How is the market value of a player determined?

Alan. Rojas-López[1] and Andrea Ruiz-Alvarez[1]

[1] *Tecnologico de Monterrey, Campus Guadalajara, Escuela de Ingeniería.*

Publication date: 24/11/2023

Abstract— In Tthe realm of football analytics , this study employs Principal Component Analysis (PCA) to unravel the intricate tapestry of player statistics within the La Liga database and discern their correlation with market values. The primary objective is to distill the multitude of performance metrics into a concise set of principal components that encapsulate the essence of a player's contribution on the pitch. Through rigorous analysis, the research explores the potential existence of a minimum correlation between these principal components and the market value of La Liga players.By leveraging PCA, the researchers reduce the dimensionality of the dataset, transforming diverse player statistics—such as assists, red and yellow cards, minutes played, Man of the Match recognitions, shots per game, passing accuracy, aerial duels won, ratings, and substitute appearances—into a compact representation. This reduction not only facilitates a clearer interpretation of player performance but also enables the identification of underlying patterns that may influence a player's market value. Preliminary findings suggest that the application of PCA in the La Liga player database yields a set of principal components exhibiting minimal correlation with market values. This discovery challenges conventional wisdom and prompts a reevaluation of the traditional metrics used to assess a player's worth. The implications of such a revelation extend beyond statistical nuances, offering insights that could redefine the criteria for evaluating player contributions and negotiating transfer values in the football industry. As the study unravels the implications of this minimum correlation, it contributes to the evolving landscape of football analytics by introducing a nuanced perspective on the relationship between player statistics and market values. The findings not only advance the understanding of player valuation in La Liga but also pave the way for future research that explores alternative metrics and methodologies in the realm where the beautiful game intersects with quantitative analysis.

Keywords— PCA, linear regression, eigenvalues, eingenvectors, MOTM, pearson coeffficent

I. Introduction

Footbal , beyond its role as a global sport, is a complex ecosystem where players are not only athletes but valuable assets strategically maneuvered in the competitive landscape of various leagues. La Liga, Spain's premier football division, stands as a microcosm of this dynamic environment, boasting a rich history, unparalleled skill, and a fervent fan base. In this report, we delve into the intricate web of factors that contribute to the market value of La Liga players, employing mathematical analysis to dissect and comprehend the nuances underlying their worth.

La Liga: A Sporting Tapestry:

La Liga, officially known as the Primera División, represents the pinnacle of Spanish football. Renowned for its technical finesse, tactical brilliance, and captivating matches, La Liga has been the breeding ground for some of the world's most iconic footballers. With clubs like FC Barcelona and Real Madrid leading the charge, the league has become a global spectacle, drawing in fans and talent from every corner of the world.

Market Value: Beyond the Pitch:

In the realm of football, a player's market value transcends the boundaries of mere athleticism. It is a metric that encapsulates a multitude of variables, ranging from on-field performance to commercial appeal. But what constitutes this elusive market value? How do factors like goals, assists, and overall contribution on the pitch translate into a quantifiable figure? These are the questions that propel our exploration into the intricate arithmetic that underpins the financial valuations of La Liga players.

Goals: The Currency of Victory:

At the heart of football, goals stand as the ultimate expression of triumph. In La Liga, each goal is not just a point on the scoreboard but a testament to a player's skill, precision, and ability to outwit formidable opponents. Analyzing the distribution of goals across teams and players unveils strategic nuances, shedding light on the attacking prowess that often separates the victors from the vanquished.

Assists as a Measure of Impact:

In our analysis, we will dissect one key element of a player's performance: assists. An assist, often the unsung hero of a goal, goes beyond the scorecard, revealing the collaborative and strategic prowess of a player. By examining the data surrounding assists, we aim to unravel the correlation between these contributions and the overall market value of La Liga players. Are these metrics mere statistical artifacts, or do they serve as reliable indicators of a player's

Contact data: ,

financial worth?

Red and Yellow Cards: Balancing Aggression and Composure

Discipline is a critical aspect of a player's profile. Red and yellow cards, often overlooked in traditional analyses, provide insights into a player's temperament and the balance between aggression and composure. Understanding the disciplinary aspect is crucial for assessing a player's reliability and impact on team dynamics.

Minutes Played: Endurance, Consistency, and Influence

The time a player spends on the pitch is more than a mere statistic; it is a testament to their endurance, consistency, and influence on the game. Analyzing minutes played offers a nuanced perspective on a player's role within the team and their capacity to make a lasting impact.

Man of the Match: Consistent Excellence Recognized

The "Man of the Match" accolade is more than an honor; it signifies consistent excellence and game-changing performances. Identifying players who frequently receive this recognition provides a qualitative measure of their influence and ability to shine when it matters most.

Shots per Game: Unveiling Offensive Prowess

Beyond goal-scoring statistics, shots per game unveil a player's offensive intent and goal-scoring potential. This metric provides a quantitative measure of a player's goal-scoring threat, shedding light on their ability to create opportunities and convert them into tangible results.

Passing Accuracy: Precision in Distribution

In the intricate dance of football, passing accuracy is the conductor's wand. Analyzing passing accuracy goes beyond completion percentages; it delves into the precision with which a player distributes the ball, dictating the flow of the game and influencing team strategy.

Aerial Duels Won: Dominance in the Air

Aerial duels are physical battles that often sway the momentum of a game. The number of aerial duels won reflects a player's ability to assert dominance in the air, crucial in both offensive and defensive scenarios. This metric unveils the physicality and aerial prowess that contribute to a player's overall impact.

Rating: Comprehensive Assessment of Performance

Player ratings, synthesized from various performance aspects, offer a comprehensive assessment of a player's overall contribution. Understanding the factors that contribute to these ratings provides a holistic view of a player's impact, combining both quantitative and qualitative dimensions.

Substitute Appearances: Versatility and Impact off the Bench

The role of substitutes is often underrated. The number of substitute appearances unveils a player's versatility, adaptability, and impact coming off the bench. This metric provides insights into a player's ability to change the dynamics of a game, even when not starting.

As we embark on this mathematical exploration, we aim to unravel the intricate relationships between these key performance metrics and the market values of La Liga players. By dissecting these components, we endeavor to provide a comprehensive understanding of the mathematical underpinnings in the realm where the beautiful game meets analytical rigor.

The Intersection of Mathematics and Football:

Football, with its fluidity and unpredictability, may seem worlds apart from the precision of mathematics. However, in this report, we bridge these seemingly disparate realms, using mathematical tools and analytical frameworks to discern patterns, correlations, and hidden insights within the intricate fabric of La Liga football. Through this lens, we embark on a journey to demystify the variables shaping the market value of players, offering a quantitative perspective on the beautiful game.

As we unravel the mathematical underpinnings of La Liga players' market value, we invite readers to join us in exploring the fascinating interplay between sport and analytics, where numbers become the silent narrators of a story written on the soccer pitch.

II. Objective

Our primary objective in this report is to unravel the intricate web of variables and components that significantly influence the market value of La Liga players, employing the powerful tool of Principal Component Analysis (PCA) as our guiding compass. In the kaleidoscope of football metrics, from goals and assists to possession percentages and player ratings, we aim to distill the essence of their collective impact on the financial valuations of players. PCA, a sophisticated statistical technique, allows us to sift through the myriad variables, identify patterns, and extract the principal components that carry the most weight in explaining the variance within our dataset. By doing so, we seek to uncover the underlying structures and relationships that define a player's market value in La Liga. Our analytical journey involves not only the application of PCA but also a meticulous interpretation of the results, teasing out the nuanced contributions of each variable. Through this methodical process, we aim to provide a comprehensive understanding of the key factors that elevate certain players to higher market values, thus contributing to the broader discourse on the intersection of mathematical analysis and the economics of football in La Liga.

In the format of this publication, the sections and subsections of the document are not numbered and added with the traditional commands of LaTeX, that is

III. Hypothesis

Based on our preliminary exploration of the intricate dynamics within La Liga, we hypothesize that there exists a significant and meaningful correlation between the various variables and components encapsulated in a player's statistical profile and their corresponding market value. We anticipate that individual performance metrics such as goals, assists, and player ratings, alongside broader team statistics like possession percentages, collectively contribute to the overall financial valuation of players in the league. This hypothesis stems from the inherent logic that a player's on-field contributions, whether in scoring goals, creating opportunities for teammates, or influencing the overall dynamics of a match, are likely to be reflected in their market value. We posit that through the application of PCA, we will be able to identify the principal components that bear the strongest correlation with market value, thereby shedding light on the essential

factors that drive the economic valuations of La Liga players. This hypothesis forms the foundation of our analytical approach as we embark on unraveling the intricate relationship between the quantitative dimensions of player performance and their corresponding market worth in the context of Spanish football.

IV. PCA

Principal Component Analysis (PCA): Unveiling Patterns in La Liga Player Valuation

Principal Component Analysis (PCA) stands as a powerful statistical method employed in multivariate analysis to uncover latent structures and patterns within datasets. In our context of exploring the variables contributing to the market value of La Liga players, PCA becomes an invaluable tool for dimensionality reduction and feature extraction. This technique enables us to transform a set of correlated variables into a new set of uncorrelated variables, known as principal components, ordered by their ability to explain the maximum variance in the original data. [1]

The fundamental goal of PCA is to identify the principal components that capture the most critical information within the dataset. These principal components are linear combinations of the original variables, ensuring that each subsequent component accounts for the maximum variance orthogonal to the preceding ones. Mathematically, the transformation of the original variables

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{21}X_2 + \ldots + a_{p1}X_p \\ Z_2 &= a_{12}X_1 + a_{22}X_2 + \ldots + a_{p2}X_p \\ &\vdots \\ Z_p &= a_{1p}X_1 + a_{2p}X_2 + \ldots + a_{pp}X_p \end{aligned}$$

Here, the coefficients a_{ij} are the loadings, representing the contribution of each original variable to the corresponding principal component. The key insight lies in the fact that the first principal component (Z_1) explains the maximum variance in the data, followed by the second (Z_2), and so on. [2]

Step 1: Standardize the Data

Given a dataset with n observations and p variables, let the data matrix be denoted as X with dimensions $n \times p$. The first step is often to standardize the data by subtracting the mean and dividing by the standard deviation for each variable:

$$Z = \frac{(X - \mu)}{\sigma}$$

Where:

- Z is the standardized data matrix.
- X is the original data matrix.
- μ is the mean vector (size p).
- σ is the standard deviation vector (size p).

Step 2: Compute the Covariance Matrix

Next, calculate the covariance matrix C of the standardized data:

$$C = \frac{1}{n-1} Z^T Z$$

Where:

- C is the covariance matrix.
- Z^T is the transpose of Z.

Step 3: Calculate Eigenvectors and Eigenvalues

Compute the eigenvectors V and eigenvalues Λ of the covariance matrix C. The eigenvectors represent the directions of maximum variance, and the eigenvalues indicate the magnitude of variance in those directions.

$$CV = \Lambda V$$

Where:

- V is a matrix of eigenvectors.
- Λ is a diagonal matrix of eigenvalues.

Step 4: Choose Principal Components

Sort the eigenvalues in descending order and choose the top k eigenvectors to form the matrix P, where k is the desired number of principal components.

$$P = [v_1, v_2, \ldots, v_k]$$

Step 5: Project the Data onto Principal Components

Project the standardized data Z onto the matrix P to obtain the transformed data matrix Y:

$$Y = ZP$$

Where:

- Y is the transformed data matrix.

Result

The matrix Y represents the dataset in a new space where each column is a principal component. These components are orthogonal (uncorrelated) and capture the maximum variance in the data.

In our analysis, we will apply PCA to the dataset comprising La Liga player statistics, seeking to identify the principal components that most significantly influence market value. The resulting principal components will serve as our refined set of variables, allowing us to focus on the essential dimensions of player performance that contribute to their financial valuation. Through this method, PCA provides not only a means of simplifying the analysis but also a powerful lens for uncovering the underlying structures that connect the diverse variables within the intricate realm of La Liga football.

V. Application of PCA on La Liga Player Dataset: Unraveling Market Value Factors

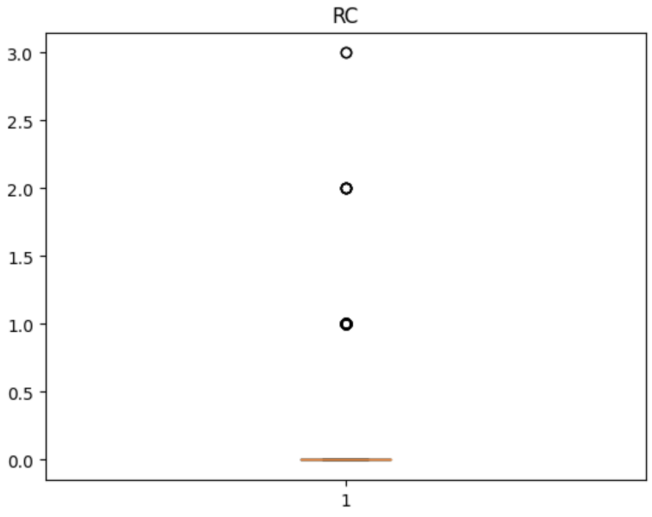
Implementation of PCA on La Liga Player Dataset

Application of PCA on La Liga Player Dataset: Unraveling Market Value Factors

In our pursuit of understanding the intricate factors influencing the market value of La Liga players, we undertook a systematic application of Principal Component Analysis (PCA). This method allowed us to distill complex player statistics into essential components, offering insights into the nuanced relationships within the dataset.

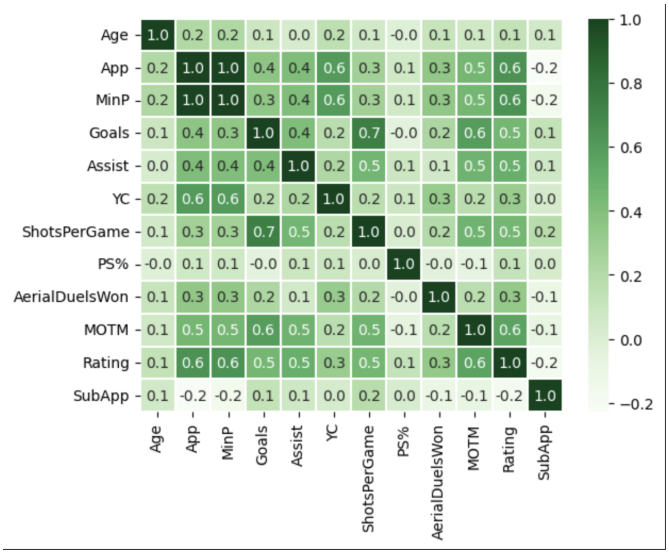
1. **Data Cleaning and Outlier Elimination:** Our journey commenced with meticulous data cleaning, addressing missing values and eliminating outliers to ensure the robustness of our analysis. Particular attention was paid to the identification and removal of outliers that could potentially skew our results.

2. **Variable Analysis and Red Cards Elimination:** Subsequent to data cleaning, we analyzed variable distributions through box plots. Notably, we observed a lack of clear evidence pointing to a single factor predominantly influencing player market value. Given this, we entertained the hypothesis that a combination of factors collectively shapes market value. Consequently, we eliminated the 'Red Cards' variable, deemed non-informative due to negligible variance.



3. **Consideration for PCA:** With attributes seemingly unrelated to market value, we recognized the potential applicability of PCA. This technique, designed to uncover patterns in high-dimensional data, became a suitable avenue for our exploration.

4. **Correlation Matrix and Heatmap Visualization:** To understand the interrelationships between variables, we constructed a correlation matrix and visualized it through a heatmap. The analysis revealed a 1-1 correlation between 'Minutes Played' and 'Appearances,' leading us to eliminate 'Appearances' seeing they were not giving any new important information.



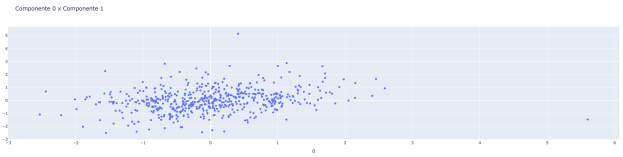
5. **Standardization of Data:** Standardizing all data became imperative to ensure that each variable contributed proportionately to the PCA, maintaining the integrity of the analysis.

6. **Eigenvectors and Eigenvalues Calculation:** The core of PCA lies in determining eigenvectors and eigenvalues. These vectors provide the direction of components carrying information, with higher eigenvalues signifying increased relevance. By graphing the percentage of eigenvalues, normalized by the sum of all values, we identified the most informative components. [3]

7. **Principal Component Analysis:** After selecting the relevant eigenvectors, we proceeded to create a dataframe of these principal components, assigning names and providing interpretations based on the information they encapsulate. - The most relevant component is going to be called youth. Why? Age is the heavier variable that explains this component; followed by negative rating; followed by the yellow ca ards which affect negatively. This makes sense because young players tend to have a high market value. And, players that perform badly (negative rating), are not as valuable. -The second component will be called the supersub. This is because the components that affect the most are subapp; yellow cards, and rating. Both subbApp and rating are of high relevance; meanig that substitute players that perform well on the pitch, are more relevant. -The third component will be called: no.9. This is the pure deffinition of a classical 9 striker. Low shots per game, high goals and many aerial duels won. The definition of accuracy and efficiency in the goal making. -The fourth component is called: The veteran. The component that affects the most is MinP: meaning experience on the field, followed by goals, and aerial duels won (negative). - The fifth component is called the centre back. The components that affect the most are: AerialDuelsWon, Rating(negatively), MOTM (negatively) and assists (negatively). -The component number 6 is called the offensive: With assists, shots per game and MOTM (negative), being the principal components. -The last component is called the playmaker: With assists and rating being the more relevant components. This means we can explain up to 85% of data with this Principal Components. [4]

8. **Final DataFrame Creation and Linear Regression Analysis:** Multiplying the transposed eigenvectors by the

standardized dataset resulted in the final dataframe. However, scatterplots of the principal components revealed no discernible clusters. Subsequently, a linear regression analysis aimed at predicting the 'Market Value' using principal components only achieved a predictive capacity of 29%.

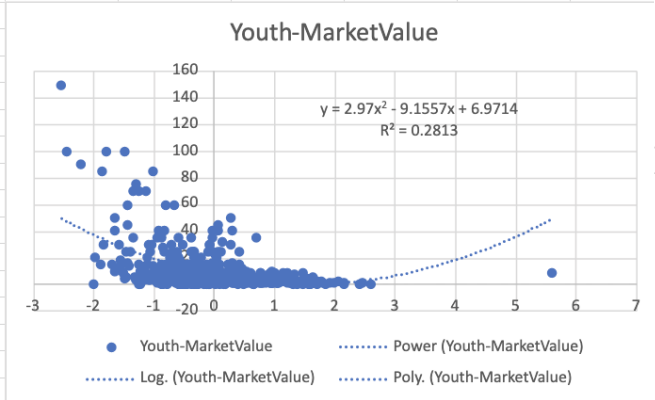


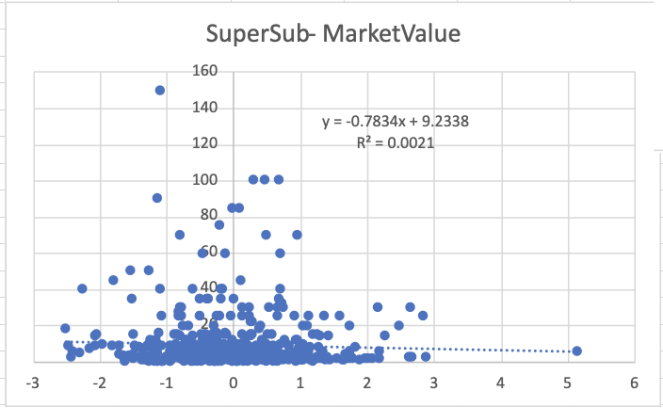
9. **Eigenanalysis for Predictive Modeling:** Given the limitations of linear regression, we proceeded to eigenanalysis, attempting to predict the 'Market Value' using a regression of principal components. This approach provides a nuanced understanding of how these components collectively influence a player's market value, transcending the constraints of linear regression. [5]

