UNIVERSIDADE DA CORUÑA

Modelización Estadística de Datos de Alta Dimensión

Escalado multidimensional de la empleabilidad provincial española

facultade de informática da coruña

El presente estudio ha sido realizado en su integridad por *Jorge Crespo Rivas* (j.crespo.rivas@udc.es) y *Jesús Estévez Amoedo* (j.esteveza@udc.es), estudiantes del Grado en Ciencia e Ingeniería de Datos de la Universidade da Coruña e integrantes del grupo de prácticas número 3 de la asignatura Modelización Estadística de Datos de Alta Dimensión.

Contents.

INTRODUCCIÓN.

El estudio de la empleabilidad a escala subnacional resulta esencial para comprender las dinámicas laborales, las diferencias territoriales en oportunidades económicas y la eficacia de las políticas públicas. Analizar cómo varían indicadores demográficos y económicos entre provincias permite identificar patrones estructurales (movilidad poblacional, niveles de actividad y desempleo, y diferencias en el desempeño macroeconómico) que no son observables mediante análisis parciales y que, por tanto, requieren un tratamiento multivariante riguroso.

Para este trabajo se emplea una base de datos provincial correspondiente al año 2017, extraída del Instituto Nacional de Estadística y disponible en el fichero de trabajo. Las variables consideradas resumen aspectos demográficos y del mercado laboral que, tratadas conjuntamente, permiten captar la heterogeneidad regional y las relaciones complejas entre las dimensiones demográfica, laboral y económica, aspectos que no resultan evidentes mediante análisis univariantes aislados.

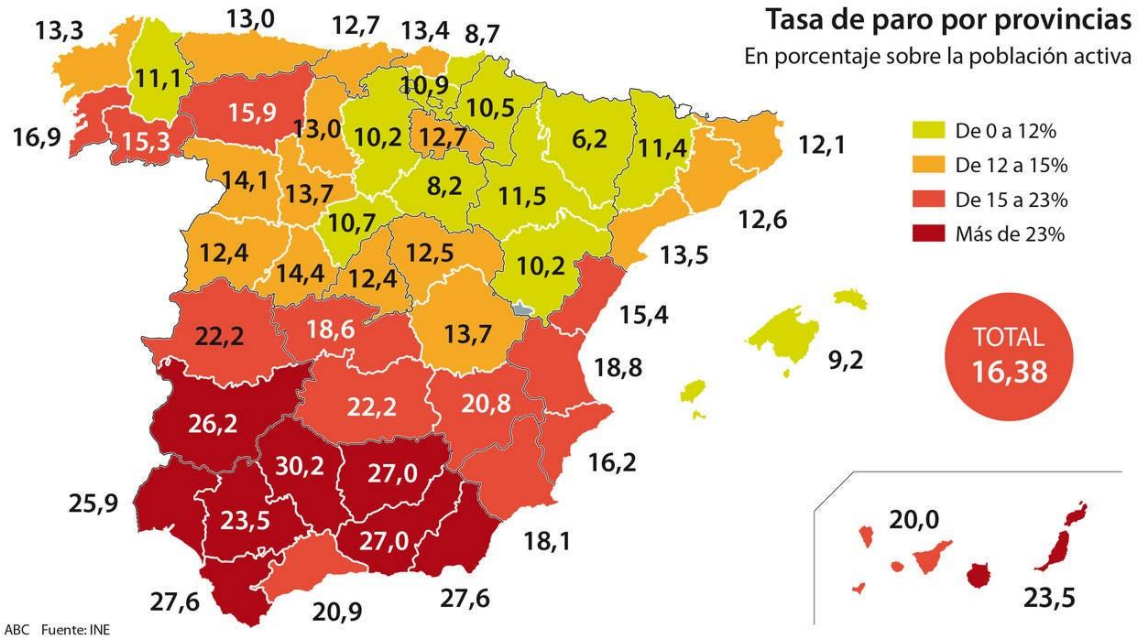

Figure 1: Tasa de paro por provincias en % sobre población activa en 2017. [1]

El objetivo central de este trabajo es encontrar una representación en un espacio euclídeo de dimensión reducida en la que cada punto represente una provincia y las distancias entre puntos aproximen, en la medida de lo posible, una medida razonable de disimilitud entre las observaciones. Este planteamiento corresponde a la familia de métodos de escalado multidimensional (MDS) o coordenadas principales, cuyo propósito es facilitar la visualización y la interpretación de estructuras de semejanza/diferencia en datos multivariantes. Además, el trabajo abordará cuestiones teóricas y prácticas relevantes: la posible no unicidad de la solución (debida a rotaciones y reflexiones en el espacio), el efecto de cambios de unidad en las variables sobre la configuración final y la elección de la dimensión de representación que balancee el resultado respecto al conjunto de datos. Se incluirá también la representación gráfica de las configuraciones (con etiquetas identificativas por provincia), la exploración de la necesidad de rotaciones o simetrías para facilitar la interpretación, y la extensión del procedimiento para representar, alternativamente, las propias variables en el mismo marco geométrico.

Como tal, esta investigación combina un enfoque descriptivo y computacional con un fundamento teórico riguroso (basado siempre en la materia contenida en apuntes y clases de la asignatura) para evaluar si la representación espacial obtenida resulta útil para captar patrones relevantes en la empleabilidad provincial y qué limitaciones emergen de la metodología aplicada. A lo largo de esta memoria se documentarán las decisiones metodológicas, los cálculos intermedios y las interpretaciones derivadas de las gráficas finales, permitiendo una valoración crítica de la aplicabilidad del método al conjunto de datos considerado.

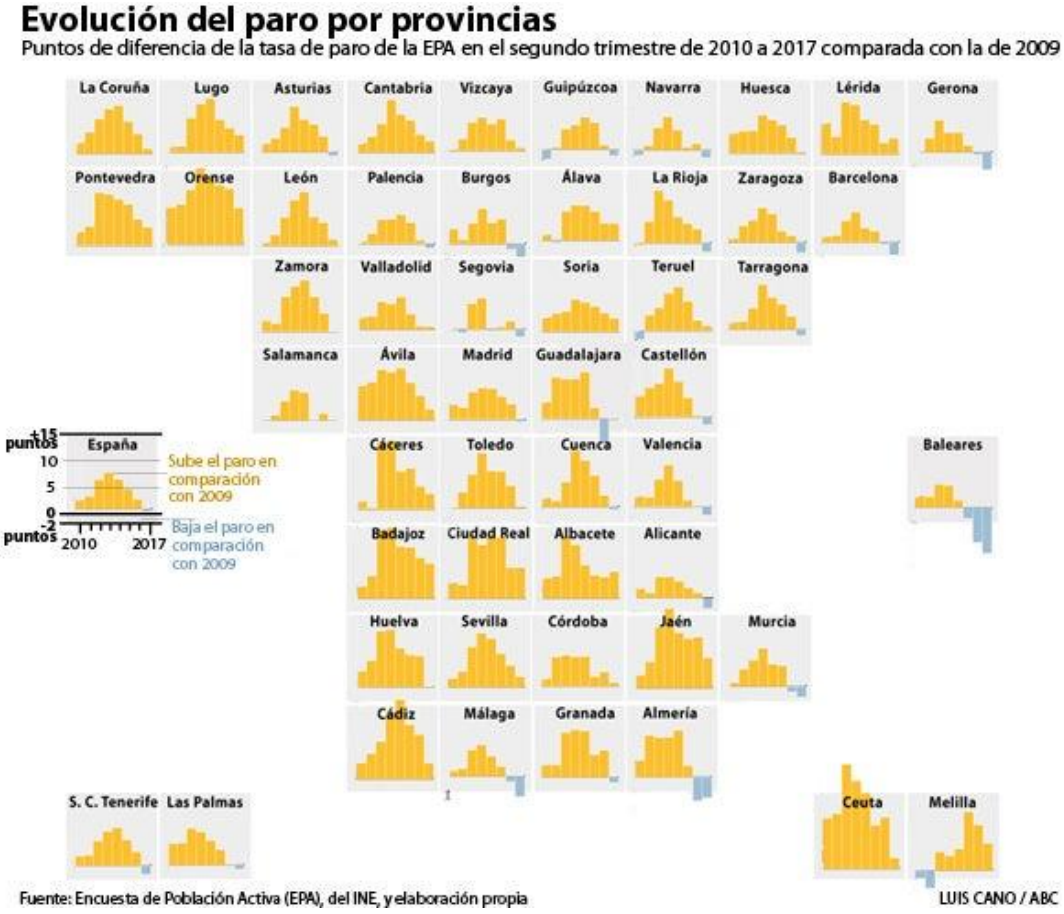

Figure 2: Evolución puntos diferenciales en tasa de paro 2010-2017 vs. 2009. [2]

EJERCICIO 1: Realizar un estudio descriptivo (univariante y multivariante) de las variables que se consideren de interés.

