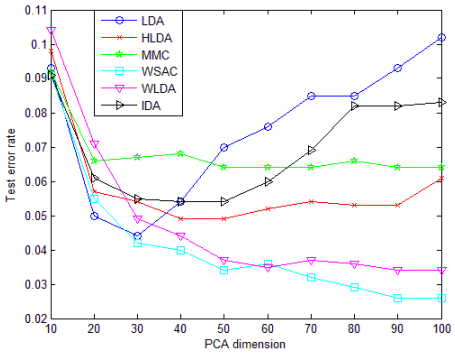
Supplements to WSAC

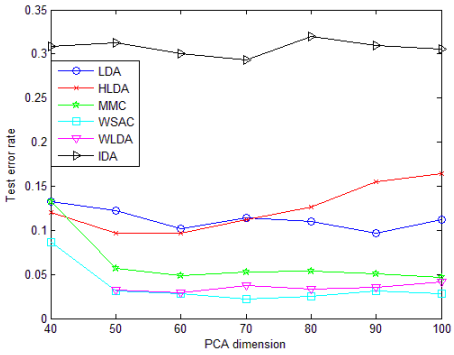
Leilei Yang, Songcan Chen

a. Chosen number of PCs

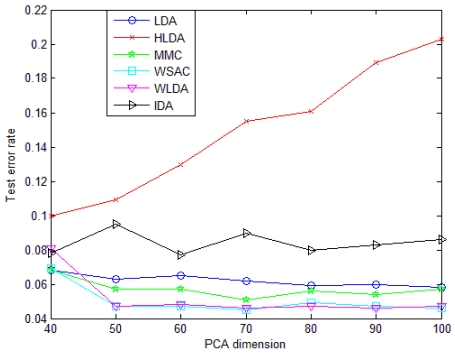
We will show the results of each method using different number of PCs on different dataset in the Fig.1 below. And from the figure below we can easily draw some conclusions as well as witness the competitiveness of our method. Considering the trends of these curves and the number of the training data points in each dataset, we can choose the number of each dataset. (mfeat-factor:40, AR:60, ORL:60, COIL:40, Yale:30)



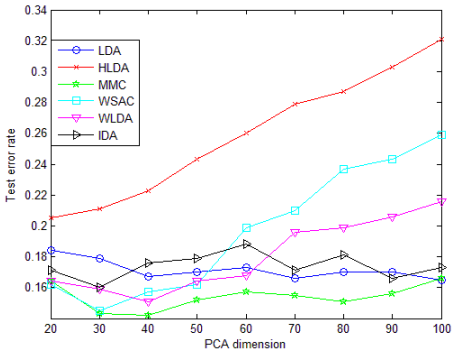
(a) mfeat-factor (40, 40)



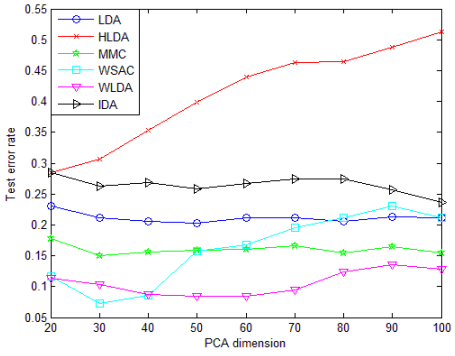
(b) AR (160, 60)



(c) ORL (160, 60)



(d) COIL (80, 40)



(e) Yale (60, 30)

Fig. 1 (a), (b), (c), (d), (e) give the DR performance of compared methods with varying principal components of each dataset. The x-axis is the number of principal components, and the y-axis is the error rate yielded by the nearest-neighbor classifier after using DR methods. (Digits in the parentheses respectively represent the number of training data and the chosen number of principal components)

b. Description

Due to of the lack of space as short paper, we omit the necessary deduction steps in the manuscript. Instead, we show the necessary deduction steps and partial experimental results of kernelized version as follows.

According to[5], the original space X can be mapped into a higher dimensional space F through a nonlinear mapping function Φ which can be induced from a corresponding kernel $k(\bullet,\bullet)$:

(1) $\Phi: X \to F$

$$x \to \Phi(x) .$$

Then the average-case within-class scatter matrix S_w^Φ in the space F is defined as:

$$\begin{aligned} S_w^\Phi &= S_t^\Phi - S_b^\Phi \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} \left(\Phi(x_i)-\Phi(x_j)\right)\left(\Phi(x_i)-\Phi(x_j)\right)^T - \sum_{k=1}^{m} n_k \left(\bar{m}_k-\bar{m}\right)\left(\bar{m}_k-\bar{m}\right) \right), \end{aligned} \tag{2}$$

and while the worst-case between-class scatter matrix S_{ij}^Φ as:

$$S_{ij}^\Phi = \left(\bar{m}_i-\bar{m}_j\right)\left(\bar{m}_i-\bar{m}_j\right)^T, \tag{3}$$

where $\bar{m} = \left(\sum_{i=1}^{n} \Phi(x_i) \right) / n$ is the mean of the whole dataset and $\bar{m}_k = \left(\sum_{x_i \in C_k} \Phi(x_i) \right) / n_k$ is the class mean of C_k. S_t^Φ is the total scatter and S_b^Φ is the average between-class scatter.

