A GENERALIZATION OF LINEAR DISCRIMINANT ANALYSIS IN MAXIMUM LIKELIHOOD FRAMEWORK

Nagendra Kumar,Andreas G. Andreou, CLSP, Johns Hopkins University
Nagendra Kumar, JHU, 105 Barton Hall, 3400 N. Charles St., Baltimore, MD, 21218

Keywords— **Fisher-Rao, Linear Discriminant Analysis, Clustering, Dimension Reduction, Classification, Heteroscedastic**

Abstract— **The Fisher–Rao linear discriminant analysis (LDA) is a valuable tool for multi-class classification and data reduction. We investigate LDA within the maximum likelihood framework and propose a general formulation to handle heteroscedasticity. Small size numerical experiments with randomly generated data verify the validity of our formulation.**

I. Introduction

Linear discriminant analysis (LDA) is a mathematical tool widely used for dimensionality reduction and multi-class classification [1], [2], [3]. Our interest in LDA [4] stems from our desire to use auditory features in speech recognition and from the encouraging results obtained by Brown [5]. However, inconsistent modeling assumptions between LDA, and the models used for recognition yield final systems with non-optimum performance.

LDA can be derived as a maximum likelihood method for normal populations with different means, and common co-variance matrices [6]. Hastie [7] has further generalized this approach to the case where class distributions are Gaussian mixture models. However the constraint of common co-variance matrices is still maintained.

In this work we present a generalization where the constraint of equality of the co-variance matrices is relaxed [8]. Surprisingly, despite the past observations of relationship between maximum likelihood models and LDA [6], to the best of our knowledge, the proposed generalization of LDA to handle heteroscedasticity has not been reported in the literature.

II. Dimension Reduction Through Linear Projections

The problem of dimensionality reduction through linter projections can be described as follows: Let x be an n dimensional observation vector. We seek a linear transformation $\Re^n \rightarrow \Re^p$ $(p < n)$ of the form $y_p = \theta_p^T x$ where θ_p is an $n \times p$ matrix. Let θ be a non-singular linear transformation which is partitioned as

$$\theta = [\theta_p \theta_{n-p}] = \left[\vec{\theta_1} \ldots \vec{\theta_n}\right] \tag{1}$$

where θ_p are the first p columns of θ and θ_{n-p} corresponds to the remaining $n-p$ columns and $\vec{\theta_i}$ corresponds to the i'th column of θ. Then, observation vector dimension reduction can be viewed as a two step procedure. First a non-singular linear transformation is applied to x to obtain $y = \theta^T x$, and then in the second step only the first p rows of y are retained to give y_p.

Dimensionality reduction as described above, is useful in practical pattern classification applications, where the ultimate objective the design of a system that puts the vector of observations (features) in different classes on the basis of the observed data. LDA attempts to choose the linear transformation θ_p in such a way so as to retain the maximum amount of class discrimination information in the early stages of the system.

Let there be a total of J classes, and let $g(i) \rightarrow \{1 \ldots J\}$ indicate the class that is associated with x_i. Let $\{x_i\}$ be the set of training examples available. Then $\sum_{g(i)=j} 1 = N_j$ associated with class j, and $\sum_{j=1}^{J} N_j = N$ is the total number of training examples. If $\bar{X}$ is the sample mean where:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2}$$

The total sample variance $\bar{T}$ is defined as

$$\bar{T} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})(x_i - \bar{X})^T \tag{3}$$

The class means $\bar{X}_j$, the class variances $\bar{W}_j$ and the overall pooled variance $\bar{W}$ are defined as:

$$\bar{X}_j = \frac{1}{N_j} \sum_{g(i)=j} x_i \tag{4}$$

$$\bar{W}_j = \frac{1}{N_j} \sum_{g(i)=j} (x_i - \bar{X}_j)(x_i - \bar{X}_j)^T \quad j = 1 \ldots J \tag{5}$$

$$\bar{W} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{W}_j \tag{6}$$

LDA maximizes the ratio of the overall variance to the within class variance [3], [9].

$$\hat{\vec{\theta}}_1 = \text{argmax}_{\vec{\theta}_1} \frac{\vec{\theta}_1^T \bar{T} \vec{\theta}_1}{\vec{\theta}_1^T \bar{W} \vec{\theta}_1} \tag{7}$$

If before LDA, x undergoes a linear transformation, any such full rank linear transformation will appear both in the numerator and the denominator of this ratio (as a multiplier of $\bar{T}$ and $\bar{W}$), and thus divide out. Hence, linear discriminants are invariant to any full-rank linear transformation of the input [9]. It can be shown that the solution to the above equation corresponds to the right eigenvector of $\bar{W}^{-1}\bar{T}$ that has the largest eigenvalue [3]. 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 the eigenvectors, a p dimensional uncorrelated observations are obtained.

