Information Fusion 59 (2020) 44–58

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Overview and comparative study of dimensionality reduction techniques for high dimensional data

Shaeela Ayesha, Muhammad Kashif Hanif*, Ramzan Talib

Department of Computer Science, Government College University, Faisalabad, Pakistan

ARTICLE INFO

Keywords:
Dimensionality reduction
Features
High dimensional data
Linear techniques
Nonlinear techniques

ABSTRACT

The recent developments in the modern data collection tools, techniques, and storage capabilities are leading towards huge volume of data. The dimensions of data indicate the number of features that have been measured for each observation. It has become a challenging task to analyze high dimensional data. Different dimensionality reduction techniques are available in literature to eliminate irrelevant and redundant features. Selection of an appropriate dimension reduction technique can help to enhance the processing speed and reduce the time and effort required to extract valuable information. This paper presents the state-of-the art dimensionality reduction techniques and their suitability for different types of data and application areas. Furthermore, the issues of dimensionality reduction techniques have been highlighted that can affect the accuracy and relevance of results.

1. Introduction

Over the past few years, a huge volume of digital data is continuously being generated in different application areas. Moreover, the size, heterogeneity, complexity, and dimensionality of data are growing exponentially [1]. Applications of High Dimensional Data (HDD) [2,3] have been found in various domains including biomedical, web, education, medicine, business, and social media [4]. The massive amount of new HDD is continuously evolving in different formats (e.g., text [5], digital images [6], speech signals [7], and videos [8]).

For Machine Learning (ML) models, high dimensionality of data can raise many issues for accurate classification, pattern recognition, and visualization [9]. Learning in high-dimensional spaces or with a large number of features can become difficult due to high computational complexity. The term curse of dimensionality means that if the amount of data for which to train a model is fixed, then increasing dimensionality can lead to overfitting. This issue can be avoided by bringing in exponentially more data for each additional dimension [10].

The Dimensionality Reduction (DR) [11] can be performed through feature extraction (feature transformation) and feature selection. Feature extraction transforms the original HDD sets to new reduced data sets by removing redundant and irrelevant features. New feature set preserves maximum information of the original data set. Feature selection [12] selects the subset of features from input data that are most relevant to the given problem [6,7,13,14]. It becomes a tiresome task to manually extract the suitable and relevant features for different applications. Selection of an appropriate technique for DR can reduce time and effort to select and extract relevant features for analysis.

Dimensionality Reduction Techniques (DRTs) offer an efficient way to reduce the number of input variables (dimensions) before applying ML models. Many DRTs are available that can be applied to reduce computation time and to make efficient use of computing resources. DRTs can be applied at the pre-processing stage before data analysis and development of ML model. A wide variety of DRTs [11,15] are available for different types of data. The major difficulty in applying DRTs is each DRT has been developed to maintain certain aspects of the original data. Therefore, a particular DRT may be suitable for some type of data or application and may be inappropriate for other. Furthermore, some DRTs have been designed under certain constraints that are scope and application limited.

Moreover, change in internal parameters and functions can modify the core functionality of DRT. This study gives a brief overview of the DRTs and their available variants. Furthermore, we identified various application areas and data sets for which DRTs have been successfully applied. Moreover, different issues that may arise during the DR process have been highlighted.

Rest of the paper is organized in different sections. The Section 2 explores different DRTs and variants of these techniques. Section 3 presents a comparative study of DRTs and application areas with available solutions. Section 4 highlights the issues and challenges of DRTs. Section 5 gives a brief discussion. At the end, we conclude the outcomes.

* Corresponding author.
E-mail addresses: ayesha.csgcuf@gmail.com (S. Ayesha), mkashifhanif@gcuf.edu.pk (M.K. Hanif), ramzantalib@gcuf.edu.pk (R. Talib).

https://doi.org/10.1016/j.inffus.2020.01.005
Received 21 May 2019; Received in revised form 8 January 2020; Accepted 13 January 2020
Available online 17 January 2020
1566-2535/© 2020 Elsevier B.V. All rights reserved.

Table 1
List of discriminative features of DRTs.

Features	Description
Supervised Learning [16]	Makes use of labeled data input and target output data are known.
Unsupervised Learning [17]	Discovers patterns from unlabeled data.
Semi-supervised Learning [18]	For both labeled and unlabeled data can be used.
Linear Technique [19]	Represents each reduced dimension as a linear combination of the original dimension.
Non-linear Technique [20]	Each reduced dimension results from a nonlinear transformation of the original dimensions.
Local [21]	Based on neighborhood relationship.
Global [21]	All the pairwise distances are considered.
Parametric [4]	Parametric methods based on an explicit number of parameters involved.
Non-parametric [4]	Non-parametric methods do not assume a specific functional form, rather they can be directly assigned to different data points.
Manifold Learning [21]	Imposes additional restrictions, while mapping high dimensional data to low dimensions. It deals with local and global approaches.
Intrinsic Dimension [22]	Consists of a minimum number of variables required to define the data without significant loss of information.
Curse of Dimensionality [23,24]	The curse of dimensionality is the expression of all phenomena that appears with HDD. This can rise in difficulty when training models with high-dimensional data.

2. Dimensionality reduction techniques

DR is a process of transforming high dimensional representation of data in low dimension representations. With the immense increase in HDD, use of different DRTs has become common in many application areas. Moreover, many new techniques are evolving continuously. DRTs transform original data set having high dimensionality and converts it into a new data set representing low dimensionality while preserving the original meanings of the data as much as possible [25]. The low dimensional representation of the original data helps to overcome the issue of curse of dimensionality [23,24]. The low dimensional data can be easily processed, analyzed, and visualized.

Formally, DRT transforms the high dimensional data $\boldsymbol{Y} = [y_1, y_2, \cdots, y_m] \in \mathbb{R}^{m \times p}$ having p dimensions and m observations into low dimensional data $\boldsymbol{Z} = [z_1, z_2, \cdots, z_m] \in \mathbb{R}^{m \times k}$ where $k < \; < p$ in ideal case. DRT can have implicit, explicit or inverse mapping to reconstruct a sample from the low-dimensional representation [26].

DRTs can be used to extract only relevant features that are useful for analysis while eliminating redundant and unnecessary features. The use of DRTs can reduce the computation time and storage space requirements. For example, reducing dimensions from 100 to 2D or 3D will certainly reduce storage requirement. Moreover, 2D or 3D will provide better interpretation and visualization of data [27]. By eliminating redundant and irrelevant features, DRTs can become helpful for integration of various data sets [28]. Table 1 shows a list of discriminative features of DRTs that will be useful for further understanding. This study provides an overview of linear and nonlinear DRTs.

2.1. Linear dimension reduction techniques

Linear Dimension Reduction Techniques (LDRTs) use simple linear functions to transform HDD into lower dimensions. This section provides a brief overview of different LDRTs and their variants to explore the suitability of specific LDRTs for different applications [19].

2.1.1. Principal component analysis

PCA was originally introduced by Pearson [29] and developed independently by Hotelling [30]. PCA is an unsupervised linear mapping based on an eigen vector search and suitable for Gaussian data. PCA provides different strategies for reducing the dimensionality of feature space and preserves the maximum amount of variance of the original data [31,32]. PCA can be computed using different algorithms including eigen values, latent variable analysis, factor analysis or Linear Regression (LR) [28,33,34]. Major applications of PCA include image and speech processing, visualization, exploratory data analysis, and robotic sensor data [33,35].

The goal of PCA is to identify a set of uncorrelated features known as Principal Components (PCs). The first PC holds the largest amount of variance of original data, while the second PC represents the second largest variance, and so on. First few k PCs hold the largest amount variation of the original data and reduces the dimensions of data from p to k. Only prominent PCs holding maximum information can be used to visualize data in low dimensional space. Let $\boldsymbol{Y} \in \mathbb{R}^{m \times p}$ be an $m \times p$ matrix having m observations and p features. The PCs $z_i \in \mathbb{R}^m$ can be computed as a linear weighted combination of features [34].

$$\boldsymbol{Z} = \boldsymbol{YW} \tag{1}$$

Here $\boldsymbol{Z} = [z_1, z_2, \cdots, z_m] \in \mathbb{R}^{m*p}$ and $\boldsymbol{W} = [w_1, w_2, \cdots, w_p \in \mathbb{R}^{p*p}]$. The transformed data have the smallest reconstruction error and maximum variance among all possible projections [36].

Variants of PCA have been developed for different data, types and structures. Kambhatla and Leen [37] introduced an extension to PCA using local linear approach called Local PCA (LPCA). Experimental results showed better performance of LPCA when compared with NN and PCA for speech and image data. Locantore et al. [38] developed Robust PCA (RPCA) for functional data to enhance the robustness of the traditional algorithm. Hubert et al. [39] proposed Robust PCA (ROBPCA) based on Projection Pursuit (PP) using robust scatter matrix estimation. ROBPCA provided more accurate and computationally fast results than classical PCA. In order to work with outliers and missing elements in the data, Serneels and Verdonck [40] introduced an Expectation Robust PCA and experimental results indicate its suitability for different sizes of data sets.

Vidal et al. [41] proposed Generalized PCA (GPCA) to deal with HDD space having an unknown number of subspaces. GPCA was applied on different data sets for clustering faces, temporal video segmentation, and 3D motion segmentation. Li et al. [42] introduced an improved 2D-PCA for online palm print recognition system and achieved accuracy of approximately 99.72%. Wang et al. [43] proposed an incremental (2D-PCA) for the tracking of moving objects from videos. Qiu et al. [44] used 2D-PCA for the classification of SAR images and attained better results than classical PCA.

For feature extraction and recognition of tensor objects, Lu et al. [45] developed a Multi-linear PCA (MPCA). MPCA based face recognition system has shown better results than original PCA and 2D-PCA. Di et al. [46] introduced a Multilevel Functional PCA to extract multilevel intra-core and inter-subject geometric components of functional data. Moreover, Happ and Greven [47] proposed a Multivariate Functional PCA (MFPCA) to manage multivariate functional data from different domains having different dimensions. Performance of MFPCA is not affected with data sparsity and data with measurement error. Monforte et al. [48] applied MFPCA for the development of an automatic robotic arm system to achieve human like grip from a robotic system.

Metsalu and Vilo [49] introduced an interactive module called ClustVis. ClustVis visualizes clusters of multivariate data using PCA and heatmap and provides an interactive environment to makes modification easier. The experimental results can be viewed and stored in different formats. Tree-structured Multi-linear PCA (TM-PCA) reduces the dimensions of input sequences and sentences that makes the sentence classification easy and simple. TM-PCA with SVM has shown better performance than recurrent neural network for text data classification [50].

Zou et al. [51] proposed Sparse PCA (SPCA) to manage sparsity of gene expression data. Journée et al. [52] developed a Generalized Power method for Sparse PCA (GP-SPCA) to overcome the curse of dimensionality issue. Fabris et al. [53] used SPCA for the description of glucose variability in type 2 diabetes. Yi et al. [54] proposed Joint Sparse PCA (JSPCA) to select useful features from the data and to efficiently deal with outliers. A new version of SPCA uses a hybrid approach based on robust and scalable algorithm [34].

PCA has been extensively used for exploring HDD sets to identify dominant trends. Enhancing visualization and exploring patterns specific to one data set is a challenging task. Abid et al. [55] proposed Contrastive PCA (cPCA) to get low dimensional structures from the given data and find better insights. cPCA has shown suitability for application areas where PCA has already been used. Cardot and Degras [56] discussed different methods for online PCA such as perturbation techniques, incremental methods, and stochastic optimization. Fan et al. [57] analyzed PCA for big data and focused on community detection, ranking, mixture model, and manifold learning. Random Permutation PCA (RP-PCA) and RP-2D-PCA have been applied for efficient image recognition in a biometric system [58].

2.1.2. Singular value decomposition

Singular Value Decomposition (SVD) is an unsupervised LDRT that is closely related to PCA. SVD can be used to solve metric equations and data reduction problems [59–62]. Researchers have used SVD in different areas including digital image processing [63], taxonomic classification of biological sequences [64], pattern recognition [65], gene expression data [28,32,61], signal processing [66,67], NLP [68], bioinformatics [64], and text summarization [69].

SVD is specifically developed for matrix decomposition. SVD can be applied to any real-world matrix. It factorizes matrix $\boldsymbol{Y} \in \mathbb{R}^{m\times p}$ into $\boldsymbol{USV}^T$. Matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ are two orthogonal matrices having dimensions $m\times k$ and $p\times k$, respectively. Matrix $\boldsymbol{S}$ is a $k\times k$ diagonal matrix containing singular values of $\boldsymbol{Y}$. Number of singular values obtained is k [70]. One important feature of SVD is to reconstruct original matrix $\boldsymbol{Y}$ using matrices $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{S}$. SVD is computationally hard but can be improved with random sampling, sensitive to outliers, and nonlinearities in data. Results are not always the best for visualization and interpretation can becomes difficult. SVD follows a de-facto standard for DR process in generic data sets and provides an optimal level of DR for linear projections [71,72].

To work with orthogonal and sparsity issue, constrained SVD was proposed that can incorporate multiple constraints to efficiency of classical SVD[73]. Husson et al. [74] introduced a multi-level SVD based imputation method. This method is useful for efficient pre-processing and management of diverse nature of data collected from multiple sources. This new approach can be useful in different fields such as education, medical and life science.

2.1.3. Latent semantic analysis

Over the past few years, better exploration of scientific knowledge from text documents has gained popularity. With an explosive growth of text documents, term similarities in different fields and their association are creating many issues. A term may cause conflict in identification and classification. For example, a Fly and fly both have the same spellings but one can be used as noun and other as a verb [75]. Latent Semantic Analysis (LSA) is an unsupervised linear mapping designed for text documents based on the PCA or SVD computation. LSA is a vector based technique used to compare and represent the text of HD corpus data into lower dimensions [5,76,77]. LSA is useful to retrieve relevant documents from a huge collection of documents. Deerwester et al. [78] performed indexing using LSA to discover relevant documents based on terms found in the queries.

