

A Comparative Study on Linear Dimensionality Reduction Techniques For Big Data Analytics

Devinder Kaur, Oshin Bhardwaj, Surbhi Sharma,
devinder.21696@lpu.co.in, oshin.24954@lpu.co.in,
surbhi.23540@lpu.co.in School of Computer Science &
Engineering
Lovely Professional University, Jalandhar, India

Abstract - With a rapid increase in advance and smart infrastructure in our daily lives, a huge amount of data generation occurs from various smart devices. This data usually falls under the category of big data as it is generated in large volumes with high speed at regular intervals. So, it becomes very challenging to represent and process such a large amount of data. Big data handling mainly involves data pre-processing and data processing. Finally, decision making is done on this processed data. Furthermore, data pre-processing plays a very crucial role in the efficient big data processing and data analytics. The data which is generated in large amount from various devices is of heterogeneous nature and uncertain. Big data pre-processing involves various steps, such as feature selection, feature extraction, data compression, reduction of duplicate and null values, and dimensionality reduction. High dimensionality of data is one of the main problem at this stage of data preprocessing to deal with. In this paper, a survey is done on various linear dimensionality reduction techniques. However, to choose an appropriate technique for a specified data is very important and it highly depends upon the nature of data we are dealing with. Dimensionality reduction techniques mainly fall under the category of lossless and lossy reductions. In this paper, we have reviewed lossy techniques to reduce the high dimensionality of big data. Various techniques namely Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Linear Discriminant Analysis (LDA), etc are reviewed and compared depending upon various parameters. These techniques are implemented and simulated on a dataset using python programming for various iterations. While performing the dimensionality reduction reconstruction error should be minimum and reduction should be maximum. So in this paper, we have reviewed linear techniques with respect to various parameters such as reconstruction error and reduction ratio.

Keywords: Big data, dimensionality reduction techniques, data compression, Cyber-Physical Systems, data analytics

I. INTRODUCTION

From the past few years we have witnessed the emergence of big data everywhere. Big data is basically the large data sets which are so large in volume that it becomes very difficult to deal with them using the traditional methods. According to the big data definition, the 5 main characteristics of big data are volume, velocity, veracity, variety and value. Volume defines the size of any data which is found to be very large in case of big data generally in petabytes and exabytes or even more than that. Velocity defines the speed of data with which it gets generated. Generally, the mobility of data defines its velocity. Today, a massive amount of data gets generated at real time within a fraction of seconds such as credit card transactions, posts going viral on social media like facebook, twitter, instagram etc. Moreover, Majority of the data which gets generated is heterogeneous in nature. Structured, unstructured, semi structured are the main types of big data. Everyday, every minute a large amount of data gets generated from various devices such as mobile phones, social networking sites, multimedia, sensor, software logs, RFIDs etc. So, it becomes very difficult to handle such massive datasets. Big data face many challenges during its processing due to its characteristics.

A. Challenges of big data

According to the latest stats, data generation is growing at faster rates than ever before. The huge amount of big data generated at such a high speed poses multiple dimensions. Due to the heterogeneous nature of big data, problem of high dimensionality arises in data sets. The heterogeneous nature of data involves structured, unstructured, and semi structured type which comes under the variety characteristic of big data. Generally, high dimensionality comes from applications such as, multimedia, mobiles, social networking sites etc. The different

variety of data such as texts, xmls, photos, videos, audios, web pages, etc. make the dimensionality of data multiple and thus highly complicated. The number of attributes in datasets has become huge along with the huge number of instances which are very difficult to handle using traditional data handling techniques. It makes the data more complex and to handle such data increases both the time and space complexity. The main challenges faced by big data include capturing the data, data storage, data analysis, and visualization and a few more to name. One of main challenges that it faces at the early stages of data analysis after storage is of high dimensionality. It becomes very important to reduce the dimensions of data to lower dimensions so that data visualization becomes easy. The dimensionality of big data remains a bottleneck for the provision of quality of service. The variety character present in the big data makes it high dimensional. The data which is generated is of heterogeneous nature as it includes video, audio, text, xml etc. type of data involved in it. Moreover each type of data further involve various types of attributes which makes the data multidimensional. In order to reduce the dimensions, we have a very power technique which is known as dimensionality reduction. It is widely used in data analytics and machine learning to visualize data more effectively. So, dimensionality reduction is basically used to represent high order or high dimensional data using lower orders or dimensions. This domain which deals with such challenges is also known as feature engineering. Feature selection and feature extraction are two main types of feature engineering. Dimensionality reduction also can be considered as a sub branch of feature engineering. Figure no. 1 shows 5 main characteristics of big data.

