Aprendizaje No Supervisado y Texto como Datos

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

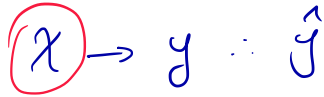
1. Motivation
2. Principal Component Analysis
   - What are PCAs?

3. Text as Data

LDA

Agenda

Motivation



- One way to think about almost everything we do is as dimension reduction.
- We are trying to learn from high-dimensional X some low-dimensional summaries that contain the information necessary to make good decisions.
- We have a high-dimensional X, and you try to model it as having been generated from a small number of components/factors.
- We are attempting to simplify X for its own sake.

Motivation

- Unsupervised learning is often much more challenging.
- The exercise tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response.
- Unsupervised learning is often performed as part of an exploratory data analysis.
- Furthermore, it can be hard to assess the results obtained from unsupervised learning methods, since there is no universally accepted mechanism for performing cross- validation or validating results on an independent data set.
- There is no way to check our work because we don't know the true answer: the problem is unsupervised.

Agenda

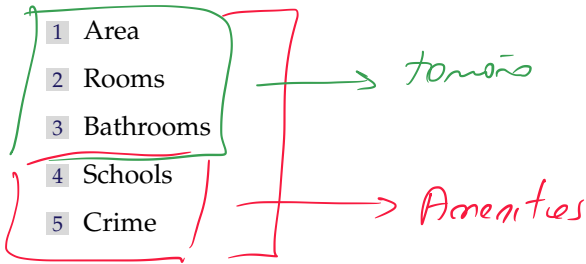
Agenda

Principal Component Analysis

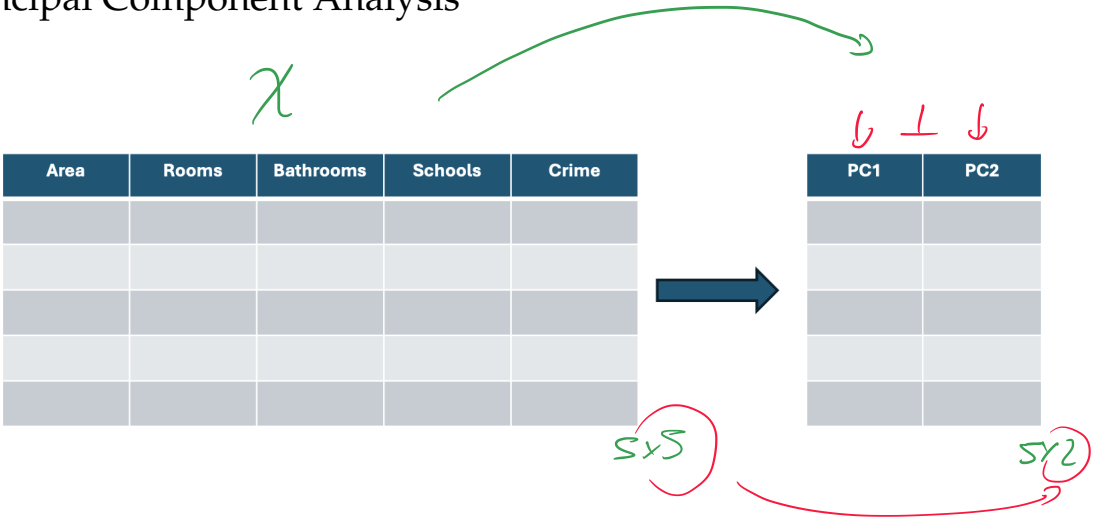
- PCA is an unsupervised learning technique that allows to
  - reduce the dimensionality of data sets,
  - while preserving as much "variability" as possible.
- It is an unsupervised approach, it involves only a set of variables/features $X_1, X_2, \ldots, X_p$, and no associated response Y.

Principal Component Analysis

- For example:



