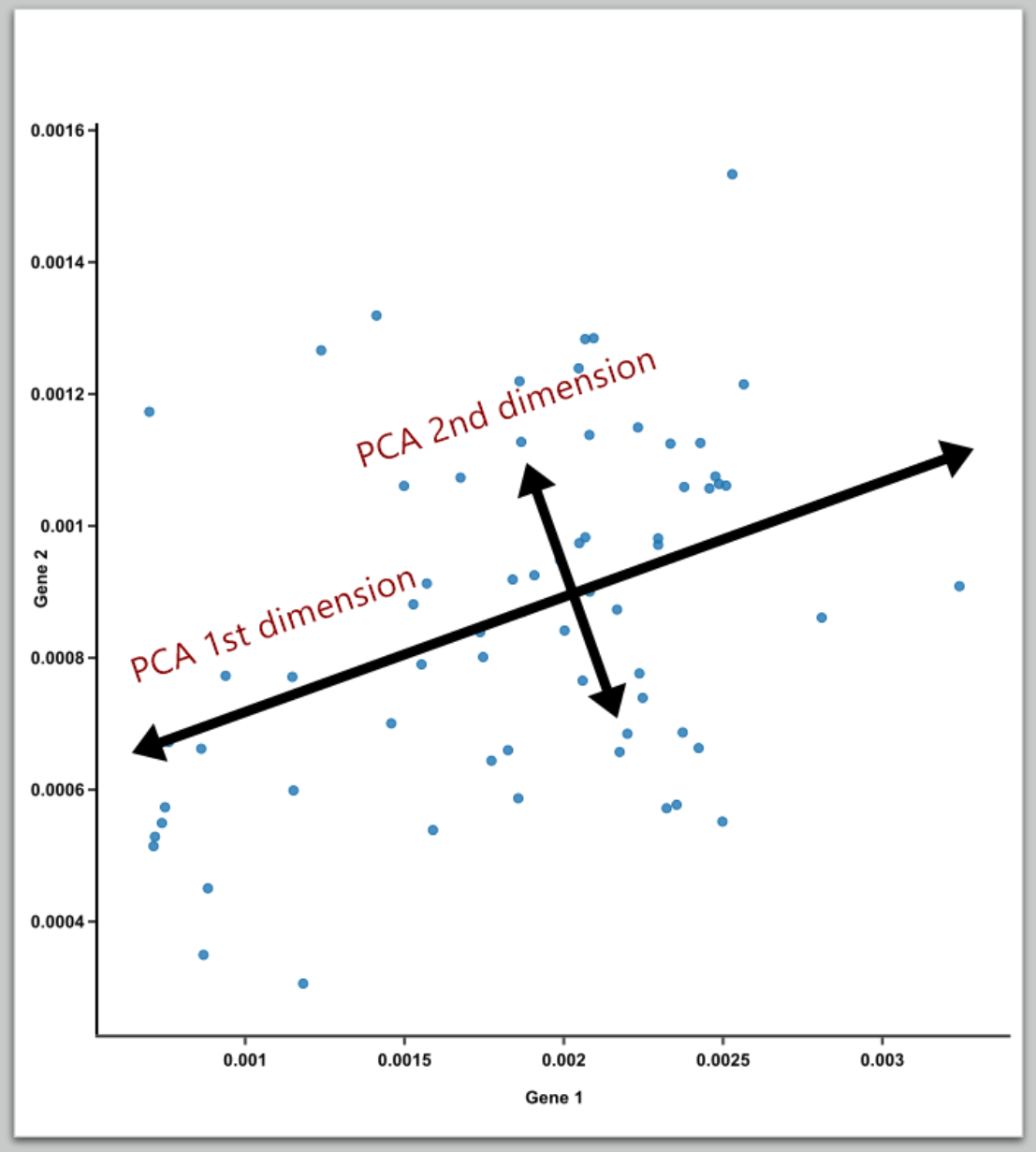
Machine Learning
PCA

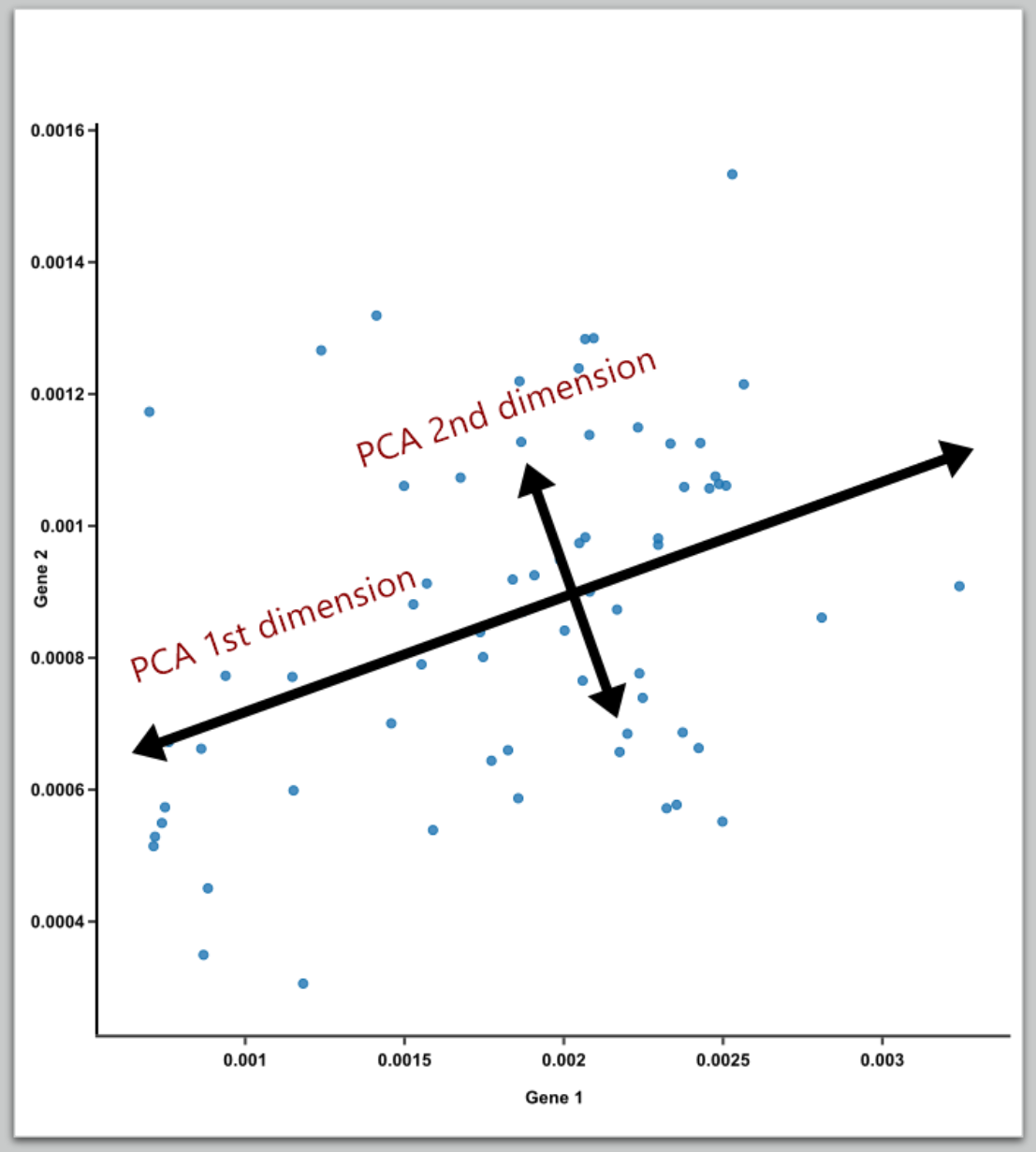
¿Qué es el PCA?

- Principal Component Analysis (PCA)
- Método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.
- Supóngase que existe una muestra con n individuos cada uno con p variables (X1, X2, …, Xp), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes (z<p) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores.
- Cada una de estas z nuevas variables recibe el nombre de componente principal.



¿Qué es el PCA?

