Aprendizaje No Supervisado

El aprendizaje no supervisado es un conjunto de herramientas estadísticas destinadas al análisis descriptivo para las que solo contamos con un conjunto de características $X_1, X_2, \ldots, X_p$ medidas en n observaciones. No estamos interesados en la predicción, **porque no tenemos una variable de respuesta asociada *Y***, sino que buscamos responder preguntas del tipo:

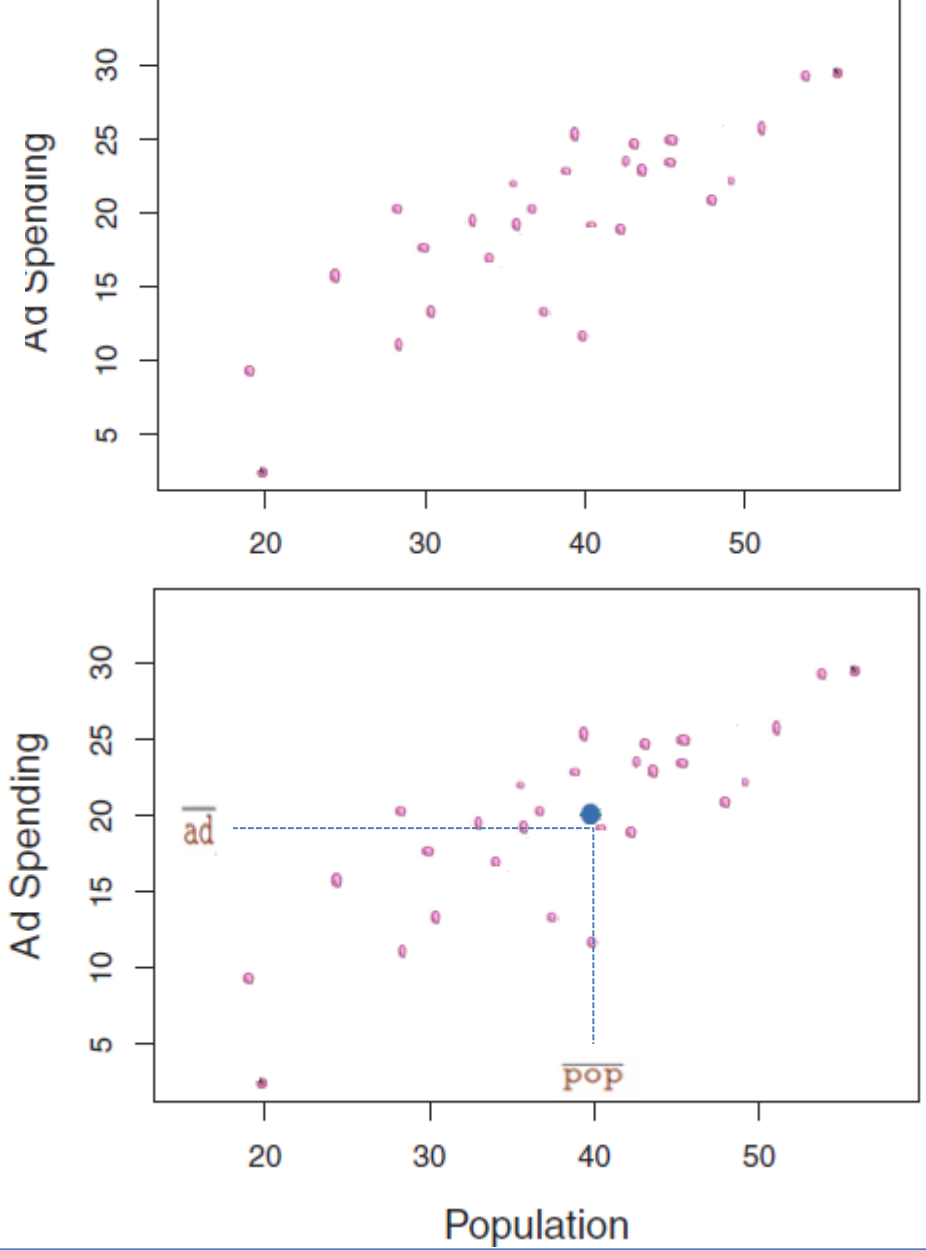
¿Hay alguna forma informativa de **visualizar** los datos?
¿Podemos armar subgrupos entre las **variables?**
¿Podemos armar subgrupos entre las **observaciones**?

Dificultades

No existe un mecanismo universalmente aceptado para realizar la **validación** cruzada o validar resultados en un conjunto de datos independientes. La razón de esta diferencia con los vistos regresión y clasificación es simple: si ajustamos un modelo predictivo utilizando una técnica de aprendizaje supervisado, entonces es posible verificar nuestro trabajo al ver qué tan bien nuestro modelo predice la respuesta *Y* en las observaciones que no se usan para ajustar el modelo. Sin embargo, en el aprendizaje no supervisado, no hay manera de verificar nuestro trabajo porque no sabemos la verdadera respuesta.

