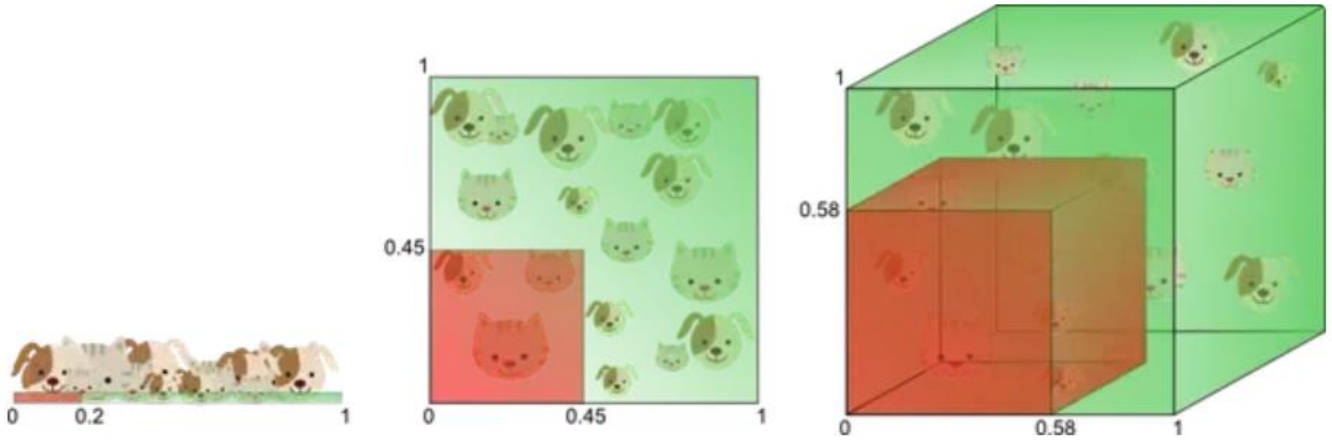
Metodología no supervisada aplicada

Problemas de alta dimensión

Cuando la cantidad de variables a analizar es tanta, que el exceso de información nos impide extraer conocimiento útil podemos decir que estamos frente a un problema de alta dimensionalidad.

Un término popularizado últimamente para referirse a este problema es la llamada *curse of dimensionality*, la cual hace referencia a que cuando podemos pensar que ganamos al tener más variables, estas impiden extraer conocimiento útil y además estadísticos como el promedio pierden capacidad de represental la población



Problemas del alta dimensión

La solución más evidente para este problema es la reducción de variables. Tal como vimos en clases anteriores, se pueden ir eliminando variables mediante IV, IG, selección secuencial o regresión penalizada (LASSO), sin embargo esto sólo tiene sentido en metodologías supervisadas.

Al no saber la importancia de una variable respecto a nuestro target, buscamos **reducir variables con la menor pérdida posible de información**. Respecto a este problema, el álgebra lineal tiene mucho que aportar en esto, en particular veremos dos formas muy similares de abordar ese problema.

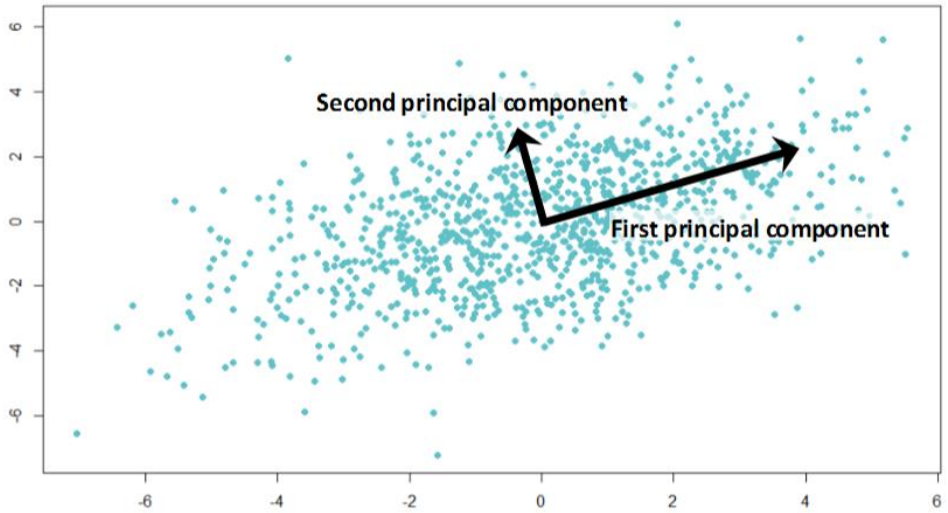
- **PCA** de sus siglas en inglés Principal Components Analysis, consiste en armar nuevas variables que sean combinaciones lineales de las originales.
- **Kernel PCA** es una modificación del PCA a fin de captar más características de la data original que la que presenta a simple vista.

