Técnicas de reducción de la dimensión de la información: Componentes Principales.

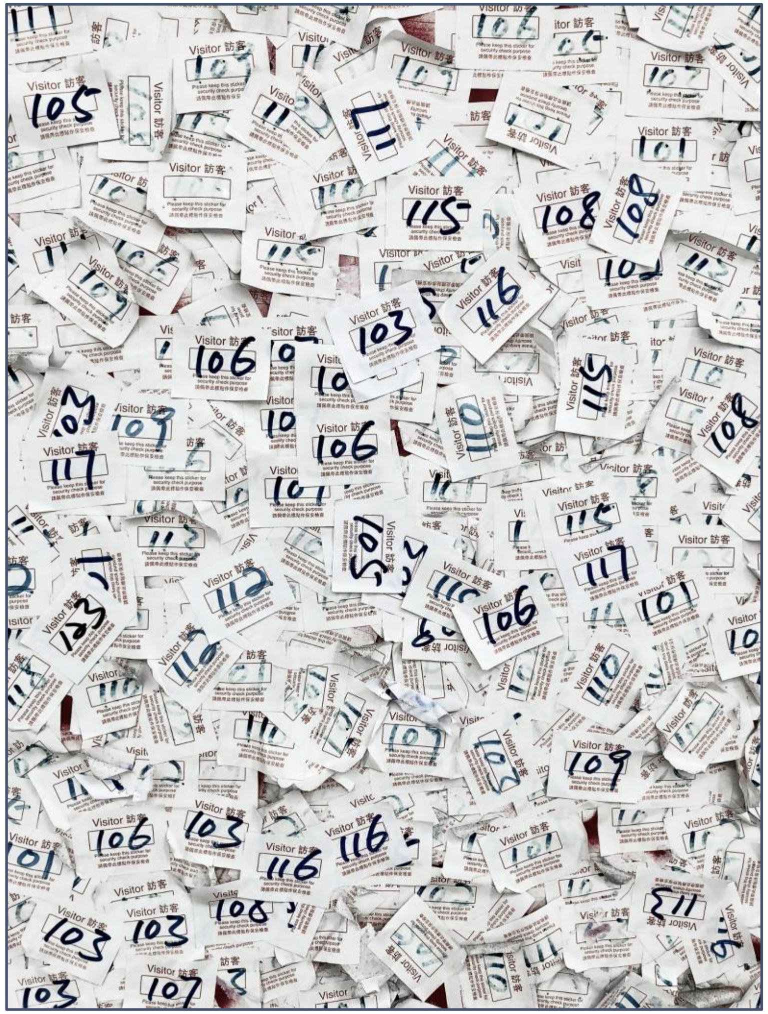
Miguel Ángel Tarancón Morán & Consolación Quintana Rojo

Área de Estadística Económica y Empresarial

Departamento de Economía Aplicada 1

Universidad de Castilla – La Mancha

Contenidos.



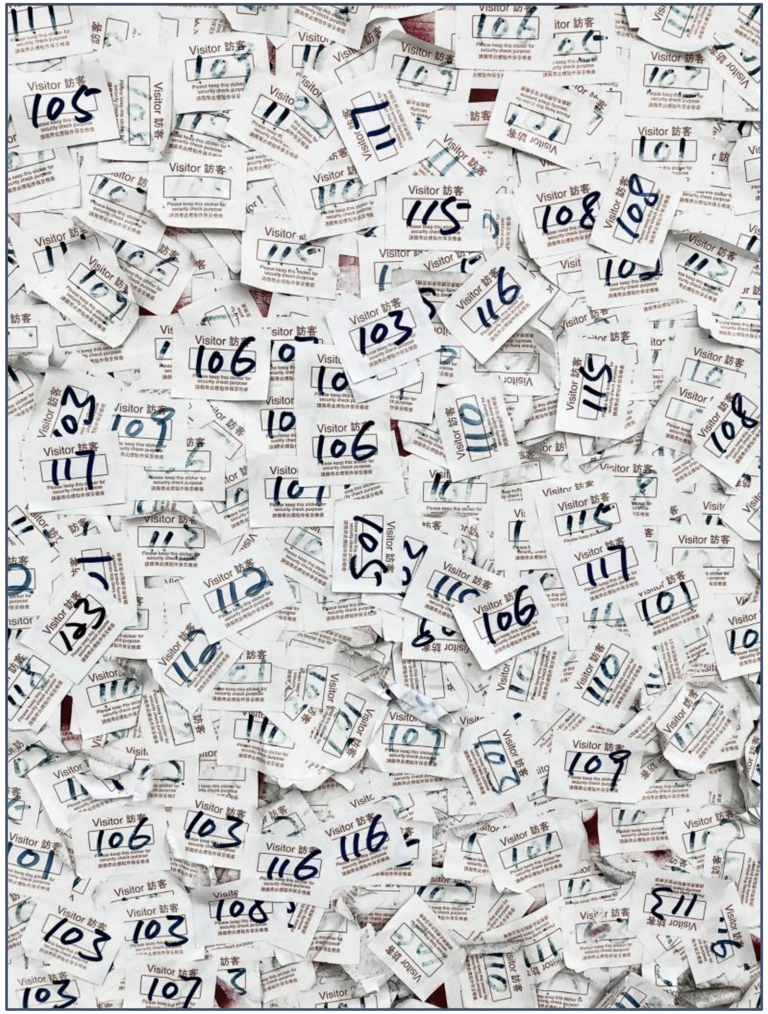
- Muchas veces, en la caracterización de los casos de una población a estudiar (por ejemplo, las empresas de un sector económico), observamos que hay muchas variables que caracterizan a tales casos.

