Reducción de la dimensión

Existen varias razones por las cuales nos puede interesar reducir la dimensionalidad del problema:

- En muchos algoritmos de aprendizaje, la complejidad depende tanto de las dimensiones del input como del tamaño de la muestra de datos, y podría interesarnos reducir la memoria y el cálculo requeridos

- Cuando se decide que un input no es necesario, se ahorra el costo de extraerlo

- Los modelos más simples son más robustos en conjuntos de datos pequeños

- Cuando los datos pueden ser explicados con pocas características, se puede tener una mejor idea del proceso subyacente y esto permite una extracción de conocimientos

- Cuando los datos pueden ser representados en pocas dimensiones sin pérdida de información, pueden graficarse y analizarse visualmente su estructura

Principio de parsimonia (o *navaja de Ockham*):

De un conjunto de variables explicativas que forman parte del modelo a estudiar, debe seleccionarse la combinación más reducida y simple posible

Hay dos métodos principales para reducir la dimensionalidad:

- **selección de características**
  - Estamos interesados en encontrar *k* de las *p* dimensiones que nos den de la mayor cantidad de información y en descartar las otras *p-k* dimensiones
  - Veremos el método de *selección de conjuntos*

- **extracción de características**
  - Estamos interesados en encontrar un conjunto de dimensión *k* que sea combinación de las *p* dimensiones originales.

- Los métodos más conocidos y usados son *análisis de componentes principales* (PCA) y *análisis discriminante lineal* (LDA), los cuales se basan en métodos de proyección lineal.

Selección de conjuntos

- En este método estamos interesados en encontrar el “mejor” subconjunto en un conjunto de características.

- El mejor subconjunto contiene el mínimo número de características que más contribuyen a la precisión.

- Hay 2^p subconjunto posibles de p variables, pero no se pueden testear todos salvo que *p* sea pequeño.

Hay dos métodos:

- **Selección hacia adelante** (forward):
  se comienza sin variables y se van agregando una a una, en cada paso agregando la variable que produzca el mayor decrecimiento del error, hasta que cualquier agregado no produzca decrecimiento del error (o decrezca muy poco)

- **Selección hacia atrás** (backward):
  se comienza con todas las variables y se van removiendo una a una, en cada paso removiendo la variable que produzca el mayor decrecimiento del error (o lo incremente sólo muy poco), hasta que cualquier remoción adicional incremente el error significativamente.

- El control del error puede hacerse con conjunto de validación distinto del conjunto de entrenamiento porque queremos testear la generalización de la precisión.

- Denotemos con F al conjunto de características de dimensiones de input, x_i, $i = 1, \cdots, p$.

- $E(F)$ denota al error que se comete una muestra de validación cuando sólo se usan los inputs en F.
  Dependiendo de la aplicación, el error puede ser el error cuadrático medio o el error de mala clasificación.

Algoritmo selección hacia adelante

1. Comenzamos sin características: $F = \emptyset$.

2. En cada paso, para toda posible x_i, se entrena a nuestro modelo y se calcula $E(F \cup x_i)$ sobre el conjunto de validación.

1. Se elige el input x_j que minimiza dicho error:

$$j = \arg \min_i E(F \cup x_i)$$

2. Agregamos x_j a F si $E\left(F \cup x_j\right) < E(F)$

3. Nos detenemos si al remover alguna característica no decrece el error.

- Un algoritmo similar se puede aplicar para la búsqueda hacia atrás.

- La complejidad de la búsqueda hacia atrás es la misma que la de la búsqueda hacia adelante, excepto que entrenar a un sistema con más características es más costoso que entrenar a uno con pocas, y por lo tanto es preferible la búsqueda hacia adelante si se espera que haya muchas características que no son útiles.

Aplicación a la regresión lineal

- La *selección hacia adelante* se comienza con un modelo con la única variable independiente que tenga la mayor correlación lineal con la variable dependiente *Y*.

- Luego se agrega al modelo aquella variable independiente que cumpla alguno de estos criterios equivalentes:

  1. Tenga la correlación parcial muestral más alta con la variable dependiente, ajustando con las variables independientes que ya están en la ecuación

2. Al agregar esa variable se obtiene el mayor incremento posible de R^2 que con cualquier otra variable

3. La variable agregada producirá el mayor valor del estadístico *F* que con cualquier otra variable que no esté aún en el modelo

- Así, se comienza con un conjunto de variables de tamaño 1 y en cada paso se añade otra variable al modelo de acuerdo con el criterio anterior, hasta que se cumple alguno de los siguientes criterios de parada:

1. Se obtiene un conjunto de tamaño *p* predefijado

2. Los test *F* de cada una de las variables que aún no están en el modelo tienen todos un valor menor que un valor prefijado *F-in*.