Análisis de componentes principales

Supongamos que nuestras columnas de datos son $x_1, x_2, \cdots x_m$ es decir, m variables. Nuestro objetivo es crear k variables de la forma

$$y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots \alpha_{1m}x_m$$
$$y_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots \alpha_{2m}x_m$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$y_k = \alpha_{k1}x_1 + \alpha_{k2}x_2 + \cdots \alpha_{km}x_m$$

Con $k < m$
Estas nuevas variables son constructos de las iniciales. La idea es que las combinaciones, con pocas de ellas captemos la mayor información posible de la data inicial.
La intuición detrás de esto es simplemente cambiar la base del espacio vectorial que generan los datos, de tal forma que menos dimensiones capten la información



En el ejemplo de la imagen, la primera componente capta la mayor tendencia de los datos, la segunda un tanto menos. Mientras más correlacionados estén los datos más se puede reducir la dimensionalidad dada la redundancia de la información. Geométricamente hablando los valores propios generan estas bases.

Análisis de componentes principales

La obtención de las componentes se basa en las correlaciones, en efecto dada la matriz de correlaciones, se sabe que esta es simétrica y definida positiva, por lo que existe su descomposición espectral

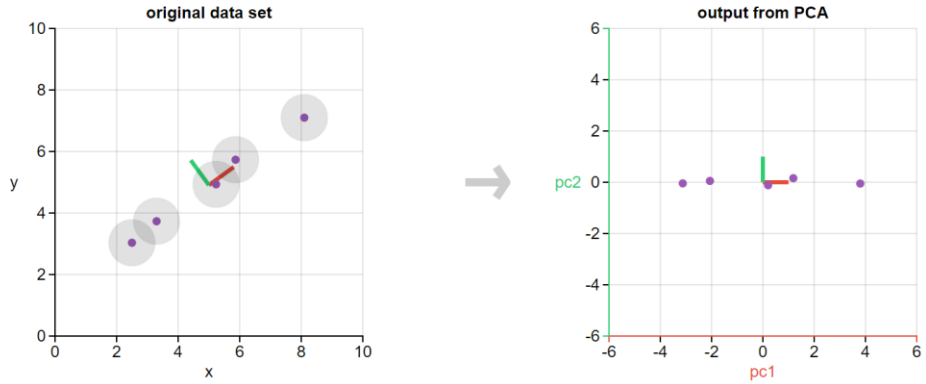
$$M = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1m} \\ \vdots & \ddots & \vdots \\ \rho_{m1} & \cdots & \rho_{mm} \end{bmatrix}$$

$$M = Q\Sigma Q^{-1}$$

En esta descomposición Q es una matriz con los vectores propios como columnas y Σ una matriz diagonal con los valores propios en su diagonal, además se da que

$$\sum_{i=1}^{m} \lambda_i = m$$

Luego transformando usando los vectores propios se obtienen las nuevas características.

