Análisis Multivariante:

Análisis de Componentes Principales

Edith Johana Medina Hernández – edithjoh@gmail.com

Análisis Multivariado Enfocado a la Gestión de Riesgos

- El Análisis de Componentes Principales es una técnica de reducción de la dimensión que describe la información de un conjunto de variables observadas mediante un conjunto de variables más pequeño (las componentes principales).

- Las nuevas variables (componentes principales) son incorreladas, y se obtienen en orden decreciente de importancia.

- Esperamos que sólo unas pocas recojan la mayor parte de la información de los datos.

- La transformación es, en realidad, una rotación ortogonal en el espacio p-dimensional.

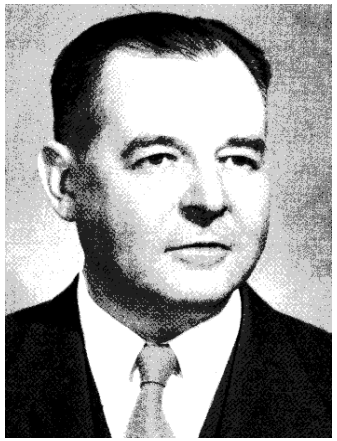
- El ACP puede entenderse, entonces, como la búsqueda del subespacio de mejor ajuste a los datos.

Análisis de Componentes Principales (PCA ó ACP)

Antecedentes

Pearson, K. (1901). On lines and planes of closets fit to systems of points in the space. *Philosophical Magazine*, **2**: 559-572.

Pearson trata de encontrar una matriz de menor dimensión que la original, que mejor resuma la información de los datos originales, en el sentido de los mínimos cuadrados.

