STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **3**, March 2015, pp 66–78.
Published online in International Academic Press (www.IAPress.org)

The 95% confidence intervals of error rates and discriminant coefficients

Shuichi Shinmura [1], *

[1] *Faculty of Economics, Seikei University, Japan*

Received: 15 November 2014; Accepted: 13 February 2015
Editor: David G. Yu

Abstract Fisher proposed a linear discriminant function (Fisher's LDF). From 1971, we analysed electrocardiogram (ECG) data in order to develop the diagnostic logic between normal and abnormal symptoms by Fisher's LDF and a quadratic discriminant function (QDF). Our four years research was inferior to the decision tree logic developed by the medical doctor. After this experience, we discriminated many data and found four problems of the discriminant analysis. A revised Optimal LDF by Integer Programming (Revised IP-OLDF) based on the minimum number of misclassification (minimum NM) criterion resolves three problems entirely [13, 18]. In this research, we discuss fourth problem of the discriminant analysis. There are no standard errors (SEs) of the error rate and discriminant coefficient. We propose a k-fold cross-validation method. This method offers a model selection technique and a 95% confidence intervals (C.I.) of error rates and discriminant coefficients.

Keywords Linear Discriminant Function (LDF), Logistic Regression, Support Vector Machine (SVM), Number of Misclassifications (NM), Minimum NM (MNM), Revised IP-OLDF based on MNM criterion, Revised IPLP-OLDF, Revised LP-OLDF, Linear Separable Data and Model, K-fold Cross-Validation.

AMS 2010 subject classifications 62J07,90C08,68R05

DOI: 10.19139/soic.v3i1.109

1. Introduction

Fisher [3] described the linear discriminant function (Fisher's LDF) and founded the discriminant theory. Following Fisher's LDF, the quadratic discriminant function (QDF) and multi-class discrimination using Mahalanobis distance were proposed. Statistical software packages implement these discriminant functions based on the variance-covariance matrices. However, Fisher never formulated two equations of standard errors (SEs) of error rate and discriminant coefficient of Fisher's LDF. After these discriminant functions, the Framingham study developed a logistic regression. On the other hand, the discriminant functions using the mathematical programming (MP) were proposed. Stam [21] summarized these L_p-norm discriminant functions and lamented "Why statistical users rarely used MP-based functions?" honestly. One of the reasons is that there is no examination of real data. Vapnik [24] introduced three kinds of a support vector machine (SVM) such as a hard margin SVM (H-SVM), a soft margin SVM (S-SVM) and a kernel SVM. Statistical users accept SVM examined by the real data.

We used these discriminant functions in many applications, such as medical diagnosis, pattern recognition, genome discrimination, and rating property and bond. However, there are many problems in the discriminant analysis. It may be the simplicity of the discriminant rule: If $y_i * f(\mathbf{x}_i) > 0$, $\mathbf{x}_i$ is classified to class1/class2 correctly. If $y_i * f(\mathbf{x}_i) < 0$, $\mathbf{x}_i$ is misclassified. There are three serious problems hidden in this simplistic scenario [18].
Problem 1: We cannot discriminate between cases where $\mathbf{x}_i$ lies on the discriminant hyperplane $(f(\mathbf{x}_i) = 0)$

*Correspondence to: Faculty of Economics, Seikei University, 3-3-1 Kichijoji-kitamachi Musashino Tokyo, 180-8633 Japan. Email: sshinmura@gmail.com

ISSN 2310-5070 (online) ISSN 2311-004X (print)

correctly. We ignored this unresolved problem until now. If some cases lie on the discriminant hyperplane, the numbers of misclassification (NMs) or error rates may not be correct. Many researchers treat these cases belong to class1 without valid reason. Some statisticians believe it is decided by dice because statistics is the study of probability. Both treatments are not logical.

Problem 2: H-SVM tells us the discrimination of the linear separable data clearly. However, there is no research about it because of two reasons. H-SVM can recognize the linear separable data, but it can apply only to the linear separable data, and S-SVM sometimes cannot acknowledge the linear separable data. For this reason, there are few linear separable data. IP-OLDF based on MNM criterion finds the Swiss bank note data is linearly separable [4, 10, 11, 12]. In addition to this data, the pass/fail determination by exam scores [15] is good research data for the discrimination of linear separable data, because of two reasons. We can easily obtain it, and there is a trivial LDF, MNM of which is zero. Some statisticians believe that the purpose of discriminant analysis is to discriminate overlapping data correctly. However, the definition of overlap is uniquely defined by the condition of "$MNM > 0$" in the world of LDF. All LDFs except H- SVM and Revised IP-OLDF cannot define the overlap theoretically because they cannot discriminate the linear separable data correctly.

Problem 3: We assume that cases vary in statistics. If some variable is constant, we cannot compute the inverse matrices of the variance- covariance matrices. For this reason, most statistical packages exclude these variables from the discriminant analysis. However, JMP [8] adopts the generalized inverse matrix technique and can compute the inverse matrix. However, we found a defect, when the value of a variable belonging to one class is constant, and the value of another class varies. QDF and a regularized discriminant analysis (RDA) [5] misclassify all cases of class2 to class1. After the end of Dec. in 2012, JMP released modified RDA, but QDF enhanced by the generalized inverse matrices cannot discriminate this particular example correctly until 2014. If we add small random noise to the variable with a constant value, third problem is resolved. There is no need to exclude the constant variable in the discriminant analysis.

