Análisis Inteligente de Datos

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Profesora: Mónica Cantoni

Ayudantes: Cecilia Oliva - Fabiana Rossi - Pamela Pairo

Clase 06 – Análisis de Componentes Principales

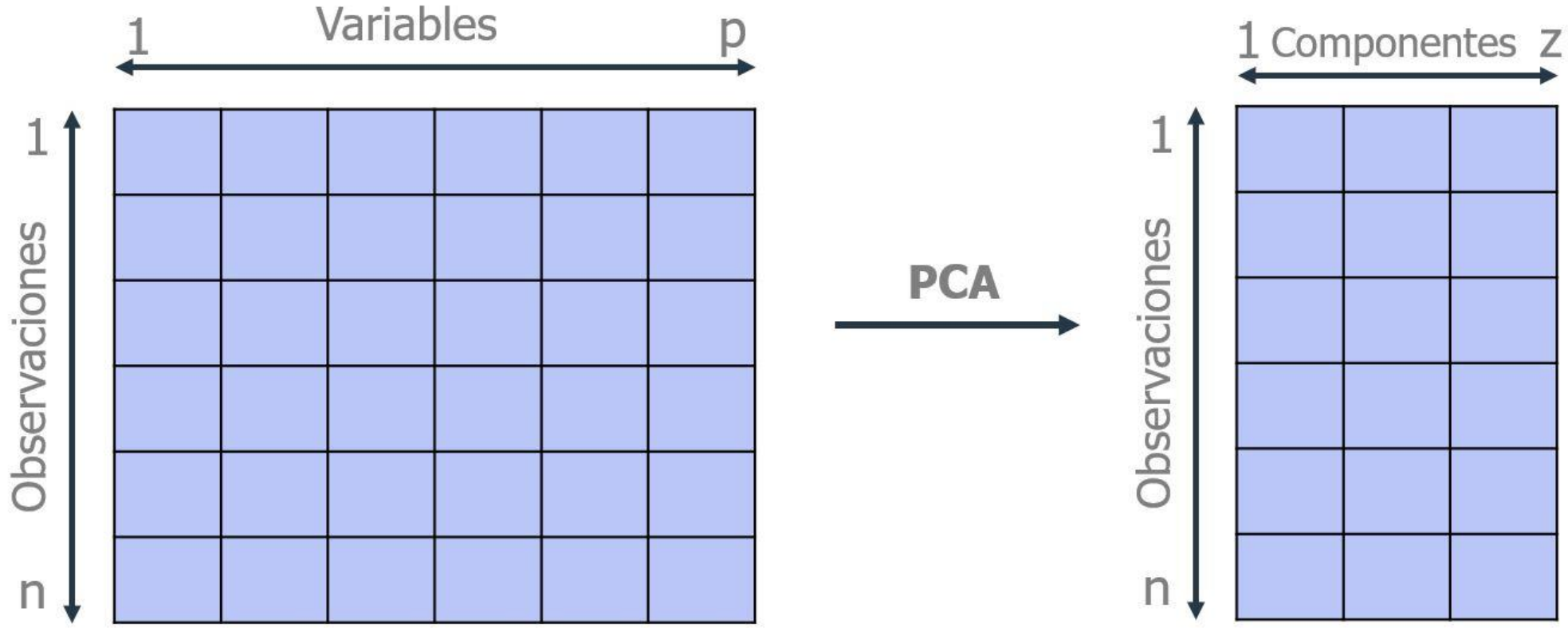
Introducción al Análisis de Componentes Principales

El Análisis de Componentes Principales conocido como PCA es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.

Supóngase que existe una muestra con n individuos cada uno con p variables $(x_1, x_2, \dots, x_p)$ es decir, el espacio muestral tiene p variables

PCA pertenece a la familia de técnicas conocida como aprendizaje no supervisado.

Componentes principales



Nociones previas

Vector en $\mathbb{R}^n$

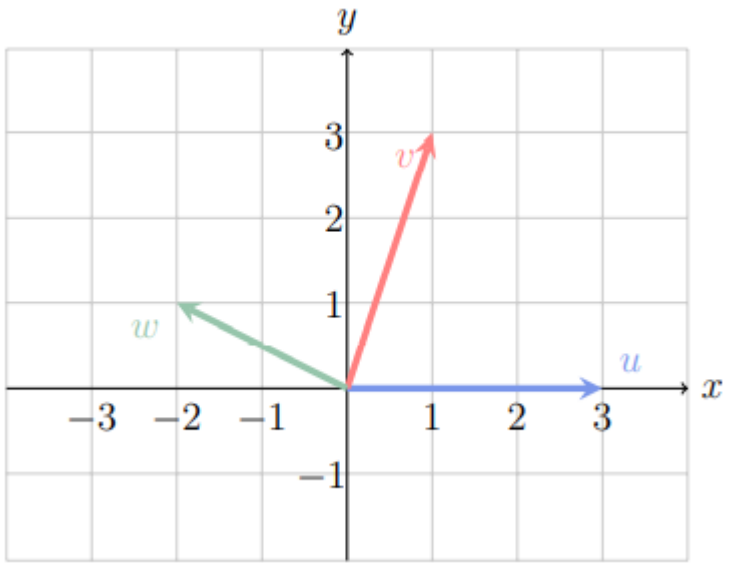
Punto de coordenadas $P = (v_1, v_2, \dots, v_n)$,

Vector $v = (v_1, v_2, \dots, v_n)$,

Vector es un segmento orientado cuyo punto inicial es el origen de coordenadas de $\mathbb{R}^n$ y el punto final es el punto P.

Nociones previas

Ejemplo:
Vectores en $\mathbb{R}^2$ cuyo origen de coordenadas es el punto $p = (0,0)$ y el punto final es $p = (x,y)$
$u = (3,0), v = (1,3)$ y $w = (-2,1)$



Vectores en coordenadas

Espacio vectorial $\mathbb{R}^n$, si $v = (v_1, v_2, \dots, v_n)$ y $w = (w_1, w_2, \dots, w_n)$ son dos vectores en $\mathbb{R}^n$ $y \alpha \in \mathbb{R}$

Se definen las operaciones en el espacio vectorial suma y el producto por un escalar de la siguiente manera:

$$v + w = (v_1 + w_1, v_2 + w_2, \dots, v_n + w_n)$$

$$\alpha v = (\alpha v_1, \alpha v_2, \dots, \alpha v_n)$$

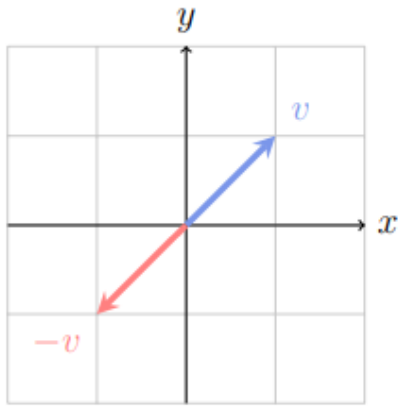
Combinación lineal de vectores v y w si existen escalares α y β tales que $u = (\alpha v + \beta w)$

Geométricamente, las posibles combinaciones lineales de un único vector son los múltiplos de ese vector (comparten la misma dirección pertenecen a la misma recta)

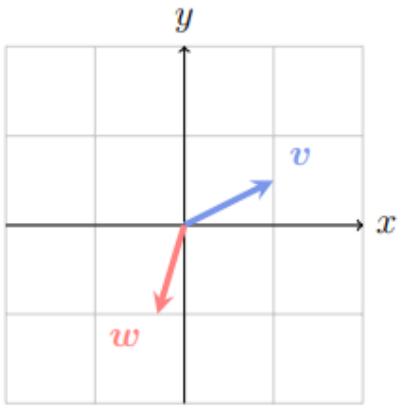
Si dos vectores no nulos en el plano $\mathbb{R}^2$ no tienen la misma dirección cualquier otro vector del plano puede escribirse como combinación lineal de ellos.

Dependencia lineal

- Los vectores son linealmente dependientes (l.d.)
- $\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n = 0$ con algún $\alpha_i \neq 0$
- Los vectores son linealmente independientes(l.i.)
- $\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n = 0$ implica que todos los escalares son iguales a cero.
- $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$



(a) Vectores l.d.



(b) Vectores l.i.

Enfoque estadístico

- Dos vectores (variables) son l.d. cuando la información que brinda uno de ellos es redundante con la información que brinda el otro. Uno de ellos es múltiplo de otro.
- Ejemplo: Salario en pesos y salario en dólares.
- Si tres vectores son l.d. significa que la información que brindan es una combinación lineal de la información de las otras dos variables.

