Exploración multivariante de datos ómicos

Descriptivo, PCA y Clustering

Alex Sanchez-Pla

Departamento de Genética, Microbiología y Estadística Universitat de Barcelona

Versión 2024-03-25

# Esquema

1. Introducción y Motivación
2. Exploración de datos multivariantes
3. Reducción de la dimensión
4. Descubrimiento de grupos en los datos
5. Referencias y recursos

2 / 45

# Introducción y motivación

3 / 45

# ¿El enfoque adecuado para Ómicas?

Estadística Clásica

 Regresión múltiple

 Análisis discriminante  ANOVA

Las tablas de datos son *largas* y *no*

*muy anchas ("más indivíduos que variables)*

# La naturaleza de los datos ómicos

Los datos ómicos tienen estructura distinta a la habitual en estadística multivariante.

Miden características distintas

 Espectros GC/MS, Expresión, Concentración...

Aunque tienen aspectos en común

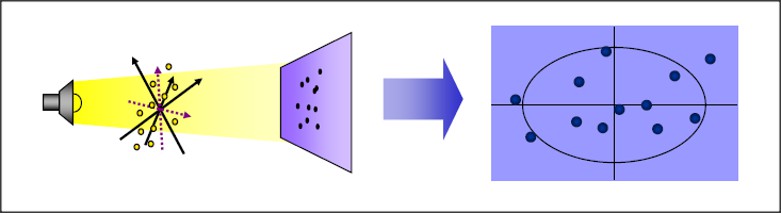


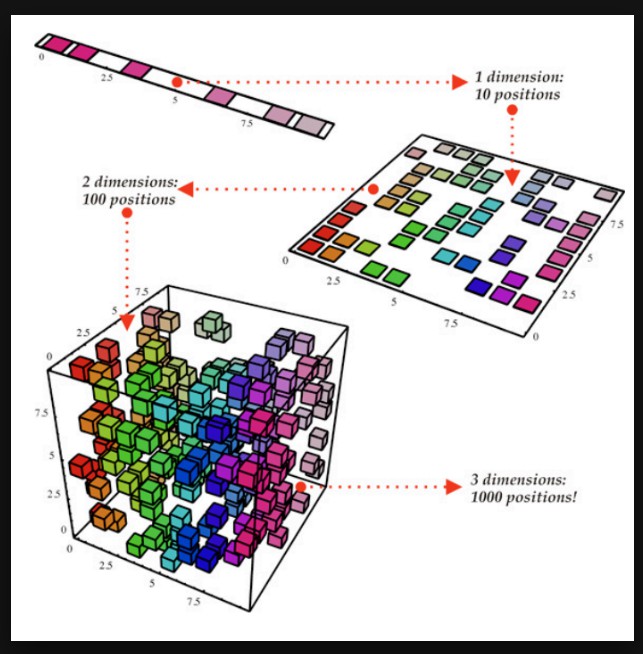
 Muchas variables (**K**) medidas simultáneamente

 Pocas muestras (**N**) analizadas

La mayoría de ellos son de alto rendimiento

# Una aproximación (más) adecuada

Con datos ómicos es habitual aplicar métodos denominados de *proyección* o también de *reducción de la dimensión*

Presentan ventajas interesantes

 Examinan TODAS las variables juntas

 Evita la pérdida de información

 Encuentra tendencias *subyacentes*

= "variables latentes"

# ¿Qué es una proyección?

Variables = ejes en un espacio multidimensional

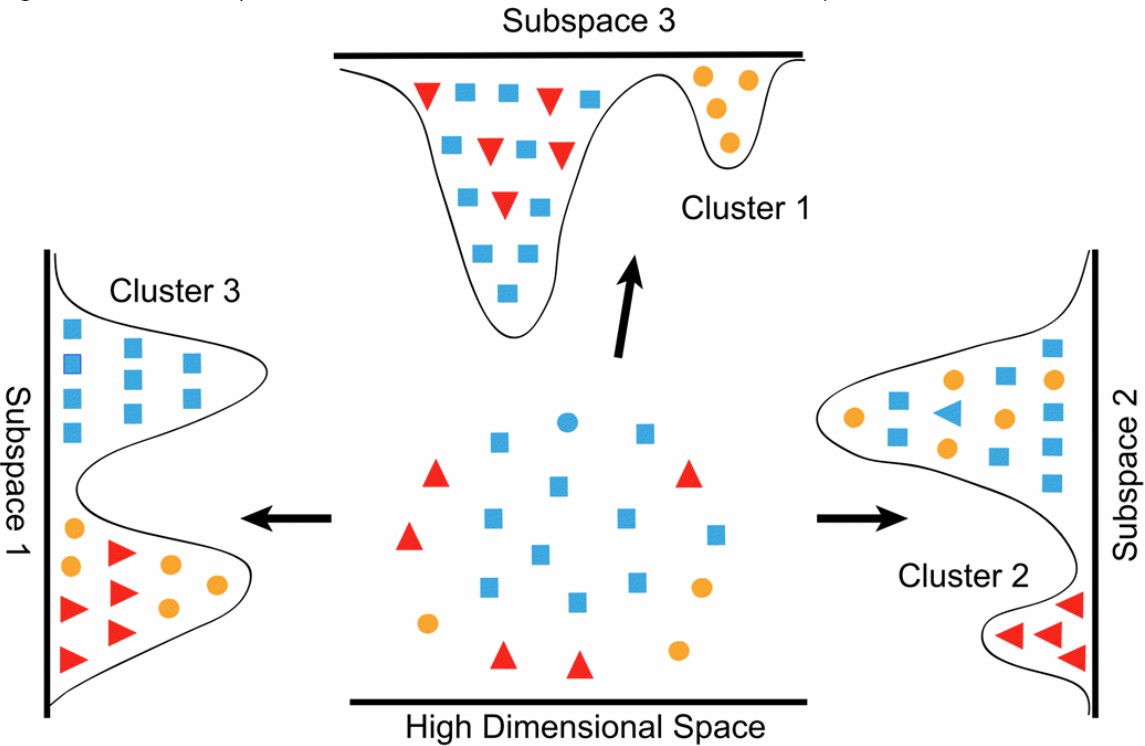
Observaciones = puntos en el espacio multidimensional

Proyectando puntos en (hiper)planos

*puede ser posible*

 Simplificar/Mejorar la visualización

 Resaltar agrupaciones inherentes o naturales de los datos



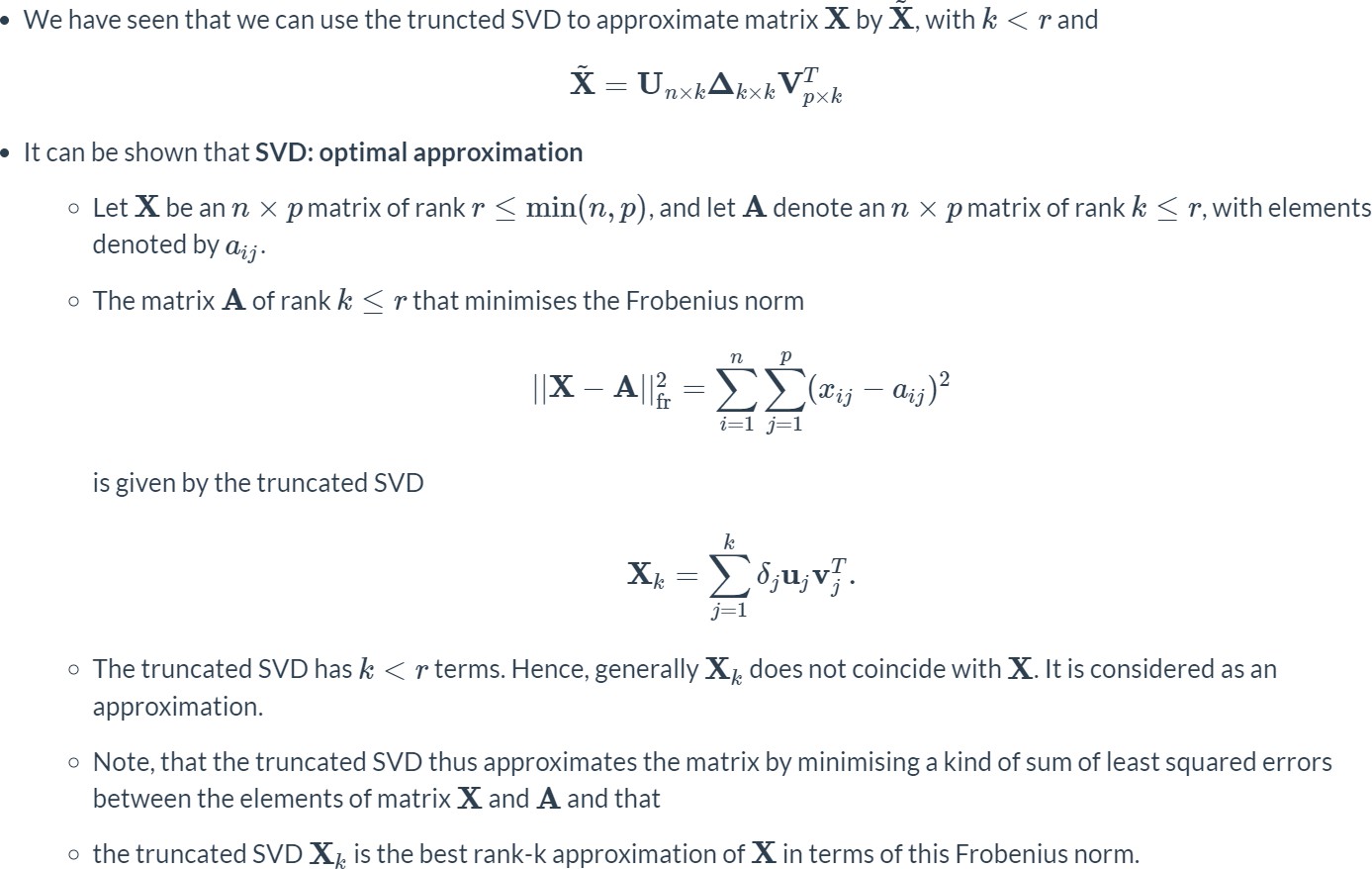
[Data Science post: What is dimensionality](https://datascience.stackexchange.com/questions/130/what-is-dimensionality-reduction-what-is-the-difference-between-feature-selecti) [reductio](https://datascience.stackexchange.com/questions/130/what-is-dimensionality-reduction-what-is-the-difference-between-feature-selecti)

 Destacar la relación entre estos grupos.