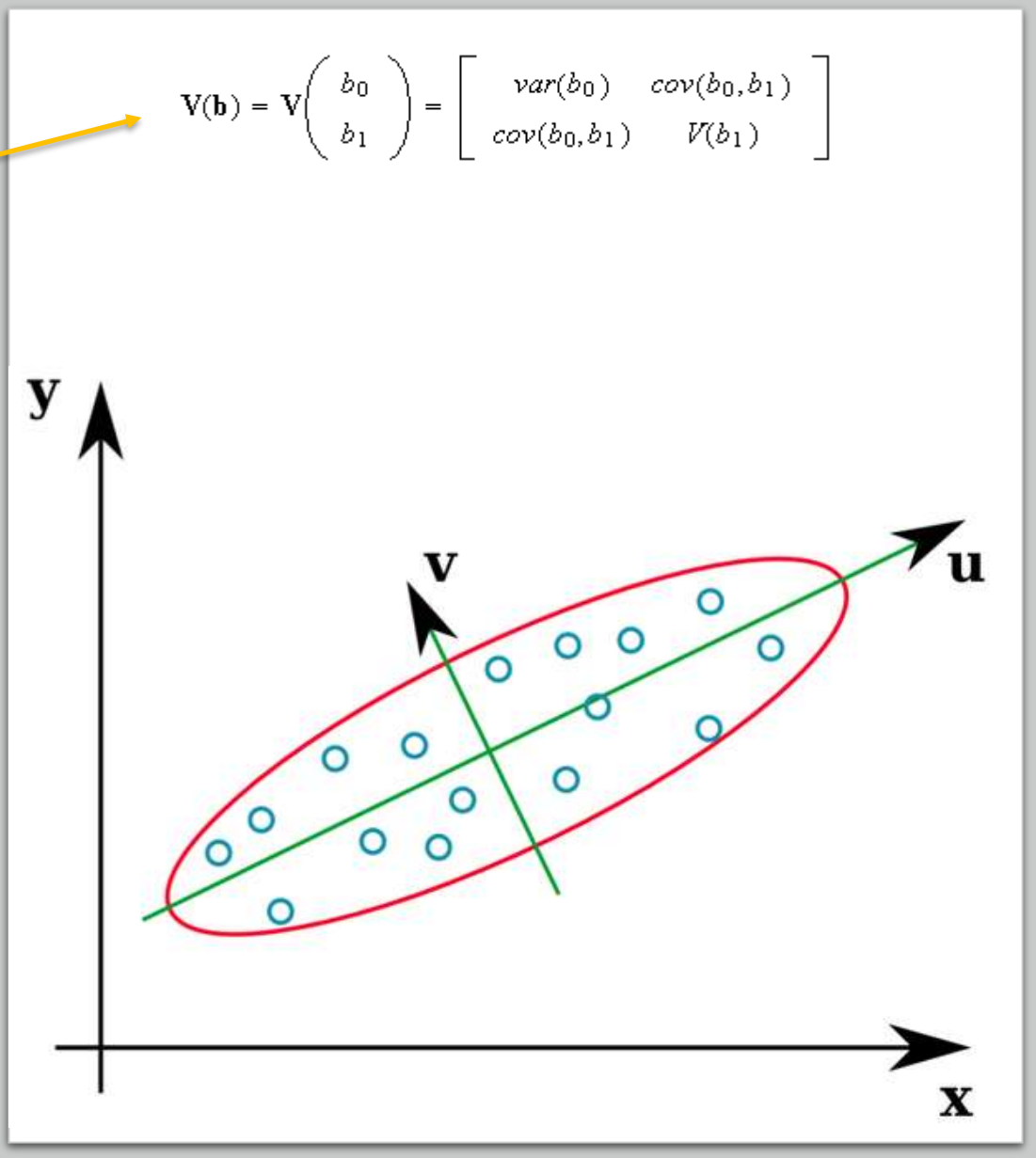
- El método de PCA permite por lo tanto “condensar” la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, clustering, etc.



