

Análisis de Componentes Principales

Jocelyn Dunstan Escudero

jdunstan@uc.cl

Departamento de Ciencia de la Computación
& Instituto de Matemática Computacional
Pontificia Universidad Católica de Chile

Santiago, Chile



22 de octubre de 2025

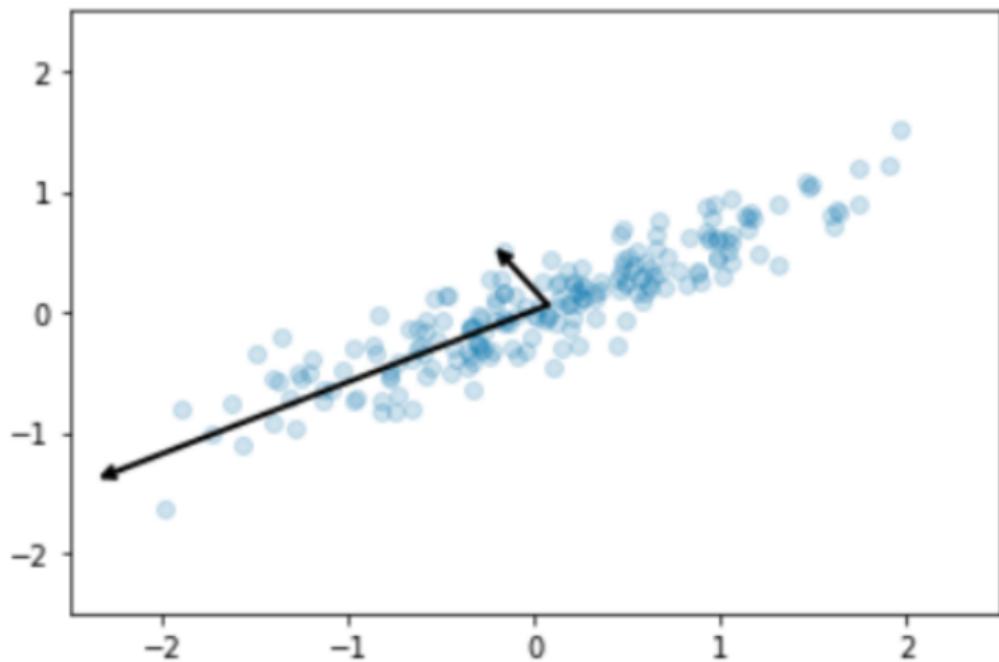
Objetivos

- Entender el análisis de componentes principales para describir un conjunto de datos en términos de nuevas componentes no correlacionadas.
- Conocer algunos *trade-offs* o compensaciones.



Principal component analysis (PCA)





Idea central

- Agrupar los predictores de forma que encontremos nuevos “ejes” en los que se explique mejor la varianza de los datos.



Idea central

- Agrupar los predictores de forma que encontremos nuevos “ejes” en los que se explique mejor la varianza de los datos.
- Pedimos que estos nuevos “ejes” sean perpendiculares entre sí. Encontramos estos “ejes” o componentes como un problema de álgebra lineal (descomposición de valores propios)



Idea central

- Agrupar los predictores de forma que encontremos nuevos “ejes” en los que se explique mejor la varianza de los datos.
- Pedimos que estos nuevos “ejes” sean perpendiculares entre sí. Encontramos estos “ejes” o componentes como un problema de álgebra lineal (descomposición de valores propios)
- El número de componentes principales será igual o menor al número de predictores.



Intuición detrás del cálculo de PCA



Hasta 3 dimensiones podemos visualizar observaciones y tratar de encontrar individuos parecidos, pero qué hacemos si hay más

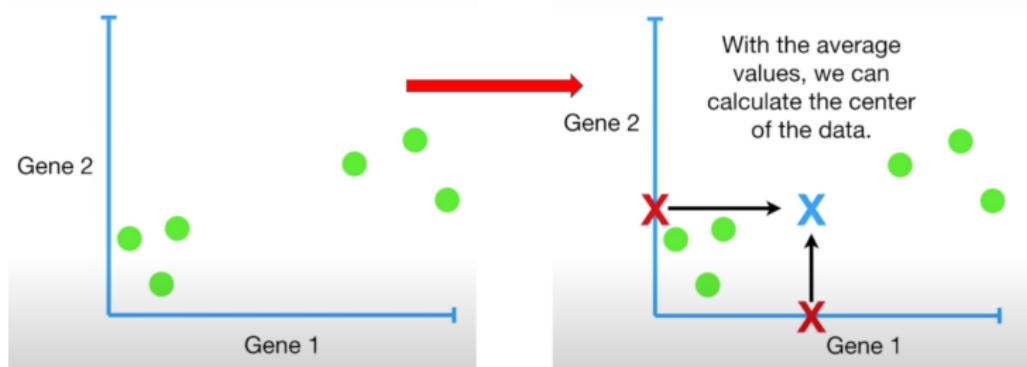
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Primero transformamos los datos de modo que tengan media cero