3. Cuando al agregar una nueva variable independiente, esta junto con las que ya están en el modelo son muy cercanas a ser colineales.

El test F

- Consideremos 2 modelos lineales: uno con k parámetros y otro con p parámetros, con $k < p$.

- El test F puede usarse para comparar estos modelos viendo si el modelo de k parámetros es un submodelo del otro.

- Sea

$$F_k = \frac{\left(RSS_k - RSS_p\right)/(p-k)}{RSS_p/(n-p)}$$

donde RSS_j es la suma de los cuadrados de los residuos para el modelo con j parámetros, es decir:

$$RSS_j = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{j-1} x_{i,j-1}\right)^2$$

- Entonces diremos que preferimos el modelo de k parámetros si

$$F_k > \mathcal{F}_{p-k,n-p,\alpha}$$

Ejemplo

- Consideremos el siguiente conjunto de datos generados artificialmente:

	X_1	X_2	X_3	X_4	Y
1	0.76	0.05	0.87	0.87	0.14
2	0.47	0.40	1.16	1.20	1.06
3	0.46	0.45	1.03	1.05	1.14
4	0.55	0.26	0.89	0.90	0.50
5	0.55	0.86	1.69	1.72	1.45
6	0.38	0.52	1.19	1.23	1.02
7	0.39	0.31	0.86	0.88	0.78
8	0.46	0.14	0.76	0.78	0.53
9	0.10	0.41	0.52	0.53	0.54
10	0.95	0.26	1.23	1.23	0.69
11	0.17	0.97	1.39	1.42	1.49
12	0.98	0.82	1.96	1.98	1.92
13	0.23	0.22	0.71	0.74	0.13
14	0.74	0.33	1.09	1.09	1.04
15	0.61	0.73	1.63	1.67	1.09
16	0.62	0.85	1.56	1.57	1.35
17	0.51	0.97	1.56	1.57	1.60
18	0.81	0.16	1.18	1.20	0.47
19	0.03	0.76	1.08	1.12	0.79
20	0.77	0.38	1.22	1.23	1.14

- Ajustemos todos los modelos con todos los subconjuntos posibles de variables:

  - Con una variable

variables	R^2	F	β_0	β_1	β_2	β_3	β_4
X_1	0,0323	0,6013	0,7715	0,3265			
X_2	0,7053	43,0783	0,2689		1,3698		
X_3	0,7072	43,4700	-0,3565			1,1026	
X_4	0,7070	43,4357	-0,3684				1,0942

Los valores de R^2 y de F son similares para los modelos con las variables X_2, X_3 y X_4, pero como el mayor valor se obtiene para el ajuste con X_3, consideramos el modelo:

$$Y = -0.3565 + 1.1026\, X_3$$

o Con dos variables:

variables	R^2	F	β_0	β_1	β_2	β_3	β_4
X_1 X_2	0,8304	41,6125	-0,1340	0,6552	1,4867		
X_1 X_3	0,7589	26,7496	-0,2916	-0,4635		1,2548	
X_1 X_4	0,7485	25,2993	-0,3046	-0,4102			1,2213
X_2 X_3	0,8083	35,8333	-0,1879		0,7807	0,6334	
X_2 X_4	0,8047	35.0335	-0,1898		0,7776		0,6258
X_3 X_4	0,7073	20,5435	-0,3618			0,6469	0,4526

El mejor ajuste se obtendría con:

$$Y = -0.1340 + 0.6552\, X_1 + 1.4867\ \ X_2$$

o Con tres variables:

variables	R^2	F	β_0	β_1	β_2	β_3	β_4
X_1 X_2 X_3	0,8307	26,1697	-0,1164	0,7371	1,5888	-0,0941	
X_1 X_2 X_4	0,8305	26,1364	-0,1220	0.7027	1,5483		-0,0562
X_2 X_3 X_4	0,8176	23,9096	-0.1270		0,8630	42,5350	-3,6442

