Artículo
Revista digital Matemática, Educación e Internet (www.cidse.itcr.ac.cr/revistamate/). Vol. 11, Nº 2. Enero – Agosto 2011.

Implementación del Análisis en Componentes Principales con el software estadístico R

Lic. Juan José Fallas M.
jfallas@itcr.ac.cr
Instituto Tecnológico de Costa Rica

Lic. Jeffry Chavarría M.
jchavarria@itcr.ac.cr
Instituto Tecnológico de Costa Rica

Resumen.

El Análisis en Componentes Principales (ACP) constituye la técnica base para el Análisis Multivariado de Datos. Su objetivo principal es reducir la cantidad de variables, manteniendo la máxima cantidad de información, presente en una tabla de datos de variables cuantitativas. En el presente artículo se expone un panorama general sobre la estructura que fundamenta un ACP y se implementa un caso concreto en el software estadístico R. Para ello es necesario un conocimiento básico de este software..

Palabras claves: Componentes principales, ACP, nube de puntos, inercia, software R..

Abstract.

Principal Component Analysis (PCA) represents the basic technique in Multivariate Data Analysis. PCA is a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. This article explains the general procedure to perform a PCA by showing an example using the statistical package R. It is required a basic knowledge of this software.

KeyWords: Principal components, PCA, inertia, software R.

1.1 Introducción.

El Análisis en Componentes Principales (ACP) constituye la técnica base en el Análisis Multivariado de Datos. Su objetivo principal es reducir la cantidad de variables, manteniendo la máxima cantidad de información presente en una tabla de datos de variables cuantitativas, mediante la construcción de nuevas variables denominadas *componentes principales* que contienen, en gran medida, la información de las variables originales. Entre otras cosas, por medio de un ACP se busca determinar las principales relaciones que existen entre los individuos, entre las variables y entre los individuos y las variables.

A pesar de que hoy en día existen muchos paquetes estadísticos que realizan el ACP y las demás técnicas multivariadas, se tiene la desventaja de que la poca intervención por parte del usuario desvirtúa la esencia del ACP, pues se enfatiza en el fin y se dejan de lado las ideas que sustentan esta técnica. En muchos casos

la aplicación del ACP se reduce a la interpretación de los gráficos generados por estos paquetes, obviando la importancia de tener como mínimo un panorama general sobre la estructura teórica que fundamenta un ACP. Por este motivo, en este artículo se abordan las ideas principales de esta técnica y se implementan en el software estadístico R.

1.2 El ACP

A partir de n individuos $x_1,\ldots,x_n$ sobre los cuales se han observado o medido p variables cuantitativas denotadas $\mathrm{x}^1,\ldots,\mathrm{x}^p$ se construye la tabla de datos:

	x^1	x^2	$\ldots$	x^p
x_1	x_1^1	x_1^2	$\ldots$	x_1^p
x_2	x_2^1	$\ddots$		
$\vdots$	$\vdots$			
x_n	x_n^1			x_n^p

Tabla 1.1 Tabla de datos.

la cual se denota X. Cada individuo se visualiza como un vector de dimensión p. Para el primer individuo de la tabla X se tiene el vector $\left(x_1^1,x_1^2,\ldots,x_1^p\right)$, donde x_1^i corresponde al valor observado de este individuo en la variable cuantitativa x^i. De esta manera, a los individuos se les visualiza como vectores en el espacio vectorial $\mathbb{R}^p$ y, por lo tanto, a este espacio se le denomina el *espacio de los individuos*. Análogamente, las variables se identifican como vectores del espacio vectorial $\mathbb{R}^n$ y, por lo tanto, a éste se le denomina el *espacio de las variables*.

Para efectos de simplificar cálculos, como etapa preliminar de un ACP se procede a centrar y estandarizar las variables, esto es que todas las variables tengan media cero y varianza 1. Entiéndase, para la variable cuantitativa x^j, restar a cada entrada del vector correspondiente la media $\overline{\mathrm{x}^j}$ y dividir entre la desviación σ_{x^j}. Con esta consideración, el ACP se denomina *ACP normado* y tiene la particularidad de que todas las variables tienen el mismo peso o ponderación en el estudio[1].

Gráficamente los individuos son puntos en el espacio $\mathbb{R}^p$, que definen una nube de n puntos. Para el caso $n=7$ y $p=3$ la nube de puntos se visualiza como se muestra en la figura 1.

[1]En caso que se quiera que las variables tengan diferentes pesos, se debe recurrir al ACP no normado. Sin embargo, este enfoque no corresponde al objetivo de este artículo, pues implica detallar más ciertos aspectos teóricos del ACP. En particular, se debe conocer el concepto de métrica que se emplea sobre los espacios vectoriales $\mathbb{R}^n$ y $\mathbb{R}^p$. Luego, manipulando la métrica definida sobre el espacio de las variables se logra incorporar diferentes pesos sobre cada una de ellas.

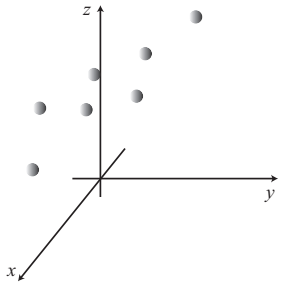

Figura 1.1 Representación en $\mathbb{R}^3$ de la nube de puntos.

