Hindawi
Advances in Mathematical Physics
Volume 2021, Article ID 9967368, 8 pages
https://doi.org/10.1155/2021/9967368

Research Article

Dimension Reduction Big Data Using Recognition of Data Features Based on Copula Function and Principal Component Analysis

Fazel Badakhshan Farahabadi,[1] **Kianoush Fathi Vajargah**,[2] **and Rahman Farnoosh**[3]

[1] *Department of Statistics, Islamic Azad University, Science and Research Branch, Tehran, Iran*
[2] *Department of Statistics, Islamic Azad University, Tehran North Branch, Iran*
[3] *School of Mathematics, Iran University of Science and Technology, Tehran 16844, Iran*

Correspondence should be addressed to Kianoush Fathi Vajargah; fathi_kia10@yahoo.com

Received 12 March 2021; Accepted 16 June 2021; Published 12 July 2021

Academic Editor: Sandro Wimberger

Nowadays, data are generated in the world with high speed; therefore, recognizing features and dimensions reduction of data without losing useful information is of high importance. There are many ways to dimension reduction, including principal component analysis (PCA) method, which is by identifying effective dimensions in an acceptable level, reducing dimension of data. In the usual method of principal component analysis, data are usually normal, or we normalize data; then, the principal component analysis method is used. Many studies have been done on the principal component analysis method as a step of data preparation. In this paper, we propose a method that improves the principal component analysis method and makes data analysis easier and more efficient. Also, we first identify the relationships between the data by fitting the multivariate copula function to data and simulate new data using the estimated parameters; then, we reduce the dimensions of new data by principal component analysis method; the aim is to improve the performance of the principal component analysis method to find effective dimensions.

1. Introduction

In many real-world programs, reduction of high-volume data is of high importance and necessity as a prestage of data processing. For example, in data mining programs, dimensionality reduction is considered one of the most important stages to remove data redundancy, to increase precision of measurement, and to improve decision making process. Analyzing high-volume data is intrinsically difficult via high-volume computations for many learning algorithms as well as data processing. In dimensionality reduction methods, extraction of data features is highly important. A highly used method to reduce dimension reduction of data in data mining and in the data preparing phase is the principal component analysis method. The PCA method can be used if the original variables are correlated, homogeneous, if each component is guaranteed to be independent and if the dataset is normally distributed [1, 2]. The critical issues for the majority of dimensionality reduction studies are how to provide a convenient way to generate correlated multivariate random variables without imposing constrain to specific types of marginal distributions. An appropriate approach to this problem is to use Copula's theory [3, 4]. In this paper, we first use the copula function to study the correlation and relationships between data to determine and eliminate irrelevant properties and simulate new data using the estimated parameter; then, by using the PCA method, we reduce the dimensions of data [4–6].

1.1. Principal Component Analysis (PCA). Principal component analysis method has been first developed by Karl Pearson in 1901. The analysis includes analyzing special values of the covariance matrix. Analyzing principal components upon mathematics definition is an orthogonal transformation taking data to a new system of coordinates so that the largest data variance would be placed on the first coordinate axis; the second largest variance would be placed on the second coordinate axis and etc. Principal component analysis is aimed at transferring dataset X with m dimensions to data Y with l dimensions. Therefore, it is assumed that matrix X is formed of vectors $X_1, X_2, \cdots, X_n$ each of which placed in m column in matrix X. So, the data matrix would be in form of $m \times n$. Principal components are just related to covariance matrix Σ (correlation matrix p) of random variables $X_1, X_2, \cdots, X_n$ [7].

1.2. Calculating Empirical Mean and Covariance Matrix and Data Normalization. To calculate covariance matrix, data have to be normalized first. To do so, the primarily vector of empirical mean would be calculated as follows:

$$U_m = \frac{1}{n}\sum_{i=1}^{n} X_{[m,i]}. \tag{1}$$

Clearly, the empirical mean would be applied on matrix lines.

Then, the distance matrix to mean would be obtained as follows:

$$B = X - uh, \tag{2}$$

where h is a vector with size of $1 \times n$ and value equal to 1 in each of the entries.

