Journal of Mathematical Modeling
Vol. 13, No. 4, 2025, pp. 803-815. Research Article

JMM

Dimension reduction by identifying and removing redundant variables using copula function

Kianoush Fathi Vajargah[†*], Hamid Mottaghi Golshan[‡], Fazel Badakhshan Farahabadi[†]

[†]*Department of Statistics, Islamic Azad University, North Tehran Branch, Tehran, Iran*
[‡]*Department of Mathematics, Islamic Azad University, Shahriar Branch, Shahriar, Iran*
Email(s): K_fathi@iau-tnb.ac.ir, Ha.Mottaghi@iau.ac.ir, fazelbadakhshan@gmail.com

Abstract. In today's world, rapid developments in science and engineering are increasingly adding up to larger amounts of data; as a result, numerous problems have emerged in the analysis of big data. Hence, data dimensionality reduction can accelerate data analysis and even yield better results without losing any useful data. A copula represents an appropriate model of dependence to compare multivariate distributions and better detect the relationships of data. Therefore, a copula is employed in this study to identify and delete noisy data from the original data. Then, it is compared to the principal component analysis to show its superiority.

Keywords: Copula function, Gaussian copula function (normal), Classification, Principal component analysis, data analysis, Parkinson's disease.
AMS Subject Classification 2010: 62H05, 62H25, 62H20.

1 Introduction

Dimensionality reduction is an essential step in the data mining process, particularly for big data, and is especially important in the modern period due to the growing volume of data in scientific, industrial, and commercial domains. Reducing dimensionality not only helps classification algorithms perform better, but also simplifies and clarifies the data analysis process. Numerous techniques have been put out in recent decades to minimize the dimensionality of data. Principal component analysis (PCA), a feature extraction-based technique, is one of the most popular of these approaches. By identifying and eliminating components with lower variance, it reduces the dimensionality of the data. Because of its ease of use and effectiveness, PCA has been applied in several disciplines.

The PCA might not always be the best option for dimensionality reduction, though, particularly if the data has a complicated structure or if some of the components that have been eliminated have

*Corresponding author
Received: 16 August 2024 / Revised: 15 April 2025 / Accepted: 26 April 2025
DOI: 10.22124/jmm.2025.28169.2484

significant characteristics [18–20]. In order to reduce the dimensionality of data, several techniques have been employed in addition to PCA, including self-organizing mapping (SOM), multidimensional scaling (MDS), and linear discriminant analysis (LDA). The type of data and the application's goal determine which of these approaches is the best one. Each has benefits and drawbacks of its own. For further details on these techniques, see [3–5, 10, 11, 14, 16]. In this work, we provide a novel feature selection-based approach to data dimensionality reduction. Our suggested approach determines whether dimensions have a high structural correlation with one another by estimating the community parameters and using the detail function. It then reduces the dimensionality of the data and gets it ready for analysis by eliminating noise and superfluous dimensions.

This study proposes a method based on feature selection to identify the dimensions with highly structural correlations by using a copula and estimating population parameters. The proposed method then reduces data dimensionality and analyzes data by eliminating noisy data.

This study seeks to show how the proposed dimensionality method improves classification efficiency and facilitates data analysis. For this purpose, the novel dimensionality reduction method is first introduced and then compared with the PCA. After that, the data reduced through these two methods will be evaluated in naïve Bayes classification and k-nearest neighbour (KNN) techniques.

Advantages of the suggested approach are as follows:

- Accurate identification of related dimensions: By utilizing the detail function, our suggested approach may precisely identify the dimensions having a structural correlation with one another. This guarantees the preservation of significant data characteristics throughout the dimensionality reduction procedure [1].

- Elimination of noise and superfluous dimensions: Our suggested approach enhances data quality and boosts data mining algorithms' accuracy by eliminating noise and unnecessary dimensions [8].

- Analyzing data is made easier and more efficient with our suggested technique, which reduces data dimensions and eliminates noise and superfluous dimensions. We will look at the specifics of the suggested approach, its implementation, and the outcomes of tests on various data in the remaining sections of this article.

