INFORMATICA, 1992, Vol.3, No.3, 348–359

# BAYES ESTIMATORS OF THE DISCRIMINANT SCORES IN THE STATISTICAL GROUP CLASSIFICATION PROBLEMS

Mirosław KRZYŚKO

Institute of Mathematics
Adam Mickiewicz University
60–769 Poznań, Poland

**Abstract.** The rules of classification of the group of $N$ independent observations into one of $k$ normal populations are considered. In the case when parameters are not known, Bayes estimators of the discriminant scores are suggested.

**Key words:** statistical group classification, Bayes estimation.

**1. Introduction.** In the standard classification problem one wishes to assign an individual to one of $k$ populations on the basis of a vector $x$ of observed characteristics for that individual.

Abusev and Lumelsky (1980, 1987) were taking into consideration the problem of classification not a single observation but the group $X_0 = \{x_{01}, \ldots, x_{0N_0}\}$ of $N_0$ independent observations. Such problems arise in a medical and technical diagnostic, in particular in epidemiology and quality control.

In this paper we construct the rules of classification of the group $X_0 = \{x_{01}, \ldots, x_{0N_0}\}$ of $N_0$ independent observations into one of $k$ normal populations $N_p(\mu_i, \Sigma_i)$, $i = 1, \ldots, k$. These rules are based on knowledge of the discriminants scores $\ln[q_i f_i(\overline{x}_0, A_0)]$, where $f_i(\overline{x}_0, A_0)$ is the density function of the joint distribution of the sufficient statistics

$$\overline{x}_0 = N_0^{-1} \sum_{j=1}^{N_0} x_{0j}, \qquad A_0 = \sum_{j=1}^{N_0} (x_{0j} - \overline{x}_0)(x_{0j} - \overline{x}_0)',$$

and $q_i$ is the prior probability that the classifying sample $X_0$ was drawn from the $i$-th population, $i = 1, \ldots, k$.

This knowledge is usually unavailable. Recently the author (Krzyśko, 1991) constructed the unique minimum variance unbiased estimators of the discriminant scores.

In this paper we propose Bayes estimators of these functions. In addition, some properties of these estimators are investigated.

**2. The classification rules.** The determinant of a matrix $A$ is denoted by $|A|$, $A > 0$ means that $A$ is a positive definite matrix. By $X \sim N_p(\mu, \Sigma)$ we mean that $X$ is a random vector with a $p$-dimensional normal distribution whose density is denoted by

$$n_p(x \mid \mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \quad \Sigma > 0. \tag{1}$$

By $A \sim W_p(\nu, \Sigma)$ we mean that $A$ is a random $p \times p$ symmetric matrix having a Wishart distribution whose density is denoted by

$$\begin{aligned} w_p(A \mid \nu, \Sigma) = {} & 2^{-\frac{p\nu}{2}} \Gamma_p^{-1}\left(\frac{\nu}{2}\right) |\Sigma|^{-\frac{\nu}{2}} [\eta(A)]^{\frac{\nu-p-1}{2}} \\ & \times \exp\left\{-\frac{1}{2} tr(A\Sigma^{-1})\right\}, \end{aligned} \tag{2}$$

for $\nu \geqslant p$ and $\Sigma > 0$, where

$$\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left[t - \frac{1}{2}(j-1)\right] \tag{3}$$

is the multivariate gamma function.

Under condition that the sample $X_0 = \{x_{01}, \ldots, x_{0N_0}\}$ was drawn from $N_p(\mu_i, \Sigma_i)$ we have

$$\bar{x}_0 \sim N_p(\mu_i, N_0^{-1}\Sigma_i), \qquad A_0 \sim W_p(N_0 - 1, \Sigma_i),$$

and the density function of the joint distribution of the sufficient statistics $(\bar{x}_0, A_0)$ has the following form:

$$\begin{aligned} f_i(\bar{x}_0, A_0) = {} & (\pi^{-1} 2^{-N_0} N_0)^{\frac{p}{2}} \Gamma_p^{-1}\left(\frac{1}{2}(N_0 - 1)\right) |\Sigma_i|^{-\frac{N_0}{2}} |A_0|^{\frac{N_0-p-2}{2}} \\ & \times \exp\left\{-\frac{1}{2}\left[N_0(\bar{x}_0 - \mu_i)'\Sigma_i^{-1}(\bar{x}_0 - \mu_i) + tr(A_0\Sigma_i^{-1})\right]\right\}, \end{aligned} \tag{4}$$

for $N_0 - p - 1 \geqslant 0$, $A_0 > 0$, $\Sigma_i > 0$ and $i = 1, \ldots, k$.