Covariance matrix Σ with $m \times m$ dimensions would be obtained as follows:

$$\sum = E[B \otimes B] = E[B \cdot B^*] = \frac{1}{n} B \cdot B^*, \tag{3}$$

where E is arithmetic mean, $\bigotimes$ is an external coefficient, and B^* is the matrix B conjugate transpose.

Consider $X' = [X_1, X_2, \cdots, X_n]$ random vector and assume that this random vector has matrix covariance Σ with special values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Consider following linear compositions:

$$\begin{cases} Y_1 = l_1' X = l_{11}X_1 + l_{21}X_2 + \cdots + l_{n1}X_n, \\ Y_2 = l_2' X = l_{12}X_1 + l_{22}X_2 + \cdots + l_{n2}X_n, \\ \vdots \\ Y_n = l_n' X = l_{1n}X_1 + l_{2n}X_2 + \cdots + l_{nn}X_n. \end{cases} \tag{4}$$

Using relationship (4), we have

$$\text{var}\ (Y_i) = l_i' \sum l_i, \quad \text{cov}\ (Y_i, Y_k) = l_i' \sum l_k, \quad i, k = 1, 2, \cdots, n. \tag{5}$$

Its principal components are $Y_1, Y_2, \cdots, Y_n$ unrelated linear compositions; variances of which in relationship (5) would be large to the extent possible. The first principal component of a linear composition has maximum variance. Clearly, var $(Y_1) = l_1' \Sigma l_1$ can be maximized through multiplying each l_1 by a constant. That is, the first principal component of linear composition is $l_1' X$ which maximizes var (Y_1) with consideration of $l_1' l_1' = 1$. The second principal component of linear composition is $l_2' X$ which maximizes var (Y_2) with consideration of $l_2' l_2 = 1$ and cov $(l_1' X, l_2' X) = 0$, continuously to the n^{th} principal component.

According to relationship (5), we have

$$\sum_{i=1}^{n} \text{var}(X_i) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n = \sum_{i=1}^{n} \text{var}(Y_i), \tag{6}$$

and ratio of total variance to K^{th} component $(k = 1, 2, \cdots, n)$ is

$$\left(\text{Total share of population variance related to principal } K^{\text{th}} \text{ component}\right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}. \tag{7}$$

If for large n, the highest maximum variance of total population (80 or 90%) could be attributed to the first several components; these components can be replaced by n primary variables, losing not much information [2, 8–10].

2. Copula Function

In general, the copula function is the link function of multivariate distributions and their marginal distributions. The copula function is a multivariate distribution, marginal distribution which follows uniform distribution of [0,1] interval [11–13].

2.1. Characteristics of Copula Function. Assume the following characteristics for $C : I^2 \longrightarrow I$:

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

$$C(u, 0) = C(0, v) = 0, C(u, 1) = u, C(1, v) = v \tag{8}$$

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

$$C(U_2, v_2) + C(U_1, v_1) - C(U_1, v_2) - C(U_2, v_1) \geq 0 \tag{9}$$

Such function like C implied in the two above conditions is called the copula function [14].

2.2. Sklar's Theorem. It is indicated by Sklar's theorem that if joint distribution function like H would be available with marginal distributions F and G, then, there would be copula function C available. That is, for every $X_i, X_j \in \mathbb{R}$, we have

$$H(X_i, X_j) = C(F(X_i), G(X_j)), \tag{10}$$

and if F and G would be continuous, then, copula function C would be unique. Otherwise, C would be defined as unique on $\mathrm{Rang}(F) \times \mathrm{Rang}(G)$.

The most important application of the copula function is formulation of a proper method to produce distribution of random related multivariate variables and to provide a solution for the problem of density estimation transformation [15].

For reversible transformation of n continuous random variables $X_1, X_2, \cdots, X_n$ based on their distribution function to n independent variables with uniform distribution $U_1 = F_1(X_1)$, $U_2 = F_2(X_2), \cdots, U_n = F_n(X_n)$, the probability density function $X_1, X_2, \cdots, X_n$ would be equal to $f(X_1, \cdots, X_n)$ and joint probability density function U_1, $U_2, \cdots, U_n$ would be equal to $C(U_1, \cdots, U_n)$. Therefore, probability density function $f(X_1, \cdots, X_n)$ can provide a nonparametric form (unknown distribution). Here, probability density function $C(U_1, \cdots, U_n)$ for U_1, $U_2, \cdots, U_n$ would be estimated instead of $X_1, X_2, \cdots, X_n$, so that problem of density estimation becomes simpler. Then, it would be simulated so that random samples $X_1, X_2, \cdots, X_n$ would be obtained through reverse transformation $X_i = F^{-1}(U_i)$.