Análisis de Componentes Principales

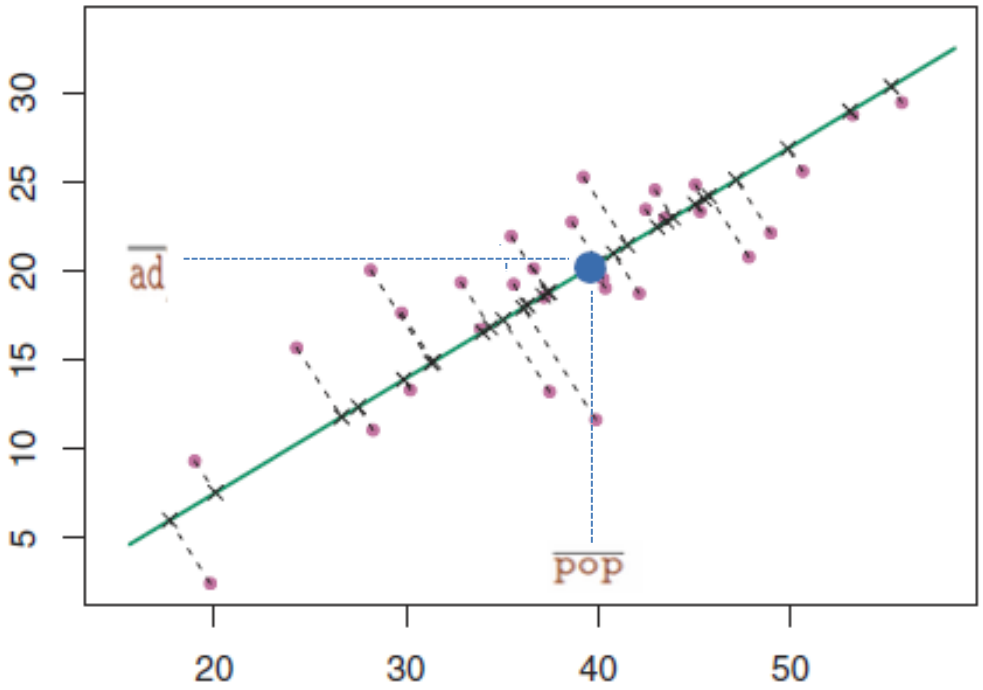
Supongamos que como parte de un **análisis exploratorio de datos (**EDA por sus siglas enn inglés) deseamos visualizar *n* observaciones, cuyo vector de características $X_1, X_2, \ldots, X_p$ tiene *p* dimensiones. Podríamos hacerlo examinando diagramas de dispersión bidimensionales de los datos, cada uno de los cuales contenga dos de las *p* características. Si *p* es grande, entonces ciertamente no será posible explorar todas las combinaciones; además, **lo más probable es que ninguno de estos diagramas sea informativo, ya que cada uno contiene solo una pequeña fracción de la información total presente en el conjunto de datos**. Por ello es sería de mucha utilidad contar con una representación de baja dimensión de los datos (por ejemplo bidimensional) que capture la mayor cantidad de información posible.



