Acta Scientific COMPUTER SCIENCES

Volume 5 Issue 1 January 2023

Review Article

Review of High-Dimensional Data Reduction Methods

Mahmoud Rokaya*

Department of Information Technology, Taif University, Saudi Arabia

***Corresponding Author:** Mahmoud Rokaya, Department of Information Technology, Taif University, Saudi Arabia.

Received: November 14, 2022

Published: December 23, 2022

Abstract

In the current decade, most of the computational problems came to be problems with high dimensional data. Correct data reduction will relax a load of computation to an acceptable range in time and space. Most of the available data reduction methods are built on the statistical background. Few of them adopted machine learning. Few works considered ensemble learning as a method to merge different methods to get a superior method to all merged individual reduction methods. This work will present the history of high-dimensional data reduction methods. It will analyze the recent developments in methods of reduction data schemes, especially ensemble methods.

Keywords: Dimensionality Reduction; Random Projection; PCA; Curvilinear Component Analysis (CCA); Projected Support Points (PSPs); Sequential Ensemble (SEMSE); Projection Pursuit; Particle Swarm Optimization (PSO); Genetic Algorithm (GA); Quadratic Discriminant Analysis (QDA)

Introduction

Transforming data with a higher number of dimensions into data with a significantly lower number of dimensions and keeping the most meaningful characteristics of the original data is called dimensionality reduction. Dimensionality reduction is required for many reasons. Among these reasons, sparse data causes the data to be computationally intractable. Dimensionality reduction requirements appear whenever a large volume of observations is faced. This case is frequently faced with signal processing, social media data, natural language processing, voice recognition, bioinformatics, neuroinformatics, and many other cases. Processing requirements, improving the storage and transfer time become a required characteristic of the huge amount of data that could be collected due to increasing the number of users and the advances in technologies of collecting and storing the data. So, the reduction of the dimensionality of the data transferred from being the desired action into a mandatory requirement in almost all applications. However, using the reduced data instead of the original one directly is still a challenge [1]. Figure 1 illustrates the idea behind the dimensionality reduction. The set represents the original data, represented the output of dimensionality reduction of , the function can be seen as a function with high complexity in time and space. The cost of applying the function on the reduced data is very small compared to the cost of applying the same function on the original input data . Certainly, function is complex, unless there will be no benefit from the reduction process. The cost of the reduction process and the cost of applying the function on must be lower than applying the function on . Also, another requirement that should be fulfilled by the reduction of the process is the equivalence of the output of applying on and applying on the reduced data .

Through the last few decades, many dimensionality methods are proposed. For example, Hierarchical subspace sampling, random projection, Principal Component Analysis (PCA), Curvilinear Component Analysis (CCA), Projected Support Points (PSPs), sequential ensemble-based framework (SEMSE), Projection Pursuit, double-phases particle swarm optimization (PSO), Genetic Algorithm (GA) for dimensionality reduction, Ant Colony Optimization (ACO) and quadratic discriminant analysis (QDA). Most of these methods will be explored through this review. Some of these methods like PCA depend on a mathematical model to reduce each correlated feature into its best representative feature.

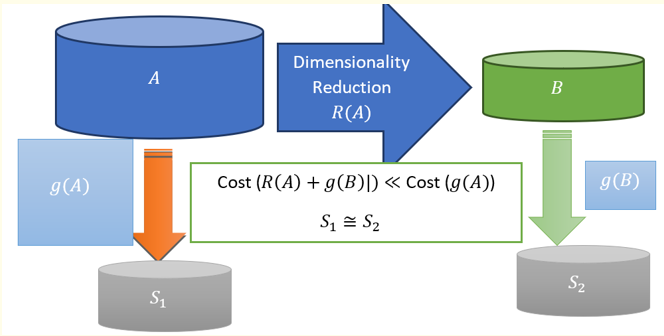

Figure 1: Dimensional Data Reduction concept.

PCA performance will be affected if there are a high number of noisy features. Other methods depend on distinguishing between trusted and outlier features through ensemble learning to add as much as of trusted features and reduce as much as possible of outlier features. Other methods adopted the behavior of insects like ants to gain the most suitable path to reach a target. The target here is the lowest number of features, for example, ACO and PSO methods. Finally, as an optimization problem, GA and QDA presented a solution to the dimensionality reduction problem. Applications of dimensionality reduction are almost everywhere.