LSA aims to learn semantic representation of text and infer associations among words [79]. LSA takes $m\times p$ words from co-occurrence matrix $\boldsymbol{Y} \in \mathbb{R}^{m\times p}$ as input. Each matrix entry y_{ij} represents local frequency of a given word i for a given document j. The co-occurrence counts are transformed to weight to determine information about the meaning of a document. The transformed matrix of weighted terms has dimension factors and can be reduced with SVD computation. For instance, LSA $[\boldsymbol{USV}^T] = SVD(\boldsymbol{Y}, k)$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices. Finally, a new matrix is reconstructed from the reduced decomposed matrix and similarity measures are calculated from the reconstructed matrix [76,79]. In LSA, singular values computed using SVD are used to represent the dimensions of meaning for words and passages in the language. The dimensionality of these semantic representations can be reduced by replacing some of the singular values with 0.

LSA is being used in numerous applications for categorization of terms and documents, summarization of large collections of documents [80,81], searching words and their meanings, finding pairs of synonym and antonym, educational applications [76,82,83], to analyze the professional's competency pattern for successful business management [84], to analyze communication among patient and physician to prescribe medication [85]. LSA can also be applied for text [77], images and videos[1] and to identify medical records [86]. Furthermore, Santilli et al. [87] performed an analogous comparison of twitter posts using LSA.

Hofmann introduced a Probabilistic LSA (PLSA) model for effective information retrieval, Natural Language Processing (NLP), ML, and related areas [88,89]. Many categories of text documents and linguistic data collections were accessed through automatic indexing of documents and results indicate a substantial and consistent improvement of the probabilistic method over standard LSA. Zhai and Geigle [90] provided a broad overview and applications of PLSA. Si and Jin [91] proposed a Regularized Probabilistic LSA (RP-LSA) model to adjust model flexibility and avoid the over fitting issue. Tu et al. [92] introduced hk-LSA for reducing the dimensions of text documents. Uysal and Gunal [93] introduced a Genetic Algorithm based on Latent Semantic Features (GALSFs) to improve the text classification. The proposed approach combined the feature selection and extraction phases. Discriminative PLSA approach is used for face recognition and achieved success in recognizing face based on single training sample [94].

2.1.4. Locality preserving projections

Locality Preserving Projections (LPP) is an unsupervised LDRT that relies on the linear approximation of the nonlinear Laplacian Eigen map [95]. Graphs can be developed using Laplacian graph notation [95,96]. Linear projective maps are used to solve variation issues and optimal preservation of neighborhood structure of data. This method preserves the distance between samples when projecting data to lower dimensions [96–98].

LPP finds $w_1, w_2, \cdots, w_m$ vectors to map high dimensional data $\boldsymbol{Y}$ into low dimensional data by Eq. 2 [95]. $w^T YDY^T w$ should be equal to one.

$$\text{argmin}_w = \sum_{ij} \left\| w^T y_i - w^T y_j \right\|_2^2 A_{ij} \tag{2}$$

$\boldsymbol{A}$ is an affine similarity matrix and computed using Eq. 3.

$$A_{ij} = exp\left(-\frac{\left\| y_i - y_j \right\|^2}{\beta} \right) \tag{3}$$

where β is average squared distance between all pairs. Eq. 2 can be reduced to find eigen value decomposition

$$YLY^T w = \lambda YDY^T w \tag{4}$$

where $L = D - A$ and λ_i are eigen vectors.

LPP has been applied in different application areas including image retrieval [99], image and video classification, face recognition [16,100], pattern recognition [101], automatic speech recognition [6], and computer vision [98]. Huang et al. [102] introduced a Locality-Regularized Linear Regression Discriminant Analysis (LL-RDA) based on LL Regression Classification (LLRC) [103]. LL-RDA was derived by maximizing the inter-class reconstruct of local scatters at the same time minimizing the intra-class reconstruction of local scatters.

He [99] introduced an incremental semi-supervised LPP (SLPP) learning algorithm using PCA and LPP for image retrieval. SLPP provides better representation in lower dimensions and preserves semantic information of the image. Moreover, Xia and Rui-xia [16] applied supervised LPP for human face recognition. Cheng et al. [104] proposed Supervised Kernel LPP (SKLPP) to enhance face recognition accuracy. SKLPP increased the approximation for face manifold learning and enhanced within class relations.

Yu et al. [105] proposed Discriminant LPP (DLPP) for face recognition that offers better performance than original LLP. Furthermore, it reduced noise in images and transformation difference without sacrificing much of intrinsic difference. DLPP has shown good performance for pattern recognition [105]. However, DLPP cannot solve the small sample size problem. Lu et al. [106] introduced a Matrix Exponential-Based Discriminative LPP (MED-LPP) to solve the small sample size problem efficiently. Zhong et al. [107] presented DLPP based on L1-norm maximization to improve performance for pattern recognition. The proposed method was efficient in the presence of outliers and resolve the issue of small sample size.

For efficient image recognition, Zhu and Zhu [100] developed an Orthogonal DLPP (ODLPP). Experiments on Yale Face and AR face data showed that ODLPP performed better than Eigenface, Fisherface, Laplasianface, and Orthogonal LPP (OLPP). Lu et al. [108] introduced Discriminant LPP based on Maximum Margin Criterion (DLPP/MMC). Experiments performed on ORL, Yale and Pie face data sets showed DLPP/MMC performed better than DLPP, LPP, LDA + PCA, LDA, and PCA methods and achieved higher image recognition accuracy. For Facial expression recognition, Turan et al. [109] presented a Soft Locality Preserving Map (SLPM) method. SLPM effectively reduce the dimensions of the feature vectors and enhance the discriminative power of the extracted features for expression recognition. SLPM feature generation method enhanced the generalizability of the underlying manifold learning for facial expression recognition. Wang et al. [110] proposed a Grassmann manifold based on the LPP (GLPP). Experimental results showed the effectiveness of GLPP for image/video classification when compared with HD Grassmann manifold.

To extract useful features and pattern recognition, Li et al. [101] introduced a Kernel based Self-Optimized Locality Preserving Discriminant Analysis (KSLPDA) which performed better than PCA, LDA, LPP, SLPP and SKLPP. Wong and Zhao [98] proposed Supervised Optimal LPP (SOLPP) and Normalized Laplacian-based Supervised Optimal LPP (NL-SOLPP) to enhance the efficiency of feature extraction in computer vision applications. SOLPP and NL-SOLPP achieved higher recognition efficiency than orthogonal LPP [100] and uncorrelated LPP [111].

Traditional LPP cannot be implemented using 2D image vectors because of the singularity matrix issue. Chen et al. [112] proposed 2D-LPP to enhance the image recognition using 2D image matrices instead of 1D vector. 2D-LPP saves local information and helps to detect an intrinsic manifold structure of the image. They showed 2D-LPP achieved higher recognition rate and better performance than LPP, 2D-PCA, and 2D-LDA when applied for the same number of dimensions. In another work, Xu et al. [97] suggested an improvement in 2D-LLP method for face recognition. 2D-LPP [112] is an unsupervised method which is unable to use discriminative information of sparse data. Wan et al. [113] proposed Sparse 2D Discriminant LPP (S2DD-LPP) based on 2D Discriminant LPP (2DD-LPP) and elastic net regression. They showed S2DDLPP has improved performance when compared with other methods (2D-PCA, 2D-PCA-L1, 2D-LDA, 2DD-LPP and 2DD-LPP-L1) for the ORL, Yale, and FERET face data sets.

2.1.5. Independent component analysis

Independent Component Analysis (ICA) is an unsupervised LDRT widely used for exploration of multi-channel data [114]. ICA is used to model data as a linear mixture of non-Gaussian independent source. ICA is a multi-variate approach used to perform linear transformation in such a way that the resulting components are uncorrelated and independent. Some applications of ICA include data analysis and compression, Bayesian detection, source localization, and blend and mixed source separation and identification [114].

ICA is a process of extracting independent components from the linear transformations of the original data. For this purpose, maximum likelihood [115] and minimization of mutual information in between components can be used [116]. Assume that the observed data $\boldsymbol{Y} = (y_1, y_2, \cdots, y_m)^T$ is composed of using linear transformation of $m \times p$ matrix $\boldsymbol{T}$ and non-Gaussian component vector $\boldsymbol{s}_i = (s_1, s_2, \cdots, s_n)^T$.

$$\boldsymbol{Y}_i = \boldsymbol{T}\boldsymbol{s}_i \tag{5}$$

ICA finds a linear mapping $\boldsymbol{W}$ of the source vector $\boldsymbol{s}_i$ such that each component of an estimate $\boldsymbol{v}$ is as independent as possible [117].

$$\boldsymbol{v}_i = \boldsymbol{W}\boldsymbol{Y}_i = \boldsymbol{W}\boldsymbol{T}\boldsymbol{s}_i \tag{6}$$

ICA uses an objective function to measure the degree of dependence between the components. Many algorithms have been proposed to estimate independent components based on different objective functions [118]. Hyvarinen [119] suggested a Fast fixed point ICA (FastICA) for separating complex values and linearly mixed source signals. FastICA can be used for Blend Source Separation (BSS) and feature extraction. BSS is being used in applications of biomedical, signal processing, finance, communication, remote sensing, and many other [120,121]. Akkalkotkar and Brown [116] introduced Mixed ICA/PCA through Reproducibility Stability (MIPReSt) approach that uses an iterative estimation method to rank different sources. It is used to determine the dimensions of non-Gaussian subspaces from mixture of data.

Yang et al. [122] applied the functions of Ranking and Averaging ICA by Reproducibility (RAICAR) to address the issues faced by spatial ICA for functional Magnetic Resonance Imaging (fMRI). When the signal mixture contains both Gaussian and non-Gaussian sources, Gaussian sources cannot be recovered by ICA and influence the estimate of non-Gaussian sources. Radüntz et al. [123] used IICA-based feature extraction method for automatic EEG artifact elimination. Capola ICA (CICA) is based on Hoeffding measure of dependence for classification of time series data [118]. temporal ICA (tICA) separates global noise and useful global signals when capturing fMRI data [124].

To predict variations in stock market, Ince and Trafalis [125] proposed a mixed method that combines the ICA and kernel methods. In this approach, ICA was used to choose important factors after defining inputs and kernel approaches were applied to forecast the variations and directions of the stock market.

Traditional clustering methods usually ignore the issues like lack of data normality and small temporal observations. Nascimento et al. [126] proposed a hybrid approach that combines ICA and hierarchical clustering called ICAclust. ICAclust does not require to define the number of clusters in advance. For temporal gene expression data, ICAclust has shown the better result than traditional k-mean clustering [126].

Other variants of ICA include Probabilistic ICA (PICA) for fMRI [127], Faster ICA under orthogonal constraint [128], Sparse Gaussian ICA (SGICA) [129] and Super Gaussian BSS using Fast-ICA with ChebyshevPade approximation [121].

2.1.6. Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a supervised LDRT which applies the linear combination of features as a linear classifier for DR [130–134]. LDA captures only global geometrical structure information and ignores the geometrical variation of local data points for the same class. Some of the application of LDA are face recognition [135–138], early detection of diseases [139], text recognition [137], person re-identification [140], hand movement classification [141], automatic diagnosis of machine operations [142], motor imagery EEG [143], and ground water redox conditions [144].

The aim of LDA is to find a linear transformation matrix $\boldsymbol{W} \in \mathbb{R}^{m\times k}(k << p)$ to map high dimensional data $\boldsymbol{Y} \in \mathbb{R}^{m\times p}$ into lower dimensional data $\boldsymbol{Z} \in \mathbb{R}^{m\times k}$.

$$\boldsymbol{Z} = \boldsymbol{W}^T\boldsymbol{Y}. \tag{7}$$

Optimal projection matrix can be computed using Eq. 8 [145].

$$min\ Tr\left(\frac{\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W}}\right) \tag{8}$$

where *Tr*(.) is the trace of the matrix. $\boldsymbol{S}_w$ and $\boldsymbol{S}_t$ can be computed as

$$\boldsymbol{S}_w = \frac{1}{m}\sum_{1}^{c} m_i(\overline{\boldsymbol{y}_i} - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y}_i} - \overline{\boldsymbol{y}})^T \tag{9}$$

$$\boldsymbol{S}_t = \frac{1}{m}\sum_{1}^{c}\sum_{1}^{m_i}(\overline{\boldsymbol{y}_{ij}} - \overline{\boldsymbol{y}_i})(\overline{\boldsymbol{y}_{ij}} - \overline{\boldsymbol{y}_i})^T \tag{10}$$

Here, $\overline{\boldsymbol{y}}$ is the mean of all the samples. Zhang et al. [146] proposed Local Intra-class Geometrical Variation Preserving LDA (LIPLDA) that captures local information. LIPLDA performed well and improved recognition accuracy for both global and local geometrical information. Liu et al. [147] proposed Local LDA based on k-Nearest Neighbor (kNN) that uses an affinity matrix which assigns weight according to the importance of sample structural information.

LDA cannot perform well for non-Gaussian and small sample size data. To address these issues, Subclass Discriminant Analysis (SDA) [148] and Mixture Subclass DA (MSDA) [149] has been developed. Moreover, Ran et al. [138] and Xin et al. [18] proposed Generalized EDA (GEDA) and Semi-supervised Regularized Discriminant Analysis (SRDA) to deal with the small sample size issue, respectively.

To reduce the overlap between models of sub-classes, Wan et al. [150] proposed Separability-oriented SDA (SSDA) that uses hierarchical clustering for dividing a class into separability-oriented criteria. Ye et al. [151] proposed 2D-LDA method to overcome the singularity issue implicitly and achieved better recognition accuracy. Recent advancements showed 2D-LDA as a successful matrix based DR method. However, 2D-LDA cannot deal with singularity issue. To overcome the singularity issue, Li et al. [152] proposed a Generalized Lp-norm 2D-LDA (G2D-LDA) framework with regularization. G2D-LDA achieved better generalization performance and solve the singularity problem. To accomplish better classification, Baudat and Anouar [153] introduced a Generalized Discriminant Analysis (GDA) using the kernel functionality.

2.1.7. Projection pursuit

Projection Pursuit (PP) introduced by Friedman and Turkey [154] is an unsupervised LDRT [154]. PP is a classical DRT which has been widely used for exploratory data analysis. PP is a non-parametric technique that finds low dimensional projections and explores interesting patterns for analysis [155].

