia es de la forma

$$\mathcal{R} = \left\{\mathbf{y} : (\mathbf{y} - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{y} - \overline{\mathbf{x}}) \leq K\right\} \tag{2.1}$$

donde $\overline{\mathbf{x}} = \dfrac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, $\mathbf{S} = \dfrac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ y el factor de tolerancia $K = K(q, \delta, n, d)$ es función del tamaño de la muestra de referencia n, de la dimensión d, así como de los parámetros q y δ.

El siguiente resultado muestra que si tomamos estimadores de posición y escala afín equivariantes, para el cálculo de la constante K podemos suponer $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = \mathbf{I}_d$, es decir, suponer en el caso de la región antes descripta que $\overline{\mathbf{x}} \sim N_d\left(\mathbf{0}, \frac{1}{n}\mathbf{I}_d\right)$ y $(n-1)\mathbf{S} \sim W_d(n-1, \mathbf{I}_d)$, donde $W_d(m, \boldsymbol{\Sigma})$ es la distribución Wishart $d-$dimensional con m grados de libertad y matriz de escala $\boldsymbol{\Sigma}$.

Proposición 2.2.1. *Sean $\mathbf{x}_i$, $1 \leq i \leq n$ vectores aleatorios independientes con distribución elíptica $P_{1,\boldsymbol{\theta}}$, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, o sea, $\mathbf{C}^{-\frac{1}{2}}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{z}_i$ tiene distribución esférica G con $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$. Sean $\mathbf{t}_n = \mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ y $\mathbf{V}_n = \mathbf{V}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ estimadores de posición y escala afín equivariantes. Definamos la región*

$$\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left\{\mathbf{y} : (\mathbf{y} - \mathbf{t}_n)' \mathbf{V}_n^{-1} (\mathbf{y} - \mathbf{t}_n) \leq K\right\} .$$

Sea K el factor de tolerancia asociado a $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, es decir, la constante tal que

$$p_n(\boldsymbol{\theta}) = P_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n) \geq q\right] \geq \delta \quad \forall \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) .$$

Entonces, $p_n(\boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$ y para el cálculo de K es posible suponer $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = \mathbf{I}_d$. Más precisamente, el factor de tolerancia es la constante tal que

$$P_{n,\boldsymbol{\theta}_0}\left[P_{1,\boldsymbol{\theta}_0}\left(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right) \geq q\right] \geq \delta \ ,$$

donde $\boldsymbol{\theta}_0 = (\mathbf{0}, \mathbf{I}_d)$.

DEMOSTRACIÓN. Por ser los estimadores afín equivariantes se verifica $\mathbf{t}_n\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) = \boldsymbol{\mu} + \mathbf{C}\,\mathbf{t}_n\left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right)$ y $\mathbf{V}_n\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) = \mathbf{C}\,\mathbf{t}_n\left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right)\mathbf{C}'$, donde $\mathbf{z}_i = \mathbf{C}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)$ con $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$. El resultado se deduce inmediatamente usando que la región $\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \boldsymbol{\mu} + \mathbf{C}\mathcal{R}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$. □

El objetivo es entonces hallar la constante K que satisface

$$P_{n,(\mathbf{0},\mathbf{I}_d)}\left[P_{1,(\mathbf{0},\mathbf{I}_d)}\left(\left(\mathbf{x} - \overline{\mathbf{x}}\right)\mathbf{S}^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}\right) \leq K\right) \geq q\right] \geq \delta \tag{2.2}$$

donde $\mathbf{x} \sim N_d\left(\mathbf{0}, \mathbf{I}_d\right)$, $P_{n,(\mathbf{0},\mathbf{I}_d)}$ representa la probabilidad calculada respecto de la distribución de $(\overline{\mathbf{x}}, \mathbf{S})$ cuando $\mathbf{x}_i \sim N_d\left(\mathbf{0}, \mathbf{I}_d\right)$ y $P_{1,(\mathbf{0},\mathbf{I}_d)}$ representa la probabilidad inducida por $\mathbf{x}$.

2.2.1. Aproximaciones para el factor de tolerancia

En esta sección describiremos las propuestas dadas por varios autores a fin dar un valor aproximado para el factor de tolerancia, debido a la complicación de su computo numérico a través de la resolución de (2.2).

Ignorando los términos de orden 2 en $\dfrac{1}{n}$, John (1962) obtuvo la siguiente aproximacion para el cálculo del factor K

$$K = \frac{d\,(n-1)\;g\left(q, \dfrac{d}{n}, d\right)}{g(1-\delta, 0, (n-1)d)} \ , \tag{2.3}$$

donde $g(p, \nu, m)$ es el percentil p de la distribución χ^2 con m grados de libertad y parámetro de no centralidad ν. Fuchs y Kennet (1987) mostraron que cuando el tamaño de muestra es grande, $n \geq 50$, y el número de variables es moderado ($d < 7$) el uso del percentil de la χ^2 central en el numerador conduce a resultados similares. Luego, la aproximación del factor de tolerancia K dada por estos autores es

$$K = \frac{d\,(n-1)\;g(q, 0, d)}{g(1-\delta, 0, (n-1)d)} \ .$$

A través de un estudio de Monte Carlo, Krishnamoorthy y Mathew (1999) comparan distintos métodos para calcular el factor de tolerancia K. Ellos muestran que la

aproximación propuesta por John (1962) y dada por (2.3), que es la que usualmente se utiliza en la literatura, es inadecuada cuando $d \geq 2$. Las aproximaciones basadas en la media geométrica de los autovalores de la matriz Wishart y en la media armónica son mejores que la anterior. La aproximación basada en la media geométrica para el factor de tolerancia K está dada por

$$K = \frac{c_1 \, (n-1) \, g\left(q, \frac{d}{n}, d\right)}{G\left(1-\delta, \frac{d\,(n-d)}{2}\right)} \, ,$$

donde

$$c_1 = \frac{d}{2}\left[1 - \frac{(d-1)(d-2)}{2n}\right]^{\frac{1}{d}}$$

y $G(\alpha, m)$ es el percentil α de una distribución $\Gamma(m, 1)$. Por otra parte, la aproximación basada en la media armónica para el factor de tolerancia K está dada por

$$K = \frac{d \, (n-1) \, g\left(q, \frac{d}{n}, d\right)}{g(1-\delta, 0, (n-1)d - d(d+1) + 2)} \, . \tag{2.4}$$

Las aproximaciones más precisas se deben a Krishnamoorthy y Mathew (1999) y consisten en modificar la aproximación de la media armónica de los autovalores de la matriz Wishart (2.4). La aproximación de estos autores está dada por

$$K = \frac{c_2 \, g\left(q, \frac{d}{n}, d\right)(n-1)}{g(1-\delta, 0, c_3)} \, ,$$

donde

$$c_2 = \frac{d(c_3 - 2)}{n-d-2} \qquad\qquad c_3 = \frac{d(n-d-1)(n-d-4) + 4(n-2)}{n-2} \, .$$

En todas las aproximaciones analíticas, el grado de precisón depende de parámetros fundamentales como el tamaño de la muestra de referencia (n) y de la dimensión (d), no existiendo una aproximación uniformemente mejor.

Krishnamoorthy y Mathew (1999) proponen un método de Monte Carlo para estimar el factor de tolerancia K que puede describirse como sigue

(i) $j = 1$

(ii) Se generan n vectores aleatorios $\mathbf{x}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$. Se calcula $\overline{\mathbf{x}}$ y $\mathbf{S}$.

(iii) Se generan $R = 1200$ vectores aleatorios $\mathbf{y}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$ y para cada uno de ellos se evalúa la forma cuadrática $Q_i = (\mathbf{y}_i - \overline{\mathbf{x}})' \, \mathbf{S}^{-1} \, (\mathbf{y}_i - \overline{\mathbf{x}})$.

(iv) Se ordenan las formas cuadráticas $Q^{(1)} \leq \cdots \leq Q^{(R)}$ y se busca el percentil q de la forma cuadrática $Q^{(R\,q)}$. Denotemos u_j a este percentil.

(v) $j = j + 1$.

(vi) Se repiten (ii) a (v) $N = 1200$ veces, conservando u_j en cada iteración.

(vii) Se ordenan los valores de u_j, $u^{(1)} \leq \cdots \leq u^{(N)}$ y se busca el percentil δ, $u^{(N\,\delta)} = K^*$, K^* es una aproximación al factor de tolerancia K.

Cuando $n > d + 1$, Guttman (1970) obtiene una aproximación para la media y la varianza de la variable aleatoria probabilidad de cubrimiento

$$C(\overline{\mathbf{x}}, \mathbf{S}) = P_{1,(\mathbf{0}, \mathbf{I}_d)} \left((\mathbf{x} - \overline{\mathbf{x}}) \, \mathbf{S}^{-1} \, (\mathbf{x} - \overline{\mathbf{x}}) \leq K \right)$$

usando un desarrollo de Taylor de primer orden alrededor de $(\mathbf{0}, \mathbf{I}_d)$. La esperanza y varianza aproximadas resultan ser

$$\begin{aligned} \mu_C(K) &= \psi_d(K) - K^{\frac{d}{2}} e^{-\frac{K}{2}} \frac{1}{2^{\frac{d}{2}+1} n \Gamma(\frac{d}{2})} \\ \sigma_C^2(K) &= K^d e^{-K} \frac{1}{d \; 2^{d-1} \; n \; \Gamma^2(\frac{d}{2})} \end{aligned}$$

con $\psi_d(K) = P(W_d \leq K)$ donde $W_d \sim \chi_d^2$. Luego, utilizando esa aproximación para la esperanza y varianza expresadas como funciones de la constante K, Guttman (1970) aproxima la distribución de C por una distribución Beta de parámetros (a, b), $\mathcal{B}(a, b)$. Para dicha distribución el valor esperado y la varianza están dados por $\dfrac{a}{a+b}$ y $\dfrac{a\,b}{(a+b)^2(a+b+1)}$. Por lo tanto igualando estos valores a μ_C y a σ_C^2, se obtiene un sistema de ecuaciones que permiten despejar los parámetros a y b en función del factor de tolerancia K, más precisamente, se tiene

$$\begin{aligned} a &= \frac{\mu_C^2 \; (1 - \mu_C) - \mu_C \, \sigma_C^2}{\sigma_C^2} \\ b &= \frac{\mu_C \; (1 - \mu_C)^2 - (1 - \mu_C) \; \sigma_C^2}{\sigma_C^2} \; . \end{aligned}$$

De esta forma, se puede aproximar para tamaños de muestra grandes el factor de tolerancia por la constante K que satisface la ecuación

$$\int_q^1 \frac{\Gamma(a(K) + b(K))}{\Gamma(a(K))\Gamma(b(K))} t^{a(K)-1} (1-t)^{b(K)-1} dt = \delta \; . \tag{2.5}$$

Esta aproximación resulta ser no conservadora, en el sentido que la cobertura real es menor que la teórica, especialmente para valores de muestra n pequeños y dimensiones d grandes. La subestimación del factor de tolerancia encontrado puede deberse a tomar como punto de valuación en la aproximación a aquel punto con mayor cobertura $(\mathbf{0}, \mathbf{I}_d)$.

En la Tabla A.1 se compara la cobertura real y la teórica para 16 factores de tolerancia solución de (2.5) con $n = 100$ y $d = 4$ para niveles de confianza y de cobertura de 0.75, 0.90, 0.95 y 0.99.

La cobertura real se calculó mediante el siguiente algoritmo:

(i) Se generan n vectores aleatorios $\mathbf{x}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$. Se calcula $\overline{\mathbf{x}}$ y $\mathbf{S}$.

(ii) Para evaluar la cobertura se generaron R vectores aleatorios $\mathbf{y}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$ y para cada uno de ellos se verifica si yace o no dentro de la region $\mathcal{R}$

$$\mathcal{R} = \left\{\mathbf{y} : (\mathbf{y} - \overline{\mathbf{x}})' \, \mathbf{S}^{-1} \, (\mathbf{y} - \overline{\mathbf{x}}) \leq K\right\}$$

Si $\mathbf{y}_i \in \mathcal{R}$ se define $c_i = 1$ en caso contrario, $c_i = 0$. Se calcula la cobertura promedio $\overline{c}$ como $\overline{c} = \dfrac{1}{R}\displaystyle\sum_{i=1}^{R} c_i$.

(iii) Se repiten (i) y (ii) N veces, conservando $\overline{c}$ en cada iteración.

(iv) Ordenamos las coberturas promedio $\overline{c}^{(1)} \leq \cdots \leq \overline{c}^{(N)}$ y nos quedamos con la $N(1-\delta)-$ésima cobertura promedio, $\overline{c}^{(N\,(1-\delta))} = \pi$, la que aproxima a la cobertura real con nivel de confianza δ.

Se tomó $N = R = 1000$.

Un examen detallado de esta aproximación revela que el principal problema en la falta de precisión radica en la sobreestimación de la esperanza de la probabilidad de cobertura (la media de la variable aleatoria C) y no en la forma de la distribución.

Para ver esto se consideraron 27 combinaciones de n, d y K que se describen en la Tabla A.2. Para cada una de ellas se calcularon 1000 realizaciones de la variable aleatoria C. Cada realización surge de estimar, como se describió anteriormente, una vez generados $\overline{\mathbf{x}}$ y $\mathbf{S}$, la probabilidad $P_{1,(\mathbf{0}, \mathbf{I}_d)}\left((\mathbf{x} - \overline{\mathbf{x}})\, \mathbf{S}^{-1} \, (\mathbf{x} - \overline{\mathbf{x}}) \leq K\right)$.

Para evaluar la bondad del ajuste de la distribución Beta propuesta por Guttman (1970), se realizaron los gráficos cuantil-cuantil de los datos generados, comparándolos con los de la distribución Beta teórica con parámetros estimados a partir de la media muestral y la varianza empírica. Estos gráficos se presentan en las Figuras B.1 a B.5 en la que se aprecia que la distribución empírica no difiere de una Beta.

La Figura B.7 muestra las medias empíricas y la subestimación producida por las medias aproximadas por Guttman (1970), para las combinaciones de parámetros descriptas en la Tabla A.2. En todos los casos, la aproximación sobreestima el cubrimiento promedio, lo que produce una subestimación en el cálculo del factor de tolerancia K. La aproximación de las varianzas tampoco es buena, pero no posee un sesgo sistemático, como en el caso de la media.

2.2.2. Propuesta de un paso para el cálculo del factor de tolerancia

Una alternativa para el cálculo del factor de tolerancia es utilizar como factor inicial el valor deducido del método propuesto por Guttman (1970), evaluar mediante simulaciones la subestimación de la cobertura promedio y en base a ésta calcular mediante una aproximación lineal el verdadero valor del factor.

Así, este estimador en un paso se calcula de la siguiente manera.

Sea K_0 el factor inicial calculado por Guttman (1970), la aproximación para el valor esperado de la probabilidad de cubrimiento está dada por

$$\mu_C(K_0) = \psi_d(K_0) - K_0^{\frac{d}{2}} e^{-\frac{K_0}{2}} \frac{1}{2^{\frac{d}{2}+1}\, n\, \Gamma(\frac{d}{2})} \ .$$

Sin embargo, el cubrimiento real (generalmente menor) estimado por Monte Carlo es una función del factor de tolerancia que indicaremos $\widetilde{\mu}(K_0)$. Buscamos un valor K_1 que nos dé un cubrimiento real igual a $\mu_C(K_0)$. Como no podemos conocer la derivada de $\widetilde{\mu}'(K)$ de $\widetilde{\mu}$, la aproximamos por la de $\mu_C(K)$ y entonces usando un desarrollo de Taylor de orden 1 obtenemos

$$\mu_C(K_0) = \widetilde{\mu}(K_1) = \widetilde{\mu}(K_0) + (K_1 - K_0)\, \mu'_C(K_0) \ .$$

De esta forma, la aproximación de un paso para el factor de tolerancia resulta ser

$$K_1 = K_0 + \frac{1}{\mu'_C(K_0)} \left(\mu_C(K_0) - \widetilde{\mu}(K_0)\right) \ ,$$

donde

$$\mu'_C(t) = f_d(t) - \frac{1}{2^{\frac{d}{2}+2}\, n\, \Gamma(\frac{d}{2})}\, t^{\frac{d}{2}} e^{-\frac{t}{2}} \left(\frac{d}{t} - 1\right)$$

con $f_d(t)$ la densidad de una χ^2_d evaluada en t.

La mejora en la probabilidad de cobertura puede verse en la Tabla A.3, donde se comparan, para $n = 100$ y $d = 4$, tres factores: el de Guttman (1970) K_0, la aproximación en un paso K_1 y el factor de la media harmónica de Krishnamoorthy

y Mathew (1999) K_2. Tambien se dan las coberturas generadas por los mismos, π_0 y π_1 y π_2, respectivamente.

Como puede verse de la Tabla A.3 y de la Tabla 4 de Krishnamoorthy y Mathew (1999), el procedimiento de un paso mejora notablemente la probabilidad de cobertura y da resultados del factor de tolerancia cercanos a los de estos autores, con menor costo computacional ya que los valores de N y R pueden tomarse menores a los considerados por esos autores. Siendo la razón de esto último la mayor velocidad de convergencia del promedio, utilizado para estimar la probabilidad de cobertura, en relación a la velocidad de convergencia del percentil muestral involucrado en la estimación de K. Adicionalmente, cada una de las N iteraciones es más rápida pues se evita el ordenamiento necesario para calcular cuantiles.

En conclusión, la única forma de garantizar para cualquier combinación de n, d, q y γ el correcto cálculo del factor de tolerancia es mediante Monte Carlo. Sin embargo, el procedimiento de un paso asegura una mejora si no se desea emplear un método con alto costo computacional.

Capítulo 3

Regiones de Tolerancia Multivariadas Robustas

3.1. Introducción

Aunque para la familia de distribuciónes normal, el problema de las regiones de tolerancia parece resuelto utilizando los procedimientos descriptos en el Capítulo 2, estos procedimientos no son distribuciónalmente robustos ya que son extremadamente sensibles a pequeños alejamientos del supuesto de normalidad. Más precisamente, la presencia de una sola observación atípica puede modificar en gran medida la región hallada, debido a cambios en la estimación de la matriz de covarianza y de la media, alterando la cobertura real de dicha región o el nivel de confianza de la misma. La pregunta es, por lo tanto, si el contenido q es todavía una cota inferior válida para la probabilidad de cobertura de la región de tolerancia basada en la media muestral $\overline{\mathbf{x}}$ y la matriz de covarianza muestral $\mathbf{S}$. En general, la respuesta es negativa. En el caso univariado, los límites de tolerancia no son ni siquiera aproximadamente válidos cuando nos alejamos de la hipótesis de normalidad, como fue observado por Butler(1982), Canavos y Koutraouvaelis (1984), Fernholz y Gillespie (2001) y las referencias citadas allí. Fernholz (2002) extendió la propuesta de Fernholz y Gillespie (2001) para incluir estimadores robustos en los extremos del intervalo a fin de obtener intervalos de tolerancia robustos.

En este Capítulo estudiaremos, en primer lugar, la sensibilidad del procedimiento clásico a observaciónes atípicas. Basado en ese hecho y con el objetivo de obtener procedimientos menos sensibles a datos atípicos, propondremos regiones de tolerancia basadas en estimadores robustos de posición y escala. Para el caso particular de los estimadores de Donoho–Stahel, se calculan numéricamente los factores de tolerancia para distintos valores de cobertura q y niveles de confianza δ. Un análisis de

sensibilidad preliminar análogo al realizado con el estimador clásico permite mostrar la ventaja de utilizar este tipo de procedimiento. Finalmente, un estudio de simulación permitirá comparar el comportamiento de las regiones clásicas y robustas bajo distintas distribuciónes.

3.2. Estudio de sensibilidad de las regiones de tolerancia clásicas

Para mostrar la falta de robustez en el caso multivariado contaminamos muestras de tamaño $n = 30$ con distribución $N(\mathbf{0}, \mathbf{I}_d)$, en dimensiones $d = 2$, 3, 4 y 5, cambiando un elemento de la muestra por un dato atípico externo $\mathbf{x}$ (outlier) a distancia $\Delta(\mathbf{x}) = \|\mathbf{x}\| = 2$, 4, 8 y 16 del centro de la distribución, en la dirección $\mathbf{e}_1 \in R^d$. Para cada una de estas muestras contaminadas se consideró la región de tolerancia definida por (2.1) con el factor de tolerancia construído para muestras normales con una cobertura teórica $q = 0.95$ y un nivel de confianza $\delta = 0.95$.