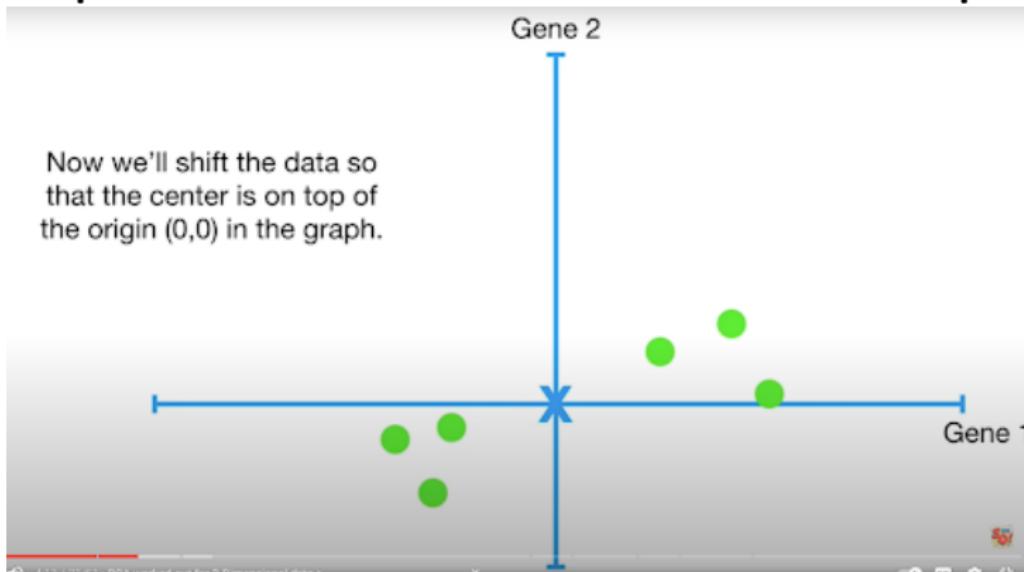


<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Note que esto no cambia la ubicación relativa de los puntos

Now we'll shift the data so that the center is on top of the origin (0,0) in the graph.

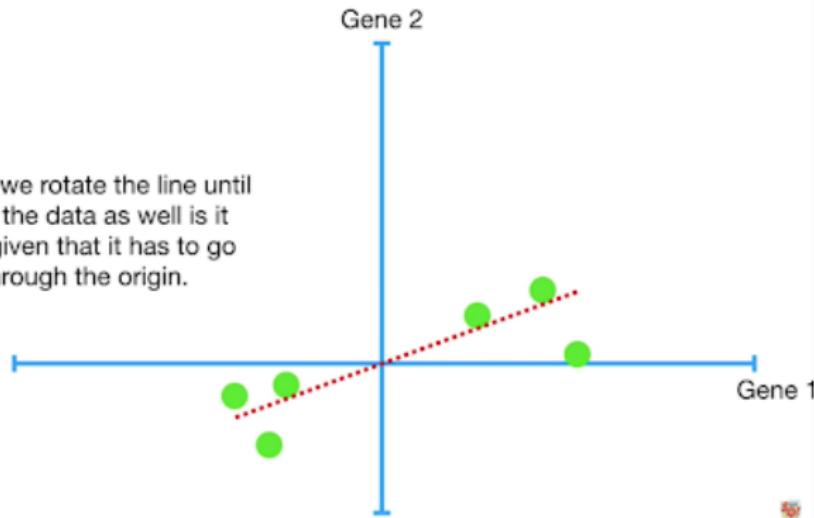


<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Trato de encontrar la recta que mejor aproxima los datos

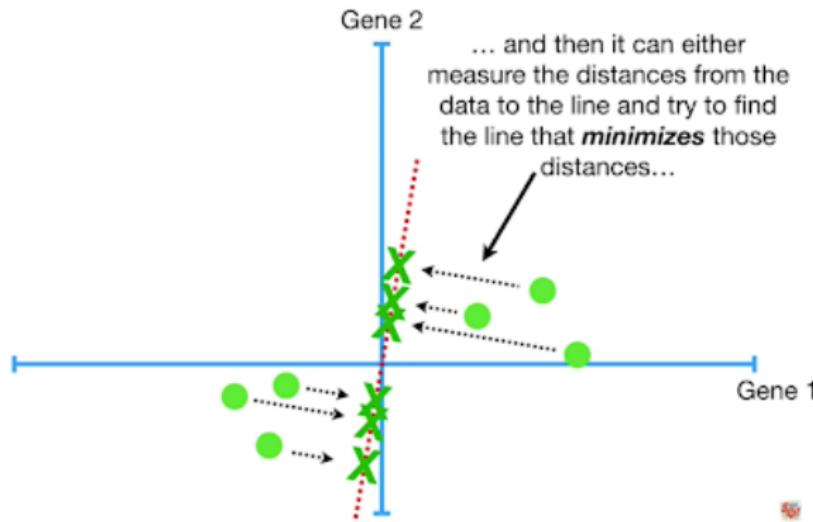
...then we rotate the line until it fits the data as well as it can, given that it has to go through the origin.



