

Reducción de la factorialidad. Análisis de Componentes Principales.

Ricardo Alberich, Juan Gabriel Gomila y Arnau Mir

Introducción

- El problema central del análisis de datos es la reducción de la dimensionalidad.
- Es decir, si es posible describir con precisión los valores de las p variables por un pequeño subconjunto $r < p$ de ellas con una pérdida mínima de información.
- Éste es el objetivo del análisis de componentes principales: dadas n observaciones de p variables, se analiza si es posible representar esta información con menos variables.
- Para alcanzar dicho objetivo, vamos a realizar un ajuste ortogonal por mínimos cuadrados.

Análisis de Componentes Principales

Introducción: Matriz (tabla) de datos.

Ind.	x_1	x_2	\dots	x_k	v_1	v_2
1						
2						
3						
\vdots						
n						
s_1						
s_2						

- Donde las variables x_1, \dots, x_n describen una realidad común de los n individuos observados.

Introducción: Matriz (tabla) de datos.

- Las variables v_1 , v_2 son de perfil (o explicativas) y los individuos s_1 , s_2 son individuos suplementarios o ilustrativos.
- Tanto los individuos como las variables suplementarias ayudan a interpretar la variabilidad de los datos.

Objetivos del análisis

Objetivos del análisis

- Reducción de la dimensionalidad (factorialidad).
- Lo que se busca es un espacio de variables más reducido y fácil de interpretar.
- El problema es que si reducimos el número de variables es posible que no representemos toda la variabilidad de los datos originales.
- El idea básica es:

Consentir una pérdida de información para lograr una ganancia en la significación

Análisis Factorial

- Algunos autores consideran el ACP como una parte del Análisis Factorial
- En las técnicas de AF se postula que la variabilidad total se puede explicar mediante distintos tipos de factores:
- factores comunes subyacentes (F_i).
- factores específicos de las variables (E_i).
- fluctuaciones aleatorias (A_i).

[illegible]

Análisis Factorial

- Podríamos decir que en un Análisis Factorial se fija a priori la cantidad de varianza de cada variable que debe quedar interpretada por los factores comunes.
- Este valor recibe el nombre de comunalidad y se suele representar como h_i^2 .

Así tenemos

- h_i^2 comunalidad de la variable X_i , es la varianza explicada por F_1, F_2, \dots, F_k
- $s_i^2 - h_i^2$ es la varianza de la variable X_i que se queda en los factores específicos y aleatorios.

Var. observada = Var. común + Var. específica y aleat..

El problema de los Componentes Principales

El problema de los Componentes Principales

Todos los factores son comunes

$$X_1 = \alpha_{11}CP_1 + \alpha_{12}CP_2 + \cdots + \alpha_{1p}CP_p$$

$$X_2 = \alpha_{21}CP_1 + \alpha_{22}CP_2 + \cdots + \alpha_{2p}CP_p$$

.....

$$X_p = \alpha_{p1}CP_1 + \alpha_{p2}CP_2 + \cdots + \alpha_{pp}CP_p$$

Se trata de encontrar unas nuevas variables CP_1, \dots, CP_p , a las que llamaremos componentes principales, de forma que:

- Se cumplan las condiciones anteriores.
- El origen de las variables esté situado en el vector de medias o centro de gravedad de las observaciones.
- Sean incorreladas entre si $Cor(CP_i, CP_j) = 0$ para $i \neq j, i, j = 1, \dots, p$.
- $Var(CP_1) > Var(CP_2) > \dots > Var(CP_p)$ y hagan máximas

Tipos de A.C.P.

Tipos de A.C.P:

- Sobre los datos centrados: a cada variable se le resta su media $x_i - \bar{x}_i$.
- Sobre los datos tipificados $\frac{x_i - \bar{x}_i}{s_i}$.
- En el primer caso las variables centradas tienen media cero y la misma varianza que las variables originales: se le suele llamar ACP de covarianzas.
- En el segundo caso las variables tipificadas tienen media cero y varianza 1: se le suele llamar ACP de correlaciones o normado.

