Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

Reducción de la Dimensión

DIMENSIÓN

La definición de dimensión es muy precisa en términos matemáticos.

DIMENSIÓN

La definición de dimensión es muy precisa en términos matemáticos.

Definición

Un conjunto de vectores $\{v_1, \ldots, v_q\}$ en una espacio vectorial se dicen linealmente independientes (l.i.) si cumplen la siguiente propiedad,

$$a_1 v_1 + \cdots + a_q v_q = 0 \;\Rightarrow\; a_1 = \cdots = a_q = 0.$$

DIMENSIÓN

La definición de dimensión es muy precisa en términos matemáticos.

Definición

Un conjunto de vectores $\{v_1, \ldots, v_q\}$ en una espacio vectorial se dicen linealmente independientes (l.i.) si cumplen la siguiente propiedad,

$$a_1 v_1 + \cdots + a_q v_q = 0 \;\Rightarrow\; a_1 = \cdots = a_q = 0.$$

Por ejemplo, $(1, 2)$ y $(1, 3)$ vectores de $\mathbb{R}^2$ linealmente independientes.

DIMENSIÓN

Defincion

Decimos que $\{v_1, \ldots, v_q\}$ generan el subespacio S si para todo $x \in S$, se tiene que existen $a_1, \ldots, a_q$ números reales no todos nulos tales que

$$x = a_1 v_1 + \ldots a_q v_q.$$

DIMENSIÓN

Defincion

Decimos que $\{v_1, \ldots, v_q\}$ generan el subespacio S si para todo $x \in S$, se tiene que existen $a_1, \ldots, a_q$ números reales no todos nulos tales que

$$x = a_1 v_1 + \ldots a_q v_q.$$

Estamos diciendo que cualquier elemento del subespacio se puede escribir como una combinación lineal de $v_1, \ldots, v_q$.

DIMENSIÓN

Defincion

Sea $\mathbb{V}$ un espacio vectorial, decimos que el subconjunto $\{v_1, \ldots, v_p\}$ es una base de $\mathbb{V}$ si

- $v_1, \ldots, v_p$ generan $\mathbb{V}$.
- $v_1, \ldots, v_p$ son linealmente independientes.

DIMENSIÓN

Defincición

Sea $\mathbb{V}$ un espacio vectorial, decimos que el subconjunto $\{v_1, \ldots, v_p\}$ es una base de $\mathbb{V}$ si

- $v_1, \ldots, v_p$ generan $\mathbb{V}$.
- $v_1, \ldots, v_p$ son linealmente independientes.

La dimensión del espacio se define como la cantidad de vectores en una base. Es decir, si $\{v_1, \ldots, v_p\}$ es una base de $\mathbb{V}$, entonces $dim(\mathbb{V}) = p$.

DIMENSIÓN

Defincición

Sea $\mathbb{V}$ un espacio vectorial, decimos que el subconjunto $\{v_1, \ldots, v_p\}$ es una base de $\mathbb{V}$ si

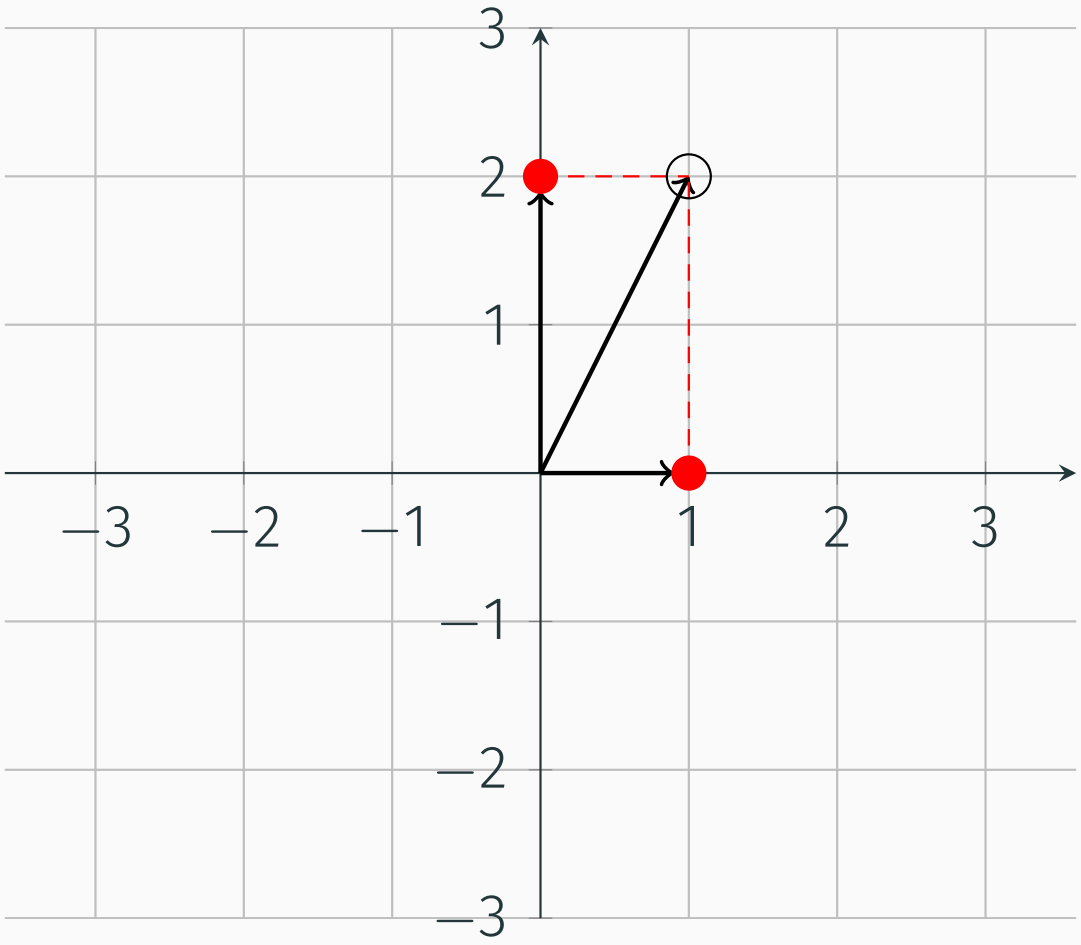
- $v_1, \ldots, v_p$ generan $\mathbb{V}$.
- $v_1, \ldots, v_p$ son linealmente independientes.

La dimensión del espacio se define como la cantidad de vectores en una base. Es decir, si $\{v_1, \ldots, v_p\}$ es una base de $\mathbb{V}$, entonces $dim(\mathbb{V}) = p$.

Es rápido de ver que $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ es una base de $\mathbb{R}^3$.

