Journal of the Korean Data & Information Science Society
2017, **28**(5), 1179–1189

http://dx.doi.org/10.7465/jkdi.2017.28.5.1179
한국데이터정보과학회지

Discriminant analysis using empirical distribution function

Jae Young Kim[1] · Chong Sun Hong[2]

[1]Research Institute of Applied Statistics, Sungkyunkwan University
[2]Department of Statistics, Sungkyunkwan University

Received 18 July 2017, revised 5 September 2017, accepted 7 September 2017

Abstract

In this study, we propose an alternative method for discriminant analysis using a multivariate empirical distribution function to express multivariate data as a simple one-dimensional statistic. This method turns to be the estimation process of the optimal threshold based on classification accuracy measures and an empirical distribution function of data composed of classes. This can also be visually represented on a two-dimensional plane and discussed with some measures in ROC curves, surfaces, and manifolds. In order to explore the usefulness of this method for discriminant analysis in the study, we conducted comparisons between the proposed method and the existing methods through simulations and illustrative examples. It is found that the proposed method may have better performances for some cases.

Keywords: Accuracy, AUC, classification, HUM, threshold, VUS.

1. Introduction

Discriminant analysis, which was developed by Fisher in the 1930s, is a multivariate analysis method that classifies the characteristics of observations belonging to K (≥ 2) populations and assigns new observations belonging to a specific population out of several populations (Hong, 2012, p. 103). When a population is composed of samples extracted from several classes, the discriminant analysis can be used to classify and identify which classes of individuals have been extracted, and the method has been studied in various ways (Critchley and Vitiello, 1991; Fung, 1992; Fung, 1996; Jhun and Choi, 2009).

In order to apply the discriminant analysis to data, some constraints should be assumed. For example, the linear and quadratic discriminant analyses resort to the normality. In some situation which does not satisfy this assumption, an alternative discriminant analysis method is proposed using the multivariate empirical distribution function defined by Hong *et al.* (2017). The multivariate empirical distribution function has the advantage that multivariate data can be expressed as a one-dimensional distribution function. Therefore,

[1] Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul 03063 Korea.
[2] Corresponding author : Professor, Department of Statistics, Sungkyunkwan University, Seoul 03063 Korea. E-mail: cshong@skku.edu

a discriminant analysis data composed of several classes could be represented as a mixed distribution consisting of multivariate empirical distribution functions for the number of classes, and the discriminant analysis can be performed by estimating and identifying the thresholds of the mixed distributions.

Similar to classifying with a discriminant function in existing discriminant analysis methods, this discriminant analysis method estimates classification points by using a multivariate empirical distribution function of sample data that composes the mixed distribution. This is performed using an optimal threshold maximizing the discriminant power and minimizing the data loss. Furthermore, this discriminant method can be connected to various classification accuracy measures to estimate the optimal threshold, and it is a useful method to evaluate the classification performance of the mixed distribution model.

This method also provides a visual representation of the change in the classification and misclassification rates of the model, so that it has the advantage of enabling a discussion with ROC curves, surfaces and manifolds as well as AUC (area under ROC curve), VUS (volume under the ROC surface), and HUM (hyper-volume under the ROC manifold), which are commonly used in credit evaluation and medical statistics.

The study of the optimal threshold in the mixed distribution model depends on the number of classes in the data. For example, it is assumed that a parameter space is divided into two classes named 'Bad (or default)' and 'Good (or non-default)'. This can be expressed as a 2×2 confusion matrix in order to compare classes in which each object actually belongs and classes resulting from the predicted results. The ROC curve can be expressed when the TPR (true positive rate; or recall, sensitivity) and FPR (false positive rate; or false alarm rate, 1-specificity) respectively correspond to the X and Y axes of the graph (Pepe, 2003; Tasche, 2006). In addition, the AUC statistic is determined that the mixed distribution model is better classified as its value is closer to 1.

There are well known classification accuracy measures such as the J index of Youden (1950), the TA (total accuracy) of Lambert and Lipkovich (2008), the amended closest to (0, 1) of Perkins and Schisterman (2006) and others. In this study, we use the TR (true rate) of Hong and Joo (2010) and BA (balanced accuracy) of Velez *et al.* (2007) which are the arithmetic mean of the sensitivity and specificity. The TR has an advantage in terms of the accuracy because the sum of the misclassifications is smaller than for TA, and it has a linear relationship with the J index, amended closest to (0, 1), and Kolmogorov-Smirnov statistic (Yoo and Hong (2011) for more detail).