The sample $X_0 = \{x_{01}, \ldots, x_{0N_0}\}$ is assigned to that population for which its discriminant score (see, e.g., Rao, 1965, p.488)

$$u_i(\bar{x}_0, A_0) = q_i f_i(\bar{x}_0, A_0) \tag{5}$$

is the highest, where $q_i$ is the prior probability that the classifying sample $X_0$ was drawn from the $i$-th population and the density $f_i(\bar{x}_0, A_0)$ is given by (4), $i = 1, \ldots, k$.

Such a rule is shown to minimize the Bayes risk.

Taking the natural logarithm of $q_i f_i(\bar{x}_0, A_0)$ and omitting the factor

$$(\pi^{-1} 2^{-N_0} N_0)^{\frac{p}{2}} \Gamma_p^{-1}\left(\frac{1}{2}(N_0 - 1)\right) |A_0|^{\frac{N_0 - p - 2}{2}},$$

common to all $i$, we see that the equivalent discriminant score (5) for the $i$-th population is

$$u_i(\bar{x}_0, A_0) = -\frac{1}{2} N_0 \left[\Delta_i^2(\bar{x}_0) + \ln|\Sigma_i| + N_0^{-1} tr\left(A_0 \Sigma_i^{-1}\right)\right] + \ln q_i, \tag{6}$$

involving the mean $\mu_i$ and the covariance matrix $\sum_i$ of the $i$-th population, where

$$\Delta_i^2(\bar{x}_0) = (\bar{x}_0 - \mu_i)' \Sigma_i^{-1} (\bar{x}_0 - \mu_i), \qquad i = 1, \ldots, k. \tag{7}$$

The function (6) is quadratic in $\bar{x}_0$ and may be called a quadratic discriminant score.

The sample $X_0$ is assigned to that population for which the quadratic discriminant score has the highest value.

If the populations do not differ in the covariance matrices, then the sample $X_0$ is assigned to that population for which its discriminant score

$$e_i(\bar{x}_0, A_0) = q_i f_i(\bar{x}_0, A_0) \tag{8}$$

is the highest, where the density $f_i(\bar{x}_0, A_0)$ is given by (4) with $\Sigma_i = \Sigma$, $i = 1, \ldots, k$.

Taking the natural logarithm of $q_i f_i(\overline{x}_0, A_0)$ and omitting the terms common to all $i$, we see that the equivalent discriminant score for the $i$-th population may be written

$$e_i(\overline{x}_0) = N_0\left[\mu_i'\Sigma^{-1}\overline{x}_0 - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i\right] + \ln q_i, \tag{9}$$

which is linear in $\overline{x}_0$ and may be called a linear discriminant score.

We see that the classification rules are based on knowledge of the discriminant scores (6) or (9).

Since this knowledge is usually unavailable it is necessary to estimate these functions.

**3. The estimators of the quadratic discriminant scores.** We shall now treat the case in which we have a sample from each of $k$ normal populations and we wish to use that information in classifying another sample as coming from one of the $k$ populations. Suppose that we have a sample $x_{i1}, x_{i2}, \ldots, x_{iN_i}$ from $N_p(\mu_i, \Sigma_i)$, where $N_i - p - 1 \geqslant 0$, $i = 1, \ldots, k$. Clearly, our best estimate of $\mu_i$ is $\overline{x}_i = \frac{1}{N_i}\sum_{j=1}^{N} x_{ij}$ and of $\Sigma_i$ is $S_i$ defined by $S_i = \frac{1}{N_i - 1}A_i$, where $A_i = \sum_{j=1}^{N_i}(x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)'$, $i = 1, \ldots, k$.

