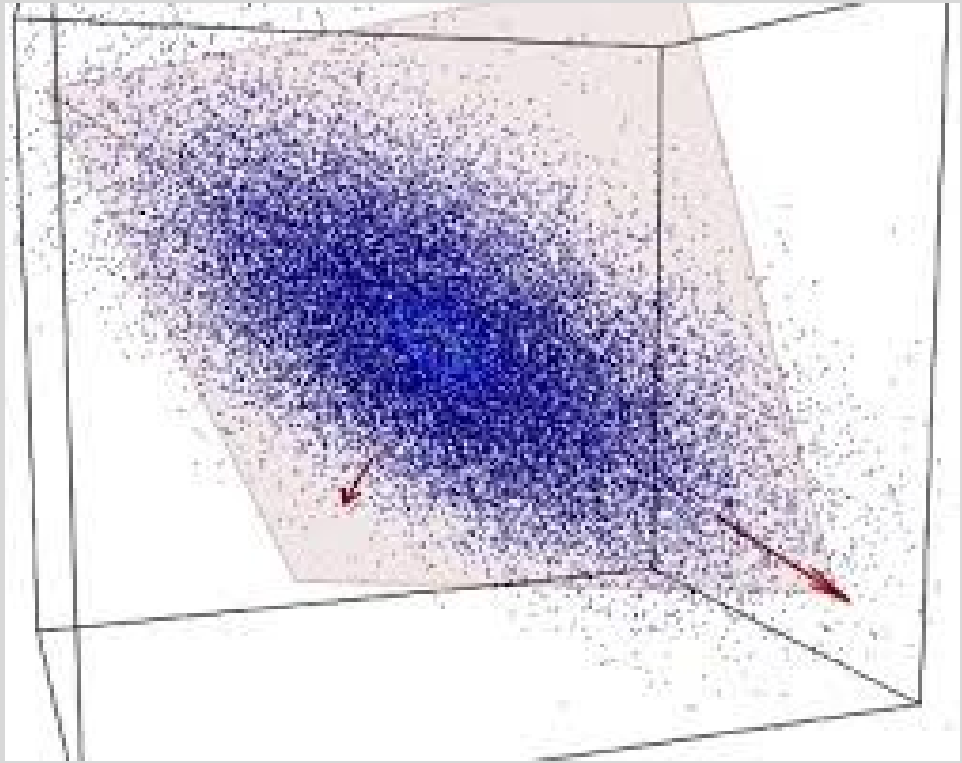


DigitalHouse > Coding School

DATA SCIENCE

Unidad: 4
Modulo: 7

Análisis de Componentes Principales

2017
Buenos Aires

OBJETIVOS DE LA CLASE

1

SOBRE LOS DATOS

Analizar formato y preprocesamiento para aplicar el algoritmo de PCA

2

SOBRE LA TÉCNICA

Entender cómo se calculan y qué problema resuelven los componentes principales, interpretación geométrica.

3

SOBRE LOS RESULTADOS

Interpretar los resultados del algoritmo y determinar la cantidad óptima de componentes.

4

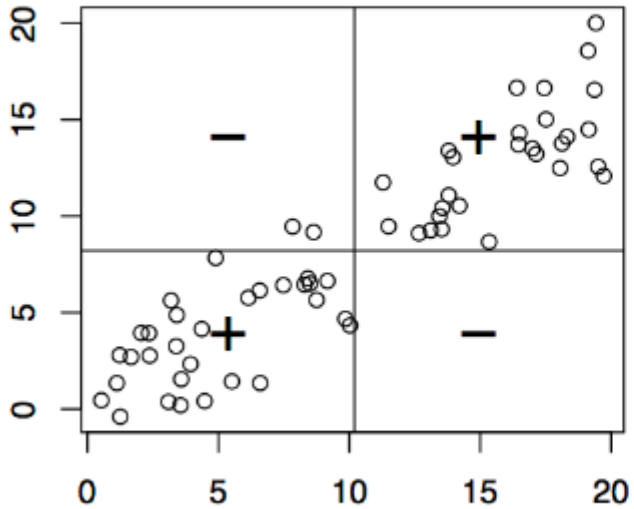
APLICACIONES

Conocer y experimentar con las distintas aplicaciones prácticas de componentes principales

