Sample for Ph.D candidates (This is an excerpt from the first chapter of my undergraduate thesis, the original is in Spanish)

Salvador Garcia

15 February 2017

1 Introduction

The Fisher's Linear Discriminant Analysis (FLDA) is one statistical method that is used to reduce the dimensionality of a dataset as a previous step in a classification problem. The dimensionality reduction for this method is special because the final objective is to separate two or more classes into a new space. The method was proposed by Ronald A. Fisher (1936), and it was generalized by C. R. Rao (1948).

The FLDA is used in machine learning, in particular for problems of pattern recognition. It can be applied to areas like medicine, finance and computer science. The optimization problem of the FLDA consists in the following: Given a set of observations that have an assigned class, we try to find the best matrix V^{**} that projects each observation into a $k-dimensional$ subspace, with k much lower than the original size space. When the best matrix is found, then simpler clusterization methods can be used (as knn) to make a better classification model.

The FLDA finds the optimal solution of a trace ratio, with the Between-class scatter matrix in the numerator and the Within-class scatter matrix in the denominator. To maximize this objective function was considered computationally expensive in the past, so it was replaced by simpler formulations. The objective of this work is to use the method of Newton-Lanczos as an efficient computational solution. This solves iteratively the maximization problem very fast because of the Newton iterations. To measure the performance of this method in the problem, it will be compared with the precision and the computational time with other linear classification algorithms.

The first chapter of this thesis works with the optimization problem used in the FLDA, for this the basic concepts will be defined. Then, an equivalent formulation will be used, one that is much easier to solve. This new formulation includes a new objective function $f(\rho)$ and a one-dimentional argument ρ. At the end, the existence and uniqueness conditions for the solution of $f(\rho)$ will be given.

In the second chapter, the method Newton-Lanczos will be introduced. This finds the optimum ρ for the proposed $f(\rho)$. As the first step, the Lanczos method will be described with its computational cost and the advantages over the traditional methods to calculate eigenpairs. Then, the Newton algorithm and the Lanczos procedure will be used together to find an efficient solution. At the last part, the optimality conditions will be given.

In the third chapter the numerical experiments will be presented. The used datasets were from the companies *Otto Group* y *State Farm*. These were selected because they have a very high dimensionality and its computationally expensive to use traditional methods. In addition, a comparison in terms of time and precision with the traditional methods like the Linear Discriminant Analysis and the Multiple Linear Regression will be made.

For this thesis the computational language R was used. In addition, the libraries *ProjectTemplate*, *tidyr*, *ggplot2*, *stringr* and *dplyr*. All the computations were made using the library *LAPACK (Fortran)* in the version for OS X 10.11.4 (*liblapack.dylib)*.

2 Fisher's Linear Discriminant Analysis

The main objective of this chapter is to introduce the Fisher's Linear Discriminant Analysis (FLDA), which optimize the trace $\mathrm{Tr}(V^TAV)/\mathrm{Tr}(V^TBV)$ over the set of orthogonal matrices V with A,B positive definite matrices. To give a solution to the problem, different formulations of the problem have been given in statistical learning and pattern recognition books, but the solution to the original formulation is not given because is complicated to compute [6] [5]. As a consequence, different reformulations of the problem are now popular: maximize the trace of the between-class scatter matrix with a restriction over the within-scatter matrix (similar to the PCA problem); or maximize the ratio of determinants (Trace determinant ratio) or maximize the trace of the ratio of these matrices (Ratio trace maximization). [1] [3] [4] [2]. At the end, all the formulations are simplified objective functions of the original one.

In this chapter, an efficient solution for the FLDA using the Newton-Lanczos method will be given. In the first section, the problem will be introduced as a problem of Machine Learning. In the second, the theory will be given to follow easily the text. In the third section, an intuitive solution for the problem in the case of one dimension projectors will be presented and then it will be generalized to p dimensions. At last, the necessary and sufficient conditions for the optimization problem will be derived.

2.1 Machine Learning

Machine learning is founded in two research areas: Computer Sciences and Statistics. From the first, it takes the questions: How can we build machines that solves problems? and, Given the actual technology, What kind of problems are feasible to solve? On the other hand, from Statistics it tries to answer: What conclusions can be inferred from the dataset? and How can we manage the uncertainty of this method? [4]. The joint work from both areas to try to answer these questions helps to build a computational statistical framework of machine learning.