Recordemos que dada una matriz de datos \mathbf{X} ($n \times p$ es decirde n individuos y p variables) representábamos por $\tilde{\mathbf{X}}$ la matriz de datos centrada. Entonces:

- La matriz de covarianzas de \mathbf{X} viene dada por

$$\mathbf{S} = 1/n \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$$

ACP covarianzas:

ACP covarianzas:

- Sea **S** la matriz de covarianzas de orden p . Calculamos sus valores propios

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

y los correspondientes vectores propios ortonormales
(perpendiculares y de norma 1)

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$$

- Las direcciones de los componentes principales quedan determinadas por su respectivo vector.
- Cálculo de las coordenadas de la nueva matriz de datos respecto a las nuevas variables *CP*:

Ejemplo

Vamos a realizar un ACP sobre el ejemplo de la estatura de un niño recién nacido visto en el tema anterior de regresión.

Recordemos los datos:

x_1	x_2	x_3	x_4	Sexo
78	48.2	2.75	29.5	Niña
69	45.5	2.15	26.3	Niña
77	46.3	4.41	32.2	Niña
88	49	5.52	36.5	Niño
67	43	3.21	27.2	Niña
80	48	4.32	27.7	Niña
74	48	2.31	28.3	Niña
94	53	4.3	30.3	Niño
102	58	3.71	28.7	Niño

Ejemplo

Donde:

- x_1 : edad en días
- x_2 : estatura al nacer en cm.
- x_3 : peso en Kg. al nacer
- x_4 : aumento en tanto por ciento de su peso con respecto de su peso al nacer.
- El sexo es una variable de perfil que intentaremos explicar con nuestro análisis de componentes principales.

Código para la carga de datos

```
n = 9
p = 4
X = matrix(c(78,48.2,2.75,29.5,69,45.5,2.15,26.3,
77,46.3,4.41,32.2, 88,49,5.52,36.5, 67,43,3.21,27.2,
80,48,4.32,27.7, 74,48,2.31,28.3, 94,53,4.3,30.3,
102,58,3.71,28.7),nrow=n,byrow=T)
Datos= as.data.frame(X)
names(Datos) = paste("x",c(1:p),sep="")
Sexo = as.factor(c("Niña","Niña","Niña","Niño",
"Niña","Niña","Niña","Niño","Niño"))
Datos$Sexo=Sexo
```

El siguiente código dibuja un diagrama matricial de las variables.

```
pairs(Datos[,1:4],pch=21,
bg = c("red", "blue")[unclass(Datos$Sexo)],
main="Diagrama matricial de las variables.
\n Azul: Niña, Rojo: Niño")
```

Ejemplo

- La matriz de covarianzas de los datos anteriores es:

$$\mathbf{S} = \begin{pmatrix} 119333 & 43133 & 6148 & 12511 \\ 43133 & 17193 & 1148 & 1886 \\ 6148 & 1148 & 1111 & 2428 \\ 12511 & 1886 & 2428 & 8624 \end{pmatrix}$$

- Los valores propios son: 136615, 8861, 0738, 0047
- Los vectores propios ortonormales correspondientes a los valores propios, son las columnas de la siguiente matriz:

$$\begin{pmatrix} 0934 & -0022 & 0256 & 0247 \\ 0339 & 0354 & -0661 & -0568 \\ 0047 & -0248 & 0566 & -0785 \\ 0097 & -0902 & -0421 & -0013 \end{pmatrix}$$

Ejemplo

- Las expresiones de las variables nuevas CP_i en función de las antiguas, notemos que se calculan sobre los datos centrados, son:

$$CP_1 = 0934 \cdot \tilde{X}_1 + 0339 \cdot \tilde{X}_2 + 0047 \cdot \tilde{X}_3 \\ + 0097 \cdot \tilde{X}_4,$$

$$CP_2 = -0022 \cdot \tilde{X}_1 + 0354 \cdot \tilde{X}_2 - 0248 \cdot \tilde{X}_3 \\ - 0902 \cdot \tilde{X}_4,$$

$$CP_3 = 0256 \cdot \tilde{X}_1 - 0661 \cdot \tilde{X}_2 + 0566 \cdot \tilde{X}_3 \\ - 0421 \cdot \tilde{X}_4,$$

$$CP_4 = 0247 \cdot \tilde{X}_1 - 0568 \cdot \tilde{X}_2 - 0785 \cdot \tilde{X}_3 \\ - 0013 \cdot \tilde{X}_4$$