Álgebra lineal

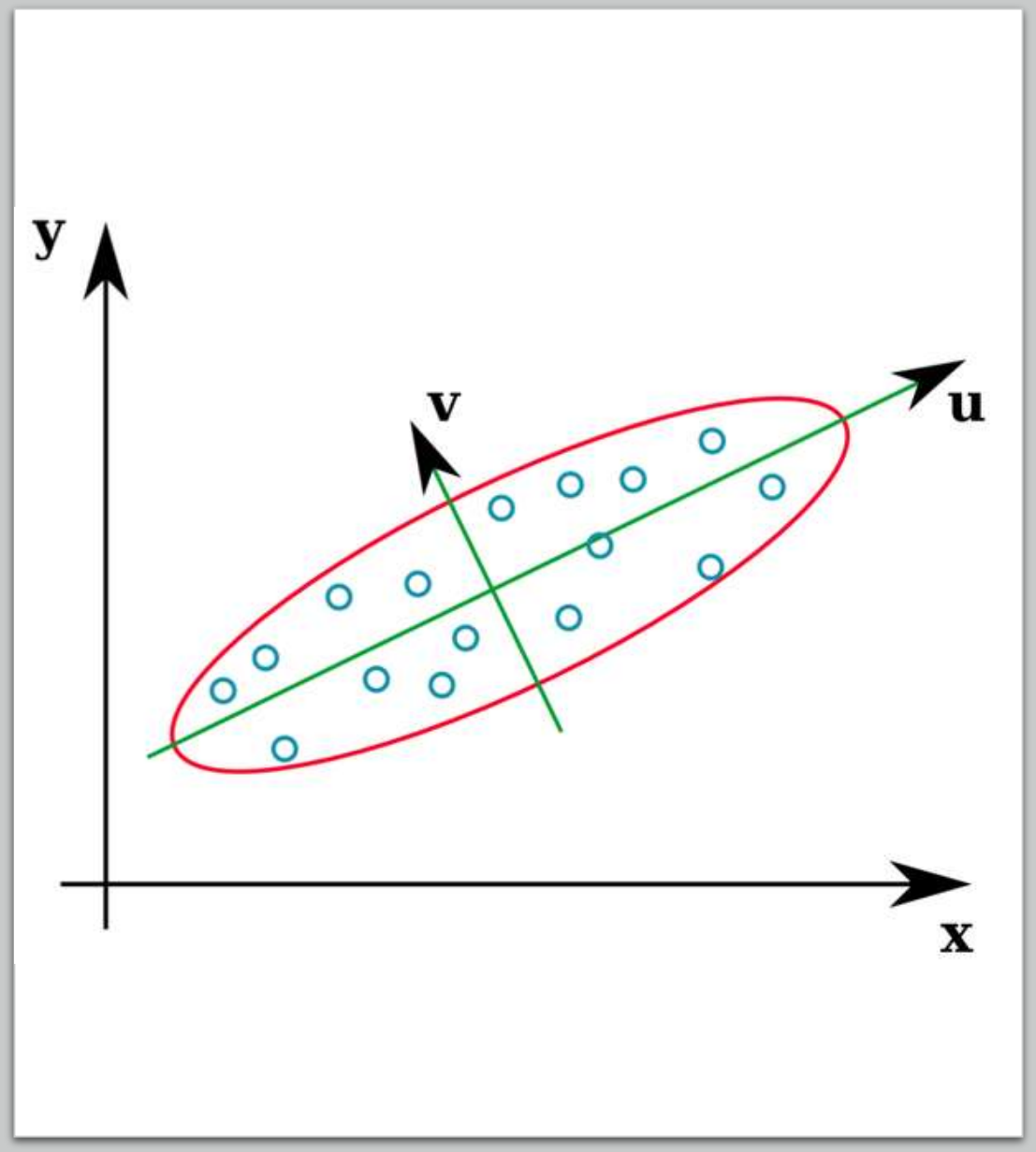
- Para calcular el PCA debemos obtener la matriz de covarianza de nuestros datos.
- Realizamos transformaciones lineales de nuestra matriz de covarianza: transformaciones de los puntos en el plano.
- Las transformaciones nos devuelven dos vectores (u, v, en la imagen).
  - Vectores propios (dirección): Eigenvectors
- Las transformaciones nos devuelven los valores propios.
- Aquellos valores más altos, son los que representan la mayor varianza de nuestros datos.

$$\mathbf{V(b)} = \mathbf{V}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} var(b_0) & cov(b_0, b_1) \\ cov(b_0, b_1) & V(b_1) \end{bmatrix}$$



