Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

DETECCIÓN DE ATIPICIDADES EN ANÁLISIS DE DATOS

Autor: Héctor Hugo de la Torre Díaz
Tutor: Luis Ángel García Escudero
Año 2023

Agradecimientos

Me gustaría dedicar unas palabras de agradecimiento a todas las personas que me han acompañado y apoyado durante estos 5 años de trayectoria universitaria.

En primer lugar, a Luis Ángel, mi tutor, por su dedicación y ayuda constante para guiar este Trabajo Fin de Grado.

A mi familia, por su paciencia, comprensión y apoyo en todo momento.

A mis amigos y compañeros, tanto de Valladolid como del Erasmus, que me han acompañado en muchas aventuras y me han animado a continuar cuando más lo necesitaba.

Por último, a los voluntarios de ESN Valladolid, por hacerme redescubrir mi ciudad natal y ser mi segunda familia. Gracias por tanto.

Resumen

Identificar y tratar observaciones atípicas es un paso fundamental en cualquier análisis estadístico, dado que una atipicidad puede influir negativamente en el resultado del análisis y marcar la diferencia a la hora de evitar consecuencias catastróficas. Además, la detección de atipicidades es importante puesto que esas observaciones atípicas pueden tener un interés en sí mismo, por tener características especiales. Tanto es así que la detección de atipicidades puede considerarse una disciplina transversal en muchos campos desde la medicina, la economía, la ciberseguridad o la industria, entre otras. Cada día se siguen desarrollando nuevas técnicas para detectar anomalías en conjuntos de datos de todo tipo, cada una con sus características propias y útiles en función del problema a tratar. En este trabajo se han recopilado algunos de los principales métodos no supervisados y supervisados de detección de atipicidades en análisis de datos numéricos, su implementación en R junto a las librerías recomendadas y su aplicación a conjuntos de datos.

Abstract

Identifying and dealing with outliers is fundamental for any statistical analysis, since an outlier could negatively influence the analysis results and make a difference in avoiding catastrophic consequences. Furthermore, the detection of anomalies is important because those atypical observations may have inherent interest due to their special characteristics. So much so that outlier analysis could be considered a cross-discipline in many fields such as medicine, economics, cybersecurity or industry, to give some examples. Every day new techniques continue to be developed in order to detect anomalies in datasets of all types, having each technique its own characteristics and focused on the problem to be addressed. In this work some of the main unsupervised and supervised methods for outlier detection have been compiled, along with their implementation in R, the useful packages and their application to datasets.

Índice general

Capítulo 1

Introducción

Outliers, residuales, atípicos... Todas estas palabras en términos de estadística hacen referencia al mismo tipo de dato, aquel valor que es significativamente diferente del resto de los datos que constituyen nuestro conjunto de datos (en contraste con los valores aparentemente normales, a veces referidos como "inliers"). Una posible definición formal de qué es un outlier es la siguiente (Hawkins 1980):

"Un outlier es una observación que se desvía tanto de las demás observaciones provocando sospechas de que fue generada por un mecanismo diferente".

La terminología puede diferir en función de los diferentes tipos de usuarios, al igual que la definición de un "outlier" es adaptada al caso práctico en el que se está aplicando. Por lo general, independientemente del ámbito de aplicación y la técnica de detección de outliers usada, el análisis de valores atípicos puede verse como un problema de clasificación (usualmente no supervisado). Identificar y clasificar un outlier no es siempre fácil, ni existe un método perfecto que lo logre. De hecho, suele ser un criterio muchas veces bastante subjetivo, dada la dificultad de definir cómo de distinta debe ser una observación para ser considerada atípica. Además, algunos autores diferencian entre "ruido" o valor "atípico débil" respecto a la anomalía o valor "atípico fuerte", existiendo una zona gris donde al final prevalece la decisión del autor para elegir el método de detección y su posible eliminación del conjunto de datos según considere.

La mayoría de los algoritmos actuales para detección de atipicidades producen una salida perteneciente a uno de estos dos tipos:

- Un valor numérico o *score* que cuantifique la probabilidad o la tendencia de una observación cualquiera para ser considerada outlier o atípico. A mayor *score*, mayor probabilidad de que dicha observación sea atípica.
- Una etiqueta dicotómica, indicando si la observación es o no un outlier. En aquellos casos donde se requiera tomar una decisión de si una observación cualquiera es o no atípica, podemos convertir en etiquetas dicotómicas las puntuaciones o *scores* que cuantificaban cómo de atípica es una observación fijando un umbral.

Además, se debe tener en cuenta la naturaleza de las variables y sus relaciones en el conjunto de datos. Este permite utilizar variables categóricas, numéricas o mixtas. También, entender el contexto y la procedencia de esos datos ayuda a obtener mejores resultados.

La detección y análisis de outliers puede aplicarse en infinidad de campos como la medicina, economía, movilidad o meteorología, entre otros.

Por ejemplo, una de las aplicaciones típicas del análisis de outliers es la detección de fraudes en finanzas. Con estas técnicas se pretende limitar el gran impacto económico que supone en nuestra sociedad el uso de forma fraudulenta de tarjetas de crédito, reclamaciones falsas de seguros, manipulación del mercado de valores... Por desgracia, seguimos viendo noticias donde se destapan casos de fraude financiero por todo el mundo. Por ejemplo, se conoce que la fiscalía europea abrió investigaciones por un volumen de fraude cercano a los 10.000 millones de euros en 2022 (Gómez 2023). Es un tema candente a día de hoy, donde van surgiendo, a medida que pasa el tiempo, nuevos métodos y enfoques para afrontar los nuevos casos de fraude en nuestra sociedad. También es frecuente aplicar la detección de atipicidades a la detección de anomalías en el mercado de valores, lo que puede marcar oportunidades de negocio o anticipar problemáticas.

En medicina, identificar una anomalía o un dato atípico puede suponer la diferencia entre detectar si un paciente tiene una determinada enfermedad o no. No diagnosticar una patología cuando esta sí se da, o diagnosticarla erróneamente supone un coste muy alto que es inasumible. Por ello, se utilizan métodos de detección de anomalías en datos obtenidos por sensores o imágenes para clasificar al paciente correctamente (Gaspar et al. 2011).

En controles de calidad para fábricas, el análisis de outliers es utilizado tanto para detectar fallos en características individuales del producto, como dentro de todo el proceso industrial de fabricación. Uno de los posibles objetivos es estimar cuándo el proceso de fabricación está siendo anómalo, analizando fallos o defectos en las piezas fabricadas. De la misma forma, podemos extender la aplicación del análisis de outliers a la detección de fallos en sistemas mecánicos o la detección de fallas estructurales. En ocasiones, esto da lugar a procedimientos de detección temprana de fallos y a establecer alertas previas a fallos más catastróficos.

En el campo de la ciberseguridad, donde a día de hoy son muy comunes intrusiones debidas a brechas de seguridad, los algoritmos de detección de outliers proporcionan una herramienta eficaz para localizar estas intrusiones anómalas. Sean intrusiones en sistemas locales, en servidores, o en red, el problema es el mismo: determinar comportamientos anómalos (atípicos) que permitan mejorar la seguridad y prevenir futuros ataques.

Comportamientos bruscos en el tiempo, en el clima, o ecosistema, entre otras aplicaciones de la estadística en Ciencias Naturales, son estudiados mediante técnicas de detección de atipicidades. Más en concreto, registros de temperaturas anómalos en España, temperaturas más altas del agua del mar Mediterráneo en la primavera, y el aumento considerable de las olas de calor en España son fenómenos que han sido estudiados mediante el análisis de datos atípicos (Crespo Garay 2021).

Existen otras muchas aplicaciones de la detección de outliers, desde el análisis a las redes sociales, el estudio de la movilidad en una ciudad o región, en astronomía, ...

En esta memoria se exponen varios de los métodos y técnicas más populares para la detección y manejo de atípicos, así como su implementación en R y su aplicación a ejemplos de conjuntos de datos abiertos. En concreto, en el Capítulo 2 se tratarán métodos no supervisados de detección de outliers, como el análisis de valores extremos, métodos basados en clustering, en los modelos lineales y los métodos basados en proximidad.

El Capítulo 3 expone las técnicas de agregación o "ensembles" como combinaciones de algoritmos para incrementar la robustez y precisión en el problema de clasificación de atipicidades, así como ejemplos de ensembles comunes y útiles. Por otro lado, el Capítulo 4 describe métodos de análisis supervisado aplicado a outliers, junto a las dificultades que conllevan estos métodos y las posibles soluciones a dichas dificultades.

La implementación en R de muchos de los métodos descritos en los anteriores capítulos se muestra en el Capítulo 5, a través de ejemplos de datos abiertos junto a los paquetes empleados en cada caso.

Finalmente, en el Capítulo 6 se extraen las conclusiones obtenidas de realizar este trabajo y se exponen algunas posibles líneas futuras de investigación.

Capítulo 2

Métodos no supervisados

En los métodos no supervisados no se dispone de ejemplos de observaciones que hayan sido previamente etiquetadas ya como "outliers", al contrario de lo que sucede con métodos supervisados, donde se puede "aprender" a clasificar atipicidades gracias a que algunas observaciones con sus correspondientes valores están previamente etiquetadas como tal.

2.1. Análisis de valores extremos

Los primeros métodos diseñados para la detección de atípicos están basados en modelos probabilísticos y estadísticos. Fueron desarrollados mucho antes del inicio de la computación como la ciencia que conocemos ahora, y se siguen aplicando actualmente, debido a su sencillez y utilidad. A continuación, se expone el análisis de valores extremos para datos univariantes y datos multivariantes.

2.1.1. Caso univariante

Para el caso unimodal, podemos decir que una observación es extrema si ésta se encuentra en uno de los extremos de la distribución, zonas conocidas generalmente como "colas" y que tienen baja densidad. Para casos como el mostrado en la Figura 2.1, cuya función de densidad es f_X, las colas de dicha distribución quedan definidas de la siguiente manera:

$$\{x : f_X(x) \leq c\} = \{x \leq c_1\} \cup \{x \geq c_2\},$$

para valores $c_1 < c_2$ que sirven para acotar las colas.

La desigualdad de Chebychev nos ayuda a fijar unas cotas para una distribución aleatoria cualquiera. Es una desigualdad débil dado que no requiere apenas de asunciones sobre la variable aleatoria (incluso no hace falta que sea unimodal).

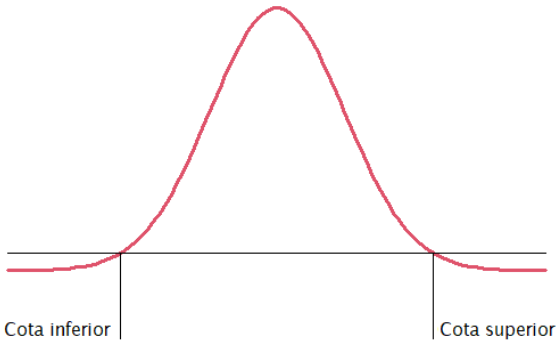


Figura 2.1: Ejemplo de distribución univariante

Desigualdad de Chebychev

La desigualdad de Chebychev nos dice que para una variable aleatoria X con media $\mu = E[X]$ y varianza $\sigma^2 = \text{Var}[X]$, se tiene que:

$$P(|X - \mu| > k) \leq \frac{\sigma^2}{k^2},$$

para cualquier valor de la constante $k > 0$.

La desigualdad de Chebychev es directa desde la desigualdad de Markov, que dice que para cualquier variable positiva X se tiene

$$P(X \geq a) \leq \frac{E[X]}{a}$$

para cualquier $a > 0$. Nótese que

$$E[X] = \int_0^\infty x f_X(x) dx \geq \int_0^a x f_X(x) dx \geq \int_0^a a f_X(x) dx = aP(X \geq a)$$

La desigualdad de Chebychev surgiría de aplicar la desigualdad de Markov con la variable $Y = (X - E[X])^2$ y $a = k^2$.

Como consecuencia de la desigualdad de Chebychev, tenemos

$$P\left(\left|\frac{X - \mu}{\sigma}\right| > k\right) \leq \frac{1}{k^2}$$

y, por tanto, podemos decir que x puede ser atípico si

$$\left\{x : \left|\frac{X - \mu}{\sigma}\right| > k\right\}$$

cuando se considere un $k > 0$ relativamente grande puesto que $1/k^2$ se hace bastante pequeño.

La Tabla 2.1 muestran las correspondientes acotaciones de las probabilidades de esas colas para diferentes valores de la constante k.

k	Acotación de $P(\lvert\frac{X-\mu}{\sigma}\rvert > k)$
2	0.2500
3	0.1111
4	0.0625
5	0.0400
6	0.0279

Tabla 2.1: Acotaciones para distintos k en la aplicación de la desigualdad de Chebychev

La versión frecuentista de la desigualdad de Chebychev, para una muestra de datos $\{x_i\}_{i=1}^n$, con

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

y

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

sería

$$fr\{x_i : |x_i - \overline{x}| > k \cdot s\} \geq \frac{1}{k^2},$$

donde $fr(\cdot)$ denota la frecuencia relativa. Podemos, pues, etiquetar como atípicos las observaciones x_i de la muestra $\{x_i\}_{i=1}^n$ que verifican $|z_i| > k$ para

$$z_i = \frac{x_i - \overline{x}}{s}$$

para un k alto ($k = 3$ o $k = 6$). En el caso de $k = 6$, estamos ante la conocida regla "six-sigma" que es muy utilizada en Control Estadístico de Procesos (SPC) en la Ingeniería.

Además de la desigualdad de Chebychev, existen otras desigualdades que proporcionan cotas más finas, como la desigualdad de Chernoff para la suma de variables de Bernoulli, pero que no serán consideradas en esta memoria.

Distribución Normal

Podemos afinar la acotación que proporciona la desigualdad de Chebychev si conocemos la distribución que ha podido generar los datos. Por ejemplo, para una variable aleatoria X que sigue una distribución normal X de parámetros μ y σ^2, y que denotaremos como $X \sim N(\mu, \sigma^2)$, con función de densidad:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

entonces las cotas de esta distribución unimodal y simétrica son las siguientes:

$$\{x : f_X(x) \leq c_1\} = \left\{x : \left|\frac{x-\mu}{\sigma}\right| \geq c_2\right\},$$

para constantes c_1 y c_2 positivas. Podemos, por tanto, considerar scores basados en

$$s(x) = \left| \frac{x - \mu}{\sigma} \right|. \tag{2.1}$$

Tanto es así que, si $X \sim N(\mu, \sigma^2)$, tenemos

$$P(X \notin [\mu \pm z_{\alpha/2} \cdot \sigma]) = \alpha \text{ con } z_\alpha = \Phi^{-1}(1 - \alpha)$$

obteniendo, según el valor de α elegido, los intervalos $[\mu \pm 1.96\sigma]$ para $\alpha = 0.05$, $[\mu \pm 2.24\sigma]$ para $\alpha = 0.025$, $[\mu \pm 2,58\sigma]$ para $\alpha = 0.01$, y así sucesivamente. Esto nos permite marcar límites para $s(x)$ en (2.1) para marcar valores en las "colas" de la distribución normal basándonos en $z_{\alpha/2}$ para distintos valores de α.

Aun asumiendo normalidad, para aplicar los resultados anteriores al problema de detección de outliers, es importante notar que los parámetros μ y σ^2 son desconocidos y habrá que estimarlos a partir de la muestra $\{x_i\}_{i=1}^n$ por $\hat{\mu} = \overline{x}$ y $\hat{\sigma} = s = \sqrt{s^2}$. Esto nos lleva a que el intervalo

$$[\overline{x} \pm z_{\alpha/2} \cdot s]$$

deba incluir de forma aproximada una fracción $1 - \alpha$ de las observaciones en nuestros datos $\{x_i\}_{i=1}^n$, cuando se cumpla la hipótesis de normalidad y n sea grande (para que $\overline{x}$ y s sean buenos estimadores de los parámetros μ y σ poblacionales desconocidos). Este intervalo es notablemente menos amplio que los obtenidos simplemente aplicando la desigualdad de Chebychev. En ocasiones, la aplicación del Teorema Central del Limite puede garantizar esa suposición de normalidad para la distribución que ha generado los datos $\{x_i\}_{i=1}^n$.

En resumen, la forma principal para determinar si un punto es atípico, bajo la suposición de normalidad, es el cálculo de z-scores (2.1) o, equivalentemente, el uso de datos estandarizados o tipificados z_i con

$$z_i = \frac{x_i - \overline{x}}{s},$$

y marcar atípicos cuando $|z_i|$ exceda 1.96 ($\alpha = 0.05$), 2.24 ($\alpha = 0.025$) o 2.58 ($\alpha = 0.01$).

También se pueden buscar atípicos asociados a valores pequeños de

$$P(|Z| > |z_i|) = 2(1 - \Phi(|z_i|)),$$

con Z siendo la $N(0, 1)$ y Φ la función de distribución de Z.

