High-Dimensional Data Visualization by PCA and LDA

Chaur-Chin Chen
Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

Abbie Hsu
Institute of Information Systems & Applications, National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT: Data grow so fast that people may not have enough time to read all of the details, instead, most people want to know the structure via the graphical visualization for the data at hand. Thus motivated, this paper introduces methods to summarize data by graphical plots. We assume that the collected data have been represented in a form of pattern matrix, each row vector represents the features derived from an object under study. The goal of this paper is to review the multivariate statistical techniques: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), to provide simple Matlab codes to demonstrate how to visualize high-dimensional data sets in 2D and 3D plots by means of PCA and LDA. Experiments on the 8OX character data set, a microarray gene expression data, and Wine data set normalized by a z-score transform are demonstrated.

1 INTRODUCTION

People produce huge amount of data in daily life. By collecting and analyzing big data (WebBigData), they want to improve life or to replace human labor, such as predict economic circumstances and identify diseases. With the rapidity of computing speed and the substantial increase of storage space, it is an important issue to reveal the structure of high-dimensional data by visualization. This paper aims to investigate two popular methods, Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) for data visualization (Bishop 2006).

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of possibly correlated observations into a set of linearly uncorrelated components called principal components. On the other hand, LDA, a generalization of Fisher's linear discriminant (Bishop 2006), is a method to find a linear combination of observations which characterizes or separates two or more classes of objects. In other words, PCA and LDA are two widely used multivariate statistical methods used for dimensionality reduction, feature selection, or classification in the area of Machine Learning and Pattern Recognition (Bishop 2006). This paper aims to provide simple Matlab codes (Hanselman & Littlefield 2005) for implementations by PCA and LDA with the emphasis on the 2D and 3D visualization of high-dimensional data, including Munson’s character data set (Jain & Dubes 1988), Alon’s colon cancer data set (WebArray), and a Wine data set (WebWine).

2 PRINCIPAL COMPONENT ANALYSIS AND LINEAR DISCRIMINANT ANALYSIS

2.1 *Principal Component Analysis (WebPCA)*

Principal Component Analysis is a method of multivariate statistical analysis for dimensionality reduction on high-dimensional data sets. We give the problem statement and provide a computational solution as follows.

2.1.1. *Problem Statement and Its Solution*

Let X be an m-dimensional random vector with covariance matrix C. The problem is to consecutively find the unit vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ such that $y_i = \mathbf{x}^t \mathbf{a}_i$ with $Y_i = \mathrm{X}^t \mathbf{a}_i$ satisfies

(a)var(Y_1) is the maximum.
(b)var(Y_2) is the maximum subject to cov(Y_2,Y_1)=0.
(c) var(Y_k) is the maximum subject to cov(Y_k,Y_i)=0,

where$1<i<k\leq m$.

Y_i is called the *i*-th principal component.

To fulfill the above statements, we let (λ_i,$\mathbf{u}_i$) be the pairs of eigenvalues and eigenvectors of C such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ and $\|\mathbf{u}_i\|_2 = 1, \forall 1 \leq i \leq m$.

Then it can be shown that $\mathbf{a}_i = \mathbf{u}_i$ and $\text{var}(Y_i) = \lambda_i$ for $1 \le i \le m$ (Jolliffe 1986).

2.1.2 *Computing Principal Components*

Given observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{n} \in R^m$, the following procedures show how to compute principal components.

(1) Compute the *mean vector* $\mathbf{u} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}_\mathrm{i}/\mathrm{n}$

(2) Compute the *covariance matrix* $C = \sum_{i=1}^{\mathrm{n}} (\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^t /\mathrm{n}$.

(3) Compute the eigenvalue/eigenvector pairs $(\lambda_i, \mathbf{u}_i)$ of C, $1 \le \mathrm{i} \le \mathrm{m}$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_m \ge 0$.

(4) Compute the first *d* principal components $y_i^{(j)} = \mathbf{x}_i^t \mathbf{u}_j$ for each observation $\mathbf{x}_i$, $1 \le i \le n$, along the direction $\mathbf{u}_j$, $1 \le \mathrm{j} \le \mathrm{d}$.

