SUT Journal of Mathematics
Vol. 48, No. 1 (2012), 55–69

Constrained linear discriminant rule via the Studentized classification statistic based on monotone missing data

Nobumichi Shutoh, Masashi Hyodo, Tatjana Pavlenko and Takashi Seo

(Received July 13, 2011; Revised May 1, 2012)

Abstract. This paper provides an asymptotic expansion for the distribution of the Studentized linear discriminant function with k-step monotone missing training data. It turns out to be a certain generalization of the results derived by Anderson [1] and Shutoh and Seo [12]. Furthermore we also derive the cut-off point constrained by a conditional probability of misclassification using the idea of McLachlan [8]. Finally we perform Monte Carlo simulation to evaluate our results.

AMS 2010 *Mathematics Subject Classification.* 62H30, 62H10.

Key words and phrases. Asymptotic expansion, linear discriminant analysis, probabilities of misclassification, monotone missing data.

§1. Introduction

Discriminant analysis is well known as one of the statistical procedures for assigning a p–dimensional sample vector $\boldsymbol{x}$ to one of two groups, $\Pi^{(1)}$ and $\Pi^{(2)}$. In this paper, we primarily discuss the linear discrimination when these two groups are derived by distributions $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ and $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$, respectively. Since $\boldsymbol{\mu}^{(g)}$ and Σ are usually unknown, we construct their estimators using the training data $T = \{\boldsymbol{x}_j^{(g)}\}_{j=1}^{N_1^{(g)}}$ from the gth group $\Pi^{(g)}$, $g = 1, 2$. Then consider the linear discriminant function (or Wald-Anderson's plug-in criterion)

$$W = (\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})' S^{-1} \left[\boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}^{(1)} + \overline{\boldsymbol{x}}^{(2)}) \right], \tag{1.1}$$

where

$$\overline{\boldsymbol{x}}^{(g)} = \frac{1}{N_1^{(g)}}\sum_{j=1}^{N_1^{(g)}}\boldsymbol{x}_j^{(g)},\ S = \frac{1}{n_1}\sum_{g=1}^{2}\sum_{j=1}^{N_1^{(g)}}(\boldsymbol{x}_j^{(g)}-\overline{\boldsymbol{x}}^{(g)})(\boldsymbol{x}_j^{(g)}-\overline{\boldsymbol{x}}^{(g)})',$$

and $n_1 = N_1^{(1)} + N_1^{(2)} - 2$. Using the discriminant function (1.1) a new observation $\boldsymbol{x}$ is to be assigned to $\Pi^{(1)}$ if $W > c$, where c is a cut-off point. Classification accuracy is usually measured by the misclassification probabilities that are defined as

$$\begin{aligned} e(2|1) &= \mathrm{E}_T[\Pr(W \le c|T, \boldsymbol{x} \in \Pi^{(1)})], \\ e(1|2) &= \mathrm{E}_T[\Pr(W > c|T, \boldsymbol{x} \in \Pi^{(2)})]. \end{aligned}$$

Several authors have been interested in evaluating these probabilities. For example, Anderson [1] derived an asymptotic expansion of the Studentized version of W and investigated relation between the corresponding misclassification probabilities and the cut-off point c. He also proposed the cut-off point c such that

$$\mathrm{E}_T[\Pr(W \le c|T, \boldsymbol{x} \in \Pi^{(1)})] = \alpha + \mathrm{O}(n_1^{-2}),$$

where α is a value given by experimenters. Further, McLachlan [8] considered the discrimination where one type of error is generally regarded as more serious than the other such as medical applications associated with the diagnosis of diseases, and then, McLachlan [8] proposed the cut-off point c such that

$$\mathrm{Pr}_T[\Pr(W \le c|T, \boldsymbol{x} \in \Pi^{(1)}) < M] = 1 - \beta + \mathrm{O}(n_1^{-2}),$$

where $1 - \beta$ is the desired level of confidence and M is an upper bound.

The above-mentioned results have been developed for the case where all data is observed. However, the datasets often suffer from missing observation by some reasons. In particular, k-step monotone missing data is often observed owing to dropout. Further, variables can be reordered to arrange non-monotone missing data to a dataset that is similar to k-step monotone missing data consisting of the following sample vectors from $\Pi^{(g)}$ $(g = 1, 2)$:

$$\begin{gathered}
\begin{pmatrix} \boldsymbol{x}_{11}^{(g)} \\ \boldsymbol{x}_{21}^{(g)} \\ \vdots \\ \boldsymbol{x}_{k-1,1}^{(g)} \\ \boldsymbol{x}_{k1}^{(g)} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{x}_{1N_1^{(g)}}^{(g)} \\ \boldsymbol{x}_{2N_1^{(g)}}^{(g)} \\ \vdots \\ \boldsymbol{x}_{k-1,N_1^{(g)}}^{(g)} \\ \boldsymbol{x}_{kN_1^{(g)}}^{(g)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_{1,N_1^{(g)}+1}^{(g)} \\ \boldsymbol{x}_{2,N_1^{(g)}+1}^{(g)} \\ \vdots \\ \boldsymbol{x}_{k-1,N_1^{(g)}+1}^{(g)} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{x}_{1,N_{[2]}^{(g)}}^{(g)} \\ \boldsymbol{x}_{2,N_{[2]}^{(g)}}^{(g)} \\ \vdots \\ \boldsymbol{x}_{k-1,N_{[2]}^{(g)}}^{(g)} \end{pmatrix} \\
, \ldots, \\
\begin{pmatrix} \boldsymbol{x}_{1,N_{[k-2]}^{(g)}+1}^{(g)} \\ \boldsymbol{x}_{2,N_{[k-2]}^{(g)}+1}^{(g)} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{x}_{1,N_{[k-1]}^{(g)}}^{(g)} \\ \boldsymbol{x}_{2,N_{[k-1]}^{(g)}}^{(g)} \end{pmatrix}, \left(\boldsymbol{x}_{1,N_{[k-1]}^{(g)}+1}^{(g)} \right), \ldots, \left(\boldsymbol{x}_{1,N_{[k]}^{(g)}}^{(g)} \right),
\end{gathered} \tag{1.2}$$

