# hw3

*Pablo, Román, Sofia*

*25/2/2020*

```
##
## Ejericios Pablo:

## [1] 5 2 9

##
## Ejericios Roman:

## [1] 2 8 3

##
## Ejericios Sofia:

## [1] 4 6 1
```

## Ejercicio 2

Obtenemos la matriz muestral de varianzas y covarianzas, luego obtenemos los componentes principales muestral.

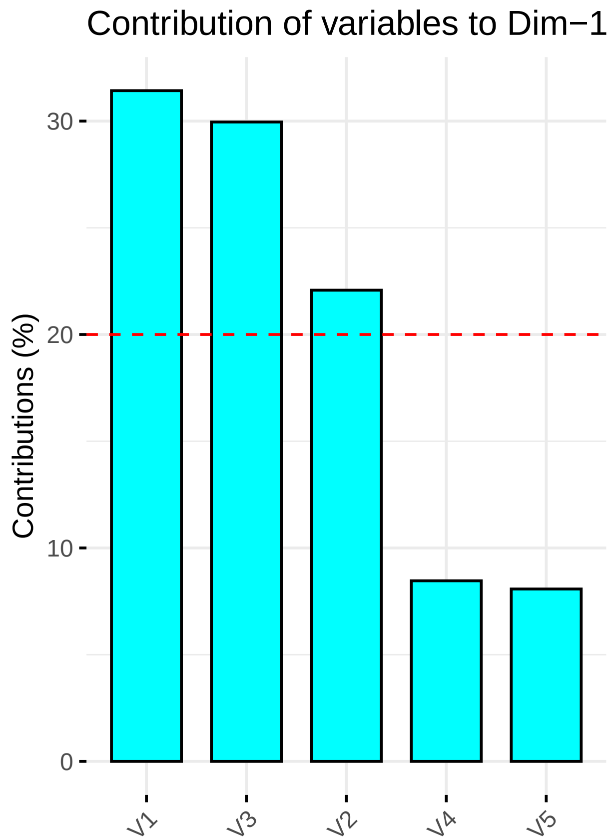
```
## La matriz de varianzas y covarianzas es:
```

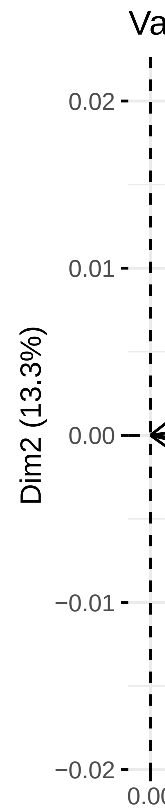
| | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|
| V1 | 0.0016299 | 0.0008167 | 0.0008101 | 0.0004422 | 0.0005140 |
| V2 | 0.0008167 | 0.0012294 | 0.0008276 | 0.0003869 | 0.0003109 |
| V3 | 0.0008101 | 0.0008276 | 0.0015561 | 0.0004873 | 0.0004625 |
| V4 | 0.0004422 | 0.0003869 | 0.0004873 | 0.0008023 | 0.0004085 |
| V5 | 0.0005140 | 0.0003109 | 0.0004625 | 0.0004085 | 0.0007587 |

```
##
##
## Los componentes principales son:

##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## V1  0.561  0.739  0.126  0.284  0.208
## V2  0.470         0.468 -0.688 -0.281
## V3  0.547 -0.654  0.114  0.500
## V4  0.291 -0.113 -0.610 -0.438  0.582
## V5  0.284        -0.617        -0.728
##
```

```
##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings        1.0    1.0    1.0    1.0    1.0
## Proportion Var     0.2    0.2    0.2    0.2    0.2
## Cumulative Var     0.2    0.4    0.6    0.8    1.0
```




Veamos a detalle los componentes principales de los datos

Vemos que por un “análisis de codo”, las primeras **tres componentes principales** son las que contiene la mayor variabilidad. Veamos las comunalidades de las CP para encontrar una interpretación.

```
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## V1  0.561  0.739  0.126  0.284  0.208
## V2  0.470         0.468 -0.688 -0.281
## V3  0.547 -0.654  0.114  0.500
## V4  0.291 -0.113 -0.610 -0.438  0.582
## V5  0.284        -0.617        -0.728
##
##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings        1.0    1.0    1.0    1.0    1.0
## Proportion Var     0.2    0.2    0.2    0.2    0.2
## Cumulative Var     0.2    0.4    0.6    0.8    1.0
```

Podemos ver que la primer componente es una ponderación de todas las acciones de la NYSE, mientras que la segunda componente distingue 3 grupos: La pprimera acción, las acciones 2 y 5 pero por último las acciones 3 y 4.

