Text as Data - Modelado de Tópicos

Ciencia de Datos y Econometría Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

El problema: descubrir de qué se habla

Supongamos que tenemos un conjunto de documentos: artículos de prensa, reseñas, o ensayos. Cada documento contiene miles de palabras, y lo que nos gustaría saber es algo tan simple como:

> *¿Cuáles son los temas principales de este conjunto de textos?*

Queremos, por ejemplo, descubrir que unos textos hablan de política, otros de deportes, y otros de economía, **sin leerlos uno a uno**.

Esto es lo que llamamos **modelado de tópicos** (topic modeling): encontrar patrones de palabras que tienden a aparecer juntas y que nos permiten describir los textos en pocas dimensiones conceptuales.

Nuestra representación: texto como datos

Sabemos que cada documento puede representarse como un vector de palabras. Si construimos una **matriz documento término (DTM)**, tenemos:

- Cada fila = un documento.
- Cada columna = una palabra.
- Cada celda = cuántas veces aparece esa palabra en ese documento (o su peso TF-IDF).

Esta matriz es **muy grande y muy poco densa** (casi todo son ceros), pero contiene toda la información sobre qué palabras co-ocurren en qué documentos.

¿Qué sería un "tópico"?

En esta representación, un **tópico** puede pensarse como un patrón recurrente de uso de palabras:

- El tema *deportes* podría tener alta frecuencia en palabras como *gol*, *equipo*, *partido*, *estadio*.
- El tema *economía* podría estar asociado a *precio*, *inflación*, *mercado*, *inversión*.

Cada documento entonces puede verse como una **combinación de tópicos**: algunos hablan mucho de uno, otros mezclan varios.

Cómo buscar esos patrones: una mirada algebraica

Una forma de pensar este problema es la siguiente:

- Tenemos una matriz enorme X (documentos $\times$ palabras).
- Queremos encontrar **combinaciones de palabras** que expliquen las **mayores diferencias entre documentos**.
- Dicho de otro modo, buscamos direcciones en el espacio de palabras donde los documentos se separen más.

Y eso, en estadística y econometría, tiene un nombre conocido: **maximizar la varianza**.

De los tópicos a la varianza: la intuición de PCA

Imaginemos los documentos como puntos en un espacio donde cada eje es una palabra. Si dos palabras suelen aparecer juntas (por ejemplo *precio* y *mercado*), los puntos estarán alineados en una dirección particular.

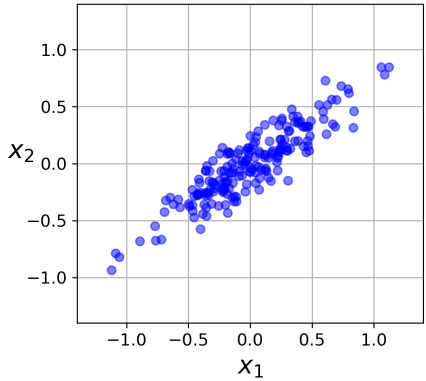
PCA busca precisamente eso:

> *La dirección en la que los documentos difieren más, es decir, donde la* ***varianza*** *es máxima.*

Esa dirección, una combinación lineal de palabras, puede interpretarse como un **tema** o **tópico latente**.

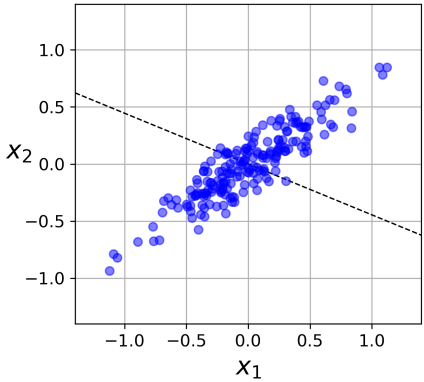
Visualizando la idea

En este gráfico, cada punto azul representa un documento descrito por dos palabras X_1 y X_2. Los documentos están fuertemente correlacionados (cuando se habla mucho de X_1, también se habla mucho de X_2):



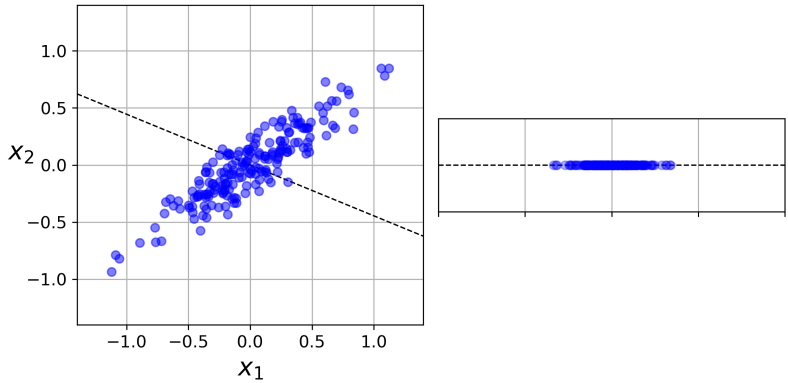
Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.