Ejemplo: Nadadores

Nadador	Tramo 1	Tramo 2	Tramo 3	Tramo 4
1	10	10	13	12
2	12	12	14	15
3	11	10	14	13
4	9	9	11	11
5	8	8	9	8
6	8	9	10	9
7	10	10	8	9
8	11	12	10	9
9	14	13	11	11
10	12	12	12	10
11	13	13	11	11
12	14	15	14	13
13	10	10	12	13
14	15	14	13	14

Ejemplo: Nadadores

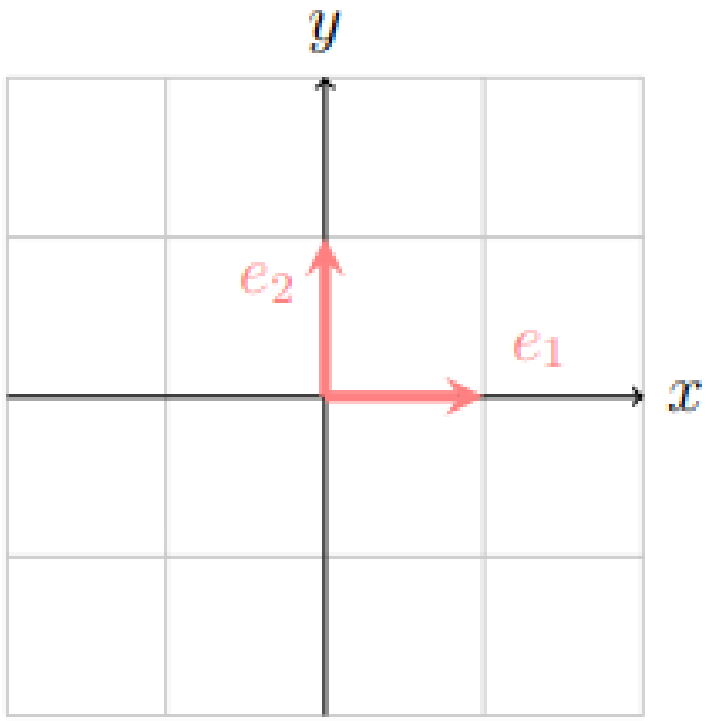
Podríamos estar interesados, por ejemplo, en:

- $w_1 = \frac{1}{2}v_1 + \frac{1}{2}v_2$ el tiempo medio empleado por cada nadador en los primeros dos tramos.
- $w_2 = \frac{1}{2}v_3 + \frac{1}{2}v_4$ el tiempo medio empleado por cada nadador en los últimos dos tramos.
- $w_3 = w_1 - w_2$ la diferencia entre los dos promedios anteriores.

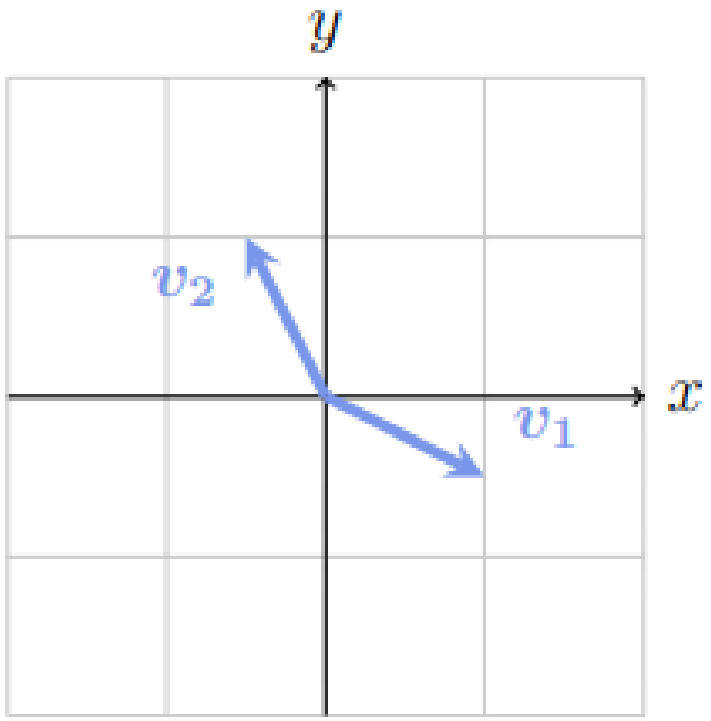
Ejemplo: Nadadores

Nadador	Tramo 1	Tramo 2	Tramo 3	Tramo 4	w1	w2	w3
1	10	10	13	12	10	12.5	-2.5
2	12	12	14	15	12	14.5	-2.5
3	11	10	14	13	10.5	13.5	-3
4	9	9	11	11	9	11	-2
5	8	8	9	8	8	8.5	-0.5
6	8	9	10	9	8.5	9.5	-1
7	10	10	8	9	10	8.5	1.5
8	11	12	10	9	11.5	9.5	2
9	14	13	11	11	13.5	11	2.5
10	12	12	12	10	12	11	1
11	13	13	11	11	13	11	2
12	14	15	14	13	14.5	13.5	1
13	10	10	12	13	10	12.5	-2.5
14	15	14	13	14	14.5	13.5	1

