Análisis de componentes principales de datos económicos de una serie de países

Presentación en vídeo:

http://personales.upv.es/asala/YT/V/paises1.html

http://personales.upv.es/asala/YT/V/paises2.html

http://personales.upv.es/asala/YT/V/paises3.html

*Datos tomados de página 72 de "*100 problemas resueltos de estadística multivariante (implementados en Matlab)*"; Autoras: Amparo Baillo, Áurea Grané, Delta Publicaciones (con permiso de la segunda de las autoras). Datos numéricos en: http://halweb.uc3m.es/esp/Personal/personas/agrane/libro/ficheros_datos/capitulo_4/paises_mundo.txt

Este código funcionó correctamente en Matlab **2020a**

Objetivo: realizar un análisis de componentes principales (PCA) de datos económicos. Observar las clasificaciones resultantes a partir de los componentes con más varianza.

Tabla de Contenidos

1.- Carga, visualización y preprocesado de datos....1
Transformación para reducir la asimetría con respecto a la media....4
Estandarización de las variables (media 0, vza 1)....6
2.- Análisis de componentes principales....7
Análisis de los componentes de mayor varianza (1 y 2)....11
Análisis de los últimos componentes principales (menor varianza)....13
Conclusiones....15
Apéndice: funciones auxiliares....15

1.- Carga, visualización y preprocesado de datos

```
fid=fopen("paises_mundo_n.txt");
data=textscan(fid,"%s%f%f%f%f%f%f%f%f%f%f");
fclose(fid);
```

El fichero contiene datos de 96 paises de los indicadores:

X1 = Tasa anual de crecimiento de la población,

X2 = Tasa de mortalidad infantil por cada 1000 nacidos vivos,

X3 = Porcentaje de mujeres en la población activa,

X4 = PNB en 1995 (en millones de dólares),

X5 = Producción de electricidad (en millones kW/h),

X6 = Lineas telefónicas por cada 1000 habitantes,

X7 = Consumo de agua per cápita,

X8 = Proporción de la superficie del país cubierta por bosques,

X9 = Proporción de deforestación anual,

X10 = Consumo de energía per cápita,

X11 = Emisión de CO2 per capita.

```
data
```

data = 1×12 cell

	1	2	3	4	5	6	7	8
1	96×1 cell	96×1 double	96×1 double	96×1 double	96×1 double	96×1 double	96×1 double	96×1 double

```
nombres=data{1}'
```

nombres = 1×96 cell

```
'Albania'   'Angola'    'ArabiaSaudi''Argelia'   'Argentina' 'Australia' 'Austri ···
```

```
datos_paises=[data{2:end}]
```

datos_paises = 96×11

10^6 ×

```
    0.0000    0.0000    0.0000    0.0022    0.0039    0.0000    0.0001    0.0001 ···
    0.0000    0.0001    0.0000    0.0044    0.0010    0.0000    0.0001    0.0000
    0.0000    0.0000    0.0000    0.1335    0.0910    0.0001    0.0005    0.0000
    0.0000    0.0000    0.0000    0.0446    0.0199    0.0000    0.0002    0.0000
    0.0000    0.0000    0.0000    0.2784    0.0660    0.0002    0.0010    0.0000
    0.0000    0.0000    0.0000    0.3379    0.1672    0.0005    0.0009    0.0000
    0.0000    0.0000    0.0000    0.2165    0.0533    0.0005    0.0003    0.0000
    0.0000    0.0001    0.0000    0.0286    0.0099    0.0000    0.0002    0.0000
    0.0000    0.0000    0.0000    0.2507    0.0722    0.0005    0.0009    0.0000
    0.0000    0.0001    0.0000    0.0020    0.0000    0.0000    0.0000    0.0000
     ⋮
```

```
[npaises,nindicadores]=size(datos_paises)
```

```
npaises = 96
nindicadores = 11
```

Analicemos los estadísticos usuales:

```
media=mean(datos_paises)
```

```
media = 1×11
10^5  ×
    0.0000    0.0004    0.0004    1.1659    0.6926    0.0017    0.0051    0.0003 ···
```

```
varianza=var(datos_paises)
```

```
varianza = 1×11
10^10  ×
    0.0000    0.0000    0.0000    4.9998    1.8219    0.0000    0.0000    0.0000 ···
```

```
desvtippaises=std(datos_paises)
```

```
desvtippaises = 1×11
10^5  ×
    0.0000    0.0003    0.0001    2.2360    1.3498    0.0020    0.0057    0.0002 ···
```