According to Sklar's theorem, one copula function with n unique dimensions C is available in $[0,1]^n$ with uniform marginal distribution $U_1, U_2, \cdots, U_n$. That is, every function F with margins $F_1, F_2, \cdots, F_n$ can be written as follows:

$$\forall (X_1, \cdots, X_n) \in \mathbb{R}^n, F(X_1, \cdots, X_n) = C(F_1(X_1), \cdots, F_n(X_n)). \tag{11}$$

To evaluate a copula function selected via an estimated parameter and to avoid defining any hypothesis on distributions, empirical distribution function can be used. An empirical copula function is useful to study the dependence structure of multivariate random vectors. In general, empirical copula function is as follows:

$$C_{ij} = \frac{1}{n}\sum_{k=1}^{n} I_{\left(U_{kj} \le U_{ij}\right)}, \tag{12}$$

where $I_{(\bullet)}$ would be an indicator function [16].

2.3. Gaussian Copula Function. Difference between Gaussian copula function and normal joint distribution function is that the first one authorizes various distribution functions to be used for joint distribution [14]. However, in probability theory and statistics, normal multivariate distribution is considered the generalization of one-dimensional normal distribution [17].

Gaussian copula function is defined as

$$C(\Phi(X_1), \cdots, \Phi(X_n)) = \frac{1}{|\sum|^{1/2}} \exp\left\{\frac{-1}{2} X^t \left(\overset{-1}{\sum} - I\right) X\right\}, \tag{13}$$

where $\Phi(X_i)$ is a standard Gaussian function and X_i has standard normal distribution and Σ is a correlation matrix. As a result, $C(U_1, \cdots U_n)$ copula function would be called a Gaussian copula function.

3. Methodology

In the research, a two-stage method would be used for dimensionality reduction. That is, primarily empirical copula function and fit of Gaussian copula function to data would be used to estimate parameter p for variables X_1, $X_2, \cdots, X_n$. Important advantages of using the copula function in multivariate distributions is that correlation between variables would be considered by these functions, and in fact, there would be no need for independence of variables; instead, the correlation structure between variables would be even considered by these functions [18]. For estimation purposes, generating function is available with dependence unscaled value available in it. The correlation coefficient value has to be specified. To do so, the Pearson correlation coefficient will be used and defined as follows for two X_i and X_j variables:

$$\rho = \frac{\mathrm{cov}\left(X_i, X_j\right)}{\sigma_{X_i}\sigma_{X_j}}, \tag{14}$$

where σ_{X_i} and σ_{X_j} are standard deviations of X_i and X_j, respectively.

Then, those data with lower correlation compared to others would be eliminated and using estimated function and Gaussian copula function for $X_1, X_2, \cdots, X_m$, where m uniform variables $U_1 = F_1(X_1)$, $U_2 = F_2(X_2), \cdots, U_n = F_m(X_m)$ would be generated $(m \le n)$ and placed instead of $X_1, X_2, \cdots, X_m$ in the principal component analysis method. After dimensionality reduction, the results would be compared through applying the method on raw data [16, 19].

4. Numerical Results

During past 30 years, increasing prevalence of urinary stone disease has been observed. About 80% of kidney stones are from calcium oxalate type. Here, 79 urine samples would be analyzed to see if some of physical features of urine are related to formation of calcium oxalate or not. These data include following columns (variables), which is available at https://cran.r-project.org/web/packages/cond.

Using Gaussian copula function, correlation values of variables would be obtained as follows:

Considering Table 1, it is observed that correlation of variable $X2$ is lower than other variables; so, it would be eliminated at the first stage. After estimation of parameters, new data would be generated. Figure 1 shows the copula function for main data and data generated by this method.

Now, data would be generated based on estimated parameters. To specify whether data are generated correctly or not, diagram QQPlot would be drawn.

