POWER LINEAR DISCRIMINANT ANALYSIS

Makoto Sakai1,2, Norihide Kitaoka2,3, Seiichi Nakagawa2

1 DENSO CORPORATION, Nisshin 470-0111, Japan
2 Toyohashi University of Technology, Toyohashi 441-8580, Japan
3 Nagoya University, Nagoya 464-8601, Japan

msakai@rlab.denso.co.jp, {kitaoka, nakagawa}@slp.ics.tut.ac.jp

ABSTRACT

Dimensionality reduction is one of the important preprocessing steps to handle high-dimensional data. Linear discriminant analysis (LDA) is a classical and popular approach for this purpose. LDA finds an optimal linear transformation, which maximizes the ratio of the variance in the between-class distance to the variance in the within-class distance. On the other hand, in order to overcome the limitation in LDA resulting from the assumption of equal covariance, several heteroscedastic extensions, such as heteroscedastic discriminant analysis (HDA), have been proposed. However, it is difficult to find one particular criterion suitable for any kind of data set in carrying out dimensionality reduction while preserving discriminative information.

In this paper, we propose a new generalized framework which we call power linear discriminant analysis (PLDA). PLDA can describe various criteria including LDA and HDA with one parameter. Numerical results show the effectiveness for various data sets.

1. INTRODUCTION

The data treated in pattern recognition tend to have high dimensions to increase information used in post-processing. Such data may have strong correlations among dimensions, and may include nonessential information. In addition, high-dimensional data require a heavy computational load. Thus, dimensionality reduction without losing essential information is a very important issue.

Linear discriminant analysis (LDA) is widely used for this purpose [1, 2]. LDA is a powerful tool to preserve discriminative information and is an effective method of dimensionality reduction for various kinds of data. LDA assumes each class has the same within-class covariance [3]. However, this assumption does not necessarily hold for a real data set.

In order to overcome this limitation, several methods have been proposed [4, 5, 6]. In the speech recognition community, Kumar et al. incorporated the maximum likelihood estimation as an objective function to estimate parameters for different Gaussians with unequal covariances [4]. Saon et al. proposed another objective function similar to Kumar’s and showed its relationship with a constrained maximum likelihood estimation [5]. Both Kumar’s and Saon’s heteroscedastic extensions are called heteroscedastic discriminant analysis (HDA). Also, in the machine vision/learning community, De la Torre et al. proposed an objective function which uses divergence as a measure of the distance between two classes unlike HDA and accumulates pairwise distances [6]. The effectiveness of these methods for some data sets has been experimentally shown. However, it is difficult to find one particular criterion suitable for any kind of data set.

In this paper, we focus on LDA and Saon’s HDA, and give a new interpretation of them. Then, we propose a new framework which we call power linear discriminant analysis (PLDA). PLDA can describe various criteria including LDA and HDA with one parameter. Numerical results show the effectiveness for various data sets.

The paper is organized as follows: Classical LDA and its extensions are reviewed in Section 2. Then, a new framework of PLDA is proposed in Section 3. Numerical results are presented in Section 4. Finally, conclusions and future work are given in Section 5.

2. LINEAR DIMENSIONALITY REDUCTION

In this section, we formulate the problem of linear dimensionality reduction and review conventional methods. The generic problem of linear dimensionality reduction is as follows: Given n-dimensional samples \(x_j \in \mathbb{R}^n (j = 1, 2, \ldots, N)\), let us find a transformation matrix \(B \in \mathbb{R}^{p \times n}\) that maps these samples to p-dimensional samples \(z_j \in \mathbb{R}^p (j = 1, 2, \ldots, N) (p < n)\), where \(N\) denotes the number of samples and \(z_j = B^T x_j\).

