Análisis de Datos Económicos y Sociales de Países en la Década de los 90

Herramientas de Visualización de datos

Integrante	LU	Correo electrónico
Tomás Curti	327/19	tomasacurti@gmail.com
Agustin de Otazua	277/21	agusdeotazua@gmail.com

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2160 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (54 11) 4576-3359

http://www.fcen.uba.ar

1. Consigna

Vamos a realizar el ejercicio 4 de la práctica 2, pero únicamente analizando los casos sin escalar y usando logaritmo, también sin escalar.

El archivo paises_mundo.csv tiene indicadores económicos y sociales de 96 paıses en algún momento de la década de los 90. Las variables son la tasa de mortalidad infantil cada 1000 nacidos vivos (mortinf), producto nacional bruto (PNB), producción de electricidad (prod elec), consumo de energía per capita (cons energía) y emisión de CO2 per capita (CO2). Se tiene la siguiente hoja de ruta para realizar un análisis de los datos:

1. Realizar un scatterplot de pares. Comentar sobre la linealidad y de los datos y su ajuste con una distribución Normal.
2. Obtener y graficar los scores de PCA en sus dos primeras coordenadas. ¿Como se puede interpretar la ubicación de los países?
3. Calcular la proporción de variabilidad total acumulada por las dos primeras coordenadas de los scores.
4. Realizar un heatmap de la correlación muestral entre los scores y los datos. ¿Que variables son las que inciden mas en cada componente principal? ¿Tiene sentido calcular coeficientes de correlación?

Se propone recorrer esta ruta partiendo de tres escenarios distintos:

1. Los datos originales, sin estandarizarlos por columnas.
2. Tomar logaritmo natural de los datos y usarlos sin estandarizar por columnas.

2. Introducción

Este informe se centra en el análisis de datos económicos y sociales de 96 países en algún momento de la década de los 90. Las variables incluidas son la tasa de mortalidad infantil (mortinf), el producto nacional bruto (PNB), la producción de electricidad (prod elec), el consumo de energía per cápita (cons energia), y las emisiones de CO2 per cápita (CO2). Este análisis corresponde al ejercicio 2.4 de la Práctica 2, en el cual se propone una hoja de ruta detallada para explorar y comprender los datos, aplicando técnicas como el Análisis de Componentes Principales (PCA) y la evaluación de la influencia de las variables en las componentes principales. Además, se explora la variación en los resultados al considerar tres escenarios diferentes: datos originales sin estandarización, datos con logaritmo natural sin estandarización y datos con logaritmo natural y estandarización por columnas. Este análisis proporcionará una visión más completa de los patrones presentes en los indicadores económicos y sociales de los países en la década de los 90.

3. Desarrollo

3.1. Scatterplot y distribución

3.1.1. Sin escalamiento

Primero hacemos un scatterplot entre los 5 pares de variables.

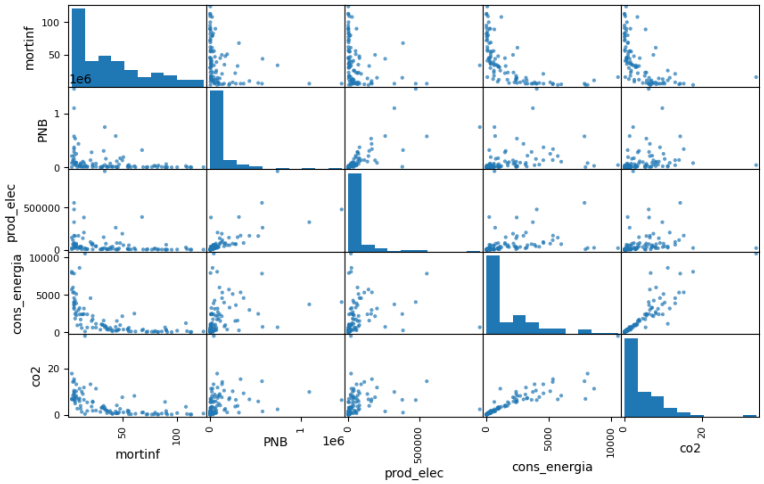

Figura 1: scatter plot entre las 5 variables (sin escalar).

Observamos que muy pocos pares de variables presentan una relación lineal. Uno de estos es (co2, cons_energia) y en menor medida (prod_elec, PNB).

