On maximization of Fisher criterion

Takumi Enomoto

In this document, we tackle with maximization problem of Fisher criterion. First we describe the detailed setting.
Consider a classification problem on 2-class labelling (C_1, C_2). Our feature spece is $\mathbb{R}^d$, and $i \in C_k$ means the label of data $x_i \in \mathbb{R}^d$ is k.
We denote

- $N_k = \#C_k$, the number of data labelled k,
- $\mu_k = \frac{1}{N_k} \sum_{i \in C_k} x_i$, the mean within the class C_k,
- $\Delta\mu = \mu_1 - \mu_2$, the difference between means of each class,
- $S_B = \Delta\mu\Delta\mu^\top$, variance between classes,
- $S_W = \sum_{k=1,2} \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^\top$, variance within each classes.

In addition, we assume S_W is invertible (so positive definite) and $\Delta\mu \neq 0$. If the latter condition is violated, then we cannot use Fisher criterion. Our matter is sepating two distributtions for each classes. In this case, however, the projected distributions overlap each other for any direction.
Under the above notations, we define *Fisher criterion* J by

$$J(w) = \frac{w^T S_B w}{w^T S_W w},$$

where $w \in \mathbb{R}^d \setminus 0$ is a weight in linear classification function. Our problem is the maximization of this function on weight space. Note that the denominator of J is non-zero, because S_W is positive definite.

Remark 1. Trivially J is homogeneous, that is, it holds that $J(\lambda w) = J(w)$ for any $\lambda > 0$, so we can regard it with a (continuous) function on the sphere S^{d-1}. Hence J has the maximum and minimum because S^{d-1} is compact. This ensures the existence of maximizer of J. In the below discussion, we consder J is a function of $\mathbb{R}^d \setminus 0$ in order to use elementaty differential calculations.

For derivation of our maximizer, we consider the gradient of J. Set $f(w) = \frac{1}{2} w^\top S_B w$ and $g(w) = \frac{1}{2} w^\top S_W w$, then $J = \frac{f}{g}$ and

$$\nabla J(w) = \frac{1}{g^2}(g\nabla f - f\nabla g).$$

Hence at its local extreme points it holds that $g\nabla f - f\nabla g = 0$. Now from elementaty calculation, $\nabla f(w) = S_B w$ and $\nabla g(w) = S_W w$, so this condition is written as

$$S_B w = J(w) \cdot S_W w.$$

This means that our maximization problem is reduced to generalized eigenvalue problem for the pair (S_B, S_W) The eigenvalues are the extreme values of J and the eigenvectors are its extreme points, which are candidates of our disired maximizer.
Now we assume that S_W is invertible, so this is written again as

$$S_W^{-1} S_B w = \lambda w.$$

This is an ordinal eigenvalue problem, but we need to take care that $S_W^{-1} S_B$ is not necessarily a normal matrix.
To avoid this issue, we consider some variable change given from the positivity S_W. We can consider the root of S_W and its inverse, so by using variable change $v = {S_W}^{\frac{1}{2}} w$ we have

$${S_W}^{-\frac{1}{2}} S_B {S_W}^{-\frac{1}{2}} v = \lambda v.$$

Here the matrix ${S_W}^{-\frac{1}{2}} S_B {S_W}^{-\frac{1}{2}}$ is positive semidefinite, so we can solve the eigenvalue problem for this matrix easily.

Remark 2. Same discussion can be applied to generalized Reighley quotient $\frac{x^\top A x}{x^\top B x}$ in the case that A and B are positive and B is invertible.

In our case, recall $S_B = \Delta\mu\Delta\mu^\top$. Set $\nu = {S_W}^{-\frac{1}{2}} \Delta\mu$, then we have

$${S_W}^{-\frac{1}{2}} S_B {S_W}^{-\frac{1}{2}} = \nu\nu^\top.$$

We can find the eigenvalues and the eigenvectors of $\nu\nu^T$ directly as follows:

- the eigenvalues are $\{\nu^\top \nu, 0\}$,
- the eigenvector for $\nu^\top \nu$ is ν,
- the eigenvectors for 0 are the nonzero vectors orthogonal to ν.

Reverse the variable by $w = {S_W}^{-\frac{1}{2}} v$, we solve the eigenvalue problem for $S_W^{-1} S_B$, which is original one, as follows:

- the eigenvalues are $\{J_0 = \Delta\mu^\top {S_W}^{-1} \Delta\mu, 0\}$,
- the eigenvector for J_0 is ${S_W}^{-\frac{1}{2}} \nu = {S_W}^{-1} \Delta\mu$.

Note that S_W is not orthogonal so we have no orthogonal eigenspace decomposition of $S_W^{-1} S_B$. Recall the set of this eigenvalues are coinside with the set of the extreme values of Fisher criterion J. This set is consisted from $\{J_0 > 0, 0\}$, so J_0 is the maximum of J at ${S_W}^{-1} \Delta\mu$.
As a wrap-up, we state our conclusion:

Proposition 1. *The maximizer of Fisher criterion J is ${S_W}^{-1} \Delta\mu$.*