2 Preliminaries

The PCA is mainly employed to describe variations in a set of correlated variables $x^{'} = (x_1, \ldots, x_q)$ based on a new set of uncorrelated variables $y^{'} = (y_1, \ldots, y_q)$, in which every y_i represents a linear combination of x's. The new variables are considered in a descending order of importance. In other words, y_1 calculates the highest rate of variations in initial data among all linear combinations of x. After that, y_2 is selected to calculate the residual variations in a way that it is uncorrelated with y_1. The new variables defined in this process $(y_1, \ldots, y_d)$ are the principal components [1, 2, 8, 9, 20].

In the PCA, a small number of initial components are generally expected to explain a large proportion of variations in initial variables $x_1, \ldots, x_q$. Therefore, these variables are employed to prepare an appropriate summary with a lower dimension for various reasons [25].

2.1 Finding principal components of population

Assume that x is a random $q \times 1$ vector with a mean μ and a specific positive definite covariance matrix (Σ). It is also assumed that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$ (eigenvalues) [6] and that $H = [h_1, \ldots, h_q]$ is an orthogonal (i.e. $H^{'}H = HH^{'} = I_q$) $q \times q$ matrix such that

$$H^{'}\Sigma H = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_q) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_q \end{pmatrix},$$

which results from the spectral decomposition [7, 12].

Hence, h_i is the eigenvector of Σ corresponding to λ_i. The following equation is now considered

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_q \end{pmatrix} = H^{'}X = \begin{pmatrix} h_1^{'} \\ \vdots \\ h_q^{'} \end{pmatrix} X = \begin{pmatrix} h_1^{'}X \\ \vdots \\ h_q^{'}X \end{pmatrix} = \begin{pmatrix} h_{11}X_1 + h_{12}X_2 + \cdots + h_{1q}X_q \\ \vdots \\ h_{q1}X_1 + h_{q2}X_2 + \cdots + h_{qq}X_q \end{pmatrix}.$$

The components of $U_1, \ldots, U_q$ from U are called the principal components of X. In other words

$$\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_q \end{pmatrix} = \begin{pmatrix} l1.st - PC \\ 2.nd - PC \\ \vdots \\ q.th - PC \end{pmatrix}.$$

Moreover, $\operatorname{cov}(U) = \Lambda$; therefore, $U_1, \ldots, U_q$ are all uncorrelated, and $\operatorname{var}(U_i) = \lambda_i$ for $i = 1, \ldots, q$.

The first principal component is $U_1 = h_1^{'}X$ with the variance of λ_1; the second principal component is $U_2 = h_2^{'}X$ with the variance of λ_2; and so on.

In addition, the principal components have the following optimization features.

1. The first principal component (U_1) is a normal linear combination of components of X with the highest variance possible. This maximum variance is λ_1. Out of all normal linear combinations of components of X that are uncorrelated with U_1, the second principal component (U_2) has the maximum variance of λ_2. Generally, the kth principal component ($U_k : k = 1, \ldots, q$) has the maximum variance of λ_k out of all the normal linear combinations of components of X that are uncorrelated with $U_1, \ldots, U_{k-1}$ [20].

2. The ratio of the total variance for the kth component is as follows:

$$\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_q}, \qquad k = 1, \ldots, q.$$

3. If the highest variance of a population (e.g. 80–90%) is attributed to a few of the initial components for a large q, these components can then replace q initial variables without losing much information [1].

2.2 Copula

In brief, a copula is a function that links multivariate distributions to their marginal distributions. In other words, a copula is a multivariate distribution, the marginal distributions of which follow a normal distribution within (0, 1).

A copula is used for various reasons. First, it is a method for measuring the free-scale dependence. Second, it is a starting point for developing joint distributions with known margins. In fact, a considerable number of general studies on copulas analyze the dependence of random variables, for they allow us to distinguish between the dependence of variables and the effects of marginal distributions. This characteristic resembles the bivariate normal distribution where there are no links between its mean vector and its covariance matrix, both of which indicate the distribution simultaneously [17].

2.3 Main features of a Copula

Assume that $C: I^2 \to I$ has the following features:

1) For every $u, v \in [0,1]$, we have

$$C(u,0) = C(0,v) = 0, C(u,1) = u, C(1,v) = v.$$

2) For every $0 \le v_1 < v_2 \le 1, 0 \le u_1 < u_2 \le 1$, we have

$$C(U_2,v_2) + C(U_1,v_1) - C(U_1,v_2) - C(U_2,v_1) \ge 0.$$

Such a function like C implied in the two above conditions is called a copula function [15].