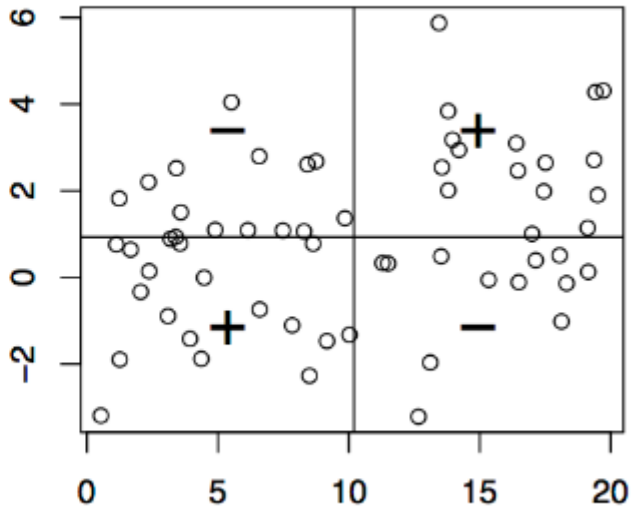
ANÁLISIS DE COMPONENTES PRINCIPALES

- Decimos que dos variables X e Y, tienen covarianza positiva cuando se encuentran por encima de su media al mismo tiempo y tienen covarianza negativa cuando al mismo tiempo, una está por debajo y otra por encima.

Covarianza positiva



Covarianza cercana a cero.



La covarianza se mide como:

$$Cov_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)}$$

La covarianza de un conjunto de datos con p variables se puede representar con una matriz de p x p llamada **matriz de varianzas y covarianzas**:

	^GSPC	^IXIC	XOM	C	GE	MSFT	K	GM
^GSPC	0.633	0.929	0.505	0.495	0.448	0.258	0.261	1.226
^IXIC	0.929	1.737	0.340	0.584	0.507	0.482	0.211	1.842
XOM	0.505	0.340	3.253	-0.421	-0.017	0.268	0.318	2.197
C	0.495	0.584	-0.421	1.923	0.688	0.176	0.277	-0.242
GE	0.448	0.507	-0.017	0.688	1.834	0.761	0.232	0.049
MSFT	0.258	0.482	0.268	0.176	0.761	1.945	0.181	1.315
K	0.261	0.211	0.318	0.277	0.232	0.181	1.045	0.688
GM	1.226	1.842	2.197	-0.242	0.049	1.315	0.688	9.429

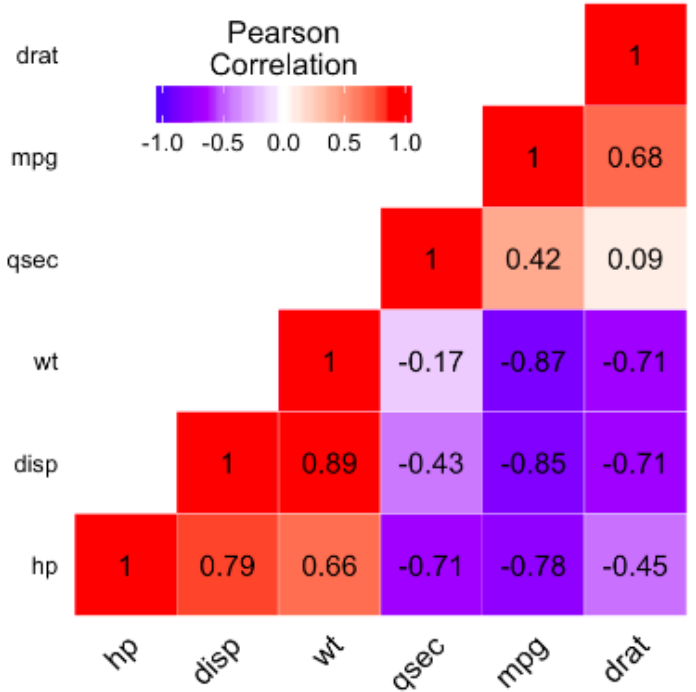
* En la diagonal se encuentra la varianza de cada feature

