

Análisis de Componentes Principales

- Ejemplo:
 - Se ha observado un conjunto de variables X_1 a X_8 en un grupo de casos.
- Objetivos
 - Resumir las 8 variables del archivo en sólo unas pocas variables (componentes) no correlacionadas capaces de capturar la máxima información posible de los datos.

Análisis de Componentes Principales

- Procedimiento

- Comprobar si los datos son apropiados para el análisis.
- Si la respuesta es afirmativa, realizar ACP y encontrar un número adecuado de componentes que resuma la información original
- Interpretar los resultados del análisis

- Resultados

- Matriz de comunalidades
- Tabla de varianza total explicada
- Matriz de componentes
- Gráfico de componentes.

Análisis de Componentes Principales

¿Es apropiado el ACP para los datos?

```
> round(cor(a),2)#matriz de correlaciones de x1 a x8
```

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.00	0.97	0.98	0.10	-0.26	0.26	0.21	0.30
x2	0.97	1.00	0.99	0.11	-0.31	0.31	0.29	0.35
x3	0.98	0.99	1.00	0.15	-0.33	0.30	0.28	0.33
x4	0.10	0.11	0.15	1.00	-0.84	0.25	0.30	0.23
x5	-0.26	-0.31	-0.33	-0.84	1.00	-0.35	-0.41	-0.34
x6	0.26	0.31	0.30	0.25	-0.35	1.00	0.97	0.99
x7	0.21	0.29	0.28	0.30	-0.41	0.97	1.00	0.96
x8	0.30	0.35	0.33	0.23	-0.34	0.99	0.96	1.00

Se observa fuerte asociación entre las variables x1, x2 y x3. Entre x4 y x5. Entre x6, x7 y x8. Se concluye que es adecuado el uso de ACP

El determinante de la matriz de correlaciones es casi nulo

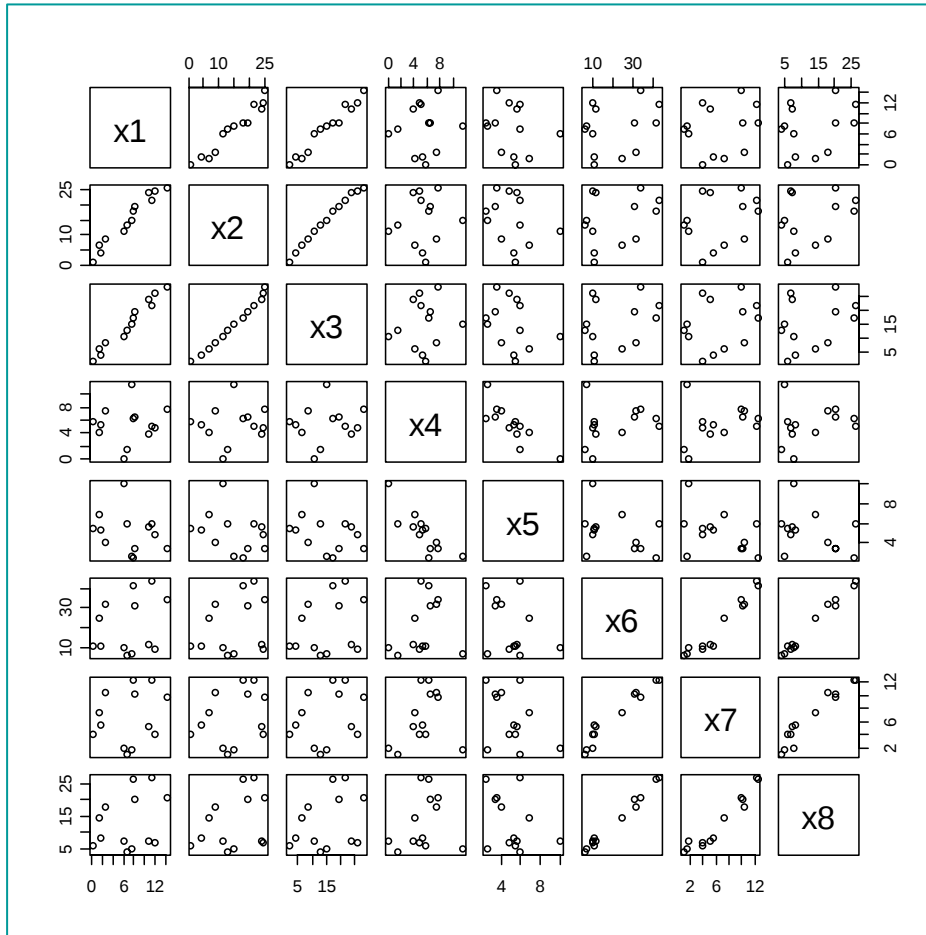
```
>det(cor(a))
```

```
[1] 1.31792e-08
```

Análisis de Componentes Principales

¿Es apropiado el ACP para los datos?