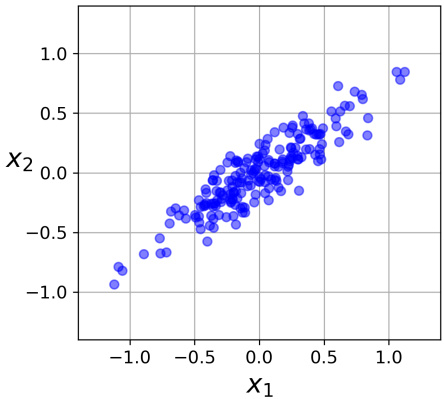
Principal Component Analysis



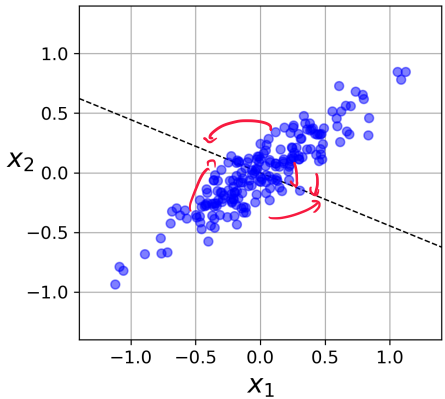
Principal Component Analysis

- PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation.
- The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting.
- PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension.