* En el resto de la matriz se encuentran las covarianzas

— La correlación es una versión estandarizada (dividida por los desvíos estándar) de la covarianza:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

* La correlación está acotada entre 1 y -1.
* Siempre que la covarianza es positiva, la correlación es positiva y viceversa.
* Mientras que la correlación no tiene unidades físicas, la covarianza sí.



SI TENGO VARIABLES CORRELACIONADAS Y QUIERO REDUCIR LA DIMENSIÓN DE MIS DATOS
¿QUÉ PODRÍA HACER?

Repaso: combinaciones lineales.

Una combinación lineal de dos o más columnas es la suma de sus componentes multiplicados por escalares.
Por ejemplo, los datos en negro representan los tiempos que tardaron 8 corredores en recorrer cada tramos de una carrera

corredor	km4	km8	km12	km16	w1	w2	w3
1	10	10	13	12	10	12,5	-2,5
2	12	12	14	15	12	14,5	-2,5
3	11	10	14	13	10,5	13,5	-3
4	9	9	11	11	9	11	-2
5	8	8	9	8	8	8,5	-0,5
6	8	9	10	9	8,5	9,5	-1
7	10	10	8	9	10	8,5	1,5
8	11	12	10	9	11,5	9,5	2

En violeta se encuentran tres posibles **combinaciones lineales** de los datos:

- w1 = ½ * km4 + ½* km8
- w2 = ½ * km12 + ½ * km16
- w3 = ½ * km4 + ½ * km8 - ½ * km12 +½ * km16

¿Qué son los componentes principales?

- Son combinaciones lineales de las variables originales
- Específicamente, son las combinaciones que tienen la **máxima varianza posible** y al mismo tiempo **no están correlacionados** entre sí.
- El coeficiente de correlación entre los distintos componentes es 0.
- Pueden existir tantos componentes principales como variables en el dataset.
- Los pesos de cada variable en la nueva componente se llaman “loadings”.

En el ejemplo anterior:

- w3 = ½ * km4 + ½* km8 - ½* km12 - ½* km16

Si esta combinación lineal fuera un componente principal, los loadings serían:

[½,½,-½,-½]

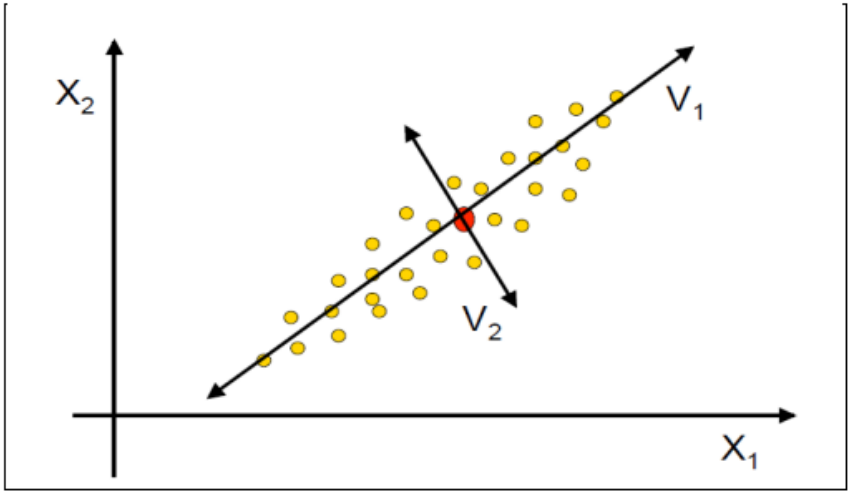
¿Cómo se eligen los loadings?

