計測自動制御学会東北支部 第 340 回研究集会 (2022.12.23)
資料番号 340-1

カーネル法による線形主成分判別分析

Kernelized Linear Principal Component Discriminant Analysis

曲凌暁 *，裴岩 **

○ Qu Lingxiao*, Pei Yan**

*会津大学 大学院コンピュータ理工学研究科,
**会津大学 コンピュータ・サイエンス部門

*Graduate School of Computer Science and Engineering, University of Aizu,
**Division of Computer Science, University of Aizu

キーワード： 主成分分析 (Principal Component Analysis)，線形判別式分析 (Linear Discriminant Analysis)，カーネル法 (Kernel method)

連絡先： 〒 965-8580 日本国福島県会津若松市一箕町鶴賀 会津大学 裴岩,
Tel.: +81-(242)-37-2765 E-mail: peiyan@u-aizu.ac.jp

1. Introduction

Principal Component Analysis (PCA) [26] and Linear Discriminant Analysis (LDA) [8] are two statistical learning algorithms using coordinate transformation for dimension reduction of high-dimensional data. They are applied widely in feature extraction, data classification, and clustering problems.

PCA is an unsupervised learning method, pursuing some projection directions with the maximum sum of the total variance of projected data. It is useful for feature mining and extraction by reducing higher dimensional data to lower dimensional data [1]. LDA is a supervised learning method with the objective that finding the projection direction with the maximum distance of projections of class means and the minimum distance between the projections of sample data in each class and the projections of corresponding class mean [12]. LDA has good performance in data classification and clustering problems.

PCA and LDA are used to solve linear problems in low-dimensional space. Kernel methods [10] can be used to increase data dimension, and transform the nonlinear models in low dimensional space to linear ones in high dimensional space. It simplifies the complexity of problems using the transformation. There is research on kernel PCA [21], referred to as KPCA in the following, and Generalized Discriminant Analysis (GDA) [3]. These researches perform PCA and LDA in high dimensional space by using kernel methods to handle nonlinear problems [17].

KPCA has many extensions of its techniques,

algorithms, and applications. For example, sparse kernel PCA [23], robust kernel PCA [18], incremental kernel PCA [4], adaptive kernel PCA [6], and streaming kernel PCA [9], etc. KPCA has been applied in kinds of scenes, including face recognition [14], image modling [13], fault detaction [5], and geostatistics [20], etc. There are also many pieces of research on GDA, which apply GDA with different approaches and extend GDA to multiple situations. For instance, GDA based on distance [2], GDA of matrix exponential approach [29], modified GDA preventing eigenvalue degenerating [30], as well as GDA with generalized singular value composition [11], etc. In addition, the applications of GDA to tree-structured classification [15], feature extraction with DNN [22], face recognition [16], and under-sampled problems [27, 28], etc., have been hot topics for a long time.

Pei proposed a framework of data analysis methods called Linear Principal Component Discriminant Analysis (LPCDA) using the same characteristics of PCA and LDA [19]. It has the advantages of both supervised and unsupervised learning methods. In this paper, we extend LPCDA to solve nonlinear problems in high dimensional space using kernel methods. It can be used for both classification and clustering problems, so we refer to it as a series of semi-supervised learning methods. We propose a framework of data analysis methods for nonlinear problems in high dimensional space by combining three objectives of KPCA and GDA as follows,

1) pursuing projection vectors of the maximum total variance of projected data in feature space;

2) pursuing projection vectors of maximum sum of distances between projected class means in feature space;

3) pursuing projection vectors of minimum sum of distances between projected class data in each class and the corresponding projected class mean in feature space.

It is referred to as Kernelized Linear Principal Component Discriminant Analysis (KLPCDA). The implementation of kernel extension in LPCDA and the presence of varieties of data analysis methods in high dimensional space is the originality of this work.

The following of this paper is organized as below. We illustrate the theoretical implementation of our proposed KLPCDA in the section "Kernelized Linear Principal Component Discriminant Analysis". We show our experiment and evaluation in the section "Evaluation". The analysis and discussion of evaluation results and the problems that remained are presented in the next section. Finally, we conclude the whole work and present the potential study subjects.

2. Kernelized Linear Principal Component Discriminant Analysis

A series of data analysis methods called Linear Principal Component Discriminant Analysis, which is a uniform framework based on Principal Component Analysis and Linear Discriminant Analysis [19], was proposed, as shown in Table 1. In this work, we establish the kernel implementation of proposed methods in Linear Principal Component Discriminant Analysis, by using the kernel trick. It pursues a projected direction v in feature space H and sat-

isfies: (a) with the maximum total variance of projected samples [21]; (b) with the maximum sum of the distance of the projected group's center points [3]; and (c) with minimum sum of the inner variance of the projected groups [3].

Given n training samples of column vectors $x_1, x_2, \cdots, x_n \in R^d$ in original space.

Given $n = \sum_{i=1}^{L} n_i$ training samples f column vectors with labels $1, \cdots, L$: $x_1^{(1)}, \cdots, x_{n_1}^{(1)}, \cdots, x_j^{(i)}, \cdots, x_{n_L}^{(L)}$ in original space, where $x_j^{(i)} \in R^d$, $i \in [1, L]$, $j \in [1, n_i]$ is the j-th sample in class i. n_i and L are the number of i-th class and total labels, respectively. They are mapped into a feature space and referred respectively as $X^T = \left[\phi(x_1), \phi(x_2), \cdots, \phi(x_n)\right]$, and $X^T = \left[\phi(x_1^{(1)}), \cdots, \phi(x_{n_1}^{(1)}), \cdots, \phi(x_j^{(i)}), \cdots, \phi(x_{n_L}^{(L)})\right]$.

