

# Harmonic Mean for Subspace Selection

Wei Bian and Dacheng Tao

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, P.R. China  
{cswbian, csdct}@comp.polyu.edu.hk

## Abstract

*Under the homoscedastic Gaussian assumption, it has been shown that Fisher's linear discriminant analysis (FLDA) suffers from the class separation problem when the dimensionality of subspace selected by FLDA is strictly less than the class number minus 1, i.e., the projection to a subspace tends to merge close class pairs. A recent result shows that maximizing the geometric mean of Kullback-Leibler (KL) divergences of class pairs can significantly reduce this problem. In this paper, to further reduce the class separation problem, the harmonic mean is applied to replace the geometric mean for subspace selection. The new method is termed maximization of the harmonic mean of all pairs of symmetric KL divergences (MHMD). As MHMD is invariant to rotational transformations, an efficient optimization procedure can be conducted on the Grassmann manifold. Thorough empirical studies demonstrate the effective of harmonic mean in dealing with the class separation problem.*

## 1. Introduction

Fisher's linear discriminant analysis (FLDA) is one of the most well-known linear subspace selection methods. However, under the homoscedastic Gaussian assumption, FLDA suffers from the class separation problem [6][7]: when the dimensionality of the subspace is strictly less than  $c-1$ , wherein  $c$  is the class number, close classes will be merged together in the selected subspace.

Many approaches have been developed to reduce this problem. Loog et al. [2] proposed the approximate pairwise accuracy criterion (aPAC), which uses a weighting function to emphasize the close class pairs in order to reducing the merging of close class pairs. Lotlikar and Kothari [3] developed the fractional-step

FLDA, which essentially is also a weighting approach but selects a subspace through fractional steps.

Recently, Tao et al. [6][7] proposed the maximization of the geometric mean of all pairs of Kullback-Leibler (KL) divergences (MGMD) for subspace selection, which maximizing the geometric mean of KL divergences between different class pairs. This geometric mean method was shown to perform much better than aPAC and the fractional-step FLDA in dealing with the class separation problem. However, as will see in this paper there is still room to further reducing the class separation problem.

In this paper, we introduce a new discriminative subspace selection criterion, the maximization of the harmonic mean of all pairs of symmetric KL divergences (MHMD), which maximizes the harmonic mean of the symmetric KL divergences between all class pairs. Mathematical analysis interprets why MHMD performs better than MGMD. Moreover, by suitable transforms, we show MHMD is equivalent to a rotational invariant optimization problem with orthonormal constraint. Based on this property, an efficient optimization algorithm is developed based on a conjugate gradient step on the Grassmann manifold. Preliminary experiments on both synthetic data and real data from the COIL-20 object image database, and two datasets in the UCI machine learning repository demonstrate the harmonic mean criterion significantly reduces the class separation problem and performs better than the geometric mean criterion.

The rest of this paper is organized as follows. In Section 2, we present the harmonic mean criterion and analyze why the harmonic mean performs better than the geometric mean in reducing the class separation problem. In Section 3, we present a Grassmann manifold based optimization algorithm to efficiently find a solution based on this new criterion. In Section 4, experiments on both synthetic data and real data are shown. Section 5 concludes.

## 2. Harmonic mean for subspace selection

In this Section, we first present the harmonic mean criterion for discriminative subspace selection and show why it performs better than the geometric mean in reducing the class separation problem.

### 2.1 The harmonic mean criterion

The harmonic mean criterion for subspace selection is to maximize the harmonic mean of all pairwise distances between all class pairs. It has been proved that FLDA is equivalent to maximizing the arithmetic mean of all pairwise distances [6][7]. Therefore, the harmonic mean criterion can be seen as a variation of FLDA. In our setting, we use the symmetric KL divergence to describe the pairwise distance between classes, where the symmetric KL divergence between two pdfs  $p_1$  and  $p_2$  is

$$SD(p_1 \parallel p_2) = \int (p_1 - p_2) \ln \left( \frac{p_1}{p_2} \right) du \quad (1)$$

