Optimal discriminant functions for normal populations

Hirofumi Wakaki [a], Makoto Aoshima [b]

[a] *Department of Mathematics, Hiroshima University, 1-3-1, Kagamiyama, Higashi–Hiroshima 739-8526, Japan*

[b] *Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan*

Abstract

A class of discriminant rules which includes the Fisher's linear discriminant function and the likelihood ratio criterion are defined. Using asymptotic expansions of the distributions of the discriminant functions in this class, we derive a formula of cut-off points which satisfy some conditions on misclassification probabilities, and derive the optimal rules for some criteria. Some numerical experiments are carried out to examine the performance of the optimal rules for finite numbers of samples.

Key words: linear discriminant function, W–rule, Z–rule, asymptotic expansion.
AMS 2000 subject classifications : 62H30, 62H20

1 Introduction

We consider a problem of classifying an observation vector $\boldsymbol{x}$ into one of two normal populations $\Pi_1 : N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\Pi_2 : N_p(\boldsymbol{\mu}_2, \Sigma)$, where $\boldsymbol{\mu}_i$ is the mean vector of $\Pi_i\,(i = 1, 2)$ and Σ is the common covariance matrix. Suppose that the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$ are unknown and the training samples

$$\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{iN_i} \quad (i = 1, 2)$$

from Π_i are available. Let the sample mean vectors $\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2$ and the pooled sample covariance matrix S be given by

$$\overline{\boldsymbol{x}}_1 = \frac{1}{N_i}\sum_{j=1}^{N_i} \boldsymbol{x}_{ij}\,(i = 1, 2), \quad S = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)',$$

where $n = N_1 + N_2 - 2$. Then the W–rule is based on the statistic

$$[\boldsymbol{x} - \tfrac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)]' S^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2),$$

which is proposed by Wald [9] and Anderson [1,3]. The Z–rule was introduced by Kudo [6,7] and John [5] as a competitor to the Wald–Anderson's W-rule. The Z–rule is based on the statistic

$$\frac{N_1}{N_1+1}(\boldsymbol{x}-\overline{\boldsymbol{x}}_1)'S^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}}_1)-\frac{N_2}{N_2+1}(\boldsymbol{x}-\overline{\boldsymbol{x}}_2)'S^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}}_2). \tag{1.1}$$

Das Gupta [4] showed that the Z–rule is minimax in the class of invariant classification rules for certain type of risk functions. Let Δ be the Mahalanobis distance between two populations :

$$\Delta^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).$$

Let $P_i(\phi)$ be the probability that a random vector $\boldsymbol{x}$ from Π_i $(i=1,2)$ is misclassified by a classification rule ϕ. Then the risk considered by Das Gupta is given by

$$risk(\phi)=\tfrac{1}{2}\left\{l\left(\frac{N_1}{N_1+1}\Delta\right)P_1(\phi)+l\left(\frac{N_2}{N_2+1}\Delta\right)P_2(\phi)\right\},$$

where $l()$ is a certain function defined on $(0,\infty)$. It seems that the factor $\frac{N_i}{N_i+1}$ in the loss function $l()$ is required only for the theory. In this paper we consider natural criteria. Let π_i be the prior probability that $\boldsymbol{x}$ comes from Π_i, and let c_i be the cost of misclassification of $\boldsymbol{x}$ which comes from Π_i. If the prior probabilities are known the risk of a classification rule ϕ is defined as the expected cost of misclassification :

$$risk_1(\phi)=c_1\pi_1P_1(\phi)+c_2\pi_2P_2(\phi), \tag{1.2}$$

which is called *the total risk* in the following sections. Our interest is in whether the Z–rule is still optimal for this risk, and how we can find a classification rule superior to both the W–rule and the Z–rule if these rules are not optimal.

It is difficult to derive the exact values of $risk_1$ for the W–rule and the Z–rule since the exact distribution functions of their discriminant functions are very complicated. One way of comparing the performance of these rules is to approximate the risks by using asymptotic expansions when the sample sizes tend to infinity. Moreover we can find a classification rule which is superior to both the W–rule and the Z–rule by deriving the asymptotic expansion formula of the risk in a certain class of classification rules which includes both the W–rule and the Z–rule as in the following sections.

If the prior probabilities are unknown, we consider minimax criterion:

$$\begin{aligned} risk_2(\phi) &= \max\{c_1\pi_1P_1(\phi)+c_2\pi_2P_2(\phi)\mid 0\le\pi_1\le 1,\pi_1+\pi_2=1\} \\ &= \max\{c_1P_1(\phi),c_2P_2(\phi)\}. \end{aligned} \tag{1.3}$$

If the misclassification of $\boldsymbol{x}$ which comes from Π_1 is serious, one may require to control the misclassification probability P_1. In such case the problem is to find the classification rule which minimizes $P_2(\phi)$ under the condition that

$$P_1(\phi) \leq \alpha$$

for a given constant α. In section 4, we treat the problem of this type as well as the problem of finding the optimal rules with respect to the minimax criterion in the class of classfication rules defined in section 2.

When the sample sizes are large relative to the dimension, the differences among the classification rules are small since the classification rules considered in this paper are asymptotically equivalent. Therefore the new method derived in this paper will be useful when the sample sizes are small and the dimension is relatively large in practical point of view. We show some results of numerical experiments in section 5.

2 Class of discriminant functions

First we prepare some notations. For arbitrary two p dimensional vector $\boldsymbol{x} = (x_i), \boldsymbol{y} = (y_i)$ and arbitrary symmetric matrix $A = (a_{ij})$ of size p, we denote the $m = 2p + p(p+1)/2$ dimensional vector of elements in $\boldsymbol{x}, \boldsymbol{y}$ and A without redundancy as

$$< \boldsymbol{x}, \boldsymbol{y}, A >= (x_1, \cdots, x_p, y_1, \cdots, y_p, a_{11}, a_{22}, \cdots, a_{pp}, a_{12}, a_{13}, \cdots, a_{p-1,p})'.$$

When A is nonsingular, the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$ associated with A is denoted as

$$q(\boldsymbol{x}, \boldsymbol{y}, A) = (\boldsymbol{x} - \boldsymbol{y})' A^{-1} (\boldsymbol{x} - \boldsymbol{y}),$$

which is often simply denoted as $q(\boldsymbol{t})$ for $\boldsymbol{t} =< \boldsymbol{x}, \boldsymbol{y}, A >$. Now, we define a class of discriminant functions by generalizing the W–rule and the Z–rule. If $\boldsymbol{\theta} =< \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma >$ is known, Bayes rule is based on difference between Mahalanobis distances of $\boldsymbol{x}$ from two populations :

$$q(\boldsymbol{x}, \boldsymbol{\mu}_1, \Sigma) - q(\boldsymbol{x}, \boldsymbol{\mu}_2, \Sigma).$$

The W–rule is given by only replacing $q(\boldsymbol{x}, \boldsymbol{\mu}_i, \Sigma)$ with $q(\boldsymbol{x}, \overline{\boldsymbol{x}}_i, S)$. While the Z–rule multiplies the weighting term $N_i/(N_i + 1)$ to $q(\boldsymbol{x}, \overline{\boldsymbol{x}}_i, S)$ before taking the difference. It seems natural to attach the weighting terms because the performance of the estimated Mahalanobis distance depends on the sample sizes. Our problem is not to estimate the Mahalanobis distance, but to obtain good classification rules with respect to the risks given in §1. The best weighting terms may depend on the risk function. Therefore we consider the rule based

on the inequality :

$$d_a(\boldsymbol{x};\boldsymbol{T}) := \frac{1}{2}\Big\{(1+a)q(\boldsymbol{x},\overline{\boldsymbol{x}}_1,S) - (1-a)q(\boldsymbol{x},\overline{\boldsymbol{x}}_2,S)\Big\} \leq b, \tag{2.1}$$

where $\boldsymbol{T} =< \overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2, S >$. Note that the form of the pair of weights $1+a$ and $1-a$ is not restrictive because the inequality $a_1 q(\boldsymbol{x},\overline{\boldsymbol{x}}_1,S) - a_2 q(\boldsymbol{x},\overline{\boldsymbol{x}}_2,S) \leq c$ is equivalent with the above inequality with $a = (a_1 - a_2)/(a_1 + a_2)$ and $b = c/(a_1 + a_2)$ if a_1 and a_2 are positive.

Let $\phi_{a,b}(\boldsymbol{T})$ be the rule to classify $\boldsymbol{x}$ into Π_1 if (2.1) holds, and to classify $\boldsymbol{x}$ into Π_2 otherwise.

Consider to minimize the $risk_1$ given by (1.2). Let

$$c_0 = \frac{c_2\pi_2}{c_1\pi_1}. \tag{2.2}$$

Then minimizing $risk_1(\phi_{a,b}(\boldsymbol{T}))$ is equivalent with minimizing

$$P(a,b) := P_1(\phi_{a,b}(\boldsymbol{T})) + c_0 P_2(\phi_{a,b}(\boldsymbol{T})). \tag{2.3}$$

Note that the above probabilities P_1 and P_2 are with respect to the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{T}$. Since the exact distribution function of $d_a(\boldsymbol{x};\boldsymbol{T})$ is too complicated to handle, we consider to approximate the misclassification probabilities by using asymptotic expansions up to the order n^{-2}, where $n = N_1 + N_2 - 2$. We assume that N_1/N_2 tends to some positive constant when $n \to \infty$.