Now let $\tilde{X} = \left(\Phi(x_1),\ldots,\Phi(x_n)\right)$, $M = \left(\bar{m}_1,\ldots,\bar{m}_m\right)$, $H_n = I_n - \frac{1}{n} 1_n 1_n^T$ be the $n \times n$ identity matrix where 1_n is an $n \times 1$ column vector of all ones, $D = diag\left(n_1,\ldots,n_m\right)$ be a diagonal matrix whose (i,i) element is the number of data points in class i, E an $n \times m$ indicator matrix whose (i,j) element equals to 1 if x_i is from class j and 0 otherwise, L_{ij} an $m \times 1$ column vector with ith element being 1, jth element being -1, and the others all equaling to 0. Then it is easy to see that $M = \tilde{X}ED^{-1}$ where D^{-1} denotes the inverse of matrix D if D is nonsingular and the pseudo-inverse otherwise.

From the definitions of S_w^Φ and S_{ij}^Φ, we can rewrite them in matrix form as:

$$S_w^\Phi = \tilde{X}\left(\frac{1}{n}\left(H_n - H_n E D^{-1} E^T H_n \right) \right) \tilde{X}^T \tag{4}$$

and

$$S_{ij}^\Phi = \tilde{X}\left(ED^{-1}L_{ij}L_{ij}^T D^{-1}E^T \right)\tilde{X}^T . \tag{5}$$

Thus we can formulate a nonlinear form of WSAC by changing the projection matrix as $W=\tilde{X}A$, then our method can be formulated as

$$\begin{aligned} \max_{A} \quad & J(A)=\min_{i,j\in\mathbb{N}_m,i<j} Tr\left(A^T K\left(ED^{-1}L_{ij}L_{ij}^T D^{-1}E^T\right)KA\right) \\ s.t. \quad & Tr(A^T K\left(\frac{1}{n}\left(H_n - H_n ED^{-1}E^T H_n\right)\right)KA)\le 1 \end{aligned} \tag{6}$$

where $K=\tilde{X}\tilde{X}^T$. As a result, we can get the nonlinear form of WSAC for whichit can be optimized similarly.

c. Experiments

It needs to mention that in the original works [2,3], respectively involving HLDA and WLDA, such two DAs were not kernelized and just compared with other linear DA methods including LDA. Consider this fact, thus we just also compare the remaining three kernerlized methods, i.e. kernel LDA, kernel MMC and kernel WSAC. In this experiment, we follow your suggestion to adopt the Gaussian kernel $k\left(x_1,x_2\right)=\exp\left(-\frac{\left\|x_1-x_2\right\|_2^2}{\sigma^2}\right)$, where $\sigma^2=\frac{S}{n^2}\sum_{i,j=1}^{n}\left(x_i-x_j\right)^T\left(x_i-x_j\right)$, n being the number of training points and $0\le S\le 2$ being a scale parameter. We follow the same experiments settings on the image datasets to the previous ones. From the results in the Table I, we can draw similar conclusions to those in linear versions: our method still retains its competitiveness.

Table I. AVERAGE TESTERROR OF DIFFERENTLDA-BASED METHODS (STANDARD DEVIATIONSARE INPARENTHESES). THE 1STAND 2ND PERFORMANCES ARE DENOTED IN BOLD AND UNDERLINED RESPECTIVELY. THE VALUES OF S WHEN EVERY METHOD GETS THE BEST RESULTARE ALSO GIVEN IN THE TABLE.

	Kernel LDA	Kernel MMC	Kernel WSAC
ORL	<u>0.0837(0.0220)</u>, S=1.5	0.0944(0.0229), S=0.3	**0.0581(0.0189)**, S=1.5
COIL	0.2028(0.0376), S=1.5	<u>0.2025(0.0257)</u>, S=0.3	**0.1465(0.0265)**, S=1.5
Yale	<u>0.2248(0.0530)</u>, S=1.8	0.2695(0.0266), S=0.5	**0.1986(0.0329)**, S=1.5

3 References

[1] B. Xu, K. Huang, C. Liu, "Dimensionality Reduction by Minimal Distance Maximization," *2010 20th International Conference on Pattern Recognition (ICPR)*, pp.569-572, 23-26, Aug. 2010

[2] Yu Zhang, Dit-Yan Yeung, "Worst-case Linear Discriminant Analysis," *Advances in Neural Information Processing Systems 23 (NIPS)*, pp. 2568-2576. Vancouver, Canada, 2010.

[3] R. P. W. Duin, M. Loog, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.26, no.6, pp. 732-739, June 2004

[4] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," In S. Thrun, L. K. Saul, and B. Scholkopf, editors, *Advances in Neural Information Processing Systems16*, Vancouver, British Columbia, Canada, 2003.

[5] G. Baudat and FatihaAnouar. "Generalized discriminant analysis using a kernel approach," Neural Computation, 12(10):2385-2404, 2000.