Introducción al Análisis Multivariante

Vectores aleatorios, técnicas de análisis multivariante, distancias estadísticas

Curso 2011-2012

Considero que el cerebro de cada cual es como una pequeña pieza vacía que vamos amueblando con elementos de nuestra elección. Un necio echa mano de cuanto encuentra a su paso, no encuentra cabida o, en el mejor de los casos, se halla tan revuelto con las demás cosas que resulta difícil dar con él. El operario hábil selecciona con sumo cuidado el contenido de ese vaso disponible que es su cabeza. Sólo de herramientas útiles se compondrá su arsenal, pero éstas serán abundantes y estarán en perfecto estado. Constituye un grave error el suponer que las paredes de la pequeña habitación son elásticas o capaces de dilatarse indefinidamente. A partir de cierto punto, cada nuevo dato añadido desplaza necesariamente a otro que ya poseíamos. Resulta por tanto de inestimable importancia vigilar que los hechos inútiles no arrebaten espacio a los útiles.

Sherlock Holmes en "Estudio en Escarlata"

Introducción

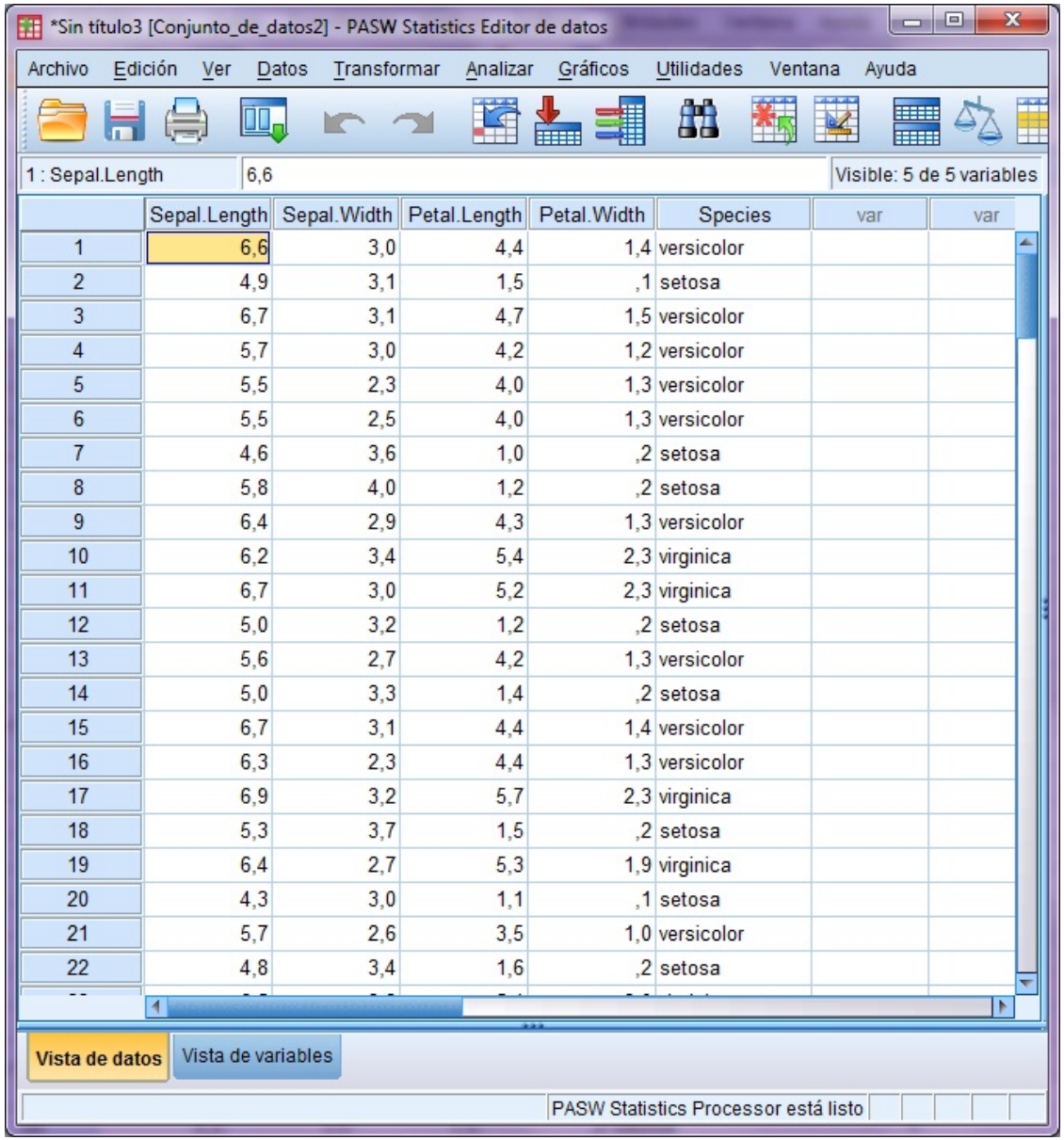
- Son muchas las situaciones reales en las que necesitamos tener en cuenta varias variables de forma simultánea.
- Podemos pensar en problemas sencillos en todas las disciplinas en los que registremos datos de más de una variable sobre distintos elementos o individuos de una muestra. Por ejemplo:
  - Si los individuos son organismos, podemos recoger datos de diferentes medidas morfológicas o psicológicas.
  - En ecología se suele disponer de distintas medidas químicas obtenidas sobre los individuos de la muestra.
  - ...
- En algunas ocasiones puede resultar adecuado estudiar cada una de las variables de interés de forma individual. Sin embargo, en general las variables están relacionadas entre sí de tal manera que los análisis individuales proporcionan poca información sobre la estructura del conjunto de datos.