El mejor ajuste se obtendría con:

$$Y = -0.1164 + 0.7371\,X_1 + 1.5888\,X_2 - 0.0941\,X_3$$

o Y se utilizan todas las variables, el ajuste sería:

variables	R^2	F	β_0	β_1	β_2	β_3	β_4
X_1 X_2 X_3 X_4	0.8645	23.9185	-0.2354	4.0855	4.6604	-29.7863	26.5627

$$Y = -0.2354 + 4.0855\,X_1 + 4.6604\,X_2 - 29.7863X_3 + 26.5627\;X_4$$

- Si se hiciera un **procedimiento de selección hacia adelante** para ajustar con a lo sumo dos variables se obtendría:

Paso 1:

variables	R^2	F
X_1	0,0323	0,6013
X_2	0,7053	43,0783
X_3	0,7072	43,4700
X_4	0,7070	43,4357

elegimos X_3 para incorporar al modelo

Paso 2:

variables	R^2	F
X_1 X_3	0,7589	26,7496
X_2 X_3	0,8083	35,8333
X_3 X_4	0,7073	20,5435

elegimos X_2 para incorporar al modelo

Autovalores y autovalores

Sea $A \in K^{n\times n}$ donde $K = \mathbb{R}$ o $K = \mathbb{C}$.
Se dice que $\lambda \in K$ es un **autovalor** de A si existe $\boldsymbol{v} \in K^n$, $\boldsymbol{x} \neq \mathbf{0}$ tal que $A\boldsymbol{v} = \lambda\boldsymbol{v}$
El vector $\boldsymbol{v}$ se llama **autovector** asociado a λ.

Teorema: (método para calcular autovalores y autovectores)

1. λ es un autovalor de A sí y sólo si $\det(A - \lambda I) = 0$

2. $\boldsymbol{v}$ es el autovector asociado a λ si $(A - \lambda I)\boldsymbol{v} = \mathbf{0}$

Algunas propiedades:

- $P(\lambda) = \det(A - \lambda I)$ es un polinomio de grado n que se denomina polinomio característico.

- Si $A \in \mathbb{R}^{n \times n}$ es una matriz simétrica entonces:

  a. Todos sus autovalores son reales.

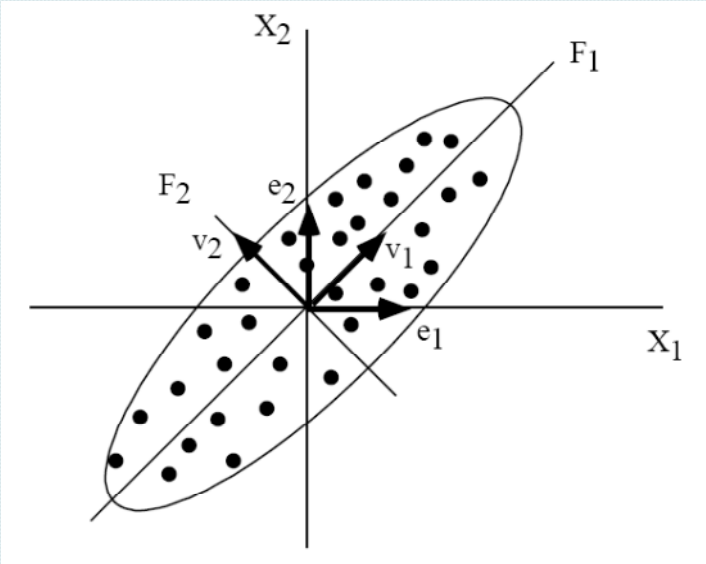
  b. Existe una matriz ortogonal $V \in \mathbb{R}^{n \times n}$ (i.e. $VV^t = I$) tal que $V^t A V = \Lambda$ donde

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \qquad V = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$$

  siendo $\lambda_1 \geq \cdots \geq \lambda_n$ los autovalores de A con autovectores asociados respectivos $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$.

  c. $\det(A) = \prod_{i=1}^{n} \lambda_i$

Análisis de componentes principales



- El Análisis de Componentes Principales (PCA) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables).

- Es un método de proyección en el que estamos interesados en encontrar una función que transforme el espacio original de dimensión p en uno de menor dimensión k de modo que la pérdida de información sea mínima.

- La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente.

- Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente.
  A éstos se les denominará **componentes principales**.

- Las componentes principales serán una combinación lineal de las variables originales, y además serán independientes entre sí.

- La proyección de $\boldsymbol{x}$ en la dirección de $\boldsymbol{w}$ es

$$z = \boldsymbol{w}^t \boldsymbol{x}$$

- La componente principal es $\boldsymbol{w}_1$, siendo $\boldsymbol{w}_1$ la dirección tal que al proyectar la muestra en $\boldsymbol{w}_1$ se evidencia la mayor variabilidad entre los puntos de la misma

- Para obtener una única solución, pedimos que $\|\boldsymbol{w}_1\| = 1$.