Under the homoscedastic Gaussian assumption, the pdf of  $i$ th class is  $p_i \sim N(\bar{x} | \bar{\mu}_i, \Sigma)$ , wherein  $\bar{\mu}_i$  is the mean vector and  $\Sigma$  is the common covariance matrix of all classes. Suppose the subspace to be selected is denoted by the matrix  $W$ , then the pdf of the  $i$ th class in the projected subspace is  $p_i \sim N(\bar{y} | W^T \bar{\mu}_i, W^T \Sigma W)$ . By definition (1), we obtain the symmetric KL divergence between the  $i$ th class and the  $j$ th class in the projected subspace  $W$  as:

$$SD_W(p_i \parallel p_j) = \text{tr} \left( (W^T \Sigma W)^{-1} (W^T D_{ij} W) \right) \quad (2)$$

where matrix  $D_{ij} = (\bar{\mu}_i - \bar{\mu}_j)(\bar{\mu}_i - \bar{\mu}_j)^T$ .

Based on the divergence defined in (2), we define the maximization of the harmonic mean of all pairs of symmetric KL divergences (MHMD) criterion for  $c$ -class subspace selection as:

$$\max_W \left[ \sum_{1 \leq i < j \leq c} q_i q_j (SD_W(p_i \parallel p_j))^{-1} \right]^{-1} \quad (3)$$

or equivalently

$$\max_W H(W) = - \sum_{1 \leq i < j \leq c} q_i q_j (SD_W(p_i \parallel p_j))^{-1} \quad (4)$$

where  $q_i$  is the prior probability of the  $i$ th class, and can be set as the number of samples in the  $i$ th class. Criterion (3) or (4) is motivated by the harmonic mean on numbers  $H = N(a_1^{-1} + a_2^{-1} + \dots + a_N^{-1})^{-1}$ . We denote the solution of (4) by  $W_H$ .

### 2.2. Comparison with the geometric mean

The geometric mean criterion [6][7], i.e., MGMD, is a variation of FLDA. Under the homoscedastic Gaussian assumption, it is defined as

$$\max_W G(W) = \sum_{1 \leq i < j \leq c} q_i q_j \log(SD_W(p_i \parallel p_j)). \quad (5)$$

Next we give an explanation of why criterion (4) works better than criterion (5). Here, we simplify the pairwise symmetric KL divergence in the projected subspace  $SD_W(p_i \parallel p_j)$  as  $SD_{ij}^W$ . Partial derivatives of the objective functions in criteria (4) and (5) w.r.t. the pairwise divergence  $SD_{ij}^W$  are respectively:

$$\frac{\partial H}{\partial SD_{ij}^W} = \frac{q_i q_j}{(SD_{ij}^W)^{-2}} \quad \text{and} \quad \frac{\partial G}{\partial SD_{ij}^W} = \frac{q_i q_j}{(SD_{ij}^W)^{-1}}. \quad (6)$$

From both partial derivatives in (6), we can see the difference between criteria (4) and (5): the responses of the objective functions in criteria (4) and (5) w.r.t the increment of the pairwise divergence  $SD_{ij}^W$  are with different order. That is, although both partial derivatives  $\partial H / \partial SD_{ij}^W$  and  $\partial G / \partial SD_{ij}^W$  are sensitive to small  $SD_{ij}^W$  (i.e., the smaller the  $SD_{ij}^W$  the bigger the derivatives), the harmonic mean criterion emphasizes small symmetric divergences more than geometric mean criterion (the former is of order  $-2$  and the later is of order  $-1$ ). An interesting problem here is: what is the optimal order. Of course it is not simply  $-2$ . But this discussion will need further studies on the relation between the order and the classification accuracy, which beyond the scope of this paper. Our thorough experimental results show that order  $-2$  is better than  $-1$  in reducing the class separation problem.

## 3. Optimization on Grassmann manifold

In this section, we first introduce the conjugate gradient method on Grassmann manifold, and show how the MHMD problem can be solved by this optimization method according to the property of rotational invariant.