ESTUDIO DESCRIPTIVO UNIVARIANTE

Para la realización de este ejercicio, en primer lugar, visualizaremos las primeras filas de los datos con los que trabajaremos.

```
> head(employment)
                inm.1000   em.1000 activos empleados en.paro     ipc      pib
Albacete         9.658594 13.084991   57.16     41.55   27.30 101.391 17383.24
Alicante/Alacant 10.258191  9.872011   57.35     43.07   24.91 101.047 17562.79
Almería          12.998179 11.081435   63.52     40.85   35.70 101.246 17287.07
Araba/Álava      15.118920 11.433495   61.09     50.93   16.64 100.306 34053.30
Asturias          6.237551  7.617976   51.63     40.90   20.78 100.756 19505.32
Ávila            16.740323 20.871446   53.30     39.80   25.33 101.999 17622.24
```

Figure 3: Primeras filas del dataset "employment" asignado.

Tras ejecutar la función head, observamos que la base de datos consta de 7 variables, todas ellas de carácter numérico (cuantitativas), siendo representadas inm.1000 y em.1000 por cada mil (‰), por otra parte, activos, empleados y en.paro por cada cien (%) y, por último, siendo ipc y pib índices/indicadores.

También analizaremos el sumario de los datos analizados.

```
> summary(employment)
    inm.1000         em.1000         activos        empleados       en.paro     
 Min.   : 6.238   Min.   : 7.080   Min.   :48.16   Min.   :32.78   Min.   :14.05  
 1st Qu.: 8.816   1st Qu.: 9.094   1st Qu.:55.49   1st Qu.:39.96   1st Qu.:18.42  
 Median :10.342   Median :11.860   Median :57.75   Median :43.05   Median :23.17  
 Mean   :11.387   Mean   :12.617   Mean   :57.92   Mean   :43.90   Mean   :24.19  
 3rd Qu.:13.210   3rd Qu.:14.097   3rd Qu.:60.84   3rd Qu.:48.00   3rd Qu.:29.37  
 Max.   :27.724   Max.   :26.380   Max.   :65.84   Max.   :53.26   Max.   :42.34  
      ipc              pib        
 Min.   : 99.83   Min.   :14957  
 1st Qu.:100.72   1st Qu.:17518  
 Median :100.93   Median :19016  
 Mean   :100.95   Mean   :20832  
 3rd Qu.:101.18   3rd Qu.:24376  
 Max.   :102.00   Max.   :34053  
```

Figure 4: Sumario del dataset "employment".

Para realizar el estudio, hemos decidido descartar como variables útiles tanto el pib como el ipc. La información que nos brindan estos indicadores podría ser conveniente en otros contextos en los que se tuviera más información de cada provincia, como por ejemplo la población o el coste medio de la vivienda, pero no en este caso.

Analizando de manera aislada las variables representadas por cada mil (haciendo uso de la función barplot) obtenemos los siguientes resultados:

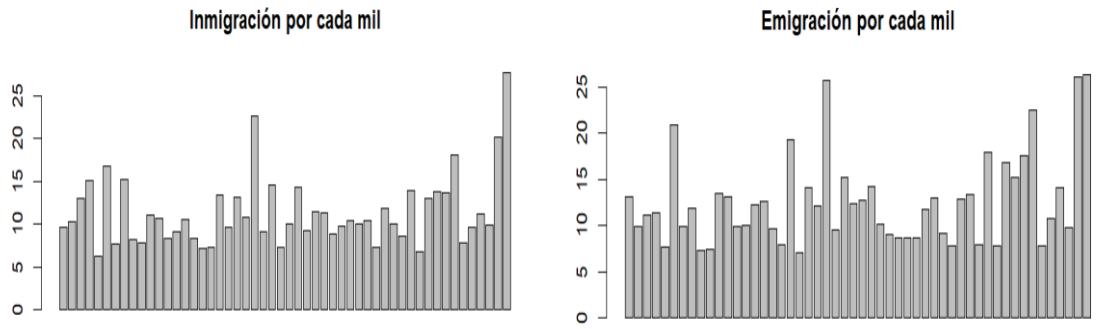

Figure 5: Barplot de las variables inmigración y emigración.

A partir de este análisis se observa que, tanto en términos de inmigración como de emigración, la mayoría de las provincias no supera los 20 habitantes por cada mil habitantes, si bien la emigración tiende a ser sensiblemente mayor. Existen algunas provincias atípicas que muestran niveles de inmigración y emigración notablemente superiores a la media; en ningún caso se registran cifras inferiores, lo cual es coherente dado que tasas por debajo de 5 por mil serían extraordinariamente bajas. En la práctica, la tendencia predominante sitúa a la gran mayoría de las provincias en un intervalo aproximado de 5 a 15 habitantes por cada mil.

De manera alternativa, y mediante el uso de la función boxplot, hemos analizado también de forma aislada las variables representadas de manera porcentual, obteniendo así los siguientes gráficos:

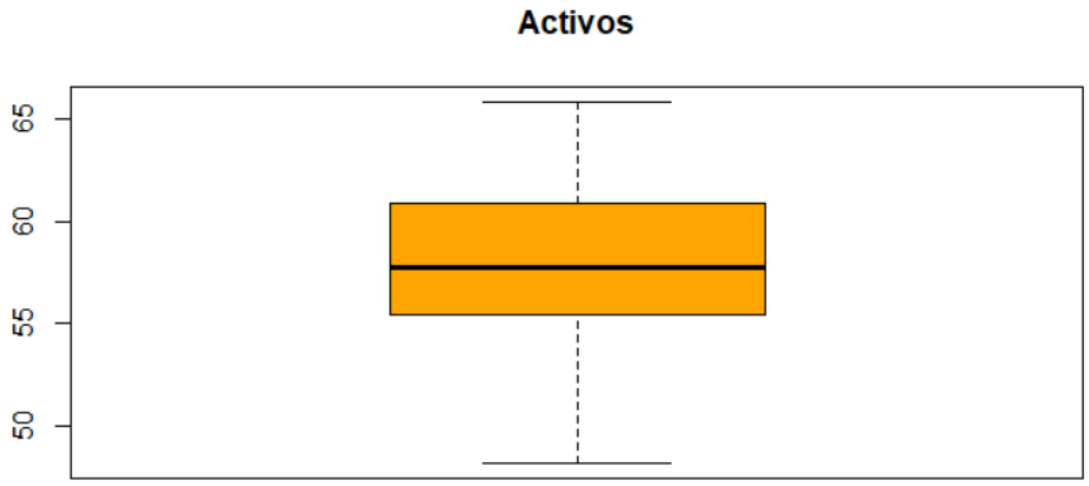

Figure 6: Boxplot de población activa.

Para los activos, la mediana se encuentra aproximadamente en 57.5% y la caja es estrecha (abarca solo un 5% de la totalidad), lo cual indica que la inmensa mayoría de las provincias cuentan con más de la mitad de la población en activo.

A mayores, no nos hemos encontramos con atípicos y los bigotes apenas se alejan de la caja, lo cual sugiere que no hay provincias con una cantidad muy baja o alta de activos comparado con el resto.

