Tema 4
Análisis de Componentes Principales

Aurea Grané
Departamento de Estadística
Universidad Carlos III de Madrid

Se han observado p variables $X_1, X_2, \ldots, X_p$ sobre una muestra de n individuos. La matriz de datos muestrales es

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

En adelante supondremos que $\mathbf{X}$ es una matriz centrada. (Si no lo fuera, la transformación $\mathbf{H}\,\mathbf{X}$, donde $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\,\mathbf{1}'$ es la matriz de centrado, daría lugar a tal configuración).

Problema: ¿Podemos describir la "información" contenida en estos datos mediante algún conjunto de variables menor que el de variables originales?

Idea: Si una variable es función de otras, contiene información redundante.

Por tanto, si las p variables observadas están **fuertemente correlacionadas**, será posible sustituirlas por menos variables sin gran pérdida de "información".

Esta reducción de la dimensión va a permitir:

- Simplificar posteriores análisis, que se harán a partir de un menor número de variables que el original.
- Una representación gráfica de los individuos en dimensión reducida (generalmente, 1 ó 2).
- Examinar e interpretar las relaciones entre las variables observadas.

Definición y obtención de las componentes principales

Sean $\mathbf{X} = [X_1, \ldots, X_p]$ y $\mathbf{S} = \mathsf{var}(\mathbf{X})$ su matriz de covarianzas.

Puesto que $\mathbf{S} \geq 0$ y simétrica, su descomposición espectral es

$$\mathbf{S} = \mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}',$$

donde $\mathbf{T}'\,\mathbf{T} = \mathbf{T}\,\mathbf{T}' = \mathbf{I}$, con $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_p]$ y $\mathbf{\Lambda} = \mathsf{diag}(\lambda_1, \ldots, \lambda_p)$, con $\lambda_1 > ... > \lambda_p$.

Las *componentes principales* de $\mathbf{X}$ son las nuevas variables

$$Y_j = \mathbf{X}\,\mathbf{t}_j, \quad j = 1, \ldots, p.$$

Para cada j, la nueva variable Y_j se construye a partir del j-ésimo autovector de $\mathbf{S}$.

Propiedades de las componentes principales

- Las componentes principales tienen **varianza decreciente**:

$$\left.\begin{array}{c} \mathsf{var}(Y_1) = \mathsf{var}(\mathbf{X}\,\mathbf{t}_1) = \mathbf{t}_1'\,\mathbf{S}\,\mathbf{t}_1 = \lambda_1\,\mathbf{t}_1'\,\mathbf{t}_1 = \lambda_1 \\ \mathsf{var}(Y_2) = \mathsf{var}(\mathbf{X}\,\mathbf{t}_2) = \mathbf{t}_2'\,\mathbf{S}\,\mathbf{t}_2 = \lambda_2\,\mathbf{t}_2'\,\mathbf{t}_2 = \lambda_2 \\ \vdots \\ \mathsf{var}(Y_p) = \mathsf{var}(\mathbf{X}\,\mathbf{t}_p) = \mathbf{t}_p'\,\mathbf{S}\,\mathbf{t}_p = \lambda_p\,\mathbf{t}_p'\,\mathbf{t}_p = \lambda_p \end{array}\right\} \text{ con } \lambda_1 > ... > \lambda_p$$

- y están **incorrelacionadas** unas con otras:
  $\mathsf{cov}(Y_i, Y_j) = \mathsf{cov}(\mathbf{X}\,\mathbf{t}_i, \mathbf{X}\,\mathbf{t}_j) = \mathbf{t}_i'\,\mathbf{S}\,\mathbf{t}_j = \lambda_j\,\mathbf{t}_i'\,\mathbf{t}_j = 0$, para $i \neq j$, puesto que $\mathbf{T}$ es una matriz ortogonal.

