SUT Journal of Mathematics
Vol. 55, No. 1 (2019), 11–23

An estimator of misclassification probability for multi-class Euclidean distance classifier in high-dimensional data

Hiroki Watanabe, Takashi Seo and Masashi Hyodo

(Received May 28, 2018; Revised March 5, 2019)

Abstract. Aoshima and Yata [1] observed that classification accuracy of Euclidean distance-based classifiers have good performance at high dimensions. For practical use, it is necessary to estimate the misclassification probability using the training data set. Although cross-validation is usually used for such problems, it does not necessarily have good estimation accuracy at high dimension. In this paper, we propose a new estimator of misclassification probabilities at high-dimensional settings. Our estimator is obtained using the asymptotic multivariate normality of discriminant functions at high-dimensional settings. Finally, we numerically justify the high accuracy of our proposed estimator in finite sample applications, inclusive of high-dimensional scenarios.

AMS 2010 *Mathematics Subject Classification.* 62H12, 62H30.

Key words and phrases. Asymptotic normality, Euclidean distance-based classifier, multi-class classification, probability of misclassification.

§1. Introduction

In this paper, we focus on the multi-class classification concerned with the allocation of a p-dimensional random vector $\boldsymbol{X}$ to one of the several populations, $G_1, G_2, \ldots, G_q$ using training data sets

$$\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}, \boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}, \ldots, \boldsymbol{X}_{q1}, \boldsymbol{X}_{q2}, \ldots, \boldsymbol{X}_{qn_q}.$$

Here, $\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in_i}$ is a p-dimensional random sample from the i-th population G_i. For $p \le \sum_{i=1}^{q} n_i - q$, a natural extension of Fisher linear discriminant exists using multiple discriminant analysis (See, Johnson and Wichern [8]). However, when $p > \sum_{i=1}^{q} n_i - q$, it cannot be used due to the singularity of pooled sample covariance matrix. In this case, the Euclidean distance-based classifier is often used. Recently, Aoshima and Yata [1] claimed that classification accuracy of Euclidean distance-based classifiers have good performance at high dimensions with some assumptions. Furthermore, Aoshima and Yata [3] proposed a quadratic discriminant function in non-sparse setting and discusses its asymptotic properties. This discriminant rules can be discriminated by using both the differences of mean vectors and covariance matrices.

We focus on evaluating the probabilities of misclassification of Euclidean distance-based classifiers. Let $\boldsymbol{X}$ be test data generated from one of the several populations $G_1, G_2, \dots, G_q$. The Euclidean distance-based discriminant function is defined as

$$U_{ji} = \|\boldsymbol{X} - \overline{\boldsymbol{X}}_j\|^2 - \|\boldsymbol{X} - \overline{\boldsymbol{X}}_i\|^2 - \frac{\mathrm{tr}(S_j)}{n_j} + \frac{\mathrm{tr}(S_i)}{n_i}$$

for $i \neq j,\ i, j \in \{1, 2, \dots, q\}$, where S_i and S_j are the sample covariance matrices. Using this function, the classification rule for test data $\boldsymbol{X}$ is given by

$$\boldsymbol{X} \in \mathcal{R}_k \Rightarrow \boldsymbol{X} \sim G_k.$$

where, the region $\mathcal{R}_k$ $(k \in \{1, 2, \dots, q\})$ is defined by

$$\mathcal{R}_k = \{\boldsymbol{X} \in \mathbb{R}^p\ ;\ U_{jk} > 0,\ j = 1, 2, \dots, q,\ j \neq k\}.$$

Here, the notation "$\boldsymbol{X} \sim G_\ell$" means $\boldsymbol{X}$ generated from G_ℓ. Then, the misclassification probability of an observation from G_k is