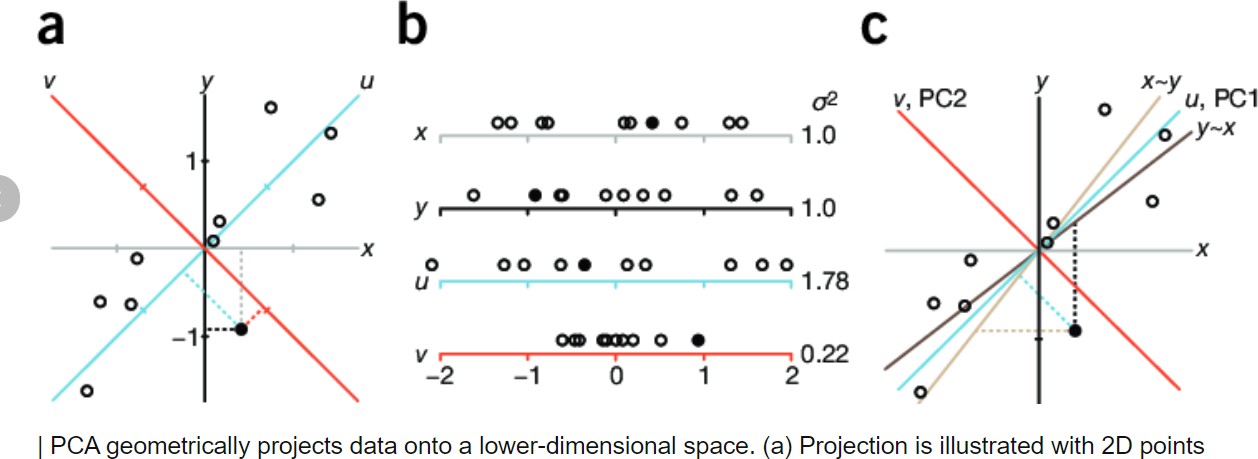
 Detectar problemas como efectos de lote o valores atípicos

7 / 45

# Enfoque geométrico y algebráico

Las proyecciones pueden siempre considerarse desde un punto de vista algebráico o geométrico

 Algebraicamente: La información de las observaciones se resume en algunas variables nuevas (latentes)

 Geométricamente: El conjunto de puntos en un espacio dimensional K (K = número de variables) se *aproxima* mediante un (hiper)plano, es decir un espacio de menor dimensión en donde se proyectan los puntos.

# Organización de los contenidos

El análisis exploratorio inicial

 Descripción y posibles transformaciones

Los métodos de proyección/reducción de dimensiones o de descomposicion matricial.

 Nos centraremos en Análisis de componentes principales (PCA), pero mencionaremos aotras aproximaciones

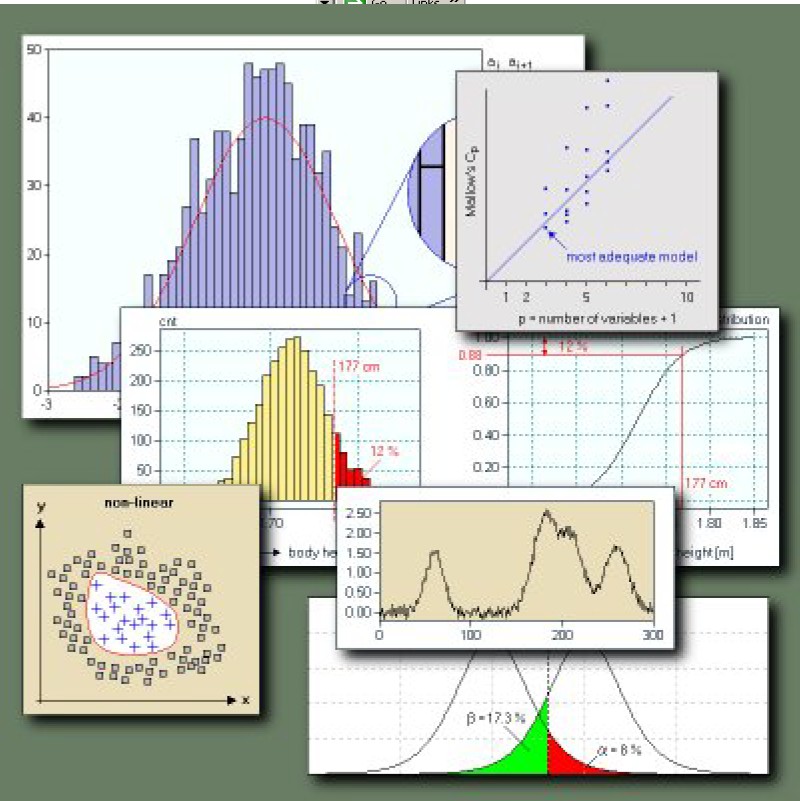
Los métodos de búsqueda de patrones o clases

 Análisis de conglomerados y metodos aglomerativos (KMeans/PAM)

# Exploración de datos multivariantes

10 / 45

# I. La descripción de los datos

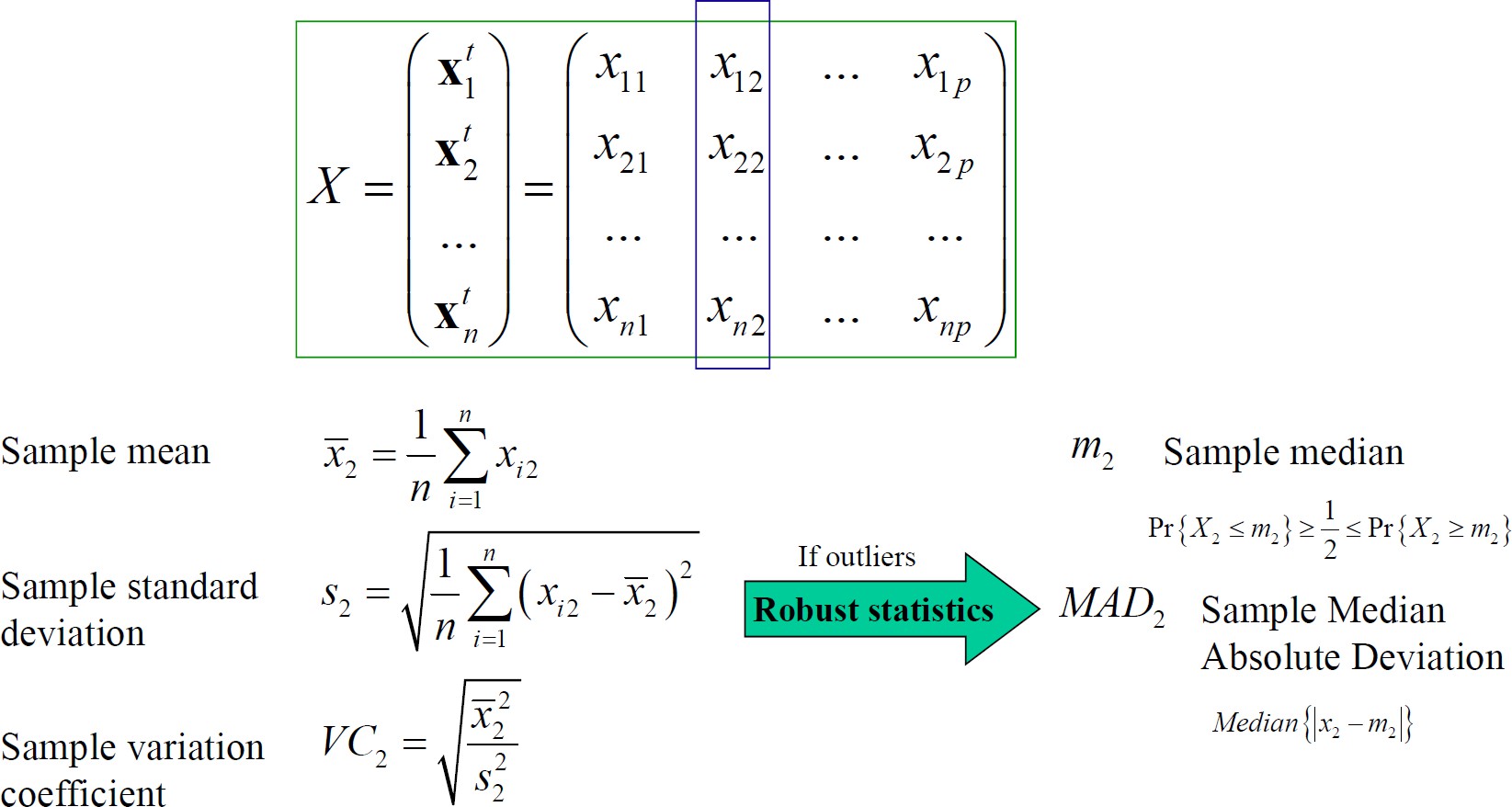
Comenzaremos cualquier análisis estadístico observando los datos

 Cuáles/Cuántas variables,  Cuáles/Cuántas muestras  ¿Faltan valores?

