Principal Discriminants Analysis for small-sample-size problems: application to chemical sensing

M. Wang, A. Perera and R. Gutierrez-Osuna

Department of Computer Science, Texas A&M University, USA, mqwang@tamu.edu, {aperera,rgutier}@cs.tamu.edu

Abstract

Two dimensionality reduction techniques are widely used to analyze data from chemical sensor arrays: Fisher's Linear Discriminants Analysis (LDA) and Principal Components Analysis (PCA). LDA finds the directions of maximum discrimination in classification problems, but has a tendency to overfit when the ratio of training samples to dimensionality is low, as is commonly the case in chemical sensor array problems. PCA is more robust to overfitting but, being a variance model, fails to capture discriminatory information in low-variance sensors. In this article we propose a hybrid model, termed Principal Discriminants Analysis (PDA), which incorporates both LDA and PCA criteria by means of a regularization parameter. The model is characterized on a synthetic dataset and validated with experimental data from an array of 15 metal-oxide sensors exposed to five varieties of roasted coffee beans. Our results show that PDA provides higher predictive accuracy than LDA or PCA alone. In addition, the model is able to find a trade-off between discriminant- and variance-based projections according to where information is located in the distribution of the data.

Index Terms—**Gas sensor arrays, principal component analysis, linear discriminant analysis, regularization.**

INTRODUCTION

The conventional feature extraction approach in chemical sensor arrays consists of allowing the sensors to stabilize following exposure to the analytes, and using their steady state values as a feature vector for further processing. However, several authors [1],[2] have shown that additional information can be extracted from the sensors if their time-dependent evolution is included in the feature-extraction stage. Still, feature extraction is an open problem, to where recent publications [3] continue to focus on the development of new techniques for chemical sensors signals. When transient information is used, the dimensionality of the feature space may become larger than the number of samples acquired for each target compound, resulting in a sparsely sampled feature space. Though not as highly dimensional as the datasets often encountered in image processing (e.g. face recognition [4], [5]), sensor-transient datasets are still prone to suffer from strong overfitting effects. In addition, there is a significant amount of noise and cross-sensitivity due to thermal oscillations, sampling differences across experiments, humidity and other interferents.

In order to facilitate the task of the classifier, the dimensionality of the feature space is usually reduced by means of a projection technique. The most common technique, Principal Component Analysis (PCA) [6],[7],

finds a subspace that contains most of the variance in the dataset. Fisher's Linear Discriminant Analysis (LDA) [8], is a supervised method that uses within-class variance information to build a projection. Though a powerful method, LDA has several drawbacks [9]. First, as a supervised technique, LDA has a tendency to overfitting in small-sample-size problems, where the dimensionality is higher than the number of vectors in the training set [10], [11]. A second, more structural problem with LDA occurs when the dataset has more information in the variance than in the mean. In this case, LDA also tends to produce poor results.

This paper proposes a hybrid criterion that incorporates both LDA and PCA criteria by means of a regularization parameter. We will show that this approach not only reduces the over-fitting effects of LDA, but also improves the classification performance of either PCA or LDA by finding a balance between discriminatory information in the mean and the variance for a given dataset.

PRINCIPAL DISCRIMINANTS ANALYSIS

The proposed method, termed Principal Discriminant Analysis (PDA), is based on the eigenvalue decomposition of the hybrid matrix H defined by:

$$H = (1-\varepsilon) \cdot S_w^{-1} S_b + \varepsilon \cdot S_T \qquad (1)$$

where S_w, S_b are the within- and between-scatter matrices [12], S_T is the total data covariance, and $\varepsilon \in [0,1]$ is a regularization parameter obtained through cross-validation.

For $\varepsilon=0$ the eigenvalues of H correspond to those of the LDA solution, whereas for $\varepsilon=1$ PDA finds the conventional PCA projection. For intermediate and increasing values of ε, the LDA solution is regularized by gradually incorporating variance information. An additional benefit of this regularization method is an increase in the number of non-zero eigenvalues beyond the upper limit of LDA (i.e., the number of classes minus one).

