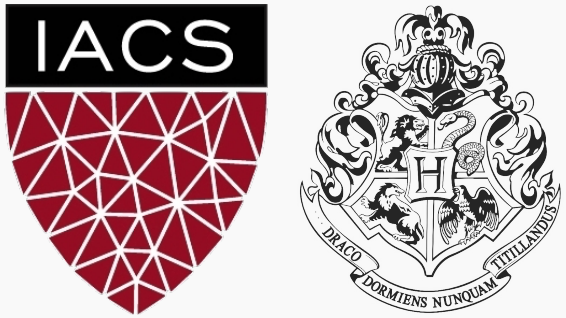
Lecture #6: Dimension Reduction

CS 109A, STAT 121A, AC 209A: Data Science

Pavlos Protopapas Kevin Rader



Lecture Outline

Review

More on Interaction Terms

High Dimensionality

Principal Components Analysis (PCA)

PCA for Regression (PCR)

PCA vs Variable Selection

Review

Overfitting and Regularization

More on Interaction Terms

Interaction Terms: A Review

Recall that an interaction term between predictors X_1 and X_2 can be incorporated into a regression model by including the multiplicative (i.e. cross) term in the model, for example

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon.$$

Example

Suppose X_1 is a binary predictor indicating whether a NYC ride pickup is a tax or an Uber, X_2 is the times of day of the pickup and Y is the length of the ride.

What is the interpretation of β_3?

Including Interaction Terms in Models

Recall that to avoid overfitting, we sometimes elect to exclude a number of terms in a linear model.

It is standard practice to always include the ***main effects*** in the model. That is, we always include the terms involving only one predictor, $\beta_1 X_1, \beta_2 X_2$ etc.

Question: Why are the ***main effects*** important?

Question: In what type of model would it make sense to include the interaction term without one of the main effects?

How Many Interaction Terms?

Example

Our NYC taxi and Uber dataset has 1.1 billion taxi trips and 19 million Uber trips. Each trip is described by $p = 23$ predictors (and 1 response variable). How many interaction terms are there?

- Two-way interactions: $\binom{p}{2} =$
- Three-way interactions: $\binom{p}{3} =$
- Etc

The total number of interaction terms (including main effects) is $\sum_{k=1}^{p} \binom{p}{k} = 2^p \approx 8.3$ million.

What is problem with building a model that includes all possible interaction terms?

Model Unidentifiability

Previously, we had been using samples of 100k observations from the dataset to build our models. If we include all possible interaction terms, our model will have 8.3 mil parameters. **We will not be able to uniquely determine 8.3 mil parameters with only 100k observations**. In this case, we call the model ***unidentifiable***.

In practice, we can:

- increase the number of observation
- consider only scientifically important interaction terms
- perform variable selection
- perform another ***dimensionality reduction*** technique like PCA

High Dimensionality

When Does High Dimensionality Occur?

The problem of high dimensionality can occur when the number of parameters exceeds or is close to the number of observations. This can occur when we consider lots of interaction terms, like in our previous example. But this can also happen when the number of main effects is high.

For example:

- When we are performing polynomial regression with a high degree and a large number of predictors.
- When the predictors are genomic markers in a computational biology problem.
- When the predictors are the counts of all English words appearing in a text.

A Framework For Dimensionality Reduction

One way to reduce the dimensions of the feature space is to create a new, smaller set of predictors by taking linear combinations of the original predictors.

We choose $Z_1, Z_2, \ldots, Z_m$, where $m < p$ and where each Z_i is a linear combination of the original p predictors

$$Z_i = \sum_{j=1}^{p} \phi_{ji} X_j$$

for fixed constants ϕ_{ji}. Then we can build a linear regression regression model using the new predictors

$$Y = \beta_0 + \beta_1 Z_1 + \ldots + \beta_m Z_m + \epsilon.$$

Notice that this model has a smaller number ($m < p$) of parameters.

A Framework For Dimensionality Reduction

A method of dimensionality reduction includes 2 steps:

1. Determine a optimal set of new predictors $Z_1, \ldots, Z_m$, for $m < p$.
2. Express each observation in the data in terms of these new predictors. The transformed data will have m columns rather than p.

