Original Article

A Review Paper on Dimensionality Reduction Techniques

Faizan Riyaz Mulla[1], Dr. Anil Kumar Gupta[2]

[1]Department of Computer Science Engineering, AISSMS College of Engineering, Pune, India. E-mail: faizanmulla2000@gmail.com

[2]Centre for Development of Advanced Computing (C-DAC) Pune, India. E-mail: akgupta5592@gmail.com

Abstract

Dimensionality Reduction (DR) is the process of reducing the numerous features or random variables under consideration to a limited number of features by obtaining a set of principal variables. These techniques cater great values in machine learning, which come in handy to simplify a classification or a regression dataset, thereby yielding a better-performing predictive model. Techniques used for DR include Feature Selection methods, Matrix Factorization, AutoEncoder methods, and Manifold Learning. Merits of DR include data compression, reduced space of storage, and removal of redundant features. This paper attempts to review various techniques used to carry out dimensionality reduction while providing an exhaustive comparative study over the merits and demerits of each of the techniques used under the empirical experiments performed by the authors whose work is being reviewed.

Keywords: Dimensionality Reduction (DR), PCA, DP-PCA, ICA, SVD, LDA, Feature Selection, Feature Extraction, Autoencoders, Isomap, Umap, t-SNE, k-PCA, and Factor Analysis.

DOI: 10.47750/pnr.2022.13.S03.198

INTRODUCTION

Concerning data science, DR has a long history dealing with how it was developed over a while in combination with the various techniques involving data interpretation and data visualization—catering to primary purposes of removal of noise in data and compressing and transforming complex dimensional space data into simpler dimensional space while maintaining the meaningful properties of prototypical data.

The DR techniques reviewed include:

1. Principal Component Analysis (PCA)
2. Distributed Parallel-PCA (DP-PCA)
3. Linear Discriminant Analysis (LDA)
4. Independent Component Analysis (ICA)
5. Singular Value Decomposition (SVD)
6. Feature Selection (FS)
7. Feature Extraction (FE)
8. Autoencoders (AE)
9. Isometric Mapping (ISOMAP)
10. Kernal PCA (k-PCA)
11. (11)Uniform Manifold Approximation and Projection (UMAP)
12. T-SNE
13. Factor Analysis (FA)

While dealing with data of high dimensions, the unprocessed raw data is often scattered and thinly distributed, giving rise to “The Curse of high Dimensionality.”

The DR techniques resolve this problem of high dimensionality and make the dataset computationally tractable. These techniques are primarily categorized under two types, namely linear and non-linear.

Some of the linear DR methods are PCA, Factor Analysis (FA), LDA, and Truncated SVD. While as other non-linear methods, also known as Manifold Learning, include Kernel PCA, t-SNE, MDS, and ISOMAP.

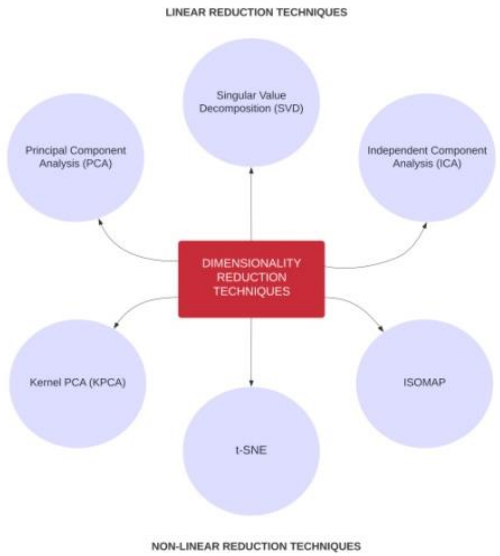

Fig. 1. Various linear and non-linear techniques are used in DR.

PRINCIPAL COMPONENT ANALYSIS (PCA)

A. BRIEF INTRODUCTION

PCA has been a popular and most commonly used linear DR technique. Karl Pearson earlier proposed it in the year 1901 [1]. While the method was introduced for non-random variables, it later extended to random variables [2].

B. WORKING PRINCIPLE

Its main working principle involves the decomposition of the eigenvalue of the covariance data matrix, $M\{XX^T\} = M\Lambda M^T$ [10]. Here the columns of matrix M are represented by eigenvectors, and the corresponding eigenvalues are present in Λ. Now the data is projected into subspace with top eigenvectors corresponding to the largest eigenvalue as follows:

$$X^{PCA} = M_k^T X \text{ [10]}$$

Where M_k contains top k eigenvectors.

The main idea is the mapping of features that are n-dimensional to k dimensions (k < n), which are new orthogonal features, namely principal components [3].

In this, we input the n-dimensional sample set $X = (x^1, x^2, .. x^m)$. After this, we centralize all the sample set $x^{(i)} = x^{(i)} - \frac{1}{m}\sum_{j-1}^{m} x^{(j)}$ [3]. Now using the matrix M which depicts the covariance and eigenvalue decomposition XX^T, we are able to find the eigen-vectors $(\omega_1, \omega_2, \omega_3, .., \omega_k)$ which correspond to the k-significant eigenvalues and form the eigenvector matrix W.

Finally, we convert the input data set $x^{(i)}$ into a unique dataset given by $z^{(i)} = W^T x^{(i)}$ [3] and receive the output sample set $X' = (z^1, z^2, .. z^m)$ [3].