$\overline{ad}$ Promedio de Publicidad

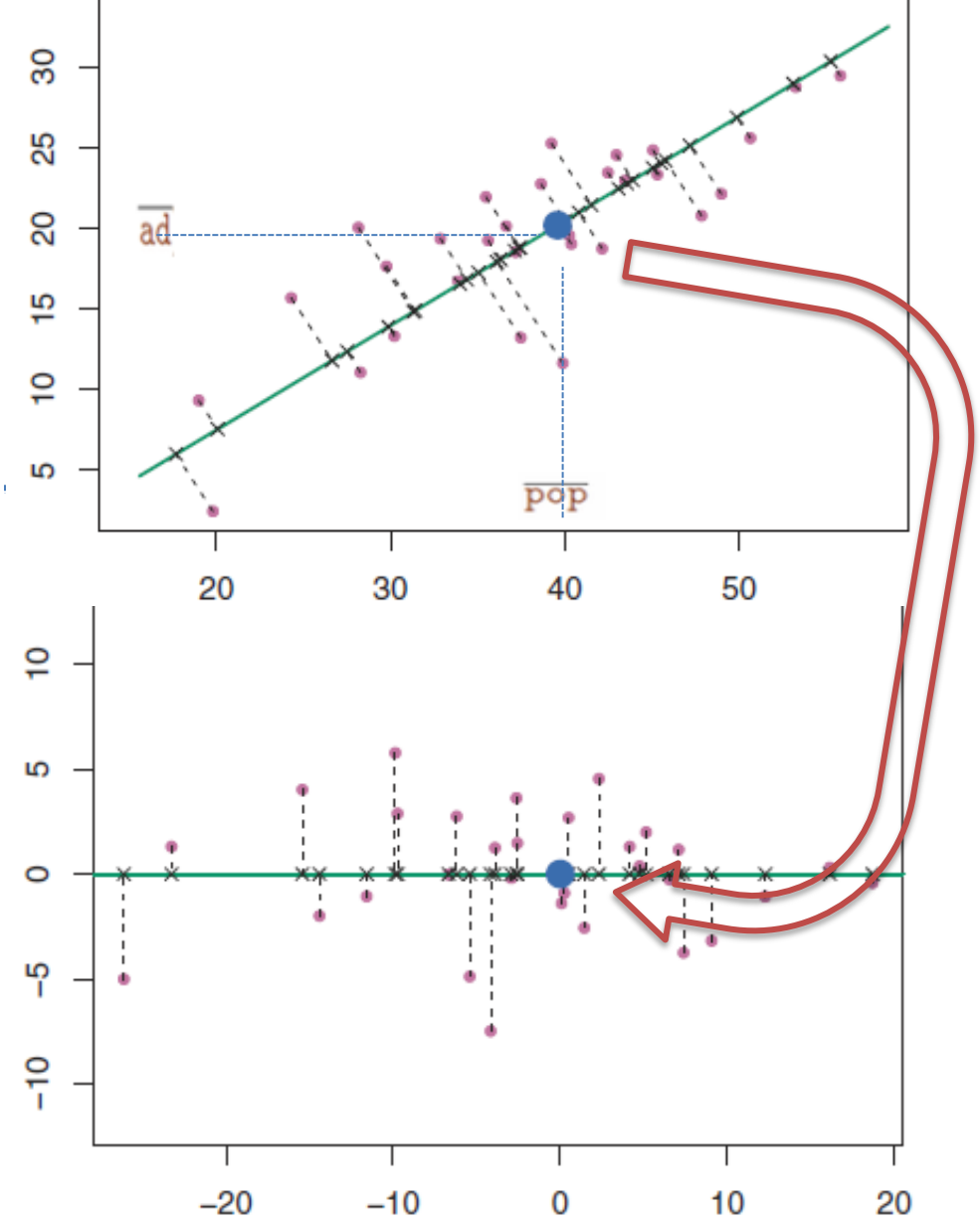
$\overline{pop}$ Promedio de Población

La primer componente principal de un conjunto de características $X_1, X_2, \ldots, X_p$ es la **combinación lineal normalizada** de las características que tenga la **mayor varianza**. En el ejemplo tenemos los atributos *tamaño de la población* y *gasto en publicidad* para 100 ciudades. La línea continua verde representa la dirección de la primera componente principal de los datos**. Podemos ver a simple vista que esta es la dirección a lo largo de la cual existe la mayor variabilidad en los datos.** Es decir, si *proyectamos* las 100 observaciones sobre esta línea las proyecciones resultantes tendrán la mayor varianza posible. Proyectar un punto en una línea simplemente implica encontrar la ubicación en la línea más cercana al punto (de forma perpendicular)