Por otro lado, dado que en la diagonal de la figura se aprecian los histogramas de cada variable y estos son fuertemente asimétricos (sesgados a izquierda), podemos concluir que la distribución de dichas variables no es gaussiana.

3.1.2. Con logaritmo natural

Aplicamos logaritmo natural a los datos y vemos cómo quedan los scatterplots.

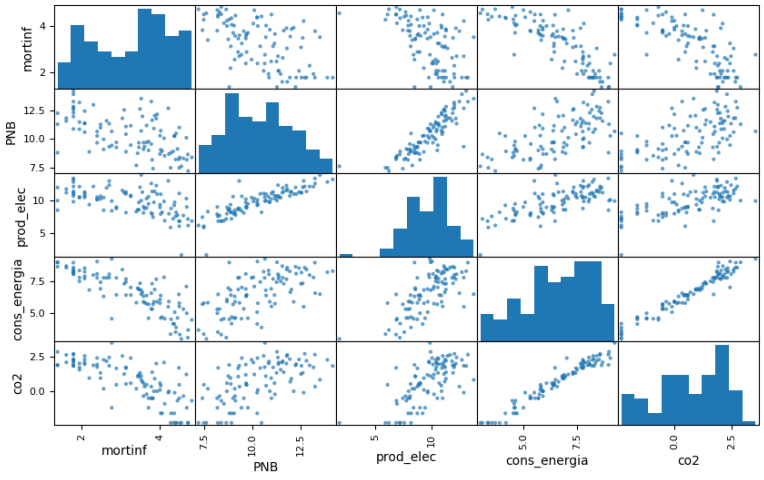

Figura 2: scatter plot entre las 5 variables (Aplicando logaritmo natural).

Ahora las relaciones entre las variables se ven mucho más lineales que en el caso anterior. Además, las distribuciones son más simétricas y se parece más a gaussianas.

3.2. PCA

3.2.1. Sin escalamiento

Estandarizamos las observaciones, obtuvimos las componentes principales muestrales mediante PCA y proyectamos dicha muestra a las mismas. Luego realizamos un scatterplot de las dos primeras coordenadas de cada observación, etiquetando el país asociado.

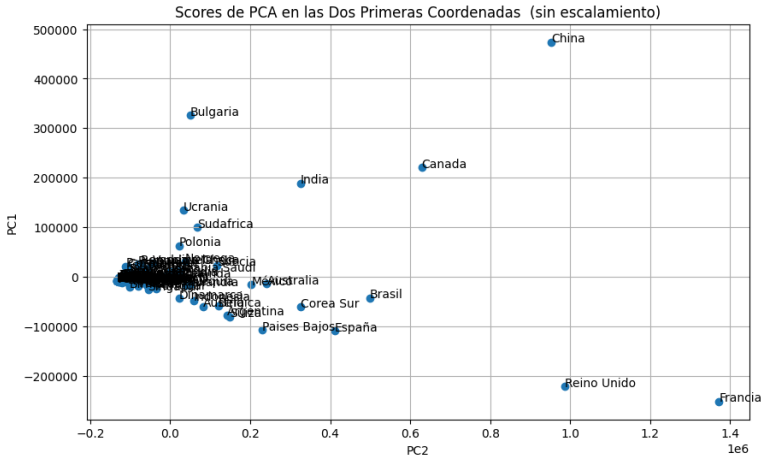

Figura 3: scatter plot de las 2 primeras coordenadas luego de aplicar PCA, sin escalar.

Lo primero que se aprecia es que los scores están acumulados en una región. Lo segundo es que las escalas de cada componente principal difieren notablemente.

Para entender mejor qué representa cada componente principal, realizamos un biplot.

