On Generalizations of Linear Discriminant Analysis

*Nagendra Kumar and Andreas G. Andreou
Report JHU/ECE–96-07

April 8 , 1996

Abstract

Fisher–Rao Linear Discriminant Analysis (LDA), a valuable tool for multi-group classification and data reduction, has been investigated in the maximum likelihood framework. It has been shown that the LDA solution is a special case from the more general class of solutions. Generalizations of the LDA formulation have been proposed to handle the case where the within class variances are unequal, and their performance has been examined on randomly generated test data.

1 Introduction

Statistical pattern classification deals with classifying objects into different categories, based on certain observations made on the objects. For example on might want to classify furniture into different catagories of classes such as tables, desks and chairs. Here the classes are assumed to be mutually exclusive and exhaustive. The possible information available about the object is in terms on certain measurements made on the object. For example whenever a new piece of furniture arrives, there may be machines to measure the height, the width and the weight of the furniture. These three quantities would then form an observation or feature vector x. The three dimensional space to which the feature vectors belong is called the feature space. For the example above, is is known then none of the three features could possibly take a negative value. However, for the sake of convenience in classifier design, it is

*Electrical and Computer Engineering, Johns Hopkins University, Baltimore MD 21218 USA

often assumed that the feature space is the whole real axis. It is assumed that if the object (furniture) belongs to a certain class, then features take values according to a probability distribution that depends on the class. A statistical pattern classifier then assigns vectors x from $\Re_n$ (n dimensional feature space) to one of J classes. The assignment is performed so as to minimize the probability of error. Let $\Omega = \{\omega_1 \ldots \omega_J\}$ be the set of possible classes that could have given rise to the observation $x \in \Re_n$. Then based on the observation alone, assigning x to a class ω_j such that

$$P(\omega_j/x) > P(\omega_i/x) \, \forall i \neq j \tag{1}$$

minimizes the probability of error [1]. Decomposing (1) into priors and posterior, and noting that $P(x)$ cancels out on both the sides, one could write

$$P(x/\omega_j)P(\omega_j) > P(x/\omega_i)P(\omega_i) \, \forall i \neq j \tag{2}$$

if $P(x/\omega_j)$ and $P(\omega_j)$ were known, we could use (2) to design the classifier. However what is available are just some design examples in the form of training data. The approach to the classifier design is to estimate the priors $P(\omega_j)$ and the parameters of the class-conditional densities on the basis of the training data, and use them in the classifier described by (2) as if they were the true distributions.

The number of features that one may consider in a practical pattern classification problem is large. In the example of furniture, one could possibly think of several other features such as packaging material, reflectivities and color etc. In speech recognition, the dimensionality of acoustic features based on auditory models may range from a few tens to several hundreds. It is reasonable to believe that each of these features helps in some discrimination. If the parameters of the recognition-model were completely known, adding new features should not degrade the performance of pattern recognition system. At most, if the new features do not contain any information, they will be ignored.

Unfortunately, for pattern-classification problems, it has frequently been observed in practice that, beyond a certain point, the inclusion of additional features degrades the performance. The basic source of the problem can always be traced to the fact that the training data is finite, and increase in dimensionality increases the effective number of independent model parameters [2, 3] without increasing the amount of information available. This problem is often called "the curse of dimensionality"

One way to counter this problem is to find a transformation which maps the input feature vectors to output feature vectors of lower dimension, and which preserves as much discrimination information as possible about the classes associated with the input vectors. For simplicity only linear transformations may be considered.

A new method for finding the best linear transformations is described in this report. The linear projections are optimal in the maximum likelihood sense if it is assumed that the training data $x_i, i = 1 \ldots N$ are independent random occurrences and that the class conditional densities $P(x/\omega_j)$ are normally distributed with mean μ_j and variance matrix Σ_j. Section 2 gives a brief introduction to principle component analysis and multi-group linear discriminant analysis. The basic shortcomings of the two approaches have been pointed out. Section 3 describes and demonstrates the technique of Maximum likelihood parameter estimation. Using maximum likelihood technique, an objective function for finding the optimal linear projections in the case of heteroscedastic class distributions has been developed in section 4.

2 Dimension reduction through linear projections

One method commonly used for data compression and dimension reduction is principal component analysis (PCA) (also known as Karhunen-Loéve transform) [1]. The first principal component of a sample vector is the direction along which there is the largest variance over all samples. The hope behind this approach is that the direction along which there is maximum variation is also most likely to contain the information about class discrimination. The nth principal component is chosen to be a linear combination of the input features that has the largest variance under the constraint that it is also uncorrelated to the previous $n-1$ constraints.

One problem with principal components is that it's not independent of the parameter scale. If the nth parameter of the input feature vector is multiplied by some constant (greater than one), it cannot possibly affect the information contained in the nth parameter about the class with which the input feature vector is associated. But as this nth parameter is multiplied by larger and larger value, the variance of this parameter will become larger, and the first principal component will point in the direction of nth parameter although it may not contain any class discrimination information.

The general form of the problem is shown in figure 1. Suppose that the problem at hand is two way classification, and the within class distributions are Gaussians with equal variance in a two dimensional sample space. The ellipses in the figure represent contours of equal probability density for the two Gaussians. The line marked PCA is in the direction of maximum variance for each of the Gaussians. In this example it is

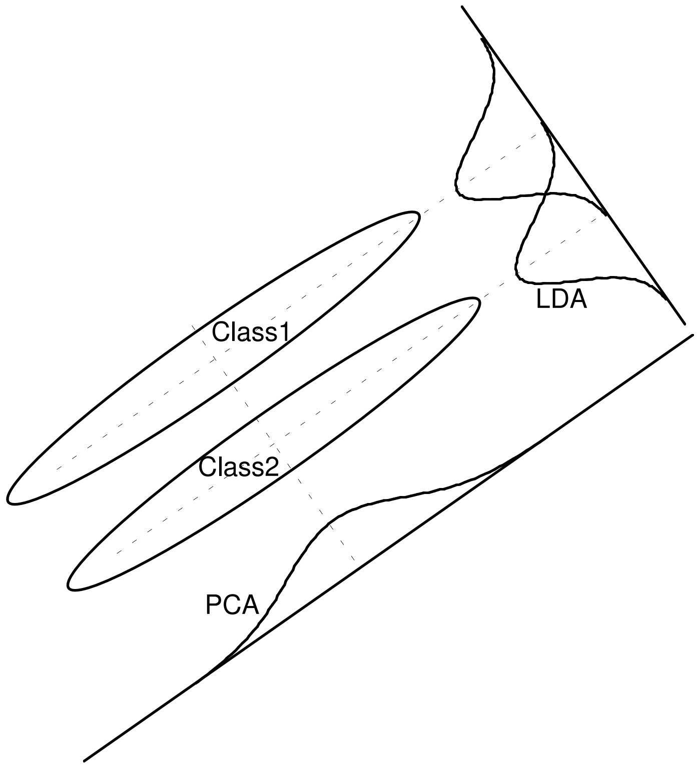

Figure 1: An illustration of how PCA fails in the case of two Gaussian classes

also in the direction of the maximum variance of the mixture of the two Gaussians, and hence in the direction of the first principal component. But a projection on this line contains no class discrimination information.

One way to solve this problem is to use methods such as linear discriminant analysis (LDA) that maximize the ratio of the overall variance to the within class variance [4, 5]. Since any full rank linear transformation will appear both in the numerator and the denominator of this ratio, and thus divide out, linear discriminants are invariant to any full-rank linear transformation of the input. let T be the ensemble sum-of-squares and cross products matrix as defined earlier. Let W be the average within class covariance matrix $W = \sum_{j=1}^{J}(N_j/N)W_j$. Then the between group variance $B = T - W$. Then it is desired to choose a projection $y = \hat{\theta}^T X$ that