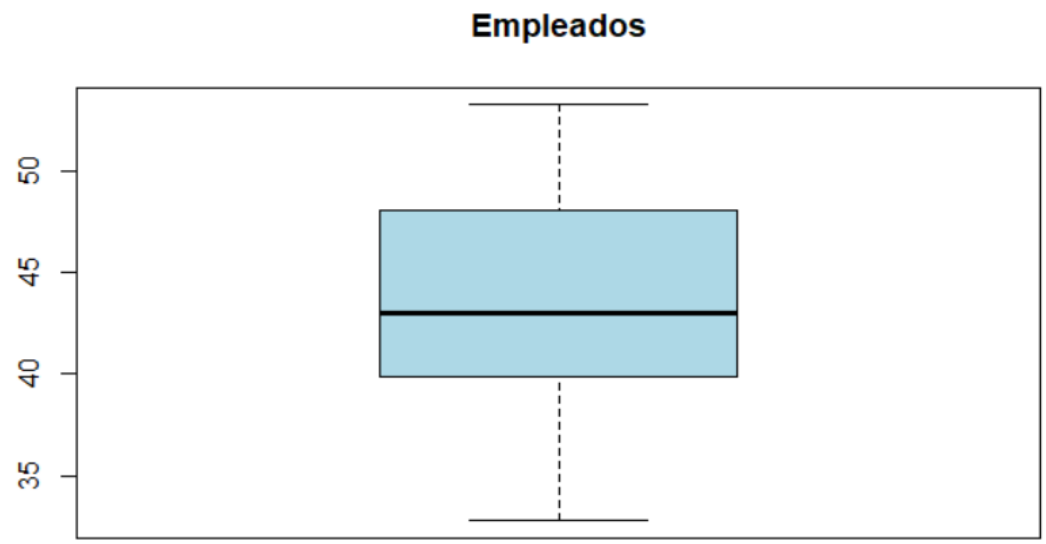

Figure 7: Boxplot de población empleada.

Para los empleados, la mediana con respecto a la caja es la más asimétrica de las tres analizadas (rondando el 43% aproximadamente), aunque tampoco lo suficiente como para destacarlo o tenerlo en cuenta. Los bigotes en este caso tampoco se encuentran especialmente distantes.

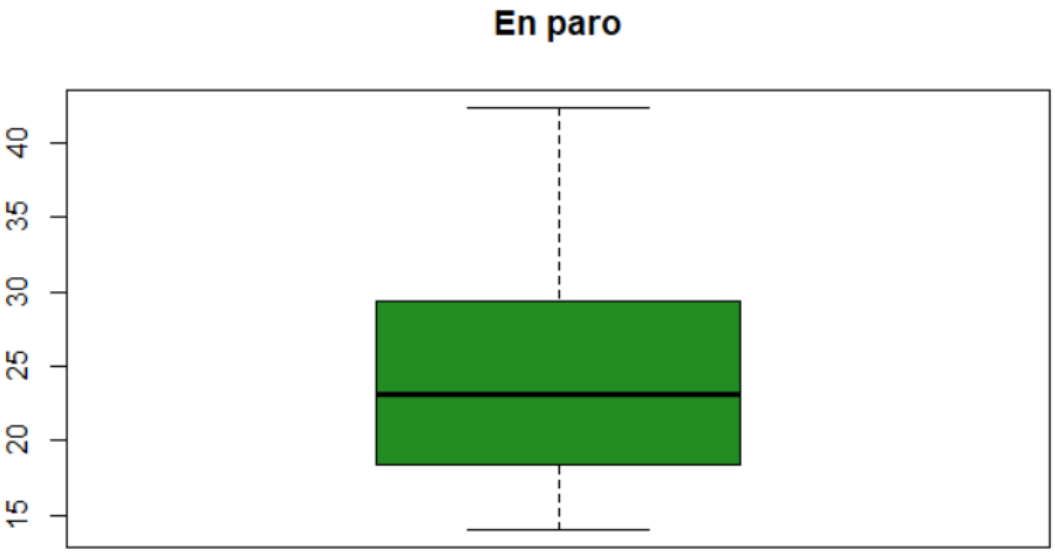

Figure 8: Boxplot de población en paro.

Con relación al paro, la caja también es estrecha y la mediana simétrica como en los casos anteriores. Esta nos indica que la mayoría de los datos (el 50% de las observaciones) están contenidos entre estos valores, lo que sugiere que la tasa de paro en la mayoría de las provincias no es extremadamente baja ni excesivamente alta.

El bigote superior dista bastante de la caja, lo que indica que hay algunas provincias con una tasa de paro alta, y que el paro en la mayoría de las provincias es una cuestión moderada.

Todos estos datos han sido analizados realizando un análisis descriptivo univariante, lo cual resulta muy limitante a la hora de interpretar los resultados. Mediante el estudio univariante de las variables es posible que se esté perdiendo información relevante que podría explicar patrones o comportamientos (como por ejemplo el paro en una región, que podría depender de factores que no se reflejan si solo se observa la tasa de paro de manera aislada).

A continuación, y dadas dichas circunstancias, se realizará el estudio descriptivo multivariante.

ESTUDIO DESCRIPTIVO MULTIVARIANTE

Para analizar estos datos de manera multivariante hemos decidido que podría ser interesante realizar un gráfico comparando la diferencia entre inmigrantes/emigrantes con la tasa de paro.

Tras observar el gráfico, vemos inapreciable algún tipo de patrón o relación entre dicha diferencia, asique, a modo de confirmación, denotamos que la correlación entre ambas variables es muy pequeña.

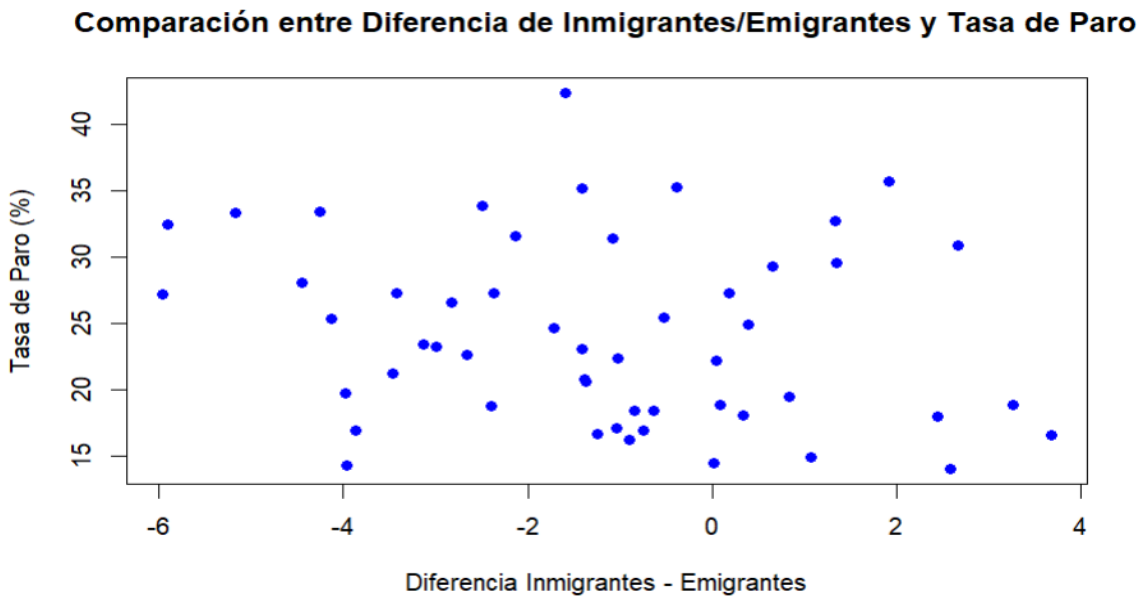

Figure 9: Gráfico inm/em vs tasa de paro.

En un primer momento, pensamos que sería una buena idea comparar estas 2 variables, ya que podría haber indicios de que en las provincias en las que entra más gente de la que se va habría un aumento en la demanda del trabajo, por lo que incrementaría el paro.

Como es coherente, la diferencia entre inmigrantes y emigrantes tampoco es explicativa con respecto a la tasa de empleabilidad.