Estadística robusta

Sin embargo, el problema de usar z-scores basados en $z_i = (x - \overline{x})/s$ para detectar atipicidades es que $\overline{x}$ y s están muy afectados por las observaciones atípicas, con s sobrestimando la dispersión. A modo de ejemplo, consideramos el siguiente conjunto de datos que representan alturas de individuos en metros $\{1.66, 1.72, 1.83, 1.76, 1.69, 173\}$, donde la última observación es errónea por

haber dado la altura en centímetros en lugar de en metros. La media $\overline{x}$ y desviación típica s para este conjunto de datos es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1.66 + 1.72 + 1.83 + 1.77 + 1.69 + 173}{6} \approx 30.28$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \approx 69.92$$

Dando lugar a los siguientes $|z_i|$:

$$|z_i| \approx \{0.4193, 0.4084, 0.4069, 0.4077, 0.4089, 2.0412\}$$

Para estos casos, donde los outliers pueden quedan enmascarados (masking), es mejor utilizar estimadores robustos $\widehat{\mu}$ y $\widehat{\sigma}$ en lugar de $\overline{x}$ y s, que pueden estar muy negativamente afectado por, incluso, una única observación atípica.

La alternativa es usar estimadores robustos de localización y escala como, por ejemplo, la mediana en lugar de la media, y la desviación absoluta media (MAD) o el estimador Q_n (P. Rousseeuw y Croux 1993) en lugar de la desviación típica muestral. Para estos datos tendríamos:

$$\hat{\mu} = \text{med}\{x_i\}_{i=1}^{n} = 1.745$$

y

$$\hat{\sigma} = \text{MAD}(\{x_i\}_{i=1}^{n}) = 1.4826 \cdot \text{med}\{|x_i - \text{med}\{x_i\}_{i=1}^{n}|\}_{i=1}^{n} \approx 0.1038$$

o, alternativamente,

$$Q_n = 2.2219\{|x_i - x_j| : i < j\}_{(k)} \text{ con } k = \binom{h}{2} \text{ y } h = \lfloor \frac{n}{2} \rfloor + 1,$$

donde

$$\{|x_i - x_j| : i < j\}_{(1)} \leq \{|x_i - x_j| : i < j\}_{(2)} \leq ... \leq \{|x_i - x_j| : i < j\}_{(\binom{n}{2})}$$

son todos los valores ordenados de las posibles $\binom{n}{2}$ posibles diferencias $|x_i - x_j|$ con $i < j$.

Las constantes 1.4826 (para el MAD) y 2.2219 (para Q_n) son así definidas para tener consistencia en el modelo normal. Con estos valores robustificados los $|z_i|$ son:

$$|z_i| \approx \{0.8190, 0.2409, 0.8190, 0.2409, 0.5230, 1650.1416\}$$

De hecho, existe un método gráfico de detección de outliers, los boxplots, que emplean estimadores robustos como son la mediana y el rango intercuartílico.

Boxplots

Los boxplots o diagramas de caja son uno de los principales recursos gráficos para visualizar valores extremos. El boxplot resume la información de la distribución basándose en la mediana y los cuartiles que permiten definir el rango intercuartílico o IQR, equivalente a la distancia entre el tercer y el primer cuartil (IQR $= Q_3 - Q_1$). Se calculan las cotas para marcar outliers de la siguiente forma:

$$\text{Cota inferior } = Q_1 - 1.5 \cdot \text{IQR} \quad \text{Cota superior } = Q_3 + 1.5 \cdot \text{IQR}.$$

Aquellos puntos menores que esa cota inferior, o mayores que esa cota superior son considerados como outliers y marcados como "∘" en los boxplots. Por ejemplo, en la Figura 2.2 se muestra un boxplot donde aparecen 3 outliers.

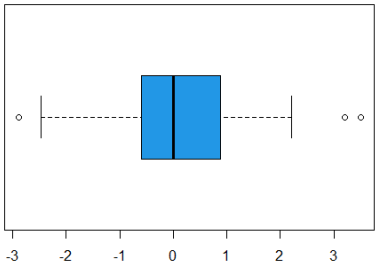


Figura 2.2: Gráfico Boxplot

2.1.2. Caso multivariante

Tratar con datos multivariantes, es decir, $\{x_i\}_{i=1}^n$ con $x_i \in \mathbb{R}^p$, es más complicado porque en $\mathbb{R}^p$ no hay un orden natural que nos diga cuando $x \leq y$ para $x, y \in \mathbb{R}^p$. Una primera forma de analizar datos multivariantes, de tal forma que podamos extender los z-scores a más de una dimensión, el es uso de la distancia de Mahalanobis $d(x; \mu, \mathbf{\Sigma})$.

Distancia de Mahalanobis

La distancia de Mahalanobis $d(x; \mu, \mathbf{\Sigma})$ sirve para medir cómo de lejos está un valor x respecto al centroide de los datos, y está definida como sigue:

$$d(x; \mu, \mathbf{\Sigma}) = \sqrt{(x-\mu)'\mathbf{\Sigma^{-1}}(x-\mu)}$$

siendo μ el vector p-dimensional de medias y $\mathbf{\Sigma}$ la matriz de covarianzas de tamaño $p \times p$. Para el caso univariante, se corresponde con los z-scores:

$$d(x; \mu, \sigma^2) = \sqrt{(x-\mu)'\frac{1}{\sigma^2}(x-\mu)} = \left|\frac{x-\mu}{\sigma}\right|$$

Si X sigue una distribución normal p-variante con vector de medias μ y matriz de covarianzas $\mathbf{\Sigma}$, es decir, $X \to N_p(\mu, \mathbf{\Sigma})$, podemos definir los extremos de forma análoga tal y como se desarrolló para la distribución normal univariante en 2.1.1 de la forma:

$$\{x : f_X(x) \leq c_1\} = \{x : (x-\mu)'\mathbf{\Sigma}^{-1}(x-\mu) \leq c_2\} = \{x : d^2(x; \mu, \mathbf{\Sigma}) \leq c_2\}$$

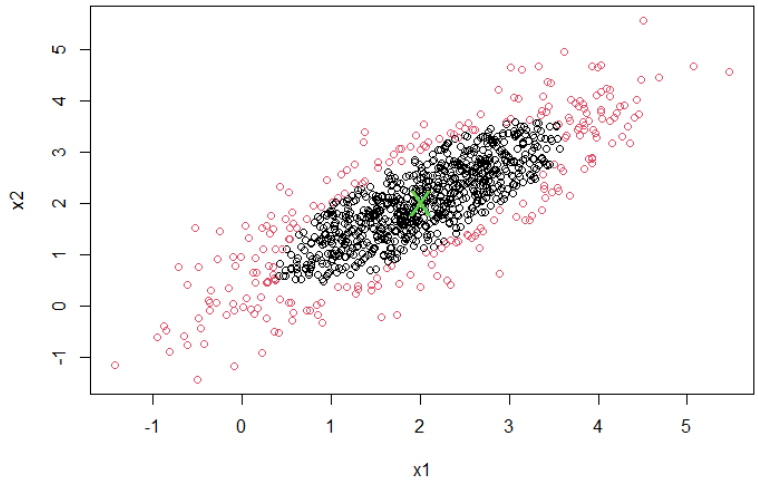


Figura 2.3: En color rojo se muestran el 25 % de observaciones con mayores distancias de Mahalanobis para una muestra de una normal bivariante con μ señalado por "×".

Si X sigue una $N_p(\mu, \mathbf{\Sigma})$, entonces $d^2(X; \mu, \mathbf{\Sigma}) = (X-\mu)'\mathbf{\Sigma^{-1}}(X-\mu)$ sigue una distribución χ^2_p, y se pueden obtener puntos de corte para los scores $d(x; \mu, \mathbf{\Sigma})$ teniendo en cuenta que:

$$P(d^2(X; \mu, \mathbf{\Sigma}) \geq \chi^2_{p,\alpha}) = \alpha,$$

para $\chi^2_{p,\alpha} = F^{-1}_{\chi^2_p}(1-\alpha)$.

De nuevo, los parámetros poblacionales μ y $\mathbf{\Sigma}$ son desconocidos y debemos usar la muestra $\{x_1, ..., x_n\}$ y los estimadores

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

y

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})'$$

para estimar esos parámetros desconocidos. $\mathbf{S}$ es la conocida matriz de varianzas-covarianzas muestral. Con esos estimadores tendremos de forma aproximada que, para $X \sim N_p(\mu, \mathbf{\Sigma})$, se debe tener:

$$P(X \notin \{x : d^2(x; \overline{x}, \mathbf{S}) \geq \chi^2_{p,1-\alpha}\}) \approx \alpha.$$

Se podría afinar aún más teniendo en cuenta que si $X_1, ..., X_n$ es una muestra aleatoria simple de una $N_p(\mu, \mathbf{\Sigma})$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})',$$

entonces

$$\frac{(n-1)^2}{n} d^2(X_i; \overline{X}, \mathbf{S}) \sim \text{Beta}\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$$

(Deliang y Liang 2021).

Si p es grande, puede ser que no exista $\mathbf{S}$ (la matriz $\mathbf{S}$ no sea invertible) y no se pueda calcular la distancia de Mahalanobis. En tal caso, podemos utilizar una técnica de regularización reemplazando $\mathbf{S}$ por $\mathbf{S}+\lambda I_p$, para un valor pequeño del parámetro de regularización $\lambda > 0$ y siendo I_p la matriz identidad en $\mathbb{R}^p$, de forma que sí exista $(\mathbf{S} + \lambda I_p)^{-1}$. La distancia de Mahalanobis regularizada sería $d^2(x; \overline{X}, \mathbf{S} + \lambda I_p)$.

Distancias de Mahalanobis robustas

En analogía a lo visto en el caso univariante, podemos reemplazar $\overline{x}$ y $\mathbf{S}$, que son poco robustos, por estimadores robustos como $\widehat{\mu}_{\text{MCD}}$ y $\widehat{\Sigma}_{\text{MCD}}$ que son derivados del estimador de *Minimum Covariance Determinant* (MCD). El MCD busca una fracción $h < n$ de observaciones, indexadas por un subconjunto de índices $\mathcal{H} = \{i_1, ..., i_h\} \subset \{1, ..., n\}$, tales que su determinante

$$\left| \frac{1}{h} \sum_{i \in \mathcal{H}} (x_i - \overline{x}_{\mathcal{H}})(x_i - \overline{x}_{\mathcal{H}})' \right|,$$

con

$$\overline{x}_{\mathcal{H}} = \frac{1}{h} \sum_{i \in \mathcal{H}} x_i$$

sea lo menor posible entre todos los subconjunto de índices posibles con h índices.

Estos estimadores $\widehat{\mu}_{\text{MCD}}$ y $\widehat{\Sigma}_{\text{MCD}}$ no están afectados por las observaciones más atípicas (cuyos subíndices no aparecen en el $\mathcal{H}$ óptimo porque incrementarían mucho el determinante). Por tanto, $d^2(x; \widehat{\mu}_{\text{MCD}}, \widehat{S}_{\text{MCD}})$ es más efectivo que $d^2(x; \overline{x}, \mathbf{S})$ para detectar observaciones atípicas, de forma análoga a lo que sucedía al usar la mediana y el MAD al calcular los z-values.

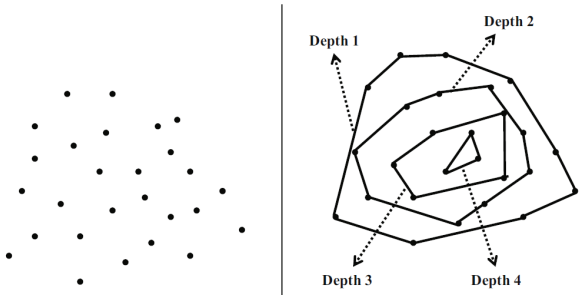
Existen otros métodos que no asumen ninguna distribución a priori, como los métodos basados en profundidades o "depth", en ángulos, los "bagplots", vistos como una generalización de los boxplots a datos bidimensionales y tridimensionales.

Basados en profundidades (depth)

Dentro de un análisis basado en profundidades, las observaciones son analizadas utilizando el concepto de "profundidad" o depth. Las observaciones con mayor profundidad son las observaciones menos extremas y esta profundidad se suele determinar atendiendo a conceptos geométricos. Esto se traduce en buscar las observaciones más profundas, etiquetar cada observación según su profundidad, y aquellas observaciones menos profundas son las susceptibles de ser atípicas. Las principales técnicas basadas en profundidades son el *Tukey depth* y *Convex Hull depth.*

El *Tukey depth* es una medida de profundidad para $x_i \in \mathbb{R}^p$ respecto a $\{x_1, ..., x_n\}$, considerando el menor número de observaciones en $\{x_1, ..., x_n\}$ incluidas en cualquier semiespacio que contenga a la observación x_i.

Convex hull es un algoritmo iterativo, donde en cada iteración t, todas las observaciones en los vértices de la "envolvente convexa" o *convex hull* son eliminadas del conjunto de datos, asignando a dichas observaciones la profundidad t. El algoritmo es repetido hasta que el conjunto de datos termina vacío, siendo la profundidad el indicador de atipicidad (*score*). Una profundidad baja (en los límites de la "envolvente conexa") para una observación es señal de atipicidad. En la Figura 2.4 se muestra un ejemplo de *Convex hull*.



(a) Data points (b) Convex hulls

Figura 2.4: Detección de atipicidades basado en profundidades (C.C. Aggarwal 2017)

Basados en ángulos

Por lo general, los puntos atípicos, al encontrarse en los extremos del conjunto de datos, pueden abarcar los demás puntos con ángulos de una pequeña variabilidad, es decir, todos los demás puntos se encuentran en direcciones similares respecto al valor atípico (véase la Figura 2.5) (C.C. Aggarwal 2017). En cambio, los puntos cercanos al centro de la nube de puntos, para encerrar a todo el junto de datos, necesitan de ángulos mucho más grandes (o prácticamente 360º si hay puntos en todas las demás direcciones).

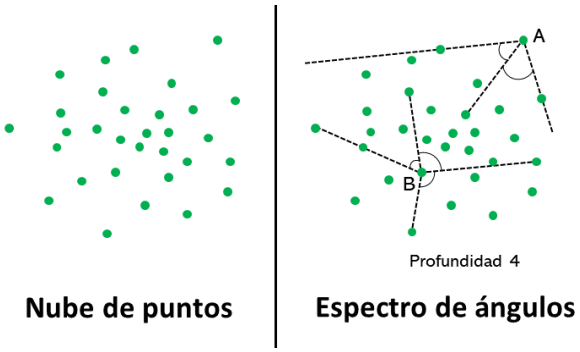


Figura 2.5: Detección de atipicidades basado en ángulos

Para decidir cuando un punto es outlier podemos utilizar como medida de amplitud el coseno ponderado. Se requieren 3 datos x_i, x_j, x_k, de los que se calcularán los vectores $\overrightarrow{x_i x_j} = x_j - x_i$ y $\overrightarrow{x_i x_k} = x_k - x_i$.

$$\cos(\overrightarrow{x_i x_j}, \overrightarrow{x_i x_k}) = \frac{< \overrightarrow{x_i x_j}, \overrightarrow{x_i x_k} >}{||\overrightarrow{x_i x_j}||||\overrightarrow{x_i x_k}||}$$

Variando los datos x_j y x_k, calculando los diferentes valores del coseno ponderado dejando siempre fijo el dato x_i, podemos aproximar la varianza de todo el espectro de ángulos o "angle-based outlier factor" (ABOF). Este valor, es el mínimo en el caso de los outliers (al tener un espectro de ángulos con menor amplitud). Así podemos decidir si clasificar un valor como atípico basándonos en:

$$\mathrm{ABOF}(x_i) = \mathrm{Var}\{\cos(\overrightarrow{x_i x_j}, \overrightarrow{x_i x_k})\}_{j,k \in \{1,\ldots,n\}}$$

Para reducir costes y tiempo de computación, manteniendo un punto fijo x_i, se puede utilizar k-vecinos más próximos para seleccionar k puntos (en lugar de calcular todos los posibles vectores desde el punto x_i) y aproximar el ABOF para ese punto x_i.

Bagplots

Los bagplots, introducidos en P. J. Rousseeuw, Ruts y Tukey (1999) son una generalización de los boxplot a 2 e incluso 3 dimensiones, permitiendo observar gráficamente la localización, variabilidad, asimetría y outliers de un conjunto de datos. En la Figura 2.6 se muestra un bagplot, compuesto de 2 polígonos solapados. El polígono interior, llamado también "bolsa", es construido según la profundidad *Tukey depth*, conteniendo como máximo el 50 % de las observaciones más profundas del conjunto de datos. El polígono exterior, o "valla", es construido multiplicando por un factor (normalmente 3) el polígono interior, donde todos los puntos fuera de la "valla" son marcados como atípicos.