2.4 Sklar's theory

Assume that H is a joint probability distribution function with marginal distributions of F and G. Then C is a copula if the following equation is true for every $x, y \in \mathbb{R}$,

$$H(x,y) = C(F(x), G(y)).$$

If F and G are continuous, then the copula C is unique; otherwise, C is defined as unique on $\text{Rang}(F) \times \text{Rang}(G)$. Conversely, if C is a copula with marginal univariate distributions F and G, then H is a function with margins F and G.

According to Sklar's theory, if F and G have normal distributions, then $H(x,y) = C(x,y)$. It represents a copula of bivariate distribution with a normal marginal distribution within (0, 1). In other words, a copula is a bivariate distribution function with normal marginal distributions within (0, 1).

Assume that c, g, f, and h are density functions of distributions C, G, F, and H, respectively. Based on Sklar's theory, the following equation is true:

$$h(x,y) = c(F(x), G(y)).f(x).g(y),$$

where $c(u,v) = \dfrac{\partial^2 C(u,v)}{\partial u \partial v}$ [23].

The important application of a copula is to present an appropriate method for generating distributions of correlated random multivariate variables and offer a solution to the problem of density estimation conversion. To show the problem of reversible transforms of m-dimensional random continuous variables $X_1,\ldots,X_m$ based on their distribution function into m normal independent variables $U_1 = F_1(X_1),\ldots,F_m(X_m)$, it should be assumed that $f(x_1,\ldots,x_m)$ and $c(u_1,\ldots,u_m)$ are the density probability function of $x_1,\ldots,x_m$ and the joint density function of $U_1,\ldots,U_m$, respectively. The density probability function $c(u_1,\ldots,u_m)$ is estimated for $U_1,\ldots,U_m$ instead of $x_1,\ldots,x_m$ in this case to simplify the density estimation problem because the estimation of the density probability function $f(x_1,\ldots,x_m)$ may be a nonparametric form (i.e. an unknown distribution). The random samples $x_1,\ldots,x_m$ are then obtained by simulation using the inverse transform $X_i = F^{-1}(U_i)$.

The scalar field theory indicates that there is a unique m-dimensional copula in $[0,1]^m$ with standard normal marginal distributions $U_1,\ldots,U_m$, whereas Nelson stated that every distribution function F with margins $F_1,\ldots,F_m$ could be written as follows:

$$\forall(X_1,\ldots,X_m) \in \mathbb{R}^m, \qquad F(X_1,\ldots,X_m) = C(F_1(X_1),\ldots,F_m(X_m)).$$

To evaluate a copula selected with an estimated parameter and avoid defining any hypotheses on $F_i(X_i)$, the empirical distribution function of a marginal distribution $F_i(X_i)$ can be employed to transform m samples of X into m samples of U [17,24].

An empirical copula is useful for analyzing the dependence structure of a multivariate vector. It is officially defined as the following equation:

$$C_{ij} = \frac{1}{m}\left(\sum_{k=1}^{m} I_{(v_{kj}\leq v_{ij})}\right).$$

3 New method

A unique approach to dimensionality reduction of multidimensional data is proposed in this section. This technique estimates an infinite multivariate copula distribution in particular kinds of marginal distributions of random variables displaying data dimensions using the copula theory. A comprehensive and unscaled explanation of dependency is provided by a copula-based model. Comparing the dependency of random variables using this model may be made easier by estimating the copula parameters. This dependence is then used to clean the original data, removing any extraneous numbers or noisy data.

3.1 Gaussian copula

The difference between a Gaussian copula and a joint normal distribution is that the Gaussian copula allows us to have different types of a distribution function for a joint distribution. However, according to the probability theory, the multivariate normal distribution is the generalization of a one-dimensional normal distribution.

The standard multivariate Gaussian copula is defined as below:

$$c(\Phi(X_1),\ldots,\Phi(X_m)) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}X^T(\Sigma^{-1}-I)X\right),$$

where $\Phi(x_i)$ is the standard distribution of $f_i(x_i)$, whereas $X_i \sim N(0,1)$ and Σ are the correlation matrices. As a result, $c(u_1,\ldots,u_m)$ is called the Gaussian copula, and the joint density is obtained from the following equation:

$$c(u_1,\ldots,u_m) = \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\xi^T(\Sigma^{-1}-I)\xi\right],$$

where $u_i = \Phi(x_i)$ and $\xi = (\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_m))^T$ [21].