where $p \equiv p_1 + \cdots + p_k$, $N_{[i]}^{(g)} = N_1^{(g)} + \cdots + N_i^{(g)}$, and $\boldsymbol{x}_{k-i+1,j}^{(g)}$ denotes a p_{k-i+1}–dimensional sample vector from $\Pi^{(g)}$ for $i = 1, \ldots, k$, $j = 1, \ldots, N_{[i]}^{(g)}$ and $g = 1, 2$. Then, we assume a large-sample asymptotic framework for k-step monotone missing data:

$$N_1^{(g)} \to \infty,\ N_{[\ell]}^{(g)} \to \infty,\ \frac{N_1^{(2)}}{N_1^{(1)}} \to q_1,\ \frac{N_{[\ell]}^{(2)}}{N_{[\ell]}^{(1)}} \to q_{[\ell]}\ (\ell = 2, \ldots, k,\ g = 1, 2),$$

where q_1 and $q_{[\ell]}$ are positive constants, respectively. Kanda and Fujikoshi [5] suggested the asymptotic approximation for the misclassification probabilities for the datasets in (1.2) for $k = 2$ (i.e., 2-step monotone missing data), assuming that covariance matrix is known. Recently, a solution addressing the same problem when a common covariance matrix Σ is unknown was given by Batsidis et al. [2] and Shutoh and Seo [12] in the case of 2-step monotone missing data. In this study, we generalize the results derived by Anderson [1] and McLachlan [8] to the case of k-step monotone missing data. Although Shutoh [10, 11] focused on the approximations for the misclassification probabilities with a given cut-off point under k-step monotone missing data, we now suggest a technique for specifying a cut-off point c for any desired constraint on the upper bound of the conditional misclassification probability. Testing procedure in the case of monotone missing patterns in multivariate data has been studied; see e.g., Chang and Richards [3] and Koizumi and Seo [6, 7]. Distributional properties of the estimators based on k-step monotone missing data were extensively studied in Kanda and Fujikoshi [4].

The paper will proceed as follows: In Section 2, we embed a k-step monotone missing data scheme into discrimination framework. In Section 3, we present asymptotic expansion of the Studentized version of W under the missing data assumption and derive the constrained discriminant rule. In Section 4, we evaluate properties of the technique suggested for specifying the cut-off point c by using Monte Carlo simulation, and compare our results with the methods due to Anderson [1] and McLachlan [8]. In Section 5, we conclude with discussion.

§2. Notation

In this paper, we assume k-step monotone missing scheme for training data, meaning the sample vectors coming from $\Pi^{(g)}$ stated in (1.2) can be represented as

$$\boldsymbol{x}^{(g)}_{(k-i+1)j} = \begin{pmatrix} \boldsymbol{x}^{(g)}_{1j} \\ \vdots \\ \boldsymbol{x}^{(g)}_{k-i+1,j} \end{pmatrix} \sim N_{p_{[k-i+1]}}(\boldsymbol{\mu}^{(g)}_{(k-i+1)}, \Sigma_{(k-i+1)}) \tag{2.1}$$
$$(g=1,2,\ i=1,\dots,k,\ j=N^{(g)}_{[i-1]}+1,\dots,N^{(g)}_{[i]}),$$

where $\boldsymbol{x}^{(g)}_{\alpha j}$ for $\alpha=1,\dots,k$ and $j=1,\dots,N^{(g)}_{[k-\alpha+1]}$ is p_α–dimensional partitioned sample vector. Here $\boldsymbol{\mu}^{(g)}_\alpha$ is p_α–dimensional partitioned vector of $\boldsymbol{\mu}^{(g)}$, $\Sigma_{\alpha\beta}$ is $p_\alpha \times p_\beta$ partitioned matrix of Σ for $\alpha=1,\dots,k,\ \beta=1,\dots,k$,

$$\boldsymbol{\mu}^{(g)}_{(k-i+1)} = \begin{pmatrix} \boldsymbol{\mu}^{(g)}_1 \\ \vdots \\ \boldsymbol{\mu}^{(g)}_{k-i+1} \end{pmatrix},$$
$$\Sigma_{(k-i+1)} = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1,k-i+1} \\ \vdots & \ddots & \vdots \\ \Sigma_{k-i+1,1} & \cdots & \Sigma_{k-i+1,k-i+1} \end{pmatrix}, \tag{2.2}$$

$p_{[k-i+1]} = \sum_{\alpha=1}^{k-i+1} p_\alpha$, $N^{(g)}_{[i]} = \sum_{\ell=1}^{i} N^{(g)}_\ell$ and $N^{(g)}_{[0]} \equiv 0$.