The objective (a) is presented by Eq. (1). The objective (b) is presented by Eq. (2) and (3), and (c) is presented by Eq. (4) and (5). In these equations, m_i, m_j and m represent the means of the projections of class i, j and all samples, respectively. $\mathbf{m_i}$ and $\mathbf{m}$ represent the means of class i and all samples. These objectives are all implemented in the feature space. Baudat etc. didn't consider each group's center in objective (c) [3]. We consider it and pursue the distance between projections of each group's samples and this group's center is the minimum, as shown in Eq.s (4) and (5).

$$v = \arg \max_{v \in R^d, ||v||=1} \sigma^2 = \arg \max_{v \in R^d, ||v||=1} v^T C v. \tag{1}$$

$$\begin{aligned} v &= \arg \max_{v \in R^d, ||v||=1} \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{n_i}{n} \frac{n_j}{n} (m_i - m_j)^2 \\ &= \arg \max_{v \in R^d, ||v||=1} v^T S_b v. \end{aligned} \tag{2}$$

Table 1 Methods of principal component discriminant analysis. There are three objectives in LPCDA [19], i.e., $\arg\max v^T C v$, $\arg\max v^T S_b v$, and $\arg\min v^T S_w v$. With the combinations of these three objectives, there are seven methods in LPCDA, making up the proposed uniform framework, containing the PCA and LDA. This table is adopted from reference [19].

Method No.	target function	meaning
1	$\frac{v^T C v + v^T S_b v}{v^T S_w v}$	LPCDA: $\frac{v^T C v + v^T S_b v}{v^T S_w v}$
2	$v^T C v + v^T S_b v$	LPCDA: $v^T C v + v^T S_b v$
3	$\frac{v^T S_b v}{v^T S_w v}$	linear discriminant analysis
4	$v^T C v$	principal component analysis
5	$\frac{v^T C v}{v^T S_w v}$	LPCDA: $\frac{v^T C v}{v^T S_w v}$
6	$v^T S_b v$	LPCDA: $v^T S_b v$
7	$v^T S_w v$	LPCDA: $v^T S_w v$

$$S_b = \sum_{i=1}^{L} \frac{n_i}{n} (\mathbf{m_i} - \mathbf{m})(\mathbf{m_i} - \mathbf{m})^T. \tag{3}$$

$$\begin{aligned} v &= \arg \min_{v \in R^d, |v|=1} \sum_{i=1}^{L} \frac{n_i}{n} \sum_{l=1}^{n_i} (v^T \phi(x_l^{(i)}) - m_i)^2 \\ &= \arg \min_{v \in R^d, |v|=1} v^T S_w v. \end{aligned} \tag{4}$$

$$S_w = \sum_{i=1}^{L} \frac{n_i}{n} \sum_{l=1}^{n_i} (\phi(x_l^{(i)}) - \mathbf{m_i})(\phi(x_l^{(i)}) - \mathbf{m_i})^T. \tag{5}$$

We will introduce the theoretical implementation of kernelizing five out of seven methods in Table 1, where No.3 and No.4 methods have been kernelized before, that is kernel PCA and GDA, respectively. We weight the objectives (a), (b), and (c) by parameter α, β, and γ, that is, $\alpha v^T C v$, $\beta v^T S_b v$, $\gamma v^T S_w v$, respectively. In the following deduction, we will transform the combinations of three objectives to eigenvalue problems containing matrices C, S_b, S_w by La-

grange multiplier. However, in feature space,

$$C = \frac{1}{n} X^T X,\ S_b = X^T B X,\ S_w = X^T W X, \tag{6}$$

are unknown, where

$$B = \frac{1}{n} diag(\frac{1}{n_1}\mathbf{1}_{n_1\times n_1}, \cdots, \frac{1}{n_L}\mathbf{1}_{n_L\times n_L}) - \frac{1}{n^2}\mathbf{1}_{n\times n},$$
$$W = diag(P_1 I_{n_1}, \cdots, P_L I_{n_L}) - diag(\frac{1}{n}\mathbf{1}_{n_1\times n_1}, \cdots, \frac{1}{n}\mathbf{1}_{n_L\times n_L}),$$

can be known [3, 21]. $\mathbf{1}_{n_i\times n_i}, i \in [1, L]$ are a $n_i \times n_i$ matrix with all entries 1. In order to solve the eigenvalue problems, we have to apply Wanba theory [25]

$$v = X^T \alpha, \tag{7}$$

and kernel matrix $K = XX^T$ to the deduction [10]. Schölkopf etc. [21] implemented centering in high-dimensional space by computing centered kernel matrix $\widetilde{K}$ from K,

$$\widetilde{K} = K - \frac{1}{n}\mathbf{1}_{n\times n}K - \frac{1}{n}K\mathbf{1}_{n\times n} - \frac{1}{n^2}\mathbf{1}_{n\times n}K\mathbf{1}_{n\times n},$$

where $\mathbf{1}_{n\times n}$ are a $n\times n$ matrix with all entries 1. We replace K with the centered $\widetilde{K}$ to operate the centering in practice.

2.1 Method No.1

The target function is $v = \arg\max \frac{\alpha v^T C v + \beta v^T S_b v}{\gamma v^T S_w v}$. To simplify the calculation, we set α, β, and $\gamma = 1$ in the following. We can solve the problem by the Lagrange multiplier. About the objective (c), we set it as a constraint $||v^T S_w v|| = 1$ in the following calculation. The objective of Method No.1 can be presented by $v = \arg\max_{||v^T S_w v||=1} v^T C v + v^T S_b v$. From Lagrange multiplier, it can be deduced as follows,

$$\begin{aligned} f(v,\lambda) &= v^T C v + v^T S_b v - \lambda(v^T S_w v - 1), \\ \frac{\partial f}{\partial v} &= 2Cv + 2S_b v - 2\lambda S_w v = 0, \\ \frac{\partial f}{\partial \lambda} &= v^T S_w v - 1 = 0. \end{aligned} \tag{8}$$