- Las covarianzas entre cada componente principal y las variables originales X_i son: $\mathsf{Cov}(Y_j, [X_1, \ldots, X_p]) = \lambda_j\,\mathbf{t}_j', \quad j = 1, \ldots, p.$

  Utilizando que $\mathbf{Y} = \mathbf{X}\,\mathbf{T}$ y la descomposición espectral de $\mathbf{S}$:

$$\mathsf{Cov}(\mathbf{Y}, \mathbf{X}) = \frac{1}{n}\,\mathbf{Y}'\,\mathbf{X} = \frac{1}{n}\,\mathbf{T}'\,\mathbf{X}'\,\mathbf{X} = \mathbf{T}'\,\mathbf{S} = \mathbf{T}'\,(\mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}') = \mathbf{\Lambda}\,\mathbf{T}'$$

  La fila j de esta matriz proporciona las covarianzas entre Y_j y las variables originales $X_1, \ldots, X_p$.

  Por ejemplo, la covarianza entre Y_1 y $X_1, \ldots, X_p$ es $\lambda_1\,\mathbf{t}_1'$.

- La correlación entre Y_j y la variable original X_i es

$$\mathsf{corr}(Y_j, X_i) = \frac{\mathsf{cov}(Y_j, X_i)}{\sqrt{\mathsf{var}(Y_j)\ \mathsf{var}(X_i)}} = \frac{\lambda_j\,t_{ij}}{\sqrt{\lambda_j\,s_{ii}}} = t_{ij}\,\sqrt{\frac{\lambda_j}{s_{ii}}},$$

  donde t_{ij} es el elemento i-ésimo del autovector $\mathbf{t}_j$.

Representación de los individuos

Con las nuevas coordenadas dadas por las componentes principales, el individuo i-ésimo, es decir, la fila $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ de la matriz de datos $\mathbf{X}$, se expresa como

$$\mathbf{y}_i' = \mathbf{x}_i'\,\mathbf{T} = (\mathbf{x}_i'\,\mathbf{t}_1, \ldots, \mathbf{x}_i'\,\mathbf{t}_p).$$

La matriz de datos transformados es $\mathbf{Y} = \mathbf{X}\,\mathbf{T}$, que representa las "observaciones" de las nuevas variables (componentes principales) sobre los n individuos de la muestra.

Esta transformación puede interpretarse geométricamente considerando los n individuos como n puntos del espacio $\mathbb{R}^p$.

Consideremos la distancia euclídea (al cuadrado) entre los individuos i-ésimo y j-ésimo, en las nuevas coordenadas:

$$\begin{aligned} d^2_{Euclid}(i,j) &= (\mathbf{y}'_i - \mathbf{y}'_j)(\mathbf{y}_i - \mathbf{y}_j) = (\mathbf{x}'_i\,\mathbf{T} - \mathbf{x}'_j\,\mathbf{T})(\mathbf{T}'\,\mathbf{x}_i - \mathbf{T}'\,\mathbf{x}_j) \\ &= (\mathbf{x}'_i - \mathbf{x}'_j)\,\mathbf{T}'\,\mathbf{T}\,(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}'_i - \mathbf{x}'_j)(\mathbf{x}_i - \mathbf{x}_j) \end{aligned}$$

Ignorando orientaciones, podemos pensar la transformación como una *rotación* en $\mathbb{R}^p$.

El primero de los nuevos ejes (la primera componente principal) es la dirección a lo largo de la cual la dispersión de los puntos-individuos es máxima. Sucesivamente, cada componente principal es aquella dirección, ortogonal a las anteriores, a lo largo de la cual hay dispersión máxima.

Reducción de la dimensión

La variación total de $\mathbf{X}$ se define como $\mathsf{tr}(\mathbf{S}) = \sum_{i=1}^p \lambda_i$.