 Obtener estadísticas y gráficos de resumen simples

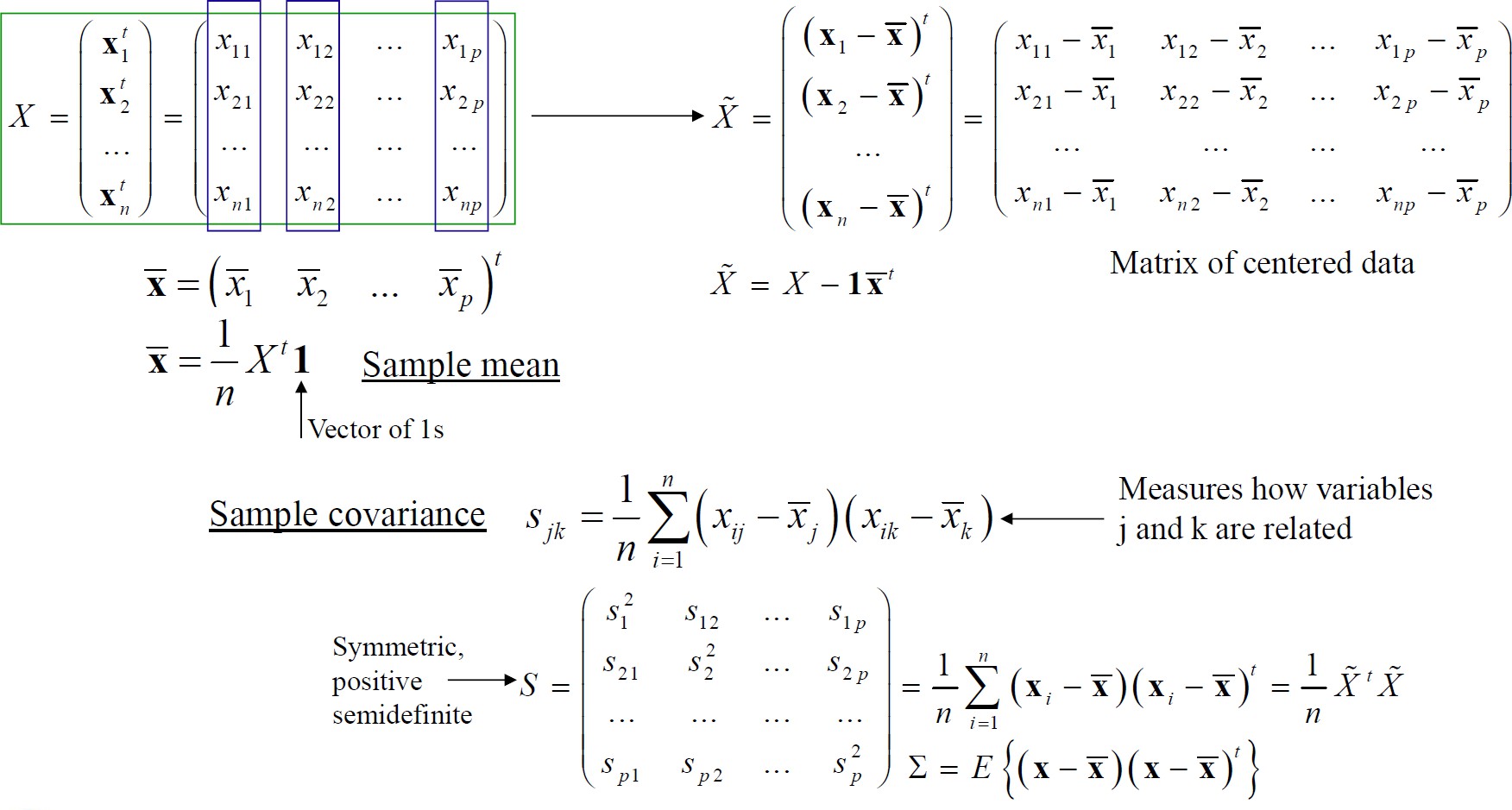
Podemos calcular los estadísticos descriptivos habituales, empezando por medidas univariantes y progresando a estadísticos bivariantes o multivariantes

# Resúmenes numéricos univariantes



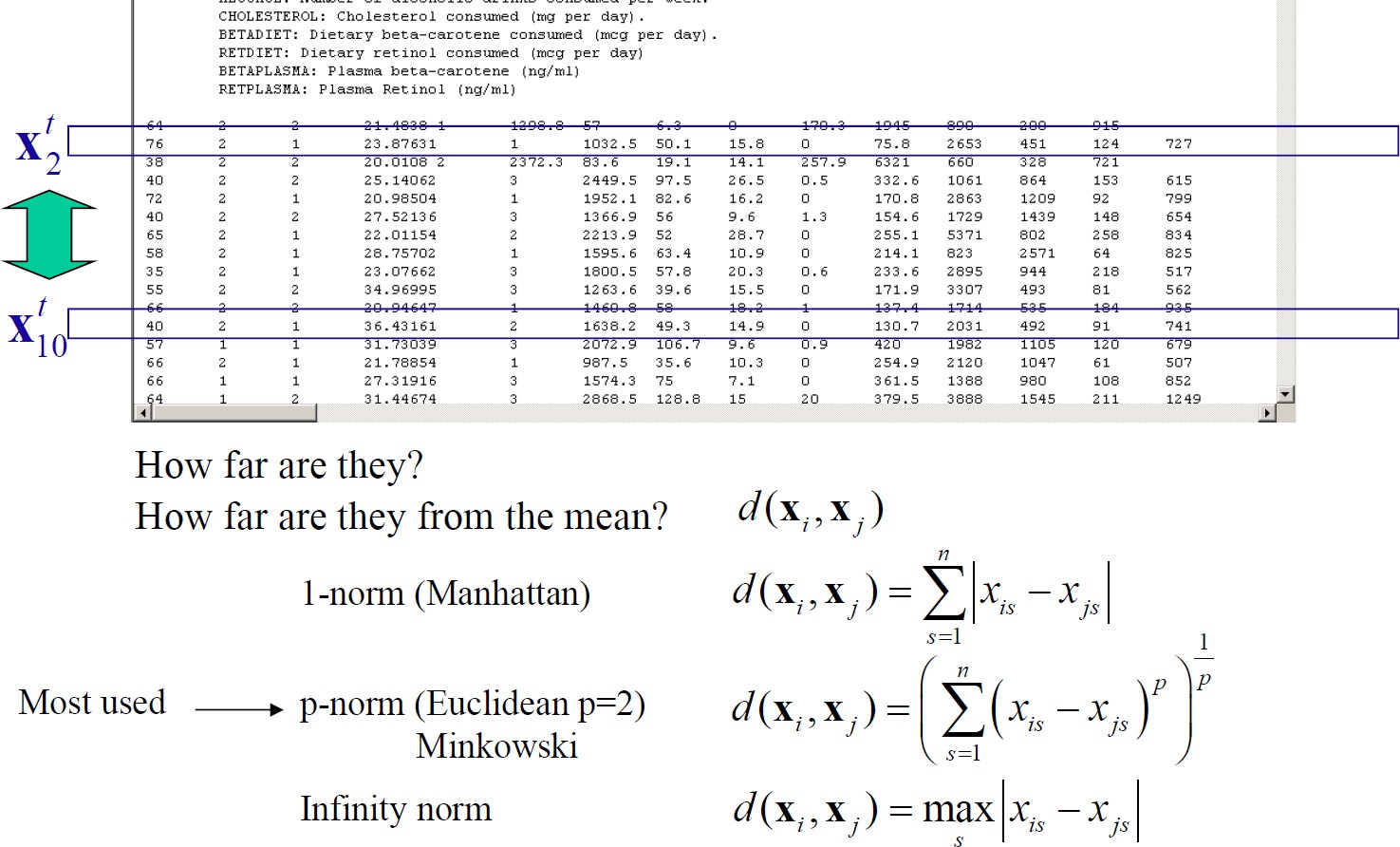
[Fuente: Curso de Analisis Multivariante de Datos. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Resúmenes numéricos bivariantes



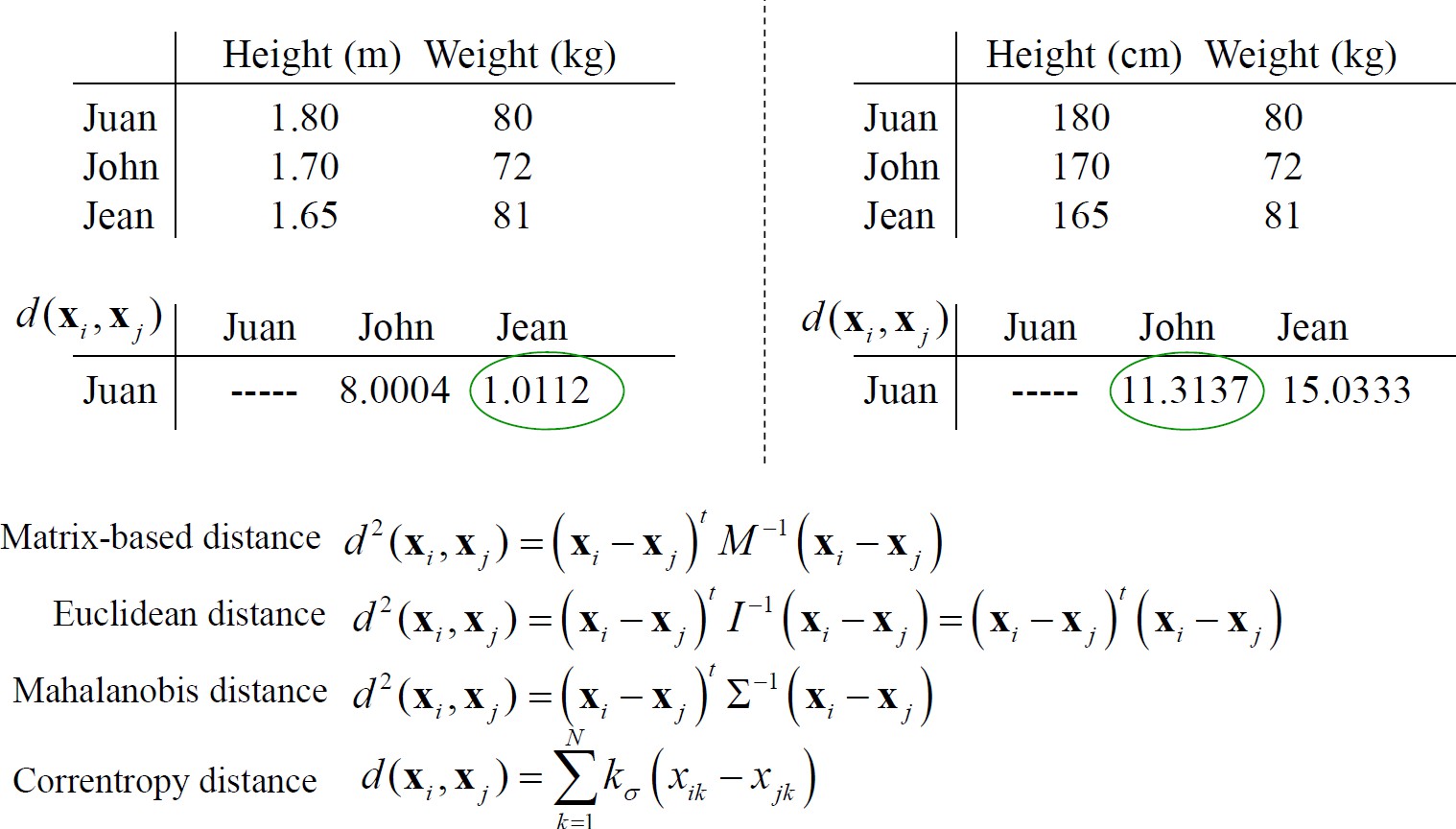
[Fuente: Curso de Analisis Multivariante de Datos. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Distancias multivariantes (1)



