# Geometric Linear Discriminant Analysis

*Mark Ordowski and Gerard G. L. Meyer*
Center for Language and Speech Processing
The Johns Hopkins University
3400 North Charles Street
Baltimore, MD 21218, USA
marko@clsp.jhu.edu, gglmeyer@jhu.edu

## ABSTRACT

When it becomes necessary to reduce the complexity of a classifier, dimensionality reduction can be an effective way to address classifier complexity. Linear Discriminant Analysis (LDA) is one approach to dimensionality reduction that makes use of a linear transformation matrix. The widely used Fisher's LDA is "sub-optimal" when the sample class covariance matrices are unequal, meaning that another linear transformation exists that produces lower loss in discrimination power. In this paper, we introduce a geometric approach to Linear Discriminant Analysis (GLDA) that can reduce the number of dimensions from *n* to *m* for any number of classes. GLDA is able to compute a better linear transformation matrix than Fisher's LDA for unequal sample class covariance matrices and is equivalent to Fisher's LDA when those matrices are equal or proportional. The classification problems we present in this paper demonstrate and strongly suggest that geometric LDA can generate the "optimal" classifier in a lower dimension.

## 1 INTRODUCTION

The optimal algorithmic approach to LDA is where the linear transformation minimizes the loss of discrimination power. In the framework of a parameterized classifier [1], GLDA is an approach that appears to be optimal and independent of the relationship between sample class covariance matrices. Since the rate of misclassifications is related to discrimination power of a classifier, we use the number of misclassified test vectors to compare the GLDA classifiers to a Fisher's LDA classifier and the optimal classifier.

After some preliminaries, we present the geometric approach to LDA that reduces an $n\times 1$ feature vector to a scalar. This approach is not amenable to a closed form solution, so we describe the iterative algorithm to find the geometric discriminant. Before providing the generalized solution for GLDA, that is finding a transformation matrix that reduces a $n\times 1$ vector to a $m\times 1$ vector where $m<n$, we present a simple classification problem where GLDA finds the "optimal" classifier in a lower dimension. Finally, we compare the number of misclassifications made by the GLDA classifier, the Fisher's LDA, and the optimal classifier for problems where the sample class covariance matrices are proportional or heteroscedastic.

## 2 PRELIMINARIES

The notation used in this paper and the definition of a parameterized classifier are given. For completeness, we briefly present Fisher's linear discriminant for two reasons. First, we compare GLDA to this widely used approach for LDA. Second, we use the Fisher discriminant as the starting point for the iterative process that finds the geometric discriminant.

### 2.1 Notation and Assumptions

1. We have $c$ classes $\omega_1,\omega_2,\ldots,\omega_c$ where $c$ is known.
2. The feature vector $\mathbf{x}$ is an $n\times 1$ vector in a Euclidean vector space $E^n$, and $n$ is known.
3. Each class has $n_i$ training vectors with $N=\sum_{i=1}^{c} n_i$.
4. The class conditional densities $\Pr(\mathbf{x}|\omega_i)$ are multivariate normal, $N(\mu_i,\Sigma_i)$, where the mean is a column vector.
5. A-priori probabilities $\Pr(w_i)$ are assumed to be equal.

### 2.2 Parameterized Classifiers

A possible approach to reduce classifier complexity consists of using a $n\times m$ matrix, $\Theta_m$, to classify the transformed observation $\mathbf{y}=\Theta_m^T\mathbf{x}$ instead of $\mathbf{x}$. This parameterization of the classifier creates a class of classifiers, of which one can be chosen that is the best solution to the problem. In this paper, we present a geometric approach to LDA that calculates the transformation matrix $\Theta_m$.

