# Estadística no Paramétrica: Proyecto Final

Paula Rodríguez Díaz y Felipe González Casabianca

21 de mayo de 2017

Este documento cuenta como entregable para el proyecto final del curso: *Estadística no Paramétrica y remuestreo*, dictado en el primer semestre del año 2017 por Adolfo Quiroz en la Universidad de los Andes.

## 1. Profundidades

### Resumen

En esta sección se hablará del concepto de profundidad en estadística, definiendo y explicando las profundidades más usadas. En la siguiente sección se mostrarán ejemplos gráficos comparando algunas de las profundidades mencionadas acá.

### 1.1. Introducción

El concepto de profundidad fue mencionado por primera vez por John Tukey en 1975 ([Tukey, 1975]), como una herramienta para visualizar datos bivariados y desde entonces se ha extendido a datos multivariados con una serie de aplicaciones. La noción de profundidad hace referencia a qué tan *interno*, *propio* o *central* es un cierto elemento dado una muestra multivariada.

Este concepto permite entonces una noción de ordendenamiento (del centro hacia afuera) entre datos multivariados permitiendo hacer análisis no paramétrico para datos en varias dimensiones.

Antes de proseguir con la definición de profundidad y algunos ejemplos, cabe mencionar que un aspecto crucial de trabajar con estas nociones es su demanda computacional. Para que la noción de centro sea útil es necesario que su cómputo sea eficiente, por lo que existe un esfuerzo colaborativo para encontrar nociones y algoritmos eficientes que permitan su uso en dimensiones altas.

**Definición 1.** Una función $D : \mathbb{R}^d \times (\mathbb{R}^d)^n \longrightarrow \mathbb{R}^+$ se considera de profundidad si cumple las siguientes 5 condiciones:

1. **Invariante bajo traslación:** $D(x + x_0, \bar{X} + x_0) = D(x, \bar{X})$ para todo $x_0 \in \mathbb{R}^d$

2. **Invariante bajo aplicaciones lineales:** Para cualquier transformación lineal invertible $A \in \mathbb{R}^d \times \mathbb{R}^d$ se tiene que $D(Ax, A\bar{X}) = D(x, \bar{X})$

3. **Nula en el infinito:** $\lim_{\|x\| \to \infty} D(x, \bar{X}) = 0$

4. **Monótona decreciente en rayos:** Sea $x^* \in \mathbb{R}^d$ tal que $D(x^*, \bar{X})$ es máximo, entonces cualquier $z \in \mathbb{R}^d$ fijo, la función $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ definida como $\psi(\alpha) = D(x^* + z\alpha, \bar{X})$ es monótona decreciente.

5. **Semicontinua por arriba:** los conjuntos de nivel $D_\alpha(X) = \{z \in \mathbb{R}^d : D(z, X) \geq \alpha\}$ son cerrados para todo $\alpha$

Los literales 1 y 2 nos dicen que deformar linealmente o trasladar la muestra no tiene efecto sobre la noción de profundidad. El cuarto nos va servir como fundamento para el algoritmo que construye los contornos que veremos en la siguiente sección.

Adicionalmente, nos referiremos a un **contorno** de radio $r > 0$, para una muestra y una profundidad dada, como el conjunto de elementos del espacio que tiene profundidad $r$.

Veamos las centralidades principales de este resumen. Por simplicidad, escribimos únicamente $D(x)$ y asumimos que existe una muestra general, de elementos de tipo $x$ asociada.

## 1.2. Profundiad de Mahalanobis

La primera noción de profundidad fue propuesta por Mahalanobis en 1936. Para esta noción de profundidad se tiene que los puntos con mismo valor de profundidad forman el contorno de una elipse.

**Definición 2. Profundidad de Mahalanobis 1936** Dada una muestra $S_n \in \mathbb{R}^\rho$ en dimensión $\rho$ con distribución $F$ y segundo momento finito, la versión poblacional de la profundidad de Mahalanobis de un elemento $x \in \mathbb{R}^\rho$, denotada $D^\rho_m(F, x)$,está dada por:

$$D^\rho_m(F, x) = \left[1 + (x - \mu)\Sigma^{-1}(x - \mu)\right]^{-1}$$

Donde $\mu$ es el vector de medias de $F$ y $\Sigma$ la matriz de covarianza de $F$. La versión muestral de la profundidad de Mahalanobis de un elemento $x \in \mathbb{R}^\rho$, denotada $D^\rho_m(x)$, está dada por:

$$D^\rho_m(x) = \left[1 + (x - \overline{X})S^{-1}(x - \overline{X})\right]^{-1}$$

Donde $\overline{X}$ es el vector de media muestral y $S$ la matriz de covarianza muestral.