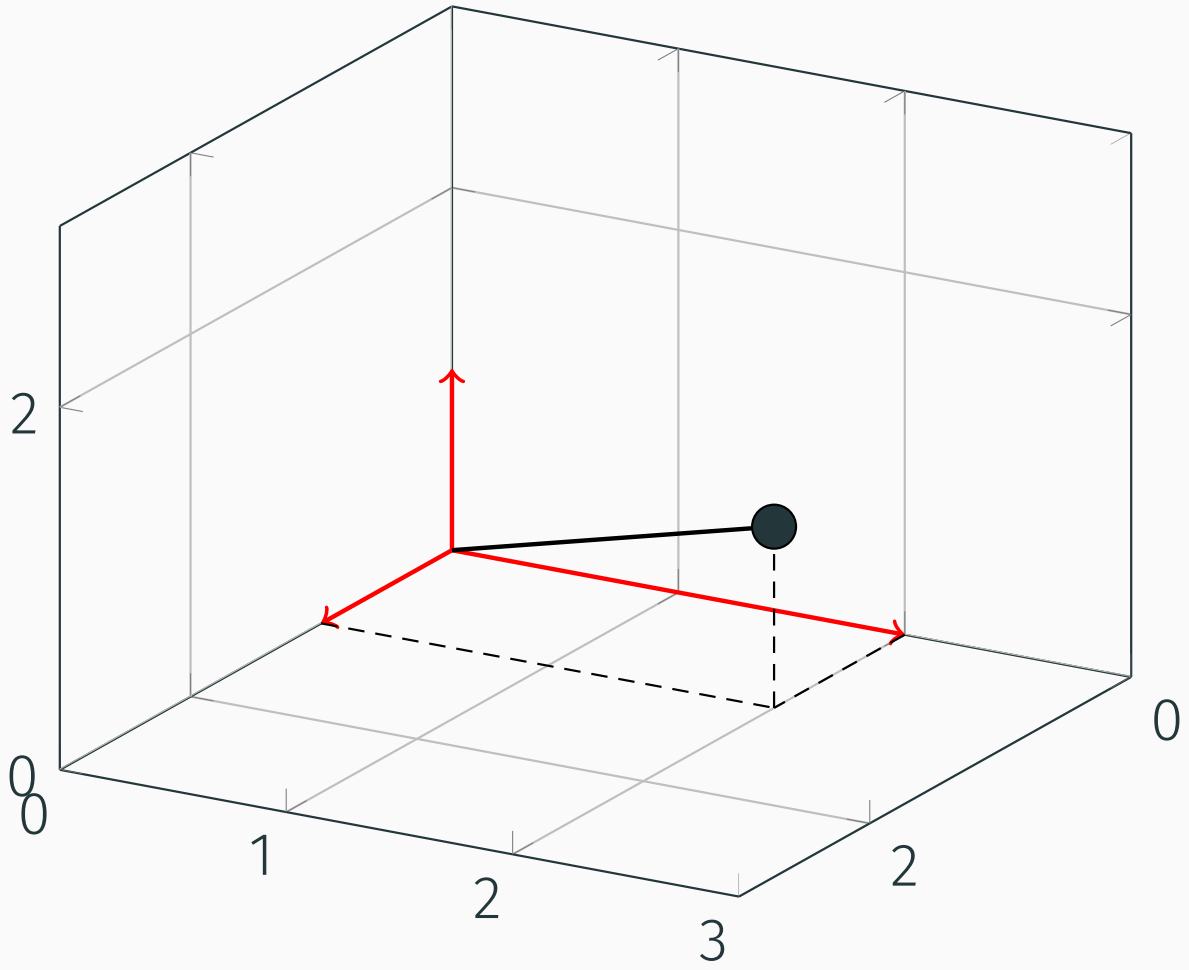
DIMENSIÓN

En criollo ...



DIMENSIÓN

En criollo ...



DIMENSIÓN

¿De qué hablamos cuando hablamos de dimensión en los datos?

DIMENSIÓN

¿De qué hablamos cuando hablamos de dimensión en los datos?

En general nos referimos a la cantidad de variables o “features”de nuestro dataset.

DIMENSIÓN

¿De qué hablamos cuando hablamos de dimensión en los datos?

En general nos referimos a la cantidad de variables o “features”de nuestro dataset.

Brutalmente dicho, si registramos los datos en forma de tabla donde las observaciones están representadas en las filas, nos referimos a la dimensión como la cantidad de columnas.

DIMENSIÓN

¿De qué hablamos cuando hablamos de dimensión en los datos?

En general nos referimos a la cantidad de variables o “features”de nuestro dataset.

Brutalmente dicho, si registramos los datos en forma de tabla donde las observaciones están representadas en las filas, nos referimos a la dimensión como la cantidad de columnas.

Graficamente representamos cada fila como un vector de dimensión (cant. de columnas) cuando queremos dibujar.

DIMENSIÓN

¿Qué es un técnica de reducción de dimensión?

DIMENSIÓN

¿Qué es un técnica de reducción de dimensión?

Es encontrar una representación de los datos en dimensión menor que la cantidad de variables o “features”que estemos analizando.

DIMENSIÓN

¿Qué es un técnica de reducción de dimensión?

Es encontrar una representación de los datos en dimensión menor que la cantidad de variables o “features”que estemos analizando.

Geometricamente, es aplicar una transformación a los datos a un espacio de dimensión (bien definida) menor.

DIMENSIÓN

¿Qué es un técnica de reducción de dimensión?

Es encontrar una representación de los datos en dimensión menor que la cantidad de variables o “features”que estemos analizando.

Geometricamente, es aplicar una transformación a los datos a un espacio de dimensión (bien definida) menor.

¿Por qué? ...

VENTAJAS

- Reduce el espacio de almacenamiento requerido.
- Disminuye los tiempos de procesamiento.
- Elimina características que no son relevantes o no aportan verdadera información al análisis.
- En dimensiones bajas (2D o 3D) se puede hacer una visualización de los datos

VENTAJAS

- Reduce el espacio de almacenamiento requerido.
- Disminuye los tiempos de procesamiento.
- Elimina características que no son relevantes o no aportan verdadera información al análisis.
- En dimensiones bajas (2D o 3D) se puede hacer una visualización de los datos

Además ...

Siempre nos acecha la MALDICIÓN DE LA DIMENSIÓN!

PCA

Componentes Principales es una técnica de reducción de dimensión. Supongamos que nuestro conjunto de datos D vive en $\mathbb{R}^p$.