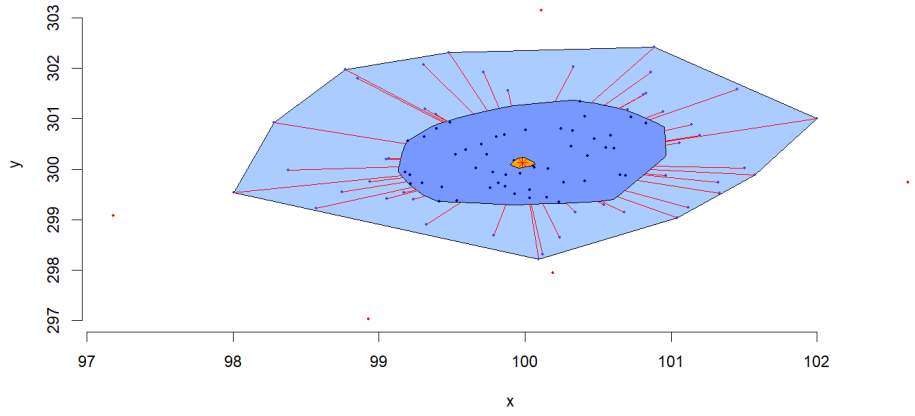


Figura 2.6: Gráfico Bagplot

2.2. Clustering

Un outlier puede verse también como un concepto relativamente complementario al concepto de cluster. Es decir, un outlier es una observación individual, un dato que es distinto de las demás observaciones pero en cambio, un cluster es un grupo o grupos de observaciones que son similares. Por ejemplo, en la Figura 2.7, se representa una distribución bimodal, donde la línea negra representa el límite o corte a partir del cual una observación es considerada atípica. Véase que hay una zona con outliers que no son extremos, entre los dos clusters y por debajo de la línea negra.

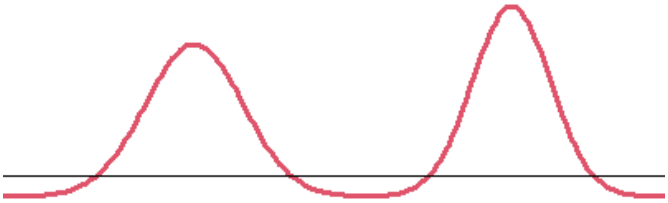


Figura 2.7: Ejemplo de distribución bimodal

Además, en ocasiones los outliers aparecen en pequeños clusters propios. Una posible explicación es que la anomalía causante de generar observaciones atípicas puede haberse repetido varias veces.

Para un conjunto de datos, se realiza una partición en K clusters particionando los índices $\{1, ..., n\}$ en $C_1 \cup ... \cup C_K$ de forma que $C_k \cap C_{k'} = \emptyset$ para $k \neq k$. Podemos tomar ahora

$$\overline{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i$$

$$\mathbf{S}_k = \frac{1}{n_k - 1} \sum_{i \in C_k} (x_i - \overline{x}_k)(x_i - \overline{x}_k)'$$

donde n_k es el número de índices en C_k.

Hay varias opciones a la hora de identificar outliers con métodos clusters. En lugar de utilizar la distancia de Mahalanobis $d(x; \overline{x}, \mathbf{S})$ a nivel global, se pueden usar la distancia $d(x; \overline{x}_k, \mathbf{S}_k)$, con las $\overline{x}_k$ y $\mathbf{S}_k$ definidos anteriormente. De esta forma, las observaciones atípicas son aquellas que cumplen la siguiente expresión:

$$\min_{1 \leq k \leq K} d^2(x; \overline{x}_k, \mathbf{S}_k) \geq \chi^2_{p-1,\alpha} \tag{2.2}$$

2.2.1. Verosimilitud de clasificación

Hay métodos de Análisis Cluster que asumiendo densidades multivariantes de localización μ_k y matrices de dispersión $\mathbf{\Sigma}_k$ en los K componentes que generan los clusters maximizan la verosimilitud de "clasificación"

$$\prod_{k=1}^{K} \prod_{i \in c_k} \phi(x_i; \mu_k, \mathbf{\Sigma}_k)$$

donde $\phi(.; \mu, \mathbf{\Sigma})$ sería la densidad de una normal p-variante con parámetros μ y $\mathbf{\Sigma}$ en todas las posibles particiones $\{1, ..., n\} = C_1 \cup ... \cup C_K$, con $C_k \cap C_{k'} = \emptyset$ para $k \neq k'$, $\mu_k \in \mathbb{R}^p$ y $\mathbf{\Sigma}_k$ matrices semidefinidas positivas y simétricas para $k = 1, ..., K$. Existen algoritmos CEM (Classification EM) que permiten buscar los parámetros $\hat{\mu}_1, ..., \hat{\mu}_K, \hat{\mathbf{\Sigma}}_1, ..., \hat{\mathbf{\Sigma}}_K$ óptimos (junto a la partición óptima).

Dado que las particiones quedan definidas como

$$C_k = \{i : \phi(x_i; \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k) = \max_{1 \leq l \leq K} \phi(x_i; \hat{\mu}_l, \hat{\mathbf{\Sigma}}_l)\},$$

se pueden usar como "scores" para identificar las atipicidades aquellas observaciones con menor valor de

$$s(i) = \underset{1\leq k\leq K}{\text{máx}}\, \phi(x_i; \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k).$$

En el caso de que ϕ se corresponda con la densidad normal p-variante el criterio se asemeja mucho a utilizar la expresión (2.2).

2.2.2. Verosimilitudes de mixtura

Podemos ajustar el análisis cluster para mixturas mediante el algoritmo EM. En este caso, la función de verosimilitud a maximizar se corresponde con:

$$\prod_{i=1}^{n}\sum_{k=1}^{K} p_k\phi(x_i; \mu_k, \mathbf{\Sigma}_k)$$

donde p_k es la probabilidad de que la observación x_i haya sido generada por la componente k-ésima, para $k = 1, ..., K$.

Tras ajustar los parámetros óptimos $\hat{p}_1, ..., \hat{p}_K, \hat{\mu}_1, ..., \hat{\mu}_K$ y $\hat{\mathbf{\Sigma}}_1, ..., \hat{\mathbf{\Sigma}}_K$, el cálculo de los scores $s(i)$ se calcula usando

$$s(i) = \sum_{k=1}^{K} \hat{p}_k\phi(x_i; \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k)$$

siendo atípicas aquellas observaciones con un valor bajo de $s(i)$.

2.2.3. Clustering robusto

Las observaciones extremas pueden afectar negativamente los parámetros y la partición en clusters y, de nuevo, las observaciones atípicas se pueden ocultar. Como solución a este problema, se pueden usar verosimilitudes de clasificación o de mixtura recortadas.

El uso de verosimilitudes recortadas puede verse como una extensión del procedimiento de *Minimum Covariance Determinant* ya visto. Por ejemplo, en el caso de verosimilitudes de clasificación recortadas, se buscan posibles particiones $\{1, ..., n\} = C_0 \cup C_1 \cup ... \cup C_K$, con $C_k \cap C_{k'} = \emptyset$ para $k \neq k'$, donde además se pide que C_0 incluya un número α de índices, que son óptimamente seleccionados. Por supuesto, se buscan también parámetros $\mu_k \in \mathbb{R}^p$ y $\mathbf{\Sigma}_k$ matrices simétricas y semidefinidas positivas que resulten óptimos. Con esos ingredientes, el objetivo es maximizar la verosimilitud recortada

$$\prod_{k=1}^{K}\prod_{i\in C_k} \phi(x_i; \mu_k, \mathbf{\Sigma}_k),$$

donde el producto se realiza para $k = 1, ..., K$ y no para $k = 0, 1, ..., K$. De esta forma no se tienen en cuenta las observaciones con índices en C_0 y, justo, esa proporción α de observaciones con índices en C_0 son las observaciones que consideramos más potencialmente atípicas. Podemos también usar (2.2), basándose en los estimadores estimados robustamente, para etiquetar observaciones atípicas.

Este procedimiento de verosimilitud de clasificación recortada, junto a una serie de restricciones sobre las matrices $\hat{\mathbf{\Sigma}}_1, ..., \hat{\mathbf{\Sigma}}_K$ que no serán descritas, nos llevan al procedimiento TCLUST introducido en García-Escudero et al. 2008. Ese procedimiento usa una modificación del algoritmo CEM, donde una proporción α de observaciones son descartadas en las iteraciones del algoritmo CEM.

La Figura 2.8 muestra el resultado de aplicar la metodología TCLUST para $K = 3$ y $\alpha = 0.1$. Las observaciones en negro corresponden a la fracción del 10 % de observaciones más potencialmente atípicas detectadas por TCLUST.

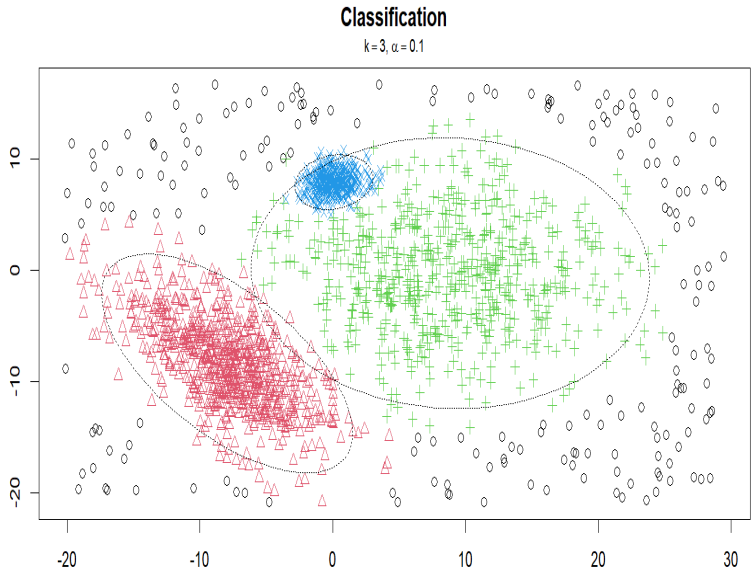


Figura 2.8: Clustering robusto usando TCLUST con $K = 3$ y $\alpha = 0.1$

La verosimilitud de mixturas puede ser también robustificada mediante recortes de una forma parecida. En este caso se usa un algoritmo EM que incorpora eliminación de observaciones.

2.3. Modelos lineales

Normalmente, suele existir una alta correlación entre variables en un conjunto de datos. Las técnicas que vamos a tratar a continuación tratan de explotar estas relaciones usando modelos lineales. Distinguiremos dos posibilidades en este tipo de técnicas: las técnicas de regresión y la técnica de Análisis de Componentes Principales.

2.3.1. Regresión

La idea es que existan algunas variables que nos resulten de especial interés y que sepamos que están relacionadas con otras variables (covariables). Las técnicas de regresión nos permiten saber si los valores observados en estas variables importantes toman valores atípicos atendiendo a los valores que conocemos de las covariables.

En un modelo lineal de regresión, los valores observados son ajustados a un modelo compuesto por un sistema lineal de ecuaciones. Debido a que por lo general no existe una solución exacta al sistema lineal de ecuaciones (sistema sobre-determinado), se busca encontrar aquellos coeficientes que minimizan el error cuadrático medio o MSE de los valores predichos respecto a los valores reales.

Entonces, en nuestro modelo lineal de regresión, los outliers son aquellos valores cuyo valor predicho por el modelo, que depende de las covariables, tiene una mayor desviación que otros puntos respecto a su valor observado. Esta desviación es cuantificable y permite medir la atipicidad de un valor y establecer scores.

Para ajustar el valor y_i por un modelo lineal de regresión de d covariables (regresoras) se emplea la siguiente expresión:

$$y_i = \beta_0 + \sum_{i=1}^{d} \beta_i \cdot x_{ij} + \varepsilon_i,$$

para $i = 1, ..., n$, donde $x_{i1}, ..., x_{id}$ son los valores observados en las covariables y $\beta_0, \beta_1, ..., \beta_d$ son los coeficientes que deben ajustarse en función de los valores observados y ε_i sería un error aleatorio.

Los residuales de la regresión $e_i = y_i - \hat{y}_i$, donde

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_d x_{id},$$

pueden usarse para ver si el valor y_i es atípico (respecto a los valores conocidos en las variables explicativas). Los residuales e_i siguen una distribución aproximadamente normal y pueden utilizarse sus z-values para decidir que y_i pueden considerarse como atípicos. De hecho, se sabe que $E(z_i) = 0$ y $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$ con h_{ii} el elemento i-ésimo de la matriz "hat" $\mathbf{H}$ definida como

$$\mathbf{H} = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X},$$

para $\mathbf{X}$ la matriz de diseño en la regresión (que tiene por columnas los valores de las d variables explicativas y una primera columna con valores iguales a 1). El parámetro σ^2 se estima con el MSE y se usa

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

(residuos estandarizados) para marcar atípicos cuando $|r_i| > z_{\alpha/2}$. De hecho, se suele usar la corrección de Bonferroni si no hay sospecha "a priori" de que la observación y_i es atípica, para corregir que se realizan "comparaciones múltiples", y así se usa $|r_i| > z_{\alpha/2n}$, donde n es el número de observaciones en la regresión.

Es importante notar que los r_i solo siguen una distribución $N(0, 1)$ de forma aproximada. Incluso, tampoco siguen una distribución t_{n-p}, aunque el estimador de σ^2 asociado al MSE tenga $n - p$ grados de libertad. El problema reside en que e_i, que aparece en el numerador de r_i, también aparece al calcular el MSE y no se tiene la independencia que se necesita para tener una distribución t_{n-p}. Por tanto, es preferible comenzar desde los residuales $e_{i,-i}$ obtenidos mediante validación cruzada. De esta forma, se obtienen nuevos estimadores $\hat{\beta}_{-i}$ al ajustar una nueva regresión donde

se han eliminado el elemento y la fila i-ésima y la predicción por validación cruzada $\hat{y}_{i,-i} = \beta_0 + x_i\hat{\beta}_{\mathbf{i}}$. Por tanto, se tendría $e_{i,-i} = y_i - \hat{y}_{i,-i}$, que se puede ver que verifica

$$e_{i,-i} = \frac{e_i}{(1-h_{ii})}$$

y esta relación nos permite definir unos residuales "estudentizados" definidos como

$$t_i = \frac{e_i}{\sqrt{MSE_{-i}(1-h_{ii})}},$$

que ahora sí siguen distribuciones t_{n-p-1} (se pierde un grado de libertad). Siguiendo este criterio, las observaciones se etiquetan como atípicas cuando verifiquen

$$|t_i| > t_{n-p-1,\alpha/2n}.$$

Puede darse el caso de que los valores de las variables predictoras x_i también sean atípicos con *leverages* h_{ii} altos y la observación y_i pueda ocultarse por poder ser de "influencia". Vemos que los leverages h_{ii} situados en la diagonal de la matriz $\mathbf{H}$ juegan un papel muy importante en la detección de atípicos y puntos de influencia porque nos permiten ver lo "rara" que es la observación i-ésima en sus variables explicativas en x_i (de hecho h_{ii} puede reescribirse usando la distancia de Mahalanobis).

Los puntos de influencia pueden ser muchas veces detectados con estos h_{ii}, como medida "potencial" de influencia, o mediante herramientas como los DFFITS, como método "efectivo". No obstante, desgraciadamente, grupos de observaciones atípicas pueden "ocultarse" en herramientas como los DFFITS en un fenómeno conocido como enmascaramiento o "masking". Este fenómeno de enmascaramiento hace muy interesante el uso de la regresión robusta que trataremos a continuación.

Regresión robusta

En la Figura 2.9 se muestra un ejemplo donde un valor de y_i atípico o punto influyente (en azul) provoca una desviación considerable en los valores $\hat{y}_i$ ajustados. La recta roja representa la recta de regresión ajustada según lo desarrollado anteriormente. En cambio, la recta verde representa la recta de regresión robusta ajustada según el estimador robusto LTS que se expondrá a continuación.

Una forma de evitar los puntos influyentes a la hora de ajustar la regresión es utilizar el estimador *Least Trimmed Squares* o LTS (P. Rousseeuw 1984) en lugar de mínimos cuadrados para los coeficientes B_i del modelo de regresión. Este estimador, común para efectuar regresión robusta, y a diferencia de los métodos cuadrados, ordena los residuales de los mínimos cuadrados una vez calculados, tomando los h residuales menores (o lo que es lo mismo, descartando los $n-h$ residuales de mayor valor). En resumen, el LTS se obtiene de la siguiente forma:

$$\min_{\beta} \sum_{i=1}^{h} r^2(\beta)_{(i)},$$

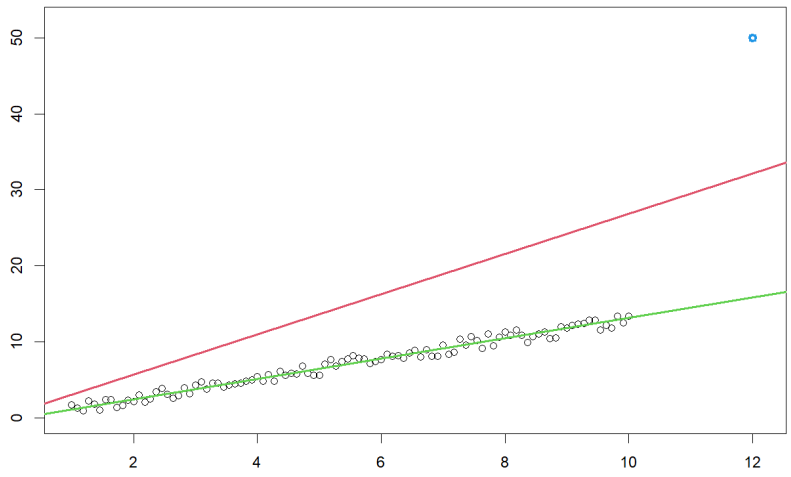


Figura 2.9: Regresión robusta (verde) y no robusta (rojo) con un outlier (azul) influyendo en la estimación de y_i

donde

$$r^2(\beta)_{(1)} \leq ... \leq r^2(\beta)_{(n)}$$

son los residuales al cuadrado ordenados y usar $r^2(\beta)$ hace patente que el residual depende del valor de los parámetros β considerados. El resultado va a depender del valor h que se elija, siendo preferible un valor h alto, de forma que el modelo se ajuste a la mayoría de los datos, descartando únicamente aquellos residuales con un valor notablemente distinto a los demás. Los residuales obtenidos por esta regresión robusta LTS son utilizados para marcar atípicos porque el β óptimo obtenido ya no está afectado por observaciones atípicas o puntos de ruptura, salvo que el nivel de contaminación exceda a $n - h$.