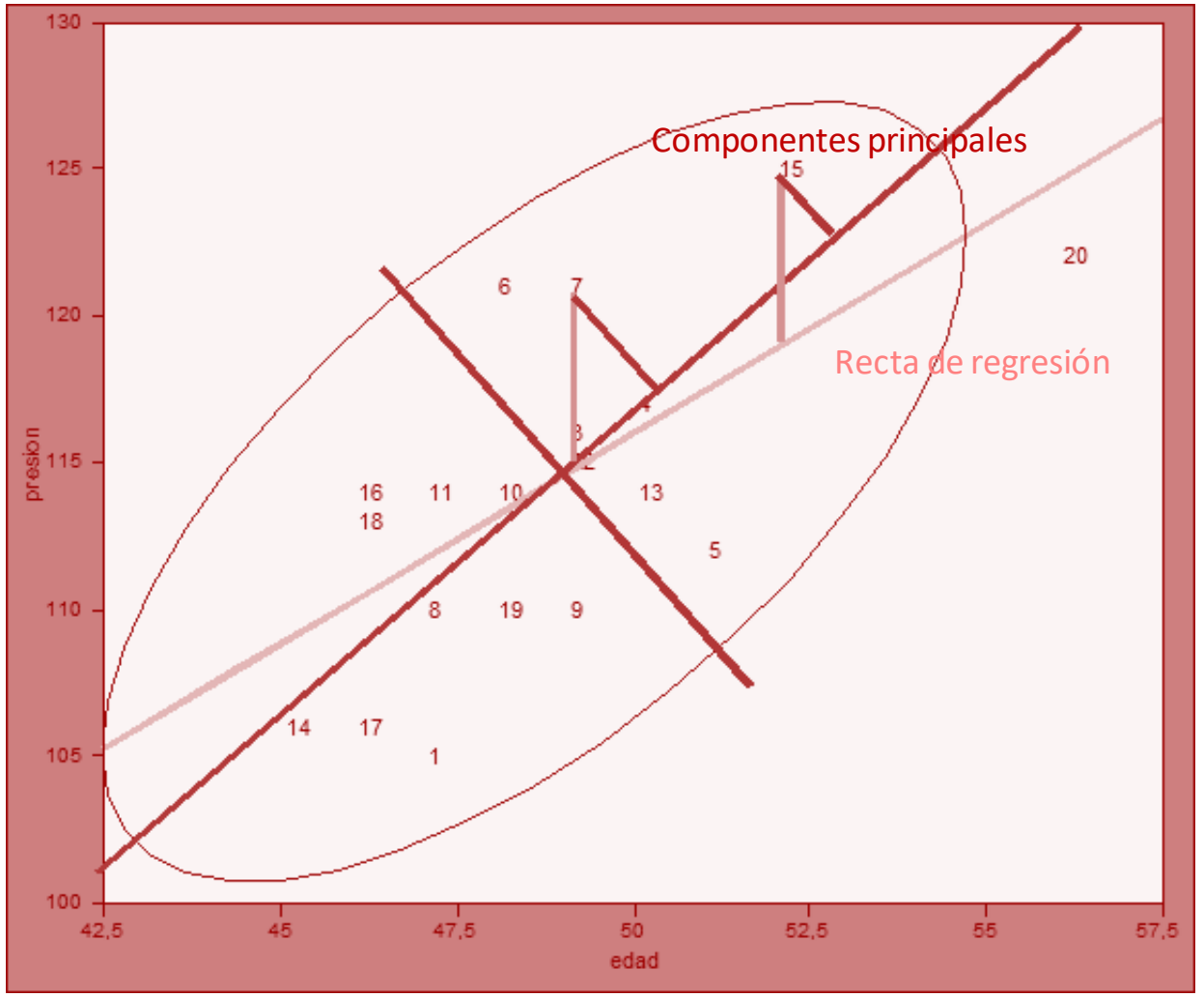


Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**:417-441,498-520.

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, **1**: 27-35

La aproximación de Hotelling obtiene sucesivamente combinaciones lineales de variables con varianza máxima.

