Análisis de Componentes Principales

Luis Escobar

15 September 2018

Brevísimo repaso

Para entender el procedimiento que se realiza con el método de componentes principales, vale la pena hacer un breve recuento de algunos conceptos y médidas estadísticas importantes: Normalización, Matriz de Covarianzas, Matriz de Correlaciones, Eigenvalores y Eigenvectores.

Normalización

Recordemos que podemos normalizar (i.e., transformar una serie de datos para que tenga media cero y varianza 1) restando a cada punto de la serie la media de los datos y escalando con la varianza. El escalamiento se usa principalmente cuando puede haber problemas de escalas en nuestros datos. Por ejemplo, si estamos realizando un análisis en el que tenemos una variable que mida el salario de una muestra de personas, y otra que mida su estatura, nos encontraremos con que, dado que el salario generalmente se encuentra en miles de pesos, y la estatura se encuentra en metros, en cualquier análisis que realicemos, los resultados estarán sumamente sesgados por la magnitud de la variable "salario", por lo que, para hacer comparables las dos variables, es una buena práctica normalizar los datos. Para normalizar nuestros datos, hacemos:

$$\frac{X - \overline{X}}{Var(X)} \tag{1}$$

Matriz de Covarianzas.

Mientras que la media y la varianza son medidas de tendencia central y de dispersión, respectivamente, para una serie de datos, existen medidas similares para la interacción entre dos o más series de datos. Una de esas medidas es la covarianza. La covarianza de dos series se calcula como:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{2}$$

La covarianza de una serie consigo mismo da como resultado su varianza. Cuando se tienen dos o más series, podemos acomodar las covarianzas en una matriz.

Matriz de Correlaciones

Definimos la correlación entre dos variables como:

$$corr_{xy} = \frac{s_{xy}}{Var(x)Var(y)} \tag{3}$$

Es decir, es el cociente de la covarianza entre el producto de sus varianzas. El coeficiente de correlación nos dice qué tan "sincronizados" son los movimientos entre las variables. Si la correlación es igual a 1 (valor máximo) quiere decir que el aumento (disminución) de una serie tiene un movimiento en la misma dirección y

de la misma magnitud en la otra serie, análogamente, si la correlación es cero el movimiento es en dirección contraria. Cuando tenemos dos o más series también es posible crear una matriz de correlaciones.

Eigenvalores

Sea A una matriz de tamaño $n \times n$ y sea I la matriz identidad del mismo tamaño. Los eigenvalores de A son los valores $\lambda_1, \lambda_2, \ldots, \lambda_n$ que satisfacen la ecuación polinomial $|A - \lambda I| = 0$ (donde $||$ es el determinante).

Por otro lado, los eigenvectores son aquellos que satisfacen $A\mathbf{x} = \lambda\mathbf{x}$ o lo que es lo mismo $(A - \lambda I)\mathbf{x} = 0$. Si tenemos n eigenvalores, entonces tenemos n eigevectores asociados a cada eigenvalor.

La importancia de los eigenvalores-eigenvectores reside en que, cualquier matriz A puede ser reconstruida a partir de ellos de la siguiente forma:

$$A = \sum_{i=1}^{n} \lambda_i e_i e_i^{'} \tag{4}$$

donde (λ_i, e_i) son los pares de eigenvalores-eigenvectores de una matriz. A la ecuación anterior se le conoce como la **descomposición espectral** de una matriz. *IMPORTANTE*, para que la descomposición espectral sea posible, es necesario que nuestra matriz sea simétrica y cuadrada, ésto para garantizar que todos los eigenvalores sean positivos.

