# Dimensionality Reduction in Data Mining: A Copula Approach

**Article** *in* Expert Systems with Applications · December 2016

DOI: 10.1016/j.eswa.2016.07.041



**5 authors**, including:

Rima Houari
Institut National des Sciences Appliquées de …
**9** PUBLICATIONS **20** CITATIONS
SEE PROFILE

Ahcène Bounceur
Université de Bretagne Occidentale
**123** PUBLICATIONS **455** CITATIONS
SEE PROFILE

Tahar Kechadi
University College Dublin
**333** PUBLICATIONS **2,091** CITATIONS
SEE PROFILE

Abdelkamel A Kamel Tari
Université de Béjaïa
**45** PUBLICATIONS **65** CITATIONS
SEE PROFILE

# Dimensionality Reduction in Data Mining: A Copula Approach

Rima Houari[a], Ahcène Bounceur[b,*], M-Tahar Kechadi[c], A-Kamel Tari[a], Reinhardt Euler[b]

[a] *LIMED Laboratory, University of Bejaia, Faculty of exact sciences, Computing Department, 06000, Bejaia, Algeria*
[b] *Lab-STICC Laboratory, University of Brest, 20 Avenue Victor Le Gorgeu, 29238 Brest, France*
[c] *Lab-CASL Laboratory,University College Dublin, Belfield, Dublin 4, Ireland*

**Abstract**

The recent trends in collecting huge and diverse datasets have created a great challenge in data analysis. One of the characteristics of these gigantic datasets is that they often have significant amounts of redundancies. The use of very large multi-dimensional data will result in more noise, redundant data, and the possibility of unconnected data entities. To efficiently manipulate data represented in a high-dimensional space and to address the impact of redundant dimensions on the final results, we propose a new technique for the dimensionality reduction using Copulas and the LU-decomposition (Forward Substitution) method. The proposed method is compared favorably with existing approaches on real-world datasets: Diabetes, Waveform, two versions of Human Activity Recognition based on Smartphone, and Thyroid Datasets taken from machine learning repository in terms of dimensionality reduction and efficiency of the method, which are performed on statistical and classification measures.

*Keywords:* Data mining, Data pre-processing, Multi-dimensional Sampling, Copulas, Dimensionality reduction.

## 1. Introduction

High-dimensionality data reduction, as part of a data pre-processing-step, is extremely important in many real-world applications. High-dimensionality reduction has emerged as one of the significant tasks in data mining applications and has been effective in removing duplicates, increasing learning accuracy, and improving decision making processes. High-dimensional data are inherently difficult to analyse, and computationally intensive for many learning algorithms and multi-dimensional data processing tasks.

Moreover, most of these algorithms are not designed to handle large, complex, and diverse data such as real-world datasets. One way of dealing with very large sizes of data is to use a high-dimensionality reduction method, which not only reduces the size of the data that will be analysed but also removes redundancies. In this paper, we propose a new approach which reduces the size of the data by eliminating redundant attributes based on sampling methods. The proposed technique is based on the theory of Copulas and the LU-decomposition method (Forward Substitution). A Copula provides a suitable model of dependencies to compare well-known multivariate data distributions to better distinguish the relationship between the data. The detection of dependencies is thereafter used to determine and to eliminate the irrelevant and/or redundant attributes. We have already presented part of this work in (Houari et al., 2013b)(Houari et al., 2013a). The critical issues for the majority of dimensionality reduction studies based on sampling and probabilistic representation (Colomé et al., 2014)(Fakoor & Huber, 2012) are how to provide a convenient way to generate correlated multivariate random variables without imposing constraints to specific types of marginal distributions; to examine the sensitivity of the results of different assumptions about the data distribution;

*Corresponding author. Tel.: +33 (0) 2 98 01 62 17.
*Email addresses:* ri.houari@gmail.com (Rima Houari), Ahcene.Bounceur@univ-brest.fr (Ahcène Bounceur ), tahar.kechadi@ucd.ie (M-Tahar Kechadi), tarikamel59@gmail.com (A-Kamel Tari), Reinhardt.Euler@univ-brest.fr (Reinhardt Euler)

to specify the dependencies between the random variables; to reduce the redundant data and remove the variables which are linear combinations of others; and to maintain the integrity of the original information. For these reasons, the main goal of this paper is to propose a new method for dimensionality reduction based on sampling methods addressing the challenges mentioned before. The paper uses both statistical and classification methods to improve the efficiency of the proposed method. In the statistical part, a standard deviation of the final dimensionality reduction results will be computed for all databases with each dimensionality reduction method studied (PCA, SVD, SPCA, and our approach). However, the effectiveness of dimensionality reduction in classification methods will be improved using from one side the full set of dimensions and from the other the reduced set of provided data in terms of precision and recall, for the three classifiers: Artificial Neural Network (ANN), $k$-nearest neighbors ($k$-NN), naïve Bayesian.

The paper is organized as follows: In Section 2 we discuss some related work. Basic concepts are presented in Section 3. Section 4 presents the modeling system designed to overcome the problem of the curse of dimensionality. Section 5 describes the proposed method. The experimental results are given in Section 6, and finally, Section 7 concludes the paper.

## 2. Related work

The process of data mining typically consists of three steps carried out in succession: data preprocessing, data analysis (which involves considering various models and choosing the best one based on the problem to be solved and its features), and result interpretation (finally the selected model is tested on new data to predict or estimate the expected outcome). In this work, we will focus on the data pre-processing step.

Data pre-processing is an important step in the data mining process, as it counts for more than 60% on average of the effort in building the process. However, it is often loosely controlled, resulting in out-of-range values (outliers or noise), redundant and missing values, etc. Analyzing data that have not been carefully pre-processed can produce erroneous and misleading results. Thus, pre-processing the data and representing them in a format that is acceptable for the next phase is the first and foremost task before performing any analysis, as we all know "no quality data, no quality results". Data pre-processing main tasks include data cleaning, integration, transformation, reduction, and discretisation. Data cleaning deals with noise, outliers, inconsistencies, and missing values. For instance, we have already proposed a new approach that fairly estimates the missing values in (Houari et al., 2013c). Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization, are often applied depending on the learning algorithm used. Data reduction obtains a reduced representation of the dataset that is much smaller in volume, yet produces the same (or almost the same) results (Han & Kamber, San Francisco, 2006). In this paper, we will focus on high-dimensionality reduction. We first describe the most common techniques for dimensionality reduction, and then discuss sampling as a way to reduce the sizes of very large data collections while preserving their main characteristics.

### *2.1. Linear dimensionality reduction*

Principal Component Analysis (PCA) is a well established method for dimensionality reduction. It derives new variables (in decreasing order of importance) that are linked by linear combinations of the original variables and are uncorrelated. Several models and techniques for data reduction based on PCA have been proposed (Sasikala & Balamurugan, 2013). (Zhai et al., 2014) proposed a maximum likelihood approach to the multi-size PCA problem. The covariance based approach was extended to estimate errors within the resulting PCA decomposition. Instead of making all the vectors of fixed size and then computing a covariance matrix, they directly estimate the covariance matrix from the multi-sized data using nonlinear optimization. (Kerdprasop et al., 2014) studied the recognition accuracy and the execution times of two different statistical dimensionality reduction methods applied to the biometric image data, which are: PCA and linear discriminant analysis (LDA). The learning algorithm that has been used to train and recognize the images is a support vector machine with linear and polynomial kernel functions. The main drawback of reducing dimensionality with PCA is that it can only be used if the original variables are correlated, and homogeneous, if each component is guaranteed to be independent and if the dataset is normally distributed. If the original variables are not normalized, PCA is not effective.

The Sparse Principal Component Analysis (SPCA) (Zou et al., 2006) is an improvement of the classical method of PCA to overcome the problem of correlated variables using the LASSO technique. LASSO is a promising variable selection technique, producing accurate and sparse models. SPCA is based on the fact that PCA can be written as a regression problem where the response is predicted by a linear combination of the predictors. Therefore, a large number of coefficients of principal components become zero, leading to a modified PCA with sparse loading. Many studies on data reduction based on SPCA have been presented. (Shen & Huang, 2008) proposed an iterative algorithm named sparse PCA via regularized SVD (sPCA-rSVD) that uses the close connection between PCA and singular value decomposition (SVD) of the data matrix and extracts the PCs through solving a low rank matrix approximation problem. (Bai et al., 2015) proposed a method based on sparse principal component analysis for finding an effective sparse feature principal component (PC) of multiple physiological signals. This method identifies an active index set corresponding to the non-zero entries of the PC, and uses the power iteration method to find the best direction.

Singular Value Decomposition (SVD) is a powerful technique for dimensionality reduction. It is a particular case of the matrix factorization approach and it is therefore also related to PCA. The key issue of an SVD decomposition is to find a lower dimensional feature space by using the matrix product $U \times S \times V$, where $U$ and $V$ are two orthogonal matrices and $S$ is a diagonal matrix with $m \times m$, $m \times n$, and $n \times n$ dimensions, respectively. SVD retains only $r \ll n$ positive singular values of low effect to reduce the data, and thus $S$ becomes a diagonal matrix with only $r$ non-zero positive entries, which reduces the dimensions of these three matrices to $m \times r$, $r \times r$, and $r \times n$, respectively. Many studies on data reduction have been presented which are built upon SVD, such as the ones used in (Zhang et al., 2010) and (Watcharapinchai et al., 2009). (Lin et al., 2014) developed a dimensionality reduction approach by applying the sparsified singular value decomposition (SSVD). Their paper demonstrates how SSVD can be used to identify and remove nonessential features in order to facilitate the feature selection phase, to analyze the application limitations and the computational complexity. However, the application of SSVD on large datasets showed a loss of accuracy and makes it difficult to compute the eigenvalue decomposition of a matrix product $A^T A$, where $A$ is the matrix of the original data.

### *2.2. Nonlinear dimensionality reduction*

A vast literature devoted to nonlinear techniques has been proposed to resolve the problem of dimensionality reduction, such as manifold learning methods, e.g., Locally Linear Embedding (LLE), Isometric mapping (Isomap), Kernel PCA (KPCA), Laplacian Eigenmaps (LE), and a review of these methods is summarized in (Gisbrecht & Hammer, 2015; Wan et al., 2016). KPCA (Kuang et al., 2015) is a nonlinear generalization of PCA in a high-dimensional kernel space constructed by using kernel functions. By comparing with PCA, KPCA computes the principal eigenvectors using the kernel matrix, rather than the covariance matrix. A kernel matrix is done by computing the inner product of the data points. LLE (Hettiarachchi & Peters, 2015) is a nonlinear dimensionality reduction technique based on simple geometric intuitions. This algebraic approach computes the low-dimensional neighborhood preserving embeddings. The neighborhood is preserved in the embedding based on a minimizing cost function in input space and output space, respectively. Isomap (Zhang et al., 2016) explores an underlying manifold structure of a dataset based on the computation of geodesic manifold distances between all pairs of data points. The geodesic distance is determined as the length of the shortest path along the surface of the manifold between two data points. It first constructs a neighborhood graph between all data points based on the connection of each point to all its neighbors in the input space. Then, it estimates geodesic distances of all pairs of points by calculating the shortest path distances in the neighborhood graph. Finally, multidimensional scaling (MDS) is applied to the arising geodesic distance matrix to find a set of low-dimensional points that greatly match such distances.