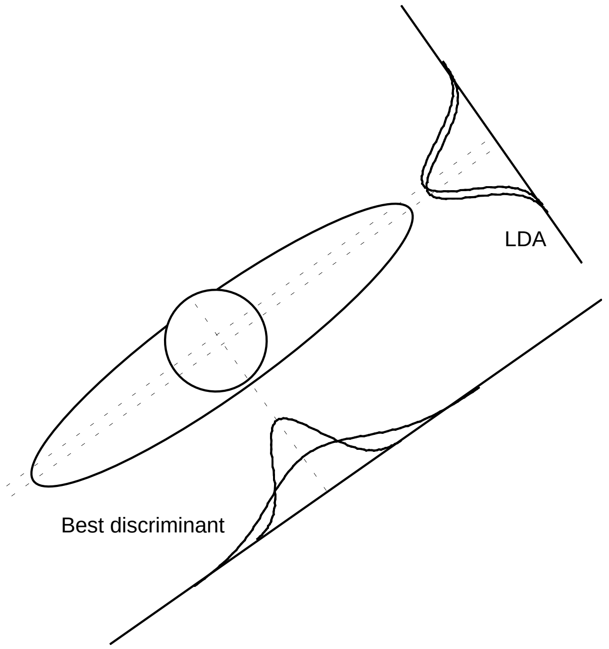

Fig. 1. An illustration of how LDA fails in the case of two Gaussian classes

However, LDA may fail when the within class distributions are heteroscedastic. This is schematically depicted in figure 1. The two classes have *almost* the same mean, but the variances are very different in one direction. In this case a classifier would perform better if a projection is taken along the direction in which the variances are different, and un-equal variance models can be used in the classifier design. Since the LDA method pools the within class variances, it would choose the projection marked as LDA in the figure, which is clearly not the best.

In this paper, we provide a solution to this problem by elaborating on earlier work of Campbell [6] who first noted the relationship between an LDA transformation and the estimation of maximum likelihood parameters of a Gaussian model with some a-priori assumptions on the structure of the models. The first assumption was that all the class discrimination information resides in a p dimensional sub-space of the n dimensional observation space, while the second assumption constraints the within class variances to be equal for all the classes. Therefore, LDA projections are best suited for a Gaussian-model based classifier which assumes that the observation vector was generated from that, while its performance is not optimal when the the class distributions are heteroscedastic. The generalization the handle heteroscedasticity is obtained by dropping the equal variance assumption in the model. There is no closed form solution for θ and numerical optimization techniques have to be used to find the optimal projections. But this is not much different than LDA, where again numerical methods are required to find the eigen-vectors. The objective function is simplified to speed up the numerical optimization.

III. Generalizations of LDA

As in the previous section, let θ be a non-singular linear transformation that transforms the data variables x into new variables y. For dimension reduction we will assume a model that only the first p components of y carry any class discrimination information. This is equivalent to assuming that the class means lie in a p-dimensional subspace, and the remaining $n-p$ dimensional subspace is homogeneous with respect to class means and variances. Also, the full rank linear transformation θ is such that the first p columns of θ span the p-dimensional sub-space in which the class means and probably the class variances are different. 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 as follows

$$\mu_j = \begin{bmatrix} \mu_{j,1} \\ \vdots \\ \mu_{j,p} \\ \mu_{0,p+1} \\ \vdots \\ \mu_{0,n} \end{bmatrix} = \begin{bmatrix} \mu_j^p \\ \mu_0 \end{bmatrix} \tag{8}$$

$$\Sigma_j = \begin{bmatrix} \Sigma^p_{j(p \times p)} & 0 \\ 0 & \Sigma^{(n-p)}_{(n-p \times n-p)} \end{bmatrix} \tag{9}$$

where μ_0 is the common term in all the means, and μ_j^p are different for each class. Σ_j have also been partitioned in the corresponding manner, such that $\Sigma^{(n-p)}$ is common for all classes, while Σ_j^p could be different for different classes.