Next, when the class is divided into three categories, the ROC surface in three-dimensional space can be expressed. The VUS is the statistic whose value is between ⅙ and 1, and the closer it gets to 1, the higher the classification models' discriminant power becomes. In this case, J_3, AC_3, and TR_3 are defined as the classification accuracy by expanding the method for two classes (Hong and Jung, 2013), and TR_3 is used as a classification accuracy measure for three classes in this study. When there are more than four classes in a sample, the ROC manifold can be extended and constructed in a similar manner. In addition, the discriminant power can be calculated using the HUM which is a classification accuracy measure for multidimensional space (Hong and Jung, 2014). In this study, we will not discuss when there are four or more classes in a sample, but this can be extended to the concepts and methods discussed above.

This paper consists of 4 sections. Section 2 briefly explains the multivariate empirical distribution function proposed by Hong *et al.* (2017) and proposes a process of discriminant analysis using a multivariate empirical distribution function. In Section 3, we perform the proposed discriminant analyses for various kinds of simulation and illustrative examples, and we also explore the advantages and disadvantages via comparison with existing discriminant analysis. Finally, Section 4 summarizes the results of this study and discusses further research.

2. Alternative discriminant analysis

Assuming that the cumulative distribution function for the p (≥ 2)-variate random sample of size $n, \{(X_{1i}, \ldots, X_{pi}),\ i = 1, \ldots, n\}$ is estimated as $\hat{F}(x_1, \ldots, x_p)$, Hong *et al.* (2017) defined the multivariate empirical distribution function as

$$\hat{F}_n \equiv \hat{F}_n(x_1, \ldots, x_p) = \frac{1}{n}\sum_{i=1}^{n} I(\hat{F}_i \leq f),$$

where the random variable $\hat{F}_i = \hat{F}(X_{1i}, \ldots, X_{pi})$ and f is a constant value of $F(x_1, \ldots, x_p)$.

The i^{th} random sample vector in k^{th} class with sample size n_k is expressed as $\{\underline{X_i}^k = (X_{1i}^k, \ldots, X_{pi}^k),\ i = 1, \ldots, n_k, k = 1, \ldots, K\}$. When the mean vector and the variance-covariance matrix of the class k are $\hat{\mu}_k$ and $\hat{\Sigma}_k$ respectively, the weighted mean vector $\hat{\mu}$ is $\sum_{k=1}^{K} \lambda_k \hat{\mu}_k$ and the alternative pooled variance-covariance matrix $\hat{\Sigma}$ of the whole sample can be proposed such as

$$\hat{\Sigma} = [\frac{1}{pK}\sum_{k=1}^{K}(\lambda_k \hat{\Sigma}_k)^{-1}]^{-1},$$

where $\lambda_k = n_k / \sum_{k=1}^{K} n_k$. The alternative discriminant analysis using the multivariate empirical distribution function is performed in the following steps.

Table 2.1 Illustrative data

	(X, Y)	*Class*	$\hat{F}_i$	$\hat{F}_{n_1}^1$	$\hat{F}_{n_2}^2$
1	(5.04, 2.78)	1	0	0.1	0
2	(4.12, 3.53)	1	0.0011	0.2	0
3	(3.06, 4.63)	1	0.0057	0.3	0
4	(3.31, 4.49)	1	0.0085	0.4	0
5	(3.14, 4.76)	2	0.0094	0.4	0.09
6	(4.18, 4.10)	1	0.0148	0.5	0.09
7	(3.58, 5.29)	1	0.0597	0.6	0.09
8	(3.69, 5.09)	1	0.0623	0.7	0.09
9	(4.81, 4.51)	1	0.1024	0.8	0.09
10	(4.51, 4.86)	1	0.1662	0.9	0.09
11	(5.53, 4.58)	1	0.1722	1	0.09
12	(4.71, 5.13)	2	0.3035	1	0.18
13	(4.41, 5.75)	2	0.3451	1	0.27
14	(4.66, 5.33)	2	0.3572	1	0.36
15	(4.69, 5.78)	2	0.4771	1	0.45
16	(4.96, 5.49)	2	0.5184	1	0.55
17	(4.92, 6.65)	2	0.6420	1	0.64
18	(6.01, 5.49)	2	0.7401	1	0.73
19	(5.46, 6.24)	2	0.8416	1	0.82
20	(6.57, 5.70)	2	0.8585	1	0.91
21	(6.03, 6.45)	2	0.9577	1	1