Base de un vector

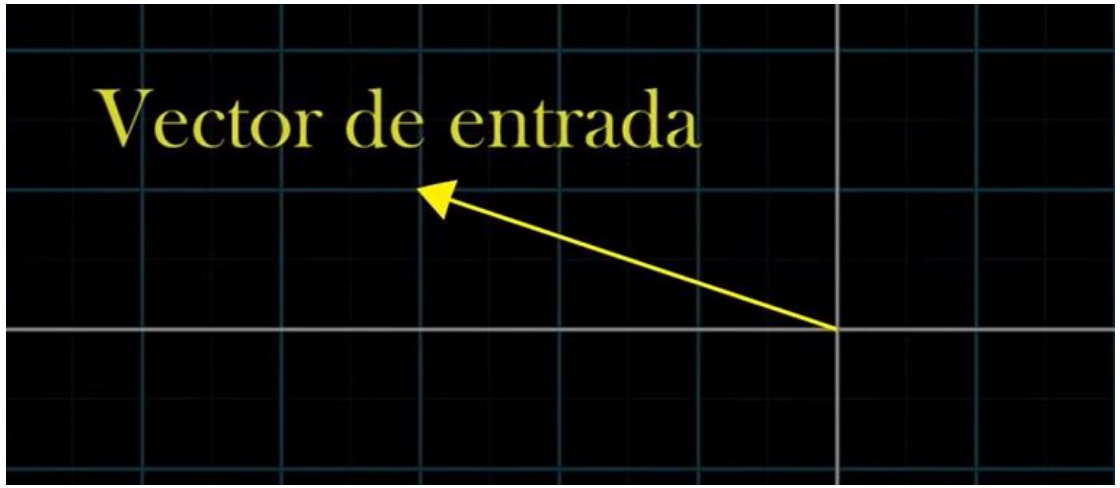


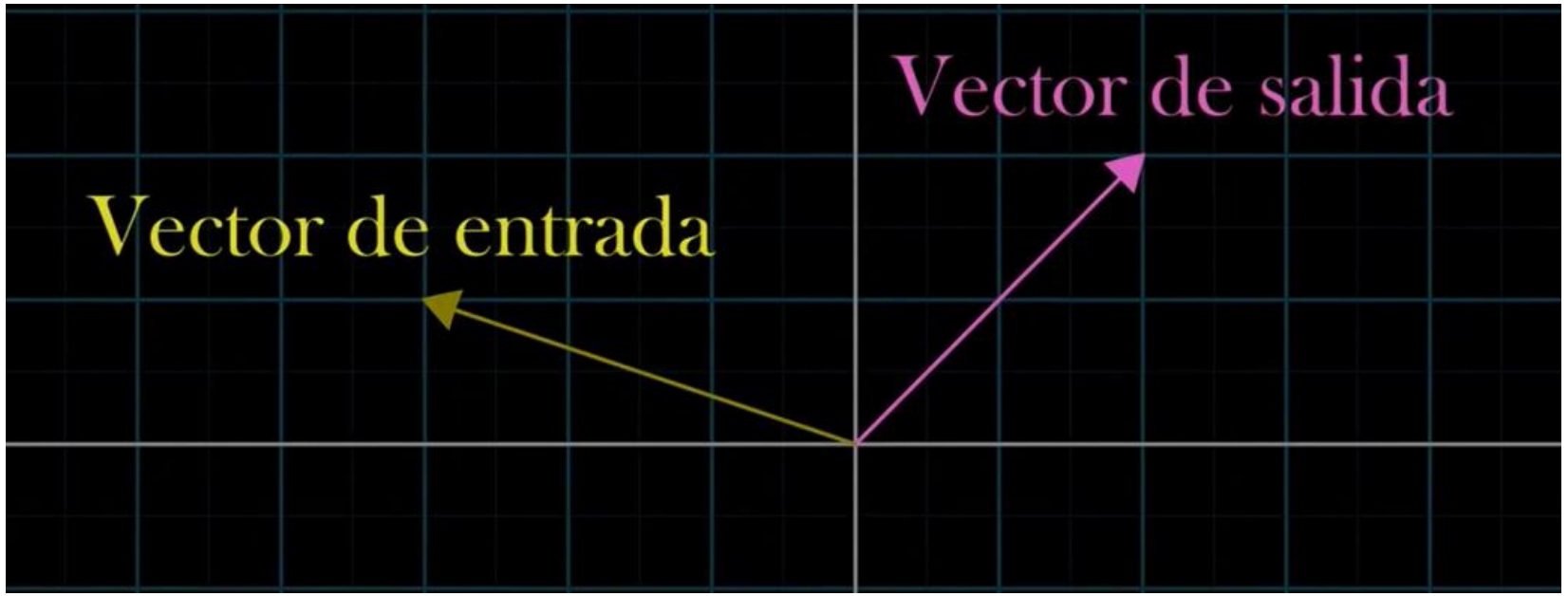
(a) Base canónica



(b) Base $\{v_1, v_2\}$

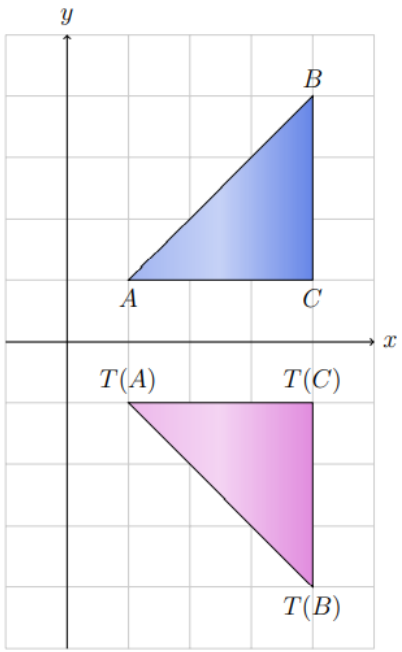
Transformaciones



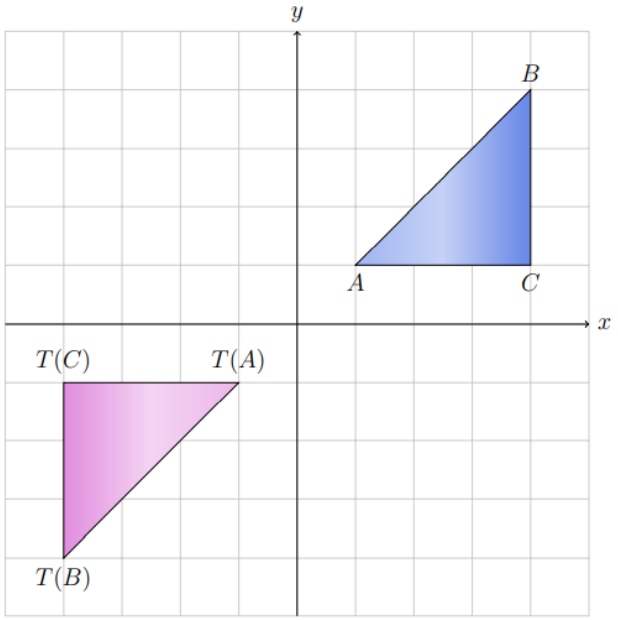


Transformaciones

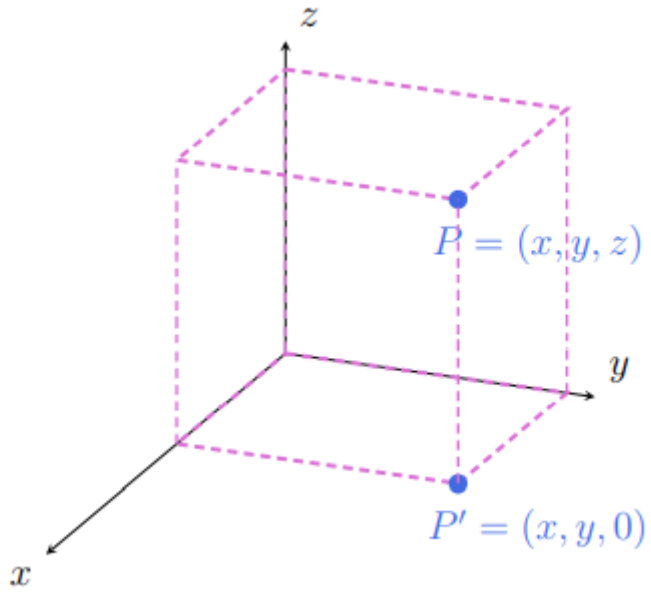
Simetría



Rotación



Proyección ortogonal



Autovalores (valores propios) y autovectores (vectores propios)

Ejemplo

Polinomio característico $Det(A - \lambda I) = 0$

$$A = \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix}$$

$$Det(A - \lambda I) = \left| \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \left| \begin{pmatrix} 2 - \lambda & 3 \\ 3 & -6 - \lambda \end{pmatrix} \right| =$$

$$(2 - \lambda)(-6 - \lambda) - 9 = \lambda^2 + 4\lambda - 21.$$

$$\lambda_1 = 3 \ \ \lambda_2 = -7$$

Autovalores (valores propios) y autovectores (vectores propios)

Para hallar los autovectores resolvemos los siguientes sistemas homogéneos

Para $\lambda_1 = 3$

$$(A - 3I)\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \leftrightarrow \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 3\begin{pmatrix} x \\ y \end{pmatrix}$$

Resolviendo el sistema

Las dos ecuaciones son equivalentes entre sí

Nos quedamos con la primera ecuación $-x + 3y = 0$

$$x = 3y$$

Autovalores son $(3y, y) = y(3,1)$

Se realiza lo mismo para $\lambda_2 = -7$

Autovalores son $(x, -3x) = x(1, -3)$

La norma de un vector es la longitud del vector sirve para cuantificar el tamaño del vector

$$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{x_1^2} + \boldsymbol{x_2^2} + \ldots + \boldsymbol{x_n^2}}$$

Cuando se quiere cuantificar la variabilidad de un conjunto de datos se cuantifica el tamaño de la matriz de covarianzas.

Relación entre autovalores, traza y determinante

Matriz cuadrada $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$

Traza suma de los elementos de la diagonal $tr(A) = a + d$

Determinante $Det(A) = ad - cb$

La traza y el determinante dan una idea del tamaño de la variabilidad del conjunto y están relacionadas con los **autovalores** de la matriz

Polinomio característico

$$xA(\lambda) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - (a + d)\lambda + ad - bc$$
$$= \lambda^2 - tr(A)\lambda + \det(A)$$

Las raíces del polinomio, $\lambda_1 \; y \; \lambda_2$, son los **autovalores** de A.

Motivación del problema de reducción de la dimensión

Ejemplo:

Factores de riesgo de sufrir enfermedad coronaria

- X1: presión arterial media en mm/Hg,
- X2: edad en años,
- X3: peso en kg,
- X4: superficie corporal en m2 ,
- X5: tiempo transcurrido desde el diagnóstico de hipertensión en años
- X6: pulsaciones por minuto,
- X7: medida asociada al stress.