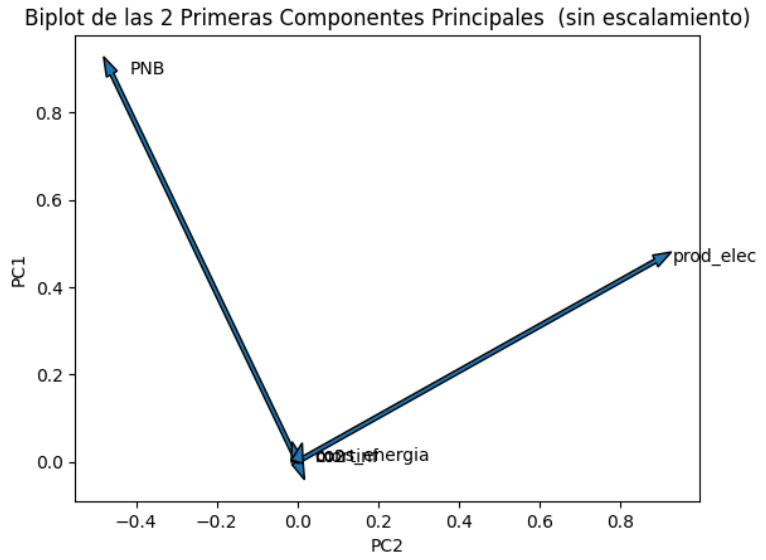

Figura 4: Biplot de las 2 primeras componentes principales, sin escalar.

Observamos que las variables PNB y prod_elec acaparan las mayores magnitudes, mientras que las flechas de las demás variables apenas se aprecian.

Dado que la primera componente principal tiene estas dos variables con score positivo, y siendo estas de magnitud mayor al resto de las variables, es razonable que den valores grandes los scores de

dicha componente. En contraposición, para la segunda componente principal estas dos variables tienen scores de signo opuesto, lo que explicaría que se cancelen y den valores pequeños.

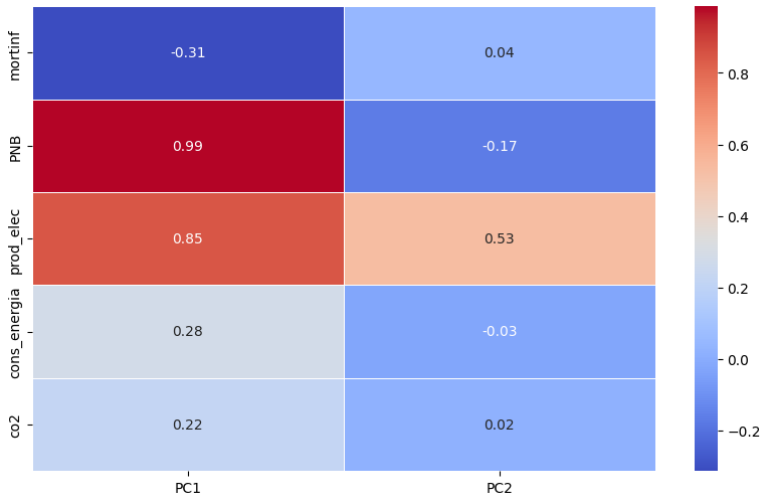

Figura 5: Correlación entre los scores y las variables originales, sin escalar.

Otra forma de verlo es calculando la varianza explicada por cada componente. Obtenemos que la primera explica un 90 % y la segunda casi un 10 %, por lo que la primera es notablemente más relevante. Cabe destacar que ambas explican prácticamente toda la variabilidad de los datos.

Por último realizamos una matriz de correlación entre los scores de estas dos componentes principales y la variables.

Nuevamente, vemos esta relación fuerte que existe con las variables PNB y prod_elec, especialmente en la primera componente principal, mientras que para las demás variables la correlación es baja.

3.2.2. Con logaritmo natural

Aplicamos logaritmo natural a los datos, tampoco escalamos y repetimos el procedimiento.

En el scatterplot ya no se observa un cúmulo claro, ni tampoco una componente tiene valores notablemente mayores que la otra. Esto está relacionado con que el logaritmo .^escaló"las variables más grandes, como PNB.

Realizando el biplot observamos que ahora las variables que antes eran poco significativas ahora tienen la misma representatividad que las mayores.

Si consideramos que las variables están más o menos a la misma escala aplicando logaritmo, se prevee que los valores de la componente 1 sean mayores en magnitud que los de la segunda ya que, excepto la de mortinf, tiene las flechas apuntando en la misma dirección, mientras que en el otro están en sentidos opuestos.

Lo volvemos a ver en las varianzas explicadas. La primera explica un 79 % y la segunda casi un 15 %.

