Trabajo Fin de Máster

Medidas de profundidad y geometría

Carlos Mena López

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Medidas de profundidade e xeometría

Título en español: Medidas de profundidad y geometría

English title: Depth measures and geometry

Modalidad: Modalidad A

Autor/a: Carlos Mena López, Universidad de Santiago de Compostela

Director/a: Alberto Rodríguez Casal, Universidad de Santiago de Compostela; Beatriz Pateiro López, Universidad de Santiago de Compostela

Breve resumen del trabajo:

En el contexto unidimensional, podemos entender la mediana como el punto "más profundo" de una distribución, donde al hablar de profundidad nos referimos a una medida del grado de centralidad de un punto con respecto a la distribución. De este modo, a medida que nos alejamos de la mediana, la profundidad de los puntos iría disminuyendo. La generalización de esta idea en el caso multidimensional no es inmediata pues, entre otras cosas, requeriría de una definición multivariante de mediana. Han surgido así en la literatura diferentes definiciones de medidas de profundidad en el contexto multidimensional, en las que confluyen aspectos de geometría y estadística. El objetivo de este trabajo el alumno o la alumna estudie el concepto de medida de profundidad, analice cuáles son las propiedades deseables de una medida de profundidad y haga una revisión de las principales propuestas en la literatura.

Don/doña Alberto Rodríguez Casal, Profesor/a Titular de Universidad de la Universidad de Santiago de Compostela y don/doña Beatriz Pateiro López, Profesor/a Titular de Universidad de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Medidas de profundidad y geometría

fue realizado bajo su dirección por don/doña Carlos Mena López para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Las Palmas de Gran Canaria, a 04 de julio de 2023.

RODRIGUEZ CASAL ALBERTO - 34896829X
Firmado digitalmente por RODRIGUEZ CASAL ALBERTO - 34896829X
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-34896829X, givenName=ALBERTO, sn=RODRIGUEZ CASAL, cn=RODRIGUEZ CASAL ALBERTO - 34896829X
Fecha: 2023.07.06 13:52:23 +02'00'

El/la director/a:
Don/doña Alberto Rodríguez Casal

PATEIRO LOPEZ BEATRIZ - 46898344X
Firmado digitalmente por PATEIRO LOPEZ BEATRIZ - 46898344X
Fecha: 2023.07.05 13:59:51 +02'00'

El/la director/a:
Don/doña Beatriz Pateiro López

MENA LOPEZ CARLOS - 54134050P
Firmado digitalmente por MENA LOPEZ CARLOS - 54134050P
Fecha: 2023.07.06 13:02:21 +01'00'

El/la autor/a:
Don/doña Carlos Mena López

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Capítulo 1

Introducción

Las medidas de profundidad multivariante nacen a finales del siglo pasado como una herramienta no paramétrica útil en la inferencia con datos multivariantes. Concretamente, buscan ofrecer una medida cuantitativa del grado de centralidad de un punto con respecto a una cierta distribución de probabilidad o conjunto de datos. De esta forma, se puede tratar el punto o conjunto de puntos más profundos como una medida de posición central que se propone como un equivalente multidimensional de la mediana.

A partir de esta región central, la profundidad disminuye en todas direcciones en el espacio, permitiéndonos generar divisiones del espacio basadas en su nivel de profundidad, creando las conocidas como regiones centrales o de profundidad. Así, podemos utilizar estos conjuntos para clasificar datos y obtener una medida de orden en el espacio multivariante.

Comparando con el contexto unidimensional, se podría ver como una generalización del concepto de los cuantiles de una distribución. Si bien la interpretación del orden en un espacio de dimensión 2 o mayor no es tan inmediata como el orden natural en una dimensión, en numerosas ocasiones resulta útil tener una noción del mismo basada en el grado de centralidad.

Por este motivo se desarrollaron y siguen desarrollando las conocidas funciones de profundidad, que al aplicar a un conjunto de datos nos ofrecerían una medida cuantitativa del grado de profundidad de una observación. Posteriormente se adaptaron también algunas de estas funciones para trabajar con respecto a distribuciones de probabilidad paramétricas, expandiendo el enfoque no paramétrico inicial de la metodología.

La profundidad rápidamente se aplicó a diversos campos de la estadística, como análisis exploratorio de datos multivariantes y detección de datos atípicos con la creación de herramientas como el bagplot. También se aplicó las medidas de profundidad para crear el DD-plot, una herramienta gráfica utilizada para contrastes de hipótesis que posteriormente serviría como la base para el desarrollo del DD-classifier, una extensión del DD-plot utilizada en problemas de clasificación. Finalmente, se aplicaría al ámbito de la regresión, en la forma de la regresión profunda, un método alternativo para la regresión lineal con propiedades interesantes de robustez frente a outliers.

Si bien tras su aparición en la década de los 70 se postuló como una metodología interesante y con aplicación en diversos ámbitos de la estadística, la profundidad no se generalizó como una herramienta de uso común debido a la presencia de un gran inconveniente. En general las distintas medidas de profundidad propuestas en la literatura eran computacionalmente muy costosas por lo que su uso en espacios de alta dimensionalidad era impensable, e incluso en dimensiones menores se presentaba problemático ante conjuntos de datos de gran y no tan gran tamaño.

Es por estos motivos que no ha sido hasta las últimas décadas cuando han experimentado un resurgimiento en su popularidad. En gran parte, debido a que numerosos avances en informática, principalmente en la capacidad de procesamiento, han permitido reducir notablemente los tiempos de cálculo, pero también gracias a que en la literatura posterior se han desarrollado algoritmos alternativos para el cálculo de las funciones de profundidad que se demostraron más eficientes que las propuestas originales. Actualmente ya contamos con herramientas razonables en el espacio bidimensional, además

de importantes avances en 3 y 4 dimensiones, pero dimensiones mayores siguen siendo un frente abierto que explorar en busca de soluciones satisfactorias.

A lo largo del presente documento, en el Capítulo 2 se postularán las características básicas y propiedades generales de una función de profundidad y de las regiones centrales o de profundidad que se generan a raíz de su aplicación. A continuación, en el Capítulo 3 se explorarán algunas de las medidas más populares formuladas en la literatura, donde se presentará una definición de las mismas, se estudiará el cumplimiento de las propiedades enunciadas en el capítulo anterior y se desarrollará métodos de cálculo con respecto a conjuntos de datos y, cuando sea posible, distribuciones de probabilidad predefinidas.

Continuaremos en los Capítulos 4 y 5, donde se realizará una revisión a algunos de los algoritmos alternativos de cálculo de las distintas medidas de profundidad presentadas en el Capítulo 2 y un estudio de simulación para estudiar el desempeño en distintas situaciones de una de las medidas.

Finalmente, en el Capítulo 6 concluiremos presentando algunas de las aplicaciones que se han desarrollado con las medidas de profundidad, además de ilustrando su funcionamiento con su aplicación a conjuntos de datos reales.

Capítulo 2

Conceptos generales

A lo largo de este segundo capítulo, se ofrecerá una definición formal del concepto de función de profundidad y se postularán una serie de propiedades que deben cumplir para resultar de utilidad en su uso. Además, se procederá similarmente para un concepto derivado de las mismas, las regiones centrales o de profundidad, que no son más que regiones del espacio bidimensional definidos por su profundidad respecto a la medida de probabilidad subyacente.

2.1. Propiedades

En uno de los artículos más relevantes de esta metodología, Zuo y Serfling (2000) definen de forma simple las funciones de profundidad como: *"cualquier función D que ofrezca una ordenación de puntos $x \in \mathbb{R}^d$ basada en P del centro hacia el exterior"*, siendo P una distribución de probabilidad.

Con respecto a su definición formal, podemos, por ejemplo, encontrar una en Cascos et al. (2011), quienes postulan que, para una distribución de probabilidad P perteneciente a $\mathbb{R}^d$, una función de profundidad D, es una función acotada, $D(\cdot; P) : \mathbb{R}^d \mapsto [0, 1]$, que asigna a cada punto en $\mathbb{R}^d$ un valor en función de su grado de centralidad respecto a P.

En su artículo, Zuo y Serfling (2000) formulan una serie de propiedades deseables que se exigen a una función cualquiera para considerarla función de profundidad. Concretamente, imponen condiciones de invarianza afín, profundidad máxima en el centro, monotonía respecto del punto más profundo y que se anule en el infinito.

Dichas propiedades se pueden formular formalmente de la siguiente forma:

P.1 *"Invarianza Afín"*: La profundidad de un punto con respecto a una distribución debe ser invariante frente a cambios en localización o escala. Dado un punto $x \in \mathbb{R}^d$ y una transformación afín de la forma $T(x) = Ax + b$, siendo A una matriz invertible de dimensiones $d \times d$ y b un vector d-dimensional, entonces $D(T(x); T(P)) = D(x; P)$, siendo P la distribución de probabilidad original y $T(P)$ la distribución tras aplicar la transformación.

P.2 *"Profundidad máxima en el centro"*: Para cualquier distribución P que tenga un centro preestablecido, como pueda ser el centro de simetría en distribuciones simétricas, este debe ser el de mayor profundidad. Sea θ el centro de la distribución P, $D(\theta; P) = \sup_{x \in \mathbb{R}^d} D(x; P)$.

P.3 *"Monotonía respecto del punto más profundo"*: A medida que el punto $x \in \mathbb{R}^d$ se aleja del punto más profundo en cualquier dirección fija, su profundidad debe decrecer de forma monótona. Para cualquier P con un punto de máxima profundidad θ, $D(x; P) \leq D(\theta + \lambda(x - \theta); P)$ con $\lambda \in [0, 1]$.

P.4 *"Se anula en el infinito"*: La profundidad de un punto cualquiera x perteneciente al espacio euclídeo d-dimensional $\mathbb{R}^d$ debe tender a 0 a medida que su módulo tienda a infinito.

$\lim_{||x||\to\infty} D(x;P) = 0$, siendo $||x||$ la norma del vector x, $||x|| = \sqrt{\sum_{i=1}^{d} x_i^2}$.

Además de las 4 propiedades propuestas por Zuo y Serfling (2000), otros autores exigen propiedades adicionales a las funciones de profundidad:

P.3′ *"Cuasiconcavidad"*: Dados dos puntos $\{x, y \in \mathbb{R}^d : x, y \in D_\alpha\}$, cualquier punto $z \in \overline{xy}$, siendo $\overline{xy}$ el segmento lineal que une ambos puntos, $\overline{xy} = \{z \in \mathbb{R}^d : z = \lambda x + (1-\lambda)y\}, \lambda \in [0,1]$, también tiene un nivel de profundidad de al menos α. Formalmente, $D(z;P) \geq \text{mín}\{D(x;P), D(y;P)\}$ para cualesquiera $x, y \in \mathbb{R}^d$ y $\lambda \in [0,1]$.

P.5 *"Semicontinuidad superior"*: El conjunto de puntos con profundidad igual o superior a α, $D_\alpha(P) = \{x \in \mathbb{R}^d; D(x;P) \geq \alpha\}$ es cerrado para cualquier α.

La propiedad ***P.1*** no es más exigir que la profundidad de un punto cualquiera dependa de su posición relativa al conjunto o distribución, no del sistema de coordenadas utilizado ni la escala de medida. Por su parte, ***P.2*** pretende asegurar que las funciones de profundidad respeten los centros predefinidos en caso de que estos existan. Algunos autores, como Mosler (2013), no exigen este postulado, sino que la consideran una consecuencia del resto de propiedades.

Por su parte la propiedades ***P.3*** $-$ ***P.4*** aseguran por una parte que los conjuntos D_α sean acotados y estrellados respecto al punto de máxima profundidad y por otra que la profundidad en dichas regiones decrezca a medida que nos alejamos del centro o punto más profundo.

Definición 2.1.1 (Conjunto estrellado)**.** *Un conjunto S se considera estrellado o con forma de estrella si existe $\theta \in S$ tal que para todo $x \in S$ el segmento lineal que une ambos puntos está contenido a su vez por el conjunto. $\overline{\theta x} \subseteq S, \forall x \in S$.*

La propiedad ***P.3′*** asegura que las regiones de profundidad sean conjuntos convexos. Es una versión más restrictiva de la propiedad ***P.3***, dado que los conjuntos convexos son por definición estrellados. Si la función de densidad cumple esta propiedad se la denomina *profundidad convexa*.

Definición 2.1.2 (Conjunto convexo)**.** *Un conjunto S se considera convexo si el segmento lineal que une dos puntos cualesquiera del mismo, $x, y \in S$, pertenece a su vez al conjunto. $\overline{xy} \subseteq S, \forall x, y \in S$. Es sencillo ver que un conjunto convexo no es más que un conjunto estrellado con respecto a todos sus puntos.*

Finalmente, la propiedad ***P.5*** es una restricción técnica que en ocasiones se añade por utilidad.

2.2. Regiones centrales

El conjunto de puntos $x \in \mathbb{R}^d$ con profundidad igual o superior a un cierto valor $\alpha \in [0,1]$ se conoce como la región de profundidad D_α. Estas regiones, si D cumple las condiciones mencionadas anteriormente, tienen una serie de propiedades relevantes enunciadas entre otros por Cascos et al. (2011) y Mosler (2013).

- La región de profundidad para un $\alpha_{\text{máx}} = \sup_{x \in \mathbb{R}^d} D(x;P)$ se corresponde con los puntos de mayor profundidad y es a su vez la más interna. $D_{\alpha_{\text{máx}}}(P) = \{x \in \mathbb{R}^d : D(x;P) = \alpha_{\text{máx}}\}$.
- Las regiones de profundidad D_α forman conjuntos anidados de forma que $D_\beta(x;P) \subseteq D_\alpha(x;P)$ si $0 \leq \alpha \leq \beta$.
- Las regiones D_α presentan invarianza afín y son acotadas, cerradas, estrelladas y, si además se cumple ***P.3′***, convexas.

Otra aplicación interesante es que podemos definir la profundidad de un punto $x \in \mathbb{R}^d$ como el conjunto D_α más interno en el que está contenido. Formalmente, $D(x;P) = \sup\{\alpha : x \in D_\alpha(P)\}$.

Capítulo 3

Medidas de profundidad

En el presente capítulo se expondrán tres metodologías distintas para calcular la profundidad. En primer lugar se presentará la profundidad de Mahalanobis, debido a que la simplicidad de su cálculo la convierte en un buen punto introductorio, y se dedicará el resto del capítulo a las profundidades semiespacial y simplicial, dos medidas clásicas y ampliamente recomendadas por su versatilidad.

Para cada medida se postulará una definición de la misma y se expondrán algunas de sus características generales, se comprobará su cumplimiento de la propiedades enunciadas en el capítulo anterior y se desarrollará al método de cálculo con respecto a conjuntos de datos y distribuciones de probabilidad paramétricas.

3.1. Profundidad de Mahalanobis

3.1.1. Definición

Dado un punto $x \in \mathbb{R}^d$, Liu y Singh (1993) definen la profundidad de Mahalanobis como el inverso de la distancia de Mahalanobis al cuadrado más 1:

$$M_hD(x,P) = \frac{1}{1+(x-\mu_P)'\Sigma_P^{-1}(x-\mu_P)},$$

siendo μ_P y Σ_P el vector de medias y la matriz de varianzas-covarianzas de la distribución P. Es trivial ver que se maximiza cuando la distancia de Mahalanobis es 0, es decir, en el vector de medias que actuará siempre como el centro.

Además, en el centro $M_hD = 1$, por lo que $M_hD \in (0,1]$ cumpliendo la acotación inferior y superior que enunciaban Cascos et al. (2011) en su definición de función de profundidad. En las siguientes medidas veremos que estos límites son bastante generales y que bajo determinadas circunstancias es posible hallar cotas superiores más restrictivas para otras funciones de profundidad.

Al ser esta función una transformación de la distancia de Mahalanobis, las regiones de profundidad generadas mediante su aplicación presentan la forma de elipses concéntricas independientemente de la distribución subyacente. Es por esto que si las regiones de profundidad de la distribución no adoptan esta forma, su uso es desaconsejable. A cambio, es una profundidad muy ligera computacionalmente, con tiempos de cálculos muy reducidos comparados con la mayoría de funciones de profundidad, lo que la convierte en una herramienta útil cuando tenga sentido su aplicación.

3.1.2. Propiedades

Es fácil ver que por la propia definición de la distancia de Mahalanobis, esta función de profundidad cumple casi todas las propiedades deseables, invarianza, convexidad, monotonía, anulación en el infinito y semicontinuidad superior.

En cuanto a la invarianza afín se demuestra por el Teorema 3.1.1

Teorema 3.1.1. *La distancia de Mahalanobis, y por extensión, la profundidad de Mahalanobis son invariantes ante transformaciones afines.*

Demostración. *Sean A una matriz invertible de dimensiones $d \times d$, b un vector d-dimensional y P una distribución con vector de medias y matriz de varianzas-covarianzas μ_P y Σ_P respectivamente:*

- $E(AP + b) = AE(P) + b = A\mu_P + b.$
- $\Sigma_{AP+b} = A'\Sigma_P A.$

Por tanto la distancia de Mahalanobis al cuadrado de un punto x perteneciente a P tras la transformación sería: $(Ax+b-(A\mu_P+b))'(A'\Sigma_P A)^{-1}(Ax+b-(A\mu_P+b)) = (x-\mu_P)'A'(A')^{-1}\Sigma_P^{-1}A^{-1}A(x-\mu_P)$ $= (x-\mu_P)'\Sigma_P^{-1}(x-\mu_P)$.

En cambio, la propiedad ***P*.2**, profundidad máxima en el centro, solo se cumple si el punto de máxima profundidad coincide con la media, dado que M_hD se maximiza en este punto.

Con respecto a la propiedad $P.3$ y $P.3'$, por definición de la distancia de Mahalanobis, M_hD generará regiones de profundidad, D_α, con forma de elípses centradas en la media independientemente de la distribución subyacente. Estas elipses serán a su vez por construcción conjuntos convexos cumpliendo así ambas propiedades.

Finalmente, para probar la propiedad $P.4$ nos referimos al Teorema 3.1.2

Teorema 3.1.2. *La profundidad de Mahalanobis tiende a 0 a medida que nos alejamos del punto central.*

Demostración. *La distancia de Mahalanobis al cuadrado $(x-\mu_P)'\Sigma_P^{-1}(x-\mu_P)$ tiende al infinito a medida que nos alejamos del punto central. Por tanto, siendo la profundidad de Mahalanobis su inversa, está tenderá a su vez a 0.*

3.1.3. Cálculo a partir de un conjunto de datos

Para calcular la profundidad de Mahalanobis de un punto $x_0 \in \mathbb{R}^d$ con respecto a un conjunto de datos $X = (x_1, ..., x_n) \in \mathbb{R}^d$, basta con sustituir μ_P y Σ_P en la expresión de $M_hD(x, P)$ por sus equivalentes muestrales, $\overline{X}$ y S_X. Formalmente:

$$M_hD(x_0, X) = \frac{1}{1+(x_0-\overline{X})'S_X^{-1}(x_0-\overline{X})}.$$

Podemos ilustrar esta metodología representando gráficamente el resultado de aplicar esta función a una muestra proveniente de una $N_2(0, I_2)$. Ver Figuras A.1 - A.6 para ver la representación de las regiones de profundidad resultantes para muestras de tamaño 200 y tamaño 500.

Finalmente, a nivel muestral o con respecto a conjuntos de datos, las medidas de profundidad obtenidas mediante la aplicación de esta función de profundidad pueden presentar problemas de falta de robustez, si se calcula de la forma expuesta en esta sección. Esto es debido a que tanto la media como la matriz de varianzas-covarianzas son estadísticos no robustos. Este problema se puede subsanar sustituyendo estos estadísticos de posición central y dispersión por otros estadísticos robustos.

3.2. Profundidad semiespacial

3.2.1. Definición

Para un punto $x \in \mathbb{R}^d$, la **Profundidad semiespacial** (*HD, Halfspace Depth*), con respecto a una medida de probabilidad P se define como la mínima masa de probabilidad contenida por un semiespacio cerrado (Definición 3.2.1) que contenga a x, de entre todos los semiespacios cerrados que contengan a x. La idea principal detrás de dicha medida fue propuesta por Tukey (1975) respecto a

una muestra, siendo la primera vez que se trabajaba con el concepto de profundidad y posteriormente sería desarrollada por Donoho y Gasko (1992) por lo que también se la conoce como *Tukey depth* o *location depth*. Es una de las funciones de profundidad más utilizadas en la práctica, debido a que presenta un buen comportamiento ante situaciones diversas y es conceptualmente sencilla.

Definición 3.2.1 (Semiespacio)**.** *En general un semiespacio cerrado H, es cualquiera de los dos conjuntos convexos en que un hiperplano de dimensión $d-1$ divide un espacio de dimensión $\mathbb{R}^d$ de forma que al unir dos puntos cualesquiera pertenecientes cada uno a uno de los dos semiespacios debamos intersecar el hiperplano. Podemos expresar el semiespacio de forma lineal, tal que $\mathbf{c}^\intercal\mathbf{x} \geq b$ (si utilizamos un mayor estricto estaríamos ante semiespacios abiertos).*

Por tanto, la HD se definiría formalmente de la siguiente forma:

$$HD(x;P) = \text{ínf}\{P(H) : H \text{ semiespacio cerrado, } x \in H\}, x \in \mathbb{R}^d.$$

Aplicada al caso univariante la profundidad de un punto cualquiera $x \in \mathbb{R}^1$ no sería más que la mínima masa de probabilidad a cada lado del punto, $HD(x;P) = \text{mín}\{F(x), 1-F(x)\}$ siendo $F(x)$ la función de distribución. Es fácil ver entonces que, en $d=1$, la profundidad semiespacial máxima es de 0.5 y se alcanza cuando $F(x) = 1-F(x) = 0.5$, es decir, en la mediana de la distribución.

