Análisis de Datos — Curso 2007/2008
Licenciatura en Biología

Técnicas de análisis para la reducción de la dimensión

1. Sean dos variables aleatorias de media 0 y matriz de varianzas y covarianzas igual a:

$$\mathbf{S} = \begin{bmatrix} 1+d & 1 \\ 1 & 1+d \end{bmatrix},$$

donde d es una constante positiva conocida. Se pide:

(a) Calcular los autovalores de la matriz $\mathbf{S}$.

(b) Calcular las componentes principales.

(c) Calcular la proporción de variabilidad explicada por cada componente principal.

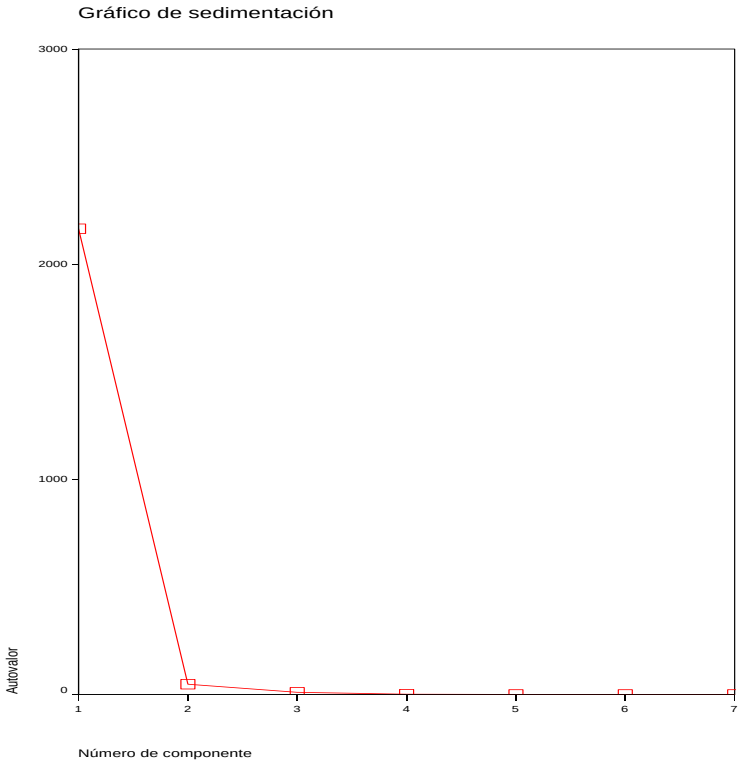
(d) Calcular las correlaciones entre las componentes y las variables originales.

(e) Interpretar las componentes en función de d.

2. En las siguientes salidas de SPSS presentamos un análisis de componentes principales de los datos del fichero `limes.xls` que contiene las siguientes medidas de limas de Tahiti: día de recolección, diámetro del fruto, tamaño del fruto, peso del fruto, volumen del fruto, volumen del zumo, peso del zumo y peso de la cáscara.

Varianza total explicada

		Autovalores iniciales		
	Componente	Total	% de la varianza	% acumulado
Bruta	1	2164,939	97,373	97,373
	2	47,024	2,115	99,488
	3	10,351	,466	99,953
	4	,911	4,099E-02	99,994
	5	7,107E-02	3,196E-03	99,997
	6	3,668E-02	1,650E-03	99,999
	7	2,339E-02	1,052E-03	100,000



(a) ¿Cuántas componentes principales se deben utilizar si queremos que expliquen más del 90 % de la variabilidad total?

(b) Calcule los vectores propios que definen las componentes principales del apartado (a).

(c) Interprete las componentes principales obtenidas en apartado (b).

(d) (Ejercicio de ordenador) Repita los apartados (a, b y c) con el fichero `limes.xls` para el análisis de componentes principales normado.