Diferencia con regresión



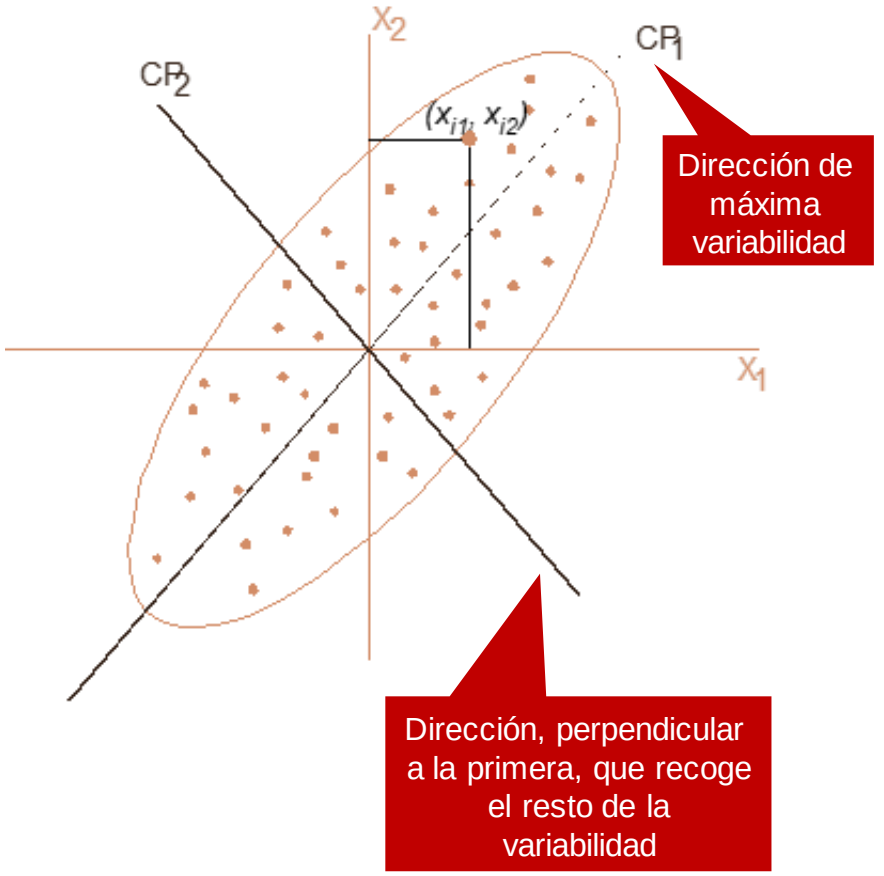
Análisis de Componentes Principales

- El Análisis de Componentes Principales es una técnica de reducción de la dimensión que describe la información de un conjunto de variables numéricas observadas mediante un conjunto de variables más pequeño (las componentes principales).

- Las nuevas variables (componentes principales) son incorreladas, y se obtienen en orden decreciente de importancia.

- El ACP puede entenderse como la búsqueda del subespacio de mejor ajuste a los datos y a diferencia de la regresión, en este método se espera que las variables estén correlacionadas entre sí

Principal Component Analysis (PCA)

