ANÁLISIS MULTIVARIADO SOBRE EL DIAGNÓSTICO DEL CÁNCER DE MAMA EN EL ESTADO DE WISCONSIN

Oscar J Layton Neyfer L Gomez [a]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

Se desarrolla la descripción de tipo multivariado para el conjunto de datos pertenecientes al diagnóstico del cáncer de mama en el estado de Wisconsin proporcionados por el *Repositorio de aprendizaje automático de la UCI*, en miras a realizar un análisis con soporte a los diversos métodos estadísticos desarrollados en la teoría que al ser usados junto con el software Estadístico R permitirán concluir en base a las distintas representaciones gráficas y resultados obtenidos.

Introducción

El uso de medidas estadísticas univariadas ha sido parte importante en caracterizar la distribución muestral de la población, ahora bien los datos referentes a las biopsias mamarias son de tipo mutivariado, cada variable lleva consigo una relación implícita lo cual conlleva hacer un análisis relacionando todas las variables propias de su estado en la naturaleza, para llevar a cabo tal análisis se adopta de una estructura que permita hacer relación de las diferencias en cuestión, como primer aspecto característico que permita conocer la información obtenida conociéndose su origen y datos mas relevantes.Todo lo anterior haciendo un estudio detallado a los datos extraídos por medio de un análisis descriptivo.

Como segundo aspecto, se da lugar a la metodología multivariada no paramétrica usando estadísticas basadas en métodos robustos, todo lo anterior sujeto al concepto de profundidad, su uso permite la representación gráfica de la información suministtrada por las observaciones multivariadas, como lo permite el bagplot, el cual permitirá observar características importantes para establecer de que distribución provienen los datos y proporcionará metodologías para la detección de outliers.

Biopsia por aspiración con aguja fina

El conjunto de datos relativos al diagnóstico del cáncer de mama en el estado de Wisconsin fueron recopilados a partir de las diversas características medidas en una imagen digitalizada de un aspirado con aguja fina (FNA) en masas mamarias, procedimiento que describe características de los núcleos celulares presentes en la imagen. Tal proceso de medición en núcleos celulares es proporcionado por Cruz & Martinez de Larios (2002) quienes además de considerarlo importante destacan su uso desde a mediados del siglo pasado, procedimiento que viene destacándose para detectar lesiones de tipo clínico, sin embargo su uso a lo largo del tiempo se ha generalizado para la evaluación de anormalidades mastográficas, conllevando a la detección del cáncer.

El uso de la técnica para el análisis de masas mamarias, se ha generalizado por los resultados obtenidos puesto que se ha encontrado un alto grado de concordancia con los diagnósticos histopatológicos, generando una evaluación rápida y a tiempo del diagnostico. Las mediciones son realizadas a partir de características a partir de imágenes (ver 1). No obstante, las imágenes digitalizadas en el procedimiento (FNA) son presentadas en la web en miras de ser utilizadas para su análisis para más información ver Mathematical Programming in Machine (2001)

[a]Estadístico. Universidad Nacional de Colombia

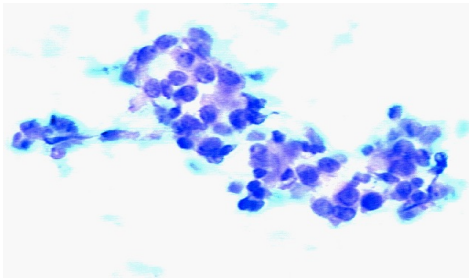


FIGURA 1: Imagen digitalizada diagnostico cáncer de mama Wilconsin, Mathematical Programming in Machine (2001)

El conjunto de datos está conformado por 569 observaciones medidas a 33 variables, para realizar el presente análisis se tomarán las tres primeras las cuales son de tipo numérico (ver tabla 1).

TABLA 1: Datos alusivos a diagnósticos de cáncer de mama en Wisconsin

ID	**Diagnosis**	**Compactness**	**Smoothness**	**Symmetry**
1	M	0.28	0.12	0.24
2	M	0.08	0.08	0.18
3	M	0.16	0.11	0.21
4	M	0.28	0.14	0.26
5	M	0.13	0.10	0.18
⋮	⋮	⋮	⋮	⋮

- X_1 : **Compactness** (compacidad): Variable medida en $(\frac{P^2}{A} - 1)$, donde P es el perímetro y A el area de la masa estudiada)
- X_2 : **Smoothness** (suavidad): La cual hace referencia a la variación local en longitudes de radio.
- X_3 : **Simetría**
- **Diagnosis:** Hace referencia al diagnostico generado y posee dos valores M: Maligno y B: Benigno.

Como etapa inicial, para un análisis descriptivo de los datos se la matriz de varianzas y correlaciones para de esta manera poder caracterizar el conjunto de datos.

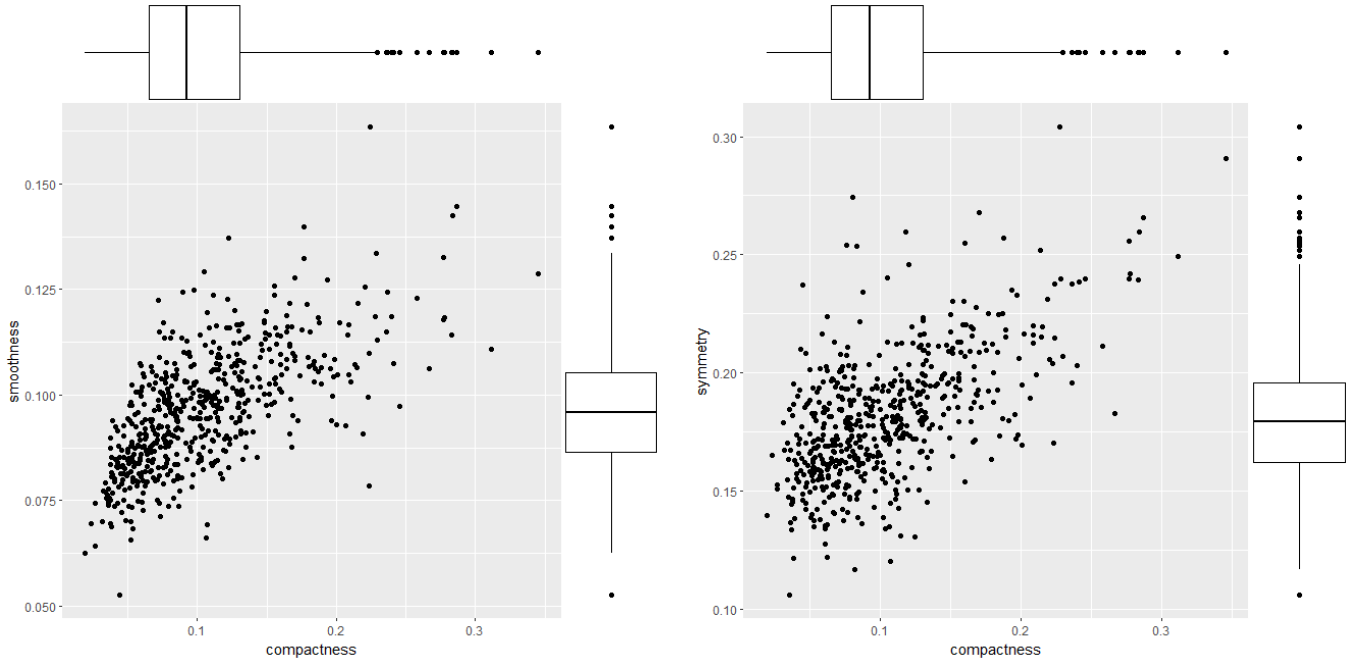
$$\overline{\mathbf{X}}' = \begin{bmatrix} \overline{X}_1 & , & \overline{X}_2 & , & \overline{X}_3 \end{bmatrix} = \begin{bmatrix} 0.10434 & , & 0.09636 & , & 0.18116 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 0.002784 & 0.000488 & 0.000870 \\ 0.000488 & 0.000197 & 0.000214 \\ 0.000870 & 0.000214 & 0.000750 \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & 0.659 & 0.602 \\ 0.659 & 1 & 0.577 \\ 0.602 & 0.577 & 1 \end{bmatrix}$$

Observando la matriz de correlaciones, se puede observar que son todas positivas, indicando la existencia de una asociación lineal esto puede ser observado en la orientación de los puntos en el eje coordenado como es el caso de el scatter plot 2a el cual caracteriza a las variables X_1 :compacidad y X_2 :suavidad el hecho de r_{12} ser positiva indica que conjuntamente los valores aumentan para valores mayores y disminuyen para menores.

En lo que corresponde al boxplot para la variable Compacidad, se puede observar que el 50 % de la información está entre $1Q = 0.06492$ y $3Q = 0.13040$, no obstante en el caso univariado los valores mayores a $Q3+1.5RI = 0.22862$ son considerados outliers para Compacidad. Así mismo, en lo que corresponde a la variable Suavidad se puede observar que el 50 % de la información está entre $1Q = .08637$ y $3Q = .10530$, no obstante los valores mayores a $Q3 + 1.5RI = 0.133695$ y menores que $Q1 - 1.5RI = 0.057975$ son considerados outliers para la variable Suavidad.

En el scatter plot 2b se observa una relación entre las variables X_1 :Compacidad y X_3 : Simetría de tipo lineal el cual es observado en la correlación r_{13} positiva, indicando tendencia a valores grandes de compacidad con valores grandes de simetría conjuntamente y valores pequeños de suavidad con pequeños en simetría.



(A) (B)

FIGURA 2: Scatter plot y boxplot

La existencia de una asociación lineal puede ser observado en la orientación de los puntos en el eje coordenado como es el caso de el scatter plot 3 el cual está caracterizado por las variables X_s : Suavidad y X_2 : Simetría, el hecho de r_{23} ser positivo indica que conjuntamente los valores aumentan para valores mayores y disminuyen para menores.

En lo que corresponde al boxplot para la variable Simetría, se puede observar que el 50 % de la información está entre $1Q = 0.1619$ y $3Q = 0.1957$, así mismo, para valores mayores a $Q3 + 1.5RI = 0.2464$ y menores que $Q1 - 1.5RI = 0.1112$ son considerados outliers para la variable Simetría.

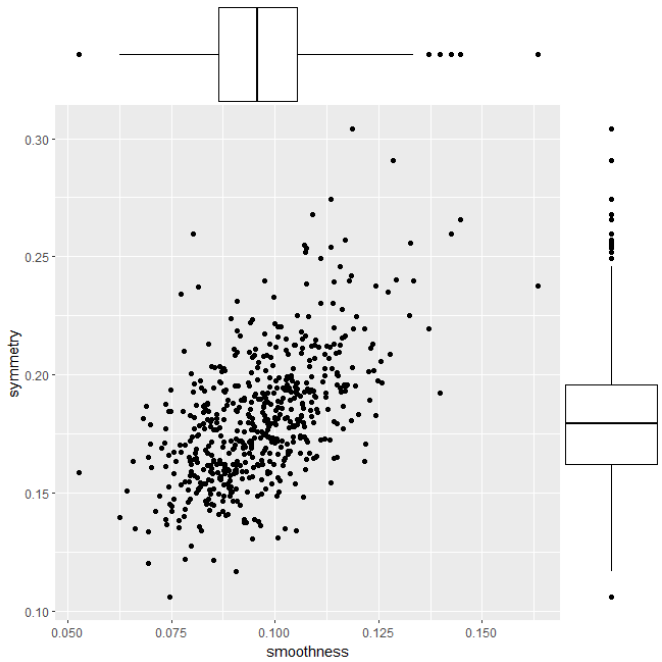


FIGURA 3: Scatter plot y boxplot para las variables Suavidad y Simetría

Según Johnson y Wichern (2002) Las cantidades s_{ik} y r_{ik} no transmiten, en general, todo lo que hay que saber sobre la asociación entre dos variables. Pueden existir asociaciones no lineales que no sean reveladas por estas estadísticas descriptivas. Covarianza y correlación proporcionan medidas de asociación lineal, o asociación a lo largo de una línea (p.9). Para establecer relaciones entre los datos multivariados se realizaran métodos no paramétricos asociados al concepto de profundidad.

No obstante el usar los métodos no paramétricos para el conjunto de datos tiene sus grandes ventajas, puesto que no se considerará si las distribuciones marginales son normales, como es el caso actual el cual se determina que las distribuciones marginales no son normales.

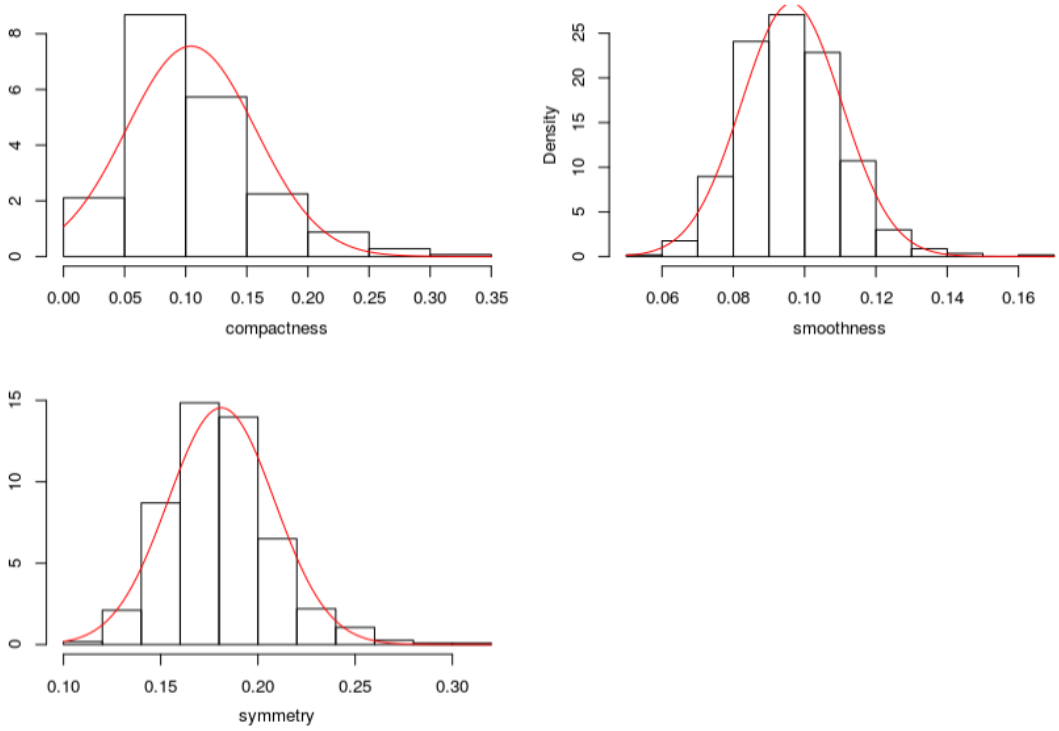


Figura 4: Bar plot de variables

En los barplot proporcionados en la figura 4, se logra identificar la forma de la distribución por variables,lográndose observar que las distribuciones son sesgadas como es el caso de la variable Compacticidad y simetría se encuentras sesgadas a derecha, lo cual indica que al parecer no provienen de distribuciones normales.

