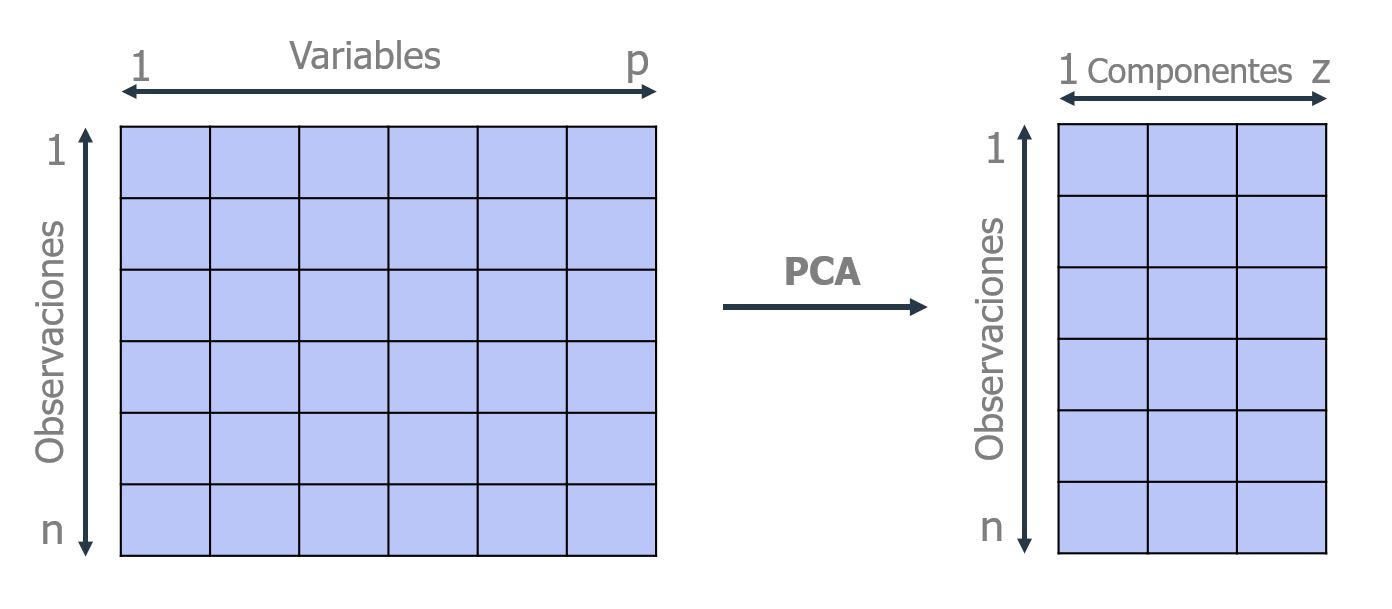
El análisis de componentes principales (*Principal Component Analysis PCA*) es un método de reducción de dimensionalidad que permite simplificar la complejidad de espacios con múltiples dimensiones a la vez que conserva su información.

Supóngase que existe una muestra con nn individuos cada uno con pp variables (X1X1, X2X2, ..., XpXp), es decir, el espacio muestral tiene pp dimensiones. PCA permite encontrar un número de factores subyacentes (z<p)(z<p) que explican aproximadamente lo mismo que las pp variables originales. Donde antes se necesitaban pp valores para caracterizar a cada individuo, ahora bastan zz valores. Cada una de estas zz nuevas variables recibe el nombre de componente principal.



El método de PCA permite por lo tanto "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Aun así, no hay que olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes. Dos de las principales aplicaciones del PCA son la visualización y el preprocesado de predictores previo ajuste de modelos supervisados.

La librería **[scikitlearn](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html)** contiene la clase sklearn.decomposition.PCA que implementa la mayoría de las funcionalidades necesarias para crear y utilizar modelos PCA. Para visualizaciones, **[Yellowbrick](https://www.scikit-yb.org/en/latest/api/features/pca.html)** ofrece funcionalidades extra.

