STATISTICS,OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **2**, June 2014, pp 114–129.
Published online in International Academic Press (www.IAPress.org)

Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP

Shuichi Shinmura *

Faculty of Economics, Seikei University, Japan

Received 17 December 2013; Accepted 10 April 2014
Editor: David G. Yu

Abstract Shinmura [12, 13, 14] proposes an optimal linear discriminant function (OLDF) using integer programming (IP) called as IP-OLDF based on the minimum number of misclassifications (MNM) criterion. It is defined on the data and discriminant coefficient spaces. We can understand the relation of a linear discriminant function (LDF) and NM clearly. This basic knowledge tells us several new facts of the discriminant theory. If data satisfies the Harr's condition [1] or general position, IP-OLDF can obtain true MNM. But if data does not satisfy it, it may not choose the true MNM, because of the unresolved problem of the discriminant analysis that all LDFs cannot discriminate the cases $\mathbf{x}_i$ on the discriminant hyper- plane $(f(\mathbf{x}_i) = 0)$ correctly.
Therefore, Revised IP-OLDF [15, 16] is developed. However, it requires large elapsed runtime (CPU) because it is solved by IP. In this paper, we show how to reduce CPU time by Revised IPLP- OLDF, NMs of which are good estimates of MNMs. It is evaluated whether NM of Revised IPLP-OLDF almost is as low as MNM by Revised IP-OLDF. And CPU time of Revised IPLP-OLDF is remarkably improved compared with Revised IP-OLDF. These results are examined by a total of 149 different discriminant functions using by real training samples and re-sampling validation samples.

Keywords Minimum number of misclassifications (MNM), Revised IP-OLDF, Revised LP-OLDF, Revised IPLP-OLDF, logistic regression, Fisher's linear discriminant function (LDF), re-sampling

DOI: 10.19139/soic.v2i2.52

1. Introduction

In this paper, four linear discriminant functions by mathematical programming (MP) are introduced. IP-OLDF is defined by IP and looks for the minimum NM

*Correspondence to: Faculty of Economics, Seikei University, Japan.
Email: shinmura@econ.seikei.ac.jp

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

(MNM) of the training data. Discriminant rule[1] is very simple. This simplicity may hide many problems of discriminant theory. On the contrary, IP-OLDF finds many new facts about discriminant theory as follows [16]:

1. It is defined on the data and discriminant coefficient spaces. Therefore, we can understand the relation of discriminant functions and NMs. If training data consists of n cases and p-features, n linear equations $(H_i(\mathbf{b}) = {}^t\mathbf{x}_i * \mathbf{b} + 1 = 0)$ divide p-coefficients space into finite convex polyhedron. Vector $\mathbf{b}$ is discriminant coefficients, and n vectors $\mathbf{x}_i (i = 1, 2, \cdots, n)$ are n cases on data space. It is very important for us to understand the relation of discriminant coefficients $\mathbf{b}$ and NMs on coefficient space. Interior points $\mathbf{b}_j$ of each convex polyhedron correspond to LDF $(f_j(\mathbf{x}) = {}^t\mathbf{b}_j * \mathbf{x} + 1)$ on data space, and have unique NM, because interior points are surrounded by specific k linear equations and NM is decided by the number of minus half-plane of $H_i(\mathbf{b}) = 0 (i = 1, 2, \cdots, n)$. We cannot find the relation of discriminant function and NM until now, because the constant is treated as free variable and define $(p + 1)$-dimensional coefficient space. Case $\mathbf{x}_i$ on data space corresponds to linear equation $H_i(\mathbf{b}) = 0$ on discriminant coefficients space, and point $\mathbf{b}_j$ on coefficients space corresponds to discriminant functions $f_j(\mathbf{x}) = {}^t\mathbf{b}_j * \mathbf{x} + 1$. This is clear by fixing the constant of discriminant function.
2. Optimal convex polyhedron is defined as convex polyhedron, NM of which has MNM. Until now, all discriminant function cannot avoid some cases on $f(\mathbf{x}) = 0$. We have no rule how to discriminate these cases into class 1/class 2 correctly. This unresolved problem is abandoned until now. This means that NMs of all discriminant functions may not be correct. If we judge $|f(\mathbf{x})| \leq 10^{-6}$ as zero and the number of cases on $f(\mathbf{x}) = 0$ are 'm', true NM may increase at least m. It is founded that IP-OLDF finds vertex of OCP, if data is general position[2] . Only Revised IP- OLDF can find the interior point of OCP directly. If data is not general position, IP-OLDF may not find the vertex of OCP. The point $\mathbf{b}_j$ on vertex or edge of convex polyhedron is not free from the unresolved problem, because there are cases $\mathbf{x}_i$ on $f_j(\mathbf{x}_i) = 0$. If LDF finds interior point $\mathbf{b}_j$ in theoretical, this function is free from the unresolved problem. This is confirmed by checking that the number of $|f(\mathbf{x})| \leq 10^{-6}$ is zero. Therefore, all discriminant functions except for Revised IP-OLDF must output NM and this number.

[1]Let $f(\mathbf{x})$ is a linear discriminant function. If $y_i * f(\mathbf{x}_i) > 0$, $\mathbf{x}_i$ is classified into class 1/class 2 correctly. If $y_i * f(\mathbf{x}_i) < 0$, $\mathbf{x}_i$ is misclassified into class 1/class 2. We cannot decide how to discriminate $f(\mathbf{x}_i) = 0$ into class 1/class 2. We call this problem as the unresolved problem of discriminant theory.
[2]General position means that the design matrix made by features satisfies Harr's condition [1].

