Análisis multivariante - PEC 1

Descripción de datos multivariantes

Juan José Rodríguez Aldavero

May 7, 2020

Contents

1 Descripción del conjunto de datos

1.a Análisis de medidas de centralización

Comenzamos describiendo el conjunto de datos a partir de una serie de medidas de centralización. El vector de medias viene descrito por la expresión

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

mientras que el vector de medianas se corresponde con el percentil 50 de las distintas variables del conjunto de datos. Los calculamos a partir de los siguientes comandos en R:

```
apply(AM20, 2, mean) #Medias
apply(AM20, 2, median) #Medianas
apply(AM20, 2, quantile) #Cuantiles
```

En la siguiente tabla se pueden ver los vectores de medias, medianas y percentiles del conjunto de datos.

($\times 10^5$)	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
Media	1,960	0,562	0,032	0,075	0,452	0,026	0,397	0,052	0,014	0,401	0,030	0,395	0,384
Mediana	1,803	0,584	0,032	0,071	0,433	0,023	0,412	0,016	0,004	0,390	0,030	0,380	0,360
Percentil 25	1,272	0,359	0,021	0,034	0,282	0,012	0,226	0,005	0,001	0,240	0,018	0,240	0,226
Percentil 50	1,803	0,584	0,032	0,071	0,433	0,023	0,412	0,016	0,004	0,390	0,030	0,380	0,360
Percentil 75	2,410	0,757	0,042	0,112	0,599	0,037	0,562	0,054	0,011	0,550	0,042	0,540	0,530
Percentil 100	6,890	0,999	0,071	0,213	1,167	0,115	0,835	1,059	0,494	0,976	0,079	0,976	0,976

Vemos como la variable A presenta unos valores en media muy superiores al resto, mientras que las variables $\{B, E, G, H, I, J, L, M\}$ presentan unos valores intermedios y las variables $\{C, D, F, K, \}$ presentan unos valores más pequeños que el resto. Esta variabilidad nos hace pensar que el conjunto de datos no está normalizado, ya que no está en una escala común.

1.b Análisis de medidas de variabilidad

Continuamos viendo un conjunto de medidas de variabilidad para los datos. En primer lugar, conviene analizar las variables del conjunto de datos de forma individual para conocer su comportamiento y, posteriormente, estudiar medidas de variabilidad multivariantes. Comenzamos describiendo los vectores de desviaciones típicas, varianzas y coeficientes de variación del conjunto de datos. Estos vienen descritos por:

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \qquad \text{desviación típica}$$

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad \text{varianza}$$

$$CV_j = \frac{s_j}{\bar{x}_j} \qquad \text{coeficiente de variación}$$

Los calculamos mediante los comandos:

```
apply(AM20, 2, sd) #Desviaciones tipicas
apply(AM20, 2, var) #Varianzas
apply(AM20, 2, sd)/apply(AM20, 2, mean) #Coeficientes de variacion
```

Para el conjunto de datos se obtienen los valores:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
Desv. típica ($\times 10^5$)	1,013	0,247	0,014	0,049	0,213	0,018	0,212	0,094	0,038	0,208	0,014	0,206	0,207
Coef. variación	0,516	0,439	0,442	0,650	0,471	0,703	0,534	1,831	2,639	0,520	0,482	0,521	0,538
Varianza ($\times 10^8$)	102,518	6,088	0,020	0,241	4,547	0,033	4,505	0,893	0,141	4,333	0,020	4,244	4,276

Los coeficientes de desviación típica y varianza no están en la misma escala al igual que en la tabla. Esto hace que no sean medidas adecuadas para comparar la variabilidad entre distintas variables. Por este motivo, podemos utilizar el coeficiente de variación ya que está normalizado. A partir de este coeficiente, vemos como las variables $\{H, I\}$ son las que presentan una mayor variabilidad, mientras que el resto de variables presentan menor variabilidad pero dentro del mismo rango. Por otro lado, podemos conocer la asimetría y homogeneidad de las variables del problema a partir de los respectivos coeficientes,

$$A_j = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{ij} - \bar{x}_j)^3}{s_j^3} \qquad \text{Coeficiente de asimetría}$$

$$K_j = \frac{1}{n}\sum_{i=1}^{n}\frac{d_{ij}^2}{s_j^4} \qquad d_{ij} = (x_{ij} - \bar{x}_j)^2 \qquad \text{Coeficiente de kurtosis}$$

Usamos los comandos de la librería *e1071*:

```
library(e1071) #Importamos la libreria
apply(AM20, 2, skewness) #Asimetrias
apply(AM20, 2, kurtosis) #Kurtosis
```

Para el conjunto de datos toman los valores:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
Asimetría	0,936	-0,157	0,013	0,375	0,416	1,005	-0,088	4,224	6,865	0,322	-0,010	0,323	0,344
Kurtosis	0,959	-1,020	-0,633	-0,762	-0,444	1,561	-0,973	26,527	61,910	-0,519	-0,744	-0,501	-0,534

Podemos observar como las variables $\{H, I\}$ presentan un alto grado de asimetría así como de kurtosis. Estos valores hacen sospechar de la **presencia de valores atípicos** (principalmente el alto grado de kurtosis, con el criterio $k_j \geq 8$).

Por el otro lado, el resto de variables presentan ambas magnitudes en el mismo rango de valores y próximas a cero, lo que nos hace suponer que son poblaciones heterogéneas y sin presencia de atípicos muy notables. En el posterior apartado de representación gráfica se interpretarán estas magnitudes estadísticas. En segundo lugar, se estudian las medidas de variabilidad multivariantes para el conjunto de datos. Empezamos por calcular la matriz de varianzas y covarianzas, que toma la forma

$$S_{ij} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) \cdot (x_{ik} - \bar{x}_k) \qquad \text{Matriz de varianzas y covarianzas}$$

se invoca en R mediante el comando:

```
S <- var(AM20)
```

y toma los valores:

($\times 10^8$)	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
A	102,518	0,254	0,070	-0,018	0,393	-0,184	0,083	-0,263	-0,090	0,656	0,047	0,806	0,758
B	0,254	6,088	0,170	1,144	4,071	0,351	5,135	0,323	0,185	3,748	0,250	3,640	3,576
C	0,070	0,170	0,020	0,026	0,148	0,007	0,130	-0,015	-0,003	0,163	0,011	0,162	0,159
D	-0,018	1,144	0,026	0,241	0,834	0,071	1,019	0,093	0,052	0,742	0,048	0,716	0,702
E	0,393	4,071	0,148	0,834	4,547	0,246	3,517	0,553	0,217	3,994	0,258	3,862	3,788
F	-0,184	0,351	0,007	0,071	0,246	0,033	0,305	0,022	0,015	0,224	0,015	0,215	0,213
G	0,083	5,135	0,130	1,019	3,517	0,305	4,505	0,337	0,186	3,181	0,211	3,085	3,025
H	-0,263	0,323	-0,015	0,093	0,553	0,022	0,337	0,893	0,161	-0,340	-0,025	-0,337	-0,334
I	-0,090	0,185	-0,003	0,052	0,217	0,015	0,186	0,161	0,141	0,056	0,003	0,054	0,058
J	0,656	3,748	0,163	0,742	3,994	0,224	3,181	-0,340	0,056	4,333	0,284	4,200	4,122
K	0,047	0,250	0,011	0,048	0,258	0,015	0,211	-0,025	0,003	0,284	0,020	0,268	0,264
L	0,806	3,640	0,162	0,716	3,862	0,215	3,085	-0,337	0,054	4,200	0,268	4,244	4,145
M	0,758	3,576	0,159	0,702	3,788	0,213	3,025	-0,334	0,058	4,122	0,264	4,145	4,276

Los valores de la diagonal principal contienen las varianzas de las variables, mientras que el resto de valores representan las covarianzas entre pares de variables.