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



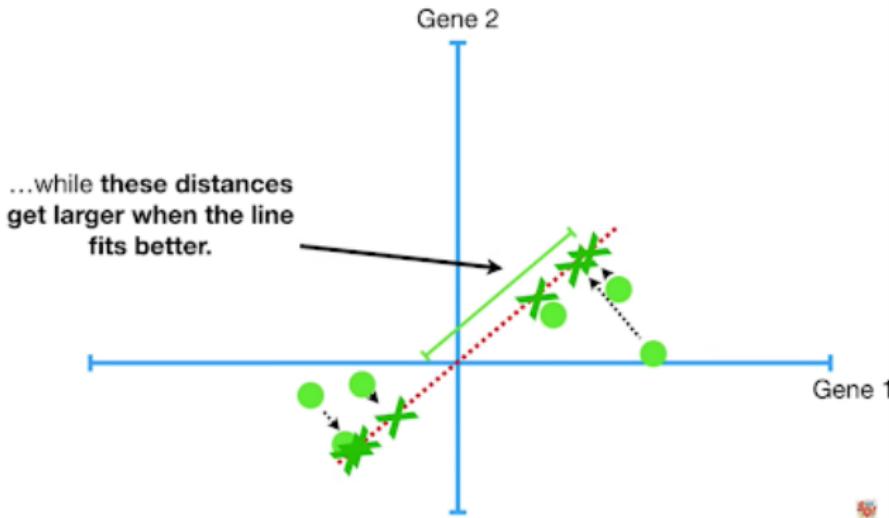
Lo cual puede ser minimizando las distancia de los puntos a la recta



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



O maximizando las proyección de los puntos en la recta! (esto es poco intuitivo)



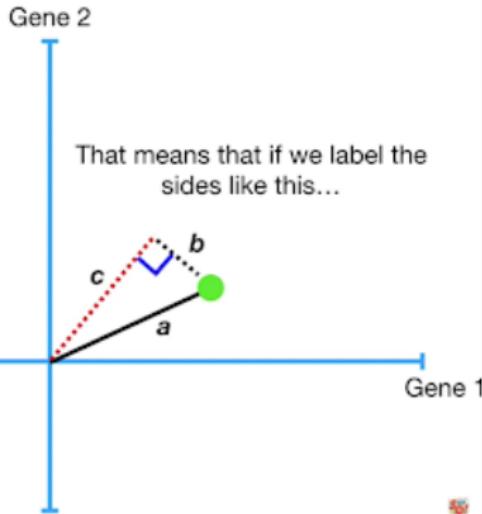
<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Consideremos un solo punto y por Pitágoras vemos que las distancias c y b están relacionadas: si una aumenta la otra disminuye

...then we can use the Pythagorean theorem to show how b and c are inversely related.

$$a^2 = b^2 + c^2$$



That means that if we label the sides like this...

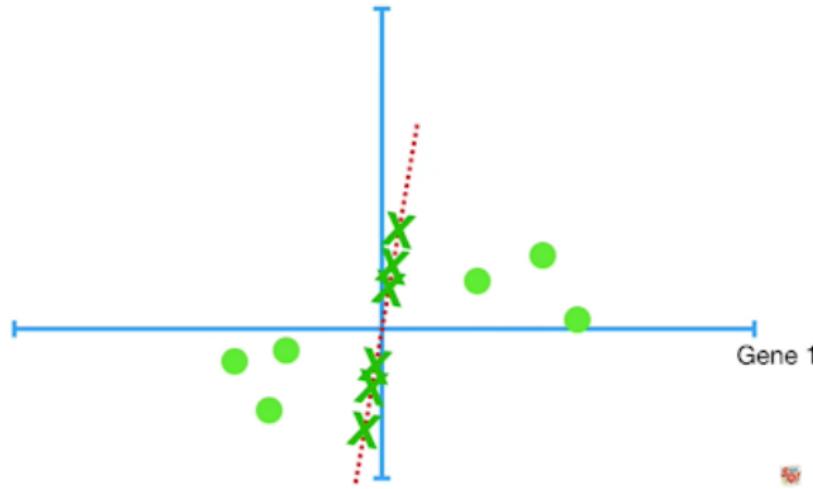


<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Cantidad a maximizar

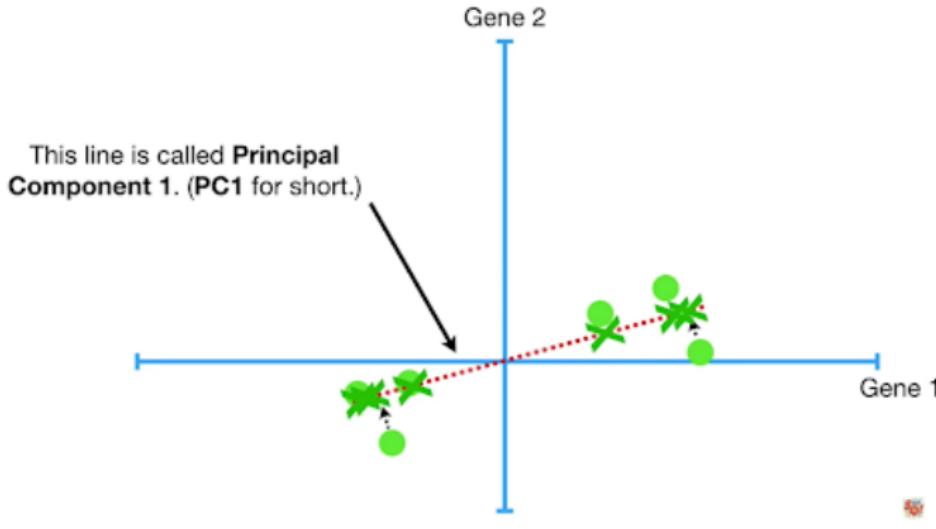
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS(distances)}$$