> plot(a) # Gráfico de las variables x1 a x8 del data frame a



Se observa fuerte asociación entre las variables x1, x2 y x3. Entre x4 y x5. Entre x6, x7 y x8. Se concluye que es adecuado el uso de ACP

Análisis de Componentes Principales

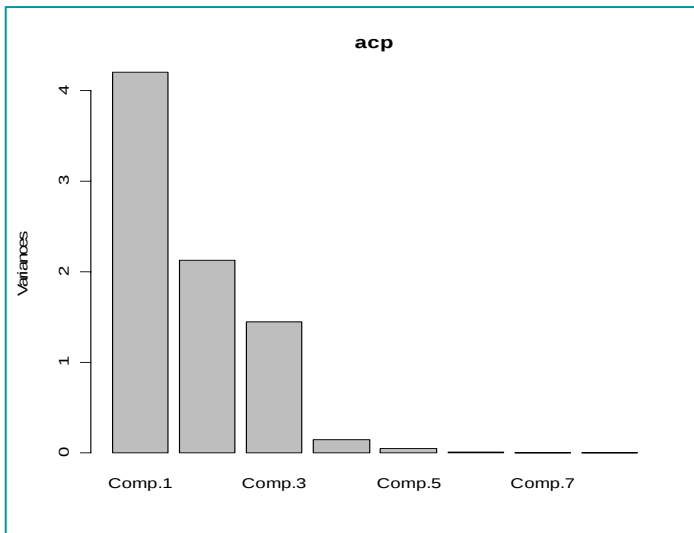
Función princomp()

```
>acp=princomp(a,cor=TRUE) #usa la matriz de correlaciones para el análisis)
>summary(acp)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.0499036	1.4594230	1.2052970	0.38124395	0.218726278
Proportion of Variance	0.5252631	0.2662394	0.1815926	0.01816837	0.005980148
Cumulative Proportion	0.5252631	0.7915026	0.9730952	0.99126353	0.997243674

	Comp.6	Comp.7	Comp.8
Standard deviation	0.118072199	0.0806303271	0.0401038291
Proportion of Variance	0.001742631	0.0008126562	0.0002010396
Cumulative Proportion	0.998986304	0.9997989604	1.0000000000



Tantas Componentes como variables (8)

Según la **Varianza total explicada** la primera componente Y1 tiene una varianza $V(Y1) = 4,202$ igual al primer autovalor. $V(Y2) = 2,130$ y $V(Y3) = 1,453$.

Varianzas de las componentes principales en valores descendentes (Valores propios o autovalores). La suma = 8 (coincide con suma de las varianzas de todas las variables tipificadas, salvo factor $n/n-1$)

Análisis de Componentes Principales

Función princomp()

>acp=**princomp**(a,cor=TRUE) #usa la matriz de correlaciones para el análisis)
>**predict**(acp) #Proporciona las puntuaciones en las componentes principales Y

Tienen interés, especialmente las de mayor varianza, tanto para comparar casos como para utilizarlas en otros análisis, sustituyendo las variables X originales.

Están incorreladas, la matriz de covarianza es:
> round(cov(predict(acp)),3)

		PC1	PC2	PC3	PC4	PC5	PC6
[1,]	-2.6635778	1.3254091	-0.6725289	-0.007877723	0.02462217	0.134964532	
[2,]	-2.0471813	1.1095352	-0.4383188	0.138268981	-0.14259286	0.204099471	
[3,]	-1.3343457	1.4316100	0.9174669	-0.177057283	-0.13901828	-0.188189303	
[4,]	0.2564168	2.1184313	-0.2819989	-0.026118381	-0.19530523	-0.018938842	
[5,]	-2.5867596	-1.1318569	2.1024888	-0.462593645	0.14505663	-0.047087571	
[6,]	-1.9901966	-1.3326656	0.2621269	0.664380683	0.37800622	-0.005820114	
[7,]	-0.4146824	-0.5833204	-2.9419248	-0.556015653	0.19264898	-0.130032932	
[8,]	2.5417560	1.4095469	0.1648108	0.552251799	0.22774004	-0.066823148	
[9,]	1.7914983	0.4628199	-0.1160498	0.270908453	-0.09869660	-0.114171791	
[10,]	2.6333447	0.3594584	1.5592602	-0.458464027	0.08784772	0.070619805	
[11,]	0.2487181	-2.0365336	0.1168405	0.183675749	-0.40630641	-0.012524457	
[12,]	0.4694430	-2.3576761	-0.4123375	0.161514217	-0.20655122	0.001158728	
[13,]	3.0955665	-0.7747582	-0.2598355	-0.282873170	0.13254885	0.172745622	

		PC7	PC8
[1,]	0.08899993	-0.012315001	
[2,]	-0.15179307	-0.015020696	
[3,]	0.08668485	-0.021462307	
[4,]	0.07686015	0.044519754	
[5,]	-0.06102748	-0.024861565	
[6,]	0.06102395	0.034957575	
[7,]	-0.05640002	0.021350829	
[8,]	-0.06924813	0.003955654	
[9,]	-0.05245509	-0.063484334	
[10,]	-0.04688297	0.050912887	
[11,]	-0.01064770	0.064022812	
[12,]	0.02246169	-0.045622684	
[13,]	0.11242390	-0.036952925	

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
PC1	4.202	0.00	0.000	0.000	0.000	0.000	0.000
PC2	0.000	2.13	0.000	0.000	0.000	0.000	0.000
PC3	0.000	0.00	1.453	0.000	0.000	0.000	0.000
PC4	0.000	0.00	0.000	0.145	0.000	0.000	0.000
PC5	0.000	0.00	0.000	0.000	0.048	0.000	0.000
PC6	0.000	0.00	0.000	0.000	0.000	0.014	0.000
PC7	0.000	0.00	0.000	0.000	0.000	0.000	0.007
PC8	0.000	0.00	0.000	0.000	0.000	0.000	0.002

Análisis de Componentes Principales

Función princomp()

>acp=**princomp**(a,cor=TRUE) #usa la matriz de correlaciones para el análisis)

Gráfico de dispersión de las componentes Y1 e Y2:

Los gráficos con las puntuaciones en las componentes principales visualizan qué casos destacan en determinadas componentes:

```
>require(MASS)
```

```
>eqscplot(predict(acp)[,1:2],type="n",xlab="primera componente principal"
```

```
,ylab="segunda componente principal")#Gráfico que usa la misma escala en los ejes
```

```
>text(predict(acp) [,1:2],labels=as.character(row.names(a)))
```

```
>title(main="Empresas en 1ª y 2ª componentes")
```

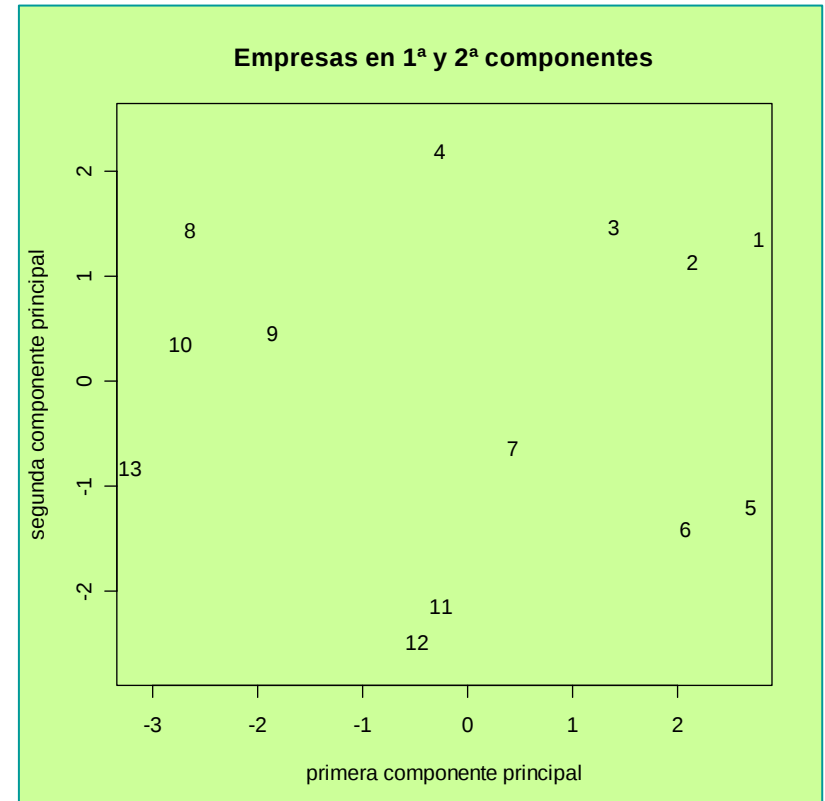
Las puntuaciones de las variables en 2 primeras componentes recogen el 79,1% de variabilidad de los datos.

Estas nuevas variables están incorrelacionadas entre sí.

Las empresas número 4 y 12 son las más diferenciadas en la segunda componente.

Las empresas 13,10 y 8 son las más alejadas de las empresas 1 y 5 en la componente primera.

Recuerde que el signo de las puntuaciones en las componentes es arbitrario.



Análisis de Componentes Principales

Análisis mediante extracción de unas pocas componentes:

97,3 % de la variabilidad de los datos se encuentra representado por las 3 primeras componentes

Matriz de componentes C

