International Journal of Applied Mathematics
and Statistical Sciences (IJAMSS)
ISSN (P): 2319–3972; ISSN (E): 2319–3980
Vol. 10, Issue 2, Jul–Dec 2021; 71–80

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

EQUIVALENCE OF FISHER DISCRIMINANT ANALYSIS AND LEAST SQUARE

Chro

Research Scholar, Department of Statistics, Kurdistan Institution for Strategic Studies and Scientific Research, Kurdistan Region, Iraq

ABSTRACT

Linear Discriminant Analysis (LDA) is a well-known method for dimensionality reduction and classification. LDA in the binary-class case has been shown to be equivalent to linear regression with the class label as the output. This implies that LDA for binary class classification can be formulated as a least square problem. However many real-world applications involves multi-class classification, where a least square formulation for LDA is desirable. Previous studies have shown certain relationship between multivariate linear regression and LDA. Many of these studies show that multivariate linear regression with a specific class indicator matrix as the output can be applied as a pre-processing step for LDA. However, directly casting LDA as a least squares problems remains open for the multi-class case.

In this paper used Fisher Linear Discriminant in an original space and finding the coefficients, compare these coefficients with the coefficients of least square method, to show that these methods are equivalent in directions, this equivalent happen when the statistics of Rayleigh Coefficient is maximized.

By using the Iris dataset was introduced by R. A. Fisher as an example for discriminant analysis, that the data report four characteristics (sepal width, sepal length, pedal width and pedal length) of three species of Iris flower with the class label as output. We took just two species to explain the equivalent between LDA and LS.

KEYWORDS: *Fisher Linear Discriminant, Least Square, Rayleigh Coefficient*

Article History

Received: *14 Oct 2021* | ***Revised:*** *20 Oct 2021* | ***Accepted:*** *26 Oct 2021*

1. INTRODUCTION

Linear Discriminant Analysis (LDA) is a traditional statistical method which has proven successful on classification and dimensionality reduction problems [(6)]. The procedure is based on an Eigen value resolution and gives an exact solution of the maximum of the inertia but this method fails for a nonlinear problem.

The original LDA formulation, known as the Fisher linear Discriminant Analysis (FLDA)[(5)] deals with binary-class classification. The key idea in (FLDA) is to look for a direction that separates the class mean well (when projected onto that direction) while achieving a small variance around these means. FLDA bears strong connections to linear regression with the class label as the output for classification. It has been shown[(3, 10)] that FLDA is equivalent to a least square problem.

Fisher's Linear Discriminant Analysis separates multivariate data with different classes nicely in the linear projection. In a two-class data separation, FDA tries to find the projection vector such that the between-class scatter matrix

is maximized and the within-class scatter matrix is minimized. Then the linear projection of this vector will ensure the greatest separatability for the two classes' data[2].

The intuition behind Fisher's linear Discriminant (FLD) consists of looking for a vector of compounds $\mathbf{w}$ such that, when a set of training samples are projected in to it, the class centres are far apart while the spread within each class is small, consequently producing a small overlap between classes[12]. This is done by maximizing a cost function known in some contexts as Rayleigh Coefficient, $J(\mathbf{w})$. The data taken from "Edgar Anderson (1935). "The irises of the Gaspé Peninsula". *Bulletin of the American Iris Society* **59**: 2–5" [4]

Theoretical Part

1. Linear Discriminant

More formally one looking for a function $f : \mathcal{X} \to \mathbb{R}^D$, such that $f(\mathbf{x})$ *and* $f(\mathbf{z})$ are similar whenever $\mathbf{x}$ *and* $\mathbf{z}$ are, and different otherwise. Similarity is usually measured by class membership and Euclidean distance. In the special case in the linear Discriminant analysis one is seeking a linear function, i.e. a set of projections

$$f(\mathbf{x}) = \mathbf{W}^T\mathbf{x} \qquad W \in \mathbb{R}^{N\times D}$$

where the matrix W is chosen, such that a contrast criterion G is optimized, in some cases with respect to a set of constraints S, i.e.

$$\max\ G(W) \ \textbf{subject to}\ W \in S \qquad \textbf{(1)}$$

This setup is absolutely equivalent to e.g. principle component analysis where the contrast criterion would be that of maximal variance (or least mean square error) and the constraint set would be that of orthogonality of the matrix W. However, PCA is an unsupervised technique and does not use any label. There is no principle that the direction found by PCA will be particularly discriminative.