We substitute these estimates for the parameters in (6) to obtain

$$\hat{u}_i^{(1)}(\overline{x}_0, A) = -\frac{N_0}{2}\left[D_i^2(\overline{x}_0) + \ln|S_i| + N_0^{-1}tr(A_0 S_i^{-1})\right] + \ln q_i, \tag{10}$$

where

$$D_i^2(\overline{x}_0) = (\overline{x}_0 - \overline{x}_i)'S_i^{-1}(\overline{x}_0 - \overline{x}_i), \qquad i = 1, \ldots, k. \tag{11}$$

It is easy to show that (10) is a consistent, but biased (asymptotically unbiased) estimator of the quadratic discriminant score (6).

Recently the author (Krzyśko, 1991) proved the following theorem.

**Theorem 1.** *For $N_i - p - 2 > 0$, the unique minimum variance unbiased estimator of the quadratic discriminant score* (6) *which*

*depends on the complete sufficient statistic* $(\bar{x}_i S_i)$ *is given by*

$$\widehat{u}_i^{(2)}(\bar{x}_0, A_0) = -\frac{1}{2}N_0\left[\frac{N_i - p - 2}{N_i - 1}D_i^2(\bar{x}_0) + \ln|S_i| + N_0^{-1}\frac{N_i - p - 2}{N_i - 1}tr(A_0 S_i^{-1}) - \sum_{n=1}^{p}\psi\left[\frac{1}{2}(N_i - n)\right] - pN_i^{-1} + p\ln\frac{N_i - 1}{2}\right] + \ln q_i, \tag{12}$$

*where* $D_i^2(\bar{x}_0)$ *is given by* (11), $i = 1, \ldots, k$, *and*

$$\psi(x) = \frac{d\ln\Gamma(x)}{dx}$$

*is the psi (digamma) function* (see Abramowitz and Stegun, 1965, p.258).

We shall now consider the Bayes estimator of the quadratic discriminant score given by the formula (6).

For a given $x$ the function $u_i(x)$ is a function of the unknown parameters $\mu_i$, $\Sigma_i$, $i = 1, \ldots, k$. Until now the parameters $\mu_i$, $\Sigma_i$, $i = 1, \ldots, k$, were treated as fixed. In the Bayesian approach they will be treated as random variables which distributions reflect the subjective personal belief in the probable values. Assume that before any observation has been made on $X$ our belief in the values of the parameters $\delta_i = (\mu_i, \Sigma_i^{-1})$ is represented by the density of a prior distribution $l(\delta_i)$, $i = 1, \ldots, k$.

Sometimes it is easier to use the improper prior densities $(\int l(\delta_i)\, d\delta_i = \infty)$. Such distributions can be considered if only

$$\int l(\delta_i) f(x \mid \delta_i)\, d\delta_i < \infty,$$

i.e., the posterior distribution is defined.

We will consider the prior distribution that is worked out from the Jeffreys invariance rule (see Jeffreys, 1961, p.179). The rule assumes that the parameters $\delta_i = (\mu_i, \Sigma_i^{-1})$ are independent, i.e.,

$$l(\mu_i, \Sigma_i^{-1}) = l_1(\mu_i) l_2(\Sigma_i^{-1})$$

and that

$$\ell_1(\mu_i) \propto |\mathcal{J}(\mu_i)|^{\frac{1}{2}}, \qquad \ell_2(\Sigma_i^{-1}) \propto |\mathcal{J}(\Sigma_i^{-1})|^{\frac{1}{2}},$$

where $\mathcal{J}(\mu_i)$, $\mathcal{J}(\Sigma_i^{-1})$ are the information matrices for $\mu_i$ and $\Sigma_i^{-1}$ while $\propto$ is the sign of proportionality.

The elements $\mathcal{J}_{mn}$ of the information matrix $\mathcal{J}$ are given by the formula

$$\mathcal{J}_{mn} = -E\left(\frac{\partial^2 \ln f(x \mid \delta_i)}{\partial \delta_{im} \partial \delta_{in}}\right), \qquad m, n = 1, \ldots, p.$$