Unas de las desventajas que presenta esta noción de profundidad es que depende en la existencia de segundo momento, su alta complejidad computacional y que no es una medida robusta dada su dependencia de la media y matriz de covarianza.

## 1.3. Profundiad de Tukey

Esta profundidad fue propuesta por John Tukey en 1975 ([Tukey, 1975]) y, al igual que muchas de las nociones que surgieron después, tiene una motivación completamente geométrica. La definición de la profundidad de Tukey o profundidad de localización es la siguiente:

**Definición 3. Profundidad de Tukey 1975** Dada una muestra $S_n \in \mathbb{R}^\rho$ de tamaño $N$ en dimensión $\rho$ con distribución $F$, la versión muestral de la profundidad de Tukey de un elemento $x \in \mathbb{R}^\rho$, denotada $D_t^\rho(x)$, es el mínimo promedio de elementos de $S_n$ de un lado de un semiespacio (half-space) cerrado cuyo borde incluya a $x$.

Para el caso unidimensional, la profundidad de Tukey se puede ver como:

$$D_t^1(x) = \frac{1}{N} \text{mín}\{|L|, |R|\}$$

donde

$$R = \{\alpha \in S_n \mid \alpha \geq x\} \text{ y } L = \{\alpha \in S_n \mid \alpha \leq x\}$$

y en el caso bidimensional, la profundiad de Tukey de un punto $x$ es el mínimo número de datos de la muestran que quedan al lado de rectas que pasan por $x$.

La versión poblacional de la profundidad de Tukey de un elemento $x \in \mathbb{R}^\rho$, denotada $D_t^\rho(F, x)$ está dada por:

$$D_t^\rho(F, x) = \inf_H \{F(H) : H \text{ es un semiespacio cerrado en} \mathbb{R}^\rho \text{y } x \in H\}$$

## 1.4. Profundidad según capas convexas

Esta profundidad fue propuesta por Barnett en 1976 y no cuenta con una versión poblacional. Aún así, la versión muestral consiste en que dada una muestra $S_n \in \mathbb{R}^\rho$ en dimensión $\rho$, la profundidad de un elemento $x \in \mathbb{R}^\rho$ corresponde al nivel de la capa convexa, construida con los datos de la muestra, a la cual $x$ pertenece.

## 1.5. Profundiad de Oja

Esta profundidad fue propuesta por Hannu Oja en el año 1983 ([Oja, 1983]) y está basada en el cálculo de *símplices* usando elementos de la muestra.

**Definición 4. Profundidad de Oja 1983** Dada una muestra $S_n \in \mathbb{R}^\rho$ en dimensión $\rho$ con distribución $F$, la versión muestral de la profundidad de Oja de un elemento $x \in \mathbb{R}^\rho$, denotada como $D_o^\rho(x)$, es el inverso del promedio muestral de todos los *volumenes* de los *simplices* cerrados, seleccionando como vértices $\rho$ elementos de $S_n$ y $x$.

Note que para el caso unidimensional, esta noción de profundidad se puede definir como:

$$D_o^1(x) = \sum_{i=1}^{n} |x - x_i|$$

Para el caso bidimensional, la profunidad de Oja es la suma de las áreas de todos los triángulos formados con dos elementos de la muestra y $x$.

