10.2478/v10048-008-0013-z
MEASUREMENT SCIENCE REVIEW, Volume 8, Section 1, No. 3, 2008

On the Discriminant Analysis in the 2-Populations Case

František Rublík

Institute of Measurement Science, Slovak Academy of Sciences,
Dúbravská cesta 9, 84104 Bratislava, Slovakia
E-mail: umerrubl@savba.sk

The empirical Bayes Gaussian rule, which in the normal case yields good values of the probability of total error, may yield high values of the maximum probability error. From this point of view the presented modified version of the classification rule of Broffitt, Randles and Hogg appears to be superior. The modification included in this paper is termed as a WR method, and the choice of its weights is discussed. The mentioned methods are also compared with the K nearest neighbours classification rule.

Keywords: discriminant analysis, nearest neighbour rule, Gaussian Bayes classification rule, maximum probability error.

1. Introduction

Suppose that $\mathbf{X_1} = (\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1})$, $\mathbf{X_2} = (\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2})$ are observed data consisting of m-dimensional column vectors; here $\mathbf{X_1}$ is a random sample from an m-dimensional population Π_1 and $\mathbf{X_2}$ a random sample from another m-dimensional population Π_2. Let $\mathbf{Z}$ denote an m-dimensional observed column vector, which is assumed to belong to one of the populations. The task of the discriminant analysis is to classify $\mathbf{Z}$ by means of $\mathbf{X_1}$, $\mathbf{X_2}$ either as belonging to the population Π_1 or to Π_2. The quality of the classification rule is judged by means of probability $P(2|1)$ that the element belonging to Π_1 is assigned to Π_2, by probability $P(1|2)$ that the element belonging to Π_2 is assigned to Π_1 and by the probability of the total error

$$PTE = p_1 P(2|1) + p_2 P(1|2)\,, \tag{1}$$

where p_j denotes the probability that the observed value $\mathbf{Z}$ belongs to the population Π_j (i.e., the prior probability of the population Π_j). The quantity (1) is used to express the probability of the wrong decision in the whole classification process. In addition to this, we prefer to consider the quantity

$$MPE = \max\{P(2|1)\,,\, P(1|2)\} \tag{2}$$

because this maximum probability of error assigns the same importance to both populations regardless of the frequency with which the observed vector $\mathbf{Z}$ is generated by the particular population Π_1 or Π_2.

The aim of the paper is to present a classification rule which can be expected to yield good results when the probability (2) is chosen as a criterion of quality. This new rule, proposed and implemented by the author of this paper in the technical report [5], is described in Section 2. Classical rules, with which we compare the new rule by means of simulation results, are described in Section 3. Simulation estimates of the power of the mentioned classification rules are presented in Section 4 and these results are discussed in Section 5.

In what follows we use for $j = 1, 2$ the notation

$$\overline{\mathbf{X}}_\mathbf{j} = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{X}_{ji}\,, \quad \mathbf{S_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} (\mathbf{X}_{ji} - \overline{\mathbf{X}}_\mathbf{j})(\mathbf{X}_{ji} - \overline{\mathbf{X}}_\mathbf{j})'. \tag{3}$$

2. Weighted Ranks Rule

Now we are going to present the modification of the rank based classification procedure, described in [2] and [3]. The modification consists in using a new target discriminant function employed in computation of the ranks and in involving weights in the decision process for the sake of achieving a better PTE.

The idea of the classification rule follows the approach worked out in [2] and [3] in the sense that the observation which has to be classified is added to a training set and the obtained ranks are used for classification. More precisely, observed vector $\mathbf{Z}$ is first added to $\mathbf{X_1}$ observations and the rank $R_{\mathbf{X_1}}(\mathbf{Z})$ of the target function $D_{\mathbf{X_1}}(\mathbf{Z})$ is computed, then $\mathbf{Z}$ is added to $\mathbf{X_2}$, the rank $R_{\mathbf{X_2}}(\mathbf{Z})$ of the target function $D_{\mathbf{X_2}}(\mathbf{Z})$ is obtained and the classification is carried out by means of the values of these ranks. Thus, add the observation $\mathbf{Z}$ to the $\mathbf{X_1}$ observations and compute