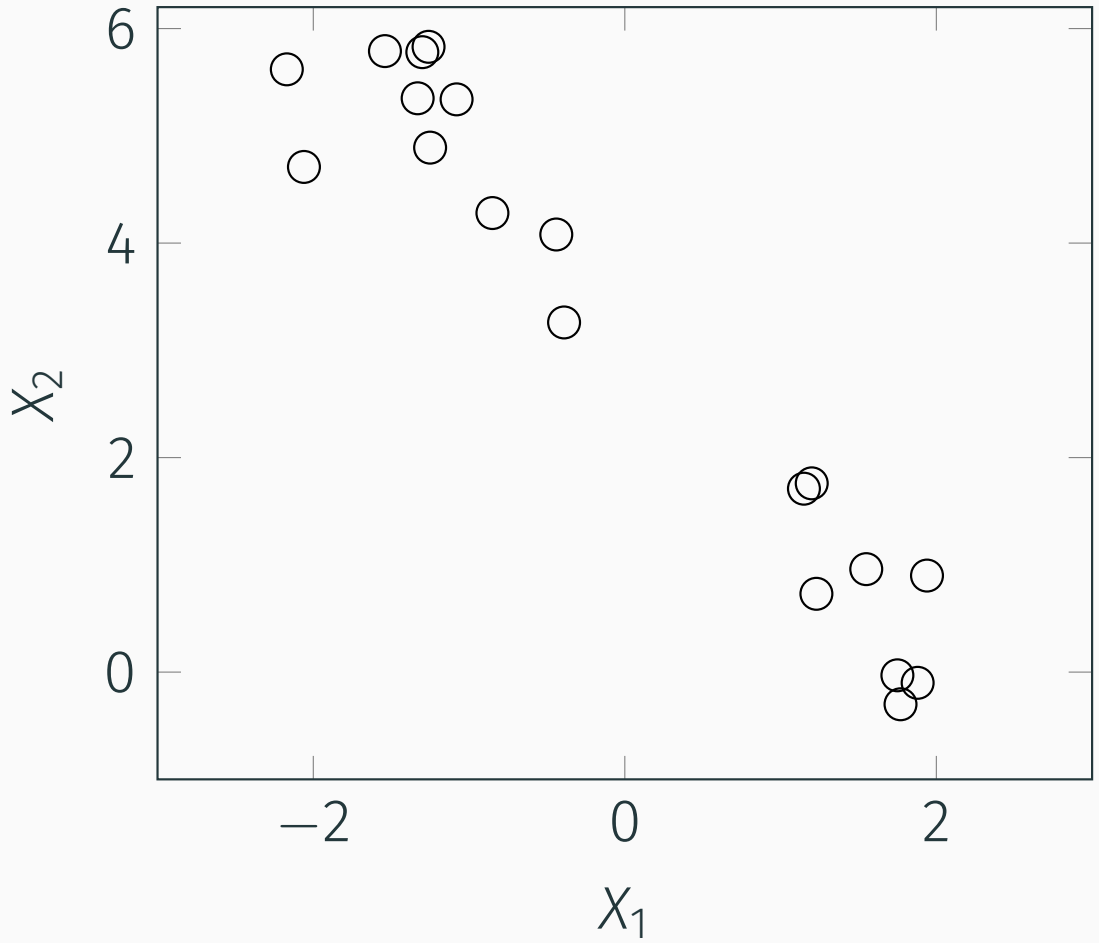
PCA

Componentes Principales es una técnica de reducción de dimensión. Supongamos que nuestro conjunto de datos D vive en $\mathbb{R}^p$.

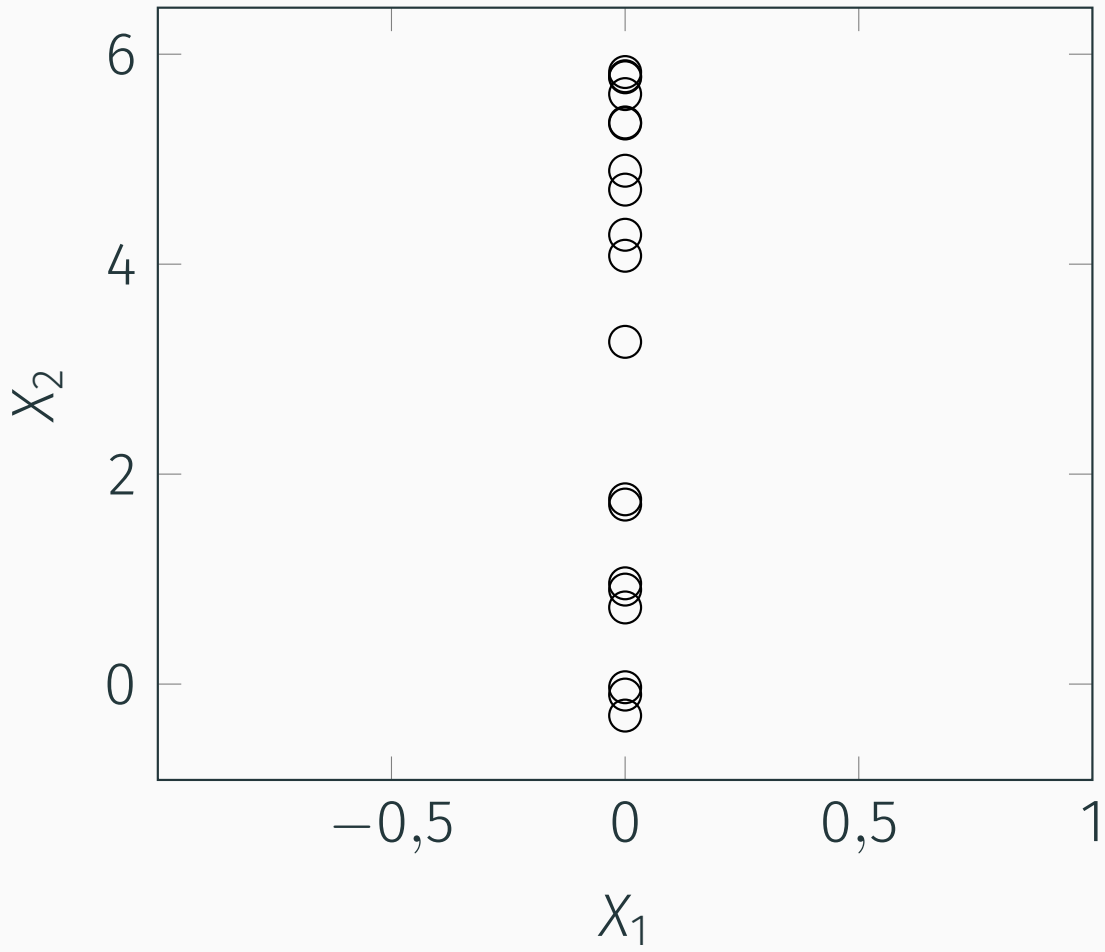
La técnica de componentes principales se basa en contstruir una transformación lineal (función) que lleve nuestros datos a un $\mathbb{R}^q$ con $q << p$.