Tabla 2: Prueba de normalidad univariada

	Test	Variable	Statistic	p value	Normality
1	Shapiro-Wilk	compactness	0.9170	<0.001	NO
2	Shapiro-Wilk	smoothness	0.9875	1e-04	NO
3	Shapiro-Wilk	symmetry	0.9726	<0.001	NO

Para determinar si en realidad no son normales se procede a realizar los respectivos QQplot lo cuales se presentan en la figura 5, donde se puede identificar que para todas las variables (Compacticidad, Suavidad y Simetría) los datos no se encuentran alrededor de la línea recta, lo cual da indicios de no cumplimiento del supuesto de normalidad. Para confirmar este hecho se realizan las pruebas de Shapiro-Wilk (Ver Tabla 2) para contrastar la hipótesis:

$$H_0 : F_n(w) = N(\mu_0, \sigma_0^2) \qquad vs \qquad H_1 : F_n(w) \neq N(\mu_0, \sigma_0^2)$$

Hay evidencia estadistica para rechazar H_0 por lo cual se afirma el hecho de que las distribuciones no son normales.

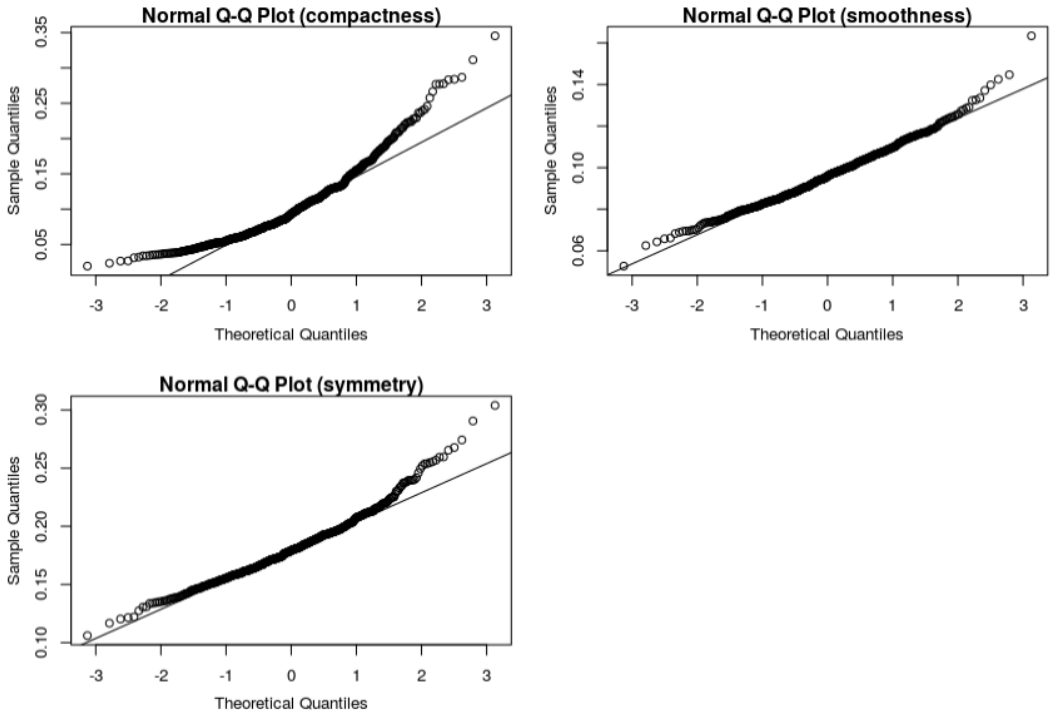


FIGURA 5: QQ plot por variables

Chi-square plot

Para determinar si la distribución de los datos es normal multivariada (Normal 3-variada) se calcula la fracción de los puntos dentro de un contorno, para de esta manera compararlo con la probabilidad teórica es un procedimiento útil, Conociendo que:

$$(\mathbf{X}-\mu)' \sum^{-1} (\mathbf{X}-\mu) \leqslant \chi_p^2(\alpha)$$

Remplazando $\sum^{-1}$ y μ por sus estimaciones $\mathbf{S^{-1}}$ y $\overline{\mathbf{X}}$ respectivamente, se calculan las distancias al cuadrado:

$$d_j^2 = (\mathbf{X_j} - \overline{\mathbf{X}})' S^{-1} (\mathbf{X_j} - \overline{\mathbf{X}}) \qquad con \qquad j = 1, ...569$$

Para generar el Chi-square plot se consideran los pasos descritos por Johnson y Wichern (2002) ordenando las distancias al cuadrado de menor a mayor $d_{(1)}^2 \leq d_{(2)}^2 \leq d_{(3)}^2, ..., \leq d_{(n)}^2$ y graficando los pares $\left(q_3(\frac{j-\frac{1}{2}}{569}), d_{(j)}^2\right)$, siendo $q_3(\frac{j-\frac{1}{2}}{569})$ el cuantil de un chi-cuadrado. obteniéndose la siguiente gráfica 6

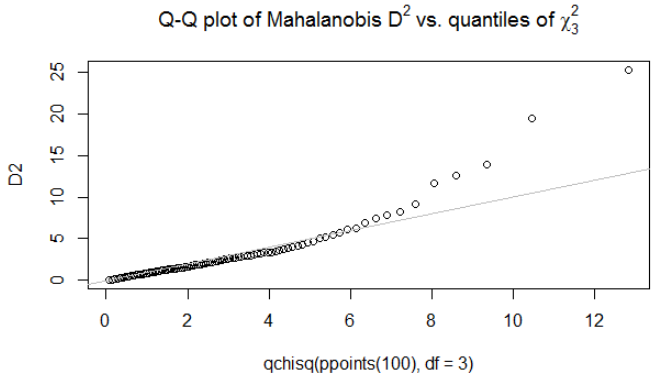


FIGURA 6: Chi-square plot con distancias ordenadas

Observándose la gráfica se puede ver que los puntos no se encuentran en la linea recta, las distancias mahalanobis mayores son demasiado grandes en relación a distancias esperadas de las poblaciones normales trivariadas. Por consiguiente los datos provenientes de diagnostico de cáncer de mama en el estado de Wisconsil no provienen de una distribución normal trivariada.

Outliers

En el caso univariado la detección de datos atípicos se realizó observando los boxplot por variable, sin embargo en el caso bivariado no es tan simple, han de considerarse todas las variables al mismo tiempo, este procedimiento de encontrar las observaciones atípicas puede hacerse por medio del Chi-square plot siendo los candidatos las observaciones con mayor distancia Mahalanobis, no obstante el observar los scatter plot darán algún indicio.

En la figura 7 se muestran situaciones con muchas observaciones inusuales, pese a esto en la esquina superior derecha de los scatter plots se elimina un patrón de tipo elíptico por tales datos. En consecuencia, se procede a detectar las observaciones atípicas considerando los pasos determinados por Johnson y Wichern (2002) los cuales establecen estandarizar el conjunto de datos:

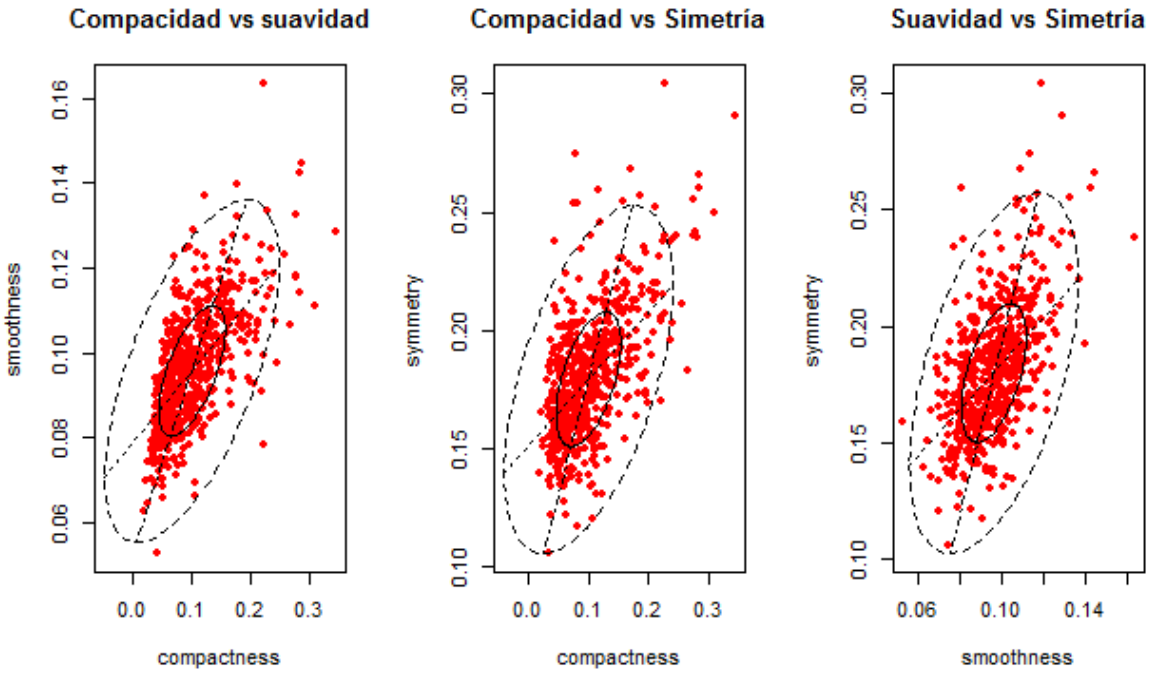


Figura 7: Patrones elípticos en scatterplot

Para determinar las observaciones inusuales, se procede a estandarizar las observaciones y calcular las distancias Estadísticas en una lista para poder contrastarse por un percentil apropiado de la distribución de chi-cuadrado con 3 grados de libertad. En consecuencia se buscan las distancias mayores al estadístico $\chi^2_3(0.005) = 12.83$. La tabla 3 revela los resultados llegándose a 17 observaciones outliers multivariantes de los datos provenientes al diagnóstico de cáncer de mama en Wisconsil.

Tabla 3: Valores estandarizados y que son posibles outliers

Obs	z_1	z_2	z_3	d^2
4	3.40	3.28	2.87	13.95
13	2.68	0.07	2.14	13.95
26	2.34	1.58	4.48	21.35
61	-0.45	1.21	3.40	22.02
77	0.01	2.33	2.16	13.34
79	4.57	2.29	4.00	25.32
83	3.07	0.71	0.06	15.90
113	2.25	-1.28	-0.39	20.19
123	3.46	3.44	3.08	15.08
151	-0.54	1.22	2.66	15.56
182	3.39	1.26	2.13	13.38
259	3.92	1.03	2.49	20.21
289	0.26	-1.15	2.86	19.18
425	-0.40	0.79	2.65	13.49
444	-1.13	-1.06	2.05	14.89
505	2.27	4.77	2.07	24.22
521	0.34	2.90	1.41	13.08

No obstante es importante distinguir los outliers dentro de los datos, la gráfica que permite identificar los outliers gráficamente se presenta a continuación (ver figura 8), cabe notar que las observaciones atípicas son coloreadas de color negro:

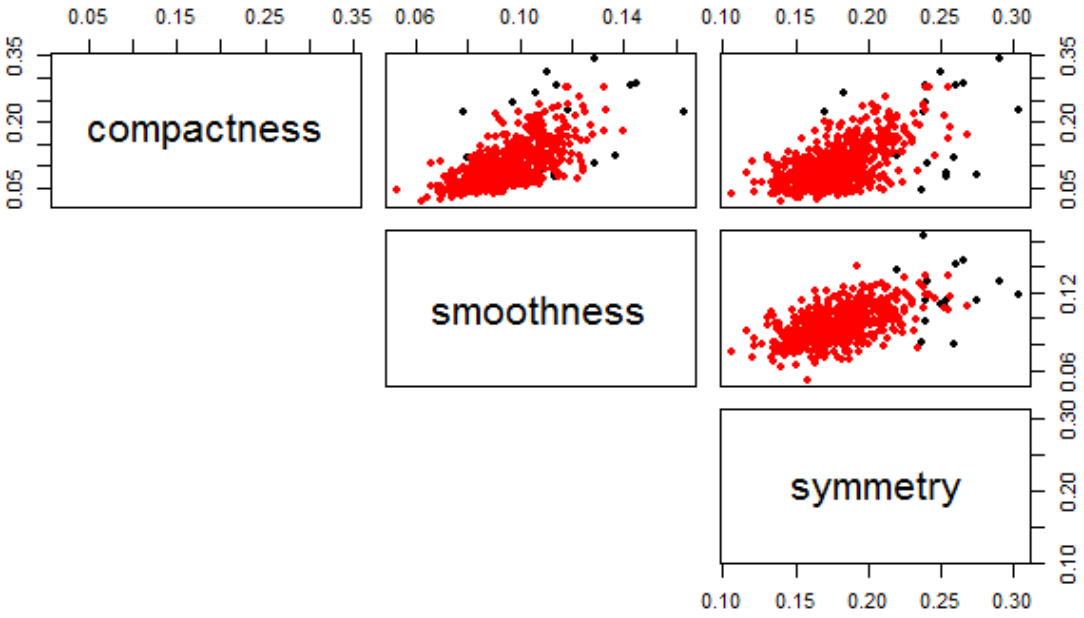


Figura 8: Observaciones outliers multivariadas en el conjunto de datos

Metodología (Métodos no paramétricos basados en profundidad)

Teniendo en cuenta que el análisis multivariado clásico se basa en gran medida en diversos supuestos como el de la normalidad, existen muchas situaciones en la que su cumplimiento es de poca frecuencia y más si se tratase en su forma multivariada, actualmente existe otro tipo enfoque para abordar este tipo de problemas como lo es el enfoque no paramétrico basado en profundidad, el cual será abordado en su parte fundamental en el análisis de los diagnósticos de cáncer de Wisconsin. Además se de ser parte articulador permitirá como lo señala Liu, R., Parelius, J. y Singh, K. (1999) en determinar características cuantitativas, gráficas multivariadas de distribución junto con métodos de inferencia.