$$\tilde{\mathbf{X}}_\mathbf{1} = \tfrac{1}{n_1+1}\Big(\sum_{i=1}^{n_1} \mathbf{X}_{1i} + \mathbf{Z}\Big)\,,$$

$$\tilde{\mathbf{S}}_\mathbf{1} =$$
$$= \tfrac{1}{n_1+1}\Big(\sum_{i=1}^{n_1} (\mathbf{X}_{1i} - \tilde{\mathbf{X}}_\mathbf{1})(\mathbf{X}_{1i} - \tilde{\mathbf{X}}_\mathbf{1})' + (\mathbf{Z} - \tilde{\mathbf{X}}_\mathbf{1})(\mathbf{Z} - \tilde{\mathbf{X}}_\mathbf{1})'\Big)$$

the estimates based on the extended sample. Define the functions of the argument $\mathbf{U} \in R^m$ by the formula

$$\tilde{t}_1(\mathbf{U}) = (\mathbf{U} - \tilde{\mathbf{X}}_\mathbf{1})'\,(\tilde{\mathbf{S}}_\mathbf{1})^{-1}(\mathbf{U} - \tilde{\mathbf{X}}_\mathbf{1}\Big)\,,$$
$$t_2(\mathbf{U}) = (\mathbf{U} - \overline{\mathbf{X}}_\mathbf{2})'\,(\mathbf{S_2})^{-1}(\mathbf{U} - \overline{\mathbf{X}}_\mathbf{2})\,,$$

put

$$D_{\mathbf{X_1}}(\mathbf{U}) = \begin{cases} t_2(\mathbf{U}) - \tilde{t}_1(\mathbf{U}) & \text{if } t_2(\mathbf{U}) > \tilde{t}_1(\mathbf{U}), \\ \log\Big(\dfrac{t_2(\mathbf{U})}{\tilde{t}_1(\mathbf{U})}\Big) & \text{if } t_2(\mathbf{U}) \le \tilde{t}_1(\mathbf{U}), \end{cases}$$

and compute the ranks $R_{\mathbf{X_1}}(\mathbf{Z}), R_{\mathbf{X_1}}(\mathbf{X}_{11}), R_{\mathbf{X_1}}(\mathbf{X}_{12}), \ldots, R_{\mathbf{X_1}}(\mathbf{X}_{1n_1})$ of the numbers $D_{\mathbf{X_1}}(\mathbf{Z}), D_{\mathbf{X_1}}(\mathbf{X}_{11}), D_{\mathbf{X_1}}(\mathbf{X}_{12}), \ldots, D_{\mathbf{X_1}}(\mathbf{X}_{1n_1})$ in their ordering according to the magnitude. Further, add the observation $\mathbf{Z}$ to the $\mathbf{X_2}$ observations, compute

$$\tilde{\mathbf{X}}_\mathbf{2} = \tfrac{1}{n_2+1}\Big(\sum_{i=1}^{n_2} \mathbf{X}_{2i} + \mathbf{Z}\Big)\,,$$

$$\tilde{\mathbf{S}}_\mathbf{2} =$$
$$= \tfrac{1}{n_2+1}\Big(\sum_{i=1}^{n_2} (\mathbf{X}_{2i} - \tilde{\mathbf{X}}_\mathbf{2})(\mathbf{X}_{2i} - \tilde{\mathbf{X}}_\mathbf{2})' + (\mathbf{Z} - \tilde{\mathbf{X}}_\mathbf{2})(\mathbf{Z} - \tilde{\mathbf{X}}_\mathbf{1})'\Big)$$

and define the functions of the argument $\mathbf{U} \in R^m$ by the formula

$$\tilde{t}_2(\mathbf{U}) = (\mathbf{U} - \tilde{\mathbf{X}}_\mathbf{2})'\,(\tilde{\mathbf{S}}_\mathbf{2})^{-1}(\mathbf{U} - \tilde{\mathbf{X}}_\mathbf{2}\Big)\,,$$
$$t_1(\mathbf{U}) = (\mathbf{U} - \overline{\mathbf{X}}_\mathbf{1})'\,(\mathbf{S_1})^{-1}(\mathbf{U} - \overline{\mathbf{X}}_\mathbf{1})\,.$$

