# Pattern Recognition with Applications to Biomedical Images
# Independent Study in Mathematics – CSUN Spring 2006
# Fisher

P. Perona & K. F. Stevenson

January 22, 2007

## 1 Introduction

Consider $\mathcal{N}$ data points $X_i \in \mathbb{R}^D$. The points belong to $\mathcal{G}$ classes. Look for a linear transformation $\alpha : \mathbb{R}^D \to \mathbb{R}$ (i.e. to a one-dimensional space) such that the points $\alpha(X_i) = X_i a = Z_i \in \mathbb{R}$ are easy to classify.

(RMK: Notice that this is the opposite order from our usual notation: Normally we have a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ and $L(\vec{x}) = [L][\vec{x}]$ (i.e. we multiply on the left by the $m \times n$ matrix of $[L]$ representing $L$ wrt the standard basis). For notational reasons we are taking the transpose of EVERYTHING! So, $\alpha(X_i) = ([\alpha][X_i])^T = [X_i]^T[\alpha]^T$. Thus $a = [\alpha]^T$. As $[\alpha]$ is a $1 \times D$ matrix, then $a$ is a $D \times 1$ matrix. Thus $aX_i$ makes sense $((1 \times D)(D \times 1))$ and it is a real number.)

Assume that the $\mathcal{N}$ data points $X_i$ have zero mean. (What does this mean? Want $\frac{1}{\mathcal{N}} \sum_1^{\mathcal{N}} X_i = \vec{0}$.)

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 $A_i^j$ the $i, j$-th element of $A$. Moreover:

$$
\begin{aligned}
[i] &= g \qquad g \text{ is the number of the class of the vector } X_i && (1)\\
X &= \begin{bmatrix} X_1 \\ \dots \\ X_{\mathcal{N}} \end{bmatrix} \in \mathbb{R}^{\mathcal{N}\times D} \qquad \text{the data AS ROWS} && (2)\\
n &= \begin{bmatrix} n_1, \dots, n_{\mathcal{G}} \end{bmatrix} \qquad \text{number of data points in each group} && (3)\\
N &= \operatorname{diag}(n) = \begin{bmatrix} n_1, 0, 0, \dots, 0 \\ 0, n_2, 0, \dots, 0 \\ \dots \\ 0, 0, \dots, 0, n_{\mathcal{G}} \end{bmatrix} && (4)\\
G_i^j &= \delta([i], j) \qquad i.e. \qquad G = [G_i^j] && (5)\\
G &= \begin{bmatrix} 1, 0, \dots, 0 \\ 1, 0, \dots, 0 \\ \dots \\ 1, 0, \dots, 0 \\ \dots \\ 0, \dots, 0, 1 \\ \dots \\ 0, \dots, 0, 1 \\ \dots \\ \dots \\ 0, \dots, 0, 1 \end{bmatrix} \in \mathbb{R}^{\mathcal{N}\times\mathcal{G}} \text{ if we reorder the } X_i \text{ so that the first } n_1 \text{ are in class 1, etc} && (6)\\
N &= \operatorname{diag}(n) = G^T G \qquad \text{WHY?} && (7)\\
M_g &= \frac{1}{n_g} \sum_{[i]=g} X_i \qquad \text{Mean of the class } g && (8)\\
M &= \begin{bmatrix} M_1 \\ \dots \\ M_{\mathcal{G}} \end{bmatrix} \in \mathbb{R}^{\mathcal{G}\times D} \qquad \text{Matrix collecting the means of each class} && (9)\\
M &= N^{-1} G^T X \qquad \text{WHY?} && (10)\\
GM &= \text{ WHAT?} && (11)\\
X_0 &= X - GM = DX \qquad \text{the data, each referred to group's mean} && (12)\\
D &= I - G N^{-1} G^T && (13)\\
X_0 &= DX \qquad \text{WHY?} && (14)\\
A &= U_A L_A V_A^T \qquad \text{the singular value decomposition of A, IGNORE NOW.} && (15)\\
& && (16)
\end{aligned}
$$

# 2 Optimization problem

In order for the points $Z_i = aX_i$ to be easy to classify one would like to simultaneously maximize the *between-clusters distance* and minimize the *within cluster distance* inside that

projected space. These quantities may be defined as:

**Between-clusters distance in projected space** – Consider the means $M_g$ of each class $g$. One would like to maximize their spread around the overall mean (the origin, since $X$ is zero-mean). Let $m_g$ be the mean of class $g$ after projecting via $\alpha$. Let

$$\text{spread}(M) = \sum_{g=1}^{\mathcal{G}} m_g^2.$$

Show that

$$\text{spread}(M) = (a^T M^T)(Ma).$$

If we wish to weigh each mean according to the number of points in the class then we take:

$$\text{spread}(GM) = \sum_{g=1}^{\mathcal{G}} m_g^2 n_g^2.$$

Show that

$$\text{spread}(M) = (a^T M^T G^T)(GMa).$$

Now let:

$$B \quad = \quad (GM)^T(GM) = X^T G N^{-1} G^T X \tag{17}$$

Thus

$$\text{spread}(GM) = a^T B a.$$

**Within-clusters distance** – Consider the spread of the points around each group's center. The distance across class $g$ within the projection is

$$\text{classspread}(g) = \sum_{[i]=g} ||X_i a - M_g a||^2 = \sum_{[i]=g} ||X_{0,i} a||^2.$$

We want that for each $g$ this is SMALL, so we can look at all of them at once:

$$\text{classspread} = \sum_{i}^{\mathcal{N}} ||X_i a - M_{[i]} a||^2 = \sum_{[i]=g} ||X_{0,i} a||^2.$$

Show that

$$\text{classspread} = (a^T X_0^T)(X_0 a).$$

Now let:

$$W \quad = \quad X_0^T X_0 = X^T D^T D X \tag{18}$$

Thus

$$\text{classspread} = a^T W a.$$

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

$$J(a) = \frac{a^T B a}{a^T W a} \tag{19}$$

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

$$DJ(a) = \frac{2Baa^TWa - 2Waa^TBa}{(a^TWa)^2} = 0 \tag{20}$$

$$\text{define} \quad \lambda \doteq \frac{a^T B a}{a^T W a} \tag{21}$$

$$\Rightarrow Ba = \lambda W a \qquad a^T W a \neq 0 \tag{22}$$

$$\tag{23}$$

Therefore in order to find the value of $a$ we need to solve the *generalized eigenvector problem* $Ba = \lambda W a$ 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:

$$Ba = \lambda W a \qquad a^T W a \neq 0 \tag{24}$$

$$W^\dagger = V_{X_0} L_{X_0}^{\dagger 2} V_{X_0}^T \tag{25}$$

$$\Rightarrow W^\dagger B a = \lambda a \tag{26}$$

$$\text{define}(U_{WB}, L_{WB}, V_{WB}) \doteq \text{SVD}(W^\dagger B) \tag{27}$$

$$\Rightarrow a = V_{WB}^1 \tag{28}$$

# 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.

You will find Fisher linear discriminants discussed in B. Ripley *Pattern recognition and neural networks*, Cambridge University Press, 1996 (SFL library, call n. **qa 76.87 r56 1996**).