3. MNM decreases monotonously $MNM_q \geq MNM_{q+1}$. MNM_q is MNM of q-features, and MNM_{q+1} is MNM of $(q+1)$-features added one feature to existed q-features. Proof is very simple because OCP in q-coefficients space is included in $(q+1)$- coefficients space. This means an important fact. If $MNM_q = 0$, MNM of all models including these q-features are zero. Flury & Rieduyl [4] collect 200 genuine and counterfeit Swiss bank note data having 6 features and write a textbook about discriminant theory. IP-OLDF finds MNM of two-features (x_4, x_6) is zero. Therefore, 16 models including (x_4, x_6) are zero. It is concluded that Fisher's LDF and QDF based on the variance covariance matrices almost cannot recognize linear separable data [20].

In this research, two comparisons are tried. First, Revised IP- OLDF resolves problems of discriminant theory. But, this requires more CPU time, because this is solved by IP. Therefore, Revised IPLP-OLDF that looks for good estimate of MNM is developed. The CPU times and NMs of Revised IPLP-OLDF are compared with Revised IP-OLDF. It is concluded that the CPU time of Revised IPLP-OLDF is faster than Revised IP-OLDF, and error rates of Revised IPLP-OLDF are less than equal those of Revised IP-OLDF in the validation sample. Secondly, Revised IPLP-OLDF is compared with Fisher's LDF and logistic regression by 100-fold cross validations using 100 re-sampling samples [18, 19].

2. Linear Discriminant Functions (LDF)

2.1. Fisher's LDF and logistic regression

Fisher [3] introduces Fisher's LDF based on the maximization of ratio (between classes / within class). If we admit Fisher's assumption that the distributions of two classes are normal distributions such as $F_1(\mathbf{x} : \mathbf{m}_1, \Sigma_1)$ and $F_1(\mathbf{x} : \mathbf{m}_2, \Sigma_2)$, and variance covariance matrices are same($\Sigma_1 = \Sigma_2$), the same Fisher's LDF is derived by the plug in rule such as $log(F_1/F_2) = 0$. If variance covariance matrices of two classes are not same ($\Sigma_1 \neq \Sigma_2$), QDF is introduced. Multi-class discrimination and MT (Mahalanobis - Taguchi) theory [22] in QC are defined by Mahalanobis distance. Variance covariance matrix plays an important role in the discrimination theory. Model selection technique is achieved by the sweep operator [5]. But several serious problems are found as follows.

1. In general, NMs or error rates[3] of Fisher's LDF and QDF are worse than logistic regression. Therefore, users in medical and economic field use logistic regression instead of Fisher's LDF and QDF. This is reason

[3] See [24] about the relation of sample and population error rates under Fisher's assumption.

why logistic regression is free from specific distribution such as normal distribution.

2. NMs of LDF and QDF are not zero for linear separable data such as Swiss bank note data and 18 pass/fail determinations of exams [17]. Latter results are as follows. Error rates of Fisher's LDF are from 2.2% to 16.7%. Error rate of QDF is from 0.8% to 10.8% [20]. These problems are caused by the reason why real data does not satisfy Fisher's assumption.

2.2. IP-OLDF

IP-OLDF is defined in (1). Vector $\mathbf{b}$ is p-discriminant coefficients. From n cases, we obtain the optimal coefficients $\mathbf{b}$ that minimizes $\sum e_i$ by IP. The constant of linear equation $(H_i(\mathbf{b}) = {}^t\mathbf{x}_i * \mathbf{b} + 1)$ is fixed to 1 for $i = 1, \ldots, n$. This notation can show the relation of LDFs and NMs. Decision variable e_i is 0/1 integer variable for $\mathbf{x}_i$. If $\mathbf{x}_i$ is classified into class 1/class 2 correctly, $e_i = 0$ and $y_i * ({}^t\mathbf{x}_i * \mathbf{b} + 1) \geq 0$. But, if there are cases on the discriminant hyper-plane $({}^t\mathbf{x}_i * \mathbf{b} + 1) = 0$, IP-OLDF treats $e_i = 0$ and $y_i * ({}^t\mathbf{x}_i * \mathbf{b} + 1) \geq 0$ nevertheless we cannot judge which classes these cases belong to. For misclassified case $\mathbf{x}_i$, $e_i = 1$ and $y_i * ({}^t\mathbf{x}_i * \mathbf{b} + 1) \geq -10000$. This means that binary integer variable choose $({}^t\mathbf{x}_i * \mathbf{b} + 1) = 0$ or $({}^t\mathbf{x}_i * \mathbf{b} + 1) = -10000$ as the linear discriminant hyper- planes for classified /misclassified cases. Therefore, we get MNM as optimal solution if data is general position. If data is not general position, object function may not be true MNM[4].

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

where
$i = 1, 2, \cdots, n$ (n is the sample size);
$y_i = 1/-1$ for $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip}) \in$ class 1/class 2;
$e_i : 0/1$, the decision variable corresponding to each $\mathbf{x}_i$;
$\mathbf{b}$: p-discriminant coefficients vector;
M: Big M constant such as 10000.
This model in (1) is very important that the constant is fixed to be 1. We can exchange $\mathbf{b}$ and $\mathbf{x}$, and obtain $H_i(\mathbf{b}) = {}^t\mathbf{x}_i * \mathbf{b} + 1 = {}^t\mathbf{b} * \mathbf{x}_i + 1 = f(\mathbf{x}_i)$. Case $\mathbf{x}_i$ on data space corresponds to linear equation $H_i(\mathbf{b})$. Point $\mathbf{b}_j$ on discriminant coefficient space corresponds to linear discriminant function $f_j(\mathbf{x}) = {}^t\mathbf{b}_j * \mathbf{x} + 1$. Therefore, this model is considered on both data and discriminant coefficient spaces. The linear hyper-plane $H_i(\mathbf{b})$ divides the p-coefficients space into two half planes. If $\mathbf{b}_j$ is included in the plus half plane of $H_i(\mathbf{b})(y_i * ({}^t\mathbf{x}_i * \mathbf{b}_j + 1) > 0)$, it means that $y_i * ({}^t\mathbf{x}_i * \mathbf{b}_j + 1) = y_i * ({}^t\mathbf{b}_j * \mathbf{x}_i + 1) = y_i * f_j(\mathbf{x}_i) > 0$.

