HETEROSCEDASTIC DISCRIMINANT ANALYSIS

Submitted to

Dr. Joseph Picone
ECE8990 — Special Topics in ECE
Pattern Recognition

By
Issac Alphonso
Januarary 25, 2000

HETEROSCEDASTIC DISCRIMINANT ANALYSIS

Issac Alphonso

Critical Review Paper
ECE 8990 - Special Topics in ECE
Pattern Recognition
email: {alphonso}@isip.msstate.edu

ABSTRACT

Heteroscedastic discriminant analysis (HDA) is a technique that maximizes the class discrimination in the projected space, similar to linear discrimination analysis (LDA), without the assumption of equal sample covariances. The description of the algorithm looks promising given the fact that LDA is known to be inappropriate for classes with unequal sample covariances. This review will focus on two things, the fact that HDA fails to live up to its billing as a better discriminator and the maximum likelihood linear transformation that is employed along with HDA.

1. INTRODUCTION

Most speech recognition applications perform some form of per-processing on the data in order to extract features that can best model the data. Principle component analysis (PCA) is used by some applications because of its ability to pick out those dimensions that best represent the data and thereby reducing the dimensionality. Linear discrimination analysis is preferred in speech applications primarily because PCA seeks the directions that are efficient for representation where as LDA seeks directions that are efficient for discrimination. The paper in review, “Maximum Likelihood Discriminant Feature Spaces,” was written by G. Saon, M. Padmanabhan, R. Gopinath and S. Chen. In the paper the authors defined a new objective function, which is an extension of the work done by K. Nagendra [2], that they claim maximizes the class discrimination in the HDA space. However, the experimental results clearly show that HDA performs worse that LDA. The main reason for the improvement in the word error rate, on the Switchboard and Voicemail tasks, is due to the maximum likelihood linear transform (MLLT). In my opinion this paper should never have been published as all of the relevant information content could have been obtained from the references [2, 3].

This paper has been organized as follows: section 2 will describe the basics of LDA and explain why LDA perform best under the equal sample covariance assumption. Section 3 will describe the extension of LDA to HDA and section 4 will describe the MLLT. Finally section 5 will focus on the why the focus of this paper should have been the MLLT and not HDA.

2. FISHER’S LINEAR DISCRIMINANT

Linear discrimination analysis considers the problem of classifying *n* *d*-dimensional samples by reducing it into a more manageable *p*-dimension space ($p < n$) [6]. In two-dimensions LDA can be thought of as the projection of the samples onto a line. The goal of linear discrimination is to move the line around and find an orientation for which the projected samples are well separated.

If we have a set of *n* *d*-dimensional samples $\mathbf{x}_1 ... \mathbf{x}_n$ and if we use the samples in the set to form a linear combination of the components of $\mathbf{x}$, we obtain the scalar dot product $y = \boldsymbol{W}^t \boldsymbol{x}$ As you can see the direction of the vector **W** is importance in discriminating between the classes. Hence, our goal is simply the matter of finding the best possible direction of **W**.

In order to determine the best possible direction for **W** we define the *scatter matrices* $\mathbf{S_i}$ and $\mathbf{S_w}$. $\mathbf{S_i}$ is defined as a measure of the variability or scatter of the samples within the class

$$\boldsymbol{S}_i = \sum_{\boldsymbol{x} \in D_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^{\boldsymbol{t}}$$

and $\mathbf{S_w}$ is a measure of the total within-class variability or scatter and is given by

$$\boldsymbol{S}_w = \sum_{i=1}^{c} \boldsymbol{S}_i$$

Apart from the within-class scatter we define another *scatter matrix* called the between-class scatter. The between-class scatter $\mathbf{S_b}$ is a measure of the variability of the various class means w.r.t to the global mean and is given by

$$\boldsymbol{S}_b = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^t$$

Using the *scatter matrices* we define an objective function ***J*(.)** such that maximizing the objective function leads to the optimal value for **W**.

$$\boldsymbol{J}(\boldsymbol{W}) = \frac{\left|\boldsymbol{W}^t \boldsymbol{S}_b \boldsymbol{W}\right|}{\left|\boldsymbol{W}^t \boldsymbol{S}_w \boldsymbol{W}\right|}$$

Intuitively it can be seen that in order to maximize ***J*(.)** the class means need to be as far apart as possible (between-class) and the samples within the classes need to be tightly clustered (within-class). It can be shown that a vector **W** that maximizes ***J*(.)** must satisfy

$$\boldsymbol{S}_b \boldsymbol{W} = \lambda \boldsymbol{S}_w \boldsymbol{W}$$

Where λ represents the eigen values and **W** represents the eigen vectors of the between-class to the within-class ratio. In order to reduce the dimensionality we select the eigen vectors with the *p* largest eigen values (*p* < *n).*

Hence, LDA gives the value of **W** that yields the maximum ratio of the between-class scatter to within-class scatter. However, the estimate **W** can be shown to be dependent on the sample covariances i.e., given unequal sample covariances linear discrimination does not yield the direction of maximum discrimination.

