Minería de Datos y Modelización Predictiva (II)

Fernández Hernández, Alberto. 54003003S

1/2/2021

## PARTE 1. ANÁLISIS DE COMPONENTES PRINCIPALES

### Apartado 1

**Calcular la matriz de correlaciones, y su representación gráfica ¿Cuáles son las variables más correlacionadas de forma inversa?**

De forma previa a la matriz de correlación, debemos cargar el conjunto de datos, además de renombrar los nombres de fila empleando los nombres de provincia de la primera variable:

provincias <- as.data.frame(read\_excel("Provincias.xlsx"))  
rownames(provincias) <- provincias$Provincia; provincias <- provincias[, -1] # Eliminamos finalmente la columna "Provincia"

Una vez cargado el conjunto de datos, calculamos la matriz de correlación:

matriz.correlaciones <-cor(provincias)  
knitr::kable(matriz.correlaciones, digits =2,caption = "Correlaciones")

Correlaciones

Poblacion

Mortalidad

Natalidad

IPC

NumEmpresas

Industria

Construccion

CTH

Infor

AFS

APT

TasaActividad

TasaParo

Ocupados

PIB

CANE

TVF

VS

Poblacion

1.00

-0.34

0.11

0.33

0.99

0.96

0.98

1.00

0.94

0.99

0.98

0.33

0.01

1.00

0.98

0.10

0.99

0.57

Mortalidad

-0.34

1.00

-0.74

0.19

-0.31

-0.28

-0.30

-0.33

-0.26

-0.31

-0.30

-0.73

-0.46

-0.33

-0.30

0.02

-0.33

-0.25

Natalidad

0.11

-0.74

1.00

-0.25

0.11

0.09

0.09

0.10

0.11

0.10

0.11

0.47

0.38

0.11

0.11

-0.12

0.08

-0.03

IPC

0.33

0.19

-0.25

1.00

0.36

0.42

0.40

0.36

0.30

0.32

0.33

0.09

-0.58

0.36

0.36

-0.19

0.34

0.19

NumEmpresas

0.99

-0.31

0.11

0.36

1.00

0.97

0.99

0.99

0.96

0.99

0.99

0.33

-0.06

1.00

0.99

0.04

0.98

0.54

Industria

0.96

-0.28

0.09

0.42

0.97

1.00

0.97

0.98

0.89

0.95

0.93

0.29

-0.08

0.96

0.94

0.12

0.97

0.57

Construccion

0.98

-0.30

0.09

0.40

0.99

0.97

1.00

0.99

0.96

0.98

0.98

0.34

-0.11

0.99

0.99

0.03

0.98

0.56

CTH

1.00

-0.33

0.10

0.36

0.99

0.98

0.99

1.00

0.93

0.98

0.97

0.33

-0.01

0.99

0.97

0.09

0.99

0.58

Infor

0.94

-0.26

0.11

0.30

0.96

0.89

0.96

0.93

1.00

0.97

0.99

0.31

-0.11

0.96

0.99

-0.07

0.91

0.41

AFS

0.99

-0.31

0.10

0.32

0.99

0.95

0.98

0.98

0.97

1.00

0.99

0.32

-0.03

0.99

0.99

0.09

0.98

0.54

APT

0.98

-0.30

0.11

0.33

0.99

0.93

0.98

0.97

0.99

0.99

1.00

0.33

-0.08

0.99

1.00

-0.01

0.96

0.48

TasaActividad

0.33

-0.73

0.47

0.09

0.33

0.29

0.34

0.33

0.31

0.32

0.33

1.00

0.03

0.35

0.33

-0.12

0.33

0.26

TasaParo

0.01

-0.46

0.38

-0.58

-0.06

-0.08

-0.11

-0.01

-0.11

-0.03

-0.08

0.03

1.00

-0.05

-0.10

0.39

0.01

0.10

Ocupados

1.00

-0.33

0.11

0.36

1.00

0.96

0.99

0.99

0.96

0.99

0.99

0.35

-0.05

1.00

0.99

0.05

0.98

0.54

PIB

0.98

-0.30