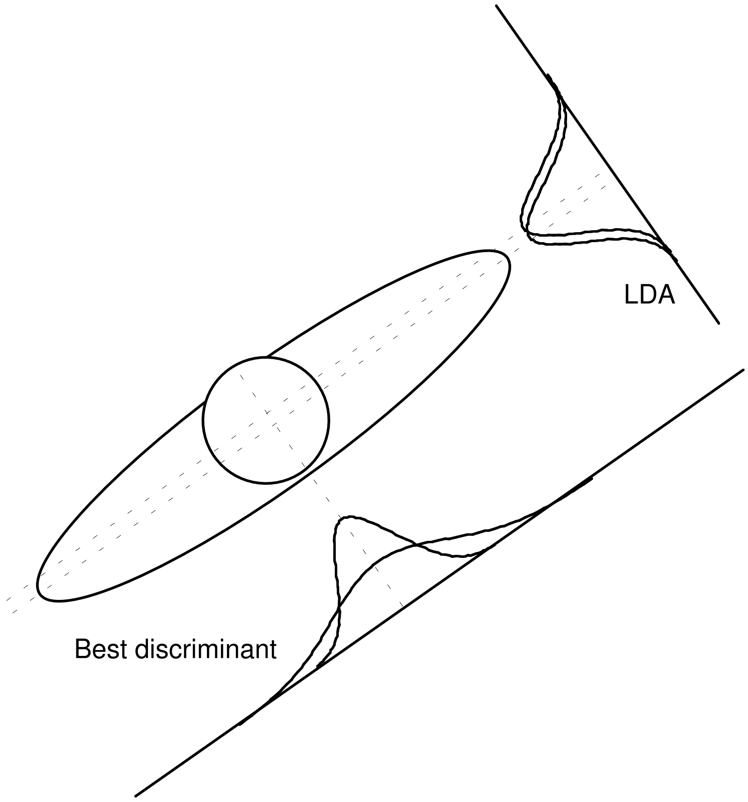

Figure 2: An illustration of how LDA fails in the case of two Gaussian classes

maximizes the ratio of between group to within group variance.

$$\hat{\theta} = \mathrm{argmax}_{\theta} \frac{\theta^T B \theta}{\theta^T W \theta} \tag{3}$$

It can be shown that the solution to the above equation corresponds to the right eigenvector of $W^{-1}B$ or $W^{-1}T$ that has the largest eigenvalue. By choosing the eigenvectors corresponding to the largest p eigenvalues, and letting $y_p = \hat{\theta}_p^T X$ where $\hat{\theta}_p$ is a $n \times p$ matrix of eigenvectors, we can get p dimensional uncorrelated features. The first linear discriminant direction is shown by the line marked LDA in figure 1. As one can clearly see, this indeed is the direction that maximally discriminates between the classes.

However LDA may also fail when the within class distributions are heteroscedastic. One such situation is depicted in figure 2. The two classes have *almost* the same mean, but the variances are very different in one direction. Due to this reason a classifier would perform better if a projection is taken along the direction in which the variances are different. However since LDA method pools the within group variances, it would choose the projection marked as LDA.

At this point there is one more important issue that needs to be considered. The choice of the optimal projection is closely tied to the classifier model that is being used for the subsequent classifier design. Consider the case of figure 2. Suppose that Gaussian models are being used to model each class, and the class variances are assumed to be equal for all classes. Then the classifier performance would still be best if the projection is chosen along the direction marked LDA. Of course this performance would be much worse compared to the best possible performance, due to bad modeling assumptions. However if the classifier model allowed for unequal variances, then the best choice of projection is different.

Campbell [6] has shown that if the the variance matrices Σ_j are assumed to be the same for all classes ($\Sigma_j = \Sigma$), then LDA is equivalent to imposing constraints on the class means that they all lie in a p-dimensional subspace and then performing maximum likelihood parameter estimation. Solution is chosen along the direction of the canonical variates that diagonalize both the pooled within-group and the between group variance matrices. Thus in effect it is assumed that a $n-p$ dimensional linear subspace of the feature space contains no class discrimination information and that in the p dimensional subspace that does contain the class discrimination information, the information is entirely in the class means, as the variances are assumed to be equal. Therefore LDA projections are best suited for a Gaussian model classifier that assumes equal within class variances.

3 Maximum Likelihood Parameter Estimation

Prior probabilities in (2) are taken as the relative frequency of occurrence. Hence the estimation problem basically reduces to the parameter estimation for the class conditional densities. Parameters would be estimated through maximum likelihood parameter estimation given that the classes are known for the training data. The estimate is defined to be the one that maximizes the likelihood of the training samples actually observed. Since maximizing the logarithm of a function is equivalent to maximizing the function itself, often the logarithm of the likelihood (log-likelihood) is maximized. Writing down in terms of the parameters the log-likelihood of the training data is

$$\begin{aligned}\log P(x_1,\dots,x_N) &= \log\{\prod_{i=1}^{N} P(x_i;\mu_{g(i)},\Sigma_{g(i)})\} \quad (4)\\ &= \sum_{i=1}^{N}(-\frac{1}{2}(x_i-\mu_{g(i)})^T\Sigma_{g(i)}^{-1}(x_i-\mu_{g(i)}) - \\ &\qquad \frac{1}{2}\log((2\pi)^n|\Sigma_{g(i)}|)) \quad (5)\end{aligned}$$

Here $g(i) = k$ if $x_i \in \omega_k$. Note that Σ_j are restricted to the class of symmetric positive definite matrices. Local maximum can be found by setting the derivatives of the log-likelihood w.r.t. the parameters to zero. However we would first rewrite the summation in (5) as

$$\sum_{j=1}^{J}\sum_{g(i)=j} -\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j) - \frac{1}{2}\log((2\pi)^n|\Sigma_j|) \quad (6)$$

Let N_j be the number of data points in the training data that belong to the class j ($\sum_{j=1}^{J} N_j = N$). Then (6) can be rewritten as

$$\frac{-Nn}{2}\log 2\pi - \frac{1}{2}\sum_{j=1}^{J}\left[\left(\sum_{g(i)=j}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\right) - \frac{N_j}{2}\log|\Sigma_j|\right] \quad (7)$$

Differentiation of the above expression with respect to the means and the variances can be performed using the matrix differentiation results listed in Appendix A. this leads to the equations

$$\begin{aligned}\frac{\partial}{\partial\mu_j}\log P(\{x_i\}) &= \sum_{g(i)=j}\Sigma_j^{-1}(x_i-\mu_j) \quad j = 1\dots J \quad (8)\\ \frac{\partial}{\partial\Sigma_j^{-1}}\log P(\{x_i\}) &= \frac{N_j\Sigma_j}{2} - \frac{1}{2}\sum_{g(i)=j}(x_i-\mu_j)(x_i-\mu_j)^T \quad j = 1\dots J \quad (9)\end{aligned}$$

For convenience, we shall define $\bar{X}_j = \frac{1}{N_j}\sum_{g(i)=j} x_i$, $W_j = \sum_{g(i)=j}(x_i - \bar{X}_j)(x_i - \bar{X}_j)^T$, $\bar{W}_j = W_j/N_j$, $W = \sum_{j=1}^{J} W_j$, $\bar{W} = W/N$, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $T = \sum_{i=1}^{N}(x_i - \bar{X})(x_i - \bar{X})^T$, $\bar{T} = T/N$ and $B = T - W = \sum_{j=1}^{J} N_j(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^T$.

By setting the derivatives equal to zero, these equations can be easily solved to yield the estimates of μ_j and Σ_j as

$$\hat{\mu}_j = \bar{X}_j \quad j = 1 \dots J \tag{10}$$

$$\hat{\Sigma}_j = \frac{W_j}{N_j} \quad j = 1 \dots J \tag{11}$$

Since the maximum likelihood values of the parameters have been computed, they can be used in (2) for the classification decision. In the case where feature dimension is reduced (for example by using methods in section 2), the parameters are estimated for the reduced data.

4 Generalizations of LDA

As mentioned in section 2, LDA assumes the within-class variances to be equal. However it is not necessary to assume all the within class variances to be the same. In the this section I present a generalization of LDA that can handle heteroscedastic distributions by dropping the equal variance assumptions. Due to computational simplicity, within class variances are often assumed to be diagonal. This is especially true in the case of speech recognition using hidden Markov models. Therefore the optimal projections for this case have also been considered. However due to increased complexity of the problem, eigen-analysis is not sufficient, and other numerical techniques have to be used to find the optimal projections. It has been shown that the objective function can be greatly simplified to speed up the numerical optimization.

