https://datascience-project.eu/

Plantilla de ficha de formación

Título	Análisis de Correspondencias (AC)
Palabras clave (metaetiqueta)	AC, variables cualitativas, inercia explicada, valores propios.
Idioma	Español
Objetivos / Metas / Aprendizaje resultados	**Este módulo tiene como objetivo presentar y explicar Técnica de Análisis de Correspondencias** **Al final de este módulo serás capaz de:** **- Conocer la lógica del AC;** **- Conoce los requisitos** **- realizar un AC** **- realizar un AC en R con el paquete FactorMineR**

Curso de formación:	
Alfabetización en ciencia de datos	
de visualización de datos y análisis visual	X
Introducción a la ciencia de datos para las ciencias humanas y sociales	
Ciencia de datos para siempre	
Periodismo de datos y storytelling	

Descripción	En este módulo de formación se le presentará la técnica de análisis multidimensional denominada Análisis de Correspondencia, AC. El análisis de correspondencias es una forma de escalado multidimensional, que esencialmente construye una especie de modelo espacial que muestra las asociaciones entre un conjunto de variables categóricas. Si el conjunto incluye solo dos variables, el método generalmente se denomina Análisis de Correspondencias Simple (ACS). Si el análisis involucra más de dos variables, generalmente se denomina Análisis de Correspondencias Múltiple (ACM). En este módulo nos ocuparemos del análisis de correspondencias simple. El objetivo de este análisis es reducir la dimensionalidad del fenómeno investigado

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	preservando la información contenida en el mismo. La técnica es aplicable a fenómenos medidos con variables cualitativas. La última parte del módulo estará dedicada a la aplicación de AC con el software R.
Contenido dispuesto en 3 niveles	**1. INTRODUCCIÓN** El análisis de correspondencias, AC, es una técnica de análisis multidimensional que es capaz de traducir casi cualquier tipo de tabla que consiste en datos numéricos en forma gráfica. El objeto de las CA son las matrices de contingencia, cuyos elementos indican el número de veces que se han detectado juntas las características de dos magnitudes diferentes. El objetivo principal del AC es analizar las relaciones entre dos variables y cualitativas observadas en un conjunto de unidades estadísticas. Esto se hace a través de la identificación de un espacio "óptimo", es decir, de una dimensión reducida que representa la síntesis de la información estructural contenida en los datos originales. El propósito del análisis es sacar a la luz el entrecruzamiento de vínculos o correspondencias que existen entre los datos bajo examen. **2. REQUISITOS PARA EL ANÁLISIS DE COINCIDENCIA** Para realizar el análisis de correspondencias es importante analizar las variables a utilizar para tener claras algunas de sus características. En concreto, las variables deben cumplir los siguientes requisitos: - *Las **variables** deben ser **Cualitativas** :* Las variables cualitativas son variables que no están representadas por números, sino por modalidades, por ejemplo: género, nivel de educación, estado civil, etc. Estas modalidades, también llamadas **categorías, deben ser exhaustivas y mutuamente excluyentes . Mutuamente excluyente significa que las modalidades variables no deben contener el mismo tipo de información. Por ejemplo, para la variable "color de cabello" no se pueden ingresar los modos "cabello oscuro" y "cabello castaño", ya que cabello oscuro también significa cabello castaño y viceversa. Estaustivo significa que las modalidades de una variable deben tener en cuenta todas las posibilidades. Por ejemplo, para la variable "nivel de estudios" se insertan las modalidades "diploma", "licenciatura", "título**

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

de segundo nivel". Estas tres modalidades no tienen en cuenta todos los posibles niveles de educación.

- ***Las variables deben ser interdependientes*** **:**
  Antes de realizar el análisis de las correspondencias es necesario verificar el grado de interdependencia entre las **dos variables consideradas, ya que si fueran independientes no tendría sentido realizar el análisis de los partidos.**
  Para hacer esto, realice la prueba de chi-cuadrado :
  H_0**:las dos variables son independientes**
  H_1**:las dos variables no son independientes**
  Para interpretar los resultados de la prueba observamos el p-valor:
  p-valor < 0,05: se rechaza la hipótesis nula y en consecuencia las variables se consideran con cierto grado de dependencia.

3. Cómo conducir CA

Después de verificar los requisitos de CA, puede pasar al análisis real.

3.1) Tablas de contingencia

En el análisis de correspondencias trabajamos con tablas de contingencia, que contienen las frecuencias conjuntas de las modas de las dos variables cualitativas X e Y. Estas matrices están formadas siempre por enteros nunca negativos que son conteos, es decir, simples registros de lo ocurrido. Además, ambas variables categóricas juegan un papel simétrico en el que todos los elementos tienen la misma naturaleza.

$X \backslash Y$	y_1	y_2	y_3	
x_1				
x_2		$n_{i,j}$		$n_{i.}$
x_3				
		$n_{.j}$		n

X, Y son las variables cualitativas.

x_1, x_2 , x_3 : son las modas de la variable de X

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

y_1, y_2 , y_3 : son las modas de la variable de Y

$n_{i,j}$: son las frecuencias conjuntas absolutas, es decir, las frecuencias de los pares, por ejemplo$n_{1,1}$: $X = x_1; Y = y_1$

$n_{i\cdot}$: son las filas marginales:$n_{i.} = \sum_{j=1}^{C} n_{i,j}$

$n_{\cdot j}$: son los marginales de columna:$n_{.j} = \sum_{i=1}^{R} n_{i,j}$

Estos son la suma de la fila (o columna) fija de las frecuencias conjuntas en los modos de Y (para las columnas en los modos de X).