2.1.1 The learning process

A machine *learns* given a task (T), a measure of the performance (R) and a type of experience (E) if the system improves the performance (R) in the task (T) with this experience (E) [4]. With the data, we try to model a structure in order that the machine improves their performance when it receives more information. The diversity of the task, as well as the number of applications is big. For example:

- *Spam/no-spam classification*, Here, (E) are the emails, (T) the task to classify correctly the *spam* y (R) the proportion of correctly classified emails.
- *Face recognition/classification*, Here, (E) are the faces of distinct peoples, (T) is the correct classification of the faces and (R) is then measured as the percentage of correct classified faces.

The learning process has many applications and distinct assumptions. For this reason, is useful to provide a framework that groups all the methods given some criteria. The classification used here is the proposed for T. Hastie [3]. This divides the methods into two groups: Supervised learning and unsupervised learning. The first assumes the an output variable helps to build the structure of the model. Examples of this are the Linear Regression, the decision and classification trees (CART) and the Support Vector Machines (SVM). On the other hand, the unsupervised learning only uses the information of the independent variables. For example, cluster analysis, association rules and some dimensionality reduction methods.

After this first classification, subgroups of the supervised methods are made depending on the output variable [1] When the model considers a quantitative variable, then it receives the name of regression, and when it is a categorical variable it receives the name of classification. On the other hand, the unsupervised learning has two main branches [3]. The first is called segmentation, in which all the observations that are in the same group are homogeneous between them, but between groups they are heterogeneous. And the second group is the dimensionality reduction methods, where we try to project the observations of the dataset into a subspace of the original space generated from the dataset.

The Fisher's LDA is a branch of the supervised learning, in particular of the classification methods. Alternative linear methods for this, is the Logistic Regression, the Classic Linear Discriminant Analysis and the Support Vector Machines.

2.2 Dispersion matrices

First of all, the nomenclature will be defined. $\mathbb{C} = k$ with $k = 1,2,3,...,K$ will be the set of the K distinct classes where the each observation x_i with $i = 1,...,N$ can be assigned and N is the total number of observations. Let's define C_k as the subset of the observations $1,2,3,...,N$ that belongs to the class k. This way, let's define as N_k to the number of observations in class k or equivalently,

[1] In this text will be used the terms input variable and output variables as the dependent and independent variables respectively

the cardinality of C_k. As the last definition $w_i = V^T x_i$ is the data multiplied with the matrix V, or equivalently, the projected observation. Then, the means of each group k is μ_k and the global mean of all the observation is defined as μ:

$$\mu_k = \frac{1}{N_k} \sum_{i \in C_k} x_i \tag{1}$$

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

On the other hand, let's define the mean of the projected observations w_i as:

$$\widetilde{\mu_k} = \frac{1}{N_k} \sum_{i \in C_k} w_i \tag{3}$$

$$\widetilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} w_i \tag{4}$$

The Fisher's Linear Discriminant Analysis is used in its formula the dispersion matrices. In specific, the covariance matrix, the scatter matrix of all the observations, the within-class scatter and the between-class scatter. It is important to analyze the terminology and the formulas that will be used in this thesis in order to understand the logic behind the formulation of the optimization problem.

Σ will represent the covariance matrix of all the observations. Let's define as $\widehat{\Sigma}$ the unbiased estimator of Σ which is divided by $N-1$:

$$\widehat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \tag{5}$$

If this matrix is not scaled by the $N-1$ factor, then it is known as Scatter matrix. In this thesis, is represented as S_T, with the sub index T meaning that is taking all the observations for the formula:

$$S_T = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \tag{6}$$

If the matrix is just considering the observations that belongs to the class k, this will be represented as S_k, with the sub index k meaning that it is taking the scatter of just the observations with class k:

$$S_k = \sum_{i=1 \in C_k} (x_i - \mu)(x_i - \mu)^T$$

In the same way we are going to define the Within-class scatter matrix, as the sum over all the S_k scatter matrix presented above.

$$S_I = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{7}$$

Now, let's define the Between-class scatter matrix as the sum of all the squared differences of the mean of each class with respect to the global mean of all the observations. This is multiplied by the number of observations in each class k (N_k).