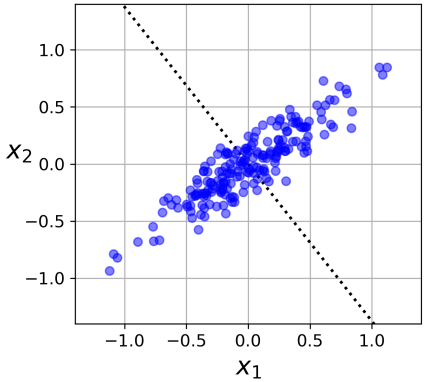
Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.



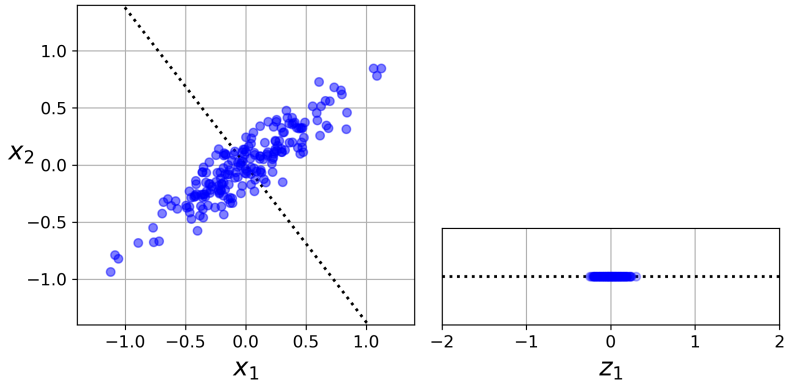
Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.



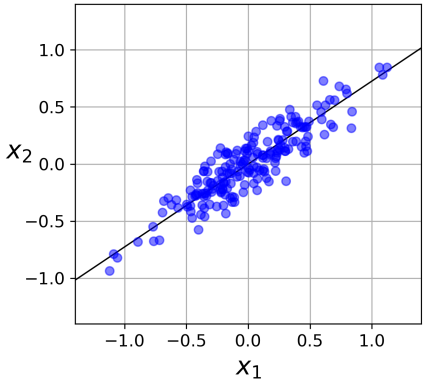
Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.



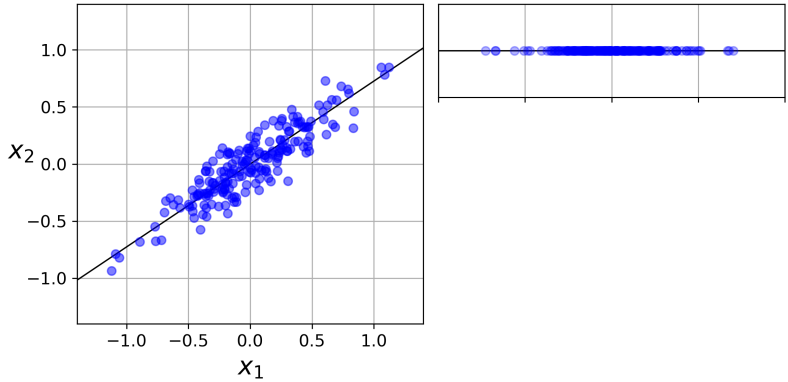
Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.



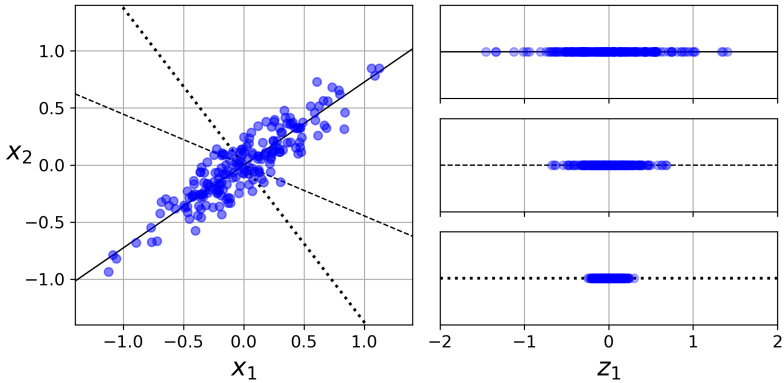
Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.