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



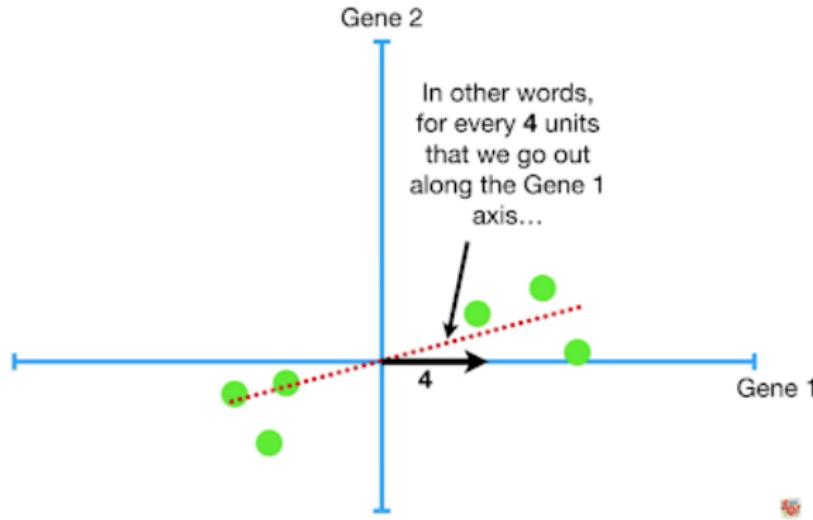
En este caso la pendiente de la recta es 0.25 lo que indica cómo los genes 1 y 2 construyen la componente



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Construimos el vector unitario de PC1



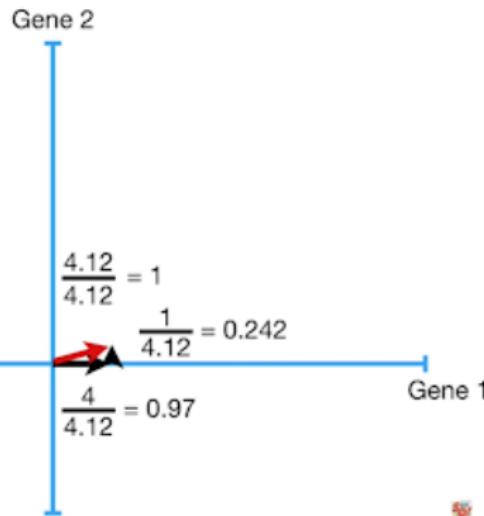
<https://www.youtube.com/watch?v=FgakZw6K1QQ>



The new values change our recipe...

To make PC1
Mix **0.97** parts Gene 1
with **0.242** parts Gene 2

...but the ratio is the same: we still use 4 times as much Gene 1 as Gene 2.



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



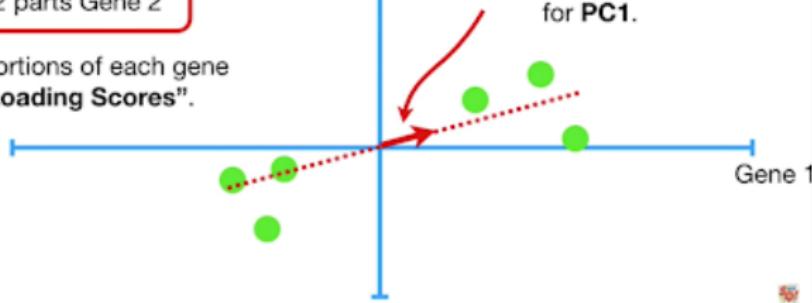
Segunda componente principal

To make PC1
Mix 0.97 parts Gene 1
with 0.242 parts Gene 2

...and the proportions of each gene
are called "Loading Scores".

Gene 2

Terminology Alert!!! This 1 unit long vector, consisting of **0.97** parts Gene 1 and **0.242** parts Gene 2, is called the "**Singular Vector**" or the "**Eigenvector**" for **PC1**.

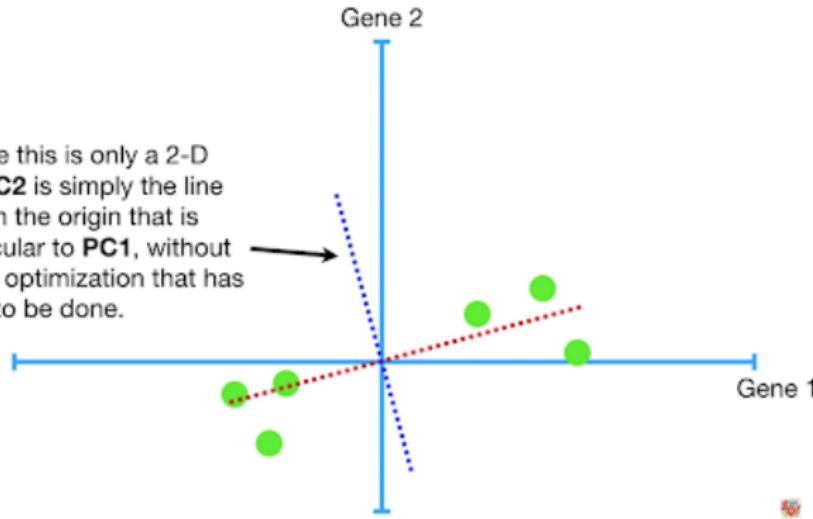


<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Segunda componente principal