Posteriormente se probará que el máximo de $HD(x;P)$ es $\frac{1}{2}$ en cualquier dimensión, lo que es fácil de ver a nivel intuitivo, dado que al observar los dos semiespacios generados por un hiperplano cualquiera siempre que exista uno que encierre una masa de probabilidad superior a 0.5 inevitablemente el otro debe ser inferior a este valor para poder sumar 1, ateniéndonos a las propiedades básicas de la probabilidad, y por tanto la HD no puede ser superior a $\frac{1}{2}$.

Cabe destacar, en cualquier caso, que este resultado tanto en dimensión $d=1$ como en dimensiones superiores solo es generalizable en el caso absolutamente continuo. En el caso de las distribuciones discretas no necesariamente tiene que existir un valor x que cumpla $F(x) = 0.5$ (véase por ejemplo la distribución uniforme discreta con un número impar de posibles valores), por lo que en estas situaciones el valor exacto de la cota superior variará en cada caso, aunque si se cumpla en cualquier caso que el máximo debe ser igual o menor a 0.5.

3.2.2. Propiedades

Con respecto a las propiedades exigidas a las funciones de profundidad, la profundidad semiespacial las cumple todas, como demuestran por ejemplo Zuo y Serfling (2000), Cascos et al. (2011) o Mosler (2013). En primer lugar, por ejemplo, es fácil ver que la distribución relativa de los puntos es invariante a transformaciones afines y por tanto HD es, a su vez, invariante.

En cuanto a la propiedad ***P.2***, la demostraremos a través de los Teoremas 3.2.1 y 3.2.2 tal y como recoge Rousseeuw y Ruts (1999):

Teorema 3.2.1. *Para cualquier distribución de probabilidad P en $\mathbb{R}^d$ para la cual exista una densidad se puede demostrar que $\max_x HD(x;P) \leq \frac{1}{2}$ obteniendo así que $HD(x;P) \in [0, \frac{1}{2}]$.*

Demostración. *Por definición, cualquier hiperplano pasando por el punto θ, $\mathbf{c}^\intercal\theta$, divide el espacio en dos semiespacios cerrados que podemos denotar de forma genérica $H^+_{c,\theta}, H^-_{c,\theta}$ tal que $H^+_{c,\theta} = \{x \in \mathbb{R}^d : \mathbf{c}^\intercal\mathbf{x} \geq \mathbf{c}^\intercal\theta\}$ y $H^-_{c,\theta} = \{x \in \mathbb{R}^d : \mathbf{c}^\intercal\mathbf{x} \leq \mathbf{c}^\intercal\theta\}$. Para un cierto $\theta \in \mathbb{R}^d$ y P con densidad asociada, cualquier hiperplano que contenga x divide por tanto el espacio en dos semiespacios, $H^+_{c,\theta}, H^-_{c,\theta}$, y así su profundidad será la ínfima masa de probabilidad de todos los posibles pares $H^+_{c,\theta}, H^-_{c,\theta}$, $HD(\theta) = \text{ínf } min\{P(H^+_{c,\theta}), P(H^-_{c,\theta})\}$. Por construcción $P(H^+_{c,\theta}) + P(H^-_{c,\theta}) - P(H^+_{c,\theta} \cap H^-_{c,\theta}) = 1$ y como $H^+_{c,\theta} \cap H^-_{c,\theta}$ es un hiperplano su probabilidad es 0. De esta forma $P(H^+_{c,\theta}) = 1 - P(H^-_{c,\theta})$ y por tanto $min\{P(H^+_{c,\theta}), P(H^-_{c,\theta})\} \leq \frac{1}{2}$.*

A continuación, probamos que si la distribución de probabilidad tiene un punto considerado central, por ejemplo el caso de simetría angular, la cota superior anterior se alcanza en ese punto.

Definición 3.2.2 (Simetría angular)**.** *Una distribución P es angularmente simétrica respecto a un punto $\theta_0 \in \mathbb{R}^d$ si las variables aleatorias $\frac{X-\theta_0}{||X-\theta_0||}$ y $-\frac{X-\theta_0}{||X-\theta_0||}$ están equidistribuidas.*

De acuerdo a Zuo y Serfling (2000), si una distribución P es angularmente simétrica, entonces también presenta simetría semiespacial.

Definición 3.2.3 (Simetría semiespacial)**.** *Una distribución P presenta simetría semiespacial respecto a un punto $\theta_0 \in \mathbb{R}^d$ si todos los hiperplanos que pasan por θ_0 dividen el espacio $\mathbb{R}^d$ en dos semiespacios equiprobables.*

Teorema 3.2.2. *Para cualquier distribución de probabilidad P para la cual exista una densidad y presente simetría angular con respecto a algún $x = \theta_0$, entonces $\max_x HD(x; P) = HD(\theta_0; P) = \frac{1}{2}$.*

Demostración. *Si P presenta simetría angular respecto a un punto θ_0, para un semiespacio H^+_{c,θ_0} generado por un hiperplano cualquiera que contenga θ_0, $P(H^+_{c,\theta_0}) = P(intH^+_{c,\theta_0})$, siendo $intH^+_{c,\theta_0}$ el interior del semiespacio cerrado o equivalentemente el semiespacio abierto, (el hiperplano tiene probabilidad asociada 0). Por simetría angular $P(intH^+_{c,\theta_0}) = P(-intH^+_{c,\theta_0}) = P(intH^-_{c,\theta_0})$. De esta forma $P(intH^+_{c,\theta_0}) = P(intH^-_{c,\theta_0}) = \frac{1}{2}$. Así $\max HD(x; P) = HD(\theta_0; P) = \frac{1}{2}$.*

Y por tanto se alcanza la máxima profundidad en el punto central.

Con respecto a la propiedad ***P.3***, siguiendo de nuevo a Zuo y Serfling (2000), su cumplimiento queda establecido en el Teorema 3.2.3:

Teorema 3.2.3. *Dadas una distribución P con punto más profundo θ y un punto cualquiera $x \in \mathbb{R}^d$, $HD(x, P) \leq HD(\theta + \lambda(x - \theta); P), \forall \lambda \in (0, 1)$.*

Demostración. *Consideremos el ínfimo en la definición de la profundidad semiespacial entre todos los semiespacios que no contienen a θ. Para cualquier semiespacio cerrado con $\theta + \lambda(x - \theta)$ en su frontera, $H_{\theta+\lambda(x-\theta)}$, por el Teorema del hiperplano separador siempre existe un H_x tal que $H_x \subset H_{\theta+\lambda(x-\theta)}$ y por tanto $HD(x; P) \leq HD(\theta + \lambda(x - \theta); P)$.*

Teorema 3.2.4 (Hiperplano Separador)**.** *Sean X e Y dos subconjuntos convexos no vacíos disjuntos de $\mathbb{R}^d$. Entonces existe un vector v distinto de 0 y un número real c tal que $\mathbf{v}^\intercal x \geq c$ y $\mathbf{v}^\intercal y \leq c$ para todo $x \in X$ e $y \in Y$. El hiperplano $\mathbf{v}^\intercal \cdot = c$ separa ambos conjuntos.*

En su artículo, Rousseeuw y Ruts (1999) prueban además que la profundidad semiespacial cumple la propiedad ***P.3′***.

Finalmente, de acuerdo a Zuo y Serfling (2000), podemos comprobar el cumplimiento de ***P.4*** mediante el Teorema 3.2.5:

Teorema 3.2.5. *La profundidad semiespacial tiende a 0 a medida que nos alejamos del punto central.*

Demostración. *Sea X una variable aleatoria d-dimensional y el punto x un punto cualquiera perteneciente a $\mathbb{R}^d$, y sean $||X||$ y $||x||$ sus respectivas normas, es trivial ver que $P(||X|| \geq ||x||) \to 0$ cuando $||x|| \to \infty$ y que para cada x y X existe un semiespacio cerrado H_x tal que $H_x \subset \{||X|| \geq ||x||\}$, por tanto $HD(x; P) \to 0$ cuando $||x|| \to \infty$ probando la propiedad **P.4**.*

Con estas propiedades, obtenemos que las regiones de profundidad de la profundidad semiespacial cumplen todas las características enunciadas para las regiones centrales.

3.2.3. Cálculo para una distribución dada

Si P fuese una distribución con expresión conocida, podemos tratar de calcular la masa de probabilidad dejada por el hiperplano a partir de ella. Por ejemplo, si P fuese una normal estándar bidimensional $N_2(0, I_2)$, sabemos que es una distribución con simetría angular, por lo que la profundidad máxima se alcanzará en el centro, $(0, 0)$ en este caso, y será igual a $\frac{1}{2}$. Por otra parte si

consideramos un punto cualquiera (x_1, x_2), el hiperplano que lo contenga en su frontera y deje menor masa de probabilidad a un lado será el tangente a la circunferencia de radio $\sqrt{x_1^2 + x_2^2}$ centrada en $(0, 0)$. Por dicha simetría, además, podemos afirmar que la profundidad de todos los puntos pertenecientes a dicha circunferencia serán igualmente profundos y por tanto, podemos simplificar el cálculo, trasladándonos al punto $(\sqrt{x_1^2 + x_2^2}, 0)$. Finalmente, dado que las distribuciones marginales de una $N_2(0, I_2)$ son, a su vez, $N(0, 1)$ la profundidad de dicho punto será la masa de probabilidad que deja a su derecha evaluado en la marginal de x_1. Formalmente: $HD(x_i; N(0, I)) = 1 - \phi(\sqrt{x_1^2 + x_2^2})$ siendo ϕ la función de distribución de una normal estándar.

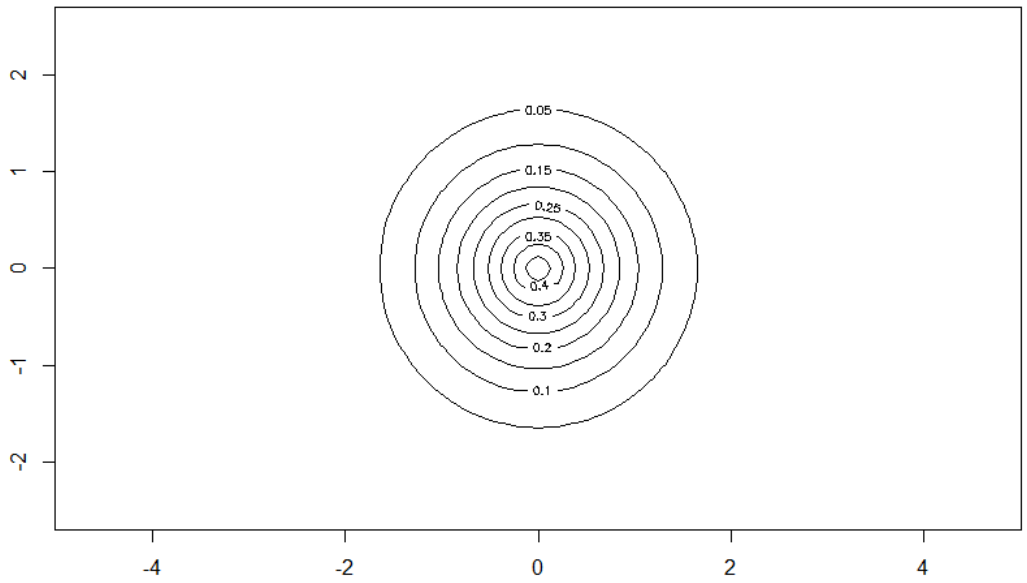
A partir de aquí, podemos extraer otro resultado: Todos los puntos $x_i = (x_{i,1}, x_{i,2})' \in \mathbb{R}^2$ tal que $\sqrt{x_{i,1}^2 + x_{i,2}^2} = c$ siendo c una constante cualquiera, son equiprofundos y por tanto los contornos de profundidad tendrán la forma de círculos concéntricos definidos por:

$HD(x_i; N_2(0, I_2)) = \alpha \Rightarrow 1 - \phi(\sqrt{x_{i,1}^2 + x_{i,2}^2}) = \alpha \Rightarrow 1 - \alpha = \phi(\sqrt{x_{i,1}^2 + x_{i,2}^2}) \Rightarrow \sqrt{x_{i,1}^2 + x_{i,2}^2} = \phi^{-1}(1 - \alpha)$ con $\alpha \in [0, \frac{1}{2}]$.

La expresión anterior podemos cambiar el signo $=$ por $\leq$ para hallar las regiones de profundidad en vez de los contornos.

Expresado formalmente $D_\alpha = \{x \in \mathbb{R}^2 : HD(x; N_2(0, I_2)) \geq \alpha\} = \{x \in \mathbb{R}^2 : \sqrt{x_1^2 + x_2^2} \leq \phi^{-1}(1 - \alpha)\}$. Ver Figura 3.1.

Figura 3.1: Contornos de Profundidad de $N_2(0, I_2)$



De forma más general, y dada la propiedad de invarianza afín, la profundidad de un punto x con respecto a una normal multivariante $N_d(\mu_d, \Sigma_d)$ es igual a la profundidad del punto $\Sigma^{-1/2}(x - \mu_d)$ con respecto a una normal multivariante estándar $N_d(0, I_d)$. De esta forma, podemos utilizar esta relación para construir las regiones de profundidad de la siguiente forma: $D_\alpha = \{x \in \mathbb{R}^d : \sqrt{(x - \mu_d)'\Sigma_d^{-1}(x - \mu_d)} \leq \phi^{-1}(1 - \alpha)\}$.

De esta forma, al adoptar las regiones de profundidad la forma de elipses concéntricas, la profundidad de un punto cualquiera depende solo de su distancia de Mahalanobis al centro, tal y como se puede apreciar en la expresión general enunciada en el párrafo anterior. De esta forma, cabría esperar que en el caso de la distribución normal, y de forma más general, cualquier distribución angularmente simétrica, la profundidad de Mahalanobis y la profundidad semiespacial ofrezcan resultados similares. Existirán en cualquier caso, ligeras diferencias entre ambas, dado que la profundidad de Mahalanobis realiza una transformación a la medida de distancia, mientras que en el caso de la semiespacial se trabaja sobre la propia distancia.

3.2.4. Cálculo a partir de un conjunto de datos

La profundidad semiespacial de un punto $x_0 \in \mathbb{R}^d$ con respecto a un conjunto de datos $x_1, ..., x_n \in \mathbb{R}^d$ se definiría como el mínimo número de puntos contenidos por un semiespacio cerrado con x_0 en su borde.

Si denotamos el hiperplano de la siguiente manera $\mathbf{c}^\intercal \cdot = b$ tal que $\mathbf{c}^\intercal x_0 = b$, la expresión formal de la profundidad semiespacial muestral sería:

$$HD(x_0; x_1, ..., x_n) = n^{-1} \min_{c \neq 0} \#\{i | \mathbf{c}^\intercal \mathbf{x_i} \geq \mathbf{c}^\intercal \mathbf{x_0}\}.$$

El cálculo de esta profundidad implica encontrar el conjunto de parámetros $\mathbf{c}^\intercal$ para el cual se minimiza la expresión anterior, y a medida que aumenta la dimensionalidad se vuelve computacionalmente muy costoso.

Podríamos, por ejemplo, testear si nuestro conjunto de datos procede de una normal y calcular la profundidad como en la subsección anterior estimando los parámetros correspondientes, media y matriz de varianzas-covarianzas. Si no es posible, o preferimos mantener el enfoque no paramétrico se han propuesto otros algoritmos alternativos (Ver sección Algoritmos) que permiten obtener tanto el valor exacto como aproximaciones de forma más eficientes.

Finalmente, Donoho y Gasko (1992), demuestran que, siendo $x_1, ..., x_n$ una muestra proveniente de la distribución de probabilidad P y con la consecuente distribución empírica P_n, la profundidad muestral convergerá a la poblacional de forma casi segura a medida que el tamaño muestral tienda a infinito. Formalmente, $\sup_x |HD(x; P_n) - HD(x; P)| \to 0$ cuando $n \to \infty$.

Si por ejemplo, aplicásemos la profundidad semiespacial a muestras procedentes de la normal estándar bivariante vemos en las Figuras A.7 - A.12 que obtenemos contornos con la forma aproximada de elipses concéntricas, en línea con la propiedad de convergencia recién mencionada.

En este caso la profundidad se ha calculado a través del paquete *"ddalpha"* de R utilizando uno de los algoritmos propuestos por Dyckerhof y Mozharovskyi (2016).

3.3. Profundidad simplicial

3.3.1. Definición

La **Profundidad simplicial** (*SD, simplicial Depth*), propuesta por Liu (1990), postula que, para un punto $x \in \mathbb{R}^d$, la profundidad con respecto a una medida de probabilidad P se define como la probabilidad de que dicho punto esté contenido en un símplex cerrado formado por $d+1$ observaciones aleatorias independientes de P.

$$SD(x, P) = P(x \in S\{X_1, X_2, ..., X_{d+1}\}), x \in \mathbb{R}^d.$$

Siendo $S(X_1, X_2, ..., X_{d+1})$ el símplex cerrado formado por $d+1$ observaciones.

Definición 3.3.1 (símplex)**.** *Definido formalmente, si $X_1, X_2, ..., X_{d+1}$ son puntos independientes afines, el símplex de dimensión $d+1$ es un politopo (objeto geométrico regular de caras planas) formado*

por la envoltura convexa de los $d+1$ vértices. Por ejemplo, en $d = 1$ serán segmentos lineales, en $d = 2$ adoptarán la forma de triángulos y en $d = 3$ tetraedros.

El punto x que tenga la profundidad máxima es considerado el centro y Liu (1990) demuestra que, en distribuciones de probabilidad con simetría angular, la profundidad de dicho punto será 2^{-d}, 0.25 en el caso bivariante por ejemplo.

Aplicado al entorno univariante, la profundidad simplicial se definiría como la probabilidad de que el punto θ este comprendido entre dos observaciones aleatorias e independientes de una distribución F, X_1 y X_2. Formalmente quedaría definida como $SD(x, F) = P(x \in \overline{X_1X_2}), x \in \mathbb{R}^1$, siendo $\overline{X_1X_2}$ el segmento lineal cerrado que une ambas observaciones.

De acuerdo a Liu (1988), en el caso que F sea continua esto se traduce a $SD(x, F) = 2F(X)(1 - F(x))$ por lo que es fácil ver que alcanza su valor máximo en 0.5 y se alcanza a su vez cuando $F(x)$ es igual a 0.5, es decir, en la mediana. Al igual que para la profundidad semiespacial, esta cota superior solo tiene sentido, tal y como matiza la propia autora, en el caso de distribuciones continuas, dado que en en distribuciones discretas no tiene por que haber un punto tal que $F(x) = 0.5$, y por tanto la cota superior será distinta.

Si bien esta función de profundidad, al igual que la semiespacial, se muestra bastante versátil, su uso no está tan extendido dado que es computacionalmente más compleja y las regiones de profundidad que genera no son necesariamente convexas, lo que a veces dificulta su aplicabilidad.

3.3.2. Propiedades

Aplicada a funciones de distribución continuas con simetría angular, Liu (1990) demuestra a través de los siguientes Teoremas que cumple las cuatro propiedades deseables de una función de profundidad propuestas por Zuo y Serfling (2000). En cambio, en este último artículo demuestran mediante una serie de contraejemplos que si la distribución subyacente es discreta, no se garantiza el cumplimiento de ***P.2*** y ***P.3***.

En primer lugar, es trivial ver que presenta invarianza afín pues al igual que en el caso anterior depende de la posición relativa del punto respecto a la distribución o conjunto de puntos y esta es invariante a cambio de localización o escala.

Con respecto a ***P.2***:

Teorema 3.3.1. *Si P es una distribución absolutamente continua en $\mathbb{R}^d$ y angularmente simétrica respecto a $\theta_0 \in \mathbb{R}^d$, entonces $SD(\theta_0; P) = \sup_{x \in \mathbb{R}^d} SD(x; P) = 2^{-d}$.*

Continuando con ***P.3***:

Teorema 3.3.2. *Si P es una distribución absolutamente continua en $\mathbb{R}^d$ y angularmente simétrica respecto al origen, entonces $SD(\alpha x; P)$ es monótonamente no creciente en $\alpha \geq 0$ para todo $x \in \mathbb{R}^d$.*

Finalmente, respecto a ***P.4***:

Teorema 3.3.3. *Para cualquier distribución P y punto x pertenecientes a $\mathbb{R}^d$, $\sup_{||x|| \geq M} D(x; P) \to 0$ cuando $M \to \infty$.*

3.3.3. Cálculo para una distribución dada

Los algoritmos de cálculo para la profundidad simplicial, trabajan con respecto a un conjunto de datos S y no con respecto a una distribución de probabilidad P.

Sin embargo, Liu (1990) demuestra que $\sup_{x \in \mathbb{R}^d} |SD(x; P_n) - SD(x; P)| \to 0$ de forma casi segura si $n \to \infty$ y P es una distribución absolutamente continua con densidad acotada. Además demuestra que si la profundidad máxima de P se alcanza en un punto único y la densidad no se anula en el entorno de dicho punto entonces el máximo muestral converge al poblacional también de forma casi segura.

De esta forma, podemos calcular una aproximación razonable de la profundidad con respecto a una distribución P como la profundidad respecto a una muestra proveniente de dicha distribución, siempre y cuando el tamaño muestral sea suficientemente elevado.