Selection of the regularization parameter ε is performed through cross-validation. Values of $\varepsilon \sim 1$ will indicate that discriminatory information is contained mostly in the variance of the dataset. On the other hand, $\varepsilon \sim 0$ will be characteristic of problems where most of the information is in the mean, and the variance of each class contains no discriminant information (i.e., all classes have equal covariance). Overall, the model will bias the solution towards PCA, LDA, or strike a balance between both, depending on where information is located in the dataset at hand.

RESULTS

We have characterized the model proposed in eq. (1) on two independent datasets. The first dataset set was generated artificially in order to establish proof of concept, and characterize the performance of PDA against PCA and LDA. The second dataset consists of experimental data from a gas sensor array exposed to the headspace of different roasted coffees. The objective of these experiments is to determine if a regularization method such as PDA is able to balance the contributions of the PCA and LDA solutions in datasets where information is located in both subspaces. In what follows, results will be given as a function of two parameters: the regularization parameter ε, and the ratio of number of samples per class to dimensionality ρ:

$$\rho = \frac{N_C}{D} \tag{2}$$

These parametric variations allow us to characterize the performance of the model as a function of the sparseness of the feature space, measured by the parameter ρ, and the structure of the projection (e.g., mean vs. variance), controlled by the regularization parameter ε. At low ρ, (e.g. $\rho=0.3$, 66 samples per class at $D=200$), LDA will tend to have poor predictive power as a result from over-fitting. Therefore, a sweep over ρ values will allow us to evaluate the relative sensitivity of PDA to dataset sparseness. Similarly, a sweep over ε will allow us to analyze the performance of the model as it evolves from a PCA projection towards an LDA projection, e.g. for datasets with large ρ and information primarily in the means, the model should favor the LDA solution.

Synthetic database

The first dataset is a small-sample-set problem with three Gaussian classes and fixed dimensionality $D=200$. We define an initial distribution with three overlapping classes $\mu_1^0 = \mu_2^0 = \mu_3^0 = [000]^T$, $\Sigma_1^0 = \Sigma_2^0 = \Sigma_3^0 = I$, where μ^0_i and Σ^0_i are the mean and covariance matrix for class i, respectively. This initial distribution represents the case where the classes are non separable, or a Bayes error rate of 2/3. Distributions with increasing levels of separability are obtained by morphing the initial distribution with a final distribution defined by:

$$\mu_1^f = \begin{bmatrix} -3 \\ 0 \\ 2 \end{bmatrix}, \mu_2^f = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}, \mu_3^f = \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix} \tag{3a}$$

$$\Sigma_1^f = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, \Sigma_2^f = \begin{bmatrix} 1.45 & -1.47 & 0 \\ -1.47 & 3.15 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \tag{3b}$$

$$\Sigma_3^f = \begin{bmatrix} 1.45 & 1.47 & 0 \\ 1.47 & 3.15 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \tag{3c}$$

according to the following interpolation procedure:

$$\Sigma_i = \beta \cdot \Sigma_i^f + (1-\beta) \cdot \Sigma_i^0,$$
$$\mu_i = \beta \mu_i^f + (1-\beta) \mu_i^0, \tag{4}$$

where β controls the amount of discriminatory information in the dataset. For the results shown in this section, four experiments were performed, corresponding to values of $\beta=\{0.25, 0.5, 0.75, 1\}$. These "intrinsic" distributions were generated in a three-dimensional space. To obtain the target dimensionality D=200, small Gaussian noise $N(0,\alpha)$ was added to the remaining 197 dimensions ($\alpha=0.1$).