La idea del ACP es proyectar ortogonalmente los puntos que conforman la nube en un espacio de dimensión menor, de manera que la pérdida de información sea mínima. A continuación se muestra el espacio de proyección óptimo de dimensión dos[2] correspondiente a la nube de puntos de la figura 1.

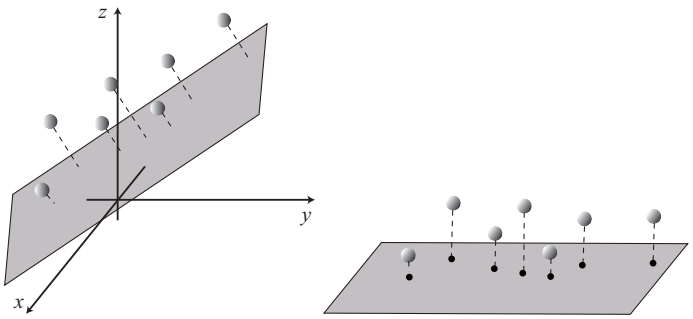

Figura 1.2 Proyección ortogonal óptima.

Así, en lugar de analizar la nube como un todo, se analiza la información presente en los puntos generados en el espacio sobre el que se está proyectando (puntos en negro en la figura 2).

En términos gráficos, este efecto en el agrupamiento de las proyecciones se mide a través de la *inercia de la nube de puntos*. Intuitivamente, la inercia es una medida de la dispersión de la nube alrededor de un punto denominado *centro de gravedad*. Este punto en un ACP normado corresponde al vector nulo. Entre más distanciadas queden las proyecciones entre sí, se dice que la inercia es mayor y entre más inercia exista a la hora de proyectar menos es la pérdida de información. Por esta razón, en la construcción del ACP éste se plantea como un problema de optimización en el que se busca un espacio de dimensión menor sobre el cual proyectar y en el que la *inercia sea máxima*.

Por ejemplo, para el caso planteado en las figuras 1 y 2, si se realiza la proyección sobre otro espacio, como el que se muestra en la figura 3, los puntos proyectados se agrupan más. Este fenómeno provoca que haya pérdida de información, lo cual se pudo evitar con una proyección adecuada de la nube de puntos.

[2]En este caso corresponde a un plano, pero en la mayoría de los estudios en los que se aplica el ACP la dimensión del espacio es muy grande, de manera que es imposible representar gráficamente la nube de puntos.

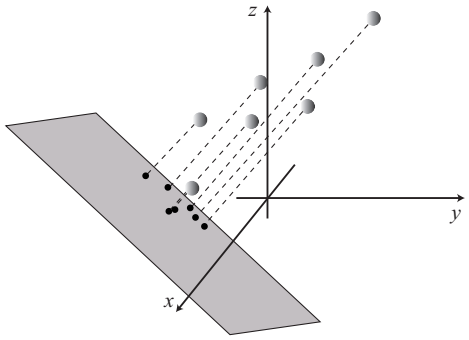

Figura 1.3 Proyección ortogonal no óptima.

Como solución al problema de optimización que se genera en el ACP se determina que los vectores propios de la matriz de correlaciones $\Re$, calculada a partir de la tabla de datos X, son la clave para construir el espacio óptimo de proyección. En este sentido, la matriz de correlaciones es:

$$\Re = \begin{pmatrix} 1 & r\left(x^1,x^2\right) & \dots & r\left(x^1,x^p\right) \\ r\left(x^1,x^2\right) & 1 & & r\left(x^1,x^p\right) \\ \vdots & \vdots & 1 & \vdots \\ r\left(x^1,x^p\right) & r\left(x^2,x^p\right) & \dots & 1 \end{pmatrix}$$

donde $r\left(x^i,x^j\right)$ es la correlación entre x^i y x^j, y se define como:

$$r\left(x^i,x^j\right) = \frac{\operatorname{cov}\left(x^i,x^j\right)}{\sigma_{x^i} \cdot \sigma_{x^j}}$$

con $\operatorname{cov}\left(x^i,x^j\right)$ la covarianza entre x^i y x^j, y σ_{x^i} la desviación estándar de x^i. Si p_i corresponde a los pesos o ponderaciones de los individuos, entonces la covarianza y la desviación estándar están respectivamente dadas por:

$$\operatorname{cov}\left(x^i,x^j\right) = \sum_{k=1}^{n} p_i \left(x_k^i - \overline{x^i}\right)\left(x_k^j - \overline{x^j}\right) \quad \text{y} \quad \sigma_{x^i} = \sqrt{\sum_{i=1}^{n} p_i \left(x_k^i - \overline{x^i}\right)^2}$$

Al diagonalizar la matriz $\Re$ de dimensión $p \times p$ se obtienen p valores propios $\lambda_1, \dots, \lambda_p$ tales que

$$\lambda_1 \geq \dots \geq \lambda_p \geq 0$$

y p vectores propios $u_1, \dots, u_p$ que son ortogonales dos a dos. A partir de cada vector propio u_i se genera un subespacio de dimensión uno (una recta) denotado $\triangle_{u_i}$. Estos subespacios satisfacen que $\triangle_{u_i} \cap \triangle_{u_j} = \{\mathbf{0}\}$ para $i \neq j$.

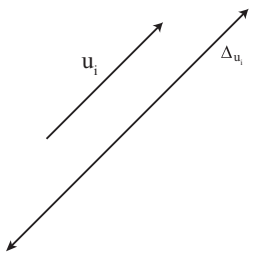

Figura 1.4 Espacio generado.