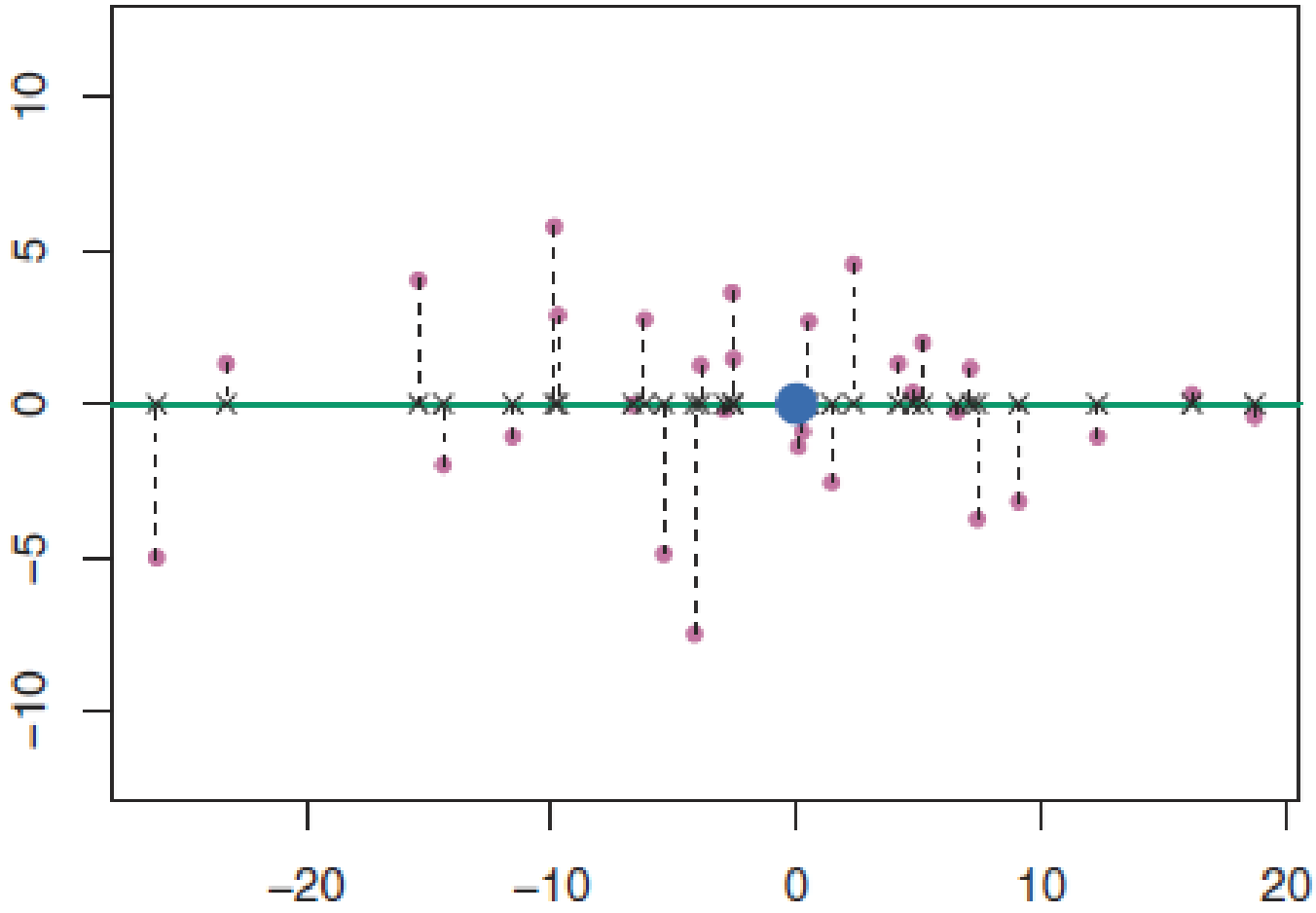
$\overline{ad}$ Promedio de Publicidad

$\overline{pop}$ Promedio de Población



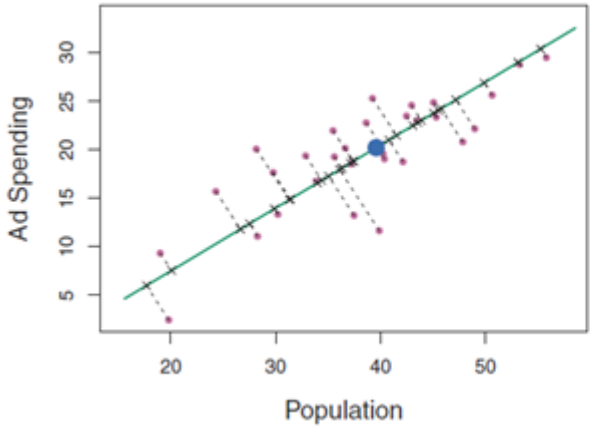
Paso el pto $(\overline{ad}, \overline{pop})$ a ser el centro (0,0)

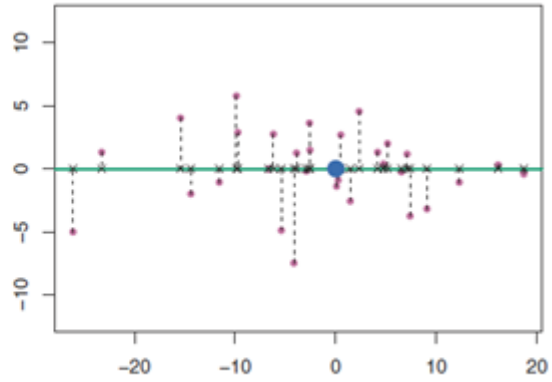
El eje *x* es la Primera Componente principal (PC_1)



Primera Comp. Ppal.

$$Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}})$$

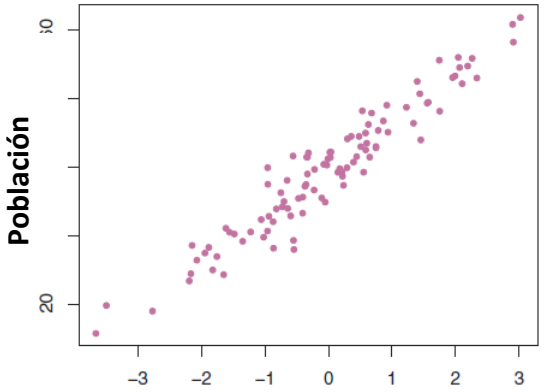




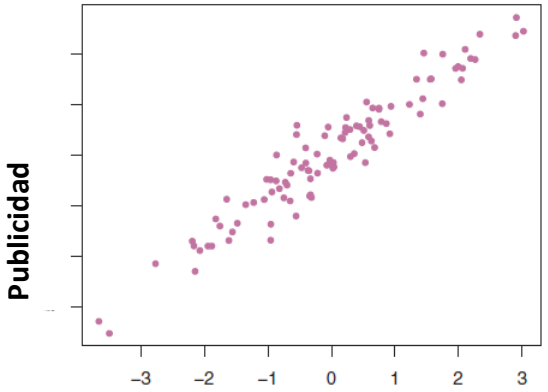
Primera Comp. Ppal.