- Entonces, si $z_1 = \boldsymbol{w}_1^t \boldsymbol{x}$, con $\text{cov}(\boldsymbol{x}) = \Sigma$, entonces:

$$\text{var}(z_1) = \boldsymbol{w}_1^t \Sigma \boldsymbol{w}_1$$

- Buscamos $\boldsymbol{w}_1$ tal que se maximice var(z_1) sujeto a la restricción $\|\boldsymbol{w}_1\| = 1$.

- Usando multiplicadores de Lagrange, tenemos que:

$$\max_{\boldsymbol{w}_1} \boldsymbol{w}_1^t \Sigma \boldsymbol{w}_1 - \alpha(\boldsymbol{w}_1^t \boldsymbol{w}_1 - 1)$$

- Derivando con respecto a $\boldsymbol{w}_1$ e igualando a cero, obtenemos:

$$2\Sigma \boldsymbol{w}_1 - 2\alpha \boldsymbol{w}_1 = 0 \quad \Longrightarrow \quad \Sigma \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1$$

- Es decir, $\boldsymbol{w}_1$ es autovector de Σ con autovalor α.

- Como queremos maximizar $\boldsymbol{w}_1^t \Sigma \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1^t \boldsymbol{w}_1$, elegimos el autovector que corresponda al mayor autovalor para maximizar la varianza.
  - Por lo tanto, la componente principal es el autovector de la matriz de covarianza muestral con el mayor autovalor $\lambda_1 = \alpha$.

- La segunda componente principal, $\boldsymbol{w}_2$ debería también maximizar la varianza, tener norma 1 y ser ortogonal a $\boldsymbol{w}_1$ (este último se pide para que la proyección $z_2 = \boldsymbol{w}_2^t \boldsymbol{x}$ no esté correlacionada con z_1).

- Para la segunda componente principal tenemos:

$$\max_{\boldsymbol{w}_2} \boldsymbol{w}_2^t \Sigma \boldsymbol{w}_2 - \alpha(\boldsymbol{w}_2^t \boldsymbol{w}_2 - 1) - \beta(\boldsymbol{w}_2^t \boldsymbol{w}_1 - 0)$$

- Derivando con respecto a $\boldsymbol{w}_2$ e igualando a cero, obtenemos:

$$2\Sigma \boldsymbol{w}_2 - 2\alpha \boldsymbol{w}_2 - \beta \boldsymbol{w}_1 = 0 \qquad (*)$$

- Premultiplicando por $\boldsymbol{w}_1^t$, resulta:

$$2\boldsymbol{w}_1^t \Sigma \boldsymbol{w}_2 - 2\alpha \boldsymbol{w}_1^t \boldsymbol{w}_2 - \beta \boldsymbol{w}_1^t \boldsymbol{w}_1 = 0$$

- Como $\boldsymbol{w}_1^t \boldsymbol{w}_2 = 0$, y además

$$\boldsymbol{w}_1^t \Sigma \boldsymbol{w}_2 = (\boldsymbol{w}_1^t \Sigma \boldsymbol{w}_2)^t = \boldsymbol{w}_2^t \Sigma \boldsymbol{w}_1 = \boldsymbol{w}_2^t \lambda_1 \boldsymbol{w}_1 = \lambda_1 \boldsymbol{w}_2^t \boldsymbol{w}_1 = 0$$

reemplazando se obtiene que $\beta = 0$ y la ecuación (*) se reduce a :

$$\Sigma \boldsymbol{w}_2 = \alpha \boldsymbol{w}_2$$

lo cual implica que $\boldsymbol{w}_2$ es un autovector de Σ con el segundo mayor autovalor $\lambda_2 = \alpha$.

- Análogamente, podemos mostrar que las otras componentes son los autovectores con los autovalores decrecientes.

- Como Σ es simétrica, para dos autovalores diferentes, los autovectores son ortogonales.

- Si Σ es definida positiva (i.e. $\boldsymbol{x}^t \Sigma \boldsymbol{x} > 0 \quad \forall \boldsymbol{x} \neq \boldsymbol{0}$), entonces sus autovalores son positivos.

- Si Σ es singular, entonces su rango es $k < p$, y $\lambda_i = 0$ para $i = k+1, \cdots, p$.