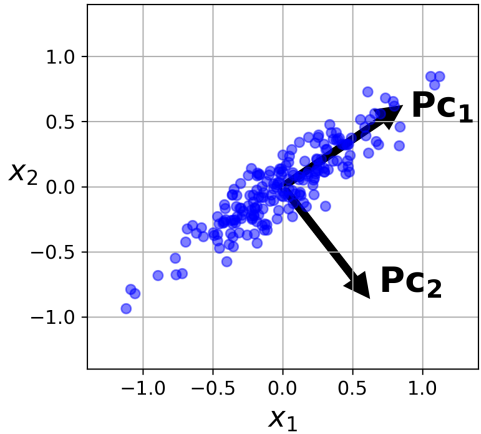


Visualizando la idea (continuación)

PCA busca una nueva dirección, PC_1, donde los puntos estén más esparcidos. Esa es la dirección de **máxima varianza**, el eje donde los documentos se diferencian más.



Visualizando la idea (continuación)



Qué hace PCA en la práctica

PCA toma la matriz de datos (por ejemplo, la DTM) y busca **combinaciones lineales de palabras**:

$$PC_1 = \delta_{1,1}X_1 + \delta_{2,1}X_2 + \cdots + \delta_{k,1}X_k$$

donde los coeficientes $\delta_{j,1}$ (los *loadings*) indican qué tan importante es cada palabra para esa dimensión.

- El primer componente (PC_1) explica la mayor parte de las diferencias entre documentos.
- El segundo componente (PC_2) explica la siguiente mayor parte, y así sucesivamente.

¿Qué son los componentes principales?

La utilidad de la reducción de dimensionalidad puede no ser obvia en dos dimensiones, pero se vuelve mucho más clara cuando tenemos datos altamente dimensionales.

Supongamos que tenemos n observaciones, cada una con k variables o atributos, representadas por $X_1, X_2, \ldots, X_k$.

Por ejemplo, consideremos 500 artículos de prensa de los principales medios colombianos (El Tiempo, El Espectador, Semana) publicados durante 2024, donde cada artículo se representa mediante un vocabulario de 2,000 palabras únicas más frecuentes (excluyendo stopwords). Aquí, n serían los 500 artículos y k las 2,000 palabras del vocabulario

La solución: reducción de dimensionalidad

Necesitamos entonces un método que nos permita encontrar una representación de baja dimensionalidad que contenga la mayor información posible.

Intuitivamente, el PCA plantea que cada observación se encuentra en un espacio k-dimensional, pero no todas estas dimensiones son igualmente informativas.

Por lo tanto, el PCA busca representar, o mejor dicho, proyectar, los datos originales a un espacio de menor dimensión de forma tal de retener la mayor cantidad de información posible.

Componentes principales

Estas nuevas dimensiones, encontradas por el PCA y llamadas componentes, son combinaciones lineales de las variables originales. El primer componente entonces tomará la forma:

$$f_1 = \delta_{11}X_1 + \delta_{12}X_2 + \cdots + \delta_{1k}X_k$$

donde f_1 denota el primer componente principal y los δ_{ij} son conocidos como pesos o *loadings* del primer componente principal.

Esta ecuación claramente muestra que el primer componente principal es una combinación lineal de las variables originales.

La pregunta clave

Dado que solo contamos datos sobre $X_1, X_2, \ldots, X_k$, surge la pregunta:

¿Cómo se calculan estos δ's que me permiten construir los componentes?

Cálculo del primer componente: setup

Formalmente, supongamos que X es una matriz $n \times k$ que contiene los datos de n observaciones y k variables, cada una centrada para tener media cero.

La matriz X tiene asociada una matriz de covarianza $S = \text{Var}(X)$, que es una matriz cuadrada de orden k.

El objetivo del primer componente principal es encontrar la combinación lineal de las variables originales que maximice la varianza, preservando así la mayor cantidad de información posible.

Example