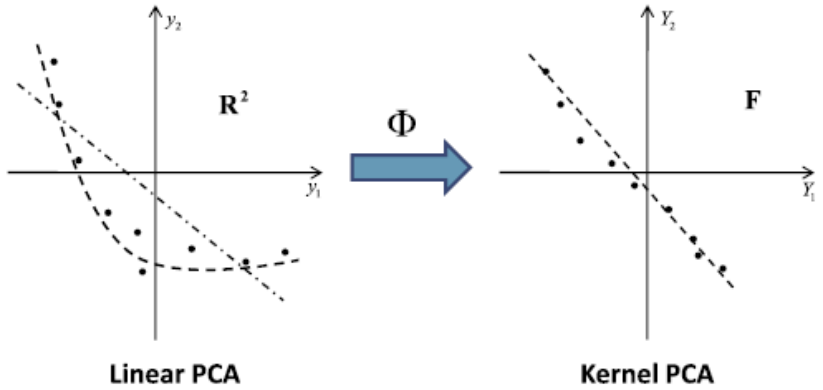


En general debiéramos obtener menos variables pero se perderá la variabilidad en un porcentaje, la cantidad recatada será.

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{m} \lambda_j}$$

Análisis de componentes principales

El PCA toma sólo combinaciones lineales de los datos y toma como supuesto que tienen ese comportamiento. ¿Qué pasa si no ocurre eso necesariamente? Una forma de abordar esto es previamente transformar los datos.



El concepto es bastante rudimentario pues es generar transformaciones de los datos y aplicar PCA. Estas transformaciones se generan mediante una **expansión kernel.** Esta expansión es un aumento de dimensionalidad de características mediante una familia de transformaciones.

Una expansión kernel es un mapeo de características basada en productos internos. Algunos ejemplos de kernels son

Lineal:

$$K(x, y) = x^T y$$

Polinomial:

$$K(x, y) = (x^T y + r)^n$$

Gaussiano (o RBF):

$$K(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}}$$

Si tenemos dos variables x1 y x2, y queremos hacer una expansión polinomial de segundo orden, obtendremos todas las combinaciones $x_1^2, x_2^2, x_1 x_2$, etc.

Sistemas de recomendación

Los sistemas de recomendación corresponden a técnicas, metodologías y algoritmos a fin de generar una sugerencia a un usuario que idealmente cumplan las siguientes características.

Relevancia es un objetivo evidente, al usuario se le debiera recomendar algo que si le interesa y no un artículo al azar

Novedad Además de ser algo que le interese al usuario, ideal que sea algo novedoso, un recomendador que siempre recomienda algo que ya consumió no cumple con esta característica

Serendipia hace referencia a la capacidad de sorprender al usuario con la recomendación. Cuando uno es capaz de sugerir algo nuevo al usuario que le termina gustando implica una ganancia aún mayor.

Variedad las recomendaciones debieran mostrarle un abanico de opciones al usuario con tal de darle capacidad de elección.

Sistemas de recomendación

Para generar las recomendaciones se deben de tener los datos en forma de matriz de ranking, de las cuales hay en dos tipos, las binarias y las no binarias

	Queso	Pan	Palta	Mayonesa
Pedro	1	0	1	1
Juan	0	0	0	1
Diego	1	1	1	0

Matriz de ranking binaria

	Dark	Modern Family	GoT	El Marginal
Pedro	3	2	5	5
Juan	4	5	1	3
Diego	1	1	1	0

Matriz de ranking no binaria

Además dependiendo del enfoque, puede ser una matriz de Usuario-Item, Item-Item o Usuario-Usuario en algunos caso. Las primeras dos son las más frecuentes.
Dentro de las metodologías más habituales para generar recomendaciones se pueden mencionar las siguientes

Popular No tiene mucha ciencia, consiste en recomendar la categoría más consumida. Siempre está presente pues al menos tendrá una tasa de efectividad más alta que otro escogido al azar. Falla en cuanto a la novedad y la serendipia.
Reglas de asociación Utilizando indicadores de asociación entre ellos se ordenan listas de recomendaciones. En general cuesta definir las reglas, se requiere muchos análisis a posteriori.
Filtros colaborativos familia de metodologías que buscan asociar un objeto o usuario con objetos o usuarios parecidos. En base a esto último están los UBCF (usuarios) y los IBCF (items)


CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
piccolo