Ejemplo


- La nueva matriz de datos respecto de las nuevas variables será:

$$\mathbf{CP} = \tilde{\mathbf{X}}\mathbf{u} = \begin{pmatrix} -3054 & 0201 & -0827 & 0280 \\ -12719 & 2480 & -0333 & 0103 \\ -4293 & -3295 & -0025 & -0228 \\ 7373 & -6736 & -0183 & 0029 \\ -15299 & 0565 & 1029 & 0183 \\ -1354 & 1319 & 1463 & -0321 \\ -6997 & 1411 & -1460 & -0233 \\ 13677 & 0437 & 0629 & 0282 \\ 22666 & 3618 & -0292 & -0095 \end{pmatrix}$$

- Se puede observar que si se multiplican escalarmente dos columnas cualesquiera, el resultado es nulo. Es decir, las columnas de la nueva matriz de datos son ortogonales dos a dos.

Ejemplo

Como podemos observar, nuestro análisis ha explicado la variable de perfil sexo ya que distingue entre niños y niñas con las dos primeras componentes.



ACP_files/figure-beamer/plotACP1-1.pdf

El siguiente código dibuja todos los componentes

```
pairs(solacp$scores,pch=21,  
bg = c("red", "blue")[unclass(Datos$Sexo)],  
main="Diagrama matricial de  
los componentes principales")
```

ACP correlaciones.

ACP correlaciones.

Sea \mathbf{R} la matriz de correlaciones de orden p . Calcularemos sus valores propios

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

y los correspondientes vectores propios ortonormales.

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$$

Las direcciones de los componentes principales quedan determinadas por el vector propio correspondiente.

Cálculo de las coordenadas de la nueva matriz de datos respecto de las nuevas variables CP:

$$\mathbf{CP} = \mathbf{Z}\mathbf{u},$$

donde \mathbf{Z} es la matriz de datos tipificados y \mathbf{u} es la matriz de los vectores propios.

Ejemplo

- Realicemos un análisis ACP de correlaciones con el ejemplo anterior.
- La matriz tipificada de datos es:

$$\mathbf{Z} = \begin{pmatrix} -0275 & -0139 & -0836 & -0045 \\ -1099 & -0791 & -1405 & -1135 \\ -0366 & -0598 & 0739 & 0874 \\ 0641 & 0054 & 1792 & 2338 \\ -1282 & -1393 & -0400 & -0829 \\ -0092 & -0188 & 0654 & -0658 \\ -0641 & -0188 & -1254 & -0454 \\ 1190 & 1018 & 0635 & 0227 \\ 1922 & 2224 & 0075 & -0318 \end{pmatrix}$$

Ejemplo

- La matriz de correlaciones **R** vale, en este caso:

$$\mathbf{R} = \begin{pmatrix} 1000 & 0952 & 0534 & 0390 \\ 0952 & 1000 & 0263 & 0155 \\ 0534 & 0263 & 1000 & 0784 \\ 0390 & 0155 & 0784 & 1000 \end{pmatrix}$$

- Los valores propios de dicha matriz son:

2560, 1229, 0208, 000325

- La matriz de los vectores propios es:

$$\begin{pmatrix} 0573 & 0359 & -0038 & 0736 \\ 0478 & 0578 & 0145 & -0646 \\ 0499 & -0459 & -0707 & -0201 \\ 0442 & -0572 & 0691 & -0029 \end{pmatrix}$$

Ejemplo

- Las expresiones de las variables nuevas CP_i en función de las antiguas Z_i son:

$$CP_1 = 0573 \cdot Z_1 + 0478 \cdot Z_2 + 0499 \cdot Z_3 \\ + 0442 \cdot Z_4,$$

$$CP_2 = 0359 \cdot Z_1 + 0578 \cdot Z_2 - 0459 \cdot Z_3 \\ - 0572 \cdot Z_4,$$

$$CP_3 = -0038 \cdot Z_1 + 0145 \cdot Z_2 - 0707 \cdot Z_3 \\ + 0691 \cdot Z_4,$$

$$CP_4 = 0736 \cdot Z_1 - 0646 \cdot Z_2 - 0201 \cdot Z_3 \\ - 0029 \cdot Z_4$$

Ejemplo

- La nueva matriz de datos respecto de las nuevas variables será:

$$\mathbf{CP} = \mathbf{Zu} = \begin{pmatrix} -0661 & 0231 & 0550 & 0057 \\ -2209 & 0443 & 0137 & 0018 \\ 0259 & -1316 & 0008 & -0058 \\ 2319 & -1899 & 0332 & 0008 \\ -1965 & -0608 & -0444 & 0061 \\ -0107 & -0065 & -0941 & -0058 \\ -1282 & 0497 & 0570 & -0085 \\ 1585 & 0594 & -0189 & 0084 \\ 2061 & 2122 & -0023 & -0027 \end{pmatrix}$$

- Se puede observar que si calculamos el producto escalar de dos columnas cualesquiera, el resultado es nulo. Es decir, las columnas de la nueva matriz de datos son ortogonales dos a dos.

Propiedades ACP covarianzas.

Propiedades ACP covarianzas.

Sea \mathbf{X} una matriz de datos $n \times p$ y sea

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{pmatrix}$$

su matriz de covarianzas.

Recordemos que s_i^2 es la varianza de la variable \mathbf{x}_i y que s_{ij} son las covarianzas de la variables \mathbf{x}_i y \mathbf{x}_j .

Además la Varianza Total = $tr(\mathbf{S}) = \sum_{i=1}^p s_i^2$

Propiedades ACP covarianzas.

- $Var(\mathbf{CP}_i) = \lambda_i$. La varianza de cada componente principal es su valor propio.
- $\sum_{i=1}^n Var(\mathbf{CP}_i) = \sum_{i=1}^n \lambda_i = tr(\mathbf{S}) = \sum_{i=1}^n s_i^2$. Por lo tanto los componentes principales reproducen la varianza total
- Los componentes principales tienen correlación cero entre sí (son *incorrelados*) por lo tanto su matriz de covarianzas es

$$\mathbf{S}_{CP} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

Propiedades ACP covarianzas.

- $\det(\mathbf{S}_{CP}) = \prod_{i=1}^n \lambda_i = \det(\mathbf{S})$. Luego los componentes principales conservan la varianza generalizada.
- La proporción de varianza explicada por la componente j -ésima es

$$\frac{\lambda_j}{\sum_{i=1}^n \lambda_i}$$

Además al ser *incorrelados* la proporción de varianza explicada por los k primeros componentes es

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i}$$

- $cov(\tilde{\mathbf{X}}_i, \mathbf{CP}_j) = \lambda_j u_{ji}$; $corr(\tilde{\mathbf{X}}_i, \mathbf{CP}_j) = \frac{\sqrt{\lambda_j} u_{ji}}{s_i}$ donde u_{ji} es la i -ésima componente del vector propio \mathbf{u}_j .

Ejemplo

Vamos a comprobar las propiedades anteriores con nuestro ejemplo. Recordemos las matrices de datos de las variables originales **X** (centradas) y de las variables en componentes principales **CP**:

$$\tilde{\mathbf{X}} = \begin{pmatrix} -3000 & -0578 & -0881 & -0133 \\ -12000 & -3278 & -1481 & -3333 \\ -4000 & -2478 & 0779 & 2567 \\ 7000 & 0222 & 1889 & 6867 \\ -14000 & -5778 & -0421 & -2433 \\ -1000 & -0778 & 0689 & -1933 \\ -7000 & -0778 & -1321 & -1333 \\ 13000 & 4222 & 0669 & 0667 \\ 21000 & 9222 & 0079 & -0933 \end{pmatrix},$$
$$\mathbf{CP} = \begin{pmatrix} -3054 & 0201 & -0827 & 0280 \\ -12719 & 2480 & -0333 & 0103 \\ -4293 & -3295 & -0025 & -0228 \\ 7373 & -6736 & -0183 & 0029 \\ -15299 & 0565 & 1029 & 0183 \\ -1354 & 1319 & 1463 & -0321 \\ -6997 & 1411 & -1460 & -0233 \\ 13677 & 0437 & 0629 & 0282 \\ 22666 & 3618 & -0292 & -0095 \end{pmatrix}$$

