AID - Trabajo Práctico Nro. III

Análisis de Componentes Principales

Ejercicio 1:

Sea Σ la matriz de varianzas y covarianzas poblacionales:

$$\Sigma = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

Correspondiente al vector aleatorio $X = (X_1, X_2, X_3)'$ de media 0.

a- Hallar los autovalores y autovectores de la matriz de varianzas y covarianzas.

b- Escribir la expresión de las componentes principales $Y=(Y_1, Y_2, Y_3)'$ e indique que proporción de la variabilidad explica cada una de ellas.

c- Hallar las primeras dos componentes principales correspondientes a la observación X=(2,2,1)

Ejercicio 2:

Considerando los datos de la base **chalets.xls**, se pide:

a- Graficar el boxplot de cada una de las variables. Indicar, si se observa, la presencia de valores atípicos.

b- Graficar los diagramas de dispersión de las variables de a pares. Estimar la presencia de correlación entre variables a partir de estos gráficos, indicando si le parece fuerte y el signo de las mismas.

c- Calcular el vector de medias y la matriz de varianzas y covarianzas muestral.

d- Hallar la matriz de correlación muestral. Verificar las estimaciones realizadas visualmente.

e- A partir de estas observaciones, le parece razonable pensar en un análisis de componentes principales para reducir la dimensión del problema?.

f- Hallar la primera componente principal y graficar sus coeficientes mediante barras verticales.

g- Indicar qué porcentaje de la variabilidad total logra explicar esta componente.

h- Explicar si se trata de una componente de tamaño o de forma. Es posible ordenar las promotoras en función de esta componente?. Si la respuesta es afirmativa, cual es la mayor y cual la menor; si es negativa, explicar por qué no es posible ordenarlos.

Ejercicio 3:

Dado el siguiente conjunto de datos:

$$X = \begin{bmatrix} 3 & 6 \\ 5 & 6 \\ 10 & 12 \end{bmatrix}$$

a. Calcule:
   - La matriz de covarianza.
   - Los autovalores y autovectores.
   - Las componentes principales y su contribución porcentual a la varianza total.
   - Grafique los datos en R^2 en la base original.
   - Y en la base de los dos primeros ejes.

b. Repita los cálculos con los datos estandarizados.

c. Interprete los resultados obtenidos

d. Verifique que los dos vectores

$$a_1 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix} \qquad a_2 = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$$

son ortogonales entre sí. Represente gráficamente estos dos vectores en un gráfico bidimensional y trace rectas desde el origen hasta la ubicación de cada uno de los vectores en el gráfico.

Ejercicio 4:

Sea S la matriz de varianzas y covarianzas poblacionales:

$$S = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

Correspondiente al vector aleatorio X = (X_1,X_2,X_3)' donde:

X_1: puntuación media obtenida en las asignaturas de econometría

X_2: puntuación media obtenida en las asignaturas de derecho

X_3: puntuación media obtenida en asignaturas libres

Para alumnos de la carrera de economía.

a. Calcule los autovalores de la matriz S.
b. Interprete la segunda componente principal sabiendo que el autovector correspondiente: e_1=(0,5744; -0,5744; 0,5744)
c. Como se debería interpretar si un estudiante tuviera segunda una puntuación en la componente principal muy inferior a la de sus compañeros?.
d. ¿Cuántas componentes principales serán necesarias para explicar al menos el 80% de la variabilidad total del conjunto?

Ejercicio 5:

El siguiente conjunto de datos se refiere a 20 observaciones de **suelo**, donde se midió:

x_1: contenido de arena, x_2: contenido de cieno,

x_3: contenido de arcilla, x_4: contenido de materia orgánica

x_5: acidez, según PH.

	x_1	x_2	x_3	x_4	x_5
1	77,3	13,0	9,7	1,5	6,4
2	82,5	10,0	7,5	1,5	6,5
3	66,9	20,6	12,5	2,3	7,0
4	47,2	33,8	19,0	2,8	5,8
5	65,3	20,5	14,2	1,9	6,9
6	83,3	10,0	6,7	2,2	7,0
7	81,6	12,7	5,7	2,9	6,7
8	47,8	36,5	15,7	2,3	7,2
9	48,6	37,1	14,3	2,1	7,2
10	61,6	25,5	12,9	1,9	7,3
11	58,6	26,5	14,9	2,4	6,7
12	69,3	22,3	8,4	4,0	7,0

13	61,8	30,8	7,4	2,7	6,4
14	67,7	25,3	7,0	4,8	7,3
15	57,2	31,2	11,6	2,4	6,5
16	67,2	22,7	10,1	3,3	6,2
17	59,2	31,2	9,6	2,4	6,0
18	80,2	13,2	6,6	2,0	5,8
19	82,2	11,1	6,7	2,2	7,2
20	69,7	20,7	9,6	3,1	5,9

Compare los resultados del Análisis en Componentes Principales para la matriz de covarianza y para la matriz de correlación. Por ejemplo:

a. Los porcentajes de variabilidad que logran explicar cada una de las componentes son los mismos?.
b. Cambia el orden de las componentes?
c. Cambian los loadings de las componentes?
d. Cuál de los dos análisis le parece más adecuado y por qué?.