Por otra parte, Liu (1990) sí que ofrece alguna información sobre las características de la profundidad simplicial para ciertas distribuciones paramétricas. Principalmente, estipula que si la distribución subyacente P es elíptica, es decir, la densidad en un punto x es función de $(x-\mu)'V^{-1}(x-\mu)$ siendo μ un parámetro de localización y la matriz invertible V un parámetro de dispersión, los contornos de profundidad de P serán, a su vez, elipses anidadas centradas en μ.

3.3.4. Cálculo a partir de un conjunto de datos

En el caso de que tuviésemos una muestra $X = (x_1, ..., x_n) \in \mathbb{R}^d$, el procedimiento para calcular la SD en un punto x_0 con respecto a dicho conjunto de datos es de la siguiente forma:

1. Obtenemos todos los posibles símplex **S** formados por $d+1$ observaciones de la muestra. Estos serían por tanto $\binom{n}{d+1}$ posibles símplex.
2. Generar una variable indicadora $I(x_0 \in S)$ igual a 1 si x_0 está contenido en un determinado símplex o no.
3. Calcular la proporción de símplices en los que $I(x_0 \in S) = 1$.

Formalmente: $SD(x_0; X) = \binom{n}{d+1}^{-1} \sum_{1 \le i_1 < i_2 < i_{d+1} \le n} I(x \in S\{x_{i_1}, x_{i_2}, ..., xi_{d+1}\})$.

Si, por ejemplo, trabajásemos con una muestra $X \in \mathbb{R}^2$, nuestros símplex serían triángulos formados por tres puntos cualesquiera de la muestra, y obtendríamos la siguiente expresión:

$$SD(x_0; X) = \binom{n}{3}^{-1} \sum_{1 \le i < j < k \le n} I(x \in S\{x_i, x_j, x_k\}), x \in \mathbb{R}^2.$$

La profundidad simplicial muestral resultante es un U-estadístico de grado $d+1$ de acuerdo a Liu (1988), por lo que es insesgada por definición.

Definición 3.3.2 (U-estadístico)**.** *Hoeffding (1948). Dada una función $h(x_1, ..., x_m)$ y una muestra $X_1, ..., X_n$ i.i.d con distribución F y $n \ge m$, un U-estadístico con kernel h y grado m se define como el valor medio de la función evaluada en todas las posibles permutaciones $P_{m,n}$ de tamaño m de la muestra.*

$$U_n(h) = (\tfrac{n!}{(n-m)!})^{-1} \sum_{P_{m,n}} h(X_{i_1}, ..., X_{i_m}).$$

Si la función kernel h es simétrica, se puede simplificar como el promedio entre todas las combinaciones $C_{m,n}$ de m elementos de la muestra

$$U_n(h) = \binom{n}{m}^{-1} \sum_{C_{m,n}} h(X_{i_1}, ..., X_{i_m}).$$

Para el caso de distribuciones discretas mencionado anteriormente Burr et al. (2006) proponen una formulación revisada donde la profundidad se construye como el promedio de la proporción de símplex cerrados y abiertos que contienen al punto x, o expresado de otra forma, el promedio entre el de la proporción de símplex con x en su interior y con x en su frontera. Esta nueva definición presenta un comportamiento mejor en este caso, pero las propiedades $P.2$ y $P.3$ aún no se cumplen de forma general, por lo que sigue sin ser una medida de aplicabilidad universal.

Este procedimiento, que implicaría generar todos los posibles símplex y contar cuantos incluyen un punto dado, es computacionalmente muy costoso, pues si trabajásemos, por ejemplo, con una muestra de 200 individuos y 4 variables, tendríamos $\binom{200}{5} = 2.54 \cdot 10^9$ posibles símplex en los que evaluar nuestro punto de interés.

Si aplicamos esta medida a muestras procedentes de una $N_2(0, I_2)$ como en los casos anteriores (Ver Figuras A.13 - A.16.) observamos que obtenemos círculos concéntricos tal y como enunciábamos que debería ocurrir para distribuciones con curvas de nivel de forma elíptica como es el caso de la normal. Por otra parte, y como se puede apreciar fácilmente en dichas figuras, las regiones de profundidad D_α obtenidas a través de la profundidad simplicial constituyen conjuntos estrellados pero no convexos. volvemos a obtener alrededor del punto de simetría. Cálculos realizados a través del paquete *"ddalpha"* de R utilizando el algoritmo básico.

Capítulo 4

Algoritmos

El cálculo de la profundidad en dimensión superior a dos es altamente complejo, incluso en $d = 2$ se presentan dificultades en el cálculo exacto. Por ello han ido surgiendo algoritmos que calculen aproximaciones a los valores reales de la profundidad e incluso algoritmos alternativos para el cálculo del valor exacto. En cualquier caso no se suele trabajar en dimensiones superiores a 3 o 4.

Uno de los artículos fundamentales en este aspecto es el de Rousseeuw y Ruts (1996), quienes proponen un primer algoritmo alternativo para el cálculo exacto tanto de la profundidad semiespacial como de la simplicial en $\mathbb{R}^2$.

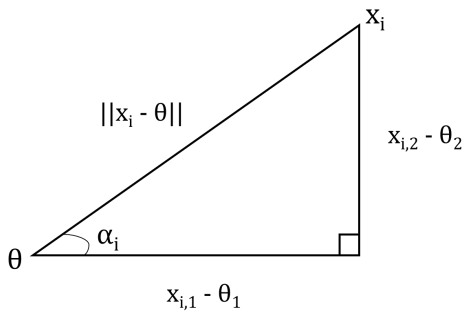
Dado un conjunto de datos $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^2$ y un punto cualquiera $\theta \in \mathbb{R}^2$ que no tiene porque pertenecer a X, estamos interesados en calcular, en primer lugar, su profundidad simplicial $SD(\theta, X)$.

Para ello, Rousseeuw y Ruts (1996) proponen el siguiente algoritmo. Primero generan un vector $\vec{\alpha}$ con el ángulo formado entre cada observación y θ. Para ello se basan en las propiedades de los triángulos rectángulos, que especifican que podemos calcular cualquiera de los dos ángulos no rectos del mismo como el arcoseno del cociente entre el cateto opuesto y la hipotenusa o el arcocoseno del cociente entre el adyacente y la hipotenusa.

Aplicado en nuestro caso, si visualizamos el triángulo formado por θ y un $x_i \in X$ (Ver Figura 4.1), cada cateto es una de las dos dimensiones del vector diferencia y la hipotenusa la norma de dicho vector. Así pues, podemos calcular $\alpha_i = \arcsin \frac{(x_{i,2}-\theta_2)}{||(x_i-\theta)||} = \arccos \frac{(x_{i,1}-\theta_1)}{||(x_i-\theta)||}$.

En particular, los autores proponen utilizar el arcoseno cuando $\frac{(x_{i,1}-\theta_1)}{||(x_i-\theta)||} > \frac{(x_{i,2}-\theta_2)}{||(x_i-\theta)||}$ y el arcocoseno cuando sucede al contrario. Si existe algún $x_i = \theta$ se retira por ahora para utilizarlo más adelante en el algoritmo.

Figura 4.1: Triángulo Rectángulo

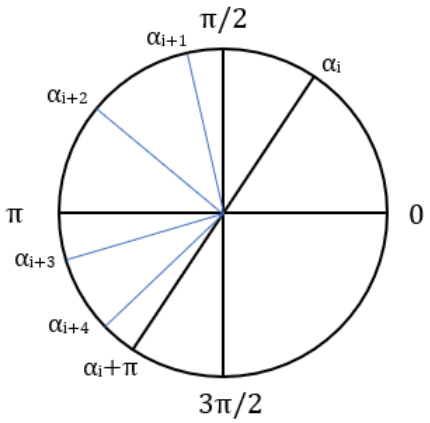


A continuación, se ordena este vector α_i en orden creciente y se calcula la máxima diferencia entre

dos posiciones consecutivas. Si el resultado es superior a π, θ está fuera de la nube de puntos y su profundidad es 0. Si θ no está fuera de la nube, en el vector de ángulos ordenados le restamos a todos los elementos el valor del primero, α_1 (rotamos la nube de puntos sobre θ hasta que $x_{1,2} - \theta_2 = 0$, lo que está permitido por invarianza afín), y obtenemos una nueva secuencia $0 = \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n < 2\pi$.

Una vez construida la secuencia anterior, calculamos h_i para cada individuo, donde h_i es el mayor entero tal que $\alpha_i \leq \alpha_{i+1} \leq \cdots \leq \alpha_{i+h_i} \leq \alpha_i + \pi$, es decir, el número de observaciones contenidos por el semicírculo formado por α_i y $\alpha_i + \pi$. Por ejemplo en la Figura 4.2 el h_i correspondiente a α_i sería de 4.

Figura 4.2: Cálculo h_i



La primera consecuencia del nuevo vector $\vec{h}$ es que un triángulo formado por tres observaciones, $\Delta(x_i, x_j, x_k)$, no incluye a θ si y solo si existe un arco circular menor a π que contenga todos los ángulos del símplice $(\alpha_i, \alpha_j, \alpha_k)$, es decir, que si el primero en aparecer es por ejemplo α_i, ambos α_j y α_k tendrían que encontrarse entre sus correspondientes h_i ángulos. De esta forma podemos medir la profundidad simplicial como el complementario de los símplices que no contienen a θ.

$SD(\theta, X) = \binom{n}{3} - \sum_{i=1}^{n} \binom{h_i}{2}$

En la expresión anterior, los autores deciden por conveniencia que $\binom{h_i}{2} = 0$ si $h_i < 2$, . En el caso de que existiesen $x_i = \theta$, a este valor de la profundidad habría que sumarle todos los símplices donde 1, 2 o los 3 vértices son iguales a θ en función del número de observaciones que se encuentren en esta situación.

En el mismo artículo, los autores adaptan el algoritmo anterior para el cálculo de la profundidad semiespacial. Concretamente, dados de nuevo el punto θ y el conjunto de datos $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^2$, postulan que su profundidad semiespacial se puede calcular como

$$HD(\theta, X) = \frac{1}{n} \min_i \{\min(k_i, n - k_i)\}$$

Desgranando la expresión anterior, $k_i = F(i) - G(i)$, donde

$$F(i) = \#\{j : 0 \leq \alpha_j < \alpha_i + \pi\},$$
$$G(i) = \#\{j : 0 \leq \alpha_j < \alpha_i\}$$

En ambos casos, los ángulos α_i y α_j se refieren al vector ordenado y rotado sobre θ sobre el que calculábamos los h_i. Los $F(i)$ indicarían el número de observaciones con ángulos menores que $\alpha_i + \pi$ y $G(i)$ el número de observaciones menores a α_i. De esta forma, podemos obtener una relación entre los h_i y los $F(i)$ de la forma $h_i = F(i) - i$, donde es trivial ver que las observaciones entre α_i y $\alpha_i + \pi$ son iguales al número de observaciones hasta $\alpha_i + \pi$ menos el valor de la posición de α_i en el vector.

De la misma forma, siempre que en el vector $\vec{\alpha}$ no existan empates, podemos expresar $G(i) = i - 1$ y por tanto, $k_i = F(i) - G(i) = h_i + i - (i - 1) = h_i + 1$.

De esta forma, k_i representaría lo mismo que h_i, pero incluyendo el propio α_i. Si extrapolásemos de vuelta a los puntos originales, los puntos contenidos en este semicírculo son aquellos contenidos por uno de los semiespacios generados por el hiperplano que une θ con x_i, incluyendo al propio x_i, y, por tanto, $n - k_i$ son el resto de puntos o aquellos incluidos en el otro semiespacio. Así calculamos la profundidad semiespacial como el mínimo absoluto de todos estos valores. En el caso de que tengamos $x_i = \theta$, se considera que está incluido en todos los hiperplanos considerados. Tanto en este caso como en el caso de la profundidad simplicial, el algoritmo devuelve el valor exacto de la profundidad correspondiente.

En el caso de la profundidad semiespacial, destacan también trabajos como Rousseeuw y Struyf (1998), Dyckerhof y Mozharovskyi (2016) o Cuesta Albertos y Nieto Reyes (2008). Estos últimos, plantean la profundidad semiespacial de otra forma, dada una distribución de probabilidad P, denominan Π_v a la proyección de $\mathbb{R}^d$ en el espacio unidimensional generado por vector $\vec{v} \in \mathbb{R}^d$, y de la misma forma, $P \circ \Pi_v^{-1}$ a la distribución marginal de P con respecto a dicho espacio unidimensional. Dado un conjunto de datos $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$ y el vector $\vec{v}$, $P \circ \Pi_v^{-1}$ sería la distribución del vector $X^* = \{x_1^*, \ldots, x_n^*\} \in \mathbb{R}^1$ donde, $x_i^* = v_1 x_{i,1} + \cdots + v_d x_{i,d}$.

De esta forma definen $HD(x; P) = \text{ínf}\{D_1(x^*; P \circ \Pi_v^{-1}) : v \in \mathbb{R}^d\}, x \in \mathbb{R}^d$, donde D_1 es la profundidad semiespacial unidimensional $D_1(x; P) = \text{mín}\{P(-\infty, x], P[x, \infty)\}$. De esta forma, la profundidad semiespacial de un punto se define como la ínfima profundidad unidimensional de entre las posibles proyecciones unidimensionales de x, con respecto a las marginales unidimensionales de P.

A partir de aquí, los autores proponen la profundidad semiespacial aleatorizada, donde, en vez de calcular el ínfimo con respecto a todas las posibles proyecciones, se calcula como la media de los valores de D_1 con respecto a k proyecciones seleccionadas de forma aleatoria. Demuestran, además, que se pueden obtener buenas aproximaciones con valores relativamente bajos de k, lo que reduce notablemente los tiempos de computación. Por ejemplo, sugieren que en $d = 2$, bastaría con un valor de k incluso inferior a 10.

En cuanto a la profundidad simplicial, Afshani et al. (2015) y Cheng y Ouyang (2001) proponen algoritmos aproximativos que producen mejoras en los tiempo de computación hasta dimensión 4. En cualquier caso, no se conoce un algoritmo más rápido que generar todos los símplices y contar en cuantos está contenido el punto si $d > 4$.

Capítulo 5

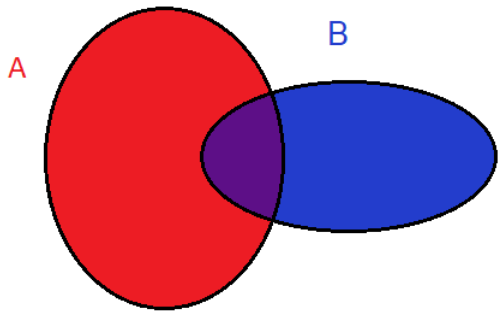
Estudio de simulación

A lo largo de este quinto capítulo se realizarán una serie de simulaciones a través del software R con el objetivo de evaluar el rendimiento de las medidas y algoritmos expuestos a lo largo del presente trabajo. Nos centraremos en dos aspectos principales a la hora de extraer resultados, el tiempo de computación requerido en cada caso y como de precisa es la aproximación de los contornos muestrales a los teóricos.

Para poder medir como de bien aproximan las regiones muestrales su contrapartida teórica, tomaremos los conjuntos dados por las regiones de profundidad D_α y calcularemos el área del conjunto de la diferencia simétrica entre ambos (Definición 5.0.1). A tal efecto generaremos una rejilla de puntos para los que evaluaremos en cada caso su pertenencia o falta de la misma tanto al conjunto teórico como al muestral.

Definición 5.0.1 (Diferencia Simétrica)**.** *Dados dos conjuntos, A y B, se conoce como diferencia simétrica entre ambos conjuntos $A \triangle B$ al conjunto formado por aquellos elementos que pertenecen o bien a A o a B pero no a ambos. Formalmente, $A \triangle B = (A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$, siendo $\cup$, $\cap$ y $\setminus$ los operadores unión, intersección y diferencia respectivamente. Ver Figura 5.1, donde el conjunto diferencia simétrica se correspondería con la unión de los conjuntos A y B, menos el conjunto intersección, coloreado en púrpura.*

Figura 5.1: Diferencia Simétrica



A partir de aquí, podemos aproximar el área de cada punto de la siguiente forma: Si tenemos una rejilla bidimensional, donde el eje x está comprendido en el intervalo (a, b) y el eje y por el intervalo (c, d) el área total de la rejilla será el área de un rectángulo, $base \cdot altura$, en este caso $(b - a)(d - c)$

unidades cuadradas. Si tenemos g puntos en cada eje, podemos aproximar el área que representa cada punto como el cociente entre el área total y el número de puntos, $\frac{(b-a)(d-c)}{g^2}$.

De esta forma, sabiendo el número de puntos incluidos en el conjunto de la diferencia simétrica, podemos aproximar el área de dicho conjunto.

Para poder formular unas conclusiones más completas y robustas, dichas simulaciones se realizarán bajo una serie de parámetros que variarán en las distintas realizaciones del experimento a fin de observar los resultados bajo distintas condiciones. En concreto, además de comparar distintas distribuciones de probabilidad subyacentes, serán variables a tener en cuenta el tamaño de la muestra n, y el nivel de profundidad α. Además, simularemos cada situación un total de 1000 veces, de forma que reduzcamos el efecto de la aleatoriedad en los resultados. Por último, destacar que a lo largo del experimento se utilizará sólo la profundidad semiespacial, dado que de las tres medidas presentadas es la más utilizada en la práctica pues como comentábamos en capítulos anteriores es más versátil que que la profundidad de Mahalanobis y más rápida y aplicable que la simplicial.

5.1. Distribución normal estándar bivariante

En primer lugar comenzaremos con uno de los escenarios más sencillos, donde la distribución de probabilidad subyacente es una normal estándar bivariante, $N_2(0, I_2)$. En este caso, utilizaremos una rejilla de 150 puntos en cada eje comprendidos ambos entre -2.5 y 2.5, obteniendo así un total de $150^2 = 22500$ puntos con un área aproximada de $\frac{5^2}{150^2} = 1.11 \cdot 10^{-3}$ unidades cuadradas (uds^2) cada uno. De esta forma obtenemos una rejilla bastante fina, de forma que los resultados sean lo más exactos posibles.

Con respecto a los tamaños muestrales, estos se establecerán en $n \in \{50, 500, 5000\}$ de forma que podamos comprobar que la profundidad muestral tiende a la poblacional cuando $n \to \infty$. Finalmente, trabajaremos con tres niveles de profundidad $\alpha \in \{0.1, 0.25, 0.4\}$ que, dado que la profundidad semiespacial está contenida en el intervalo $[0, 0.5]$, nos ofrecerá una visión global del rendimiento, desde las regiones más pequeñas hasta las más externas.

Se utilizarán cuatro métodos para el cálculo de las regiones de profundidad:

- Las regiones teóricas se calcularán utilizando la metodología expuesta en el apartado *"Cálculo para una distribución dada"* de la sección dedicada a la profundidad semiespacial, donde se demostraba que la región de profundidad D_α de una normal multivariante estaba compuesta por todos aquellos puntos cuya distancia de Mahalanobis a la media fuera inferior o igual al cuantil $1 - \alpha$ de una normal estándar univariante. En este caso se calcula la distancia de Mahalanobis con los parámetros poblacionales de la distribución.

- Las regiones muestrales paramétricas se construyen de la misma forma que las teóricas, salvo que para el cálculo de la distancia de Mahalanobis se utilizan los parámetros estimados a partir de la muestra de tamaño n de la distribución teórica.

- Las regiones muestrales no paramétricas se construyen primero mediante el algoritmo de cálculo exacto desarrollado por Dyckerhof y Mozharovskyi (2016) e implementado a través del paquete *"dd-alpha"* de R. Se construirán sin realizar ninguna hipótesis sobre la distribución subyacente ni sus parámetros y seleccionando aquellos puntos con profundidad igual o superior a α.

- Las regiones muestrales no paramétricas se calcularán, además, mediante la aproximación aleatoria propuesta por Cuesta Albertos y Nieto Reyes (2008) seleccionando $k = 10$ proyecciones. Se realizarán, al igual que las exactas, a través del paquete *"dd-alpha"* de R.

A continuación se presentarán las tablas con los resultados de las simulaciones, tanto las relativas al área del conjunto diferencia simétrica como las relativas al tiempo de computación.

En cuanto a la diferencia simétrica, el resultado expresado para cada método será la media de las áreas obtenidas en cada una de las 1000 simulaciones, siendo el conjunto diferencia simétrica indicador

de las regiones contenidas por el conjunto de profundidad teórico o el resultante de la aplicación del método muestral indicado, pero no por ambos. Dicho resultado estará expresado en unidades cuadradas y se acompañará entre paréntesis del error relativo, calculado como el cociente entre el error y el área del conjunto teórico.

Por otra parte, el tiempo de computación indicará el tiempo medio necesario para el cálculo de la profundidad, obtenido como el tiempo total utilizado para las 1000 simulaciones dividido entre el número de las mismas.

Antes de presentar las tablas, cabe destacar, primero, para facilitar la interpretación que las áreas de los conjuntos D_α teóricos son de 5.1597, 1.4292 y 0.2016 uds^2 para los niveles de profundidad $\alpha = \{0.1, 0.25, 0.4\}$ respectivamente. Para calcular estas áreas podemos aprovechar que los contornos de profundidad para una normal estándar bivariante forman circunferencias concéntricas con radio $\phi^{-1}(1-\alpha)$ para un nivel de profundidad α dado. Con este resultado podemos calcular el área como el área de un círculo tal que $A = \pi r^2$.

En cuanto a los tiempos de computación, tener en consideración que las simulaciones se han realizado utilizando un ordenador personal de sobremesa con un procesador *AMD Ryzen 5 2600 Six-Core Processor* de 3.40 GHz y 16 GB de RAM.