It must be mentioned that the estimated covariance matrix C is nonnegative definite, all of its eigenvalues are real and nonnegative. For most of the practical data sets, fewer eigenvalues dominate the others, that is, $\rho_k = \frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_k+\cdots+\lambda_m} \ge$ 85% for $1 \le \mathrm{k} << \mathrm{m}$. In applications, it usually selects $\mathrm{k} = 2$ or 3 for the purpose of visualization.

2.2 *Linear Discriminant Analysis (WebLDA)*

Linear Discriminant Analysis is another method of multivariate statistical analysis for dimensionality reduction or classification with the category of each object is tagged. We give the problem statement followed by a computational solution below.

Given a set of training patterns $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from K categories, where $n_1 + n_2 + \cdots + n_K = n$. Define the mean vector $\mathbf{u}$, the between-class scatter matrix B, the within-calss scatter matrix W, and the total scatter matrix T as follows.

$$\mathbf{u} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}_\mathrm{i}/\mathrm{n} \tag{1}$$

$$\mathrm{B} = \sum_{\mathrm{i}=1}^{\mathrm{n}} n_i (\mathbf{u}_i - \mathbf{u})(\mathbf{u}_i - \mathbf{u})^t \tag{2}$$

$$\mathrm{W} = \sum_{\mathrm{i}=1}^{\mathrm{K}} \sum_{\mathbf{x} \in \boldsymbol{\omega}_i} (\mathbf{x} - \mathbf{u}_i)(\mathbf{x} - \mathbf{u}_i)^t \tag{3}$$

$$\mathrm{T} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^t \tag{4}$$

Note that B+W=T. Define a criterion

$$\rho = (\mathbf{v}^t \mathrm{B}\mathbf{v}/\mathbf{v}^t \mathrm{W}\mathbf{v}) \tag{5}$$

A classical discriminant analysis finds an optimal set of discriminant vectors by

[a] Look for a unit vector $\mathbf{u}_1$ which maximizes ρ, where $\mathbf{u}_1$ is the eigenvector corresponding to the largest eigenvalue of the following equations

$$\mathrm{B}\mathbf{u} = \lambda \mathrm{W}\mathbf{u} \tag{6}$$

[b] Look for a unit vector $\mathbf{u}_2$ which maximizes ρ subject to $\mathbf{u}_2^t \mathrm{W} \mathbf{u}_1$=0.

[c] Sequentially seek for a unit vector $\mathbf{u}_k$ which maximizes ρ subject to $\mathbf{u}_k^t \mathrm{W} \mathbf{u}_j$=0 for k>2, and $1 \le \mathrm{j} < k$.

In practical applications, we compute the first three generalized eigenvectors corresponding to the largest three generalized eigenvalues from eq. (6) for our usage.

3 DATA DESCRIPTION

We implement PCA and LDA for visualization on three high-dimensional data sets which are introduced as follows.

3.1 *Munson's 8OX Data Set (Jain & Dubes 1988)*

The first data set, 8OX, is extracted from Munson's handprinted Fortran character set (Jain & Dubes 1988). The 45 8-dimensional pattern vectors were derived from 45 handprinted characters written by 15 persons, each person wrote the characters "8," "O," and "X" once. A handprinted character is interpreted as a binary image placed on a 24×24 grid and a pattern vector consisting of eight features, represents the distances (counted in the number of pixels) measured from the eight directions: East, Northeast, North, Northwest, West, Southwest, South, and Southeast directions, respectively.

3.2 *Colon Cancer Data Set (Alon et al. 1999)*

The colon cancer data set was one of the gene expression data sets collected by Kent Ridge Biomedical Data Set (WebArray). The colon data set contains 62 patient samples (as patterns in our study). Among which, 40 samples come from tumor biopsies (labelled as "negative") and 22 samples come from normal biopsies (labelled as "positive") which are healthy parts of the colons of the same patients.

Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels (Alon et al. 1999). The best 46 genes (as features in our study) with the biological information unaltered are selected by applying Fisher's Linear Discriminant Analysis (De la Bastida 2013).

3.3 *Wine Data Set (WebWine)*

The Wine data were found in UCI Machine Learning Repository, a website contains a collection of databases, domain theories, and data generators that are used by the machine learning community. The data are the results of a chemical analysis of wines. The wines were grown in the same region in Italy, but were derived from three different cultivars.