Put

$$D_{\mathbf{X_2}}(\mathbf{U}) = \begin{cases} t_1(\mathbf{U}) - \tilde{t}_2(\mathbf{U}) & \text{if } t_1(\mathbf{U}) > \tilde{t}_2(\mathbf{U}), \\ \log\left(\dfrac{t_1(\mathbf{U})}{\tilde{t}_2(\mathbf{U})}\right) & \text{if } t_1(\mathbf{U}) \le \tilde{t}_2(\mathbf{U}), \end{cases}$$

and compute the ranks $R_{\mathbf{X_2}}(\mathbf{Z}), R_{\mathbf{X_2}}(\mathbf{X}_{21}), R_{\mathbf{X_2}}(\mathbf{X}_{22}), \ldots, R_{\mathbf{X_2}}(\mathbf{X}_{2n_2})$ of the numbers $D_{\mathbf{X_2}}(\mathbf{Z}), D_{\mathbf{X_2}}(\mathbf{X}_{21}), D_{\mathbf{X_2}}(\mathbf{X}_{22}), \ldots, D_{\mathbf{X_2}}(\mathbf{X}_{2n_2})$ in their ordering according to the magnitude. In accordance with the importance of the population $\mathbf{\Pi}_2$, choose its weight $we > 0$. The classification rule is

$$\begin{aligned} &\textit{If} \quad \frac{R_{\mathbf{X_1}}(\mathbf{Z})}{n_1+1} \ge we\frac{R_{\mathbf{X_2}}(\mathbf{Z})}{n_2+1} \textit{ classify } \mathbf{Z} \in \mathbf{\Pi}_1, \\ &\quad \textit{otherwise classify } \mathbf{Z} \in \mathbf{\Pi}_2. \end{aligned} \tag{4}$$

In this rule the inequality $\ge$ is employed to simplify computation because the validity of the equality occurs in practice either very rarely or not at all. Typically, the sizes of the training sets are not equal and for the sake of the clarity of the use of the weights we shall label by Π_1 the population with a larger sample size (hence from now on the inequality $n_1 \ge n_2$ holds), the rule (4) will be referred to as the *WR method* (WR stands for weighted ranks). We remark that the WR method turned out to be useful in the classification of the observed data concerning doping control studied in [5]. In what follows we shall investigate three particular values of the weight:

$$we = 1\,, \quad we = \frac{n_2}{n},\ n = n_1 + n_2, \quad we = \frac{n_2}{n_1}\,.$$

The choice $we = 1$, which will be called the *equal weights*, can be expected to result in smaller MPE from (2), while the other two values of the weight will often lead to a smaller total error (1).

3. Some Alternative Classical Rules

Let for $j = 1, 2$

$$\hat{p}_j = \frac{n_j}{n_1 + n_2}\,.$$

The following classification rule

$$\begin{aligned} &\textit{Assign } \mathbf{Z} \textit{ to } \Pi_j \textit{ if} \\ &(\mathbf{Z} - \overline{\mathbf{X}}_\mathbf{j})' {\mathbf{S_j}}^{-1} (\mathbf{Z} - \overline{\mathbf{X}}_\mathbf{j}) + \log(det(\mathbf{S_j})/(\hat{p}_j^2)) = \\ &= \min_t \left[(\mathbf{Z} - \overline{\mathbf{X}}_\mathbf{t})' {\mathbf{S_t}}^{-1} (\mathbf{Z} - \overline{\mathbf{X}}_\mathbf{t}) + \log(det(\mathbf{S_t})/(\hat{p}_t^2)) \right], \end{aligned} \tag{5}$$

is obtained by plugging the estimates $\overline{\mathbf{X}}_\mathbf{j}$, $\mathbf{S_j}$ and $\hat{p}_j$ of the mean, covariance matrix and the probability p_j, into the Bayes classification rule (here log denotes the logarithm to the base e). It is well-known (cf. [1], pp. 321-322), that the mentioned Bayes rule minimizes the probability PTE of the total error; the rule (5) will be referred to as *EBN* (*empirical Bayes normal rule*). Since the plugged estimates are consistent, it is logical to use (5) provided that the sampled populations Π_1, Π_2 are normally distributed. Thus, in this setting the parameters of the populations are unknown, but their form is known. Since the knowledge of the type of the distribution is not always at disposal, the inference on classification is in practice sometimes based on the EBN rule even if the assumption of the normality is not fulfilled. Aspects of the use of the EBN rule are investigated in the next sections by means of simulation results in connection with some non-parametric procedures.