Los factores de tolerancia utilizados, fueron calculados por simulación utilizando el procedimiento de Krishnamoorthy y Mathew (1999) con $N = 1000$ y $R = 1000$, y se presentan en la Tabla A.4.

Mediante un estudio de Monte Carlo, se calculó la cobertura real π de cada región como se describió en la Sección 2.2.1 del Capítulo 2 con $N = 1000$ y $R = 1000$ y el incremento del volumen, $\mathcal{I}$, medido en % de la región contaminada con respecto al volúmen de la región sin contaminar. Los resultados se presentan en la Tabla A.5.

Como se desprende de la Tabla A.5, el principal problema consiste no sólo en la modificación de la cobertura real, ya que q deja de ser una cota válida, sino en el aumento del volúmen del elipsoide de tolerancia, como consecuencia del incremento en la varianza generalizada estimada ($\det(\mathbf{S})$). Esto explica que la cobertura real aumente, con la muestra de referencia contaminada, aún cuando la región esté centrada lejos del cero (verdadero centro) como resultado de la falta de robustez de la media como estimador de posición.

Sea $\mathbf{S} = \boldsymbol{\beta}\boldsymbol{\Lambda}\boldsymbol{\beta}'$ donde $\boldsymbol{\beta}$ son los autovalores de $\mathbf{S}$ y $\boldsymbol{\Lambda}$ es la matriz de autovalores. Para la región de tolerancia

$$\begin{aligned} \mathcal{R} &= \left\{\mathbf{y} : (\mathbf{y} - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{y} - \overline{\mathbf{x}}) \leq K\right\} \\ &= \left\{\mathbf{y} : (\boldsymbol{\beta}'\mathbf{y} - \boldsymbol{\beta}'\overline{\mathbf{x}})' \frac{\boldsymbol{\Lambda}^{-1}}{K} (\boldsymbol{\beta}'\mathbf{y} - \boldsymbol{\beta}'\overline{\mathbf{x}}) \leq 1\right\}, \end{aligned}$$

el volúmen $\mathcal{V}$ se calcula de la siguiente forma

$$\mathcal{V}(\mathcal{R}) = \int_{\mathcal{R}} 1 \, d\mathbf{y} = \prod_{i=1}^{d} \sqrt{K\lambda_i} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} = K^{\frac{d}{2}} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} [\det(\mathbf{S})]^{\frac{1}{2}}$$

Por lo tanto, el volúmen es proporcional a la raíz cuadrada del determinante de la matriz de covarianza estimada ($\mathbf{S}$). El incremento del volúmen de la región, en presencia de datos atípicos, sumado al corrimiento del centro de la misma como consecuencia del dato atípico externo, es especialmente problemático cuando se usa la región de tolerancia con fines de clasificación de nuevas muestras. El incremento de volúmen $\mathcal{I}$ se define como

$$\mathcal{I} = \sqrt[d]{\frac{[\det(\mathbf{S}_c)]^{\frac{1}{2}}}{[\det(\mathbf{S})]^{\frac{1}{2}}}}$$

donde $\mathbf{S}_c$ indica a la matriz de covarianza muestral calculada con la muestra contaminada.

Para tener una idea del incremento esperado de volúmen calcularemos $\det(E(\mathbf{S}))$, cuando la muestra está contaminada.

Proposición 3.2.1. *Sean $\mathbf{x}_i$, $1 \leq i \leq n$ vectores aleatorios independientes tales que $\mathbf{x}_i \sim N_d(0, \mathbf{I}_d)$ para $1 \leq i \leq n-1$ y $\mathbf{x}_n \sim N_d(\boldsymbol{\mu}, \mathbf{I}_d)$. Consideremos los estimadores clásicos de posición y escala*

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad y \qquad \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \ .$$

Entonces, la esperanza de los estimadores clásicos está dada por

$$E(\overline{\mathbf{x}}) = \frac{1}{n}\boldsymbol{\mu} \tag{3.1}$$

$$E(\mathbf{S}) = \mathbf{I}_d + \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\mu}' \tag{3.2}$$

$$\det(E(\mathbf{S})) = \mathbf{I}_d + \frac{1}{n}\|\boldsymbol{\mu}\|^2 \tag{3.3}$$

Demostración. (3.1) se deduce fácilmente ya que

$$E(\overline{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} E(\mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n-1} E(\mathbf{x}_i) + \frac{1}{n}E(\mathbf{x}_n) = \frac{1}{n}\boldsymbol{\mu} \ .$$

Por otra parte, como

$$\begin{aligned}(n-1)E(\mathbf{S}) &= \sum_{i=1}^{n} E\left((\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\right) = \sum_{i=1}^{n} E(\mathbf{x}_i\mathbf{x}_i') - nE(\overline{\mathbf{x}}\overline{\mathbf{x}}') \\ &= \sum_{i=1}^{n-1} E(\mathbf{x}_i\mathbf{x}_i') + E(\mathbf{x}_n\mathbf{x}_n') - nE(\overline{\mathbf{x}}\overline{\mathbf{x}}') \ ,\end{aligned}$$

usando que $E(\mathbf{x}_i\mathbf{x}_i') = \mathbf{I}_d$ para $1 \leq i \leq n-1$, $E(\mathbf{x}_n\mathbf{x}_n') = \boldsymbol{\mu}\boldsymbol{\mu}' + \mathbf{I}_d$ y $E(\overline{\mathbf{x}}\overline{\mathbf{x}}') = \frac{1}{n^2}\boldsymbol{\mu}\boldsymbol{\mu}' + \frac{1}{n}\mathbf{I}_d$ ya que $\overline{\mathbf{x}} \sim N_d\left(\frac{1}{n}\boldsymbol{\mu}, \frac{1}{n}\mathbf{I}_d\right)$, se deduce

$$\begin{aligned}(n-1)E(\mathbf{S}) &= (n-1)\mathbf{I}_d + \boldsymbol{\mu}\boldsymbol{\mu}' + \mathbf{I}_d - n\left(\frac{1}{n^2}\boldsymbol{\mu}\boldsymbol{\mu}' + \frac{1}{n}\mathbf{I}_d\right)\\ &= (n-1)\mathbf{I}_d + \left(1 - \frac{1}{n}\right)\boldsymbol{\mu}\boldsymbol{\mu}'\\ &= (n-1)\mathbf{I}_d + \frac{n-1}{n}\boldsymbol{\mu}\boldsymbol{\mu}'.\end{aligned}$$

Con lo cual

$$\begin{aligned}\det(E(\mathbf{S})) &= \det\left(\mathbf{I}_d + \left(1 - \frac{1}{n}\right)\boldsymbol{\mu}\boldsymbol{\mu}'\right)\\ &= \left(1 + \frac{1}{n}\mathrm{tr}(\boldsymbol{\mu}\boldsymbol{\mu}')\right). \square\end{aligned}$$

Por otra parte, si contaminamos con una masa puntual tenemos el siguiente resultado

Proposición 3.2.2. *Sean $\mathbf{x}_i$, $1 \leq i \leq n$ vectores aleatorios independientes tales que $\mathbf{x}_i \sim N_d(0, \mathbf{I}_d)$ para $1 \leq i \leq n-1$ y $\mathbf{x}_n \sim \Delta_{\boldsymbol{\mu}}$, con $\Delta_{\boldsymbol{\mu}}$ la masa puntual en $\boldsymbol{\mu}$. Entonces, se verifica*

$$\begin{aligned}E(\overline{\mathbf{x}}) &= \frac{1}{n}\boldsymbol{\mu}\\ E(\mathbf{S}) &= \frac{n-1}{n}\mathbf{I}_d + \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\mu}'\\ \det(E(\mathbf{S})) &= \left(\frac{n-1}{n}\right)^d\left(1 + \frac{1}{n-1}\|\boldsymbol{\mu}\|^2\right)\end{aligned}$$

Demostración. La demostración es idéntica a la de la Proposición 3.2.1 observando que $E(\overline{\mathbf{x}}\overline{\mathbf{x}}') = \frac{1}{n^2}\boldsymbol{\mu}\boldsymbol{\mu}' + \frac{n-1}{n^2}\mathbf{I}_d$. $\square$

Así, en el caso de contaminaciones puntuales el incremento del volúmen esperado podría ser medido por

$$\sqrt{\left(1 + \frac{1}{n-1}\|\boldsymbol{\mu}\|^2\right)},$$

que sólo depende de la norma del vector $\boldsymbol{\mu}$.

Para estudiar la sensibilidad a datos atípicos internos (inliers), contaminamos muestras de tamaño $n = 30$ con distribución $N(\mathbf{0}, \mathbf{I}_d)$ en dimensiones $d = 2$, 3, 4 y 5,

con 1, 2, 3, o 4 datos atípicos de ubicados en el cero. Para cada una de estas muestras contaminadas se consideró la región de tolerancia definida a través del factor de tolerancia asociado a una cobertura teórica $q = 0.95$ y un nivel de confianza $\delta = 0.95$ para datos normales que se presentan en la Tabla A.4. Luego, mediante Monte Carlo se calculó la cobertura real de cada región y el incremento del volúmen con respecto al volúmen de la región sin contaminar. Los resultados se muestran en la Tabla A.6 que permite apreciar que la introducción de datos atípicos internos (inliers) tiene como resultado la reducción de la cobertura real como consecuencia de la reducción del volúmen de la región, dejando de ser válida la cota q pedida para la cobertura. Ambos efectos se agravan con el aumento de la dimensión.

Estos hechos muestran la necesidad de definir regiones de tolerancia que no sean tan sensibles a unas pocas observaciónes atípicas.

3.3. Regiones de tolerancia robustas

Un procedimiento "*plug–in*"puede utilizarse para obtener regiones de tolerancia robustas, que no se vean afectadas por la presencia de unos pocos datos atípicos. Dicho procedimiento consiste en reemplazar los estimadores de posición y escala clásicos ($\overline{\mathbf{x}}$ y $\mathbf{S}$ respectivamente) por estimadores robustos. Más precisamente, sean $\mathbf{t}_n = \mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ y $\mathbf{V}_n = \mathbf{V}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ estimadores robustos de posición y escala. Definamos la región

$$\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left\{\mathbf{y} : (\mathbf{y} - \mathbf{t}_n)' \mathbf{V}_n^{-1} (\mathbf{y} - \mathbf{t}_n) \leq K\right\} \tag{3.4}$$

donde la constante K se elige de modo tal que

$$p_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = P_{n,\boldsymbol{\theta}}\left[P_{1,\boldsymbol{\theta}}(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n) \geq q\right] \geq \delta \quad \forall \quad \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ ,$$

cuando $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\mathbf{x}_i \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $1 \leq i \leq n$, siendo $P_{n,\boldsymbol{\theta}}$ la distribución de $(\mathbf{t}_n, \mathbf{V}_n)$ y $P_{1,\boldsymbol{\theta}}$ la de $\mathbf{x}$. Diremos entonces que la región $\mathcal{R}$ es *una región de tolerancia robusta*.

Para el cálculo del factor de tolerancia, surge el problema de la dependencia de $p_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ en los parámetros desconocidos de posición y escala. Para evitar ese problema y utilizando la Proposición 2.2.1, se deberán elegir estimadores robustos de posición y escala multivariados afín equivariantes. Es decir, si $\mathbf{t}_n = \mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ y $\mathbf{V}_n = \mathbf{V}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ son estimadores robustos de posición y escala afín equivariantes, el factor de tolerancia K de la región $\mathcal{R}$ definida en (3.4) resuelve

$$P_{n,(\mathbf{0}, \mathbf{I}_d)}\left[P_{1,(\mathbf{0}, \mathbf{I}_d)}(\mathbf{x} \in \mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_n) | \mathbf{x}_1, \ldots, \mathbf{x}_n) \geq q\right] \geq \delta \ . \tag{3.5}$$

Por otra parte, para comparar los factores de tolerancia robustos definidos por (3.5) con los factores de tolerancia clásicos dados en la Tabla 4 de Krishnamoorthy y

Mathew (1999), es necesario que los estimadores de posición y escala sean Fisher–consistentes para la distribución normal. Es decir, que si $\mathbf{x}_i \sim N_d(0, \mathbf{I}_d)$, $1 \leq i \leq n$, $\mathbf{V}_n \xrightarrow{c.t.p.} \mathbf{I}_d$ y $\mathbf{t}_n \xrightarrow{c.t.p.} \mathbf{0}$.

Una revisión sobre estimadores robustos multivariados de posición y escala puede verse en Maronna y Yohai (1998).

3.4. Estimadores de posición y escala robustos

Existen numerosas propuestas de estimadores robustos de posición y escala. La primera de ellas es el M–estimador propuesto por Maronna (1976). La mayor desventaja de los M–estimadores multivariados de escala, con función de escores monótona, es que su punto de ruptura decrece con la dimensión. Para resolver este problema de falta se robustez, se introdujeron otras familias de estimadores robustos entre las que podemos mencionar el estimador de escala de elipsoide de mínimo volúmen (Rousseeuw y van Zomeren (1990)), el de mínimo determinante (MCD, Rousseeuw (1985)), el de Donoho (1982)–Stahel (1981) y los S-, MM- y $\tau-$estimadores (Lopuhaä (1990)). Todas estas propuestas pueden alcanzar un punto de ruptura igual a $\frac{1}{2}$. Entre ellos, sólo el de elipsoide de mínimo volúmen converge a una tasa menor, $n^{\frac{1}{3}}$, mientras que los otros convergen a una tasa del orden de $\sqrt{n}$.

Recordaremos la definición de los estimadores introducidos por Stahel (1981) y Donoho (1982) ya que son los que utilizaremos en este trabajo. Estos estimadores poseen un alto punto de ruptura en cualquier dimensión y son calculados como un promedio pesado de las observaciónes donde cada punto tiene un peso inversamente proporcional a su medida de "*atipicidad*".

Este estimador se define como sigue. Dada una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$, sea $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ y $\mathbf{a}'\mathbf{X}$ las proyecciones univariadas de los datos. Sean m y s estimadores univariados de posición y escala. Una medida de la atipicidad de $\mathbf{a}'\mathbf{x}_i$ es la distancia estandarizada

$$\frac{|\mathbf{a}'\mathbf{x}_i - m(\mathbf{a}'\mathbf{X})|}{s(\mathbf{a}'\mathbf{X})} .$$

Luego, la atipicidad del punto $\mathbf{x}_i \in I\!R^d$ se define como

$$r(\mathbf{x}_i, \mathbf{X}) = \sup_{a \in I\!R^d} \frac{|\mathbf{a}'\mathbf{x}_i - m(\mathbf{a}'\mathbf{X})|}{s(\mathbf{a}'\mathbf{X})} .$$

El estimador de Donoho–Stahel pesa a cada observación según su medida de atipicidad. Los estimadores de posición y escala multivariados quedan definidos entonces

como

$$\mathbf{t}_n = \frac{\sum_{i=1}^{n} w_i \, \mathbf{x}_i}{\sum_{i=1}^{n} w_i}, \tag{3.6}$$

$$\mathbf{V}_n = \beta \, \frac{\sum_{i=1}^{n} w_i \, (\mathbf{x}_i - \mathbf{t}_n) \, (\mathbf{x}_i - \mathbf{t}_n)'}{\sum_{i=1}^{n} w_i}, \tag{3.7}$$

donde $w_i = w\left(r^2\left(\mathbf{x}_i, \mathbf{X}\right)\right)$ con w una función de peso no–negativa y usualmente no–creciente y β es una constante de calibración para obtener Fisher–consistencia.

Los estimadores de posición y escala multivariados resultan ser afín equivariantes si los estimadores univariados m y s lo son. Por otra parte, si deseamos estimadores Fisher–consistentes cuando $\mathbf{x}_i \sim N\left(\mathbf{0}, \mathbf{I}_d\right)$, la constante β debe elegirse igual a

$$\beta = \frac{d \; E\left(w\left(W_d\right)\right)}{E\left(w\left(W_d\right) W_d\right)} \; ,$$

donde $W_d \sim \chi^2_d$.

Debido a la imposibilidad de calcular exactamente la medida de atipicidad y según lo propuesto por Stahel (1981) y lo estudiado por Maronna y Yohai (1995), se aproxima el supremo tomando un máximo sobre todas las direcciones ortogonales a los subespacios generados por $M = 1000$ submuestras de d puntos de los n puntos originales. En dimensión $d = 2$, es posible dar un algoritmo que no use técnicas de remuestreo maximizando sobre 1000 direcciones equiespaciadas (con ángulos $\ell \cdot 2\pi/1000$, $1 \leq \ell \leq 1000$).

En este trabajo, los pesos son calculados utilizando la función de Huber, o sea, la función $w(t) = w_H\left(\sqrt{t}\right)$ con w_H la función de peso de Huber dada por:

$$w_H\left(r\right) = I_{[0,c]}\left(r\right) + I_{(c,\infty)}\left(r\right)\left(\frac{c}{r}\right)^2$$

donde $c = \sqrt{\chi^2_{0{,}95,d}}$. Por otra parte, para obtener estimadores afín equivariantes, se eligieron como estimadores univariados $m(y_1, \ldots, y_n) = \underset{1 \leq i \leq n}{\text{median}}\left(y_i\right)$, la mediana de las observaciónes, y $s(y_1, \ldots, y_n) = \dfrac{1}{\Phi^{-1}(0{,}75)} \, \text{MAD}(y_1, \ldots, y_n)$, la mediana de los desvíos absolutos respecto de la mediana, escalada de modo a resultar Fisher–consistente cuando $y_i \sim N(0, 1)$.

Al utilizar la función de Huber $w(t) = w_H\left(\sqrt{t}\right)$ y para obtener Fisher–consistencia para datos normales, debe elegirse la constante de consistencia $\beta = d\,\dfrac{c_0}{c_2}$, donde

$$\begin{aligned} c_0 &= E\left(w(W_d)\right) = P\left(W_d < c^2\right) + c^2\, E\left(\frac{1}{W_d} I_{(c^2,\infty)}(W_d)\right) \\ &= P\left(W_d < c^2\right) + c^2\,\frac{1}{2\,(d-2)}\,\left(1 - P\left(W_{d-2} < c^2\right)\right) \quad \text{si} \quad d \neq 2 \\ c_2 &= E\left(w(W_d)W_d\right) = E\left(W_d I_{(0,c^2)}(W_d)\right) + c^2\, E\left(I_{(c^2,\infty)}(W_d)\right) \\ &= d\, P\left(W_{d+2} < c^2\right) + c^2\,\left(1 - P\left(W_d < c^2\right)\right)\ . \end{aligned}$$

Los valores de β se dan en la Tabla A.7.

3.5. Obtención del factor de tolerancia para la regiones de tolerancia robustas

Mediante el procedimiento descripto en la Sección 3.3 y utilizando los estimadores de Donoho–Stahel introducidos en la Sección 3.4, podemos definir una región de tolerancia robusta reemplazando los estimadores clásicos de posición y escala, por sus análogos robustos. La región entonces queda definida por

$$\mathcal{R}_{\mathrm{DS}} = \left\{\mathbf{y} : (\mathbf{y} - \mathbf{t}_n)'\,\mathbf{V}_n^{-1}\,(\mathbf{y} - \mathbf{t}_n) \leq K_{\mathrm{DS}}\right\} \tag{3.8}$$

donde $\mathbf{t}_n = \mathbf{t}_n\,(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ y $\mathbf{V}_n = \mathbf{V}_n\,(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ están definidos en (3.6) y (3.7).

Un algoritmo de simulación, análogo al considerado por Krishnamoorthy y Mathew (1999), fue desarrollado en MATLAB para aproximar el valor del factor de tolerancia K_{DS} para distintas combinaciones de n, d, q y δ. Ese algoritmo se describe brevemente a continuación

(i) Para cada j, se generan n vectores aleatorios $\mathbf{x}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$. Se calcula $\mathbf{t}_n$ y $\mathbf{V}_n$ definidos en (3.6) y (3.7).

(ii) Se generan R vectores aleatorios $\mathbf{y}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$ y para cada uno de ellos se evalúa la distancia de Mahalanobis al centro robusto según $\mathbf{V}_n$, es decir, $DM_i = (\mathbf{y}_i - \mathbf{t}_n)'\,\mathbf{V}_n^{-1}\,(\mathbf{y}_i - \mathbf{t}_n)$.

