Juan Menduiña

Maestría en Econometría - UTDT
Examen Final - Análisis Estadístico Multivariado

Este trabajo está basado en los Indicadores de Desarrollo Mundial que publica el Banco Mundial. Estos indicadores comprenden una selección de variables económicas, sociales y ambientales conformada a partir de información del Banco Mundial y otras agencias. La base de datos completa cubre más de 1400 indicadores referidos a más de 200 economías a partir del año 1960 y se actualiza de manera periódica. Se puede acceder a estos datos través de este link:

https://databank.worldbank.org/source/world-development-indicators#

El objetivo de este trabajo es aplicar herramientas del análisis multivariado para realizar una descripción completa que permita comprender las similitudes y diferencias entre los países latinoamericanos y las interrelaciones entre las características analizadas en términos de desarrollo. Se podrán consultar y utilizar los datos disponibles, directamente, desde la página web indicada más arriba o bien utilizar los datos de una consulta previa que abarca un período de 15 años recientes (años 2008 a 2022) y que se adjunta en formato CSV y Excel.

Ejercicio 1.

A partir de la base de datos disponible, elegir un subconjunto de 5 a 20 indicadores y un año de referencia para efectuar el análisis. Ésta será la matriz de datos que se utilizará durante todo el examen. Cualquier subconjunto de información es válido, no hay elecciones incorrectas; sin embargo, verificar que la cantidad de observaciones disponibles supere a la cantidad de variables elegidas.

Se seleccionaron los siguientes 10 indicadores para el año 2020:

- Access to electricity (% of population) - EG_ELC_ACCS_ZS.
- Age dependency ratio (% of working-age population) - SP_POP_DPND.
- CO2 emissions (metric tons per capita) - EN_ATM_CO2E_PC.
- Current account balance (% of GDP) - BN_CAB_XOKA_GD_ZS.
- GDP (constant 2015 US$) - NY_GDP_MKTP_KD.
- GDP per capita (constant 2015 US$) - NY_GDP_PCAP_KD.
- General government final consumption expenditure (% of GDP) - NE_CON_GOVT_ZS.
- Gross capital formation (% of GDP) - NE_GDI_TOTL_ZS.
- Individuals using the Internet (% of population) - IT_NET_USER_ZS.
- Life expectancy at birth, total (years) - SP_DYN_LE00_IN.

La cantidad de observaciones es 20:

- Antigua y Barbuda;
- Argentina;
- Bahamas;
- Belice;
- Bolivia;
- Brazil;
- Chile;
- Colombia;
- Costa Rica;
- República Dominicana;
- Ecuador;
- El Salvador;
- Guatemala;
- Haiti;
- Honduras;
- Jamaica;
- Mexico;
- Nicaragua;
- Paraguay;
- Perú.

Por lo tanto, $n= 20 > p= 10$, por lo que la cantidad de observaciones disponibles supera a la cantidad de variables elegidas.

Ejercicio 2.

Realizar un breve análisis descriptivo de la información disponible para el conjunto de variables y año elegidos. Calcular y describir distintas medidas de variabilidad global de los datos.

Media, varianza y coeficiente de variación:

Stats	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
Mean	-1.733542	95.45054	2.131974	69.47973	15.73852	21.02839	2.32e+11	7363.095	72.81105	50.18109
Variance	48.99248	145.0107	2.340354	276.2175	10.03417	77.42326	2.01e+23	2.70e+07	15.72179	44.86341
CV	-4.037665	.12616	.717561	.2392035	.201269	.4184366	1.931874	.7059046	.054457	.1334769

Por lo tanto, se puede observar que las variables “Current account balance (% of GDP)” y “Life expectancy at birth, total (years)” son las que tienen un mayor y menor coeficiente de variación, respectivamente. Ambas resultados son, relativamente, esperables, debido a la alta volatilidad que suele tener la cuenta corriente y a la, esperable, baja volatilidad que tiene la esperanza de vida.

Medidas de variabilidad global:

C= matriz de correlaciones; S= matriz de varianzas y covarianzas.

Varianza total C= 10.
Varianza media C= 1.
Varianza generalizada C= 0,00012383.
Varianza efectiva C= 0,40670729.

Varianza total S= 2,014e+23.
Varianza media S= 2,014e+22.
Varianza generalizada S= 1,696e+39.
Varianza efectiva S= 8373,9535.

Por lo tanto, se puede observar que, para el caso de la matriz de correlaciones, la varianza total (que es igual a la traza de la matriz de correlaciones) es igual a 10 y la varianza media (que es igual a la varianza total sobre la cantidad de variables) es igual a 1. Al no tener en cuenta la estructura de dependencia entre las variables, se computan la varianza generalizada (que es igual al determinante de la matriz de correlaciones) y la varianza efectiva (que es igual a la raíz p -variables- de la varianza generalizada), siendo esta última la que serviría para comparar conjuntos de datos con distinta cantidad de variables y que, en este caso, nos indica que hay cierta relación relevante entre estas variables.

Ejercicio 3.

Calcular la matriz de varianzas y covarianzas y la matriz de correlaciones muestrales e interpretar, brevemente, sus resultados.

Matriz de varianzas y covarianzas:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
BN_CAB_XOK~S	48.9925									
EG_ELC_ACC~S	-16.436	145.011								
EN_ATM_CO2~C	-8.52574	7.27781	2.34035							
IT_NET_USE~S	-69.2809	122.143	20.1715	276.217						
NE_CON_GOV~S	-5.60217	24.3135	1.15328	21.8565	10.0342					
NE_GDI_TOT~S	-36.3966	.277449	7.04133	61.5438	2.12191	77.4233				
NY_GDP_MKT~D	3.5e+11	9.2e+11	7.1e+10	1.8e+12	2.5e+11	-8.7e+11	2.0e+23			
NY_GDP_PCA~D	-28977.4	25054.4	7193.31	70761.1	3412.84	16813.2	3.0e+14	2.7e+07		
SP_DYN_LE0~N	-8.11475	27.2763	2.846	44.9305	4.71054	10.3208	1.4e+11	11872.2	15.7218	
SP_POP_DPND	31.1587	-33.3573	-6.38174	-81.6805	-10.6715	-25.6618	-6.0e+11	-21994.7	-14.4968	44.8634

Matriz de correlaciones muestrales:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
BN_CAB_XOK~S	1.0000									
EG_ELC_ACC~S	-0.1950	1.0000								
EN_ATM_CO2~C	-0.7962	0.3951	1.0000							
IT_NET_USE~S	-0.5956	0.6103	0.7934	1.0000						
NE_CON_GOV~S	-0.2527	0.6374	0.2380	0.4152	1.0000					
NE_GDI_TOT~S	-0.5910	0.0026	0.5231	0.4208	0.0761	1.0000				
NY_GDP_MKT~D	0.1107	0.1704	0.1029	0.2384	0.1764	-0.2198	1.0000			
NY_GDP_PCA~D	-0.7965	0.4003	0.9047	0.8191	0.2073	0.3676	0.1280	1.0000		
SP_DYN_LE0~N	-0.2924	0.5713	0.4692	0.6818	0.3750	0.2958	0.0812	0.5761	1.0000	
SP_POP_DPND	0.6646	-0.4136	-0.6228	-0.7337	-0.5030	-0.4354	-0.1998	-0.6318	-0.5459	1.0000

Por un lado, se puede observar cómo “Life expectancy at birth, total (years)” se correlaciona negativamente con todas las variables excepto con “Current account balance (% of GDP)”, mientras que ésta, además, se correlaciona negativamente con todas las variables excepto con “GDP (constant 2015 US$)”. Por otro lado, se observa cómo el resto de las variables se correlacionan positivamente entre sí, excepto entre “Individuals using the Internet (% of population)” y “GDP (constant 2015 US$)”.

Ejercicio 4.

Realizar un análisis de componentes principales utilizando tanto la matriz de varianzas y covarianzas como la matriz de correlaciones. Describir, de manera general, los resultados obtenidos en cada caso y comparar los resultados destacando diferencias y/o similitudes entre ambos casos. En particular, describir la proporción de varianza explicada por cada componente principal. Elegir los resultados de uno de estos análisis para continuar con la próxima consigna.

