ACEAT-493
Data Visualization by PCA, LDA, and ICA

Tsun-Yu Yang[a], Chaur-Chin Chen[a,b]
[a]Institute of Information Systems and Applications, National Tsing Hua U., Taiwan
[b]Department of Computer Science, National Tsing Hua University, Taiwan
*Corresponding author: cchen@cs.nthu.edu.tw

ABSTRACT

While the internet applications are widely used, the amount of data information has a rapid growth. Data Science or Big data analysis has become a popular issue nowadays. Data visualization provides intuitive methods to reveal some important properties of high-dimensional data, for example, clustering tendency.

This paper studies three linear dimensionality reduction methods to project high-dimensional data into lower-dimensional subspace: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are linear mappings for the usage of projection. Independent Component Analysis (ICA) is originally proposed to solve the blind source separation problem. Like a linear mapping, ICA also computes a mapping matrix for data projection. Unlike PCA and LDA, ICA could extract the independent sources in a mixture of non-Gaussian distributions.

We review the background of PCA, LDA, and ICA by algorithmic approaches and illustrate the 2D projection associated with a K-means clustering result on three data sets: IRIS, 8OX, and Thyroid data. Experimental results help us reveal the structure of high-dimensional data in some sense.

Keyword: Big Data Analysis, Data Science, ICA, LDA, PCA.

1. Introduction

The Moore's law is an observation which indicates over the history of computing hardware, the number of transistors in a dense integrated circuit doubles approximately every two years [1]. Although this trend has continued for more than half a century, its doubts about the ability of projection to remain valid into the indefinite future have been expressed. On the other hand, the huge data need to be processed and analyzed by software have an exponential growth, for example, medical study, customer behaviors and many other fields. The size of database could easily grow from current terabytes to petabytes, exabytes, zettabytes, or even yottabytes in the near future. The term of Big data is mentioned to describe this phenomenon. How to effectively illustrate and process Big data is an emerging issue.

Big data analysis costs plenty of resources in data processing even if the time and space complexity of the algorithms is low. We are interested in the way to represent the features of observed data in a low-dimensional space for visualization. The extracted features should narrate the characteristics of a dataset precisely and make the dataset readable [2]. This work studies dimensionality reduction from high-dimensional data to lower-dimensional space for visualization.

According to Friedman [3], the main goal of data visualization is to communicate information clearly and effectively through graphical means. There are two major advantages in data visualization: (1) Make the observation clustering characteristics display in a more intuitive way. (2) Data visualization converts abstract high-dimensional data into visual diagrams which makes the evaluation of data feasible and transmits the result of analysis handily.

Because of being the most capable with human visual ability, data visualization generally displays in 2-D or 3-D pictures. It means that when we deal with high-dimensional data, we should process dimensionality reduction on dataset before data visualization through feature selection or extraction. We study three linear methods: Principal Component Analysis (PCA) [4], Linear Discriminant Analysis (LDA) [5], and Independent Component Analysis (ICA) [6]. Figure 1 illustrates that a linear transform from high-dimensional data into a 2 or 3 dimensional space, we can easily display the aspect of an original dataset in a 2D or 3D diagrams [7].

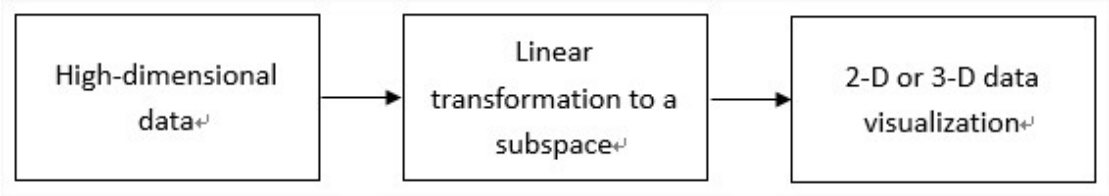


Figure 1. A Diagram of Data Reduction for Visualization.

