

# Aprendizaje No Supervisado

César Gómez

21 de enero de 2022

# El reto del aprendizaje no supervisado

- En el contexto de *aprendizaje supervisado* típicamente se tiene acceso a un conjunto de  $p$  atributos  $X_1, \dots, X_p$  medidos sobre  $n$  observaciones, además de una respuesta  $Y$  también medida sobre las mismas  $n$  observaciones.  
El objetivo es predecir  $Y$  utilizando  $X_1, \dots, X_p$ .

# El reto del aprendizaje no supervisado

- En el contexto de *aprendizaje supervisado* típicamente se tiene acceso a un conjunto de  $p$  atributos  $X_1, \dots, X_p$  medidos sobre  $n$  observaciones, además de una respuesta  $Y$  también medida sobre las mismas  $n$  observaciones.  
El objetivo es predecir  $Y$  utilizando  $X_1, \dots, X_p$ .
- En el contexto de **aprendizaje no supervisado** (ANS), solo se tiene acceso a un conjunto de atributos  $X_1, \dots, X_p$  medidos sobre  $n$  observaciones.

# El reto del aprendizaje no supervisado

- En el contexto de *aprendizaje supervisado* típicamente se tiene acceso a un conjunto de  $p$  atributos  $X_1, \dots, X_p$  medidos sobre  $n$  observaciones, además de una respuesta  $Y$  también medida sobre las mismas  $n$  observaciones.  
El objetivo es predecir  $Y$  utilizando  $X_1, \dots, X_p$ .
- En el contexto de **aprendizaje no supervisado** (ANS), solo se tiene acceso a un conjunto de atributos  $X_1, \dots, X_p$  medidos sobre  $n$  observaciones.
- En ANS no estamos interesados en predicción, por que no disponemos de una respuesta  $Y$  asociada.

- De forma general el objetivo consiste en descubrir patrones interesantes sobre el conjunto de medidas en los atributos  $X_1, \dots, X_p$ .
- Por ejemplo, hay alguna forma informativa de visualizar los datos?
- Es posible descubrir subgrupos entre las variables o entre las observaciones?.

El aprendizaje no supervisado consiste en un conjunto de diversas técnicas para responder preguntas como las anteriores. Nos concentraremos principalmente en:

- **Análisis de componentes principales. PCA** por sus siglas en Inglés.
- **Clustering o agrupamiento.**

- El aprendizaje no supervisado es llevado a cabo comúnmente como parte de un *análisis exploratorio de datos*.

- El aprendizaje no supervisado es llevado a cabo comúnmente como parte de un *análisis exploratorio de datos*.
- En ANS puede ser difícil, evaluar los resultados, debido a que no hay un mecanismo universalmente valido para llevar a cabo validación cruzada o validar los resultados en un conjunto de datos de prueba independiente. Esto se debe a que no hay una variable respuesta.

Algunas aplicaciones del ANS incluyen, por ejemplo:

- El aprendizaje no supervisado es llevado a cabo comúnmente como parte de un *análisis exploratorio de datos*.
- En ANS puede ser difícil, evaluar los resultados, debido a que no hay un mecanismo universalmente valido para llevar a cabo validación cruzada o validar los resultados en un conjunto de datos de prueba independiente. Esto se debe a que no hay una variable respuesta.

Algunas aplicaciones del ANS incluyen, por ejemplo:

- Un investigador en cáncer puede medir los niveles de expresión génica en 100 pacientes diagnosticados con cáncer de mama. Entonces el investigador puede estar interesado en encontrar subgrupos entre los pacientes para tener una mejor comprensión de la enfermedad.



- En un sitio de ventas online, pueden interesarse también en identificar subgrupos de clientes con características de compra similares basados en las historias de compras de estos clientes. Entonces, publicidad y promociones pueden ser enviadas a estos clientes dependiendo del perfil y gusto de estos últimos.

# Análisis de componentes principales. PCA

- El **análisis de componentes principales**, (PCA por sus siglas en inglés, *principal component analysis*) es una técnica para reducir la dimensión de una matriz de datos  $\mathbf{X}$  de tamaño  $n \times p$ .

# Análisis de componentes principales. PCA

- El **análisis de componentes principales**, (PCA por sus siglas en inglés, *principal component analysis*) es una técnica para reducir la dimensión de una matriz de datos  $\mathbf{X}$  de tamaño  $n \times p$ .
- Las componentes principales hacen referencia a las direcciones en el espacio de atributos  $X_1, \dots, X_p$  sobre las cuales los datos presentan la mayor variabilidad, en breve definimos de forma más precisa las *componentes principales*.