### 2.3 Fisher's Linear Discriminant Analysis

Fisher's linear discriminant [2] is found by maximizing one of the criterion functions in (1).

$$f(\theta)=\frac{\theta^T\mathbf{S}_B\theta}{\theta^T\mathbf{S}_W\theta} \quad \text{or} \quad f(\theta)=\frac{\theta^T\mathbf{S}_T\theta}{\theta^T\mathbf{S}_W\theta} \tag{1}$$

Equation (2) defines the matrix $\mathbf{S}_B$, which is the between class scatter matrix, and the matrix $\mathbf{S}_W$, which is the within class scatter matrix. $\mathbf{S}_T$ is the total scatter matrix where $\mathbf{S}_T=\mathbf{S}_B+\mathbf{S}_W$. If there is a limited amount of training, then the criterion function containing the matrix $\mathbf{S}_T$ may provide a more accurate discriminant since more data went into its estimation.

$$\begin{aligned}\mathbf{S}_B&=\frac{1}{N}\sum_{i=1}^{c} n_i(\mu_i-\mu)(\mu_i-\mu)^T \quad \text{with } \mu=\sum_{i=1}^{c}\mu_i\\ \mathbf{S}_W&=\frac{1}{N}\sum_{i=1}^{c} n_i\Sigma_i\end{aligned} \tag{2}$$

The maximization of equation (1) is obtained by computing the eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_B$ or $\mathbf{S}_w^{-1}\mathbf{S}_T$. There are at most $rank(\mathbf{S}_B)-1$ eigenvectors that have non-zero eigenvalues. The $rank(\mathbf{S}_B)$ is at most $c$, the number of classes to discriminate. So, there are at most $c-1$ linear independent eigenvectors for the transformation matrix $\theta_f$ that can be used to reduce the dimension of the problem from $E^n$ to $E^{c-1}$.

# 3 GEOMETRIC LDA

In this section, we present the framework and algorithm to find a geometric linear discriminant, $\mathbf{z}_g$, that will map a column vector $\mathbf{x} \in E^n$ to a scalar $y = \mathbf{z}_g^T\mathbf{x}$ using geometric ideas. The foundation of this approach is a geometric function, $f(\mathbf{z})$, that has a maximum that is correlated with minimum classification error. We then present a simple classification problem to show a case where GLDA generates the "optimal" classifier in a lower dimension.

## 3.1 Geometric Framework

The geometric discriminant, $\mathbf{z}_g$, is found by considering the hyperellipsoids that are defined by contours of constant conditional densities with $\gamma_i > 0$ and $i = 1,2,\ldots,c$.

$$H(\gamma_i) = \left\{\mathbf{x} \in E^n \middle| \Pr(\mathbf{x}|\omega_i) \geq \gamma_i\right\} \tag{3}$$

So, given any non-zero vector $\mathbf{z}$ in $E^n$, let $\underline{\mathbf{v}}_i$ and $\overline{\mathbf{v}}_i$ be vectors in hyperellipsoids $H(\gamma_i)$ that satisfy

$$\mathbf{z}^T\underline{\mathbf{v}}_i \leq \mathbf{z}^T\mathbf{v} \leq \mathbf{z}^T\overline{\mathbf{v}}_i \quad \forall \mathbf{v} \in H(\gamma_i) \tag{4}$$

If the covariance matrix is positive definite, then the sets $H(\gamma_i)$ are strictly convex, closed and bounded, and the points $\mathbf{z}^T\underline{\mathbf{v}}_i$ and $\mathbf{z}^T\overline{\mathbf{v}}_i$ are well defined.

The necessary and sufficient conditions to find both $\underline{\mathbf{v}}_i$, the lower bound, and $\overline{\mathbf{v}}_i$, the upper bound, given a transformation vector $\mathbf{z}$, are described by a boundary point on the hyperellipsoid as given by the equations in (5).

$$\begin{aligned} \Sigma_i^{-1}(\mathbf{v}_i - \mu_i) &= -\lambda\mathbf{z} \\ 0.5(\mathbf{v}_i - \mu_i)^t\Sigma_i^{-1}(\mathbf{v}_i - \mu_i) &= \tilde{\gamma}_i \end{aligned} \tag{5}$$

When the two equations in (5) are solved simultaneously, the lambda has a $\pm$ root. The vector $\underline{\mathbf{v}}_i$ comes from the negative root of lambda and the vector $\overline{\mathbf{v}}_i$ uses the positive root of lambda.