The density function of a data point under the model above is given as

$$P(x_i) = \frac{|\theta|}{\sqrt{(2\pi)^n |\Sigma_{g(i)}|}} e^{\frac{(y_i - \mu_{g(i)})^T \Sigma_{g(i)} (y_i - \mu_{g(i)})}{2}} \tag{10}$$

where $y = \theta^T x$. The log-likelihood L_F of the data under the linear transformation and constrained Gaussian model assumption for each class is

$$\begin{aligned} \log \ & L_F(\mu_j, \Sigma_j, \theta | x_i) \\ &= -\frac{1}{2} \sum_{i=1}^{N} \{ (\theta^T x_i - \mu_{g(i)})^T \Sigma_{g(i)}^{-1} (\theta^T x_i - \mu_{g(i)}) \\ &\quad + \log((2\pi)^n |\Sigma_{g(i)}|) \} + log|\theta| \end{aligned} \tag{11}$$

The subscript F is to remind us that Σ_j^P are different and full co-variance matrices. The above likelihood function can now be maximized with respect to various parameters. A straight-forward maximization with respect to various parameters can be a time consuming task. However the task can be considerably simplified by first calculating the optimal values of the mean and variance parameters in terms of the linear transformation θ. Differentiating the likelihood equation with respect to the parameters μ_j and Σ_j gives the mean and variance estimates as

$$\hat{\mu}_j^p = \theta_p^T \bar{X}_j \tag{12}$$

$$\hat{\mu}_0 = \theta_{n-p}^T \bar{X} \tag{13}$$

$$\hat{\Sigma}_j^p = \frac{1}{N_j} (\theta_p^T W_j \theta_p) \quad j = 1 \ldots J \tag{14}$$

$$\hat{\Sigma}^{n-p} = \frac{1}{N} \theta_{n-p}^T T \theta_{n-p} \tag{15}$$

Note that the $\mu_j, \ j = 1 \ldots J$ can be calculated if θ is known, and $\sigma_j, \ j = 1 \ldots J$ if $\mu_j, \ j = 1 \ldots J$ and θ are known. Therefore, we would first like to solve for θ. Substituting the values of the optimized μs and σs in (11) gives the likelihood of the data in terms of θ which can be simplified and then maximized with respect to θ to give

$$\begin{aligned} \hat{\theta}_F &= \text{argmax}_\theta \{ -\frac{N}{2} \log |(\theta_{n-p}^T \bar{T} \theta_{n-p})| \\ &\quad - \sum_{j=1}^{J} \frac{N_j}{2} \log |(\theta_p^T \bar{W}_j \theta_p)| + N \log |\theta| \} \end{aligned} \tag{16}$$

where $\hat{\theta}$ is the estimate of the parameter θ. At this point one may choose to use only the first p columns of $\hat{\theta}$ to obtain the best discriminating projection under the Gaussian model assumption.

A. Σ constrained to diagonal

Due to computational simplicity, within class variances are often assumed to be diagonal. This is especially true in the case of speech recognition where the number of models is so enormous that invariably diagonal variance matrices are assumed. Therefore the optimal projections for this case have also been considered. Suppose we further assume that Σ_j^p and $\Sigma^{(n-p)}$ are diagonal matrices such that $\Sigma_j = \text{Diag}(\sigma_j^1 \ldots \sigma_j^p \ \sigma^{p+1} \ldots \sigma^n)$. Then in terms of the matrix partitions above the log-likelihood of the data can be written as

$$\begin{aligned} &\log \ L_D(\mu_j, \Sigma_j, \theta | \{x_i\}) = \\ &\frac{-Nn}{2} \log 2\pi + N \log |\theta| - \frac{N}{2} \sum_{k=p+1}^{n} \log |\sigma^k| \\ &- \sum_{j=1}^{J} \frac{N_j}{2} \sum_{k=1}^{p} \log |\sigma_j^k| - \frac{1}{2} \sum_{j=1}^{J} \sum_{g(i)=j} \sum_{k=1}^{p} \frac{(\vec{\theta}_k^T x_i - \mu_{j,k})^2}{\sigma_j^k} \\ &+ \frac{1}{2} \sum_{j=1}^{J} \sum_{g(i)=j} \sum_{k=p+1}^{n} \frac{(\vec{\theta}_k^T x_i - \mu_{0,k})^2}{\sigma^k} \end{aligned} \tag{17}$$

