PCA 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?
   - Factor Model Interpretation
3. Principal Component Regression (PCR)
4. Text as Data

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:
  1. Area
  2. Rooms
  3. Bathrooms
  4. Schools
  5. Crime

Principal Component Analysis

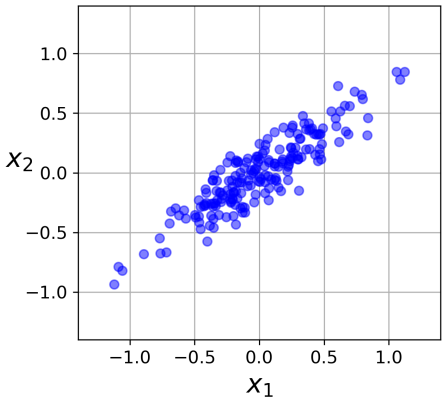
Area	Rooms	Bathrooms	Schools	Crime

PC1	PC2

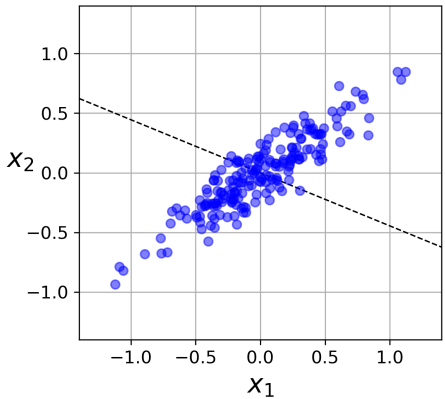
Principal Component Analysis

- 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.