Assume that $X \sim N(\mu_i, \Sigma_i)$, $i = 1, \ldots, k$. Then

$$L_i = \ln f(x \mid \mu_i, \Sigma_i^{-1}) = \frac{1}{2} \ln |\Sigma_i^{-1}| - \frac{1}{2}(x - \mu_i)' \Sigma_i^{-1}(x - \mu_i) + C$$

for some constant $C$.
First suppose that $\Sigma_i^{-1}$ is fixed. Then

$$\frac{\partial L_i}{\partial \mu_i} = -\frac{1}{2}\left(2\Sigma_i^{-1}\mu_i - 2\Sigma_i^{-1}x\right)$$

and

$$-E\left(\frac{\partial^2 L_i}{\partial \mu_i \partial \mu_i'}\right) = -E(-\Sigma_i^{-1}) = \Sigma_i^{-1} = \text{const}, \qquad i = 1, \ldots, k.$$

Hence the information matrix is constant and the prior density of $\mu_i$ is

$$\ell_1(\mu_i) \propto \text{const}.$$

Now assume that $\mu_i$ is fixed. Writing $L_i$ in the form

$$L_i = \frac{1}{2} \ln |\Sigma_i^{-1}| - \frac{1}{2} tr(x - \mu_i)(x - \mu_i)' \Sigma_i^{-1} + C$$

and using the formulas (see for instance Press, 1972, p.41)

$$\frac{d}{dX}|X| = |X| X^{-1} \qquad \text{for any nonsingular matrix } X,$$

$$\frac{d}{dX} tr(A'X) = A \qquad \text{for } p \times q \text{ matrix } A \text{ and } q \times p \text{ matrix } X.$$

we get

$$\frac{\partial L_i}{\partial \Sigma_i^{-1}} = \frac{1}{2}\Sigma_i - \frac{1}{2}(x-\mu_i)(x-\mu_i)', \qquad i = 1,\ldots,k.$$

If $\Sigma_i^{-1} = (\sigma_{mn}^{(i)})$, then

$$\frac{\partial^2 L_i}{\partial \sigma_{mn}^{(i)} \partial \sigma_{kl}^{(i)}} \propto |\Sigma_i|^{p+1}$$

and

$$|\mathcal{J}(\Sigma_i^{-1})| \propto |\Sigma_i|^{p+1}, \qquad i = 1,\ldots,k.$$

Hence the prior density of the matrix $\Sigma_i^{-1}$ satisfies the condition

$$\ell_2(\Sigma_i^{-1}) \propto |\Sigma_i|^{\frac{p+1}{2}}, \qquad i = 1,\ldots,k.$$

Finally, if $\mu_i$ and $\Sigma_i$ are a priori independent, the prior density implied by the Jeffrys argument is

$$\ell(\mu_i, \Sigma_i^{-1}) \propto |\Sigma_i|^{\frac{p+1}{2}}, \qquad i = 1,\ldots,k. \tag{13}$$

If the quadratic loss function is used, then the Bayes estimator of the discriminant score is the expected value of that function with respect to the joint posterior distribution of the parameters which occur in it (see e.q. Ferguson, 1967, p.46).

The random vector $\overline{x}_i$ has the normal distribution $N_p(\mu_i, N_i^{-1}\Sigma_i)$ and the matrix $A_i$ has the Wishart distribution $W_p(N_i - 1, \Sigma_i)$. Since $\overline{x}_i$ and $A_i$ are independently distributed, then the density of the joint distribution of $(\overline{x}_i, A_i)$ is

$$f\left(\overline{x}_i, A_i \mid \mu_i, \Sigma_i^{-1}\right) \propto |A_i|^{\frac{N_i-p-2}{2}} |\Sigma_i^{-1}|^{\frac{N_i}{2}} \times \exp\left\{-\frac{1}{2} tr\Sigma_i^{-1}\left[A_i + N_i(\overline{x}_i - \mu_i)(\overline{x}_i - \mu_i)'\right]\right\}$$

which upon combining with the assumed joint prior density (13) yields

$$g\left(\mu_i, \Sigma_i^{-1} \mid \overline{x}_i, A_i\right) \propto |\Sigma_i^{-1}|^{\frac{N_i-p-1}{2}} \times \exp\left\{-\frac{1}{2} tr\Sigma_i^{-1}\left[A_i + N_i(\overline{x}_i - \mu_i)(\overline{x}_i - \mu_i)'\right]\right\}$$

as the joint posterior density for $\mu_i$ and $\Sigma_i^{-1}$, $i = 1, \ldots, k$.