Matriz de componentes

	Bruta						
	Componente						
	1	2	3	4	5	6	7
Diametro fruto	,638	-,008	,071	,007	-,011	,145	,099
Tamaño fruto	,726	-,064	-,013	,003	,265	-,007	,017
Peso fruto	29,199	-1,221	2,356	-,185	-,002	-,003	-,003
Volumen fruto	31,069	-2,479	-2,088	,040	-,004	,001	,000
Volumen zumo	12,987	4,383	-,197	,090	,020	,088	-,082
Peso zumo	13,090	4,418	-,216	,053	-,020	-,088	,081
Peso cascara	2,475	-,812	,591	,930	-,002	-,005	,002

3. En una empresa se ha realizado un reconocimiento médico a 420 trabajadores. El reconocimiento incluía un análisis en el que se han medido las concentraciones en sangre (expresadas

en mg/dl) de las siguientes sustancias: glucosa (x_1), dos indicadores del funcionamiento renal (creatinina y ácido úrico: x_2 y x_3), dos indicadores de lípidos sanguíneos (colesterol total y triglicéridos x_4 y x_5) y un indicador del funcionamiento hepático (bilirrubina, x_6). Las medias y varianzas muestrales de estas variables se indican en la siguiente tabla:

	x_1	x_2	x_3	x_4	x_5	x_6
$\bar{x}_i$	85	0,8	5,5	175	70	0,6
s_i^2	256	0,09	6,25	1600	900	0,09

(a) Supongamos que se toma el vector de variables $(x_1, \ldots, x_6)$ como un indicador del estado de salud ¿Cómo se puede definir una medida de la distancia de cada individuo al núcleo central de la muestra? Una vez definida esta medida ¿cómo se podrían identificar los individuos "atípicos"?

(b) Se ha realizado un análisis de componentes principales a partir de estos datos,utilizando para ello la matriz de correlaciones (en lugar de la matriz inicial de covarianzas), lo que equivale a utilizar variables tipificadas (es decir, divididas por su desviación típica para que tengan varianza 1). Se han obtenido los siguientes autovalores:

$$\lambda_1^R = 1,4 \quad \lambda_2^R = 1,25 \quad \lambda_3^R = 1,2 \quad \lambda_4^R = 1,1 \quad \lambda_5^R = 0,9 \quad \lambda_6^R = 0,15$$

Calcule el porcentaje de variabilidad explicado por cada componente principal.

(c) Comente brevemente las posibles ventajas o inconvenientes que presentaría el uso de la metodología de componentes principales en este caso. Como conclusión, ¿aconsejaría la utilización de esta técnica?

(d) ¿Cuál es el valor de la covarianza entre las dos primeras componentes principales? Si cree que esta cantidad no puede calcularse con la información disponible, indique qué información adicional necesitaría.

(e) Se ha realizado un análisis factorial de estos datos, con un única variable latente (o factor). En este análisis se han considerado también las variables tipificadas (y, por tanto, se ha utilizado la matriz de correlaciones). Se han obtenido los siguientes resultados:

Variable	x_1	x_2	x_3	x_4	x_5	x_6
Comunalidad ($h_j^2 = \lambda_{j1}^2$)	0,2	0,3	0,32	0,2	0,35	0,4

¿Qué porción de la variabilidad total está explicada por la variable latente?

(f) ¿Apoyan estos resultados la idea de que se puede explicar globalmente el estado de salud de un individuo mediante una única variable latente?

4. Se incluyen a continuación los datos correspondientes al análisis de 315 personas, en cada una de las cuales se han medido 8 variables: edad, índice de Quetelet (índice de masa corporal), y las cantidades diarias de ingesta de calorías, grasa, fibra, colesterol, retinol (vitamina A) y betacaroteno (un derivado de la vitamina A).

Estadísticos descriptivos

		edad	quetelet	calorias	grasa	fibra	colester	betadiet	retdiet
N		315	315	315	315	315	315	315	315
Media		50,146	26,1574	1796,6546	77,0333	12,7886	242,4606	2185,6032	832,7143
Mediana		48,000	24,7353	1666,8000	72,9000	12,1000	206,3000	1802,0000	707,0000
Desv. típ.		14,5752	6,01355	680,34743	33,82944	5,33019	131,9916	1473,88655	589,28903
Varianza		212,437	36,163	462872,632	1144,431	28,411	17421,78	2172341,55	347261,561
Percentiles	25	39,000	21,7885	1333,8000	53,9000	9,1000	154,9000	1114,0000	479,0000
	50	48,000	24,7353	1666,8000	72,9000	12,1000	206,3000	1802,0000	707,0000
	75	63,000	28,9498	2106,4000	95,3000	15,6000	308,9000	2863,0000	1047,0000

