Journal of Multivariate Analysis 140 (2015) 234–244

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Asymptotic properties of the misclassification rates for Euclidean Distance Discriminant rule in high-dimensional data

Hiroki Watanabe[a], Masashi Hyodo[b,*], Takashi Seo[a], Tatjana Pavlenko[c]

[a] *Department of Mathematical Information Science, Tokyo University of Science, Japan*

[b] *Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, Japan*

[c] *Department of Mathematics, KTH Royal Institute of Technology, KTH Royal Institute of Technology, Sweden*

ARTICLE INFO

Article history:
Received 23 February 2014
Available online 27 May 2015

AMS subject classifications:
62H30
62E20
62H10
62H12

Keywords:
High-dimensional framework
Conditional error rate
Expected error rate

ABSTRACT

Performance accuracy of the Euclidean Distance Discriminant rule (EDDR) is studied in the high-dimensional asymptotic framework which allows the dimensionality to exceed sample size. Under mild assumptions on the traces of the covariance matrix, our new results provide the asymptotic distribution of the conditional misclassification rate and the explicit expression for the consistent and asymptotically unbiased estimator of the expected misclassification rate. To get these properties, new results on the asymptotic normality of the quadratic forms and traces of the higher power of Wishart matrix, are established. Using our asymptotic results, we further develop two generic methods of determining a cut-off point for EDDR to adjust the misclassification rates. Finally, we numerically justify the high accuracy of our asymptotic findings along with the cut-off determination methods in finite sample applications, inclusive of the large sample and high-dimensional scenarios.

1. Introduction

In this paper, we focus on the discrimination problem which is concerned with the allocation of a given object, $\boldsymbol{x}$, a random vector represented by a set of features $(x_1, \ldots, x_p)$, to one or two populations, Π_1 and Π_2 given by $\mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathcal{N}_p(\boldsymbol{\mu}_2, \Sigma)$, respectively, where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and common covariance matrix Σ is non-singular. Let $\{\boldsymbol{x}_{gj}\}_{j=1}^{N_g}$ be a random sample of independent observations drawn from gth population $\mathcal{N}_p(\boldsymbol{\mu}_g, \Sigma)$, $g = 1, 2$. Let also $N = N_1 + N_2$ denote the total sample size and set $n = N - 2$. We are interested to explore the discrimination procedure that can accommodate $p > n$ cases, with the main focus on the performance accuracy in the asymptotic framework that allows p to grow together with n.

Clearly, the classical discriminant procedures, like Fisher linear discriminant rule, cannot be used when $p > n$ since the sample covariance matrix is singular and hence cannot be inverted. An intuitively appealing alternative considered in this study focuses on geometrical properties of the sample space and re-formulates the classification problem in terms of the *Euclidean distance discriminant rule* (EDDR): assign a new observation $\boldsymbol{x}$ to the "nearest" population Π_g, i.e. assign to Π_g if it is on average closer to the data from Π_g than to the data from the other population. Matusita's papers (see [3,4]) are perhaps the oldest references dealing with the discriminant rule based on distance measures, including the case when the multivariate distributions underlying the data are not specified.

[*] Corresponding author.
E-mail address: hyodoh_h@yahoo.co.jp (M. Hyodo).

http://dx.doi.org/10.1016/j.jmva.2015.05.008
0047-259X/

Recently, Aoshima and Yata [2] have been considered the EDDR for the high-dimensional multi-class problem with different class covariance matrices. In particular, they derived asymptotic conditions which ensure that the expected misclassification rate converges to zero. Recent paper by Srivastava [8] used the Moore–Penrose inverse of the estimated covariance matrix and suggested a second-order approximation of the expected error rate in high-dimensional data.

We, in this study, focus on the asymptotic behavior of the misclassification rates of EDDR. Continuing with the normality assumption, with $\boldsymbol{\mu}_g$ acting as the center of the Π_g's distribution we define

$$T_0(\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{\mu}_2\|^2 - \|\boldsymbol{x}-\boldsymbol{\mu}_1\|^2, \tag{1.1}$$

and its sample based version as

$$\widetilde{T}(\boldsymbol{x}) = \|\boldsymbol{x}-\bar{\boldsymbol{x}}_2\|^2 - \|\boldsymbol{x}-\bar{\boldsymbol{x}}_1\|^2 \tag{1.2}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\bar{\boldsymbol{x}}_g$'s denote the sample mean vectors, $g = 1, 2$. Hence, each term in (1.1) and (1.2) represents the distance between the observed vector $\boldsymbol{x}$ and the centroid of Π_g's or its sample based counterpart.

The natural advantage of using $\widetilde{T}(\boldsymbol{x})$ for classifying high-dimensional data is its ability to mitigate the effect of dimensionality on the performance accuracy. Indeed, as it is seen from (1.2), $\widetilde{T}(\boldsymbol{x})$ utilizes only the marginal distribution of the p variables, thereby naturally reducing the effect of large p in implementations. But the dimensionality has impact on the classification accuracy. To show this, we first point out that classifier $\widetilde{T}(\boldsymbol{x})$ has a bias. In fact,

$$\mathrm{E}[\widetilde{T}(\boldsymbol{x})|\boldsymbol{x}\in\Pi_g] = (-1)^{g-1}\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2 + \frac{N_1-N_2}{N_1N_2}\mathrm{tr}\,\Sigma, \quad g=1,2,$$

and thus the impact of dimensionality is implied by the quantity $(N_1-N_2)\mathrm{tr}\,\Sigma/(N_1N_2)$. In this study, we introduce the bias-corrected version $\widetilde{T}(\boldsymbol{x})$ defined as

$$T(\boldsymbol{x}) = \|\boldsymbol{x}-\bar{\boldsymbol{x}}_2\|^2 - \|\boldsymbol{x}-\bar{\boldsymbol{x}}_1\|^2 - \frac{N_1-N_2}{N_1N_2}\mathrm{tr}\,S, \tag{1.3}$$

where the subtraction of $(N_1-N_2)/(N_1N_2)\mathrm{tr}\,S$ in (1.3) is to guarantee that $\mathrm{E}[T(\boldsymbol{x})|\boldsymbol{x}\in\Pi_g] = (-1)^{g-1}\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2$, $g=1,2$. Here, $S=(1/n)\sum_{g=1}^{2}\sum_{j=1}^{N_g}(\boldsymbol{x}_{gj}-\bar{\boldsymbol{x}}_g)(\boldsymbol{x}_{gj}-\bar{\boldsymbol{x}}_g)'$.