Clustering and classification of data are the main domains for dimensionality reduction. DNA gene expression profiles and clustering subjects based on brain dysfunctions are examples of clustering problems that need dimensionality reduction. Classifying multimedia data and tweets or blogs on social media is another example of fields that needs dimensionality reduction of data. The review is divided into two parts. The first part covers the main methods for dimensionality reduction and the other part covers some applications of dimensionality reduction.

Main dimensionality reduction methods

Principal component analysis (PCA)

In 1901, Karl Pearson presented PCA for the first time. Later, the method was independently presented and given the name Harold Hotelling. Also, in the field of signal processing, PCA was given the name discrete Karhunen–Loève transform (KLT). Depending on the field, PCA was given many names such as proper orthogonal decomposition (POD), singular value decomposition (SVD), and eigenvalue decomposition (EVD). PCA details can be found in many references [2]. There are a number of PCA variations based on the data nature and the application such as Sparse PCA [3], Nonlinear PCA [4], and Robust PCA [5]. Also, there is several similar algorithms for PCA such as independent component analysis [6] and Network component analysis [7].

Hierarchical subspace sampling

To create a reduced representation of the data Hierarchical Subspace Sampling was proposed by Aggarwal in 2002**.** The method details can be found in [1]. Based on the hierarchical approach and subspace sampling procedure the reduction process can be more effective. Subspace sampling is used for providing a compact representation and hard bounds of the expected error of the approximation. The compression factor becomes better when the size of the data increased. For sure, this property gives the approach a competing feature in contrast to other approaches. The locality-specific multi-dimensional representation enables the use of the reduced data in other applications such as selectivity estimation and nearest neighbor search. Space subsampling finds the significant local properties of the data and this in turn can help in finding an effective solution of the mentioned problems. For selectivity estimation methods, traditional methods such as histograms achieve low accuracy even with a limited number of dimensions. From the other hand, Subspace Sampling provides more accurate results on color-histogram data sets with a dimensionality of more than 50.

Random projection

Random projection, in its core idea, depends on the Johnson-Lindenstrauss lemma [8]. Given dimensional original data we need

to reduce it dimensionality to k- dimensional data where (k≪d). dimensions are called the random projection of the d- dimension the original data set $X_{d\times N}$ and its random projection data set $X_{K\times N}$ satisfy the following equation:

$$X^{RP}_{k\times N} = R_{k\times d} X_{d\times N} \quad \text{---------(1)}$$

Where each vector in the matrix $R_{k\times d}$ is normalized in terms of each column in R is a unit vector. The problem of finding the matrix R for k and d dimension is of order O(dkN). Moreover, if the matrix X has many zeros (sparse) and the max number of non-zero elements in each column of is the complexity is reduced to be O(ckN) [9]. Note that, in general, R is not orthogonal, however, for high dimensions R^TR would be an approximation of identity matrix. This note is enough for avoiding calculating as orthogonal matrix which is a have a high complexity [8]. In practice, R is randomly chosen based on Gaussian distribution which can also relaxed using similar distribution such as

$$r_{ij} = \sqrt{3}\begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases} \quad \text{---------- (2)}$$

For each element $r_{ij} \in R$. Equation (3) slightly reduces the computations cost when compared to PCA, SVD and LSI. Johnson, William used random projection in clustering application for image and text data [9].

Curvilinear component analysis (CCA)

In Curvilinear Component Analysis (CCA), a data set X (Input set) with high dimension is mapped to a lower dimension data set Y (output space or latent space) through training a self-organizing neural network. For each neuron two weights are attached. One for X dimensionality and another one Y dimensionality. is the projection of X in a lower dimension space. So, each neuron can be seen as a transfer agent between the original vector X and its projection Y in the latent space.

The CCA algorithm can be summarized in the following steps:

Input vector quantization: many ways can be used for input quantization. One common method is to use unsupervised classical neural network.

The CCA projection: Calculate the distance between each two points x_i, x_j in the input space X.

$$D_{ij} = \|x_i - x_j\| \quad \text{---------- (3)}$$

In the latent space (output space) the distance L_{ij} between the projection y_i, y_j of the input x_i, x_j must be constrained to be equal to

$$L_{ij} = \|y_i - y_j\| = D_{ij} \quad \text{----------(4)}$$

To this end, right Bregman divergence can be used to calculate the error to penalize long distance on the output side of each pair of neurons. right Bregman divergence (RBD_{ij}) is given by:

$$RBD_{ij} = \lambda^2 y_j \left[e^{-\frac{D_{ij}}{\lambda}} - e^{-\frac{L_{ij}}{\lambda}} + (D_{ij} - L_{ij}) \frac{e^{\frac{L_{ij}}{\lambda}}}{\lambda} \right] \quad \text{----------(5)}$$

Where λ represents a threshold to give short and long between-point distances . Equation (5) guide the neurons to reduce the distance in the latent space to converge to the distance in the input space.