In this research, we discuss fourth problem that the discriminant analysis is not the inferential statistics such as a regression analysis. Fisher never formulated two equations of SE of the discriminant coefficient and error rate. Until now, no statisticians are successful in formulating these SEs by traditional approach based on the normal distribution. For this reason, we propose the "K-fold cross-validation for small sample" method by the computer-intensive approach. This method is a combination of re-sampling technique and k-fold cross- validations. By this new method, we can obtain the 95% confidence intervals (C.I.) of error rates and discriminant coefficients [16, 17]. Although Konishi and Honda [6] proposed the "bootstrap methods [1] for error rate estimation in discriminant analysis", a new approach is more helpful for researchers who wish to discriminate their small sample. In addition to these results, we discuss the new model selection technique. There are many model selection techniques such as stepwise methods and all possible combination of independent variables and statistics such as AIC, BIC and Cp statistics in the regression analysis. On the other hand, a leave-one-out (LOO) method is used as model selection technique in the discriminant analysis [7]. Our approach may be helpful for model selection procedure as follows. We select the best model that has the minimum mean of error rates (M2) in the validation samples and compare eight LDFs with this model. Some statisticians believe MNM is foolish criterion because it overfits the training sample and overestimates the validation sample. They say the generalization ability of Fisher's LDF is good because it assumes the Fisher's assumption without examination of real data. This claim has been proven to be a complete mistake by our method. We showed that the means of error rates (M1) in the training sample and M2 of Revised IP-OLDF were less than those of Fisher's LDF[20]. Revised IP-OLDF had resolved first and second problems [14]. In this research, it can resolve fourth problem of the discriminant analysis.

2. Method

In this research, we compare the mean of error rates "M1 & M2" and 95% C.I. of error rates and discriminant coefficients of eight LDFs. Eight LDFs are as follows: logistic regression, Fisher's LDF, Revised IP-OLDF, Revised IPLP-OLDF [19], Revised LP-OLDF, H-SVM and two S-SVMs those are SVM4 (penalty $c = 10^4$) and SVM1 (penalty $c = 1$).

2.1. Existing Discriminant Functions

Fisher defined Fisher's LDF to maximize the variance ratio (between/within classes). If we accept Fisher's assumption, the same LDF is obtained in equation (1) by the plug-in rule.

$$\text{LDF} : f(\mathbf{x}) =^{t} \{\mathbf{x} - (\mathbf{m}_1 + \mathbf{m}_2)/2\}\Sigma^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{1}$$

Where
$\mathbf{m}_1, \mathbf{m}_2$: the means of class1 and class2;
Σ: the pooled variance-covariance matrix.

This equation defines Fisher's LDF explicitly. For this reason, statistical software packages adopt this equation. Statistician who derived QDF known that most real data did not satisfy Fisher's assumption.When the variance-covariance matrices of two classes are not the same ($\Sigma_1 \neq \Sigma_2$), QDF can be used. In addition, Mahalanobis distance is used as multi- class discrimination and MT theory [22] in QC. We applied these functions in many areas but cannot obtained the functions if some independent variables remain constant. There are three examples. First, some variables, which belong to both classes, are the same constant. Second, some variables, which belong to both classes, are different but constant. Third, some variable, which belongs to only one class, is constant. Most statistical software packages exclude all variables in these three examples. On the other hand, JMP enhances QDF using the generalized inverse matrix technique. QDF of JMP can treat the first and second examples correctly, but cannot handle the third example properly until 2014.

Recently, the logistic regression in the following equation has been used for the heavy users such as the medical and economic areas:

$$\log(p/(1-p)) = f(\mathbf{x}). \tag{2}$$

Where
p: the probability belongs to class1.

It adapts the maximum likelihood estimation developed by Fisher and offers the 95% C.I. of logistic regression coefficients by numerical approach. If the data is linearly separable, Firth [2] tells us that the estimations are unstable. In this example, JMP output the warning. We can confirm that the data is linearly separable if JMP output the warning and "MNM=0" is confirmed by Revised IP-OLDF. Usually, NM of logistic regression is decided at the level "p=0.5". In this research, we choose the minimum NM (This is not equal to MNM) on the ROC curve by JMP.

Vapnik proposed three different SVM models. H-SVM in equation (3) indicates the discrimination of linear separable data clearly.

$$\text{MIN} = \|b\|^2/2; \quad y_i * ({}^t\mathbf{x}_i\mathbf{b} + b_0) \geq 1; \tag{3}$$

Where
$y_i = 1/-1$ for $\mathbf{x}_i \in$ class1/class2;
$\mathbf{x}_i$: p-independent variables (p-variables);
$\mathbf{b}$: p-discriminant coefficients;
b_0: the constant and free variable.

Some statisticians misunderstand that the discrimination of linear separable data is very easy. In statistics, there is no technical term for linear separable data. However, the condition of "MNM=0" is the same as being linearly separable. Note that "NM=0" does not imply the data is linearly separable. It is unfortunate that there has been no research on linear separable data because of two reasons. First, H-SVM can apply only to linear separable data, and S-SVM may not analyze it correctly. Second, there is few linear separable research data because H-SVM or Revised IP-OLDF must discriminate all possible models in order to find linear separable models. For this reason, there are few linear separable models until now. Real data are rarely linearly separable. For this reason, S-SVM has been defined in equation (4) with two objects.

$$\mathrm{MIN} = \|\mathbf{b}\|^2/2 + c * \Sigma e_i; \quad y_i * ({}^t\mathbf{x}_i\mathbf{b} + b_0) \geq 1 - e_i; \tag{4}$$

Where
c: penalty c;
e_i: non-negative decision variable.

These two objects are combined by defining some "**penalty c**." The Markowitz portfolio model [23] to minimize the risk and maximize the return is the same as S-SVM. However, the return is incorporated as a constraint, and the objective function minimizes the risk. The decision maker chooses a solution on the efficient frontier. On the contrary, S-SVM does not have a rule to determine "c" correctly; nevertheless, it can be solved by an optimization solver. In this research, two S-SVMs such as SVM4 ($c = 10^4$) and SVM1 ($c = 1$) are examined. We know the "M1 & M2" of SVM4 are almost better than SVM1.