Podemos pensar en los valores de la componente principal Z_1 como **un resumen** de las dos variables (pop y presup) en un sólo número para cada observación !!

La primer componente principal parece **capturar la mayor parte de la información (VARIABILIDAD) contenida** en los predictores pop y publicidad

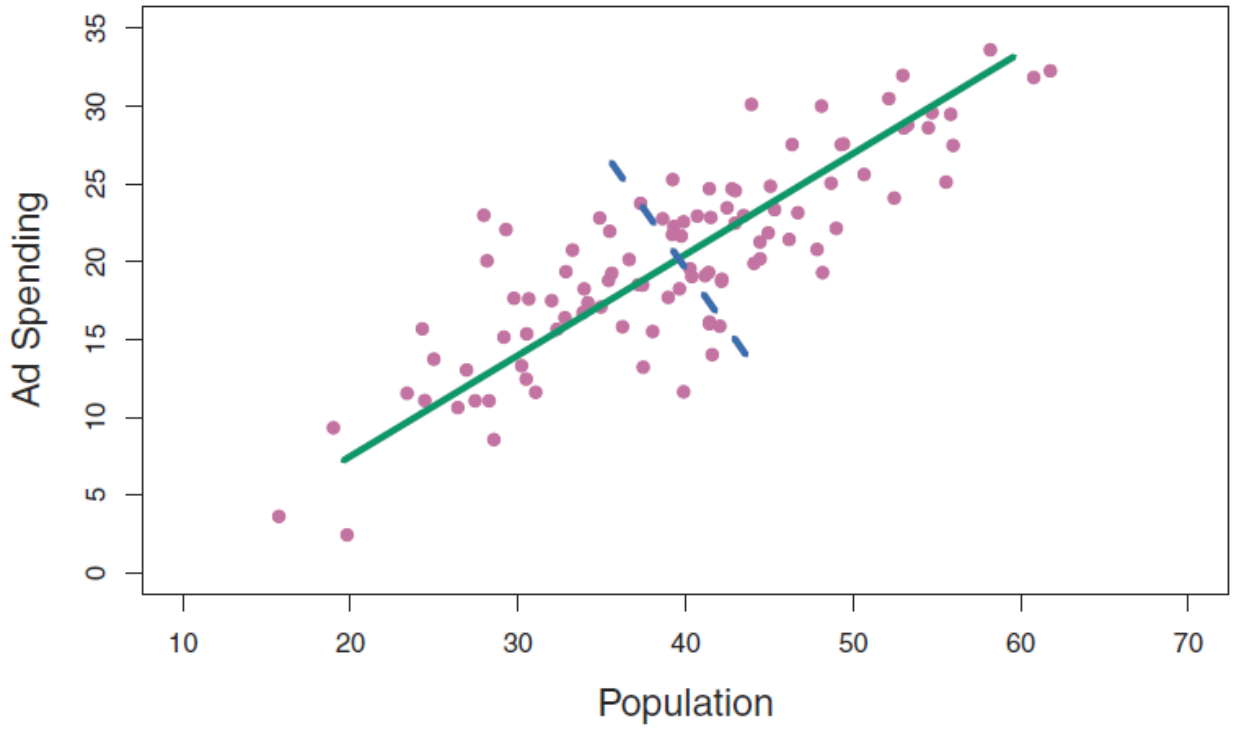


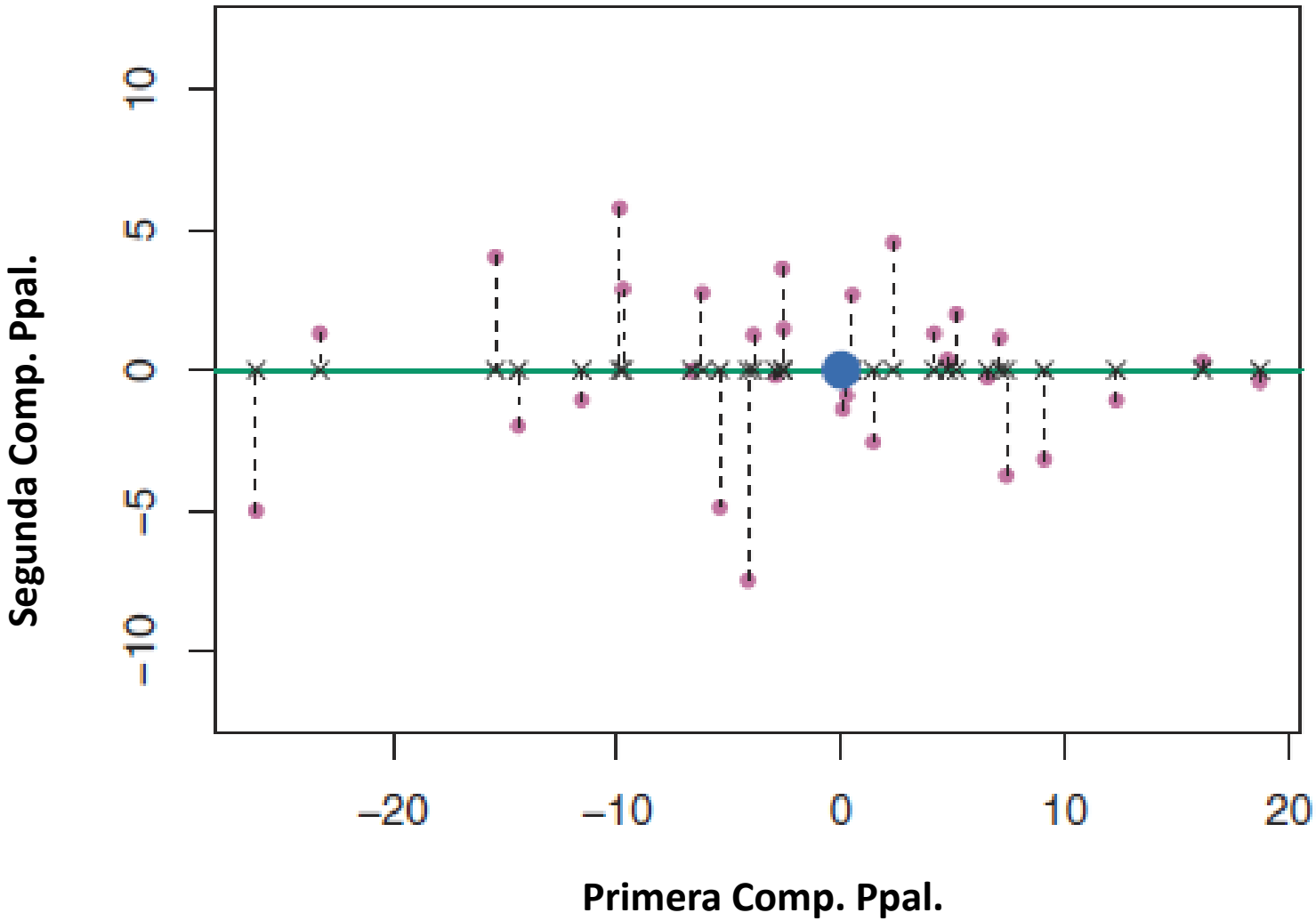
Primera Comp. Ppal.