$$S_E = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{8}$$

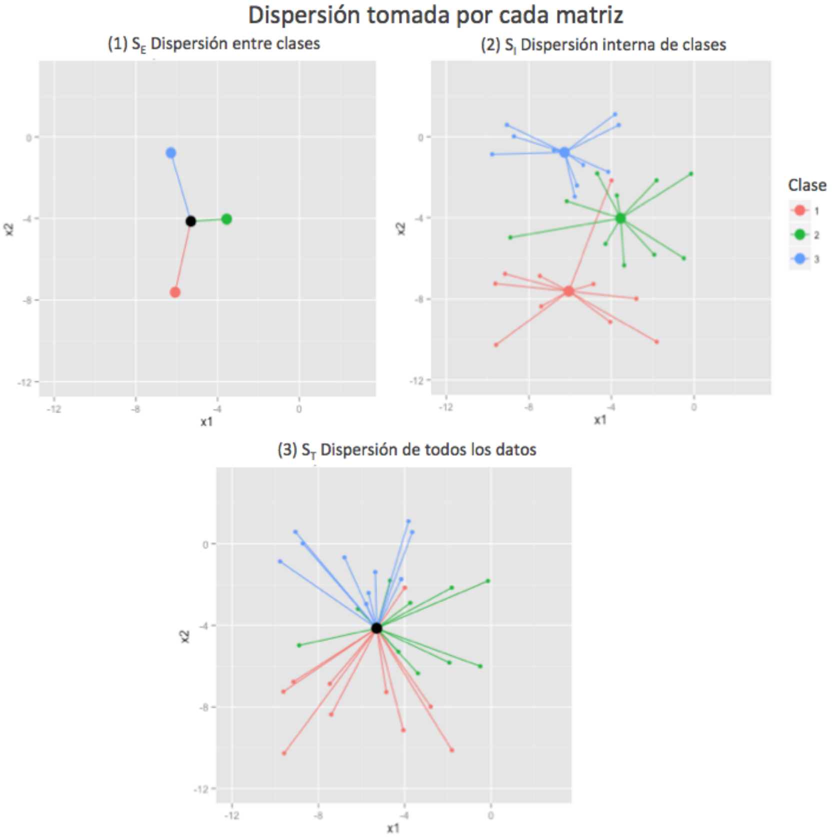

Figure 1: In the plot (1) it is represented the distances between the mean of all the observations (black points) and the means of each class (color points). The plot (2) represents S_I: the squared difference between each observation and the mean of all the observations that are in the same class. Finally, in the plot (3), it is represented S_T, the scatter of all the data with respect to the global mean

Between the Within-class scatter matrix S_I, the Between-class scatter matrix and the total scatter matrix there is an important relationship. It is always true that $S_T = S_I + S_E$; then, if we add the information for each class to the observations, the total scatter matrix can be factored as the sum of the Within-class scatter matrix and the Between-class scatter matrix.

As an example of the relation $S_T = S_I + S_E$, 10 normally distributed observations were generated for each class. (The correlation coefficient for the generated data is -0.005). Then, calculating the scatter matrices from the formulas (1.6), (1.7) and (1.8), we have the following equivalences:

Class	**Distribution x1**	**Distribution x2**
1	N(-5, 2.5)	N(-8, 2)
2	N(-3, 2.5)	N(-4, 2)
3	N(-7, 2.5)	N(-1, 2)

S_I	S_E	S_T
$\begin{bmatrix} 186.05 & 2.78 \\ 2.78 & 94.58 \end{bmatrix}$	$\begin{bmatrix} 46.13 & -4.15 \\ -4.15 & 234.57 \end{bmatrix}$	$\begin{bmatrix} 232.18 & -1.36 \\ -1.36 & 329.16 \end{bmatrix}$

This is important because, as seen in the next chapter, the proposed problem can be rewritten just with the total scatter matrix in the numerator of the trace ratio instead of the between-class scatter matrix. Now, a very common problem that arises in machine learning problems is that the cost of the computation can become intractable when the dimensionality of the observations increases. In the Fisher's Linear Discriminant Analysis it is required that the scatter matrices of the observations be recomputed at each iteration and also, to calculate the inverse or pseudo-inverse of a matrix. These steps can become quite expensive. As a solution to the problem of dimensionality (or commonly named curse of dimensionality) here we propose to use the Principal Component Analysis as a prior step to reduce the dimensionality of the original data. This method is very easy to calculate and the standard linear algebra libraries are optimized to do this [5]. Due to the objective of this thesis no more techniques will be considered for the dimensionality reduction, but books as [3] and [1] propose distinct methods to solve this problem.

Now, coming back to the original problem, we want to find the projection that keeps together the observations that belong to the same class, at the same time that the centroid of each class is far. When this projection is obtained, then is easy to find an hyper-plane that separates the different classes.

The optimization problem can be formulated as:

$$\max_{\substack{V \in \mathbb{R}^{n\times p} \\ V^T V = I}} \frac{\operatorname{Tr}(V^T S_E V)}{\operatorname{Tr}(V^T S_I V)} \tag{9}$$

The solution for this problem does not have a closed form, so in the literature distinct formulations have arose to solve it analytically [6] [2]. Some of these formulations are listed below:

$$\max_{\substack{V \in \mathbb{R}^{n\times p} \\ V^T S_I V = I}} \operatorname{Tr}(V^T S_E V) \tag{10}$$

$$\max_{\substack{V \in \mathbb{R}^{n\times p} \\ V^T V = I}} \operatorname{Tr}\left(\frac{V^T S_E V}{V^T S_I V}\right) \tag{11}$$

$$\max_{\substack{V \in \mathbb{R}^{n\times p} \\ V^T V = I}} \frac{|V^T S_E V|}{|V^T S_I V|} \tag{12}$$

with $|\bullet| = det(\bullet)$ y $\operatorname{Tr}(\bullet) = Trace(\bullet)$.