Introducción

- Ya hemos considerado datos multivariantes al hacer regresión múltiple. Aun así, hay muchas otras técnicas que permiten analizar datos multivariantes.
- Las técnicas de análisis multivariante incluyen tanto métodos puramente descriptivos que tienen por objetivo extraer información de los datos disponibles, como métodos de inferencia que, a través de la construcción de modelos, pretenden obtener conclusiones sobre la población que ha generado los datos.

Introducción

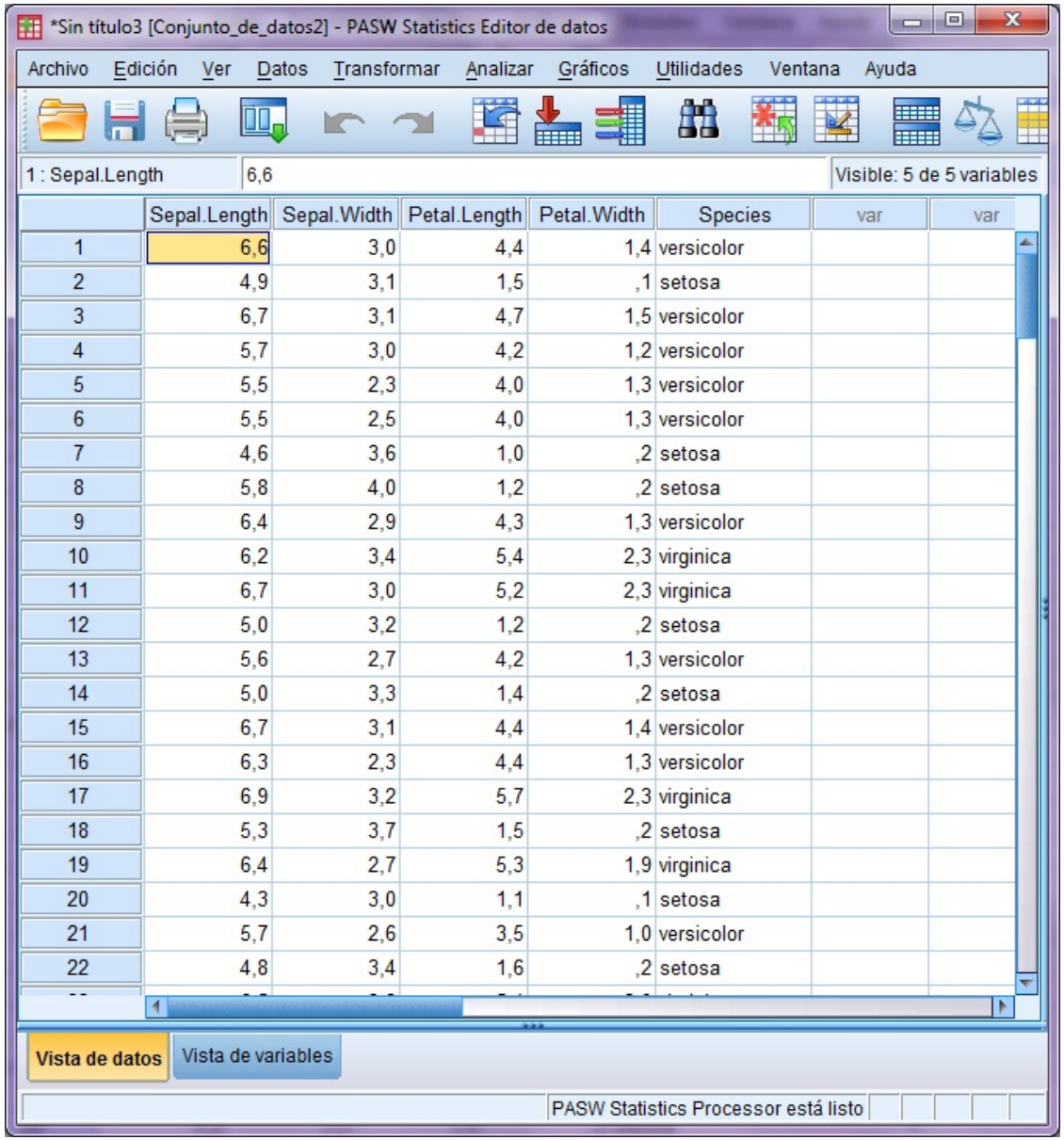
Ejemplo de datos multivariantes: datos Iris de Fisher