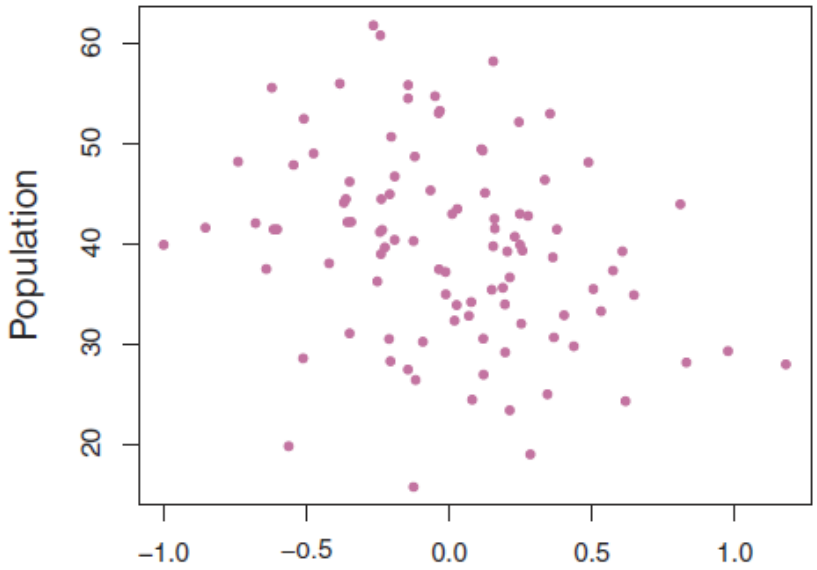
Primera Comp. Ppal.

Hasta ahora nos hemos concentrado en la primer componente principal. En general, uno puede construir hasta *p* componentes principales distintas. La segunda componente principal Z_2 será la **combinación lineal** de las variables que no esté **correlacionada** con Z_1 (o sea que sus direcciones sean *perpendiculares* u *ortogonales*) y tenga la mayor varianza sujeta a esta restricción.