(iii) Se ordenan las formas cuadráticas $DM^{(1)} \leq \cdots \leq DM^{(R)}$ y se busca el percentil q de la forma cuadrática $DM^{(R\,q)}$. Denotemos u_j a este percentil.

(iv) Se repiten (i) a (iii) N veces, conservando u_j en cada iteración.

(v) Se ordenan los valores de u_j, $u^{(1)} \leq \cdots \leq u^{(N)}$ y se busca el percentil δ, $u^{(N\,\delta)}$ que es una aproximación al factor de tolerancia K_{DS}.

Esta distancia cuadrada $u^{(N\,\delta)}$ tiene la propiedad muestral de ser aquella que determina regiones con cobertura de al menos q 100 % en al menos el δ 100 % de los casos.

En las Tablas A.8, A.9 y A.10 se presentan, para $d = 2, 3, 4$ respectivamente, los valores de la constante de tolerancia K_{DS} que cumple (3.5) para distintos valores de q, δ y n. Se observa que estos valores son mayores que los factores de tolerancia clásicos, debido a la pérdida de eficiencia del estimador robusto.

Es de esperar que, como consecuencia de la menor eficiencia de los estimadores robustos de posición y escala, se obtengan regiones de mayor volumen. Para comparar los volúmenes de ambas regiones calculamos la raiz $1/d$ del cociente entre el volúmen de la región robusta y el de la región clásica. Las comparaciones para muestras de referencia de tamaños $n = 20, 30, 50, 100$, dimensiones $d = 2, 3, 4, 5, 8$, cobertura teórica de $q = 0.95$ y un nivel de confianza $\delta = 0.95$ se dan en la Tabla A.11. Los cálculos fueron hechos utilizando los pasos (i) a (iv) para el cálculo de la cobertura real descriptos en la Sección 2.2.1 del Capítulo 2, se tomaron $N = 1000$ $R = 1000$ y $M = 1000$.

Se ve que el precio pagado por la robustez en términos de incremento de volúmen de la región no es excesivo, excepto en configuraciones extremas como $d = 8$ y $n = 20$, donde el número de observaciónes no permiten obtener una buena aproximación para el supremo en la medida de atipicidad.

Como consecuencia de la ganancia en robustez, la región robusta presenta en todos los casos un volúmen mayor que la región clásica. Este incremento de la región se debe principalmente a la pérdida eficiencia de los estimadores multivariados de posición y escala utilizados.

3.6. Cálculo del error

Si R y N son lo suficientemente grandes (> 1000) podemos tener una idea del error cometido en el cálculo del factor de tolerancia mediante Monte Carlo. Tomaremos como medida del error la diferencia entre el K obtenido y el que obtendríamos si, en lugar de tomar los estadísticos de orden $R\,q$ y $N\,\delta$ utilizados en el algoritmo anterior, tomásemos $\rho(R, q)$ y $\rho(N, \delta)$, donde $\rho(R, q)$ es el estadístico de orden que satisface

$$P\left(DM \leq DM^{\rho(R,q)}\right) \geq q \quad \text{y} \quad P(K \leq K^{\rho(N,q)}) \geq \delta.$$

Como no conocemos las distribuciónes ni de DM ni de K calculamos $\rho(R,q)$ no paramétricamente como el valor ρ que satisface

$$P(Bi(R,q) \geq \rho + 1) \leq 1 - q \; .$$

Como R y N son lo suficientemente grandes y $q \leq$ 0,99 utilizando la aproximación normal obtenemos $\rho(R,q) = R\,q + 1.96\sqrt{R\,q\,(1-q)}$.

Por ejemplo, si $R = 5000$ y $q = 0.99$ obtenemos $R\,q = 4950$ y $\rho(R,q) = 4964$ mientras que si $R = 10000$ y $q = 0.99$ obtenemos $R\,q = 9900$ y $\rho(R,q) = 9920$.

Los valores estimados de estos errores se dan en la Tabla A.12, para niveles de cobertura $q = 0.90, 0.95$ y 0.99, niveles de confianza $\delta = 0.90, 0.95$ y 0.99, dimensiones $d = 2, 3, 4$ y distintos tamaños de muestra n.

3.7. Estudio de sensibilidad de la regiones de tolerancia robustas

En esta sección se presenta un estudio preliminar que permite en forma análoga a lo hecho para las regiones clásicas analizar el impacto de datos atípicos en la región de tolerancia robusta propuesta.

Se consideraron las regiones $\mathcal{R}_{\mathrm{DS}}$ definidas por (3.8) con los factores de tolerancia dados en las Tablas A.8, A.9 y A.10. El efecto por el agregado de un dato atípico externo se observa en la Tabla A.13 análoga a de la Tabla A.5 construída para la región clásica.

Se observa que el incremento del volúmen de la región de tolerancia es moderado, lo que produce un ligero incremento de la cobertura real, gracias a que el estimador de posición mantiene la región centrada. De esta forma, la presencia de datos atípicos externos severos sólo afecta las regiones de tolerancia incrementando ligeramente el volúmen de las mismas. En el caso clásico, el incremento de la cobertura real era mayor y se producía a expensas de un aumento desmedido del volumen.

Es razonable que el impacto del agregado de inliers sea más importante para la región robusta, pues los datos atípicos internos reciben un peso relativamente elevado dado que estos valores son los más "centrales" en todas las proyecciones y, por lo tanto, reciben una mayor ponderación en el cálculo de la matriz de covarianzas estimada. Este efecto se observa en la Tabla A.14.

3.8. Estudio de la cobertura de las regiones clásicas y robustas para distribuciónes alternativas

Tanto en el caso clásico como en el caso robusto, los factores de tolerancia fueron calculados con la finalidad de que la cobertura y la confianza real de las regiones coincidieran con las teóricas cuando las observaciónes que conforman la muestra de referencia provienen de una distribucón normal multivariada. Es interesante plantearse cual es el comportamiento de estas regiones cuando las observaciónes provienen de otras distribuciónes multivariadas, distintas a la normal, como ser la distribución esférica t, d-variada con g grados de libertad, que indicaremos $\mathcal{T}_g(d)$. La densidad de esta distribución está dada por

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{g+d}{2}\right)}{\Gamma\left(\frac{g}{2}\right)(g\pi)^{\frac{d}{2}}\left(1+\frac{1}{g}\|\mathbf{x}\|^2\right)^{\frac{g+d}{2}}} .$$

En el caso particular $g = 1$, esta distribución se llama distribución Cauchy d-variada, $\mathcal{C}_d$, y su densidad es entonces,

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{\frac{d+1}{2}}\left(1+\|\mathbf{x}\|^2\right)^{\frac{d+1}{2}}} .$$

Para generar muestras con distribución t multivariada, usamos el siguiente resultado que puede hallarse en Muirhead (1982).

Lema 3.8.1 *Sean $X_1, \ldots, X_n, Z$ ($Z > 0$) variables aleatorias tales que dado $Z = z$ la distribución de $\mathbf{X} = (X_1, \ldots, X_n)$ es $N(\mathbf{0}, z\,\mathbf{I})$. Si Z tiene distribución F la densidad de $\mathbf{X}$ es*

$$f(\mathbf{x}) = \int_0^\infty (2\pi\, z)^{\frac{n}{2}} \exp\left(-\frac{1}{2z}\sum_{i=1}^n x_i^2\right) dF(z) . \tag{3.9}$$

La densidad dada por (3.9), es esférica y se denomina una mezcla de normales. Se sigue que podemos generar observaciónes con densidad (3.9) generando $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I}_d)$ independiente de $Z \sim F$ y definiendo $\mathbf{x} = Z^{\frac{1}{2}}\mathbf{y}$. Observemos que si $P(Z = 1) = 1 - \epsilon$ y $P(Z = \sigma^2) = \epsilon$, $\mathbf{x}$ tiene una densidad normal contaminada, mientras que si $\frac{g}{Z} \sim \chi^2_g$, entonces $\mathbf{x} \sim \mathcal{T}_g(d)$.

Para estudiar el comportamiento de las regiones clásicas y robustas ante distintas distribuciónes alternativas, se evaluó la cobertura real cuando las muestras de referencia tienen distribución G mediante el siguiente algoritmo

(i) Se generan n vectores aleatorios $\mathbf{x}_i \sim G$. Se calculan los estimadores robustos de posición y escala multivariados $\mathbf{t}_n$ y $\mathbf{V}_n$ definidos en (3.6) y (3.7).

(ii) Se generan R vectores aleatorios $\mathbf{y}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$ y para cada uno de ellos se verifica si yace o no dentro de la region $\mathcal{R}_{\mathrm{DS}}$

$$\mathcal{R}_{\mathrm{DS}} = \left\{\mathbf{y} : (\mathbf{y} - \mathbf{t}_n)' \, \mathbf{V}_n^{-1} \, (\mathbf{y} - \mathbf{t}_n) \leq K\right\}$$

Si $\mathbf{y}_i \in \mathcal{R}_{\mathrm{DS}}$ se define $c_i = 1$ en caso contrario, $c_i = 0$. Se calcula la cobertura promedio $\overline{c}$ como $\overline{c} = \dfrac{1}{R}\displaystyle\sum_{i=1}^{R} c_i$.

(iii) Se repiten (i) y (ii) N veces, conservando $\overline{c}$ en cada iteración.

(iv) Se ordenan las coberturas promedio $\overline{c}^{(1)} \leq \cdots \leq \overline{c}^{(N)}$ y se preserva la $N(1-\delta)-$ésima cobertura promedio, $\overline{c}^{(N\,(1-\delta))} = \pi_{\mathrm{DS}}$, que aproxima a la cobertura real con nivel de confianza δ.

De igual forma se procede en el caso clásico, donde se indicará por π_{C} a la cobertura así obtenida.

Los resultados de cobertura real para las regiones clásica y robusta, con $\delta = 0.95$ y $q = 0.95$, se presentan en las Tablas indicadas entre paréntesis para las siguientes distribuciónes G

- la distribución normal sin contaminar $G = N(\mathbf{0}, \mathbf{I}_d)$ (Tabla A.11).
- la distribución t–multivariada con g grados de libertad $G = \mathcal{T}_g(d)$, con $g = 1$, 2, 3 (Tablas A.15, A.16 y A.17, respectivamente).
- la distribución $G = (1-\epsilon)\, N(\mathbf{0}, \mathbf{I}_d) + \epsilon\, \mathcal{C}_d$ con $\epsilon = 0.05$ y 0.10 (Tablas A.18 y A.19, respectivamente).
- la distribución $G = (1-\epsilon)\, N(\mathbf{0}, \mathbf{I}_d) + \epsilon\, N(\mathbf{0}, 25\, \mathbf{I}_d)$ con $\epsilon = 0.05$ y 0.10 (Tablas A.20 y A.21, respectivamente).
- la distribución normal con outliers a distancia 8 y 16 en la dirección $\mathbf{e}_1 \in R^d$. (Tablas A.22 y A.23, respectivamente).

Se tomaron datos en dimensión $d = 2$, 3, 4, 5 y 8 y tamaños de muestra iguales a $n = 20$, 30, 50 y 100.

Con la finalidad de lograr un mejor entendimiento de la diferencia entre ambas regiones, y sabiendo que en el caso de la distribución Cauchy, por ejemplo, la media

muestral también tiene una distribución Cauchy, calculamos la distancia al cero (centro de la distribución) de las 2 regiones. Esto se realiza mediante el cociente entre la mediana de las normas de los centros de las regiones clásicas ($\|\overline{\mathbf{x}}\|$) y la mediana de las normas de los centros de la región robusta ($\|\mathbf{t}_n\|$) sobre las N replicaciones. También, se reporta en las Tablas la raíz $1/d$ del cociente entre la mediana del volúmen de la región clásica ($\mathcal{V}_{\mathrm{C}}$) y la mediana del volúmen de la región robusta ($\mathcal{V}_{\mathrm{DS}}$).

En el caso de la distribución Cauchy, la Tabla A.15 permite observar que, tanto para las regiones clásicas como para las regiones robustas, la cobertura real es superior a la teórica. El fenómeno más destacable es, sin embargo, el desmedido incremento del volúmen de las regiones clásicas en relación a las robustas, como consecuencia de la falta de robustez de los estimadores utilizados en la construcción de dichas regiones. Otra diferencia importante consiste en la mejora en el centrado de la región, debido al estimador de posición robusto, que surge de la observación de los elevados valores de la última columna de la Tabla A.15. Tomando el caso particular $d = 4$ y $n = 30$, vemos que con ambas regiones obtenemos, con un nivel de confianza del 95 %, probabilidades de cobertura muy superiores a la teórica del 95 %. La mayor probabilidad de cobertura de la región clásica se acompaña con un incremento de más de 81 veces (2.99^4) el volúmen de la región y con un centro que se encuentra 6 veces más alejado del 0 que el centro de la región robusta.

Las Tablas A.16 y A.17 dan los resultados cuando las muestras se generan con las distribuciónes $\mathcal{T}_2(d)$ y $\mathcal{T}_3(d)$, respectivamente. Como era de esperar, a medida que aumentan los grados de libertad de la distribución t, el efecto del incremento del volúmen de la región clásica en relación a la robusta se atenúa, al igual que la diferencia en los centrados de ambas regiones. Esto se debe al hecho que al aumentar los grados de libertad nos acercamos a la normalidad, donde encontrabamos el fenómeno opuesto, más moderado, en el que la menor eficiencia de los estimadores de posición y escala redundaban en un ligero incremento del volúmen de la región robusta.

Con la finalidad de analizar el comportamiento de ambas regiones en situaciones más realistas, en las cuales se espera que sólo una pequeña proporción de la muestra se aleje de la normalidad, contaminamos la muestra de referencia con un 5 % y un 10 % de observaciones con distribución Cauchy (Tablas A.18 y A.19, respectivamente). En ellas vemos que, en casi todos los casos, la región robusta se halla más cerca de la cobertura teórica, con volúmenes inferiores y un mejor centrado, de lo que lo está la región clásica. Este fenómeno se hace más importante a medida que aumentan tanto la dimensión d, como el tamaño de muestra n y el porcentaje de contaminación. Especial atención requieren los casos $n = 20$ en los cuales lo anteriormente dicho no es válido. Esto se debe a que la contaminación no es tan importante como para revertir el efecto de mayor volumen y probabilidad de cobetura de la región robusta ya observado en la tabla A.11. Similares conclusiones pueden obtenerse cuando la contaminación se realiza con datos provenientes de una distribución normal pero con una matriz de

escala mayor ($25\,\mathbf{I}_d$), (Tablas A.20 y A.21, respectivamente).

Capítulo 4

Función de Influencia de la Probabilidad de Cobertura

4.1. Introducción

En adelante, pensaremos a los estimadores en base a los cuales construimos nuestras regiones de tolerancia, como funcionales en el espacio de las funciones de distribución evaluados en la distribución empírica.

Recordemos que la función de influencia describe el efecto, en el funcional de interés, de una contaminación infinitesimal en el punto $\mathbf{x}$, estandarizada por la masa de la contaminación.

El funcional del que nos interesa calcular la función de influencia es la probabilidad de cobertura (Pc), para un nivel de confianza fijo δ, una cobertura teórica particular q, y una constante de corte K, es decir,

$$\begin{aligned} Pc\left(G,F\right) &= P_F\left(\left(\mathbf{x}-\mathbf{T}\left(G\right)\right)'\mathbf{V}\left(G\right)^{-1}\left(\mathbf{x}-\mathbf{T}\left(G\right)\right)\leq K\right) \text{ con x}\sim F \\ &= \int I_{\mathcal{R}(G)}\left(\mathbf{x}\right)dF\left(\mathbf{x}\right) \end{aligned}$$

donde

- $\mathcal{R}(G)=\{\mathbf{x}:\ \left(\mathbf{x}-\mathbf{T}\left(G\right)\right)'\mathbf{V}\left(G\right)^{-1}\left(\mathbf{x}-\mathbf{T}\left(G\right)\right)\leq K\}$
- $\mathbf{T}\left(G\right)$ es el parámetro de posición de la distribución G,
- $\mathbf{V}\left(G\right)$ es la matriz de escala de la distribución G y
- K es el factor de tolerancia.

Para un tamaño de muestra n dado, un nivel δ, una cobertura teórica q, la constante K dependía de δ, q, n y de la dimensión d. Con la notación de los capítulos anteriores $K = K(\delta, q, n, d)$. En este Capítulo supondremos la constante fija.

Es importante observar que este funcional depende de dos distribuciónes, F y G. Más precisamente, la probabilidad de cobertura $Pc(G, F)$ depende de la distribución respecto de la cual calculamos la cobertura F y de la distribución de la muestra de referencia G con la cual estimamos los parámetros de posición y escala. No hacemos supuestos sobre la distribución G, ya que la región de tolerancia puede evaluarse sobre cualquier conjunto de datos, aunque esperamos que utilizando estimadores clásicos la región de tolerancia sea mala si la distribución de los datos está lejos de la distribución normal. Cuando se desean regiones de coberturas parámetricas, en muchas situaciones como las que hemos considerado se supone $F = G = N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Es decir, bajo el modelo, los datos con los que construímos la región siguen el modelo normal y en ese caso, supondremos que $\mathbf{T}(G) = \boldsymbol{\mu}$, $\mathbf{V}(G) = \boldsymbol{\Sigma}$, o sea, que el procedimiento de estimación proviene de funcionales Fisher–consistentes. Para evaluar la función de influencia sólo consideraremos contaminación en la muestra de referencia G y no en la de los datos futuros F que permite calcular la probabilidad de cobertura. Es decir, es la función de distribución G de la cual proviene la muestra de referencia la que contaminamos en el punto $\mathbf{x}$, pues nos interesa saber cual es el efecto en la probabilidad de cobertura, de alejamientos de la normalidad en la muestra de referencia. Cuando evaluemos la función de influencia, así como en discriminación, supondremos que ambas distribuciónes coinciden, lo cual es una suposición hecha frecuentemente en la literatura en forma implícita.

Para simplificar el cálculo de la función de influencia, supondremos que $F = N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. El Lema siguiente da una expresión para la probabilidad de cobertura, en forma análoga a la dada por Croux y Joossens (2004) para la probabilidad total de mala clasificación.

Lema 4.1.1. *Sea F una distribución elipsoidal con parámetro de posición $\boldsymbol{\mu}$ y matriz de escala $\boldsymbol{\Sigma}$ y sean $\mathbf{T}$ y $\mathbf{V}$ funcionales de posición y escala multivariados. Supongamos que $\mathbf{x} = \boldsymbol{\mu} + \mathbf{C}\mathbf{z}$ donde $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$ y $\mathbf{z} \sim F_0$ con F_0 esférica. Sea $\mathbf{A}(G)$ la matriz $\mathbf{A}(G) = \mathbf{C}^{-1}\mathbf{V}(G)(\mathbf{C}^{-1})'$ y consideremos $\boldsymbol{\beta}(G)$ la matriz de autovectores de $\mathbf{A}(G)$, $\boldsymbol{\beta}(G)'\boldsymbol{\beta}(G) = \mathbf{I}_d$, $\boldsymbol{\Lambda}(G) = diag(\lambda_1(G), \ldots, \lambda_d(G))$, la matriz de autovalores, $\lambda_1(G) \geq \cdots \geq \lambda_d(G) \geq 0$, es decir, $\mathbf{A}(G) = \boldsymbol{\beta}(G)\boldsymbol{\Lambda}(G)\boldsymbol{\beta}(G)'$. Entonces,*

$$\begin{aligned} Pc(G, F) &= P_{F_0}\left((\mathbf{z} - \boldsymbol{\tau}(G))'\mathbf{A}(G)^{-1}(\mathbf{z} - \boldsymbol{\tau}(G)) \leq K\right) \\ &= P_{F_0}\left(\sum_{i=1}^{d}\left(\frac{z_i - \tau_i(G)}{\sqrt{\lambda_i(G)}}\right)^2 \leq K\right) \end{aligned} \tag{4.1}$$

donde $\boldsymbol{\tau}(G) = \boldsymbol{\beta}(G)'\mathbf{C}^{-1}(\mathbf{T}(G) - \boldsymbol{\mu})$.