The criterion function for the geometric discriminant is constructed by comparing scalars from two mapping functions. So, class separability is determined by using maps $f_{(i,j)\min}(\mathbf{z})$ and $f_{(i,j)\max}(\mathbf{z})$ defined by equation (6) for each pair of indices $(i,j)$ where $i = 1,2,\ldots,c-1;\ j = i+1, i+2,\ldots,c$.

$$\begin{aligned} f_{(i,j)\min}(\mathbf{z}) &= \max\left\{\mathbf{z}^t\underline{\mathbf{v}}_i, \mathbf{z}^t\underline{\mathbf{v}}_j\right\} - \min\left\{\mathbf{z}^t\overline{\mathbf{v}}_i, \mathbf{z}^t\overline{\mathbf{v}}_j\right\} \\ f_{(i,j)\max}(\mathbf{z}) &= \max\left\{\mathbf{z}^t\overline{\mathbf{v}}_i, \mathbf{z}^t\overline{\mathbf{v}}_j\right\} - \min\left\{\mathbf{z}^t\underline{\mathbf{v}}_i, \mathbf{z}^t\underline{\mathbf{v}}_j\right\} \end{aligned} \tag{6}$$

The geometric criterion function $f(\mathbf{z})$ to be maximized is

$$f(\mathbf{z}) = \frac{\min\left\{f_{(i,j)\min}(\mathbf{z})\right\}}{\max\left\{f_{(i,j)\max}(\mathbf{z})\right\}} \tag{7}$$

## 3.2 Iterative Algorithm

In general, equation (7) does not have a closed form solution[1] and the transformation vector, $\mathbf{z}_g$, must be found using an iterative algorithm. An adjustable step steepest ascent algorithm is used to maximize the criterion function of (7). The iterative approach begins with an initial guess of the solution by using the Fisher discriminant $\theta_f$, an initial step length $\lambda_1 > 0$, and a small scalar $\varepsilon \approx 10^{-5}$ used as a stop rule. The iterative algorithm that maximizes (7) is:

Given $\mathbf{z}_1 = \theta_f, \lambda_1 > 0, \varepsilon > 0$

Step 0: Set $i = 1$.

Step 1: Let $\mathbf{x}_i = \mathbf{z}_i + \lambda_i\nabla f(\mathbf{z}_i)$.

Step 2: $\mathbf{z}_{i+1}, \lambda_{i+1} = \begin{cases} \mathbf{x}_i, \lambda_i & \text{if } f(\mathbf{x}_i) > f(\mathbf{z}_i) \\ \mathbf{z}_i, \lambda_i/2 & \text{else} \end{cases}$

Step 3: If $\left\|\nabla f(\mathbf{z}_i)\right\| \leq \varepsilon$ stop, else go to Step 4.

Step 4: Set $i = i+1$, and go to Step 1.

## 3.3 Demonstration of Geometric LDA

Consider a classification problem with two classes in $E^2$ where each class is normally distributed and the model parameters are given in (8). To illustrate the difference between the geometric LDA and Fisher LDA, we compare the decision regions in $E^2$ resulting from the projection onto the respective discriminants from (9). Figure 1 is an intuitively pleasing result. The geometric discriminant correctly captures the major axis of the ellipse from class 2, which results in a reduced number of misclassified test samples. In fact, we searched all possible linear discriminants and found that the optimal discriminant is indeed $\mathbf{z}_g$ from (9).

$$\begin{aligned} P(\omega_1) = P(\omega_2) = \frac{1}{2};\quad \mu_1 = \begin{bmatrix} 7 \\ 5 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \\ \Sigma_1 = \begin{bmatrix} 27 & 2 \\ 2 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \end{aligned} \tag{8}$$

[1] When $\Sigma_1 = \Sigma_2 = \Sigma$, an analytical solution can be found for the transformation vector $\mathbf{z}$, and it is the Fisher discriminant.