### *2.3. Sampling dimensionality reduction*

Other widely used techniques are based on sampling. They are used for selecting a representative subset of relevant data from a large dataset. In many cases, sampling is very useful because processing the entire dataset is computationally too expensive. In general, the critical issue of these

strategies is the selection of a limited but representative sample from the entire dataset. Various random, deterministic, density biased sampling, pseudo-random number generator and sampling from non-uniform distribution strategies exist in the literature (Rubinstein & Kroese, 2011). However, very little work has been done on the Pseudo-random number generator and sampling from non-uniform distribution strategies, especially in the multi-dimensional case with heterogeneous data. Naïve sampling methods are not suitable for noisy data which are part of real-world applications, since the performance of the algorithms may vary unpredictably and significantly. The random sampling approach effectively ignores all the information present in the samples which are not part of the reduced subset (Whelan et al., 2010). An advanced data reduction algorithm should be developed in multi-dimensional real-world datasets, taking into account the heterogeneous aspect of the data. Both approaches (Colomé et al., 2014)(Fakoor & Huber, 2012) are based on sampling and a probabilistic representation from uniform distribution strategies. The authors of (Fakoor & Huber, 2012) proposed a method to reduce the complexity of solving Partially Observable Markov Decision Processes (POMDP) in continuous state spaces. The paper uses sampling techniques to reduce the complexity of the POMDPs by reducing the number of state variables on the basis of samples drawn from these distributions by means of a Monte Carlo approach and conditional distributions. The authors in (Colomé et al., 2014) applied dimensionality reduction to a recent movement representation used in robotics, called Probabilistic Movement Primitives (ProMP), and they addressed the problem of fitting a low-dimensional, probabilistic representation to a set of demonstrations of a task. The authors fitted the trajectory distributions and estimated the parameters with a model-based stochastic using the maximum likelihood method. This method assumes that the data follow a multivariate normal distribution which is different from the typical assumptions about the relationship between the empirical data. The best we can do is to examine the sensitivity of results for different assumptions about the data distribution and estimate the optimal space dimension of the data.

### *2.4. Similarity measure dimensionality reduction*

There are other widely used methods for data reduction based on similarity measures (Wencheng, 2010)(Pirolla et al., 2012)(Zhang et al., 2010). According to (Dash et al., 2015), the presence of redundant or noisy features degrades the classification performance, requires huge memory, and consumes more computational time. (Dash et al., 2015) proposes a three-stage dimensionality reduction technique for microarray data classification using a comparative study of four different classifiers, multiple linear regression (MLR), artificial neural network (ANN), $k$-nearest neighbor ($k$-NN), and naïve Bayesian classifier to observe the improvement in performance. In their experiments, the authors reduce the dimension without compromising the performance of such models. (Deegalla et al., 2012) proposed a dimensionality reduction method that employs classification approaches based on the $k$-nearest neighbor rule. The effectiveness of the reduced set is measured in terms of the classification accuracy. This method attempts to derive a minimal consistent set, i.e., a minimal set which correctly classifies all the original samples (Whelan et al., 2010). (Venugopalan et al., 2014) discussed the ongoing work in the field of pattern analysis for bio-medical signals (cardio-synchronous waveform) using a Radio Frequency Impedance Interrogation (RFII) device for the purpose of user identification. They discussed the feasibility of reducing the dimensions of these signals by projecting them into various sub-spaces while still preserving inter-user discriminating information, and they compared the classification performance using traditional dimensionality reduction methods such as PCA, independent component analysis (ICA), random projections, or $k$-SVD-based dictionary learning. In the majority of cases, the authors see that the space obtained based on classification carries merit due the dual advantages of reduced dimension and high classification.

Developing effective clustering methods for high-dimensional datasets is a challenging task (Whelan et al., 2010). (Boutsidis et al., 2015) studied the topic of dimensionality reduction for $k$-means clustering that encompasses the union of two approaches: 1) A feature selection-based algorithm selects a small subset of the input features and then the $k$-means is applied on the selected features. 2) A feature extraction-based algorithm constructs a small set of new artificial features and then the $k$-means is applied on the constructed features. The first feature extraction method is based on random projections and the second is based on fast approximate SVD factorization. (Sun et al., 2014) devel-

oped a tensor factorization based on a clustering algorithm (k-mean), referred to as Dimensionality Reduction Assisted Tensor Clustering (DRATC). In this algorithm, the tensor decomposition is used as a way to learn low-dimensional representation of the given tensors and, simultaneously, clustering is conducted by coupling the approximation and learning constraints, leading to the PCA Tensor Clustering and Non-negative Tensor Clustering models.

In this study, we develop a sampling-based dimensionality reduction technique that can deal with very high-dimensional datasets. The proposed approach takes into account the heterogeneous aspects of the data, and it models the different multivariate data distributions using the theory of Copulas. It maintains the integrity of the original information and reduces effectively and efficiently the original high-dimensional datasets.

## 3. Basic Concepts

This section aims to introduce the basic concepts used in our approach. Our dimensionality reduction technique is based on probabilistic and sampling models, therefore, one needs to recall some fundamental concepts. These include the notions of a Probability Density Function (PDF), Cumulative Distribution Function (CDF), a random variable used to generate samples from a probability distribution, a Copula to model dependencies of data without imposing constraints to specific types of marginal probability density functions, and dependence and rank correlations of multivariate random variables to measure dependencies of the dimensions. In the following table we give the basic notations used throughout this paper.

Table 1: Primitives and their definitions

| Primitive | Definition |
|---|---|
| $X$ | $n \times m$ data matrix (random variable). |
| $X^i$ | $i^{th}$ row of the matrix $X$. |
| $X_j$ | $j^{th}$ column of the matrix $X$. |
| $F_j(.)$ | CDF of the $j^{th}$ column. |
| $f_j(.)$ | PDF of the $j^{th}$ column. |
| $C$ | Gaussian Copula of the matrix $X$. |
| $c$ | Density associated with C. |
| $C_{ij}$ | Empirical Copula of the matrix $X$. |
| $\Sigma$ | Correlation matrix of C. |
| $X^t$ | Transposed matrix of $X$. |
| $v_{ij}$ | Value of the $i^{th}$ row and $j^{th}$ column. |

Let $f$ be the Probability Density Function (PDF) of a random variable $X$. The probability distribution of $X$ consists in calculating the probability $P(X_1 \leq x_1, X_2 \leq x_2, ..., X_m \leq x_m)$, $\forall (X_1, ..., X_m) \in R^m$. It is completely specified by the CDF $F$ which is defined in (Rubinstein & Kroese, 2011) as follows:

$$F(x_1, x_2, ..., x_m) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_m \leq x_m) \quad (1)$$

### *3.1. Random Variable Generation*

In this section, we address the problem of generating a sample from a one-dimensional cumulative distribution function $CDF$ by calculating the inverse transform sampling. To illustrate the problem, let $X$ be a continuous random variable with a $CDF$ $F(x) = P[X \leq x]$, and $U$ be a continuous uniform distribution over the interval $[0, 1]$. The transform $X = F^{-1}(U)$ denotes the inverse transform sampling function of a given continuous uniform variable $U = F(X)$ in $[0, 1]$, where $F^{-1}(u) = min\{x, F(x) \geq u\}$ (Rubinstein & Kroese, 2011). So the simple steps used for generating a sample $X \sim F$ are given as follows (Rubinstein & Kroese, 2011):

1. Generate $U \sim U[0, 1]$;
2. Return $X = F^{-1}(U)$.

The usual problem is how to combine one-dimensional distribution functions to form multivariate distributions and how to estimate and simulate their density $f(x_1, x_2, ..., x_m)$ to obtain the required number of random samples of $X_{i,i=1,...,m}$, especially in high-dimensional spaces. This problem will be explained in the following section.

### *3.2. Modeling with Copulas*

The first usage of Copulas is to provide a convenient way to generate correlated multivariate random variable distributions and to present a solution for the difficulties of transformation of the density estimation problem.

To illustrate the problem of invertible transformations of $m$-dimensional continuous random variables $X_1, ..., X_m$ according to their $CDF$, into $m$ independently uniformly-distributed variables $U_1 = F_1(X_1), U_2 = F_2(X_2), ..., U_m = F_m(X_m)$, let $f(x_1, x_2, ..., x_m)$ be the probability density function of $X_1, ..., X_m$, and let $c(u_1, u_2, ..., u_m)$ be the joint probability density function of $U_1, U_2, ..., U_m$. In general, the estimation of the probability density

function $f(x_1, x_2, ..., x_m)$ can provide a nonparametric form (unknown families of distributions). In this case, we estimate the probability density function $c(u_1, u_2, ..., u_m)$ of $U_1, U_2, ..., U_m$ instead of that $X_1, ..., X_m$ to simplify the density estimation problem, and then simulate it to achieve the random samples $X_1, ..., X_m$ by using the inverse transformations $X_i = F_i^{-1}(U_i)$.

Sklar's Theorem showed that there exists a unique m-dimensional Copula $C$ in $[0, 1]^m$ with standard uniform marginal distributions $U_1, ..., U_m$. (Nelsen, 2007) states that every distribution function $F$ with margins $F_1, ..., F_m$ can be written $\forall(X_1, ..., X_m) \in \mathbb{R}^m$ as:

$$F(X_1, ..., X_m) = C(F_1(X_1), ..., F_m(X_m)). \quad (2)$$

To evaluate the suitability of a selected Copula with estimated parameter and to avoid the introduction of any assumptions on the distribution $F_i(X_i)$, one can utilize an empirical $CDF$ of a marginal $F_i(X_i)$, to transform $m$ samples of $X$ into $m$ samples of $U$. An empirical Copula is useful for examining the dependence structure of multivariate random vectors. Formally, the empirical Copula is given by the following equation:

$$C_{ij} = \frac{1}{m}(\sum_{k=1}^{m} I_{(v_{kj} \leq v_{ij})}), \quad (3)$$

where the function $I_{(arg)}$ is the indicator function, which equals 1 if $arg$ is true and 0 otherwise. Here, $m$ is used to keep the empirical $CDF$ less than 1, where $m$ is the number of observations. In the following, we will focus on the Copula that results from a standard multivariate Gaussian Copula.