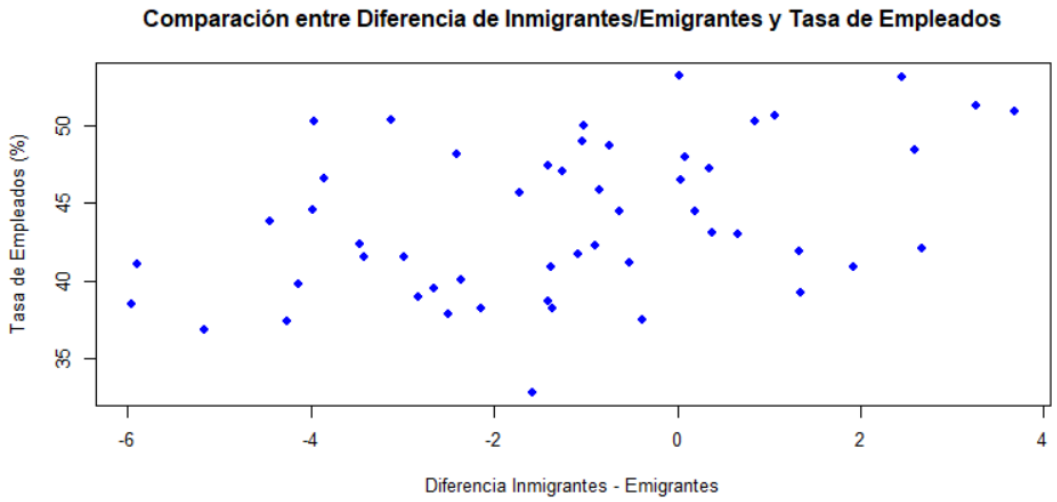

Figure 10: Gráfico inm/em vs tasa de empleo.

Es debido a esto que hemos decidido incluir el PIB como variable de interés, y compararlo con las tasas de empleados y de paro en busca de una relación. Un resultado coherente y esperable ante este análisis sería que las provincias con más PIB tuvieran también una mayor tasa de empleabilidad y una menor tasa de paro, debido a que este índice mide el valor de todos los bienes y servicios producidos en la economía anual de cada provincia.

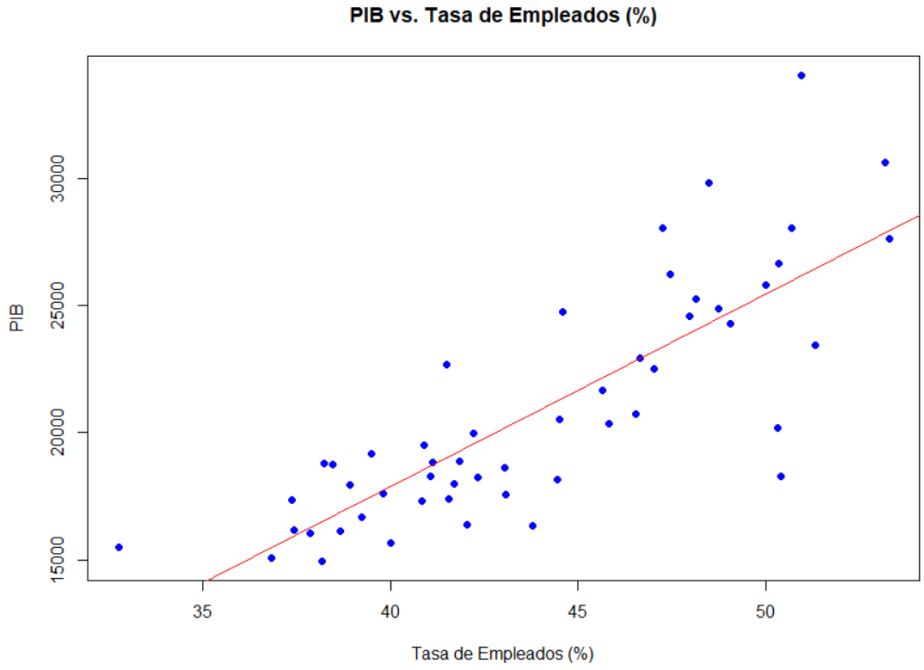

Figure 11: Gráfico PIB vs tasa de empleo.

Efectivamente, tras analizar la gráfica denotamos una apreciable relación entre PIB y tasa de empleabilidad. Por norma general, cuanto más PIB tiene una provincia, mayor es su tasa de empleo, como habíamos estimado anteriormente.

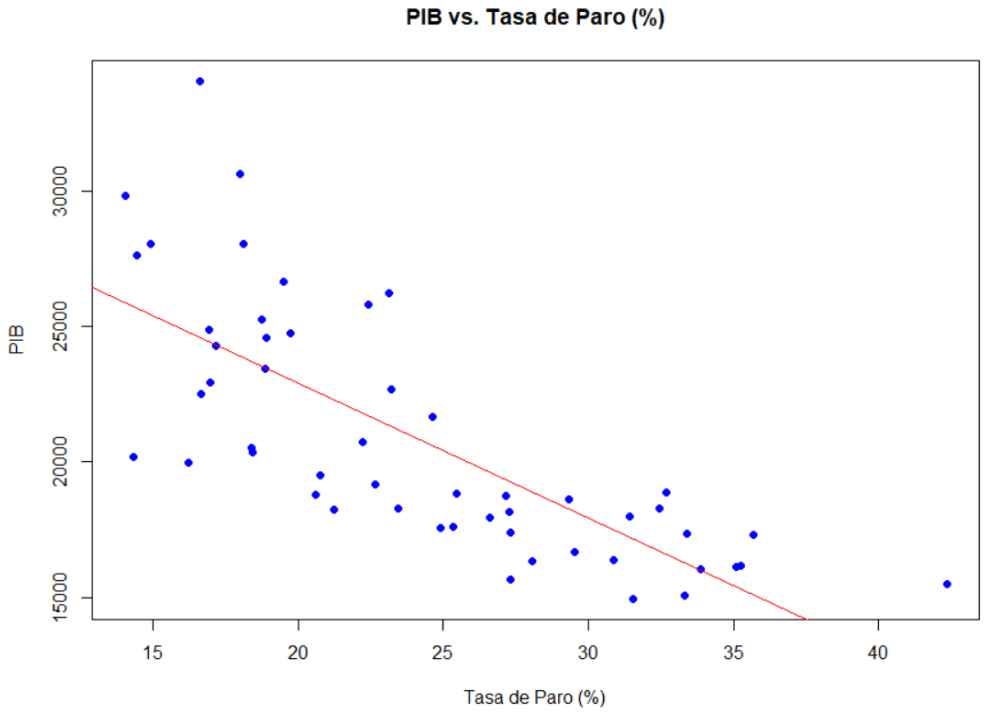

Figure 12: Gráfico PIB vs tasa de paro.

Como también era de esperar, a menor PIB, generalmente se le asocia una menor tasa de paro. Concluimos de esta manera nuestro estudio multivariante.

EJERCICIO 2: Calcular la matriz de distancias de las observaciones de la muestra.

Para el cálculo de la matriz de distancias hemos empleado la librería StatMatch y la función mahalanobis.dist de R, obteniendo así el siguiente resultado:

```
> Distancias.emp
                    Albacete Alicante/Alacant  Almería Araba/Álava Asturias    Ávila  Badajoz Balears, Illes Barcelona  Bizkaia   Burgos  Cáceres    Cádiz Cantabria
Albacete                  NA          2.120772 4.680367    4.941293 2.896628 2.833361 1.676642       3.887639  3.815114 3.401468 1.946732 2.213539 4.180445 1.8823911
Alicante/Alacant    2.120772                NA 4.650005    4.802153 2.866744 3.315578 1.830818       2.536960  3.953873 3.804122 3.133052 2.077203 4.558379 1.5264730
Almería             4.680367          4.650005       NA    5.670927 5.339424 5.346923 5.323894       5.778885  6.150243 5.248541 4.986773 5.636751 6.103339 4.6893884
Araba/Álava         4.941293          4.802153 5.670927          NA 4.998245 4.864594 5.170694       4.537181  4.530197 3.281017 3.874138 5.165275 5.068070 4.8888012
Asturias            2.896628          2.866744 5.339424    4.998245       NA 4.392432 3.279631       4.205888  3.414768 2.510000 2.959333 2.307752 4.325154 2.4624308
Ávila               2.833361          3.315578 5.346923    4.864594 4.392432       NA 3.210215       4.812575  5.874794 4.875215 3.548651 3.768572 5.446231 3.2355638
Badajoz             1.676642          1.830818 5.323894    5.170694 3.279631 3.210215       NA       3.720847  4.188772 4.031083 3.199424 2.097324 3.696824 2.5297615
Balears, Illes      3.887639          2.536960 5.778885    4.537181 4.205888 4.812575 3.720847             NA  3.450813 4.226292 3.896507 3.087327 5.145998 3.3495556
Barcelona           3.815114          3.953873 6.150243    4.530197 3.414768 5.874794 4.188772       3.450813        NA 2.335999 2.939173 3.228781 4.046503 3.9946391
Bizkaia             3.401468          3.804122 5.248541    3.281017 2.510000 4.875215 4.031083       4.226292  2.335999       NA 2.164400 3.445513 3.999470 3.4353525
Burgos              1.946732          3.133052 4.986773    3.874138 2.959333 3.548651 3.199424       3.896507  2.939173 2.164400       NA 2.979867 4.505723 2.4686486
Cáceres             2.213539          2.077203 5.636751    5.165275 2.307752 3.768572 2.097324       3.087327  3.228781 3.445513 2.979867       NA 3.529411 2.6027443
Cádiz               4.180445          4.558379 6.103339    5.068070 4.325154 5.446231 3.696824       5.145998  4.046503 3.999470 4.505723 3.529411       NA 5.2863271
Cantabria           1.882391          1.526473 4.689388    4.888801 2.462431 3.235564 2.529762       3.349556  3.994639 3.435352 2.468649 2.602744 5.286327       NA
Castellón/Castelló  1.396358          2.533047 3.974411    4.279007 3.489442 3.062083 2.582139       3.932101  3.899864 3.218723 1.686063 3.246882 4.694173 2.0925249
Ciudad Real         1.909632          3.391345 5.270154    4.822178 3.967838 3.010933 2.060910       5.017798  4.623944 3.893437 2.834124 3.325672 3.630331 3.4609348
Córdoba             1.823541          2.679692 4.980056    4.912145 2.817727 4.026749 1.775185       4.028288  3.170341 3.079861 2.693929 1.972884 2.538608 3.0770914
Coruña, A           2.330520          2.047768 4.797944    4.892754 1.358476 3.934641 2.926897       3.600445  3.498093 2.794209 2.542937 2.391105 4.922801 1.2281218
Cuenca              2.432102          3.804714 5.927017    4.647375 3.582877 2.508783 3.057108       5.022357  4.593507 3.662165 2.597188 3.006881 3.970550 3.6201981
```