From Eq. (8), we obtain

$$(C + S_b)v = \lambda S_w v, \tag{9}$$

which is a generalized eigenvalue problem. The objective can be deduced as

$$\begin{aligned} v &= \arg \max_{||v^T S_w v||=1} v^T C v + v^T S_b v \\ &= \arg \max_{||v^T S_w v||=1} \lambda. \end{aligned}$$

Therefore, the optimal v corresponds to the maximum eigenvalue λ of Eq. (9). From Eq. (6), and Wahba Theory Eq. (7), with kernel matrix $K = XX^T$, the eigenvalue problem (9) can be dealt as below.

$$\begin{aligned} (\frac{1}{n}X^T X + X^T B X)v &= \lambda X^T W X v, \\ \frac{1}{n}XX^T X v + XX^T B X V &= \lambda XX^T W X v, \\ \frac{1}{n}XX^T XX^T \alpha + XX^T BXX^T \alpha &= \lambda XX^T WXX^T \alpha, \\ (WK)^{-1}(\frac{1}{n}K + BK)\alpha &= \lambda\alpha. \end{aligned} \tag{10}$$

In Eq. (10), matrices K, W, B, and parameter n are all known. Hence we can solve this eigenvalue problem and find eigenvalues λ and eigenvectors α. We suppose $||v|| = 1$ for target $v = X^T\alpha$, so it has to be normalized as

$$v = \frac{X^T\alpha}{||X^T\alpha||} = \frac{X^T\alpha}{\sqrt{\alpha^T K \alpha}}.$$

Suppose there are new data x', its projections onto v in a feature space are

$$v^T\phi(x') = \frac{1}{\sqrt{\alpha^T K\alpha}}\alpha^T \begin{bmatrix} k(x_1^{(1)}, x') \\ \vdots \\ k(x_{n_1}^{(1)}, x') \\ \vdots \\ k(x_1^{(L)}, x') \\ \vdots \\ k(x_{n_L}^{(L)}, x') \end{bmatrix}.$$

2.2 Method No.2

The target function of this method is $v = \arg\max \alpha v^T Cv + \beta v^T S_b v$. With Lagrange multiplier, it's objective can be presented as $v = \arg\max_{v\in R^d, ||v||=1} v^T Cv + v^T S_b v$, and can be calculated to obtain this eigenvalue problem,

$$(C + S_b)v = \lambda v. \tag{11}$$

The objective can be simplified as

$$\begin{aligned} v &= \arg \max_{v\in R^d, ||v||=1} v^T Cv + v^T S_b v \\ &= \arg \max_{v\in R^d, ||v||=1} \lambda. \end{aligned}$$

The target v corresponds to the largest λ of Eq. (11). By applying kernel matrix, Eq. (6), and Eq. (7), we can transform Eq. (11) and solve the eigenvalue problem Eq. (12).

$$\begin{aligned} (\frac{1}{n}X^T X + X^T BX)v = \lambda v, \\ (\frac{1}{n}K + BK)\alpha = \lambda\alpha. \end{aligned} \tag{12}$$

2.3 Method No.3

Method NO.3 is GDA and its implementation in feature space using the kernel approach has been finished in the year 2000 [3]. While the authors didn't handle the centering in feature space and didn't take the distance of projections between class samples and the corresponding class means into consideration, which is dealt with by us in the following.

The target function is $v = \arg\max \frac{\beta v^T S_b v}{\gamma v^T S_w v}$. As mentioned in Eq. (5), it contains the class means. The objective can be presented as $v = \arg\max_{||v^T S_w v||=1} v^T S_b v$. By Lagrange multiplier,

$$S_b v = \lambda S_w v. \tag{13}$$

With kernel matrix, Eq. (6), and Eq. (7), we calculate Eq. (13) and obtain the solvable eigenvalue problem Eq. (14).

$$\begin{aligned} X^T BXv = \lambda X^T WXv, \\ (WK)^{-1}BK\alpha = \lambda\alpha. \end{aligned} \tag{14}$$

2.4 Method No.4

Method No.4 is kernel PCA and it was implemented in 1996 [21]. The target function is $v = \arg\max \alpha v^T Cv$. By Lagrange multiplier, the objective can be presented as $v = \arg\max_{v\in R^d, ||v||=1} v^T Cv$, and be transformed to the eigenvalue equation Eq. (15)

$$\frac{1}{n}X^T Xv = \lambda v. \tag{15}$$

From Eq. (6) and (7), and kernel matrix, Eq. (15) can be deduced a solvable one Eq. (16).

$$\frac{1}{n}K\alpha = \lambda\alpha. \tag{16}$$

2.5 Method No.5

For this method, the target function is $v = \arg\max \frac{\alpha v^T Cv}{\gamma v^T S_w v}$. From Lagrange multiplier, the objective can be denoted as $v = \arg\max_{||v^T S_w v||=1} v^T Cv$. By further calculation, we obtain