In the next part of this chapter, the original problem (1.9) will be solved when $p = 1$, and for this reason, the generalized Rayleigh quotient is introduced. For the generalization to p dimensions the problem will be reformulated and the existence and uniqueness will be determined. Afterwards, a new function $f(\rho)$ will be defined. This is useful because is easier to work numerically with this new function instead of working with the original trace ratio. At last, using the eigenvalues of S_I y S_E the optimal value ρ^* will be upper and lower bounded.

2.3 Trace ratio problem

In this text, when using the term projection it will always refer to the orthogonal transformation given by the matrix $V = [V_1|V_2|...|V_p]$ with $V \in \mathbb{R}^{n\times p}$. [2] Let $X = [X_1|X_2|...|X_m]$ the matrix of observations of size $\mathbb{R}^{n\times m}$, then when we expand $V^T X$:

$$V^T X = \begin{pmatrix} V_1^T(X_1) & V_1^T(X_2) & \dots & V_1^T(X_m) \\ V_2^T(X_1) & V_2^T(X_2) & \dots & V_2^T(X_m) \\ \vdots & \vdots & \ddots & \vdots \\ V_p^T(X_1) & V_p^T(X_2) & \dots & V_p^T(X_m) \end{pmatrix}$$

As can be seen, each entry of the matrix $V^T X$ is equal to $V_i^T X_j$: a dot product between a column of V and a row of X. On the hand, we have that $proj_{\mathbf{V_i}}\mathbf{X_j} = \frac{X_j \cdot V_i}{||V_i||^2} V_i$. Like $||V_i|| = 1$, then the formula is simplified as $proj_{\mathbf{V_i}}\mathbf{X_j} = (X_j \cdot V_i)V_i$; equivalently a vector of norm $(X_j \cdot V_i)$ in the direction of V_i. Because of this reason, this is the projected observations: each observation is projected by each column of V.

The trace ratio problem is easy to analyze when $V \in \mathbb{R}^{n\times p}$ projects to a lower dimensional space. For example, when $p = 2$, then the problem is equivalent to find the best projection to a plane, and when $p = 1$, to the best projection of a vector. As an example, a synthetic dataset was created, where each $x_i \in \mathbb{R}^3$. In this dataset the distributions are normally distributed and are projected in$\mathbb{R}^2$ and $\mathbb{R}^1$. These data can be observed in the figure 1.3.

[2] This matrix has orthogonal columns, then assuming that each column has norm 1, then $V^T V = I_p$