$$D \subset \mathbb{R}^p \xrightarrow{\text{PCA}} \mathbb{R}^q.$$

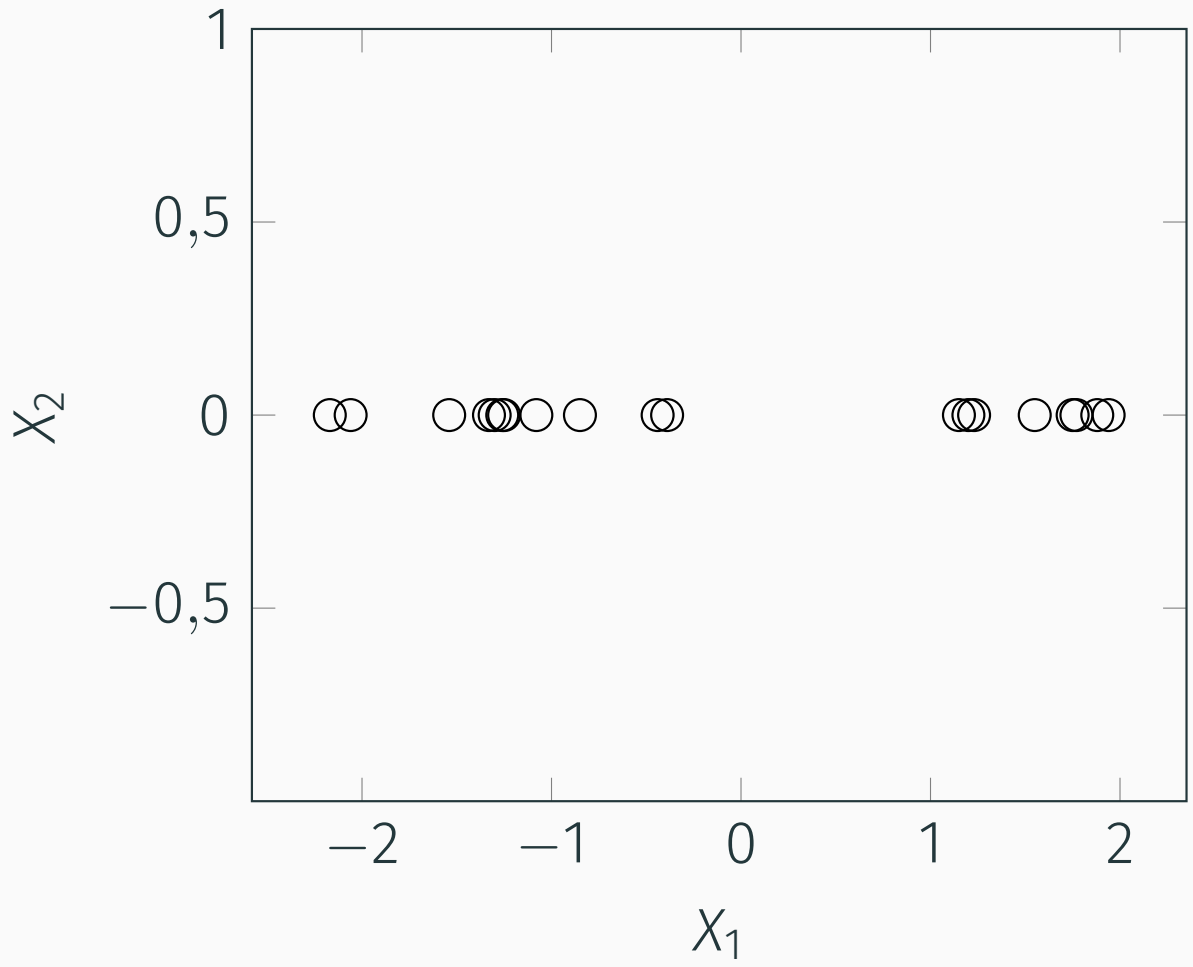
MOTIVACIÓN



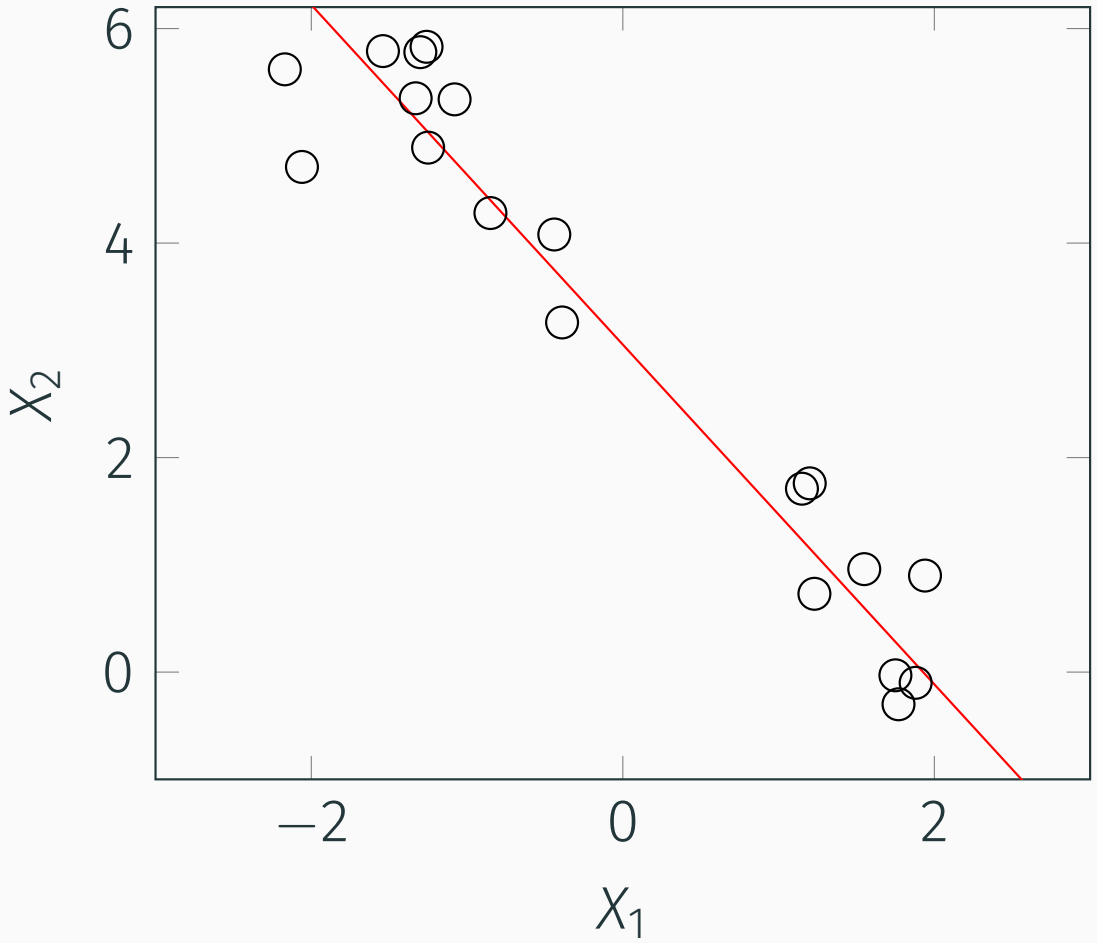
MOTIVACIÓN



MOTIVACIÓN



MOTIVACIÓN



Transformaciones Lineales

TRANSFORMACIÓN LINEAL

Definición

Una transformación lineal es una función $T : \mathbb{R}^n \to \mathbb{R}^m$ que cumple dos propiedades:

- $\forall\, v, w \in \mathbb{R}^n, \;\; T(v + w) = T(v) + T(w)$.
- $\forall \lambda \in \mathbb{R},\; v \in \mathbb{R}^n, \;\; T(\lambda v) = \lambda T(v)$.

Mini Ejercicio: mostrar que para toda transformación lineal $T(0) = 0$.

Coloquialmente: una transformación lineal es una función que manda vectores en vectores.

EJEMPLO

Definimos $T : \mathbb{R}^2 \to \mathbb{R}^3$ como $T(x_1, x_2) = (x_1, x_2, x_1 + x_2)$

Si tomamos el vector $v = (2, 3)$ en $\mathbb{R}^2$ y le aplicamos T nos lleva al vector $(2, 3, 5)$ en $\mathbb{R}^3$.

Es decir, $T(2, 3) = (2, 3, 2 + 3)$.

Observemos que podemos representar a T con una matriz.

$$T(x_1, x_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_1 + x_2 \end{pmatrix}$$

MATRIZ ASOCIADA

Las transformaciones lineales que se definen en espacios de dimensión finita tienen la propiedad de tener una matriz asociada como vimos en el ejemplo.

MATRIZ ASOCIADA

Las transformaciones lineales que se definen en espacios de dimensión finita tienen la propiedad de tener una matriz asociada como vimos en el ejemplo.

Propiedad

Es decir, si $T: \mathbb{R}^n \to \mathbb{R}^m$ es una transformación lineal, entonces existe una matriz $A \in \mathbb{R}^{m \times n}$ tal que $T(v) = Av$ para todo $v \in \mathbb{R}^n$.

$$T(x_1, \ldots, x_n) = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \ddots & \ldots & a_{2n} \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Autovalores

Ahora nos centraremos en las transformaciones lineales de $\mathbb{R}^n$ en sí mismo.