Now, using the representations (2.1) and (2.2), the linear discriminant function is defined by

$$W_k = (\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)})'\widehat{\Sigma}^{-1}\left[\boldsymbol{x} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}^{(1)} + \widehat{\boldsymbol{\mu}}^{(2)})\right],$$

where $\widehat{\boldsymbol{\mu}}^{(g)}$, $g=1,2$ and $\widehat{\Sigma}$ are certain estimators of $\boldsymbol{\mu}^{(g)}$, $g=1,2$ and Σ (the details are found in section 3 of Shutoh [10]). Then, the misclassification probabilities can be expressed as

$$\begin{aligned} e_k(2|1) &= \mathrm{E}_{T_k}[\Pr(W_k \leq c|T_k, \boldsymbol{x} \in \Pi^{(1)})], \\ e_k(1|2) &= \mathrm{E}_{T_k}[\Pr(W_k > c|T_k, \boldsymbol{x} \in \Pi^{(2)})], \end{aligned}$$

where T_k is k-step monotone missing training data. Note that Shutoh [10] considered a special case $c = 0$ under a large-sample framework as well as under a high-dimensional framework. Assume henceforth that $\boldsymbol{x} \in \Pi^{(1)}$. The symmetry of our discrimination rule allows us to obtain the results for $\boldsymbol{x} \in \Pi^{(2)}$ by the same arguments.

For the case of $e_k(2|1)$, we consider $\mathrm{E}_{T_k}[\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})]$, since

$$Z_k = V_k^{-\frac{1}{2}}(\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)})'\widehat{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(1)})$$

is distributed as the standard normal distribution given $\widehat{\boldsymbol{\mu}}^{(1)}$, $\widehat{\boldsymbol{\mu}}^{(2)}$, $\widehat{\Sigma}$ and $\boldsymbol{x} \in \Pi^{(1)}$, where $u = [c-(1/2)D_k^2]/D_k$, $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0, 1)$,

$$\begin{aligned} D_k^2 &= (\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)})'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)}), \\ F_k &= (\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)})'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{\mu}^{(1)}), \\ V_k &= (\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)})'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)}). \end{aligned} \tag{2.3}$$

For the details of D_k^2, F_k and V_k, see Shutoh [9, 10].

§3. Main results

3.1. Asymptotic expansion for the distribution of Studentized version of W_k

To derive asymptotic expansion of $e_k(2|1) = \mathrm{E}_{T_k}[\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})]$, we now expand $\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})$ using perturbation method, as in Shutoh [9]. Let A be a $p \times p$ matrix. Then, for a large m,

$$\left(I - \frac{1}{\sqrt{m}}A\right)^{-1} = I + \sum_{i=1}^{\infty} m^{-\frac{i}{2}}A^i.$$

By applying this expansion to (2.3), we can obtain the following stochastic expansions in the form

$$\begin{aligned} D_k^2 &\equiv \Delta^2 + \frac{D_{k1}}{\sqrt{n}} + \frac{D_{k2}}{n} + \frac{D_{k3}}{n\sqrt{n}} + \mathrm{O}_p(n^{-2}), \\ F_k &\equiv \frac{F_{k1}}{\sqrt{n}} + \frac{F_{k2}}{n} + \frac{F_{k3}}{n\sqrt{n}} + \mathrm{O}_p(n^{-2}), \\ V_k &\equiv \Delta^2 + \frac{V_{k1}}{\sqrt{n}} + \frac{V_{k2}}{n} + \frac{V_{k3}}{n\sqrt{n}} + \mathrm{O}_p(n^{-2}), \end{aligned}$$

where $n = n_{[k]}$, $n_{[\ell]} = \sum_{i=1}^{\ell}(N_i^{(1)} + N_i^{(2)} - 2)$ $(\ell = 2, \ldots, k)$, $\Delta^2 = \boldsymbol{\delta}'\Sigma^{-1}\boldsymbol{\delta}$ and $\boldsymbol{\delta} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$. Furthermore, using the above expressions and Taylor expansion of $\Phi(\cdot)$, we can determine w_{k1}, w_{k2} and w_{k3} using the following equation,

$$(3.1)\ \Phi((uD_k + F_k)V_k^{-\frac{1}{2}}) \equiv \Phi(u) + \phi(u)\left[\frac{w_{k1}}{\sqrt{n}} + \frac{w_{k2}}{n} + \frac{w_{k3}}{n\sqrt{n}}\right] + \mathrm{O}_p(n^{-2}),$$

where $\phi(u)$ is the density function of $N(0,1)$ (see Shutoh [9] for the details).

Theorem 1. *The cumulative distribution function of the Studentized version of the linear discriminant function* $[W_k - (1/2)D_k^2]/D_k$ *under* $\boldsymbol{x} \in \Pi^{(1)}$ *is expanded as*