- A veces, el contar con tantas variables hace difícil la caracterización de estos casos. Esto ocurre cuando hay **variables que aportan una información muy parecida** sobre estos agentes.

- Estas técnicas tratan de **reducir el número de variables** que caracterizan los agentes del entorno; **perdiendo la menor cantidad global de información** posible.

- Una de estas técnicas, utilizada habitualmente, es la técnica de **Componentes Principales** (CP).

Introducción.

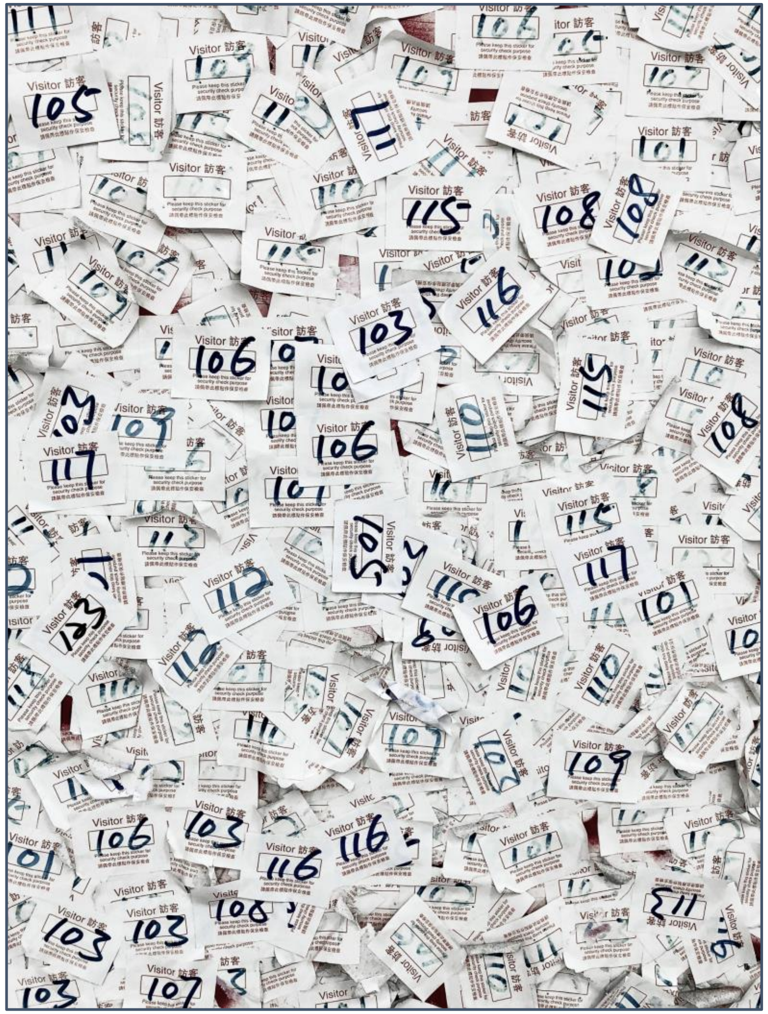


- El objetivo del análisis de **Componentes Principales (CP)** consiste en, a partir de una población caracterizada por **p** variables, reducir el número de variables a **h**, con h < p, de manera que las h nuevas variables, denominadas *componentes principales*:

  - Están **incorrelacionadas** entre sí.

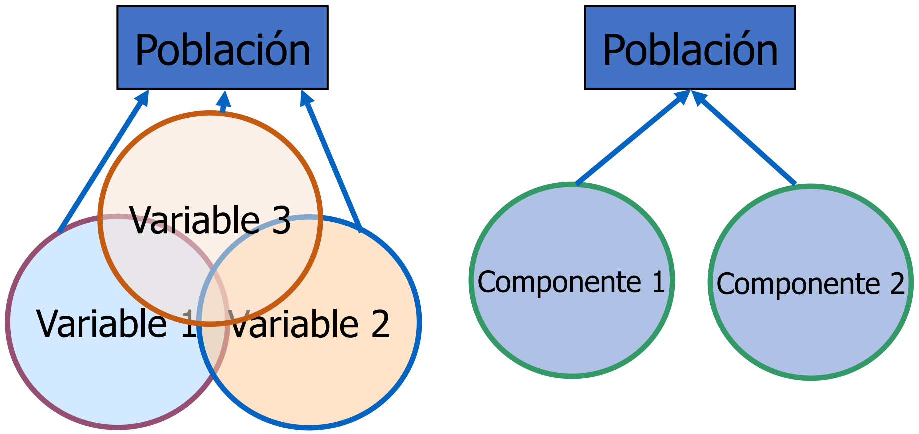
  - Son una **combinación lineal** de las variables originales.