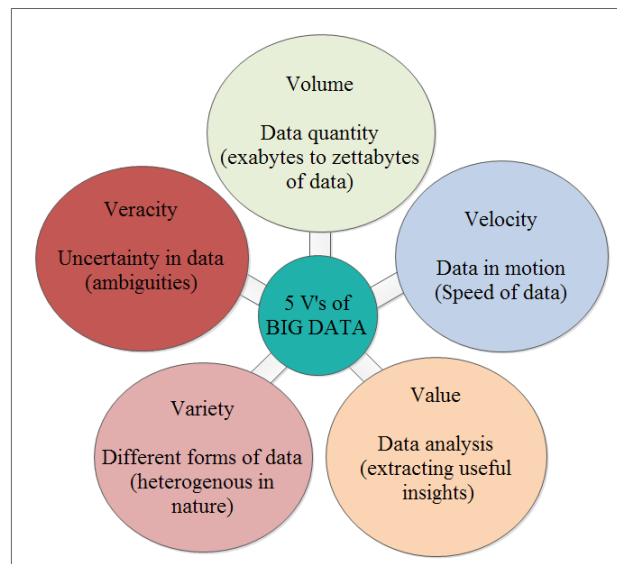


Fig. 1. 5 Vs of big data

B. Research Contribution of this paper

Major contributions of this paper are as given below.

- 1) Comparative analysis of various dimensionality reduction techniques is done.
- 2) A case study for a flower dataset is taken to generate the lower dimensions.
- 3) The analysis is being done using extensive simulations on realistic traces of high dimensional data using various parameters.

C. Organization

The remaining paper is organized as follows. Following subsection represents big data challenges and various existing dimensionality reduction techniques. Section II represents the rationale of this work. Section III involves the results and discussions related to the comparative analysis of various techniques discussed in this paper. Finally, Section IV concludes the paper.

D. Existing big data dimensional reduction techniques

There are many techniques present to deal with the dimensionality reduction such as PCA (Principal Component analysis), LDA (Linear Discriminant analysis, Singular Value Decomposition (SVD), kernel PCA, a few more to name. The dimensionality reduction techniques can be further viewed as for linear and non linear type of data. A data which is highly dimensional and non linear in nature can be dealt with non-linear techniques such as High dimensional singular value decomposition (HOSVD), Multi-linear PCA, Tucker decompositions are also used for to deal with this problem. Each technique has its own pros and cons depending upon the type of data they are used for to reduce the dimensionality.

1) Singular Value Decomposition (SVD)

SVD is one of the main dimensionality reduction techniques in the field of linear techniques. SVD is also known by the name of factor analysis. SVD is basically a mathematical procedure which is used to decompose a matrix into three other matrices. It has many applications in various fields specifically in the field of signal processing and statistics which deals with the representation, interpretation, and analysis of data. Apart from this, it has spread its branches in other fields as well such as in psychology and sociology, astronomy, climate and atmospheric science. Although, it has a very important role to play in the field of machine learning and data analytics. SVD involves factorization of a real or complex matrices. De Souza, *et al.* in

[1] presented a lossy data compression technique based upon SVD in smart distribution systems. They utilized the proposed scheme to reduce the volume of data in smart grids and then to transmit it through communication network in an effective manner. Authors inferred that SVD is a very competitive data reduction technique which also maintains the originality of data after the reconstruction. Similarly, authors in [2] used a high-order singular value decomposition (HOSVD) technique to decompose smart grid data. HOSVD is a non-linear form of SVD, which uses SVD as its main technique and is widely used to decompose high-dimensional data. In the similar manner, authors in [3], [4], [5] represented various schemes based upon singular value decomposition for different fields. In SVD, original matrix M is being decomposed into three other matrices. Out of which two of them are orthogonal matrices and the third one is the diagonal matrix which actually define the singular values. This third matrix is known as singular matrix. It can be defined using following equation:

$$M = USV^*(1)$$

Here, M represents the original matrix, S represents the singular matrix. U represents orthogonal matrix.