Definición

Sea $T : \mathbb{R}^n \to \mathbb{R}^n$ una transformación lineal decimos que $v \neq 0$ es un autovector de T con autovalor $\lambda_v \in \mathbb{R} - \{0\}$ si $T(v) = \lambda_v v$.

Autovalores

Ahora nos centraremos en las transformaciones lineales de $\mathbb{R}^n$ en sí mismo.

Definición

Sea $T : \mathbb{R}^n \to \mathbb{R}^n$ una transformación lineal decimos que $v \neq 0$ es un autovector de T con autovalor $\lambda_v \in \mathbb{R} - \{0\}$ si $T(v) = \lambda_v v$.

- $A \in \mathbb{R}^{n \times n}$ es simétrica si $A' = A$.
- $A \in \mathbb{R}^{n \times n}$ es definida positiva si $z'Az > 0$ para todo $z \neq 0$.
- Si A es simétrica y positiva, entonces tengo una base ortonormal de autovectores.

Componentes Principales

DEFINICION DURA

Sea $X \in \mathbb{R}^p$ un vector aleatorio, tal que $E(X) = \mu$ y $Var(X) = \Sigma$. Sean,

- $\lambda_1 \geq \cdots \geq \lambda_p$ los autovalores de Σ.
- $\gamma_1, \ldots, \gamma_p$ los autovectores de Σ asociados a los correspondientes autovalores.
- $\Gamma = (\gamma_1, \ldots, \gamma_p)$ la matriz que tiene como columnas los autovectores.
- $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$

Observar que

$$\Gamma'\Sigma\Gamma = \Lambda, \quad \text{y} \quad \Gamma\Gamma' = Id_p.$$

DEFINICION DURA

Notar que como $\{\gamma_1, \ldots, \gamma_p\}$ es una base de $\mathbb{R}^p$ podemos escribir a X como

$$X = \mu + \sum_{j=1}^{p} \gamma_j'(X - \mu)\gamma_j = \mu + \sum_{j=1}^{p} \langle X - \mu, \gamma_j \rangle \gamma_j.$$

Sea el vector $v = \Gamma'(X - \mu)$, las coordenadas de $v = (v_1, \ldots, v_p)$ se llaman las componentes principales de X.

DEFINICION DURA

Notar que como $\{\gamma_1, \ldots, \gamma_p\}$ es una base de $\mathbb{R}^p$ podemos escribir a X como

$$X = \mu + \sum_{j=1}^{p} \gamma_j'(X - \mu)\gamma_j = \mu + \sum_{j=1}^{p} \langle X - \mu, \gamma_j \rangle \gamma_j.$$

Sea el vector $v = \Gamma'(X - \mu)$, las coordenadas de $v = (v_1, \ldots, v_p)$ se llaman las componentes principales de X.

Observación: la $j-$ésima componente principal $v_j = \gamma_j'(X - \mu) = \langle \gamma_j, X - \mu \rangle$, se corresponde con la proyección ortogonal de $(X - \mu)$ sobre la dirección de γ_j.

DIBUJITO BLANDO

HACER UN DIBUJITO PARA QUE NO SE MUERAN!

Propiedad: información

Las componentes principales $v_1, \ldots, v_p$ son no correlacionadas y $Var(v_j) = \lambda_j$. O sea,

$$Var(v) = \Lambda = diag(\lambda_1, \ldots, \lambda_p).$$

La propiedad nos dice que cada componente aporta información que las componentes anteriores no aportaron.

En particular si X tiene distribución normal multivariada cada componente es independiente de las demás.

Propiedad: optimalidad

Sea H_0 el subespacio generado por $\gamma_1, \ldots, \gamma_q$ y sea H otro subespacio de dimensión q. Llamemos $\pi(X, H)$ a la proyección ortogonal de X sobre el subespacio H. Entonces,

$$E\left[||x - \pi(x, H_0)||^2\right] \leq E\left[||x - \pi(x, H)||^2\right].$$

Esta propiedad nos dice que las componentes principales nos dan el mejor ajuste lineal sobre un subespacio de dimensión menor.

PROPIEDADES

$$\underset{||a||=1}{\text{máx}}\ Var(a'X) = Var(v_1), \tag{1}$$

es decir el máximo se alcanza en γ_1

PROPIEDADES

$$\max_{||a||=1} Var(a'X) = Var(v_1), \tag{1}$$

es decir el máximo se alcanza en γ_1

$$\max_{\substack{||a||=1 \\ Cov(a'x, v_j)=0 \ 1\leq j\leq k-1}} Var(a'X) = Var(v_k), \tag{2}$$

alcaza el máximo en γ_k y la condición sobre la covarianza me asegura que no agrego información repetida.

PROPIEDADES

$$\underset{||a||=1}{\text{máx}}\ Var(a'X) = Var(v_1), \tag{1}$$

es decir el máximo se alcanza en γ_1

$$\underset{\substack{||a||=1 \\ Cov(a'x,v_j)=0\ 1\leq j\leq k-1}}{\text{máx}} \qquad Var(a'X) = Var(v_k), \tag{2}$$

alcaza el máximo en γ_k y la condición sobre la covarianza me asegura que no agrego información repetida.

$$\sum_{j=1}^{p} Var(v_j) = \sum_{j=1} \lambda_j = traza(\Sigma). \tag{3}$$

EN PALABRAS

Tenemos nuestro vector aleatorio X y queremos escribirlo en un subespacio de dimensión menor q.

EN PALABRAS

Tenemos nuestro vector aleatorio X y queremos escribirlo en un subespacio de dimensión menor q.

Considero como primer vector unitario para la base el que genere la combinación lineal con mayor varianza. Ese vector, por (1), coincide con el autovector de Σ con autovalor más grande (γ_1).

EN PALABRAS

Tenemos nuestro vector aleatorio X y queremos escribirlo en un subespacio de dimensión menor q.

Considero como primer vector unitario para la base el que genere la combinación lineal con mayor varianza. Ese vector, por (1), coincide con el autovector de Σ con autovalor más grande (γ_1).

Ahora necesito otro vector para la base, que genere la combinación lineal con mayor varianza, pero no repita información, es decir, $Cov(a'X, \gamma_1) = 0$. Por (2), ese vector coincide con γ_2.

Siguiendo voy incorporando cada componente.

EN PALABRAS

Otra observación importante que tenemos que hacer es que de esta forma conseguimos la transformación lineal que planteamos al comienzo. Esta transformación se corresponde con la proyección ortogonal sobre el subespacio generado por los autovectores de Σ.

EN PALABRAS

Otra observación importante que tenemos que hacer es que de esta forma conseguimos la transformación lineal que planteamos al comienzo. Esta transformación se corresponde con la proyección ortogonal sobre el subespacio generado por los autovectores de Σ.

¿Cómo escribo a X?

$$\mu + \sum_{j=1}^{q} \gamma_j'(x - \mu)\gamma_j.$$

¿CUÁNTAS COMPONENTES?

Vimos que si tomamos todas las componentes el total de variabilidad que aporta cada combinación lineal es la traza de Σ (3), es decir, la suma de sus autovalores.