$$e_k = 1 - \Pr(\boldsymbol{X} \in \mathcal{R}_k | \boldsymbol{X} \sim G_k).$$

However, it is generally difficult to obtain an exact value for e_k. Therefore, there are many studies dealing with asymptotic approximations for misclassification probability. For the Fisher linear discriminant rule, see, e.g., Fujikoshi and Seo [5]. Tonda et al. [13] derived asymptotic unbiased estimator of misclassification probability for two-class linear discriminant rule. Shutoh [9, 10, 11] and Shutoh and Seo [12] discussed the asymptotic properties of two-class linear discriminant function for monotone missing samples. The approximation of e_k is derived by using the asymptotic normality of the discriminant function U_{ji}. Aoshima and Yata [1] showed the asymptotic normality of U_{ji} under some conditions, and obtained the approximation of the upper bound of e_k by combining the asymptotic normality and Boole's inequality. Aoshima and Yata [3] showed the different type of asymptotic normality of U_{ji}. Recently, Watanabe et al. [14] obtained a plug-in type estimator of e_k by estimating unknown parameters contained in the approximate value of e_k for two-class classification. We extend Watanabe et al.'s results to multiple groups. For that, $(q-1)$-dimensional distribution of $(U_{1k}, \ldots, U_{k-1k}, U_{k+1k}, \ldots, U_{qk})^\top$ is necessary. In general it is difficult to derive an exact distribution, so we consider an asymptotic distribution. The asymptotic normality of univariate U_{ji} has already been shown in Aoshima and Yata [1, 3]. However, in general, univariate asymptotic normality of U_{ji} is not a sufficient condition of multivariate asymptotic normality of $(U_{1k}, \ldots, U_{k-1k}, U_{k+1k}, \ldots, U_{qk})^\top$. In this paper, we newly obtain the asymptotic multivariate normality for $(U_{21}, U_{31}, \ldots, U_{q1})^\top$. By using this result, it is possible to construct an approximation that is not the upper bound of e_k. Further, we propose the plug-in type estimator of misclassification probability of e_k using asymptotic multivariate normality.

The rest of the paper is organized as follows. In Section 2, we demonstrate the asymptotic multivariate normality for several Euclidean discriminant functions and propose the plug-in estimator of misclassification probability of e_k. In Section 3, we summarize the results of numerical experiments, justifying the validity of the suggested estimators for the data along with a number of high-dimensional scenarios. Finally, in Section 4, we conclude this paper.

§2. Main result

We show the asymptotic multivariate normality for a p-dimensional vector whose components are several Euclidean discriminant functions U_{jk} for $j, k \in \{1, 2, \ldots, q\},\ j \neq k$. For simplicity of notation, we only deal with e_1. Then, $k = 1$ and $j \in \{2, 3, \ldots, q\}$.

2.1. Statistical model and some moments of discriminant function

Assume that the data are generated by the following model:

$$\boldsymbol{X} = \Sigma_1^{1/2} \boldsymbol{Z}_{10} + \boldsymbol{\mu}_1,\ \forall_{\ell \in \{1,2,\ldots,q\}, t \in \{1,2,\ldots,n_\ell\}}\ \boldsymbol{X}_{\ell t} = \Sigma_\ell^{1/2} \boldsymbol{Z}_{\ell t} + \boldsymbol{\mu}_\ell,$$

where $\boldsymbol{e}_s^\top \boldsymbol{Z}_{\ell t}$ are iid random variables s.t. fourth moment is bounded. Here, $\boldsymbol{e}_s$ denotes

$$\boldsymbol{e}_s = (0 \cdots 0 \overset{s}{\check{1}} 0 \cdots 0)^\top.$$

If $\boldsymbol{e}_s^\top \boldsymbol{Z}_{\ell t} \sim \mathcal{N}(0, 1)$, then the condition is trivially true. Under this model, the population mean vector and covariance matrix of $\boldsymbol{X}_{\ell 1}$ are $\mathrm{E}(\boldsymbol{X}_{\ell 1}) = \boldsymbol{\mu}_\ell$ and $\mathrm{var}(\boldsymbol{X}_{\ell 1}) = \Sigma_\ell$, respectively. Let $\forall_{i,i' \in \{1,2,\ldots,q\}}\ \boldsymbol{\delta}_{ii'} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}$. The mean of U_{j1}, variance of U_{j1}, and covariance of $(U_{j1}, U_{j'1})$ for $j, j' \in \{2, 3, \ldots, q\}$ are

$$\begin{aligned}
\mu_j &= \mathrm{E}(U_{j1}) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_j\|^2 = \|\boldsymbol{\delta}_{1j}\|^2, \\
\sigma_j^2 &= \mathrm{var}(U_{j1}) = 4\left\{\Delta_{1j} + \frac{\|\Sigma_1\|_F^2}{n_1} + \frac{\mathrm{tr}(\Sigma_1\Sigma_j) + \Delta_{j1}}{n_j}\right\} + \frac{2\|\Sigma_1\|_F^2}{n_1(n_1-1)} \\
&\qquad + \frac{2\|\Sigma_j\|_F^2}{n_j(n_j-1)}, \\
\sigma_{jj'} &= \mathrm{cov}(U_{j1}, U_{j'1}) = 4\left(\Delta_{1jj'} + \frac{\|\Sigma_1\|_F^2}{n_1}\right),
\end{aligned}$$