### *3.3. Gaussian Copula*

The difference between the Gaussian Copula and the joint normal $CDF$ is that the Gaussian Copula allows to have different marginal $CDF$ types from the joint distribution (Nelsen, 2007). However, in probability theory and statistics, the multivariate normal distribution is a generalization of the one-dimensional normal distribution. The Gaussian Copula is defined as follows:

$$C(\Phi(x_1), ..., \Phi(x_m)) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp(\frac{-1}{2} X^t(\Sigma^{-1} - I)X). \quad (4)$$

where $\Phi(x_i)$ is the CDF standard Gaussian distribution of $f_{i(x_i)}$, i.e., $X_i \sim N(0, 1)$, and $\Sigma$ is the correlation matrix. The resulting Copula $C(u_1, ..., u_m)$ is called Gaussian Copula. The density associated with $C(u_1, ..., u_m)$ is obtained with the following equation:

$$c(u_1, ..., u_m) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp\left[\frac{-1}{2}\xi^t(\Sigma^{-1} - I)\xi\right], \quad (5)$$

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

### *3.4. Dependence and Rank Correlation*

Since the Copula of a multivariate distribution describes its dependence structure, it might be appropriate to use measures of dependence which are Copula-based. The Pearson correlation measures the relationship $\Sigma = cov(X_i, X_j)/(\sigma_{X_i}\sigma_{X_j})$ where $cov(X_i, X_j)$ is the covariance of $X_i$ and $X_j$ while $\sigma_{X_i}, \sigma_{X_j}$ are the standard deviations of $X_i$ and $X_j$.

Kendall rank correlation (also known as Kendall's coefficient of concordance) is a nonparametric test that measures the strength of dependence between two random samples $X_p^i, X_{p'}^i$ of $n$ observations. The notion of concordance can be defined by the following equation:

$$\begin{aligned}\tau = P[(X_p^i - X_p^j)(X_{p'}^i - X_{p'}^j) > 0] - \\ P[(X_p^i - X_p^j)(X_{p'}^i - X_{p'}^j) < 0].\end{aligned} \quad (6)$$

For the Gaussian Copula, Kendall's $\tau$ can be calculated as follows:

$$\tau = \frac{2}{\pi} \arcsin \Sigma_{X_i X_j}. \quad (7)$$

## 4. System Modeling

To overcome the problem of reducing a large set of variables, we will identify and remove the redundant dimension and variables which are linear combinations of others. In this section, we will show how to use a binary linear programming formulation to find a lower linear space of dimensions (columns $X_j$) of the original matrix for a maximization of redundant columns. The idea is that for a given set of data $X = X_1, X_2, ..., X_m$, redundancies exist which may be eliminated while retaining most of the relevant dimension. After eliminating these redundancies, we can represent the data nearly as completely in a $k$-dimensional space.

*4.1. Decision Variables*

Let $Y = Y_1, Y_2, ..., Y_m$ be the decision variables, i.e., $Y_{j,j=1,...,m}$ takes 0/1 indicator variables for each column $X_j \in X$, where $Y_j = 1$ indicates that $X_j$ is redundant, $Y_j = 0$ otherwise.

We make $Y_j = 1$ to reduce the redundancy among the variables of a given dataset $X$ as in the original $m$ dimensions, by eliminating dependent columns, requiring measures of dependence which are based on the multivariate Gaussian Copula.

$$Y_j = \begin{cases} 1, & \text{if } X_j \text{ is redundant} \\ 0, & \text{otherwise.} \end{cases}$$

Considering the key relation in the theory of Copula under absolute continuity assumptions of dimensions, the correlation matrix fully characterizes the joint distribution of dimensions $X_{j,j=1,...,m} \in X$. To decide the dependence between the continuous dimensions, the correlation matrix of different dimensions is evaluated, and a threshold value of the correlation matrix is considered to compare the dimensions $X_1, X_2, ..., X_m$. Note, that the correlation matrix ($\Sigma$) is the $m \times m$ data matrix, and we assumed that $Y$ has always the same dimension as $\Sigma$. Throughout this development, $Y$ will be a $m \times m$ data matrix containing 0/1 indicator variables which represent clearly the dependence comparison of the redundancies data. In the next part 4.2, we will define the objective function to model the redundancy problem to the optimality by eliminating less important dimensions taking the value 1 indicator variable in $Y$.

*4.2. Objective function*

After defining the decision variables, our main goal is to transform the high-dimensional problem into a low-dimensional space with an optimal solution maximizing the number of columns to be eliminated. The redundant columns are defined by the function $f(Y_1, ..., Y_m) = \sum_{j=1}^{m} Y_j$, where $Y_j \in \{0, 1\}, j = 1, ..., m$. Our objective can be formulated as follows:

$$\begin{aligned} &Max(\sum_{j=1}^{m}(Y_j)). \\ &Y_j \in \{0, 1\}, j = 1, ..., m. \end{aligned} \tag{8}$$

The low-dimensional of $X_1, X_2, ..., X_m$ is obtained by maximizing the less important dimensions which are represented by the sum of $Y = Y_1, Y_2, ..., Y_m$ under the set of constraints that will be defined in subsection 4.3, where the components of $Y_{j,j=1,...,m}$ are the binary decision variables.

*4.3. Constraints*

In order to eliminate the $(m - k)$ dimensional data redundancies and providing an optimal subspace, preserving the linear independence between $X_{i,i=1,...,k}$, the objective function can be optimized initially under the following constraints:

$$\begin{cases} \sum\limits_{k \in B_c^{(*)}} \alpha_k X_k = 0 \Leftrightarrow \alpha_k = 0; \forall k \in B_c^{(*)} \\ \\ B_c^{(*)} = \{j \in \{1, ..., m\} \ / \ Y_j = 0\}. \\ \\ Y_j \in \{0, 1\}; \forall j \in \{1, ..., m\}. \end{cases} \tag{9}$$

where $k$ denotes the dimension of the subspace of $B_c^{(*)}$, and $\alpha$ is a vector representing the coefficients of the linear combination of dimensions.

The first constraint consists to verify the linear independence of the dimensions $X_{i,i=1,...,k}$ belonging to the subset $B_c^{(*)}$, i.e., $\alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k = 0$, where the only solution to this system of equations is: $\alpha_1 = 0, \ldots, \alpha_k = 0$.

In the second constraint, $Y_j = 0$ indicates that $X_{i,i=1,...,k}$, is not redundant. Given a set of vectors $X_1, ..., X_k \in B_c^{(*)}$ with $Y_j = 0$. $B_c^{(*)}$ provides an optimal linearly independent dimension matrix in $k$-dimensional space, where $k < m$.

The third constraint represents the integrality constraint, that shows that $Y_j$ is the m-dimensional vector of binary decision variables.

Solving this optimization problem is complex, especially in the Big Data setting. Future research will be towards the investigation of a more formal approach for the determination of a solution. Also it would be interesting to examine the possibility of using metaheuristics or a hybrid approach. Therefore, we felt inspired by this mathematical model to propose a new solution based on sampling and algebraic methods, to treat the problem of dimensionality reduction in very large datasets, which will be presented in the next section.

## 5. Proposed Approach

The approach presented in this paper for dimensionality reduction in very large datasets is based on the theory of Copulas and the LU-decomposition method (Forward Substitution). The main goal of the method is to reduce the dimensional spaces of data without losing important/interesting information. On the other hand, the goal is to estimate the multivariate joint probability distribution without imposing constraints on specific types of marginal distributions of dimensions. Figure 1 shows an overview of the proposed reduction method which operates in two main steps.

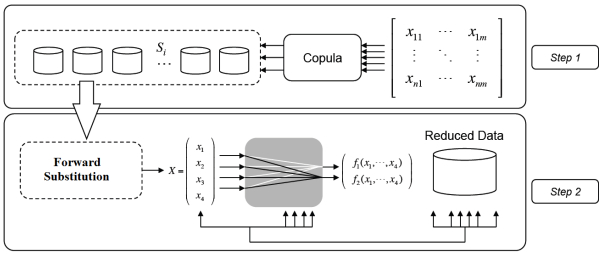



Figure 1: Overview of the proposed reduction method.

In the first step, large raw datasets are decomposed into smaller subsets when calculating the data dependencies using a Copula by taking into account heterogeneous data and removing the data which are strongly dependent. In the second step, we want to reduce the space dimensions by eliminating dimensions that are linear combinations of others. Then we will find the coefficients of the linear combination of dimensions by applying the LU-decomposition method (Forward Substitution) to each subset to obtain an independent set of variables in order to improve the efficiency of data mining algorithms. The two different steps of the proposed method are as follows (See also Figure 1):

- Step 1: Construction of dependent sample subsets $S_{i,(i=1,\ldots,k')}$

  In order to decompose the real-world dataset into smaller dependent sample subsets, we will consider the vectors which are linearly dependent in the original data.

  We first calculate the empirical Copula to better observe the dependencies between variables. According to the marginal distributions from the observed and approved empirical Copula, we can determine the theoretical Copula, that links univariate marginal distributions to their joint multivariate distribution function, and then we will regroup dimensions having the strong correlation relationship in each sample subset $S_{i,(i=1,\ldots,k')}$ by estimating the parameters of the Copula. In this paper, we have presented the Gaussian Copula that corresponds to our experimental results. An illustration of the Copula method is given in Figure 2.

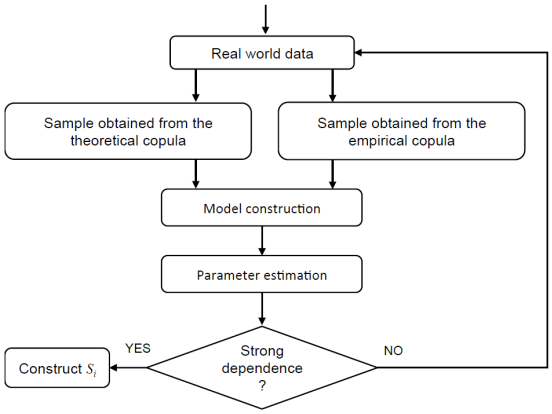



Figure 2: Construction of the subsets $S_{i,(i=1,\ldots,k')}$.

