The 100-fold Cross Validation for Small Sample Method

Shuichi Shinmura
Faculty of Economics, Seikei Univ.
Tokyo, Japan
e-mail: shinmura@econ.seikei.ac.jp

Abstract—**We establish a new theory of discriminant analysis by mathematical programming (MP) and develop three MP-based optimal linear discriminant functions (Optimal LDFs). Those are Revised IP-OLDF based on a minimum number of misclassification (minimum NM, MNM) criterion by integer programming (IP), Revised LP-OLDF by linear programming (LP) and Revised IPLP-OLDF that is a mixture model of Revised LP-OLDF and Revised IP-OLDF. We evaluate these LDFs with two support vector machines (SVMs), Fisher's LDF and logistic regression. Although we could compare these LDFs by six different small samples, we could not validate these LDFs by the validation samples. Therefore, we developed "100-fold cross validation for small sample" method that is a combination of k-fold cross validation and re-sampling sample (The Method). By this break-through, we can validate seven LDFs with the 95% confidence interval (CI) of error rates and the discriminant coefficients in the training and validation samples. Especially, we can select the best model with minimum mean error rates in the validation sample (M2) instead of the leave-one-out (LOO) procedure. We compared seven LDFs using six different datasets and showed that the best models of Revised IP-OLDF are better than the other six best models by the Method.**

Keywords- ***Fisher's LDF; logistic regression; two SVMs; three Optimal LDFs (OLDFs); Best Model; LOO.***

I. INTRODUCTION

We establish a new theory of discriminant analysis by MP-based OLDFs [28]. In statistics, the discrimination means the method to classify class/object categories by independent variables. On the other hand, classification of cases by independent variables is cluster analysis. Three OLDFs, namely Revised IP-OLDF, Revised LP-OLDF, and Revised IPLP-OLDF [22] are validated with hard-margin SVM (H-SVM), soft-margin SVM (S-SVM) [29], Fisher's LDF [3] and logistic regression [1] by the "100-fold cross validation for small sample" method (The Method). It is a combination of k-fold cross validation and re-sampling sample. If we fix k=100, we can obtain the 95% confidence intervals (CIs) of error rates and the discriminant coefficients in the training and validation samples [17] [18] [20] [23]-[25]. When we fixed k=10 at first, we noticed we were not able to get 95% CIs. LOO procedure [6] cannot offer 95% CIs. There are four serious problems with discriminant analysis [21] [26]. Only Revised IP-OLDF [12] - [16] can discriminate the cases on the discriminant hyperplane exactly. Other LDFs cannot discriminate these cases correctly (Problem1). All LDFs except for H-SVM and Revised IP-OLDF cannot discriminate linear separable data (LSD) theoretically (Problem2). Problem3 is the defect of the generalized inverse matrix and effects the quadratic discriminant function (QDF) and regularized discriminant analysis (RDA) [5]. Although statisticians developed discriminant functions based on the variance-covariance matrices, we found many defects. Most statisticians misunderstand that the discriminant analysis is the inferential statistics as same as the regression analysis. Although Fisher proposed Fisher's LDF and established the theory of discriminant analysis, he never proposed the standard error (SE) of error rate and discriminant coefficients (Problem4), nevertheless Fisher's LDF assume Fisher's assumption. In this paper, we discuss on Problem4 and propose the Method using iris data [2] because it is relevant evaluation data of discriminant analysis. Because the iris data is not LSD, we cannot discuss H-SVM for this data.

In Section 2, we explain five MP-based LDFs. In our research, we compare two statistical LDFs and five MP-based LDFs by the Method. We code the Method of Fisher's LDF and logistic regression by JMP script [7] and do not discuss in this paper. We discuss five MP-based LDFs coded by LINGO [8].

In Section 3, we explain the Method. By this break-through, we can validate seven LDFs by the 95% CIs and best models. Genuine statisticians established the inferential statistics by their creative brain and theoretical distribution. Because the Method is a computer-intensive approach by computer power and software of MP and statistics, we had better consider the Method is not traditional inferential statistics that is more straightforward than LOO procedure.

In Section 4, we explain the results of iris data by the Theory because Fisher's LDF is most suitable for iris data. Fisher proposed Fisher's LDF under Fisher's assumption that two classes have the same normal distributions and two different means. Because statisticians have difficulty to develop good test statistics for Fisher's assumption, we usually obtain MP-based LDFs and logistic regression better results than Fisher's LDF for many real data, most of whom may not satisfy Fisher's assumption.

In Section 5, we summarize the results by the Theory using CPD data [9], Swiss banknote data [4], student data [11], six pass/fail determination using exam scores [19] and Japanese automobile data [28].

II. MP-BASED LDFS BY LINGO

A. The Iris Data in Excel

We explain the Method using iris data that is critical evaluation data in the discriminant analysis. It consists of three species as follows: setosa, versicolor, and virginica. Each species has 50 cases. There are four variables, such as: X1 (petal width), X2 (petal length), X3 (sepal width) and X4 (sepal length). Because we can separate the setosa from other two species by the scatter plot quickly, we usually omit the setosa and focus on the two-class discrimination of versicolor (y_i =1) and virginica (y_i = -1) in Table 1. All values of class2 are changed negative values. We define Excel range name 'IS' that is "B2:F101." LINGO can retrieve 'IS' array values by "IS = @OLE( );" function and use it as LINGO array name 'IS.' Next, we define the Excel range name 'CHOICE' that is "I2:M16" in Table 2. Fifteen rows correspond the models from the full model (X1, X2, X3, X4) to the 1-variable model (X1). If the model includes the variable, the value is '1,' otherwise it is '0.'

