Reducción de dimensiones

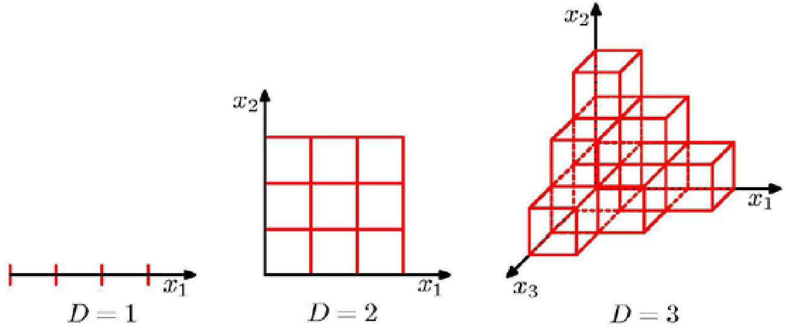
Una Visión General

Antonio Alvarez
aqalag94@gmail.com
05.06.2020

Maldición de la dimensión

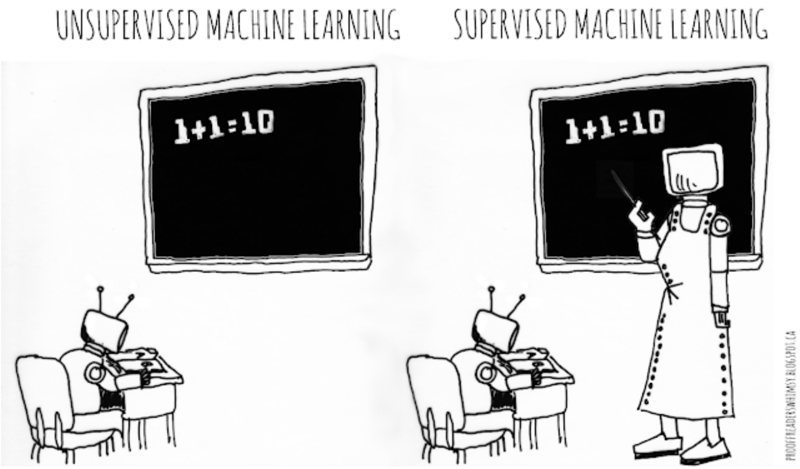
La maldita data

- Mientras aumenta el numero de dimensiones, el volumen de datos puede aumentar mas rápido
- ... haciendo a los datos **sparse**: la mayoría de los puntos están muy "lejos" unos de otros
- Para tener resultados estadísticos confiables, necesitamos exponencialmente mas observaciones que características



Aprendizaje no supervisado

- No existe un objetivo simple
- Suele ser mas subjetivo
- Datos sin etiquetas
- Estas técnicas son muy importante en diversas aplicaciones



Proyección

- En la practica, no todos las columnas están distribuidas *uniformemente*
- Así, la mayoría de estas columnas están un subespacio de baja dimensionalidad
- Algebraicamente representado como una combinación linear:

$$Z = \phi_1 X_1 + \cdots + \phi_p X_p$$

En 3 dimensiones

Antecedentes estadísticos

De un vector (o variable), podemos determinar:

- El promedio *total* de las variables:

$$\mu_A = \frac{1}{n}(a_1 + \ldots + a_n)$$

- La varianza y covarianza de los datos

$$Var(A) = \frac{1}{n-1}\left((a_1 - \mu_A)^2 + \cdots + (a_n - \mu_A)^2\right)$$

$$Cov(A, B) = \frac{1}{n-1}\left((a_1 - \mu_A)(b_1 - \mu_B) + \cdots + (a_n - \mu_A)(b_n - \mu_B)\right)$$

Antecedentes estadísticos

Podemos calcular el promedio total de las variables como un vector:

$$\overrightarrow{\mu} = \frac{1}{n}(\overrightarrow{x_1} + \ldots + \overrightarrow{x_n})$$