2.2. Optimal Linear Discriminant Functions (OLDFs)

At first, IP-OLDF can be defined as in the following equation:

$$\mathrm{MIN} = \Sigma e_i; \quad y_i * ({}^t\mathbf{x}_i\mathbf{b} + 1) \geq -M * e_i; \tag{5}$$

Where
e_i : 0/1 integer decision variable;
M: 10,000 (Big M constant).

This notation is defined on p-dimensional coefficient space because the constant is fixed to 1. We can understand the relation between the LDFs and NMs in both data and coefficients space. And n linear equations made by n constraints ($H_i(\mathbf{b}) = y_i * ({}^t\mathbf{x}_i\mathbf{b} + 1) = 0$) divide the coefficients space into finite convex polyhedron clearly [11].

This basic understanding leads several new facts of the discriminant theory as follows:

1) LDF corresponding to the interior point of the convex polyhedron has unique NM, and there is no case on the discriminant hyperplane. LDF corresponding to the vertex or edge are not free from the first problem because there may be cases on the discriminant hyperplane. For this reason, error rates and NMs of all LDFs may not be correct because they cannot choose the interior points theoretically.

2) Only Revised IP-OLDF in equation (6) are free from the first problem. It can always find the interior point of the optimal convex polyhedron (OCP). Only LDF corresponding to the interior point of the OCP has exact MNM.

3) If MNM$_q$=0, all MNM including these q-independent variables (q-variables) are zero.

$$\mathrm{MIN} = \Sigma e_i; \quad y_i * ({}^t\mathbf{x}_i\mathbf{b} + b_0) \geq 1 - M * e_i; \tag{6}$$

Where
e_i: 0/1 integer decision variable;
b_0: free decision variable.

If e_i is a non-negative real variable, we utilize **Revised LP-OLDF**, which is an L$_1$-norm LDF. Its elapsed runtime (CPU time) is faster than that of Revised IP-OLDF. If we choose a large positive number as the penalty c of S-SVM, the result may be almost the same as that given by Revised LP-OLDF. It is the reason the first term of the objective value in the equation (4) becomes meaningless.

Revised IPLP-OLDF is a combined model of Revised LP-OLDF and Revised IP-OLDF. In the first step, Revised LP-OLDF is applied for all cases, and e_i is fixed to 0 for cases that are discriminated correctly by Revised LP-OLDF. In the second step, Revised IP-OLDF is used for cases, e_i of which are not zero in the first step. It is important that e_i changes from real decision variable to binary integer variable. For this reason, Revised IPLP-OLDF can be expected to obtain an estimate of MNM faster than Revised IP-OLDF for large samples [19].

2.3. *Original Data and Re-sampling Data*

Until now, we find three kinds of original data that are linearly separable. Those are the Swiss bank note data, the data of the pass/fail determination by exam scores and the Japanese 44 cars data [18]. IP- OLDF found that the Swiss bank note data is linearly separable by 2- variables (X4, X6). For this reason, we consider 16 linear separable models including 2-variables (X4, X6). The 47 models that remain are not linearly separable. Our claim is that we focus on these 16 linear separable models because we need not select the best model among 47 models. In addition to this merit, to compare the results of eight LDFs by these 16 models is more straightforward and clear than to compare eight LDFs by 47 models.

After 2010, we teach the statistical preliminary course for approximately 130 freshmen, attended. Midterm and final exams consisted of 100 questions with ten choices. We consider two discriminations using 100 item scores and four testlet scores as independent variables. If the pass mark is 50 points, we can easily obtain **a trivial LDF** ($f = T1 + T2 + T3 + T4 - 50$). If $f \geq 0$ or $f < 0$, the students pass or fail the exam, respectively. In this example, students on the discriminant hyperplane pass the exam because their score is exactly 50. This example indicates that there is **no first problem because independent variables decide the discriminant rule.**

Table 1 shows the discrimination of four testlet scores for a 10% level of the midterm exams. 'p' denotes the number of independent variables selected by the forward stepwise method. In 2012, the 2- variables model (T4, T2) was linearly separable. There are 15 discriminant models by all combination of variables, only four LDFs of which are linearly separable. In this research, we focus on four linear separable models such as (T1, T2, T3, T4), (T1, T2, T4), (T2, T3, T4), (T2, T4). In addition, we compare of eight LDFs by these four models. Those comparisons are very clear because four models are linearly separable. On the other hand, when we compare eight LDFs by eleven models, we may not judge the results positively. Some statisticians claim that the purpose of discriminant analysis is to discriminate overlapping data, not linear separable data. Until now, we cannot define the overlapping status in the world of LDF. Users of Fisher's LDF and QDF misunderstand all 15 models are overlapping because NMs are not zero. Only MNMs show four models including (T2, T4) are not overlapping, and 11 models are overlapping. It is difficult for us to define the overlapping status of the world of non- linear discrimination.

Table 1. NMs of four discriminant functions by forward stepwise in 2012 midterm exams at the 10% level.

p	Var.	Model	Revised IP-OLDF	Logistic	LDF*1	QDF*1
1	**T4**	T4	4	8	14(6)	12(6)
2	**T2**	T2,T4	**0**	**0**	**11(1)**	**9(1)**
3	T1	T1,T2,T4	0	0	12(1)	8(1)
4	T3	T1-T4	0	0	12(1)	1(0)

*1: This table was calculated in 2012. NMs of LDF and QDF remarkably decreased at the end of Dec. in 2014. The NMs in parenthesis and other tables are computed again in 2014. The algorithm of JMP may improved for the linear separable data.