Datos

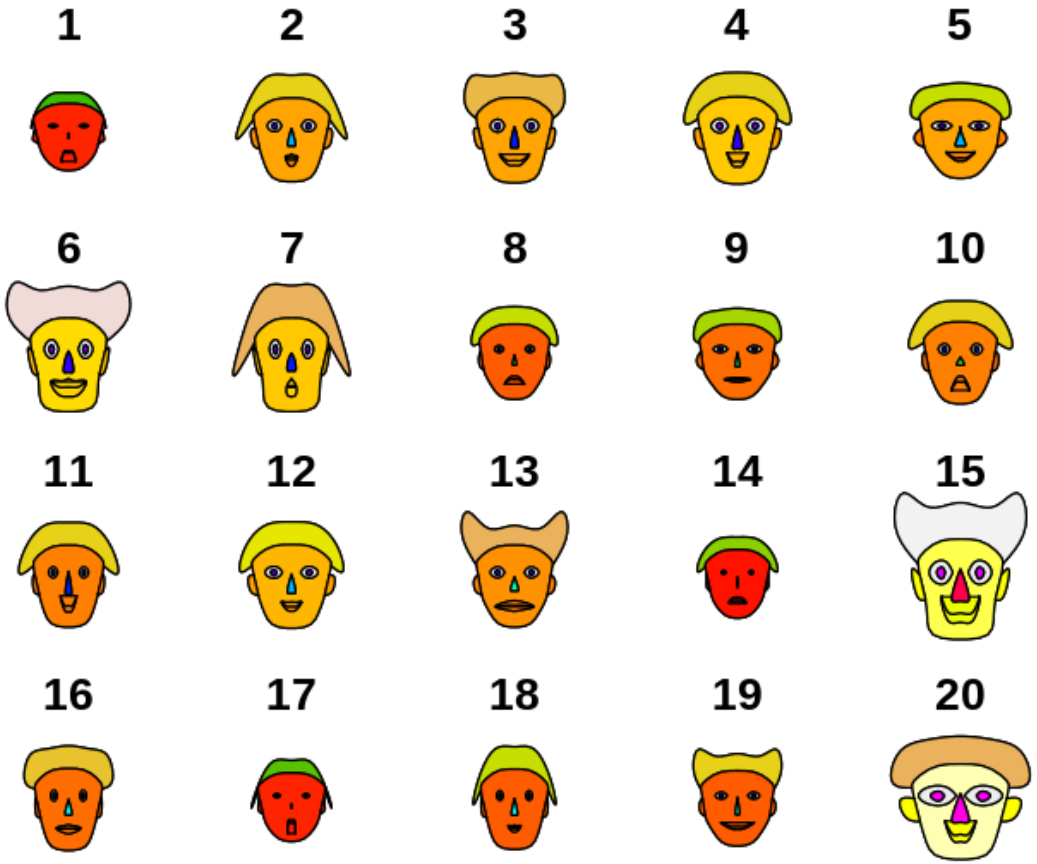
Caso	Presión	Edad	Peso	Superficie	Tiempo	Pulso	Stress	Sexo	sexo_lab
1	105	47	85.4	1.75	5.1	63	33	F	1
2	115	49	94.2	2.1	3.8	70	14	F	1
3	116	49	95.3	1.98	8.2	72	10	F	1
4	117	50	94.7	2.01	5.8	73	99	F	1
5	112	51	89.4	1.89	7	72	95	F	1
6	121	49	99.5	2.25	9.3	71	10	F	1
7	121	48	99.8	2.25	2.5	69	42	F	1
8	110	47	90.9	1.9	6.2	66	8	F	1
9	110	49	89.2	1.83	7.1	69	62	M	2
10	114	48	92.7	2.07	5.6	64	35	M	2
11	114	47	94.4	2.07	5.3	74	90	M	2
12	115	50	94.1	1.98	5.6	71	21	M	2
13	114	49	91.6	2.05	10.2	68	47	M	2
14	106	45	87.1	1.92	5.6	67	80	M	2
15	125	52	101.3	2.19	10	76	98	M	2
16	114	46	94.5	1.98	7.4	69	95	M	2
17	106	46	87	1.87	3.6	62	18	M	2
18	113	46	94.5	1.9	4.3	70	12	M	2
19	110	48	90.5	1.88	9	71	99	M	2
20	122	56	95.7	2.09	7	75	99	M	2

PCA Factores de riesgo

```
[ ] Importance of components:
                           PC1    PC2     PC3     PC4     PC5     PC6
    Standard deviation     7.6284 2.8207 1.91826 1.82213 0.32526 0.04855
    Proportion of Variance 0.7944 0.1086 0.05023 0.04532 0.00144 0.00003
    Cumulative Proportion  0.7944 0.9030 0.95320 0.99852 0.99997 1.00000
    Importance of components:
                           PC1    PC2     PC3     PC4     PC5     PC6
    Standard deviation     7.6284 2.8207 1.91826 1.82213 0.32526 0.04855
    Proportion of Variance 0.7944 0.1086 0.05023 0.04532 0.00144 0.00003
    Cumulative Proportion  0.7944 0.9030 0.95320 0.99852 0.99997 1.00000
    Standard deviations (1, .., p=6):
    [1] 7.62838924 2.82066407 1.91826271 1.82213284 0.32526410 0.04855139

    Rotation (n x k) = (6 x 6):
                    PC1         PC2          PC3         PC4         PC5
    Presión    0.70226831 -0.21284356 -0.314858408  0.05926993  0.59494998
    Edad       0.21961612  0.42752948 -0.595557787  0.44723251 -0.46067549
    Peso       0.53149633 -0.45503300  0.222900905 -0.20018644 -0.64717780
    Superficie 0.01466270 -0.01863294 -0.005173188 -0.00373921  0.09082754
    Tiempo     0.09602579  0.40220856 -0.292824381 -0.86127213 -0.03805518
    Pulso      0.40824463  0.63461975  0.640872384  0.12079683  0.07263385
                       PC6
    Presión    -0.070018705
    Edad        0.045380664
    Peso        0.043105583
    Superficie  0.995563877
    Tiempo      0.004828898
    Pulso       0.003022153
```

Representación gráfica – Caras de Chernoff
Ejemplo factores de riesgo de enfermedad cardiovascular



PCA Nadadores

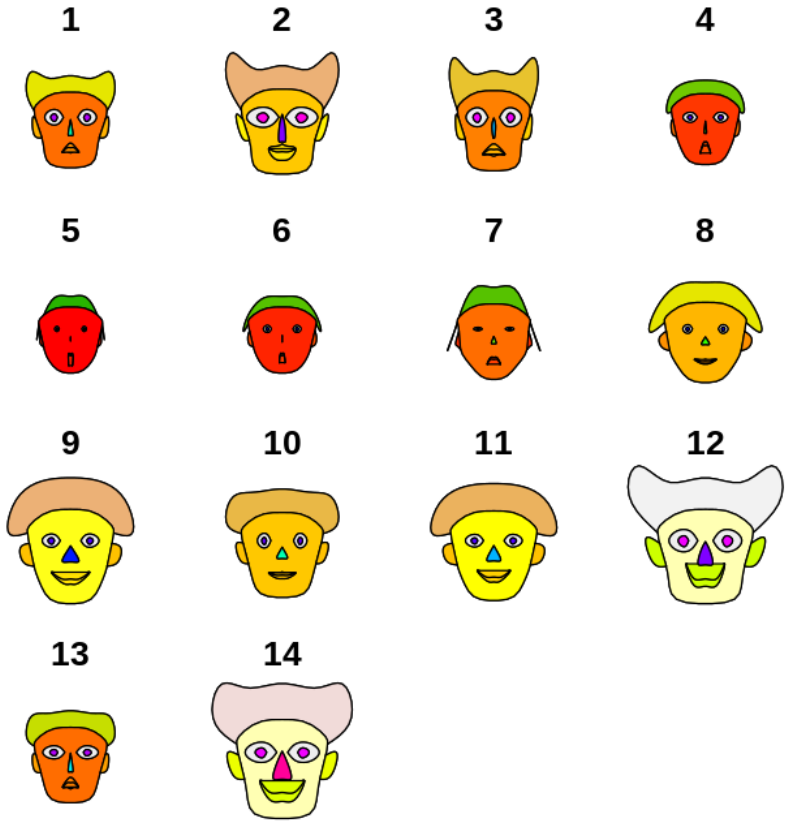
Análisis de Componentes Principales

Ejemplo Nadadores

```
[17] nad=read_excel("/content/nadadores.xlsx")
     nadadores=data.frame(nad[,2:5])
     nad.pca.cov=prcomp(nadadores, center=TRUE, scale=FALSE)
     nad.pca.cor=prcomp(nadadores, center=TRUE, scale=FALSE)
     summary(nad.pca.cov)
     summary(nad.pca.cor)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     3.5840 1.9858 0.7030 0.42241
Proportion of Variance 0.7356 0.2258 0.0283 0.01022
Cumulative Proportion  0.7356 0.9615 0.9898 1.00000
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     3.5840 1.9858 0.7030 0.42241
Proportion of Variance 0.7356 0.2258 0.0283 0.01022
Cumulative Proportion  0.7356 0.9615 0.9898 1.00000
```

Representación gráfica – Caras de Chernoff
Ejemplo: Nadadores



Análisis de componentes principales

- La información se presenta en forma tabular
- Todas las variables juegan el mismo papel no existen variables independientes ni dependientes.
- El objetivo es reducir la dimensión del problema, descartar información redundante.
- Permite visualizar información multidimensional.
- La matriz de covarianzas representa la variabilidad del conjunto de datos y está influenciada por las unidades de medición. Es recomendable basarse en la matriz de correlaciones.
- Se debe determinar el número de componentes principales a seleccionar para finalizar el proceso.
- Los análisis confirmatorios permiten evaluar la estabilidad de las componentes principales y al mismo tiempo brindan la posibilidad de detectar outliers.
- Puede ser útil para detectar anormalidad en las observaciones.