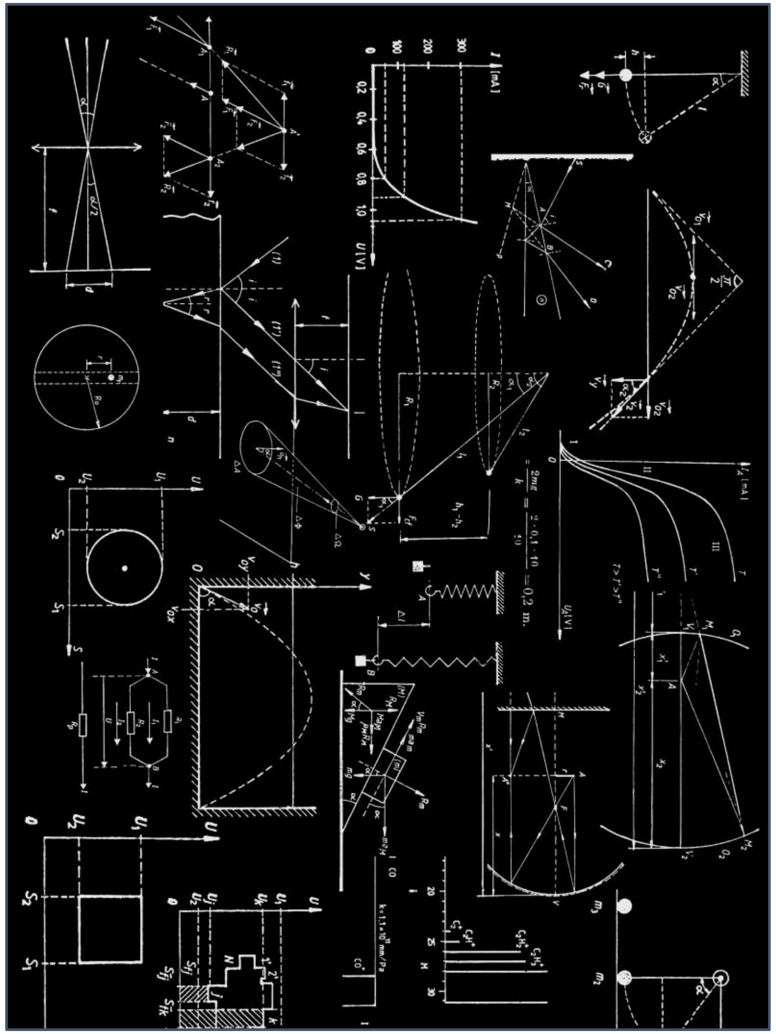
  - En conjunto, recogen la **mayor cantidad posible de información** (varianza) contenida en el conjunto de las variables originales sobre la población a caracterizar.



- Una premisa para que aplicar CP tenga sentido es que las **variables originales estén altamente correlacionadas**, es decir, que contengan información redundante; ya que en caso contrario no haría falta someter a los datos a esta técnica de reducción de la dimensión de la información.

- En definitiva, se trata de intentar eliminar la **información redundante** que tenemos acerca de una población.





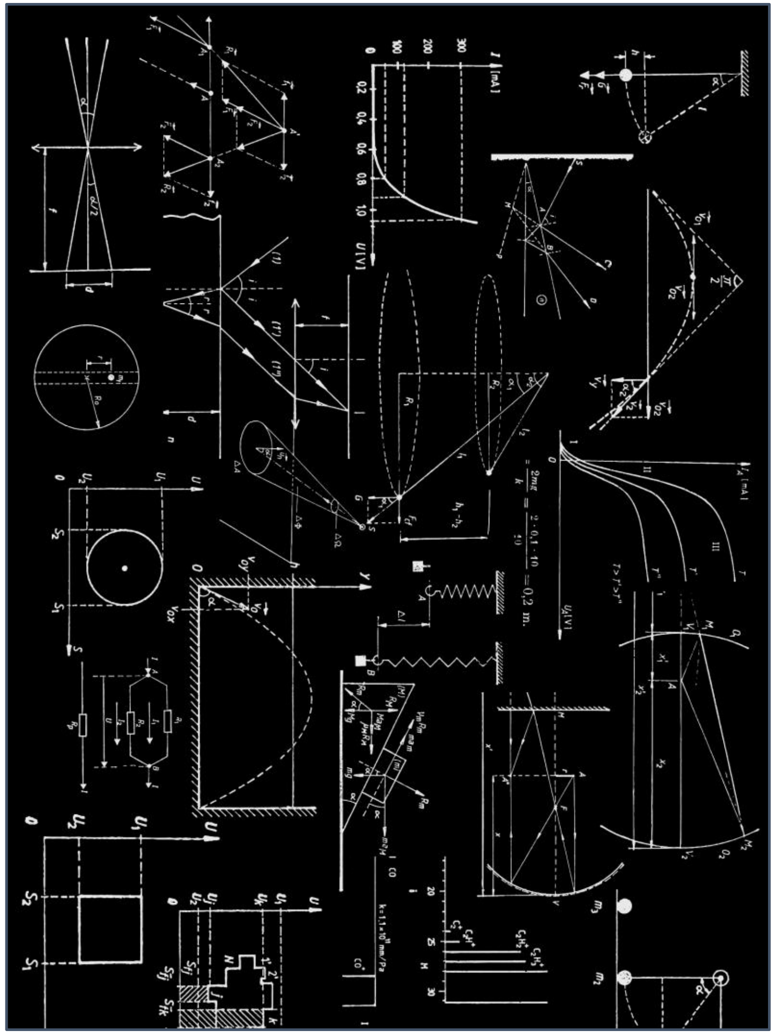
- **Matriz de datos:**

$$\begin{array}{cc} & \overbrace{\hphantom{x_{11}\quad x_{12}\quad \cdots \quad x_{1p}}}^{\text{p variables originales}} \\ & \hphantom{x_{11}\quad x_{12}\;} x_j \\ \text{n individuos} & \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \end{array}$$

- **Ejemplo:**

	variable 1	variable 2	variable 3
caso 1	23	13	6,5
caso 2	31	21	5
caso 3	24	14	4,5
caso 4	35	23	5
caso 5	52	41	7,5
caso 6	46	37	6
media	35,17	24,83	5,75
desv. típica	11,75	11,70	1,13

Obtención de componentes principales.