An analysis determined the quantities of 13 continuous constituents (as features) found in each of the three types of wines (as classes in our study). Within

178 instances (as patterns in our study), Class 1 includes 59 patterns, Class 2 includes 71 patterns, and Class 3 includes 48 patterns.

The numbers of features, patterns, and categories in the above data sets are summarized in Table 1.

Table 1. A Summary of Data Sets in Experiments.

Data	# of features	# of Patterns	# of Categories
8OX	8	45 (15, 15, 15)	3
Colon	46	62 (22, 40)	2
Wine	13	178 (59, 71, 48)	3

4 EXPERIMENTS WITH MATLAB CODES

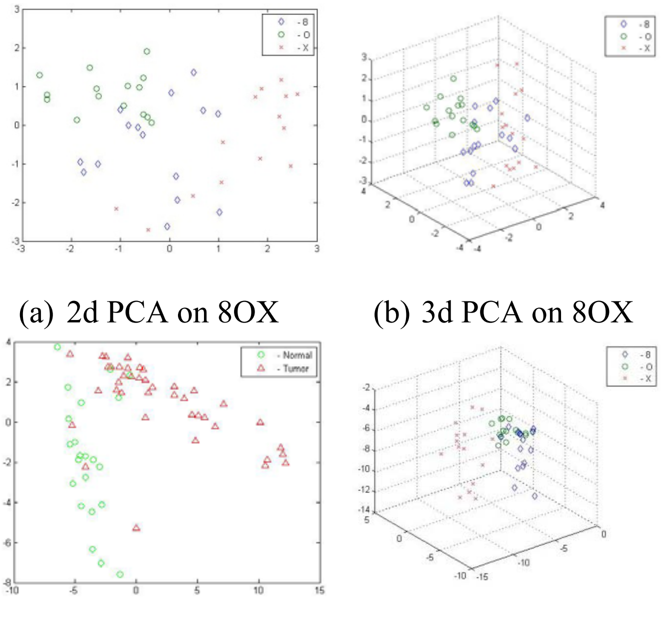
We illustrate 2D and 3D plots by PCA and LDA on the aforementioned three data sets which help visualize the clustering tendency of high-dimensional data which may also help reveal the distance of objects in different categories. To overcome the huge difference in scale between different features, we adopt a z-score transform (Jain & Dubes 1988) to convert a feature into a Gaussian-like distribution by

$$Z=(X-u)/s, \qquad (7)$$

where u is the sample mean and s is the corresponding standard deviation.

4.1 *Results on 8OX data set*

Figure 1 shows the projection plots by PCA and LDA on 8OX data set for visualization.



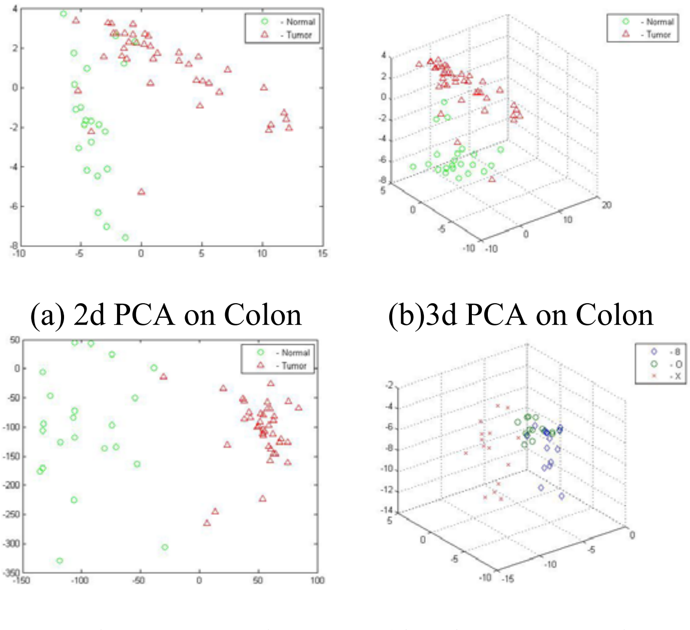
(a) 2d PCA on 8OX (b) 3d PCA on 8OX

(c) 2d LDA on 8OX (d) 3d LDA on 8OX

Figure 1. Projection by PCA and LDA on 8OX data.