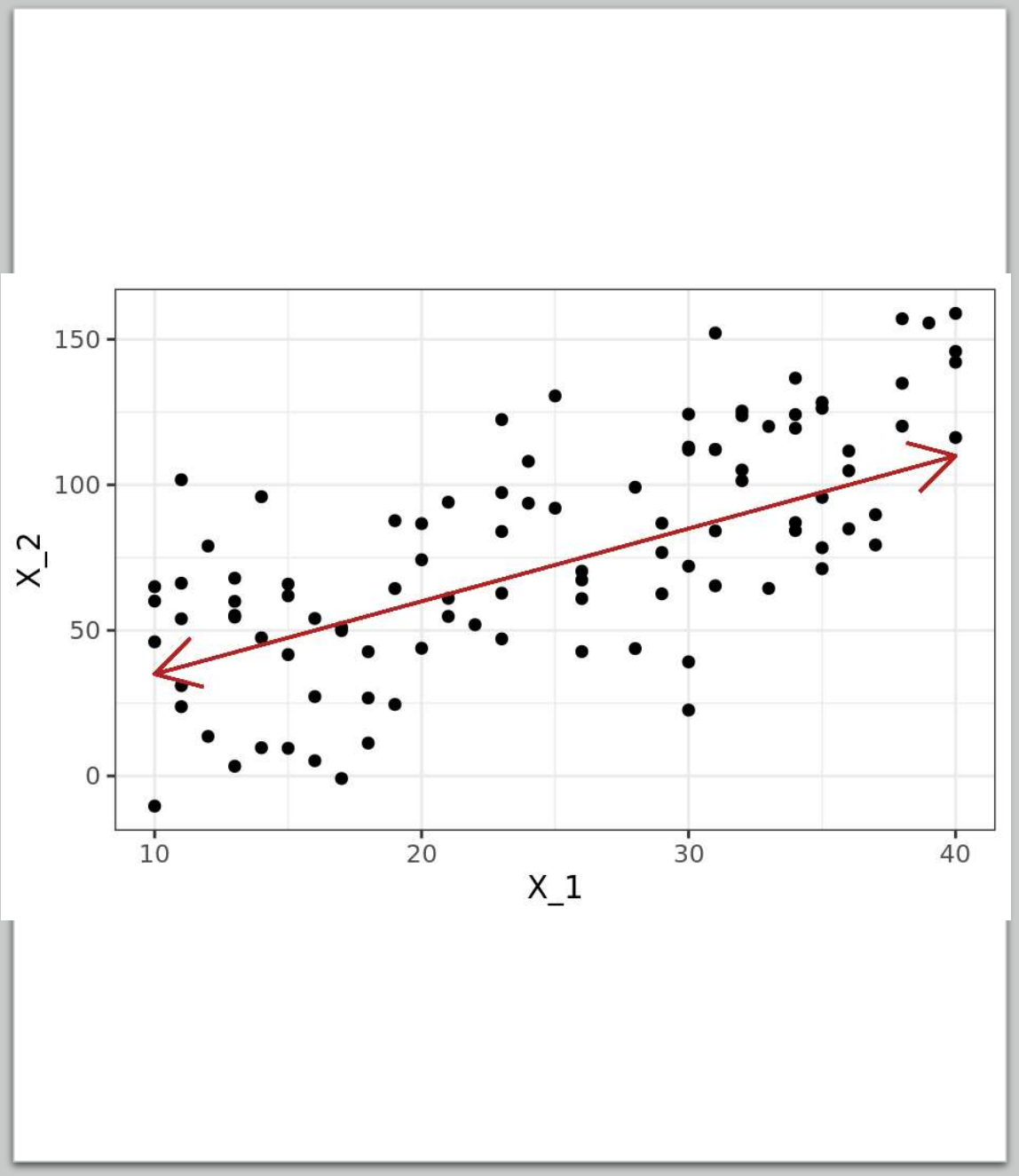
Álgebra lineal

- PCA aplica dos conceptos matemáticos.
- Eigenvectors:
  - Los eigenvectors de una matriz son todos aquellos vectores que, al multiplicarlos por dicha matriz, resultan en el mismo vector o en un múltiplo entero del mismo.
- Eigenvalue:
  - Cuando se multiplica una matriz por alguno de sus eigenvectors se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número.
  - A todo eigenvector le corresponde un eigenvalue y viceversa.
- En el método PCA, cada una de las componentes se corresponde con un eigenvector, y el orden de componente se establece por orden de eigenvalue.