Es posible comprobar si existen variables redundantes observando los autovalores de esta matriz. Para ello usamos el comando:

```
ev <- eigen(S)$values #Autovalores
```

($\times 10^{14}$)	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}	λ_{13}
Autovalores	102,542	23,976	3,457	1,343	0,216	0,121	0,092	0,079	0,014	0,010	0,006	0,001	0

Vemos que el último valor, λ_{13} toma el valor $\lambda_{13} = 2,56 \cdot 10^{-8}$ muchos órdenes de magnitud menor que el resto de autovalores por lo que podemos tomarlo como cero. Esto significa que la matriz de varianzas y covarianzas es linealmente dependiente y existe una variable redundante que es linealmente proporcional al resto. La identificaremos posteriormente calculando los coeficientes de correlación múltiple R^2. A partir de esta matriz, podemos describir las siguientes medidas de varianza:

$$T = tr(S) = \sum_{i=1}^{p} s_i^2 \qquad \text{Varianza total}$$

$$\bar{s}^2 = \frac{1}{p}\sum_{i=1}^{p} s_i^2 \qquad \text{Varianza media}$$

$$VG = \det(S) \qquad \text{Varianza generalizada}$$

$$VE = \sqrt[p]{\det(S)} \qquad \text{Varianza efectiva}$$

calculadas mediante los comandos:

```
T <- sum(diag(S)) #Varianza total
s2 <- T / 13 #Varianza media
VG <- det(S) #Varianza generalizada
VE <- det(S)^(1/13) #Varianza efectiva
```

en nuestro caso toman los valores:

$$T = 1,319 \cdot 10^{10}$$

$$\bar{s}^2 = 1,014 \cdot 10^{9}$$

$$VG = 4,980 \cdot 10^{79}$$

$$VE = 1,351 \cdot 10^{6}$$

vemos como la varianza generalizada, a pesar de ser el determinante de una matriz (casi) linearmente dependiente, no toma el valor cero. Esto puede deberse a la alta sensibilidad del determinante a los cambios de escala unido al hecho de que la matriz no es linealmente dependiente al completo. En efecto, si tomamos el determinante de la matriz de varianzas y covarianzas sin tener en cuenta el factor de escala $\times 10^8$ obtenemos el valor

$$\det(S) = 6,156 \cdot 10^{-25}$$

que captura mejor el hecho de que es aproximadamente linealmente dependiente. Por último, podemos calcular los coeficientes de asimetría y homogeneidad para el caso multivariante utilizando los coeficientes generalizados para el caso multivariante. Las expresiones toman la forma:

$$A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3 \qquad \text{Coeficiente de asimetría multivariante}$$

$$K_p = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2 \qquad \text{Coeficiente de kurtosis multivariante}$$

donde d_{ij} es la distancia de Mahalanobis. Para calcularlos hemos usado la librería *Multivariate Normality (MVN)* a partir del método *mvn*. En particular, calculamos los coeficientes de asimetría y kurtosis de Mardia.

```
library(MVN)
mardia <- mvn(AM20, mvnTest="mardia")
mardia$multivariateNormality
```

con los valores:

$$A_p = 42360.11$$

$$K_p = 316.56$$

1.c Análisis de dependencias lineales

Continuamos mediante el cálculo de las dependencias lineales entre variables del conjunto de datos. Existen varias medidas para estudiar la dependencia lineal en término de que variables se utilizen. Como pueden existir correlaciones lineales internas y redundancias, comenzamos describiendo los coeficientes de correlación múltiple, R^2, para cada una de las variables. Este coeficiente es

$$R_j^2 = \frac{1}{s^{jj}s_{jj}} \qquad \text{Coeficiente de correlación múltiple}$$

donde

$$s_{jj} = \{S\}_{jj} \qquad s^{jj} = \{S^{-1}\}_{jj}$$

Para calcularlo hemos recurrido a un procedimiento alternativo, calculando cada uno de los ajustes lineales de una variable con todas las demás ya que *R* calcula automáticamente el valor de R^2 para cada uno de estos ajustes. Por ejemplo, para calcular R^2 para la variable A utilizaríamos los comandos:

```
modelA <- lm(A ~ B+C+D+E+F+G+H+I+J+K+L+M, data = AM20)
summary(modelA) #Obtenemos el valor de R^2 para este ajuste
```

haciendo esto para todas las variables obtenemos los valores:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
R^2	0,033	0,971	0,430	0,974	1	0,651	0,988	1	0,278	1	0,936	0,978	0,947

vemos como las variables $\{E, H, J\}$ toman un valor $R^2 = 1$ lo que significa que presentan una relación lineal exacta con otras variables del conjunto. Por este motivo, algunas de ellas son redundantes y pueden ser eliminadas.

Para conocer cuantas variables son redundantes, calculamos el rango de la matriz de varianzas y covarianzas como el número máximo de autovalores de la matriz distintos de cero. A partir de la sección anterior, vemos que $rg(S) = 12$ y existen 12 variables independientes, por lo tanto es necesario borrar una de las variables redundantes anteriores. Procediendo a eliminar la variable *E*, vemos como las redundancias internas desaparecen y dejan de existir variables con $R^2 = 1$.

```
qr(S)$rank #Rango de S
S <- subset(S, select=-c(E)) #Eliminar la variable E
```

Repitiendo el procedimiento para los R^2:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
R^2	0,033	0,971	0,430	0,974	-	0,651	0,988	0,372	0,278	0,985	0,936	0,978	0,947

A partir de ahora realizamos los cálculos para este nuevo conjunto de datos sin la variable *E*. Para este nuevo conjunto de variables calculamos la matriz de correlación, que describe todas las correlaciones existentes entre pares de variables y cuyos elementos son los coeficientes de correlación lineales entre variables