n = es el número de muestra, que se puede obtener sumando los marginales de fila o columna:$n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{i,j} \quad \forall\, i,j$

Puede cambiar de frecuencias absolutas a frecuencias relativas dividiendo cada frecuencia absoluta por n:$f_{i,j} = \frac{n_{i,j}}{n}$

3.2) Matriz de Perfil de Fila y Matriz de Perfil de Columna

La matriz de perfiles de fila se obtiene dividiendo las frecuencias absolutas (o frecuencias relativas) por los márgenes de fila respectivos. Por lo tanto:

$$\frac{n_{i,j}}{n_i} = \frac{f_{i,j}}{f_{i.}} \quad \forall\, i,j$$

La tabla de contingencia será:

$\frac{f_{i,j}}{f_{i.}} = \frac{n_{i,j}}{n_{i.}}$	1 1 1
profilo medio	1

En los márgenes de la fila tenemos todos 1 y esto representa la suma de los perfiles de fila.

En los márgenes de la columna se encuentran los perfiles medios que se obtienen sumando las frecuencias relativas por columna; o promediando los elementos de la matriz de perfil de fila, por columna.

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Este es un promedio ponderado, donde las masas están representadas por las filas marginales $f_{i.}$.

Trabajar con frecuencias pierde dimensión, por lo que el espacio fila se representa por un espacio C-1 dimensiones, es decir

Se puede construir una **diagonal matriz de marginales de fila $\boldsymbol{D_R}$**, que tiene perfiles de fila en la diagonal mayor. La matriz diagonal de filas marginales es una matriz **R· R,** que tiene dimensiones iguales a las filas y en la diagonal mayor contiene las filas marginales de la tabla de frecuencias relativas. Una matriz diagonal es una matriz cuyo elemento genérico en la diagonal mayor es el marginal de fila, encima o debajo de ella, son todos ceros. Siempre es una matriz simétrica y cuadrada. Con la matriz diagonal de márgenes de fila se puede construir el **arreglo de perfiles de fila** : se obtiene dividiendo las frecuencias relativas por los márgenes de fila $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$**.** Las dimensiones de **F** son R· C, mientras que $\boldsymbol{D_R}$tiene dimensión R· R, como no se puede hacer la división entre matrices, se calcula la inversa $\boldsymbol{D_R}$y se multiplica por **F** , resolviendo así el problema de dimensionalidad: $\boldsymbol{D_R^{-1}} \cdot \boldsymbol{F}$.

Lo mismo ocurre con las columnas, con algunas pequeñas diferencias.

La matriz de perfiles de columna se construye dividiendo las frecuencias absolutas por los márgenes de columna relativos:

$$\frac{n_{i,j}}{n_{.j}} = \frac{f_{i,j}}{f_{.j}} \qquad \forall\, i,j$$

La tabla de contingencia que obtendrás será:

$\frac{f_{i,j}}{f_{.j}} = \frac{n_{i,j}}{n_{.j}}$			perfilo medio
1	1	1	1

En este caso en los marginales de la columna tendrás todo 1 y en los marginales de la fila tendrás el perfil medio de la columna. En este caso las masas están representadas por los marginales de las columnas $f_{.j}$.

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Obviamente, incluso en el espacio de columnas se trabaja en menos de una dimensión, por lo que el espacio de columnas es R-1.

Se puede construir una matriz diagonal de columnas marginales $\boldsymbol{D_C}$, que tenga perfiles de columna en la diagonal mayor. La matriz diagonal de marginales de columna es una matriz **C·C,** que tiene dimensiones iguales a las columnas y en la diagonal mayor contiene los marginales de columna de la tabla de frecuencias relativas. Una matriz diagonal es una matriz cuyo elemento genérico en la diagonal mayor es el marginal de columna, arriba o abajo de ella, son todos ceros. Siempre es una matriz simétrica y cuadrada. Con la matriz diagonal de marginales de columna se puede construir la **matriz de perfiles de columna** : se obtiene dividiendo las frecuencias relativas entre los marginales de columna $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$**.** Las dimensiones de **F** son R· C, $\boldsymbol{D_C}$teniendo dimensión C·C, dado que no se puede hacer la división entre matrices, se calcula la inversa $\boldsymbol{D_C}$y posmultiplica a **F** , resolviendo así el problema de dimensionalidad: $\boldsymbol{F \cdot D_C^{-1}}$.

3.3) Distancias

En el análisis de correspondencias es necesario entender qué distancia hay entre los valores, esto para entender si las modalidades están lejos o cerca entre sí y por lo tanto si se parecen o no. Puedes hacer esto observando las frecuencias: cuanto más bajas son, más cerca están y viceversa. Existen varios métodos para calcular la distancia: **distancia euclidiana** y **distancia chi-cuadrado** .

La **distancia euclidiana** es la más sencilla y premia las distancias más altas a expensas de las más bajas. Se calcula haciendo la diferencia de las frecuencias relativas elevándolas al cuadrado.

Para perfiles de fila:

$$d_{(i,i')} = \sqrt{\sum_{j=1}^{C} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right)^2}$$

Para perfiles de columna:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

$$d_{(j,j')} = \sqrt{\sum_{i=1}^{R} \left(\frac{f_{i,j}}{f_{.j}} - \frac{f_{i,j'}}{f_{.j'}} \right)^2}$$

La **distancia chi-cuadrado** premia las distancias más bajas porque las frecuencias con número bajo son reponderadas con respecto a las filas, insertando en la fórmula la inversa de la marginal de columna (respecto a las columnas, insertando en la fórmula la inversa de la marginal de fila). La desventaja de la distancia chi-cuadrado es que el recíproco de los marginales de columna (o fila) puede tender a cero y, por lo tanto, una sola respuesta puede contribuir en exceso al cálculo de la distancia.