Let θ be a non-singular linear transformation that transforms the data variables x into new variables y. The purpose of introducing the linear transform is to allow us to impose certain constraints on the means and the variances in order to simplify the statistical model. We would like these constraints to be not too restrictive, and at the same time allow us some reduction in the recognizer complexity. One such constraint is to assume that the means lie in a p-dimensional subspace. If the within class variances are assumed to be equal, then the assumption is equivalent that all the discrimination information is contained in a p dimensional linear subspace of the n dimensional feature space, and the remaining $n-p$ dimensional subspace may be ignored. We also choose the full rank linear transformation θ such that the first p columns of θ span the same p-dimensional sub-space. Since θ allows for rotations, the constraints on the mean are not too restrictive. For the sake of notational convenience we would partition the parameter space of the means μ_j, variances Σ_j and the linear transformation θ as follows:

$$\mu_j = \begin{bmatrix} \mu_{j,1} \\ \vdots \\ \mu_{j,p} \\ \mu_{0,p+1} \\ \vdots \\ \mu_{0,n} \end{bmatrix} = \begin{bmatrix} \mu_j^p \\ 0_{n-p\times 1} \end{bmatrix} + \begin{bmatrix} 0_{p\times 1} \\ \mu_0 \end{bmatrix} \tag{12}$$

$$\Sigma_j = \begin{bmatrix} \Sigma^p_{j(p\times p)} & \Sigma_{j(p\times n-p)} \\ \Sigma_{j(n-p\times p)} & \Sigma^{(n-p)}_{j(n-p\times n-p)} \end{bmatrix} \tag{13}$$

$$\theta = [\theta_p \theta_{n-p}] = \left[\vec{\theta}_1 \ldots \vec{\theta}_n\right] \tag{14}$$

where μ_0 is the common term in all the means (first p entries are zero), and μ_j^p are different for each class (last $n-p$ entries are zero). Σ_j have also been partitioned in the corresponding manner. The first p columns of theta are written as θ_p and the remaining as θ_{n-p} or as $\vec{\theta}_k$ to denote the k'th column.

Then the log-likelihood of the data under the linear transformation and Gaussian model assumption is

$$\frac{-Nn}{2}\log 2\pi + N\log|\theta| - \frac{1}{2}\sum_{j=1}^{J}\left(\sum_{g(i)=j}(\theta^T x_i - \mu_j)^T \Sigma_j^{-1}(\theta^T x_i - \mu_j)\right) - \frac{N_j}{2}\log|\Sigma_j| \tag{15}$$

Constraints will now be placed on the variances Σ_j and the means μ_j, so that the best discriminant projections are obtained for the models we wish to consider.

4.1 Σ constrained to diagonal

Suppose we assume the following. $\Sigma_{j(p\times n-p)}$ is zero; and $\Sigma_j^{(n-p)} = \Sigma^{(n-p)}\ \forall j$; and Σ_j^p and $\Sigma^{(n-p)}$ are diagonal matrices such that $\Sigma_j = \text{Diag}(\sigma_j^1 \ldots \sigma_j^p\ \sigma^{p+1} \ldots \sigma^n)$. These assumptions in effect are that the discrimination information lies entirely in a p-dimensional subspace, and is contained both in the value of the class means, and the variances (but the co-variance matrix is diagonal). Then in terms of the matrix partitions above the log-likelihood of the data can be written as

$$\log P_D(\{x_i\}) = \frac{-Nn}{2}\log 2\pi + N\log|\theta| - \frac{N}{2}\sum_{k=p+1}^{n}\log|\sigma^k|$$
$$-\sum_{j=1}^{J}\frac{N_j}{2}\sum_{k=1}^{p}\log|\sigma_j^k| - \frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}\sum_{k=1}^{p}\frac{(\vec{\theta}_k^T x_i - \mu_{j,k})^2}{\sigma_j^k}$$
$$+\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}\sum_{k=p+1}^{n}\frac{(\vec{\theta}_k^T x_i - \mu_{0,k})^2}{\sigma^k} \tag{16}$$

To find a maxima, we would differentiate the above expression w.r.t. the various parameters.

$$\frac{\partial}{\partial \mu_j^p}\log P_D(\{x_i\}) = -\sum_{g(i)=j}(\Sigma_j^p)^{-1}(\theta_p^T x_i - \mu_j^p)$$
$$j = 1\ldots J \tag{17}$$
$$\frac{\partial}{\partial \mu_0}\log P_D(\{x_i\}) = -\sum_{i=1}^{N}(\Sigma^{n-p})^{-1}(\theta_{n-p}^T x_i - \mu_0) \tag{18}$$
$$\frac{\partial}{\partial \sigma_j^k}\log P_D(\{x_i\}) = -\frac{N_j}{2\sigma_j^k} + \sum_{g(i)=j}\frac{(\vec{\theta}_k^T x_i - \mu_{j,k})^2}{2(\sigma_j^k)^2}$$
$$j = 1\ldots J, k = 1\ldots p \tag{19}$$
$$\frac{\partial}{\partial \sigma^k}\log P_D(\{x_i\}) = -\frac{N}{2\sigma^k} + \sum_{i=1}^{N}\frac{(\vec{\theta}_k^T x_i - \mu_{0,k})^2}{2(\sigma^k)^2}$$
$$k = p+1\ldots n \tag{20}$$
$$\frac{\partial}{\partial \theta}\log P_D(\{x_i\}) = N\theta^{-T} + \sum_{j=1}^{J}\sum_{g(i)=j} -x_i x_i^T \theta \Sigma_j^{-1}$$
$$+\sum_{j=1}^{J}\sum_{g(i)=j}\frac{1}{2}x_i\mu_j^T\Sigma_j^{-1} + \frac{1}{2}\Sigma_j^{-1}\mu_j x_i^T \tag{21}$$

Setting the derivatives equal to zero and solving for the values of μ_j and Σ_j gives the following.

$$\mu_j^p = \theta_p^T \bar{X}_j \tag{22}$$

$$
\begin{aligned}
\mu_0 &= \theta_{n-p}^T \bar{X} &(23)\\
\Sigma_j^p &= \mathrm{Diag}(\theta_p^T \bar{W}_j \theta_p) \quad j = 1 \ldots J &(24)\\
\Sigma^{n-p} &= \mathrm{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p}) &(25)
\end{aligned}
$$

Note that the μ's can be calculated if θ is known, and σ's if μ's are known. Therefore, we would first like to solve for θ. Substituting values of all the maximized parameters except θ gives the likelihood of the data as

$$
\begin{aligned}
\log P_{D,\theta}(\{x_i\}) &= \frac{-Nn}{2}\log 2\pi - \frac{N}{2}\log|\mathrm{Diag}(\theta_{n-p}^T \bar{T}\theta_{n-p})| - \sum_{j=1}^{J}\frac{N_j}{2}\log|\mathrm{Diag}(\theta_p^T \bar{W}_j \theta_p)| \\
&\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i - \bar{X}_j)^T \theta_p \mathrm{Diag}(\theta_p^T \bar{W}_j \theta_p)^{-1}\theta_p^T (x_i - \bar{X}_j)) \\
&\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i - \bar{X})^T \theta_{n-p} \mathrm{Diag}(\theta_{n-p}^T \bar{T}\theta_{n-p})^{-1}\theta_{n-p}^T (x_i - \bar{X})) \\
&\quad + N\log|\theta| \qquad (26)
\end{aligned}
$$

The maximum likelihood estimate for θ can now be found by maximizing the above likelihood numerically. Finding a numerical solution for θ is discussed in section 4.4. However before we discuss how to obtain a numerical solution that maximizes (26) it is possible to make certain observations about the nature of the linear transformation θ that maximizes the likelihood.

Note that from (21)

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\log P_D(\{x_i\})\frac{\theta^T}{N} &= I_n - \sum_{j=1}^{J}\frac{N_j}{N}(\bar{W}_j + \bar{X}_j \bar{X}_j^T)\theta\Sigma_j^{-1}\theta^T + \\
&\quad \sum_{j=1}^{J}\frac{N_j}{2N}(\bar{X}_j \mu_j^T \Sigma_j^{-1} + \Sigma_j^{-1}\mu_j \bar{X}_j^T)\theta^T \qquad (27)
\end{aligned}
$$

To solve for θ at the extremums the values of μ_j and Σ_j from (22, 23, 24, 25) are now substituted in (27) and the result equated to zero. Since Σ_j have been partitioned into four parts, the resulting implicit equations for θ are best written by dividing them into the following four equations.

$$
I_p = \sum_{j=1}^{J}\frac{N_j}{N}\theta_p^T W_j \theta_p (\mathrm{Diag}(\theta_p^T W_j \theta_p))^{-1} \qquad (28)
$$

$$
\begin{aligned}
0 &= \sum_{j=1}^{J} \frac{N_j}{N} \theta_{n-p}^T W_j \theta_p (\mathrm{Diag}(\theta_p^T W_j \theta_p))^{-1} && (29) \\
I_{n-p} &= \theta_{n-p} T \theta_{n-p} (\mathrm{Diag}(\theta_{n-p}^T T \theta_{n-p}))^{-1} && (30) \\
0 &= \theta_p T \theta_{n-p} (\mathrm{Diag}(\theta_{n-p}^T T \theta_{n-p}))^{-1} && (31)
\end{aligned}
$$

Since multiplying by the inverse of the diagonal entries corresponds to normalizing the diagonal entries of a matrix, (30) means that $\theta_{n-p} T \theta_{n-p}$ is a diagonal matrix. Also (31) simply means $\theta_p T \theta_{n-p} = 0$. Equations (28) and (25) are more difficult to interpret, due to summation. However an analogy could be drawn by considering the case when Σ_j are all assumed to be equal. This leads to the very well known case that has the Fisher-Rao LDA as a solution.