$$Cv = \lambda S_w v. \tag{17}$$

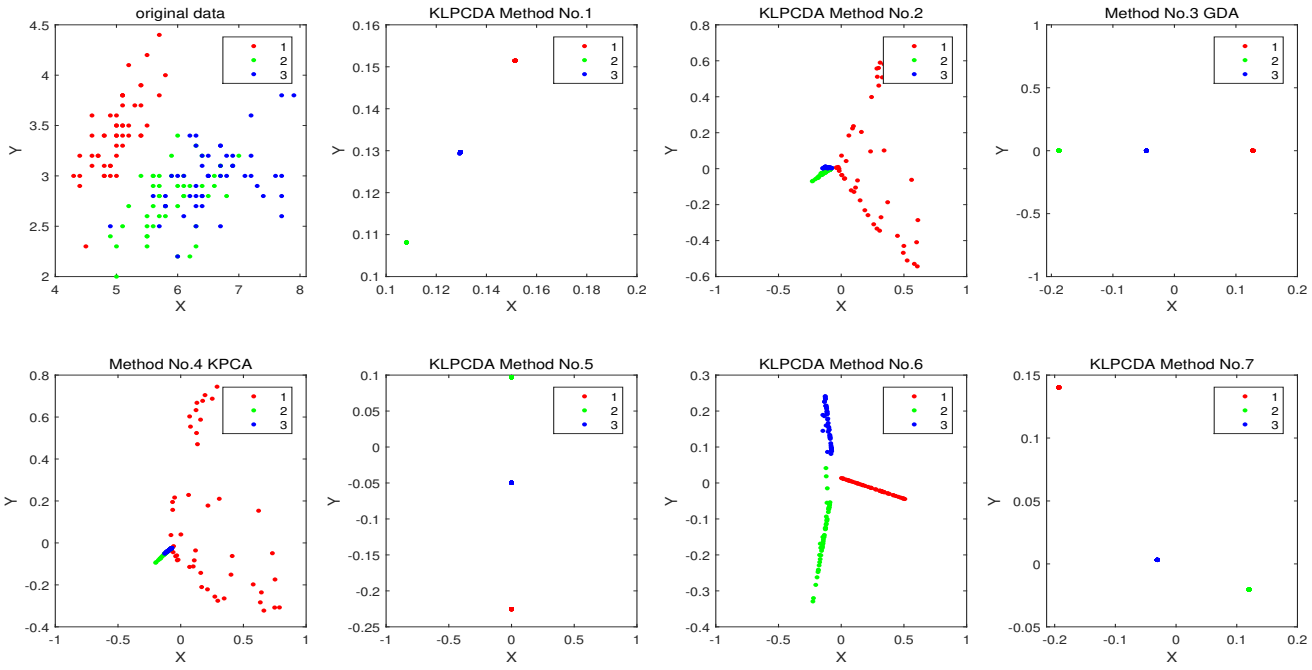

Fig. 1 Projections of seven kernelized data analysis methods on Fisher Iris data set by using Gaussian kernel with $\sigma = 0.2$. The labels $1, 2, 3$ correspond to the classes Iris-setosa, Iris-versicolor, and Iris-virginica, respectively. The X axis is the first principal component, and the Y axis is the second principal component.

With Eq. (6), kernel matrix, and Wahba Representer Theory, we deduce Eq. (17) and obtain the eigenvalue equation (18).

$$\frac{1}{n}X^T X v = \lambda X^T W X v,$$
$$\frac{1}{n}(WK)^{-1}K\alpha = \lambda\alpha. \tag{18}$$

We can work out α and λ.

2.6 Method No.6

The target function of this method is $v = \arg\max \beta v^T S_b v$. With the Lagrange multiplier, we can present and deduce our objective as $v = \arg\max_{v\in\mathbb{R}^d, ||v||=1} v^T S_b v$, and the eigenvalue equation

$$S_b v = \lambda v. \tag{19}$$

We transform the above eigenvalue equation Eq. (19) to the following solvable one with the application of Eq. (6) and (7), and kernel matrix.

$$X^T B X X^T \alpha = \lambda X^T \alpha,$$
$$BK\alpha = \lambda\alpha. \tag{20}$$

We can solve the eigenvalue problem Eq. (20).

2.7 Method No.7

The target function of this method is $v = \arg\min v^T S_w v$. By Lagrange multiplier, the objective can be represented as $v = \arg\min_{v\in\mathbb{R}, ||v||=1} v^T S_w v$, and be deduced to

$$S_w v = \lambda v. \tag{21}$$

With the usage of Wahba Representer Theory, the kernel matrix, and Eq. (6), we deduce the

Table 2 Total variance of projected samples, each projected inner-class variance, and sum of square-distances between projected class-means conducted on Fisher Iris data set of Gaussian kernel, $\sigma = 0.2$, for each method. The bold data shows the optimal results of same items among seven methods.

Method No.	Total variance	Total variance of class 1	Total variance of class 2	Total variance of class 3	distance between two classes
1	6.2901e-04	7.6961e-11	1.8332e-10	9.2548e-09	0.0056
2	0.0814	0.1562	0.0016	4.4162e-04	0.2624
3	0.0167	1.7943e-11	4.2740e-11	2.1577e-09	0.1496
4	**0.0830**	0.1749	0.001279	3.4076e-04	0.2227
5	0.0175	2.2640e-13	5.3928e-13	2.7225e-11	0.1563
6	0.0574	0.0235	0.0075	0.0030	**0.4129**
7	0.0216	**4.4864e − 31**	**1.4725e − 31**	**2.0750e − 31**	0.1927

Table 3 Total variance of projected samples, each projected inner-class variance, and sum of square-distances between projected class-means conducted on Fisher Iris data set of Polynomial kernel, $r = 3$, for each method. The bold data shows the optimal results of same items among seven methods.

Method No.	Total variance	Total variance of class 1	Total variance of class 2	Total variance of class 3	distance between two classes
1	1.6533e-04	2.2965e-06	2.8006e-05	8.4096e-05	0.0011
2	1.0052e+05	714.0925	1.5101e+04	4.8193e+04	7.1050e+05
3	1.2196e-08	9.3949e-11	1.8823e-09	5.9085e-09	8.5849e-08
4	1.0060e+05	778.7099	1.5358e+04	4.8745e+04	7.0861e+05
5	9.2758e+04	179.2304	1.1506e+04	4.0608e+04	6.7552e+05
6	**1.0142e + 05**	1.4489e+03	1.5166e+04	4.8512e+04	**7.1525e + 05**
7	7.5838e-25	**1.1093e − 25**	**2.2541e − 25**	**7.9985e − 25**	3.1617e-24

eigenvalue equation (21) and obtain

$$X^T W X v = \lambda v,$$
$$WK\alpha = \lambda\alpha. \qquad (22)$$

The eigenvalue problem Eq. (22) is solvable.