Figure 13: Distancias de Mahalanobis entre provincias.

Debido al elevado número de provincias y, por tanto, el gran tamaño de la matriz, hemos adjuntado solamente el principio de esta. A mayores, hemos incluido en el análisis la distancia mínima y máxima de los elementos menos y más distantes entre sí respectivamente.

```
> min_distancia
[1] 0.9165571
> max_distancia
[1] 7.074022
>
> which(Distancias.emp == min_distancia, arr.ind = TRUE)
         row col
Zaragoza  50  33
Navarra   33  50
> which(Distancias.emp == max_distancia, arr.ind = TRUE)
        row col
Melilla  52   3
Almería   3  52
```

Figure 14: Operaciones restantes de distancias entre provincias.

Como podemos observar, usando la distancia de Mahalanobis, Zaragoza y Navarra resultan las provincias más parecidas entre sí, mientras que Melilla y Almería serían las más distantes.

EJERCICIO 3: ¿Hay una única solución del problema?

Es cierto que existe otra alternativa, por ejemplo, la distancia euclídea; no obstante, su uso no sería del todo apropiado. Las variables presentan elevadas correlaciones en casos como inm.1000 y em.1000, además de contenerse variables con gran variabilidad como el PIB. La distancia euclídea mide directamente la separación entre dos observaciones sin considerar las correlaciones entre ellas ni las diferencias de escala, lo que puede resultar problemático en este contexto. En cambio, la distancia de Mahalanobis es más adecuada para estos casos, ya que tiene en cuenta las correlaciones y ofrece un tratamiento más robusto frente a los outliers, que afectan de manera significativa a la métrica euclídea.

EJERCICIO 4: ¿Variaría sustancialmente la solución si las variables se expresasen en otras unidades? ¿Cómo variaría, si es el caso?

Si nos enfocamos exclusivamente en la distancia de Mahalanobis, podemos afirmar que las variaciones en las magnitudes de las variables no afectan de manera significativa al análisis, puesto que esta métrica incorpora en su formulación las correlaciones existentes entre las distintas variables. De este modo, permite una comparación más equilibrada entre observaciones, independientemente de las unidades de medida o de la escala de los datos, lo que resulta especialmente útil en contextos donde las variables presentan diferentes órdenes de magnitud o alta interdependencia.

Por el contrario, al emplear la distancia euclídea, los cambios en la magnitud de las variables adquieren una relevancia notable. Por ejemplo, si una de las variables (como podría ser un índice) se expresa en millones, su peso dominaría sobre el resto, distorsionando los resultados. Este tipo de desequilibrio puede tener consecuencias importantes en métodos estadísticos y de aprendizaje como el análisis clúster o el escalamiento multidimensional, ya que ambos dependen directamente de las distancias entre observaciones. En consecuencia, la elección de la métrica de distancia resulta un aspecto clave para garantizar la validez e interpretabilidad de los resultados obtenidos.

EJERCICIO 5: Realizar el procedimiento para encontrar la configuración de puntos pedida, sin utilizar la función propia de R. ¿Cómo se definen las matrices necesarias en cada paso?

El procedimiento que hemos desarrollado sigue un enfoque de análisis de escalamiento multidimensional clásico (MDS).

Se calcula la matriz de distancias euclídeas entre las observaciones escaladas. Esto produce una matriz D de tamaño n*n, donde n representa el número de observaciones. Cada elemento D_{ij} representa la distancia euclídea entre las observaciones i y j.

```
> D
                   Albacete Alicante/Alacant  Almería Araba/Álava Asturias
Albacete           0.000000         1.160694 2.263781    5.386146 2.716571
Alicante/Alacant   1.160694         0.000000 2.431649    4.757566 2.130888
Almería            2.263781         2.431649 0.000000    5.579795 4.338550
Araba/Álava        5.386146         4.757566 5.579795    0.000000 5.200729
Asturias           2.716571         2.130888 4.338550    5.200729 0.000000
Ávila              2.968059         3.773376 4.154654    6.623631 4.850359
Badajoz            1.406025         1.776401 2.568240    6.308361 2.624215
Balears, Illes     4.360542         3.581106 4.292057    2.421230 4.604191
Barcelona          5.088451         4.183050 5.193850    2.796876 4.369198
Bizkaia            4.106897         3.276356 4.822610    2.635941 3.008821
Burgos             2.798792         2.363892 3.674836    2.807083 3.274446
Cáceres            1.538067         1.318236 2.795855    5.372331 2.209579
Cádiz              3.748549         3.679257 3.455038    6.959178 4.038579
Cantabria          1.871607         1.391676 3.510317    4.332477 2.120377
Castellón/Castelló 1.605735         1.547757 2.349109    4.121993 3.229254
Ciudad Real        1.369973         2.248069 2.486465    6.361652 3.239400
Córdoba            1.927684         1.902619 2.409573    6.000681 2.658896
Coruña, A          2.361868         1.629135 3.945943    4.443705 1.231355
Cuenca             2.066844         2.822257 3.560046    5.965060 3.640602
```

Figure 15: Primeras filas y columnas de la matriz D.

Como en los apartados anteriores, y teniendo en cuenta las dimensiones considerablemente grandes de la matriz obtenida, se ha optado por presentar únicamente una parte representativa de la misma. Concretamente, se muestran solo las primeras filas y columnas, lo que permite ilustrar de manera clara la estructura y el formato de los datos sin sobrecargar el documento con información excesiva. Este mismo criterio se ha seguido para las capturas y representaciones posteriores incluidas en este ejercicio, con el fin de mantener la coherencia en la presentación de los resultados y facilitar su comprensión, garantizando al mismo tiempo la legibilidad y la pertinencia del contenido mostrado.