3.4) Espacio de Filas y Espacio de Columnas

En el **espacio de filas,** los dos componentes son:

- Perfil de fila:$\mathbf{D_R^{-1} \cdot F}$
- Métrico:$\mathbf{D_C^{-1}}$

Comencemos con la fórmula:

$$\boldsymbol{\Psi_{n\times 1} = X_{n\times p} \cdot u_{p\times 1}}$$

Haciendo las sustituciones apropiadas:

$$\boldsymbol{\Psi = D_R^{-1} \cdot F \cdot D_C^{-1} \cdot u}$$

El objetivo del análisis de correspondencias es el conjunto de ejes unitarios que permiten maximizar las distancias entre las proyecciones de los perfiles de fila. Por lo tanto, debemos buscar aquellos vectores que maximicen las proyecciones. Dado que los vectores $\boldsymbol{u}$pueden ser infinitos, se agrega la restricción de la norma unitaria.

$$\boldsymbol{u^T \cdot \mathbf{D_C^{-1}} \cdot u = 1}$$

Problema de maximización: Maximizar la inercia explicada (variación explicada), que corresponde a la variabilidad para variables cuantitativas.

$$\begin{cases} \text{MAX: } \left\{ \hat{\psi}^T D_R \hat{\psi} \right\} \\ v^T D_C^{-1} v = 1 \end{cases}$$

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Para resolver el problema de maximización con restricciones, utilice el método de los multiplicadores de Lagrange:

$$\mathcal{L}(\upsilon, \lambda) = (\hat{\psi}^T \boldsymbol{D}_R \hat{\psi}) - \lambda(\upsilon^T \boldsymbol{D}_C^{-1} \upsilon - 1)$$

$\boldsymbol{\lambda}$= multiplicador de Lagrange, que es un escalar;

***u*=** vector de pesos que buscamos

Realizando las sustituciones necesarias, tendremos:

$$\mathcal{L}(\upsilon, \lambda) = (\boldsymbol{D}_R^{-1} \boldsymbol{F} \boldsymbol{D}_C^{-1} \upsilon)^T \boldsymbol{D}_R (\boldsymbol{D}_R^{-1} \boldsymbol{F} \boldsymbol{D}_C^{-1} \upsilon) - \lambda(\upsilon^T \boldsymbol{D}_C^{-1} \upsilon - 1)$$

Realizamos las operaciones de transposición, sustituimos a $\boldsymbol{D_R} \cdot \boldsymbol{D_R^{-1}}$ por la matriz identidad $\boldsymbol{I}$ y $[(-\boldsymbol{\lambda}) \cdot (-1)]$ la reemplazamos por $\boldsymbol{\lambda}$. Entonces podemos eliminar la transpuesta de las matrices diagonales $\mathbf{D_C^{-1}}$ y $\mathbf{D_R^{-1}}$, ya que la transpuesta de una matriz diagonal no cambia. Conseguir:

$$\mathcal{L}(\upsilon, \lambda) = \upsilon^T \boldsymbol{D}_C^{-1} \boldsymbol{F}^T \boldsymbol{D}_R^{-1} \boldsymbol{F} \boldsymbol{D}_C^{-1} \upsilon - \lambda \upsilon^T \boldsymbol{D}_C^{-1} \upsilon + \lambda$$

Calculamos las derivadas parciales, derivando el lagrangiano respecto a $\boldsymbol{u}$ y las igualamos a 0:

$$\frac{\partial \mathcal{L}}{\partial \upsilon} = 0 \rightarrow \frac{\partial \mathcal{L}}{\partial \upsilon} = 2\boldsymbol{F}^T \boldsymbol{D}_R^{-1} \boldsymbol{F} \boldsymbol{D}_C^{-1} \upsilon - 2\lambda \upsilon = 0$$

Multiplica la ecuación por $\mathbf{D_C^{-1}}$:

$$\boldsymbol{F}^T \boldsymbol{D}_R^{-1} \boldsymbol{F} \boldsymbol{D}_C^{-1} \upsilon = \lambda \upsilon$$

Si reemplazamos la transposición de perfiles de fila y la matriz de perfiles de columna con $\boldsymbol{S}$, podemos escribir la ecuación característica como:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

$$Sv = \lambda v$$

Maximizar la inercia explicada de los perfiles de fila equivale a descomponer esta matriz en autovalores y autovectores de la misma. El primer valor propio está asociado con el primer vector propio que explica la inercia máxima. Los autovectores que se extraigan posteriormente , se extraerán ortogonalmente poniendo la restricción de ortogonalidad

$$\boldsymbol{u_1^T \cdot D_C^{-1} \cdot u_2 = 0}$$

Usamos la restricción de ortogonalidad para poder elegir el segundo componente que explicará la inercia que no explica el primer componente. Obviamente, la primera componente extraída explica la máxima inercia, es decir, la máxima elongación de la nube de puntos.