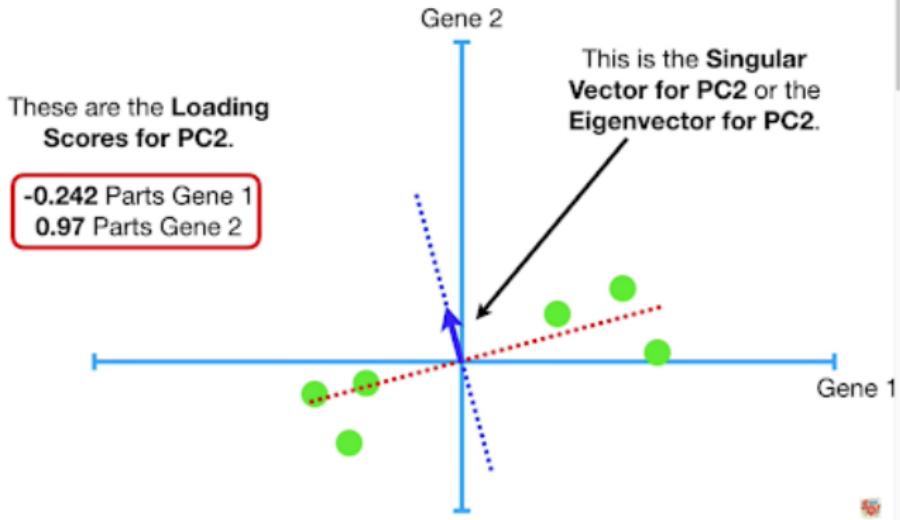
Because this is only a 2-D graph, **PC2** is simply the line through the origin that is perpendicular to **PC1**, without any further optimization that has to be done.



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Graficando en los nuevos ejes

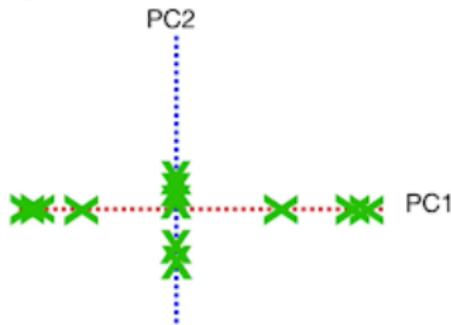


<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Porcentaje de la varianza explicada

...then we use the projected points
to find where the samples go in
the PCA plot.



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Scree plot (considerando que tenemos 3PC)

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

That means that the total variation around both PCs is **15 + 3 = 18...**

$$\frac{SS(\text{distances for PC1})}{n - 1} = \text{Variation for PC1}$$

PC2 ...and that means PC1 accounts for $15 / 18 = 0.83 = 83\%$ of the total variation around the PCs.

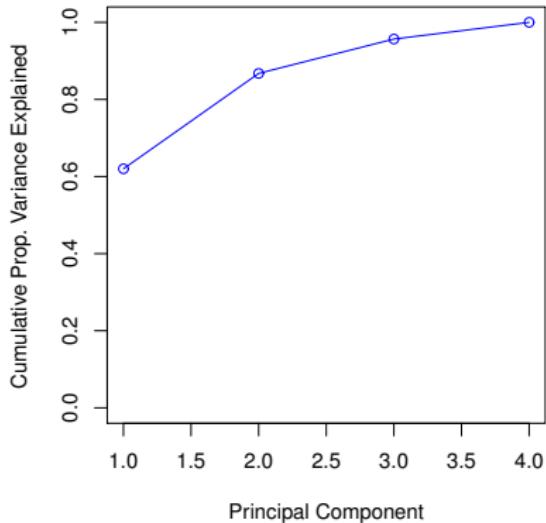
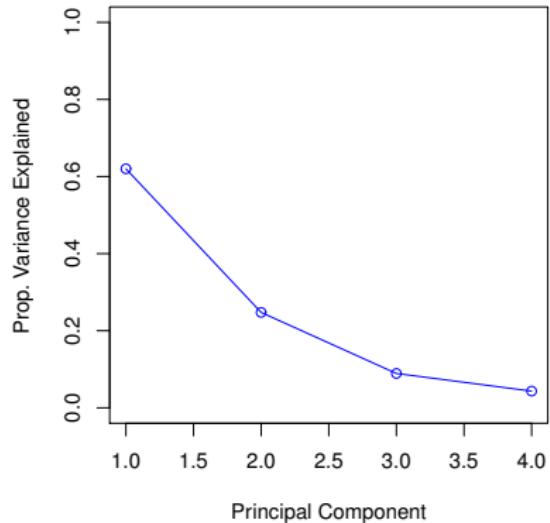
$$\frac{SS(\text{distances for PC2})}{n - 1} = \text{Variation for PC2}$$