Model performance was estimated using a nearest-neighbor (1NN) classifier, and a validation set containing 600 samples. The results are shown in Fig. 1. The intensity of the images in the left column denotes classification rate in % as a function of ρ and ε. Each case (i-iv) in Fig. 1 represents a different value of β, which controls the amount of discriminatory information in the mean and the variance of the data. The upper and bottom rows in Fig. 1 shows results for the dataset configuration with high ($\beta=0.25$) and low ($\beta=1$) class overlap, respectively. Bayes error rates, shown in Table 1, are estimated with a 1NN classifier using a separate dataset with 2000 training samples and 2000 validation samples (on the 3-dimensional space). Since the 1NN error rate is bounded by twice the Bayes error [12], the worst-case Bayes error can then be estimated as half the 1NN empirical error rate.
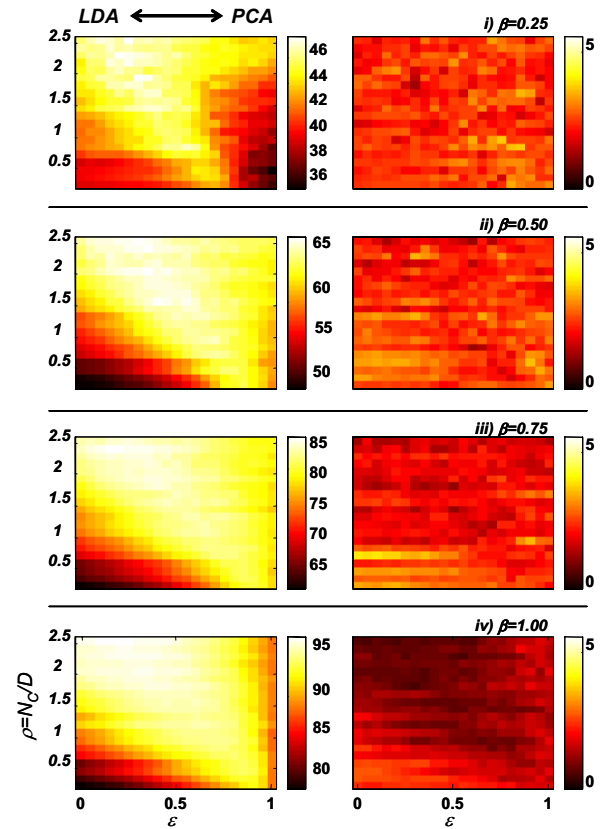

Fig. 1. (a) Classification rate in %, and (b) its standard deviation for the synthetic dataset.

Table 1. Bayes errors for the four cases under study

Case	β	Bayes Error	Case	β	Bayes Error
i)	0.25	0.26	iii)	0.75	0.07
ii)	0.50	0.16	iv)	1.00	0.01

As shown in Fig. 1, for all values of β, the classification performance of LDA ($\varepsilon=0$) increases with ρ, whereas the performance of PCA ($\varepsilon=1$) is lower, as could be expected. Note that, for any given ratio of samples/dimensions ρ, the maximum performance occurs at intermediate values of ε. This indicates that PDA is regularizing LDA in the small-sample-size case, or otherwise finding a balance between discriminant-based and variance-based projections.

The stability of the projection is illustrated in Fig. 1(b) in terms of the standard deviation of the 1NN performance as a function of (ε,ρ). Note that the standard deviation is lower than the difference in performance between PDA and either PCA and LDA, particularly in the small-sample cases (low ρ). It is also interesting to note that the standard deviation of the classifications decreases with the Bayes error.

These results indicate that PDA is able to find an improved projection as a balance between the PCA and LDA solutions. These results can be explained as follows. First, since the synthetic dataset contains significant discriminatory information in the covariance of the data, the second term in equation (1) allows PDA to capture some of this information. Second, LDA aligns the discriminant planes in the directions of the sub-space defined by the class-conditional means, while minimizing the within-class covariance of the projections. Overall, there exists a trade-off between mean and variance information, such that combination of the two solutions can provide better discrimination than either solution alone.

Gas sensor array dataset

The proposed PDA model was also validated on experimental data from a sensor-array system developed at NC State University [13]. The sensor array contained twelve metal-oxide sensors from Capteur (Dodcot, UK) (sensors AA20, AA25, CT05, CT23, G06, G07, CT03, CT04, CT22, LG09, LG10 and LG21) and three metal-oxide sensors from Figaro Engineering Inc. (Osaka, Japan) (sensors TGS2600, TGS2610 and TGS2620). The sensor array was exposed to five coffee beans varieties, namely Sulawesy, Kenya, Arabian, Sumatra and Colombia. Coffee beans samples were held in 30ml bottles, and the dynamic headspace was extracted with a constant flow of 0.1 lpm. The sampling cycle consisted of a wash cycle of 30 seconds, a reference cycle of 180 seconds, a sample cycle of 60 seconds and a final wash cycle of 10 seconds. A total of 225 samples were acquired over a period of five days, 45 samples for each variety of coffee. An example of the transient response of the sensor array is shown in Fig. 3(a). These transient responses were acquired at 10Hz. A 30-dimensional feature vector was obtained by selecting two samples ($t=16,56s$) from the transient waveform of each sensor. A PCA scatterplot of the data is shown in Fig. 3(b). From the figure, it is noticeable the high amount of drift in the data; the main clusters are not related with class information but rather with different days of acquisition. These first two projections contain 75% of the total variance in the dataset, which indicates that discriminatory information is contained in low-variance channels.