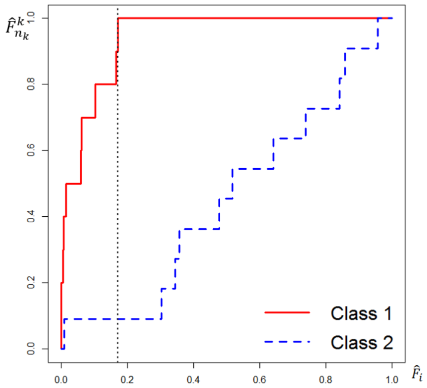

Figure 2.1 $(\hat{F}_i, \hat{F}_{n_k}^k)$ plot

Discriminant analysis using empirical distribution function

Step 1. For i^{th} random sample vector $\underline{x_i}^k$, the normal cumulative distribution function,

$$\hat{F}_i \equiv \Phi(\underline{x_i}^k; \hat{\mu}, \hat{\Sigma}),\ i = 1, \ldots, n_1, \ldots, n_K,$$

is estimated with the sample's weighted mean vector $\hat{\mu}$ and the sample's variance-covariance matrix $\hat{\Sigma}$.

Step 2. The empirical distribution function of k class, $\hat{F}^k_{n_k}$, is calculated as follows.

$$\hat{F}^k_{n_k} = \hat{F}(\underline{x_i}^k) = \frac{1}{n_k}\sum_{j=1}^{n_k} I(\hat{F}_j \leq f),\ i = 1, \ldots, n_k.$$

Step 3. The discriminant analysis is performed using $\{(\hat{F}_i, \hat{F}^k_{n_k}), k = 1, \ldots, K\}$. Among many classification accuracy measures, the threshold could be determined using the following true rate:

$$\mathrm{TR}_k = \frac{2}{k}[\sum_{i=1}^{k-1} TR_{i,i+1} - \frac{(k-2)}{2}] \text{ for all } K \text{ classes,}$$

where $\mathrm{TR}_{a,b}$ is the true rate for $\hat{F}^a_{n_a}$ and $\hat{F}^b_{n_b}$.

We assume random sample of size 21 in the second column of Table 2.1, which consists of 10 ($= n_1$) samples in Class 1 and 11 ($= n_2$) samples in Class 2 ($K = 2$) to understand the discriminant analysis proposed in this study. The normal cumulative distribution $\{\hat{F}_i = \Phi(x_i, y_i; \hat{\mu}, \hat{\Sigma}), i = 1, \ldots, 21\}$ is shown in the fourth column of Table 2.1 using the weighted mean vector $\hat{\mu} = \begin{pmatrix}4.638 & 5.078\end{pmatrix}^{\mathrm{T}}$ and variance-covariance matrix $\hat{\Sigma} = \begin{pmatrix}0.5896 & 0.0413\\ 0.0413 & 0.3230\end{pmatrix}$.

In Step 2, the bivariate empirical distribution function of the k^{th} class, $\{\hat{F}^k_{n_k}, k = 1, 2\}$, corresponding to $\hat{F}_i$ is obtained and summarized in the fifth and sixth columns of Table 2.1. Note that $\hat{F}^1_{n_1}$ increases by 0.1 ($= 1/n_1$) if the sample belongs to Class 1, and $\hat{F}^2_{n_2}$ increases by 0.09 ($= 1/n_2$) if the sample belongs to Class 2.

In Step 3, we used the classification accuracy TR in this study to discriminate using $\hat{F}^k_{n_k}$. The TR is defined as the maximum of the arithmetic mean of the sensitivity and specificity, that can be represented as $\hat{F}^1_{n_1}$ and $(1 - \hat{F}^2_{n_2})$, respectively, in Table 2.1 (see Hong and Joo (2010) for more detail). And for K = 2, 3, and 4, the TR could be derived as the followings :

$$\begin{aligned}
\mathrm{TR}_2 &= \mathrm{TR}_{1,2} && \text{for } K = 2,\\
\mathrm{TR}_3 &= \frac{2}{3}(\sum\nolimits_{i=1}^{2} TR_{i,i+1} - \frac{1}{2}) && \text{for } K = 3,\\
\mathrm{TR}_4 &= \frac{2}{4}(\sum\nolimits_{i=1}^{3} TR_{i,i+1} - 1) && \text{for } K = 4.
\end{aligned}$$

And the TR has the following linear relationship with J index, amended closest to (0,1)