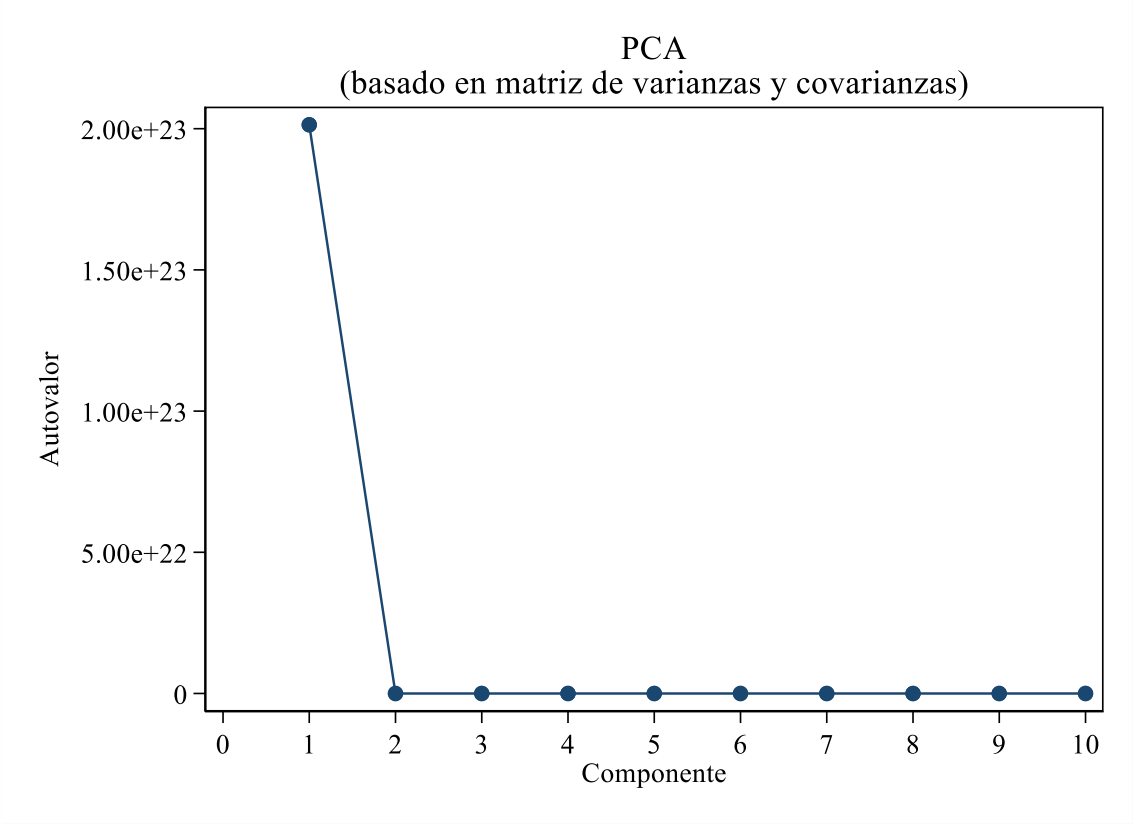
Basado en matriz de varianzas y covarianzas:

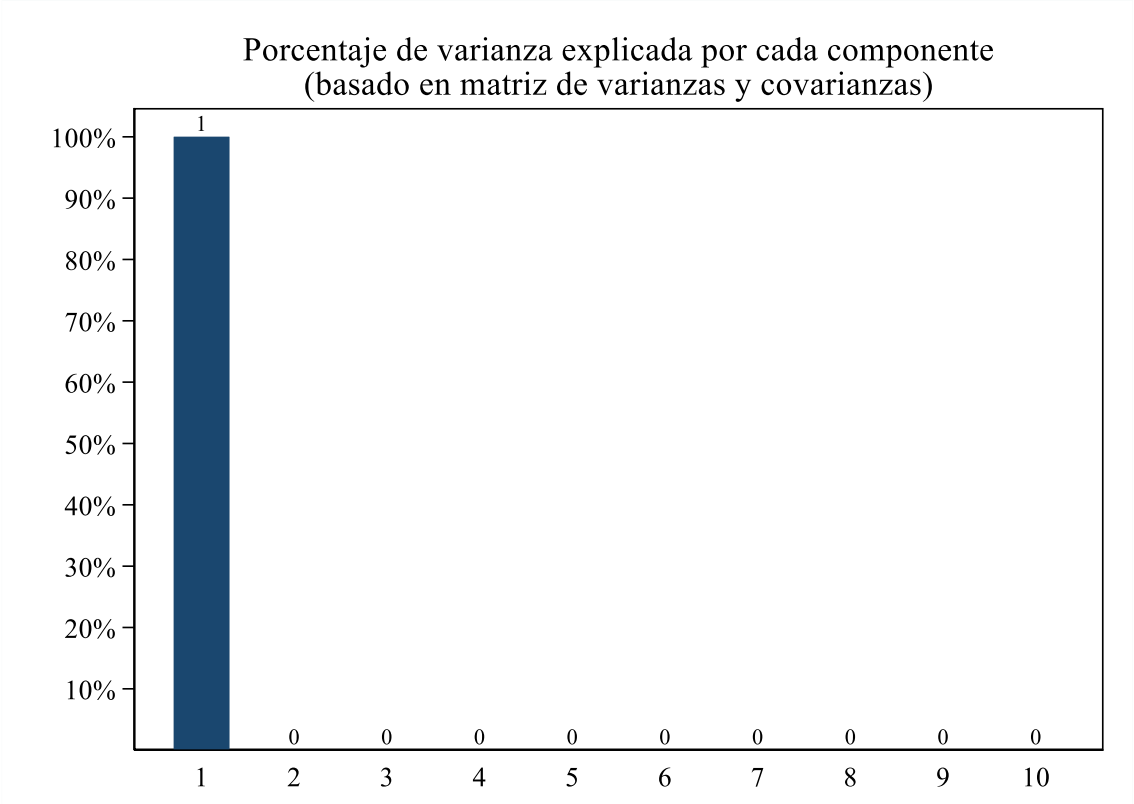
```
Principal components/covariance                 Number of obs    =         20
                                                Number of comp.  =          1
                                                Trace            =   2.01e+23
    Rotation: (unrotated = principal)           Rho              =     1.0000
```

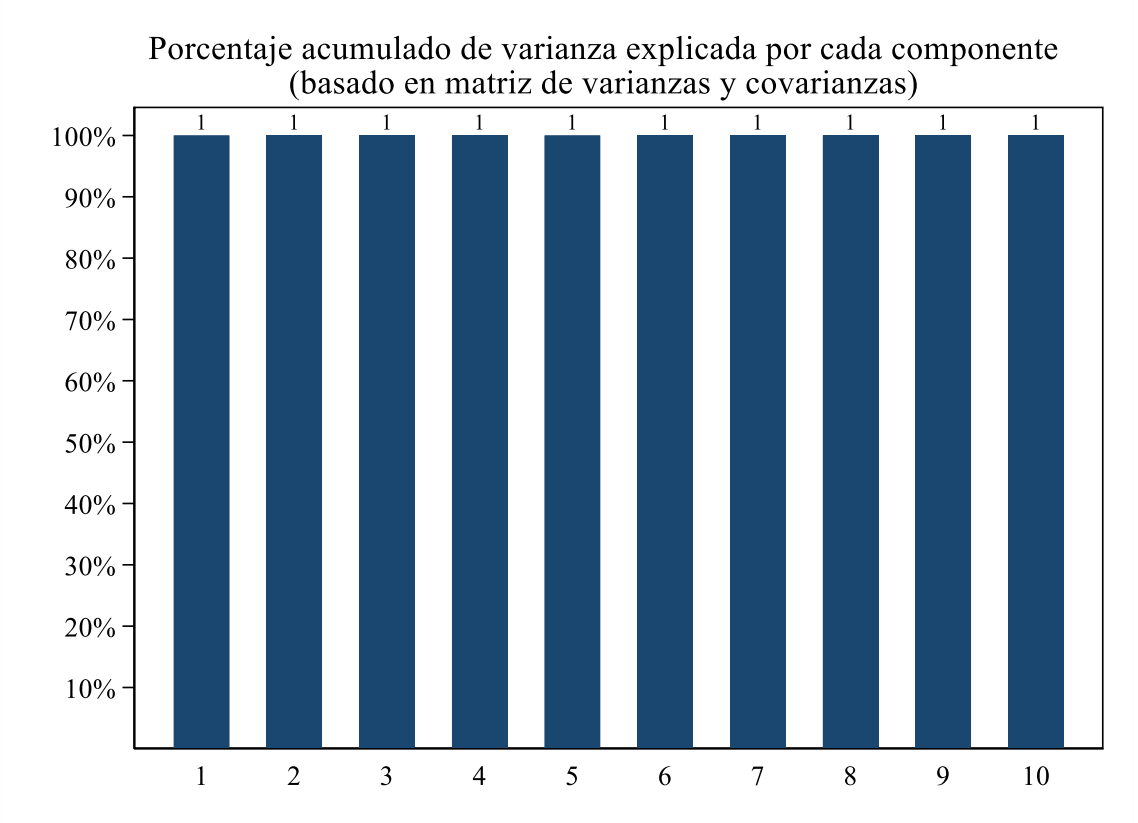
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.01402e+23	2.01402e+23	1.0000	1.0000
Comp2	0	0	0.0000	1.0000
Comp3	0	0	0.0000	1.0000
Comp4	0	0	0.0000	1.0000
Comp5	0	0	0.0000	1.0000
Comp6	0	0	0.0000	1.0000
Comp7	0	0	0.0000	1.0000
Comp8	0	0	0.0000	1.0000
Comp9	0	0	0.0000	1.0000
Comp10	0	.	0.0000	1.0000

Principal components (eigenvectors)

Variable	Comp1	Unexplained
BN_CAB_XOK~S	0.0000	48.39
EG_ELC_ACC~S	0.0000	140.8
EN_ATM_CO2~C	0.0000	2.316
IT_NET_USE~S	0.0000	260.5
NE_CON_GOV~S	0.0000	9.722
NE_GDI_TOT~S	-0.0000	73.68
NY_GDP_MKT~D	1.0000	33554432
NY_GDP_PCA~D	0.0000	26573043
SP_DYN_LE0~N	0.0000	15.62
SP_POP_DPND	-0.0000	43.07


