Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Análisis Multivariante
Cali, septiembre de 2018

Laboratorio 1: Construcción de la Inercia

Kevin Steven García[a], Alejandro Vargas[b], Alejandro Soto[c]

1. Introducción

En este informe se llevara a cabo un proceso de análisis multivariado con conceptos como la Inercia, centrado y estandarización de los datos,el cálculo de la matriz de varianzas y covarianzas y la matriz de correlaciones, descomposición en valores y vectores propios, construcción de las componentes principales o las proyecciones de la matriz de datos con los vectores propios asociados, y además, se construiran gráficos tanto en segunda como en tercera dimensión donde se representarán las plantas de acuerdo a sus 4 variables. Este proceso se le realizará a la siguiente matriz que contiene datos de 10 plantas acerca de 4 variables que son, X1: altura de la planta (en metros), X2: longitud radicular (en cms), X3: área foliar (en cms^3) y X4: peso en pulpa (en gramos)

Planta	X1	X2	X3	X4
1	1.89	53	4.5	120
2	1.92	62	5.2	136
3	1.95	71	5.4	145
4	2.11	75	6.1	156
5	1.78	58	5.0	127
6	2.12	77	6.5	160
7	2.06	82	7.6	169
8	2.19	78	7.3	164
9	1.93	60	5.5	142
10	1.84	72	5.7	147

2. Punto 1

- Matriz de datos original:
  Primero, obtenemos el centro de gravedad o individuo promedio, en este caso, obtendremos la planta promedio de la siguiente manera:

$$g = (\bar{X}1, \bar{X}2, \bar{X}3, \bar{X}4)$$

  Entonces:

$$g = (1.979, 68.8, 5.88, 146.6)$$

  Posteriormente, calculamos la inercia con la formula:

$$I(N) = \sum p_i \|x_i - g\|_M^2$$

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co
[c]Código: 1532457. E-mail: asotomurillo@gmail.com

Lo que nos dio:

$$I(N) = 317.9118$$

- Matriz con datos centrados:
  Para la matriz con datos centrados (a cada valor de la variable se le resta su media correspondiente) sabemos que el centro de gravedad g está en el origen de coordenadas en R^p, como en nuestro caso tenemos 4 variables el centro de gravedad será

$$g = (0, 0, 0, 0)$$

Entonces, la formula para calcular la inercia se convierte en

$$I(N) = \sum p_i \|x_i\|_M^2$$

Haciendo uso de la formula obtuvimos el mismo resultado anterior:

$$I(N) = 317.9118$$

3. Punto 2

Para la matriz de datos estandarizada, es importante recordar que el proceso de estandarización se debe realizar con la desviación estándar sin la corrección de Bessel, es decir, se debe dividir sobre n y no sobre n-1.
Con esta matriz, el centro de gravedad g también es el origen de coordenadas en R^p (R^4 en nuestro caso)

$$g = (0, 0, 0, 0)$$

Aplicando la formulas de la inercia con el centro de gravedad anterior, obtenemos

$$I(N) = 4$$

La diferencia que se logra apreciar entre el calculo de la inercia utilizando la matriz de datos original o centrada y la matriz de datos estandarizada es que en los primeros dos casos, dicha inercia da como resultado la suma de las varianzas de cada variable, mientras que en el segundo caso (matriz de datos estandarizada)la inercia da como resultado el numero de variables involucradas, 4 en nuestro caso.

