Ejemplo:

Se dispone de las calificaciones en 9 asignaturas de los 29 alumnos de un curso, según se indica a continuación:

Lista de variables

Estadística 1 (ST1), Estadística 2 (ST2), Estadística 3 (ST3),
Investigación Operativa (IOP), Informática (INF), Matemáticas (MAT),
Economía (ECO), Gestión (GES) e Inglés (ING).

Listado de casos

	ST1	ST2	GES	ST3	IOP	INF	MAT	ECO	ING
1	,3	,3	1,0	,0	1,7	,6	,6	,6	,3
2	3,4	2,0	,3	1,0	1,0	,3	,3	,3	,6
3	2,4	2,7	,6	1,2	1,3	1,3	1,0	1,3	1,0
4	1,0	,6	1,7	,6	3,1	1,0	1,7	1,0	2,0
5	1,7	3,1	2,4	2,9	5,5	2,4	3,1	1,7	1,3
6	,6	1,0	1,3	,0	,3	1,7	2,4	3,1	3,7
7	2,0	1,7	3,7	,0	,6	2,0	2,0	5,1	5,1
8	5,8	1,3	3,1	2,3	4,4	3,7	3,7	4,8	4,4
9	5,1	6,2	2,0	4,8	5,8	3,1	5,8	3,4	3,1
10	4,1	5,8	3,4	4,0	5,1	4,8	1,3	5,5	5,5
11	3,7	2,4	4,8	3,4	6,2	4,1	4,8	5,8	6,2
12	1,3	3,4	5,1	1,1	2,0	5,5	2,7	6,5	6,5
13	5,5	4,1	2,7	3,0	3,7	5,8	4,4	7,5	7,2
14	6,8	7,5	6,2	6,3	7,5	2,7	3,4	2,7	1,7
15	6,2	6,5	4,1	4,2	2,4	3,4	5,5	6,2	5,8
16	2,7	5,1	7,5	2,8	4,1	6,5	6,5	6,8	6,8
17	3,1	4,4	5,8	2,8	3,4	6,8	5,1	8,6	8,2
18	4,8	5,5	4,4	3,8	4,8	7,5	6,8	8,2	8,6
19	7,9	7,9	7,9	6,9	8,2	5,1	6,2	2,4	2,4
20	7,2	7,2	8,6	5,9	7,2	8,9	7,5	7,9	7,9
21	6,5	6,8	8,2	4,8	2,7	8,6	9,6	8,9	9,6
22	4,4	4,8	5,5	5,0	6,5	8,2	8,6	9,6	10,0
23	7,5	3,7	9,3	4,9	6,8	9,6	8,2	10,0	9,3
24	8,9	8,9	6,8	7,8	8,9	6,2	4,1	2,0	2,7
25	8,2	9,3	6,5	7,8	9,3	4,4	7,2	4,4	4,1
26	9,3	8,6	7,2	7,4	7,9	7,2	7,9	4,1	3,4
27	9,6	9,6	8,9	9,2	10,0	7,9	8,9	3,7	4,8
28	8,6	8,2	10,0	7,2	8,6	10,0	9,3	9,3	8,9
29	10,0	10,0	9,6	8,7	9,6	9,3	10,0	7,2	7,5

Si acudimos a los estadísticos descriptivos básicos de estas calificaciones, observaremos que las notas (a excepción de las de ST3) se encuentran estandarizadas a media 5,124 y desviación típica 2,904, siendo, por tanto, las dispersiones originales en cada variable del mismo orden. En esta situación, el ACP va a centrarse en simplificar claramente la diversidad de las notas casi exclusivamente a partir de las relaciones de dependencia existentes entre las variables originales. Nótese que en otras situaciones en las que las diferencias de varianzas son notables (por ejemplo, cuando se deben a que los datos se miden en distintas escalas , o cuando las variables son de muy diferente naturaleza), estas diferencias de varianza también van a ser explicadas por las CP.

