Tecnológico de Monterrey

Campus Estado de México

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 101)

Actividad 2.Componentes Principales

Juan Pablo Castañeda Serrano A01752030

PARTE I

1. Calcule las matrices de varianza-covarianza S con **cov**(X) y la matriz de correlaciones con **cor**(X)

```
# 1. Leer el archivo y calcular matrices
data = pd.read_csv('paises_mundo.csv')
cov_matrix = data.cov()
corr_matrix = data.corr()
print("cov_matrix = ")
print(cov_matrix)
print("corr_matrix = ")
print(corr_matrix)
print(90*"*")
```

```
cov_matrix =
                CrecPobl      MortInf  PorcMujeres          PNB95  ...      PropBosq    PropDefor     ConsEner       EmisCO2
CrecPobl        1.538298 2.195026e+01    -6.078026  -8.933379e+04  ...     -3.887018     0.336197 -8.384169e+02 -1.137877
MortInf        21.950263 1.032859e+03    -9.249342  -2.269332e+06  ...    -14.663158    12.762961 -4.442568e+04 -94.855000
PorcMujeres    -6.078026 -9.249342e+00    76.983224   2.813114e+05  ...     65.178947     0.268059  2.855207e+02 -2.150132
PNB95      -89333.788772 -2.269332e+06 281311.418421   4.999786e+10  ... 247431.122807 -58062.027632  1.415628e+08 250167.323509
ProdElec   -49739.635746 -1.043435e+06 226024.813487   2.247791e+10  ...  70359.785965 -31803.401546  6.801296e+07 139277.888640
LinTelf      -136.907895 -4.381366e+03    449.975000   2.039550e+07  ...    248.715789    -99.404605  3.426262e+05    638.570000
ConsAgua      -48.270921 -1.288211e+03  -1568.313487   1.097481e+07  ...  -2220.757895    -67.437928  2.092242e+05    486.932763
PropBosq       -3.887018 -1.466316e+01     65.178947   2.474311e+05  ...    401.003509     2.625263 -5.153439e+03    -12.897193
PropDefor       0.336197  1.276296e+01      0.268059  -5.806203e+04  ...      2.625263     1.817253 -1.051522e+03     -2.632487
ConsEner     -838.416886 -4.442568e+04    285.520724   1.415628e+08  ...  -5153.438596 -1051.521875  5.014395e+06  10286.159781
EmisCO2        -1.137877 -9.485500e+01     -2.150132   2.501673e+05  ...    -12.897193    -2.632487  1.028616e+04     27.268614
```

```
corr_matrix =
             CrecPobl   MortInf  PorcMujeres     PNB95  ProdElec   LinTelf  ConsAgua  PropBosq  PropDefor  ConsEner   EmisCO2
CrecPobl     1.000000  0.550679    -0.558527 -0.322122 -0.297111 -0.563212 -0.067730 -0.156503   0.201079 -0.301877 -0.175689
MortInf      0.550679  1.000000    -0.032801 -0.315792 -0.240537 -0.695589 -0.069756 -0.022784   0.294593 -0.617311 -0.565208
PorcMujeres -0.558527 -0.032801     1.000000  0.143388  0.190851  0.261670 -0.311062  0.370967   0.022663  0.014532 -0.046928
PNB95       -0.322122 -0.315792     0.143388  1.000000  0.744761  0.465396  0.085415  0.055259  -0.192623  0.282725  0.214251
ProdElec    -0.297111 -0.240537     0.190851  0.744761  1.000000  0.286645  0.180477  0.026031  -0.174784  0.225019  0.197600
LinTelf     -0.563212 -0.695589     0.261670  0.465396  0.286645  1.000000  0.105939  0.063371  -0.376238  0.780684  0.623937
ConsAgua    -0.067730 -0.069756    -0.311062  0.085415  0.180477  0.105939  1.000000 -0.192992  -0.087058  0.162598  0.162274
PropBosq    -0.156503 -0.022784     0.370967  0.055259  0.026031  0.063371 -0.192992  1.000000   0.097250 -0.114925 -0.123336
PropDefor    0.201079  0.294593     0.022663 -0.192623 -0.174784 -0.376238 -0.087058  0.097250   1.000000 -0.348338 -0.373962
ConsEner    -0.301877 -0.617311     0.014532  0.282725  0.225019  0.780684  0.162598 -0.114925  -0.348338  1.000000  0.879655
EmisCO2     -0.175689 -0.565208    -0.046928  0.214251  0.197600  0.623937  0.162274 -0.123336  -0.373962  0.879655  1.000000
**********
```