Existen alternativas para el estimador LTS, como los S-estimadores (P. Rousseeuw y Yohai 1984) y los MM-estimadores (Yohai 1987).

2.3.2. Análisis de Componentes Principales (PCA)

En la Sección 2.3.1 suponíamos que había unas variables privilegiadas (respuesta) y covariables (explicativas), pero en ocasiones no existen esas observaciones "privilegiadas" y queremos dar un tratamiento simétrico a todas las variables.

Para un conjunto de datos con p variables y centrado (la media de cada variable o dimensión es 0), realizamos una aproximación del mismo a una dimensión inferior q, tal que $q \ll p$. De esta forma, el Análisis de Componentes Principales indica las q diferentes direcciones (autovectores) donde al proyectar los datos estos quedan mejor representados y se pierda la menor cantidad de información posible. Además, es un método más estable a la presencia de outliers respecto de la regresión.

Las componentes principales están incorreladas entre sí. Aquellas componentes principales o autovectores con un mayor autovalor, son las que más varianza explican sobre los datos, permitiéndonos eliminar aquellos autovectores que "explican" muy poco de la varianza y por ende

reducir la dimensionalidad de los datos. En cambio, no hay un método infalible sobre cuantas componentes mantener, sino que depende de que límite de "varianza explicada" de los datos se elige.

Algunas de las propiedades más características del Análisis de Componentes Principales (ACP) son las siguientes:

- Si los datos son transformados a un nuevo sistema de ejes acorde a los autovectores ortogonales obtenidos por ACP, la varianza de los datos transformados en los nuevos ejes se corresponden con el respectivo autovalor. También las covarianzas muestrales de los datos en esta nueva representación son igual a 0.
- Dado que las varianzas son pequeñas para los datos transformados en los autovectores con autovalores pequeños, desviaciones significantes de los datos transformados respecto a la media en estas direcciones o autovectores representan outliers.

Cálculo de los scores

Sea $\mathbf{X}$ la matriz cuyas n filas vienen dadas por $\{x_1, ..., x_n\} \subset \mathbb{R}^p$ y supongamos que los datos están centrados, de tal forma que las medias de las columnas de $\mathbf{X}$ son todas iguales a 0 (un centrado de las variables es necesario). Las componentes principales se obtienen buscando los autovectores unitarios (norma 1) u_1, ..., u_q de la matriz $\mathbf{X'X}$ asociados a sus autovalores mayores $\lambda_j \geq \lambda_2 \geq ... \geq \lambda_q$ (esta matriz $\mathbf{X'X}$ coincide en dados centrados, salvo el factor $1/n$, con la matriz de varianzas-covarianzas muestral $\mathbf{S}$).

Una vez obtenidos las componentes principales, podemos computar el score normalizado para cada observación x_i tal y como sigue:

$$s(x_i) = \sum_{j=1}^{q} \frac{(x_i' \cdot u_j)^2}{\lambda_j}.$$

Por ejemplo, en la Figura 2.10, se han mantenido $q = 3$ componentes principales, ordenados según sus autovalores de mayor a menor. Aquellos valores que no se alinean con los ejes de los componentes principales pueden considerarse atípicos. Es decir, aquellos valores cuya distancia respecto de la proyección en los ejes de los componentes principales es mayor, pueden sen considerados como residuales, obteniendo un score alto según la fórmula descrita anteriormente.

Análisis de Componentes Principales robusto

Si $\hat{x}_i$ denota la proyección de x_i en el espacio generado por los autovectores $u_1, ..., u_q$, a diferencia de la regresión, el ACP minimiza

$$\sum_{i=1}^{n} ||x_i - \hat{x_i}||^2.$$

Esta minimización, al igual que la regresión, está afectada por observaciones atípicas (por estar basada en mínimos-cuadrados) y dichas observaciones pueden enmascararse.

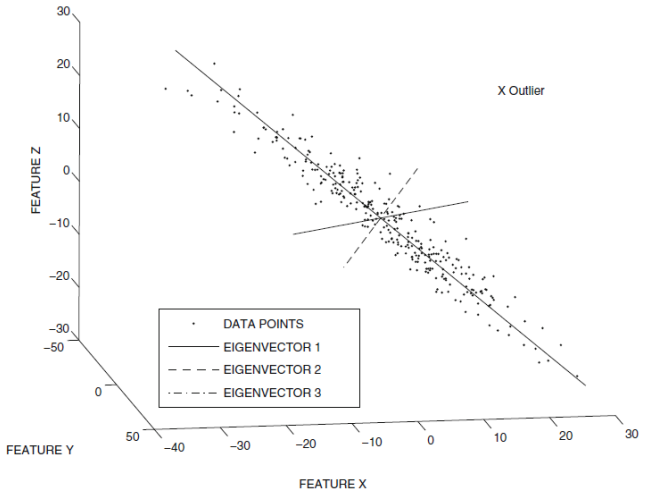


Figura 2.10: Análisis de Componentes Principales con q = 3 (C.C. Aggarwal 2017)

Se han desarrollado versiones robustas de los componentes principales, algunas basadas en sustituir la matriz de covarianzas **S** muestral sobre la que se buscan autovectores y autovalores por un estimador más robusto como el $\hat{\mathbf{\Sigma}}_{\text{MCD}}$ (Croux y Haesbroeck 1999). Alternativamente, el ROBPCA (Hubert, P. Rousseeuw y Branden 2005) está basado en técnicas de búsqueda de proyecciones o "*Projection Pursuit*".

2.3.3. No linealidad

La filosofía de los componentes principales puede generalizarse a extensiones no lineales. Para calcular componentes principales, se necesitan productos escalares $\langle x_i, x_{i'} \rangle$ en la matriz $\mathbf{X'X}$. De forma que, con $x \in \mathbb{R}^p$, si trabajamos con transformaciones no lineales $x \to h(x) \in \mathbb{R}^m$ necesitamos $\langle \Phi(x_i), \Phi(x_{i'}) \rangle$. En estos casos, es recomendable usar el "kernel trick", fijando un núcleo $K(\cdot, \cdot)$ (puede ser polinómico, de bases radiales, ...) tal que $K(x_i, x_{i'}) = \langle \Phi(x_i), \Phi(x_{i'}) \rangle$ sin necesidad de llegar a fijar explícitamente la función $\phi(\cdot)$. Así se gana flexibilidad en la determinación de atípicos abandonando la linealidad.

One Class SVM

La idea que subyace bajo el algoritmo de Support Vector Machine o SVM es encontrar el hiperplano que mejor separe las diferentes clases de un conjunto de datos dentro de un espacio de dimensionalidad alta, a la vez que el margen o distancia entre las diferentes observaciones a ese hiperplano sea máximo, es decir, se deje un mayor margen. En la mayoría de las ocasiones hay que permitir ciertas violaciones de esos márgenes que son penalizados con un parámetro de coste.

En el caso de detección de atipicidades, donde a priori no hay ninguna observación clasificada como outlier, partimos de la asunción de que al inicio todas las observaciones pertenecen a una misma clase "normal" (de ahí SVM One-class). Asimismo, se asume que una función kernel (desconocida) se usa para separar las observaciones originales entre la clase "normal" y una nueva clase que contendrá únicamente los outliers, situándolos en el centro (origen) del nuevo espacio de

mayor dimensión (Schölkopf et al. 1999). El hiperplano será mejor cuanto mayor sea la validez de esta última asunción. Para clasificar una observación, se utiliza la siguiente función de decisión:

$$\text{sign}(\boldsymbol{\beta} \cdot \Phi(x) - \beta_0)$$

donde $\Phi(.)$ es la función desconocida usada para transformar los datos al nuevo espacio de alta dimensionalidad, $\boldsymbol{\beta}$ es el vector de coeficientes para $\Phi(\cdot)$ y β_0 es el sesgo o bias. Un valor positivo indica que la observación sigue perteneciendo a la clase "normal", mientras que un valor negativo indica que la observación pertenece en realidad a la clase de 'outliers'.

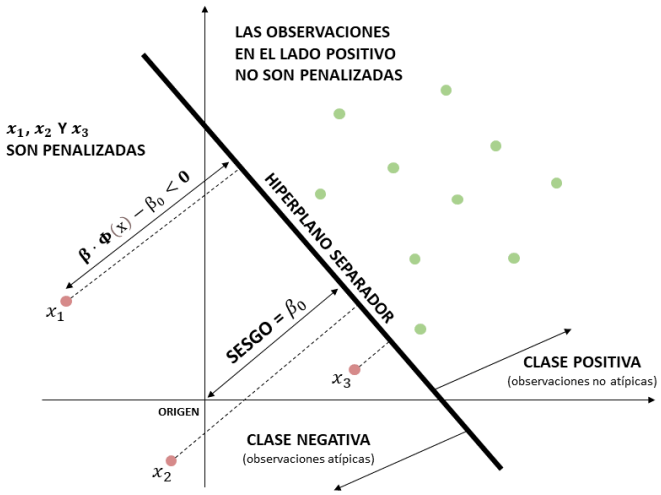


Figura 2.11: Ejemplo de SVM One-class con los outliers en rojo y las demás observaciones en verde

Como se puede ver en el ejemplo de la Figura 2.11, el hiperplano separa el espacio en dos zonas: positiva (para la clase normal), y negativa (para la clase de outliers), donde además se encuentra el origen. El sesgo β_0 controla la distancia del plano al origen. Aquellos puntos cuya función de decisión $\boldsymbol{\beta} \cdot \Phi(x) - \beta_0 < 0$ son clasificados como outliers.

Para sortear el hecho de que la función $\Phi(\cdot)$ es desconocida, a la hora de definir la función objetivo a optimizar (y obtener el hiperplano separador), se utilizan variables de holgura ξ_i con costes (Amer, Goldstein y Abdennadher 2013).

2.4. Basados en proximidad

Los métodos basados en proximidad estudian la relación de proximidad entre las observaciones del conjunto de datos a través de la distancia entre ellas o la densidad local de las observaciones.

2.4.1. Basados en distancias

Los métodos basados en distancias son aquellos que emplean las distancias

$$\{d(x_i, x_{i'})\}_{i,i'=1,\ldots,n}$$

para detectar outliers y construir scores. Entre los algoritmos más populares destacan los k-vecinos más próximos y sus variantes, como la versión ponderada y la versión armónica de los k-vecinos más próximos.

k-vecinos más próximos

Para un conjunto de datos $D = \{x_1, ..., x_n\}$, el score basado en la distancia para una observación cualquiera x_i se corresponde a la distancia a su k i-ésimo vecino más próximo en $D - \{x_i\}$, siendo el score $D_{k(i)} = d(x_i, x_{k(i)})$ con $X_{k(i)}$ el k-vecino más próximo de x_i en D. El parámetro k es elegido por el usuario, donde los scores obtenidos pueden variar significativamente dependiendo de k. Nótese que para el score de la observación X_i, esta misma observación no es incluida en el cálculo de las distancias a los k-vecinos más próximos.

k-vecinos más próximos ponderados

Es muy habitual no conocer el valor exacto de k a utilizar, siendo necesario buscar alternativas más robustas a las posibles variaciones del valor de k.

Para esta problemática muy común en la práctica, es más conveniente promediar los valores obtenidos de las distancias a los k-vecinos más próximos, para un rango de k elegido por el usuario.

$$\frac{d(x_i, x_{1(i)}) + d(x_i, x_{2(i)}) + ... + d(x_i, x_{k(i)})}{k} = \frac{D_1(x_i) + ... + D_k(x_i)}{k}$$

Este promedio puede ser equitativo (media de los k-vecinos más próximos), resultando en unos scores menos sensibles a variaciones en el valor de k.

k-vecinos más próximos armónicos

Puede utilizarse también la media armónica de las distancias de los k-vecinos más próximos.

$$\sqrt[k]{d(x_i, x_{1(i)}) + d(x_i, x_{2(i)}) + ... + d(x_i, x_{k(i)})} = \sqrt[k]{D_1(x_i) + ... + D_k(x_i)}$$

Este cálculo es menos utilizado y se debe utilizar con precaución, eliminando las observaciones repetidas para conseguir unos scores robustos. Además, es recomendable utilizar un k alto, incluso $k = n$ (y no depender más del parámetro k).

2.4.2. Basados en densidades

Los métodos basados en densidades locales van a guardar cierta relación con los métodos descritos anteriormente, ya que si hay una densidad alta para una observación es porque tiene observaciones próximas. Los métodos más empleados se basan en correcciones de la distancia local, como el *Local Outlier Factor* (LOF), histogramas o el uso de métodos de suavizado.

Local Outlier Factor (LOF)

El *Local Outlier Factor* puede interpretarse como un score basado en las variaciones de las distintas densidades locales calculado a través de distancias normalizadas. Para una observación x_i y $D_k(x_i)$ la distancia al k-vecino más próximo de x_i, la distancia de accesibilidad $R_k(x_i, x_j)$ a la observación x_j respecto a la observación x_i se define como:

$$R_k(x_i, x_j) = \text{máx}\{d(x_i, x_j), D_k(x_j)\}.$$

Nótese que ésta distancia de accesibilidad no es simétrica entre x_i y x_j.

Definimos el k-vecindario $L_k(x_i)$ como el conjunto de puntos con distancias menores a la distancia al k-vecino más próximo de x_i. Generalmente, $L_k(x_i)$ contendrá k puntos, aunque puede darse el caso de que contenga más de k puntos debido a empates en la distancia $D_k(x_j)$.

Una vez obtenida la distancia de accesibilidad, se calcula la distancia media de accesibilidad $AR_k(x_i)$ como la media de las distancias de accesibilidad de la observación x_i a las demás observaciones del vecindario $L_k(x_i)$, es decir,

$$AR_k(x_i) = media\{R_k(x_i, x_j) \text{ para todo } x_j \in L_k(x_i)\}.$$

Ahora podemos definir el *Local Outlier Factor* como el ratio medio de las distancias medias de accesibilidad $AR_k(x_i)$ en el k-vecindario de x_i:

$$\text{LOF}_k(x_i) = media\left\{ \frac{AR_k(x_i)}{AR_k(x_j)} \text{ para todo } x_j \in L_k(x_i) \right\}.$$

Si las observaciones están distribuidas homogéneamente dentro del mismo cluster, el LOF para cualquiera de dichas observaciones tendrá un valor menor que 1. Valores mayores de 1 para una observación indica que estamos seguramente ante un atípico.

Histograma

Un histograma es la representación gráfica de la distribución de frecuencias de una variable numérica continua. La variable es típicamente discretizada en rejillas con celdas de igual tamaño entre los valores mínimo y máximo de dicha variable. Así podemos identificar que rango de celdas concentran la mayor frecuencia asociadas a una mayor densidad subyacente. También se identifican aquellas celdas incluyendo una o muy pocas observaciones e identificando a estas observaciones en estas celdas como posibles atípicos.

Un ejemplo de un histograma cualquiera es el mostrado en la Figura 2.12, correspondiente a datos normales que han sido generados aleatoriamente, apreciándose que en los extremos hay una menor densidad y cuyas celdas podrían contener los potenciales atípicos.

Métodos de suavizado

Los métodos de suavizado se basan en una función de densidad estimada $\hat{f}_h(x)$ mediante el uso de funciones núcleo y un parámetro h para controlar el suavizado. Esta estimación o aproximación a la densidad real permite obtener scores donde valores bajos indican atipicidad.

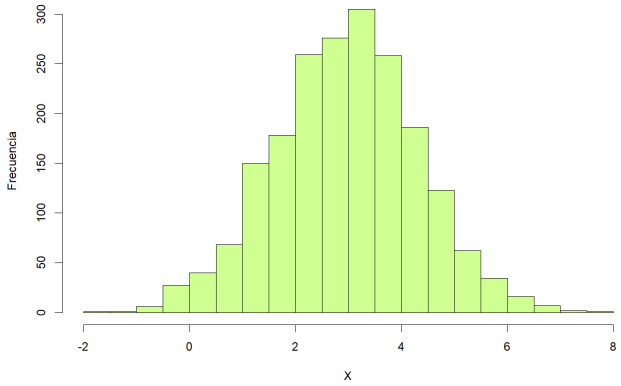


Figura 2.12: Ejemplo de histograma de datos normales generados aleatoriamente

La función de densidad estimada $\hat{f}_h(x)$ suele obtenerse como

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

donde h es el parámetro de suavizado (h mayores proporcionan un suavizado mayor) y $K : \mathbb{R} \to \mathbb{R}^+$ es el núcleo. Así, se marcan como atípicos aquellas observaciones con $s(i) = \hat{f}_h(x_i)$ más pequeño.