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
ST1	5,124	2,9404	29
ST2	5,124	2,9404	29
GES	5,124	2,9404	29
ST3	4,131	2,7577	29
IOP	5,124	2,9404	29
INF	5,124	2,9404	29
MAT	5,124	2,9404	29
ECO	5,124	2,9404	29
ING	5,124	2,9404	29

Por su parte, la estructura de relaciones entre las 9 variables observadas sobre estos 29 individuos puede examinarse inicialmente a través de su matriz de correlaciones de Pearson, en la que se puede observar la estrecha relación (alta correlación) existente entre algunas de estas variables, lo que induce a pensar en la posibilidad de representar la información en un espacio más reducido y claro, como nos lo va a proporcionar la aplicación del ACP.

Correlaciones de Pearson

	ST1	ST2	GES	ST3	IOP	INF	MAT	ECO	ING
ST1	1	,872	,757	,936	,828	,659	,738	,286	,273
ST2	,872	1	,750	,945	,804	,634	,713	,255	,252
GES	,757	,750	1	,783	,726	,869	,844	,586	,586
ST3	,936	,945	,783	1	,932	,666	,763	,248	,247
IOP	,828	,804	,726	,932	1	,600	,669	,179	,173
INF	,659	,634	,869	,666	,600	1	,875	,809	,818
MAT	,738	,713	,844	,763	,669	,875	1	,673	,684
ECO	,286	,255	,586	,248	,179	,809	,673	1	,985
ING	,273	,252	,586	,247	,173	,818	,684	,985	1

El análisis de Componentes Principales aplicado a este caso arroja el siguiente resultado sobre las varianzas explicadas por cada nueva componente

Varianza total explicada por el ACP

		Autovalores iniciales(a)		
		Total	% de la varianza	% acumulado
Componente	1	53,618	69,838	69,838
	2	16,864	21,966	91,804
	3	1,854	2,415	94,219
	4	1,460	1,902	96,121
	5	1,220	1,589	97,710
	6	1,068	1,391	99,100
	7	,529	,689	99,789
	8	,121	,157	99,947
	9	,041	,053	100,000

Observemos que todos los autovalores de la matriz de varianzas y covarianzas de los datos son estrictamente positivos y no hay ninguno que sea igual a cero; lo que significa que todas

las variables aportan algo de información. Es decir, que no hay ninguna que sea absolutamente despreciable porque todas tienen algo de información nueva que no tenían alguna de las anteriores.

Sin embargo, las magnitudes de los autovalores son comparativamente muy diferentes. Así que, como cada autovalor indica la varianza que explica la correspondiente componente principal, la primera componente, cuyo autovalor 53,618 es máximo, reproduce del orden de 100 veces diversidad que la séptima componente, por ejemplo, cuyo autovalor es solo 0,529. Luego parece lógico pensar en una posible reducción de la dimensión del problema y retener sólo un número reducido de las primeras componentes principales. Para ello, recurramos a los criterios estudiados.

Si sumamos todos los autovalores, obtendremos la suma de todas las varianzas, y que coincide con la suma de las varianzas de las variables originales. Como no nos interesa tanto la cantidad absoluta, sino la que relativamente explica cada componente principal, al lado de los autovalores aparece el porcentaje del total de la varianza que explica la componente correspondiente. De esta forma, la primera componente principal explica un 69,838% de la suma total de varianzas (proporción del 53,618 frente a la suma de todos los autovalores).Es decir, más de la mitad de la variabilidad de los datos se debe a esa dimensión primera que estamos analizando. La segunda es un 21,966%, y las restantes explican sustancialmente menos de la dispersión total.

