



PCA CON REGRESIÓN

Benitez Patiño Ian
Cañas Taboada Andrea
García García Adrián
Maya Cortés Angel
Rodríguez Herrera Dante

¿QUÉ ES EL PCA?

Técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos al transformarlo en un conjunto de componentes lineales no correlacionadas. Ayuda a identificar patrones subyacentes en los datos al resumir la información en nuevas variables llamadas componentes principales.



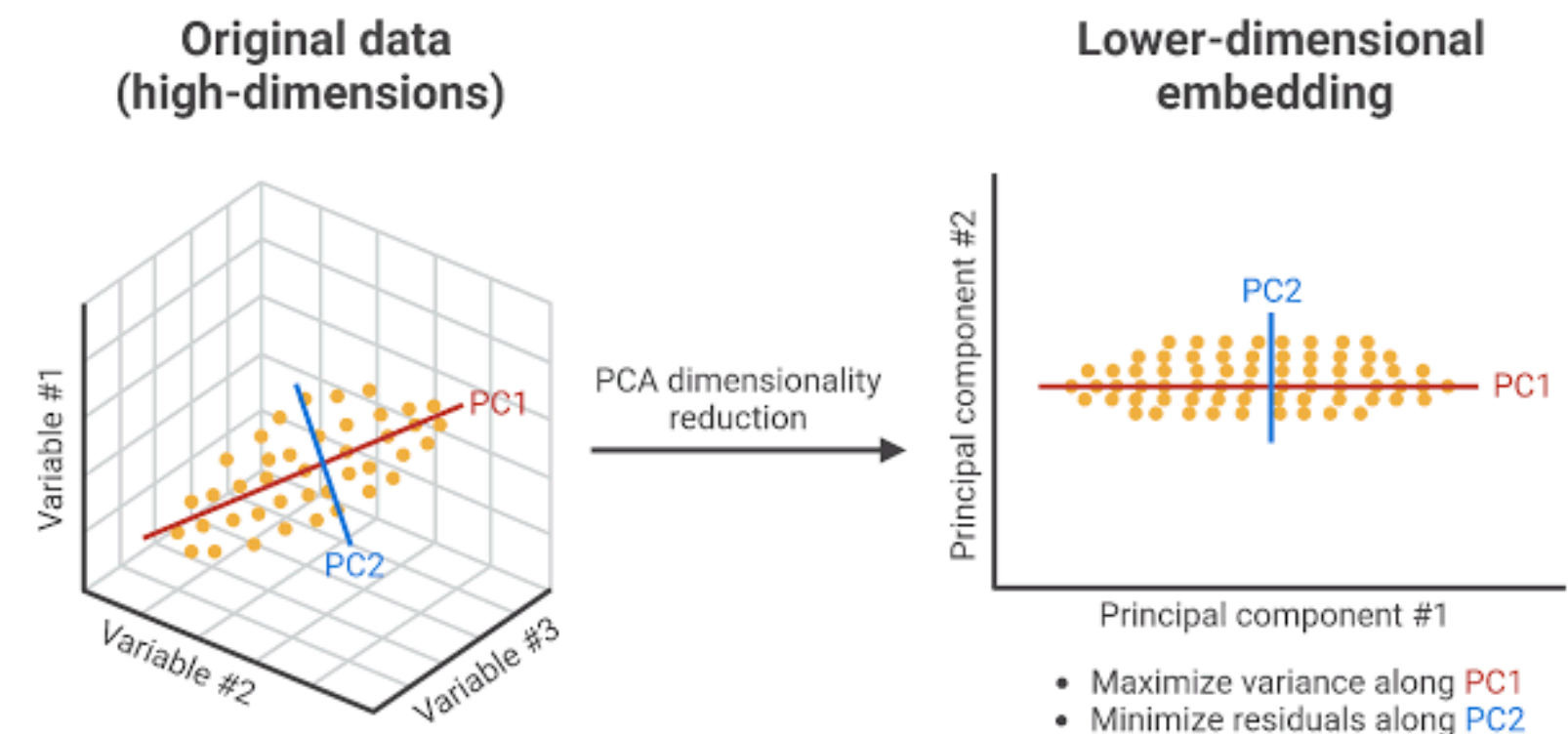
Es comúnmente usado en análisis de datos y aprendizaje automático para simplificar la interpretación y visualización de conjuntos de datos complejos.