Tenemos UN objetivo y DOS restricciones:

- **OBJETIVO**: que la nueva variable (**la componente principal**, combinación lineal de las otras variables) tenga la máxima varianza posible a lo largo del dataset.

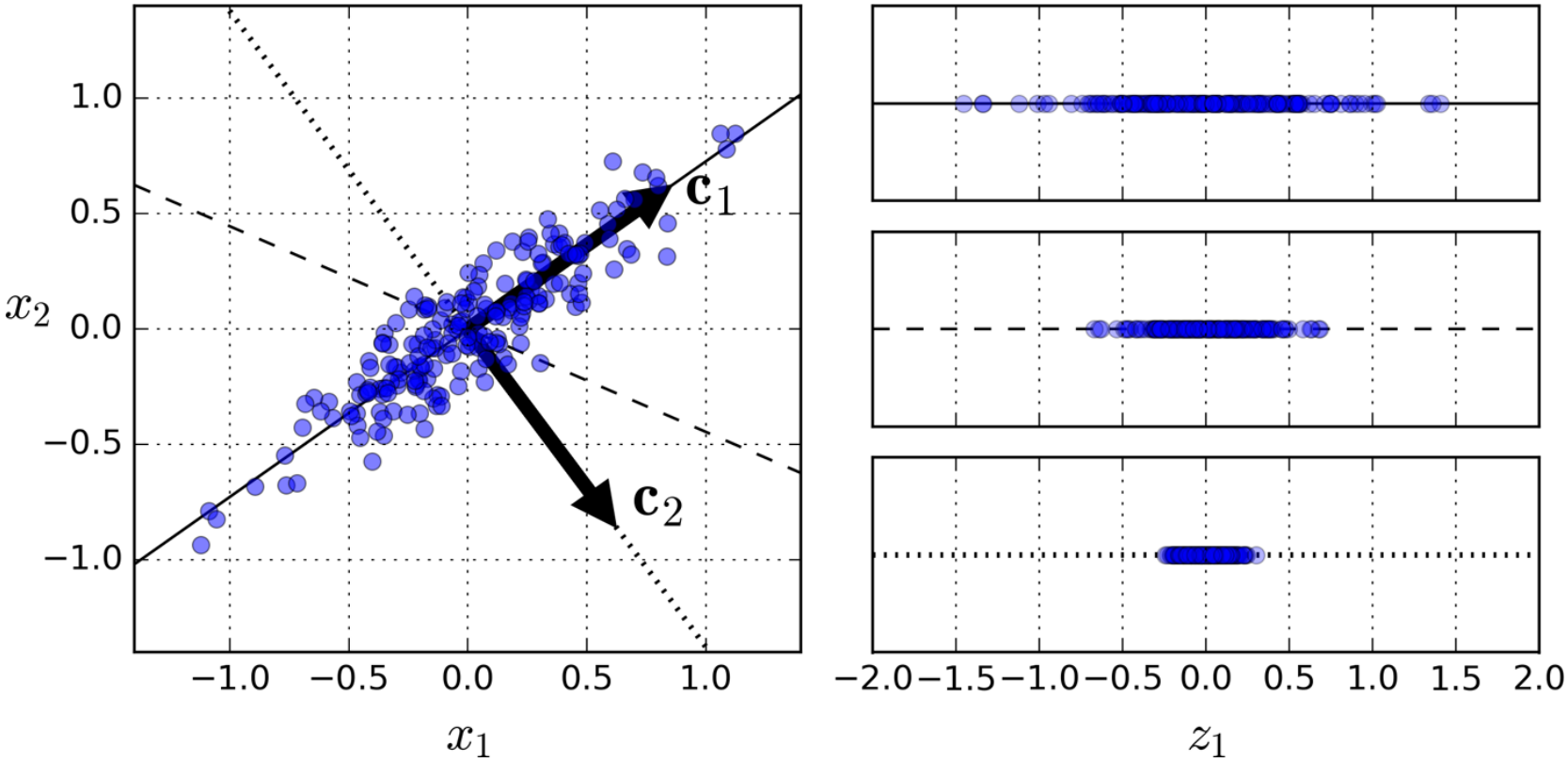
- **RESTRICCIÓN 1**: Que la suma de los loadings al cuadrado sea 1. *¿Por qué?*