Es común que se "centren" los datos para tener promedio 0. Sea B una matriz de $m \times n$ donde su columna i es $\overrightarrow{x_i} - \overrightarrow{\mu_i}$:

$$B = |\overrightarrow{x_1} - \overrightarrow{\mu_1}| \ldots |\overrightarrow{x_n} - \overrightarrow{\mu_n}|$$

Así, podemos definir la matriz de covarianza como:

$$S = \frac{1}{n-1} BB^T$$

Esta matriz **siempre** sera cuadrada y podrá encontrarse los autovalores y autovectores

Componentes principales

Se puede solucionar este problema por **descomposición de valores singulares** de la matriz de datos estandarizada

Si $\mathbf{X}$ es una matriz de $I \times J$ dimensiones con rango L donde $L \leq \min\{I, J\}$, entonces la **DVS** de $\mathbf{X}$ es:

$$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^{\mathrm{T}}$$

donde $\mathbf{U}$ es la matriz izquierda de vectores singulares $I \times L$, $\mathbf{V}$ es la matriz derecha de vectores singulares $I \times L$ y $\mathbf{\Delta}$ es la matriz diagonal de los valores singulares.

La *inercia de la columna* es definida como la suma de los elementos cuadrados de $\mathbf{\Delta}$:

$$\lambda_j^2 = \sum_i^J x_{i,j}^2$$

La suma de todas las inercias resulta en la *inercia total*, denotada por $\mathcal{I}$

Componentes principales

- Dado un plano hiperdimensional, identifica los axis que capturan la mayor inercia posible

El primer componente esta definido como la combinación linear:

$$\mathbf{w}_{(1)} = \arg\max_{||\mathbf{w}||=1} \left\{ \sum_i (\mathbf{x}_{(i)} \cdot \mathbf{w})^2 \right\} = \arg\max_{||\mathbf{w}||=1} \left\{ ||\mathbf{X}\mathbf{w}||^2 \right\}$$

, el cual define la *máxima* varianza capturada. De ahí los demás componentes deben de satisfacer:

$$\mathbf{w}_k = \arg\max_{||\mathbf{w}||=1} \left\{ ||\hat{\mathbf{X}}_k \mathbf{w}||^2 \right\}$$

donde $\hat{\mathbf{X}}_k$ es la sustracción de $\mathbf{X}$ por el k componente. Es decir:

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)} \mathbf{w}'_{(s)}$$

R

Antecedentes de álgebra linear

Teorema 1: Si A es cualquier matriz con dimensiones $m \times n$, entonces la matriz A^TA de $n \times n$ o AA^T de $m \times m$ es simétrica

Por lo tanto, basado en el **Teorema Espectral**, si A es simétrica, entonces es A es *diagonizable* y solo tiene *autovalores* (λ) reales. Es decir:

$$A\overrightarrow{v_i} = \lambda\overrightarrow{v_i}$$

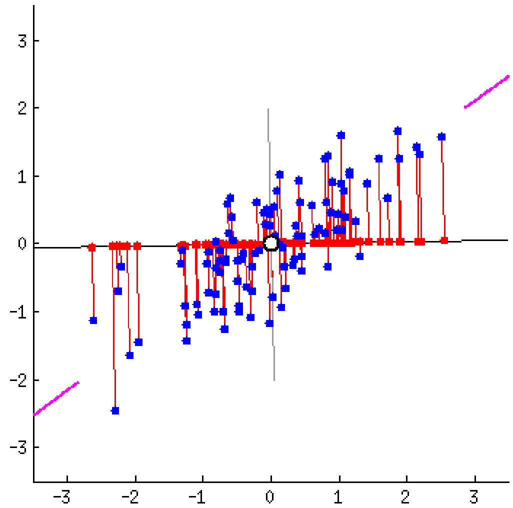
Proposición 1: Las matrices A^TA y AA^T tienen los *mismos* autovalores