C. MERITS AND DEMERITS OF THE TECHNIQUE

Emphasizing some of the merits of this technique is that this algorithm comes from the class of unsupervised machine learning, is quite effective in the removal of redundant features and noise as it does not consider the label category and maps that data which is collected towards the direction (bases) having significant variance [8]. Moreover, since the algorithm reduces the numerous features in the model, it overcomes the hurdle of data over-fitting. However, this also makes the model vulnerable to under-fitting if much of the variance in collected data gets suppressed.

Concerning the research conducted by C. Yumeng and F. Yinglan [9], a few of the shortcomings of this technique are that the reduction results are not good enough, are time-consuming, have high information loss, and yield mediocre prediction accuracy scores. They thus proposed an improved version of the PCA technique, which had its foundation relied on the entropy methods' weight and called it EW-PCA.

DISTRIBUTED PARALLEL PCA (DP-PCA)

A. BRIEF INTRODUCTION

In a traditional PCA method, data normalization is required to be done before constructing a PCA because PCA is reactive to the relative scalability of various variables [4]. For example, concerning PCA, one would need to compute each process attribute's mean and variance to normalize each process variable to zero means and unit variance. And as we are in a time where there is an ever increase in big data, it would not be feasible for us to perform computation over each attribute owing to its massive size, and hence there is an urge for parallel computing.

B. WORKING PRINCIPLE

Following the work provided by L. Wang [5], the paper uses a calculation of the "mean" method to achieve standardization of the unaltered dataset to be fed to the PCA. Then the paper improves the solving of the matrix of correlation coefficient for PCA, followed by designing a distributed parallel data dimensionality scheme for data reduction.

Here for an input data set represented by $X=(x_1, x_2, x_3, .., x_n)$, a mean dealing matrix M is created using:

$m_i^j = \frac{x_i^j}{\overline{x^j}}$, where $\overline{x^j}$ =($\sum_{k=1}^{n} x_k^j$) / n and i,j=1,2,..,d. [5]

$$M = \begin{bmatrix} m_1^1 & \cdots & m_1^d \\ \vdots & \ddots & \vdots \\ m_n^1 & \cdots & m_n^d \end{bmatrix} \text{ [5]}$$

Now constructing a square matrix G using,

$$g_i^j = \sum_{k=1}^{n} m_k^i m_k^j \text{ [5]}$$

$$G = \begin{bmatrix} g_1^1 & \cdots & g_1^d \\ \vdots & \ddots & \vdots \\ g_n^1 & \cdots & g_n^d \end{bmatrix} \text{ [5]}$$

Then calculating linear sum vector B and square sum vector F as:

$b^j = \sum_{k=1}^{n} x_k^j$, f = $\sum_{k=1}^{n} (x_k^j)^2$;

where B = [$b^1, b^2, .., b^d$] and F = [$f^1, f^2, .., f^d$] [6].

Using Pearson's method to solve the coefficient of correlation 'r.'

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}} \text{ [7]}$$

The equation gets deduced to:

$$r_i^j = \frac{n.g_i^j - b^i.b^j}{\sqrt{(n.f^i - (b^i)^2)}.\sqrt{(n.f^j - (b^j)^2)}}$$

Hence utilizing the above formula, we derive the correlation matrices for coefficient as R

$$R = \begin{bmatrix} r_1^1 & \cdots & r_1^d \\ \vdots & \ddots & \vdots \\ r_n^1 & \cdots & r_n^d \end{bmatrix} \text{ [5]}$$

Using R, we find the eigenvalues and eigenvectors and determine the first k-eigenvectors for constructing the matrix Q of projection. Finally, we map the original dataset on the matrix Q, which is required to get the dimensionally reduced dataset.

C. MERITS AND DEMERITS OF THE TECHNIQUE

According to Linlin [5], DP-PCA prevails over the fallibility of PCA as it has a much lesser computation time. DP-PCA also uses the "mean" method, which serves as the elimination of influence on dimensions and reflects the variation information of the unaltered data. DP-PCA is observed to refine the matrix of correlation coefficient and thus enhance the calculation speed. Linlin [5] advocates that DP-PCA achieves better DR and manifestly concludes that DP-PCA reduces processing time and is undoubtedly a better algorithm than PCA.

J. Zhu and Z. Ge [4], using the map-reduce platform, established in multiple blocks, successfully implemented DP-PCA to perform distributed and parallel computations.

Table 1- Merits and Demerits of PCA vs. DP-PCA

Technique	Merits	Demerits	Author Remarks
PCA	Popular technique and widely used[1] The method extends to non-random and random variables[2] Overcomes data over-fitting Effective Preprocessing step for data compression Noise Removal	Data Standardization before constructing PCA[5] Responsive to the correlative scaling of various variables[4] Lower Computational Speed[5] Higher Processing Time[5]	It seeks a transformation procedure orthogonally for extraction of the set of uncorrelated linear component- principal from the possibly correlated variables[4] It performs reduction over a single block requiring long processing time and causes insufficient memory usage[5]
DP-PCA	Parallel and distributed execution[4] Higher Computational Speed[5] Lower Processing time[5] Lesser Information loss	None when compared against PCA	MapReduce platform is established over several blocks to perform distributed and parallel computing[4] High Scalability dealing with large datasets[5] Achieves better dimensionality reduction effect[5]

LINEAR DISCRIMINANT ANALYSIS (LDA)

A. BRIEF INTRODUCTION

LDA has been a traditional linear learning algorithm. Fisher first proposed it in 1936 and was earlier known as "Fisher Linear Discriminant (FLD)" [11]. LDA was later adopted in artificial intelligence(AI) domains and recognition of patterns by Belhumeur in the year 1996.