$$\mathbf{z}_g = \begin{bmatrix} 0.9993 \\ 0.0378 \end{bmatrix} \qquad \theta_f = \begin{bmatrix} 0.5855 \\ 0.8107 \end{bmatrix} \tag{9}$$

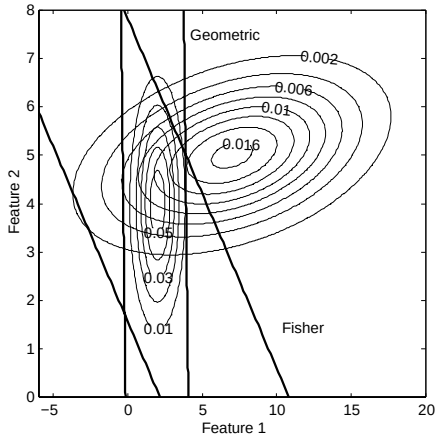


Figure 1: Contour plot for Gaussian distributions defined by equation (8). The decision boundaries are plotted in $E^2$ for the geometric discriminant, the vertical lines, and Fisher's discriminant, the diagonal lines.

Simulating this classification problem 100 times, the following table is a comparison of the average error rate for the optimal classifiers in $E^2$ and $E$, geometric LDA, and Fisher's LDA classifiers in $E$. We used Matlab to generate $N = 10{,}000$ vectors defined by equation (8) and classified those vectors using the four different classifiers. The geometric discriminant has an error rate that is 6% higher than the optimal classifier in $E^2$ and matches the optimal classifier in $E$ whereas Fisher's LDA classifier in $E$ is 56% higher than the optimal classifier in $E^2$.

| | Optimal (in $E^2$) | Optimal (in $E$) | Geometric (in $E$) | Fisher (in $E$) |
|---|---|---|---|---|
| Total Error | 11.0% | 11.7% | 11.7% | 17.2% |

## 4 GENERALIZED GEOMETRIC LDA

In this section, we find an $n \times m$ orthogonal transformation matrix $\Theta_m$, which transforms a vector $\mathbf{x}$ in $E^n$ to a vector $\mathbf{y}$ in $E^m$, where $m < n$. Instead of extending geometric discriminant analysis to calculate the geometric matrix $\Theta_m$ for $m > 1$, we define a procedure that uses GLDA described in section 3.1 and the iterative algorithm in section 3.2 to calculate $\Theta_m$ via stages.

The procedure has two basic steps: 1) find a geometric discriminant given a vector space; 2) reduce that space such that the next geometric discriminant is orthogonal to the previous. At the heart of the process, an intermediate projection matrix followed by a rotation is used to make the training vectors orthogonal to a given geometric discriminant and to remove the flatness in the class distributions created by the projection. This projection is the key to the process that makes the next geometric discriminant orthogonal to the previous.

For completeness, we describe the process of generating the orthogonal transformation matrix, $\Theta_m$, from the $i^{\text{th}}$ step, where $i = 1,2,...,m$. The $i^{\text{th}}$ step begins with the training vectors $\mathbf{x} \in E^{n-i+1}$ and $\Theta_{i-1} = \left[\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \ldots, \tilde{\mathbf{z}}_{i-1}\right]$, where $\tilde{\mathbf{z}}_j$ is a column vector containing the geometric discriminant $\mathbf{z}_j$ augmented with $j-1$ zeros, $j = 1, \ldots, i-1$. The state of the data going into the $i^{\text{th}}$ step becomes more clear after the following two paragraphs.

We first calculate the geometric discriminant, $\mathbf{z}_i$, using model parameters from $E^{n-i+1}$ space. Since it is desirable to have an orthogonal $\Theta_m$, we apply an intermediate transformation, $\mathbf{y} = \mathbf{A}^T\mathbf{x}$, so that every vector $\mathbf{y}$ is orthogonal to $\mathbf{z}_i$ (i.e. $\mathbf{z}_i^T\mathbf{y} = 0$). It can be shown[2] that the matrix, $\mathbf{A} = \mathbf{I} - \mathbf{z}_i\,\mathbf{z}_i^T$, is a projection where every vector $\mathbf{y}$ is orthogonal to the geometric discriminant $\mathbf{z}_i$.