Luego, la suma directa[3] de estos subespacios genera el espacio de proyección óptima E, en el cual la inercia es máxima.

$$E = \triangle_{u_1} \oplus \triangle_{u_2} \oplus \ldots \oplus \triangle_{u_p}$$

Además, cada suma directa $\triangle_{u_i} \oplus \triangle_{u_j}$, $i \neq j$, genera el *plano principal* $i \times j$. Estos planos serán fundamentales para proyectar los individuos y las variables, y así analizar las relaciones que se generan entre ellos. A cada subespacio $\triangle_{u_i}$ se le denomina *eje principal.*

Los valores propios de $\Re$ tienen la característica de que su suma corresponde a la inercia de la nube de puntos (denotada I), y a su vez coincide con p, la cantidad de variables en el estudio, esto es:

$$I = \lambda_1 + \ldots + \lambda_p = p \tag{1.1}$$

A su vez, se define el porcentaje de inercia explicada por el eje principal $\triangle_{u_i}$ como

$$\left(\frac{\lambda_j}{I} \cdot 100\right) \% \tag{1.2}$$

Este porcentaje indica qué tanta información explica un eje principal sobre la nube de puntos. De esta manera, los ejes principales generados por los vectores propios asociados a los valores propios más pequeños son los que acumulan menos inercia y, por lo tanto, acumulan poca información sobre los individuos y las variables. Esto tiene como consecuencia que los últimos ejes y planos principales carezcan de importancia para el estudio.

Finalmente, se construyen las *componentes principales* que corresponden a nuevas variables $\mathrm{c}^1, \ldots, \mathrm{c}^q$ que sintetizan la información presente en las variables originales $\mathrm{x}^1, \ldots, \mathrm{x}^p$. La k-ésima componente principal c^k se define como:

$$\mathrm{c}^k = X_{CE} \cdot u_k \tag{1.3}$$

donde X_{CE} corresponde a la tabla de datos X centrada y estandarizada y u_k es el k-ésimo vector propio de la matriz R.

En general, como se mencionó para los ejes y los planos principales, las primeras componentes principales acumulan la mayoría de la información sobre las variables originales. Por esta razón, las últimas componentes siempre se descartan y así se produce una reducción en la dimensión del problema. Además, por medio de las componentes principales se construyen las coordenadas de las proyecciones de los individuos y las variables en los diferentes planos principales.

A continuación se analiza con detalle un caso en el que se aplica el Análisis en Componentes Principales, apoyado con el software estadístico R. En este ejemplo se abordan algunos otros conceptos que complementan el análisis.

1.2.1 Aplicación del ACP.

A continuación se incluye una tabla que muestra el promedio en quices y las notas obtenidas en los cuatro exámenes parciales por 10 estudiantes del curso Matemática General del Instituto Tecnológico de Costa Rica, durante el primer semestre del 2008. Los 10 estudiantes fueron tomados aleatoriamente del grupo de estudiantes que asistieron a los cuatro parciales.

[3]Dados dos espacios vectoriales U y V tales que $U \cap V = \{\mathbf{0}\}$ entonces la suma directa de U y V se denota $U \oplus V$ y se define como $U \oplus V = \{a + b \text{ t.q. } a \in U \text{ y } b \in V\}$

	PQ	P1	P2	P3	P4
A01	63.30	95.17	90.00	90.52	44.45
A02	89.60	90.33	92.50	94.83	85.19
A03	57.20	59.68	42.50	15.52	12.04
A04	71.20	83.88	82.50	67.25	64.82
A05	78.20	85.49	46.67	58.97	85.19
A06	75.70	93.55	90.00	56.04	81.49
A07	79.80	91.94	92.50	87.59	88.89
A08	65.60	59.68	56.67	37.07	25.93
A09	93.50	91.94	95.00	87.94	77.78
A10	67.10	59.68	83.34	65.18	50.00

Tabla 1.2 Tabla de datos.

Así, la matriz de datos X contempla $m = 10$ individuos y $n = 5$ variables.

$$X = \begin{pmatrix} 63.30 & 95.17 & 90.00 & 90.52 & 44.45 \\ 89.60 & 90.33 & 92.50 & 94.83 & 85.19 \\ 57.20 & 59.68 & 42.50 & 15.52 & 12.04 \\ 71.20 & 83.88 & 82.50 & 67.25 & 64.82 \\ 78.20 & 85.49 & 46.67 & 58.97 & 85.19 \\ 75.70 & 93.55 & 90.00 & 56.04 & 81.49 \\ 79.80 & 91.94 & 92.50 & 87.59 & 88.89 \\ 65.60 & 59.68 & 56.67 & 37.07 & 25.93 \\ 93.50 & 91.94 & 95.00 & 87.94 & 77.78 \\ 67.10 & 59.68 & 83.34 & 65.18 & 50.00 \end{pmatrix}$$