[4]Liitschwager & Wang [6] proposes LDF based on MNM criteria, but their model could not find the true MNM because the constraints are inaccurate.

Linear discriminant function $f_j(\mathbf{x})$ discriminates $\mathbf{x}_i$ correctly on data space. If $\mathbf{b}_i$ is included in the minus half plane of $H_i(\mathbf{b})(y_i * ({}^t\mathbf{x}_i * \mathbf{b}_j + 1) < 0)$, $f_j(\mathbf{x})$ misclassifies $\mathbf{x}_i$ on data space. N linear hyper-planes divide the discriminant coefficients space into a finite number of convex polyhedrons. Interior points of this convex polyhedron are included in the plus or minus half plane of each $H_i(\mathbf{b}) = 0$. Therefore, the interior points of same convex polyhedron have a unique NM. LDFs corresponding to these interior points classify the same cases correctly and misclassify others. If we choose LDF corresponding to the interior points, those are free from the unresolved problems. If data is general position, IP– OLDF stops the optimization by choosing just p constraints that become $H_i(\mathbf{b}) = 0$ out of n constraints. Interior points $\mathbf{b}_j$ of OCP are located in the plus side of $H_i(\mathbf{b}) = 0$ that composes OCP. If data is not general position, IP-OLDF may choose over $(p + 1)$ constraints. We cannot discriminate these $(p + 1)$ cases in class 1/class 2 theoretically. Until now, this important recognition is disregard.

2.3. Revised IP-OLDF and Revised LP-OLDF

Revised IP-OLDF is defined in (2). The constant of this discriminant function is a free variable b_0. The right-hand constant of the constraints are changed to $(1 - M * e_i)$. If $e_i = 0$, $\mathbf{x}_i$ is classified by SVs $(y_i * ({}^t\mathbf{x}_i * \mathbf{b} + b_0) \geq 1)$. If $e_i = 1$ for the misclassified cases, the constraints are relaxed $(y_i * ({}^t\mathbf{x}_i * \mathbf{b} + b_0) \geq -9999)$. The Big M constant is very important to prevent cases from being on the discriminant hyper-plane, because the misclassified cases by SVs are extracted to alternative SV $(y_i * ({}^t\mathbf{x}_i * \mathbf{b} + b_0) = -9999)$ and there are no cases in $|y_i * ({}^t\mathbf{x}_i * \mathbf{b} + b_0)| \leq 1$.

$$MIN = \sum e_i; \quad y_i * ({}^t\mathbf{x}_i * \mathbf{b} + b_0) \geq 1 - M * e_i. \tag{2}$$

where b_0: constant term (free variable).

Revised LP-OLDF is defined by changing e_i from 0/1 decision variable to real nonnegative variable. This method is one of the L1- norm methods [7, 21]. The object function is the summation of the distances from the discriminant hyper-plane of the misclassified cases, because $e_i = 0$ for the classified cases. It is as same as S-SVM[5] if penalty c is large positive number.

2.4. Revised IPLP-OLDF

Revised IPLP-OLDF is defined in two phases as follows: In the first phase, Revised LP-OLDF is applied to all cases, and these cases are categorized in two groups: cases that are classified correctly $(e_i = 0)$ and cases that are not classified $(e_i = 1)$ by SVs. In the second phase, Revised IP-OLDF is applied to latter cases.

[5] S-SVM has two objects as same as the portfolio model [23].

The CPU time may be reduced because Revised IP-OLDF analyzes restricted cases. This method is called as Revised IPLP-OLDF.

3. Comparison of Revised IP-OLDF and Revised IPLP-OLDF

In this study, four kinds of real data are used as the training samples: The student data [16] consists of 40 students with five features. The object variable consists of two classes: 25 students who pass the exam and 15 students who fail. All combinations of features $(31 = 2^5 - 1)$ are investigated. Iris data [2] consists of 100 cases with 4 features. The object variable consisted of two species: 50 versicolor and 50 virginica. All combinations of features $(15 = 2^4 - 1)$ are investigated. CPD data [11] consists of 240 patients with 19 features. The object variable consists of two classes: 180 pregnant women whose babies are born by natural delivery and 60 pregnant women whose babies are born by Caesarian section. Forty models selected by forward and backward stepwise methods are investigated, because there are $(2^{19} - 1)$ models by all combinations of features. The Swiss bank notes data [4] consists of 200 cases with 6 features. The object variable consists of two kinds of bills: 100 genuine and 100 counterfeit bills. A total of $(63 = 2^6 - 1)$ models are investigated. Four kinds of re-sampling data are generated by Speakeasy. These samples consisted of 20,000 cases and those are used as the validation samples. Revised IP-OLDF and Revised IPLP-OLDF are applied to both the training and validation samples by LINGO (Optimization Modeling Software for Linear, Nonlinear, and Integer Programming) Ver.10 [10] developed by LINDO Systems Inc. in 2008. And, both CPU times are compared from Table1 to Table4. In addition to this results, the NMs of 135 models by Revised IPLP-OLDF are compared with 135 NMs of LDF and logistic regression by 100-folf cross validation in Table5 by LINGO Ver.14 in 2014.

3.1. Swiss Bank Notes Data

Table 1 shows the result of the Swiss bank note data (bank data). The first column (Var.) shows the 63 models from 6-features (p=6) to 1-feature (p=1). In the same number of features (p), those are arranged in descending order of R-squares. Here, x_1 is length of bill (mm); x_2 and x_3 are width of left and right edges (mm); x_4 and x_5 are bottom and top margin widths (mm); x_6 is length of image diagonal (mm). Variable name is shown by only suffix number in the table. The third column (IP) shows MNM by Revised IP-OLDF. IP-OLDF finds that MNM of(x_4, x_6) is zero. Therefore, 16 models including (x_4, x_6) are linear separable. The fourth column (EC1) shows NM of the re- sampling data (or validation data) obtained by 63 discriminant functions of Revised IP-OLDF. The fifth column (%) shows the difference of two error rates as defined by the formula (EC1/20000 - IP/200)$*$100. This column means one of "generalization ability index of each

Revised IP-OLDF". Six differences are greater than 4%. We had better considered about generalization ability of each model in addition to whole models.