PCA was introduced by K. Pearson in early 20th century [8]. It linearly transforms data into a lower-dimensional subspace by obtaining the maximized variance of the data in a low-dimensional representation. LDA was first proposed by R.A. Fisher in 1936 which uses class label to compute the between-class scatter matrix and within-class scatter matrix. The criterion defined by these two matrices will give a good separation among different classes [9]. ICA was first proposed by P. Comon [6]. ICA divides a mixed signal into additive subcomponents of non-Gaussian signals that are statistically independent from each other [10]. This paper adopts an algorithmic approach to design and implement PCA, LDA, and ICA for dimensionality reduction

on three data sets: IRIS, 8OX, and Thyroid [11].

The remaining of this paper is organized as follows. Section 2 gives a background review of PCA, LDA, and ICA. Section 3 describes the data sets and illustrates the experimental results. Section 4 draws the conclusion.

2. Background Review for PCA, LDA, and ICA

2.1 Principal Component Analysis (PCA)

PCA is a dimensionality reduction method which is like fitting an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipse is short, the variance along that axis is also small. By ignoring that axis and its corresponding fewer principal components from the representation of the dataset, we only lose a little amount of information.

In practice, PCA is an orthogonal linear transform, which converts an original data into a new coordinate system such that the largest variance of the projected data comes to lie along the first axis, the second largest variance lies on the second axis, and so on. We define these coordinates as the principal components [12].

Considering n observations in a dataset, each observation is *m*-dimensional by ignoring the class label. Let $\overrightarrow{x_{\mathrm{i}}} \in \mathbb{R}^m, \mathrm{i} = 1, 2, \ldots, \mathrm{n}$. We depict the steps of computing principal components as follows.

(a) Compute the m-dimensional mean vector $\vec{\mu}$ by

$$\vec{\mu} = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{x_{\mathrm{i}}} \tag{1}$$

(b) Compute the estimated covariance matrix C for the observed data by

$$\mathrm{C} = \frac{1}{n}\sum_{i=1}^{n} (\overrightarrow{x_{\mathrm{i}}} - \vec{\mu})(\overrightarrow{x_{\mathrm{i}}} - \vec{\mu})^t \tag{2}$$

(c) Compute n eigenvalues and corresponding eigenvectors of C, $(\lambda_i, \overrightarrow{v_i}), 1 \leq i \leq \mathrm{n}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

(d) Compute the first d principal components by

$$y_i^{(j)} = \overrightarrow{x_{\mathrm{i}}}^t \overrightarrow{u_{\mathrm{j}}} \tag{3}$$

for each observation $\vec{x}_i, 1 \leq i \leq n,$ along the direction $\overrightarrow{u_{\mathrm{j}}},\ j = 1, 2, \cdots, d.$

Because the estimated covariance matrix C is nonnegative definite, so all of its eigenvalues must be real and nonnegative. In general, a few larger eigenvalues dominate the others in most of the practical data sets, that is,

$$\rho_k = \frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_k+\cdots+\lambda_m} \geq 85\%\ for\ 1 \leq k \ll m \tag{4}$$

where ρ_k is sometimes called the percentage retained in data representation.

2.2 Linear Discriminant Analysis (LDA)

LDA is another dimensionality reduction method such that feature clusters are most separable after the transformation. It requires the class label of each observation [13]. Consider a set of observations $\{\vec{x}\}$ with K classes, the linear discriminant analysis (LDA) or Fisher discriminant analysis (FDA) can be depicted as follows. Suppose an observation may come from the *ith* class of K classes with each class containing n_i observations, the within-class scatter matrix S_w can be written as

$$S_w = \sum_{i=1}^{K} \sum_{\vec{x} \in \omega_i} (\vec{x} - \overrightarrow{\mu_i})(\vec{x} - \overrightarrow{\mu_i})^t \quad (5)$$