A continuación, en las Tablas 5.1 y 5.2 se pueden observar los resultados relativos a los errores y los tiempos de ejecución. En el todos los casos la diferencia simétrica ente los conjuntos teóricos y los muestrales paramétricos se denotará TP y se procederá de igual manera para los no paramétricos exacto y aproximado que se establecerán como TNPE y TNPA.

Área Conjunto Diferencia Simétrica (uds^2)			
TP	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	1.2011 (23.27 %)	0.3841 (7.44 %)	0.1167 (2.26 %)
$\alpha = 0.25$	0.5201 (36.39 %)	0.1625 (11.37 %)	0.05 (3.498 %)
$\alpha = 0.4$	0.1687 (83.68 %)	0.0572 (28.37 %)	0.019 (9.42 %)
TNPE	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	1.450 (28.1 %)	0.4825 (9.35 %)	0.1496 (2.899 %)
$\alpha = 0.25$	0.6409 (44.84 %)	0.205 (14.34 %)	0.0647 (4.53 %)
$\alpha = 0.4$	0.1802 (89.38 %)	0.0731 (36.26 %)	0.0219 (10.86 %)
TNPA	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	1.5121 (29.31 %)	0.5338 (10.34 %)	0.2862 (5.55 %)
$\alpha = 0.25$	0.6032 (42.2 %)	0.2104 (14.72 %)	0.0903 (6.32 %)
$\alpha = 0.4$	0.1917 (95.09 %)	0.0690 (34.23 %)	0.0258 (12.797 %)

Cuadro 5.1: Áreas de los conjuntos diferencia simétrica y error relativo para una $N_2(0, I_2)$. TP, TNPE y TNPA indican si el método utilizado para la profundidad muestral es el paramétrico, no paramétrico exacto o no paramétrico aproximado respectivamente

Tiempo Medio Computación (sg)			
TP	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	$3.09 \cdot 10^{-3}$	$3.12 \cdot 10^{-3}$	$4.11 \cdot 10^{-3}$
$\alpha = 0.25$	$3.33 \cdot 10^{-3}$	$3.31 \cdot 10^{-3}$	$4.10 \cdot 10^{-3}$
$\alpha = 0.4$	$3.17 \cdot 10^{-3}$	$3.08 \cdot 10^{-3}$	$3.83 \cdot 10^{-3}$
TNPE	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	0.1549	1.7468	19.3
$\alpha = 0.25$	0.1559	1.7569	19.9
$\alpha = 0.4$	0.1563	1.7677	19.1
TNPA	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	0.0173	0.1138	1.0516
$\alpha = 0.25$	0.0174	0.1106	1.0576
$\alpha = 0.4$	0.0168	0.1116	1.1211

Cuadro 5.2: Tiempo medio de computación para una $N_2(0, I_2)$. TP, TNPE y TNPA indican si el método utilizado para la profundidad muestral es el paramétrico, no paramétrico exacto o no paramétrico aproximado respectivamente

A partir de las tablas anteriores podemos extraer varios resultados principales. En primer lugar, el método paramétrico es, con bastante diferencia, el más rápido y además presenta los mejores resultados. En concreto, obtiene errores entre un 11.99 y 59.22 por ciento más pequeños que el no paramétrico aproximado dependiendo de los parámetros, y de entre un 6.38 y un 21.99 por ciento con respecto al exacto. Por su parte, en el mejor de los casos, este procedimiento es casi seis veces más rápido que el aproximado y casi 50 veces más rápido que el exacto.

Además, el error disminuye notablemente en todos los casos al aumentar el tamaño muestral. Este resultado va acorde a la teoría expuesta previamente, pues se había postulado que las regiones muestrales tendían a las teóricas al aumentar n, e incluso de forma intuitiva cabría esperar que a mayor n mayor y mejor información disponible y por tanto más precisos serán los estimadores.

Otro resultado inmediato es el aumento del error cuanto más interior es la región D_α considerada. La explicación más inmediata es el reducido tamaño de estas regiones, lo que hace que cualquier pequeño error sea porcentualmente más significativo, por lo que se debe exigir un grado de precisión mayor al trabajar con estos conjuntos. Concretamente, vemos que para los tres métodos considerados el porcentajes de error para $\alpha = 0.4$ es de, al menos, el doble que para $\alpha = 0.25$.

En cualquier caso, es razonable el buen desempeño del método paramétrico, pues los datos sobre los que estamos trabajando son una muestra proveniente de la distribución teórica que estamos asumiendo para dicho método y de la misma forma no cabría esperar el mismo rendimiento cuando no podamos asumir dicha hipótesis. Aun así si sometemos nuestro conjunto de datos a un contraste de hipótesis y podemos asumir normalidad, este método se postula como una opción a tener en cuenta.

Por otra parte, los métodos no paramétricos son considerablemente más lentos, acentuándose di-

cha diferencia a medida que aumenta el tamaño muestral. Especialmente el método exacto acaba acercándose a las 6 horas de tiempo de computación en la muestra de 5000 observaciones. Además, la diferencia entre el método aproximado y el exacto es más que notoria en este aspecto, pero no tanto en cuanto a precisión. El método exacto consigue obtener errores más pequeños en todos los casos, (salvo para $\alpha = 0.4$ con $n = 500$ y $\alpha = 0.25$ con $n = 50$ donde el aproximado es ligeramente mejor aunque de forma mínima) pero a costa de un tiempo computacional más de 10 veces superior. Concretamente el método exacto obtiene errores entre un 5.99 y un 47.72 por ciento menores respecto al aproximado.

Es reseñable también que, aunque es cierto que estamos evaluando la profundidad en una rejilla de más de $2 \cdot 10^4$ puntos, los tiempos de computación en los casos no paramétricos crecen muy rápidamente incluso en el caso bidimensional. Mientras que el método paramétrico calcula la profundidad de todos los puntos en milésimas de segundo el método aproximado aumenta hasta el segundo y el exacto hasta los 20 segundos. Si bien parece razonable calcular la profundidad de un conjunto de 22500 puntos con respecto a otros 5000 en tan sólo 20 segundos, en la práctica no es raro trabajar con conjuntos aún mayores, sin contar el hecho de limitar el cálculo a $\mathbb{R}^2$ siendo la dimensionalidad uno de los mayores causantes de aumentar los tiempos de computación. Queda patente así, una de las principales desventajas de esta metodología, que como hemos mencionado en este trabajo es su coste computacional.

Finalmente, en esta situación queda a costa del usuario decidir cuál de los dos aspectos valora en mayor medida y escoger en consecuencia, pero por norma general el algoritmo aproximativo parece una opción más razonable cuando el método paramétrico no pueda ser aplicado.

5.2. Presencia de outliers

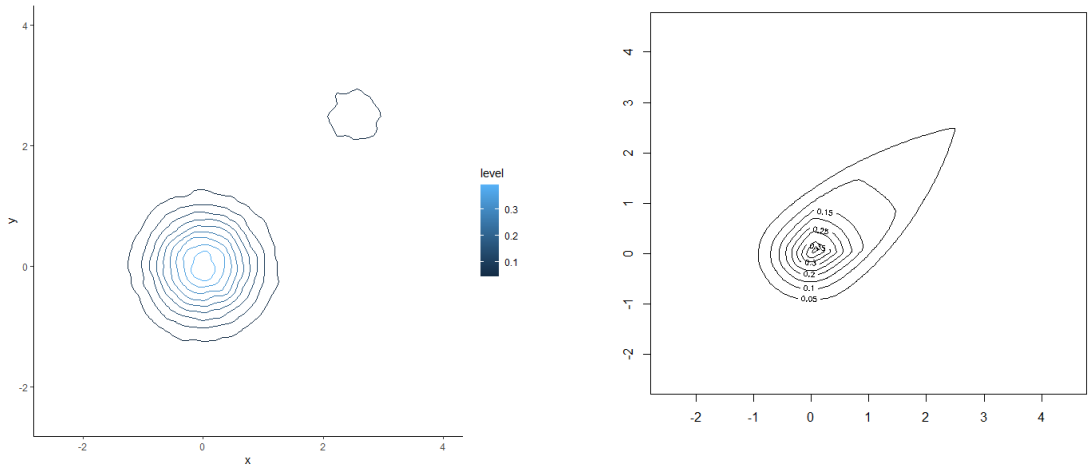
Una vez estudiado el caso más simple, podemos explorar qué sucede cuando los datos no son tan fieles a una distribución proveniente de un modelo paramétrico concreto. Para ello, continuando con el ejemplo de la normal bidimensional, podemos utilizar modelos de mixturas de normales para generar datos con diferentes características.

En primer lugar, podemos interesarnos por el caso donde el conjunto de datos presenta un reducido número de valores con un comportamiento diferente al de la mayoría. Estos datos atípicos afectarán a las regiones de profundidad, pero por su poca probabilidad deberían ser notorios principalmente en las más externas.

Para ello, generaremos una nueva variable aleatoria Z_1, mixtura de dos normales. Específicamente, tendremos una normal centrada en $\mu_1 = (0, 0)$ y con una matriz de varianzas covarianzas, $\Sigma_1 = \frac{1}{3} I_2$ y otra normal centrada en $\mu_2 = (2.5, 2.5)$ con una $\Sigma_2 = \frac{1}{4} I_2$. A estas normales les asignaremos pesos $\gamma = (0.9, 0.1)$ de forma que obtengamos una distribución basada en la primera componente, pero contaminada con una pequeña probabilidad con observaciones provenientes de la segunda componente. Formalmente, $Z_1 = \frac{9}{10} N_2(\mu_1, \Sigma_1) + \frac{1}{10} N_2(\mu_2, \Sigma_2)$.

De esta forma, estudiaremos el efecto que puedan tener los outliers sobre los tres métodos muestrales anteriores. En este caso, los contornos de profundidad teóricos se aproximarán aplicando el método no paramétrico exacto a una muestra de Z_1 de tamaño $n = 1 \cdot 10^5$, que debería ser una muestra lo suficientemente grande como para eliminar la aleatoriedad intrínseca al trabajo con muestras. En la Figura 5.2 podemos observar el dichos contornos junto a las curvas de nivel de la densidad.

Figura 5.2: Contornos de Densidad y Profundidad Z_1



(a) Contornos de Densidad

(b) Contornos de Profundidad

Para esta primera mixtura, los contornos de densidad mantienen la forma de circunferencias concéntricas para la primera componente de la mixtura, pero la introducción de la segunda componente genera una nueva región de densidad mayor que 0 alejada de las primeras. Esto se ve reflejado en los contornos de profundidad haciendo que las circunferencias presenten extensiones hacia la región de la segunda componente de la mixtura, rompiendo la forma elíptica. Dado que la segunda componente tiene mucho menos peso que la primera, este efecto en las regiones de profundidad es más notorio a medida que trabajamos con regiones más externas o menos profundas.

En esta mixtura, la rejilla utilizada seguirá siendo de 150 puntos por eje, pero estos ahora estarán comprendidos en el intervalo $(-2.5, 4.5)$, por lo que cada punto representará un área aproximada de $\frac{7^2}{150^2} = 2.17 \cdot 10^{-3}$, que es una aproximación ligeramente peor que el caso anterior pero debería seguir siendo lo suficientemente fiable.

En este caso, el área de las regiones de profundidad teóricas serían de 2.9444, 0.5466 y 0.06098 uds^2 para los niveles $\alpha = \{0.1, 0.25, 0.4\}$ respectivamente. Dichas áreas se han calculado a través de la aproximación utilizada para el área del conjunto diferencia simétrica, es decir, calculando cuantos puntos de la rejilla caen en el interior de cada contorno y multiplicando dicho valor por el área de cada punto, debido a que ya al contrario que para la normal estándar no disponemos de una expresión explícita que nos permita calcularlas.

Si calculásemos las áreas de las regiones de profundidad de la primera componente de la mixtura, una normal estándar centrada en $\mu_1 = (0, 0)$ y con matriz de varianzas-covarianzas $\Sigma_1 = \frac{1}{3} I_2$, que sería nuestra distribución sin contaminación de datos atípicos, podríamos utilizar de nuevo la expresión $A = \pi r^2$. En este caso las áreas serían de 1.7199, 0.4764 y 0.0672 para los niveles $\alpha = \{0.1, 0.25, 0.4\}$ respectivamente. Observamos que al introducir los datos atípicos las regiones de profundidad para α igual a 0.1 y 0.25 han multiplicado su tamaño por un factor de 1.711 y 1.147 cada una. En cambio la región más interna se reduce ligeramente, aunque quizás pueda deberse a la aproximación utilizada en el caso con outliers.

Finalmente, habrá un cambio notable con respecto al caso anterior, dado que, tras ver los elevados tiempos de computación, el número de simulaciones se reducirán a la mitad, es decir 500, que deberían ser suficientes para seguir obteniendo resultados fiables.

Para esta primera mixtura podemos observar los resultados en las Tablas 5.3 y 5.4:

Área Conjunto Diferencia Simétrica (uds^2)			
TP	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	1.1231 (38.14 %)	0.741 (25.17 %)	0.6528 (22.17 %)
$\alpha = 0.25$	0.5327 (97.45 %)	0.4508 (82.47 %)	0.4256 (77.86 %)
$\alpha = 0.4$	0.1452 (238.11 %)	0.1313 (215.32 %)	0.1283 (210.39 %)
TNPE	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	1.2661 (43 %)	0.6183 (20.999 %)	0.3165 (10.75 %)
$\alpha = 0.25$	0.2614 (47.82 %)	0.0836 (15.29 %)	0.0283 (5.18 %)
$\alpha = 0.4$	0.0596 (97.74 %)	0.0251 (41.16 %)	$7.49 \cdot 10^{-3}$ (12.28 %)
TNPA	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	1.4957 (50.79 %)	0.7512 (25.51 %)	0.4664 (15.84 %)
$\alpha = 0.25$	0.267 (48.84 %)	0.0961 (17.58 %)	0.0548 (10.02 %)
$\alpha = 0.4$	0.0669 (109.71 %)	0.0266 (43.64 %)	0.0129 (21.15 %)

Cuadro 5.3: Áreas de los conjuntos diferencia simétrica y error relativo para la mixtura Z_1. TP, TNPE y TNPA indican si el método utilizado para la profundidad muestral es el paramétrico, no paramétrico exacto o no paramétrico aproximado respectivamente

Tiempo Medio Computación (sg)			
TP	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	$3.12 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$	$3.94 \cdot 10^{-3}$
$\alpha = 0.25$	$3.02 \cdot 10^{-3}$	$3.22 \cdot 10^{-3}$	$3.98 \cdot 10^{-3}$
$\alpha = 0.4$	$3.12 \cdot 10^{-3}$	$3.16 \cdot 10^{-3}$	$4 \cdot 10^{-3}$
TNPE	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	0.15966	1.7402	18.77
$\alpha = 0.25$	0.1595	1.7404	25.24
$\alpha = 0.4$	0.1602	1.7329	18.62
TNPA	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	0.0166	0.11	1.0198
$\alpha = 0.25$	0.0166	0.1098	1.0223
$\alpha = 0.4$	0.0173	0.1081	1.0438

Cuadro 5.4: Tiempo medio de computación para la mixtura Z_1. TP, TNPE y TNPA indican si el método utilizado para la profundidad muestral es el paramétrico, no paramétrico exacto o no paramétrico aproximado respectivamente

Cuando trabajamos con una distribución con presencia de outliers, vemos que el método paramétrico empieza a tener problemas de precisión. Si bien presenta errores relativos no demasiado altos para el nivel de profundidad $\alpha = 0.1$, para regiones más internas el error se dispara notablemente, con errores superiores al 200 % para el contorno más interno para cualquier tamaño muestral.

Por su parte los métodos no paramétricos rinden bastante mejor en general, aunque los errores en muestras pequeñas es significativo. Además, para esta distribución concreta observamos un comportamiento peculiar, ambos métodos no paramétricos estiman mejor la región intermedia, $\alpha = 0.25$, que la interna o la externa. Una posible explicación la encontraríamos en el hecho de que el área más interna tiene un área muy reducida, haciendo que cualquier error sea porcentualmente más significativo, mientras que el área más externa es la más afectada por los outliers que a su vez influyen en el error cometido. En cambio la región intermedia no sufre tanto ninguno de los dos efectos, obteniendo mejores resultados.

Por su parte, los tiempos de ejecución mantienen el mismo comportamiento observado en la simulación anterior, con el método paramétrico siendo notablemente más rápido y el no paramétrico exacto mucho más lento.

De esta forma parece que ante la presencia de datos atípicos los métodos no paramétricos muestran un comportamiento más robusto, haciéndolos la opción más razonable en estos casos.

Si se buscase una mayor precisión sin sacrificar tanto el tiempo de computación, se podría experimentar aumentando el número de proyecciones del método no paramétrico aproximado, que seguirá siendo más rápido que el exacto mientras el aumento de k no sea demasiado elevado.

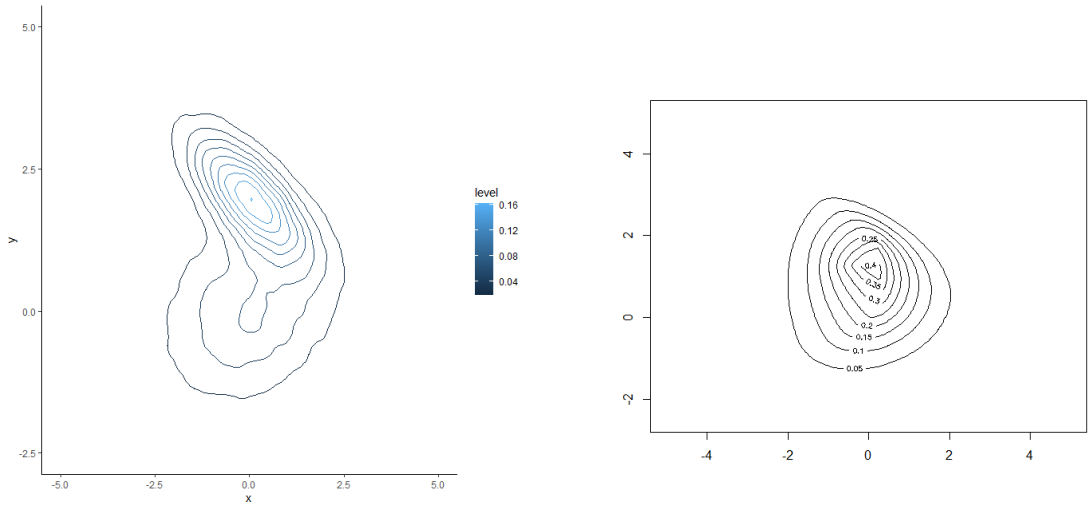
5.3. Mixtura equiprobable de normales

Tras comprobar como la presencia de datos atípicos traía consigo perdidas en la precisión de los métodos para el cálculo de la profundidad, especialmente el método paramétrico, podemos preguntarnos que pasaría si nuestra mixtura no simulase contaminación por outliers, sino crease una distribución completamente distinta a las normales de origen.

Con este objetivo, podemos modificar los pesos utilizados al crear la variable y generar una mixtura de dos normales con igual probabilidad, obteniendo así una nueva variable, Z_2. En el cálculo de la profundidad para esta distribución cabría esperar que el método paramétrico, que recordemos asume normalidad, vuelva a obtener tasas de error significativamente elevadas, pero queda estudiar como se comportan los métodos no paramétricos en estas circunstancias.

En este caso podemos, por ejemplo, crear dos subpoblaciones separadas en el espacio y con distinta dispersión tal que $Z_2 = \frac{1}{2}N_2(\mu_1, \Sigma_1) + \frac{1}{2}N_2(\mu_2, \Sigma_2)$, con $\mu_1 = (0, 0)$, $\mu_2 = (0, 2)$, $\Sigma_1 = \left(\begin{smallmatrix} 2 & 0.3 \\ 0.3 & 1 \end{smallmatrix}\right)$ y $\Sigma_2 = \left(\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 0.5 \end{smallmatrix}\right)$, obteniendo una distribución con la forma que se puede observar en la Figura A.17, donde cada componente de la mixtura está representado de un color, además de sus contornos de densidad y de profundidad teóricos en la Figura 5.3. Estos contornos de profundidad tampoco pueden ser calculados de forma exacta como en la primera simulación por lo que los aproximamos de nuevo mediante la aplicación de los contornos no paramétricos exactos a una muestra de Z_2 con $n = 1 \cdot 10^5$.

Figura 5.3: Contornos de Densidad y Profundidad Z_2



(a) Contornos de Densidad

(b) Contornos de Profundidad

Podemos observar fácilmente a partir de estos gráficos que, mientras que en el caso de la normal bivariante tanto los contornos de densidad como los de profundidad forman elipses concéntricas centradas en la media, la forma de los contornos en el caso de la mixtura es significativamente distinta entre los de densidad y los de profundidad. Mientras que los primeros son, efectivamente, elipses concéntricas en la parte superior (la correspondiente al segundo componente de la mixtura), el primer componente de la mixtura forma un saliente en la parte inferior que rompe las elipses más externas. Por su parte los contornos de profundidad sí son concéntricos, pero presentan una forma solo vagamente elíptica.

Con respecto a las simulaciones, en este caso mantendremos en número de simulaciones a 500, pero como esta nueva distribución está distribuida por un mayor espacio, en esta caso la rejilla tendrá como limites $[-5, 5]$ y $[-2.5, 5]$ para x e y respectivamente. Para compensar esto, aumentaremos el número de puntos por eje de 150 a 175, obteniendo en total 30625 puntos con un área aproximada de

$2.45 \cdot 10^{-3} uds^2$ cada uno.

