

# Trabajo práctico 2

Luciana Gattuso

---

## Ejercicio 1

$$S = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

A partir de la matriz S de covarianzas se calculan los autovalores con la función Eigen de Rstudio arrojando los siguientes resultados:

$$\lambda_1 = 6, \lambda_2 = 3, \lambda_3 = 6$$

Los autovectores

$$V_1 = (-0.408, -0.408, -0.816) \quad V_2 = (-0.577, -0.577, -0.577) \quad V_3 = (7.071, -7.071, -1.110)$$



Las variables Y de las componentes principales

$$Y_1 = (-0.408 X_1, -0.408 X_2, -0.816 X_3)$$

$$Y_2 = (-0.577 X_1, -0.577 X_2, -0.577 X_3)$$

$$Y_3 = (7.071 X_1, -7.071 X_2, -1.110 X_3)$$

Sobre la varianza explicada como la división de cada autovalor sobre la suma de los autovalores:

*Fórmula a modo de ejemplo: `variabilidad <- A$values[1]/(A$values[1]+A$values[2]+A$values[3])`*

$$Y_1 = 0.545$$

$$Y_2 = 0.272$$

$$Y_3 = 0.181$$



## Ejercicio 2

1. Analizar la distribución de las variables, utilizar gráficos scatterplot y boxplot.

Se analiza el conjunto de datos de 19 países europeos con las variables: área, Producto Bruto Interno, tasa de inflación, esperanza de vida, gasto militar, crecimiento de la población y la tasa de desempleo. En las siguientes figuras se visualizan diagramas de caja para cada una de las variables con el objetivo de entender la distribución de los datos y detectar aquellas que presentan valores atípicos.

Figura 1

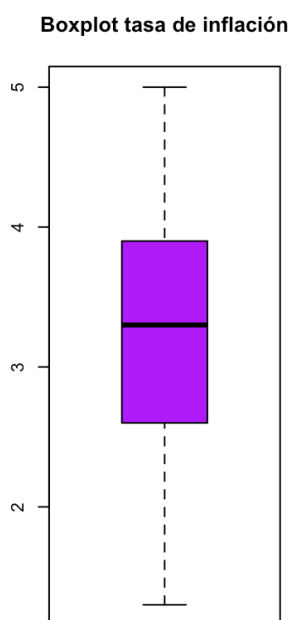


Figura 2

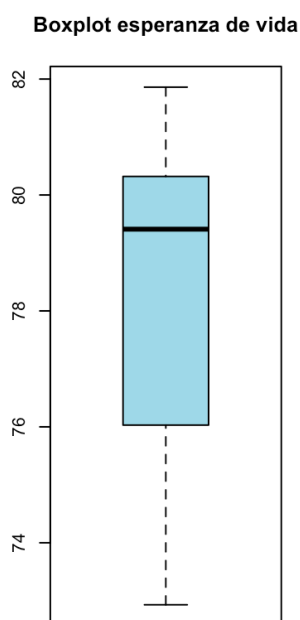
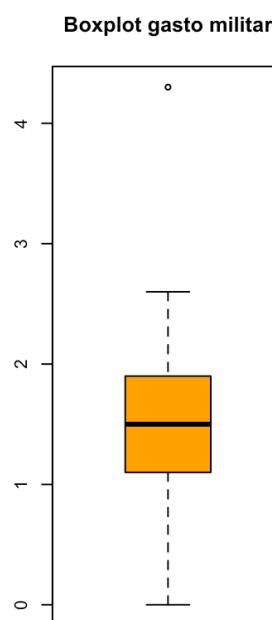


Figura 3



Si bien los datos de las variables se encuentran en un rango de valores distintos es posible compararlas según las asimetrías de las distribuciones. Por ejemplo, en el caso de la tasa de inflación y el gasto militar la distribución es simétrica a diferencia de la esperanza de vida que presenta una asimetría negativa con valores concentrados debajo de la mediana. Indicando que la mayoría de los países se concentran en un mismo nivel de inflación y de gasto militar, pero a nivel de esperanza de vida los datos están más dispersos. Se puede visualizar un solo outlier en gasto militar que corresponde a Grecia