La variación total de $\mathbf{Y} = \mathbf{X}\,\mathbf{T}$ es igual a la variación total de $\mathbf{X}$:

$$\mathsf{tr}(\mathsf{var}(\mathbf{Y})) = \mathsf{tr}\left(\frac{1}{n}\mathbf{T}'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{T}\right) = \mathsf{tr}(\mathbf{T}'\,\mathbf{S}\,\mathbf{T}) = \mathsf{tr}(\mathbf{T}'\,\mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}'\,\mathbf{T}) = \sum_{i=1}^p \lambda_i.$$

puesto que, $\mathbf{S} = \mathbf{T}\,\mathbf{\Lambda}\,\mathbf{T}'$, donde $\mathbf{T}$ es una matriz ortogonal.

Cuando el cociente (**porcentaje de variabilidad explicada**)

$$P_q = \frac{\sum_{i=1}^q \lambda_i}{\mathsf{tr}\,\mathbf{S}} \times 100, \quad q < p,$$

es cercano a 100%, entonces las variables $Y_1, \ldots, Y_q$ pueden reemplazar a $X_1, \ldots, X_p$ sin gran pérdida de información, en términos de "variación total".

Ejemplo 1: Problema 4.2

La Tabla siguiente contiene información sobre chalets construidos por diez promotoras que operan a lo largo de la costa española:

Promotora	X_1 =Duración media hipoteca (años)	X_2 =Precio medio (millones euros)	X_3 =Superficie media (m^2) de cocina
1	8.7	0.3	3.1
2	14.3	0.9	7.4
3	18.9	1.8	9.0
4	19.0	0.8	9.4
5	20.5	0.9	8.3
6	14.7	1.1	7.6
7	18.8	2.5	12.6
8	37.3	2.7	18.1
9	12.6	1.3	5.9
10	25.7	3.4	15.9

Considerando solamente las variables X_1 y X_2 realizar un análisis de componentes principales.

El vector de medias y la matriz de covarianzas son:

$$\bar{\mathbf{x}} = (19.05, 1.57)', \quad \mathbf{S} = \begin{pmatrix} 56.9685 & 5.1705 \\ 5.1705 & 0.8941 \end{pmatrix}.$$

Los autovalores y autovectores de $\mathbf{S}$ son:

$$\mathbf{\Lambda} = \mathsf{diag}(57.4413, 0.4213), \quad \mathbf{T} = \begin{pmatrix} 0.9958 & -0.0911 \\ 0.0911 & 0.9958 \end{pmatrix}.$$

Por tanto, las componentes principales serían:

$$Y_1 = 0.9958\,X_1 + 0.0911\,X_2, \quad Y_2 = -0.0911\,X_1 + 0.9958\,X_2,$$

y los porcentajes de variabilidad explicados por cada componente son:

$$\frac{57.4413}{57.8626} \times 100 = 99.27\%, \quad \frac{0.4213}{57.8626} \times 100 = 0.73\%$$

Las correlaciones entre Y_1 y las variables originales son:

$$\begin{aligned}
\mathsf{corr}(Y_1, X_1) &= t_{11}\sqrt{\frac{\lambda_1}{s_{11}}} = 0.9958\sqrt{\frac{57.4413}{56.9685}} = 0.9999 \\
\mathsf{corr}(Y_1, X_2) &= t_{21}\sqrt{\frac{\lambda_1}{s_{22}}} = 0.0911\sqrt{\frac{57.4413}{0.8941}} = 0.7302
\end{aligned}$$

Las correlaciones entre Y_2 y las variables originales son:

$$\begin{aligned}
\mathsf{corr}(Y_2, X_1) &= t_{12}\sqrt{\frac{\lambda_2}{s_{11}}} = -0.0911\sqrt{\frac{0.4213}{56.9685}} = -0.0078 \\
\mathsf{corr}(Y_2, X_2) &= t_{22}\sqrt{\frac{\lambda_2}{s_{22}}} = 0.9958\sqrt{\frac{0.4213}{0.8941}} = 0.6836
\end{aligned}$$

Observemos la primera componente con más detalle:

$$Y_1 = 0.9958\, X_1 + 0.0911\, X_2.$$

Esta componente es esencialmente X_1. Esto es debido a que la varianza de X_1 ($s_{11} = 56.9685$) es mucho mayor que la varianza de X_2 ($s_{22} = 0.8941$) y, por tanto, gran parte de la variabilidad del sistema queda explicada por X_1.