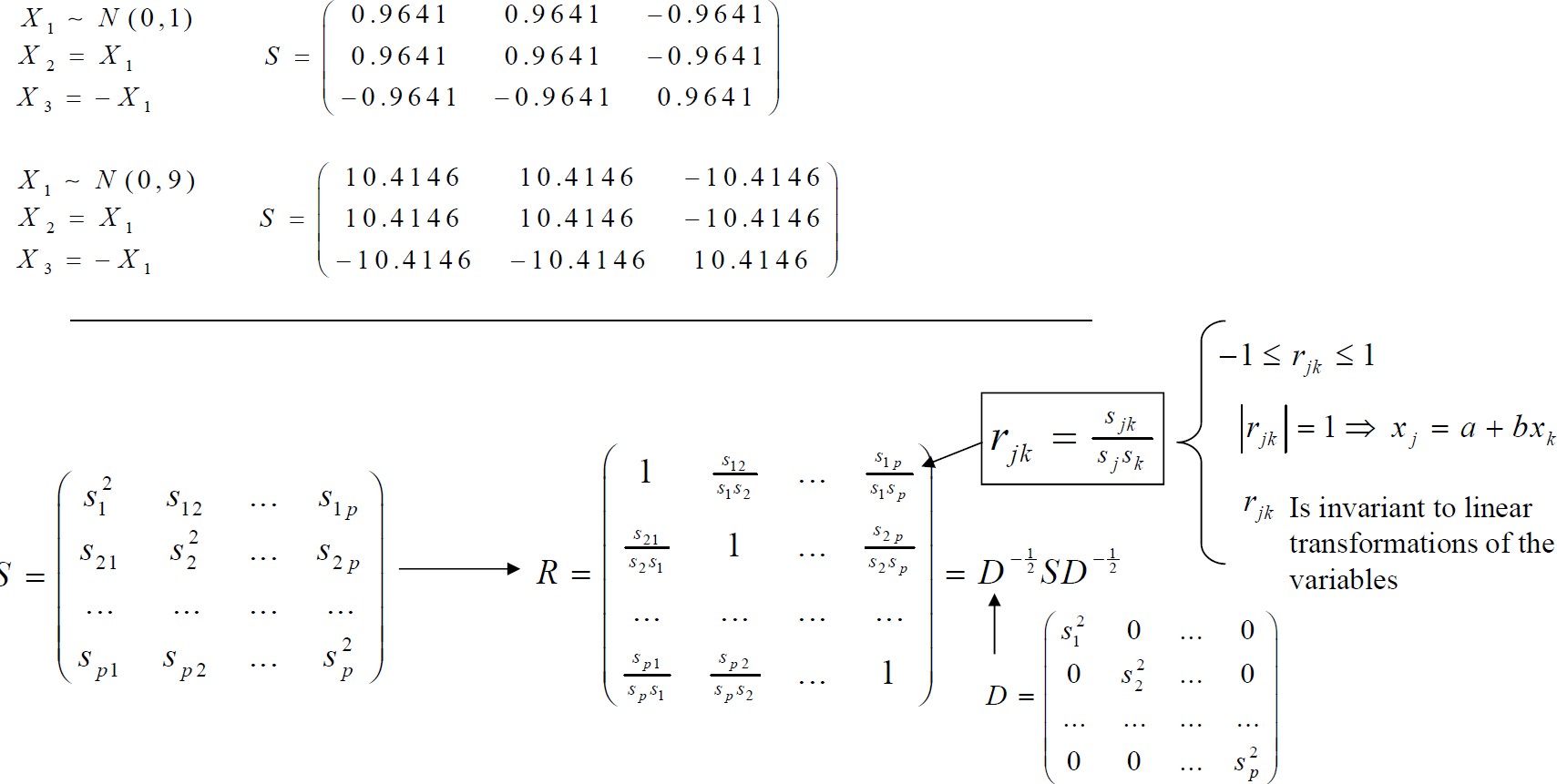
[Fuente: Curso de Analisis Multivariante de Datos. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Distancias multivariantes (2)



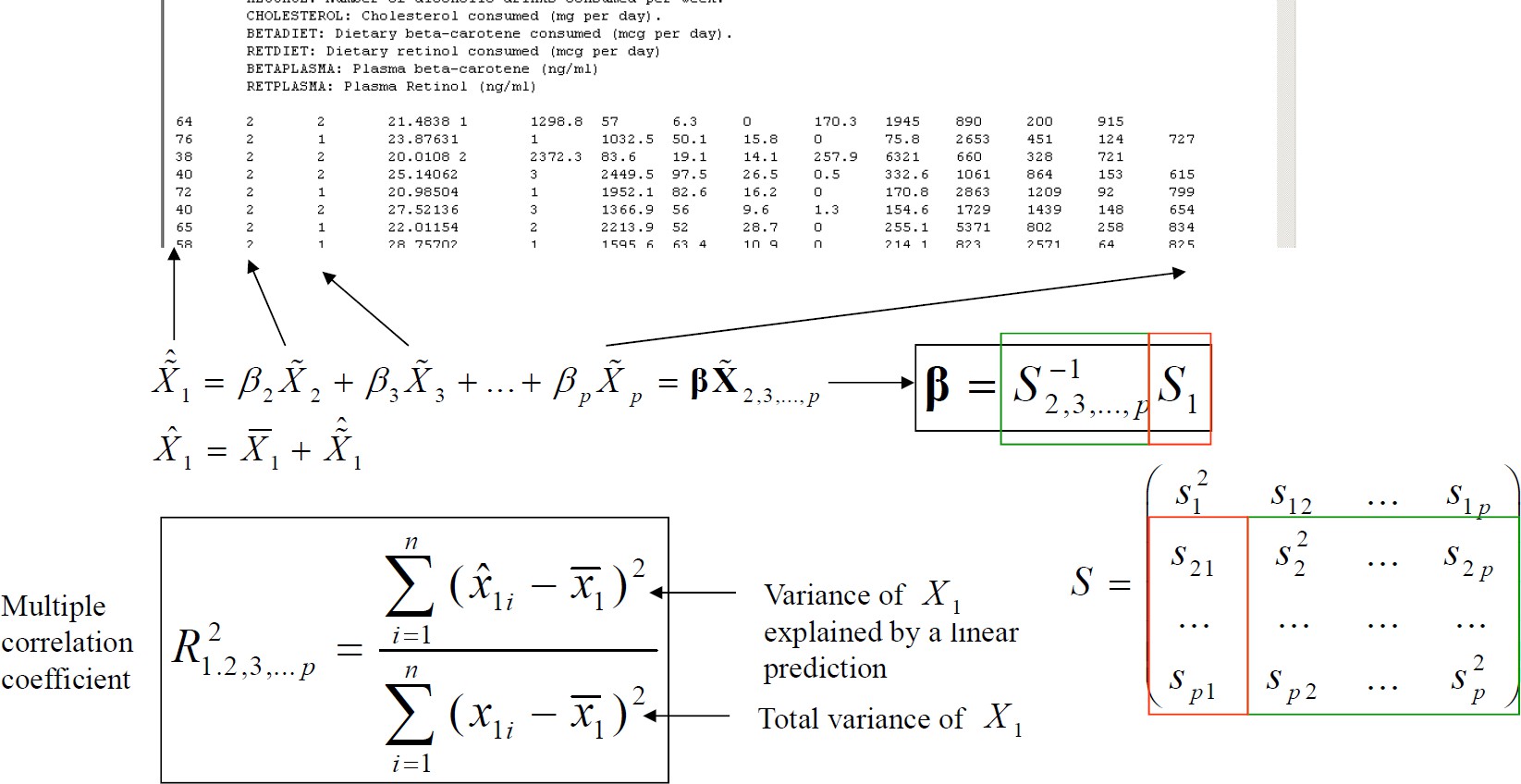
[Fuente: Curso de Analisis Multivariante de Datos. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Dependencia bivariante