Now, the EDDR given by $T(\boldsymbol{x})$ places a new observation $\boldsymbol{x}$ to Π_1 if $T(\boldsymbol{x}) > \tilde{c}$, and to Π_2 otherwise, where $\tilde{c}$ is an appropriate cut-off point. Then, for a specific $\tilde{c}$, the performance accuracy of EDDR will be represented by the pair of misclassification rates that result. Precisely, we define the conditional misclassification rate of EDDR by

$$ce(2|1) = \Pr(T(\boldsymbol{x}) \le \tilde{c}|\boldsymbol{x}\in\Pi_1,\ \bar{\boldsymbol{x}}_1,\ \bar{\boldsymbol{x}}_2,\ S)$$

and its expected version by $e(2|1) = \mathrm{E}[ce(2|1)]$, where the expectation is taken with respect to $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2$ and S. Our main objective is to derive characteristic properties of both conditional and expected misclassification rate in high-dimensional data.

In many practical problems one type of misclassification rate is generally regarded as more serious than the other, examples include e.g. medical applications associated with the diagnosis of diseases. In such a case, it might be desired to determine the cut-off $\tilde{c}$ to obtain a specified probability of the error, or at least to approximate a specified probability. Then, one might base the choice of $\tilde{c}$ on the expected misclassification rate. This method, denoted in what follows by **M1**, suggests to set a cut-off point $\tilde{c}$ such that

$$\textbf{M1}:\ e(2|1) = \mathrm{E}[ce(2|1)] = \alpha,$$

where α is a value given by experimenters.

On the other hand, one may exploit the confidence of the conditional error rate when determining $\tilde{c}$; we denote this method by

$$\textbf{M2}:\ \Pr(ce(2|1) < eu) = 1-\beta,$$

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

Both determination methods **M1** and **M2** have been established by using large sample approximation, see [1,5,6]. In this study, we extend the consideration to the high-dimensional case. Our main theoretical results provide the asymptotically unbiased and consistent estimator of $e(2|1)$ and the limit distribution of $ce(2|1)$ under general assumptions covering the case when $p > n$. In fact, **M1** and **M2** procedures can be considered as specific examples of using our generic results in the theory of EDDR in high-dimensions.

The remaining part of the paper is organized as follows. In Section 2, we derived the asymptotically unbiased and consistent estimator of $e(2|1)$. Further, the limiting approximations of the cut-off point defined by **M1** are established by using this estimator. In Section 3, two estimators of the confidence-based cut-off point defined by **M2** are proposed, for which the asymptotic normality of the conditional error rate is shown. Section 4 summaries the results of numerical experiments justifying the validity of the suggested cut-off estimators for various strength of dependence underlying the data along with a number of high-dimensional scenarios where p far exceeds the sample size. We conclude in Section 5.

2. Evaluation of the expected misclassification rate

Getting the closed-form expression for the expected error is too demanding, therefore we first shall derive its asymptotic approximation, and then based on this result, propose the consistent and asymptotically unbiased estimator of $e(2|1)$ in high dimensions. We further show how these results can be used to provide the cut-off by the determination procedure **M1**.

Let $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2,\ a_i = \operatorname{tr} \Sigma^i/p\ (i = 1, \dots, 8),\ \Delta_i = \boldsymbol{\delta}'\Sigma^i\boldsymbol{\delta}\ (i = 1, \dots, 7)$ and $\Delta_0 = \boldsymbol{\delta}'\boldsymbol{\delta}$. We make the following assumptions for the consistency and unbiasedness of the estimator of $e(2|1)$.

(A1): $N_1, N_2,\ p \to \infty$ with $0 < \lim_{(N_1,N_2,p)\to\infty} \frac{p}{n} = r_0 < \infty$,
$N_1, N_2 \to \infty$ with $0 < \lim_{(N_1,N_2)\to\infty} \frac{N_i}{n+2} = r_i < 1\ (i = 1, 2)$,
(A2): $0 < \lim_{p\to\infty} \Delta_i = \Delta_i^* < \infty\ (i = 0, 1),\ 0 < \lim_{p\to\infty} a_i = a_i^* < \infty\ (i = 1, 2)$,
(A3): $\lim_{(N_1,N_2,p)\to\infty} \frac{\Delta_3}{n} \to 0,\ \lim_{(N_1,N_2,p)\to\infty} \frac{a_4}{n} \to 0$.

Assume henceforth $\boldsymbol{x} \in \Pi_1$. The symmetry of our classification rule makes the probability of error if the mean of $\boldsymbol{x}$ is $\boldsymbol{\mu}_1$ the same as that under $\boldsymbol{\mu}_2$. Then for the conditional distribution of $T(\boldsymbol{x})$ given $(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, S)$ it holds that

$$T(\boldsymbol{x})|(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, S) \sim \mathcal{N}\left(-2U - \frac{N_1 - N_2}{N_1N_2}\operatorname{tr} S, 4V\right),$$

where

$$U = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\bar{\boldsymbol{x}}_1 - \boldsymbol{\mu}_1) - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),$$
$$V = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\Sigma(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

Now the expected error rate $e(2|1)$ of $T(\boldsymbol{x})$ can be expressed in terms of U and V as

$$e(2|1) = \mathrm{E}[ce(2|1)] = \mathrm{E}\left[\Phi\left(\frac{U + (N_2^{-1} - N_1^{-1})p\hat{a}_1/2 + c}{\sqrt{V}}\right)\right], \tag{2.1}$$

where the expectation is with respect to U, V and $\hat{a}_1,\ c = \tilde{c}/2,\ \hat{a}_1 = \operatorname{tr} S/p$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