4. Punto 3

La matriz de varianzas y covarianzas de la matriz de datos es:

	Altura planta	Longitud radicular	Área foliar	Peso pulpa
Altura planta	0.01796556	0.9842222	0.1056444	1.746222
Longitud radicular	0.98422222	96.6222222	8.8733333	151.8
Área foliar	0.10564444	8.8733333	0.9951111	15.246667
Peso pulpa	1.746222	151.8	15.246667	255.6

5. Punto 4

La matriz de correlaciones es

	Altura planta	Longitud radicular	Área foliar	Peso pulpa
Altura planta	1	0.7470236	0.7901154	0.8148896
Longitud radicular	0.7470236	1	0.9049248	0.9659460
Área foliar	0.7901154	0.9049248	1	0.9560018
Peso pulpa	0.8148896	0.9659460	0.9560018	1

6. Punto 5

- Matriz de varianzas y covarianzas:
  La descomposición de esta matriz en valores y vectores propios, nos dio
  Los valores propios

$$\lambda_1 = 348.3709, \lambda_2 = 4.778776, \lambda_3 = 0.08000646, \lambda_4 = 0.005602631$$

  Los vectores propios

λ_1	λ_2	λ_3	λ_4
-0.005759331	-0.01300728	0.001249545	0.9998980345
-0.517026145	0.85544272	-0.028891125	0.0081861926
-0.050711126	-0.06431056	-0.996640617	0.0001167914
-0.854446708	-0.51372435	0.076623899	-0.0117001348

- Matriz de correlaciones:
  La descomposición de esta matriz en valores y vectores propios, nos dio
  Los valores propios

$$\lambda_1 = 3.59564465, \lambda_2 = 0.29512043, \lambda_3 = 0.09190433, \lambda_4 = 0.01733060$$

  Los vectores propios

λ_1	λ_2	λ_3	λ_4
-0.4636832	0.8742229	0.11689440	0.08407071
-0.5046999	-0.3941588	0.61315298	0.46255844
-0.5089282	-0.1956428	-0.77874615	0.31027468
-0.5208339	-0.20517575	0.06271807	-0.82625748

7. Punto 6

Las descomposiciones no son idénticas, esta diferencia se debe a que las dos matrices son diferentes, la primera es la de varianzas y covarianzas (presentan escalas muy grandes en sus valores) y la segunda es de correlaciones (el máximo valor que puede presentar es 1 que es la máxima correlación y además esta matriz tiene la propiedad de que su traza o la suma de los valores de su diagonal da como resultado el número de variables en cuestión).Las dos me explican aspectos diferentes sobre las variables,por ende sus descomposiciones son diferentes sin embargo, estas representan la misma cantidad de información contenida en los datos (los valores propios obtenidos tienen casi la misma proporción con respecto al total en ambos casos), ya que las correlaciones que componen la matriz se obtienen a partir de las varianzas y covarianzas, por lo tanto, se podría decir que las dos matrices guardan cierta relación pero no son iguales, por ello se explica su diferencia en las descomposiciones.

8. Punto 7

De acuerdo a los datos originales es mejor utilizar la matriz de correlaciones, ya que estos datos fueron medidos en diferentes unidades (metros, centímetros, centímetros cuadrados, gramos), y la matriz de correlaciones trabaja con los datos estandarizados, es decir, pone las variables en las mismas unidades o en la misma escala, por lo tanto es mejor para hacer un análisis multivariado correcto, evitando que el proceso se vea influenciado por grandes cantidades de una variable u otra dependiendo de su escala o unidad de medida.

9. Punto 8

- Proyección de la matriz de datos originales asociada a los vectores propios de la matriz de varianzas y covarianzas:
  Esta proyección la podemos encontrar haciendo el producto $A \cdot u$ donde A es la matriz de datos originales y u es la matriz compuesta por los vectores propios correspondientes a la matriz de varianzas y covarianzas. En nuestro caso:

$$\begin{pmatrix} 1.89 & 53 & 4.5 & 120 \\ 1.92 & 62 & 5.2 & 136 \\ 1.95 & 71 & 5.4 & 145 \\ 2.11 & 75 & 6.1 & 156 \\ 1.78 & 58 & 5.0 & 127 \\ 2.12 & 77 & 6.5 & 160 \\ 2.06 & 82 & 7.6 & 169 \\ 2.19 & 78 & 7.3 & 164 \\ 1.93 & 60 & 5.5 & 142 \\ 1.84 & 72 & 5.7 & 147 \end{pmatrix} \cdot \begin{pmatrix} -0.005759331 & -0.01300728 & 0.001249545 & 0.9998980345 \\ -0.517026145 & 0.85544272 & -0.028891125 & 0.0081861926 \\ -0.050711126 & -0.06431056 & -0.996640617 & 0.0001167914 \\ -0.854446708 & -0.51372435 & 0.076623899 & -0.0117001348 \end{pmatrix}$$