- Matriz de correlaciones:

	variable x1	variable x2	variable x3
variable x1	1,0000	0,9961	0,6065
variable x2	0,9961	1,0000	0,6092
variable x3	0,6065	0,6092	1,0000

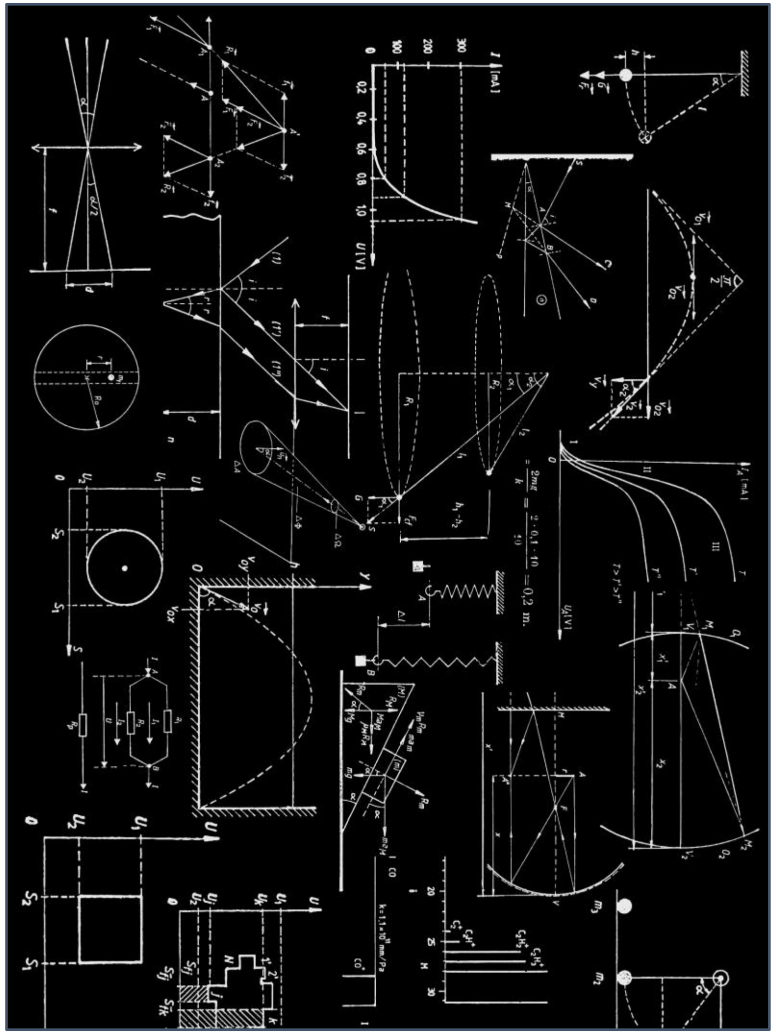
- Usualmente las variables originales han sido previamente **tipificadas** (media 0 y desviación típica 1):

Ejemplo: Esta es la matriz X con la que trabajamos

	variable 1	variable 2	variable 3
caso 1	23	13	6,5
caso 2	31	21	5
caso 3	24	14	4,5
caso 4	35	23	5
caso 5	52	41	7,5
caso 6	46	37	6
media	35,17	24,83	5,75
desv. típica	11,75	11,70	1,13

	variable 1	variable 2	variable 3
caso 1	-1,035	-1,011	0,664
caso 2	-0,354	-0,328	-0,664
caso 3	-0,950	-0,926	-1,107
caso 4	-0,014	-0,157	-0,664
caso 5	1,432	1,381	1,550
caso 6	0,922	1,040	0,221
media	0,00	0,00	0,00
desv. típica	1,00	1,00	1,00

Obtención de componentes principales.