Definición de las componentes

- Las componentes principales permiten resolver problemas de multidimensionalidad.
- Las componentes principales resultan de la combinación lineal de las variables originales y definen un nuevo espacio de representación de las observaciones.
- Las correlaciones entre las componentes principales y las variables originales analizan las similitudes y diferencias entre los individuos de ese nuevo espacio.

Análisis de Componentes Principales PCA

Notación

n = cantidad de casos

x_i= variables originales

$\bar{x}_i$ = media de la variable original

x_{ic}= variables centradas $x_{ic} = x_i - \bar{x}_i$

$\bar{x}_{ic}$ = media de la variable centrada

s_i^2= varianza de la variable original

s_{ic}^2 varianza de la variable centrada

s_i = desviación estándar de la variable original

s_{ic} = desviación estándar de la variable centrada

$SSCP$ = matriz de suma de cuadrados de los productos cruzados

S = matriz de covarianzas

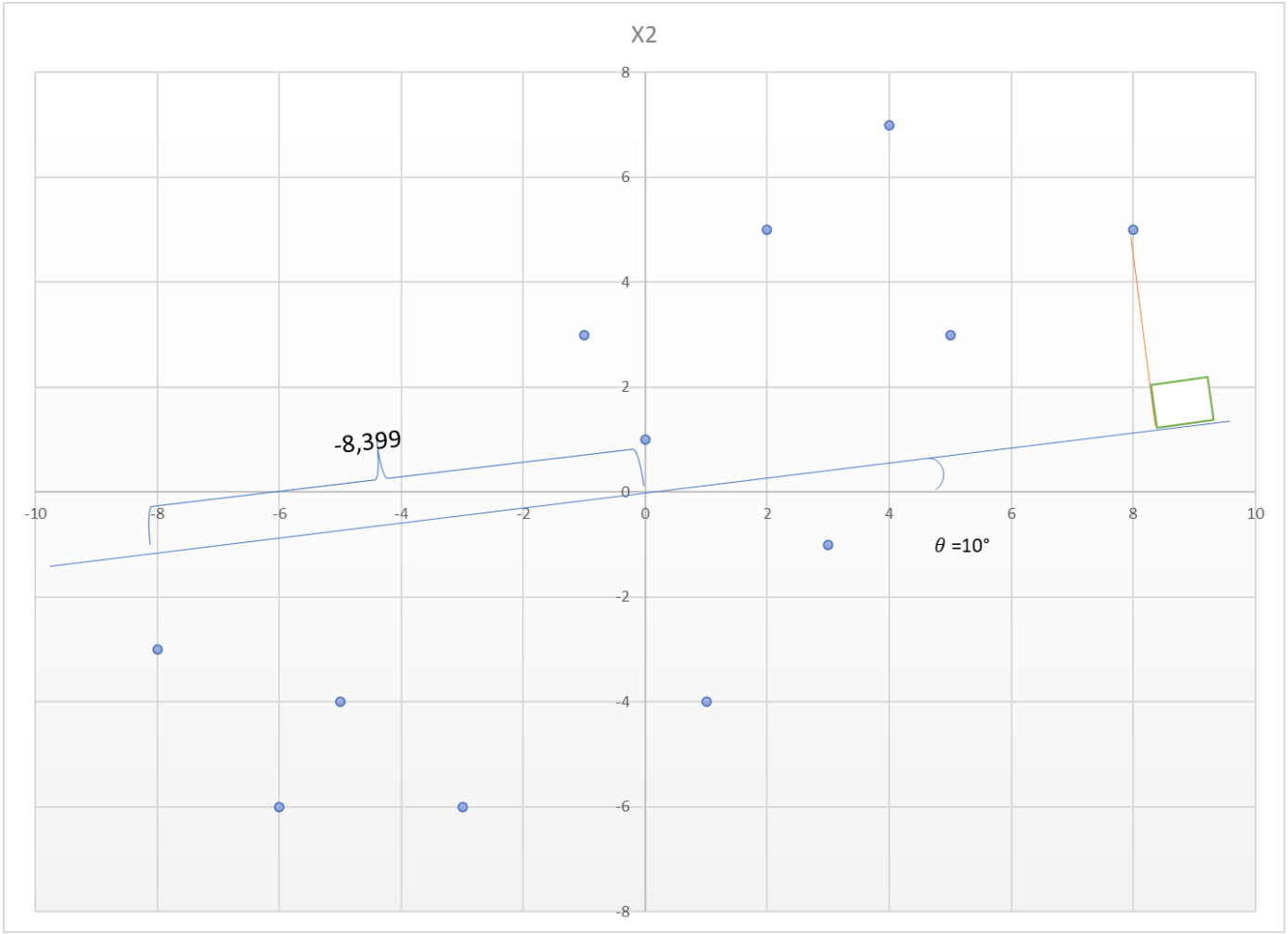
$R = \rho$ = matriz de correlaciones

θ = ángulo de rotación en radianes

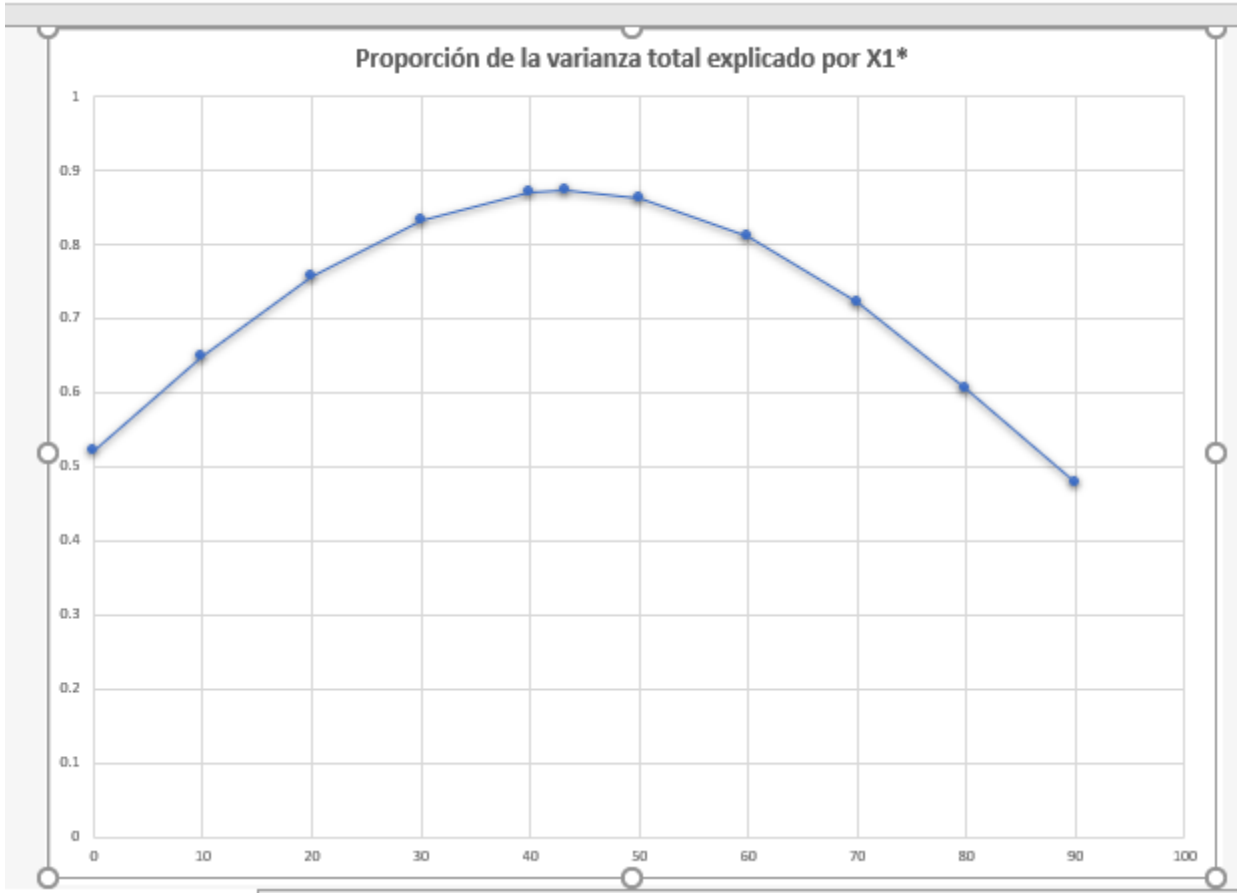
x_i^* = variable proyectada

$$x_1^* = \cos(\theta)\, x_1 + sen(\theta) x_2$$
$$x_2^* = -\text{sen}(\theta)\, x_1 + cos(\theta) x_2$$