2. Calcule los valores y vectores propios de cada matriz (Vectores cortados por espacio)

```
# 2. Valores y vectores propios
eigenvalues_cov, eigenvectors_cov = np.linalg.eig(cov_matrix)
eigenvalues_corr, eigenvectors_corr = np.linalg.eig(corr_matrix)

print("eigenvalues_corr = ")
print(eigenvalues_corr)
print("eigenvalues_cov = ")
print(eigenvalues_cov)
print(90*"*")
```

```
eigenvectors_cov =
[[-1.65816773e-06  4.70678461e-07 -1.26373574e-04 -1.92840782e-05
   5.53739707e-03  1.24345612e-02 -5.35908939e-03 -8.39081044e-02
   6.77835779e-02  9.87288738e-01 -1.15809096e-01]
 [-4.04813855e-05 -1.77425442e-05 -8.22538212e-03 -2.49325727e-03
   9.44030203e-02  9.91751509e-01 -2.25801958e-02 -7.89112785e-02
   1.63783592e-02 -2.09249092e-02  4.26487185e-04]
 [ 5.73909610e-06 -1.08454281e-05 -1.31814909e-04  5.53830717e-03
  -3.14036410e-02  8.55299154e-02  1.13648080e-01  9.85649803e-01
   1.46846382e-02  8.34432427e-02  8.24146538e-03]
 [ 8.88037597e-01  4.59763191e-01 -2.60220711e-03 -3.89358789e-04
   3.32740919e-04 -8.62100481e-06  7.56647696e-06  1.21724819e-05
   3.97146871e-07  2.72399679e-07  4.27445058e-07]
 [ 4.59763574e-01 -8.88040472e-01 -5.69489558e-04  1.09630500e-03
  -2.20781916e-04  1.95540800e-05 -1.54465759e-05 -2.55899806e-05
  -1.05947081e-06 -2.08685783e-07 -1.35388052e-06]
 [ 3.50434128e-04  4.01617912e-04  6.19424889e-02  7.64117441e-03
  -9.92140449e-01  9.10962152e-02 -4.74868183e-02 -3.41681151e-02
```

```
eigenvectors_corr =
[[ 0.31411941  0.34835747  0.07352541  0.10062784 -0.34674
  -0.09481963  0.16289743  0.44028717  0.32972147 -0.18392
 [ 0.39239544 -0.04136238  0.17759254 -0.17487096  0.38959
  -0.32307802  0.63980408  0.13398483 -0.08340489 -0.08656
 [-0.11654632 -0.58283641 -0.16686305  0.167868   -0.42854
   0.05209889  0.53108671 -0.05865031 -0.186541    0.16835
 [-0.29539377 -0.17690839  0.53343025  0.15247432 -0.34911
  -0.44913216 -0.1490207   0.26248209  0.14110658  0.04653
 [-0.25896472 -0.17356372  0.61438847 -0.12366382  0.33770
   0.50343911  0.10827458  0.17389644  0.07521971  0.02821
 [-0.44608293 -0.02719077 -0.1517725  -0.44992596  0.20997
```

```
eigenvalues_corr =
[4.02987902 1.92999195 1.37041115 0.06935866 0.14632819 0.16806846
 0.32680096 0.57130511 0.86451597 0.79414057 0.72919997]
eigenvalues_cov =
[6.16357629e+10 6.58161227e+09 4.63625593e+06 3.10723182e+05
 1.21601494e+04 5.13776704e+02 3.62788506e+02 4.54208136e+01
 5.80086834e+00 4.76808294e-01 1.43801991e+00]
**********
```