$$r_{jk} = \frac{S_{jk}}{s_j s_k}$$

```
cor(AM20) #Matriz de correlaciones
```

	A	B	C	D	F	G	H	I	J	K	L	M
A	1	0,010	0,049	-0,004	-0,100	0,004	-0,027	-0,024	0,031	0,032	0,039	0,036
B	0,010	1	0,491	0,945	0,779	0,980	0,138	0,200	0,730	0,709	0,716	0,701
C	0,049	0,491	1	0,382	0,289	0,436	-0,115	-0,058	0,560	0,547	0,559	0,550
D	-0,004	0,945	0,382	1	0,790	0,979	0,200	0,282	0,726	0,688	0,708	0,692
F	-0,100	0,779	0,289	0,790	1	0,788	0,130	0,220	0,589	0,570	0,572	0,563
G	0,004	0,980	0,436	0,979	0,788	1	0,168	0,234	0,720	0,696	0,706	0,689
H	-0,027	0,138	-0,115	0,200	0,130	0,168	1	0,454	-0,173	-0,189	-0,173	-0,171
I	-0,024	0,200	-0,058	0,282	0,220	0,234	0,454	1	0,072	0,061	0,070	0,075
J	0,031	0,730	0,560	0,726	0,589	0,720	-0,173	0,072	1	0,955	0,979	0,958
K	0,032	0,709	0,547	0,688	0,570	0,696	-0,189	0,061	0,955	1	0,910	0,894
L	0,039	0,716	0,559	0,708	0,572	0,706	-0,173	0,070	0,979	0,910	1	0,973
M	0,036	0,701	0,550	0,692	0,563	0,689	-0,171	0,075	0,958	0,894	0,973	1

vemos como se trata de una matriz simétrica donde los elementos de la diagonal principal son las correlaciones de un elemento consigo mismo y por tanto equivalen a la unidad y el resto de elementos se corresponden con las correlaciones entre pares de valores teniendo en cuenta la influencia de el resto de variables. Existen variables con un alto grado de correlación, como por ejemplo los pares $\{B, D\}$, $\{D, G\}$, $\{J, K\}$ o $\{J, L\}$. A su vez existen pares de variables con correlación negativa, sin embargo esta suele ser pequeña. Es interesante observar como las variables que tenían un alto grado de kurtosis y asimetría, $\{H, I\}$, presentan un grado de correlación muy reducido con el resto de variables, lo que de nuevo puede ser una señal de datos atípicos distorsionando las variables.De todas maneras, para estas dos variables las medidas pueden no ser estadísticamente significativas debido a su alto grado de kurtosis (no son variables normalmente distribuidas).

Seguimos describiendo las correlaciones parciales entre todos los pares de variables mediante la matriz de correlación parcial. Los elementos de esta matriz son los coeficientes de correlación parcial

$$r_{jk;1,2\ldots,p} = -\frac{s^{jk}}{\sqrt{s^{jj}s^{kk}}}$$

y describen la correlación entre las variables j y k sin tener en cuenta la influencia de todas las demás $1, 2, \ldots, p$. Usamos la librería *ppcor*:

```
library(ppcor)
pcor(AM20, method = c("pearson"))
```

Los resultados son los siguientes:

	A	B	C	D	F	G	H	I	J	K	L	M
A	1	0,028	0,005	0,007	-0,165	-0,003	-0,003	-0,003	-0,032	0,026	0,034	0,000
B	0,028	1	0,222	-0,344	0,142	0,793	0,048	-0,020	0,075	-0,048	-0,045	0,028
C	0,005	0,222	1	-0,189	-0,109	0,011	0,048	-0,056	0,056	0,003	0,021	0,021
D	0,007	-0,344	-0,189	1	0,141	0,806	0,160	0,172	0,329	-0,278	-0,159	0,012
F	-0,165	0,142	-0,109	0,141	1	-0,035	-0,031	0,031	-0,010	0,034	-0,011	0,022
G	-0,003	0,793	0,011	0,806	-0,035	1	-0,061	-0,085	-0,235	0,212	0,134	-0,029
H	-0,003	0,048	0,048	0,160	-0,031	-0,061	1	0,372	-0,027	-0,076	-0,026	-0,023
I	-0,003	-0,020	-0,056	0,172	0,031	-0,085	0,372	1	-0,070	0,073	0,027	0,039
J	-0,032	0,075	0,056	0,329	-0,010	-0,235	-0,027	-0,070	1	0,770	0,707	0,045
K	0,026	-0,048	0,003	-0,278	0,034	0,212	-0,076	0,073	0,770	1	-0,395	0,020
L	0,034	-0,045	0,021	-0,159	-0,011	0,134	-0,026	0,027	0,707	-0,395	1	0,560
M	0,000	0,028	0,021	0,012	0,022	-0,029	-0,023	0,039	0,045	0,020	0,560	1

vemos como las correlaciones han cambiado significativamente. Por ejemplo, la mayor parte de los pares de variables que antes estaban correlados fuertemente ahora dejan de estarlo (observar $\{B, D\}$, $\{J, K\}$ o $\{J, L\}$).

Por último calculamos el coeficiente de dependencia efectiva, que se puede describir como la cantidad de variabilidad del conjunto descrita por relaciones lineales entre variables. Se define como:

$$D(R) = 1 - \sqrt[p-1]{\det R}$$

se puede interpretar como que cuanto mayor sea el determinante de la matriz de correlaciones, más incorrelacionadas estarán las variables (si $\det R = 0$ estarán perfectamente correlacionadas y viceversa para $\det R = 1$). Esta variable expresa en tanto por ciento la cantidad de variabilidad del conjunto de datos explicada mediante relaciones lineales entre variables. Tenemos que

$$D(R) = 0,814$$

lo que quiere decir que un 81,4% de la variabilidad de los datos se debe a relaciones lineales entre variables, mientras que el 18,6% restante puede deberse a relaciones no-lineales o a ruido.

2 Representación gráfica de los datos

Continuamos la práctica haciendo un análisis gráfico del conjunto de datos. El método más sencillo es calcular histogramas para cada una de las variables y así estudiar su distribución. Esto nos permite también visualizar la asimetría y kurtosis de la distribución. Por otro lado, un método gráfico para estudiar pares de variables es el diagrama de dispersión, que muestra los puntos del conjunto de datos en términos de sus valores en este par de variables.

Podemos hacer ambos gráficos simultáneamente mediante una matriz de dispersión, en la cual los elementos de la diagonal principal son histogramas y el resto diagramas de dispersión. Desafortunadamente, las variables del conjunto son demasiadas como para mostrarlas todas simultáneamente, por tanto nos limitaremos a realizar la matriz de dispersión a cada una de las mitades del conjunto de datos. Utilizamos el método *pairs.panels* de la libraría *psych* de R, la cual muestra la correlación de Pearson en la mitad superior de la matriz.

```
library(psych)
AM20_1 <- subset(AM20, select=-c(H,I,J,K,L,M))
AM20_2 <- subset(AM20, select=-c(A,B,C,D,F,G))
pairs.panels(AM20_1)
pairs.panels(AM20_2)
```

En la figura 2.1 observamos la matriz de dispersión para las primeras seis variables del conjunto. Vemos como siguen distribuciones aproximadamente normales, con diferentes grados de kurtosis y asimetría. Por ejemplo, las variables A y F tienen una asimetría negativa pero poco grado de kurtosis, mientras que el resto de variables presentan un mayor grado de kurtosis. La mitad inferior muestra los diagramas de dispersión por pares de las variables, y la mitad superior la correlación de Pearson entre ambas. Vemos como existe una correlación muy fuerte entre algunos pares de variables, a saber $\{B, D\}$, $\{B, G\}$, y $\{D, G\}$, mientras que otros pares de variables como $\{A, D\}$ o $\{A, G\}$ están absolutamente descorrelacionados.