respectively. Here, $\Delta_{i_1i_2} = \|\Sigma_{i_1}^{1/2}\ \boldsymbol{\delta}_{i_1i_2}\|^2$ and $\Delta_{i_1i_2i_3} = \boldsymbol{\delta}_{i_1i_2}^\top\Sigma_{i_1}\ \boldsymbol{\delta}_{i_1i_3}$ for $i_1, i_2, i_3 \in \mathbb{N}$.

2.2. Asymptotic normality

To obtain asymptotic normality, we make asymptotic frameworks for some parameters.

Let n_j, $\mathrm{tr}(\Sigma_1\Sigma_j)$, $\mathrm{tr}\{(\Sigma_1\Sigma_j)^2\}$, $\mathrm{tr}\{(\Gamma_1^\top\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Gamma_1) \odot (\Gamma_1^\top\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Gamma_1)\}$, $\|\Sigma_1\|_F$, $\|\Sigma_1^2\|_F$, $\|\Sigma_j\|_F$ and Δ_{1j} be functions of p for $j \in \{2, 3, \ldots, q\}$. Here, "$A \odot B$" denotes Hadamard product of same size matrices A and B. Then, we assume (A1)–(A3).

(A1) $\min\{n_1, n_2, \ldots, n_q\} \to \infty$ and $n_j/n_1 \in (0, \infty)$.

(A2) $\mathrm{tr}(\Sigma_1\Sigma_j)/\|\Sigma_1\|_F^2 \in (0, \infty)$, $\Delta_{j1}/\Delta_{1j} \in (0, \infty)$, $\|\Sigma_j\|_F^2/\|\Sigma_1\|_F^2 \in (0, \infty)$.

(A3) $\|\Sigma_1^2\|_F = o(\|\Sigma_1\|_F^2)$, $\sqrt{\mathrm{tr}\{(\Sigma_1\Sigma_j)^2\}} = o(\mathrm{tr}(\Sigma_1\Sigma_j))$, $\mathrm{tr}\{(\Gamma_1^\top\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Gamma_1) \odot (\Gamma_1^\top\boldsymbol{\delta}_{1j}\boldsymbol{\delta}_{1j}^\top\Gamma_1)\} = o(\Delta_{1j})$.

Here, for a function $f(\cdot)$,"$f(p) \in (0, \infty)$ as $p \to \infty$" implies $\liminf_{p\to\infty} f(p) > 0$, $\limsup_{p\to\infty} f(p) < \infty$. In practical use, the assumption $\|\Sigma_1^2\|_F = o(\|\Sigma_1\|_F^2)$ in (A3) is often not appropriate. This assumption can be called as the non strongly spiked eigenvalue (NSSE) model in Aoshima and Yata [2]. However, it is natural to assume the strongly spiked eigenvalue (SSE) model for microarray data analysis. When NSSE assumption is not satisfied, we recommend a data transformation technique which is proposed in Aoshima and Yata [2]. This transformation reduce the discussion under SSE model to the discussion under NSSE model.

We consider the standardized Euclidean discriminant functions as follows:

$$T_j = \frac{U_{j1} - \mu_j}{\sigma_j}, \qquad \text{for} \quad j \in \{2, 3, \ldots, q\}.$$

We show the asymptotic normality of $\boldsymbol{T} = (T_2, T_3, \ldots, T_q)^\top$. Then, mean vector $\mathrm{E}(\boldsymbol{T}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{T}) = (\rho_{jj'}) =: R$, where $\rho_{jj'} = \sigma_{jj'}/(\sigma_j \sigma_{j'})$. The multivariate asymptotic normality of $\boldsymbol{T}$ is given by the following theorem.