Proposición 2: Los autovalores de A^TA y AA^T son valores *positivos*

R

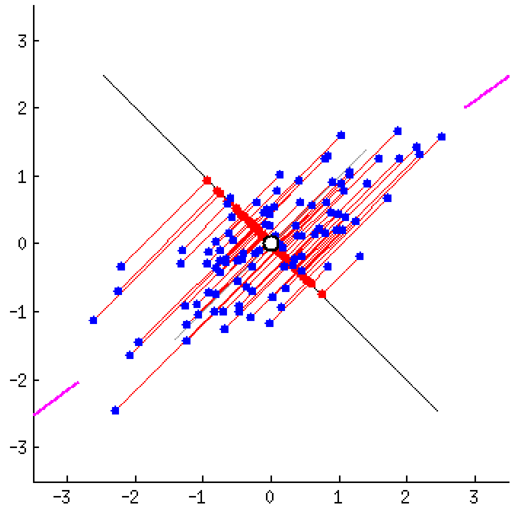
Explicación en pizarra

En 2 dimensiones



GIF en: https://i.stack.imgur.com/Q7HIP.gif

En 2 dimensiones



GIF en: https://i.stack.imgur.com/lNHqt.gif

Objetivos de PCA

1. Extraer la información mas importante de la tabla de datos
2. Comprimir el tamaño de la base al mantener solo la información mas importante
3. Simplificar la descripción de los datos
4. Analizar la estructura de las observaciones y de las variables

Es importante estandarizar

- Así como en cluster, diferentes columnas pueden tener diferentes rangos
- Es importante que todos estén centrados en 0
- Solo en poca ocasiones, no se estandariza

PCA es caro...

- Multiplicación de matrices suelen ser pesadas
- Pensar en el numero de observaciones que se tenga
- Tener una muestra puede ser mas adecuado en ciertos casos

Introducción a `airquality`

- Lecturas diarias de la calidad del aire en Nueva York, de Mayo a Septiembre 1973

- Descrita en el libro seminal *Graphical Methods for Data Analysis*

```r
data("airquality")

air <- as_tibble(airquality)

glimpse(air)
```

```
## Observations: 153
## Variables: 6
## $ Ozone   <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14, 18, 1...
## $ Solar.R <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256, 290,...
## $ Wind    <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, 8.6, 13.8, 20.1, 8.6, 6.9...
## $ Temp    <int> 67, 72, 74, 62, 56, 66, 65, 59, 61, 69, 74, 69, 66, 68, 58,...
## $ Month   <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,...
## $ Day     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...
```