To minimize RBD_{ij} the stochastic gradient algorithm can be used which is given by:

$$y_j = y_j - \alpha \frac{L_{ij} - D_{ij}}{L_{ij}} e^{-\frac{L_{ij}}{\lambda}} \quad \text{---------- (6)}$$

Where is the learning rate.

CCA has some limitations. For example, with each new execution, the same inputs in will have different equivalent points in the latent output space . In other words, CCA is variant since the only constrain is the distance between points not the points itself. CCA has a variation called on CCA. Details of on CCA can be found at [10].

Sequential ensemble-based framework SEMSE

An example of dimensionality reduction based on outliers scores is SEMSE. Based on ensemble approach SEMSE develops a set of individual agents. Each agent calculates the outlier score for each dimension. The system aggregates the results of the individual agents to calculate the final outlier score for each dimension. The dimensionality reduction is achieved through determining the outliers scores. the SEMSE performs the following steps.

- In iteration , suppose that $X=\{x_1, x_2, x_3, \ldots, x_N\}$ is a set of data items. Each item in X is described by a set of D features. For each $x_k \in X$,v $x_k=\{x_{k1}, x_{k2}, x_{k3}, \ldots, x_{kD}\}$. The outlier score is given by $y^{i-1} \in R^N$. y^{i-1} is calculated in iteration i-1.

- Let η^i defines an outlier thresholding function to produce a set of outliers L^i of $R^i \in R^{Li\times(D+1)}$
- y^{i-1} as the output corresponds to the features. Based on a sparse regression model ψ^i on R^i, a new set of data S^i and a set of optimal features M^i are generated.
- An outlier score function ϕ^i on S^i is used to compute the new outlier score vector y^i
- SEMSE iterates these steps to compose a sequential ensemble model, as illustrated in figure 2.
- To find the final outlier scores, bagging is applied to aggregate the output set of sequential ensemble models.

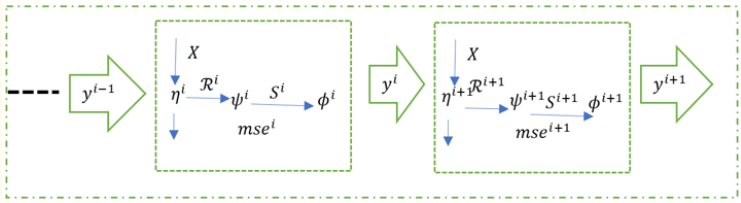

Figure 2: SEMSE steps to compose the individual agents.

According to the Aggarwal 2017, learning bias variance can be reduced by the help of bagging stage in SEMSE. SEMSE has the merit of generalization. A specific individual set of agents is generated as a result of using three specific instances of the three components. Namely, and . For example see [11] for A SEMSE Instance: CINFO.

Projection pursuit

projection pursuit is proposed to support dimensionality reduction when intuitive properties of high dimensional data spaces are faced. For example, the phenomenon of distance concentration. Fisher›s discriminant ratio, Fisher›s linear discriminant analysis, and a variant of projection pursuit are examples of the dimensionality reduction methods that fight against distance concentration. The distance concentration for a given set of data is said to be exist if the difference between the largest distance between points and the smallest distance is bounded by a given small number and this is equivellent to say that the relative variance of a given set of data converge to zero when the number of points tends to . The idea behind projection pursuit can be summarized as follows.

Assume that $x_1^{(m)},x_2^{(m)},x_3^{(m)},\ldots,x_N^{(m)}$ is a random sample from a finite sequence of data distribution F_m,m=1,2,3,.... The relative variance is given by $Var(\|x^{(m)}\|^s)/E(\|x^{(m)}\|^s)^2$, where is the variance and E is the expectation of the distribution F_m.||.|| is the Euclidean distance and will be assumed to equal 2.