En la figura 2.2 realizamos el mismo trabajo pero para las últimas 6 variables del conjunto de datos. Vemos como las últimas cuatro variables están muy correlacionadas entre sí, mientras que las variables H e I son peculiares ya que presentan un alto grado de ocurrencias en la base del histograma pero también presentan una cola muy larga (alto grado de asimetría y kurtosis). Es sobre estas dos variables donde se sospecha la presencia de atípicos.

Continuamos la parte de procedimientos gráficos aplicando métodos de disminución de la dimensionalidad para poder visualizar todo el conjunto de datos en función de dos variables. Esto se puede entender como una proyección lineal sobre el subespacio definido por ambas variables. Estas variables se definirán a partir de un criterio, que normalmente suele ser minimizar una magnitud como por ejemplo minimizar la correlación entre variables (PCA) o maximizar la kurtosis multivariante del conjunto final de datos. Calculamos la proyección del conjunto de datos que minimiza la kurtosis final. Para ello usamos el método *MinSkew* de la librería *MultiSkew*.

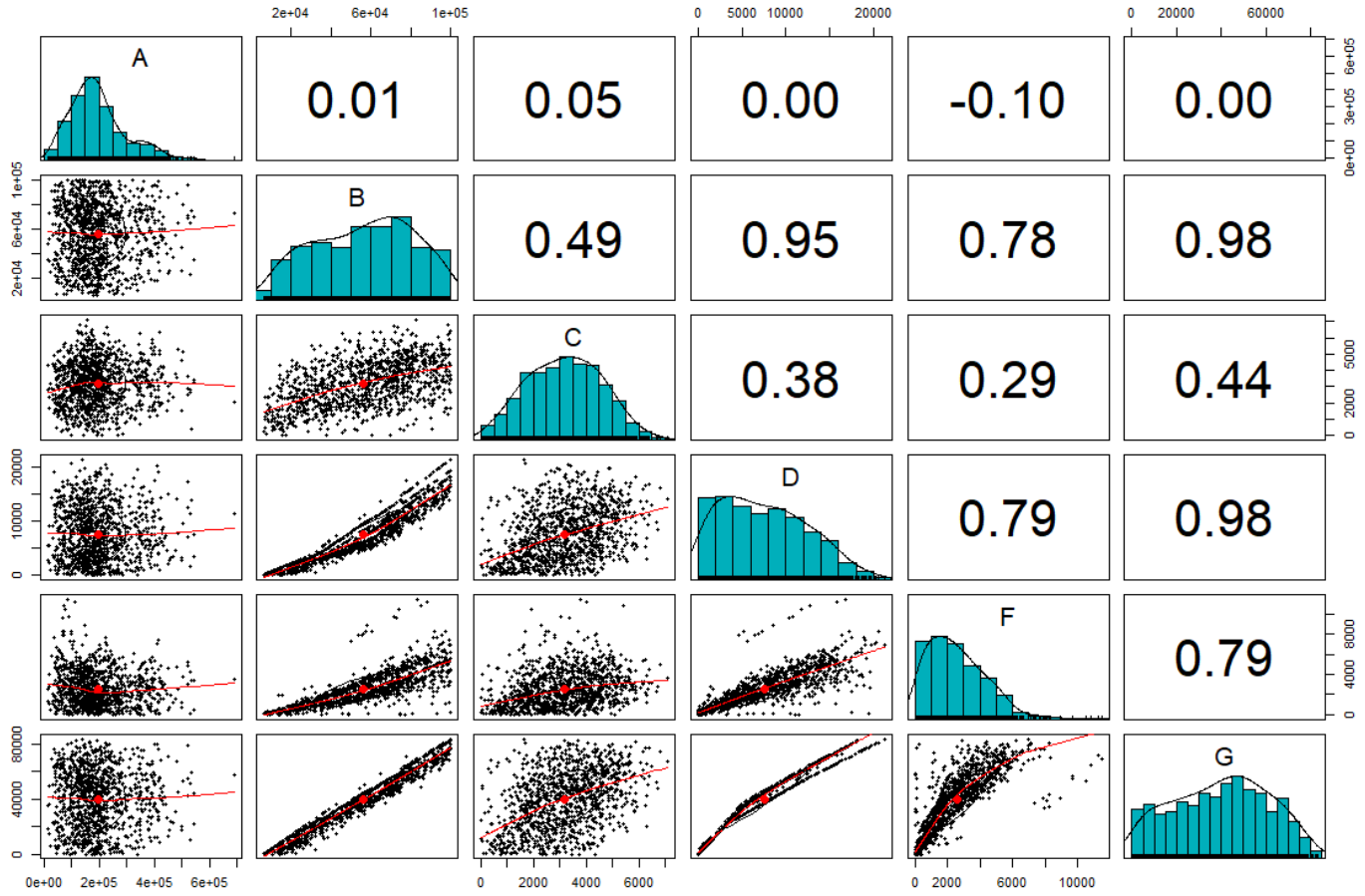

Figure 2.1: *Matriz de dispersión de las primeras 6 variables*

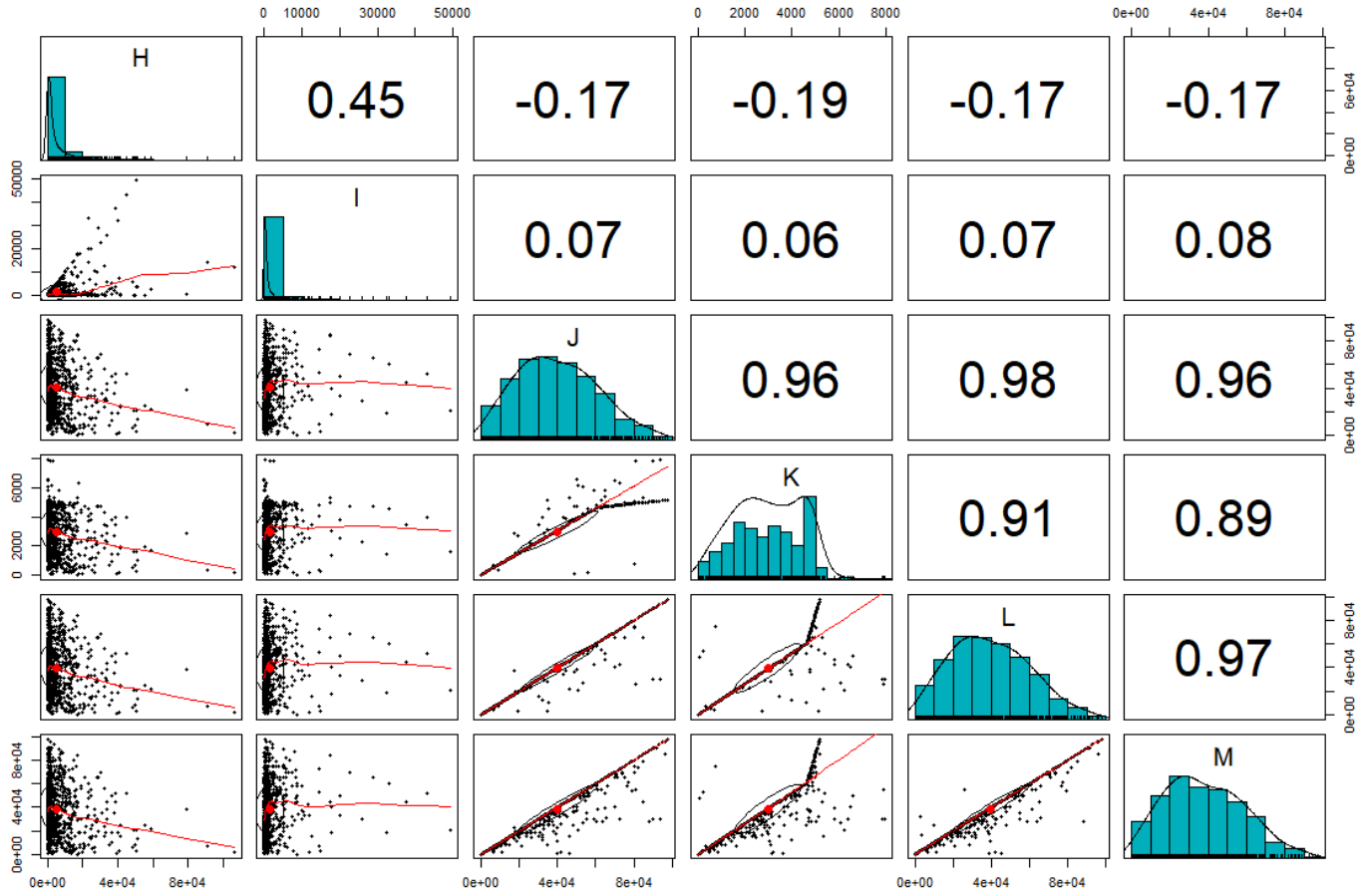

Figure 2.2: *Matriz de dispersión de las segundas 6 variables*

```
library(MultiSkew)
AM20_matrix <- data.matrix(AM20)
MinSkew(AM20_matrix, 2)
plot(Projections) #Contiene los valores de la proyeccion
```

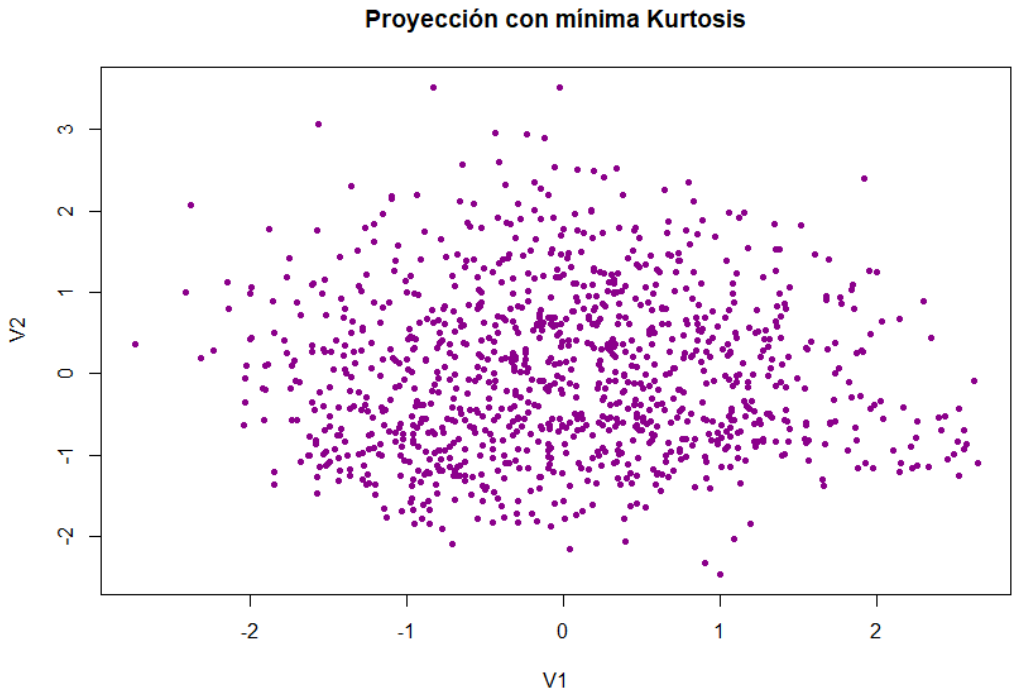