- **RESTRICCIÓN 2**: Que cada componente sea "*no correlacionado*" de todas las demás, es decir, ortogonales y de covarianza 0. *En 2 dimensiones ¿Cuántas formas tengo de elegir la dirección del segundo una vez que elegí la del primero?*

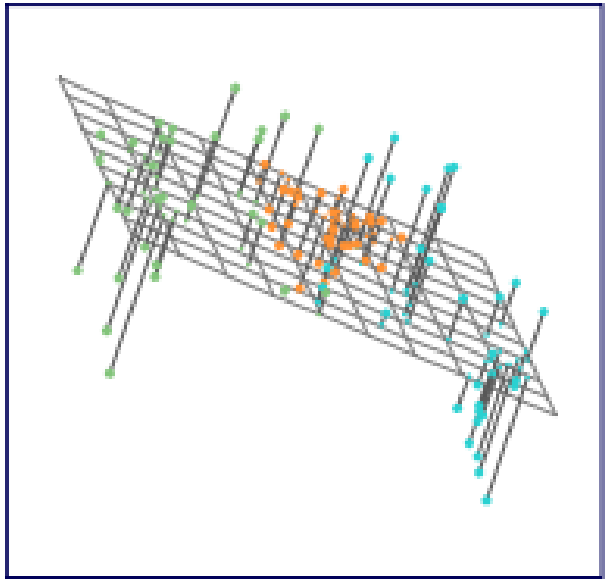


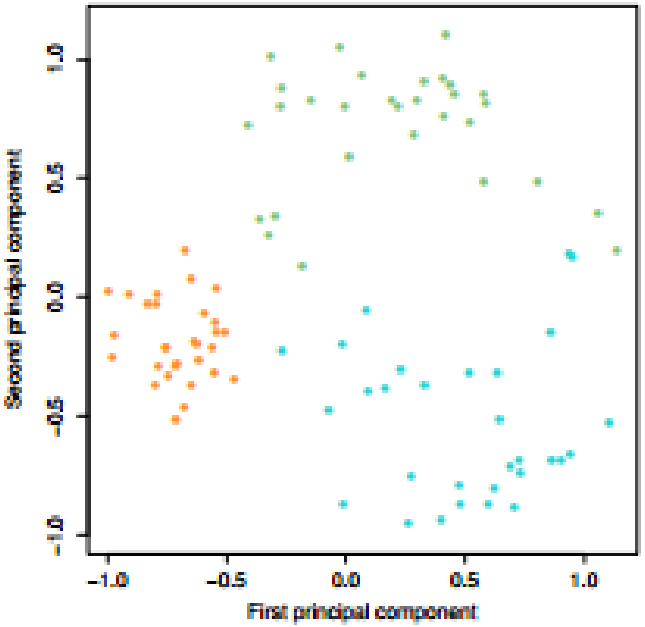
- V1 y V2 son ortogonales
- V1, es el vector que maximiza la varianza y queda una sola forma de elegir la dirección de V2 (ortogonal)
- Si dos puntos se encuentran alejados en el plano X1,X2 ¿Cuán alejados están en la proyección sobre V1? ¿Y en la proyección sobre v2?