(AC) and Kolmogorov-Smirnov (KS) statistics:

$$\begin{aligned}\text{TR} &= \tfrac{1}{2}(1+J)\\ &= \tfrac{1}{2}(1+\text{KS}), \quad \textit{where } J = \text{KS} = max[\hat{F}^1_{n_1} - \hat{F}^2_{n_2}],\\ &= 1-\tfrac{1}{2}\text{AC}, \quad \textit{where } \text{AC} = \sqrt{\frac{(\hat{F}^2_{n_2})^2 + (1-\hat{F}^1_{n_1})^2}{(\frac{\hat{F}^2_{n_2}}{1-\hat{F}^1_{n_1}+\hat{F}^2_{n_2}})^2 + (\frac{\hat{F}^1_{n_1}}{1-\hat{F}^1_{n_1}+\hat{F}^2_{n_2}})^2}} = min[1-\hat{F}^1_{n_1} + \hat{F}^2_{n_2}].\end{aligned}$$

Therefore, Class 1 and 2 are classified at the 11^{th} sample with a maximum TR value of 0.954 $(= (1.0+(1-0.09))/2)$. In Figure 2.1, $\hat{F}_i$ and $\hat{F}^k_{n_k}$ are represented as two step functions and a dotted vertical line on the $\hat{F_{11}}$ is shown as the maximum value for TR.

The result of the discriminant analysis using a multivariate empirical distribution function for the sample in Table 2.1 is found with only the 5^{th} sample of Class 2 $(3.14, 4.76)$, which is incorrectly classified as Class 1, the misclassification error rate is just 0.04762 $(= 1/(n_1+n_2))$. And the TPR (= 10/10) and FPR (= 1/11) can be obtained, so that TR appears as 0.954 $(= (10/10 + 10/11)/2)$, which is the same value obtained for Figure 2.1. From these results, it is found that discriminant analysis using empirical distribution is related to ROC curve.

3. Comparison with simulated and illustrative examples

3.1. Simulation results when $K = 2$

We assume three models with two classes as follows:

$$\textbf{(Model 1)} \quad \underline{x_1} \sim N\left(\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix}\right), \quad \underline{x_2} \sim N\left(\begin{pmatrix}4\\4\end{pmatrix}, \begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix}\right)$$

$$\textbf{(Model 2)} \quad \underline{x_1} \sim N\left(\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}1 & 0.6\\0.6 & 1.2\end{pmatrix}\right), \quad \underline{x_2} \sim N\left(\begin{pmatrix}4\\4\end{pmatrix}, \begin{pmatrix}1 & 0.7\\0.7 & 0.9\end{pmatrix}\right)$$

$$\textbf{(Model 3)} \quad \underline{x_1} \sim N\left(\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right), \quad \underline{x_2} \sim N\left(\begin{pmatrix}4\\5\end{pmatrix}, \begin{pmatrix}0.9 & \text{-0.7}\\\text{-0.7} & 0.8\end{pmatrix}\right)$$

The three models are used for a comparison of results of the proposed discriminant analysis and an existing discriminant analysis by performing a Monte Carlo simulation. The sizes of Class 1 and 2 are 20 $(= n_1)$ and 30 $(= n_2)$, respectively, and the number of iterations is set to 1,000. The existing discriminant analysis performs LDA or QDA whether a homogeneity of variance test is satisfied. And the comparison of the two methods is performed by APER (apparent error rate = the number of errors / the number of total observations).

Figure 3.1 shows the scatter plots, graphs for $(\hat{F}^2_{n_2}, \hat{F}^1_{n_1})$ and $\{(\hat{F}_i, \hat{F}^k_{n_k}), k = 1, 2\}$ for each model. It is found that the graphs of the empirical distribution function $(\hat{F}^2_{n_2}, \hat{F}^1_{n_1})$ in the middle of Figure 3.1 are expressed as ROC curves. Using the ROC curve in Figure 3.1, it is possible to measure the discrimination power using the AUC statistic. The AUCs for models 1 to 3 are 97.0%, 91.8% and 99.7%, respectively. Hence one can say that the discriminant powers are very well since these values of AUC are close to 1.