### 3.1. Conjugate gradient on the Grassmann manifold

Considering a problem  $\min F(Y)$ , where variable  $Y \in R^{n \times p}$ , with orthonormal constraint  $Y^T Y = I$ , if it also satisfies rotational invariant property  $F(Y) = F(YP)$  where  $P$  is any  $p \times p$  orthogonal matrix, then it can be efficiently solved by the

unconstrained problem  $\min F(Y)$  on the Grassmann manifold which is defined as the quotient set of all  $n \times p$  orthogonal matrix by identifying those matrices whose columns span the same subspace [1]. The Grassmann manifold provides an efficient approach to solving the optimization problem by exploiting the geometric properties of orthogonality and rotation invariance. Figure 1 summaries the conjugate gradient steps for solving  $\min F(Y)$  on the Grassmann manifold [1], where  $F_{Y_k}$  is the gradient  $\partial F / \partial Y$  at point  $Y_k$ .

**Figure 1. Conjugate gradient on Grassmann manifold**

**Algorithm Minimizing  $F(Y)$  on Grassmann Manifold**  
Step.1 Initialize with  $Y_0$  such that  $Y_0^T Y_0 = I$ , and compute  $G_0 = F_{Y_0} - Y_0 Y_0^T F_{Y_0}$ , and set  $H_0 = -G_0$ .  
Step.2 For  $k = 0, 1, \dots$   
Step.2.1 Minimize  $F(Y_k(t))$  over  $t$  by linear search where  $Y_k(t) = Y_k V \cos(\Sigma t) V^T + U \sin(\Sigma t) V^T$  and  $U \Sigma V^T$  is the compact SVD of  $H_k$ . Set  $t_k = t_{\min}$  and  $Y_{k+1} = Y_k(t_k)$ , where  $t_{\min}$  is the solution of linear search.  
Step.2.2 Update  $G_{k+1} = F_{Y_{k+1}} - Y_{k+1} Y_{k+1}^T F_{Y_{k+1}}$ .  
Step.2.3 Parallel transport  $H_k$  and  $G_k$  to the point  $Y_{k+1}$ :  
 $\tau H_k = (-Y_k V \sin \Sigma t_k + U \cos \Sigma t_k) \Sigma V^T$ ,  
 $\tau G_k = G_k - (Y_k V \sin \Sigma t_k + U (I - \cos \Sigma t_k)) U^T G_k$ .  
Step.2.4 Compute the new search direction  
 $H_{k+1} = -G_{k+1} + \gamma_k \tau H_k$ , where  $\gamma_k = \frac{\langle G_{k+1} - \tau G_k, G_{k+1} \rangle}{\langle G_k, G_k \rangle}$   
and  $\langle \Delta_1, \Delta_2 \rangle = \text{tr}(\Delta_1^T \Delta_2)$ .  
Step.2.5 Set  $H_{k+1} = -G_{k+1}$  if  $k+1 \equiv 0 \pmod{p(n-p)}$ , and go back to Step.2, until convergence.

### 3.2. Solving MHMD

We show that the criterion (4) is equivalent to a rotational invariant optimization problem with an orthonormal constraint, and therefore can be solved by conducting conjugate gradient steps on the Grassmann manifold. First, rewrite (4) by submitting (2) into it

$$\max_W H(W) = - \sum_{1 \leq i < j \leq c} q_i q_j \left( \text{tr} \left( (W^T \Sigma W)^{-1} (W^T D_{ij} W) \right) \right)^{-1} \quad (7)$$

Then, by an invertible transform  $V = \Sigma^{1/2} W$ , and letting  $A_{ij} = \Sigma^{-1/2} D_{ij} \Sigma^{-1/2}$ , problem (7) is equivalent to

$$\max_V H(V) = - \sum_{1 \leq i < j \leq c} q_i q_j \left( \text{tr} \left( (V^T V)^{-1} (V^T A_{ij} V) \right) \right)^{-1} \quad (8)$$