In order to proceed to asymptotic approximation of $e(2|1)$, we need some preparatory stochastic evaluation of U and V. We introduce the auxiliary random variables

$$\boldsymbol{z}_1 = N^{-\frac{1}{2}}\Gamma'\Sigma^{-\frac{1}{2}}(N_1\bar{\boldsymbol{x}}_1 + N_2\bar{\boldsymbol{x}}_2 - N_1\mu_1 - N_2\mu_2),$$
$$\boldsymbol{z}_2 = \left(\frac{N}{N_1N_2}\right)^{-\frac{1}{2}}\Gamma'\Sigma^{-\frac{1}{2}}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 - \mu_1 + \mu_2),$$

and observe that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are independent and identically distributed as $\mathcal{N}_p(\mathbf{0}, I_p)$, where Γ is an orthogonal matrix such that $\Sigma = \Gamma\Lambda\Gamma'$ and Λ is a diagonal matrix of eigenvalues of Σ. By means of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, we further define

$$U_0 = -\frac{1}{2}\Delta_0, \tag{2.2}$$

$$U_1 = \frac{1}{\sqrt{N}}\delta'\Gamma\Lambda^{\frac{1}{2}}\boldsymbol{z}_1 - \left(\frac{N_1}{NN_2}\right)^{\frac{1}{2}}\delta'\Gamma\Lambda^{\frac{1}{2}}\boldsymbol{z}_2 + \frac{1}{(N_1N_2)^{\frac{1}{2}}}\boldsymbol{z}_1'\Lambda\boldsymbol{z}_2 - \frac{N_1 - N_2}{2N_1N_2}(\boldsymbol{z}_2'\Lambda\boldsymbol{z}_2 - pa_1), \tag{2.3}$$

$$U_2 = \frac{(N_1 - N_2)p}{2N_1N_2}(\hat{a}_1 - a_1), \tag{2.4}$$

and observe that by using (2.2)–(2.4) the numerator in (2.1) can be decomposed as

$$U + \frac{(N_1 - N_2)p\hat{a}_1}{2N_1N_2} = U_0 + U_1 + U_2. \tag{2.5}$$

By analogy with U, V can also be decomposed by first defining V_0 and V_1 as

$$V_0 = \Delta_1 + \frac{Npa_2}{N_1N_2},$$

$$V_1 = 2\left(\frac{N}{N_1N_2}\right)^{\frac{1}{2}}\delta'\Gamma\Lambda^{\frac{3}{2}}\boldsymbol{z}_2 + \frac{N}{N_1N_2}(\boldsymbol{z}_2'\Lambda^2\boldsymbol{z}_2 - pa_2) \tag{2.6}$$

and then observing that $V = V_0 + V_1$. Let

$$U_0^* = -\frac{1}{2}\Delta_0^*, \qquad V_0^* = \Delta_1^* + \frac{r_0 a_2^*}{r_1 r_2}.$$

To evaluate the second moments, we apply Lemma A.3 (see supplemental material, Appendix A) and obtain

$$\mathrm{E}\left[\left(U + \frac{(N_1 - N_2)p\hat{a}_1}{2N_1N_2} - U_0\right)^2\right] = H_U(\Delta_1, a_2) + o(n^{-1}),$$
$$\mathrm{E}[(V - V_0)^2] = H_V(\Delta_3, a_4) + o(n^{-1}),$$

where

$$H_U(\Delta_1, a_2) = \frac{1}{N_2}\Delta_1 + \frac{(N_1^2 + N_2^2)pa_2}{2N_1^2N_2^2}, \qquad H_V(\Delta_3, a_4) = \frac{4N}{N_1N_2}\Delta_3 + \frac{2N^2pa_4}{(N_1N_2)^2}.$$

Under the assumptions (A1)–(A3), it holds that

$$\mathrm{E}\left[\left(U + \frac{(N_1 - N_2)p\hat{a}_1}{2N_1N_2} - U_0\right)^2\right] \to 0, \qquad \mathrm{E}[(V - V_0)^2] \to 0. \tag{2.7}$$

Chebyshev's inequality, (2.7), $U_0 = U_0^* + o(1)$ and $V_0 = V_0^* + o(1)$ imply that

$$U + \frac{(N_1 - N_2)p\hat{a}_1}{2N_1N_2} \xrightarrow{P} U_0^*, \qquad V \xrightarrow{P} V_0^*, \tag{2.8}$$

where $\xrightarrow{P}$ denotes convergence in probability.

Since $\Phi(\cdot)$ in (2.1) is a continuous function of U and V, it follows from (2.8), by the continuous mapping theorem, that

$$\left|\Phi\left(\frac{U + (N_1 - N_2)p\hat{a}_1/(2N_1N_2) + c}{\sqrt{V}}\right) - \Phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}}\right)\right| \xrightarrow{P} 0.$$

On the other hand, it naturally holds that

$$\left|\Phi\left(\frac{U + (N_1 - N_2)p\hat{a}_1/(2N_1N_2) + c}{\sqrt{V}}\right) - \Phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}}\right)\right| < 1.$$

Hence, by the dominated convergence theorem we have

$$\mathrm{E}\left[\left|\Phi\left(\frac{U + (N_1 - N_2)p\hat{a}_1/(2N_1N_2) + c}{\sqrt{V}}\right) - \Phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}}\right)\right|\right] \to 0. \tag{2.9}$$

Further, by applying Jensen's inequality to (2.9) we get

$$\begin{aligned}&\left|\mathrm{E}\left[\Phi\left(\frac{U + (N_1 - N_2)p\hat{a}_1/(2N_1N_2) + c}{\sqrt{V}}\right)\right] - \Phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}}\right)\right| \\ &\quad \le \mathrm{E}\left[\left|\Phi\left(\frac{U + (N_1 - N_2)p\hat{a}_1/(2N_1N_2) + c}{\sqrt{V}}\right) - \Phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}}\right)\right|\right] \to 0.\end{aligned}$$

The above results are summarized in the following lemma.

Lemma 2.1. *Under assumptions* (A1)–(A3)

$$e(2|1) \to \Phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}}\right). \tag{2.10}$$

In words, Lemma 2.1 provides a closed form expression for the limiting term of $e(2|1)$. Hence, to identify the cut-off point for $T(\boldsymbol{x})$, we derive a consistent and unbiased estimator of $e(2|1)$ by plugging-in consistent estimators of U_0^* and V_0^* into the right hand side of (2.10).