En este caso conviene **estandarizar** los datos y realizar un nuevo análisis de componentes principales. Esto es equivalente a realizar el análsis a partir de la matriz de correlaciones $\mathbf{R}$.

La matriz de correlaciones $\mathbf{R}$ es:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.7245 \\ 0.7245 & 1.0000 \end{pmatrix}$$

y sus autovalores y autovectores son:

$$\tilde{\mathbf{\Lambda}} = \mathsf{diag}(1.7245, 0.2755), \quad \tilde{\mathbf{T}} = \begin{pmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{pmatrix}.$$

Por tanto, las componentes principales son:

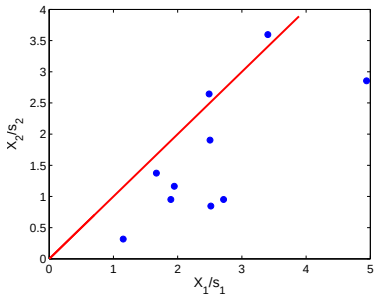
$$\tilde{Y}_1 = 0.7071\, X_1 + 0.7071\, X_2, \quad \tilde{Y}_2 = -0.7071\, X_1 + 0.7071\, X_2$$

y los porcentajes de variabilidad explicados por cada componente son:

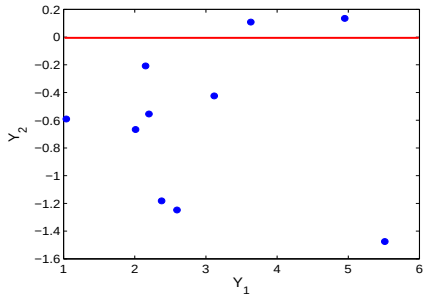
$$\frac{1.7245}{2} \times 100 = 86.22\%, \quad \frac{0.2755}{2} \times 100 = 13.78\%$$

Rotación de los ejes

Gráfico de dispersión variables estandarizadas



Representación en componentes principales



Ejemplo 2: Problema 4.4

La Tabla siguiente contienen 11 indicadores económicos y sociales de 96 países. Las variables observadas son:

X_1 = Tasa anual de crecimiento de la población, X_2 = Tasa de mortalidad infantil por cada 1000 nacidos vivos, X_3 = Porcentaje de mujeres en la población activa, X_4 = PNB en 1995 (en millones de dólares), X_5 = Producción de electricidad (en millones kW/h), X_6 = Líneas telefónicas por cada 1000 habitantes, X_7 = Consumo de agua per cápita, X_8 = Proporción de la superficie del país cubierta por bosques, X_9 = Proporción de deforestación anual, X_{10} = Consumo de energía per cápita, X_{11} = Emisión de CO2 per cápita.