PCA
(basado en matriz de varianzas y covarianzas)
Autovalor
2.00e+23
1.50e+23
1.00e+23
5.00e+22
0
0 1 2 3 4 5 6 7 8 9 10
Componente


Porcentaje de varianza explicada por cada componente
(basado en matriz de varianzas y covarianzas)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
1
0 0 0 0 0 0 0 0 0
1 2 3 4 5 6 7 8 9 10



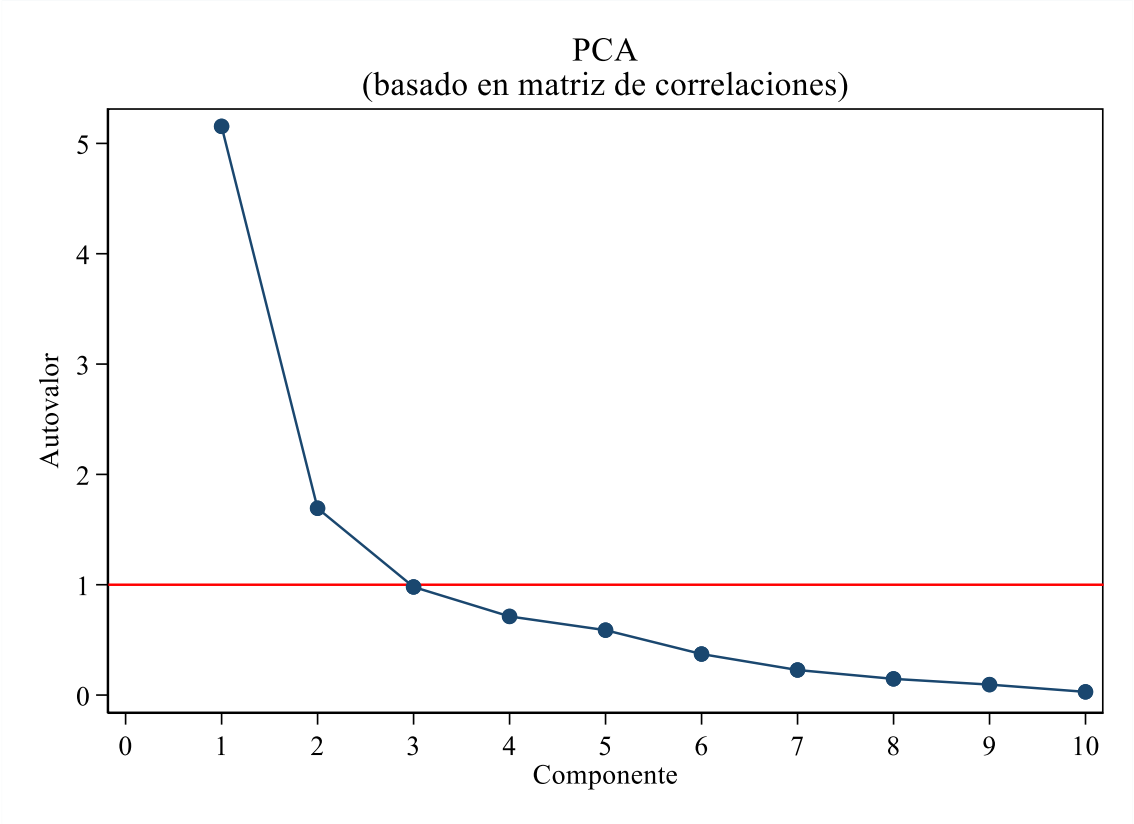
Por lo tanto, basando el análisis de componentes principales en la matriz de varianzas y covarianzas, toda la varianza es explicada por un solo componente principal, lo cual se debe a la, relativamente, alta varianza de la variable "GDP (constant 2015 US$)".

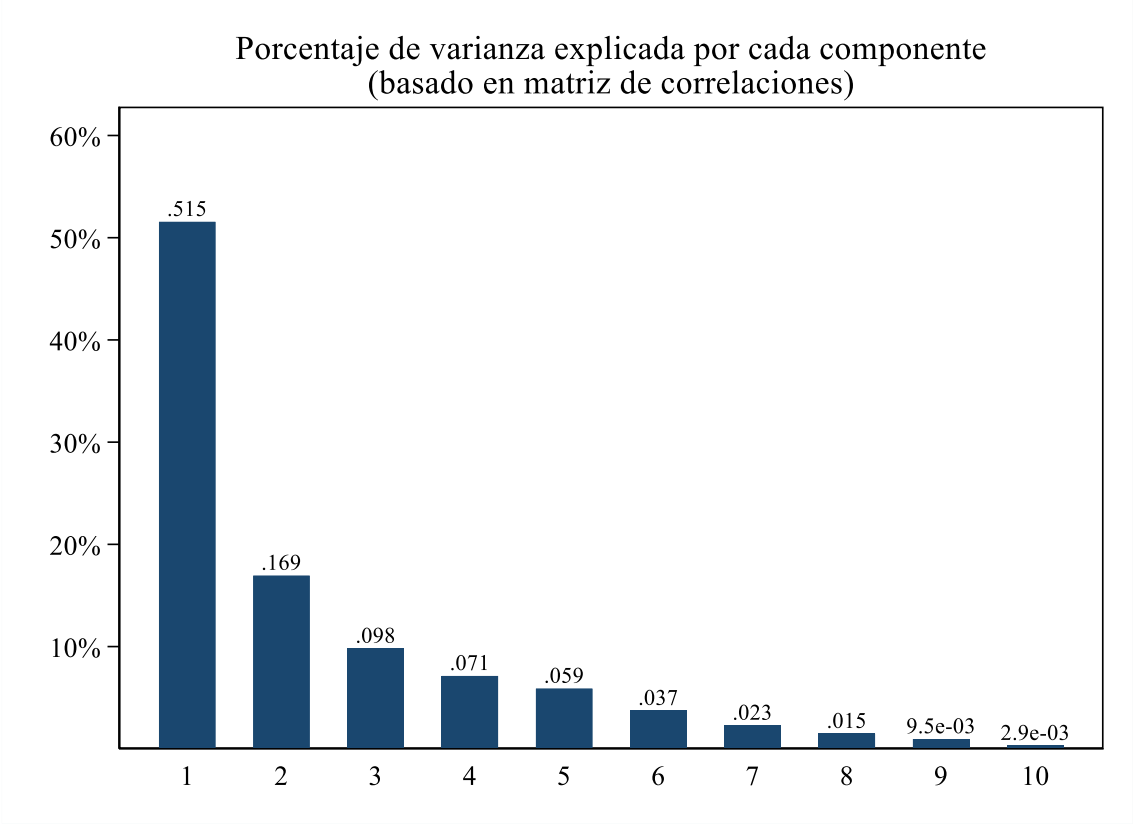
Basado en matriz de correlaciones:

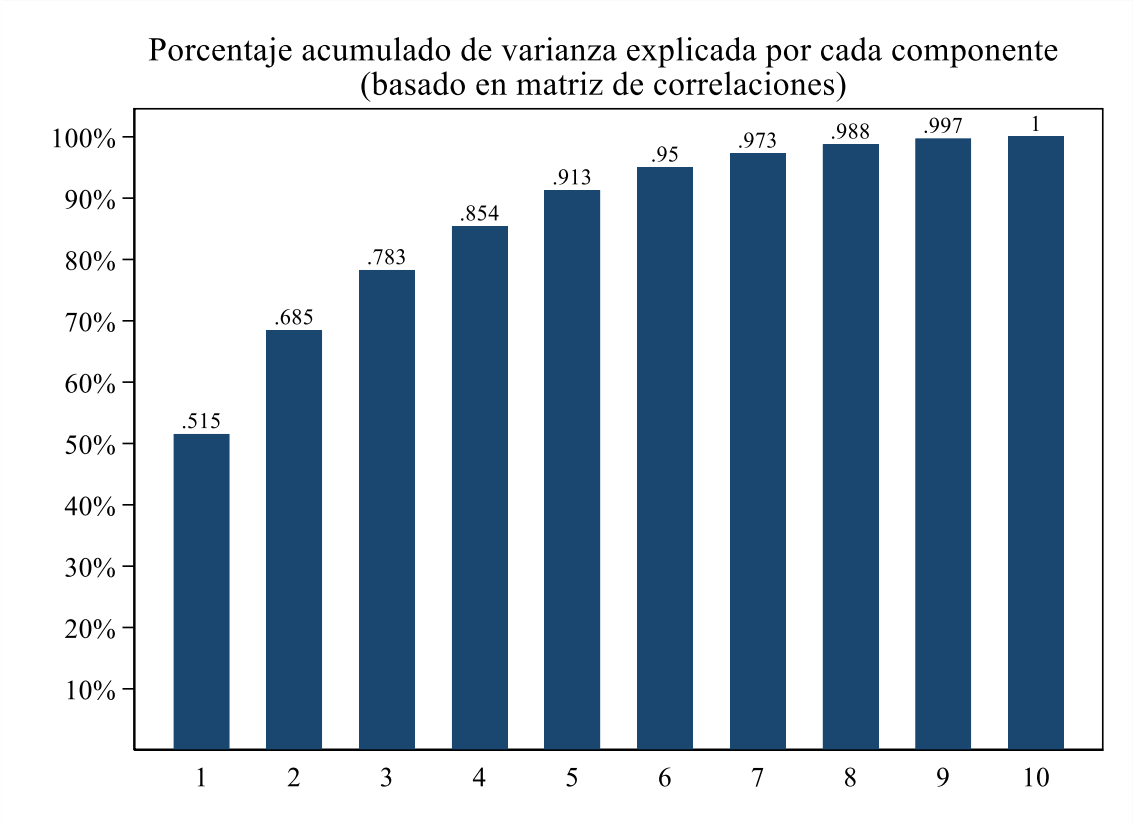
```
Principal components/correlation                  Number of obs    =         20
                                                  Number of comp.  =         10
                                                  Trace            =         10
    Rotation: (unrotated = principal)             Rho              =     1.0000

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      5.15483      3.46141             0.5155       0.5155
           Comp2 |      1.69343      .713389             0.1693       0.6848
           Comp3 |      .980039       .26705             0.0980       0.7828
           Comp4 |      .712989      .124529             0.0713       0.8541
           Comp5 |       .58846      .216419             0.0588       0.9130
           Comp6 |      .372041      .144929             0.0372       0.9502
           Comp7 |      .227112     .0803226             0.0227       0.9729
           Comp8 |      .146789     .0517453             0.0147       0.9876
           Comp9 |     .0950439      .065779             0.0095       0.9971
          Comp10 |     .0292649            .             0.0029       1.0000
    --------------------------------------------------------------------------
```