1.2.1.1 Centrado y estandarizado. Primero se procede a cargar la tabla 2 y posteriormente se realiza el proceso de centrado y estandarizado, tal y como se muestra en las figuras 5 y 6.

```
> Tabla <-read.table("C:/Users/Juan José/Desktop/GuiaDidactica/Notas.txt",header=TRUE)
> Tabla
   ALU   PQ    P1    P2    P3    P4
1  A01 63.3 95.17 90.00 90.52 44.45
2  A02 89.6 90.33 92.50 94.83 85.19
3  A03 57.2 59.68 42.50 15.52 12.04
4  A04 71.2 83.88 82.50 67.25 64.82
5  A05 78.2 85.49 46.67 58.97 85.19
6  A06 75.7 93.55 90.00 56.04 81.49
7  A07 79.8 91.94 92.50 87.59 88.89
8  A08 65.6 59.68 56.67 37.07 25.93
9  A09 93.5 91.94 95.00 87.94 77.78
10 A10 67.1 59.68 83.34 65.18 50.00
>
```

Figura 1.5 Tabla de datos.

```
> TablaCentradaEstandarizada<-scale(Tabla[,2:6])
> TablaCentradaEstandarizada
              PQ         P1         P2          P3         P4
 [1,] -0.9368927  0.9240324  0.6298027  0.95283719 -0.6285669
 [2,]  1.3403973  0.6054005  0.7525043  1.12094593  0.8665180
 [3,] -1.4650855 -1.4123818 -1.7015276 -1.97248884 -1.8179557
 [4,] -0.2528398  0.1807775  0.2616980  0.04520604  0.1189756
 [5,]  0.3532830  0.2867687 -1.4968613 -0.27774996  0.8665180
 [6,]  0.1368106  0.8173829  0.6298027 -0.39203269  0.7307347
 [7,]  0.4918254  0.7113917  0.7525043  0.83855446  1.0023014
 [8,] -0.7377381 -1.4123818 -1.0060549 -1.13194516 -1.3082177
 [9,]  1.6780944  0.7113917  0.8752059  0.85220598  0.5945843
[10,] -0.6078546 -1.4123818  0.3029257 -0.03553296 -0.4248918
attr(,"scaled:center")
    PQ     P1     P2     P3     P4
74.120 81.134 77.168 66.091 61.578
attr(,"scaled:scale")
      PQ       P1       P2       P3       P4
11.54881 15.18994 20.37463 25.63817 27.24929
> |
```

Figura 1.6 Tabla centrada y estandarizada.

Como se muestra en la figura 6, además de la tabla centrada y estandarizada, R muestra los promedios y las desviaciones de los datos de cada una de las columnas.

1.2.1.2 La matriz de correlaciones. Como en el ACP normado se debe diagonalizar la matriz de correlaciones de la *TablaCentradaEstandarizada,* entonces se procede a realizarlo en R con el comando *cor,* tal y como se muestra en la figura 7.

```
> MatrizCorrelaciones <-cor(TablaCentradaEstandarizada)
> MatrizCorrelaciones
          PQ        P1        P2        P3        P4
PQ 1.0000000 0.6251779 0.5332789 0.6706295 0.8337751
P1 0.6251779 1.0000000 0.6360671 0.7560867 0.7736149
P2 0.5332789 0.6360671 1.0000000 0.8317164 0.5562036
P3 0.6706295 0.7560867 0.8317164 1.0000000 0.6977386
P4 0.8337751 0.7736149 0.5562036 0.6977386 1.0000000
> |
```

Figura 1.7 Cálculo de la matriz de correlaciones.

1.2.1.3 Valores y vectores propios. Posteriormente, se buscan los valores y vectores propios de la matriz de correlaciones con el comando:

```
eigen(MatrizCorrelaciones)
```

los cuales se extraen mediante las instrucciones:

```
valorespropios <-eigen(MatrizCorrelaciones)$values

vectorespropios <-eigen(MatrizCorrelaciones)$vectors
```

Así, se tienen los siguientes resultados:

```
> valorespropios<-eigen(MatrizCorrelaciones)$values
> valorespropios
[1] 3.7713036 0.6256033 0.3424250 0.1434121 0.1172560
> vectorespropios<-eigen(MatrizCorrelaciones)$vectors
> vectorespropios
           [,1]        [,2]        [,3]       [,4]       [,5]
[1,] -0.4343801  0.50490905 -0.52820680 -0.1812110  0.4945108
[2,] -0.4507848  0.01258555  0.78869866 -0.0518923  0.4146038
[3,] -0.4202008 -0.64450661 -0.28844598  0.5410692  0.1791243
[4,] -0.4698121 -0.34835460 -0.09288376 -0.6934827 -0.4103412
[5,] -0.4591458  0.45625514  0.08440094  0.4368015 -0.6189485
>
```

Figura 1.8 Cálculo de los vectores y valores propios.

Por lo tanto, los valores propios están dados por:

$$\lambda_1 = 3.7713036, \lambda_2 = 0.6256033, \lambda_3 = 0.342425, \lambda_4 = 0.1434121 \text{ y } \lambda_5 = 0.117256$$

y a su vez los dos primeros vectores propios son:

$$u_1 = \begin{pmatrix} -0.4343801 \\ -0.4507848 \\ -0.4202008 \\ -0.4698121 \\ -0.4591458 \end{pmatrix} \text{ y } u_2 = \begin{pmatrix} 0.50490905 \\ 0.01258555 \\ -0.64450661 \\ -0.34835460 \\ 0.45625514 \end{pmatrix}$$