The objective of PP is to find k dimensional projection $\boldsymbol{A} = [a_1, a_2, \cdots, a_k] \in \mathbb{R}^{p\times k}(k < p)$ so that the projected data maximize the predefined objective function δ called the projection pursuit index [154,156]. The projection pursuit index measures the degree of interestingness of the projected data. Moreover, PP index is flexible for various pattern recognition tasks such as cluster analysis and classification [156].

$$\text{argmax}_A\,\delta(\boldsymbol{Y}\boldsymbol{A}) \tag{11}$$

where $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$. PP decomposes the problem into a sequence of k optimization problems. Each optimization problem computes one base in $\boldsymbol{A}$. The first base a_1 is found by searching p dimensional unit length vector. The projected data maximizes the one-dimension PP index. Then, PP tries to remove all the information in that direction from the original data in order to avoid finding the same projection direction in subsequent iterations. The process is repeated until all k bases are computed [156]. Popular methods for PP optimization are gradient techniques [157,158], Newton-Raphson menthod [159,160], random search [158,161,162], genetic algorithm [163], simulated annealing [164], and particle swarm optimization [165]. The PP index can be used for supervised analysis [164,166–168], clustering analysis [161,162,169,170], and regression analysis [171,172].

Intrator introduced a hybrid classification and regression approach using unsupervised learning exploratory PP and supervised learning PP regression to reduce cost and complexity issue [173]. Exploratory PP combines a collection of data analytic techniques for low dimensional representation [174]. A PP based method was developed for the detection of multiple outliers [175]. Bolton and Krzanowski applied PP clustering for low dimensional representation of data [176]. Bingham and Mannila used Random Projection (RP) for DR of image and text data [177]. Experimental results indicate that RP was computationally less expensive when compared with other methods and was not affected by the curse of dimensionality [177].

Based on PP method, Lee et al. proposed a tree-based PP algorithm for classification. The key benefit of the proposed approach was to find correlation between features. It offers visual representation of differences found in groups in the form of 1D representation that provides help for interpretation of results [178]. Espezua et al. introduced a PP framework for the reduction of HDD with small sample size. The proposed framework comprises of compaction and PP phases [156].

To solve issues of non-linearity and HDD, Projection Pursuits Dynamic Cluster (PPDC) based on memetic algorithm was proposed [179]. The proposed approach was helpful to overcome the linear constraint of classical PP and to reduce the dimensionality of data. Projection Pursuits Random Forest (PPRF) algorithm can be used to solve the classification problems [180]. In this method, a tree is developed by splitting linear combinations of randomly selected features. PPRF performed better than Random Forest (RF) when classes were separated using linear combination of features or in situation when correlation between features increases.

2.2. Non-Linear dimensionality reduction techniques

Over the past few decades, a variety of NonLinear DRTs (NLDRTs) have been developed to work with applications having complex nonlinear structures [20,181,182]. To model relations present in the data in a nonlinear manner, kernel methods, kernel methods also known as "Kernel Trick" can be used [183]. Kernel trick avoids explicit mapping to learn a nonlinear function. Moreover, graph based approaches and manifold learning [184] NLDRTs have also been discussed.

2.2.1. Kernel principal component analysis

Kernel Principal Component Analysis (KPCA) introduced by Schölkopf et al. [185] is an extension of conventional PCA to work with HD feature space using kernel method. Instead of calculating the covariance of matrix, KPCA computes the principal eigen vectors of the kernel matrix. Kernel property makes PCA suitable for nonlinear mapping [11,183]. KPCA can extract nonlinear principal components using less computation power. KPCA offers good encoding for data having nonlinear manifold [186].

KPCA transforms the input data $\boldsymbol{Y}$ from original input space to kernel space for each data point using nonlinear transformation. Inner product of new feature vectors are used to form a kernel matrix K. Then, PCA is used on the centralized K to estimate the covariance matrix of the new feature vectors. Popular kernels include Radial, Gaussian, Polynomial, and Hyperbolic tangent [35]. Radial basis kernel with bandwidth b can be computed as [187]:

$$K(y_i, y_j) = exp\left(-\frac{\left\|y_i - y_j\right\|^2}{2b^2}\right). \quad (12)$$

Block Adaptive KPCA (BAKPCA) has been developed to dynamically and non-iteratively add new blocks and remove old blocks of data. BAKPCA showed improvement in signal processing and process monitoring [188]. To monitor the nonlinear dynamic processes, a Dynamic KPCA was introduced that achieved higher accuracy with minimum delay [189]. Greedy KPCA enhances the efficiency of SVM classifier [190]. Greedy KPCA is not suitable for denoising. To overcome the limitations of KPCA, an incremental KPCA was proposed to improve computation speed and storage utilization [191].

Washizawa [192] introduced Subset KPCA (SKPCA) to reduce the computational complexities of KPCA for DR and classification. To deal with outliers and to avoid wrong classification results, Robust KPCA has been proposed and applied for protein classification [193]. Zhang and Ma [194] proposed a Multi-Scale KPCA (MSKPCA) that developed a fault diagnostic method for nonlinear process monitoring. discriminative PCA (dPCA) can be used for discriminative analysis of multiple datasets [195]. dPCA was applied on applied health data, sensor data, and face images dataset.

2.2.2. Multidimensional scaling

Multidimensional Scaling (MDS) introduced by Kruskal and Wish [196] is an unsupervised NLDRT which aims to preserve a measure of similarity (dissimilarity or distance) between pairs of data points. MDS preserves an ideal pair wise distance in low dimensional space globally [196,197]. MDS has been used for exploratory analysis, multivariate analysis, and visualization. MDS works well when input distance matrix embeds the elements in d dimensional space such that the pairwise distances are preserved in the embedded space. Transformation is based on optimization of stress function, which is a sum of square errors between the dissimilarities and their corresponding embedding inter-vector distances [198]. Saeed et al. presented a brief review of MDS methods used in real world applications [199].

Consider a distance matrix $\boldsymbol{D}$ which contains the distance between data points of input data $\boldsymbol{Y}$. The objective of MDS is to find the low dimensional coordinates of each data point. The inter-point distances d'_{ij} of theses data points should be close to d_{ij}. MDS can be formulated as an optimization problem.

$$\delta(\boldsymbol{Z}) = min\left(\sum_{i=1}^{m}\sum_{j=1}^{m}(d_{ij} - d'_{ij})^2\right) \quad (13)$$

The closeness is measured by stress function which is calculated by sum-of-squares error. One of the commonly used stress function is:

$$Stress = \frac{\sum_{i<j}(d_{ij} - d'_{ij})^2}{\sum_{i<j}(d'_{ij})^2} \quad (14)$$

Traditional approaches to solve MDS were vulnerable to outliers. Mandanas and Kotropoulos proposed a unified framework that preserves MDS as maximization correntropy criteria using half-quadratic optimization for multiplicative and additive form [200]. In this way, MDS can manage dissimilarity matrix in the presence of outliers. The reason is correntropy criteria is closely related to M-estimators. Ma et al. introduced a neighbor preserving DR algorithm known as Localized MDS with BFS (LMB) to create low dimensional representation of data having latent manifold structure. LMB was applied on local neighborhood of data to take compact representation of neighborhood to get global reduction [201].

Chen and Buja introduced Local MDS (LMDS) that construct global structure using local information and was applied for drawing graphs and proximity analysis [202]. They have proposed solution to select tuning parameters and to identify the local embedding issue. In another study, Mohamed et al. proposed LMDS for 3D non-rigid shape retrieval [203].

MDS suffers a breakdown when the noise level increases. The reason is MDS depends on the number of dimensions and noise level. Peterfreund and Gavish introduced an extremely simple variant of MDS called MDS+ [204]. MDS+ acts as a unique asymptotically optimal shrinkage function. MDS+ was applied to derive shrinkage nonlinearity from the eigen values of the MDS similarity matrix. Furthermore, MDS+ overcomes the issue of external estimation for embedding dimensions. It calculates the optimal number of lower dimensions into which the data can be embedded. Blouvshtein and Cohen-Or suggested an outlier detection mechanism using Robust MDS (RMDS) based on geometric reasoning [198]. Hanley et al. proposed MDS-T that was an analytic technique used for the analysis of psychological data [205].

2.2.3. Isomap

Isomap is one of the most popular NLDRT used to find an intrinsic structure of data from nonlinear manifold learning. The distinguishing property of Isomap is to get the lower dimension representation of data, while preserving the geodesic distance [189,206,207]. Geodesic distance indicates the distance between two points in manifold. Isomap computes the geodesic distances between data points using a neighborhood graph. Each data point is connected with its k nearest neighbors. The shortest distance between two points indicate good estimate. One can obtain low dimensional representation of data points by applying MDS on resulting matrix [1]. Isomap has shown a remarkable performance for nonlinear dimensionality reduction in various research domains [207–211].

Isomap identifies the neighbors on the manifold $\boldsymbol{M}$ based on pairwise Euclidean distance $\boldsymbol{D}_Y$. Then, neighbors are stored in weighted matrix $\boldsymbol{D}_G$ which contains the distances between neighbors. Isomap computes the pairwise distance geodesic distance $\boldsymbol{D}_M$ between all pairs of data points on the manifold $\boldsymbol{M}$ using shortest path algorithms. The low dimensional representation $\boldsymbol{Z}$ of high dimensional data $\boldsymbol{X}$ can be computed using classical scaling on pairwise geodesic matrix. The vectors $\boldsymbol{z}_i$ are selected so that cost function $\boldsymbol{E}$ is minimized.

$$\boldsymbol{E} = \left\|\sigma(\boldsymbol{D}_G - \boldsymbol{D}_Z)\right\|_F \quad (15)$$

where function σ converts distances to inner products and can be defined as:

$$\sigma(\boldsymbol{D}) = -\frac{HSH}{2} \quad (16)$$

Here, $\boldsymbol{S}$ and $\boldsymbol{H}$ are matrix of squared distances and centering matrix, respectively.

Monitoring and identification of uncertainties and anomalous events from the crowd data is a challenging task. Moreover, feature management became computationally expensive for a large number of feature sets. Isomap was used to detect irregularities from large-scale real time video analytics efficiently [209].

Jiang et al. [212] proposed a Compressive Sensing method with Gaussian Mixture Random Matrix (CS-GMRM) to transform HD vector data into lower dimensions. CS-GMRM eliminates the need for the training process and preserves the matrix information of the original vector space. Najafi et al. [213] introduced a path based Isomap method to enhance the time and memory complexity. This approach uses the geodesic path to obtain the low dimensional embedding. Shi et al. [214] presented Landmark Isomap (L-Isomap) to enhance the scalability of Isomap. Furthermore, Suganya et al. [215] combined Isomap with SVM to achieve better results when compared with KPCA.

Li et al. [216] presented a parallel implementation of Isomap using Graphics Processing Unit (GPU). They have used Isomap for visualization and DR of Hyper Spectral Images (HSIs). For feature extraction and DR, a Sparse and Low-Rank Isometric Linear Embedding (SLRNILE) can be used to enhance the classification accuracy of HSI [14]. Manifold learning Isomap was successfully applied for speech summarization [210], prediction of the good and bad condition of urban road traffic [217], and the identification of cracks in the material [218]. Zheng et al. [211] introduced a sequential simulation based framework for generating super resolution, remote sensing images using Isomap and Multiple Point Statistics (MPS).

2.2.4. Locally linear embedding

Locally Linear Embedding (LLE) is an unsupervised NLDRT which aims preserve only local properties of data. LLE explores the local features of the manifold for data points as linear combination of reconstruct weight. LLE shows different data points using graph representation. LLE uses eigen vector method for nonlinear DR process and can successfully work for the identification of manifold structures [219]. Performance of LLE can be enhanced using multiple linearly independent local weight vectors for each neighborhood.

LLE identifies the neighbors of each data point $\boldsymbol{y}_i$. The neighbors can be identified using different algorithms. Then, it computes the weight matrix $\boldsymbol{W}$ of data $\boldsymbol{Y}$ from neighbors. The objective is to minimize the cost function.

$$\sum_{i=1}^{m} \left\| \boldsymbol{y}_i - \sum_{j=1}^{m} \boldsymbol{w}_{ij} \boldsymbol{y}_j \right\|^2 \tag{17}$$

Then low dimensional embedding is made for vector z_i so that cost function is minized.

$$\sum_{i=1}^{m} \left\| \boldsymbol{z}_i - \sum_{j=1}^{m} \boldsymbol{w}_{ij} \boldsymbol{z}_j \right\|^2 \tag{18}$$

Abdel-Mannan et al. [220] proposed Hessian Locally Linear Embedding (HLLE) based on Incremental LEE (ILEE) for dynamically adding new data and performing DR while preserving significant features of the original data set. Zhang and Wang [221] introduced Modified LLE (MLLE) using multiple weights. Supervised LLE and semi-supervised LLE was introduced by Kouropteva et al. [222] and Zhang and Chau [223], respectively, to enhance plant classification based on leave images. Hettiarachchi and Peters [224] suggested a Multiple Manifold LLE (MM-LLE) approach to learn multiple manifolds for multiple classes. For instance, neighborhood selection based on supervised learning in individual manifold learning. MM-LLE has performed well for object classification and recognition.

2.2.5. Self-Organizing map

Self-Organizing Map (SOM) introduced by Teuvo Kohonen in 1981-82 is an unsupervised NLDRT based on cognitive learning [225,226]. SOMs are versatile for analyzing nonlinear projections, multivariate and complex data sets [227]. For a data set with mixed type features, the dissimilarity of two instances can be measured on numerical and categorical features separately. The dissimilarity can be computed using the combination of squared Euclidean distance on numeric features and number of mismatches on categorical features. Numeric features are usually normalized before distance calculation to ensure that all features have equal influence on distance [228].

SOM groups similar data instances into 2D or 3D lattices called output map. Different data instances appear separately on the output map. Only important input space properties can be extracted from that output map [229]. SOMs are similar to vector quantization. A SOM consists of a map of neurons $n_1, n_2, \cdots, n_j$, each equipped with a weight vector $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_j$ on a p-dimensional map. SOM compares the distance to the weights of each neuron for each $\boldsymbol{y}_i$. The neuron with lowest distance to its weight is chosen and plotted in direction of $\boldsymbol{y}_i$.

$$\delta' = \min_{1 \le k \le j} \|\boldsymbol{y}_i - \boldsymbol{w}_k\|^2 \tag{19}$$

Real life applications of SOM include automatic organization of a massive collection of documents [230], intrusion detection [231], development of risk-based prioritization for stagnation [229], recognition of protein folds [232], market data analysis [233], the classification of fMRI [4], removal of noise from 6D synthetic spectral image [234], and predication rate for weather and crop production [235].