Una buena forma de tomar un criterio para dejar de agregar componentes es ver qué proporción me aporta cada una de ellas con respecto al total.

$$prop = \frac{\sum_{j=1}^{q} \lambda_j}{traza(\Sigma)}.$$

INFERENCIA

En la práctica μ y Σ son desconocidos y tenemos que estimarlos a partir de una muestra aleatoria $X_1, \ldots, X_n$.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad Q = \sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})', \quad \hat{\Sigma} = \frac{Q}{n}$$

Luego calculamos los autovectores de $\hat{\Sigma}$ y procedemos igual que con la versión poblacional.

INFERENCIA

En la práctica μ y Σ son desconocidos y tenemos que estimarlos a partir de una muestra aleatoria $X_1, \ldots, X_n$.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad Q = \sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})', \quad \hat{\Sigma} = \frac{Q}{n}$$

Luego calculamos los autovectores de $\hat{\Sigma}$ y procedemos igual que con la versión poblacional.

Si no ponga princomp en R y problema resuelto ;)

EJEMPLO

Apliquemos lo que estuvimos viendo a un dataset. Usaremos el conjunto de datos decathlon2 en la libreria *factorextra* de R.

Tenemos variables que se corresponden al desempeño de distintos atletas en dos competencias deportivas. Nosotros tomaremos solamente a modo ilustrativo.

El dataset contiene 10 variables que se corresponden a las pruebas de declatón y 27 observaciones que se corresponden con cada participante. Cada celda se corresponde con el puntaje obtenido en la prueba (tiempo, distancia, etc).

EJEMPLO

	X100m	Long.jump	Shot.put	High.jump
SEBRLE	11.04	7.58	14.83	2.07
CLAY	10.76	7.40	14.26	1.86
BERNARD	11.02	7.23	14.25	1.92
YURKOV	11.34	7.09	15.19	2.10
ZSIVOCZKY	11.13	7.30	13.48	2.01
McMULLEN	10.83	7.31	13.76	2.13