En **el espacio de columnas** dos componentes son:

- Perfil de columna:$\mathbf{F} \cdot \mathbf{D_C^{-1}}$
- Métrico:$\mathbf{D_R^{-1}}$

Comencemos con la fórmula:

$$\boldsymbol{\varphi_{p\times 1} = \left(X_{n\times p}^T\right)_{p\times n} \cdot v_{n\times 1}}$$

Reemplazamos y obtenemos

$$\boldsymbol{\varphi = \mathbf{D_C^{-1}} F^T \mathbf{D_R^{-1}} v}$$

El problema de maximización a resolver con multiplicadores de Lagrange es:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

$$\begin{cases} \text{MAX: } \{\hat{\varphi}^T \boldsymbol{D}_C \hat{\varphi}\} \\ \boldsymbol{\nu}^T \boldsymbol{D}_R^{-1} \boldsymbol{\nu} = 1 \end{cases}$$

Procediendo como en el espacio de las filas, finalmente obtendremos:

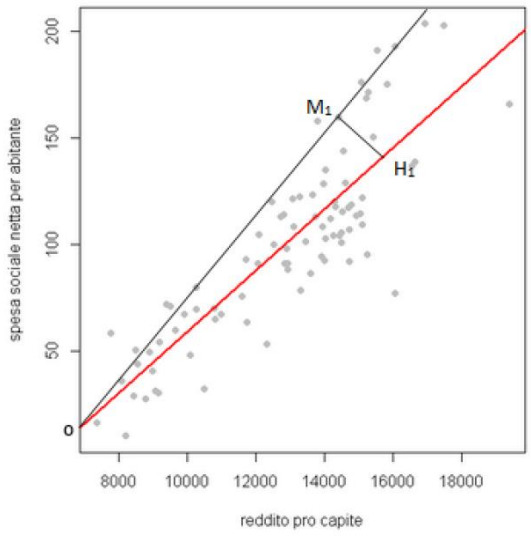
$$\boldsymbol{F}\boldsymbol{D}_C^{-1}\boldsymbol{F}^T\boldsymbol{D}_R^{-1}\boldsymbol{\nu} = \mu\boldsymbol{\nu}$$

Sustituyendo la matriz de perfiles de columna y la métrica transpuesta de perfiles de fila por $\boldsymbol{S}^*$ obtenemos la ecuación característica:

$$\boldsymbol{S}^*\boldsymbol{\nu} = \mu\boldsymbol{\nu}$$

Maximizar geométricamente la inercia explicada, es decir, hacer que la información perdida sea lo más pequeña posible y la información observada lo más grande posible, será: hacer la distancia lo más pequeña posible M_1H_1 y la distancia OH_1 lo más grande posible.

Figura 1.3: Diagramma di dispersione



Por tanto, debemos encontrar la recta f (en rojo) interpolando los puntos del espacio vectorial, para que la distancia entre todos los

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

puntos del espacio y los puntos proyectados ortogonalmente sobre la recta f sea la mínima posible.

Los valores propios en el espacio de filas corresponden a los vectores propios en el espacio de columnas, por lo que los valores propios de **S** corresponden a los de $\boldsymbol{S}^*$. Los vectores propios son iguales entre sí excepto por una constante. Entonces, cuando tenemos que maximizar, no necesitamos descomponer en valores propios y vectores propios **S** y $\boldsymbol{S}^*$, solo hacerlo con uno. La cantidad de inercia explicada es igual ya sea que calculemos **S** o $\boldsymbol{S}^*$, la relación entre los dos espacios está representada por las **fórmulas de transición** :

$$\boldsymbol{S} \to \boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F} \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon} \equiv \boldsymbol{S}^* \to \boldsymbol{\upsilon} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Espacio de filas :

$$\hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon}$$

Con:

$$\boldsymbol{\upsilon} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Aplicando las sustituciones adecuadas:

$$\frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\upsilon} \to \frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F}' \hat{\boldsymbol{\psi}}$$

Conseguir:

$$\sqrt{\lambda} \hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{F}' \hat{\boldsymbol{\psi}} \to \hat{\boldsymbol{\psi}} = \frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\boldsymbol{\psi}} \to \sqrt{\lambda} \hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\boldsymbol{\psi}}$$

Para el espacio de las filas, por lo tanto:

$$\sqrt{\lambda} \hat{\boldsymbol{\psi}} = \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\boldsymbol{\psi}} \equiv \hat{\boldsymbol{\psi}} = \sqrt{\lambda} \hat{\boldsymbol{\psi}}$$

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Espacio de columna:

$$\hat{\psi} = D_R^{-1}\nu$$

Dónde:

$$\nu = \frac{1}{\sqrt{\lambda}} F D_C^{-1} v$$

Aplicando las sustituciones adecuadas:

$$\frac{1}{\sqrt{\lambda}} D_R^{-1} F D_C^{-1} v \rightarrow \frac{1}{\sqrt{\lambda}} D_R^{-1} F \hat{\psi}$$

Conseguir:

$$\frac{1}{\sqrt{\lambda}} D_R^{-1} F \hat{\psi} \rightarrow \sqrt{\lambda}\hat{\psi} \rightarrow D_R^{-1} F \hat{\psi}$$

Para el espacio de la columna:

$$\sqrt{\lambda}\hat{\psi} = D_R^{-1} F \hat{\psi} \equiv \hat{\psi} = \sqrt{\lambda}\hat{\psi}$$

4) Ejemplo con software R

Queremos estudiar la posible relación entre las distribuciones del ganado y las diferentes regiones italianas. Los datos se refieren al año 2011, recogidos por los bancos disponibles en la web del Istat.

Hipótesis: las distintas regiones, según las características territoriales y las necesidades de la población, optan por criar unas cabezas de ganado frente a otras.

