

# AID - Trabajo Práctico Nro. II

6 de febrero de 2022

**Ejercicio 1.** Sea la matriz de varianzas y covarianzas poblacionales:

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

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 los loadings de la primer componente principal.
- d) Hallar los scores de 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. 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:*

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

- a) Calcule la matriz de covarianza, los autovalores y autovectores.
- b) Las componentes principales y su contribución porcentual a la varianza total.
- c) Grafique los datos en  $\mathbb{R}^2$  en la base original y en la base de los dos primeros ejes.

- d) Repita los cálculos con los datos estandarizados. Interprete los resultados obtenidos
- e) Verifique que los dos primeros autovectores 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:

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

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

Los datos corresponden a un conjunto de 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 de la tabla 1 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.

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

- 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. Se pide:*

- 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. ¿Qué indica el autovalor para una componente principal?

**Tabla 1:** Suelos

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
1	77.3	13	9.7	1.5	6.4
2	82.5	10	7.5	1.5	6.5
3	66.9	20.6	12.5	2.3	7
4	47.2	33.8	19	2.8	5.8
5	65.3	20.5	14.2	1.9	6.9
6	83.3	10	6.7	2.2	7
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	7
13	61.8	30.8	7.4	2.7	6.4
14	67.7	25.3	7	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
18	80.2	13.2	6.6	2	5.8
19	82.2	11.1	6.7	2.2	7.2
20	69.7	20.7	9.6	3.1	5.9

- d) ¿Cuántas componentes son necesarias para explicar el 80 % de la varianza total? Realice el grafico de sedimentación, fundamente su respuesta con este gráfico.
- e) ¿Cuál es la expresión de la primer componente principal?
- f) ¿Cómo queda expresada la primer componente principal? (en función del autovector correspondiente y de las variables).
- g) Encuentre las coordenadas del pájaro 11 en las nuevas componentes.
- h) Represente gráficamente en el plano. (Eje 1 vs 2, 1 vs 3, 2 vs 3). Interprete los tres primeros ejes.
- i) Realice un gráfico donde se observen los gorrones en los nuevos ejes 1 y 2, y resalte con distinto color el grupo de los que sobrevivieron.
- j) Utilice el Análisis en Componentes Principales como método para encontrar outliers.

**Ejercicio 7.** *Con el objetivo de obtener 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

- NE: número de consultas externas
- ES: número de estancias
- ES: número de estancias

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 listadas por los servicios del hospital de Algeciras( Andalucía, Espana):

**Tabla 2:** Por servicio del Hospital

servicios	NI	MO	RE	NE	ICM	ES	IF
<b>Cirugía</b>	2158	3.8	3.4	8567	1.17	21879	1.05
<b>Tocoginecología</b>	5146	0.3	3.1	3782	0.52	22068	0.87
<b>Hematología</b>	489	4.1	6.8	11005	1.68	4980	0.95
<b>Cardiología</b>	677	2.2	3.9	2161	1.3	8587	0.83
<b>Digestivo</b>	698	5.9	3.2	9473	1.06	7189	1.01
<b>Medicina Interna</b>	4171	12.5	5.5	21563	1.04	47909	1.02
<b>Neumología</b>	562	5.1	4.4	2659	1.47	5098	0.68
<b>Otorrinolaringología</b>	650	2.1	2.3	22024	0.87	3161	0.86
<b>Oftalmología</b>	990	0	0.2	21752	0.82	1096	0.5
<b>Pediatría</b>	3752	0.3	2.1	8273	0.51	12152	1
<b>Psiquiatría</b>	622	0	13.3	27000	1.37	6776	0.6
<b>Traumatología</b>	1410	0.7	1.5	13290	1.16	14948	1.17
<b>Urología</b>	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.

Verificar que las primeras dos componentes principales son:

$$Y_1 = 0,5380NI + 0,5126ES + 0,4081IF + 0,2635MO - 0,1561NE - 0,2535RE - 0,3511ICM$$

$$Y_2 = 0,5524MO + 0,4952RE + 0,4696ICM + 0,3756ES + 0,2867NE + 0,05778IF - 0,04908NI$$

- Grafique las cargas y explicar la interpretación de las componentes principales.
- Qué porcentaje de variabilidad logra captar cada una de ellas?. Grafique el scree plot.
- Le parece adecuado considerar dos componentes principales?.
- Hallar la correlación entre las nuevas variables y las originales.
- 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.
- Representar un biplot y buscar servicios similares, asociaciones entre las variables. Verificar en este grafico la representación de las variables originales en las componentes.