Using the same method as before, maximizing the likelihood with respect to $\mu_j, \ \Sigma_j \ j = 1 \ldots J$, we get

$$\hat{\mu}_j^p = \theta_p^T \bar{X}_j \tag{18}$$

$$\hat{\mu}_0 = \theta_{n-p}^T \bar{X} \tag{19}$$

$$\hat{\Sigma}_j^p = \text{Diag}(\theta_p^T \bar{W}_j \theta_p) \quad j = 1 \ldots J \tag{20}$$

$$\hat{\Sigma}^{n-p} = \text{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p}) \tag{21}$$

Substituting values of all the above maximized parameters except θ in (17) gives the likelihood of the data in terms of θ. The maximum likelihood estimate for θ can now be found by maximizing the likelihood numerically. It can be shown that this maximization can be simplified to the following

$$\begin{aligned} \hat{\theta}_D &= \text{argmax}_\theta \{ -\frac{N}{2} \log |\text{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p})| \\ &\quad - \sum_{j=1}^{J} \frac{N_j}{2} \log |\text{Diag}(\theta_p^T \bar{W}_j \theta_p)| + N \log |\theta| \} \end{aligned} \tag{22}$$

Where $\hat{\theta}_D$ is the estimator for θ the optimal transformation if one wishes to diagonal variance Gaussian models to model each class.

B. Σ's constrained to be equal

If we additionally require that $\Sigma_j = \Sigma, \; \forall j$, then the maximum likelihood parameter estimates can be written as follows

$$\begin{aligned} \hat{\mu}_j^p &= \theta_p^T \bar{X}_j & (23) \\ \hat{\mu}_0 &= \theta_{n-p}^T \bar{X} & (24) \\ \hat{\Sigma}^p &= \mathrm{Diag}(\theta_p^T \bar{W} \theta_p) \quad j = 1 \ldots J & (25) \\ \hat{\Sigma}^{n-p} &= \mathrm{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p}) & (26) \end{aligned}$$

and

$$\begin{aligned} \hat{\theta}_E &= \mathrm{argmax}_\theta \{ -\frac{N}{2} \log |\mathrm{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p})| \\ &\quad -\frac{N}{2} \log |\mathrm{Diag}(\theta_p^T \bar{W} \theta_p)| + N \log |\theta| \} \qquad (27) \end{aligned}$$

Here equation (27) is obtained by inserting the values of μ_j and Σ that maximize the likelihood, and then dropping the constant terms in the log-likelihood. It can be shown that the solution obtained by taking the eigenvectors corresponding to largest p eigenvalues of $\bar{W}^{-1}\bar{T}$ also maximizes the above expression, thus asserting the claim that LDA is the maximum likelihood parameter estimate of a constrained model.

IV. Computing $\hat{\theta}$

Optimal values for $\hat{\theta}_F$, $\hat{\theta}_D$ and $\hat{\theta}_E$ must be computed numerically. First it should be pointed out that the optimal solution for $\hat{\theta}$ is not unique. In fact it is easy to show that

Proposition 1: The maximum value of the log-likelihood (maximized with respect to μ_j and Σ_j) is independent of linear scaling of the columns of θ.

PROOF: W.L.O.G. we will consider the case of the log-likelihood under the equal variance assumption. It can be written as

$$\begin{aligned} \log L_E(\theta | \{x_i\}) &= \frac{-Nn}{2}(1 + \log 2\pi) - N \log |\theta| \\ &\quad + \frac{N}{2} \log |\mathrm{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p})| \\ &\quad + \frac{N}{2} \log |\mathrm{Diag}(\theta_p^T \bar{W} \theta_p)| \qquad (28) \end{aligned}$$

Now for any $n \times n$ matrices θ and A,

$$|\mathrm{Diag}(\theta_p^T A \theta_p)| = \prod_{i=1}^{p} \vec{\theta}_i^T A \vec{\theta}_i$$

Let us assume that i'th column of θ, $\vec{\theta}_i$ is scaled by a linear factor $\alpha > 0$ to give $\acute{\theta}$. W.L.O.G. we also assume that $i \le p$.

