PROBLEMAS TEMA 3: ANÁLISIS DE COMPONENTES PRINCIPALES. LICENCIADO EN ECONOMÍA

Problema 1 Calcular para la matriz de covarianzas

$$S_Y = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix},$$

(a) Las componentes principales Z_1 y Z_2.

(b) La proporción de variabilidad total explicada por la primera componente principal.

(c) Las comunalidades de las dos variables originales, cuando solo se retiene la primera componente principal.

(d) Las correlaciones de las dos variables originales con la primera componente principal.

Problema 2 Convertir la matriz de covarianzas del Problema 1 en una matriz de correlaciones R.

(a) Determinar las componentes principales Z_1 y Z_2 de R.

(b) Calcular la proporción de variabilidad total explicada por Z_1.

(c) Compara las componentes calculadas en (a) con las obtenidas en el Problema 1. ¿Son las mismas? ¿Deberían serlo?

(d) Calcular las comunalidades de las variables originales, cuando solo se retiene la primera componente principal.

(e) Calcular las correlaciones de las dos variables con la primera componente principal.

Problema 3 Sea la matriz de covarianzas

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

Determinar las componentes principales Z_1, Z_2 y Z_3. ¿Qué puedes comentar acerca de los vectores propios (y componentes principales) asociados a valores propios que son iguales?

Problema 4 (Sept. 2004) Sea la matriz de covarianzas

$$S_Y = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

proporcionar:

(a) Componentes principales.

(b) Proporción de variabilidad explicada por cada una de las componentes obtenidas.

c) Comunalidades de las variables originales si solo se retiene una componente principal. ¿Qué variable/s está/n mejor explicada/s?

d) Correlaciones de las variables originales con todas las componentes principales. ¿Qué variable/s representa/n mejor a cada componente?

Problema 5 (Feb. 2005) Con el fin de encontrar un índice económico de comparación de restaurantes españoles, se han medido tres variables sobre 30 restaurantes, X_1 =Gasto en publicidad, X_2 =Gasto en personal y X_3 =Beneficios netos. La matriz de covarianzas de dichas variables es

$$S_X = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

(a) Calcúlese las componentes principales.

(b) ¿Qué proporción de variabilidad total explica cada componente?.

(c) Si solo se desea obtener un índice que represente la información de las tres variables, ¿cuál sería dicho índice? ¿Qué variable es la mejor representada por dicho índice?.

(d) ¿Cómo está de correlado dicho índice con las variables X_1, X_2 y X_3?.

Problema 6 (Sept. 2005) Se midieron las variables X_1:Ventas y X_2:Beneficios (en millones de $) a las 10 mayores corporaciones industriales estadounidenses, y se obtuvieron el siguiente vector de medias y matriz de covarianzas ($\times 10^5$),

$$\bar{X} = \begin{bmatrix} 62{,}309 \\ 2{,}927 \end{bmatrix}, \quad S = \begin{bmatrix} 10005{,}2 & 255{,}76 \\ 255{,}76 & 14{,}30 \end{bmatrix}.$$

(a) Determinar las componentes principales. Observa la diferencia de magnitud entre los coeficientes de X_1 y de X_2. ¿A qué crees que es debido? ¿Qué solución propones a esto?

(b) Proporcionar las varianzas de las componentes principales, y el porcentaje de variabilidad explicada por la primera componente.

(c) Calcular las correlaciones entre la primera componente y las variables X_1 y X_2. ¿Qué interpretación se le puede dar a la primera componente?

Problema 7 (Feb. 2006) Se han medido tres variables relacionadas con la calidad de vida de los ciudadanos en las provincias españolas, X_1 : Índice de precios, X_2 : Índice de densidad del tráfico y X_3 : Superficie media de parques por 1000 viviendas. Se han calculado las varianzas y covarianzas de dichas variables, y se ha obtenido la siguiente matriz

$$S_X = \begin{pmatrix} 3 & -1 & -2 \\ -1 & 3 & -2 \\ -2 & -2 & 4 \end{pmatrix}.$$

(a) Sin hacer cálculos, ¿Cuántas componentes principales hay como máximo? ¿Cómo se definen dichas componentes principales?

(b) Comprueba que los valores propios de la matriz de covarianzas son 0, 4 y 6.

(c) Utilizando el Análisis de Componentes Principales, proporciona un solo índice que represente a las tres variables de calidad de vida lo mejor posible.

(d) De la variabilidad total de las provincias en las tres variables medidas, ¿qué porcentaje recoge/explica el índice obtenido en (c)?

(e) De la variabilidad que presentan las provincias en cada variable por separado, ¿qué porcentaje explica/recoge el índice calculado en (c)?

(f) ¿Cuál es la correlación del índice obtenido en (c) con las tres variables de la calidad de vida? Según los resultados obtenidos ¿valores altos del índice indican mejor calidad o peor calidad de vida?