As the true distribution of Π_1, Π_2 is not always known (this occurs often when some biological phenomena are observed), methods constructed without a reference to the type of distribution are often useful. To this class of nonparametric rules belongs the WR rule, constructed in Section 2. One of the most known nonparametric rules is the *nearest neighbour rule*, which will be referred to as *NN*. This rule assigns the observation to the population which has a larger number of observations amongst the ones that are closest to the classified vector. More precisely, the experimenter chooses a positive odd integer $K < n_1 + n_2$. Further, let

$$v^{(1)} \le v^{(2)} \le \ldots \le v^{(n_1+n_2)}$$

denote the ordering of the numbers $\{\|\mathbf{Z} - \mathbf{X_{ji}}\|;\ j = 1, 2\,,\ i = 1, \ldots, n_j\}$ according to their magnitude. Put

$$h(1) = \#\Big\{t \in \{1, \ldots, K\};\ v^{(t)} = \|\mathbf{Z} - \mathbf{X}_{1i}\|\,,\ 1 \le i \le n_1\Big\}$$

and

$$h(2) = \#\Big\{t \in \{1, \ldots, K\};\ v^{(t)} = \|\mathbf{Z} - \mathbf{X}_{2i}\|\,,\ 1 \le i \le n_2\Big\}\,,$$

i.e., $h(j)$ denotes the number of observations from the jth populations amongst the K closest. If $h(1) > h(2)$, then $\mathbf{Z}$ is assigned to Π_1, if $h(1) < h(2)$, then it is assigned to Π_2. Since the number K is odd, this procedure has always a unique result (it cannot occur that $h(1) = h(2)$). Our experience from simulations shows that for sample sizes not exceeding several hundred, the choice of K ranging approximately from $0.05(n_1 + n2)$ to $0.1(n_1 + n_2)$ usually yields good results.

4. Some Simulation Results

The effect of the rules was investigated by means of simulation estimates always obtained from $N = 5000$ trials. In the case that there was only a tiny overlap between the populations, the mentioned procedures yielded only mildly different results and therefore we focus on situations with larger probabilities of error. In the following tables we present several simulation estimates of the probabilities of the error, each case representing some particular type of the situation, differing mainly in the behaviour of the tail probabilities of the underlying distributions. The best result for the given distribution and sample sizes is printed in boldface letters.

Case 1 is a sampling from the normal distribution $\Pi_1 = N_3(\mu_1, \Sigma_1)$ and $\Pi_2 = N_3(\mu_2, \Sigma_2)$ where $\mu_1 = (9, 8, 10)'$, $\Sigma_1 = diag(2.3, 3, 4.2)$, $\mu_2 = (8, 6, 11)'$ and $\Sigma_2 = diag(3, 4.2, 2)$. Case 2 is a sampling from the Cauchy distributions $\Pi_1 = C_3(\mu_1, \Sigma_1)$ and $\Pi_2 = C_3(\mu_2, \Sigma_2)$ with independent components, location parameters $\mu_1 = (29, 32, 8)'$, $\mu_2 = (27.5, 21, 16)'$ and the scale parameter matrices $\Sigma_1 = diag(1.3, 2.6, 2.3)^{1/2}$, $\Sigma_2 = diag(2.9, 3, 1.7)^{1/2}$. Case 3 is a sampling from the distribution of $\Sigma\varepsilon + \mu$, where ε has independent components, each having the Pareto density $f(x) = (m-1)/x^m$ if $x \ge 1$ and 0 elsewhere (with $m = 1.85$), and the population Π_j, $j = 1, 2$, has the location parameters $\mu_1 = (29, 32, 4)'$, $\mu_2 = (27.5, 7.5, 30.8)'$, and the scale parameter matrices $\sigma_1 = diag(1.1, 2, 2)^{1/2}$, $\sigma_2 = diag(2.9, 3, 1.5)^{1/2}$. As usual, $diag$ denotes here the diagonal matrix with the given diagonal.