2.4.3. Limitaciones

En el peor de los casos, los métodos basados en distancias requieren realizar operaciones del orden de $O(n^2)$. Por ello, dado que no sería factible recurrir a estos métodos cuando n es grande, se han desarrollado técnicas que permitan reducir el efecto de estas limitaciones y minimicen el esfuerzo computacional requerido, como realizar submuestreo o "*early termination*" en lugar de efectuar todos los cálculos para todas las observaciones. (C.C. Aggarwal 2015). siendo ineficiente.

Además, la calidad de los outliers encontrados con estos métodos disminuye cuando la dimensionalidad del conjunto de datos es mayor (maldición de la dimensionalidad). Este problema es bastante conocido en la estimación no-paramétrica de densidades.

Por lo general, los métodos basados en proximidades usan todas las variables disponibles en nuestros datos y, por tanto, un posible ruido o falta de interés de algunas variables utilizadas puede influir negativamente en dichos métodos y ocultar outliers.

Capítulo 3

Ensembles

Algunos algoritmos para la detección de outliers funcionan mejor para ciertos conjuntos de datos, mientras que con otros algoritmos obtenemos mejores resultados para otros. Tanto es así que puede darse el caso de que individualmente cada algoritmo detecte algún tipo particular de anomalía, de tal forma que algunos algoritmos detecten como outlier alguna observación que otros algoritmos no detectan. Un método de agregación o "ensemble" combina varios algoritmos, denominados "detectores base" dentro del ensemble, para obtener una salida unificada. Así se logra un nuevo método (de métodos) más robusto y una detección de outliers generalmente más precisa.

Para construir un ensemble, es clave la elección de los algoritmos que lo forman y la metodología a seguir a la hora de combinar los scores que típicamente están en diferentes escalas. Nótese que algunas técnicas trabajan con scores normalizados, otras con distancias en bruto, algunos algoritmos utilizan scores bajos para clasificar outliers en lugar de scores altos, etc. Esto se debe tener en cuenta de cara al resultado final que proporcione el ensemble.

3.1. Clasificación

Podemos distinguir principalmente dos tipos de ensembles:

- *Centrados en el modelo*: Los algoritmos utilizados o detectores base son diferentes entre sí o se utilizan diferentes configuraciones de parámetros de un mismo algoritmo.
- *Centrados en los datos*: Los algoritmos utilizados o detectores base son los mismos aplicados a variaciones del conjunto de datos. Estas variaciones pueden ser obtenidas a través de un muestreo o remuestreo (por ejemplo, *bootstrap*, añadiendo ruido al conjunto de datos) o ponderado las diferentes observaciones del conjunto de datos.

También podemos clasificar los ensembles atendiendo a la independencia computacional de los algoritmos o detectores base utilizados:

- *Independientes*: Los algoritmos se ejecutan de forma independiente, combinando las salidas al final. Este suele ser el enfoque más común.

- *Secuenciales*: Los algoritmos se ejecutan en secuencia, es decir, la salida de un algoritmo es la entrada para el siguiente algoritmo, creando un modelo más refinado de detección de outliers en cada paso.

Las dos clasificaciones mostradas son compatibles entre sí, pudiendo haber ensembles centrados en el modelo tanto independientes como en secuencia, y lo mismo para ensembles centrados en los datos.

3.2. Combinación de scores

Para poder combinar los scores de los distintos algoritmos de un ensemble, y obtener una salida unificada, debemos normalizar sus valores.

3.2.1. Normalización de scores

Existen dos formas principales para normalizar scores, el escalado y la estandarización:

Escalado

Los scores son escalados al intervalo $[0, 1]$, por ejemplo, de la siguiente forma:

$$S_j(i) = \frac{s_j(i) - \text{mín}_j\{s_j(i)\}}{\text{máx}_j\{s_j(i)\} - \text{mín}_j\{s_j(i)\}},$$

donde $s_j(i)$ corresponde con el score de la observación x_i por el algoritmo j y $S_j(i)$ sería el valor del score escalado. Desgraciadamente, el escalado es muy sensible a los valores extremos (máximo y mínimo) de los scores, por lo que suele utilizarse más la estandarización como recurso de normalización.

Estandarización

Los scores están estandarizados de la siguiente manera:

$$S_j(i) = \frac{s_j(i) - \mu_j}{\sigma_j},$$

donde μ_j y σ_j se corresponden con $\mu_j = media_j\{s_j(i)\}$ y $\sigma_j = \sqrt{\text{Var}_j\{s_j(i)\}}$. A través de la estandarización, obtenemos z-values (aunque no podamos asumir que dichos scores sigan una distribución normal).

3.2.2. Unificación de scores

Una vez los scores están normalizados, se deben combinar de alguna forma para proporcionar un score o resultado final. Las técnicas más comunes son elegir el score máximo de entre todos los scores obtenidos por los algoritmos para la misma observación, o, alternativamente, elegir el score en la mediana o promediar los scores para elaborar el score final. Nótese que con el score máximo se favorece que una observación x_i sea considerada como atípica si alguno de los algoritmos utilizados la detectan como atípica.

3.3. Modelos

La utilización de diferentes métodos estadísticos y su combinación suele derivar en una aleatoriedad en la determinación de las observaciones atípicas. En consecuencia, podemos intentar estimar su MSE o error cuadrático medio. A su vez, podemos descomponer el MSE en dos componentes: el "sesgo" y la "varianza". En función de las técnicas utilizadas para construir ensembles, los métodos se centrarán en reducir uno de los dos componentes: sesgos y varianza.

3.3.1. Basados en la reducción de la Varianza

Los métodos de reducción de varianza se centran en promediar los scores obtenidos de múltiples ejecuciones de un algoritmo aleatorizado (ya sea por muestreo aleatorio del conjunto de datos o de los algoritmos y sus parámetros). Se basa en la idea de que el promedio de un conjunto de variables aleatorias tiene menor variabilidad (cuando no son demasiado dependientes), resultando en una mejora del rendimiento del ensemble.

Bagging

El bagging (bootstrap aggregating) realiza muestreos con reemplazamiento tipo *bootstrap*, obteniendo varias muestras del mismo tamaño que el conjunto de datos inicial. El detector de outliers se aplica a cada nueva muestra, utilizando por tanto un ensemble centrado en los datos. Se obtienen scores diferentes para un mismo punto, que son promediados para obtener el resultado final. Además, utilizando detectores base inestables (es decir, algoritmos que proporcionan scores bastante diversos para las mismas observaciones en cada ejecución), se consigue mejorar aún más el rendimiento general del ensemble porque se promueve la independencia de los scores usados en el promedio.

Rotated Bagging

Es una variante del bagging (C.C Aggarwal y Sathe 2015). Aquí, los datos son rotados según un sistema de ejes aleatorio, para posteriormente seleccionar un conjunto de variables menor que en el conjunto de datos inicial. Esta rotación y posterior reducción de la dimensionalidad permite exponer outliers en algunas de las nuevas proyecciones creadas.

Los pasos del rotated bagging son los siguientes:

1. Determinar un sistema de ejes rotados aleatoriamente.
2. Muestrear $r = 2 + \frac{\sqrt{p}}{2}$ direcciones del sistema de ejes rotados (p se corresponde con el número total de dimensiones) y proyectar los datos en estas r direcciones.
3. Ejecutar el detector de outliers sobre los datos proyectados.
4. Estandarizar los scores obtenidos y combinarlos promediando los resultados.

3.3.2. Basados en la reducción del sesgo

El sesgo o bias puede verse como el error inherente al uso del modelo por la consideración de suposiciones más restrictivas de lo que realmente se requiere. Un ejemplo de técnica basada en la reducción del sesgo es el prunning o podado:

Prunning o podado

Está basado en la eliminación iterativa de outliers. Es decir, entrenar los algoritmos con un subconjunto del conjunto de datos original (conjunto de entrenamiento), localizar los posibles outliers, eliminarlos del conjunto de entrenamiento (si los hubiese), y volver a entrenar los algoritmos repitiendo los pasos anteriores. Se fija un umbral con un score alto para clasificar una observación como outlier y posteriormente eliminarla (técnica conservadora). Este proceso se repite hasta que ninguna observación supera el umbral o hasta alcanzar un número prefijado de iteraciones del proceso.

Asimismo, hay variantes del prunning que utilizan el coeficiente de correlación de Pearson para seleccionar los algoritmos que determinarán que observaciones son outliers, como la versión desarrollada en Rayana y Akoglu 2016.

3.3.3. Combinación de técnicas para reducir varianza y sesgo o bias

En los últimos años se han desarrollado también técnicas de ensembles que permiten aprovechar las ventajas de reducir tanto la varianza como el sesgo sin afectar a la precisión del ensemble y sin requerir de enormes esfuerzos computacionales. Un ejemplo de este propuesto es el Average-of-Maximum (C.C Aggarwal y Sathe 2015).

Average-of-Maximum

Por lo general, usar un promedio de scores contribuye a reducir la variabilidad, mientras que utilizar los scores máximos en la detección de outliers contribuye más a reducir sesgos. El "Average-of-Maximum" divide los m algoritmos del ensemble en m/q cubos con q componentes o algoritmos en cada cubo. De cada cubo se obtienen los scores máximos, y estos son promediados entre los m/q cubos utilizados. Se debe seleccionar un valor q menor que m/q.

Capítulo 4

Detección supervisada de outliers

Hasta ahora, todas las técnicas descritas en esta memoria se han centrado en el problema no supervisado de detección de atipicidades donde no hay información disponible al inicio sobre anomalías ya detectadas en el conjunto de datos. Aportar al conjunto de datos de entrenamiento esta información sobre que observaciones resultaron outliers puede mejorar notablemente la precisión y rendimiento de la detección de outliers. Si bien es verdad que prácticamente las anomalías nos son desconocidas y no disponemos casi nunca de ejemplos, siempre que sea posible se debe incorporar esa información disponible al conjunto de datos, aunque solo sea una parte muy pequeña.

La detección de outliers supervisada es bastante más difícil que un problema general de clasificación supervisada por los retos que se detallan a continuación.

4.1. Etiquetado de clases

Etiquetar manualmente las observaciones es bastante costoso. El aprendizaje activo trata de identificar únicamente aquellas observaciones sospechosas de ser outliers, más difíciles de obtener y más informativas a la hora de entrenar nuestro modelo. De esta forma, se reduce el numero de observaciones necesarias por etiquetar, y por ende, el coste. Además, esta técnica es útil en casos en las que existen pocos ejemplos de outliers con los que empezar a construir el modelo.

Por el general, se sigue un proceso iterativo donde un experto debe etiquetar algunas de las observaciones para entrenar el modelo, aplicarlo al conjunto entero de datos, y mostrar los resultados al experto para que modifique las etiquetas que considere necesarias. Este ciclo se repite hasta que la modificación de etiquetas no influye significativamente en la precisión del modelo, o una vez el límite de coste prefijado de etiquetar iterativamente observaciones es superado (Pelleg y Moore 2004).

El experto deberá etiquetar aquellas observaciones de mayor ambigüedad o incertidumbre, aquellas que estén en el límite de ser consideradas outliers u observaciones normales, o aquellas observaciones que se ajustan poco al modelo. Dichas observaciones son las que mayor información aportan al modelo y contribuyen a maximizar su rendimiento sin necesidad de etiquetar todo el conjunto de datos.

4.2. Clases raras o clases desbalanceadas

Uno de los problemas que presenta el análisis de detección de outliers supervisado es que es sensible al coste de clasificar erróneamente un observación. En ocasiones, es preferible clasificar una observación normal como outlier (falso positivo) a clasificar un outlier como observación normal (falso negativo) y, en otras ocasiones, puede ser justo lo contrario. Por ejemplo, en ocasiones, los outliers pueden ser un grupo muy reducido de observaciones pero de gran importancia práctica en campos como la medicina y detección de tumores, o por ejemplo si se trata de una intrusión en una red informática. En estos casos, puede ser preferible aceptar algún falso positivo y poder garantizar que el número de falsos negativos sea extremadamente pequeño.

En consecuencia, se deben usar otras métricas mas allá de la precisión total (donde el hecho de que el número de atipicidades comúnmente sea muy reducido no refleja adecuadamente la calidad del análisis) y ponderar de forma muy distinta los errores asociados a la clase de outliers respecto a la clase normal. Para ello existen varios tipos de algoritmos expuestos a continuación que tratan de atajar este problema.

4.2.1. Aprendizaje sensible al coste

Tomando como premisa que la clasificación errónea de un outlier puede conllevar un coste bastante elevado, se puede modificar la función objetivo buscando maximizar la precisión "ponderada". Por ejemplo, se puede minimizar la tasa de clasificaciones erróneas ponderadas mediante la minimización de $c_1 n_1 + c_2 n_2$, donde c_1 es el coste de un falso positivo (una observación normal detectada como outlier), n_1 el número de falsos positivos, c_2 el coste de un falso negativo (una observación atípica no detectada como outlier) y n_2 el número de falsos negativos (con generalmente $c_2 \gg c_1$).

Esta modificación o ponderación de las tasas de clasificación erróneas puede ser aplicada con bastantes algoritmos de clasificación como, por ejemplo, el Naive Bayes (NB), la discriminación lineal y cuadrática, los árboles de clasificación, los Random Forests y el SVM.

4.2.2. Remuestreo adaptado

La base del algoritmo es remuestrear los datos para "aumentar" la proporción relativa de la clase rara (en este caso las observaciones atípicas). Este remuestreo puede ser: con o sin reemplazamiento, sobremuestreando la clases rara o outliers o, alternativamente, inframuestreando la clase normal y manteniendo todas las observaciones de la clase rara (también conocido como selección de un lado), o ambas.

4.2.3. Sobremuestreo sintético (SMOTE)

Para cada clase rara o minoritaria, se localizan sus k-vecinos más próximos. En función del nivel de sobremuestreo necesario, una parte de esos vecinos son escogidos aleatoriamente. Después, se genera una observación sintética en el "segmento" que une la observación inicial de la clase minoritaria con su vecino seleccionado. La posición exacta de la observación sintética es escogida según una distribución uniforme aleatoria (Chawla et al. 2002).

4.2.4. Boosting

En este caso, la idea principal es dar mayor importancia a aquellas observaciones dudosas de ser o no ser outliers. El algoritmo da comienzo asignando pesos iguales a cada observación, pero estos pesos son actualizados en cada iteración en función de la precisión obtenida en la ronda anterior, aumentando los pesos de las observaciones mal clasificadas. El resultado final es la media ponderada de las predicciones de todas las iteraciones del algoritmo. Un ejemplo de implementación es el algoritmo Adaboost (Schapire y Singer 1998). Esta metodología del boosting puede ser adaptada al problema de clases raras.

4.3. Clases contaminadas

En muchas ocasiones, la única clase etiquetada son las observaciones "supuestamente normales" pero encontrándose entre esas observaciones "supuestamente normales" también algunas anomalías o outliers no detectadas aún (clase contaminada por valores que deberían ser etiquetados como outliers). También puede verse desde la perspectiva de tener una clase normal con ruido en lugar de una clase negativa etiquetada como tal.

Esencialmente hay dos tipos de métodos para tratar con clases contaminadas. El primero se basa en heurísticas para identificar observaciones "normales" para etiquetarlas como tal, y utilizar estas observaciones "normales" para posteriormente entrenar el modelo junto a las observaciones etiquetadas como outliers. El otro método es el uso de pesos en las observaciones sin etiquetar, siendo una opción menos común.

4.4. Información parcial

Es posible que, debido a la cantidad de datos disponibles o a su dificultad para conseguir una muestra de datos pulida y no contaminada, no dispongamos de observaciones de todas las clases que queremos. En este caso, considerado más bien un problema semi-supervisado, hay partes de los datos que podemos utilizar para entrenar nuestro modelo, mientras que otras partes faltan.

Existen formas de solucionar este problema. Le et al. 2011 ofrece algunas posibles soluciones cuando estamos tratando con información parcial.

Capítulo 5

Código en R

A continuación, se exponen la implementación de algunas de las técnicas presentadas en esta memoria en R en conjuntos de datos de ejemplo junto a los paquetes requeridos para su uso. Este capítulo sigue la estructura de los Capítulos 2 a 4, de forma que el código de R mostrado trata de seguir las técnicas vistas en cada sección.

5.1. Métodos no supervisados

5.1.1. Análisis de valores extremos

R proporciona muchos métodos base para calcular outliers en conjuntos de datos univariantes, sin necesidad de utilizar librerías externas. Para el caso multivariante, en los ejemplos expuestos a continuación, sí ha sido necesario instalar librerías disponibles en CRAN para facilitar la detección de atipicidades.