Table 1: Estimation of parameter ρ for variables of urine.

	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
$X1$	1	-0.30856	0.83231	0.57256	0.81165	0.54872
$X2$		1	-0.25167	-0.09762	-0.27985	-0.12147
$X3$			1	0.77226	0.81012	0.58452
$X4$				1	0.45542	0.43444
$X5$					1	0.58813
$X6$						1

$X1$is urine gravity, $X2$ is urine pH, $X3$ is urine osmolarity (it is corresponding to unit of solute concentration), $X4$ is urine conductivity (it is corresponding to concentration of charged ions in solution), $X5$ is urea concentration (mM/liter), and $X6$ is calcium concentration (mM/liter).

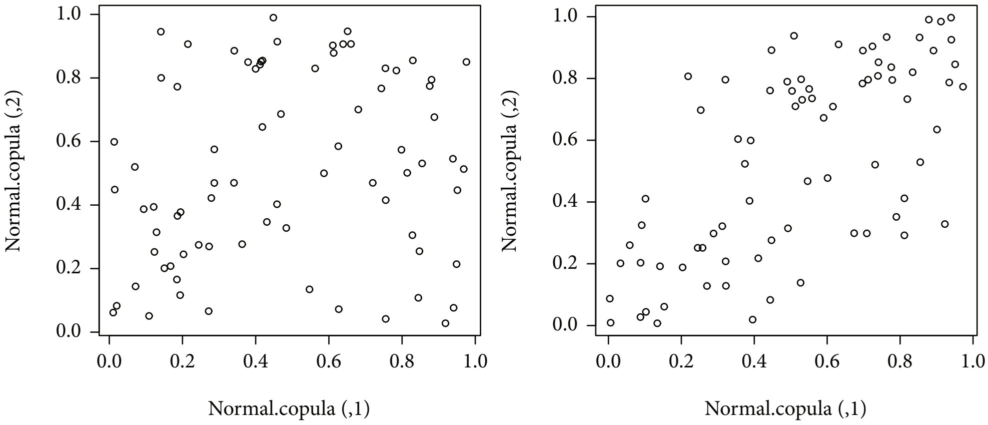

Figure 1: Diagram of copula function for generated data based on main and reduced data.

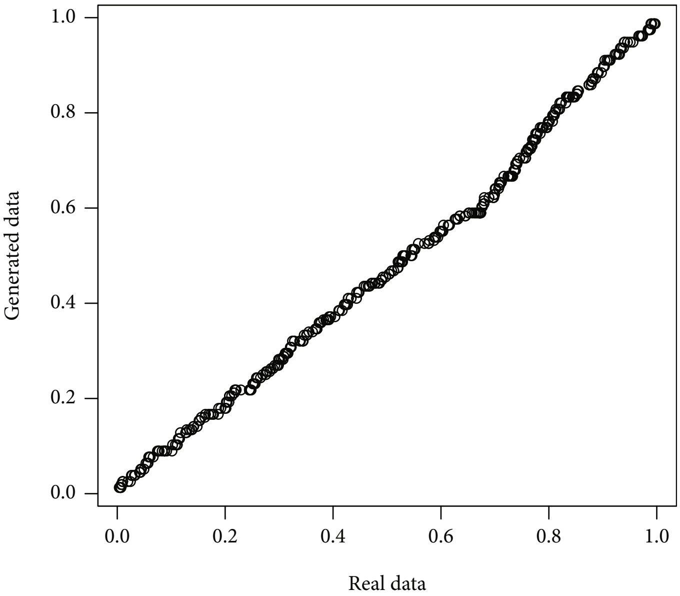

Figure 2: QQ Plot diagram of real and generated data for data of example 1.

Correct data generation is shown by Figure 2. In the second stage, after elimination of the $X2$ variable on data generated, principal component analysis would be done. In Figure 2, principal components for primary data and those generated by copula function are shown after reduction of the $X2$ variable. Figure 3 shows principal components for main data and the data generated.

Ratios of population variance related to principal components are provided in following table. Its screen plot is as follows.