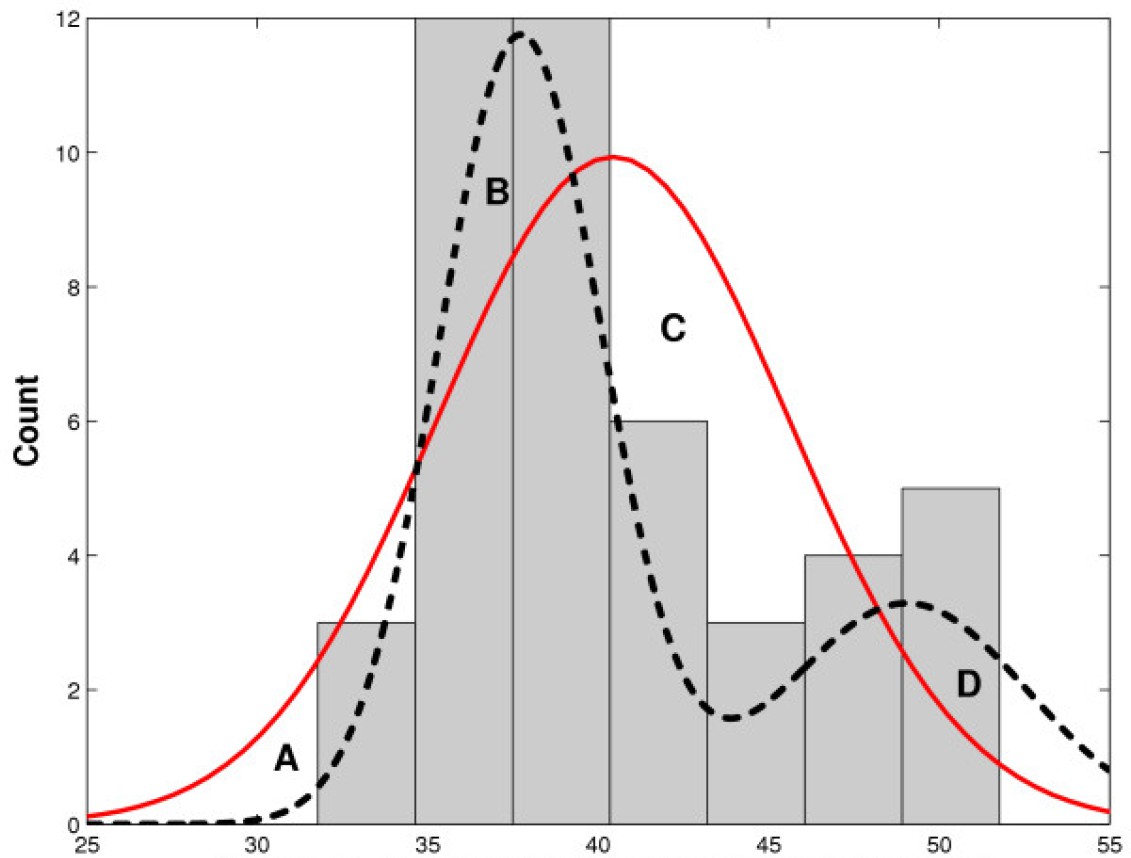
[Fuente: Curso de Analis is Multivariante de Datos. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Dependencia multivariante



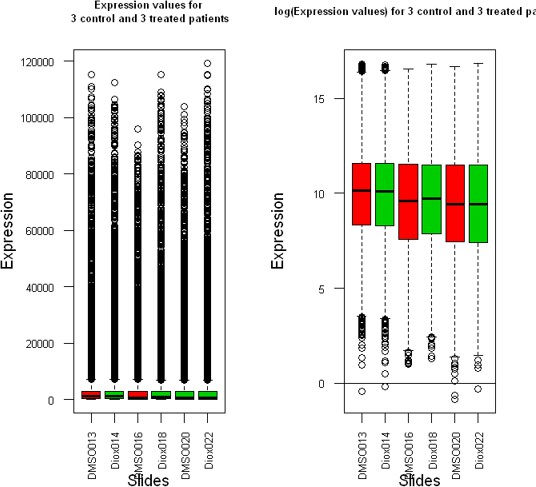
[Fuente: Curso de Analisis Multivariante de Datos. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# La visualización de los datos

Es posible visualizar las variables una a una

 Histograma, Boxplts, Violin plots, diagramas de barras

Pero lo habitual es empezar por gráficos en 2 o 3 dimensiones

 Boxplot múltiple, Scatterplots, MosaicplotsLos gráficos suelen ser 2D o 3D

 Los mapas de calor o *heatmaps*

son muy populares

18 / 45

# Visualizando múltiples variables

Para trabajar con tres o más variables existen diferentes alternativas

 Obviamente una es empezar por lo anterior y representar las variables ighnorando su relación o estudiarla por pares

 Otra alternativa es trabajar en dimensiones reducidas, es decir representar por ejemplo 2 a 2 o 3 a 3 algunas de las primeras components.

Esta es la aproximación que se discute en la sección siguiente, dedicada a la reducción de la dimensión.

19 / 45

# Ejemplo de exploración de datos

Este enlace [(Enlace al workflow)](https://bioconductor.org/packages/release/bioc/vignettes/POMA/inst/doc/POMA-demo.html) contiene un workflow completo de unos datos de metabolómica usando el paquete POMA de Bioconductor.



20 /

 Los primeros pasos contienen muestran aspectos como la exploración de los datos, las transformaciones y la detección e imputación de valores faltantes.

 Obsevése el uso de

, una

SummarizedExperiment

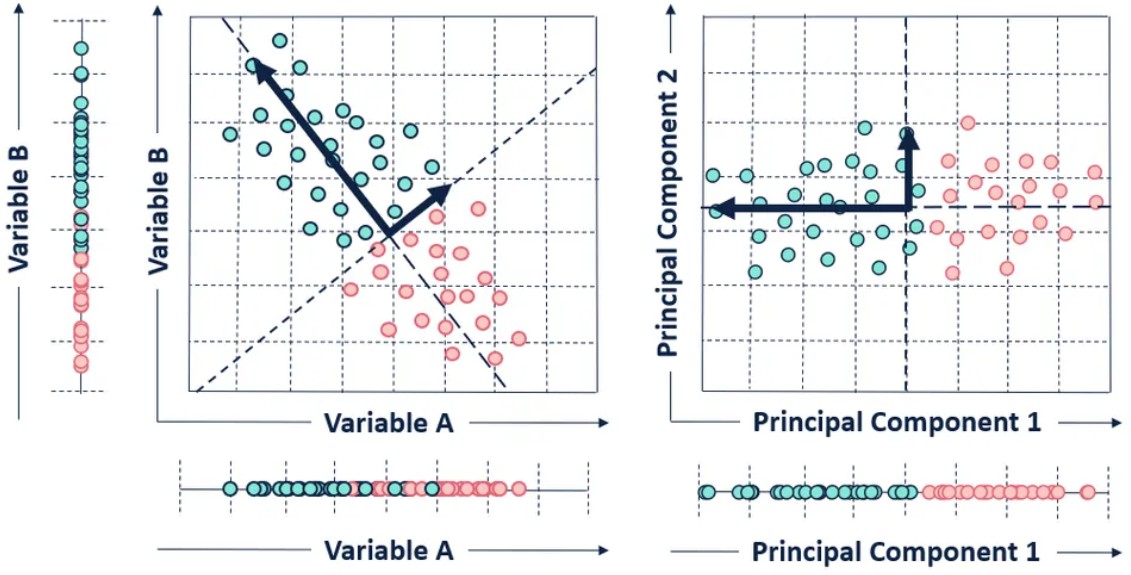
generalización de la clase ExpressionSet, que se pensó originalmente para microarrays.

45

# Reducción de la dimensión

21 / 45

# Análisis de Componentes Principales

Dada una matriz de ***datos KxN*** que contiene ***K*** variables (mediciones *probablemente* correlacionadas) en ***N*** muestras (objetos/individuos...)

El PCA lleva a cabo una transformación de las variables en ***K*** nuevas componentes que

 Reflejan las diferentes fuentes de variabilidad de los datos, pero

 no están correlacionadas, es decir, *cada componente representa una fuente diferente de variabilidad,*

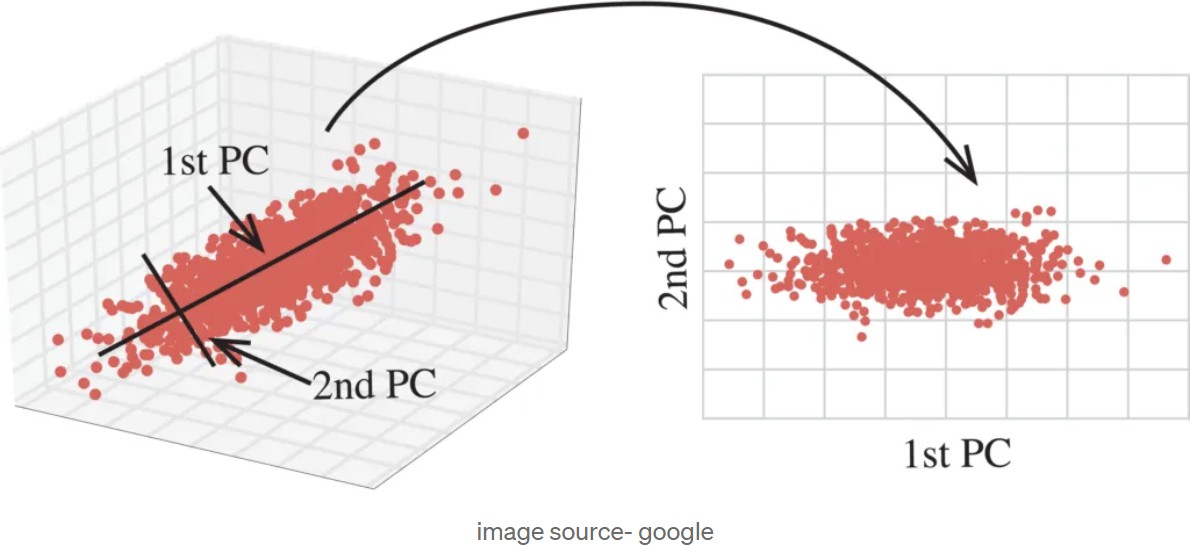
[Fuente](https://towardsdatascience.com/tidying-up-with-pca-an-introduction-to-principal-components-analysis-f876599af383)

# Diseñados para mejorar

Estas nuevas componentes se construyen de tal forma que *tienen una capacidad explicativa decreciente*: cada componente explica más que la siguiente (menos que la anterior).

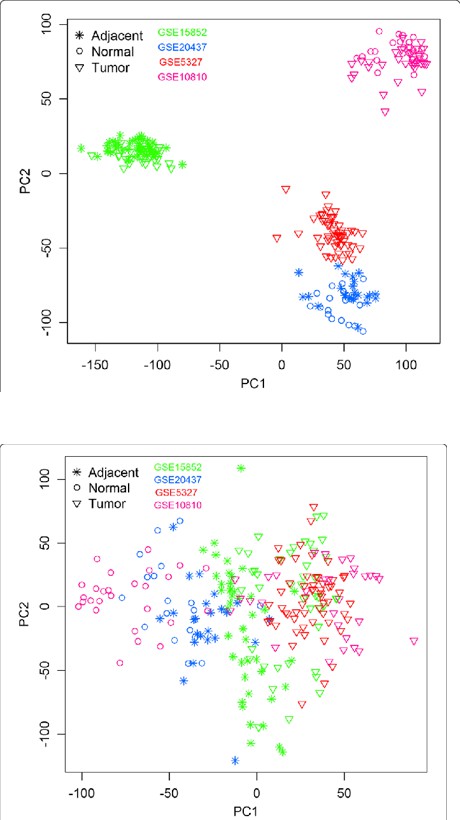
 Esto *puede permitir* quedarse con las primeras componentes y descartar el resto.

