Fisher 判别分析中投影空间的优化问题

王锦贵 丑纪范
(黑龙江省气象台) (兰 州 大 学)
郭秉荣
(武 汉 大 学)

提 要

本文对 Fisher 意义下的判别分析中泛函值 $\lambda = \boldsymbol{C}^T \cdot B \cdot \boldsymbol{C}/\boldsymbol{C}^T \cdot W \cdot \boldsymbol{C}$ 与判别效果（指错判个数）之间的非同一性进行了讨论，指出在 m 个变量构成的 m 维空间中，由广义特征问题 $(B - \lambda W)\boldsymbol{C} = 0$ 的特征向量所决定的投影空间不一定是判别效果最优的空间，即，泛函 λ 取极值与判别效果取极值不是等价问题。

本文构造了一个与 Fisher 判别效果相一致，但又无解析形式的泛函：

$$I = I(\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_r)$$

它建立在线性函数集合之上，I 取极值与判别效果取极值是等价的。

文章提出了可采用最优化方法根据泛函 λ 值取极值的解求得泛函 I 值取极值的解，从而解决了 Fisher 判别分析中投影空间的优化问题。实例计算表明，在优化的投影空间中进行 Bayes 准则判别其效果要优于由广义特征问题 $(B - \lambda W)\boldsymbol{C} = 0$ 所决定的投影空间中所进行的判别。

一、引 言

Fisher 在本世纪三十年代提出的判别分析方法在许多学科中有着广泛的应用，在天气预报中，这一方法也用得相当普遍[1,2]。Fisher 判别的中心思想是把高维空间的点集合向低维空间投影，其投影方向是由广义特征问题 $(B - \lambda W)\boldsymbol{C} = 0$ 的特征向量决定的，其后的判别便在投影空间中利用 Bayes 准则进行。

近年来，判别分析的理论和方法取得了一定的进展，但是也还存在着一些基本问题，例如在逐步判别中"判别效果"实际上是指 Wilks 量 $U = |W|/|T|$，并且 U 越小则认为判别效果越好。但在实际计算中，人们却关心样本错分个数，错分个数越少则认为效果越好。应当指出，U 值与错分个数虽有密切关系，但它们又不是一回事，即具有非同一性。

本文讨论了在 Fisher 意义下的判别分析中泛函 $\lambda = \boldsymbol{C}^T \cdot B \cdot \boldsymbol{C}/\boldsymbol{C}^T \cdot W \cdot \boldsymbol{C}$ 与判别效果之间的非同一性，并且提出了 Fisher 判别分析中投影空间的优化问题。

1983 年 3 月 1 日收到，1984 年 1 月 23 日收到修改稿。

二、判别效果最优的投影空间形成原理

Fisher 意义下的判别分析在许多著作中都有详细介绍[3,4]，为了便于问题的讨论和比较，现将 Fisher 判别的基本原理和计算公式简述如下：

由 m 个表示事物特征的变量 $(x_1, x_2, \cdots, x_m)$ 构成 m 维空间，每个事物视为 R^m 空间中的一个点，而事物的全体形成 R^m 中的一个集合，把这个集合向一维空间投影，也就是作线性变换

$$Z_{gk} = \sum_{i=1}^{m} C_i x_{igk} \tag{1}$$

$$g = 1, 2, \cdots, G \qquad (分G类)$$

$$k = 1, 2, \cdots, n_g$$

$$n_1 + n_2 + \cdots + n_G = N \qquad (N个样本)$$

记 $\boldsymbol{C} = (C_1, C_2, \cdots, C_m)^T$.

所谓要使事物在投影空间中有利于分类，也就是使泛函 $\lambda = b(z)/w(z)$ 取极大值.

$$w(z) = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (Z_{gk} - \bar{Z}_g)^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} C_i C_j \tag{2}$$

$$b(z) = \sum_{g=1}^{G} n_g (\bar{Z}_g - \bar{Z})^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} b_{ij} C_i C_j \tag{3}$$

$$w_{ij} = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (x_{igk} - \bar{x}_{ig})(x_{jgk} - \bar{x}_{jg}) \tag{4}$$

$$b_{ij} = \sum_{g=1}^{G} n_g (\bar{x}_{ig} - \bar{x}_i)(\bar{x}_{jg} - \bar{x}_j) \tag{5}$$