- El primer autovector, es decir el que corresponde al autovalor más grande, $\boldsymbol{w}_1$, se denomina componente principal y explica la mayor parte de la varianza; el segundo, corresponde al segundo autovalor más grande es la segunda componente principal, y así siguiendo.

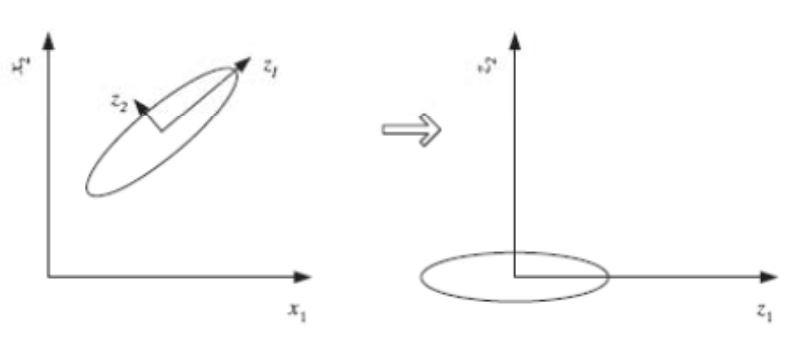
- Definimos:

$$\boldsymbol{z} = W^t(\boldsymbol{x} - \hat{\boldsymbol{\mu}})$$

  donde las k columnas de W corresponden a las k componentes principales de S, el estimador de Σ.

- Se resta la media muestral $\hat{\boldsymbol{\mu}}$ a $\boldsymbol{x}$ para centrar los datos en el origen antes de proyectar.

- Después de esta transformación lineal, obtenemos un espacio k-dimensional cuyas variables son los autovectores y las varianzas sobre estas nuevas variables son los autovalores.



- El análisis de componentes principales centra a la muestra y luego rota los ejes en la dirección de mayo varianza.

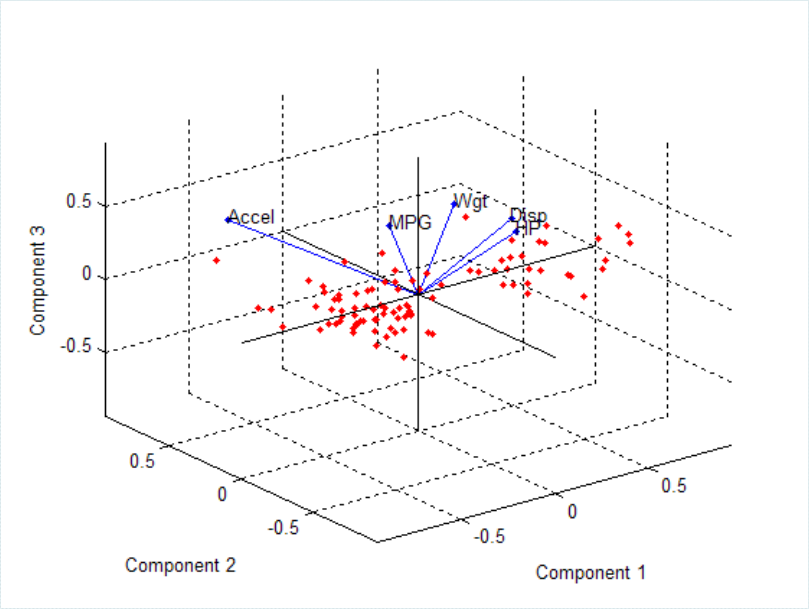
- Como $\det(S) = \prod_{i=1}^{p} \lambda_i$, se puede pensar que algunos autovalores tienen poca contribución a la varianza y podrían ser descartados.

  - Entonces, se podría considerar las k componentes que explicaran por ejemplo el 90% de la varianza.

- Así **la proporción de varianza** explicada por las k componentes centrales es:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

Biplot



- Un biplot permite representar en un mismo plot las observaciones y las variables de datos multivariados.

- Permite visualizar la magnitud y el signo de la contribución de las primeras dos (en 2D) o tres (en 3D) de las componentes principales, y cómo cada observación se representa en términos de estas componentes.

Ejemplo

- Supongamos que deseamos conocer cuáles son los factores relacionados con el riesgo de enfermedad coronaria.

- Del conocimiento previo sabemos que el riesgo está relacionado con la **presión arterial**, la **edad**, la **obesidad**, el tiempo que se ha sido **hipertenso**, el **pulso**, y el **stress**.