Table 1. Result of Bank Data (Only 6 models are shown in $p = 4, 3, 2$)

Var.	p	IP	EC1	%	LP	IPLP	EC2	%
1,2,3,4,5,6	6	**0**	0	0.0	0	0	0	0.0
2,3,4,5,6	5	**0**	0	0.0	0	0	0	0.0
1,3,4,5,6	5	**0**	95	0.5	0	0	0	0.0
1,2,4,5,6	5	**0**	799	**4.0**	0	0	0	**0.0**
1,2,3,4,6	5	**0**	807	**4.0**	0	0	531	**2.7**
1,2,3,4,5	5	1	371	1.4	2	1	389	1.4
1,2,3,5,6	5	1	115	0.1	1	1	115	0.1
3,4,5,6	4	**0**	0	0.0	0	0	0	0.0
2,4,5,6	4	**0**	0	0.0	0	0	0	0.0
1,4,5,6	4	**0**	95	0.5	0	0	0	0.0
2,3,4,6	4	**0**	0	0.0	0	0	0	0.0
1,3,4,6	4	**0**	1303	**6.5**	0	0	531	**2.7**
1,2,4,6	4	**0**	1303	**6.5**	0	0	531	**2.7**
4,5,6	3	**0**	0	0.0	0	0	0	0.0
3,4,6	3	**0**	0	0.0	0	0	0	0.0
1,4,6	3	***0***	*1303*	***6.5***	0	0	531	**2.7**
2,4,6	3	**0**	0	0.0	0	0	0	0.0
3,4,5	3	2	198	0.0	3	2	198	0.0
2,4,5	3	2	198	0.0	2	2	198	0.0
4,6	2	**0**	0	0.0	0	0	0	0.0
4,5	2	3	277	-0.1	4	3	282	-0.1
3,6	2	1	115	0.1	1	1	115	0.1
5,6	2	1	115	0.1	1	1	115	0.1
2,6	2	1	115	0.1	1	1	115	0.1
1,6	2	1	115	0.1	1	1	115	0.1
6	1	2	211	0.1	2	2	211	0.1
4	1	16	1589	-0.1	16	16	1589	-0.1
5	1	**47**	4758	0.3	**45**	**47**	4758	0.3
3	1	43	4331	0.2	43	43	4331	0.2
2	1	48	4791	0.0	53	48	4791	0.0
1	1	55	11869	**31.8**	87	55	11869	**31.8**

The LP column shows NMs by Revised LP-OLDF in the first phase of Revised IPLP-OLDF. NM of 1- variable (x_5) is 45 and is less than MNM=47. MNM is the lower limit of NMs of all LDFs. This shows that Revised LP-OLDF is not free from the unresolved problem. The IPLP column shows the estimates of MNM by Revised IPLP-OLDF in the second phase. All 63 results of both functions are same. The EC2 column shows NMs in the validation samples. The second "%" column shows the difference of the two error rates by the formula (EC2/20000 - IPLP/200)$*$ 100. By comparison of two '%' columns tell us that the values of Revised IPLP-OLDF are less than those of Revised IP-OLDF. This may show that generalization ability of Revised IPLP-OLDF is better than Revised IP-OLDF in the whole models. MP-based models are solved by fixing some cases on the discriminant hyper-plane or SVs. Therefore, these discriminant functions

cannot count NM correctly because some cases lie on the discriminant hyper-plane. In the case of Revised LP-OLDF, some cases are fixed on the SVs. But, if $e_i = 1/10000 = 0.0001$, $\mathbf{x}_i$ lies on the discriminant hyper-plane. This is examined in future research.

The CPU times of Revised IP-OLDF and Revised IPLP-OLDF of 63 models are 133,399 seconds and 2688 seconds, respectively. Revised IPLP-OLDF is approximately 50 times faster than Revised IP-OLDF.

3.2. Iris Data

Table 2. Result of Iris Data

Var.	p	IP	EC1	%	LP	IPLP	EC2	%
1,2,3,4	4	1	204	0.0	2	1	204	0.0
2,3,4	3	2	411	0.1	2	2	411	0.1
1,3,4	3	2	414	0.1	2	2	414	0.1
1,2,4	3	4	799	0.0	7	4	799	0.0
1,2,3	3	2	402	0.0	3	2	402	0.0
2,4	2	5	1020	0.1	6	5	1024	0.1
3,4	2	3	622	0.1	6	3	622	0.1
1,3	2	4	817	0.1	5	4	823	0.1
1,4	2	5	1024	0.1	6	5	1024	0.1
2,3	2	6	1209	0.0	6	6	1213	0.1
1,2	2	25	4924	-0.4	27	25	4975	-0.1
4	1	6	1232	0.2	6	6	1232	0.2
3	1	7	1413	0.1	7	7	1408	0.0
1	1	**27**	5362	0.2	**25**	**27**	5362	-0.2
2	1	**37**	7351	0.2	**34**	**37**	7351	-0.2

Table 2 shows the result of the iris data. The first column (Var.) shows the 15 models from p=4 to p=1. x_1 through x_4 mean sepal length (x_1), sepal width (x_2), petal length (x_3), petal width (x_4), and species (x_5). The third column (IP) shows MNM by Revised IP- OLDF. The fourth column (EC1) shows NM of the re-sampling data by obtained 31 discriminant functions of Revised IP-OLDF. The fifth column (%) is defined by the formula (EC1/20000 - IP/100)$*$ 100. The LP column shows NMs by Revised LP-OLDF. NMs of two 1-var models such as (x_1) and (x_2) are less than MNMs. The IPLP column shows the estimates of MNM by Revised IPLP-OLDF in the second phase. All 15 results of both functions are same. The EC2 column shows NM in the validation samples. The second '%' column is defined by the formula (EC2/20000 - IPLP/100) $*$ 100. All absolute values of both '%' columns are less than 0.4%. This implies us that both Revised IP-OLDF and Revised IPLP-OLDF are good generalization ability for iris data. The CPU times of Revised IP- OLDF and Revised IPLP-OLDF of the 15 models

are 446 seconds and 30 seconds. Revised IPLP-OLDF is approximately 15 times faster than Revised IP-OLDF.