2.1. Linear Discriminant Analysis

Within-class and between-class covariance matrices are defined as follows [1, 2]:

\[
\Sigma_w = \frac{1}{N} \sum_{k=1}^{c} \sum_{x_j \in D_k} (x_j - \mu_k) (x_j - \mu_k)^T
\]

\[
= \sum_{k=1}^{c} F_k \Sigma_k, \quad (1)
\]

\[
\Sigma_B = \sum_{k=1}^{c} P_k (\mu_k - \mu) (\mu_k - \mu)^T, \quad (2)
\]

where \(c\) denotes the number of classes, \(D_k\) denotes the subset of samples labeled as class \(k\), \(\mu\) is the mean vector for all the classes, \(\mu_k\) is the mean vector in the class \(k\), \(\Sigma_k\) is the covariance matrix in the class \(k\), and \(P_k\) is the class weight:

\[
\mu = \frac{1}{N} \sum_{j=1}^{N} x_j, \quad (3)
\]

\[
\mu_k = \frac{1}{N_k} \sum_{x_j \in D_k} x_j, \quad (4)
\]

\[
\Sigma_k = \frac{1}{N_k} \sum_{x_j \in D_k} (x_j - \mu_k) (x_j - \mu_k)^T, \quad (5)
\]

\[
P_k = \frac{N_k}{N}, \quad (6)
\]

where \(N_k\) is the number of samples labeled as class \(k\) and \(N\) is the number of all samples.

In LDA, the objective function is defined as follows:

\[
J_{LDA}(B) = \frac{|B^T \Sigma_B B|}{|B^T \Sigma_w B|}. \quad (7)
\]
LDA finds a transformation matrix $\mathbf{B}$ that maximizes the objective function:

$$
\mathbf{B}_{LDA} = \arg \max_{\mathbf{B}} \, J_{LDA}(\mathbf{B}).
$$

(8)

The optimization problem in Eq. (8) is equivalent to finding all the eigenvectors that satisfy the following equation:

$$
\Sigma_b \varphi = \lambda \Sigma_w \varphi,
$$

(9)

where $\lambda$ denotes an eigenvalue and $\varphi$ denotes an eigenvector.

### 2.2. Heteroscedastic Extensions

LDA is not the optimal transformation when the class distributions are heteroscedastic. Campbell has shown that LDA is related to the maximum likelihood estimation of parameters for a Gaussian model with equal class covariance [3]. However, this condition is not necessarily satisfied for a real data set.

In order to overcome this limitation, several extensions have been proposed. Kumar et al. incorporated the maximum likelihood estimation of parameters for differently distributed Gaussians [4]. Saon et al. proposed another objective function similar to Kumar’s and showed its relationship with a constrained maximum likelihood estimation [5]. Both Kumar’s and Saon’s heteroscedastic extensions are called heteroscedastic discriminant analysis (HDA). Also, De la Torre et al. proposed an objective function which uses divergence as a measure of the distance between two classes unlike HDA and accumulates pairwise distances [6].

In this paper, we focus on Saon’s HDA objective function:

$$
J_{HDA}(\mathbf{B}) = \prod_{k=1}^{c} \left( \frac{|\mathbf{B}^T \Sigma_b \mathbf{B}|}{|\mathbf{B}^T \Sigma_k \mathbf{B}|} \right)^{N_k} \propto \left( \frac{\tilde{\Sigma}_b}{\sum_{k=1}^{c} \tilde{\Sigma}_k} \right)^{\sum_{k=1}^{c} P_k \Sigma_k},
$$

(10)

The solution to maximize Eq.(10) is not analytically obtained. Therefore, its maximization is performed using a numerical optimization technique.

See also [5] concerning the relationship between Saon’s and Kumar’s objective functions.

### 2.3. Dependency on Data Set

In Figure 1, two-dimensional two- or three-class data samples are projected onto a one-dimensional subspace by LDA and HDA. Figure 1(a) shows that HDA has higher separability than LDA for the data set used in [5]. On the other hand, as shown in Figure 1(b), LDA has higher separability than HDA for another data set. Figure 1(c) shows the case with another data set where both LDA and HDA have low separabilities. Thus, LDA and HDA do not always classify the given data set appropriately. All results show that the separabilities of LDA and HDA depend significantly on data sets.

### 3. GENERALIZATION OF DISCRIMINANT ANALYSIS

As shown above, it is difficult to separate appropriately every data set with one particular criterion such as LDA and HDA. Here, we concentrate on providing a framework which integrates various criteria.