- Para la investigación seleccionamos al azar 20 pacientes hipertensos en los que medimos las siguientes variables:
  x_1: presión arterial media (mm Hg)
  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 stress

- Tratamos de estudiar la situación del grupo de pacientes en relación a los factores de riesgo y las posibles interrelaciones entre las distintas variables.

- Inicialmente queremos describir el conjunto de pacientes utilizando simultáneamente todas las variables

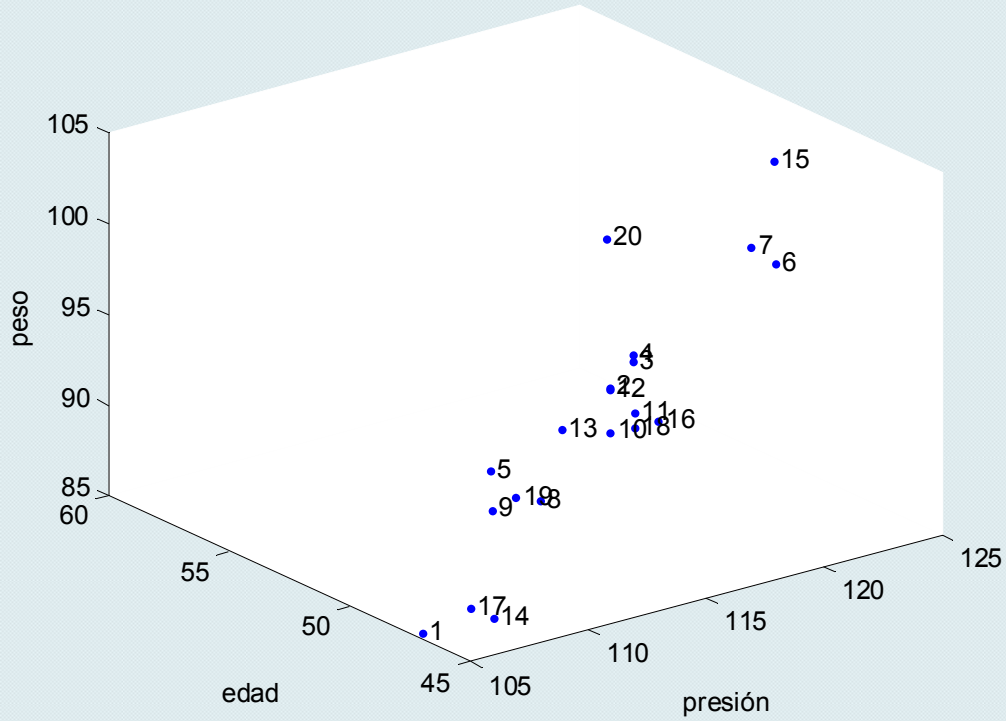
- Los datos obtenidos se muestran en la tabla siguiente:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	105	47	85,4	1,75	5,1	63	33
2	115	49	94,2	2,1	3,8	70	14
3	116	49	95,3	1,98	8,2	72	10
4	117	50	94,7	2,01	5,8	73	99
5	112	51	89,4	1,89	7	72	95
6	121	48	99,5	2,25	9,3	71	10
7	121	49	99,8	2,25	2,5	69	42
8	110	47	90,9	1,9	6,2	66	8
9	110	49	89,2	1,83	7,1	69	62
10	114	48	92,7	2,07	5,6	64	35
11	114	47	94,4	2,07	5,3	74	90
12	115	49	94,1	1,98	5,6	71	21
13	114	50	91,6	2,05	10,2	68	47
14	106	45	87,1	1,92	5,6	67	80
15	125	52	101,3	2,19	10	76	98
16	114	46	94,5	1,98	7,4	69	95
17	106	46	87	1,87	3,6	62	18
18	113	46	94,5	1,9	4,3	70	12
19	110	48	90,5	1,88	9	71	99
20	122	56	95,7	2,09	7	75	99

