1-INTRODUCCION

En 1974 Tukey introdujo el concepto de **profundidad (depth)** o **semiespacio de profundidad (halfspace depth)** de un punto relativo a un conjunto de datos multivariado. [13,14]

Para un conjunto unidimensional de datos $X = \{X_1, \ldots X_n\}$ se define la profundidad (depth) de un valor x relativo a un conjunto X como el mínimo número de datos a la derecha y a la izquierda de x.[3]

$$depth_1(x, X) = \min(\#\{i : X_i \le x\}, \#\{i : X_i \ge x\}) \quad (1)$$

Ejemplo:

$X = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21\}$

Si queremos calcular $depth(6, X)$, hacemos:

$\#\{i : x_i \le 6\} = \#\{1, 3, 5\} = 3$

$\#\{i : x_i \ge 6\} = \#\{7, 9, 11, 13, 15, 17, 19, 21\} = 8$

Luego $depth(6, X) = \min(3, 8) = 3$

En el caso d-dimensional, Donoho y Gasko [3], definen la profundidad como:

$$depth_d(x, X) = \min_{|u|=1} depth_1(u^T x; \{u^T X_i\}) = \min_{|u|=1} \#\{i : u^T X_i \ge u^t x\} \quad (2)$$

Como $u \in R^d$ es un vector unitario, entonces el conjunto $\{u^T X_i\}$ es la proyección ortogonal del conjunto de datos X sobre un espacio de dimensión 1.

Es decir la profundidad de un punto $x \in R^d$ relativo a un conjunto X de R^d es la menor de las profundidades de x relativas a las proyecciones ortogonales del conjunto de datos X sobre espacios unidimensionales.

Rousseeaw y Ruts [12] la llaman **location depth** y la notan $ldepth(\theta, X)$, es decir

$$\boxed{ldepth(\theta, X) = \min_{|u|=1} \#\{i : u^T X_i \le u^t \theta\}} \quad (3)$$

Como **profundidad (depth)** definen

$$\boxed{depth(\theta, X) = \frac{1}{n} ldepth(\theta, X)} \quad (4)$$

Es decir

$$\boxed{depth\ (x, X) = \frac{1}{n} \min_{|u|=1} \#\{i : u^T X_i \le u^t x\}} \quad (5)$$

Esta es la notación que vamos a usar.

Tukey consideró además los contornos de profundidad (**contours of depth)**, que en realidad son regiones de profundidad para indicar la forma de un conjunto de datos en R^2 y sugirió que a partir del concepto de profundidad (depth) se podría definir un concepto análogo al rango para el caso multivariado.

De hecho en el caso unidimensional (si no hay datos repetidos) el valor mínimo y máximo de la muestra son los puntos con $ldepth = 1$, los cuartiles 1 y 3 son los puntos con $ldepth \approx n/4$ y la mediana tiene $ldepth \approx n/2$.

Si calculamos las profundidades para cada uno de los elementos del ejemplo, obtenemos:

$ldepth(1, X) = ldepth(21, X) = 1$
$ldepth(3, X) = ldepth(19, X) = 2$
$ldepth(5, X) = ldepth(17, X) = 3$
$ldepth(7, X) = ldepth(15, X) = 4$
$ldepth(9, X) = ldepth(13, X) = 5$
$ldepth(11, X) = 6$

Observamos:

$$med = 11 \; y \; ldepth(11, X) = 6 \approx \frac{n}{2}$$

$$Q_1 = 5 \quad Q_3 = 17 \; y \; ldepth(5, X) = ldepth(17, X) = 3 \approx \frac{n}{4}$$

$$ldepth(1, X) = ldepth(21, X) = 1$$

1 y 21 son los valores mínimo y máximo respectivamente del conjunto.

La propuesta inicial de Tukey dio lugar a un número interesante de posibilidades.
En primer lugar permite definir la mediana en el caso multivariado. Como en el caso unidimensional la mediana es el valor de "máxima profundidad", en más dimensiones sería razonable pensar que el punto $x \in R^d$ con máxima profundidad será la mediana multivariada.
Segundo, el contorno de profundidad $n/4$ es una región convexa cuya forma indica la escala y la correlación de los datos.
En tercer lugar se puede definir una media podada, promediando aquellos puntos que tienen una profundidad (depth) dada, por ejemplo $\text{ldepth} \geq n/10$
La mediana, la media podada y la covarianza podada estimada tienen dos importantes propiedades: son invariantes por transformaciones afines y son robustas en altas dimensiones.

En las secciones siguientes vamos a ver que la noción de profundidad (depth) conduce a estimadores que son **invariantes por transformaciones afines de los datos y tienen alto punto de ruptura.**

2. PROFUNDIDAD (DEPTH)

Lema 2.1: ldepth es invariante por transformaciones afines, es decir:

$$ldepth(Ax + b; \{AX_i + b\}) = ldepth(x, X)$$

para todo $b \in R^d$ y $A \in R^{d \times d}$ no singular.

Demostración:

Si $H_{u,x} = \{y \in R^d : u^t y \geq u^t x\}$ es un semiespacio en R^d, vale

$x_i \in H_{u,x} \Leftrightarrow Ax_i + b \in AH_{u,x} + b \quad \forall b \in R^d; \quad \forall A \in R^{d \times d}$ no singular.

Entonces $\min_{|u|=1}\{i : x_i \in H_{u,x}\} = \min_{|u|=1}\{i : Ax_i + b \in AH_{u,x} + b\}$, luego

$$ldepth(x, X) = \min_{|u|=1} ldepth_1(u^T x; \{u^T X_i\}) = \min_{|u|=1} \#\{i : u^T X_i \geq u^t x\} = \min_{|u|=1} \#\{i : X_i \in H_{u,x}\} =$$
$$= \min_{|u|=1}\{i : AX_i + b \in AH_{u,x} + b\} = \min_{|u|=1} \#\{i : u^t(AX_i + b) \geq u^t(Ax + b)\} =$$
$$= ldepth(Ax + b, \{AX_i + b\})$$

Definición 2.1: se define [3] la región de profundidad k como

$$\boxed{D_k = \{x \in R^d : ldepth(x, X) \geq k\}} \quad (6)$$

o equivalentemente la región de profundidad α como [10]

$$\boxed{D_\alpha = \{x \in R^d : depth(x, X) \geq \alpha\}} \text{ con } \alpha \in [0,1] \quad (7)$$

Lema 2.2: D_k es la intersección de los semiespacios cerrados de dimensión d que contienen al menos $n+1-k$ puntos del conjunto $X^{(n)} = \{X_1, ..., X_n\}$. [12]

Demostración: $\subseteq$) Si x no pertenece a la intersección, existe un semiespacio cerrado que contiene al menos $n+1-k$ puntos de $X^{(n)}$ y que no contiene a x, entonces x pertenece a un semiespacio abierto que contiene a lo sumo $n-(n+1-k) = k-1$ observaciones. Entonces $ldepth(x, X^{(n)}) \leq k-1 \Rightarrow x \notin D_k$ lo que es un absurdo que provino de suponer que x no pertenecía a la intersección.

$\supseteq$) Supongamos que $x \notin D_k$; entonces $ldepth(x, X^{(n)}) < k$, luego x pertenece a un semi- espacio cerrado que contiene menos de k observaciones de $X^{(n)}$. El complemento de este semiespacio

cerrado será un abierto de R^d que contiene al menos $n+1-k$ puntos de $X^{(n)}$, a partir del cual podemos obtener un semiespacio cerrado que no contiene a x y que tiene al menos $n+1-k$ puntos de $X^{(n)}$.

Lema2.3: las regiones de profundidad forman una sucesión decreciente de conjuntos convexos encajados, es decir $D_{k+1} \subset D_k$.[3]

Demostración: D_k es convexo, por ser la intersección de semiespacios cerrados.

Los semiespacios de R^d que contienen por lo menos $n+1-k$ puntos de $X^{(n)}$, en particular contienen $n-k$ puntos de $X^{(n)}$. Luego D_k es la intersección de una subfamilia de la familia que define a D_{k+1}

Como los puntos en D_k satisfacen un subconjunto de las condiciones que los puntos en D_{k+1} deben satisfacer $D_{k+1} \subset D_k$.

Una pregunta que podemos hacernos es: ***¿Cuántas regiones*** D_k ***hay?***

Es decir ***¿cuál es el valor máximo de profundidad (depth) para un conjunto de datos?***

Para $d = 1$ la mediana es el valor de máxima profundidad, tiene *depth* $n/2$.

Vamos a ver que para $d > 1$ la máxima profundidad puede ser menor a $n/2$, ésto dependerá de la forma del conjunto de datos.

Definición 2.2: $k^*(X) = \max_x ldepth(x, X)$ (8)

Es decir es la máxima profundidad alcanzada sobre los puntos de R^d. [3]

Definición 2.3: $k^+(X) = \max_i ldepth(X_i, X)$ (9)

Es decir es la máxima profundidad alcanzada sobre los puntos de X . [3]

Definición 2.4: diremos que un conjunto de datos está en "posición general" si ninguno de sus puntos es combinación lineal afín de los otros (es decir no más de dos puntos pertenecen a una recta, no más de tres puntos a un mismo plano, etc.)

Definición 2.5: Notaremos con $\lceil a \rceil$ al entero más cercano y mayor o igual que a y con $\lfloor a \rfloor$ al entero más cercano y menor o igual que a.