Then the new value of the likelihood can be written as

$$\begin{aligned} \log \;\; L_E(\acute{\theta} | \{x_i\}) &= \frac{-Nn}{2}(1 + \log 2\pi) - N \log(\alpha |\theta|) \\ &\quad + \frac{N}{2} \log(\alpha^2 |\mathrm{Diag}(\theta_p^T \bar{W} \theta_p)|) \\ &\quad + \frac{N}{2} \log |\mathrm{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p})| \\ &= \frac{-Nn}{2}(1 + \log 2\pi) + \frac{N}{2} \log |\mathrm{Diag}(\theta_{n-p}^T \bar{T} \theta_{n-p})| \\ &\quad + \frac{2N \log \alpha}{2} + \frac{N}{2} \log |\mathrm{Diag}(\theta_p^T \bar{W} \theta_p)| \\ &\quad - N \log |\theta| - N \log \alpha \\ &= \log L_E(\theta | \{x_i\}) \qquad (29) \end{aligned}$$

Thus proposition 1 follows. Similarly it can be shown that non-singular transformations of θ_p, or θ_{n-p} do not affect log-likelihood $L_F(.)$.

Therefore sometimes it is easier to optimize if the norm of columns of θ is explicitly constrained to some non-zero number, say 1. We have then been able to perform the required optimization using quadratic programming algorithms such as those available in MATLAB™ optimization toolbox. Analytic expressions for the derivative of the likelihood are explicitly provided for the optimization routines. Even though we use quadratic optimization techniques the likelihood surface is not necessarily quadratic and the optimization algorithms occasionally fail. It has however been possible to recover by slightly perturbing θ and re-starting the optimization.

V. Experiments

Experiments with randomly generated data have been performed to illustrate the method. Five dimensional data has been randomly generated for four classes with Gaussian distribution for each class. The means vectors and the variance matrices for each of the classes are also randomly generated. Then the dimension is reduced to two ($p = 2$) using the method described in the previous section. The optimization was performed using the standard MATLAB™ optimization tool-box. Since the optimization has to be performed iteratively, it is useful to first find the linear discriminants, and use those as the initial guess. The analytical derivatives are supplied explicitly. The results are shown in figures 2, 3 and 4.

Some interesting observations can be made from these projections. Note that two of the four classes (dots and stars) significantly overlap in the case where variances are assumed to be equal (Fig. 2). In

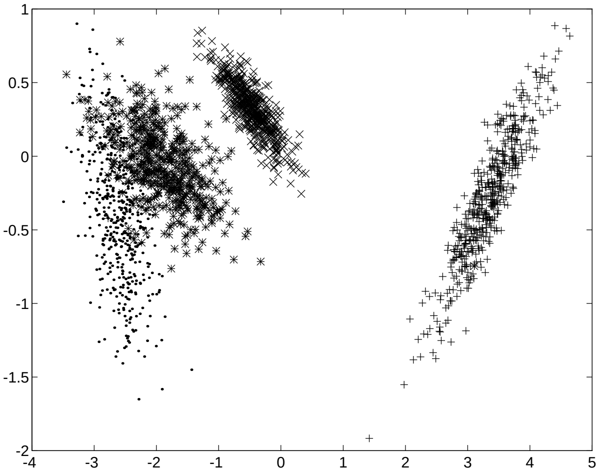

Fig. 2. Optimal Projection of a five dimensional data with four classes to a two dimensional subspace. Variances are assumed to be equal.

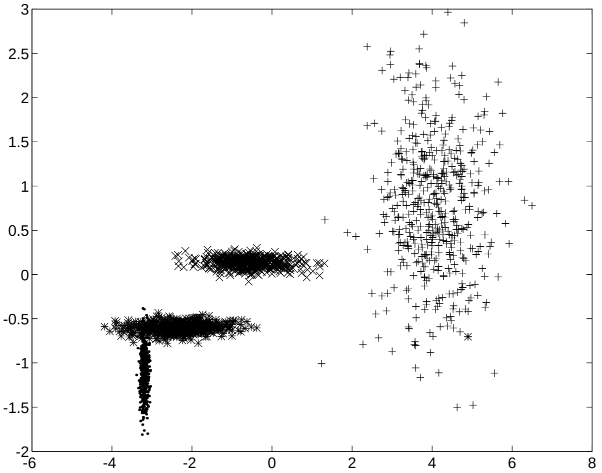

Fig. 3. Optimal Projection of a five dimensional data with four classes to a two dimensional subspace. Variances are assumed to be diagonal and unequal.

the projections shown in figures 3 and 4, this overlap has been significantly reduced. Also note that in the case of LDA (figure 2), 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 the difference in variance is clear from the data plotted in figure 4. When the variance along the discriminating projections are constrained to be diagonal (figure 3) 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-class co-variance matrices diagonal is in general 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-class co-variance matrices are as close to diagonal as possible.