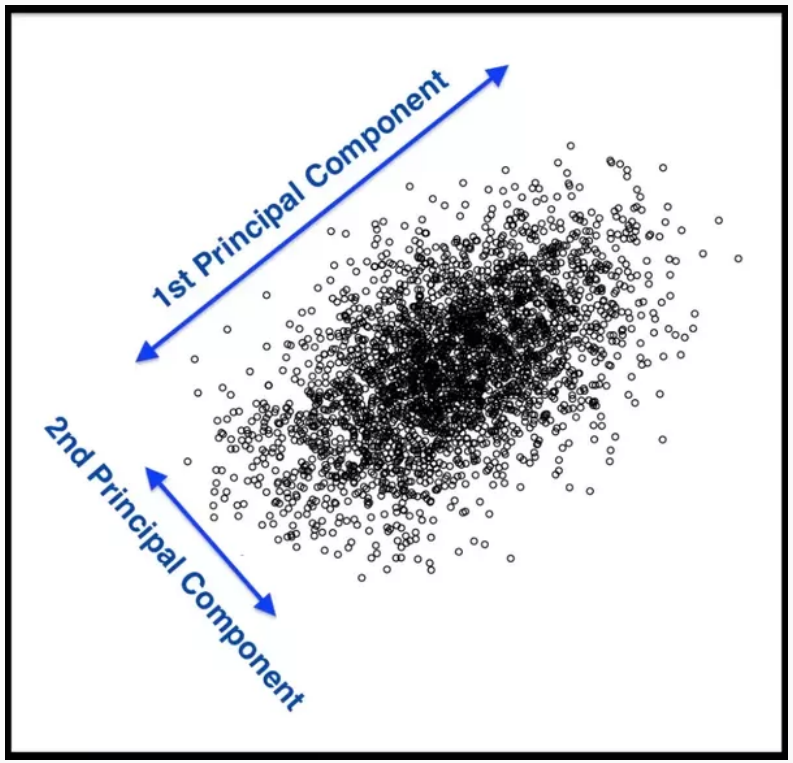
Thereafter, we can fit a model using the new predictors.

The method for determining the set of new predictors (what do we mean by an optimal predictors set) can differ according to application. We will explore a way to create new predictors that captures the variations in the observed data.

Principal Components Analysis (PCA)

Principal Components Analysis (PCA)

Principal Components Analysis (PCA) is a method to identify a new set of predictors, as linear combinations of the original ones, that captures the 'maximum amount' of variance in the observed data.



Principal Components Analysis (PCA)

Definition

Principal Components Analysis (PCA) produces a list of p ***principle components*** $(Z_1, \ldots, Z_p)$ such that

- Each Z_i is a linear combination of the original predictors, and it's vector norm is 1
- The Z_i's are pairwise orthogonal
- The Z_i's are ordered in decreasing order in the amount of captured observed variance.

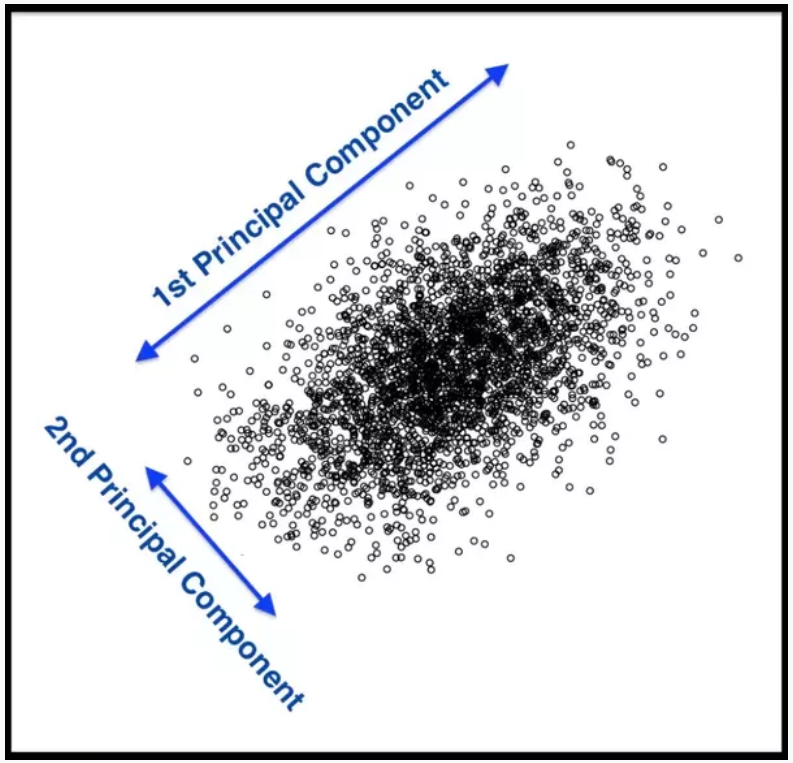
  That is, the observed data shows more variance in the direction of Z_1 than in the direction of Z_2.