TABLE I. THE IRIS DATA IN EXCEL

	A	B	C	D	E	F
1	species	X1	X2	X3	X4	y
2	versicolor	7	3.2	4.7	1.4	1
…	versicolor	…	…	…	…	1
51	versicolor	5.7	2.8	4.1	1.2	1
52	virginica	-6.3	-3.3	-6	-2.5	-1
…	virginica	…	…	…	…	-1
101	virginica	-5.2	-3	-5.1	-1.8	-1

TABLE II. RANGE NAME SUCH AS CHOICE

SN	p	X1	X2	X3	X4	c
1	4	1	1	1	1	1
2	3	0	1	1	1	1
3	3	1	0	1	1	1
4	3	1	1	0	1	1
5	3	1	1	1	0	1
6	2	0	0	1	1	1
…	…	…	…	…	…	…
8	2	1	0	0	1	1
12	1	0	0	0	1	1
13	1	0	0	1	0	1
14	1	0	1	0	0	1
15	1	1	0	0	0	1

After optimization, we output three arrays, such as the "NM, ZERO and VARK100" to the Excel range name by "@OLE( ) = NM, ZERO, VARK100;" function. "NM: (N2: N16)" stores 15 NMs. "ZERO:(O2:O16)" stores the number of cases on discriminant hyperplane of 15 models. "VARK100: (P2:T16)" stores the coefficients of 15 models.

B. Five LDFs to Solve Original Data by LINGO

In this paper, we explain the model by LINGO, which is the solver developed by LINDO Systems Inc. [8]. We develop six LDFs; those are Revised IP-OLDF (RIP), Revised IPLP-OLDF (IPLP), Revised LP-OLDF (LP), H-SVM and two S-SVM (SVM4 for penalty c=10^4 and SVM1 for penalty c=1). In this paper, we consider two S-SVMs are different LDFs. The Revised IP-OLDF in (1) can find the right MNM by "MIN=Σe_i" because it can directly find the interior point of an optimal convex polyhedron (OCP) [10]. If case $\mathbf{x_i}$ is classified, e_i=0. If case $\mathbf{x_i}$ is misclassified, e_i=1. Because the discriminant score becomes negative for the misclassified case, Revised IP-OLDF selects alternative support vector, such as "y_i* ($^t\mathbf{x_i}$ $\mathbf{b}$+ b_0) = 1 - M*e_i=-9999" instead of "y_i*($^t\mathbf{x_i}\mathbf{b}$+b_0) =1" for misclassified cases.

$$\text{MIN}=\Sigma e_i; \qquad (1)$$

y_i* ($^t\mathbf{x_i}$ $\mathbf{b}$+ b_0) >= 1 - M* e_i ;
$\mathbf{b}$: p independent variables, b_0: the intercept,
$\mathbf{x_i}$: (1*p) case vector if data is (n*p),
($^t\mathbf{x_i}$ $\mathbf{b}$+ b_0): the discriminant score,
M: big M constant, such as 10000,
y_i: y_i= 1 for class 1 and y_i = -1 for class2,
e_i: 0/1 integer variable corresponding $\mathbf{x_i}$.

We can define this model in 'SUBMODEL' section of LINGO. 'RIP' is the sub-model name of Revised IP-OLDF. We can solve and control this IP model by this name. "@SUM and @FOR" are two essential LINGO loop functions. "@SUM (N(i): E(i))" means "$\Sigma_{i=1}{}^n$ E(i)". "@FOR(N(i):" defines n constraints, such as "@SUM(P1(j): IS(i, j) * VARK(j) * CHOICE(k, j)) >= 1-BIGM*E(i)); for i=1,…,n". "@FOR(P1(j): @FREE (VARK(j))); for j=1,…,p" defines the discriminant coefficient **b** as the free decision variable." "@FOR(N(i): @BIN(E(i))); for i=1,…,n" defines that 'e_i' are 0/1 integer variables. By these function, we can define a compact model. If we insert '!' before "@FOR(N(i): @BIN(E(i)));", it changes the only comment, and 'e_i' becomes non-negative real decision variable by the default. This model is Revised LP-OLDF. Therefore, we define the model of Revised LP-OLDF named 'LP' that is the second SUBMODEL.

```
SUBMODEL RIP (or LP):
 MIN=ER;  ER=@SUM(N(i):E(i));
  @FOR(N(i):
  @SUM(P1(j):IS(i,j)*VARK(j)*CHOICE(k,j))
          >= 1-BIGM*E(i));
  @FOR(P1(j): @FREE(VARK(j)));
  (or !) @FOR(N(i): @BIN(E(i)));
ENDSUBMODEL
```

Third, we define Revised IPLP-OLDF. In the first stage, we discriminate the data by Revised LP-OLDF. In the second phase, we discriminate the restricted cases misclassified by Revised LP-OLDF. Therefore, we must distinguish two alternatives stored in the array ‘CONSTANT’ and Revised IP-OLDF discriminate only the misclassified cases by the “SUBMODEL IPLP” that is restricted Revised IP-OLDF.

```
SUBMODEL IPLP:
  MIN=ER;  ER=@SUM(N(i):E(i));
 @FOR(N(i):@SUM(P1(j):IS(i,j)*VARK(j)*CHOICE(k,j))
        >= 1-BIGM*E(i));
 @FOR(P1(j): @FREE(VARK(j)));
 @FOR(N(I)| CONSTANT(i)#GT#0:@BIN(E(I)));
  @FOR(N(I)| CONSTANT(i)#EQ#0:E(I)=0);
ENDSUBMODEL
```