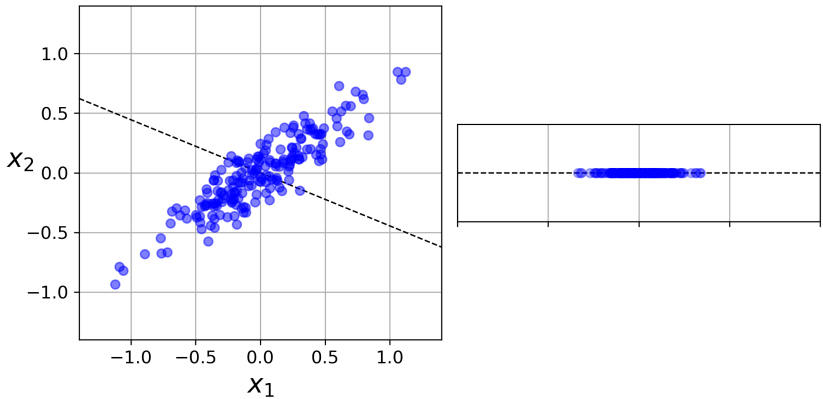
Principal Component Analysis



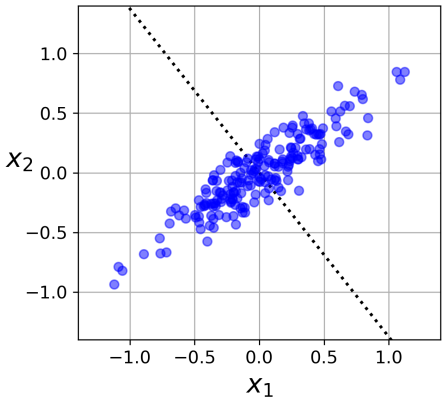
Principal Component Analysis



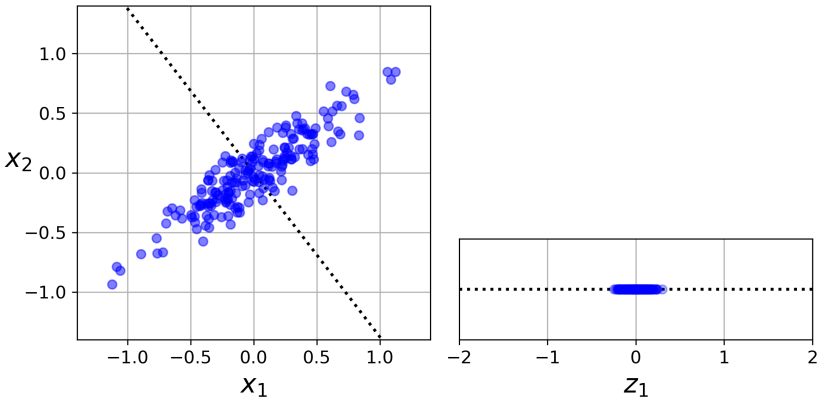
Principal Component Analysis



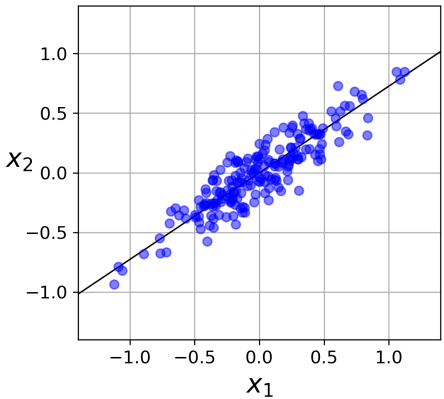
Principal Component Analysis



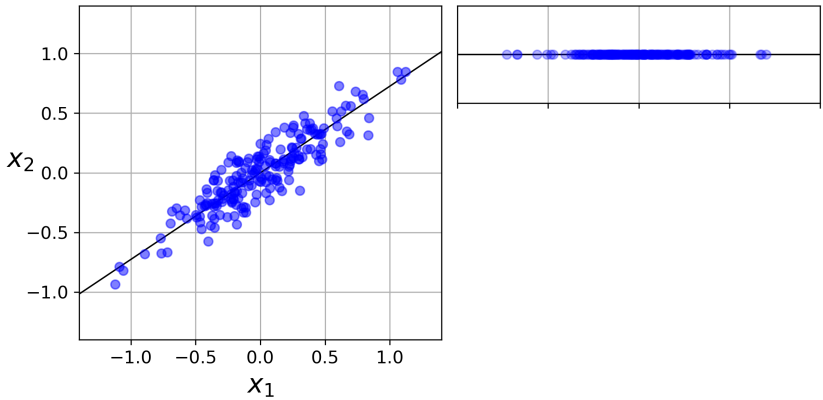
Principal Component Analysis



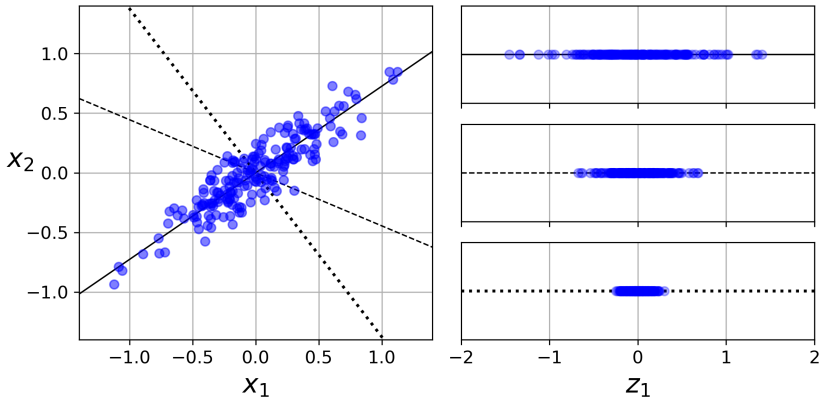
Principal Component Analysis



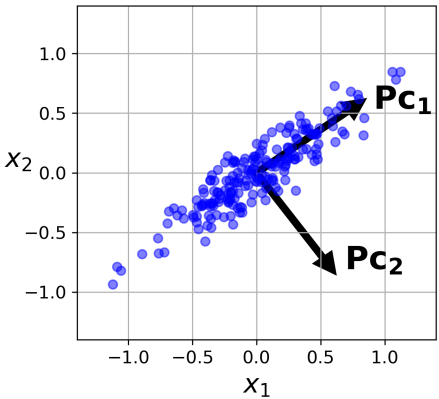
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



Principal Component Analysis

- Example: X_1, X_2

$$F_1 = \delta_{11} X_1 + \delta_{21} X_2 \tag{1}$$

- The δ coefficients are called loadings or rotations
- $\sum_{j=1}^{2} \delta_{j1}^2 = 1$

Agenda

Factor Model Interpretation

- We can map each observation to the K factors
- Suppose we have p regressors and K=1

$$x_i = hf_i \tag{2}$$

- h is $p \times 1$
- fi 1×1 and is the factor
- h are the factor loadings
- In this model, the factor f_i affects all regressors $x_j i$
- But the magnitude is specific to the regressor and captured by h

Factor Model Interpretation

Test Scores

$$x_i = hf_i \tag{3}$$

- x_i is a set of test scores for an individual student
- f_i is the student's latent ability
- h is how ability affects the different test scores
  - Some tests may be highly related to ability
  - Some tests may be less related
  - Some may be unrelated (random?)