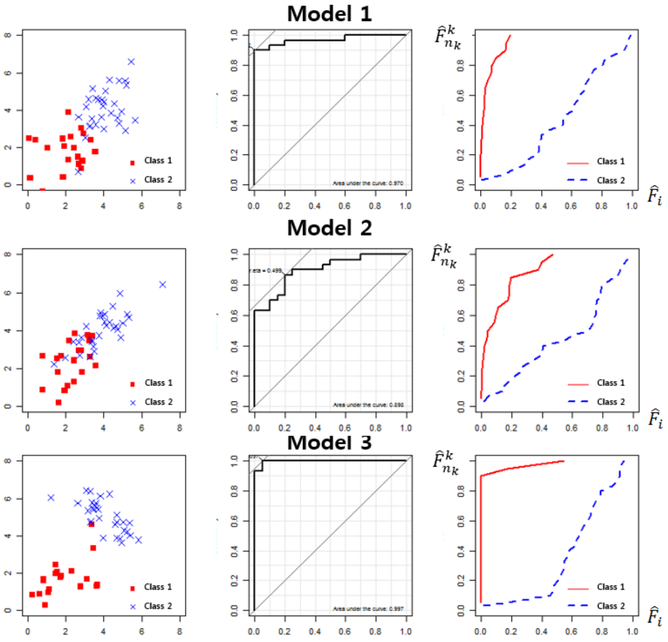

Figure 3.1 Scatter plot, ROC curve, and $(\hat{F}_i, \hat{F}^k_{n_k})$ plot

The right graphs depict the step functions through $\hat{F}_i$ and $\hat{F}^k_{n_k}$. They enable comprehensive and objective comparisons of the samples. The final results for the comparison are shown in Table 3.1 by using the mean and standard deviation according to the type of models and discriminant analysis methods. The sums of the misclassification are also shown through each discriminant analysis, and in the case of Model 1, the number of misclassifications of the discriminant analysis using multivariate empirical distributions is 0.716 $(= 4.040 - 3.324)$, which is less than the existing discriminant analysis. Similarly, in Model 2 and 3, the numbers of misclassifications of the proposed discriminant analysis in the study are 1.046 $(= 6.058 - 5.012)$ and 0.064 $(= 1.448 - 1.384)$, respectively. In addition, the standard deviation of the simulations is also generally small.

Exploring above results through TR, we found that the TRs in the proposed discriminant analysis are better than in existing discriminant analysis. The TR in the proposed discriminant analysis according to a Model 1 is obtained as 0.934 $(= (18.704/20 + 27.972/30)/2)$, whereas it appears as 0.917 $(= (17.920/20 + 28.130/30)/2)$ in the case of existing discriminant analysis. And the TRs of Model 2 and 3 have 0.899 and 0.971 for the proposed method, while those of Model 2 and 3 have 0.871 and 0.963, respectively. As a result of this simulation, the discriminant analysis using the multivariate empirical distribution could be seen to be well discriminated.

Table 3.1 Comparison of discriminant analyses when K=2

			Actual		Misclassification
			Class 1	Class 2	
Model 1	Proposed discriminant analysis	Class 1	18.704 (1.264)	2.028 (1.683)	3.324
		Class 2	1.296 (1.264)	27.972 (1.683)	
	Existing discriminant analysis	Class 1	17.920 (1.306)	1.856 (1.167)	4.040
		Class 2	2.080 (1.331)	28.130 (1.269)	
Model 2	Proposed discriminant analysis	Class 1	17.940 (1.523)	2.952 (1.864)	5.012
		Class 2	2.060 (1.523)	27.048 (1.864)	
	Existing discriminant analysis	Class 1	16.662 (1.641)	2.720 (1.451)	6.058
		Class 2	3.338 (1.641)	27.280 (1.451)	
Model 3	Proposed discriminant analysis	Class 1	19.320 (0.841)	0.704 (0.914)	1.384
		Class 2	0.680 (0.841)	29.269 (0.914)	
	Existing discriminant analysis	Class 1	18.552 (1.060)	0.000 (0.000)	1.448
		Class 2	1.448 (1.060)	30.000 (0.000)	

3.2. Simulation results when $K = 3$

We assume the following two models in the case of three classes. The data from 100 samples (30, 40, and 30 samples of each class) are used, and simulations are performed in the analogous methods, as shown in Section 3.1.

$$\textbf{(Model 1)} \quad \underline{x_1} \sim N\left(\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}1 & 0.2\\0.2 & 1\end{pmatrix}\right), \quad \underline{x_2} \sim N\left(\begin{pmatrix}4\\4\end{pmatrix}, \begin{pmatrix}1 & 0.3\\0.3 & 1\end{pmatrix}\right), \quad \underline{x_3} \sim N\left(\begin{pmatrix}6\\6\end{pmatrix}, \begin{pmatrix}1 & 0.4\\0.4 & 1\end{pmatrix}\right)$$

$$\textbf{(Model 2)} \quad \underline{x_1} \sim N\left(\begin{pmatrix}2\\3\end{pmatrix}, \begin{pmatrix}0.9 & -0.8\\-0.8 & 0.9\end{pmatrix}\right), \quad \underline{x_2} \sim N\left(\begin{pmatrix}3\\5\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right), \quad \underline{x_3} \sim N\left(\begin{pmatrix}5\\7\end{pmatrix}, \begin{pmatrix}0.6 & -0.4\\-0.4 & 0.8\end{pmatrix}\right)$$

To visually display the assumed models, the scatter plots of each model and the graphs of $\{(\hat{F}_i, \hat{F}^k_{n_k}),\ k = 1, 2, 3\}$ are shown respectively on the left and right sides of Figure 3.2. Based on the right graph in Figure 3.2, the proposed discriminant analysis makes it possible for an optimal classification criterion since three empirical distribution function graphs $(\hat{F}_i, \hat{F}^k_{n_k})$ for each class are represented by three non-decreasing functions.