Theorem 1. *Under (A1)–(A3), $\boldsymbol{T} \rightsquigarrow \mathcal{N}_{q-1}(\mathbf{0}, \lim_{p\to\infty} R)$. Here, "$\rightsquigarrow$" denotes the convergence in distribution.*

Proof. From Cramér–Wold theorem(See, Cramér and Wold [4]), it is sufficient to show $\boldsymbol{\beta}^\top \boldsymbol{T} \rightsquigarrow \mathcal{N}(0, \lim_{p\to\infty} \boldsymbol{\beta}^\top R \boldsymbol{\beta})$ for any $q-1$ dimensional nonrandom vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_{q-1})^\top \in \mathbb{R}^{q-1}/\{\mathbf{0}\}$. We define

$$\epsilon_s = 2\sum_{j=2}^{q} \beta_j \sigma_j^{-1} (\boldsymbol{\delta}_{1j}^\top \Sigma_1^{1/2} \boldsymbol{e}_s + \overline{\boldsymbol{Z}}_1^\top \Sigma_1 \boldsymbol{e}_s - \overline{\boldsymbol{Z}}_j^\top \Sigma_j^{1/2} \Sigma_1^{1/2} \boldsymbol{e}_s) Z_s,$$

where $Z_s = \boldsymbol{e}_s^\top \boldsymbol{Z}_{10}$, $\overline{\boldsymbol{Z}}_i = n_i^{-1} \sum_{t=1}^{n_i} \boldsymbol{Z}_{it}$ for $i \in \{1, 2, \ldots, q\}$. Then, under (A1) and (A2), $\boldsymbol{\beta}^\top \boldsymbol{T} = \sum_{s=1}^{p} \epsilon_s + o_p(1)$. Let $\mathcal{F}_0 = \sigma\{\overline{\boldsymbol{Z}}_1, \overline{\boldsymbol{Z}}_2, \ldots, \overline{\boldsymbol{Z}}_q\}$ and $\mathcal{F}_{s-1} = \sigma\{\overline{\boldsymbol{Z}}_1, \overline{\boldsymbol{Z}}_2, \ldots, \overline{\boldsymbol{Z}}_q, Z_1, Z_2, \ldots, Z_{s-1}\}$ for $s \geq 2$. Then, (ϵ_s) is a martingale difference sequence. Under (A1) and (A2), there exists $\lim_{p\to\infty} \boldsymbol{\beta}^\top R \boldsymbol{\beta} (\in (0, \infty))$. Let $\sigma^2 = \lim_{p\to\infty} \boldsymbol{\beta}^\top R \boldsymbol{\beta}$. Also, under (A1) and (A3),

$$\sum_{s=1}^{p} \mathrm{E}(\epsilon_s^2 | \mathcal{F}_{s-1}) = \sigma^2 + o_p(1), \quad \sum_{s=1}^{p} \mathrm{E}(\epsilon_s^4) = o(1).$$

Applying the martingale central limit theorem(See, Hall and Heyde [6]), we prove asymptotic normality of $\boldsymbol{\beta}^\top \boldsymbol{T}$. □

2.3. The estimator of misclassification probability

Using Theorem 1, we propose the asymptotic approximation of misclassification probability as follows:

$$\tilde{e}_1 = 1 - F(\boldsymbol{r}, R), \tag{2.1}$$

where $F(\boldsymbol{r}, R) = \int_{\mathcal{D}} (2\pi)^{-(q-1)/2} |R|^{-1/2} e^{-\boldsymbol{w}^\top R^{-1} \boldsymbol{w}/2} d\boldsymbol{w}$. Here,

$$\mathcal{D} = \{\boldsymbol{w} \in \mathbb{R}^{q-1} \ ; \ \boldsymbol{e}_1^\top \boldsymbol{w} + r_1, \boldsymbol{e}_2^\top \boldsymbol{w} + r_2, \ldots, \boldsymbol{e}_{q-1}^\top \boldsymbol{w} + r_{q-1} > 0\}.$$

and $\boldsymbol{r} = (\mu_2/\sigma_2, \mu_3/\sigma_3, \ldots, \mu_q/\sigma_q)^\top$.