3.2 Rank correlations

Spearman's rho (ρ) and Kendall's tau (τ) are the two criteria for rank correlation [15]. For random variables X_1 and X_2, Spearman's ρ is defined as below:

$$\rho_s(X_1,X_2) = \rho(F_1(X_1),F_2(X_2)).$$

In fact, the correlation matrix of Spearman's ρ shows the correlations of Spearman's ρ in pairs. It is a positive and known matrix, which is shown as follows:

$$\rho_s(X_1,X_2) = 12\int_0^1\int_0^1 (C(u_1,u_2)-u_1u_2)du_1du_2.$$

Moreover, the following equation is true for the bivariate Gaussian copula:

$$\rho_s(X_1,X_2) = \frac{6}{\pi}\arcsin\frac{\rho}{2} \simeq \rho,$$

where ρ is the Pearson's linear correlation coefficient [24].

Regarding random variables X_1 and X_2, Kendall's τ is defined as follows:

$$\rho_\tau(X_1,X_2) = E\left[\text{sign}((X_1-\tilde{X}_1)(X_2-\tilde{X}_2))\right],$$

where $(\tilde{X}_1,\tilde{X}_2)$ is independent of (X_1,X_2) but has an equal joint distribution with it.

Furthermore, Kendall's τ can be written as follows:

$$\rho_\tau(X_1,X_2) = P((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)>0) - P((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)<0).$$

When $\rho_\tau(X_1,X_2)=0$, both of the above probabilities are equal.

If X_1 and X_2 have continuous margins, the following equation is true:

$$\rho_\tau(X_1,X_2) = 4\int_0^1\int_0^1 C(u_1,u_2)dC(u_1,u_2) - 1.$$

Moreover, the following equation can be written for a bivariate Gaussian copula:

$$\rho_\tau(X_1,X_2) = \frac{2}{\pi}\arcsin\rho,$$

where ρ is Pearson's correlation coefficient.

Spearman's ρ and Kendall's τ are instances of rank correlation, which depend only on a bivariate copula and not on margins; therefore, they are not constant under monotonically increasing variations. They range within [0, 1].

3.3 Copula estimation

There are several methods for estimating a copula:

1. Maximum Likelihood Estimation (MLE): This method is often considered difficult to use, for there are many parameters to estimate.

2. Pseudo-MLE: There are two types of pseudo-MLE, i.e. parametric pseudo-MLE and semi parametric pseudo-MLE. They are used more often than MLE. In pseudo-MLE, the margins are estimated through the cumulative distribution function, and the copula is then estimated through MLE.

3.4 MLE

Consider $Y = (Y_1, \ldots, Y_m)$ a random diagram. Assume that $F_{Y_1}(\cdot|\theta_1), \ldots, F_{Y_m}(\cdot|\theta_m)$ is a parametric model for marginal distribution functions and that $c_Y(\cdot|\theta_C)$ is a parametric model for copula Y. The following equation is true:

$$f_Y(y) = f_Y(y_1, \ldots, y_m) = c_Y(F_{Y_1}(y_1), \ldots, F_{Y_m}(y_m)) \prod_{j=1}^{m} f_{Y_j}(y_j).$$

Assume that an instance of IID is $Y_{1:n} = (Y_1, \ldots, Y_n)$. The likelihood logarithm is then obtained

$$\begin{aligned} \log L(\theta_1, \ldots, \theta_m, \theta_C) &= \log \textstyle\prod_{i=1}^{n} f_Y(y_i) \\ &= \sum_{i=1}^{n} (\log[c_Y(F_{Y_1}(y_{i,1}|\theta_1), \ldots F_{Y_m}(y_{i,m}|\theta_m)|\theta_C)] \\ &\quad + \log(f_{Y_1}(y_{i,1}|\theta_1)) + \cdots + \log(f_{Y_m}(y_{i,m}|\theta_m))). \end{aligned}$$

ML estimators $\widehat{\theta_1}, \ldots, \widehat{\theta_2}, \widehat{\theta_C}$ are obtained from the maximization of the above equation based on $\theta_1, \ldots, \theta_m, \theta_C$.