显然，$\boldsymbol{C}$ 应该满足方程

$$\partial\lambda/\partial\boldsymbol{C} = 0 \tag{6}$$

即

$$(B - \lambda W)\boldsymbol{C} = 0 \tag{7}$$

$$B = (b_{ij})_{m\times m} \qquad W = (w_{ij})_{m\times m}$$

于是，求投影方向 $\boldsymbol{C}$ 的问题就变成了对方程(7)求广义特征值及其相应的特征向量的问题. 求解出的特征值大小也就表明它对应的特征向量(投影方向)对 G 类母体判别分类的能力大小.

(7)式有 R 个非零特征值 $(R \leqslant \min(m, (G-1))$，对应的 R 个特征向量为：

$$\boldsymbol{C}_i^0 = (C_{1i}^0, C_{2i}^0, \cdots, C_{mi}^0)^T$$

$$i = 1, 2, \cdots, R$$

用前 r 个特征向量 $\boldsymbol{C}_1^0, \boldsymbol{C}_2^0, \cdots, \boldsymbol{C}_r^0$ 构成的投影空间记为 $H_0^r (1 \leqslant r \leqslant R)$. 在 H_0^r 中判别用 Bayes 准则. 假定各类母体均服从正态分布

$$f_g(\boldsymbol{Z}) = |\Sigma_g^{-1}|/(2\pi)^{r/2} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{Z} - \bar{\boldsymbol{Z}}_g)^T \Sigma_g^{-1} (\boldsymbol{Z} - \bar{\boldsymbol{Z}}_g)\right] \tag{8}$$

均值向量与协方差矩阵均用样本估计，即

$$\bar{\boldsymbol{Z}} = (Z_1, Z_2, \cdots, Z_r)^T$$

$$\bar{\boldsymbol{Z}}_g = (\bar{\boldsymbol{Z}}_{1g}, \bar{\boldsymbol{Z}}_{2g}, \cdots, \bar{\boldsymbol{Z}}_{rg})^T$$

$$\bar{\boldsymbol{Z}}_{ig} = \frac{1}{n_g}\sum_{k=1}^{n_g} Z_{igk} = \frac{1}{n_g}\sum_{k=1}^{n_g}\sum_{j=1}^{m} C_{ij}x_{jgk}$$

$$i = 1, 2, \cdots, r \tag{9}$$

$$\Sigma_g = \begin{bmatrix} \sigma_{11}\sigma_{12}\cdots\sigma_{1r} \\ \sigma_{21}\sigma_{22}\cdots\sigma_{2r} \\ \vdots\ \ \vdots\ \ \ \vdots \\ \sigma_{r1}\sigma_{r2}\cdots\sigma_{rr} \end{bmatrix}_g \tag{10}$$

$$(\sigma_{ij})_g = \frac{1}{n_g - 1}\sum_{k=1}^{n_g}(Z_{igk} - \bar{Z}_{ig})(Z_{igk} - \bar{Z}_{ig})$$

$$i, j = 1, 2, \cdots, r \tag{11}$$

判别分类直接依赖于后验概率

$$P(g/\boldsymbol{Z}) = q_g \cdot f_g(\boldsymbol{Z}) \Big/ \sum_{n=1}^{G} q_n \cdot f_n(\boldsymbol{Z}) \tag{12}$$

$q_n(n = 1, 2, \cdots, G)$ 为先验概率．对(12)式两边取对数，并略去与 g 无关的项，导到 G 个非线性判别函数(下式已用观测频率 n_g/N 来代替先验概率 q_g)

$$Y_g(\boldsymbol{Z}) = \ln n_g + \frac{1}{2}\ln|\Sigma_g^{-1}| - \frac{1}{2}\boldsymbol{Z}^T \cdot \Sigma_g^{-1}\boldsymbol{Z}$$

$$+ \bar{\boldsymbol{Z}}^T \cdot \Sigma_g^{-1}\bar{\boldsymbol{Z}}_g - \frac{1}{2}\bar{\boldsymbol{Z}}_g^T \cdot \Sigma_g^{-1}\bar{\boldsymbol{Z}}_g \tag{13}$$