- La *primera componente principal* de un conjunto de atributos  $X_1, \dots, X_p$  corresponde a la combinación lineal, normalizada de los atributos

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p. \quad (1)$$

Que posee la mayor varianza.

Es normalizada en el sentido de que  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .

A los coeficientes  $\phi_{11}, \phi_{21}, \dots, \phi_{p1}$  se los denomina, pesos de la primera componente principal.

- La *primera componente principal* de un conjunto de atributos  $X_1, \dots, X_p$  corresponde a la combinación lineal, normalizada de los atributos

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p. \quad (1)$$

Que posee la mayor varianza.

Es normalizada en el sentido de que  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .

A los coeficientes  $\phi_{11}, \phi_{21}, \dots, \phi_{p1}$  se los denomina, pesos de la primera componente principal.

- Al vector  $\phi_1 = [\phi_{11}, \phi_{21}, \dots, \phi_{p1}]^T$  como el *vector de pesos de la primera componente principal*.

# Como se calculan las componentes principales

Dada una matriz de datos  $n \times p$ ,  $\mathbf{X}$ . Se asume que las columnas de  $\mathbf{X}$  han sido estandarizadas, (centradas y escaladas), se busca la combinación lineal de los valores de los atributos de la muestra de la forma

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i,2} + \cdots + \phi_{p1}x_{i,p}, \quad i = 1, \dots, n. \quad (2)$$

que posee la máxima varianza muestral, sujeto a la restricción  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .

De forma más precisa, se tiene que las componentes del vector  $\phi_1 = [\phi_{11}, \phi_{21}, \dots, \phi_{p1}]^T$  corresponden a la solución del problema de optimización

$$\underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \quad \text{sujeto a} \quad \sum_{j=1}^p \phi_{j1}^2 = 1. \quad (3)$$

- Se referirá a los valores  $z_{11}, z_{21}, \dots, z_{n1}$  como los *scores* de la primera componente principal.



- Se referirá a los valores  $z_{11}, z_{21}, \dots, z_{n1}$  como los *scores* de la primera componente principal.
- Estos no son más que las componentes o proyecciones de cada observación  $\mathbf{x}_i$  sobre la primera dirección principal. O simplemente la coordenada en  $Z_1$  de cada una de las observaciones  $\mathbf{x}_i$

- Se referirá a los valores  $z_{11}, z_{21}, \dots, z_{n1}$  como los *scores* de la primera componente principal.
- Estos no son más que las componentes o proyecciones de cada observación  $\mathbf{x}_i$  sobre la primera dirección principal. O simplemente la coordenada en  $Z_1$  de cada una de las observaciones  $\mathbf{x}_i$
- El vector  $\phi_1 = [\phi_{11}, \phi_{21}, \dots, \phi_{p1}]^T$  define una dirección en el espacio de atributos  $X_1, \dots, X_p$  en que los datos poseen máxima variación.

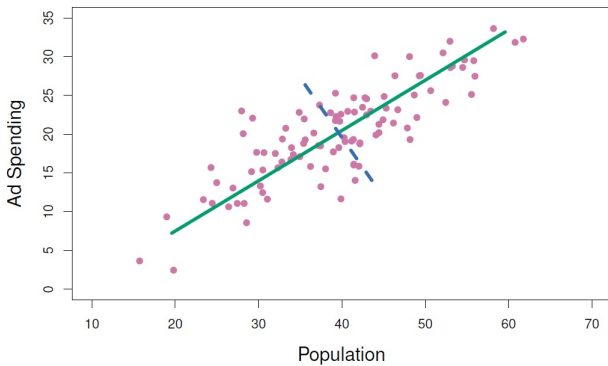


Figura 1: Primeras 2 direcciones principales.

- Después de que se ha determinado la primera componente principal  $Z_1$ , La segunda componente principal  $Z_2$  es la combinación lineal de los atributos  $X_1, \dots, X_p$ , es decir

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \dots + \phi_{p2}X_p. \quad (4)$$

que presenta máxima variabilidad y es no correlacionada con  $Z_1$ .

- Después de que se ha determinado la primera componente principal  $Z_1$ , La segunda componente principal  $Z_2$  es la combinación lineal de los atributos  $X_1, \dots, X_p$ , es decir

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \dots + \phi_{p2}X_p. \quad (4)$$

que presenta máxima variabilidad y es no correlacionada con  $Z_1$ .