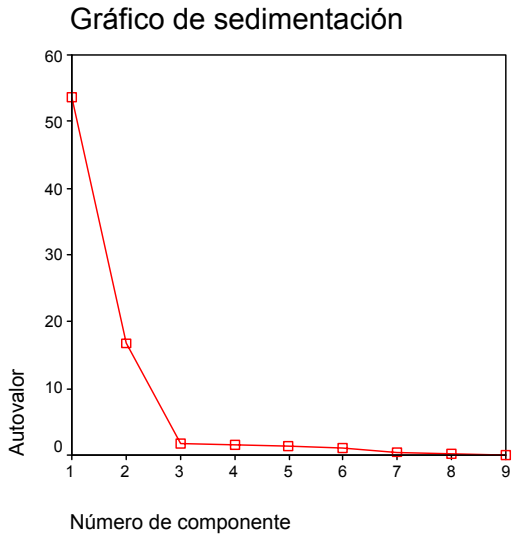
Como las componentes están ordenadas de más explicativas a menos explicativas, podemos acumular fácilmente la cantidad de dispersión total que explican entre las primeras componentes consideradas, lo que aparece en la columna siguiente, donde podemos observar que entre las dos primeras componentes principales explican el 91,804% (69,838% de la primera CP +21,966% de la segunda).

Observemos que si nos quedamos con estas dos primeras CP, aunque perdamos algo de información, conseguimos retener más del 90% de la diversidad de los datos originales, con prácticamente la quinta parte de variables (2 CP frente a las 9 variables originales); lo que desde un punto de vista práctico es valioso tanto por la simplicidad de análisis como de cómputo que provee.

Si acudimos a los criterios de selección estudiados, el criterio de la media exige comparar los autovalores con la media de los autovalores, que puede comprobarse es $\overline{\lambda} = 8{,}5306$. Aún en el caso que recurramos a la corrección de rebajar este umbral, por ejemplo a su 80%, esto es a $0.80 \cdot \overline{\lambda} = 6{,}8244$, sólo hay dos autovalores mayores que la unidad: los dos primeros.

Si acudimos al criterio de la proporción conveniente de la varianza explicada, y establecemos ésta en un 80% ó 90% (proporciones que se consideran en este ámbito bastante exigentes), observamos que se requieren esas dos mismas primeras componentes para explicar al menos esa diversidad (91,804% de la varianza total).

Y si finalmente recurrimos al gráfico de sedimentación, observamos que la zona de erosión es abrupta para las dos primeras componentes principales, comenzando el valle a partir de la tercera componente; lo que nos vuelve a sugerir que retengamos sólo las dos primeras.



Así que cualquiera de los criterios nos confirma que hay dos componentes principales que explican sustancialmente más que las variables originales y con las que reproducimos bastante bien la diversidad inicial.

En la siguiente tabla se presentan los coeficientes de las 9 componentes principales. Sus columnas marcan, por tanto, la dirección de los autovectores de la matriz de varianzas y covarianzas, y han sido normalizados para que su módulo sea justamente la dispersión típica explicada; es decir, la raíz cuadrada de su autovalor asociado.

Matriz de componentes

	Componente								
	1	2	3	4	5	6	7	8	9
ST1	2,539	-1,128	,487	,167	-,031	-,812	,026	,005	-,039
ST2	2,491	-1,197	,737	-,284	,298	,540	-,005	-,033	-,063
GES	2,726	,072	-,558	-,901	,075	-,157	,238	,017	,005
ST3	2,448	-1,229	,110	,186	,058	,116	,027	,079	,169
IOP	2,372	-1,383	-,806	,614	,204	,160	,094	-,020	-,062
INF	2,700	,938	-,223	-,096	,149	-,062	-,627	-,040	,007
MAT	2,734	,395	,038	,067	-,983	,212	,021	-,028	-,016
ECO	1,892	2,191	,171	,256	,254	-,027	,245	-,217	,036
ING	1,891	2,213	,141	,223	,164	,063	,095	,252	-,035