We now have a problem, $\mathbf{y} \in E^{n-i+1}$, where the class distributions are restricted to a hyperplane in $k$ dimensions, where $k = n - i + 1$. It is trivial to remove the $k^{\text{th}}$ dimension by finding a $k \times k$ transformation matrix [3], $\mathbf{R}_i$, that rotates the problem, $\mathbf{v} = \mathbf{R}_i\mathbf{y}$, so that the geometric discriminant $\mathbf{z}_i$ is aligned in the direction of the $k^{\text{th}}$ principal axis. This will make all the data values equal to zero in the $k^{\text{th}}$ dimension. We now have a problem in $k-1$ dimensions with $\tilde{\mathbf{z}}_i$ defined as the geometric discriminant $\mathbf{z}_i$ augmented with $n-k$ zeros. We now can create a $n \times i$ geometric matrix $\Theta_i = \left[\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \ldots, \tilde{\mathbf{z}}_i\right]$.

We repeat the process described above until $m$ geometric discriminants have been found. However, the above process will generate geometric discriminants that have an incorrect orientation to each of the previously generated discriminants. The correct orientation for each geometric discriminant can be recovered by applying the appropriate rotation matrix to unravel the $i-1$ rotations. Equation (10) is the appropriate rotation matrix to insure that all the geometric discriminants are properly oriented to the standard orthonormal basis.

$$\hat{\mathbf{z}}_i = \left( \prod_{j=i-1}^{1} \mathbf{R}_j \right)^T \tilde{\mathbf{z}}_i, \qquad \text{for } i = 1,2,\ldots,m \tag{10}$$

We now have the $n \times m$ geometric transformation matrix $\Theta_m = \left[\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_m\right]$, where $\hat{\mathbf{z}}_i$ is a $n \times 1$ vector and $\Theta_m^T\Theta_m = \mathbf{I}$.

## 5 Comparison of Classifiers in $E^5$

In this final section we demonstrate that the geometric approach to LDA can substantially reduce the misclassification

[2] Show $\mathbf{z}_i^T\mathbf{y} = 0$. $\mathbf{z}_i^T\mathbf{A} = \mathbf{z}_i^T\left(\mathbf{I} - \mathbf{z}_i\mathbf{z}_i^T\right) = \mathbf{z}_i^T - \mathbf{z}_i^T\mathbf{z}_i\mathbf{z}_i^T = \mathbf{z}_i^T - \mathbf{z}_i^T = 0$.

rate. We present a series of classification problems with two-classes in $E^5$ where the sample class covariance matrices are equal or proportional, and heteroscedastic respectively. We used GLDA to create a classifier in $E$, $E^2$, $E^3$, and $E^4$. These four classifiers were compared to Fisher's LDA classifier in $E$ and the optimal classifier in $E^5$. We used the mean and standard deviation of the total number of misclassified test vectors as the comparison point.

Using Matlab, the source model defined by $\Pr(\mathbf{x}|\omega_1)$ and $\Pr(\mathbf{x}|\omega_2)$ was used to generate $N = 50{,}000$ training vectors. The geometric matrix, $\Theta_i$ for $i = 1,2,3,4$, the Fisher transformation vector $\theta_f$, and the class model parameters for the optimal classifier were estimated from the training set. The source model was then used to generate $N = 50{,}000$ test vectors to evaluate the classifiers. This process was repeated 15 times.