**Álgebra lineal**

En esta sección se describen dos de los conceptos matemáticos que se aplican en el PCA: *eigenvectors* y *eigenvalues*. Se trata simplemente de una descripción intuitiva con la única finalidad de facilitar el entendimiento del cálculo de componentes principales.

**Eigenvectors**

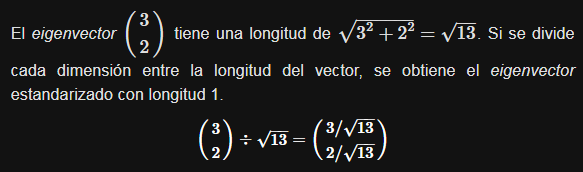
Los *eigenvectors* son un caso particular de multiplicación entre una matriz y un vector. Obsérvese la siguiente multiplicación:



El vector resultante de la multiplicación es un múltiplo entero del vector original. 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. Los *eigenvectors* tienen una serie de propiedades matemáticas específicas:

* Los *eigenvectors* solo existen para matrices cuadradas y no para todas. En el caso de que una matriz *n* x *n* tenga *eigenvectors*, el número de ellos es *n*.
* Si se escala un *eigenvector* antes de multiplicarlo por la matriz, se obtiene un múltiplo del mismo *eigenvector*. Esto se debe a que, si se escala un vector multiplicándolo por cierta cantidad, lo único que se consigue es cambiar su longitud pero la dirección es la misma.
* Todos los *eigenvectors* de una matriz son perpendiculares (ortogonales) entre ellos, independientemente de las dimensiones que tengan.

Dada la propiedad de que multiplicar un *eigenvector* solo cambia su longitud pero no su naturaleza de *eigenvector*, es frecuente escalarlos de tal forma que su longitud sea 1. De este modo se consigue que todos ellos estén estandarizados. A continuación se muestra un ejemplo:



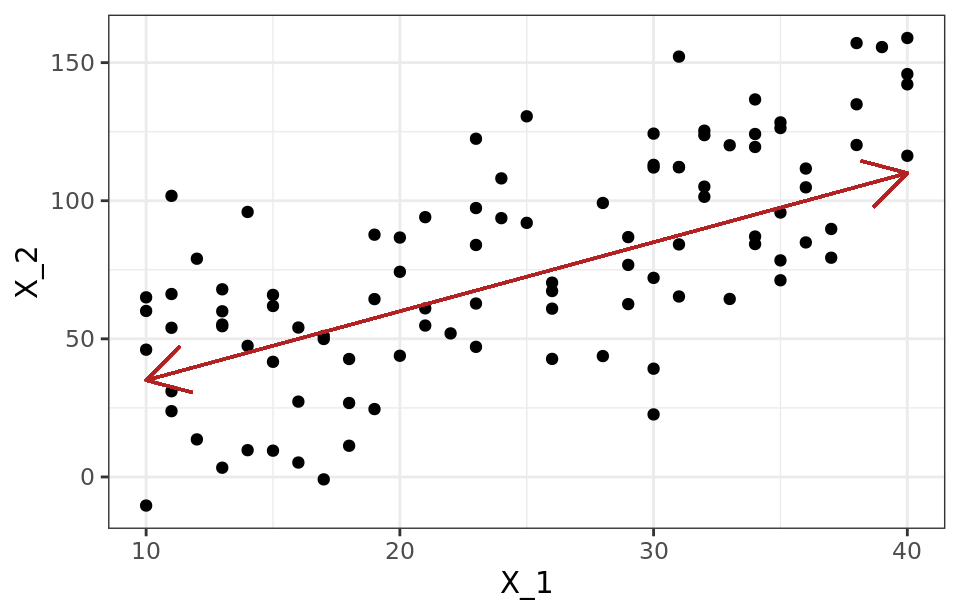
**Eigenvalue**

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. Al valor por el que se multiplica el *eigenvector* resultante se le conoce como *eigenvalue*. A todo *eigenvector* le corresponde un *eigenvalue* y viceversa.