La versión poblacional de la profundidad de Oja de un elemento $x \in \mathbb{R}^\rho$, denotada como $D_o^\rho(F, x)$, está dada por el valor esperado sobre $F$ del volumen del simplicial generado por $x$ y $\rho$ elementos de la muestra de la siguiente manera:

$$D_o^\rho(F, x) = \left[1 + \mathrm{E}[\text{volumen}(S(x, X_1, ...X_d))]\right]^{-1}$$

## 1.6. Profundiad de Liu

Esta profundidad fue propuesta por Regina Y. Liu en 1990 (Liu et al. [1990]) y al igual que la profundiad de Oja, está basada en *simplices* formados por elementos de la muestra. Se define a continuación:

**Definición 5. Profundidad de Liu 1990** Dado una muestra $S_n \in \mathbb{R}^\rho$ en dimensión $\rho$ con distribución $F$, la versión muestral de la profundidad de Liu de un elemento $x \in \mathbb{R}^\rho$, denotada como $D_l^\rho(x)$, es la cantidad promedio de *simplices* cerrados con vértices en $S_n$ que contienen a $x$

Al igual que las definiciones anteriores, damos las nociones en las primeras dimensiones.

En el caso unidimensional esta profundidad es:

$$D_l^1(x) = \mid \{(\alpha, \beta) \in S_n \times S_n \mid \alpha < \beta,\ \alpha \leq x \leq \beta\} \mid$$

En el caso bidimensional, esta noción corresponde al numero de triángulos formados con tres elementos distintos de $S_n$ que incluyen a $x$ en su interior o frontera.

La versión poblacional de la profundidad de Liu de un elemento $x \in \mathbb{R}^\rho$, denotada como $D_l^\rho(F, x)$, está dada por la probabilidad de que $x$ esté en simplicial generado por $\rho + 1$ elementos aleatorios de la muestra.

$$D_l^\rho(F, x) = F\left[x \in S[X_1, ...X_{\rho+1}]\right]$$

## 1.7. Computación de profundidades

A pesar de que las definiciones de profundidad presentadas anteriormente fuesen intuitivas y sencillas de exponer, la computación de las mismas puede ser bastante compleja, sobretodo para dimensiones altas. Rausseeuw y Hubert (Rousseeuw and Hubert [2015]) calcularon la complejidad computacional para el

cómputo de la profundidad de Tukey y Liu de un dato de una muestra bivariada obteniendo los siguientes resultados[1]:

Cuadro 1: Comlejidad computacional para el cálculo de profundidades

| Profundidad | Complejidad Tiempo |
|---|---|
| Tukey | $O(nlog(n))$ |
| Liu | $O(nlog(n))$ |
| Oja | $O(n^6)$ |

También calcularon la complejidad computacional para el cálculod del punto con mayor profundidad para una muestra bivariada, obteniendo los siguientes resultados:

Cuadro 2: Complejidad computacional para el cálculo de profundidad máxima.

| Mediana | Complejidad Tiempo |
|---|---|
| Tukey | $O(nlog^3(n)))$ |
| Liu | $O(n^4)$ |
| Oja | $O(nlog^3(n))$ |

## 2. Experimentos

En esta sección veremos una comparación gráfica entre algunas de las distintas nociones de profundidad explicadas en la sección anterior. Se realizarán los siguientes experimentos:

1. Comparación de las profundidades sobre una muestra bivariada normal de 300 datos con media cero y varianza $I$.
2. Comparación de las profundidades sobre la misma muestra anterior con 80 elementos de ruido (ejemplo bimodal).
3. Comparación de las profundidades sobre una muestra bivariada de 300 elementos con distribución uniforme.

Todos los ejemplos siguientes se realizaron con el uso de las librerías *depth*(Genest et al. [2017]) y *ddalpha*(Pokotylo et al. [2016]). La primera mencionada, aunque más reciente, está principalmente concentrada en la profundidad de Tukey. La segunda es más general y con soporte a más centralidades, recomendamos esta última para proyectos con profundidades.

[1]El dato para la profundidad de Oja fue extraído de Liu et al. [2006] y corresponde a la versión fuerza bruta del algoritmo.