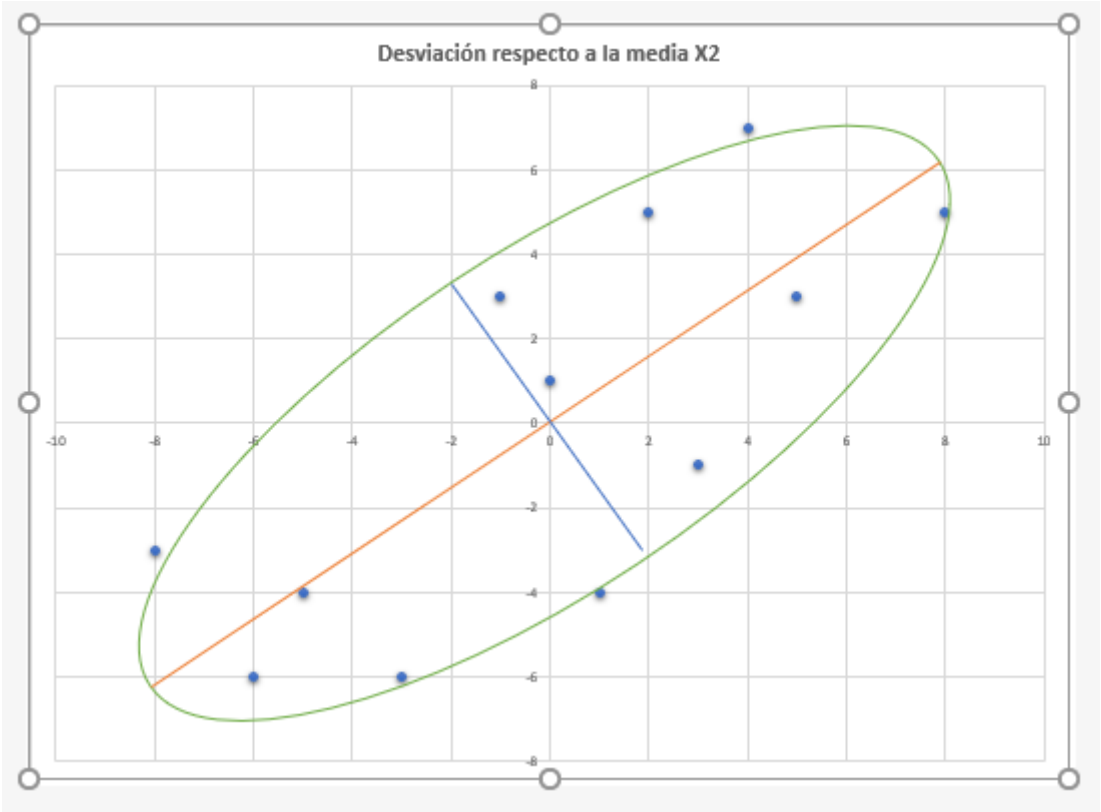
Proyecciones



Proporción de la varianza total explicada



Autovalores y Autovectores



Representación gráfica - Dispersograma

- Dispersograma
- Dispersograma 3D
- Dispersograma 3D por grupos
- Ejes principales

Variabilidad de las componentes principales

- La variabilidad total del conjunto de datos, es la suma de las varianzas de cada una de las variables; es decir, la traza de la matriz de covarianzas de las variables originales. Simbólicamente, la variabilidad total se calcula como

$$tr(\Sigma) = \Sigma_{11} + \Sigma_{22} + \cdots + \Sigma_{pp}.$$

Cantidad de componentes principales

Criterio 1: Porcentaje de variabilidad explicada

Es un porcentaje de variabilidad mínimo que se desea explicar y se toman las primeras m componentes que alcanzan este porcentaje de explicación. Si se desea explicar el q% de la variabilidad, se eligen k componentes de manera tal que:

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{tr(\Sigma)} \geq \frac{q}{100}$$

Siendo

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_{k-1}}{tr(\Sigma)} < \frac{q}{100}$$

No se trabaja con la matriz de covarianza de las variables Σ, se trabaja su estimación $\hat{\Sigma}$ que es la matriz de covarianza muestral.

Se puede calcular los autovalores y los autovectores de esta matriz.

Criterio 2: Criterio de Kaiser

Este criterio consiste en retener las m principales componentes tales que sus autovalores resulten iguales o mayores a 1

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 1 \; y \; \lambda_{m+1} < 1$$

Algunos autores recomiendan utilizar

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0.7$$

Extendiendo este criterio a la matriz de covarianzas se eligen las m primeras componentes tales que:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq \frac{tr(\Sigma)}{p} \quad \text{y } \lambda_{m+1} < \frac{tr(\Sigma)}{p}$$

Puede considerarse utilizar como cota inferior a $\frac{0.7}{p} tr(\Sigma)$

Criterio 3: Criterio del bastón roto

- Si la proporción de variabilidad explicada por $Y_1, Y_2, \ldots, Y_m$ se estabiliza a partir de un cierto valor de m, entonces aumentar la dimensión no aportaría cambios significativos.
- El gráfico de sedimentación o ***scree plot*** representa la secuencia de valores propios de la matriz de covarianzas ordenados de mayor a menor.
- La sugerencia de este criterio es seleccionar las componentes previas a la zona de acumulación de sedimentos.

Criterio 4: Prueba de esfericidad

- Se aplica sólo cuando $\frac{n}{p} < 5$
- Si las observaciones provienen de una distribución Normal p-variada y las variables son independientes no existen direcciones de máxima variabilidad.
- La distribución tiene forma de esfera.
- El test se basa en el estadístico Chi-Cuadrado y se aplica en forma secuencial.
- H0 a partir de m no hay direcciones de máxima variabilidad, a partir de m la distribución es esférica.

Estimación de las componentes principales

La primer componente principal capta el 73% de la varianza total.

Las dos primeras componentes logran captar el 96% de la variabilidad total del conjunto

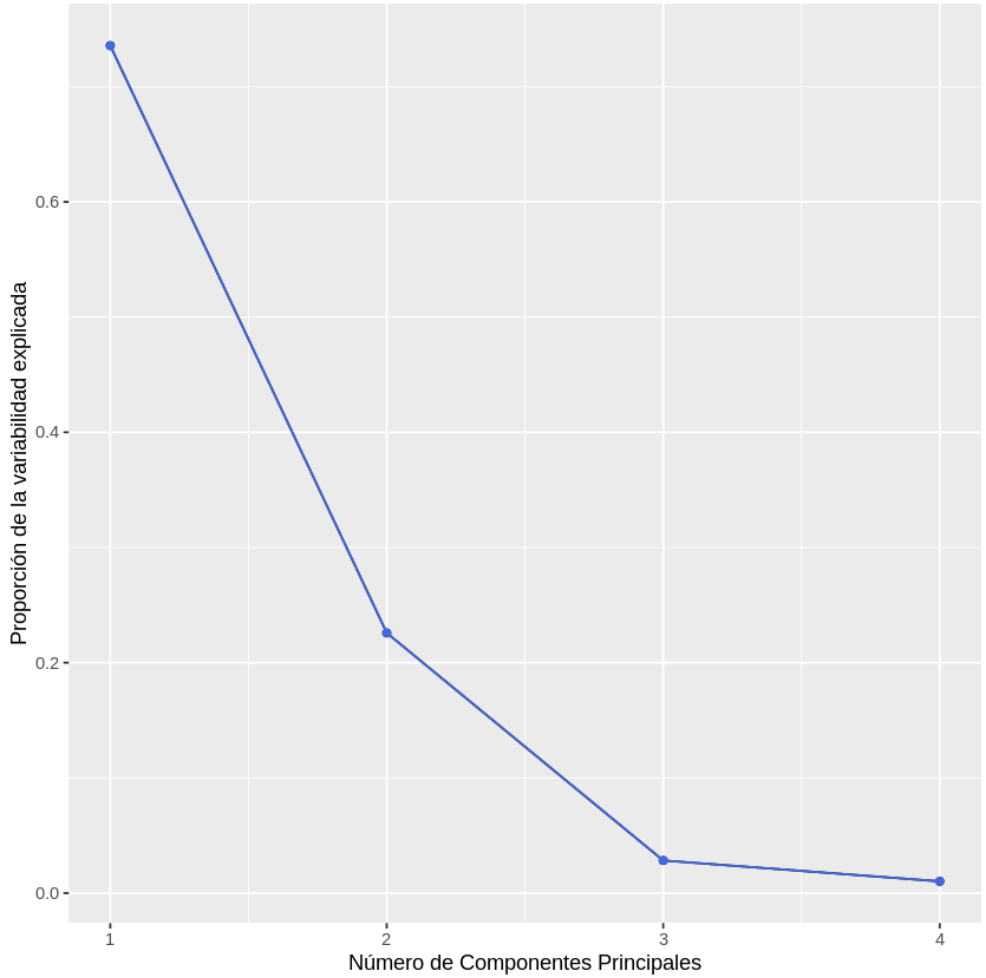
Los autovalores disminuyen considerablemente a partir de la tercer componente y alcanzan valores por debajo de 1.