En particular, si $F_0 = N_d(\mathbf{0}, \mathbf{I}_d)$, *se tiene*

$$Pc(G,F) = \int I_{\mathcal{S}}(\mathbf{y}) \prod_{i=1}^{d} \sqrt{\lambda_i(G)}\, \varphi\left(\sqrt{\lambda_i(G)}\, y_i + \tau_i(G)\right)\, d\mathbf{y}. \tag{4.2}$$

donde $\mathcal{S} = \{\mathbf{y} : \sum_{i=1}^{d} y_i^2 \leq K\}$ *y* $\varphi(t)$ *indica la densidad de una variable con distribución* $N(0,1)$.

Demostración. Como $\mathbf{x} = \boldsymbol{\mu} + \mathbf{C}\mathbf{z}$ y $\mathbf{A}(G) = \mathbf{C}^{-1}\mathbf{V}(G)(\mathbf{C}^{-1})' = \boldsymbol{\beta}(G)\boldsymbol{\Lambda}(G)\boldsymbol{\beta}(G)'$ es fácil ver que

$$\begin{aligned} Pc(G,F) &= P_F\left((\mathbf{x} - \mathbf{T}(G))'\mathbf{V}(G)^{-1}(\mathbf{x} - \mathbf{T}(G)) \leq K\right) \\ &= P_{F_0}\left((\mathbf{z} - \mathbf{C}^{-1}(\mathbf{T}(G) - \boldsymbol{\mu}))'\mathbf{A}(G)^{-1}(\mathbf{z} - \mathbf{C}^{-1}(\mathbf{T}(G) - \boldsymbol{\mu})) \leq K\right) \\ &= P_{F_0}\left((\boldsymbol{\beta}(G)'\mathbf{z} - \boldsymbol{\tau}(G))'\boldsymbol{\Lambda}(G)^{-1}(\boldsymbol{\beta}(G)'\mathbf{z} - \boldsymbol{\tau}(G)) \leq K\right) \end{aligned}$$

Como F_0 es esférica tenemos que $\boldsymbol{\beta}(G)'\mathbf{z}$ tiene la misma distribución que $\mathbf{z}$ luego

$$\begin{aligned} Pc(G,F) &= P_{F_0}\left((\mathbf{z} - \boldsymbol{\tau}(G))'\boldsymbol{\Lambda}(G)^{-1}(\mathbf{z} - \boldsymbol{\tau}(G)) \leq K\right) \\ &= P_{F_0}\left(\sum_{i=1}^{d} \frac{1}{\lambda_i(G)} (z_i - \tau_i(G))^2 \leq K\right) \\ &= P_{F_0}\left(\sum_{i=1}^{d} \left(\frac{z_i - \tau_i(G)}{\sqrt{\lambda_i(G)}}\right)^2 \leq K\right). \end{aligned}$$

Si $F_0 = N_d(\mathbf{0}, \mathbf{I}_d)$, sean $y_i = \dfrac{z_i - \tau_i(G)}{\sqrt{\lambda_i(G)}}$, entonces $y_i \sim N\left(\dfrac{-\tau_i(G)}{\sqrt{\lambda_i(G)}}, \dfrac{1}{\lambda_i(G)}\right)$ independientes. Indiquemos por F_1 la distribución de $\mathbf{y}$. Usando (4.1), se obtiene

$$\begin{aligned} Pc(G,F) &= P_{F_1}\left(\sum_{i=1}^{d} y_i^2 \leq K\right) = E_{F_1}(I_{\mathcal{S}}(\mathbf{y})) \\ &= \int I_{\mathcal{S}}(\mathbf{y}) \prod_{i=1}^{d} f_{Y_i}(y_i)\, d\mathbf{y} \\ &= \int I_{\mathcal{S}}(\mathbf{y}) \prod_{i=1}^{d} \sqrt{\lambda_i(G)}\, \varphi\left(\sqrt{\lambda_i(G)}\, y_i + \tau_i(G)\right)\, d\mathbf{y}\ . \square \end{aligned}$$

4.2. Función de influencia

Como es bien sabido, la función de influencia es una medida de la robustez respecto de outliers o datos atípicos. Esencialmente, la función de influencia es la derivada de orden uno de la versión funcional del estimador. A partir de ella, varios autores han construído medidas diagnóstico de detección de outliers. Utilizando la definición de Hampel (1974), tenemos que la función de influencia, para el funcional Pc, en el punto $\mathbf{x}$ y la distibución G se define como:

$$\begin{aligned} \mathrm{IF}\left(\mathbf{x}, Pc, G\right) &= \lim_{\epsilon \to 0} \frac{Pc\left(\left(1-\epsilon\right)G + \epsilon\Delta_{\mathbf{x}}, F\right) - Pc\left(G, F\right)}{\epsilon} \\ &= \frac{\partial}{\partial \epsilon} Pc\left(G_{\epsilon,\mathbf{x}}, F\right)\Big|_{\epsilon=0} \end{aligned}$$

donde, de ahora en adelante, indicamos por $G_{\epsilon,\mathbf{x}} = \left(1-\epsilon\right)G + \epsilon\Delta_{\mathbf{x}}$ a la distribución contaminada con $\Delta_{\mathbf{x}}$ la masa puntual en $\mathbf{x}$.

Es conveniente tener en cuenta que el funcional $Pc\left(G_{\epsilon,\mathbf{x}}, F\right)$ depende de los funcionales de posición y escala $\mathbf{T}\left(G_{\epsilon,\mathbf{x}}\right)$ y $\mathbf{V}\left(G_{\epsilon,\mathbf{x}}\right)$. Es la distribución G de estos últimos dos funcionales la que se halla contaminada en una proporción ϵ por una masa de probabilidad puntual en $\mathbf{x}$.

Calcularemos la función de influencia en dos situaciones distintas. En la primera, las distribuciónes G y F son tales que los autovalores de la matriz $\mathbf{A}(G)$ son todos distintos. En la segunda, $F = G$ y $\mathbf{T}$ y $\mathbf{V}$ son funcionales Fisher–consistentes, es decir, $\mathbf{T}(G) = \boldsymbol{\mu}$ y $\mathbf{V}(G) = \boldsymbol{\Sigma}$. En ese caso, $\boldsymbol{\tau}(G) = \mathbf{0}$ y $\mathbf{A}(G) = \mathbf{I}_d$. Por lo tanto, existe un problema para calcular la función influencia de los autovectores de $\mathbf{A}(G)$ en $G = F$ ya que no están univocamente determinados. Para resolver ese problema definimos los autovectores de $\mathbf{A}(G_{\epsilon,\mathbf{x}})$, para ϵ en un entorno de 0 de modo tal que los elementos diagonales de la matriz $\boldsymbol{\beta}(G_{\epsilon,\mathbf{x}})$ sean positivos. Esto asegura la continuidad en $G = F$ de $\boldsymbol{\beta}(G)$.

Teorema 4.2.1. *Sean* $\mathbf{T}$ *y* $\mathbf{V}$ *funcionales de posición y escala multivariados. Sea* $F = N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, *con* $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$ *y* G *una distribución tal que los autovalores,* $\lambda_1 \geq \cdots \geq \lambda_d$, *de* $\mathbf{A}(G) = \mathbf{C}^{-1}\mathbf{V}(G)(\mathbf{C}^{-1})'$ *son todos distintos. Sea* $\boldsymbol{\beta}_j$ *el autovector asociado a* λ_j. *Indiquemos por* $\boldsymbol{\tau} = \boldsymbol{\tau}(G) = \boldsymbol{\beta}'\mathbf{C}^{-1}(\mathbf{T}(G) - \boldsymbol{\mu})$ *y por* F_1 *la distribución de* $\mathbf{y} = (y_1, \ldots, y_d)'$, *donde* $y_1, \ldots, y_d$ *son independientes tales que* $y_i \sim N\left(\dfrac{-\tau_i\left(G\right)}{\sqrt{\lambda_i\left(G\right)}}, \dfrac{1}{\lambda_i\left(G\right)}\right)$ *para* $1 \leq i \leq d$.

Supongamos que existen IF$(\mathbf{x}, \mathbf{T}, G)$ *y* IF$\left(\mathbf{x}, \mathbf{V}, G\right)$. *Entonces, la función de influ-*

encia de $Pc(G)$ está dada por:

$$\begin{aligned} IF\left(x, Pc, G\right) &= \sum_{j=1}^{d} \boldsymbol{\beta}_j' \mathbf{C}^{-1} IF\left(\mathbf{x}, \mathbf{V}, G\right) \left(\mathbf{C}^{-1}\right)' \boldsymbol{\beta}_j \times \\ &\quad \times \frac{1}{2\lambda_j} \left[P_{F_1}\left(\mathcal{S}\right) - E_{F_1}\left(I_{\mathcal{S}}(\mathbf{y}) y_j \left(\lambda_j \, y_j + \sqrt{\lambda_j} \tau_j \right) \right) \right] - \\ &- \sum_{j=1}^{d} \left(\sqrt{\lambda_j} E_{F_1}\left(y_i I_{\mathcal{S}}\left(\mathbf{y}\right) \right) + \tau_j P_{F_1}\left(\mathcal{S}\right) \right) \times \\ &\quad \times \left(\sum_{i \neq j} \frac{\boldsymbol{\beta}_i' \mathbf{C}^{-1} IF\left(\mathbf{x}, \mathbf{V}, G\right) \left(\mathbf{C}^{-1}\right)' \boldsymbol{\beta}_j}{\lambda_j - \lambda_i} \, \tau_i + \boldsymbol{\beta}_j' \mathbf{C}^{-1} IF(\mathbf{x}, \mathbf{T}, G) \right) . \end{aligned}$$

Observación 4.2.1. Observemos que si $\mathbf{T}(G) = \boldsymbol{\mu}$, $\tau_i = 0$ para todo $1 \leq i \leq d$, con lo cual se obtiene

$$\mathrm{IF}\left(x, Pc, G\right) = \sum_{j=1}^{d} \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{V}, G\right) \left(\mathbf{C}^{-1}\right)' \boldsymbol{\beta}_j \, \frac{P_{F_1}\left(\mathcal{S}\right) - \lambda_j E_{F_1}\left(I_{\mathcal{S}}(\mathbf{y}) y_j^2 \right)}{2\lambda_j} .$$

Por lo tanto, la función de influencia de la probabilidad de cobertura no depende de la función de influencia del funcional de posición ni de la expresión asociada a la influencia de los autovectores. Se elimina en esta expresión el problema de tener autovalores cercanos. Esa expresión es la que obtendremos en el siguiente Teorema que da la función de influencia cuando $G = F$. En particular, si $\boldsymbol{\Sigma} = \mathbf{I}_d$, $T(G) = \boldsymbol{\mu}$ y $\mathbf{V}(G) = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, con $\lambda_1 > \ldots > \lambda_d$, se tiene

$$\mathrm{IF}\left(x, Pc, G\right) = \sum_{j=1}^{d} \mathrm{IF}\left(\mathbf{x}, \mathbf{V}, G\right)_{jj} \, \frac{P_{F_1}\left(\mathcal{S}\right) - \lambda_j E_{F_1}\left(I_{\mathcal{S}}(\mathbf{y}) y_j^2 \right)}{2\lambda_j} .$$

Teorema 4.2.2. *Sea $F = G = N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, con $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$. Sean $\mathbf{T}$ y $\mathbf{V}$ funcionales Fisher–consistentes en G, es decir, $\mathbf{T}(G) = \boldsymbol{\mu}$ y $\mathbf{V}(G) = \boldsymbol{\Sigma}$.*

Supongamos que existen $IF(\mathbf{x}, \mathbf{T}, G)$ y $IF\left(\mathbf{x}, \mathbf{V}, G\right)$. Entonces, la función de influencia de $Pc(G)$ está dada por:

$$\begin{aligned} IF\left(x, Pc, G\right) &= \frac{1}{2} \left(P\left(W_d \leq K\right) - \frac{1}{d} E\left(W_d I_{(0,K]}\left(W_d\right) \right) \right) tr\left(IF\left(\mathbf{x}, \mathbf{V}, G\right) \boldsymbol{\Sigma}^{-1} \right) \\ &= \frac{1}{2} \left(P\left(W_d \leq K\right) - P\left(W_{d+2} \leq K\right) \right) tr\left(IF\left(\mathbf{x}, \mathbf{V}, G\right) \boldsymbol{\Sigma}^{-1} \right) , \end{aligned}$$

donde W_d es una variable aleatoria tal que $W_d \sim \chi_d^2$.

En particular, si $\mathbf{\Sigma} = \mathbf{I}_d$ *se verifica*

$$IF(x, Pc, G) = \frac{1}{2}\left(P\left(W_d \leq K\right) - P\left(W_{d+2} \leq K\right)\right) tr\left(IF\left(\mathbf{x}, \mathbf{V}, G\right)\right) .$$

Este Teorema muestra que la probabilidad de cobertura tendrá función de influencia acotada si la matriz de escala utilizada tiene influencia acotada.

Para probar estos Teoremas necesitaremos el siguiente Lema.

Lema 4.2.1. *Sea* $\mathbf{\Lambda} = diag\left(\lambda_1, \ldots, \lambda_d\right)$ *y* $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)'$. *Sea*

$$H(\mathbf{\Lambda}, \tau) = \int I_{\mathcal{S}}\left(\mathbf{y}\right) \prod_{i=1}^{d} \sqrt{\lambda_i}\, \varphi\left(\sqrt{\lambda_i}\, y_i + \tau_i\right)\, d\mathbf{y}$$

donde φ *es la densidad de una variable* $N(0,1)$ *y* $\mathcal{S} = \{\mathbf{y} : \|\mathbf{y}\|^2 \leq K\}$. *Sea* F_1 *la distribución de* $\mathbf{y} = (y_1, \ldots, y_d)'$, *donde* $y_1, \ldots, y_d$ *son independientes tales que* $y_i \sim N\left(\frac{-\tau_i\left(G\right)}{\sqrt{\lambda_i\left(G\right)}}, \frac{1}{\lambda_i\left(G\right)}\right)$ *para* $1 \leq i \leq d$. *Entonces, la función* H *es derivable y*

$$\begin{aligned}
\frac{\partial H}{\partial \tau_i} &= -\left[\sqrt{\lambda_i}\, E_{F_1}\left(I_{\mathcal{S}}(\mathbf{y}) y_i\right) + \tau_i P_{F_1}\left(\mathcal{S}\right)\right] \\
\frac{\partial H}{\partial \lambda_i} &= \frac{1}{2\lambda_i}\left[P_{F_1}\left(\mathcal{S}\right) - E_{F_1}\left(I_{\mathcal{S}}(\mathbf{y}) y_i \left(\lambda_i y_i + \sqrt{\lambda_i}\, \tau_i\right)\right)\right] .
\end{aligned}$$

Demostración. La demostración es análoga a la del Lema 1 de Croux y Joossens (2004). Utilizando que $\varphi'(t) = -t\varphi(t)$, la definición de H y el hecho de que podemos derivar bajo el signo integral obtenemos:

$$\begin{aligned}
\frac{\partial H}{\partial \tau_i} &= \int I_{\mathcal{S}}\left(\mathbf{y}\right) \prod_{j \neq i} \sqrt{\lambda_j}\, \varphi\left(\sqrt{\lambda_j}\, y_j + \tau_j\right) \sqrt{\lambda_i}\, \frac{\partial \varphi\left(\sqrt{\lambda_i}\, y_i + \tau_i\right)}{\partial \tau_i}\, d\mathbf{y} \\
&= -\int I_{\mathcal{S}}\left(\mathbf{y}\right) \prod_{j=1}^{d} \sqrt{\lambda_j}\, \varphi\left(\sqrt{\lambda_j}\, y_j + \tau_j\right) \left(\sqrt{\lambda_i}\, y_i + \tau_i\right)\, d\mathbf{y} \\
&= -\int I_{\mathcal{S}}\left(\mathbf{y}\right) \left(\sqrt{\lambda_i}\, y_i + \tau_i\right) f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} .
\end{aligned}$$

Por otra parte,

$$\begin{aligned}
\frac{\partial H}{\partial \lambda_i} &= \int I_{\mathcal{S}}(\mathbf{y}) \prod_{j \neq i} \sqrt{\lambda_j}\, \varphi\left(\sqrt{\lambda_j}\, y_j + \tau_j\right) \frac{\partial \left\{\sqrt{\lambda_i}\, \varphi\left(\sqrt{\lambda_i}\, y_i + \tau_i\right)\right\}}{\partial \lambda_i}\, d\mathbf{y} \\
&= \int I_{\mathcal{S}}(\mathbf{y}) \prod_{j=1}^{d} \sqrt{\lambda_j}\, \varphi\left(\sqrt{\lambda_j}\, y_j + \tau_j\right) \left[\frac{1}{2\lambda_i} - \frac{1}{2\lambda_i}\left(\lambda_i\, y_i + \sqrt{\lambda_i}\tau_i\right) y_i\right]\, d\mathbf{y} \\
&= \frac{1}{2\lambda_i}\left[P_{F_1}(\mathcal{S}) - \int I_{\mathcal{S}}(\mathbf{y})\, y_i \left(\lambda_i\, y_i + \sqrt{\lambda_i}\tau_i\right) f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}\right],
\end{aligned}$$

lo que prueba el resultado. □

Demostración del Teorema 4.2.1. Para calcular la función de influencia usaremos (4.2). Como $Pc(G,F) = H(\boldsymbol{\Lambda}(G), \tau(G))$, aplicando la regla de la cadena obtenemos:

$$\begin{aligned}
\mathrm{IF}(x, Pc, G) &= \left.\frac{\partial Pc\left(G_{\epsilon,\mathbf{x}}, F\right)}{\partial \epsilon}\right|_{\epsilon=0} \\
&= \sum_{i=1}^{d} \left.\frac{\partial H}{\partial \lambda_i}\right|_{(\boldsymbol{\Lambda},\boldsymbol{\tau})} \left.\frac{\partial \lambda_i\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0} + \sum_{i=1}^{d} \left.\frac{\partial H}{\partial \tau_i}\right|_{(\boldsymbol{\Lambda},\boldsymbol{\tau})} \left.\frac{\partial \tau_i\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0}, \qquad (4.3)
\end{aligned}$$

donde $(\boldsymbol{\Lambda}, \boldsymbol{\tau}) = (\boldsymbol{\Lambda}(G), \boldsymbol{\tau}(G))$. Como los autovalores de $\mathbf{A}(G) = \mathbf{C}^{-1}\mathbf{V}(G)(\mathbf{C}^{-1})'$ tienen multiplicidad 1, por el Lema 3 de Croux y Haesbroeck (2000) tenemos

$$\begin{aligned}
\mathrm{IF}\left(\mathbf{x}, \boldsymbol{\beta}_j, G\right) &= \sum_{i \neq j} \frac{\boldsymbol{\beta}_i' \mathrm{IF}(\mathbf{x}, \mathbf{A}, G)\, \boldsymbol{\beta}_j}{\lambda_j - \lambda_i}\, \boldsymbol{\beta}_i \\
&= \sum_{i \neq j} \frac{\boldsymbol{\beta}_i' \mathbf{C}^{-1} \mathrm{IF}(\mathbf{x}, \mathbf{V}, G)\, (\mathbf{C}^{-1})' \boldsymbol{\beta}_j}{\lambda_j - \lambda_i}\, \boldsymbol{\beta}_i \qquad (4.4) \\
\mathrm{IF}(\mathbf{x}, \lambda_j, G) &= \boldsymbol{\beta}_j' \mathrm{IF}(\mathbf{x}, \mathbf{A}, G)\, \boldsymbol{\beta}_j \\
&= \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}(\mathbf{x}, \mathbf{V}, G)\, (\mathbf{C}^{-1})' \boldsymbol{\beta}_j\,. \qquad (4.5)
\end{aligned}$$