若 $Y_{g^*}(\boldsymbol{Z}) = \max_{1\leqslant g\leqslant G}\{Y_g(\boldsymbol{Z})\}$，则把个体划归第 g^* 类．

下面我们讨论投影空间的优化问题．在 Fisher 判别中，H_0^r 是根据泛函 λ 达极大值的要求得到的，在这样的投影空间中进行判别分析时，错判个数不一定最少，所谓投影空间的优化问题就是要在 R^m 空间中寻找新的投影空间 H_*^r，r 个投影方向记为 $(\boldsymbol{C}_1^*, \boldsymbol{C}_2^*, \cdots, \boldsymbol{C}_r^*)$，使在其中的判别效果最优．为此，构造一个新泛函

$$I = \sum_{g=1}^{G}\sum_{k=1}^{n_g} L(g^*/g)_k$$

$$L(g^*/g)_k = 0 \quad \text{当 } g = g^*$$

$$L(g^*/g)_k = 1 \quad \text{当 } g \neq g^*$$

显然，对事物的一个个体 $\boldsymbol{X} = (x_1, x_2, \cdots, x_m)^T$ 在 R^m 中原属第 g 类，而经过某种线性变换后映射到投影空间中为 $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_r)^T$，根据 G 个判别函数判为第 g^* 类．对于确定的线性变换就能得到确定的投影空间，也就能确定在 R^m 中原属第 g 类的个体在投影空间中判别分类的类别 g^*，因而也就能确定 $L(g^*/g)_k$ 是取 0 或 1，此理推广到事物的全体，就得到 I 是定义在线性函数(线性变换)集合上的一个泛函：

$$I = I(\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_r) \tag{14}$$

问题归结为求解使(14)式取极小的 $C_1^*, C_2^*, \cdots, C_r^*$. 由 $C_1^*, C_2^*, \cdots, C_r^*$ 构成的 H_*^r 被称为优化的投影空间. 这也等价于在 $m \times r$ 维的向量空间中寻求 I 值的极小点，或者是认为解如下的非线性方程组：

$$\partial I / \partial C_i = 0 \quad i = 1, 2, \cdots, r \tag{15}$$

由上述可以看出，经典的 Fisher 意义下的判别是在 H_*^r 的判别空间中进行，而优化的 Fisher 意义下的判别是在 H_*^r 的判别空间中进行，两者的判别准则都基于 Bayes 原理.

在最优化方法中，最小二乘法、共轭斜量法和变尺度方法都是寻求泛函取极值的方法，但在计算过程中都需要计算导数，这就需要知道 I 的解析表达式，而 I 的解析表达式是无法描写的，故只能采用最优化方法中的不需计算导数只需计算泛函值的直接方法——单纯形法求得 I 取极值的解[5].

单纯形法的基本思想就是在不计算导数的情况下，先算出若干个点的泛函值，从它们的分布可以看出泛函变化的大概趋势，为寻求它的极值作参考.

三、利用单纯形法求 H_*^r 的计算过程

I 是 $m \times r$ 元函数，它的极值点为 $m \times r$ 维空间 $\phi^{m\times r}$ 中一个向量，该向量的分量就构成了 $C_1^*, C_2^*, \cdots, C_r^*$.

记

$$\begin{aligned} C_0^\dagger &= (C_1^{0T}, C_2^0, \cdots, C_r^0)^T \\ &= (C_{11}^0, C_{21}^0, \cdots C_{m1}^0, C_{12}^0, C_{22}^0, \\ &\quad \cdots C_{m2}^0, \cdots, C_{1r}^0, C_{2r}^0, \cdots, C_{mr}^0)^T \\ &= (C_1^0, C_2^0, \cdots, C_{m\times r}^0)^T \end{aligned}$$

初始单纯形必须由 $\phi^{m\times r}$ 空间中的 $m \times r + 1$ 个点构成，且取 $C_1^\dagger - C_0^\dagger$, $C_2^\dagger - C_0^\dagger$, $\cdots$, $C_{m\times r}^\dagger - C_0^\dagger$ 为线性独立的向量，否则我们取极值点的范围便局限在一个低维空间中.

$$\begin{aligned} C_j^\dagger &= C_0^\dagger + (\delta - 1) C_j^0 e_j \\ j &= 1, 2, \cdots, m \times r \end{aligned}$$

δ 为初始步长因子，e_j 为 $\phi^{m\times r}$ 空间中第 j 个单位坐标向量.