- En este caso los pesos  $\phi_{j2}$  del vector que define la segunda dirección principal  $\phi_2 = [\phi_{12} \ \phi_{22} \ \dots \ \phi_{p2}]$  resuelve el problema de optimización

$$\underset{\phi_{12}, \dots, \phi_{p2}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j2} x_{ij} \right)^2 \right\}$$

sujeto a

$$\sum_{j=1}^p \phi_{j1}^2 = 1,$$

$$\sum_{j=1}^p \phi_{j1} \phi_{j2} = 0.$$

(5)

## Relación con autovalores

Dada una matriz simétrica  $\Sigma$  como por ejemplo la matriz de covarianza de un vector  $X = (X_1, X_2, \dots, X_p)$ , entonces existe una matriz  $\Phi$  tal que

$$\Sigma = \Phi \Lambda \Phi^T, \quad (6)$$

$$= \Phi \begin{bmatrix} \lambda_1 & 0 & \cdots & & \\ 0 & \lambda_2 & 0 & \cdots & \\ \vdots & & & \ddots & \\ 0 & 0 & \cdots & 0 & \lambda_p \end{bmatrix} \Phi^T, \quad (7)$$

donde  $\Phi^{-1} = \Phi^T$ , es decir  $\Phi \cdot \Phi^T = \Phi^T \cdot \Phi = I$ .

- Las columnas de  $\Phi$  representan los **autovectores** de  $\Sigma$  correspondientes a los autovalores  $\lambda_i$  que pueden ser ordenados de la forma  $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ .
- También puede mostrarse que si

$$\Sigma = \Phi \Lambda \Phi^T, \quad (8)$$

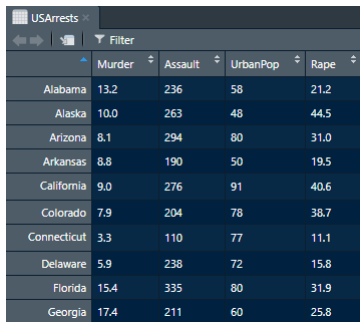
entonces

$$\begin{aligned} \text{Var}(X) = \text{Tr}(\Sigma) &= \text{Tr}(\Lambda) \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p. \end{aligned} \quad (9)$$

(10)



# Ejemplo

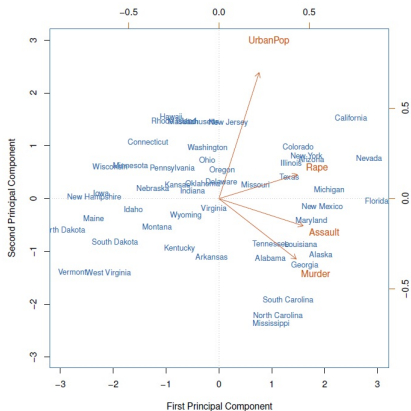


	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8

Figura 2: Base de datos USArrest (ISLR).

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

**Cuadro 1:** Los vectores de pesos de las componentes principales  $\phi_1$  y  $\phi_2$ , para los datos USArrests.



**Figura 3:** Las 2 primeras componentes principales para los datos USAarrests, esta figura se denomina biplot ilustra tanto los scores sobre las 2 primeras componentes principales como los pesos sobre dichas componentes.

## Otra interpretación de las componentes principales

- Los primeras  $M$  direcciones definidas por los vectores de pesos sobre las componentes principales proporcionan la mejor aproximación  $M$  dimensional (en términos de distancia Euclídea) para el conjunto de observaciones  $x_{ij}$ , de forma tal que

$$x_{ij} \approx \sum_{m=1}^M z_{im} \phi_{jm}, \quad (11)$$

(asumiendo que la matriz de datos original  $\mathbf{X}$  posee las columnas centradas).

## Otra interpretación de las componentes principales

- Los primeras  $M$  direcciones definidas por los vectores de pesos sobre las componentes principales proporcionan la mejor aproximación  $M$  dimensional (en términos de distancia Euclídea) para el conjunto de observaciones  $x_{ij}$ , de forma tal que

$$x_{ij} \approx \sum_{m=1}^M z_{im} \phi_{jm}, \quad (11)$$

(asumiendo que la matriz de datos original  $\mathbf{X}$  posee las columnas centradas).