Luego, usando que $\tau_j\left(G_{\epsilon,\mathbf{x}}\right) = \boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right)' \mathbf{C}^{-1}\left(\mathbf{T}\left(G_{\epsilon,\mathbf{x}}\right) - \boldsymbol{\mu}\right)$, donde $\boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right)$ es el $j-$ésimo autovector de la matriz $\mathbf{A}\left(G_{\epsilon,\mathbf{x}}\right)$, obtenemos

$$\begin{aligned}
\left.\frac{\partial \tau_j\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0} &= \left.\frac{\partial}{\partial \epsilon}\left\{\boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right)' \mathbf{C}^{-1}\left(\mathbf{T}\left(G_{\epsilon,\mathbf{x}}\right) - \boldsymbol{\mu}\right)\right\}\right|_{\epsilon=0} \\
&= \mathrm{IF}\left(\mathbf{x}, \boldsymbol{\beta}_j, G\right)' \mathbf{C}^{-1}\left(\mathbf{T}(G) - \boldsymbol{\mu}\right) + \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}(\mathbf{x}, \mathbf{T}, G) \\
&= \sum_{i \neq j} \frac{\boldsymbol{\beta}_i' \mathbf{C}^{-1} \mathrm{IF}(\mathbf{x}, \mathbf{V}, G)\, (\mathbf{C}^{-1})' \boldsymbol{\beta}_j}{\lambda_j - \lambda_i} \tau_i + \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}(\mathbf{x}, \mathbf{T}, G) \qquad (4.6)
\end{aligned}$$

Luego, usando (4.5), (4.6), de (4.3) deducimos

$$\begin{aligned}\mathrm{IF}\left(\mathbf{x}, Pc, G\right) &= \sum_{j=1}^{d} \left.\frac{\partial H}{\partial \lambda_j}\right|_{(\boldsymbol{\Lambda},\boldsymbol{\tau})} \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{V}, G\right) \left(\mathbf{C}^{-1}\right)' \boldsymbol{\beta}_j \\ &+ \sum_{j=1}^{d} \left.\frac{\partial H}{\partial \tau_j}\right|_{(\boldsymbol{\Lambda},\boldsymbol{\tau})} \left[\sum_{i \neq j} \frac{\boldsymbol{\beta}_i' \mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{V}, G\right) \left(\mathbf{C}^{-1}\right)' \boldsymbol{\beta}_j}{\lambda_j - \lambda_i} \tau_i + \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{T}, G\right)\right]\end{aligned}$$

con lo cual usando el Lema 4.2.1, se obtiene el resultado. □

Demostración del Teorema 4.2.2. Es análoga a la del Teorema 4.2.1. Observemos que por (4.3) sólo debemos calcular $\left.\frac{\partial \tau_j\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0}$ y $\left.\frac{\partial \lambda_j\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0}$. Usando que $\tau_j\left(G\right) = 0$ ya que $\mathbf{T}(G) = \boldsymbol{\mu}$ y que elegimos los autovectores de $\mathbf{A}(G)$ de modo que sus elementos diagonales sean positivos, obtenemos por la continuidad de $\boldsymbol{\beta}\left(G_{\epsilon,\mathbf{x}}\right)$

$$\begin{aligned}\left.\frac{\partial \tau_j\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0} &= \lim_{\epsilon \to 0} \boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right)' \mathbf{C}^{-1} \frac{\mathbf{T}\left(G_{\epsilon,\mathbf{x}}\right) - \mathbf{T}(G)}{\epsilon} \\ &= \boldsymbol{\beta}_j\left(G\right)' \mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{T}, G\right) \ . \end{aligned} \tag{4.7}$$

Por otra parte, como $\mathbf{A}(G) = \mathbf{I}_p$, $\lambda_j(G) = 1$ y $\boldsymbol{\beta}_j(G) = \mathbf{e}_j$ el $j-$ésimo vector de la base canónica, usando la ortogonalidad de los autovectores tenemos

$$\begin{aligned}\left.\frac{\partial \lambda_j\left(G_{\epsilon,\mathbf{x}}\right)}{\partial \epsilon}\right|_{\epsilon=0} &= \lim_{\epsilon \to 0} \frac{\boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right)' \mathbf{A}(G_{\epsilon,\mathbf{x}}) \boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right) - 1}{\epsilon} \\ &= \lim_{\epsilon \to 0} \boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right)' \frac{\mathbf{A}(G_{\epsilon,\mathbf{x}}) - \mathbf{I}_d}{\epsilon} \boldsymbol{\beta}_j\left(G_{\epsilon,\mathbf{x}}\right) \\ &= \boldsymbol{\beta}_j' \mathrm{IF}\left(\mathbf{x}, \mathbf{A}, G\right) \boldsymbol{\beta}_j = \mathrm{IF}\left(\mathbf{x}, \mathbf{A}, G\right)_{jj} \ . \end{aligned} \tag{4.8}$$

Luego, usando (4.7) y (4.8) de (4.3) deducimos

$$\begin{aligned}\mathrm{IF}\left(\mathbf{x}, Pc, G\right) &= \sum_{j=1}^{d} \left.\frac{\partial H}{\partial \lambda_j}\right|_{(\mathbf{I}_d,0)} \left(\mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{V}, G\right) \left(\mathbf{C}^{-1}\right)'\right)_{jj} \\ &+ \sum_{j=1}^{d} \left.\frac{\partial H}{\partial \tau_j}\right|_{(\mathbf{I}_d,0)} \boldsymbol{\beta}_j' \mathbf{C}^{-1} \mathrm{IF}\left(\mathbf{x}, \mathbf{T}, G\right) \ .\end{aligned}$$

Sea $F_0 = N_d(\mathbf{0}, \mathbf{I}_d)$, usando el Lema 4.2.1 y que $\tau_i(G) = 0$, $\lambda_i = 1$ deducimos que

$$\begin{aligned}
\left.\frac{\partial H}{\partial \tau_i}\right|_{(\mathbf{I}_d,0)} &= -E_{F_0}\left(I_{\mathcal{S}}(\mathbf{y})y_i\right) = 0 \\
\left.\frac{\partial H}{\partial \lambda_i}\right|_{(\mathbf{I}_d,0)} &= \frac{1}{2}P_{F_0}(\mathcal{S}) - \frac{1}{2}E_{F_0}\left(I_{\mathcal{S}}(\mathbf{y})y_i^2\right) \\
&= \frac{1}{2}P_{F_0}(\mathcal{S}) - \frac{1}{2d}E_{F_0}\left(I_{\mathcal{S}}(\mathbf{y})\sum_{i=1}^{d} y_i^2\right) .
\end{aligned}$$

Por otra parte, como $P_{F_0}(\mathcal{S}) = P(W_d \leq K)$ y $E_{F_0}\left(I_{\mathcal{S}}(\mathbf{y})\sum_{i=1}^{d} y_i^2\right) = E\left(W_d I_{(0,K]}(W_d)\right)$, se obtiene el resultado, ya que $E\left(W_d I_{(0,K]}(W_d)\right) = d\,P(W_{d+2} \leq K)$. □

Corolario 4.2.1. *Sea $F = G = N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sean $\mathbf{T}(G) = \int \mathbf{x}\, dG$ y $\mathbf{V}(G) = \int (\mathbf{x} - \mathbf{T}(G))(\mathbf{x} - \mathbf{T}(G))'\, dG$, los funcionales clásicos de posición y de escala. Entonces, la función de influencia de $Pc(G, F)$ está dada por:*

$$IF(x, Pc, G) = \frac{1}{2}\, c_d\, \left((\mathbf{x} - \boldsymbol{\mu})'\, \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - d\right) ,$$

donde $c_d = P(W_d \leq K) - P(W_{d+2} \leq K)$ con W_d una variable aleatoria tal que $W_d \sim \chi^2_d$.

De la última expresión se observa que la función de influencia de la probabilidad de cobertura sólo depende de $\mathbf{x}$ a través de la norma de Mahalanobis, siendo una función creciente de la misma. En cuanto a la constante K, cuanto más grande es K más se acerca a 0 la constante c_d, y con ésta la función de influencia. Por otro lado, dejando fija la constante K, vemos que la función no esta acotada, mostrando la sensibilidad a observaciónes atípicas del procedimiento clásico.

Es interesante observar que contaminaciones en puntos con norma de Mahalanobis iguales a $\sqrt{d}$ no ejercen ninguna influencia en el funcional, en tanto que contaminaciones con normas menores a $\sqrt{d}$ poseen una influencia negativa y contaminaciones con normas mayores a $\sqrt{d}$ poseen una influencia positiva.

De lo anterior se deduce que el impacto de un dato atípico interno conduce a una disminución en la probabilidad de cobertura, siendo el caso más extremo $\mathbf{x} = \boldsymbol{\mu}$, donde la función toma el valor $-\frac{1}{2}\, d\, c_d$.

Por otro lado, el impacto de un dato atípico externo produce un aumento (no acotado) de la probabilidad de cobertura.

La Figura B.8.(a) muestra la función de influencia del procedimiento clásico cuando $d = 2$ para 5 factores de tolerancia distintos ($K = 2$, 4, 6, 8 y 10). Este gráfico permite visualizar las observaciónes descriptas.

En general, la función de influencia cuando $F = G = N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, puede obtenerse del Teorema 1 y del Lemma 1 de Croux y Haesbroek (2000) que da una expresión para la función de influencia de un funcional de escala robusto. Según ese resultado, si $G = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\Delta^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, es la distancia de Mahalanobis, entonces, para cualquier funcional de escala afín equivariante, $\mathbf{V}(G)$ para el que exista su función de influencia, existen dos funciones $\alpha_{\mathbf{V}}$ y $\gamma_{\mathbf{V}} : [0, \infty) \to I\!R$ tales que $\mathrm{IF}(\mathbf{x}, \mathbf{V}, G) = \alpha_{\mathbf{V}}(\Delta(\mathbf{x}))(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})' - \gamma_{\mathbf{V}}(\Delta(\mathbf{x}))\boldsymbol{\Sigma}$. De esta propiedad se deduce el siguiente resultado

Corolario 4.2.2. *Sea $F = G = N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sean $\mathbf{T}$ y $\mathbf{V}$ funcionales robustos afín equivariantes y Fisher–consistentes en G, es decir, $\mathbf{T}(G) = \boldsymbol{\mu}$ y $\mathbf{V}(G) = \boldsymbol{\Sigma}$. Entonces, la función de influencia de $Pc(G, F)$ está dada por:*

$$IF(x, Pc, G) \;=\; \frac{1}{2}\, c_d \,\left[\alpha_{\mathbf{V}}(\Delta(\mathbf{x}))\,\Delta^2(\mathbf{x}) - d\,\gamma_{\mathbf{V}}(\Delta(\mathbf{x}))\right] ,$$

donde $c_d = P(W_d \le K) - P(W_{d+2} \le K)$ con W_d una variable aleatoria tal que $W_d \sim \chi^2_d$.

Una revisión sobre estimadores robustos multivariados de posición y escala puede verse en Maronna y Yohai (1998). Entre los estimadores robustos de escala afín equivariantes podemos mencionar el $S-$estimador de escala (Lopuhaä (1990)). El funcional asociado a estos estimadores está definido como la solución $(\mathbf{T}(G), \mathbf{V}(G))$ de minimizar el determinante de $\mathbf{V}(G)$, $\det(\mathbf{V}(G)$ entre todos los $(\mathbf{t}, \mathbf{V})$ tales que

$$E\left(\rho(\left[(\mathbf{x} - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x} - \mathbf{t})\right]^{\frac{1}{2}}\right) \le b\ .$$

Estos estimadores surgen como una extensión de los estimadores de elipsoide de mínimo volúmen introducidos por Rousseeuw y van Zomeren (1990). Para obtener un estimador con alto punto de ruptura y con una tasa de convergencia $\sqrt{n}$ a la distribución normal, uno debe elegir una función de escores ρ acotada.

Para este estimador, tenemos las expresiones siguientes para las funciones $\alpha_{\mathbf{V}}$ y

$\gamma_{\mathbf{V}}$ que permiten calcular la función de influencia

$$\begin{aligned}\alpha_{\mathbf{V}}(t) &= \frac{d}{\gamma}\frac{\Psi(t)}{t} \\ \gamma_{\mathbf{V}}(t) &= \frac{1}{\gamma}\Psi(t)t - \frac{2}{\omega}\left(\rho(t) - b\right)\end{aligned}$$

donde $\Psi(t) = \rho'(t)$, $\gamma = \frac{1}{d+2}\left[E\left[\Psi'\left(\|\mathbf{x}\|\right)\|\mathbf{x}\|^2\right] + (d+1)\omega\right]$, $b = E\rho\left(\|\mathbf{x}\|\right)$ y $\omega = E\left[\Psi\left(\|\mathbf{x}\|\right)\|\mathbf{x}\|\right]$ con $\mathbf{x} \sim N\left(\mathbf{0}, \mathbf{I}_d\right)$. Por lo tanto, la probabilidad de cobertura tendrá influencia acotada usando S–estimadores de posición y escala si $\eta(t) = t\Psi(t)$ y $\rho(t)$ son funciones acotadas, lo cual es un requerimiento estándar para obtener estimadores de escala con punto de ruptura positivo.

En los capítulos anteriores hemos considerado como estimador de escala el estimador de Donoho–Stahel, que es un estimador de alto punto de ruptura. Para obtener la función de influencia cuando usamos este estimador robusto, utilizaremos los resultados de Gervini (2002).

Dados funcionales univariados m y s de posición y escala, definamos la medida de atipicidad $r^2\left(\mathbf{x}, G\right)$ como

$$r^2\left(\mathbf{x}, G\right) = \sup_{\mathbf{a}\in I\!R^d}\left|\frac{\mathbf{a}'\mathbf{x} - m\left(G^{\mathbf{a}}\right)}{s(G^{\mathbf{a}})}\right| ,$$

con $G^{\mathbf{a}}$ la distribución de $\mathbf{a}'\mathbf{x}$ cuando $\mathbf{x} \sim G$, o sea, es la distribución marginal univariada asociada a G en la dirección proyectada por el vector $\mathbf{a}$. El funcional asociado a los estimadores de posición y escala de Donoho–Stahel puede escribirse como

$$\begin{aligned}\mathbf{T}(G) &= \frac{E_G\left(w\left(r^2\left(\mathbf{x}, G\right)\right)\mathbf{x}\right)}{E_G\left(w\left(r^2\left(\mathbf{x}, G\right)\right)\right)} \\ \mathbf{V}(G) &= \beta\frac{E_G\left(w\left(r^2\left(\mathbf{x}, G\right)\right)\left(\mathbf{x} - \mathbf{T}(G)\right)\left(\mathbf{x} - \mathbf{T}(G)\right)'\right)}{E_G\left(w\left(r^2\left(\mathbf{x}, G\right)\right)\right)}\end{aligned}$$

donde β es una constante de consistencia calibrada de tal manera que cuando $G = N_d(0, I_d)$, $\mathbf{V}\left(G\right) = \mathbf{I}_d$.

Los estimadores de Donoho–Stahel resultan afín equivariantes si los funcionales univariados de posición m y escala s lo son. Para una distribución esférica G tenemos entonces $\mathbf{T}(G) = 0$ y

$$\mathbf{V}(G) = \beta\frac{E\left(w\left(\frac{R^2}{s_0}\right)R^2\right)}{d\,E\left(w\left(\frac{R^2}{s_0}\right)\right)}\mathbf{I}_d$$

donde $s_0 = s\left(G^{\mathbf{e}_1}\right)$, $R^2 = \|\mathbf{x}\|^2$ con $\mathbf{x} \sim G$. En particular, si $G = N_d(0, I_d)$ y s es un funcional de escala tal que $s(\Phi) = 1$ con Φ la distribución $N(0,1)$, obtenemos Fisher–consistencia eligiendo

$$\beta = \frac{d\, E\left(w\left(W_d\right)\right)}{E\left(w\left(W_d\right) W_d\right)} \,,$$

donde $W_d \sim \chi^2_d$.

En los capítulos anteriores hemos elegido como función w la función $w(t) = w_H\left(\sqrt{t}\right)$ con w_H la función de peso de Huber dada por:

$$w_H\left(r\right) = I_{[0,c]}\left(r\right) + I_{(c,\infty)}\left(r\right)\left(\frac{c}{r}\right)^2$$

donde $c = \sqrt{\chi^2_{0{,}95,d}}$. Por otra parte, la medida de atipicidad $r(\mathbf{x}, G)$ se tomaba eligiendo m la mediana y $s(G^{\mathbf{a}}) = \dfrac{1}{\Phi^{-1}(0{,}75)}\,\mathrm{MAD}(G^{\mathbf{a}})$, la mediana de los desvíos absolutos respecto de la mediana, escalada de modo que $s(\Phi) = 1$.

Corolario 4.2.3. *Sea $F = G = N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. Sean $\mathbf{T}$ y $\mathbf{V}$ los funcionales robustos afín equivariantes asociados a los estimadores de Donoho–Stahel, tales que $\mathbf{T}(G) = \boldsymbol{\mu}$ y $\mathbf{V}(G) = \boldsymbol{\Sigma}$. Supongamos que se cumplen las siguientes condiciones*

W1. *$w : [0,\infty) \to [0,\infty)$ es acotada y $w(u^2)u^2$ es acotada.*

W2. *w es derivable en casi todo punto y $\eta(u^2) = w'(u^2)u^4$ es acotada.*

Entonces, si tomamos como medida univariada de posición m a la mediana y como medida de escala a $s(\cdot) = \dfrac{1}{\Phi^{-1}(0{,}75)}\,\mathrm{MAD}(\cdot)$, la función de influencia de $Pc(G,F)$ está dada por:

$$\begin{aligned}
IF\left(\boldsymbol{\mu}, Pc, G\right) &= 0 \\
IF\left(\mathbf{x}, Pc, G\right) &= \beta\,\frac{1}{2}\,c_d\,\left[\frac{c_1}{c_0} g(\Delta(\mathbf{x})) + \frac{w\left(\Delta^2(\mathbf{x})\right)\Delta^2(\mathbf{x}) - c_2}{c_0}\right] \quad si \quad \mathbf{x} \neq \boldsymbol{\mu} \\
&= \frac{d}{2\,c_2}\,c_d\,\left[c_1\, g(\Delta(\mathbf{x})) + w\left(\Delta^2(\mathbf{x})\right)\Delta^2(\mathbf{x}) - c_2\right] \quad si \quad \mathbf{x} \neq \boldsymbol{\mu}\,,
\end{aligned}$$

con $\Delta^2(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$,

$$\begin{aligned} c_d &= P\left(W_d \leq K\right) - P\left(W_{d+2} \leq K\right) \\ c_0 &= E\left(w(W_d)\right) \\ c_1 &= -2E\left(w'(W_d)W_d^2\right) = -2E\left(\eta(W_d)\right) \\ c_2 &= E\left(w(W_d)W_d\right) \\ g(t) &= \frac{0{,}5 - F_{\mathcal{B}\left(\frac{1}{2},\frac{d-1}{2}\right)}\left(\dfrac{\left[\Phi^{-1}(0{,}75)\right]^2}{t^2}\right)}{2\Phi^{-1}(0{,}75)\,\varphi\left(\Phi^{-1}(0{,}75)\right)}\,, \end{aligned}$$

donde W_d *es una variable aleatoria tal que* $W_d \sim \chi^2_d$, φ *indica la densidad de una variable* $N(0,1)$ *y* $\mathcal{B}\left(\frac{1}{2},\frac{d-1}{2}\right)$ *es la distribución Beta de parámetros* $\frac{1}{2}$ *y* $\frac{d-1}{2}$.