$$\Phi(u) + \frac{\phi(u)}{n}b_{k1}(u) + \mathrm{O}(n^{-2}), \tag{3.2}$$

where $b_{k1}(u) = c_0 + c_1 u + c_3 u^3$,

$$\begin{aligned}
c_0 &= \frac{p-1}{r_1\Delta}(1+q_1) \\
&\quad + \sum_{\ell=2}^{k}\frac{p_{[k-\ell+1]} - \Delta_{k-\ell+1}^2}{\Delta}\left(\frac{1+q_{[\ell]}}{r_{[\ell]}} - \frac{1+q_{[\ell-1]}}{r_{[\ell-1]}}\right), \\
c_1 &= -\frac{1}{r_1}\left(p - \frac{1}{4} + \frac{1}{2}q_1\right) \\
&\quad - \sum_{\ell=2}^{k}\Delta_{k-\ell+1}^2\left\{\frac{1}{r_{[\ell]}}\left(p_{[k-\ell+1]} + \frac{3}{2} + \frac{1}{2}q_{[\ell]} - \frac{7}{4}\Delta_{k-\ell+1}^2\right)\right. \\
&\quad \left. - \frac{1}{r_{[\ell-1]}}\left(p_{[k-\ell+1]} + \frac{3}{2} + \frac{1}{2}q_{[\ell-1]} - \frac{7}{4}\Delta_{k-\ell+1}^2\right)\right\}, \\
c_3 &= -\frac{1}{4r_1} - \sum_{\ell=2}^{k}\frac{\Delta_{k-\ell+1}^4}{4}\left(\frac{1}{r_{[\ell]}} - \frac{1}{r_{[\ell-1]}}\right), \\
\boldsymbol{\delta}_{(k-\ell+1)} &= \boldsymbol{\mu}_{(k-\ell+1)}^{(1)} - \boldsymbol{\mu}_{(k-\ell+1)}^{(2)}, \\
\delta_{k-\ell+1}^2 &= \boldsymbol{\delta}'_{(k-\ell+1)}\Sigma_{(k-\ell+1)}^{-1}\boldsymbol{\delta}_{(k-\ell+1)}, \quad \Delta_{k-\ell+1} = \delta_{k-\ell+1}/\Delta,
\end{aligned}$$

r_1 *denotes the limit of* n_1/n *and* $r_{[\ell]}$ *denotes the limit of* $n_{[\ell]}/n$, *for* $\ell = 2, \ldots, k$.

Proof. Noting that

$$\begin{aligned}
\Pr\left[\frac{W_k - \frac{1}{2}D_k^2}{D_k} \le u \middle| T_k, \boldsymbol{x} \in \Pi^{(1)}\right] &= \Pr[Z_k \le (uD_k + F_k)V_k^{-\frac{1}{2}} | T_k, \boldsymbol{x} \in \Pi^{(1)}] \\
&= \Phi((uD_k + F_k)V_k^{-\frac{1}{2}}),
\end{aligned}$$

(3.2) follows from the expectations for the multivariate normal distribution and Wishart distribution that $\mathrm{E}_{T_k}(w_{k1}) = 0$, $\mathrm{E}_{T_k}(w_{k2}) = b_{k1}(u)$ and the expectations of the terms included in $w_{k3}/(n\sqrt{n})$ are either 0 or $\mathrm{O}(n^{-2})$. □

We notice that the cumulative distribution function of the Studentized version of the linear discriminant function $[W_k + (1/2)D_k^2]/D_k$ under $\boldsymbol{x} \in \Pi^{(2)}$ is expanded as

$$\Phi(u') - \frac{\phi(u')}{n} b'_{k1}(-u') + \mathrm{O}(n^{-2}),$$

where $u' = [c + (1/2)D_k^2]/D_k$ and $b'_{k1}(u)$ is obtained by inverting q_1 and $q_{[\ell]}$ in $b_{k1}(u)$. For $k = 2$, Theorem 1 coincides with the result derived by Shutoh and Seo [12].

Generalizing Anderson [1], we can obtain the cut-off point c using the above results.

Theorem 2. *For a given α, the cut-off point c which satisfies*

$$\mathrm{E}_{T_k}[\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})] = \alpha + \mathrm{O}(n^{-2}) \tag{3.3}$$

is $c = u_0 D_k + (1/2)D_k^2$, where $u_0 = z_{1-\alpha} - b_{k1}(z_{1-\alpha})/n$, $z_{1-\alpha}$ is the upper 100α percentage point of $N(0,1)$.

Proof. If we put $u_0 = a + b/n$ into (3.2) and combine with (3.3), then we can determine $a = z_{1-\alpha}$ and $b = -b_{k1}(z_{1-\alpha})$. The outline of the proof is similar to that of Theorem 9.8.1 in Siotani et al. [13]. □

In practical applications, unknown parameters Δ^2 and $\delta_{k-\ell+1}^2$ should be replaced by their asymptotically unbiased estimators proposed in Shutoh [11]: $\{(n_1 - p - 1)D_{k,k}^2\}/n_1$ and $\{(n_{[\ell]} - p_{[k-\ell+1]} - 1)D_{k,k-\ell+1}^2\}/n_{[\ell]}$, where

$$D_{k,\alpha}^2 = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_1^{(1)} - \widehat{\boldsymbol{\mu}}_1^{(2)} \\ \vdots \\ \widehat{\boldsymbol{\mu}}_\alpha^{(1)} - \widehat{\boldsymbol{\mu}}_\alpha^{(2)} \end{pmatrix}' \begin{pmatrix} \widehat{\Sigma}_{11} & \cdots & \widehat{\Sigma}_{1\alpha} \\ \vdots & \ddots & \vdots \\ \widehat{\Sigma}_{\alpha 1} & \cdots & \widehat{\Sigma}_{\alpha\alpha} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{\mu}}_1^{(1)} - \widehat{\boldsymbol{\mu}}_1^{(2)} \\ \vdots \\ \widehat{\boldsymbol{\mu}}_\alpha^{(1)} - \widehat{\boldsymbol{\mu}}_\alpha^{(2)} \end{pmatrix},$$

$\widehat{\boldsymbol{\mu}}_\alpha^{(g)}$ is p_α–dimensional partitioned vector of the estimator of $\boldsymbol{\mu}^{(g)}$ based on k-step monotone missing data and $\widehat{\Sigma}_{\alpha\beta}$ is $p_\alpha \times p_\beta$ partitioned matrix of the estimator of Σ based on the same.