Assume that the classes we are trying to discriminate are multinormal with equal covariances. The discriminant function for the classes can be given by

$$g_i(\boldsymbol{x}) = (-\tfrac{1}{2})(\boldsymbol{x} - \mu_i)^t \Sigma^{-1} (\boldsymbol{x} - \mu_i) + \log P(\omega_i)$$

$$g_i(\boldsymbol{x}) = \boldsymbol{w}^t{}_i \boldsymbol{x} + w_{i0}$$

Where $\mathrm{w_{io}}$ is a constant involving the **w** and the class priors

$$\boldsymbol{w}_i = \Sigma^{-1} \mu_i$$

$$w_{i0} = \left(-\frac{1}{2}\right) \mu^t{}_i \, \Sigma^{-1} \mu_i + \log P(\omega_i)$$

Using the individual class means and a common sample covariance yields a transform $\mathbf{w}_\mathrm{i}$ in the same direction as that of **W** [5].

3. EXTENSION TO HDA

Using linear discrimination analysis as an initial estimate the authors incorporated the individual weighted contributions of the classes using a new objective function. The objective function uses the original transform **W** generated by LDA and maximizes it with respect to the individual covariance of each class. The objective function is given by

$$\prod_{j=1}^{J} \frac{\left|\boldsymbol{W}\boldsymbol{S}_b\boldsymbol{W}^t\right|^{N_j}}{\left|\boldsymbol{W}\Sigma_j\boldsymbol{W}^t\right|} = \frac{\left|\boldsymbol{W}\boldsymbol{S}_b\boldsymbol{W}^t\right|^{N}}{\prod_{j=1}^{J} \left|\boldsymbol{W}\Sigma_j\boldsymbol{W}^t\right|^{N_j}}$$

Taking the log of the objective function gives us the following discriminant function

$$H(\boldsymbol{W}) = \sum_{j=1}^{J} -N_j \log\left|\boldsymbol{W}\Sigma_j\boldsymbol{W}^t\right| + N\log\left|\boldsymbol{W}\boldsymbol{S}_b\boldsymbol{W}^t\right|$$

Maximizing the discriminant function by taking the derivative gives us the following result

$$\frac{d}{d\boldsymbol{W}}H(\boldsymbol{W}) = \sum_{j=1}^{J} -2N_j\left(\boldsymbol{W}\Sigma_j\boldsymbol{W}^t\right)^{-1}\boldsymbol{W}\Sigma_j + 2N(\boldsymbol{W}\boldsymbol{S}_b\boldsymbol{W}^t)^{-1}\boldsymbol{W}\boldsymbol{S}_b$$

The above result however has no closed form solution and so it must be solved numerically.

Solving the above result numerically requires the use of constrained quadratic optimization algorithms [1]. A loose upper bound on the error used to maximize H(W) is given by $e^{\frac{-H(W)}{2N}}$

Hence, the authors hope that minimizing the upper bound on the error rate will maximize H.

4. LIKELIHOOD TRANSFORM

There are several places in the paper where the authors use a maximum likelihood (ML) linear transform to show how HDA does better that LDA in reducing the word error rate. In this section we will take a closer look at the MLLT and how it improves classification performance.

In speech recognition we generally assume that the underlying distribution is normal. There are several reasons for doing this among them being the minimum number of parameters of a gaussian, high entropy rate of a gaussian distribution and the fact that any distribution can be approximated by a mixture of gaussians.

Let $p(x, \{\mu_j\}, \{\Sigma_j\})$ represent a gaussian used to model class *j* of our training set. The ML principle maximizes the likelihood that the estimated mean and covariance are close to their true values. We can then use the trained gaussians to test for class membership on the test samples. In most cases the classes in the original space have a high overlap and hence classification yields poor results. However, in some cases the data can be transformed to a new space where the class separability is maximized. We can then model the training set in the feature space and use it to classify the test samples. However, it is difficult to compare the likelihood of a test sample given that the classes have been modeled in the transformed space. The problem is one of scaling and if we let A_j represent the transform we can always choose A_j such that $|A_j| = 1$ for every class. If we let A_j be a volume preserving transform we see that the likelihoods are equal.