Aunque la librería *ddaplha* tiene soporte para el dibujo de contornos, no encontramos esta libreria sino luego de realizar los experimentos con *depth*, la cual solo ofrece la construcción de contornos para la profundidad de Tukey. Por lo tanto, desarrollamos el siguiente esquema para la construcción de contornos:

Dada una muestra $\bar{X}$, una centralidad D y un número positivo r, se procede de la siguiente manera:

1. Sea $\delta = \text{máx}(dist(x, y) \mid x, y \in \bar{X})$
2. Encontrar la mediana de la muestra según D. Esto traduce al elemento $x^* \in \bar{X}$ tal que $D(x, \bar{X})$ es máximo. Esto se puede hacer directamente a través de las librerías mencionadas.
3. Se selecciona un ángulo $\theta \in [0, 2\pi)$
4. Se realiza una búsqueda binaria entre $x^*$ y elemento que este a una distancia $\delta$ de el en dirección $\theta$, a lo largo de la linea recta que los une, hasta encontrar un elemento $x_\theta$ que tenga profundidad $r$.
5. Se repite este procedimiento para varios ángulos hasta obteniendo así un conjunto de elementos todos con profundidad $r$, alrededor de $x^*$.
6. Finalmente se unen estos puntos para obtener una aproximación del contorno de radio $r > 0$

Note que dicho procedimiento funciona gracias a que las profundidades son monótonas decrecientes sobre rayos. Aunque dicha propiedad funciona sobre un elemento con profundidad máxima del espacio, que no necesariamente está dentro de la muestra, la mediana se encuentra lo suficientemente cerca a dicho punto para que este procedimiento funcione para radios en cierto rango $r$.

Para las gráficas que se encuentran a continuación, se pueden ver cuatro contornos. Estos corresponden a cuatro valores uniformemente escogidos entre la mínima y la máxima profundidades de la muestra.