Now we finish the theoretical deduction of our proposed five methods, kernel PCA and GDA.

3. Evaluation

In this section, we evaluate our proposed KLPCDA using Fisher Iris [7] and wine [24] data sets. We choose two kernel methods, Gaussian kernel

$$k(x,z) = exp(\frac{-|x-z|^2}{2\sigma^2}),$$

and Polynomial kernel

$$k(x,z) = (< x,z > +1)^r.$$

By calculating the total variance, the total variance of each class, and the distance between projections of each class mean, we evaluate each method visually and quantitatively.

3.1 Evaluation on Fisher Iris Data

In this data set, there are 150 instances with four attributes: sepal length, sepal width, petal length, and petal width. It contains 3 classes of 50 instances each: Iris-setosa, Iris-versicolor, and Iris-virginica. In this data set, one class is linearly separable from the other two and the latter are not linearly separable from each other. In the following evaluations, all the samples are centered in feature space by a centered kernel matrix.

Figure 1 and Figure 2 show the classification results of projections of the samples onto the first two principal component axes of five methods in our proposed KLPCDA, KPCA, and GDA on Fisher Iris data set by using Gaussian kernel and Polynomial kernel, respectively. Table 2 and Table 3 present the quantitative

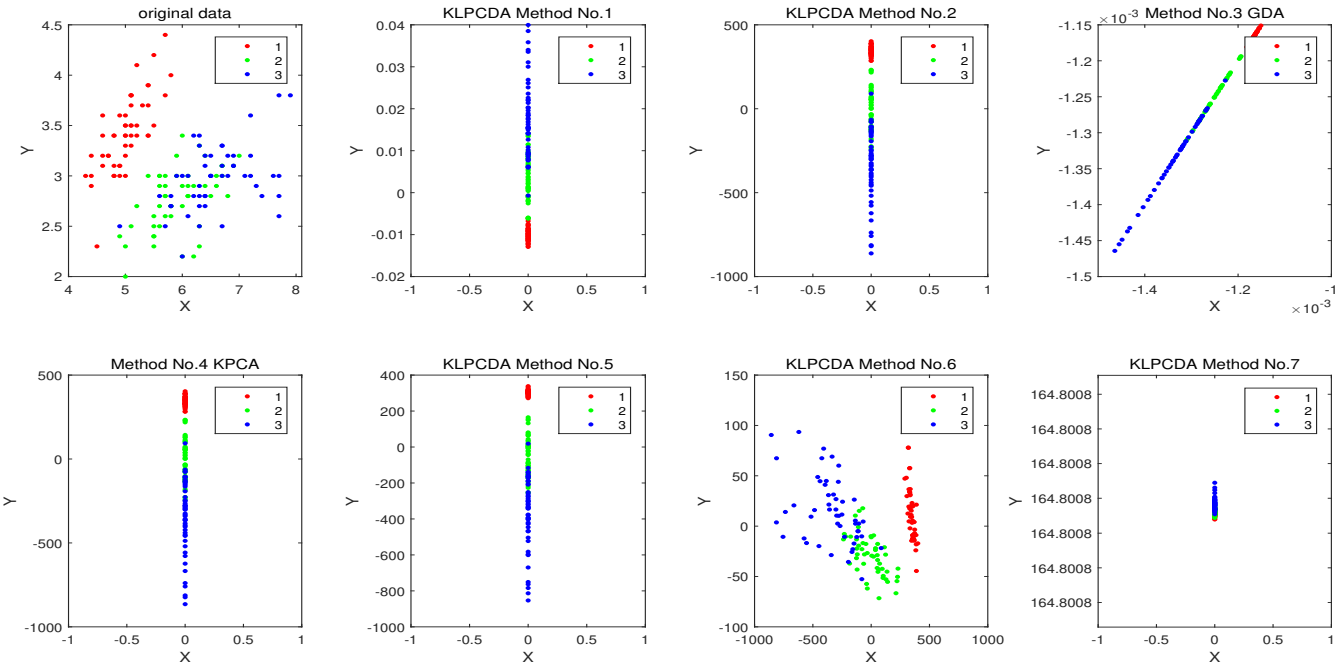

Fig. 2 Projections of seven kernelized data analysis methods on Fisher Iris data set by using Polynomial kernel with $r = 3$. The labels $1, 2, 3$ corresponds to the classes Iris-setosa, Iris-versicolor, and Iris-virginica, respectively. The X axis is the first principal component, and the Y axis is the second principal component.

evaluations of seven methods in KLPCDA with Gaussian kernel and Polynomial kernel, respectively.

3.2 Evaluation on Wine Data

In this dataset, there are 178 wine samples of three classes labeled 1, 2, and 3. It has thirteen attributes: 1) Alcohol, 2) Malic acid, 3) Ash, 4) Alcalinity of ash, 5) Magnesium, 6) Total phenols, 7) Flavanoids, 8) Nonflavanoid phenols, 9) Proanthocyanins, 10) Color intensity, 11) Hue, 12) OD280/OD315 of diluted wines, and 13) Proline. Class 1, 2, and 3 have 59, 71, and 48 samples, respectively.

Figure 3 and Figure 4 present the visual evaluation of our proposed five methods in KLPCDA, GDA, and KPCA, conducted on Wine data set by using Gaussian kernel and Polynomial kernel, respectively. Table 4 and Table 5 present the quantitative evaluation of seven methods in KLPCDA with Gaussian kernel and Polynomial kernel, respectively.