$$= \begin{pmatrix} -130.1751 & -16.62244 & 3.181117 & 0.9201849 \\ -148.5351 & -17.18845 & 3.449468 & 0.8367372 \\ -160.8887 & -14.12624 & 3.679773 & 0.8351320 \\ -172.3921 & -16.40253 & 3.709623 & 0.8992407 \\ -138.7661 & -15.97202 & 3.074571 & 0.7692845 \\ -176.8643 & -16.77240 & 3.559692 & 0.8788583 \\ -187.1949 & -17.18867 & 3.008472 & 0.7546226 \\ -180.8401 & -18.02421 & 3.040072 & 0.9103302 \\ -152.6430 & -22.00111 & 3.668014 & 0.7601980 \\ -163.1292 & -14.31611 & 3.505000 & 0.7099642 \end{pmatrix}$$

- Proyección de la matriz de datos centrados asociada a los vectores propios de la matriz de varianzas y covarianzas:
  Esta proyección la podemos encontrar haciendo el producto $A \cdot u$ donde A es la matriz de datos centrados y u es la matriz compuesta por los vectores propios correspondientes a la matriz de varianzas y covarianzas. En nuestro caso:

$$\begin{pmatrix} -0.089 & -15.8 & -1.38 & -26.6 \\ -0.059 & -6.8 & -0.68 & -10.6 \\ -0.029 & 2.2 & -0.48 & -1.6 \\ 0.131 & 6.2 & 0.22 & 9.4 \\ -0.199 & -10.8 & -0.88 & -19.6 \\ 0.141 & 8.2 & 0.62 & 13.4 \\ 0.081 & 13.2 & 1.72 & 22.4 \\ 0.211 & 9.2 & 1.42 & 17.4 \\ -0.049 & -8.8 & -0.38 & -4.6 \\ -0.139 & 3.2 & -0.18 & 0.4 \end{pmatrix} \cdot \begin{pmatrix} -0.005759331 & -0.01300728 & 0.001249545 & 0.9998980345 \\ -0.517026145 & 0.85544272 & -0.028891125 & 0.0081861926 \\ -0.050711126 & -0.06431056 & -0.996640617 & 0.0001167914 \\ -0.854446708 & -0.51372435 & 0.076623899 & -0.0117001348 \end{pmatrix}$$

$$= \begin{pmatrix} 30.97 & 0.24 & -0.21 & 0.09 \\ 12.61 & -0.33 & 0.06 & 0.01 \\ 0.25 & 2.74 & 0.29 & 0.01 \\ -11.25 & 0.46 & 0.32 & 0.07 \\ 22.38 & 0.89 & -0.31 & -0.06 \\ -15.72 & 0.09 & 0.17 & 0.05 \\ -26.05 & -0.33 & -0.38 & -0.07 \\ -19.70 & -1.16 & -0.35 & 0.08 \\ 8.50 & -5.14 & 0.28 & -0.07 \\ -1.99 & 2.55 & 0.12 & -0.12 \end{pmatrix}$$