Sabemos por el *teorema de Anderson* que $\hat{\lambda} \sim \mathcal{N}(\lambda, 2\lambda^2/n), \text{si} \quad n \to \infty$. Por lo que podemos construir un intervalo de confianza bonferronizado (ya que por teorema $\lambda_i \perp \lambda_j \quad \forall i \neq j$.)

$$\lambda_i \in \frac{\hat{\lambda_i}}{1 \pm z_{\alpha/2m}\sqrt{2/n}}.$$

Por lo que tendremos un conjunto de intervalos de confianza al nivel del 90% para $\lambda_{1,2,3}$ dados por

1. 0.0458595, 0.0853459
2. 0.0215263, 0.040061
3. 0.0207552, 0.038626

Es importante ver que podemos **reducir la dimensión** de los datos en *3 componentes principales*, ya que en estas tres primeras componentes obtenemos el 85% de la variablidad de los datos.

### FALTA REPLICAR

# Ejercicio 3

## 3a. Comparar estimaciones

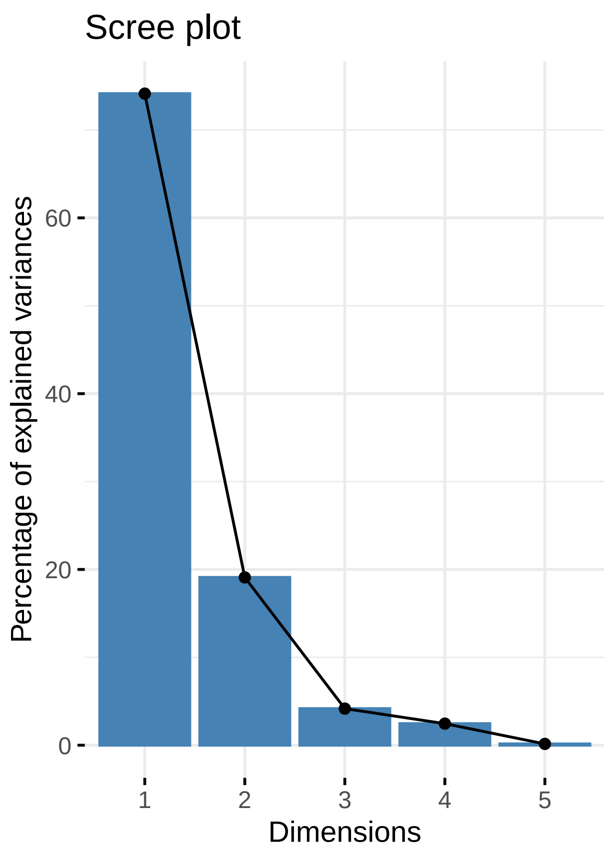
Veamos:

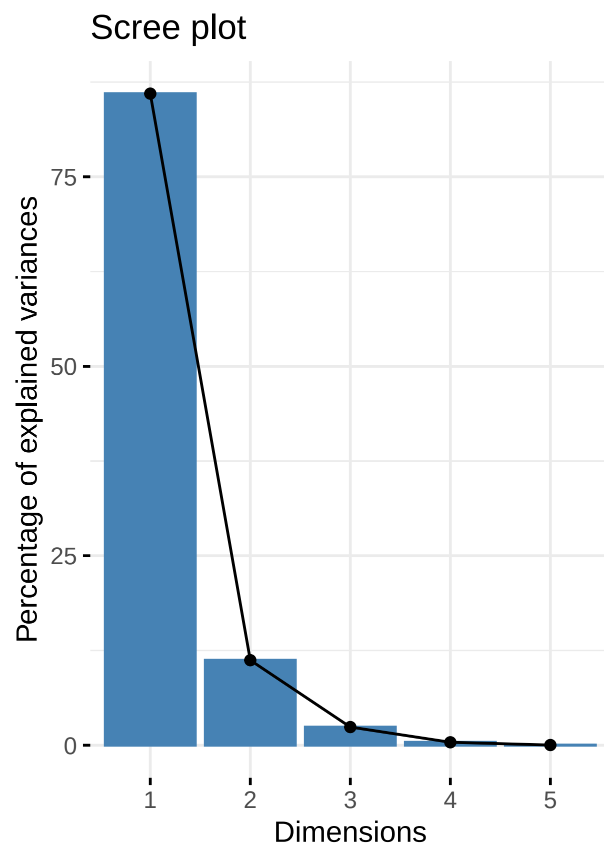
- El *error relativo para los scores* de las CP's, calculado por la $||.||_\infty$ es: 2.8311423
- El vector con los errores relativos para $\lambda$'s estimadas es