4. Analysis and Discussion

4.1 Discussion on Results from Fisher Iris Data

In Figure 1 of Gaussian kernel on Fisher Iris dataset, we choose parameter $\sigma = 0.2$. Our proposed methods NO.1, NO.5, NO.6, and NO.7 have perfect classification results as GDA, all of which completely separate three classes. Method NO.2 has a similar but a little bit better result than KPCA. Each of our proposed methods can classify data clearly and perform not inferior and even better than GDA

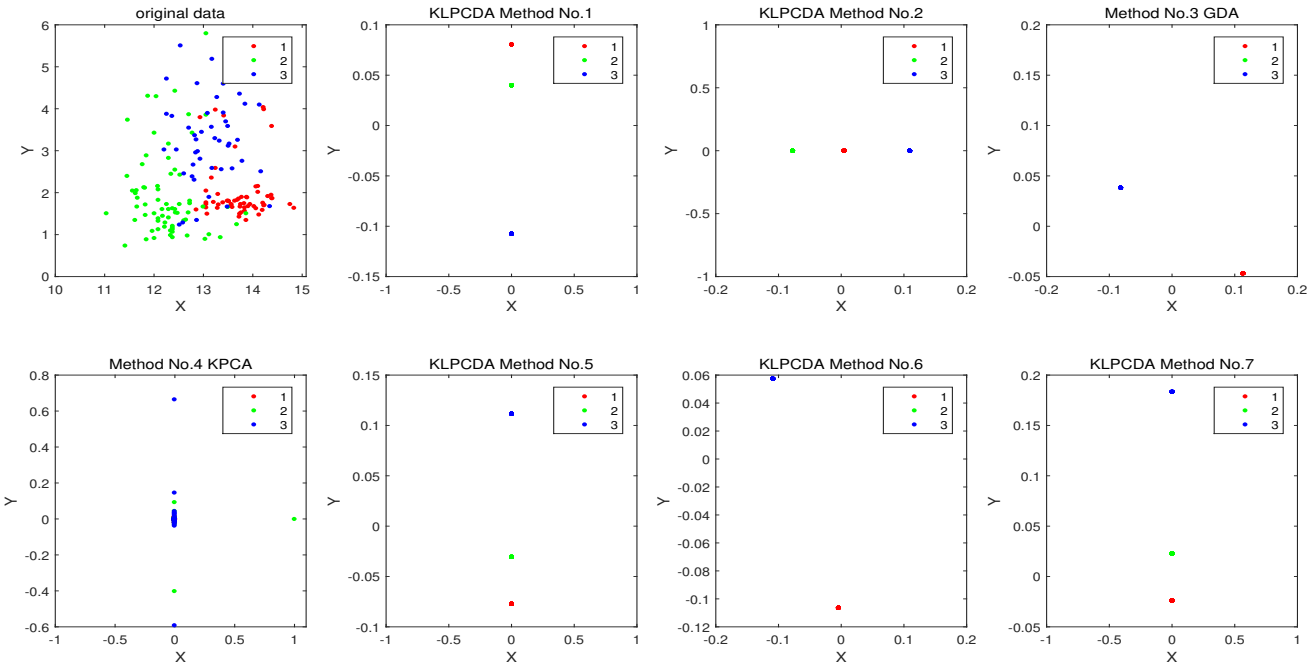

Fig. 3 Projections of seven kernelized data analysis methods on Wine data set by using Gaussian kernel with $\sigma = 0.2$. The X axis is the first principal component, and the Y axis is the second principal component.

or KPCA. Table 2 presents the quantitative evaluation results, from which we know KPCA has the maximum total variance which is tightly larger than that of our proposed method NO.2. Our proposed method NO.7 has the minimum sum of each inner-class variance. Our proposed method NO.6 has the maximum sum of square distances between projected class means. Our proposed methods NO.5 and NO.7 have better values on every evaluation metric item than GDA, and our proposed methods have better values on some of the evaluation metric items than KPCA.

Figure 2 and Table 3 show the evaluation results conducted on the Fisher Iris data set with the polynomial kernel, where we set the parameter as $r = 3$. From the figure, we can visually observe that the classification results of methods NO.1, NO.2, and NO.5 are nearly the same as that of KPCA, which have different projection directions from that of GDA. In general, our proposed methods NO.1, NO.2, NO.5, and NO.6 have not a bad performance but are inferior to that of using a Gaussian kernel. Meanwhile, these methods have similar classification effects to KPCA and GDA. From Table 3, our proposed method NO.6 has the maximum total variance of projected data and sum of square distances between projected class means, while method NO.7 has the minimum sum of each projected inner-class variance. Consequently, it indicates that our proposed method NO.7 has advantages over GDA and KPCA on inner-class variances. And method NO.6 has better values on evaluation metric items of total variances and the sum of the square distance between class means than GDA and KPCA.

Table 4 Total variance of projected samples, each projected inner-class variance, and sum of square-distances between projected class-means conducted on Wine data set of Gaussian kernel, $\sigma = 0.2$, for each method. The bold data shows the optimal results of same items among seven methods.

Method No.	Total variance	Total variance of class 1	Total variance of class 2	Total variance of class 3	distance between two classes
1	0.0057	4.5842e-17	**2.47689e − 04**	0.0053	0.0154
2	0.0056	**2.4168e − 32**	0.0010	0.0085	0.0082
3	**0.0146**	6.1464e-16	0.0058	0.0121	**0.0616**
4	0.0113	3.1197e-05	0.0201	0.0202	8.2849e-04
5	0.0057	4.3761e-17	3.2754e-04	**0.0049**	0.0164
6	0.0113	2.4011e-31	0.0047	0.0086	0.0508
7	0.0069	1.7207e-31	3.2759e-04	0.0063	0.0190

Table 5 Total variance of projected samples, each projected inner-class variance, and sum of square-distances between projected class-means conducted on Wine data set of Polynomial kernel, $r = 3$, for each method. The bold data shows the optimal results of same items among seven methods.