Ejemplo

- La matriz de los vectores propios de la matriz **S** era:

$$\begin{pmatrix} 0934 & -0022 & 0256 & 0247 \\ 0339 & 0354 & -0661 & -0568 \\ 0047 & -0248 & 0566 & -0785 \\ 0097 & -0902 & -0421 & -0013 \end{pmatrix}.$$

- Las varianzas de las variables *CP* son las siguientes:

$$\begin{aligned} \text{Var}(\mathbf{CP}_1) &= 136615, & \text{Var}(\mathbf{CP}_2) &= 8861, \\ \text{Var}(\mathbf{CP}_3) &= 0738, & \text{Var}(\mathbf{CP}_4) &= 00468, \end{aligned}$$

valores que corresponden a los valores propios de la matriz de covarianzas \mathbf{S} .

- La traza de la matriz **S** vale: $\text{tr}(\mathbf{S}) = 146261$. Si sumamos los 4 valores propios, su valor coincide con el valor anterior:
 $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 146261$.

Ejemplo

- La matriz de covarianzas de las variables **CP** vale:

$$\text{cov}(\mathbf{CP}) = \begin{pmatrix} 136615 & 0000 & 0000 & 0000 \\ 0000 & 8861 & 0000 & 0000 \\ 0000 & 0000 & 0738 & 0000 \\ 0000 & 0000 & 0000 & 0047 \end{pmatrix}$$

Podemos observar que es una matriz diagonal con los valores propios de la matriz **S** en la diagonal.

- El determinante de las matrices de covarianzas de $\tilde{\mathbf{X}}$ y **CP** vale 41785, valor que coincide con el producto de los valores propios de la matriz **S**:

$$136615 \cdot 8861 \cdot 0738 \cdot 00468 = 41785$$

Ejemplo

- La proporción de varianza explicada por los componentes viene dada en la tabla siguiente:

Variables	Varianza Explicada
CP₁	$136615/146261 = 0934$
CP_{1,2}	$(136615 + 8861)/146261 = 0995$
CP_{1,2,3}	$(136615 + 8861 + 0738)/146261 = 0999$
CP_{1,2,3,4}	1

Ejemplo

- La matriz de covarianzas entre las variables $\tilde{\mathbf{X}}$ y \mathbf{CP} vale:

$$\text{cov}(\tilde{\mathbf{X}}, \mathbf{CP}) = \begin{pmatrix} 127653 & -0198 & 0189 & 0012 \\ 46377 & 3138 & -0488 & -0027 \\ 6422 & -2195 & 0417 & -0037 \\ 13283 & -7989 & -0311 & -0001 \end{pmatrix}$$

Recordemos la matriz de vectores propios de la matriz $\tilde{\mathbf{S}}$:

$$\begin{pmatrix} 0934 & -0022 & 0256 & 0247 \\ 0339 & 0354 & -0661 & -0568 \\ 0047 & -0248 & 0566 & -0785 \\ 0097 & -0902 & -0421 & -0013 \end{pmatrix}.$$

Ejemplo

- Si multiplicamos la primera columna de la matriz anterior \$(

0934

0339

0047

0097

\$porelvalorpropio136615 de la matriz **S** obtenemos la primera columna de la matriz $cov(\tilde{\mathbf{X}}, \mathbf{CP})$:

$$136615 \cdot \begin{pmatrix} 0934 \\ 0339 \\ 0047 \\ 0097 \end{pmatrix} = \begin{pmatrix} 127652 \\ 46377 \\ 6422 \\ 13283 \end{pmatrix}$$

- En general, podemos escribir:

$$\mathbf{u} \cdot \text{diag}(\lambda) = cov(\tilde{\mathbf{X}}, \mathbf{CP}),$$

donde **u** es la matriz formada por los vectores propios de la matriz **S** y $\text{diag}(\lambda)$ es una matriz diagonal con los valores propios de la matriz

Propiedades ACP covarianzas.

- La primera componente principal es la recta que conserva mayor inercia de la nube de puntos.
- Las dos primeras componentes principales forman el plano que conserva mayor inercia de la nube de puntos.
- Lo mismo sucede con los espacios formados por las k primeras componentes

Propiedades ACP correlaciones.

Propiedades ACP correlaciones.