3.3. Student Data

Table 3. Result of Student Data

Var.	p	IP	EC1	%	LP	IPLP	EC2	%
1,2,3,4,5	5	3	2004	3	4	3	2004	3
1,2,3,5	4	3	2004	3	4	3	2004	3
1,2,3,4	4	3	2004	3	5	3	2004	3
1,3,4,5	4	3	2004	3	6	3	2004	3
1,2,4,5	4	4	2099	0	6	4	2099	0
2,3,4,5	4	3	2004	3	6	3	2004	3
1,2,3	3	3	2004	3	4	3	2004	3
1,3,5	3	3	2004	3	4	3	2004	3
1,3,4	3	5	2486	0	7	5	2486	0
1,2,4	3	5	2486	0	7	5	2486	0
1,2,5	3	3	2004	3	6	3	2004	3
2,3,4	3	4	2637	3	7	4	2637	3
2,3,5	3	3	2004	3	4	3	2004	3
3,4,5	3	3	2004	3	4	3	2004	3
1,4,5	3	6	3720	4	8	6	3720	4
2,4,5	3	5	2808	2	7	5	2808	2
1,3	2	5	2831	2	5	5	2831	2
1,2	2	5	2486	0	9	5	2486	0
2,3	2	5	3034	3	7	5	3034	3
3,4	2	5	2808	2	7	5	2808	2
3,5	2	4	2637	3	7	4	2637	3
1,5	2	4	2401	2	6	4	2401	2
1,4	2	**7**	3587	0	**6**	**7**	3587	0
2,4	2	**6**	3464	2	**5**	**6**	3464	2
2,5	2	**6**	3757	4	**3**	**6**	3757	4
4,5	2	13	6290	-1	13	13	6290	-1
3	1	8	4527	3	8	8	4527	3
1	1	**7**	3587	0	**6**	**7**	3587	0
2	1	**7**	4641	**6**	**3**	**7**	3628	**1**
4	1	13	6290	-1	13	13	6290	-1
5	1	15	10000	13	15	15	10000	13

Table 3 shows the result of the student data. The first column (Var.) shows the 31 models from p=5 to p=1. x_1 through x_5 mean the hours of study per day, number of days drinking per week, spending money per month, sex (0/1 dummy

variable), and smoking (0/1 dummy variable). The third column (IP) shows MNM of the student data. The fourth column (EC1) shows NM of the re-sampling data by Revised IP-OLDF. The fifth column (%) shows the difference of the two error rates by the formula (EC1/20000 - IP/40) $*$ 100. Three 2-variable models and two 1-variable models in LP column are less than MNMs in IP column. The IPLP column shows estimates of MNMs by Revised IPLP-OLDF. All 31 results of both functions are same. The EC2 column shows NM of the re-sampling data by Revised IPLP-OLDF. The second '%' column is defined by the formula (EC2/20000 - IPLP/40) $*$ 100. Both values of '%' columns except for 1-variable (x_2) are same. Absolute values of both '%' columns are larger than other data sets. The CPU times of Revised IP-OLDF and Revised IPLP-OLDF are 20 seconds and 40 seconds. Revised IPLP- OLDF is slower than Revised IP-OLDF, because all features are integers and many values are overlaps.

3.4. CPD Data

Table 4 shows the result of CPD data. The first column (p) shows the 40 models from p=1 to p=19. "F, B, f, and b in the column Type" show the forward (F) and backward (B) models for the full models, and forward (f) and backward (b) models for the 16-variables model dropped three variables (x_4, x_7 and x_{14}) that relate to multicolinearities. The features are as follows. x_1: age of a pregnant woman, x_2: number of times of a delivery, x_3: number of the sacrum, x_4: anteroposterior distance at the pelvic inlet, x_5: anteroposterior distance at the wide pelvis, x_6: anteroposterior distance at the narrow pelvis, x_7: the shortest anteroposterior distance, x_8: fetal biparietal diameter, and x_9: x_7-x_8, x_{10}: anteroposterior distance at the pelvic inlet, x_{11}: biparietal diameter at the pelvic inlet, x_{12}: x_{13}-x_{14}, x_{13}: area at the pelvic inlet, x_{14}: area of the fetal head, x_{15}: area at the bottom length of the uterus, x_{16}: abdominal circumference, x_{17}: external conjugate, x_{18}: interprochanteric diameter, and x_{19}: lateral conjugate. Small random noises are added to x_9 and x_{12}. The fourth column (IP) shows MNM by Revised IP-OLDF. The fifth column (EC1) shows NM of the re-sampling data by Revised IP-OLDF. The sixth column (%) is defined by the formula (EC1/20000 - IP/240) $*$100. All NMs in LP column are greater than equal those in IP column. The IPLP column shows the estimates of MNM in the second phase. All 40 results of both functions are same. The EC2 column shows NM of the re-sampling data. The second '%' column is defined by the formula (EC2/20000 - IPLP/240) $*$ 100. Comparison of two '%' columns are as follows. There are 32 models, both '%' of which are same. Seven '%' of Revised IP-OLDF are greater than those of Revised IPLP-OLDF. And only one error rate of Revised IPLP-OLDF is greater than Revised IP-OLDF. The CPU times of Revised IP-OLDF and Revised IPLP-OLDF of the 40 models are 38,170 seconds and 380 seconds. Revised IPLP-OLDF is approximately 100 times faster than Revised IP-OLDF. This large difference in CPU time may be

caused by the multicolinearity, because it may require a long time to check the convergence.