4.2 Σ's constrained to be equal

If we additionally require that $\Sigma_j = \Sigma \; \forall j$ the equations (24), (28) and (29), would change to the following

$$
\begin{aligned}
\Sigma^p &= \frac{1}{N} \mathrm{Diag}(\theta_p^T W \theta_p) && (32) \\
I_p &= \theta_p^T W \theta_p (\mathrm{Diag}(\theta_p^T W \theta_p))^{-1} && (33) \\
0 &= \theta_{n-p}^T W \theta_p (\mathrm{Diag}(\theta_p^T W \theta_p))^{-1} && (34)
\end{aligned}
$$

Suppose we know some real $\hat{\theta}$ such that both $\hat{\theta}^T T \hat{\theta}$ and $\hat{\theta}^T W \hat{\theta}$ are diagonal matrices. Under our usual notation let $\hat{\theta}_p$ denote the first p columns and so on. Then such a $\hat{\theta}$ is solution to the equations (33) and (34). This kind of $\hat{\theta}$ is called the generalized eigenvector of the matrices T and W and correspond to the well known Fisher-Rao LDA solution. However the constraints above require only $\theta_p^T W \theta_p$ to be diagonal. $\theta_{n-p}^T W \theta_{n-p}$ may not necessarily be diagonal and (30), (31), (33) and (34) may still be satisfied. Similar argument holds for $\theta_p^T T \theta_p$. Hence one can conclude that the choice of θ is not unique. It is reasonable indeed, because given any projection θ_p, any full rank linear transform of θ_p is also an equally good projection.

Also note that eigenvectors are usually scaled to have the same norm. There is nothing in the above equations that would restrict the columns of θ to have the same norm. Therefore for numerical reasons it is desirable to fix the columns of θ to have the same norm when solving for the maximum likelihood solution. Note that the equations above do give any hints as to what portion of the columns of θ becomes θ_p, and and any such combination is an acceptable solution. Since it is known that the likelihood is maximized when those columns of θ that correspond to larger eigenvalues of $W^{-1}B$ form θ_p, it is reasonable to assume that the other solutions are saddle points of the likelihood surface. Also, as would also be demonstrated through numerical examples, there are other solutions to the above equations that give the same value for the maximum likelihood as the generalized eigenvectors.

To get a numerical solution for $\hat{\theta}$, the estimated values of μ_j and Σ are substituted in (16) to get the log-likelihood of the data in terms of θ as

$$\begin{aligned}
\log P_{E,\theta}(\{x_i\}) &= \frac{-Nn}{2}\log 2\pi + \frac{N}{2}\log|\mathrm{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p})| + \frac{N}{2}\log|\mathrm{Diag}(\theta_p^T\bar{W}\theta_p)| \\
&\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X}_j)^T\theta_p\mathrm{Diag}(\theta_p^T\bar{W}\theta_p)^{-1}\theta_p^T(x_i-\bar{X}_j)) \\
&\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X})^T\theta_{n-p}\mathrm{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p})^{-1}\theta_{n-p}^T(x_i-\bar{X}_j)) \\
&\quad -N\log|\theta| \qquad (35)
\end{aligned}$$

This likelihood can be maximized with respect to θ.

4.3 Σ_j unequal

The constraint or the model assumption that the Σ's be diagonal is very useful from the practical point of view, because then matrix inversion can be replaced by simple division. However under certain circumstances it may be desirable to model the data as having full variance-covariance matrices that are not simultaneously diagonalizable. In that case, the basic problem can be re-formulated as in the previous section, but with dropping off the assumption that Σ_j^p and $\Sigma^{(n-p)}$ are diagonal matrices. Then log-likelihood of the data can be written as

$$\begin{aligned}
\log P_U(x_i) &= \frac{-Nn}{2}\log 2\pi + \frac{N}{2}\log|\Sigma^{n-p}| + \sum_{j=1}^{J}\frac{N_j}{2}\log|\Sigma_j^p| + N\log|\theta| \\
&\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}(\theta_p^T x_i-\mu_j^p)^T(\Sigma_j^p)^{-1}(\theta_p^T x_i-\mu_j^p) +
\end{aligned}$$

$$(\theta_{n-p}^T x_i - \mu_0)^T (\Sigma^{n-p})^{-1} (\theta_{n-p}^T x_i - \mu_0) \tag{36}$$

Again we would differentiate the above expression w.r.t. the various parameters.

$$\begin{aligned}
\frac{\partial}{\partial \mu_j^p} \log P_U(\{x_i\}) &= -\sum_{g(i)=j} (\Sigma_j^p)^{-1} (\theta_p^T x_i - \mu_j^p) \quad j = 1 \ldots J && (37)\\
\frac{\partial}{\partial \mu_0} \log P_U(\{x_i\}) &= -\sum_{i=1}^{N} (\Sigma^{n-p})^{-1} (\theta_{n-p}^T x_i - \mu_0) && (38)\\
\frac{\partial}{\partial (\Sigma_j^p)^{-1}} \log P_U(\{x_i\}) &= N_j \Sigma_j^p - \sum_{g(i)=j} (\theta_p^T x_i - \mu_j^p)(\theta_p^T x_i - \mu_j^p)^T \\
&\quad j = 1 \ldots J && (39)\\
\frac{\partial}{\partial (\Sigma^{n-p})^{-1}} \log P_U(\{x_i\}) &= N \Sigma^{n-p} - \sum_{i=1}^{N} (\theta_{n-p}^T x_i - \mu^{n-p})(\theta_{n-p}^T x_i - \mu^{n-p})^T && (40)\\
\frac{\partial}{\partial \theta_p^T} \log P_U(\{x_i\}) &= \sum_{j=1}^{J} \sum_{g(i)=j} (\Sigma_j^p)^{-1} (\theta_p^T x_i x_i^T - \mu_j^p x_i^T) - (\theta^{-T})_p && (41)\\
\frac{\partial}{\partial \theta_{n-p}^T} \log P_U(\{x_i\}) &= \sum_{i=1}^{N} (\Sigma^{n-p})^{-1} (\theta_{n-p}^T x_i x_i^T - \mu_0 x_i^T) - (\theta^{-T})_{n-p} && (42)\\
& && (43)
\end{aligned}$$

Setting these derivatives to zero and solving for μ_j and Σ_j gives the following.

$$\begin{aligned}
\mu_j^p &= \theta_p^T \bar{X}_j && (44)\\
\mu_0 &= \theta_{n-p}^T \bar{X} && (45)\\
\Sigma_j^p &= \frac{1}{N_j} (\theta_p^T W_j \theta_p) \quad j = 1 \ldots J && (46)\\
\Sigma^{n-p} &= \frac{1}{N} \theta_{n-p}^T T \theta_{n-p} && (47)\\
& && (48)
\end{aligned}$$

Substituting the values of the these parameters in (36) gives the likelihood of the data in terms of θ as

$$
\begin{aligned}
\log P_{U,\theta}(\{x_i\}) \;=\; & \frac{-Nn}{2}\log 2\pi - \frac{N}{2}\log|(\theta_{n-p}^T \bar{T}\theta_{n-p})| - \sum_{j=1}^{J}\frac{N_j}{2}\log|(\theta_p^T \bar{W}_j\theta_p)| \\
& -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X}_j)^T\theta_p(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T(x_i-\bar{X}_j)) \\
& -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X})^T\theta_{n-p}(\theta_{n-p}^T\bar{T}\theta_{n-p})^{-1}\theta_{n-p}^T(x_i-\bar{X})) \\
& +N\log|\theta| \qquad (49)
\end{aligned}
$$

This expression can now be maximized to get the maximum likelihood solution for θ. To make some general observations about the nature of θ, substitute the values of μ_j and Σ_j in the equations (41) and (42) to get the following equations:

$$
\begin{aligned}
I_p &= \sum_{j=1}^{J}\frac{N_j}{N}\theta_p W_j\theta_p(\theta_p^T W_j\theta_p)^{-1} \qquad (50)\\
0 &= \sum_{j=1}^{J}\frac{N_j}{N}\theta_{n-p} W_j\theta_p(\theta_p^T W_j\theta_p)^{-1} \qquad (51)\\
I_{n-p} &= \theta_{n-p}T\theta_{n-p}(\theta_{n-p}^T T\theta_{n-p})^{-1} \qquad (52)\\
0 &= \theta_p T\theta_{n-p}(\theta_{n-p}^T T\theta_{n-p})^{-1} \qquad (53)
\end{aligned}
$$

Equations (50) and (52) are obviously redundant. The other two equations indicate the constraints that the full rank linear transform θ will have to satisfy.