Figure 2.3: *Matriz de dispersión de la proyección con mínima kurtosis*

En esta figura 2.3 observamos como el resultado presenta una distribución muy semejante a la normal, consecuencia de minimizar la kurtosis del conjunto final. Los coeficientes de las nuevas variables son

	A	**B**	**C**	**D**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
V1	-0,170	-0,082	-0,965	0,098	0,009	-0,126	-0,001	0,009	-0,048	-0,020	-0,039	-0,053
V2	-0,119	0,762	-0,002	-0,075	-0,056	-0,564	0,018	-0,003	0,148	0,053	0,171	0,154

Vemos como la variable dominante para la primera componente de mínima kurtosis es la C, mientras que para la segunda una combinción de las variables B y G.

Podemos hacer un procedimiento análogo proyectando los datos sobre sus dos componentes principales. Este método busca proyectar los datos de forma que se distorsionen lo menor posible (menor pérdida de información), entendiendo esto como la proyección que conserve al máximo posible la variabilidad del problema original. Para ello usamos el método *prcomp* de la librería *stats* instalada por defecto sobre el conjunto de datos. Es muy importante estandarizar los datos con el método *scale* ya que hemos visto que la primera variable es mucho más grande que el resto y dominaría el procedimiento.

```
library(stats)
scaled.AM20 <- scale(AM20)
PCA <- prcomp(scaled.AM20)
X <- PCA$x
plot(X) #Representa las dos primeras columnas PC1 y PC2
summary(PCA) #Varianza acumulada de las componentes
```

El resultado es:

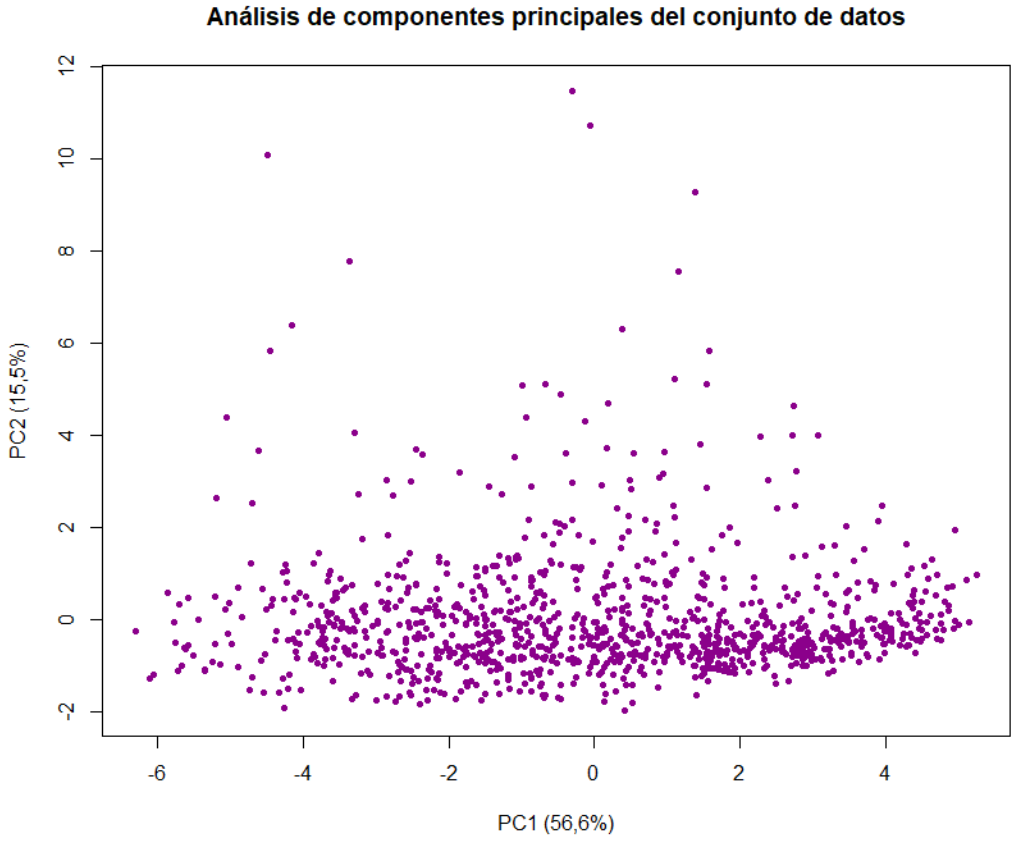

Figure 2.4: *Proyección sobre las dos componentes principales*

Estas dos primeras componentes principales explican el 72,1% de la varianza total de los datos y nos permiten hacernos una idea de su estructura a grandes rasgos. Vemos como no existen grupos o clusters claramente diferenciados, lo que puede significar que el conjunto de datos procede de la misma fuente, sin embargo están concentrados en una zona del gráfico. Si tuviéramos una variable categórica conteniendo etiquetas para cada punto del conjunto de datos, podríamos interpretar el gráfico con mayor profundidad ya que se podrían distinguir las regiones donde se concentra cada conjunto de datos.

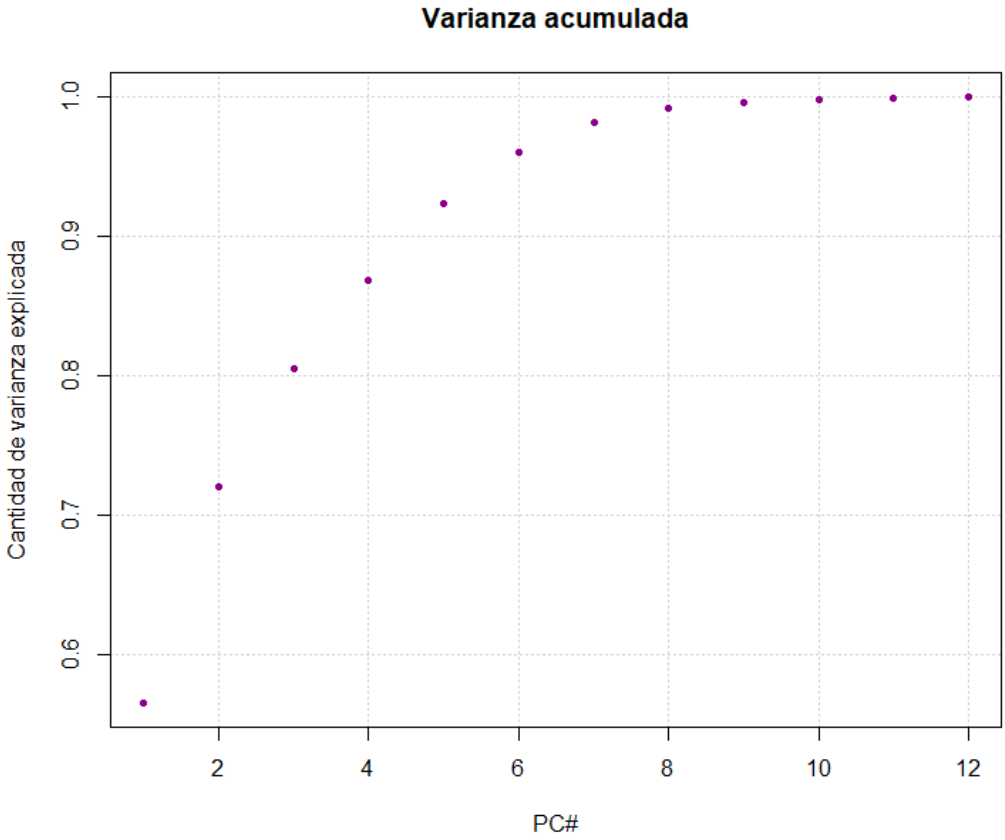

Figure 2.5: *Varianza explicada por las componentes principales*

Vemos como las primeras cinco componentes principales explican más del 90% de la variabilidad del problema. La combinación lineal de variables originales que da lugar a las componentes principales es:

	A	**B**	**C**	**D**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
PC1	-0,01	-0,35	-0,23	-0,35	-0,29	-0,35	0,01	-0,07	-0,36	-0,35	-0,35	-0,35
PC2	-0,10	0,17	-0,22	0,23	0,22	0,21	0,61	0,51	-0,18	-0,19	-0,19	-0,19

Si conocieramos el significado de cada una de las variables originales se podría dar una interpretación estadística de estas nuevas variables.

3 Detección de valores atípicos

Buscamos valores atípicos del conjunto de datos usando dos métodos. El primero se basa en un análisis univariante de las variables que se sospeche que puedan contener estos valores.

El método estará basado en el criterio

$$crit(x) = \frac{|x_i - med(x)|}{MEDA(x)} \qquad MEDA(x) = med|x_i - med(x)|$$

para cada instancia de la variable. Si esta métrica es superior a 4.5 consideraremos dicho valor como atípico.

```
fun_outlier <- function(x) {apply(x, 2, function(y)
    abs(y-mean(y))/mad(y))}
outliers <- fun_outlier(AM20)
outliers.df <- as.data.frame(outliers)
#Contamos el numero de atipicos para cada variable
outliersI <- nrow(subset(outliers.df, outliers.df$I>4.5)) #103
outliersJ <- nrow(subset(outliers.df, outliers.df$J>4.5)) #101
#Representamos los histogramas
hist(outliersI$I)
hist(outliersJ$J)
```

Realizando este análisis sobre las variables, obtenemos los siguientes resultados:

Variable	**A**	**B**	**C**	**D**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
Número de atípicos	0	1	0	0	3	0	101	103	0	0	0	0

vemos como efectivamente siguiendo este criterio las variables *H* e *I* tienen un mayor número de atípicos. Esto era de esperar viendo sus valores para la asimetría y kurtosis.