First we consider a and b as constants. Then $P(a,b)$ defined by (2.3) can be expanded as

$$P(a,b) = R_0 + \frac{1}{n}R_1 + \frac{1}{n^2}R_2 + O(n^{-3}),$$

where R_0, R_1 and R_2 are the functions of $a, b, c_0, \Delta, \sqrt{r_1}, \sqrt{r_2}$, with $r_i = n/N_i$ $(i = 1, 2)$. Neglecting the terms of the order $O(n^{-3})$, the optimal values of a and b are obtained as the solution of the system of equations :

$$\begin{aligned} s_1\Big(a,b,\frac{1}{n}\Big) &:= \frac{\partial R_0}{\partial a} + \frac{1}{n}\frac{\partial R_1}{\partial a} + \frac{1}{n^2}\frac{\partial R_2}{\partial a} = 0, \\ s_2\Big(a,b,\frac{1}{n}\Big) &:= \frac{\partial R_0}{\partial b} + \frac{1}{n}\frac{\partial R_1}{\partial b} + \frac{1}{n^2}\frac{\partial R_2}{\partial b} = 0. \end{aligned}$$

Since the limiting value R_0 corresponds to the risk of the Bayes rule $d_0(\boldsymbol{x};\boldsymbol{\theta})$ which includes unknown parameters, R_0 is minimized at $(a,b) = (0, b_0)$, where $b_0 = -\log c_0$. Hence $s_1(0, b_0, 0) = s_2(0, b_0, 0) = 0$. Therefore, the theorem of

implicit function will show that the optimal values of a and b can be expanded as

$$a_{opt} = 0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots, \quad b_{opt} = b_0 + \frac{b_1}{n} + \frac{b_2}{n^2} + \cdots, \tag{2.4}$$

where a_1, a_2, b_1 and b_2 are the functions of c_0, r_1, r_2 and Δ. Since Δ is unknown and should be estimated, we consider a class of classification rules

$$\mathcal{C} = \{\phi_{\alpha,\beta} : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\} \tag{2.5}$$

where $\mathcal{A}$ is the set of all functions given by

$$\alpha(D^2) = \tfrac{1}{n}\alpha_1(D^2) + \tfrac{1}{n^2}\alpha_2(D^2)$$

with arbitrary C^1-class function $\alpha_1(\,)$ and continuous function $\alpha_2(\,)$, and $\mathcal{B}$ is the set of all functions given by

$$\beta(D^2) = \beta_0(D^2) + \tfrac{1}{n}\beta_1(D^2) + \tfrac{1}{n^2}\beta_2(D^2)$$

with arbitrary C^2-class function $\beta_0(\,)$, C^1-class function $\beta_1(\,)$ and continuous function $\beta_2(\,)$. Here, $\phi_{\alpha,\beta}$ is the classification rule that classify $\boldsymbol{x}$ into Π_1 if

$$\frac{1}{2}\Big\{(1+\alpha(D^2))q(\boldsymbol{x}, \overline{\boldsymbol{x}}_1, S) - (1-\alpha(D^2))q(\boldsymbol{x}, \overline{\boldsymbol{x}}_2, S)\Big\} \le \beta(D^2) \tag{2.6}$$

and classify $\boldsymbol{x}$ into Π_2 otherwise, where $D^2 = q(\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2, S)$.

3 Minimizing the total risk

In this section we consider to minimize $risk_1$ in the class $\mathcal{C}$ of classification rules given by (2.5). It is equivalent with minimizing

$$P(\alpha, \beta) = P_1(\phi_{\alpha,\beta}) + c_0 P_2(\phi_{\alpha,\beta})$$

where c_0 is given by (2.2). As $n \to \infty$, $P(\alpha, \beta)$ converges to $P(0, \beta_0(\Delta^2))$ which has the minimum at

$$\beta_0(\Delta^2) \equiv b_0 = -\log c_0. \tag{3.1}$$

In order to make the problem simple, we assume without any loss of generality that $\Sigma = I_p$ and $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = (\frac{1}{2}\Delta, 0, \cdots, 0)'$ because the joint distribution of $q(\boldsymbol{x}, \overline{\boldsymbol{x}}_1, S), q(\boldsymbol{x}, \overline{\boldsymbol{x}}_2, S)$ and D^2 is invariant under the group of Affine transformations:

$$\boldsymbol{x} \mapsto A\boldsymbol{x} + \boldsymbol{b},\ \boldsymbol{x}_{ij} \mapsto A\boldsymbol{x}_{ij} + \boldsymbol{b}\ (i = 1, 2;\ j = 1, \cdots, N_i)$$

with arbitrary nonsingular matrix A and vector $\boldsymbol{b}$.

Let the conditional distribution function of $d_a(\boldsymbol{x};\boldsymbol{T})$ given $\boldsymbol{T}=\boldsymbol{t}$ for $\boldsymbol{x}$ which comes from Π_i be denoted as

$$F_i(y;a,\boldsymbol{t},\Delta)=\Pr\{d_a(\boldsymbol{x};\boldsymbol{T})\le y|T=\boldsymbol{t},\boldsymbol{x}\sim\Pi_i\}\quad(i=1,2). \tag{3.2}$$

Let

$$Q_c(\boldsymbol{t},a,b;\Delta)=1-F_1(b;a,\boldsymbol{t},\Delta)+cF_2(b;a,\boldsymbol{t},\Delta). \tag{3.3}$$

Then the risk is represented as $P(\alpha,\beta)=RE[Q_{c_0}(\boldsymbol{T},\alpha(D^2),\beta(D^2);\Delta)]$. In the following, we simply denote Q_{c_0} as Q_0.

The difference between the risks of $\phi_{\alpha,\beta}$ and the plug-in rule ϕ_{0,b_0} can be expanded as in the following lemma.

Lemma 1.

$$\begin{aligned}
&P(\alpha,\beta)-P(0,b_0)\\
&\quad=\frac{1}{n^2}\Big\{\alpha_1'(\Delta^2)q^{(\boldsymbol{t})}(\Delta)'J(\Delta)Q_0^{(a,\boldsymbol{t})}(\Delta)+\beta_1'(\Delta^2)q^{(\boldsymbol{t})}(\Delta)'J(\Delta)Q_0^{(b,\boldsymbol{t})}(\Delta)\\
&\qquad+\frac{1}{2}\mathrm{tr}\Big[\Big(\alpha_1(\Delta^2)Q_0^{(a,\boldsymbol{t},\boldsymbol{t}')}(\Delta)+\beta_1(\Delta^2)Q_0^{(b,\boldsymbol{t},\boldsymbol{t}')}(\Delta)\Big)J(\Delta)\Big]\\
&\qquad+\frac{1}{2}\Big(\alpha_1(\Delta^2)^2Q_0^{(a,a)}(\Delta)+2\alpha_1(\Delta^2)\beta_1(\Delta^2)Q_0^{(a,b)}(\Delta)\\
&\qquad\qquad+\beta_1(\Delta^2)^2Q_0^{(b,b)}(\Delta)\Big)\Big\}+O(n^{-3}),
\end{aligned} \tag{3.4}$$

where

$$J(\Delta)=RCov[\sqrt{n}(\boldsymbol{T}-\boldsymbol{\theta})],$$

α_1' *and* β_1' *are the derivatives of* α_1 *and* β_1*, respectively,* $Q_0^{(a,a)}(\Delta),Q_0^{(a,b)}(\Delta)$*, and* $Q_0^{(b,b)}(\Delta)$ *are scalars,* $q^{(\boldsymbol{t})}(\Delta),Q_0^{(a,\boldsymbol{t})}(\Delta)$ *and* $Q_0^{(b,\boldsymbol{t})}(\Delta)$ *are* $(q\times 1)$*-vectors,* $Q_0^{(a,\boldsymbol{t},\boldsymbol{t}')}(\Delta)$ *and* $Q_0^{(b,\boldsymbol{t},\boldsymbol{t}')}(\Delta)$ *are* $(q\times q)$*-matrices given by*

$$\begin{aligned}
&q^{(\boldsymbol{t})}(\Delta)=\frac{\partial}{\partial\boldsymbol{t}}q(\boldsymbol{t})|_0,\ Q_0^{(a,a)}(\Delta)=\frac{\partial^2}{(\partial a)^2}Q_0(\boldsymbol{t},a,b;\Delta)|_0,\\
&Q_0^{(a,b)}(\Delta)=\frac{\partial^2}{\partial a\partial b}Q_0(\boldsymbol{t},a,b;\Delta)|_0,\ Q_0^{(b,b)}(\Delta)=\frac{\partial^2}{(\partial b)^2}Q_0(\boldsymbol{t},a,b;\Delta)|_0,\\
&Q_0^{(a,\boldsymbol{t})}(\Delta)=\frac{\partial^2}{\partial a\partial\boldsymbol{t}}Q_0(\boldsymbol{t},a,b;\Delta)|_0,\ Q_0^{(b,\boldsymbol{t})}(\Delta)=\frac{\partial^2}{\partial b\partial\boldsymbol{t}}Q_0(\boldsymbol{t},a,b;\Delta)|_0,\\
&Q_0^{(a,\boldsymbol{t},\boldsymbol{t}')}(\Delta)=\frac{\partial^3}{\partial a\partial\boldsymbol{t}\partial\boldsymbol{t}'}Q_0(\boldsymbol{t},a,b;\Delta)|_0,\ Q_0^{(b,\boldsymbol{t},\boldsymbol{t}')}(\Delta)=\frac{\partial^3}{\partial b\partial\boldsymbol{t}\partial\boldsymbol{t}'}Q_0(\boldsymbol{t},a,b;\Delta)|_0.
\end{aligned}$$

Here $|_0$ *denotes that the derivatives are evaluated at* $\boldsymbol{t}=\boldsymbol{\theta}$ *or* $(\boldsymbol{t},a,b)=(\boldsymbol{\theta},0,b_0)$.