Caso univariante

Se ha generado aleatoriamente un conjunto de datos normal contaminado como ejemplo para la implementación de los métodos univariantes de "análisis de valores extremos" expuestos en la Sección 2.1.1. Este conjunto de datos contiene dos outliers en los valores $x = 30$ y $x = 300$. El segundo es tan extremo como para afectar notablemente a los estadísticos clásicos de media y desviación típica.

```
set.seed(40)
x <- rnorm(100, mean = 0, sd = 1)
x <- c(x, 30, 300)
```

Se ha tomado $k = 5$ para la desigualdad de Chebychev, de forma que $P(|\frac{X-\mu}{\sigma}| > 5) \leq 0.04$. En cambio, para el cálculo de los scores suponiendo normalidad, se ha tomado $\alpha = 0.001$, de modo que $P(X \notin [\overline{x} \pm 3.29 \cdot s]) \approx 0.001$.

```
estandarizados<-(x-mean(x))/sd(x)
estandarizados[101:102]
```

```
plot(estandarizados)
abline(h=5,col=2,lty=2)
abline(h=qnorm(1-0.001/2),col=3,lty=2)
```

Los p-valores asociados a estos z-scores se calculan con el siguiente código:

```
plot(pnorm(abs(estandarizados), 0, 1, lower.tail = FALSE),ylab="p-values")
abline(h=0.001,col=2,lty=2)
```

La Figura 5.1 muestra los z-scores y las lineas horizontales los valores para marcar observaciones atípicas usando Chebychev en rojo (4 % como máximo fuera) y suponiendo normalidad en verde (0.1 % como máximo fuera). La parte derecha de la Figura 5.1 muestra los p-valores encontrados (suponiendo normalidad).

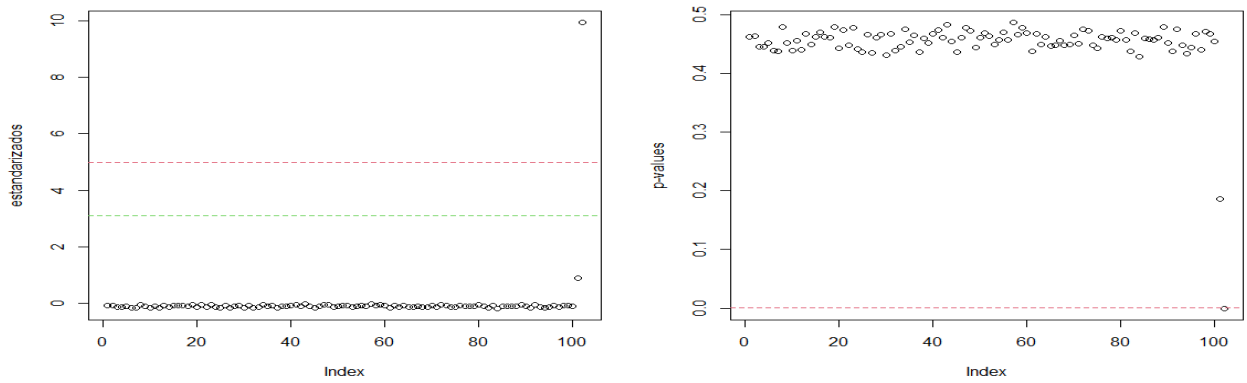


Figura 5.1: En la gráfica de la izquierda se muestran los z-scores usando la media y la desviación típica. Las lineas horizontales marcan atípicos usando Chebychev (rojo) y suponiendo normalidad (rojo). En la parte derecha se muestran los p-valores.

Se ve que la observación $x = 300$ sí es detectada como atípica pero la observación también atípica en $x = 30$ no lo es. El problema es que ese atípico tan extremo en $x = 300$ afecta muy notablemente a la media y la desviación típica muestral y "enmascara" (masking) el otro atípico.

Se ha repetido el cálculo de los z-scores pero utilizando estimadores robustos: la mediana y la desviación absoluta media (MAD) en sustitución de la media y desviación típica, respectivamente.

```
estandarizados_rob<-(x-median(x))/mad(x)
plot(estandarizados_rob)
estandarizados_rob[101:102]
abline(h=5,col=2,lty=2)
abline(h=qnorm(0.999),col=3,lty=2)
plot(pnorm(abs(estandarizados_rob), 0, 1, lower.tail = FALSE),ylab="p-values")
abline(h=0.001,col=2,lty=2)
```

La figura 5.2 muestra que ahora sí que son detectados perfectamente los dos atípicos. Esto muestra la ventaja de considerar estimadores robustos para evitar enmascaramientos.

Por último, basándonos en estos mismos datos univariantes, se ha construido el boxplot. En la Figura 5.3 se detectan 2 observaciones atípicas (en esta figura también se ha realizado "zoom"

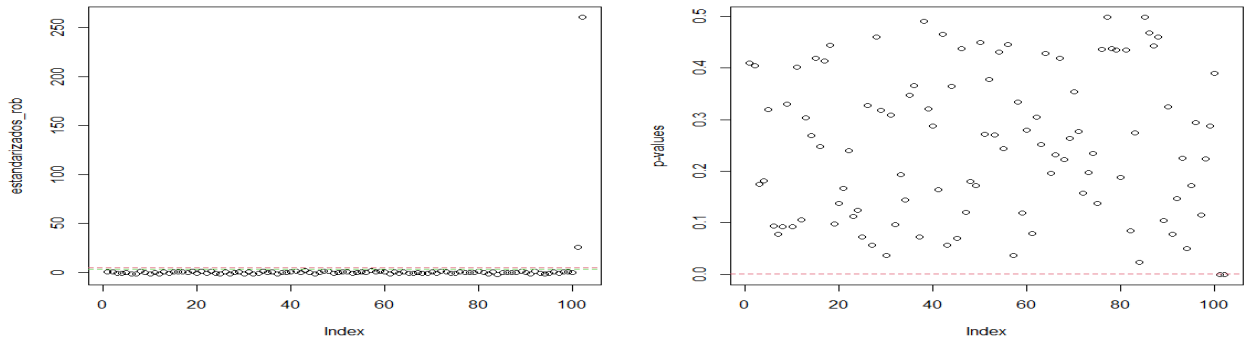


Figura 5.2: z-scores y p-valores (robustos) basados en la mediana y la desviación absoluta media (MAD) en sustitución de la media y desviación típica.

eliminando el outlier más "obvio" de $x = 300$), aquellas observaciones x_i tal que $x_i \notin [Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$.

```
boxplot <- boxplot(x, horizontal = TRUE, col=4, bty="n")
boxplot$out
boxplot(x, horizontal = TRUE, col=4, bty="n",ylim=c(-3,32))
```

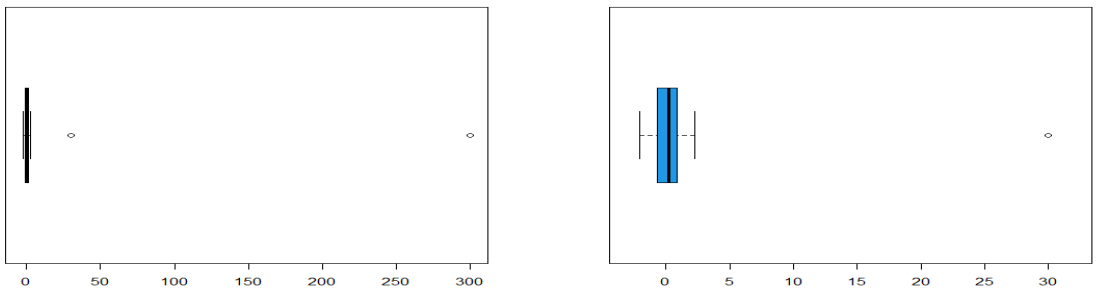


Figura 5.3: Boxplot pora los datos univariantes. A la derecha se realiza "zoom" eliminando el outlier en $x = 300$.

5.1.2. Caso multivariante

Para el caso multivariante, se ha generado un conjunto de datos normales bidimensionales a través de la función `mvrnorm` del paquete `MASS` (Venables y Ripley 2002), que será utilizada para testear implementaciones de "análisis de valores extremos" multivariantes. Se han añadido tres observaciones atípicas en $(-20, 10)$, $(0, 20)$ y $(-20, -20)$. Estas tres atípicas corresponden a las observaciones con índices 1001, 1002 y 1003.

```
library(MASS)
set.seed(50)
mu  <- c(3,3)
cov <- cbind(c(2,1),c(1,2))
X <- mvrnorm(1000, mu, Sigma = cov)
X <- rbind(X, c(-20, 10),c(0, 20),c(-20, -20))
```

Se han calculado los scores basándose en la distancia de Mahalanobis y los estimadores de μ y Σ obtenidos mediante la media muestral $\overline{x}$ y la matriz de covarianzas muestral $\mathbf{S}$. Con estos estimadores se usa $\alpha = 0.001$, se calcula la distancia de Mahalanobis al cuadrado $d^2(X; \overline{x}, \mathbf{S})$ y se marcan las observaciones con distancias de Mahalanobis al cuadrado mayores que $\chi^2_{2,0.999}$.

```
mu_est  <- apply(X,2,mean)
cov_est <- cov(X)
d2<- mahalanobis(X, mu_est, cov_est)
plot(d2,ylab="Mahalanobis al cuadrado")
abline(h=qchisq(0.999,2),col=2,lty=2)
```

El resultado de este código aparece en Figura 5.4, donde vemos que (en este caso) se detectan perfectamente las tres observaciones atípicas añadidas. Posteriormente veremos casos en los que es notablemente mejor alternativas robustas al uso de $\overline{x}$ y $\mathbf{S}$, por ejemplo, basadas en el MCD.

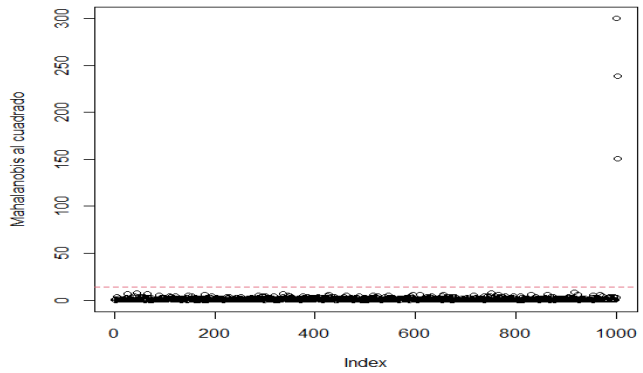


Figura 5.4: Distancias de Mahalanobis para los datos normales contaminados. La linea roja horizontal corresponde al límite $\chi^2_{2,0.999}$.

El paquete `DepthProc` (Kosiorowski y Zawadzki 2022) proporciona bastantes funcionalidades relacionadas con métodos basados en profundidades, entre ellas implementa la técnica *Tukey depth* vista en la Sección 2.1.2. Siguiendo con el conjunto de datos anterior, he aquí el código en R que permite obtener las observaciones atípicas basándose en profundidades según *Tukey depth*, donde se han seleccionado las 3 observaciones con menor profundidad.

```
library(DepthProc)
tukey <- depthTukey(X)
sel <- which(tukey <=sort(tukey)[5])
plot(X,xlab="x1",ylab="x2")
points(X[sel,],col=2,pch=19,cex=1.5)
order(tukey)[1:7]
#  180  656  830  955 1001 1002 1003
```

En la Figura 5.5 se han marcado en rojo las 7 observaciones con menor profundidad. Nótese que se detectan las tres observaciones atípicas añadidas (1001, 1002 y 1003) pero también se han detectado otras que no parecen tan atípicas, como para merecer atención.

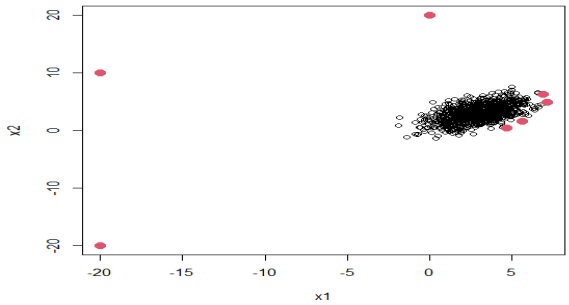


Figura 5.5: Resultado de aplicar el depth de Tukey y marcando en rojo las 7 observaciones con menor profundidad.

El paquete `abodOutlier` (Jiménez 2015) proporciona una implementación sencilla para la detección de outliers basada en ángulos, a través de la función `abod`. A continuación se muestra su uso para el mismo conjunto de datos bivariantes, donde se ha decidido utilizar la opción de k-vecinos más próximos para seleccionar los puntos con los que calcular los ángulos (en lugar de calcular todos los ángulos posibles) y, así, acelerar los cálculos y reducir el tiempo de ejecución del método. Se ve que se han identificado correctamente las 3 observaciones atípicas añadidas.

```
library(abodOutlier)
scores <- abod(X, method="knn", k = 15)
order(scores)[1:3]
### [1] 1001 1002 1003
```

Aunque el uso del depth de Tukey nos detectó algunas observaciones no tan atípicas (Figura 5.5), se ha usado la técnica de "bagplot" con otras nociones de profundidad mediante el paquete **mrfDepth**.

```
library(mrfDepth)
bagplot(compBagplot(X))
bagplot(compBagplot(X, type = "projdepth"))
bagplot(compBagplot(X, type = "sprojdepth"))
```

Los resultados se muestran en la Figura 5.6, donde se ha usado el "halfspace depth" en el gráfico superior izquierdo, "projection depth" en el gráfico superior derecho y una corrección para asimetría en el gráfico inferior. Se ve que que los métodos basados en "projection depth" son útiles detectando los tres atípicos.

5.1.3. Clustering

El paquete **tclust** (Fritz, Garcia-Escudero y Mayo-Iscar 2012) implementa funciones para realizar clustering robusto mediante el uso de TCLUST visto en la Sección 2.2.3. Para el conjunto de datos `swissbank`, incluido en **tclust**, se muestra el uso de la función `tclust` para $k = 2$ grupos. El grupo 0 se corresponde con las observaciones recortadas, o en nuestro caso, las atípicas.

```
library(tclust)
data(swissbank)
```

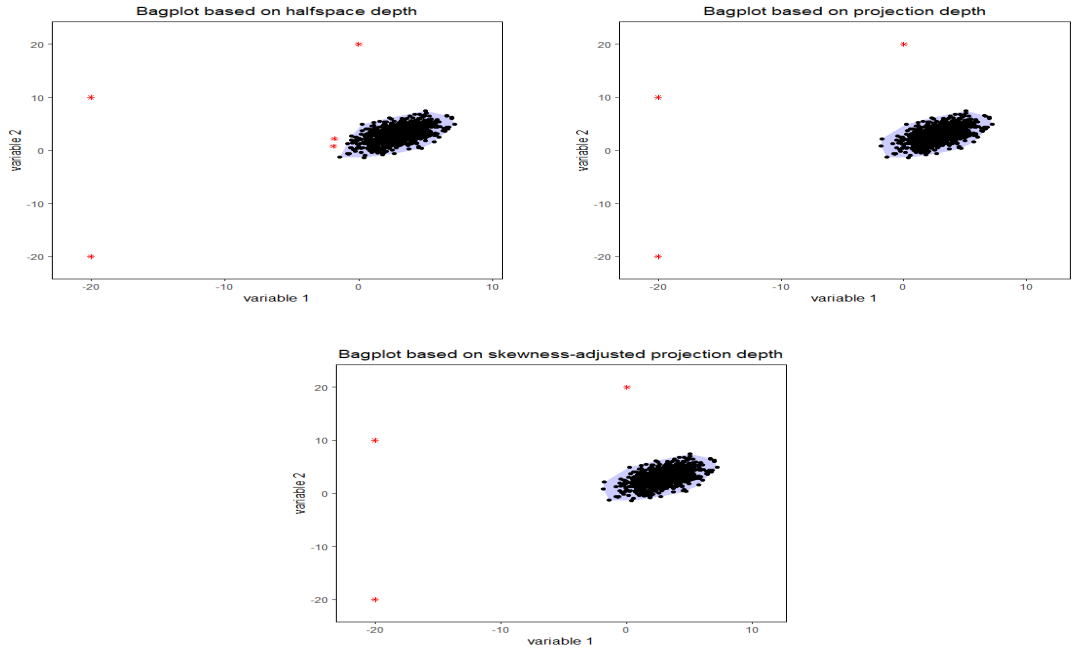


Figura 5.6: Bagplots y detección de atípicos obtenidos mediante el paquete **mrfDepth**.

```
clus <- tclust (swissbank, k = 2, alpha = 0.08, restr.fact = 50)
pairs (swissbank, col = clus$cluster + 1)
plot (swissbank[, 4], swissbank[, 6], col = clus$cluster + 1)
which(clus$cluster == 0)
### [1] 1 40 71 138 148 160 161 162 167 168 171 180 182 187 192 194
```

En la Figura 5.7 se muestra el resultado de `tclust` por pares de variables a la izquierda, con los dos grupos detectados marcados en color rojo y verde, mas las observaciones atípicas "recortadas" en negro, y a la derecha, se ha ampliado el par formado por las variables `IF_Lower` (eje de abscisas) y `Diagonal` (eje de ordenadas), donde se aprecia el cluster formado por atípicos y que pueden corresponder a otra banda de falsificadores.

5.1.4. Modelos lineales

El paquete `robustbase` (Maechler et al. 2023) proporciona métodos para realizar estadística robusta, entre ellos, regresión robusta a través del estimador LTS.