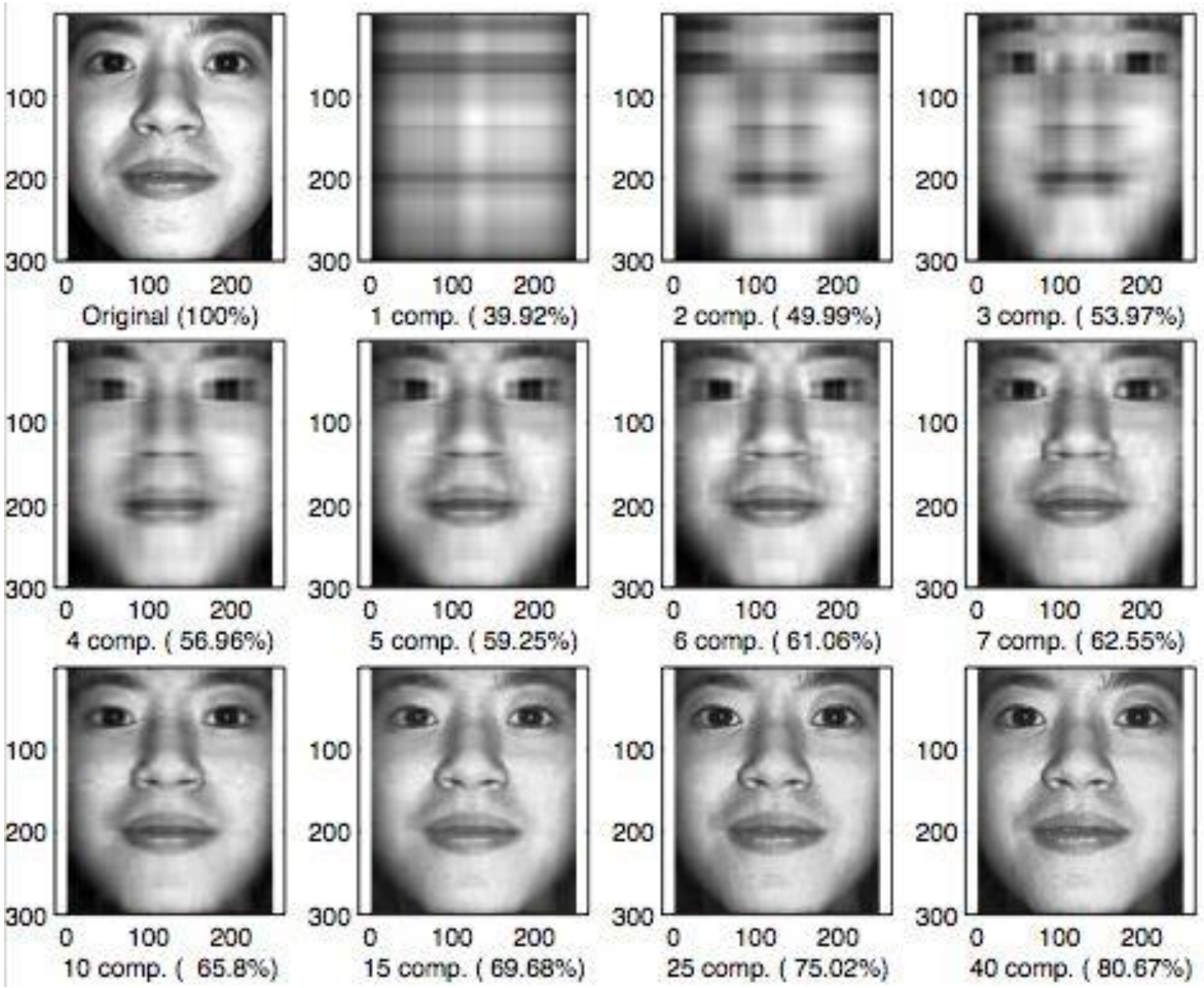


Ejemplo Imágenes (1)

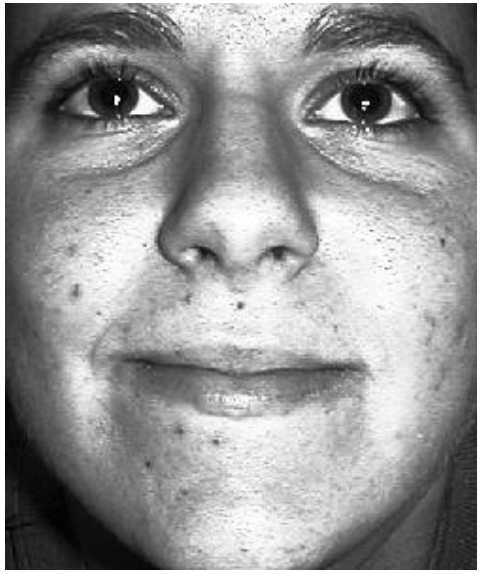
Datos: Imágenes de una cara (300x250 pixels) codificados en una escala de grises de 0 a 256.

