# R - PCA

## Analisis de componentes principales

Ariel Vallarino

```
inegi <- read.csv("inegi.csv")  # Importo datos
```

## Limpieza de datos:

Es importante limpiar bien los datos y asegurarse de que esten en unidades equivalentes.

```
# Recorro columnas de interes y las divido por la Poblacion
for (col in c("Divorcios", "DefuncionesGenerales", "Nacimientos", "Divorcios", "Matri
monios")) {
  inegi[ ,col] <- inegi[ ,col] / inegi[ ,"Poblacion"]
}
head(inegi)
```

```
##                 Estado Poblacion  PIBpc Secundarias
## 1       Aguascalientes   1184996  84.70         348
## 2      Baja California   3155070  83.07         595
## 3  Baja California Sur    637026  94.64         156
## 4             Campeche    822441 395.55         318
## 5 Coahuila de Zaragoza   2748391 106.05         553
## 6               Colima    650555  76.49         169
##   IndiceAprovechamientoSecundaria PorcentajeAnalfabetas
## 1                            85.3                  3.26
## 2                            86.0                  2.57
## 3                            87.9                  3.21
## 4                            78.5                  8.31
## 5                            75.9                  2.63
## 6                            81.8                  5.13
##   DefuncionesGenerales Nacimientos    Divorcios Matrimonios
## 1          0.004444741  0.02272835 1.009104e-09 0.005240524
## 2          0.004676917  0.01992697 3.309067e-10 0.005557721
## 3          0.004299668  0.02019384 1.833406e-09 0.004483333
## 4          0.004819799  0.02230312 1.682414e-09 0.006404107
## 5          0.005530509  0.02181858 4.855929e-10 0.005597457
## 6          0.005713583  0.02063161 1.554740e-09 0.005207861
##   PorcentajePartosHospitales PorcentajeAguaPotable PorcentajeAguaEntubada
## 1                       97.1                  98.0                   98.9
## 2                       65.7                  93.3                   95.9
## 3                       95.2                  86.7                   92.4
## 4                       87.0                  89.5                   90.3
## 5                       90.3                  97.9                   98.2
## 6                       98.5                  97.9                   98.5
##   PorcentajeElectricidad PorcentajeParedesSolidas PorcentajePisoTierra
## 1                   99.2                    92.3                  1.7
## 2                   98.5                    77.0                  3.3
## 3                   96.7                    90.3                  5.8
## 4                   96.8                    80.7                  4.7
## 5                   99.1                    84.8                  1.6
## 6                   99.0                    94.7                  4.5
```

### Eliminar datos innecesarios.

como por ej: Estado y Poblacion (las 2 primeras columnas)

```
x <- inegi[ ,-(1:2)]
row.names(x) <- inegi$Estado
```

Importo libería:

```
# install.packages("FactoMineR")
library("FactoMineR")
```

```
## Warning: package 'FactoMineR' was built under R version 3.3.2
```