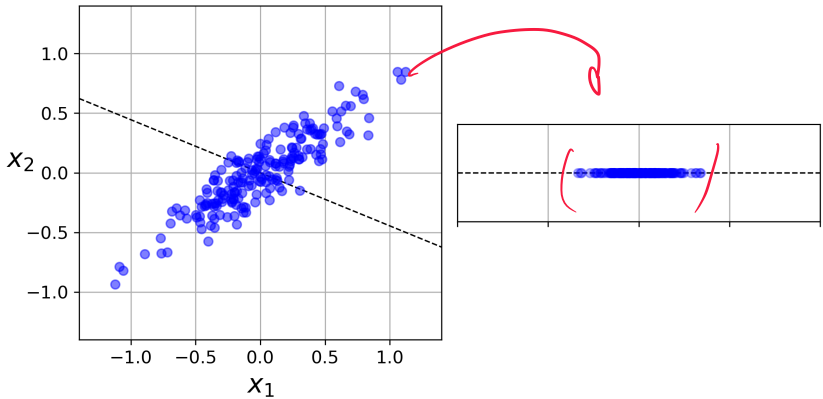
Principal Component Analysis



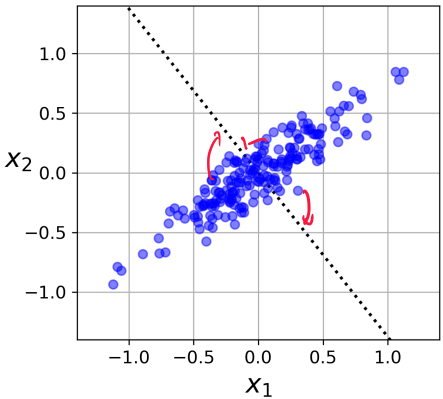
Principal Component Analysis



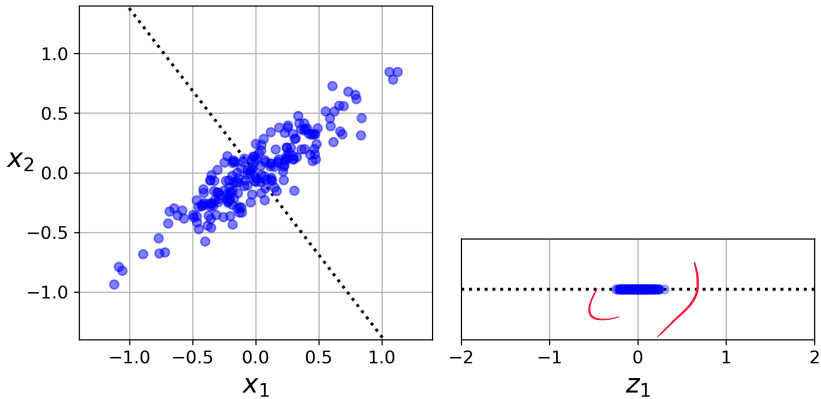
Principal Component Analysis



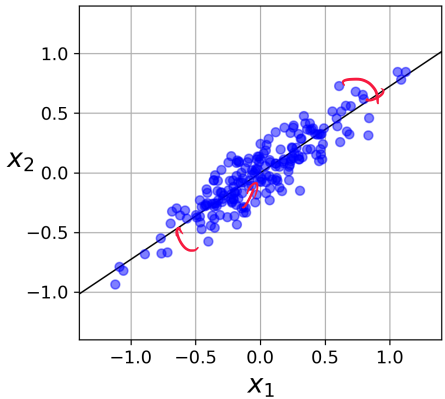
Principal Component Analysis



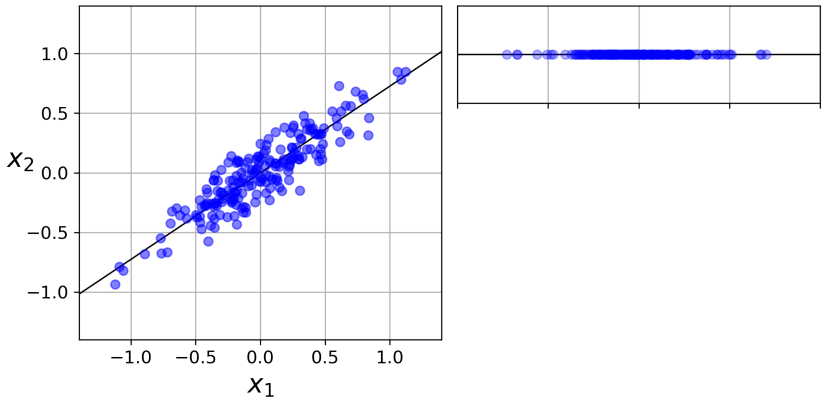
Principal Component Analysis



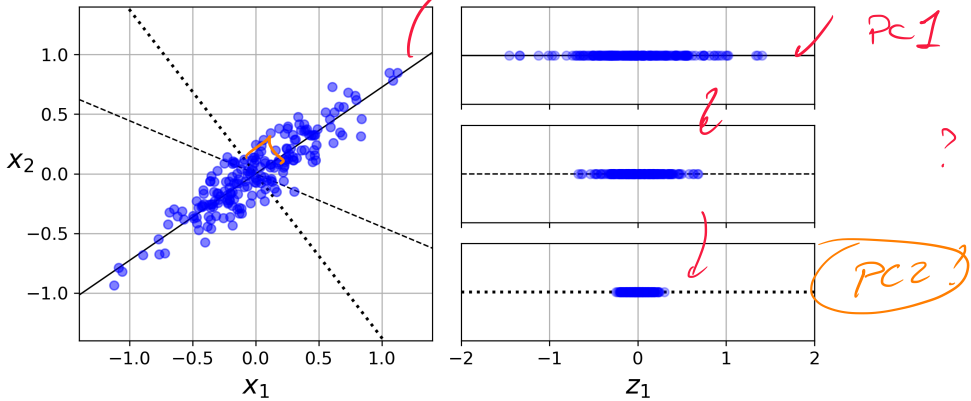
Principal Component Analysis



Principal Component Analysis

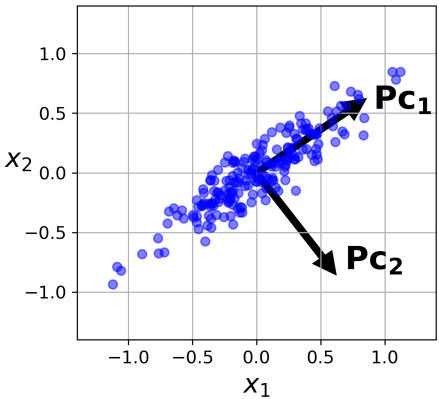


Principal Component Analysis



$$PC1 = \delta_{11} x_1 + \delta_{21} x_2$$

Principal Component Analysis



Principal Component Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$PC_1 = \delta_{11} X_1 + \delta_{21} X_2 + \cdots + \delta_{p1} X_p \tag{1}$$

- The δ coefficients are called loadings or rotations—these are properties of the model and are shared across all observations.
- By normalized we mean that $\sum_{j=1}^{p} \delta_{j1}^2 = 1$

Principal Component Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$PC_1 = \delta_{11}X_1 + \delta_{21}X_2 + \cdots + \delta_{p1}X_p \tag{1}$$

- The δ coefficients are called loadings or rotations—these are properties of the model and are shared across all observations.
- By normalized we mean that $\sum_{j=1}^{p} \delta_{j1}^2 = 1$
- In our example:

$$PC_1 = \delta_{11}Area + \delta_{21}Rooms + \delta_{31}Bathrooms + \delta_{41}Schools + \delta_{51}Crime \tag{2}$$

Motivation

- Given a $n \times p$ data set X, how do we compute the first principal component?

$$PC_1 = X\delta_1 \tag{3}$$
$$= \delta_{11}X_1 + \delta_{21}X_2 + \cdots + \delta_{p1}X_p \tag{4}$$

- The idea is to preserve the most information possible
- In other words, we are going to try to generate an index that reproduces (the best it can) the information (variability) of the original variables
- How we do that?
- Maximize the variance

Principal Component Analysis

- The problem then looks like

$$max V(PC_1) = max V(X\delta_1) \tag{5}$$

$$\text{subject to} \tag{6}$$

$$\delta_1 \delta_1' = 1$$

- Let us call the solution to this problem δ_1^*
- $PC_1^* = X\delta_1^*$ is the 'best' linear combination of X.