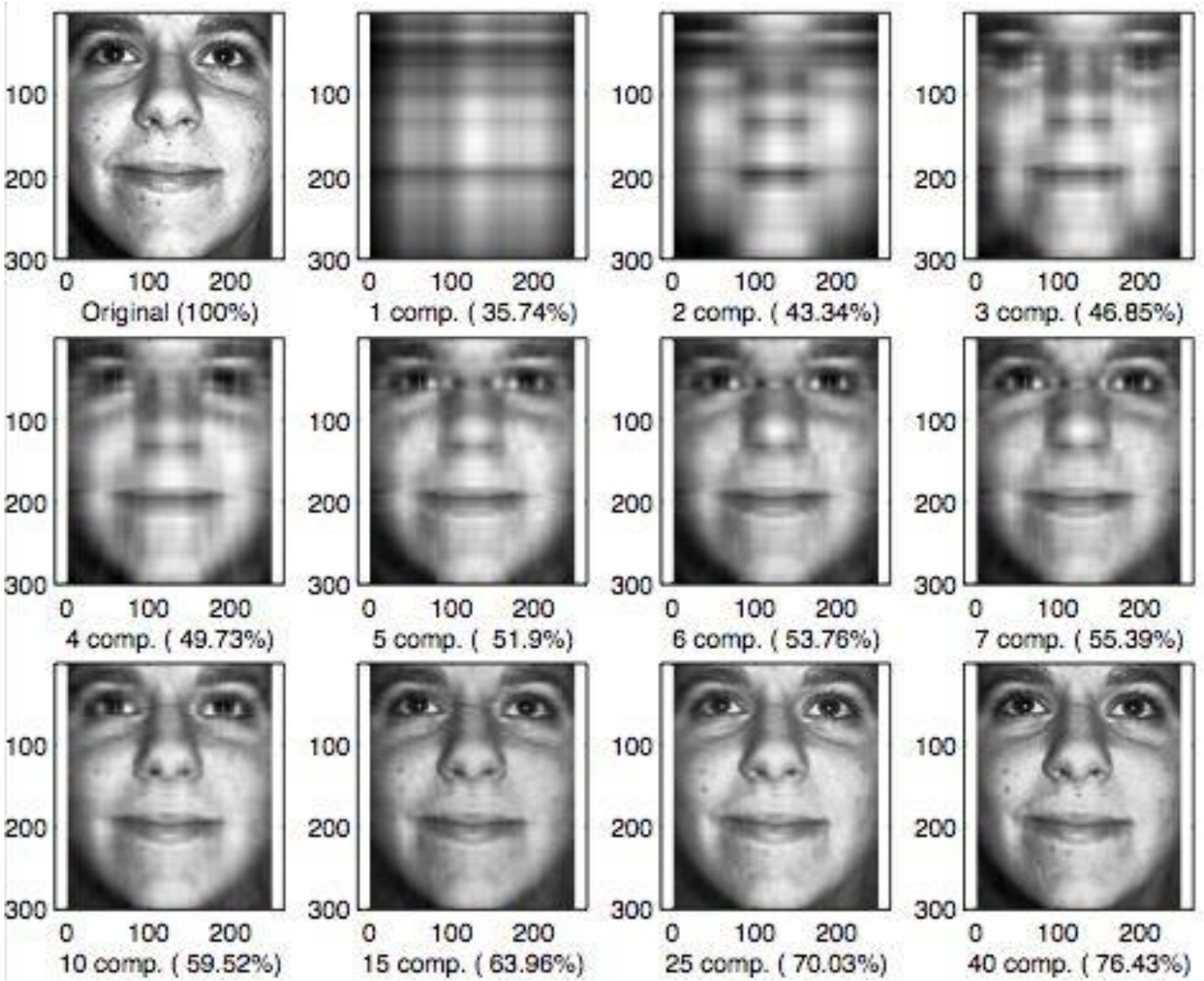
103	104	106	106	123	120	114	111	127	134	125	132	125	123	139	137	147	154	160
98	103	113	116	114	114	111	109	118	123	127	129	129	129	132	142	149	154	160
103	101	106	108	108	113	123	120	116	116	118	123	134	137	134	144	144	147	152
116	114	113	114	111	116	132	129	125	120	123	129	129	129	139	139	142	142	147
104	108	111	113	113	114	120	120	120	120	125	127	125	127	139	142	149	152	149
108	109	111	108	104	111	118	116	123	127	132	134	137	137	137	149	142	144	157
106	108	109	108	103	109	120	120	129	134	134	134	142	142	144	147	144	152	157
101	103	100	103	106	109	123	125	129	134	134	134	142	142	144	139	157	160	157
100	100	100	104	109	113	123	125	127	129	137	139	142	144	142	142	157	160	154
97	100	106	109	114	116	125	129	132	129	134	137	142	142	147	149	147	147	160
97	100	109	111	116	118	127	132	137	134	127	129	137	142	142	152	139	144	160
101	104	109	113	114	118	114	116	127	129	129	132	142	154	152	166	178	182	182
103	108	109	109	118	120	120	123	137	134	147	147	166	178	160	152	154	163	188
103	106	109	111	118	125	129	132	149	144	152	149	142	142	132	123	108	111	116
108	113	123	129	139	147	142	139	125	113	101	97	83	82	83	83	87	87	88
113	116	123	127	132	134	125	114	97	89	83	82	75	77	79	89	104	104	100
116	127	132	125	103	97	83	78	77	79	88	94	103	101	100	92	94	104	87
118	125	118	108	88	84	77	78	83	85	85	89	84	87	82	69	67	69	63
108	97	84	84	87	87	82	75	63	60	65	56	43	44	41	37	28	27	31
83	79	77	74	69	67	54	49	38	35	25	27	30	29	28	27	27	27	21
78	77	74	67	59	54	44	38	24	21	28	32	38	36	36	36	31	28	21

