CLASSIFICATION PROBABILITY ANALYSIS OF PRINCIPAL COMPONENT NULL SPACE ANALYSIS

Namrata Vaswani and Rama Chellappa

Dept. of Electrical and Computer Engineering and Center for Automation Research,
University of Maryland, College Park, MD 20742, USA
{namrata,rama}@cfar.umd.edu

Abstract

In a previous paper [1], we have presented a new linear classification algorithm, Principal Component Null Space Analysis (PCNSA) which is designed for problems like object recognition where different classes have unequal and non-white noise covariance matrices. PCNSA first obtains a principal components space (PCA space) for the entire data and in this PCA space, it finds for each class 'i', an M_i dimensional subspace along which the class's intra-class variance is the smallest. We call this subspace an Approximate Null Space (ANS) since the lowest variance is usually "much smaller" than the highest. A query is classified into class 'i' if its distance from the class's mean in the class's ANS is a minimum. In this paper, we discuss the PCNSA algorithm more precisely and derive tight upper bounds on its classification error probability. We use these expressions to compare classification performance of PCNSA with that of Subspace Linear Discriminant Analysis (SLDA) [2].

1. INTRODUCTION

Within the last several years, many algorithms have been proposed for object and face recognition problems. For problems like face recognition under small pose variations which involve discriminating similar type of objects, different classes have similar class covariance matrices (in particular similar directions of low and high intra-class variance). On the other hand for "non-similar type" classification problems like object recognition or face recognition under large pose variation, the minimum variance direction for one class might be a maximum variance direction for another. Existing linear classification algorithms like principal component analysis (PCA), linear discriminant analysis (LDA) and subspace LDA (SLDA) are optimal for the "similar type" classification problems. PCA [3] yields projection directions that maximize the total scatter but do not minimize the within class variance of each class. LDA [4] encodes discriminatory information by finding directions that maximize the ratio of between class scatter to within-class (or intra-class) scatter. In [2], PCA and LDA are combined to yield a subspace LDA (SLDA) based classification algorithm for face recognition which uses PCA first for dimensionality reduction and then LDA. In [5], performance of PCA and LDA is compared as a function of the amount of training data available and results are shown on different face databases.

1.1. Problem Formulation

Consider a P-dim data sample $\mathbf{Y}$ from class i (denote by C_i). (Henceforth, we refer to class i as C_i).

$$(\mathbf{Y})_{P\times 1}|\{Y \in C_i\} \sim \mathcal{N}(\mu_{full,i}, \Sigma_{full,i}) \tag{1}$$

The data sample projected in an L-dim PCA space with projection matrix, $(W^{PCA})_{P\times L}$, is

$$(\mathbf{X})_{L\times 1} \triangleq W^{PCA^T}(\mathbf{Y} - \bar{\mu}_{full}) \sim \mathcal{N}(\mu_i, \Sigma_i)\text{where} \tag{2}$$

$(\mu_i)_{L\times 1} = W^{PCA^T}(\mu_{full,i} - \bar{\mu}_{full})$, $(\Sigma_i)_{L\times L} = W^{PCA^T}\Sigma_{full,i}W^{PCA}$. $\mu_{full,i}, \Sigma_{full,i}$ are the class mean and class covariance in the original P-dimensional data space while μ_i, Σ_i are the class mean and covariance in the reduced PCA space (assumed L dimensional). For high dimensional data like images, the real dimensionality of data (with noise removed) is much smaller than P and PCA helps to remove directions with only noise and retain directions with large between class variance. In this work, we address the classification problem for the most general class covariance matrices (unequal, non-white) with eigenvalue decomposition $\Sigma_i = U_i\Lambda_iU_i^T$. LDA, on the other hand, assumes same eigenvectors for all classes ($U_i = U$) i.e. similar directions of low and high variance while PCA when used for classification assumes $U_i = I, \Lambda_i = \sigma_i^2 I$ i.e. the class covariances are white in PCA space.

2. PRINCIPAL COMPONENT NULL SPACE ANALYSIS

Assumptions:

1. For all classes, Σ_i has a high enough condition number, $R = \lambda_{max}/\lambda_{min}$ and hence an approximate null space exists. This would happen for most real classification problems especially the "non-similar type" ones.
2. Distance of class mean of a class,j, from class mean of class i in ANS of class i (denoted by N_i) is "significantly" greater than zero,i.e. $||N_i^T(\mu_j - \mu_i)|| > \rho||\mu_j - \mu_i|| > 0$[1].

Algorithm [1]:

[1] If this condition is not satisfied for two classes i and j, and if their null spaces coincide, i.e. $N_i = N_j$, we would have $d_i(\mathbf{X}) = d_j(\mathbf{X})$ always causing the algorithm to fail always

1. **Obtain PCA Space:** Evaluate the sample mean, $\bar{\mu}_{full}$ and covariance, Σ_{full} of the training data of all classes taken together as one sample set. Obtain the PCA projection matrix, $(W^{PCA})_{P \times L}$
2. Project the training data samples of each class into PCA space. Evaluate for each class i, the class mean, μ_i and the class covariance, Σ_i in PCA space.
3. **Obtain Class ANS:** Evaluate the approximate null space $(N_i)_{L \times M_i}$ for each class i as the M_i trailing eigenvectors of Σ_i (having eigenvalues $\lambda \leq 10^{-4}\lambda_{max}$, to satisfy assumption 1).
4. **Obtain Valid Classification Directions in ANS:** Now $N_i = [e_{i,1}|e_{i,2}|...e_{i,k}...|e_{i,M_i}]$. A null space direction, e, is a valid classification direction only if the distance between class means along that direction is "significantly" greater than zero (assumption 2) i.e. $e = e_{i,k}$ satisfies $|(\mu_i - \mu_j)^T e| > \rho ||\mu_i - \mu_j||$, $\forall j \neq i$, $0 < \rho < 1$ or equivalently, $\theta \triangleq \cos^{-1}(\frac{|(\mu_i - \mu_j)^T e|}{||\mu_i - \mu_j||}) < \theta_0 < \frac{\pi}{2}$. The PCNSA classification matrix for class i (${W_i}^{NSA}$) is chosen as those columns of N_i which satisfy this condition.
5. **Classification:** Project the query $\mathbf{Y}$ into the PCA space, $\mathbf{X} = W^{PCA^T}(\mathbf{Y} - \bar{\mu}_{full})$. The most likely class, c, is given by $c = \arg\min_i d_i(\mathbf{X})$, where

$$d_i(\mathbf{X}) \triangleq ||{W_i}^{NSA^T}(\mathbf{X} - \mu_i)|| \tag{3}$$

3. CLASSIFICATION ERROR PROBABILITY

We obtain error probability bound for PCNSA for a two class problem. We first evaluate error probability assuming a one dimensional ANS per class so that $W_i^{NSA} = (N_i)_{L\times 1}$. We then show how this can be extended to the general case of M dimensional ANS per class. The two class error probability expressions provide a union bound for the multi-class error probability. Define E_i as the event that error occurs given query $\mathbf{X} \in C_i$ (class i). The average error probability is $P_{e,avg} = (P(E_1) + P(E_2))/2$.

3.1. One-dim ANS per class

Using PCNSA's class specific metric ((3)), the error event E_1 is

$$E_1 \triangleq \{{d_2}^2(\mathbf{X}) < {d_1}^2(\mathbf{X}) | \mathbf{X} \in C_1\} \tag{4}$$

Now since each class has a one dimensional ANS, ${W_1}^{NSA} = N_1$ and $d_1(\mathbf{X}) = |N_1^T(\mathbf{X} - \mu_1)|$ is a scalar. Using (2), $N_1^T(\mathbf{X} - \mu_1)|\{\mathbf{X} \in C_1\} \sim \mathcal{N}(0, \lambda_{ANS,1})$. To upper bound on $P(E_1)$, let

$$\Delta = k\sqrt{\lambda_{ANS,1}} \tag{5}$$

Then, $P({d_1}^2(\mathbf{X}) > \Delta^2 | \mathbf{X} \in C_1) = 2(1 - \Phi(k)) \triangleq g(k)$ (6)

where $\Phi(.)$ is the cdf of an $\mathcal{N}(0,1)$ random variable. We choose k large enough so that $g(k)$ is small. For $k = 10$, $g(k) = 10^{-23}$. Now the error event E_1 (defined in (4)) can be split as [2],

$$\begin{aligned} E_1 &= \{d_2^2 \leq d_1^2, d_1^2 \leq \Delta^2\} \cup \{d_2^2 \leq d_1^2, d_1^2 > \Delta^2\} \\ &\subseteq \{{d_2}^2(\mathbf{X}) \leq \Delta^2\} \cup \{{d_1}^2(\mathbf{X}) > \Delta^2\}. \end{aligned} \tag{7}$$

Thus, $P(E_1) \leq P({d_2}^2(\mathbf{X}) \leq \Delta^2) + g(k)$. Now $d_2(\mathbf{X}) = |N_2^T(\mathbf{X} - \mu_2)|$. Using (2) we get,
$\mathbf{Z} \triangleq \frac{N_2^T(\mathbf{X} - \mu_1)}{\sqrt{N_2^T \Sigma_1 N_2}} \sim \mathcal{N}(0,1)$. So defining,

$$\alpha \triangleq |N_2^T(\mu_1 - \mu_2)|, \text{ and } \sigma \triangleq \sqrt{N_2^T \Sigma_1 N_2}, \tag{8}$$

we get, $P({d_2}^2(\mathbf{X}) < \Delta^2) = P(\frac{\alpha - \Delta}{\sigma} < \mathbf{Z} < \frac{\alpha + \Delta}{\sigma})$. Thus

$$\begin{aligned} P(E_1) &\leq P(\frac{\alpha - \Delta}{\sigma} < \mathbf{Z} < \frac{\alpha + \Delta}{\sigma}) + g(k) \\ &= \Phi(\frac{\alpha + \Delta}{\sigma}) - \Phi(\frac{\alpha - \Delta}{\sigma}) + g(k) \\ &= \int_{\frac{\alpha}{\sigma}(1 - \frac{\Delta}{\alpha})}^{\frac{\alpha}{\sigma}(1 + \frac{\Delta}{\alpha})} \mathcal{N}(z; 0, 1) dz + g(k) \end{aligned} \tag{9}$$

3.2. $\mathbf{M_i}$-dim ANS per class

In this case N_1 and N_2 are $L \times M_i, i = 1, 2$ dim matrices. Define

$$\Delta^2 = k^2 (\sum_{j=1}^{M_1} \lambda^2_{ANS,1,j}) \tag{10}$$

Error event E_1 is as defined in (4) and can be bounded using exactly the same logic as in (7). Thus we have

$$P(E_1) \leq P({d_2}^2(\mathbf{X}) < \Delta^2 | \mathbf{X} \in C_1) + P({d_1}^2(\mathbf{X}) > \Delta^2 | \mathbf{X} \in C_1) \tag{11}$$

First consider $P({d_1}^2(\mathbf{X}) > \Delta^2)$. Define

$$Z_{N_1} \triangleq N_1^T(\mathbf{X} - \mu_1) \sim \mathcal{N}(0, \Lambda_{ANS,1}), \ \Lambda_{ANS,1} \text{ diagnol} \tag{12}$$

then ${d_1}^2(\mathbf{X}) = ||Z_{N_1}||^2$. It is easy to see that

$$\{{d_1}^2(\mathbf{X}) > \Delta^2\} \subseteq \{\cap_j A_j\}^c, \ A_j = \{Z^2_{N_1,j} < k^2 \lambda^2_{ANS,1,j}\} \tag{13}$$

By (12), the components of the vector Z_{N_1} are independent and hence the events A_j are independent. Also, $P(A_j) = 1 - g(k)$ where $g(k)$ is defined in (6). Thus using (13),

$$P({d_1}^2(\mathbf{X}) > \Delta^2) \leq 1 - \prod_{j=1}^{M_1} P(A_j) = 1 - (1 - g(k))^{M_1} \triangleq g_{M_1}(k) \tag{14}$$

Now consider $P({d_2}^2(\mathbf{X}) \leq \Delta^2)$. Define

$$\begin{aligned} \beta &\triangleq N_2^T(\mu_2 - \mu_1) \text{ and } \Sigma \triangleq N_2^T \Sigma_1 N_2 \\ Z_{N_2} &\triangleq N_2^T(\mathbf{X} - \mu_1) \sim \mathcal{N}(0, \Sigma), \end{aligned} \tag{15}$$

then ${d_2}^2(\mathbf{X}) = ||Z_{N_2} - \beta||^2$. Let $\Sigma = USU^T$ is the eigenvalue decomposition of Σ. U is the $M_2 \times M_2$ matrix of eigenvectors and $S = diag(\sigma_j^2)$ is a diagnol matrix of its eigenvalues. Using U to diagnolize Z_{N_2}, we get

$$Z_{N_2}^{indep} = U^T Z_{N_2} \sim \mathcal{N}(0, S), \ S \text{ diagnol} \tag{16}$$

$$\text{Also define, } \alpha \triangleq |U^T \beta| \tag{17}$$

Since U is orthonormal, $||Z_{N_2}^{indep} - \alpha|| = ||Z_{N_2} - \beta||$ and so

$$P({d_2}^2(\mathbf{X}) \leq \Delta^2) = P(||Z_{N_2}^{indep} - \alpha||^2 < \Delta^2) \tag{18}$$

[2] Assume $\mathbf{X} \in C_1$ everywhere

Now, it is easy to see that

$$\{||Z_{N_2}^{indep} - \alpha||^2 < \Delta^2\} \subseteq \bigcap_j B_j, B_j = \{(Z_{N_2,j}^{indep} - \alpha_j)^2 < \Delta^2\}$$

The events $\{B_j\}$ are independent since elements of the vector $Z_{N_2}^{indep}$ are independent. Using (9), $P(B_j) = P(\alpha_j - \Delta < Z_{N_2,j} < \alpha_j + \Delta) = [\Phi(\frac{\alpha_j+\Delta}{\sigma_j}) - \Phi(\frac{\alpha_j-\Delta}{\sigma_j})]$, where $\sigma_j^2 = S_{j,j}$.

$$\text{Thus, } P({d_2}^2(\mathbf{X}) \leq \Delta^2) \leq \prod_{j=1}^{M_2}[\Phi(\frac{\alpha_j + \Delta}{\sigma_j}) - \Phi(\frac{\alpha_j - \Delta}{\sigma_j})] \quad (19)$$

Finally, combining (11), (14) and (19), we get

$$P(E_1) \leq \prod_{j=1}^{M_2}[\Phi(\frac{\alpha_j + \Delta}{\sigma_j}) - \Phi(\frac{\alpha_j - \Delta}{\sigma_j})] + g_{M_1}(k) \quad (20)$$

4. COMPARISON WITH SUBSPACE LDA

4.1. Subspace LDA (SLDA)

Now SLDA [2] computes a PCA space for the training data of all classes out together. In PCA space, it performs LDA, i.e. it computes the most discriminant directions, W^{LDA}, as $W^{LDA} = \arg\min_{W:W^TW=1}(W^T\Sigma_b W)/(W^T\Sigma_w W)$, where $\Sigma_b = (\sum_{i=1}^{K}(\mu_i - \bar{\mu}))/K$ and $(\Sigma_w = \sum_{i=1}^{K}\Sigma_i)/K$. The classification metric is $d_i(\mathbf{X}) = ||W^{LDA^T}(\mathbf{X} - \mu_i)||$. The error event for a two class problem (one dimensional W^{LDA}) is $E_1 \triangleq \{{d_2}^2(\mathbf{X}) < {d_1}^2(\mathbf{X})|\mathbf{X} \in C_1\}$. Error probability [6] follows using Gaussian hypothesis testing:

$$P(E_1) = 1 - \Phi(\frac{\hat{\alpha}}{\hat{\sigma}}) \quad = \quad \int_{\frac{\hat{\alpha}}{\hat{\sigma}}}^{\infty} \mathcal{N}(z;0,1)dz \quad (21)$$

$$\hat{\alpha} \triangleq \frac{|W^{LDA^T}(\mu_2 - \mu_1)|}{2}, \ \hat{\sigma} \quad \triangleq \quad \sqrt{W^{LDA^T}\Sigma_1 W^{LDA}} \quad (22)$$

4.2. Comparison

Looking at expressions (9) and (22), it is clear that PCNSA error probability can be made small if either $\frac{\Delta}{\alpha} = k\frac{\sqrt{\lambda_{ANS,i}}}{|(\mu_i - \mu_j)^T N_i|}$ tends to zero or $\frac{\alpha}{\sigma}$ tends to infinity. On the other hand, the LDA error probability goes to zero if and only if $\frac{\hat{\alpha}}{\hat{\sigma}}$ goes to infinity.

We now compare the error probabilities for a best and a worst case situation for LDA. We make some simplifying assumptions to reduce the number of variables. We assume a two dimensional PCA space and each class having a one dimensional ANS and one direction of maximum variance. Also, we assume that the eigenvalues of covariance matrices of both classes are equal, i.e. $\lambda_{max,1} = \lambda_{max,2} = \lambda_{max}$ and $\lambda_{ANS,1} = \lambda_{ANS,2} = \lambda_{min}$. Now for the classes to be linearly separable, for any orientation of the $(\mu_1 - \mu_2)$ direction w.r.t. the ANS-1 N_1 direction, the distance between the means should be at least of the order of $2\sqrt{\lambda_{max}}$. In our analysis below, we let $||\mu_1 - \mu_2|| = \sqrt{\lambda_{max}}$. With these assumptions, the error probability expressions can be reduced to a function of 3 variables: the condition number, $R = \lambda_{max}/\lambda_{min}$, the angle between N_1 and N_2, denoted by ψ and the angle made by the line joining the means (the vector $\mu_1 - \mu_2$) with N_2, denoted by θ. In two dimensions these two angles automatically fix the angle between the direction of $(\mu_1 - \mu_2)$ and N_1. We study two extreme cases of ψ, $\psi = 0^o$ (case 1) and $\psi = 90^o$ (case 2) which correspond to best case and worst case scenarios for LDA. We show that PCNSA works well in both these extreme cases as long as the assumptions of section 2 are satisfied and fails completely when they are not.

Intuitive Comparison: We first provide an intuitive comparison the two cases ($\psi = 0, 90^o$) using figure 1(a) and (b). In both figures, the condition number R is set to a large value (assumption 1 of section 2). We have $\theta \approx 0$ in figure 1(a) and $\theta \approx 45^o$ in 1(b), both being far from 90^o (assumption 2).

Figure 1(a) (case 1), is a best case scenario for both PCNSA and LDA since Y axis is the ANS direction for both classes and the common LDA direction(P^{LDA}) is close to the Y axis (ANS direction for either class). Thus variance of both classes along P^{LDA} is small. Also variance of class 1 along ANS of class 2 and vice versa is small too and $\theta = 0^o$ is far from 90^0. But in figure 1(b) (case 2), the maximum variance direction of one class coincides with the ANS of the other. This is the worst case for LDA but PCNSA works very well in this case. This case demonstrates the need for the PCNSA algorithm. Here, the Y axis is ANS direction for class 1 but a maximum variance direction for class 2 and vice versa for X axis. Thus P^{LDA} is along the direction $(\mu_1 - \mu_2)$. Along P^{LDA} both classes have a large enough variance. So LDA will have a high error probability in this case. The region for error event E_1 is ZRV and for E_2 is XRT. But PCNSA will still work well because the integration region for E_1 is only those parts of ellipse 1 which are closer to M^2 along N_2 (X axis) than to M^1 along N_1 (Y axis) and similarly for E_2. Thus the error region is the small region PQRS for both E_1 and E_2.

Comparing $P(error)$ as a function of R and θ: Now in case 1 ($\psi = 0^o$), $N_1 = N_2 = [0\ 1]^T$. Using the simplifying assumptions and definitions (8), $\Sigma_1 = \Sigma_2 = diag\{\lambda_{max}, \lambda_{min}\}$, $\alpha = \sqrt{\lambda_{max}}\cos\theta$ and $\sigma = \sqrt{\lambda_{min}}$. $R = \lambda_{max}/\lambda_{min}$ is the condition number of either class's covariance matrix. Substituting in (9), we get $P(E_1^{NSA}) \leq \int_{\sqrt{R}\cos\theta - k}^{\sqrt{R}\cos\theta + k}\mathcal{N}(z;0,1)dz \triangleq P(E^{NSA,bound})$. and the same expression for $P(E_2^{NSA})$ so that $P(E_{avg}^{NSA}) = P(E_1^{NSA})$. We also evaluate $P(E^{LDA})$ using (22). MATLAB is used to evaluate W^{LDA} for different values of R and θ. Both $P(E^{NSA,bound})$ and $P(E^{LDA})$ are plotted in figure 2(a), for $\theta \in [0, 90^o]$, and $R = 10^3, 10^4, 10^5$. This is a best case scenario for both SLDA and PCNSA as long as θ is bounded away from 90^o (assumption 2 of section 2 satisfied). We have for both NSA and LDA

$$\begin{aligned} \lim_{R\to\infty} P(E, R, \theta) = 0, \quad \forall \quad |\theta| < \theta_0 < 90^o \\ \text{But, } \lim_{\theta\to 90^o}\lim_{R\to\infty} P(E^{NSA,bound}, R, \theta) &= 1 \\ \text{and, } \lim_{\theta\to 90^o}\lim_{R\to\infty} P(E^{LDA}, R, \theta) &\approx 0.31 \qquad (23) \end{aligned}$$

i.e. when θ tends to 90^o, PCNSA fails completely while the performance of LDA degrades gracefully.

Now in case 2 ($\psi = 90^o$), $N_1 \perp N_2$ i.e. $N_1 = [0\ 1]^T$ and $N_2 = [1\ 0]^T$. So $\Sigma_1 = diag\{\lambda_{max}, \lambda_{min}\}$ while $\Sigma_2 = diag\{\lambda_{min}, \lambda_{max}\}$. This gives $P(E_1^{NSA}) \leq \int_{\cos\theta - \frac{k}{\sqrt{R}}}^{\cos\theta + \frac{k}{\sqrt{R}}}\mathcal{N}(z;0,1)dz$. For LDA, $\Sigma_w = \frac{\Sigma_1 + \Sigma_2}{2} = diag\{\frac{\lambda_{max}+\lambda_{min}}{2}, \frac{\lambda_{max}+\lambda_{min}}{2}\}$ so that W^{LDA} is along $(\mu_1 - \mu_2)$ i.e. $W^{LDA} = [\cos\theta\ \sin\theta]^T$. Thus we have $P(E_1^{LDA}) = \int_{\frac{\sqrt{R}}{2(\sqrt{R}\cos^2\theta+\sin^2\theta)}}^{\infty} \mathcal{N}(z;0,1)dz$. The expressions for $P(E_2)$ for both PCNSA and LDA have the "cos" replaced by "sin". Case 2, as also discussed earlier, is the worst case for LDA. The average error probabilities are plotted in figure 2(b). The LDA error probability in this case converges to a non-zero value which depends on θ, i.e. we get,

$$\lim_{R\to\infty} P(E^{LDA}, R, \theta) = \frac{\int_{\frac{\sec^2\theta}{2}}^{\infty} \mathcal{N}(z;0,1)dz + \int_{\frac{cosec^2\theta}{2}}^{\infty} \mathcal{N}(z;0,1)dz}{2} \tag{24}$$

The above limit is approximately the LDA curve (dotted line) shown in figure 2(b). PCNSA still works very well in this case, i.e. we have (using (**??**))

$$\lim_{R\to\infty} P(E^{NSA}, R, \theta) = 0 \quad \forall\ \theta \tag{25}$$

although the rate of convergence is much slower than in case 1.

5. DISCUSSION AND FUTURE WORK

Thus from the above analysis, we can conclude that PCNSA fails for small values of R (no null space) or when the distance between class means projected along ANS becomes small ($\theta \to 90^o$). We have included checks in steps 3 and 4 of our algorithm (section **??**) to avoid these two situations. For all other cases, its performance is superior or as good as SLDA as long as the query data follows the training data distribution. By evaluating the error probability expressions, one can choose between LDA and PCNSA for a given application or even use different algorithms for different class pairs in a multi-class classification problem. Also, since PCNSA defines a class specific metric, it has a better ability to detect "new" (untrained) classes. This has also been observed experimentally.

But in real applications, the model is never exact and so the ANS calculation is never exact. ANS is more sensitive to model variation than the LDA classification directions. So, performance of PCNSA when compared with LDA in real applications is not as good as that predicted by the analytical expressions. As part of future work, we hope to do a perturbation analysis similar to that done by [6]. We also intend to present results on real classification applications of using PCNSA and combining PCNSA-LDA using error probability expressions (not given here due to lack of space).

6. REFERENCES

[1] N. Vaswani, "A linear classifier for gaussian class conditional distributions with unequal covariance matrices," in *International Conference on Pattern Recognition*, 2002.

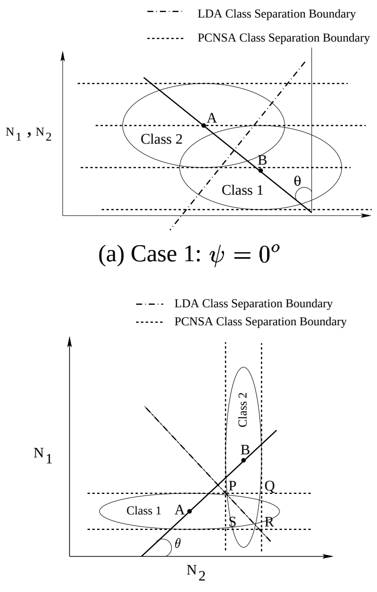
[2] W. Zhao, R. Chellappa, and P.J. Phillips, "Subspace linear discriminant analysis for face recognition," *IEEE Trans. on Image Processing*, 1999.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol 3, no. 1 1991.

[4] P. Belhumeur, J. Hespanha, and D. Kreigman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, July 1997.

[5] Alex M. Martinez and Avinash C. Kak, "Pca versus lda," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, February 2001.

[6] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical performance analysis of linear discriminant classifiers," in *IEEE CVPR*, 1998.

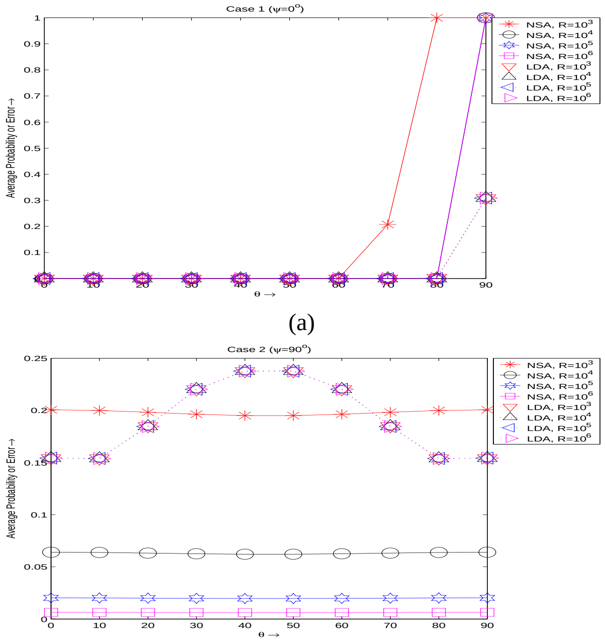


LDA Class Separation Boundary
PCNSA Class Separation Boundary
N1 , N2
Class 2
Class 1
A
B
θ
LDA Class Separation Boundary
PCNSA Class Separation Boundary
N1
N2
Class 2
Class 1
A
B
P
Q
R
S
θ

(a) Case 1: $\psi = 0^o$

(b) Case 2: $\psi = 90^o$

Fig. 1. In **Case 1**, ANS directions (Y-axis) of both classes coinciding. In **Case 2**, Y axis is ANS for class 1 & maximum variance direction for class 2, vice versa for X axis.



(a)

(b)

Fig. 2. Average probability of error as a function of θ for different values of condition number R for (a) Case 1 (b) Case 2. As can be seen the LDA error probability does not vary much with R in either case (curves for all R values are coincident) and also does not degrade much as $\theta \to 90^o$.