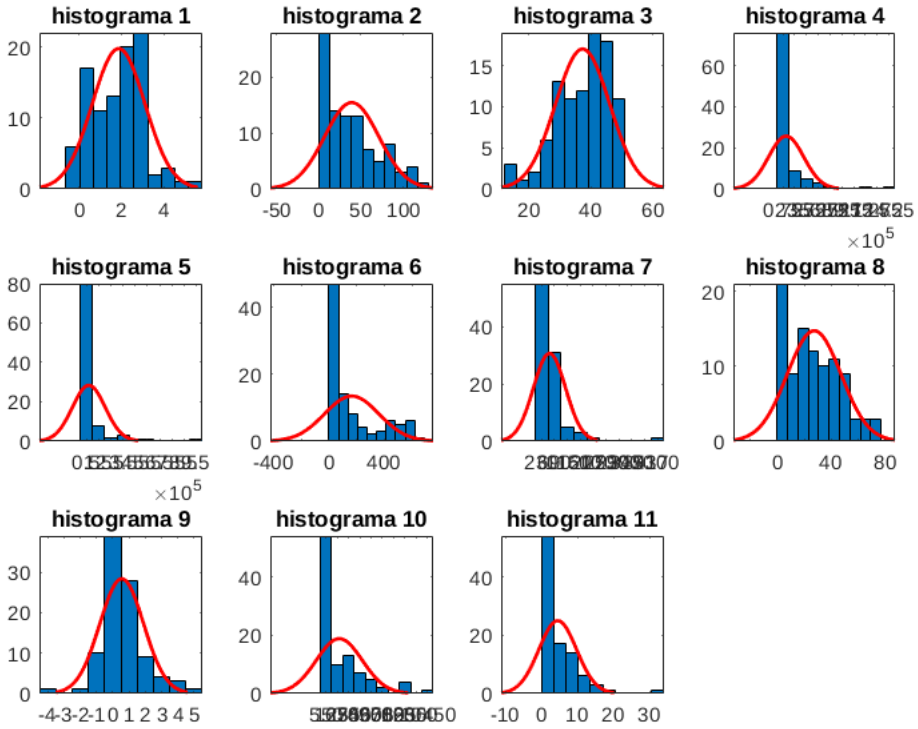
Hay que recalcar que la asimetría en algunos registros es muy fuerte:

```
skpaises=skewness(datos_paises)
```

```
skpaises = 1×11
    0.2531    0.8208   -0.7293    3.6457    3.8554    1.1550    4.1217    0.4813 ···
```

Visualizamos histogramas (superponiendo la distribución normal de misma media y varianza):

```
for i=1:11
    subplot(3,4,i)
    histfit(datos_paises(:,i))
    title(sprintf('histograma %d',i))
    axis tight
end
```

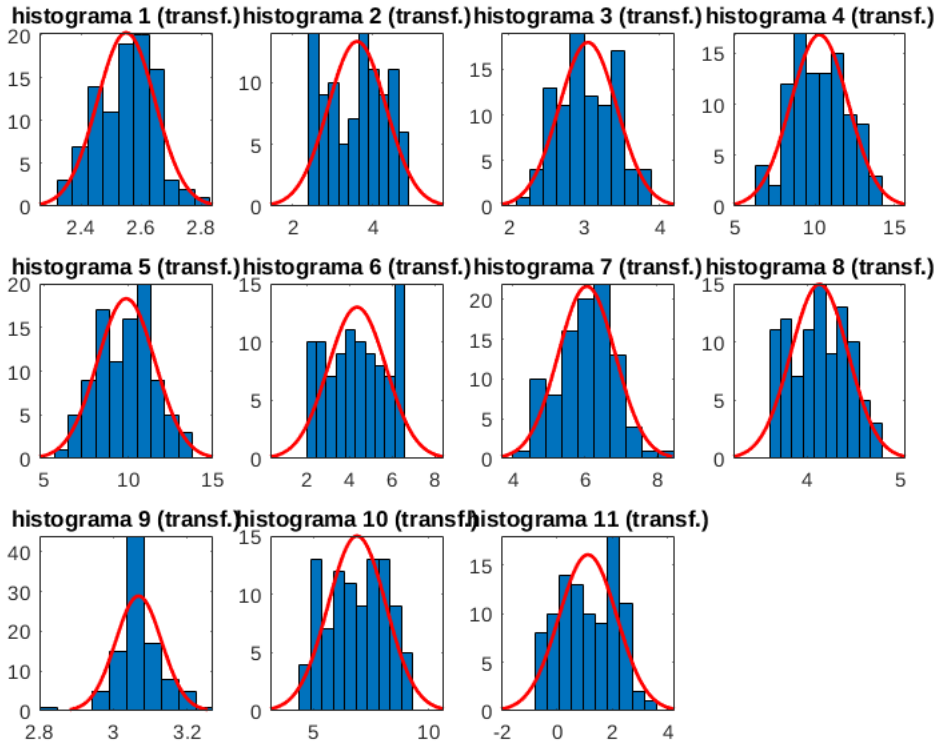



```
figure()
```

Transformación para reducir la asimetría con respecto a la media

Vamos a intentar que se asemejen más a una distribución normal o, al menos, centrada, en los más asimétricos