### 3.1. Relationship between LDA and HDA

From a different viewpoint, LDA and HDA objective functions can be rewritten as

$$
J_{LDA}(\mathbf{B}) = \frac{|\mathbf{B}^T \Sigma_b \mathbf{B}|}{|\mathbf{B}^T \Sigma_w \mathbf{B}|} = \frac{\tilde{\Sigma}_b}{\sum_{k=1}^{c} P_k \Sigma_k},
$$

(11)

$$
J_{HDA}(\mathbf{B}) = \prod_{k=1}^{c} \left( \frac{|\mathbf{B}^T \Sigma_b \mathbf{B}|}{|\mathbf{B}^T \Sigma_k \mathbf{B}|} \right)^{N_k} \propto \left( \frac{\tilde{\Sigma}_b}{\sum_{k=1}^{c} \tilde{\Sigma}_k} \right)^{\sum_{k=1}^{c} P_k \Sigma_k},
$$

(12)

where $\tilde{\Sigma}_b = \mathbf{B}^T \Sigma_b \mathbf{B}$ and $\tilde{\Sigma}_k = \mathbf{B}^T \Sigma_k \mathbf{B}$ are the between-class and class $k$ covariance matrices in the projected space, respectively.

Both numerators denote determinants of the between-class covariance matrix. In Eq. (11), the denominator can be viewed...
as a determinant of the weighted arithmetic mean of the class covariance matrices. Similarly, in Eq. (12), the denominator can be viewed as a determinant of the weighted geometric mean of the class covariance matrices. Thus, the difference between LDA and HDA is the definitions of the mean of the class covariance matrices.

3.2. Power Linear Discriminant Analysis

As described above, Eqs. (11) and (12) give us a new integrated interpretation of LDA and HDA. As extension of this interpretation, their denominators can be replaced by a determinant of the weighted harmonic mean, or a determinant of the root mean square.

In the econometric literature, a more general definition of a mean is often used, called the weighted mean of order m [7]. We extend this notion to a determinant of a matrix mean and propose a new objective function as follows:

$$J_{PLDA}(B, m) = \left| \sum_{k=1}^{c} P_k \Sigma_k^{m} \right|^{1/m}, \quad (13)$$

where \( m \) denotes a control parameter. Intuitively, as \( m \) becomes larger, the classes with larger variances become dominant in the denominator of Eq. (13). Contrarily, as \( m \) becomes smaller, the classes with smaller variances become dominant. Thus, varying a parameter \( m \), the proposed objective function can represent various objective ones. Some typical objective functions are enumerated below.

- \( m = 2 \) (root mean square)
  $$J_{PLDA}(B, 2) = \left| \sum_{k=1}^{c} P_k \Sigma_k \right|^{1/2}.$$  

- \( m = 1 \) (arithmetic mean)
  $$J_{PLDA}(B, 1) = \sum_{k=1}^{c} P_k \Sigma_k = J_{LDA}(B).$$

- \( m = 0 \) (geometric mean)
  $$J_{PLDA}(B, 0) = \prod_{k=1}^{c} \Sigma_k^{P_k}, \quad \propto J_{HDA}(B).$$

- \( m = -1 \) (harmonic mean)
  $$J_{PLDA}(B, -1) = \left| \sum_{k=1}^{c} \frac{1}{P_k \Sigma_k^{-1}} \right|^{1/m}.$$  

We call this new discriminant analysis formulation Power Linear Discriminant Analysis (PLDA). Figure 1(c) shows that PLDA can have a higher separability for a data set with which LDA and HDA have lower separability. To maximize the PLDA objective function with respect to \( B \), we can use numerical optimization techniques such as the Nelder-Mead method [8] and SANN method [9]. These methods need no derivatives of the objective function. However, it is known that these methods converge slowly. In some special cases, using a matrix differential calculus [10], the derivatives of the objective function are derived. Hence, we can use some fast convergence methods, such as the quasi-Newton method and conjugate gradient method [11].

3.2.1. Order \( m \) constrained to be an integer

Assuming that a control parameter \( m \) is constrained to be an integer, the derivatives of the PLDA objective function are formulated as follows:

$$\frac{\partial}{\partial B} \log J_{PLDA}(B, m) \left| \sum_{k=1}^{c} P_k \Sigma_k \right| - \frac{1}{m} \log \left| \sum_{k=1}^{c} P_k \Sigma_k \right|$$

$$= 2 \Sigma B \Sigma^{-1} - 2 D_m, \quad (14)$$

where

$$D_m = \begin{cases} \frac{1}{m} \sum_{k=1}^{c} P_k \Sigma_k B \Sigma_k^{-1}, & \text{if } m > 0 \\ - \frac{1}{m} \sum_{k=1}^{c} P_k \Sigma_k B \sum_{j=1}^{m} Y_{m, j, k}, & \text{otherwise} \end{cases}$$

and

$$X_{m, j, k} = \Sigma_k^{m-j} \left( \sum_{l=1}^{c} P_l \Sigma_l^m \right)^{-1} \Sigma_k^{-1},$$

$$Y_{m, j, k} = \Sigma_k^{m+j-1} \left( \sum_{l=1}^{c} P_l \Sigma_l^m \right)^{-1} \Sigma_k^{-j}.$$