Matriz de correlaciones

		edad	quetelet	calorias	grasa	fibra	colester	betadiet	retdiet
Correlación	edad	1,000	-,017	-,177	-,169	,045	-,114	,072	-,010
	quetelet	-,017	1,000	,004	,049	-,088	,110	-,007	,032
	calorias	-,177	,004	1,000	,872	,465	,659	,243	,402
	grasa	-,169	,049	,872	1,000	,276	,710	,143	,412
	fibra	,045	-,088	,465	,276	1,000	,154	,483	,215
	colester	-,114	,110	,659	,710	,154	1,000	,116	,443
	betadiet	,072	-,007	,243	,143	,483	,116	1,000	,053
	retdiet	-,010	,032	,402	,412	,215	,443	,053	1,000

Análisis factorial

Matriz factorial(a)

	Factor 1	Factor 2
edad	,044	-,209
quetelet	-,087	,080
calorias	,471	,808
grasa	,282	,913
fibra	,999	-,006
colester	,159	,731
betadiet	,483	,013
retdiet	,217	,393

Método de extracción: Máxima verosimilitud.
a 2 factores extraídos. Requeridas 9 iteraciones.

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
28,025	13	,009

Varianza total explicada

	Sumas de las saturaciones al cuadrado de la extracción		
Factor	Total	% de la varianza	% acumulado
1	1,616	20,194	20,194
2	2,224	27,806	48,000

Método de extracción: Máxima verosimilitud.

Análisis de componentes principales

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,065	38,315	38,315	3,065	38,315	38,315
2	1,384	17,302	55,617	1,384	17,302	55,617
3	1,009	12,612	68,229	1,009	12,612	68,229
4	,944	11,806	80,035			
5	,671	8,389	88,424			
6	,521	6,511	94,935			
7	,303	3,782	98,717			
8	,103	1,283	100,000			

Método de extracción: Análisis de Componentes principales.

Matriz de componentes(a)

	Componente 1	Componente 2	Componente 3
edad	-,176	,416	,553
quetelet	,053	-,304	,800
calorias	,922	-,022	-,099
grasa	,892	-,201	-,049
fibra	,523	,668	-,046
colester	,799	-,290	,102
betadiet	,347	,727	,114
retdiet	,594	-,136	,158

Método de extracción: Análisis de componentes principales.
a 3 componentes extraídos

Determina razonadamente si son ciertas o falsas las siguientes afirmaciones:

(a) Entre las personas observadas hay 157 con una edad superior a 50.146 años.

(b) El porcentaje de personas con obesidad severa (índice de Quetelet superior a 40) excede el 25 %.

(c) El grado de asociación lineal entre las diferentes variables observadas es bastante bajo, con la excepción del par grasa-colesterol.

(d) El modelo de análisis factorial con dos factores se ajusta bien a los datos pero el porcentaje de variabilidad explicada con ambos factores es inferior al 50 %.

5. En un estudio médico se han tomado datos de 768 mujeres mayores de 21 años de una comunidad india en Arizona (Estados Unidos). Para cada mujer se han medido las siguientes variables:

x_1: Concentración plasmática de glucosa al cabo de dos horas en un test de tolerancia a la glucosa

x_2: Presión sanguínea diastólica

x_3: Grosor de la piel en el pliegue del triceps

x_4: Concentración de insulina al cabo de dos horas en un test de tolerancia a la glucosa

x_5: Índice de masa corporal

x_6: Valor de "pedigree diabético" (evaluación de factores hereditarios relacionados con la diabetes)

x_7: Edad

Se ha realizado un análisis exploratorio de estas variables, junto con un análisis de componentes principales (utilizando en este último la matriz de correlaciones). Al final de este enunciado se incluyen algunos de los resultados obtenidos. Teniendo en cuanta estos resultados, responder razonadamente a las siguientes preguntas:

(a) Calcular la varianza de la primera componente principal y su covarianza con la variable x_5.

(b) A partir de la información disponible ¿parece razonable suponer la normalidad en la variable vectorial $(x_1, \ldots, x_7)$?

(c) ¿Parece aconsejable en este caso utilizar las componentes principales como método para reducir la dimensión?

(d) ¿Cuál es la suma de las varianzas de las componentes principales? Explica por qué se obtendrá este valor siempre que se realice un análisis de componentes principales (basado en la matriz de correlaciones) con 7 variables.