Las áreas de los conjuntos de profundidad teóricos para Z_2 son, a su vez, de 7.969, 2.248 y 0.164 uds^2 para los niveles $\alpha = \{0.1, 0.25, 0.4\}$ respectivamente, calculadas de nuevo utilizando la misma aproximación que para el área del conjunto diferencia simétrica.

A continuación, en las Tablas 5.5 y 5.6 podemos observar los resultados de error y tiempo medio de ejecución.

Área Conjunto Diferencia Simétrica (uds^2)			
TP	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	2.319 (29.1 %)	1.401 (17.58 %)	1.1779 (14.78 %)
$\alpha = 0.25$	0.892 (39.68 %)	0.4385 (19.51 %)	0.359 (15.97 %)
$\alpha = 0.4$	0.3327 (202.87 %)	0.2881 (175.67 %)	0.2749 (167.62 %)
TNPE	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	2.166 (27.18 %)	0.6978 (8.76 %)	0.2385 (2.99 %)
$\alpha = 0.25$	1.039 (46.22 %)	0.3403 (15.14 %)	0.1119 (4.98 %)
$\alpha = 0.4$	0.1835 (112.8 %)	0.0932 (56.38 %)	0.0355 (21.65 %)
TNPA	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	2.504 (31.42 %)	0.8752 (10.98 %)	0.5142 (6.45 %)
$\alpha = 0.25$	1.03 (45.82 %)	0.4071 (18.11 %)	0.1974 (8.78 %)
$\alpha = 0.4$	0.2432 (148.29 %)	0.1223 (74.57 %)	0.0692 (42.2 %)

Cuadro 5.5: Áreas de los conjuntos diferencia simétrica y error relativo para la mixtura Z_2. TP, TNPE y TNPA indican si el método utilizado para la profundidad muestral es el paramétrico, no paramétrico exacto o no paramétrico aproximado respectivamente

Cuando aplicamos la metodología a esta variable Z_2 vemos que tal y como esperábamos el método paramétrico obtiene tasas de error notablemente superiores a los otros métodos destacando el caso del contorno más profundo, donde el conjunto diferencia simétrica tiene un área mayor al conjunto teórico para cualquier tamaño muestral. Al contrario que en la simulación con outliers, el método paramétrico si obtiene tasas de error inferiores al 200 % para la región más interna, pero sigue alcanzando su mínimo en 167 %, muy lejos de ser aceptable.

Por su parte, la diferencia entre los métodos no paramétricos continúa en la línea vista hasta ahora, donde el aproximado sacrifica precisión a cambio de velocidad. Si bien el aumento de error parece ser una cuantía asumible a cambio de la notable mejora en los tiempos de computación obtenida, sigue sin demostrar la superioridad de un método frente a otro y, como se ha enunciado anteriormente, quedaría en manos del usuario valorar que aspecto considera más importante a la hora de elegir uno u otro.

Tiempo Medio Computación (sg)			
TP	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	$5.22 \cdot 10^{-3}$	$4.88 \cdot 10^{-3}$	$5.34 \cdot 10^{-3}$
$\alpha = 0.25$	$4.28 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$
$\alpha = 0.4$	$4.12 \cdot 10^{-3}$	$4.42 \cdot 10^{-3}$	$5.82 \cdot 10^{-3}$
TNPE	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	0.215	2.781	28.8
$\alpha = 0.25$	0.214	2.406	26.2
$\alpha = 0.4$	0.222	2.415	26.6
TNPA	$n = 50$	$n = 500$	$n = 5000$
$\alpha = 0.1$	0.0229	0.155	1.42
$\alpha = 0.25$	0.0226	0.153	1.427
$\alpha = 0.4$	0.0232	0.15	1.426

Cuadro 5.6: Tiempo medio de computación para la mixtura Z_2. TP, TNPE y TNPA indican si el método utilizado para la profundidad muestral es el paramétrico, no paramétrico exacto o no paramétrico aproximado respectivamente

Capítulo 6

Aplicaciones

En este sexto y último capítulo se expondrán algunas de las distintas aplicaciones que se han desarrollado a partir del concepto de profundidad y se ilustrarán mediante su aplicación a dos datasets clásicos. En particular, comenzaremos analizando su utilidad en el ámbito del análisis exploratorio de datos multivariantes y la detección de outliers.

Continuaremos presentando el DD-plot, una herramienta gráfica para la comparaciones de distribuciones mediante profundidad y desarrollaremos su aplicación en contraste de hipótesis multivariante y en clasificación. Cerraremos el capítulo exponiendo el uso de la profundidad como una alternativa para la elección de ajustes en regresión lineal.

Con el fin de ilustrar dichas aplicaciones, utilizaremos dos conjuntos de datos clásicos obtenidos del repositorio de Machine Learning de UCI de la Universidad de California, a través de *cars* y *wine* respectivamente (Schlimmer (1987), Cortez et al. (2009)). En particular utilizaremos el conjunto *wine* para ilustrar el uso en clasificación y el de *cars* para el resto de aplicaciones.

El primer conjunto, *cars*, incluye información relativa a 205 vehículos y 26 variables divididas en el modelo del vehículo, su precio y diversas especificaciones técnicas. Estos datos fueron obtenidos originalmente en Estados Unidos en la década de los 80 con el objetivo de estudiar que características influenciaban el riesgo que las aseguradoras asignaban a un determinado vehículo. En este capítulo, con el objetivo de ilustrar las distintas aplicaciones y por simplicidad, trabajaremos en el entorno bivariante y nos centraremos a tal fin en las variables precio y citympg.

La variable precio indica el precio del vehículo en el mercado y se encuentra comprendido entre 5118 y 45400 dólares, aunque el 75 % de las observaciones se encuentran por debajo de 16500 dólares, lo que nos indica la presencia de asimetría, fenómeno habitual al trabajar con variables monetarias del tipo precios o ingresos.

Por su parte, citympg (city miles per gallon) es una medida utilizada en Estados Unidos, Reino Unido y Canadá que indica las millas que es capaz de realizar el vehículo en ciudad por cada galón de combustible que consume. En Europa en cambio se suele expresar a la inversa, como Litros consumidos por cada 100 km recorridos. Las equivalencias correspondientes serían, $1mpg = 0.4251km/L = 235.2146L/100km$.

En cuanto al mpg de un vehículo, valores más altos implican mayor eficiencia y es de esperar que exista una relación inversa entre esta variable y el precio, dado que los vehículos más económicos o de calle presentan diseños de motor más centrados en la eficiencia, mientras que vehículos de más alta gama suelen mejorar el rendimiento a costa de dicha eficiencia. En este caso la variable se encuentra en valores entre 13 y 49 mpg, aunque al igual que con el precio el tercer cuartil es de 30 mpg, lo que nos pone de nuevo sobre aviso de asimetría. Con esto, ya podemos empezar a pensar en la posible presencia de outliers y su influencia en la regresión lineal por mínimos cuadrados clásica.

Por otro lado, el conjunto *wine* incluye 6497 observaciones de una determinada variedad de vino portugués. Para cada una de ellas incluye 14 variables, divididas en características químicas y de composición y una valoración entre 3 y 9, siendo 3 los peores vinos y 9 los mejores. Esta última

variable podemos transformarla en una dummy que indique si es o no de buena calidad, considerando como casos positivos o vinos buenos aquellos con calidad superior a 5. Obtendremos un clasificador que divida a los vinos en buenos o malos en función de su grado de alcohol y su nivel de ácido acético, las dos variables con las que presenta mayor correlación del conjunto.

6.1. Análisis exploratorio y detección de outliers

La aplicación más inmediata de la profundidad la obtenemos en análisis exploratorio para datos multivariantes. El cálculo del punto más profundo o región más profunda nos ofrece una medida de tendencia central robusta frente a valores atípicos. Del mismo modo, el calculo de las regiones D_α nos permite obtener la aproximación multivariante de los cuantiles.

En este sentido, Rousseeuw, Ruts y Tukey (1999) desarrollan el bagplot como un equivalente bivariante del diagrama de cajas. El bagplot de un conjunto bivariante de datos incluye el punto de máxima profundidad o central, así como dos conjuntos denominados Bolsa y Lazo que ofrecen el primero una medida de dispersión y el segundo una indicación de posibles outliers. Al igual que este último, el bagplot nos permite visualizar gráficamente varias características de los datos. Como acabamos de mencionar, el centro y la Bolsa nos permiten visualizar gráficamente localización y dispersión, pero a su vez, la forma de la Bolsa nos puede dar información acerca de la correlación entre las variables (su orientación puede indicar correlación positiva o negativa) o de su simetría o asimetría a través de su forma.

Para su construcción y dado un conjunto de datos $X = (x_1, ..., x_n) \in \mathbb{R}^d$, emplean la profundidad semiespacial y determinan en primer lugar el punto θ que maximice $HD(\theta; X)$ que es la mediana o punto central. En caso de que el punto de máxima profundidad sea no singular escogen el promedio de los puntos que cumplan la condición de máxima profundidad.

A continuación calculan el α_0 tal que la región de profundidad D_{α_0} sea el conjunto convexo que contenga el 50 % de las observaciones más profundas. A este conjunto es al que denominan la Bolsa, B, y sería el equivalente a la caja del boxplot unidimensional.

Finalmente, para la creación del segundo conjunto, el Lazo, Rousseeuw, Ruts y Tukey (1999) sugieren "inflar" la bolsa con respecto al punto central por un factor de 3 y considerar como outliers todas las observaciones que queden fuera de esta región. En cambio, Cascos et al. (2011) sugieren un método alternativo, se traza una línea recta entre el punto de máxima profundidad y todos los puntos fuera de B, y se nomina como outliers aquellos para los que la línea recorra al menos el doble de distancia fuera de la Bolsa que dentro. Finalmente, generan el lazo como la envolvente convexa de todos los puntos que no hayan sido denominados outliers.

En su artículo, Rousseeuw, Ruts y Tukey (1999), postulan también que, dado que la profundidad semiespacial es invariante ante transformaciones afines, al aplicarle una de estas transformaciones a los datos el bagplot sufrirá una transformación acorde haciendo que los puntos que pertenecían a la Bolsa sigan perteneciendo y los que se habían considerado outliers siendo teniendo dicha consideración.

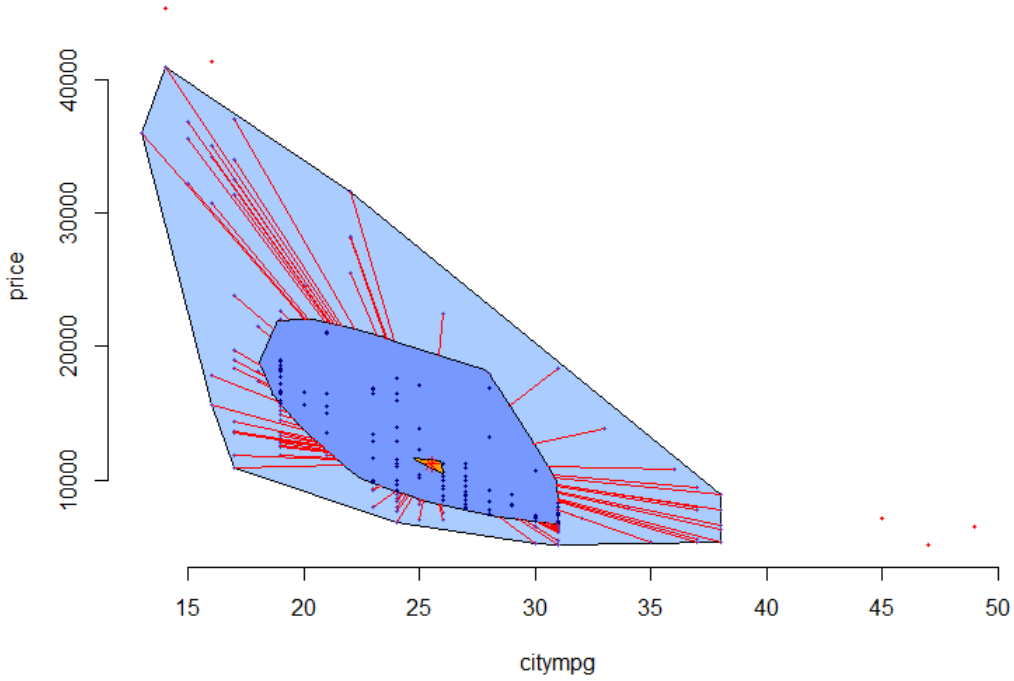
Podemos ver en la Figura 6.1, un ejemplo de bagplot aplicado a nuestro conjunto de datos de vehículos, donde apreciamos en amarillo el centro de la distribución, en azul oscuro la bolsa y en azul más claro el lazo. Observamos que el centro de la distribución se encuentra en valores muy bajos de la variable precio siendo un claro indicio de asimetría en la distribución. Esto es debido a que encontramos muchos más vehículos de gamas bajas en circulación que de gamas medias y sobre todo altas. En cambio, la eficiencia en el consumo de combustible si parece presentar una distribución más simétrica, a excepción de tres vehículos con una eficiencia anormalmente alta.

Son estos tres vehículos de eficiencia inusualmente alta, junto a los dos vehículos más caros del conjunto, los que han sido marcados como posibles outliers mediante la aplicación del bagplot.

La forma del bagplot, por su parte, nos parece indicar una relación inversa entre ambas variables, y estos 5 posibles outliers muestran un comportamiento acorde a esta relación. De esta manera parece que su inclusión como outliers se debe a su valor extremo, que sería entonces necesario comprobar si se trata de errores que deben ser subsanados o dichos valores son correctos.

Finalmente, podemos observar en las Figuras A.18 y A.19 otros bagplots, generados juntando la variable precio con otras dos variables del conjunto, *"wheelbase"*, que indica la distancia entre los ejes delantero y trasero del vehículo, y *"carheight"* que indica su altura. En ellos podemos observar como sería un bagplot en situación de dependencia positiva y mayor presencia de outliers y como sería en situación de independencia.

Figura 6.1: Bagplot datos *cars*



Otra aplicación interesante para la detección de outliers aparece propuesta por Mosler (2013), quien propone una relación inversa entre una función de profundidad cualquiera $D(x;P)$ y una "outlyingness function" $Out(x;P)$. Para ello propone transformar el grado de profundidad en una medida del "outlyingness", o grado de outlier de un punto de la siguiente forma:

$$Out(x;P) = \frac{1}{D(x;P)} - 1$$

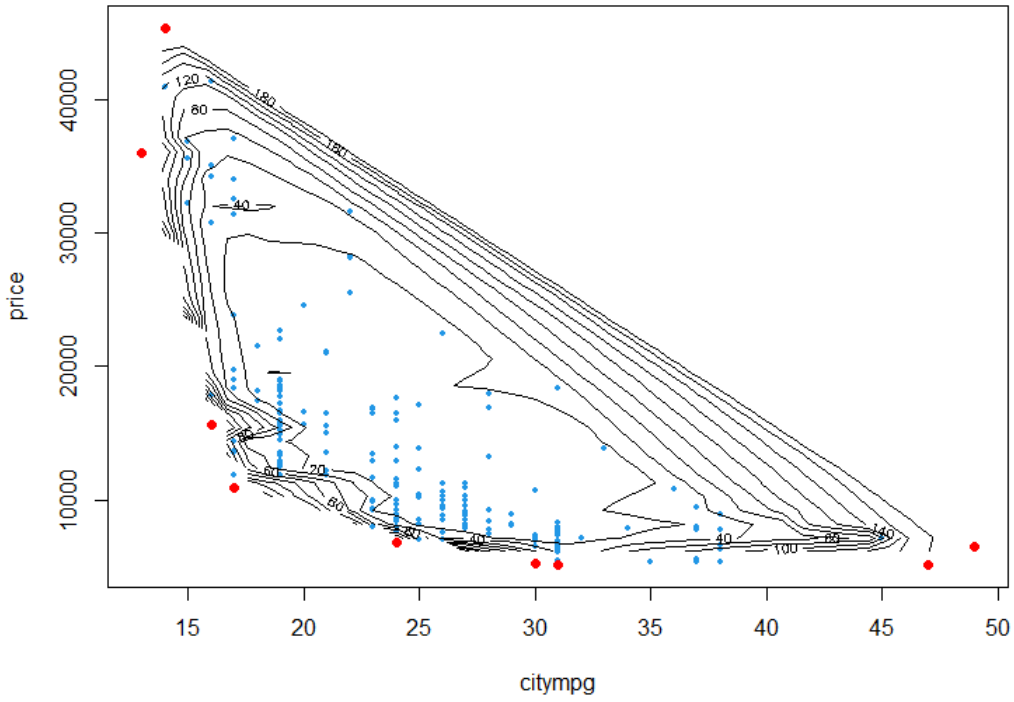
Es trivial ver entonces ver que, para una distribución de probabilidad o conjunto de datos con profundidad máxima α_{max}, la función $Out(x;P)$ está contenida en el intervalo $[\frac{1}{\alpha_{max}} - 1, \infty)$. En la sección de las propiedades generales enunciábamos que una función de profundidad cualquiera estará contenida de forma general en el intervalo $[0,1]$ por lo que el intervalo resultante sería $Out(x;P) \in [0,\infty)$. En cualquier caso, ya hemos visto como para medidas de profundidad concretas y bajo determinados supuestos podemos acotar α_{max} en mayor medida. Por ejemplo, la profundidad semiespacial para distribuciones angularmente simétricas alcanza su máximo en $\alpha = 0.5$, por lo que el mínimo de $Out(x;P)$ sería 1 en lugar de 0.

En cualquier caso, $Out(x;P)$ tomará su valor mínimo en el centro de la distribución y tenderá a infinito cuanto más nos alejemos. De esta forma, podemos relacionarlo con las regiones de profundidad

y establecer que las observaciones fuera de la región D_α tienen un grado de "outlyingness" de $\frac{1}{\alpha} - 1$ y podrían ser considerados como outliers de nivel α. Es trivial ver, además, que al no ser la función más que una transformación de la función de profundidad, $Out(x; P)$ cumplirá las propiedades que cumpla la profundidad utilizada para su cálculo.

Ilustrado de nuevo con el conjunto de datos *cars*, podemos ver los valores de $Out(x; P)$ calculados a través de la profundidad semiespacial representados gráficamente junto sus contornos en la Figura 6.2. En este caso, los datos que alcanzan los valores más elevados son aquellos que se encuentran marcados en rojo en el scatterplot y que incluyen 3 de los detectados anteriormente, además de varios nuevos por la frontera izquierda del conjunto. En este caso, de nuevo, no parece que nuestros outliers sean individuos con comportamientos anómalos, sino mediciones extremas en las variables. Teniendo en cuenta ambas medidas, parece que los individuos que más deberíamos tener en cuenta a la hora de trabajar con el conjunto de datos son los dos que presentan los mayores valores de mpg en ciudad y el vehículo con el mayor precio.

Figura 6.2: Outlyingness *cars*



6.2. DD-Plot

Otro ejemplo lo obtenemos en Liu, Parelius et al. (1999), quienes proponen el DD-plot como una herramienta gráfica para comparar distribuciones conceptualmente similar al gráfico cuantil-cuantil o qq-plot del entorno unidimensional.

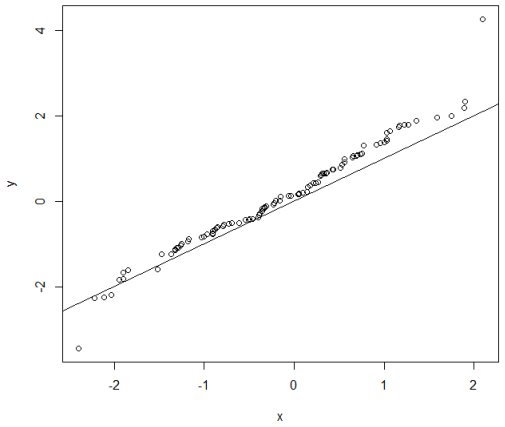
En el entorno univariante podemos utilizar el qq-plot para comparar gráficamente dos distribuciones mediante la representación conjunta de sus cuantiles. Dadas dos distribuciones de probabilidad, dos

conjuntos de datos o incluso una distribución y un conjunto, calculamos los cuantiles de ambos y los representamos cada uno en un eje. Si ambas distribuciones son iguales los cuantiles deberían quedar representados en el gráfico en torno a la bisectriz, mientras que si presentan un patrón alejado de esta podemos afirmar que las distribuciones no coinciden.

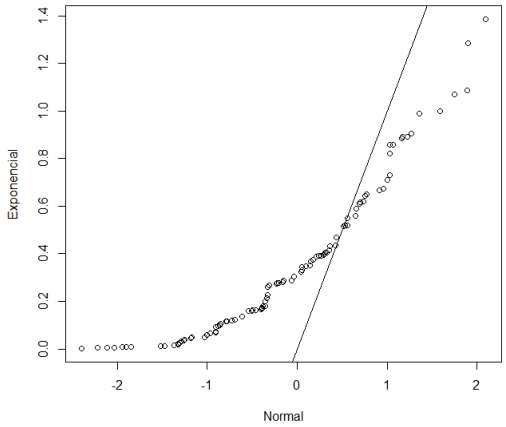
De esta forma, podemos comparar los cuantiles de un conjunto de datos con los de una distribución teórica en particular como una herramienta gráfica para testear si los datos provienen o no de la distribución teórica.

Por ejemplo, en la Figura 6.3 podemos ver los qq-plots resultantes de comparar dos muestras provenientes de la misma distribución, en este caso la normal estándar, y de dos distribuciones distintas, la normal estándar y la Exponencial con parámetro $\lambda = 1$.