4.2 *Results on Colon cancer data set*

Figure 2 shows the projection plots by PCA and LDA on colon cancer data set for visualization.



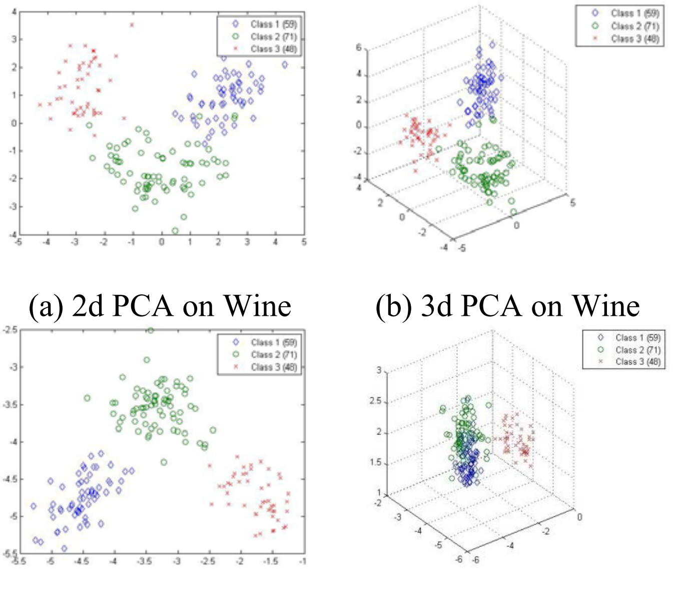
(a) 2d PCA on Colon (b)3d PCA on Colon

(c) 2d LDA on Colon (d) 3d LDA on Colon

Figure 2. Projection by PCA and LDA on Colon cancer data.

4.3 *Results on Wine data set*

Figure 3 shows the projection plots by PCA and LDA on wine data set for visualization.



(a) 2d PCA on Wine (b) 3d PCA on Wine

(c) 2d LDA on Wine (d) 3d LDA on Wine

Figure 3. Projection by PCA and LDA on the Wine data set.

4.4 *Discussion*

The visualization on the projection results by PCA and LDA demonstrates that the patterns by LDA in different categories are more separable than those obtained by PCA due to that LDA utilizes the category information but PCA does not. In computation, LDA may encounter the singularity problem of the within-class scatter matrix, whereas, the problem could be resolved by applying the preconditioning and diagonal shifting strategies (Heath 2002, Hsu 2014**)**. We adopt the latter strategy to avoid the near-singularity problem which did not occur in the aforementioned three test data sets.

5 MATLAB CODES FOR PCA AND LDA

Simple Matlab codes (Hsu 2014) for implementing PCA on colon cancer data set consisting of 62 patients (22 normal and 40 tumor cases) with 46 features in two categories are listed below. The input data set derived from (Alon et al. 1999) could be found in (De la Bastida 2015).

5.1 *Matlab code for 2d PCA on colon data set*

```
% Filename: pcacolon.m
%       PCA on colon62x46.txt
fin=fopen('colon62x46.txt','r');
d=46;   N=62;
fgetl(fin); fgetl(fin); fgetl(fin);
% read the input data
A=fscanf(fin,'%f\t',[d+1 N]);   A=A';
Z=A(:,1:d);     % remove labels in the last column
% Do z-score transform
u=mean(Z,1); s=std(Z,1);
for j=1:d
       u0=u(j); s0=s(j);
       X(:,j)=(Z(:,j)-u0)/s0;
end
K=2;   Y=PCA(X,K);     % call PCA function
X1=Y(1:22,1); Y1=Y(1:22,2);
X2=Y(23:62,1); Y2=Y(23:62,2);
plot(X1,Y1,'Og',X2,Y2,'^r');
legend('- Normal','- Tumor')
```

5.2 *Matlab code for function PCA*

```
function Y=PCA(X,K)
C=cov(X);
[U D]=eig(C);
L=diag(D);
[E index]=sort(L,'descend');
Xproj=zeros(d,K);   % initiate a projection matrix
for j=1:K
     Xproj(:,j)=U(:,index(j));
end
Y=X*Xproj;          % first K principal components
```