Chaudhary et al. [236] developed Community SOM (CSOM) to enhance the overall learning process. Isa et al. [237] proposed a hybrid approach based on probability distribution and SOM with Naive Bayes classifier to improve the classification accuracy for text documents. SOM based clustering can achieve better classification accuracy for huge volume of documents. In classical SOM, there was need to predefine the map topology and it became difficult to find the hierarchical relationships in data. Yang et al. [238] developed a novel algorithm based on several text mining approaches to enhance the efficiency of SOM and an automatic organization of text documents.

Traditional SOM uses Mean Square Error (MSE) as objective function that affected the performance of SOM for noisy data. A correntropy based approach was used in place of MSE that enhances the efficiency of SOM in the presence of noise. Weighted-SOM (W-SOM) was employed for accurate crop and weather prediction, which is the combination of both SOM and learning vector quantization. The prediction accuracy was enhanced by minimizing the Within Class Error (WCE) among clusters [235]. A multistage Visual Analytical (VA) method with SOM flow can be used to iteratively refine the cluster to analyze time series data [239]. Liu et al. [240] used SOM for ecological informatics to obtain useful information from the behavioral data. SOMs of single cell spectra enables characterizing the heterogeneity between and within cell populations that lie along distinct pathways [241].

2.2.6. Learning vector quantization

Learning Vector Quantization (LVQ) is a supervised NLDRT which is similar to SOM. LVQ is a competitive based neural network introduced by Kohonen [226]. Classification can be performed using distance between input vectors. The class borders enhance the classification accuracy of LVQ.

LVQ consists of an input layer, a single LVQ layer and an output layer. The output layer has nodes equal to distinct classes. LVQ parameterizes the classification using C classes prototypes. These are chosen as representatives of the respective classes. They are characterized by their location in feature space $\boldsymbol{s}_i$ and the respective class labels $c(\boldsymbol{s}_i) \in 1, 2, \cdots, C$. For given distance measure $d'(\boldsymbol{s}, \boldsymbol{y})$, data point is assigned to a class $c(\boldsymbol{s}_i)$ with $d'(\boldsymbol{s}_i, \boldsymbol{y}) \le d'(\boldsymbol{s}_j, \boldsymbol{y})$ for all $j \neq i$. Distance measures can also be specified using a matrix.

LVQ has gained popularity in different application areas due to its classification accuracy [242]. For feature extraction from still pictures of human faces, combination of Gabor filter and LVQ approach can be applied to recognize different facial expression [243]. Fitria et al. [244] employed LVQ based artificial neural network classifiers for several kinds of signal processing methods to recognize and classify arrhythmia from the ECG signals. Hammer et al. [245] proposed a general framework to combine different variants of LVQ to improve the classification accuracy for different types of data. PCA and LVQ were combined by Hu et al. [246] for classification of mobile learning (mlearning) strategies. Selection of an appropriate dissimilarity measure for a classification problem has become a time consuming and challenging task. To solve this issue, an interpretative framework using median variation of the Generalized LVQ (GLVQ) based on dissimilarities has been developed by Nebel and Kaden [247] to enhance the classification accuracy.

LVQ and Robust Soft LVQ (RSLVQ) can be applied only on vector data. It cannot work well with complex data sets having pairwise relationships. Hofmann et al. [248] introduced a Kernel based RSLVQ approach that can handle more complex non vector data using general Gram matrix. However, downside of this approach is high computational complexity due to full Gram matrix. Schleif [249] presented a hybrid approach to extract random Fourier features for an LVQ matrix. This technique provides reasonably smaller and discriminative feature set.

2.2.7. T-Stochastic neighbor embedding

t-Stochastic Neighbor Embedding (t-SNE) introduced by Hinton and Roweis [250] is an unsupervised NLDRT used to reduce HDD in low dimensional space. t-SNE is based on matching distances between distributions [250]. t-SNE is non-parametric and well suited for visualization of data sets having nonlinear structures [251]. It captures most of the local structure of HDD by revealing a global structure and has functionality to work with manifold learning. For low dimensional transformation, t-SNE relies on conditional probability.

For input data $\boldsymbol{Y}$, SNE computes the conditional probabilities p_{ij}. The goal of t-SNE is to have k dimensional map that reflects the similarities q_{ij} between two points $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$.

$$p_{j|i} = \frac{exp(-\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|^2)/2\sigma_i^2}{\sum_{k\neq i} exp(-\left\|\boldsymbol{y}_i - \boldsymbol{y}_k\right\|^2)/2\sigma_i^2} \tag{20}$$

$$q_{ij} = \frac{1 + \left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2)^{-1}}{\sum_{k\neq i}(1 + \left\|\boldsymbol{z}_i - \boldsymbol{z}_k\right\|^2)^{-1}} \tag{21}$$

t-SNE a distribution in the embedding space.

$$E_{t-SNE} = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{22}$$

where

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2m} \tag{23}$$

Konstorum et al. conducted a comparative analysis of four DRTs, i.e., PCA, Isomap, t-SNE, and Diffusion Maps by implementing them on a benchmark mass cytometry data set [252]. The results of these reductions were compared for computation time, residual variance, neighborhood proportion error (NPE), and for 2D visualization. It was found that t-SNE and Diffusion Maps were the two most effective methods for preserving local distance relationships among cells and providing informative visualizations.

Xie et al. proposed a multi-view Stochastic Neighbor Embedding (m-SNE) to systematically combine the structures (features) from different sources for a unified representation and future processing [253]. Experimental results have shown the effectiveness of m-SNE for scene recognition and data visualization. Platzer applied t-SNE for DR and visualization of large biological data [254]. Gisbrecht et al. introduced a kernel-based t-SNE to enhance the functionality of classical non-parametric t-SNE. This approach preserves the functionality of the classical t-SNE and makes it suitable for parametric NLDRTs [255].

Example

This section analyzes and visualizes different DRTs. In this study, we used ElectroCardioGram (ECG) signal for heartbeat data derived from PhysioNet MIT-BIH Arrhythmia database [256]. This dataset has 188 dimensions and 1094446 observations. This dataset has class imbalance and other problems. For this purpose, data pre-processing was performed. Uniform class distribution was produced by re-sampling with bootstrapping based method to reduce the bias. Then, we used random 70/30 train/test data split.

In this work, we have implemented different linear and nonlinear techniques for ECG data and visualized the results using scatter plot(Fig. 1). We have considered only first four components for each technique. Each component is shown by different color. These techniques were compared using computational time. Nonlinear techniques like Isomap, MDS, and t-SNE have worst execution time when compared with other techniques. Many nonlinear DRTs have shown excellent ability to perform dimension reduction. However, they are vulnerable to noise [257]. The need for data sampling can have an adverse impact on their future practical implementations. Overall, most methods were able to represent the general features of the ECG data set. Visualizations generated by LLE, SOM, and t-SNE showed the most apparent divergence from other methods.

3. Comparative study

This section presents core functionalities that different DRTs and their variants offer. For this purpose, this section has been divided into different subsections. Next subsection summarizes the different properties of DRTs. This study also explores the positive and negative aspects of each technique. Then issues and challenges of HDD are discussed. Afterwards, application areas where DRTs can be used are presented.

3.1. Properties of DRTs

The following findings have been identified from the studied literature. Parameters used for the comparative study have been taken from Table 1. The parameter legends used in Table 2 are: Learning Type (S=Supervised, US=Unsupervised, SS=Semi-supervised), Linearity (L=Linear, Nl=Non-Linear), Data distribution (Gu=Gaussian, NG=Non Gaussian), Parametric (P=Parametric, NP=Non-Parametric), Span (Lo=Local, Gl=Global, M=Manifold (Local+ Global)), Data Type (V=Vector, Di=Distance), Deterministic Property (De=Deterministic, ND=Non-Deterministic), and Objectives. The "+" symbol indicates that base DRT naturally hold this property and can perform that particular task. The "*" symbol indicates that variant of the base DRT provide the mentioned property.

3.2. Common issues and challenges of HDD and solutions using DRTs

Selection of an appropriate DRT can decrease response time and enhance the efficiency of the algorithm. DRTs can reduce computation time and cost by eliminating redundant features. HD and a huge volume of data may lead toward ML model over-fitting. DRTs reduces the dimensionality of data by eliminating irrelevant features that make it feasible to train data for ML models.

Different issues of HDD have been identified based on articles studied. Normally, outliers have values much higher than all other values in data set or have a distance that lies beyond the normal distribution. It becomes difficult to analyze a particular set of values in the presence of outliers [40]. Different DRTs have been developed to solve this problem for different types of data.

Another big issue that can affect the original data is noise. Sometimes, it becomes difficult to identify noise from real data. DLPP [105] offers facility to remove noise from image data. Sparsity means that most of data values are blank. If there are too many sparse values in data, then it becomes difficult to get accurate results after processing the data. Some version of PCA can work with sparse data. Moreover, unknown subspaces are difficult to identify and manage with many DRTs. Now variants of some DRTs can solve this issue. Another common issue of HDD is curse of dimensionality [23]. It covers all phenomena that appear with HDD and most of the time decreases the performance of the learning algorithm.

The small sample size issue raises when the number of dimensions exceeds than the number of observations. Processing and analysis of such data becomes difficult. Several DRTs are available to solve small sample size problem for different types of data. Multivariate data is an essential part of HDD. DRTs can make effective use of multivariate data by extracting valuable relevant information. Common issue with multivariate data is missing values. Another issue is the absence of class labels [258]. Singularity issue affects the results of analysis when data cannot be elaborated in 1D or vector form.

The data from environment or other sources can be mixed with original data. Process of separating irrelevant features from useful features is known as BSS. Many DRTs offer this facility. However, the

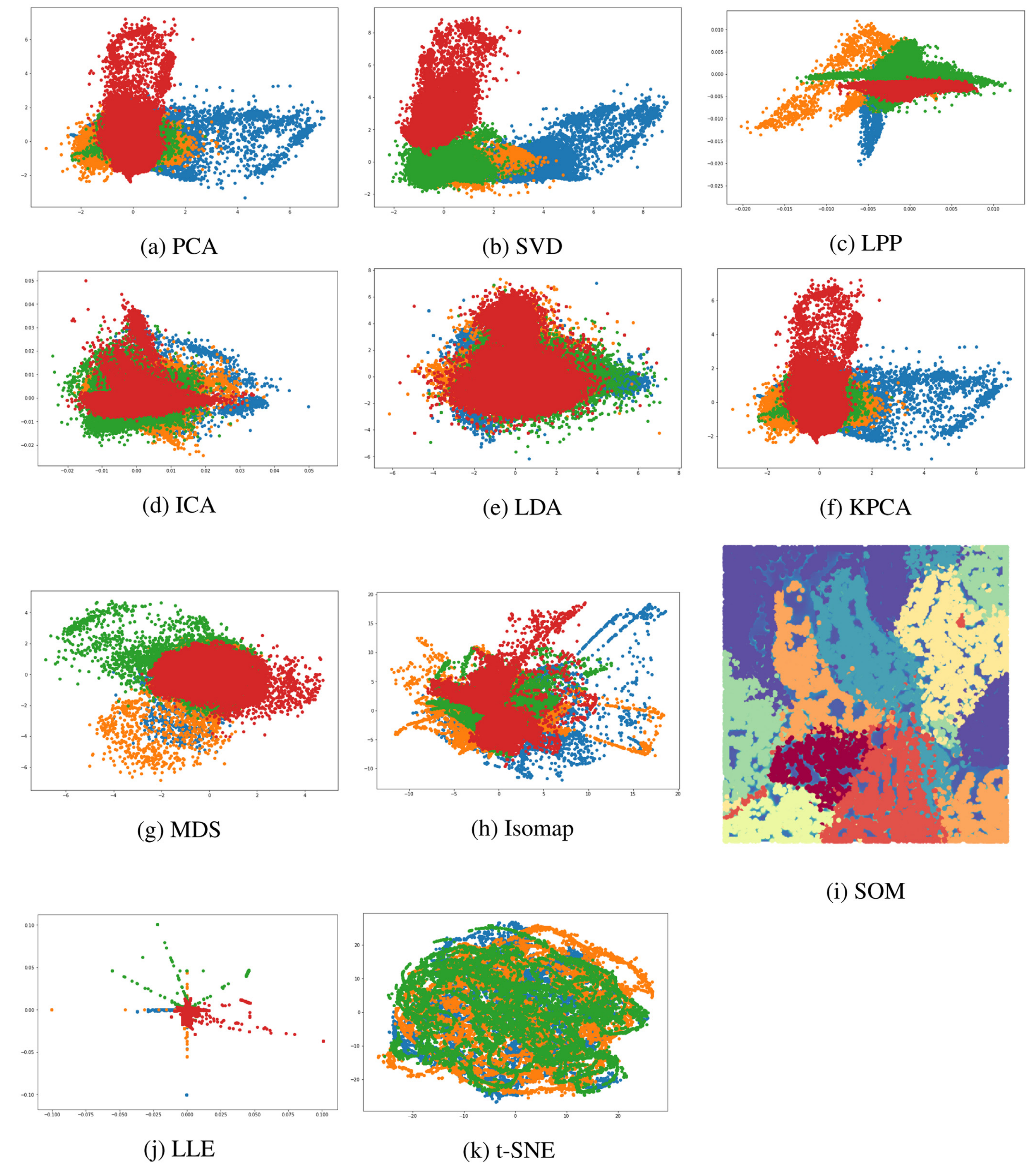

Fig. 1. Comparison of different linear and nonlinear techniques for ECG data.

removal of useful features along with irrelevant features can lead towards wrong classification. For this reason, BSS issue can severely affect the classification accuracy. This issue is commonly found in image and signal data [121]. Table 3 presents a list of common issues found in HDD and available solutions by applying DRTs and variants of DRTs.

3.3. Available DRTs and different application areas

According to this study, various DRTs and their variants have been developed to solve different types of issues. Each technique has advantages and disadvantages to solve different problems. The application area and type of data for which a specific technique was developed also

Table 2
A Comparative Study of DRTs and their Variants.