Cuadro 1: cuatro columnas del dataset

EJEMPLO

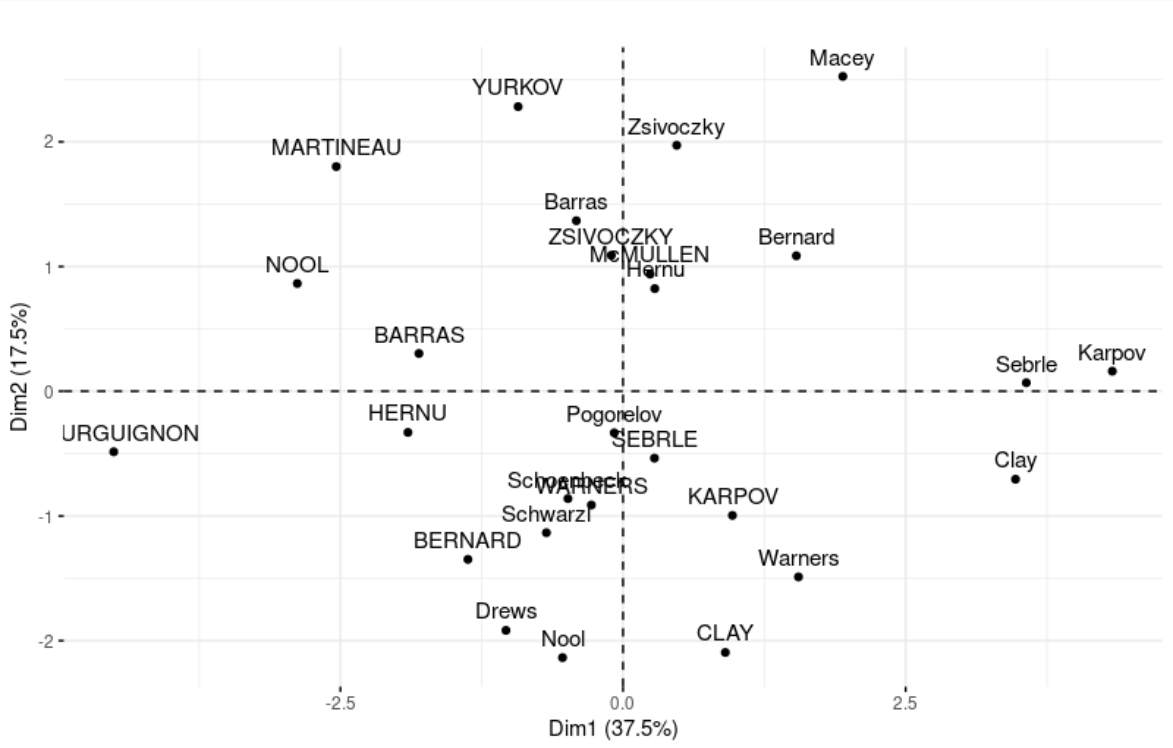


Figura 1: Dos componentes

EJEMPLO

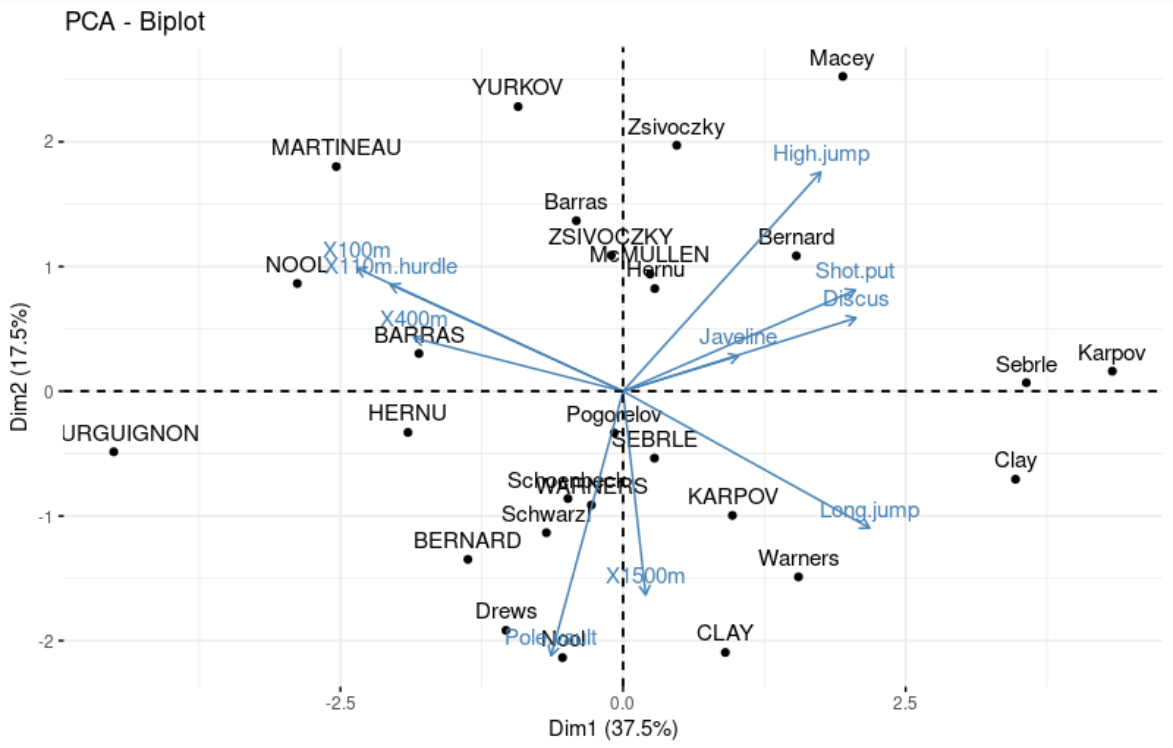


Figura 2: Biplot

EJEMPLO

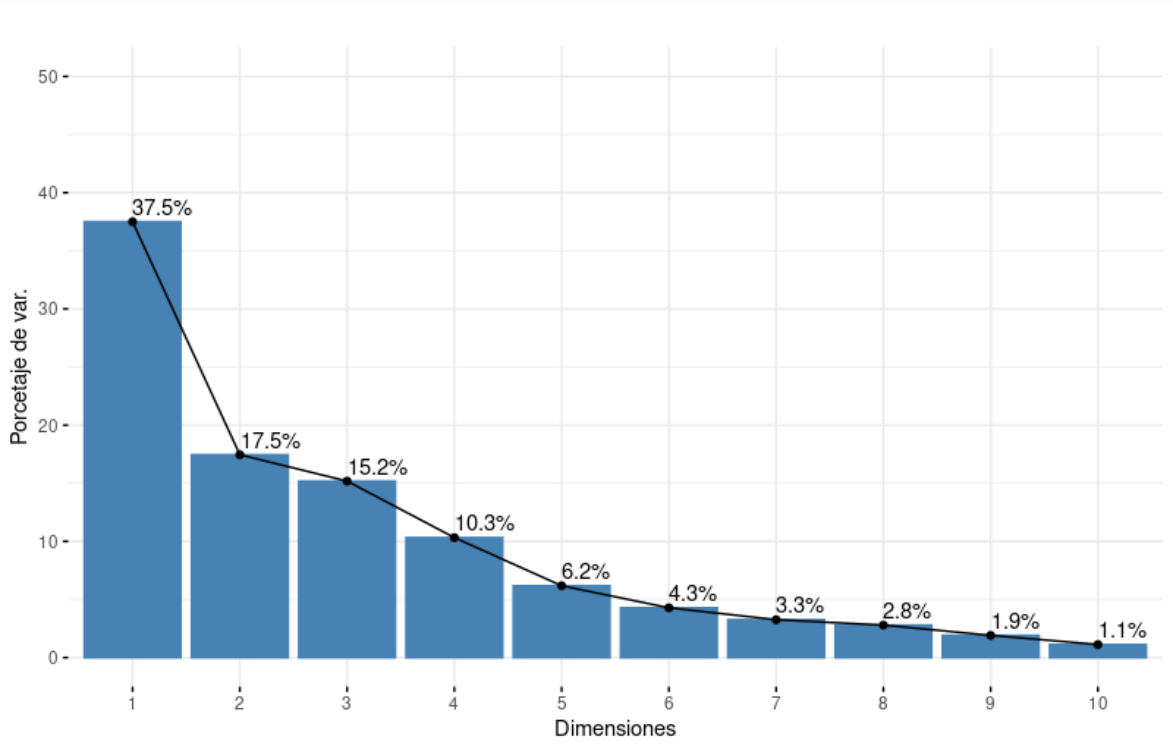


Figura 3: Variabilidad aportada por cada componente

EJEMPLO

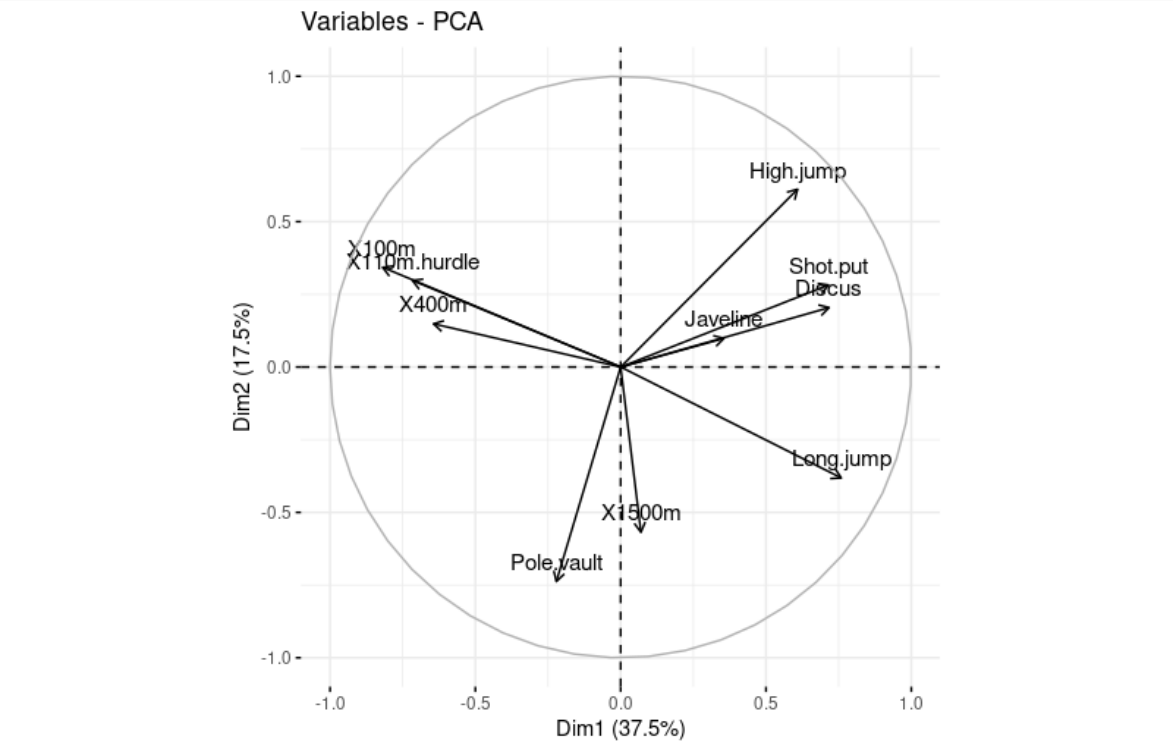


Figura 4: Var plot

EJEMPLO

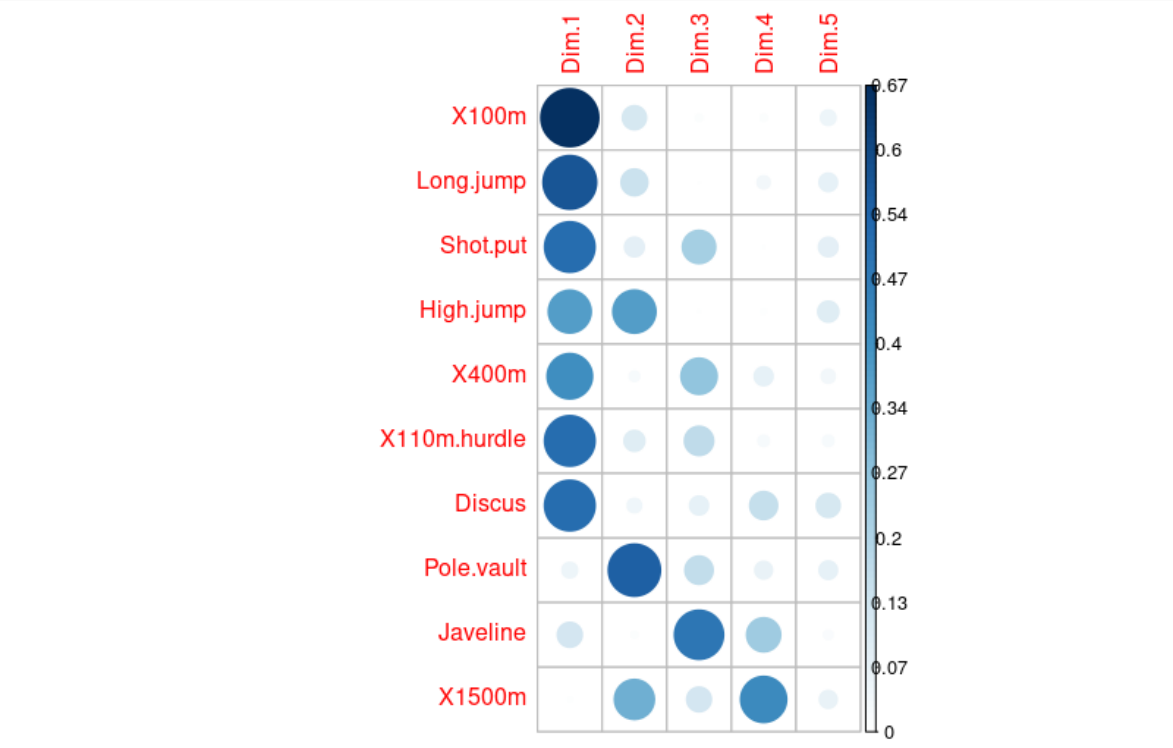