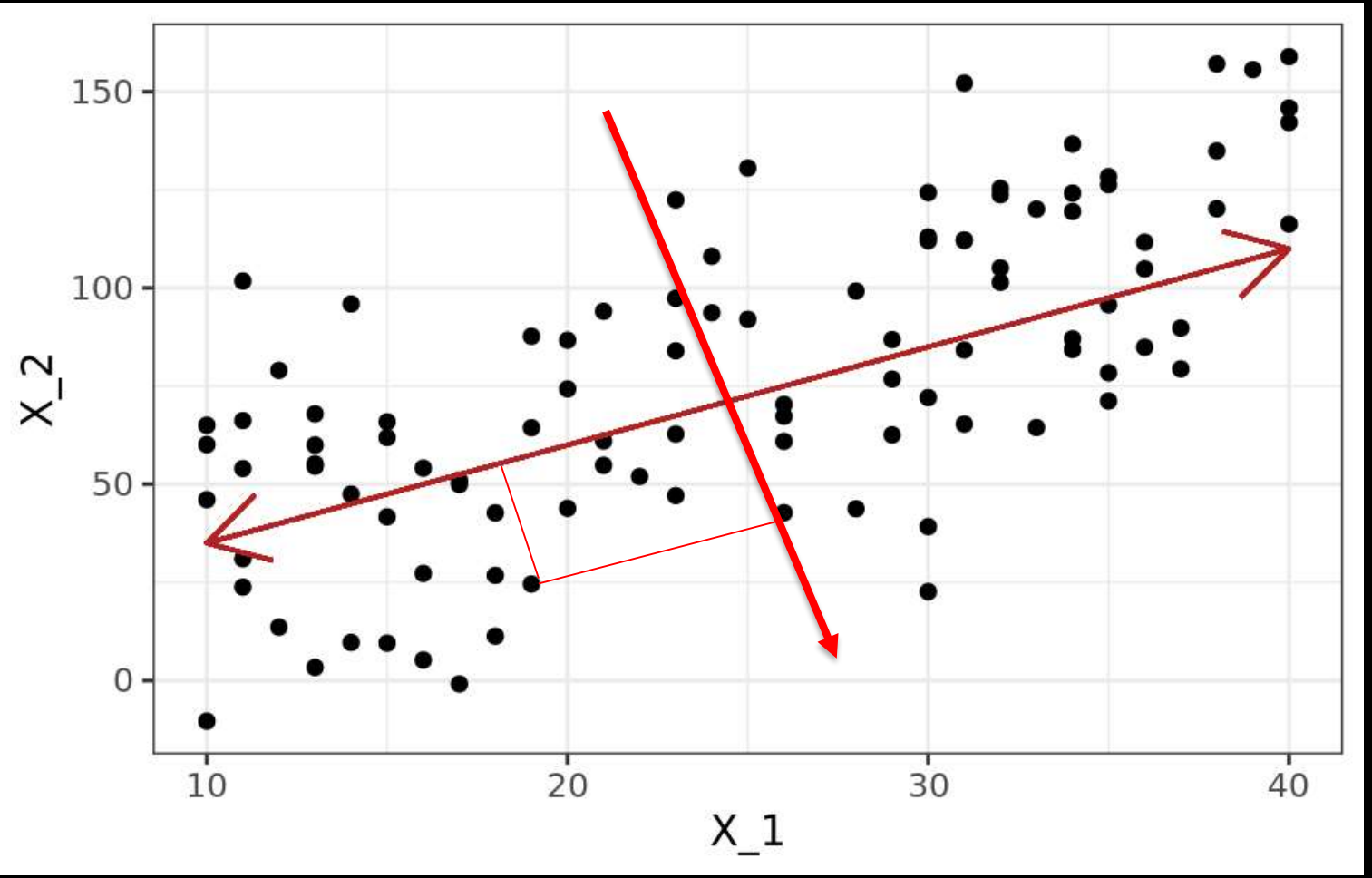


¿Cómo funciona?

- Una forma intuitiva de entender el proceso de PCA consiste en interpretar las componentes principales desde un punto de vista geométrico.