Proposición 2.1: si X es un conjunto de datos en posición general vale

$$\left\lceil \frac{n}{d+1} \right\rceil \leq k^*(X) \leq \left\lceil \frac{n}{2} \right\rceil \quad [3]$$

Observación: si el conjunto de datos es casi simétrico, el máximo valor de profundidad será mucho mayor que $\frac{n}{d+1}$ y estará muy próximo a $\frac{n}{2}$

Más adelante veremos que la cota inferior se alcanza si el conjunto de datos es un conjunto estratégicamente anidado de d simplices.

Acerca de $k^+(X)$ solo podemos decir que en general $1 \leq k^+(X) \leq k^*(X)$

Para poder demostrar la proposición 2.1 necesitamos definiciones y lemas previos.

Vimos que $H_{u,x} = \{y \in R^d : u^t y \geq u^t x\}$ es un semiespacio en R^d, con interior $(H_{u,x})^0 = \{y \in R^d : u^t y > u^t x\}$ y borde $\delta(H_{u,x}) = \{y \in R^d : u^t y = u^t x\}$

Definición 2.6: dado un conjunto de datos $X_i, i = 1, ..., n$ se define la distribución empírica $P_n(S) = n^{-1} \# \{i : X_i \in S\}$ para todo conjunto medible S.

Definición 2.7: se define en $H_{u,x}$ la métrica $\mu_H(P_n, P) = \sup_{u,x} \left| P_n(H_{u,x}) - P(H_{u,x}) \right|$

Observación: μ_H tiene la propiedad de Glivenco-Cantelli: si $\{X_i\}$son independientes e idénticamente distribuidos con distribución de probabilidad P entonces $\mu_H(P_n, P) \xrightarrow[n\to\infty]{} 0$ c.s..[7]

Definición 2.8:[3] se define la **probabilidad proyectada** como:

$$\boxed{\Pi(x) = \inf_u P(H_{u,x})} \quad (10)$$

Es decir $\Pi(x)$ es la mínima probabilidad alcanzada entre los semiespacios que contienen a x.

Definición 2.9:[3] la **probabilidad empírica proyectada** se define como:

$$\boxed{\Pi_n(x) = \inf_u P_n(H_{u,x})} \quad (11)$$

Observaciones:

1) Podemos relacionar el concepto de profundidad y de probabilidad empírica proyectada.
Como $P_n(H_{u,x}) = n^{-1}\#\{i : X_i \in H_{u,x}\}$, tomando ínfimo en ambos miembros nos queda $\inf P_n(H_{u,x}) = n^{-1} ldepth(x, X)$, luego

$$\boxed{n^{-1} ldepth(x, X^{(n)}) = \Pi_n(x)} \quad (12)$$

2) $\sup_x \left| \Pi_n(x) - \Pi(x) \right| \le \mu_H(P_n, P)$

Por Glivenko-Cantelli $\mu_H(P_n, P) \xrightarrow[c.s]{} 0$, luego

$$\boxed{n^{-1} ldepth(x, X^{(n)}) \xrightarrow[c.s.]{} \Pi(x)} \quad (13)$$

Definición 2.10: para toda distribución P sobre R^d y para todo $x \in R^d$, Rousseeaw y Ruts [10] definen con la notación de (4):

$$\boxed{depth(x) = \inf_{|u|=1} P(H_{u,x})} \quad (14)$$

Lema 2.4: Π es una función semicontinua superiormente de x [16]. Si además P es absolutamente continua, Π es continua.[3]

Demostración: para cada semiespacio cerrado H, la funcional lineal $P \longrightarrow P(H)$ es semicontinua superiormente (s.c.s) para la convergencia débil. Además, $P(.-v) \xrightarrow[v \xrightarrow{w} 0]{} P$ [1]

Como $H_{u,x} = H_{u,w} + (x-w)$, si llamamos $f_u(x) = P(H_{u,x})$, f_u es semicontinua superiormente.

Por otro lado tenemos $\Pi(x) = \inf_u P(H_{u,x}) = \inf_u f_u(x)$.

Luego Π es s.c.s., por ser el ínfimo de una colección de funciones s.c.s.

Debemos probar que si P es absolutamente continua Π es continua, pero como Π es s.c.s., bastará probar que es semicontinua inferiormente (s.c.i.)

Sea $(x_n)_{n\geq 1}$ una sucesión tal que $x_n \xrightarrow[n\to\infty]{} x_0$, y u_n una sucesión de direcciones tales que $P(H_{u_n,x}) \leq \Pi(x_n) + 1/n$. Como las u_n están contenidas en la esfera unitaria en R^d, entonces existe un punto de acumulación, es decir se puede extraer una subsucesión convergente tal que $u_n \to u$.

Si hacemos $P(H_{u,x_0}) - P(H_{u_n,x_n}) = \int (I_{H_{u,x_0}} - I_{H_{u_n,x_n}})\, dP$

Como la diferencia en valor absoluto de los indicadores está mayorada por 1, podemos aplicar el teorema de convergencia mayorada.[16]

Si $u_n \xrightarrow[n\to\infty]{} u$ y $x_n \xrightarrow[n\to\infty]{} x_0 \Rightarrow P(H_{u,x_0}) - P(H_{u_n,x_n}) \longrightarrow 0$

Luego $\lim_{n\longrightarrow\infty} \inf \Pi(x_n) = \lim_{n\longrightarrow\infty} \inf P(H_{u_n,x_n}) = P(H_{u,x_0}) \geq \Pi(x_0)$, y por lo tanto es s.c.i.

Definición 2.11: una distribución de probabilidades es centrosimétrica respecto de x_0 si: $P(x_0 + S) = P(x_0 - S)$ para todo conjunto medible S.

Lema 2.5: Si P es centrosimétrica respecto de x_0, entonces $\Pi(x_0) \geq \frac{1}{2}$. Si además P es absolutamente continua, $\Pi(x_0) = \frac{1}{2}$. [3]

Demostración: Como $H_{u,x_0} = x_0 + H_{u,0}$ y P es centrosimétrica con respecto a x_0, tenemos $P(H_{u,x_0}) = P(x_0 + H_{u,0}) = P(x_0 - H_{u,0}) = P(H_{-u,x_0})$

Como $H_{u,x_0} \cup H_{-u,x_0} = R^d$

$2P(H_{u,x_0}) = P(H_{u,x_0}) + P(H_{-u,x_0}) \geq 1 \quad \Rightarrow \quad P(H_{u,x_0}) \geq \frac{1}{2}$

Si P es absolutamente continua $P(\delta H_{u,x_0}) = 0$ para todo semiespacio.

Luego $P(H_{u,x_0}) + P(H_{-u,x_0}) = 1 \quad \Rightarrow \quad P(H_{u,x_0}) = \frac{1}{2}$

Lema 2.6: $\max_x \Pi(x) \geq 1/(d+1)$

Antes de demostrarlo vamos a demostrar el siguiente lema.

Lema2.7: para toda probabilidad P y $\alpha > 0$

$D_\alpha = \cap\{H : H \text{ es un semiespacio cerrado con} P(H^c) < \alpha\}$ [10]

Demostración: $\subset$) si $x \in D_\alpha \quad \Leftrightarrow \quad \inf_{|u|=1} P(H_{u,x}) \geq \alpha \quad \Leftrightarrow \quad \forall\, u / |u| = 1\, P(H_{u,x}) \geq \alpha$

Supongamos que existe un semiespacio cerrado H con $P(H^c) < \alpha$ tal que $x \notin H$. Entonces existe un vector unitario u (ortogonal al borde de H) tal que $H_{u,x} \subset H^c$, luego $P(H_{u,x}) \leq P(H^c) < \alpha$,que es una contradicción .

$\supset$) Supongamos que $x \in \{H : H \text{ es un semiespacio cerrado con} P(H^c) < \alpha\}$ y $x \notin D_\alpha$; entonces existe un vector unitario $u / P(H_{u,x}) = \alpha_0 < \alpha$.

A $H_{u,x}$ podemos escribirlo como la intersección de una sucesión decreciente de semiespacios abiertos

$H^n_{u,x} = \{y \in R^d : u^t y > u^t x - 1/n\}$

Como $H_{u,x} = \cap \downarrow H^n_{u,x} \quad \Rightarrow \quad \alpha_0 = P(H_{u,x}) = lim \downarrow P(H^n_{u,x})$ porque P es una medida finita. Luego existe un $m / P(H^m_{u,x}) < \alpha$. Si llamamos $H = (H^m_{u,x})^c \quad \Rightarrow \quad P(H^c) < \alpha$ absurdo, pues habíamos supuesto $x \notin H$.

Demostración del lema 2.6: [10] sea $\alpha^* = \max_x \Pi(x) \leq 1$ y sea $\varepsilon > 0$ si consideramos las bolas cerradas $B_n = \{x \in R^d : |x| \leq n\}$; como $\cup B_n \uparrow R^d \Rightarrow P(B_n) \uparrow 1$, entonces existe un m tal que $P(B_m) > 1 - \varepsilon$, para simplificar la notación vamos a llamar $B = B_m$. Por otro lado tenemos que $D_{\alpha^* + \varepsilon} = \cap\{H : H \text{ es un semiespacio cerrado con} P(H^c) < \alpha^* + \varepsilon\} = \phi$

Pero todos los semiespacios H con $P(H^c) < \alpha$ pertenecen al conjunto $\{H : H \text{ es un semiespacio cerrado con} P(H^c) < \alpha^* + \varepsilon\}$ que entonces no son vacíos.