Figura 6.3: Ejemplos qq-plot con distribuciones iguales y con distribuciones distintas



(a) Dos muestras de una $N(0,1)$ con $n = 100$



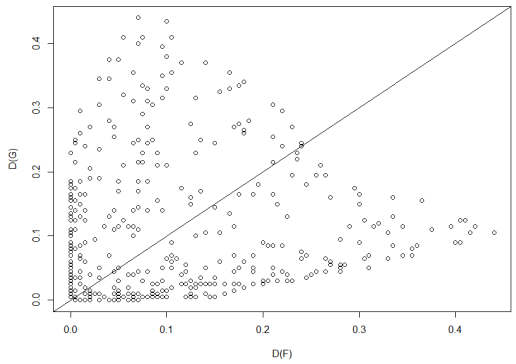
(b) Muestra de una $N(0,1)$ y muestra de una $Exp(1)$ ambas con $n = 100$

A partir de aquí, podemos trazar la equivalencia con el DD-plot, de *"Depth-Depth plot"*, que no sería más que repetir el mismo procedimiento pero, para poder comparar distribuciones en $\mathbb{R}^d, d > 1$, sustituyendo los cuantiles por la profundidad con respecto a cada distribución.

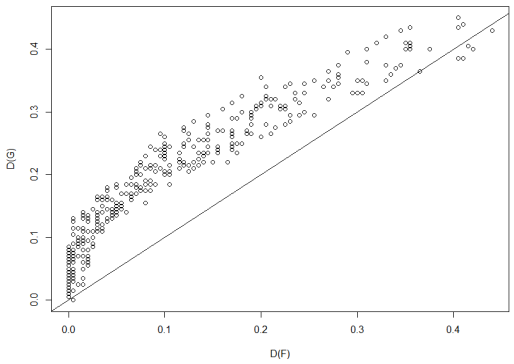
Dadas dos distribuciones de probabilidad, que podemos denotar como F y G, definimos el DD-Plot como el resultado de representar gráficamente el resultado de aplicar una medida de profundidad a ambas distribuciones. Formalmente, $DD(F,G) = \{(D(x;F), D(x;G)) \text{ para todo } x \in \mathbb{R}^d\}$.

En el gráfico resultante, perteneciente a $\mathbb{R}^2$ independientemente de la dimensión de los datos originales, representamos en cada eje una de las dos profundidades. Si ambas distribuciones son idénticas, entonces las observaciones se acumularán en la bisectriz, y otros patrones indicarán diferencias entre ellas. Entre otros, Liu, Parelius et al. (1999) plantean que si la diferencia entre ambas distribuciones es de localización, el DD-plot resultante presentará dispersión en torno a la bisectriz, mayor cuanto más nos alejemos del origen. En cambio, si la diferencia es en la escala de la dispersión, la nube de puntos resultantes formará un arco que se aleja de la bisectriz. Ver Figura 6.4 para ver los DD-plot comparando una $N_2((0,0), I_2)$ frente a una $N_2((1,1), I_2)$ (Subfigura a)) y una $N_2((0,0), I_2)$ frente a una $N_2((0,0), \left(\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right))$ (Subfigura b)). En su artículo, Liu, Parelius et al. (1999) también investigan los efectos de cambios en simetría y kurtosis, entre otros resultados.

Figura 6.4: DD-plots diferencia en localización y diferencia en escala



(a) DD-plot Dos normales con distinta localización



(b) DD-plot Dos normales con distintas escala

6.2.1. Contraste de hipótesis

El DD-plot puede ser utilizado como una herramienta gráfica para contrastes de hipótesis multivariantes. Concretamente podemos, por ejemplo, testear si una muestra $X_n = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$ proveniente de una distribución desconocida, pero con distribución empírica F_n procede o no de una distribución teórica G a través del DD-plot construido tal que $DD(F_n, G) = \{(D(x; F_n), D(x; G)) \text{ para todo } x \in X_n\}$. De nuevo, si la hipótesis nula es cierta y X_n proviene de G, los puntos es en el DD-plot resultante deberán concentrarse en torno a la bisectriz.

De la misma forma, dadas dos muestras $X_n = \{x_1, \ldots, x_n\}$ e $Y_n = \{y_1, \ldots, y_n\}$ con sendas distribuciones empíricas F_n y G_n, podemos construir $DD(F_n, G_n) = \{(D(z; F_n), D(z; G_n)) \text{ para todo } z \in \{X_n \cup Y_n\}\}$ para testear si ambas muestras proceden o no de una misma distribución subyacente.

Aplicado a nuestro dataset de vehículos, podemos comparar nuestro conjunto de datos compuesto por las variables precio y citympg, X, con una $N_2(\overline{X}, S_X)$, como una primera forma visual de testear normalidad, siendo $\overline{X}$ el vector de medias del conjunto y S_X su matriz de varianzas-covarianzas. En la Figura 6.5 podemos ver el resultado utilizando las profundidades semiespacial, simplicial y de Mahalanobis.

En este caso, no parece razonable asumir normalidad para nuestros datos. Mediante el DD-plot realizado con la profundidad semiespacial podríamos llegar a aceptarla, pero los realizados mediante profundidad simplicial o de Mahalanobis presentan patrones claros de rechazo, por lo que parece lógico rechazar la hipótesis nula en este caso.

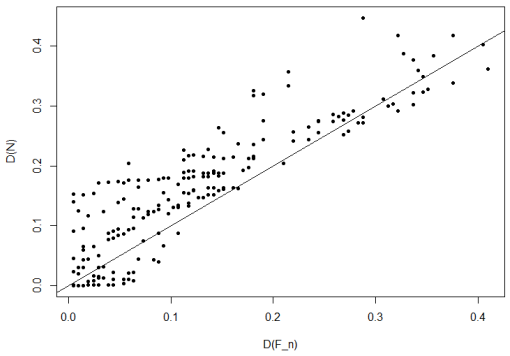
6.2.2. Clasificación

A partir del DD-plot, Li et al. (2012) desarrollan el DD-classifier, que no es más que una extensión para su uso en clasificación. Dadas dos muestras $X_n = \{x_1, \ldots, x_n\}$ e $Y_n = \{y_1, \ldots, y_n\}$ con distribuciones empíricas F_n y G_n, el DD-Classifier busca obtener el polinomio de grado k y origen en el $(0, 0)$ que mejor permita separar las observaciones procedentes de cada muestra en el gráfico DD-plot.

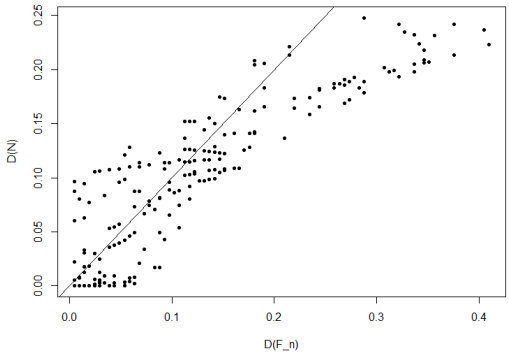
Sea $r_a(x) = \sum_{i=1}^{k_0} a_i x^i$ el polinomio, k_0 el grado preestablecido y $a = (a_1, \ldots, a_{k_0})$ los coeficientes del polinomio, buscamos el vector a que minimice el tasa de clasificación incorrecta derivada de la siguiente regla de clasificación:

$$\begin{cases} D(x, G_n) > r_a(D(x, F_n)) \implies \text{asignar } x \text{ a } G \\ D(x, G_n) < r_a(D(x, F_n)) \implies \text{asignar } x \text{ a } F \end{cases} \tag{6.1}$$

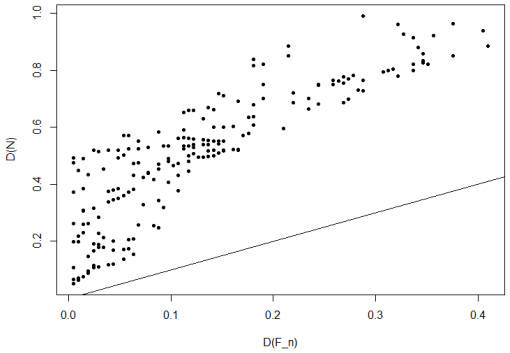
Figura 6.5: DD-plots Contraste normalidad



(a) semiespacial



(b) simplicial



(c) Mahalanobis

A la hora de aplicar el DD-classifier nos enfrentamos a la elección de dos parámetros fundamentales, la medida de profundidad a utilizar y el grado del polinomio. En cuanto a la profundidad los autores sugieren elegir aquella que mejor se adapte a los objetivos de nuestro estudio o en su defecto utilizar Validación Cruzada *(CV)* para seleccionar aquella que ofrezca la menor tasa de clasificación incorrecta. En particular sugieren la profundidad semiespacial como una opción razonable en la mayoría de situaciones.

Por otra parte, para seleccionar el grado del polinomio también recomiendan utilizar técnicas de Validación Cruzada para elegir el que ofrezca mejores resultados.

En este caso, utilizaremos el conjunto de datos *wine*, y trataremos de clasificarlos en buenos o malos (variable quality ≤ 5 ó quality ≥ 6) en función de su grado de alcohol y su nivel de ácido acético. Utilizaremos la profundidad semiespacial y seleccionaremos el grado del polinomio separador mediante CV. En la Figura 6.6 podemos observar la nube de puntos de los vinos en función de las dos variables elegidas y con el color señalizando su calidad. Podemos ver como, sobre todo en el centro, están bastante entremezclados, lo que podría penalizar el rendimiento del clasificador.

Para ello dividiremos las 6497 observaciones del conjunto en dos muestras de forma aleatoria. El 80 % de las observaciones conformarán la muestra de entrenamiento y se utilizarán para entrenar el clasificador. El 20 % restante será la muestra de test con la que comprobaremos su eficacia. Puede verse el DD-plot resultante en la Figura 6.7.

Figura 6.6: Scatter plot Datos *wine*

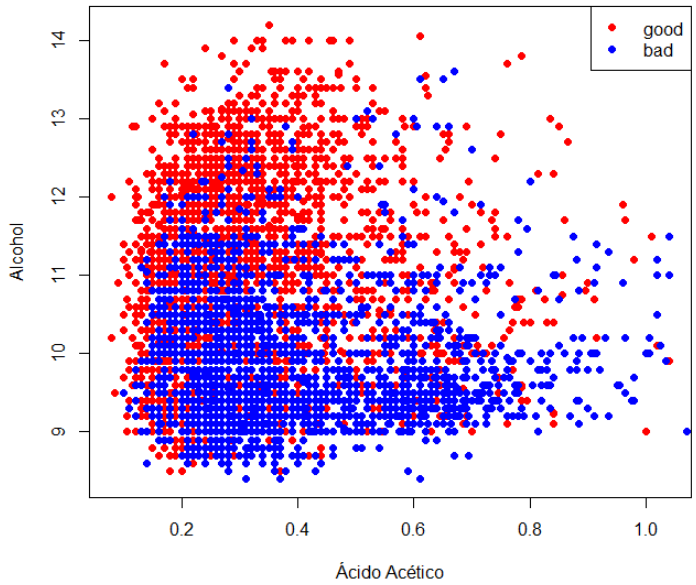
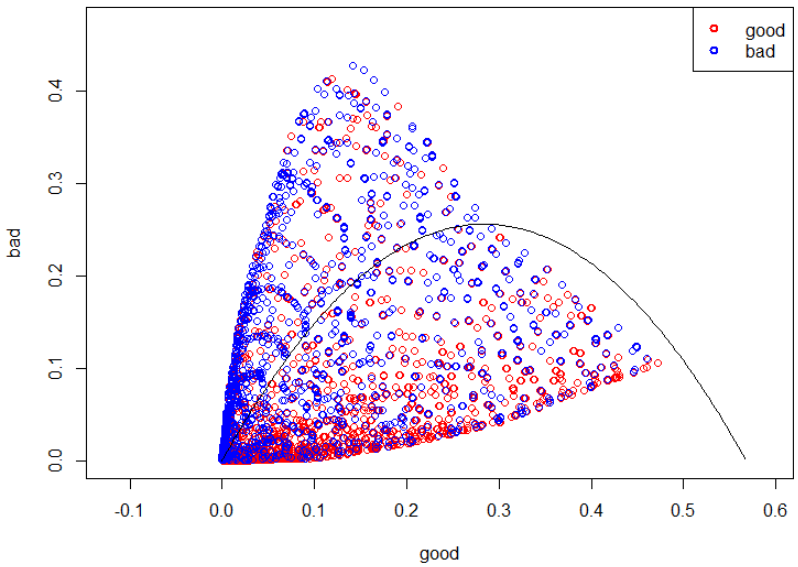


Figura 6.7: DD-Classifier datos *wine* con 2 variables y HD

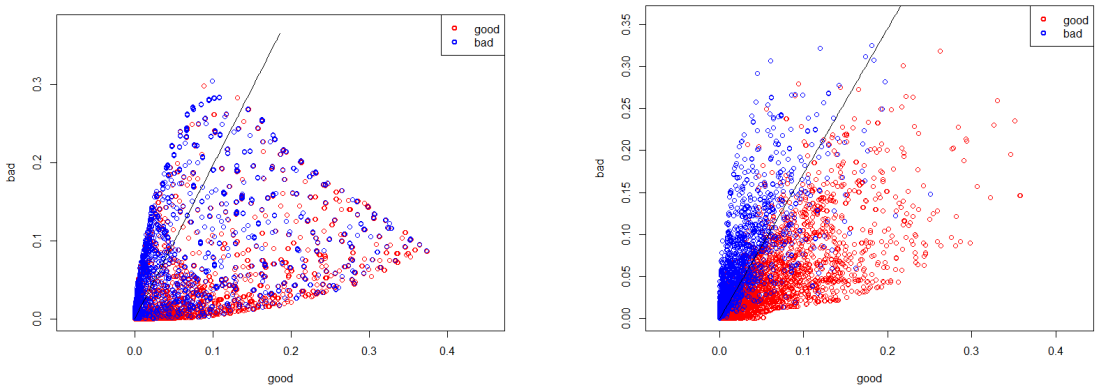


Podemos observar en el gráfico la regla de clasificación generada con la muestra de entrenamiento

y podemos observar que hay un porcentaje de la muestra que está bastante entremezclado en el centro del DD-plot, dificultando obtener un buen polinomio separador. El seleccionador por CV ha obtenido que la menor tasa de clasificación incorrecta se obtiene con un polinomio de grado 2 y, al aplicarlos a la muestra de test obtenemos un valor de dicha tasa de 26.769 %, que no es un valor demasiado bueno, por lo que sería conveniente buscar otro método de clasificación o nuevas variables explicativas. Podríamos por ejemplo, seleccionar la profundidad de Mahalanobis, lo cual nos seleccionaría un polinomio de grado 5 y elevaría la tasa de clasificación incorrecta al 28 %. Podemos ver el DD-plot resultante de este segundo intento en la Figura 6.8 Subfigura a).

Por otra parte, si en lugar de cambiar la profundidad, añadiésemos todas las demás variables presentes en el conjunto de datos (como indicábamos al principio del capítulo estas variables son características químicas de los vinos como por ejemplo la densidad, el pH o el nivel de sulfatos entre otras), obtendríamos como clasificador un polinomio de grado 1 (Figura 6.8 Subfigura b)) y una tasa de clasificación incorrecta de 8.54 %, mejorando notablemente con respecto a los casos anteriores.

Figura 6.8: DD-plots diferencia en localización y diferencia en escala



(a) DD-Classifier datos *wine* con 2 variables y M_hD (b) DD-Classifier datos *wine* dataset completo y HD

A partir de dicho gráfico podemos explicar la regla de clasificación de la siguiente forma, si calculamos la profundidad de una nueva observación con respecto al conjunto de vinos de mala calidad y con respecto al conjunto de vinos de buena calidad y el resultado cae a la izquierda del polinomio lo asignamos al grupo de vinos de mala calidad. Formalmente resultaría en la expresión que mencionábamos anteriormente, $D(x, G_n) > r_a(D(x, F_n))$, donde una observación se asignaría al grupo malo si su profundidad con respecto a ese grupo es mayor que el resultado de evaluar el polinomio en el valor de su profundidad respecto a los vinos buenos.

6.3. Regresión

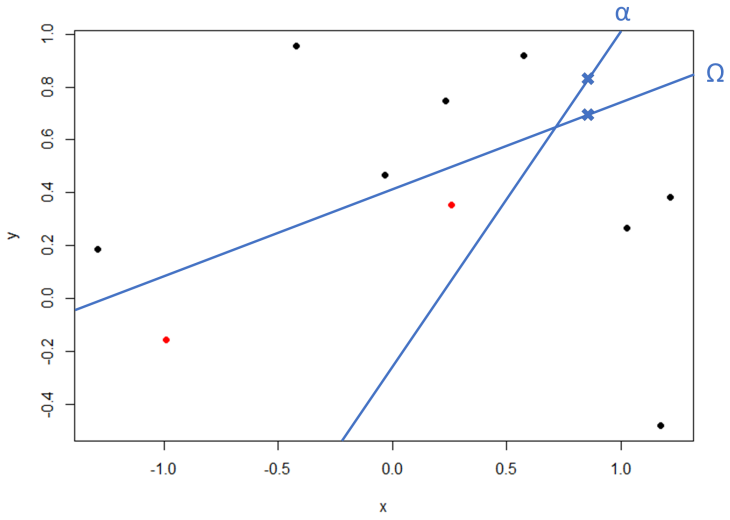
Finalmente, Rousseeuw y Hubert (1999) desarrollan la "regression depth", como una forma de seleccionar los parámetros de una recta de regresión mediante profundidad. Aplicado al entorno bivariante, y dada una regresión simple con coeficientes $\vec{\beta} = (\beta_1, \beta_2)$ y residuos $r_i(\vec{\beta}) = y_i - \beta_1 - \beta_2 x_i$ ajustada a una muestra $Z_n = \{(x_i, y_i); i = 1, \ldots, n\}$, definimos primero el concepto de "nonfit" como aquel conjunto de parámetros $\vec{\beta}$ para los cuales existe un punto v_0 que no coincida con ningún x_i tal que:

$r_i(\vec{\beta}) < 0 \,\forall\, x_i < v_0$ *y* $r_i(\vec{\beta}) > 0 \,\forall\, x_i > v_0$

ó
$r_i(\vec{\beta}) > 0 \,\forall\, x_i < v_0$ y $r_i(\vec{\beta}) < 0 \,\forall\, x_i > v_0$.

Expresado gráficamente, se considera un "nonfit" si existe un punto v_0 sobre el cual podemos rotar la recta de regresión hasta que sea vertical sin atravesar ninguna de las observaciones de la muestra. De esta forma, cualquier recta que se encuentra siempre por encima o por debajo de la nube de puntos será un "nonfit". A partir de aquí desarrollan el concepto de "regression depth", $rdepth(\vec{\beta}, Z_n)$ como el mínimo número de observaciones que hay que eliminar o cambiar de signo para que la recta se convierta en un "nonfit". Es trivial, entonces, ver que $rdepth(\vec{\beta}, Z_n) \in [0, n]$ A raíz de este resultado, se deriva que la recta de regresión que mejor se ajusta a una muestra es aquella que alcance un mayor valor de profundidad. Como podemos ver en la Figura 6.9, la recta de regresión α es un "nonfit", dado que podemos rotarla sobre el punto marcado por la cruz hasta la posición vertical sin atravesar ninguna observación. En cambio, la recta Ω necesita atravesar como mínimo los dos puntos marcados en rojo, por lo que tiene una profundidad de $rdepth(\Omega, Z_n) = 2$.

Figura 6.9: Fit y Nonfit en $\mathbb{R}^2$



A nivel de cálculo, Rousseeuw y Hubert (1999) proponen el siguiente algoritmo en $\mathbb{R}^2$. Dadas una recta de regresión caracterizada por el vector de parámetros $\vec{\beta}$, un conjunto de datos $Z_n = \{(x_i, y_i)\}$ y unos residuos derivados de dicho ajuste $r_i(\vec{\beta})$, calculamos la profundidad del modelo $rdepth(\vec{\beta}, Z_n)$ de la siguiente forma. Primero ordenamos la muestra de forma que $x_1 \leq x_2 \leq \cdots \leq x_n$, y obtenemos:

$$rdepth(\vec{\beta}, Z_n) = \min_{1 \leq i \leq n} (\min\{L^+(x_i) + R^-(x_i); R^+(x_i) + L^-(x_i)\}).$$

En la expresión anterior, $L^+(x_i) = \#\{j : x_j \leq x_i \,\text{y}\, r_j \geq 0\}$ sería el número de puntos a la izquierda de x_i por encima de la recta de regresión y $R^-(x_i) = \#\{j : x_j > x_i \,\text{y}\, r_j \leq 0\}$ el número de puntos a la derecha de x_i por debajo de la recta. Así, podemos ver $L^+(x_i) + R^-(x_i)$ como el número de observaciones que atravesaríamos si rotamos la recta de regresión hasta la posición vertical sobre x_i en sentido horario. Observar que L y R indican si contamos a la izquierda o derecha de x_i *(left/right)*, y los exponentes $^+$ y $^-$ si consideramos los de residuo positivo negativo respectivamente.

De esta forma, $L^-(x_i)$ y $R^+(x_i)$ se definen como el inverso de los casos anteriores, $L^-(x_i)$ son las observaciones a la izquierda que caen por debajo de la recta de regresión y $R^+(x_i)$ las observaciones a la derecha de x_i que caen por encima de la recta de regresión. Consecuentemente, estaríamos computando

el inverso del caso anterior, es decir, las observaciones que atravesaríamos si rotamos la recta de regresión hasta la posición vertical sobre x_i en sentido antihorario.