4.4 Numerical Solution for θ

From the previous discussion it is clear that once θ is known, there are closed form expression for both the means μ_j and the variances Σ_j. However there is no closed form expression for θ based on the data. Therefor on has to use numerical methods for finding θ. For that purpose one could try to use the general methods for solving non-linear equations like (28), (29), (30), and (31). However as noted earlier, such a solution may not necessarily correspond the the global maximum likelihood solution. Therefore it is better to explicitly optimize the log-likelihood of the data. The number of optimization parameter is greatly reduced by eliminating the mean and the variance. This is done by replacing them with their respective maximum likelihood estimates given θ as in equations (26),(35) and (49). This section explains how the evaluation of the log-likelihood function can be further simplified in order to reduce the expense of numerical optimization.

Proposition 1 *Let F be any full rank n by n matrix. let t be any n by p rank p matrix ($p < n$). Then* Trace $(t(t^T F t)^{-1} t^T F) = p$

Proof: Let $A = t(t^T F t)^{-1} t^T F$. $t^T F t$ is rank p, and hence A is rank p. Therefore $n - p$ eigenvalues of A are zero. Now $At = t$. Since t is rank p, this means that the remaining p eigenvalues of A are exactly 1. Since trace of a matrix is equal to the sum of the eigenvalues, hence the claim. Proof:

Proposition 2 *Let F be any full rank n by n matrix. let t be any n by p matrix with nonzero columns ($p < n$). Then* Trace $(t(\mathrm{Diag}(t^T F t))^{-1} t^T F) = p$

Proof: Let $A =$ Trace $(t(\mathrm{Diag}(t^T F t))^{-1} t^T F) = p$. Let us denote columns of t by t_i ($t = [t_1 \dots t_p]$). Then

$$A = \mathrm{Trace}\left(\sum_{i=1}^{p} t_i (t_i^T F t_i)^{-1} t_i^T F\right) \quad (54)$$

$$= \sum_{i=1}^{p} \mathrm{Trace}\left(t_i (t_i^T F t_i)^{-1} t_i^T F\right) \quad (55)$$

Now from preposition 1, it follows that Trace $t_i (t_i^T F t_i)^{-1} t_i^T F = 1$. Hence preposition 2 follows.

Proposition 3

$$\log P_{U,\theta}(\{x_i\}) = -\frac{N}{2} \log |(\theta_{n-p}^T \bar{T} \theta_{n-p})| - \sum_{j=1}^{J} \frac{N_j}{2} \log |(\theta_p^T \bar{W}_j \theta_p)|$$

$$+ N \log |\theta| - \frac{Nn}{2}(1 + \log 2\pi) \quad (56)$$

Proof: It can be easily verified that

$$\begin{aligned}
& \frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X}_j)^T\theta_p(\theta_p^T W_j\theta_p)^{-1}\theta_p^T(x_i-\bar{X}_j)) \\
= & \frac{1}{2}\sum_{j=1}^{J}\text{Trace }(\theta_p(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T W_j) \\
= & \sum_{j=1}^{J}\frac{N_j}{2}\text{Trace }(\theta_p(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T\bar{W}_j) \qquad (57)
\end{aligned}$$

which from preposition 1 is equal to

$$\sum_{j=1}^{J}\frac{N_j p}{2} = \frac{Np}{2} \qquad (58)$$

Similarly it follows that

$$\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X})^T\theta_{n-p}(\theta_{n-p}^T\bar{T}\theta_{n-p})^{-1}\theta_{n-p}^T(x_i-\bar{X}_j)) = \frac{N(n-p)}{2} \qquad (59)$$

Hence the claim follows from (49).

Proposition 4

$$\begin{aligned}
\log P_{D,\theta}(\{x_i\}) = & \frac{N}{2}\log|\text{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p})| - \sum_{j=1}^{J}\frac{N_j}{2}\log|\text{Diag}(\theta_p^T\bar{W}_j\theta_p)| \\
& -\frac{Nn}{2}(1+\log 2\pi) + N\log|\theta| \qquad (60)
\end{aligned}$$

Proof: Using preposition 2 it follows that:

$$\begin{aligned}
& \frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X}_j)^T\theta_p\text{Diag}(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T(x_i-\bar{X}_j)) \\
= & \frac{1}{2}\sum_{j=1}^{J}\text{Trace }(\theta_p\text{Diag}(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T W_j) \qquad (61) \\
= & \frac{Np}{2} \qquad (62)
\end{aligned}$$

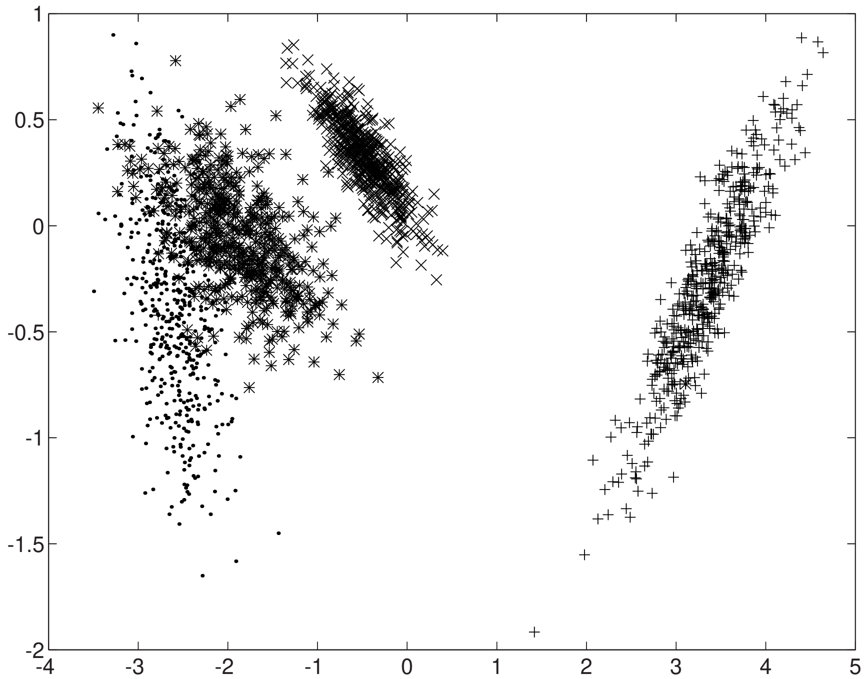

Figure 3: Optimal Projection of a five dimensional data with four classes to a two dimensional subspace. Variances are assumed to be equal.

Similarly

$$\begin{aligned} & -\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X})^T\theta_{n-p}\mathrm{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p})^{-1}\theta_{n-p}^T(x_i-\bar{X})) \\ = & \frac{N(n-p)}{2} \end{aligned} \tag{63}$$

Hence the claim follows from (26).

Proposition 5

$$\begin{aligned} \log P_{E,\theta}(\{x_i\}) = & \frac{N}{2}\log|\mathrm{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p})| + \frac{N}{2}\log|\mathrm{Diag}(\theta_p^T\bar{W}\theta_p)| \\ & -\frac{Nn}{2}(1+\log 2\pi) - N\log|\theta| \end{aligned} \tag{64}$$

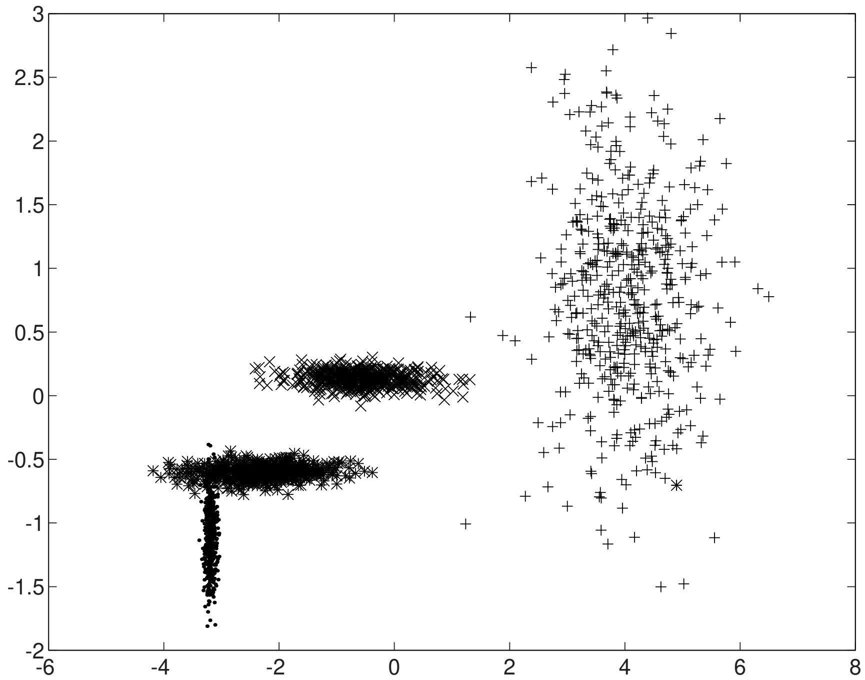

Figure 4: Optimal Projection of a five dimensional data with four classes to a two dimensional subspace. Variances are assumed to be diagonal and unequal.