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10	Unexplained
BN_CAB_XOK~S	-0.3457	0.3523	-0.0310	0.2515	0.3384	0.1305	-0.1454	0.4319	0.4179	0.4223	0
EG_ELC_ACC~S	0.2670	0.4609	-0.2893	0.2048	-0.2269	0.4188	-0.3991	-0.4447	-0.0196	0.0802	0
EN_ATM_CO2~C	0.3883	-0.1899	0.2000	0.0925	-0.2274	0.2758	0.1107	0.1416	0.7221	-0.2883	0
IT_NET_USE~S	0.4058	0.0823	0.0973	0.1608	0.0846	0.1137	-0.3019	0.6479	-0.4437	-0.2556	0
NE_CON_GOV~S	0.2240	0.4265	-0.3959	-0.5386	-0.1443	0.0226	0.4644	0.2777	0.0230	0.0809	0
NE_GDI_TOT~S	0.2350	-0.4485	-0.2104	-0.2462	0.5766	0.4897	0.0081	-0.0962	-0.0489	0.2340	0
NY_GDP_MKT~D	0.0675	0.4278	0.7486	-0.2521	0.2835	0.1862	0.1393	-0.2170	-0.0744	-0.0376	0
NY_GDP_PCA~D	0.3896	-0.1321	0.2482	0.2608	-0.2762	-0.1085	0.2291	0.0275	-0.1389	0.7352	0
SP_DYN_LE0~N	0.3120	0.1898	-0.2068	0.5004	0.4821	-0.3191	0.3944	-0.1916	0.0412	-0.2129	0
SP_POP_DPND	-0.3670	-0.0337	-0.0130	0.3573	-0.1811	0.5745	0.5226	0.0701	-0.2780	-0.1352	0







Por lo tanto, basando el análisis de componentes principales en la matriz de correlaciones, los primeros cuatro componentes principales explican el 85,41% de toda la varianza. Considerando la varianza media (igual a 1), se seleccionarían los dos primeros

componentes principales (ya que estos tienen un autovalor mayor a 1), mientras que, considerando un umbral de 80% de varianza a explicar, se seleccionarían los cuatros primeros componentes principales.

Se elige el análisis de componentes principales basado en la matriz de correlaciones para continuar con la siguiente consigna y se sugiere extraer los dos primeros componentes principales.

Ejercicio 5.

Tomando en cuenta las primeras dos componentes principales del análisis elegido en el ejercicio anterior, calcular los coeficientes de correlación de cada componente principal con respecto a cada variable original. Interpretar la representación de las primeras 2 componentes principales en términos de su correlación con las variables que las definen.

Coeficientes de correlación de cada componente principal con respecto a cada variable original:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D	u_1	u_2
BN_CAB_XOK~S	1.0000											
EG_ELC_ACC~S	-0.1950	1.0000										
EN_ATM_CO2~C	-0.7962	0.3951	1.0000									
IT_NET_USE~S	-0.5956	0.6103	0.7934	1.0000								
NE_CON_GOV~S	-0.2527	0.6374	0.2380	0.4152	1.0000							
NE_GDI_TOT~S	-0.5910	0.0026	0.5231	0.4208	0.0761	1.0000						
NY_GDP_MKT~D	0.1107	0.1704	0.1029	0.2384	0.1764	-0.2198	1.0000					
NY_GDP_PCA~D	-0.7965	0.4003	0.9047	0.8191	0.2073	0.3676	0.1280	1.0000				
SP_DYN_LE0~N	-0.2924	0.5713	0.4692	0.6818	0.3750	0.2958	0.0812	0.5761	1.0000			
SP_POP_DPND	0.6646	-0.4136	-0.6228	-0.7337	-0.5030	-0.4354	-0.1998	-0.6318	-0.5459	1.0000		
u_1	-0.7849	0.6062	0.8816	0.9213	0.5085	0.5336	0.1532	0.8847	0.7083	-0.8332	1.0000	
u_2	0.4584	0.5997	-0.2471	0.1071	0.5550	-0.5836	0.5567	-0.1719	0.2469	-0.0439	0.0000	1.0000

Por lo tanto, se puede observar que el primer componente principal se correlaciona positivamente con todas las variables excepto con “Age dependency ratio (% of working-age population)” y “Current account balance (% of GDP)”, mientras que el segundo componente principal se correlaciona positivamente con “Current account balance (% of GDP)”, “Access to electricity (% of population)”, “Individuals using the Internet (% of population)”, “General government final consumption expenditure (% of GDP)”, “GDP (constant 2015 US$)” y “Life expectancy at birth, total (years)”, y negativamente con “CO2 emissions (metric tons per capita)”, “Gross capital formation (% of GDP)”, “GDP per capita (constant 2015 US$)” y “Age dependency ratio (% of working-age population)”.

En resumen, se puede decir que:

- el primer componente principal tomará valores altos en aquellos países en donde las variables “Age dependency ratio (% of working-age population)” y “Current account balance (% of GDP)” resulten menos importantes, en términos relativos, a las restantes; y
- el segundo componente principal tomará valores altos en aquellos países en donde las variables que ponderan con un signo positivo en el autovector resulten más importantes, en términos relativos, a las restantes.

Ejercicio 6.

Calcular el valor de las primeras 2 componentes principales para cada una de las observaciones (países). Realizar un breve análisis de estadística descriptiva de cada una de ellas.

Componente principal 1:

```
                    Scores for component 1
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -4.727859      -4.727859
 5%    -3.682592      -2.637325
10%    -2.368377       -2.09943       Obs                  20
25%    -1.382419      -2.093045       Sum of wgt.          20

50%    -.1675797                      Mean           -2.79e-09
                        Largest       Std. dev.       2.270426
75%      1.08631       1.302309
90%     3.427007        2.32834       Variance        5.154834
95%     4.689173       4.525674       Skewness        .3488936
99%     4.852673       4.852673       Kurtosis        3.349865
```

Por lo tanto, se puede observar que el menor valor de este componente es -4,727859 (correspondiente a Haití, que tiene valores relativamente bajos en las variables que ponderan con un signo positivo), mientras que el mayor valor de este componente es 4,852673 (correspondiente a Antigua y Barbuda, que tiene valores relativamente bajos en las variables que ponderan con un signo negativo).