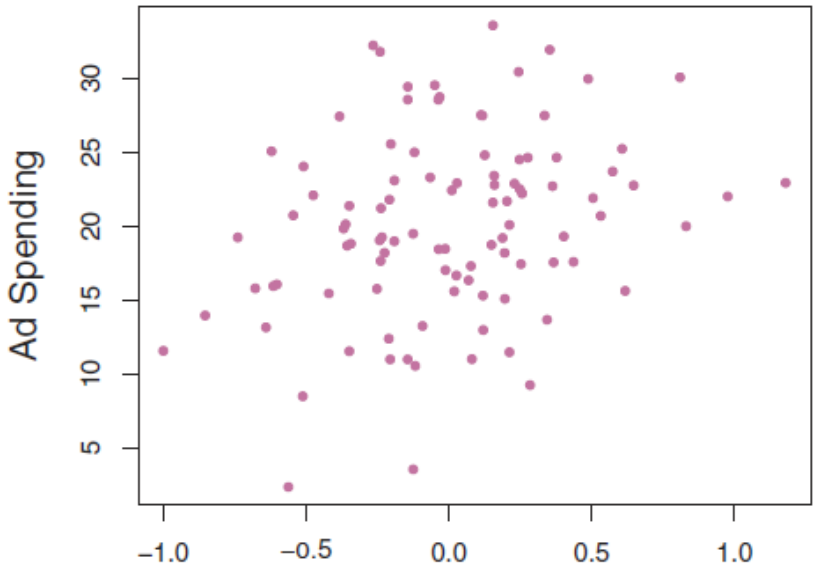



Segunda Comp. Ppal.
10
5
0
−5
−10
−20
−10
0
10
20
Primera Comp. Ppal.

Vemos que h**ay poca relación** entre la segunda componente principal y las dos variables, lo que sugiere de nuevo que, en este caso, solo se necesita la primer componente principal para representar con precisión las var *pop* y *public*.



Segunda Comp. Ppal.



Segunda Comp. Ppal.

Determinación de los Componentes Principales

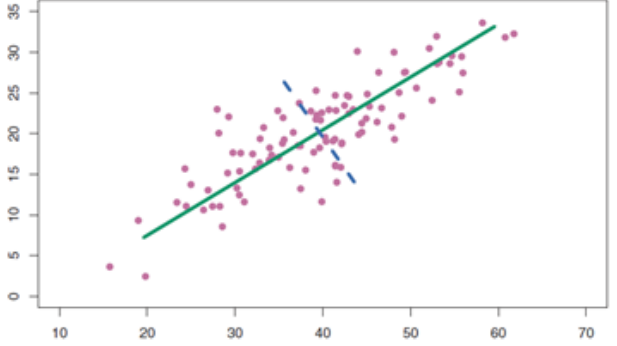
Sabemos que cada una de las *n* observaciones pertenece a un espacio *p*-dimensional, pero no todas estas dimensiones son igualmente *interesantes*. **PCA busca un pequeño número de dimensiones que sean lo más *interesantes* posible**, donde el concepto de *interesante* se mide por **el nivel de variabilidad de las observaciones a lo largo de cada dimensión**.
Para ello vamos a armar combinaciones lineales (CL) de las *p* variables.

La primer componente principal de un conjunto de variables $X_1, X_2, \ldots, X_p$ es la combinación lineal normalizada de ellas que tenga la **mayor varianza**.

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p \quad \text{donde} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

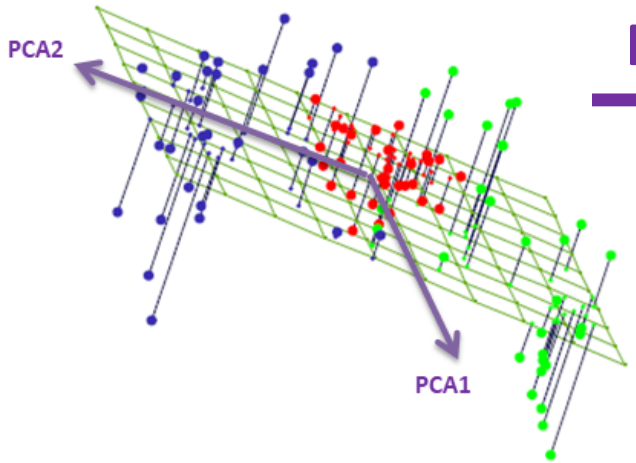
Como sólo estamos interesados en la **varianza** asumimos que cada una de las variables **x** en la matriz ***X*** se ha centrado para tener una media de cero (es decir, las medias de las columnas de ***X*** son cero). Luego buscamos la CL ***z*** de las *p* carac $z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$
que evaluadas sobre las *n* observaciones tenga la mayor varianza muestral. Como hicimos que cada una de las *p* variables tenga media=0 (cada una de las columnas de la matriz de observaciones) entonces deberíamos encontrar los coeficientes $\phi_{11}, \ldots, \phi_{p1}$ que maximicen la siguiente fórmula $\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$ sujeta a la restricción $\sum_{j=1}^{p}\phi_{j1}^2 = 1$ **(1)**

La segunda componente principal será la combinación lineal de $X_1, X_2, \ldots, X_p$ que maximice la varianza de combinaciones lineales Z_2 que no están correlacionadas con Z_1. Para encontrar $\boldsymbol{\phi}_2$, resolvemos un problema similar a (1) con $\boldsymbol{\phi}_2$ reemplazando $\boldsymbol{\phi}_1$, y con la restricción adicional de que $\boldsymbol{\phi}_2$ sea ortogonal a $\boldsymbol{\phi}_1$