As U_0^* and V_0^* are functions of Δ_0^*, Δ_1^* and a_2^*, we begin by obtaining their consistent estimators.

Lemma 2.2. *Let estimators of Δ_0^*, Δ_1^*, a_2^* be defined as*

$$\widehat{\Delta}_0 = \widehat{\boldsymbol{\delta}}'\widehat{\boldsymbol{\delta}} - \frac{Np}{N_1N_2}\hat{a}_1, \tag{2.11}$$

$$\widehat{\Delta}_1 = \widehat{\boldsymbol{\delta}}'S\widehat{\boldsymbol{\delta}} - \frac{Np}{N_1N_2}\hat{a}_2, \tag{2.12}$$

$$\hat{a}_2 = \frac{n^2}{p(n+2)(n-1)}\left(\operatorname{tr} S^2 - \frac{(\operatorname{tr} S)^2}{n}\right), \tag{2.13}$$

respectively, where $\widehat{\boldsymbol{\delta}} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$. Then under assumptions (A1)–(A3)

$$\widehat{\Delta}_0 \xrightarrow{P} \Delta_0^*, \qquad \widehat{\Delta}_1 \xrightarrow{P} \Delta_1^*, \qquad \hat{a}_2 \xrightarrow{P} a_2^*.$$

Proof. To show consistency of a_1 and a_2, we use exact expressions for the variances of these estimators derived in [7] as

$$\mathrm{E}[(\hat{a}_1 - a_1)^2] = \frac{2a_2}{np}, \tag{2.14}$$

$$\mathrm{E}[(\hat{a}_2 - a_2)^2] = \frac{8(n+2)(n+3)(n-1)^2}{pn^5}a_4 + \frac{4(n+2)(n-1)}{n^4}(a_2^2 - p^{-1}a_4). \tag{2.15}$$

Then by applying Chebyshev's inequality to (2.14), (2.15), $a_1 - a_1^* = o(1)$ and $a_2 - a_2^* = o(1)$, it can be seen that

$$\hat{a}_1 \xrightarrow{P} a_1^*, \qquad \hat{a}_2 \xrightarrow{P} a_2^*. \tag{2.16}$$

To show consistency of $\widehat{\Delta}_0$ and $\widehat{\Delta}_1$, we first consider the following random variables

$$\widetilde{\Delta}_0 = \widehat{\boldsymbol{\delta}}'\widehat{\boldsymbol{\delta}} - \frac{Np}{N_1N_2}a_1, \qquad \widetilde{\Delta}_1 = \widehat{\boldsymbol{\delta}}'S\widehat{\boldsymbol{\delta}} - \frac{Np}{N_1N_2}a_2$$

and evaluate the first two moments of $\widehat{\boldsymbol{\delta}}'\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\delta}}'S\widehat{\boldsymbol{\delta}}$. We rewrite

$$\widehat{\boldsymbol{\delta}}'\widehat{\boldsymbol{\delta}} = \boldsymbol{\delta}'\boldsymbol{\delta} + 2\left(\frac{N}{N_1N_2}\right)^{1/2}\boldsymbol{\delta}'\Sigma^{1/2}\boldsymbol{z} + \frac{N}{N_1N_2}\boldsymbol{z}'\Sigma\boldsymbol{z},$$

and

$$\widehat{\boldsymbol{\delta}}'S\widehat{\boldsymbol{\delta}} = \boldsymbol{\delta}'S\boldsymbol{\delta} + 2\left(\frac{N}{N_1N_2}\right)^{1/2}\boldsymbol{\delta}'S\Sigma^{1/2}\boldsymbol{z} + \frac{N}{N_1N_2}\boldsymbol{z}'\Sigma^{1/2}S\Sigma^{1/2}\boldsymbol{z}, \tag{2.17}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I_p)$. Then it easily follows that

$$\mathrm{E}[\widetilde{\Delta}_0] = \Delta_0^* + o(1), \qquad \mathrm{E}[\widetilde{\Delta}_1] = \Delta_1^* + o(1) \tag{2.18}$$

and

$$\mathrm{Var}[\widetilde{\Delta}_0] = \frac{4N}{N_1N_2}\Delta_1 + \frac{2N^2p}{N_1^2N_2^2}a_2, \tag{2.19}$$

$$\mathrm{Var}[\widetilde{\Delta}_1] = \frac{2a_2^2N^2p^2}{nN_1^2N_2^2} + \frac{4a_2\Delta_1Np}{nN_1N_2} + \frac{2a_4N^3p}{nN_1^2N_2^2} + \frac{2\Delta_1^2}{n} + \frac{4\Delta_3N^2}{nN_1N_2}. \tag{2.20}$$

By applying Chebyshev's inequality to (2.18)–(2.20), we obtain

$$\widetilde{\Delta}_0 \xrightarrow{P} \Delta_0^*, \qquad \widetilde{\Delta}_1 \xrightarrow{P} \Delta_1^*. \tag{2.21}$$

Finally, from (2.16), (2.21), $a_1 - a_1^* = o(1)$ and $a_2 - a_2^* = o(1)$, we see that consistency of $\widetilde{\Delta}_0$ and $\widetilde{\Delta}_1$ imply consistency of $\widehat{\Delta}_0$ and $\widehat{\Delta}_1$. □

Now by substituting the estimators of Δ_0^*, Δ_1^*, a_2^* into the limiting term in Lemma 2.1. The consistent estimator of $e(2|1)$ is given by $\Phi((\widehat{U}_0 + c)\widehat{V}_0^{-1/2})$, where $\widehat{U}_0 = -\widehat{\Delta}_0/2$ and $\widehat{V}_0 = \widehat{\Delta}_1 + Np\hat{a}_2/(N_1N_2)$.

The following theorem is provided by the consistency of estimators $\widehat{\Delta}_0$, $\widehat{\Delta}_1$ and $\hat{a}_2$, continuous mapping theorem and dominated convergence theorem.