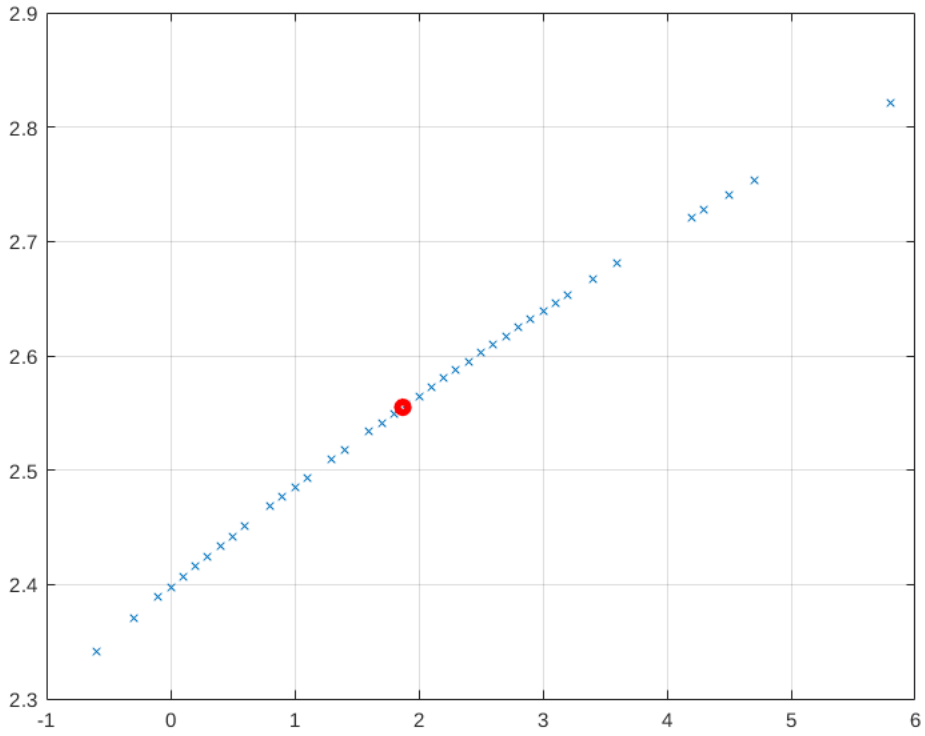
```
datos2=datos_paises;
valoresmin=min(datos_paises);valoresmax=max(datos_paises);
offset(1)=11;offset(2)=8;offset(3)=60;offset(4)=-800;offset(5)=450;
offset(6)=7;offset(7)=72;offset(8)=38;offset(9)=21;offset(10)=100;offset(11)=.5;
for i=1:11
   subplot(3,4,i)
   datos2(:,i)=log(offset(i)+sign(skpaises(i))*datos_paises(:,i));
   histfit(datos2(:,i))
   title(sprintf('histograma %d (transf.)',i))
   axis tight
end
```




```
figure()
```

Por ejemplo, la transformación en el dato 7, que era el más descentrado es:

```
tst=1;
plot(datos_paises(:,tst),datos2(:,tst),'x'), grid on
hold on
plot(media(tst),log(offset(tst)+sign(skpaises(tst))*media(tst)),'or','LineWidth',4)
hold off
```



Ahora están más centrados los datos:

```
skpaises%antes
```

```
skpaises = 1×11
    0.2531    0.8208   -0.7293    3.6457    3.8554    1.1550    4.1217    0.4813 ···
```

```
sk2=skewness(datos2) %después
```

```
sk2 = 1×11
   -0.0018    0.0002    0.0101    0.0008   -0.0017    0.0036    0.0008    0.0093 ···
```

Estandarización de las variables (media 0, vza 1)

Hagamos transformación lineal de los datos d a media cero y desviación típica unidad $z = (d - \bar{d})/\sigma_d$, resultando:

```
datosNorm=zscore(datos2);
mean(datosNorm)
```

```
ans = 1×11
10^-13 ×
   -0.0095   -0.0012   -0.0119    0.0085   -0.0103   -0.0047    0.0071    0.0213 ···
```

```
std(datosNorm)
```

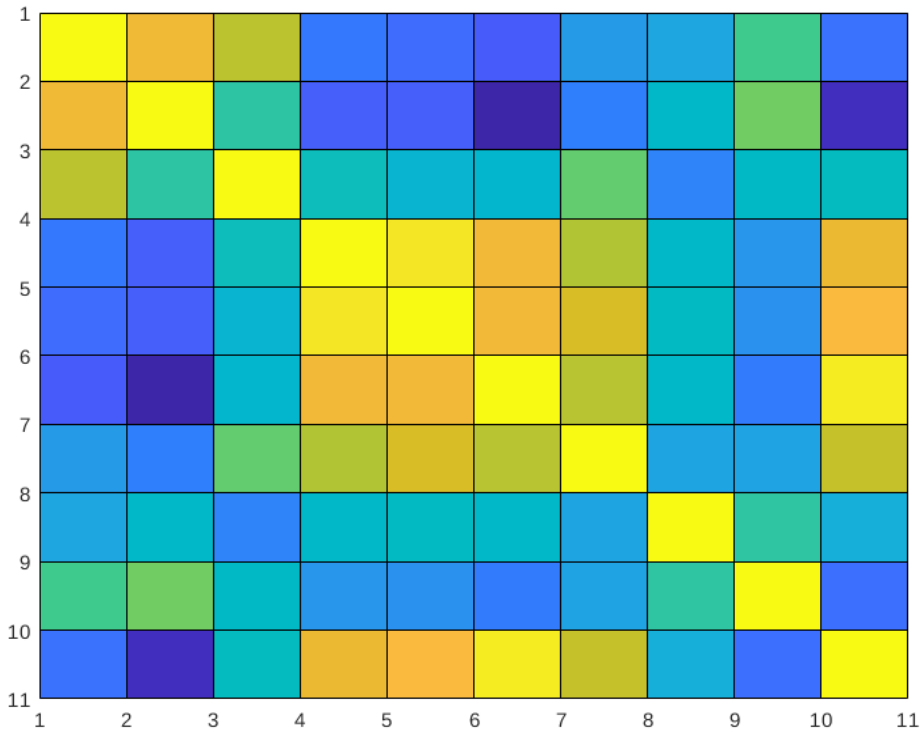
```
ans = 1×11
    1.0000    1.0000    1.0000    1.0000    1.0000    1.0000    1.0000    1.0000 ···
```