- Cuando  $M = \min(n - 1, p)$ , entonces la representación es exacta

$$x_{ij} = \sum_{m=1}^M z_{im} \phi_{jm}, \quad (12)$$

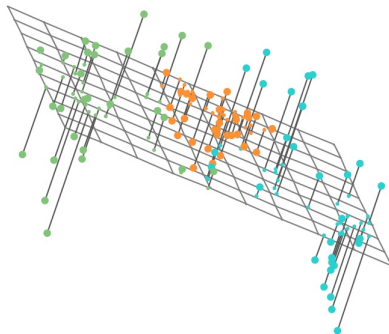


Figura 4

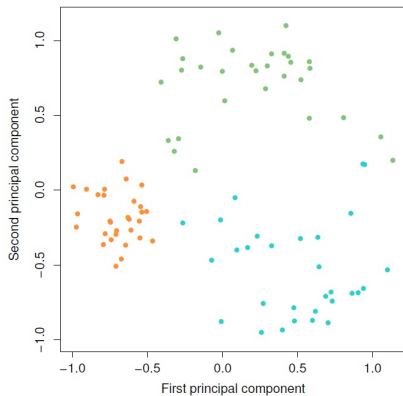
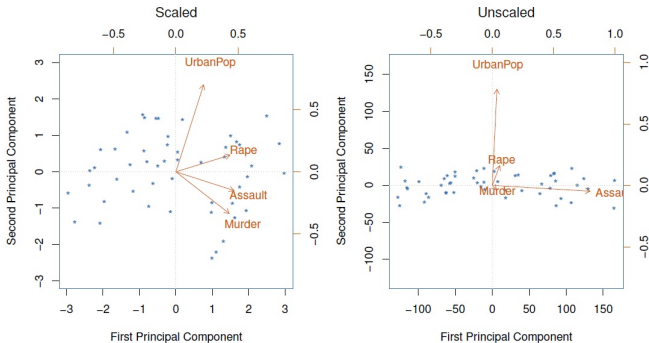


Figura 5

# Escalado de las variables en PCA



**Figura 6:** En general se recomienda estandarizar las variables de forma que todas tengan, desviación estándar igual a 1.



# Proporción de varianza explicada

- Cuanta de la información en el conjunto de datos se pierde por proyectar las observaciones solo sobre las primeras componentes principales?.

# Proporción de varianza explicada

- Cuanta de la información en el conjunto de datos se pierde por proyectar las observaciones solo sobre las primeras componentes principales?.
- Cuanta de la varianza presente en los datos no está contenida en las primeras componentes principales?

# Proporción de varianza explicada

- Cuanta de la información en el conjunto de datos se pierde por proyectar las observaciones solo sobre las primeras componentes principales?.
- Cuanta de la varianza presente en los datos no está contenida en las primeras componentes principales?
- De forma más general estamos interesados en la *proporción de varianza explicada* (PVE) por cada componente principal.

# Proporción de varianza explicada

- Cuanta de la información en el conjunto de datos se pierde por proyectar las observaciones solo sobre las primeras componentes principales?.
- Cuanta de la varianza presente en los datos no está contenida en las primeras componentes principales?
- De forma más general estamos interesados en la *proporción de varianza explicada* (PVE) por cada componente principal.
- La **varianza total** presente en un conjunto de datos (suponiendo que las variables están centradas para tener una media de cero) está definida por

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2, \quad (13)$$

- En su lugar la varianza explicada por la  $m$ -ésimo componente principal es

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} x_{ij} \right)^2, \quad (14)$$

y por lo tanto, la PVE por la  $m$ -ésima componente principal es

$$\frac{\sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2} \quad (15)$$

- En total hay  $M = \min(n - 1, p)$  componentes principales y la PVE total suma uno.

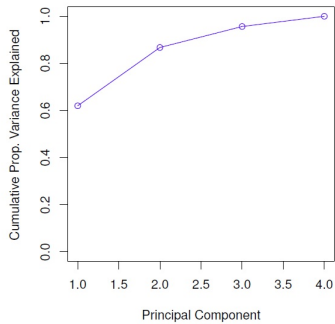
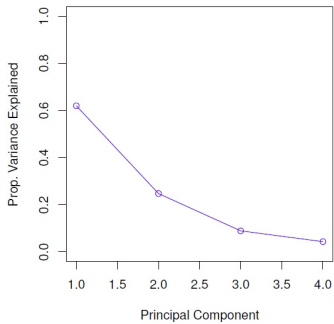


Figura 7: .