Theorem 2.1. *Under assumptions* (A1)–(A3)

$$\Phi\left((\widehat{U}_0+c)\widehat{V}_0^{-1/2}\right)\xrightarrow{P} e(2|1)\quad \textit{and}\quad \mathrm{E}\left[\Phi\left((\widehat{U}_0+c)\widehat{V}_0^{-1/2}\right)\right]\to e(2|1).$$

By the results of Theorem 2.1 and Lemma 2.1, the **M1**-based cut-off point for EDDR using $T(\boldsymbol{x})$ is provided by

$$\hat{c}_1=\widehat{V}_0^{1/2}z_\alpha-\widehat{U}_0,$$

where z_α is the α-percentile of $\mathcal{N}(0,1)$ and $\alpha\in(0,1)$.

3. Asymptotic distribution of the conditional misclassification rate

Our objective in this section is to establish the asymptotic distribution of $ce(2|1)$, for which we need some auxiliary notations and assumptions. We begin by modifying the high-dimensional asymptotic framework from Section 2 by replacing the Assumption (A3) with (B3) as follows:

$$\begin{aligned}
&\text{(B3)}:\ 0<\lim_{p\to\infty}\Delta_i=\Delta_i^*<\infty\quad (i=2,3),\qquad 0<\lim_{p\to\infty}a_i=a_i^*<\infty\quad (i=3,4),\\
&\qquad \frac{p}{n}=r_0+o\left(\frac{1}{\sqrt{n}}\right),\qquad \frac{N_i}{n+2}=r_i+o\left(\frac{1}{\sqrt{n}}\right)\quad (i=1,2),\\
&\qquad \Delta_i=\Delta_i^*+o\left(\frac{1}{\sqrt{p}}\right)\quad (i=0,1),\qquad a_2=a_2^*+o\left(\frac{1}{\sqrt{p}}\right),\\
&\qquad \lim_{(N_1,N_2,p)\to\infty}\frac{\Delta_i}{\sqrt{n}}\to 0\quad (i=4,5),\qquad \lim_{(N_1,N_2,p)\to\infty}\frac{a_i}{\sqrt{n}}\to 0\quad (i=5,6).
\end{aligned}$$

As $ce(2|1)$ is a function of the variable set of (U,V), we first obtain the joint asymptotic distribution of (U,V).

Lemma 3.1. *Let*

$$\widetilde{U}=U+\frac{(N_1-N_2)p\hat{a}_1}{N_1N_2}.$$

Then under assumptions (A1), (A2) *and* (B3) *the following holds*

$$\sqrt{n}\left\{\begin{pmatrix}\widetilde{U}\\ V\end{pmatrix}-\begin{pmatrix}U_0^*\\ V_0^*\end{pmatrix}\right\}\xrightarrow{\mathcal{D}}\mathcal{N}_2(\mathbf{0},\Theta),$$

where

$$\Theta=\lim_{(N_1,N_2,p)\to\infty} n\begin{pmatrix}H_U(\Delta_1,a_2) & H_{UV}(\Delta_2,a_3)\\ H_{UV}(\Delta_2,a_3) & H_V(\Delta_3,a_4)\end{pmatrix},$$

$$H_{UV}(\Delta_2,a_3)=-\frac{2}{N_2}\Delta_2-\frac{N(N_1-N_2)pa_3}{(N_1N_2)^2},$$

and $\xrightarrow{\mathcal{D}}$ *denotes convergence in distribution.*

Proof. Let d_1 and d_2 denote two non-random values which satisfy $0<\lim_{(N_1,N_2,p)\to\infty}|d_1|<\infty$ and $0<\lim_{(N_1,N_2,p)\to\infty}|d_2|<\infty$, and introduce the random variable

$$Q=\sqrt{n}\left\{d_1\left(\widetilde{U}-U_0^*\right)+d_2\left(V-V_0^*\right)\right\}.$$

The asymptotic normality of Q would imply that the joint distribution of $\widetilde{U}$ and V is asymptotically normal. Thus, Lemma 3.1 will be proven if we show the normal convergence of Q under (A1), (A2) and (B3). We introduced the following notations

$$\begin{aligned}
\omega_1&=\frac{d_1\sqrt{n}}{\sqrt{N}}\Lambda^{\frac{1}{2}}\Gamma'\boldsymbol{\delta},\\
\omega_2&=\frac{2d_2\sqrt{nN}}{\sqrt{N_1N_2}}\Lambda^{3/2}\Gamma'\boldsymbol{\delta}-\frac{d_1\sqrt{nN_1}}{\sqrt{NN_2}}\Lambda^{1/2}\Gamma'\boldsymbol{\delta},\\
\Omega_3&=\frac{d_1\sqrt{n}}{\sqrt{N_1N_2}}\Lambda,\\
\Omega_4&=\frac{d_2\sqrt{n}N}{N_1N_2}\Lambda^2-\frac{d_1\sqrt{n}(N_1-N_2)}{2N_1N_2}\Lambda.
\end{aligned}$$

Now, since $\hat{a}_1 - a_1 = O_p(n^{-1})$ by (2.14),

$$Q = \omega_1' \boldsymbol{z}_1 + \omega_2' \boldsymbol{z}_2 + \boldsymbol{z}_1' \Omega_3 \boldsymbol{z}_2 + \boldsymbol{z}_2' \Omega_4 \boldsymbol{z}_2 + o_p(1)$$

under the assumptions (A1), (A2) and (B3). Note also that

$$\begin{aligned}
\omega_1' \omega_1 &= \frac{d_1^2 n}{N} \boldsymbol{\delta}' \Sigma \boldsymbol{\delta}, \\
\omega_2' \omega_2 &= \frac{4 d_2^2 n N}{N_1 N_2} \boldsymbol{\delta}' \Sigma^3 \boldsymbol{\delta} + \frac{d_1^2 n N_1}{N N_2} \boldsymbol{\delta}' \Sigma \boldsymbol{\delta} - \frac{4 d_1 d_2 n}{N_2} \boldsymbol{\delta}' \Sigma^2 \boldsymbol{\delta}, \\
\operatorname{tr} \Omega_3^2 &= \frac{d_1^2 n}{N_1 N_2} \operatorname{tr} \Sigma^2, \\
\operatorname{tr} \Omega_4^2 &= \frac{d_2^2 n N^2}{N_1^2 N_2^2} \operatorname{tr} \Sigma^4 + \frac{d_1^2 n (N_1 - N_2)^2}{4 N_1^2 N_2^2} \operatorname{tr} \Sigma^2 - \frac{d_1 d_2 n (N_1^2 - N_2^2)}{N_1^2 N_2^2} \operatorname{tr} \Sigma^3.
\end{aligned}$$