In the ‘CALC’ section, we insert the below statements for Revised IPLP-OLDF that is a mixture model of Revised LP-OLDF and restricted Revised IP-OLDF.

```
@SOLVE(LP);
@FOR(N(i):@IFC(E(I)#EQ#0:CONSTANT(i)=0;  @ELSE
   CONSTANT(i)=1;));
MNM=0; ER1=0; MNM2=0; ER2=0;
@FOR(P1(J):VARK(J) =0; @RELEASE( VARK( J)));
@SOLVE(IPLP);
```

S-SVM has two objects in (2). These two objects are combined by defining some “penalty c.” We must define the value of penalty in the CALC section. In this research, two S-SVMs, such as SVM4 and SVM1, are examined. We know the mean error rates of SVM4 are almost better than SVM1. If we delete the second object “c* Σe$_i$” and “-M*e,” it becomes H-SVM that is not used in this paper.

$$\begin{aligned} &\mathrm{MIN} = \|\mathbf{b}\|^2/2 + c^* \Sigma e_i\,; \qquad (2)\\ &\quad y_i^* ({}^t\mathbf{x_i}\,\mathbf{b}+ b_0) >= 1-\ M^*e_i\,;\\ &\quad \mathbf{b},\ \mathbf{x_i},\ ({}^t\mathbf{x_i}\,\mathbf{b}+ b_0),\ y_i,\text{: same in (1)}\\ &\quad c : \text{penalty } c,\ e_i\text{: non-negative decision variable.}\end{aligned}$$

```
SUBMODEL SSVM:
MIN=ER;ER=@SUM(P(J1):
VARK(j1)^2)/2+Penalty*@SUM(N(i):E(i));
 @FOR (N(i): @SUM(P1(j):IS(i,j)*VARK(j)*
    CHOICE(k,j)) >= 1-E(i));
 @FOR (P1(j): @FREE(VARK(j)));
ENDSUBMODEL
```

If we insert five LDFs before the ‘CALC’ section, we can easily discriminate the original data by five LDFs.

C. *Discrimination of the Iris Data by LINGO*

We can discriminate the iris data by MP-based LDFs using “SETS, DATA, five SUBMODELs, CALC, and second DATA” sections. In the ‘SETS’ section, “P, P1, N and ERR(MS)” are one-dimensional sets, element numbers of those are 4, 5, 100 and 15 defined in ‘DATA’ section. Set ‘P1’ has one-dimensional array ‘VARK’ that stores the discriminant coefficient of one discriminant model. Set ‘N’ has two one-dimensional arrays. ‘E’ stores the 100 binary integer values of ‘e$_i$’ and ‘CONSTANT’ stores 100 discriminant scores. “MS” has the two one-dimensional arrays. ‘NM’ and ‘ZERO’ store the number of misclassifications (NM) and the number of cases on the discriminant hyperplane. If we discriminate the data by RIP, ‘NM’ column shows MNMs of 15 models. From ‘ZERO’ column, we can confirm Revised IP-OLDF is free from the Problem1. Because other LDFs cannot avoid the Problem1, all LDFs must check these numbers. Now, we cannot trust the output of NMs by statistical LDFs. ‘VARK100’ stores 15 coefficients of Revised IP-OLDF.

```
MODEL:
SETS:
 P; P1: VARK; P2; N: E, CONSTANT; MS: NM, ZERO;
 D(N, P1):IS; MB(MS, P1):CHOICE;
 VP(MS, P1):VARK100;
ENDSETS
DATA:
  P=1..4; P1=1..5; N=1..100; MS=1..15;
  CHOICE, IS = @OLE( );
ENDDATA
```

```
! Here, insert six SUBMODELs (LDFs).
```

```
CALC:
@SET('DEFAULT'); @SET('TERSEO',2);
  K=1; G=1; LEND=@SIZE(MS);
@WHILE(K#LE#LEND:
@FOR( P1( J): VARK( J) = 0;
@RELEASE( VARK( J)));NM=0; Z=0; Penalty=10000;
@SOLVE(RIP); !Change the submodel name.;
 @FOR(P1(J1): VARK100(@SIZE(MS)*(G-1) +K, J1)
       =VARK(J1)*CHOICE(k,J1));
 @FOR(n(I):  CONSTANT(i)= @SUM(P1(J1): IS(i,J1)
                    *VARK(J1)*CHOICE(k,J1)));
 @FOR(n(I): @IFC(CONSTANT(i) #EQ#0: Z=Z+1));
 @FOR(n(I): @IFC(CONSTANT(i) #LT#0: NM=NM+1));
    NM(K)=NM; ZERO(K)=Z; K=K+1);
ENDCALC
DATA:
        @OLE( )=NM, ZERO, VARK100;
ENDDATA
END
```

III. THE THEORY

A. *The Method Outlook*

In this paper, we proposed the Method, as follows [17].