Furthermore, the problem (8) can be reformulated into the following style with an orthonormal constraint

$$\max_V H(V) = - \sum_{1 \leq i < j \leq c} q_i q_j \left( \text{tr} (V^T A_{ij} V) \right)^{-1} \quad (9)$$

s.t.  $V^T V = I$

Problem (9) is rotationally invariant, that is for an arbitrary orthogonal matrix  $P$ ,  $H(V) = H(VP)$ . Thus, we can get the solution of (9) by solving an unconstrained problem on the Grassmann manifold. Note we have to change (9) into min problem first by changing the sign of  $H(V)$  for the consistency with the algorithm in figure 1. The gradient of  $H(V)$  is

$$\frac{\partial H}{\partial V} = 2 \sum_{1 \leq i < j \leq c} q_i q_j \left( \text{tr} (V^T A_{ij} V) \right)^{-2} A_{ij} V. \quad (10)$$

Suppose the optimal solution of (9) is  $V^*$ , then we get the optimal solution of (7) as  $W_H = \Sigma^{-1/2} V^*$ .

## 4. Experiments

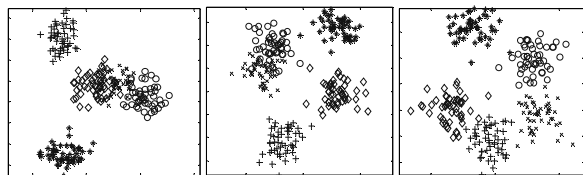
In this section, we compare performance of MHMD with FLDA and MGMD in handling the class separation problem in FLDA. Preliminary experiments are conducted on both synthetic data and real datasets from the Columbia Object Image Library (COIL)-20 object database [4] and the UCI machine learning repository [5]. In the following experiments, for both MGMD and MHMD, each trial runs five times with different initials and the best solution is chosen for classification.

### 4.1. Synthetic data experiment

We consider a five-classes classification problem: each class is a 10-dimensional Gaussian distribution, the common covariance matrix is an identity matrix  $I_{10}$ , and different class means are randomly sampled from a 10-dimensional zero-mean Gaussian with the covariance matrix  $2I_{10}$ . For thorough statistical evaluation, we run 500 trials independently. In each trial, we generate 50 samples for each class as the training set, and 100 samples for each class as the testing set. The nearest neighbor rule with the Mahalanobis distance is utilized for classification. The average classification accuracies (AA) associated with standard deviations (SD) is shown in the Table 1. Besides, for illustration of the ability of MHMD in dealing with the class separation problem, we show the 2-dimensional visualization of the five classes in Figure 2 for different methods. Both Table 1 and Figure 2 show MHMD performs better than MGMD and FLDA. Figure 1 shows MGMD balances distances of class pairs in the projected subspace.

**Table 1. The synthetic data experiment**

Method	1-D	2-D	3-D	4-D
FLDA(AA)	68.16	90.36	97.61	99.63
MGMD(AA)	77.38	96.30	97.86	99.63
MHMD(AA)	<b>79.58</b>	<b>97.48</b>	<b>99.04</b>	99.63
FLDA(SD)	8.17	6.18	2.59	0.71
MGMD(SD)	4.80	2.60	2.43	0.71
MHMD(SD)	3.47	1.80	1.13	0.71



FLDA MGMD MHMD  
**Figure 2. 2-D visualization of the five classes**

#### 4.2. Real data experiments

We first compare the performance of MHMD with MGMD and FLDA on the COIL-20 database [4]. This database contains 1440 images of 20 objects with 72 images of each. For statistical justification, 100 trials were run independently. In each trial, we randomly select 5 images of each object as the training set, and use the rest for testing. The nearest neighbor rule with the Mahalanobis distance is applied for classification. Average classification accuracies associated with the standard deviations is given in Table 2, which shows the MHMD is better than MGMD in handling the class separation problem in FLDA.