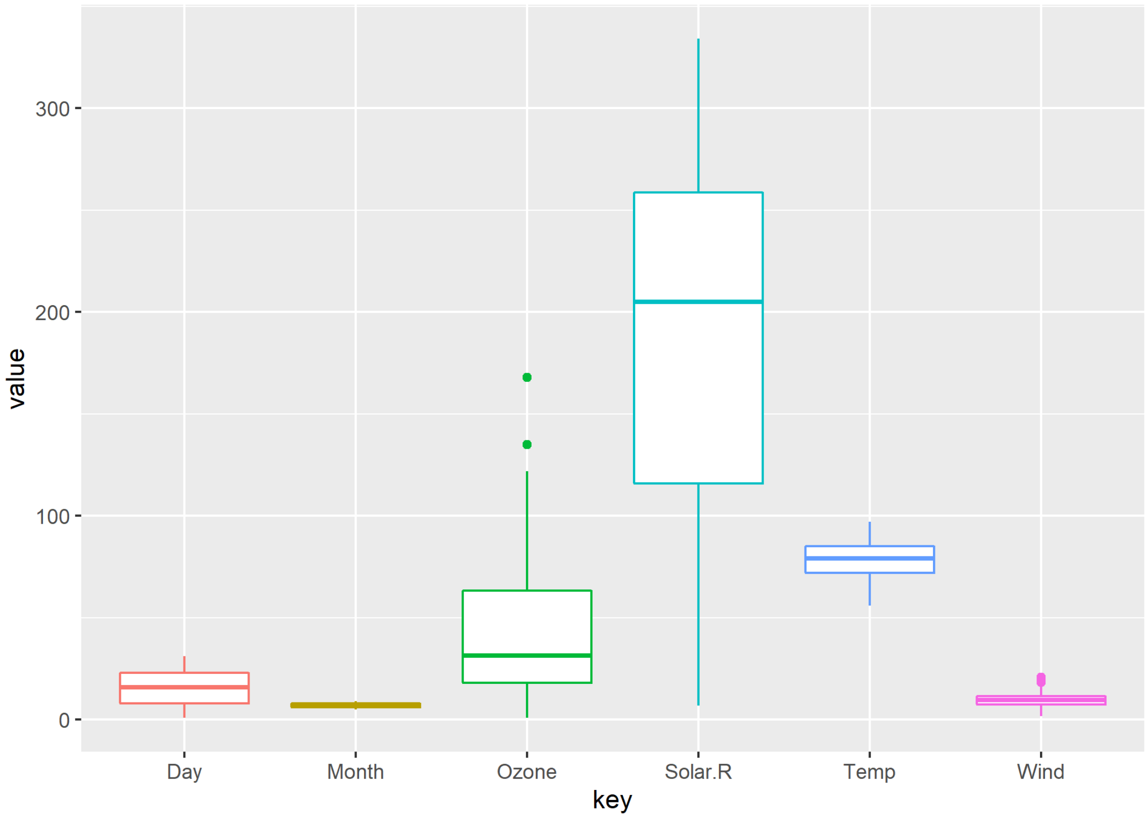
summary

```
summary(air)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

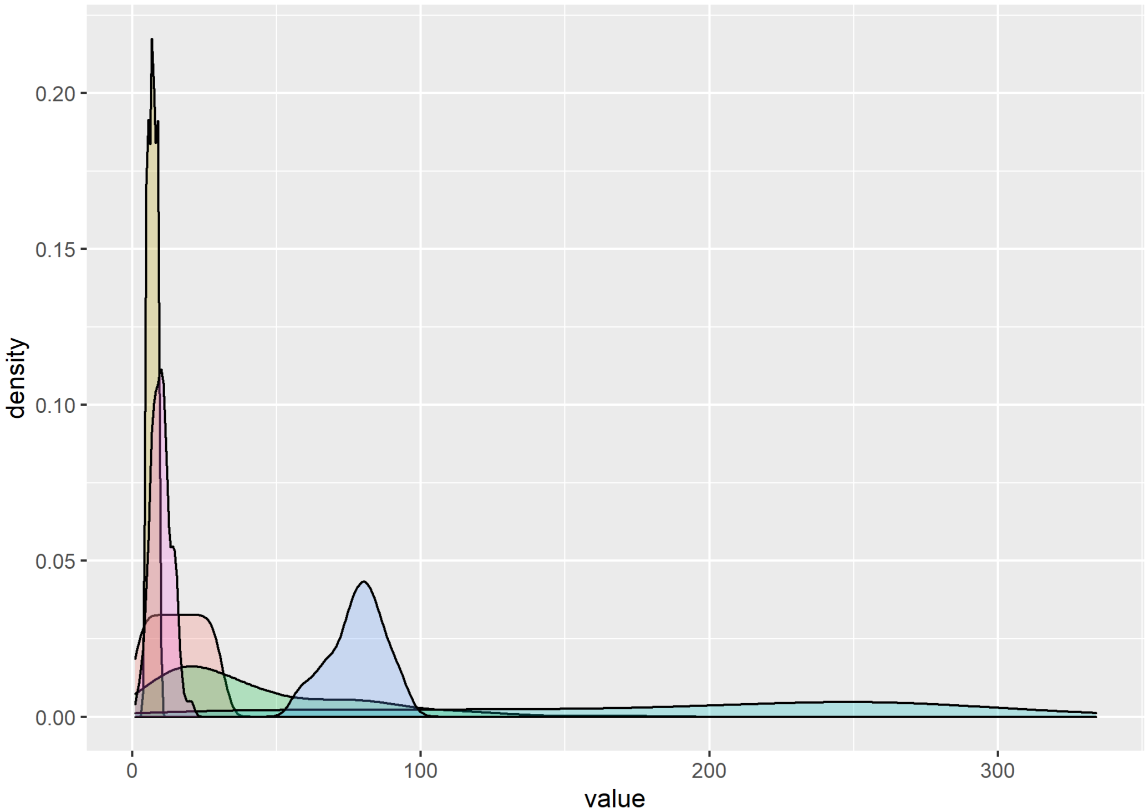
EDA