2.4. *K-fold Cross-Validation*

Fisher's LDF is re-defined by the plug-in rule under Fisher's assumption. Also, if we consider $y_i = 1/-1$ as the object variable and analyze the data by the regression analysis, obtained regression coefficients are proportional to equation (1) by second plug-in rule. For this reason, we can use model selection techniques of regression analysis. Some statisticians misunderstand the discriminant analysis is the same as the regression analysis by the reason derived from the normal distribution. We cannot obtain two SEs of error rates and discriminant coefficients. Although there are many model selection techniques such as stepwise approaches and statistics such as Cp, AIC and BIC in the regression analysis, we can use only the LOO method in the discriminant analysis. In this paper, we propose the "k-fold cross-validation for small sample" method as follows. In addition, we consider the best model that has minimum M2 in the validation sample. In this research, we fix $k = 100$ in order to obtain the 95% C.I. of error rates and discriminant coefficients.

1) We copy the original data 100 times. In addition, we add one variable, the value of which is generated by the uniform random number. Now, we can obtain a large re-sampling sample.

2) We consider this data as pseudo-population. We are sorting the data in ascending order in each class. In addition, we add the sub- sample numbers from 1 to 100. This operation is the same effect by re-sampling from pseudo-population or original sample directly.

3) The sub-samples from 1 to 100 are the 100 training samples. In addition, pseudo-population is the validation sample. This method has two merits. The relation of the validation sample and training samples equals to the relation of population and samples. The validation sample has the same character as the original data that prevents the mistakes. For this reason, we can control the quality of data sets.

This method gives us much information, compared with the LOO methods. We can evaluate eight LDFs by "M1 & M2" and the 95% C.I. of error rates and discriminant coefficients. Until now, users of discriminant analysis cannot use this useful information.

3. Pass/Fail Determination by Exam Scores

In this research, we discuss the discrimination of linear separable data. We examined the pass/fail determination by midterm exam at the 10% level in 2012. Pass mark is 36 points. For this reason, a trivial LDF such as f=T1+T2+T3+T4-36 is a linear separable model. If $f \geq 0$ or $f < 0$, student pass or fail the exam, respectively. There are 15 discriminant models made by 4-variables such as T1, T2, T3, and T4. Because MNM of (T2, T4) model is zero, four MNMs including (T2, T4) are zero. In addition, other eleven MNMs are not zero. In section 3.1, we focus on four linear separable models. We compare the "M1s & M2s" of eight LDFs including H-SVM. In section 3.2, we discuss the ranges for error rates of four linear separable models and the ranges in four out of eleven models. In section 3.3, we discuss the 95% C.I. of discriminant coefficients of four linear separable models.

3.1. The Mean of Error Rates

Table 2(1). Four M1s and M2s of eight LDFs by exam scores.

RIP		M1	M2	Diff.	Model	
14m8s	1	0	0.879	0.879	1,2,3,4	
	2	0	1.121	1.121	1,2,4	
	3	0	0.887	0.887	2,3,4	
	4	0	0.863	0.863	2,4	
	5	0.734	2.444	1.710	1,3,4	
	6	2.242	4.613	2.371	1,2,3	
	7	1.782	3.444	1.661	1,4	
	8	2.282	3.250	0.968	3,4	
	9	4.831	6.556	1.726	1,2	
	10	4.403	6.411	2.008	2,3	
	11	4.944	7.266	2.323	1,3	
H-SVM		M1	M2	Diff.	M1Diff.	M2Diff.
5m23s	1	0	0.806	0.806	0	-0.073
	2	0	1.153	1.153	0	0.032
	3	0	0.895	0.895	0	0.008
	4	0	0.895	0.895	0	0.032
SVM4		M1	M2	Diff.	M1Diff.	M2Diff.
9m23s	1	0	0.806	0.806	0	-0.073
	2	0	1.153	1.153	0	0.032
	3	0	0.887	0.887	0	0.000
	4	0	0.903	0.903	0	0.040
	5	1.363	2.839	1.476	0.629	0.395
	6	3.992	5.782	1.790	1.750	1.169
	7	2.919	4.242	1.323	1.137	0.798
	8	3.113	3.790	0.677	0.831	0.540
	9	6.556	7.565	1.008	1.726	1.008
	10	6.976	7.992	1.016	2.573	1.581
	11	6.750	8.145	1.395	1.806	0.879

Tables 2(1),2(2),2(3) show the results by 100-fold cross-validations. "M1 and M2" columns are the means of error rates in the training and validation samples. The first column shows eight LDFs. Those are RIP (Revised IP-OLDF), H-SVM, SVM4, SVM1, IPLP (Revised IPLP- OLDF), LP (Revised LP-OLDF), Logistic (logistic regression) and Fisher's LDF. Only Fisher's LDF depends on the normal distribution. Four rows from 1 to 4 correspond to four linear separable models. Four models are (T1, T2, T3, T4), (T1, T2, T4), (T2, T3, T4), and (T2, T4) in this order. 'Model' column shows the suffix of a variable. We omit four 1-variable models from the table. For this reason, seven models from 5 to 11 are showed in the table.

Table 2(2). Four M1s and M2s of eight LDFs by exam scores.