Para realizar una descripción de la distribución multivariante, surge principalmente el problema de poder determinar una forma de ordenación entre objetos de R^p, esta necesidad conlleva el gran aporte del concepto de profundidad el cual según Zuo y Serfling. (2000) es una función D (x; P) que proporciona una ordenación hacia el centro. Tukey (1975) uno de los pioneros en este campo propone una profundidad de "medio espacio" y sugiere su papel en la definición de estadísticas multivariadas de orden y rango univariable a través de "contornos" inducidos por la profundidad.(p.1). Por consiguiente se presenta la siguiente definición la cual sera artifice de los procedimientos subsecuentes:

Función de profundidad: Dada una distribución de probabilidad P en R^d, una función de profundidad es una función acotada que asigna a cada punto de R^d su grado de centralidad respecto de P.

$$D\,(\,.\,;P):\ R^d\ \mapsto\ [0,1]$$

La función suele llamarse de profundidad se satisface las siguientes propiedades:

- $D(Ax+b;P_{AX+b}) = D(x;P_x)$ para cualquier x en R^d, cualquier matriz de dimensión $d \times d$ no singular y cualquier vector b de dimensión d, donde donde F_x es la función de distribución del vector aleatorio X.
- $Lim_{\|x\|\to\infty} = D(x;P) = 0$.
- $D(\theta;P)Sup_{x\in R^d}D(x;P)$ Para todo P con θ el centro.
- Si θ es el punto con mayor profundidad entonces: $D(x;P) \leqslant D(\theta+\lambda(x-\theta);P)$ para cualquier $0 \leq \lambda \leq 1$.
- El conjunto de puntos cuya profundidad es al menos α es cerrado. $D_\alpha = D_\alpha(P) = \{x \in R^d : D(x;P) \geq \alpha\}$, D_α constituye una región central y aveces se le considera $\alpha - trimmed$.

La Profundidad de datos y ordenación de observaciones multivariadas puede realizarse por medio de cualquier función de profundidad, de las cuales se destacan:

Función profundidad Mahalanobis M_{MAH} está definida como:

$$M_{MAH}(x;P) = \left[1 + (x-\mu)\sum\nolimits_F^{-1}(x-\mu)\right]^{-1}$$

Donde μ y $\sum_F$ son el vector de medias y la matriz de dispersión de F respectivamente. La versión muestral de M_HD es obtenida remplazando μ_F y $\sum_F$ por sus estimaciones muestrales.

Función profundidad Euclidiana D_{EUK} está definida como:

$$D_{EUK}(x;P) = [1 + \|x - \overline{x}\|]^{-1}$$

Projection depth $D(x,P)_{PRO}$ Para todo $x \in R^d$ con $d > 1$

$$PD_F(x) = \left[1 + sup_{||u||=1}\frac{|u^{'}x - Med(u^{'}X)|}{MAD(u^{'}X)}\right]^{-1}$$

A continuación se presentan las distancias por medio de las funciones Mhalanobis, Euclidea, Profundidad y de Tuckey:

Tabla 4: Tipos de distancias para las primeras 20 observaciones del diagnóstico de cáncer de mama en Wisconsil

	D.Euclidiana	D.Mahalanobis	D.Poyeccion	D.Tukey
1	0.97	0.08	0.13	0.01
2	1.00	0.50	0.37	0.15
3	1.00	0.43	0.29	0.12
4	0.96	0.07	0.13	0.00
5	1.00	0.68	0.41	0.20
6	0.99	0.16	0.22	0.01
7	1.00	0.93	0.58	0.34
8	0.99	0.25	0.27	0.05
9	0.99	0.15	0.21	0.02
10	0.98	0.12	0.17	0.01
11	1.00	0.42	0.36	0.09
12	1.00	0.74	0.42	0.22
13	0.98	0.07	0.12	0.00
14	1.00	0.40	0.32	0.10
15	0.98	0.14	0.17	0.02
16	0.99	0.23	0.24	0.04
17	1.00	0.39	0.34	0.07
18	0.99	0.22	0.21	0.04
19	1.00	0.43	0.35	0.10
20	1.00	0.57	0.38	0.15
⋮	⋮	⋮	⋮	⋮

Para hacer una distinción de adentro hacia afuera, se calculan las profundidades de toda la muestra ver 4 $\{x_1, ..., x_n\}$, por consiguiente se procede a ordenar de menor a mayor, sea el punto $X_{[i]}$ el punto de la muestra asociado con la mayor profundidad (mediana). A continuación se presenta la tabla 5 que presenta las medianas dependiendo del tipo de función usada.

Tabla 5: Mediana por diferentes tipos de funciones de profundidad

	compactness	smoothness	symmetry
Euclideana	0.1041	0.1008	0.1813
Mahalanobis	0.1117	0.0974	0.1807
Proyección	0.1034	0.0978	0.1752
Tukey	0.1022	0.0942	0.1769

Se puede observar que las medianas dependiendo de la función de profundidad son diferentes, sin embargo no varían mucho una de las otras.

Estimadores robustos

Como se ha observado anteriormente se han encontrado diferencias entre las estimaciones de la mediana multivariada dependiendo de la función de profundidad seleccionada, en ocasiones anteriores se identifica a la mediana como el centro de las observaciones, es decir el punto más profundo es el estimador de la mediana multivariante, sin embargo en el sentido robusto Lopez, (2010) especifica que tal mediana deja de ser eficiente en este sentido se encuentran los conceptos de los L-Estadísticos univariantes, que consisten en una ponderación de los elementos muestrales de forma que sea posible eliminar la influencia de los puntos más externos que puedan tener un elevado índice de atipicidad (p.30).

Ahora bien tomando la mediana multivariante como un estimador de localización de las observaciones, puede extenderse en el caso robusto como la media recortada. Es necesario es necesario primero definir un proceso aleatorio basado en los estadísticos de orden por profundidad, en este sentido López, (2010)

Dados los estadísticos de orden para una muestra aleatoria de tamaño n $X[1], X[2], ..., X[n]$ se define:

$$\xi_n(t) = \begin{cases} X_{[i]} & si \quad \frac{i-1}{n} < t < \frac{i}{n} \\ X_{[1]} & si \quad t = 0 \end{cases}$$

Luego se define el PL-Estadístico:

$$PL_n = \int_0^1 \overline{\xi}_n(t) w(t) dt = \int_0^1 \xi_n(t) \overline{w}(t) dt$$

Donde $\overline{w}(t)$ es el peso medio dentro de cada clase de equivalencia, así mismo López (2010) afirma que el cálculo muestral de estadísticos de este tipo es trivial en caso de no haber empates entre puntos, ya que tras la ordenación de datos se promedian los $n(1-\alpha)$ puntos más profundos si dicha cantidad es entera y en caso contrario los $n(1-\alpha)+1$ puntos, donde denota la parte entera. Si hubiera empates y, por lo tanto, clases de equivalencia se asignara un peso $1/[n(1-\alpha)]$. (p.32)

No obstante para el calculo de estos estadísticos robustos se usará la función **CovLP** del la librería **Dephproc** el cual estima los parámetros de localización y dispersión mediante los estadísticos PL.

generándose:

```
    Robust Estimate of Location:
compactness   smoothness     symmetry
    0.09462      0.09462      0.18160

Robust Estimate of Covariance:
      [,1]       [,2]       [,3]
[1,]  0.0027247  0.0004795  0.0008489
[2,]  0.0004795  0.0001955  0.0002107
[3,]  0.0008489  0.0002107  0.0007392
```

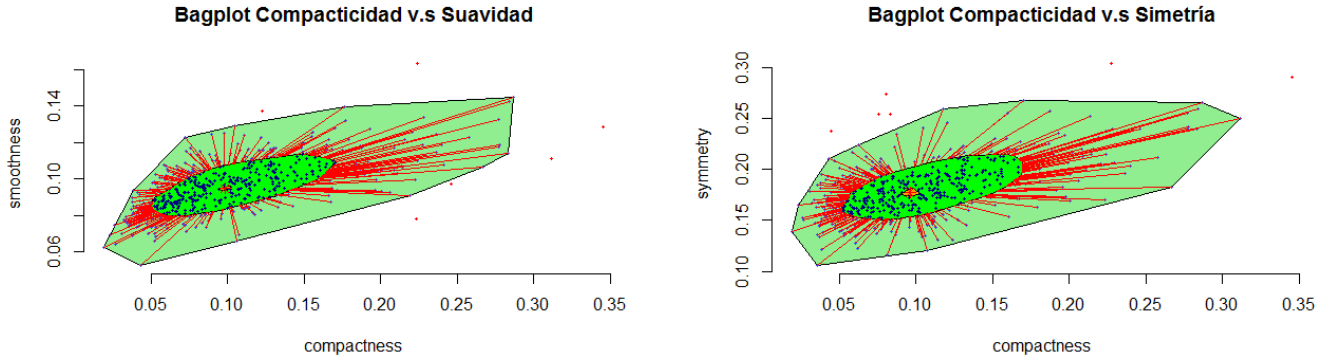
Contornos de profundidad y Bagplot

En lo que se refiere a la ordenación de observaciones por medio de la profundidad, López, A. (2010) establece que existen métodos gráficos que proporcionan la posibilidad de llevar a cabo análisis estadísticos exploratorios, dando lugar a diagramas de puntos y a curvas. No obstante el uso de representaciones gráficas sera el objetivo para caracterizar la distribución de los datos del diagnostico de cáncer; luego se da lugar al Bagplot el cual según Rousseeuw, Ruts, y Tukey (1999) es una generalización del boxplot para el caso bivariado y lleva consigo la ubicación de las observaciones a un punto central llevando al concepto de profundidad.

Para lograr un análisis de las observaciones se usará el uso de representaciones gráficas entre ellas el Bagplot (ver figuras 9 y 10) el cual es desarrollado por la librería **DepthProc** el bagplot permite visualizar medidas estadísticas bivariadas diagnostico de cáncer de mama en Wisconsin permitiendo, como lo destaca Rousseeuw et al., (1999) visualizar la ubicación, propagación, correlación, asimetría y colas de los datos (P.1). En este orden de ideas, se representa un bagplot por cada par de variables

- El la figura 9a se evidencia el bagplot para las variables Compacticidad y Suavidad, la región central (verde fluorescente) contiene el 50 % de los datos, que es el equivalente al boxplot en el caso univariado. Dentro de la región verde clara están los datos considerados intrínsecos a la nube de puntos.
- Para las variables Suavidad y asimetría se presenta el bagplot 9b el cual genera una forma asimétrica especificándose por la mediana de las observaciones

- En referencia a los bagplot presentado en la figura 9a el cual hace alusión a las variables Suavidad y simetría, se logra observar que es el que presenta mayor simetría con respecto al punto mas profundo, mostrándose la menor cantidad de valores atípicos.



(A) (B)

FIGURA 9: Bagplots por pares de variables

No obstante, cabe aclarar que fuera de la región del bagplot se ubican los datos outlier de la distribución bivariada conjunta (pares de variables), sin embargo, para determinar los outliers (en general) de la distribución multivariada conjunta ha de establecerse una comparación más general. Adicionalmente, cabe resaltar que el punto más profundo es un estimador de la mediana poblacional como se mencionó anteriormente este punto se resalta en el gráfico como punto rojo más central de la representación.

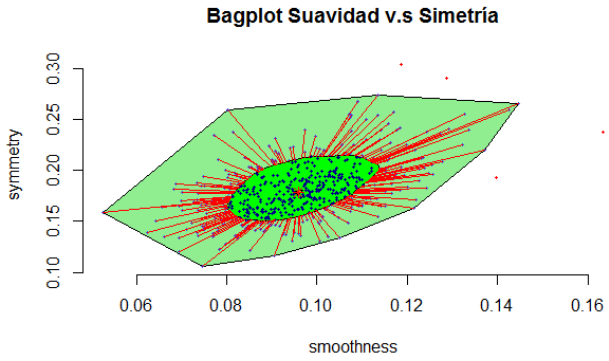


FIGURA 10: Bagplot

Para caracterizar la distribución multivariada se realizan la prueba de Mardia Skewness para identificar si la distribución multivariada es normal, usando la librería **MVN** se obtienen los resultados los cuales conducen a que las observaciones no proviene de una normal multivariada. Lo cual se esperaba puesto que las distribuciones univariadas no eran normales.

TABLA 6: Prueba de normalidad multivariada

	Test	Statistic	p value	Result
1	Mardia Skewness	235.568241179259	5.83465806202868e-45	NO
2	Mardia Kurtosis	12.9061675554888	0	NO

Contornos

El gráfico de contornos permitirá evaluar gráficamente la forma que tiene la distribución bivariante que representan las observaciones analizadas.

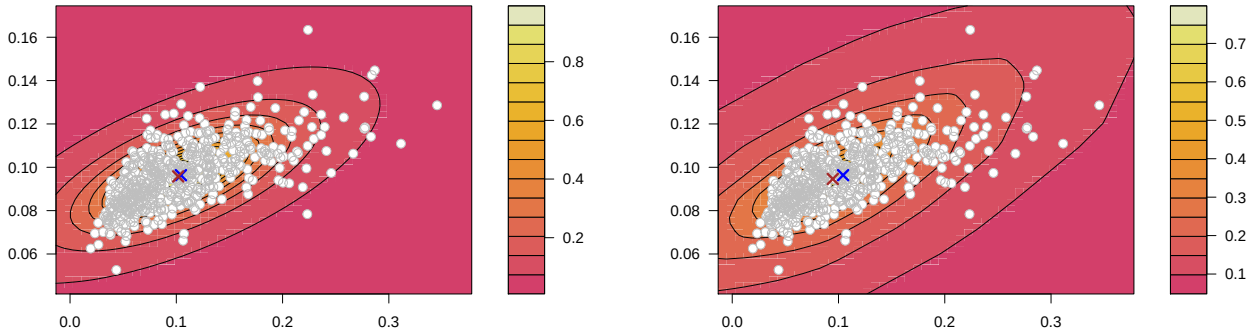
Un gráfico de contornos hace referencia a: $R(t) = \{x \in R^d : D(x) > t\}$ al cual se le denomina el **contorno de profundidad t** o conjunto de niveles; En este caso presentamos los gráficos de contorno para cada par de variables.