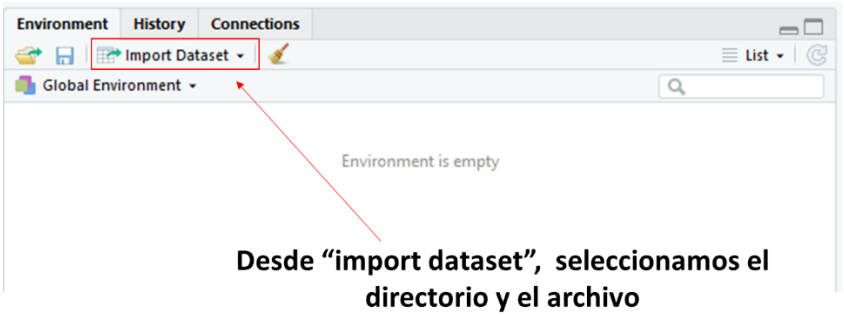
Conjunto de datos:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Regione	Bovini	Ovini	Caprini	Equini	Suini	Conigli	Totale
Piemonte	23516	2303	3418	2370	2429	1392	35428
Valle d'Aosta	1585	347	284	53	16	11	2296
Liguria	1642	1126	549	949	258	924	5448
Lombardia	15480	2592	3175	3647	4346	1191	30431
Trentino Alto Adige	10482	2279	2424	1513	3292	266	20256
Veneto	16007	1642	1207	2429	3634	1907	26826
Friuli-Venezia Giulia	1539	83	207	280	1477	117	3703
Emilia-Romagna	8522	1315	908	3161	1541	308	15755
Toscana	4392	4918	607	2163	2046	1764	15890
Umbria	3132	2734	667	1245	4107	1924	13809
Marche	2940	1877	342	383	7103	1786	14431
Lazio	9256	8678	1624	3535	6849	4269	34211
Abruzzo	5588	6590	1710	1362	10241	2450	27941
Molise	2976	2510	610	534	3943	60	10633
Campania	10971	6248	3675	1448	15145	6708	44195
Puglia	3010	1918	826	691	759	921	8125
Basilicata	3156	7426	3562	1280	6137	2606	24167
Calabria	5496	3701	3505	1839	21522	2087	38150
Sicilia	7387	4963	1088	1930	821	63	16252
Sardegna	8200	12880	3171	3333	9324	523	37431
Totale	145277	76130	33559	34145	104990	31277	425378

Importamos el conjunto de datos:



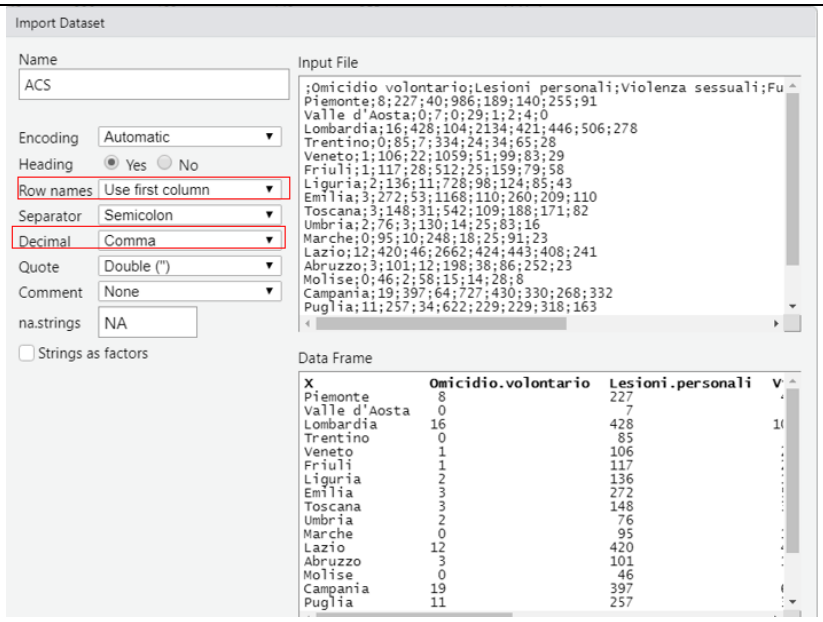
Desde “import dataset”, seleccionamos el directorio y el archivo

En el campo ***row names***, seleccionamos: “***use first column***" para tener las etiquetas de individuos y variables en los gráficos.

En el campo ***decimal*** seleccionamos "***comma***".

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.



Con el comando:

X<- as.matrix (nombre_del_conjunto de datos)

Atribuimos a **X**, como objeto, el conjunto de datos utilizado en el análisis.

Antes de poder realizar el AC es necesario establecer el grado de interdependencia entre las dos variables consideradas, esto se debe a que en el caso de que sean independientes puede no tener sentido continuar con el AC. Para verificar esto realizamos la prueba de chi-cuadrado.

El comando es:

Nombre del objeto en R<- chisq.test (X)

```
        Pearson's Chi-squared test

data:  X
X-squared = 126691.2, df = 95, p-value < 2.2e-16
```

Se puede observar que el ***valor de p*** es inferior al nivel de significación más utilizado, es decir, 0,05. Por lo tanto, podemos rechazar la hipótesis nula de independencia estadística entre las dos variables y podemos continuar con el análisis.

Co-funded by the Erasmus+ Programme of the European Union

Ahora queremos crear una matriz de frecuencias relativas **F**.

Calculamos el número de muestra, con el comando:

n<-sum(X)

y luego dividiendo la matriz inicial (por lo tanto todas las frecuencias conjuntas) por el número de muestra obtenemos la matriz **F** . Dominio:

F<-X/n

El siguiente paso es obtener las tablas **de perfiles fila y columna**. Para ello, en primer lugar, es necesario calcular las sumas de fila y columna. Respectivamente los comandos son:

umrow<-apply(F,1,sum)
sumcol<-apply(F,2,sum)

Luego calculamos la matriz diagonal de los marginales de fila y su inversa con los comandos:

Dr<-diag(sumrow)
Dr_inv<-solve(Dr)

Ahora podemos calcular perfiles fila. En términos matriciales, premultiplicamos la inversa de la matriz diagonal de la fila marginal a la matriz de frecuencias relativas. El comando a utilizar es:

Pr <- Dr_inv %*%F

Lo mismo para los perfiles de columnas, recordando que en este caso se debe postmultiplicar la inversa de la matriz de columnas por la matriz de frecuencias relativas.