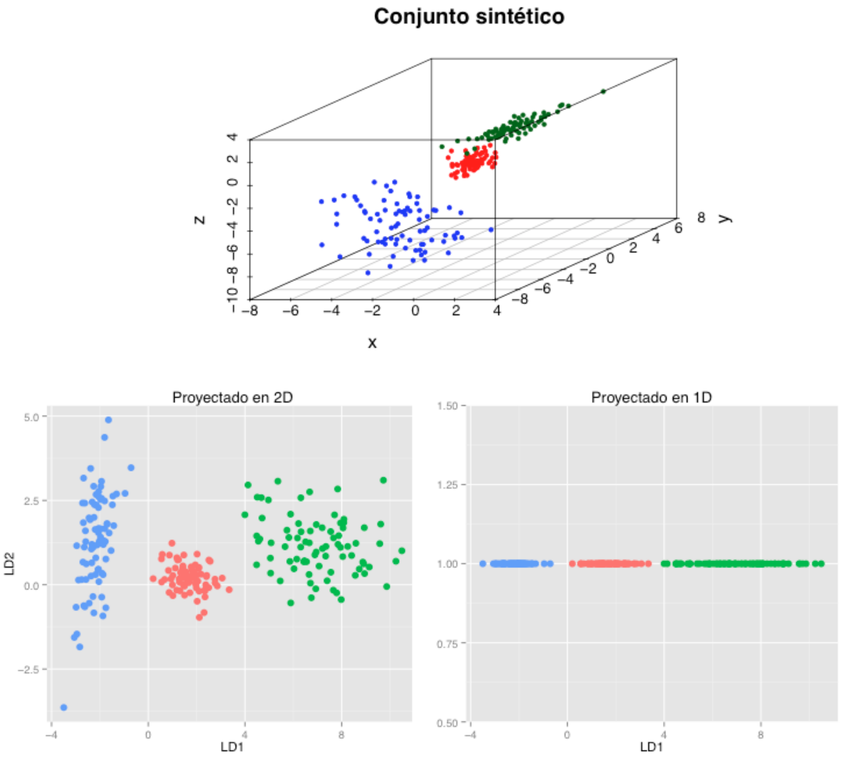

Figure 2: In the above graph it is shows the original data in $\mathbb{R}^3$. In the graph in the lower left it is showing the best projection in $\mathbb{R}^2$ and in the right the best projection in $\mathbb{R}$

2.3.1 Solution when p = 1

The problem (1.9) is quite straightforward when $V \in \mathbb{R}^n$. v is a one dimensional projection:

$$\max_{v\in\mathbb{R}^n} \frac{v^T S_E v}{v^T S_I v} \tag{13}$$

Again, $x_i \in \mathbb{R}^n$ are the original observations with $i = 1,\ldots,N$. Then $w_i \in \mathbb{R}$ are the projected observations with the vector v. This way $w_i = v^T x_i$. It is convenient to define $\widehat{\mu}_k = v^T \mu_k$ y $\widehat{\mu} = v^T \mu$ as the mean of the class and the mean of all the projected observations. The projected Between-class scatter matrix and the projected Within-class scatter matrix are now defined as:

Between-class scatter matrix Φ_E of the projected observations w_i:

$$\Phi_E = \sum_{k=1}^{K} N_k(\widehat{\mu}_k - \widehat{\mu})^2$$

$$\Phi_E = \sum_{k=1}^{K} N_k(v^T \mu_k - v^T \mu)^2$$

$$\Phi_E = \sum_{k=1}^{K} N_k v^T (\mu_k - \mu)(\mu_k - \mu)^T v$$

Because of the distributive property of the matrices, we have that $vAv + vBv = v(A+B)v$, then:

$$\Phi_E = v^T \big[\sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T \big] v \tag{14}$$

Within-class scatter matrix Φ_I of the projected observations w_i:

$$\Phi_I = \sum_{k=1}^{K} \sum_{i \in C_k} (w_i - \widehat{\mu}_k)^2$$

$$\Phi_I = \sum_{k=1}^{K} \sum_{i \in C_k} (v_i^T x_i - v_i^T \mu_k)^2$$

$$\Phi_I = \sum_{k=1}^{K} \sum_{i \in C_k} v^T (x_i - \mu_k)(x_i - \mu_k)^T v$$

Using the distributive property of the matrices:

$$\Phi_I = v^T \big[\sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \big] v \tag{15}$$

The formulas of Φ_I y Φ_E can be expressed in function of the original dispersion matrices S_I y S_E. This way:

$$\Phi_E = f(S_E) = v^T S_E v$$
$$\Phi_I = f(S_I) = v^T S_I v$$

We have that $\Phi_I, \Phi_E \in \mathbb{R}$, then, when maximizing the ratio $\frac{\Phi_E}{\Phi_I}$ with respect to v has as a result a projection that preserves the observations that belong to the same class close to each other, at the same time that the centroids for each class is far away from each other. For the one dimensional case, a closed-form solution can be found. The theory associated with this problem is related with the Generalized Rayleigh Quotient, which, under the constraints for this problem, can be transformed to a normal Rayleigh Quotient.

Proposition 1. *The solution to the maximization of the Rayleigh Quotient:*

$$\max_{v \in \mathbb{R}^n} \frac{v^T A v}{v^T v}$$

when A is symmetric, is obtained when v is the eigenvector associated with the largest eigenvalue of A.

Proof. First, is easy to see that the if we factorize c, the following two maximization problems are equivalents when $c \neq 0$ and $c \in \mathbb{R}$:

$$\max_{v \in \mathbb{R}^n} \frac{v^T A v}{(v^T v)}$$

$$\max_{v \in \mathbb{R}^n} \frac{(cv)^T A (cv)}{(cv)^T (cv)}$$

Without loss of generality, it is assumed that $||v|| = 1$. When $v \in \mathbb{R}^{n \times 1}$, the problem is equivalent to the maximization of the numerator with the equality constraint $v^T v = 1$:

$$\max_{\substack{v \in \mathbb{R}^n \\ v^T v = 1}} v^T A v$$

This problem is more easy and can be solved with the associated Lagrangian function:

$$\mathscr{L}(v, \lambda) = v^T A v - \lambda (v^T v - 1)$$
$$\frac{\partial \mathscr{L}(v, \lambda)}{\partial v} = (A + A^T) v - 2\lambda v = 0$$

Like A is a symmetric matrix, then $A + A^T = 2A$, and the solution is:

$$Av = \lambda v \tag{16}$$

This is the generalized eigenvalue problem. Then, the optimum value is when v is the eigenvector associated with the largest eigenvalue of A.