定义：

$$I_h = I(C_h^\dagger) = \max_{0<j<m\times r} \{I_j\}$$

$$I_l = I(C_l^\dagger) = \min_{0<j<m\times r} \{I_j\}$$

$$I_g = I(C_g^\dagger) = \max_{\substack{0<j<m\times r \\ j\neq h}} \{I_j\}$$

单纯形法求极值点包括下述四种运算过程

(1) 反射

先求单纯形除 $C_h^\dagger$ 外的中心点

$$C_c^{\phi} = \frac{1}{m \times r}\left(\sum_{j=0}^{m\times r} C_j^{\phi} - C_h^{\phi}\right)$$

再求 C_h^{ϕ} 关于 C_c^{ϕ} 的反射点

$$C_r^{\phi} = 2C_c^{\phi} - C_h^{\phi}$$

(2) 延伸

若 $I_r = I(C_r^{\phi}) < I_g$，且当

$$(1-\mu)I_h + \mu I_r < I_l$$

时，则将向量 $C_r^{\phi} - C_h^{\phi}$ 延伸得

$$C_e^{\phi} = (1-\mu)C_h^{\phi} + \mu C_r^{\phi}$$

若 $I_e < I_r$，则以 C_e^{ϕ} 代替 C_h^{ϕ}，否则以 C_r^{ϕ} 代替 C_h^{ϕ}. 式中的 μ 为延伸因子，$\mu > 1$.

(3) 收缩

若 $I_r > I_g$，则将向量 $C_r^{\phi} - C_h^{\phi}$ 收缩. 令 $C_r^{\phi} = (1-\theta)C_h^{\phi} + \theta C_r^{\phi}$ (θ'为压缩因

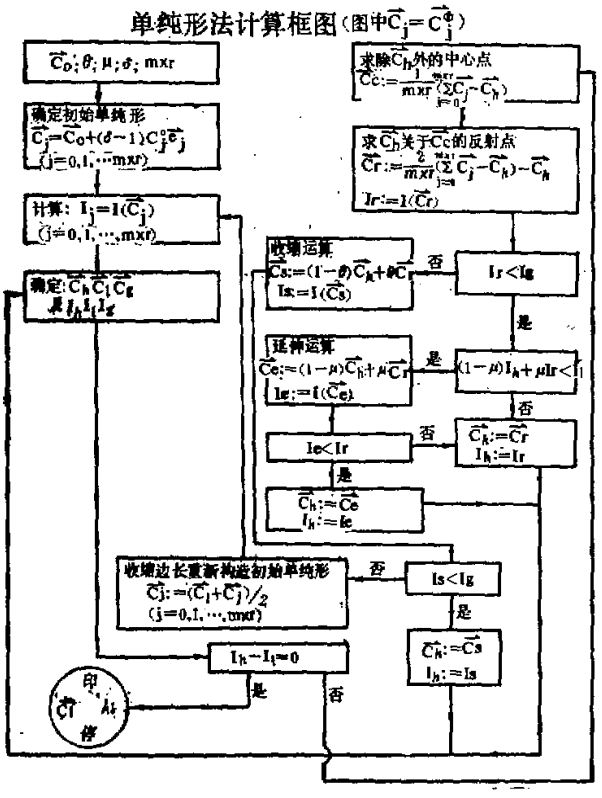

图 1 单纯形法计算框图

图中 $C_j = C_j^{\phi}$

子，$0 < \theta < 1$，$\theta \neq 0.5$）. 若 $I_r < I_g$，则以 C_r^ϕ 代替 C_k^ϕ.

(4) 收缩边长

若 $I_r \geq I_g$，则收缩初始单纯形边长，令

$$C_k^\phi = \frac{1}{2}(C_j^\phi + C_l^\phi) \quad k, j = 1, 2, \cdots, m \times r$$

构成新的单纯形重新开始. 计算流程见框图.(图 1).

说明：初始步长因子 $\delta \neq 1$，根据经验一般来说，$\delta = 0.5 \sim 2.0$；$\theta = 0.25$ 或 0.75；取 $\mu = 1.2 \sim 2.0$.