3.2.2. \( \Sigma_k \) constrained to be diagonal

Because of computational simplicity, the covariance matrices in the class \( k \) are often assumed to be diagonal [4, 5]. Since a diagonal matrix multiplication is commutative, the derivatives of the PLDA objective function are simplified as follows:

$$\frac{\partial}{\partial B} \log J_{PLDA}(B, m) \left| \sum_{k=1}^{c} P_k \Sigma_k \right| - \frac{1}{m} \log \left| \sum_{k=1}^{c} P_k \text{diag} \left( \Sigma_k \right)^m \right|$$

$$= 2 \Sigma B \Sigma^{-1} - 2 \left( \sum_{k=1}^{c} P_k \Sigma_k \text{diag} \left( \Sigma_k \right)^{-m} \right) \left( \sum_{k=1}^{c} P_k \text{diag} \left( \Sigma_k \right)^m \right)^{-1}, \quad (15)$$

where \( \text{diag} \) is an operator which sets zero to off-diagonal elements. When \( m \) is equal to zero, the PLDA objective function corresponds to the diagonal HDA (DHDA) objective function introduced in [5].

4. NUMERICAL RESULTS

In this section, we experimentally evaluate the performance of the conventional and proposed dimensionality reduction methods. We used the glass, iris, and wine benchmark problems from the UCI machine learning repository [12]. The first data set includes 214 samples belonging to 6 different classes. We only used 175 samples belonging to 3 different classes for training and test. Each class has 70, 76, and 29 samples, respectively. The dimension of this data set is nine. The second data set includes 150 samples belonging to 3 different classes, each with 50 samples. The dimension of this data set is four. The third data set includes 178 samples belonging to 3 different classes. Each class has 59, 71, and 48 samples, respectively. The dimension of this data set is thirteen.
Each data set was randomly partitioned into a training set consisting of two-thirds of the whole set and into a test set consisting of one-third. We used the training set to compute the optimal discriminant and then tested its performance using the test set. To reduce the variability, the splitting method was repeated 100 times, and the resulting accuracies were averaged.

All data were centered and scaled. Then, the dimensionality reduction was performed using principal component analysis (PCA), LDA, HDA, and PLDA for the three data sets. The dimensions of the glass, iris, and wine data sets were reduced to two. In HDA and PLDA except when \( m \) is equal to one, to optimize Eq. (13), we used the BFGS algorithm as a numerical optimization technique [11]. Eq. (14) was used to compute a gradient. The LDA transformation matrix was used for the initial gradient. For all the experiments, a Gaussian classifier was used for classification after reducing dimensionality.

The results with the glass, iris, and wine data sets are presented in Table 1, 2, and 3, respectively. The best results are highlighted in bold. In the case of the glass data set, PLDA with \( m = 25 \) gives the highest accuracy as shown in Table 1. For the iris data set in Table 2, the highest accuracy is obtained by PLDA with a different control parameter (\( m = 13 \)). For the wine data set in Table 3, the highest accuracy is obtained by PLDA with \( m = 10 \). All results show that each data set has a different optimal control parameter, and PLDA with the optimal control parameter consistently outperforms the other methods.