<https://www.youtube.com/watch?v=FgakZw6K1QQ>



Scree plot



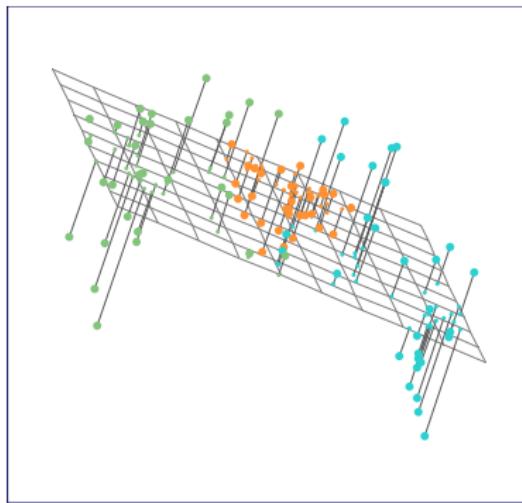
<https://www.statlearning.com/>



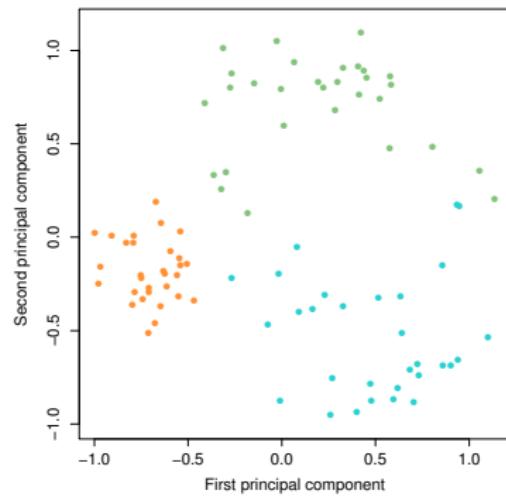
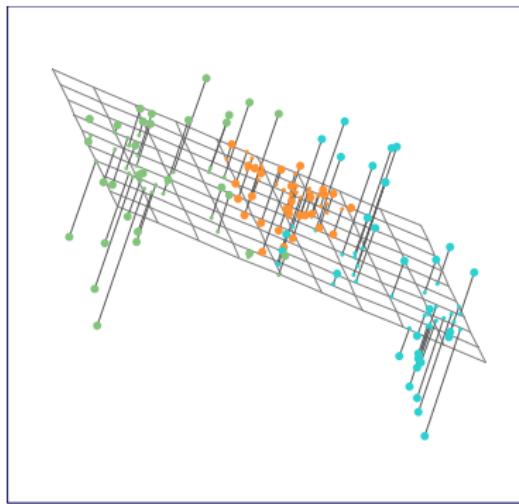
Interpretando datos graficados en componentes principales