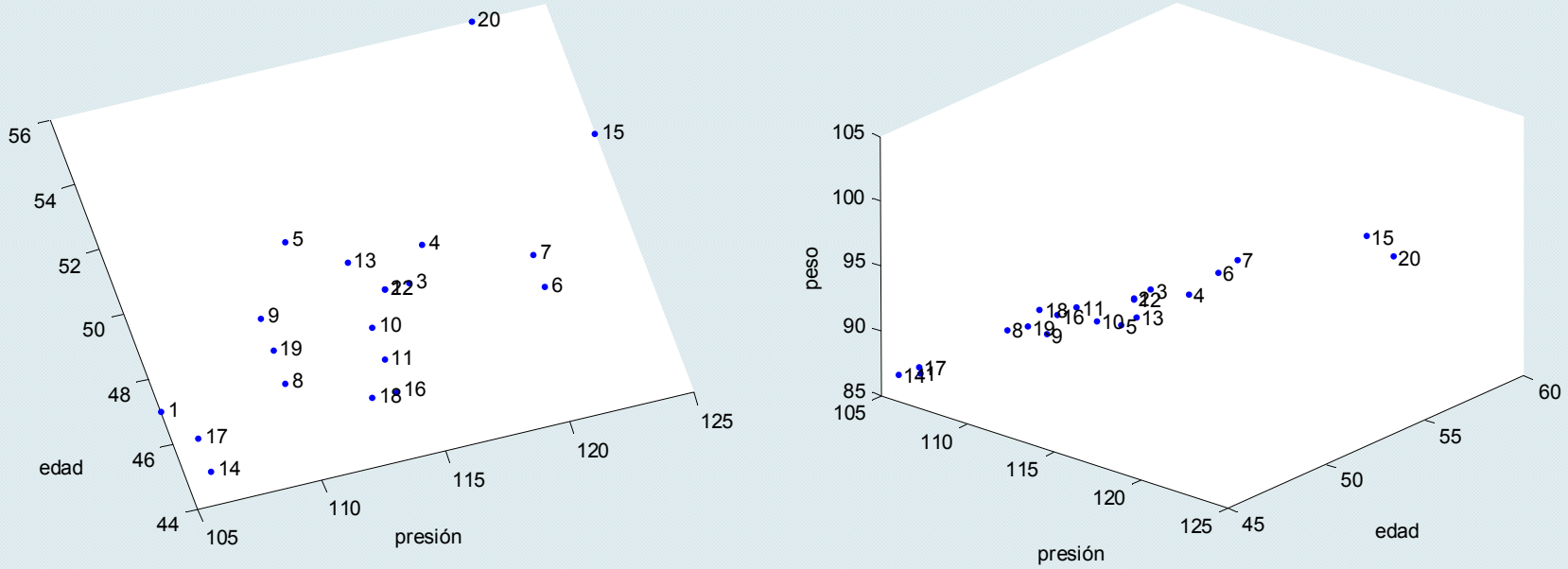
- La dimensión inicial es 7, pero ¿será posible describir el conjunto de datos utilizando un número menor de dimensiones, aprovechando las interrelaciones entre las variables?

- ¿Es posible definir un índice general que cuantifique la situación de riesgo?

- Consideremos las tres primeras variables: **presión arterial**, **edad** y **peso**