**Table 3. The “Image Segmentation” dataset**

Method	1-D	2-D	3-D	4-D	5-D	6-D
FLDA	50.14	66.33	81.91	87.52	89.43	89.71
MGMD	69.14	79.48	81.95	<b>88.00</b>	89.52	89.71
MHMD	<b>68.19</b>	<b>84.33</b>	<b>83.33</b>	87.95	<b>89.67</b>	89.71

Then, we evaluate MHMD on two datasets in the UCI machine learning repository [5]. They are the “Image Segmentation” data set, which consists of 210 training samples and 2100 testing samples from 7 classes in  $R^{19}$ ; and the “Landsat Satellite” dataset, which consists of 4435 training samples and 2000 testing samples from 6 classes in  $R^{36}$ . Classification accuracies are shown in Tables 3 and 4, respectively. Experimental results show the effectiveness of the proposed MHMD.

**Table 2. The COIL-20 dataset**

Method	2-D	4-D	6-D	8-D	10-D	12-D	14-D	16-D	18-D
FLDA(AA)	58.19	74.89	81.00	83.41	84.59	85.15	85.35	85.39	85.37
MGMD(AA)	62.36	75.43	81.24	<b>83.61</b>	84.78	85.30	<b>85.49</b>	85.51	85.37
MHMD(AA)	<b>64.13</b>	<b>78.21</b>	<b>81.51</b>	<b>83.61</b>	<b>84.96</b>	<b>85.31</b>	<b>85.49</b>	<b>85.52</b>	85.37
FLDA(SD)	4.49	3.03	2.47	2.42	2.54	2.56	2.48	2.42	2.36
MGMD(SD)	4.02	2.94	2.44	2.39	2.52	2.53	2.49	2.42	2.36
MHMD(SD)	3.71	3.04	2.51	2.41	2.52	2.53	2.47	2.40	2.36

**Table 4. The “Landsat Satellite” dataset**

Method	1-D	2-D	3-D	4-D	5-D
FLDA	55.45	72.35	82.65	83.10	83.95
MGMD	<b>67.75</b>	79.85	<b>82.85</b>	83.45	83.95
MHMD	65.45	<b>79.95</b>	82.75	<b>83.55</b>	83.95

#### 5. Conclusion

In this paper, we present a new subspace selection method, termed the maximization of the harmonic mean of all pairs of symmetric KL divergences (MHMD), and further introduce an efficient algorithm for finding an optimal solution of MHMD based on a conjugate gradient step on the Grassmann manifold. Experiments on the synthetic dataset and real datasets from COIL-20 object image database, and two datasets in the UCI machine learning repository demonstrate that the harmonic mean outperforms the geometric mean in dealing with the classification problem in FLDA.

#### Acknowledgements

The work was supported by Hong Kong Research Grants Council General Research Fund (under project number 528708), the Competitive Research Grants for Newly Recruited Junior Academic Staff 2007-2008 with the Hong Kong Polytechnic University (under project number A-PC0A), and National Science Foundation of China (under project number 60703037).

#### References

- [1] A. Edelman, T. A. Arias and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303-353, October 1998.
- [2] M. Loog. Approximate pairwise accuracy criteria for multiclass linear dimension reduction: generalizations of the Fisher criterion. *IEEE T-PAMI*, 26(7):762-766, July 2001.
- [3] R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE T-PAMI*, 22(6):623-627, June 2000.
- [4] S. A. Nene, S. K. Nayar and H. Murase. Columbia object image library: COIL-20. Columbia University, 1996.
- [5] D. J. Newman, S. Hettich, C. L. Blake and C. J. Merz. UCI repository of machine learning databases. Dept. of Information and Computer Science, University of California, Irvine, 1998.
- [6] D. Tao, X. Li, X. Wu, and S. J. Maybank. General Averaged Divergences Analysis. *IEEE ICDM*, 2007.
- [7] D. Tao, X. Li, X. Wu, and S. J. Maybank. Geometric Mean for Subspace Selection in Multiclass Classification. *IEEE T-PAMI*, in press.
- [8] X. Li, et al., "Discriminant Locally Linear Embedding with High Order Tensor Data," *IEEE T-SMC-B*, 2008.