Calculamos el mínimo entre ambos pares para cada i y la profundidad del ajuste será el mínimo global. De esta forma, estamos calculando, evaluado en cada observación, el mínimo número de observaciones que la recta de regresión atravesaría hasta estar en posición vertical si la rotamos tanto en sentido horario como antihorario.

A continuación, postulan la siguiente igualdad, de interés para el cálculo de la profundidad:

$$\max_{\vec{\beta}} rdepth(\vec{\beta}, Z_n) = \max_{i<j} rdepth(\vec{\beta}^{ij}, Z_n).$$

Donde $\vec{\beta}^{ij}$ es la recta que une dos observaciones x_i y x_j, por lo que podemos limitar el cálculo de la profundidad solo a aquellos conjuntos de parámetros que ofrezcan rectas que atraviesen dos observaciones de la muestra, en vez de todas las posibles combinaciones de parámetros.

A raíz de este resultado definen el estimador de la regresión más profunda, $T_r^*(Z_n)$, como aquel ajuste $\vec{\beta}$, que maximice la expresión $rdepth(\vec{\beta}, Z_n)$. Formalmente:

$$T_r^*(Z_n) = \operatorname*{argmax}_{\vec{\beta}} rdepth(\vec{\beta}, Z_n) = \operatorname*{argmax}_{\vec{\beta}^{ij}} rdepth(\vec{\beta}^{ij}, Z_n).$$

Adicionalmente, estipulan que si el conjunto de parámetros $\vec{\beta}^{ij}$ que maximiza la expresión anterior es no singular y existen más de un conjunto de parámetros en lo que se maximiza la profundidad, se selecciona el promedio de dichos valores.

Finalmente, generalizan el estimador al caso d-dimensional, tal que para una muestra $Z_n = \{(x_{i,1}, \ldots, x_{i,d-1}, y_i), i = 1, \ldots, n\} \subset \mathbb{R}^d$, obtengamos una regresión lineal múltiple en la forma del hiperplano $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{d-1} x_{i,d-1}$.

En este caso, se considera que un conjunto de parámetros $\vec{\beta}$ es un "nonfit" si en uno de los dos semiespacios generados por él, todos los residuos son positivos y en el otro son todos negativos. En consecuencia, la profundidad de la regresión en el caso $d > 2$ se define de la misma forma. Sea el hiperplano dado por el conjunto de parámetros $\vec{\beta}$, podemos denotar los dos semiespacios en los que divide el espacio como $L(\vec{\beta})$ y $R(\vec{\beta})$ al igual que en el caso anterior, $L^+(\vec{\beta})$ serán las observaciones en dicho semiespacio con residuo positivo y $L^-(\vec{\beta})$ las que tienen residuos negativos, construyéndose $R^+(\vec{\beta})$ y $R^-(\vec{\beta})$ de la misma forma. Así pues, la profundidad de la regresión será

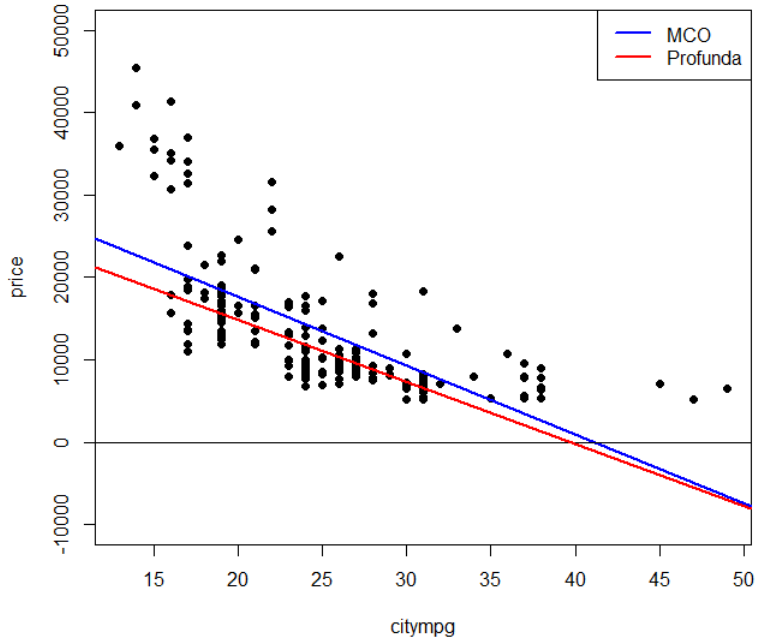
$$rdepth(\vec{\beta}, Z_n) = \min_{\vec{\beta}}(\min\{L^+(\vec{\beta}) + R^-(\vec{\beta}); R^+(\vec{\beta}) + L^-(\vec{\beta})\}),$$

y nuestro estimador aquel conjunto $\vec{\beta}$ que maximice la expresión anterior.

Podemos aplicar esta metodología al caso del conjunto de datos de vehículos, donde ya habíamos detectado previamente la presencia de valores atípicos. Los ajustes por Mínimos Cuadrados Ordinarios tradicionales son comúnmente sensibles a la presencia de outliers por lo que una metodología diferente puede resultar de utilidad. En la Figura 6.10 podemos ver el ajuste por mínimos cuadrados obtenido y el de máxima profundidad.

En este caso vemos que la regresión lineal clásica se ve empujada hacia arriba por un pequeño grupo de datos con valores más alejados del origen que la mayoría, incluyendo el grupo de 5 outliers detectado previamente. En cambio la regresión que maximiza la profundidad es menos sensible a los datos influyentes ya que trata de ajustarse lo máximo posible al centro de la nube de puntos. En el caso de la regresión lineal tendríamos $\hat{y}_i = 34395.4404 - 837.3964 x_i$, mientras que en la regresión profunda sería $\hat{y}_i = 29946.6826 - 755.4697 x_i$, siendo como decíamos la coordenada en el origen la mayor diferencia. En cambio, vemos que ninguno de los dos modelos es capaz de ofrecer predicciones de precios para vehículos con valores de citympg superiores a 40, donde en ambos casos se predicen valores de precios negativos.

Figura 6.10: Regression Depth y Regresión por MCO



Apéndice A

Figuras

Figura A.1: Contornos de Profundidad $M_h D$ Muestrales de $N_2(0, I_2)$ con $n = 200$.

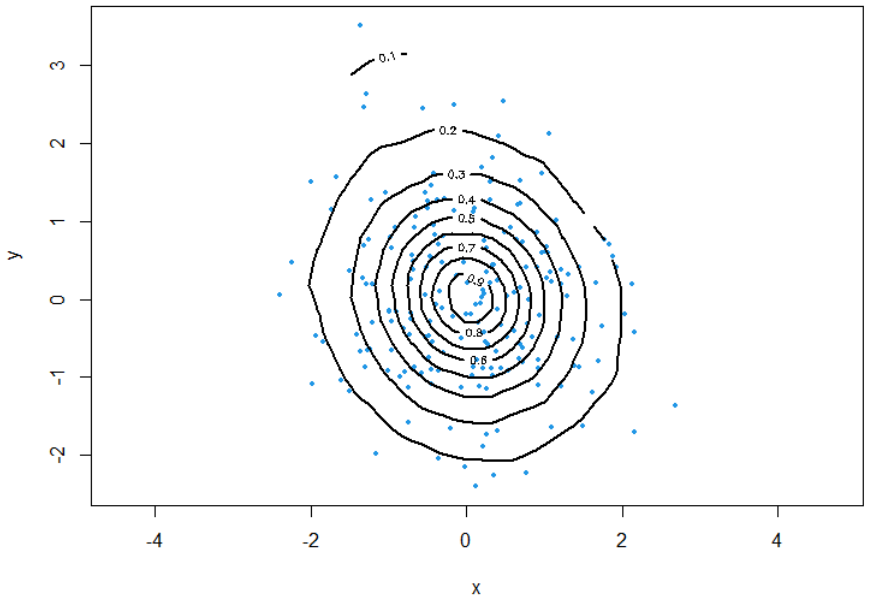


Figura A.2: Regiones de Profundidad M_hD Muestrales de $N_2(0, I_2)$ con $n = 200$.

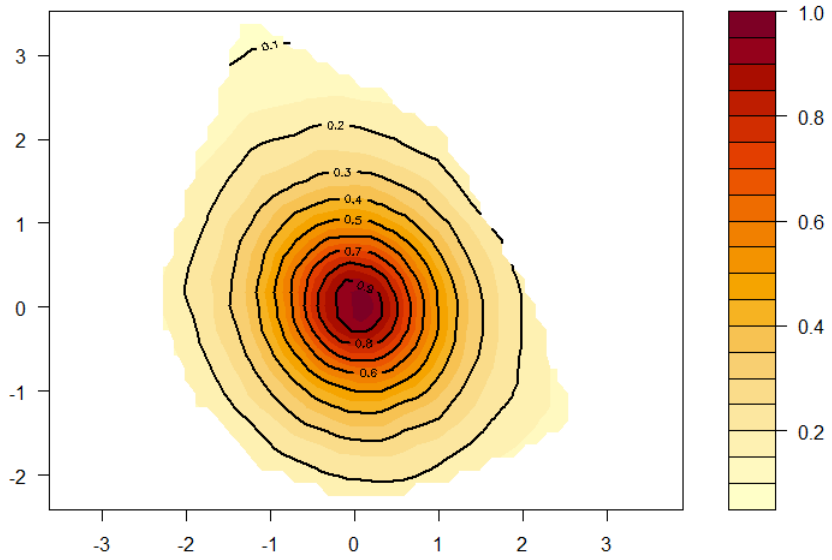


Figura A.3: Comparación Contornos de Profundidad M_hD Muestrales de $N_2(0, I_2)$ con $n = 200$.

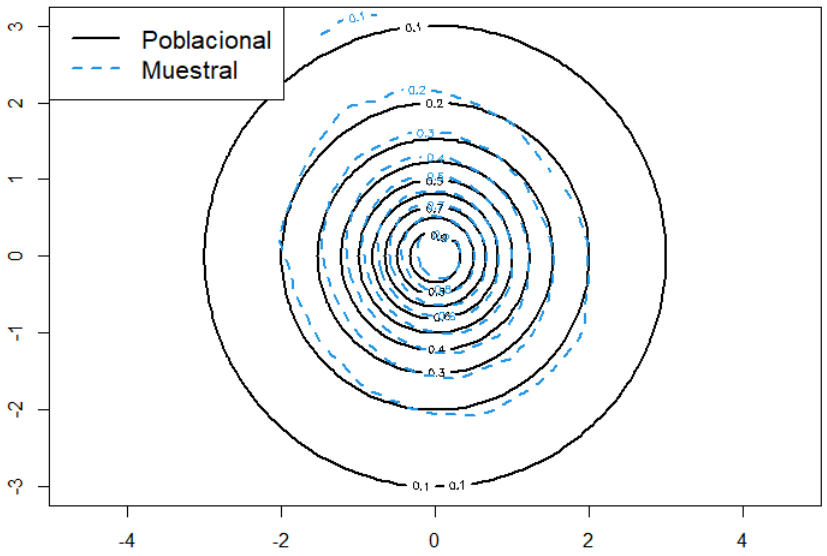


Figura A.4: Contornos de Profundidad M_hD Muestrales de $N_2(0, I_2)$ con $n = 500$.

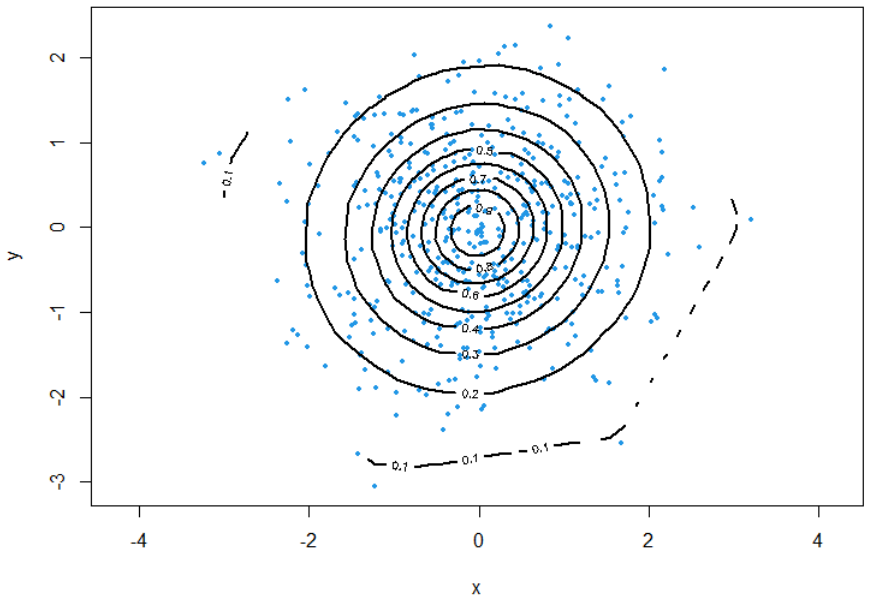


Figura A.5: Regiones de Profundidad M_hD Muestrales de $N_2(0, I_2)$ con $n = 500$.

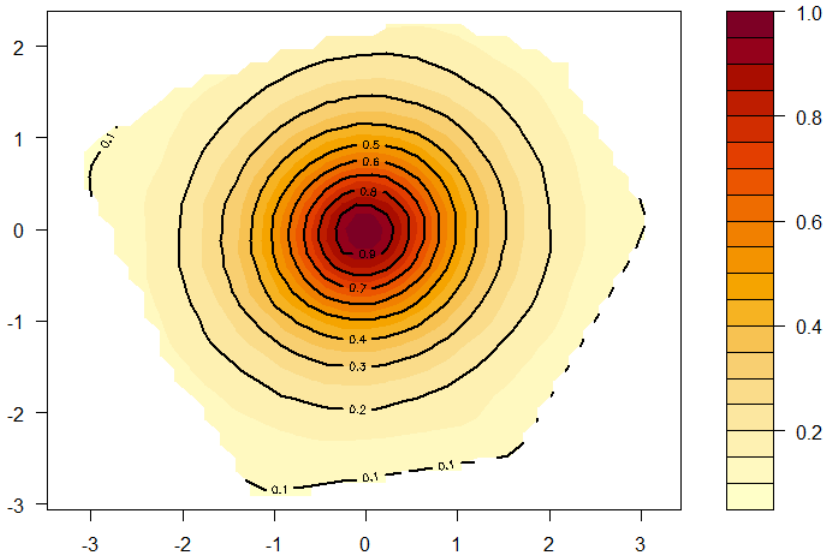


Figura A.6: Comparación Contornos de Profundidad M_hD Muestrales de $N_2(0, I_2)$ con $n = 500$.

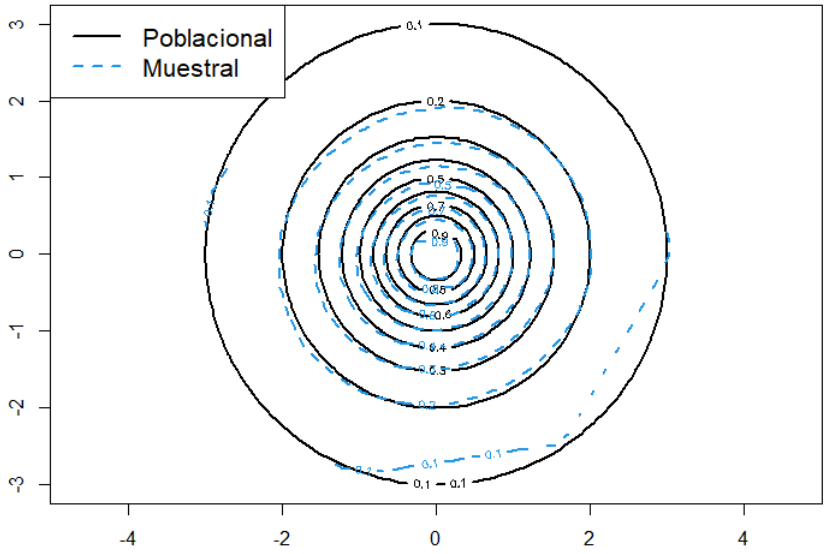


Figura A.7: Contornos de Profundidad HD Muestrales de $N_2(0, I_2)$ con $n = 200$.

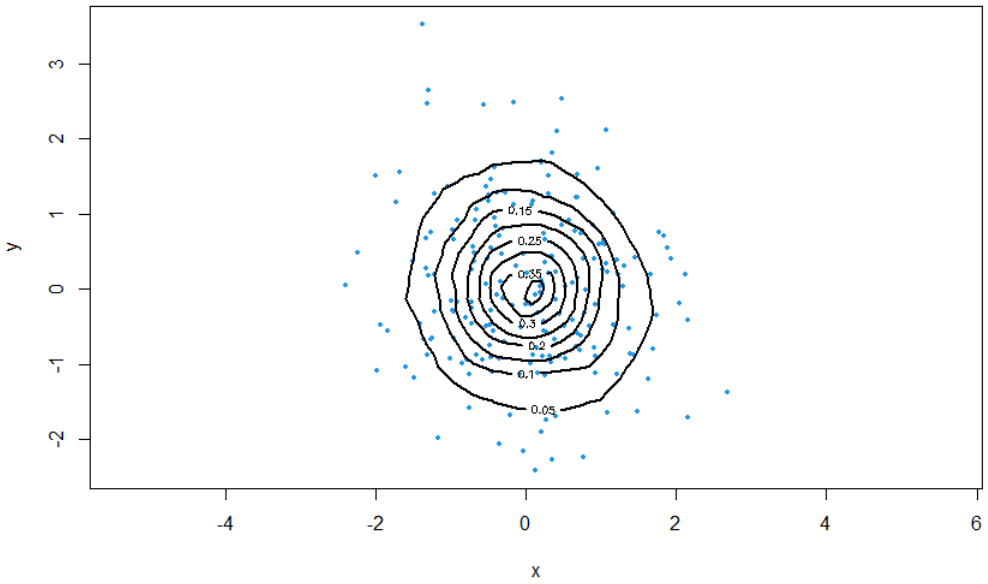


Figura A.8: Regiones de Profundidad HD Muestrales de $N_2(0, I_2)$ con $n = 200$.

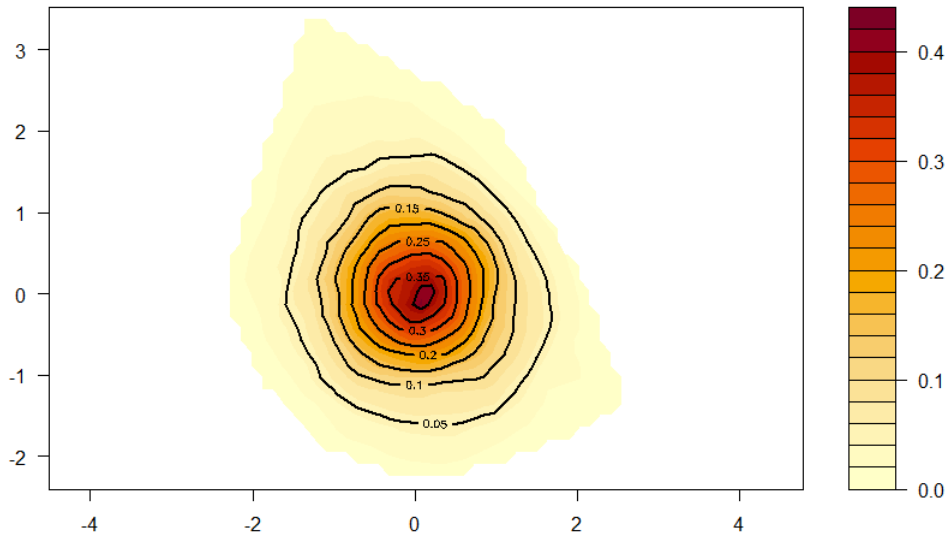


Figura A.9: Comparación Contornos de Profundidad HD Muestrales de $N_2(0, I_2)$ con $n = 200$.

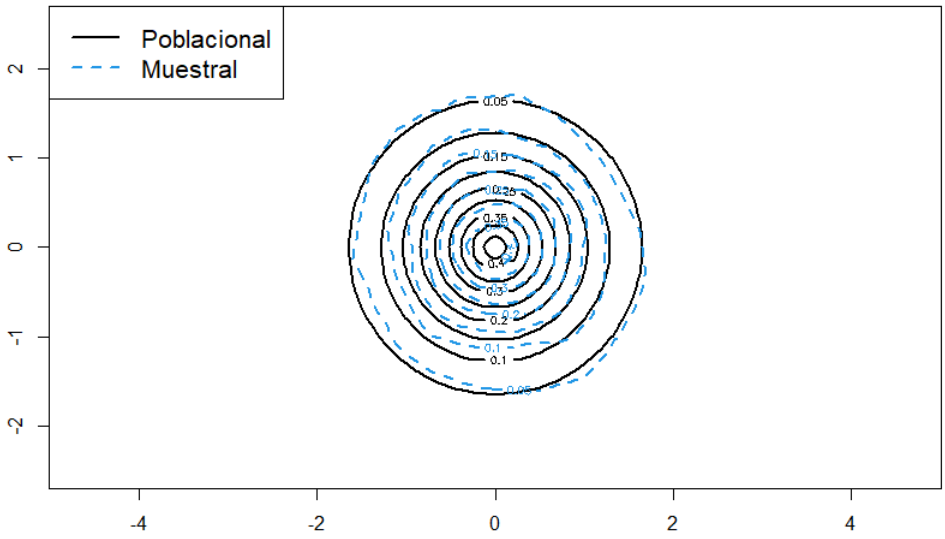


Figura A.10: Contornos de Profundidad HD Muestrales de $N_2(0, I_2)$ con $n = 500$.