En este gráfico, la medida de centralidad adoptada sera adoptada según la medida de profundidad que se esté considerando; mientras cada curva al rededor denota cierta frontera de "centralidad" o profundidad en este caso (X roja sera el punto mas central, X representa la media).Se procede a visualizar los contornos para cada par de variables, haciendo uso de la profundidad de Mahalanobis y la Proyección.

Mientras la distancia al centro calculada en función de la profundidad Euclideana, arroja un gráfico con todas sus observaciones en el centro de la distribución (sugeriría que no hay datos atípicos), y además dichos contornos son circulares, inconvenientes para la estructura de los datos que ni siquiera parecen provenir de distribuciones normales; las gráficas de contornos que arrojan tanto el uso de la profundidad de Mahalanobis, como el de la Proyección, parecen hacer un mejor uso de la información disponible.

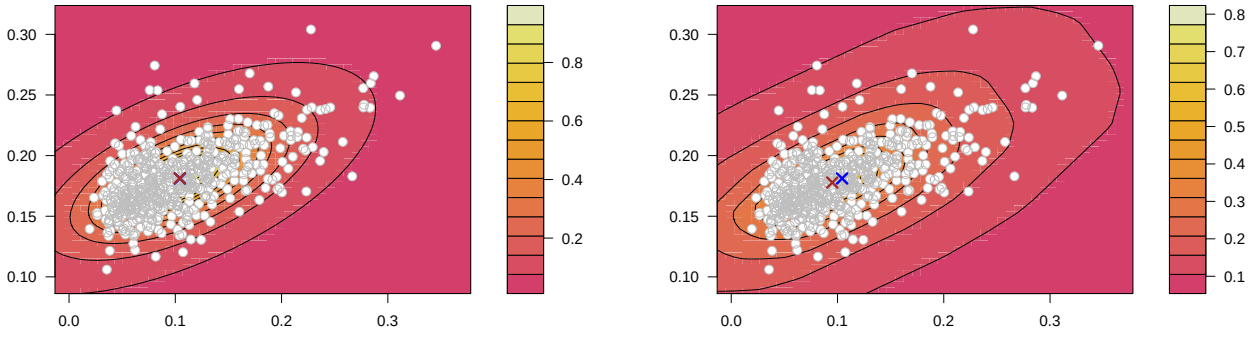
Las gráficas de contornos para las variables Compacidad y Suavidad 11, haciendo uso de las profundidades de Mahalanobis 11a y de Proyección 11b, manifiestan que la distribución multivariada parece no simétrica

Para Compacidad y Simetría también se debe considerar algún tipo de asimetría 12; mientras al observar los contornos para las variables Suavidad y simetría 13 las observaciones parecen çomportarse un poco mejorrespecto a las regiones planteadas por las dos profundidades, según estos criterios este par de variables parecieran ajustarse a una distribución elíptica.



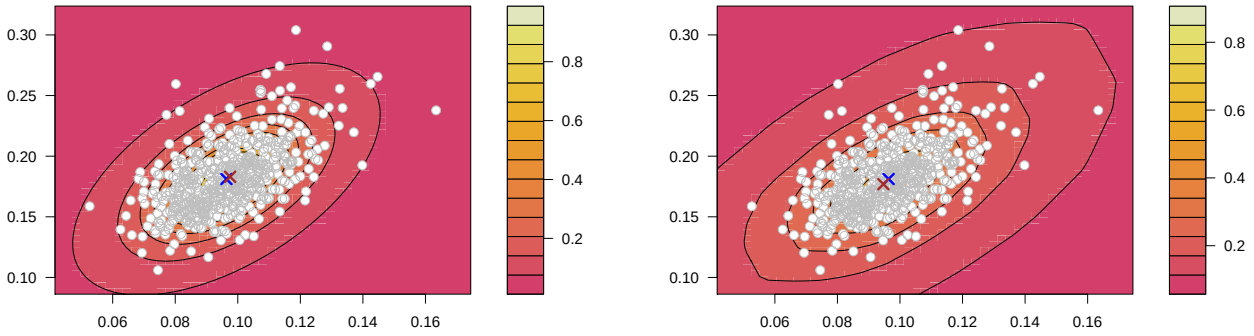
(A) Mahalanobis depth (B) Projection depth

FIGURA 11: Contornos Compacidad (eje x), Suavidad(eje y)



(A) Mahalanobis depth (B) Projection depth

FIGURA 12: Contornos Compacidad (eje x), Simetría (eje y)



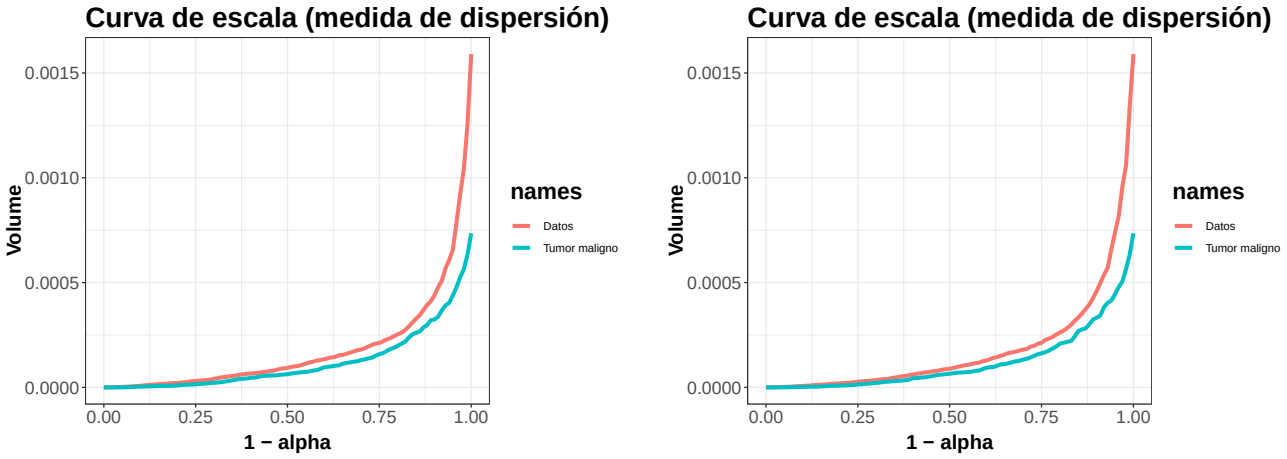
(A) Mahalanobis depth (B) Projection depth

FIGURA 13: Contornos Suavidad (eje x), Simetría (eje y)

En los gráficos anteriores de los contornos para cada par de variables, se observa claramente como los puntos mas cercanos al parámetro de centralidad son los que presentan profundidades mayores. En todas las gráficas se observa que hay variables que varían mas que otras dando la apariencia elíptica de los contornos, aunque la distancia de Mahalanobis muestra ciertos valores por fuera de estos contornos, los contornos usando la Profundidad de proyección se ajustan un poco mejor a los datos, permite evaluar la simetría de la distribución bivariada.

Curvas de escala

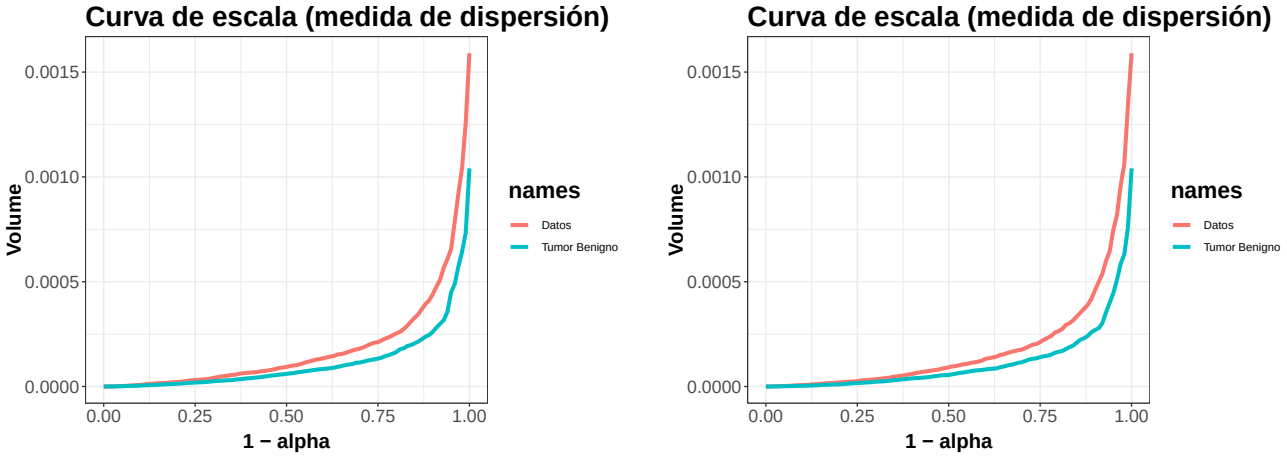
Las curvas de escala permiten observar la proporción del volumen de la nube de puntos que se encuentra contenido a una profundidad de $\alpha = [0, 1]$.



(A) Mahalanobis depth (B) Projection depth

FIGURA 14: Curvas de escala Datos vs Maligno

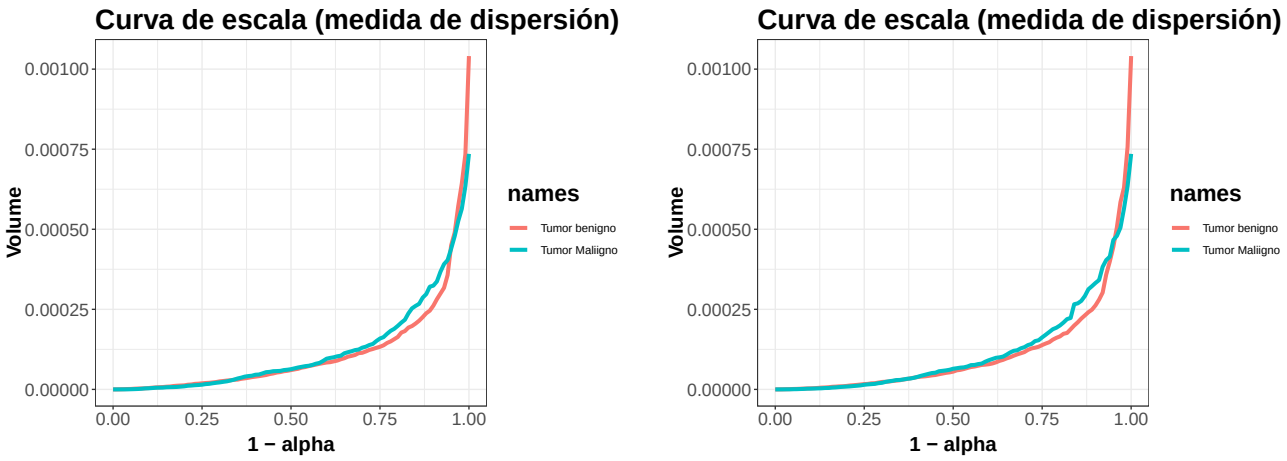
La primera gráfica 14 muestra las curvas de escala para las observaciones completas (curva roja) y las observaciones que presentaron tumores malignos (curva azul); se aprecia como entre α se hace mas grande la concentración de datos completos, con respecto a las observaciones con tumores malignos, donde las oservaciones que presentaron tumores malignos, parecen menos variables que las observaciones totales. Algo similar sucede en la segunda gráfica 15 donde esta vez la curva azul representara el volumen de los datos en α.



(A) Mahalanobis depth (B) Projection depth

FIGURA 15: Curvas de escala Datos vs Benigno

En la tercera gráfica se comparan las curvas de escala para las pacientes diagnosticadas con Cáncer (Azul) y las diagnosticadas sin Cáncer (Rojo). Las observaciones de pacientes diagnosticadas con cáncer parecen estar menos concentradas respecto al punto de mayor profundidad, en comparación con las observaciones de pacientes sanas, esto sugiere que la distribución de las pacientes que presentan cáncer respecto a las variables de estudio presenta una mayor variabilidad que las observaciones de pacientes sanas.



(A) Mahalanobis depth (B) Projection depth

FIGURA 16: Curvas de escala Benigno vs Maligno

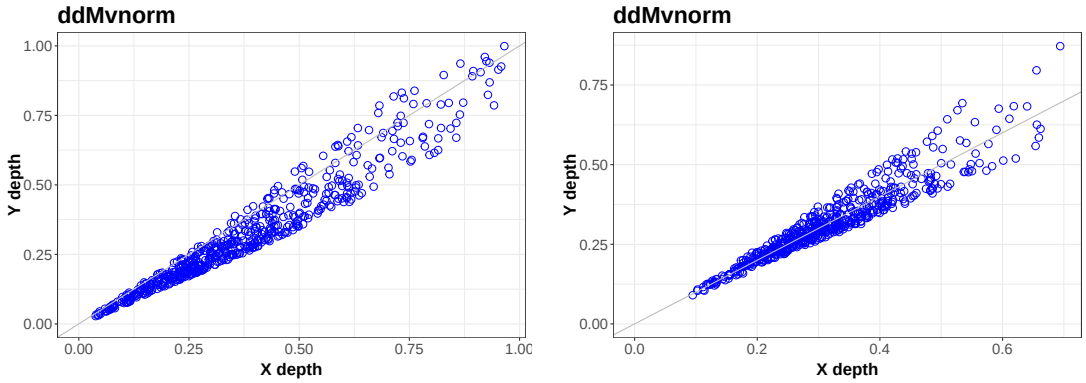
DD-Plot

El uso del DD-plot permite comparar dos muestras multivariadas mediante el uso del concepto de profundidad, y observar en dos dimensiones el comportamiento conjunto de los datos.

Mediante la aplicación del algoritmo DD-Plot al conjunto de datos respectivo a cáncer de mama, se busca evidenciar gráficamente las similitudes y diferencias que tienen las variables observadas en las pacientes diagnosticadas o no con cáncer de mama.