SVM1		M1	M2	Diff.	M1Diff.	M2Diff.
8m53s	1	0	0.806	0.806	0	-0.073
	2	0.718	1.589	0.871	0.718	0.468
	3	0.129	0.976	0.847	0.129	0.089
	4	0.758	1.710	0.952	0.758	0.847
	5	1.653	3.153	1.500	0.919	0.710
	6	4.355	6.000	1.645	2.113	1.387
	7	2.984	4.274	1.290	1.202	0.831
	8	3.105	3.774	0.669	0.823	0.524
	9	6.540	7.524	0.984	1.710	0.968
	10	7.242	8.185	0.944	2.839	1.774
	11	6.855	8.153	1.298	1.911	0.887
IPLP		M1	M2	Diff.	M1Diff.	M2Diff.
13m16s	1	0	1.000	1.000	0	0.121
	2	0	1.081	1.081	0	-0.040
	3	0	0.927	0.927	0	0.040
	4	0	0.847	0.847	0	-0.016
	5	0.734	2.435	1.702	0.000	-0.008
	6	2.315	4.718	2.403	0.073	0.105
	7	1.782	3.492	1.710	0.000	0.048
	8	2.347	3.395	1.048	0.065	0.145
	9	4.839	6.548	1.710	0.008	-0.008
	10	4.476	6.274	1.798	0.073	-0.137
	11	5.113	7.137	2.024	0.169	-0.129
LP		M1	M2	Diff.	M1Diff.	M2Diff.
4m23s	1	0	0.960	0.960	0	0.081
	2	0	1.105	1.105	0	-0.016
	3	0	0.952	0.952	0	0.065
	4	0	0.798	0.798	0	-0.065
	5	1.363	2.847	1.484	0.629	0.403
	6	3.911	5.645	1.734	1.669	1.032
	7	2.887	4.185	1.298	1.105	0.742
	8	3.073	3.710	0.637	0.790	0.460
	9	6.468	7.395	0.927	1.637	0.839
	10	6.460	7.468	1.008	2.056	1.056
	11	6.508	7.927	1.419	1.565	0.661

Revised IP-OLDF analyzes 1,500 training samples as follows. First, we discriminate the 100 training samples ($n = 124$ cases) by 100 discriminant models of (T1, T2, T3, T4) and obtain 100 NMs. From these NMs, 100 error rates are calculated by dividing $n = 124$ and the mean of error rate 'M1' is calculated by 100 error rates. In this manner, other fourteen M1s are computed. After optimization, these 1,500 Revised IP-OLDFs are applied to the validation samples ($N = 12,400$ cases). In addition, we compute fifteen M2s and the percentiles such as 0%, 2.5%, 50%, 97.5% and 100% of error rates and discriminant coefficients. In this way, LINGO [9] analyzes six MP-based LDFs. JMP [8] script analyzes logistic regression and Fisher's LDF. However, it outputs only 1500 NMs. Excel computes the error rates and percentiles.

RIP, H-SVM, SVM4, IPLP, LP, and Logistic can recognize four linear separable models correctly because four M1s are zero. SVM1 and Fisher's LDF cannot recognize three and four linear separable models.

We define a model with a minimum M2 as the best model among fifteen models. RIP, IPLP, LP and Fisher's LDF choose the fourth model (T2, T4) as the best model. H-SVM, SVM4, SVM1, and Logistic select the first model (T1, T2, T3, T4). Minimum values of M2 are 0.863, 0.806, 0.806, 0.806, 0.847, 0.798, 0.774 and 9.913%, respectively. Logistic has the smallest value 0.774%. In this data, we cannot choose the same best models. Nevertheless, the best model of Revised IP-OLDF has the minimum value of M2 among eight LDFs by other data.

"Diff." column is the difference between (M2 - M1). Some statisticians claim a model with a small value of "Diff." has good generalization ability. Although the fourth model of Fisher's LDF has the minimum value 0.372, the values of M1 and M2 are 9.54 and 9.913 those are very wrong. For this reason, this statistics is not useful to check the generalization ability of discriminant models. "M1Diff.& M2Diff." columns are the differences of (M1 & M2 of seven LDF - those of RIP). There are ten minus values of "M2Diff." for four linear separable models. This fact tells us ten 'M2' values of Revised IP-OLDF are worse than six LDFs except for Fisher's LDF within 0.1%. On the other hand, four models of Fisher's LDF are 9.05% worse than Revised IP-OLDF. Six ranges of "M2Diff." are $[0.395, 1.581]$, $[0.524, 1.774]$, $[\underline{-0.137}, 0.145]$, $[0.403, 1.032]$, $[0.559, 1.625]$ and $[6.238, 9.595]$, respectively. Only minimum value of Revised IPLP-OLDF is minus. This result may suggest Revised IP-OLDF and Revised IPLP-OLDF are superior to other six LDFs for seven models.

Table 2(3).Four M1s and M2s of eight LDFs by exam scores.

Logistic		M1	M2	Diff.	M1Diff.	M2Diff.
12m	1	0	0.774	0.774	0	-0.105
	2	0	1.089	1.089	0	-0.032
	3	0	0.847	0.847	0	-0.040
	4	0	0.911	0.911	0	0.048
	5	1.589	2.831	1.242	0.855	0.387
	6	4.121	5.459	1.338	1.879	0.846
	7	3.250	4.258	1.008	1.468	0.814
	8	3.677	4.028	0.350	1.395	0.778
	9	6.935	7.777	0.842	2.105	1.221
	10	7.653	8.036	0.383	3.250	1.625
	11	7.048	7.825	0.776	2.105	0.559
Fisher's LDF		M1	M2	Diff.	M1Diff.	M2Diff.
15m	1	9.637	10.541	0.904	9.637	9.662
	2	9.887	10.546	0.659	9.887	9.425
	3	9.476	10.091	0.615	9.476	9.204
	4	9.540	9.913	0.372	9.540	9.050
	5	11.444	12.039	0.595	10.710	9.595
	6	12.363	12.683	0.320	10.121	8.070
	7	11.766	12.165	0.398	9.984	8.721
	8	10.806	11.029	0.223	8.524	7.779
	9	13.185	13.465	0.279	8.355	6.908
	10	12.492	12.649	0.157	8.089	6.238
	11	16.282	16.482	0.200	11.339	9.216