Dc<-diag(sumcol)
Dc_inv<-solve(Dc)
Pc<-F%*%Dc_inv

Ahora podemos calcular las distancias entre los puntos. Como ya se mencionó, hay dos tipos de distancia: **euclídea** y **chi-cuadrado** .

Distancia euclídea para perfiles fila:

d_euc_r <- dist (rbind (Pr [1,], Pr [2,]))

Co-funded by the Erasmus+ Programme of the European Union

Distancia euclídea para perfiles columna:

d_euc_c <- dist (rbind (Pr [,1], Pr [,2]))

Distancia chi-cuadrado para perfiles fila:

d_r <-pr[1,]-pr[2,]
d<-d_r^2/ sumcol
d_chi_r <-sqrt(sum(d))

Distancia chi-cuadrado para perfiles columna:
dc<-Pr[,1]-Pr[,2]
dc<-dc^2/sumrow
d_chi_c<-sqrt(sum(dc))

La ecuación característica de la matriz **de perfiles fila es**:

S<-t(Pr)%*%Pc

Como la matriz S no es simétrica, es necesario diagonalizarla para obtener **S_tilde :**

A<-t(F)%*%Dr_inv%*%F #simmetria

Dc_12<-diag(sumcol^(-1/2))

S_tilde<-Dc_12%*%A%*%Dc_12

Ahora tenemos que maximizar la inercia explicada descomponiendo la matriz en autovalores y vectores propios:

AC<-eigen(S_tilde)

lambda<-as.matrix(AC$values)

lambda<-lambda[-1,]

w<-AC$vectors

u<-Dc^(1/2)%*%w

u<-u[,-1]

Co-funded by the Erasmus+ Programme of the European Union

La ecuación característica de la matriz de **perfiles de columna es:**

S_star<-F%*%Dc_inv%*%t(F)%*%Dr_inv

Para pasar de ***u*** a ***v***, usamos fórmulas de transición (ya que la cantidad de inercia explicada es igual por filas y columnas).

sq_lambda<-diag((sqrt(lambda))^(-1))

v<-F%*%Dc_inv%*%u%*%sq_lambda

Calculamos factores y coordenadas, primero por filas y luego columnas:

fp_r<-Dc_inv%*%u

fp_c<-Dr_inv%*%v

PHI_coord<-Dc_inv%*%t(F)%*%fp_c

PSI_coord<-Dr_inv%*%F%*%fp_r

Representamos la gráfica con las coordenadas principales:

PRINCOORD<-rbind(PSI_coord,PHI_coord)

rows<-row.names(X);columns<-colnames(X)

plot(PRINCOORD[,1],PRINCOORD[,2],type="n",main="Main Coordinates",xlab="Axis1",ylab="Axis2")+ text(PRINCOORD[1:20,1],PRINCOORD[1:20,2],labels=rows,col="spring green4")

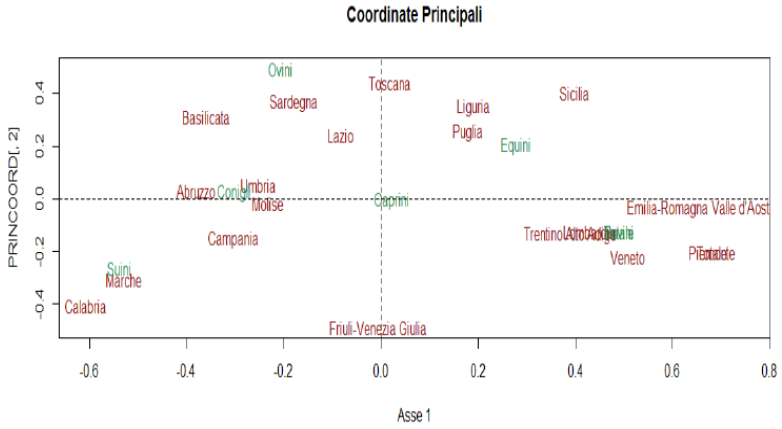
text(PRINCOORD[21:29,1],PRINCOORD[21:29,2],labels=columns,col="violetred")

abline(h=0,v=0,lty=2,lwd=1.5)

Y así obtenemos:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.



Mirando este gráfico podemos decir, por ejemplo, que en regiones como Abruzzo, Molise o Umbría se crían principalmente conejos.

Seleccionamos los componentes:

inertia<-sum(diag(S))-1

sum(lambda)

in_exp<-lambda/inertia

in_exp_<-cumsum(in_exp)

Y visualizamos los resultados obtenidos:

```
> inerzia
[1] 0.2978321
> in_exp
[1] 0.58571295 0.23305781 0.10382933 0.04875445 0.02864546
> in_exp_cum
[1] 0.5857130 0.8187708 0.9226001 0.9713545 1.0000000
```

La primera dimensión por sí sola explica el 58,57% de la variabilidad, y las tres primeras juntas explican el 92,26% de la variabilidad global de los datos.