DD-plot (Datos vs Normal multivariada)

Los DD-Plot son un algoritmo que se vale del concepto de profundidad y su aplicación para brindar una representación gráfica en dos dimensiones, en especial son de gran ayuda para realizar comparaciones entre muestras, ya que con ellos se pueden diagnosticar cambios en la localización, en la escala y en la forma con sencillos diagramas bidimensionales.(López, 2010. p28). No obstante se procede a determinar si la distribución multivariada perteneciente a los datos de las mujeres diagnosticadas con cáncer es igual a la distribución multivariada de los datos provenientes de mujeres que no fueron diagnosticadas con cáncer.

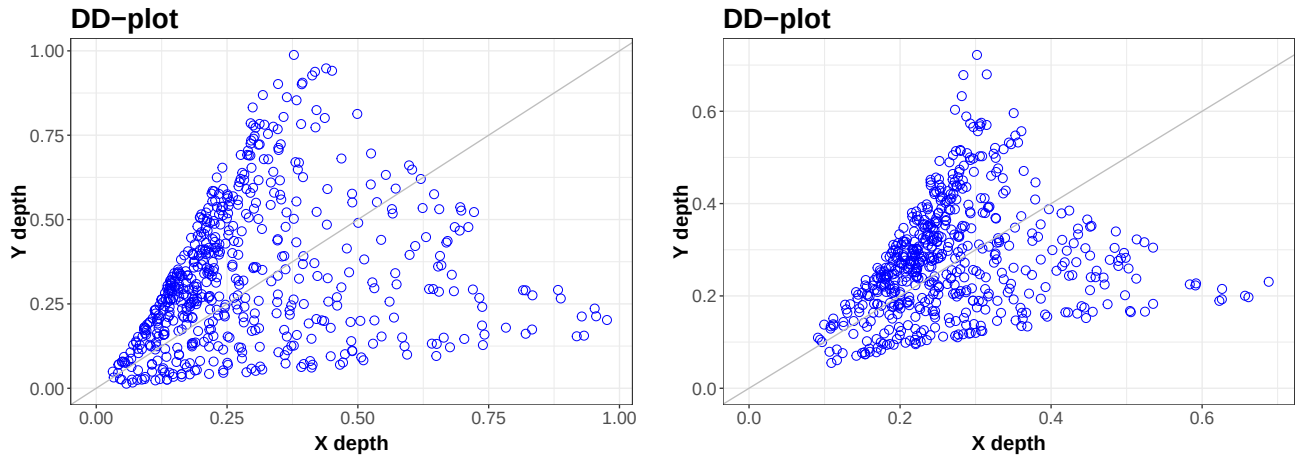


(A) Mahalanobis depth (B) Projection depth

FIGURA 17: DD-plot (Datos vs Normal multivariada)

En la primera figura 17 se aprecian las diferencias entre la matriz de datos referente a nuestro análisis y un conjunto de observaciones que provienen de una distribución normal multivariada.
Se puede concluir que la matriz de datos original parece provenir de una distribución no normal multivariada.

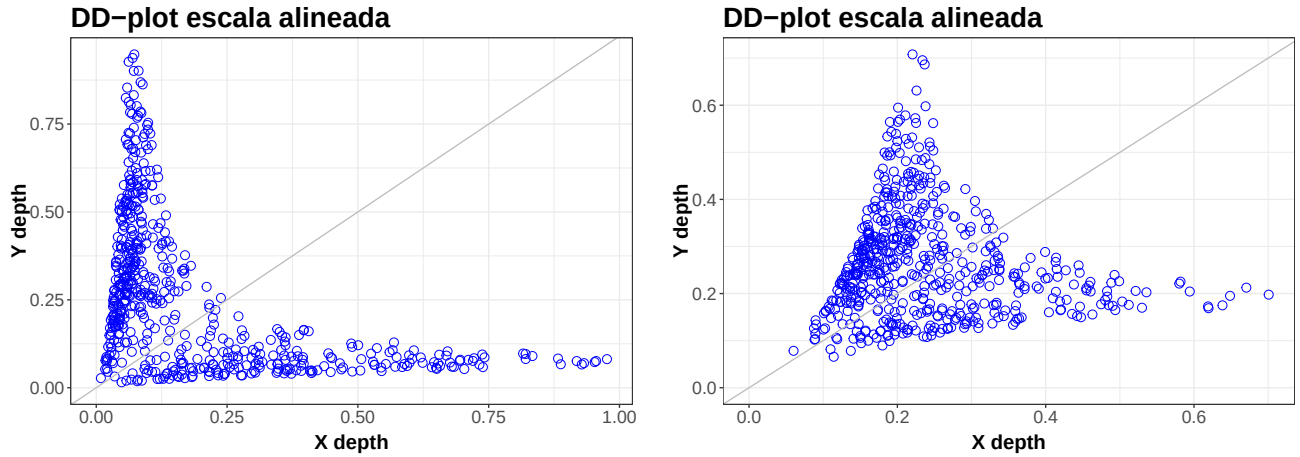
Posteriormente se hace uso del DD-plot para observar similitudes y diferencias entre las observaciones que tuvieron diagnósticos benignos como malignos.



(A) Mahalanobis depth (B) Projection depth

FIGURA 18: DD-plot (Maligno (eje x) vs Benigno (eje y)

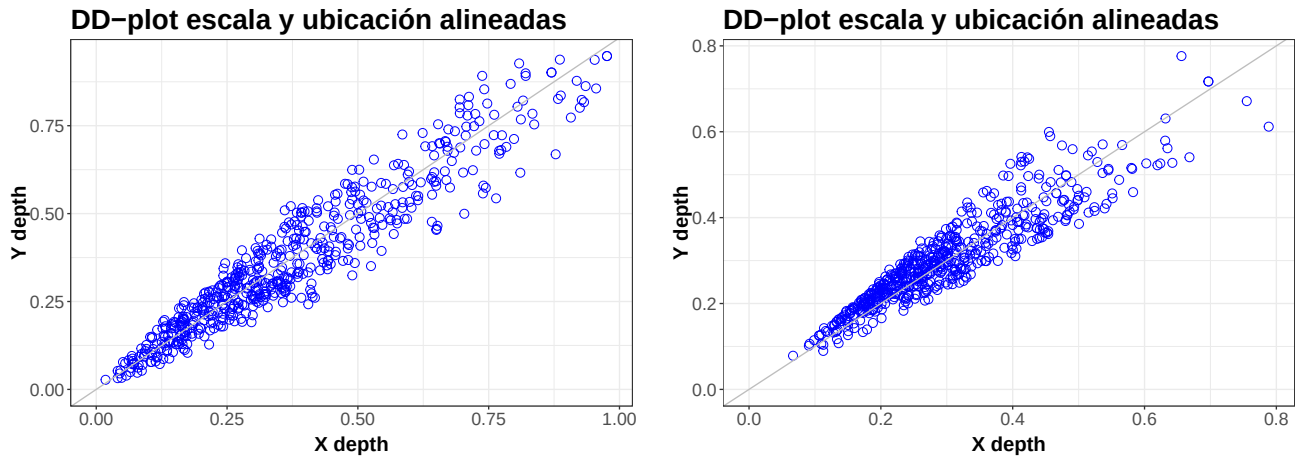
En el primer gráfico DD 18 para las profundidades de Mahalanobis y de Proyección, los datos parecen provenir de distribuciones con diferentes parámetros de centralidad.



(A) Mahalanobis depth (B) Projection depth

FIGURA 19: DD-plot (Maligno (eje x) vs Benigno (eje y)

Asumiendo que los parametros de escala de ambas distribuciones son los mismos, la evidencia de parametros de localización diferentes se hace mas evidente



(A) Mahalanobis depth (B) Projection depth

FIGURA 20: DD-plot (Maligno (eje x) vs Benigno (eje y)

En los gráficos anteriores se puede observar el comportamiento conjunto de los dos grupos de observaciones multivariadas respectivas al cáncer de mama; lo anterior observando las profundidades de cada uno de los puntos respecto a cada "nube de puntos.º distribución de la que provienen las muestras. Se observan las diferentes representaciones gráficas respecto a la profundidad, y en ellas se señala que las dos distribuciones presentan un cambio de ubicación central entre las distribuciones .

En el primer gráfico (Figura 17) se ve claramente dos grupos de puntos, cuya característica es que entre mas profunda es una observación respecto a una de las nubes de puntos esta se hace mucho menos profunda respecto a la otra nube de puntos. Esto indica que el parámetro de centralidad parece ser diferente entre las dos distribuciones o en este caso tipo de diagnóstico (Maligno y Benigno).

Detección de outliers

Con el objetivo de identificar datos atípicos del conjunto de datos multivariados, se procede a usar la libreria MVN (Multivariate Normality Tests) desarrollada por Korkmaz, S. Goksuluk, D. y Zararsiz, G. (2014) la cual es usada para realizar las diversas representaciones gráficas y pruebas para normalidad multivariada. cuya correlación es positiva

La siguiente tabla 7 muestra los datos atípicos calculados a partir de la distancia de Mahalanobis al cuadrado, con respecto a la función de distribución empírica; el calculo de la distancia es basado en el estimador MCD (Minimum covariance determinant) que es un estimador altamente robusto para los parámetros de ubicación y dispersión multivariados (Hubert y Debruyne 2010, p1). Al observar las tablas 3 y 21 se observan datos considerados atípicos que están presentes en ambas (ver filas resaltadas en tabla 21)

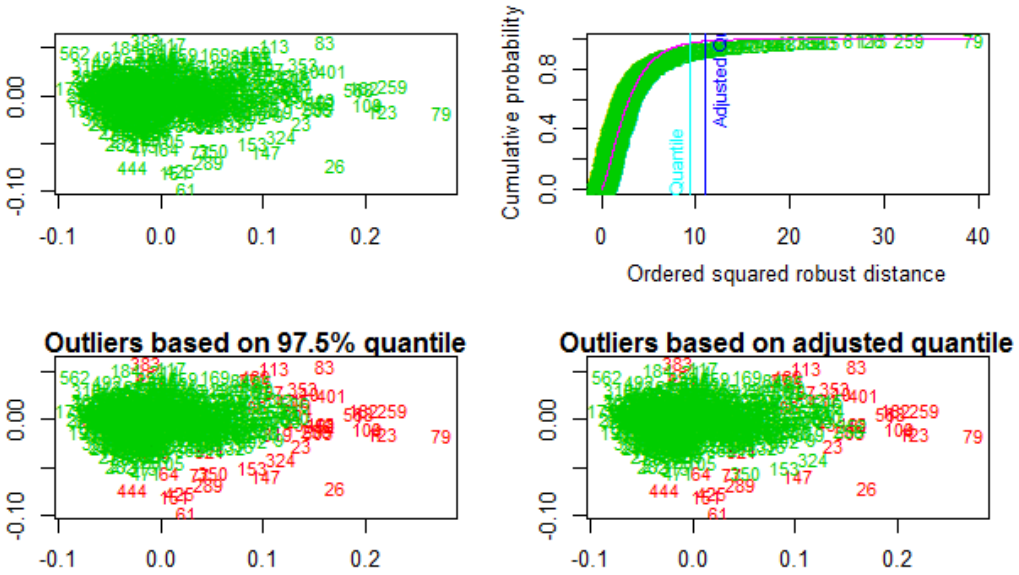


FIGURA 21: outliers basados en el estimador MCD

TABLA 7: Outliers desde el concepto profundidad

	compactness	smoothness	symmetry
1	0.28	0.12	0.24
4	**0.28**	**0.14**	**0.26**
10	0.24	0.12	0.20
13	**0.25**	**0.10**	**0.24**
23	0.21	0.11	0.25
26	**0.23**	**0.12**	**0.30**
43	0.22	0.09	0.23
61	**0.08**	**0.11**	**0.27**
64	0.09	0.08	0.23
77	**0.10**	**0.13**	**0.24**
79	**0.35**	**0.13**	**0.29**
83	**0.27**	**0.11**	**0.18**
106	0.18	0.14	0.19
109	0.28	0.13	0.26
113	**0.22**	**0.08**	**0.17**
123	**0.29**	**0.14**	**0.27**
147	0.17	0.11	0.27
151	**0.08**	**0.11**	**0.25**
182	**0.28**	**0.11**	**0.24**
191	0.24	0.11	0.24
257	0.21	0.09	0.19
259	**0.31**	**0.11**	**0.25**
273	0.20	0.09	0.17
289	**0.12**	**0.08**	**0.26**
353	0.24	0.11	0.20
383	0.11	0.07	0.12
401	0.26	0.12	0.21
425	**0.08**	**0.11**	**0.25**
431	0.22	0.10	0.20

Inferencias sobre el vector de medias

Según las gráficas de las distribuciones marginales de las variables (Compacticidad, simetría y Suavidad) se reflejaba la no normalidad (ver figura 4) junto con los qqplot (ver figuras 5) y verificando este supuesto bajo las pruebas de Shapiro-Wilk se afirmaba la no normalidad (ver tabla 2), este es un supuesto indispensable en la teoría clásica puesto que su uso sin este supuesto conllevaría inferencias no válidas, una forma de solucionar este problema es considerando transformaciones de los datos para de esta manera realizar análisis de la teoría normal. Según Johnson y Wichern (2002) las transformaciones no son más que una reexpresión de los datos en diferentes unidades.

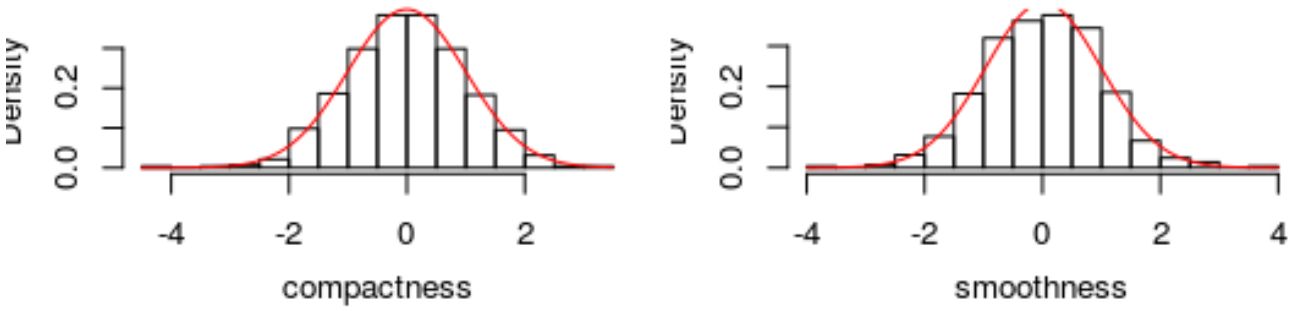
En muchos casos, la elección de una transformación para mejorar la aproximación a la normalidad no es obvio, no existe una única transformación a los datos algunas son mejores que otras, uno de las transformaciones es comunmente usadas es la de Box-Cox la cual tiene como objetivo homogeneizar la varianza, sin embargo en la mayoría de casos al usarse se logra corregir conjuntamente el problema de la no normalidad. (Melo et al., 2018. p256).