 Es decir es posible utilizar los valores de las primeras componentes principales para obtener una representación de los datos *en dimensión reducida*, es decir inferior a la original



[En este ejemplo](https://medium.com/alliedoffsets/unveiling-the-power-of-pca-turbocharge-your-data-science-with-dimensionality-reduction-fa9d1fc6e510) pasamos de tres a dos dimensiones sin casi perder calidad de la representación

# Revelando información latente

En general se considera que estas transformaciones

 sirven para revelar los patrones dominantes y las principales tendencias en los datos

 a la vez que permiten visualizar agrupaciones naturales

o erróneas como las debidas a efecto batch.

La imagen muestra el grafico de las dos primeras componentes antes (arriba) y despues )abajo) de ajustar para eliminar un efecto batch

# ¿Cómo funciona el PCA?

Supongamos una matriz de datos 2xN de dos variables correlacionadas.

Al estar los datos correlacionados, es difícil separar cada fuente de variabilidad

Si K fuera mucho más alto, sería aún más difícil.

Para obtener las nuevas bases se plantea el problema matemático de optimización consistente en *buscar, de forma secuencial, una transfomación de la matriz original en la que los vectores resultantes sean ortogonales dos a dos*, es decir

 Las nuevas variables (nuevos ejes de coordenadas) son independientes las unas de las otras.

 En este caso la matriz de varianzas covarianzas será diagonal, puesto que la covarianza de variables independientes es cero.

# Calculando las PCs

Bajo condicones generales se ve que esto se obtendrá *diagonalizando* la matriz de varianzas-covarianzas de los datos centrados cálculo éste que resulta en

 Un vector de valores propios (VAPs) proporcionales a la varianza de cada variable en las nuevas coordenadas

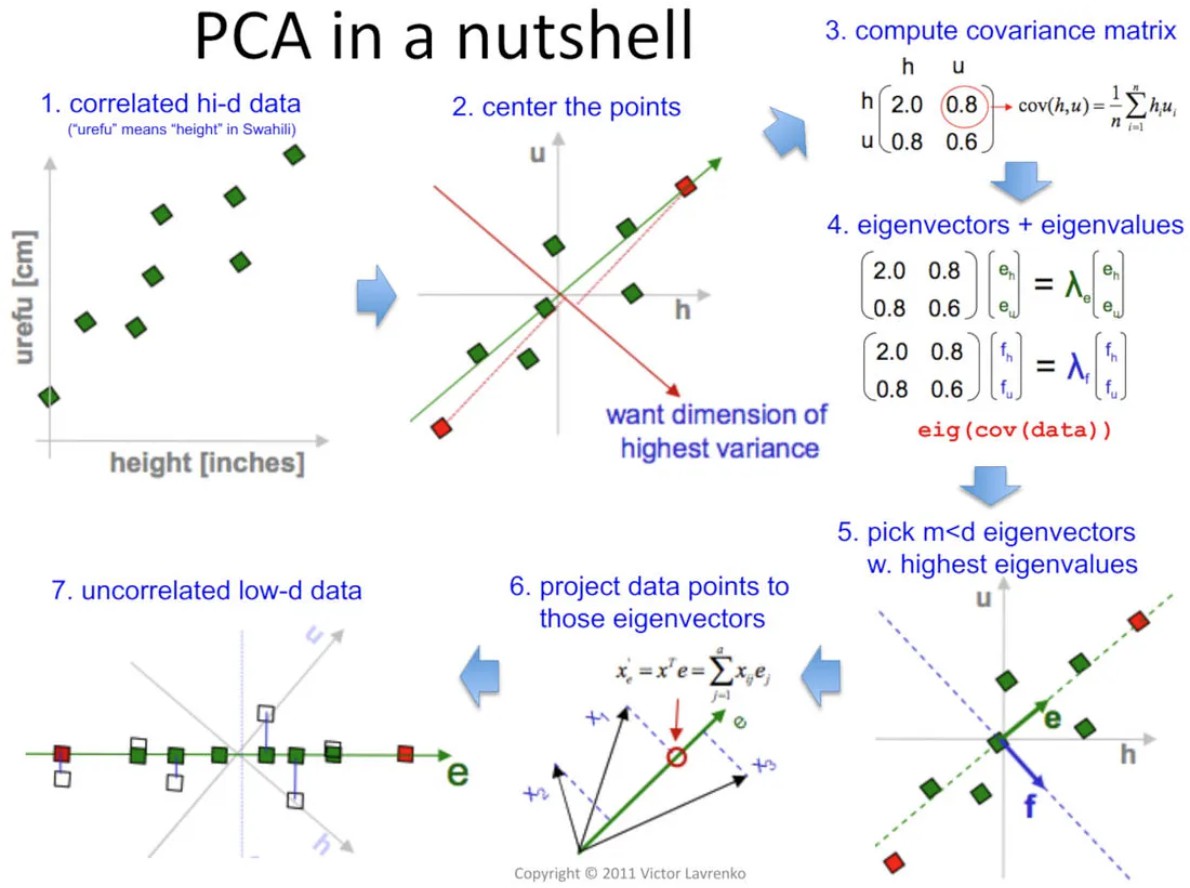
 Una matriz de vectores propios (VEPs) que son, precisamente, las puntuaciones o

*scores* de las observaciones en estas nuevas coordenadas.

Sin entrar en los detalles matemáticos, obsérvese que al transformar la matriz de varianzas en diagonal se consigue que las covarianzas de las nuevas variables pasen a valer 0, es decir se obtienen variables independientes.

Además los valores propios (la diagonal de la nueva matriz) quedan ordenados de mayor a manor, lo que corresponde a la idea de uqe las nuevas variables tiene, cada una una variabilidad mayor que la siguiente, y por lo tanto explica un porcentaje mayor que las siguientes.

# Obtención de las PCs



Ejemplo con R

Los dos capítulos del texto abierto "[Applied Multivariate Statistics with R](https://uw.pressbooks.pub/appliedmultivariatestatistics/)" presentan, de forma muy asequible.

 [La diagonalización de una matriz ("EigenAnalysis")](https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/eigenanalysis/)

 [Como esto se aplica en el PCA](https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/pca/)

Puede verse también un ejemplo en el primer caso del recopilatorio [Casos y ejemplos](https://aspteaching.github.io/AMVCasos/) [de Análisis multivariante con R](https://aspteaching.github.io/AMVCasos/)

 [Ejemplo PCA-1: Búsqueda de factores latentes en datos ecológicos](https://aspteaching.github.io/AMVCasos/#ejemplo-pca-1-b%C3%BAsqueda-de-factores-latentes-en-datos-ecol%C3%B3gicos1)

# Interpretación de las PCs

La primera componente principal es

 Una combinación lineal (CL) de las variables originales

 Que va en la dirección de mayor variabilidad en los datos  Explica la *cantidad máxima de variación en los datos*

La 2º y sucesivas PC también es una CL de todas las variables originales aunque con otros coeficientes calculados de forma que:

 Se alinea con la *siguiente dirección de mayor variabilidad*, ***ortogonalmente*** a las PCs anteriores.

 Explican la *cantidad máxima de variación restante*

La interpretación de las nuevas componentes puede resultar clara, correspondiendo a dimensiones biológicas reales o, en ocasiones, no serlo en absoluto.

 Para facilitarla, los programas que calculan el PCA calculan también la correlación entre las nuevas componentes y las variables originales.

# Representación en el espacio PCA

Las puntuaciones o *scores* son las nuevas coordenadas en el sistema ortogonal definido por los PC y se han obtenido para que cada PC explique la *cantidad máxima de* ***variación restante*** *en los datos*

Esto significa que no es necesario utilizar todos los PC para visualizar los datos en este nuevo sistema de coordenadas, por lo que puede realizarse una visualizacón en una dimensión inferior.]

 A menudo, tomando las primeras PCs explicará un alto porcentaje de variabilidad.

 Por lo general, solo se toman los primeros 2 o 3, pero

Siempre debe verificarse que ***con estos pocos es bastante***

 El gráfico denominado "scree plot" junto con un criterio de "cambio de pendiente" o del codo, suelen resultar útiles.

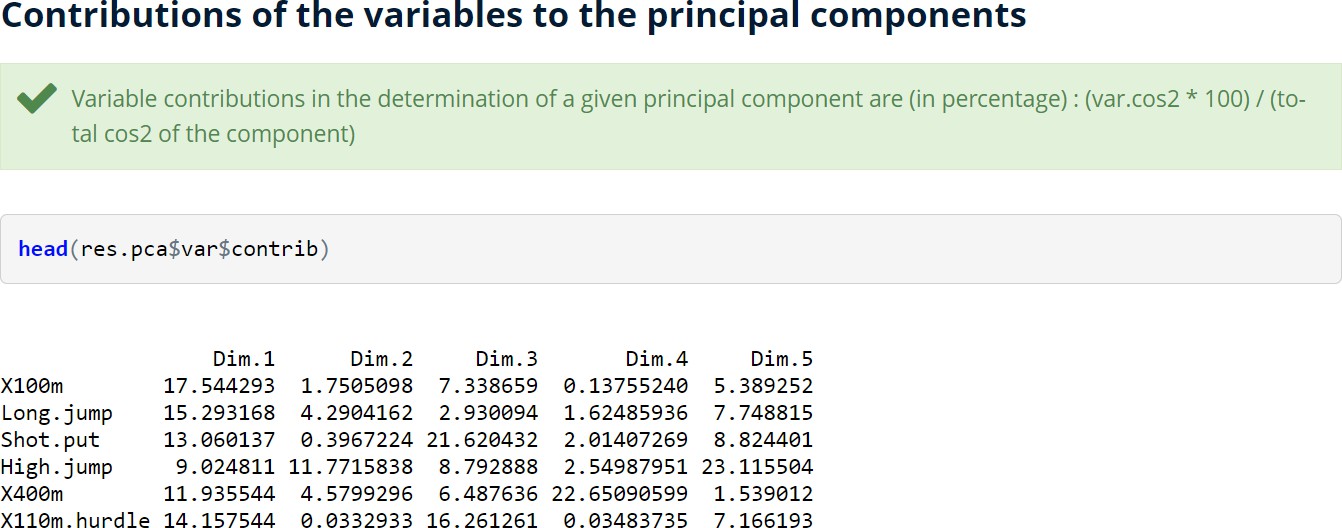
# Interpretación de las PCs

Las PCs pueden interpretarse observando cuáles de las variables originales contribuyen más a su variabilidad

 Cuanto más se correlaciona una variable con un PC, mayor es su influencia.