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	var	var
1	6,6	3,0	4,4	1,4	versicolor		
2	4,9	3,1	1,5	,1	setosa		
3	6,7	3,1	4,7	1,5	versicolor		
4	5,7	3,0	4,2	1,2	versicolor		
5	5,5	2,3	4,0	1,3	versicolor		
6	5,5	2,5	4,0	1,3	versicolor		
7	4,6	3,6	1,0	,2	setosa		
8	5,8	4,0	1,2	,2	setosa		
9	6,4	2,9	4,3	1,3	versicolor		
10	6,2	3,4	5,4	2,3	virginica		
11	6,7	3,0	5,2	2,3	virginica		
12	5,0	3,2	1,2	,2	setosa		
13	5,6	2,7	4,2	1,3	versicolor		
14	5,0	3,3	1,4	,2	setosa		
15	6,7	3,1	4,4	1,4	versicolor		
16	6,3	2,3	4,4	1,3	versicolor		
17	6,9	3,2	5,7	2,3	virginica		
18	5,3	3,7	1,5	,2	setosa		
19	6,4	2,7	5,3	1,9	virginica		
20	4,3	3,0	1,1	,1	setosa		
21	5,7	2,6	3,5	1,0	versicolor		
22	4,8	3,4	1,6	,2	setosa		

Introducción

Ejemplo de datos multivariantes: datos Iris de Fisher



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	var	var
1	6,6	3,0	4,4	1,4	versicolor		
2	4,9	3,1	1,5	,1	setosa		
3	6,7	3,1	4,7	1,5	versicolor		
4	5,7	3,0	4,2	1,2	versicolor		
5	5,5	2,3	4,0	1,3	versicolor		
6	5,5	2,5	4,0	1,3	versicolor		
7	4,6	3,6	1,0	,2	setosa		
8	5,8	4,0	1,2	,2	setosa		
9	6,4	2,9	4,3	1,3	versicolor		
10	6,2	3,4	5,4	2,3	virginica		
11	6,7	3,0	5,2	2,3	virginica		
12	5,0	3,2	1,2	,2	setosa		
13	5,6	2,7	4,2	1,3	versicolor		
14	5,0	3,3	1,4	,2	setosa		
15	6,7	3,1	4,4	1,4	versicolor		
16	6,3	2,3	4,4	1,3	versicolor		
17	6,9	3,2	5,7	2,3	virginica		
18	5,3	3,7	1,5	,2	setosa		
19	6,4	2,7	5,3	1,9	virginica		
20	4,3	3,0	1,1	,1	setosa		
21	5,7	2,6	3,5	1,0	versicolor		
22	4,8	3,4	1,6	,2	setosa		

Introducción

¿Qué podemos hacer con técnicas de análisis multivariante?