Case 1	$n_1 = 80$ $n_2 = 60$			
	$P(2\|1)$	$P(1\|2)$	PTE	MPE
EBN		0.186	0.317	**0.242** 0.317
	K			
NN	7	0.213	0.341	0.268 0.341
NN	11	0.195	0.335	0.255 0.335
	we			
WR	1	0.253	0.247	0.250 **0.253**
WR	n_2/n	0.168	0.345	0.244 0.345
WR	n_2/n_1	0.219	0.282	0.246 0.282

Case 1	$n_1 = 160$ $n_2 = 60$			
	$P(2\|1)$	$P(1\|2)$	PTE	MPE
EBN		0.080	0.488	**0.191** 0.488
	K			
NN	11	0.076	0.540	0.202 0.540
NN	17	0.062	0.559	0.198 0.559
	we			
WR	1	0.248	0.245	0.247 **0.248**
WR	n_2/n	0.124	0.409	0.202 0.409
	n_2/n_1	0.152	0.364	0.210 0.364

Case 2	$n_1 = 80$ $n_2 = 60$			
	$P(2\|1)$	$P(1\|2)$	PTE	MPE
EBN		0.308	0.417	0.355 0.417
	K			
NN	7	0.092	0.118	0.103 0.118
NN	11	0.097	0.110	**0.103 0.110**
	we			
WR	1	0.142	0.152	0.146 0.152
WR	n_2/n	0.099	0.221	0.151 0.221
WR	n_2/n_1	0.121	0.178	0.146 0.178

Case 2	$n_1 = 160$ $n_2 = 60$			
	$P(2\|1)$	$P(1\|2)$	PTE	MPE
EBN		0.337	0.450	0.377 0.450
	K			
	11	0.054	0.161	**0.084** 0.161
	17	0.066	0.153	0.090 **0.153**
	we			
WR	1	0.136	0.172	0.145 0.172
WR	n_2/n	0.076	0.273	0.130 0.273
WR	n_2/n_1	0.086	0.243	0.129 0.243

Case 3	$n_1 = 80$ $n_2 = 30$			
	$P(2\|1)$	$P(1\|2)$	PTE	MPE
EBN		0.219	0.371	0.260 0.371
	K			
NN	7	0.036	0.155	**0.068** 0.155
NN	9	0.041	0.149	0.070 0.149
	we			
WR	1	0.113	0.122	0.116 **0.122**
WR	n_2/n	0.054	0.230	0.102 0.230
WR	n_2/n_1	0.058	0.203	0.098 0.203

5. Discussion and conclusions

It can be seen from the tables that for unbalanced sampling the use of the rule which is best from the point of view of PTE may result in an unacceptably high MPE (the case 1 when MPE equals 0.488), whereas a classification rule with a mildly worse PTE may yield strikingly better MPE. Thus, while the EBN rule remains the best choice for normal populations as far as the PTE is concerned, this rule should not be used either for strongly unbalanced sample sizes or for populations differing from the Gaussian type; in such cases, the use of nonparametric competitors can be recommended. If the normality of the distributions is questionable and one aims to reduce the PTE, then either the NN method or the WR method (proposed in the previous text) with a suitably chosen weight can be used. If, however, the populations are considered to be of equal importance, then, with the exception of approximately balanced sampling from normal distributions, the use of the WR method with we=1 is recommendable because its MPE attains good values under various circumstances. It should be noted here that distributions not resembling any theoretical model often occur in practice and then a comparison of the WR method with various weights and with other mentioned competitors, based on cross-validation, can be useful.

Acknowledgement

This reserch was supported by the grant VEGA 1/3016/26 from the Scientific Grant Agency of the Slovak Republic.

REFERENCES

[1] Anděl, J.(1985). *Matematická statistika.* Praha: SNTL/ALFA. (In Czech)

[2] Broffitt, J. D., Randles, R. H., Hogg, R. V. (1976). Distribution-free partial discriminant analysis. *JASA* 71(356), 934 - 939.

[3] Randles, R. H., Broffitt, J. D., Ramberg, J. S. Hogg, R. V. (1978). Discriminant analysis based on ranks. *JASA* 73(362), 379 - 384.

[4] Seber, G. A. F. (1984). *Multivariate Observations.* New York: Wiley & Sons.

[5] Witkovský, V., Rublík, F., Arendacká, B., Grendár, M., Farkaš, I., Bajla, I., Holländer, I. (2005). *Alternative approaches to band classification in Epo images for the GASepo software.* Seibersdorf, Austria: Austrian Research Centers.