四、实例计算

预报因子资料(1954—1979)

x_1——亚欧地区 3 月份平均纬向环流指数；

x_2——亚欧地区 1 月份平均经向环流指数；

x_3——北半球 500 百帕 6 月份平均极涡中心位置；

x_4——北半球 500 百帕 3 月份平均极涡中心强度

x_5——500 百帕 2 月份平均东亚大槽平均位置预报量资料(1955—1980)

y——甘肃省 5 月和 6 月平均降水量

此例中，y 分为 3 类，总体样本 $N = 26$，其中 $n_1 = 8$，$n_2 = 10$，$n_3 = 8$.

我们先计算在 H_0^1 中错判个数

$$B = \begin{bmatrix} 0.44174 & 0.20227 & 0.73982 & -7.88031 & 5.21027 \\ & 0.09470 & -0.37327 & -4.01942 & 3.27731 \\ & & 244.69615 & 127.39038 & -296.13846 \\ & & & 221.76346 & -268.99615 \\ & & & & 443.21538 \end{bmatrix}$$

$$W = \begin{bmatrix} 0.91560 & -0.20625 & -4.09675 & 1.67800 & 1.13050 \\ & 0.23800 & 6.20250 & 1.04250 & -3.12000 \\ & & 613.15000 & 66.22500 & -48.40000 \\ & & & 561.77500 & -102.85000 \\ & & & & 473.90000 \end{bmatrix}$$

于是求得 $\lambda_1^0 = 3.44262$，$\lambda_2^0 = 0.83710$.

$$C_1^0 = (0.70553, 2.21206, -0.02142, -0.01279, 0.02830)^T$$

$$C_2^0 = (-0.71302, -0.39290, -0.02346, 0.00970, 0.02430)^T$$

由计算知道，在 H_0^1 中错判 8 个，在 H_0^2 中错判 3 个. 在应用最优化方法求得的 H_*^1 中，$\lambda_1^* = 3.13374$

$$C_1^* = (0.61610, 2.92975, -0.02217, -0.01152, 0.01539)^T$$

错判 2 个. 这里的 H_*^1 是当 $\delta = 0.5$、$\mu = 1.8$、$\theta = 0.75$ 时求得的. 在 H_*^2 中，

$$\lambda_1^* = 2.93094,\quad \lambda_2^* = 0.54249$$

$$\boldsymbol{C}_1^* = (0.57116,\ 2.49047,\ -0.02279,\ -0.01036,\ 0.02291)^T$$

$$\boldsymbol{C}_2^* = (-0.82619,\ -0.43594,\ -0.02157,\ 0.00913,\ 0.02594)^T$$

错判1个. 这里的 H_*^2 是当 $\delta = 0.5$、$\mu = 1.4$、$\theta = 0.25$ 时求得的.

从上面的计算结果看出,对由 x_1—x_5 构成的空间而言,在 H_*^1 中的错判个数比在 H_0^1 中减少6个;在 H_*^2 中错判个数比在 H_0^2 中减少2个,同时还看出,与 H_*^r 对应的 λ_r^* 值均比与 H_0^r 对应的 λ_r^0 值小。

五、关于泛函 λ 与判别效果的讨论

类似于逐步判别分析中 U 值与判别效果之间存在着非同一性问题, Fisher 意义下的判别分析中 λ 值与判别效果之间也存在着非同一性问题. 从理论上分析看, 其原因是泛函 λ 要依赖于 W 和 B 两矩阵,但不依赖于先验概率,而判别分类依赖于 Σ_g 矩阵和先验概率. 下面的计算也表明,不能认为 λ 值大则相应的投影空间中的判别效果就好,而 λ 值小则相应的投影空间中的判别效果就差.

在这里,我们仍选用实例计算中的资料,设构成 R_1^4 的变量为 x_1, x_2, x_3, x_4; 而构成 R_2^4 的变量为 x_1, x_2, x_3, x_5. 在两空间中被判别的对象同为 y.

由方程(7)求得 R_1^4 的最大特征值 $\lambda = 2.81020$, 在相应的投影空间中错判个数为3个,而 R_2^4 中最大特征值 $\lambda = 3.29098$, 相应的投影空间中的错判个数为5个,因此,根据最大特征值的大小来筛选判别因子的方法在实际应用中是有局限性的.