This method has a few setbacks:

1. There are too many parameters to estimate, especially for large values of m. As a result, optimization can be difficult.

2. If any of the univariate parametric distributions $F_{Y_i}(\cdot|\theta_i)$ are defined incorrectly, bias can emerge in univariate distributions and the copula [15].

3.5 Pseudo-MLE

Pseudo-MLE helps solve the above mentioned MLE drawbacks. This method has two steps:

1. The marginal distribution functions are first estimated to define $\widehat{F}_{Y_j}$, for $j = 1, \ldots, m$. For this purpose, the following two methods can be adopted:

   - The empirical distribution function is defined as below for $y_{1,i}, \ldots, y_{n,j}$:

$$\widehat{F}_{Y_i}(y) = \frac{\sum_{i=1}^{n} I_{\{y_{i,j} \leq y\}}}{n+1}.$$

- A parametric model is developed with $\widehat{\theta}_j$ obtained from the univariate conventional MLE.

2. The parameters of copula θ_C are obtained by maximizing the following expression:

$$\sum_{i=1}^{n} \log[c_Y(\widehat{F}_{Y_1}(y_{i,1}), \ldots, \widehat{F}_{Y_m}(y_{i,m})|\theta_C].$$

It should be noted that the above expression is obtained directly from the likelihood logarithm only by using marginal distributions in Step 1 and using the parameters of θ_C that were not estimated [15].

This method consists of two steps:

Step 1: In this step, pseudo-MLE is adopted to link the univariate marginal distributions to their joint multivariate distribution function. After that, the copula parameters are estimated to place the dimensions with strong correlation in a smaller set. If ρ of a copula is greater than 0.7 for two random continuous variables X_1 and X_2, these variables are strongly correlated; thus, they are placed in the subset of interest.

Step 2: In the second step, the dimensions of this subset are analyzed to delete the dimensions that are the linear combinations of the subset dimensions. Finally, the greatest value of ρ is selected from the subset, and the rest of the dimensions are deleted, for the other dimensions behave like the selected dimension and can be used as additional dimensions.

4 Real data on Parkinson's patients

Numerical comparisons were drawn by using a dataset including 195 individuals (both healthy and affected with Parkinson's disease). The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders. For this purpose, 22 biomedical audio variables were employed to process every participant's voice. Hence, the research data consisted of 22 biomedical audio variables and one classification variable (healthy individuals and patients with Parkinson's disease). The variables are as follows: (X_1) is MDVP: Fo(Hz): Average vocal fundamental frequency, (X_2) is MDVP: Fhi(Hz): Maximum vocal fundamental frequency, (X_3) is MDVP: Flo(Hz): Minimum vocal fundamental frequency, (X_4) is MDVP: Jitter(%), (X_5) is MDVP: Jitter(Abs), (X_6) is MDVP: RAP, (X_7) is MDVP: PPQ, (X_8) is DDP: Jitter: Several measures of variation in fundamental frequency, (X_9) is MDVP: Shimmer, (X_{10}) is MDVP: Shimmer(dB), (X_{11}) is Shimmer: APQ3, (X_{12}) is Shimmer: APQ5, (X_{13}) is MDVP: APQ, (X_{14}) is Shimmer: DDA : Several measures of variation in amplitude, (X_{15}) is NHR, (X_{16}) is HNR: Two measures of ratio of noise to tonal components in the voice, (X_{17}) is RPDE, (X_{18}) is D2: Two nonlinear dynamical complexity measures, (X_{19}) is DFA: Signal fractal scaling exponent, (X_{20}) is spread1, (X_{21}) is spread2 and (X_{22}) is PPE: Three nonlinear measures of fundamental frequency variation.

Status : Classification variable with two status, i.e. N for healthy individuals and P for patients with Parkinson's disease[1].