Hence, conditional on $\Sigma_i^{-1}$, $\mu_i \sim N_p(\bar{x}_i, N_i^{-1}\Sigma_i)$. Also, marginally $\Sigma_i^{-1} \sim W_p(N_i - 1, (N_i - 1)^{-1}S_i^{-1})$. Now, the evaluation of $E[u_i(\bar{x}_0, A_0) \mid \bar{x}_0, A_0]$ may be most conveniently obtained by taking conditional and then unconditional expectations, viz.,

$$E[u_i(\bar{x}_0, A_0) \mid \bar{x}_0, A_0] = -\frac{N_0}{2} E_{\Sigma_i^{-1}}\Big\{E_{\mu_i}[\Delta_i^2(\bar{x}_0) \mid \Sigma_i^{-1}, \bar{x}_0] - \ln|\Sigma_i^{-1}| + \frac{1}{N_0} tr A_0 \Sigma_i^{-1} \mid \bar{x}_0, A_0\Big\} + \ln q_i, \qquad i = 1, \ldots, k.$$

This requires the evaluation of terms such as $E(\ln|B|)$ where $B \sim W_p(\nu, \Omega)$. Enis and Geisser (1970) showed that

$$E(\ln|B|) = \sum_{j=1}^{p} \psi\Big[\frac{1}{2}(\nu + 1 - j)\Big] + \ln|2\Omega|.$$

Now, it is well known that if the $p$-vector $X$ is distributed as $N_p(\delta, \Omega)$ then $Y = (X - a)'\Omega^{-1}(X - a)$ is distributed as $\chi^2(\beta, p)$ where $\beta = (\delta - a)'\Omega^{-1}(\delta - a)$ and $\chi^2(\beta, p)$ denotes the random variable having the noncentral $\chi^2$ distribution with $p$ degrees of freedom and noncentrality parameter $\beta$.

Moreover, $EY = E[\chi^2(\beta, p)] = \beta + p$. Also, it is easy to show that if $B \sim W_p(\nu, \Omega)$ then for any nonnull vector of real constants $a' = (a_1, \ldots, a_p)$, $Y = \frac{a'Ba}{a'\Omega a} \sim \chi_\nu^2 = \chi^2(0, \nu)$. Hence, $E[a'Ba] = \nu a'\Omega a$.

Since $\Sigma_i^{-1} \sim W_p[N_i - 1, (N_i - 1)^{-1}S_i^{-1}]$, we obtain

$$E\Big[\ln|\Sigma_i^{-1}|\Big] = \ln|S_i^{-1}| + \sum_{j=1}^{p}\Big\{\psi[\frac{1}{2}(N_i - j)] + \ln 2(N_i - 1)^{-1}\Big\}.$$

Further, $\mu_i \sim N_p(\bar{x}_i, N_i^{-1}\Sigma_i)$ conditional on $\Sigma_i^{-1}$, thus

$$E_{\Sigma_i^{-1}}\Big\{E_{\mu_i}[\Delta_i^2(\bar{x}_0) \mid \Sigma_i^{-1}, \bar{x}_0] \mid \bar{x}_0, A_0\Big\} = E_{\Sigma_i^{-1}}\Big\{N_i^{-1}(\beta_i + p) \mid \bar{x}_0, A_0\Big\} = (\bar{x}_0 - \bar{x}_i)'S_i^{-1}(\bar{x}_0 - \bar{x}_i) + pN_i^{-1}.$$

Hence

$$E[u_i(\bar{x}_0, A_0) \mid \bar{x}_0, A_0] = \hat{u}_i^{(1)}(\bar{x}_0, A_0) - 2^{-1}N_0\Big\{pN_i^{-1} - \sum_{j=1}^{p}\psi[\frac{1}{2}(N_i - j)] + p\ln 2^{-1}(N_i - 1)\Big\} = \hat{u}_i^{(3)}(\bar{x}_0, A_0),$$

where $\hat{u}_i^{(1)}(\bar{x}_0, A_0)$ is given by (10).
This proves the following theorem.