  The dependence between two continuous random variables $X_1$ and $X_2$ is defined as follows: If the correlation parameter $\rho$ is greater than 0.5, then $X_1$ and $X_2$ are positively correlated, meaning that the values of $X_1$ increase as the values of $X_2$ increase (i.e., the more each attribute implies the other). Hence, a higher value may indicate that $X_1$ and $X_2$ are positively dependent, and probably have a highly redundant attribute, then these two samples will be made as in the same subset $S_{i,(i=1,\ldots,k')}$. When the parameter of the Copula $\rho$ of the two continuous random variables $X_1$ and $X_2$ is greater than 0.7, then $X_1$ and $X_2$ have a strong dependence. If the resulting value is equal or less than 0, then $X_1$ and $X_2$ are independent and there is no correlation between them.

  The output of the sample subset $S_{i,(i=1,\ldots,k')}$ represents a matrix that retains only dependent samples of the original matrix in order to detect, and remove a maximum of the redundant dimensions, which are linear combinations of others, in the second step.

- Step 2: LU-decomposition method

  The key idea behind the use of the For-

ward Substitution method is to solve the linear system equations as given by the samples $S_{i,(i=1,...,k')}$ with an upper-triangular coefficient matrix in order to find the coefficients of linear sample combinations and to provide a low linear space $(X_{i,i=1,...,k})$ of the original matrix as shown in Figure 3.

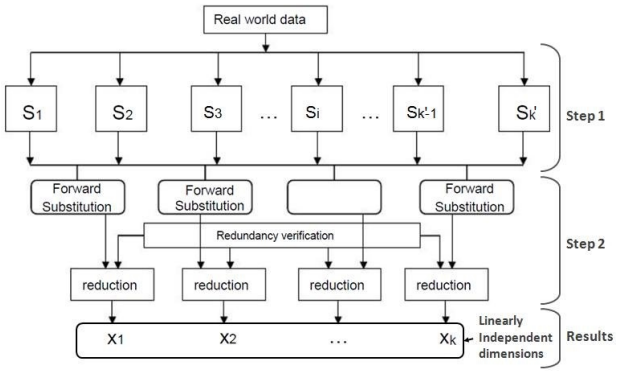



Figure 3: Schema of dimensionality reduction.

The LU decomposition method is an efficient procedure for solving a system of linear equations $\alpha \times \acute{S} = C$, and it can help accelerate the computation. When $C$ is a column vector in the dependent sample subsets $S_{i,(i=1,...,k')}$, and $\alpha_j$ is an output vector representing the relationship between dimensions or the coefficients of the linear combination of dimensions, $\acute{S}_{i,(i=1,...,k'-1)}$ induces a lower triangular matrix without column $C$. We conclude that each matrix $\acute{S}_{i,(i=1,...,k'-1)}$ induces a lower triangular matrix of the following form:

$$(\acute{S}) \begin{cases} \alpha_1 x_{11} & = & c_1 \\ \alpha_1 x_{21} + \alpha_2 x_{22} & = & c_2 \\ \vdots & & \vdots \\ \alpha_1 x_{n1} + \alpha_2 x_{n2} + ... + \alpha_n x_{nn} & = & c_n \end{cases}$$

From the above equations, we see that $\alpha_1 = c_1/x_{11}$. Thus, we compute $\alpha_1$ from the first equation and substitute it into the second to compute $\alpha_2$,..., etc. Repeating this process, we reach equation $i$, $2 \leq i \leq n$, using the following formula:

$$\alpha_i = \frac{1}{x_{ii}} \left[ c_i - \sum_{j=1}^{i-1} \alpha_j x_{ij} \right], i = 2, ..., n. \quad (10)$$

The algorithm 1 used for this resolution makes $(n \times (n-1))/2$ additions and subtractions, $(n \times (n-1))/2$ multiplications and $n$ divisions to calculate the solution, a global number of operations in the order of $n^2$.

**Algorithm 1** Dimensionality linear combination reduction method

**Input:** Vector $C$ and a lower triangular matrix $\acute{S}$;
**Output:** Vector $\alpha$.
begin
  $\alpha_1 = c_1/x_{11}$
  **for i := 2 to** $n$ **do**
  $\alpha_i = c_i$;
    **for j := 1 to** $i-1$ **do**
    $\alpha_i = \alpha_i - x_{ij}\alpha_j$
    end
  $\alpha_i = \alpha_i/x_{ii}$
end

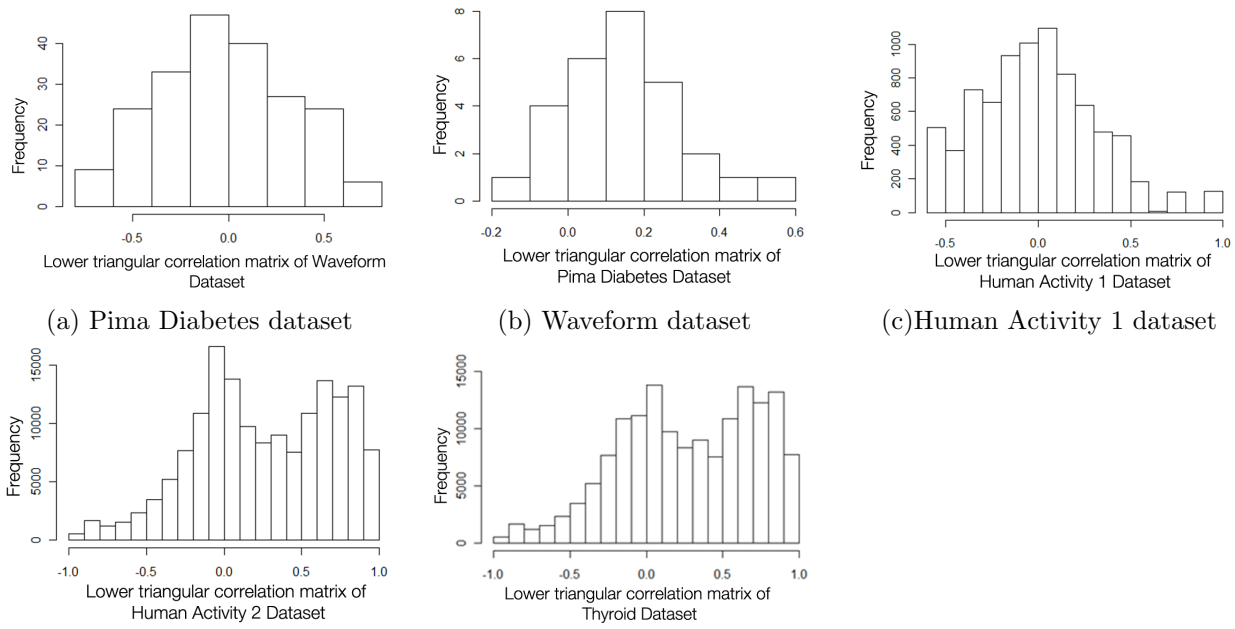
## 6. EXPERIMENTAL RESULTS

Our experiments were performed on the following real-world datasets taken from the machine learning repository (Lichman, 2013), which are from different application domains, including the Healthcare dataset, which have proven helpful in both medical diagnoses and in improving our understanding of the human body.

### *6.1. Data Source*

We have selected four different datasets whose characteristics are discussed below.

*Pima Diabetes Database.* The dataset was selected from a larger dataset held by the National Institute of Diabetes and Digestive and Kidney Diseases. All patients in this database are Pima-Indian women at least 21 years old. There are 268 (34.9%) cases which had a positive test for diabetes and 500 (65.1%) which had a negative such test. Some clinical cases are also used in this database: 1) Number of times pregnant, 2) Plasma glucose concentration after 2 hours in an oral glucose tolerance test, 3) Diastolic blood pressure (mm Hg), 4) Triceps skin fold thickness (mm), 5) 2-Hour serum insulin (mu U/ml), 6) Body mass index, 7) Diabetes pedigree function and 8) Age.




(a) Pima Diabetes dataset (b) Waveform dataset (c)Human Activity 1 dataset

(d)Human Activity 2 dataset (e) Thyroid Disease dataset

Figure 4: Global histograms of correlation coefficients for Pima Diabetes, Waveform, Human Activity Recognition using Smartphone 1 and 2, and Thyroid Disease datasets.

*Waveform Database.* These are data given by David Aha in the year 1988. They are generated by a waveform database generator. The database contains 33367 instances (rows) and 21 attributes with continuous values between 0 and 6. The main goal of our analysis is to reduce the redundant waves or those which are a combination of other waves.

*Human Activity Recognition using Smartphone Datasets.* The experiments have been carried out with a group of 30 volunteers within an age bracket of 19 and 48 years. Each person performed six activities (walking, walking-upstairs, walking-downstairs, sitting, standing, laying) carrying a smartphone (Samsung Galaxy S II) on the waist.

The goal of our analysis is to reduce the redundancy between the variables for each database. The descriptions of the two datasets used are given as follows :

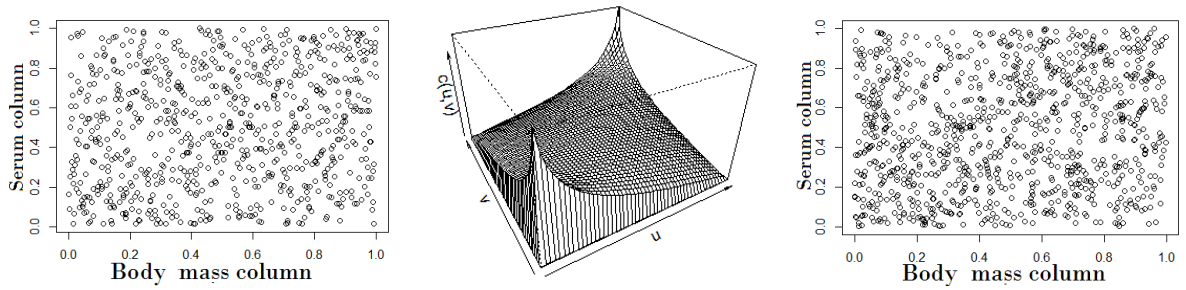
1. Total-acc-x-train: The first dataset was obtained from a study on the acceleration signal in standard gravity which includes 7352 rows and 128 attributes.

2. Body-acc-x-train: The second dataset was obtained from a study on the sensor signals (accelerometer and gyroscope). The sensor acceleration signal, which has gravitational and body motion components, was separated into body acceleration and gravity using a Butterworth low-pass filter. This database includes 7352 rows and 384 attributes.