Ejercicio 6:

La tabla **gorriones.xls** contiene datos de 49 aves, 21 de los cuales sobrevivieron a una tormenta. Le solicitamos:

a. Estandarice las variables y calcule la matriz de covarianzas para las variables estandarizadas.
b. Verifique que ésta es la matriz de correlación de las variables originales.
c. Le parece adecuado en este caso un análisis de componentes principales.
d. ¿Qué indica el autovalor para una componente principal?
e. ¿Cuántas componentes son necesarias para explicar el 80% de la varianza total?
f. Realice el grafico de sedimentación, cuantas componentes elegiría en función de este gráfico.
g. ¿Cómo queda expresada la primer componente principal? (en función del autovector correspondiente y de las variables).
h. ¿Cuál es el valor de λ_3? ¿Qué proporción de la varianza total explica este autovalor?
i. Calcule el tercer autovector.

j. Encuentre las coordenadas del pájaro 11 en las nuevas componentes.

k. Exprese los datos en términos de las nuevas componentes.

l. Represente gráficamente en el plano. (Eje 1 vs 2, 1 vs 3, 2 vs 3).

m. Interprete los tres primeros ejes.

n. Realice un gráfico donde se observen los gorriones en los nuevos ejes 1 y 2, y resalte con distinto color el grupo de los que sobrevivieron.

o. Utilice el Análisis en Componentes Principales como método para encontrar outliers.

Ejercicio 7:

Con el objetivo de oObtener índices útiles para la gestión hospitalaria basados en técnicas estadísticas multivariantes descriptivas se se recogió información del Hospital de Algeciras correspondiente a los ingresos hospitalarios del periodo 2007-2008.
Se estudiaron las variables habitualmente monitorizadas por el Servicio Andaluz de Salud, del Sistema Nacional de Salud Español:

NI: número de ingresos,
MO: mortalidad,
RE: número de reingresos,
NE: número de consultas externas,
ICM: índice *case-mix*, Un índice case-mix superior o inferior a 1 indicará mayor o menor complejidad de los enfermos atendidos en el servicio con respecto al patrón de referencia.
ES: número de estancias
IF: índice funcional.

Las variables se midieron en un total de 22486 ingresos.

En la siguiente tabla se aprecia la DISTRIBUCIÓN DE LOS VALORES OBTENIDOS EN LAS VARIABLES POR LOS SERVICIOS DEL HOSPITAL DE ALGECIRAS, ANDALUCÍA, ESPAÑA:

servicios	NI	MO	RE	NE	ICM	ES	IF
Cirugía	2158	3,8	3,4	8567	1,17	21879	1,05
Tocoginecología	5146	0,3	3,1	3782	0,52	22068	0,87
Hematología	489	4,1	6,8	11005	1,68	4980	0,95
Cardiología	677	2,2	3,9	2161	1,3	8587	0,83
Digestivo	698	5,9	3,2	9473	1,06	7189	1,01
Medicina interna	4171	12,5	5,5	21563	1,04	47909	1,02
Neumología	562	5,1	4,4	2659	1,47	5098	0,68
Otorrinolaringología	650	2,1	2,3	22024	0,87	3161	0,86
Oftalmología	990	0	0,2	21752	0,82	1096	0,5
Pediatría	3752	0,3	2,1	8273	0,51	12152	1
Psiquiatría	622	0	13,3	27000	1,37	6776	0,6
Traumatología	1410	0,7	1,5	13290	1,16	14948	1,17
Urología	1161	2	3,9	4767	0,79	8959	1,01

La idea central del ACP es conseguir la simplificación de un conjunto de datos, generalmente cuantitativos, procedentes de un conjunto de variables interrelacionadas.
Este objetivo se alcanza obteniendo, a partir de combinaciones lineales de las variables originalmente medidas, un nuevo conjunto de igual número de variables, no correlacionadas, llamadas componentes principales (CP) en las cuales permanece la variabilidad presente en los datos originales, y que al ordenarlas decrecientemente por su varianza, nos permiten explicar el fenómeno de estudio con las primeras CP.

a- Verificar que las primeras dos componentes principales son:

Y_1=0.5380 NI+0.5126 ES+0.4081 IF+0.2635 MO-0.1561 NE- 0.2535 RE- 0.3511 ICM.

Y_2=0.5524 MO+0.4952 RE+0.4696 ICM+0.3756 ES+0.2867NE+0.05778 IF- 0.04908 NI.

Grafique las cargas y explicar la interpretación de las componentes principales.

b- Qué porcentaje de variabilidad logra captar cada una de ellas?. Grafique el scree plot.

c- Le parece adecuado considerar dos componentes principales?.

d- Hallar la correlación entre las nuevas variables y las originales.

e- Ordenar los servicios en función de su puntuación en cada una de las dos primeras componentes principales. Indicar cuales son los servicios más demandados y los más complejos.

f- Representar un biplot y buscar servicios similares, asociaciones entre las variables. Verificar en este grafico la representación de las variables originales en las componentes.