Proof: The proof is very similar to the ones above and hence omitted.

The expressions in prepositions 3, 4 or 5 can be maximized to find the best projections for the case of interest. In a practical implementation, the entire likelihood function is normalized by N. For numerical stability norm of θ is also constrained.

To demonstrate the generalizations, five dimensional random data has been generated for four classes with Gaussian distribution for each class. The means vectors and the variance matrices for each of the classes is also randomly generated using a random number generator. The results of implementation are shown in figures 3, 4 and 5. The optimization was performed using the standard MATLAB optimization toolbox. Since the optimization has to be performed iteratively, it is useful to first find the linear discriminants, and use those as the inital guess. The analytical derivatives are supplied explicitly.

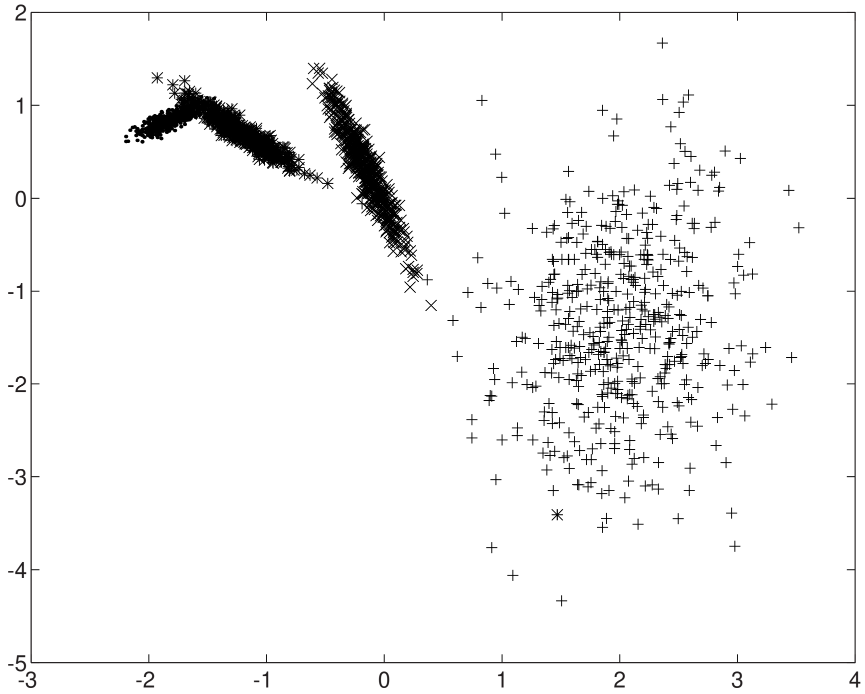

Figure 5: Optimal Projection of a five dimensional data with four classes to a two dimensional subspace. Variances are assumed to be unequal.

Some interesting observations can be made from these projections. Note that two of the four classes significantly overlap in the case where variances are assumed to be equal 3. In the projections shown in figures 4 and 5, this overlap has significantly reduced. Also note that in the case of LDA (figure 3), the within class variances do not appear to be too much different. Hence one may be tempted to incorrectly conclude that assuming equal variances is not a very bad assumption. However such a thought ma not arise in the mind of a person looking at figure 5. When the variance along the discriminating projections are constrained to be diagonal (figure 4) a projection is indeed found for which the major and the minor axis of the within class distributions are almost parallel to the horizontal and the vertical axis. Finding a linear transformation that would make all the within-group co-variance matrices diagonal is in general is impossible when the number of groups is more than two. However since the columns of θ are not constrained to be orthogonal, it has been possible to find a transformation for which the within-group co-variance matrices are as close to diagonal as possible.

5 On choice of the model class and order

Under certain circumstances there may be compelling reasons to believe that the underlying probability densities are Gaussian. However more often than not, the assumption is made due to the lack of any other better assumption. However the system designer well understands that Gaussian distributions span only a small subset of the set of all probability density functions. If the true probability measure does not belong to the class of Gaussian distributions, then even with infinite amount of training data, we may not hope to achieve the minimum probability of error that one would achieve if the true underlying probability measure were known. The reason is that we are always biased towards the assumption of a Gaussian distribution, while the true distribution is not Gaussian. Hence our classifier is only approximate.

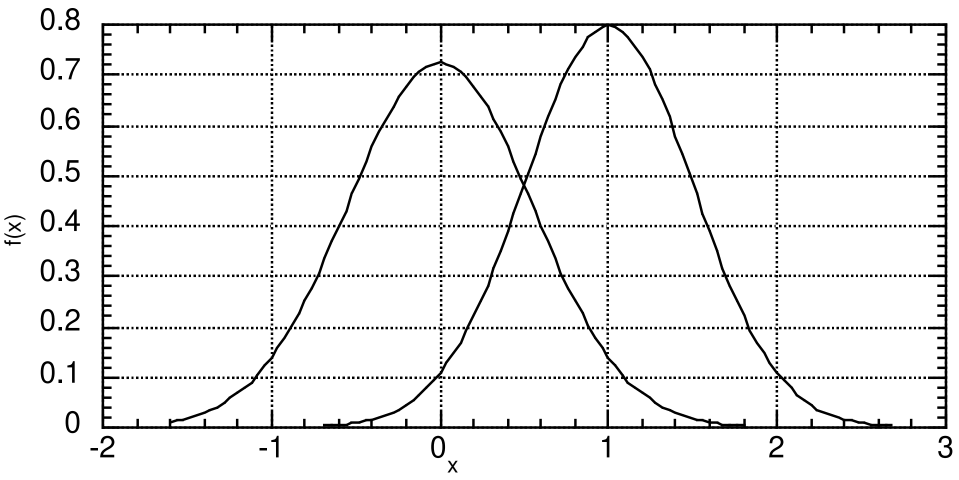

Figure 6: Two group classification problem: different means and only slightly different variances

Now consider the other extreme. Suppose the true distribution is Gaussian. Only the parameters are unknown. Now if we had infinite amount of training data, then we could get the correct parameters with probability one. However the amount of training data is finite. Hence our estimates have a variance about the true estimate.

Now suppose we consider an even more restricted class of probability measures, for example in which variance of all the all the classes are the same, (although the means are different). This class of distributions is a proper subset of the case considered in the previous paragraph. It can easily be shown that the maximum likelihood estimators for the means and the variance for this class are given as

$$\mu_j = \bar{X}_j \quad j = 1 \dots J \tag{65}$$

$$\Sigma = \frac{W}{N} \tag{66}$$

This new estimate of Σ depends on more data, and hence has lesser variance compared to the original individual estimates of variances. However since we are now constrained to a smaller class of probability distributions the bias in the distribution function is larger. If the advantage gained in terms of reduced variance outweighs the disadvantage due to increased bias, on an average, we would expect the recognition performance to improve.