By combining these terms, we now obtain the asymptotic variance of Q as

$$\sigma_Q^2 = \lim_{(N_1, N_2, p) \to \infty} n\{d_1^2 H_U(\Delta_1, a_2) + 2 d_1 d_2 H_{UV}(\Delta_2, a_3) + d_2^2 H_V(\Delta_3, a_4)\}$$

and observe that (A1), (A2) and (B3)

$$0 < \sigma_Q^2 < \infty. \tag{3.1}$$

Furthermore, the following convergence results hold

$$\omega_1' \Omega_3 \omega_2 \to 0, \qquad \omega_2' \Omega_4 \omega_2 \to 0, \qquad \operatorname{tr} \Omega_3^2 \Omega_4 \to 0 \quad \text{and} \quad \operatorname{tr} \Omega_4^3 \to 0. \tag{3.2}$$

Now by using (3.1) and (3.2), and by applying (A.1) from Lemma A.1 (see supplemental material, Appendix A), we obtain

$$\frac{\omega_1' \Omega_3 \omega_2}{\sigma_Q^3} \to 0, \qquad \frac{\omega_2' \Omega_4 \omega_2}{\sigma_Q^3} \to 0, \qquad \frac{\operatorname{tr} \Omega_3^2 \Omega_4}{\sigma_Q^3} \to 0 \quad \text{and} \quad \frac{\operatorname{tr} \Omega_4^3}{\sigma_Q^3} \to 0. \tag{3.3}$$

(3.3) in combination with Lemma A.1 shows that the asymptotic normality of Q holds, which completes the proof. □

Now we are ready to state our main results on the distribution of $ce(2|1)$. Besides the distribution of the latter we also find the asymptotic distribution of the logit transform of $ce(2|1)$. Our motivation to make this particular type of transform will be clear below.

Theorem 3.1. *Let the logit transform of $ce(2|1)$ be defined by*

$$\ell(2|1) = \log \frac{ce(2|1)}{1 - ce(2|1)}$$

and let the operator $\nabla_{(u,v)}(\cdot)$ for a function $f(u, v)$ be defined as

$$\nabla_{(u,v)} f(u, v) = \left(\frac{\partial f}{\partial u}, \frac{\partial f}{\partial v} \right)'.$$

Then in the framework (A1), (A2) and (B3) $ce(2|1)$ and $\ell(2|1)$ are asymptotically normal, i.e.

$$\text{(i) } \sqrt{n}(ce(2 \mid 1) - e_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \tau^2\right),$$

$$\text{(ii) } \sqrt{n}(\ell(2 \mid 1) - \ell_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \tau_\ell^2\right)$$

with

$$e_0 = \Phi\left(\frac{U_0^* + c}{V_0^{*1/2}} \right), \qquad \ell_0 = \log \frac{e_0}{1 - e_0}, \qquad \tau^2 = \nabla'_{(U_0^*, V_0^*)} \Theta \nabla_{(U_0^*, V_0^*)}, \qquad \tau_\ell^2 = \frac{\tau^2}{(1 - e_0) e_0},$$

where $\nabla_{(U_0^*, V_0^*)}$ is defined as

$$\nabla_{(U_0^*, V_0^*)} = \left(V_0^{*-1/2} \phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}} \right), -\frac{(U_0^* + c)}{2 V_0^{*3/2}} \phi\left(\frac{U_0^* + c}{\sqrt{V_0^*}} \right) \right)'.$$

Proof. By using asymptotic normality of $(\widetilde{U}, V)$ and by applying Lemma A.4 (see supplemental material, Appendix A) to the function

$$g(\widetilde{U}, V) = \Phi\left(\frac{\widetilde{U}+c}{V^{1/2}}\right)$$

it easily follows that

$$\nabla_{(\tilde{u},v)} g(\tilde{u}, v) = \left(\frac{\partial g}{\partial \tilde{u}}, \frac{\partial g}{\partial v}\right)' = \left(v^{-1/2}\phi\left(\frac{\tilde{u}+c}{\sqrt{v}}\right), -\frac{(\tilde{u}+c)}{2v^{3/2}}\phi\left(\frac{\tilde{u}+c}{\sqrt{v}}\right)\right)'.$$

Then we obtain

$$\sqrt{n}(g(\widetilde{U}, V) - \Phi((U_0^*+c){V_0^*}^{-1/2})) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \nabla'_{(U_0^*,V_0^*)}\Theta\nabla_{(U_0^*,V_0^*)}\right).$$

The statement (ii) can be proven similarly. □

Now we are ready to explore the determination method **M2** which chooses the cut-off point c to get the desired level of confidence $1-\beta$ of a pre-specified upper bound eu. By the asymptotic normality of $ce(2|1)$ and $\ell(2|1)$, we propose to set the cut-off points for the EDDR using $T(\boldsymbol{x})$ as

$$\begin{aligned}&\text{(i)} \quad c_{2,1} \text{ s.t. } c_{2,1} = -U_0^* + {V_0^*}^{1/2} z_{\gamma},\\ &\text{(ii)} \quad c_{2,2} \text{ s.t. } c_{2,2} = -U_0^* + {V_0^*}^{1/2} z_{\gamma_\ell},\end{aligned} \tag{3.4}$$

where

$$\gamma = eu - \frac{\tau}{\sqrt{n}} z_{1-\beta}, \qquad \gamma_\ell = \frac{eu}{(1-eu)\exp(\tau_\ell z_{1-\beta}/\sqrt{n}) + eu}.$$

Remark 3.1. If $\gamma \notin [0, 1]$ then (i) is not defined. This motivates our logit transform of $ce(2|1)$ which yields the result (ii) where $\gamma_\ell \in [0, 1]$ always.