*Thyroid Disease Diagnosis Problem.* A Thyroid dataset contains measurements of the amounts of different hormones produced by the thyroid gland. The UCI machine learning directory contains 6 database versions. We chose a dataset that contains three types of thyroid diagnosis problems which are assigned to the values that correspond to hyper-thyroid, hypo-thyroid and normal function of the thyroid gland. Each type has different attributes that contain patients information and laboratory tests. The information about these attributes includes: age, sex, thyroxine, query on thyroxine, antithyroid-medication, sick, pregnant, thyroid surgery, I131 treatment, query hyperthyroid, lithium, goiter, tumor, hypopituitary, psych, TSH, T3, T4U, FTI, BGreferral source: WEST, STMW, SVHC, SVI, SVHD and others. This database includes 7200 (patients) rows and 561 attributes (patients information and laboratory tests).

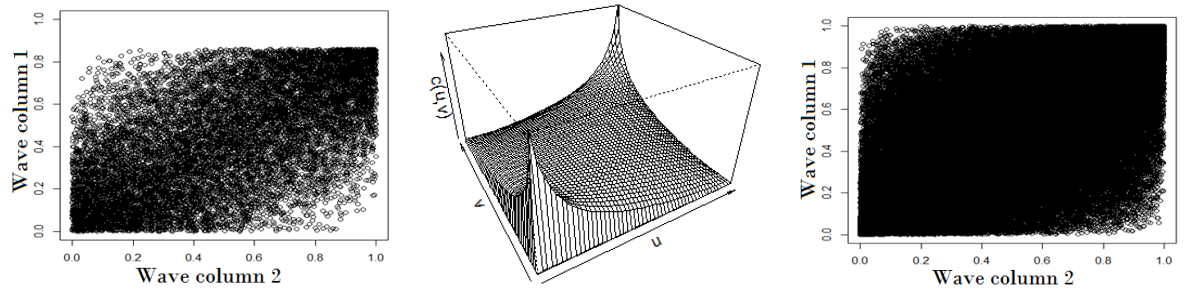
### *6.2. Validation of the Gaussian Copula hypothesis*

Figures (5, 6, 7, 8, 9) (A), show the empirical Copula samples of each database used, generated



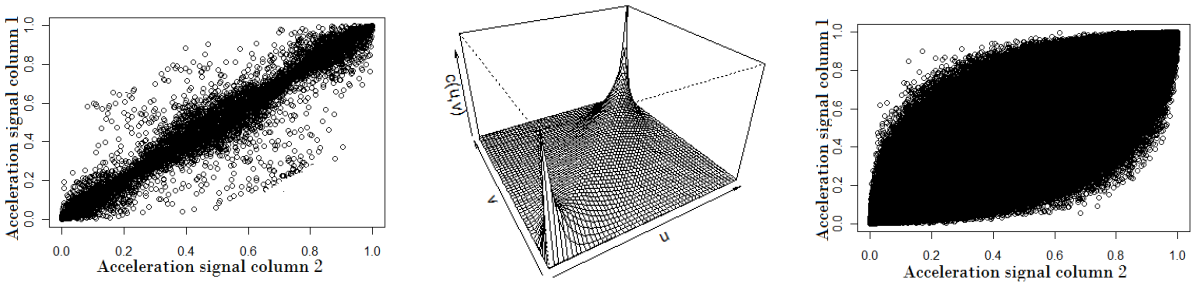

(A) Empirical Copula (B) Copula density (C) Theoretical Copula

Figure 5: Bivariate empirical and theoretical Copula ($\rho = 0.39$) with the corresponding densities of the Pima Diabetes dataset.



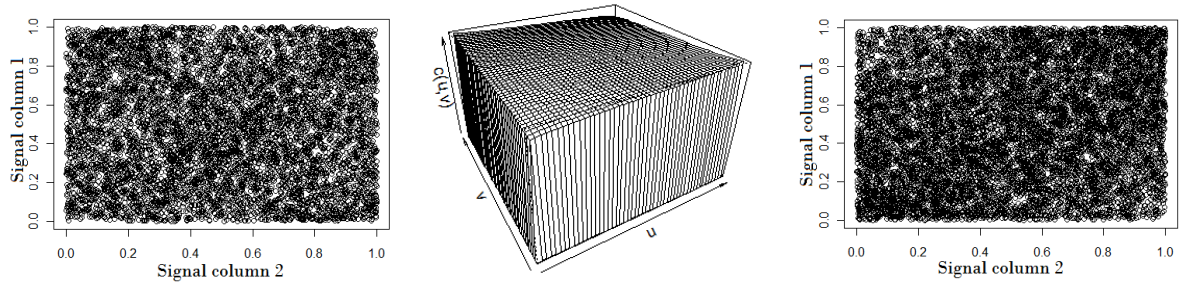

(A) Empirical Copula (B) Copula density (C) Theoretical Copula

Figure 6: Bivariate empirical and theoretical Copula ($\rho = 0.508$) with the corresponding densities of the Waveform dataset.



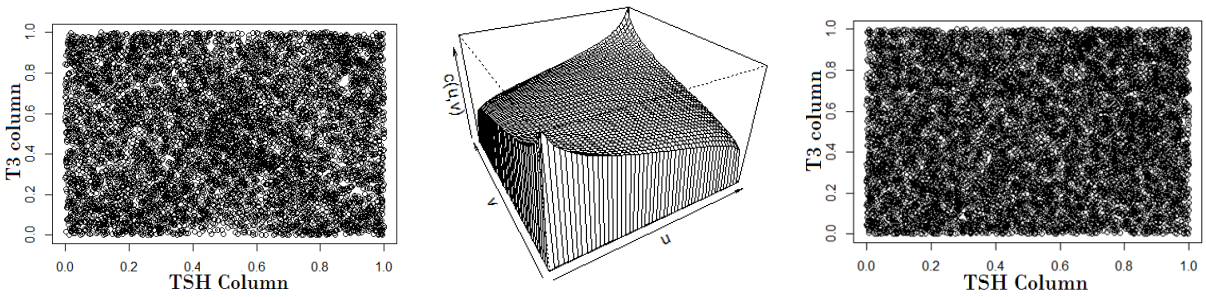

(A) Empirical Copula (B) Copula density (C) Theoretical Copula

Figure 7: Bivariate empirical and theoretical Copula ($\rho = 0.9$) with the corresponding densities of the Human Activity Recognition using Smartphone dataset 1.




(A) Empirical Copula (B) Copula density (C) Theoretical Copula

Figure 8: Bivariate empirical and theoretical Copula ($\rho = -0.012$) with the corresponding densities of the Human Activity Recognition using Smartphone dataset 2.




(A) Empirical Copula (B) Copula density (C) Theoretical Copula

Figure 9: Bivariate empirical and theoretical Copula ($\rho = 0.14$) with the corresponding densities of the Thyroid Disease dataset.

with the formula(3) as given in the previous subsection 3.2. By comparing each empirical bivariate Copula distribution with the bivariate Gaussian Copula distribution, we can observe that they have the same parametric form of a Gaussian Copula (Figures (5, 6, 7, 8, 9 (C)). In order to formally select the appropriate Copula, we use the Kolmogorov-Smirnov (K-S) goodness-of-fit test. K-S is applicable to any bivariate or multi-dimensional Copula. This test can be used to quantify a distance between the empirical distribution samples and the cumulative distribution function to decide whether they have the same distribution. In order to verify that the global Copula is Gaussian (case of our results), K-S goodness-of-fit test compares the empirical Copula obtained based on the $m$ data samples with the standard multivariate Gaussian Copula generated with the similar correlation matrix of the empirical Copula, then decide if both Copulas come from the same family of Copulas.

### *6.3. Dependence Structure*

Different spreads are explained by the different levels of dependence existing between each pair. As shown in Figure 4, various combinations of correlation between the variables of the two versions of Human Activity Recognition using Smartphone datasets, and the Thyroid Disease dataset are positive, and, therefore, these variables are dependent. On the other side, the plot of a Copula density can be used to examine the dependence. An illustration of some Copulas (densities) is given in Figures (5, 6, 7, 8, 9)(B). We noticed for Figures (8, 9) (B) (with corresponding correlation coefficients $\rho = -0.012$, $\rho = 0.14$), that the variables $X_1$ and $X_2$ are independent; the corresponding Copula density is a horizontal surface, indicating equal probability in any pair $(u, v)$. On the other hand, when $X_1$ and $X_2$ are positively dependent ($\rho \geq 0.5$), as shown in Figures 6(B) and 7(B), most of the Copula density will fall upon the main diagonal.

### *6.4. Data reduction*

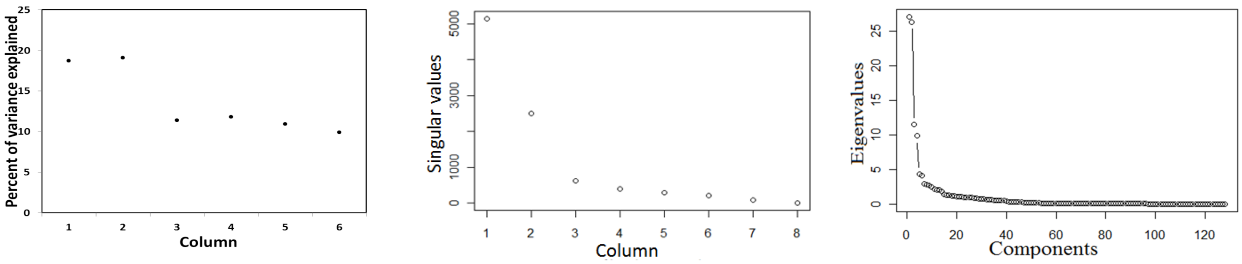
By implementing different reduction techniques like SVD, PCA, SPCA and our approach, Figures (10, 11, 12, 13, 14), and Figures (5, 6, 7, 8, 9)(C), show the graphical results obtained without data reduction.

In these Figures (5, 6, 7, 8, 9) (C), we have generated a sample from a standard multivariate Gaussian distribution that has the same correlation matrix as the sample of the empirical Copula, where the correlation matrix is shown as a histogram in Figure 4.

The databases were also analysed using the SVD method. We have shown the diagonal matrix $S$, as obtained for each database, in Figures (10, 11, 12, 13, 14)(B). The diagonal entries $(s_1, s_2, ....., s_m)$ of the matrix $S$ have the property $s_1 \geq s_2 \geq .... \geq s_m$. The reduction process with SVD is performed by retaining only the $r \ll m$ positive non-zero eigenvalues on the diagonal matrix.