B. WORKING PRINCIPLE

LDA belongs to the group of supervised learning algorithms. It attempts to detect a combination of attributes of two kinds of objects or events, linearly, to categorize and distinguish them [3]. For each given n-dimensional dataset $X = (x^1, x^2, .. x^n)$[3], it finds the corresponding expected value or a category label $y_{(i)}$.

The working ideology of LDA is given under the definition "the variance inside a class is the smallest after projecting while as the variance between classes is the largest [12] ". The LDA classification methodology assumes that the sample data of each category conform to Gaussian distribution(GD), which is conducive to LDA projection [3]. On the arrival of a recent sample, we are able to project it and put the projected sample attribute into the GD density and the probability function of each category to calculate the probability that it belongs to this category. The prediction

category is the one that corresponds to the highest probability.

C. MERITS AND DEMERITS OF THE TECHNIQUE

The merits of this technique are that it always had a rich internal structure [13] and is able to use the probabilistic algorithm effectively for the model's training. Also, LDA capable for performing DR, suitable for massive datasets. So the LDA can be utilized in many areas [15][16][17][18]. LDA is also capable of serving as a classification algorithm rather than being limited to a DR technique.[3]

The traditional LDA model is vulnerable to the singularity problem when dealing with high-dimensional and low sample size data [14]. Some limitations of LDA observed are that the classification accuracy of the LDA is witnessed to be unsatisfactory when used alone[13]. Hence, this technique is often combined with some other techniques like PCA and is utilized in a hybrid system to solve the issue of singularity and give optimum accuracy scores.

INDEPENDENT COMPONENT ANALYSIS (ICA)

A. BRIEF INTRODUCTION

ICA has received considerable attention due to its varied applications, from source-channel separation equalizer to recognition of speech and functional MRI [19]. In addition, its utilization in rectilinear analysis of mixtures for remote sensing images are also discovered [20][21][22][23].

B. WORKING PRINCIPLE

The fundamental ideology of this technique is that it assumes the data to be mixed up linearly with a set of sources that are independent and separate. To demix these signal sources under their independence that is statistically measured is done using mutual information [24]. The ICA aims at finding a demixing matrix V that separates the signal source vector s into a group of statistically independent sources [24]. Several different criteria have been coined to measure source independence [19]. Nevertheless, they all originated from mutual information for measuring the discrepancy between two unexpected sources [25].

C. MERITS AND DEMERITS OF THE TECHNIQUE

This technique has been a valuable supplement of the classical PCA. However, unlike the PCA, which attempts to decorrelate the components in a vector, ICA methods are aimed at making the components as independent as possible [23]. As a result, ICA innovations in remote sensing has become an evolving topic in recent years.

Even though PCA, alongside MNF, utilize the concept to rank-organize eigenvalues to their PCs, some limitations of ICA include no indistinguishable guide present for it to rank-organize ICA-generated ICs [24]. Therefore, J. Wang and C. Chang introduced three IC rank organization and selection criteria. Per these different requirements, three techniques, ICA-DRI, ICA-DR2, and ICA-DR3, pick out a group of most desirable ICs to carry out DR [24].

SINGULAR VALUE DECOMPOSITION (SVD)

A. BRIEF INTRODUCTION

SVD belongs to the class of supervised dimensionality methods [27]. The SVD classifies the data into linearly independent components for performing dimensionality reduction. In a paper, Y. Jaradat and colleagues [26] introduced SVD, an effective matrix breakdown methodology. SVD is the underlying computing engine of several approaches like PCA, eigendecomposition, matrix decomposition, Cholesky decomposition, etc.

B. WORKING PRINCIPLE

In SVD, the breaking down of the matrix dataset X is presented by $X = U\,SV^T$ [10] and 'U' contains singular vectors on the left, 'V' contains singular vectors on the right side, and 'S' consists of the values that are singular. The altered data is then projected onto the spatiality defined by k left singular vectors, which correspond to k-most significant singular values to get the DR results from data : $X^{SVD} = U_k^r\,SV^T$ [10] ; Where U_k Contains the k left singular value.

C. MERITS AND DEMERITS OF THE TECHNIQUE

A few of the merits of SVD are that it is the most effective and widely used matrix factorization technique in computation [26]. SVD is considered the computational engine of many data-driven algorithms and applications. SVD is used in PCA, where high dimensional data is decomposed into lower-dimensional data's statistically dominant patterns. SVD is more commonly used in clustering and classifications, signal processing, orthogonal decomposition, dynamic mode decomposition, etc.[26]

A. Winursito and R. Hidayat [27] compared the accuracies of SVD and PCA methods when they were used in a hybrid system combined with Mel Frequency Cepstral Coefficients (MFCC), a popular feature extraction method used in the recognition systems of speech.

Table 2- Comparison PCA used with MFCC [27]

n-dimensional data (PCA)	Accuracy(%)
10	86.43
12	85.71
14	64.29

Table 3- Comparison SVD used with MFCC [27]

n-dimensional data (SVD)	Accuracy(%)
10	88.57
12	88.57
14	90.71

It is observed that SVD+MFCC gives a better accuracy score than PCA+MFCC for the given n-dimensional data range 10-14.