Figura 4

Boxplot crecimiento de la población

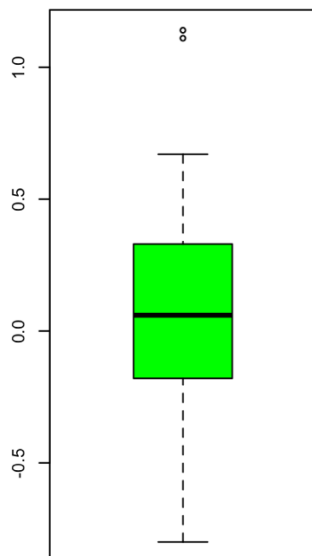
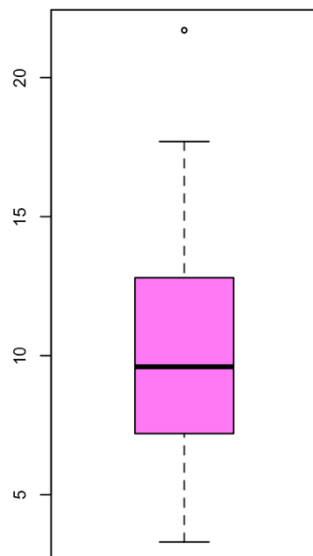


Figura 5

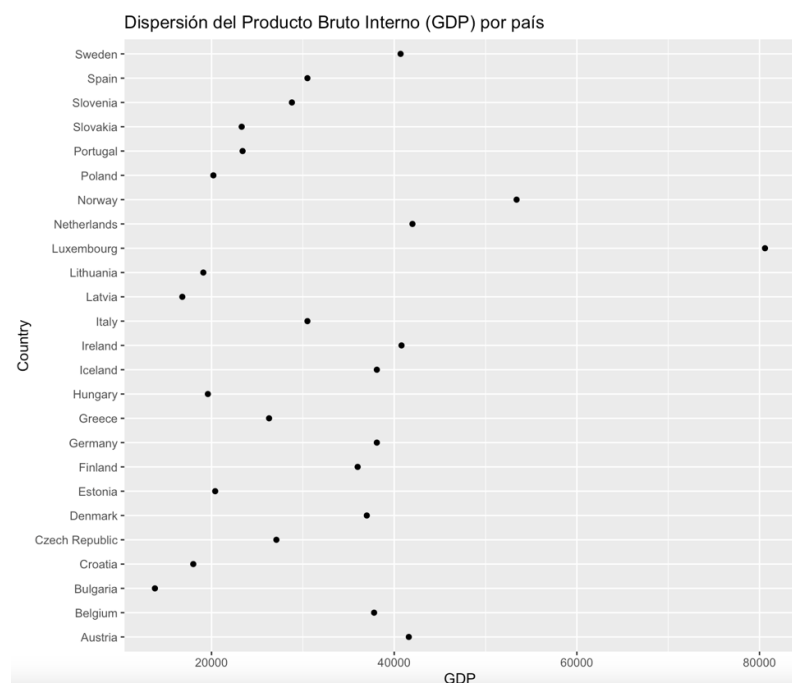
Boxplot tasa de desempleo



Respecto del crecimiento de la población cuenta con una distribución simétrica y la tasa de desempleo asimétrica positiva. Luxemburgo e Irlanda tienen un crecimiento poblacional muy superior al resto de los países de Europa representados como outliers. España es el valor atípico de la tasa de desempleo en el boxplot.

Se seleccionó el Producto Bruto Interno para graficar la dispersión de los datos por país.