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     3.5840 1.9858 0.7030 0.42241
Proportion of Variance 0.7356 0.2258 0.0283 0.01022
Cumulative Proportion  0.7356 0.9615 0.9898 1.00000
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7098 0.9573 0.34831 0.19690
Proportion of Variance 0.7309 0.2291 0.03033 0.00969
Cumulative Proportion  0.7309 0.9600 0.99031 1.00000
Standard deviations (1, .., p=4):
[1] 3.5840449 1.9858213 0.7030236 0.4224119

Rotation (n x k) = (4 x 4):
                PC1        PC2        PC3        PC4
Tramo..1 0.5646186 -0.4377769  0.1981595  0.6710365
Tramo..2 0.5070915 -0.4847300 -0.2170857 -0.6787996
Tramo..3 0.4282533  0.5107620 -0.7218770  0.1860507
Tramo..4 0.4905742  0.5590263  0.6264984 -0.2330800
```

Estimación de las componentes principales

- Las dos últimas componentes explican una proporción de variabilidad mucho menor que la que explican las dos primeras.
- La segunda componente explica algo más el 20% de la variabilidad total.



Estimación de las componentes principales

- **Criterio 1:** Si se quiere explicar el 75% de la variabilidad total del conjunto de los nadadores se deben considerar las dos primeras componentes.
- **Criterio 2:** Si se consideran los autovalores mayores que 1 de las variables estandarizadas se debe tomar una sola componente. Si se consideran los mayores que 0.7 se deberían tomar las dos primeras. Los autovalores corresponden a la varianza de la componente y por lo tanto deben elevarse al cuadrado.
- **Criterio 3:** El grafico de sedimentación presenta un quiebre en la segunda componente, coincide con los criterios anteriores.
- **Criterio 4:** El criterio de esfericidad elige las dos primeras componentes principales y rechaza con nivel 0.05 en el tercer test. La secuencia de los p-valores es: 0.2805 - 0.1806 - 0.0908 -0.0456

Escalas de medida

- Se toman **variables estandarizadas** (matriz estandarizada por columnas) cuando las escalas de medida de las variables son diferentes. Se centran las variables y se dividen por el desvío estándar.
- En este caso las componentes estarían calculadas por las **matriz de correlaciones.**
- Cuando las componentes principales se calculan a partir de la **matriz de covarianzas** los factores de carga **dependen de la escala de medida** de las variables y son difíciles de interpretar.
- Si se calculan a partir de la matriz de correlaciones las cargas (loading) son las correlaciones entre las componentes principales y las variables originales.
- Las contribuciones relativas del factor al elemento son las correlaciones al cuadrado. Miden la contribución del elementos a la componente principal.
- Las **componentes son combinaciones lineales de las variables originales** y se espera que pocas recojan la mayor parte de la variabilidad de los datos, obteniéndose así una reducción de la dimensión del problema.

Cargas o loadings

	PC1	PC2	PC3	PC4
Tramo..1	0.5646186	-0.4377769	0.1981595	0.6710365
Tramo..2	0.5070915	-0.4847300	-0.2170857	-0.6787996
Tramo..3	0.4282533	0.5107620	-0.7218770	0.1860507
Tramo..4	0.4905742	0.5590263	0.6264984	-0.2330800

Si la carga (coeficiente o *loading)* de una de las variables en la componente principal es positiva significa correlación positiva entre la variable y la componente. El coseno del ángulo formado por la componente y la variable es positivo.
Si la carga es positiva un individuo que tenga una puntuación alta en esa variable tendrá valores más altos en esa componente que otro individuo que tiene menor valor en esa variable y valores similares al primero en las restantes variables.
Si la carga es negativa, la variable se correlaciona en forma negativa con la primera componente.
Cuando la carga de una variable es negativa para dos individuos con puntuaciones similares en las restantes variables el que tenga puntuación más alta de los dos en esta variable se ubicará en un valor menor de la componente.
Variables originales estandarizadas

$$Z_i = \frac{x_i - \bar{x}}{s_x}$$

Puntajes o scores para la primera componente principal

$$Y_1 = 0.56Z_1 + 0.51Z_2 + 0.43\,Z_3 + 0.49Z_4$$

Scores para la segunda componente principal

$$Y_2 = -0.44Z_1 - 0.48Z_2 + 0.51\,Z_3 + 0.55Z_4$$

Datos estandarizados

Nadador	Tramo 1 est.	Tramo 2 est.	Tramo 3 est.	Tramo 4 est.
1	-0.5458	-0.5832	0.7479	0.3357
2	0.3531	0.3774	1.2715	1.7456
3	-0.0963	-0.5832	1.2715	0.8057
4	-0.9952	-1.0635	-0.2992	-0.1343
5	-1.4446	-1.5438	-1.3463	-1.5442
6	-1.4446	-1.0635	-0.8227	-1.0742
7	-0.5458	-0.5832	-1.8698	-1.0742
8	-0.0963	0.3774	-0.8227	-1.0742
9	1.2520	0.8576	-0.2992	-0.1343
10	0.3531	0.3774	0.2244	-0.6042
11	0.8026	0.8576	-0.2992	-0.1343
12	1.2520	1.8182	1.2715	0.8057
13	-0.5458	-0.5832	0.2244	0.8057
14	1.7015	1.3379	0.7479	1.2756

Puntajes o scores

PC1	PC2	PC3	PC4
-0.0424	1.1107	-0.2696	0.0433
1.8559	1.1645	0.3049	-0.1938
0.6790	1.4109	-0.2274	0.3344
-1.2579	0.7918	0.1402	-0.0110
-2.9392	0.0080	-0.0879	0.1432
-2.2122	0.2592	-0.2209	-0.1978
-2.0171	-0.9473	0.5654	-0.0462
-0.7957	-1.1142	-0.2569	-0.2048
0.8640	-1.2438	0.1738	0.2945
0.1809	-0.5419	-0.5731	0.1668