Feature Selection Techniques

A. Brief Introduction

Two primary goals of the FS technique are to detect self-explanatory features having high distinguishable robustness betwixt task-detecting (TD) classes and to minimize the feature vector's dimensionality acuired from the feature extraction step [28]. This phase is proved vital due to the various benefits it provides, (i) minimizing computing difficulty, (ii) removing the issue of scalability, and (iii) enhancing the predictive accuracy of diagnosis.

Two main types of FS methods are discussed below:

1) Filter-Based FS.

2) Wrapper-Based FS.

B. Working Principle (Filter-Based FS)

It mainly consists of three main stages: (i)Generation of the attribute set, (ii) measurement step, and (iii) testing by a learning algorithm [28]. It uses statistical tests taken to determine the group of attributes having the most significant predictive power. Fisher Criterion Score (F-Score) is considered to select the optimal features.

C. Merits and Demerits of the Technique (Filter-Based FS)

Merits of using the filter-based FS approach is that the F-Score considered calculates the significance of each attribute autonomously of the other features by analogizing its association with the outcoming feature [28].

A setback to be emphasized here is that the F-Score only inspects the distinguishable power of each attribute. It cannot detect the distinguishability of several attributes. Therefore, attributes having lesser score values will not be considered, even if they complement the top attributes and are relevant [28].

D. Working Principle (Wrapper-Based FS)

Most of its working principle is the same as the filter-based FS technique. However, unlike the measurement phase in the filter-based FS approach, an algorithmic learning technique is utilized in the wrapper-based FS approach. Due to this learning algorithm, the wrappers help attain optimum FS outcomes in several cases [28].

E. Merits and Demerits of the Technique (Wrapper-Based FS)

While using a learning algorithm in the wrapper-based FS approach, we find better accuracy scores when compared with the filter-based FS technique. However, on the other hand, the learning technique used here is the primary reason why wrappers perform slowly as compared to their filter-based counterpart techniques [28]. Hence it is observed here that the accuracy of the wrapper-based approach is coming at the cost of computation time of the learning technique.

Table 4- A Comparison of Feature Selection Techniques by A. M. Alhassan [29]

FS techniques	Type	Advantage	Disadvantage
Fisher score	Filter	It measures the relevance of attribute subsets efficiently.	The dependency of one attribute over other attribute is not considered.
Correlation	Filter	A multivariate filtering technique prioritizes the attributes over the interrelational heuristic evaluation function.	Incapable of finding solid interactions.
Backward elimination technique	Wrapper	Effectively identifies the interdependencies among attribute subsets.	Classifiers execute several times to check the quality of attributes.
Sequential forward selection	Wrapper	Simple and avoids overfitting effectively.	The selected subsets are specific to the classifiers under consideration.

Feature Extraction

A. Brief Introduction

Feature extraction transforms unaltered data into numerous attributes that can be processed, reserving the information in the unaltered dataset. The most preferred FE methods include Gradient attributes, structural attributes, regional attributes, projection histograms, Zernike moments, and zoning.[31]

B. Working Principle

Unlike the FS techniques consisting of selecting a group of essential attributes while maintaining necessary information, in FE, transforming the unaltered group of attributes into a reduced group of attributes is carried out by maximizing

disjuncture between classes in the compressed data space [30].

C. MERITS AND DEMERITS OF THE TECHNIQUE

As existing FE methods have limitations related to time and accuracy, S. Ansari[31] presents a faster, more efficient, and optimized feature vector for handwriting recognition. These attributes combine structural attributes, regional attributes, and gradient attributes. Consequently, S. Ansari[31] claims that the extraction technique totals 91 attributes, which was the highest compared to the attributes generated from all other techniques.

AUTOENCODERS

A. BRIEF INTRODUCTION

In 2006, Hinton and Salakhutdinov[32] identified an approach from deep learning that applies to DR, utilizing neural nets that self-learn and predict from their input. These networks are known as autoencoders. It yields a non-linear DR that surpasses SVD-based techniques once trained.

B. WORKING PRINCIPLE

For an input x and the projection z, the applied remodeling x', the AE is assembled using two networks:

An encoder is described by a function $f(x) = z$ [33] for which x belongs to the input values; z is the network's output. A decoder is given by a function $g(z) = x'$[33] Here z is the input value while x' is the network's output.

The training focuses on reducing the distance betwixt the applied remodeling and its input.

$$g(f(x)) = x'[33]$$

It is vital to create a threshold capacity for the AE model to resemble its incoming variables over its generated variables and withdraw the valuable properties.

The autoencoder is given training for both encoder and decoder. Still, when applied to DR, the decoder part is discarded, and the encoder's output is treated as the data's projection [33].

C. MERITS AND DEMERITS OF THE TECHNIQUE

Autoencoders utilizing neural net would require longer computation time and resources for training than the techniques based on SVD. It shows that neural nets are not worthy regarding applications having limited resources prototyping [33].

The neural networks resolve by settling for a miniature value of an objective function and between training. In contrast, SVD-based techniques like PCA are deterministic and exact.