The results for the three classes are summarized in Table 3.2, which shows the mean and standard deviation, and the sum of the misclassifications according to each model. In the case of Model 1, the difference of the misclassification errors in the proposed discriminant analysis and the existing discriminant analysis is 1.676 ($= 14.362 - 12.686$). In Model 2, we also found that the difference in the misclassification errors is 0.99 ($= 10.988 - 9.998$), which means that the proposed discriminant analysis has fewer misclassification errors than that of existing ones.

From a perspective of TR_3 ($= 2(\text{TR}_{12} + \text{TR}_{23} - 1/2)/3$), TR_3 in proposed discriminant analysis is 0.880 ($= 2((27.496/30 + 36.470/40)/2 + (36.140/40 + 27.184/30)/2 - 1/2)/3$), which is greater than TR_3 value in existing discriminant analysis of 0.859 ($= 2((26.388/30 + 36.542/40)/2 + (36.708/40 + 26.006/30)/2 - 1/2)/3$) in case of Model 1. Also, in Model 2

Table 3.2 Comparison of discriminant analyses when K=3

			Actual			Misclassification
			Class 1	Class 2	Class 3	
Model 1	Proposed discriminant analysis	Class 1	27.496 (1.803)	3.503 (2.296)	0.006 (0.077)	12.686
		Class 2	2.494 (1.803)	32.640 (3.223)	2.816 (1.912)	
		Class 3	0.010 (0.099)	3.830 (2.381)	27.178 (1.912)	
	Existing discriminant analysis	Class 1	26.388 (1.695)	3.458 (1.513)	0.006 (0.077)	14.362
		Class 2	3.612 (1.695)	33.250 (2.284)	3.994 (1.721)	
		Class 3	0.000 (0.000)	3.292 (1.612)	26.000 (1.721)	
Model 2	Proposed discriminant analysis	Class 1	29.730 (0.560)	3.736 (1.951)	0.000 (0.000)	9.998
		Class 2	0.270 (0.560)	31.968 (2.868)	1.696 (1.435)	
		Class 3	0.000 (0.000)	4.296 (2.301)	28.304 (1.435)	
	Existing discriminant analysis	Class 1	29.678 (0.648)	6.122 (1.813)	0.000 (0.000)	10.988
		Class 2	0.322 (0.648)	29.516 (2.526)	0.182 (0.488)	
		Class 3	0.000 (0.000)	4.362 (1.536)	29.818 (0.449)	

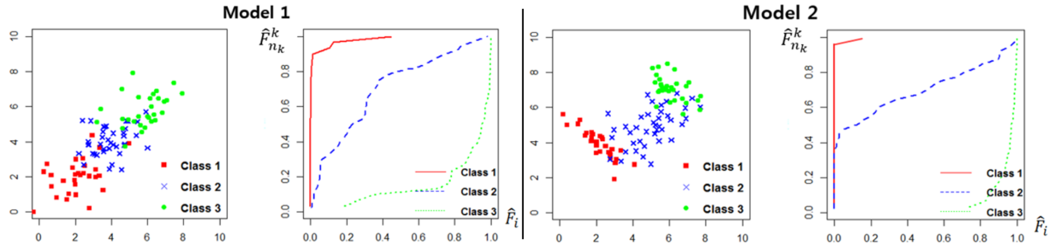

Figure 3.2 Scatter plot and $(\hat{F}_i, \hat{F}^k_{n_k})$ plot

the result of TR$_3$ (= 0.911) proposed in this study are slightly better than TR$_3$ (= 0.907) of existing method .

The multivariate empirical distribution functions $\hat{F}^1_{n_1}$, $\hat{F}^2_{n_2}$, and $\hat{F}^3_{n_3}$ could be used to express the ROC surfaces. The discrimination power in the three-dimensional ROC surfaces can be measured using the VUS, and it can be seen that the discrimination power is reasonably good because the VUSs are obtained as 84.2% and 65.2%, respectively.

With the results of this simulation, the proposed discrimination analysis method has better performances than the existing discriminant method. The standard deviations of the proposed discriminant analysis in Table 3.2 are generally large, so it might be said that the proposed method is more sensitive to the data even for the three classes.