```
air %>% gather(key, value) %>%
  ggplot(aes(key, value, color=key)) + geom_boxplot() +
  guides(color=FALSE)
```



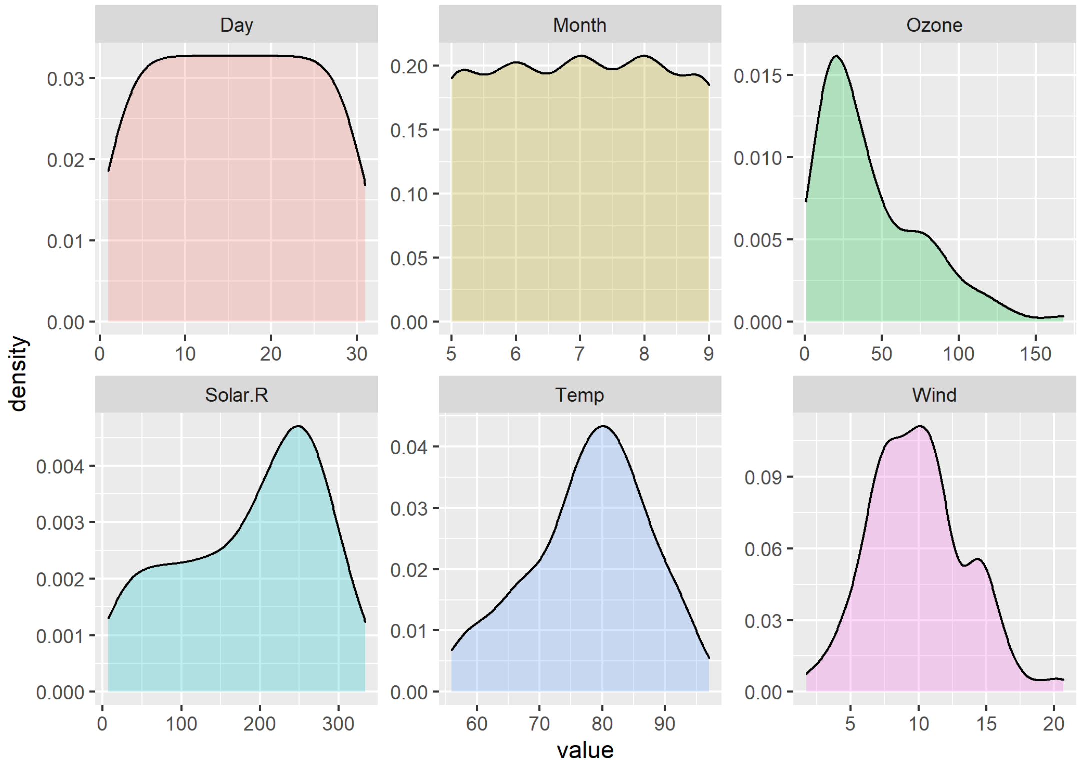
EDA

```
air %>% gather(key, value) %>%
  ggplot(aes(value, fill=key)) + geom_density(alpha=0.25) +
  guides(fill=FALSE)
```



EDA

```
air %>% gather(key, value) %>%
  ggplot(aes(value, fill=key)) + geom_density(alpha=0.25) +
  facet_wrap(~key, scales = "free") +
  guides(fill=FALSE)
```



PCA en R

Varios paquetes, varias funciones

1. `stats::prccomp`
2. `stats::princomp`
3. `ade4::PCA`
4. **`FactoMineR::PCA`**