3. Calcule la proporción de varianza explicada por cada componente

```
# 3. Proporción de varianza explicada
total_variance = np.sum(np.diag(cov_matrix))
explained_variance_ratio_cov = eigenvalues_cov / total_variance
explained_variance_ratio_corr = eigenvalues_corr / np.sum(eigenvalues_corr)

print("total_variance = ")
print(total_variance)
print("explained_variance_ratio_corr = ")
print(explained_variance_ratio_corr)
print("explained_variance_ratio_cov = ")
print(explained_variance_ratio_cov)
print(90*"*")
```

```
total_variance =
68222335252.70132
explained_variance_ratio_corr =
[0.36635264 0.17545381 0.12458283 0.00630533 0.01330256 0.01527895
 0.02970918 0.05193683 0.07859236 0.0721946  0.06629091]
explained_variance_ratio_cov =
[9.03454311e-01 9.64729842e-02 6.79580362e-05 4.55456679e-06
 1.78242937e-07 7.53091641e-09 5.31773802e-09 6.65776295e-10
 8.50288738e-11 6.98903508e-12 2.10784328e-11]
**********
```

4. Acumule los resultados anteriores

```
# 4. Acumule los resultados
cum_explained_variance_cov = np.cumsum(explained_variance_ratio_cov)
cum_explained_variance_corr = np.cumsum(explained_variance_ratio_corr)
print("cum_explained_variance_cov = ")
print(cum_explained_variance_cov)
print("cum_explained_variance_corr = ")
print(explained_variance_ratio_corr)

print(90*"*")
```

```
cum_explained_variance_cov =
[0.90345431 0.9999273  0.99999525 0.99999981 0.99999999 0.99999999
 1.         1.         1.         1.         1.        ]
cum_explained_variance_corr =
[0.36635264 0.17545381 0.12458283 0.00630533 0.01330256 0.01527895
 0.02970918 0.05193683 0.07859236 0.0721946  0.06629091]
**************
```

5. Según los resultados anteriores, ¿qué componentes son los más importantes? ¿qué variables son las que más contribuyen a la primera y segunda componentes principales? ¿por qué lo dice? ¿Cómo influyen las unidades de las variables?

Componentes más importantes

Matriz de correlación: Las tres primeras componentes son las más importantes basándonos en los valores propios.

Matriz de varianza-covarianza: Las dos primeras componentes son las más importantes basándonos en los valores propios.

Variables que más contribuyen a las componentes principales

Para la matriz de varianza-covarianza (S):

Primer componente (columna 0 de eigenvectors_cov): Las variables con los valores más grandes en magnitud son PNB95 y ProdElec. Estas dos variables son las que más contribuyen a la primera componente principal.

Segunda componente (columna 1 de eigenvectors_cov): Al igual que antes, las variables PNB95 y ProdElec tienen los valores más grandes en magnitud y, por lo tanto, son las que más contribuyen.

Para la matriz de correlaciones (R):

El primer componente(columna 0 de eigenvectors_corr): Las variables con mayores valores en magnitud son CrecPobl, MortInf, LinTelf. Estas variables tienen la mayor contribución a la primera componente principal.

Segundo componente (columna 1 de eigenvectors_corr): Las variables con mayores valores en magnitud son CrecPobl, PorcMujeres y PropBosq. Estas variables son las que más contribuyen a la segunda componente principal.

Interpretación

Los vectores propios indican cómo se ponderan las variables originales para formar la componente principal. Un valor grande en magnitud para una variable en un vector propio indica que esa variable es una contribución significativa a esa componente.

Para la matriz S: Las variables PNB95 y ProdElec dominan tanto la primera como la segunda componente. Esto podría indicar que estas variables tienen grandes varianzas (o co-varianzas con otras variables) en comparación con las demás, y esto es reflejado en las componentes principales.

Para la matriz R: Las variables que más contribuyen cambian entre componentes. Por ejemplo, CrecPobl tiene una alta contribución en ambas componentes, mientras que otras como MortInf y LinTelf son importantes para la primera componente y PorcMujeres y PropBosq para la segunda. Dado que esta matriz está normalizada por la varianza de las variables, esta diferencia sugiere estructuras de correlación subyacentes entre las variables.