The approximation (2.1) include the unknown values $\|\boldsymbol{\delta}_{1j}\|^2$, $\mathrm{tr}(\Sigma_1\Sigma_j)$, $\|\Sigma_i\|_F^2$, Δ_{1j}, Δ_{j1} and $\Delta_{1jj'}$. We prepare unbiased estimators of these unknown values as follows:

$$
\begin{aligned}
\widehat{\mu}_j &= \|\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_j\|^2 - \frac{\mathrm{tr}(S_1)}{n_1} - \frac{\mathrm{tr}(S_j)}{n_j},\ \widehat{\mathrm{tr}(\Sigma_1\Sigma_j)} = \mathrm{tr}(S_1S_j),\\
\widehat{\|\Sigma_i\|_F^2} &= \frac{(n_i-1)[(n_i-1)(n_i-2)\mathrm{tr}(S_i^2) + \{\mathrm{tr}(S_i)\}^2 - n_iK_i]}{n_i(n_i-2)(n_i-3)},\\
\widehat{\Delta}_{1j} &= V_{1jj} - \frac{2U_{1j}}{(n_1-1)(n_1-2)} - \frac{\mathrm{tr}(S_1S_j)}{n_1}\\
&\qquad + \frac{2n_1K_1 - (n_1-1)\left\{\mathrm{tr}(S_1)\right\}^2 - (n_1-1)^2\mathrm{tr}(S_1^2)}{n_1(n_1-2)(n_1-3)},\\
\widehat{\Delta}_{j1} &= V_{j11} - \frac{2U_{j1}}{(n_j-1)(n_j-2)} - \frac{\mathrm{tr}(S_jS_1)}{n_j}\\
&\qquad + \frac{2n_jK_j - (n_j-1)\left\{\mathrm{tr}(S_j)\right\}^2 - (n_j-1)^2\mathrm{tr}(S_j^2)}{n_j(n_j-2)(n_j-3)},\\
\widehat{\Delta}_{1jj'} &= V_{1jj'} - \frac{U_{1j}+U_{1j'}}{(n_1-1)(n_1-2)}\\
&\qquad + \frac{2n_1K_1 - (n_1-1)\left\{\mathrm{tr}(S_1)\right\}^2 - (n_1-1)^2\mathrm{tr}(S_1^2)}{n_1(n_1-2)(n_1-3)}.
\end{aligned}
$$

Here, for $i_1, i_2, i_3 \in \{1, 2, \ldots, q\}$,

$$
\begin{aligned}
K_{i_1} &= \frac{1}{n_{i_1}-1}\sum_{t=1}^{n_{i_1}} \|\boldsymbol{X}_{i_1t} - \overline{\boldsymbol{X}}_{i_1}\|^2,\\
V_{i_1i_2i_3} &= (\overline{\boldsymbol{X}}_{i_1} - \overline{\boldsymbol{X}}_{i_2})^\top S_{i_1}(\overline{\boldsymbol{X}}_{i_1} - \overline{\boldsymbol{X}}_{i_3}),\\
U_{i_1i_2} &= (\overline{\boldsymbol{X}}_{i_1} - \overline{\boldsymbol{X}}_{i_2})^\top \sum_{t=1}^{n_{i_1}} (\boldsymbol{X}_{i_1t} - \overline{\boldsymbol{X}}_{i_1})(\boldsymbol{X}_{i_1t} - \overline{\boldsymbol{X}}_{i_1})^\top(\boldsymbol{X}_{i_1t} - \overline{\boldsymbol{X}}_{i_1}).
\end{aligned}
$$

The unbiased estimator $\widehat{\Delta}_{1jj'}$ is newly obtained in this paper. The unbiased estimator $\widehat{\|\Sigma_i\|_F^2}$ was proposed by Himeno and Yamada [7]. The unbiased estimator $\widehat{\Delta}_{1j}$ and $\widehat{\Delta}_{j1}$ were proposed by Watanabe et al. [14]. However, some estimators do not always take appropriate values. We note that $\Delta_{1j}, \Delta_{1j'}, \Delta_{j1} > 0$ and $\Delta_{1jj'} \in [-\sqrt{\Delta_{1j}\Delta_{1j'}}, \sqrt{\Delta_{1j}\Delta_{1j'}}]$. We truncate estimators of these parameters so that they take appropriate values. Thus, we obtain $\widetilde{\Delta}_{1j} = \max\{0, \widehat{\Delta}_{1j}\}$, $\widetilde{\Delta}_{1j'} = \max\{0, \widehat{\Delta}_{1j'}\}$, $\widetilde{\Delta}_{j1} = \max\{0, \widehat{\Delta}_{j1}\}$ and

$$
\widetilde{\Delta}_{1jj'} = \min\left\{\sqrt{\widetilde{\Delta}_{1j}\widetilde{\Delta}_{1j'}}, \max\left\{\widehat{\Delta_{1jj'}}, -\sqrt{\widetilde{\Delta}_{1j}\widetilde{\Delta}_{1j'}}\right\}\right\}.
$$

