Serdica

Bulgariacae mathematicae publicationes

Сердика

Българско математическо списание

NONPARAMETRIC ESTIMATION ASSOCIATED WITH DISCRIMINANT ANALYSIS

M. KRZYŚKO

Let us assume that in the population π_i, the observed random vector $\mathbf{X}_i$ has a p-dimensional normal distribution with the parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}$, for $i=1, 2$. We have an observation $\mathbf{x}$, which we wish to classify to one of the two populations π_1 or π_2. Of many approaches to the problem of discrimination thus formulated, we shall here select a decision-theoretic approach. The linear discriminant function obtained within this approach is a function of the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}\,(i=1, 2)$. We are here concerned with the case in which the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are not known and in which some estimators of the linear discriminant function are used. We shall introduce two types of estimators, examine their properties and make some comparisons between them.

1. The linear discriminant function. Let us assume that the probability density function of a p-dimensional random vector $\mathbf{X}_i$ observed in the population π_i is of the form

$$(1)\qquad f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})=(2\pi)^{-p/2}\,|\,\boldsymbol{\Sigma}\,|^{-1/2}\exp\,[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)],\quad i=1, 2.$$

The known a priori probability that the event of the observation to be classified comes from population π_i will be denoted by $q_i\,(q_i>0,\ q_1+q_2=1)$ and the loss arising from a misclassification of an observation $\mathbf{x}$ into population π_j whilst it really belongs to population π_i, by $S(j|i)$ for $i, j=1, 2$. If $S(j|i)$ is the so-called simple loss function of the form

$$S(j|i)=\begin{cases}0, & \text{if}\quad j=i,\\ 1, & \text{if}\quad j\neq i,\end{cases}$$

then the Bayes risk r is expressed by

$$r=1-\sum_{i=1}^{2} q_i \int_{R_i} f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})d\mathbf{x}$$

and the optimal (in the sense of minimizing the value of r) classification region R_i, that is the set of those $\mathbf{x}$'s for which we can state that the observation under classification belongs to population π_i has the form:

$$R_i=\{\mathbf{x}: q_i f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\geqq q_j f(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}),\ j=1, 2, j\neq i\}$$

or, equivalently,

$$R_i=\{\mathbf{x}: v_{ij}(\mathbf{x})\geqq \ln(q_j/q_i),\ j=1, 2,\ j\neq i\},$$

SERDICA Bulgaricae mathematicae publicationes. Vol. 9, 1983, p. 107-114.

where

$$v_{ij}(\mathbf{x}) = \ln [f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})/f(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})] = \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i), \quad i, j = 1, 2;\ j \neq i. \tag{2}$$

The function $v_{ij}(\mathbf{x})$, given by (2) is called *linear discriminant function* In order to use that function it is necessary to know the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$, $i = 1, 2$. When these parameters are not known, the following two procedures are actually used.

2. The frequency-related estimators of the density function. Let $\bar{\mathbf{x}}_i$ and $\mathbf{S}$ denote the usual estimators of the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ obtained from two samples of size N_1 and N_2 respectively. If, in the density function of the form (1), we replace the unknown parameters by their estimators, we obtain the *frequency related estimator of the density function* $f(\mathbf{x}/\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ of the form

$$p(\mathbf{x}|\bar{\mathbf{x}}_i, \mathbf{S}) = (2\pi)^{-p/2} |\mathbf{S}|^{-1/2} \exp[-\frac{1}{2} D_i^2(\mathbf{x})], \tag{3}$$

where $D_i^2(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)$, $i = 1, 2$.

If we use the estimator (3), the linear discriminant function takes the form

$$u_{ij}(\mathbf{x}) = \frac{1}{2} D_j^2(\mathbf{x}) - \frac{1}{2} D_i^2(\mathbf{x}); \ i, j = 1, 2;\ j \neq i. \tag{4}$$

Another type of the density function estimator is the Bayes estimator.

3. Bayes estimator of the density function. When the quadratic loss function is used, then the Bayes estimator of the density function $f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is the expected value of that function with respect to the a posteriori distribution of the parameters which occur in it.