La matriz de correlaciones dos a dos es:

```
cov_paises2=cov(datosNorm) %correlaciones dos a dos
```

```
cov_paises2 = 11×11
    1.0000    0.6390    0.4929   -0.4354   -0.4904   -0.5803   -0.2384   -0.1690 ···
    0.6390    1.0000    0.1343   -0.5704   -0.5682   -0.9063   -0.4016   -0.0271
    0.4929    0.1343    1.0000    0.0321   -0.0598   -0.0484    0.3049   -0.3706
   -0.4354   -0.5704    0.0321    1.0000    0.8946    0.6494    0.4681   -0.0134
   -0.4904   -0.5682   -0.0598    0.8946    1.0000    0.6456    0.5638    0.0062
   -0.5803   -0.9063   -0.0484    0.6494    0.6456    1.0000    0.4850   -0.0157
   -0.2384   -0.4016    0.3049    0.4681    0.5638    0.4850    1.0000   -0.1814
   -0.1690   -0.0271   -0.3706   -0.0134    0.0062   -0.0157   -0.1814    1.0000
    0.2162    0.3337   -0.0065   -0.2724   -0.3020   -0.4222   -0.1908    0.1363
   -0.4656   -0.8457    0.0134    0.6255    0.6706    0.9239    0.5096   -0.1069
      ⋮
```

```
pcolor(cov_paises2), axis ij
```



2.- Análisis de componentes principales

```
[U,S,V]=svd(datosNorm,'econ');
size(U)
```

```
ans = 1×2
    96    11
```

```
size(S)
```

```
ans = 1×2
    11    11
```

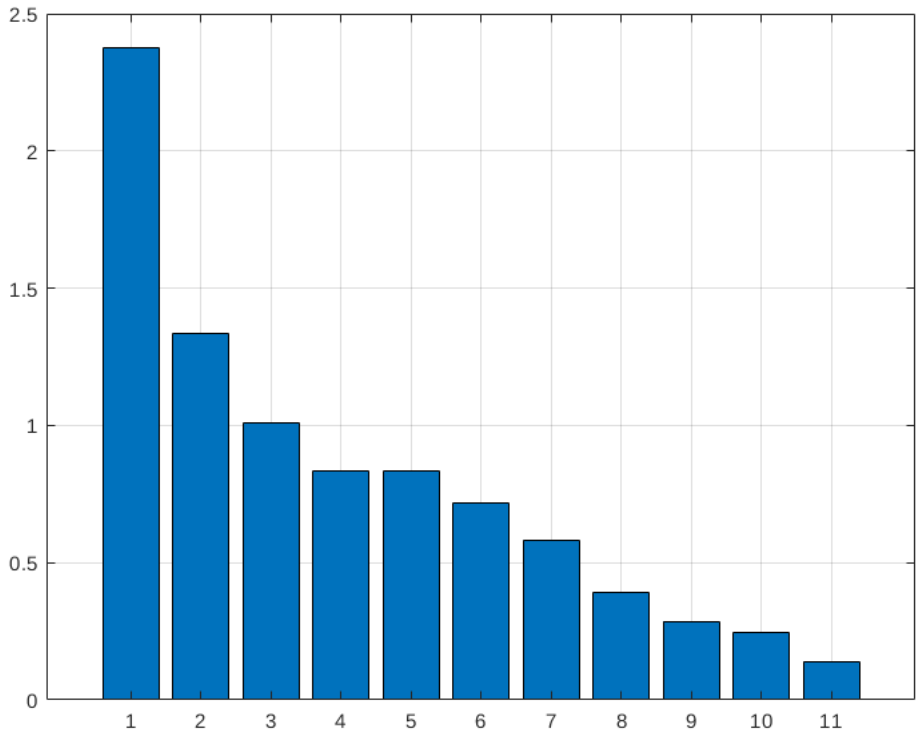
```
size(V)
```

```
ans = 1×2
    11    11
```

```
stdcp=diag(S)'/sqrt(95)
```

```
stdcp = 1×11
    2.3769    1.3332    1.0078    0.8339    0.8321    0.7191    0.5824    0.3901 ···
```

```
bar(stdcp), grid on
```