DRTs	Learning Type			Linearity		Distribution		Parametric		Span			Data Type		Deterministic		Objectives
	S	US	SS	Le	Nl	Gu	NG	P	Np	Lo	Gl	M	V	Di	De	Nd	
PCA	*	+		+	*	+	*	+		+		*	+			+	Preserve Variance
SVD		+		+		+		+	*				+				Optimal DR
LSI		+		+			+	+		+			*	+		+	Classification Accuracy
PP	*	+	*	+			+			+		*	*	*			Manifold Learning
ICA		+		+		*	+		+			+					Explore Multi-channel Data
LPP	*	+		+	*	+	*			+		+	*	*			Manifold Learning
LDA	+		*	+	*	+	*	+		*	+	*	+		+		Classification Performance
KPCA	*	+			*	*				*	*	+	+			+	Manifold Extraction
MDS	*	+			+				+	*	+	*	+	+	+		Manifold Extraction
Isomap		+			+				+		+	+	+	+	+		Manifold Extraction
LLE		+							+	+			+		+		Manifold Extraction
SOM		+			+			+				*	+		+		Prediction Accuracy
LVQ	+				+			+		+	*		+	*	+		Classification Accuracy
t-SNE	*	+	*		+	*	+	*	+	*	*	*	*	+	*	+	Neighborhood Preservation

Table 3
Issues in HDD and solutions using DRTs.

Issues	DRTs	Reference
Outliers	ER-PCA, JSPCA, DLPP based on L1-norm, PP, Robust KPCA, MDS	[40], [54], [107], [175], [193]
Noise Removal	DLPP, DSDV, MDS+, Isomap, SOM	[105], [259], [204], [213], [234]
Sparsity	SPCA,JSPCA	[51], [54]
Unknown subspace	GPCA	[41]
Curse of dimensionality	GP-SPCA	[52]
Computation cost	RP-LSA, SOM	[260], [235]
Response Time	2D-PCA, RP-LSA	[42], [260]
Computation Time	ROB-PCA, GP-SPCA, Fast-ICA, RP-LSA, KPCA, Isomap	[39], [52], [260], [119], [191], [209]
Over-fitting	RPLSA	[91]
SSS	MEDLPP, DLPP, SRDA, GEDA	[18], [107], [106], [138]
Multi-variate	MFPCA	[47]
Singularity	2DLDA, G2DLDA, 2D-LPP	[112], [151], [152]
BSS	ICA, FastICA	[120], [121], [119]

Table 4
Utilization of different DRT variants in various application areas.

Application Areas	DRTs	References
Face and image recognition	SKLPP, MPCA, DLPP, 2D-LPP, S2DDLPP, ODLPP, DLPP/MMC	[104], [45], [105], [112], [113], [100], [108], [127]
Pattern recognition	KSLPDA, DLPP, LIPLDA, 2DLDA	[101], [105], [146], [151]
Enhance text classification	GALSF, RLSA	[93], [261]
Image/Video classification and clustering	GLPP, SOM	[110], [204], [239]
Gene expression data	ICAClust	[126]
Pattern classification	SRDA, RAD	[141], [138]
Visualization	MDS, SOM, t-SNE	[203], [251], [254]
Signal processing	IICA, FastICA, LVQ	[119], [67], [123], [262], [244]
Time Series	PCA+SOM	[233]

matters. Findings of Section 2 with respect to application areas are summarized in Table 4.

DRTs are available for various types of data such as image, signal, text, etc., Table 5 shows the data types for which DRTs have been applied. Data type information was collected from datasets used in different studies. The "+" symbol indicates that a particular DRT can work with this type of data. The "-" symbol indicates that DRT cannot be used for the mentioned data type. Blank means that no information is available in the studied literature.

4. Issues of dimensionality reduction techniques

Despite the above mentioned benefits, some limitations and issues are associated with DRTs. Moving from high dimensional to low dimensional data poses numerous challenges. Selection of a suitable method according to the type of data is a big issue which need to be addressed. It is essential to find a suitable mechanism to attain the highest level of accuracy when combining the output of several DRTs.

Another issue is identification of redundancy level as HDD can have many redundant features. It becomes difficult to identify the level of redundancy and remove it without affecting the performance in low dimensional mapping. Many DRTs offer this facility. However, some features may be redundant but important for analysis and decision making. One may get non-redundant and low dimensional data sets which are easy to manage after applying DRTs. There is a possibility that important information may be lost in low dimensional data set that were essential for analysis. In such a situation, selection of suitable of DRT can become a challenging task.

Many DRTs can enhance visualization of data using few dimensions [9]. Sometimes, important dimensions are not selected during the DR process and affect the results adversely. For example, a feature has lower value but has significant impact on the prediction accuracy, missing such a feature will lead toward a wrong decision. Most of the DRTs cannot work with input data directly and need pre-processing such as PCA and SOM. There is need to normalize data to get precise results. It is claimed that use of DRTs reduces computation time, but some DRTs

Table 5
Overview of supportive DRTs for different data type.

DRTs	Text	Image	Signals	Audio/Video	Time Series	Structured
PCA	+	+	+	+	+	+
SVD	+	+	+	+	+	+
LSA	+					
PP	+	+				
ICA	-	+	+	+	+	+
LPP	+	+	+	+	+	+
LDA	+	+	+			
KPCA		+	+	+	+	
MDS	+	+	+	+	+	+
Isomap	+	+	+	+		
LLE		+	+	+		
SOM	+	+	+	+	+	+
LVQ	+	+	+	+	+	+
t-SNE	+	+	+	+	+	+

(i.e., SVD and some nonlinear techniques) are computationally expensive.

A common problem that may rise during DR process is to fix how many features to select for analysis. Another important aspect to consider is to identify the nature of data and features that will be used for analysis. Finding the relevant and important features has become a problematic task. It entails domain, knowledge, and human expertise to extract most relevant features for future processing and selection of ML models for classification. Interpretation of results is another reason that makes the DR application infeasible for applications which requires interpretation. It is not possible to retain full information of HDD set after applying DRTs. It depends on the given data set and nature of the problem to identify the most relevant features for processing.

In some conditions, DR is not possible as most of the HD features are important for analysis. The reason is the features can be inter-related. Dealing with interdependency of variables is also a big issue which needs to be handled. many DRTs are available and helpful for taking only relevant features for analysis by removing noise. Level of noise also affects the performance of DRTs. The last aspect to consider is minor changes in input can affect the results. For example, values of particular variable can be intentionally set to very high or very low. This may lead to wrong classification or predictions. Similarly, selection of a single wrong feature can lead towards wrong results. For example, for weather forecast, region information is most important, but wrong region information will certainly give wrong information.

5. Discussion

Over the past few decades, DRTs have gained lots of attention for pattern recognition, computer vision, face and object recognition, motion detection, and classification. This comparative study covers a broader perspective about applications and types of data. We explored the functionalities of original DRTs and their available variants that can enhance the suitability of DRTs for linear and nonlinear applications.

DRTs have been studied by many researchers to solve different issues of HDD and data pre-processing [1]. Comparative study of DRTs has been presented in [5,11,12,15] but information about new variant of some DRTs has not been presented. In this study, we discussed seven linear and seven NLDRTs and their variants for text, image, and signal data. LDRTs and their variants have been used to manage different types of data structures and application areas[19]. Maaten and Hinton [251] reviewed different versions of t-SNE but it was limited to data visualization. They analyzed linear and non-linear DRTs for image data only. NLDRTs have been reviewed in [20]. A comparative analysis of four DRTs including PCA, Isomap, t-SNE, and Diffusion Maps by implementing them on a benchmark mass cytometry data set is presented in [252].

PCA and its different versions (i.e., robust PCA, sparse PCA) are still widely in use due to its simplicity and efficiency. The presence of outliers can adversely affects the performance of PCA. SVD is an unsupervised DR method suitable for sparse data but is computationally expensive for some applications. SVD has been successfully used for text summarization, image processing, classification and noise removal. LSI is an unsupervised vector-based technique frequently used for text summarization, categorization, IR, NLP, and topic modeling. PP is an unsupervised method that uses the non-Gaussian data and works efficiently even in the presence of outliers.

ICA originally works well with Gaussian data, but its variants have the capability to deal with Gaussian and non-Gaussian data. Potential applications of ICA are signal processing, BSS (fastICA), feature extraction, gene expression data, remote sensing, communication, and informatics. LLP is a graph-based technique which is most suitable for face recognition and image/video classification [6,7]. Presence of noise and outlier does not influence its performance as in PCA. Different versions of LLP (SKLLP, DLPP, ODLLP, GLPP, S2DDLPP) have enhanced its recognition capabilities and reduced noise. The singularity issue in a classical LDA is solved by a variant of LDA (G2DLDA). Another issue, small sample size has been solved using EDA, GEDA, and SRDA. LDA is being used for the text, image, medical, data for feature extraction, and pattern classification.

NLDRTs have been successfully used to enhance the efficiency of nonlinear transformations. KPCA, a variant of most widely used PCA, has been developed to deal with nonlinearities in data. These techniques are mostly used at pre-processing stage for feature extraction and DR. MDS is an unsupervised NLDRTs that uses topology structure and preserves a pair-wise scalar product. It is commonly used for DR, data exploration, and visualization. Classical MDS is sensitive to outliers. Isomap preserves the topological structure of data. LIsomap has enhanced the scalability. It has been successfully used for crowed monitoring. SLRNILE provides better classification accuracy but is computationally expensive.

LEE preserves the local structure of data in lower dimensions. LEE is suitable only for well sampled manifolds. SOM is an extension of locally linear PCA. SOM is an excellent tool for data exploration and visualization. It is suitable for multidimensional data found in images, text, time series, ecology, etc. LVQ uses labeled data and performs classification based on the distance between input vectors. t-SNE was used to visualize and extract features of HD using supervised methods [255]. MDS preserves distances and a t-SNE preserves the neighborhood structure of data points.

6. Conclusion

We conclude that most of the DR applications need human-machine collaboration to achieve classification and prediction accuracy. As more data became available, need for DRTs also increased to reduce uncertainty in decision making. Linear techniques use linear transformation and requires less computation power. Nonlinear techniques can have high computation time and cost in some situations, but have been successfully implemented for many complex application areas (i.e., biomedical, audio and video data). This study will be helpful for selecting suitable DRT according to type of data. In future research work, we will design algorithms to resolve different issues highlighted in this work.

Credit authorship contribution statement

Shaeela Ayesha: Conceptualization, Formal analysis, Methodology, Writing - original draft. **Muhammad Kashif Hanif:** Formal analysis, Supervision, Validation, Writing - review & editing. **Ramzan Talib:** Supervision, Validation.

Declaration of competing interest

The authors have declared no conflict of interest.

CRediT authorship contribution statement

Shaeela Ayesha: Conceptualization, Formal analysis, Methodology, Writing - original draft. **Muhammad Kashif Hanif:** Formal analysis, Supervision, Validation, Writing - review & editing. **Ramzan Talib:** Supervision, Validation.