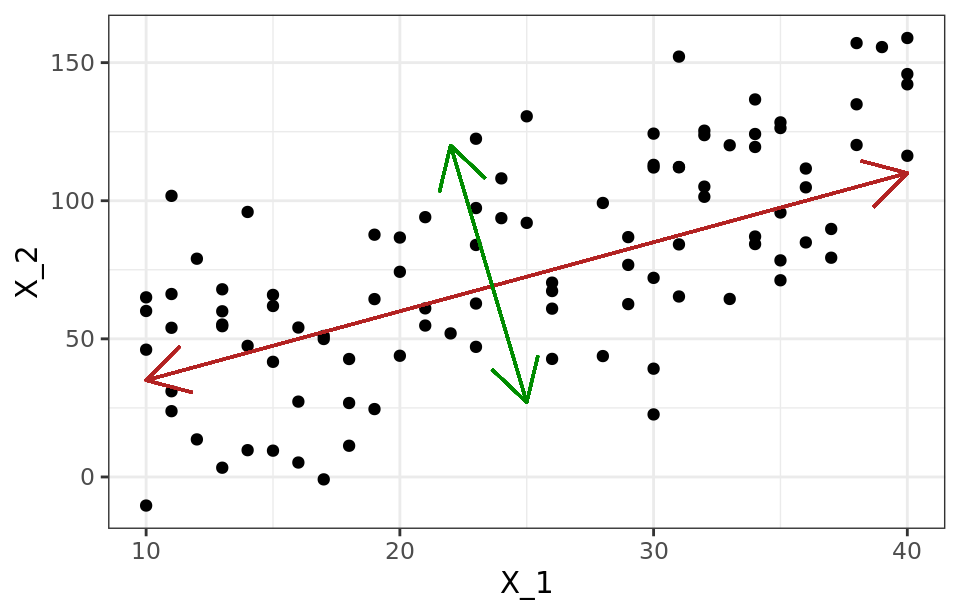
En el método PCA, cada una de las componentes se corresponde con un *eigenvector*, y el orden de componente se establece por orden decreciente de *eigenvalue*. Así pues, la primera componente es el *eigenvector* con el *eigenvalue* más alto.

**Interpretación geométrica de las componentes principales**

Una forma intuitiva de entender el proceso de PCA es interpretar las componentes principales desde un punto de vista geométrico. Supóngase un conjunto de observaciones para las que se dispone de dos variables (X1X1, X2X2). El vector que define la primera componente principal (Z1Z1) sigue la dirección en la que las observaciones tienen más varianza (línea roja). La proyección de cada observación sobre esa dirección equivale al valor de la primera componente para dicha observación (*principal component score*, zi1zi1).



La segunda componente (Z2Z2) sigue la segunda dirección en la que los datos muestran mayor varianza y que no está correlacionada con la primera componente. La condición de no correlación entre componentes principales equivale a decir que sus direcciones son perpendiculares/ortogonales.

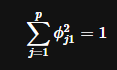


**Cálculo de las componentes principales**

Cada componente principal (ZiZi) 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 (X1X1, X2X2, ..., XpXp) es la combinación lineal normalizada de dichas variables que tiene mayor varianza:



Que la combinación lineal sea normalizada implica que:



Los términos ϕ11ϕ11, ..., ϕ1pϕ1p reciben en el nombre de *loadings* y son los que definen las componentes. Por ejemplo, ϕ11ϕ11 es el *loading* de la variable X1X1 de la primera componente principal. Los *loadings* pueden interpretarse como el peso/importancia que tiene cada variable en cada componente y, por lo tanto, ayudan a conocer que tipo de información recoge cada una de las componentes.