Visualización en componentes principales

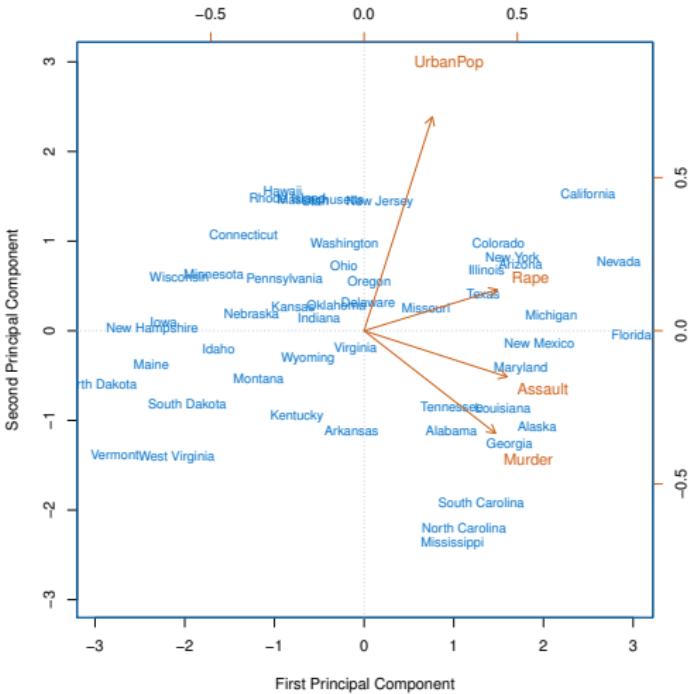


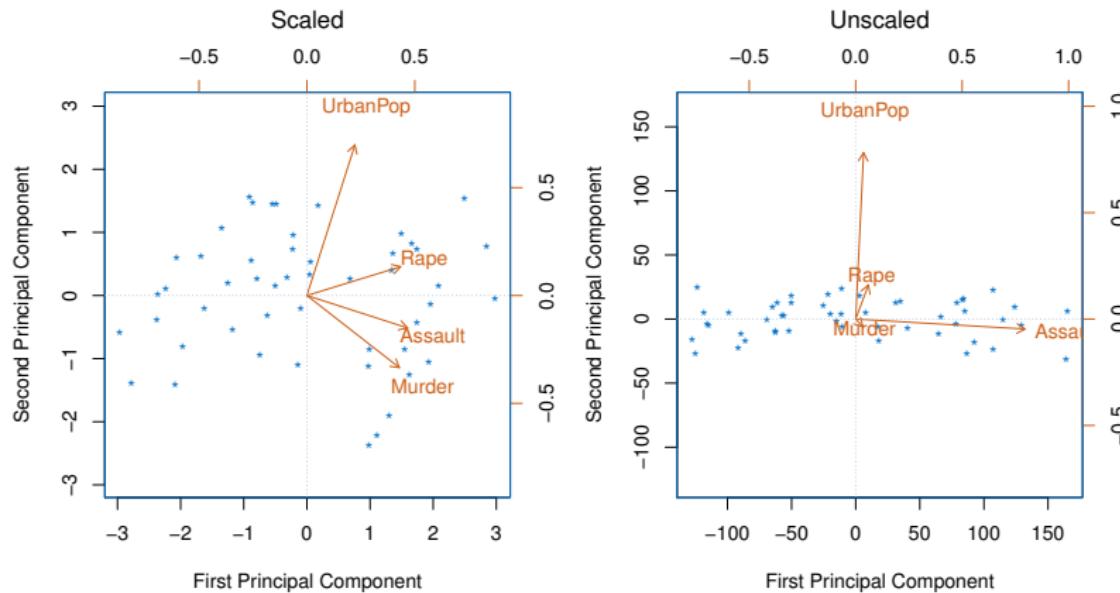
Visualización en componentes principales



<https://www.statlearning.com/>

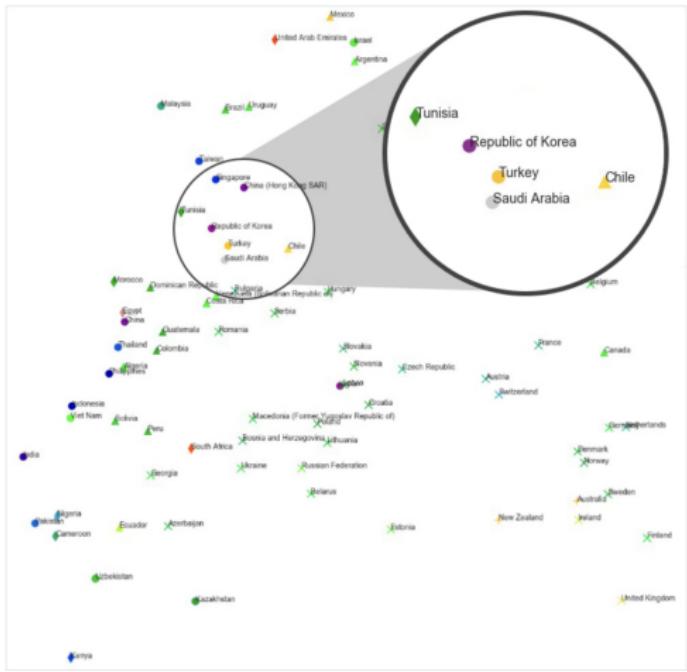






<https://www.statlearning.com/>





*Los colores indican la prevalencia de obesidad | Imagen referencial cortesía de Jocelyn Dunstan



<https://www.emol.com/noticias/Tecnologia/2019/07/30/956345/Chilena-que-proyecto-tasas-de-obesidad-en-72-.html>

html

Free-lunch



There is no free lunch in statistics

No one method dominates all others over all possible data sets



Trade-offs



Recordemos que estamos haciendo

Sea X un vector con características (*input variables*) e Y la variable respuesta (*output variable*)



Recordemos que estamos haciendo

Sea X un vector con características (*input variables*) e Y la variable respuesta (*output variable*)

Diremos que X se relaciona con Y via f :

$$Y = f(X) + \varepsilon,$$

donde ε es un error aleatorio que no depende X y tiene promedio cero.



¿Por qué quisieramos estimar f ?

- Predicción
- Inferencia



¿Por qué quisieramos estimar f ?

- Predicción:

$$\hat{Y} = \hat{f}(X),$$

donde \hat{Y} es la predicción y \hat{f} es la estimación de f .



¿Por qué quisieramos estimar f ?

- Predicción:

$$\hat{Y} = \hat{f}(X),$$

donde \hat{Y} es la predicción y \hat{f} es la estimación de f .

Trataremos de minimizar la distancia entre $f(x)$ y $\hat{f}(x)$ con la elección del modelo. Podemos usar métodos paramétricos (asumir la forma de la relación entre X e Y), o no paramétrico.



¿Por qué quisieramos estimar f ?

- Predicción:

$$\hat{Y} = \hat{f}(X),$$

donde \hat{Y} es la predicción y \hat{f} es la estimación de f .

Trataremos de minimizar la distancia entre $f(x)$ y $\hat{f}(x)$ con la elección del modelo. Podemos usar métodos paramétricos (asumir la forma de la relación entre X e Y), o no paramétrico.

- Inferencia:



¿Por qué quisieramos estimar f ?

- Predicción:

$$\hat{Y} = \hat{f}(X),$$

donde \hat{Y} es la predicción y \hat{f} es la estimación de f .