Para una descripción mucho más detallada de los eigenvalores y los eigenvectores véase (Johnson and Wichern 1988)

Ejemplo (a mano, en el salón)

Análisis de Componentes Principales

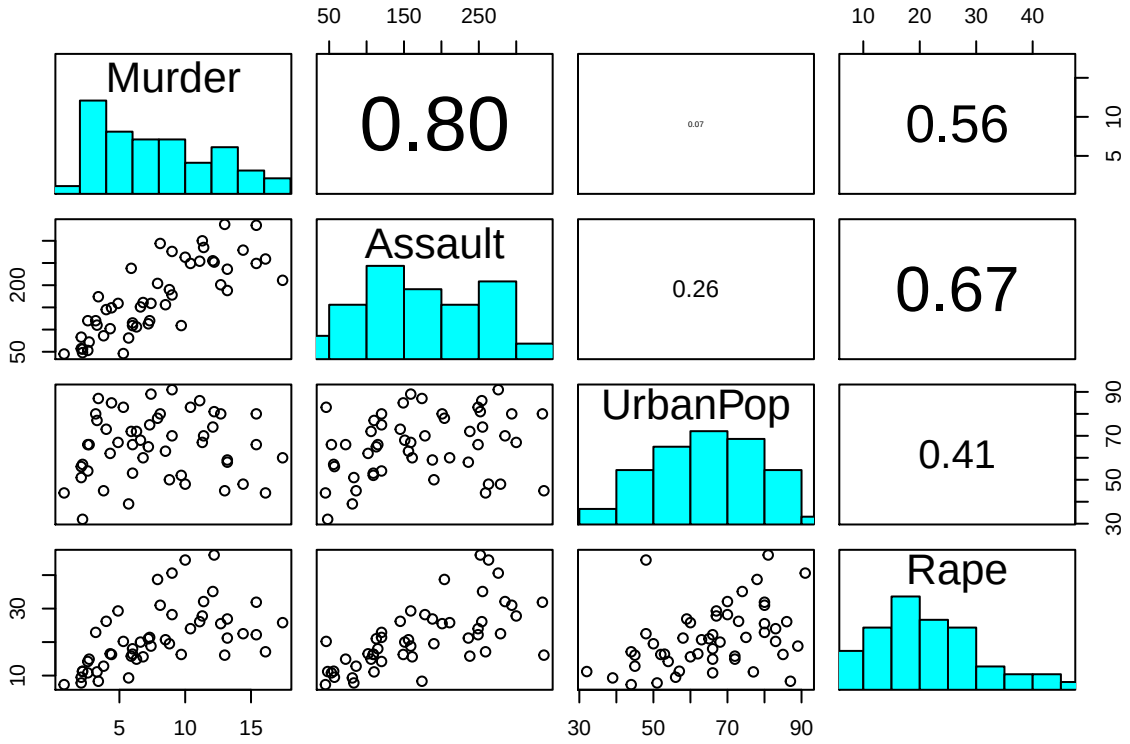
El análisis de componentes principales es una de las herramientas más usadas para el análisis de datos, sea en minería de datos, machine learning, ciencia de datos, etc. Su importancia reside en que es una herramienta poderosa que nos permite extraer información de un conjunto de datos sumamente complejos. El objetivo principale del Análisis de Componentes Principales (Principal Components Analysis, o PCA) es el de reducción de dimensionalidad (dimensionality reduction). Por ejemplo, consideremos el subconjunto de datos "USArrests". Dichos datos contienen estadísticas de número de arrestos por cada 100,000 habitantes para cada estado en EE.UU. en tres categorías: asesinato (Murder), asalto (Assault) y violación (Rape), además, contiene el porcentaje de la población de cada estado que vive en un área urbana.

```
suppressMessages(library(tidyverse))
suppressMessages(library(gridExtra))
data("USAccDeaths")
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Es claro que cada uno de los puntos en nuestro conjunto de datos no es visible, pues se encuentran en un espacio de cuatro dimensiones. A lo más, podemos aspirar a ver la relación que guardan entre cada par de observaciones