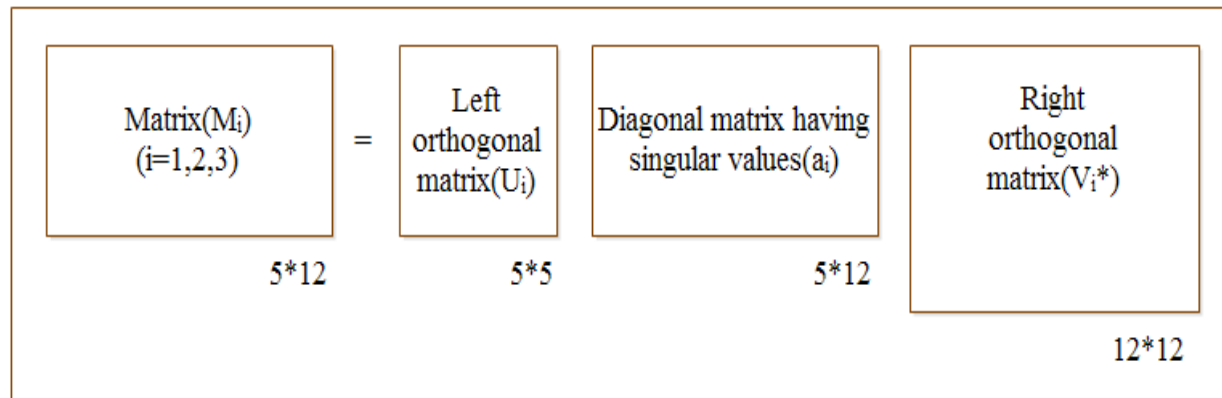
V^* represents the transpose conjugate of matrix V . The diagonal matrix S , with non-negative entries has i singular values denoted by σ_i . The singular values (σ_i) in the diagonal matrix S are as given below.

$$\begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_n \end{bmatrix}$$

$$S = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_i & 0 & \dots & 0 \end{pmatrix} \quad (3)$$

$$\text{where; } \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_i > 0 \quad (4)$$

Figure 2 shows an example of SVD, in which a matrix of dimensions 5*12 is taken. This matrix is decomposed into three other matrices as shown in the right side, from which central matrix defines the singular matrix having singular values with respective dimensions. Further, Top n singular values can be extracted from it for better reduction ratio using truncated-SVD. As, reduction ratio and reconstruction error is a trade-off to balance the



originality of data or information. So, singular values can be extracted accordingly for the data reconstruction.

Fig. 2. Singular value decomposition

2) Principal component analysis (PCA)

PCA is one of another very popular technique in the field of dimensionality reduction. It is a statistical mechanism that uses an orthogonal transformation to convert a set of observations of correlated attributes or columns into a set of values which are linearly uncorrelated. These uncorrelated values are called as principal components (PCs). PCA extracts the dominant patterns from the matrix in the form of principal components. The first set of PCs captures the maximum variance, followed by second, third, and so on. Final PCs results into the lower dimensions of data. Furthermore, there are various factors upon which the results of the matrix depend such as scaling of the matrix, variance scaling etc. It is also known as linear PCA. It has a wide range of applications in different fields. PCA is used as a main tool in data analytics and predictive models. PCA is a very similar procedure to factor analysis. PCA deals with the problem of over-fitting in data, which gets generated in the data due to larger number of attributes. Number of principal components can be equal or less than the number of attributes or features present in the data. Basically, it depends upon the original dimensions of the data. PC1 is given the highest priority from the different components.

PC1 and PC2 should preserve the property of orthogonality. i.e. PC1 should be totally independent from PC2. i.e. different PCs should not have any dependency upon each other. The mathematical procedure of obtaining principal components from the high dimensional raw data can be defined using following steps:

- 1) Collect the data and represent it.
- 2) Calculate the mean for each dimension present in the dataset.
- 3) Obtain the covariance matrix for the given dataset.
- 4) Find the eigenvectors and respective eigenvalues for each eigenvalue.
- 5) Sort the eigenvectors in the decreasing order of eigenvalues.
- 6) Choose n eigenvectors with the highest eigenvalues to form a m n dimensional matrix M.
- 7) Use this m n eigenvector matrix to transform the samples onto the new subspace.

3) Linear discriminant analysis (LDA)

LDA is also one of another linear technique to deal with the problem of dimensionality reduction. It is also known as normal discriminant analysis (NDA), or discriminant function analysis. It is a generalization of Fisher's linear discriminant. LDA is a method which is widely used in various fields such as in statistics, and pattern recognition. It is used to find a linear combination of features that characterizes two or more classes of objects or attributes. The resulting combination may be used as a linear classifier, or for dimensionality reduction before later classification. LDA is an unsupervised algorithm. LDA in comparison to the PCA, it tends to preserve as much as discriminatory power as possible. Mathematically, LDA can be used to reduce the higher dimensions

of data by using steps as given below.