Proof. It holds that

$$\frac{\partial}{\partial h} Q_c(\boldsymbol{t}, a, b; \Delta)|_c = 0 \quad (h = \boldsymbol{t}, a, b), \tag{3.5}$$

where $|_c$ denotes that the derivative is evaluated at $(\boldsymbol{t}, a, b) = (\boldsymbol{\theta}, 0, -\log c)$, because $Q_c(\boldsymbol{t}, a, b; \Delta)$ has the minimum at that point. (3.4) is given by the Taylor expansion of $Q_0(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta)$ at $\boldsymbol{T} = \boldsymbol{\theta}$ using (3.5) followed by taking expectations term by term. □

It is difficult to find the best choice of α_1 and β_1 such that (3.4) is minimized for all Δ since (3.4) includes $\alpha_1'(\Delta^2)$. However, we can find α_1 and β_1 which improve the plug-in rule as in the following lemma.

Lemma 2. *Let γ be defined by*

$$\gamma(\Delta^2) = -\frac{q^{(\boldsymbol{t})}(\Delta)' J(\Delta) Q_0^{(a,\boldsymbol{t})}(\Delta)}{q^{(\boldsymbol{t})}(\Delta)' J(\Delta) Q_0^{(b,\boldsymbol{t})}(\Delta)}, \tag{3.6}$$

let

$$B_1(\Delta^2; \alpha_1, \tilde{\beta}_1, \gamma) = \gamma(\Delta^2)\alpha_1(\Delta^2) + \tilde{\beta}_1(\Delta^2)$$

for arbitrarily chosen function $\tilde{\beta}_1$, and set

$$\beta_{\alpha_1}(\Delta^2) = b_0 + \frac{1}{n} B(\boldsymbol{t}; \alpha_1, \tilde{\beta}_1, \gamma). \tag{3.7}$$

Then neglecting the terms of order $O(n^{-3})$, $P(\alpha_1, \beta_{\alpha_1})$ is minimized at

$$\begin{aligned}\alpha_1(\Delta^2) = &-\tfrac{1}{2}\{Q_0^{(a,a)}(\Delta) + 2Q_0^{(a,b)}(\Delta)\gamma(\Delta^2) + Q_0^{(b,b)}(\Delta)\gamma(\Delta^2)^2\}^{-1} \\ &\Big\{2\tilde{\beta}_1(\Delta^2)(Q_0^{(a,b)}(\Delta) + Q_0^{(b,b)}(\Delta)\gamma(\Delta^2)) \\ &\quad + \mathrm{tr}[J(\Delta)(Q_0^{(a,\boldsymbol{t},\boldsymbol{t}')}(\Delta) + Q_0^{(b,\boldsymbol{t},\boldsymbol{t}')}(\Delta)\gamma(\Delta^2) + 2\gamma'(\Delta^2)Q_0^{(b,\boldsymbol{t})}(\Delta)q^{(\boldsymbol{t})}(\Delta)')]\Big\}.\end{aligned} \tag{3.8}$$

Proof. Substituting $\beta_1(\Delta^2) = B_1(\Delta^2; \alpha_1, \tilde{\beta}_1, \gamma)$ to (3.4), we obtain

$$
\begin{aligned}
&RE_{\boldsymbol{T}}[Q(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta)] - RE_{\boldsymbol{T}}[Q(\boldsymbol{T}, 0, \beta_0; \Delta)] \\
&= \tfrac{1}{2}\{\tilde{\beta}_1(\Delta^2)^2 Q_0^{(b,b)}(\Delta) + \tilde{\beta}_1(\Delta^2)\mathrm{tr}[J(\Delta)Q_0^{(b,\boldsymbol{t},\boldsymbol{t}')}(\Delta)]\} \\
&\quad + (\tilde{\beta}_1{}'(\Delta^2))q^{(\boldsymbol{t})}(\Delta)' J(\Delta) Q_0^{(b,\boldsymbol{t})}(\Delta) \\
&\quad + \tfrac{1}{2}\{\alpha_1(\Delta^2)^2 (Q_0^{(a,a)}(\Delta) + 2Q_0^{(a,b)}(\Delta)\gamma(\Delta^2) + Q_0^{(b,b)}(\Delta)\gamma(\Delta^2)^2)\} \\
&\quad + \tfrac{1}{2}\alpha_1(\Delta^2)\Big\{2\tilde{\beta}_1(\Delta^2)(Q_0^{(a,b)}(\Delta) + Q_0^{(b,b)}(\Delta)\gamma(\Delta^2)) \\
&\qquad + \mathrm{tr}[J(\Delta)(Q_0^{(a,\boldsymbol{t},\boldsymbol{t}')}(\Delta) + Q_0^{(b,\boldsymbol{t},\boldsymbol{t}')}(\Delta)\gamma(\Delta^2) + 2\gamma'(\Delta^2)Q_0^{(b,\boldsymbol{t})}(\Delta)q^{(\boldsymbol{t})}(\Delta)')]\Big\} \\
&\quad + O(n^{-3}),
\end{aligned}
\tag{3.9}
$$

where γ' and $\tilde{\beta}_1{}'$ are the derivatives of γ and $\tilde{\beta}_1$, respectively. (3.9) does not include the derivative α_1' and is the quadratic polynomial of $\alpha_1(\Delta^2)$, which has the minimum at $\alpha_1(\Delta^2)$ given by (3.8). □

Actual calculations of the derivatives using lemma 5 in appendix A, show that

$$
\gamma(\Delta^2) = p - 1 + \frac{b_0^2}{\Delta^2} + \frac{1}{4}\Delta^2 \tag{3.10}
$$

and

$$
Q_0^{(a,b)}(\Delta) + Q_0^{(b,b)}(\Delta)\gamma(\Delta^2) = 0, \tag{3.11}
$$

which shows that (3.8) does not depend on the choice of $\tilde{\beta}_1$, and the optimal function is given by

$$
\alpha_1(D^2) = \frac{r_2 - r_1 - 2b_0}{2}. \tag{3.12}
$$

The following theorem gives a way to improve an arbitrarily chosen classification rule in the class $\mathcal{C}$.

Theorem 3. *Let ϕ_{α^*,β^*} be a classification rule in the class $\mathcal{C}$ given by* (2.5) *where*

$$
\begin{aligned}
\alpha^*(D^2) &= \frac{1}{n}\alpha_1^*(D^2) + \frac{1}{n^2}\alpha_2^*(D^2), \\
\beta^*(D^2) &= b_0 + \frac{1}{n}\beta_1^*(D^2) + \frac{1}{n^2}\beta_2^*(D^2).
\end{aligned}
$$

Set

$$
\beta(D^2) = b_0 + \frac{1}{n}\Big\{\beta_1^*(D^2) + \gamma(D^2)[\alpha_1(D^2) - \alpha_1^*(D)^2]\Big\}. \tag{3.13}
$$

Then

$$
risk_1(\phi_{\alpha,\beta}) < risk_1(\phi_{\alpha^*,\beta^*})
$$

up to terms of $O(n^{-3})$*, where* $\alpha(D^2) = \frac{1}{n}\alpha_1(D^2)$.

Proof. Choose

$$\tilde{\beta}_1(D^2) = \beta_1^*(D^2) - \gamma(D^2)\alpha_1^*(D^2)$$

for (3.7). Then

$$\beta_{\alpha_1^*}(D^2) = \beta^*(D^2) \quad \text{and} \quad \beta_{\alpha_1}(D^2) = \beta(D^2).$$

Hence Lemma 2 leads the desired results. □

In the case of $c_1\pi_1 = c_2\pi_2$, the Z–rule classify $\boldsymbol{x}$ in to Π_1 if (1.1) is less than or equal to $b_0 = 0$. The inequality is equivalent with

$$\frac{1}{2}(1 + a_Z)q(\boldsymbol{x}, \overline{\boldsymbol{x}}_1, S) - \frac{1}{2}(1 - a_Z)q(\boldsymbol{x}, \overline{\boldsymbol{x}}_2, S) \leq 0$$

where

$$a_Z = \frac{N_1 - N_2}{2N_1N_2 + N} = \frac{r_2 - r_1}{2n} + O(n^{-2}),$$

which shows the optimality of the Z–rule in our framework.

In the case of $c_1\pi_1 \neq c_2\pi_2$, one may use the Z–rule with cut off point b_0 instead of 0 in the above inequality, since it is asymptotically optimal. However, this rule can be improved by using (3.12) and (3.13) with $\beta_1^*(D^2) = 0$.