DEMOSTRACIÓN. Sea $F_0 = N_d\left(\mathbf{0}, \mathbf{I}_d\right)$, luego si $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$, se verifica que $\mathrm{IF}(\mathbf{x}, \mathbf{V}, G) = \mathbf{C}\mathrm{IF}\left(\mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right), \mathbf{V}, F_0\right)\mathbf{C}'$. Por lo tanto,

$$\mathrm{tr}\left(\mathrm{IF}\left(\mathbf{x}, \mathbf{V}, G\right)\boldsymbol{\Sigma}^{-1}\right) = \mathrm{tr}\left(\mathrm{IF}\left(\mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right), \mathbf{V}, F_0\right)\right)\ .$$

El resultado se obtiene del Teorema 3 de Gervini (2002) y de las expresiones dadas en el apéndice de su trabajo para el caso de la mediana y la MAD ya que $\mathrm{IF}\left(\mathbf{0}, \mathbf{V}, F_0\right) = 0$ y si $\mathbf{z} \neq \mathbf{0}$

$$\mathrm{IF}\left(\mathbf{z}, \mathbf{V}, F_0\right) = \beta\left\{\alpha(\|\mathbf{z}\|)\left(\frac{\mathbf{z}\mathbf{z}'}{\|\mathbf{z}\|^2} - \frac{\mathbf{I}_d}{d}\right) + \left[\frac{c_1}{c_0}g(\|\mathbf{z}\|) + \frac{w\left(\|\mathbf{z}\|^2\right)\|\mathbf{z}\|^2 - c_2}{c_0}\right]\frac{\mathbf{I}_d}{d}\right\}$$

para una función α. Tomando traza se obtiene el resultado ya que el primer término se anula y $\beta = d\,\dfrac{c_0}{c_2}$. □

La condición **W1** asegura que la función influencia de la cobertura es acotada. En particular, cuando $w(t) = w_H\left(\sqrt{t}\right)$, tenemos $w'(t) = \dfrac{w'_H\left(\sqrt{t}\right)}{2\sqrt{t}}$. Por lo tanto,

$$\begin{aligned} c_0 &= P\left(W_d < c^2\right) + c^2\ E\left(\frac{1}{W_d}I_{(c^2,\infty)}(W_d)\right) \\ &= P\left(W_d < c^2\right) + c^2\ \frac{1}{2\,(d-2)}\ \left(1 - P\left(W_{d-1} < c^2\right)\right) \quad \text{si} \quad d \neq 2 \\ c_1 &= 2\,c^2\ E\left(\,I_{(c^2,\infty)}(W_d)\right) = 2\,c^2\ \left(1 - P\left(W_d < c^2\right)\right) \\ c_2 &= E\left(W_d I_{(0,c^2)}(W_d)\right) + c^2\ E\left(I_{(c^2,\infty)}(W_d)\right) \\ &= d\ P\left(W_{d+2} < c^2\right) + c^2\ \left(1 - P\left(W_d < c^2\right)\right) \end{aligned}$$

con $c = \sqrt{\chi^2_{0,95,d}}$.

La Figura B.8.(b) muestra la función de influencia de la probabilidad de cobertura cuando utilizamos los estimadores de Donoho–Stahel para $d = 2$ y 5 factores de tolerancia distintos ($K = 2$, 4, 6, 8 y 10). Como puede observarse la función de influencia es acotada, con una discontinuidad en 0 debido a la discontinuidad de la influencia de los funcionales univariados de posición y escala. En efecto, la función de influencia en $\mathbf{x} = \boldsymbol{\mu}$ es 0 pero

$$\lim_{\mathbf{x}\to\boldsymbol{\mu}} \mathrm{IF}\left(\mathbf{x}, Pc, G\right) = -\frac{d}{2\,c_2}\,c_d\,\left[c_1\,\frac{1}{4\Phi^{-1}(0{,}75)\,\varphi\left(\Phi^{-1}(0{,}75)\right)} + c_2\right]$$

mostrando la discontinuidad en $\mathbf{0}$.

Teniendo en cuenta que la escala en la Figura B.8 no permite apreciar el efecto de inliers en el estimador clásico, en la Figura B.9 graficamos la función de influencia de la probabilidad de cobertura en el rango $0 \leq \Delta(\mathbf{x}) \leq 2$, que permite apreciar dicho efecto. Como muestra el gráfico, ambos estimadores son sensibles a este tipo de datos, aunque el efecto está acotado y es mayor para los estimadores robustos, como se observó en la Sección 3.7 del Capítulo 3.

Vale la pena destacar que una medida diagnóstico asociada a la probabilidad de cobertura, estaría definida a través de

$$D(\mathbf{x}) \;=\; \frac{1}{2}\,c_d\,\left(\left(\mathbf{x}-\widehat{\boldsymbol{\mu}}\right)'\,\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{x}-\widehat{\boldsymbol{\mu}}\right) - d\right)\,,$$

donde $\widehat{\boldsymbol{\mu}}$ y $\widehat{\boldsymbol{\Sigma}}$ son estimadores robustos de posición y escala multivariados, por ejemplo, los estimadores de Donoho–Stahel. Esta medida coincide salvo un cambio de coordenadas con la medida usual utilizada para detectar datos multivariados atípicos que se basa en una versión robusta de la distancia de Mahalanobis introducida por Rousseeuw y van Zomeren (1990).

Capítulo 5

Métodos de remuestreo para regiones de tolerancia multivariadas

Cuando existen sospechas que la distribución de origen no es normal, una posibilidad para estimar la verdadera proporción de cobertura de la región hallada, ya sea clásica o robusta, consiste en calcular mediante remuestreo (Bootstrap) la proporción de cobertura verdadera. este método fue propuesto para el caso univariado por Fernholz y Gillespie (2001).

El objetivo de este Capítulo es describir como se pueden adaptar esas ideas al caso multivariado. No ahondaremos en la teoría que lleva a mostrar que propiedades de dicho procedimiento que será objeto de trabajo futuro.

5.1. Región de tolerancia multivariada con $q-$cobertura corregida

Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra aleatoria de una distribución F y sea F_n la función de distribución empírica. Definamos la siguiente medida de discrepancia entre la cobertura real (desconocida) y la cobertura muestral

$$D_n = \sqrt{n}\left(P_{F_n}\left(\left(\mathbf{x}-\mathbf{t}_n\right)\mathbf{V}_n^{-1}\left(\mathbf{x}-\mathbf{t}_n\right)\leq K\right) - P_F\left(\left(\mathbf{x}-\mathbf{t}_n\right)\mathbf{V}_n^{-1}\left(\mathbf{x}-\mathbf{t}_n\right)\leq K\right)\right) ,$$

donde $P_{F_n}\left(\left(\mathbf{x}-\mathbf{t}_n\right)\mathbf{V}_n^{-1}\left(\mathbf{x}-\mathbf{t}_n\right)\leq K\right)$ es la proporción de la muestra de referencia que queda contenida en la región de tolerancia (cobertura muestral). Como no conocemos la distribución original, remuestreamos la muestra de referencia para obtener

un estimador de D_n.

$$D_n^* = \sqrt{n}\left(P_{F_n^*}\left(\left(\mathbf{x}-\mathbf{t}^*\right)\mathbf{V}^{*-1}\left(\mathbf{x}-\mathbf{t}^*\right)\leq K\right) - P_{F_n}\left(\left(\mathbf{x}-\mathbf{t}^*\right)\mathbf{V}^{*-1}\left(\mathbf{x}-\mathbf{t}^{*-1}\right)\leq K\right)\right)$$

donde F_n^*, $\mathbf{t}^*$ y $\mathbf{V}^*$ corresponden respectivamente, a la función de distribución, el estimador de posición y el estimador de escala sobre la muestra boostrap tomada de la muestra original.

El método consiste en, dada la muestra de referencia, la proporción de cobertura teórica q y el nivel de confianza δ requeridos, calcular los estimadores de posición y escala (clásicos o robustos) elegir el factor K correspondiente, generar B submuestras de la muestra de referencia y para cada una calcular D_n^*. Elegir el cuantil δ de los D_n^*, al que llamaremos d_δ^*. Así la cobertura corregida por remuestreo estará dada por

$$q_n^* = P_{F_n}\left(\left(\mathbf{x}-\mathbf{t}_n\right)\mathbf{V}_n^{-1}\left(\mathbf{x}-\mathbf{t}_n\right)\leq K\right) - \frac{d_\delta^*}{\sqrt{n}}$$

Un procedimiento similar puede aplicarse para corregir el factor de tolerancia K en lugar de corregir la cobertura q.

Capítulo 6

Conclusiones

En esta tesis, se propuso un procedimiento para obtener regiones de tolerancia menos sensibles a un porcentaje de observaciones atípicas. La propuesta es una propuesta de tipo *plug–in* y consiste en reemplazar los estimadores clásicos por estimadores robustos y equivariantes. Para poder comparar los factores de tolerancia clásicos y robustos los estimadores deben ser además Fisher–consistentes para la distribución normal.

Según lo observado en las simulaciones, la región robusta propuesta posee, en oposición a la región clásica, un comportamiento más estable, en términos de cobertura y centrado, ante la presencia de datos atípicos externos. Frente a datos atípicos internos la ganancia en estabilidad es mucho menor. Sin embargo, son los datos atípicos externos los que potencialmente tienen una influencia no acotada, en tanto que los internos están naturalmente acotados y son los menos graves. Esta diferencia en estabilidad de la probabilidad de cobertura fue justificada teoricamente en el Capítulo 4, donde se calculó la función de influencia del funcional probabilidad de cobertura. En el caso clásico esta función es no acotada mientras que al utilizar funcionales robustos de posición y escala con influencia acotada, la función de influencia del funcional probabilidad de cobertura resulta acotada. Por otra parte, la función de influencia nos sugirió una medida diagnóstico que coincide, salvo un cambio de coordenadas, con la distancia de Mahalanobis.

La mejora en estabilidad y centrado se logra con un ligero incremento en el volumen, y por ende en la cobertura, por lo que en el caso estrictamnente normal, la región robusta puede pensarse como conservativa en términos de la probabilidad de cobertura.

En cuanto al cálculo de los factores que determinan la región, para ninguna de las dos regiones existen expresiones explícitas. Si se pretende un cálculo preciso de los factores debe recurrirse a Monte Carlo. Los tiempos computacionales involucrados son superiores para la región robusta, no obstante, la práctica indica que el cómputo

de un factor robusto con un grado de precisión razonable ($N = 1000, R = 1000$ y $M = 1000$) en una computadora personal Pentium 4, demora aproximadamente 20 minutos. Incluso en aplicaciones como la detección de cambios en control de calidad, en las cuales sucesivas regiones de tolerancia deben ser calculadas en simultáneo con algún proceso industrial, el muestreo puede realizarse de modo tal de mantener fijos los parámetros n, d, q y δ, necesitándose asísólo un factor de tolerancia.

Apéndice A

Apéndice 1. Tablas

K	δ	q	π
5.664	0.75	0.75	0.7233
8.158	0.75	0.90	0.8780
9.930	0.75	0.95	0.9336
13.84	0.75	0.99	0.9832
5.913	0.90	0.75	0.7208
8.507	0.90	0.90	0.8742
10.35	0.90	0.95	0.9320
14.39	0.90	0.99	0.9823
6.067	0.95	0.75	0.7196
8.721	0.95	0.90	0.8753
10.60	0.95	0.95	0.9310
14.72	0.95	0.99	0.9813
6.368	0.99	0.75	0.7160
9.130	0.99	0.90	0.8717
11.07	0.99	0.95	0.9277
15.33	0.99	0.99	0.9794

Tabla A.1: Cobertura real (π) y teórica (q) para $n = 100$, $d = 4$ y distintos valores de δ. El factor de tolerancia K es la aproximación de Guttman (1970).

Comb.	n	d	K	Comb.	n	d	K	Comb.	n	d	K
1	25	2	6	10	25	2	9	19	25	2	12
2	50	2	6	11	50	2	9	20	50	2	12
3	100	2	6	12	100	2	9	21	100	2	12
4	25	4	6	13	25	4	9	22	25	4	12
5	50	4	6	14	50	4	9	23	50	4	12
6	100	4	6	15	100	4	9	24	100	4	12
7	25	8	6	16	25	8	9	25	25	8	12
8	50	8	6	17	50	8	9	26	50	8	12
9	100	8	6	18	100	8	9	27	100	8	12

Tabla A.2: Valores de n, d y K para los que se efectuó la comparación de la cobertura media con una variable con distribución Beta.

δ	q	K_0	π_0	K_1	π_1	K_2	π_2
0.75	0.75	5.664	0.7233	5.965	0.7483	6.083	0.7570
0.75	0.90	8.158	0.8780	8.715	0.8996	8.787	0.9006
0.75	0.95	9.930	0.9336	10.710	0.9492	10.717	0.9492
0.75	0.99	13.840	0.9832	15.360	0.9903	14.996	0.9888
0.90	0.75	5.913	0.7208	6.241	0.7474	6.466	0.7622
0.90	0.90	8.507	0.8742	9.111	0.8973	9.340	0.9048
0.90	0.95	10.350	0.9320	11.190	0.9484	11.391	0.9518
0.90	0.99	14.390	0.9823	16.010	0.9897	15.939	0.9894
0.95	0.75	6.067	0.7196	6.414	0.7469	6.710	0.7686
0.95	0.90	8.721	0.8753	9.343	0.8973	9.693	0.9068
0.95	0.95	10.600	0.9310	11.490	0.9476	11.822	0.9530
0.95	0.99	14.720	0.9813	16.470	0.9896	16.543	0.9895
0.99	0.75	6.368	0.7160	6.741	0.7442	7.203	0.7723
0.99	0.90	9.130	0.8717	9.803	0.8952	10.405	0.9098
0.99	0.95	11.070	0.9277	12.030	0.9459	12.690	0.9565
0.99	0.99	15.330	0.9794	17.210	0.9889	17.757	0.9906

Tabla A.3: Cobertura π_0, π_1 y π_2 obtenidas, para $n = 100$ y $d = 4$, usando el factor de tolerancia de Guttman (K_0), el procedimiento de un paso (K_1) y el factor de tolerancia de Krishnamoorthy y Mathew (K_2), respectivamente. q indica la cobertura teórica y δ el nivel de confianza.

d	K
2	9.8752
3	13.1367
4	16.4330
5	20.0726

Tabla A.4: Factores de tolerancia clásicos K calculados por simulación cuando $q = \delta = 0.95$.

Δ	$d = 2$		$d = 3$		$d = 4$		$d = 5$	
	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$ (%)
2	0.9570	1.0333	0.9570	1.0219	0.9557	1.0164	0.9534	1.0129
4	0.9697	1.1143	0.9641	1.0748	0.9614	1.0554	0.9606	1.0455
8	0.9783	1.3374	0.9715	1.2131	0.9685	1.1570	0.9658	1.1222
16	0.9806	1.7687	0.9752	1.4657	0.9708	1.3300	0.9676	1.2564

Tabla A.5: Cobertura real (π) e Incremento de volúmen ($\mathcal{I}$) por inclusión de un dato atípico con norma Δ, para $q = \delta = 0.95$ y $n = 30$.

Δ	$d = 2$		$d = 3$		$d = 4$		$d = 5$	
de inliers	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$ (%)
1	0.9439	0.9832	0.9427	0.9813	0.9411	0.9815	0.9407	0.9810
2	0.9359	0.9608	0.9333	0.9627	0.9303	0.9606	0.9295	0.9610
3	0.9293	0.9465	0.9276	0.9448	0.9204	0.9431	0.9180	0.9431
4	0.9206	0.9256	0.9137	0.9247	0.9082	0.9238	0.9064	0.9227

Tabla A.6: Cobertura real (π) e Incremento de volúmen ($\mathcal{I}$) por inclusión de inliers, para $q = \delta = 0.95$ y $n = 30$.

d	β
2	1.0413708
3	1.0070053
4	1.0000000
5	0.9957046
6	0.9927844
7	0.9906613
8	0.9890429
9	0.9877651
10	0.9867286
15	0.9835156

Tabla A.7: Valores de la constante de consistencia β para distintas dimensiones d.

Estimadores por Monte Carlo de los Factores de Tolerancia Robustos para $d = 2$									
	$q = 0.90$			$q = 0.95$			q =0.99		
n	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$
20	10.8320	12.3821	17.5341	14.5522	16.7106	24.5253	23.1711	27.8769	40.9653
25	9.3069	10.2947	13.2224	12.3985	13.9580	18.1720	19.9671	22.3422	29.4218
30	8.2326	9.2963	11.3772	10.9862	12.2417	15.4112	17.4303	19.8460	25.5177
35	7.9080	8.6992	10.1594	10.4656	11.4732	14.0209	16.5040	18.2180	22.6908
40	7.2711	7.8236	9.3063	9.7533	10.4301	12.1758	15.3556	16.8527	19.6146
45	7.1068	7.5837	8.6729	9.3606	10.0652	11.5020	14.7729	16.3259	19.1649
50	6.7816	7.2303	8.1386	8.8651	9.4666	11.0186	14.1042	15.2943	17.4805
55	6.5337	6.9445	7.7602	8.5696	9.1817	10.3032	13.6083	14.6165	17.0323
60	6.4745	7.0355	7.8174	8.6042	9.1507	10.3930	13.4796	14.5166	16.6398
65	6.3193	6.6702	7.5251	8.2748	8.7425	9.8564	12.8604	13.8477	15.9314
70	6.1470	6.6540	7.3719	8.1761	8.6563	9.8496	12.7697	13.5505	15.6173
75	6.1106	6.4738	7.2753	8.0202	8.5651	9.5449	12.7843	13.4542	15.2029
80	6.0064	6.2933	6.9345	7.8461	8.2518	9.1011	12.3369	13.0866	14.7179
85	5.9857	6.2590	6.8931	7.8599	8.1117	9.1025	12.2277	12.9145	14.3738
90	5.9770	6.3030	7.0116	7.8260	8.2017	9.2849	12.2239	12.8636	14.7869
95	5.9036	6.1823	6.7342	7.7652	8.1079	8.8956	12.1545	12.7148	13.8342
100	5.8052	6.0258	6.6421	7.6159	7.9355	8.6379	11.9852	12.4499	13.5771

Tabla A.8: Factores de tolerancia robustos K_{DS} en dimensión $d = 2$, para niveles de cobertura $q = 0.90$, 0.95 y 0.99 y niveles de confianza $\delta = 0.90$, 0.95 y 0.99.

Estimadores por Monte Carlo de los Factores de Tolerancia Robustos para $d = 3$									
	$q = 0.90$			$q = 0.95$			q =0.99		
n	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$
20	18.4888	21.1397	27.4670	24.2531	27.9458	36.5284	38.3513	44.9395	58.5057
25	14.5141	16.2015	21.2053	18.8563	21.4465	28.2591	29.2044	34.0270	45.4828
30	12.4591	13.6268	17.5833	16.0860	17.4503	23.0519	24.5519	27.5803	36.2613
35	11.3296	12.1778	15.2488	14.5182	15.6294	19.9587	22.2818	24.6301	30.5704
40	10.6179	11.4505	14.0012	13.5417	14.6996	17.9198	20.4660	22.4652	29.1666
45	10.1765	10.7696	12.3920	12.9928	13.7569	16.0137	19.5297	21.1643	24.4763
50	9.7695	10.2317	11.2996	12.3856	13.0009	14.5262	18.7228	19.8433	22.9663
55	9.3373	9.9091	11.1179	11.8649	12.5228	14.4079	17.8622	18.9811	22.5006
60	9.1459	9.8136	10.7406	11.6818	12.4917	13.7287	17.5797	19.0078	21.9252
65	8.8516	9.2392	10.7574	11.1660	11.7582	13.6343	16.7690	17.6115	20.9971
70	8.6261	9.1126	9.9451	10.9276	11.5222	12.5902	16.3966	17.4556	18.8966
75	8.6470	9.0168	9.6662	10.9072	11.5021	12.3449	16.3814	17.0689	18.5640
80	8.3896	8.8135	9.3718	10.7288	11.1830	11.9359	15.8844	16.5346	18.8273
85	8.3207	8.7319	9.4782	10.5832	11.0636	12.0174	15.8495	16.5357	17.7586
90	8.1590	8.5613	9.3359	10.3339	10.8568	11.9504	15.4399	16.0606	17.8178
95	8.1348	8.5243	9.1853	10.3238	10.6891	11.6131	15.3301	16.0695	18.0892
100	7.9569	8.2413	8.9158	10.0528	10.4174	11.2309	15.1451	15.7830	17.7933

Tabla A.9: Factores de tolerancia robustos K_{DS} en dimensión $d = 3$, para niveles de cobertura $q = 0.90$, 0.95 y 0.99 y niveles de confianza $\delta = 0.90$, 0.95 y 0.99.