We shall denote that estimator by $h(\mathbf{x}|\bar{\mathbf{x}}_i, \mathbf{S})$, $i = 1, 2$. We have

$$h(\mathbf{x}|\bar{\mathbf{x}}_i, \mathbf{S}) = \int\int f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) t(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}|\bar{\mathbf{x}}_i, \mathbf{S}) d\boldsymbol{\mu}_i \, d\boldsymbol{\Sigma}, \tag{5}$$

where $t(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}|\bar{\mathbf{x}}_i, \mathbf{S})$ is the density function of the a posteriori distribution of the parameters $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Assume that the density function of the joint a priori distribution of the parameters $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is a Jeffreys function [5] of the form

$$g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}^{-1}) \propto |\boldsymbol{\Sigma}|^{(p+1)/2};$$

we obtain [4]:

$$h(\mathbf{x}|\bar{\mathbf{x}}_i, \mathbf{S}) = c_i [1 + N_i(N_i+1)^{-1}(N_1+N_2-2)^{-1} D_i^2(x)]^{-(N_1+N_2-1)/2}, \tag{6}$$

where

$$c_i = [\pi N_i^{-1}(N_1+N_2-2)(N_i+1)]^{-p/2} \frac{\Gamma[(N_1+N_2-1)/2]}{\Gamma[(N_1+N_2-2)/2] |(N_1+N_2-2)\mathbf{S}|^{1/2}}, \quad i = 1, 2.$$

The function $h(\mathbf{x}|\bar{\mathbf{x}}_i, \mathbf{S})$ is the density function of the p-dimensional t distribution [3].

If we use the estimator (6), then the discriminant function takes the following form:

(7)
$$w_{ij}(\mathbf{x}) = \ln [h(\mathbf{x}|\bar{\mathbf{x}}_i, \mathbf{S})/h(\mathbf{x}|\bar{\mathbf{x}}_j, \mathbf{S})]$$
$$= \frac{N_1+N_2-1}{2} \ln [1+N_j(N_j+1)^{-1}(N_1+N_2-2)^{-1}D_j^2(\mathbf{x})]$$
$$-\frac{N_1+N_2-1}{2} \ln [1+N_i(N_i+1)^{-1}(N_1+N_2-2)^{-1}D_i^2(\mathbf{x})] + \frac{p}{2} \ln \frac{N_1(N_j+1)}{N_j(N_i+1)},$$
$$i, j=1, 2; \quad j \neq i.$$

4. A comparison of the two types of estimators. We shall now consider the consistency and the mean bias of the estimators $u_{ij}(\mathbf{x})$ and $w_{ij}(\mathbf{x})$ of the linear discriminant function $v_{ij}(\mathbf{x})$. As $\bar{\mathbf{x}}_i$ and $\mathbf{S}$ are estimators from samples aken from the normal population $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$,

$$p \lim_{N_i \to \infty} \bar{\mathbf{x}}_i = \boldsymbol{\mu}_i, \quad p \lim_{N_1, N_2 \to \infty} \mathbf{S} = \boldsymbol{\Sigma},$$

where p lim denotes the asymptotic convergence in probability. Hence,

$$p \lim_{N_1, N_2 \to \infty} (\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) = (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

and

$$p \lim_{N_1 N_2 \to \infty} u_{ij}(\mathbf{x}) = v_{ij}(\mathbf{x}).$$

Consequently, $u_{ij}(\mathbf{x})$ is a consistent estimator of the linear discriminant function $v_{ij}(\mathbf{x})$. Similarly,

$$p \lim_{N_1, N_2 \to \infty} \frac{\ln \left[1+\frac{N_i}{(N_i+1)(N_1+N_2-2)} D_i^2(\mathbf{x})\right]}{\frac{1}{N_1+N_2-1}} = (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

and

$$\lim_{N_1, N_2 \to \infty} \ln \frac{N_i(N_j+1)}{N_j(N_i+1)} = 0.$$

Hence,

$$p \operatorname{lin}_{N_1, N_2 \to \infty} w_{ij}(\mathbf{x}) = v_{ij}(\mathbf{x}).$$