References

[1] L. Gao, J. Song, X. Liu, J. Shao, J. Liu, J. Shao, Learning in high-dimensional multimedia data: the state of the art, Multimedia Syst. 23 (3) (2017) 303–313.
[2] D. Amaratunga, J. Cabrera, High-dimensional data, J. Natl. Sci. Found. 44 (1) (2016) 3–9.
[3] T. Ortner, P. Filzmoser, M. Rohm, C. Breiteneder, S. Brodinova, Guided projections for analyzing the structure of high-dimensional data, J. Comput. Graphical Stat. (2018) 1–34.
[4] S. Bahrami, M. Shamsi, A non-parametric approach for the activation detection of block design fmri simulated data using self-organizing maps and support vector machine, J. Med. Signals Sens. 7 (3) (2017) 153.
[5] B. Tang, M. Shepherd, E. Milios, M.I. Heywood, Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering, in: International Workshop on Feature Selection for Data Mining, volume 39, 2005, pp. 81–88.
[6] Y. Tang, R. Rose, A Study of Using Locality Preserving Projections for Feature Extraction in Speech Recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP., IEEE, 2008, pp. 1569–1572.
[7] C.-D. Chang, C.-C. Wang, B.C. Jiang, Singular value decomposition based feature extraction technique for physiological signal analysis, J. Med. Syst. 36 (3) (2012) 1769–1777.
[8] S. Ng, Principal component analysis to reduce dimension on digital image, Procedia Comput. Sci. 111 (2017) 113–119.
[9] P. Naik, M. Wedel, L. Bacon, A. Bodapati, E. Bradlow, W. Kamakura, J. Kreulen, P. Lenk, D.M. Madigan, A. Montgomery, Challenges and opportunities in high-dimensional choice data analyses, Mark. Lett. 19 (3–4) (2008) 201.
[10] W.K. Vong, A.T. Hendrickson, D.J. Navarro, A. Perfors, Do additional features help or hurt category learning? the curse of dimensionality in human learners, Cogn. Sci. 43 (3) (2019) e12724.
[11] L.V.D. Maaten, E. Postma, J.V. Herik, Dimensionality reduction: a comparative, J. Mach. Learn. Res. 10 (2009) 66–71.
[12] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[13] A. Cano, S. Ventura, K.J. Cios, Multi-objective genetic programming for feature extraction and data visualization, Soft Comput. 21 (8) (2017) 2069–2089.
[14] W. Sun, G. Yang, B. Du, L. Zhang, L. Zhang, A sparse and low-rank near-isometric linear embedding method for feature extraction in hyperspectral imagery classification, IEEE Trans. Geosci. Remote Sens. 55 (7) (2017) 4032–4046.
[15] I.K. Fodor, A survey of dimension reduction techniques, Lawrence Livermore National Lab., CA (US), 2002 Tech. rep..
[16] Z. Zhang, F. Yang, K. Xia, R. Yang, A supervised lpp algorithm and its application to face recognition [j], J. Electron. Inf. Technol. 3 (2008) 8.
[17] W. Liao, A. Pizurica, P. Scheunders, W. Philips, Y. Pi, Semisupervised local discriminant analysis for feature extraction in hyperspectral images, IEEE Trans. Geosci. Remote Sens. 51 (1) (2013) 184–198.
[18] Y. Xin, Q. Wu, Q. Zhao, Q. Wu, Semi-supervised Regularized Discriminant Analysis for Eeg-based Bci System, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2017, pp. 516–523.
[19] J.P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: survey, insights, and generalizations, J. Mach. Learn. Res. 16 (1) (2015) 2859–2900.
[20] X.L. Zhang, Nonlinear Dimensionality Reduction of Data by Deep Distributed Random Samplings, in: Asian Conference on Machine Learning, volume 2015, pp. 221–233.
[21] A. Gisbrecht, B. Hammer, Data visualization by nonlinear dimensionality reduction, Wiley Interdiscip. Rev.: Data Min.Knowl. Discov. 5 (2) (2015) 51–73.
[22] S. Buchala, N. Davey, T.M. Gale, R.J. Frank, Analysis of linear and nonlinear dimensionality reduction methods for gender classification of face images, Int. J. Syst. Sci. 36 (14) (2005) 931–942.
[23] M. Verleysen, D. François, The Curse of Dimensionality in Data Mining and Time Series Prediction, in: International Work-Conference on Artificial Neural Networks, Springer, 2005, pp. 758–770.
[24] L. Chen, Curse of Dimensionality, in: Encyclopedia of Database Systems, Springer, 2009, pp. 545–546.
[25] A. Juvonen, T. Sipola, T. Hämäläinen, Online anomaly detection using dimensionality reduction techniques for http log analysis, Comput. Netw. 91 (2015) 46–56.
[26] G. Kraemer, M. Reichstein, M.D. Mahecha, Dimred and coranking unifying dimensionality reduction in r, R J. 10 (1) (2018) 342–358.
[27] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, D. Filliat, State representation learning for control: an overview, Neural Netw. 108 (2018) 379–392.
[28] C. Meng, O.A. Zeleznik, G.G. Thallinger, B. Kuster, A.M. Gholami, A.C. Culhane, Dimension reduction techniques for the integrative analysis of multi-omics data, Brief. Bioinform. 17 (4) (2016) 628–641.
[29] K. Pearson, Liii. on lines and planes of closest fit to systems of points in space, London Edinburgh Dublin Philos.Mag. J. Sci. 2 (11) (1901) 559–572.
[30] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417.
[31] S. Deegalla, H. Boström, K. Walgama, Choice of Dimensionality Reduction Methods for Feature and Classifier Fusion with Nearest Neighbor Classifiers, in: 15th International Conference on Information Fusion (FUSION), IEEE, 2012, pp. 875–881.
[32] S. Ahmadkhani, P. Adibi, Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework, IET Comput. Vision 10 (3) (2016) 193–201.
[33] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.
[34] N.B. Erichson, P. Zheng, K. Manohar, S.L. Brunton, J.N. Kutz, A.Y. Aravkin, Sparse principal component analysis via variable projection. arXiv preprint arXiv:1804.00341.
[35] C. Ecse, Dimensionality reduction. pca. kernel pca, Lecture lides:COMP-652 and ECSE-608.
[36] L. Wiskott, Lecture notes on principal component analysis, 2013,
[37] N. Kambhatla, T.K. Leen, Dimension reduction by local principal component analysis, Neural Comput. 9 (7) (1997) 1493–1516.
[38] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, K. Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, et al., Robust principal component analysis for functional data, Test 8 (1) (1999) 1–73.
[39] M. Hubert, P.J. Rousseeuw, K.V. Branden, Robpca: a new approach to robust principal component analysis, Technometrics 47 (1) (2005) 64–79.
[40] S. Serneels, T. Verdonck, Principal component analysis for data containing outliers and missing elements, Comput. Stat. Data Anal. 52 (3) (2008) 1712–1727.
[41] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (gpca), IEEE Trans Pattern Anal Mach Intell 27 (12) (2005) 1945–1959.
[42] Q. Li, Z.-D. Qiu, D.-m. Sun, L.L. Liu, Online palmprint identification based on improved 2d pca, Dianzi Xuebao(Acta Electronica Sinica) 33 (10) (2005) 1886–1889.
[43] T. Wang, I.Y. Gu, P. Shi, Object Tracking Using Incremental 2D-pca Learning and Ml Estimation, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, IEEE, 2007, pp. I–933.
[44] C. Qiu, H. Ren, H. Zou, S. Zhou, Performance Comparison of Target Classification in Sar Images Based on Pca and 2D-pca Features, in: 2nd Asian-Pacific Conference on Synthetic Aperture Radar, IEEE, 2009, pp. 868–871.
[45] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Mpca: multilinear principal component analysis of tensor objects, IEEE Trans. Neural Netw. 19 (1) (2008) 18–39.
[46] C.-Z. Di, C.M. Crainiceanu, B.S. Caffo, N.M. Punjabi, Multilevel functional principal component analysis, Ann. Appl. Stat. 3 (1) (2009) 458.
[47] C. Happ, S. Greven, Multivariate functional principal component analysis for data observed on different (dimensional) domains, J. Am. Stat. Assoc. (2018) 1–11.
[48] M. Monforte, F. Ficuciello, B. Siciliano, Multifunctional Principal Component Analysis for Human-like Grasping, in: Human Friendly Robotics, Springer, 2019, pp. 47–58.
[49] T. Metsalu, J. Vilo, Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap, Nucleic Acids Res. 43 (W1) (2015) W566–W570.
[50] Y. Su, Y. Huang, C.C.J. Kuo, Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis, in: 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 585–590.
[51] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graphical Stat. 15 (2) (2006) 265–286.
[52] M. Journée, Y. Nesterov, P. Richtárik, R. Sepulchre, Generalized power method for sparse principal component analysis, J. Mach. Learn. Res. 11 (Feb) (2010) 517–553.
[53] C. Fabris, A. Facchinetti, G. Fico, F. Sambo, M.T. Arredondo, C. Cobelli, M.E.P. Consortium, Parsimonious description of glucose variability in type 2 diabetes by sparse principal component analysis, J. Diabetes Sci. Technol. 10 (1) (2016) 119–124.
[54] S. Yi, Z. Lai, Z. He, Y. Cheung, Y. Liu, Joint sparse principal component analysis, Pattern Recognit. 61 (2017) 524–536.
[55] A. Abid, M.J. Zhang, V.K. Bagaria, J. Zou, Exploring patterns enriched in a dataset with contrastive principal component analysis, Nat. Commun. 9 (1) (2018) 2134.
[56] H. Cardot, D. Degras, Online principal component analysis in high dimension: which algorithm to choose? Int. Stat. Rev. 86 (1) (2018) 29–50.
[57] J. Fan, Q. Sun, W.-X. Zhou, Z. Zhu, Principal component analysis for big data, arXiv preprint arXiv:1801.01602.
[58] N. Kumar, S. Singh, A. Kumar, Random permutation principal component analysis for cancelable biometric recognition, Appl. Intell. 48 (9) (2018) 2824–2836.
[59] G.H. Golub, C. Reinsch, Singular Value Decomposition and Least Squares Solutions, in: Linear Algebra, Springer, 1971, pp. 134–151.
[60] C.F.V. Loan, Generalizing the singular value decomposition, SIAM J. Numer. Anal. 13 (1) (1976) 76–83.
[61] K. Simek, Properties of a singular value decomposition based dynamical model of gene expression data, Int. J. Appl. Math. Comput. Sci. 13 (2003) 337–345.
[62] K. Modarresi, Unsupervised feature extraction using singular value decomposition, Procedia Comput. Sci. 51 (2015) 2417–2425.
[63] L. Cao, Singular Value Decomposition Applied to Digital Image Processing, in: Division of Computing Studies, Arizona State University Polytechnic Campus, Mesa, Arizona State University polytechnic Campus, 2006, pp. 1–15.
[64] A.R. Santos, M.A. Santos, J. Baumbach, J.A.M. Culloch, G.C. Oliveira, A. Silva, A. Miyoshi, V. Azevedo, A Singular Value Decomposition Approach for Improved Taxonomic Classification of Biological Sequences, in: BMC Genomics, volume 12, BioMed Central, 2011, p. S11.
[65] A. Lassiter, S. Gugercin, Handwritten digit classification and reconstruction of marred images using singular value decomposition,

[66] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.
[67] Y. Wang, L. Zhu, Research and Implementation of Svd in Machine Learning, in: IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), IEEE, 2017, pp. 471–475.
[68] P. Mills, Singular value decomposition (svd) tutorial: Applications, examples, exercises, 2017, https://blog.statsbot.co/singular-value-decomposition-tutorial-52c695315254, (Accessed on 09/04/2019).
[69] J. Steinberger, K. Ježek, Text Summarization and Singular Value Decomposition, in: International Conference on Advances in Information Systems, Springer, 2004, pp. 245–254.
[70] G. Strang, Introduction to Linear Algebra, Wellesley-Cambridge Press, Massachusetts, 2003.
[71] N.B. Erichson, S. Voronin, S.L. Brunton, J.N. Kutz, Randomized matrix decompositions using r, arXiv preprint arXiv:1608.02148.
[72] R. Silvério-Machado, B.R. Couto, M.A. dos Santos, Retrieval of enterobacteriaceae drug targets using singular value decomposition, Bioinformatics 31 (8) (2014) 1267–1273.
[73] V. Guillemot, D. Beaton, A. Gloaguen, T. Löfstedt, B. Levine, N. Raymond, A. Tenenhaus, H. Abdi, A constrained singular value decomposition method that integrates sparsity and orthogonality, PloS one 14 (3) (2019) e0211463.
[74] F. Husson, J. Josse, B. Narasimhan, G. Robin, Imputation of mixed data with multilevel singular value decomposition, J. Comput. Graphical Stat. (2019) 1–26.
[75] R. Talib, M.K. Hanif, S. Ayesha, F. Fatima, Text mining: techniques, applications and issues, Int. J. Adv. Comput. Sci. Appl. 1 (7) (2016) 414–418.
[76] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Process. 25 (2–3) (1998) 259–284.
[77] P. Wiemer-Hastings, K. Wiemer-Hastings, A. Graesser, Latent Semantic Analysis, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence, 2004, pp. 1–14.
[78] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[79] P. Reidy, An introduction to latent semantic analysis, Ind. J. Comput. Sci. Eng.
[80] N.E. Evangelopoulos, Latent semantic analysis, Wiley Interdiscip. Rev. Cognit. Sci. 4 (6) (2013) 683–692.
[81] O. Dokun, E. Celebi, Single-document summarization using latent semantic analysis, Int. J. Scientif. Res. Inf. Syst. Eng. (IJSRISE) 1(2)
[82] Z. Zhiyuan, Latent semantic analysis,Unknown.
[83] Q.V. Dang, Implementing an Individualized Recommendation System Using Latent Semantic Analysis, in: Proceedings of the 6th International Conference on Information and Education Technology, ACM, 2018, pp. 239–243.
[84] O. Müller, T. Schmiedel, E. Gorbacheva, J.V. Brocke, Towards a typology of business process management professionals: identifying patterns of competences through latent semantic analysis, Enterprise Inf. Syst. 10 (1) (2016) 50–80.
[85] S.R. Vrana, D.T. Vrana, L.A. Penner, S. Eggly, R.B. Slatcher, N. Hagiwara, Latent semantic analysis: a new measure of patient-physician communication, Social Sci. Med. 198 (2018) 22–26.
[86] D. Gefen, J. Miller, J.K. Armstrong, F.H. Cornelius, N. Robertson, A. Smith-McLallen, J.A. Taylor, Identifying patterns in medical records through latent semantic analysis, Commun. ACM 61 (6) (2018) 72–77.
[87] S. Santilli, L. Nota, G. Pilato, The Use of Latent Semantic Analysis in the Positive Psychology: A Comparison with Twitter Posts, in: IEEE 11th International Conference on Semantic Computing (ICSC), IEEE, 2017, pp. 494–498.
[88] T. Hofmann, Probabilistic Latent Semantic Analysis, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.
[89] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach Learn 42 (1–2) (2001) 177–196.
[90] C. Zhai, C. Geigle, A Tutorial on Probabilistic Topic Models for Text Data Retrieval and Analysis, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2018, pp. 1395–1398.
[91] L. Si, R. Jin, Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2005, pp. 622–631.
[92] H.T. Tu, T.T. Phan, K.P. Nguyen, An Adaptive Latent Semantic Analysis for Text Mining, in: International Conference on System Science and Engineering (ICSSE), IEEE, 2017, pp. 588–593.
[93] A.K. Uysal, S. Gunal, Text classification using genetic algorithm oriented latent semantic features, Expert Syst. Appl. 41 (13) (2014) 5938–5947.
[94] D. Zhou, D. Yang, X. Zhang, S. Huang, S. Feng, Discriminative probabilistic latent semantic analysis with application to single sample face recognition, Neural Process. Lett. 49 (3) (2019) 1273–1298.
[95] X. He, S. Yan, Y. Hu, H.J. Zhang, Learning a Locality Preserving Subspace for Visual Recognition, in: Ninth IEEE International Conference on Computer Vision, IEEE, 2003, pp. 385–392.
[96] X. He, P. Niyogi, Locality Preserving Projections, in: Advances in Neural Information Processing Systems, 2004, pp. 153–160.
[97] Y. Xu, G. Feng, Y. Zhao, One improvement to two-dimensional locality preserving projection method for use with face recognition, Neurocomputing 73 (1–3) (2009) 245–249.
[98] W.K. Wong, H. Zhao, Supervised optimal locality preserving projection, Pattern Recognit. 45 (1) (2012) 186–197.
[99] X. He, Incremental Semi-supervised Subspace Learning for Image Retrieval, in: Proceedings of the 12th Annual ACM International Conference on Multimedia, ACM, 2004, pp. 2–8.
[100] L. Zhu, S. Zhu, Face recognition based on orthogonal discriminant locality preserving projections, Neurocomputing 70 (7–9) (2007) 1543–1546.
[101] J.-B. Li, J.-S. Pan, S.M. Chen, Kernel self-optimized locality preserving discriminant analysis for feature extraction and recognition, Neurocomputing 74 (17) (2011) 3019–3027.
[102] P. Huang, T. Li, Z. Shu, G. Gao, G. Yang, C. Qian, Locality-regularized linear regression discriminant analysis for feature extraction, Inf. Sci. (Ny) 429 (2018) 164–176.
[103] D. Brown, H. Li, Y. Gao, Locality-regularized Linear Regression for Face Recognition, in: ICPR, 2012, pp. 1586–1589.
[104] J. Cheng, Q. Liu, H. Lu, Y.W. Chen, Supervised kernel locality preserving projections for face recognition, Neurocomputing 67 (2005) 443–449.
[105] W. Yu, X. Teng, C. Liu, Face recognition using discriminant locality preserving projections, Image Vis. Comput. 24 (3) (2006) 239–248.
[106] G.-F. Lu, Y. Wang, J. Zou, Z. Wang, Matrix exponential based discriminant locality preserving projections for feature extraction, Neural Netw. 97 (2018) 127–136.
[107] F. Zhong, J. Zhang, D. Li, Discriminant locality preserving projections based on l1-norm maximization, IEEE Trans. Neural Netw. Learn. Syst. 25 (11) (2014) 2065–2074.
[108] G.-F. Lu, Z. Lin, Z. Jin, Face recognition using discriminant locality preserving projections based on maximum margin criterion, Pattern Recognit. 43 (10) (2010) 3572–3579.
[109] C. Turan, K.-M. Lam, X. He, Soft locality preserving map (slpm) for facial expression recognition, arXiv preprint arXiv:1801.03754.
[110] B. Wang, Y. Hu, J. Gao, Y. Sun, H. Chen, B. Yin, Locality preserving projections for grassmann manifold, arXiv preprint arXiv:1704.08458.
[111] H. Zhao, S. Sun, Z. Jing, Local-information-based uncorrelated feature extraction, Opt. Eng. 45 (2) (2006) 20505.
[112] S. Chen, H. Zhao, M. Kong, B. Luo, 2D-lpp: a two-dimensional extension of locality preserving projections, Neurocomputing 70 (4–6) (2007) 912–921.
[113] M. Wan, G. Yang, C. Sun, M. Liu, Sparse two-dimensional discriminant locality-preserving projection (s2ddlpp) for feature extraction, Soft Comput. (2018) 1–8.
[114] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (3) (1994) 287–314.
[115] D.T. Pham, P. Garat, Blind separation of mixture of independent sources through a quasi-maximum likelihood approach, IEEE Trans. Signal Process. 45 (7) (1997) 1712–1725.
[116] A. Akkalkotkar, K.S. Brown, An algorithm for separation of mixed sparse and gaussian sources, PloS one 12 (4) (2017) e0175775.
[117] E. Shchurenkova, Dimension reduction using independent component analysis with an application in business psychology, University of British Columbia, 2017 Ph.d. thesis.
[118] J. Rahmanishamsi, A. Dolati, M.R. Aghabozorgi, A copula based ica algorithm and its application to time series clustering, J. Classif. 35 (2) (2018) 230–249.
[119] A. Hyvarinen, Fast and fixed-point algorithms for independent component analysis, IEEE Trans. Neural Netw. 10 (3) (1999) 626–634.
[120] J. Neto, Independent component analysis, 2015, http://www.di.fc.ul.pt/~jpn/r/ica/index.html, (Accessed on 09/06/2019).
[121] X.-s. He, F. He, A.l. He, Super-gaussian bss using fast-ica with chebyshev–pade approximant, Circuits Syst. Signal Process. 37 (1) (2018) 305–341.
[122] Z. Yang, S. La Conte, X. Weng, X. Hu, Ranking and averaging independent component analysis by reproducibility (raicar), Hum. Brain Mapp. 29 (6) (2008) 711–725.
[123] T. Radüntz, J. Scouten, O. Hochmuth, B. Meffert, Automated eeg artifact elimination by applying machine learning algorithms to ica-based features, J. Neural Eng. 14 (4) (2017) 46004.
[124] M.F. Glasser, T.S. Coalson, J.D. Bijsterbosch, S.J. Harrison, M.P. Harms, A. Anticevic, D.C.V. Essen, S.M. Smith, Using temporal ica to selectively remove global noise while preserving global signal in functional mri data, Neuroimage 181 (2018) 692–717.
[125] H. Ince, T.B. Trafalis, A hybrid forecasting model for stock market prediction, Econ. Comput. Econ. Cybern. Stud. Res. 51 (3) (2017) 263–280.
[126] M. Nascimento, F.F.e. Silva, T. Sáfadi, A.C.C. Nascimento, T.E.M. Ferreira, L.M.A. Barroso, C.F. Azevedo, S.E.F. Guimarães, N.V.L. Serão, Independent component analysis (ica) based-clustering of temporal rna-seq data, PloS one 12 (7) (2017) e0181195.
[127] C.F. Beckmann, S.M. Smith, Probabilistic independent component analysis for functional magnetic resonance imaging, IEEE Trans. Med. Imaging 23 (2) (2004) 137–152.
[128] P. Ablin, J.-F. Cardoso, A. Gramfort, Faster ica under orthogonal constraint, arXiv preprint arXiv:1711.10873.
[129] N. Abrahamsen, P. Rigollet, Sparse gaussian ica, arXiv preprint arXiv:1804.00408.
[130] S.J. Press, S. Wilson, Choosing between logistic regression and discriminant analysis, J. Am. Stat. Assoc. 73 (364) (1978) 699–705.
[131] P.A. Lachenbruch, M. Goldstein, Discriminant analysis, Biometrics (1979) 69–85.
[132] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. (7) (1997) 711–720.
[133] G. Martos, Discriminant analysis in r, 2017, https://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html, (Accessed on 09/06/2019).
[134] A. Tharwat, T. Gaber, A. Ibrahim, A.E. Hassanien, Linear discriminant analysis: a detailed tutorial, AI Commun. 30 (2) (2017) 169–190.
[135] H. Yu, J. Yang, A direct lda algorithm for high-dimensional data with application to face recognition, Pattern Recognit. 34 (10) (2001) 2067–2070.
[136] D. Zhou, X. Yang, Face Recognition Using Improved-lda, in: International Conference Image Analysis and Recognition, Springer, 2004, pp. 692–699.