A modo de comparación, empleando el conjunto de datos `starsCYG`, se van a obtener las observaciones atípicas realizando regresión con mínimos cuadrados y regresión robusta LTS. Este conjunto está basado en el diagrama de Hertzsprung-Russell del cúmulo estelar CYG OB1, que contiene 47 estrellas en la dirección de Cygnus. La variable x es el logaritmo de la temperatura efectiva en la superficie de la estrella (`log.Te`) y la variable y es el logaritmo de su intensidad luminosa (`log.light`).

```
library(robustbase)
```

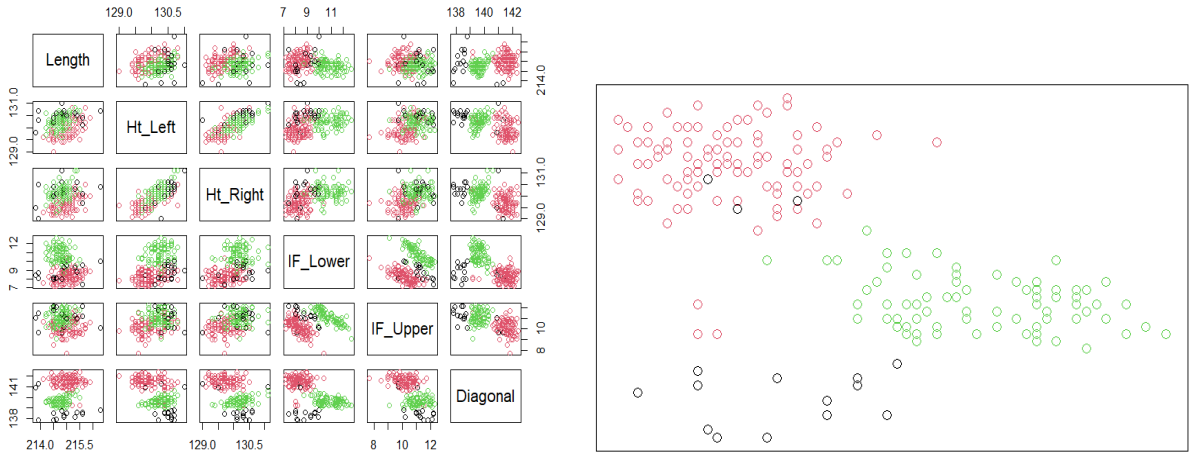


Figura 5.7: TCLUST para `swissbank` con $k = 2$

```
data(starsCYG)
lm.stars <- lm(log.light ~ log.Te, data = starsCYG)
lts.stars <- ltsReg(log.light ~ log.Te, data = starsCYG)
plot(starsCYG, xlab = "", ylab = "")
legend("bottomleft", legend=c("Regresión clas.", "Regresión LTS"), col=c(1,2),
       lwd=rep(2,2), lty=rep(1,2))
points(starsCYG[which(abs(lts.stars$residuals)>3),], col=2, pch=19)
abline(lm.stars$coefficients, lwd=2)
abline(lts.stars$coefficients,lwd=2, col=2)
```

A diferencia de la regresión con estimador de mínimos cuadrados clásica, donde no hay observaciones con un residual alto, con regresión robusta LTS hemos detectado 4 atípicos (cuatro estrellas "gigantes"), que como puede observarse en la Figura 5.8 son claramente puntos de influencia. La recta roja representa la recta obtenida mediante LTS, mientras que la recta negra representa la recta obtenida por mínimos cuadrados que se ve afectada por esos puntos de influencia.

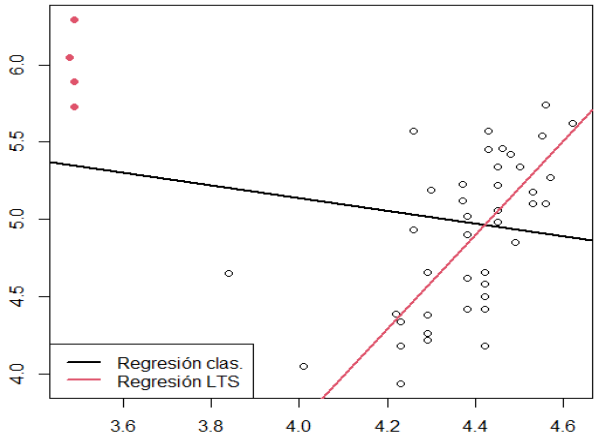


Figura 5.8: Rectas de regresión obtenidas con mínimos cuadrado (negro) y LTS (rojo)

Se pueden dar gráficos de distintos diagnósticos asociados a esta regresión LTS usando el siguiente código:

```
plot(lts.stars,which="rfit")
plot(lts.stars,which="rdiag")
```

El gráfico de la izquierda de la Figura 5.9 muestra el típico plot de residuales, pero usando residuales del LTS debidamente estandarizados. Por otro lado, el gráfico de la derecha de la Figura 5.9 muestra esos mismos residuales del LTS estandarizados en el eje de las y y la distancia robusta usando MCD de las variables explicativas en el eje de las x. Las observaciones más "raras" en las variables explicativas (que pueden ser las de influencia más negativa en regresión o "bad leverage points") tendrán valores altos en ese eje de las x. Este gráfico de la derecha, de esta forma, permite distinguir entre los denominados "good" y "bad leverage points". Otros diagnósticos de influencia, como los h_{ii} y los DFFIT, no son tan útiles como este gráfico para detectar enmascaramiento por "bad leverage points".

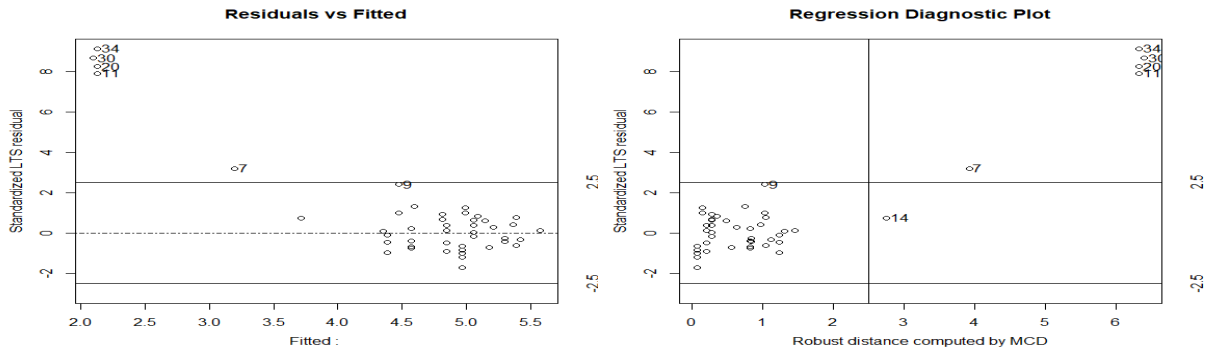


Figura 5.9: Diagnósticos asociados a la regresión LTS útiles para detectar atipicidades y puntos de influencia.

El paquete **mt** (Wanchang 2022) tiene muchas opciones útiles para realizar un análisis de datos, entre las cuales utilizaremos la función `pca.outlier` que, como indica su nombre, se basa en los componentes principales para la detección de atipicidades. Para ilustrar su funcionamiento, usaremos un subconjunto de los datos `Animals` incluidos en le paquete **MASS**. Este conjunto de datos `Animals` incluye el peso medio del cerebro y el peso medio del cuerpo para animales terrestres (donde se incluyen algunos dinosaurios ya extinguidos). Tomando como referencia dicho conjunto, a continuación, se muestra el uso de `pca.outlier` del paquete **mt** en R para un nivel de confianza de $\alpha = 0.05$.

```
data(Animals, package ="MASS")
brain <- Animals[c(1:24, 26:25, 27:28),]
library(mt)
pca <- pca.outlier(log(brain), adj = -0.5, conf.level=0.95, xlim=c(-5, 5))
pca$outlier
pca$plot
#  Dipliodocus Brachiosaurus
#            6             25
```

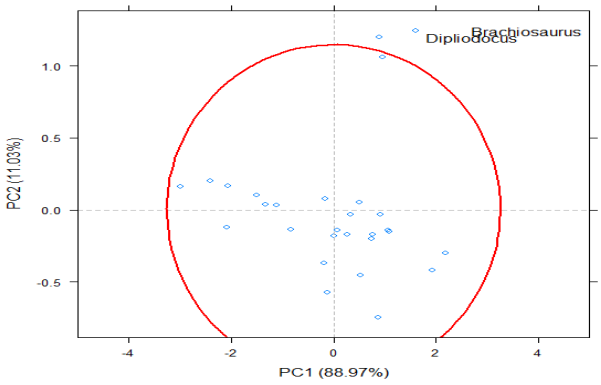


Figura 5.10: Análisis de Componentes Principales sobre `Animals` en escalas logarítmicas.

La Figura 5.10 muestra el resultado del análisis de componentes principales, tomando las dos primeras componentes que recogen (por estar en dimensión $p = 2$) el 100 % de la variabilidad total. El círculo rojo representa los valores límite a partir del cual una observación es considerada atípica, detectándose 2 atípicos, que corresponden a dos dinosaurios (Dipliodocus y Brachiosaurus).

De nuevo, para ver el interés de usar métodos robustos, un análisis más correcto se da en Figura 5.11 y 5.12 con el resultado de aplicar el código:

```
library(robustbase)
mcd <- covMcd(log(brain))
plot(mcd, which = "tolEllipsePlot", classic = TRUE)
plot(mcd, which = "distance", classic = TRUE)
```

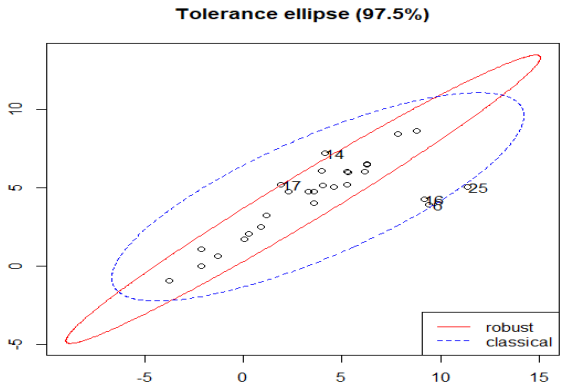


Figura 5.11: Elipses de tolerancia usando estimadores no robustos (azul) y robustos (rojo) basados en el MCD.

En la Figura 5.11 se proporcionan elipses de tolerancia de garantía 97.5 % usando la media $\overline{x}$ y la matriz de covarianza muestra $\mathbf{S}$ marcado en color azul (no-robustos) y el correspondiente al uso de $\hat{\mu}_{\text{MCD}}$ y $\hat{\mathbf{\Sigma}}_{\text{MCD}}$ (robustos) marcado en rojo. Podemos ver que los grandes dinosaurios enmascaran a dos observaciones "14" y "17" también algo atípicas de mucho interés. Estas observaciones observaciones "14" y "17" corresponden al "Human" y al "Rhesus monkey" que destacan por un logaritmo del peso del cerebro mayor que el que les corresponde al logaritmo del peso de su cuerpo.

La Figura 5.12 muestra las distancias de Mahalanobis robustas y no robustas junto a sus valores de corte que proporciona el paquete **robustbase** en R.

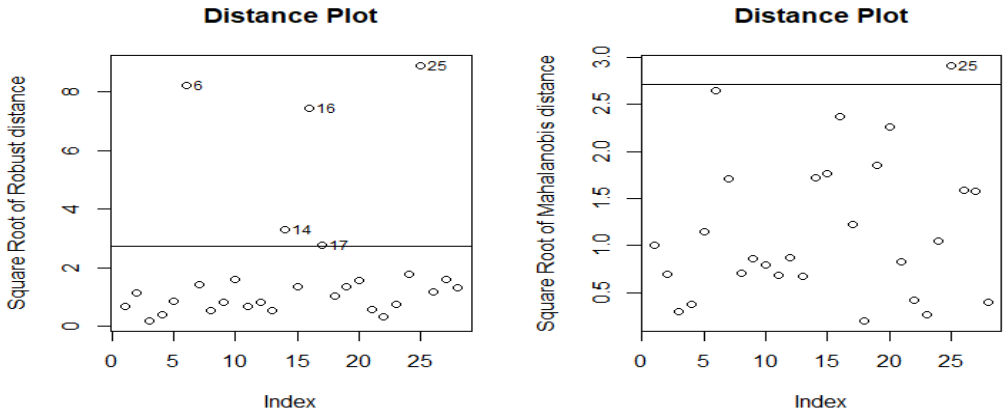


Figura 5.12: Distancias de Mahalanobis robustas y no robustas y puntos de corte proporcionados por el paquete **robustbase**.

Para utilizar "One-Class SVM" en R, el paquete **e1071** (Meyer et al. 2023) nos proporciona varios tipos de SVM, entre los cuales está dispone la opción de "One Classification", que es la que nos permite detectar outliers. El parámetro `nu` nos sirve para indicar el porcentaje de anomalías. En este caso, utilizaremos un subconjunto del conjunto de datos `swissbank`, quedándonos con los billetes falsificados (forged) en los que se sospecha que pueden haber existido varias "fuentes" de contaminación. Existen parámetros como el tipo de núcleo y sus parámetros que deben ser también elegidos correctamente (lo que no es trivial en este problema no-supervisado). Se han usado núcleos de "bases radiales" con $\gamma = 0.1$.

```
swissbank.forged <- swissbank[101:200,]
library(e1071)
x <- swissbank.forged
model <- svm(x, y=NULL, type='one-classification', nu=0.1,
             gamma=0.1, kernel = 'radial')
pred <- fitted(model)
labels <- rep(1,100)
labels[which(!pred)] <- 2
pairs(x,col=labels)
```

La figura 5.13 proporciona un gráfico de variables "por pares" con las observaciones que el SVM marca como atípicas en color rojo.

5.1.5. Basados en proximidad

El paquete `adamethods` (Vinue y Epifanio 2020) proporciona la implementación en R para ciertos algoritmos de detección de atipicidades, como la función `do_kkno` que implementa el algoritmo de k-vecinos más próximos, calculando los scores de atipicidad para cada observación. Cuando mayor sea el score más anómalo es dicha observación.

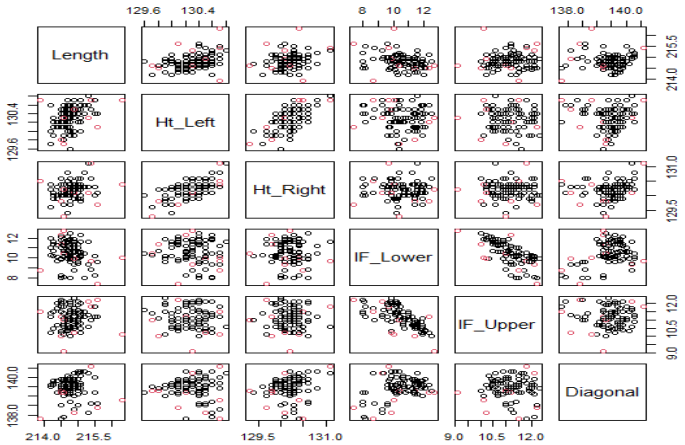


Figura 5.13: Se marcan en color rojo las observaciones que el "One-Class SVM" en **e1071** marca como atípicas para `nu=0.1`.

A modo de ejemplo, disponemos del conjunto de datos `mtcars`, con 32 observaciones de modelos de coches y 11 variables sobre el rendimiento y consumo de un automóvil.

```
library(adamethods)
data(mtcars)
data <- as.matrix(mtcars)
outl <- do_knno(data, 3, 2)
data[outl,]
###                    mpg  cyl disp hp  drat wt   qsec  vs  am gear carb
### Maserati Bora      15.0   8  301 335 3.54 3.57 14.60  0  1    5    8
### Ford Pantera L     15.8   8  351 264 4.22 3.17 14.50  0  1    5    4
### Cadillac Fleetwood 10.4   8  472 205 2.93 5.25 17.98  0  0    3    4
```

El paquete `dbscan` (Hahsler, Piekenbrock y Doran 2019) ofrece varias opciones en R para realizar clustering basado en densidades. Uno de los métodos que proporciona es el cálculo del Local Outlier Factor (LOF) para un conjunto de datos, con el número de vecinos a tener en cuenta como parámetro.

El siguiente ejemplo, muestra su aplicación para el conjunto de datos `mtcars` utilizado anteriormente. Aquellas observaciones con un LOF alto (generalmente LOF > 3) son más atípicas.

```
library(dbscan)
lof <- lof(mtcars, minPts = 3)
summary(lof)
data[lof > 3,]
###               mpg  cyl disp hp  drat wt    qsec vs am gear carb
### AMC Javelin   15.2   8  304 150 3.15 3.435 17.3  0  0    3    2
### Ferrari Dino  19.7   6  145 175 3.62 2.770 15.5  0  1    5    6
### Maserati Bora 15.0   8  301 335 3.54 3.570 14.6  0  1    5    8
```

Según los resultados de ejecutar la función `do_knno`, las observaciones atípicas son "Maserati

Bora", "Ford Pantera L", y "Cadillac Fleetwood". Utilizando LOF, las observaciones atípicas obtenidas, a diferencia de los resultados anteriores, son "AMC Javelin", "Ferrari Dino", y la única en la que coinciden ámbos métodos, "Maserati Bora".