To simplify the presentation we will in the following only consider one-dimensional Discriminant functions, i.e. f is of the form $f = (\mathbf{w}\cdot\mathbf{x})$. However, most results can easily be generalized to the multidimensional case.(10, 12)

2. Fisher's Discriminant

Probably the most well known example of a linear Discriminant is Fisher's Discriminant Fisher's idea was to look for a direction $\mathbf{w}$ that separates the class means well (when project onto the found direction) while achieving a small variance around these means[8]. The hope is that it is easy to decide for either of the two classes from this projection with a small error. The quantity measuring the difference between the means is called *between class variance* and the quantity measuring the variance around these class means is called *within class variance*, respectively. Then the goal is to find a direction that maximizes the between class variance while minimizing the within class variance at the same time. This is illustrated in Figure bellow.

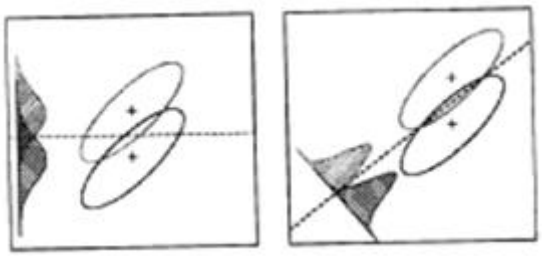

Figure1: Fisher Discriminant Analysis[13]

As shown in left graph, the two-class data is linearly projected onto a direction (vector) of $\vec{m}_1 - \vec{m}_2$, which is not a good separation as there are a lot data from two class overlap with each other. For the right graph, the two-class data are separated in a nice way that they have minimum overlapping.

To describe this mathematically let $\mathcal{X}$ denote the space of observations (e.g. $\mathcal{X} \subseteq \mathbb{R}^N$) and $\mathcal{Y}$ the set of possible labels (here $\mathcal{Y}$ = {+1, -1}). Furthermore, let $\mathcal{Z} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_M, y_M)\} \subseteq \mathcal{X} \times \mathcal{Y}$ denote the training sample of size M and denote by $\mathcal{Z}_1 = \{(\mathbf{x}, y) \in \mathcal{Z} | y = 1\}$ and $\mathcal{Z}_2 = \{(\mathbf{x}, y) \in \mathcal{Z} | y = -1\}$ the split in to the two classes of size $M_i = |\mathcal{Z}_i|$. Define $\mathbf{m}_1$ *and* $\mathbf{m}_2$ to be the empirical class means. i.e.

$$\mathbf{m}_i = \frac{1}{M_i} \sum_{x \in \mathcal{Z}_i} \mathbf{x}$$

Similarly, we can compute the means of the data projected onto some direction $\mathbf{w}$ by

$$\begin{aligned} «_i &= \frac{1}{M_i} \sum_{x \in \mathcal{Z}_i} \mathbf{W}^T \mathbf{x} \\ &= \mathbf{W}^T \mathbf{m}_i \end{aligned} \tag{2}$$

i.e. the means $«_i$ of the projection means $\mathbf{m_i}$. The variances $\pm_1^2, \pm_2^2$ of the projected data can be expressed as

$$\pm_i^2 = \sum_{x \in Z_i} \left(\mathbf{W}^T \mathbf{x} - «_i\right)^2 \tag{3}$$