*1) Let n be the number of cases and p be the number of variables including the intercept y_i (y_i = 1 for class1; y_i = -1 for class2). We copy the original data (n-cases by p-variables) 100 times and generate pseudo-population sample (100*n cases by p-variables).*

*2) We add the random number to this sample (100*n cases by (p+1)-variables) and sort it in ascending order by the random number. We divide this sample by 100 sub-samples and add the sub-sample number from 1 to 100.*

*3) We use 100 sub-samples as the training samples (n cases by (p+1)-variables) and the pseudo-sample as the validation sample (100*n cases by (p+1)-variables). This operation implies us that we re-sample 100 sub-samples from the pseudo-population. If we consider one sub-sample is the training sample, and other 99 sub-samples is the validation sample, we cannot estimate results uniformly because 100 validation samples are different. Moreover, if we fix the validation sample uniquely, we can control the training samples and validation sample very easy. For example, we can validate the validation sample generated by the original data because both samples are the same distribution. Moreover, we can get 95% CIs of the discriminant coefficients and propose the best model as model selection procedure instead of LOO procedure.*

B. How to generate the re-sampling sample and prepare the data in Excel file

We generate re-sampling sample from the original iris data and evaluate seven LDFs by the Method. Each species compose of 50 cases with 4-variables and classifier y_i. We copy each species 100 times. We add the random number (R column) as the seventh variable and sort it in ascending order in Table 3. Variable names "A1:X1 and R" are located in cells "A1 and G1." We consider this dataset is a pseudo-population and the validation sample that has the same statistics values, such as the average and range as the original data. We can control many research datasets and reduce mistakes.

TABLE III. RESAMPLING SAMPLE: ES

A1:X1	B1:X2	C1:X3	D1X4	y_i	SS	R
x(1,1)	x(2,1)	x(3,1)	x(4,1)	1	1	
				1	….	
				1	1	
				1	….	
				1	100	
				1	…	
x(1,5000)	x(2,5000)	x(3,5000)	x(4,5000)	1	100	
-x(1,5001)	-x(2,5001)	-x(3,5001)	-x(4,5001)	-1	1	
				-1	….	
				-1	1	
				-1	….	
				-1	100	
				-1	…	
-x(1,10000)	-x(2,10000)	-x(3,10000)	-x(4,10000)	-1	100	

Next, we divide this sample into 100 sub-samples and add the sub-sample number (SS column) from 1 to 100 as the sixth variable. Each sub-sample consists 100 cases and seven variables in Table 3. Six variables excluding 'R' are input by "ES= @OLE ();" in the 'DATA' section of next 'D.' The '@OLE ( )' function input the data ES on Excel range name, such as "A2: F10001" if the cell of 'X1' is located in `A1', and define the LINGO array ES. The 100 sub-samples play the training samples, and a total re-sampling sample is used as the validation sample. We consider the validation sample is a suede-population, and the training samples are the samples from the suede-population. We should fix the validation sample uniquely and evaluate the training samples by suede-population.

C. Set Notation Model by LINGO

Fisher never formulated the equation of SE for error rate and discriminant coefficient. If we discriminate the data by the Method, we can easily calculate the 95% CIs of error rates and discriminant coefficients. We obtain the Philosopher's Stone to validate seven LDFs by six small datasets. 'SET' section defines six one-dimensional sets, such as P, P1, P2, N, MS, and G100. "P, P1, and P2" are the number of independent variables, the number of (independent variables + intercept) and the number of (independent variables + intercept + sub-sample No.), respectively. These dimensions of elements in 'DATA' section are 4, 5 and 6, respectively. Only 'P1' defines one-dimensional array named 'VARK' with 5-elements that store the discriminant coefficients of the training sample.

Five sets "N, N2, MS, MS100 and G100" are one-dimensional sets, the elements of those are 100, 10000, 15, 1500 and 100 elements, respectively. Two-dimensional set 'D(N, P1):' with 100*5 has the same size array 'IS' that stores the 100 sub-sample with p-variables as the training samples. "D2(N2, P2):" with 10000*6 has the same size array 'ES' that stores the resampling-sample as the validation sample. The set ERR(MS, G100) with 15*100 has four arrays. The IC and IC_2 store MNM or NM in the training and validation samples. The EC and EC_2 store the number of cases on the discriminant hyperplanes in both samples. The set SS(N2, MS) with 10000*15 has the array SCORE2 that stores the discriminant scores of 15 models. The set VVV(MS100, P2) with 1500*6 has the array VARK100 that stores 1500 coefficients of the 100 training samples.

In the DATA section, we define nine parameters values and input two arrays, such as CHOICE and ES. The 'CHOICE' stores the pattern of 15 models showed in Table 2. The 'ES' stores the validation sample in Table 3. In CALC section, the training sample IS with 100*5 chooses 100 rows of ES by the sub-sample number (SS column).