Factor Model Interpretation

Test Scores

$$x_i = \sum_{m=1}^{k} h_m f_{mi} \tag{4}$$

- Interpretation
- There are more than one form of ability
- i.e. literary and mathematical
- In labor economics, there has been hypothesized a distinction between cognitive and non-cognitive ability which has been very useful in explaining wage patterns (some jobs require one or the other, and some both (e.g. surgeon)

Factor Interpretation: Examples

Agenda

Principal Component Regression (PCR)

- Now that you've learned how to fit factor models, what are they good for?
- In some settings, as in the previous political science example, the factors themselves have clear meaning and can be useful in their own right for understanding complex systems.
- More commonly, unfortunately, the factors are of dubious origin or interpretation.
- However, they can still be useful as inputs to a regression system.
- Indeed, this is the primary practical function for PCA, as the first stage of principal components regression (PCR).

Principal Component Regression (PCR)

- The concept of PCR is simple:
  - Instead of doing $y \to X$,
  - Use a lower-dimension set of principal components as covariates.
- This is a fruitful strategy for a few reasons:
  - PCA reduces dimension, which is usually good.
  - The PCs are independent, so you have no multicollinearity and the final regression is easy to fit.
  - You might have far more unlabeled x_i than labeled (x_i, y_i) pairs. This last point is especially powerful.
  - You can use unsupervised learning (PCA) on a massive bank of unlabeled data and use the results to reduce dimension and facilitate supervised learning on a smaller set of labeled observations.

Principal Component Regression (PCR)

- The algorithm is straightforward.

```
1 mypca <- prcomp(X, scale=TRUE)
2 z <- predict(mypca)[,1:K]
3 reg <- lm(y ., data=as.data.frame(z))
```

Principal Component Regression (PCR)

- The disadvantage of PCR is that PCA will be driven by the dominant sources of variation in X.
- If the response is connected to these dominant sources of variation, PCR works well.
- If it is more of a "needle in the haystack response," driven by a small number of inputs, then PCR will not work well.
- In practice, you do not know what scenario you are in until you try both PCR and, say, a lasso regression on the raw X inputs.

Principal Component Regression (PCR)

- How many PC do we use?
  - When PCA was used as a dimensionality reduction tool *per se* we had some guidelines...
- Should we do the same here?

Principal Component Regression (PCR)

- How many PC do we use?
  - When PCA was used as a dimensionality reduction tool *per se* we had some guidelines...
- Should we do the same here?
- In PCR the approach is slightly different
  - Construct $min(n-1, p)$ components
  - Use K fold crossvalidation adding 1 PC at a time
  - Choose the model with the lowest out of sample MSE
- Because the PCs are ordered (by their variance) and independent, this works better than subset selection on the raw dimensions of X_i.

Principal Component Regression (PCR)

- An alternative mechanism is run a lasso on the full set of PCs (works best in practice).
- This procedure makes it easy to incorporate other information in addition to the PCs.
- For example, one tactic that works well in practice is to put both PC and Xs into the lasso model matrix.
  - This then allows the regression to make use of the underlying factor structure in X and still pick up individual X_j signals that are related to y.
  - This hybrid strategy is a solution to the disadvantage of PCR mentioned earlier—that it will only pick up dominant sources of variation in X.

Principal Component Regression (PCR)

Summary of the steps

- Given a sample of regression input observations x_i, accompanied by output labels y_i for some subset of these observations:
  1. Fit PCA on the full set of X inputs to obtain PC of length $min(n-1, p)$.
  2. For the labeled subset, run a lasso regression for y on f (PC).
     - Alternatively, regress y on f and Xs to allow simultaneous selection between PCs and raw inputs.
  3. To predict for a new X_{new}, use the rotations from step 1 to get $f = \delta X_{new}$ and then feed these scores into the regression fit from step 2.

PCR Example

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.

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, you 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

propiedad	description
1	Vendo casa de 2 pisos con terraza pequeña primer piso 1 alcoba grande con clóset
2	Venta de hermoso apartamento con solo 2 años de construido, tercer piso con ascensor
3	Venta de hermoso apartamento moderno, con amplios espacios, balcón con excelente vista panorámica, 8vo piso con ascensor

Text as Data

propiedad	description
1	Vendo casa de 2 pisos con terraza pequeña primer piso 1 alcoba grande con clóset
2	Venta de hermoso apartamento con solo 2 años de construido, tercer piso con ascensor
3	Venta de hermoso apartamento moderno, con amplios espacios, balcón con excelente vista panorámica, 8vo piso con ascensor

Docs	8vo	alcoba	amplios	años	apartamento	ascensor	balcón	casa	clóset	con	construido,	espacios,	excelente	grande	hermoso	moderno,	panorámica,	pequeña	piso	pisos	primer	solo	tercer	terraza	vendo	venta	vista
1	0	1	0	0	0	0	0	1	1	2	0	0	0	1	0	0	0	1	1	1	1	0	0	1	1	0	0
2	0	0	0	1	1	1	0	0	0	2	1	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	0
3	1	0	1	0	1	1	1	0	0	3	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	1	1

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

Factor Interpretation: Examples