CPU times of eight LDFs are in the first column. Until now, MP- solver was slower than statistical software. This slowness increased the research time, by that reducing the feasibility of research using MP-solver. However, the MP-solver is a more powerful tool now. For this reason, we can study discriminant analysis by the computer-intensive approach instead of the traditional approach restricted by the theoretical distributions. In 2009, Revised IPLP-OLDF was faster than Revised IP-OLDF when tested on LINGO Ver.10 [9]. Reversal of CPU time occurred for Revised IP-OLDF and Revised IPLP-OLDF, when tested on LINGO Ver.14. We expect that Revised IPLP-OLDF may be faster than Revised IP-OLDF for a larger data set. On the other hand, CPU times of logistic regression and Fisher's LDF are about 12 minutes and 15 minutes as recorded by means of a wrist watch, respectively. Although both statistical LDFs output a large amount of information, they were nonetheless faster than Revised IP- OLDF until 2013.

3.2. The Ranges of Error Rates

Table 3 shows the ranges of error rates instead of the 95% C.I. of error rates. Upper and lower rows in each LDF display the ranges of the training and validation samples, respectively. RIP, H-SVM, SVM4, IPLP, and Logistic have the same results. These five LDFs and LP can recognize four linear separable models correctly. SVM1 cannot recognize three linear separable model and Fisher's LDF cannot recognize four linear separable models. The maximum value 3.23% of the fourth model of LP is smaller than other LDFs. We can conclude Fisher's LDF are worse in the training and validation samples. Other seven LDFs almost have the same results.

Table 3. The ranges of error rates.

	T1, T2, T3, T4		T1, T2, T4		T2, T3, T4		T2, T4	
	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
RIP*2	0	0	0	0	0	0	0	0
	0	4.84	0	3.23	0	4.84	0	4.84
SVM1	0	0	0	2.42	0	0.81	0	3.23
	0	4.84	0.81	3.23	0	4.84	0	4.84
LP	0	0	0	0	0	0	0	0
	0	4.84	0	3.23	0	4.84	0	3.23
LDF	4.03	16.13	4.03	16.94	4.03	16.13	4.03	17.74
	6.40	12.82	8.01	12.82	8.00	13.60	8.00	13.60

Upper: training sample, Lower: validation sample.

*2: Revised IP-OLDF, H-SVM, SVM4, Revised IPLP-OLDF, and Logistic regression are the same.

Table 4 shows the ranges of four models among eleven models. Four models are (T1, T3, T4), (T1, T2, T3), (T1, T4), and (T3, T4). We can understand all minimum and maximum values of RIP are less than equal other six LDFs. From **Table 2** and **Table 3**, we do not permit the superiority of RIP. However, RIP may be superior to other LDFs for the overlapping data. We must examine by other data (**Future Work 1**).

Table 4. The ranges of error rates about four non-linear separable models.

	T1, T3, T4		T1, T2, T3		T1, T4		T3, T4	
	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
RIP	0	2.419	0	5.645	0	4.032	0	5.645
	1.613	4.839	3.226	8.065	2.419	6.452	2.419	7.258
SVM4	0	5.645	0	8.871	0	7.258	0	8.065
	1.613	5.645	3.226	8.065	2.419	6.452	2.419	8.065
SVM1	0	5.645	1.613	10.484	0	7.258	0	8.065
	2.419	5.645	4.032	8.065	3.226	6.452	2.419	8.065
IPLP	0	2.419	0	6.452	0	4.032	0	8.065
	1.613	4.839	3.226	8.065	2.419	6.452	2.419	8.065
LP	0	5.645	0	8.871	0	7.258	0	8.065
	1.613	5.645	3.226	8.065	2.419	6.452	2.419	8.065
Logistic	0	5.645	0	11.290	0	8.065	0	8.871
	1.613	4.839	3.226	8.032	2.379	6.452	2.419	7.194
Fisher's LDF	4.839	17.742	5.645	20.161	4.839	17.742	4.839	17.742
	7.960	16.839	9.621	15.185	9.548	16.839	7.960	15.984

4. The 95% C.I. of Discriminant Coefficients

4.1. Six MP-Based LDFs

We examine the 95% C.I. of the coefficient about first and fourth models. We discuss which model is the best model. **Table 5(1),5(2)** show the median and 95% C.I. of six MP-based LDFs. Fisher's LDF and logistic regression by JMP script do not output 100 discriminant coefficients. If the 95% C.I. include zero, we can judge the pseudo population coefficient is zero. If the value at 2.5% is greater than 0 or the value at 97.5% is less than 0, we estimate the pseudo population coefficient a positive or negative value. Following this judgment, all coefficients of T2 and T4 are positive values and all constants are negative values. This result implies noteworthy results because most results imply us the coefficients of two variables (T2, T4) are significant at the 5% level. For this reason, we can conclude the two variables model (T2, T4) is the best model selected by RIP, IPLP, LP and Fisher's LDF. We must investigate this fact with other data (**Future Work 2**).

Table 5(1). The 95% C.I. of six LDFs.