The between-class scatter matrix S_b can be defined by the sample covariance of the class means

$$S_b = \sum_{i=1}^{K} (\overrightarrow{\mu_i} - \vec{\mu})(\overrightarrow{\mu_i} - \vec{\mu})^t \quad (6)$$

where $\vec{\mu}$ is the mean vector of all observations. The class separation in a direction $\vec{w}$ in this case will be given by

$$\rho = \frac{\vec{w}^t S_b \vec{w}}{\vec{w}^t S_w \vec{w}} \quad (7)$$

Seeking a vector $\vec{w}$ to maximize ρ in Eq. (7) is equivalent to solving a generalized eigenvalue/eigenvector system given as follows

$$S_b \vec{w} = \lambda S_w \vec{w} \quad (8)$$

2.3 Independent Component Analysis (ICA)

ICA is essentially a multivariate, parallel version of projection pursuit method which is like PCA [15]. Whereas projection pursuit extracts a series of signals one at a time from a set of M signal mixtures, ICA extracts M signals in parallel. This property makes ICA a more robust method than PCA [16]. ICA attempts to decompose a multivariate signal into independent non-Gaussian signals [17]. ICA separation of a mixed signal gives good results if the following two assumptions are satisfied.

(i) The source signals are independent of each other.

(ii) The values in each source signal have non-Gaussian distributions.

We chose *FastICA* [19] as ICA practical method or called an approach of Maximization of non-Gaussianity. *FastICA* is widely used in many applications. Before the *FastICA* algorithm could be applied, the input vector data $\vec{w}$ should be centered and whitened. First, input the data by subtracting the mean vector of $\vec{\mu}$ for the sample $\{\vec{w}\}$.

$$\vec{v} = \vec{w} - \vec{\mu} \quad (9)$$

$\vec{v}$ is the centered vector of $\vec{w}$ and $\vec{\mu}$ is the sample mean vector. The next step is whitening the data by

$$\overrightarrow{w_h} = UD^{-\frac{1}{2}}U^t\vec{v} \tag{10}$$

where U is the orthogonal matrix of eigenvectors and D is the diagonal matrix of eigenvalues, so that

$$\mathrm{E}\{\overrightarrow{w_h}\overrightarrow{w_h}^t\} = \mathrm{I} \tag{11}$$

After the preprocessing, the *FastICA* algorithm could extract multiple independent components. We follow the definition of *FastICA* by Hyvärinen [19] to give two functions f and g as follows.

$$\begin{aligned} &\text{(a)}\ \ f(u) = \log(\cosh(u));\ f'(u) = \tanh(u);\ f''(u) = 1 - \tanh^2(u) \\ &\text{(b)}\ \ g(u) = -e^{-u^2/2};\ g'(u) = ue^{-u^2/2};\ g''(u) = (1-u^2)e^{-u^2/2} \end{aligned} \tag{12}$$

$f(u)$ and $g(u)$ are non-quadratic functions, which obtain approximations of negentropy that give a very good compromise between the properties of the two classical non-Gaussianity measures given by kurtosis and negentropy.

Consider an observation matrix $\mathrm{X} \in \mathbb{R}^{m\times n}$, where X contains *n* columns of *m*-dimensional observed vector $\{\vec{\mathrm{w}}_i,\ i=1,2,\ldots,n\}$. Given a number of desired components K, K<*m*. An ICA algorithm can be stated as follows.

Input: $\mathrm{X} \in \mathbb{R}^{m\times n}$: *n m*-dim sample vectors, and K: no. of independent components.
Output: $\mathrm{W} \in \mathbb{R}^{K\times m}$: a projection matrix to extract K independent components.
Output: $\mathrm{S} \in \mathbb{R}^{K\times n}$: each column represents a vector of K independent components.

for $\mathrm{i} = 1, 2, \cdots, \mathrm{K}$

$\overrightarrow{w_\mathrm{i}} \leftarrow random\ vector\ of\ length\ m$

while $\overrightarrow{w_\mathrm{i}}$ changes, do {

$$\overrightarrow{w_\mathrm{i}} \leftarrow \frac{1}{n}Xg'\left(\overrightarrow{w_\mathrm{i}}^t X\right)^t - \frac{1}{n}g''\left(\overrightarrow{w_\mathrm{i}}^t X\right)\mathbf{1_m}\overrightarrow{w_\mathrm{i}}$$

$$\overrightarrow{w_\mathrm{i}} \leftarrow \overrightarrow{w_\mathrm{i}} - \sum_{j=1}^{i-1}\left(\overrightarrow{w_\mathrm{i}}^t\overrightarrow{w_\mathrm{j}}\right)\overrightarrow{w_\mathrm{j}}$$

$$\overrightarrow{w_\mathrm{i}} = \frac{\overrightarrow{w_\mathrm{i}}}{\|\overrightarrow{w_\mathrm{i}}\|_2}$$

}

endfor

Output: $\mathrm{W} = \begin{bmatrix} \overrightarrow{w_1}^t \\ \overrightarrow{w_2}^t \\ \vdots \\ \overrightarrow{w_K}^t \end{bmatrix} \in \mathbb{R}^{K\times m}$, S=WX.