Figura 5: Correlación entre variables y componentes

EJEMPLO

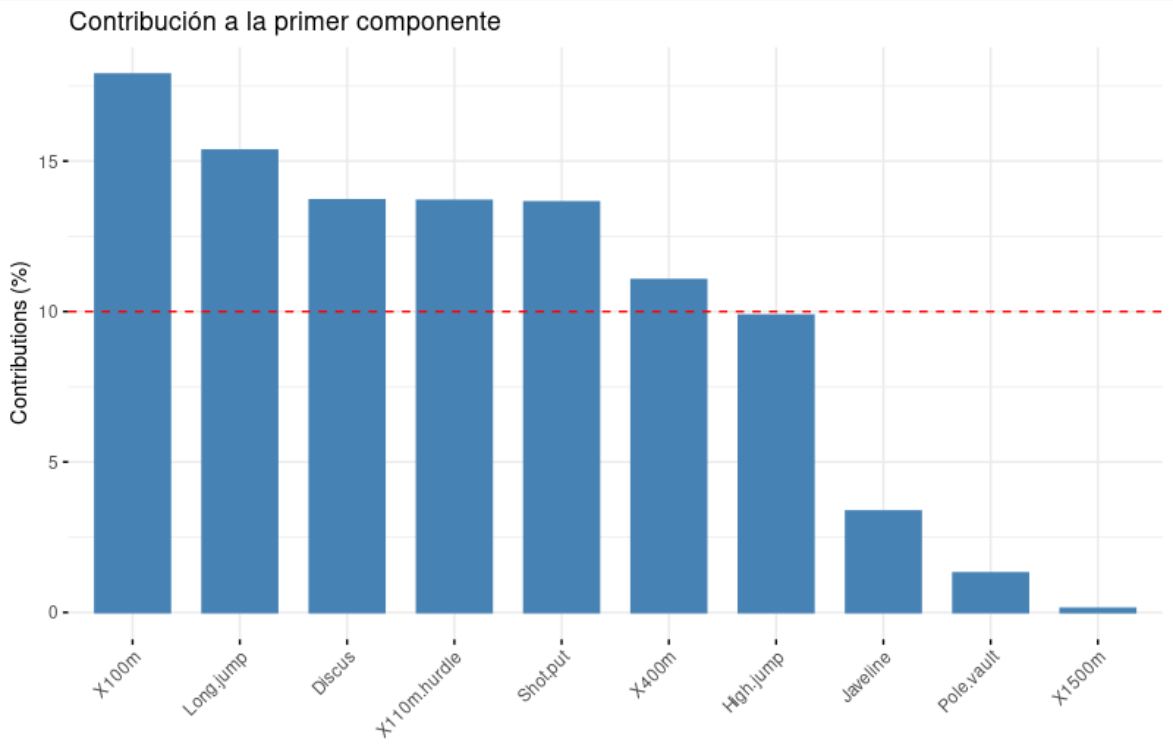


Figura 6: Contribución de las variables a la primer componente

EJEMPLO

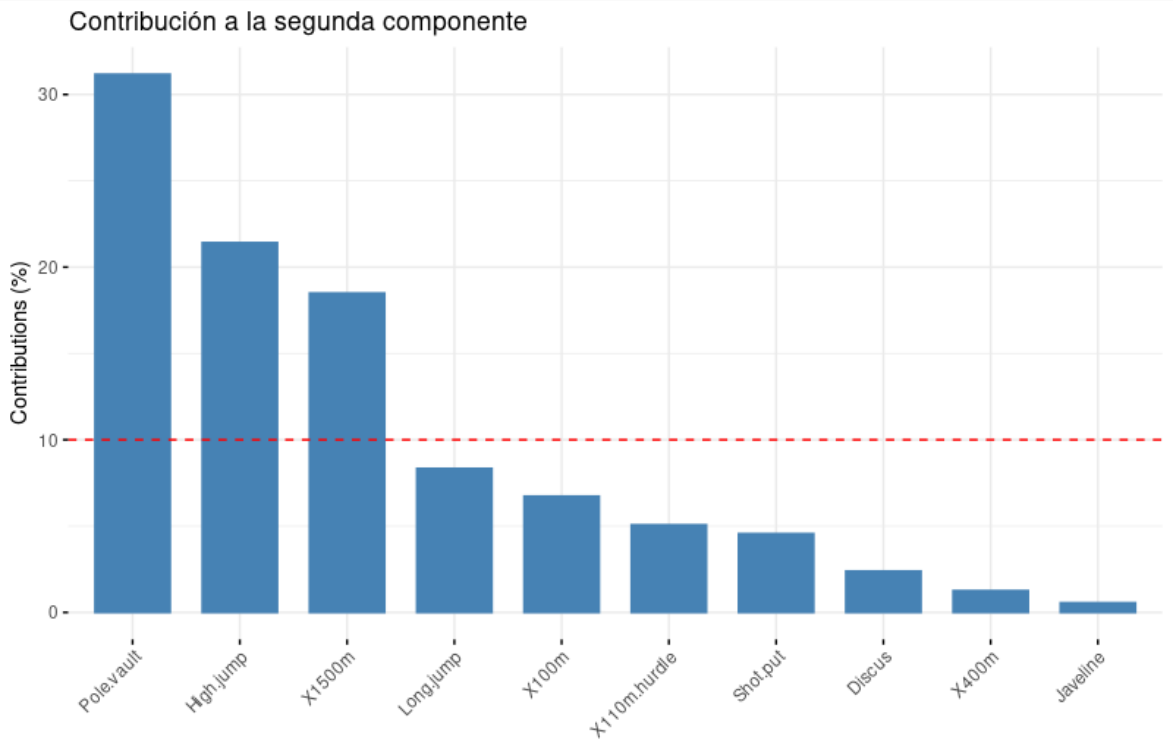


Figura 7: Contribución de las variables a la segunda componente

PARA SEGUIR PROFUNDIZANDO

Sparse PCA.

Robust PCA.

Functional PCA.

Principal Directions (PCA in Manifolds).

COFFEE BREAK!