Ejemplo Imágenes



Ejemplo Imágenes (2)





Compresión

Tamaño original:
300*250 = 75000

10 componentes: 300*10+250*10=5500
Factor de compresión :13.64
Calidad muy baja

15 componentes: 300*15+250*15 = 8250
Factor de compresión : 9
Calidad Baja

25 componentes: 300*25+250*25 = 14050
Factor de compresión : 5.34
Calidad media

25 componentes: 300*40+250*40 = 22000
Factor de compresión : 3.41
Calidad alta

Datos

Disponemos de una matriz $\mathbf{X}_{nxp}$ que contiene las medidas de *p* variables cuantitativas tomadas sobre *n* individuos. Para simplificar el resto de la exposición supondremos, sin pérdida de generalidad, que las columnas de **X** tienen media cero, es decir que se le ha restado la media de cada columna de forma que el origen se sitúa en el centro de gravedad de la nube de puntos.

Todas las variables tienen el mismo papel, es decir, el conjunto no se divide en variables dependientes e independientes.

$$X = (X_{ij}) = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Métodos de obtención

1.- Buscando aquella combinación lineal de las variables que maximiza la variabilidad. (Hottelling).

2.- Buscando el subespacio de mejor ajuste por el método de los mínimos cuadrados. (Minimizando la suma de cuadrados de las distancias de cada punto al subespacio). (Pearson).

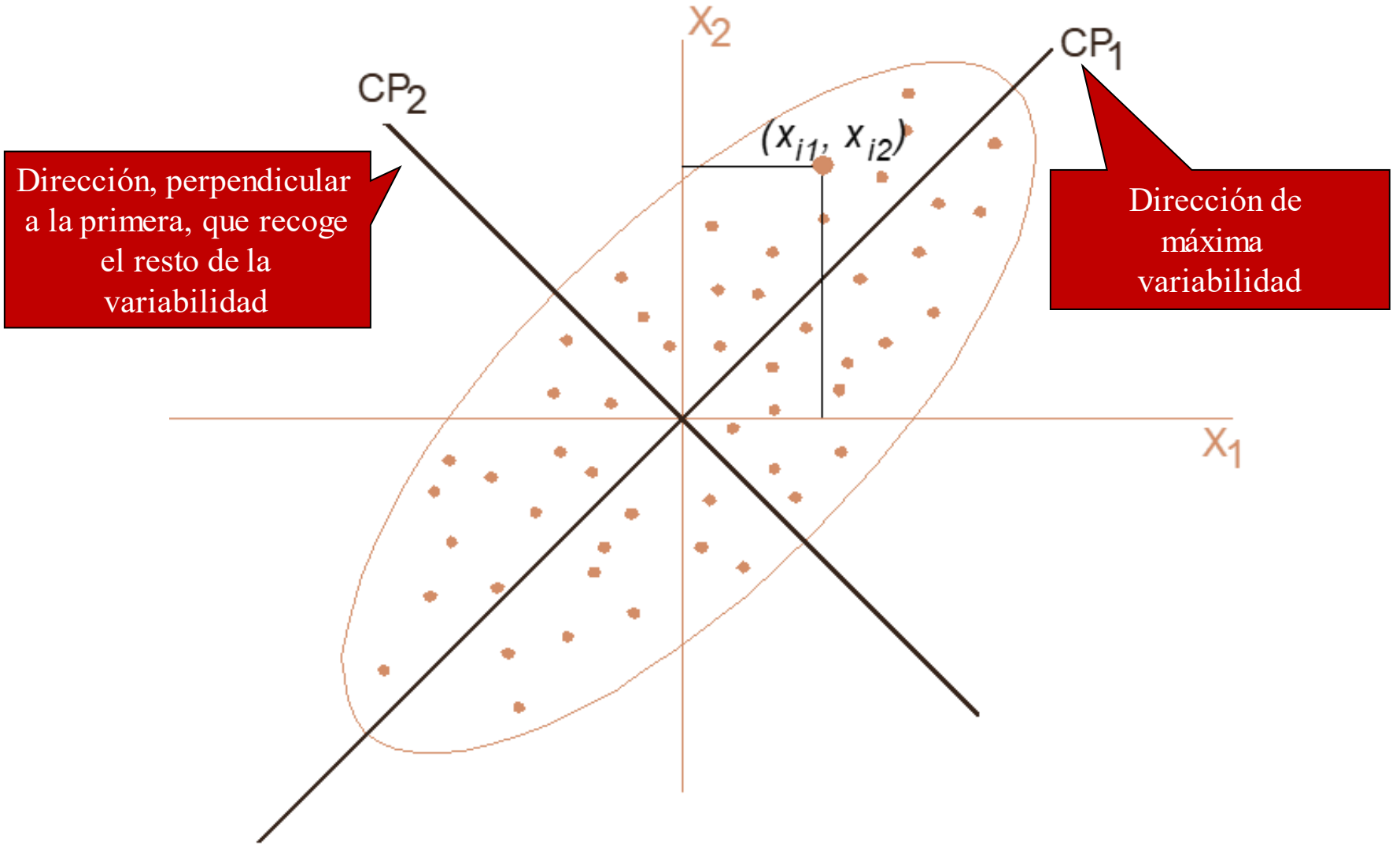
3.- Minimizando la diferencia entre las distancias euclídeas entre los puntos calculadas en el espacio original y en el subespacio de baja dimensión. (Coordenadas principales, Gower).