[137] C.H. Park, H. Park, A comparison of generalized linear discriminant analysis algorithms, Pattern Recognit. 41 (3) (2008) 1083–1097.
[138] R. Ran, B. Fang, X. Wu, S. Zhang, A simple and effective generalization of exponential matrix discriminant analysis and its application to face recognition, IEICE Trans. Inf. Syst. 101 (1) (2018) 265–268.
[139] H. Xiong, W. Cheng, W. Hu, J. Bian, Z. Guo, Fwda: a fast wishart discriminant analysis with its application to electronic health records data classification, arXiv preprint arXiv:1704.07790.
[140] L. Wu, C. Shen, A. van den Hengel, Deep linear discriminant analysis on fisher networks: a hybrid architecture for person re-identification, Pattern Recognit. 65 (2017) 238–250.
[141] A. Krasoulis, K. Nazarpour, S. Vijayakumar, Use of Regularized Discriminant Analysis Improves Myoelectric Hand Movement Classification, in: 8th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2017, pp. 395–398.
[142] M. Kedadouche, Z. Liu, M. Thomas, Bearing Fault Feature Extraction Using Autoregressive Coefficients, Linear Discriminant Analysis and Support Vector Machine under Variable Operating Conditions, in: Advances in Condition Monitoring of Machinery in Non-Stationary Operations, Springer, 2018, pp. 339–352.
[143] A. Tyagi, V. Nehra, A comparison of feature extraction and dimensionality reduction techniques for eeg-based bci system, IUP J. Comput. Sci. 11 (1) (2017) 51.
[144] S. Wilson, M. Close, P. Abraham, Applying linear discriminant analysis to predict groundwater redox conditions conducive to denitrification, J. Hydrol. (Amst) 556 (2018) 611–624.
[145] C.-N. Li, Y.-H. Shao, W. Yin, M.Z. Liu, Robust and sparse linear discriminant analysis via an alternating direction method of multipliers, IEEE Trans. Neural Netw. Learn. Syst.
[146] D. Zhang, Y. Zhao, M. Du, A new supervised dimensionality reduction algorithm using linear discriminant analysis and locality preserving projection, WSEAS Trans. Inf. Sci. Appl. E-ISSN (2013) 2224–3402.
[147] Z. Liu, G. Liu, J. Pu, X. Wang, H. Wang, Orthogonal sparse linear discriminant analysis, Int. J. Syst. Sci. (2018) 1–11.
[148] M. Zhu, A.M. Martinez, Subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1274–1286.
[149] N. Gkalelis, V. Mezaris, I. Kompatsiaris, Mixture subclass discriminant analysis, IEEE Signal Process. Lett. 18 (5) (2011) 319–322.
[150] H. Wan, H. Wang, G. Guo, X. Wei, Separability-oriented subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 409–422.
[151] J. Ye, R. Janardan, Q. Li, Two-dimensional Linear Discriminant Analysis, in: Advances in Neural Information Processing Systems, 2005, pp. 1569–1576.
[152] C.-N. Li, Y.-H. Shao, W.-J. Chen, N.Y. Deng, Generalized two-dimensional linear discriminant analysis with regularization, arXiv:1801.07426.
[153] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.
[154] J.H. Friedman, J.W. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Comput. 100 (9) (1974) 881–890.
[155] T. Jiang, H. Jia, H. Yuan, N. Zhou, F. Li, Projection pursuit: a general methodology of wide-area coherency detection in bulk power grid, IEEE Trans. Power Syst. 31 (4) (2015) 2776–2786.
[156] S. Espezua, E. Villanueva, C.D. Maciel, A. Carvalho, A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets, Neurocomputing 149 (2015) 767–776.
[157] M.C. Jones, R. Sibson, What is projection pursuit? J. R. Stat. Soc. Ser. A (1987) 1–37.
[158] P.J. Huber, Projection pursuit, Ann. Stat. (1985) 435–475.
[159] G. Nason, Three-dimensional projection pursuit, J. R. Stat. Soc. Ser. C (Applied Statistics) 44 (4) (1995) 411–430.
[160] G.P. Nason, Design and choice of projection indices, University of Bath, 1992 Ph.d. thesis.
[161] C. Posse, Projection pursuit exploratory data analysis, Comput. Stat. Data Anal. 20 (6) (1995) 669–687.
[162] C. Posse, Tools for two-dimensional exploratory projection pursuit, J. Comput. Graphical Stat. 4 (2) (1995) 83–100.
[163] Q. Guo, W. Wu, F. Questier, D. Massart, C. Boucon, S. De Jong, Sequential projection pursuit using genetic algorithms for data mining of analytical data, Anal. Chem. 72 (13) (2000) 2846–2855.
[164] E.-K. Lee, D. Cook, S. Klinke, T. Lumley, Projection pursuit for exploratory supervised classification, J. Comput. Graphical Stat. 14 (4) (2005) 831–846.
[165] A. Berro, S.L. Marie-Sainte, A. Ruiz-Gazen, Genetic algorithms and particle swarm optimization for exploratory projection pursuit, Ann. Math. Artif. Intell. 60 (1–2) (2010) 153–178.
[166] M. Grochowski, W. Duch, Projection Pursuit Constructive Neural Networks Based on Quality of Projected Clusters, in: International Conference on Artificial Neural Networks, Springer, 2008, pp. 754–762.
[167] E.-K. Lee, D. Cook, A projection pursuit index for large p small n data, Stat. Comput. 20 (3) (2010) 381–392.
[168] L.O. Jimenez, D.A. Landgrebe, Hyperspectral data analysis and supervised feature reduction via projection pursuit, IEEE Trans. Geosci. Remote Sens. 37 (6) (1999) 2653–2667.
[169] I. Perisic, C. Posse, Projection pursuit indices based on the empirical distribution function, J. Comput. Graphical Stat. 14 (3) (2005) 700–715.
[170] D. Peña, F.J. Prieto, Cluster identification using projections, J. Am. Stat. Assoc. 96 (456) (2001) 1433–1445.
[171] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Stat. Assoc. 76 (376) (1981) 817–823.
[172] P. Hall, et al., On projection pursuit regression, Ann. Stat. 17 (2) (1989) 573–588.
[173] N. Intrator, Combining exploratory projection pursuit and projection pursuit regression with application to neural networks, Neural Comput. 5 (3) (1993) 443–455.
[174] L. Jimenez, D. Landgrebe, High Dimensional Feature Reduction via Projection Pursuit, 2, I. E.E.E. Press, Piscataway NJ, 1994, pp. 1145–1147.
[175] J.-X. Pan, W.-K. Fung, K.T. Fang, Multiple outlier detection in multivariate data using projection pursuit techniques, J. Stat. Plan. Inference 83 (1) (2000) 153–167.
[176] R. Bolton, W. Krzanowski, Projection pursuit clustering for exploratory data analysis, J. Comput. Graphical Stat. 12 (1) (2003) 121–142.
[177] E. Bingham, H. Mannila, 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, ACM, 2001, pp. 245–250.
[178] Y.D. Lee, D. Cook, J. Park, E.K. Lee, et al., Pptree: projection pursuit classification tree, Electron. J. Stat. 7 (2013) 1369–1386.
[179] H. Zhang, C. Wang, W. Fan, A projection pursuit dynamic cluster model based on a memetic algorithm, Tsinghua Sci. Technol. 20 (6) (2015) 661–671.
[180] N.d. Silva, D. Cook, E.K. Lee, A projection pursuit forest algorithm for supervised classification. arXiv preprint arXiv:1807.07207.
[181] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, N. Koudas, Non-linear Dimensionality Reduction Techniques for Classification and Visualization, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 645–651.
[182] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer Science & Business Media, 2007.
[183] K.Q. Weinberger, F. Sha, L.K. Saul, Learning a Kernel Matrix for Nonlinear Dimensionality Reduction, in: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, 2004, p. 106.
[184] V.D. Silva, J.B. Tenenbaum, Global versus local methods in nonlinear dimensionality reduction, Adv. Neural Inf. Process. Syst. (2003) 721–728.
[185] B. Schölkopf, A. Smola, K.R. Müller, Kernel Principal Component Analysis, in: International Conference on Artificial Neural Networks, Springer, 1997, pp. 583–588.
[186] R. Osadchy, Kpca: Unsupervised learning, Lectture lides,
[187] N. Pochet, F. De Smet, J.A. Suykens, B.L. De Moor, Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction, Bioinformatics 20 (17) (2004) 3185–3195.
[188] L. Xie, Z. Li, J. Zeng, U. Kruger, Block adaptive kernel principal component analysis for nonlinear process monitoring, AIChE J. 62 (12) (2016) 4334–4345.
[189] J.A. Lee, A. Lendasse, M. Verleysen, Nonlinear projection with curvilinear distances: isomap versus curvilinear distance analysis, Neurocomputing 57 (2004) 49–76.
[190] V. Franc, V. Hlaváč, Greedy Kernel Principal Component Analysis, in: Cognitive Vision Systems, Springer, 2006, pp. 87–105.
[191] T.-J. Chin, D. Suter, Incremental kernel principal component analysis, IEEE Trans. Image Process. 16 (6) (2007) 1662–1674.
[192] Y. Washizawa, Subset Kernel Principal Component Analysis, in: Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on, IEEE, 2009, pp. 1–6.
[193] M. Debruyne, T. Verdonck, Robust kernel principal component analysis and classification, Adv. Data Anal. Classif. 4 (2–3) (2010) 151–167.
[194] Y. Zhang, C. Ma, Fault diagnosis of nonlinear processes using multiscale kpca and multiscale kpls, Chem. Eng. Sci. 66 (1) (2011) 64–72.
[195] J. Chen, G. Wang, G.B. Giannakis, Nonlinear dimensionality reduction for discriminative analytics of multiple datasets, IEEE Trans. Signal Process. 67 (3) (2018) 740–752.
[196] J.B. Kruskal, M. Wish, Multidimensional Scaling, Number 07–011 in Sage University Paper Series on Quantitative Applications in the Social Sciences, 1978.
[197] J.J. Meulman, The integration of multidimensional scaling and multivariate analysis with optimal transformations, Psychometrika 57 (4) (1992) 539–565.
[198] L. Blouvshtein, D. Cohen-Or, Outlier detection for robust multi-dimensional scaling, arXiv preprint arXiv:1802.02341.
[199] N. Saeed, H. Nam, M.I.U. Haq, D.B.M. Saqib, A survey on multidimensional scaling, ACM Comput. Surv. (CSUR) 51 (3) (2018) 47.
[200] F.D. Mandanas, C.L. Kotropoulos, Robust multidimensional scaling using a maximum correntropy criterion, IEEE Trans. Signal Process. 65 (4) (2017) 919–932.
[201] Y. Ma, K. He, J. Hopcroft, P. Shi, Neighbourhood-preserving dimension reduction via localised multidimensional scaling, Theor. Comput. Sci. 734 (2018) 58–71.
[202] L. Chen, A. Buja, Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis, J. Am. Stat. Assoc. 104 (485) (2009) 209–219.
[203] H.H. Mohamed, S. Belaid, W. Naanaa, L.B. Romdhane, Local commute-time guided mds for 3d non-rigid object retrieval, Appl. Intell. (2018) 1–11.
[204] E. Peterfreund, M. Gavish, Multidimensional scaling of noisy high dimensional data, arXiv preprint arXiv:1801.10229.
[205] A.W. Hanley, A.K. Baker, R.T. Hanley, E.L. Garland, The shape of self-extension: mapping the extended self with multidimensional scaling, Pers. Individ. Differ. 126 (2018) 25–32.
[206] O. Kayo, Locally linear embedding algorithm–extensions and applications,
[207] Z.A. Sani, A. Shalbaf, H. Behnam, R. Shalbaf, Automatic computation of left ventricular volume changes over a cardiac cycle from echocardiography images by nonlinear dimensionality reduction, J. Digit Imaging 28 (1) (2015) 91–98.
[208] M. Frasca, A. Rizzo, L. Gallo, L. Fortuna, M. Porfiri, Dimensionality reduction in epidemic spreading models, EPL (Europhysics Letters) 111 (6) (2015) 68006.
[209] A.S. Rao, J. Gubbi, M. Palaniswami, Anomalous Crowd Event Analysis Using Isometric Mapping, in: Advances in Signal Processing and Intelligent Recognition Systems, Springer, 2016, pp. 407–418.
[210] S.-H. Liu, K.-Y. Chen, B. Chen, H.-M. Wang, W.L. Hsu, Leveraging Manifold