To perform dimensionality reduction we select the top m principle components of PCA as our new predictors and express our observed data in terms of these predictors.

The Intuition Behind PCA

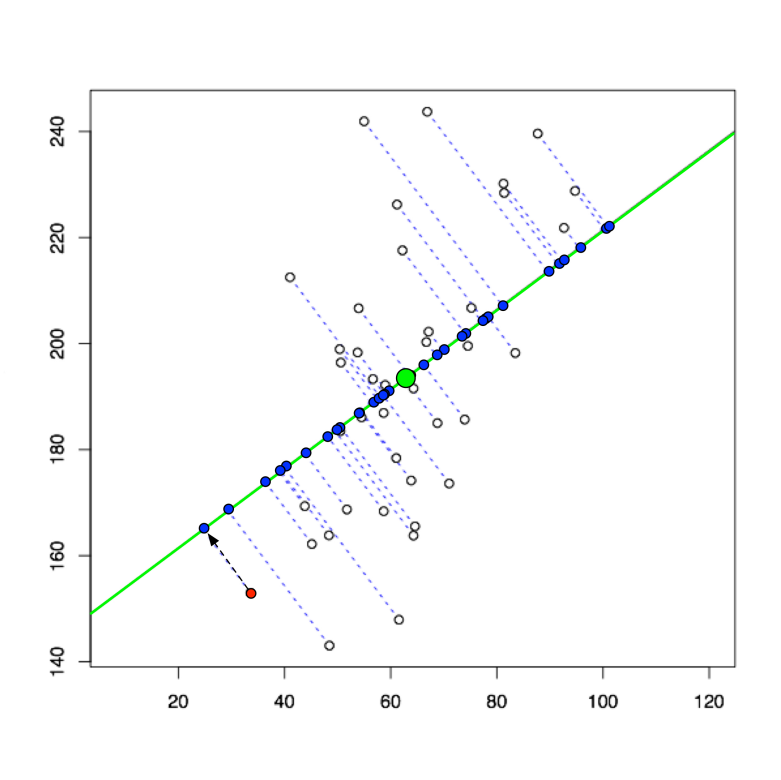
Top PCA components capture the most of amount of variation (interesting features) of the data.

Each component is a linear combination of the original predictors - we visualize them as vectors in the feature space.



The Intuition Behind PCA

Transforming our observed data means projecting our dataset onto the space defined by the top m PCA components, these components are our new predictors.



PCA for Regression (PCR)

PCA for Data Preprocessing

PCA vs Variable Selection

The Geometry Behind PCA and Variable Selection

Bibliography

1. Bolelli, L., Ertekin, S., and Giles, C. L. ***Topic and trend detection in text collections using latent dirichlet allocation***. In European Conference on Information Retrieval (2009), Springer, pp. 776-780.
2. Chen, W., Wang, Y., and Yang, S. ***Efficient influence maximization in social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (2009)***, ACM, pp. 199-208.
3. Chong, W., Blei, D., and Li, F.-F. ***Simultaneous image classification and annotation. In Computer Vision and Pattern Recognition***, 2009. CVPR 2009. IEEE Conference on (2009), IEEE, pp. 1903-1910.
4. Du, L., Ren, L., Carin, L., and Dunson, D. B. ***A bayesian model for simultaneous image clustering, annotation and object segmentation***. In Advances in neural information processing systems (2009), pp. 486-494.
5. Elango, P. K., and Jayaraman, K. ***Clustering images using the latent dirichlet allocation model.***
6. Feng, Y., and Lapata, M. ***Topic models for image annotation and text illustration***. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010), Association for Computational Linguistics, pp. 831-839.
7. Hannah, L. A., and Wallach, H. M. ***Summarizing topics: From word lists to phrases.***
8. Lu, R., and Yang, Q. ***Trend analysis of news topics on twitter***. International Journal of Machine Learning and Computing 2, 3 (2012), 327.