Tiene usos importantes como lo son:

- Preprocesamiento de datos
- Eliminación de multicolinealidad
- Visualización de datos

PCA
PCA
PCA

Principal Component Analysis (PCA) Transformation



1. Estandarización de datos: Si es necesario, se estandarizan los datos para que tengan media cero y varianza unitaria.
2. Cálculo de la matriz de covarianza o de correlación: Se calcula la matriz de covarianza o correlación según el tipo de datos y el objetivo del análisis.
3. Descomposición de la matriz: Se realiza la descomposición de la matriz de covarianza o correlación en sus vectores y valores propios.
4. Selección de componentes principales: Se seleccionan los componentes principales basados en los valores propios, que indican la varianza explicada por cada componente.
5. Transformación de datos: Los datos originales se transforman en el espacio de los componentes principales seleccionados.
6. Interpretación y análisis: Se interpreta y analiza la varianza explicada por cada componente principal para comprender los patrones subyacentes en los datos.



Ejemplo en PYTHON



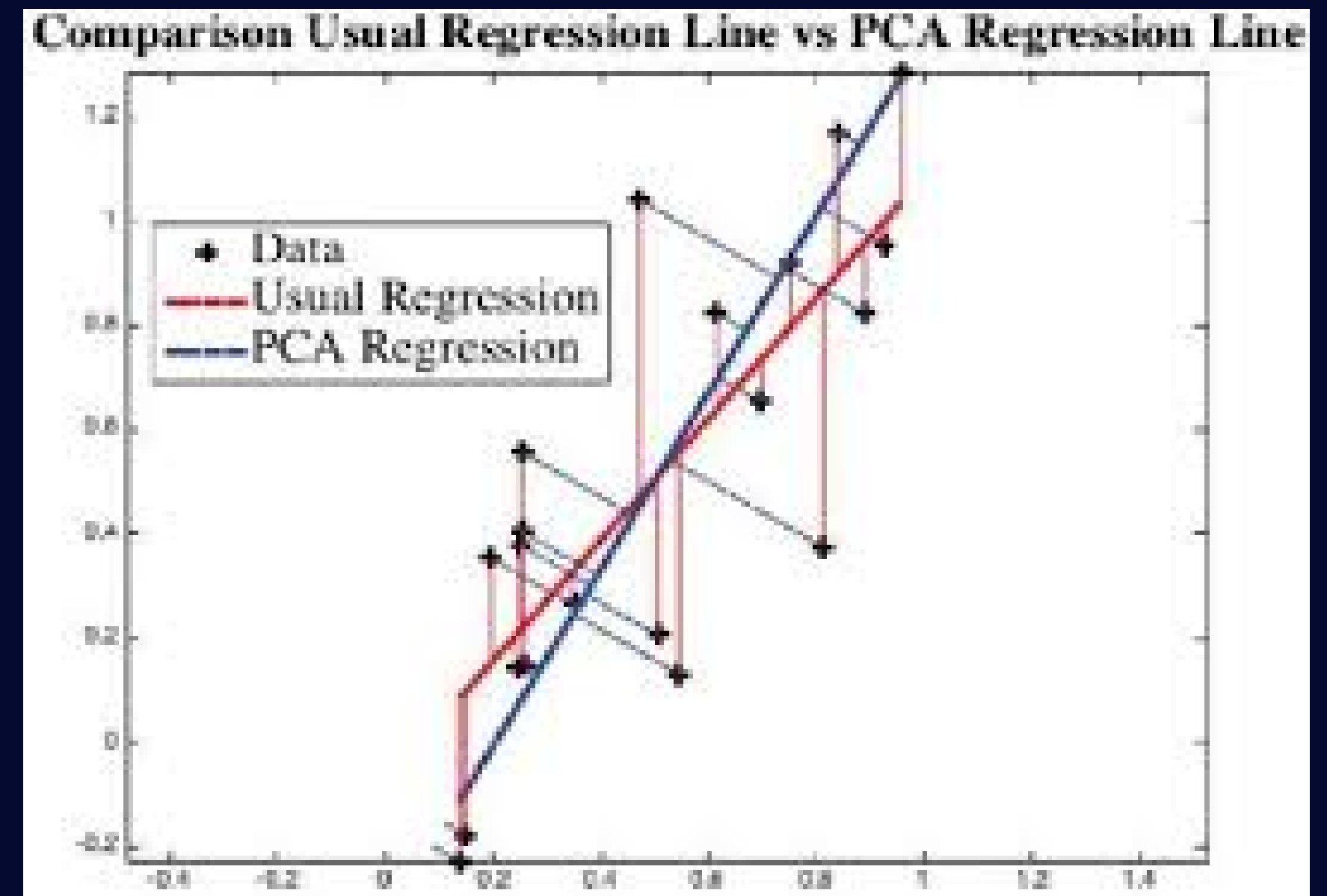
PCA CON REGRESIÓN LINEAL

- Consiste en ajustar un modelo de regresión lineal por mínimos cuadrados empleando como predictores las componentes generadas a partir de un Principal Component Analysis
- Dado que el PCA es útil eliminando información redundante, si se emplean como predictores las componentes principales se puede mejorar el modelo de regresión



La identificación del número óptimo de componentes principales que se emplean como predictores en PCR puede identificarse por *cross validation*.