Dado un set de datos XX con *n* observaciones y *p* variables, el proceso a seguir para calcular la primera componente principal es:

* Centrar 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.
* Se resuelve un problema de optimización para encontrar el valor de los *loadings* con los que se maximiza la varianza. Una forma de resolver esta optimización es mediante el cálculo de *eigenvector-eigenvalue* de la matriz de covarianzas.

Una vez calculada la primera componente (Z1Z1), se calcula la segunda (Z2Z2) repitiendo el mismo proceso pero añadiendo la condición de que la combinación lineal no pude estar correlacionada con la primera componente. Esto equivale a decir que Z1Z1 y Z2Z2 tienen que ser perpendiculares. EL proceso se repite de forma iterativa hasta calcular todas las posibles componentes (min(*n*-1, *p*)) o hasta que se decida detener el proceso. El orden de importancia de las componentes viene dado por la magnitud del *eigenvalue* asociado a cada *eigenvector*.

**Escalado de las variables**

El proceso de PCA identifica las direcciones con mayor varianza. Como la varianza de una variable se mide en sus mismas unidades elevadas al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media cero y desviación estándar de uno, aquellas variables cuya escala sea mayor dominarán al resto. De ahí que sea recomendable estandarizar siempre los datos.

**Reproducibilidad de las componentes**

El proceso de PCA estándar es determinista, genera siempre las mismas componentes principales, es decir, el valor de los *loadings* resultantes es el mismo. La única diferencia que puede darse es que el signo de todos los *loadings* esté invertido. Esto es así porque el vector de *loadings* determina la dirección de la componente, y dicha dirección es la misma independientemente del signo (la componente sigue una línea que se extiende en ambas direcciones). Del mismo modo, el valor específico de las componentes obtenido para cada observación (*principal component scores*) es siempre el mismo, a excepción del signo.

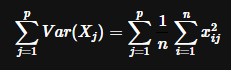
**Influencia de outliers**

Al trabajar con varianzas, el método PCA es muy sensible a *outliers*, por lo que es recomendable estudiar si los hay. La detección de valores atípicos con respecto a una determinada dimensión es algo relativamente sencillo de hacer mediante comprobaciones gráficas. Sin embargo, cuando se trata con múltiples dimensiones el proceso se complica. Por ejemplo, considérese un hombre que mide 2 metros y pesa 50 kg. Ninguno de los dos valores es atípico de forma individual, pero en conjunto se trataría de un caso muy excepcional.

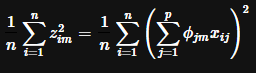
**Proporción de varianza explicada**

Una de las preguntas más frecuentes que surge tras realizar un PCA es: ¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión? o lo que es lo mismo ¿Cuanta información es capaz de capturar cada una de las componentes principales obtenidas? Para contestar a estas preguntas se recurre a la proporción de varianza explicada por cada componente principal.

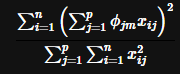
Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como



y la varianza explicada por la componente *m* es



Por lo tanto, la proporción de varianza explicada por la componente *m* viene dada por el ratio



Tanto la proporción de varianza explicada, como la proporción de varianza explicada acumulada, son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar en los análisis posteriores. Si se calculan todas las componentes principales de un set de datos, entonces, aunque transformada, se está almacenando toda la información presente en los datos originales. El sumatorio de la proporción de varianza explicada acumulada de todas las componentes es siempre 1.

**Número óptimo de componentes principales**

Por lo general, dada una matriz de datos de dimensiones *n* x *p*, el número de componentes principales que se pueden calcular es como máximo de *n*-1 o *p* (el menor de los dos valores es el limitante). Sin embargo, siendo el objetivo del PCA reducir la dimensionalidad, suelen ser de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos. No existe una respuesta o método único que permita identificar cual es el número óptimo de componentes principales a utilizar. Una forma de proceder muy extendida consiste en evaluar la proporción de varianza explicada acumulada y seleccionar el número de componentes mínimo a partir del cual el incremento deja de ser sustancial.