```
> C=cor(a,predict(acp)[,1:3]) #Matriz de componentes
```

La **matriz de componentes** muestra las correlaciones entre cada variable y cada una de las 3 componentes.

	Comp.1	Comp.2	Comp.3
x1	-0.7280141	-0.6706139	-0.040499311
x2	-0.7725832	-0.6257327	-0.045165677
x3	-0.7712993	-0.6327374	-0.006976095
x4	-0.4386716	0.2949344	0.809722115
x5	0.6031344	-0.2073072	-0.721199455
x6	-0.7984840	0.5024789	-0.316867759
x7	-0.7910934	0.5358628	-0.248069440
x8	-0.8128367	0.4676518	-0.333601277

Correlaciones de la variable X1 con las 3 componentes principales

La variable X1 presenta coeficientes de correlación altos con las componentes 1 y 2 (-0,728 y -0,671, respectivamente) y correlación insignificante con la componente 3 (igual a -0,0405).

Puede observarse que no hay una estructura clara. Por ejemplo, las variables X1, X2 y X3 presentan correlaciones altas en las componentes 1 y 2.

Análisis de Componentes Principales

Análisis mediante extracción de unas pocas componentes:

97,3 % de la variabilidad de los datos se encuentra representado por las 3 primeras componentes

Comunalidades:

El valor de la **comunalidad** de cada variable de la matriz de datos X, lo podemos interpretar como el coeficiente de correlación múltiple entre cada variable observada (X_i) y todas las componentes principales extraídas.