El método PCR no deja de ser un ajuste lineal por mínimos cuadrados que emplea componentes principales como predictores, para que sea válido se tienen que cumplir las condiciones requeridas para la regresión por mínimos cuadrados.



EJEMPLO PRÁCTICO

https://cienciadedatos.net/documentos/35_principales_component_analysis#PCR:_PCA_aplicado_a_regresión_lineal

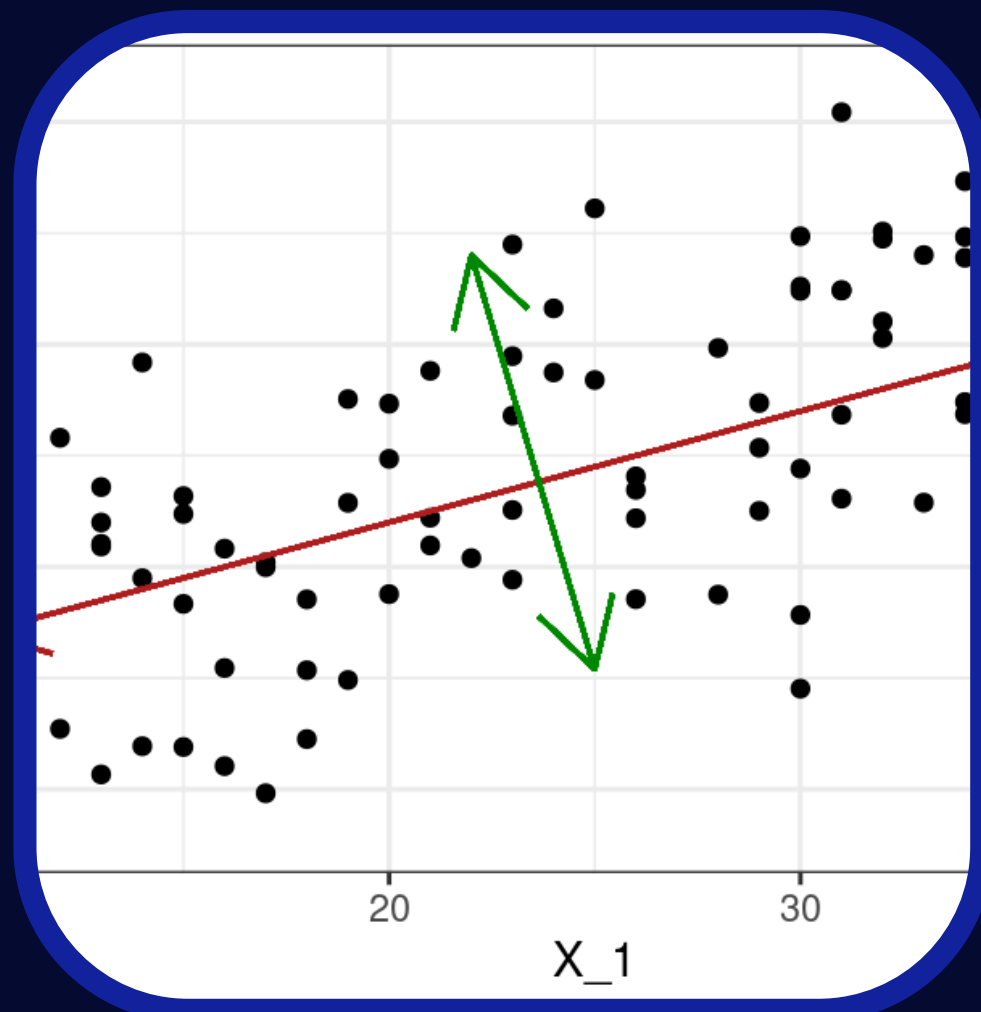
CONSIDERACIONES Y LIMITACIONES

Algunas consideraciones y limitaciones a tener en cuenta al implementar el PCA con Regresión son:

- Pérdida de interpretabilidad de las variables originales.
- Pérdida de información.
- Selección del número de componentes.
- Suposiciones de linealidad.
- Interpretación de resultados de Regresión.

Pérdida de interpretabilidad de las variables originales

Al reducir la información mediante el PCA, las variables originales se transforman en componentes principales que puede ser más difíciles de interpretar, en términos del problema real que se está modelando



Posible pérdida de información

Se busca capturar la mayor variabilidad en los datos en un número reducido de componentes principales, pero esto puede llevar a una pérdida de información relevante que puede ser importante al utilizar la regresión.