Principal Component Analysis

- The first main component? Are there others?
- Let's consider the following problem:

$$\max_{\delta_2} \ \delta_2' S \delta_2 \tag{7}$$

$$\text{st} \tag{8}$$

$$\delta_2' \delta_2 = 1 \tag{9}$$

$$\delta_2' \delta_1 = 0 \tag{10}$$

- $PC_2^* = X\delta_2^*$ is the second principal component : the best linear combination which is orthogonal to the best initial linear combination.
- Recursively, using this logic you can form q main components.
- Note that algebraically we could construct $q = p$ factors, actually the number of PC are $min(n-1, p)$

Relative importance of factors

- Let $PC_j = X\delta_j$, $j = 1, \ldots, K$ be the j-th principal component.

$$V(PC_j) = \delta_j' S \delta_j \tag{11}$$
$$= p_j' P' \Lambda P p_j \tag{12}$$
$$= \lambda_j \tag{13}$$

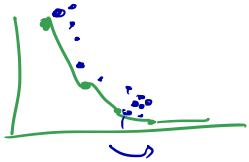
(the variance of the j-th principal component is the j-th ordered eigenvalue of S).

- We can show that the total variance of X is the sum of the variances of x_j , $j = 1, \ldots, p$, that is $trace(S)$

Selection of factors

- Although a matrix X of dimension $n \times p$ generally has $min(n-1, p)$ different principal components.
- In practice, we are generally not interested in all the components, but rather stay with the first ones that allow us to visualize or interpret data.
- Indeed, we would like to keep the minimum number that allows us a good understanding of the data.
- The natural question that arises here is whether there is an established way to determine the number of principal components to use.
- Unfortunately, there is no accepted objective way in the literature to answer it.

Selection of factors



- However, there are three simple approaches that can guide you in deciding the number of relevant major components.

1) - Visual examination of screeplot
3) - Kaiser criterion.
2) - Proportion of variance explained.

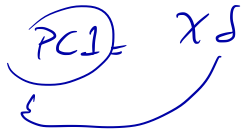
$\sum \lambda_i = p$

$p = 5 \qquad \lambda_1 = 3 \quad \lambda_2 = 1 \quad \lambda_3 < 1 \quad \lambda_4 < 1 \quad \lambda_5 < 1$

$\lambda > 0$

$\frac{3}{5} \qquad \frac{1}{5}$

PCA as a Factor Model



$$x_i = hf_i + u_i \tag{14}$$

- h y u_i son $k \times 1$
- f_i es el factor que es 1×1
- h son los loadings (cargas, pesos) factoriales.
- u_i son los errores idiosincráticos.

Factor Model Interpretation

Test Scores

$$x_i = hf_i \tag{15}$$

- x_i es un conjunto de calificaciones en exámenes para un estudiante dado.
- f_i es la habilidad latente del estudiante.
- h es cómo la habilidad afecta las diferentes calificaciones en exámenes.
  - Algunos exámenes pueden estar altamente relacionadas con la habilidad.
  - Algunos exámenes pueden estar menos relacionadas.
  - Algunos exámenes no estar relacionadas (¿aleatorias?).

Factor Model Interpretation

Test Scores

$$x_i = \sum_{m=1}^{r} h_m f_{mi} + u_i = H_0 f_i + u_i \tag{16}$$

- Los loadings etan normalizadas para ser ortonormales $\rightarrow$ factores no correlacionados
- Las cargas factoriales h_m son los eigen vectors de Σ_x (varianza de x) asociados con los mayores r eigen values.
- Ejemplo de calificaciones en exámenes (cont.)
  - Existen más de una forma de "habilidad". Ej. literaria y matemática.
  - En la economía laboral, se ha hipotetizado una distinción entre habilidad cognitiva y no cognitiva que ha sido muy útil para explicar los patrones salariales (algunos trabajos requieren una u otra, y algunos ambas, por ejemplo, cirujano).