Una diferencia con respecto al caso sin escalar, es que la matriz de correlación entre los scores y las variables es más homogénea en cuanto a magnitudes.

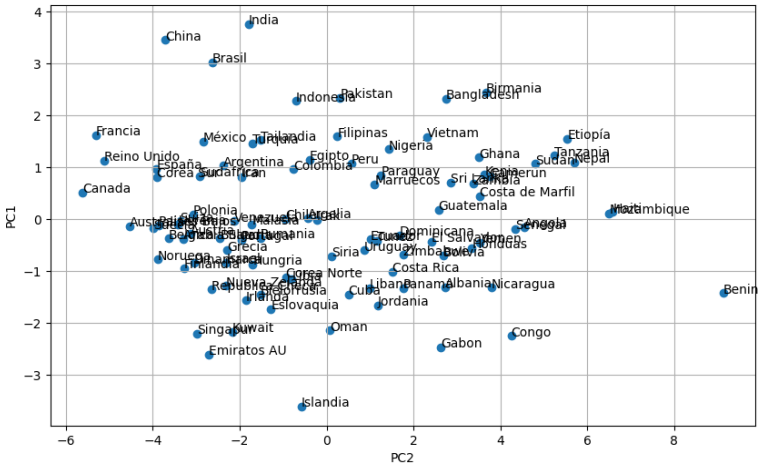

Figura 6: scatter plot de las 2 primeras coordenadas luego de aplicar PCA, con logaritmo y sin escalar.

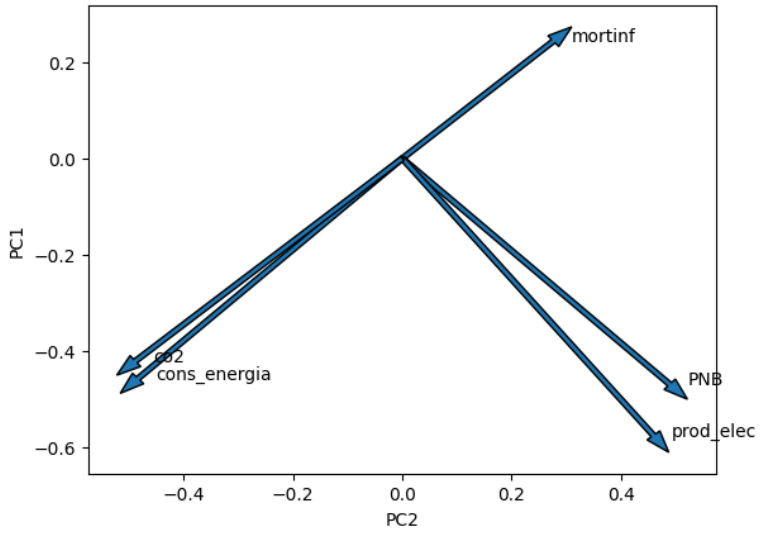

Figura 7: Biplot de las 2 primeras componentes principales, con log y sin escalar.

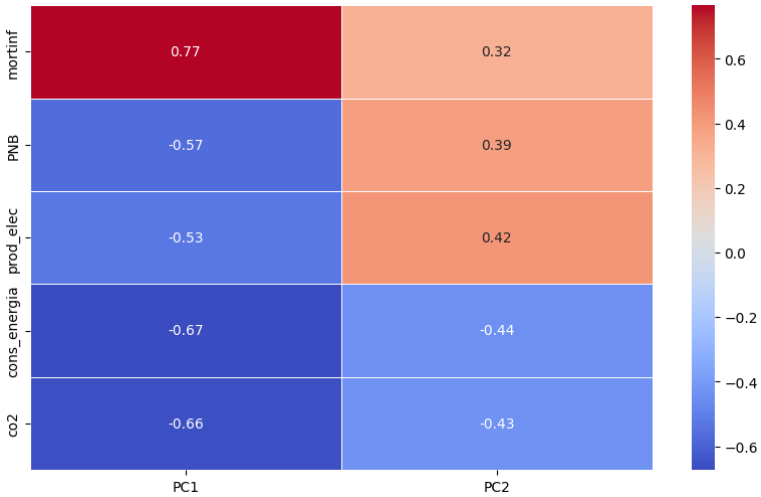

Figura 8: Correlación entre los scores y las variables originales, con log y sin escalar.