Then maximizing the between class variance and minimizing the within class variance can be achieved by maximizing

$$G(\mathbf{W}) = \frac{(«_1 - «_2)^2}{\pm_1^2 + \pm_2^2} \tag{4}$$

Which will yield a direction W such that the ratio of between-class variance (i.e. separation) and within class variance (i.e. overlap) is maximal. Now, substituting the expression (2) for the means and the expression (3) for the variance into above equation (4) yields

$$G(\mathbf{W}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{5}$$

Where we define the between and within class scatter matrices $S_B \; and \; S_W$ as

$$S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \;\; S_W = \sum_{i=1,2} \sum_{x \in Z_i} (\mathbf{x} - \mathbf{m}_i)^2 \tag{6}$$

And the vector of coefficients as

$$\mathbf{W} = S_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \tag{7}$$

It is straight forward to check that (4) is absolutely equivalent to (5). This perfectly fits into the framework (1) with an empty constraints set S. The equation $G(W)$ is often referred to as a Rayleigh coefficient.[7, 8, 9, 10, 12]

3. Connection to Least Square

The Fisher Discriminant problem described above bears strong connections to least squares approaches for classification. Classically, one is looking for a linear Discriminant function, now including a bias term, i.e.

$$f(\mathbf{x}) = \mathbf{W}^T \mathbf{x} + b \tag{9}$$

such that on the training sample the sum of squares error between the outputs $f(\mathbf{x}_i)$ and the known targets y_i is small, i.e. in a (linear) least squares approach one minimizing the sum of square

$$\mathrm{E}(\mathbf{W}, b) = \sum_{(\mathrm{x},\mathrm{y}) \in Z} (f(\mathbf{x}) - y)^2 = \sum_{(\mathrm{x},\mathrm{y}) \in Z} (\mathbf{W}^T \mathbf{x} + b - y)^2 \tag{10}$$

The least squares problem $\min_{\mathbf{W},b} \mathrm{E}(\mathbf{W}, b)$ can be written in matrix rotation as

$$\min_{W,b} \left\| \begin{bmatrix} X_1^T & 1_1 \\ X_2^T & 1_2 \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ b \end{bmatrix} - \begin{bmatrix} -\mathbf{1}_1 \\ \mathbf{1}_2 \end{bmatrix} \right\|^2 \tag{11}$$

where $X = [X_1 \;\; X_2]$ is a matrix containing all training examples partitioned according to the labels ± 1, and 1_i is a vector of once of corresponding length. The solution to a least rectangle problem of the form $\|A\mathbf{x} - b\|^2$ can be computed by using the *pseudo-inverse* of A, i.e. $\mathbf{X}^* = A^{\dagger} b = (A^T A)^{-1} A^T b$ assuming that $A^T A$ is not singular. Then $A^{\dagger} A = I$ and thus a necessary and sufficient condition for the solution $\mathbf{X}^*$ to the least square problem is $(A^T A)\mathbf{x}^* = A^T b$. Appling this to (11) yields

$$\begin{bmatrix} X_1 & X_2 \\ 1_1^T & 1_2^T \end{bmatrix} \begin{bmatrix} X_1^T & 1_1 \\ X_2^T & 1_2 \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ b \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ 1_1^T & 1_2^T \end{bmatrix} \begin{bmatrix} -\mathbf{1}_1 \\ \mathbf{1}_2 \end{bmatrix}$$

multiplying these matrices and using the definition of the sample means and within class scatter for Fisher yields:

$$\begin{bmatrix} S_W + M_1\mathbf{m_1}\mathbf{m_1^T} & M_1\mathbf{m_1} + M_2\mathbf{m_2} \\ (M_1\mathbf{m_1} + M_2\mathbf{m_2})^T & M_1 + M_2 \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ b \end{bmatrix} = \begin{bmatrix} M_2\mathbf{m_2} - M_1\mathbf{m_1} \\ M_2 - M_1 \end{bmatrix} \tag{12}$$

