PCA Group Literature Review

Curtis Fox, Eddie Pierce, and Andrew Mercer

Principal Component Analysis (PCA) is a widely used statistical method for reducing the dimensionality of multivariate data. The goal of PCA is to transform the original variables into a new set of uncorrelated variables called principal components. This literature review aims to explore the methods and applications of PCA used in previous data analysis. While we will review the use of PCA in several analytical fields, our use of PCA focuses on National Football League data. Our analysis can be found below.

In the simplest terms, PCA acts to rotate the axis of different variables to better understand the variance in the data. The variables with the most important relationships are kept as they explain the most variance in the dataset. In this way PCA is used to decrease the dimensionality of a dataset. The first Principal component is calculated by adding the variables once they have been multiplied by a weighted value. The next Principal component is calculated using another variable as long as that variable is uncorrelated to the first. This continues until the number of Principal Components equals the number of variables or until you have reached a determined cutoff point (Holland 2008). The four options provided by Holland to be used as a cutoff point are if the increase of variance explained from one variable to the next is minimal. The second is to set a limit for variance explained, possibly 90%, and ignore all PCs after this limit is reached. The third is to use a correlation matrix and the variance is less than 1. Finally, the fourth, is to ignore the last PCs in the list if their variance explained is equal. In all cases the distribution of the variables data needs to be normally distributed for PCA to work correctly.

PCA is a powerful tool that can be applied to seemingly any field. Our review has found the application of PCA in remote sensing, public health, finance, and sports. In a paper on remote sensing by Chavez (1989), the author proposed a method for extracting spectral contrast from Landsat Thematic Mapper Image data. Chavez used Selective Principal Component Analysis (SPCA), a modification of the traditional PCA. This modification was used to extract principal components that are more informative for specific image analysis applications. The author demonstrated that SPCA can improve the discrimination between land cover classes in remote sensing applications. In the field of public health Nguessan et al. (2019) performed a systematic review and meta-analysis of morbidity and mortality among the United States homeless population using PCA. The authors applied PCA to a set of variables related to health outcomes among homeless individuals. The study showed that PCA can be a useful tool for summarizing complex health data and identifying the most important factors that contribute to health outcomes. Oualy (2021) investigated the relationship between income inequality and socio-political instability in sub-Saharan Africa using PCA. The author used PCA to construct an index of socio-political instability based on political violence, government stability, and human rights violations. The study demonstrated that PCA can be an effective method for combining multiple indicators into a single index and identifying the underlying factors that contribute to socio-political instability. In the field of Finance, Fang et al. (2018) proposed a stable systemic risk ranking method for China's banking sector based on PCA. The authors applied PCA to a set of variables related to the financial performance of Chinese banks and identified the most significant principal components that explain the variance in the data. The study showed how PCA can be used to measure systemic risk in the banking sector and to identify the most important factors that contribute to financial stability. Finally, PCA is often used in Climate and weather analysis. Jolliffe 1990 used PCA to reduce dimensionality related to meteorology and climatology. The author outlines the importance of interpreting the PCA results in the context of the original data used. In another article related to climate, Horel (1984) used PCA for certain geophysical phenomena where cross-spetral analysis comes up short. The author used complex time series data to compute complex PCs from cross-covariance and cross-correlation matrices.

The examples above show the use of PCA in several fields. Our focus is the use of PCA in sports. Due to the complex nature of sports statistics, PCA has been used by many authors in the past to reduce the dimensionality of sports data sets. Moura et al. (2014) analyzed European football game-related team statistics using PCA and clustering analysis. The authors applied PCA to a set of variables related to offensive and defensive performance and identified the most significant principal components that explain the variance in the data. The study showed that PCA can be a useful tool for summarizing complex sports data and identifying the most important factors that contribute to overall team performance. An article by Pino-Ortega et al. (2021) conducted a review of the most relevant variables for training design, performance analysis, and talent identification in European football, basketball, and rugby using PCA. The authors applied PCA to a set of variables related to physical fitness, technical skills, and tactical performance and identified the most significant principal components that explain the variance in the data. The study showed that PCA can be a useful tool for summarizing complex sports data and identifying the most important factors that contribute to performance. Like Pino-Ortega’s article, Jones 2016 used PCA to analyze individual player’s performance in the United States College Football League. The author used Principal Components Analysis to try to find a better statistical method to rank professional quarterback draft prospects than the commonly used Quarterback Rating. The target variable was log transformed. After PCA, the model included only two variables after starting from a very large dataset.

Our interest was also in sports data. We decided to use a dataset using the American National Football League for our analysis. This too has also been used many times in previous studies. Watnik et. al. 2001 used a PCA analysis completed on an NFL dataset from the year 2000 regular season. PCA was used to determine each team’s strength. The PCA reduced the dataset to 2 Principal components. The authors found their best regression, with an R square value of 83% when using only the AFC teams. Using all teams and all variables resulted in a R square of only 21%.

While football has remained one of the most watched sports for some time, the popularity of fantasy sports has exploded more recently. To capitalize on fantasy sports, several betting companies have started hosting fantasy football competitions. With these competitions comes the use of analysis to find a competitive edge. With the large datasets involved Sugar et. al. attempts a PCA to reduce dimensionality and extract the first 2 principal eigenvectors. The author then projected the 7-dimensional RB and WR scaled data onto the reduced 2-dimensional PCA space. Visualizing the PCA the author sees that the distribution of WRs and RBs suggests that the x-axis roughly corresponds to a player’s receiving ability, while the y-axis corresponds to their rushing ability. When used with clustering analysis, the PCA values show a way to generate more effective subsets without the risk of overfitting the models. The authors best results show that they were able to reduce the training set for the model by a factor of 3 without significantly increasing variance.

In conclusion, PCA is a powerful statistical method for reducing the dimensionality of multivariate data and identifying the most significant factors that contribute to the variance in the data. This literature review demonstrated the wide range of applications of PCA in various fields, including remote sensing, public health, sports, and finance. PCA can be a useful tool for summarizing complex data, identifying the most important factors that contribute to the outcomes of interest, and providing insights for decision-making.