Figure 1 illustrates the relationship of classification rules in our class. The horizontal and the vertical axes in the figure represent the sets of C^1–class functions for α_1 and β_1, respectively. Since α_2 and β_2 do not appear in (3.4), we identify a rule $\phi_{\alpha,\beta}$ with a point (α_1, β_1) in the plane. Then W–rule and Z–rule with cut off point b_0 can be represented by points $\mathrm{W}(0, 0)$ and $\mathrm{Z}(\frac{1}{2}(r_2 - r_1), 0)$, respectively. The line with ordinate intercept $\mathrm{B}(\tilde{\beta}_1, 0)$ and slope γ represents the subclass of classification rules

$$\Big\{\phi_{\alpha,\beta};\, \alpha(D^2) = \frac{1}{n}\alpha_1(D^2) + O\Big(\frac{1}{n^2}\Big),\ \beta(D^2) = b_0 + \frac{1}{n}B_1(\boldsymbol{t}; \alpha_1, \tilde{\beta}_1, \gamma) + O\Big(\frac{1}{n^2}\Big)\Big\}.$$

We have seen that the optimal rule in this subclass lies on the vertical line h through the point $\mathrm{A}(\frac{1}{2}(r_2 - r_1 - 2b_0), 0)$. Therefore the rule corresponding to the point C is superior to the W–rule, and the rule corresponding to the point D is superior to the Z–rule. We cannot find the best point on h. The superiority on h depends on the unknown parameter Δ.

Remark 1. When the training sample is coming from the same distribution as the data to classify, then the prior probabilities can also be estimated from the data. Similar approach can be applied to this problem, which is remained for future.

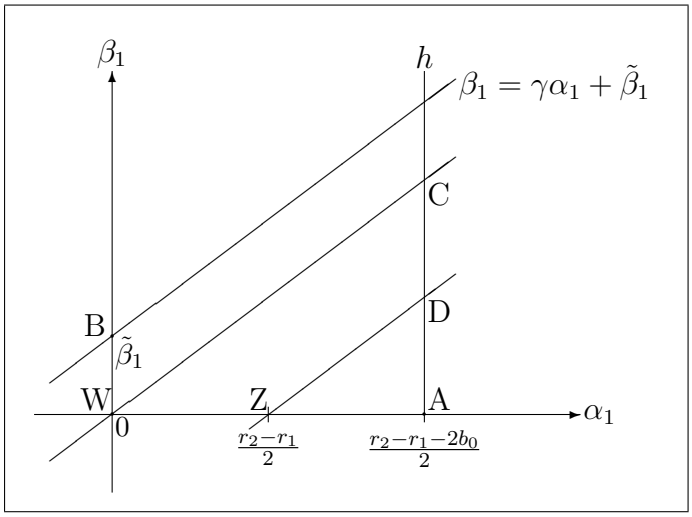


Fig. 1. Relationship of classification rules

Remark 2. Lemma 1 and 2 do not depend on the assumption of normality, but on the fact that the discriminant function converges to the optimal Bayes discriminant functions when $(a, b, \boldsymbol{t})$ converges to $(0, b_0, \boldsymbol{\theta})$. Suppose that the populations Π_1 and Π_2 are not normal and have the density functions $f_1(\boldsymbol{x}; \boldsymbol{\theta})$ and $f_2(\boldsymbol{x}; \boldsymbol{\theta})$, respectively, with known functions f_1 and f_2. The Bayes rule is based on

$$d_0(\boldsymbol{x}; \boldsymbol{\theta}) = \log f_1(\boldsymbol{x}; \boldsymbol{\theta}) - \log f_2(\boldsymbol{x}; \boldsymbol{\theta}),$$

which is simply estimated by $d_0(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$, the plug–in version, where $\hat{\boldsymbol{\theta}}$ is some consistent estimator of $\boldsymbol{\theta}$ based on the training samples. The Z–rule was derived as the likelihood ratio, treating the classfication problem as the testing problem for normal populations (see Kudo [6,7], John [5] or Anderson [1,3]). Let $d_z(\boldsymbol{x}; \boldsymbol{T})$ be logarithm of the likelihood ratio, where $\boldsymbol{T}$ is the training samples. Then we can define a class of discriminant functions :

$$d_a(\boldsymbol{x}; \boldsymbol{T}) := \Big\{1 - \frac{1}{n}a(\hat{\boldsymbol{\theta}})\Big\}d_0(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) + \frac{1}{n}a(\hat{\boldsymbol{\theta}})d_z(\boldsymbol{x}; \boldsymbol{T}) - b(\hat{\boldsymbol{\theta}}). \tag{3.14}$$

Lemma 2 will be applied to find a rule superior to both plug–in rule and the rule based on the likelihood ratio test.

4 Unknown prior probabilities

When the prior probabilities are unknown, one criterion of choosing classification rule is the minimax criterion. Let $\phi_{\alpha,\beta}$ be a classification rule in the class $\mathcal{C}$ defined by (2.5). If $c_1 P_1(\phi_{\alpha,\beta}) \neq c_2 P_2(\phi_{\alpha,\beta})$ we can reduce $risk_2(\phi_{\alpha,\beta})$ by decreasing or increasing β so as to decrease $|c_1 P_1(\phi_{\alpha,\beta}) - c_2 P_2(\phi_{\alpha,\beta})|$. Therefore

our problem of finding the optimal rule with respect to the minimax criterion is to minimize $P_2(\phi_{\alpha,\beta})$ under the condition

$$c_1 P_1(\phi_{\alpha,\beta}) - c_2 P_2(\phi_{\alpha,\beta}) = 0. \tag{4.1}$$

If the misclassification of $\boldsymbol{x}$ which comes from Π_1 is serious, one may require to control the misclassification probability P_1. In such case the problem is to minimize $P_2(\phi_{\alpha,\beta})$ under the condition that $P_1(\phi_{\alpha,\beta}) \leq u$ for specified constant u. If $P_1(\phi_{\alpha,\beta}) < u$, we can reduce $P_2(\phi_{\alpha,\beta})$ by decreasing β so as to make

$$P_1(\phi_{\alpha,\beta}) = u \tag{4.2}$$

hold.

The above two problems can be treated in the same manner. Consider the condition:

$$P_1(\phi_{\alpha,\beta}) - kP_2(\phi_{\alpha,\beta}) = u. \tag{4.3}$$

If we set $k = c_2/c_1, u = 0$ we obtain (4.1), and if $k = 0$ (4.3) corresponds to (4.2). Therefore our problem is to minimize $P_2(\phi_{\alpha,\beta})$ under the condition (4.3).

4.1 Derivation of the cut–off point

First we derive β. Since the limiting value of the left side of (4.3) is

$$1 - \Phi\Big(\frac{\beta_0}{\Delta} + \frac{\Delta}{2}\Big) - k\Phi\Big(\frac{\beta_0}{\Delta} - \frac{\Delta}{2}\Big) \tag{4.4}$$

we define $\beta_0(D^2)$ be the solution of

$$\beta_0 \ : 1 - \Phi\Big(\frac{1}{D}\beta_0 + \frac{D}{2}\Big) - k\,\Phi\Big(\frac{1}{D}\beta_0 - \frac{D}{2}\Big) = u. \tag{4.5}$$

Let

$$G(\boldsymbol{t}, a, b; \Delta) = 1 - F_1(b; a, \boldsymbol{t}, \Delta) - kF_2(b; a, \boldsymbol{t}, \Delta). \tag{4.6}$$

Then the left side of (4.3) with using (4.5) can be expanded as

$$\begin{aligned} &RE[G(\boldsymbol{T}, n^{-1}\alpha_1(D^2), \beta_0(D^2) + n^{-1}\beta_1(\Delta); \Delta)] \\ &\quad = u + n^{-1}\Big\{G_{01}(\Delta) + \alpha_1(\Delta)G^{(a)}(\Delta) + \beta_1(\Delta)G^{(b)}(\Delta)\Big\} + O(n^{-2}), \end{aligned}$$

where $G_{01}(\Delta)$ is a function of k, r_1, r_2 and Δ. Therefore, for each α_1,

$$\beta_1(D^2;\alpha_1) = -\{G^{(b)}(D)\}^{-1}\{G^{(a)}(D)\alpha_1(D^2) + G_{01}(D)\} \tag{4.7}$$

make (4.3) hold up to the order $O(n^{-1})$. Similarly, the left hand of (4.3) with (4.5) and (4.7) can be expanded as

$$\begin{aligned} &RE[G(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta)] \\ &\quad = u + n^{-2}\Big\{G_{02}(\Delta) + \alpha_2(\Delta)G^{(a)}(\Delta) + \beta_2(\Delta)G^{(b)}(\Delta)\Big\} + O(n^{-3}), \end{aligned}$$

where $G_{02}(\Delta)$ is a function of k, r_1, r_2, Δ and $\alpha_1(\Delta)$, which gives the third term of β so as to make (4.3) hold up to the order $O(n^{-2})$:

$$\beta_2(D^2;\alpha_1,\alpha_2) = -\{G^{(b)}(D)\}^{-1}\{G^{(a)}(D)\alpha_2(D^2) + G_{02}(D)\}. \tag{4.8}$$

Actual forms of β_1 and β_2 are given in Appendix B.