## 2.1. Experimento 1

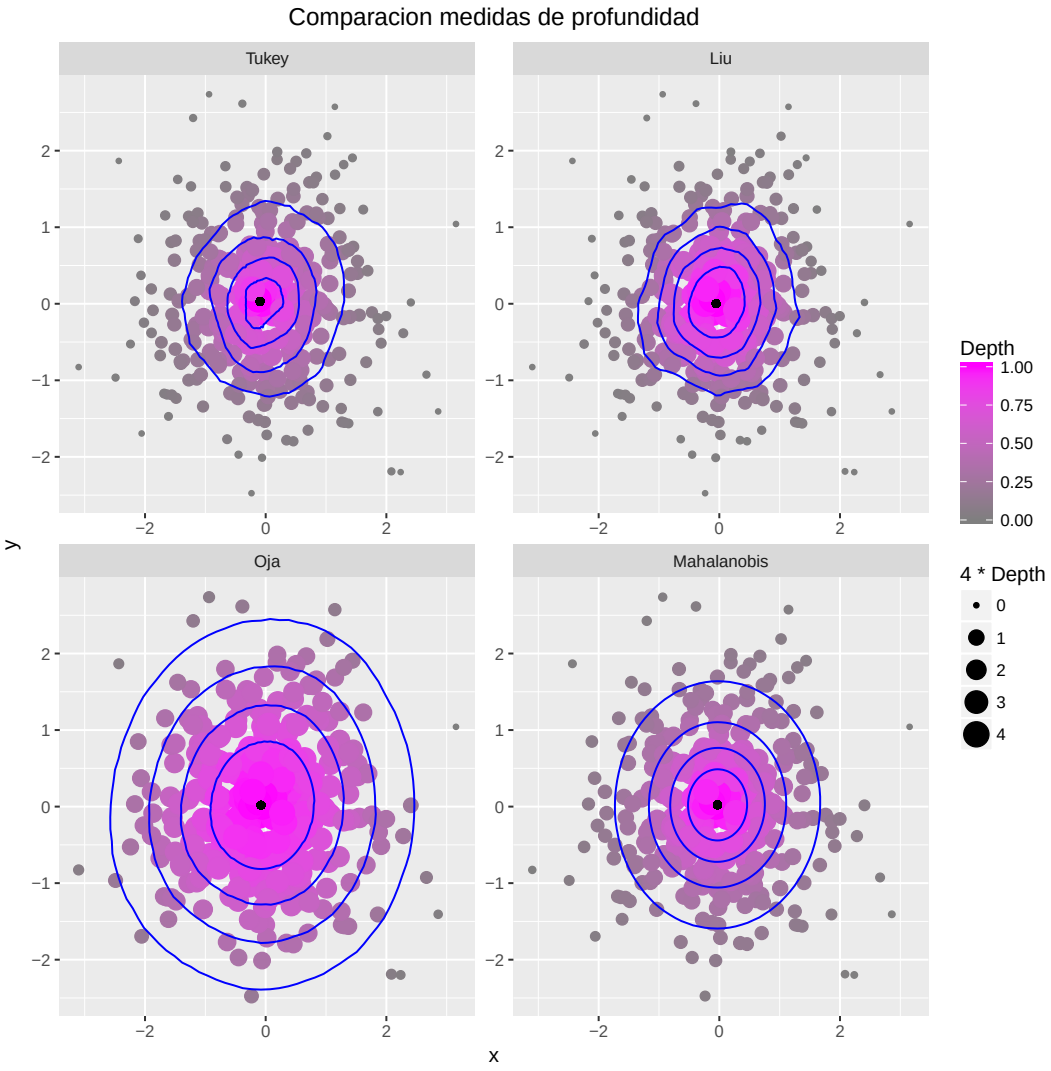


Figura 1: Comparación entre profundidades para una muestra normal bivariada $(0, I)$. Todas las muestras tienen una tendencia elíptica, cabe notar como la centralidad de Oja admite mas elementos como parte de la muestra, mientras que las otras tres profundidades excluyen a muchos elementos a medida que se alejan del elemento central.