4.- Regresiones alternadas (Métodos Biplot).

Pearson, K. (1901). On lines and planes of closets fit to systems of points in the space. *Philosophical Magazine*, **2**: 559-572.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24:417-441,498-520.
.

ACP Bidimensional (Fig)



Escalas de medida

Si las escalas de medida de las variables son muy diferentes, la variabilidad estaría dominada por las variables con magnitudes mayores de forma que las primeras componentes pueden mostrar simplemente las diferencias en la escala. En este caso conviene tomar **la matriz X estandarizada** por columnas y centrando y dividiendo por la desviación típica. En este caso las componentes estarían colocadas sobre la matriz de correlaciones.

Interpretación de resultados

- Diagramas de dispersión que representan los valores de los individuos en las primeras componentes principales.
- Interpretación de distancias en términos de similitud.
- Búsqueda de clusters (grupos) y patrones.
- Interpretación de las componentes utilizando las correlaciones con las variables originales. Las posiciones de los individuos se interpretan después en relación a la interpretación dada a las componentes.

Ejemplo de Resultados de ACP

n=20 pacientes
p=7 variables

X_1=Presión arterial media (mmHg)
X_2=Edad (años)
X_3=Peso (kg.)
X_4=Superficie corporal (m^2)
X_5=Duración de la hipertensión (años)
X_6=Pulso (pulsaciones/minuto)
X_7=Medida del estres (0-100)

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	TOTAL
λ_j	3,908	1,470	0,708	0,521	0,308	0,080	0,002	7
% Var.	55,832	21,003	10,125	7,452	4,399	1,154	0,032	100%
% Var. acum.	55,382	76,835	86,961	94,414	98,813	99,968	100	