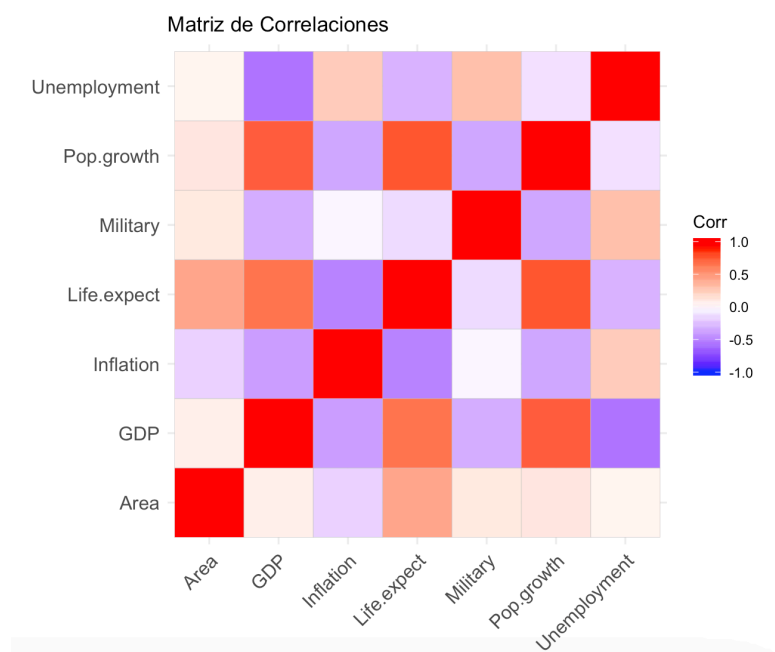
Figura 6



En este caso, los datos se concentran entre los valores 2000 a 4000 con un solo país aislado - Luxemburgo - que supera los 6000.

Previo al análisis de componentes principales se continuó con la exploración de los datos utilizando una matriz de correlaciones y se calculó el determinante de la correlación (0.02) indicando una alta correlación entre las variables.

Figura 7



Puede observarse que las variables crecimiento demográfico y esperanza de vida están muy correlacionadas con el PBI (o crecimiento económico), así como el crecimiento demográfico con la esperanza de vida. El gráfico también muestra una correlación negativa entre la tasa de desempleo y el PBI.

2. Realice un análisis de PCA utilizando las matrices de covarianzas y de correlaciones.
3. Calcular la varianza explicada por las componentes principales en ambos casos.
4. ¿Qué matriz piensa que es más apropiada para realizar el análisis?

El objetivo de este ejercicio es consolidar un índice basado en las variables (área, Producto Bruto Interno, tasa de inflación, esperanza de vida, gasto militar, crecimiento de la población y tasa de desempleo) que permita determinar qué país de Europa es el ideal para vivir. Se busca reducir el número de variables evitando la pérdida de información.

Utilizando la matriz de covarianzas, se realizó una transformación de componentes principales. La primera componente tiene una variabilidad tan alta que la proporción acumulada de la segunda componente es un número muy cercano a uno.

Figura 8

Importance of components:							
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.474e+05	1.424e+04	3.827	1.789	0.8188	0.6826	0.1757
Proportion of Variance	9.908e-01	9.250e-03	0.000	0.000	0.0000	0.0000	0.0000
Cumulative Proportion	9.908e-01	1.000e+00	1.000	1.000	1.0000	1.0000	1.0000

El uso de una matriz de correlación permite una estandarización de los datos y lograr resultados más representativos en las componentes principales. En este caso, es necesario de más de una componente principal para superar el 0.99 de varianza acumulada.