### Table 1. Classification accuracy and standard deviation (SD) for the glass data set.

<table>
<thead>
<tr>
<th>method</th>
<th>accuracy ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o dim. reduction</td>
<td>59.93 ± 5.81</td>
</tr>
<tr>
<td>PCA</td>
<td>48.77 ± 4.97</td>
</tr>
<tr>
<td>PLDA ( m = -5 )</td>
<td>46.44 ± 5.70</td>
</tr>
<tr>
<td>HDA ( (m = 0) )</td>
<td>45.46 ± 5.88</td>
</tr>
<tr>
<td>LDA ( (m = 1) )</td>
<td>56.70 ± 5.48</td>
</tr>
<tr>
<td>PLDA ( m = 5 )</td>
<td>58.87 ± 5.54</td>
</tr>
<tr>
<td>PLDA ( m = 10 )</td>
<td>59.13 ± 5.63</td>
</tr>
<tr>
<td>PLDA ( m = 15 )</td>
<td>59.19 ± 5.60</td>
</tr>
<tr>
<td>PLDA ( m = 20 )</td>
<td>59.24 ± 5.65</td>
</tr>
<tr>
<td>PLDA ( m = 25 )</td>
<td>59.29 ± 5.53</td>
</tr>
<tr>
<td>PLDA ( m = 30 )</td>
<td>59.13 ± 5.55</td>
</tr>
<tr>
<td>PLDA ( m = 40 )</td>
<td>58.70 ± 5.71</td>
</tr>
</tbody>
</table>

### Table 2. Classification accuracy and standard deviation for the iris data set.

<table>
<thead>
<tr>
<th>method</th>
<th>accuracy ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o dim. reduction</td>
<td>97.52 ± 2.16</td>
</tr>
<tr>
<td>PCA</td>
<td>96.34 ± 2.04</td>
</tr>
<tr>
<td>PLDA ( m = -10 )</td>
<td>95.20 ± 4.67</td>
</tr>
<tr>
<td>PLDA ( m = -5 )</td>
<td>95.92 ± 3.31</td>
</tr>
<tr>
<td>HDA ( (m = 0) )</td>
<td>97.26 ± 2.09</td>
</tr>
<tr>
<td>LDA ( (m = 1) )</td>
<td>97.54 ± 2.00</td>
</tr>
<tr>
<td>PLDA ( m = 5 )</td>
<td>97.60 ± 2.06</td>
</tr>
<tr>
<td>PLDA ( m = 10 )</td>
<td>97.64 ± 2.10</td>
</tr>
<tr>
<td>PLDA ( m = 13 )</td>
<td>97.68 ± 2.09</td>
</tr>
<tr>
<td>PLDA ( m = 15 )</td>
<td>97.64 ± 2.10</td>
</tr>
<tr>
<td>PLDA ( m = 20 )</td>
<td>97.58 ± 2.03</td>
</tr>
</tbody>
</table>

### Table 3. Classification accuracy and standard deviation for the wine data set.

<table>
<thead>
<tr>
<th>method</th>
<th>accuracy ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o dim. reduction</td>
<td>98.03 ± 2.10</td>
</tr>
<tr>
<td>PCA</td>
<td>75.10 ± 4.76</td>
</tr>
<tr>
<td>PLDA ( m = -10 )</td>
<td>93.93 ± 2.99</td>
</tr>
<tr>
<td>PLDA ( m = -5 )</td>
<td>94.05 ± 3.21</td>
</tr>
<tr>
<td>PLDA ( m = -5 )</td>
<td>94.20 ± 3.76</td>
</tr>
<tr>
<td>PLDA ( m = -3 )</td>
<td>95.57 ± 3.67</td>
</tr>
<tr>
<td>HDA ( (m = 0) )</td>
<td>97.62 ± 1.81</td>
</tr>
<tr>
<td>LDA ( (m = 1) )</td>
<td>98.44 ± 1.43</td>
</tr>
<tr>
<td>PLDA ( m = 3 )</td>
<td>98.27 ± 1.57</td>
</tr>
<tr>
<td>PLDA ( m = 5 )</td>
<td>98.11 ± 1.71</td>
</tr>
<tr>
<td>PLDA ( m = 7 )</td>
<td>98.10 ± 1.73</td>
</tr>
<tr>
<td>PLDA ( m = 10 )</td>
<td>97.96 ± 1.79</td>
</tr>
</tbody>
</table>

### 5. CONCLUSIONS

In this paper a new framework is proposed for integrating various criteria to reduce dimensionality. The new framework which we call power linear discriminant analysis (PLDA) includes LDA and Saon’s HDA criteria as special cases. The experimental results on real data sets show that PLDA can reduce dimensionality while preserving discriminative information. Future work includes the application of PLDA to large data sets with high-dimensional data such as in speech recognition and face recognition.

### 6. REFERENCES


---

MLRepository.html