Sea \mathbf{X} una matriz de datos $n \times p$ y sea

$\mathbf{R} =$

$$\begin{matrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{matrix}$$

La matriz de correlaciones verifica las siguientes propiedades :

- Recordemos que la diagonal es 1 pues es la varianza de los datos tipificados y que r_{ij} son las correlaciones lineales de las variables \mathbf{x}_i y \mathbf{x}_j .
- Además la Varianza Total $= \text{tr}(\mathbf{R}) = p$

Propiedades ACP correlaciones.

- $Var(\mathbf{CP}_i) = \lambda_i$. El valor propio del componente es igual a su varianza
- $\sum_{i=1}^n var(\mathbf{CP}_i) = \sum_{i=1}^n \lambda_i = tr(\mathbf{R}) = p$. Por lo tanto los componentes principales reproducen la varianza total y ésta es igual al numero de variables p .
- Los componentes principales tienen correlación cero entre sí (son *in correlados*) por lo tanto su matriz de covarianzas (que este caso es igual a la de correlaciones es

$$\mathbf{S}_{CP} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

Propiedades ACP correlaciones.

- $\det(\mathbf{S}_{CP}) = \prod_{i=1}^n \lambda_i = \det(\mathbf{R})$. Luego los componentes principales conservan la varianza generalizada.
- La proporción de varianza explicada por cada componente es

$$\frac{\lambda_i}{p}$$

Además al ser *incorreladas* la proporción de varianza explicada por los k primeros componentes es

$$\frac{\sum_{i=1}^k \lambda_i}{p}$$

- $\text{corr}(\mathbf{Z}_i, \mathbf{CP}_j) = \sqrt{\lambda_j} u_{ji}$ donde u_{ji} es la i -ésima componente del vector propio \mathbf{u}_j .

Ejemplo

Vamos a comprobar las propiedades anteriores con nuestro ejemplo. Recordemos las matrices de datos estandarizada **Z** y de las variables en componentes principales **CP**:

$$\mathbf{Z} = \begin{pmatrix} -0275 & -0139 & -0836 & -0045 \\ -1099 & -0791 & -1405 & -1135 \\ -0366 & -0598 & 0739 & 0874 \\ 0641 & 0054 & 1792 & 2338 \\ -1282 & -1393 & -0400 & -0829 \\ -0092 & -0188 & 0654 & -0658 \\ -0641 & -0188 & -1254 & -0454 \\ 1190 & 1018 & 0635 & 0227 \\ 1922 & 2224 & 0075 & -0318 \end{pmatrix},$$
$$\mathbf{CP} = \begin{pmatrix} -0661 & 0231 & 0550 & 0057 \\ -2209 & 0443 & 0137 & 0018 \\ 0259 & -1316 & 0008 & -0058 \\ 2319 & -1899 & 0332 & 0008 \\ -1965 & -0608 & -0444 & 0061 \\ -0107 & -0065 & -0941 & -0058 \\ -1282 & 0497 & 0570 & -0085 \\ 1585 & 0594 & -0189 & 0084 \\ 2061 & 2122 & -0023 & -0027 \end{pmatrix}$$

Ejemplo}

- Las varianzas de las variables \mathbf{CP}_i son las siguientes:

$$\begin{aligned} \text{Var}(\mathbf{CP}_1) &= 2560, & \text{Var}(\mathbf{CP}_2) &= 1229, \\ \text{Var}(\mathbf{CP}_3) &= 0208, & \text{Var}(\mathbf{CP}_4) &= 000325, \end{aligned}$$

valores que corresponden a los valores propios de la matriz \mathbf{R} . * Se puede comprobar que su suma vale 4 que es el valor de p en nuestro caso. * Si calculamos la matriz de covarianzas de las variables \mathbf{CP} obtenemos una matriz diagonal donde ésta contiene los valores propios de la matriz \mathbf{R} calculados anteriormente:

$$\text{cov}(\mathbf{CP}) = \mathbf{S}_{\mathbf{CP}} = \begin{pmatrix} 2560 & 0000 & 0000 & 0000 \\ 0000 & 1229 & 0000 & 0000 \\ 0000 & 0000 & 0208 & 0000 \\ 0000 & 0000 & 0000 & 0003 \end{pmatrix}$$