Figura 9

Importance of components:							
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.761	1.1754	0.9606	0.8832	0.73004	0.46295	0.2592
Proportion of Variance	0.443	0.1974	0.1318	0.1114	0.07614	0.03062	0.0096
Cumulative Proportion	0.443	0.6404	0.7722	0.8837	0.95979	0.99040	1.0000

Las varianzas de ambos casos están graficadas en las figuras 9 y 10. Está, a su vez, permite visualizar que la sobre representación de la primera componente principal en la figura 9. A diferencia de la matriz de correlación, que si bien la primer componente cuenta con una significativa variabilidad, el resto de las componentes también cuentan con un peso visual.



Figura 10

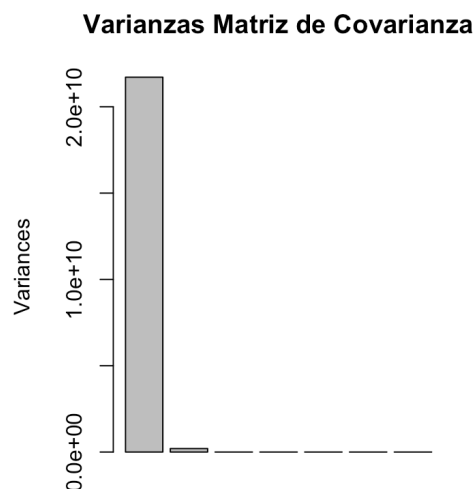
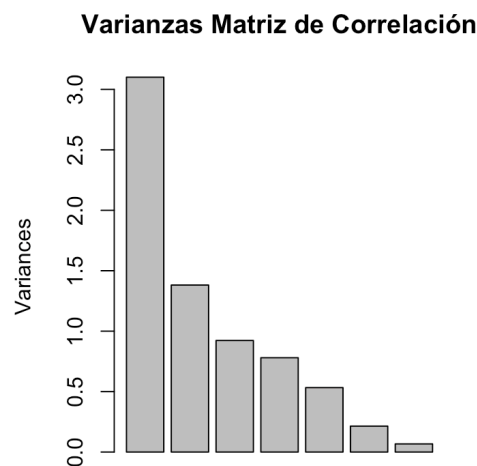


Figura 11



Se utilizó la matriz de covarianzas y se ordenaron los resultados en un nuevo dataframe para poder visualizar qué países tienen un índice más alto en la primera componente.

Figura 12

18	Greece	-18934.03
19	Italy	150471.38
20	Poland	161753.53
21	Norway	173072.25
22	Finland	187309.16
23	Germany	206198.58
24	Sweden	299485.67
25	Spain	354497.61