Influencia de las unidades

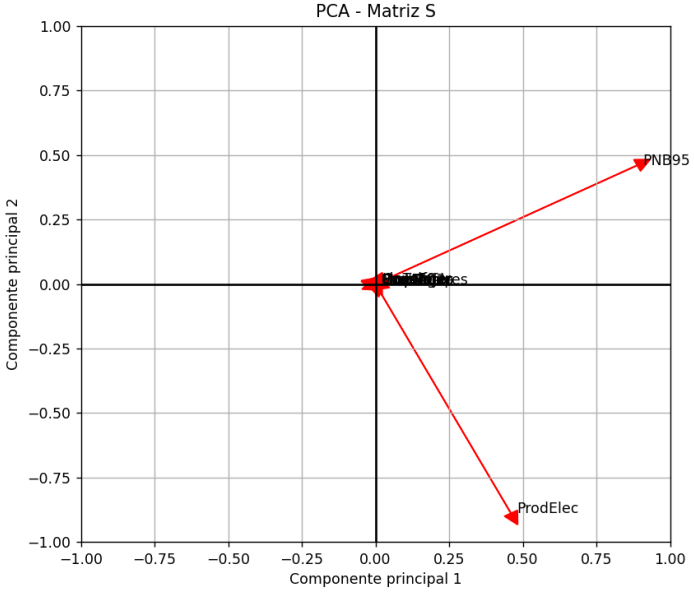
Las unidades de las variables definitivamente influyen cuando se utiliza la matriz de varianza-covarianza. Esto se puede ver en cómo las variables PNB95 y ProdElec dominan las componentes principales. Si estas variables tienen unidades que son magnitudes más grandes que las demás, sus varianzas (o co-varianzas) también serán grandes, lo que les dará un peso desproporcionado en el análisis PCA basado en S.

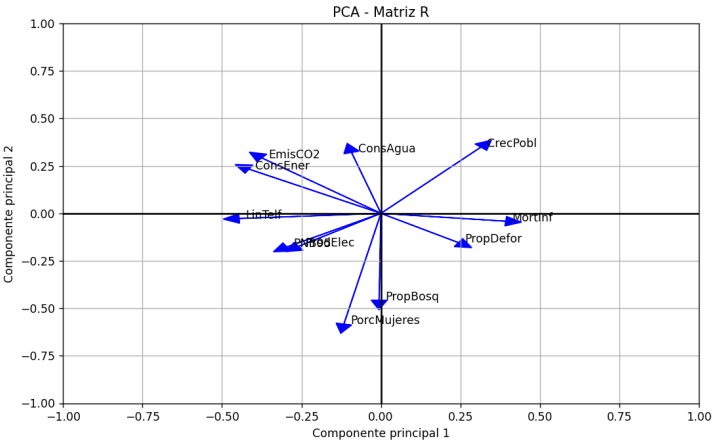
Por otro lado, al utilizar la matriz de correlaciones, las unidades ya no influyen porque todas las variables están normalizadas por sus desviaciones estándar. Por lo tanto, el análisis PCA basado en R proporciona un vistazo a las estructuras de correlación subyacentes independientemente de las magnitudes originales de las variables.

Para concluir, si se sospecha que las unidades de las variables pueden influir en el análisis, es recomendable usar la matriz de correlaciones en lugar de la matriz de varianza-covarianza.