Si consideramos $C = \{B \cap H : H \text{ es un semiespacio cerrado con} P(H^c) < \alpha^* + \varepsilon\}$, C es una familia infinita de conjuntos compactos y convexos en R^d con $\cap C = \phi$. Por el teorema de selección de Helly existen conjuntos $C_1, ...C_{d+1}$ en C / $\cap_{j=1}^{d+1} C_j = \phi, \quad \cup_{j=1}^{d+1} C_j = R^d$ [15]

Para cada $j \quad P(C_j^c) = P((B \cap H)^c) = P(B^c \cup H^c) < \alpha^* + 2\varepsilon$

Luego $1 = P(R^d) = \sum_{j=1}^{d+1} P(C_j^c) < (d+1) \cdot (\alpha^* + 2\varepsilon) \; \forall \varepsilon > 0 \quad \Rightarrow \quad \alpha^* \geq 1/(d+1)$

Estos tres lemas nos dicen que bajo simetría el valor de máxima profundidad es $\cong \frac{n}{2}$ **y es siempre** $\geq \frac{n}{d+1}$

Demostración de la proposición 2.1: si X está en posición general existe una dirección de proyección de forma tal que no hay empates en el conjunto de datos proyectados $\{u^t X_i\}$. En esta proyección la máxima profundidad es $\left\lceil \frac{n}{2} \right\rceil$.

Como $ldepth_d(x, X) = \min_{|u|=1} ldepth_1(u^t x, \{u^t X_i\}) \leq ldepth_1(v^t x, \{v^t X_i\}) \leq \left\lceil \frac{n}{2} \right\rceil$

Luego $k^*(X) \le \left\lceil \frac{n}{2} \right\rceil$.

Para la otra desigualdad usamos que $\Pi_n(x) = \frac{1}{n} \cdot ldepth(x, X)$. Si

$$\max_x \Pi(x) \ge \frac{1}{d+1} \quad \Rightarrow \quad ldepth(x, X) \ge \frac{n}{d+1} \quad \forall x \quad \Rightarrow \quad \max_x ldepth(x, X) \ge \frac{n}{d+1} \ge \left\lceil \frac{n}{d+1} \right\rceil$$

Luego $k^*(X) \ge \left\lceil \frac{n}{d+1} \right\rceil$

Proposición 2.2: sea $X^{(n)} = \{X_1, ..., X_n\}$ una muestra proveniente de una distribución de probabilidades absolutamente continua y centrosimétrica. Entonces $n^{-1}k^*(X^{(n)})$ converge en probabilidad y casi seguramente a $1/2$ cuando n tiende a infinito.
Si además P tiene densidad positiva en x_0, entonces $n^{-1}k^+(X^{(n)})$ converge en probabilidad y casi seguramente a 1/2.

Demostración: como P es centrosimétrica y absolutamente continua , por el lema 2.5, $\Pi(x_0) = 1/2$

Por otro lado $n^{-1} ldepth(x_0, X^n) = \Pi_n(x_0) \xrightarrow[c.s.]{} \Pi(x_0) = \frac{1}{2}$

Si P es absolutamente continua con probabilidad 1 y $X^{(n)}$ está en posición general, entonces por la proposición 2.1 $\left\lceil \frac{n}{d+1} \right\rceil \le k^*(X) \le \left\lceil \frac{n}{2} \right\rceil$.

Por definición de $k^*(X)$, $k^*(X) \ge ldepth(x_0, X^{(n)})$.

Luego vale $n^{-1} ldepth(x_o, X) \le n^{-1}k^*(X) \le n^{-1}\left\lceil \frac{n}{2} \right\rceil$, entonces $n^{-1}k^*(X) \xrightarrow[c.s]{} \frac{1}{2}$

Sea X_{i_n} el punto de $X_1, ..., X_n$ más cercano a x_0; es decir $X_{i_n} = \arg\min |X_i - x_0|$. Vamos a probar que $\{X_{i_n}\}_{n=1}^{\infty}$ converge casi seguramente a x_0 o equivalentemente que para todo $\varepsilon > 0$ $P(A_n) = P(\{w / |X_{i_n}(w) - x_0| > \varepsilon\}) \xrightarrow[n\to\infty]{} 0$.[5]

Sea $Y_i = |X_i - x_0|$ y $Z_n = \min\{Y_1, ..., Y_n\}$, entonces $F_{Z_n}(Z) = 1 - (1 - F_Y(Z))^n$. Ahora

$$P(A_n) = P(\{\left|X_{i_n} - x_0\right| > \varepsilon\}) = 1 - P(\left|X_{i_n} - x_0\right| \le \varepsilon) = 1 - P(Z_n \le \varepsilon) =$$
$$= 1 - F_{Z_n}(\varepsilon) = (1 - F_Y(\varepsilon))^n = (a(\varepsilon))^n$$

Por la densidad positiva de P en x_0, $F_Y(\varepsilon) > 0$, entonces $a(\varepsilon) < 1$. Luego

$\sum_{n=1}^{\infty} (a(\varepsilon))^n < \infty$ y por el lema de Borel-Cantelli $P(A^\infty) = 0$ [5].

Ahora bien $\{A_n\}_{n=1}^{\infty}$ es una sucesión decreciente, entonces $A^\infty = A_\infty = \bigcap_{n=1}^{\infty} A_n$ y por lo tanto $P(\bigcap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n) = 0$ como queríamos demostrar.

Por la definición de $k^+(X)$ y como $\left|\Pi_n(X_{i_n}) - \Pi(X_{i_n})\right| \le \mu_H(P_n, P)$ vale

$$n^{-1}k^+(X^{(n)}) \ge n^{-1}ldepth(x_{i_n}, X^{(n)}) = \Pi_n(X_{i_n}) \ge \Pi(X_{i_n}) - \mu_H(P_n, P)$$

Como P es absolutamente continua, por el lema 2.5 $\Pi(X_{i_n}) \xrightarrow[c.s.]{} \Pi(x_0)$. Por Glivenko-Cantelli y como $k^+ \le k^*$ se tiene $n^{-1}k^+(X^{(n)}) \xrightarrow[cs]{} \frac{1}{2}$.

Resumiendo, si X es un conjunto casi simétrico entonces la profundidad máxima es cercana a 1/2 independientemente de la distribución.

Entonces tendremos a lo sumo $n/2$ regiones de profundidad si el conjunto de datos es casi simétrico, pero muchas menos para conjuntos asimétricos.

Otra pregunta que podemos hacernos es: ***¿Qué forma tienen las regiones de profundidad?***

Esto dependerá del conjunto de datos. Si el conjunto de datos proviene de una muestra aleatoria de una distribución elíptica los contornos (borde de las regiones) serán casi elipses.

Observación: La convergencia de las regiones es en el sentido de la convergencia de conjuntos en la medida de Hausdorff $h(A, B) = \sup(\rho(A, B), \rho(B, A))$ donde A y B son conjuntos cerrados no vacíos y $\rho(A, B) = \sup_{x\in A} d(x, B)$. [2]

Lema 2.8: si $X^{(n)} = \{X_1, ..., X_n\}$ es una muestra aleatoria de una distribución elíptica simétrica, las $\lfloor n\alpha \rfloor$ regiones de profundidad de $X^{(n)}$ convergen en probabilidad y casi seguramente a una elipsoide de la misma forma que la de la distribución patrón y en una escala que depende de α

Por ejemplo si la muestra proviene de una distribución normal estándar Φ_d en R^d, la forma límite de los $\lfloor n\alpha \rfloor$ contornos será una esfera de radio $R_\alpha = \Phi^{-1}(1-\alpha)$ donde Φ^{-1} es la inversa de la distribución normal estándar univariada.

Demostración del lema 2.8: vamos a hacerlo en el caso en que la distribución elíptica es una normal estandar bivariada. [10]

En este caso la función de densidad $f(x,y) = (2\pi)^{-1} \exp(-(x^2 + y^2)/2)$ es estrictamente positiva en R^2.

Por simetría la probabilidad proyectada de un punto arbitrario $\Pi(x,y)$ es igual a la masa del semiespacio H a la derecha de la línea vertical que pasa por $(r,0) = (\sqrt{x^2 + y^2}, 0)$.

Luego $\Pi(x,y) = 1 - \Phi(\sqrt{x^2 + y^2}) = \Phi(-\|x\|)$ como $\Pi_n(x) \longrightarrow \Pi(x)$ uniformemente $D_\alpha = \{(x,y) \in R^2 : 1 - \Phi(\sqrt{x^2 + y^2}) \geq \alpha\}$, haciendo las cuentas nos queda $D_\alpha = \left\{(x,y) : x^2 + y^2 \leq (\Phi^{-1}(1-\alpha))^2\right\}$ que es una circunferencia de centro (0,0) y radio $\Phi^{-1}(1-\alpha)$.

En resumen, las regiones de profundidad son convexos anidados independientes del sistema de coordenadas.