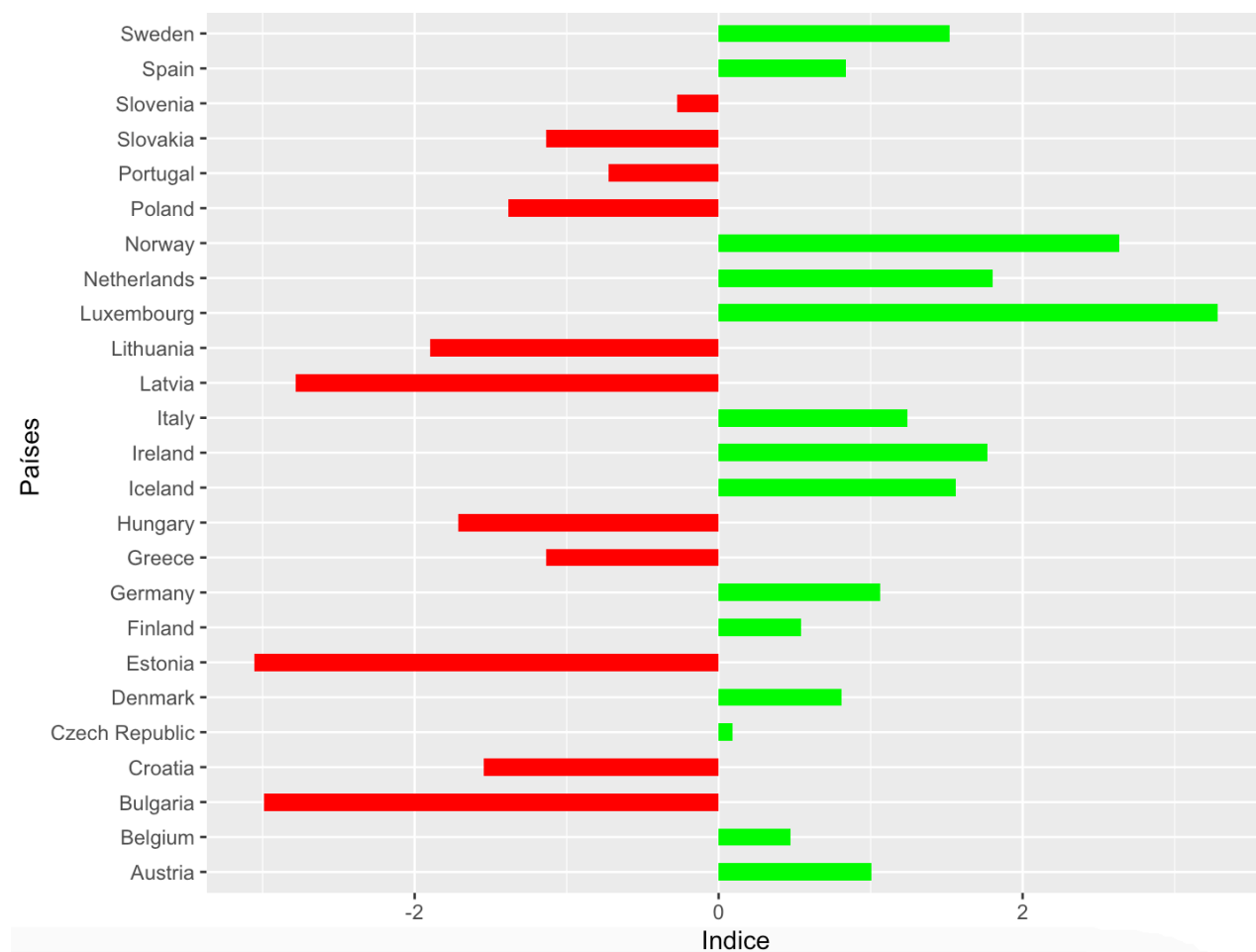
El resultado es llamativo, ya que posiciona como principal país para vivir a España, justamente el país - outlier- con más desempleo del conjunto. Entonces al utilizar la matriz de covarianza con una variable con una alta variabilidad, el resto de las componentes quedan relegadas y las conclusiones pueden ser engañosas. En cambio, estandarizando los datos o utilizando la matriz de correlación los resultados son distintos. 

Figura 13

18	Germany	1.0640489
19	Italy	1.2403741
20	Sweden	1.5189487
21	Iceland	1.5602432
22	Ireland	1.7661397
23	Netherlands	1.8056520
24	Norway	2.6376796
25	Luxembourg	3.2810475

Este nuevo índice posiciona a Luxemburgo como el país ideal para vivir del índice – en la figura 6 fue mencionado por tener el PBI más alto-. La matriz de correlación es más apropiada para realizar este tipo de análisis porque previene el sobre dimensionamiento de variables garantizando resultados veraces.