Estimadores por Monte Carlo de los Factores de Tolerancia Robustos para $d = 4$									
	$q = 0.90$			$q = 0.95$			q =0.99		
n	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$
20	27.4182	31.5431	44.5483	35.4487	41.4550	57.3808	55.7374	66.4408	92.8933
25	21.4547	24.6844	31.8190	27.4777	31.5780	42.3042	42.9968	49.9301	65.0356
30	16.9623	18.3120	21.6688	21.5732	23.2288	28.3680	32.7613	35.3599	44.2545
35	15.1949	16.6681	20.3717	19.0612	21.2592	26.2953	28.8685	32.1940	41.4370
40	13.5806	14.2707	16.4577	16.9452	18.0497	21.1071	24.9294	27.1523	33.2087
45	12.6335	13.3466	15.3875	15.8920	16.6360	19.7838	23.3600	25.0936	29.0160
50	12.3287	13.0614	14.1136	15.4366	16.2419	17.9883	22.6729	24.0837	27.8435
55	11.6009	12.1972	13.8174	14.5084	15.2489	17.5803	21.2475	22.5057	25.8496
60	11.4914	11.9824	13.1846	14.1753	14.9044	16.5707	20.7960	22.0699	24.7157
65	11.1283	11.6137	12.5504	13.7444	14.3994	15.5499	19.8585	21.0003	22.9426
70	10.7022	11.1047	12.1436	13.1528	13.7946	15.1036	19.0215	19.8970	22.1491
75	10.6953	11.0952	11.8944	13.1943	13.7044	14.7181	19.0660	20.0513	21.7096
80	10.3104	10.7255	11.3242	12.7346	13.3265	14.0519	18.5037	19.1277	20.5951
85	10.3025	10.6632	11.5255	12.6941	13.1263	14.0832	18.3106	19.0707	20.5999
90	10.2673	10.6312	11.4723	12.6865	13.1973	14.2593	18.2493	19.1498	20.6151
95	10.1159	10.4224	11.0732	12.4098	12.8850	13.6946	17.9134	18.7162	20.1619
100	9.9515	10.2544	10.7488	12.2488	12.6392	13.4576	17.7022	18.4893	19.9511

Tabla A.10: Factores de tolerancia robustos K_{DS} en dimensión $d = 4$, para niveles de cobertura $q = 0.90$, 0.95 y 0.99 y niveles de confianza $\delta = 0.90$, 0.95 y 0.99.

		Clásica		Robusta		
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\nu_{\mathrm{C}\cdot}}{\nu_{\mathrm{DS}}}}$
2	20	12.1744	0.9525	16.7106	0.9540	1.0754
2	30	9.8752	0.9460	12.2417	0.9500	1.0455
2	50	8.3989	0.9440	9.4666	0.9470	1.0206
2	100	7.4187	0.9470	7.9355	0.9500	1.0137
3	20	16.6939	0.9480	27.9458	0.9520	1.1030
3	30	13.2222	0.9485	17.4503	0.9470	1.0440
3	50	10.4174	0.9485	13.0009	0.9530	1.0240
3	100	9.6736	0.9490	10.1994	0.9475	1.0023
4	20	21.3464	0.9480	41.4550	0.9475	1.1409
4	30	16.9176	0.9510	23.2288	0.9525	1.0453
4	50	13.4051	0.9480	16.2419	0.9535	1.0341
4	100	11.6219	0.9490	12.6392	0.9490	1.0075
5	20	27.2366	0.9445	64.0125	0.9400	1.2097
5	30	20.0054	0.9490	29.1497	0.9440	1.0571
5	50	15.9508	0.9510	19.1330	0.9525	1.0224
5	100	13.6632	0.9480	14.8441	0.9505	1.0085
8	20	56.1119	0.9520	240.0697	0.9375	1.5496
8	30	32.9768	0.9455	57.5583	0.9490	1.1159
8	50	23.9586	0.9475	28.8758	0.9500	1.0228
8	100	20.8338	0.9480	20.5229	0.9490	1.0062

Tabla A.11: Factores de tolerancia clásicos (K_{C}) y robustos (K_{DS}) y cociente de los volumenes de las regiones robustas y clásicas. Datos Normales, para $q = \delta = 0.95$.

Estimadores de los Errores para $d = 2$									
	$q = 0.90$			$q = 0.95$			q =0.99		
n	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$	$\delta = 0.90$	$\delta = 0.95$	$\delta = 0.99$
20	1.7520	2.1138	5.9243	2.5690	3.4850	7.5434	7.8852	10.4353	22.2891
25	1.2336	1.4314	4.5678	1.9353	2.4213	5.4864	5.5349	8.3741	14.5812
30	1.1023	1.4010	2.5037	1.5916	2.3008	3.8286	5.0497	6.7551	10.1117
40	0.9044	1.0428	1.9548	1.3633	1.6685	2.8780	4.2577	4.9906	7.5734
50	0.7445	0.9585	1.7806	1.2879	1.6175	2.2103	4.1237	4.4089	6.0743
60	0.7628	0.9675	1.2924	1.1743	1.4055	1.7157	3.6254	4.4446	6.1464
70	0.7888	0.7616	0.9661	1.0590	1.2537	1.4157	3.4059	4.0601	5.5313
80	0.6029	0.7115	1.2388	1.0588	1.1582	1.6414	3.1249	3.4364	5.3757
90	0.5968	0.8164	1.1236	1.0209	1.1652	1.6388	3.2054	4.0591	3.9688
100	0.6250	0.6748	0.7866	0.9416	1.0955	1.2725	2.9833	3.4005	4.3486
Estimadores de los errores para $d = 3$									
20	2.7003	3.7088	6.5933	3.8074	5.6231	9.5200	10.2223	13.5992	26.2589
25	1.8556	2.6075	5.3146	2.5367	4.0700	7.8785	8.1773	10.6413	22.6478
30	1.4077	1.7900	3.3808	2.1091	3.0130	5.0629	6.8856	8.8798	11.9726
40	1.1344	1.4239	3.2362	1.5913	2.2518	4.9221	5.3517	7.4948	13.4127
50	0.9397	1.0141	1.8828	1.3692	1.5701	2.6440	4.6968	5.6341	6.0445
60	0.9442	0.9506	1.7425	1.4164	1.5834	2.2311	4.4738	4.7529	7.6495
70	0.7507	0.8438	1.1375	1.2070	1.4458	1.3620	3.8601	4.1139	5.4864
80	0.7595	0.8275	1.0154	1.0628	1.2956	2.1344	3.6641	4.1204	4.7080
90	0.7522	0.8284	1.1714	1.0543	1.1653	1.7349	3.5210	3.9224	5.0378
100	0.6773	0.7489	1.0145	1.0905	1.1628	1.5454	3.4251	3.5681	4.6563
Estimadores de los errores para $d = 4$									
20	3.8366	4.9273	8.7218	5.4870	7.6228	17.4972	17.5523	22.9152	42.7468
25	2.5764	3.4971	7.6949	4.3068	5.4465	14.0137	11.7727	15.3437	37.4955
30	1.6771	2.0198	4.5521	2.6732	3.1916	7.1120	8.1311	10.6082	15.7616
40	1.2045	1.5697	2.5827	2.0043	2.4182	4.2547	5.8899	6.5636	9.2005
50	1.1664	1.2286	1.7791	1.6176	1.8961	3.0372	5.2196	6.3935	8.9328
60	0.9491	1.1108	2.0230	1.5600	1.7077	2.8523	4.5808	4.6759	7.8257
70	0.8699	1.0323	1.2608	1.2595	1.4100	1.8614	4.2445	4.4811	6.1601
80	0.8634	0.9377	1.0714	1.2441	1.3905	1.6752	3.7855	4.5481	5.4673
90	0.8487	0.9020	1.3132	1.2085	1.4357	2.2819	3.7971	4.1658	5.5464
100	0.7883	0.8804	1.1118	1.1875	1.3041	1.6148	3.9164	4.2943	5.4480

Tabla A.12: Estimadores de los errores de los factores de tolerancia robustos K_{DS}, para niveles de cobertura $q = 0.90$, 0.95 y 0.99, niveles de confianza $\delta = 0.90$, 0.95 y 0.99 y dimensiones $d = 2, 3, 4$.

Δ	$d=2$		$d=3$		$d=4$		$d=5$	
	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$ (%)
2	0.9583	1.0309	0.9516	1.0183	0.9499	1.0119	0.9575	1.0077
4	0.9607	1.0467	0.9575	1.0321	0.9542	1.0247	0.9583	1.0226
8	0.9633	1.0508	0.9592	1.0383	0.9566	1.0311	0.9616	1.0263
16	0.9608	1.0535	0.9616	1.0362	0.9600	1.0295	0.9608	1.0230

Tabla A.13: Cobertura real (π) e Incremento de volúmen ($\mathcal{I}$) de las regiones robustas por inclusión de un dato atípico con norma Δ, para $q = \delta = 0.95$ y $n = 30$.

Cantidad de inliers	$d=2$		$d=3$		$d=4$		$d=5$	
	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$	π	$\mathcal{I}$ (%)	π	$\mathcal{I}$ (%)
1	0.9455	0.9727	0.9355	0.9669	0.9310	0.9648	0.9250	0.9628
2	0.9285	0.9434	0.9120	0.9314	0.9090	0.9228	0.8930	0.9185
3	0.9050	0.9106	0.8865	0.8878	0.8625	0.8701	0.8465	0.8540
4	0.8800	0.8701	0.8440	0.8422	0.8025	0.8176	0.7600	0.7872

Tabla A.14: Cobertura real (π) e Incremento de volúmen ($\mathcal{I}$) de las regiones robustas por agregado de inliers, para $q = \delta = 0.95$ y $n = 30$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\mathcal{V}_{\mathrm{C}\cdot}}{\mathcal{V}_{\mathrm{DS}}}}$	$\frac{\|\overline{\mathbf{x}}\|}{\|\mathbf{t}_n\|}$
2	20	12.1744	1.0000	16.7106	0.9960	2.7778	4.3498
2	30	9.7920	1.0000	12.2417	0.9980	3.1331	4.9244
2	50	8.3989	1.0000	9.4666	0.9980	4.0020	6.1716
2	100	7.4187	1.0000	7.9355	0.9990	5.9750	9.4369
3	20	16.6939	1.0000	27.9458	0.9940	2.4637	4.7695
3	30	13.2222	1.0000	17.4503	0.9970	3.0593	5.7492
3	50	10.9711	1.0000	13.0009	0.9980	3.9036	6.6528
3	100	9.6736	1.0000	10.4174	0.9990	5.3589	8.8928
4	20	21.3464	0.9990	41.4550	0.9930	2.3600	5.1893
4	30	16.9176	1.0000	23.2288	0.9940	2.9915	6.0277
4	50	13.4051	1.0000	16.2419	0.9980	3.6996	7.4199
4	100	11.6219	1.0000	12.6392	0.9990	4.9280	9.1556
5	20	27.2366	0.9985	64.0125	0.9860	2.1726	5.0921
5	30	20.0054	1.0000	29.1497	0.9935	2.7466	5.8958
5	50	15.9508	1.0000	19.1330	0.9980	3.4553	6.9667
5	100	13.6632	1.0000	14.8441	1.0000	4.8052	9.7124
8	20	56.1119	0.9960	240.0697	0.9425	1.7047	4.7189
8	30	32.9768	1.0000	57.5583	0.9865	2.4991	6.9566
8	50	23.9586	1.0000	28.8758	0.9970	3.0707	7.6931
8	100	19.5602	1.0000	20.8338	1.0000	4.0777	10.6633

Tabla A.15: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución Cauchy $d-$ variada. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\nu_{\mathrm{C}\cdot}}{\nu_{\mathrm{DS}}}}$	$\frac{\|\overline{\mathbf{x}}\|}{\|\mathbf{t}_n\|}$
2	20	12.1744	0.9950	16.7106	0.9805	1.2756	1.5943
2	30	9.7920	0.9960	12.2417	0.9850	1.3770	1.6817
2	50	8.3989	0.9980	9.4666	0.9880	1.5240	1.7894
2	100	7.4187	1.0000	7.9355	0.9930	1.6028	1.8072
3	20	16.6939	0.9920	27.9458	0.9820	1.2870	1.6639
3	30	13.2222	0.9960	17.4503	0.9840	1.4265	1.7984
3	50	10.9711	0.9990	13.0009	0.9905	1.5235	1.8929
3	100	9.6736	1.0000	10.4174	0.9920	1.6523	1.9467
4	20	21.3464	0.9920	41.4550	0.9715	1.2753	1.7438
4	30	16.9176	0.9970	23.2288	0.9825	1.4347	1.7523
4	50	13.4051	0.9990	16.2419	0.9890	1.5223	1.8904
4	100	11.6219	1.0000	12.6392	0.9930	1.6210	1.8976
5	20	27.2366	0.9860	64.0125	0.9750	1.2133	1.6914
5	30	20.0054	0.9940	29.1497	0.9785	1.3951	1.8109
5	50	15.9508	0.9980	19.1330	0.9860	1.5054	1.8997
5	100	13.6632	1.0000	14.8441	0.9940	1.6286	1.9849
8	20	56.1119	0.9850	240.0697	0.9455	0.9877	1.5191
8	30	32.9768	0.9920	57.5583	0.9720	1.3502	1.8526
8	50	23.9586	0.9980	28.8758	0.9860	1.4794	1.9352
8	100	19.5602	1.0000	20.8338	0.9950	1.5556	1.9178

Tabla A.16: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución $\mathcal{T}_2(d)$. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\nu_{\mathrm{C}\cdot}}{\nu_{\mathrm{DS}}}}$	$\frac{\|\overline{\mathbf{x}}\|}{\|\mathbf{t}_n\|}$
2	20	12.1744	0.9880	16.7106	0.9760	1.1260	1.2941
2	30	9.7920	0.9880	12.2417	0.9765	1.1477	1.2558
2	50	8.3989	0.9935	9.4666	0.9810	1.2231	1.3065
2	100	7.4187	0.9960	7.9355	0.9835	1.2583	1.3278
3	20	16.6939	0.9875	27.9458	0.9780	1.0994	1.2331
3	30	13.2222	0.9890	17.4503	0.9770	1.1938	1.3162
3	50	10.9711	0.9940	13.0009	0.9815	1.2263	1.3501
3	100	9.6736	0.9970	10.4174	0.9850	1.2856	1.3747
4	20	21.3464	0.9860	41.4550	0.9690	1.0709	1.2550
4	30	16.9176	0.9910	23.2288	0.9770	1.1946	1.3047
4	50	13.4051	0.9930	16.2419	0.9805	1.2134	1.3738
4	100	11.6219	0.9970	12.6392	0.9860	1.2690	1.3459
5	20	27.2366	0.9795	64.0125	0.9575	1.0263	1.2297
5	30	20.0054	0.9875	29.1497	0.9700	1.1920	1.3743
5	50	15.9508	0.9950	19.1330	0.9830	1.2418	1.3721
5	100	13.6632	0.9980	14.8441	0.9880	1.2659	1.4181
8	20	56.1119	0.9765	240.0697	0.9525	0.8367	1.1343
8	30	32.9768	0.9840	57.5583	0.9705	1.1424	1.3359
8	50	23.9586	0.9930	28.8758	0.9780	1.2349	1.4239
8	100	19.5602	0.9980	20.8338	0.9890	1.2502	1.3946

Tabla A.17: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución $\mathcal{T}_3(d)$. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\nu_{\mathrm{C}\cdot}}{\nu_{\mathrm{DS}}}}$	$\frac{\|\overline{\mathbf{x}}\|}{\|\mathbf{t}_n\|}$
2	20	12.1744	0.9610	16.7106	0.9600	0.9599	1.0483
2	30	9.7920	0.9570	12.2417	0.9580	0.9986	1.1352
2	50	8.3989	0.9530	9.4666	0.9530	1.0328	1.1408
2	100	7.4187	0.9610	7.9355	0.9530	1.1008	1.3178
3	20	16.6939	0.9545	27.9458	0.9495	0.9235	1.0183
3	30	13.2222	0.9590	17.4503	0.9545	1.0018	1.1108
3	50	10.9711	0.9545	13.0009	0.9530	1.0324	1.1956
3	100	9.6736	0.9620	10.4174	0.9540	1.0973	1.3717
4	20	21.3464	0.9520	41.4550	0.9480	0.8991	0.9963
4	30	16.9176	0.9580	23.2288	0.9450	0.9923	1.0702
4	50	13.4051	0.9585	16.2419	0.9570	1.0219	1.1730
4	100	11.6219	0.9600	12.6392	0.9545	1.1054	1.2708
5	20	27.2366	0.9530	64.0125	0.9500	0.8489	0.8987
5	30	20.0054	0.9540	29.1497	0.9500	0.9834	1.0300
5	50	15.9508	0.9580	19.1330	0.9525	1.0322	1.1578
5	100	13.6632	0.9630	14.8441	0.9560	1.0940	1.2787
8	20	56.1119	0.9555	240.0697	0.9605	0.6659	0.7926
8	30	32.9768	0.9520	57.5583	0.9445	0.9262	0.9483
8	50	23.9586	0.9570	28.8758	0.9545	1.0175	1.1092
8	100	19.5602	0.9600	20.8338	0.9550	1.0896	1.2826

Tabla A.18: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución $0{,}95\,N(\mathbf{0}, \mathbf{I}_d) + 0{,}05\mathcal{C}_d$. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\nu_{\mathrm{C}\cdot}}{\nu_{\mathrm{DS}}}}$	$\frac{\|\overline{\mathbf{x}}\|}{\|\mathbf{t}_n\|}$
2	20	12.1744	0.9620	16.7106	0.9595	1.0092	1.2260
2	30	9.7920	0.9610	12.2417	0.9595	1.0547	1.3079
2	50	8.3989	0.9630	9.4666	0.9560	1.1900	1.4578
2	100	7.4187	0.9700	7.9355	0.9590	1.3981	1.8004
3	20	16.6939	0.9610	27.9458	0.9580	0.9582	1.1489
3	30	13.2222	0.9620	17.4503	0.9540	1.0727	1.2395
3	50	10.9711	0.9670	13.0009	0.9605	1.1840	1.4368
3	100	9.6736	0.9750	10.4174	0.9580	1.3780	1.7698
4	20	21.3464	0.9510	41.4550	0.9445	0.9386	1.0944
4	30	16.9176	0.9670	23.2288	0.9610	1.0745	1.2703
4	50	13.4051	0.9650	16.2419	0.9595	1.1698	1.4181
4	100	11.6219	0.9715	12.6392	0.9580	1.3426	1.8048
5	20	27.2366	0.9510	64.0125	0.9470	0.8749	1.0057
5	30	20.0054	0.9590	29.1497	0.9485	1.0385	1.1910
5	50	15.9508	0.9650	19.1330	0.9580	1.1630	1.4284
5	100	13.6632	0.9730	14.8441	0.9600	1.2988	1.6645
8	20	56.1119	0.9580	240.0697	0.9485	0.6895	0.8779
8	30	32.9768	0.9590	57.5583	0.9530	0.9795	1.0942
8	50	23.9586	0.9645	28.8758	0.9565	1.1337	1.3902
8	100	19.5602	0.9710	20.8338	0.9600	1.2704	1.7727

Tabla A.19: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución $0{,}90\, N(\mathbf{0}, \mathbf{I}_d) + 0{,}10\mathcal{C}_d$. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\mathcal{V}_{\mathrm{C}\cdot}}{\mathcal{V}_{\mathrm{DS}}}}$	$\frac{\Vert\overline{\mathbf{x}}\Vert}{\Vert\mathbf{t}_n\Vert}$
2	20	12.1744	0.9645	16.7106	0.9595	1.0519	1.2873
2	30	9.7920	0.9670	12.2417	0.9580	1.1511	1.3146
2	50	8.3989	0.9720	9.4666	0.9610	1.2108	1.3870
2	100	7.4187	0.9800	7.9355	0.9620	1.3020	1.3716
3	20	16.6939	0.9610	27.9458	0.9625	1.0102	1.1857
3	30	13.2222	0.9680	17.4503	0.9605	1.1352	1.3027
3	50	10.9711	0.9705	13.0009	0.9625	1.2299	1.2912
3	100	9.6736	0.9830	10.4174	0.9630	1.3061	1.3384
4	20	21.3464	0.9640	41.4550	0.9550	0.9911	1.1543
4	30	16.9176	0.9720	23.2288	0.9610	1.1435	1.2454
4	50	13.4051	0.9700	16.2419	0.9630	1.2111	1.3921
4	100	11.6219	0.9815	12.6392	0.9630	1.2903	1.3609
5	20	27.2366	0.9580	64.0125	0.9610	0.9253	1.0750
5	30	20.0054	0.9650	29.1497	0.9580	1.1070	1.2912
5	50	15.9508	0.9730	19.1330	0.9630	1.2069	1.4072
5	100	13.6632	0.9810	14.8441	0.9660	1.2875	1.3973
8	20	56.1119	0.9595	240.0697	0.9475	0.7188	0.9663
8	30	32.9768	0.9610	57.5583	0.9615	1.0201	1.1362
8	50	23.9586	0.9700	28.8758	0.9610	1.1644	1.3456
8	100	19.5602	0.9800	20.8338	0.9660	1.2477	1.3736

