Principal Component Analysis (PCA)

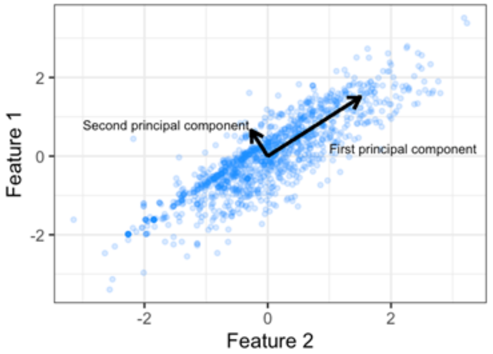
Introduction

Although having very large datasets is often thought of as a good thing, too many variables can bring about the “curse of dimensionality,” which is the idea that an enormous amount of variables or features can give rise to problems with model interpretability, data noise, and data visualization. To address these issues, we may apply PCA, a popular and effective method which reduces the number of dimensions while still retaining most of the variance. Common use cases for this technique are:

- Dimensionality Reduction, data compression
- Transformation of correlated variables into orthogonal, uncorrelated PCs (collinearty handling, feature engineering)
- Improving model generalizability

Data Requirements

PCA does not have strict requirements that must be met to perform the computation, however, there are some underlying assumptions that can indicate whether or not its application will be effective. In general, PCA is best applied on datasets with a large number features of numerical (continuous) data where the structure of large variances in the data is important.



In this 2D example we can see that the original feature space (feature 1 and feature 2) does not capture the most variability. Building a new axis from the first and second PCs (black arrows) better captures the variability in this data. The relationship between data points has not changed, but we can consider that we are simply describing the same data in terms of a new (rotated) axis.

Method Overview

The main idea behind PCA is to take a “cloud” of high dimensional data points and find a new set of axes, or basis vectors, which better describes the variance of the data and reveals underlying lower dimensional correlations (highly correlated data points will tend to cluster together). We can think of this as happening by rotating the axes on which we view our high dimensional data. This means that the relationships between our datapoints are preserved, but the way we are viewing it, the perspective from our axes, has changed; that is, we have moved from viewing our data in feature space, its original columns, to viewing it in principal component space.

The principal components (PCs) which form our new axes are derived from linear combinations of the original features. This means that we may think of PCA in terms of a change of basis or rotation. To carry out PCA, we must first subtract the mean from each feature to ensure that the data have a mean of zero and is centered about the origin. Centering the data also means that the first PC will describe the direction of maximum variance. Next, PCs are computed via eigendecomposition of the centered data covariance matrix, or equivalently, a singular value decomposition of our centered data. Recall from linear algebra that associated with each eigenvalue is an eigenvector: in PCA, the eigenvalue describes how much variance is explained by its respective eigenvector, the PC (eigenvalue and eigenvector can be thought of as the scale and direction respectively of maximal variance). After computing the PCs, we then select a subset of our ordered PCs as a new set of basis vectors on which to project our original data into “PC space.” By selecting only a subset of PCs to derive the new data set, we reduce the dimensionality of the data while still retaining as much variance as possible.

Vishesh Saharan, Jonathan Ramos

Resources

1. https://browse.arxiv.org/pdf/1404.1100.pdf
2. https://ourarchive.otago.ac.nz/bitstream/handle/10523/7534/OUCS-2002-12.pdf?sequence=1&isAllowed=y
3. https://personal.utdallas.edu/~herve/abdi-awPCA2010.pdf
4. https://www.youtube.com/watch?v=fkf4IBRSeEc&t=1s&ab_channel=SteveBrunton
5. http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf
6. https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf
7. https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf

Resources for R

1. https://uc-r.github.io/pca
2. https://www.r-bloggers.com/2021/05/principal-component-analysis-pca-in-r/
3. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp