Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

ADVANCING DATA ANALYSIS THROUGH FUZZY PRINCIPAL COMPONENT ANALYSIS: A COMPREHENSIVE EXPLORATION IN CANCER DATA INSIGHTS

HASSANIA HAMZAOUI*
LPAIS, Faculty of Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
*Corresponding Author Email: hassania.hamzaoui@usmba.ac.ma

BOUCHRA DAOUDI
LPAIS, Faculty of Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
Email: bouchra.daoudi@usmba.ac.ma

MOUNIR GOUIOUEZ
LPAIS, Faculty of Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
Email: mounir.gouiouez@gmail.com;

Abstract

This article introduces a novel paradigm in quantitative data analysis by proposing an algorithm that integrates fuzzy set theory with the established Principal Component Analysis (PCA) methodology. This integration is designed to optimize and enhance the latter's ability to capture and interpret complex patterns within quantitative data that may exhibit varying degrees of fuzziness, and imprecision in real-world datasets. Fuzzy set theory, pioneered by Zadeh in 1965, provides a formalism for representing and manipulating uncertainty and imprecision in data. By incorporating this theory into the PCA algorithm, we seek to resolve the practical limitations posed by the deterministic nature of traditional PCA. The proposed algorithm's efficacy is evaluated using a real-world dataset focused on cancer, with results systematically compared against those obtained through the PCA algorithm. The assessment extends to two additional datasets concerning cardiovascular disease and diabetes, ensuring the generalizability and robustness of the findings. Empirical evidence substantiates that the Fuzzy Principal Component Analysis algorithm surpasses the PCA algorithm in terms of efficiency and performance. This superiority persists even when confronted with datasets of increased dimensionality. This research contributes to augmenting analytics capacity for handling imprecise and uncertain data, and substantiating these enhancements through empirical validation on diverse datasets.

Keywords: Fuzzy Set Theory, Fuzzy C-Means Clustering Algorithm, Membership Function, Fuzzy Principal Component Analysis.

1. INTRODUCTION

Descriptive statistics, a fundamental component in data analysis, comprises a suite of methodologies designed to elucidate, simplify, summarize, and organize complex datasets. Principal Component Analysis (PCA), a prominent technique among descriptive statistical methods, stands as a venerable approach established in 1933, persisting across diverse scientific domains due to its efficacy in dimensionality reduction and visualization of multidimensional datasets [1–5]. The

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

continual application of PCA underscores its enduring significance in rendering intricate datasets more accessible.

Simultaneously, Fuzzy sets, introduced by Lotfi A. Zadeh, extend classical set notation, offering a formalized framework for representing and manipulating imprecise data. This extension facilitates the handling of uncertainties pervasive in real-life problems, making it an indispensable tool in contemporary data analysis [6, 7]. Furthermore, in the realm of big data analysis or situations where data contains uncertain or ambiguous information, the incorporation of fuzzy logic becomes paramount. Fuzzy logic allows for a more nuanced representation of data, acknowledging the inherent imprecision and vagueness often present in real-world datasets. The ability of fuzzy sets to capture and manage uncertainties provides a significant advantage in scenarios where traditional binary logic may fall short. It enables a more flexible and adaptive approach to data analysis, accommodating the inherent complexities of large and diverse datasets. By embracing fuzzy logic, analysts can better navigate the intricacies of uncertain information, leading to more robust and reliable insights in the face of the inherent uncertainties that characterize big data environments. The application of fuzzy logic, therefore, emerges as a crucial strategy for enhancing the accuracy and effectiveness of data analysis methodologies.

This paper delves into the integration of fuzzy set theory with Principal Component Analysis through Fuzzy Principal Component Analysis (FPCA) [8–11, 13, 17]. The integration of fuzzy logic with Principal Component Analysis (PCA) is motivated by a desire to address the inherent limitations of PCA when confronted with the intricacies of modern datasets. PCA, a widely used technique for dimensionality reduction and feature extraction, has proven effective in many scenarios. However, it comes with certain drawbacks that become more pronounced in the context of contemporary datasets. Another challenge with PCA is its reliance on linear combinations of variables, which may not adequately capture the non-linear relationships present in many real-world datasets. Fuzzy logic introduces a more flexible framework that can capture non-linear relationships and complex dependencies among variables, enhancing the capability of PCA to extract meaningful patterns from data that exhibit intricate and non-linear structures.

To empirically substantiate the efficacy of FPCA, we present a comparative analysis using three real-world datasets: The Wisconsin Breast Cancer dataset, the Cardiovascular Disease dataset, and the Diabetes dataset. Through rigorous experimentation, the paper delineates the superior performance of the FPCA algorithm.

The subsequent sections of the paper are organized as follows: Section 2 furnishes a concise overview of fuzzy sets as a generalization and extension of classical sets. Section 3 introduces Fuzzy Principal Component Analysis, elucidating its foundations in traditional PCA while incorporating principles of fuzzy logic, notably

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

the fuzzy c-means clustering algorithm. Section 4 expounds upon the advantages inherent in the proposed approach. Section 5 offers a detailed discussion of the experimental results derived from the analysis of three real-world datasets. The paper concludes in Section 6, summarizing key findings and delineating potential avenues for future research.

2. NOTION OF FUZZY SETS

The concept of the fuzzy set, introduced by Lotfi A. Zadeh in 1965, stands as a fundamental generalization of classical set theory. In contrast to the crisp, well-defined boundaries of classical sets, fuzzy sets provide a more flexible and inclusive frame-work to represent the inherent fuzziness and imprecision found in many real-world situations. Zadeh's motivation was rooted in the recognition that in everyday life, the distinctions between categories are often blurred, and objects may exhibit partial membership to multiple sets simultaneously.

The cornerstone of the fuzzy set theory lies in the use of membership functions, an extension of the characteristic functions employed in classical set theory. These membership functions serve as a mathematical tool to describe the degree or strength of membership of an element in a fuzzy set. Unlike the binary nature of classical sets, where an element either entirely belongs or does not belong to a set, membership functions in fuzzy sets allow for a continuum of gradations. This continuum reflects the varying degrees to which an element participates or belongs to a fuzzy set, capturing the nuanced and gradual nature of real-world concepts. It illustrates the gradations in the membership of an element to a subset. Given a subset S of the reference set X, a fuzzy subset S is defined as the set of pairs:

$$S = \{(s, \mu_S(s)), s \in X\}$$

Where $\mu_S(x) \in [0, 1]$ represents the degree of membership of the element s of X in S.

In the literature [6, 19], there are several types of membership functions, such as:

- the triangular fuzzy membership function;
- the trapezoidal fuzzy membership function;
- the Gaussian fuzzy membership function;
- etc.

Fuzzy logic stands as the fundamental underpinning for the practice of fuzzy reasoning, providing a robust framework for drawing conclusions when faced with uncertainty. Its significance becomes particularly pronounced in scenarios where information is inherently vague or incomplete. In contrast to traditional binary logic which insists on a rigid true or false categorization, fuzzy logic embraces the inherent shades of uncertainty that are pervasive in real-world datasets. This flexibility allows for a more nuanced and realistic representation of the complexities

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

present in the information landscape. The adaptability of fuzzy logic is key to its effectiveness. By accommodating the gradations of uncertainty, it sidesteps the limitations of an oversimplified true/false dichotomy. This adaptability enables a simpler and more intuitive method for deriving conclusions from information that lacks absolute clarity. In the face of intricate and uncertain data, fuzzy logic becomes a powerful tool, facilitating more informed decision-making by providing a structured yet flexible approach to reasoning. Essentially, it acts as a navigational guide through the complexities of uncertain data, allowing for a more comprehensive and nuanced understanding of the information at hand.

3. PRINCIPAL COMPONENT ANALYSIS VS FUZZY PRINCIPAL COMPONENT ANALYSIS

3.1 Principal Component Analysis

The factorial method, principal component analysis (PCA) [1–5], consists in reducing the dimension of the quantitative data space while preserving the maximum quantity of information. This multidimensional method was proposed in 1933 by Hotelling.

The steps of PCA can be summarized as follows:

Step 1: Data standardization;

Step 2: Calculate of correlation matrix C;

Step 3: Calculate of factorial axes from the C matrix;

Step 4: Projection of data onto factorial planes;

Step 5: Interpretation of results;

This method is highly and extremely applicable for data description and redimensioning, but it requires a few improvements to make it more robust. We therefore propose to introduce fuzzy set theory concepts into the PCA algorithm.

3.2 Fuzzy Principal Component Analysis

FPCA [8–11, 13, 17] is a fuzzified or a fuzzy version of PCA, in which the principal components are extracted taking into account the degree of membership of the samples. In other words, FPCA is a combination of fuzzy logic and a principal component analysis algorithm. This logic is used in the Fuzzy C-Means (FCM) algorithm [12, 15, 20], which is one of the most widely used fuzzy clustering methods. The main objective of FCM in FPCA is to find the optimal fuzzy partitioning (the membership degrees), which is obtained by minimizing the following function:

$$J_m(U,V) = \sum_{a=1}^{n}\sum_{b=1}^{c}(u_{ab})^m \|x_a - v_b\|^2 \qquad (1)$$

Where

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

- $X = (x_1, x_2, \ldots, x_n)$ is the dataset.
- $V = (v_1, v_2, \ldots, v_c)$ are the cluster centres', v_b is the b^{th} cluster center.
- $U = (u_{ab})_{n\times c}$ is a fuzzy partition matrix, $u_{ab} \in [0,1]$ is the membership degree of data point x_a to the fuzzy cluster *b.*
- $\|x_a - v_b\|$ is the Euclidean norm between x_a and v_b.
- $m > 1$ fuzziness parameter, it controls the fuzzy degree of membership of each data.

3.3 Methodology of FPCA

The implementation of FPCA on a dataset requires the following steps:

- Before processing data, it is essential and important to understand and examine the data to be analyzed.
- Data pre-processing phase which is the essential step in the overall data analysis
- Process. In our case, we need two tasks: data cleaning to denoise data and solve or resolve the problem of missing data, and data transformation, including attribute or data type transformation, dataset scaling and data standardizing or normalization.
- The aim of using the Fuzzy C-Means (FCM) algorithm is to extract the degree of
- Membership of samples (step 1 to step 4 in the algorithm 3.4).
- Finally, we proceed to the last step based on the Fuzzy C-Means algorithm (step 5 and step 6 in the algorithm 3.4).

We summarize the main steps of the method in the figure 1:

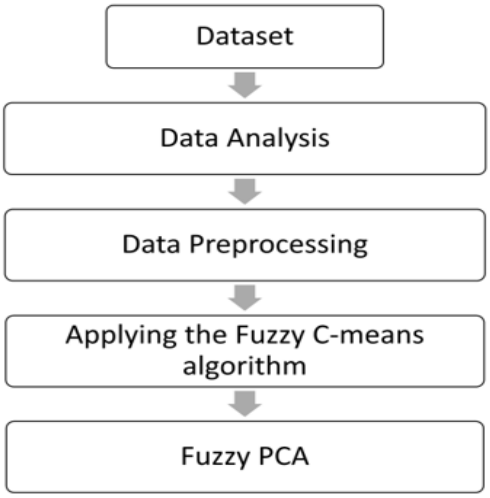


Fig 1: General steps of the FPCA method

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

3.4 Algorithmic steps for FPCA

Firstly, we fix c the number of clusters, tolerance value ε, and choose the fuzziness parameter m.

- **Step 1:** Initialize all membership degrees u_{ab} with random values ranging between 0 and 1 such that:

$$\sum_{b=1}^{c} u_{ab} = 1, \quad \forall a = 1, \ldots, n. \tag{2}$$

- **Step 2:** Calculate the cluster centers by the formula:

$$v_b = \frac{\sum_{a=1}^{n} u_{ab}^{m} . x_a}{\sum_{1=a}^{n} u_{ab}^{m}}, \quad b = 1,2, \ldots, c. \tag{3}$$

- **Step 3:** Update the memberships matrix such they satisfy the constraint (2) by the formula:

$$u_{ab} = \left(\sum_{k=1}^{c} \left(\frac{\|x_a - v_b\|}{\|x_a - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1}, a = 1, \ldots, n \text{ and } b = 1, \ldots, c. \tag{4}$$

- **Step 4:** Repeat steps 2 and 3 until the algorithm converges, i.e., the difference between the current and previous membership matrix is less than the tolerance value ε, or the number of iterations reaches a maximum value.

- **Step 5:** Calculate the fuzzy covariance matrix C using the membership matrix determined above.

$$C_{kl} = \frac{\sum_{a=1}^{n} (x_{ak} - \bar{x}_k)(x_{al} - \bar{x}_l) u_{ak}^{m} u_{al}^{m}}{\sum_{a=1}^{n} u_{ak}^{m} u_{al}^{m}}. \tag{5}$$

  Where u_{ak} and u_{al} are the membership degrees of the data x_k and x_l, respectively.

- **Step 6:** Determine the eigenvalues and eigenvectors of the fuzzy covariance matrix C as usual; these are the fuzzy principal components and the corresponding dispersion values.

4. CONTRIBUTION OF THE PROPOSED APPROACH

Our research pioneers a groundbreaking approach that brings a substantial enhancement to the conventional Principal Component Analysis (PCA) method. This innovation stems from the integration of fuzzy set theory, giving rise to what we term Fuzzy Principal Component Analysis. In outlining the key contributions of our approach, we unveil a methodology that surpasses the limitations of traditional PCA and offers a more refined and adaptable framework for data analysis. The synergy of fuzzy set theory and PCA introduces a novel perspective, allowing for a more nuanced representation of data, particularly in scenarios where uncertainties and imprecisions are prevalent. Through this integration, we aim to provide a robust and effective tool that advances the capabilities of PCA, fostering a deeper

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

understanding of complex datasets and opening new avenues for insightful analysis. The key contributions of our approach are outlined below:

4.1 Robust Dimensionality Reduction

PCA has long been a powerful tool for dimensionality reduction, allowing researchers to explore complex datasets while retaining crucial information. However, our Fuzzy Principal Component Analysis takes this a step further by incorporating fuzzy set theory, providing a nuanced understanding of data imprecision and uncertainty.

By considering the degree of membership of samples, FPCA ensures a more robust dimensionality reduction, preserving essential information even in the presence of fuzzy and imprecise data.

4.2 Overcoming Practical Limitations

Traditional PCA methods, while effective, exhibit limitations such as data loss and poor linear combinations. FPCA, by leveraging fuzzy logic and the Fuzzy C-Means (FCM) algorithm, overcomes these practical challenges. The incorporation of fuzzy clustering allows for a more adaptive and nuanced analysis, leading to improved accuracy and performance, especially when confronted with datasets of increased dimensionality.

4.3 Versatility in Handling Diverse Datasets

In an era of exponential growth in data availability, our approach recognizes the need for flexible methodologies that can handle diverse datasets. The fusion of statistical methods with fuzzy set theory provides an adaptable framework capable of addressing the uncertainty rates associated with modern datasets.

The empirical validation con- ducted on real-world datasets related to cancer, cardiovascular disease, and diabetes demonstrates the adaptability and robustness of Fuzzy Principal Component Analysis across various domains.

4.4 Advancements in Analytical Capacity

By augmenting PCA with fuzzy set theory, our research contributes to advancing analytical capacity in handling imprecise and uncertain data. The proposed FPCA algorithm showcases superior efficiency and performance compared to traditional PCA, making it a valuable tool for researchers and practitioners seeking enhanced insights from their datasets.

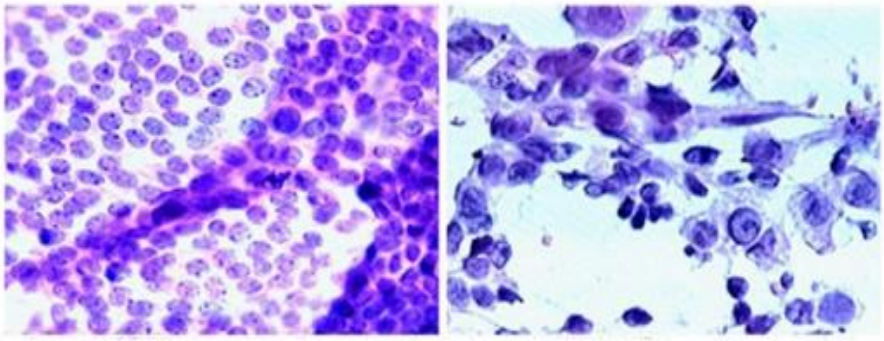
To sum up, our approach presents a significant step forward in the field of quantitative data analysis, offering a more sophisticated and adaptable method for capturing and interpreting complex patterns within real-world datasets. The empirical evidence substantiates the efficacy of Fuzzy Principal Component Analysis, reinforcing its potential as a valuable addition to the analytical vision in diverse fields.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

5. RESULTS AND DISCUSSION

To illustrate the performance of FPCA described above, we use the Breast Cancer Wisconsin dataset [2] [14, 21–23]:

The Breast Cancer Wisconsin (Diagnostic) dataset from UCI machine learning repository, contains features extracted from the digitized image of a fine needle aspiration of a breast mass. They are used for extracting features from the cell nuclei present in the digitized images [14, 23]. Here is an example of a digitized image (figure 2) (colored parts correspond to cell nuclei).



Smear with BENIGN diagnosis - uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size

Smear with MALIGNANT diagnosis - nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size

Fig 2: Pictures captured from glass layers with breast mass smears obtained by the Fine Needle Aspirate (FNA)

This dataset contains 30 variables (features), 569 individuals (images) and two classes: Malignant (212 samples) and Benign (357 samples) signifying if a patient has breast cancer or not. Ten features are calculated for each nuclei (the size, the shape and the texture), and three values are determined for each feature: the mean value, the largest (or worst) value and the standard error [23].

a) Radius (mean of distances from center to points on perimeter)
b) Texture (standard deviation of grayscale values)
c) Perimeter
d) Area
e) Smoothness (local variation in radius length)
f) Compactness (perimeter / area - 1.0)
g) Concavity (severity of concave parts of contour)
h) Concave points (number of concave portions of contour)
i) Symmetry
j) Fractal dimension (” coastline approximation” - 1)

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

The objective is to visualize the two classes of cells (malignant and benign), but as these cells are characterized by 30 variables this visualization becomes very difficult (high dimension), so we use the multidimensional data analysis methods PCA and FPCA to reduce the dimension and to construct factorial plans where we can separate the two classes of cells.

The results obtained by applying the PCA and FPCA are presented in the tables 1 and 2.

Table 1: The first seven eigenvalues and their proportion for PCA

PCA			
Component	**Eigenvalue**	**Variability (%)**	**Cumulative variability (%)**
Dim1	13.28	44.27	44.27
Dim2	5.69	18.97	63.24
Dim3	2.81	9.39	72.63
Dim4	1.98	6.60	79.23
Dim5	1.64	5.49	84.72
Dim6	1.20	4.02	88.74
Dim7	0.67	2.25	90.99

Table 2: The first seven eigenvalues and their proportion for FPCA

FPCA			
Component	**Eigenvalue**	**Variability (%)**	**Cumulative variability (%)**
Dim1	26.75	89.19	89.19
Dim2	2.35	7.84	97.03
Dim3	0.83	2.78	99.81
Dim4	0.05	0.17	99.98
Dim5	1.36×10^{-3}	4.53×10^{-3}	99.99
Dim6	1.09×10^{-5}	3.65×10^{-5}	99.99
Dim7	5.95×10^{-8}	1.98×10^{-7}	99.99

According to table 1, we notice that when we apply classical PCA to the dataset, five principal components (Dim1 - Dim5) are found to describe almost and only 84.72 % of the total variance of the dataset. In other words, the cumulative proportion of the first five components is 84.72% and the proportion of information lost is 15%.

However, for FPCA, only two axes are used to present 97.03% of information (table 2). As a result, the components derived from FPCA explain a much greater proportion of the variance through just two axes than their conventional PCA counterparts.

The number of components required to present and visualize this data set for each method is therefore clear.

The PCA load diagram or the loadings scatter plot allows us to visualize the contribution of the original data to the principal components.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

Principal components are linear combinations of the original variables, constructed to maximize the variance of the data.

In this case, principal components are constructed from the characteristics of the cells (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension for the three values: the mean value, the largest (or worst) value and the standard error) in each image.

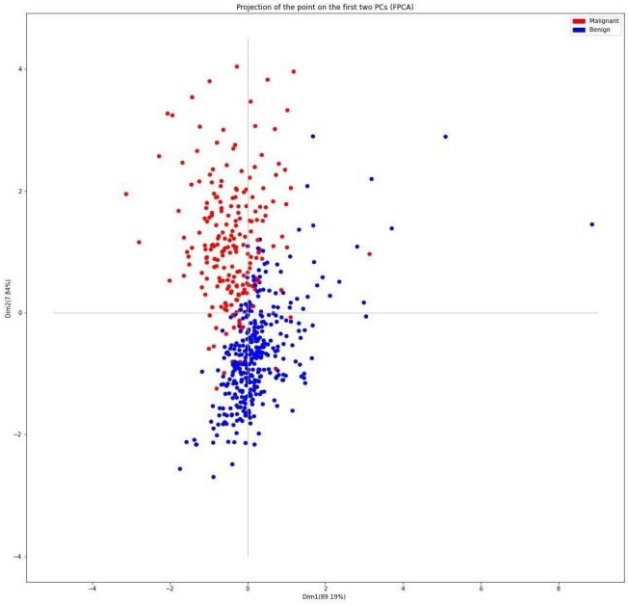


Fig 3: Scatter plot of loadings corresponding to the first two principal components (FPCA)

In the case of FPCA, the two classes are well presented on the first factorial plane in the score graph (figure 3) i.e., we can easily separate cancerous cells from non-cancerous cells by the first axis (Dim1).

Moreover, the two classes are homogeneous. So we can deduce that we need only one plan to present almost all information when applying FPCA.

In addition, the percentage of information lost is only 2.97% which means that this plan presents 97.03% of the information (we can mention that most of the variables contribute to the construction of the first two principal components).

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

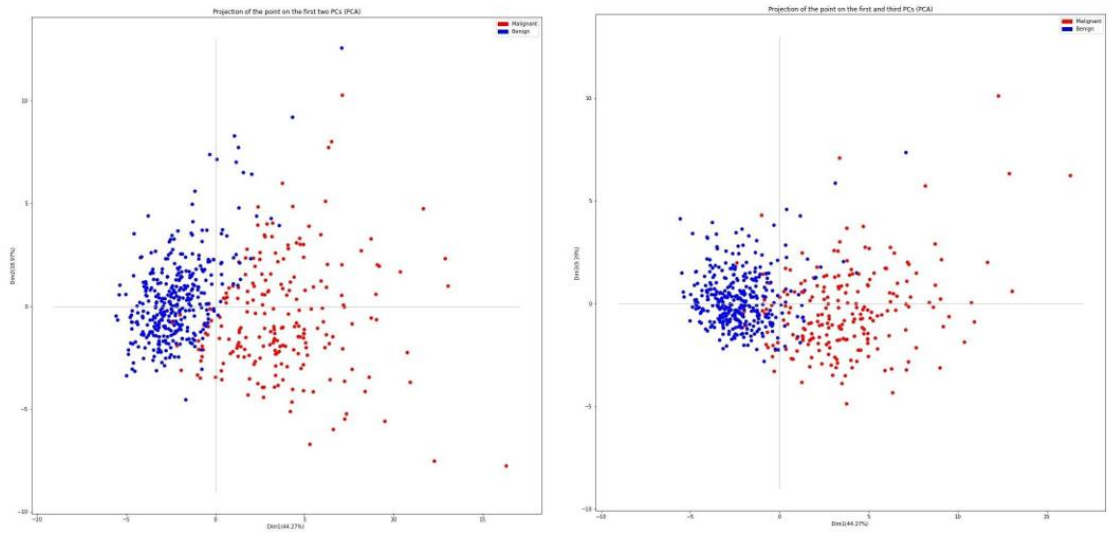


Fig 4: Scatter plot of loadings corresponding to the PCA (Dim1-Dim2 and Dim1-Dim3 factorial planes)

Contrary to the PCA, as we see in figures (4, 5, 6, 7 and 8) we need 10 graphics to visualize only 84.72 % of the information and this result is not satisfactory (for example the first factorial plane presents only 63.24% of the total information).

We remark also for PCA that in the first four planes, we can easily separate the two classes even though they are dispersed, but internally (figures 4 and 5).