Tabla A.20: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución $0{,}95\,N(\mathbf{0}, \mathbf{I}_d) + 0{,}05\,N(\mathbf{0}, 25\,\mathbf{I}_d)$. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\mathcal{V}_{\mathrm{C}\cdot}}{\mathcal{V}_{\mathrm{DS}}}}$	$\frac{\Vert\overline{\mathbf{x}}\Vert}{\Vert\mathbf{t}_n\Vert}$
2	20	12.1744	0.9770	16.7106	0.9715	1.2718	1.5478
2	30	9.7920	0.9820	12.2417	0.9750	1.3477	1.6342
2	50	8.3989	0.9890	9.4666	0.9730	1.4440	1.6286
2	100	7.4187	0.9960	7.9355	0.9750	1.5147	1.6865
3	20	16.6939	0.9780	27.9458	0.9665	1.1768	1.4892
3	30	13.2222	0.9820	17.4503	0.9660	1.3355	1.5747
3	50	10.9711	0.9880	13.0009	0.9725	1.4306	1.6146
3	100	9.6736	0.9960	10.4174	0.9750	1.5457	1.6508
4	20	21.3464	0.9715	41.4550	0.9660	1.1350	1.4800
4	30	16.9176	0.9790	23.2288	0.9660	1.3155	1.5700
4	50	13.4051	0.9870	16.2419	0.9730	1.4421	1.6532
4	100	11.6219	0.9955	12.6392	0.9750	1.5272	1.6767
5	20	27.2366	0.9765	64.0125	0.9630	1.0739	1.3234
5	30	20.0054	0.9770	29.1497	0.9670	1.2821	1.5510
5	50	15.9508	0.9860	19.1330	0.9730	1.4193	1.6428
5	100	13.6632	0.9960	14.8441	0.9770	1.5274	1.7042
8	20	56.1119	0.9690	240.0697	0.9640	0.7895	1.2064
8	30	32.9768	0.9740	57.5583	0.9665	1.1723	1.4651
8	50	23.9586	0.9840	28.8758	0.9700	1.3610	1.6582
8	100	19.5602	0.9960	20.8338	0.9770	1.4863	1.6779

Tabla A.21: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución $0{,}90\,N(\mathbf{0}, \mathbf{I}_d) + 0{,}10\,N(\mathbf{0}, 25\,\mathbf{I}_d)$. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\mathcal{V}_{\mathrm{C}\cdot}}{\mathcal{V}_{\mathrm{DS}}}}$	$\frac{\lVert\overline{\mathbf{x}}\rVert}{\lVert\mathbf{t}_n\rVert}$
2	20	12.1744	0.9820	16.7106	0.9645	1.2545	1.5281
2	30	9.8752	0.9770	12.2417	0.9615	1.2132	1.4838
2	50	8.3989	0.9750	9.4666	0.9580	1.1618	1.2640
2	100	7.4187	0.9690	7.9355	0.9510	1.0986	1.1271
3	20	16.6939	0.9730	27.9458	0.9620	1.0893	1.3154
3	30	13.2222	0.9680	17.4503	0.9550	1.1169	1.2739
3	50	10.4174	0.9690	13.0009	0.9590	1.0932	1.1640
3	100	9.6736	0.9640	10.4174	0.9510	1.0685	1.1104
4	20	21.3464	0.9690	41.4550	0.9575	1.0027	1.1511
4	30	16.9176	0.9710	23.2288	0.9590	1.0715	1.1179
4	50	13.4051	0.9675	16.2419	0.9585	1.0491	1.0903
4	100	11.6219	0.9615	12.6392	0.9500	1.0409	1.0538
5	20	27.2366	0.9650	64.0125	0.9580	0.9139	1.0417
5	30	20.0054	0.9640	28.1497	0.9545	1.0277	1.0641
5	50	15.9508	0.9605	19.1330	0.9560	1.0380	1.0742
5	100	13.6632	0.9610	14.8441	0.9555	1.0297	1.0289
8	20	56.1119	0.9570	240.0697	0.9585	0.6795	0.8313
8	30	32.9768	0.9580	57.5583	0.9575	0.9383	0.9098
8	50	23.9586	0.9610	28.8758	0.9560	1.0128	1.0075
8	100	20.8338	0.9580	20.8338	0.9540	1.0148	1.0334

Tabla A.22: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución cuando se incluye un dato atípico de norma 8. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

		Clásica		Robusta		Relaciones	
d	n	K_{C}	π_{C}	K_{DS}	π_{DS}	$\sqrt[d]{\frac{\mathcal{V}_{\mathrm{C}\cdot}}{\mathcal{V}_{\mathrm{DS}}}}$	$\frac{\lVert\overline{\mathbf{x}}\rVert}{\lVert\mathbf{t}_n\rVert}$
2	20	12.1744	0.9805	16.7106	0.9690	1.6882	2.9401
2	30	9.8752	0.9800	12.2417	0.9630	1.6022	2.4456
2	50	8.3989	0.9785	9.4666	0.9590	1.4923	2.0272
2	100	7.4187	0.9785	7.9355	0.9560	1.3357	1.6271
3	20	16.6939	0.9750	27.9458	0.9635	1.3270	2.1714
3	30	13.2222	0.9740	17.4503	0.9565	1.3527	1.9568
3	50	10.4174	0.9735	13.0009	0.9580	1.2922	1.6924
3	100	9.6736	0.9740	10.4174	0.9510	1.2161	1.3533
4	20	21.3464	0.9700	41.4550	0.9595	1.1696	1.8142
4	30	16.9176	0.9740	23.2288	0.9600	1.2391	1.7141
4	50	13.4051	0.9685	16.2419	0.9575	1.1902	1.5008
4	100	11.6219	0.9690	12.6392	0.9530	1.1497	1.2435
5	20	27.2366	0.9690	64.0125	0.9610	1.0440	1.5945
5	30	20.0054	0.9665	28.1497	0.9505	1.1543	1.5223
5	50	15.9508	0.9685	19.1330	0.9580	1.1496	1.3649
5	100	13.6632	0.9690	14.8441	0.9550	1.1106	1.2225
8	20	56.1119	0.9635	240.0697	0.9520	0.7410	1.1539
8	30	32.9768	0.9600	57.5583	0.9585	1.0135	1.2114
8	50	23.9586	0.9650	28.8758	0.9550	1.0785	1.2185
8	100	20.8338	0.9650	20.8338	0.9540	1.0657	1.1192

Tabla A.23: Probabilidad de cobertura real para las regiones de tolerancia clásica y robusta (π_{C} y π_{DS}, respectivamente) asociadas a los factores de tolerancia K_{C} y K_{DS}, obtenida bajo una distribución cuando se incluye un dato atípico de norma 16. Los factores de tolerancia corresponden a los obtenidos para datos con distribución normal para $q = \delta = 0.95$.

Apéndice B

Apéndice 2. Figuras

Combinación 1

Combinación 2

Combinación 3

Combinación 4

Figura B.1: QQ–plot entre la distribución empírica de la probabilidad de cobertura y la distribución Beta, para las combinaciones 1-4.

Combinación 5

Combinación 6

Combinación 7

Combinación 8

Figura B.2: QQ–plot entre la distribución empírica de la probabilidad de cobertura y la distribución Beta, para las combinaciones 5-8.

Combinación 9

Combinación 10

Combinación 11

Combinación 12

Figura B.3: QQ–plot entre la distribución empírica de la probabilidad de cobertura y la distribución Beta, para las combinaciones 9-12.

Combinación 13

Combinación 14

Combinación 15

Combinación 16

Figura B.4: QQ–plot entre la distribución empírica de la probabilidad de cobertura y la distribución Beta, para las combinaciones 13-16.

Combinación 17

Combinación 18

Combinación 19

Combinación 20

Figura B.5: QQ–plot entre la distribución empírica de la probabilidad de cobertura y la distribución Beta, para las combinaciones 17-20.

Figura B.7: Medias empíricas y medias aproximadas por Guttman (1970).

(a)

(b)

Figura B.8: Función de influencia de la probabilidad de cobertura cuando se utilizan los estimadores clásicos (a) y los estimadores de Donoho–Stahel (b). Se grafican las funciones correspondientes a $K = 2$ (en negro), $K = 4$ (en verde), $K = 6$ (en azul claro), $K = 8$ (en rojo) y $K = 10$ (en azul oscuro).

(a)

(b)

Figura B.9: Función de influencia de la probabilidad de cobertura cuando se utilizan los estimadores clásicos (a) y los estimadores de Donoho–Stahel (b). Se grafican las funciones correspondientes a $K = 2$ (en negro), $K = 4$ (en verde), $K = 6$ (en azul claro), $K = 8$ (en rojo) y $K = 10$ (en azul oscuro).

Apéndice C

Apéndice 3. Programas

function [cw, xrf]=SDE(x,maxrep)
% Funcion que calcula el estimador de Donoho-Sathel
% xrf: estimador de posicion
% cw: estimador de escala
% x: es la matriz de datos cno n filas y p columnas
% cds: es la const para huber con q=2
% maxrep: es la cantidad de direcciones

```
[n,p]=size(x); % Dimensiones de la matriz
const=SESGOSDE(p); % Constante correctora del sesgo
cds=chi2inv(0.95,p)^0.5;
cte=0.674; % constante del MAD
p1=p-1; % Cantidad de vectores para armar el hiperplano
pv=zeros(n,1); % Vector de pesos maximos
rep=1;
while rep <= maxrep %recorre las direcciones
    jvec=ranpn1(n,p);
    xsm=x(jvec,:); %construye el hiperplano
    xarr=xsm(2:p,:);
    xres=ones(p1,1)*xsm(1,:);
    y=xarr-xres;
    if rank(y)<p1 %controla si los puntos son l.i.
```

```
        continue;
    end
    rep=rep+1;
    v=null(y);%calcula la direccion ortogonal
    z=x*v;
    med1=median(z);
    mvec=med1.*ones(n,1);
    zm=abs(z-mvec);
    med2=median(zm);
    pe=cte*(zm./med2);
    pv=max(pv,pe);% conserva el maximo
end
w=huber(pv,cds);%calcula los pesos
w=w';
m=sum(w);
wb=w*ones(1,p);
wx=wb.*x;
xrf=sum(wx)/m;%estima la media
xd=sqrt(wb).*(x-ones(n,1)*xrf);
cw=const.*(xd'*xd)/m;%estima la covarianza
```

```
function [k,error]=FACROB(n,d,q,a,R,N,M)
% Funcion para hallar los Factores K ROBUSTOS, devuelve la
% matriz de factores y la matriz de desvios
% en ambas las filas representan niveles de confianza y las
% columnas las proporciones de conertura
% basada en el SDE
% R es para la cobertura
% N es para la confianza
% M es la cantidad de direcciones del SDE
% FACROB(30,2,[.90,.95,.99],[.10,.05,.01],500,500,500)

tic
q=q.*100;
a=a.*100;
cob=zeros(N,length(q)); % Guardo cuantiles q de distancias
cob2=zeros(N,length(q)); %Guardo cuant. q+1.96*sqrt(q*(1-q)/R)
for i=1:N        % Bucle para la confianza
    mat=mvnrnd(zeros(1,d),eye(d),n); % genero n norm. d-variadas
    [vari,media]=SDE(mat,M); % Est centro y var robustos(SDE)
    mat2= mvnrnd(zeros(1,d),eye(d),R); % muestra prop. de cob
    y=(mat2-repmat(media,R,1));
    maha=sum((y*inv(vari).*y)');
    cob(i,:)=prctile(maha,q);
    cob2(i,:)=prctile(maha,(q+1.96.*sqrt(q.*(100-q)./R)));
end
k=prctile(cob,100-a);
kcons=prctile(cob2,(100-a)+1.96*sqrt(a.*(100-a)/N));
error=kcons-k;
toc
```

function mat=MVT(n,d,g)
% Funcion que genera n variables t-MULTIVARIADAS
% en dimension d con g grados de libertad
% si g=1 es una CAUCHY multivariada

```
y=chi2rnd(g,n,1);
mat=mvnrnd(zeros(1,d),eye(d),n).*repmat(sqrt(g./y),1,d);
```

function ses=SESGOSDE(p) % Hasta dimension 10
% Constantes de sesgo para el SDE hasta dimenson 10

```
ses=1.0413708*(p==2)+1.0070053*(p==3)+
1.0000000*(p==4)+0.9957046*(p==5)+
0.9927844*(p==6)+0.9906613*(p==7)+
0.9890429*(p==8)+0.9877651*(p==9)+
0.9867286*(p==10)+0.9835156*(p>10);
```

function w=huber(u,cds)
% Funcion de Huber con q=2

```
w=min(1,(cds./u).^2);
```

```
function salida=COMPFACCONT(n,d,a,q,eps,tipo)
% Funcion que compara para un n, d, a y q la cobertura real, el
% volumen de la region ROBUSTA y la region CLASICA bajo la
% contaminacion de:
% TIPO=1 (eps%) de una Cauchy
% TIPO=2 (eps%) de una N(0,25I)
% con probabilidad de cobertura medida por la normal.
% salida=COMPFACCONT(20,4,.05,.95,.1,1)

tic
load('c:/Andres/Tolerancia/Factores/KR.mat');
load('c:/Andres/Tolerancia/Factores/KC.mat');
N=1000; % Cantidad de muestras
M=1000; % Cantidad de direcciones
R=1000; % Cantidad de puntos p/estimar la prob de cob.
a=a.*100;
q=q.*100;
aes=[50,25,10,5,1];
qes=[50,45,90,95,99];
secu=1:5;
p1=secu(aes==a);
p2=secu(qes==q);
cob=zeros(N,2);     % proporcion de cobertura
vol=zeros(N,2); % Volumen de la region
lejo=zeros(N,2); % Lejania al cero
dete=zeros(N,2); % Determinante de la region
k1=KR(p1,p2,n,d);
k2=KC(p1,p2,n,d);
for i=1:N
    if tipo==1
    matcont=MVT(n,d,1); % genero n t cuachys
    else
```

```
    matcont=mvnrnd(zeros(1,d),25.*eye(d),n); % genero n normales
    end
    matnor=mvnrnd(zeros(1,d),eye(d),n); % genero n normales mult
    cc=repmat(binornd(1,eps,n,1),1,d); % genero n bernoullis con p=eps
    mat=cc.*matcont+(1-cc).*matnor; % contamino
% ***** Region Robusta ******
    [vari,media]=SDE(mat,M); % Est centro y var robustos (SDE)
    dete(i,1)=det(vari);
    vol(i,1)=sqrt(dete(i,1).*(k1^d)).*pi^(d./2)./gamma(d./2+1);
    lejo(i,1)=sqrt(media*media');
    mat2=mvnrnd(zeros(1,d),eye(d),R); % muestra prop. de cob NORMAL
    y=(mat2-repmat(media,R,1));
    maha=sum((y*inv(vari).*y)');
    cob(i,1)=sum(maha<k1)/R;
% ***** Region Clasica ******
    media=mean(mat); % calculo la media
    vari=cov(mat); % calculo la matriz de covarianzas
    dete(i,2)=det(vari);
    vol(i,2)=sqrt(dete(i,2).*(k2^d)).*pi^(d./2)./gamma(d./2+1);
    lejo(i,2)=sqrt(media*media');
    mat2=mvnrnd(zeros(1,d),eye(d),R); % muestra prop. de cob NORMAL
    y=(mat2-repmat(media,R,1));
    maha=sum((y*inv(vari).*y)');
    cob(i,2)=sum(maha<k2)/R;
end
toc
salida=[k1,k2;prctile(cob,a);median(vol);median(lejo)];
```

Bibliografía

[1] Butler, R. (1982) Nonparametric interval and point prediction using data trimmed by a Grubbs–type outlier rule. *Ann. Statist.*, **10**, 197-204

[2] Canavos, G. & Koutraouvaelis, I. (1984) The robustness of two sided tolerance limits for normal distributions. *J. Qual. Technol.*, **16**, 144-149.

[3] Chew, V. (1966). Confidence, prediction and tolerance regions for the multivariate normal distribution. *J. Am. Statist. Assoc.*, **61**, 605-617.

[4] Croux, C. & Haesbroeck, G. (2000). Principal Component Analysis based on Robust estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika*, **87**, 603-18.

[5] Croux, C. & Joossens, K. (2004). Influence of observations on the misclassification probability in quadratic discriminant analysis. In Press in *J. Multiv. Anal.*.

[6] Cuevas, A. & Baillo, A. (2003). Parametric versus nonparametric tolerance regions in detection problems. *Working Paper 03-70, Statistics and Econometrics Series 17.*

[7] Donoho, D. L. (1982). *Breakdown Properties of Multivariate Location Estimators.* Ph.D. qualifying paper, Harvard University.

[8] Fernholz, L. & Gillespie, J. (2001). Content-corrected tolerance limits based on bootstrap. *Technometrics*, **43**, 147-155.

[9] Fernholz, L.T. (2002). Robustness Issues regarding Content Corrected Tolerance Limits. *Metrika*, **55**, 53-66.

[10] Fraser, D. (1951). Nonparametric tolerance regions. University of Toronto.

[11] Fraser, D. & Guttman, I. (1956). Tolerance regions. *Ann. Math. Statist.*, **27**, 162-179.

[12] Fuchs, C. & Kenett, S. (1987). Multivariate tolerance regions and $F-$tests. *J. Qual. Technol.*, **19**, 122-131.

[13] Fuchs, C. & Kenett, S. (1988). Appraisal of ceramic substrates by multivariate tolerance regions. *The Statician*, **37**, 401-411.

[14] Gervini, D. (2002). The influence function of the Donoho–Stahel estimator of multivariate location and scale. *Stat. Probab. Letters*, **60**, 425-435.

[15] Guttman, I. (1970). Construction of $\beta-$content tolerance regions at confidence level γ for large samples for k-variate normal distribution. *Ann. Math. Statist.*, **41**, 376-400.

[16] John, S. (1962). A tolerance region for multivariate normal distributions. *Sankhya*, Serie A, **25**, 363-368.

[17] Krishnamoorthy, K. & Mathew, T. (1999). Comparison of approximation methods for computing tolerance factors for a multivariate normal population. *Technometrics*, **41**, 234-249.

[18] Lopuhaä, H. P. (1990). *Estimation of Location and Covariance with High Breakdown Point.* Ph. D. Thesis. Delft University of Technology, Netherlands.

[19] Maronna, R. A. (1976). Robust M–Estimators of Multivariate Location and Scatter. *Ann. Statist.,* **4**, 51-67.

[20] Maronna, R. A. & Yohai, V. J. (1995). The Behavior of the Stahel–Donoho Robust Multivariate Estimator. *J. Amer. Statist. Assoc.,* **90**, 330-341.

[21] Maronna, R. A. & Yohai, V. J. (1998). Robust Estimation of Multivariate Location and Scatter . In *Encyclopedia of Statistical Sciences Update Volume 2,* eds. S. Kotz, C. Read and D. Banks, New York, Wiley, 589-596.

[22] Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory.* Wiley: New York.

[23] Murphy, R. (1948). Non–parametric tolerance limits. *Ann. Math. Statist.*, **19**, 551-589.

[24] Odeh, R. & Owen, D. (1980). *Tables for Normal Tolerance Limits, Sampling Plans and Screening.*New York: Marcel Dekker.

[25] Proschan, F. Confidence and tolerance intervals for the normal distribution. *J. Am. Statist. Assoc.*, **48**, 550-564.

[26] Rousseeuw, P. (1985). Multivariate estimation with High Breakdown Point. In *Mathematical Statistics and Applications,* (W. Grossmann, G. Pflug, I. Vincze and W. Werz, eds.), Dordrecht: Reidel, 283-297.

[27] Rousseeuw, P. J. & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Am. Statist. Assoc.*, **85**, 633-639.

[28] Stahel, W. (1981). *Robust estimation: Infinitesimal Optimality and Covariance Matrix Estimation.* Thesis (in German), ETH, Zurich.

[29] Yohai, V. & Maronna, R. (1995). The behavior of the Stahel-Donoho multivariate estimators. *J. Am. Statist. Assoc.*, **85**, 330-341.