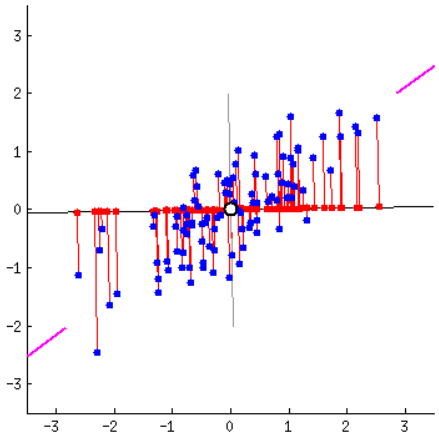
Joia Zhang
Mentor: Jerry Wei
SPA-DRP
Winter 2021

This quarter, I was very fortunate to do a project on principal component analysis (PCA) with my mentor Jerry Wei. PCA was invented by Karl Pearson in 1901, and PCA is still very relevant today. In fact, it's one of the most widely used techniques for dimensionality reduction.

Let's say one has a huge dataset with hundreds of variables. The goal of PCA is to have fewer variables while still retaining the majority of the information. This can be done by projecting the data to a few directions that account for the most variance. Taking the below animation for an example, we have the two-dimensional data and our goal is to reduce to a one-dimensional number line. The best number line occurs in the diagonal direction, which accounts for the most variance and gives the most suitable one-dimensional representation of the original data.



Step 1: Preprocess data

The first step in principal component analysis is to preprocess the data. It is essential to centralize the data by subtracting the mean from each data point so that the covariance can be easily calculated later on. It's optional to divide each variable by the standard deviation, but the plus side of doing so is that standardizing will prevent large scale variables from dominating other variables.

$$X = \frac{value - mean}{standard\ deviation}$$

Step 2: Compute covariance matrix

Once the data is centralized, it becomes much easier to calculate the covariance matrix. All we need to do is multiply the original data with its transpose to get the

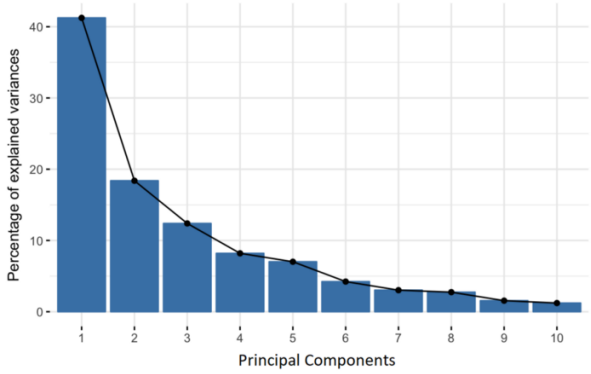
symmetric covariance matrix $S = XX^T = \begin{bmatrix} \text{Cov(x, x)} & \text{Cov(x, y)} & \text{Cov(x, z)} \\ \text{Cov(y, x)} & \text{Cov(y, y)} & \text{Cov(y, z)} \\ \text{Cov(z, x)} & \text{Cov(z, y)} & \text{Cov(z, z)} \end{bmatrix}$ where each covariance is defined as $\text{Cov(x, y)} = \Sigma \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$ so the covariance matrix describes the linear relationships between variables.

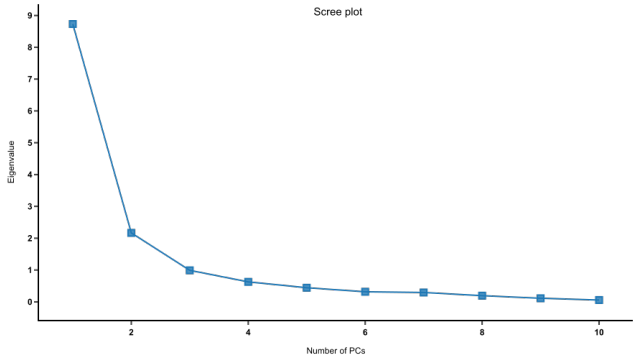
Step 3: Compute eigenvectors and eigenvalues

The symmetric covariance matrix can be decomposed using singular value decomposition into three elements $S = PDP^T$ where P is a projection matrix composed of eigenvectors and D is a diagonal matrix of eigen values. Each eigenvalue corresponds to an eigenvector inside of the projection matrix P, and the eigenvalues reflect how much variance is carried by the corresponding eigenvectors. Principal components are the eigenvectors of S and the eigenvectors with the largest eigenvalues will be picked.

Step 4: Select number of PCs to keep

The next step is to choose how many PCs to keep based on the percentage of variance explained by the PCs.





Step 5: Recast data along PCs

The final step is to recast the data along the PCs. This can be done by matrix multiplying P and the original data X to produce the transformed data $X' = PX$ and selecting the first several columns. The principal components contain a maximal amount of variance and form the new axes of the transformed data. The dimensionality of the data is thereby reduced to the number of PCs used. The PCs (eigenvectors) contain coefficients of linear combinations of the original variables in the data. These linear combinations form a new basis and the data becomes lower dimension.

Acknowledgements

Special thanks to Jerry Wei and the SPA-DRP Program.

Sources

- https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf
- https://en.wikipedia.org/wiki/Principal_component_analysis#:~:text=PCA%20was%20invented%20in%201901,Harold%20Hotelling%20in%20the%201930s
- https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues
- https://chartio.com/learn/charts/what-is-a-scatter-plot/
- https://builtin.com/data-science/step-step-explanation-principal-component-analysis
- https://www.nytimes.com/interactive/2016/05/04/upshot/electoral-map-trump-clinton.html
- https://www.forbes.com/sites/katiasavchuk/2014/07/09/are-americas-richest-families-republicans-or-democrats/?sh=12c46ef73e83
- https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063
- Data: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjMs7agzJjvAhVfJzQIHSIgAu8QFjAAegQIBBAD&url=https%3A%2F%2Fwww2.census.gov%2Fprograms-surveys%2Fcps%2Ftables%2Ftime-series%2Fhistorical-income-households%2Fh08.xls&usg=AOvVaw0MqyTNSJP8VVd2wWKeoHeo