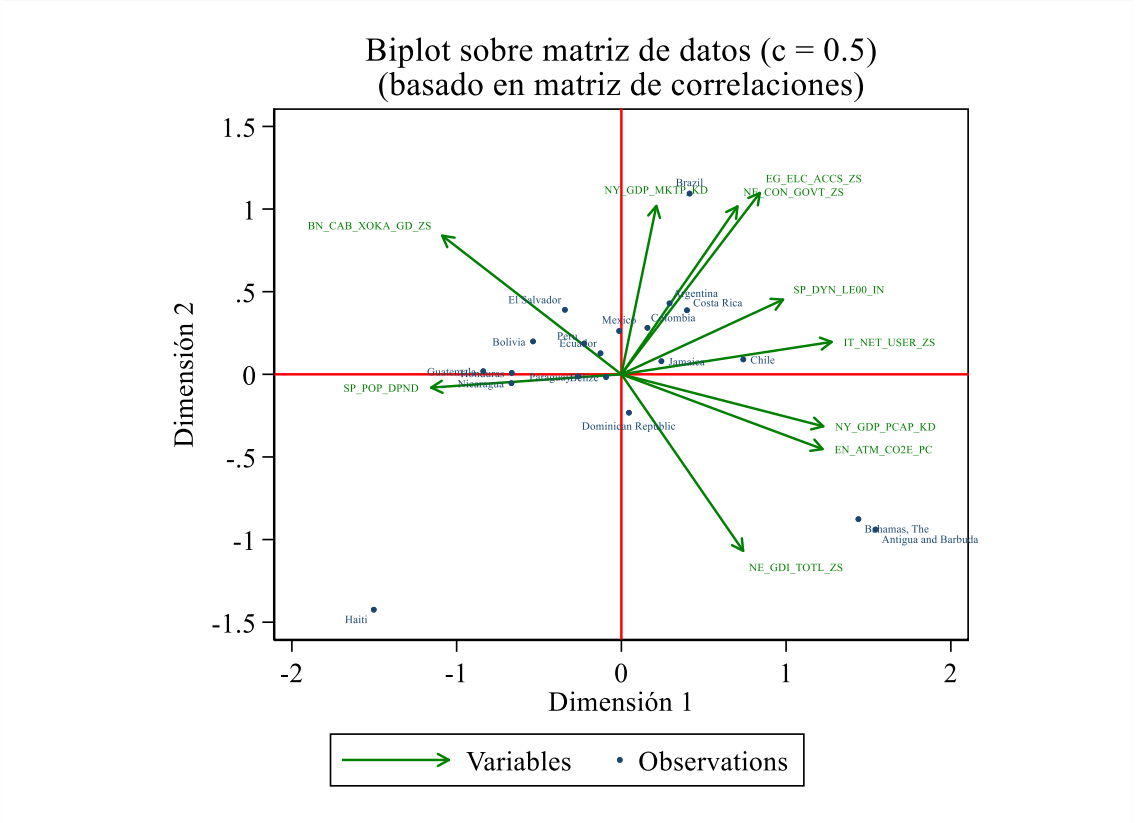
Componente principal 2:

```
                    Scores for component 2
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -3.391772      -3.391772
 5%    -2.814482      -2.237192
10%    -2.162097      -2.087002       Obs                  20
25%    -.0837795      -.5527226       Sum of wgt.          20

50%     .2018057                      Mean           -2.05e-09
                        Largest       Std. dev.       1.301318
75%     .6475806       .9231878
90%     .9760195       .9294568       Variance        1.693428
95%     1.813812       1.022582       Skewness       -.9398758
99%     2.605042       2.605042       Kurtosis        4.357151
```

Por lo tanto, se puede observar que el menor valor de este componente es -3,391772 (correspondiente a Haití, que tiene valores relativamente bajos en las variables que ponderan con un signo positivo), mientras que el mayor valor de este componente es 2,605042 (correspondiente a Brasil, que tiene valores relativamente bajos en las variables que ponderan con un signo negativo).

Ejercicio 7.

Realizar un gráfico biplot para la representación conjunta de filas y columnas (equivalente a observaciones y variables) de la matriz de datos y describir los resultados obtenidos en términos de la distribución de las observaciones (países) en el subespacio de las primeras dos componentes.

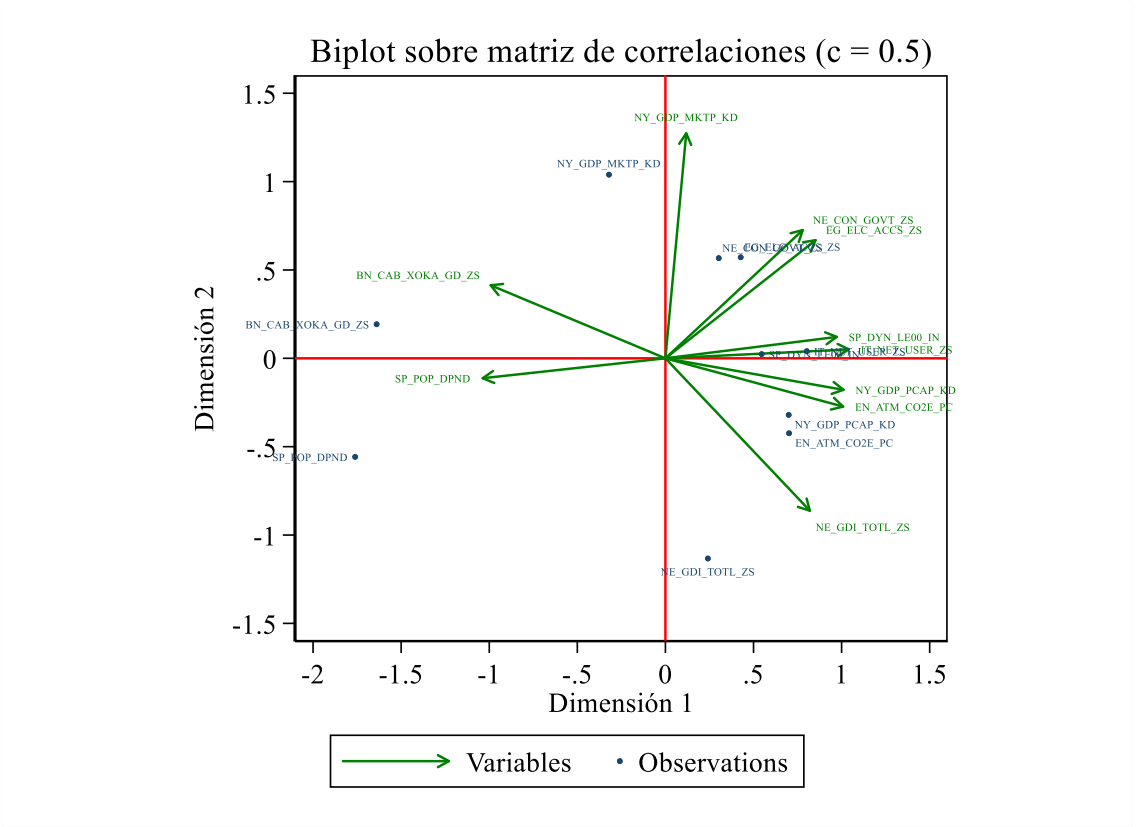


Por lo tanto, se puede observar que:

- Haití tiene un valor negativo muy alto en ambos componentes principales;
- Antigua y Barbuda y Bahamas tiene un valor positivo muy alto en el componente 1 y negativo muy alto en el componente 2;
- Brasil tiene un valor positivo muy alto en el componente 2 y un valor positivo no muy alto en el componente 1; y
- el resto de los países tienen valores más cercanos al origen de este subespacio de estas primeras dos componentes.

Ejercicio 8.

Utilizando la representación biplot adecuada, calcular una aproximación de dimensión 2 para la matriz de varianzas y covarianzas (o matriz de correlaciones, según corresponda) original e interpretar los resultados obtenidos.



Por lo tanto, se puede observar que la dirección y la longitud de las flechas coinciden, en gran medida, con el valor de la observación correspondiente.

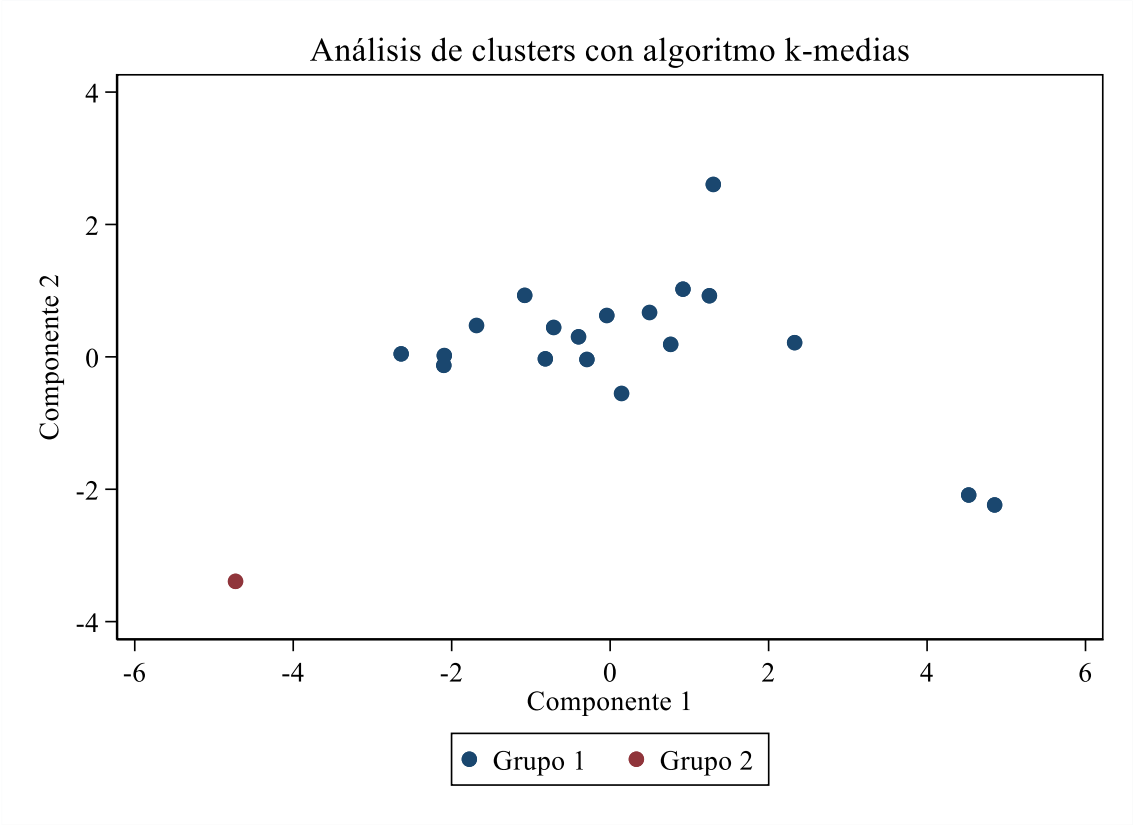
Tomando en cuenta las primeras 2 (dos) componentes principales que representaban el mayor porcentaje de varianza explicada y que se calcularon en los ejercicios anteriores, resolver los siguientes ejercicios.