D. Total Model with CALC Section by LINGO

After we define the "SETS and DATA" section, we insert six LDFs described in 'B' of Section 2. We divide two parts of the 'CALC' section. The first part is the default setting of output, global search, QP, multi-thread, etc.

```
MODEL: The Method for the Iris data;
SETS:
   P; P2; P1: VARK;
   N:;  N2: ; MS : ; MS100 : ;  G100 :;
   D (N, P1):IS;
   D2 (N2, P2):ES;
   MB (MS, P1): CHOICE;
   ERR(MS, G100):IC, IC_2, EC, EC_2;
   SS(N2, MS):SCORE2;
   VVV(MS100, P2):VARK100;
ENDSETS
DATA:
  P=1..4; P1=1..5; P2=1..6;
  N=1..100; N2=1..10000;
  MS=1..15; G100=1..100; MS100=1..1500 ;
  BIGM=10000;  ! for SVM4;
  CHOICE, ES=@OLE();
ENDDATA
! Here, insert five LDFs described in 'B' of Section 2.
```

```
CALC:
! Reset all options to default; @SET('DEFAULT');
! @SET('TERSEO',1);!Allow for minimal output;
@SET('TERSEO',2);
!Global solver (1:yes, 0:no); @SET('GLOBAL',1);
!Quadratic recognition (1:yes, 0:no);@SET('USEQPR',1);
!Multisarts (1:Off, >1 number of starts);
@SET('MULTIS',1);
!Number of threads; !@SET('THRDS',4);
!Print output immediately (1:yes, 0:no);
@SET('OROUTE',1);
!No need to compute dual values; @SET('DUALCO',0);

K=1; Lend=@SIZE(MS);
@WHILE (K#LE#Lend: f=1;
@WHILE (f#LE#100:
   @FOR(D(i, j): IS(i, j)=ES( @SIZE(N)*(f-1)+i, j));
      MNM=0; ER1=0;MNM2=0;ER2=0;
   @FOR( P1( J): VARK( J) = 0;@RELEASE( VARK( J)));

   @SOLVE ( RIP );! Set the submodel name here;

   @FOR(P1(j): VARK100(100*(k-1)+f,j)=VARK(j));
    VARK100 (100*(k-1)+f, @SIZE(P2))=K;
@FOR(n(l):SCORE(l)=@SUM(P1(j):IS(l,j)*VARK(j)*
      CHOICE (k, j)));
@FOR(n2(nn):SCORE2(nn,K)=@SUM(P1(j):ES(NN, j)*
      VARK(j)*CHOICE(k, j)));
@FOR(n(l): @IFC(SCORE(l)#LT#0:  MNM=MNM+1));
@FOR(n2(nn):
   @IFC(SCORE2(nn,k)#LT#0:ER1=ER1+1 ));
@FOR(n(l):@IFC(SCORE(l)#EQ#0: MNM2=MNM2+1));
@FOR(n2(nn):@IFC(SCORE2(nn,k)#EQ#0:ER2=ER2+1 );
 IC(K,f)=MNM;EC(k,f)=ER1;
 IC_2(K,f)=MNM2;
 EC_2(K,f)=ER2;
 f=f+1);
ENDCALC
```

```
DATA:
  @OLE( )=IC, EC, IC_2, EC_2, VARK100, SCORE2;
ENDDATA
END
```

The second part of CALC section controls the optimization models that consist two loops. The big loop is repeated 15 iterations by "K=1,…,15." The small loop is repeated 100 iterations by "f=1,…,100." If we set "@SOLVE (RIP);," we can discriminate the iris re-sampling sample by Revised IP-OLDF. If we replace this command by "`SOLVE(SSVM);," S-SVM discriminate the datasets. We can choose SVM4 or SVM1 by setting "Penalty=10000 or 1" in Calc section. In the second DATA section, we output six results on Excel arrays. "IC and EC" are the 100 MNMs in the training samples and 100 NMs in the validation samples. "IC_2 and EC_2" are the 100 numbers of cases on the discriminant hyperplane in the both samples. From these figures, we calculate the mean error rates, such as "M1 and M2" in the both samples. 'VARK100' are the 1500 discriminant coefficients of 15 models. We can calculate the 95% CI of discriminant coefficients. "SCORE2" are the 10000 discriminant scores.

IV. Results of Iris Data

A. *Results of Original Data*

We investigate all combinations of discriminant models ($15 = 2^4 - 1$). Table 4 shows the 15 models from 4-variables model to four 1-variable models. The column 'SN' is the sequential number of models. The column 'Var.' denotes the suffix of variable name. The column 'RIP' is the MNMs of Revised IP-OLDF. We can confirm "MNM monotonously decreases ($MNM_k \geq MNM_{(k+1)}$)." For example, the forward stepwise technique of the regression analysis chooses the variable as follows: X4, X2, X3, and X1 in this order. The MNM of four models decreases as follows: 6, 3, 2, 1. We can confirm the monotonous decrease of MNM by other model sequences, such as X1, X2, X3, X4 in this order. The MNM of four models decreases as follows: 37, 25, 2, 1. Therefore, we cannot choose the model having minimum MNM as the best model because we always choose the full model. Six discriminant functions represent the following abbreviations in the table. SVMs are SVM4/SVM1. Revised LP-OLDF is LP. Revised IPLP-OLDF is IPLP. The logistic regression is 'Logi.' Fisher's LDF is LDF. Six columns after 'RIP' are the difference (Diff2) between (NMs of seven discriminant functions – MNM). We omitted Revised IPLP-OLDF from the table because NMs are the same as MNMs. All NMs of each model should be greater than equal to MNM because MNM is the minimum NM in the training samples. The last row shows the number of models with a minus value of 'Diff2'. Revised LP-OLDF has two minus values. This fact means that Revised LP-OLDF is not free from the Problem1. We cannot judge the Problem1 by models having "Diff2 >= 0," because we must check 'ZERO.' Although this data is expected to give the right results for Fisher's LDF, QDF and RDA, these functions based on variance-covariance matrices are not superior to MP-based LDFs. Bold numbers of 'Diff2s' among each seven discriminant functions are maximum values. There are 23 maximum values among Fisher's LDF, QDF and RDA. On the other hand, there are 15 maximum values among SVM4, SVM1, and Logi. Roughly speaking, we judge Fisher's LDF, QDF and RDA are inferior to other LDFs, although this judgment is not clear.