# PCA

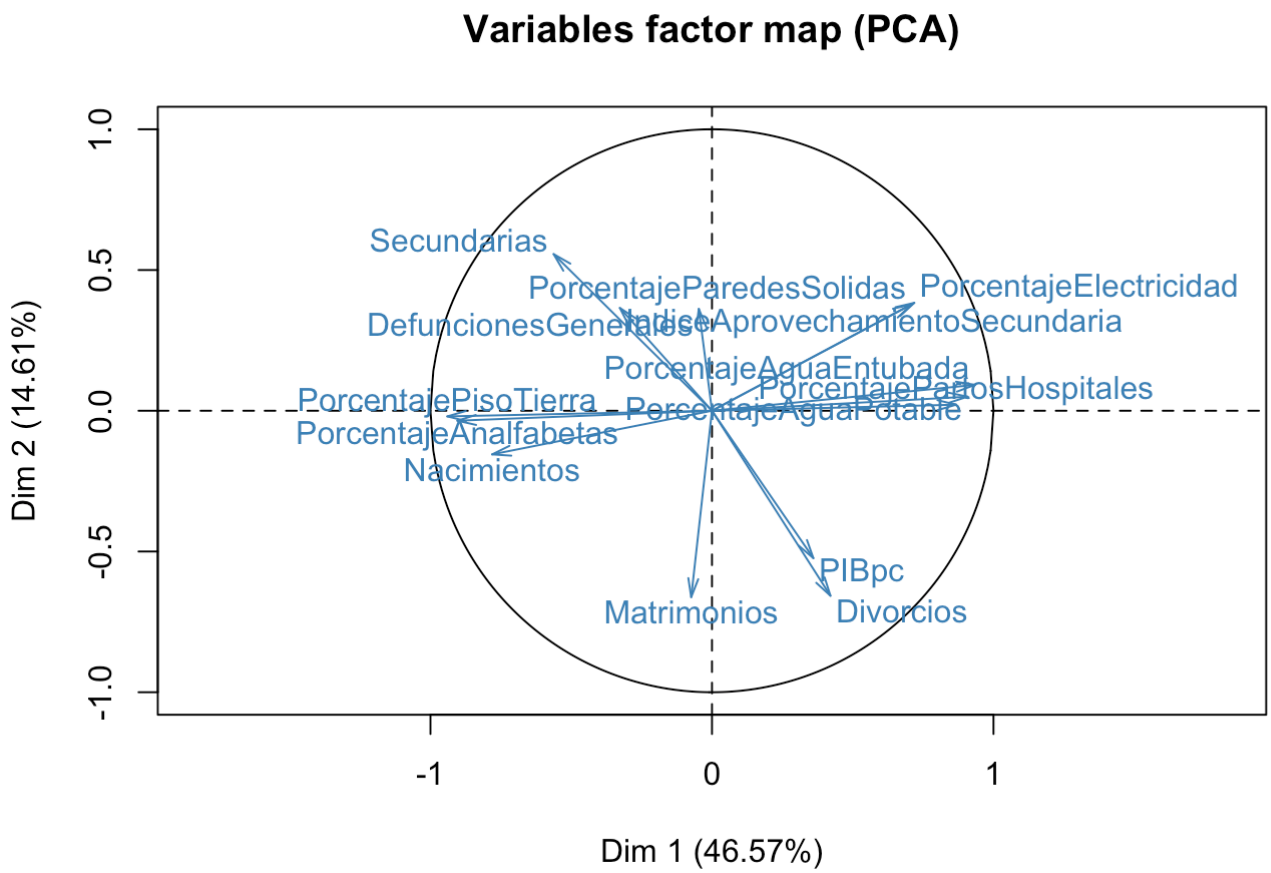
Maximizar la varianza de una cobinacion lineal sujeto a una norma
Por defecto PCA trae seteado el parametro SCALE.UNIT = TRUE para normalizar las variables

```
# Calculo PCA y muestro resultado
model <- PCA(x, graph = FALSE)
print(model)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 32 individuals, described by 14 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

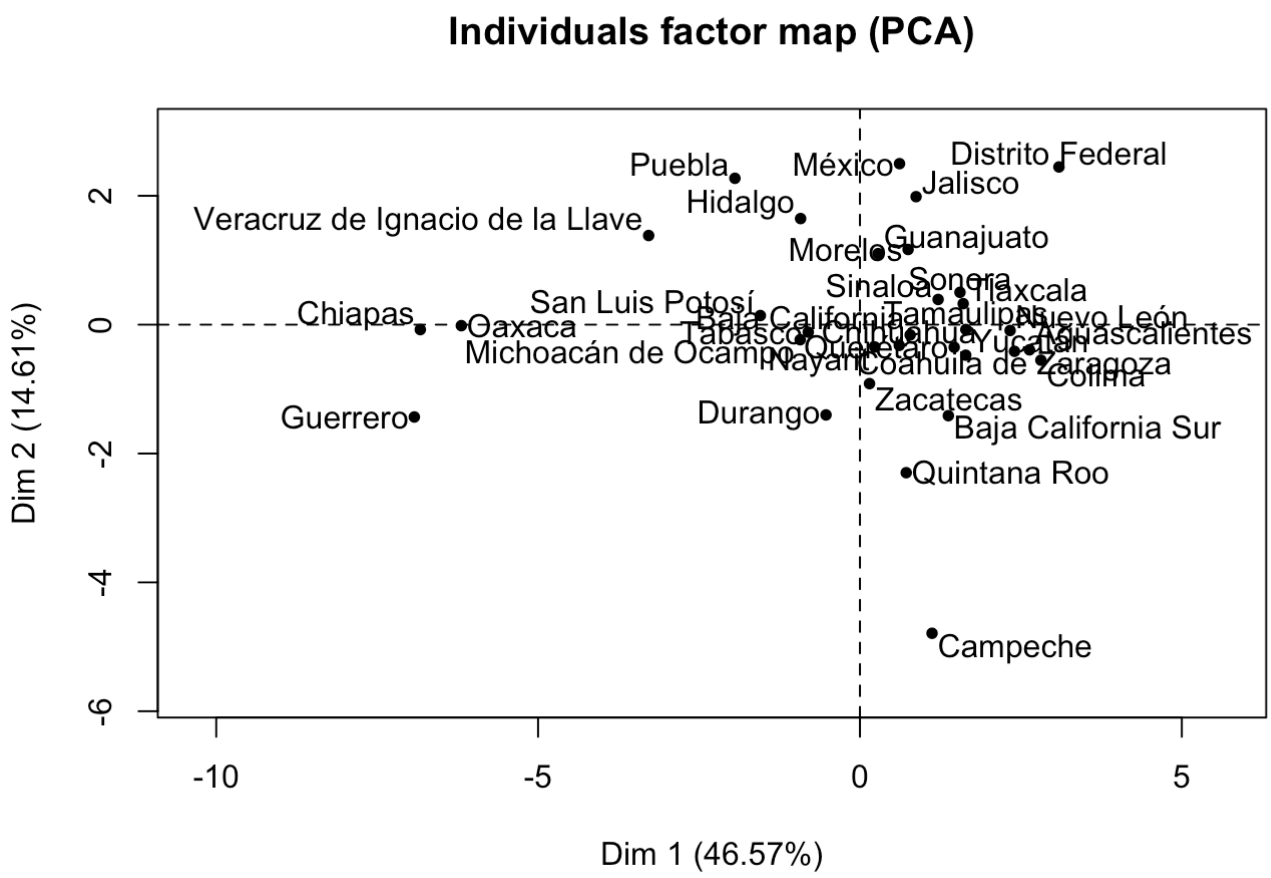
**Grafico de PCA para las Variables:**