País	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1. Albania	1	30	41	2199	3903	12	94	53	0	341	1.2
2. Angola	3	124	46	4422	955	6	57	19	0.7	89	0.5
3. Arabia Saudi	4.3	21	13	133540	91019	96	497	1	0	4566	13.1
4. Argelia	2.5	34	24	44609	19883	42	180	2	0.8	906	3
5. Argentina	1.3	22	31	278431	65962	160	1043	22	0.1	1504	3.5
6. Australia	1.4	6	43	337909	167155	510	933	19	0	5341	15.3
7. Austria	0.6	6	41	216547	53259	465	304	47	-0.4	3301	7.2
8. Bangladesh	2	79	42	28599	9891	2	220	6	4.1	64	0.2
9. Bélgica	0.3	8	40	250710	72236	457	917	20	-0.3	5120	10.1
10. Benin	3	95	48	2034	6	5	26	45	1.3	20	0.1
...						...					
61. Mozambique	1.8	113	48	1353	490	3	55	22	0.8	40	0.1
...						...					
86. Tailandia	1.3	35	46	159630	71177	59	602	25	3.5	769	2
87. Tanzania	3.1	82	49	3703	1913	3	40	38	1.2	34	0.1
88. Túnez	2.1	39	30	16369	6714	58	381	4	-1.9	595	1.6
89. Turquia	1.9	48	35	169452	78322	212	585	26	0	957	2.5
90. Ucrania	0.1	15	49	84084	202995	157	673	16	-0.3	3180	11.7
91. Uruguay	0.6	18	40	16458	7617	196	241	4	-0.6	629	1.6
92. Venezuela	2.4	23	33	65382	73116	111	382	52	1.2	2186	5.7
93. Vietnam	2.2	41	49	17634	12270	11	414	26	1.5	101	0.3
94. Yemen	4.2	100	29	4044	2159	12	335	8	0	206	0.7
95. Zambia	2.6	109	45	3605	7785	8	186	43	1.1	149	0.3
96. Zimbabue	2.8	55	44	5933	7334	14	136	23	0.7	438	1.8

Observemos que:

- las unidades de medida de las variables X_i son muy distintas (porcentajes, dólares, kWh, ...). Recordemos que los cambios de unidades (transformaciones lineales) afectan a la varianza de la variable:
$$\xi_i = a\, X_i \Rightarrow \mathsf{var}(\xi_i) = a^2\ \mathsf{var}(X_i)$$
y, como consecuencia, a las componentes principales.
- las elevadas varianzas de X_4 y X_5 hacen prever que un análisis de componentes principales realizado a partir de la matriz de covarianzas $\mathbf{S}$ dará como resultado una primera y segunda componentes principales que coincidirán básicamente con estas dos variables observadas.

Para obtener unas componentes principales que **no dependan de las unidades** en que han sido medidas las variables originales, deberíamos estandarizar a media cero y varianza unidad las variables originales X_i.

Esto es equivalente a realizar el análisis de componentes principales a partir de la matriz de correlaciones $\mathbf{R}$:

$$\mathbf{R} = \tilde{\mathbf{T}}\,\tilde{\mathbf{\Lambda}}\,\tilde{\mathbf{T}}',$$

donde $\tilde{\mathbf{T}}'\,\tilde{\mathbf{T}} = \tilde{\mathbf{T}}\,\tilde{\mathbf{T}}' = \mathbf{I}$, y $\tilde{\mathbf{\Lambda}} = \mathsf{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p)$, con $\tilde{\lambda}_1 > \ldots > \tilde{\lambda}_p$.

Con la diferencia que ahora la representación de individuos es:

$$\tilde{\mathbf{Y}} = \mathbf{X}\,\mathbf{S}_0^{-1}\,\tilde{\mathbf{T}},$$

donde $\mathbf{S}_0 = \mathsf{diag}(s_1, \ldots, s_p)$, siendo $s_i = \sqrt{\mathsf{var}(X_i)}$, para $i = 1, \ldots, p$.

Siguiendo con el segundo ejemplo, las dos primeras componentes principales obtenidas a partir de $\mathbf{R}$ son:

	Y_1	Y_2
X_1	-0.3141	**0.3484**
X_2	**-0.3924**	-0.0414
X_3	0.1165	**-0.5828**
X_4	0.2954	-0.1769
X_5	0.2590	-0.1736
X_6	**0.4461**	-0.0272
X_7	0.0924	**0.3206**
X_8	0.0057	**-0.4574**
X_9	-0.2437	-0.1541
X_{10}	**0.4150**	0.2329
X_{11}	**0.3745**	0.2917

Las variables X_2, X_6, X_{10} y X_{11} son las que más contribuyen en la primera componente principal, que puede interpretarse como un *índice de riqueda o de desarrollo.*