Representación geométrica: tres representaciones alternativas



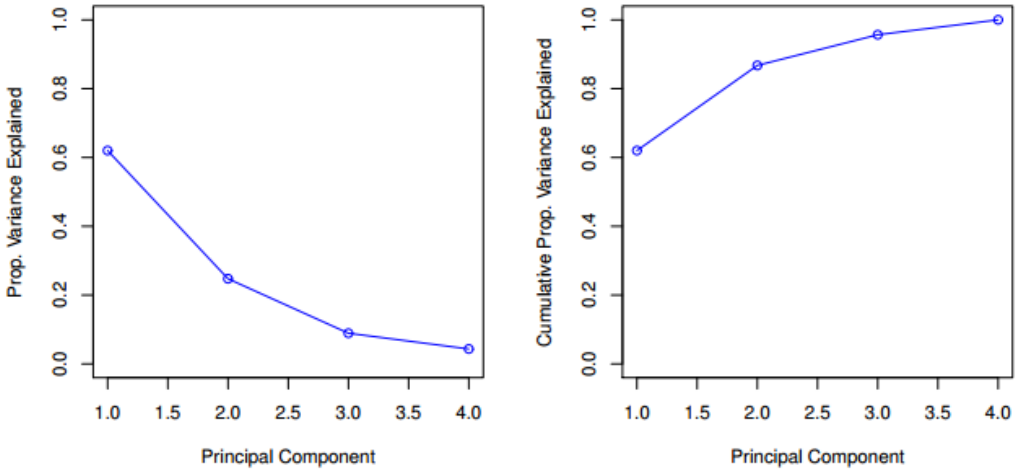
PCA en 3 dimensiones





El objetivo de PCA es buscar el plano más cercano a los datos...
¿Cuál es la diferencia con regresión lineal?

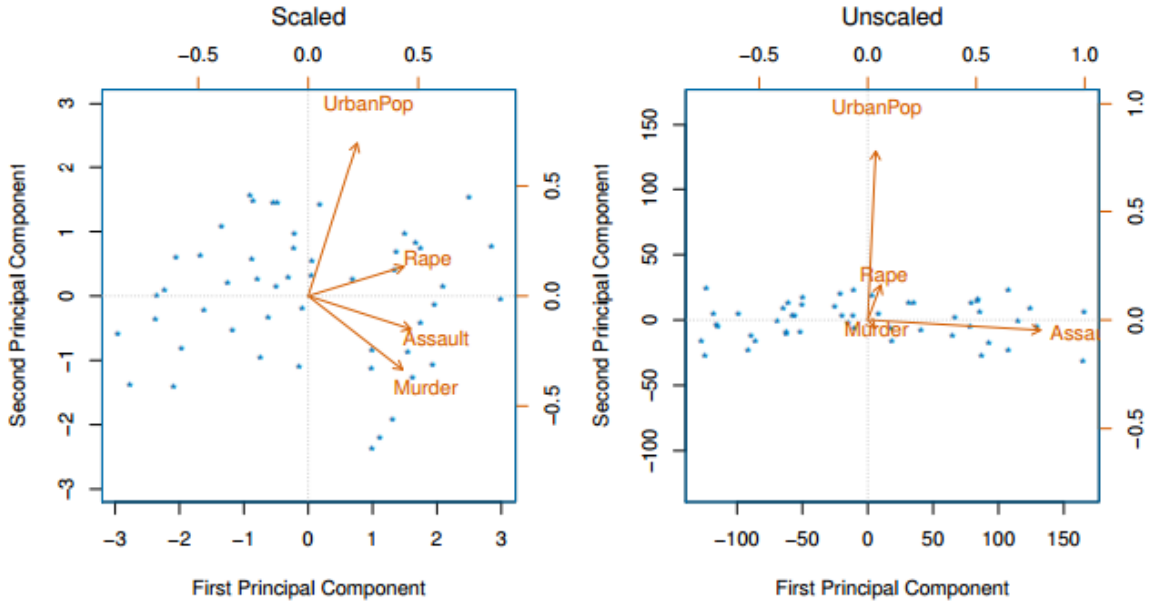
A medida que voy generando componentes principales, **cada vez tengo menos libertad** para elegir el siguiente porque el coeficiente de correlación entre todos tiene que ser 0.
Si se mide la **suma de las varianzas originales** (estandarizadas) y la **suma de la varianza de los CP** a medida que agrego nuevos, la varianza explicada es cada vez menor.