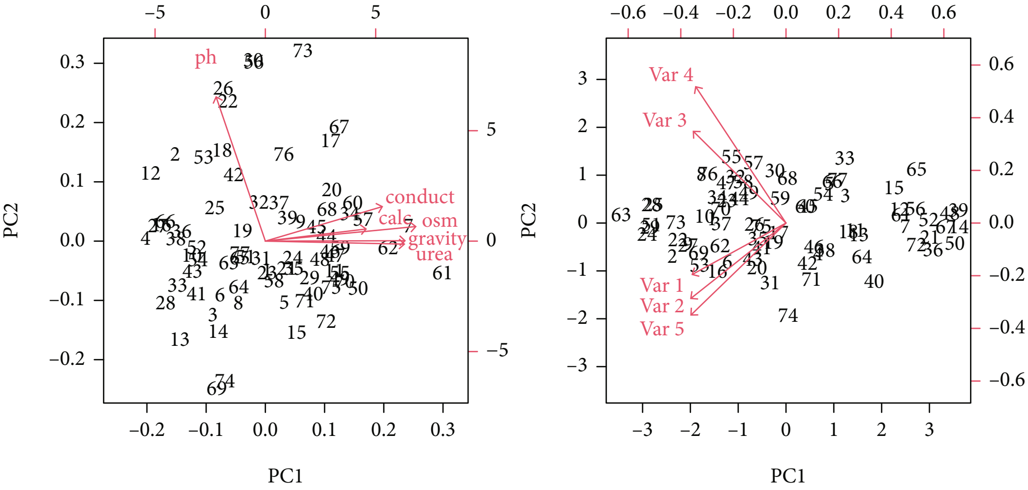

Figure 3: Diagram of principal components for raw and generated data through recommended method.

Table 2: Ratios of population variance related to principal components for main data.

PC1	PC2	PC3	PC4	PC5	PC6
0.61817360	0.15701415	0.11567297	0.07879801	0.02912841	0.00121285

Table 3: Ratios of population variance related to principal components for data generated through the recommended method.

PC1	PC2	PC3	PC4	PC5
0.73414280	0.07840848	0.07583399	0.06866719	0.04294755

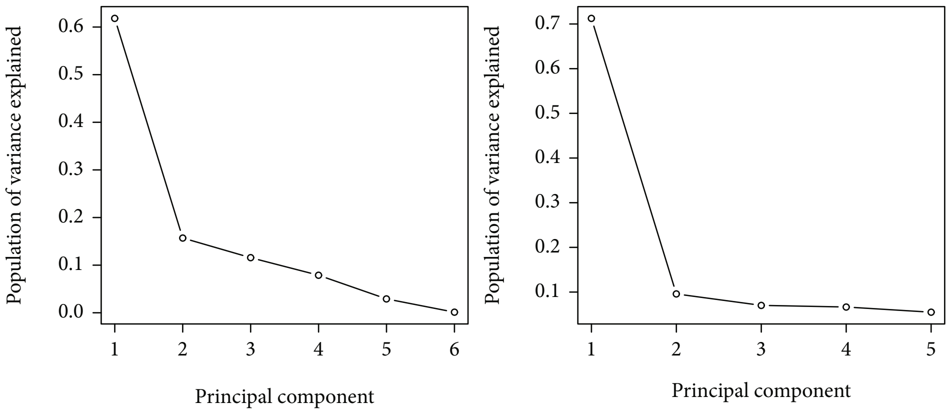

Figure 4: Diagram of population variance ration related to principal components for main data and generated data through recommended method.

Considering Tables 2 and 3 as well as Figure 4, it is observed that in dimensionality reduction method presented in the research, two first components include more than 80% of population variances and first component includes more than 70% of population.

Example 1. To recognize image resolution in a rectangular monitor, its display would be divided into different boxes and numbers of black and white dots in these boxes would be measured. Images of these characters have been made based on 20 different images, and each box from within these 20 boxes has been randomly selected. A file including 20000 unique simulators have been produced. Each stimulator has been transformed and scaled to 7 following numerical variables so that they would be placed within 0-15 range, (which is available at https://cran.r-project.org/web/packages/mlbench/index.html).

There are 2000 observations available from these variables.

Table 4: Estimation of parameter ρ for variables of resolution in a rectangular monitor.