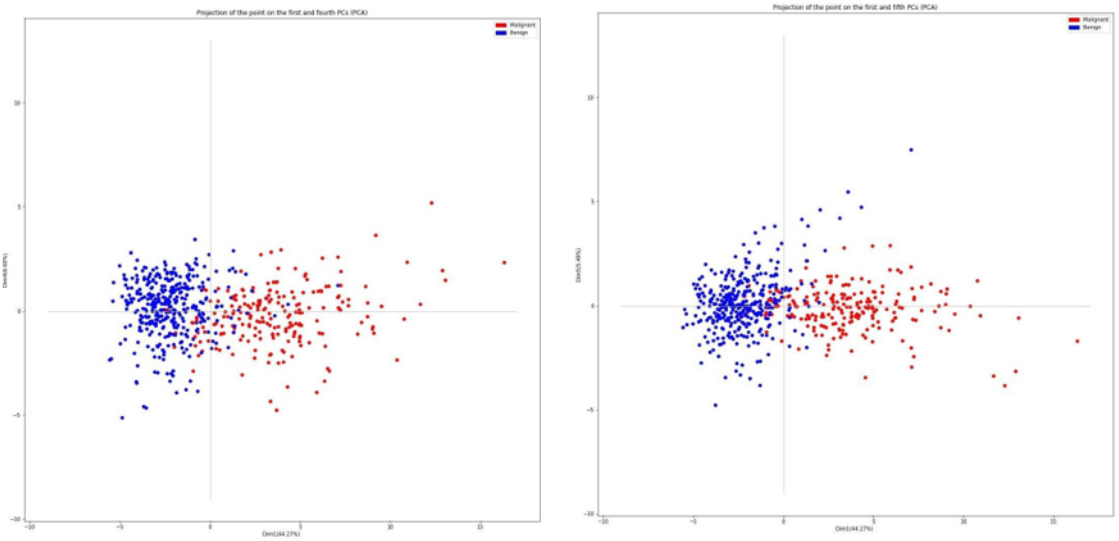


Fig 5: Scatter plot of loadings corresponding to the PCA (Dim1-Dim4 and Dim1-Dim5 factorial planes)

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

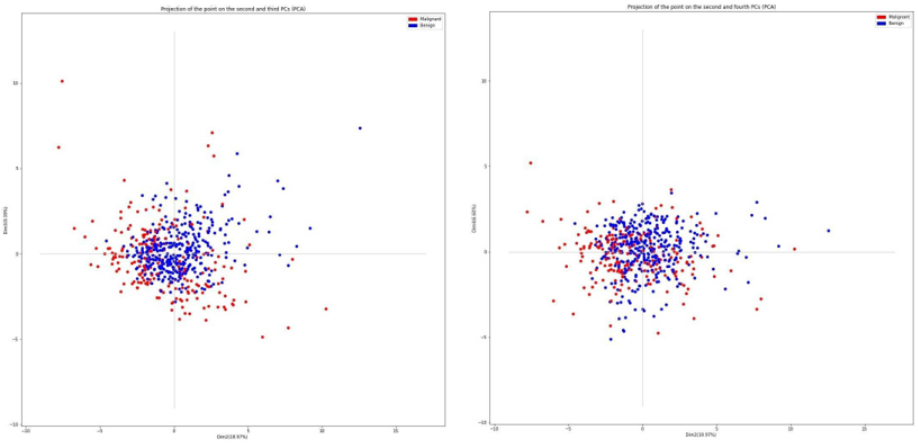


Fig 6: Scatter plot of loadings corresponding to the PCA (Dim2-Dim3 and Dim2-Dim4 factorial planes)

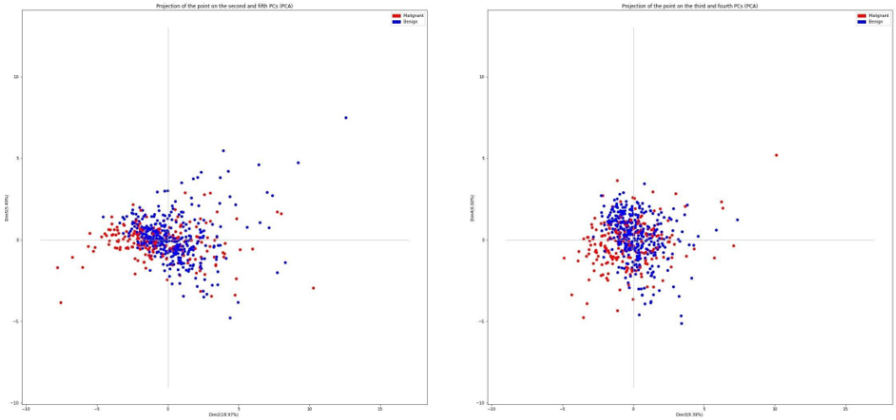


Fig 7: Scatter plot of loadings corresponding to the PCA (Dim2-Dim5 and Dim3-Dim4 factorial planes)

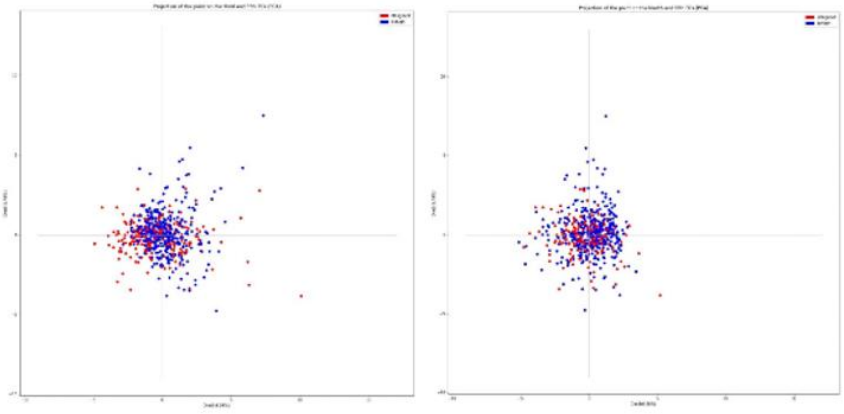


Fig 8: Scatter plot of loadings corresponding to the PCA (Dim3-Dim5 and Dim4-Dim5 factorial planes)

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

In the cases of figures 6, 7 and 8, it is hard and difficult or even impossible, to separate the two classes by eye. This is a poor representation of the data, as the two classes cannot be clearly visualized.