Ejemplo}

- El determinante de la matriz \mathbf{S}_{CP} vale: $\det(\mathbf{S}_{CP}) = 000213$, valor que coincide con el producto de los valores propios de la matriz \mathbf{R} :

$$2560 \cdot 1229 \cdot 0208 \cdot 000325 = 000213$$

- La proporción de varianza explicada por los componentes viene dada en la tabla siguiente:

Variables	Varianza Explicada
\mathbf{CP}_1	$2560/4 = 0640$
$\mathbf{CP}_{1,2}$	$(2560 + 1229)/4 = 0947$
$\mathbf{CP}_{1,2,3}$	$(2560 + 1229 + 0208)/4 = 0999$
$\mathbf{CP}_{1,2,3,4}$	1

Ejemplo}

- La matriz de correlaciones entre las variables **Z** y **CP** vale:

$$\text{corr}(\mathbf{Z}, \mathbf{CP}) = \begin{pmatrix} 0916 & 0398 & -0017 & 0042 \\ 0764 & 0641 & 0066 & -0037 \\ 0798 & -0509 & -0323 & -0011 \\ 0706 & -0634 & 0315 & -0002 \end{pmatrix}$$

La matriz de vectores propios de la matriz **R** era:

$$\begin{pmatrix} 0573 & 0359 & -0038 & 0736 \\ 0478 & 0578 & 0145 & -0646 \\ 0499 & -0459 & -0707 & -0201 \\ 0442 & -0572 & 0691 & -0029 \end{pmatrix}$$

Ejemplo}

- Si multiplicamos la primera columna de la matriz anterior

$\begin{pmatrix} 0573 \\ 0478 \\ 0499 \\ 0442 \end{pmatrix}$ por la raíz cuadrada del primer valor propio de la

matriz \mathbf{R} , $\sqrt{2560}$, obtenemos la primera columna de la matriz $\text{corr}(\mathbf{Z}, \mathbf{CP})$:

$$\sqrt{2560} \cdot \begin{pmatrix} 0573 \\ 0478 \\ 0499 \\ 0442 \end{pmatrix} = \begin{pmatrix} 0916 \\ 0764 \\ 0798 \\ 0706 \end{pmatrix}$$

Ejemplo}

- En general, podemos escribir:

$$\mathbf{u} \cdot \text{diag}(\sqrt{\lambda}) = \text{corr}(\mathbf{Z}, \mathbf{CP}),$$

donde \mathbf{u} es la

matriz formada por los vectores propios de la matriz $\tilde{\mathbf{R}}$ y $\{\text{diag}\}$

$(\sqrt{\lambda})$ es una matriz diagonal con la raíz cuadrada de los valores propios de la matriz $\tilde{\mathbf{R}}$ en la diagonal.

Propiedades ACP correlaciones

- La primera componente principal es la recta que conserva mayor inercia de la nube de puntos.
- Los dos primeros componentes principales forma el plano que conserva mayor inercia de la nube de puntos.
- Lo mismo sucede con los espacios formados por los k primeros componentes

Etapas de un ACP

Etapas de un ACP

- Determinar las variables e individuos que intervienen en el análisis, las variables de perfil y los individuos ilustrativos.
- Decidir si se realiza el análisis sobre los datos brutos (matriz de covarianzas) o sobre los datos tipificados (matriz de correlaciones):
- Cuando las variables originales \mathbf{X} están medidas en distintas unidades, conviene aplicar el análisis de correlaciones. Si están en las mismas unidades, ambas alternativas son posibles.
- Si las diferencias entre las varianzas son informativas y queremos tenerlas en cuenta en el análisis, no debemos estandarizar las variables.

Etapas de un ACP

- Reducción de la dimensionalidad; tenemos que decidir cuántas componente retenemos. La cantidad de varianza retenida será:

Comp.	Valor propio	Cantidad retenida
Cp_1	λ_1	$\lambda_1 / \sum_{i=1}^p \lambda_i$
Cp_2	λ_2	$(\lambda_1 + \lambda_2) / \sum_{i=1}^p \lambda_i$
Cp_3	λ_3	$(\lambda_1 + \lambda_2 + \lambda_3) / \sum_{i=1}^p \lambda_i$
...
Cp_p	λ_p	$(\lambda_1 + \dots + \lambda_p) / \sum_{i=1}^p \lambda_i = 1$