	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$
$X1$	1	0,7960	0,8788	0,7439	0,7282	-0,0263	0,0296
$X2$		1	0,7044	0,8203	0,6148	0,0784	-0,0754
$X3$			1	0,7089	0,8156	0,0648	0.0119
$X4$				1	0.0119	0,0618	-0,0190
$X5$					1	0,1196	-0,0278
$X6$						1	-0,4227
$X7$							1

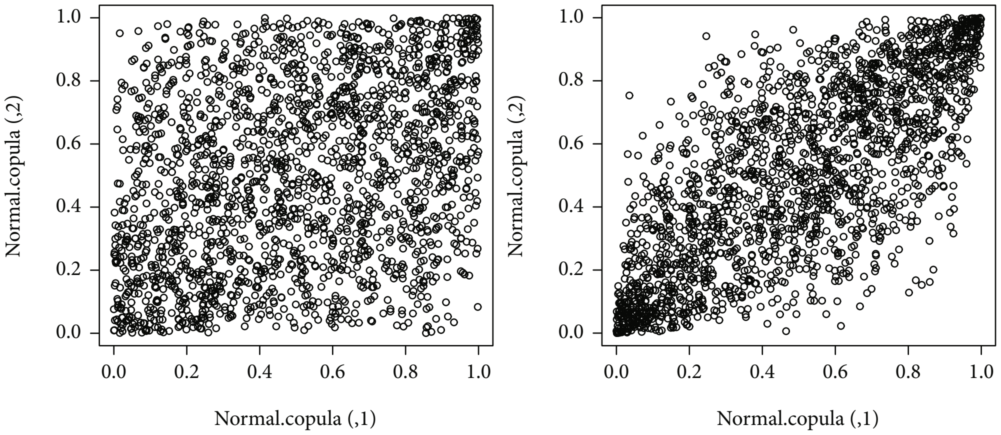

Figure 5: Diagram of copula function for main and reduced data.

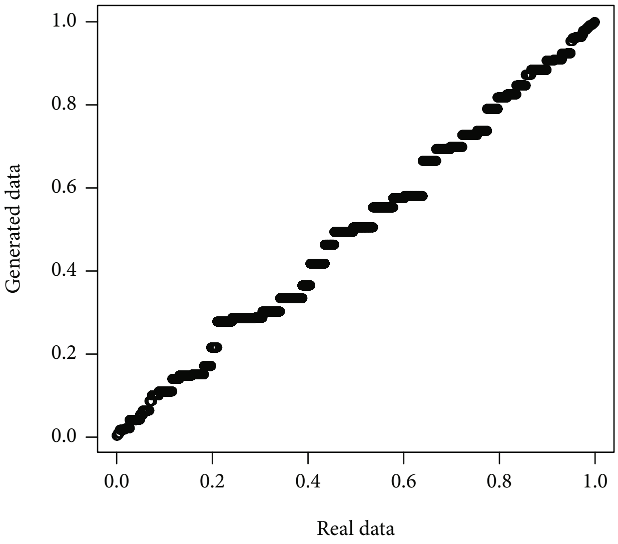

Figure 6: QQ Plot diagram of real and generated data for data of example 2.

Using Gaussian copula function, correlation values of variables would be obtained as follows:

X is the box. $X1$ is the horizontal location of box, $X2$ is the vertical location of box (y.box), $X3$is width of box (width), $X4$ is the height of box (height), $X5$ is the total numbers of dots in the box (onpix), $X6$ is the mean value of x in dots of the box (x .bar), and $X7$ is the mean value of y in dots of box (y.bar).

Considering Table 4, it is observed that correlation between variables $X6$ and $X7$ is less compared to other variables. So, these two would be eliminated at first stage and then Gaussian copula function would be fitted to reduced data and new data would be generated through estimated parameter, which is shown in Figure 5.

Now, data would be generated. QQPlot would be as follows.

Now, principal component analysis would be done on generated data. Diagrams of principal components are as follows (Figure 6).

Table 5: Ratio of population variance related to principal components for main data.

PC1	PC2	PC3	PC4	PC5	PC6
0.55123270	0.20018487	0.09008126	0.07169468	0.05074973	0.02071375

Table 6: Ratio of population variance related to principal components for data reduced through recommended method.