Table 4. Result of CPD Data

p	Type	Var.	IP	EC1	%	LP	IPLP	EC2	%	Sign
1	FBfb	12	20	2142	2.4	20	20	2142	2.4	=
2	FBfb	9,12	13	1815	3.7	17	13	1815	3.7	=
3	FBfb	9,12,18	12	**1647**	**3.2**	18	12	**1524**	**2.6**	>
4	Ffb	9,12,15,18	10	1285	2.3	13	10	1285	2.3	=
4	B	9,12,13,18	11	1468	2.8	19	11	1468	2.8	=
5	Ff	9,12,15,17,18	10	1468	2.8	19	10	1468	2.8	=
5	b	2,9,12,15,18	7	1043	2.3	13	7	1043	2.3	=
5	B	9,1214,18	11	1468	2.8	18	11	1468	2.8	=
6	B	9,1215,18	9	1136	1.9	13	9	1136	1.9	**=**
6	b	1,2,9,12,15,18	7	1043	2.3	14	7	1043	2.3	=
6	Ff	2,9,12,15,17,18	7	1043	2.3	14	7	1043	2.3	=
6	DOC1	5,9,13,14,17,18	12	1533	2.7	18	12	1533	2.7	=
6	DOC2	7,9,13,14,17,18	11	1361	2.2	17	11	1361	2.2	=
7	B	9,1215,17,18	9	1136	1.9	13	9	1136	1.9	=
7	Ffb	1,2,9,12,15,17,18	6	887	1.9	14	6	887	1.9	=
8	F	1,2,7,9,12,15,17,18	6	887	1.9	12	6	887	1.9	=
8	B	1,9,1215,17,18	8	980	1.6	15	8	980	1.6	=
8	fb	1,2,8,9,12,15,17,18	6	887	1.9	12	6	887	1.9	=
9	B	1,2,9,1215,17,18	6	887	1.9	13	6	887	1.9	=
9	F	1,2,5,7,9,12,15, 17,18	4	408	0.4	8	4	408	0.4	=
9	fb	1,2,5,8,9,12,15,17,18	4	539	1.0	9	4	539	1.0	=
10	B	1,2,7,9,1215,17,18	6	887	1.9	14	6	887	1.9	=
10	F	1,2,5,7,9,12,15,1719	4	**539**	**1.0**	7	4	**408**	**0.4**	>
10	fb	1,2,5,8,9,12,15,1719	3	370	0.6	8	3	370	0.6	=
11	B	1,2,5,7,9,1215,17,18	4	408	0.4	9	4	408	0.4	=
11	F	1,2,5,7,9,12,13,15,1719	4	**539**	**1.0**	9	4	**408**	**0.4**	>
11	fb	1,2,5,8,9,12,13,15,1719	3	370	0.6	8	3	370	0.6	=
12	FB	1,2,5,7,9,1215,1719	4	**539**	**1.0**	9	4	**408**	**0.4**	>
12	fb	1,2,5,8,9,12,13,1519	3	370	0.6	8	3	370	0.6	=
13	FB	1,2,4,5,7,9,1215,1719	3	**240**	**-0.1**	8	3	**370**	**0.6**	<
13	fb	1,2,5,8,9,1113,1519	3	390	0.7	9	3	390	0.7	=
14	FB	1,2,4,5,7,9,1115,1719	3	370	0.6	7	3	370	0.6	=
14	fb	13,5,8,9,1113,1519	2	214	0.2	7	2	214	0.2	=
15	FB	1,2,4,5,7,9,1119	3	370	0.6	8	3	370	0.6	=
15	fb	13,5,813,1519	2	202	0.2	5	2	202	0.2	=
16	FB	1,2,4,5,79,1119	2	202	0.2	5	2	202	0.2	=
16	fb	13,5,6,813,1519	2	214	0.2	5	2	202	0.2	=
17	FB	1,2,4,5,719	2	**334**	**0.8**	8	2	**214**	**0.2**	>
18	FB	1,2,419	2	**334**	**0.8**	5	2	**214**	**0.2**	>
19	FB	119	2	**221**	**0.3**	6	2	**102**	**0.3**	>

4. Comparison Revised IPLP-OLDF with Fisher's LDF and logistic regression by 100-fold cross validation

One hundred re-sampling samples are generated by four data. NMs of Revised IPLP-OLDF are compared with those of Fisher's LDF and logistic regression

by 100-fold cross validations [18, 19]. The results of Revised IPLP-OLDF are obtained by LINGO Ver.14 in 2014. The results of Fisher's LDF and logistic regression are obtained by JMP Ver.10 [8]. All possible models of Iris ($2^4 - 1 = 15$ models), Student ($2^5 - 1 = 31$ models), Swiss bank note ($2^6 - 1 = 63$ models) data are computed. There are ($2^{19} - 1$) models of CPD data. Therefore, only 26 models selected by the forward and backward stepwise methods are computed. At first, 100 NMs are computed for 135 different models. And, mean of error rates are computed by 135 models. Next, these 13,500 discriminant functions are applied for validation samples and computed mean error rates for validation samples. Last, four differences are computed in Table 5.