```
pairs(USArrests, upper.panel = panel.cor, diag.panel = panel.hist)
```



Resultaría muy útil poder explicar los datos con un subconjunto especial de los mismos que nos permitiera realizar el análisis clásico que aún es observable. Es aquí donde entra el Análisis de Componentes Principales (de aquí en adelante PCA). En el PCA se busca explicar la estructura de varianza-covarianza de un conjunto de datos a través de combinaciones lineales de las variables que forman dicho conjunto, con la finalidad de reducir la cantidad (y la dimensión) de los datos a análizar, así como facilitar la interpretación del conjunto de datos en su totalidad, tarea que resulta sumamente compleja cuando se tiene un gran número de variables, el cual puede ser tan grande como se quiera. Es importante mencionar que sólo a través de la totalidad de las variables es posible explicar en su totalidad la varianza de los datos, sin embargo, con PCA es posible explicar **la mayoría** de la varianza con unas cuántas variables. Generalmente, el PCA es un paso intermedio en un proceso de investigación, pues los resultados pueden ser utilizados en otros procedimientos, por ejemplo, análisis de regresión o de clustering.

Obtenemos las siguientes definiciones de (Johnson and Wichern 1988): En un conjunto de datos tenemos $X_1, X_2, \ldots, X_p$ variables aleatorias. Las componentes principales son combinaciones lineales de dichas p variables, y representan los ejes de un nuevo sistema de coordenadas sobre el que se proyectan los puntos originales del conjunto de datos.

Sea $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ con matriz de covarianzas Σ, cuyos eigenvalores son $\lambda_1, \lambda_2, \ldots, \lambda_p$. Las componentes principales son las combinaciones lineales

$$\begin{aligned} Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{5}$$

tales que su varianza $Var(Y_i) = a_i'\Sigma a_i$ es lo más grande posible y que estén no correlacionadas entre sí. La primer componente principal es aquella con varianza máxima, la segunda componente es aquella con la

segunda varianza más grande y así sucesivamente.

Un resultado importante dice que el calculo de las componentes principales depende únicamente de la matriz de covarianzas. Así, tenemos que si nuestro conjunto de datos tiene matriz de covarianzas Σ, con eigenvalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ y eigenvectores $e_1, e_2, \ldots, e_p$, entonces, calculamos la i-ésima componente como:

$$Y_i = e_i^{'}\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p \tag{6}$$

así, $Var(Y_i) = e_i^{'}\Sigma e_i = \lambda_i$

Una de las propiedades más importantes de las componentes es la cantidad de variabilidad del conjunto que logran capturar. Una propiedad que cumplen los eigenvalores de la matriz de covarianzas dice que, siendo $(\lambda_i, e_i) \quad \forall \quad i \in \{1, 2, \ldots, p\}$ los pares de eigenvalores-eigenvectores de Σ y $Y_i = e_i^{'}X \quad \forall \quad i \in \{1, 2, \ldots, p\}$ las componentes, entonces:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i) \tag{7}$$

lo anterior nos dice que la suma de los eigenvalores es la suma de la varianza total de las variables que forman nuestros datos, por lo que la variabilidad explicada por cada componente es el eigenvalor asociada a la misma. Es decir, la proporción de la varianza total explicada por la k-ésima componente es:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \tag{8}$$

Ejemplos (en clase)

El siguiente ejemplo sale de https://uc-r.github.io/pca, y retomaremos el data frame "USArrests".

Primero, para ver el impacto de la diferencia de escalas en las variables, vemos cómo difieren las varianzas:

```
apply(USArrests, 2, var)
```

```
##     Murder    Assault   UrbanPop       Rape
##   18.97047 6945.16571  209.51878   87.72916
```

La columna de arrestos por asaltos tiene una variable claramente mucho mayor que el resto de las variables, por lo que los resultados estarían sumamente sesgados hacia esa variable. Como mencionamos, es necesario escalar los datos.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
USArrests_Scaled <- apply(USArrests, 2, scale)
head(USArrests_Scaled)
```

```
##          Murder   Assault   UrbanPop         Rape
## [1,] 1.24256408 0.7828393 -0.5209066 -0.003416473
## [2,] 0.50786248 1.1068225 -1.2117642  2.484202941
## [3,] 0.07163341 1.4788032  0.9989801  1.042878388
```

```
## [4,] 0.23234938 0.2308680 -1.0735927 -0.184916602
## [5,] 0.27826823 1.2628144  1.7589234  2.067820292
## [6,] 0.02571456 0.3988593  0.8608085  1.864967207
```

Calculando las componentes paso a paso.