0.6308	-1.0394	0.0923	-0.0204
2.5733	-0.4693	-0.4505	-0.3071
-0.0647	1.0710	0.4077	-0.1415
2.5453	-0.4601	0.4020	0.1403

Estadísticos descriptivos

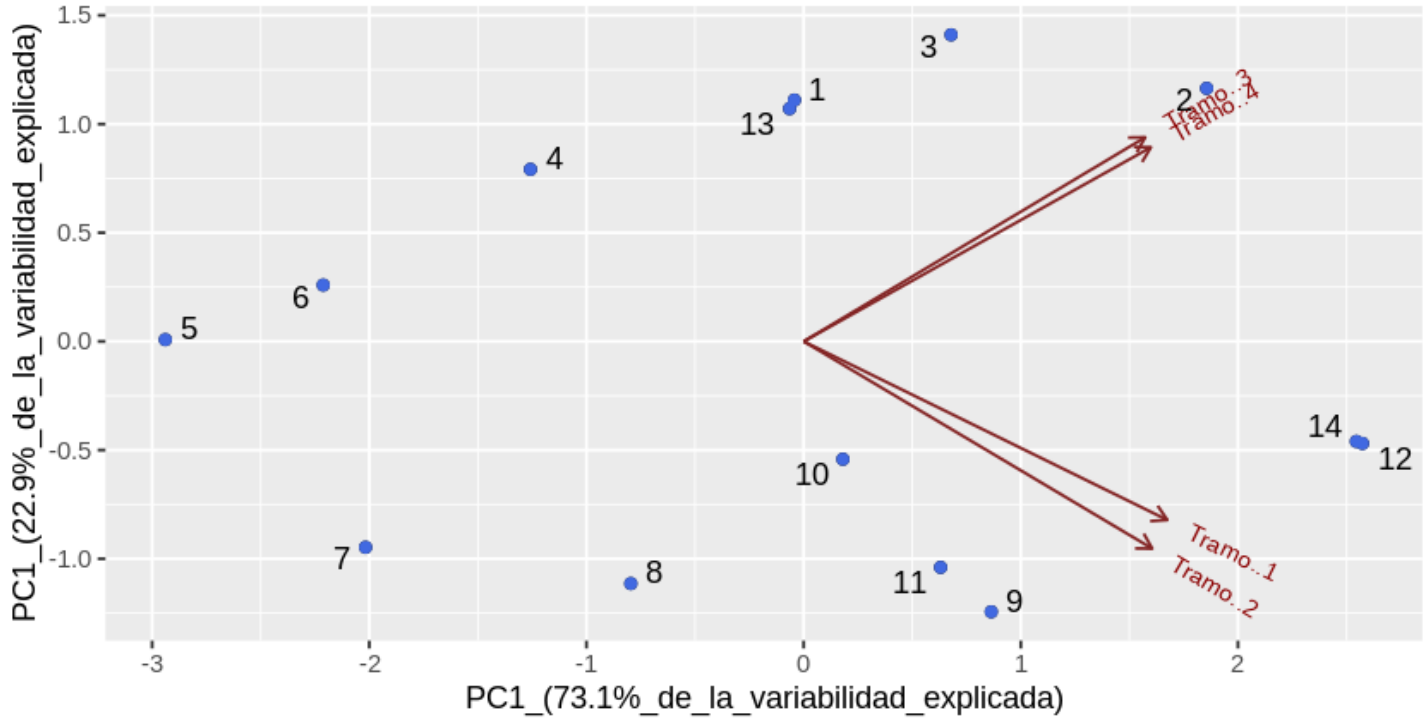
Tramo	Media	Desvío	n
1	11.21	2.22	14
2	11.21	2.08	14
3	11.57	1.91	14
4	11.29	2.12	14

Interpretación de las componentes principales

- La primera componente tiene todas las cargas positivas, por lo cual se la considera una componente de tamaño. Es decir que un individuo tendrá puntuación alta en esta componente si ha tardado mucho en todos los tramos o bien si la suma de tiempos que le ha llevado correr la carrera completa es alta. Por el contrario, los individuos que han hecho “buenos tiempos” tendrán valores bajos en esta componente. Esta componente podría denominarse ‘rapidez’.
- La segunda componente es en cambio, un contraste, se dice que es una componente de forma. Contrasta los tiempos de los primeros dos tramos con los de lo últimos dos. Un individuo tendrá alta esta componente si tardó poco al principio y desaceleró en los últimos tramos. Por el contrario, si un individuo gastó toda su energía en los dos primeros tramos y tarda mucho en los dos últimos porque está cansado, su segunda componente será baja. Esta componente podría denominarse ‘experiencia en carreras’.

Biplot

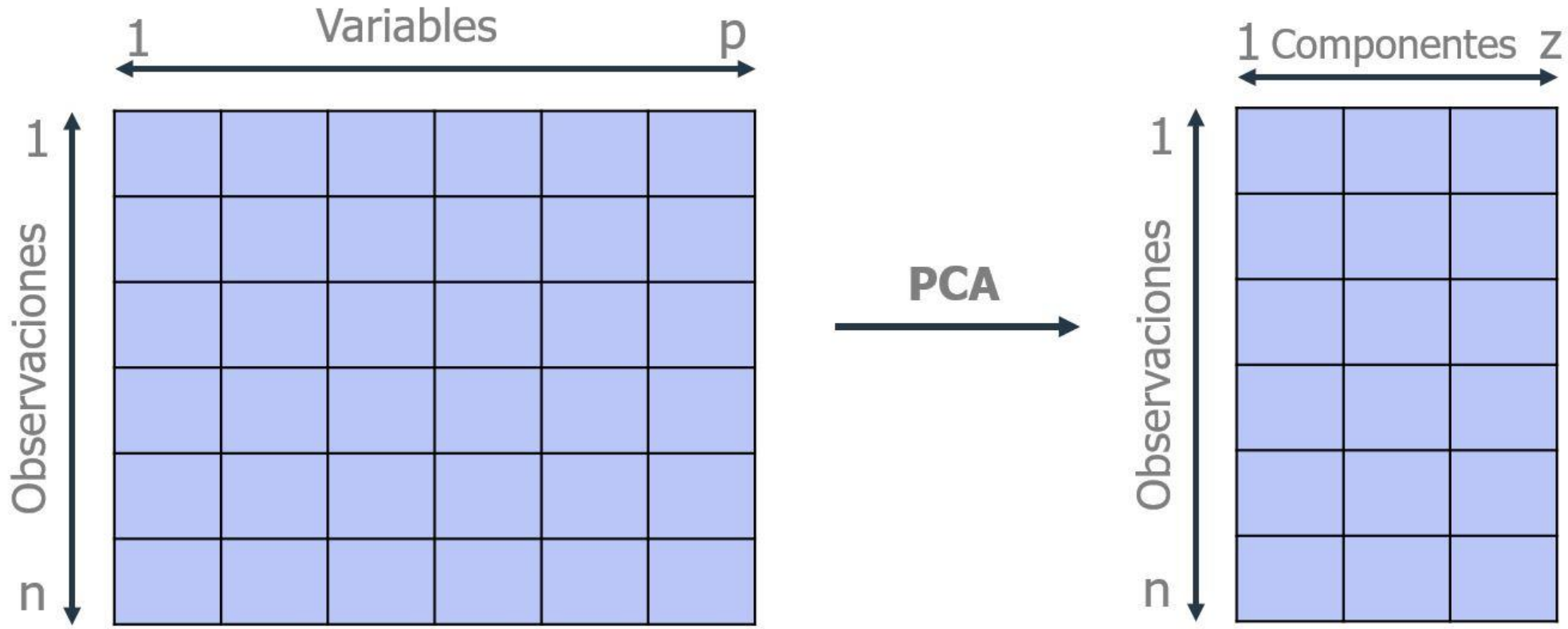
- Se representan simultáneamente las variables y los valores de cada individuo en pares de componentes principales. Biplot.
- Si son pocas observaciones tiene sentido graficar.



Biplot

- Biplot facilita:
- La interpretación de las distancias entre individuos
- La búsqueda de grupos o patrones.
- La explicación de las componentes principales utilizando las correlaciones con las variables originales.
- El estudio de las posiciones relativas de los individuos entre sí y respecto de las componentes principales graficadas.

Componentes principales



Componentes principales clásico

- El análisis de componentes principales (ACP) tiene por objetivo representar apropiadamente la información provista por un grupo de ***n*** observaciones donde se consideran ***p*** variables, reduciendo el número de estas pero resignando una baja cantidad de información. Esta representación se realiza a través de la creación de ***z*** nuevas variables no observables que resultan ser combinaciones lineales de las p variables originales ($z < p$).

Componentes principales robustas

- Los outliers univariados o multivariados pueden distorsionar la información de la matriz de covarianza muestral y conducir a resultados erróneos.
- Las técnicas robustas alternativas
- Métodos de Bootstrap requieren menos supuestos pero tienen alto costo computacional
- Otras alternativas: reemplazo del vector de medias de la matriz de covarianzas obtenida con el método clásico por el vector de medias y matriz de covarianza obtenido con el método robusto.

Alternativas robustas

- MDC Minium covarianza Determinant
- Estimador de Stahel-Donoho
- MVE Minimum volumen ellpsoid Elipsoide de mínimo volumen (Elipsoide de un estimador “recortado” aplicado sobre las distancias de Mahalanobis)

Estimador de Stahel-Donoho

- Utiliza una ponderación de las observaciones en función de su medida de alejamiento del conjunto general de datos
- La ponderación está basada en proyecciones univariadas sobre la dirección en la cual el alejamiento es máximo
- Para el caso multivariado este estimador tiene dificultades

MDC

- Otra posibilidad de obtener un estimador robusto es computar un estimador truncado.
- Se consideran solo las primeras di más pequeñas.
- La estimaciones obtenidas en este caso reciben el nombre de Mínimo Determinante de la Covariancia (MCD).
- El estimador MCD está definido por un subconjunto de h observaciones $\{x_{i1}, \dots x_{ih}\}$cuya matriz de covariancias, tenga el menor determinante a través de todos los posibles subconjuntos de tamaño h.
- Existe un algoritmo muy rápido debido a Rousseeuw y Van Driessen.
- Se ha demostrado que el proceso de iteración propuesta, converge en un número finito de pasos a un mínimo (local).
- El estimador MCD de localización y escala resultan ser la media aritmética y un múltiplo de la matriz de covariancias de la muestra de ese subconjunto h de observaciones.
- Una recomendable elección para h es [(n+p+1)/2] porque logra el punto de ruptura máximo.

MVE Elipsoide de volumen mínimo

- Esta estimación busca el elipsoide de volumen mínimo que contiene al menos la mitad de los puntos del conjunto de datos.
- Su nombre se debe a que trata de hallar entre todas las elipsoides que contengan al menos la mitad de los puntos, aquella que tenga menor volumen, es decir, un determinante menor de la matriz de covariancias.
- La tasa de consistencia del MVE es baja y la estimación de localización se da como el centro de este elipsoide y la estimación de covariancia es proporcionada por su forma.
- El MVE fue el estimador más popular por su alto punto de ruptura, pero más tarde fue sustituido por el MCD, principalmente debido a la disponibilidad de un algoritmo eficiente para su cálculo.

¿Preguntas?