**Theorem 2.** *If the density function of the joint prior distribution of the parameters* $(\mu_i, \Sigma_i^{-1})$ *is a Jeffreys* (1961) *function of the form* (13), *then for* $N_i - p - 2 > 0$

$$\hat{u}_i^{(3)}(\bar{x}_0, A_0) = \hat{u}_i^{(1)}(\bar{x}_0, A_0) - 2^{-1}N_0 \times \left\{pN_i^{-1} - \sum_{j=1}^{p} \psi[\tfrac{1}{2}(N_i - j)] + p \ln 2^{-1}(N_i - 1)\right\} \quad (14)$$

*is the Bayes estimator of the quadratic discriminant score* (6).

REMARK 1. Since for large $x$ the function $\psi(x)$ is asymptotically equal to $\ln x + o(1)$, it is clear that

$$\lim_{N_i \to \infty} \left\{p \ln 2^{-1}(N_i - 1) - \sum_{j+1}^{p} \psi[\tfrac{1}{2}(N_i - j)]\right\} = 0$$

and $\hat{u}_i^{(3)}(\bar{x}_0, A_0)$ is a consistent estimator of the quadratic discriminant score (6).

REMARK 2. If the sample sizes are equal, then the classification of a given sample $X_0$ by the estimators (10), (12) and (14) gives the same result.

**4. The estimators of the linear discriminant scores.** Let us assume that $\Sigma_1 = \ldots = \Sigma_k = \Sigma$. Then we define $S$ by

$$S = \frac{1}{N-k} A,$$

where

$$A = \sum_{i=1}^{k} A_i, \qquad N = \sum_{i=1}^{k} N_i \quad \text{and} \quad N - k - p - 1 > 0.$$

We substitute the estimators $(\bar{x}_i, S)$ for the parameters in (9) to obtain

$$\hat{e}_i^{(1)}(\bar{x}_0) = N_0\left(\bar{x}_i' S^{-1} \bar{x}_0 - \frac{1}{2}\bar{x}_i' S^{-1} \bar{x}_i\right) + \ln q_i, \qquad i = 1, \ldots, k. \quad (15)$$

It is easy to show that $\hat{e}_i^{(1)}(\overline{x}_0)$ is a consistent but biased (asymptotically unbiased) estimator of the linear discriminant score (9).

**Theorem 3.** (Krzyśko, 1991). *For $N-k-p-1>0$, the unique minimum variance unbiased estimator of the linear discriminant score* (9) *which depends on the complete sufficient statistic $(\overline{x}_i, S)$ is given by*

$$\hat{e}_i^{(2)}(\overline{x}_0) = \frac{N_0(N-k-p-1)}{N-k}\left(\overline{x}_i' S^{-1}\overline{x}_0 - \frac{1}{2}\overline{x}_i' S^{-1}\overline{x}_i\right) + \frac{pN_0}{2N_i} + \ln q_i, \quad i = 1,\ldots,k. \tag{16}$$

We shall now consider the Bayes estimator of the linear discriminant score.

The linear discriminant score (9) is the function of the unknown parameters $(\mu_i, \Sigma^{-1})$, $i = 1,\ldots,k$.

Suppose that the density function of the joint prior distribution of the parameters $(\mu_i, \Sigma^{-1})$ is Jeffreys (1961) function of the form

$$l(\mu_i, \Sigma^{-1}) \alpha |\Sigma|^{\frac{p+1}{2}}, \qquad i = 1,\ldots,k. \tag{17}$$

The random vector $\overline{x}_i$ has the normal distribution $N_p(\mu_i, N_i^{-1}\Sigma_i)$ and the matrix $A$ has the Wishart distribution $W_p(N-k, \Sigma)$. Since $\overline{x}_i$ and $A$ are independently distributed, then the density of the joint distribution of $(\overline{x}_i, A)$ is

$$f(\overline{x}_i, A \mid \mu_i, \Sigma^{-1}) \propto |A|^{\frac{N-k-p-1}{2}} |\Sigma|^{-\frac{N-k+1}{2}} \times \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(A + N_i(\overline{x}_i - \mu_i)(\overline{x}_i - \mu_i)')\right]\right\}$$

which upon combining with the assumed joint prior density (17) yields

$$g(\mu_i, \Sigma^{-1} \mid \overline{x}_i, A) \propto |A|^{\frac{N-k-p-1}{2}} |\Sigma|^{-\frac{N-k+1}{2}} \times \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(A + N_i(\overline{x}_i - \mu_i)(\overline{x}_i - \mu_i)')\right]\right\}$$

as the joint posterior density for $\mu_i$ and $\Sigma^{-1}$, $i = 1,\ldots,k$. Hence, conditional on $\Sigma^{-1}$, $\mu_i \sim N_p(\overline{x}_i, N_i^{-1}\Sigma)$.