TABLE IV. MNM AND EIGHT DIFF2

SN	Var.	RIP	SVMs	LP	Logi.	LDF	QDF	RDA
1	1,2,3,4	1	1/0	1	1	**2**	**2**	**2**
2	2,3,4	2	0/**2**	0	0	**2**	**2**	1
3	1,3,4	2	0/0	0	0	1	1	**2**
4	1,2,4	4	**3**/1	**3**	0	1	2	1
5	1,2,3	2	2/4	2	2	5	**6**	4
6	2,4	3	1/1	**3**	0	0	2	2
7	3,4	5	**3**/2	1	1	**3**	0	2
8	1,3	4	1/**3**	1	0	2	2	2
9	1,4	6	**1/1**	0	0	**1**	0	0
10	2,3	5	0/0	**1**	0	**1**	**1**	**1**
11	1,2	25	2/2	2	0	0	**4**	**4**
12	4	6	**0/0**	**0**	**0**	**0**	**0**	**0**
13	3	7	0/0	0	0	**1**	0	0
3	1	37	0/0	-3	0	0	**3**	**3**
15	2	27	**5/5**	-2	0	**5**	**5**	**5**
		-	0	**2**	0	0	0	0

1) *Diff2 of Revised IPLP-OLDF is omitted from the table because all values are zero.*
2) *Column 'SVMs' denotes both values of SVM4/SVM1.*

We cannot select the best model by MNM or error rate in the training samples. Until now, we have two options to choose a good model from the original data or training sample. The first option is the LOO procedure. The second option is to evaluate models by the model selection statistics of regression analysis. Table 5 is the result of all possible combination of models. The column 'Model' shows 15 models from 4-variables model to 1-variable model. The column 'p' indicates the number of variables. Within the same 'p,' models are descending order of "R-square (R2)". The column 'Rank' is the ranking within the same number of 'p.' This procedure is very powerful because we can overlook all models and simulate the forward and backward stepwise techniques. Both techniques choose the same models, such as: (X4) -> (X4, X2) -> (X4, X2, X3) -> (X4, X2, X3, X1). Therefore, we can easily choose a good model among these four models. Model selection statistics, such as AIC, BIC, and Cp statistics, choose the full model as a good model. However, these statistics usually select different models by other data. Therefore, we cannot usually decide a good model by these statistics uniquely.

TABLE V. THE RESULT OF ALL POSSIBLE COMBINATION

Model	p	Rank	R2	AIC	BIC	Cp
1,2,3,4	4	1	0.78	143.49	158.22	5.00
2,3,4	3	1	0.77	148.70	161.09	10.37
1,3,4	3	2	0.76	151.80	164.18	13.59
1,2,4	3	3	0.73	163.89	176.27	27.16
1,2,3	3	4	0.70	174.19	186.58	40.09
2,4	2	1	0.72	163.52	173.52	27.39
3,4	2	2	0.72	165.00	175.00	29.19
1,3	2	3	0.70	172.71	182.71	39.07
1,4	2	4	0.69	176.43	186.43	44.12
2,3	2	5	0.63	192.14	202.14	67.61
1,2	2	6	0.25	263.97	273.97	237.44
4	1	1	0.69	174.27	181.83	42.12
3	1	2	0.62	193.68	201.25	71.72
1	1	3	0.24	262.02	269.59	236.18
2	1	4	0.09	280.07	287.63	301.87

B. *Results by the Method*

Table 6 shows the results of 15 models by the Method. The first 15 models of RIP show all possible combination of models from a 4-variables model to a 1-variable model shown in column 'Model'. "M1 and M2" columns are the mean of error rates in the both samples. 'M1' decreases monotonously the same as MNM, because M1 is the average of 100 MNMs. Therefore, M1 of the full model is always minimum value theoretically. We can confirm this fact by the values of M1 in the table. Although M2 of the full model happen to be the minimum value, and it is 2.72, this may be caused by the reason this data has only four variables. We consider the model with minimum M2 is the best model. We claim the best model has good generalization ability. The column 'Diff' is the difference between (M2 - M1). Because a 1-variable model (X4) has a minimum value of 'Diff,' these statistics is not useful to choose the best model. We confirmed this fact by many types of research.

We summarize 15 models of other LDFs in two rows. The first row corresponds to the full model. All LDFs choose the full model as their best models. Those M2s are 3.03, 3.00, 2.98, 2.70, 3.07, and 3.18 %, respectively. The second row corresponds to the model with minimum 'Diff.' Last two columns, such as "M1Diff & M2Diff" are the differences between (M1/M2 of other LDFs – those of RIP). If we focus on 'M2Diff' of the full model, those are 0.31, 0.28, 0.26, -0.02, 0.35 and 0.46 % higher than Revised IP-OLDF, respectively. Therefore, six LDFs are not so bad than Revised IP-OLDF. The values of 'M2Diff' are almost less than those of 'M1Diff.' This fact may imply that Revised IP-

OLDF over-fit the training sample. We observed this defect only in this data. The column ‘Diff’ is the difference between (M2-M1). We misunderstand the model with a minimum value of ‘Diff’ has good generalization ability. If we check the ‘Diff,’ we can understand this claim is not right. Especially, although ‘Diff’ of Fisher’s LDF is -0.42%, this result is caused by the high value of M1, such as 40.72%. We claim the full model of Revised IPLP-OLDF has good generalization ability among seven LDFs. CPU times showed in full model rows tell us Fisher’s LDF and logistic regression are slower than MP-based LDFs.