PC1	PC2	PC3	PC4	PC5
0.79028555	0.05470965	0.05363693	0.05154868	0.0498199

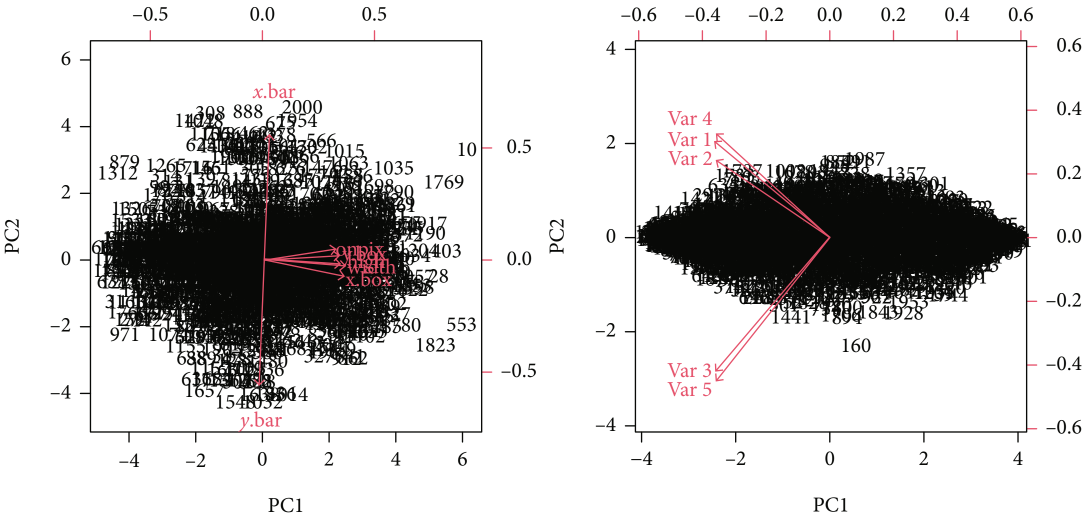

Figure 7: Diagrams of population variance ratios related to principal components for main data and recommended method.

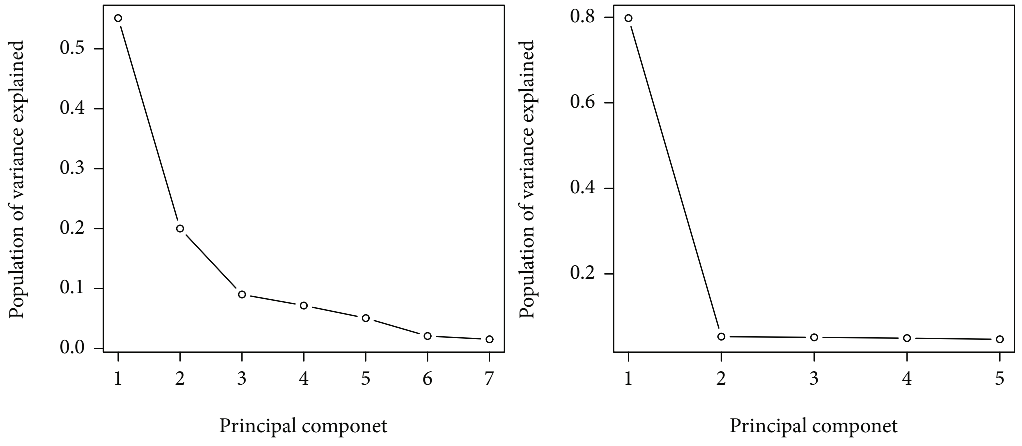

Figure 8: QQPlot diagram of real and generated data.

Screen plot of population variance ratio related to principal components for both methods are as follows.

According to Tables 5 and 6 as well as Figure 7, it is observed that ratio of population variance for the first two components in the recommended method includes almost 85% of population and the first component includes almost 80% of population, whereas, for main data, ratio of population variance for the three first components includes almost 85% of population.