Ejercicio 9.

Realizar un análisis de clusters para hallar la jerarquía de agrupación de los países considerados en la matriz de datos, en base a la información de las dos primeras componentes principales. Utilizar dos métodos distintos (se puede elegir entre los encadenamientos simple, completo, promedio, etc.).

Siguiendo la regla empírica de añadir un grupo más si el estadístico F de reducción de variabilidad es mayor que 10, se elegiría un solo grupo. Para darle más sentido al presente análisis de *clusters*, se generarán dos grupos.

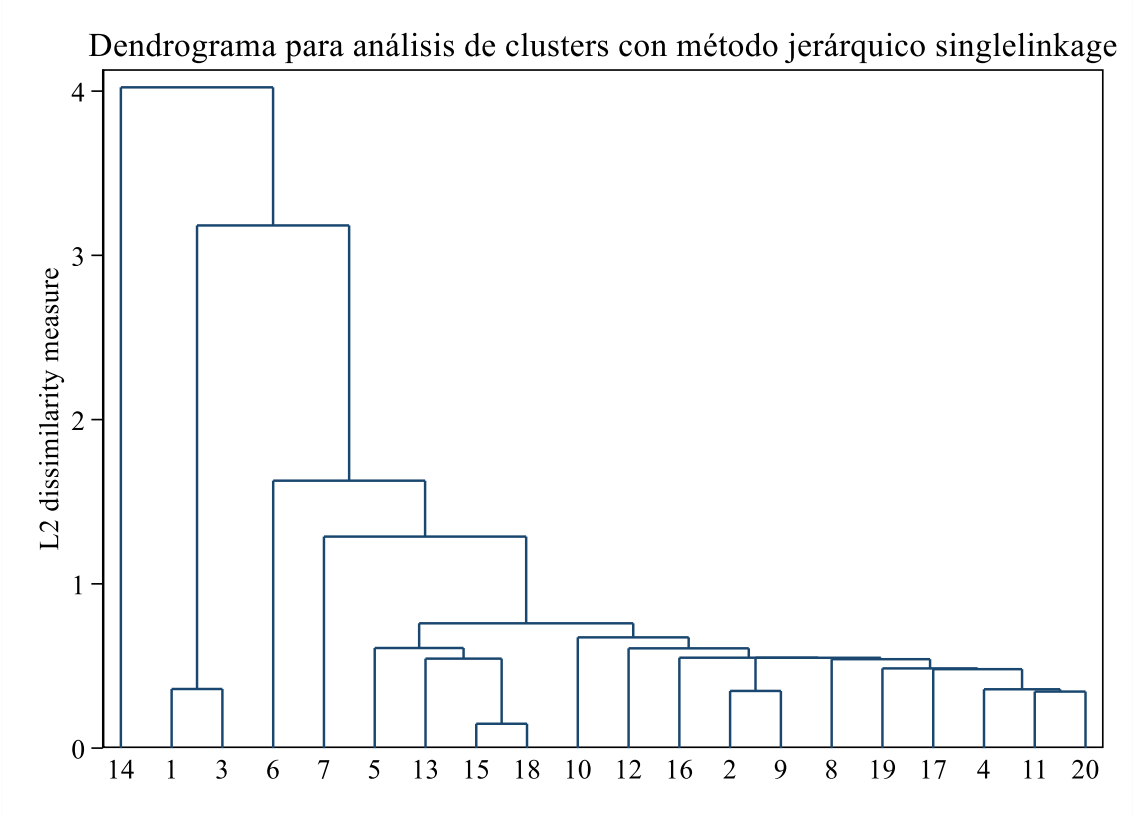
En primer lugar, se realiza un análisis de *clusters* empleando el algoritmo de las k-medias.

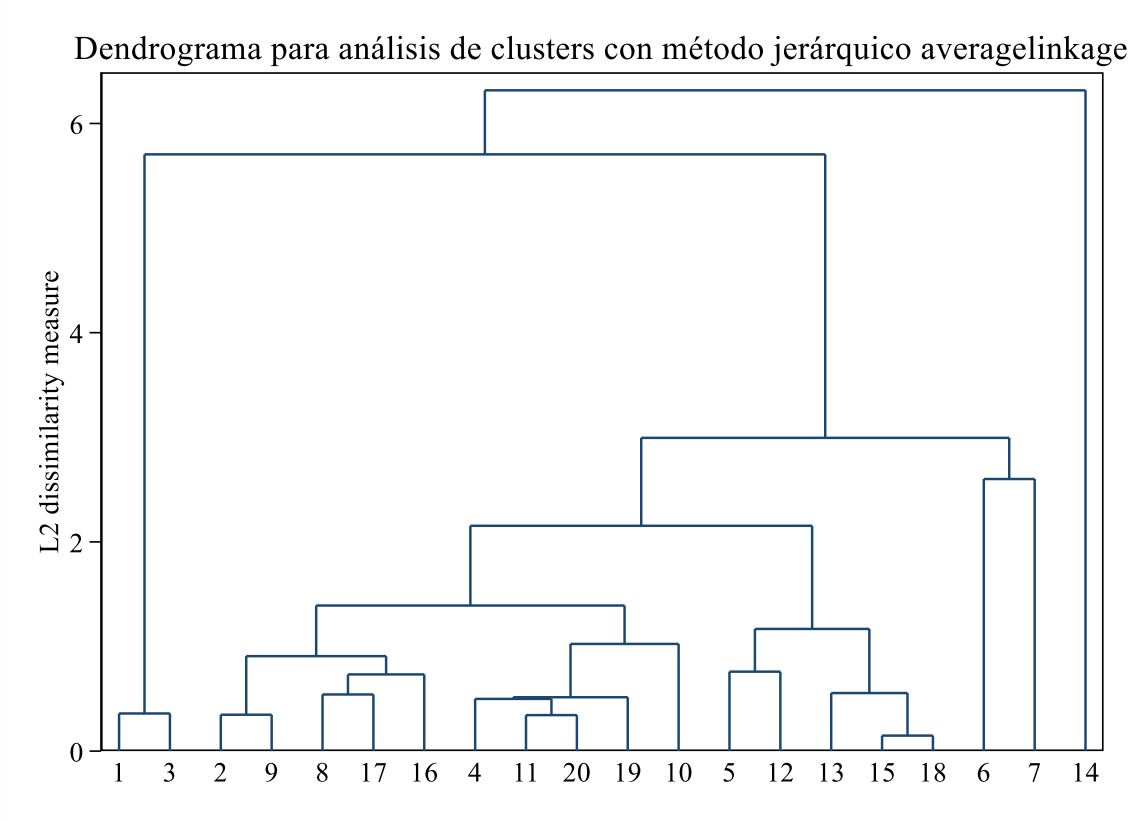


En segundo lugar, se realiza un análisis de *clusters* empleando métodos jerárquicos, trabajando con dos medidas de distancias entre *clusters* (*singlelinkage* y *averagelinkage*).

Ejercicio 10.

En cada caso, realizar una representación gráfica de los resultados obtenidos utilizando un dendrograma.




Dendrograma para análisis de clusters con método jerárquico averagelinkage
L2 dissimilarity measure
0
2
4
6
1 3 2 9 8 17 16 4 11 20 19 10 5 12 13 15 18 6 7 14

Ejercicio 11.

Interpretar los resultados obtenidos tomando en cuenta el gráfico biplot que se realizó para resolver las últimas consignas de la primera parte del examen.

Con ambos métodos jerárquicos (como así también con el algoritmo de las k-medias), si se consideran dos grupos, se tendría un grupo conformado por Haití y otro grupo conformado por el resto de los países en la matriz de datos. Tomando en cuenta el gráfico *biplot*, lo que se pudo observar es que Haití tiene un valor negativo muy alto en las dos primeras componentes principales, respecto al resto de los países, pudiendo considerarse una observación atípica en la matriz de datos, con lo cual es razonable que esta observación no sea agrupada con otra/s y que, por sí sola, conforme un grupo.

En el dendrograma, la altura de las líneas verticales y el ancho de las líneas horizontales dan pistas visuales sobre la fuerza de la agrupación. En particular, las líneas verticales más largas indican grupos con observaciones más disímiles, mientras que las más cortas indican grupos con observaciones menos disímiles; las líneas horizontales más anchas indican grupos más distintos, mientras que las más angostas indican grupos más parecidos. Por lo tanto, el análisis vertical da cuenta de que tan distintos son los países en los grupos (lo que se puede interpretar como variabilidad intra grupo), mientras que el análisis horizontal da cuenta de que tan distintos son los grupos (lo que se puede interpretar como variabilidad inter grupos).