La magnitud de la contribución de cada variable a una PC és su *carga ("loading") en dicha PC*

 Las cargas son los *cosenos del ángulo entre las variables y los PC*

31 / 45

# El PCA en resumen

EL PCA realiza una transformación en un nuevo conjunto de coordenadas ortogonales que se denominan coordenadas principales.

Estas nuevas coordenadas (PCs)

 explican la variabilidad observada, con capacidad decreciente (la mayor la de la 1era PC, la menor, la de la última PC).

 Y, por el hecho de ser ortogonales cada PC explica aspectos distintos de los datos.

El PCA puede verse como un método para desentrañar tendencias o patrones subyacentes en los datos

 El concepto de *variable latente* se relaciona con ella

El PCA puede proporcionar una visión general de los datos que se van a revelar

 Variables dominantes  Tendencias

 Patrones como valores atípicos, grupos, clústeres

32 / 45

# Descubriendo grupos en los datos

33 / 45

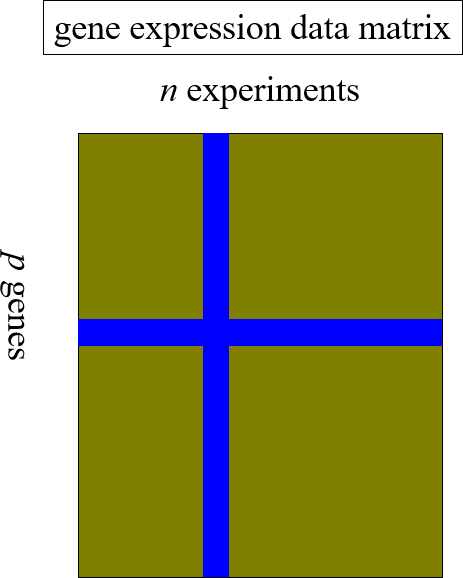
# Recordemos: los datos ómicos

En un experimento de datos ómicos, una vez preprocesados los datos, y es habitual tenerlos resumidos en una matriz de datos ómicos, por ejemplo de expresión génica donde

 Las filas representan **genes o más genéricamente *features***

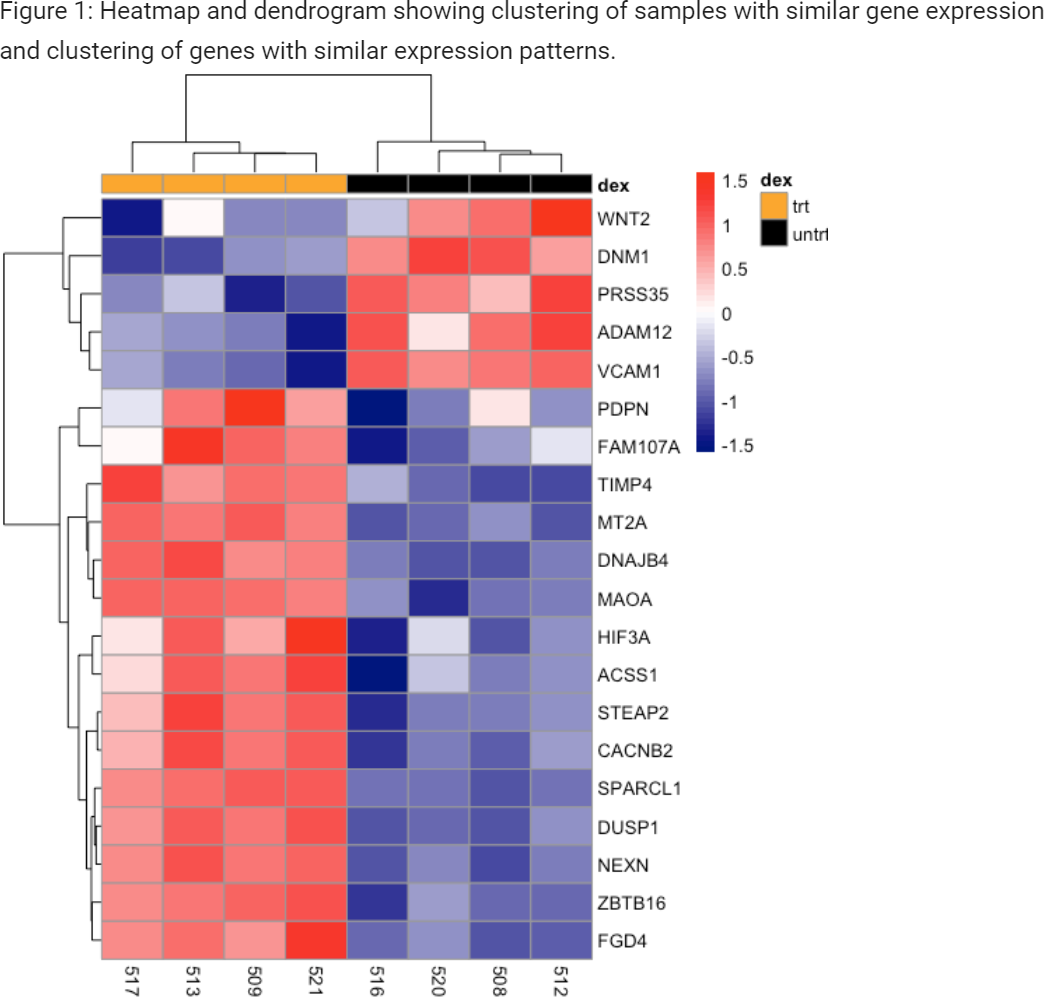
 Las columnas representan **individuos, muestras u observaciones**

Las observaciones pueden corresponder o no a condiciones experimentales distintas (e.g. tratamiento/control) o a grupos de muestras distintos (e.g. varones/hembras).-



34 / 45

# Patrones en los datos

Cuando hablamos de encontrar patrones en los datos podemos referirnos a identificar

 **Grupos de variables relacionadas** que muestran *cambios coordinados* entre condiciones.

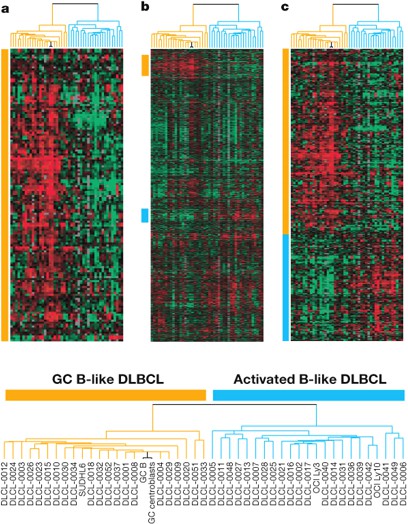
 Esto puede sugerir co- expresión, co-regulación, etc

 **Grupos de muestras** que muestran patrones de expresión similares

Esto puede sugerir la existencia de (sub-)clases no identificadas *previamente* entre las muestras

35 / 45

# Beneficios de la agrupación

Es habitual que el proceso de clustering conduzca a resultados interpretables.

Resulta especialmente útil

 Para identificar patrones en el tiempo o en el espacio.

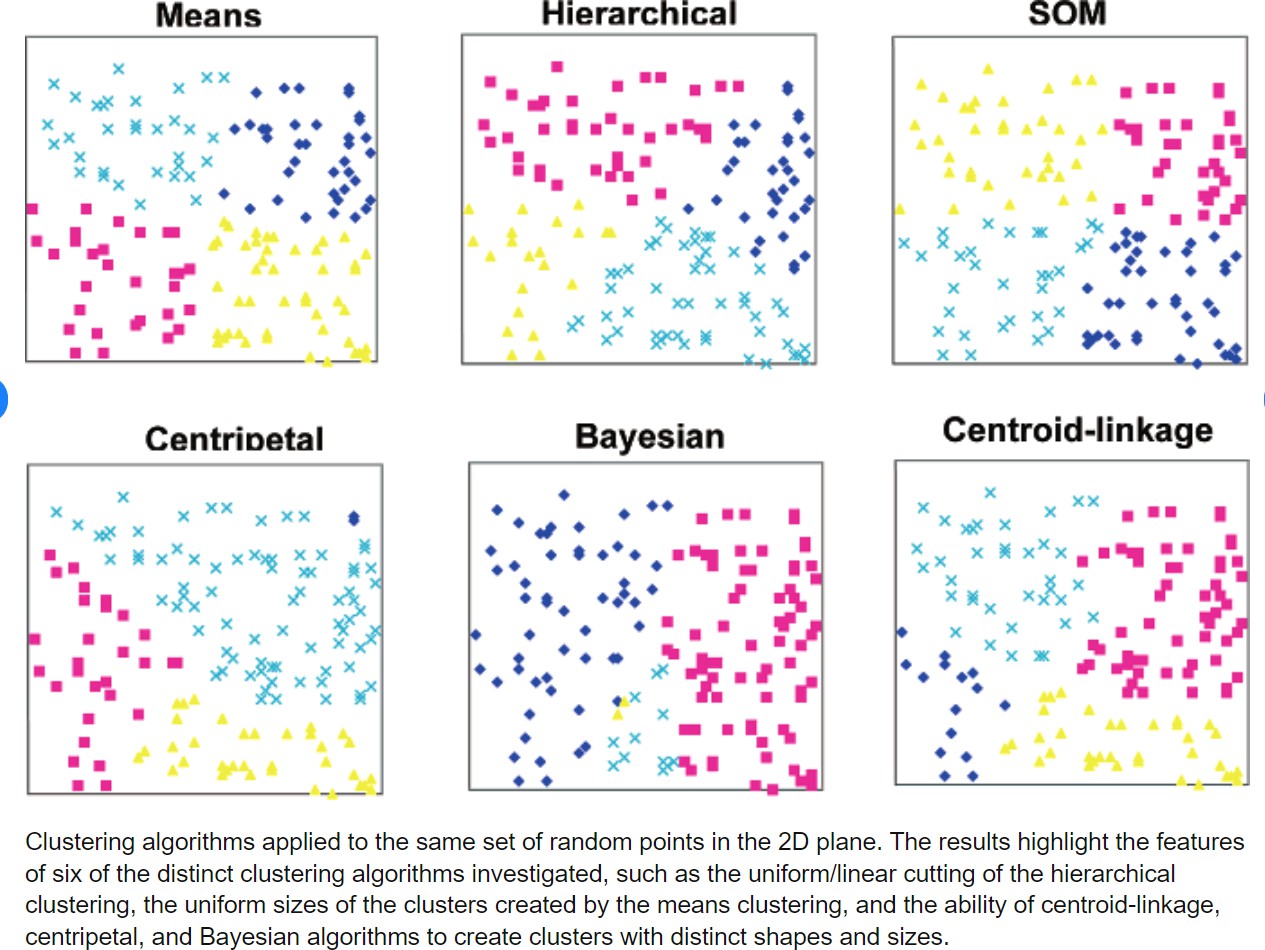
 Cuando se buscan nuevas subclases de muestras celulares (tumores, etc.).

La agrupación refuerza la señal cuando se toman promedios dentro de grupos de genes (Eisen).

[Fuente](https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r/Lesson5_intro_to_ggplot/)

36 / 45

Inconvenientes del *clustering*

Es una técnica exploratoria

Por lo general, no se dispone de significación.

Diferentes enfoques a menudo producen diferentes agrupaciones

Difícil decidir cuál es el mejor (o el "real")

Difícil evitar la tentación de seleccionar la agrupación que mejor se ajuste a nuestras hipótesis.

Cualquier conjunto de datos se puede agrupar en clústeres

Es difícil decidir si la agrupación es real o aleatoria.

37 / 45

# Las componentes del problema

Ante cualquier problema de este tipo debemos tomar distintas deciones

 ¿Qué medida de similitud o disimilitud?

 ¿Qué algoritmo de agrupación usar?

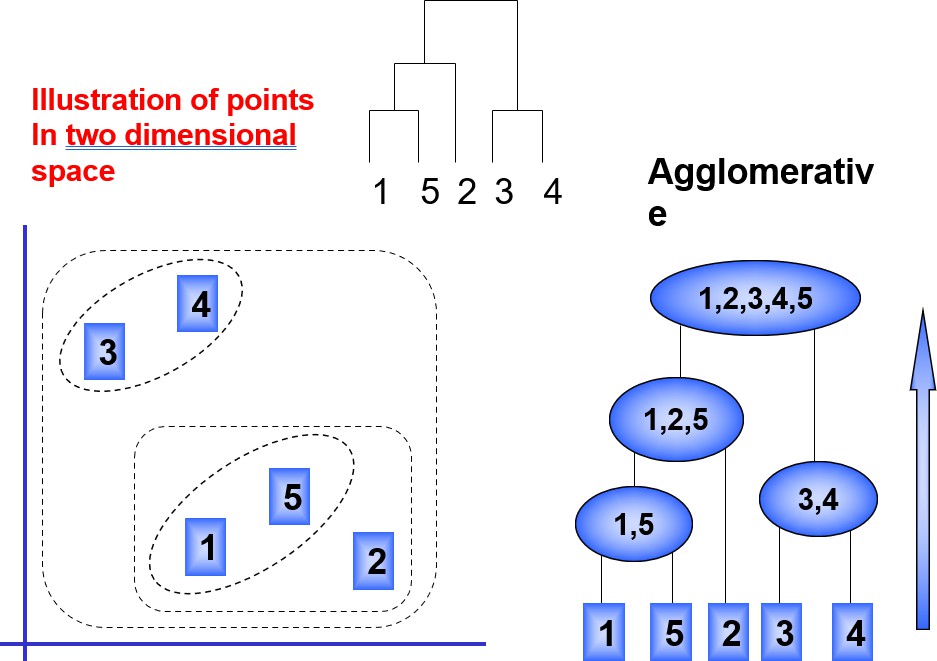
A menudo empezamos por algoritmos básicos para Y continuamos con otros más elaborados

 ¿Como se decidirá el número de grupos?

 ¿Qué indivíduos y que variables utilizar

Suele ser interesante filtar para eliminar variables redundantes y poco informativas, así como datos defectuosos (outliers o con mucho smissings)

# Cluster jerárquico

Los métodos de agrupamiento jerárquico producen un árbol o dendrograma.

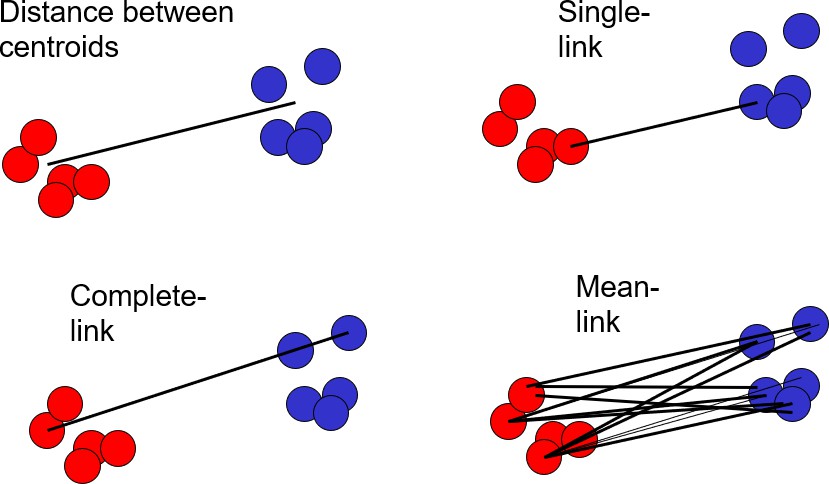
Evitan especificar cuántos clústeres son apropiados proporcionando una partición para cada *k* obtenido de cortar el árbol en algún nivel.

El árbol se puede construir de dos maneras distintas

 de abajo hacia arriba: agrupación aglomerada;

 De arriba hacia abajo: agrupación divisoria.

# Distancias para la agrupación jerárquica

Se basan en las distancias por pares entre los miembros de los clústeres

 **Average link**: promedio de las diferencias entre pares

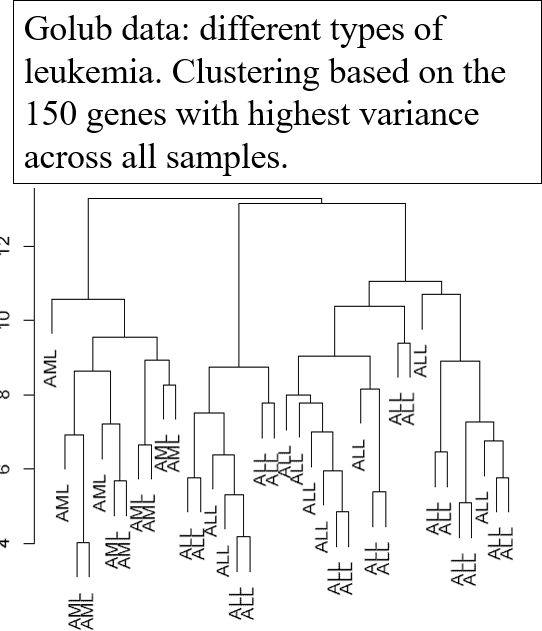
 **Single link**: mínimo de disimilitudes por pares.

 **Complete link**: máximo de disimilitudes por pares.

 **Distancia entre centroides**

El método de distancia elegido influye en los resultados.

# Ejemplo de agrupación jerárquica

En este ejemplo se ilustran algunas caracteristicas del clustering jerarquico

 La altura de un nodo en el dendrograma representa la distancia de los dos clústeres secundarios.

 Hay pérdida de información: n objetos tienen n(n-1)/2 distancias por pares, el árbol tiene n-1 nodos internos.

 El orden de las hojas no está definido únicamente por el dendrograma: 2n-2 opciones posibles.

# Resumen: Metodos de clustering

Proporcionan métodos potente y flexibles para descubrir grupos en los datos y por lo tanto

Son un potente recurso para el análisis exploratorio de datos.

La naturaleza del problema que resuelven, los hace propensos a resultados ambíguos (¿que descubren si no hay clases que descubrir?)

Por no decir que cambios en los métodos y los parametros se acompañan de resultados distintos.

La conclusión es la habitual en estos temas: Debemos usarlestos métodos con cuidado y siempre con una buena comprensión de como operan, validando siempre que sea posible los resultados que nos proporcionan.

# Referencias y recursos

43 / 45

# Referencias y recursos externos

[STHDA Statistical Tools for High Throughput Data Analysis](http://www.sthda.com/english/)

 Esta web contiene multitud de artículos sobre estadística y machine learning agrupados por temas. De hecho el autor la usa para promocionar sus cursos y libros pero la mayoría de materiales son buenos y libres

 En concreto destacamos dos series de artículos sobre los temas que nos ocupan. Para introducirse al tema bastan los primeros capítulos de cada serie, pero todas son relevantes para vuestra formación.

[Cluster Analysis in R. A practical guide](https://www.datanovia.com/en/blog/cluster-analysis-in-r-practical-guide/)

[Principal components methods in R](http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/)

# Referencias y recursos propios

[Casos y ejemplos de análisis multivariante con R](https://aspteaching.github.io/AMVCasos/)

Apuntes de Análisis Multivariante para la asignatura de grado "Diseño de Experimentos y Analisis de Datos" de la facultad de biologia (UB)

 Análisis de components principals

 Anàlisis de cluster

Materiales de análisis multivariante de la asignatura Analisis Multivariante del máster UOC-UB

 Resúmenes de Álgebra lineal