[1] Available at: https://archive.ics.uci.edu/ml/datasets/parkinsons

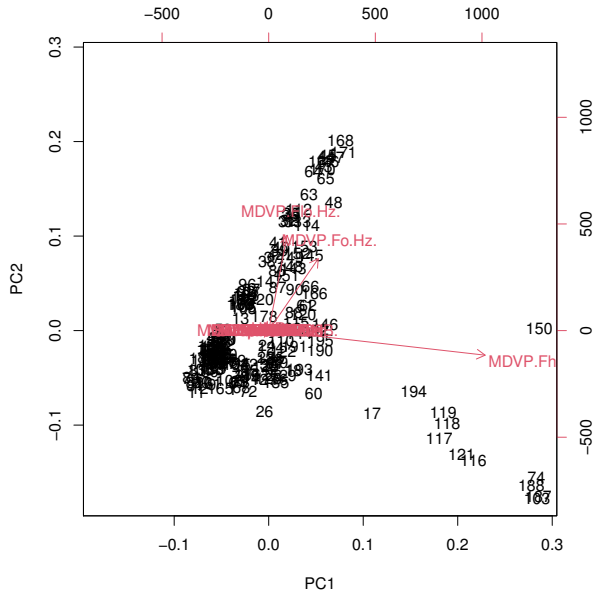


Figure 1: Principal components of the population.

In this section, the naïve Bayes classifier and KNN ($k = 5$) [13] are applied first to all data and then to the data reduced by the PCA. After that, the proposed copula-based method is adopted. The results are then compared.

For classification, the data were divided into training and test sets on a 70-to-30 proportion. The accuracy criterion is then defined as the following equation to compare the classification methods [22].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}},$$

where

- TP is a sample belonging to the positive class and is identified as a member of this class.
- FN is a sample belonging to the positive class and is identified as a member of the negative class.
- TN is a sample belonging to the negative class and is identified as a member of this class.
- FP is a sample belonging to the negative class and is identified as a member of the positive class.

The PCA is used first. Figure 1 shows the components of this method. Figure 2 demonstrates the variance ratio of population components. Accordingly, the first five components account for nearly 90% of the total variance of population. The copula-based method is now employed:

Step 1: After the Gaussian copula is estimated, the sets of strongly correlated variables are as follows:

$$\{X_1, X_2\}, \{X_4, \ldots, X_{16}, X_{22}\}.$$

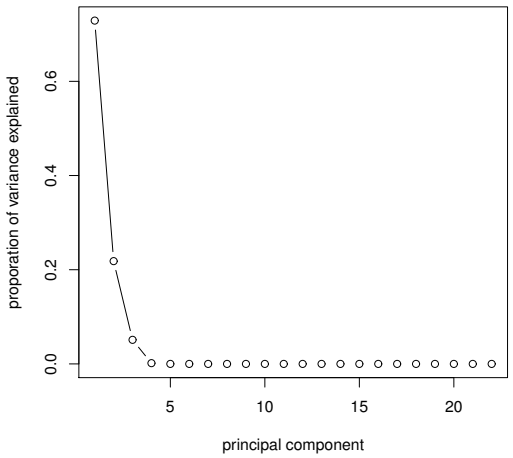


Figure 2: The proportion of variance explained by principal. components

Table 1: The accuracy results of different classification methods for the original data and the reduced data.

	Full Data	PCA	Copula
naïve Bayes	0.661	0.6271	0.7627
KNN	0.9138	0.8103	0.9137

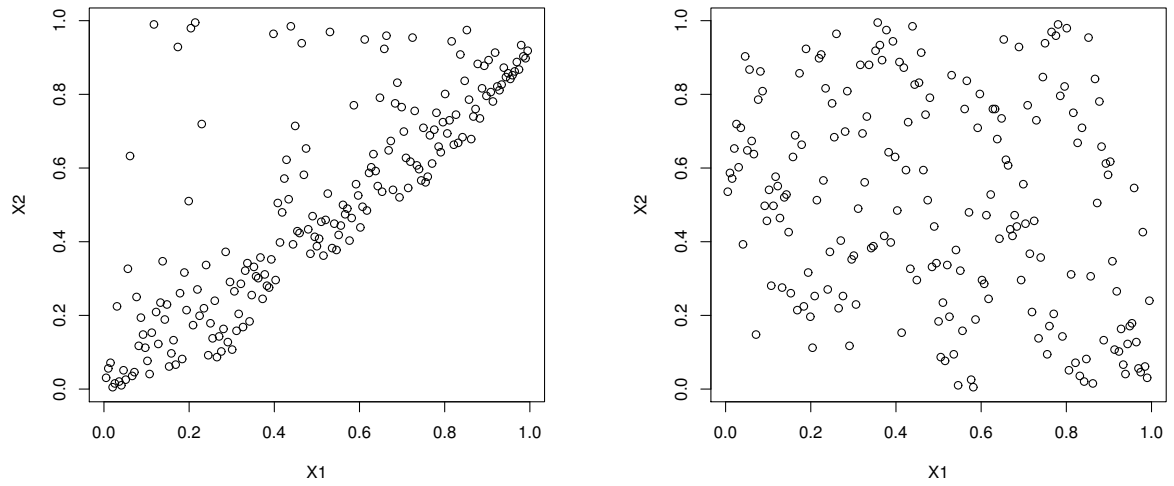
Step 2: The dimensions of these sets should now be checked in terms of linear dependence to ensure that they have nonlinear relationships. In other words, if $\sum_{k\in\{1,2\}} \alpha_k X_k = 0$, then $\alpha_k = 0$ for all $k \in \{1,2\}$. Moreover, if $\sum_{k\in\{4,5,6,7,8,9,10,11,12,13,14,15,16,22\}} \alpha_k X_k = 0$, then $\alpha_k = 0$ for all $k \in \{4,5,6,7,8,9,10,11,12,13,14,15,16,22\}$. As a result, the reduced dimensions are as follows:

$$\{X_1, X_4, X_{17}, X_{18}, X_{20}, X_{21}\}.$$

Figures 3(a)-3(b) illustrate the scattering of original data and that of the data reduced through the copula: The naïve Bayes classifier and KNN are also applied to the original data and the data reduced through the copula and the PCA. Table1 reports the accuracy results in different cases.

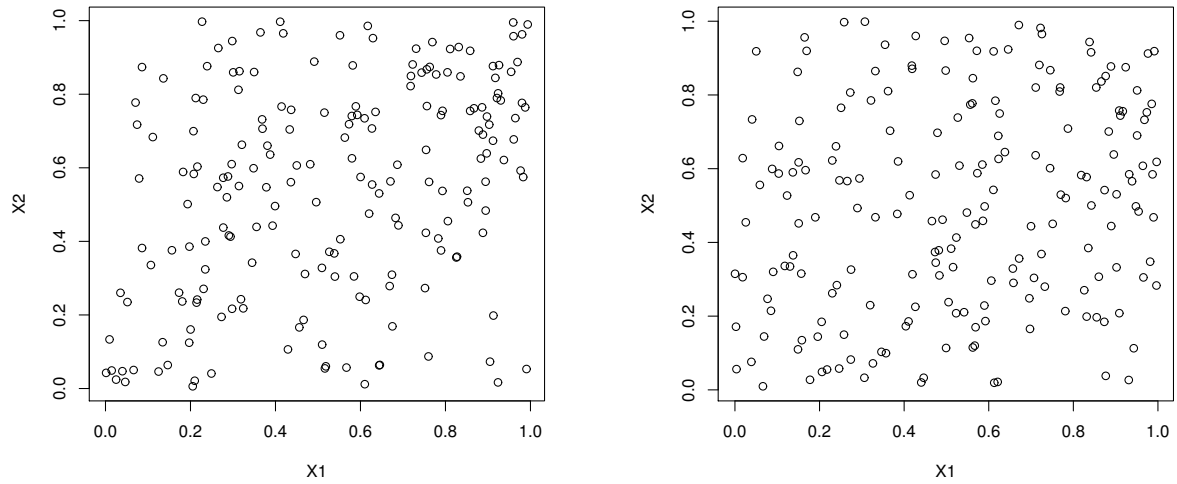
5 Conclusion

According to Figures 3-4, the scattering of the population distribution for the reduced data exceeded that of the original data. Therefore, it provides more information about the population by eliminating the noisy and additional data. As a result, the copula-based method accelerates analysis. Table 1 indicated that the accuracy of the naïve Bayes classifier increased for the data reduced through the copula method



(a) Scattering of empirical distribution for original data (b) Scattering of empirical distribution for reduced data

Figure 3: Scattering of empirical distribution of original data and that of the data reduced through the copula.



(a) Scattering of population distribution for original data (b) Scattering of population distribution for reduced data

Figure 4: Scattering of population distribution for original data and that of the data reduced through the copula.

in comparison with that of the original data; therefore, this method improved. Moreover, the number of data decreased; as a result, analysis was facilitated. In KNN, the accuracy of the original data is nearly equal to that of the data reduced through the copula. In general, the KNN outperformed the naïve Bayes

classifier on the data. Evidently, the copula-based dimensionality method was very effective and efficient in classification techniques; therefore, it can be useful for other data mining and analysis techniques for big data.

