PCA Report

Donal Loitam

July 27, 2022

Contents

1 Introduction

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets
- It transforms a large set of variables into a smaller one(combinations of the older variables) that still contains most of the information in the large set.
- **Accuracy-Simplicity Trade off:** Reducing the number of variables of a data set naturally comes at the expense of accuracy. As an added benefit, each of the "new" variables after PCA are all independent of one another
- Smaller data sets are easier to explore and visualize and make analyzing data much easier and faster

2 Key Points of the Algorithm

- STANDARDIZATION :
  - The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
  - It is critical to perform standardization prior to PCA because the latter is quite sensitive regarding the variances of the initial variables.
  - Ex. A variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

- COVARIANCE MATRIX COMPUTATION :
    - We compute this matrix to idemtify if there is any significant correlation between any 2 variables because if it is the case, then we can drop one of them as it would be redundant
    - Example of a covariance matrix having three variables x, y, z

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix} \tag{2}$$

- COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS: :
    - Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.
    - These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated/perpendicular.
    - Geometrically, principal components represent the directions of the data that explain a **maximum variance**, or, the lines that capture most information of the data.
    - Larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has
    - PCA tries to put maximum possible information in the first component, then in the second and so on, until having something like shown in the **scree plot** below.

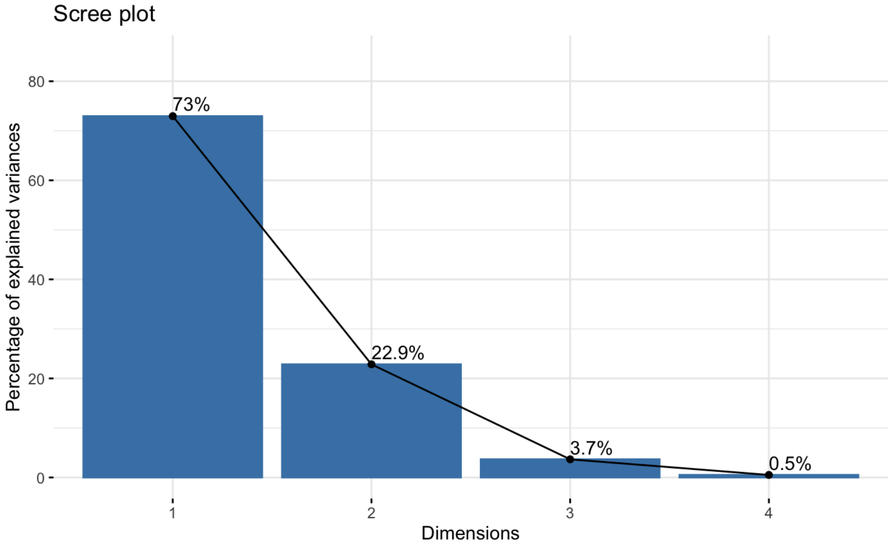

Figure 1: Scree plot for 4 dimensions

- **How PCA Constructs the Principal Components :** Let's assume that the scatter plot of our data set is as shown below, can we guess the first principal component ? Yes, it's approximately the line that matches the purple marks because it goes through the **origin(mean of all the datapoints)** and it's the line in which the projection of the points (red dots) is the most spread out. Or mathematically speaking, it's the line that

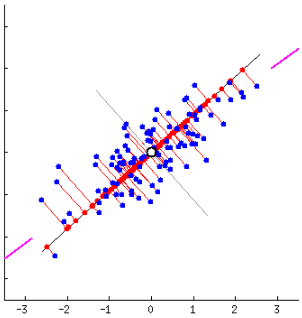

Figure 2: Finding the PCA1

maximizes the variance (the average of the squared distances from the projected points (red dots) to the **origin**). The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

3 Why does it work ?

- First,The eigenvectors of the Covariance matrix are actually the directions of the axes where there is the most variance(most information)
- Second, eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.
- Finally, we make an assumption that more variability in a particular direction correlates with explaining the behavior of the dependent variable. Lots of variability usually indicates signal, whereas little variability usually indicates noise. Thus, the more variability there is in a particular direction is, theoretically, indicative of something important we want to detect

4 Some Questions

1. Explain the Curse of Dimensionality?
Ans.

- As the number of features increase, the number of samples increases, hence, the model becomes more complex
- The more the number of features, the more the chances of overfitting.

2. What are the pros and cons of Dimensionality Reduction ?
Ans. Some advantages for Dimensionality Reduction :-

- BETTER VISUALISATION : Dimensionality Reduction helps us visualize the data on 2D plots or 3D plots.
- REDUCED SPACE AND TIME COMPLEXITY : Fewer dimensions mean less computing. Less data means that algorithms train faster.

While some drawbacks are :-

- LESS INTERPRETABLE FEATURE : Transformed features are often hard to interpret
- Some information is lost, possibly degrading the performance of subsequent training algorithms.

3. Is rotation necessary in PCA?
Ans : Yes, rotation (orthogonal) is necessary to account the maximum variance of the training set. If we don't rotate the components, the effect of PCA will diminish and we'll have to select more number of components to explain variance in the training set.
4. Assumptions taken while applyng PCA ?
Ans.

- There needs to be a **linear relationship** between all variables. The reason for this assumption is that a PCA is based on Pearson correlation coefficients.
- For getting reliable results by using the PCA algorithm, we require a large enough sample size i.e, we should have sampling adequacy.
- We need to have adequate correlations between the variables to be reduced to a smaller number of components.

5. How can you evaluate the performance of a dimensionality reduction algorithm on your dataset?
Ans.
Intuitively, a dimensionality reduction algorithm performs well if it eliminates a lot of dimensions from the dataset without losing too much information
Alternatively, if you are using dimensionality reduction as a preprocessing step before another Machine Learning algorithm (e.g., a Random Forest classifier), then you can simply measure the performance of that second algorithm; if dimensionality reduction did not lose too much information, then the algorithm should perform just as well as when using the original dataset.