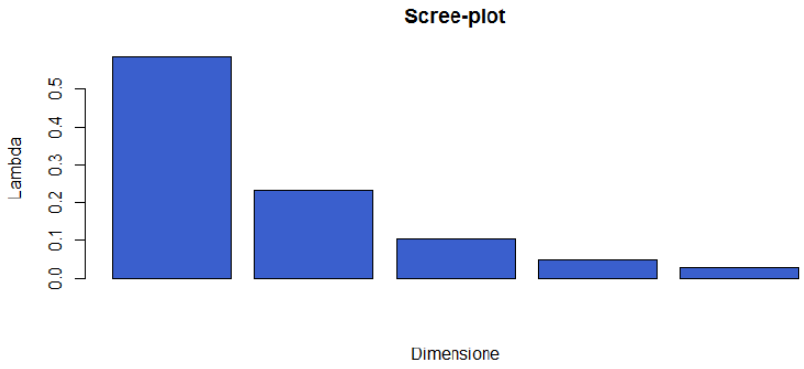
Los resultados obtenidos se pueden visualizar gráficamente con el **scree-plot de la inercia explicada**:

screeplot<-barplot(in_exp,main="Scree-plot inertia", xlab="Size", ylab="Lambda", col="lightblue")

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Figura 1.10: Scree-plot dell'inerzia spiegata



Para estudiar la calidad de la representación, procedemos como sigue:

- para evaluar cuánto influye o participa una categoría en el eje factorial calculamos **las contribuciones absolutas** , **ca,** tanto para filas como para columnas:

ca_r <-Dr%*%fp_c^2

ca_c <-DC%*%fp_r^2

para evaluar la calidad de la representación calculamos las **contribuciones relativas, cr.** Éstas dan una mejor medida de la representación de los puntos sobre los ejes y vienen dadas por el coseno del ángulo formado por el vector de proyección del punto y el vector relativo i (o j) en el *punto* i *(* o *j*) en su espacio **G<-matrix(sumcol,20,9,byrow=T)**

di<-(Pr-G)^2%*%Dc_inv

d_ig<-apply(di,1,sum)

cos2r<-PSI_coord^2/d_ig

H<-matrix(sumrow,20,9)

dj<-Dr_inv%*%(Pc-H)^2

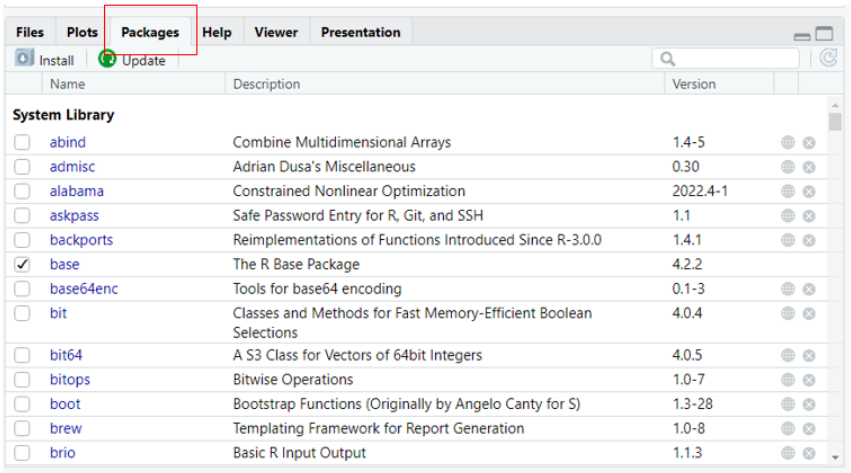
d_jh<-apply(dj,2,sum)

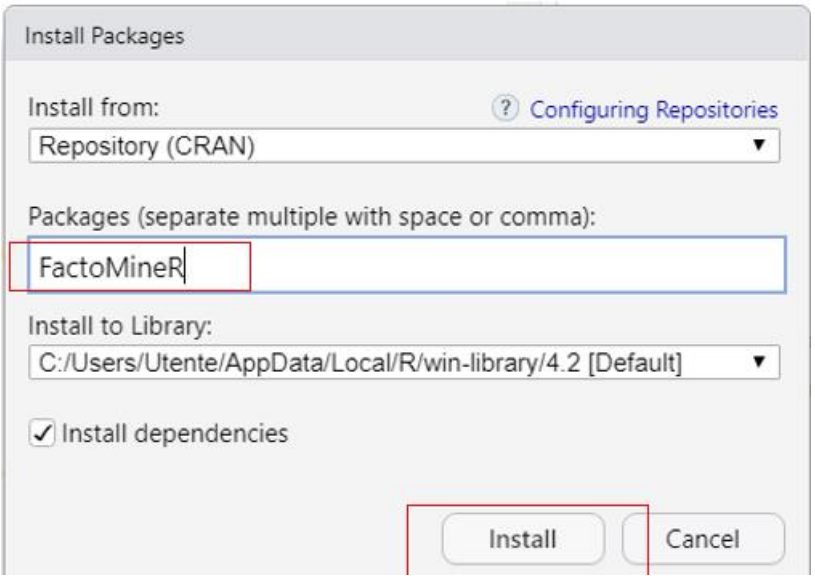
COS2C<-PHI_coord^2/d_jh

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

R posibilita emplear un paquete llamado **FactoMineR** para el análisis de correspondencias, que agrega información sobre puntos y variables y permite crear un gráfico bidimensional conjunto. Para poder usar este paquete de R primero debes descargarlo:





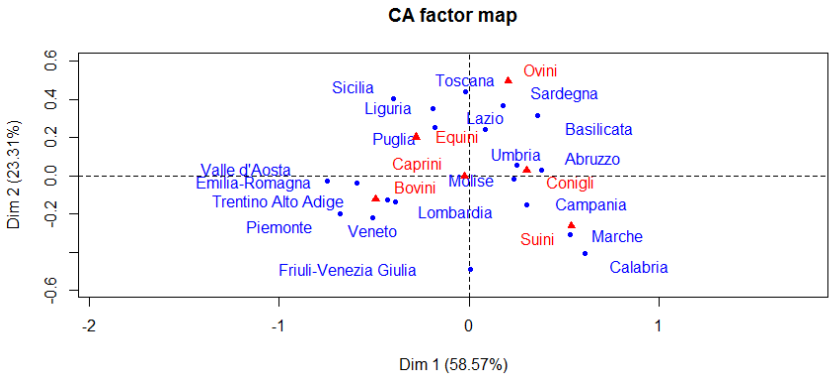
Después de instalarlo, debe llamarlo con el comando:

library(FactoMineR)

Pasemos a la creación del gráfico bidimensional para puntos y variables:

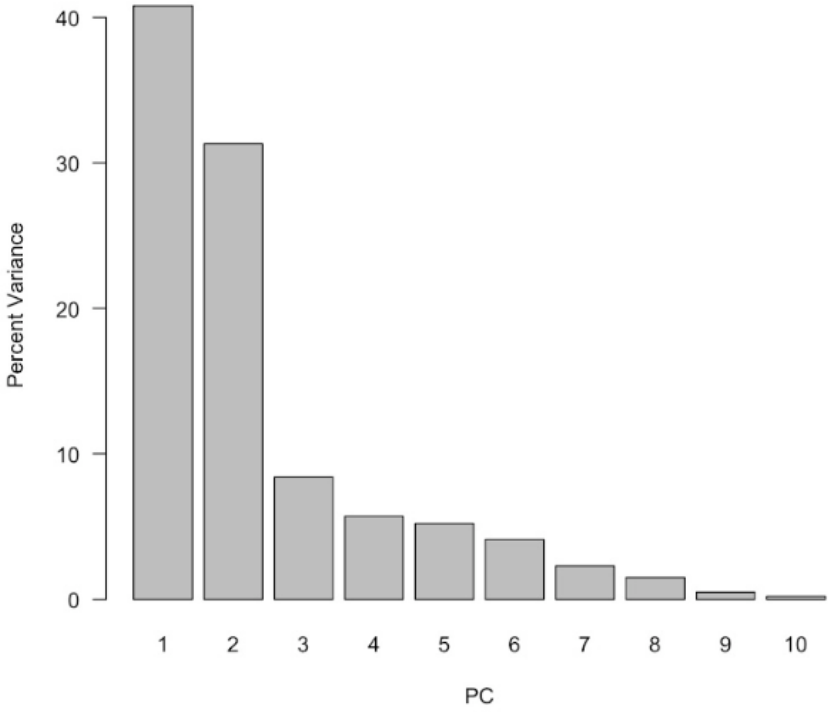
Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	CA(X, ncp = 5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL) Gráficamente tendremos:  Interpretación de resultados: Podemos decir que se confirma la hipótesis inicial. En particular, las regiones más dedicadas a la ganadería ovina parecen ser Toscana, Cerdeña y Basilicata, y esto puede explicarse por el hecho de que estas regiones son áreas de montaña y trashumancia. Los caballos se crían principalmente en Puglia, Liguria y Sicilia porque estos animales siempre se han utilizado para el trabajo en el campo. El ganado está presente en Trentino Alto-Adige, Veneto, Piamonte, Lombardía y Emilia-Romaña. De hecho, estas regiones tienen una tradición de crianza más desarrollada para uso alimentario. Los conejos aparecen principalmente en Umbria, Abruzzo y Molise. En cambio, parece que los cerdos se crían más en Marche, Campania y Molise. Estas regiones también tienen una tradición de cría más desarrollada para uso alimentario.Las cabras, en cambio, se colocan en medio de los ejes, probablemente porque no hay regiones que prefieran su cría.
Autoevaluación (preguntas y respuestas de opción múltiple)	1. El Análisis de Correspondencias tiene como objetivo: A) La agregación de las unidades estadísticas según su distancia B) La reducción de la dimensionalidad de un fenómeno complejo C) La descripción de un conjunto de datos 2. La matriz de datos inicial de un AC debe ser: A) Con datos cualitativos B) Con datos estandarizados C) Con datos cuantitativos

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	3. Los componentes extraídos en el Análisis de Correspondencias: A) Son combinaciones lineales de las variables de partida B) Tienen la propiedad de equidistribución c) Todos tienen valores propios mayores que 1 4. Con cuántas dimensiones explicarías el siguiente fenómeno?  A. Una B. Dos C. Tres
Recursos (videos, enlaces a referencias)	Pozzolo P., *Analisi delle componenti principali: da dove partire*, https://paolapozzolo.it/analisi-delle-componenti-principali-criteri/ Gilardone A., *Analisi delle componenti principali: 7 passaggi da eseguire* https://adrianogilardone.com/analisi-delle-componenti-principali/ Gilardone A., https://www.youtube.com/watch?v=OksC-g4K2gY Vardanega A., L'Analisi in componenti principali https://www.agnesevardanega.eu/wiki/r/analisi_esplorativa/analisi_in_componenti_principali

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

	Zakaria Jaadi, *A Step-by-Step Explanation of Principal Component Analysis (PCA)*, https://builtin.com/data-science/step-step-explanation-principal-component-analysis Ian T. Jolliffe and Jorge Cadima, *Principal component analysis: a review and recent developments*, https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 Science Snippets Blog, *What Is Principal Component Analysis (PCA) and How It Is Used?*, 2020 https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186
Material relacionado	
PPT relacionado	
Bibliografía	
Proporcionado por	[UNISALENTO/DEMOSTENE CENTRO ESTUDIO]

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.