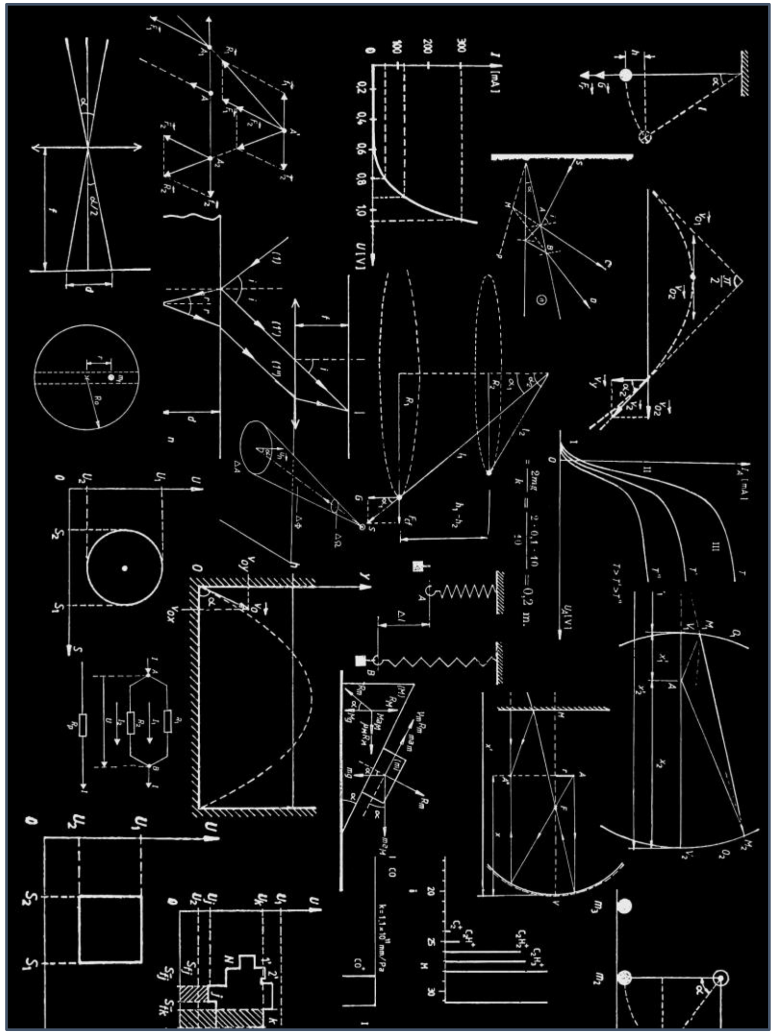
- Las variables originales contienen **información** sobre el comportamiento de los casos de nuestra población o muestra, comportamiento que los caracteriza y distingue.

- Esta información se concreta estadísticamente en las **varianzas** de las diferentes variables. La suma de varianzas (**comunalidad**) nos da una medida de la información que el conjunto de variables poseen sobre los casos.

- Las *componentes* son **combinaciones lineales** de las variables originales. **Hay tantas “componentes” como variables originales.**

- Las componentes **recogen la misma información que las variables originales** (misma suma de varianzas, *comunalidad*); pero de un modo diferente, ya que las componentes están **incorrelacionadas**: no comparten información “duplicada”.

Obtención de componentes principales.



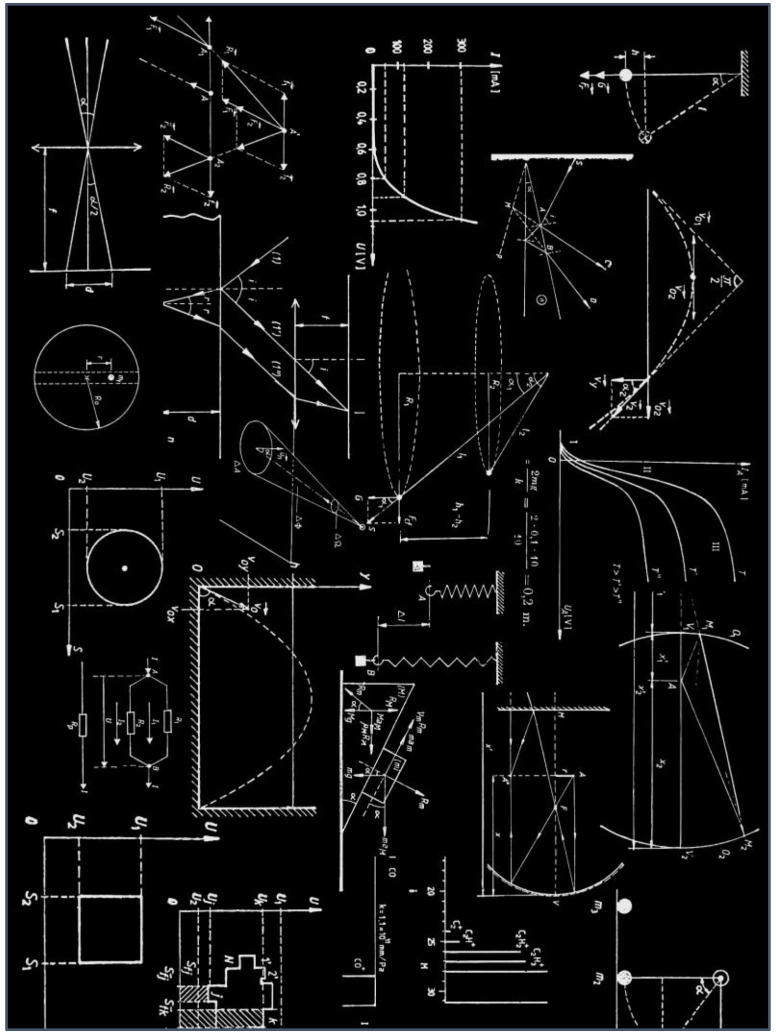
- Componente y_k para el caso *i*:

$$y_{ik} = x_{i1}v_{1k} + x_{i2}v_{2k} + \cdots + x_{ip}v_{pk}$$

- *Obtener las componentes* implica obtener el valor de los coeficientes v_{jk} (**cargas**) para cada una de las p componentes.

- Estos coeficientes se obtienen según un método iterativo, de modo que **la primera componente es la que recoge mayor parte de la comunalidad o suma de las varianzas de las variables originales**; la segunda componente es la segunda que mayor parte de “comunalidad” recoge, y así sucesivamente hasta la componente *p*.