Figura 14

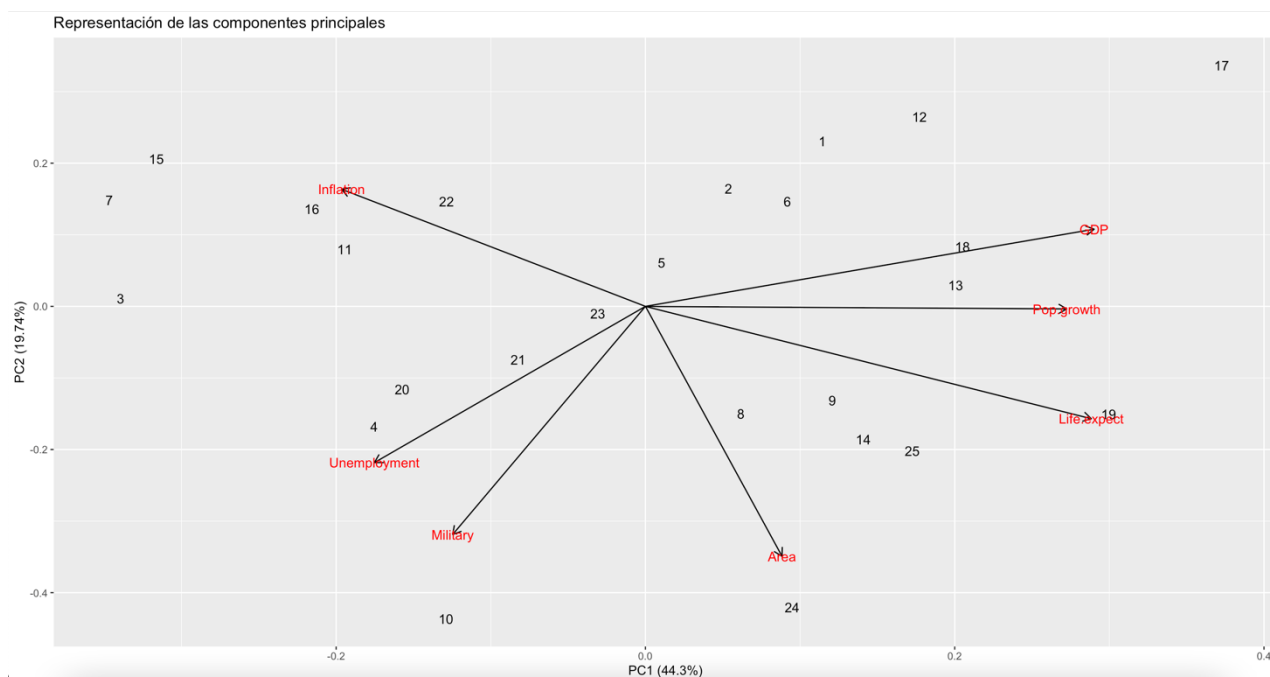


A modo de resumen, la figura 13 grafica el índice dado por la primera componente principal donde los países que superan el cero son recomendados para vivir y, aquellos con un índice menor a cero que figuran en rojo son desaconsejados para vivir.

1. Realizar el gráfico biplot, interpretar el gráfico y la primera componente principal.

Se graficó la representación de las componentes principales en un gráfico biplot. En el eje x, se visualiza la primera componente principal y en la y la segunda. Las flechas representan las variables, la variabilidad de los variables está graficada en la longitud de los vectores.

Figura 15



El conjunto de datos se encuentra ordenado por el índice de la primer componente principal, es decir que los países con número de índice más alto reflejan una mejor posición o son óptimos para vivir. La inflación y el desempleo se encuentran del lado izquierdo del gráfico en contraposición con el PBI, expectativa de vida y crecimiento poblacional. Asimismo, tomando valores negativos respecto a la PC1, pero positivos en la PC2 se encuentran Latvia (3) y Polonia (7). España (19) está bien posicionado respecto de la PC1 con un alto nivel de expectativa de vida, pero una alta tasa de desempleo. En el caso de Noruega (24) posicionado cerca del vector área y la PC1. Luxemburgo (25) se posiciona un poco más arriba en el gráfico reflejando un país de área más chica que Noruega, pero una mejor expectativa de vida – por su cercanía al vector-. Existen registros como Irlanda (22) que está bien posicionado en el índice sin embargo se posiciona negativamente frente a la PC1 por un nivel relativamente alto de inflación y de desempleo.