4.2 The optimal rule

Let $c(\Delta) = \exp\{-\beta_0(\Delta)\}$ with β_0 given by (4.5). Then under the condition (4.3), minimizing $P_2(\phi_{\alpha,\beta})$ is equivalent with minimizing
$RE[Q_{c(\Delta)}(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta)]$ since

$$\begin{aligned} &RE[Q_{c(\Delta)}(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta)] \\ &\quad = RE\Big[G(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta) + \{c(\Delta) + k\}F_2(\beta(D^2); \alpha(D^2), \boldsymbol{T}, \Delta)\Big] \\ &\quad = u + \{c(\Delta) + k\}P_2(\phi_{\alpha,\beta}). \end{aligned}$$

Theorem 4. *$P_2(\phi_{\alpha,\beta})$ has the minimum in $\mathcal{C}$ at β derived in the previous subsection, and $\alpha(\Delta^2) = \frac{1}{n}\alpha_1(\Delta^2)$ under the condition* (4.3)*, neglecting the terms of the order $O(n^{-3})$, where*

$$\alpha_1(\Delta^2) = \frac{1}{2}(r_2 - r_1 - 2\beta_0(\Delta^2)). \tag{4.9}$$

Proof. Expanding $Q_{c(\Delta)}(\boldsymbol{T}, 0, \beta(D^2); \Delta)$ at $\boldsymbol{T} = \boldsymbol{\theta}$ and then taking the expectations we obtain

$$\begin{aligned} &RE_{\boldsymbol{T}}[Q_{c(\Delta)}(\boldsymbol{T}, \alpha(D^2), \beta(D^2); \Delta)] - RE_{\boldsymbol{T}}[Q_{c(\Delta)}(\boldsymbol{T}, 0, \beta(D^2); \Delta)] \\ &\quad = \frac{1}{2n^2}\Big[\frac{2}{\Delta}(p-1)\phi(y_1)\Big\{\alpha_1(\Delta^2)^2 + \alpha_1(\Delta^2)\Big(r_1 - r_2 + 2\beta_0(\Delta^2)\Big)\Big\} \\ &\qquad + G_{03}(\Delta)\Big] + O(n^{-3}), \end{aligned} \tag{4.10}$$

which does not include $\alpha_1'(\Delta^2)$, and has the minimum at $\alpha_1(\Delta^2)$ given by (4.9). □

It is interesting that (4.9) has the same form as (3.12).

Remark 3. The result of Theorem 4 depends on the assumption of normality. In the case considered in Remark 2, the terms of the order $O(n^{-2})$ in the asymptotic expansion of P_2 for the classification rule given by (3.14) generally includes $a'(\boldsymbol{\theta})$. However, the method to derive the cut–off point given in subsection 4.1 can be applied.

5 Numerical studies

This section gives some results of Monte Carlo experiments to compare the new classification rules obtained in section 3 and section 4 with the W–rule and the Z–rule.

The values of N_1, N_2, p and Δ were chosen as follows:

$$
\begin{aligned}
&(N_1, N_2); (10,10), (10,15), (10,20), (15,15), (15,20), (20,20),\\
&p; 6, 8, 10, 12,\\
&\Delta; \Phi(-\Delta/2) = 0.1, 0.2, 0.3,
\end{aligned}
$$

The expected misclassification probabilities $P_1(\phi_{a,b}(\boldsymbol{T}))$, $P_2(\phi_{a,b}(\boldsymbol{T}))$ are estimated based on 1000,000 times of iteration. So the standard deviation is at most 0.5%. Here we used a pseudo–random number generator named *Mersenne Twister* which provides a period of $2^{19937} - 1$ and 623–dimensional equidistribution, and is sufficient for our purpose (see Matsumoto and Nishimura [8]).

5.1 The total risk

First we examine the total risk . We can assume that the costs c_1 and c_2 are equal to one. We compare the values of $\pi_1 P_1(\phi_{a,b}(\boldsymbol{T})) + \pi_2 P_2(\phi_{a,b}(\boldsymbol{T}))$ for the W–rule, the Z–rules and the optimal rules which corresponding to the points A, C and D in figure 1 in section 3 when $\pi_1 = 1/3$ and $1/2$.

Let $b_0 = -\log(\pi_2/\pi_1)$. Then the coefficients (a, b) for the classification rules

are

$$
\begin{aligned}
&\text{W–rule} : (a, b) = (0, b_0),\\
&\text{Z–rule} : (a, b) = (a_z, b_0),\\
&\text{Oo–rule} : (a, b) = (\tfrac{1}{n}a_o, b_0),\\
&\text{Ow–rule: } (a, b) = \Big(\tfrac{1}{n}a_o, b_0 + \tfrac{1}{n}\big(\gamma(D^2)a_o\big)\Big),\\
&\text{Oz–rule: } (a, b) = \Big(\tfrac{1}{n}a_o, b_0 - \tfrac{1}{n}\big(\gamma(D^2)b_0\big)\Big),
\end{aligned}
$$

where

$$
a_z = \frac{N_1 - N_2}{N_1 + N_2 + 2N_1N_2}, \quad a_o = \frac{r_2 - r_1}{2} - b_0,
$$

and $\gamma(D^2)$ is given by (3.10).

Figure 5.1 gives the values ($\times 100\%$) of $(P_1 + 2P_2)/3$ for the five rules when $p = 10, 12$. In the figure, q is the value of $(P_1 + 2P_2)/3$ for the Bayes rule, that is,

$$
q = \frac{1}{3}\Big\{1 - \Phi\Big(\frac{b_0}{\Delta} + \frac{\Delta}{2}\Big)\Big\} + \frac{2}{3}\Phi\Big(\frac{b_0}{\Delta} - \frac{\Delta}{2}\Big).
$$

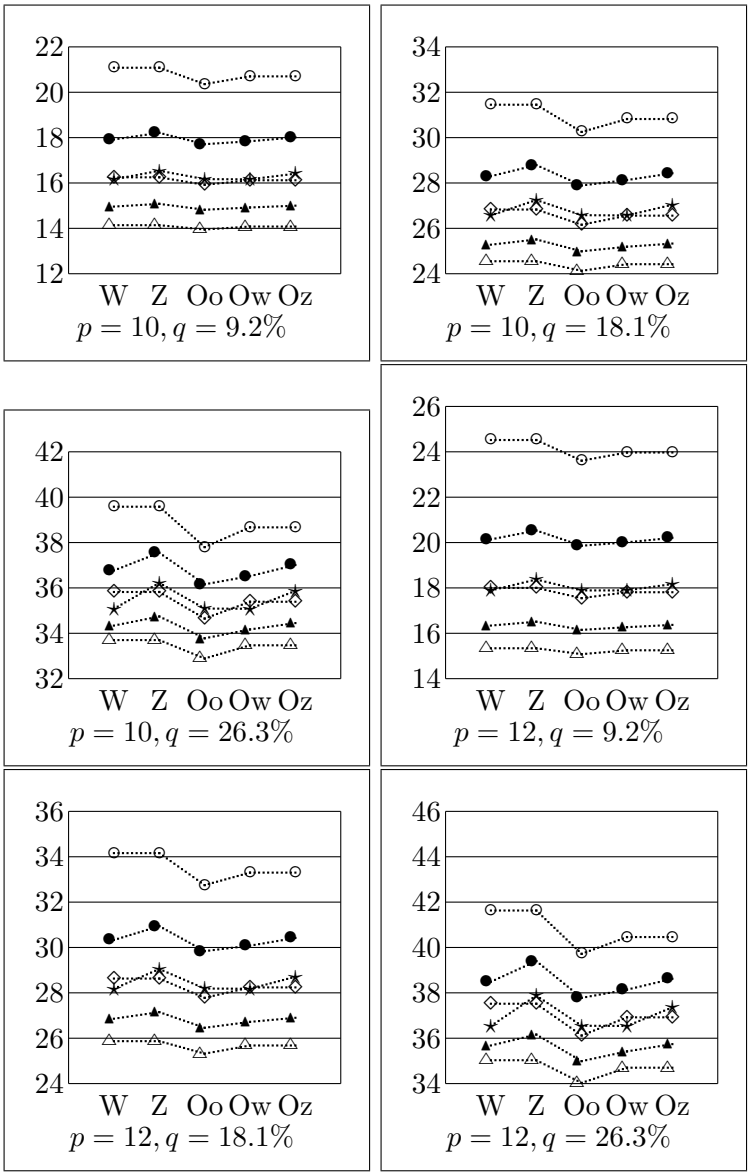
We can see that the Ow–rule performs better than the W–rule and the Oz–rule performs better than the Z–rule in all cases, and the Oo–rule has the best performance. We consider that the one of the reason of the superiority of the Oo–rule is that the coefficient a_o and the cut–off point b_0 do not depend on the samples.

When $\pi_1 = \pi_2$, the performance of the five rules were almost same. When $p = 6, 8$, we could see similar results, but the differences between the rules got smaller a little. When the sample sizes are small relative to the dimension or the Mahalanobis distance is small, the differences between the classification rules become clear. In that cases we recommend to use the Oo–rule.

5.2 The minimax criterion

We compare the values of $\max\{c_1P_1(\phi_{a,b}(\boldsymbol{T})), c_2P_2(\phi_{a,b}(\boldsymbol{T}))\}$ for five rules when $(c_1, c_2) = (1, 1)$ and $(1/2, 1)$.