1.2.1.4 Análisis de la inercia. En el caso en estudio, de acuerdo con (1) se tiene que la inercia de la nube de puntos es

$$I(N) = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = 5$$

Así, por (2) se tiene que el porcentaje de inercia explicada por cada uno de los ejes principales es:

$$\text{Eje } 1 \rightarrow \frac{\lambda_1}{5} = \frac{3.7713036}{5} = 0.75426 \rightarrow 75.43\%$$

$$\text{Eje } 2 \rightarrow \frac{\lambda_2}{5} = \frac{0.6256033}{5} = 0.12512 \rightarrow 12.51\%$$

$$\text{Eje } 3 \rightarrow \frac{\lambda_3}{5} = \frac{0.342425}{5} = 0.068484 \rightarrow 6.85\%$$

$$\text{Eje } 4 \rightarrow \frac{\lambda_4}{5} = \frac{0.1434121}{5} = 0.02868 \rightarrow 2.87\%$$

$$\text{Eje } 5 \rightarrow \frac{\lambda_5}{5} = \frac{0.117256}{5} = 0.023451 \rightarrow 2.34\%$$

La información anterior se resume en la tabla 3.

	Valores propios	% var explicada	% varianza acumulada
PQ	3.7713036	75.43	75.43
P1	0.6256033	12.51	87.94
P2	0.342425	6.85	94.79
P3	0.1434121	2.87	97.66
P4	0.117256	2.34	100

Tabla 1.3 Tabla resumen.

A pesar de que el problema original se ubica en el espacio vectorial $\mathbb{R}^5$, note que los dos primeros vectores propios, u_1 y u_2, asociados a los valores propios λ_1 y λ_2, generan dos subespacios cuya suma directa corresponde al primer plano principal (espacio de dimensión dos) que conserva un 87.94% de la información original de la nube de puntos. Desde otro punto de vista, si se trabaja en el primer plano principal existe una pérdida del 12.06% de la información, lo cual es aceptable considerando que el problema se está reduciendo de cinco a dos dimensiones.

Finalmente, se hace énfasis en que como el análisis se restringe al primer plano principal (al menos en este caso), entonces basta calcular las dos primeras componentes principales.

1.2.1.5 Componentes principales. Como el ACP es normado, de (3) se sigue que para el cálculo de las dos primeras componentes principales se deben realizar los productos matriciales de la matriz de datos centrada y estandarizada (X_{CE}) por los dos primeros vectores propios. Estos cálculos se muestran a continuación:

$$c^1 = \begin{pmatrix} -0.9368 & 0.9240 & 0.6298 & 0.9528 & -0.6285 \\ 1.3403 & 0.6054 & 0.7525 & 1.1209 & 0.8665 \\ -1.4650 & -1.4123 & -1.7015 & -1.9724 & -1.8179 \\ -0.2528 & 0.1807 & 0.2616 & 0.0452 & 0.1189 \\ 0.3532 & 0.2867 & -1.4968 & -0.2777 & 0.8665 \\ 0.1368 & 0.8173 & 0.6298 & -0.3920 & 0.7307 \\ 0.4918 & 0.7113 & 0.7525 & 0.8385 & 1.0023 \\ -0.7377 & -1.4123 & -1.0060 & -1.1319 & -1.3082 \\ 1.6780 & 0.7113 & 0.8752 & 0.8522 & 0.5945 \\ -0.6078 & -1.4123 & 0.3029 & -0.0355 & -0.4248 \end{pmatrix} \begin{pmatrix} -0.4343801 \\ -0.4507848 \\ -0.4202008 \\ -0.4698121 \\ -0.4591458 \end{pmatrix} = \begin{pmatrix} -0.43328 \\ -2.0958 \\ 3.7495 \\ -0.15752 \\ 0.078908 \\ -0.84387 \\ -1.7047 \\ 2.5124 \\ -2.0908 \\ 0.98521 \end{pmatrix}$$

$$c^2 = \begin{pmatrix} -0.9368 & 0.9240 & 0.6298 & 0.9528 & -0.6285 \\ 1.3403 & 0.6054 & 0.7525 & 1.1209 & 0.8665 \\ -1.4650 & -1.4123 & -1.7015 & -1.9724 & -1.8179 \\ -0.2528 & 0.1807 & 0.2616 & 0.0452 & 0.1189 \\ 0.3532 & 0.2867 & -1.4968 & -0.2777 & 0.8665 \\ 0.1368 & 0.8173 & 0.6298 & -0.3920 & 0.7307 \\ 0.4918 & 0.7113 & 0.7525 & 0.8385 & 1.0023 \\ -0.7377 & -1.4123 & -1.0060 & -1.1319 & -1.3082 \\ 1.6780 & 0.7113 & 0.8752 & 0.8522 & 0.5945 \\ -0.6078 & -1.4123 & 0.3029 & -0.0355 & -0.4248 \end{pmatrix} \begin{pmatrix} 0.50490905 \\ 0.01258555 \\ -0.64450661 \\ -0.34835460 \\ 0.45625514 \end{pmatrix} = \begin{pmatrix} -1.4861 \\ 0.20427 \\ 0.19681 \\ -0.25554 \\ 1.6388 \\ 0.14342 \\ -0.062519 \\ 0.055582 \\ 0.26656 \\ -0.70141 \end{pmatrix}$$