0.11

0.36

0.99

0.94

0.99

0.97

0.99

0.99

1.00

0.33

-0.10

0.99

1.00

-0.01

0.96

0.47

CANE

0.10

0.02

-0.12

-0.19

0.04

0.12

0.03

0.09

-0.07

0.09

-0.01

-0.12

0.39

0.05

-0.01

1.00

0.15

0.34

TVF

0.99

-0.33

0.08

0.34

0.98

0.97

0.98

0.99

0.91

0.98

0.96

0.33

0.01

0.98

0.96

0.15

1.00

0.67

VS

0.57

-0.25

-0.03

0.19

0.54

0.57

0.56

0.58

0.41

0.54

0.48

0.26

0.10

0.54

0.47

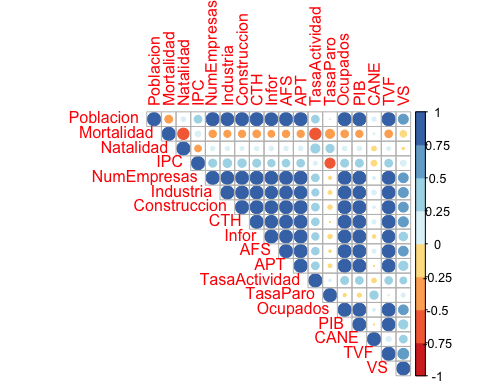
0.34

0.67

1.00

Así como su representación gráfica:

corrplot(matriz.correlaciones, type="upper", col=brewer.pal(n=8, name="RdYlBu"))



Analizando la gráfica de correlación, de forma inversa nos encontramos múltiples variables aunque la mayoría de ellas con valores muy cercanos a 0. No obstante, si debemos destacar los siguientes casos:

1. **MORTALIDAD - NATALIDAD**: presenta una elevada correlación negativa (-0.74). Es decir, aquellas provincias donde aumenta/disminuye la tasa de natalidad, la tasa de mortalidad se reduce/aumenta, respectivamente.
2. **MORTALIDAD - TASA\_ACTIVIDAD**: con una correlación muy similar a la anterior (-0.73), esto es, en provincias donde existe una alta tasa de mortalidad, el porcentaje de personas en “edad laboral” disminuye.
3. **MORTALIDAD - TASA\_PARO**: con una correlación bastante menor a las dos anteriores, aunque también es significativa (-0.46): a medida que aumenta la tasa de mortalidad, el número de parados se reduce (y viceversa).

En relación con la mortalidad, **podemos decir que a medida que aumenta la mortalidad en la provincia, disminuyen las tasas de natalidad, de actividad y de paro (aunque esta última en menor medida)**.

1. **IPC - TASA\_PARO**: una correlación inferior a -0.5, por lo que también es significativa (-0.58): **aquellas provincias donde el índice de precios al consumidor es mayor, la tasa de paro se reduce**. Dicho de otro modo, en aquellas provincias donde el coste de vida es mayor, la tasa de desempleo es menor.

Aunque no se trata de una correlación negativa, quisiera destacar la correlación existente entre *NumEmpresas*, *Industria*, *Construccion*, *CTH*, *Infor*, *AFS* y *APT*, con correlaciones muy cercanas a 1, por lo que puede que dichas variables acaben concentradas en torno a la misma componente principal.

### Apartado 2

**Realizar un análisis de componentes principales sobre la matriz de correlaciones, calculando 7 componentes. Estudiar los valores de los autovalores obtenidos y las gráficas que los resumen. ¿Cuál es el número adecuado de componentes?**

Con los datos de provincias, realizamos el análisis con un total de siete componentes, tal y como se muestra a continuación:

analisis.comp <- PCA(provincias, scale.unit = TRUE, ncp = 7, graph = TRUE)