Note: For visualization, K= 2 or 3 is used.

3. Experiments

We apply the PCA, LDA and ICA on IRIS, 8OX and Thyroid data [11] to demonstrate 2D projection for visualization by Matlab codes [21].

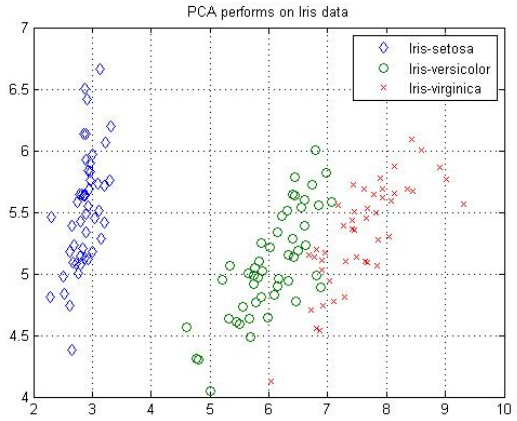
3.1 Dimensionality Reduction on IRIS Data

IRIS dataset is commonly used for a study of pattern classification, which was originally used in Fisher's experiment [5]. It consists of 3 IRIS flowers: Setosa, Versicolor, and Virginica, each class contains 50 samples of four features which are the measurements of the sepal length, sepal width, petal length, and petal width.

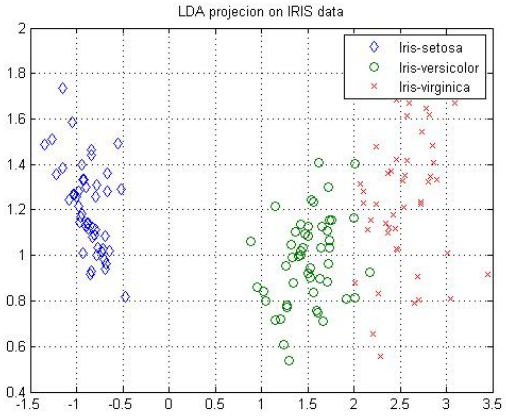
Figure 2 illustrates the results of PCA, LDA, and ICA applied on IRIS data.

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	14	36

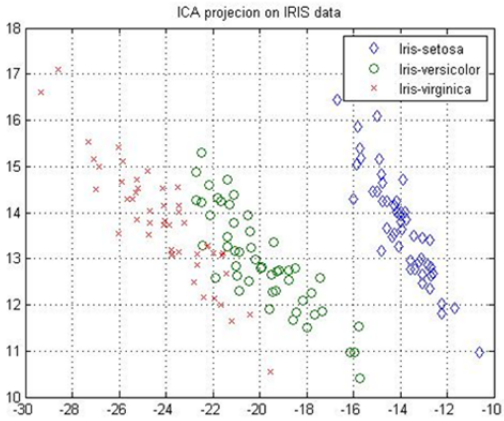
(a)



(b)



(c)



(d)

Figure 2. Results of (a) K-means, (b) PCA, (c) LDA, (d) ICA on IRIS Data.

3.2 Dimensionality Reduction on 8OX Data

The 8OX data set is derived from Munson's hand printed Fortran character set. Included are 15 patterns from each of the characters '8', 'O', 'X' consisting of 8 measurements which are the distances between edge and the character from eight

directions: *east, northeast, north, northwest, west, southwest, south,* and *southeast.*
Figure 3 illustrates the results of PCA, LDA, and ICA applied on 8OX data.