PARTE II

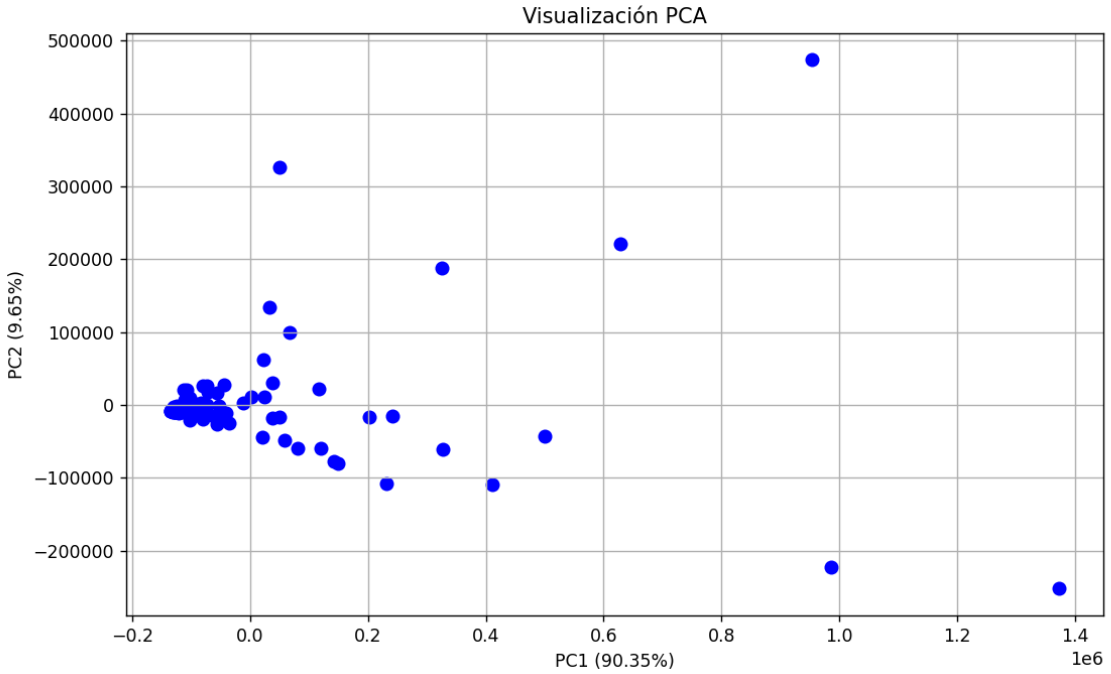
1. Obtenga las gráficas de respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes e interprete los resultados en término de agrupación de variables (puede ayudar "índice de riqueza", "índice de ruralidad")



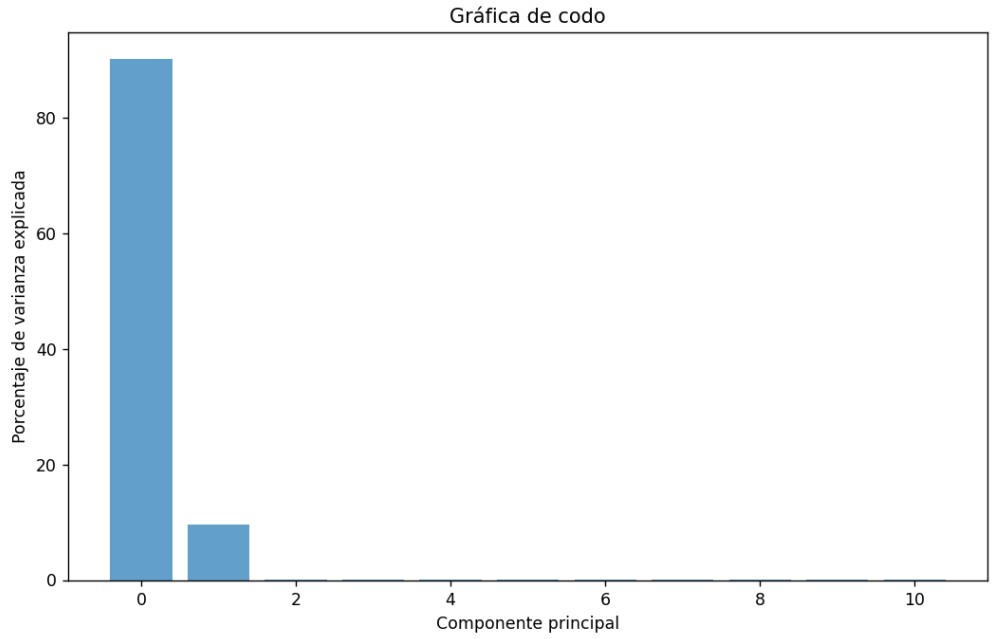


PARTE III:

1. Gráfica PCA
   Este gráfico muestra cómo se distribuyen las observaciones en el espacio definido por las dos primeras componentes principales Cómo podemos ver, hay una gran agrupación de datos cerca del cero, haciendo que nuestros outliers destaquen de manera dramática



2. Gráfica de codo
   La gráfica de codo muestra la cantidad de varianza explicada por cada componente principal. Como podemos ver, nuestros outliers referentes a la aportación de la riqueza son los que más destacan



3. Contribución de cada variable al PC1
   Esta gráfica muestra la importancia relativa (o contribución) de cada variable al primer componente principal. Las variables con contribuciones más altas son las que tienen un impacto mayor en ese componente. Esto te permite entender qué variables están más relacionadas con las variaciones capturadas por ese componente principal.