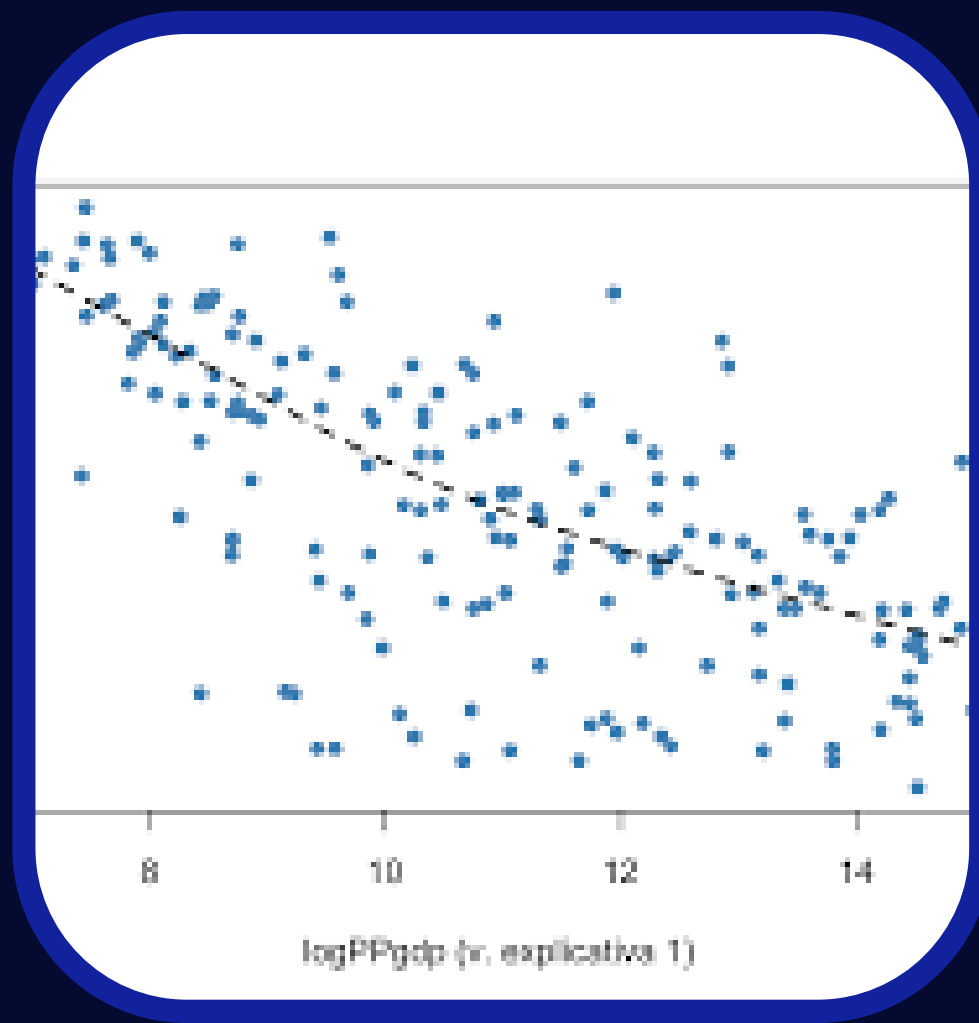


Selección del número de componentes

La elección del número de componentes principales requiere de técnicas (análisis de la varianza explicada o métodos de *validación cruzada*) para determinar un número óptimo.

Suposición de linealidad

Se asume linealidad en la relación entre variables, lo cual puede no ser válido en todos los casos, especialmente en problemas con relaciones no lineales.



Interpretación de los resultados de regresión

La interpretación de los coeficientes de regresión puede ser más compleja debido a la transformación de las variables originales en componentes principales.



NOISISTIC CONCO

- Reducción de dimensionalidad: ayuda a reducir el número de variables en un conjunto de datos.
- Conservación de información: Los componentes principales están ordenados por la cantidad de varianza que explican.
- **Mejora del rendimiento del modelo:** el PCA puede mejorar el rendimiento de los modelos de regresión lineal al reducir la complejidad del modelo y aumentar su capacidad de generalización.

ANÁLISIS DE COMPONENTES PRINCIPALES

- **Facilita la interpretación de los resultados:** Ayuda a identificar las variables más importantes para predecir la variable objetivo.
- **Desventajas:** Sensibilidad a valores atípicos en los datos, la interpretación de los componentes principales puede no ser siempre intuitiva y pérdida de datos.



GRACIAS