The similar result under $\boldsymbol{x} \in \Pi^{(2)}$ can be also obtained if we consider

$$\mathrm{E}_{T_k}[\Phi((u^* D_k^* + F_k^*)\{V_k^*\}^{-\frac{1}{2}})] = \alpha + \mathrm{O}(n^{-2}), \tag{3.4}$$

where $u^* = -(c + (1/2)\{D_k^*\}^2)/D_k^*$,

$$\Pr(W_k > c | T_k, \boldsymbol{x} \in \Pi^{(2)}) = \Phi((u^* D_k^* + F_k^*)\{V_k^*\}^{-\frac{1}{2}}),$$

$\{D_k^*\}^2$, F_k^*, and V_k^* are the statistics obtained by interchanging 1 and 2 in the superscript for D_k^2, F_k, and V_k, respectively. By noting that $D_k^2 = \{D_k^*\}^2$, the cut-off point c which satisfies (3.4) can be obtained by inverting q_1 and $q_{[\ell]}$ in u_0 for Theorem 2.

3.2. Constrained discriminant rule with Studentized version of W_k

The another main result is derived in this subsection. Similarly to Theorem 9.6.5 in Siotani et al. [13], we consider the following characteristic function:

$$c(t) \;=\; \mathrm{E}_{T_k}\left[\exp\left\{\mathrm{i}t\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})\right\}\right],$$

where $\mathrm{i} = \sqrt{-1}$. Using (3.1), we can obtain the following lemma.

Lemma 1. $\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})$ *follows asymptotically normal distribution with mean* ξ_k *and variance* σ_k^2*, where* $b_{k2}(u) = d_0 + d_2 u^2$*,*

$$\begin{aligned}
\xi_k &= \Phi(u) + \frac{\phi(u)}{n} b_{k1}(u), \quad \sigma_k^2 \;=\; \frac{\{\phi(u)\}^2}{n} b_{k2}(u), \\
d_0 &= \frac{1+q_1}{r_1} + \sum_{\ell=2}^{k} \Delta_{k-\ell+1}^2 \left\{\frac{1+q_{[\ell]}}{r_{[\ell]}} - \frac{1+q_{[\ell-1]}}{r_{[\ell-1]}}\right\}, \\
d_2 &= \frac{1}{2}\left\{\frac{1}{r_1} + \sum_{\ell=2}^{k} \Delta_{k-\ell+1}^4 \left(\frac{1}{r_{[\ell]}} - \frac{1}{r_{[\ell-1]}}\right)\right\}.
\end{aligned}$$

Proof. We can show that

$$\begin{aligned}
c(t) &= \exp\left[\mathrm{i}t\mathrm{E}_{T_k}\left\{\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})\right\}\right] \\
&\quad \times \left[1 + \frac{(\mathrm{i}t)^2}{2!}\mathrm{Var}_{T_k}\{\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})\} + R\right],
\end{aligned}$$

where R is the remainder term starting with $(\mathrm{i}t)^3$. Furthermore, we have

$$\begin{aligned}
c(t) &= \exp\left[\mathrm{i}t\mathrm{E}_{T_k}\left\{\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})\right\}\right. \\
&\quad \left. - \frac{t^2}{2}\mathrm{Var}_{T_k}\left\{\Phi((uD_k + F_k)V_k^{-\frac{1}{2}})\right\}\right](1 + R')
\end{aligned}$$

by noting that

$$\log\left[1+\frac{(\mathrm{i}t)^2}{2!}\mathrm{Var}_{T_k}\{\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})\}+R''\right]$$

is expanded as

$$\frac{(\mathrm{i}t)^2}{2!}\mathrm{Var}_{T_k}\{\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})\}+R''',$$

where R', R'' and R''' are also the remainder terms starting with $(\mathrm{i}t)^3$, respectively. The required moments of $\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})$ are

$$\begin{aligned}
\mathrm{E}_{T_k}\left[\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})\right] &= \Phi(u)+\phi(u)\left[\frac{1}{\sqrt{n}}\mathrm{E}_{T_k}(w_{k1})\right.\\
&\quad \left.+\frac{1}{n}\mathrm{E}_{T_k}(w_{k2})+\frac{1}{n\sqrt{n}}\mathrm{E}_{T_k}(w_{k3})\right]+\mathrm{O}(n^{-2}),\\
\mathrm{Var}_{T_k}\left[\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})\right] &= \frac{\{\phi(u)\}^2}{n}\mathrm{Var}_{T_k}(w_{k1})+\frac{2\{\phi(u)\}^2}{n\sqrt{n}}\\
&\quad \times\{\mathrm{E}_{T_k}(w_{k1}w_{k2})-\mathrm{E}_{T_k}(w_{k1})\mathrm{E}_{T_k}(w_{k2})\}\\
&\quad +\mathrm{O}(n^{-2}).
\end{aligned}$$

The first moment of the terms up to $\mathrm{O}_p(n^{-\frac{3}{2}})$ has been already derived in Theorem 1. Similarly, we have

$$\mathrm{Var}_{T_k}\left[\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})\right] = \frac{\{\phi(u)\}^2}{n}b_{k2}(u)+\mathrm{O}(n^{-2})$$

by using the moments of the multivariate normal distribution and Wishart distribution. □