Tras esto, transformamos la matriz de distancias D en una matriz A utilizando la fórmula $-\frac{1}{2}D^2$. Esto se basa en el hecho de que ahora trabajamos con matrices de productos escalares en lugar de distancias.

```
> A
                    Albacete Alicante/Alacant     Almería Araba/Álava
Albacete           0.0000000       -0.6736048  -2.5623525  -14.505285
Alicante/Alacant  -0.6736048        0.0000000  -2.9564581  -11.317215
Almería           -2.5623525       -2.9564581   0.0000000  -15.567054
Araba/Álava      -14.5052846      -11.3172150 -15.5670535    0.000000
Asturias          -3.6898803       -2.2703422  -9.4115087  -13.523789
Ávila             -4.4046879       -7.1191835  -8.6305754  -21.936245
Badajoz           -0.9884532       -1.5778009  -3.2979276  -19.897712
Balears, Illes    -9.5071631       -6.4121590  -9.2108772   -2.931178
Barcelona        -12.9461666       -8.7489534 -13.4880382   -3.911258
Bizkaia           -8.4333027       -5.3672540 -11.6287856   -3.474094
Burgos            -3.9166187       -2.7939934  -6.7522109   -3.939859
Cáceres           -1.1828246       -0.8688724  -3.9084019  -14.430968
Cádiz             -7.0258085       -6.7684642  -5.9686429  -24.215077
Cantabria         -1.7514558       -0.9683809  -6.1611630   -9.385180
Castellón/Castelló -1.2891930      -1.1977759  -2.7591565   -8.495414
Ciudad Real       -0.9384134       -2.5269072  -3.0912550  -20.235308
Córdoba           -1.8579838       -1.8099789  -2.9030219  -18.004089
Coruña, A         -2.7892105       -1.3270406  -7.7852328   -9.873257
Cuenca            -2.1359218       -3.9825666  -6.3369645  -17.790971
```

Figure 16: Primeras filas y columnas de la matriz A.

A continuación, se crea la matriz H, que es una matriz cuadrada de tamaño n*n definida como $H = I - \frac{1}{n}11^T$, siendo I la matriz identidad de tamaño n, y siendo 1 el vector columna de unos de tamaño n.

```
> H
             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
 [1,]  0.98076923 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
 [2,] -0.01923077  0.98076923 -0.01923077 -0.01923077 -0.01923077 -0.01923077
 [3,] -0.01923077 -0.01923077  0.98076923 -0.01923077 -0.01923077 -0.01923077
 [4,] -0.01923077 -0.01923077 -0.01923077  0.98076923 -0.01923077 -0.01923077
 [5,] -0.01923077 -0.01923077 -0.01923077 -0.01923077  0.98076923 -0.01923077
 [6,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077  0.98076923
 [7,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
 [8,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
 [9,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[10,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[11,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[12,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[13,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[14,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[15,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[16,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[17,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[18,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
[19,] -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077 -0.01923077
```

Figure 17: Primeras filas y columnas de la matriz H.

El resultado obtenido se considera coherente, dado que en toda la matriz únicamente se identifican dos valores distintos: uno situado en la diagonal principal y otro presente en el resto de las posiciones. Esta configuración confirma que los datos se encuentran correctamente centrados.

Finalmente, la matriz B, conocida también como matriz de Gram, se obtiene mediante la aplicación de la matriz de centrado H a la matriz A, conforme la expresión B = HAH.

```
> B
             [,1]        [,2]         [,3]         [,4]         [,5]         [,6]
 [1,]  2.319516434  0.99469382  1.900465848 -5.550623051   0.54382305   3.44908505
 [2,]  0.994693819  1.01708089  0.855142483 -3.013771217   1.31214333   0.08337164
 [3,]  1.900465848  0.85514248  6.606120281 -4.469090071  -3.03450348   1.36649947
 [4,] -5.550623051 -3.01377122 -4.469090071 15.589806600  -2.65494048  -7.44732677
 [5,]  0.543823050  1.31214333 -3.034503480 -2.654940477   6.14789021  -1.99503126
 [6,]  3.449085046  0.08337164  1.366499467 -7.447326771  -1.99503126  13.38802938
 [7,]  3.303791335  2.06322592  3.137618882 -8.970322249   2.76317934   2.62023984
 [8,] -3.944096862 -1.50031056 -1.504509031  9.267033258  -3.12203430  -6.65666694
 [9,] -4.701270946 -1.15527551 -3.099840560 10.968783238   0.61413666 -13.03082413
[10,] -3.260144419 -0.84531348 -4.312325444  8.334209723   2.56084364  -8.54241169
[11,] -1.526944856 -1.05553735 -2.219235147  5.084960292  -1.05713682  -1.73142963
[12,]  1.365966422  1.02870082  0.783691083 -5.247031490   2.02185715   1.64581208
[13,]  2.171496233  1.77762277  5.371963751 -8.382627207   2.95643244  -1.56697835
[14,]  0.452394925  0.58425210 -1.814010382 -0.546184414   1.87003765   0.12685825
[15,]  0.573775445  0.01397483  1.247113885  0.002699737  -1.43688645   0.53670681
[16,]  3.757233392  1.51752185  3.747693678 -8.904516455   1.36297722   5.08962212
[17,]  2.345534803  1.74232191  3.443798582 -7.165425471   2.58284109  -0.38189674
[18,]  0.053295115  0.86424726 -2.799425245 -0.395606193   3.99857446  -2.19404064
[19,]  2.635929652  0.13806702  0.578188835 -6.383974590   0.05904861   9.13153361
```

Figure 18: Primeras filas y columnas de la matriz de Gram.

EJERCICIO 6: ¿Qué dimensión tiene sentido considerar para la representación de los datos? Razonar la respuesta.

Tras la obtención de los autovalores y el cálculo de la varianza explicada (dividiendo cada autovalor por la suma total de los autovalores, obteniendo la proporción de la varianza explicada por cada dimensión y estandarizando así los autovalores para que sumen 1), calculamos la varianza acumulada.

```
> lambda
 [1]  1.534389e+02  1.022723e+02  6.017464e+01  2.790801e+01  9.866932e+00
 [6]  3.302406e+00  3.683196e-02  1.935066e-14  1.153574e-14  1.130905e-14
[11]  1.128121e-14  6.393169e-15  6.050391e-15  4.874027e-15  4.795807e-15
[16]  4.307736e-15  3.506665e-15  3.262556e-15  3.090679e-15  2.738214e-15
[21]  2.314455e-15  1.987000e-15  1.917315e-15  1.753648e-15  1.326380e-15
[26]  1.074171e-15  7.053156e-16  5.275607e-16 -5.741470e-17 -1.296412e-16
[31] -3.162966e-16 -4.204168e-16 -7.272210e-16 -1.045842e-15 -1.116279e-15
[36] -1.274491e-15 -1.356267e-15 -1.386180e-15 -1.528852e-15 -2.382676e-15
[41] -2.422924e-15 -2.619286e-15 -2.691385e-15 -2.942264e-15 -4.503454e-15
[46] -5.432623e-15 -5.661511e-15 -5.884452e-15 -6.477130e-15 -8.302571e-15
[51] -9.912579e-15 -1.761415e-14
```

Figure 19: Lambda.

```
> var_exp
 [1]  4.298008e-01  2.864771e-01  1.685564e-01  7.817371e-02  2.763847e-02
 [6]  9.250436e-03  1.031708e-04  5.420352e-17  3.231299e-17  3.167800e-17
[11]  3.160004e-17  1.790804e-17  1.694787e-17  1.365274e-17  1.343363e-17
[16]  1.206649e-17  9.822592e-18  9.138811e-18  8.657365e-18  7.670068e-18
[21]  6.483068e-18  5.565825e-18  5.370630e-18  4.912179e-18  3.715351e-18
[26]  3.008883e-18  1.975674e-18  1.477761e-18 -1.608255e-19 -3.631407e-19
[31] -8.859850e-19 -1.177638e-18 -2.037034e-18 -2.929529e-18 -3.126833e-18
[36] -3.570004e-18 -3.799068e-18 -3.882858e-18 -4.282499e-18 -6.674161e-18
[41] -6.786903e-18 -7.336935e-18 -7.538893e-18 -8.241635e-18 -1.261472e-17
[46] -1.521743e-17 -1.585857e-17 -1.648306e-17 -1.814322e-17 -2.325650e-17
[51] -2.776633e-17 -4.933935e-17
```

Figure 20: Varianza acumulada.

Mediante el cálculo acumulado de las proporciones, representado por la variable var_exp_acum, se obtuvo la fracción de varianza total explicada a medida que se van incorporando progresivamente las distintas dimensiones al análisis. Este procedimiento permite evaluar de forma detallada cómo contribuye cada componente adicional a la explicación de la variabilidad total de los datos, proporcionando una visión más completa sobre la importancia relativa de cada dimensión y facilitando la determinación del número óptimo de componentes a conservar para una adecuada representación de la información original.

```
> var_exp_acum
 [1] 0.4298008 0.7162778 0.8848342 0.9630079 0.9906464 0.9998968 1.0000000 1.0000000
 [9] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[17] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[25] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[33] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[41] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[49] 1.0000000 1.0000000 1.0000000 1.0000000
```

Figure 21: Fracción de varianza total explicada incorporando dimensiones.