Factor Interpretation: Examples

Agenda

Text as Data: The Big Picture

- We generate vast quantities of raw unstructured text.
- As the costs of storage drop and as more conversations and records move to digital platforms, we accumulate massive corpora that track communications:
  - customer conversations,
  - product descriptions or reviews,
  - news,
  - comments, blogs, tweets, etc...
- The information in text is a rich complement to the more structured variables contained in a traditional transaction or customer database.
- Social scientists have also woken up to the potential of such data and recent years have seen an explosion in studies that make use of text as data.

Text as Data: Motivation

THE JOURNAL OF FINANCE • VOL. LXII, NO. 3 • JUNE 2007

Giving Content to Investor Sentiment: The Role of Media in the Stock Market

PAUL C. TETLOCK*

ABSTRACT

I quantitatively measure the interactions between the media and the stock market using daily content from a popular *Wall Street Journal* column. I find that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. These and similar results are consistent with theoretical models of noise and liquidity traders, and are inconsistent with theories of media content as a proxy for new information about fundamental asset values, as a proxy for market volatility, or as a sideshow with no relationship to asset markets.

Text as Data: Motivation

Econometrica, Vol. 78, No. 1 (January, 2010), 35–71

WHAT DRIVES MEDIA SLANT? EVIDENCE FROM U.S. DAILY NEWSPAPERS

BY MATTHEW GENTZKOW AND JESSE M. SHAPIRO[1]

We construct a new index of media slant that measures the similarity of a news outlet's language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms' actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to consumer preferences, which account for roughly 20 percent of the variation in measured slant in our sample. By contrast, the identity of a newspaper's owner explains far less of the variation in slant.

KEYWORDS: Bias, text categorization, media ownership.

Text as Data

- To analyze text, we need to transform it into data that can be input to numeric regression and factorization algorithms.
  - Bag of Words (BoW) and DTMs
  - Word Embeddings

Text as Data: BOW

1) ▶ *Juan tiene una vaca.*
2) ▶ *Pedro tiene una caballo.*
3) ▶ *La Economía va a tener una desaceleración.*

[Juan, tiene, una, vaca]
1 1 1 1

Economía	Juan	Pedro	9 nuevo	vaca	caballo
0	1	0		5	0
0	0	1		0	1
3	0	0		0	0

Word Counts vs TF-IDF

DTM

[ ]

- $f_{i,j}$: número de veces que aparece la palabra i en el documento j
- $tf - idf_{ij} = tf_{ij} \times \left(\log\left(\frac{N}{df_i}\right)\right)$
  donde:
  - tf_{ij} es la frecuencia palabra i en el documento j
  - df_{ij} es el número de documentos que contienen la palabra i
  - N es el número de documentos

$$TF = \frac{f_{ij}}{\sum f_{ij}}$$

Topic Models

- Text is super high dimensional
- Some times unsupervized factor model is a popular and useful strategy with text data
- You can first fit a factor model to a giant corpus and use these factors for supervised learning on a subset of labeled documents.
- The unsupervised dimension reduction facilitates the supervised learning

Topic Models: Examples PCA

Topic Models: Latent Dirichlet Allocation

Overview

- The approach of using PCA to factorize text was common before the 2000s.
- Versions of this algorithm were referred to under the label latent semantic analysis.
- However, this changed with the introduction of topic modeling, also known as Latent Dirichlet Allocation (LDA), by Blei et al. in 2003.
- These authors proposed you take the bag-of-words representation seriously and model token counts as realizations from a probabilistic framework.

Topic Models: Latent Dirichlet Allocation

Paso 1: Crear los Tópicos

Para cada uno de los K tópicos:

- Generamos una distribución de palabras:

$$\phi_k \sim \text{Dirichlet}(\beta)$$

- ϕ_k es un vector de probabilidades sobre el vocabulario.

Topic Models: Latent Dirichlet Allocation

Paso 2: Mezcla de Tópicos por Documento

Para cada documento $d = 1, \ldots, D$:

- Generamos una distribución de tópicos:

$$\theta_d \sim \text{Dirichlet}(\alpha)$$

- θ_d representa la proporción de cada tópico en el documento d.

Si $K = 3$ $\quad \theta_1 = (0{,}7;\ 0{,}2;\ 0{,}1)$

Topic Models: Latent Dirichlet Allocation

Paso 3: Generación de Palabras

Para cada palabra $n = 1, \ldots, N_d$ del documento d:

1. Elegimos un tópico:

$$z_{d,n} \sim \text{Multinomial}(\theta_d)$$

2. Elegimos una palabra desde ese tópico:

$$w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$$

Este proceso se repite para cada palabra del documento.

Topic Models: Examples LDA