TABLE VI. THE COEFFICIENTS OF SEVEN LDFS

RIP	M1	M2	Diff.	Model	
1 12m11s	0.56	**2.72**	2.16	X1, X2, X3, X4	
2	0.96	3.03	2.07	X2, X3, X4	
3	1.37	3.42	2.05	X1, X3, X4	
4	2.68	5.07	2.39	X1, X2, X4	
5	1.55	3.70	2.15	X1, X2, X3	
6	3.61	5.79	2.18	X2, X4	
7	2.44	4.39	1.95	X3, X4	
8	2.91	4.82	1.91	X1, X3	
9	4.23	5.69	1.46	X1, X4	
10	4.29	7.03	2.74	X2, X3	
11	22.74	27.27	4.53	X1, X2	
12	5.40	6.08	**0.68**	X4	
13	5.88	7.25	1.37	X3	
14	25.75	28.24	2.49	X1	
15	35.67	38.93	3.26	X2	
SVM4	M1	M2	Diff.	M1Diff.	M2Diff.
1 8m43s	1.21	**3.03**	1.82	0.65	0.31
12	6.00	6.06	**0.06**	0.60	-0.02
SVM1	M1	M2	Diff.	M1Diff.	M2Diff.
1 8m42s	2.23	**3.00**	0.77	1.67	0.28
12	6.16	6.28	**0.12**	0.76	0.20
LP	M1	M2	Diff.	M1Diff.	M2Diff.
1 4m20s	1.15	**2.98**	1.83	0.59	0.26
12	5.74	5.83	**0.09**	0.34	-0.25
IPLP	M1	M2	Diff.	M1Diff.	M2Diff.
1 16m39s	0.56	**2.70**	2.14	0.00	-0.02
12	5.44	6.08	**0.64**	0.04	0.00
Logistic	M1	M2	Diff.	M1Diff.	M2Diff.
1 18m	1.36	**3.07**	1.71	1.50	0.35
15	40.68	40.30	**-0.38**	5.01	1.37
LDF	M1	M2	Diff.	M1Diff.	M2Diff.
1 16m	2.76	**3.18**	0.42	2.20	0.46
15	40.72	40.30	**-0.42**	5.05	1.37

Table 7 shows the three percentiles of the discriminant coefficients and the intercept. To fix the “intercept=1,” we divide the original five coefficients by the value of (original intercept + 0.00001) to avoid the zero divide if original “intercept=0.” By fixing the intercept, we can understand the meaning of the 95% CI of coefficients clearly [26]. Before adjusting the intercept, we struggle many 95% CI of coefficients include 0 because the signs of intercept almost have both plus and minus values [24]. Although Shinmura [17] proposed this idea, we could not obtain good results because we did not fix the intercept. Four 95% CI of the full model of Revised IP-OLDF includes zero, and we cannot reject the null hypothesis at 5% level. On the other hand, we can reject three coefficients of a 3-variables model (X2, X3, X4) at 5% level.

TABLE VII. THE 95% CI OF LDFS

	%	X1	X2	X3	X4	C
RIP	97.5	4.55	5.35	9.94	12.31	1
	50	0.06	0.11	-0.23	-0.41	1
	2.5	-5.59	-11.94	-6.93	-6.34	1
	97.5		**1.25**	**-0.06**	**-0.14**	1
	50		**0.18**	**-0.15**	**-0.54**	1
	2.5		**0**	**-0.53**	**-1.36**	1

If we choose the medians as the coefficient, we get the LDF in (3). Although we judge the full model of Revised IP-OLDF is the best model, the 95% CI of Revised IP-OLDF tells us this model may be redundant and suggest a 3-variables model as a useful model. There is a mismatch between our judgment of the model selection using M2 and the 95% CI of discriminant coefficients in the best model. We usually experienced this uncertainty in inferential statistics, also.

$$RIP= 0.18*X2-0.15*X3-0.54*X4+1. \qquad (3)$$

We cannot reject four coefficients of Revised LP-OLDF in (4), three coefficients of Revised IPLP-OLDF in (5), and two coefficients of SVM4 in (6). We can reject only four coefficients of SVM1 in (7). If we check a 3-variables model, we can reject three coefficients of four LDFs the same as Revised IP-OLDF. Before we did not fix the intercept, we lost many research time and had no knowledge about the discriminant coefficients. To summarize these results, we cannot obtain clear results of the 95% CI of the coefficient.

$$LP \ = 0.06*X1+0.13*X2-0.21*X3-0.46*X4+1 \qquad (4)$$

$$IPLP = 0.52*X1+0.11*X2-0.21*X3-0.39*X4+1 \qquad (5)$$

$$SVM4= 0.06*X1+0.13*X2-0.22*X3-0.43*X4+1 \qquad (6)$$

$$SVM1=0.08*X1+0.11*X2-0.28*X3-0.28*X4+1 \qquad (7)$$

V. CONCLUSION

In this research, we specified how to discriminate the original data and re-sampling data by the Method. We can compare five MP-based LDFs and two statistical LDFs. We obtain remarkable results.

1) We propose the new model selection procedure as the best model of each LDFs. We can easily compare and evaluate seven LDFs by the best models because we can evaluate seven LDFs by the minimum mean values of M2. In many evaluations, Revised IP-OLDF and Revised IPLP-OLDF is the best. Next, logistic regression is superior to SVM4 in many trials. M2 of SVM1 is almost greater than M2 of SVM4. Fisher's LDF are almost the worst except for the iris data.