Primero, es necesario obtener la matriz de covarianzas

```
USArrests_Cov <- cov(USArrests_Scaled)
USArrests_Cov
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

Ahora, obtenemos los eigenvalores:

```
USArrests_Eigen <- eigen(USArrests_Cov)
print(USArrests_Eigen)
```

```
## eigen() decomposition
## $values
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
##
## $vectors
##            [,1]       [,2]       [,3]        [,4]
## [1,] -0.5358995  0.4181809 -0.3412327  0.64922780
## [2,] -0.5831836  0.1879856 -0.2681484 -0.74340748
## [3,] -0.2781909 -0.8728062 -0.3780158  0.13387773
## [4,] -0.5434321 -0.1673186  0.8177779  0.08902432
```

En lo anterior, vemos los eigenvalores y los eigenvectores. Con fines ilustrativos, solo usaremos dos componentes. Primero, vemos la proporción de varianza que se explica con las dos primeras componentes:

```
USArrests_Eigen$values/sum(USArrests_Eigen$values)
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

Valores acumulados:

```
cumsum(USArrests_Eigen$values/sum(USArrests_Eigen$values))
```

```
## [1] 0.6200604 0.8675017 0.9566425 1.0000000
```

Con sólo dos componentes, explicamos el 86.7% de la varianza de los datos. Extraemos los eigenvectores; para ser consistentes con la notación, nombramos al vector de coeficientes a:

```
a <- USArrests_Eigen$vectors[, 1:2]
print(a)
```

```
##            [,1]       [,2]
## [1,] -0.5358995  0.4181809
## [2,] -0.5831836  0.1879856
## [3,] -0.2781909 -0.8728062
## [4,] -0.5434321 -0.1673186
```

Por default, los eigenvectores en R apuntan en sentido "negativo", por lo que los "volteamos" para dirigirlos al sentido "natural". Damos nombres a los renglones y las columnas.

```
a <- -a
rownames(a) <- c("Murder", "Assault", "UrbanPop", "Rape")
colnames(a) <- c("PC1", "PC2")
print(a)
```

```
##                PC1        PC2
## Murder   0.5358995 -0.4181809
## Assault  0.5831836 -0.1879856
## UrbanPop 0.2781909  0.8728062
## Rape     0.5434321  0.1673186
```

Ahora, proyectamos los datos originales en el subespacio generado por las dos componentes:

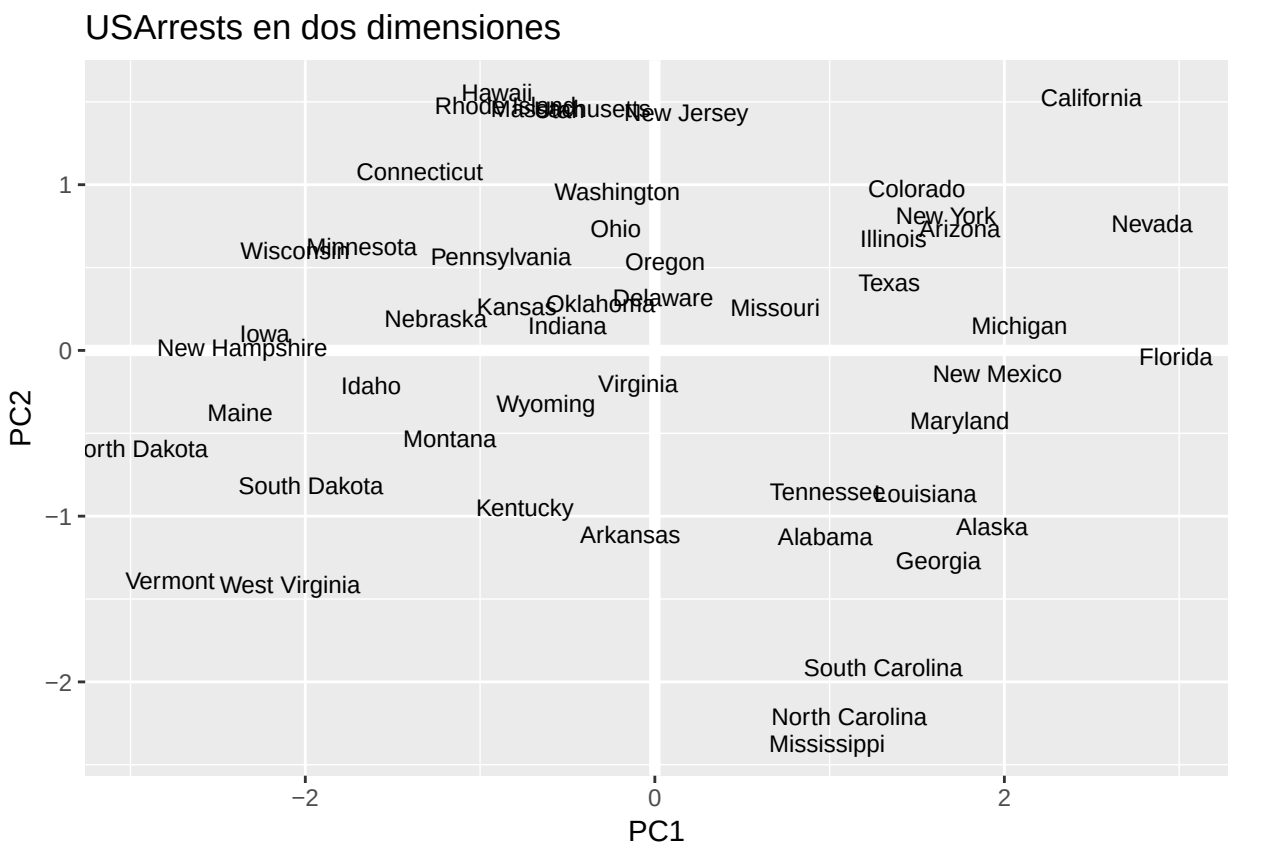
```
datos_en_subespacio <- as.matrix(USArrests_Scaled) %*% a %>% as.data.frame()
datos_en_subespacio$Estado <- rownames(USArrests)
datos_en_subespacio <- datos_en_subespacio %>% select(Estado, PC1, PC2)
print(datos_en_subespacio)
```

```
##            Estado         PC1         PC2
## 1         Alabama  0.97566045 -1.12200121
## 2          Alaska  1.93053788 -1.06242692
## 3         Arizona  1.74544285  0.73845954
## 4        Arkansas -0.13999894 -1.10854226
## 5      California  2.49861285  1.52742672
## 6        Colorado  1.49934074  0.97762966
## 7     Connecticut -1.34499236  1.07798362
## 8        Delaware  0.04722981  0.32208890
## 9         Florida  2.98275967 -0.03883425
## 10        Georgia  1.62280742 -1.26608838
## 11         Hawaii -0.90348448  1.55467609
## 12          Idaho -1.62331903 -0.20885253
## 13       Illinois  1.36505197  0.67498834
## 14        Indiana -0.50038122  0.15003926
## 15           Iowa -2.23099579  0.10300828
## 16         Kansas -0.78887206  0.26744941
## 17       Kentucky -0.74331256 -0.94880748
## 18      Louisiana  1.54909076 -0.86230011
## 19          Maine -2.37274014 -0.37260865
## 20       Maryland  1.74564663 -0.42335704
## 21  Massachusetts -0.48128007  1.45967706
## 22       Michigan  2.08725025  0.15383500
## 23      Minnesota -1.67566951  0.62590670
## 24    Mississippi  0.98647919 -2.36973712
## 25       Missouri  0.68978426  0.26070794
## 26        Montana -1.17353751 -0.53147851
## 27       Nebraska -1.25291625  0.19200440
## 28         Nevada  2.84550542  0.76780502
## 29  New Hampshire -2.35995585  0.01790055
## 30     New Jersey  0.17974128  1.43493745
## 31     New Mexico  1.96012351 -0.14141308
## 32       New York  1.66566662  0.81491072
## 33 North Carolina  1.11208808 -2.20561081
## 34   North Dakota -2.96215223 -0.59309738
```

```
## 35           Ohio -0.22369436  0.73477837
## 36       Oklahoma -0.30864928  0.28496113
## 37         Oregon  0.05852787  0.53596999
## 38   Pennsylvania -0.87948680  0.56536050
## 39   Rhode Island -0.85509072  1.47698328
## 40 South Carolina  1.30744986 -1.91397297
## 41   South Dakota -1.96779669 -0.81506822
## 42      Tennessee  0.98969377 -0.85160534
## 43          Texas  1.34151838  0.40833518
## 44           Utah -0.54503180  1.45671524
## 45        Vermont -2.77325613 -1.38819435
## 46       Virginia -0.09536670 -0.19772785
## 47     Washington -0.21472339  0.96037394
## 48  West Virginia -2.08739306 -1.41052627
## 49      Wisconsin -2.05881199  0.60512507
## 50        Wyoming -0.62310061 -0.31778662
```