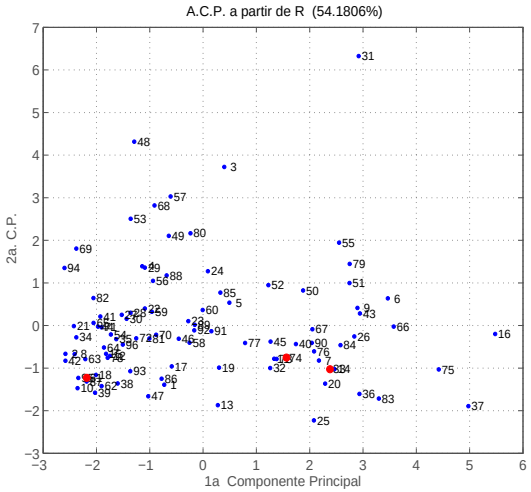
Las variables X_1, X_3, X_7 y X_8 son las que más contribuyen en la segunda componente, que puede interpretarse como un *índice de sostenibilidad/modernidad.*

El porcentaje de variabilidad explicado es: $P_2 = 54.1806$.

Las correlaciones entre las componentes principales y las variables originales son:

	Y_1	Y_2
X_1	**-0.6306**	**0.4840**
X_2	**-0.7877**	-0.0575
X_3	0.2340	**-0.8097**
X_4	0.5930	-0.2458
X_5	0.5199	-0.2411
X_6	**0.8955**	-0.0378
X_7	0.1855	**0.4454**
X_8	0.0114	**-0.6355**
X_9	-0.4891	-0.2141
X_{10}	**0.8332**	0.3235
X_{11}	**0.7519**	0.4052

donde ahora, $\mathsf{corr}(Y_j, [X_1, \ldots, X_p]) = \sqrt{\tilde{\lambda}_j}\, \tilde{\mathbf{t}}'_j$.



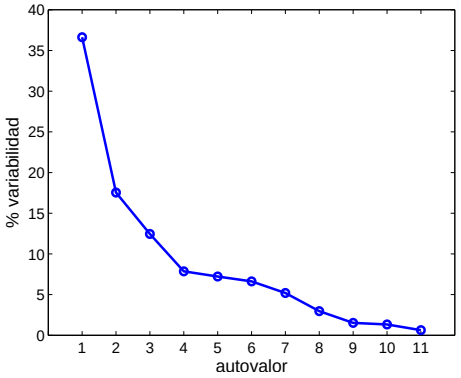
1a Componente Principal (36.64%): *índice de desarrollo.* Según este índice, Canadá (16), Francia (37) y Reino Unido (75) serían los países con mayor grado de desarrollo, mientras que Yemen (94), Haití (42) y Angola (2) serían los de menor grado.

2a Componente Principal (17.54%): *índice de sostenibilidad/modernidad.* Irán (48), Arabia Saudí (3) y Emiratos Árabes (31) son los países con un mayor valor en la segunda componente principal.

Determinación del número de componentes

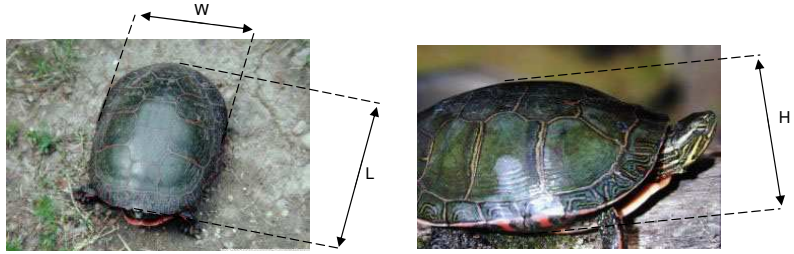
1. **Procentaje explicado**. Es el método más sencillo. Consiste en fijar un porcentaje de variabilidad explicado, por ejemplo el 90%, y considerar las sucesivas componentes principales hasta superar el porcentaje prefijado.
2. **Criterio de Kaisser**. Se excluyen aquellas componentes cuyos autovalores sean menores que $\bar{\lambda} = \mathsf{tr}(\mathbf{S})/p$, o bien menores que 1 si se han calculado las componentes a partir de $\mathbf{R}$.
3. **Modificación de Jollife**. Se ha comprobado que cuando $p \leq 20$ el criterio de Kaisser tiende a incluir pocas componentes. La modificación de Jollife excluye aquellas componentes cuyos autovalores sean menores que $0.7\,\bar{\lambda} = 0.7\ \mathsf{tr}(\mathbf{S})/p$, o bien que 0.7 si se han calculado las componentes a partir de $\mathbf{R}$.