Using the second equation in (12) to solve for b yields

$$b = \frac{M_2 - M_1 - (M_1\mathbf{m_1} + M_2\mathbf{m_2})^T \mathbf{W}}{M_1 + M_2} \tag{13}$$

Substituting this into first equation of (12) and using a few algebraic manipulations, especially the relation $a - \frac{a^2}{a+b} = \frac{ab}{a+b}$ one obtains**:**

$$\left(S_W + \frac{M_1 M_2}{M_1+M_2} S_B\right)\mathbf{W} + \frac{M_1^2+M_2^2}{M_1+M_2}(\mathbf{m}_2 - \mathbf{m}_1) = 0 \tag{14}$$

Now, since still $S_B\mathbf{W}$ is in the direction of $(\mathbf{m}_2 - \mathbf{m}_1)$, there exists a scalar $\sim \in R$ such that

$$\frac{M_1 M_2}{M_1+M_2} S_B\mathbf{W} = -\left(\frac{M_1^2+M_2^2}{M_1+M_2} - \sim\right)(\mathbf{m}_2 - \mathbf{m}_1) \tag{15}$$

Then using (15) in (14) yields:

$$S_W\mathbf{W} = \sim (\mathbf{m}_2 - \mathbf{m}_1) \Leftrightarrow \mathbf{W} = \sim S_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1). \tag{16}$$

This shows that the solution to the least square problem is in the same direction as the solution of Fisher's Discriminant, although it will have a different length. But as we already noticed, we are only interested in the *direction* on $\mathbf{W}$, not its length and hence the solutions are identical**.**[1,7,8,10,11]

Practical Part

This paper has been prepared to clear that the LDA is equivalent to least square regression entering the value of compounds $\mathbf{w}$ and a statistic Rayleigh coefficient measure the ration of projected class means to projected intra-class variance we obtain the optimal solution, means maximizing the statistic Rayleigh coefficient when a set of training sample are projected into it the class centres are far apart while the spread within each class is small, by using packages SPSS and MATLAB, and for the data see Appendix (A).

From equation (7), for LDA, the vector of coefficients i.e. Standardized Canonical Discriminant Function Coefficients are:

$$\mathbf{W} = \begin{bmatrix} -.583 \\ -.303 \\ 1.069 \\ .547 \end{bmatrix}$$

and from equation (5) the value of statistic Rayleigh Coefficient

$$G(\mathbf{W}) = 88.3829 .$$

From equation (16), for LS, the vector of coefficients are

$$\mathbf{W} = \begin{bmatrix} .150 \\ .050 \\ -.765 \\ -.337 \end{bmatrix}$$

and from equation (5) the value of statistic Rayleigh coefficient $G(\mathbf{W}) = 76.1694$

It is clear that from the values of both two coefficients and statistic of Rayleigh coefficients of Linear Discriminant and least Squares the separate of between-class scatter matrix is maximized and the within-class scatter matrix is minimized that means there are an equivalent between them, because they have the same direction, although it will have a different length. But as we already noticed, we are only interested in the *direction* on $\mathbf{W}$, not its length. To be certain from table (1) bellow of classification result from SPSS analysis clear that 100.0% of original grouped cases correctly classified in this data. In general to be the equivalent is strong there must be the ratio of misclassification is small.

Table 1: Classification Results

			Predicted Group Membership		
		SP	-1	1	Total
Original	Count	-1	50	0	50
		1	0	50	50
	%	-1	100.0	.0	100.0
		1	.0	100.0	100.0

a. 100.0% of original grouped cases correctly classified.

CONCLUSION

From the analysis of Iris Flower dataset was introduced by R. A. Fisher[4] as an example for Discriminant analysis, that the data report four characteristics (sepal width, sepal length, pedal width and pedal length) of three species of Iris flower with the class label as output. We took just two species to explain the equivalent between LDA and LS clear that there is an equivalent between the Linear Fisher Diacriminant and Least Squares method means that the separate of between-class scatter matrix is maximized and the within-class scatter matrix is minimized, because they have the same direction, although it will have a different length.