- Contrastar la hipótesis de que las medias de las variables analizadas tienen un valor específico (Inferencia sobre la media en poblaciones multivariantes).

Introducción

¿Qué podemos hacer con técnicas de análisis multivariante?

- Contrastar la hipótesis de que las medias de las variables analizadas tienen un valor específico (Inferencia sobre la media en poblaciones multivariantes).
- Representar la información mediante un número menor de variables construidas como combinaciones lineales de las originales y que expliquen la mayor parte de la variabilidad original (Análisis de Componentes Principales).

Introducción

¿Qué podemos hacer con técnicas de análisis multivariante?

- Contrastar la hipótesis de que las medias de las variables analizadas tienen un valor específico (Inferencia sobre la media en poblaciones multivariantes).
- Representar la información mediante un número menor de variables construidas como combinaciones lineales de las originales y que expliquen la mayor parte de la variabilidad original (Análisis de Componentes Principales).
- Encontrar un modelo que nos permita predecir un grupo de variables del conjunto original a partir de otro grupo de variables (Modelos de regresión multivariante).

Introducción

¿Qué podemos hacer con técnicas de análisis multivariante?

- Contrastar la hipótesis de que las medias de las variables analizadas tienen un valor específico (Inferencia sobre la media en poblaciones multivariantes).
- Representar la información mediante un número menor de variables construidas como combinaciones lineales de las originales y que expliquen la mayor parte de la variabilidad original (Análisis de Componentes Principales).
- Encontrar un modelo que nos permita predecir un grupo de variables del conjunto original a partir de otro grupo de variables (Modelos de regresión multivariante).
- Comparar las medias de las variables en dos poblaciones (Test de Hotelling)

Introducción

¿Qué podemos hacer con técnicas de análisis multivariante?

- Contrastar la hipótesis de que las medias de las variables analizadas tienen un valor específico (Inferencia sobre la media en poblaciones multivariantes).
- Representar la información mediante un número menor de variables construidas como combinaciones lineales de las originales y que expliquen la mayor parte de la variabilidad original (Análisis de Componentes Principales).
- Encontrar un modelo que nos permita predecir un grupo de variables del conjunto original a partir de otro grupo de variables (Modelos de regresión multivariante).
- Comparar las medias de las variables en dos poblaciones (Test de Hotelling)
- Comparar las medias de las variables en más de dos poblaciones (Análisis Multivariante de la Varianza).

Introducción

¿Qué podemos hacer con técnicas de análisis multivariante?