Table 5- Comparative study of PCA vs. ISOMAP vs. AutoEncoders, accuracies obtained using DR techniques over a k-NN Classifier[33]

Technique	Dataset		
	MNIST	Fashion MNIST	CIFAR-10
PCA	97.47 ± 0.07	85.55 ± 0.04	42.34 ± 0.22
ISOMAP	94.86 ± 0.15	81.16 ± 0.21	33.60 ± 0.51
DAE	97.81 ± 0.07	87.61 ± 0.10	45.76 ± 0.36

ISOMETRIC MAPPING (ISOMAP)

A. BRIEF INTRODUCTION

ISOMAP was first introduced by Tenenbaum, V.de Silva [34] in 2000. Isomap is derived from multidimensional scaling (MDS) [57], which finds an embedding that is non-linear and keeps interpoint distances consistent [33]. Isomap is a freshly proposed technique for manifold learning and non-linear DR technique. Using the Euclidean distance, ISOMAP defines the interconnectivity of point of data utilizing its nearest neighbors in the complex dimensional space [35].

B. WORKING PRINCIPLE

This technique uses interpoint geodesic distances instead of finding euclidean distances. Encoding the structure for the space of input into distances [35]. These distances are calculated by establishing a sparse graph where every vertex gets connected to its neighboring closest vertex. This distance betwixt every node pair is kept shortest and fed as the classical MOS input.

C. MERITS AND DEMERITS OF THE TECHNIQUE

Nonlinear transforms using Isomap particularly appear to be helpful in transformations to low dimensions [35]. Furthermore, the DR algorithm Isomap is robust and finds meaningful insights inside low-dimensional structures [36]. Among the various methods, Isomap is highly effective and widely used. The technique gives an easy method for finding the intrinsic geometry of a data manifold utilizing a rough estimate[37].

W. Baozhu [35] presented a method (ISO-MMI) that was used to overcome the shortcomings of traditional Isomap. It avoids the disadvantages of the linear transform algorithm, such as a high error rate and limited data. In addition, it settles the problems of Isomap, like getting the optimal objective transform function too complex and redundant.

KERNEL PRINCIPAL COMPONENT ANALYSIS (KPCA)

A. BRIEF INTRODUCTION

The KPCA is a valuable supplement of PCA utilizing methods based on kernel. While traditional DR methods like PCA and LDA are linear and incapable of reflecting the complex order interrelationships [38]. In contrast, manifold

learning techniques can obtain a more powerful and precise remodeling of non-linear manifolds.

B. WORKING PRINCIPLE

KPCA extends PCA over a kernel space. For example, for a given mapping: φ: $R^h \rightarrow F$ [38] via the given spatiality R^h to an infinite-dimension space called Hilbert space F, the set may transform to a linearly separable space in F. We theoretically implement PCA on F to obtain its non-linear distribution[38].

C. MERITS AND DEMERITS OF THE TECHNIQUE

The KPCA technique is robust in solving the compact size problem, and the data distribution is mapped from its input space R^{h} to its sample space R^{m} where m is 90 with 95% knowledge about covariance[38]. The KPCA algorithm maps the unaltered data onto an infinite-dimensional Hilbert Space through implicit space transformation [39]. The results will have linear properties, and the PCA procedure becomes compatible for extracting attributes hereafter. Considering the merits of the KPCA method in processing high-dimensional and complex data, it was used by [40] for intrusion detection alongside BP Neural nets.

UNIFORM MANIFOLD APPROXIMATION AND PROJECTION (UMAP)

A. BRIEF INTRODUCTION

L. McInnes, J. Healy, and J. Melville[41] introduced the UMAP technique. UMAPs have been a non-linear DR technique widely used for visualizations. It is also a general-purpose DR algorithm in applied machine learning without demarcation over computations on the embedding dimensions [42]. Moreover, it contains several useful features that include the transformation of new samples of data into a pretrained embedding space and using labels for supervised and semi-supervised DR.

B. WORKING PRINCIPLE

The technique mainly comprises a weighted, force-directed graph deployment into lower dimensionality.

The essential steps explained according to [43] include:

1. Given the set $X = \{x_i\}, i = 1 \ldots N$ data present in a high dimensions spatiality R^M, find the sets η_i, having k-neighbor points for every dataset points x_i ; $i = 1 \ldots N$.

2. For each 'i-th' data point, find its neighbor nearest with the minimum distance.

$$\rho_i = \min \left(d(x_i, x_j) \middle| x_j \in \eta_i, d(x_i, x_j) > 0\right) \text{ [43]}$$

3. Constructing the UMAP graph G as an undirected weighted graph with adjacency matrix:

$$B = A + A^T - A \cdot A^T \text{ [43]}$$

4. Following the data samples present in low dimension spatiality are given using the force-directed deployment of the graph using attractive F^a and repulsive F^r.

C. MERITS AND DEMERITS OF THE TECHNIQUE

Within specific clusters, UMAP can determine local internal relationships while capturing global relationships betwixt the clusters [42]. This exclusive trait of UMAP makes it preferable to t-SNE in protecting structure elements of the data in complex dimensional topological space. Moreover, this feature is fruitful in the inference of gene relationships.

Another exciting discovery is that the refinement of classification for UMAP is similar for simple and complex dimensional output spaces as compared to the other techniques [43]. The UMAP technique is effective in 2D and 3D mappings without significant losses within the quality of classification. Following the results from [42], UMAP would perform stronger when accompanied by a linear reduction method.