3. ESTIMADORES BASADOS EN LA PROFUNDIDAD -PROPIEDADES

Definición 3.1: [3] en el caso multivariado se define la **mediana de Tukey** $T_*(X)$ para $X = \{X_1, ..., X_n\}$ en R^d como el punto de mayor profundidad, es decir

$$T_*(X) = \arg\max_x ldepth(x, X)$$

Cuando la profundidad no alcanza un único máximo, se define como

$$T_*(X) = Ave\{x : ldepth(x, X) = \max_x ldepth(x, X)\}$$

Definición 3.2: un estimador análogo a la media podada con poda α es la **media profunda podada con poda** α (α-depth-trimmed mean)

$$T_\alpha(X) = Ave\{X_i \in X : ldepth(X_i, X) \geq n\alpha\}$$

Por las propiedades de la profundidad vistas en la sección 2, estos dos estimadores son estimadores consistentes del centro de simetría para cualquier distribución centrosimétrica y tienen velocidad de convergencia $n^{-1/2}$ para sus valores límites bajo condiciones débiles de regularidad. Son invariantes por transformaciones afines, es decir seleccionan el mismo punto del espacio independientemente del sistema de coordenadas.

Vamos a demostrar que además tienen buenas propiedades de robustez.

Recordemos que intuitivamente el punto de ruptura es la menor cantidad de contaminación necesaria para perturbar enteramente al estimador.

Una definición formal del punto de ruptura para una muestra finita dada por Donoho [3] es la siguiente:

Definición 3.3: sea $X^{(n)}$ un conjunto de datos de tamaño n y T un estimador. A $X^{(n)}$ le adjuntamos otro conjunto de datos $Y^{(m)}$ de tamaño m. Si para una elección estratégica de $Y^{(m)}$, $T(X^{(n)} \cup Y^{(m)}) - T(X^{(n)})$ es arbitrariamente grande, diremos que el estimador T se quiebra bajo una fracción de contaminación $m/(n+m)$. Formalmente el punto de ruptura es

$$\varepsilon^*(T, X) = min\left\{\frac{m}{n+m} : \sup_{Y^{(m)}} \left| T(X \cup Y^{(m)} - T(X) \right| = \infty\right\}$$

Por ejemplo el punto de ruptura de la media es $1/(n+1)$ es decir que una sola observación contaminada es necesaria para romper al promedio. Para la mediana es 1/2 es decir que se necesita que la mitad de los datos sean malos para romper a la mediana, y esto es lo mejor que se puede esperar para un estimador de posición.

Lema 3.1: si $k^+(X) \geq k$ entonces $T_k(X) = Ave\{X_i \in X : ldepth(X_i, X) \geq k\}$ está bien definido, su punto de ruptura está bien definido y $\varepsilon^*(T_k, X) = \frac{k}{n+k}$. [3]

Demostración: si $k^+(x) = \max_i ldepth(x_i, X) \geq k$ entonces X contiene puntos con profundidad k entonces $T_k(X)$ está bien definido.

El punto de ruptura de $T_k(X)$ estará bien definido si $T_k(X \cup Y)$ está bien definido para todo Y, es decir si en $X \cup Y$ hay puntos de profundidad k para cualquier elección de Y, pero $k^+(X \cup Y) = \max_{x_i \in X \cup Y} ldepth(X_i, X \cup Y) \geq \max_{x_i \in X \cup Y} ldepth(X_i, X) \geq \max_{x_i \in X} ldepth(X_i, X) = k^+(X)$.

Probaremos primero que $\varepsilon^*(T_k, X) \geq \frac{k}{n+k}$.

Para que T_k se rompa en X, debe ser contaminado por un conjunto $Y = \{Y_i\}$ de manera que $T_k(X \cup Y)$ esté fuera de cualquier conjunto acotado fijo, por ejemplo fuera de la cápsula convexa de X. Para esto vamos a elegir un punto contaminante Y_1 fuera de la cápsula convexa de X con $ldepth(Y_1, X \cup Y) \geq k$. Por el teorema del hiperplano que separa existe una dirección u que separa a todos los $X_i \in X \cup Y$ de $Y_1 : \max_i u^t X_i < u^t Y_1$. Pero Y_1 tiene profundidad k en $X \cup Y$ entonces hay al menos k elementos en $X \cup Y$ que están a la derecha de $u^t Y_1$. Como ninguno de estos puntos puede estar en X, entonces deben estar en Y.Luego $\#Y \geq k$ y la fracción contaminante es al menos $k/(n+k)$.

Probaremos ahora la otra desigualdad, es decir que k es una cantidad suficiente para romper a T_k en X. Elegimos como antes $Y_1, \ldots, Y_k$ en el mismo sitio fuera de la cápsula convexa de X. Para todo $u, u^t Y_1 = u^t Y_2 = \ldots = u^t Y_k$;de esta forma $ldepth(Y_i, X \cup Y) \geq k, \quad i = 1, \ldots, k$ .Luego $T_k(X \cup Y)$ es un

promedio sobre un conjunto que contiene a todos los Y_i . Como podemos elegir a Y con norma arbitrariamente grande, $T_k(X \cup Y)$ se puede hacer arbitrariamente grande.

Proposición 3.1: Sea $X^{(n)} = \{X_1, ..., X_n\}$ una muestra aleatoria de tamaño n de una distribución absolutamente continua y centrosimétrica en R^d , $d > 2$ y sea $\alpha < 1/3$. Con probabilidad 1, para todo n suficientemente grande, T_α está bien definido , el punto de ruptura de $T_\alpha(X^{(n)})$ está bien definido y

$$\varepsilon^*(T_\alpha, X) \xrightarrow[c.s.]{} \alpha \quad [3]$$

Observación: La limitación $\alpha < 1/3$ es real. De hecho , aun podando lo máximo posible T_* no puede dar un punto de ruptura mayor que $1/3$.

Demostración : para $\alpha < 1/3$ tomamos $\beta \in ((3/2)\alpha, 1/2)$, con esta elección de β cuanto más cercano está α a 1/3, más próximo está β a 1/2. Por la proposición 2.2 ; ésto implica que existe una variable aleatoria positiva $n_0(\beta)$ que es casi seguramente finita y $k^+(X^{(n)})/n > \beta \ \forall n > n_0(\beta)$

Sea Y el conjunto que contiene m puntos de contaminación $m \le n/2$

$\forall n > n_0(\beta)$, $k^+(X \cup Y) > k^+(X) > \beta n$. Luego $T_\alpha(X \cup Y)$ estará bien definido si y solo si $k^+(X \cup Y) \ge \lfloor \alpha(n+m) \rfloor$; pero $\beta n > \lfloor \alpha(n+m) \rfloor \ \forall m \le n/2 \Rightarrow \forall n > n_0(\beta)$ $T_\alpha(X \cup Y)$ está bien definida.

Para m y n fijos, $T_\alpha(X \cup Y) = T_k(X \cup Y)$ donde $k = \lfloor \alpha(n+m) \rfloor$. Por la proposición 3.1 Y puede ser elegido de manera que T_k se estropee si y solo si se contamina con una cantidad $m \ge k$, en este caso es $m \ge \lfloor \alpha(n+m) \rfloor$

$\alpha < 1/3, m = n/2$ es solución de la inecuación, además, luego el menor valor de m que es solución de la inecuación es $m = \lfloor (\alpha/(1-\alpha))n \rfloor$ o $m = \lceil (\alpha/(1-\alpha))n \rceil$

Luego $\forall n > n_0(\beta)$ el punto de ruptura está bien definido y $\varepsilon^*(T_\alpha, X) = \dfrac{m}{n+m} = \alpha + O\left(\dfrac{1}{n}\right)$ c.s.

Proposición 3.2: Sea $X^{(n)} = \{X_1, ..., X_n\}$ una muestra aleatoria de tamaño n con distribución absolutamente continua y centrosimétrica en R^d, $d > 2$ el punto de ruptura de $T_*(X^{(n)}) \xrightarrow[c.s]{} 1/3$ cuando n tiende a infinito. [3]

Demostración: primero vamos a demostrar que el valor límite del punto de ruptura es al menos 1/3.

m puntos contaminantes son suficientes para romper al estadístico, si son tales que $T_*(X \cup Y)$ está fuera de la cápsula convexa de X. Pero por el teorema de separación de un hiperplano y usando el mismo argumento de la demostración del lema 3.1, para que $T_*(X \cup Y)$ esté afuera de la cápsula convexa de X, el número de puntos contaminantes debe ser al menos la profundidad de $T_*(X \cup Y)$. Luego para que m puntos contaminantes causen ruptura se tiene que verificar que $m \geq k^*(X)$

Como $ldepth(T_*(X \cup Y), X \cup Y) = k^*(X \cup Y) \geq k^*(X) \Rightarrow m \geq k^*(X)$

Entonces $\varepsilon^*(T_*, X) = \dfrac{m}{n+m} \geq k^*(X)/(n + k^*(X))$

Como $k^*(X) = n/2(1 + o_p(1)) \Rightarrow k^*(X)/(1 + k^*(X)) \xrightarrow[c.s]{} 1/3$

Entonces $lim\inf_n \varepsilon^* \geq 1/3$ c.s.

Ahora probaremos que el valor límite del punto de ruptura es a lo sumo 1/3.

Sea N una medida tal que $N(S) = \#\{i : X_i \in S\}$. Sea x_0 el punto de centrosimetría de P y $k^0 = \max_u N(H_{u,x})$. Observemos que $k^0 = n/2(1 + o_p(1))$ pues $N(H_{u,x_0}) = nP_n(H_{u,x_0})$ y por absoluta continuidad y centrosimetría de P, $P(H_{u,x_0}) = 1/2$ para todo u.