	8	O	X
8	2	13	0
O	0	15	0
X	5	0	10

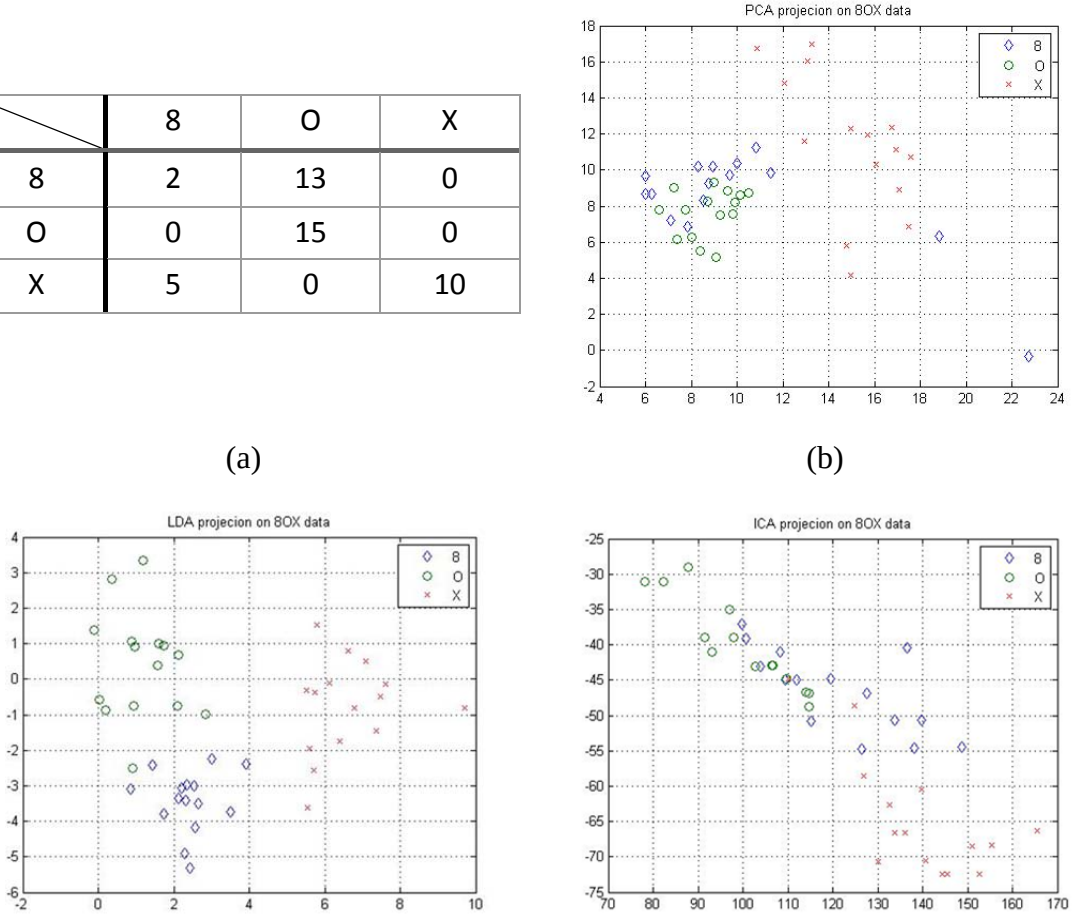
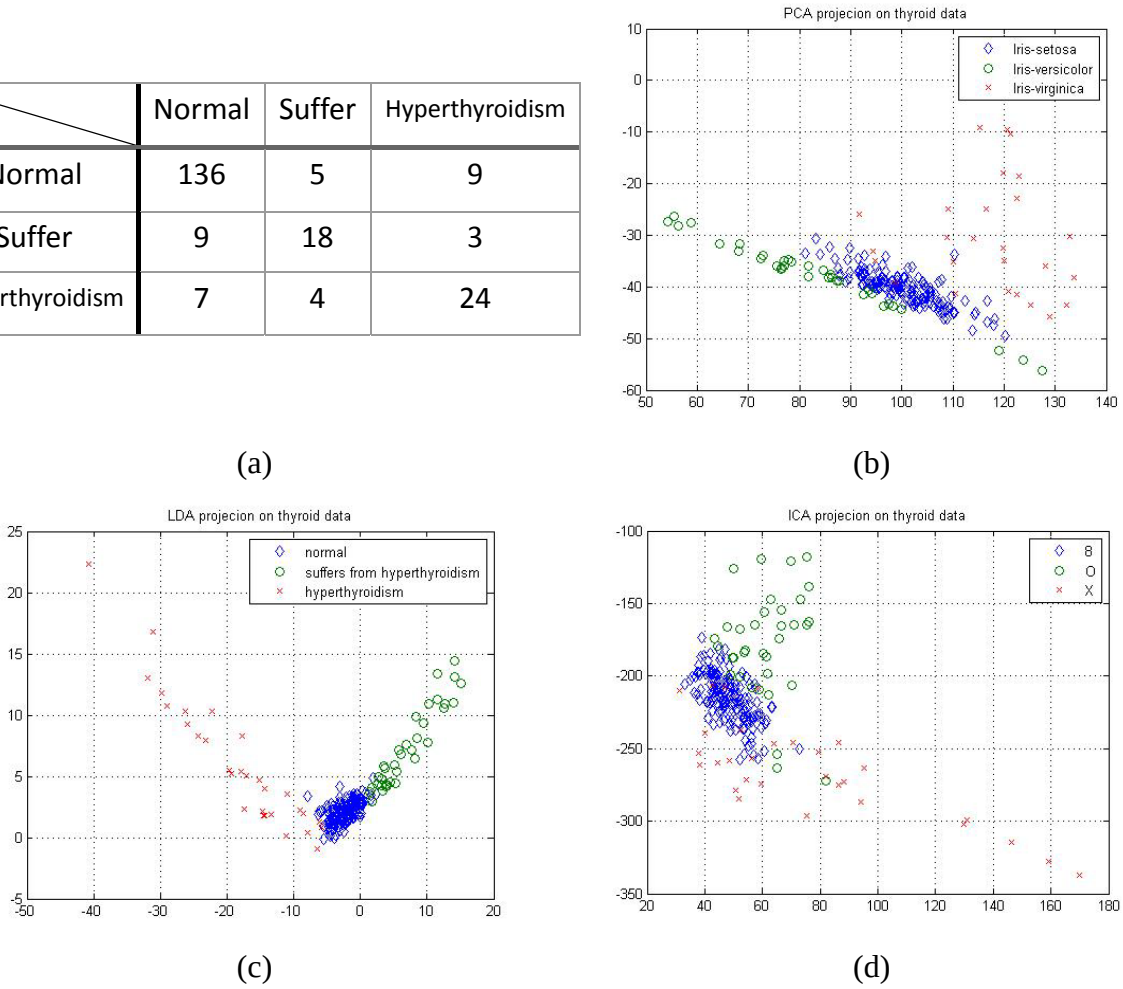