T-SNE

A. BRIEF INTRODUCTION

T-SNE and UMAP are projection-based methods for DR. The t-SNE is perhaps a more commonly used reduction technique [42]. T-SNE aims at protecting and maintaining the local structures inside the embeddings while revealing the global information, like the existence of clusters at several scales [44]. T-SNE is mainly utilized for the visualization of complex data. Its application ranges from life sciences to deep-learning technique analytics.

B. WORKING PRINCIPLE

T-SNE learns from overall distances present in between the data samples in the complex dimension spatiality as symmetric joint-probability distribution 'P' [44]. Similarly, a distribution Q over joint-probability is computed, which depicts the low-dimensional similarity. The objective is to obtain a depiction embedding where Q depicts P in the simple dimensional space.

It can be obtained by improving the positions in the simple dimension spatiality to reduce the cost function C presented by the Kullback-Leibler (KL) divergence betwixt the distributions P and Q joint-probability:

$$C(P,Q) = KL(P||Q) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} p_{ij} \ln\left({}^{p_{ij}}/_{q_{ij}}\right) \text{ [44]}$$

Where probability p_{ij} and q_{ij}, depicts the likelihood of these data samples present in complex dimensional spatiality.

C. MERITS AND DEMERITS OF THE TECHNIQUE

This technique is a modification of neighbor stochastically generated embedding that is much effortless to optimize by

minimizing the propensity to crown points at the centroid of the map and producing outstandingly finer visualization [42].

All inclusively, UMAP and t-SNE results of prediction along with linear regression are commensurable, but refining time taken by UMAP is remarkably lesser to t-SNE, thus conserving the cost of the CPU and memory's computation notably. (Refer Table-7) Also, [44] mentions that t-SNE is steerable computationally because the user defines the approximation level to specific, exciting areas.

Table 6- Comparision UMAP vs.T-SNE FAULT ANALYSIS given by [42]

FAULT ANALYSIS	AVERAGE (%) FAULT WITH PCA	AVERAGE (%) FAULT WITHOUT PCA
UMAP	14.39%	15.63%
t-SNE	15.11%	16.27%

Table 7- Comparision UMAP vs. T-SNE TIME ANALYSIS given by [42]

SIZE OF SAMPLE	1200	2800
UMAP	4.7 sec	7.1 sec
t-SNE	14.8 sec	36.1 sec

FACTOR ANALYSIS (FA)

A. BRIEF INTRODUCTION

FA [45], [46] has been a technique for the linear DR of Gaussian random variables. It is in close relation to PCA [47], which extracts the subset space given by the most significant eigenvectors in the matrix of covariance. FA has many potential applications like segmental HMM[53] in speech recognition. Furthermore, in model adaptation [54], [55], [56], the parameters are recalculated to exact new testing conditions, presenting a more efficient and powerful substitute for full covariance matrices.

B. WORKING PRINCIPLE

To design the structural orientation of covariance of complex dimensional data, FA utilizes a small number of parameters [48]. For example, concerning automatic recognition of speech, the constraints are selected in two ways: i) by increasing the occurrence of signals of speech, or ii) by effectively reducing the number of fault occurrences of classification [49], [50].

FAs working definition is different from probabilistic PCA only in the distribution conditionality of the variable observed such that the latent variable has a covariance of diagonal rather than an isotropic one[58]. Moreover, FA converts multiple measured variables into a few unrelated comprehensive indicators[59]. So, FA can reduce dimensionality and simplify data [60].

C. MERITS AND DEMERITS OF THE TECHNIQUE

Although FA is closely related to PCA and does not suffer from two crucial disadvantages of PCA [48], viz :

i) Undefined a proper density model outside the subset spatiality of PCs.

ii) Discrepancies present componentwise excluding this subset spatiality that is uniformly modeled, even if the data does not permit such an assumption;

Factor analysis is not vulnerable to each of these demerits, although it does contain the classical PCA ([51], [52]) as a particular resticting case. Furthermore, in comparison with full covariance matrices, factored ones are simpler to exploit and more effective in overfitting [48]. Finally, the predictive performance of automatic recognition of speech is measured using speed, memory, and accuracy; fortunately, FA helps in all three aspects.

CONCLUSION

This paper reviews various DR techniques used to effectively reduce the high-dimensional features by getting a principal variable set. First, a brief introduction of various techniques used and how they were developed is provided. Additionally, the paper explicitly discusses the working phenomenon, pros, and cons of the techniques and their performance compared to their similar counterparts.

Performing the DR over the unaltered dataset is very crucial as it has the potential to alter the outcomes of a machine learning classifier. Dimensionally reduced data gives optimum classification results. Several linear and non-linear techniques are discussed in the paper. We observe that a hybrid system that utilizes two or more techniques (like PCA in combination with other techniques) generally has a lesser error percentage and, in most cases, outperforms a single standalone DR technique.

A dimensionally reduced dataset helps bring about data compression for storage, removal of redundant features and noise, and the non-essential features that negatively correlate with the predictive performance of a classification model. Future research towards developing more efficient hybrid DR techniques needs to be explored that help in optimizing the accuracy of prediction models and bring computational effectiveness to the results.

REFERENCES

Pearson K. On lines and planes of closest fit[J]. Philosophical Magazine, 1 901, 6.

H.Hotelling. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24: 417-441, 1 933.