Method No.	Total variance	Total variance of class 1	Total variance of class 2	Total variance of class 3	distance between two classes
1	1.9676e+14	2.6755e+14	1.1996e+14	5.3108e+12	8.4263e+14
2	7.0470e+17	9.5720e+17	4.3000e+17	1.9086e+16	3.0195e+18
3	3.3206e+03	4.5179e+03	2.0235e+03	89.4830	1.4216e+04
4	7.0471e+17	9.5738e+17	4.2994e+17	1.9075e+16	3.0193e+18
5	0	0	0	0	0
6	**7.0519e + 17**	9.5806e+17	4.3063e+17	1.9208e+16	**3.0197e + 18**
7	1.8822e-11	**1.3672e − 11**	**5.9977e − 12**	**5.4393e − 12**	6.8830e-11

4.2 Discussion on Results from Wine data

Figure 3 shows the results on Wine data set of Gaussian kernel with parameter $\sigma = 0.2$. Every of our proposed five methods as well as GDA has great classification results. It indicates that our proposed methods perform much better than KPCA. As the quantitative evaluation results in Table 4 show, GDA has the maximum total variance of projected samples and the sum of square distances between projected class means, while our proposed methods NO.2, NO.1, and NO.5 have the minimum sum of inner-class variances of projected classes 1, 2, and 3, respectively. In addition, the proposed method NO.6 receives better values than KPCA on all of the evaluation metric items. Meanwhile, except for the total variance item, all of our proposed methods have better values than KPCA, quantitatively.

From Figure 4 obtained from the Wine data set of the polynomial kernel with parameter $r = 3$, all the results are not good. During our calculation, we met complex numbers and dealt with them by deleting the image part and leaving the real part. It leads to that the data of method NO.5 are all zeros, which can not be taken into consideration. Table 5 shows that our proposed method NO.6 has the maximum total variance and the sum of the square distance between classes and method NO.7 has the minimum sum of inner-class variances. It illustrates that our proposed methods NO.6 and NO.7 have advantages over GDA and KPCA, quantitatively.

4.3 Discussion on Observations from Evaluation

Here, we make some analysis of the observation from the evaluation results. First, the evaluation performance of the Gaussian kernel

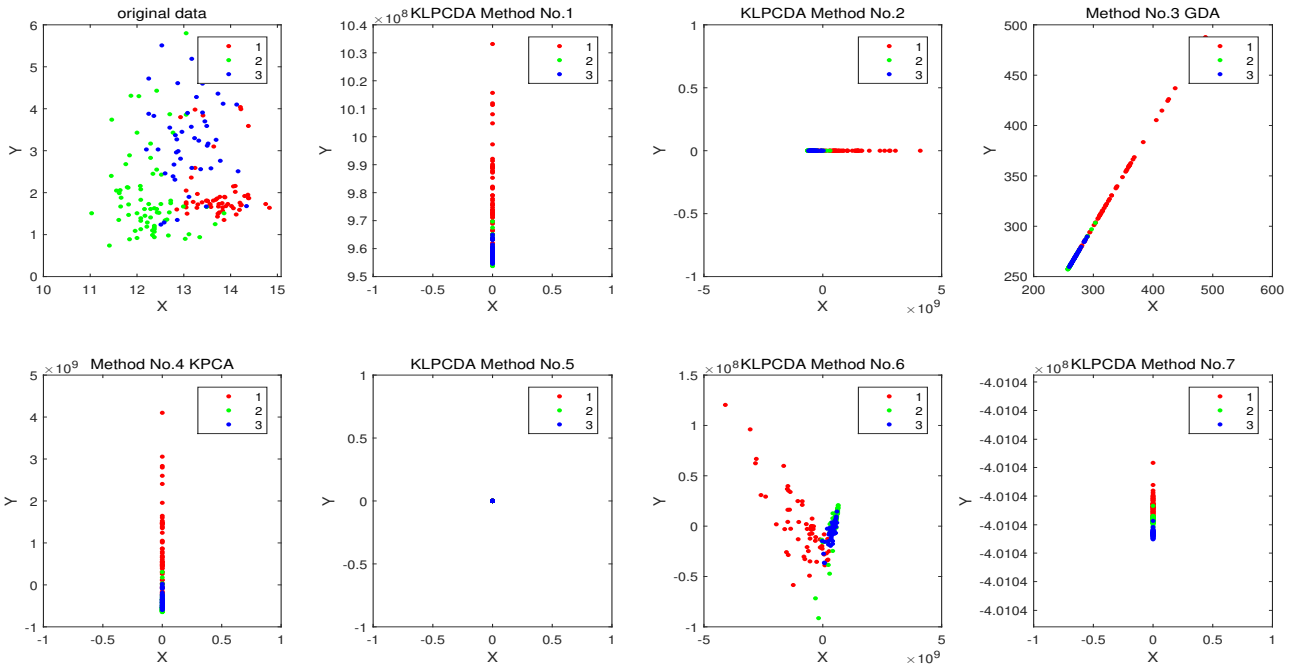

Fig. 4 Projections of seven kernelized data analysis methods on Wine data set by using Polynomial kernel with $r = 3$. The X axis is the first principal component, and the Y axis is the second principal component.

is better than that of the Polynomial kernel no matter conducting on the Fisher Iris data set or Wine data set. While there comes one problem if the parameter $\sigma = 0.2$ of the Gaussian kernel is optimal, whether the polynomial kernel performs inferior to the Gaussian kernel with any parameter. We will work on this parameter optimization problem in the future.

Second, some of our proposed methods have better performance and briefer deduction processes than GDA and KPCA under the same conditions. It can be used to replace GDA or KPCA under certain circumstances for simpler calculations. Moreover, some methods have so close projection directions that we can make use of to choose more suitable and simpler methods for different problems.