Figure 3. Results of (a) K-means, (b) PCA, (c) LDA, (d) ICA on 8OX Data.

3.3 Dimensionality Reduction on Thyroid Data

The Thyroid data set is one of the several databases about thyroid available at the UCI repository. The data set recorded 215 patients of 3 categories: (1) normal, (2) suffers from hyperthyroidism, or (3) hypothyroidism. Each sample contains five features: T3resin, Thyroxin, Triiodothyronine, Thyroidstimulating, and TSH.
Figure 4 illustrates the results of PCA, LDA, and ICA applied on Thyroid data.

	Normal	Suffer	Hyperthyroidism
Normal	136	5	9
Suffer	9	18	3
Hyperthyroidism	7	4	24



(a) (b)

(c) (d)

Figure 4. Results of (a) K-means, (b) PCA, (c) LDA, (d) ICA on Thyroid Data.

3.4 Discussion

Experiments show that LDA has better visualization performance than PCA and ICA because it utilizes the label information. There is a random factor in ICA algorithm, it means that if we choose different $\vec{w}$ vectors for calculation of independent components, the performance of ICA may obtain slightly different results.

4. Conclusion and Discussion

Data Science or Big data analysis [22] has become more important along with the rapid growth of the Internet of Things (IoT) [23]. Visualization by data dimensionality reduction help people further reveal the structure of high-dimensional Big data. This work reviews three projection methods: PCA, LDA, and ICA by algorithmic approaches with experiments illustrated on three datasets: IRIS, 8OX, and Thyroid.