Also, marginally

$$\Sigma^{-1} \sim W_p\left(N-k,(N-k)^{-1}S^{-1}\right).$$

Now, the evaluation of $E[e_i(\bar{x}_0) \mid \bar{x}_0]$ may be most conveniently obtained by taking conditional and then unconditional expectations, viz.,

$$E[e_i(\bar{x}_0) \mid \bar{x}_0] = -\frac{N_0}{2}E_{\Sigma^{-1}}\left\{E_{\mu_i}\left[(\bar{x}_0-\mu_i)'\Sigma^{-1}(\bar{x}_0-\mu_i) \mid \Sigma^{-1},\bar{x}_0\right] \mid \bar{x}_0\right\} + \ln q_i, \quad i=1,\ldots,k.$$

By a similar argument as in the proof of Theorem 2 we have

$$E\left[e_i(\bar{x}_0) \mid \bar{x}_0\right] = \hat{e}_i^{(1)}(\bar{x}_0) - \frac{pN_0}{2N_i} = \hat{e}_i^{(3)}(\bar{x}_0),$$

where $\hat{e}_i^{(1)}(\bar{x}_0)$ is given by (15), $i=1,\ldots,k$.

This proves the following theorem.

**Theorem 4.** *If the density function of the joint prior distribution of the parameters* $(\mu_i,\Sigma^{-1})$ *is a Jeffreys* (1961) *function of the form* (17), *then for* $N-k-p-1>0$

$$\hat{e}_i^{(3)}(\bar{x}_0) = N_0\left(\bar{x}_i'S^{-1}\bar{x}_0 - \frac{1}{2}\bar{x}_i'S^{-1}\bar{x}_i - \frac{p}{2N_i}\right) + \ln q_i \qquad (18)$$

*is the Bayes estimator of the linear discriminant score* (9).

REMARK 3. It is easily seen that $\hat{e}_i^{(3)}(\bar{x}_0)$ is a consistent estimator of the linear discriminant score (9).

REMARK 4. If the sample sizes are equal, then the classification of a given sample $X_0$ by the estimators (15), (16) and (18) gives the same result.

## REFERENCES

Abramowitz, M., and I.A.Stegun (eds) (1965). *Handbook of Mathematical Functions.* Dover Publications, New York.

Abusev, R.A., and Ya.P.Lumelsky (1980). Unbiased estimators and classification problems for multivariate normal populations. *Teorija Veroyatn. i Primenen.*, 25(2), 381–389 (in Russian).

Abusev, R.A., and Ya.P.Lumelsky (1987). *Statistical Group Classification.* The Perm State University Press, Perm (in Russian).

Enis, P., and S.Geisser (1970). Sample discriminants which minimize posterior squared error loss. *South. African. Statist. J.*, 4, 85–93.

Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York.

Jeffreys, H. (1961). *Theory of Probability.* 3rd ed. Oxford University Press, Oxford.

Krzyśko, M. (1991). Unbiased estimators and statistical group classification problems. In Ju. S. Kharin (Ed.), *Problems of the Computer Data Analysis and Modelling,* The Minsk State University Press, Minsk. pp. 73–78.

Press, S.J. (1972). *Applied Multivariate Analysis.* Holt, Rinehart and Winston, Inc., New York.

Rao, C.R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, New York.

Received January 1992

M. Krzyśko received Ph. D. Degree and Doctor Habilitus Degree from Adam Mickiewicz University, Poznań, Poland, in 1972 and 1977, respectively. Since 1985 full professor in the Institute of Mathematics of the Adam Mickiewicz University. Professor Krzyśko is head of Probability Theory and Mathematical Statistics Department and director of the University Computing Center. His research interests are in multivariate statistical methods and concern mainly problems of the discriminant analysis, canonical analysis and time series.