1 Introduction

We have $N$ data points $X_i \in \mathbb{R}^D$. The points belong to $G$ groups. Consider the problem of finding a linear transformation $a$ to one-dimensional space such that the points $X_i a = Z_i \in \mathbb{R}$ are easy to classify. For simplicity we will assume that the $N$ data points $X_i$ have zero mean.

Some notation: given a matrix $A$ indicate with $A_i$ and $A_j$ the $i$-th row and the $j$-th column of $A$, and with $A_{ij}$ the $i,j$-th element of $A$. Moreover:

\begin{align*}
[i] & = g \quad g \text{ is the group of point } i \\
X & = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \in \mathbb{R}^{N \times D} \quad \text{the data} \\
n & = [n_1, \ldots, n_G] \quad \text{number of data points in each group} \\
N & = \text{diag}(n) = G^T G = \begin{bmatrix} n_1, 0, 0, \ldots, 0 \\ 0, n_2, 0, \ldots, 0 \\ \vdots \\ 0, 0, \ldots, 0, n_g \end{bmatrix} \\
G_i & = \delta([i], j) \quad i.e. \quad G = \begin{bmatrix} 1, 0, \ldots, 0 \\ \vdots \\ 0, \ldots, 0, 1 \end{bmatrix} \in \mathbb{R}^{N \times G} \\
M_g & = \frac{1}{n_g} \sum_{[i]=g} X_i \quad \text{Mean of } j\text{-th coordinate in group } g \\
M & = N^{-1} G^T X = \begin{bmatrix} M_1 \\ \vdots \\ M_G \end{bmatrix} \in \mathbb{R}^{G \times D} \quad \text{Matrix collecting the means of each group} \\
A & = U_A L_A V_A^T \quad \text{the singular value decomposition of } A \\
D & = I - GN^{-1} G^T \\
X_0 & = X - GM = DX \quad \text{the data, each referred to group’s mean}
\end{align*}
2 Optimization problem

In order for the points $Z_i$ to be easy to classify one would like to simultaneously maximize the between-clusters distance and minimize the within cluster distance. These quantities may be defined as:

**Between-clusters distance** – Consider the means $M_g$ of each group $g$. One would like to maximize their spread around the overall mean (the origin, since $X$ is zero-mean):

$$B = (GM)^T (GM) = X^T G N^{-1} G^T X$$

(11)

Notice that each mean $M_g$ is counted $n_g$ times in order to reflect the frequency of group $g$.

**Within-clusters distance** – Consider the spread of the points around each group’s center. One would like to minimize:

$$W = X_0^T X_0 = X^T D^T D X$$

(12)

In order to optimize both quantities simultaneously Fisher proposed to maximize their ratio with respect to the transformation $a$:

$$J(a) = \frac{a^T B a}{a^T W a}$$

(13)

Taking the derivative with respect to $a$ and equating to zero:

$$DJ(a) = 2 B a a^T W a - 2 W a a^T B a \overline{(a^T W a)^2} = 0$$

(14)

$$\lambda = \frac{a^T B a}{a^T W a}$$

(15)

$$\Rightarrow B a = \lambda W a \quad a^T W a \neq 0$$

(16)

Therefore in order to find the value of $a$ we need to solve the generalized eigenvector problem $B a = \lambda W a$ subject to $a^T W a \neq 0$.

2.1 Eigenvector problem

Call $W^\dagger$ the generalized inverse of $W$, i.e. the inverse restricted to the subspace where $W$ is nonsingular. Then:

$$B a = \lambda W a \quad a^T W a \neq 0$$

(18)

$$W^\dagger = V_{X_0} L_{X_0}^T V_{X_0}^T$$

(19)

$$\Rightarrow W^\dagger B a = \lambda a$$

(20)

define $(U_{WB}, L_{WB}, V_{WB})$ \text{ SVD}$(W^\dagger B)$

(21)

$$\Rightarrow a = V_{WB}^T$$

(22)
2.2 Alternative approach

An equivalent approach consists in calculating a coordinate transformation \( a = Sb \) such that \( a^T W a = \|b\|^2 \). In this case one may calculate the \( b \) that maximizes the numerator, subject to \( \|b\| = 1 \). One must, however, pay attention to the fact that the solution \( b \) must not be in the null space of \( S \).

From the definition of \( W \) and \( B \) etc.:

\[
W = X_0^T X_0 = V_{X_0} L_{X_0}^2 V_{X_0}^T
\]
\[
S = V_{X_0} L_{X_0}^{-1}
\]
\[
a = Sb
\]
\[
b = S^{-1} a = L_{X_0} V_{X_0}^T a
\]
\[
a^T W a = a^T V_{X_0} L_{X_0}^2 V_{X_0}^T a = b^T b
\]
\[
a^T B a = b^T S^T B S b
\]
\[(U, L, V) = \text{SVD}(S^T B S)\]
\[
\Rightarrow b = V^1
\]
\[
\Rightarrow a = S V^1 = V_{X_0} L_{X_0}^{-1} V^1
\]

3 Code and References

Check out the Matlab function `fisherLD.m` written by Markus Weber. A prize to whoever figures out how the code works.