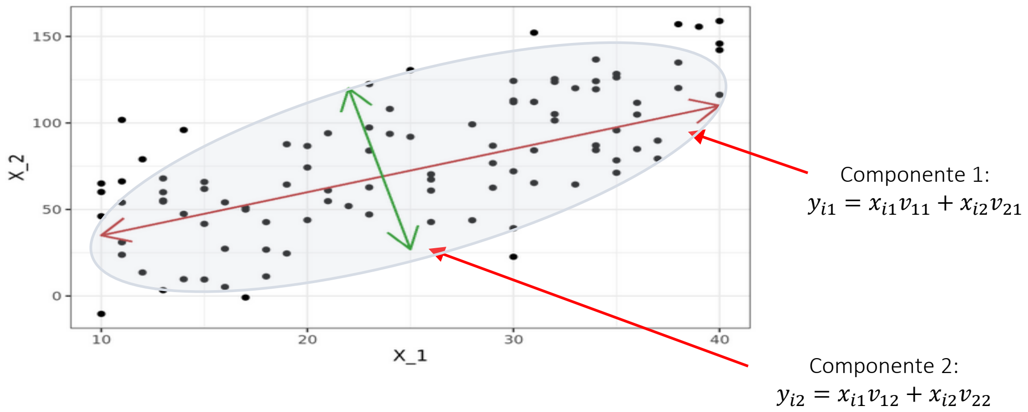
- Cada componente tiene como varianza uno de los **autovalores** de la matriz de varianzas-covarianzas de las variables originales, de mayor a menor: $\lambda_1, \lambda_2, \ldots, \lambda_p$

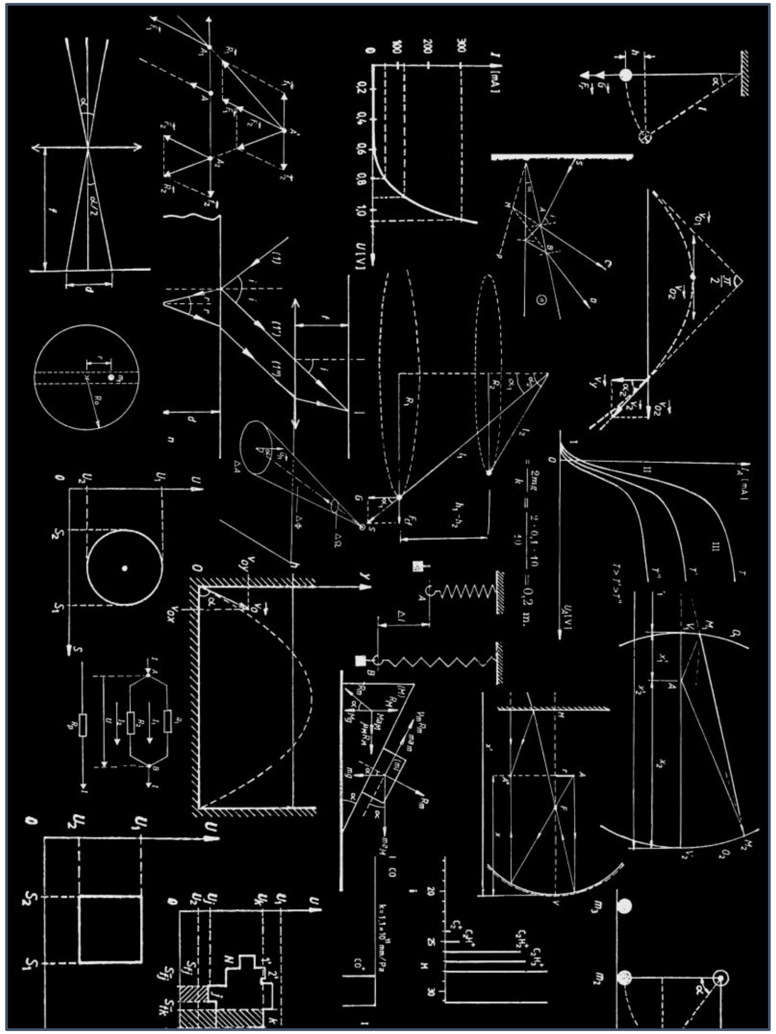


- **Ejemplo:** caso de dos únicas variables originales x_1 y x_2:

$$\text{Componente 1: } y_{i1} = x_{i1}v_{11} + x_{i2}v_{21}$$
$$\text{Componente 2: } y_{i2} = x_{i2}v_{21} + x_{i2}v_{22}$$

- Se obtienen los valores de v_{j1}, de modo que y_{i1} recoge la mayor parte de la "comunalidad" de las variables originales, que es la varianza de y_{i1}, λ_1 (mayor autovalor de la matriz de varianzas-covarianzas de las variables originales). Luego se obtienen los v_{j2}. y_{i1} y y_{i2} están incorrelacionadas (geométricamente, son ortogonales).