## 2.2. Experimento 2

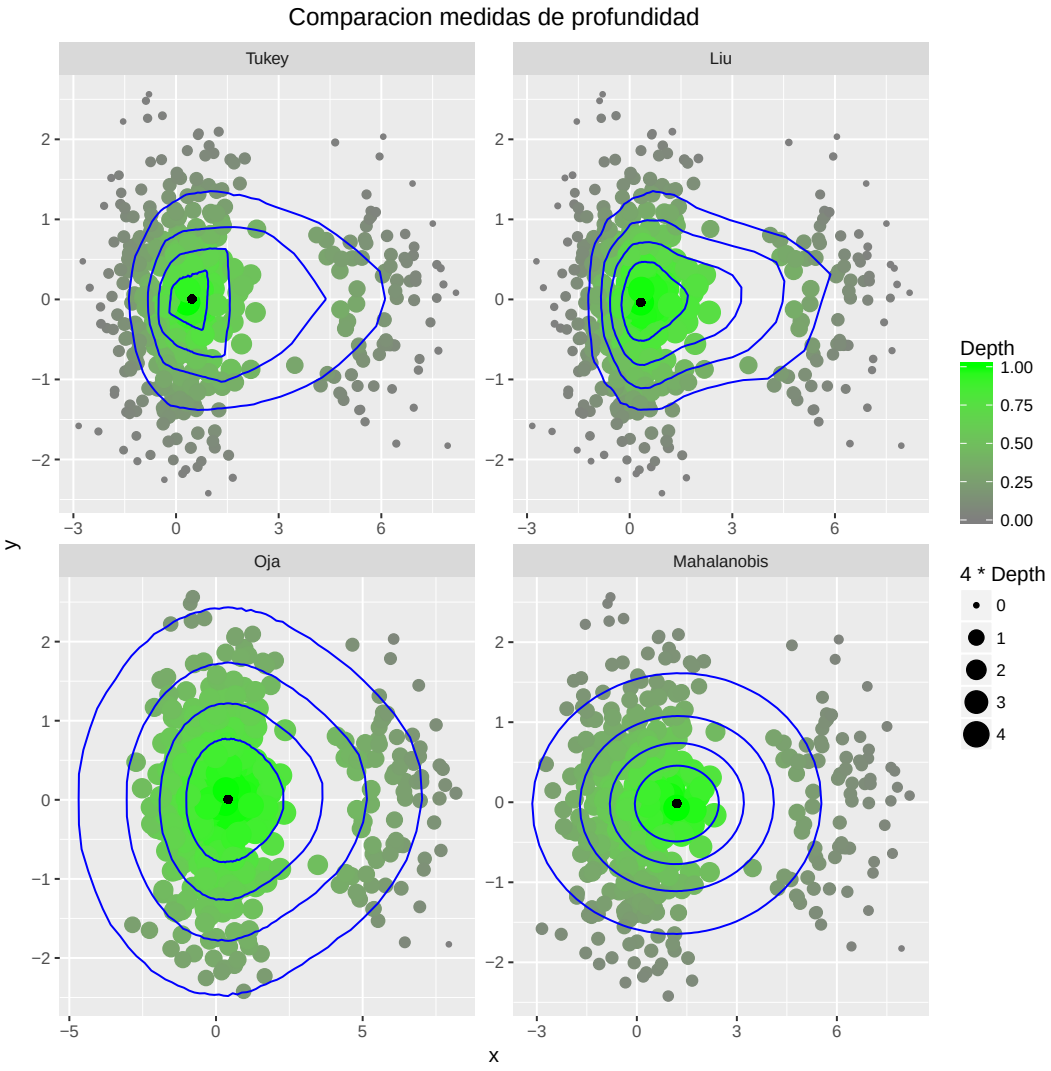


Figura 2: Comparación entre profundidades para una muestra normal bivariada $(0, I)$ con ruido. Acá cabe notar como la profundidad de Tukey y después la de Oja, aíslan muy bien la muestra principal del ruido, esforzándose por considerar como centrales, únicamente los elementos de la muestra principal.