Mean error rates of difference of (LDF – Revised IPLP-OLDF) for training samples are summarized by minimum and maximum values. Minimum and maximum values of 15 different models of iris training samples are 0.55% and 5.23%. This means that mean of error rates of LDF are from 0.55% to 5.23% worse than those of Revised IPLP-OLDF. Minimum and maximum values of 15 different models of iris validation samples are -0.6% and 2.36%. Only two models out of 15 models of LDF are better than Revised IPLP-OLDF in the validation samples. In the training samples, 135 models of LDF are worse than those of Revised IPLP-OLDF. Only 15 models of LDF are better than Revised IPLP-OLDF for validation samples. Mean error rates of difference (logistic regression – Revised IPLP - OLDF) tell us that only 3 and 33 models of logistic regression are better than Revised IPLP-OLDF for the training and validation samples, respectively. In 2014, these results are recalculated using LINGO Ver.14. The elapsed runtimes of Revised IPLP-OLDF are less than 3 minutes. The elapsed runtimes of Fisher's LDF and logistic regression by JMP are over 21 minutes. The elapsed runtimes of Revised IPLP-OLDF in Ver.13 were slower than those of Fisher's LDF and logistic regression. Reversals of CPU time have occurred for this time.

Table 5. Comparison of mean of error rates
of Revised IPLP-OLDF vs. (LDF and logistic regression)

	LDF-IPLP		Logistic regression-IPLP	
135 models	Training (0) Min/Max	Validation (15) Min/Max	Training (3) Min/Max	Validation (33) Min/Max
Iris(15)	0.55/5.23	- 0.6(2)/2.36	0.59/5.31	-0.84(2)/1.85
Bank(63)	0/5.32	-0.33(10)/3.45	0/5.4	-0.3(24)/3.64
Student(31)	1.46/8.61	-1.29(3)/7.11	-2.12 (3)/6.48	-2.89(7)/5.59
CPD(26)	3.05/7.28	2.21/6.15	0.13/3.43	0.29/1.74

5. Conclusion

The CPU times of Revised IPLP-OLDF of bank data, iris data, and CPD data are 50, 15, and 100 times faster than those of Revised IP-OLDF in 2009. All NMs

obtained by Revised IPLP-OLDF are the same as the MNMs of Revised IP-OLDF. Therefore, Revised IPLP- OLDF is useful in analyzing the huge size of data such as big data. It is compared with Fisher's LDF and logistic regression in 2014. It is concluded that only 15 models of LDF are better than Revised IP- OLDF for the validation samples. And only 3 and 33 models of logistic regression are better than Revised IPLP-OLDF in the training and validation samples. Most of statistician said Revised IP-OLDF overestimates for validation samples, because it over-fits for the training sample and means of error rates are minimum values among all LDFs in theoretically. On the other hand, most of statistician believes generalization ability of LDF is good for validation samples because it is derived from Fisher's assumption without examination by the real data. But most of real data doesn't satisfy Fisher's assumption. Pass/fail determinations of exams tell us that LDF and QDF based on variance covariance matrices cannot recognize linear separable data, nevertheless these data are trivial linear separable data [17, 20].

Appendix: 100-fold cross validation of Revised IPLP-OLDF for Iris data in Table 5.

The training samples: (100 cases by 4 features).
The validation sample: (10000 cases by 4 features).

1. Preparation of Excel

Prepare the input- and output- cell array on Excel sheet. Cell array names are used in LINGO array names.
INPUT Arrays: Two input data (cell names are ES and CHOICE). ES: re-sampling sample of iris data consists of 10,000 rows by 6 columns (4 features, y_i, sub-group number GN from 1 to 100).
For first class (y_i =1), row data = $(X_{1i}, X_{2i}, X_{3i}, X_{4i}, y_i, GN)$ for $i = 1, \ldots, 50$ and GN=$1 \ldots, 100$.
For second class (y_i =-1), row data = $(X_{1i}, X_{2i}, X_{3i}, X_{4i}, y_i, GN)$ for $i = 51, \ldots, 100$ and GN=$1 \ldots, 100$.
CHOICE: all combination of features consists of 15 rows by 5 columns (4 features and the constant). Full model and model (X4) are described as (1,1,1,1,1) and (0,0,0,1,1).

2. K-fold cross validation for iris data.

MODEL: !iris data " !.....; is a comment." ;
SETS: !Sets defines sets name/dimension/array names. P1 with 5-elements is set name and VARK with 5-elements is array name;
P/1..4/; P1/1..5/:VARK; P2/1..6/; MS/1..15/:; MS100/1..1500/:;

```
N/1..100/: E, SCORE, CONSTANT; N2/1..10000/: SCORE2;
! D2 is two dimensional set with 10000 rows * 6 columns. ES is array stored in re-
sampling sample. MB is two dimensional set with 15 rows * 5 columns. CHOICE
is array stored in the discriminant functions pattern by 0/1 values.VARK100 with
1500 rows * 6 columns is stored in the 15 * 100 discriminant coefficients. IC and
EC with 15 rows * 100 columns are stored in NMs of the training and validation
samples;
D2(N2,P2):ES; MB(MS,P1): CHOIC;
VV(MS100,P2):VARK100; G100/1..100/:;
ERR(MS,G100):IC, EC; D(N,P1):IS;
ENDSETS
DATA: ! ES and CHOICE are input from Excel and are used in LINGO's array. ;
        BIGM=10000; ES, CHOICE=@OLE();
ENDDATA
SUBMODEL LP: !Submodel defines Revised LP-OLDF. If you delete '!' of
  "! @FOR(N(I):@BIN(E(I)));" , this model is changed to Revised IP-OLDF. ;
  MIN=@SUM(N(i):E(i));
  @FOR(N(i):@SUM(P1(J1):IS(i,J1) * VARK(J1) * CHOICE(k,J1)) > 1-BIGM
* E(i));
  @FOR(P1(J1):@FREE(VARK(J1))); ! @FOR(N(I):@BIN(E(I)));
ENDSUBMODEL
SUBMODEL IPLP: !Define Revised IPLP-OLDF. ;
  MIN=@SUM(N(i):E(i));
  @FOR(N(i):@SUM(P1(J1):IS(i,J1)* VARK(J1)* CHOICE(k,J1) )>1-BIGM *
E(i));
  @FOR(P1(J1): @FREE(VARK(J1)));
  @FOR(N(I)|CONSTANT(i)#NE#0:@BIN(E(I)));
  @FOR(N(I)|CONSTANT(i)#EQ#0:E(I)=0);
ENDSUBMODEL
CALC: ! Calc section controls the complex and continuous optimizing models. ;
  @SET('DEFAULT');
  K=1;Lend=@SIZE(MS);
  @WHILE( K#LE#Lend:f=1; !15 different models;
  @WHILE( f#LE#100: !100-fold cross validation ;
  @FOR(D(i,j): IS(i,j)=ES( @SIZE(N) * (f-1)+i, j) );
  @SOLVE(LP); !1500 iterations for Revised LP-OLDF;
  @FOR(N(i):
  @IFC( E(I)#EQ#0: CONSTANT(i)=0; @ELSE CONSTANT(i)=1;));
  NM1=0; NM2=0; ! NMs of the training and the validation samples are stored
  in NM1 and NM2;
  @SOLVE(IPLP); ! The second step of Revised IPLP-OLDF. ;
  @FOR(P1(j):VARK100(100*(k-1)+f,j)=VARK(j)* CHOICE(k,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)=@SUM(P1(j):ES(NN,j)* VARK(j)*
        CHOICE(k,j)));
   @FOR(n(l):@IFC(SCORE(l)#LT#0: NM1=NM1+1));
   @FOR(n2(nn):@IFC(SCORE2(nn)#LT#0:NM2=NM2+1)); IC(K,f)=NM1;
        EC(k,f)=NM2;
   f=f+1); K=K+1);
ENDCALC
DATA: ! Output the coefficients, NMs. ;
   @OLE( )=VARK100, IC, EC;
ENDDATA
END
```