Problema 8 (Feb. 2007) Se midió el peso (X_1, en Kg.), la estatura (X_2, en cm.), el ancho de hombros (X_3, en cm.) y el ancho de caderas (X_4, en cm.) a $n = 100$ estudiantes de Economía, y se obtuvieron los siguientes resultados:

$$\bar{x}_1 = 54{,}25, \quad \bar{x}_2 = 161{,}73, \quad \bar{x}_3 = 36{,}53, \quad \bar{x}_4 = 30{,}1,$$

$$S_X = \begin{pmatrix} 44{,}7 & 17{,}79 & 5{,}99 & 9{,}19 \\ 17{,}79 & 26{,}15 & 4{,}52 & 4{,}44 \\ 5{,}99 & 4{,}52 & 3{,}33 & 1{,}34 \\ 9{,}19 & 4{,}44 & 1{,}34 & 4{,}56 \end{pmatrix}$$

Valores y vectores propios:

Valores	58.49	15.47	2.54	2.24
Vectores	0.833	0.509	0.188	0.106
	0.503	-0.855	0.020	0.123
	0.136	-0.059	0.111	-0.983
	0.187	0.074	-0.975	-0.089

(a) Seleccionar un número de componentes principales justificando adecuadamente los motivos de dicha elección. Escribir las componentes principales seleccionadas y su porcentaje de variabilidad explicado.

(b) Calcular las comunalidades de las variables e interpretarlas.

(c) Calcular las correlaciones entre las variables originales y las componentes principales.

(d) A partir de las correlaciones obtenidas en (b), interpretar el significado de las componentes principales seleccionadas.

Problema 9 (Feb. 2010) Se desea establecer un ranking entre distintos vinos españoles. Para esto, se valoran tres variables X_1 : Degustación, X_2 : Acidez total (contenido de ácido tartárico, garantía de buena elaboración) y X_3 : Sulfuro total (contenido en anhídrido sulfuroso, necesario para la conservación pero que puede afectar a la calidad). Un comité de expertos ha evaluado estas variables en $n = 60$ vinos españoles, asignando puntuaciones entre 1 y 3 sobre los niveles de estas variables. A partir de estas puntuaciones se ha calculado la siguiente matriz de covarianzas:

$$S_X = \begin{pmatrix} 4/3 & 1/3 & -1/3 \\ 1/3 & 4/3 & -1/3 \\ -1/3 & -1/3 & 4/3 \end{pmatrix}$$

(a) **(1 pto.)** Define las componentes principales de (X_1, X_2, X_3).

(b) **(1 pto.)** Calcula la primera componente principal, que será la variable que se utilizará para establecer el ránking entre los vinos.

(c) **(0.3 ptos.)** ¿Qué porcentaje de variabilidad recoge la primera componente principal?

(d) **(0.7 ptos.)** ¿Qué proporción de varianza de las variables originales está recogida por la primera componente principal? ¿Cuál de las tres está mejor representada por esta componente?

(e) **(0.6 ptos.)** Calcula los coeficientes de correlación entre cada una de las variables originales y la primera componente principal.

(f) **(0.4 ptos.)** En función de los resultados del apartado (e), ¿Cómo se caracterizan los vinos que ocuparán las primeras posiciones del ránking? ¿Y los que ocuparán las últimas posiciones?

Problema 10 (Sept. 2010) La siguiente tabla muestra los estadísticos descriptivos básicos referentes a las variables: X_1: ratio inmovilizado/activo (inmoact), X_2: ratio ventas/activos (ventact) y X_3: ratio ventas/inmovilizado (ventim) tomadas a las 229 empresas españolas con mayores ventas:

Variable	Media	Desv. típica
inmoact	0.481	0.248
ventact	1.477	1.270
ventinm	5.882	10.610

(a) **(0.5 ptos.)** A la vista de la información disponible sobre estas tres variables, justifica adecuadamente por qué sería más conveniente hacer un Análisis de Componentes Principales a partir de la matriz de correlaciones en lugar de la de covarianzas.

(b) **(0.5 pto.)** Si las covarianzas entre cada par de variables son $S^2_{X_1,X_2} = -0{,}140$, $S^2_{X_1,X_3} = -1{,}359$ y $S^2_{X_2,X_3} = 8{,}491$, escribe las matrices de covarianzas y de correlaciones entre las tres variables.

(c) **(0.5 ptos.)** Sabiendo que el mayor de los valores propios de la matriz de correlaciones es $\lambda_1 = 2{,}065$, calcula los otros valores propios.

(d) **(0.5 ptos.)** Los vectores propios asociados a cada valor propio de la matriz de correlaciones vienen dados en la siguiente tabla:

$\lambda_1 = 2{,}065$	λ_2	λ_3
$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$
-0.539	0.824	0.177
0.585	0.516	-0.625
0.607	0.234	0.760

Escribe las componentes principales de (X_1, X_2, X_3).

(e) **(0.5 ptos.)** Escribe las varianzas de las componentes principales. Utilizando los tres criterios estudiados, dí con cuántas componentes te quedarías, justificando adecuadamente tu respuesta.

(f) **(1 pto.)** ¿Cuál de las tres variables originales está mejor representada por la primera componente? Justifica tu respuesta.

(g) **(1 pto.)** Calcula los coeficientes de correlación entre cada una de las variables originales y la primera componente principal.

(h) **(0.5 ptos.)** En función de los resultados del apartado anterior, ¿cómo se caracterizan las empresas con mayores coordenadas positivas en la primera componente principal? ¿Y cómo se caracterizan las empresas que tienen las menores coordenadas negativas? ¿Puedes dar una interpretación a la primera componente principal?