Para realizarlo en R se carga la variable *MatrizX* que corresponde a la información de la variable *TablaCentradaEstandarizada* (omitiendo la información de los promedios y las desviaciones), tal y como se visualiza en la figura 1.9.

```
> MatrizX<-TablaCentradaEstandarizada[1:10,]
> MatrizX
               PQ         P1         P2          P3         P4
 [1,] -0.9368927  0.9240324  0.6298027  0.95283719 -0.6285669
 [2,]  1.3403973  0.6054005  0.7525043  1.12094593  0.8665180
 [3,] -1.4650855 -1.4123818 -1.7015276 -1.97248884 -1.8179557
 [4,] -0.2528398  0.1807775  0.2616980  0.04520604  0.1189756
 [5,]  0.3532830  0.2867687 -1.4968613 -0.27774996  0.8665180
 [6,]  0.1368106  0.8173829  0.6298027 -0.39203269  0.7307347
 [7,]  0.4918254  0.7113917  0.7525043  0.83855446  1.0023014
 [8,] -0.7377381 -1.4123818 -1.0060549 -1.13194516 -1.3082177
 [9,]  1.6780944  0.7113917  0.8752059  0.85220598  0.5945843
[10,] -0.6078546 -1.4123818  0.3029257 -0.03553296 -0.4248918
> |
```

Figura 1.9 Definición de la variable $MatrizX$.

Posteriormente se procede al cálculo de las componentes principales mediante un producto matricial, tal y como se muestra en la figura 1.10.

```
> componentes<-MatrizX %*% vectorespropios[,1:5]
> componentes
            [,1]        [,2]       [,3]         [,4]        [,5]
 [1,] -0.4332665 -1.48604028  0.9004374 -0.472743157  0.03067890
 [2,] -2.0958423  0.20427067 -0.4785680 -0.266013657  0.05233374
 [3,]  3.7494733  0.19680915  0.1804983 -0.008061755  0.31975015
 [4,] -0.1574943 -0.25551650  0.2064877  0.198652211 -0.09539418
 [5,]  0.0788845  1.63883073  0.5702646 -0.317694190 -0.39688441
 [6,] -0.8438672  0.14341992  0.4888286  0.864613351  0.22793735
 [7,] -1.7046922 -0.06252281  0.0909384  0.137400771 -0.29151597
 [8,]  2.5123482  0.05558008 -0.4393483 -0.123814416  0.14359598
 [9,] -2.0907552  0.26657456 -0.6067294 -0.198732680  0.56383980
[10,]  0.9852116 -0.70140552 -0.9128092  0.186393522 -0.55434135
> |
```

Figura 1.10 Cálculo de las componentes principales.

En la figura 10 se visualizan las cinco componentes principales que corresponden a vectores de dimensión 10×1.

1.2.1.6 Círculo de correlaciones y comunalidades. El círculo de correlaciones corresponde a la representación gráfica de la relación entre las variables. Consiste en un círculo unitario dentro del cual se irán representando una a una todas las variables. Las coordenadas para representar la variable x^k se determinan calculando la correlación entre esta variable y las componentes principales c^1 y c^2. Por ejemplo, el par ordenado con el que se representa la variable PQ en el círculo de correlaciones es[4]:

$$\left(r\left(\mathrm{PQ}, c^1\right), r\left(\mathrm{PQ}, c^2\right)\right)$$

En el caso de R, la instrucción

$$\texttt{CorreVarOriCompPrin} < -\texttt{cor}(\texttt{cbind}(\texttt{MatrizX}, \texttt{componentes}))[1:5, 6:7]$$

genera las correlaciones entre las dos primeras componentes principales y las variables originales, y la almacena en la variable *CorreVarOriCompPrin*, tal y como se muestra en la figura 11.

[4]Este cálculo se debe realizar para las demás variables.

```
> CorreVarOriCompPrin<-cor(cbind(MatrizX,TodasComponentes))[1:5,6:7]
> CorreVarOriCompPrin

PQ -0.8435594  0.39935828
P1 -0.8754170  0.00995455
P2 -0.8160235 -0.50977309
P3 -0.9123679 -0.27553139
P4 -0.8916540  0.36087542
>
```

Figura 1.11 Correlaciones entre las componentes y las variables originales.

Por lo tanto, para representar las variables en el círculo de correlaciones basta ubicar la primera columna en el eje de las abscisas y la segunda columna en el eje de las ordenadas. Este proceso se puede realizar en R directamente a través del paquete **ade4**[5]**.** En este caso se utiliza la instrucción:

```
s.corcircle(CorreVarOriCompPrin)
```

Así, se tienen las variables representadas en el círculo de correlaciones, tal y como se muestra en la figura 1.12:

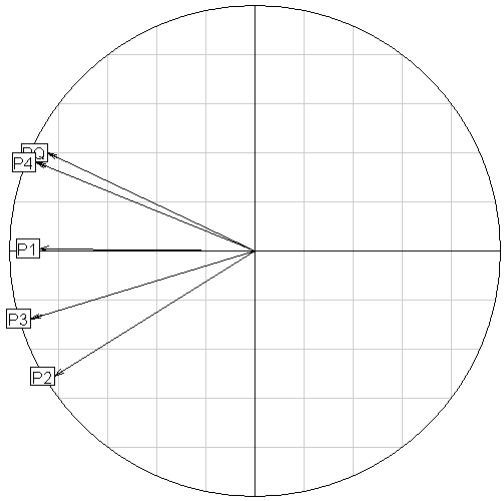

Figura 1.12 Círculo de correlaciones.