- Proyección de la matriz de datos estandarizados asociada a los vectores propios de la matriz de correlaciones:
  Esta proyección la podemos encontrar haciendo el producto $A \cdot u$ donde A es la matriz de datos estandarizados y u es la matriz compuesta por los vectores propios correspondientes a la matriz de correlaciones. En nuestro caso:

$$\begin{pmatrix} -0.70 & -1.69 & -1.46 & -1.75 \\ -0.46 & -0.73 & -0.72 & -0.70 \\ -0.23 & 0.24 & -0.51 & -0.11 \\ 1.03 & 0.66 & 0.23 & 0.62 \\ -1.56 & -1.16 & -0.93 & -1.29 \\ 1.11 & 0.88 & 0.66 & 0.88 \\ 0.64 & 1.42 & 1.82 & 1.48 \\ 1.66 & 0.99 & 1.50 & 1.15 \\ -0.39 & -0.94 & -0.40 & -0.30 \\ -1.09 & 0.34 & -0.19 & 0.03 \end{pmatrix} \cdot \begin{pmatrix} -0.46 & 0.87 & 0.12 & 0.08 \\ -0.50 & -0.39 & 0.61 & 0.46 \\ -0.51 & -0.20 & -0.78 & 0.31 \\ -0.52 & -0.21 & 0.06 & -0.83 \end{pmatrix}$$

$$= \begin{pmatrix} 2.84 & 0.70 & -0.10 & 0.15 \\ 1.31 & 0.17 & 0.01 & -0.02 \\ 0.30 & -0.17 & 0.51 & 0.02 \\ -1.25 & 0.47 & 0.39 & -0.05 \\ 2.46 & -0.46 & -0.25 & 0.11 \\ -1.75 & 0.31 & 0.21 & -0.03 \\ -2.70 & -0.66 & -0.38 & 0.05 \\ -2.63 & 0.53 & -0.30 & 0.11 \\ 1.02 & 0.18 & -0.33 & -0.34 \\ 0.42 & -1.06 & 0.23 & -0.01 \end{pmatrix}$$

Las proyecciones que se obtienen al multiplicar la matriz de datos estandarizadas por los vectores propios de la matriz de correlaciones (Z·V) significan o representan la coordenada o la posición de cada individuo en la componente o eje nuevo que se forma. Así, se ubican los individuos en el plano principal (plano que se forma entre la componente 1 y la componente 2) con las dos coordenadas o valores que arroja este producto para cada individuo.

10. Punto 9

- Gráfico en 3 dimensiones de las variables altura planta (X_1), longitud radicular (X_2) y área foliar (X_3) para la matriz original:

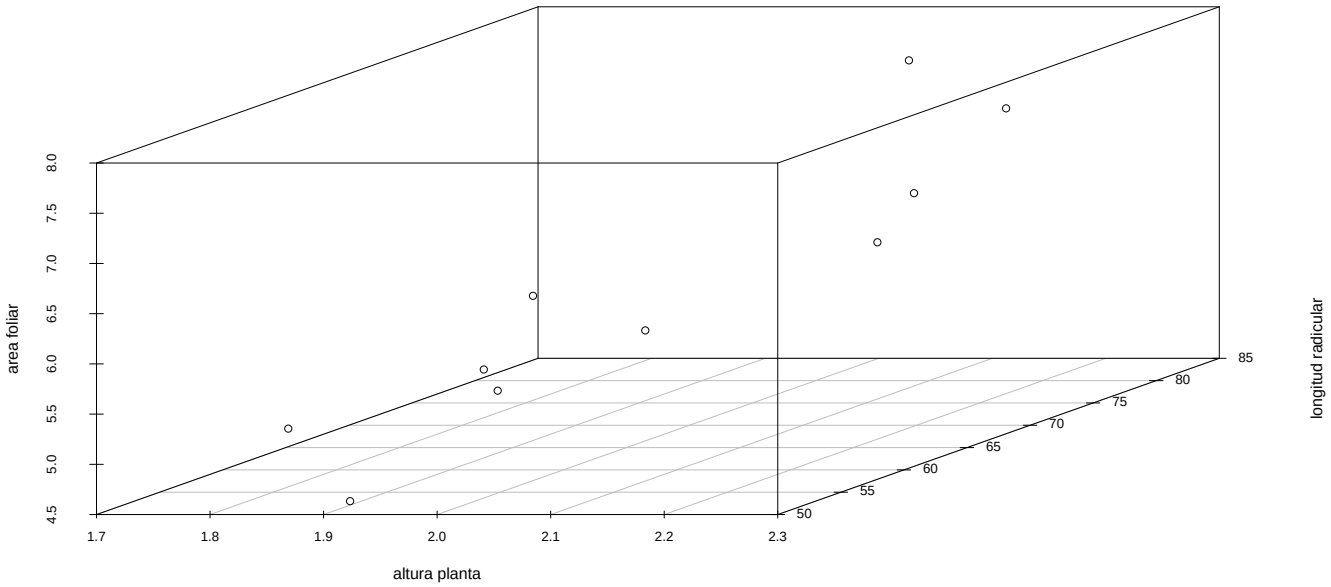

FIGURA 1: Gráfico en 3 dimensiones para la matriz de datos original

- Gráfico en 3 dimensiones de las variables altura planta (X_1), longitud radicular (X_2) y área foliar (X_3) para la matriz estandarizada:

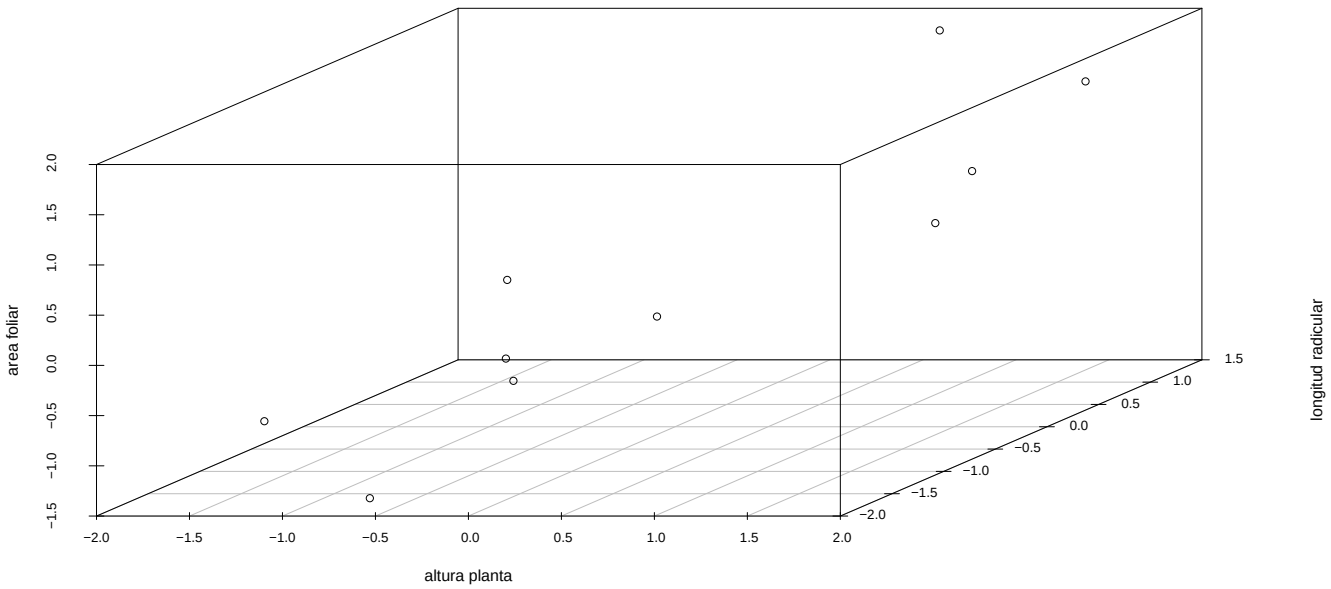

FIGURA 2: Gráfico en 3 dimensiones para la matriz de datos estandarizada

11. Punto 10

- Gráfico en 2 dimensiones de la proyección encontrada de la matriz de datos original y los vectores propios de la matriz de varianzas y covarianzas:

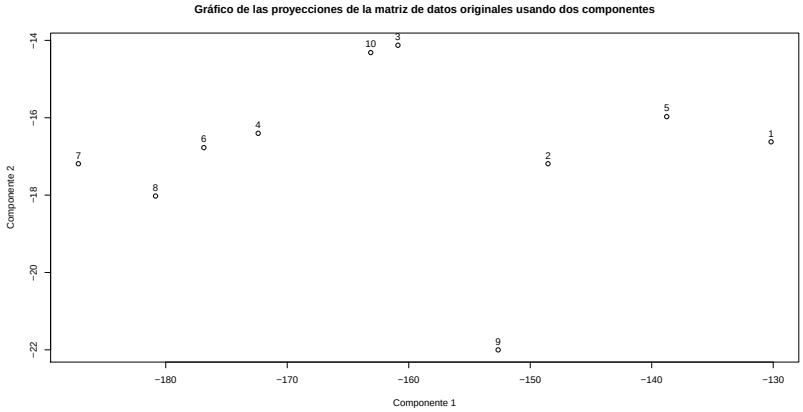

FIGURA 3: Gráfico de la proyección de la matriz de datos original

- Gráfico en 2 dimensiones de la proyección encontrada de la matriz de datos centrada y los vectores propios de la matriz de varianzas y covarianzas:

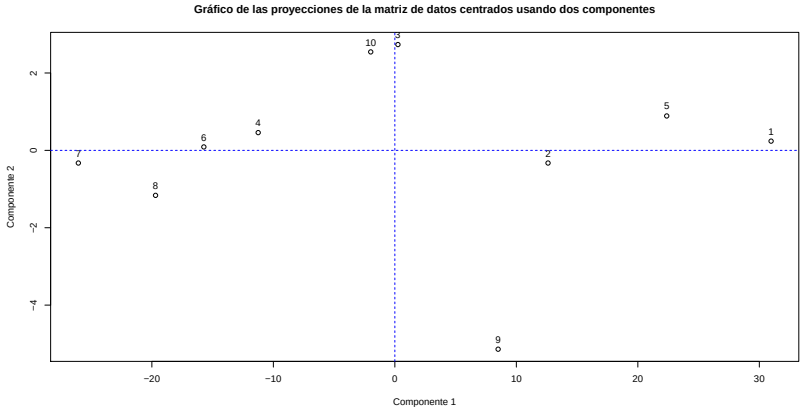

FIGURA 4: Gráfico de la proyección de la matriz de datos centrada

- Gráfico en 2 dimensiones de la proyección encontrada de la matriz de datos estandarizada y los vectores propios de la matriz de correlaciones:

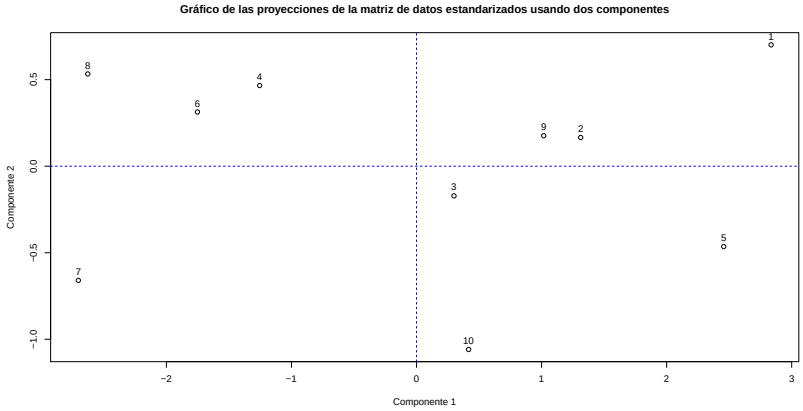

FIGURA 5: Gráfico de la proyección de la matriz de datos estandarizada

- Gráfico en 3 dimensiones (usando las 3 primeras componentes principales) de las proyecciones encontradas de la matriz de datos estandarizada y los vectores propios de la matriz de correlaciones:

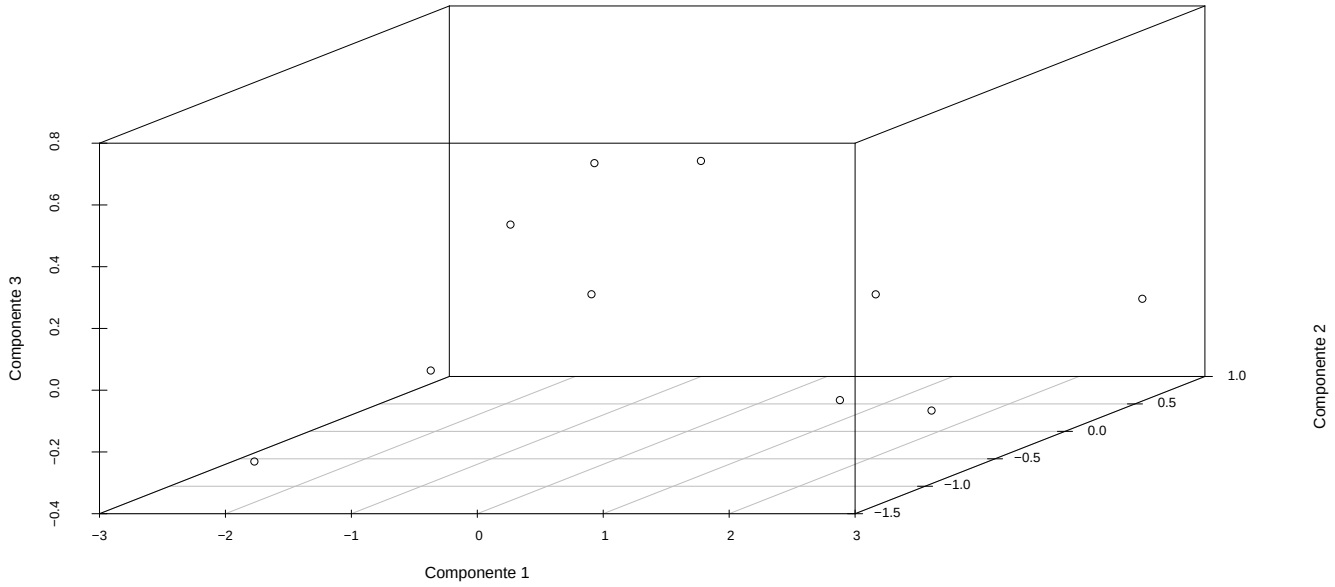

Figura 6: Gráfico de la proyección de la matriz de datos estandarizada con tres componentes principales

12. Conclusiones

Teniendo en cuenta los gráficos realizados con la matriz de datos estandarizados, podemos decir que hay plantas similares a otras en cuanto a las mediciones de las variables, según la figura 5 podemos apreciar que las plantas 9, 2, 3 y también las plantas 6 y 4 tienen características similares.

Evaluando planta por planta, vemos en la figura 5 que la mejor planta es la número 7, ya que tiene valores altos en las variables Longitud radicular, Área foliar y peso en pulpa. Las plantas 4, 6 y 8 también se podrían considerar como buenas, ya que tienen valores altos en la variable Altura de la planta. Las plantas 2,3 y 9 se podría decir que son plantas promedio (están cerca del origen); la planta 10 es mala en cuanto a su rendimiento (tiene valores bajos en casi todas las variables); y las plantas 1 y 5 son las más malas, tienen valores muy bajos en todas sus variables.

Analizando las componentes principales podemos notar que la primera componente tiene valores negativos para todos los coeficientes, lo que nos permite hacer un índice, ya que se da el factor tamaño, el cual puede interpretarse como la variable latente "productividad de la planta".

Referencias

Ligges, U. & Mächler, M. (2003), 'Scatterplot3d - an r package for visualizing multivariate data', *Journal of Statistical Software* **8**(11), 1–20.
*http://www.jstatsoft.org

Zelaya, J. T. (n.d.), *ANÁLISIS MULTIVARIADO DE DATOS*, Universidad de Costa Rica.