4. **Scree test de Cattell**. Es un método muy visual. Se consideran las $q < p$ primeras componentes hasta que los descensos de pendiente son poco significativos. Estos diagramas suelen indicar con claridad donde terminan los autovalores "grandes" y donde empiezan los "pequeños".



Ejemplo

Jolicoeur and Mosiman (1960) estudian la longitud, el ancho y la altura del caparazón de 24 tortugas *Chrysemyis picta marginata* hembra.



La tabla siguiente contiene estas variables medidas en mm.

longitud (L)	ancho (W)	altura (H)
98	81	38
103	84	38
103	86	42
105	86	42
109	88	44
123	95	46
134	100	48
136	102	49
123	92	50
133	99	51
133	102	51
133	102	51
138	98	51
138	99	51
141	105	53
149	107	55
153	107	56
147	108	57
155	117	60
158	115	62
155	115	63
159	118	63
162	124	61
177	132	67

a) Obtener el vector de medias, la matriz de covarianzas y la de correlaciones.

b) Obtener las componentes principales. Razonar si éstas deben calcularse a partir de la matriz de correlaciones o a partir de la de covarianzas.

c) ¿Qué porcentaje de variabilidad explican los nuevos ejes de representación?

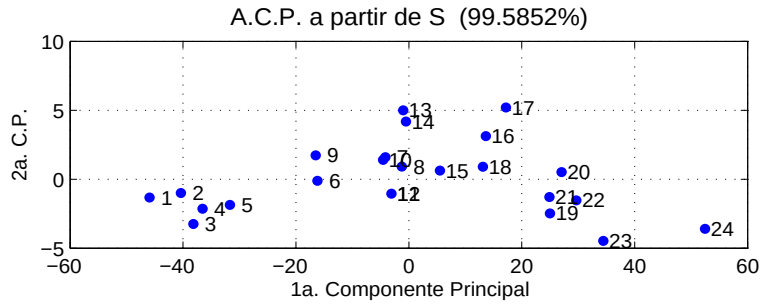
d) Interpretar las dos primeras componentes principales.

a) Llamamos `X` a la matriz de datos anterior y utilizando el programa `descrip.m`, en Matlab escribimos:

```
[m,S,R]=descrip(X)

m = 136.0417   102.5833    52.0417

S =
   432.7066   259.6840   159.0399
   259.6840   164.5764    97.6007
   159.0399    97.6007    62.0399

R=
   1.0000     0.9731     0.9707
   0.9731     1.0000     0.9659
   0.9707     0.9659     1.0000
```

b)–d) Utilizando el programa `comp.m`, en Matlab escribimos:

```
[T1,Y1,acum1,T2,Y2,acum2]=comp(X)

    T1 =
        0.8139    0.5549    0.1723
        0.4961   -0.8180    0.2911
        0.3025   -0.1514   -0.9411

 acum1 = 98.6012   99.5852  100.0000
```

Pregunta: Si los autovalores de la matriz de covarianzas son:

$$\lambda_1 = 650.1004,\ \lambda_2 = 6.4876,\ \lambda_3 = 2.7349,$$

¿cuánto valen las correlaciones entre la primera componente principal y las variables L, W y H?