Let β_0, β_1 and β_2 be defined by (4.5), (4.7) and (4.8), respectively. The coeffi-

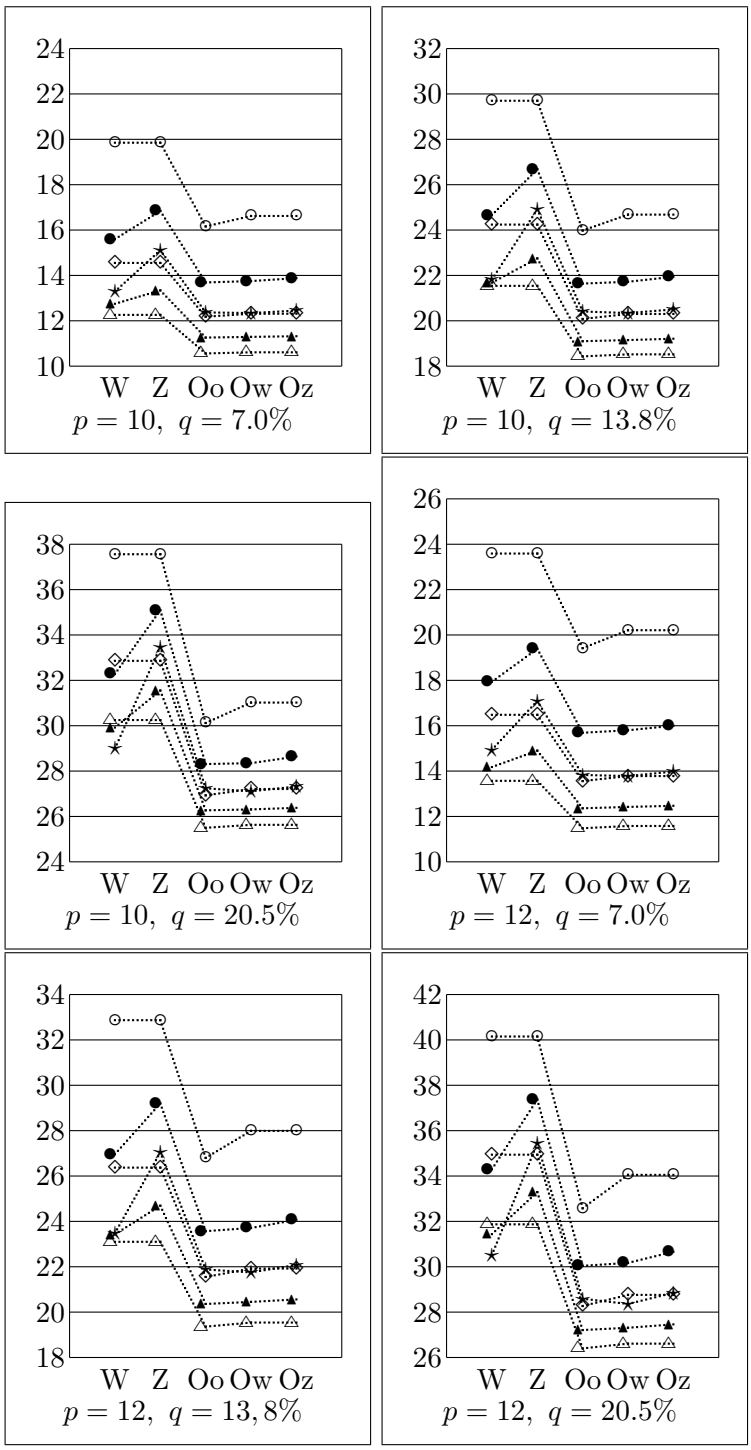


symbol	○	●	⋆	◇	▲	△
(n_1, n_2)	(10,10)	(10,15)	(10,20)	(15,15)	(15,20)	(20,20)

Figure 5.1 : Comparison of $(P_1 + 2P_2)/3$ among 5 rules.

cients (a, b) for the five rules are

W–rule : $(a, b) = (0, \beta_0(D^2))$,
Z–rule : $(a, b) = (a_z, \beta_0(D^2))$,
Oo–rule : $(a, b) = (\frac{1}{n}a_o, \beta_0(D^2) + \frac{1}{n}\beta_1(D^2; a_o) + \frac{1}{n^2}\beta_2(D^2; \alpha_o, 0))$,
Ow–rule: $(a, b) = (0, \beta_0(D^2) + \frac{1}{n}\beta_1(D^2; 0) + \frac{1}{n^2}\beta_2(D^2; 0, 0))$,
Oz–rule: $(a, b) = (a_z, \beta_0(D^2) + \frac{1}{n}\beta_1(D^2; \frac{r_2 - r_1}{2}) + \frac{1}{n^2}\beta_2(D^2; \frac{r_2 - r_1}{2}, \frac{r_1^2 - r_2^2}{4}))$,



symbol	○	●	⋆	◇	▲	△
(n_1, n_2)	(10,10)	(10,15)	(10,20)	(15,15)	(15,20)	(20,20)

Figure 5.2 : Comparison of $\max\{(\frac{1}{2}P_1, P_2\}$ among 5 rules.

where

$$a_z = \frac{N_1 - N_2}{N_1 + N_2 + 2N_1N_2}, \quad a_o = \frac{r_2 - r_1}{2} - \beta_0(D^2).$$

Figure 5.2 gives the values ($\times 100\%$) of $\max\{(P_1/2, P_2)\}$ for the five rules when $p = 10, 12$. In the figure, q is the value of $\max\{(P_1/2, P_2)\}$ for the Bayes rule, that is,

$$q = \frac{1}{2}\Big\{1 - \Phi\Big(\frac{\beta_0}{\Delta} + \frac{\Delta}{2}\Big)\Big\} = \Phi\Big(\frac{\beta_0}{\Delta} - \frac{\Delta}{2}\Big),$$

where β_0 is the solution of this equation.

We can see that the modification with using β_1 and β_2 improve the performance of the W–rule and the Z–rule. The Oo–rule performs best when the sample sizes are small ($n_1 = n_2 = 10$). In other cases of sample sizes, the three rules Oo, Ow, and Oz performs almost same.

When $p = 6, 8$, we could see similar results, but the differences between the rules got smaller a little. When the sample sizes are small relative to the dimension or the Mahalanobis distance is small, the differences between the classification rules become clear. In that cases we recommend to use the Oo–rule.

A The conditional distribution function

In this section, we give a lemma which can be used to derive the differential coefficients of the conditional distribution function $F_i(b; a, \boldsymbol{t}, \Delta)$ defined by (3.2) in section 3.

Lemma 5. *Let $\boldsymbol{t} =< \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \Gamma >$ be a constant $2p + p(p+1)/2$ vector. Suppose that Γ is positive definite. Let $\boldsymbol{x}$ be a random vector distributed as $N_p(\boldsymbol{\mu}, I_p)$. When $a \to 0$ the distribution function of $d_a(\boldsymbol{x}; \boldsymbol{t})$ can be expanded as*

$$\begin{aligned} Pr\{d_a(\boldsymbol{x}; \boldsymbol{t}) \le b\} &= \Phi\Big(y(b; \boldsymbol{\mu}, \boldsymbol{t})\Big) - \phi\Big(y(b; \boldsymbol{\mu}, \boldsymbol{t})\Big) \\ &\quad \cdot \Big\{a g_1\Big(y(b; \boldsymbol{\mu}, \boldsymbol{t}); \boldsymbol{\mu}, \boldsymbol{t}\Big) + \frac{1}{2}a^2 g_2\Big(y(b; \boldsymbol{\mu}, \boldsymbol{t}); \boldsymbol{\mu}, \boldsymbol{t}\Big)\Big\} + O(a^3), \end{aligned} \tag{A.1}$$

where Φ, ϕ are the distribution function and the density function of $N(0, 1)$,

respectively,

$$
\begin{aligned}
&y(b;\boldsymbol{\mu},\boldsymbol{t}) = \frac{b-(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)}{\sigma(\boldsymbol{t})},\\
&\bar{\boldsymbol{\eta}} = \frac{1}{2}(\boldsymbol{\eta}_1+\boldsymbol{\eta}_2),\ \sigma(\boldsymbol{t})^2 = (\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)'\Gamma^{-2}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1),\\
&g_1(y;\boldsymbol{\mu},\boldsymbol{t}) = \frac{1}{\sigma(\boldsymbol{t})}\Big\{d_0(\boldsymbol{t};\boldsymbol{\mu}) + \frac{d_1(\boldsymbol{t};\boldsymbol{\mu})}{\sigma(\boldsymbol{t})}h_1(y) + \frac{d_2(\boldsymbol{t};\boldsymbol{\mu})}{\sigma(\boldsymbol{t})^2}h_2(y)\Big\},\\
&g_2(y;\boldsymbol{\mu},\boldsymbol{t}) = \frac{1}{\sigma(\boldsymbol{t})^2}\Big\{e_0(\boldsymbol{t};\boldsymbol{\mu})h_1(y) + \frac{e_1(\boldsymbol{t};\boldsymbol{\mu})}{\sigma(\boldsymbol{t})}h_2(y),\\
&\qquad\qquad + \frac{e_2(\boldsymbol{t};\boldsymbol{\mu})}{\sigma(\boldsymbol{t})^2}h_3(y) + \frac{e_3(\boldsymbol{t};\boldsymbol{\mu})}{\sigma(\boldsymbol{t})^3}h_4(y) + \frac{e_4(\boldsymbol{t};\boldsymbol{\mu})}{\sigma(\boldsymbol{t})^4}h_5(y)\Big\}.
\end{aligned}
$$