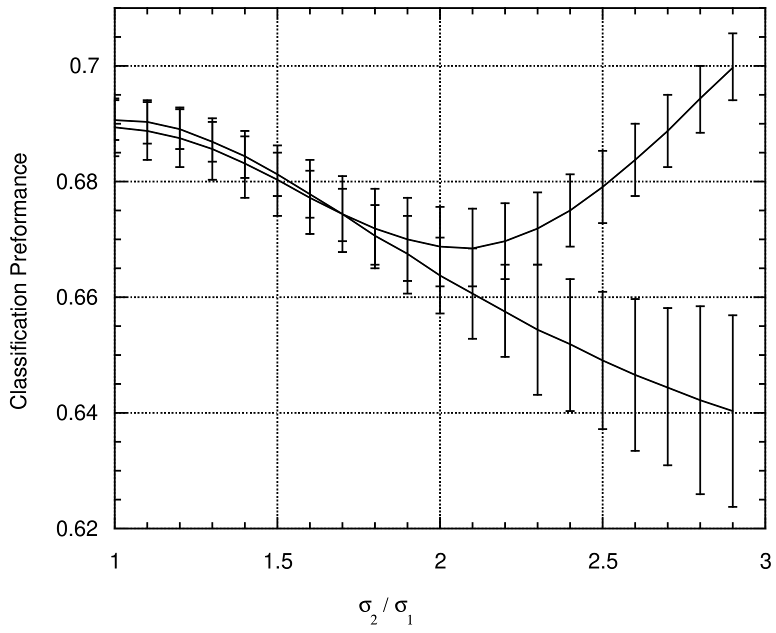

Figure 7: Classification performance for the two models versus ratio of the actual variance of the classes

For example consider figure 6. The problem depicted is to classify between two classes that have different means, and slightly different variances. In this example we would assume that only 50 training samples are available for each group. Then one could ask the question that how different should the two within-group variances (σ_1 and σ_2) should be so that it is still reasonable to assume the equal variance model. Figure 7 plots the classification performance of the models trained on 50 training samples versus the ratio of the two variances. The vertical bars denote the variance in the classification performance. To minimize the testing data influences in this monte-carlo simulation, each experiment used ten thousand test data points. The whole experiment was repeated ten thousand times to estimate the mean classification performance and the variance in the classification performance. The two curves correspond to two possible models - first in which variances are assumed to be equal, and second in which variances are assumed to be unequal.

Two important observations can be made from this plot. First that the average classification performance is better for the model that assumes equal variances as compared to the one that does not, even when the actual variances may be more than one-and-a-half times different. Of course if more training data is available, then the estimates would be better, and equal variance assumption may no longer be the best choice. Second, that if one were to make a rule for selecting a model, such a rule would be good only in a probabilistic sense, i.e. only the expected classification performance may be improved.

Similar concerns govern the choice of p - the dimensionality of the subspace which is assumed to have the discrimination information. In this case a smaller value of p covers a class probability measures that are a subset of those covered by a larger value of p. Hence a maximum likelihood criterion to choose between different p would invariably lead to the largest possible value of p, which is obviously incorrect. The problem of estimating the dimension of a model has been studied, and extensions of the maximum likelihood principle have been proposed that correspond to constructing a modified likelihood function that includes a term depending on the number of *free* parameters in the model. One such method is suggested by Akaike [7] for the slightly more general problem of choosing among different types of models with different number of parameters. His suggestion amounts to maximizing the likelihood function separately for each model γ, obtaining the likelihood $L_j(x_1, \ldots, x_n)$, and then choosing the model for which $\log L_j(x_1, \ldots, x_n) - k_j$ is largest, where k_j is the number of parameters in the model. Rissanen [8] has suggested an alternate criterion (MDL) where $\log L_j(x_1, \ldots, x_n) - k_j/2 \log n$ is maximized, where n is the sample size. It has been argued that Akaike criterion does not have good asymptotic properties [9].

When all the variance matrices are assumed to be equal, the free parameters are μ_j^p (total of Jp for all classes), $\mu_0^{(n-p)}$ (total $n-p$). Since the variance matrix is symmetric, it has only $n(n+1)/2$ free parameters. Alternately one could use the linear transformation matrix θ to diagonalize the variance matrix. Then one would associate n free parameters with the diagonal variance matrix, and the remaining $n(n-1)/2$ free parameters with the linear transformation θ. Thus the total number of free parameters is $Jp+(n-p)\ +\ n(n+1)/2$. Similarly when the variance matrices in the discriminating subspace are assumed to be diagonal, the number of free parameters can be written as $Jp+(n-p)\ +\ Jp+(n-p)+n(n-1)/2 = 2*(Jp+(n-p))+n(n-1)/2$, and when the discriminating variance matrices are assumes to full, the number of free parameters is $Jp + (n-p) + Jp(p+1)/2 + (n-p) + n(n-1)/2)$. These values for the number of free parameters can be used with the modified likelihood criterion to choose between the three class of models, and the order p.

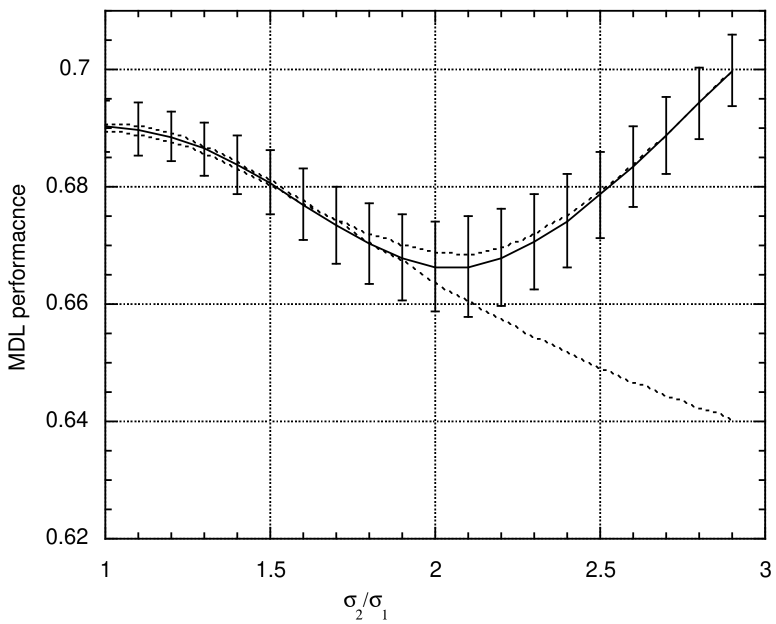

Figure 8: Classification performance for the model selected by MDL criterion versus ratio of the actual variance of the classes

MDL criterion has been applied to the example of two classes described above. The performance is measured for the model selected by using the MDL criterion, and the corresponding parameters for that model. Figure 8 displays the performance obtained using this criterion. The error bars display the variance in performance for repeated experiments. The dotted lines are the performance curves for the two models, and the dark line indicates the performance when the model is assumed to be unknown and MLD criterion is used choose between the two models. As can be seen, the average performance is close to the best possible performance if the correct model choice was known a-priori.