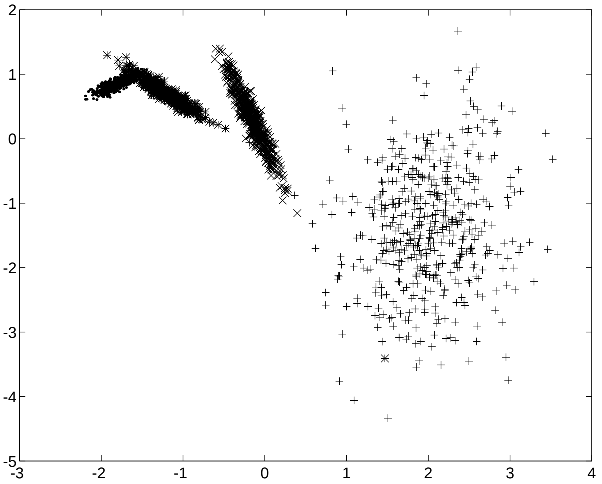

Fig. 4. Optimal Projection of a five dimensional data with four classes to a two dimensional subspace. Variances are assumed to be unequal.

VI. Discussion

Campbell [6] has proposed that

Proposition 2: If the columns of θ are restricted to the canonical variates, then the likelihood is maximized if the first p columns of θ correspond to the eigenvectors of the largest p eigenvalues of $W^{-1}T$

Although the proposition is true, one should note that the choice of restricting θ to the canonical variates, is arbitrary. Here, an alternate proof is presented and is followed by a stronger statement regarding the optimality of LDA.

PROOF: Suppose real matrix $\hat{\theta}$ is such that both $\hat{\theta}^T\bar{T}\hat{\theta}$ and $\hat{\theta}^T\bar{W}\hat{\theta}$ are diagonal matrices. Let the diagonal elements of $\hat{\theta}^T\bar{T}\hat{\theta}$ and $\hat{\theta}^T\bar{W}\hat{\theta}$ be written as $\{\alpha_1^T,\dots,\alpha_n^T\}$ and $\{\alpha_1^W,\dots,\alpha_n^W\}$ respectively. Such a $\hat{\theta}$ is called the generalized eigenvector of the matrices $\bar{T}$ and $\bar{W}$ and correspond to the well known Fisher-Rao LDA solution. This $\hat{\theta}$ also corresponds to the right eigenvectors of $\bar{W}^{-1}\bar{T}$. The eigenvalues are given as $\{\alpha_1^T/\alpha_1^W,\dots,\alpha_n^T/\alpha_1^W\}$. Assume that the columns of $\hat{\theta}$ are arranged in the order of descending magnitude of the eigenvalues of $\bar{W}^{-1}\bar{T}$.

In our usual notation let $\hat{\theta}_p$ denote the first p columns. Using proposition 1, we can assume w.l.o.g. that $|\hat{\theta}| = 1$. Then the log-likelihood of the data can be written as

$$\log L_E(\hat{\theta}|\{x_i\}) \quad = \quad \frac{-Nn}{2}(1+\log 2\pi) +$$

$$\frac{N}{2}\sum_{i=p+1}^{n}\log\alpha_i^T + \frac{N}{2}\sum_{i=1}^{p}\log\alpha_i^W \quad (30)$$

Now suppose we consider an alternate solution where the ith column of $\hat{\theta}$ $(1 \le i \le p)$ is swapped with jth column of $\hat{\theta}$ $(p < j \le n)$. Then due to the the fact that the columns were arranged in the order of descending eigenvalues, $\lambda_i^T\lambda_j^W \le \lambda_i^W\lambda_j^T$ and hence the alternate solution cannot give a better likelihood. Therefore $\hat{\theta}$ does indeed maximize the likelihood. Q.E.D.

Preposition 2 requires that the columns of $\hat{\theta}$ be the eigenvectors of $\bar{W}^{-1}\bar{T}$. However this is not a necessary constraint. In fact, a more general claim can be made.