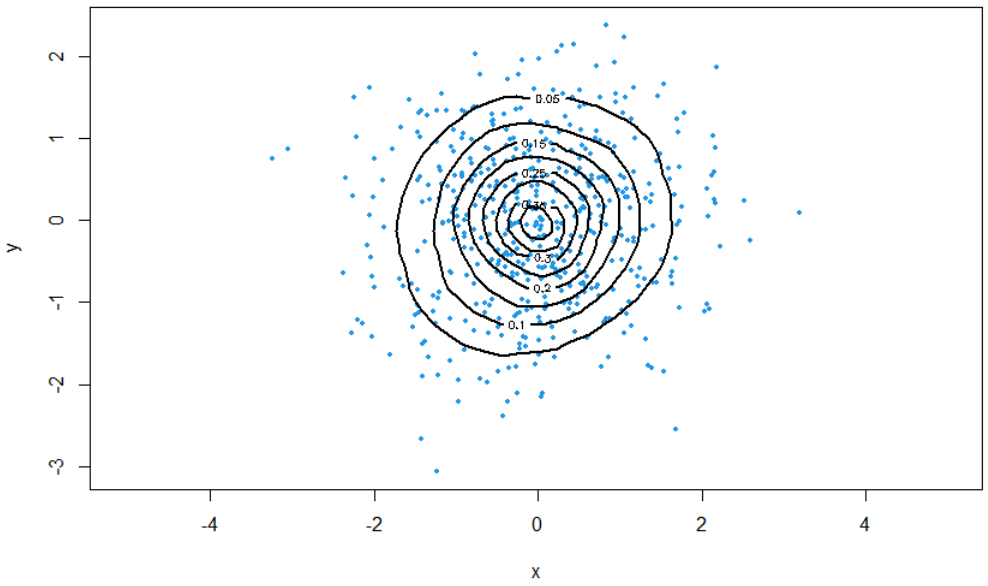


Figura A.11: Regiones de Profundidad HD Muestrales de $N_2(0, I_2)$ con $n = 500$.

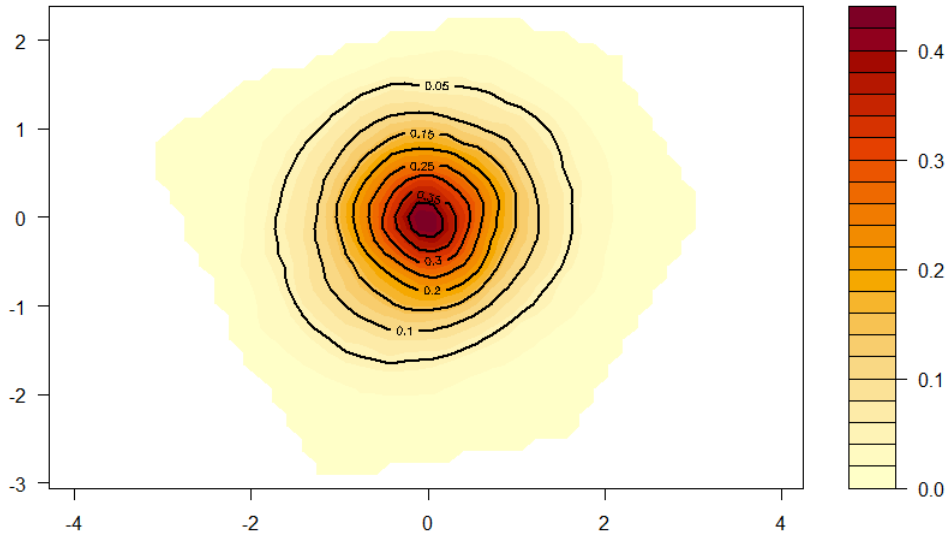


Figura A.12: Comparación Contornos de Profundidad HD Muestrales de $N_2(0, I_2)$ con $n = 500$.

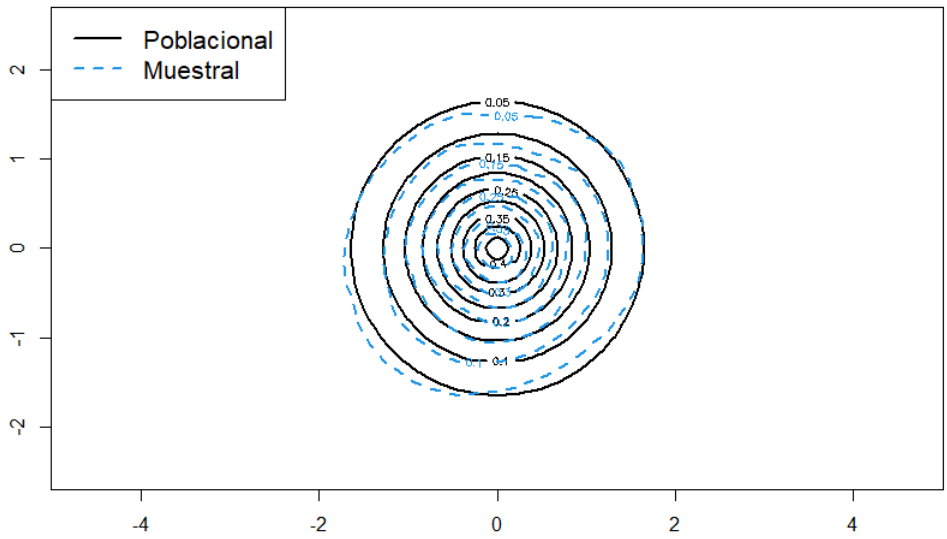


Figura A.13: Contornos de Profundidad SD Muestrales de $N_2(0, I_2)$ con $n = 200$.

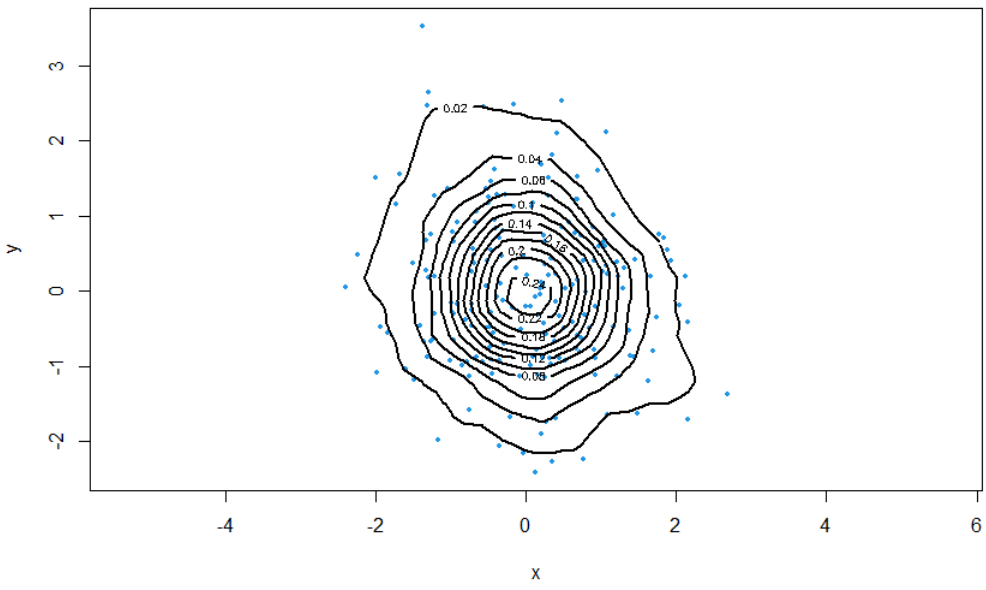


Figura A.14: Regiones de Profundidad SD Muestrales de $N_2(0, I_2)$ con $n = 200$.

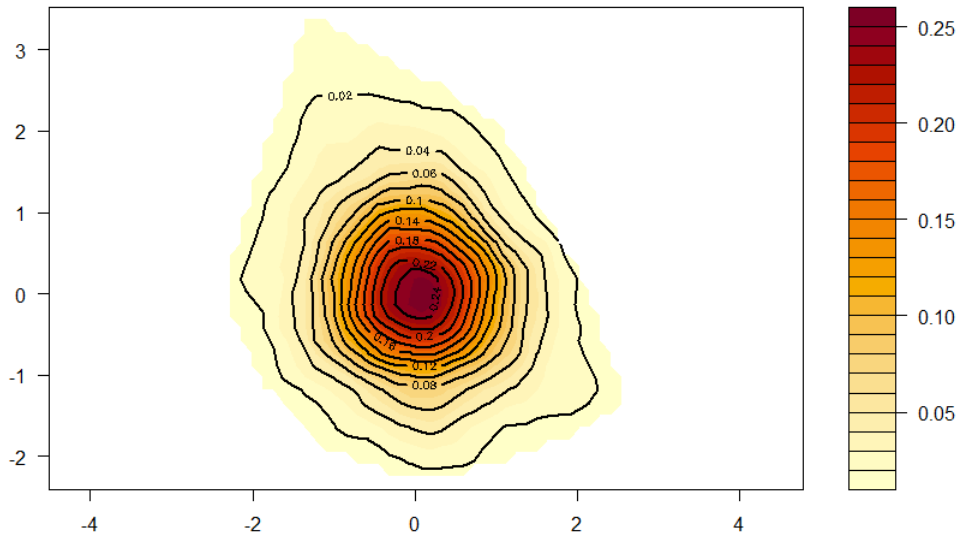


Figura A.15: Contornos de Profundidad SD Muestrales de $N_2(0, I_2)$ con $n = 500$.

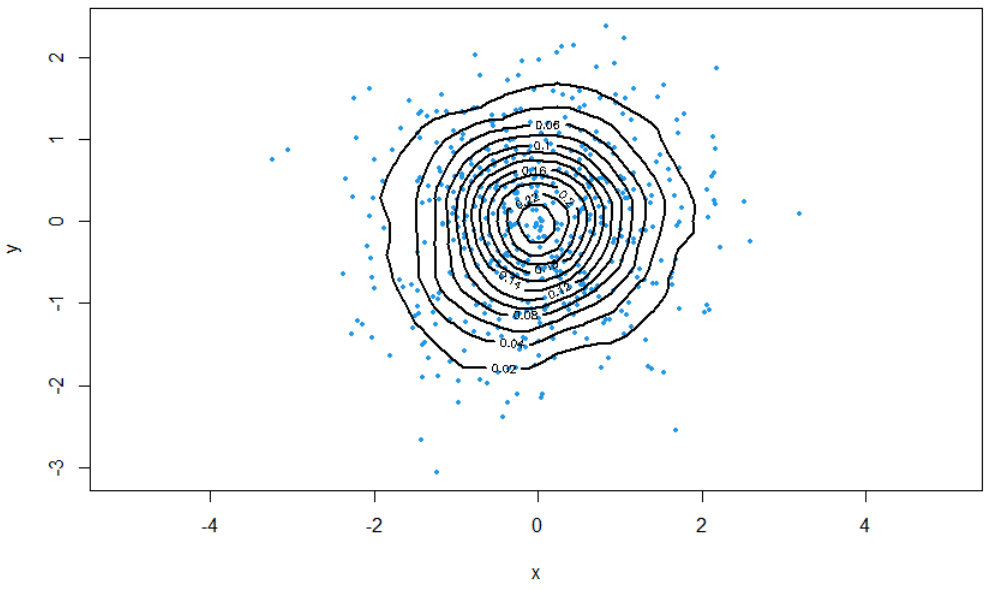


Figura A.16: Regiones de Profundidad SD Muestrales de $N_2(0, I_2)$ con $n = 500$.

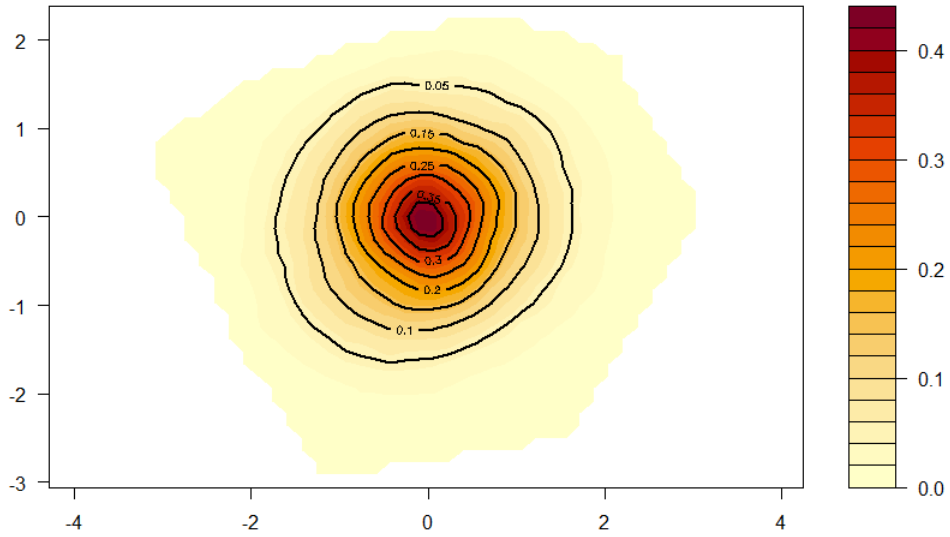


Figura A.17: Mixtura Equiprobable

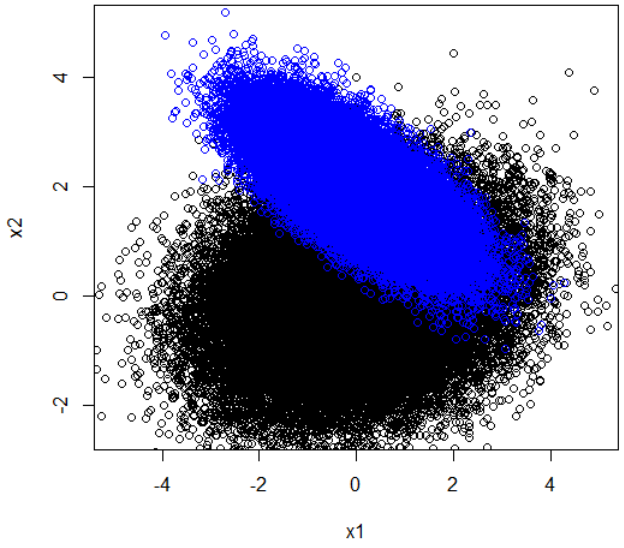


Figura A.18: Bagplot datos *cars* precio-distancia entre ejes

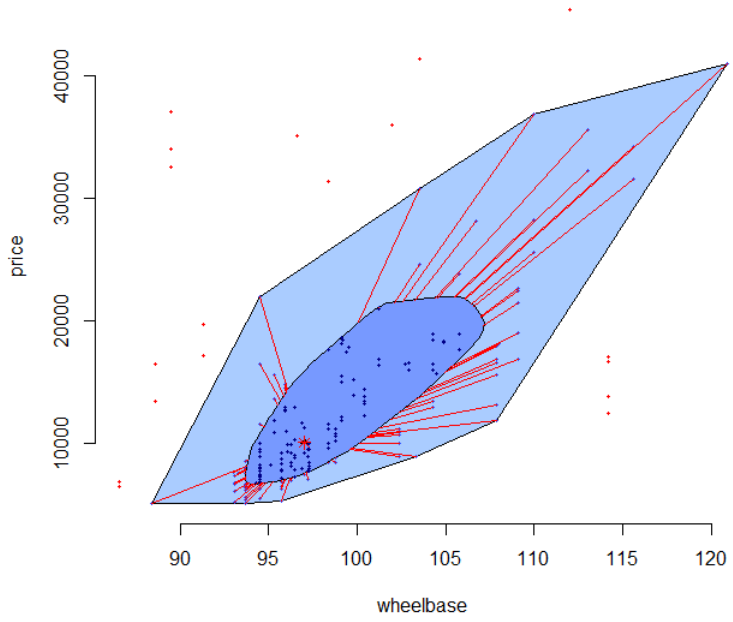
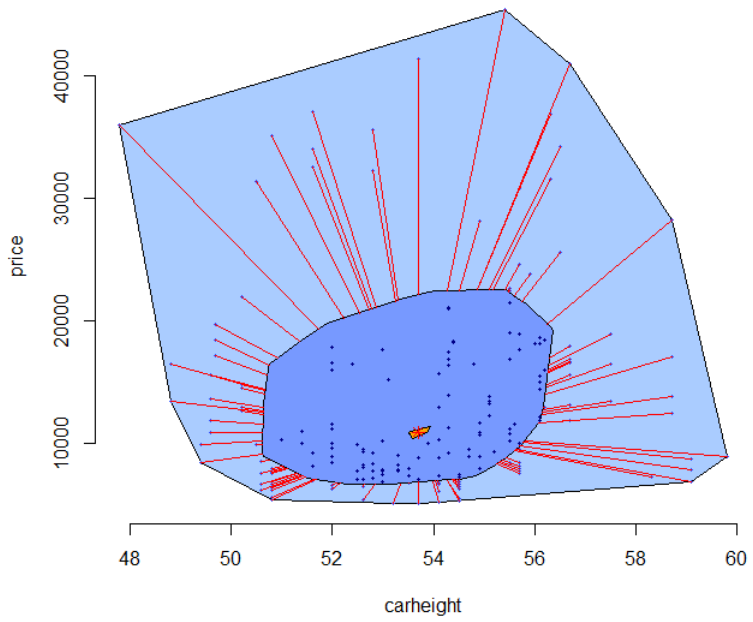


Figura A.19: Bagplot datos *cars* precio-altura



Bibliografía

Afshani, Peyman, Donald R. Sheehy y Yannik Stein (2015). "Approximating the Simplicial Depth". En: URL: http://arxiv.org/abs/1512.04856.

Burr, Michael A., Eynat Rafalin y Diane L. Souvaine (2006). "Simplicial depth: An improved definition, analysis, and efficiency for the finite sample case". En: *Dimacs Series in Discrete Mathematics and Theoretical Computer Science* 72, pág. 195.

Cascos, Ignacio, Angel López y Juan Romo (2011). "Data depth in Multivariate Statistics". En: *Boletín de Estadística e Investigación Operativa* 27.3, págs. 151-174.

Cheng, Andrew Y. y Mingqing Ouyang (2001). "On algorithms for simplicial depth". En: *Canadian Conference on Computational Geometry*.

Cortez, Paulo et al. (2009). *Wine Quality*. UCI Machine Learning Repository. DOI: 10.24432/C56S3T.

Cuesta Albertos, Juan Antonio y Alicia Nieto Reyes (2008). "The random Tukey depth". En: *Computational Statistics & Data Analysis* 52.11, págs. 4979-4988.

Donoho, David L. y Miriam Gasko (1992). "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness". En: *The Annals of Statistics* 20.4, págs. 1803-1827.

Dyckerhof, Rainer y Pavlo Mozharovskyi (2016). "Exact computation of the halfspace depth". En: *Computational Statistics & Data Analysis* 98, págs. 19-30.

Hoeffding, Wassily (1948). "A Class of Statistics with Asymptotically Normal Distribution". En: *The Annals of Mathematical Statistics* 19.3, págs. 293-325.

Li, Jun, Juan Antonio Cuesta Albertos y Regina Y. Liu (2012). "DD-Classifier: Nonparametric Classification Procedure Based on DD-Plot". En: *Journal of the American Statistical Association* 107.498, págs. 737-753.

Liu, Regina Y. (1988). "On a notion of simplicial depth". En: *Proceedings of the National Academy of Sciences* 85.6, págs. 1732-4.

Liu, Regina Y. (1990). "On a Notion of Data Depth Based on Random Simplices". En: *Annals of Statistics* 18.1, págs. 405-414.

Liu, Regina Y., Jesse M. Parelius y Kesar Singh (1999). "Multivariate analysis by data depth: descriptive statistics, graphics and inference". En: *Annals of Statistics* 27.3, págs. 783-858.

Liu, Regina Y. y Kesar Singh (1993). "A Quality Index Based on Data Depth and Multivariate Rank Tests". En: *Journal of the American Statistical Association* 88.421, págs. 252-260.

Mosler, Karl (2013). "Depth Statistics". En: *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*. Springer Berlin Heidelberg, págs. 17-34.

Rousseeuw, Peter J. y Mia Hubert (1999). "Regression Depth". En: *Journal of the American Statistical Association* 94.446, págs. 388-402.

Rousseeuw, Peter J. e Ida Ruts (1996). "Algorithm AS 307: Bivariate Location Depth". En: *Journal of the Royal Statistical Society Series C Applied Statistics* 45, págs. 516-526.

Rousseeuw, Peter J. e Ida Ruts (1999). "The depth function of a population distribution". En: *Metrika* 49, págs. 213-244.

Rousseeuw, Peter J., Ida Ruts y John W. Tukey (1999). "The Bagplot: A Bivariate Boxplot". En: *The American Statistician* 53.4, págs. 382-387.

Rousseeuw, Peter J. y Anja Struyf (1998). "Computing location depth and regression depth in higher dimensions". En: *Statistics and Computing* 8.3, págs. 193-203.

Schlimmer, Jeffrey (1987). *Automobile*. UCI Machine Learning Repository. DOI: 10.24432/C5B01C.

Tukey, John W. (1975). "Mathematics and the Picturing of Data". En: *Proceeding of the International Congress of Mathematicians* 2, págs. 523-531.

Zuo, Yijun y Robert Serfling (2000). "General notions of statistical depth function". En: *Annals of Statistics* 28.2, págs. 461-482.