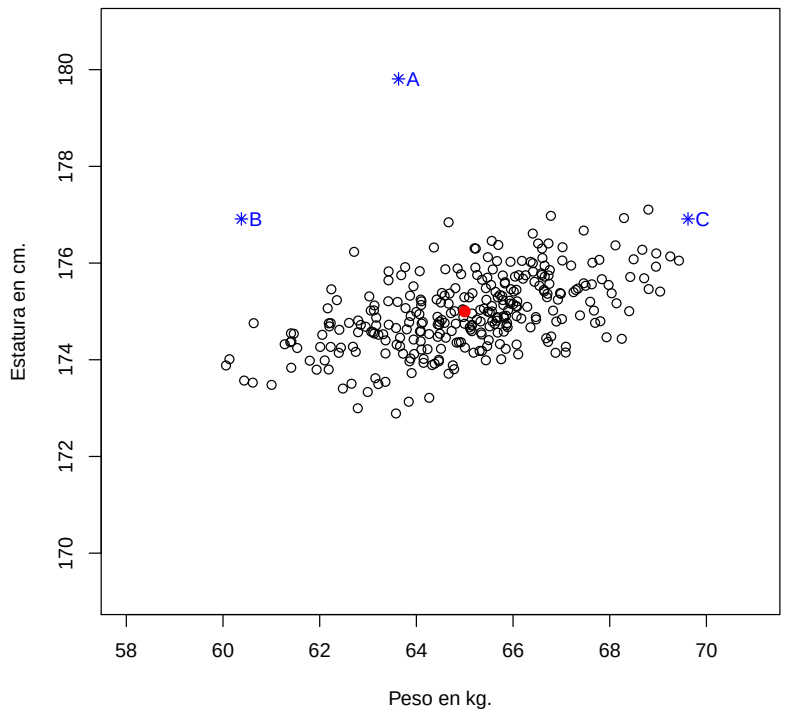
- Contrastar la hipótesis de que las medias de las variables analizadas tienen un valor específico (Inferencia sobre la media en poblaciones multivariantes).
- Representar la información mediante un número menor de variables construidas como combinaciones lineales de las originales y que expliquen la mayor parte de la variabilidad original (Análisis de Componentes Principales).
- Encontrar un modelo que nos permita predecir un grupo de variables del conjunto original a partir de otro grupo de variables (Modelos de regresión multivariante).
- Comparar las medias de las variables en dos poblaciones (Test de Hotelling)
- Comparar las medias de las variables en más de dos poblaciones (Análisis Multivariante de la Varianza).
- Clasificar en dos o más grupos a individuos en los que hemos observado varias variables (Análisis Cluster).

Distancias estadísticas

- El concepto de distancia entre objetos o individuos observados permite interpretar geométricamente muchas técnicas de análisis multivariante.
- En el caso unidimensional, la distancia entre dos puntos x e y se mide de manera natural mediante la distancia euclídea $|x - y|$.
- ¿Y cuando disponemos de una variable vectorial?

Distancias estadísticas

Ejemplo: Disponemos de los datos de peso y estatura de 300 mujeres con edades comprendidas entre 30 y 40 años. Queremos determinar la posición con respecto a la media de tres nuevas mujeres a partir de sus respectivos pesos y estaturas. Supongamos que la mujer A pesa 63 kg. y mide 180 cm. La mujer B pesa 60 kg. y mide 177 cm. La mujer C pesa 69 kg. y mide 177 cm.



Distancias estadísticas

Sean $X_i = (X_{i1}, \ldots, X_{id})$ y $X_k = (X_{k1}, \ldots, X_{kd})$ las observaciones de dos individuos i, k obtenidas al medir el vector d-dimensional $(X_1, \ldots, X_d)$.

- Se define la **distancia euclídea** entre X_i y X_j como

$$d_E(X_i, X_k) = \sqrt{\sum_{j=1}^{d} (X_{ij} - X_{kj})^2}$$

  La distancia euclídea es la más utilizada pero tiene como inconvenientes que:

  - depende de las unidades de medida de las variables (no es invariante ante cambios de escala) y
  - presupone que las variables son incorrelacionadas y de varianza unidad.