It should also be noted that all the data are concentrated on the origin, and the amount of information explained and represented is lower on the factorial planes: Dim2-Dim3 (8.5% of the information), Dim2-Dim4 (7.67% of the information), Dim2-Dim5 (7.33% of the information), Dim3-Dim4 (4.79% of the information), Dim3-Dim5 (4.45% of the information) and Dim4-Dim5 (3.62% of the information).

After analyzing the ten PCA graphs (for the load scatter plot and the correlation circle), we can see that all the information described therein can only be extracted from a single FPCA graph. So the performance and usefulness of the FPCA method for the visual synthesis of real data sets is clearly visible, as the analysis of the previous example shows.

To ensure and guarantee this result, we apply this algorithm to other real datasets.

For example, cardiovascular disease dataset and diabetes dataset.

1) Application of FPCA and PCA to cardiovascular disease dataset

We use the heart disease dataset created by the Cleveland, Hungary, Switzerland and VA Long Beach Institutes. It contains patient data, including characteristics that identify the presence of heart disease.

These data contain 15 variables (disease characteristics) from 500 patients (individuals).

The primary aim is to visualize two classes. The classes refer to a person's situation: does he/she suffer from cardiovascular disease or not (presence or absence of cardiovascular disease).

The tables 4 and 3 show the results obtained by applying FPCA and PCA.

Table 3: The first eight eigenvalues and their proportion for PCA

PCA			
Component	**Eigenvalue**	**Variability (%)**	**Cumulative variability (%)**
Dim1	2.95	19.67	19.67
Dim2	2.07	13.84	33.51
Dim3	1.76	11.74	45.25
Dim4	1.43	9.57	54.82
Dim5	1.27	8.49	63.31
Dim6	1.09	7.28	70.59
Dim7	0.97	6.50	77.09
Dim8	0.92	6.17	83.26

According to table 3 and table 4, we notice that when we apply classical PCA to the dataset, we find that the cumulative proportion of the first eight compo- nents is 83.26%. In other words, eight principal components (Dim1 - Dim8) are found (table 3) to describe only 83.26% of the information.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

Table 4: The first eight eigenvalues and their proportion for FPCA

FPCA			
Component	**Eigenvalue**	**Variability (%)**	**Cumulative variability (%)**
Dim1	14.71	98.13	98.13
Dim2	0.28	1.86	99.99
Dim3	5.39×10−16	3.59×10−15	99.99
Dim4	2.40×10−16	1.60×10−15	99.99
Dim5	1.79×10−16	1.19×10−15	99.99
Dim6	1.51×10−16	1.01×10−15	99.99
Dim7	1.21×10−16	8.11×10−16	99.99
Dim8	2.98×10−17	1.99×10−17	99.99

To analyze this amount of information, 28 factorial planes are required, which makes it very difficult and hard to interpret the results obtained.

For the FPCA, two factorial axes are sufficient to present almost all the information in the data table (99.99%). So, to interpret and visualize the analysis results, we simply need to use the first factorial plane.

2) Application of FPCA and PCA to diabetes dataset

The diabetes dataset comes from the National Institute of Diabetes and Digestive and Kidney Diseases.

768 women (individuals) aged at least 21, of Pima Indian origin, were tested for the disease diabetes. The presence of this disease was linked to 8 characteristics (variables) present in this dataset.

Even for this example, there are two classes of observations (presence or absence of diabetic disease):

The results obtained by applying the FPCA and PCA are presented in the tables 5 and 6.

Table 5: The first six eigenvalues and their proportion for PCA

PCA			
Component	**Eigenvalue**	**Variability (%)**	**Cumulative variability (%)**
Dim1	2.09	26.17	26.17
Dim2	1.73	21.64	47.81
Dim3	1.02	12.87	60.69
Dim4	0.87	10.94	71.63
Dim5	0.76	9.52	81.16
Dim6	0.68	8.53	89.69

According to table 5 and table 6, we can see that even for this example, only one factorial plane is needed to interpret 99.97% of the information for FPCA. However, for PCA, we need ten factorial planes to interpret just 81.16% of the information.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

Table 6: The first six eigenvalues and their proportion for FPCA

FPCA			
Component	**Eigenvalue**	**Variability (%)**	**Cumulative variability (%)**
Dim1	7.01	87.66	87.66
Dim2	0.98	12.3	99.97
Dim3	1.95×10^{-3}	2.44×10^{-2}	99.99
Dim4	7.53×10−5	9.42×10−4	99.99
Dim5	4.62×10−7	5.77×10−6	99.99
Dim6	1.49×10−10	1.84×10−9	99.99

In conclusion, we can see that FPCA always outperforms PCA, even when the number of variables in the data space is reduced (for example 1: 30 variables, for example 2: 15 variables and for example 3: 8 variables).

6. CONCLUSION

PCA is a method for the descriptive analysis of multidimensional data. It consists and used to minimizing the dimension of the data space in order to visualize them on factorial planes, but as the number of planes increases the interpretation of the results becomes more delicate, and this is an inconvenience of the PCA method.

We have proposed to introduce the fuzzy c-means algorithm to the PCA algorithm to achieve a more reduced and realistic analysis.

We applied FPCA and PCA to three different datasets in order to demonstrate the performance of the FPCA algorithm compared with the PCA algorithm.