Theorem 1: LDA transformation is equivalent to finding maximum-likelihood parameters of a Gaussian model which assumes that all the class discrimination information resides in a p dimensional subspace of the n dimensional feature space, and makes a further assumption that the within class variances are equal for all the classes (i.e. LDA solution is also solution of the equation (27))

PROOF: Differentiating (27) with respect to θ and equating the derivative to zero yields the following equations

$$\begin{aligned}
I_p &= \theta_p^T\bar{W}\theta_p(\mathrm{Diag}(\theta_p^T\bar{W}\theta_p))^{-1} && (31)\\
0 &= \theta_{n-p}^T\bar{W}\theta_p(\mathrm{Diag}(\theta_p^T\bar{W}\theta_p))^{-1} && (32)\\
I_{n-p} &= \theta_{n-p}\bar{T}\theta_{n-p}(\mathrm{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p}))^{-1} && (33)\\
0 &= \theta_p\bar{T}\theta_{n-p}(\mathrm{Diag}(\theta_{n-p}^T\bar{T}\theta_{n-p}))^{-1} && (34)
\end{aligned}$$

Any $\tilde{\theta}$ that satisfies (31), (32), (33), and (34) is a candidate for for the maximum-likelihood solution. However for any $\tilde{\theta}$ that satisfies (31), (32), (33), and (34), $\tilde{\theta}^T\bar{T}\tilde{\theta}$ and $\tilde{\theta}^T\bar{W}\tilde{\theta}$ have to be of the form

$$\tilde{\theta}^T\bar{W}\tilde{\theta} = \begin{bmatrix} \begin{bmatrix} \lambda_1^W & & 0 \\ & \ddots & \\ 0 & & \lambda_p^W \end{bmatrix} & 0 \\ 0 & A \end{bmatrix} \quad (35)$$

$$\tilde{\theta}^T\bar{T}\tilde{\theta} = \begin{bmatrix} B & 0 \\ 0 & \begin{bmatrix} \lambda_{p+1}^T & & 0 \\ & \ddots & \\ 0 & & \lambda_n^T \end{bmatrix} \end{bmatrix} \quad (36)$$

Where A and B are positive-definite matrices. Due to proposition 1, we can assume w.l.o.g. that $|\tilde{\theta}| = 1$. Then the log-likelihood of the data can be written as

$$\begin{aligned}
\log L_E(\tilde{\theta}|\{x_i\}) &= \frac{-Nn}{2}(1+\log 2\pi)\\
&\quad + \frac{N}{2}\sum_{i=p+1}^{n}\log\lambda_i^T + \frac{N}{2}\sum_{i=1}^{p}\log\lambda_i^W \quad (37)
\end{aligned}$$

Now let U_A and U_B be the unitary matrices that diagonalize A and B respectively. let us also define

$$U = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix} \quad (38)$$

and

$$\check{\theta} = U\tilde{\theta} \quad (39)$$

Then it is easy to verify that

$$\log L_E(\tilde{\theta}|\{x_i\}) = \log L_E(\check{\theta}|\{x_i\}).$$

However from proposition 2, $\log L_E(\check{\theta}|\{x_i\})$ cannot be greater than $\log L_E(\hat{\theta}|\{x_i\})$. Hence the proof.

Note that 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 (33), (34), (24) and (25) 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.

Acknowledgments

This research was supported by the Center for Language and Speech Processing at Johns Hopkins University and Lockheed Martin. Nagendra Kumar would like to thank Prof. Carey Priebe for his interest and encouragement.

References

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, p. 179, 1936.

[2] R. A. Fisher, "The statistical utilization of multiple measurements," *Ann. Eugen.*, vol. 8, p. 376, 1938.

[3] C. R. Rao, *Linear Statistical Inference and Its Applications.* New York: John Wiley and Sons, 1965.

[4] N. Kumar, C. Neti, and A. G. Andreou, "Application of discriminant analysis to speech recognition with auditory features," in *Proceedings of the Fifteenth Annual Speech Research Symposium*, (Johns Hopkins University, Baltimore, MD 21218), pp. 153–160, June 1995.

[5] P. F. Brown, *The Acoustic-Modelling Problem in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University, 1987.

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

[7] T. Hastie and R. Tibshrani, "Discriminant analysis by gaussian mixtures," tech. rep., AT&T Bell Laboratories, 1994.

[8] N. Kumar and A. G. Andreou, "On generalizations of linear discriminant analysis," Tech. Rep. JHU/ECE-96-07, Johns Hopkins University, 1996.

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