Con los datos originales, podemos obtener una visualización de los estados en un espacio que sí podemos percibir:

```
ggplot(as.data.frame(datos_en_subespacio), aes(PC1, PC2)) +
  modelr::geom_ref_line(h = 0) +
  modelr::geom_ref_line(v = 0) +
  geom_text(aes(label = Estado), size = 3) +
  xlab("PC1") +
  ylab("PC2") +
  ggtitle("USArrests en dos dimensiones")
```



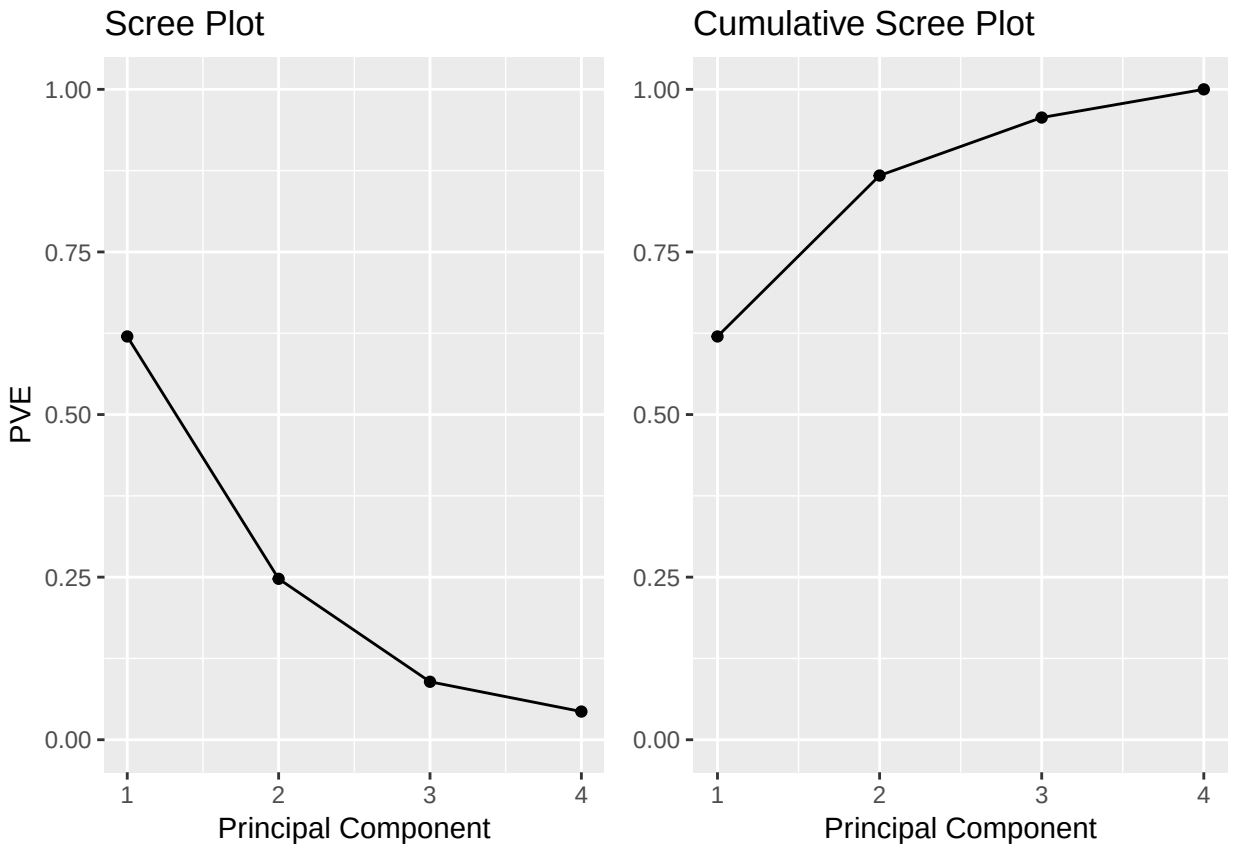
Como ya vieron, seleccionar el número óptimo de componentes es más bien arbitrario y depende del investigador. Sin embargo, una de las herramientas que se utilizan para seleccionarlas es la "ScreePlot". Donde veamos un cambio brusco en la pendiente de la gráfica es un buen punto para seleccionar el número, pues dicho cambio brusco se interpreta como un aporte marginal de cada componente a la varianza total.

```
PVE <- USArrests_Eigen$values / sum(USArrests_Eigen$values)

PVEplot <- qplot(c(1:4), PVE) +
  geom_line() +
  xlab("Principal Component") +
  ylab("PVE") +
  ggtitle("Scree Plot") +
  ylim(0, 1)

# Cumulative PVE plot
cumPVE <- qplot(c(1:4), cumsum(PVE)) +
  geom_line() +
  xlab("Principal Component") +
  ylab(NULL) +
  ggtitle("Cumulative Scree Plot") +
  ylim(0,1)

grid.arrange(PVEplot, cumPVE, ncol = 2)
```