Así, expresadas como combinaciones lineales de las variables originales, las 2 primeras componentes principales toman las expresiones (el programa muestra las de módulo igual a su desviación típica y no las de módulo unidad; lo que no tiene mayor importancia ya que es otra forma de resolver la indeterminación entre los vectores de una misma dirección):

$$\sqrt{\lambda_1}Z_1 = \sqrt{\lambda_1}(u_{11}X_1 + \cdots + u_{p1}X_p) = 2.539 \cdot ST1 + 2.491 \cdot ST2 + \cdots + 1.891 \cdot ING$$
$$\sqrt{\lambda_2}Z_2 = \sqrt{\lambda_2}(u_{12}X_1 + \cdots + u_{p2}X_p) = -1.128 \cdot ST1 - 1.197 \cdot ST2 + \cdots + 2.213 \cdot ING$$

a partir de las cuales, conocidos los valores de los individuos para las 9 asignaturas, podemos calcular las puntuaciones o valores para las componentes.

Y para interpretar el significado de las componentes, debemos recurrir a su matriz de estructura (o de correlaciones entre variables y componentes) que se reproduce a continuación.

Fijándonos en las correlaciones con las dos primeras componentes que hemos retenido, observamos que la primera componente presenta correlaciones altas y positivas con todas las variables, si bien algo menores para las asignaturas de Economía e Inglés (justo las asignaturas con menor carga numérico-matemática). Por su parte, la segunda componente presenta correlaciones positivas y altas con Economía e Inglés, también positivas aunque más bajas con la Informática y las Matemáticas, prácticamente nula con la Gestión, y negativas con todas las materias de aplicación matemática (Estadístcas e Investigación Operativa)

Matriz de estructura

	Componente								
	1	2	3	4	5	6	7	8	9
ST1	,863	-,384	,166	,057	-,011	-,276	,009	,002	-,013
ST2	,847	-,407	,251	-,097	,101	,184	-,002	-,011	-,021
GES	,927	,025	-,190	-,306	,025	-,053	,081	,006	,002
ST3	,888	-,446	,040	,067	,021	,042	,010	,029	,061
IOP	,807	-,470	-,274	,209	,070	,054	,032	-,007	-,021
INF	,918	,319	-,076	-,033	,051	-,021	-,213	-,014	,002
MAT	,930	,134	,013	,023	-,334	,072	,007	-,010	-,005
ECO	,643	,745	,058	,087	,086	-,009	,083	-,074	,012
ING	,643	,753	,048	,076	,056	,022	,032	,086	-,012

Parece pues que los individuos que presentan una puntuación alta en la componente principal primera son los que suelen tener también puntuaciones altas en todas las asignaturas, aunque especialmente en las de carácter matemático (abstracto o aplicado), y viceversa. Parece, por tanto, que los dos sentidos de la dirección de la primera componente divide a los casos entre los que tienen cierta capacidad o habilidad para sacar buenas notas en general y los que no. ¿inteligencia? ¿preparación? ¿esfuerzo?

Por otro lado, los individuos que presentan una puntuación alta para la componente segunda son los que suelen presentar frecuentemente puntuaciones también alta en Economía e Inglés y algo menos comúnmente en Informática y Matemáticas, así como los que sacan bajas notas en las materias de aplicación matemática como las Estadísticas y la Investigación Operativa, y viceversa. Parece por tanto que los dos sentidos de la dirección de la segunda componente divide a los casos entre los que tienen cierta capacidad o habilidad para sacar buenas notas en materias que requieren más razonamiento de tipo lingüístico (obsérvese que la programación de ordenadores y las matemáticas aunque con mayor requerimiento de cálculo abstracto, precisan claramente de lenguajes propios) frente a los que parecen tener más capacidad o habilidad para sacar buenas notas en materias que requieren más razonamientos de tipo cálculo numérico. ¿inteligencia, capacidad o habilidad en lenguaje versus cálculo numérico?

Nótese que la interpretación precisa de las componentes requiere un conocimiento profundo de los contenidos de las asignaturas y de los factores psicológicos que pueden causar las diferencias o comportamientos diversos. Es esta relación causa-efecto entre componentes principales y variables observadas la que de alguna forma motiva la generación de las técnicas de Análisis Factorial que se abordan en otro capítulo.