□

2.3.2 P-dimension generalization

When v is a matrix, the Generalized Rayleigh Quotient can not be written as the Rayleigh Quotient. Then, the solution given in the last section is not useful here. The problem when is projected to a p-dimension is more difficult because it does not have a closed solution, then iterative methods and alternative formulations have been proposed.

The generalization to p dimensions implies that the observations $x_i \in \mathbb{R}^n$ are now projected by the matrix $V = (V_1 | V_2 | ... | V_p)$, this way $w_i = V^T x_i$ with $w_i \in \mathbb{R}^p$ and $V_j \in \mathbb{R}^n$. Equivalently, the matrices Φ_I y Φ_E are defined as:

$$\Phi_E = \sum_{k=1}^{K} N_k || \widehat{\mu}_k - \widehat{\mu} ||_2^2$$

$$\Phi_E = \sum_{k=1}^{K} N_k ||V^T \mu_k - V^T \mu||_2^2$$

$$\Phi_E = \sum_{k=1}^{K} N_k ||V^T (\mu_k - \mu)||_2^2$$

$$\Phi_E = \sum_{k=1}^{K} N_k \left[(V_1^T (\mu_k - \mu))^2 + (V_2^T (\mu_k - \mu))^2 + ... + (V_p^T (\mu_k - \mu))^2 \right] \tag{17}$$

From this expression is important to distinguish that $V_1^T (\mu_k - \mu)$ is an scalar value because $V_1 \in \mathbb{R}^n$ y $(\mu_k - \mu) \in \mathbb{R}^n$. Another equivalent formula that is commonly used because of the algebraic properties is then presented below:

$$\Phi_E = \sum_{k=1}^{K} N_k \,\mathrm{Tr} \left[V^T (\mu_k - \mu)(\mu_k - \mu)^T V \right] \tag{18}$$

To show that these two formulations are equal, we are going to take one class as $k = k_1$. Expanding $(\bullet) = V^T (\mu_1 - \mu)(\mu_1 - \mu)^T V$ then we have a matrix $\mathbb{R}^{p \times p}$ equal to:

$$(\bullet) = \begin{pmatrix} V_1^T (\mu_1 - \mu) \\ V_2^T (\mu_1 - \mu) \\ \vdots \\ V_p^T (\mu_1 - \mu) \end{pmatrix} \begin{pmatrix} (\mu_1 - \mu)^T V_1 & (\mu_1 - \mu)^T V_2 & \dots & (\mu_1 - \mu)^T V_p \end{pmatrix}$$

$$(\bullet) = \begin{pmatrix} V_1^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_1 & \dots & V_1^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_p \\ V_2^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_1 & \dots & V_2^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_p \\ \vdots & \ddots & \vdots \\ V_p^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_1 & \dots & V_p^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_p \end{pmatrix}$$

$$(\bullet) = \begin{pmatrix} (V_1^T (\mu_1 - \mu))^2 & \dots & V_1^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_p \\ V_2^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_1 & \dots & V_2^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_p \\ \vdots & \ddots & \vdots \\ V_p^T (\mu_1 - \mu)(\mu_1 - \mu)^T V_1 & \dots & (V_p^T (\mu_1 - \mu))^2 \end{pmatrix}$$

When taking the diagonal of the above matrix, we have that $\mathrm{Tr}(V^T (\mu_1 - \mu)(\mu_1 - \mu)^T V)$ is equivalent to:

$$\mathrm{Tr}(\bullet) = (V_1^T (\mu_1 - \mu))^2 + (V_2^T (\mu_1 - \mu))^2 + \dots + (V_p^T (\mu_1 - \mu))^2$$

When we sum this expression for all the K classes, and using the linearity property of the trace, $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$, it can be rewritten as follows:

$$\Phi_E = \mathrm{Tr} \sum_{k=1}^{K} N_k \big[V^T(\mu_k - \mu)(\mu_k - \mu)^T V\big]$$

This expression is identical to (1.16). As the last step, V^T and V are factorized over all the terms:

$$\Phi_E = \mathrm{Tr}(V^T \sum_{k=1}^{K} N_k \big[(\mu_k - \mu)(\mu_k - \mu)^T\big] V)$$

or, equivalently in terms of S_E:

$$\Phi_E = \mathrm{Tr}(V^T S_E V) \tag{19}$$

Similarly, it is easy to compute the equivalent expression for the Within-class scatter matrix Φ_I.

$$\Phi_I = \mathrm{Tr}(V^T S_I V) \tag{20}$$

2.4 Solution existence and uniqueness

To show the existence and uniqueness of the solution, the matrices S_I y S_E must have the next characteristics. Let be $A = S_E$ y $B = S_I$, the first condition that is that both matrices are positive definite. The reason behind this constraint is related to the objective function, that is a quotient. As B is the denominator, if $\mathrm{Tr}(V^T BV) = 0$, then the solution will not be defined [5].