Una de las transformaciones usadas es la de la familia Johnson la cual transforma en normalidad usando la familia Z de distribuciones. la transformación de Johnson es basada en el método de los percentiles devolviendo la variable transformada, para tal fin se usa la función **RE.Johnson** de la libreria Johnson. (ver tabla 8).

Tabla 8: Datos alusivos a diagnósticos de cáncer de mama en Wisconsin, transformados

ID	**Diagnosis**	**Compactness**	**Smoothness**	**Symmetry**
1	M	2.37	1.49	1.91
2	M	-0.31	-0.79	0.10
3	M	1.06	0.96	0.99
4	M	2.44	2.81	2.29
5	M	0.69	0.35	0.09
6	M	1.18	2.03	1.04
⋮	⋮	⋮	⋮	⋮

En los barplot proporcionados en la figura 22, se logra identificar la forma de la distribución por variables después de realizar transformación a las observaciones,lográndose observar que las distribuciones son insesgadas y simétricas lo cual indica que al parecer no provienen de distribuciones normales.



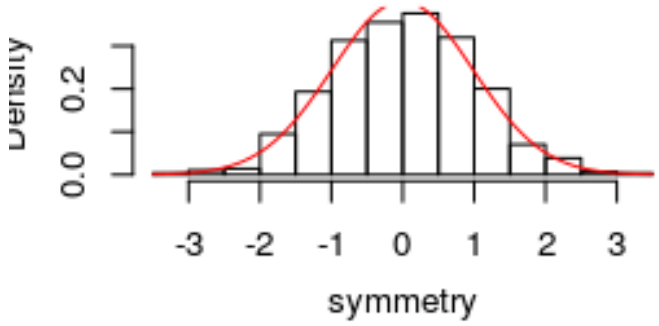


FIGURA 22: Bar plot de variables, mostrándose su distribución

Para determinar si en realidad no son normales se procede a realizar los respectivos QQplot lo cuales se presentan en la figura 23, donde se puede identificar que para todas las variables los datos se encuentran alrededor de la línea recta, lo cual da indicios del cumplimiento del supuesto de normalidad.

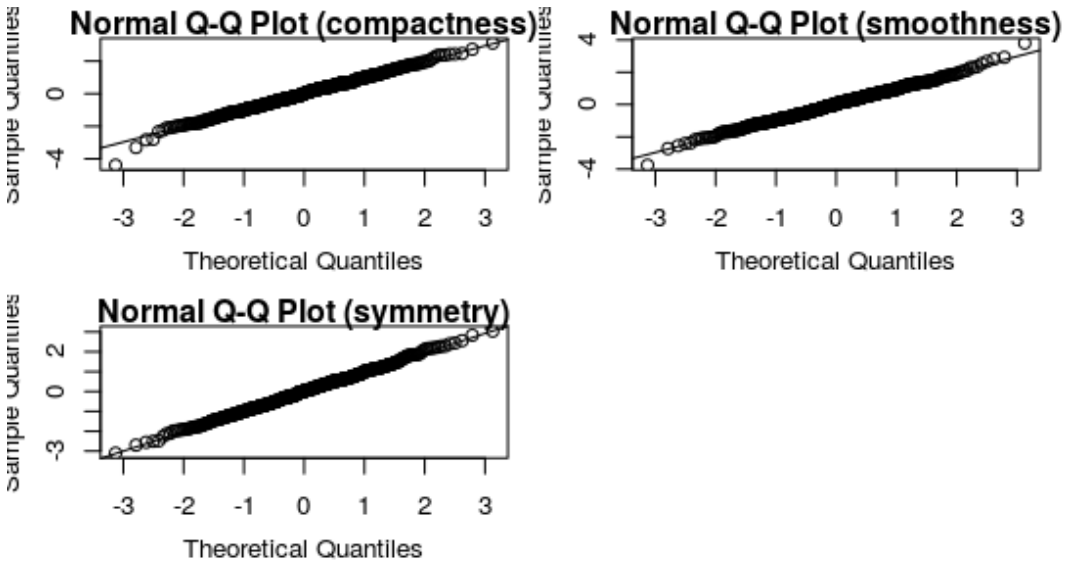


FIGURA 23: QQ plot por variables

Para confirmar este hecho se realizan las pruebas de Shapiro-Wilk (Ver Tabla 9) para contrastar la hipótesis:

$$H_0 : F_n(w) = N(\mu_0, \sigma_0^2) \qquad vs \qquad H_1 : F_n(w) \neq N(\mu_0, \sigma_0^2)$$

No hay evidencia estadistica para rechazar H_0 por lo cual se afirma el hecho de que las distribuciones son normales.

TABLA 9: Prueba de normalidad univariada para observaciones trasnformadas

	Test	Variable	Statistic	p value	Normality
1	Shapiro-Wilk	compactness	0.9960	0.1640	YES
2	Shapiro-Wilk	smoothness	0.9974	0.5249	YES
3	Shapiro-Wilk	symmetry	0.9988	0.9675	YES

Conociendo que la normalidad marginal no implica la normalidad multivariada, se genera una prueba ara identificar si las observaciones provienen de una distribucion multivarida. Para probar este hecho tomamos el test de Doornik - Hansen, Royston para probar la hipotesis:

H_0 : Las observaciones provienen de una distribución normal multivariada
vs
H_1 : Las observaciones NO provienen de una distribución normal multivariada

Tabla 10: Prueba de normalidad multivariada para observaciones trasnformadas

	Test	Statistic	df	p value	MVN
1	Doornik-Hansen	17.49	6.00	0.01	NO
2	Royston	4.83		0.18	YES

luego obserevando la tabla 10 para un nivel de significancia de $\alpha = 5\,\%$ se rechaza H_0 para la prueba Doornik-Hanzen, y no se rechaza H_0 para la prueba Royston, esto no certifica que se rechace H_0, sin embargo se considerará que provienen de una distribución 3-variada para hacer inferencias considerando que el número de observaciones es grande. Cabe aclarar que los resultados proporcionados serán contrastados con la estadistica MST.

No obstante, al considerar la normalidad de las observaciones, se tiene que: dados $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ...\mathbf{x_n}$ cada una con $\mathbf{c/X_i} \sim \mathbf{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, considerando la siguiente hipótesis entre el vector de medias:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} \qquad vs \qquad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$$

Se tiene la siguiente estadística de prueba:

$$T^2 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})' \Sigma^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})$$

Sin embargo, considerando que disponemos de la matriz de varianzas y covarianzas muestral, se tiene:

$$T^2 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})' \left(\frac{\boldsymbol{S}}{n}\right)^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})$$

$$T^2 = \sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})\sqrt{n} \quad \sim \quad T^2_{p,n-1}$$

Donde

$$T^2_{p,n-1} = \frac{(n-1)p}{n-p} F_{p,n-p}$$

Considerando lo anterior y conociendo que el conjunto de observaciones es grande puesto que n=569, se establece inferencias con respecto al vector de medias considerando el teorema del límite central (TCL):

Sean $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ...\mathbf{x_n}$ vectores aleatorios (una m.a) independiente igualmente distribuida: $\overline{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ considerando la siguiente hipótesis entre el vector de medias:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} \qquad vs \qquad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$$

$$T^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu_0})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu_0}) \leqslant c^2 \qquad ; \qquad c^2|_{H_0} \quad \sim \quad \chi^2_p$$

Intervalos de confianza

Considerando que $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ...\mathbf{x_n}$ son observaciones de una distribución 3-variada con vector de medias $\overline{\boldsymbol{x}}$ y matriz de covarianza definida positiva Σ, se tiene:

$$\boldsymbol{a}'\overline{\boldsymbol{x}} \pm \sqrt{\chi^2_p}\sqrt{\frac{\boldsymbol{a}' S \boldsymbol{a}}{n}}$$

Por consiguiente intervalos de confianza simultaneos están determinados por:

$$\overline{\boldsymbol{x}}' = \begin{bmatrix}\overline{x}_1 & , & \overline{x}_2 & , & \overline{x}_3\end{bmatrix} = \begin{bmatrix}0.10434 & , & 0.09636 & , & 0.18116\end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix}0.002784 & 0.000488 & 0.000870\\ 0.000488 & 0.000197 & 0.000214\\ 0.000870 & 0.000214 & 0.000750\end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix}1 & 0.659 & 0.602\\ 0.659 & 1 & 0.577\\ 0.602 & 0.577 & 1\end{bmatrix}$$

Intervalos de confianza del 95 % de confianza están determinados por: $\overline{x}_i \pm \sqrt{\chi^2_3(0.05)}\sqrt{\frac{s_{ii}}{n}}$ donde $\chi^2_3(0.05) = 7.81$, por lo tanto intervalos de confianza aproximados están dados por:

- $0.104 \pm \sqrt{7.81}\sqrt{\frac{0.002}{569}}$ contiene μ_1 o $0.08705418 \leqslant \mu_1 \leqslant 0.1216278$
- $0.096 \pm \sqrt{7.81}\sqrt{\frac{0.0001}{569}}$ contiene μ_2 o $0.09175678 \leqslant \mu_2 \leqslant 0.1009638$
- $0.181 \pm \sqrt{7.81}\sqrt{\frac{0.000750}{569}}$ contiene μ_3 o $0.1721886 \leqslant \mu_3 \leqslant 0.1901352$

Con respecto a lo anterior se puede afirmar que con un nivel de confianza del 95 % la media de la variable compacidad está entre 0.08705418 y 0.1216278, para la variable suavidad su media poblacional está entre 0.09175678 y 0.1009638, finalmente para la variable simetría se tiene que un intervalo de confianza del 95 % para μ_3 está entre 0.1721886 y 0.1901352.

Análisis de componentes principales

Considerando el conjunto de datos originales relativos al diagnóstico de cáncer de mama en el estado de Wisconsil el cual está compuesto por varias variables, se genera un análisis de componentes principales el cual se ocupa de explicar la estructura de varianza-covarianza del conjunto de variables a través de combinaciones lineales de estas variables. todo con el objetivo de reducir la dimensionalidad para una mayor comprensión visual y analítica, identificando nuevos ejes que son obtenidos a partir de una combinación lineal de las variables originales.

No obstante, es importante especificar que se requiere que las variables tengan una alta correlación que permita reflejar estas relaciones en los nuevos ejes, de esta manera lo primero que se debe realizar para obtener los nuevos ejes, es observar la correlación que existe entre las variables originales.

Teniendo en cuenta que las variables pueden llegar a tener diferentes escalas de medición, se podría presentar un problema ya que las diferentes unidades de medida pueden incrementar la varianza de alguna variable y esto se reflejará en el aporte a los nuevos ejes. Por consiguiente se quiere que la matriz de análisis sea una matriz escalada, de esta forma poder comparar las diferentes varianzas y observar el aporte de la información de los datos. A continuación se presenta la matriz de varianzas y covarianzas de los datos estandarizados, estos estarán dentro de la matriz Z:

	compactness	smoothness	symmetry	area
compactness	1.00	0.66	0.60	0.50
smoothness	0.66	1.00	0.56	0.18
symmetry	0.60	0.56	1.00	0.15
area	0.50	0.18	0.15	1.00

Tabla 11: Matriz de Correlaciones

En correspondencia a lo anterior, se puede identificar la existencia de alta correlación entre pares de variables como es el caso de compactness(compacidad) y smoothness(suavidad). Identificando los valores propios y sus respectivos vectores propios se tiene:

$$\lambda_1 = 2.3862324 \qquad e_1' = \begin{bmatrix}-0.5924231 & , -0.5285179 & , -0.5068592 & , -0.3358534\end{bmatrix}$$

$$\lambda_2 = 0.9366947 \qquad e_2' = \begin{bmatrix}0.1180532 & , -0.3185187 & , -0.3768539 & , 0.8617369\end{bmatrix}$$
$$\lambda_3 = 0.4439256 \qquad e_3' = \begin{bmatrix}-0.08120755 & , -0.66798855 & , 0.73476419 & , 0.08554663\end{bmatrix}$$
$$\lambda_4 = 0.2331472 \qquad e_4' = \begin{bmatrix}0.7927823 & , -0.4159399 & , -0.2473792 & , -0.3705318\end{bmatrix}$$

Ahora bien, considerando que $tra(\Sigma_X) = \sigma_{11}+\sigma_{22}+\sigma_{33}+\sigma_{44} = p = 4$ es decír $tra(\Sigma_X) = \lambda_1+\lambda_2+\lambda_3+\lambda_4 = p = 4$ y los valores propios mayores se usan para seleccionar el número de componentes a usar por medio del porcentaje de varianza explicado, este porcentaje de varianza explicado es determinado por:

(Proporción de varianza explicada por la i-esima componente) $=\frac{\lambda_i}{p}$ donde $i = 1, 2, 3, 4$ y p=4

Según los resultados anteriormente caracterizados, se procede a identificar el porcentaje de varianza explicada acumulada: $\frac{\lambda_1+\lambda_2}{p} = 83.07$, así mismo $\frac{\lambda_1+\lambda_2+\lambda_3}{p} = 94.17$, los resultados pueden observarse en la siguiente tabla:

Tabla 12: Valores propios, varianza y varianza acumulada.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.39	59.66	59.66
comp 2	0.94	23.42	83.07
comp 3	0.44	11.10	94.17
comp 4	0.23	5.83	100.00

A continuación se muestra la gráfica de los valores propios vs varianza, identificándose sus magnitudes. En este tipo de gráfico se puede elegir de manera subjetiva el número de componentes principales (ver figura 24):