Se observa cómo la proporción de varianza explicada aumenta pero cada vez menos al agregar componentes.

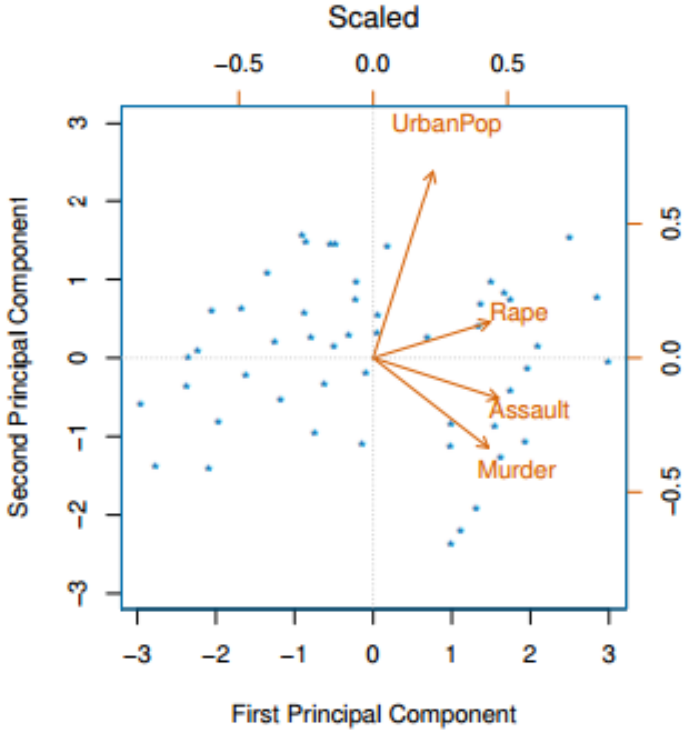
PCA se basa en la matriz de correlaciones. Por lo tanto sólo tiene sentido cuando trabajamos con **variables numéricas.** Existen otras técnicas para trabajar con datasets de variables categóricas o mixtas: análisis factorial de correspondencias y categorical PCA.

Las variables en PCA son **sensibles a la escala y deben estar estandarizadas para que todas tengan la misma varianza.** Si una variable está en una escala que le da mayor representación que a las demás puede dominar la formación de los componentes principales.



Los componentes principales, pueden ser la expresión de **variables latentes**.

Por ejemplo: en este caso una ciudad con alto valor en la primera componente es una ciudad “insegura” y una con alto valor en la segunda componente es una ciudad “grande”.



- PCA es un problema de aprendizaje NO SUPERVISADO, a diferencia de la regresión lineal ninguna variable juega el rol de target.
- Permite reducir la dimensionalidad de los datos descartando información redundante o ruido
- Nos permite lograr una representación gráfica de la información multidimensional
- Las unidades pueden influir en la variabilidad de los componentes principales. Si las escalas o las varianzas son distintas entre las X, hay que estandarizar las variables.
- Si nos encontramos con datasets demasiado grandes podemos usar la clase **IncrementalPCA()** de sklearn que va alimentando el algoritmo por mini-batches
- Hay que establecer un método para determinar la cantidad de componentes.
    - Cuando se usan como una herramienta previa de aprendizaje supervisado se puede usar cross validation.
    - Cuando se usan en aprendizaje no supervisado se puede mirar la curva de varianza explicada