- Se define la **distancia de Mahalanobis** entre X_i y X_k como

$$d_M(X_i, X_k) = \sqrt{(X_i - X_k)' \Sigma^{-1} (X_i - X_k)}$$

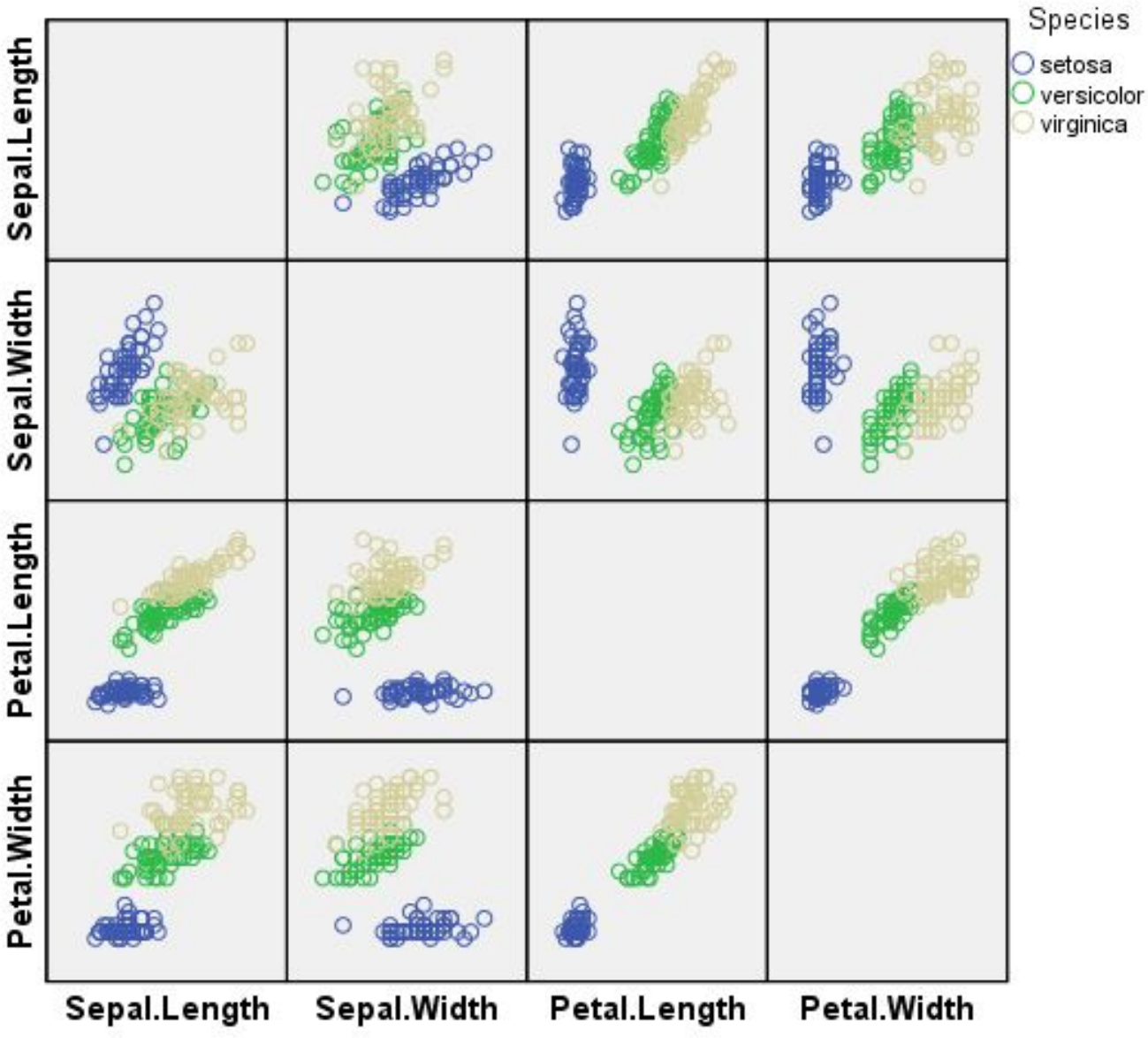
  donde Σ representa la matriz de covarianzas. Es adecuada como medida de discrepancia entre datos, porque

  - es invariante ante cambios de escala,
  - tiene en cuenta las correlaciones entre las variables