Teniendo en cuenta lo anterior, entonces, se puede mencionar que, con el método jerárquico *singlelinkage* (versus el *averagelinkage*), en general, se observan líneas verticales más cortas y líneas horizontales más angostas, indicando una menor variabilidad intra e inter grupo, respectivamente.

Ejercicio 12.

Resumir, brevemente, las principales conclusiones del análisis.

Las principales conclusiones del análisis son:

- con ambos métodos jerárquicos, Haití, por sí solo, conforma un grupo.
- con el método jerárquico *singlelinkage*, respecto al *averagelinkage*, hay una menor variabilidad intra e inter grupo.

Ejercicio 13.

Efectuar un análisis factorial para describir la variabilidad común entre las variables de la matriz de datos. El punto de partida será un modelo con un solo factor.

En primer lugar, se procede a realizar tests de normalidad multivariada:

```
Test for multivariate normality

    Mardia mSkewness =  87.95871   chi2(220) =  346.166   Prob>chi2 =  0.0000
    Mardia mKurtosis =  127.3285     chi2(1) =    1.119   Prob>chi2 =  0.2902
    Henze-Zirkler    =  1.053822     chi2(1) =   27.609   Prob>chi2 =  0.0000
    Doornik-Hansen                  chi2(20) =  107.202   Prob>chi2 =  0.0000
```

Por lo tanto, con un nivel de significancia del 1% (excepto con el test Mardia mKurtosis), estos datos aportan evidencia suficiente para indicar que no tienen una distribución normal multivariada.

A continuación, se realiza un análisis factorial (con un factor) para describir la variabilidad común entre las variables de la matriz de datos:

```
Factor analysis/correlation                       Number of obs    =         20
    Method: principal factors                     Retained factors =          1
    Rotation: (unrotated)                         Number of params =         10

    --------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
        Factor1 |      4.97431      3.61016            0.6742       0.6742
        Factor2 |      1.36415      0.79702            0.1849       0.8591
        Factor3 |      0.56713      0.16450            0.0769       0.9360
        Factor4 |      0.40263      0.12565            0.0546       0.9906
        Factor5 |      0.27698      0.15339            0.0375       1.0281
        Factor6 |      0.12358      0.14362            0.0168       1.0449
        Factor7 |     -0.02004      0.04890           -0.0027       1.0422
        Factor8 |     -0.06894      0.03563           -0.0093       1.0328
        Factor9 |     -0.10456      0.03303           -0.0142       1.0187
       Factor10 |     -0.13759            .           -0.0187       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(45) =  142.45 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Uniqueness
BN_CAB_XOK~S	-0.7879	0.3793
EG_ELC_ACC~S	0.5731	0.6715
EN_ATM_CO2~C	0.8817	0.2227
IT_NET_USE~S	0.9134	0.1658
NE_CON_GOV~S	0.4698	0.7793
NE_GDI_TOT~S	0.5155	0.7342
NY_GDP_MKT~D	0.1335	0.9822
NY_GDP_PCA~D	0.8956	0.1980
SP_DYN_LE0~N	0.6793	0.5385
SP_POP_DPND	-0.8036	0.3542

Comunalidad de las variables en el modelo factorial con un factor:

```
commonality1[10,1]
              commonality1
BN_CAB_XOK~S     .62074391
EG_ELC_ACC~S     .32847199
EN_ATM_CO2~C     .77733507
IT_NET_USE~S     .83422812
NE_CON_GOV~S     .22068814
NE_GDI_TOT~S     .26575746
NY_GDP_MKT~D     .01782569
NY_GDP_PCA~D     .80204047
SP_DYN_LE0~N     .46146383
 SP_POP_DPND     .64575751
```

Análisis de los residuos en el modelo factorial con un factor:

```
Test for multivariate normality

    Mardia mSkewness =  87.95871   chi2(220) =  346.166   Prob>chi2 =  0.0000
    Mardia mKurtosis =  127.3285     chi2(1) =    1.119   Prob>chi2 =  0.2902
    Henze-Zirkler    =  1.053822     chi2(1) =   27.609   Prob>chi2 =  0.0000
    Doornik-Hansen                 chi2(20) =  108.741   Prob>chi2 =  0.0000
```

Por lo tanto, con un nivel de significancia del 1% (excepto con el test Mardia mKurtosis), estos datos aportan evidencia suficiente para indicar que los residuos no tienen una distribución normal multivariada.

```
Test that covariance matrix is diagonal

  Adjusted LR chi2(45) =    106.98
           Prob > chi2 =    0.0000
```

Por lo tanto, con un nivel de significancia del 1%, estos datos aportan evidencia suficiente para indicar que la matriz de varianzas y covarianzas de los residuos no es diagonal, por lo que se debería aumentar el número de factores hasta que los residuos estimados verifiquen la hipótesis nula.

Estadísticos de bondad del ajuste en el modelo factorial con un factor:

```
rho2_1 = .85616482
rho2_2 = .54905014
rho2_3 = .95042033
rho2_4 = .97251968
rho2_5 = .39267303
rho2_6 = .4608879
rho2_7 = .03533362
rho2_8 = .96081202
rho2_9 = .7099788
rho2_10 = .87451226
```

Coeficiente de determinación= -0,04885902.

Ejercicio 14.

Estimar la matriz de varianzas y covarianzas que surge del modelo factorial, la cual se descompone en varianza común (debido a los factores) y varianza específica. Comparar estos resultados con respecto a la matriz S de varianzas y covarianzas muestral.

Matriz de correlaciones muestrales:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
BN_CAB_XOK~S	1.0000									
EG_ELC_ACC~S	-0.1950	1.0000								
EN_ATM_CO2~C	-0.7962	0.3951	1.0000							
IT_NET_USE~S	-0.5956	0.6103	0.7934	1.0000						
NE_CON_GOV~S	-0.2527	0.6374	0.2380	0.4152	1.0000					
NE_GDI_TOT~S	-0.5910	0.0026	0.5231	0.4208	0.0761	1.0000				
NY_GDP_MKT~D	0.1107	0.1704	0.1029	0.2384	0.1764	-0.2198	1.0000			
NY_GDP_PCA~D	-0.7965	0.4003	0.9047	0.8191	0.2073	0.3676	0.1280	1.0000		
SP_DYN_LE0~N	-0.2924	0.5713	0.4692	0.6818	0.3750	0.2958	0.0812	0.5761	1.0000	
SP_POP_DPND	0.6646	-0.4136	-0.6228	-0.7337	-0.5030	-0.4354	-0.1998	-0.6318	-0.5459	1.0000

Matriz de varianzas y covarianzas en el modelo factorial con un factor:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
BN_CAB_XOK~S	1.0000									
EG_ELC_ACC~S	-0.4515	1.0000								
EN_ATM_CO2~C	-0.6946	0.5053	1.0000							
IT_NET_USE~S	-0.7196	0.5235	0.8053	1.0000						
NE_CON_GOV~S	-0.3701	0.2692	0.4142	0.4291	1.0000					
NE_GDI_TOT~S	-0.4062	0.2955	0.4545	0.4709	0.2422	1.0000				
NY_GDP_MKT~D	-0.1052	0.0765	0.1177	0.1219	0.0627	0.0688	1.0000			
NY_GDP_PCA~D	-0.7056	0.5133	0.7896	0.8180	0.4207	0.4617	0.1196	1.0000		
SP_DYN_LE0~N	-0.5352	0.3893	0.5989	0.6205	0.3191	0.3502	0.0907	0.6084	1.0000	
SP_POP_DPND	0.6331	-0.4606	-0.7085	-0.7340	-0.3775	-0.4143	-0.1073	-0.7197	-0.5459	1.0000

Por lo tanto, se puede observar que, con un factor, las correlaciones estimadas y muestrales no son tan parecidas.

Ejercicio 15.

Repetir los dos ejercicios anteriores agregando un nuevo factor al modelo. ¿Se producen cambios en las comunalidades de las variables con respecto al modelo estimado en el Ejercicio 13? Determinar cuál de las dos especificaciones del modelo factorial (m = 1 o m = 2) resulta más adecuada para representar la estructura de asociación entre las variables. Es posible que, en función de las variables y/o países elegidos en la matriz de datos, la especificación del modelo factorial con dos factores no sea la óptima. En ese caso, agregar los comentarios que creas conveniente.

A continuación, se realiza un análisis factorial (con dos factores) para describir la variabilidad común entre las variables de la matriz de datos:

```
Factor analysis/correlation                       Number of obs    =         20
    Method: principal factors                     Retained factors =          2
    Rotation: (unrotated)                         Number of params =         19