Acknowledgements

My research starts from 1997. I do not know the long history about discriminant analysis using MP [21]. Prof. Linus Schrage sends me lists of researches in this area. He and Kevin Cunnigham, Vice President of LINDO Systems Inc., help me about my models using LINGO and What's Best! [9, 10]. In statistics, I introduce SAS into Japan in1979, and write many books and papers about SAS and JMP [7]. Especially, when I stay at IIASA located in Wien in 2013, I supervise the Japanese version of "JMP Start Statistics [8]". It is helpful for me to achieve my research after 2004. JMP division of SAS Institute Japan helps me to write the code of 100- fold cross validation of Fisher's LDF and logistic regression by JMP. I can compare Fisher's LDF and logistic regression with Revised IP-OLDF, Revised IPLP-OLDF, Revised LP-OLDF and S- SVM [25] coded by LINGO Ver.10. Linus suggests me to use the latest LINGO Ver.14.

I cannot get official research fund, because many researchers think that MNM is foolish criterion in the statistical research. I owe it by my University fund and my father's heritage. Thank you for late father. My research has begun from the doctoral thesis. Thank you for Prof. T. Tarumi, Y. Tanaka and A. Miyake [11].

REFERENCES

1. Cheney, E. W. (1966). Introduction to Approximation Theory. New York. McGraw-Hill.
2. Edgar, A. (1935). The irises of the Gaspe Peninsula. Bulletin of the American Iris Society, **59**, 2 - 5.
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. Goodnight, J. H. (1978). SAS Technical Report – The Sweep Operator: It's Importance in Statistical Computing – (R100). SAS Institute Inc. (Japanese version is translated by Shinmura.)
6. Liitschwager, J. M. & Wang, C. (1978). Integer programming solution of a classification problem. Management Science, **24/**14, 1515 - 1525.
7. Sall, J. P. (1981). SAS Regression Applications. SAS Institute Inc. (Japanese version is translated by Shinmura.)
8. Sall, J. P., Creighton, L. & Lehman, A. (2004). JMP Start Statistics, Third Edition. SAS Institute Inc. (Japanese version is edited by Shinmura.)
9. Schrage, L. (1991). LINDO Optimization Modeling System. The Scientific Press. (Japanese version is translated by Shinmura and Takamori.)
10. Schrage, L. (2006). Optimization Modeling with LINGO. LINDO Systems Inc.
11. Shinmura, S. & Miyake, A. (1979). Optimal linear discriminant functions and their application. COMPSAC79, 167 - 172.
12. Shinmura, S. (1998). Optimal Linear Discrimrnant Functions using Mathematical Programming. Journal of the Japanese Society of Computer Statistics, **11 /** 2, 89 - 101.
13. Shinmura, S. (2000). A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics, **5**, 133 - 142.
14. Shinmura, S. (2004). New Algorithm of Discriminant Analysis using Integer Programming. IPSI 2004 Pescara VIP Conference CDROM, 1 - 18.
15. Shinmura, S. (2007). Overviews of Discriminant Function by Mathematical Programming. Journal of the Japanese Society of Computer Statistics, **20/**12, 59 - 94.
16. Shinmura, S. (2010). The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing (in Japanese).
17. Shinmura, S. (2011). Problems of Discriminant Analysis by Mark Sense Test Data.Japanese Society of Applied Statistics, **40/**3157 - 172
18. Shinmura, S. (2011). Beyond Fisher's Linear Discriminant Analysis-New World of Discriminant Analysis-. 2011 ISI CD ROM, 1 - 6.
19. Shinmura, S. (2013). Evaluation of Optimal Linear Discriminant Function by 100-fold cross validation. 2013 ISI CD ROM, 1 - 6.
20. Shinmura, S. (2014). End of Discriminant Functions based on Variance Covariance Matrices. ICORES2014 (International Conference on Operations Research and Enterprise Systems 2014), 5 - 16.
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. Miyake, A. & Shinmura, S. (1976). Error rate of linear discriminant function, F.T. de Dombal & F. Gremy editors 435 - 445, North Holland Publishing Company.
25. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.