Por su parte, las *comunalidades* corresponden a una medida de la calidad de la representación de las variables en el círculo de correlaciones. Gráficamente se determina que las variables cuyo par ordenado se ubica cerca de la circunferencia están bien representadas y, por lo tanto, se pueden establecer conclusiones confiables sobre ellas. Además, entre más cerca del origen se ubique el par ordenado, se tiene menor calidad en la representación. Con este criterio, aplicado al círculo de correlaciones de la figura 12, se establece que todas las variables están bien representadas. En el caso de la variable P1 la calidad es ligeramente menor.

Numéricamente las comunalidades corresponden a la suma de los cuadrados de las correlaciones de las componentes principales con las variables originales. Por lo tanto, en el caso propuesto la información necesaria para el cálculo de las comunalidades se encuentra en la variable *CorreVarOriCompPrin*.

De esta forma, se construya la tabla 4 con las comunalidades correspondientes.

[5]En caso que el paquete no esté disponible debe utilizarse la opción *Instalar paquetes* del menú *Paquetes* para descargarlo de Internet.

	c^1	c^2	Comunalidades para el primer plano principal
PQ	−0.8435594	0.39935828	$r^2\left(PQ,c^1\right)+r^2\left(PQ,c^2\right)=0.87108\rightarrow 87.1\%$
P1	−0.8754170	0.00995455	$r^2\left(P1,c^1\right)+r^2\left(P1,c^2\right)=0.76645\rightarrow 76.6\%$
P2	−0.8160235	−0.50977309	$r^2\left(P2,c^1\right)+r^2\left(P2,c^2\right)=0.92576\rightarrow 92.5\%$
P3	−0.9123679	−0.27553139	$r^2\left(P3,c^1\right)+r^2\left(P3,c^2\right)=0.90833\rightarrow 90.8\%$
P4	−0.8916540	0.36087542	$r^2\left(P4,c^1\right)+r^2\left(P4,c^2\right)=0.92528\rightarrow 92.5\%$

Tabla 1.4 Cálculo de las comunalidades.

Basta ver los porcentajes correspondientes a las comunalidades para verificar la conclusión a la que se llegó anteriormente con el análisis gráfico del círculo de correlaciones.

1.2.1.7 Representación de los individuos. De manera análoga a como se representaron las variables en el círculo de correlaciones, los individuos pueden ser representados en un plano. En este caso, las coordenadas de los individuos son las componentes principales. Por ejemplo, el par ordenado correspondiente al individuo A01 es $(-0.43,-1.48)$, para el segundo individuo $(-2.09,0.20)$ y así sucesivamente.

La siguiente instrucción toma las componentes principales y grafica los individuos en el primer plano principal.

```
s.label(componentes,label=Tabla[,1],sub="Plano principal",possub="bottomleft")
```

En este punto, se tienen los individuos representados en el primer plano principal, tal y como se muestra en la figura 13.

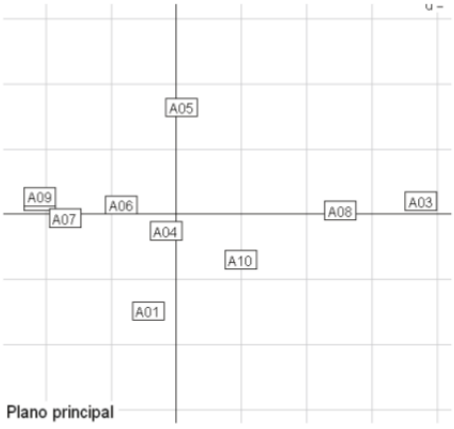

Figura 1.13 Primer plano principal.

La cercanía de los individuos A4, A06 y A10 al origen de coordenadas sugiere una mala representación de estos individuos en el plano.

1.2.1.8 Cosenos cuadrados. Los cosenos cuadrados son una medida de la calidad de la representación de los individuos sobre cada uno de los planos principales. Se denota con $\cos_k^2 1$ la calidad de la representación del k-ésimo individuo sobre el primer eje principal y $\cos_k^2 2$ la calidad de la representación del k-ésimo individuo sobre el segundo eje principal. La suma

$$\cos_k^2 1+\cos_k^2 2$$

mide la calidad de la representación del $k-$ésimo individuo sobre el primer plano principal.

En general, los cosenos cuadrados se toman en cuenta en el análisis para saber sobre cuales individuos se pueden obtener conclusiones confiables. Para ello, normalmente se aplica que un eje tendrá mucha relación con aquellos individuos cuyo coseno cuadrado sea superior o igual a un 50%. Estos individuos están particularmente bien representados sobre ese eje.