Here, $h_k(y)$ $(k=1,2,\cdots)$ is the Hermite polynomial of degree k defined by

$$\Big(\frac{d}{dy}\Big)^k\phi(y) = (-1)^k h_k(y)\phi(y)$$

and

$$
\begin{aligned}
&d_0(\boldsymbol{t};\boldsymbol{\mu}) = \mathrm{tr}[\Gamma^{-1}] + (\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-1}(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}}) + \tfrac{1}{4}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)'\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1),\\
&d_1(\boldsymbol{t};\boldsymbol{\mu}) = 2(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-2}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1), \quad d_2(\boldsymbol{t};\boldsymbol{\mu}) = (\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)'\Gamma^{-3}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1),\\
&e_0(\boldsymbol{t};\boldsymbol{\mu}) = \Big(d_0(\boldsymbol{t};\boldsymbol{\mu})\Big)^2 + 2\mathrm{tr}[\Gamma^{-2}] + 4(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-2}(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}}),\\
&e_1(\boldsymbol{t};\boldsymbol{\mu}) = 2d_0(\boldsymbol{t};\boldsymbol{\mu})d_1(\boldsymbol{t};\boldsymbol{\mu}) + 8(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-3}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1),\\
&e_2(\boldsymbol{t};\boldsymbol{\mu}) = \Big(d_1(\boldsymbol{t};\boldsymbol{\mu})\Big)^2 + 2d_0(\boldsymbol{t};\boldsymbol{\mu})d_2(\boldsymbol{t};\boldsymbol{\mu}) + 4(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)'\Gamma^{-4}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1),\\
&e_3(\boldsymbol{t};\boldsymbol{\mu}) = 2d_1(\boldsymbol{t};\boldsymbol{\mu})d_2(\boldsymbol{t};\boldsymbol{\mu}),\\
&e_4(\boldsymbol{t};\boldsymbol{\mu}) = \Big(d_2(\boldsymbol{t};\boldsymbol{\mu})\Big)^2.
\end{aligned}
$$

Proof. The characteristic function of $d_a(\boldsymbol{x},\boldsymbol{t})$ can be represented as

$$
\begin{aligned}
\psi(s) &\equiv RE[\exp\{is\ d_a(\boldsymbol{x},\boldsymbol{t})\}]\\
&= RE\Big[\exp\Big\{is(\boldsymbol{x}-\bar{\boldsymbol{\eta}})'\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1) + isa\Big\{q(\boldsymbol{x},\bar{\boldsymbol{\eta}},\Gamma) + \frac{1}{4}q(\boldsymbol{t})\Big\}\Big\}\Big]\\
&= \frac{1}{(2\pi)^{p/2}}\exp\Big\{-\frac{s^2}{2}q[\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\Gamma^2] + is(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\Big\}\\
&\quad\cdot\int\exp\Big\{-\frac{1}{2}\Big(\boldsymbol{z}-\boldsymbol{\mu}-is\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\Big)'\Big(\boldsymbol{z}-\boldsymbol{\mu}-is\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\Big)\Big\}\\
&\qquad\cdot\exp\Big[isa\Big\{q(\boldsymbol{z},\bar{\boldsymbol{\eta}},\Gamma) + \frac{1}{4}q(\boldsymbol{t})\Big\}\Big]d\boldsymbol{z}\\
&= \exp\Big\{-\frac{s^2}{2}q[\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\Gamma^2] + is(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\Big\}\exp\Big(\frac{isa}{4}q(\boldsymbol{t})\Big)\\
&\qquad\cdot RE\Big[\exp\Big\{isa\,q\Big(\boldsymbol{Z}+\boldsymbol{\mu}+is\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1),\bar{\boldsymbol{\eta}},\Gamma\Big)\Big\}\Big],
\end{aligned}
\tag{A.2}
$$

where $\boldsymbol{Z}$ is a random vector distributed as $N_p(\mathbf{0}, I_p)$. Taking the expectation term by term after expanding the exponential in (A.2) in terms of a, we obtain the expansion of the characteristic function as

$$\begin{aligned}\psi(s) = \exp\Big(-\frac{s^2}{2}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)'\Gamma^{-2}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1) + is\,(\boldsymbol{\mu}-\bar{\boldsymbol{\eta}})'\Gamma^{-1}(\boldsymbol{\eta}_2-\boldsymbol{\eta}_1)\Big)\\ \cdot\Big\{1 + isa\Big(d_0(\boldsymbol{t};\boldsymbol{\mu}) + isd_1(\boldsymbol{t};\boldsymbol{\mu}) + (is)^2 d_2(\boldsymbol{t};\boldsymbol{\mu})\Big)\\ + \frac{1}{2}(is)^2a^2\Big(e_0(\boldsymbol{t};\boldsymbol{\mu}) + ise_1(\boldsymbol{t};\boldsymbol{\mu}) + (is)^2e_2(\boldsymbol{t};\boldsymbol{\mu})\\ + (is)^3e_3(\boldsymbol{t};\boldsymbol{\mu}) + (is)^4e_4(\boldsymbol{t};\boldsymbol{\mu})\Big)\Big\} + O(a^3).\end{aligned} \tag{A.3}$$

Inverting (A.3), we obtain (A.1). □

Using lemma 5, we can calculate the derivatives used in section 3 and 4. For example, $Q_0^{(a,\boldsymbol{t})}(\Delta)$ in (3.4) can be calculated as

$$\begin{aligned}Q_0^{(a,\boldsymbol{t})}(\Delta) &= \frac{\partial^2}{\partial a\partial \boldsymbol{t}}\Big\{-F_1(b;a,\boldsymbol{t},\Delta) + cF_2(b;a,\boldsymbol{t},\Delta)\Big\}|_0\\ &= \frac{\partial}{\partial \boldsymbol{t}}\Big\{\phi\Big(y(b;\boldsymbol{\mu}_1,\boldsymbol{t})\Big)g_1\Big(y(b;\boldsymbol{\mu}_1,\boldsymbol{t});\boldsymbol{\mu}_1,\boldsymbol{t}\Big)\\ &\qquad - c_0\phi\Big(y(b;\boldsymbol{\mu}_2,\boldsymbol{t})\Big)g_1\Big(y(b;\boldsymbol{\mu}_2,\boldsymbol{t});\boldsymbol{\mu}_2,\boldsymbol{t}\Big)\Big\}\Big|_0.\end{aligned}$$

B Cut–off Point

In this section we show the actual forms of β_1 and β_2 given by (4.7) and (4.8), respectively. Since the method of calculation is similar as the one of Anderson [2], we state only the results.

$$b_1(\Delta^2) = \sum_{j=0}^{2} b_{1j}K(\Delta^2)^{j-2}, \quad b_2(\Delta^2) = \sum_{j=0}^{5} b_{2j}K(\Delta^2)^{j-5} \tag{B.1}$$

where

$K(\Delta^2) = 1 + k\exp\{\beta_0(\Delta^2)\}$,

$b_{10} = \frac{1}{2}(\Delta^2(r_1-r_2) + 2(\Delta^2 + r_1 + r_2)b_0(\Delta^2))$,

$$b_{11} = \frac{1}{16}\Big(\Delta^4 + 16(-1+p)(r_1+r_2) + 4\Delta^2(-3+4p-r_1+3r_2),$$
$$+4b_0(\Delta^2)\{-2(2\Delta^2+r_1+3r_2)+3b_0(\Delta^2)\}\Big),$$

$$b_{12} = -\frac{1}{32\Delta^2}\Big(\Delta^2\{\Delta^2(-12+\Delta^2+16p)+8(-4+\Delta^2+4p)r_2\}$$
$$+2b_0(\Delta^2)\{-\Delta^2(3\Delta^2+4(-3+4p+2r_2))+6\Delta^2 b_0(\Delta^2)-4b_0(\Delta^2)^2\}$$
$$-8a_1(\Delta^2)\{\Delta^2(-4+\Delta^2+4p)+4b_0(\Delta^2)^2\}\Big),$$

$$b_{20} = \frac{5}{8}\{\Delta^2(r_1-r_2)+2(\Delta^2+r_1+r_2)b_0(\Delta^2)\}^2,$$

$$b_{21} = \frac{1}{16\Delta^2}\Big\{\Delta^4(r_1-r_2)\{\Delta^4+8(1+2p)(r_1+r_2)+\Delta^2(4+16p-21r_1+29r_2)\}$$
$$+2\Delta^2\Big(\Delta^6+16p(r_1+r_2)^2+\Delta^4(4+16p-45r_1+55r_2)$$
$$+\Delta^2\{-47r_1^2+2r_1(2+16p+5r_2)+r_2(4+32p+53r_2)\}\Big)b_0(\Delta^2)$$
$$-4\Delta^2\{25\Delta^4+(r_1+r_2)(31r_1+19r_2)+\Delta^2(53r_1+47r_2)\}b_0(\Delta^2)^2$$
$$-8(\Delta^2+r_1+r_2)\{2\Delta^2+5(r_1+r_2)\}b_0(\Delta^2)^3\Big\},$$

$$b_{22} = \frac{1}{512\Delta^2}\Big[\Delta^2\Big\{\Delta^8+256(-1+p^2)(r_1+r_2)^2+8\Delta^6(1+4p-7r_1+9r_2)$$
$$+16\Delta^4\{-15+8p+16p^2-14r_1-54pr_1+29r_1^2$$
$$+2(9+37p-43r_1)r_2+61r_2^2\}$$
$$-128\Delta^2(r_1+r_2)\{4(1+r_1-r_2)-p(1+4p-7r_1+9r_2)\}\Big\}$$
$$+8b_0(\Delta^2)\Big\{-2\Delta^2\Big(8\Delta^6+16p(r_1+r_2)(9r_1+7r_2)$$
$$+\Delta^4(32+128p-135r_1+215r_2)$$
$$+4\Delta^2\{-37r_1^2+r_1(1+68p+20r_2)+r_2(15+60p+49r_2)\}\Big)$$
$$+b_0(\Delta^2)\Big(347\Delta^6$$
$$-4\Delta^4(-27+20p-193r_1-145r_2)-128(-1+p)(r_1+r_2)^2$$
$$+8\Delta^2\{61r_1^2+r_2(28-26p+13r_2)+r_1(28-26p+86r_2)\}$$
$$+8\{16\Delta^4+16(r_1+r_2)(2r_1+3r_2)+\Delta^2(45r_1+67r_2)\}b_0(\Delta^2)$$
$$-6\{13\Delta^2+16(r_1+r_2)\}b_0(\Delta^2)^2\Big)\Big\}\Big],$$