5.3 *Matlab code for LDA on Wine data set*

```
fin=fopen('wine178x13.txt');
d=13;   n=178;       % d features, n patterns
L(1)=59;   L(2)=130;   L(3)=178;
fgetl(fin); fgetl(fin); fgetl(fin);
A=fscanf(fin,'%f',[1+d, n]);   A=A';
n1=59;   n2=71;   n3=48;    X=A(:,1:d);
% (a) - Covariance Matrix T
X1=X(1:L(1),:);   X2=X(1+L(1):L(2),:);
X3=X(1+L(2):L(3),:);
m1=mean(X1);   m2=mean(X2);   m3=mean(X3);
mu=mean(X);    T=cov(X);
W1=cov(X1);    W2=cov(X2);    W3=cov(X3);
W=(n1-1)*W1+(n2-1)*W2+(n3-1)*W3;
B=(n1-1)*(m1-mu)'*(m1-mu)+…
     (n2-1)*(m2-mu)'*(m2-mu)+…
     (n3-1)*(m3-mu)'*(m3-mu);
s=0.0001;
C=(inv(W+s*eye(d)))*(B+eps);
% (b) - Compute Eigenvalues of W^{-1}B
[U D]=eig(C);
Lambda=diag(D);
[Cat index]=sort(Lambda,'descend');
% (c) – 3d LDA Projection for Wine data set
K=3;
Xproj=zeros(K,d);     % initiate a projection matrix
for i=1:K
     Xproj(i,:)=U(:,index(i))';
end
Y=(Xproj*X')'; %first K discriminative components
X1=Y(1:L(1),1);        Y1=Y(1:L(1),2);
Z1=Y(1:L(1),3);
X2=Y(1+L(1):L(2),1); Y2=Y(1+L(1):L(2),2);
Z2=Y(1+L(1):L(2),3);
X3=Y(1+L(2):L(3),1); Y3=Y(1+L(2):L(3),2);
Z3=Y(1+L(2):L(3),3);
plot3(X1,Y1,Z1,'d',X2,Y2,Z2,'O',X3,Y3,Z3,'X'); …
grid
legend('Class 1 (59)','Class 2 (71)','Class 3 (48)')
```

5.4 *Summary*

The result of Figure 1(c) is the output by running the Matlab code provided in sections 5.1 and 5.2 and section 5.3 lists the Matlab code for running the LDA with the output as shown in Figure 3(d).

6 CONCLUSION

The major task of this paper is to provide Matlab codes for high-dimensional data visualization by Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) with the experiments tested on three data sets, 8OX data, colon cancer data, and wine data. The results illustrate that LDA generally provides a better visualization for the clustering tendency than that obtained by PCA. However, for some applications, the category information of input data may not be available, in such cases, PCA is an alternative choice.

7 ACKNOWLEDGMENTS

This work is supported by Taiwanese Grant NSC 2221-E-007-125-MY3.

8 REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. 1999. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, 6745-6750.

De la Bastida Castillo J.D. 2013. *Software for Gene Expression Data Analysis*, M.S. Thesis, Institute of ISA, National Tsing Hua University, Hsinchu, Taiwan, May.

Bishop, C.M. 2006. *Machine Learning and Pattern Recognition*. Springer.

Jain A.K. & Dubes, R.C. 1988. *Algorithms for Clustering Data*, Prentice Hall.

Hanselman, D. & Littlefield, B. 2005. *Mastering MATLAB 7*, Pearson Prentice Hall.

Heath, M.T. 2002. *Scientific Computing: An Introductory Survey*, 2nd edition, McGraw Hill.

Hsu, A.W.L. 2014. *Principal Component Analysis and Its Applications*, M.S. Thesis, Institute of ISA, National Tsing Hua University, Hsinch, Taiwan, June.

Jolliffe, I.T. 1986. *Principal Component Analysis*, 1st edition, Springer.

WebArray 2015. http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html, consulted on March 13.

WebBigData. 2015. http://www.webopedia.com/TERM/B/big_data_analytics.html

WebLDA. 2015 http://en.wikipedia.org/wiki/Linear_discriminant_analysis, consulted on March 13.

WebPCA. 2015. http://en.wikipedia.org/wiki/Principal_component_analysis, consulted on March 13, 2015.

WebWine. 2015. http://archive.ics.uci.edu/ml/datasets/Wine, consulted on March 13, 2015.