Learning for Extractive Broadcast News Summarization, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 5805–5809.
[211] X. Zheng, Y. Yuan, X. Lu, Dimensionality reduction by spatial–spectral preservation in selected bands, IEEE Trans. Geosci. Remote Sens. 55 (9) (2017) 5185–5197.
[212] J. Jiang, X. He, M. Gao, X. Wang, X. Wu, Human action recognition via compressive-sensing-based dimensionality reduction, Optik-Int. J. Light Electron Optics 126 (9–10) (2015) 882–887.
[213] A. Najafi, A. Joudaki, E. Fatemizadeh, Nonlinear dimensionality reduction via path-based isometric mapping, IEEE Trans. Pattern Anal. Mach. Intell. 38 (7) (2016) 1452–1464.
[214] H. Shi, B. Yin, Y. Bao, Y. Lei, A Novel Landmark Point Selection Method for L-isomap, in: 12th IEEE International Conference on Control and Automation (ICCA), IEEE, 2016, pp. 621–625.
[215] S. Suganya, G. Muthumari, C. Balasubramanian, Propitiating Behavioral Variability for Mouse Dynamics Using Dimensionality Reduction Based Approach, in: International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE), IEEE, 2016, pp. 1–6.
[216] X. Li, Y. Xie, D. Hu, Z. Lan, Analysis of the geometrical evolution in on-the-fly surface-hopping nonadiabatic dynamics with machine learning dimensionality reduction approaches: classical multidimensional scaling and isometric feature mapping, J. Chem. Theory Comput. 13 (10) (2017) 4611–4623.
[217] Y. Liu, C. Xia, Z. Fan, R. Wu, X. Chen, Z. Liu, Implementation of fractal dimension and self-organizing map to detect toxic effects of toluene on movement tracks of daphnia magna, J. Toxicol. (2018).
[218] M.M. Nezhad, E. Gironacci, M. Rezania, N. Khalili, Stochastic modelling of crack propagation in materials with random properties using isometric mapping for dimensionality reduction of nonlinear data sets, Int. J. Numer. Methods Eng. 113 (4) (2018) 656–680.
[219] L.K. Saul, S.T. Roweis, An introduction to locally linear embedding, Unpublished. Available at:http://www.cs.toronto.edu/~roweis/lle/publications.html.
[220] O. Abdel-Mannan, A.B. Hamza, A. Youssef, Incremental Hessian Locally Linear Embedding Algorithm, in: Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on, IEEE, 2007, pp. 1–4.
[221] Z. Zhang, J. Wang, Mlle: Modified Locally Linear Embedding Using Multiple Weights, in: Advances in Neural Information Processing Systems, 2007, pp. 1593–1600.
[222] O. Kouropteva, O. Okun, M. Pietikäinen, Supervised Locally Linear Embedding Algorithm for Pattern Recognition, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2003, pp. 386–394.
[223] S. Zhang, K.W. Chau, Dimension Reduction Using Semi-supervised Locally Linear Embedding for Plant Leaf Classification, in: International Conference on Intelligent Computing, Springer, 2009, pp. 948–955.
[224] R. Hettiarachchi, J.F. Peters, Multi-manifold lle learning in pattern recognition, Pattern Recognit. 48 (9) (2015) 2947–2960.
[225] T. Kohonen, The self-organizing map, Proc. IEEE 78 (9) (1990) 1464–1480.
[226] T. Kohonen, The self-organizing map, Neurocomputing 21 (1–3) (1998) 1–6.
[227] T. Kohonen, Essentials of the self-organizing map, Neural Netw. 37 (2013) 52–65.
[228] N. Chen, N.C. Marques, An Extension of Self-organizing Maps to Categorical Data, in: Portuguese Conference on Artificial Intelligence, Springer, 2005, pp. 304–313.
[229] S.R. Mounce, R. Sharpe, V. Speight, B. Holden, J. Boxall, Self-organizing Maps for Knowledge Discovery from Corporate Databases to Develop Risk Based Prioritization for Stagnation, in: 11th International Conference on Hydroinformatics, Sheffield, 2015.
[230] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE Trans. Neural Netw. 11 (3) (2000) 574–585.
[231] E. De laHoz, E.D.L. Hoz, A. Ortiz, J. Ortega, B. Prieto, Pca filtering and probabilistic som for network intrusion detection, Neurocomputing 164 (2015) 71–81.
[232] O. Polat, Z. Dokur, Protein fold recognition using self-organizing map neural network, Curr. Bioinform. 11 (4) (2016) 451–458.
[233] G. Das, M. Chattopadhyay, S. Gupta, A comparison of self-organising maps and principal components analysis, Int. J. Market Res. 58 (6) (2016) 815–834.
[234] E. Merényi, J. Taylor, Som-empowered Graph Segmentation for Fast Automatic Clustering of Large and Complex Data, in: Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), 2017 12th International Workshop on, IEEE, 2017, pp. 1–9.
[235] P. Mohan, K.K. Patil, Weather and crop prediction using modified self organizing map for mysore region, Methods 193 (2017) 192–199.
[236] V. Chaudhary, R. Bhatia, A.K. Ahlawat, Community som (csom): an improved self--organizing map learning technique, Int. J. Fuzzy Syst. 17 (2) (2015) 129–132.
[237] D. Isa, V. Kallimani, L.H. Lee, Using the self organizing map for clustering of text documents, Expert Syst. Appl. 36 (5) (2009) 9584–9591.
[238] H.-C. Yang, C.-H. Lee, H.W. Hsiao, Incorporating self-organizing map with text mining techniques for text hierarchy generation, Appl. Soft Comput. 34 (2015) 251–259.
[239] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, D.A. Keim, Somflow: guided exploratory cluster analysis with self-organizing maps and analytic provenance, IEEE Trans. Vis. Comput. Graph. 24 (1) (2018) 120–130.
[240] C.-L. Liu, W.-H. Hsaio, T.-H. Chang, H.H. Li, Clustering data with partial background information, Int. J. Mach. Learn. Cybern. (2018) 1–16.
[241] V. Mirshafiee, B.A. Harley, M.L. Kraft, Visualizing intrapopulation hematopoietic cell heterogeneity with self-organizing maps of sims data, Tissue Eng. Part C: Methods 24 (6) (2018) 322–330.
[242] J. Yousefi, A. Hamilton-Wright, Characterizing emg data using machine-learning tools, Comput. Biol. Med. 51 (2014) 1–13.
[243] S. Bashyal, G.K. Venayagamoorthy, Recognition of facial expressions using gabor wavelets and learning vector quantization, Eng. Appl. Artif. Intell. 21 (7) (2008) 1056–1064.
[244] D. Fitria, M.A. Ma'sum, E.M. Imah, A.A. Gunawan, Automatic arrhythmias detection using various types of artificial neural network based learning vector quantization (lvq), Jurnal Ilmu Komputer dan Informasi 7 (2) (2014) 90–100.
[245] B. Hammer, D. Hofmann, F.-M. Schleif, X. Zhu, Learning vector quantization for (dis-) similarities, Neurocomputing 131 (2014) 43–51.
[246] S. Hu, Y. Gu, H. Jiang, Study of Classification Model for College Students' M-learning Strategies Based on Pca-lvq Neural Network, in: 8th International Conference on Biomedical Engineering and Informatics (BMEI), IEEE, 2015, pp. 742–746.
[247] D. Nebel, M. Kaden, Dissimilarity extraction in a median variant of learning vector quantization, Mach. Learn. Rep. 9 (2015) 33–40.
[248] D. Hofmann, A. Gisbrecht, B. Hammer, Efficient approximations of robust soft learning vector quantization for non-vectorial data, Neurocomputing 147 (2015) 96–106.
[249] F.M. Schleif, Small Sets of Random Fourier Features by Kernelized Matrix Lvq, in: Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), 2017 12th International Workshop on, IEEE, 2017, pp. 1–5.
[250] G.E. Hinton, S.T. Roweis, Stochastic Neighbor Embedding, in: Advances in Neural Information Processing Systems, 2003, pp. 857–864.
[251] L.v.d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605. (Nov)
[252] A. Konstorum, E. Vidal, N. Jekel, R. Laubenbacher, Comparative analysis of linear and nonlinear dimension reduction techniques on mass cytometry data, bioRxiv (2018) 273862.
[253] B. Xie, Y. Mu, D. Tao, K. Huang, M-sne: multiview stochastic neighbor embedding, IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) 41 (4) (2011) 1088–1096.
[254] A. Platzer, Visualization of snps with t-sne, PloS one 8 (2) (2013) e56883.
[255] A. Gisbrecht, A. Schulz, B. Hammer, Parametric nonlinear dimensionality reduction using kernel t-sne, Neurocomputing 147 (2015) 71–82.
[256] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.
[257] F.S. Tsai, K.L. Chan, Dimensionality Reduction Techniques for Data Exploration, in: 2007 6th International Conference on Information, Communications & Signal Processing, IEEE, 2007, pp. 1–5.
[258] W. Black, B.J. Babin, Multivariate Data Analysis: Its Approach, Evolution, and Impact, in: The Great Facilitator, Springer, 2019, pp. 121–130.
[259] U.D. Dixit, M. Shirdhonkar, Logo Based Document Image Retrieval Using Singular Value Decomposition Features, in: International Conference on Signal and Information Processing (IConSIP), IEEE, 2016, pp. 1–4.
[260] J. Lin, D. Gunopulos, Dimensionality Reduction by Random Projection and Latent Semantic Indexing, in: Proceedings of the Text Mining Workshop at the 3rd SIAM International Conference on Data Mining, 2003.
[261] S. Hao, Y. Xu, H. Peng, K. Su, D. Ke, Automated Chinese Essay Scoring from Topic Perspective Using Regularized Latent Semantic Indexing, in: 22nd International Conference on Pattern Recognition (ICPR), IEEE, 2014, pp. 3092–3097.
[262] W. Ren, G. Wen, R. Luan, Z. Yang, Z. Zhang, Single-channel Blind Source Separation and Its Application on Arc Sound Signal Processing, in: Transactions on Intelligent Welding Manufacturing, Springer, 2018, pp. 115–126.