Entonces $\left| k^0/n - 1/2 \right| \leq \sup_u \left| P_n(H_{u,x_0}) - P(H_{u,x_0}) \right| \leq \mu_H(P_n, P) \xrightarrow[p]{c.s} 0$ por Glivenko-Cantelli.

Sea $m = k^0 + 2d + 1$ vamos a probar que $\varepsilon^*(T_*, X) \leq \dfrac{m}{n+m}$

Para probarlo necesitamos dos lemas previos.

Lema 3.2: Sea $x \in R^d$ un punto arbitrario y sea x_0 un punto tal que $N(H_{u,x_0}) \leq k^0 \ \forall u$. Entonces existe una dirección u / $N((H_{u,x_0})^0) \leq k^0$ y $N((H_{u,x_0})^0)$ no contiene a y.

Lema 3.3: Sea X en posición general $N((H_{u,x_0})^0) \le k^0$, $y \notin (H_{u,x_0})^0$, existe $w / N(H_{w,x}) \le k^0 + 2d, \quad y \notin H_{w,x}$.

Sea y un punto arbitrario en R^d y sea $Y^{(m)}$ un conjunto de datos con m repeticiones de y, entonces $ldepth(x, X \cup Y) \ge m$. Vamos a ver que y es el punto de mayor profundidad de $X \cup Y$, es decir $depth(x, X \cup Y) < m \quad \forall x \ne y$ (*)

Como y es arbitrario, probaremos que $\forall\, y \in R^d \; T_*(X \cup Y) = y$ tiene solución, luego T_* se romperá con una contaminación de tamaño m. Sea

$$M(S) = \begin{cases} m & si\ y \in S \\ 0 & si\ y \notin S \end{cases}$$

Entonces $ldepth(x, X \cup Y) = \underset{u}{inf}\,(N(H_{u,x}) + M(H_{u,x})) \le \underset{u}{inf}\,\{N(H_{u,x}) : M(H_{u,x}) = 0\}$

Como $N(H_{u,x_0}) \le k^0$ para todo u por definición de k^0. Por el lema 5 existe una solución particular u con $N(\text{int}\, H_{u,x}) \le k^0, y \notin \text{int}\, H_{u,x}$.

Luego por el lema 3.3 existe $w / N(H_{w,x}) \le k^0 + 2d, \quad y \notin H_{w,x}$.

Como $y \notin H_{w,x}, M(H_{w,x}) = 0$ y $\inf\{N(H_{u,x}) : M(H_{u,x}) = 0\} \le N(H_{w,x}) \le k^0 + 2d$

Además teníamos $m > k^0 + 2d$. Estas últimas 3 conclusiones prueban (*).

Como $m = n/2(1 + o_{c.s.}(1)),\ \varepsilon^*(T_*, X) \le \dfrac{m}{n+m}$ implica $lim\sup_{n\to\infty} \varepsilon^*(T_*, X^{(n)}) \le 1/3\, c.s.$

¿Qué pasa si P no tiene distribución centrosimétrica ?

Supongamos que $k^* / n \to \beta \le 1/2\, c.s.$. El argumento de la proposición 3.2 nos dice que para cada $\alpha < \beta/(1+\beta), T_\alpha$ está bien definida y su punto de ruptura está bien definido para n suficientemente grande y converge casi seguramente a α.

Proposición 3.4: Sea X en posición general, entonces $\varepsilon^*(T_*, X) \ge \dfrac{1}{d+1}$. [3]

Demostración: como en la proposición anterior, si m puntos son necesarios para romper $T_* \Rightarrow\ m \geq k^*(X \cup Y)$.

Por el lema 2.7. $k^*(X \cup Y) \geq (n+m)/(d+1) \quad \Rightarrow \quad m/(n+m) \geq 1/(d+1)$

Luego por definición de $\varepsilon^*(T_*, X)$, $\varepsilon^*(T_*, X) \geq 1/(d+1)$

4- BAGPLOT, UN GRAFICO BIVARIADO BASADO EN LA PROFUNDIDAD

4.1 Descripción

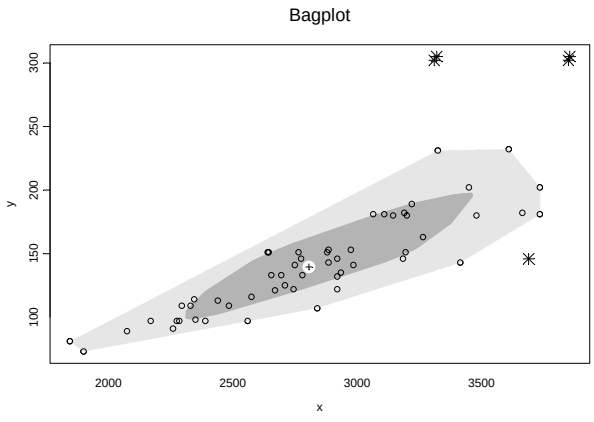
El **bagplot** es una versión bivariada del box-plot, propuesta por Rousseeaw y Ruts. [8,11] El box-plot está basado en rangos, pues la caja va desde la observación con rango $\lfloor n/4 \rfloor$ hasta la observación con rango $\lceil 3n/4 \rceil$, y la línea en el interior de la caja es la mediana, por lo tanto es necesario generalizar el concepto de rango al caso multivariado.

Como vimos antes una natural generalización del rango al caso multivariado es la noción de profundidad introducida por Tukey en 1974.

Luego el **bagplot** se basa en las regiones de profundidad y en la mediana de Tukey definidas en las Secciones 2 y 3.

En las figuras 4.1 y 4.2 se presentan dos versiones del bagplot para un mismo conjunto de datos. El bagplot está formado por una "bolsa" que contiene al 50% central del conjunto de los datos (representada por el polígono oscuro), una zona alrededor de la bolsa que denominaremos vecindad (representada por el área más clara) que contiene los inliers más extremos y una valla (cuyo gráfico es opcional y se representa con línea punteada en la figura 4.2) que separa los inliers de los outliers.

La mediana de Tukey es indicada por una cruz y los outliers están indicados por asteriscos.



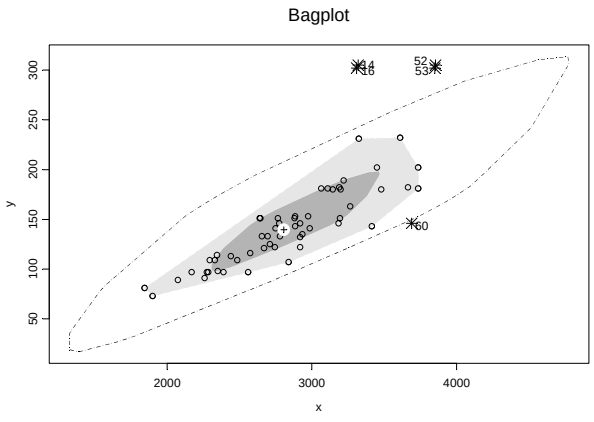


Figura 4.1

Figura 4.2

En las dos figuras fueron incluidos todos los puntos muestrales, y en la figura 4.2 se identificaron los outliers.

Estos gráficos fueron obtenidos utilizando la versión de diciembre de 1999 del algoritmo BAGPLOT para S-PLUS-2000 propuesto por Rousseeuw y Ruts (ver Apéndice I).

La figura 4.1 es la salida que se obtiene por defecto para una muestra con 15 o más datos; pero el programa permite diferentes opciones, como por ejemplo la inclusión de todos los puntos de la muestra y su identificación (clickeando sobre ellos).

El programa también calcula las coordenadas de la mediana de Tukey, que para nuestro ejemplo son (2806.63, 139.513).

El conjunto de datos usado para el ejemplo pertenece a la librería del S-PLUS. La variable x es el peso y la variable y el desplazamiento del motor de 60 autos chicos, medianos y grandes.

Como el boxplot en el caso univariado, el bagplot nos permite visualizar diferentes características del conjunto de datos: su **posición** (la mediana de Tukey), la **dispersión** (el tamaño de la bolsa), la **correlación** (la orientación de la bolsa), la **simetría** (la forma de la bolsa y la vecindad) y las **colas** (la amplitud de la vecindad y los outliers).

4.2 Visualización de estructuras de correlación: ejemplos

Para observar cómo se reflejan en el bagplot diferentes estructuras de correlación, se generaron muestras aleatorias, cada una con 50 datos, de normales bivariadas con distintos coeficientes de correlación ρ programando una función en S-PLUS (ver apéndice III)

Los bagplots obtenidos se muestran en las figuras 4.3 y 4.4.

Podemos observar que cuando $\rho = 0$ la bolsa es casi un círculo que se va deformando con forma casi elíptica hasta llegar a un boxplot univariado cuando $\rho = 1$ o $\rho = -1$.

La orientación de izquierda a derecha es ascendente cuando $\rho > 0$ y descendente cuando $\rho < 0$.

Podemos observar también en todos los gráficos la simetría de los datos por la forma de la vecindad.

En estos ejemplos no hay outliers.