We estimate σ_j and $\sigma_{jj'}$ using the following estimators:

$$\widehat{\sigma}_j = \sqrt{4\left(\widetilde{\Delta}_{1j} + \frac{\widehat{\|\Sigma_1\|_F^2}}{n_1} + \frac{\widehat{\mathrm{tr}(\Sigma_1\Sigma_j)} + \widetilde{\Delta}_{j1}}{n_j}\right) + 2\sum_{i\in\{1,j\}} \frac{\widehat{\|\Sigma_i\|_F^2}}{n_i(n_i-1)}},$$

$$\widehat{\sigma}_{jj'} = 4\left(\widetilde{\Delta}_{1jj'} + \frac{\widehat{\|\Sigma_1\|_F^2}}{n_1}\right).$$

Let $\widehat{\rho}_{jj'} = \widehat{\sigma}_{jj'}/(\widehat{\sigma}_j\widehat{\sigma}_{j'})$. Replacing the unknown value $\rho_{jj'}$ in R, we obtain the estimator $\widehat{R} = (\widehat{\rho}_{jj'})$. We note that the matrix $\widehat{R}$ is always a positive matrix. Moreover, we estimate $\boldsymbol{r}$: $\widehat{\boldsymbol{r}} = (\widehat{\mu}_2/\widehat{\sigma}_2, \widehat{\mu}_3/\widehat{\sigma}_3, \ldots, \widehat{\mu}_q/\widehat{\sigma}_q)^\top$. By substituting each estimator of unknown value in (2.1), we obtain the estimator of misclassification probability as follows:

$$\widehat{e}_1 = 1 - F\left(\widehat{\boldsymbol{r}}, \widehat{R}\right). \tag{2.2}$$

§3. Simulation Studies

We investigate the numerical performances of the asymptotic approximation $\tilde{e}_1$ and its plug-in estimator $\widehat{e}_1$ using Monte Carlo simulation. For simplicity, we treat the discrimination problem among three groups.

First, we investigate the accuracy of the asymptotic approximations

$$\text{(I)} : e_1 \approx \tilde{e}_1, \quad \text{(II)} : e_1 \approx \sum_{j=2}^{3} \Phi\left(-\frac{\|\boldsymbol{\delta}_{1j}\|^2}{\delta_j}\right),$$

where

$$\delta_j = \sqrt{\frac{4\mathrm{tr}(\Sigma_1^2)}{n_1} + \frac{4\mathrm{tr}(\Sigma_1\Sigma_j)}{n_j} + \frac{2\mathrm{tr}(\Sigma_1^2)}{n_1(n_1-1)} + \frac{2\mathrm{tr}(\Sigma_j^2)}{n_j(n_j-1)}}.$$

Here, the approximation (I) represents our proposed method based on (2.1), and the approximation (II) represents the method proposed by Aoshima and Yata [1]. (I) is derived by using the asymptotic multivariate normality for $(U_{21}, U_{31}, \ldots, U_{q1})^\top$ which is obtained Theorem 1. (II) is the approximation of the upper bound of e_1 by combining the asymptotic normality of U_{ij} and Boole's inequality. This approximation is valid under some regularity conditions. Note that (I) approximates e_1 directly, whereas (II) approximates the upper bound of e_1.

The misclassification probability e_1 is calculated via simulation with 100,000 replications.
For the distribution of $\boldsymbol{Z}_{it} = (Z_{i\ell t})$, we set the following two distributions:

$$\begin{aligned} &\text{(D1)} \quad Z_{i\ell t} \sim \mathcal{N}(0,1), \\ &\text{(D2)} \quad Z_{i\ell t} = U_{i\ell t}/\sqrt{5/4} \text{ for } U_{i\ell t} \sim \mathcal{T}_{10}. \end{aligned}$$

Note that (D1) and (D2) satisfies our moment condition. The structure of the covariance matrix is set with the following:

$$\Sigma_1 = \left(0.3^{|i-j|}\right),\ \Sigma_2 = 1.2\left(0.3^{|i-j|}\right),\ \Sigma_3 = 2.4\left(0.3^{|i-j|}\right).$$