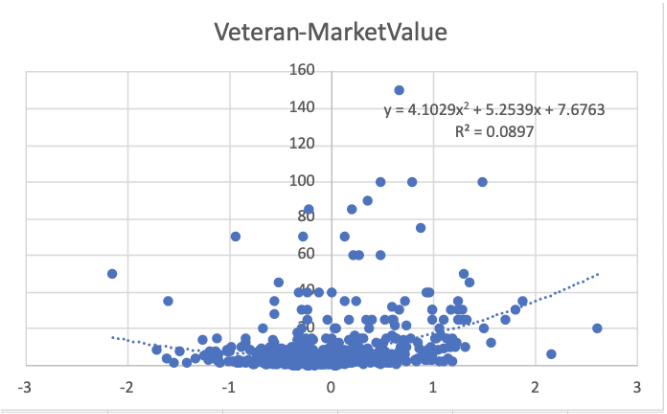
VI. Results

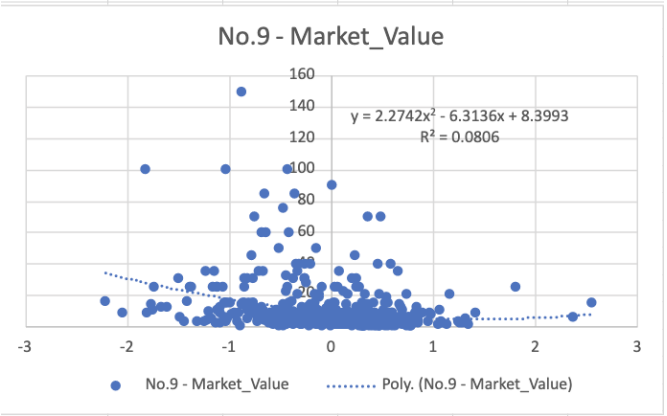
Making a multiple linear regression the Pearson coefficent is very low making it only 28% value can be explained through the variables. Intercept: 9.23376366743373 Coefficients: [-7.37457417 1.36062951 -1.80779013 2.45840802 -3.00277127 -1.11265624 2.15561661] 0.28710157914676604

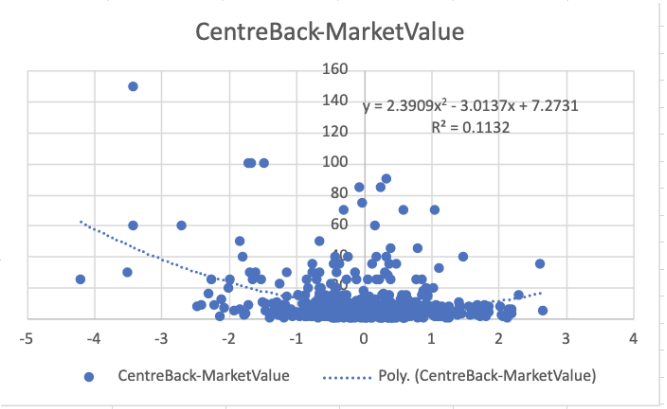
Analysing the linear regression with the database reduced by PCA, doing it with every component separated, s the Pearson coefficient is still very low in all of the variables. With this information it can be concluded that there is not a high correlation between the variables and the players market value.

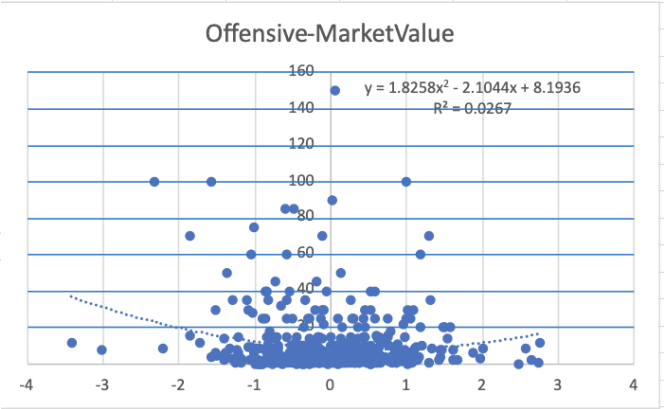












VII. Conclusions

Hence, with the presented evidence the conclusion that follows is that there is low evidence of a linear combination of the variables presented in the data-base that can explain the target variable. With PCA dimension reduction done, and linear regression with a target variable performed, the statistic fundamentals of our conclusions are set.

The low value R^2 (Pearson Coefficient) $= 0.2871$, suggests that only 28% of data can be explained with a the model: multiple linear regression with Principal Components targetting Market Value.

Based on the statistical evidence, FIFA is being called-on to reevaluate the way their players are Valuated.

References

[1] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, "Use of principal component analysis (pca) and hierarchical cluster analysis (hca) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective," *Trends in Food Science & Technology*, vol. 72, pp. 83–90, 2018.

[2] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009830049390090R

[3] P. Honeine, "An eigenanalysis of data centering in machine learning," 2014.

[4] Z. Liang and Y. Lee, "Eigen-analysis of nonlinear pca with polynomial kernels," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 6, no. 6, pp. 529–544, 2013.

[5] M. O. Faruqe and M. A. M. Hasan, "Face recognition using pca and svm," in *2009 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication*, 2009, pp. 97–101.