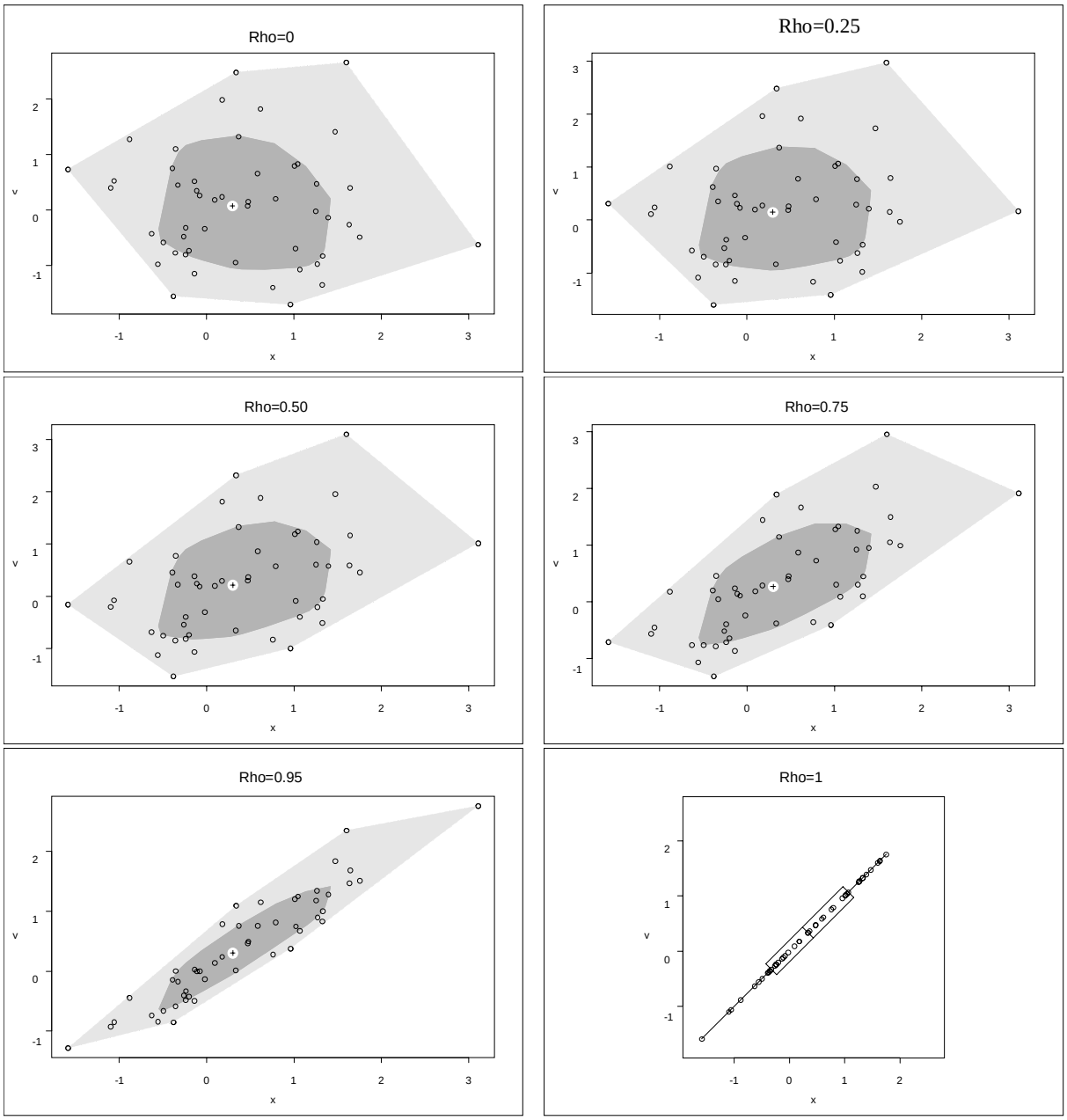


Figura 4.3

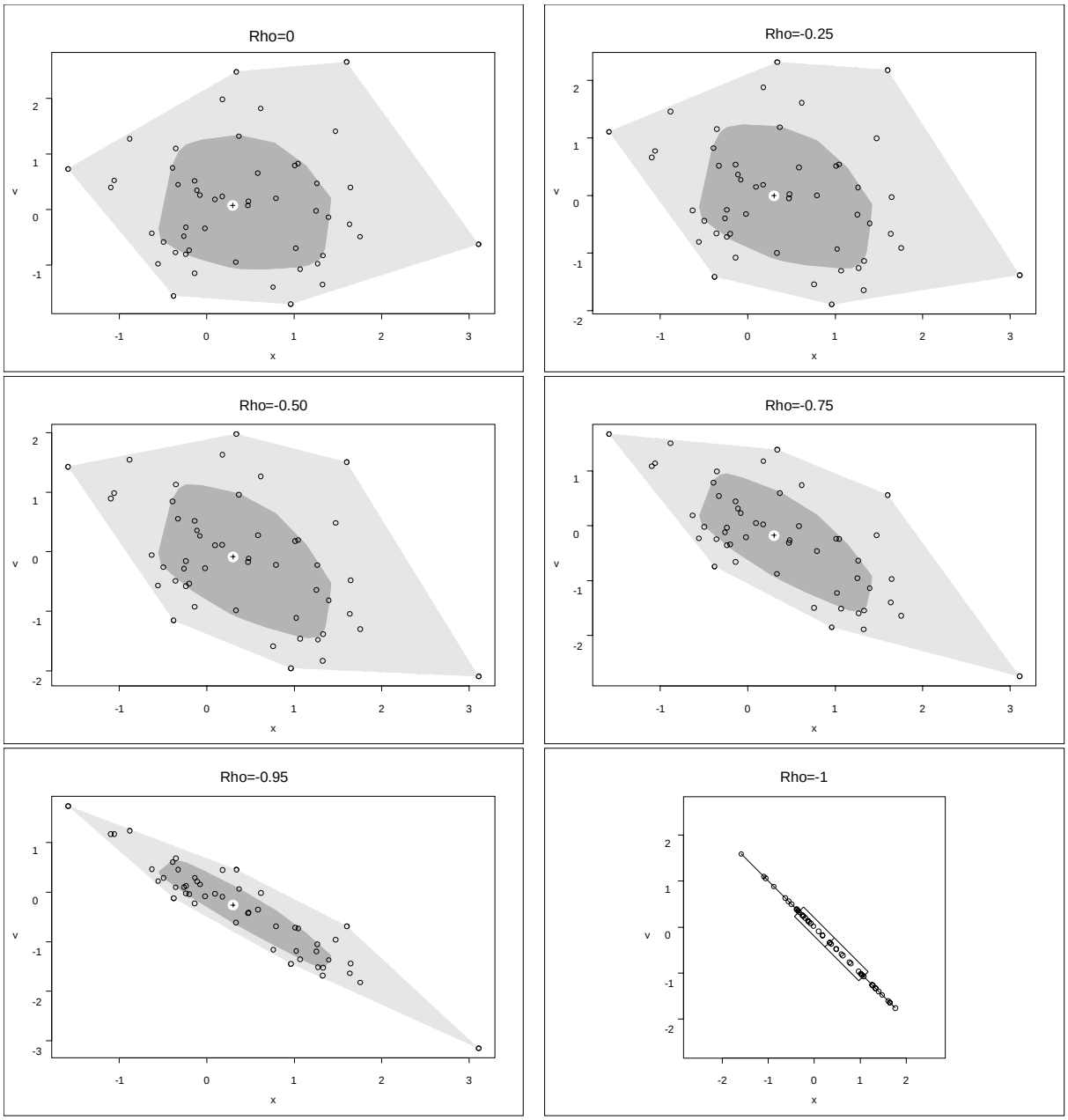


Figura 4.4

4.3 Construcción del bagplot

Para construir la **bolsa** B**,** primero se determina el valor de k tal que $\#D_k \le \lfloor n/2 \rfloor < \#D_{k-1}$ y se interpola linealmente en forma relativa a la mediana de Tukey T_*. La bolsa es un polígono convexo.

La **valla** es obtenida expandiendo la bolsa por un factor 3 relativo a T_*. Este valor 3 fue obtenido en base a simulaciones. [8]

La **vecindad** contiene todos los datos entre la bolsa y la valla, es decir rodea a la bolsa y contiene a los puntos que no son outliers.

Supongamos que $X^{(n)} = \{X_1,..., X_n\}$ está en posición general, si no, reemplazamos cada X_i por $X_i + \varepsilon_i$ donde $\varepsilon_i \approx \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\gamma & 0\\0 & \gamma\end{pmatrix}\right)$ $\gamma > 0$ pequeño.

Si el conjunto de datos no está todavía en posición general se aumenta el valor de γ .

Luego se calculan las profundidades de todas las observaciones (con la rutina LDEPTH).

Para cada valor de profundidad (ldepth) encontrado se cuenta el número de datos con esa profundidad y llamamos k^+ a la mayor profundidad encontrada. Luego se determina $k / \# D_k \le \lfloor n/2 \rfloor \le \# D_{k-1}$. Por otro lado se calcula la mediana de Tukey $T_*(X)$ (con la rutina HALFMED), usando el hecho que la mayor profundidad k^+ verifica $k^+ < \lfloor n/2 \rfloor$.

Luego se hace la transformación de los datos $X_i - T_*$.