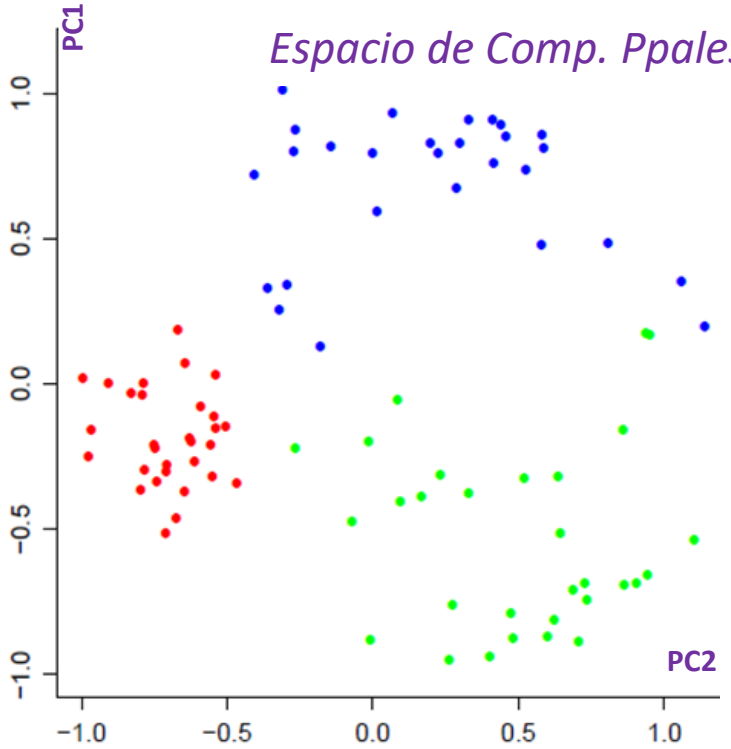
Importante: en un conjunto de datos más grande con p> 2 variables, habrá varios componentes principales distintos y se definirán de manera similar.

Espacio de Atributos

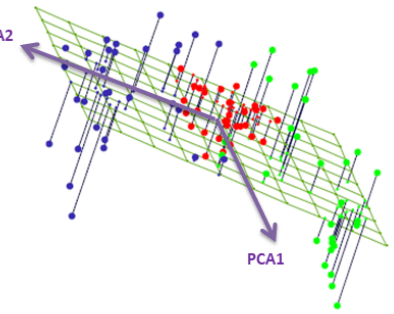


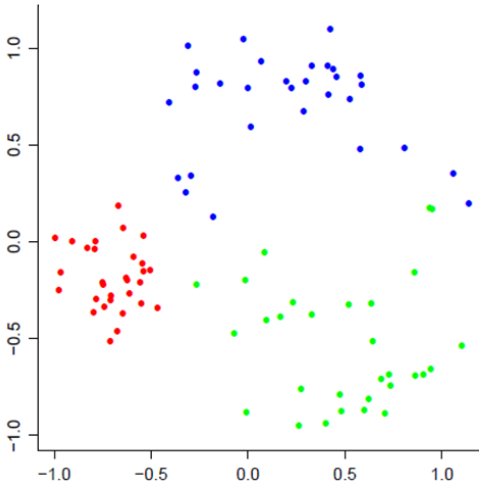
PCA

Espacio de Comp. Ppales.



Proporción de la Varianza Explicada





Vemos que esta representación bi-dimensional de los datos tridimensionales captura con éxito los patrones. **Las observaciones cercanas entre sí en el espacio tridimensional** (espacio original de atributos) **permanecen cercanas en la representación bidimensional** (espacio de *Componentes Principales*).

Ahora podemos hacer una simple pregunta **¿qué parte de la variación en los datos no está contenida en las primeras componentes principales?** Estamos interesados en conocer la proporción de varianza explicada (PVE) por cada componente principal. La varianza total presente en un conjunto de datos (asumiendo que las variables se han centrado para tener una media de cero) se define como:

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \qquad \textbf{(2)}$$

y la varianza explicada por la componente ppal. *m*

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2 \qquad \textbf{(3)}$$

Por lo tanto, el Porcentaje de Varianza Explicada (**PVE**) de la componente principal *m* está dado por:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2} \quad \textbf{(4)}$$

El PVE de cada componente principal es una cantidad positiva. Para calcular el PVE acumulado de las primeras componentes principales *M*, podemos simplemente sumar los PVE de cada una de ellos (fórmula 4) . En total, hay mín(*n* - 1, *p*), y sus PVE suman uno.

Determinación de la cantidad de Componentes Principales a utilizar

En general, una matriz de datos *X de n × p* tiene min*(n - 1, p)* componentes principales. Elegiremos la menor cantidad de componentes principales que se requieran para explicar la mayor cantidad posible de la variación de los datos. Para ello utilizaremos un Gráfico de **sedimentación** (*Scree Plot*), que nos muestra el PVE (der: PVE acumulado). La cant de CP se ubica el **codo** (*elbow*)