- **Ejemplo:** En el ejemplo inicial, teníamos 3 variables originales tipificadas.

	x_{i1} variable 1	x_{i2} variable 2	x_{i3} variable 3
$i = 1$	-1,035	-1,011	0,664
$i = 2$	-0,354	-0,328	-0,664
$i = 3$	-0,950	-0,926	-1,107
$i = 4$	-0,014	-0,157	-0,664
$i = 5$	1,432	1,381	1,550
$i = 6$	0,922	1,040	0,221

Componente 1: $y_{i1} = x_{i1}v_{11} + x_{i2}v_{21} + x_{i3}v_{31}$
Componente 2: $y_{i2} = x_{i1}v_{12} + x_{i2}v_{22} + x_{i3}v_{32}$
Componente 2: $y_{i3} = x_{i1}v_{13} + x_{i2}v_{23} + x_{i3}v_{33}$

- Cálculo de los coeficientes o **cargas**:

$$y_{i1} = x_{i1} \cdot 0{,}612 + x_{i2} \cdot 0{,}613 + x_{i3} \cdot 0{,}499$$
$$y_{i2} = x_{i1} \cdot -0{,}355 + x_{i2} \cdot -0{,}351 + x_{i3} \cdot 0{,}866$$
$$y_{i3} = x_{i1} \cdot 0{,}706 + x_{i2} \cdot -0{,}708 + x_{i3} \cdot 0{,}003$$

- Autovalor asociado a la primera componente (varianza): $\lambda_1 = 2{,}492$

- Porcentaje de la comunalidad (suma de las varianzas de las variables originales) recogido por la primera componente: 83,1%.

Retención de componentes.

- Del proceso de cálculo de las componentes se obtienen el **mismo número** de estas que de variables originales (p variables y p componentes). La diferencia es que las componentes son “variables” **incorrelacionadas** entre sí.

- Las p componentes recogen el 100% de la suma de las varianzas de las variables originales (*comunalidad*). Pero nosotros **queremos tener menos componentes** que variables originales. Estas serán las **componentes principales**.

- La suma de las varianzas de las componentes coincide con la comunalidad que, si las variables han sido tipificadas, **es igual al número de variables o componentes**.

- La **retención** de componentes consiste en decidir **con cuántas nos quedamos** (aunque se pierda un poco de comunalidad).

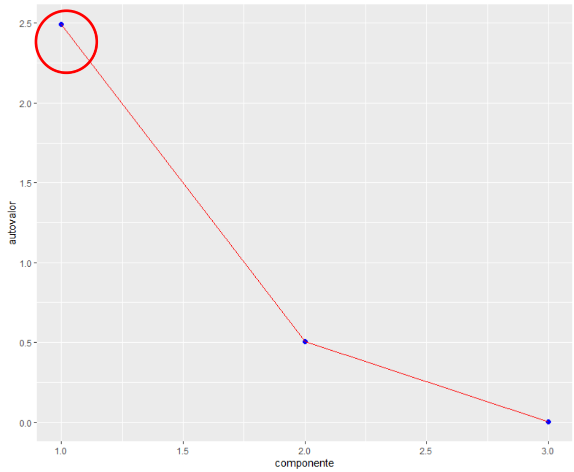
Retención de componentes.

- Hay varios procedimientos para identificar el número de componentes a retener. Uno muy extendido: retener los componentes que tengan **varianzas** (autovalores asociados) **mayores que 1**.

- En el **ejemplo**: $Var\ (y_{i1}) = \lambda_1 = \boxed{2{,}492}$

$$Var\ (y_{i2}) = \lambda_2 = 0{,}504$$

$$Var\ (y_{i3}) = \lambda_3 = 0{,}003$$



Solo se retiene la componente 1: Solo hay **1 componente principal**, que recoge el 83,1% de la comunalidad (comportamiento de los casos según las variables originales).

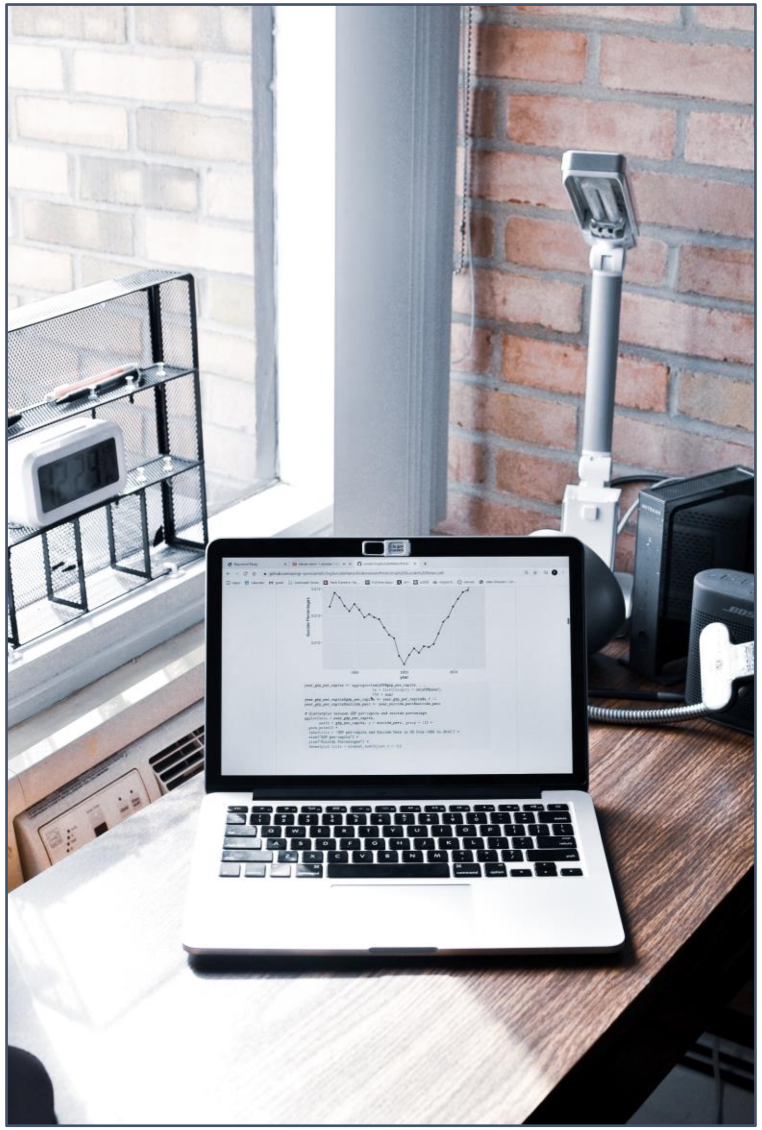
Significado de las componentes principales.