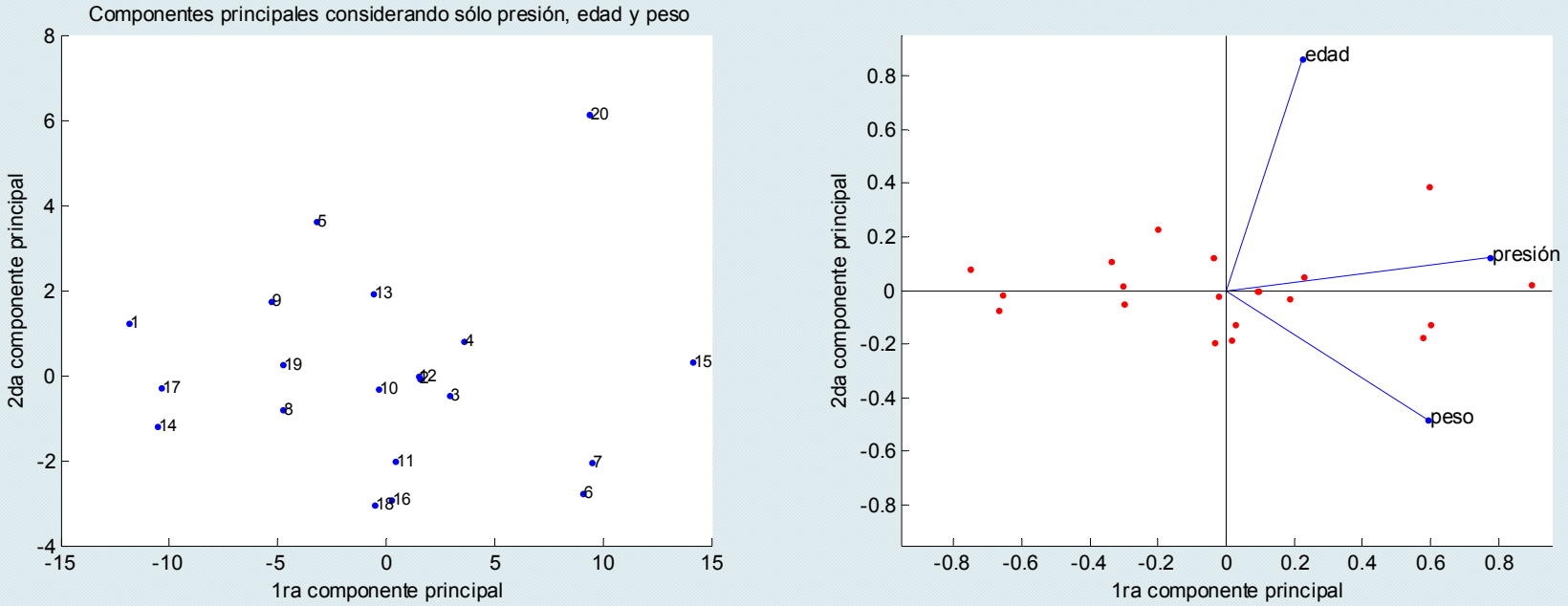


- Si movemos la figura resultante, observaremos que los puntos están casi sobre un plano.
  - Esto se pone de manifiesto en la figura siguiente en la que se ha conseguido un punto de vista desde el que los puntos parecen estar sobre una línea recta.



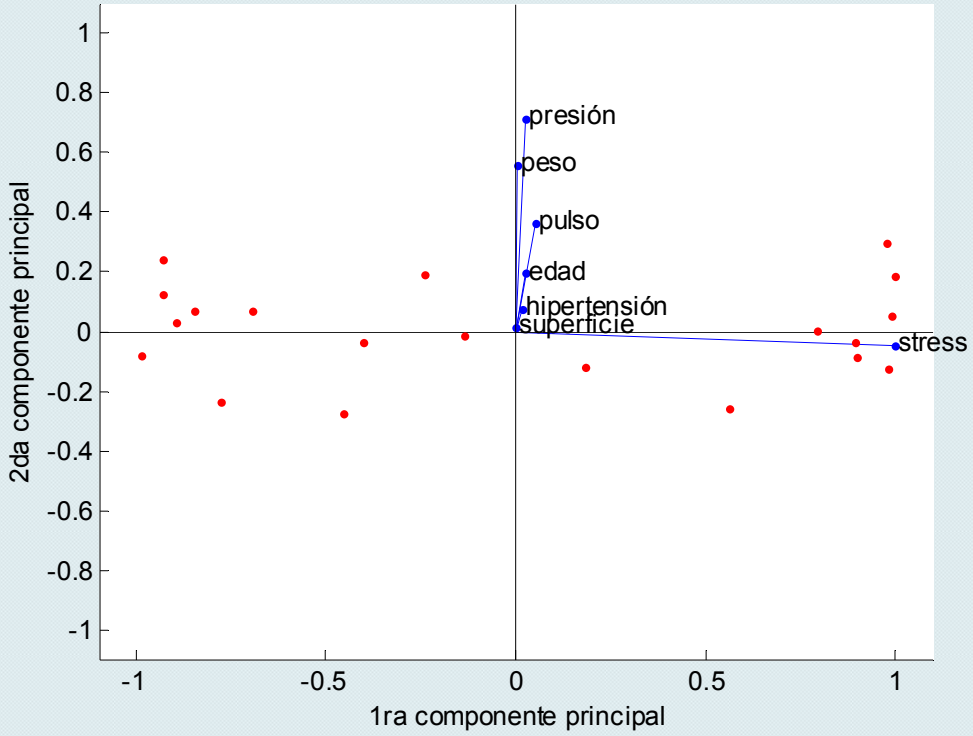
  - Este hecho pone de manifiesto que no son necesarias tres dimensiones para describir el conjunto de datos, sino solamente dos.

- La representación de las dos primeras componentes, para los datos anteriores y con sólo estas tres variables aparece en la figura siguiente.



Las dos primeras componentes absorben el 99% de la variabilidad de los datos.

- La figura siguiente muestra las dos primeras componentes principales para el conjunto de las 7 variables.



- Las dos primeras componentes absorben el 99% de la variabilidad de los datos.