## 2.3. Experimento 3

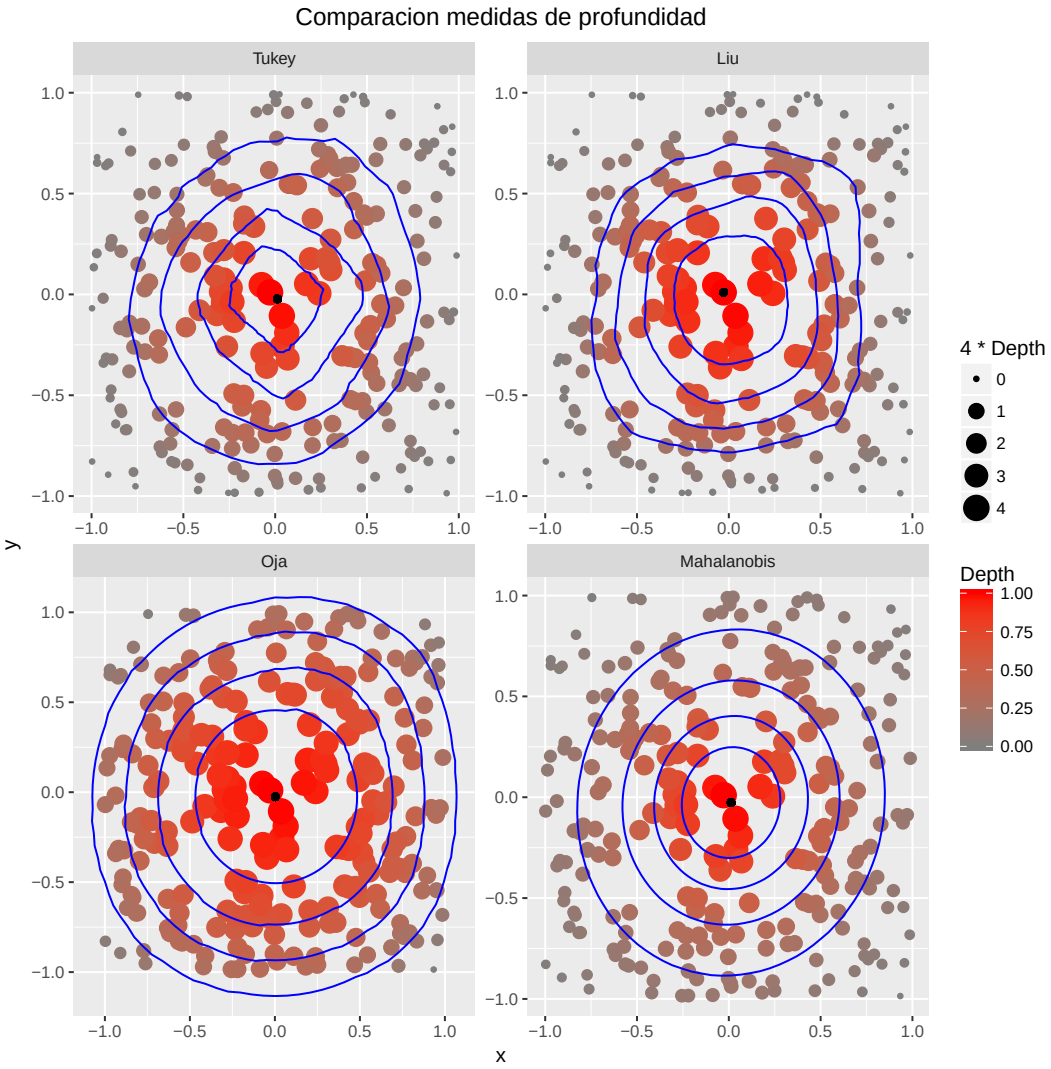


Figura 3: Comparación entre profundidades para una muestra uniforme bivariada. Para este caso, los contornos idealmente deben corresponde a rectángulos y se ve como la profundidad de Liu crea contornos que tienden a esta forma geométrica.

# 3. Aplicaciones

En esta sección se quiere comentar un poco sobre las posibles aplicaciones que pueden tener las nociones de profundidad:

- **Ordenamiento multivariado de muestras**. Durante la primera mitad del semestre, construimos pruebas no paramétricas utilizando estadísticos de orden (estadísticos basados en el ordenamiento univariado de una muestra dada). Las nociones de centralidad permiten extender la noción de orden a varias variables, organizando los elementos según su profundidad.

- **Detección de outliers.** En procesos industriales estandarizados, un producto final, con sus características asociadas, se puede ver como un elemento dentro de una muestra histórica de productos. Resulta de interés tener un mecanismo que detecte si cierta combinación de características resulta patológica para un producto como control de calidad. Una forma de enfrentar este problema es asumir que los productos se distribuyen normal y ver en que percentil se encuentra dicho producto. Sin embargo podemos, sin asumir una distribución sobre los datos, calcular la centralidad del nuevo producto respecto a la muestra histórica y decidir sobre su irregularidad a través de este valor. Nuevos productos con centralidades bajas, son indicio de que no cumplen los controles de calidad o que se alejan del estándar regular de la empresa.

- **Detección de clusters**. Un procedimiento usual para encontrar clusters dentro de muestras en un espacio multidimensional es la rutina $k-means$. Dicho algoritmo se basa en encontrar el centro geométrico de una muestra e ir actualizando los clusters según su cercanía a dicho punto. Sin embargo, este procedimiento no resulta tan eficaz al enfrentarse con muestras que no siguen una geometría circular. Este fenómeno se puede mitigar con las nociones de profundidad mencionadas. La noción de centro geométrico se puede reemplazar por la noción de mediana (según profundidad) y en lugar de asociar elementos a los centros geométricos mas cercanos vía distancia euclidiana, se pueden asociar a la muestra en la que tengan mayor profundidad. Esta opción ha sido explorada por varios autores: (Torrente and Romo)(Zhang et al. [2013])

# Referencias


Maxime Genest, Jean-Claude Masse, and Jean-Francois Plante. *depth: Nonparametric Depth Functions for Multivariate Analysis*, 2017. URL `https://CRAN.R-project.org/package=depth`. R package version 2.1-1.

Regina Y Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88 (421):252–260, 1993.

Regina Y Liu, Jesse M Parelius, Kesar Singh, et al. Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, 27(3):783–858, 1999.

Regina Y Liu, Robert Joseph Serfling, and Diane L Souvaine. *Data depth: robust multivariate analysis, computational geometry, and applications*, volume 72. American Mathematical Soc., 2006.

Regina Y Liu et al. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

Karl Mosler. Depth statistics. In *Robustness and complex data structures*, pages 17–34. Springer, 2013.

Maria Raquel Neto. The concept of depth in statistics.

Hannu Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332, 1983.

Oleksii Pokotylo, Pavlo Mozharovskyi, and Rainer Dyckerhoff. Depth and depth-based classification with r-package ddalpha. *arXiv:1608.04109*, 2016.

Peter J. Rousseeuw and Mia Hubert. Statistical depth meets computational geometry: a short survey. 2015.

Aurora Torrente and Juan Romo. Refining k-means by bootstrap and data depth.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531, 1975.

Zhanpan Zhang, Xinping Cui, Daniel R Jeske, and James Borneman. Biclustering scatter plots using data depth measures. *Statistical Analysis and Data Mining*, 6(2):102–115, 2013.