S. Feng and H. Wang, "Comparison of PCA and LDA Dimensionality Reduction Algorithms based on Wine Dataset," 2021 33rd Chinese Control and Decision Conference (CCDC), 2021, pp. 2791-2796, doi: 10.1109/CCDC52312.2021.9602325.

J. Zhu, Z. Ge, and Z. Song, "Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with big data," IEEE Trans. Ind. Informatics., 2017, doi: 10.1109/TII.2017.2658732.

L. Wang, "Research on Distributed Parallel Dimensionality Reduction

Algorithm Based on PCA Algorithm," 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019, pp. 1363-1367, doi: 10.1109/ITNEC.2019.8729427.

Xu Yajing, Wang Yuanzheng. The Improvement of the Application Method of Principle Component Analysis[J]. MATHEMATICS IN PRACTICE AND THEORY, 2016, 36(6): 68-75.

Wang Xinghua, Xu Xuanhao, Zhou Yawu, A Clustering Algorithm of Power Userload Curves Based on Pearson correlation Coefficient[J]. Heilongjiang Electric 7, 39(5) / 397 – 401.

S. Feng and H. Wang, "Comparison of PCA and LDA Dimensionality Reduction Algorithms based on Wine Dataset," 2021 33rd Chinese Control and Decision Conference (CCDC), 2021, pp. 2791-2796, doi: 10.1109/CCDC52312.2021.9602325.

C. Yumeng and F. Yinglan, "Research on PCA Data Dimension Reduction Algorithm Based on Entropy Weight Method," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2020, pp. 392-396, doi: 10.1109/MLBDBI51377.2020.00084.

M. Vikram, R. Pavan, N. D. Dineshbhai and B. Mohan, "Performance Evaluation of Dimensionality Reduction Techniques on High Dimensional Data," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 1169-1174, doi: 10.1109/ICOEI.2019.8862526

Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7.

Wei Feng. Research on feature extraction and feature selection of hyperspectral remote sensing data [D]

X. Liu, H. Xiong, and N. Shen, "A hybrid model of VSM and LDA for text clustering," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), 2017, pp. 230-233, doi: 10.1109/CIAPP.2017.8167213.

S. Ji and J. Ye, "Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection," in IEEE Transactions on Neural Networks, vol. 19, no. 10, pp. 1768-1782, Oct. 2008, doi: 10.1109/TNN.2008.2002078.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993-1022.

Bhattacharya, Indrani, Sil, Jaya, Sparse representation based query classification using LDA topic modeling, Advances in Intelligent Systems and Computing,v 469,p 621-629,2016.

Liu, Q.,Chen, E., Xiong, H.,Ge, Y., Li,Z., and Wu, X., "A Cocktail Approach for Travel Package Recommendation," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.26, NO.2, FEBRUARY 278-293,2014.

Yue Liu Shimin Wang., and Qian Cao., Research on Commodities Classification Based on LDA IMM 2015, Lancaster:DEStech Publivations 2015: 189-191.

A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Anal ysis. New York: Wiley, 2001.

J. Bayliss, J. A. Gualtieri, and R. F. Cromp, "Analyzing hyperspectral data with independent component analysis," Proc. SPIE, vol. 3240, pp. 133-143, 1997.

T. M. Tu, "Unsupervised signature extraction and separation in hyperspectral images: a noise-adjusted fast independent component analysis approach," Opt. Eng., vol. 39, no. 4, pp. 897-906, 2000.

C.-I Chang, S. S. Chiang, J. A. Smith, and I. W. Ginsberg, "Linear spectral random mixture analysis for hyperspectral imagery," IEEE Trans. Geosci. Remote Sens., vol. 40, no. 2, pp. 375-392, Feb. 2002.

X. Zhang and C. H. Chen, "New independent component analysis method using higher-order statistics with applications to remote sensing images," Opt. Eng., vol. 41, pp. 1717-1728, Jul. 2002.

Jing Wang and Chein-I Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," in IEEE Transactions on Geoscience and Remote Sensing, vol. 44, no. 6, pp. 1586-1600, June 2006, doi: 10.1109/TGRS.2005.863297.

T. Cover and J. Thomas, Elements of Information Theory. New York: Wiley, 1991.

Y. Jaradat, M. Masoud, I. Jannoud, A. Manasrah, and M. Alia, "A Tutorial on Singular Value Decomposition with Applications on Image Compression and Dimensionality Reduction," 2021 International Conference on Information Technology (ICIT), 2021, pp. 769-772, doi: 10.1109/ICIT52682.2021.9491732.

A. Winursito, R. Hidayat, A. Bejo, and M. N. Y. Utomo, "Feature Data Reduction of MFCC Using PCA and SVD in Speech Recognition System," 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 2018, pp. 1-6, doi: 10.1109/ICSCEE.2018.8538414.

K. Pavya and B. Srinivasan, "Feature selection algorithms to improve thyroid disease diagnosis," 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), 2017, pp. 1-5, doi: 10.1109/IGEHT.2017.8094070.

A. M. Alhassan, W. M. N. W. Zainon, Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis, IEEE Access. 9 (2021), 87310–87317. doi: 10.1109/ACCESS.2021.3088613

S. Ghosh and P. Pramanik, "A Combined Framework for Dimensionality Reduction of Hyperspectral Images using Feature Selection and Feature Extraction," 2019 IEEE Recent Advances in Geoscience and Remote Sensing: Technologies, Standards, and Applications (TENGARSS), 2019, pp. 39-44, doi: 10.1109/TENGARSS48957.2019.8976039.