REFERENCES

1. *Bishop, C.M. "Pattern Recognition and Machine Learning". Springer, 2006.*
2. *Cai Yundong "Glass Identification with FDA and Kernel FDA," Assignment report of "Computational Intelligence: Methods and Application," 2005.*

3. *Duda, R.O; P.E. Hart, and D. Stork. Pattern Classification. Wiley, 2000.*
4. *Edgar Anderson (1935). "The irises of the Gaspé Peninsula". Bulletin of the American Iris Society* ***59****: 2–5.*
5. *Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics* ***7****: 179–188. http://digital.library.adelaide.edu.au/coll/special//fisher/138.pdf.*
6. *Fukunaga K.," Introduction to Statistical Pattern Recognition", Academic press, INC, 2nd ed, 1990*
7. *Herbrich, H. "Learning Kernel Classifiers. Theory and Application". MIT Press, Cambridge 2002.*
8. *Jieping Ye "Least Squares Linear Discriminant Analysis". Arizona State University. 2007.*
9. *Nguyen Quang Huy, "Linear Discriminant Analysis in Data Visualization", Assignment report of "Computational Intelligence: Methods and Applications", 2005*
10. *Mika, S. Kernel Fisher Discriminants. PhD thesis, University of Technology, Berlin, 2002.*
11. *N. Ctistianini, C.M; J. Shawe-Taylor. "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods". Cambridge University Press, 2000.*
12. *Scholkopf and A.J. Smola. "Learning with Kernels. MIT Press, Cambridge, MA, 2002"*
13. *Wlodzislaw Duch. Computational Intelligence: Methods and Application, Lecture 7: Discriminant Component Analysis.*

Appendix (A)

Iris flowers dataset was introduced by R. A. Fisher [4] as an example for Discriminant analysis, that the data report four characteristics (sepal width, sepal length, pedal width and pedal length) of three species of Iris flower with the class label as output. We took just two species to explain the equivalent between LDA and LS.

Table 2

Fisher's Iris Data				
Sepal Length	**Sepal Width**	**Petal Length**	**Petal Width**	**Species**
5.1	3.5	1.4	0.2	*setosa*
4.9	3.0	1.4	0.2	*setosa*
4.7	3.2	1.3	0.2	*setosa*
4.6	3.1	1.5	0.2	*setosa*
5.0	3.6	1.4	0.2	*setosa*
5.4	3.9	1.7	0.4	*setosa*
4.6	3.4	1.4	0.3	*setosa*
5.0	3.4	1.5	0.2	*setosa*
4.4	2.9	1.4	0.2	*setosa*
4.9	3.1	1.5	0.1	*setosa*
5.4	3.7	1.5	0.2	*setosa*
4.8	3.4	1.6	0.2	*setosa*
4.8	3.0	1.4	0.1	*setosa*
4.3	3.0	1.1	0.1	*setosa*
5.8	4.0	1.2	0.2	*setosa*
5.7	4.4	1.5	0.4	*setosa*
5.4	3.9	1.3	0.4	*setosa*
5.1	3.5	1.4	0.3	*setosa*