Firstly, we considered the Wisconsin breast cancer dataset to show the performance of the FPCA algorithm versus the PCA algorithm. We found that the FPCA algorithm represents more than 95% of the information on a single factorial plane, whereas the PCA algorithm represents almost the same information on ten factorial planes.

Secondly, we used the cardiovascular disease dataset, and found that the FPCA algorithm represents than 99% of the information on a first factorial plane, whereas the PCA algorithm represents almost the same information on more than 28 factorial planes.

Thirdly, we tried the diabetes dataset, and again found that the FPCA algorithm represents over 99% of the information on a first factorial plane. As opposed to the PCA method, which requires ten factorial designs to present just 81 % of information.

We always find that the FPCA method outperforms the PCA method. This ensures the effect of fuzzy set theory in PCA.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

References

1) Sarker, I.H.: Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN COMPUT. SCI. **2**, 377 (2021).

2) https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

3) Jolliffe, I.T.; Cadima, J.: Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences **374**(2065), 20150202 (2016).

4) Tripathy, B.; Sundareswaran, A.; Ghela, S.: Principal component analysis (pca). In: Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization, 5-16. CRC Press (2021).

5) Deisenroth, M.P.; Faisal, A.A.; Ong, C.S.: Dimensionality reduction with principal component analysis. In: Mathematics for Machine Learning, 286-310. Cambridge University Press (2020).

6) Boukichou-Abdelkader, N.; Montero-Alonso, M..; Muoz-Garca, A.: Different routes or methods of application for dimensionality reduction in multicenter studies databases. Mathematics **10**, 696 (2022).

7) Mustafa B. Babanli and Jale M. Babanli.: Fuzzy Decision Method Based on Zadehs Data Aggregation Approach: 14th International Conference on Theory and Application of Fuzzy Systems and Soft Computing ICAFS-2020, Volume 1306, (2021).

8) Yue, W., Liu, X., Li, S., et al.: Knowledge representation and reasoning with industrial application using interval-valued intuitionistic fuzzy Petri nets and extended TOPSIS: Int. J. Mach. Learn. & Cyber.**12**, 9871013 (2021). https://doi.org/10.1007/s13042-020-01216-1

9) Gu, X., Han, J., Shen, Q., et al.: Autonomous learning for fuzzy systems: a review: ArtifIntell Rev **56**, 75497595 (2023).

10) Salgado, P.; Goncalves, L.; Igrejas, G.: Sliding PCA Fuzzy Clustering Algorithm. AIP Conference Proceedings **1389**(1), 1992-1995 (2011).

11) Hadri, A.; Chougdali, K.; Touahni, R.: Intrusion detection system using pca and fuzzy pca techniques. In: 2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS), 1-7 (2016)

12) Wu, H., Gu, X..: Fuzzy Principal Component Analysis Model on Evaluating Innovation Service Capability: Scientific Programming**9**, (2020).

13) Nascimento, S.; Mirkin, B.; Moura-Pires, F.: A fuzzy clustering model of data and fuzzy c-means. In: Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063), **1**, 302-307 (2000)

14) Gueorguieva, N.; Valova, I.; Georgiev, G.: Fuzzyfication of principle component analysis for data dimensionalty reduction. In: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1818-1825 (2016)

15) Sizilio, G.R.; Leite, C.R.; Guerreiro, A.M.; Neto, A.D.D.: Fuzzy method for pre- diagnosis of breast cancer from the fine needle aspirate analysis. Biomedical engineering online **11**(1), 1-21 (2012)

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301651

16) Li, Y.; Bao, T.; Shu, X.; Chen, Z.; Gao, Z.; Zhang, K.: A hybrid model integrating principal component analysis, fuzzy c-means, and gaussian process regression for dam deformation prediction. Arabian Journal for Science and Engineering **46**, 4293-4306 (2021)

17) Kiri sci, M.; Simsek, N.: A novel kernel principal component analysis with application disaster preparedness of hospital: interval-valued fermatean fuzzy set approach. The Journal of Supercomputing, 1-31 (2023)

18) Hefaidh, H.; Mbarek, D.: Using fuzzy-improved principal component analysis (pca-if) for ranking of major accident scenarios. Arabian Journal for Science and Engineering **45**, 2235-2245 (2020)

19) Tekin, A.: Classification of erythemato-squamous diseases using association rules and fuzzy c-means clustering. Arabian Journal for Science and Engineering **39**(6), 4699-4705 (2014)

20) Straccia, U.: Fuzzy sets and mathematical fuzzy logic basics. In: Foundations of Fuzzy Logic and Semantic Web Languages, 101-162. Chapman and Hall/CRC (2016)

21) Krasnov, D.; Davis, D.; Malott, K.; Chen, Y.; Shi, X.; Wong, A.: Fuzzy c-means clustering: A review of applications in breast cancer detection. Entropy **25**(7), 1021 (2023)

22) Ahmad, A.: Breast cancer statistics: recent trends. Breast cancer metastasis and drug resistance: challenges and progress, 17 (2019)

23) Alloqmani, A., Abushark, Y.B., Khan, A.I.: Anomaly detection of breast cancer using deep learning. Arabian Journal for Science and Engineering, 126 (2023)

24) Ahmad, F.K.; Yusoff, N.: Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. In: 2013 13th International Conference on Intelligent Systems Design and Applications, 121-125 (2013)