Trataremos de minimizar la distancia entre $f(x)$ y $\hat{f}(x)$ con la elección del modelo. Podemos usar métodos paramétricos (asumir la forma de la relación entre X e Y), o no paramétrico.

- Inferencia: En este caso queremos entender la asociación entre Y y X . Por ej, ¿Qué predictores están asociados con la variable respuesta?, ¿Cuál es la relación entre la respuesta y cada predictor?, ¿Puedo asumir que la relación entre predictores y respuesta es lineal?



Trade-off entre precisión e interpretabilidad



Bias-Variance trade-off

- Varianza:
 - ¿Cuánto cambia \hat{f} si lo estimo usando datos de entrenamiento distintos?



Bias-Variance trade-off

- Varianza:

- ¿Cuánto cambia \hat{f} si lo estimo usando datos de entrenamiento distintos?
- Idealmente no debería cambiar mucho. Modelos más flexibles tienden a tener mayor varianza.



Bias-Variance trade-off

- Varianza:
 - ¿Cuánto cambia \hat{f} si lo estimo usando datos de entrenamiento distintos?
 - Idealmente no debería cambiar mucho. Modelos más flexibles tienden a tener mayor varianza.
- Sesgo (*Bias*):
 - Error introducido por aproximar un problema de la vida real por un modelo más simple.

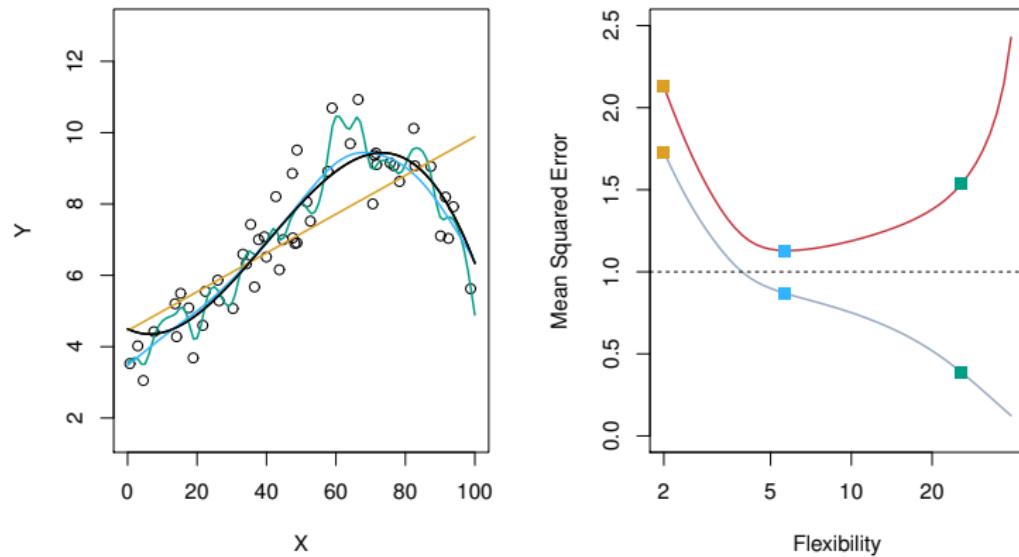


Bias-Variance trade-off

- Varianza:
 - ¿Cuánto cambia \hat{f} si lo estimo usando datos de entrenamiento distintos?
 - Idealmente no debería cambiar mucho. Modelos más flexibles tienden a tener mayor varianza.
- Sesgo (*Bias*):
 - Error introducido por aproximar un problema de la vida real por un modelo más simple.
 - Modelos más flexibles tienden a tener menor *bias*.



Bias-Variance trade-off



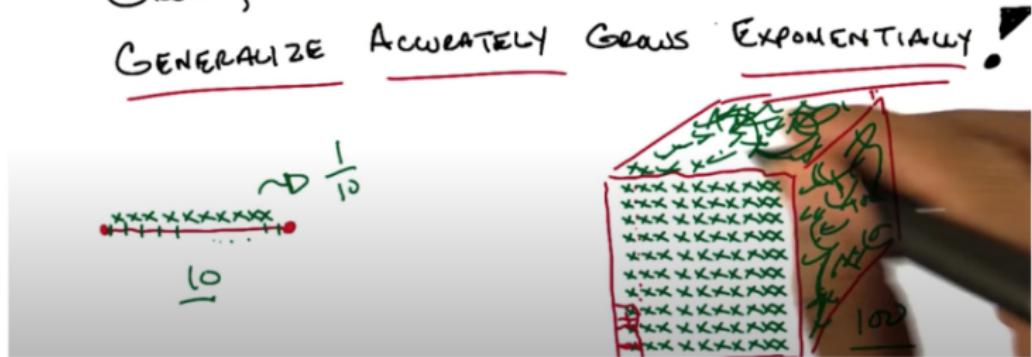
Left: Data simulated from f , shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.



Maldición de la dimensionalidad

CURSE OF DIMENSIONALITY

AS THE NUMBER OF FEATURES OR DIMENSIONS
GROWS, THE AMOUNT OF DATA WE NEED TO
GENERALIZE ACCURATELY GROWS EXPONENTIALLY!



<https://www.youtube.com/watch?v=0yPcbeiwps8>