```
#Comunalidades correspondientes a las 8 variables
# o equivalente suma de cuadrados de cada fila de C
>apply(C*C,1,sum)
      x1      x2      x3      x4      x5      x6      x7      x8
0.9814 0.9905 0.9953 0.9351 0.9269 0.9905 0.9745 0.9907
```

Proporción de varianza de X1 explicada por las 3 componentes

Correlaciones de la variable X1 con las 3 componentes principales

$$0,981 = (-0,728)^2 + (-0,671)^2 + (-0,0405)^2$$

Comunalidad de x1

Matriz de componentes

	Comp.1	Comp.2	Comp.3
x1	-0.7280141	-0.6706139	-0.040499311
x2	-0.7725832	-0.6257327	-0.045165677
x3	-0.7712993	-0.6327374	-0.006976095
x4	-0.4386716	0.2949344	0.809722115
x5	0.6031344	-0.2073072	-0.721199455
x6	-0.7984840	0.5024789	-0.316867759
x7	-0.7910934	0.5358628	-0.248069440
x8	-0.8128367	0.4676518	-0.333601277

Análisis de Componentes Principales

Solución con 3 componentes rotadas

La función **varimax()**

Varimax proporciona directamente la **matriz de componentes rotadas** (rota\$loadings).

>rota=**varimax**(C, normalize = T) # C es la matriz de componentes (correlaciones entre X e Y) en solución no rotada

La matriz de comunalidades lógicamente es la misma. La solución se busca en un espacio de la misma dimensión (3).

Las **saturaciones al cuadrado** correspondientes a las componentes. (Observe que la componente renombrada como 2 acumula mayor varianza)

	Comp.1	Comp.2	Comp.3
SS loadings	2.943	2.984	1.857
Proportion Var	0.368	0.373	0.232
Cumulative Var	0.368	0.741	0.973

La suma de las saturaciones al cuadrado de las 3 componentes es la misma, pero la variabilidad capturada por cada componente difiere

El porcentaje acumulado (explicado por las 3 componentes) es el mismo en ambas soluciones. La variabilidad de los puntos en el espacio tridimensional es la misma. Al cambiar los ejes, cambia la variabilidad relativa a cada eje, pero la total es la misma.

Análisis de Componentes Principales

Rotaciones de la solución

La función **varimax()**

Varimax proporciona directamente la **matriz de componentes rotadas** (rota\$loadings).

>rota=**varimax**(C, normalize = T) # C es la matriz de componentes (correlaciones entre X e Y) en solución no rotada

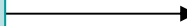
La matriz de comunalidades lógicamente es la misma. La solución se busca en un espacio de la misma dimensión (3).

Matriz C de componentes no rotados

	Comp.1	Comp.2	Comp.3
x1	-0.7280141	-0.6706139	-0.040499311
x2	-0.7725832	-0.6257327	-0.045165677
x3	-0.7712993	-0.6327374	-0.006976095
x4	-0.4386716	0.2949344	0.809722115
x5	0.6031344	-0.2073072	-0.721199455
x6	-0.7984840	0.5024789	-0.316867759
x7	-0.7910934	0.5358628	-0.248069440
x8	-0.8128367	0.4676518	-0.333601277

Matriz CR de componentes rotados

	Comp.1	Comp.2	Comp.3
x1	-0.110	-0.983	
x2	-0.169	-0.977	
x3	-0.147	-0.980	0.117
x4	-0.118		0.960
x5	0.215	0.194	-0.918
x6	-0.976	-0.145	0.130
x7	-0.961	-0.111	0.197
x8	-0.972	-0.181	0.112



La rotación mejora la interpretación de los componentes. Las variables X1, X2 y X3 están altamente correlacionadas con el primer factor o componente

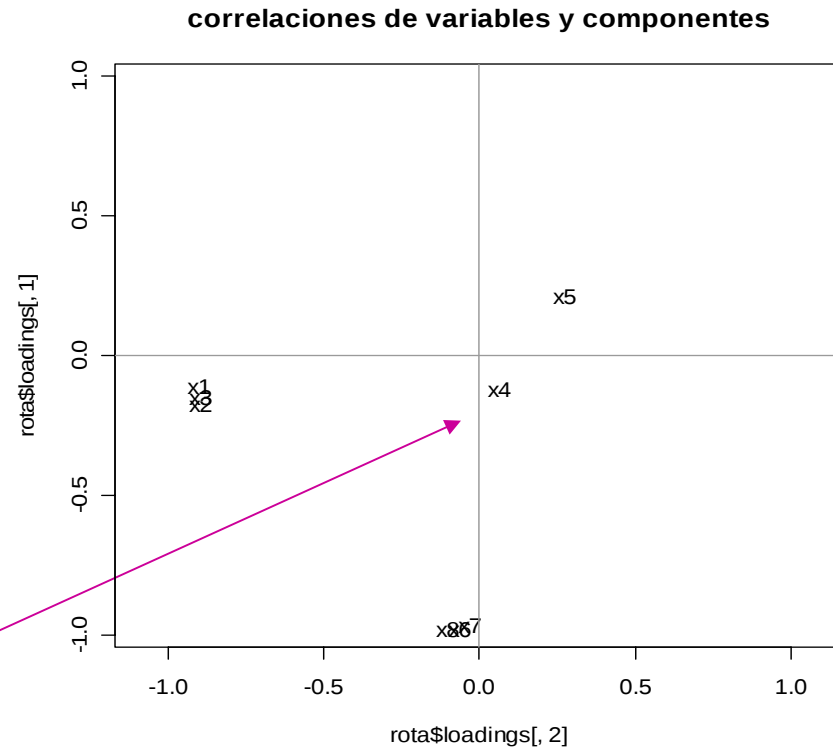
Las variables X6, X7 y X8 están altamente correlacionadas con la segunda componente principal. Las variables X4 y X5 están altamente correlacionadas con la tercera componente principal.

Análisis de Componentes Principales

El gráfico muestra las componentes 1 y 2 como ejes horizontal y vertical, respectivamente en el espacio rotado, que facilita la interpretación.

Las variables X1, X2 y X3 muestran alta correlación con la componente primera. Están alejadas del origen, horizontalmente

Las variables X4 y X5 no muestran correlación importante con ninguna. Observe que están en el centro aproximadamente, cerca del origen



Las variables X6, X7 y X8 están altamente correlacionadas con la segunda componente. Observe que están en el extremo del gráfico