Tal y como puede apreciarse en la captura, al considerar únicamente una dimensión se logra explicar aproximadamente el 42 % de la varianza total, mientras que al incorporar una segunda dimensión este valor se incrementa hasta el 72 %, continuando dicho aumento progresivamente hasta alcanzar el valor máximo de 1. Tras analizar la evolución de la varianza explicada acumulada, se ha determinado que tres dimensiones resultan suficientes para llevar a cabo una representación adecuada de los datos, ya que con ellas se alcanza una explicación del 88,48 % de la varianza total, valor que se considera satisfactorio para los fines del análisis realizado.

EJERCICIO 7: ¿Cuáles serían las coordenadas de los puntos obtenidos?

Para la obtención de las coordenadas de los puntos obtenidos calculamos los autovectores de la matriz B, extrayendo así los 3 más relevantes.

```
> autovec
             [,1]         [,2]         [,3]
 [1,] -0.106378063 -0.006769027 -0.026683923
 [2,] -0.048465976 -0.053279148  0.005229943
 [3,] -0.100962708  0.041264754  0.234931703
 [4,]  0.295302903  0.032982047  0.013377909
 [5,] -0.046050503 -0.196123433 -0.133457241
 [6,] -0.178322483  0.222198946 -0.222971673
 [7,] -0.181066264 -0.098945173  0.021935787
```

Figure 22: Autovectores de la matriz B.

```
> vec12
 [1] -0.106378063 -0.048465976 -0.100962708
 [7] -0.181066264  0.205215496  0.245161213
[13] -0.216067167  0.015556061  0.012833763
[19] -0.161411697  0.228469767  0.167997710
```

Figure 23: Autovectores de la matriz B.

Tras esto escalamos el autovector seleccionado multiplicándolo por la raíz cuadrada del autovalor correspondiente. Este escalado asegura que las coordenadas proyectadas preserven las distancias originales entre los puntos tanto como sea posible.

```
> result
 [1] -1.3177098 -0.6003502 -1.2506295  3.6579302 -0.5704296 -2.2088885 -2.2428758
 [8]  2.5420135  3.0368228  2.1659836  1.4158819 -1.3753073 -2.6764336  0.1926936
[15]  0.1589724 -2.4197066 -1.8824316  0.1925028 -1.9994139  2.8300652  2.0809951
[22] -1.6691935  0.8951389 -1.9394745  1.3437513 -2.3675056 -1.8001269  2.9647089
[29] -0.2411370  3.3564283 -0.4751507 -0.0603866  2.6423682 -1.6573782 -0.5301548
[36] -0.6047769 -0.9370143  1.6997233 -1.2726201 -0.5801228  1.4935615 -0.5330224
[43]  0.6932937  1.2780993  0.4603000 -1.2314392  0.8717854  0.8265708 -1.4217536
[50]  1.3828951 -1.0334280 -1.2836240
```

Figure 24: Resultado del escalado multidimensional.

De este modo, se obtienen las coordenadas correspondientes a los distintos puntos en la primera dimensión del espacio transformado, resultado del proceso de reducción dimensional previamente descrito. Estas coordenadas reflejan la posición relativa de cada observación dentro de dicho eje principal, permitiendo interpretar cómo se distribuyen los datos en función de la componente que mayor proporción de varianza explica. En otras palabras, esta primera dimensión concentra la mayor parte de la información relevante del conjunto de datos, ofreciendo una representación simplificada pero significativa de la estructura subyacente del mismo.

EJERCICIO 8: Representar gráficamente la solución propuesta, incluyendo etiquetas que identifiquen cada observación.

Esta representación gráfica permite visualizar las distancias existentes entre las distintas observaciones en un espacio tridimensional, generada mediante la función scatterplot3d a partir de las coordenadas calculadas previamente.

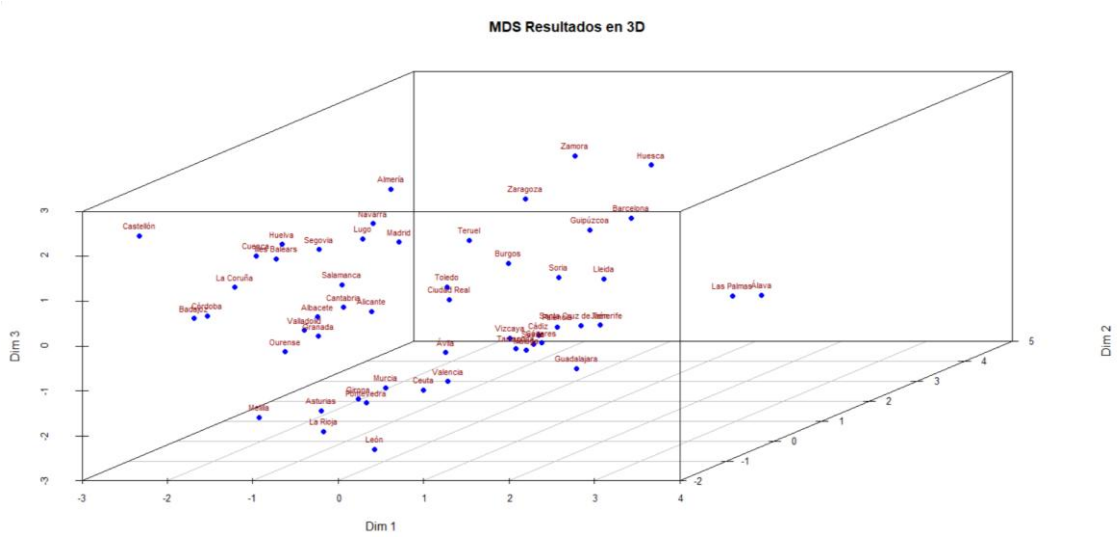

Figure 25: Representación del escalado multidimensional con etiquetas.

Asimismo, las etiquetas asociadas a cada observación se han representado en un plano bidimensional, manteniendo la correspondencia con los puntos mostrados en el gráfico. Si bien la representación es adecuada y refleja correctamente la estructura espacial de los datos, se observa que algunos elementos no pueden distinguirse con total claridad, lo que se debe principalmente a la superposición de puntos en determinadas zonas del espacio tridimensional.

EJERCICIO 9: ¿Es necesario aplicar alguna rotación o simetría? ¿Por qué? Si se ha realizado alguna rotación o simetría, representar gráficamente la nueva solución con sus etiquetas correspondientes.

Como se ha mencionado con anterioridad, la representación inicial de los datos no resultaba la más adecuada para una interpretación clara y precisa de las observaciones. Con el objetivo de mejorar la visualización tridimensional, se procedió a realizar una rotación de los ejes mediante el uso de la función cbind.

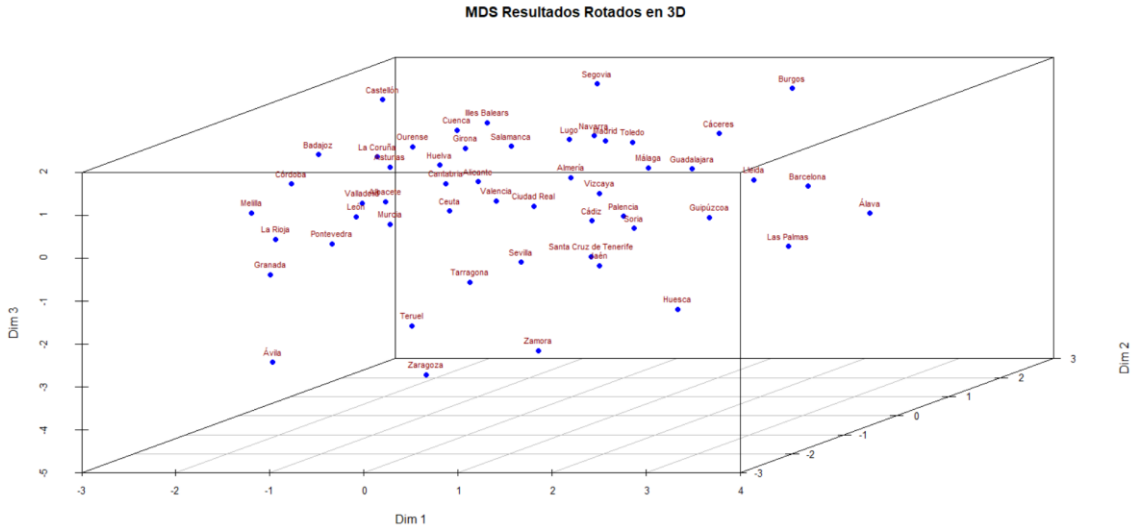

Figure 26: Representación del escalado multidimensional con rotación.