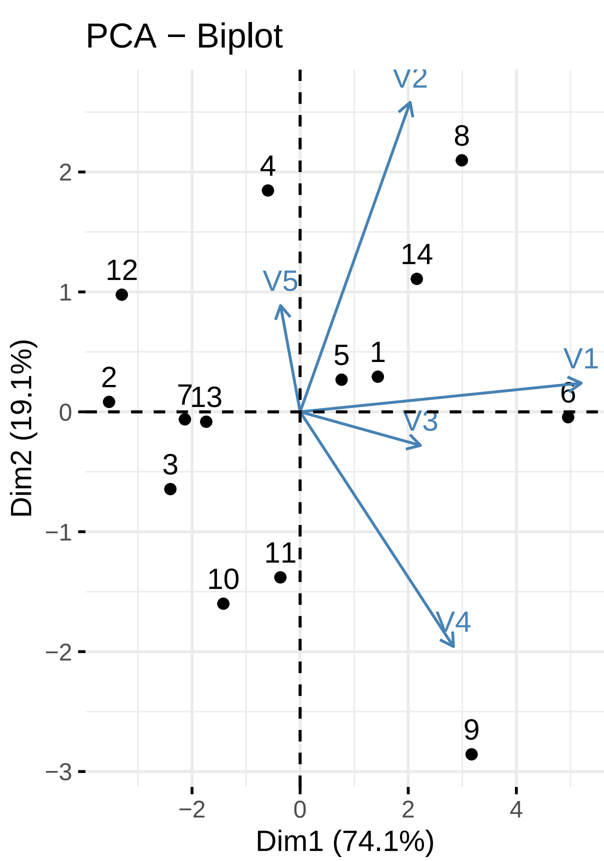
| | lambda |
|---|---|
| Comp.1 | -1.7117792 |
| Comp.2 | -0.9300589 |
| Comp.3 | -0.9090022 |
| Comp.4 | -0.0005601 |
| Comp.5 | -0.0002891 |

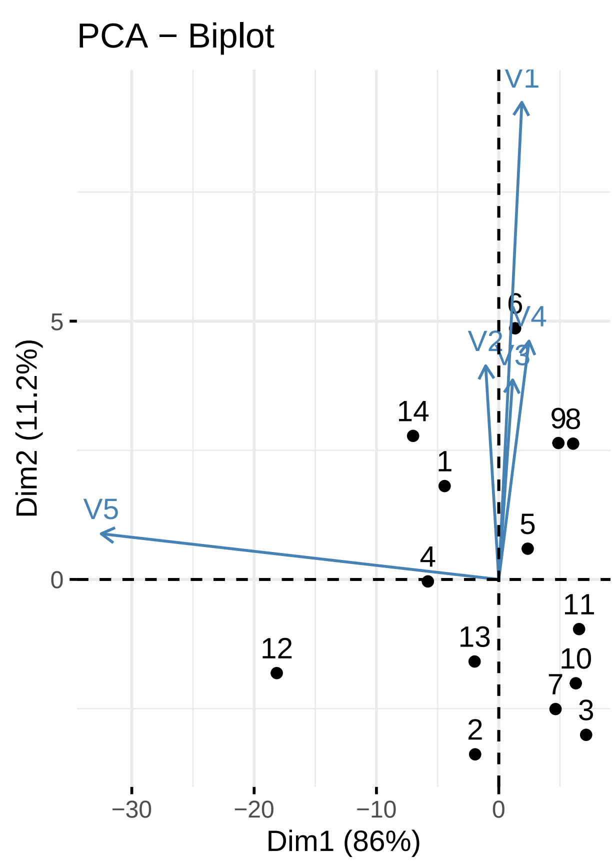
Se puede concluir que el cambio de escala SÍ afecta la estimación de los componentes principales, y por mucho.

## 3b. Interpretación



```
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## V1  0.781                0.542  0.302
## V2  0.306  0.764 -0.162 -0.545
## V3  0.334                      -0.937
## V4  0.426 -0.579  0.220 -0.636  0.172
## V5         0.262  0.962
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## V1         0.782         0.541  0.302
## V2         0.350  0.764 -0.540
## V3         0.327 -0.101        -0.938
## V4         0.391 -0.632 -0.642  0.172
## V5 -0.994
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

Para el caso de los datos originales, $Y_1$ es una ponderación de las 5 variables, dándole mayor peso a la primer variable. La segunda CP es una distinción entre la segunda variable y quinta variable contra la cuarta.

En el caso de los datos modificados, $Y_1^{'}$ se inclina *totalmente* hacia la última variable. Mientras que la segunda componente es un ponderaje de las primeras cuatro variables originales, se observa con mayor claridad en el *biplot*.

## 3c. Describir efectos

Vemos que el cambio de escala distorciona por completo los componentes principales y por ende su interpretación. Es recomendable en este caso usar $R$, o lo que es lo mismo, *estandarizar* los datos (studentizarlos).