- 1) Obtain scatter matrices for within the class and between class.
- 2) Obtain eigenvectors and corresponding eigenvalues for above scatter matrices.
- 3) Sort the eigenvalues in decreasing order and select the top n values.
- 4) Create a new matrix containing eigenvectors that map to the n values.
- 5) Obtain the new attributes or features which are known as LDA components by taking the dot product of the data and the matrix from step 4.

In this section, various dimensionality reduction techniques are discussed by different authors. In the context of PCA, [9] used PCA on a data which was subjected to uncertainty, to reduce its dimensionality. Authors discussed that, linear dimensionality reduction techniques are more advantageous comparatively to the non-linear. As, probability distribution remain intact in the former case. Zhao *et al.* in [10] used a supervised dimensionality reduction technique, i.e. LDA. In this paper, authors first used PCA to remove the null values and then applied LDA to avoid the small sample size problem. They proposed a joint novel scheme to implement PCA and LDA jointly. In the similar context, authors in [11] proposed a robust method for LDA using Bhattacharyya error bound optimization. Furthermore, authors proposed that RLDA can be easily extended to sparse LDA. Similarly in [12], authors investigated recursive problem of orthogonal LDA and orthogonal least square regressions (SLR). In [13], authors have reviewed approx. 50 papers for different techniques present for dimensionality reduction. Similarly, various authors in [6], [7], [8] presented and analyzed various compression and data decomposition techniques in order to deal with the problem of dimensionality reduction.

II. RATIONALE

After analyzing the above research proposals, it is inferred that a vast amount of data is generated in today's world from various smart devices and appliances. It becomes very difficult to store, handle, and process this huge amount of data. So, if data pre-processing is done using an appropriate technique, then it can lead to a faster and efficient data processing and big data analytics. As this processed big data is utilized for various real life applications such as to make weather and electricity predictions. So, dimensionality reduction in data pre-processing plays a very vital and crucial role to deal with the above discussed challenges of big data. Also, big data represented as tensors could be reduced to a simpler form by removing the redundant and ambiguous dimensions. Further, to draw useful data patterns and decisions from the reduced data, data analytics and smart metering comes into pictures. In order to process this information there is a requirement of a well-incorporated system such as CPSS that include all the essential factors such as physical entities, cyber and social space.

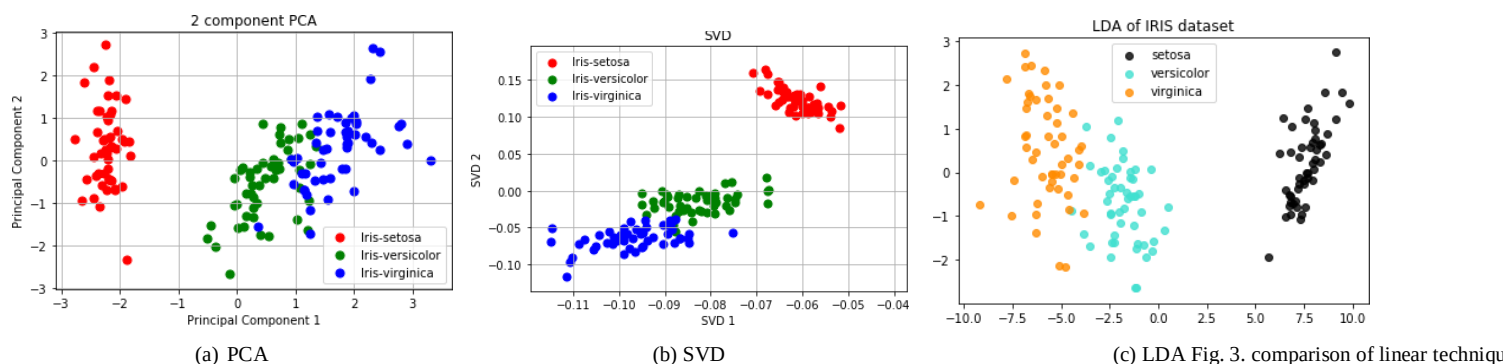
III. RESULTS AND DISCUSSIONS

In this paper, we analyzed and compared various techniques present for dimensionality reduction in big data. A standard data set for 3 different flower varieties is taken. Three linear dimensionality reduction techniques are applied over it, namely