5.2. Ensembles

El paquete `outlierensembles` (Kandanaarachchi 2021) implementa diversas formas de combinación de scores, entre ellas, tomar la media de los scores de cada observación, tomar el score máximo o el cálculo del nuevo score mediante agregación.

A modo de ejemplo, utilizando el conjunto de datos `faithful`, se va a construir un ensemble formado por 5 algoritmos distintos donde se va a utilizar el escalado como método de normalización de los scores calculados por cada algoritmo. El paquete `DDoutliers` (Madsen 2018) implementa varios métodos descritos anteriormente para el cálculo de scores, de los cuales se utilizarán los algoritmos de k-vecinos más próximos (suma y agregación), LOF, LOOP (basado en densidades) y LDOF (basado en distancias).

```
scaled_data <- scale(faithful)
s1 <- DDoutlier::KNN_AGG(scaled_data)
s2 <- DDoutlier::KNN_SUM(scaled_data)
s3 <- DDoutlier::LOF(scaled_data)
s4 <- DDoutlier::LOOP(scaled_data)
s5 <- DDoutlier::LDOF(scaled_data)
scores <- cbind.data.frame(s1, s2, s3, s4, s5)
```

Una vez tenemos los scores, se van a combinar de 3 formas distintas, para posteriormente comparar los resultados y el efecto de cada combinación en la detección de outliers. Se ha decidido combinar los scores en un score final promediando todos los scores (`average_ensemble`), calculando el score final por agregación (`threshold_ensemble`) o calculando el score final mediante un algoritmo voraz (`greedy_ensemble`).

```
ensemble_1 <- average_ensemble(scores)
ensemble_2 <- greedy_ensemble(scores, kk=12)$scores
ensemble_3 <- threshold_ensemble(scores)
select_1 <- order(ensemble_1, decreasing=TRUE)[1:12]
select_2 <- order(ensemble_2, decreasing=TRUE)[1:12]
select_3 <- order(ensemble_3, decreasing=TRUE)[1:12]
```

Se han seleccionado las 12 observaciones con mayor score final como atípicas (aproximadamente el 5 % del número de observaciones totales del conjunto de datos `faithful`).

```
plot(scaled_data, col=3, yaxt="n", xlab="", ylab="", xaxt="n")
points(scaled_data[select_1,1], scaled_data[select_1,2], col=1, pch=2) #(Average)
points(scaled_data[select_2,1], scaled_data[select_2,2], col=2, pch=3) #(Greedy)
points(scaled_data[select_3,1], scaled_data[select_3,2], col=4, pch=4) #(Threshold)
```

En la Figura 5.14 se muestra el conjunto de datos `faithful` con las observaciones no atípicas

en verde. Las observaciones atípicas detectadas mediante el promedio de scores están marcadas con "△". Aquellas observaciones atípicas según el ensemble `greedy` marcadas con "+". Por último, las observaciones atípicas según el ensemble que calcula el score final mediante agregación están representadas con "×".

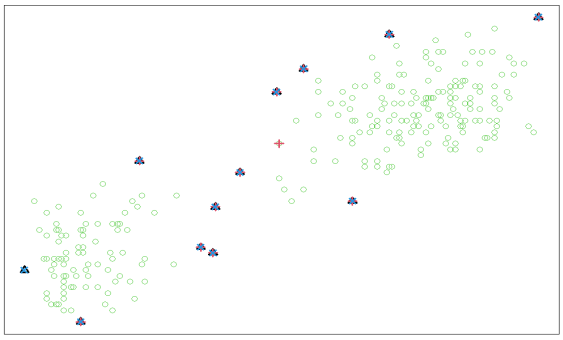


Figura 5.14: Atípicos en `faithful` según `average` (△), `greedy` (+) y `threshold` (×)

A la vista de los resultados mostrados en la Figura 5.14, los 3 ensembles han asignado los scores máximos prácticamente a las mismas observaciones a diferencia de 2: una observación entre los dos clusters únicamente marcada por el ensemble del algoritmo voraz, y otra observación en el extremo del cluster inferior izquierdo, que según el mismo ensemble basado en un algoritmo voraz no es considerada lo suficientemente atípica.

El paquete `bagged.outliertrees` (Santos 2021) implementa un ensemble formado por árboles de decisión haciendo uso del *bagging* para mejorar la robustez y reducir la variabilidad del conjunto de datos.

A través del método homónimo `bagged.outliertrees`, podemos personalizar parámetros del ensemble como número de árboles a considerar, la tasa de remuestreo, o el umbral del score a partir del cual una observación debería ser considerada atípica. Además, como detalle adicional de la librería, se implementa un método que imprime por pantalla los resultados del ensemble de una forma más legible.

A modo de ejemplo, se utilizará el conjunto de datos `hypothyroid` que ofrece la propia librería, que cuenta con 2772 observaciones y 23 variables. Se va a entrenar un ensemble basado en *bagging* formado por 10 árboles.

```
library(bagged.outliertrees)
data(hypothyroid)
ensemble <- bagged.outliertrees(hypothyroid, ntrees = 10,
                    subsampling_rate = 0.5, z_outlier = 3)
outliers <- predict(ensemble, newdata = hypothyroid,
                    min_outlier_score = 0.95, nthreads = 1)
```

De más de 2000 observaciones del conjunto de datos se han detectado entre 100 y 120 observaciones atípicas. Utilizando la orden `print(outliers)` puede obtenerse más información sobre estos atípicos.

5.3. Métodos supervisados

Para ejemplificar un problema de detección de atipicidades supervisado, se va a retomar el segundo conjunto de datos contaminado (multivariante) utilizado durante la Sección 5.1.2, generado aleatoriamente.

Para etiquetar una observación como atípica o no, se van a utilizar los resultados obtenidos del análisis de valores extremos para el caso multivariante, también en la Sección 5.1.2, para posteriormente realizar un análisis supervisado de detección de outliers una vez hemos etiquetado todo el conjunto de datos (véase la Sección 4.1).

```
[...]
d2<- mahalanobis(X, mu_est, cov_est)
class <- pchisq(d2, 2, lower.tail=FALSE) < 0.001
```

Una vez disponemos de las etiquetas para cada observación, y dado que estamos ante un caso de clases balanceadas (62 observaciones etiquetadas como atípicas vs. 938 observaciones etiquetadas como no atípicas), procederemos a utilizar boosting, en concreto, el algoritmo adaboost. Gracias al método `ada` disponible en el paquete `ada` (Culp, Johnson y Michailidis 2016), podemos aplicar el algoritmo adaboost con parámetros configurables como el número de iteraciones (fijadas en 25 en este ejemplo), o el tipo de algoritmo *boosting* a utilizar.

En la Figura 5.15 se muestran marcadas con "△" las observaciones inicialmente etiquetadas como outliers y marcadas con "×"las observaciones identificadas como atípicas por el algoritmo adaboost. A grandes rasgos, la gran mayoría de observaciones identificadas como outliers por adaboost sí estaban etiquetadas como tal (alta sensitividad). De la misma forma, varias observaciones etiquetadas como atípicas han sido identificadas como no atípicas por adaboost (varios falsos negativos).

```
library(ada)
adaboost <- ada(X, class, iter=25)
ada_pred <- predict(adaboost, as.data.frame(X))
plot(X, col=3, yaxt="n", xlab="", ylab="", xaxt="n")
points(X[class,], col=2, lwd=2, pch=2)                  # Etiquetadas
points(X[as.logical(ada_pred),], col=4, lwd=2, pch=4)   # Adaboost
```

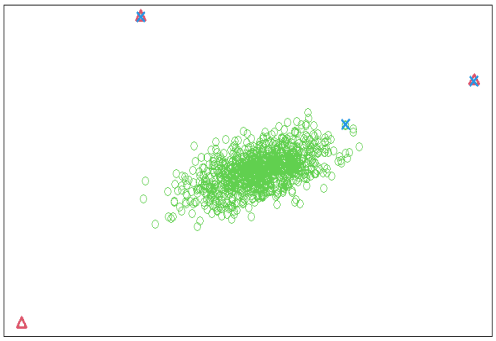


Figura 5.15: Atípicos según las etiquetas iniciales (△) y adaboost (×)

Capítulo 6

Conclusiones y direcciones futuras

La detección de atipicidades es un problema muy relevante, siendo un tema recurrente a la hora de trabajar en Análisis de datos con datos donde las anomalías pueden afectar negativamente a los resultados finales. Además, la detección de outliers tiene un claro interés en sí mismo como, por ejemplo, en campos como la medicina, finanzas, industria, ciberseguridad, climatología u movilidad, entre otros. En este TFG hemos mostrado que es un problema complejo y sin solución única, con muchas técnicas disponibles y muchas más que se van continuamente desarrollando a lo largo del tiempo, cada una con sus características propias y desventajas. Usar técnicas robustas, capaces de resistir bien la contaminación, resulta de gran utilidad para minimizar los efectos de las atipicidades y evitar que dichas atipicidades se "enmascaren" entre las observaciones no atípicas.

Una de las posibles líneas futuras de este trabajo sería profundizar en datos más "especializados". En esta memoria nos hemos centrado en la detección de atipicidades para datos incluyendo únicamente variables numéricas, es decir, tomando valores en $\mathbb{R}^p$. Por tanto, sería interesante recopilar y comprender los métodos de detección de atipicidades para datos categóricos, mixtos, series temporales, series discretas, datos espaciales, grafos y redes, por ejemplo.

Incluso, para problemas de datos tomando valores en $\mathbb{R}^p$, es importante notar que cada vez es más frecuente tratar problemas donde la dimensión del espacio p es elevada. Aunque alguna solución a esta problemática ha sido presentado en la memoria, tiene sentido pensar en técnicas robustas que no eliminen toda una fila de la matriz de datos x_i porque $x_i = (x_{i1}, ..., x_{ip})'$ incluya alguna "celda" x_{ij} atípica. Esto puede ser muy extremo porque se elimina toda la información del resto de las celdas en x_i que no fueron atípicas. Esto es especialmente preocupante cuando p sea elevado, donde es difícil pensar que ningún x_{ij} de un número p grande de celdas en x_i puede ser atípico. En esta situación es preferible eliminar (o detectar como atípicas) solo las celdas realmente atípicas en x_i y dejar toda la información de las otras celdas no atípicas.

Dentro de los paquetes contenidos en CRAN (Comprehensive R Archive Network) hay un gran número de librerías que implementan muchos de los métodos considerados en esta memoria. En el Capítulo 5 se han expuesto algunos ejemplos de aplicación de la mayoría de métodos a diversos conjuntos de datos, existiendo otras muchas implementaciones que hacen uso de otras librerías y que no se han incluido en este trabajo.

Bibliografía

Aggarwal, C.C y S. Sathe (2015). «Theoretical foundations and algorithms for outlier ensembles». En: *ACM sigkdd explorations newsletter* 17, págs. 24-47.

Aggarwal, C.C. (2015). *Data Mining*. Vol. First Edition. Springer.

– (2017). *Outlier Analysis*. Vol. Second Edition. Springer.

Amer, M., M. Goldstein y S. Abdennadher (2013). «Enhancing One-Class Support Vector Machines for Unsupervised Anomaly Detection». En: *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*. New York, NY, USA: Association for Computing Machinery.

Chawla, N.V. et al. (2002). «SMOTE: synthetic minority over-sampling technique». En: *Journal of Artificial Intelligence Research* 16, págs. 321-357.

Crespo Garay, C. (2021). *Por qué es tan peligroso el aumento de las olas de calor en España*. URL: https://www.nationalgeographic.es/medio-ambiente/2021/08/por-que-es-tan-peligroso-el-aumento-de-las-olas-de-calor-en-espana. Último acceso: 05-05-2023.

Croux, C. y G. Haesbroeck (1999). «Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies». En: *Biometrika* 87, págs. 603-618.

Culp, M., K. Johnson y G. Michailidis (2016). *ada: The R Package Ada for Stochastic Boosting*. R package version 2.0-5. URL: https://CRAN.R-project.org/package=ada.

Deliang, D. e Y. Liang (2021). «High-Dimensional Mahalanobis Distances of Complex Random Vectors». En: *Mathematics* 9, págs. 1-12.

Fritz, H., L.A. Garcia-Escudero y A. Mayo-Iscar (2012). «tclust: An R Package for a Trimming Approach to Cluster Analysis». En: *Journal of Statistical Software* 47, págs. 1-26. URL: https://cran.r-project.org/web/packages/tclust/index.html.

García-Escudero, L.A. et al. (2008). «A general trimming approach to robust Cluster Analysis». En: *The Annals of Statistics* 36, págs. 1324-1345.

Gaspar, J. et al. (2011). «A Systematic Review of Outliers Detection Techniques in Medical Data.» En: *HEALTHINF*, págs. 575-582.

Gómez, M. V. (2023). *La Fiscalía Europea abrió investigaciones por un volumen de fraude cercano a los 10.000 millones en 2022*. URL: https://elpais.com/economia/2023-03-01/la-fiscalia-europea-abrio-investigaciones-por-un-volumen-de-fraude-cercano-a-los-10000-millones-en-2022.html. Último acceso: 11-03-2023.

Hahsler, M., M. Piekenbrock y D. Doran (2019). «dbscan: Fast Density-Based Clustering with R». En: *Journal of Statistical Software* 91, págs. 1-30. URL: https://cran.r-project.org/web/packages/dbscan/.

Hawkins, D.M. (1980). *Identification of outliers*. Springer.

Hubert, M., P. Rousseeuw y K.V. Branden (2005). «ROBPCA: A new approach to robust principal component analysis». En: *Technometrics* 47, págs. 64-79.

Jiménez, J. (2015). *abodOutlier: Angle-Based Outlier Detection*. R package version 0.1. URL: https://CRAN.R-project.org/package=abodOutlier.

Kandanaarachchi, S. (2021). *outlierensembles: A Collection of Outlier Ensemble Algorithms*. R package version 0.1.0. URL: https://CRAN.R-project.org/package=outlierensembles.

Kosiorowski, D. y Z. Zawadzki (2022). *DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena*. URL: arXiv:1408.4542.

Le, T. et al. (2011). «A novel parameter refinement approach to one class support vector machine». En: *Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13-17, 2011, Proceedings, Part II 18*. Springer, págs. 529-536.

Madsen, J.H. (2018). *DDoutlier: Distance & Density-Based Outlier Detection*. R package version 0.1.0. URL: https://CRAN.R-project.org/package=DDoutlier.

Maechler, M. et al. (2023). *robustbase: Basic Robust Statistics*. URL: https://cran.r-project.org/web/packages/robustbase/index.html.

Meyer, D. et al. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-13. URL: https://CRAN.R-project.org/package=e1071.

Pelleg, D. y A. Moore (2004). «Active Learning for Anomaly and Rare-Category Detection». En: *Advances in Neural Information Processing Systems*. Ed. por L. Saul, Y. Weiss y L. Bottou. Vol. 17. URL: https://proceedings.neurips.cc/paper_files/paper/2004/file/8c59fd6fbe0e9793ec2b27971221cace-Paper.pdf.

Rayana, S. y L. Akoglu (2016). «Less is more: Building selective anomaly ensembles». En: *ACM transactions on knowledge discovery from data (tkdd)* 10, págs. 1-33.

Rousseeuw, P. (1984). «Least Median of Squares Regression». En: *Journal of the American Statistical Association* 79, págs. 871-880.

Rousseeuw, P. y C. Croux (1993). «Alternatives to Median Absolute Deviation». En: *Journal of the American Statistical Association* 88, págs. 1273-1283.

Rousseeuw, P. y V. Yohai (1984). «Robust Regression by Means of S-Estimators». En: *Springer Lecture Notes in Statistics* 26, págs. 256-272.

Rousseeuw, P. J., I. Ruts y J. W. Tukey (1999). «The Bagplot: A Bivariate Boxplot». En: *The American Statistician* 53, págs. 382-387.

Santos, R. (2021). *bagged.outliertrees: Robust Explainable Outlier Detection Based on OutlierTree*. R package version 1.0.0. URL: https://CRAN.R-project.org/package=bagged.outliertrees.

Schapire, R.E. e Y. Singer (1998). «Improved Boosting Algorithms Using Confidence - rated Predictions». En: *Machine Learning* 37, págs. 297-336.

Schölkopf, B. et al. (1999). «Support vector method for novelty detection». En: *Advances in Neural Information Processing Systems* 12.

Venables, W.N. y B.D. Ripley (2002). *Modern Applied Statistics with S*. New York: Springer.

Vinue, G. e I. Epifanio (2020). *adamethods: Archetypoid Algorithms and Anomaly Detection*. R package version 1.2.1. URL: https://CRAN.R-project.org/package=adamethods.

Wanchang, L. (2022). *mt: Metabolomics Data Analysis Toolbox*. R package version 2.0-1.19. URL: https://CRAN.R-project.org/package=mt.

Yohai, V. (1987). «High Breakdown-Point and High Efficiency Robust Estimates for Regression». En: *The Annals of Statistics* 15, págs. 642-656.