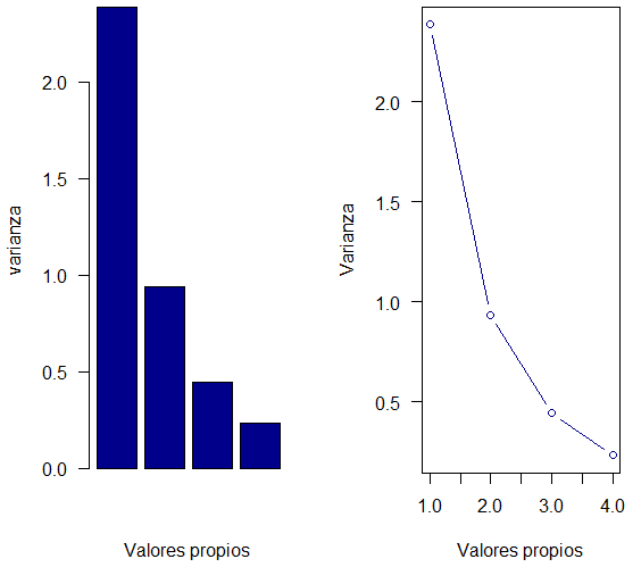


Figura 24

De acuerdo a lo mencionado anteriormente, se decide retener los dos primeros componentes principales, los cuales alcanzan a explicar el 83.07 % de la varianza total. Para construir los nuevos ejes que pertenecerán ahora a R^2, se debe tener en cuenta que las nuevas componentes son combinaciones lineales de las variables originales, esto es $Z_k = e_{k1}x_{i1} + e_{k2}x_{i2}$ con $i = 1, 2, ..., n$ y $k = 1, 2$

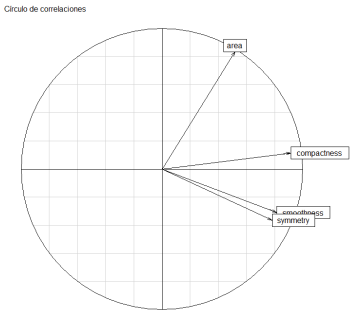


FIGURA 25

	Dim.1	Dim.2
compactness	0.92	0.11
smoothness	0.82	-0.31
symmetry	0.78	-0.36
area	0.52	0.83

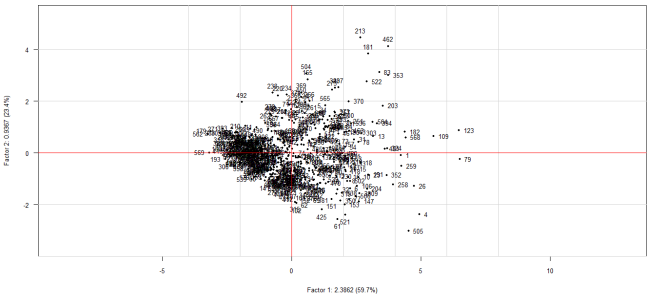


FIGURA 26: acp3

El círculo de correlaciones de la Figura 25. muestra que las variables Smothness (Suavidad) y Simmetry (Simetría), están altamente correlacionadas, como se había dicho anteriormente. Como todas las variables quedaron hacia un mismo lado del plano, es lo que se conoce como *factor tamaño*.

Se muestra el primer plano factorial de los individuos en la Figura 26. Superponiendo este gráfico con el círculo de correlaciones, Figura 25, se puede percibir que los registros que se encuentran en el primer cuadrante explican la variable Áárea. Mientras que los registros que se encuentran en el cuarto cuadrante explican las variables Smothness (Suavidad) y Simmetry (Simetría).

Análisis Factorial

A continuación se realiza un análisis factorial con el objetivo de establecer si las covarianzas o correlaciones observadas sobre un conjunto de variables pueden ser explicados en términos de un numero pequeño no observable de variables. Con este propósito, se define el modelo

$$X = \mu + \Lambda f + U$$

en el que X es un vector aleatorio de tamaño $(p \times 1)$ con media μ y matriz de covarianzas σ; Λ es una matriz de constantes (cargas) de tamaño $(p \times k)$, f es un vector columna de k componentes, $(k \leq p)$, y U un vector aleatorio de tamaño $(p \times 1)$. A los elementos f se les llama factores comunes y los elementos de U factores específicos.

Se asume que $E(f) = 0$; $cov(f) = I$; $E(U) = 0$; $cov(U) = E(UU') = \Psi$ y $cov(f, U) = 0$.

La Tabla 13 y la Tabla 14, muestran, respectivamente, los vectores propios de la matriz de correlaciones de los datos y los valores propios correspondientes, así como el porcentaje de varianza acumulada.

Tabla 13: Vectores propios asociados a la matriz de correlaciones

	1	2	3	4
1	-0.59	0.12	-0.08	0.79
2	-0.53	-0.32	-0.67	-0.41
3	-0.51	-0.38	0.73	-0.25
4	-0.34	0.86	0.09	-0.37

Tabla 14

Comp	Autovalor	Porc.Var	Acum.Porc.Var
1	2.39	59.66	59.66
2	0.94	23.43	83.09
3	0.44	11.09	94.18
4	0.23	5.82	100.00

Se decide hacer el análisis con los dos primeros factores, debido al porcentaje de varianza explicada que se alcanza con estos, cerca del 84 %, utilizando el método de componentes principales.

La Figura 27. muestra la relación de las variables con los ejes coordenados. Se puede pensar que las variables *Compactness*. *Smothness* y *Simmetry* conforman un factor, mientras que *Área* representaría otro factor. Esto también se puede ver en la matriz de correlaciones de los datos; las tres primeras variables mencionadas tienen alta correlación.

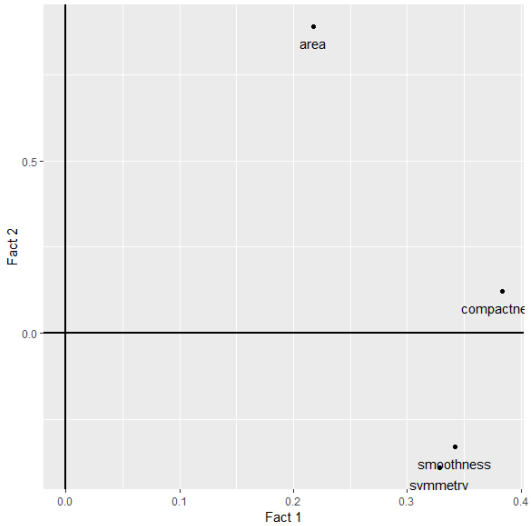


Figura 27

La Figura 28 muestra la nube de individuos clasificados según el tipo de diagnóstico B o M.

FIGURA 28

La Tabla 15 se muestra los factores **f1** y **f2** así como las comunalidades $h_i = f_{i1}^2 + f_{i2}^2$ y la varianza específica $\psi_i = 1 - h_i$

TABLA 15: Tabla relativa a la matriz de carga y comunalidad

Variable	f1	f2	Comunalidad	Var. Específica
1	-0.92	0.11	0.85	0.15
2	-0.82	-0.31	0.76	0.24
3	-0.78	-0.37	0.75	0.25
4	-0.52	0.83	0.96	0.04

Se puede ver que las primeras tres variables tienen un alto peso en la conformación del primer eje, mientras que la cuarta variable sobre el segundo factor. Teniendo en cuenta esto, se decide que no es necesario hacer ningún tipo de rotación, puesto que se conformaron grupos de variables que explican bien cada factor.

Métodos de clasificación

Clasificar lleva consigo el asignar un elemento a una característica distintiva particular generando agrupamientos diferenciados entre si, este método de clasificación es conocido como supervisada encontrándose el análisis discriminante o clasificación automática. No obstante otro método de clasificar es el no supervisado el cual conlleva a organizar grupos de elementos de acuerdo a características distintivas donde cada uno de sus individuos son homogéneos en su interior y diferentes entre grupos, este tipo de método de clasificación se le conoce como no supervisado o análisis cluster.

Considerando que el agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos es lo que se conoce como análisis cluster, ha de tenerse en cuenta que los grupos son desconocidos a priori y es necesario construirlos considerando las observaciones en conjunto, esto da cabida a lo conocido como:

- Métodos Jerárquicos: La totalidad de individuos no se agrupan en clusters de una sola vez, sino que se van haciendo particiones sucesivas a "distintos niveles de agregación o agrupamiento ". Inicialmente cada individuo es un grupo, generándose la menor distancia euclidiana entre individuos y clusters por medio de medidas de similaridad entre cluster llamados Linkages, sucesivamente se van formando grupos de mayor tamaño fusionando grupos cercanos.

  Cabe mencionar que el hecho de agrupar individuos y clusters por medio de una medida de mimilaridad (Linkage) implica el anidamiento entre grupos e individuos.

- Métodos no Jerárquicos: Una característica distintiva en este método es que se establece primeramente un número de grupos a priori y los individuos se van clasificando a uno de esos grupos.

Análsis cluster

Para iniciar el procedimiento se establece una medida de disimilaridad entre 2 observaciones, frecuentemente se emplea la distancia Euclidiana. El algoritmo inicia desde la parte inferior del dendrograma, donde cada observación es tratada como su propio cluster. En consecuencia se asocian los dos clusters que son más similares entre si para de esta manera ser fusionados, asi sucesivamente hasta que no se tenga más asociaciones es decir que se tenga solo un cluster, En la gráfica 29 se puede observar la gráfica del dendrograma para el total de observaciones.

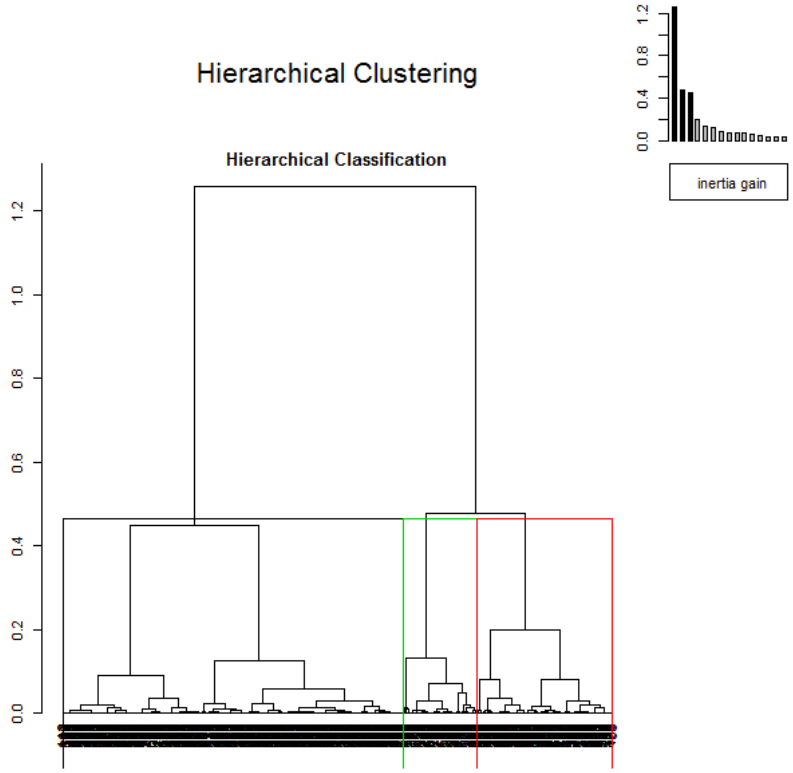


FIGURA 29: Dendrograma para los diagnósticos de cáncer de mama en el estado de Wisconsil

Es importante aclarar que la altura del corte del dendrograma controla el número de clusters a usarse, no obstante el corte para la selección de cluster se realiza por medio del estadístico r^2 o RS, el cual está definido como:

$$RS = 1 - \frac{\sum_{g=1}^{K} \sum_{i \in C_g} \left\| \boldsymbol{X_i} - \overline{\boldsymbol{X_g}} \right\|^2}{\sum_{i=1}^{n} \left\| \boldsymbol{X_i} - \overline{\boldsymbol{X}} \right\|^2}$$

RS (R cuadrado) es la razón entre la heterogeneidad entre los distintos clusters y la heterogeneidad total. Dado que la heterogeneidad total se puede descomponer como suma de la heterogeneidad entre los conglomerados y de la heterogeneidad dentro de los conglomerados.

No obstante una forma de identificar el corte del dendrograma para fijar el número de cluster a usar es 3 puesto que RS obtenido es 0.7.

Algoritmo k-means

En la sección anteriro se generaban agrupamientos por medio de observaciones de un conjunto de datos cuyo objetivo era particionar en diferentes grupos de tal manera que las observaciones dentro de cada grupo son muy similares, mientras que las observaciones en diferentes grupos son muy diferentes.

Ahora previamente conociendo el número de grupos "k", el algoritmo k-means realiza una asignación de los individuos a los grupos ya preestablecidos por medio de una aleatorización, considerando que la importancia recae en generar grupos ha de usarse una medida de heterogeneidad entre grupos (varianza) donde el objetivo es minimizar las varianzas de todas las variables en los grupos, así se obtendrán grupos más homogéneos.

El criterio de homogeneidad usado es minimizar la suma de cuadrados dentro de los grupos (SCDG), es decir minimizar la traza de $\boldsymbol{W}$:

$$\boldsymbol{W} = \sum_{g=1}^{G} \sum_{n_i}^{i=1} (\boldsymbol{X_{ig}} - \overline{\boldsymbol{X_g}})(\boldsymbol{X_{ig}} - \overline{\boldsymbol{X_g}})'$$

Los pasos para generados por el algoritmo de k-means es asignar aleatoriamente cada observación a uno de los K grupos. Por consiguiente computar el centroide o vector de medias en cada grupo, calcular las distancias Euclideas de cada elemento de cada grupo al centroide más cercano (regla de desición: asignar elemento al grupo cuya distancia al centroide es más cercana), el siguiente paso es calcular la traza en cada asignación de los individuos a los grupos esto hace que el proceso secuencialmente calcule el centroide de cada grupo y su correspondiente varianza entre grupos. Al encontrase un agrupamiento donde no se reduce la varianza entre grupos el algoritmo k-means se detiene.

