A Linear Classifier for Gaussian Class Conditional Distributions with Unequal Covariance Matrices

Namrata Vaswani
Center for Automation Research
Deptt of Electrical and Computer Engineering
University of Maryland, College Park
MD 20742, USA
namrata@cfar.umd.edu

Abstract

In this paper we present a linear pattern classification algorithm, Principal Component Null Space Analysis (PCNSA) which uses only the first and second order statistics of data for classification and compare its performance with existing linear algorithms. PCNSA first projects data into the PCA space in order to maximize between class variance and then finds separate directions for each class in the PCA space along which the class has the least variance (in an ideal situation the null space of the within class covariance matrix) which we define as the "approximate null space" (ANS) of the class. To obtain the ANS, we calculate the covariance matrix of the class data in PCA space and find its eigenvectors with least eigenvalues. The method works on the assumption that an ANS of the within-class covariance matrix exists, which is true for many classification problems. A query is classified as belonging to the class for which its distance from the class mean projected along the ANS of the class is a minimum. Results for PCNSA's superior performance over LDA and PCA are shown for object recognition.

1. Introduction

Within the last several years much progress has been made towards recognizing faces under small variations in lighting, facial expressions and pose. Both linear and non-linear algorithms have been proposed. Among linear pattern classification algorithms the most common one is principal component analysis (PCA) [1] which yields projection directions that maximize the total scatter across all classes but does not minimize the within class variance of each class and also sometimes retains directions with unwanted large variations due to lighting. An algorithm which encodes discriminatory information by finding directions that maximize the ratio of between class scatter to within-class scatter is linear discriminant analysis (LDA) [2]. [3] combines PCA and LDA to propose a subspace LDA (SLDA) based classification algorithm for face recognition which uses PCA first for dimensionality reduction and then LDA. [4] also uses subspace LDA for view based image retrieval from a database of real world objects. In [5] performance of PCA and LDA for classification is compared and superior performance of PCA for small or non-uniformly sampled training data is shown. Murase and Nayar [6] propose a compact representation of object appearance in which each object class is represented in the PCA space by a spline manifold. Independent Component Analysis (ICA) is compared with PCA in [7] for face recognition.

We present a new linear algorithm, principal component null space analysis (PCNSA) which finds separate directions for each class to minimize the within class scatter of the class and thus performs better than SLDA for applications like object recognition in which different classes have very different covariance matrix structures.

2. Motivation

PCA is optimal as a classification algorithm for applications where the within class noise can be modeled as white with the same variance in all classes. This is because PCA can filter out only out-of-band noise and not reduce noise in signal subspace. But this is the best that can be done if the noise in signal space is also white (all directions of signal space have same amount of noise). When the noise is colored but the noise (within-class variance) covariance matrix for the different classes is similar, LDA which finds common directions to simultaneously minimize the average within-class covariance matrix and maximize the between class covariance matrix (maximize signal to noise ratio) works best. Subspace LDA is a better solution because in most high dimensional(image) applications, there are a lot of directions with just noise and very little signal power (out of band noise) which can be thrown out by the initial PCA.

PCA and LDA are suitable for the 'apples from apples' type of classification problems where the noise covariance ma-

trices of all classes are similar (colored or white). In the 'apples from oranges' type applications like object recognition where this assumption is not satisfied, the minimum variance direction for one class might be a maximum variance direction for the other and hence all classes would not have really low variance along the common LDA direction. Within class variance occurs in image classification applications because of translations, rotations or affine deformations of the original template. PCNSA works on the assumption that the effect of these is not uniform in all directions (not white noise) and for all classes(different within class matrices) and finds for each class individually the directions along which the effect of these deformations is minimal. This is done by obtaining a 3-4 dimensional approximate null space (ANS) of the within-class noise covariance matrix for each class.

3. Problem Formulation

3.1. Noise Model

In this paper we assume that the data from each class has the most general Gaussian distribution (unequal means and unequal, non-white covariance matrices). Query belonging to class i (C_i), Q^i has a normal distribution

$$Q^i \sim N(M^i, \Sigma^i) \tag{1}$$

$$M^i = \begin{pmatrix} \mu^i \\ \mu \end{pmatrix} \tag{2}$$

$$\Sigma^i = \begin{pmatrix} \Sigma_w^i & 0 \\ 0 & \Sigma_{n,out} \end{pmatrix} \tag{3}$$

where M^i is the class mean, Σ_w^i is noise covariance matrix in signal subspace (directions along which μ^i's are different) [8] and $\Sigma_{n,out}$ is the out of band noise (noise along directions where there is no inter-class mean variation) which is filtered out by performing PCA. Hence, the query projected in the PCA space, $X^i = Q^i.W^{PCA}$ has a distribution

$$X^i \sim N(\mu^i, \Sigma_w^i) \tag{4}$$

3.2. Motivation

PCA is optimal as a classification algorithm for applications where the within class noise can be modeled as white i.e. $\Sigma_w^i = \sigma_i^2 I$ and hence there is no minimum noise direction in signal subspace. When the noise is colored but directions of minimum and maximum noise variance are same for all classes i.e. the eigenvalue decomposition (EVD) of $\Sigma_w^i = U^T \Lambda_i U$, SLDA is the optimal solution. But for 'apples from oranges' type classification applications like object recognition this assumption is not satisfied. The minimum variance direction for one class might be a maximum variance direction for the other, and all classes do not have really low variance along the common set of LDA directions. PCNSA addresses such applications by finding for each class separately, directions along which its noise variance is smallest. It exploits the fact that in most 'apples from oranges' type applications Σ_w^i is ill-conditioned i.e. the minimum variance is very close to zero and hence these directions can be said to form an 'approximate' null space (ANS) for the class. We use the distance from class mean projected along class ANS ($d^i(X) = ||(X - \mu^i)^T.N^i||^2$) as the classification metric. N^i is an orthogonal basis for the ANS of class i.

PCNSA works for the most general class covariance matrix, $\Sigma_w^i = U_i^T \Lambda_i U_i$ as long as (1) Σ_w^i's are ill-conditioned and hence an approximate null space exists for all classes and (2) The distance between class means projected in ANS of any class is non-zero. If (2) is not satisfied and if two classes have identical ANS's then $d^i(X)$'s would be always equal for any query X belonging to these two classes and the algorithm would fail for all queries from these two classes.

4. Algorithm

The stepwise PCNSA algorithm is given below.

- **Obtain PCA Space:** If $\mathbf{D}$ is a P^2 length observation vector, obtain a sample estimate of $M_D = E(\mathbf{D})$, and $\Sigma = Cov(\mathbf{D})$ the $P^2 \times P^2$ covariance matrix. Perform EVD to obtain the principal directions(PCA) , i.e. get W^{PCA} such that

  $$W^{PCA^T} \Sigma W^{PCA} = \Lambda_L^{max} \tag{5}$$

  where W^{PCA} is the $P^2 \times L$ PCA matrix and $\Lambda_L = Diag\{\lambda_1, \lambda_2, ...\lambda_L\}$ are the top L eigenvalues.

- Project the data of each class in PCA space and estimate sample mean and covariance of the class in PCA subspace i.e. for each class i, estimate

  $$\mu^i = E[(\mathbf{D^i} - M_D)W^{PCA}] \tag{6}$$

  $$\Sigma_w^i = Cov[(\mathbf{D^i} - M_D)W^{PCA}] \tag{7}$$

  where $\mathbf{D^i}$ is an observation of the i-th class.

- **Obtain Class ANS:** Do an EVD on Σ_w^i and keep the M least variance components i.e. obtain N^i such that

  $$N^{i^T} \Sigma_w^i N^i = \Lambda_i^{min} \tag{8}$$

  where N^i is the $L \times M$ orthonormal basis of ANS of class i and Λ_i^{min} is the $M \times M$ matrix containing the M least eigenvalues of the EVD, $M \approx 3, 4$.

- **Obtain Valid Classification Directions in ANS:** A null space direction of class i, e^i can be used in the classification metric only if condition (2) in section 3.2 is satisfied, which is equivalent to

$$|(\mu_i - \mu_j)^T e^i| > \cos\theta_0 ||\mu_i - \mu_j||, \ \ \forall j \neq i, \ \ |\theta_0| < 90^o \tag{9}$$

- **Classification:** Project the query Q first in PCA space to get $X = Q.W^{PCA}$ and choose the most probable class using the following distance metric

$$Class = \arg\min_i ||(X - \mu^i)^T.N^i||^2 \tag{10}$$

- **New Class Detection:** If distances from two or more classes are roughly equal, we conclude that the query belongs to a 'new' (untrained) class. This is because a query will have a very sharp minimum in its own class's ANS and if there is no such sharp minimum, then one can say that it does not belong to any of the trained classes. The metric we use for this is $d_{min} > td_i, \forall i \neq i_{min}$ with threshold t set at 0.5.

4.1. Discussion

We have derived error probability bounds for PCNSA and SLDA in [9] for a two class classification scenario with one ANS per class and one LDA direction. $P(E_1^{NSA})$, $P(E_1^{LDA})$ are probability of a query from class 1 being classified as belonging to class 2 using PCNSA, SLDA respectively. For PCNSA an upper bound has been evaluated while for SLDA an exact expression is obtained, hence showing the PCNSA bound to be lower than SLDA error probability suffices to show that PCNSA would have lower error probability.
The error probability expressions are:

$$P(E_1^{NSA}) \leq \int_{\frac{\alpha-\Delta}{\sigma}}^{\frac{\alpha+\Delta}{\sigma}} N(z;0,1)dz \tag{11}$$

where

$$\begin{aligned} \alpha &\triangleq |(\mu^1 - \mu^2)^T N^2| \\ \sigma &\triangleq \sqrt{N^{2T}\Sigma_w^1 N^2} \\ \Delta &\triangleq k\sqrt{\lambda_1} \end{aligned} \tag{12}$$

λ_1 is eigenvalue (variance) along ANS of class 1 (NS-1). λ_1 is very small and hence Δ is also very small.

$$P(E_1^{LDA}) = \int_{\frac{\check{\alpha}}{\check{\sigma}}}^{\infty} N(z;0,1)dz \tag{13}$$

where

$$\begin{aligned} W &\equiv W^{LDA} \\ \check{\alpha} &\triangleq \frac{|(\mu^2 - \mu^1)^T W|}{2} \\ \check{\sigma} &\triangleq \sqrt{W^T \Sigma_w^1 W} \end{aligned} \tag{14}$$

To compare the two methods, assume that α and $\check{\alpha}$ are roughly equal and large. Consider a case where the maximum within-class variance direction of class 1 is the null space (NS) direction of class 2 and vice versa. In such a case, as has been discussed in [9], both $\sigma = \sqrt{N^{2T}\Sigma_w^1 N^2}$ and $\check{\sigma} = \sqrt{W^T\Sigma_w^1 W}$ would be large. So $\frac{\check{\alpha}}{\check{\sigma}}$ would be small leading to high SLDA probability. This is the worst case for SLDA. But for PCNSA, the error probability depends on both $\frac{\alpha}{\sigma}$ and $\frac{\Delta}{\sigma}$. Even though $\frac{\alpha}{\sigma}$ would be small, but because Δ is small and σ is high, $\frac{\Delta}{\sigma}$ will be even smaller and hence the region of integration will be very small leading to a low error probability. In fact for large σ the PCNSA error probability can be approximated as $\sqrt{\frac{2}{\pi}}e^{\frac{-\alpha^2}{2\sigma^2}}\frac{\Delta}{\sigma}$. For the other extreme case where NS-1 and NS-2 directions coincide, both σ and $\check{\sigma}$ would be very small and both methods would have very small error probabilities as long as α is large. But if equation (9) is not satisfied, $\alpha = 0$ and so PCNSA would fail.
New(untrained) classes can be detected most easily using PCNSA because when a query belongs to a trained class its distance from class mean along that class's ANS is a very sharp minimum while a query belonging to a new class will have no such sharp minimum. Detecting new classes is more difficult with LDA because trained classes will also not have very sharp minimum distances from their own class means along LDA directions.
One problem with finding ANS directions for PCNSA is that one needs large amount of training data to correctly find directions along which there is almost no variation. Training data size per class should be at least 2-3 times the dimension of the PCA space to correctly estimate the lowest eigenvalues (and corresponding eigenvectors) of the class covariance matrix. This fact becomes more evident from figure 2.

The extra overhead for obtaining PCNSA or SLDA subspace in PCA space (highly reduced dimension data) is negligible compared to the initial principal eigenvectors calculation done on N^2 dimensional image data. The more important query classification time is proportional to the number of inner products (equal to total number of projection directions) to be taken. For a K class application, with L dim PCA space and 3 ANS directions per class, LDA requires a maximum of '$K-1$' N^2- dim inner products, PCNSA '$3K$' N^2-dim inner products while PCA requires 'L' N^2-

dim inner products. 'L' will be larger than '$3K$' for most applications. Hence PCNSA time complexity for classification is lower than that for PCA but higher than that for SLDA

5. Results

We compare the performance of PCNSA with that of SLDA and PCA for object recognition which is an example of the 'apples from oranges' type of problems for which PCNSA outperforms LDA and PCA.

The algorithm was tested on the Columbia Object Image Library (COIL) which contains 20 different objects and 72 views of each object taken at 5 degree apart orientations. Due to the entirely different covariance matrix structures of different objects, PCA and SLDA do not work too well while PCNSA performs very well.

Sequential testing was done by choosing 10 frames per class at a time for testing and the rest 62 for training and in this way a total of 1400 tests were carried out by choosing different test and training samples every time. Sample images from the 20 classes are shown in figure 1(a).

Table 1(a) shows the error probability (percentage of total data which got misclassified) when all 20 classes were trained and data from the same 20 classes was used for testing and new class detection was not done. In 1(b) we show results for the same data but with new class detected enabled. Now total error increases since some queries get wrongly classified as new. The probability that queries from trained classes get classified as 'new' is termed as 'miss probability'. 'Total error probability' is the total misses plus misclassifications and 'error (excluding misses)' is only misclassifications. In 1(c) we show results for training 16 classes and testing on data from all 20. 'New & Detected' is the percentage of new class queries that got detected as 'new'.

Error percentage both with and without new class detection is lowest for PCNSA and more than two times higher for SLDA and PCA. Miss probabilities are also lowest for PCNSA. In table 1(c), new class detection is best for PCNSA followed by SLDA and PCA.

6. Conclusions

A new linear algorithm for classification in colored noise, principal component null space analysis (PCNSA) is presented and its performance compared with that of PCA and subspace LDA. Superior performance of PCNSA is shown for applications with vastly different within-class covariance matrices ('apples from oranges' type problems) like object recognition for which PCA and SLDA fail. Computational complexity of PCNSA for classification is shown to be smaller than or equal to that of PCA but a little worse than SLDA. But PCNSA fails for small training data or

Figure 1: Samples from the different classes used for object recognition

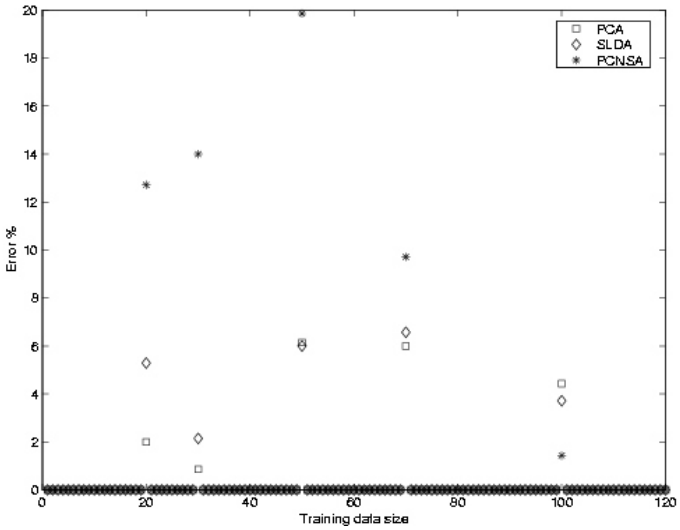

Figure 2: Error probability variation with reduced training data sizes per class

when number of classes is very large.

Performance of PCNSA can be improved further by combining it with LDA i.e. finding for each class, directions which not only minimize its variance but also maximize its distance from means of other classes.

References

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

[2] P.Belhumeur, J Hespanha, D Kreigman, "Eigenfaces vs. Fisherfaces: recognition Using Class Specific Linear Projection", *Trans PAMI*, vol. 19, no. 7, July 1997

[3] W.Zhao, R Chellappa, P.J. Phillips, "Subspace Linear Discriminant Analysis for Face Recognition", *IEEE Trans IP*, 1999

[4] D.L.Swets and J.J. Wengs, "Using Discriminant Eigenfeatures for Image Retrieval", *IEEE Trans. PAMI*, Vol 18, pp 831-836, Aug 1996

(a) NO NEW, NEW CLASS DETECTION DISABLED			
Method	PCA	SLDA	PCNSA
Error%	18.50	10.07	4.36
(b) NO NEW, NEW CLASS DETECTION ENABLED			
Method	PCA	SLDA	PCNSA
Miss%	46.21	28.71	13.43
Tot Error %	47.00	30.42	13.70
Error % (excluding misses)	0.79	1.71	0.27
(c) 4 NEW, NEW CLASS DETECTION ENABLED			
Method	PCA	SLDA	PCNSA
New&Det%	83.93	92.86	93.21
Miss%	39.64	23.00	13.07
Tot Error%	43.50	25.40	14.71
Error% (excluding misses)	3.86	2.40	1.64

Table 1: Object Recognition results for tests on data (a) from 20 trained classes (no new), (b) from 16 trained classes and 4 'new' classes

[5] Alex M. Martinez and Avinash C. Kak, "PCA versus LDA", *IEEE Trans. PAMI*, Vol 23, pp 228-233, Feb 2001

[6] Murase and Nayar, "Visual Learning and Recognition of 3-D Objects from Appearance", *IJCV*, vol 14, pp 5-24, 1995

[7] M.S. Bartlett, H.M. Lades, T.J Sejnowski, " Independent Component Representations for Face Recognition",*Proc of SPIE*, Vol 2399, pp 528-539, 1998

[8] G. Bienvenu, "Influence of the Spatial Coherence of Background Noise on High-resolution Passive Methods", *Proc ICASSP*, 1979, pp 306-309

[9] N. Vaswani, "A Linear Classifier for Gaussian Class Conditional Distributions with Unequal Covariance Matrices : Algorithm and Analysis", submitted to *Trans. PAMI*