RIP	T1	T2	T3	T4	c
97.5%	0.384	1.413	1.396	0.438	-6.253
Median	0.149	0.654	0.405	0.314	-13.142
2.5%	-0.108	0.242	-0.018	0.171	-18.378
97.5%	0.838	14.364		4	-7.670
Median	-0.036	3.335		0.842	-32.823
2.5%	-0.909	0.290		0.280	-147.360
97.5%		2.800	1.786	0.800	-5.975
Median		2.260	0.400	0.540	-22.888
2.5%		0.328	0	0.210	-32.200
97.5%		13		4	-8.467
Median		4		2	-55
2.5%		0.608		0.333	-146
H-SVM	T1	T2	T3	T4	c
97.5%	0.483	1.170	0.761	0.655	-6.581
Median	0.256	0.623	0.443	0.369	-13.994
2.5%	-0.031	0.218	-0.025	0.193	-22.351
97.5%	5.008	12.400		4	-7.829
Median	0.380	3.335		0.929	-34.687
2.5%	-0.226	0.238		0.281	-145.400
97.5%		2.8	1.108	0.858	-5.848
Median		2.103	0.400	0.521	-22.526
2.5%		0.265	0.046	0.241	-32.543
97.5%		13		4	-8.43
Median		4		2	-55
2.5%		0.548		0.339	-146
SVM4	T1	T2	T3	T4	c
97.5%	0.483	1.170	0.761	0.655	-6.582
Median	0.256	0.623	0.443	0.369	-13.989
2.5%	-0.028	0.219	-0.025	0.193	-22.351
97.5%	5.008	12.400		4	-7.825
Median	0.380	3.335		0.929	-34.687
2.5%	-0.226	0.238		0.281	-145.400
97.5%		2.800	1.108	0.858	-5.852
Median		2.103	0.400	0.521	-22.526
2.5%		0.265	0.047	0.241	-32.544
97.5%		13		4	-8.340
Median		4		2	-55
2.5%		0.548		0.339	-146
SVM1	T1	T2	T3	T4	c
97.5%	0.483	1.170	0.761	0.655	-6.582
Median	0.256	0.623	0.443	0.370	-13.994
2.5%	-0.031	0.218	-0.025	0.193	-22.351
97.5%	0.764	1.722		0.784	-7.347
Median	0.259	0.870		0.399	-14.037
2.5%	-0.107	0.215		0.239	-29.647
97.5%		2.486	1.000	0.664	-5.850
Median		1.510	0.371	0.447	-18.159
2.5%		0.265	0.046	0.234	-27.201
97.5%		2.642		0.793	-7.386
Median		1.534		0.498	-16.478
2.5%		0.448		0.282	-29.275

Table 5(2). The 95% C.I. of six LDFs.

IPLP	T1	T2	T3	T4	c
97.5%	0.579	1.366	1.462	0.655	-6.029
Median	0.171	0.651	0.418	0.341	-13.252
2.5%	-0.123	0.058	-0.048	0.179	-22.689
97.5%	3.25	14		4	-7.248
Median	0.125	3.335		0.849	-32.880
2.5%		-0.667	0.256	0.281	-147.000
97.5%		2.800	1.786	0.800	-6.529
Median		2.260	0.400	0.517	-22.526
2.5%		0.343	0.040	0.217	-32.200
97.5%		13		4	-8.783
Median		4		2	-55
2.5%		0.503		0.392	-146
LP	T1	T2	T3	T4	c
97.5%	0.412	1.677	1.428	0.451	-6.047
Median	0.149	0.705	0.405	0.330	-13.060
2.5%	-0.158	0.177	-0.036	0.172	-18.773
97.5%	3.25	14		4	-7.62
Median	0.106	3.335		0.842	-32.823
2.5%	-0.667	0.380		0.278	-147.000
97.5%		2.885	1.786	0.800	-5.975
Median		2.260	0.400	0.530	-22.526
2.5%		0.328	-0.136	0.210	-32.200
97.5%		13		4	-8.281
Median		4		2	-55
2.5%		0.548		0.339	-146

4.2. *Fisher's LDF and Logistic Regression*

Fisher never formulated the SE of discriminant coefficient. If we use class identifier y_i as the object variable in the regression analysis, we can obtain the SE of the regression coefficient. **Table 6** shows the 95% C.I. of the regression coefficient calculated by the SE from the validation sample. First two rows show the "2.5% and 97.5%" of regression coefficients. In this research, the validation sample is considered as the pseudo-population. Although the C.I. of pseudo- population may be no meaning, we count the number of 100 regression coefficients by the training samples out of the 95% C.I. The outliers show these figures. We expect these figures are about 5 cases. In eight intervals, the number of outliers is less than eight. In five intervals, the number of outliers is less than five. For this reason, we can accept this 95% C.I., and judge two Fisher's LDFs are significant. We cannot decide the best model by the 95% C. I. of the Fisher's linear discriminant coefficient.

Table 6. The 95% C.I. of the regression analysis.

LDF	T1	T2	T3	T4	c
2.5%	0.004	0.071	0.004	0.022	-0.722
97.5%	0.008	0.008	0.01	0.024	-0.645
Outlier	5	6	2	7	3
2.5%		0.075		0.024	-0.68
97.5%		0.083		0.026	-0.61
Outlier		8		4	3

JMP calculates SEs of logistic regression coefficients defined by Hessian matrix that is obtained by the maximum likelihood estimation. The numbers in parentheses in equation (7) are SEs. The values of SEs are enormous, and all 95% C.I. become considerably broad, including zero. For example, the 95% C.I. of T2 of model (T2, T4) is $[-175 - 1.96 * 6534, -175 + 1.96 * 6534] = [-12982, 12632]$. For this reason, JMP outputs a warning message. However, if we find "NM=0" on the ROC by JMP and "MNM=0" by Revised IP-OLDF, we judge it is the linear separable model. In general, an exact logistic regression supported by SAS is recommended. It avoids complex work. We have no idea to decide the best model for logistic regression by the 95% C.I. of the logistic regression coefficient.

$$\begin{aligned} Logistic1234 &= -2.6(725) * T1 - 21(2489) * T2 - 6.6(1248) * T3 - 6.97(724) * T4 + 296(26825). \\ Logistic24 &= -175(6534) * T2 - 54(2012) * T4 + 1968(73372). \end{aligned} \tag{7}$$