For practical use, the unknown parameters $\Delta_0^*, \Delta_1^*, \Delta_2^*, \Delta_3^*, a_1^*, a_2^*, a_3^*$ and a_4^* in (i)–(ii) should be replaced by their consistent estimators. To ensure consistency, the asymptotic framework (A1)–(A3) is modified by replacing (A3) with

$$(\mathrm{B}'3):\ 0 < \lim_{p\to\infty} a_i = a_i^* < \infty \quad (i = 3, \dots, 8), \qquad 0 < \lim_{p\to\infty} \Delta_i = \Delta_i^* < \infty \quad (i = 2, \dots, 7).$$

By the consistency results of Lemma A.5 and A.6 (see supplemental material, Appendix A), obtained under the assumptions (A1), (A2) and (B′3), we propose estimators for **M2**-based cut-off points derived in (3.4), as

$$\begin{aligned}&\text{(i)} \quad \hat{c}_{2,1} \text{ s.t. } \hat{c}_{2,1} = -\widehat{U}_0 + \widehat{V}_0^{1/2} z_{\hat{\gamma}},\\ &\text{(ii)} \quad \hat{c}_{2,2} \text{ s.t. } \hat{c}_{2,2} = -\widehat{U}_0 + \widehat{V}_0^{1/2} z_{\hat{\gamma}_\ell},\end{aligned} \tag{3.5}$$

where

$$\hat{\gamma} = eu - \frac{\hat{\tau}}{\sqrt{n}} z_{1-\beta}, \qquad \hat{\gamma}_\ell = \frac{eu}{(1-eu)\exp(\hat{\tau}_\ell z_{1-\beta}/\sqrt{n}) + eu}.$$

4. Simulation study

We now turn to numerical evaluation of the asymptotic results and the suggested cut-off points. The goal of the simulation experiment is threefold: to investigate the finite sample behavior of newly derived asymptotic approximations, to compare the performance of our approach under independence with that for dependent data with various dependence strength, and to investigate the effect of choice of the confidence level in combination with the upper bound specification.

The data sets for each $\Pi_g,\ g = 1, 2$ are independently generated as

$$\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \dots, \boldsymbol{x}_{1N_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu_1, \Sigma), \qquad \boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \dots, \boldsymbol{x}_{2N_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu_2, \Sigma), \tag{4.1}$$

respectively. To assess the performance for dependent data, Σ will be assumed to have band correlation $\Sigma = (\sigma_{ij})$,

$$\sigma_{ij} = \begin{cases}\rho^{|i-j|}, & |i-j| \le 50,\\ 0, & |i-j| > 50,\end{cases}$$

with ρ ranging from 0 to 0.5, which is chosen to fulfill the condition (A2). To constrain the classification complexity, we set

$$\Sigma^{-1/2}\mu_1 = (p)^{-1/2}(5^{1/2}, 5^{1/2}, \dots, 5^{1/2})' \quad \text{and} \quad \mu_2 = (0, 0, \dots, 0)',$$

through the whole simulation experiment.

Table 1
Misclassification probability of EDDR based on $\hat{c}_1$.

N	p				
	64	128	256	512	1024
$\Sigma = I$					
64	0.201793	0.202039	0.201824	0.201865	0.201870
128	0.200916	0.201113	0.201099	0.200896	0.200811
256	0.200412	0.200442	0.200533	0.200539	0.200460
$\rho = 0.2$					
64	0.201434	0.201462	0.201832	0.201741	0.201876
128	0.200718	0.201274	0.200788	0.200826	0.200614
256	0.200245	0.200183	0.200339	0.200457	0.200174
$\rho = 0.5$					
64	0.202088	0.200184	0.200799	0.201209	0.201662
128	0.200979	0.199646	0.200136	0.200289	0.200633
256	0.200343	0.199499	0.199728	0.199936	0.200197

To evaluate the effect of high-dimensionality and sample size, we let $p = 64, 128, 256, 512, 1024$ and $N_1 = N_2,\ N = 64, 128, 256$ for each choice of ρ.

First, as in the previous sections, we focus without loss of generality on evaluation of $ce(2|1)$. For each triple (p, N, ρ), we generate data according to (4.1), apply EDDR given by $T(\boldsymbol{x})$ in (1.3) with both **M1**-based cut-offs, $\hat{c}_1$ established in Section 2, and repeat the whole process independently 10^5 times. As a result, we get 10^5 conditional classification errors of $T(\boldsymbol{x})$:

$$C^{(i)} = \Phi\left(\frac{U^{(i)} + (N_2^{-1} - N_1^{-1})p\hat{a}_1^{(i)}/2 + \hat{c}_1^{(i)}}{\sqrt{V^{(i)}}}\right), \quad i = 1, \ldots, 10^5,$$

which after averaging provides *attained error rate*

$$ae(\hat{c}_1) = \frac{1}{10^5}\sum_{i=1}^{10^5} C^{(i)}.$$

This result, being summarized in Table 1, suggest that the EDDR based on $\hat{c}_1$ is optimally adaptive in a sense that its performance accuracy is closely approaching the actual value of the misclassification probability, $\alpha = 0.2$. Stably good results are obtained when varying the dependence strength ρ, in both large sample and high-dimensional cases. This provides a finite sample justification of the asymptotic framework (A1)–(A3) which allows for both $p > N$ and $p < N$ scenarios. As expected, by our asymptotic results obtained in Theorem 2.1 along with consistency of estimators of U_0^* and V_0^* suggested for the practical use of the cut-off c_1, the classification procedure remains accurate even when p grows.

To evaluate the performance of the **M2**-based cut-offs we use the simulation setting (4.1), with the same variety of covariance strength, $\beta = 0.05$, and two values of $eu = 0.2$ representing the upper bound on the actual misclassification probability. Then for each setting, the performance of the classification procedure by $T(\boldsymbol{x})$ with cut-offs $\hat{c}_{2,1}$ and $\hat{c}_{2,2}$ given in (3.5) is analyzed. Proceeding with the same simulation strategy as above for each cut-off choice, we consider the *attained confidence level*

$$acl(\hat{c}_{2,i}) = \frac{\#\left\{\Phi\left(\{U + (N_2^{-1} - N_1^{-1})\hat{a}_1/2 + \hat{c}_{2,i}\}/\sqrt{V}\right) \le eu\right\}}{10^5}, \quad i = 1, 2,$$

which is obtained by averaging the observed confidence level of $ce(2|1)$ of $T(\boldsymbol{x})$ with $\hat{c}_{2,i}$ for each, i, over 10^5 independent replicates of the data generation step, estimation of parameters and classification.