In the following theorem, we obtain the cut-off point c such that

$$\mathrm{Pr}_{T_k}\left[\Phi((uD_k+F_k)V_k^{-\frac{1}{2}})<M\right] = 1-\beta+\mathrm{O}(n^{-2}). \tag{3.5}$$

Theorem 3. *The cut-off point $c=(1/2)D_k^2+uD_k$ which satisfies (3.5) can be obtained by*

$$u=m-\frac{h_{k1}}{\sqrt{n}}-\frac{h_{k2}}{n}-\frac{h_{k3}}{n\sqrt{n}},$$

where $m = z_{1-M}$*,*

$$\begin{aligned} h_{k1} &= z_\beta b_{k2}^{\frac{1}{2}}, \quad h_{k2} \;=\; b_{k1} + \frac{1}{2} z_\beta^2 m (b_{k2} - 2d_2), \\ h_{k3} &= z_\beta b_{k2}^{\frac{1}{2}} \biggl[m b_{k1} + \frac{m d_2}{2 b_{k2}} (z_\beta^2 m d_2 - 2 b_{k1}) + \frac{z_\beta^2 b_{k2}}{3} (m^2 - 1) \\ &\qquad\qquad - \frac{3m^2}{2} (z_\beta^2 d_2 + 2c_3) + \frac{z_\beta^2 d_2}{2} - c_1 \biggr], \end{aligned}$$

$b_{k1} = b_{k1}(m)$, $b_{k2} = b_{k2}(m)$.

Proof. By Lemma 1, we express (3.5) as follows:

$$\Phi\left(\frac{M - \xi_k}{\sigma_k} \right) = 1 - \beta + \mathrm{O}(n^{-2}).$$

Thus, our goal is to derive the point u which satisfies

$$M = \xi_k + z_\beta \sigma_k. \tag{3.6}$$

It should be noted that M is specified by experimenters. We put the solution u of (3.6) as

$$u = m - \frac{h_{k1}}{\sqrt{n}} - \frac{h_{k2}}{n} - \frac{h_{k3}}{n\sqrt{n}},$$

where h_i's ($i = 1, 2, 3$) are the unknown finite constants. Since we have

$$\begin{aligned} \Phi(u) &= M + \frac{\phi(m)}{\sqrt{n}} \Bigl\{ -h_{k1} \Bigr\} + \frac{\phi(m)}{n} \Bigl\{ -h_{k2} - \frac{1}{2} m h_{k1}^2 \Bigr\} \\ &\quad + \frac{\phi(m)}{n\sqrt{n}} \Bigl\{ -h_{k3} - m h_{k1} h_{k2} - \frac{1}{6}(m^2 - 1) h_{k1}^3 \Bigr\} + \mathrm{o}(n^{-\frac{3}{2}}), \\ \phi(u) &= \phi(m) + \frac{\phi(m)}{\sqrt{n}} \Bigl\{ m h_{k1} \Bigr\} \\ &\quad + \frac{\phi(m)}{n} \Bigl\{ m h_{k2} + \frac{1}{2}(m^2 - 1) h_{k1}^2 \Bigr\} + \mathrm{o}(n^{-1}), \\ b_{k1}(u) &= b_{k1} + \frac{1}{\sqrt{n}} \Bigl\{ -(c_1 + 3c_3 m^2) h_{k1} \Bigr\} + \mathrm{o}(n^{-\frac{1}{2}}), \\ \{b_{k2}(u)\}^{\frac{1}{2}} &= b_{k2}^{\frac{1}{2}} + \frac{1}{\sqrt{n}} \Bigl\{ -m d_2 b_{k2}^{-\frac{1}{2}} h_{k1} \Bigr\} \\ &\quad + \frac{1}{n} \Bigl\{ -m d_2 b_{k2}^{-\frac{1}{2}} h_{k2} + \frac{1}{2} d_2 (1 - m^2 d_2 b_{k2}^{-1}) b_{k2}^{-\frac{1}{2}} h_{k1}^2 \Bigr\} + \mathrm{o}(n^{-1}), \end{aligned}$$

we can equate the terms of respective order $n^{-\frac{1}{2}}$, n^{-1}, and $n^{-\frac{3}{2}}$ in (3.6) and determine the unknown constants h_{ki}'s, which proves Theorem 3. □

In practical applications, unknown parameters Δ^2 and $\delta^2_{k-\ell+1}$ should be replaced by their asymptotically unbiased estimators proposed in Shutoh [11]: $\{(n_1 - p - 1)D^2_{k,k}\}/n_1$ and $\{(n_{[\ell]} - p_{[k-\ell+1]} - 1)D^2_{k,k-\ell+1}\}/n_{[\ell]}$.

The similar result under $\boldsymbol{x} \in \Pi^{(2)}$ can be also obtained if we consider

$$\Pr_{T_k}\left[\Phi((u^* D^*_k + F^*_k)\{V^*_k\}^{-\frac{1}{2}}) < M\right] = 1 - \beta + \mathrm{O}(n^{-2}). \tag{3.7}$$

The cut-off point $c = -(1/2)D^2_k + u^* D_k$ which satisfies (3.7) can be obtained by

$$u^* = m - \frac{h^*_{k1}}{\sqrt{n}} - \frac{h^*_{k2}}{n} - \frac{h^*_{k3}}{n\sqrt{n}},$$

where h^*_{ki}'s are defined as the constants inverting $q_1 = q_{[1]}$ and $q_{[\ell]}$ in h_{ki}'s for $i = 1, 2, 3$ and $\ell = 2, \ldots, k$.