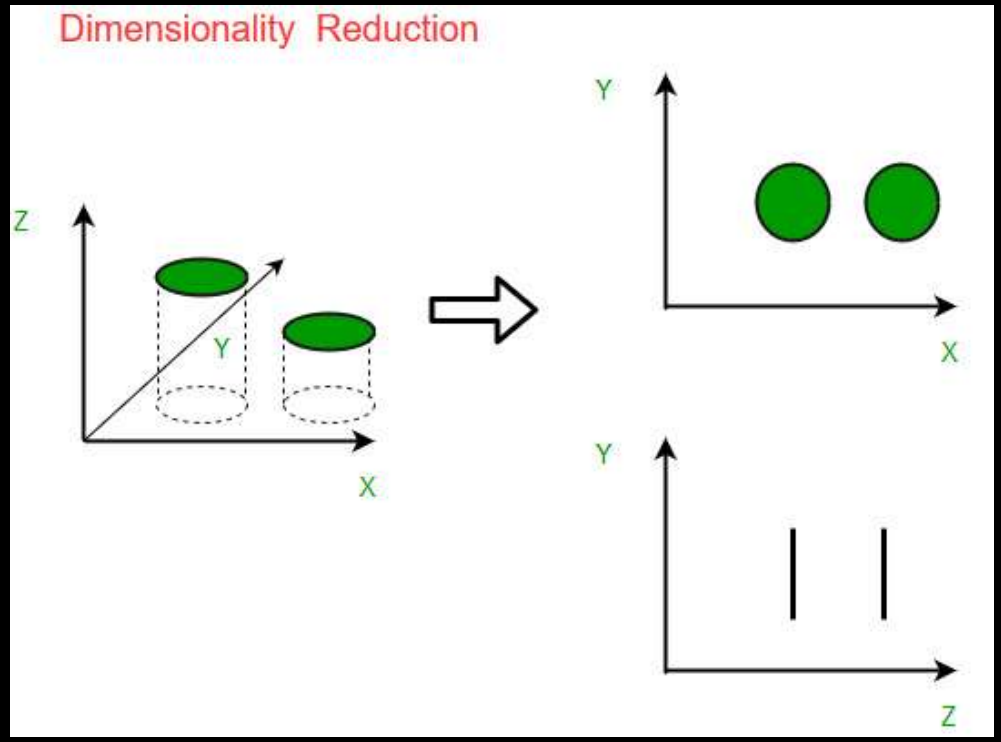
¿Cómo funciona?



¿Cómo funciona?