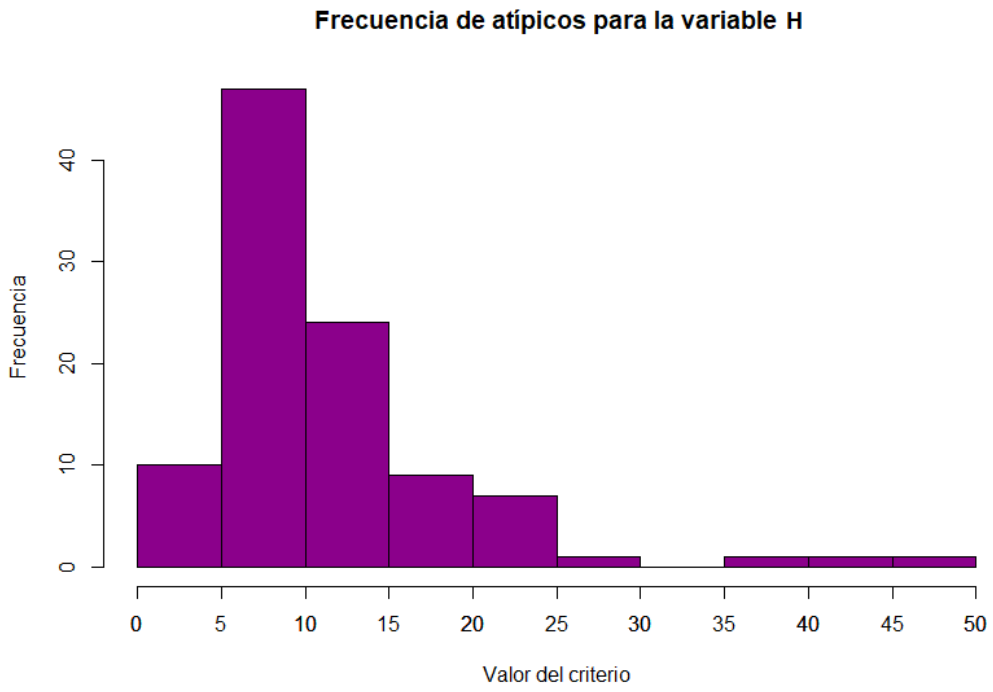

Figure 3.1: *Distribución de atípicos para la variable H*

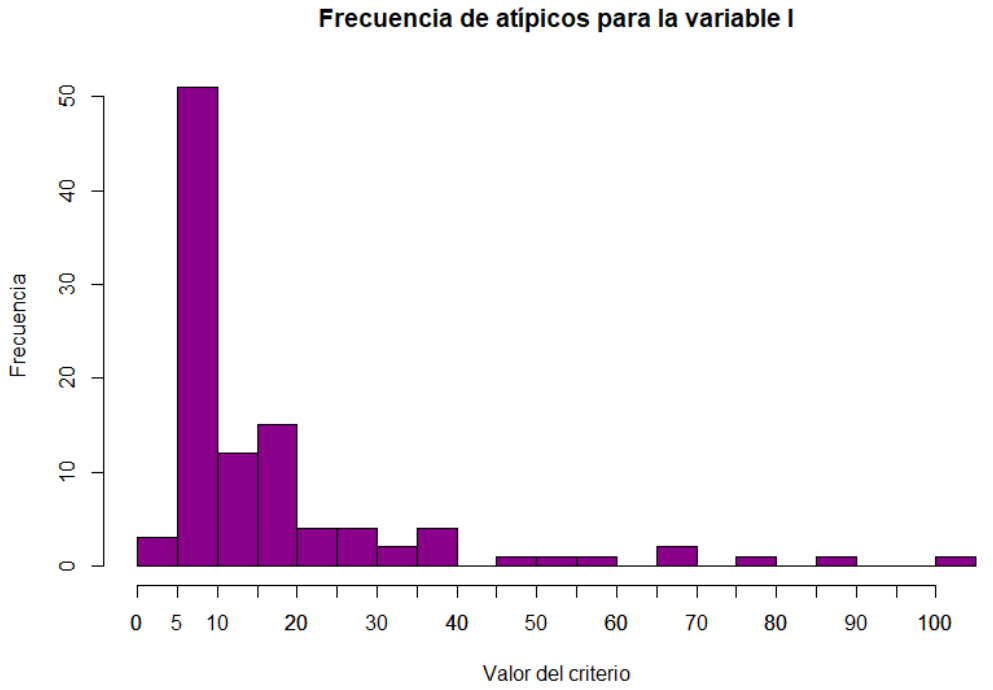

Figure 3.2: *Distribución de atípicos para la variable I*

Vemos como ambas variables tienen multitud de valores con un alto valor del criterio seleccionado para encontrar atípicos, aunque la mayor parte se encuentran cerca del valor límite de 4.5, entre 5 y 10 por lo que confirmar que son atípicos requeriría de una mayor atención.

Continuamos encontrando atípicos para el conjunto de datos estudiando todas las variables simultaneamente. Para ello aplicamos un método de cálculo basado en la distancia de Mahalanobis sobre la proyección de mínima kurtosis hayada anteriormente. Esto nos permite visualizar los valores atípicos con facilidad.

Procedemos a aplicar el procedimiento de proyección de mínima kurtosis visto en la sección anterior mediante el método *MinSkew* y después aplicamos el método de detección de outliers *maha* de la librería *OutlierDetection* sobre esta proyección:

```
library(MultiSkew)
library(OutlierDetection)
AM20_matrix <- data.matrix(AM20)
MinSkew(AM20_matrix, 2)
maha(Projections)
```

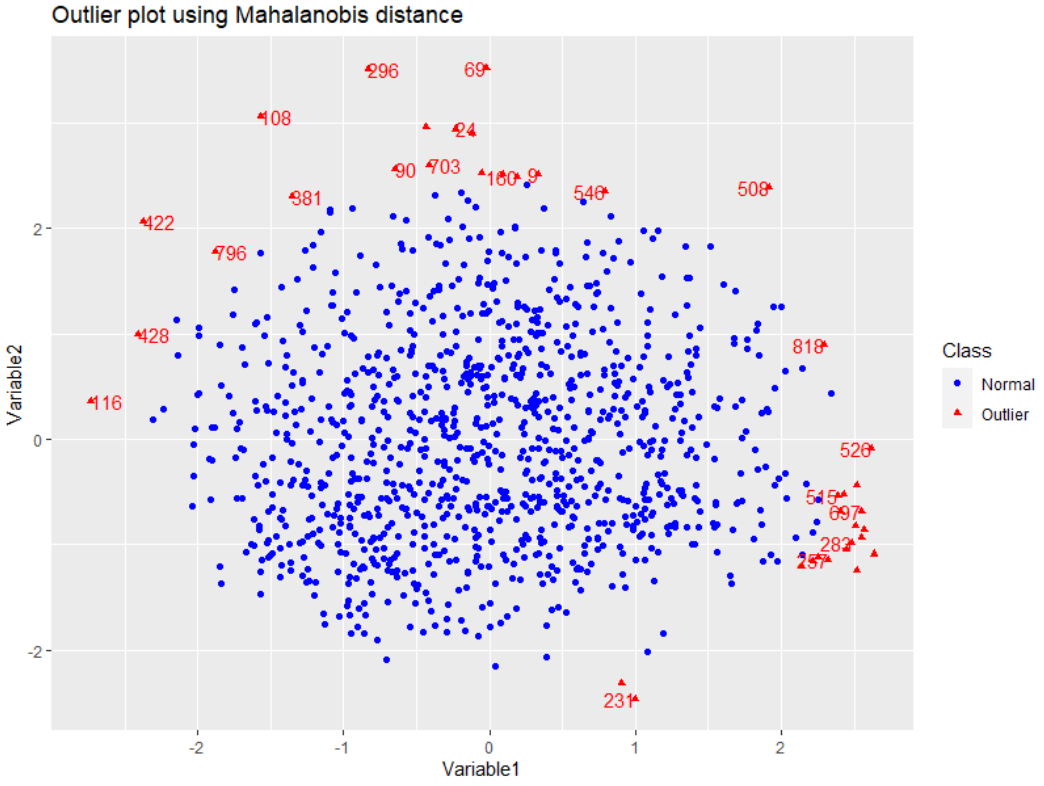

Figure 3.3: *Distribución de atípicos sobre la proyección de mínima kurtosis*

La librería utilizada nos permite aplicar más métodos. Por ejemplo, podemos usar el método basado en Robust Kernal-based Outlier Factor (RKOF).

```
dens(Projections)
```