We set the mean vectors as following two cases:

$$\begin{aligned} &\text{(M1)} \quad \boldsymbol{\mu}_1 = \boldsymbol{0},\ \boldsymbol{\mu}_2 = \left(\sqrt{30/p}, \sqrt{30/p}, \ldots, \sqrt{30/p}\right)^\top,\ \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2, \\ &\text{(M2)} \quad \boldsymbol{\mu}_1 = \boldsymbol{0},\ \boldsymbol{\mu}_2 = (-1,1,-1,1,\ldots,-1,1,0,\ldots,0)^\top,\ \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2. \end{aligned}$$

Here, in (M2), the number of non-zero elements in $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ is $\lceil\{\mathrm{tr}(\Sigma_1^2)\}^{1/2}/2\rceil$. The dimensions and sample sizes are chosen as follows:

$$p = 100, 250, 500, 1000;\ (n_1, n_2, n_3) = (20, 40, 60),\ (40, 80, 120),\ (60, 120, 180).$$

Then, we compare the true value e_1, the approximation (I) and the approximation (II) on these settings. By comparing the approximations in Table 1, it is seen that approximation (I) is closer to the true value e_1 than (II) is for all cases. In Table 2, (I) and (II) are close to true value e_1 when the sample size is relatively small, and (I) is closer to the true value than (II) when the sample size is relatively large. In situations where the dimension p is large and sample size n is small, (II) is close to the true value e_1 and is a conservative approximation.

Next, we investigate the mean squared error (MSE) of the consistent estimator $\widehat{e}_1$ on the same settings. For comparison, we consider the leave-one-out cross-validation method (CV), which is a popular method for estimating prediction errors for small samples. The MSEs of the estimators CV and $\widehat{e}_1$ are given in Tables 3-6. These tables show that $\widehat{e}_1$ has smaller MSEs than CV does for all cases. Moreover, it can be confirmed that the MSE of our estimator is not influenced even if the distribution of $Z_{i\ell t}$ is a t distribution.

Table 1: Comparison of approximations when (M1)

			(n_1, n_2, n_3)		
p			(20,40,60)	(40,80,120)	(60,120,180)
100	e_1	(D1)	0.0663	0.0552	0.0508
		(D2)	0.0662	0.0543	0.0515
	approx	(I)	0.0683	0.0557	0.0516
		(II)	0.0000	0.0000	0.0000
250	e_1	(D1)	0.0998	0.0702	0.0627
		(D2)	0.0984	0.0721	0.0627
	approx	(I)	0.1007	0.0718	0.0623
		(II)	0.0034	0.0000	0.0000
500	e_1	(D1)	0.1489	0.0979	0.0795
		(D2)	0.1469	0.0977	0.0796
	approx	(I)	0.1499	0.0983	0.0799
		(II)	0.0377	0.0032	0.0003
1000	e_1	(D1)	0.2217	0.1474	0.1134
		(D2)	0.2217	0.1464	0.1134
	approx	(I)	0.2229	0.1472	0.1146
		(II)	0.1414	0.0367	0.0104

Table 2: Comparison of approximations when (M2)

			(n_1, n_2, n_3)		
p			(20,40,60)	(40,80,120)	(60,120,180)
100	e_1	(D1)	0.3598	0.2854	0.2482
		(D2)	0.3573	0.2879	0.2459
	approx	(I)	0.3642	0.2909	0.2500
		(II)	0.3517	0.1855	0.1043
250	e_1	(D1)	0.3433	0.2629	0.2125
		(D2)	0.3460	0.2604	0.2150
	approx	(I)	0.3485	0.2626	0.2135
		(II)	0.3772	0.2093	0.1235
500	e_1	(D1)	0.3225	0.2313	0.1750
		(D2)	0.3232	0.2303	0.1753
	approx	(I)	0.3258	0.2307	0.1770
		(II)	0.3672	0.1998	0.1158
1000	e_1	(D1)	0.3183	0.2175	0.1613
		(D2)	0.3171	0.2169	0.1633
	approx	(I)	0.3201	0.2190	0.1617
		(II)	0.3773	0.2094	0.1236

Table 3: Comparison of MSEs when (M1) and (D1)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0016	0.0007	0.0004
	CV	0.0032	0.0013	0.0008
250	$\widehat{e}_1$	0.0022	0.0008	0.0005
	CV	0.0044	0.0017	0.0010
500	$\widehat{e}_1$	0.0030	0.0011	0.0006
	CV	0.0060	0.0022	0.0012
1000	$\widehat{e}_1$	0.0040	0.0015	0.0008
	CV	0.0078	0.0030	0.0016