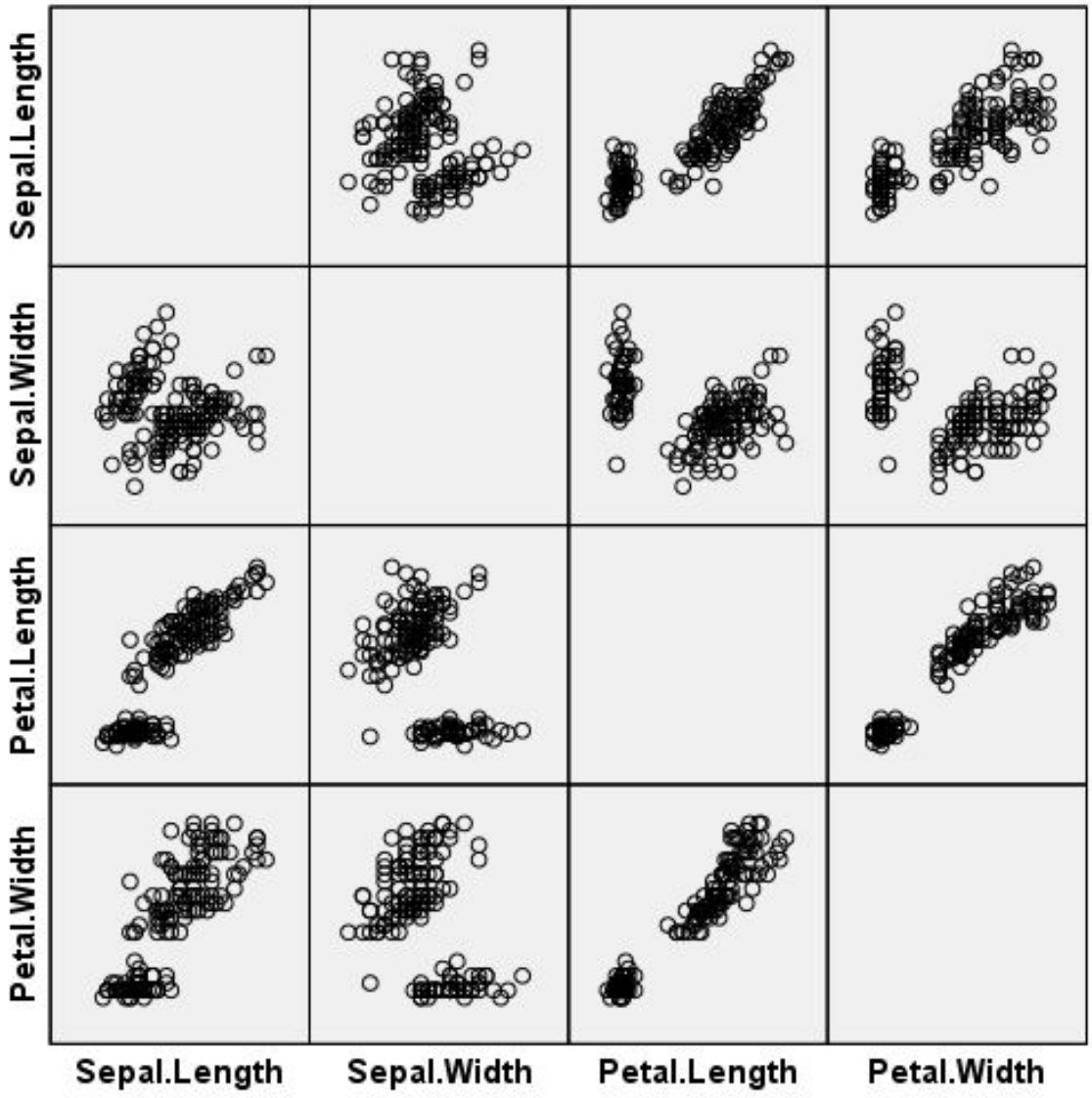
Análisis cluster

- Los métodos de análisis cluster tiene por objetivo identificar grupos de individuos con características comunes a partir de la observación de varias variables en cada uno de ellos.
- Un **cluster** es un grupo de individuos homogéneos entre sí y separados de los individuos de los otros clusters.
- El objetivo es por lo tanto ordenar los individuos en grupos de forma que el grado de asociación/similitud entre miembros del mismo cluster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters.

Análisis cluster



Análisis cluster



Análisis cluster

- Métodos basados en particiones: Producen una partición de los individuos en un número especificado de grupos. Ejemplo: Algoritmo de k-medias.
- Métodos jerárquicos:
  - Métodos divisivos. Parten de un único cluster que se va dividiendo paso a paso, hasta obtener tantos clusters como datos.
  - Métodos aglomerativos: Parten de tantos clusters como datos y en cada paso se van juntando dos clusters hasta obtener un único cluster con todos los datos.