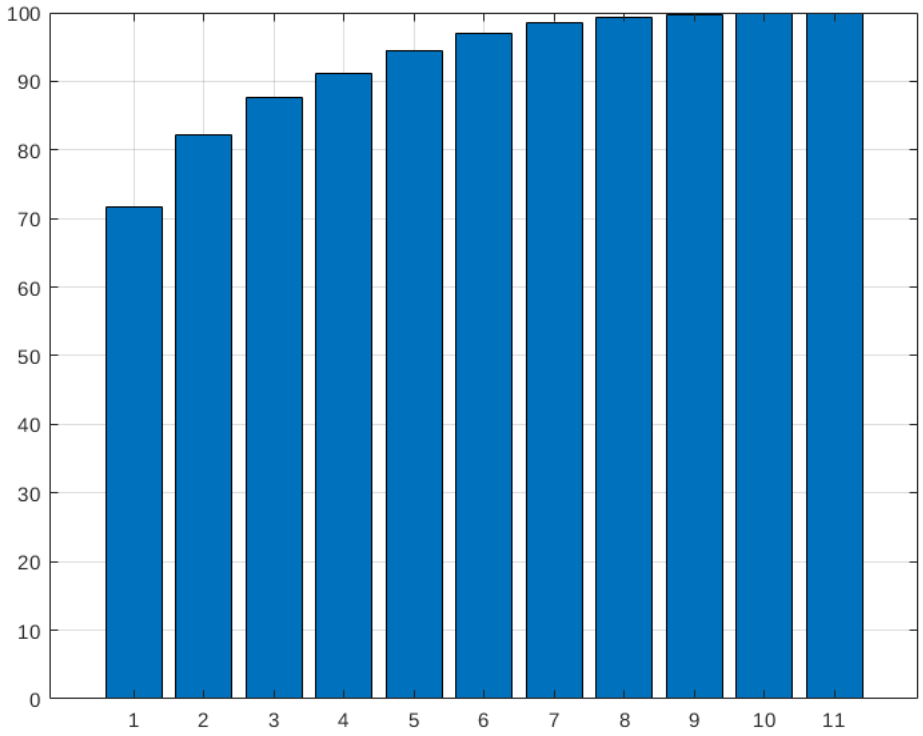
La suma de varianza de todos los componentes es igual a la varianza total:

```
tv=sum(stdcp.^2)
```

```
tv = 11.0000
```

Variabilidad (en unidades de desviación típica) explicada por cada componente:

```
ev=sqrt(cumsum(stdcp.^2)'/tv);
bar(ev*100), grid on
```

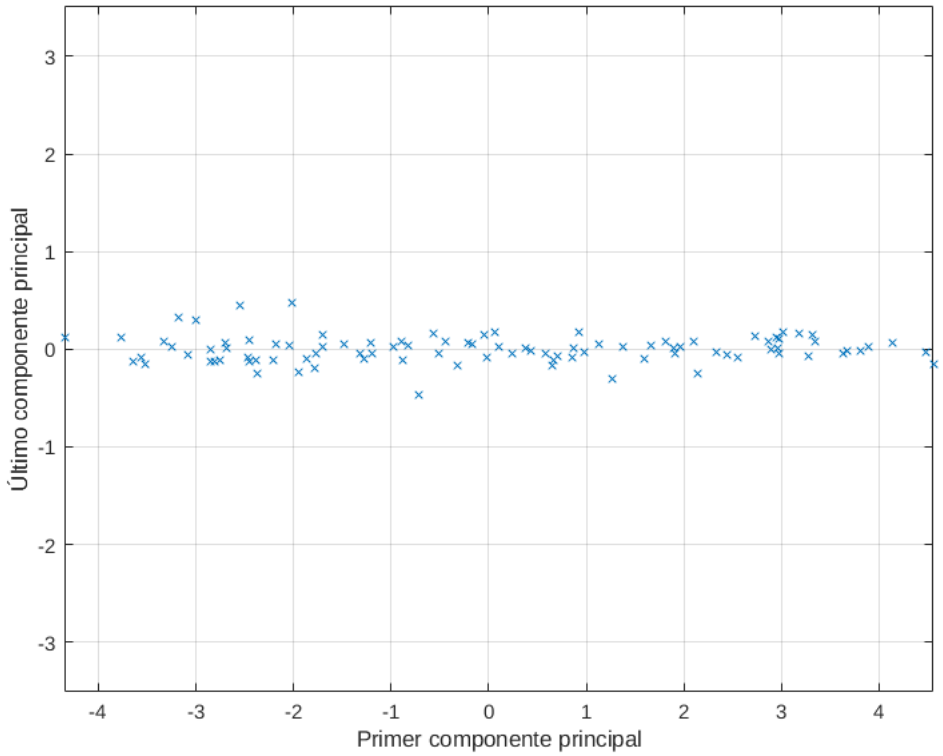


Vamos a traducir los datos a componentes principales con el cambio de variable lineal:

```
Comp_prin=datosNorm*V; %U*S
std(Comp_prin)
```

```
ans = 1×11
    2.3769    1.3332    1.0078    0.8339    0.8321    0.7191    0.5824    0.3901 ···
```

```
%comparamos el primero y el último
plot(Comp_prin(:,1),Comp_prin(:,11),'x')
axis equal, grid on
xlabel('Primer componente principal')
ylabel('Último componente principal')
```