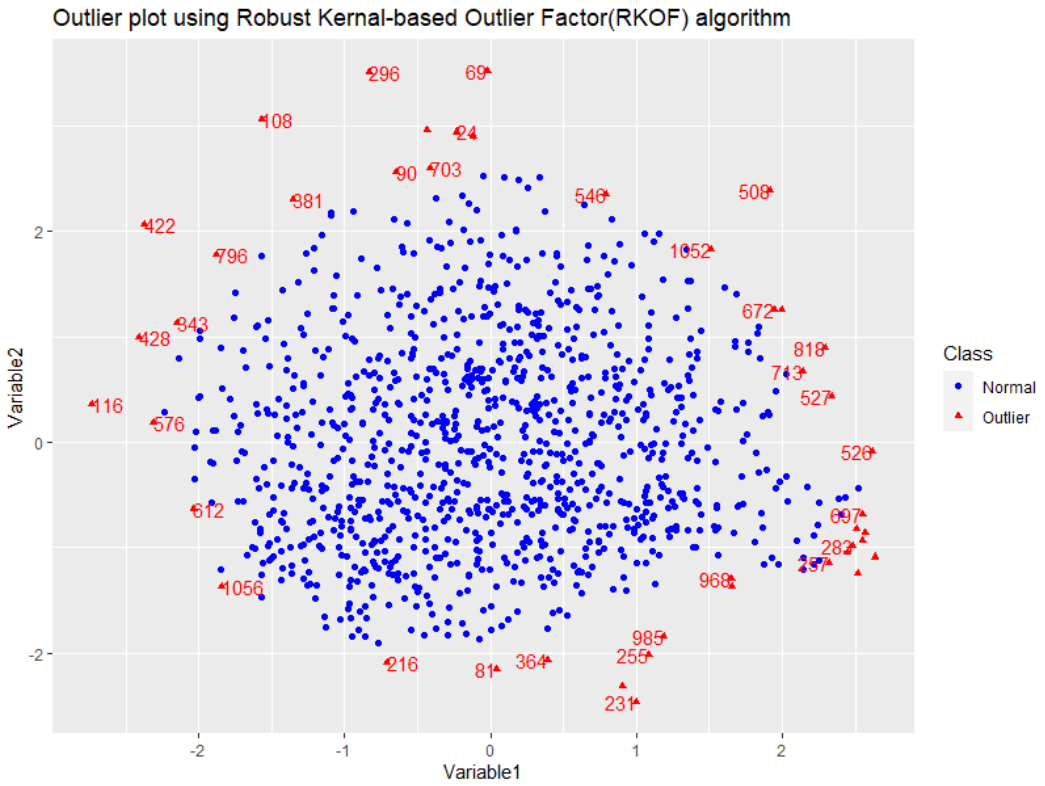

Figure 3.4: *Distribución de atípicos sobre la proyección de mínima kurtosis*

Vemos como los resultados son muy parecidos. Mostramos las probabilidades para los atípicos en las siguientes tablas:

Posición	Mahalanobis (40)	RKOF (47)
9	96%	
24	99%	100%
69	100%	100%
81		100%
86	99%	100%
90	97%	95%
108	100%	100%
116	98%	100%
160	96%	
216		100%
231	97%	100%
255		100%
257	97%	100%
283	97%	99%
296	100%	100%
322	96%	
343		100%
364		100%
381	97%	100%
406	95%	
422	99%	100%
428	97%	100%
454	96%	
458	96%	100%
502	96%	
508	99%	100%
515	95%	
526	97%	100%
527		100%
546	95%	100%
576		100%
612		98%
672		100%
674	97%	100%
697	97%	96%
703	97%	97%
713		100%
796	96%	100%
818	95%	100%
848	96%	
887	99%	100%
930	96%	
931		100%
938	98%	100%
948	97%	98%
952	97%	97%
968		100%
985		100%
1038	98%	100%
1045	98%	100%
1052		100%
1056		100%
1068		100%
1069	95%	
1070	96%	

Table 3.1: *Distribución de atípicos sobre la distribución bivariante para RKOF y Mahalanobis*

Vemos como la técnica univariante identifica una mayor cantidad de datos atípicos. Esto se debe a que las variables H e I presentan una distribución peculiar, con un alto grado de kurtosis y asimetría. Viendo el histograma del gráfico 2.2 vemos que se debe a que la mayor parte de los datos se encuentran concentrados al principio de la distribución, pero presentan una cola muy larga y estos datos se identifican como atípicos.

Por otro lado, las técnicas multivariantes detectan un menor número de atípicos porque tienen en consideración la influencia de todas las variables. La técnica que se ha empleado es una de las muchas que existen para hallar atípicos de manera multivariante. En esta técnica, hemos proyectado los datos sobre el subespacio de menor kurtosis y después hemos encontrado los atípicos en este espacio. En la bibliografía de la asignatura, se recomienda usar el criterio

$$crit(x) = \frac{|x_i - med(x)|}{MEDA(x)} \qquad MEDA(x) = med|x_i - med(x)|$$

sobre esta nueva distribución bivariante, pero por simplicidad se ha optado por recurrir a la librería *OutlierDetection*. Los métodos empleados, por distancia de Mahalanobis y por RKOF dan lugar a resultados muy similares como se puede observar en las figuras 3.3 y 3.4, aunque podemos observar como el último método da lugar a más atípicos y con un intervalo de confianza más reducido. Vemos como la mayoría de los valores coinciden entre sí como es lógico.

Es importante destacar que los atípicos detectados mediante este procedimiento deben considerarse *sospechosos* más que *definitivos*, y la consideración definitiva de cuales se toman como atípicos o no debe basarse en un estudio con mayor profundidad que el llevado a cabo aquí. Esto se debe a que la proyección bivariante que estamos considerando (mínima kurtosis) puede no capturar correctamente el comportamiento de los atípicos, y este tipo de datos puede contener información muy relevante del conjunto por lo que etiquetarlos de manera errónea puede tener consecuencias negativas sobre el análisis estadístico. Sería interesante realizar este procedimiento para otro tipo de proyecciones bivariantes (por ejemplo, máxima kurtosis o mínima/máxima asimetría) o incluso sobre proyecciones de mayor dimensión, y comparar los resultados con los obtenidos por estos dos métodos para así considerar con mayor confianza qué datos pueden ser etiquetados definitivamente como atípicos y posteriormente descartados.