5. Conclusion

Considering the two aforementioned examples, it has been observed that data generated according to the estimated parameters of the Gaussian copula distribution are consistent with the original data (see Figures 2 and 8) by using the recommended method in the research and copula function to recognize dependencies and structural dependence between variables in addition to elimination of redundant data will increase efficiency of principal component analysis method as well as speed of obtaining analysis results (see Figures 4 and 7, Tables 2, 3, 5, and 6). Considering the point that nowadays data are generated with high-speed, appropriate, and efficient methods for dimensionality reduction without losing information are of high importance and necessity, and recommended method in the research is a useful one to do so. The recommended method in the research can be also used for other dimensionality reduction techniques so that data would be prepared for more analysis, for example in data mining.

Data Availability

The data that support the findings of this study are openly available at https://cran.r-project.org/web/packages/cond and https://cran.r-project.org/web/packages/mlbench/index.html.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

All authors contributed equally. All authors read and approved the final manuscript.

References

[1] J. Forkman, J. Josse, and H.-P. Piepho, "Hypothesis tests for principal component analysis when variables are standardized," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 24, no. 2, pp. 289–308, 2019.

[2] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, article 20150202, 2016.

[3] A. Colomé, G. Neumann, J. Peters, and C. Torras, "Dimensionality reduction for probabilistic movement primitives," in *2014 IEEE-RAS International Conference on Humanoid Robots*, pp. 794–800, Madrid, Spain, 2014.

[4] R. Fakoor and M. Huber, "A sampling-based approach to reducing the complexity of continuous state space pomdps by decomposition into coupled perceptual and decision processes," in *2012 11th International Conference on Machine Learning and Applications*, pp. 687–692, Boca Raton, FL, USA, 2012.

[5] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009.

[6] D. Paul and I. M. Johnstone, "Augmented sparse principal component analysis for high dimensional data," 2012, https://arxiv.org/abs/1202.1242.

[7] L. I. Smith, "A tutorial on principal components analysis," 2002, Computer Science Technical Report No. OUCS-2002-12), 2002, https://hdl.handle.net/10523/7534.

[8] I. T. Jolliffe, "Principal components in regression analysis," in *Principal Component Analysis*, pp. 129–155, Springer, 1986.

[9] P. P. Markopoulos, S. Kundu, S. Chamadia, and D. A. Pados, "Efficient l1-norm principal-component analysis via bit flipping," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4252–4264, 2017.

[10] M. Zhai, F. Shi, D. Duncan, and N. Jacobs, "Covariance-based PCA for multi-size data," in *2014 22nd International Conference on Pattern Recognition*, pp. 1603–1608, Stockholm, Sweden, 2014.

[11] E. W. Weisstein, "Mathworld–a wolfram web resource," 2004, https://mathworld.wolfram.com/Erf.html.

[12] D. Lopez-Paz, J. M. Hernández-Lobato, and G. Zoubin, "Gaussian process vine copulas for multivariate dependence," in *International Conference on Machine Learning*, pp. 10–18, Atlanta, Georgia, USA, 2013.

[13] D. MacKenzie and T. Spears, "'The formula that killed wall street': the gaussian copula and modelling practices in investment banking," *Social Studies of Science*, vol. 44, no. 3, pp. 393–417, 2014.

[14] R. B. Nelsen, *An Introduction to Copulas*, Springer Science & Business Media, 2007.

[15] F. Durante, J. Fernandez-Sanchez, and C. Sempi, "A topological proof of Sklar's theorem," *Applied Mathematics Letters*, vol. 26, no. 9, pp. 945–948, 2013.

[16] R. Houari, A. Bounceur, M.-T. Kechadi, A.-K. Tari, and R. Euler, "Dimensionality reduction in data mining: a Copula approach," *Expert Systems with Applications*, vol. 64, pp. 247–260, 2016.

[17] D. MacKenzie and T. Spears, *The Formula that Killed Wall Street? The Gaussian Copula and the Material Cultures of Modelling*, School of Social and Political Science, University of Edinburgh, 2012.

[18] A. Lipton and A. Rennie, *Credit Correlation: Life after Copulas*, World Scientific, 2008.

[19] F. R. Pirolla, M. T. Santos, J. C. Felipe, and M. X. Ribeiro, "Dimensionality reduction to improve contentbased image retrieval: a clustering approach," in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp. 752-753, Philadelphia, PA, USA, 2012.