Estadísticos

		Glucosa	Diastolica	Piel	Insulina	IMC	Herencia	Edad
N	Válidos	768	768	768	768	768	768	768
	Perdidos	0	0	0	0	0	0	0
Media		120,8945	69,1055	20,5365	79,7995	31,9926	0,4719	33,2409
Mediana		117,0000	72,0000	23,0000	30,5000	32,0000	0,3725	29,0000
Desv. típ.		31,97262	19,35581	15,95222	115,24400	7,88416	0,33133	11,76023

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,083	29,754	29,754	2,083	29,754	29,754
2	1,320	18,862	48,616	1,320	18,862	48,616
3	1,029	14,707	63,323	1,029	14,707	63,323
4	,873	12,472	75,795	,873	12,472	75,795
5	,697	9,960	85,755	,697	9,960	85,755
6	,592	8,458	94,214	,592	8,458	94,214
7	,405	5,786	100,000	,405	5,786	100,000

Método de extracción: Análisis de Componentes principales.

Matriz de componentes[a]

	Componente						
	1	2	3	4	5	6	7
Glucosa	,540	,425	,462	-,321	-,274	-,246	-,276
Diastolica	,490	,397	-,555	,109	,316	-,427	,028
Piel	,683	-,455	-,229	-,023	,201	,290	-,386
Insulina	,666	-,274	,347	-,364	,316	,012	,359
IMC	,666	-,043	-,367	,062	-,592	,133	,218
Herencia	,409	-,118	,443	,782	,022	-,096	,015
Edad	,177	,827	,063	,094	,177	,488	,048

Método de extracción: Análisis de componentes principales.

a. 7 componentes extraídos

6. Se han obtenido 9 medidas morfológicas correspondientes a 42 comadrejas hembra capturadas en diferentes lugares de Australia (*Australian Journal of Zoology* 43, 449-458, 1995). Las variables consideradas son:

Nombre	Descripción
`cabeza`	longitud de la cabeza
`craneo`	anchura del cráneo
`longitud`	longitud total
`cola`	longitud de la cola
`pie`	longitud del pie
`oreja`	longitud de la oreja
`ojo`	distancia del canto medio al canto lateral
`pecho`	perímetro del pecho
`vientre`	perímetro del vientre

A los datos estandarizados se les ha ajustado un modelo factorial con dos factores. La correspondiente salida de `SPSS`, junto con una matriz de diagramas de dispersión para algunas de las variables, aparece en el Anexo 1. (Algunas de las cifras de la salida han sido sustituidas por letras). Responde a las preguntas siguientes:

(a) A la vista de los diagramas de dispersión, determina si las siguientes afirmaciones son verdaderas o falsas: (a1) No hay datos atípicos. (a2) Todas las correlaciones entre las variables tienden a ser positivas aunque no demasiado próximas a uno.

(b) Calcula los valores A y B que se han omitido en la salida.

(c) ¿Cuánto vale la correlación entre la longitud total y el primer factor? ¿Es aceptable, al nivel de significación 0,01, la validez del modelo factorial considerado?

(d) Con la información disponible calcula un valor estimado para la correlación entre las variables `cabeza` y `longitud`.

Matriz factorial[a,b]

	Factor	
	1	2
CABEZA	,566	,610
CRANEO	,374	,510
LONGITUD	,593	,684
COLA	-,035	,781
PIE	,946	-,177
OREJA	,802	-,430
OJO	,049	,448
PECHO	,550	,349
VIENTRE	,278	,630

Método de extracción: Máxima verosimilitud.

a. 2 factores extraídos. Requeridas 6 iteraciones.

Prueba de la bondad de ajuste[a]

Chi-cuadrado	gl	Sig.
35,627	19	,012

Varianza total explicada[a]