3.3. Illustrative example when $K = 2$

We explore the results of two methods for some illustrative examples in the case of 2 and 3 classes. For annual financial data of bankrupt and solid companies (Johnson and Wichern, 2007, p. 657), the proposed discriminant analysis method is used, and this result is compared to that of the existing method. The data is made of four kinds of random financial variables with two classes (default and non-default).

Figure 3.3 represents the graph of the normal distribution function $\hat{F}_i$ and the multivariate empirical distribution function $\{\hat{F}^k_{n_k}, k = 1, 2\}$ for the financial illustrative example. As the value of $\hat{F}_i$ increases, the differences between $\hat{F}^1_{n_1}$ and $\hat{F}^2_{n_2}$ becomes larger and larger, so it seems to be clearly divided into two classes. Calculating a TR to discriminate between two classes, it is found that Class 1 and 2 are classified at $\hat{F_{11}}$ $(= 0.038)$ with the maximum TR of 0.8209, and a dotted vertical line on the $\hat{F_{11}}$ is represented in Figure 3.3. Also the same value of the TR can be obtained as 0.8209 $(= (16/21 + 22/25)/2)$ for Table 3.3.

Table 3.3 shows the results of two discriminant analysis methods: one is the method proposed in this study and the other is an existing method using the cross-validation after a test for homogeneity of the variance-covariance matrix for all independent variables. Based on Table 3.3, whereas the number of misclassifications is 8 and the well-classification rates of Class 0 and 1 are respectively 86% and 80% for the proposed method, the existing method has results with the number of misclassification of 8 and well-classification rate of Class 0 and 1 of 86% and 92%, respectively. Therefore, unlike what we expected, the existing method has better results than those of the proposed method in the example with financial data.

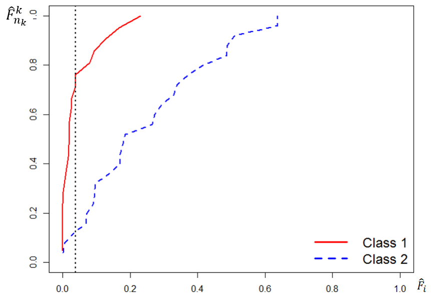

Figure 3.3 $(\hat{F}_i, \hat{F}^k_{n_k})$ plot for Bankruptcy data

Table 3.3 Comparison of discriminant analyses when K=2

		Actual	
		Class 1	Class 2
Proposed discriminant analysis	Class 1	16	3
	Class 2	5	22
Existing discriminant analysis	Class 1	18	2
	Class 2	3	23

3.4. Illustrative example when $K = 3$

The Concho Water Snake data (Johnson and Wichern, 2007, p. 667) consists of 'Tail Length' and 'Snout to Vent Length' variables with $K = 3$ (the gender variable is not used).

The graph of $\hat{F}_i$ and $\{\hat{F}^k_{n_k}, k = 1, 2, 3\}$ is shown in Figure 3.4 and the result of the comparative analysis for each method are shown in Table 3.4. The existing discriminant analysis to be compared is performed in the same method, as shown in Section 3.3. The horizontal differences among the three $\hat{F}^k_{n_k}$ in Figure 3.4 are wider than the differences in Figure 3.3, meaning that a high discriminant power can be expected. A classification accuracy TR_{12} and TR_{23}, which are the component of the linear relationship of TR_3, allow to discriminate between three classes. The maximum values of the TR_{12} and TR_{23} are 0.8744 at $\hat{F_{21}}$ $(= 0.049)$ and 0.923 at $\hat{F_{39}}$ $(= 0.292)$, respectively. In Figure 3.4, dotted vertical lines on the $\hat{F_{21}}$ and $\hat{F_{39}}$ are shown. In addition, TR_{12} and TR_{23} can have the same value as 0.874 $(= (16/17 + (1 - 5/26))/2)$ and 0.923 $(= (22/26 + (1 - 0/23))/2)$ for method propose in this study for Table 3.4.

Table 3.4 shows that the number of misclassifications is 10, and the classification accuracies for Class 1, 2, and 3 are 94%, 65% and 100%, respectively, for the proposed method. On the other hand, the existing method has results in which the number of misclassification is 11, and the classification accuracies of Class 1, 2, and 3 are 76%, 81%, and 91%, respectively. In this example, we can find that the proposed method can be seen to have better performances than the existing method.