Obviamente, el comando pca de Matlab también realiza los mismos cálculos:

```
[coef,score,latent,t2]=pca(datosNorm);
norm(abs(coef)-abs(V),"fro") %el SVD puede dar direcciones de signo contrario...
```

```
ans = 3.2775e-14
```

```
norm(abs(score)-abs(Comp_prin),"fro")
```

```
ans = 3.9524e-13
```

```
norm(latent'-stdcp.^2)
```

```
ans = 4.6899e-15
```

```
t2(7)
```

```
ans = 4.7159
```

```
datosNorm(7,:)*inv(cov_paises2)*datosNorm(7,:)'
```

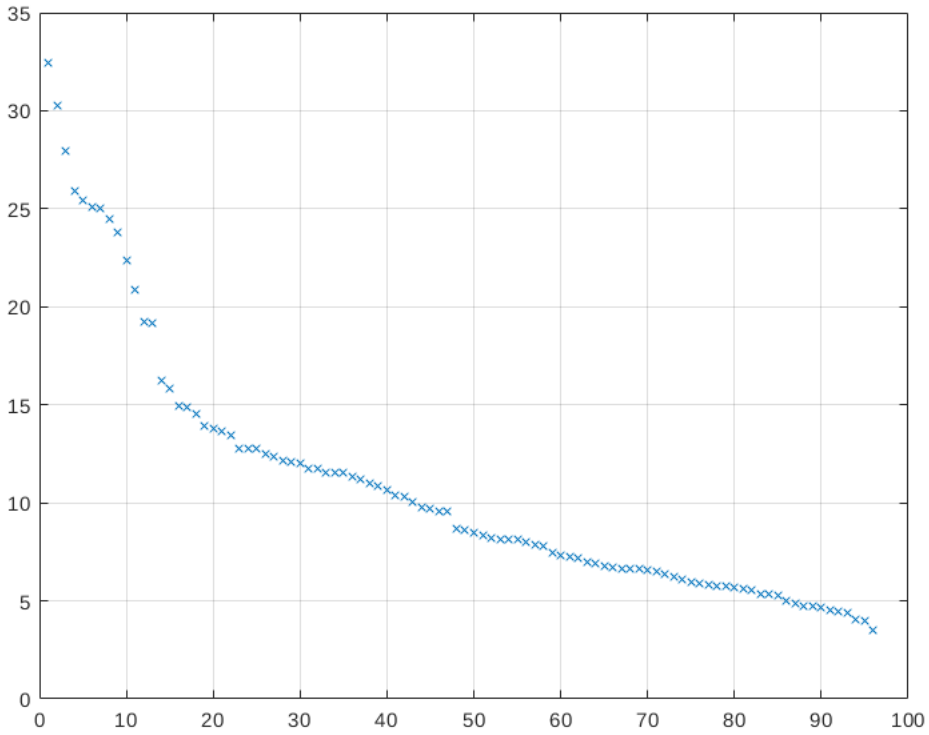
```
ans = 4.7159
```

```
sum((Comp_prin(7,:)./stdcp).^2)
```

```
ans = 4.7159
```

Los países con más resíduo T2 (más lejos en el elipsoide de varianzas-covarianzas en el que se distribuirían los datos si fueran distrib. normal multidimensional) son:

```
[val,II]=sort(t2,"descend");
plot(val,'x'), grid on
```



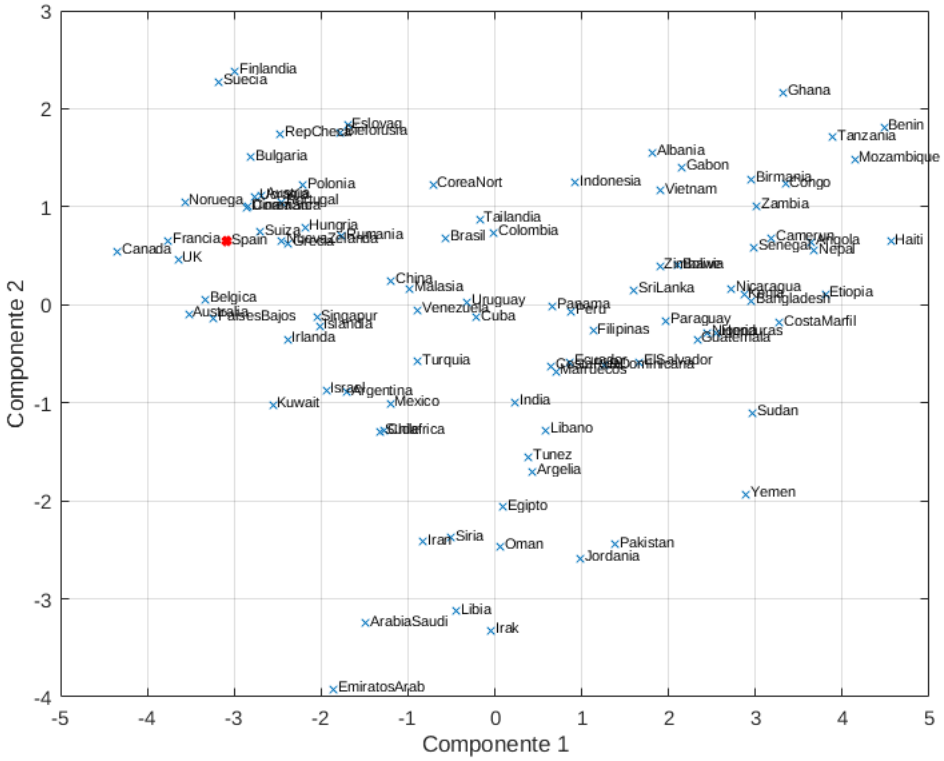

```
nombres{II(1:12)}
```

```
ans =
'Bulgaria'
ans =
'Kuwait'
ans =
'CoreaNort'
ans =
'Islandia'
ans =
'Siria'
ans =
'EmiratosArab'
ans =
'Gabon'
ans =
'Irak'
ans =
'Paraguay'
ans =
'Haiti'
ans =
'SriLanka'
ans =
'Singapur'
```