```
plot(model, choix = "var",col.var="steelblue")
```



### Grafico de PCA para los resultados individuales

```
plot(model, choix = "ind")
```



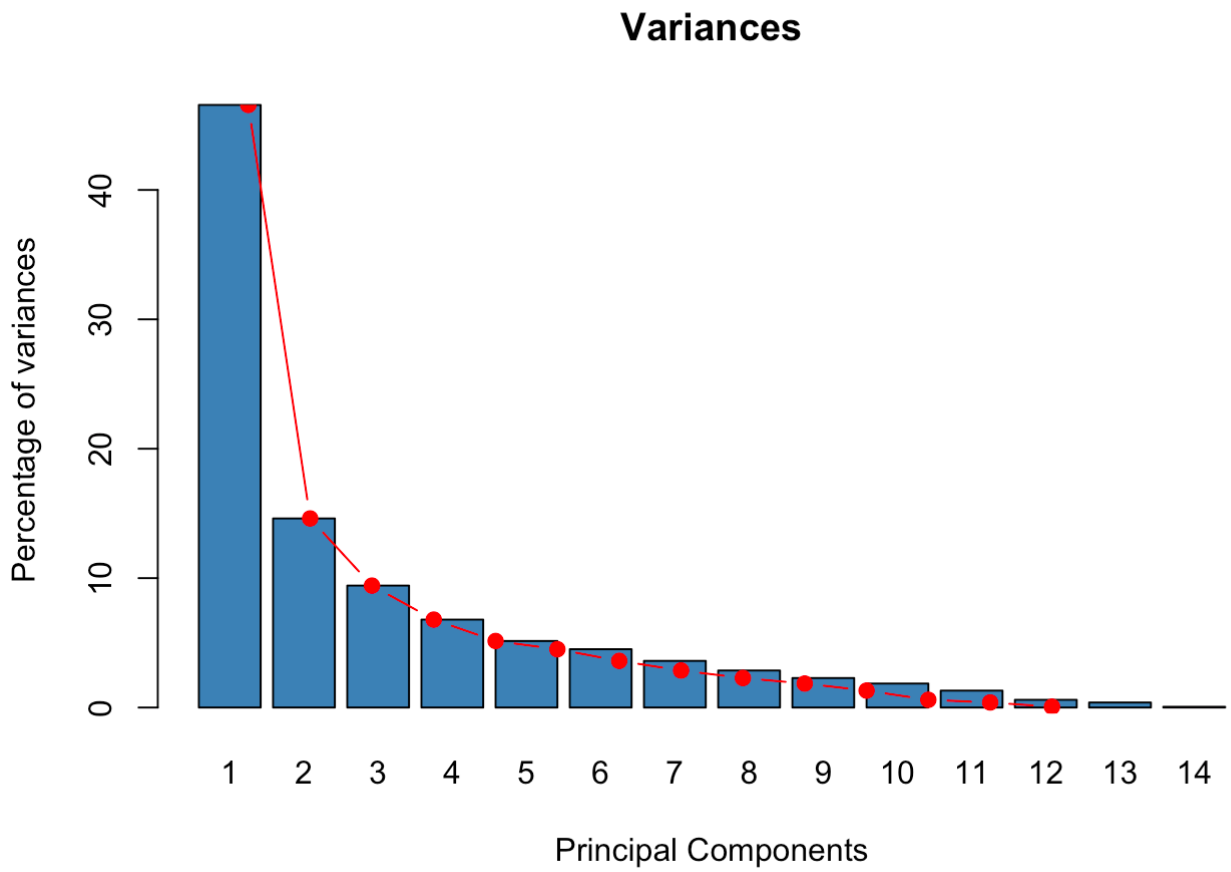

### Eigenvalores:

```
eigenvalores <- model$eig
head(eigenvalores[, 1:2])
```

```
##        eigenvalue percentage of variance
## comp 1  6.5195848              46.568463
## comp 2  2.0458108              14.612934
## comp 3  1.3193327               9.423805
## comp 4  0.9513398               6.795284
## comp 5  0.7198109               5.141506
## comp 6  0.6308262               4.505902
```

### Grafico valores

```
barplot(eigenvalores[, 2], names.arg=1:nrow(eigenvalores),
       main = "Variances",
       xlab = "Principal Components",
       ylab = "Percentage of variances",
       col ="steelblue")
# Dibulo linea
lines(x = 1:nrow(eigenvalores), eigenvalores[, 2],
      type="b", pch=19, col = "red")
```




**Aproximadamente el 60% de la informacion esta contenida en las primeras 2 componentes principales.**