Para calcular los vértices de D_k , $p_1^k,..., p_m^k$ y los de D_{k-1} , $p_1^{k-1},..., p_{m'}^{k-1}$ se usa la rutina ISODEPTH, donde m y m' son respectivamente el cardinal de los puntos más extremos de D_k y D_{k-1} . Luego se calculan los ángulos $\alpha_1,..., \alpha_m$ entre la horizontal (poniendo el origen en $T_*(X)$) y los vértices de D_k y los ángulos $\beta_1,..., \beta_{m'}$ con los vértices de D_{k-1}. Se juntan los $m + m'$ ángulos y se los ordena en forma creciente , obteniendo $\gamma_1,..., \gamma_{m+m'}$. Para cada ángulo γ_j se interpola entre D_k y D_{k-1} de la siguiente manera. Se calculan los puntos de intersección entre la recta que une $T_* = 0$ con un vértice de D_k y el lado correspondiente de D_{k-1}, sean estos puntos $q_1^k,..., q_m^k$. De la misma manera se calculan los puntos de intersección entre la recta que une $T_* = 0$ con un vértice de D_{k-1} y el lado correspondiente de D_k, sean estos puntos $q_1^{k-1},..., q_{m'}^{k-1}$. (Figura 4.5 en la última página)

Luego los vértices de la bolsa B son computados como:

$$w_j = \begin{cases} \lambda \cdot p_j^k + (1-\lambda) \cdot q_j^k & si\ \gamma_j\ es\ un\ \alpha \\ \lambda \cdot q_j^{k-1} + (1-\lambda) \cdot p_j^{k-1} & si\ \gamma_j\ es\ un\ \beta \end{cases}$$

El valor de λ se obtiene haciendo lo siguiente. Sea $n_k = \#D_k$, $n_{k-1} = \#D_{k-1}$ y $n_{\text{int}} = n_{k-1} - n_k$, luego

$$\lambda = \frac{\lfloor n/2 \rfloor - n_{\text{int}}}{n_{k-1} - n_{\text{int}}}$$

En segundo lugar se construyen los "bigotes" y las vallas.

La valla es obtenida multiplicando la bolsa por un factor 3 respecto de T_*, este valor está basado en simulaciones.

Un "bigote" es un segmento que une una observación que se encuentra entre la bolsa y la valla desde el borde de la bolsa en la dirección de T_* (esto generaliza los "bigotes" del caso univariado).

Para obtener los "bigotes" y poder detectar los outliers es necesario computar las distancias d_i desde $T_* = 0$ a un punto X_i relativo a la bolsa B. Es decir para cada observación X_i buscamos la intersección u_i de la recta L_i con el correspondiente lado de B.

Cada punto u_i es obtenido como $u_i = \lambda \cdot v_i^k + (1-\lambda) \cdot v_i^{k-1}$ y la distancia relativa $d_i = \| X_i \| / \| u_i \|$.

Para cada observación X_i se determina su tipo t_i dado por

$$t_i = \begin{cases} 1 & 0 \le d_i \le 1 \\ 2 & 1 < d_i \le 3 \\ 3 & d_i > 3 \end{cases}$$

Es decir una observación será de tipo 1 si está dentro de la bolsa, de tipo 2 si está entre la bolsa y la valla y de tipo 3 si está más allá de la valla. Éstas últimas son consideradas outliers y se las indica con un asterisco.

Para dibujar los "bigotes" se unen las observaciones de tipo 2 con su respectivo u_i.

Cuando el tamaño de la muestra es grande sólo se dibuja para cada lado de la bolsa el bigote correspondiente a la mayor distancia d_i.

En la versión utilizada para graficar el bagplot, los bigotes fueron reemplazados por la vecindad alrededor de la bolsa, cuyos vértices corresponde a los inliers más extremos, es decir sus vértices corresponden para cada lado de la bolsa a las observaciones con la mayor distancia d_i.

En lugar de dibujar la mediana de Tukey se puede dibujar el núcleo. Ésto se hace achicando la bolsa en forma relativa a T_*, los vértices del núcleo son $c_j = \xi \cdot w_j \quad j = 1,..., m + m'$ y ξ se asume como 0.1.

Cuando las observaciones están sujetas a una traslación o una transformación lineal no singular (por ejemplo una rotación), el bagplot es transformado en forma acorde, pues como vimos en la sección 2 la profundidad es invariante por transformaciones lineales afines y los conjuntos convexos se transforman en conjuntos convexos. Entonces los puntos dentro de la bolsa siguen siendo interiores y los outliers siguen siendo outliers. Para observar ésto al conjunto de datos del ejemplo de la figura 4.1 se le aplicó una rotación de 90º y 180º, los bagplots obtenidos son los de la figura 4.6 y 4.7 respectivamente.

Comparando los dos gráficos con el gráfico de la figura 4.1 podemos observar que los outliers siguen siendo los casos 14, 16, 52, 53 y 60.

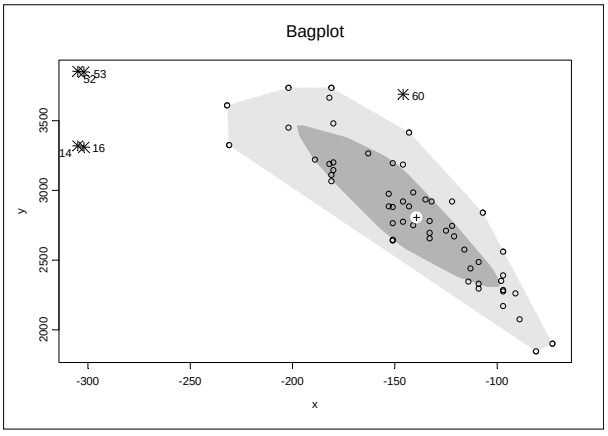


Figura 4.6

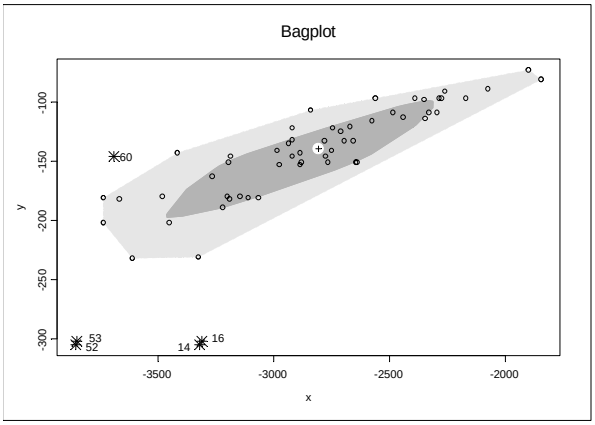


Figura 4.7

4.4 Ejemplos

Cuando el conjunto de datos es pequeño la variabilidad de la valla es muy grande para poder detectar en forma confiable los outliers, por eso para $n < 15$ el algoritmo sólo grafica la mediana de Tukey T_* y segmentos que unen T_* con los puntos de la muestra.

El ejemplo de la figura 4.8 corresponde a una muestra de tamaño 14 de una distribución uniforme en el [0,1] × [0,1] generada al azar.

Las coordenadas de la mediana de Tukey son (0.431961,0.314959).

Además, si el tamaño de la muestra es menor a 15, el algoritmo implementado en S-PLUS emite el siguiente cartel:

"The bag is only plotted when there are at least 15 observations."

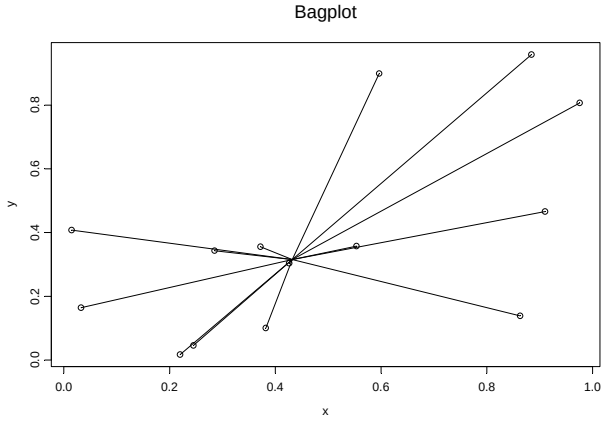


Figura 4.8

Si el conjunto de datos es casi lineal, el bagplot se reduce a un boxplot univariado. En ese caso el algoritmo implementado en S-PLUS calcula la valla superior como $Q_2 + 4(Q_3 - Q_2)$ y la valla inferior como $Q_2 + 4(Q_1 - Q_2)$, donde Q_j es el cuartil j. Esta versión del box-plot es más adecuada que la clásica para distribuciones asimétricas.

El ejemplo de la figura 4.9 corresponde a una muestra de tamaño 25 donde la variable x corresponde a una normal estandar generada al azar, y la variable y fue definida como

$y = 1.5x + 2.5$ para 22 datos y se eligieron 3 datos alejados de la recta., para que más de la mitad de los datos resultasen alineados.

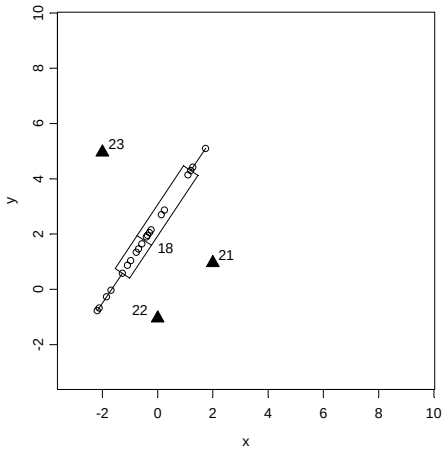


Figura 4.9

En este caso no calcula la mediana de Tukey y el algoritmo implementado en S-PLUS emite el siguiente cartel:
"At least 50% of the data lie on the line: y= 1.5 x+ 2.5"
"Therefore, the bagplot reduces to a univariate boxplot."
Los outliers son identificados con triángulos negros.