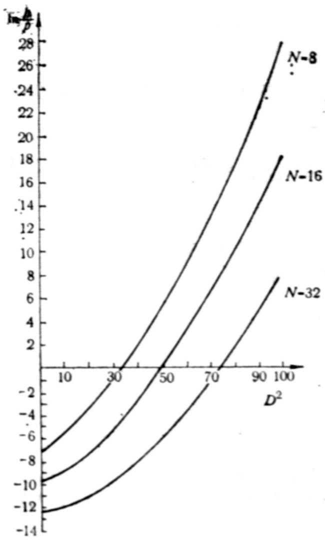
Therefore, $w_{ij}(\mathbf{x})$ is also a consistent estimator of the linear discriminant function $v_{ij}(\mathbf{x})$.

The fact that the two estimators (4) and (7) of the discriminant function $v_{ij}(\mathbf{x})$ are asymptotically equivalent does not mean, however, that for finite samples there are no substantial differences between them. A convenient way of capturing the quantitative difference of the estimators (4) and (7) is to examine the expression

(8)
$$\ln [h(\mathbf{x})|\bar{\mathbf{x}}_i, \mathbf{S})/p(\mathbf{x}\,|\,\bar{\mathbf{x}}_i, \mathbf{S})]$$
$$= \frac{1}{2} D_i^2(\mathbf{x}) - \frac{N_1+N_2-1}{2} \ln [1+N_i(N_i+1)^{-1}(N_1+N_2-2)^{-1}D_i^2(\mathbf{x})]$$
$$+ \ln \frac{\Gamma[(N_1+N_2-1)/2]}{\Gamma[(N_1+N_2-2)/2]} + \frac{p}{2} \ln \frac{2N_i}{(N_1+N_2-2)^2(N_i+1)}, \quad i=1, 2.$$

Fig. 1 shows a graph of the expression (8) as a function of the argument $D_i^2(\mathbf{x})$ for $p=4$ and $N_1=N_2=N=8$, 16, 32.

The value of the expression $\ln(h/p)$ over the interval $0 \leq D^2 \leq 100$ varies approximately from 10^{-3} to 10^{13}, from 10^{-4} to 10^8 and from 10^{-5} to 10^4 for



$N=8$, 16, 32 respectively. It can be seen that, especially for small samples, the values of the two estimators differ considerably.

We shall now find the mean bias of the estimators $u_{ij}(\mathbf{x})$ and $w_{ij}(\mathbf{x})$ of the linear discriminant function $v_{ij}(\mathbf{x})$. By mean bias we shall mean the following expressions:

$$E\{[u_{ij}(\mathbf{x})-v_{ij}(\mathbf{x})]|\,\mathbf{x}\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\}$$

and

$$E\{[w_{ij}(\mathbf{x})-v_{ij}(\mathbf{x})]|\,\mathbf{x}\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\}.$$

We have

$$E[v_{ij}(\mathbf{x})\,|\,\mathbf{x}\sim N(\boldsymbol{\mu}_i, \Sigma)] = \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) = \Delta_{ij}^2,$$

where Δ_{ij} is the Mahalanobis distance between the populations π_i and π_j. Further

$$E[u_{ij}(\mathbf{x})\,|\,\mathbf{x}\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})] = \frac{1}{2}\,\frac{N_1+N_2-2}{N_1+N_2-p-3}\,\Delta_{ij}^2 + \frac{p(N_1+N_2-2)}{2(N_1+N_2-p-3)}\left(\frac{1}{N_j}-\frac{1}{N_i}\right).$$

Therefore,

(9)
$$E\{[u_{ij}(\mathbf{x})-v_{ij}(\mathbf{x})]\,|\,\mathbf{x}\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\}$$
$$= \frac{p+1}{2(N_1+N_2-p-3)}\,\Delta_{ij}^2 + \frac{p(N_1+N_2-2)}{2(N_1+N_2-p-3)}\left(\frac{1}{N_j}-\frac{1}{N_i}\right).$$