Acknowledgements

We would like to thank to respected reviewers for their thorough reading of the manuscript and their helpful remarks that helped us to improve the manuscript.

References

[1] F. Badakhshan Farahabadi, K. Fathi Vajargah, R. Farnoosh, *Dimension reduction big data using recognition of data features based on Copula function and principal component analysis*, Adv. Math. Phys. **2021** (2021) 9967368.

[2] F. Badakhshan Farahabadi, K. Fathi Vajargah, R. Farnoosh, *Dimension reduction of big data and deleting noise and its efficiency in the decision tree method and its use in covid 19*, Int. J. Math. Model. Comput. **12** (2022) 183–190.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York, 2006.

[4] U. Braga-Neto. *Fundamentals of Pattern Recognition and Machine Learning*, Springer, Cham, 2024.

[5] D.K. Choubey, P. Kumar, S. Tripathi, S. Kumar, *Performance evaluation of classification methods with pca and pso for diabetes*, Netw. Model. Anal. Health Inform. Bioinform. **9** (2020) 5.

[6] B. Fathi Vajargah, F. Merdoust, K. Fathi Vajargah, *On computing dominant eigenpair by markov chain monte carlo method*, J. Appl. Math. Inform. **6** (2010) 2.

[7] K. Fathi Vajargah, *Comparing ridge regression and principal components regression by monte carlo simulation basedon MSE*, J. Comput. Sci. Comput. Math. **3** (2013) 25–29.

[8] K. Fathi Vajargah, H. Mottaghi Golshan, F. Badakhshan Farahabadi, *Improving the LDA linear discriminant analysis method by eliminating redundant variables for the diagnosis of COVID-19 patients*, Appl. Appl. Math. **18** (2023) p1.

[9] J. Forkman, J. Josse, and H.-P. Piepho, *Hypothesis tests for principal component analysis when variables are standardized*, J. Agric. Biol. Environ. Stat. **24** (2019) 289–308.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Computer Science and Scientific Computing, Academic Press, Boston, MA, second edition, 1990.

[11] P. Geethanjali, *Comparative study of PCA in classification of multichannel EMG signals*, Australas Phys. Eng. Sci. Med. **38** (2015) 331–343.

[12] G.H. Golub, C.F. Van Loan, *Matrix Computations*, JHU press, 2013.

[13] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*, Springer Science & Business Media, 2011.

[14] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, second edition, 2009.

[15] M. Haugh, *An Introduction to Copulas: Quantitative risk management*, Lecture Notes. New York: Columbia University, 2016.

[16] D. Hong, F. Zhang, *Weighted elastic net model for mass spectrometry imaging processing*, *Math. Model. Nat. Phenom.* **5** (2010) 115–133 .

[17] R. Houari, A. Bounceur, M.T. Kechadi, A.K. Tari, R. Euler, *Dimensionality reduction in data mining: A copula approach*, Expert Syst. Appl. **64** (2016) 247–260.

[18] I. Jolliffe, *A 50-year personal journey through time with principal component analysis*, J. Multivar. Anal. **188** (2022) 104820.

[19] I.T. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

[20] I.T. Jolliffe, J. Cadima, *Principal component analysis: a review and recent developments*, Philos. Trans. Roy. Soc. A, **374** (2016) 16.

[21] D. Lopez-Paz, J.M. Hernandez-Lobato, G. Zoubin, *Gaussian process vine copulas for multivariate dependence*, In International Conference on Machine Learning **28** (2013) 10–18.

[22] C.E. Metz, Basic principles of ROC analysis, In Seminars in nuclear medicine, **8** (1978) 283–298.

[23] R.B. Nelsen, *An Introduction to Copulas*, Springer Series in Statistics, Springer, New York, second edition, 2006.

[24] F.R. Pirolla, M.T. Santos, J.C. Felipe, M.X. Ribeiro, *Dimensionality reduction to improve content-based image retrieval: A clustering approach*, In 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (2012) 752–753.

[25] A. Zinovyev, *Overcoming complexity of biological systems: from data analysis to mathematical modeling*, Math. Model. Nat. Phenom **10** (2015) 186–206.