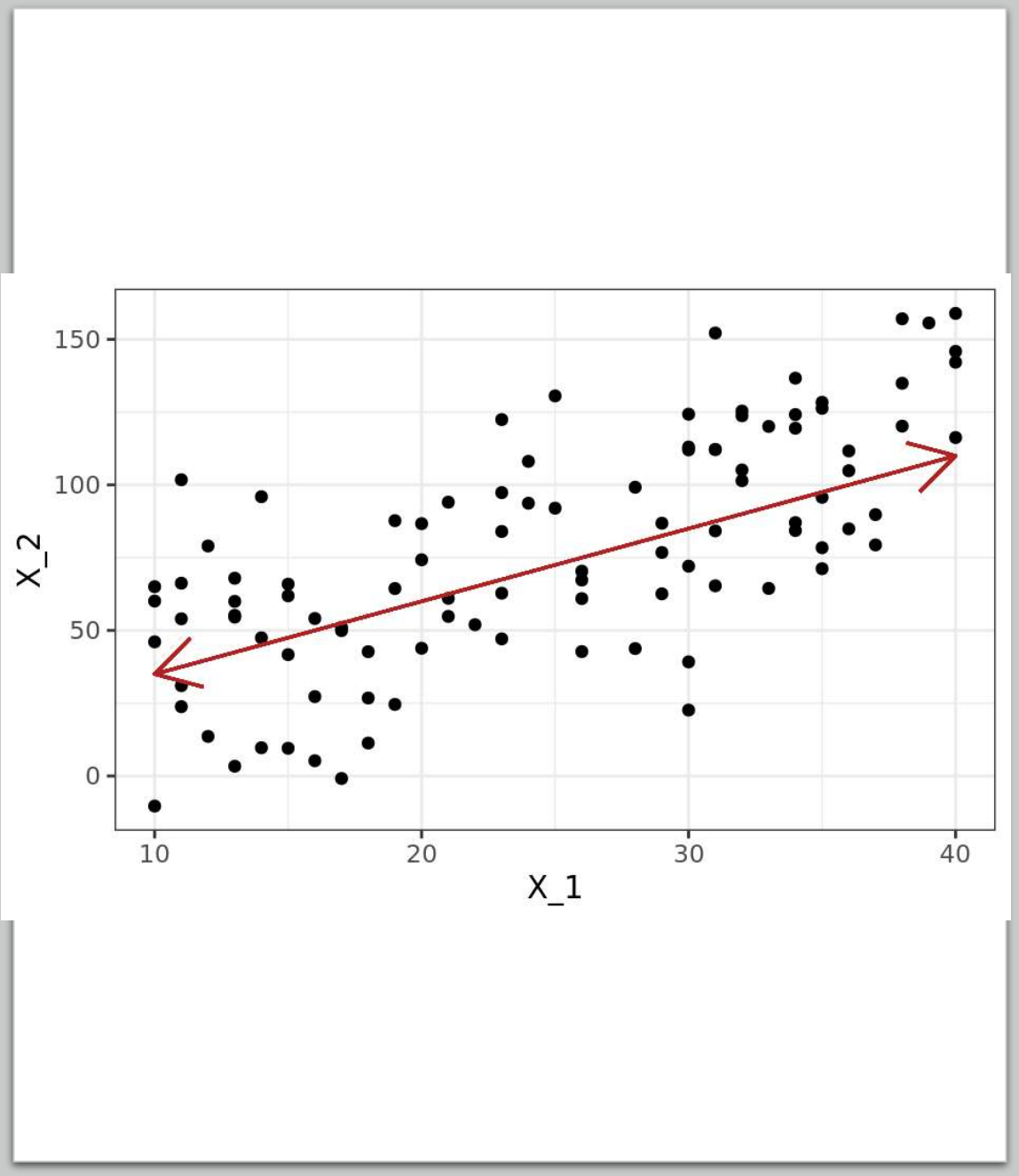
Cálculo de componentes principales

- Cada componente principal se obtiene por combinación lineal de las variables originales.
- Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales.
- La primera componente principal de un grupo de variables (X1, X2, …, Xp) es la combinación lineal de dichas variables que tiene mayor varianza.



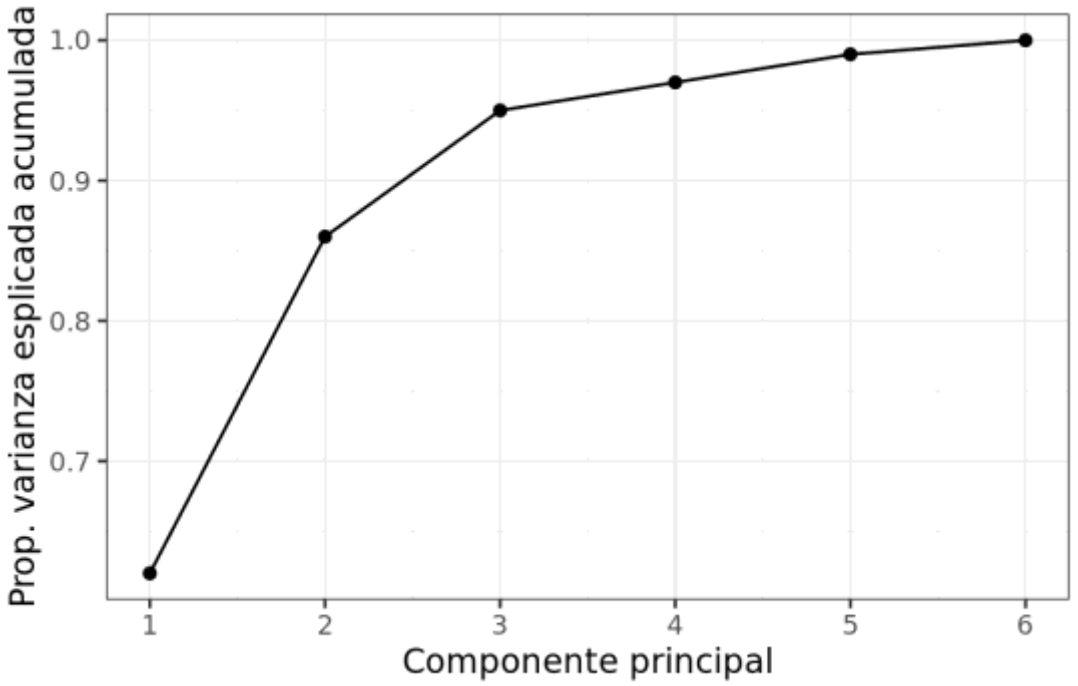
Proceso para calcular la primera componente

1) Centralización de las variables: se resta a cada valor la media de la variable a la que pertenece. Con esto se consigue que todas las variables tengan media cero.
2) Se obtiene cada componente a través de la optimización de combinaciones lineales para obtener aquellas con la máxima varianza.
3) Una vez calculada la primera, se obtiene la segunda y se repite iterativamente hasta calcular todas las posibles componentes, o hasta que se decida detener el proceso.



A TENER EN CUENTA

- Escalado de variables: PCA identifica direcciones cuya varianza es mayor. Por ello deberemos tener los datos en la misma escala.
- Influencia de outliers: al trabajar con varianzas, PCA es altamente sensible a outliers. Es muy recomendable estudiar si los hay.
- ¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión? (Varianza explicada de cada componente principal).
- Es de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos.



Preguntas