Solve the problem

$\underline{W}x = way+w\delta$ --------- (7)

x

Through the last few decades, many dimensionality methods are proposed. For example, Hierarchical subspace sampling, random projection, Principal Component Analysis (PCA), Curvilinear Component Analysis (CCA), Projected Support Points (PSPs), sequential ensemble-based framework (SEMSE), Projection Pursuit, double-phases particle swarm optimization (PSO), Genetic Algorithm (GA) for dimensionality reduction, Ant Colony Optimization (ACO) and quadratic discriminant analysis (QDA). Most of these methods will be explored through this review. Some of these methods like PCA depend on a mathematical model to reduce each correlated feature into its best representative feature.

Particle swarm optimization (PSO) for data dimensionality reduction

The idea in PSO is to simulate the behaviour of birds during hunting. Each bird is called a particle and a collection of the birds is called swarm. Each particle can be defined through three features, its velocity, position and path. The movement of each particle depends on it previous optimal position and the current global optimal path of the swarm. The steps of PSO can be summarized in figure 3.

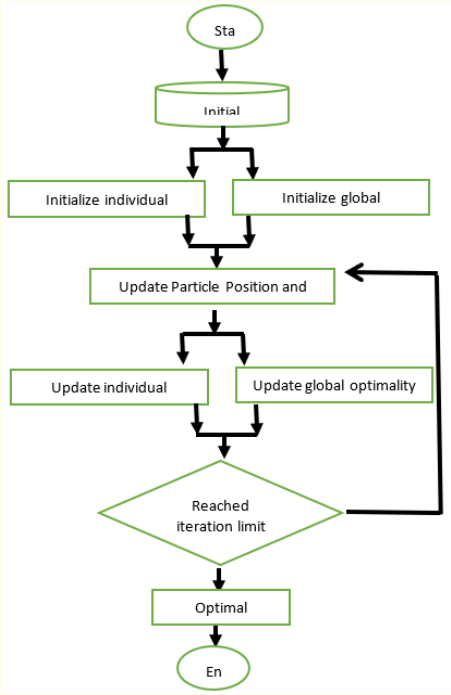

Figure 3: Simplified presentation of Particle Swarm Algorithm.

The computations converge by the tendency of each particle to move near around one optimal point.

Dimensionality reduction can be seen as a particle swarm optimization problem if the problem is stated as find the best features among a set of given features to reduce the feature data. An example of applying PSO can be found in [13].

Genetic algorithms (GAs) for dimensionality reduction

Genetic algorithm for dimensionality reduction based on considering that there is a collection of subsets of features and there is a method to measure information gain provided by each feature and there is a fitness function that will lead to the best sub feature set. An example of using genetic algorithm in dimensionality reduction is proposed by [14]. Based on the nature of the data and the problem type, The goal behind the dimensionality reduction become clear and a specific fitness function can be built. The crossover and mutation operations are used to produce the new generation and making a leap while proceeding towards the optimal solution.

Discussion

Dimensionality reductions arises because of the nature of observations in the modern applications, huge numbers of observations on a small experimental unit. Many difficulties come with high dimensional data. It is not only the high cost of computation but also the inability to perform the calculations at all when we face a sparse subspace [15]. Also, the suitability of the current analysis tools for high dimensional-data reduction are questioned when facing sparse subspace or the problem of distance concentration an additive unstructured noise in the latent space. projection pursuit for example were developed to fight the problem of data concentration and its variation to work in high sparse subspace. Detailed analysis of some traditional method shows that, even traditional methods tend to fight the concentration data issue despite they are not developed considering this issue [12].

Quality measure is an important tool to define the meaningful patterns in high-dimensional data. Many quality metrics were proposed to help the user to choose the most promising dimensions of his data for visualizing his data. However, the relation between these methods is not explored [16]. Explored these approaches to analyize the quality metrics through finding the factors that discriminate these quality measures from each other.

Dimensionality reduction methods not only differs in the accuracy and the resulted latent space but also differs in the nature of data that they can process, the issues in the data that they highlight and the application that they can applied for. [1] proposed a method of dimesionality reduction that fight the degradation in the usability of the latent space data. In order to create a reduced data, the hierarchical subspace sampling is used to find the local implicit dimensionalities for each element in the original data space. The second step is to create a reduced re4spesentation of the data.

Panthong and Srivihok highlighted the issue of using ensemble learning to solve the problem of features selection. Based on ensemble methods such as Bagging and AdaBoost, the algorithm combines the results using sequential backward selection, sequential forward selection and optimize selection. The individual learners are presented by Naïve Bayes and decisions trees. Online dimensionality reduction feature is available for a limited number of dimensionality reduction algorithms such as linear projections

and PCA [17]. To enable online dimensionality reduction. Cirrincione et, al. developed the online Curvilinear Component Analysis (online CCA). online CCA inherits the features of its predecessor CCA. Online CCA is adaptive in real time. Online CCA can track high dimensional distributions with non-stationary. For modelling very complex high dimensional data online CCA can be considered as a basic technique for complex supervised neural networks [10].