Tras experimentar con diversas combinaciones posibles, se identificó una configuración que ofrecía una mejor perspectiva general, permitiendo distinguir con mayor nitidez las distintas observaciones y evitando la superposición entre ellas. En esta nueva disposición, el eje X se mantiene sin modificaciones, mientras que el eje Y pasa a ocupar la posición que anteriormente correspondía al eje Z, y este último (eje Z) adopta ahora el lugar del antiguo eje Y, aunque con el signo invertido, lo que mejora notablemente la legibilidad del gráfico.

EJERCICIO 10: Aplicar el método utilizando directamente la función propia de R, y comprobar que se corresponden los resultados con lo obtenido siguiendo los pasos anteriores.

Después de llevar a cabo un escalamiento multidimensional utilizando la función específica que ofrece RStudio, se puede comprobar que el procedimiento desarrollado previamente ha sido plenamente correcto.

```
> cmdscale(D, k=3)
                          [,1]       [,2]       [,3]
Albacete            -1.3177098 -0.06845502 -0.20699336
Alicante/Alacant    -0.6003502 -0.53881082  0.04056988
Almería             -1.2506295  0.41730953  1.82241952
Araba/Álava          3.6579302  0.33354670  0.10377554
Asturias            -0.5704296 -1.98339185 -1.03525867
Ávila               -2.2088885  2.24709293 -1.72964281
Badajoz             -2.2428758 -1.00063031  0.17016097
Balears, Illes       2.5420135  0.57772414  1.10524847
Barcelona            3.0368228 -1.49368419  1.35932817
Bizkaia              2.1659836 -1.61331527 -0.10284265
Burgos               1.4158819  0.19054916 -0.44680907
Cáceres             -1.3753073 -0.28081055  0.09636320
Cádiz               -2.6764336 -1.37971011  2.18304641
Cantabria            0.1926936 -0.64141436 -1.02952158
Castellón/Castelló   0.1589724  0.13017661  0.05929346
Ciudad Real         -2.4197066 -0.40959929 -0.03292583
Córdoba             -1.8824316 -1.24542488  0.97122474
Coruña, A            0.1925028 -1.46430395 -0.97737213
Cuenca              -1.9994139  1.18755958 -1.19605882
Gipuzkoa             2.8300652 -1.48102593 -0.55116379
Girona               2.0809951  0.77973046  0.97449769
Granada             -1.6691935 -0.18218428  1.45876682
Guadalajara          0.8951389  4.27366136  1.05783341
Huelva              -1.9394745 -0.91861001  1.19214912
Huesca               1.3437513  1.09992120 -0.91279156
Jaén                -2.3675056 -0.90909543  0.45877654
León                -1.8001269 -0.23242519 -2.04022208
Lleida               2.9647089  0.96813410 -0.20353640
Lugo                -0.2411370 -1.14117828 -1.75151720
Madrid               3.3564283 -0.15085849  0.52812665
Málaga              -0.4751507 -0.63082351  1.78473386
Murcia              -0.0603866 -0.82791853  1.07146020
Navarra              2.6423682 -0.82462299 -0.40901975
Ourense             -1.6573782 -0.67943337 -2.17000967
Palencia            -0.5301548 -0.29034102 -1.07758145
Palmas, Las         -0.6047769 -0.47608106  1.60922590
Pontevedra          -0.9370143 -1.43023736 -0.38119917
Rioja, La            1.6997233  0.16337876 -0.32647917
```

Figure 27: Resultados del escalado multidimensional con la función cdmscale de RStudio.

Las coordenadas obtenidas a través de esta función coinciden exactamente con las calculadas en nuestro análisis anterior, lo que confirma la validez y consistencia de los pasos realizados. Este resultado respalda la precisión del método aplicado y demuestra que las transformaciones efectuadas han sido adecuadas para representar fielmente las relaciones entre las observaciones dentro del espacio multidimensional.

EJERCICIO 11: Realizar el mismo procedimiento para encontrar la configuración de puntos sobre un espacio, pero ahora para representar las variables (no las observaciones).

Para llevar a cabo un escalamiento multidimensional sobre las variables, es necesario trabajar directamente sobre la matriz de correlaciones. Al aplicar este procedimiento, se ha podido observar la existencia de una relación directa entre la matriz de distancias euclídeas calculada a partir de las variables estandarizadas y la propia matriz de correlaciones de las mismas.

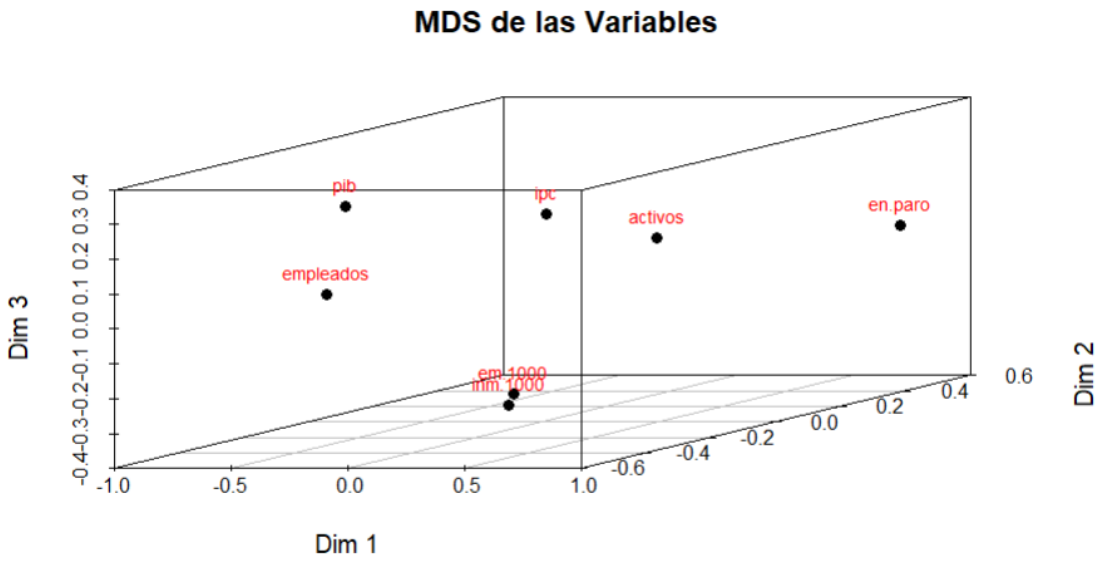

Figure 28: Resultados del escalado multidimensional con la función cdmscale de R.

Teniendo esto en cuenta, se ha optado nuevamente por seleccionar tres dimensiones con el objetivo de facilitar una visualización más clara y comprensible de los datos. Tal como se aprecia en la representación resultante, la tasa de emigración y la tasa de inmigración presentan una alta correlación entre sí, mientras que el resto de las variables muestran una relación más distante, lo que permite identificar de manera más intuitiva la estructura de interdependencias dentro del conjunto de datos.

EJERCICIO 12: A la vista de los resultados obtenidos a lo largo del ejercicio, ¿se puede afirmar que la aplicación del método a este conjunto de datos ha resultado útil? Justificar la respuesta.

En definitiva, se puede concluir de manera firme que la exploración y el análisis de los datos han resultado plenamente satisfactorios. La aplicación de la metodología seguida ha permitido obtener una visualización clara y comprensible de la información, al mismo tiempo que ha facilitado la identificación de correlaciones significativas entre las distintas variables y la observación de cómo los cambios en las magnitudes de estas afectan a los resultados. Este enfoque ha proporcionado, por tanto, una comprensión más profunda de la estructura de los datos y de las relaciones existentes entre sus componentes, confirmando la eficacia del proceso realizado.

VISUAL REFERENCES.

[1] https://www.ine.es

[2] https://www.abc.es/economia/abci-mapa-paro-espana-peores-y-mejores-provincias-para-encontrar-trabajo-201710261116_noticia.html?utm_source=chatgpt.com