Performance results, being summarized in Table 2 indicate that the cuf-off $\hat{c}_{2,2}$ based on the logit transform is in general conservative and provides better classification accuracy than $\hat{c}_{2,1}$ in both large sample and high-dimensional settings. Note that Theorem 3.1, the classification procedure based on $\hat{c}_{2,1}$ and $\hat{c}_{2,2,}$ is expected to provide an accurate asymptotic performance due to the properties of the logit transform and due to consistency of the estimators of the unknown parameters suggested in Remark 3.1. The consistency is stated under rather general asymptotic framework proposed in (B′3), covering both large sample and high-dimensional cases. These asymptotic findings are completely supported by stably good classification performance obtained for finite sample cases with both $p \geq N$ and $p < N$ for various choices of the dependence strength ρ.

5. Conclusion

This paper contributes to the asymptotic analyses of the EDDR performance in high-dimensional data, with particular focus on determining a cut-off point to adjust the probabilities of misclassification. Two generic cut-off determination

Table 2
Attained confidence level.

N		p				
		64	128	256	512	1024
$\Sigma = I$						
64	$acl(\hat{c}_{2,1})$	0.934	0.931	0.935	0.934	0.933
	$acl(\hat{c}_{2,2})$	0.956	0.954	0.956	0.956	0.955
128	$acl(\hat{c}_{2,1})$	0.939	0.938	0.938	0.938	0.938
	$acl(\hat{c}_{2,2})$	0.954	0.953	0.954	0.953	0.953
256	$acl(\hat{c}_{2,1})$	0.944	0.943	0.942	0.941	0.942
	$acl(\hat{c}_{2,2})$	0.954	0.953	0.952	0.952	0.953
$\rho = 0.2$						
64	$acl(\hat{c}_{2,1})$	0.936	0.935	0.932	0.934	0.933
	$acl(\hat{c}_{2,2})$	0.958	0.957	0.955	0.956	0.955
128	$acl(\hat{c}_{2,1})$	0.940	0.936	0.940	0.939	0.940
	$acl(\hat{c}_{2,2})$	0.955	0.958	0.955	0.953	0.954
256	$acl(\hat{c}_{2,1})$	0.944	0.944	0.944	0.943	0.941
	$acl(\hat{c}_{2,2})$	0.955	0.955	0.954	0.953	0.952
$\rho = 0.5$						
64	$acl(\hat{c}_{2,1})$	0.932	0.939	0.938	0.935	0.933
	$acl(\hat{c}_{2,2})$	0.954	0.962	0.960	0.958	0.956
128	$acl(\hat{c}_{2,1})$	0.938	0.945	0.941	0.943	0.939
	$acl(\hat{c}_{2,2})$	0.953	0.960	0.957	0.958	0.954
256	$acl(\hat{c}_{2,1})$	0.943	0.948	0.947	0.946	0.943
	$acl(\hat{c}_{2,2})$	0.953	0.959	0.958	0.957	0.954

approaches, **M1** based on the expected error and **M2** based on the upper bound of the actual misclassification probability, *eu* with the specified confidence level $1 - \beta$, are proposed.

To establish the cut-off by **M1**, an approximation of the expected misclassification rate along with its asymptotic unbiased estimator, is derived; our result extends the approach of Anderson [1] by considering a more general asymptotic set-up that allows $p > N$. Subsequently, the cut-off based on the main term of the asymptotic expression is suggested.

To set up the cut-off based on **M2**, the asymptotic normality of the conditional misclassification rate and its logit transform are established for a given β and *eu* in high-dimensions. Based on the asymptotic results, two types of cut-offs are also established. Our newly derived results extend the asymptotic consideration by McLachlan [5] to a high-dimensional case.

For both **M1** and **M2** approaches, the practically workable expressions of the theoretical cut-offs are established, for which we obtain consistent and asymptotic unbiased estimators of a set of unknown parameters. The validity of the new asymptotic results in a finite sample case is numerically shown by applying the cut-offs in the suggested EDDR classifier $T(\boldsymbol{x})$ for a range of confidence levels, various strength of correlation and a set of p and N values.

As both suggested cut-off determination procedures demonstrate stably good accuracy in high dimensions, they can generally be recommended for practical applications in distance-based classifiers, with EDDR as special case, when it is desired to set a cut-off point to achieve a specified misclassification rate.

Acknowledgments

We would like to thank the associate editor and two reviewers for many valuable comments and helpful suggestions which led to an improved version of this paper. The authors thank Professor Makoto Aoshima and Professor Yasunori Fujikoshi for extensive discussions, references and encouragements. The research of Tatjana Pavlenko is in part supported by the grant 2013-45266 VR of Sweden. The research of Takashi Seo was supported in part by Grant-in-Aid for Scientific Research (C) (23500360). The research of Masashi Hyodo is in part supported by the Stiftelsen G.S. Magnuson travel grant (2013), the Royal Swedish Academy of Sciences.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2015.05.008.

References

[1] T.W. Anderson, An asymptotic expansion of the distribution of the studentized classification statistic W, Ann. Statist. 1 (1973) 964–972.
[2] M. Aoshima, K. Yata, A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, Ann. Inst. Statist. Math. 66 (2014) 983–1010.
[3] K. Matusita, Decision rules, based on the distance for problems of fit, two samples, and estimation, Ann. Math. Statist. 26 (1955) 631–640.
[4] K. Matusita, M. Motoo, On the fundamental theorem for the decision rule based on distance $\| \ \|$, Ann. Inst. Statist. Math. 7 (1956) 137–142.

[5] G.J. McLachlan, Constrained sample discrimination with the Studentized classification statistic W, Comm. Statist. Theory Methods 6 (1977) 575–583.
[6] N. Shutoh, M. Hyodo, T. Pavlenko, T. Seo, Constrained linear discriminant rule via the Studentized classification statistic based on monotone missing data, SUT J. Math. 48 (2012) 55–69.
[7] M.S. Srivastava, Some tests concerning the covariance matrix in high dimensional data, J. Japan Statist. Soc. 35 (2005) 251–272.
[8] M.S. Srivastava, Minimum distance classification rules for high dimensional data, J. Multivariate Anal. 97 (2006) 2057–2070.