TABLE I
COMPARISON TABLE OF DIMENSIONALITY REDUCTION TECHNIQUES

Sr no.	Name	Features	Applications
1	Singular value decomposition(SVD)	- A singular matrix is formed to obtain the singular values. - SVD can be utilized to find the principal components.	- Widely used for data compression in the field of signal processing and statistics. - Commonly used for fitting solutions in linear equations.
2	principal component analysis(PCA)	- It is unsupervised in nature and a technique for linear transformation of data. - PCA gives better results in classification task of an image recognition if the number of samples for a given class are relatively small. - PCA can be pictured as a technique that finds the directions of maximal variance in a data.	- Used in the fields of Quantitative science and neuroscience. - Primarily used for dimensionality reduction in the field of face recognition, image compression, and computer vision. - Also used in the domains of finance, data mining, and psychology to find pattern and perform decision making.
3	Linear Discriminant analysis(LDA)	- In contrast to PCA, LDA tends to find a feature subspace that maximizes class separability. - LDA is supervised in nature. - Linear transformation technique Closely related to ANOVA and regression analysis.	- LDA is successfully used for identification, predictions and for pattern recognition. - It is widely used in machine learning for predictions and decision making.

PCA, SVD, and LDA. Results are as shown in the figure number 3. First figure in this shows two principal components after applying PCA on the dataset. First PC is having maximum variance. Second figure shows the implementation of SVD, having a singular matrix. And finally, third figure shows the implementation of LDA.



IV. CONCLUSION

In this paper, a comparative study of various linear dimensionality reduction techniques is presented. These techniques are broadly classified into linear and non-linear categories. As, data pre-processing plays a very crucial role in the efficient big data processing and thus data analytics. So, presenting the high dimensions of data into lower dimensions without losing the originality of data is very important. However, to choose an appropriate technique for a specified data is very important and it highly depends upon the nature of data that we are dealing with. Various techniques namely Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Linear Discriminant Analysis (LDA) are reviewed and compared depending upon various parameters such as reconstruction error and reduction ratio. These techniques are implemented and simulated on a dataset using python programming for a number of iterations. While performing the dimensionality reduction reconstruction error should be minimum and reduction should be maximum. The techniques are analyzed by taking a common data set and by applying various iterations on them. It can be stated that selecting a suitable dimensionality reduction technique is a trade-off between reducing the reconstruction error and maximizing the reduction.

REFERENCES

- [1] J. C. S. de Souza, T. M. L. Assis, and B. C. Pal, "Data compression in smart distribution systems via singular value decomposition," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 275–284, 2017.
- [2] D. Kaur, G. S. Aujla, N. Kumar, A. Y. Zomaya, C. Perera, and R. Ranjan, "Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: Sdn perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1985–1998, Oct 2018.
- [3] A. Zhang and D. Xia, "Tensor svd: Statistical and computational limits," *IEEE Transactions on Information Theory*, 2018.
- [4] L. T. Yang, X. Wang, X. Chen, L. Wang, R. Ranjan, X. Chen, and M. J. Deen, "A multi-order distributed hosvd with its incremental computing for big services in cyber-physical-social systems," *IEEE Transactions on Big Data*, 2018.
- [5] I. Kajo, N. Kamel, and Y. Ruichek, "Incremental tensor-based completion method for detection of stationary foreground objects," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.
- [6] M. T. Asif, S. Kannan, J. Dauwels, and P. Jaillet, "Data compression techniques for urban traffic data," in *Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2013 IEEE Symposium on*. IEEE, 2013, pp. 44–49.
- [7] H. Harb and C. A. Jaoude, "Combining compression and clustering techniques to handle big data collected in sensor networks," in *Communications Conference (MENACOMM), IEEE Middle East and North Africa*. IEEE, 2018, pp. 1–6.
- [8] D. Dong and J. Herbert, "Content-aware partial compression for textual big data analysis in hadoop," *IEEE Transactions on Big Data*, 2017.
- [9] J. Grtler, T. Spinner, D. Streeb, D. Weiskopf, and O. Deussen, "Uncertainty-aware principal component analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 822–831, Jan 2020.
- [10] X. Zhao, J. Guo, F. Nie, L. Chen, Z. Li, and H. Zhang, "Joint principal component and discriminant analysis for dimensionality reduction," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2019.
- [11] C. Li, Y. Shao, W. Yin, and M. Liu, "Robust and sparse linear discriminant analysis via an alternating direction method of multipliers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2019.
- [12] Y. Liu, R. Zhang, F. Nie, X. Li, and C. Ding, "Supervised dimensionality reduction methods via recursive regression," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2019.
- [13] W. Siblini, P. Kuntz, and F. Meyer, "A review on dimensionality reduction for multi-label classification," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.