For the sparse space issue in dimensionality reduction, Mak and Joseph proposed the projected support points (PSPs) algorithm. For this end, after establishing a framework, based on PSPs, they proposed two algorithms for one-shot and sequential reduction for big data majorization-minimization and subsampling for efficient optimization [18].

One of the domains that needs a real time dimensionality reduction is Surface electromyography (sEMG) since it is very important for HCI. For sure, with the huge number of features and data volume the processing will be slow. To solve this problem based on dimensionality reduction, Jie J, Liu K, Zheng, and Dai proposed a double-phases particle swarm optimization (PSO) method for dimensionality reduction. The method implements two methods. WLMRKNN and WRKNN are based on KNN for the classification task and the classification accuracy is used to evaluate the particles of PSO [19].

Cleaning data from noisy and outlier detection are important issues in dimensionality reduction, where the aim is to get a set of data representing the original data in the latent space with few or without noisy or outlier dimensions. Dimensionality reduction methods usually focus on one of these problems but not both in the same time. Pang., *et al*. proposed SEMSE and its instance CINFO to develop a method that can provide a dimensionality reduction considering reducing noisy and outlier dimensions. SEMSE stands for sequential ensemble. SEMSE sequential feature came from it successive nature in applying three learners one after another. each individual learner is a recurrent network. For detecting outliers, SMSE applies Cantelli's inequality-based outlier thresholding function. For detecting noisy dimension, SMSE applies lasso-based sparse regression by treating the original features as predictors and the outlier scores as the target feature [11].

Dimensionality reduction might depend on the domain of the problem. For example, in software engineering, dimensionality reduction was used to be the base of solving low quality of bug reports and the absences of engagement of developer's problems. Ge., *et al*. proposed high-dimensional hybrid data reduction method to help in solving these problems. The term hybrid came from combining feature selection and instance selection. Based on building a core data set that consists of bug reports which is clean from noninformative bug or redundant reports or words. Also, developers are engaged to effectively define similar bug reports [19].

Another example of developing a designed dimensionality reduction method came from the domain of multimedia. Permanasasi., *et al*. addressed the problem of dimensionality reduction. Multimedia is related to many fields including image processing, social media analysis, multimedia retrieval, data mining, database management, and so on. Multimedia has the feature of exponential growth in size since the higher the quantity of data, the higher the complexity, diversity, and dimensions. Projection Pursuit is proposed to overcome the failure of PCA when countering data with a large volume of noise. In Projection Pursuit, only important data are transferred. The data importer is decided by generating a projection index. An optimization process is generated to maximize the generated projection index [20].

Dimensionality reduction was implemented in many applications to improve the performance of the applications in terms of space, time, and accuracy. Random projections were tested in several cases by [9]. These cases include information retrieval in text documents and noisy and noiseless images as well as using a sparse random matrix. Random projections had been proven to compete with other dimensionality methods such as PCA. Also, the cost of computations was fewer in the case of Random projections than in PCA.

Using dimensionality reduction is common in classification and clustering applications. To avoid trapping in local minimum while using EM or K-means, Ding., *et al*. used repeated dimension reductions. Clustering data with high overlap such as Internet newsgroups and DNA gene expression profiles proved the effectiveness of dimensionality reduction [21].

Another important domain that depends heavily on dimensionality reduction is fMRI data analysis. Durieux and Wilderjans used dimensionality reduction to improve the clustering process. They

tried to develop a method for a brain-based categorization system based on FC patterns. ICA was used, in the first phase, for dimensionality reduction for each patient's fMRI data then the clustering algorithm, in the second phase, is applied to the reduced data [22].