§4. Simulation studies

In this section, we compare the proposed results for Theorems 2 and 3 based on the monotone missing training data T_k with the similar results for the complete training data T (i.e., the methods derived by Anderson [1] and McLachlan [8]) under $\boldsymbol{x} \in \Pi^{(1)}$.

At first, in order to evaluate our result derived in Theorem 2, we compare our result for $k = 3$ with the result derived by Anderson [1]. We give $\Delta = 1.05$ and $\alpha = 0.10$. Further, the dimensionalities are set as $p = 3$ ($p_1 = p_2 = p_3 = 1$). The sample sizes are set as $N_{(3)} \equiv N_1 = N_2 = N_3 = 10, 15, 20, 40$, where $N_\ell \equiv N^{(1)}_\ell = N^{(2)}_\ell$ ($\ell = 1, 2, 3$). For the result derived by Anderson [1], the sample size is set as $N_1 = 10, 15, 20, 40$.

Table 1. The comparison of the misclassification probabilities with the cut-off points proposed in Anderson [1] and Theorem 2.

N_1	The result of Anderson [1]	$N_{(3)}$	The result of Th.2
10	0.1071	10	**0.1055**
15	0.1032	15	**0.1026**
20	0.1019	20	**0.1015**
40	0.1006	40	**0.1005**

For all the cases we performed, the misclassification probabilities are closer to a specified α than the result derived by Anderson [1]. See the bold face we mark in Table 1.

Furthermore, we also compare our result stated in Theorem 3 for $k = 3$ with the result derived by McLachlan [8]. We select the seven cases, as listed in

Table 2. The dimensionalities are set as $p = 3$ ($p_1 = p_2 = p_3 = 1$). The sample sizes are set as $N_{(3)} = 20, 40, 100, 200$. For the result derived by McLachlan [8], the sample size is set as $N_1 = 20, 40, 100, 200$.

Table 2. The selected parameters in simulation studies for Cases 1–7.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
$1-\beta$	0.95	0.99	0.90	0.95	0.95	0.95	0.95
Δ	1.05	1.05	1.05	0.50	1.36	1.05	1.05
M	0.20	0.20	0.20	0.20	0.20	0.10	0.25

We compare the values of $1 - \beta$ listed in Table 3. In Cases 2, 4, and 6, our result also provides the level $1 - \beta$ which is closer to the specified value than the result for the complete data. These results imply that the proposed procedure is more efficient than the result derived by McLachlan [8] when β, Δ, and M are lower. By bold face, we mark the results where the suggested procedure demonstrate superior performance comparing to the method due to McLachlan [8].

Table 3. The values of the desired $1 - \beta$ level of confidence for Cases 1–7.

	The values $1-\beta$ for the result of McLachlan [8]						
N_1	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
20	0.8959	0.9430	0.8518	0.8701	0.9017	0.8562	0.9075
40	0.9196	0.9671	0.8723	0.9034	0.9225	0.8997	0.9256
100	0.9331	0.9791	0.8834	0.9262	0.9342	0.9247	0.9358
200	0.9376	0.9830	0.8866	0.9340	0.9382	0.9335	0.9391

	The values $1-\beta$ for the result of Th.3						
$N_{(3)}$	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
20	0.8975	**0.9528**	0.8474	**0.8861**	0.9001	**0.8700**	0.9062
40	0.9195	**0.9716**	0.8676	**0.9115**	0.9212	**0.9062**	0.9239
100	0.9320	**0.9808**	0.8792	**0.9273**	0.9331	**0.9272**	0.9338
200	0.9359	**0.9836**	0.8818	0.9328	0.9365	**0.9352**	0.9364

In Table 4, we compare the misclassification probabilities based on the cut-off points provided by the method due to McLachlan [8] and Theorem 3, respectively. Proposed procedure results in the misclassification probability closer to a specified M. Then it can be also observed that the other misclassification probability for the proposed procedure is lower than that for the method due to McLachlan [8].

Table 4. The misclassification probabilities with the cut-off point c for Cases 1–7.

	$e(2\mid 1)$ for the result of McLachlan [8]						
N_1	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
20	0.1256	0.1079	0.1380	0.1307	0.1243	0.0621	0.1599
40	0.1384	0.1206	0.1499	0.1408	0.1379	0.0645	0.1784
100	0.1571	0.1427	0.1656	0.1575	0.1571	0.0734	0.2011
200	0.1686	0.1568	0.1752	0.1688	0.1685	0.0795	0.2143
	$e(1\mid 2)$ for the result of McLachlan [8]						
N_1	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
20	0.5983	0.6327	0.5752	0.7870	0.4780	0.7399	0.5376
40	0.5446	0.5783	0.5247	0.7512	0.4214	0.7067	0.4786
100	0.4945	0.5196	0.4807	0.7075	0.3731	0.6674	0.4279
200	0.4704	0.4894	0.4600	0.6850	0.3499	0.6456	0.4034
	$e_3(2\mid 1)$ for the result of Th.3						
$N_{(3)}$	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
20	0.1482	0.1315	0.1583	0.1438	0.1490	0.0707	0.1892
40	0.1596	0.1457	0.1677	0.1566	0.1601	0.0749	0.2038
100	0.1737	0.1636	0.1794	0.1725	0.1738	0.0828	0.2204
200	0.1812	0.1736	0.1854	0.1807	0.1813	0.0875	0.2289
	$e_3(1\mid 2)$ for the result of Th.3						
$N_{(3)}$	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
20	0.5374	0.5672	0.5204	0.7533	0.4119	0.7003	0.4721
40	0.4982	0.5219	0.4850	0.7174	0.3742	0.6692	0.4309
100	0.4642	0.4799	0.4556	0.6820	0.3438	0.6391	0.3977
200	0.4488	0.4603	0.4426	0.6664	0.3298	0.6245	0.3829