2) Next, IP-OLDF found the Swiss banknote data is LSD, and 16 models including (X4, X6) are linear separable models. Other 47 models are not linear separable models. We can conclude H-SVM and Revised IP-OLDF can recognize LSD theoretically. Other LDFs are not free from the Problem 2. It is hard for us to find LSD occasionally. We locate the pass/fail determination of exam scores give us good research data for linearly separable models [19]. By these examinations, the error rates of Fisher's LDF are 20% worse than Revised IP-OLDF with MNM=0. Therefore, we claim the discriminant functions based on the variance-covariance matrices are fragile for the discrimination of data that has many cases nearby the discriminant hyperplane. We had better re-evaluated the old principal researchers discriminated by these functions.

3) Many statisticians struggle to select feature of microarray datasets because it has many variables (genes) (Problem5). Only Revised IP-OLDF can select feature naturally and shows that high dimensional gene space consists several small disjoint unions of gene sub-spaces those are linearly separable. Therefore, we can analyze these small gene sub-spaces by the common statistical methods [27].

4) The Method solves Problem4 for six MP-based LDFs instead of LOO [6]. Revised IP-OLDF solves Problem1, 2 and 5. H-SVM solves Problem2. Other LDFs can not solve Problem1and Problem2 theoretically.

5) We should not use the iris data for evaluation of discriminant analysis because it cannot tell us the differences of discriminant functions.

REFERENCES

[1] D. R. Cox, "The regression analysis of binary sequences (with discussion)," J Roy Stat Soc, B20, pp.215-242, 1958.

[2] A. Edgar, "The irises of the Gaspe Peninsula," Bulletin of the American Iris Society, Vol. 59, pp. 2-5, 1945.

[3] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol. 7, pp. 179–188, 1936.

[4] B. Flury and H. Rieduyl, " Multivariate Statistics: A Practical Approach," Cambridge University Press, 1988.

[5] J. H. Friedman, "Regularized Discriminant Analysis," Journal of the American Statistical Association, Vol. 84/405, pp. 165-175, 1989.

[6] P. A. Lachenbruch, and M. R. Mickey, "Estimation of error rates in discriminant analysis," Technometrics, Vol. 10, pp.1-11, 1968.

[7] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics, third ed, SAS Institute Inc. 2004.

[8] L. Schrage, Optimization Modeling with LINGO, LINDO Systems Inc. 2006.

[9] S. Shinmura, "Optimal Linearly Discriminant Functions using Mathematical Programming," Journal of the Japanese Society of Computer Statistics, Vol. 11/2, pp. 89-101, 1998.

[10] S. Shinmura, "A new algorithm of the linear discriminant function using integer programming," New Trends in Probability and Statistics, Vol. 5, pp.133-142, 2000.

[11] S. Shinmura, Optimal Linear Discriminant Function using Mathematical Programming, Dissertation, March 200, pp. 1-101, Okayama Univ., 2000.

[12] S. Shinmura, "Enhanced Algorithm of IP-OLDF," ISI2003 CD-ROM, pp.428-429, 2003.

[13] S. Shinmura, "New Algorithm of Discriminant Analysis using Integer Programming," IPSI 2004 Pescara VIP Conference CD-ROM, pp.1-18, 2004.

[14] S. Shinmura, "New Age of Discriminant Analysis by IP-OLDF –Beyond Fisher's Linear Discriminant Function-," ISI2005, pp.1-2, 2004.

[15] S. Shinmura, "Comparison of Revised IP-OLDF and SVM," ISI2009, pp.1-4, 2007.

[16] S. Shinmura, "Overviews of Discriminant Function by Mathematical Programming," Journal of the Japanese Society of Computer Statistics, Vol. 20/1-2, pp. 59-94. 2007.

[17] S. Shinmura, "The optimal linearly discriminant function," Union of Japanese Scientist and Engineer Publishing. 2010.

[18] S. Shinmura, "Beyond Fisher's Linear Discriminant Analysis –New World of the discriminant analysis-," 2011 ISI CD-ROM, pp.1-6. 2011.

[19] S. Shinmura, "Problems of Discriminant Analysis by Mark Sense Test Data," Japanese Society of Applied Statistics, Vol. 40/3, pp.157-172, 2011.

[20] S. Shinmura, "Evaluation of Optimal Linearly Discriminant Function by 100-fold cross-validation," 2013 ISI CD-ROM, pp.1-6, 2013.

[21] S. Shinmura, "End of Discriminant Functions based on Variance-Covariance Matrices," ICORE201, pp.5-16. 2014.

[22] S. Shinmura, "Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP," Statistics, Optimization and Information Computing, Vol. 2, pp. 114-129. 2014.

[23] S. Shinmura, "Comparison of Linearly Discriminant Functions by K-fold Cross-validation," Data Analytic 2014, pp.1-6, 2014.

[24] S. Shinmura, "The 95% confidence intervals of error rates and discriminant coefficients" Statistics, Optimization and Information Computing, Vol. 3, pp.66-78, 2015.

[25] S. Shinmura, "A Trivial Linear Discriminant Function. Statistics," Optimization, and Information Computing, Vol.3, pp. 322-335, 2015. DOI: 10.19139/soic.20151202.

[26] S. Shinmura, "Four Serious Problems and New Facts of the Discriminant Analysis," In E. Pinson, F. Valente, B. Vitoriano (Eds.), Operations Research and Enterprise Systems, pp.15-30, Springer (ISSN: 1865-0929, ISBN: 978-3-319-17508-9, DOI: 10.1007/978-3-319-17509-6), 2015.

[27] S. Shinmura, "Matroska Feature Selection Methods for Microarray," Biotechno 2016, pp.1-8, 2016.

[28] S. Shinmura, "New Theory of Discriminant Analysis after R. Fisher," Springer, 2016. ISBN 978-981-10-2163-3

[29] V.Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag, 1995.