S. Ansari and U. Sutar, "Optimized and efficient feature extraction method for devanagari handwritten character recognition," 2015 International Conference on Information Processing (ICIP), 2015, pp. 11 15, doi: 10.1109/INFOP 2015.7489342.

G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504-507, July 2006.

Q. Fournier and D. Aloise, "Empirical Comparison between Autoencoders and Traditional Dimensionality Reduction Methods," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. 211-214, doi: 10.1109/AIKE.2019.00044.

J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for non-linear dimensionality reduction," Science, vol. 290, no. 5500, p. 2319, 2000.

Wang Baozhu, Wu Nan, Liu Cuixiang, and Jia Kejin, "Dimensionality reduction based on Isomap and Mutual Information Maximization," 2010 The 2nd Conference on Environmental Science and Information Application Technology, 2010, pp. 829-832, doi: 10.1109/ESIAT.2010.5567461.

Y. Zou, D. Du, and G. Guo, "Image matching for weld seam tracking based on non-linear dimensionality reduction method Isomap," 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010, pp. 2826-2829, doi: 10.1109/FSKD.2010.5569249.

M. Balasubramanian, E.L. Schwartz, J.B. Tenenbaum, et al. The Isomap Algorithm and Topological Stability. Science 295(5552), 2002: 7

J. Shu, W. Liu, F. Meng, and Y. Zhang, "Multi-feature Image Retrieval by Nonlinear Dimensionality Reduction," 2014 Seventh International Symposium on Computational Intelligence and Design, 2014, pp. 6-9, doi: 10.1109/ISCID.2014.206.

H. Zhou, C. Gao and X. Liu, "Design and Optimization of Nonlinear Dimensionality Reduction Algorithm for Hyperspectral Images on Heterogeneous System," 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017, pp. 1076-1081, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.164.

D. Hongwei and W. Liang, "Research on Intrusion Detection Based on KPCA-BP Neural Network," 2018 IEEE 18th International

Conference on Communication Technology (ICCT), 2018, pp. 911-915, doi: 10.1109/ICCT.2018.8600090.

L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426, 2018.

K. Pal and M. Sharma, "Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space," 2020 Fourth International Conference on I-SMAC (loT in Social, Mobile, Analytics, and Cloud) (I-SMAC), 2020, pp. 1106-1110, doi: 10.1109/1 SMAC49090.2020.9243502.

E. Myasnikov, "Using UMAP for Dimensionality Reduction of Hyperspectral Data," 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 2020, pp. 1-5, doi: 10.1109/Far EastCon50210.2020.9271656.

N. Pezzotti, B. P. F. Lelieveldt, L. v. d. Maaten, T. Höllt, E. Eisemann and A. Vilanova, "Approximated and User Steerable tSNE for Progressive Visual Analytics," in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 7, pp. 1739-1752, 1 July 2017, doi: 10.1109/TVCG.2016.2570755.

B. Everitt, An Introduction to Latent Variable Models. London, U.K.: Chapman and Hall, 1984.

D. Rubin and D. Thayer, "EM algorithms for factor analysis," Psychometrika, vol. 47, pp. 69-76, 1982.

R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973.

L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," in IEEE Transactions on Speech and Audio Processing, vol. 8, no. 2, pp. 115-125, March 2000, doi: 10.1109/89.824696.

B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," IEEE Trans. Speech Audio Processing, vol. 5, pp. 266-277, 1997.

B. Juang and S. Katagiri, "Discriminative learning for minimum error classification," IEEE Trans. Acoust., Speech, Signal Processing, vol. 40, pp. 3043-3054, 1992.

S. Roweis, "EM algorithms for PCA and SPCA," in Advances in Neural Information Processing Systems, M. Jordan, M. Kearns, and S. Solla, Eds. Cambridge, MA: MIT Press, 1998, vol. 10, pp. 626-632.

M. E. Tipping and C. Bishop, "Mixtures of principal component analyzers," in Proc. IEEE 5th Int. Conf. Artificial Neural Networks, 1997.

M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. 4, pp. 360-378, 1996.

J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Acoust., Speech, Signal Processing, vol. 2, pp. 291-298, 1994.

C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput., Speech, Lang., vol. 9, pp. 171-185, 1995.

G. Zavaliagkos, R. Schwartz, and R. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in Proc. ICASSP 1995, pp. 676-679.

Z. Zhou, J. Mo, and Y. Shi, "Data imputation and dimensionality reduction using deep learning in industrial data," in 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Dec 2017, pp. 2329–2333.

B. Zhao, J. R. Sveinsson, M. O. Ulfarsson, and J. Chanussot, "(Semi-) Supervised Mixtures of Factor Analyzers and Deep Mixtures of Factor Analyzers Dimensionality Reduction Algorithms For Hyperspectral Images Classification," IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 887-890, doi: 10.1109/IGARSS.2019.8898932.

D. Zhang and Y. Gong, "The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure," in IEEE Access, vol. 8, pp. 220990-221003, 2020, doi: 10.1109/ACCESS.2020.3042848.

Z. Liu, Q. Zhan, and G. Tian, "A review of a comprehensive evaluation of factor analysis," Statist. Decis., vol. 5, no. 19, pp. 68-73, 2019.