Third, we proposed five methods from the linear combination of three objectives of KPCA and GDA. We will explore other combination ways according to the quantitative evaluation results by contrasting the advantages and properties of different methods, to obtain more effective data analysis methods.

Fourth, the projection directions and quantitative data results can tell us the differences and changes in the inner relationships of data before and after being projected into higher dimensional space. It can help us to inquire into the possibility of kernelizing data analysis methods in other spaces.

5. Conclusion

In this paper, we extended the five data analysis methods proposed by Pei [19] named Linear Principal Component Discriminant Analysis to high dimensional space by mapping the data into a feature space and applying kernel

methods to handle mapped data. We called it Kernelized Linear Principal Component Discriminant Analysis. It combines supervised and unsupervised learning methods so that we can refer to our KLPCDA as semi-supervised learning methods to be used for both classification and clustering problems. We evaluate two data sets and use two kernel functions, from which we illustrate the advantages of our proposed methods over KPCA and GDA.

In the future, we will continue to focus on solving three problems about our proposed KLPCDA. The first one is the parameter optimization issue. We will explore the optimal parameters of each kernel function resulting in the optimal evaluation results and attempt to exhibit the changing trend of performance as parameters change visually. This is a significant work of practical application. The second one is investigating other possible combinations or creative ways of the three objectives of KPCA and GDA, to obtain new kernelized data analysis methods with lower complexity and better classification or clustering performance or any other advantages. The last one is to apply our methods to the real scene, by using more realistic data sets to solve more practical problems and achieve more meaningful and effective results. These subjects will be involved in our future work.

参考文献

1) Abdi, H.; and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4): 433–459.

2) Anderson, M. J.; and Robinson, J. 2003. Generalized discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics*, 45(3): 301–318.

3) Baudat, G.; and Anouar, F. 2000. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10): 2385–2404.

4) Chin, T.-J.; and Suter, D. 2007. Incremental kernel principal component analysis. *IEEE transactions on image processing*, 16(6): 1662–1674.

5) Cui, P.; Li, J.; and Wang, G. 2008. Improved kernel principal component analysis for fault detection. *Expert Systems with Applications*, 34(2): 1210–1219.

6) Ding, M.; Tian, Z.; and Xu, H. 2010. Adaptive kernel principal component analysis. *Signal processing*, 90(5): 1542–1553.

7) Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188.

8) Fukunaga, K. 2013. *Introduction to statistical pattern recognition*. Elsevier.

9) Ghashami, M.; Perry, D. J.; and Phillips, J. 2016. Streaming kernel principal component analysis. In *Artificial intelligence and statistics*, 1365–1374. PMLR.

10) Hofmann, T.; Schölkopf, B.; and Smola, A. J. 2008. Kernel methods in machine learning. *The annals of statistics*, 36(3): 1171–1220.

11) Howland, P.; and Park, H. 2004. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE transactions on pattern analysis and machine intelligence*, 26(8): 995–1006.

12) Izenman, A. J. 2013. Linear discriminant analysis. In *Modern multivariate statistical techniques*, 237–280. Springer.

13) Kim, K. I.; Franz, M. O.; and Scholkopf, B. 2005. Iterative kernel principal component analysis for image modeling. *IEEE transactions on pattern analysis and machine intelligence*, 27(9): 1351–1366.

14) Kim, K. I.; Jung, K.; and Kim, H. J. 2002. Face recognition using kernel principal component analysis. *IEEE signal processing letters*, 9(2): 40–42.

15) Loh, W.-Y.; and Vanichsetakul, N. 1988. Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83(403): 715–725.

16) Lu, J.; Plataniotis, K. N.; and Venetsanopoulos, A. N. 2003. Face recognition using kernel direct discriminant analysis algorithms. *IEEE transactions on Neural Networks*, 14(1): 117–126.

17) Mika, S.; Schölkopf, B.; Smola, A.; Müller, K.-R.; Scholz, M.; and Rätsch, G. 1998. Kernel PCA and de-noising in feature spaces. *Advances in neural information processing systems*, 11.

18) Nguyen, M.; and Torre, F. 2008. Robust kernel principal component analysis. *Advances in Neural Information Processing Systems*, 21.

19) Pei, Y. 2015. Linear principal component discriminant analysis. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2108–2113. IEEE.

20) Sarma, P.; Durlofsky, L. J.; and Aziz, K. 2008. Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics. *Mathematical Geosciences*, 40(1): 3–32.

21) Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5): 1299–1319.

22) Stuhlsatz, A.; Lippel, J.; and Zielke, T. 2012. Feature extraction with deep neural networks by a generalized discriminant analysis. *IEEE transactions on neural networks and learning systems*, 23(4): 596–608.

23) Tipping, M. 2000. Sparse kernel principal component analysis. *Advances in neural information processing systems*, 13.

24) Vandeginste, B. 1990. PARVUS: An extendable package of programs for data exploration, classification and correlation, M. Forina, R. Leardi, C. Armanino and S. Lanteri, Elsevier, Amsterdam, 1988, Price: US $645 ISBN 0-444-43012-1. *Journal of Chemometrics*, 4(2): 191–193.

25) Wahba, G.; and Wang, Y. 2014. Representer Theorem. *Wiley StatsRef: Statistics Reference Online*, 1–11.

26) Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3): 37–52.

27) Ye, J.; Janardan, R.; Park, C. H.; and Park, H. 2004. An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8): 982–994.

28) Ye, J.; and Yu, B. 2005. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6(4).

29) Zhang, T.; Fang, B.; Tang, Y. Y.; Shang, Z.; and Xu, B. 2009. Generalized discriminant analysis: A matrix exponential approach. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(1): 186–197.

30) Zheng, W.; Zhao, L.; and Zou, C. 2004. A modified algorithm for generalized discriminant analysis. *Neural Computation*, 16(6): 1283–1297.