根据实例计算和上述讨论,无论是在由相同的变量构成的空间中,或是在由不同的变量构成的空间中, 对应于 λ 值大的投影空间中的判别效果不一定比对应 λ 值小的投影空间中的判别效果好.

六、问题讨论

(1) 我们在比较 H_0^r 和 H_*^r 中的判别效果时, 特别注意到 H_0^r 上 λ 值的判别显著性, 在本文第四节的实例中, $\lambda_1^0 = 3.44262$;根据显著性检验公式计算

$$\left[N - 1 - \frac{1}{2}(m + G)\right] \cdot \ln(1 + \lambda) = 31.31613$$

而 $\chi^2_{0.01}(6) = 16.82$,这说明投影空间 H_0^1 是具有判别的显著性的.

(2) 计算表明, H_*^r 可以是不唯一的, 在第四节的例子中, 当 $\delta = 2.0$, $\mu = 1.4$, $\theta = 0.25$ 时,我们还可以求得错判个数也为2的投影空间 H_*^1

$$\boldsymbol{C}_1^* = (1.11127,\ 3.43444,\ -0.02407,\ -0.018436,\ 0.018706)^T$$

(3) 最优化原理在判别分析中的应用为建立以错判个数为依据的判别方法提供了基础,特别是采用单纯形法,避免要求知道 $I = I(\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_r)$ 的具体函数形式,但必须强调指出,以 Fisher 判别分析中的 H_0^r 作为求 H_*^r 的初值是有重要意义的,因为只有这样才能保证在 H_*^r 中优于(或等于)在 H_0^r 中的判别效果. 最优化原理也可以推广运用到

逐步判别、最小方差判别等分析中去.

(4) 求得优化的投影空间，提高判别效果是有实际意义的，但求 H_* 的计算量比求 H_0 的要大，特别当 $m \times r$ 比较大时更为明显，这给计算带来一定的局限. 另外，在求解 H_* 的过程中，优化方法的效果还受到参数 δ、θ 和 μ 的影响. 在实际计算过程中，这些参数的取值往往带有一定的经验性[5].

本文曾请 李麦村，史文恩两位同志审阅，承蒙提出了宝贵意见，特此致谢.

参　考　文　献

[1] 王宗皓，李麦村，天气预报中的概率统计方法，科学出版社，1974.
[2] Miler, R. G., Statistical prodiction by discriminant analysis, *meteorological monographs*, Vol. 4, No. 25, 1962.
[3] 中国科学院计算中心概率统计组编著，概率统计计算，科学出版社，1979.
[4] 中央气象局气象科学研究所编，数理统计天气预报文集，农业出版社，1979.
[5] 南京大学数学系计算数学专业编，最优化方法，科学出版社，1979.

THE OPTIMIZATION OF PROJECT SPACE IN THE FISHER DISCRIMINANT ANALYSIS

Wang Jingui
(Meteorological Observatory of Heilongjiang Province)
Chou Jifan
(Lanzhou University)
Guo Bingyong
(Wuhan University)

Abstract

In this paper authors has studied the non-consistency between functional value $\lambda = \mathbf{C}^T \cdot B \cdot \mathbf{C} / \mathbf{C}^T \cdot W \cdot \mathbf{C}$ and the discriminant effect (misjudging number), pointing out that the project space determined by the eigenvectors of the generalized eigenvalue problem $(B-\lambda w)\ C=0$ in the space consisting of m variables is uncertain of the optimization space in discriminant effect. In other words, the maximal value of functional λ is not equivalent to the minimal value of the misjudging number.

We has constructed a functional $I=I(C_1, C_2, \cdots, C_r)$ which is consistent with Fisher discriminant effect but has no analytic form. It is set up based on a set of linear functions. In this case the minimal value of I is equivalent to the discriminant effect.

Finally, we have found that the solution of minimal value of functional I may be obtained according to the solution of the maximal value of functional λ using the optimum method. Thus, the optimization of project space in Fisher discriminant analysis has been solved. The calculations of cases have shown that the effect of discriminant analysis, which is carried out in the optimum project space by Bayes principle is better than that of the discriminant which is carried out in the project space determined by the generalized eigenvalue problem $(B-\lambda w)C=0$.