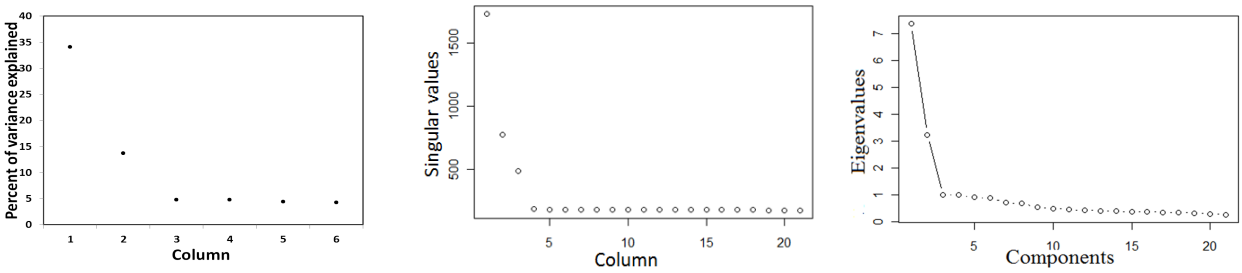
From the results obtained with PCA (Figures (10, 11, 12, 13, 14)(C)), we have subtracted the mean from each row vector of datasets and computed the co-variance matrix to calculate eigenvectors and eigenvalues of each dataset. The reduction process with the PCA method is performed by the Kaiser rule, in which we drop all components with eigenvalues under 1.

The reduction process with SPCA is performed by the number of non-zero loadings and the percent of explained variance exploiting the LASSO technique interpreting the principal component loadings as a regression problem. So, many coefficients of the principal components become zero and it is easier to



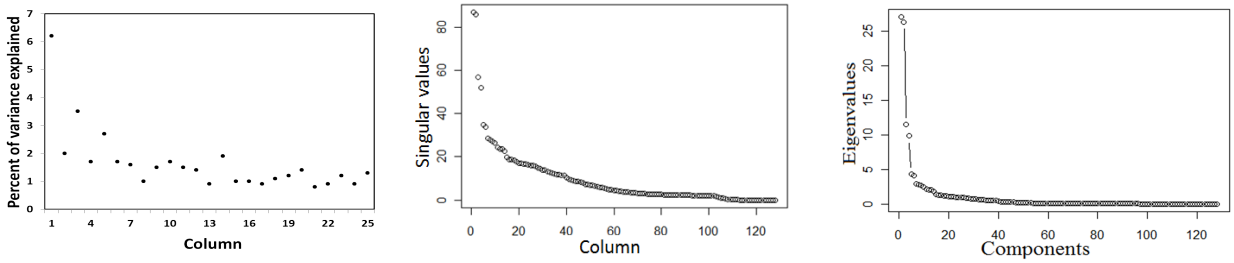

(A) SPCA (B) SVD (C) PCA

Figure 10: Results obtained from SPCA, SVD, PCA, of the Pima Diabetes dataset before dimensionality reduction.



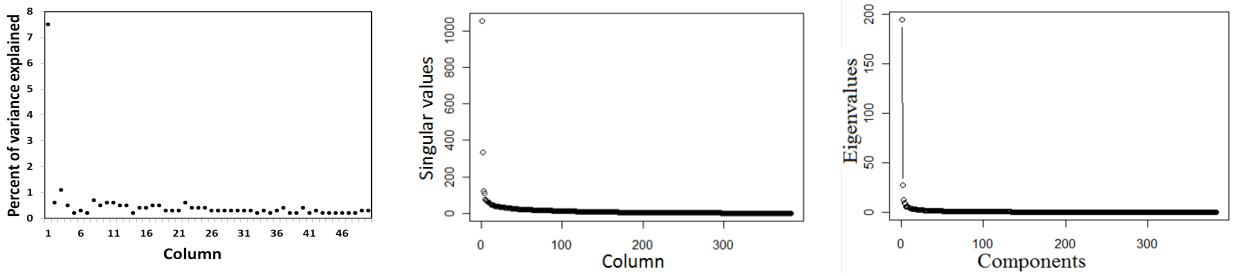

(A) SPCA (B) SVD (C) PCA

Figure 11: Results obtained from SPCA, SVD, PCA, of the Waveform dataset before dimensionality reduction.



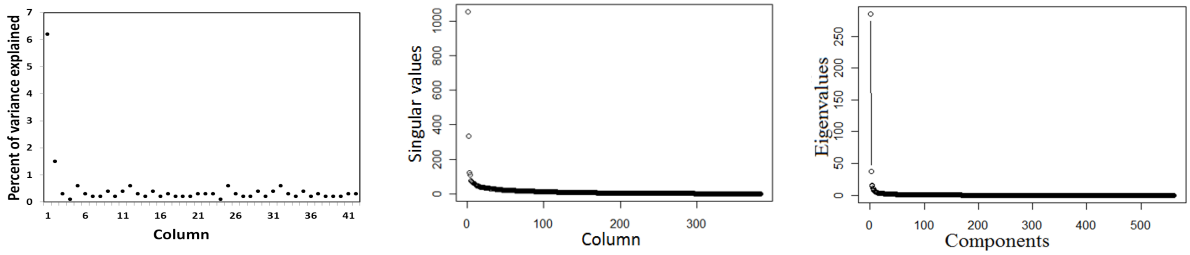

(A) SPCA (B) SVD (C) PCA

Figure 12: Results obtained from SPCA, SVD, PCA, of the Human Activity Recognition using Smartphone dataset 1 before dimensionality reduction.




(A) SPCA (B) SVD (C) PCA

Figure 13: Results obtained from SPCA, SVD, PCA, of the Human Activity Recognition using Smartphone dataset 2 before dimensionality reduction.




(A) SPCA (B) SVD (C) PCA

Figure 14: Results obtained from SPCA, SVD, PCA, of the Thyroid Disease dataset before dimensionality reduction.

extract meaningful variables. (Figures (10, 11, 12, 13, 14)(C)) show the obtained percent of explained variance.

The numerical results obtained for dimensionality reduction are shown in Table 2.

Table 2: Dimensionality reduction results (number of columns reduced.)

| Attribute Dimension Reduction | | | | |
|---|---|---|---|---|
| Methods | SVD | PCA | SPCA | PA |
| Pima Diabetes | 0 | 4 | 2 | 5 |
| Waveform | 5 | 15 | 15 | 17 |
| Human Activity 1 | 23 | 102 | 93 | 107 |
| Human Activity 2 | 112 | 337 | 338 | 339 |
| Thyroid Disease | 180 | 500 | 519 | 520 |

*6.4.1. Interpretation of the results*

According to Table 2, SVD provides the lowest number of reduced dimensions for the global dimension reduction and all databases, i.e., it reduces 0 attributes for the Pima Diabetes dataset, 5 attributes for the Waveform dataset, 23 for the Human Activity 1 dataset, 112 for the Human Activity 2 dataset and 180 attributes for the Thyroid dataset. We can observe that SPCA works very well and has good results, finds spaces with a lower dimension than those obtained with SVD but these methods are far from being optimal. We can see that in all cases, PCA and SPCA reduce respectively, 4 and 2 attributes for the Pima Diabetes dataset, 15 attributes for the Waveform dataset, 102 and 93 for the Human Activity 1 dataset, 337 and 338 for the Human Activity 2 dataset, and 500, 519 attributes for the Thyroid dataset. The results obtained by the proposed approach (PA) are much better than SVD, PCA and also SPCA for all datasets as shown in Table 2. It reduces 5 attributes for the Pima Diabetes dataset, 17 attributes for the Waveform dataset, 107 for the Human Activity 1 dataset, 339 for the Human Activity 2 dataset and 520 attributes for the Thyroid dataset. The results of our approach overcome the main weaknesses of SVD, PCA, and SPCA in a large database as follows:

- The susceptibility to find an optimal low-dimensional space, to remove the redundant data and the variables which are linear combinations of others;
- The sensitivity of examining the different assumptions about the data distribution.

The reduction process with SPCA is performed by the number of non-zero loadings and the percent of explained variance exploiting the LASSO technique and interpreting the principal component loadings as a regression problem. So, many coefficients of the principal components become zero and it is easier to extract meaningful variables. (Figures (10, 11, 12, 13, 14)(C)), show the obtained percent of explained variance.

*6.5. Efficiency of the proposed method*

To improve the efficiency of the proposed method, we have used statistical and classification measure methods. The statistical precision and bias are combined to define the performance of the dimensionality reduction method, expressed in terms of the standard deviation of a set of results. The classification accuracy is obtained by using the full set of dimensions on one side and the reduced set of provided data in terms of precision and recall on the other side.

*Statistical precision.* The goal of this part is to test the statistical efficiency, after the final dimensionality reduction of all the databases using PCA, SVD,

SPCA, and the proposed approach. The most common precision measure is the standard deviation ($sd$) measured by the formula (11).

$$sd = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}, \qquad (11)$$

where $N$ is the number of values of the sample, and $\overline{x_i}$ is the mean of the values $x_i$. We have represented and evaluated the standard deviations of each dataset after the final dimensionality reduction of PCA, SVD, SPCA, and the proposed approach. A large standard deviation may not be good, biased and makes the results of the dimensionality reduction less precise.

Figure 15 shows the global standard deviation of PCA, SVD, SPCA, and our approach for Pima Diabetes, Waveform, Human Activity Recognition using Smartphone 1 and 2, and Thyroid Disease datasets. In order to have clear graphs and an easier data visualization, we have represented the results of SVD of Thyroid and Human Activity Recognition using Smartphone 2 dataset in Figure 15(f).

Based on the results exhibited in Figure 15, we note that: PCA has the lowest standard deviations for the Pima Diabetes and Waveform dataset (Figure 15(a)(b)) in comparison to SPCA and SVD. The maximum value of the standard deviation of the PCA results achieves 1.44 for the Pima Diabetes dataset and 2.17 to the Waveform dataset. However, SVD has a small bias for the Human Activity 1 dataset (Figure 15(c)(d)(f)), in comparison to SPCA and PCA. Its maximum value of the standard deviation achieves 0.32 for Human Activity 1. For the Thyroid dataset and Human Activity 2 (Figure 15(e)(f)), we noticed that SPCA has the lowest standard deviations, in Comparison to SVD and PCA. The maximum value of the standard deviation of the SPCA results achieves 2.74 for the Thyroid dataset and 2.4 for the Human Activity 2. The results obtained by our approach are much better than SVD, PCA, and also SPCA for all datasets as shown in Figure 15. In general, the results of all simulation studies have shown that the proposed method based on Copulas is more appropriate for dimensionality reduction in large datasets. The proposed approach based on Gaussian Copulas seems to be the best for practical use because it has the smallest bias and the standard deviations are more stable, i.e., the maximum values of standard deviation achieve 0.73 for the Pima Diabetes, 0.7 for the Waveform, 0.18 for the Human Activity 1, 0.52 for the Human Activity 2 and, finally, 0.54 for the Thyroid database. Our results overcome the main weaknesses of SVD, PCA and SPCA in a large database by finding the smallest bias and maintaining the integrity of the original information.