The procedure for validating PDA is similar to the one described earlier. The number of training samples per class was increased from 23 ($\rho=0.8$) to 43 ($\rho=1.4$), and the performance of PDA was estimated for different values of ε. For each given training-set size, the rest of available samples were used as a validation set for the 1NN classifier. Results of the classification performance on the validation set are shown in Fig. 2.

A final experiment was also performed by splitting the data into a training set (50% of the data), a validation set (25%), and a test set (25%). The dimensionality of the data was set to D=45. The scatter matrices in eq. (1) were obtained from the training set, whereas the regularization parameter was obtained from the performance of the 1NN classifier on the validation set. The test set was finally employed to estimate 1NN performance on the final PDA projection. PDA performed with a 64% classification rate on the test set, whereas LDA and PCA operated at 45% and 25%, respectively. In contrast with the synthetic case, the variance of the data does not provide discriminatory information, since the performance at $\varepsilon=1$ is close to that of a random classifier. Despite the lack of discriminatory information in the variance, PDA is nonetheless able to find an intermediate value of ε with higher performance than LDA. Therefore, in this case PDA works as a regularizer for LDA rather than as a mechanism to incorporate variance information.
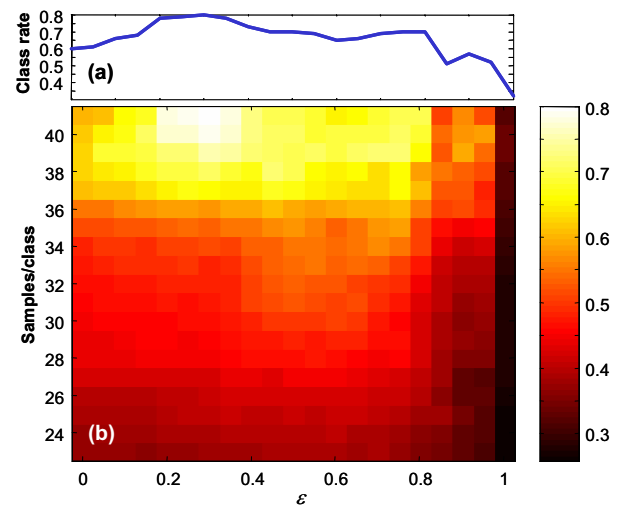

Fig. 2. (a) Classification rate at higher training set size ($\rho=1.37$, 41 samples per class) vs. ε. (b) Classification rate on validation data as a function of ρ and ε.
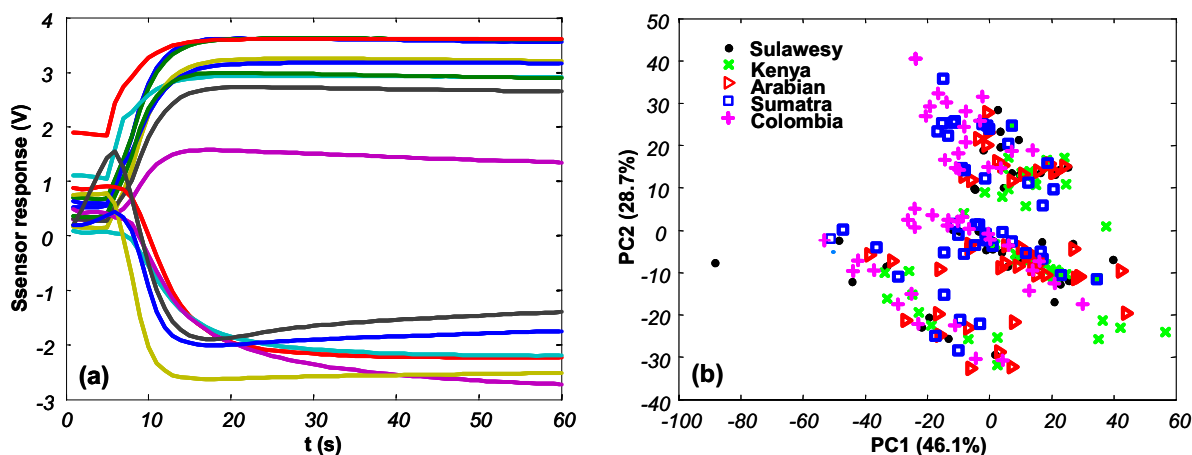
593