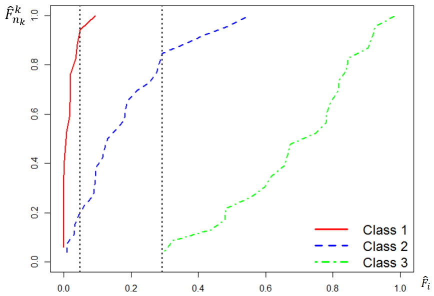

Figure 3.4 $(\hat{F}_i, \hat{F}^k_{n_k})$ plot for Snake data

Table 3.4 Comparison of discriminant analyses when K=3

		Actual		
		Class 1	Class 2	Class 3
Proposed discriminant analysis	Class 1	16	5	0
	Class 2	1	17	0
	Class 3	0	4	23
Existing discriminant analysis	Class 1	13	2	0
	Class 2	4	21	2
	Class 3	0	3	21

4. Conclusion

The discriminant analysis is a useful statistical technique that has been studied from various perspectives as a method of discriminating between classes using information of variables and determining which classes belong to a new object. In this study, we propose an alternative discriminant analysis method using the multivariate empirical distribution function of Hong *et al.* (2017), which could express multivariate data as a simple one-dimensional statistic.

This discriminant analysis method consists of the following process. For discriminant analysis data composed of K (≥ 2) classes, the univariate normal distribution functions are estimated using a weighted mean vector and the variance-covariance matrix, and the K empirical distribution functions corresponding to each class are obtained. We then draw a graph expressed by the normal distribution function and the empirical distribution functions, which explains some characteristics of each distribution. Also, the ROC curve when $K = 2$ and the ROC surface when $K = 3$ could be displayed with empirical distribution functions. Therefore, the discriminant analysis method can be regarded as determining the optimal threshold using various classification accuracy measures, including the TR.

Monte Carlo simulations in various mixed models and illustrative examples when $K = 2$ and 3 are explored to evaluate the two discriminant methods. We compare the results of the existing discriminant analysis method to that of method proposed in this study. In general, the proposed method might be found to have fewer misclassifications than the existing method for data fitted to a normal distribution.

Nonetheless, one cannot definitively conclude that the proposed method has better performances than the existing discrimination analysis due to the illustrative example in Section 3.3. In particular, when the empirical distribution functions for each class are almost the same, the proposed method has poor misclassification error rate. For example, if two random samples are generated from $(N((-1,1)^{\mathrm{T}}, I_2), N((1,-1)^{\mathrm{T}}, I_2))$, then the misclassification error rates for the proposed discriminant method are very close to 0.5, since the empirical distribution functions for each class are almost the similar. However, this research has significance in that we have implemented a method applying the empirical distribution function defined by Hong *et al.* (2017). This discriminant analysis using a multivariate empirical distribution function needs to consider many data sets followed by various distribution functions, and we will extend this work in future research.

References

Critchley, F. and Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika*, **78**, 677-690.

Fung, W. K. (1992). Some diagnostic measures in discriminant analysis. *Statistics & Probability Letters*, **13**, 279-285.

Fung, W. K. (1996). The influence of observations on misclassification probability in multiple discriminant analysis. *Communications in Statistics-Theory and Methods*, **25**, 1917-1930.

Hong, C. S. (2012). *SAS/SPSS and multivariate data analysis*, Free Academy, Paju.

Hong, C. S. and Joo, J. S. (2010). Optimal thresholds from non-normal mixture. *Korean Journal of Applied Statistics*, **23**, 943-953.

Hong, C. S. and Jung, E. S. (2013). Optimal thresholds criteria for ROC surfaces. *Journal of the Korean Data & Information Science Society*, **24**, 1489-1496.

Hong, C. S., Park, J. and Park, Y. H. (2017). Multivariate empirical distribution functions and descriptive methods. *The Korean Data & Information Science Society*, **28**, 87-98.

Jhun, M. S. and Choi, I. K. (2009). Adaptive nearest neighbors for classification. *Korean Journal of Applied Statistics*, **22**, 479-488.

Hong, C. S. and Jung, D. G. (2014). Standard criterion of hypervolume under the ROC manifold. *Journal of the Korean Data & Information Science Society*, **25**, 473-483.

Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis*, PrenticeHall International. INC., New Jersey.

Lambert, J. and Lipkovich, I. (2008). A macro for getting more out of your ROC curve. *SAS Global forum*, **231**.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, USA.

Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of "optimal"cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, **163**, 670-675.

Tasche, D. (2006). Validation of internal rating systems and PD estimates. *The Analytics of Risk Model Validation*, **28**, 169-196.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M. and Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic epidemiology*, **31**, 306-315.

Yoo, H. S. and Hong, C. S. (2011). Optimal criterion of classification accuracy measures for normal mixture. *Communications for Statistical Applications and Methods*, **18**, 343-355.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, **3**, 32-35.