5. Conclusion

In this research, we have discussed the fourth problem of discriminant analysis. Fisher never formulated two SEs of error rate and discriminant coefficient. However, some statisticians believe the discriminant analysis is inferential statistics similar to regression analysis because Fisher's LDF assumes the Fisher's assumption based on the normal distribution. This claim is not logical. Statistical software reflects the common knowledge obtained by statistical research. Statistical users can infer that discriminant analysis is not the same as traditional inferential statistics because a commercial software never display the SE of error rate and discriminant coefficients. There is a research about the error rate by the bootstrap method [6] by the computer-intensive approach. In this study, we propose the "k-fold cross-validation for small sample" method. This method can resolve the fourth problem and evaluates eight LDFs by the "M1 & M2". The procedure of this method is very straightforward. In addition, there are several merits as follows:

1) It is easy to generate the re-sampling sample by statistical software.

2) It reflects the relation of pseudo-population and samples. The training samples should be a sub-set of pseudo-population. In addition, we can control the quality of the training and validation samples very easy.

3) This method shows good results as explained in this paper. At least this method is better than the LOO method and displays the following outcomes:

1) Fisher's LDF is worst. SVM1 is second worst. H-SVM, SVM4, Revised IPLP-OLDF, Revised LP-OLDF, and logistic regression show good results for the linear separable models. On the contrary, Revised IP-OLDF is the best among eight LDFs for other eleven models that are not linear separable. This outcome may imply Revised IP-OLDF is superior to other LDFs for non-linear separable models.

2) We cannot decide the best model with minimum M2 in this data. However, the 95% C.I. of coefficients by six MP-based LDFs recommend 2-variables model of (T2, T4). We must examine this results with other data (**Future Work 3**).

3) In this research, we conclude that Fisher's LDF is the worst among the eight LDFs. We obtain the same conclusion from the CPD data and the student data [20]. We conclude that Fisher's LDF is fragile for discrimination such as the pass/fail determination and medical diagnosis of healthy and ill classes. These data have the same characteristic having many cases near the linear hyperplane. This feature does not satisfy the Fisher's assumption. In the nea future, we shall examine other data that does not satisfy the Fisher's assumption (**Future Work 4**). We plan to work on four future works using different data, further to understand the results obtained in this study.

Acknowledgments

This research started in 1997 and finished in 2014. It was achieved by "What's Best! and LINGO [9] of LINDO Systems Inc". and "SAS and JMP [8] of SAS Institute Inc". Fourth problem is our last goal after 2012.

REFERENCES

1. Efron, B., (1979). Bootstrap Methods -Another Look at the Jackknife-. The Annals of Statics, 7/1, 1-26.

2. Firth, D., (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80, 27-39.
3. Fisher, R. A., (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, 179-188.
4. Flury, B., Rieduyl, H., (1988). Multivariate Statistics: A Practical Approach. Cambridge University Press.
5. Friedman, J. H., (1989). Regularized Discriminant Analysis. Journal of the American Statistical Association, 84/405, 165-175.
6. Konishi, S., Honda, M., (1992). Bootstrap Methods for Error Rate Estimation in Discriminant Analysis. Japanese Society of Applied Statistics, 21/2, 67-100.
7. Lachenbruch, P. A., Mickey, M. R., (1968). Estimation of error rates in discriminant analysis. Technometrics 10, 1-11.
8. Sall, J. P., Creighton, L., Lehman, A., (2004). JMP Start Statistics, Third Edition. SAS Institute Inc.
9. Schrage, L., (2006). Optimization Modeling with LINGO. LINDO Systems Inc.
10. Shinmura, S., (1998). Optimal Linear Discrimrnant Functions using Mathematical Programming. Journal of the Japanese Society of Computer Statistics, 11 / 2, 89-101.
11. Shinmura, S., (2000). A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics, 5, 133-142.
12. Shinmura, S., (2004). New Algorithm of Discriminant Analysis using Integer Programming. IPSI 2004 Pescara VIP Conference, CD-ROM, 1-18.
13. Shinmura, S., (2007). Overviews of Discriminant Function by Mathematical Programming. Journal of the Japanese Society of Computer Statistics, 20/1-2, 59-94.
14. Shinmura, S., (2010). The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing (in Japanese).
15. Shinmura, S., (2011). Problems of Discriminant Analysis by Mark Sense Test Data. Japanese Society of Applied Statistics, 40/3, 157-172.
16. Shinmura, S., (2011). Beyond Fisher's Linear Discriminant Analysis - New World of Discriminant Analysis. ISI CD-ROM, 1-6.
17. Shinmura, S., (2013). Evaluation of Optimal Linear Discriminant Function by 100-fold Cross-validation. 2013 ISI CD-ROM, 1-6.
18. Shinmura, S., (2014). End of Discriminant Functions based on Variance Covariance Matrices. ICORES, 5-14, 2014.
19. Shinmura, S., (2014). Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP. Statistics, Optimization and Information Computing, vol. 2, 114-129.
20. Shinmura, S., (2014). Comparison of Linear Discriminant Function by K-fold Cross-validation. Data Analytic 2014, 1-6.
21. Stam, A., (1997). Nontraditional approaches to statistical classification: Some perspectives on Lp-norm methods. Annals of Operations Research, 74, 1-36.
22. Taguchi, G., Jugulu, R. (2002). The Mahalanobis-Taguchi Strategy-A Pattern Technology System. John Wiley & Sons.
23. Markowitz, H. M., (1959). Portfolio Selection, Efficient Diversification of Investment. John Wiley & Sons, Inc.
24. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.