T.T. Ngo gives the idea to generalize the work with the positive semi-definite matrices. To achieve this, it is necessary to find when $\mathrm{Tr}(V^T BV)$ is equal to 0. If the B matrix is diagonalized, $B = Q\Lambda_B Q^T$ with Q orthogonal and Λ_B the matrix with the eigenvalues of B in the diagonal entries, then:

$$\mathrm{Tr}(\Lambda_B) = \lambda_{B_1} + \lambda_{B_2} + ... + \lambda_{B_n} \qquad \mathit{with} \qquad \widehat{V} = Q^T V$$

This way $\widehat{V} = (\widehat{V}_1 | \widehat{V}_2 | ... | \widehat{V}_p)$ and each $\widehat{V}_i^T = (\widehat{V}_{i1}, \widehat{V}_{i2}, ..., \widehat{V}_{in})$ is a row vector. The $\widehat{V}^T$ now is equivalent to:

$$\widehat{V}^T = \begin{pmatrix} \widehat{V}_1^T \\ \widehat{V}_2^T \\ \vdots \\ \widehat{V}_p^T \end{pmatrix} = \begin{pmatrix} \widehat{V}_{11} & \widehat{V}_{12} & \cdots & \widehat{V}_{1n} \\ \widehat{V}_{21} & \widehat{V}_{22} & \cdots & \widehat{V}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{V}_{p1} & \widehat{V}_{p2} & \cdots & \widehat{V}_{pn} \end{pmatrix}$$

Then the $\mathrm{Tr}(V^T BV)$ have the next expression:

$$\mathrm{Tr}(V^T BV) = \mathrm{Tr}(V^T Q\Lambda_B Q^T V)$$

$$\mathrm{Tr}(V^TBV) = \mathrm{Tr}(\widehat{V}^T\Lambda_B\widehat{V})$$

$$V^TBV = \begin{pmatrix} \widehat{V}_1^T \\ \widehat{V}_2^T \\ \vdots \\ \widehat{V}_p^T \end{pmatrix} \begin{pmatrix} \lambda_{B_1} & 0 & \dots & 0 \\ 0 & \lambda_{B_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_{B_n} \end{pmatrix} \left(\widehat{V}_1|\widehat{V}_2|\dots|\widehat{V}_p \right)$$

Solving this matrix multiplication and calculating the trace, we have the next terms:

$$\begin{aligned} \mathrm{Tr}(V^TBV) = & \lambda_{B_1}\widehat{V}_{11}^2 & +\lambda_{B_2}\widehat{V}_{12}^2 & +\dots & +\lambda_{B_n}\widehat{V}_{1n}^2 & + \\ & \lambda_{B_1}\widehat{V}_{21}^2 & +\lambda_{B_2}\widehat{V}_{22}^2 & +\dots & +\lambda_{B_n}\widehat{V}_{2n}^2 & + \\ & \vdots & \vdots & \vdots & \vdots & \\ & \lambda_{B_1}\widehat{V}_{p1}^2 & +\lambda_{B_2}\widehat{V}_{p2}^2 & +\dots & +\lambda_{B_n}\widehat{V}_{pn}^2. & \end{aligned}$$

It is easy to see that the above expression have $p \times n$ terms, so it can be expressed with two sums. The first $j = 1, ..., p$ and the second $i = 1, ...n$:

$$\mathrm{Tr}(V^TBV) = \sum_{j=1}^{p}\sum_{i=1}^{n} \lambda_{B_i}\widehat{V}_{ji}^2$$

$$\mathrm{Tr}(V^TBV) = \sum_{i=1}^{n} \lambda_{B_i} \sum_{j=1}^{p} \widehat{V}_{ji}^2 \tag{21}$$

From the last expression, we can separate the sums over i. This way, each element i can be expressed with two factors:

$$(i)\lambda_{B_i} \tag{22}$$

$$(ii)\sum_{j=1}^{p} \widehat{V}_{ji}^2 \tag{23}$$

The idea is that $\mathrm{Tr}(V^TBV)$ is positive, so at least one term should be positive. If (1.21) and (1.22) are different from zero for at least one i, then this condition is achieved. The idea is expressed in the Lemma 1.1:

Lemma 1. *Let B be positive semi-definite and $V \in \mathbb{R}^{n\times p}$. If B has at most $p-1$ eigenvalues equal to 0, then $\mathrm{Tr}(V^TBV) = \mathrm{Tr}(\widehat{V}^T\Lambda_B\widehat{V}) \neq 0$ for any orthogonal matrix V.*

Proof. Let $\widehat{V} = [\widehat{V}_1|...|\widehat{V}_p]$ so that $\widehat{V}^T\widehat{V} = V^TQQ^TV = V^TI_nV = I_p$. This way, is easy to build a matrix $\widehat{V}' \in \mathbb{R}^{p\times p}$ selecting p from the n rows of $\widehat{V}$ so that $\widehat{V}'$ is not singular. $\widehat{V}'$ have the property that does not contains eigenvalues equal to 0; as a consequence, each row and column does not contains the $\widehat{0}$ vector. Then, at least there are p rows of $\widehat{V}$ that $\sum_{j=1}^{p} \widehat{V}_{ji}^2 \neq 0$ for each one of them. On the other hand, the lemma assumes that the matrix B has at most $p-1$ eigenvalues equal to 0, so at least one element of the sum is distinct from zero. □

Analyzing the last result, there are $n-p+1$ positive eigenvalues of B ($\lambda_{B_i} \neq 0$) and p row from $\widehat{V}$ that have norm distinct from 0. When computing the formula (1.20), at least one combination of λ_{B_i} and one of the p rows are both positive. To give an example of this, let C_i with $i = 1, ..., n-p+1$ be the eigenvalues of B and K_j with $j = 1, ..., p$ the norm of the rows of $\widehat{V}$ that are all distinct from 0.

i	λ_{B_i}	$\sum\limits_{j=1}^{p} \widehat{V}_{ji}^2$	$\lambda_{B_i} \sum\limits_{j=1}^{p} \widehat{V}_{ji}^2$
1	C_1	0	0
2	C_2	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-p$	C_{n-p}	0	0
$n-p+1$	C_{n-p+1}	K_p	$C_{n-p+1}K_p$
$n-p+2$	0	K_{p-1}	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	0	K_2	0
n	0	K_1	0

With this combination, we have that at least one term of $\sum\limits_{i=1}^{n} \lambda_{B_i} \sum\limits_{j=1}^{p} \widehat{V}_{ji}^2$ is non-zero, then $\mathrm{Tr}(V^T BV) \neq 0$. Under these circumstances, it is guaranteed that the denominator is greater than 0. Just is necessary to make sure that is less than infinite.

Lemma 2. *Let $U_p = \{V \in \mathbb{R}^{n \times p} | V^T V = I_p\}$ be a compact space with $V = (v_1, v_2, ..., v_p)$*

Proof. We have that U_p is a closed set because it contains all it limit points; on the other hand, U_p is bounded under the norm-2 and the Frobenius-norm:

Taking the norm-2 and the Frobenius norm of V:

$$\begin{aligned}
||V||_2 =& Max\{||V_x||_2 \quad | \quad ||x||_2 = 1\} \\
=& ||V_x||_2^2 \\
=& (Vx)^T (Vx) \\
=& x^T V^T V x \\
=& x^T x = 1 \\
||V||_F =& \sum_{F}^{p} ||v_i|| = p
\end{aligned}$$

Then U_p is closed and bounded, then U_p is a compact space. □

With this, we have that $\mathrm{Tr}(V^TAV)$ have a finite value because all of the entries are finite.

Lemma 3. *Given A and B two positive definite matrices, with B positive semi-definite with range larger than $n-p$, then the quotient* (1.9) *have a maximum value ρ^* [5].*

Proof. Taking the result from the lemma 1.1 we have that $\mathrm{Tr}(V^TBV) \neq 0$; on the other hand, $V \in U_p$ is a compact space. With these two properties, the value (1.9) is different from infinite. Then (1.9) have a maximum value ρ^* that is achieved with the matrix V^{**}. □

References

[1] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification.* John Wiley & Sons, 2012.

[2] Keinosuke Fukunaga. *Introduction to statistical pattern recognition.* Academic press, 2013.

[3] Trevor Hastie, Robert Tibshirani, and Friedman. *The elements of statistical learning*, volume 2. Springer, 2009.

[4] Tom Michael Mitchell. *The discipline of machine learning*. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

[5] Thanh T Ngo, Mohammed Bellalij, and Yousef Saad. The trace ratio optimization problem. *SIAM review*, 54(3):545–569, 2012.

[6] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.