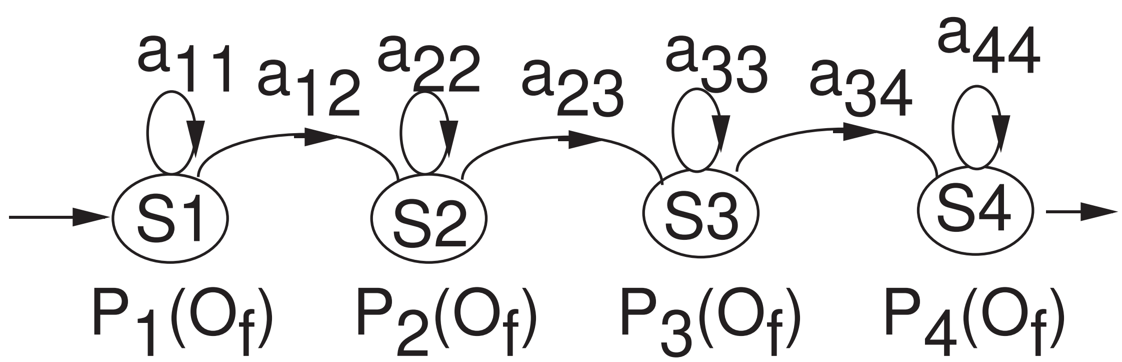

Figure 9: Example of a hidden Markov model

6 Extension to Hidden Markov Models

The maximum likelihood framework of feature dimension reduction can be generalized to hidden Markov models (HMM) [10]. Such a formulation would be useful for a data $o_1, \ldots, o_t$ which is time-series of vectors. To distinguish from the iid observations x we will denote the time-series by O, $O = o_1, \ldots, o_T$, where T is the index of final time-step. Initial state is assumed to be 1, and final state assumed to be N. Fig. 9 shows an example of a HMM. It consists of a collection of states $\{S_j\}$. At a discrete time step, the model can make a transition from state S_i to state S_j based on the transition probability a_{ij}. When any state S_j is entered, a n-dimensional feature vector is emitted, based on an underlying distribution $P_j(O)$ which can be assumed as Gaussian *with the assumption that there is a $n-p$ dimensional linear subspace in which $P_j(O)$ are exactly the same for all the states S_j*. The observations are now sequences of vectors, and each sequence can be assumed to be independent of the other. Without loss of generality, we will assume that the initial state, and the final state are fixed and known. Then such a HMM is completely specified by the state transition probability matrix $A = [a_{ij}]$, and the state probability distributions $\{P_j(O)\}$ which are in turn specified by the set $\{\theta, \mu_0, \mu_j^p, \Sigma_j^p, \Sigma^{(n-p)}\}$. To find a maximum-likelihood estimate of the HMM parameters, can be found by embedding the optimization for θ in the maximization step of the E-M algorithm [11, 12]. In particular, for the models discussed in section 4, the parameters to be estimated are $\{\theta, \mu_j, \Sigma_j\}$. Now there are additional parameters to be estimated, namely $a_{i,j}$ the transition probabilities from state i to state j. If the state transition sequence were known for the observed data, then it would be very easy to estimate the transition probabilities. Since the state transition sequence is not known, it is termed as "missing data". The expanded set of parameters is estimated in an iterative manner. An initial guess of the model parameters is assumed. The expected value of probabilities of being in a particular state at each time step (missing data) are estimated on the basis of the training data, and assuming that $\{a_{i,j}, \theta, \mu_j, \Sigma_j\}$ are known. This is the expectation step. Then $\{a_{i,j}.\theta, \mu_j, \Sigma_j\}$ are re-estimated so as to maximize the likelihood of the training data plus the transition probabilities. This is the maximization step.

The expected value of the transitions at any time is calculated using the forward-backward algorithm [11]. Let the forward probability $\alpha_j(t)$ for some model M be defined as

$$\alpha_j(t) = P(o_1, \ldots, o_t, s(t) = j|M) \tag{67}$$

That is, the joint probability of observing the first t speech vectors and being in the state j at time t. This forward probability is efficiently calculated by the recursion

$$\alpha_j(t) = \left[\sum_{i=all\ \ states} \alpha_i(t-1)a_{ij}\right] b_j(o_t) \tag{68}$$

The backward probability is defined as

$$\beta_i(t) = P(o_{t+1}, \ldots, o_T|s(t) = j, M) \tag{69}$$

This backward probability can again be efficiently computed by the backward recursion

$$\beta_i(t) = \sum_{j=all\ \ states} a_{ij} b_j(o_{t+1}) \beta_j(t+1) \tag{70}$$

Then based on the forward and backward probabilities,

$$\alpha_j(t)\beta_j(t) = P(O, s(t) = j|M) \tag{71}$$

Now the probability of being in state i at time t and state j at time $t+1$ can be written as

$$\begin{aligned}
P(s(t)=i,s(t+1)=j,|O,M) &= P(s(t)=i,s(t+1)=j,O|M)/P(O|M) && (72)\\
&= \alpha_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)/P(O|M) && (73)\\
&= \frac{\sum_{t=1}^{T}\alpha_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)}{\sum_{t=1}^{T}\alpha_i(t)\beta_i(t)} && (74)
\end{aligned}$$

And the likelihood of being in state j at time t is

$$\begin{aligned}
L_j(t) &= P(s(t)=j|O,M)\\
&= \frac{P(O,s(t)=j|M)}{P(O|M)}\\
&= \frac{\alpha_j(t)\beta_j(t)}{P(O|M)} && (75)
\end{aligned}$$

The normalizing constant $P(O|M)$ satisfies the condition $\sum_{all\ states} L_j(t)=1$ and it can be shown that $P(O|M)=\alpha_N(T)$.

Once the likelihood of being in any state is known, the weighted means and sum of square products are calculated as

$$N_j = \sum_{t=1} T L_j(t) \tag{76}$$

$$N = \sum_{all\ states} N_j = T \tag{77}$$

$$\bar{N}_j = \frac{N_j}{N} \tag{78}$$

$$\bar{X}_j = \frac{\sum_{t=1}^{T} o_t}{N} \tag{79}$$

$$\bar{X} = \frac{\sum_{t=1}^{T} L_j(t)o_t}{N_j} \tag{80}$$

$$\bar{W}_j = \sum_{t=1}^{T}\frac{L_j(t)}{N_j}(o_t-\bar{X}_j)(o_t-\bar{X}_j)^T \tag{81}$$

$$\bar{W} = \sum_{all\ states}\bar{N}_j\bar{W}_j \tag{82}$$

$$\bar{T} = \sum_{t=1}^{T}\frac{1}{N}(o_t-\bar{X})(o_t-\bar{X})^T \tag{83}$$

The above statistic can then be used to perform the constrained optimization of section 4 to find $\theta, \mu_0, \Sigma^{(n-p)}$ and $\{\mu_j^p, \Sigma_j^p\} \forall S_j$. This procedure is repeated until convergence.

It is straightforward to prove that this algorithm is indeed EM is straightforward. Along the formulation in [11] we first note that calculating $L_j(t)$ is indeed the expectation step. Second step is to note that if the probability of being in a particular state at any time $L_j(t)$ is known then the likelihood of the data is indeed maximized by estimating the parameters μ_j and Σ_j in the manner described above.

Note that mixture models are a special case of HMM's, that contain non-emitting initial and final states, and as many additional states (K) as there are mixtures in the model. Transition probability from the initial state to a "mixture" state is a_{1j} ($\sum_{j=1}^{K} a_{1j} = 1$). Each of the mixture states makes a non-emitting transition into the final state. Due to this analogy, the above generalization holds for mixture models as well.

Appendix A

Here we list some matrix differentiation results. Consider this simple example. Suppose

$$y = X^T A X \tag{84}$$

here X is a $n \times 1$ vector, and A is a $n \times n$ matrix. Suppose we wish to differentiate y with respect to X. It means that we want to evaluate the partial derivative of y with to each element of X. Then instead of enumerating each of the derivatives individually, it's more convenient to define a matrix notation where

$$\frac{\partial y}{\partial X} \equiv \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_N} \end{bmatrix} \tag{85}$$

Then by performing a term by term differentiation it can be shown that

$$\frac{\partial}{\partial X} X^T A X = (A + A^T) X \tag{86}$$

This is true even when X is an arbitrary matrix. Similarly $\frac{\partial y}{\partial A}$ would be a $n \times n$ matrix in which the ij'th entry would be $\frac{\partial y}{\partial a_{ij}}$.

$$\frac{\partial}{\partial A} X^T A X = X X^T \tag{87}$$

Now that the basic technique has been understood, I would state without proof, a list of matrix differentiation results. These results can also be found in references like [13].

$$\frac{\partial}{\partial A} x^T C^T A C y = A C y x^T + A^T C x y^T \tag{88}$$

$$\frac{\partial}{\partial A} X^T A^{-1} X = -A^{-1} X X^T A^{-1}$$

$$\frac{\partial}{\partial A} \mathrm{tr} A = I \tag{89}$$

$$\frac{\partial}{\partial A} |A| = |A| (A^T)^{-1} \tag{90}$$

$$\frac{\partial}{\partial A} \log |A| = (A^T)^{-1} \tag{91}$$

$$\frac{\partial}{\partial A} \log |A^{-1}| = -(A^T)^{-1} \tag{92}$$

Acknowledgments

This research was supported by the Center for Language and Speech Processing at Johns Hopkins University, Martin Marietta corporation, and National Security Agency (Grant number: G599-E96-2001). one of the authors (NK) would like to thank Prof. Carey Priebe for his interest and suggestions. References to this work should be either to the published version, if any, or with apriori permission from the authors.

References

[1] R. O. Duda and P. B. Hart, *Pattern Classification and Scene Analysis.* John Wiley & Sons, Inc., 1973.

[2] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[3] L. Kanal and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recognition*, pp. 225–234, 1971.

[4] A. M. Kshirsagar, *Multivariate Analysis.* Marcel Dekker, Inc., 1972.

[5] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *American Statistical Association Journal*, pp. 1159–1178, 1967.

[6] N. Campbell, "Canonical variate analysis - a general formulation," *Australian Journal of Statistics*, vol. 26, pp. 86–96, 1984.

[7] H. Akaike, "A new look at the statistical identification model," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[8] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, vol. 15 of *Series in Computer Science.* World Scientific Publishing Co. Pte Ltd., 1989.

[9] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[10] L. R. Rabiner and B. H. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, pp. 4–16, January 1986.

[11] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164–171, 1970.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via em algorithm," *Journal of the Royal Statistical Society*, vol. xx, pp. 1–38, Jan. 1977.

[13] J. Bibby and H. Toutenburg, *Prediction and Improved Estimation in Linear Models.* John Wiley & Sons, 1977.