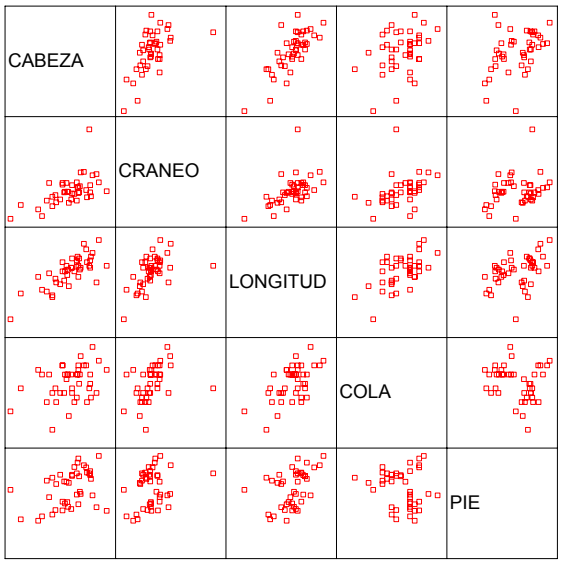
	Sumas de las saturaciones al cuadrado de la extracción		
Factor	Total	% de la varianza	% acumulado
1	2,735	A	30,393
2	2,645	29,392	59,785

Método de extracción: Máxima verosimilitud.

Comunalidades[a]

	Extracción
CABEZA	B
CRANEO	,399
LONGITUD	,820
COLA	,611
PIE	,927
OREJA	,828
OJO	,203
PECHO	,425
VIENTRE	,474

Método de extracción: Máxima verosimilitud.



7. Se ha realizado un análisis de componentes principales, utilizando la matriz de correlaciones, para los datos correspondientes a los records de atletismo conseguidos por 55 países en 8 especialidades $(X_1, \ldots, X_8)$: 100 m. (en seg.), 200 m. (en seg.), 400 m. (en s.), 800 m. (en min.), 1500 m. (min.), 5000 m. (min.), 10000 m. (min.), maratón (min.). Las salidas SPSS del análisis pueden verse a continuación:

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
X1	10,4711	,35143	55
X2	21,1038	1,37541	55
X3	46,4387	1,45702	55
X4	1,7933	,06368	55
X5	3,6982	,15591	55
X6	13,8278	,79173	55
X7	28,9967	1,80169	55
X8	136,6240	9,22703	55

Varianza total explicada

Componente	Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado
1	6,004	75,053	75,053
2	1,039	12,988	88,041
3	**A1**	**A2**	**A3**
4	,138	1,723	96,658
5	,109	1,359	98,017
6	,077	,962	98,979
7	,057	,714	99,693
8	,025	,307	100,000

Matriz de correlaciones

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1,000	,449	,841	,756	,700	,606	,635	,520
X2	,449	1,000	,287	,277	,303	,246	,233	,180
X3	,841	,287	1,000	,870	,835	,791	,785	,705
X4	,756	,277	,870	1,000	,918	,868	,868	,806
X5	,700	,303	,835	,918	1,000	,919	,934	,866
X6	,606	,246	,791	,868	,919	1,000	,964	,921
X7	,635	,233	,785	,868	,934	,964	1,000	,943
X8	,520	,180	,705	,806	,866	,921	,943	1,000

Matriz de componentes

	Componente							
	1	2	3	4	5	6	7	8
X1	,791	,389	-,413	**B**	-,140	-,014	-,019	,022
X2	,357	,854	,377	-,025	,034	,016	,001	-,006
X3	,904	,089	-,322	-,023	,260	,014	,040	-,016
X4	,947	-,027	-,115	-,237	-,100	,137	-,063	-,019
X5	,964	-,070	,052	-,131	-,074	-,142	,139	,038
X6	,946	-,183	,173	,019	,062	-,085	-,158	,069
X7	,954	-,190	,156	,088	-,033	-,066	-,019	-,125
X8	,896	-,275	,255	,152	,001	,160	,080	,039

(a) Calcular el valor que aparece indicado con la letra B en la tabla titulada "Matriz de componentes".

(b) ¿Cuántas componentes habría que utilizar para explicar al menos el 80 % de la varianza?. ¿Cómo pueden interpretarse estas componentes?

(c) ¿Cuánto vale la varianza de la segunda componente principal? Calcula los valores que aparecen indicados con las letras A1, A2 y A3 en la tabla titulada "Varianza total explicada".

(d) ¿Cuál es el valor de la primera componente principal para un país cuyos records son $(10,31 \;\; 20,06 \;\; 44,84 \;\; 1,74 \;\; 3,57 \;\; 13,28 \;\; 27,66 \;\; 128,30)$?

(e) ¿Cuánto vale la covarianza entre la primera componente principal y la variable X_3?