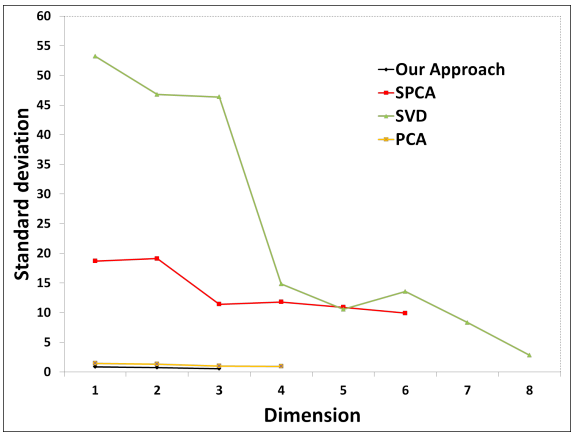
*Classification accuracy.* The goal of this part is to improve the effectiveness of dimensionality reduction before and after the final reduction of the dimensionality of Pima Diabetes and Waveform databases, by using the classication methods for PCA, SVD, Sparse PCA, and our proposed approach. The reason to use the two databases Pima Diabetes and Waveform databases in this simulation, is that we have the set of predefined classes. In both databases, we have chosen 70% of data as a training set and 30% as a testing set. Three different classication techniques such as Artificial Neural Network (ANN), $k$-nearest neighbors ($k$-NN), naïve Bayesian are used to evaluate the performance of this data reduction scheme in this paper.

- Artificial Neural Network (ANN) (Agatonovic-Kustrin & Beresford, 2000) uses the back-propagation method to train the network by adjusting the weights for each input data. After the learning, ANN constructs a mathematical relationship between the input training data and the correct class in order to predict the correct class of a new input instance.

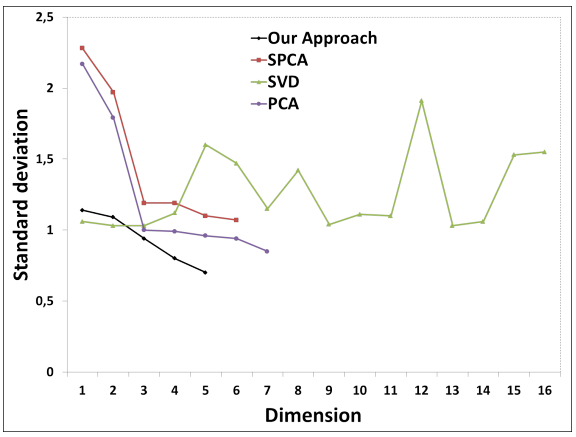
- $k$-nearest neighbors ($k$-NN) (Derrac et al., 2016) aims to classify each instance of the test sample according to its similarity with the examples of the learning set and returns the most frequent class among these k examples (neighbors).

- Naïve Bayesian (Saoudi et al., 2016) uses Bayes' rule to find the probability of a given instance from the test sample belonging to the specific class. The learning of this classifier is done by computing the mean and the variance of each dimension in each class.

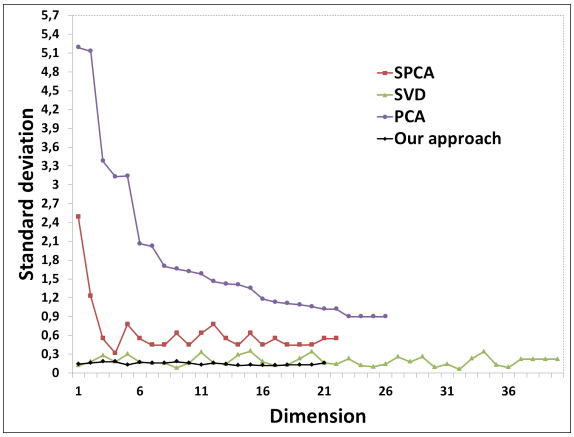
In general, the performance of a classification process can be evaluated by the following quantities: True Positives (TP), False Positives (FP), and False Negatives (FN), and the use of different metrics such as precision and recall. The precision



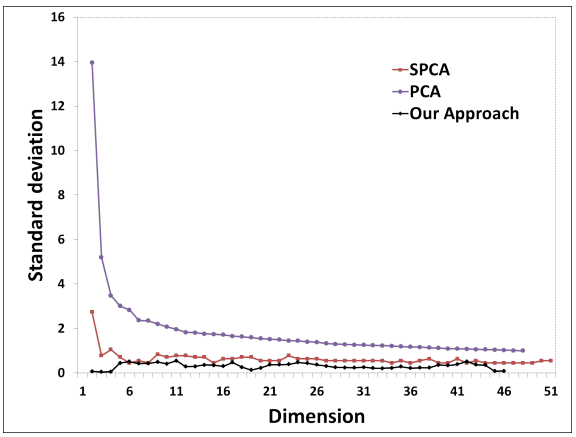

(a) Pima Diabetes dataset



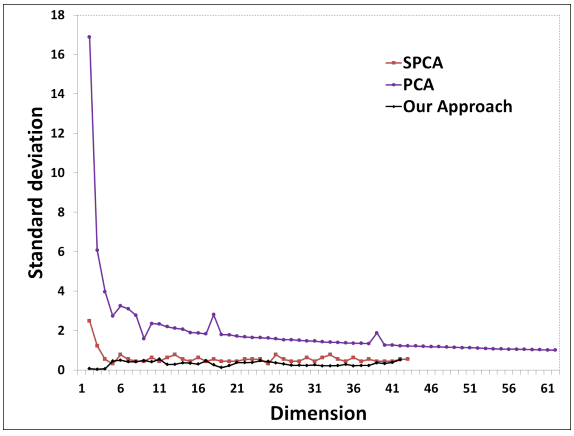

(b) Waveform dataset



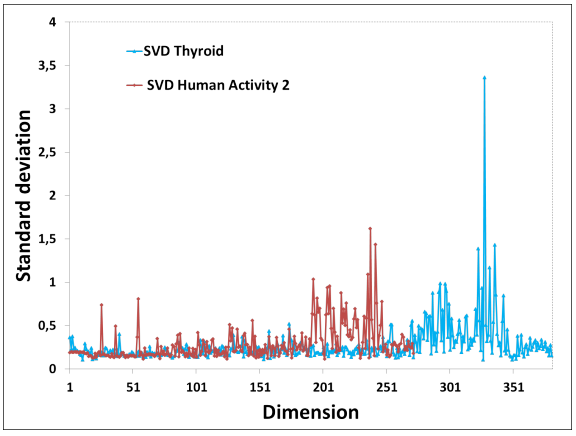

(c)Human Activity 1 dataset




(d) Human Activity 2 dataset




(e) Thyroid Disease dataset




(f) SVD results using Thyroid and Human 2 dataset

Figure 15: Global standard deviation of PCA, SVD, SPCA, and our approach for Pima Diabetes, Waveform, Human Activity Recognition using Smartphone 1 and 2, and Thyroid Disease datasets.

$P$ and the recall $R$ are measured by the following formulas:

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

Table 3 shows the results of ANN, $k$-NN and naïve Bayesian classication accuracy obtained on the Pima Diabetes and Waveform, by using on one side the full set of dimensions and on the other the reduced set provided by PCA, SVD, SPCA, and the proposed approach applied in terms of precision and recall.

*6.5.1. Interpretation of the results*

The ANN classifier accuracy is obtained by using the full set of dimensions on one side and the reduced set of provided data on the other. The results obtained with Pima Diabetes database show that the precision measure increases in the case of PCA, SPCA and our proposed approach when compared with the full data, and keeps the same result with SVD. The precision measure achieves 0.763 for PCA, 0.748 for SVD, 0.782 for SPCA and 0.783 for the proposed approach. But the SVD recall measure can be considered better than the full data, PCA and SPCA, since it achieves 0.757 for PCA, 0.722 for SVD, 0.774 for SPCA and 0.722 for the proposed approach. Observing the Waveform database in Table 3, it is clear that the proposed approach and PCA give better performance than both the SVD and SPCA methods, since it is 0.929 for PCA, 0.915 for SVD, 0.923 for SPCA, and 0.929 for the proposed approach. We see that SVD precision measure is more accurate compared with the full data, PCA, and SPCA, since it achieves 0.924 for PCA, 0.912 for SVD, 0.923 for SPCA, and 0.812 for the proposed approach.

In case of $k$-NN classifier, the first observation is that only our approach performs the precision measure better than all the other methods before and after dimensionality reduction, acquiring with the Pima Diabetes database 0.7 for PCA, 0.738 for SVD, 0.694 for SPCA and 0.783 for the proposed approach, and with the Waveform database 0.879 for PCA, 0.873 for SVD, 0.888 for SPCA and 0.899 for the proposed approach. According to the four dimensionality reductions, the recall measure is reduced after reduction, only in case of PCA and SPCA with the Waveform database, and similar to the full data in the case of SVD with the Pima Diabetes database. The recall measure acquires with the Pima Diabetes database 0.7 for PCA, 0.738 for SVD, 0.617 for SPCA and 0.783 for the proposed approach, and with the Waveform database 0.897 for PCA, 0.873 for SVD, 0.888 for SPCA and 0.899 for the proposed approach. The overall performance of ANN can be considered better in comparison with ANN in terms of precision but less accurate according to recall measures.

For the naïve Bayesian classifier, it is observed that the majority of the precision results obtained for the Pima Diabetes database is better after dimensionality reduction for all methods, it achieves 0.796 for PCA, 0.768 for SVD, 0.782 for SPCA and 0.798 for the proposed approach, but it is the opposite case with the Waveform database by acquiring the values 0.889 for PCA, 0.896 for SVD, 0.896 for SPCA and 0.923 for the proposed approach. The recall measure can be considered better after all dimensionality reduction methods for the Waveform database, however in the case of the Pima Diabetes database, it is increased for PCA and SPCA, similar to the full data with SVD, but reduced with our approach with the Pima Diabetes database, i.e., they achieve, respectively, 0.883, 0.8 for PCA, 0.875, 0.77 for SVD, 0.882, 0.787 for SPCA and 0.879, 0.731 for the proposed approach.

By comparing with PCA, SVD, and SPCA, we can see in all cases of our results, before and after the final reduction and using the ANN, $k$-NN, naïve Bayesian classifiers, that our approach yields the highest precision and lowest recall.