Table 2 Contd.,

5.7	3.8	1.7	0.3	*setosa*
5.1	3.8	1.5	0.3	*setosa*
5.4	3.4	1.7	0.2	*setosa*
5.1	3.7	1.5	0.4	*setosa*
4.6	3.6	1.0	0.2	*setosa*
5.1	3.3	1.7	0.5	*setosa*
4.8	3.4	1.9	0.2	*setosa*
5.0	3.0	1.6	0.2	*setosa*
5.0	3.4	1.6	0.4	*setosa*
5.2	3.5	1.5	0.2	*setosa*
5.2	3.4	1.4	0.2	*setosa*
4.7	3.2	1.6	0.2	*setosa*
4.8	3.1	1.6	0.2	*setosa*
5.4	3.4	1.5	0.4	*setosa*
5.2	4.1	1.5	0.1	*setosa*
5.5	4.2	1.4	0.2	*setosa*
4.9	3.1	1.5	0.2	*setosa*
5.0	3.2	1.2	0.2	*setosa*
5.5	3.5	1.3	0.2	*setosa*
4.9	3.6	1.4	0.1	*setosa*
4.4	3.0	1.3	0.2	*setosa*
5.1	3.4	1.5	0.2	*setosa*
5.0	3.5	1.3	0.3	*setosa*
4.5	2.3	1.3	0.3	*setosa*
4.4	3.2	1.3	0.2	*setosa*
5.0	3.5	1.6	0.6	*setosa*
5.1	3.8	1.9	0.4	*setosa*
4.8	3.0	1.4	0.3	*setosa*
5.1	3.8	1.6	0.2	*setosa*
4.6	3.2	1.4	0.2	*setosa*
5.3	3.7	1.5	0.2	*setosa*
5.0	3.3	1.4	0.2	*setosa*
7.0	3.2	4.7	1.4	*versicolor*
6.4	3.2	4.5	1.5	*versicolor*
6.9	3.1	4.9	1.5	*versicolor*
5.5	2.3	4.0	1.3	*versicolor*
6.5	2.8	4.6	1.5	*versicolor*
5.7	2.8	4.5	1.3	*versicolor*
6.3	3.3	4.7	1.6	*versicolor*
4.9	2.4	3.3	1.0	*versicolor*
6.6	2.9	4.6	1.3	*versicolor*
5.2	2.7	3.9	1.4	*versicolor*
5.0	2.0	3.5	1.0	*versicolor*
5.9	3.0	4.2	1.5	*versicolor*
6.0	2.2	4.0	1.0	*versicolor*
6.1	2.9	4.7	1.4	*versicolor*
5.6	2.9	3.6	1.3	*versicolor*
6.7	3.1	4.4	1.4	*versicolor*
5.6	3.0	4.5	1.5	*versicolor*
5.8	2.7	4.1	1.0	*versicolor*
6.2	2.2	4.5	1.5	*versicolor*
5.6	2.5	3.9	1.1	*versicolor*
5.9	3.2	4.8	1.8	*versicolor*
6.1	2.8	4.0	1.3	*versicolor*
6.3	2.5	4.9	1.5	*versicolor*
6.1	2.8	4.7	1.2	*versicolor*

Table 2 Contd.,

6.4	2.9	4.3	1.3	*versicolor*
6.6	3.0	4.4	1.4	*versicolor*
6.8	2.8	4.8	1.4	*versicolor*
6.7	3.0	5.0	1.7	*versicolor*
6.0	2.9	4.5	1.5	*versicolor*
5.7	2.6	3.5	1.0	*versicolor*
5.5	2.4	3.8	1.1	*versicolor*
5.5	2.4	3.7	1.0	*versicolor*
5.8	2.7	3.9	1.2	*versicolor*
6.0	2.7	5.1	1.6	*versicolor*
5.4	3.0	4.5	1.5	*versicolor*
6.0	3.4	4.5	1.6	*versicolor*
6.7	3.1	4.7	1.5	*versicolor*
6.3	2.3	4.4	1.3	*versicolor*
5.6	3.0	4.1	1.3	*versicolor*
5.5	2.5	4.0	1.3	*versicolor*
5.5	2.6	4.4	1.2	*versicolor*
6.1	3.0	4.6	1.4	*versicolor*
5.8	2.6	4.0	1.2	*versicolor*
5.0	2.3	3.3	1.0	*versicolor*
5.6	2.7	4.2	1.3	*versicolor*
5.7	3.0	4.2	1.2	*versicolor*
5.7	2.9	4.2	1.3	*versicolor*
6.2	2.9	4.3	1.3	*versicolor*
5.1	2.5	3.0	1.1	*versicolor*
5.7	2.8	4.1	1.3	*versicolor*