§5. Conclusion and discussion

This paper provided the asymptotic expansion for the Studentized version of W_k in subsection 3.1. Subsection 3.1 also derived the cut-off point c which could control the misclassification probability with a specified value. Moreover subsection 3.2 provided a certain method for determining the cut-off point which could control the conditional misclassification probability with its upper bound and the level of significance. They were certain extensions of the results derived by Anderson [1], McLachlan [8] and Shutoh and Seo [12]. Our results on specifying the significance level are more accurate in a way that they allow for reducing the upper bound of the conditional misclassification probability for a given cut-off point. In Section 4 we demonstrated the advantages of our approach by Monte Carlo simulation and showed that it can be useful for specifying the cut-off point c in practical classification problems for some cases.

For the proposed result in Theorem 3, unfortunately, the estimator of h_{k1} has bias with order n^{-1} and its bias correction is one of the future problem. Possible further studies on the classification problem with missing data could be focused on other approaches such as e.g. maximum likelihood approach or quadratic discriminant analysis.

Acknowledgements

The authors would like to extend their sincere gratitude to the Editor and the referee who gave invaluable comments and suggestions, which have greatly enhanced this paper. Any remaining errors are the authors' responsibility. This research was in part supported by grants from the Forum for Asian Studies (4440101-2010), Stockholm University, Sweden. N. Shutoh's and M. Hyodo's research was in part supported by Grant-in-Aid for JSPS Fellows (23·6926, 23·9731). T. Pavlenko's research was in part supported by grants from Swedish Research Council (421-2008-1966). T. Seo's research was in part supported by Grant-in-Aid for Scientific Research (C) (23500360).

References

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

[2] A. Batsidis, K. Zografos and S. Loukas, *Errors in discrimination with monotone missing data from multivariate normal populations*, Comput. Statist. Data Anal. **50** (2006), 2600–2634.

[3] W. Y. Chang and D. St. P. Richards, *Finite-sample inference with monotone incomplete multivariate normal data, II.*, J. Multivariate Anal. **101** (2010), 603–620.

[4] T. Kanda and Y. Fujikoshi, *Some basic properties of the MLE's for a multivariate normal distribution with monotone missing data*, Amer. J. Math. Management. Sci. **18** (1998), 161–190.

[5] T. Kanda and Y. Fujikoshi, *Linear discriminant function and probabilities of misclassification with monotone missing data*, Proc. 8th China-Japan Statist. Sympos., 142–143, 2004.

[6] K. Koizumi and T. Seo, *Simultaneous confidence intervals among k mean vectors in repeated measures with missing data*, Amer. J. Math. Management. Sci. **29** (2009), 263–275.

[7] K. Koizumi and T. Seo, *Testing equality of two mean vectors and simultaneous confidence intervals in repeated measures with missing data*, J. Japanese Soc. Comput. Statist. **22** (2009), 33–41.

[8] G. J. McLachlan, *Constrained sample discrimination with the Studentized classification statistic W*, Comm. Statist. – Theory Methods **6** (1977), 575–583.

[9] N. Shutoh, *Constrained sample discrimination with the Studentized linear discriminant function based on monotone missing training data*, Technical Report No.10–09, Hiroshima Statistical Research Group, Hiroshima University, 2010.

[10] N. Shutoh, *An asymptotic approximation for EPMC in linear discriminant analysis based on monotone missing data*, J. Statist. Plann. Inference **142** (2012), 110–125.

[11] N. Shutoh, *An asymptotic expansion for the distribution of the linear discriminant function based on monotone missing data*, J. Statist. Comput. Simulat. **82** (2012), 241–259.

[12] N. Shutoh and T. Seo, *Asymptotic expansion of the distribution of the Studentized linear discriminant function based on two-step monotone missing samples*, Comm. Statist. – Simulation Comput. **39** (2010), 1365–1383.

[13] M. Siotani, T. Hayakawa and Y. Fujikoshi, *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Inc., Ohio, U.S.A., 1985.

Nobumichi Shutoh
Department of Health Sciences, Oita University of Nursing and Health Sciences
2944-9, Megusuno, Oita-shi, Oita 870-1201, Japan
E-mail: shutoh@oita-nhs.ac.jp

Masashi Hyodo
Graduate School of Economics, The University of Tokyo
Research Fellow of the Japan Society for the Promotion of Science
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
E-mail: hyodoh_h@yahoo.co.jp

Tatjana Pavlenko
Department of Mathematics, KTH Royal Institute of Technology
SE-100 44 Stockholm, Sweden
E-mail: pavlenko@math.kth.se

Takashi Seo
Department of Mathematical Information Science, Tokyo University of Science
1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
E-mail: seo@rs.kagu.tus.ac.jp