En el siguiente ejemplo se generó una muestra de 25 datos donde en la variable x más de la mitad de los datos toman el valor 3 y la variable y corresponde a una normal estandar. El bagplot obtenido es el de la figura 4.10. Como antes no se calcula la mediana de Tukey y el algoritmo implementado en S-PLUS emite el siguiente cartel:

At least 50% of the data lie on the vertical line: x= 3
Therefore, the bagplot reduces to a univariate boxplot.
Los outliers son identificados con triángulos negros.

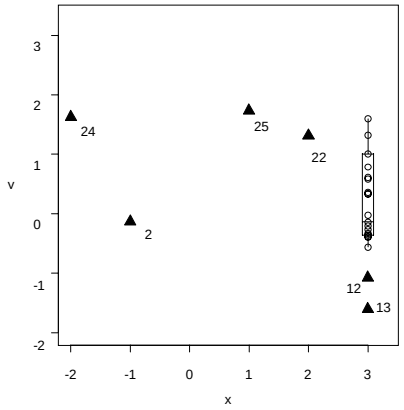


Figura 4.10

En el siguiente ejemplo se generó una muestra de 25 datos donde en la variable *y* más de la mitad de los datos toman el valor 3 y la variable *x* corresponde a una normal estandar. El boxplot obtenido es el de la figura 4.11. Como antes no se calcula la mediana de Tukey y el algoritmo implementado en S-PLUS emite el siguiente cartel:

At least 50% of the data lie on the line: y= 0 x+ 3

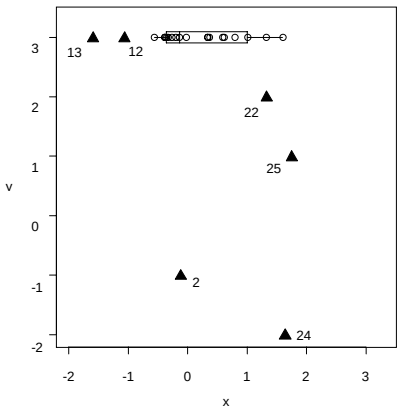


Figura 4.11

Los outliers son identificados con triángulos negros.

Cuando el número de datos es mayor que 150 el algoritmo trabaja con submuestras elegidas al azar de tamaño 150.
En el siguiente ejemplo se generó una muestra de 625 datos donde la variable *x* corresponde a Chi-cuadrado con 15 grados de libertad y la variable *y* corresponde a una normal estandar. El bagplot obtenido es el de la figura 4.12.

Las coordenadas de la mediana de Tukey son (14.1937 , -0.00103823).

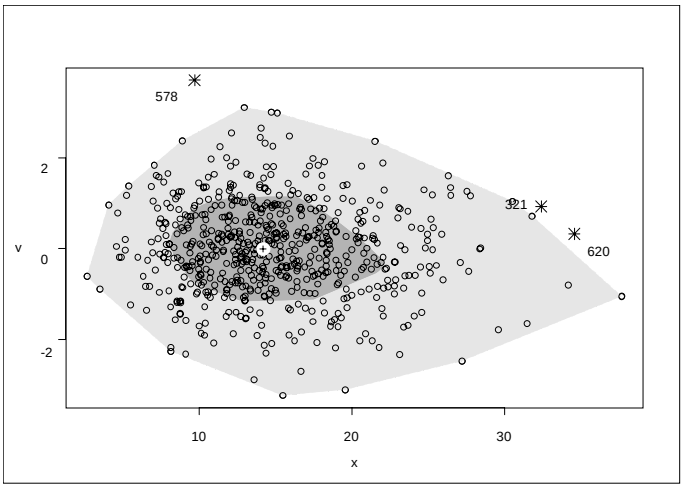


Figura 4.12

El siguiente ejemplo corresponde a una muestra de 21 niños a los que se les registró la edad en meses a la cual cada uno dijo su primera palabra y el puntaje obtenido en la prueba de Gessell tomada mucho despues. El propósito del trabajo fue determinar si la edad de la primera palabra (*x*) podría predecir el puntaje posterior del test (*y*). Los datos se muestran en la siguiente tabla:

Caso	Edad en meses	Puntajeadaptativo de Gessel
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

El bagplot obtenido es el que muestra la figura 4.13.

Las coordenadas de la mediana de Tukey son (11.5594 , 98.262)

Si observamos la figura 4.13 podemos decir que existe una correlación negativa entre las variables x e y , el conjunto de datos es asimétrico, pues la mediana de Tukey se encuentra en la parte superior izquierda de la bolsa donde la vecindad es muy estrecha, su forma además no sugiere una distribución elíptica y se observan 3 outliers, correspondientes a los casos 11, 18 y 19.

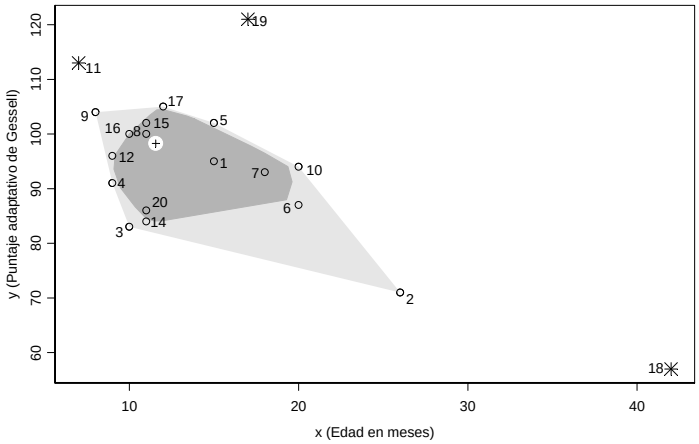


Figura 4.13

El último ejemplo corresponde a Royston y Abrams (1980) que midieron la duración promedio del ciclo menstrual en días (x) y la temperatura basal preovulatoria promedio en ºC (y) de un grupo de 21 mujeres sanas que estaban usando un método de planificación familiar natural. Los datos se muestran en la siguiente tabla:

Caso	Duración	Temperatura
1	22.9	36.44
2	26.3	36.21
3	26.6	36.71
4	26.8	36.13
5	26.9	36.25
6	26.9	36.53
7	27.5	36.41
8	27.6	36.45
9	27.6	36.53
10	28.0	36.31
11	28.4	36.63
12	28.4	36.54
13	28.5	36.52
14	28.8	36.62
15	28.8	36.40
16	29.4	36.48
17	29.9	36.39
18	30.0	36.37
19	30.3	36.77
20	31.2	36.76
21	31.8	36.50

El bagplot obtenido es el de la figura 4.14

Las coordenadas de la mediana de Tukey son (28.2539 , 36.4879).

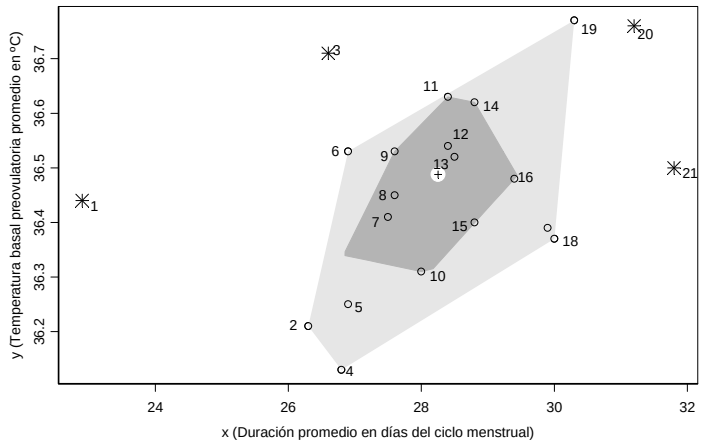


Figura 4.14

Del bagplot, podemos observar que parece existir correlación positiva entre las variables. Hay 4 outliers, los casos 1, 5, 20 y 21. y se podría pensar por la forma de la bolsa y la vecindad en una distribución elíptica.

El coeficiente de correlación de Pearson entre las variables x e y es 0.3649.

Si no consideramos los outliers el coeficiente de correlación es 0.3044.

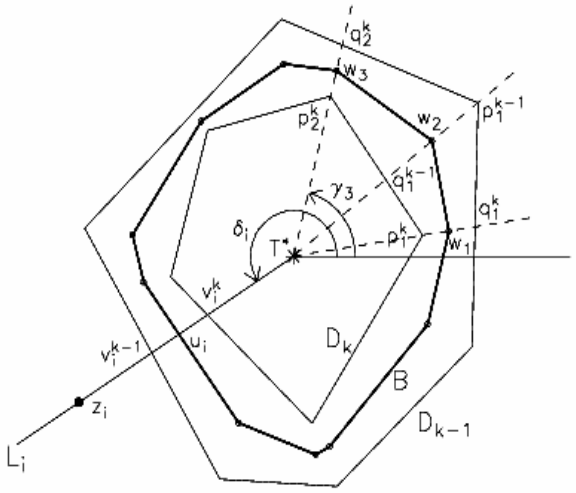

q_2^k
w_3
p_2^k
w_2
p_1^{k-1}
q_1^{k-1}
γ_3
q_1^k
δ_i
T
p_1^k
w_1
v_i^k
v_i^{k-1}
u_i
D_k
B
z_i
D_{k-1}
L_i

Figura 4.5