Conclusion

This short review highlighted the importance of dimensionality reduction in almost all applications. Many methods were proposed for dimensionality reduction even before the PC era. In the last few years, many methods were proposed such as Hierarchical subspace sampling, random projection, Principal Component Analysis (PCA), Curvilinear Component Analysis (CCA), Projected Support Points (PSPs), sequential ensemble-based framework (SEMSE), Projection Pursuit, double-phases particle swarm optimization (PSO), Genetic Algorithm (GA) for dimensionality reduction, Ant Colony Optimization (ACO) and quadratic discriminant analysis (QDA). Most of these methods will be explored through this review. These methods were proposed for many reasons. The following reasons were explored, inability to perform the calculations at all when we face a sparse subspace, nature of data, needing a real-time dimensionality reduction, cleaning data from noisy and outlier detection, and domain of the problem and their corresponding solutions. Some works are based on ensemble learning as a tool for dimensionality reduction. Ensemble learning will be suitable for dimensionality reduction if it can exceed the barrier of time and space that the dimensionality reduction was proposed to relax.

Bibliography

1. Aggarwal CC. "Hierarchical subspace sampling: a unified framework for high dimensional data reduction, selectivity estimation and nearest neighbor search". In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (2002): 452-463.

2. Jolliffe Ian T and Cadima Jorge. "Principal component analysis: a review and recent developments". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374.2065 (2016): 20150202.

3. Hui Zou., *et al*. "Sparse principal component analysis". *Journal of Computational and Graphical Statistics* 15.2 (2006): 262-286.

4. Hastie T and Stuetzle W. "Principal Curves". *Journal of the American Statistical Association* 84.406 (1989): 502-506.

5. T Bouwmans and E Zahzah. "Robust PCA via Principal Component Pursuit: A Review for a Comparative Evaluation in Video Surveillance". *Computer Vision and Image Understanding* 122 (2014): 22-34.

6. Hyvärinen Aapo. "Independent component analysis: recent advances". *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 371 (1984): 20110534.

7. Liao JC., *et al*. "Network component analysis: Reconstruction of regulatory signals in biological systems". *Proceedings of the National Academy of Sciences* 100.26 (2003): 15522-15527.

8. Johnson William B and Lindenstrauss Joram. "Extensions of Lipschitz mappings into a Hilbert space". Conference in Modern Analysis and Probability (New Haven, Conn., 1982). Contemporary Mathematics. Providence, RI: American Mathematical Society 26 (1984): 189-206.

9. Bingham E and Mannila H. "Random projection in dimensionality reduction: applications to image and text data". In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (2001): 245-250.

10. Cirrincione G., *et al*. "The online curvilinear component analysis (onCCA) for real-time data reduction". In 2015 International Joint Conference on Neural Networks (IJCNN) (2015): 1-8. IEEE.

11. Pang G., *et al*. "Sparse modeling-based sequential ensemble learning for effective outlier detection in high-dimensional numeric data". In Proceedings of the AAAI Conference on Artificial Intelligence 32.1 (2018).

12. Kabán A. "On the distance concentration awareness of certain data reduction techniques". *Pattern Recognition* 44.2 (2011): 265-277.

13. Jie J., *et al*. "High dimensional feature data reduction of multichannel sEMG for gesture recognition based on double phases PSO". *Complex and Intelligent Systems* 7.4 (2021): 1877-1893.

14. Faraoun KM and Rabhi A. "Data dimensionality reduction based on genetic selection of feature subsets". *INFOCOMP Journal of Computer Science* 6.3 (2007): 36-46.

15. Johnstone IM and Titterington DM. "Statistical challenges of high-dimensional data". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906 (2009): 4237-4253.

16. Bertini E., *et al*. "Quality metrics in high-dimensional data visualization: An overview and systematization". *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011): 2203-2212.

17. Panthong R and Srivihok A. "Wrapper feature subset selection for dimension reduction based on an ensemble learning algorithm". *Procedia Computer Science* 72 (2015): 162-169.

18. Mak S and Joseph VR. "Projected support points: a new method for high-dimensional data reduction". arXiv preprint arXiv:1708.06897 (2017).

19. Ge X., *et al*. "High-dimensional hybrid data reduction for effective bug triage". *Mathematical Problems in Engineering* (2020): 2020.

20. Permanasasi Y., *et al*. "PCA and projection pursuits on high dimensional data reduction". *In Journal of Physics: Conference Series* 1722.1 (2001): 012087. IOP Publishing.

21. Ding C., *et al*. "Adaptive dimension reduction for clustering high dimensional data". In 2002 IEEE International Conference on Data Mining, 2002. Proceedings (2002): 147-154). IEEE.

22. Durieux J and Wilderjans TF. "Partitioning subjects based on high-dimensional fMRI data: comparison of several clustering methods and studying the influence of ICA data reduction in big data". *Behaviormetrika* 46.2 (2019): 271-311.