We shall now calculate the expected value of the estimator $w_{ij}(\mathbf{x})$. If $\mathbf{x}\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, then

$$\frac{\nu_2}{\nu_1}\frac{N_i}{(N_i+1)(N_1+N_2-2)}D_i^2(\mathbf{x})\sim F_{\nu_1,\nu_2}$$

(the central F distribution with $\nu_1=p$ and $\nu_2=N_1+N_2-p-1$ degrees of freedom). Hence,

$$E\{\ln[1+\frac{N_i}{(N_i+1)(N_1+N_2-2)}D_i^2(\mathbf{x})]\mid \mathbf{x}\sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})\}=E[\ln(1+\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2})].$$

This last expected value will be calculated using the method of characteristic functions. We have

$$\varphi(t)=E\{\exp[it\ln(1+\frac{\nu_1}{\nu_2}E_{\nu_1,\nu_2})]\}=E[(1+\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2})^{it}]=\frac{\Gamma[(\nu_1+\nu_2)/2]\,\Gamma[(\nu_2-2it)/2]}{\Gamma(\nu_2/2)\,\Gamma[(\nu_1+\nu_2-2it)/2]}.$$

Hence

$$E[\ln(1+\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2})]=\frac{\varphi'(0)}{i}=\psi[\frac{1}{2}(\nu_1+\nu_2)]-\psi[\frac{1}{2}\nu_2],$$

where [1, p. 258] $\psi(x)=\frac{d\ln\Gamma(x)}{dx}$.

If $\mathbf{x}\sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, then

$$\frac{\nu_2}{\nu_1}\frac{N_j}{(N_j+1)(N_1+N_2-2)}D_j^2(\mathbf{x})\sim F_{\nu_1,\nu_2},\lambda$$

(non-central F distribution with $\nu_1=p$ and $\nu_2=N_1+N_2-p-1$ degrees of freedom and non-centrality parameter $\lambda=N_j(N_j+1)^{-1}\Delta_{ij}^2$). Hence

$$E\{\ln[1+\frac{N_j}{(N_j+1)(N_1+N_2-2)}D_j^2(\mathbf{x})]\mid \mathbf{x}\sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})\}=E[\ln(1+\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2,\lambda})].$$

Further on

$$\varphi_1(t)=E\{\exp[it\ln(1+\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2,\lambda})]\}=\sum_{m=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^m}{m!}\frac{\beta[\frac{1}{2}\nu_1+j,\frac{1}{2}(\nu_2-2it)]}{\beta[\frac{1}{2}\nu_1+j,\frac{1}{2}\nu_2]}.$$

Hence

$$E[\ln(1+\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2,\lambda})]=\frac{\varphi_1'(0)}{i}=\sum_{m=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^m}{m!}\{\psi[\frac{1}{2}(\nu_1+\nu_2)+m]-\psi[\frac{1}{2}\nu_2]\}.$$

Therefore

$$E[w_{ij}(\mathbf{x})\mid \mathbf{x}\sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})]=\frac{N_1+N_2-1}{2}\sum_{m=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^m}{m!}\{\psi[\frac{1}{2}(N_1+N_2-1)+m]$$

$$-\psi[\frac{1}{2}(N_1+N_2-p-1)]\}-\frac{N_1+N_2-1}{2}\{\psi[\frac{1}{2}(N_1+N_2-1)]-\psi[\frac{1}{2}(N_1+N_2-p-1)]\}$$

$$+\frac{p}{2}\ln\frac{N_i(N_j+1)}{N_j(N_i+1)}.$$

Using of the recurrence relation $\psi(x+1)=\psi(x)+x^{-1}$ and the identity

$$\Pr(\chi_{2m}^2\leq\lambda)=\sum_{j=m}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^j}{j!}$$

we get:

$$\sum_{m=0}^{\infty} \frac{e^{\lambda/2}(\lambda/2)^m}{m!} \{\psi\,[\tfrac{1}{2}(N_1+N_2-1)+m] - \psi\,[\tfrac{1}{2}(N_1+N_2-p-1)]\}$$

$$= \sum_{m=0}^{\infty} \frac{1}{p+m} \Pr(\chi^2_{2(m+1)} \leq \lambda) + \psi\,[\tfrac{1}{2}(N_1+N_2-1)] - \psi\,[\tfrac{1}{2}(N_1+N_2-p-1)].$$