Fig. 3. (a) Example of the sensor-array transient response. (b) PCA scatterplot for the coffee dataset.

CONCLUSIONS

This paper has presented a novel technique for dimensionality reduction that provides a balance between discriminant-based and variance-based projections. This balance is established by regularizing the Fisher's Discriminants with the PCA solution. Selection of the regularization parameter through cross-validation allows the technique to be tuned to the specific distribution of information in a given dataset. The model has been characterized on a synthetic dataset, and validated on experimental data from an array of gas sensors improving individual performance of both variance-based and discriminant-based projections. The proposed method serves dual purpose. First, PDA prevents the Fisher's Discriminant projection from overfitting the training data by regularizing with the pooled covariance matrix. Second, and disregarding rank-deficiency issues, the model successfully balances mean and variance information according to the distribution of information in the data. Both scenarios are common in sensor array systems, where experimental data is usually limited and sensor signals are prone to suffer from strong cross-selectivity to interferents such as temperature and humidity.

REFERENCES

[1] W. M. Sears, K. Colbow, and F. Consadori. "*Algorithms to improve the selectivity of thermally-cycled tin oxide gas sensors*," Sensors and Actuators B, Chemical, 19(4):333–349, 1989

[2] A. Heilig, N. Bârsan, U. Weimar, M. Schweizer-Berberich, J. W. Gardner, and W. Göpel. "*Gas identification by modulating temperatures of SnO2-based thick film sensors*," Sensors and Actuators B, Chemical, 43(1–3):45–51, September 1997

[3] C. Distante, M. Leo, P. Siciliano, and K. Persaud. "*On the study of feature extraction methods for an electronic nose*," Sensors and Actuators B, Chemical, 87:274–288, 2002.

[4] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "*Face recognition using lda-based algorithms*," IEEE Trans. Neural Networks, vol. 14, pp. 195–200, 2003.

[5] Wen-Yi Zhao, Rama Chellappa, P.J. Jonathon Phillips, and Azriel Rosenfeld "*Face Recognition: A Literature Survey*", ACM Computing Survey, vol. 35, No 4, pp 399-458, 2003.

[6] I. Jolliff, "Principal Component Analysis". Springer-Verlag, 1986.

[7] M. Kendall, "*Multivariate Analysis*". Charles Griffin&Co., 1975.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, "*Pattern Classification*". New York: John Wiley & sons, Inc, 2nd ed., 2001

[9] W. Y. Zhao, "*Discriminant component analysis for face recognition*," in Pattern Recognition, 2000. 15th Int. Conf. on, vol. 2, (Barcelona, Spain), pp. 818–821, Sept 2000.

[10] A.K. Jain and B. Chandrasekaran, "*Dimensionality and sample size consideration in pattern recognition practice,*" in Handbook of Statistics, vol. 2. P.R. Krishnaiah and L.N. Kanal, Eds. Amsterdam, The Netherlands, pp. 835-855, 1982

[11] L. Kanal and B. Chandrasekaran, "*On dimensionality and sample size in statistical* pattern classification," Pattern Recognition, vol. 10, pp. 238—255, 1971

[12] K. Fukunaga, "Statistical Pattern Recognition," Ac. Press, 1990

[13] H. T. Nagle, R. Gutierrez-Osuna, B. G. Kermani and S. Schiffman, "Environmental Monitoring", Handbook of Machine Olfaction: Electronic Nose Technology, T. C. Pearce, S. S. Schiffman, H. T. Nagle and J. W. Gardner (Eds.), Wiley-VCH, pp 457-440, 2002.