Análisis de los componentes de mayor varianza (1 y 2)

Los primeros componentes son los de mayor variabilidad, los que mejor distinguen a unos países de otros (entre las combinaciones lineales de los datos).

```
plotpca2D(1,2,Comp_prin,nombres)
```



Valores extremos del primer componente:

```
[~,I3]=sort(Comp_prin(:,1));
{nombres{I3(1:8)}} %los más bajos (izquierda)
```

```
ans = 1×8 cell
'Canada'    'Francia'    'UK'        'Noruega'    'Australia' 'Belgica'    'PaisesB ···
```

```
{nombres{I3(end:-1:89)}} %los más altos (derecha)
```

```
ans = 1×8 cell
'Haiti'     'Benin'     'Mozambique''Tanzania'  'Etiopia'    'Nepal'      'Angola' ···
```

Valores extremos del segundo componente:

```
[~,I4]=sort(Comp_prin(:,2));
{nombres{I4(1:8)}}%más bajo componente (abajo)
```

```
ans = 1×8 cell
'EmiratosArab''Irak'      'ArabiaSaudi''Libia'     'Jordania'  'Oman'       'Paki ···
```

```
{nombres{I4(end:-1:90)}}%más alto valor del componente (arriba)
```

```
ans = 1×7 cell
```

```
'Finlandia' 'Suecia'    'Ghana'     'Eslovaq'   'Benin'     'Bielorusia''RepChec ···
```

Intentemos evaluar el significado de los dos primeros componentes:

```
V(:,1:2)'
```

```
ans = 2×11
    0.2657    0.3695    0.0258   -0.3302   -0.3456   -0.3943   -0.2512    0.0209 ···
   -0.4130   -0.1163   -0.6554   -0.0386   -0.0073    0.0257   -0.3147    0.5154
```

```
%X1 = Tasa anual de crecimiento de la población,
%X2 = Tasa de mortalidad infantil por cada 1000 nacidos vivos,
%X3 = Porcentaje de mujeres en la población activa,
%X4 = PNB en 1995 (en millones de dólares),
%X5 = Producción de electricidad (en millones kW/h),
%X6 = Lineas telefónicas por cada 1000 habitantes,
%X7 = Consumo de agua per cápita,
%X8 = Proporción de la superficie del paı´s cubierta por bosques,
%X9 = Proporción de deforestación anual,
%X10 = Consumo de energı´a per cápita,
%X11 = Emisión de CO2 per capita.
```

El índice 1 combina básicamente todo excepto el %mujeres en pobl. activa y la proporción de bosque. Con intensidead por encima de 0.35 los 2(+),5(-),6(-),10(-), 11(-)... se trata de un índice de desarrollo (comunicaciones, electricidad, energía)...

El índice 2 se centra en 1-crecimiento demográfico(-), 3-mujeres activas(--), 7 consumo de agua(-) y 8 bosques(++).

Pero, los signos hay que tener en cuenta que la *corrección logarítmica de la skewness* ha "dado la vuelta" a las mujeres activas, con lo que respecto a datos originales es:

1-crecimiento demográfico(-), 3-mujeres activas(++), 7 consumo de agua(-) y 8 bosques(++).

Este índice es un índice más sociocultural, los de valor bajo tienen mucho crecimiento demográfico, pocas mujeres trabajando, pocos bosques... por el contrario, los altos son poco crecimiento, muchas mujeres trabajando, más bosques.

Análisis de los últimos componentes principales (menor varianza)

Los últimos componentes principales son el "modelo" (combinación lineal de variables que en todos los países tiene una varianza menor)... Podríamos sacar alguna conclusión sobre la exactitud de los datos comprobando qué muestras (países) se desvían más del modelo...

```
[B,I]=sort(Comp_prin(:,11).^2,'descend');
%algún dato erróneo o son paises atípicos?
{nombres{I(1:4)}}
```

```
ans = 1×4 cell
'Islandia'  'CoreaNort' 'Kuwait'    'Suecia'
```

Estos son los países que más se ajustan al modelo "medio"... ¿Quizás también por haber "falseado" o "inventado" algún dato para parecer "promedio"?

```
[Comp_prin(I(92:96),11) I(92:96)]
```

```
ans = 5×2
    0.0094   72.0000
    0.0071    7.0000
    0.0055    8.0000
    0.0032   94.0000
    0.0008   20.0000
```

```
{nombres{I(92:96)}}
```