$$p(x, \mu_j, \Sigma_j) = p(y, \mu_j, \Sigma_j)\prod_{j=1}^{J}\left|A_j\right|^{N_j}$$

When we add constrains to the ML principle, in our case a diagonal covariance, we can show that the inequalities

$$\left|diag(\bar{\Sigma})\right| \geq \left|\bar{\Sigma}\right| \quad \text{and} \quad p_{diag}(x) \leq p(x)$$

hold. Also, we can represent the equation of a normal distribution as

$$a(N, d)e^{-\frac{1}{2}\Phi}$$

where

$$\Phi = \sum_j N_j\left\{(\bar{\mu}_j - \mu_j)^t\Sigma^{-1}{}_j(\bar{\mu}_j - \mu_j) + Tr\left(\Sigma_j^{-1}\Sigma_j\right) + \log\left|\Sigma_j\right|\right\}$$

and

$$a(N, d) = 2\pi^{-\frac{Nd}{2}}$$

Hence our equation for the likelihood in the original space becomes

$$p_{diag}(x) = g(N, d)\prod_{j=1}^{J}\left|diag(\bar{\Sigma}_j)\right|^{-\frac{N_j}{2}}$$

and the likelihood in the feature space is

$$p_{diag}(y) = g(N, d)\prod_{j=1}^{J}\left|diag(\overline{A_j\Sigma}_jA_j^t)\right|^{-\frac{N_j}{2}}$$

We can see from the above equation that the best ML solution is a function of *A*. The likelihood can be maximized over *A* to obtain the best feature space in which to model the diagonal covariance constraint. By inspection we can show that one optimal choice for *A* is the eigen basis of the sample covariance where $\Lambda = A\bar{\Sigma}A^t$ and

$$p(y) = g(N,d)\prod_{j=1}^{J}|\Lambda|^{-\frac{N_j}{2}} = g(N,d)\prod_{j=1}^{J}|\bar{\Sigma}|^{-\frac{N_j}{2}} = p(x)$$

Hence, the likelihood in the original space which achieves the likelihood of a full-covariance is

$$p_{diag}(x) = g(N,d)\prod_{j=1}^{J}\left|A_j\right|^{N_j}\left|diag(\overline{A_j\Sigma}_jA_j^t)\right|^{-\frac{N_j}{2}}$$

Imposing diagonal gaussian models in the feature space reduces the storage and memory requirements. However, it comes at a loss of likelihood and does not discriminate since the model parameters are estimated independently. We can globally transform the data with a unimodular *A* and model the transformed data. The loss of likelihood can be minimized by searching for a suitable *A*. This intern forms the basis of the MLLT.

5. CONCLUSION

A closer inspection of the experimental results on the Switchboard task gives us more of an insight into performance of HDA.

Table 1: Word error rates for Switchboard.

System	WER
HDA	54.89%
LDA	43.16%
LDA+MLLT	40.46%
HDA+MLLT	39.67%

It is clear from the above results that HDA performs worse that LDA on the Switchboard task. The authors main objective in this paper is to convince us that HDA is a better discriminator that LDA given unequal sample covariances. However, the experimental data seem to suggest otherwise. HDA does however show an improvement over LDA, albeit a minor 0.79% improvement, when applied along with the MLLT. It is hard not to be skeptical about the implementation of HDA because: the improvement in performance is too small to be statistically significant and the MLLT appears to play a much larger role in the improvement than HDA.

In conclusion, I would say that although the concept behind HDA does sound promising the new objective function given by the author fails to deliver what it promises. In fact the most interesting thing about the paper is the MLLT of which not much was said. A better approach to the paper would have been to focus on the role played by the MLLT in reducing the word error rate. However, that approach was already covered in another paper [3].

REFERENCES

[1] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen, “Maximum likelihood discriminant feature spaces.”, *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. II1129-II1132, Beijing, China, 2000.

[2] N. Kumar and A. G. Andreou. “Heteroscedastic discriminant analysis and reduced rank HMM’s for improved speech recognition,” *Speech Communications*, pp. 23:283-297, 1998.

[3] R. A. Gopinath, “Maximum likelihood modeling with gaussian distributions for classification,” *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 2:661 -664, Seattle 1998.

[4] M. J. F. Gales, “Semi-tied covariance matrices for hidden Markov models,” *Proceedings of the IEEE Int. Conf. on Speech and Audio Processing,* pp. 7:272-281, 1999.

[5] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Wiley-Interscience Publishing, New York, New York, USA, 2000.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, New York, New York, USA, 1990.