Ejemplo de Resultados de ACP

X_i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
PRESION	0,48814	-0,18969	-0,00547	-0,06758	0,04693	-0,37659	-0,75969
EDAD	0,36568	0,25049	-0,15331	-0,82970	0,17236	0,02013	0,24800
PESO	0,44713	-0,33244	0,03614	0,22336	-0,15349	-0,51151	0,59427
SUPERFICIE	0,40671	-0,38985	0,00711	0,19929	0,50399	0,62021	0,06458
DURACION	0,21965	0,43261	0,86381	0,09397	0,09557	0,00801	0,02047
PULSO	0,42683	0,23457	-0,16222	0,13200	-0,73112	0,42646	-0,05144
STRESS	0,17952	0,62976	-0,45015	0,43723	0,38322	-0,17186	0,03132

$$Y_1 = 0{,}48.\text{PRESIÓN} + 0{,}36.\text{EDAD} + \ldots + 0{,}17.\text{STRESS}$$

Representaciones Biplot

Variables

Individuos

Mientras más largos los vectores, más importantes las variables en el plano representado

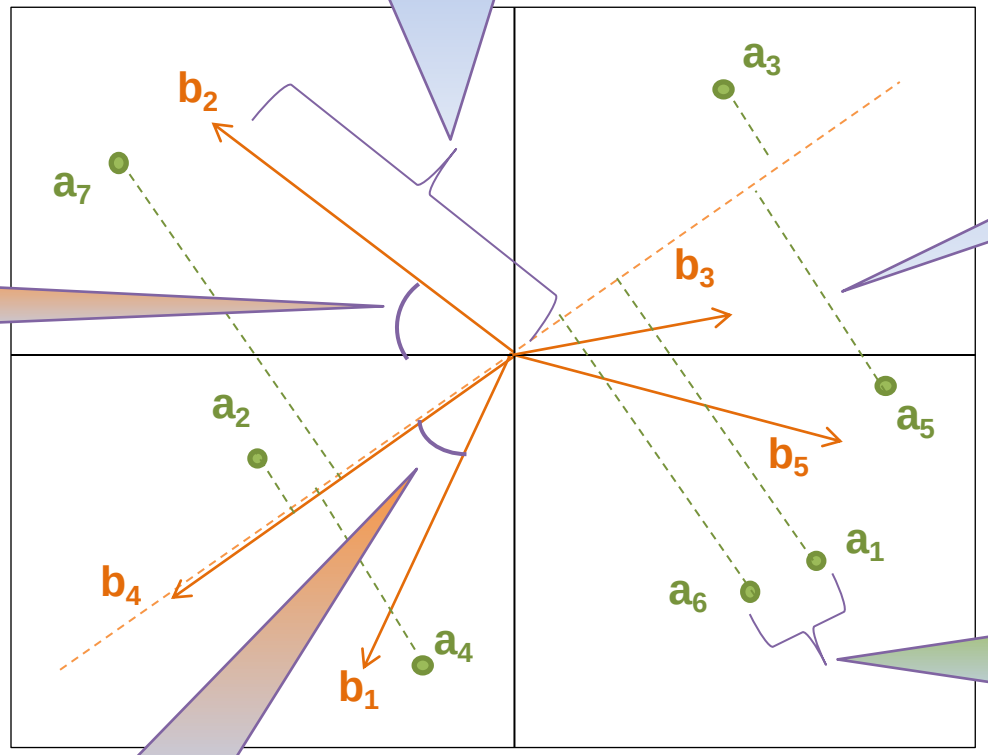
Variabilidad de la variable b_2

Orden de los individuos en relación a la variable b_4

Los individuos que están en la dirección del vector, tienen mayores valores frente a aquellos que se ubican en dirección opuesta

Contribución entre variables y factores

Las variables con ángulos más pequeños en relación a los ejes son las que más contribuyen a explicarlos



Similaridad entre individuos

Las cercanías entre individuos pueden analizarse para generar Clusters

Covariación entre variables

Dos variables con ángulo muy pequeño entre sí, están muy correlacionadas


GRACIAS
BIBLIOTECA
13