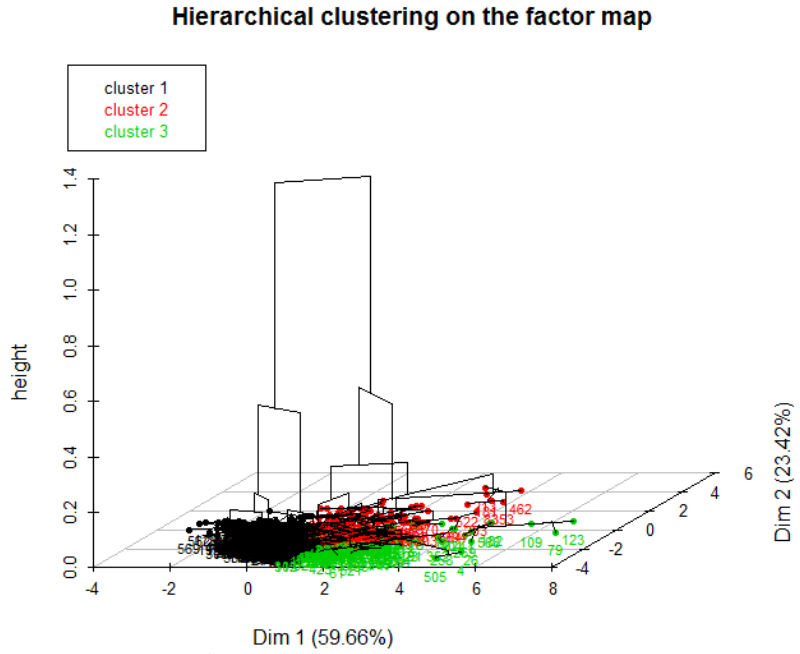


Figura 30: Agrupamiento generado ppor el algoritmo k-means

En la figura 30 se observa la clasificación realizada por el algoritmo k-means generándose 3 grupos homogéneos de tal manera que la varianza entre grupos es mínima. En la tabla 16 se muestra los vectores que representan los centroides de cada grupo.

Tabla 16: Vectores que representan el centroide de los grupos

	compactness	smoothness	symmetry
1	0.21	0.11	0.22
2	0.13	0.10	0.19
3	0.07	0.09	0.17
SSDC	0.15798	0.21018	0.25832

En relación al algoritmo se estableció la suma de cuadrados dentro de los grupos y esta esta descrita en la tabla 17 e muestra la suma de cuadrados dentro de los grupos:

Tabla 17: Suma de cuadrados dentro de los grupos

	Grupo 1	Grupo 2	Grupo 3
SSDC	0.15798	0.21018	0.25832

Según el algoritmo, se realizo el agrupamiento minimizando la varianza entre grupos, en este orden de ideas se presentan los 9 individuos y su clasificación (GRUPO1: 1 ; GRUPO2: 2 y GRUPO3:3).

Individuos	Clasificación
1	1
2	3
3	2
4	1
5	2
6	1
⋮	⋮

El k-ésimo vecino mas cercano

La clasificación vía el k-ésimo vecino mas cercano implementada mediante la función knn del paquete FNN (Fast Nearest Neighbor Search Algorithms and Applications, Alina Beygelzimer, Sham Kakadet and John Langford) requiere de una muestra de entrenamiento y otra de prueba (particiones de la muestra original donde cada individuo tien asignada una categoria) la primera sirve de soporte para generar los grupos ya que se asignan elementos cuyo grupo ya es conocido, mientras que la prueba test es a la que se le asocia su pertenencia en algún grupo vía el vecino mas cercano.

Se dividió la base original en individuos con tumor benigno y maligno respectivamente, luego se seleccionan 250 y 150 individuos de cada grupo respectivamente que conformaran la muestra de entrenamiento, los individuos restantes serán clasificados via knn.

En la Figura 31. se muestra el nivel de exactitud alcanzado para diferentes valores de k. Se puede ver que con 13, 14 0 15, se alcanza el máximo nivel, con cerca de un 93 %

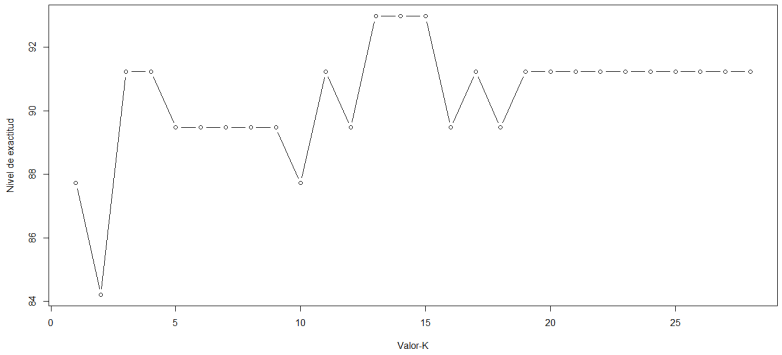


Figura 31: Gráfico de Exactitud-Algoritmo KNN

Teniendo en cuenta lo anterior, se toma un $k = 13$ y se obtiene que 85 % de los diagnósticos "Benignoïdentificados en la muestra de prueba fueron clasificados correctamente vía método del k-ésimo vecino mas cercano.
El 16 % de diagnósticos "Malignoïdentificados en la muestra de prueba fueron clasificados correctamente vía método del k-ésimo vecino mas cercano.

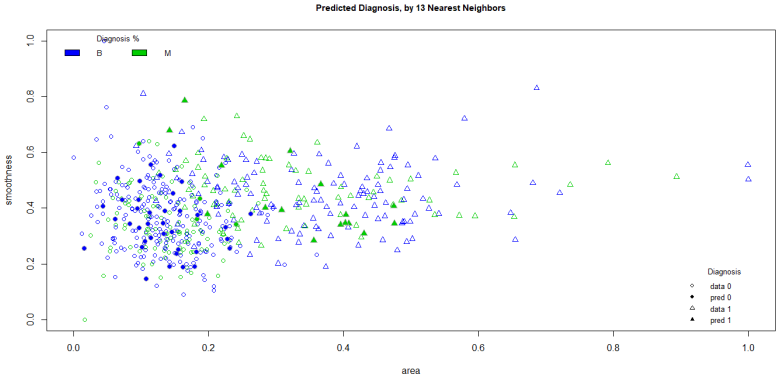


FIGURA 32: Clasificación

La Figura 32 muestra una gráfica de los datos con el conjunto de entrenamiento en formas huecas y los nuevos elementos representados por las figuras rellenadas, en la que se utilizaron las variables **Área** y **Suavidad**(Smoothness).

Manova

Teniendo en cuenta que los datos con los que se está trabajando provienen de dos poblaciones diferentes, a las cuales se les midió cuatro variables, para ilustrar el uso del análisis de varianza multivariada (**MANOVA**), a una vía, se desea probar la hipótesis nula de si la media 4-dimensional para las poblaciones según sea el diagnóstico, **Maligno** o **Benigno** son iguales.

Para hacer este tipo de análisis se asume que:

- X_{l1}, X_{l2}, ... , X_{ln} es una muestra aleatoria de tamaño n_l de una población con media μ_l, $l = 1, 2, \ldots, g$. Las variables aleatorias de poblaciones diferentes son independientes
- Todas las poblaciones tienen matriz de covarianza común Σ
- Cada población tiene una distribución normal.

Teniendo en cuenta que se tiene un tamaño de muestra grande, la tercera condición puede relajarse y utilizar el teorema de límite central. Por otra parte, en cada población, las variables no distribuyen individualmente normal, por tanto la prueba se hace con los variables transformadas

Los vectores de medias para las dos poblaciones son:

$$\mu_B = [0.08, 0.09, 0.17, 462.79]$$
$$\mu_M = [0.14, 0.10, 0.19, 978.38]$$

El modelo a una vía del **MANOVA** es el siguiente:

$$X_{lj} = \mu + \tau_l + e_{lj}$$

Con los errores $e_{ij} \sim N(0, \sigma^2)$ e independientes.
Para estimar los parámetros del modelo, las observaciones se descomponen de la siguiente manera:

$$x_{lj} = \bar{x} + (\bar{x_l}) - \bar{x} + (x_{lj} - \bar{x_l})$$

Cada término de la igualdad anterior, se corresponde con los términos de

$$x_{lj} = \hat{\mu} + \hat{\tau_l} + \hat{e}_{ij}$$

Para probar la hipótesis sobre el efecto de los tratamientos, se calculan las matrices de suma de cuadrados y productos cruzados, las cuales se muestran a continuación para los tratamientos y para los errores, **B** y **W**, respectivamente.

	compactness	smoothness	symmetry	area
compactness	0.56	0.09	0.16	4464.73
smoothness	0.09	0.01	0.03	714.65
symmetry	0.16	0.03	0.05	1284.01
area	4464.73	714.65	1284.01	35358547.15

Tabla 18

	compactness	smoothness	symmetry	area
compactness	1.02	0.19	0.33	797.75
smoothness	0.19	0.10	0.10	-216.99
symmetry	0.33	0.10	0.38	-454.96
area	797.75	-216.99	-454.96	34984591.70

Tabla 19

Se cálcula la Lambda de Wilks, Λ^* que es igual a

$$\mathbf{\Lambda^*} = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{\left|\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)'\right|}{\left|\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\bar{\mathbf{x}})(\mathbf{x}_{\ell j}-\bar{\mathbf{x}})'\right|}$$

Finalmente para probar la hipótesis sobre el efecto de los tratamientos

$$H_0 = \tau_1 = \tau_2 = 0$$

Se establece el siguiente estadístico de prueba, teniendo en cuenta el número de grupos, $l = 2$ y el número de observaciones.

$$p \geq 1 \qquad g = 2 \qquad \left(\frac{\Sigma n_\ell - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,\Sigma n_\ell - p - 1}$$

A continuación se muestran diferentes salidas del paquete estadístico *R*, de los cuales se puede concluir que se rechaza la hipótesis nula, pues se obtuvieron valores pequeños para cada uno de estos.

```
    > summary(Di_m, test = "Wilks")
          Df   Wilks approx F num Df den Df
diagnosis   1 0.41142   201.72      4    564
Residuals 567
             Pr(>F)
diagnosis < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(Di_m, test = "Hotelling-Lawley")
           Df Hotelling-Lawley approx F num Df
diagnosis   1           1.4306   201.72      4
Residuals 567
          den Df    Pr(>F)
diagnosis    564 < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(Di_m, test = "Pillai")
           Df  Pillai approx F num Df den Df
diagnosis   1 0.58858   201.72      4    564
Residuals 567
             Pr(>F)
diagnosis < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(Di_m, test = "Roy")
           Df    Roy approx F num Df den Df
diagnosis   1 1.4306   201.72      4    564
Residuals 567
             Pr(>F)
diagnosis < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Por ejemplo, utilizando el test de Wilks, se obtiene un valor $\Lambda^* = 0.41142$ y un valor p de $2.2e-16$.

Prueba de rachas de Wald-Wolfowitz Generalizada

La prueba siguiente es la generalización de la prueba de rachas de Wald-Wolfowitz basada en el minimum spanning tree (MST) propuesto por Friedman y Rafsky (1979). La aplicación de este método permite observar si un par de muestras multivariadas siguen la misma distribución de probabilidad; la distribución nula del estadístico de prueba es estimada permutando los labels muestrales y calculando el estadístico de prueba para cada uno, el estadístico de prueba es basado en el número de vértices eliminados en el MST de la muestra total que pertenecen a nodos provenientes de grupos distintos, donde la medida de distancia entre un par de individuos es generalizada via un minimal spanning tree.

Para llevar a cabo la implementación de este método se hace uso de el paquete GSAR (Gene Set Analysis in R) de R mediante su función WWtest. El estadístico de prueba en este caso es W, donde:

$$W = \frac{R - E(R)}{Var(R)}^{1/2}$$

R hace referencia a las suma de los vértices que unen a nodos que representan individuos de grupos diferentes mas 1.

El valor observado del estadístico $W = -14.89$, y el p-valor correspondiente es de 0.0009; luego se concluye que los dos grupos comparados correspondientes a diagnósticos maligno y benigno presentan distribuciones diferentes.

Referencias

- Chebana, F., & Ouarda, T. B. (2011). Depth-based multivariate descriptive statistics with hydrological applications. Journal of Geophysical Research: Atmospheres, 116(D10).
- Cruz, N. y Martinez de Larios, N. (2002). La biopsia por aspiración con aguja fina en glándula mamaria: diagnostico citológico y concordancia histológica y clínica. Rev Hospital General M Gea Gonzáles, 5, 79-84.
- Dua, D. y Graff, C. (2019). Repositorio de aprendizaje automático de la UCI. Irvine, CA: Escuela de Información y Ciencias de la Computación. Universidad de California. disponible en: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29`
- Hubert, M., y Debruyne, M. (2010). Minimum covariance determinant. Wiley interdisciplinary reviews: Computational statistics, 2(1), 36-43.
- Johnson, R, y Wichern, D. (2002). *".Applied multivariate statistical analysis"* (Vol. 5, No. 8). Upper Saddle River, NJ: Prentice hall.
- Korkmaz, S. Goksuluk, D. y Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. The R Journal, 6(2), 151-162. `https://journal.r-project.org/archive/2014/RJ-2014-031/RJ-2014-031.pdf`
- Liu, R., Parelius, J y Singh, K. (1999). 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.
- López, A. (2010). Similaridad y contraste mediante profundidad estadística (Doctoral dissertation, Tesis doctoral).`https://core.ac.uk/download/pdf/29401738.pdf`
- Mathematical Programming in Machine (2001). Grupo de optimización del Departamento de Ciencias informáticas de la Universidad de Wisconsin-Madison. datos Wisconsin Breast Cancer DatabaseDisponible en: `http://pages.cs.wisc.edu/~olvi/uwmp/mpml.html`
- Melo, O., López, L y Melo, S (2018). "Diseño de Experimentos-Métodos y Aplicaciones". Universidad Nacional de Colombia.
- Pokotylo, O., Mozharovskyi, P., Dyckerhoff, R. (2016). Depth and depth-based classification with R-package ddalpha. `arXiv:1608.04109`
- Rousseeuw, P. J., Ruts, I., y Tukey, J. W. (1999). The bagplot: a bivariate boxplot. The American Statistician, 53(4), 382-387.
- Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. Journal of statistical Planning and Inference, 123(2), 259-278.
- Zuo, Y., y Serfling, R. (2000). General notions of statistical depth function. Annals of statistics, 461-482.