The model parameters for the class distributions were written in canonical form. This form allows us to generally evaluate the GLDA approach and infer what would happen given any two-class problem. Given any non-singular matrices[3] $\Sigma_1, \Sigma_2$, there always exists a non-singular transformation [4] to put $\Sigma_1, \Sigma_2$ in canonical form. Equation (11) is the canonical form of model parameters for $N(\mu_1, \Sigma_1)$, $N(\mu_2, \Sigma_2)$ where $v_1 \geq v_2 \geq \ldots \geq v_5 > 0$.

$$\begin{aligned} \mu_1 &= 0, \quad \mu_2 = [m_1, m_2, \ldots, m_5]^T \\ \Sigma_1 &= \mathbf{I}, \quad \Sigma_2 = diag(v_1, v_2, \ldots, v_5) \end{aligned} \tag{11}$$

## 5.1 Equal or Proportional Covariance Matrices

We ran a series of simulations using model parameters defined by equation (11) where $m_i = 0.1$ and $v_i = v$ for $i = 1,2,\ldots,5$. We scaled the problem by multiplying $\Sigma_1$ by 0.01 and varied $v$ logarithmically from $10^{-3}$ to 10. Figure 2 shows the percentage of errors made by each of the six classifiers. The standard deviation of error was less than 0.1% absolute in all cases. As stated earlier, the GLDA classifier in $E^1$ is shown to be equivalent to Fisher's LDA classifier in $E^1$. Figure 2 also shows the difference between the optimal classifier in $E^5$ and the GLDA classifiers in $E^2$, $E^3$, and $E^4$.

## 5.2 Heteroscedastic Covariance Matrices

Using the model parameters defined by equation (11), we ran a series of simulations where $m_i = 0.01$ and $v_i = v / 4^{i-1}$ for $i = 1,2,\ldots,5$. We scaled the problem by multiplying $\Sigma_1$ by 0.01 and varied $v$ logarithmically from $10^{-3}$ to 10. Figure 3 shows that the GLDA classifier in $E^1$ has drastically reduced the number of errors when compared to Fisher's LDA in $E^1$. The error bars are not shown if the error was less than 0.1% absolute.

[3] This assumption can be made with no real loss of generality.

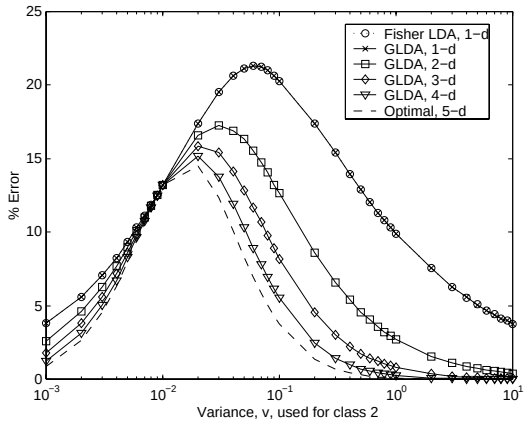


Figure 2: Class covariance matrices that are equal or proportional.

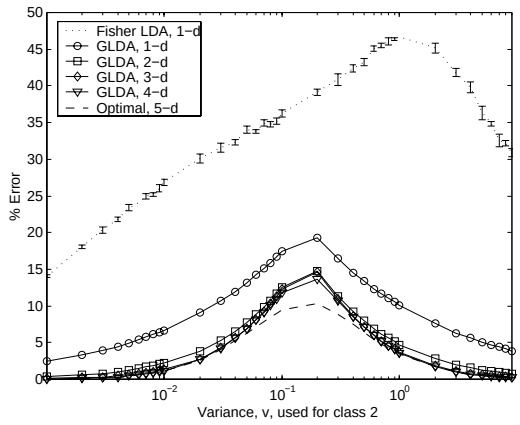


Figure 3: Class covariance matrices that are heteroscedastic.

# 6 CONCLUSION

We have presented a geometric approach to LDA. We have presented some simple classification problems that show this approach gives a linear transformation that better preserves the discrimination power of a classifier in a lower dimension. We have also provided a simple example where the geometric approach generated the optimal classifier in a lower dimension.

# 7 REFERENCES


[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. London: Academic Press, 1990.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, 1973.

[3] B. Noble and J. W. Daniel, *Applied Linear Algebra*, 3rd ed. Englewood Cliffs: Prentice-Hall, 1988.

[4] T. W. Anderson, "Appendix 1: Matrix Theory," in *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1958, pp. 338-341.