All of PCA, LDA, and ICA have a significant advantage in computing time when the data dimension is relatively low compared to the sample size, while ICA takes larger

time complexity. 3D visualization by these three methods can be easily extended with a little more computation time. We compute the W projection matrix to project the original data into a lower-dimensional subspace. For a huge sample size of Big data, a sampling technique could be adopted in a preprocessing stage.

LDA uses label information, but PCA and ICA do not. If the label information is not available, one can first use a clustering method such as K-means algorithm to find the label of each pattern vector before applying LDA.

One application of visualization by dimensionality reduction on Big data is to guide future feature acquisition for improving pattern recognition and saving storage space for speeding up data transmission via the wire/wireless transmission.

5. Acknowledgments

This work is supported by Grant NSC 101-2221-E-007-125-MY3 and MOST 104-2221-E-007-035.

REFERENCES

[1] http://en.wikipedia.org/wiki/Moore%27s_law, last access on 6/9/2015.

[2] J. Yang, D. Zhang, A.F. Frangi, and J.Y. Yang, “Two-dimensional PCA: a new approach to appearance-based face representation and recognition,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 26, no. 1, 131-137, 2004.

[3] V. Friedman, "Data Visualization and Infographics in: Graphics." Monday inspiration, 2008.

[4] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, vol. 24, no. 6, 417-441, 1933.

[5] R.A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” Annals of Eugenics, vol. 7, no.2, 179–188, 1936.

[6] P. Comon, “Independent component analysis, a new concept?” Signal Processing, vol. 36, 287-314, 1994.

[7] C.C. Chen, Y.S. Shieh, and H.T. Chu. "Face image retrieval by projection-based features." The 3rd International Workshop on Image Media Quality and its Applications. 2008.

[8] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, 559-572, 1901.

[9] A.M. Martinez and A.C. Kak, “PCA Vesus LDA” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, 228-233, 2001.

[10] A.J. Bell and T.J. Sejnowski, “The independent components of natural scenes are edge filters,” Vision Research, vol. 37, no. 23, 3327-3338, 1997.

[11] https://archive.ics.uci.edu/ml/datasets.html, last acessed on June 9, 2015.

[12]I.T. Jolliffe, “Principal Component Analysis,” Springer, 1st edition, 1986.
[13]H.R. Bittencourt, B.P.O. Pasini, D.A. de O. Moraes, B.D. dos Santos, and V. Haertel, “Comparative Analysis of Two Classes Implementing Nominal Logistic Regression,” Revista Brasileira de Biometria, vol. 27, no. 1, 115-124, 2009.
[14]W.J. Krzanowski, P. Jonathan, W.V McCarthy, and M.R. Thomas, “Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data,” Applied Statistics, vol. 44, 101–115, 1995.
[15]M. Li and B. Yuan, “2D-LDA: A statistical linear discriminant analysis for image matrix” Pattern Recognition Letters, vol. 26, no. 5, 527–532, 2005.
[16]A. Hyvärinen, “A fast fixed-point algorithm for independent component analysis,” Neural computation, vol. 9, no. 7, 1483-1492, 1997.
[17]L.J. Cao, K.S. Chua, W.K. Chong, and H.P. Lee, “A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine.” Neurocomputing, vol. 55, no.1, 321-336, 2003.
[18]J. V. Stone, "Independent Component Analysis: A Tutorial Introduction," The MIT Press Cambridge, Massachusetts, London, England; ISBN 0-262-69315-1.
[19]A. Hyvärinen, “Independent component analysis: algorithms and applications,” Neural Networks, vol. 13, no. 4–5, 411–430, 2000.
[20]S. Pang, S. Ozawa, and N. Kasabov, “Incremental Linear Discriminant Analysis for Classification of Data Streams,” IEEE Trans. on Systems, Man, and Cybernetics, vol. 35, no. 5, 905-914, 2005.
[21]http://www.cs.nthu.edu.tw/~cchen/Projection/pca8OX.m, last access on 7/23/2015.
[22]https://en.wikipedia.org/wiki/Data_science, last access on 7/23/2015.
[23]https://en.wikipedia.org/wiki/Internet_of_Things, last access on 7/23/2015.