    --------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
       Factor1  |      4.97431      3.61016            0.6742       0.6742
       Factor2  |      1.36415      0.79702            0.1849       0.8591
       Factor3  |      0.56713      0.16450            0.0769       0.9360
       Factor4  |      0.40263      0.12565            0.0546       0.9906
       Factor5  |      0.27698      0.15339            0.0375       1.0281
       Factor6  |      0.12358      0.14362            0.0168       1.0449
       Factor7  |     -0.02004      0.04890           -0.0027       1.0422
       Factor8  |     -0.06894      0.03563           -0.0093       1.0328
       Factor9  |     -0.10456      0.03303           -0.0142       1.0187
      Factor10  |     -0.13759            .           -0.0187       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(45) =  142.45 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
BN_CAB_XOK~S	-0.7879	0.4611	0.1666
EG_ELC_ACC~S	0.5731	0.5794	0.3359
EN_ATM_CO2~C	0.8817	-0.2397	0.1652
IT_NET_USE~S	0.9134	0.1415	0.1458
NE_CON_GOV~S	0.4698	0.4929	0.5364
NE_GDI_TOT~S	0.5155	-0.5022	0.4820
NY_GDP_MKT~D	0.1335	0.3609	0.8519
NY_GDP_PCA~D	0.8956	-0.1677	0.1698
SP_DYN_LE0~N	0.6793	0.2847	0.4575
SP_POP_DPND	-0.8036	-0.0614	0.3505

Comunalidad de las variables en el modelo factorial con dos factores:

```
commonality2[10,1]
              commonality2
BN_CAB_XOK~S     .83338206
EG_ELC_ACC~S     .66412751
EN_ATM_CO2~C     .83479472
IT_NET_USE~S     .85424241
NE_CON_GOV~S     .46364859
NE_GDI_TOT~S     .51795456
NY_GDP_MKT~D     .14809954
NY_GDP_PCA~D     .83017706
SP_DYN_LE0~N     .54250554
 SP_POP_DPND     .64953224
```

Por lo tanto, se producen cambios en las comunalidades de las variables con respecto al modelo estimado en el Ejercicio 13 (con un factor), aumentando en todos los casos, ya que, al considerar más factores, los valores de comunalidad son mayores (hay más variación de cada variable explicada por factores) y, en general, se explican mejor los datos, lo que se puede ver en el hecho de que la matriz de varianzas y covarianzas estimada con dos factores (versus la estimada con un factor) es más parecida a la matriz de varianzas y covarianzas muestral.

Análisis de los residuos en el modelo factorial con dos factores:

```
Test for multivariate normality

    Mardia mSkewness =  87.95871   chi2(220) =  346.166   Prob>chi2 =  0.0000
    Mardia mKurtosis =  127.3285     chi2(1) =    1.119   Prob>chi2 =  0.2902
    Henze-Zirkler    =  1.053822     chi2(1) =   27.609   Prob>chi2 =  0.0000
    Doornik-Hansen                  chi2(20) =  108.741   Prob>chi2 =  0.0000
```

Por lo tanto, con un nivel de significancia del 1% (excepto con el test Mardia mKurtosis), estos datos aportan evidencia suficiente para indicar que los residuos no tienen una distribución normal multivariada.

```
Test that covariance matrix is diagonal

  Adjusted LR chi2(45) =    103.04
           Prob > chi2 =    0.0000
```

Por lo tanto, con un nivel de significancia del 1%, estos datos aportan evidencia suficiente para indicar que la matriz de varianzas y covarianzas de los residuos no es diagonal, por lo que se debería aumentar el número de factores hasta que los residuos estimados verifiquen la hipótesis nula.

Estadísticos de bondad del ajuste en el modelo factorial con dos factores:

```
rho2_1 = .97223846
rho2_2 = .88718967
rho2_3 = .97270721
rho2_4 = .97875472
rho2_5 = .71232716
rho2_6 = .7676322
rho2_7 = .27426561
rho2_8 = .97116017
rho2_9 = .79069882
rho2_10 = .87717235
```

Coeficiente de determinación= 0,2388525.

Matriz de correlaciones muestrales:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
BN_CAB_XOK~S	1.0000									
EG_ELC_ACC~S	-0.1950	1.0000								
EN_ATM_CO2~C	-0.7962	0.3951	1.0000							
IT_NET_USE~S	-0.5956	0.6103	0.7934	1.0000						
NE_CON_GOV~S	-0.2527	0.6374	0.2380	0.4152	1.0000					
NE_GDI_TOT~S	-0.5910	0.0026	0.5231	0.4208	0.0761	1.0000				
NY_GDP_MKT~D	0.1107	0.1704	0.1029	0.2384	0.1764	-0.2198	1.0000			
NY_GDP_PCA~D	-0.7965	0.4003	0.9047	0.8191	0.2073	0.3676	0.1280	1.0000		
SP_DYN_LE0~N	-0.2924	0.5713	0.4692	0.6818	0.3750	0.2958	0.0812	0.5761	1.0000	
SP_POP_DPND	0.6646	-0.4136	-0.6228	-0.7337	-0.5030	-0.4354	-0.1998	-0.6318	-0.5459	1.0000

Matriz de varianzas y covarianzas en el modelo factorial con dos factores:

	BN_CAB~S	EG_ELC~S	EN_ATM~C	IT_NET~S	NE_CON~S	NE_GDI~S	NY~TP_KD	NY~AP_KD	SP_DYN~N	SP_POP~D
BN_CAB_XOK~S	1.0000									
EG_ELC_ACC~S	-0.1844	1.0000								
EN_ATM_CO2~C	-0.8052	0.3664	1.0000							
IT_NET_USE~S	-0.6544	0.6054	0.7714	1.0000						
NE_CON_GOV~S	-0.1428	0.5548	0.2960	0.4988	1.0000					
NE_GDI_TOT~S	-0.6377	0.0045	0.5749	0.3998	-0.0054	1.0000				
NY_GDP_MKT~D	0.0612	0.2856	0.0312	0.1730	0.2406	-0.1124	1.0000			
NY_GDP_PCA~D	-0.7829	0.4161	0.8298	0.7942	0.3380	0.5459	0.0590	1.0000		
SP_DYN_LE0~N	-0.4039	0.5543	0.5307	0.6607	0.4594	0.2072	0.1934	0.5606	1.0000	
SP_POP_DPND	0.6048	-0.4962	-0.6938	-0.7427	-0.4078	-0.3834	-0.1295	-0.7094	-0.5634	1.0000

Por lo tanto, se puede observar que, con dos factores, las correlaciones estimadas y muestrales son más parecidas (versus las estimadas con un factor).

Por último, la especificación del modelo factorial que resulta más adecuada para representar la estructura de asociación entre las variables es, considerando el BIC (Bayesian Information Criterion), m = 1 y, considerando el AIC (Akaike Information Criterion), m = 2.

Ejercicio 16.

Interpretar, brevemente, los resultados obtenidos, tomando en cuenta las eventuales diferencias y semejanzas con los resultados del cálculo de componentes principales.

En primer lugar, se recuerdan los resultados obtenidos tanto en el análisis de componentes principales como en el análisis factorial.

Análisis de componentes principales:

```
Principal components (eigenvectors)

    ------------------------------------------------
        Variable |    Comp1     Comp2  | Unexplained
    -------------+--------------------+-------------
    BN_CAB_XOK~S |  -0.3457    0.3523  |       .1738
    EG_ELC_ACC~S |   0.2670    0.4609  |       .2729
    EN_ATM_CO2~C |   0.3883   -0.1899  |       .1617
    IT_NET_USE~S |   0.4058    0.0823  |       .1397
    NE_CON_GOV~S |   0.2240    0.4265  |       .4334
    NE_GDI_TOT~S |   0.2350   -0.4485  |       .3747
    NY_GDP_MKT~D |   0.0675    0.4278  |       .6666
    NY_GDP_PCA~D |   0.3896   -0.1321  |       .1878
    SP_DYN_LE0~N |   0.3120    0.1898  |       .4373
     SP_POP_DPND |  -0.3670   -0.0337  |       .3039
    ------------------------------------------------
```

Análisis factorial:

```
Factor loadings (pattern matrix) and unique variances

    -----------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
    BN_CAB_XOK~S |  -0.7879    0.4611 |      0.1666
    EG_ELC_ACC~S |   0.5731    0.5794 |      0.3359
    EN_ATM_CO2~C |   0.8817   -0.2397 |      0.1652
    IT_NET_USE~S |   0.9134    0.1415 |      0.1458
    NE_CON_GOV~S |   0.4698    0.4929 |      0.5364
    NE_GDI_TOT~S |   0.5155   -0.5022 |      0.4820
    NY_GDP_MKT~D |   0.1335    0.3609 |      0.8519
    NY_GDP_PCA~D |   0.8956   -0.1677 |      0.1698
    SP_DYN_LE0~N |   0.6793    0.2847 |      0.4575
     SP_POP_DPND |  -0.8036   -0.0614 |      0.3505
    -----------------------------------------------
```

Por lo tanto, se puede observar que:

- el signo de los componentes principales y de los factores asociados a cada una de las variables es el mismo; y
- el valor absoluto de los factores asociados a cada una de las variables es mayor al de los componentes principales.

Siendo así, al igual que en el caso del análisis de componentes principales, se tiene que:

- el primer factor tomará valores altos en aquellos países en donde las variables "Age dependency ratio (% of working-age population)" y "Current account balance (% of GDP)" resulten menos importantes, en términos relativos, a las restantes; y
- el segundo factor tomará valores altos en aquellos países en donde las variables que ponderan con un signo positivo en el autovector (mencionadas en el Ejercicio 5) resulten más importantes, en términos relativos, a las restantes.