$$
\begin{aligned}
b_{23} = -\frac{1}{1024\Delta^2}\Big[\Delta^2\Big\{ & 3\Delta^8 + 8\Delta^6(3+12p-7r_1+13r_2) \\
& +256(-1+p)(r_1+r_2)\{r_1+3pr_1+(5+3p)r_2\} \\
& +128\Delta^2\Big(-r_1(9+p-12p^2+4r_1+7pr_1) \\
& \quad +\{-15+p(7+12p+6r_1)\}r_2+(4+13p)r_2^2\Big) \\
& +16\Delta^4\Big(-45+48p^2+r_1(-16+15r_1)+28r_2 \\
& \quad -58r_1r_2+55r_2^2+p(24-50r_1+110r_2)\Big)\Big\} \\
-2\Big\{ & 79\Delta^8 - 256(-3+p)(-1+p)(r_1+r_2)^2 \\
& +8\Delta^6(45+156p-76r_1+174r_2) \\
& -16\Delta^4\Big(37-72p+16p^2+12r_1-174pr_1+43r_1^2 \\
& \quad -2(36+63p+23r_1)r_2-73r_2^2\Big) \\
& +128\Delta^2\Big(r_1\{-12-r_1+p(17-4p+12r_1)\} \\
& \quad +\{2(-6+r_1)+p(17-4p+18r_1)\}r_2+(-1+6p)r_2^2\Big) \\
& +512\Delta^2(r_1-r_2)a_1(\Delta^2)\Big\}b_0(\Delta^2) \\
+8\Big\{ & 241\Delta^6 - 64(-1+p)(r_1+r_2)(5r_1+7r_2) \\
& +\Delta^4(324-240p+536r_1+320r_2) \\
& +8\Delta^2\Big(r_1(76-70p+41r_1)+2(46-43p+29r_1)r_2-19r_2^2\Big) \\
& -256(\Delta^2+r_1+r_2)a_1(\Delta^2)\Big\}b_0(\Delta^2)^2 \\
+16\Big\{ & 75\Delta^4 - 64(r_1+r_2)+12\Delta^2(-5+4p+12r_1+34r_2) \\
& +8\Big(13r_1^2+46r_1r_2+37r_2^2+6p(r_1+r_2)\Big)\Big\}b_0(\Delta^2)^3 \\
& -48(39\Delta^2+40r_1+56r_2)b_0(\Delta^2)^4+288b_0(\Delta^2)^5\Big],
\end{aligned}
$$

$$
\begin{aligned}
b_{24} = & \frac{1}{32\Delta^4}a_1(\Delta^2)^2\Big[\Delta^4\{\Delta^4+8\Delta^2(-1+p)+16(-1+p)(3+p)\} \\
& \quad +8\Delta^2(-4+\Delta^2+4p)b_0(\Delta^2)^2+16b_0(\Delta^2)^4\Big] \\
& +\frac{1}{8\Delta^2}a_1(\Delta^2)\Big[8\Delta^2(-1+p)(r_1-r_2) \\
& \quad +b_0(\Delta^2)\Big(\Delta^4+16(-1+p)(r_1+r_2)+4\Delta^2(-7+8p-r_1+3r_2)
\end{aligned}
$$

$$
\begin{aligned}
&+4b_0(\Delta^2)\{-2(2\Delta^2+r_1+3r_2)+3b_0(\Delta^2)\}\Big)\Big]\\
&+\frac{1}{3072\Delta^2}\Big[\Delta^2\Big\{3\Delta^8+2\Delta^6(17+48p-12r_1+36r_2)\\
&\quad+16\Delta^4\Big(-73+60p+48p^2-9r_1-18pr_1+3r_1^2+3(9+26p-6r_1)r_2+27r_2^2\Big)\\
&\quad+768(-1+p)\Big(r_1\{-4+p(4+r_1)\}+2\{-2+r_1+p(2+r_1)\}r_2+(4+p)r_2^2\Big)\\
&\quad+96\Delta^2\Big(25-56p+32p^2-18r_1+4pr_1+16p^2r_1-2r_1^2-4pr_1^2\\
&\qquad+2\{-21+2p(9+4p+2r_1)\}r_2+2(-1+6p)r_2^2\Big)\Big\}\\
&\quad+2b_0(\Delta^2)\Big\{3\Big(-15\Delta^8-8\Delta^6(13+28p-6r_1+24r_2)\\
&\qquad+256(-1+p)(r_1+r_2)\{(-2+p)r_1+(-4+p)r_2\}\\
&\qquad+16\Delta^4\{37-72p+16p^2+3r_1-30pr_1+3r_1^2-(31+14p+6r_1)r_2-9r_2^2\}\\
&\qquad+64\Delta^2[-r_1\{-13-r_1+2p(11-4p+2r_1)\}\\
&\qquad\quad+(35-46p+8p^2-4(1+p)r_1)r_2+3r_2^2]\Big)\\
&\quad+4b_0(\Delta^2)\Big(81\Delta^6-192(-1+p)(r_1+r_2)(-2+r_1+3r_2)\\
&\qquad+\Delta^4(374-240p+144r_1+72r_2)\\
&\qquad-24\Delta^2\{28-36p-23r_1+18pr_1-3r_1^2+(-39+34p-6r_1)r_2+9r_2^2\}\\
&\qquad+2b_0(\Delta^2)\Big[3\Big\{11\Delta^4+4\Delta^2(-15+12p+2r_1+24r_2)\\
&\qquad\quad+8[r_1(-7+6p+r_1)-9r_2+6(p+r_1)r_2+9r_2^2]\Big\}\\
&\qquad+b_0(\Delta^2)\{50-117\Delta^2-72r_1-216r_2+54b_0(\Delta^2)\}\Big]\Big)\Big\}\Big],
\end{aligned}
$$

$$
\begin{aligned}
b_{25}=\frac{1}{3072\Delta^6}\Big[&\Delta^6\Big\{768a_2(-4+\Delta^2+4p)\\
&-\Delta^2\{1200-224\Delta^2+5\Delta^4+96(-28+3\Delta^2)p+1536p^2\}\\
&-48\{\Delta^4+64(-1+p)^2+4\Delta^2(-7+8p)\}r_2+192(4+\Delta^2-4p)r_2^2\\
&+48a_1(\Delta^2)\Big(16(-1+p)\{\Delta^2+4(-1+p+r_2)\}\\
&\quad-\{\Delta^4+8\Delta^2(-1+p)+16(-1+p)(3+p)\}a_1(\Delta^2)\Big)\Big\}\\
&+2\Delta^4\Big\{25\Delta^6+768(-1+p)r_2^2+48\Delta^4(-14+18p+3r_2)\\
&+48\Delta^2\{25+8p(-7+4p)-44r_2+48pr_2-4r_2^2\}\\
&+48a_1(\Delta^2)\Big(-2\Delta^2(-28+\Delta^2+32p)-16(-4+\Delta^2+4p)r_2\\
&\quad+\{\Delta^4+8\Delta^2(-3+p)+16(-1+p)^2\}a_1(\Delta^2)\Big)\Big\}b_0(\Delta^2)
\end{aligned}
$$

$$
\begin{aligned}
&+8\Delta^4\Big\{384a_2-\Delta^2(-336+25\Delta^2+432p)-24(3\Delta^2+16(-1+p))r_2\\
&\quad+48a_1(\Delta^2)\Big(3\Delta^2+4(-5+6p+2r_2)-(-4+\Delta^2+4p)a_1(\Delta^2)\Big)\Big\}b_0(\Delta^2)^2\\
&+16\Delta^2\Big\{\Delta^2(-112+25\Delta^2+144p+24r_2)\\
&\quad+48a_1(\Delta^2)\Big(-3\Delta^2+(\Delta^2+4(-3+p))a_1(\Delta^2)\Big)\Big\}b_0(\Delta^2)^3\\
&-16\Delta^2\Big\{25\Delta^2+48(-2+a_1(\Delta^2))a_1(\Delta^2)\Big\}b_0(\Delta^2)^4\\
&+32\{5\Delta^2+48a_1(\Delta^2)^2\}b_0(\Delta^2)^5\Big].
\end{aligned}
$$

Acknowlegement

The authors would like to thank the associate editors and reviewers for their carefull reading, giving valuable comments and advice to improve the papers.

References

[1] Anderson, T. W. (1951). Classification by multivariate analysis, *Psychometrika*, **16**, 31-50.

[2] Anderson, T. W. (1973). An asymptotic expansion of the distribution of the Studentized classification statistic W, *Ann. Statist*, **1**, 964-972.

[3] Anderson, T. W. (1984). *An introduction to multivariate analysis* (2nd ed.), John Wiley & Sons, New York.

[4] Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations, *Ann. Math. Statist*, **36**, 1174-1184.

[5] John, S. (1960). On some classification problems. I, *Sankhyã*, **22**, 301-308.

[6] Kudo, A. (1959). The classificatory problem viewed as a two-decision problem, *Mem. Fac. Sci. Kyushu Univ. Ser. A* , **13**, 93-125.

[7] Kudo, A. (1960). The classificatory problem viewed as a two-decision problem II, *Mem. Fac. Sci. Kyushu Univ. Ser. A* , **14**, 63-83.

[8] Matsumoto, M. and Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator, *ACM Trans. on Modeling and Computer Simulation*, **8**, 3-30.

[9] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups, *Ann. Math. Statist.*, **15**, 145–162.