Table 4: Comparison of MSEs when (M1) and (D2)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0016	0.0007	0.0004
	CV	0.0031	0.0013	0.0008
250	$\widehat{e}_1$	0.0022	0.0008	0.0005
	CV	0.0044	0.0016	0.0010
500	$\widehat{e}_1$	0.0030	0.0011	0.0006
	CV	0.0060	0.0021	0.0012
1000	$\widehat{e}_1$	0.0041	0.0015	0.0008
	CV	0.0078	0.0030	0.0016

Table 5: Comparison of MSEs when (M2) and (D1)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0042	0.0022	0.0014
	CV	0.0091	0.0041	0.0026
250	$\widehat{e}_1$	0.0042	0.0021	0.0012
	CV	0.0090	0.0038	0.0023
500	$\widehat{e}_1$	0.0041	0.0019	0.0011
	CV	0.0087	0.0036	0.0020
1000	$\widehat{e}_1$	0.0042	0.0017	0.0010
	CV	0.0088	0.0035	0.0019

Table 6: Comparison of MSEs when (M2) and (D2)

		(n_1, n_2, n_3)		
p		(20,40,60)	(40,80,120)	(60,120,180)
100	$\widehat{e}_1$	0.0042	0.0022	0.0014
	CV	0.0091	0.0041	0.0026
250	$\widehat{e}_1$	0.0042	0.0021	0.0013
	CV	0.0090	0.0039	0.0023
500	$\widehat{e}_1$	0.0041	0.0019	0.0011
	CV	0.0088	0.0036	0.0020
1000	$\widehat{e}_1$	0.0042	0.0017	0.0010
	CV	0.0088	0.0035	0.0019

§4. Conclusion

We considered the multi-class classification problem for high-dimensional data. In this paper, we showed the asymptotic multivariate normality for several Euclidean distance-based discriminant functions under high-dimensional settings. Our theoretical results have been established under variance heterogeneity and nonnormality. Further, using asymptotic multivariate normality, we proposed a new estimator of misclassification probability of Euclidean distance-based discriminant rule. We confirmed that proposed estimators have good performances in high-dimensional situations through numerical simulations.

Acknowledgments

The authors would like to thank Mr. Hiroshi Furukawa for his help. The third author's research was supported in part by a Grant-in-Aid for Young Scientists (B) (17K14238) from the Japan Society for the Promotion of Science.

References

[1] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Ann. Inst. Stat. Math.*, **66**, 983–1010.

[2] Aoshima, M. and Yata, K. (2019a). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Ann. Inst. Stat. Math.*, **71**, 473–503.

[3] Aoshima, M. and Yata, K. (2019b). High-dimensional quadratic classifiers in non-sparse settings. *Methodol Comput. Appl. Probab.*, **21**, 663–682.

[4] Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.*, s1-**11**, 290–294.

[5] Fujikoshi, Y. and Seo, T. (1998). Asymptotic approximations for EPMC's of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large. *Random Operators and Stochastic Equations*, **6**, 269-280.

[6] Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and its Application.* Academic Press, New York.

[7] Himeno, T. and Yamada, T. (2014). Estimations for some functions of covariance matrix in high dimension under non-normality and its applications. *J. Multivariate Anal.*, **130**, 27–44.

[8] Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis*, 2nd ed. Prentice Hall.

[9] Shutoh, N. (2011). Asymptotic expansions relating to discrimination based on two-step monotone missing samples. *J. Statist. Plann. Inference*, **141**, 1297–1306.

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

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

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

[13] Tonda, T., Nakagawa, T. and Wakaki, H. (2017). EPMC estimation in discriminant analysis when the dimension and sample sizes are large. *Hiroshima Math. J.*, **47**, 43–62.

[14] Watanabe, H., Hyodo, M., Yamada, Y. and Seo, T. (2019). Estimation of misclassification probability for a distance-based classifier in high-dimensional data. *Hiroshima Math. J.*, **49**, 175–193.

Hiroki Watanabe
Department of Health Sciences, Oita University of Nursing and Health Sciences
2944-9, Megusuno, Oita-shi, Oita 870-1201, Japan
E-mail: `watahiro919@gmail.com`

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

Masashi Hyodo
Department of Mathematical Sciences, Graduate School of Science
Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai-shi, Osaka 599-8531, Japan
E-mail: caicmhy@gmail.com