Dependiendo del criterio de cada quien, parecería que el número óptimo de componentes es 3.

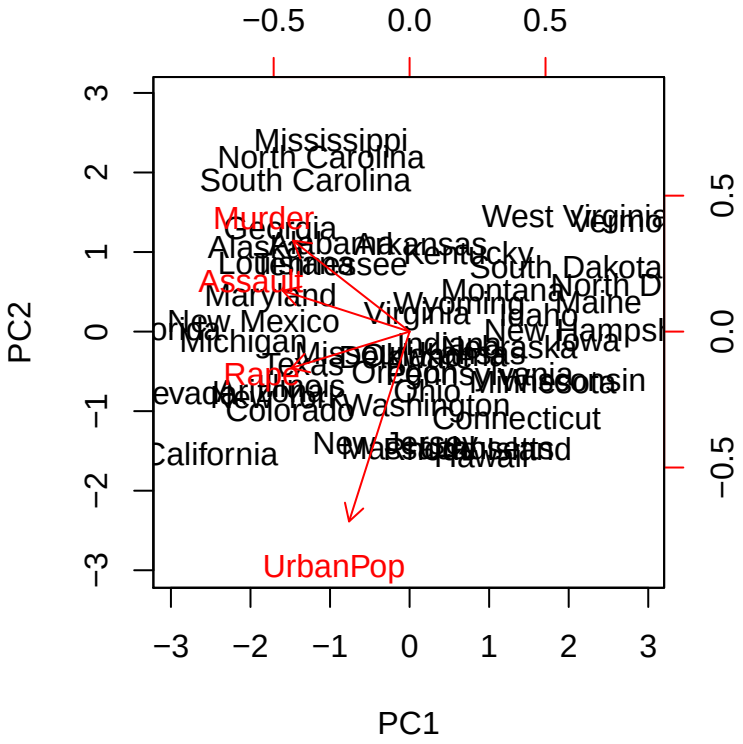
Obviamente, R tiene implementados varios métodos para calcular las componentes.

```r
resultado_pca <- prcomp(x = USArrests, scale. = T)
str(resultado_pca)
```

```
## List of 5
##  $ sdev    : num [1:4] 1.575 0.995 0.597 0.416
##  $ rotation: num [1:4, 1:4] -0.536 -0.583 -0.278 -0.543 0.418 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##   .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
##  $ center  : Named num [1:4] 7.79 170.76 65.54 21.23
##   ..- attr(*, "names")= chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ scale   : Named num [1:4] 4.36 83.34 14.47 9.37
##   ..- attr(*, "names")= chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ x       : num [1:50, 1:4] -0.976 -1.931 -1.745 0.14 -2.499 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##   .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
##  - attr(*, "class")= chr "prcomp"
```

Podemos obtener una gráfica que nos ayuda a ver la influencia de cada variable en cada componente.

```r
biplot(resultado_pca, scale = 0)
```



Referencias

Johnson, R. A., and D. W. Wichern, eds. 1988. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.