- A veces es difícil encontrar un **significado económico** a las componentes, ya que son simples combinaciones lineales de las variables originales.

- Para ello hemos de fijarnos en el valor (absoluto) de las cargas v_{jk} que las definen; teniendo en cuenta que están en la "misma escala" (porque las variables han sido tipificadas) y, por tanto, son comparables.

- En general, las componentes tendrán una interpretación asociada a las variables originales (X) con las que comparte **cargas con mayor valor absoluto**.

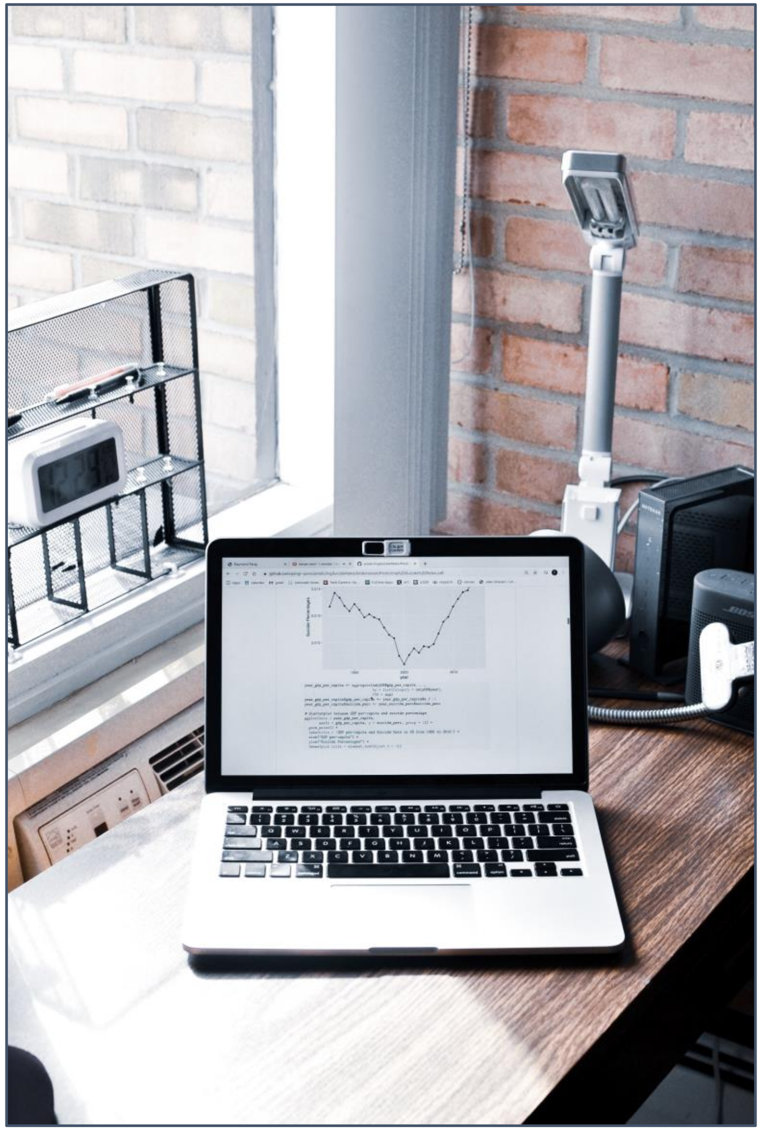
Puntuación de componentes principales.



- Con el término de “**puntuaciones**” (“scores”) nos referimos al **valor que toma cada componente principal para cada individuo o caso**.

- Estas puntuaciones son necesarias, a veces, cuando el análisis de componentes principales se usa como **etapa intermedia** de otros análisis (como el análisis clúster, o el análisis de regresión).

- Para obtener las puntuaciones de la componente k en el individuo i, no hay más que obtener los valores de las variables originales (x) para ese individuo y sustituirlas en la ecuación de la componente.



- **Ejemplo:** vimos que la primera CP era:

$$y_{i1} = x_{i1} \cdot 0{,}612 + x_{i2} \cdot 0{,}613 + x_{i3} \cdot 0{,}499$$

- Así la puntuación o score de la primera CP para el primer individuo será:

$$y_{i1} = -1{,}035 \cdot 0{,}612 - 1{,}011 \cdot 0{,}613 + 0{,}664 \cdot 0{,}499 = -0{,}922$$

¡Muchas gracias!