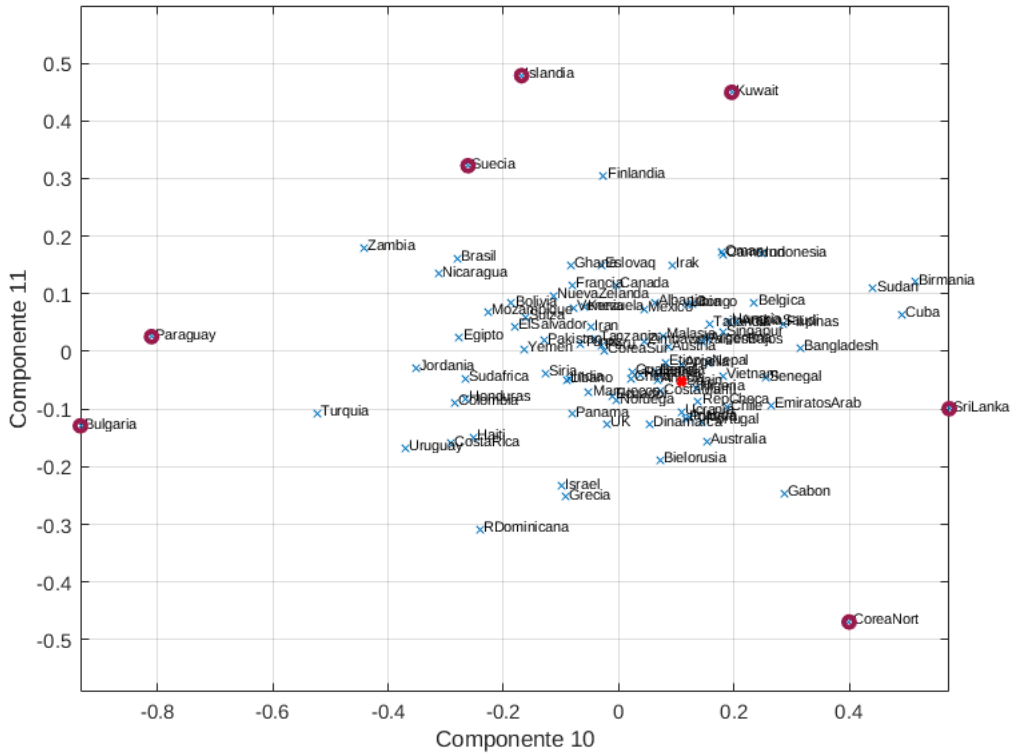
```
ans = 1×5 cell
'Peru'      'Austria'   'Bangladesh''Yemen'     'CoreaSur'
```

Si consideremos como "dentro de modelo" los dos últimos componentes principales, entonces ordenaríamos por "elipsoide" de confianza:

```
[B2,I2]=sort(Comp_prin(:,11).^2/S(11,11)^2+Comp_prin(:,10).^2/S(10,10)^2,'descend');
{nombres{I2(1:7)}}
```

```
ans = 1×7 cell
'Bulgaria'  'CoreaNort' 'Islandia'  'Paraguay'  'Kuwait'    'Suecia'    'SriLank···
```

```
plotpca2D(10,11,Comp_prin,nombres)
hold on
plot(Comp_prin(I2(1:7),10),Comp_prin(I2(1:7),11),'o','Color',[0.6 0.1 0.3],'LineWidth',
hold off, axis equal
```



Conclusiones

Hemos cargado datos, y los hemos preprocesado para centrarlos alrededor de la media con una transformación logarítmica + offset (aunque si el offset es grande comparado con el rango de los datos la distorsión sería poca). Luego los hemos trasladado a media cero y varianza unidad (estandarizado).

Hemos analizado los países que más se desviaban del "modelo que siguen casi todos" (últimos componentes) y las combinaciones que más diferencian a los países, encontrando un componente de "consumo y tecnificación" y otro más "sociocultural".

La forma de "Y" de las proyecciones sobre los primeros componentes apunta a que podría existir alguna relación no lineal no contemplada: otras transformaciones (por ejemplo, los datos en bruto) o la inclusión de otras no-linealidades (productos, potencias)... podrían dar datos diferentes.

Nótese, asimismo, que la clasificación abordada es "no supervisada". El caso "supervisado" sería determinar los componentes que más ayudan a predecir determinados índices o clasificaciones preestablecidas; éstos se obtendrían con técnicas PLS.

Apéndice: funciones auxiliares

```
function plotpca2D(comp1,comp2,Datos,Nombres)
arguments
   comp1 {mustBeInteger}
   comp2 {mustBeInteger}
   Datos
```

```
   Nombres={}
end
tf=ishold;
npaises=size(Datos,1);
plot(Datos(:,comp1),Datos(:,comp2),'x')
hold on
rr=axis;
grid on
if(~isempty(Nombres))
   for i=1:npaises
       text(Datos(i,comp1)+.01*rr(2),Datos(i,comp2)+.01*rr(4),Nombres{i},'FontSize',8)
   end
end
plot(Datos(33,comp1),Datos(33,comp2),'xr','LineWidth',3) %ESPAÑA
xlabel('Componente '+string(comp1))
ylabel('Componente '+string(comp2))
if(~tf)
   hold off
end
end
```