Los cosenos cuadrados se calculan a partir de las componentes principales, tal y como se muestra a continuación (se muestra únicamente el cálculo de las tres primeras entradas del primero y del segundo coseno cuadrado):

$$\cos_1^2 1 = \frac{(-0.4332665)^2}{(-0.4332665)^2+(-1.48604028)^2+(0.9004374)^2+(-0.472743157)^2+(0.03067890)^2} \approx 0.054709$$

$$\cos_2^2 1 = \frac{(-2.0958423)^2}{(-2.0958423)^2+(0.20427067)^2+(-0.4785680)^2+(-0.266013657)^2+(0.05233374)^2} \approx 0.927323$$

$$\cos_3^2 1 = \frac{(3.7494733)^2}{(3.7494733)^2+(0.19680915)^2+(0.1804983)^2+(-0.008061755)^2+(0.31975015)^2} \approx 0.987801$$

Las tres primeras entradas del segundo coseno cuadrado se calculan de la siguiente manera:

$$\cos_1^2 1 = \frac{(-1.48604028)^2}{(-0.4332665)^2+(-1.48604028)^2+(0.9004374)^2+(-0.472743157)^2+(0.03067890)^2} \approx 0.643589$$

$$\cos_2^2 1 = \frac{(0.20427067)^2}{(-2.0958423)^2+(0.20427067)^2+(-0.4785680)^2+(-0.266013657)^2+(0.05233374)^2} \approx 0.008809$$

$$\cos_3^2 1 = \frac{(0.19680915)^2}{(3.7494733)^2+(0.19680915)^2+(0.1804983)^2+(-0.008061755)^2+(0.31975015)^2} \approx 0.0027216$$

A partir de los cosenos cuadrados se determina la calidad del primer plano principal, tal y como se muestra en la tabla 5.
La última columna muestra que los individuos A04, A06 y A10 están mal representados en el primer plano principal y, por lo tanto, es muy poca la información que se puede deducir sobre ellos (como consecuencia de la pérdida de información). En estos casos, los individuos deben ser analizados directamente en la tabla original de datos.

1.2.2 Conclusiones

A partir de las figuras 12 y 13 correspondientes al círculo de correlaciones y el primer plano principal, se establecen las siguientes conclusiones (algunas de ellas ya han sido comentadas en el documento).

a. Las cinco variables (PQ, P1, P2, P3 y P4) están bien representadas en el primer plano principal. Esto se deduce en el círculo de correlaciones notando que los cinco puntos, correspondientes a las cinco variables, están muy cercanos a la circunferencia. Esto también se corrobora analizando las

	c^1	c^2	$\cos^2 1$ %	$\cos^2 2$ %	Calidad primer plano %
A01	-0.4332665	-1.48604028	5.47	64.3	69.82
A02	-2.0958423	0.20427067	92.7	0.88	93.61
A03	3.7494733	0.19680915	98.7	0.27	99.05
A04	-0.1574943	-0.25551650	13.6	36.01	49.69
A05	0.0788845	1.63883073	0.18	81.9	82.18
A06	-0.8438672	0.14341992	40.2	1.1	41.36
A07	-1.7046922	-0.06252281	96.1	0.12	96.28
A08	2.5123482	0.05558008	96.4	0.047	96.50
A09	-2.0907552	0.26657456	84.5	1.3	85.96
A10	0.9852116	-0.70140552	36.7	18.6	55.44

Tabla 1.5 Calidad del primer plano principal.

comunalidades calculadas en la tabla 4. La variable que tiene menor calidad de representación es P1, sin embargo, la comunalidad es de un 0.7664 que corresponde a un 76.64%, el cual es aceptable.

b. Todas las variables correlacionan positivo, lo que significa que un individuo específico mantuvo un rendimiento muy estable en las cinco evaluaciones.

c. La mayor correlación se da entre PQ y P4. Es decir, los estudiantes de buen promedio en quices tuvieron un buen rendimiento en el cuarto parcial (con respecto a la media), y viceversa, los estudiantes con promedio bajo en quices, tuvieron un bajo rendimiento en el cuarto parcial (con respecto a la media).

d. Analizando el primer plano principal, se nota que la primera componente principal separa a los estudiantes que están por encima de la media, de los que están por debajo de la media, tanto en el promedio de quices, como en los cuatro parciales.

e. Por su parte, la segunda componente separa a los de mejor rendimiento en los parciales 2 y 3, de los que tuvieron mejor rendimiento en el parcial 4 y el promedio de quices.

f. Los individuos A08 y A03, de acuerdo con su posición en el plano principal, tuvieron bajo rendimiento en todas las evaluaciones.

g. Los individuos A02 y A09 tienen un rendimiento muy parecido.

h. La primera componente principal representa muy bien a los individuos A02, A03, A07, A08 y A09.

i. La segunda componente principal representa muy bien al individuo A05 y en menor grado a A01 (igualmente muy aceptable pues para este caso $\cos^2 2 = 64.3\%$).

Bibliografía

[1] Castillo, W., González, J. y Trejos, J. (en edición). *Análisis Multivariado de Datos*. Universidad de Costa Rica.

[2] Chessel, D. et al (2009). *Package 'ade4', Analysis of Ecological Data : Exploratory and Euclidean methods in Environmental sciences*. Recuperado el 15 de enero del 2010. `http://cran.r-project.org/web/packages/ade4/ade4.pdf`

[3] Correa, J. C. & González, N. (2002). *Gráficos estadísticos con R*. Universidad Nacional de Colombia, Departamento de Matemáticas. Recuperado el 1 de febrero del 2010. `http://cran.r-project.org/doc/contrib/grafi3.pdf`

[4] Ortiz, J. & Pardo, C. E. (2004). *Análisis multivariado de datos en R*. Presentado en 'Simposio de Estadística', Universidad Nacional de Colombia. Departamento de Estadística, Cartagena. Recuperado el 20 de enero del 2010. `www.docentes.unal.edu.co/cepardot/docs/analmultir.pdf`