Using that transformation we obtain

$$E\,[w_{ij}(\mathbf{x}) \mid \mathbf{x} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})] = \frac{p}{2} \ln \frac{N_i(N_j+1)}{N_j(N_i+1)} + \sum_{m=0}^{\infty} \frac{N_1+N_2-1}{N_1+N_2-1+2m} \Pr(\chi^2_{2(m+1)} \leq \lambda).$$

Thus, the value of the mean bias of the estimator $w_{ij}(\mathbf{x})$ equals

(10) $$E\{[w_{ij}(\mathbf{x}) - v_{ij}(\mathbf{x})]\mathbf{x} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\}$$

$$= \frac{p}{2} \ln \frac{N_i(N_j+1)}{N_j(N_i+1)} + \sum_{m=0}^{\infty} \frac{N_1+N_2-1}{N_1+N_2-1+2m} \Pr(\chi^2_{2(m+1)} \leq \frac{N_j}{N_j+1} \Delta^2_{ij}) - \frac{1}{2} \Delta^2_{ij}.$$

Tabulated values of the mean biases of the estimators $u_{ij}(\mathbf{x})$ and $w_{ij}(\mathbf{x})$ for various of Δ^2, p and $N_1=N_2=N$ are contained in Table 1. In this table the

Table 1

The Mean Bias of the Estimators $u_{ij}(\underline{x})$ and $w_{ij}(\underline{x})$ When $N_1=N_2=N$

Estimator and dimension p	u $p=2$			u $p=4$			u $p=8$			w $p=2, 4, 8$		
Δ^2 \ N	16	32	64	16	32	64	16	32	64	16	32	64
1.1004	0.06	0.03	0.01	0.11	0.05	0.02	0.24	0.09	0.04	—0.01	—0.02	—0.01
2.8325	0.16	0.07	0.03	0.28	0.12	0.05	0.61	0.24	0.11	—0.13	—0.07	—0.04
6.5690	0.36	0.17	0.08	0.66	0.29	0.14	1.41	0.56	0.25	—0.45	—0.25	—0.13
10.8222	0.60	0.27	0.13	1.08	0.47	0.22	2.32	0.92	0.42	—0.93	—0.55	—0.29
21.6504	1.20	0.55	0.25	2.17	0.95	0.45	4.61	1.84	0.83	—2.92	—1.74	—0.96
38.1981	2.12	0.97	0.47	3.82	1.67	0.79	8.18	3.24	1.47	—7.02	—4.46	—2.60

values of Δ^2 are chosen in such a way that $\Phi(-\Delta/2)=0.3, 0.2, 0.1, 0.05, 0.01$ 0.001, where $\Phi(-\Delta/2)$ is the probability of misclassification.

It can be seen from Table 1 that the mean bias of the estimator $u_{ij}(\mathbf{x})$ is positive and increases with p. The mean bias of the estimator $w_{ij}(x)$ is negative and independent from p when $N_1=N_2=N$. For small p and increasing values of $\Delta^2 u_{ij}(\mathbf{x})$ is the better estimator. In other cases, especially if p, N and Δ^2 are all small, the mean bias of the estimator $w_{ij}(\mathbf{x})$ is less than mean bias of the estimator $u_{ij}(\mathbf{x})$.

REFERENCES

1. M. Abramowitz, I. A. Stegun (Eds.). Handbook of mathematical functions. New York, 1965.
2. T. W. Anderson. An introduction to multivariate statistical analysis New York, 1958.
3. E. A. Cornish. The multivariate t-distribution associated with a set of normal sample deviates. *Austral. J. Phys.*, 7, 1954, 531-542.
4. S. Geisser. Predictive discrimination. — In: ***Multivariate analysis,*** P. R. Krishnaiah (Ed.). New York, 1966, 149-163.
5. H. Jeffreys. Theory of probability. Oxford, 1961.

Inst. Math., Univ. Adama Mickiewicza
60—769 Poznan Poland

Received 7. 7. 1981