*6.5.2. Discussion of the experimental results*

In this part, the results of the dimensionality reduction will be summarized on the basis of two steps: the dimensionality reduction and the efficiency of the methods used.

1. The proposed dimensionality reduction approach is compared with SVD, PCA, and SPCA, using the real-world Pima Diabetes, Waveform, two versions of Human Activity Recognition using Smartphone, and Thyroid datasets. The results of all simulation studies have shown that the proposed method is much better than SVD, PCA and also SPCA for all datasets, the difference of rate improvement is most sensitive for all values, and also it is the most appropriate for the dimensionality reduction in large datasets by finding an

Table 3: Classification accuracy obtained before and after dimensionality reduction.

| | | | | | Dimensionality reduction | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Classification methods | Full Data | | PCA | | SVD | | SPCA | | Our approach | |
| | | | P | R | P | R | P | R | P | R | P | R |
| Database | **Pima Diabetes** | ANN | 0.748 | 0.722 | 0.763 | 0.757 | 0.748 | 0.722 | 0.782 | 0.774 | 0.783 | 0.722 |
| | | kNN | 0.738 | 0.735 | 0.7 | 0.683 | 0.738 | 0.735 | 0.694 | 0.687 | 0.783 | 0.617 |
| | | Naïve Bayes | 0.767 | 0.77 | 0.796 | 0.8 | 0.768 | 0.77 | 0.782 | 0.787 | 0.798 | 0.731 |
| | **Waveform** | ANN | 0.925 | 0.924 | 0.929 | 0.928 | 0.915 | 0.912 | 0.923 | 0.923 | 0.929 | 0.812 |
| | | kNN | 0.879 | 0.879 | 0.897 | 0.897 | 0.873 | 0.873 | 0.888 | 0.888 | 0.899 | 0.812 |
| | | Naïve Bayes | 0.903 | 0.882 | 0.889 | 0.883 | 0.896 | 0.875 | 0.888 | 0.882 | 0.923 | 0.879 |

optimal low-dimensional space, removing the redundant data and the variables which are linear combinations of other;

2. Our paper applies statistical and classification measure methods to improve the efficiency of the proposed method. In the first stage, statistical measures are used to define the performance of the dimensionality reduction method, expressed in terms of the standard deviation of a set of results. In this stage, our approach seems to be the best for practical use because it has the smaller bias and the standard deviations are more stable for all databases. In the second stage, classification measures are used to improve the effectiveness of dimensionality reduction using on one side the full set of dimensions and on the other the reduced set of provided data in terms of precision and recall, for the four dimensionality reductions used. A comparative study shows the performance of the ANN, $k$-NN, naïve Bayesian classifiers when compared the Pima Diabetes and Waveform databases without and with dimensionality reduction. It is observed that in a majority of the cases (before or after reduction), our approach outperforms significantly the performance of the other dimensionality reductions, and yields the highest precision and the lowest recall with all classifiers.

## 7. Conclusion and Future work

In this paper, we have proposed a new method for dimensionality reduction in the data pre-processing phase of mining high-dimensional data. This approach is based on the theory of Copulas (sampling techniques) to estimate the multivariate joint probability distribution without constraints of specific types of marginal distributions of random variables that represent the dimensions of our datasets. A Copula based model provides a complete and scale-free description of dependency that is thereafter used to detect the redundant values. A more extensive evaluation is made by eliminating dimensions that are linear combinations of others after having decomposed the data, and using the LU-decomposition method. We have reformulated the problem of data reduction as a constrained optimization problem. We have compared the proposed approach with well-known data mining methods using five real-world datasets taken from the machine learning repository in terms of the dimensionality reduction and the efficiency of the methods. The efficiency of the proposed method was improved by using the both statistical and classification methods. The different results obtained show the effectiveness of our approach which outperforms significantly the performance of dimensionality reduction comparing to other methods, i.e., it provided a smaller bias with more better standard deviations, a highest precision, and a lowest recall with all classifiers for all databases. Further work can be carried out in several directions. Researchers have made great efforts to improve the performance of the dimensionality reduction approach in very large datasets. However, the most serious problem is the presence of missing values in datasets. Missing values can result in loss of efficiency of the dimen-

sionality reduction approach, lead to complications in handling and analyzing the data, or distort the relationship between the data distribution. Also, it would be interesting to investigate the possibility of using metaheuristics or hybrid approaches to determine a solution of the proposed optimization problem in the Big Data setting.

## References


Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, *22*, 717–727.

Bai, D., Liming, W., Chan, W., Wu, Q., Huang, D., & Fu, S. (2015). Sparse principal component analysis for feature selection of multiple physiological signals from flight task. In *Control, Automation and Systems (ICCAS), 2015 15th International Conference on* (pp. 627–631). IEEE.

Boutsidis, C., Zouzias, A., Mahoney, M. W., & Drineas, P. (2015). Randomized dimensionality reduction for-means clustering. *Information Theory, IEEE Transactions on*, *61*, 1045–1062.

Colomé, A., Neumann, G., Peters, J., & Torras, C. (2014). Dimensionality reduction for probabilistic movement primitives. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on* (pp. 794–800). IEEE.

Dash, R., Misra, B., Dehuri, S., & Cho, S.-B. (2015). Efficient microarray data classification with three-stage dimensionality reduction. In *Intelligent Computing, Communication and Devices* (pp. 805–812). Springer.

Deegalla, S., Boström, H., & Walgama, K. (2012). Choice of dimensionality reduction methods for feature and classifier fusion with nearest neighbor classifiers. In *Information Fusion (FUSION), 2012 15th International Conference on* (pp. 875–881). IEEE.

Derrac, J., Chiclana, F., Garcia, S., & Herrera, F. (2016). Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets. *Information Sciences*, *329*, 144–163.

Fakoor, R., & Huber, M. (2012). A sampling-based approach to reduce the complexity of continuous state space POMDPs by decomposition into coupled perceptual and decision processes. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on* (pp. 687–692). IEEE volume 1.

Gisbrecht, A., & Hammer, B. (2015). Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*, 51–73.

Han, J., & Kamber, M. (San Francisco, 2006). *Data Mining: Concepts and Techniques*. 13: 978-1-55860-901-3 (2nd ed.). Diane Cerra.

Hettiarachchi, R., & Peters, J. (2015). Multi-manifold LLE learning in pattern recognition. *Pattern Recognition*, *48*, 2947–2960.

Houari, R., Bounceur, A., & Kechadi, M.-T. (2013a). A new method for dimensionality reduction of multi-dimensional data using copulas. In *Programming and Systems (ISPS), 2013 11th International Symposium on* (pp. 40–46). IEEE.

Houari, R., Bounceur, A., & Kechadi, T. (2013b). A New Approach for Dimensionality Reduction of Large Multi-Dimensional Data Based on Sampling Methods for Data Mining, .

Houari, R., Bounceur, A., Kechadi, T., Tari, A., & Euler, R. (2013c). A new method for estimation of missing data based on sampling methods for data mining. In *Advances in Computational Science, Engineering and Information Technology* (pp. 89–100). Springer.

Kerdprasop, N., Chanklan, R., Hirunyawanakul, A., & Kerdprasop, K. (2014). An empirical study of dimensionality reduction methods for biometric recognition. In *Security Technology (SecTech), 2014 7th International Conference on* (pp. 26–29). IEEE.

Kuang, F., Zhang, S., Jin, Z., & Xu, W. (2015). A novel svm by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection. *Soft Computing*, (pp. 1–13).

Lichman, M. (2013). Uci machine learning repository, university of california, irvine, school of information and computer sciences, http://archive.ics.uci.edu/ml.

Lin, P., Zhang, J., & An, R. (2014). Data dimensionality reduction approach to improve feature selection performance using sparsified svd. In *Neural Networks (IJCNN), 2014 International Joint Conference on* (pp. 1393–1400). IEEE.

Nelsen, R. B. (2007). *An introduction to copulas*. (2nd ed.). Springer Science & Business Media.

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

Rubinstein, R. Y., & Kroese, D. P. (2011). *Simulation and the Monte Carlo method* volume 707. John Wiley & Sons.

Saoudi, M., Bounceur, A., Euler, R., & Kechadi, T. (2016). Data mining techniques applied to wireless sensor networks for early forest fire detection. In *Proceedings of the International Conference on Internet of things and Cloud Computing* (p. 71). ACM.

Sasikala, S., & Balamurugan, S. A. A. (2013). Data classification using pca based on effective variance coverage (evc). In *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on* (pp. 727–732). IEEE.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, *99*, 1015–1034.

Sun, Y., Gao, J., Hong, X., Guo, Y., & Harris, C. J. (2014). Dimensionality reduction assisted tensor clustering. In *Neural Networks (IJCNN), 2014 International Joint Conference on* (pp. 1565–1572). IEEE.

Venugopalan, S., Savvides, M., Griofa, M. O., & Cohen, K. (2014). Analysis of low-dimensional radio-frequency impedance-based cardio-synchronous waveforms for biometric authentication. *Biomedical Engineering, IEEE Transactions on*, *61*, 2324–2335.

Wan, X., Wang, D., Peter, W. T., Xu, G., & Zhang, Q. (2016). A critical study of different dimensionality reduction methods for gear crack degradation assessment under different operating conditions. *Measurement*, *78*, 138–150.

Watcharapinchai, N., Aramvith, S., Siddhichai, S., & Marukatat, S. (2009). Dimensionality reduction of sift using pca for object categorization. In *Intelligent Signal*

*Processing and Communications Systems, 2008. ISPACS 2008. International Symposium on* (pp. 1–4). IEEE.

Wencheng, P. (2010). Theory and application of parameter self-optimizing intelligent sampling method. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on* (pp. 66–69). IEEE.

Whelan, M., Khac, N. A. L., Kechadi, M. et al. (2010). Data reduction in very large spatio-temporal datasets. In *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on* (pp. 104–109). IEEE.

Zhai, M., Shi, F., Duncan, D., & Jacobs, N. (2014). Covariance-based pca for multi-size data. In *Pattern Recognition (ICPR), 2014 22nd International Conference on* (pp. 1603–1608). IEEE.

Zhang, J., Nie, X., Hu, Y., Liu, S., Tian, Y., & Wu, L. (2010). A method for land surveying sampling optimization strategy. In *Geoinformatics, 2010 18th International Conference on* (pp. 1–5). IEEE.

Zhang, T., Du, Y., Huang, T., & Li, X. (2016). Stochastic simulation of geological data using isometric mapping and multiple-point geostatistics with data incorporation. *Journal of Applied Geophysics*, *125*, 14–25.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, *15*, 265–286.