Matroska Feature Selection Method for Microarray Data

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

Abstract— **We propose a Matroska feature selection method (Method 2) for microarray datasets (the datasets). We had already established a new theory of the discriminant analysis (Theory) and developed an optimal Linear Discriminant Function (OLDF) named Revised IP-OLDF. This LDF can naturally select features for the datasets. The dataset consists of several small genes subspaces that we call small Matroskas (SMs) and are linearly separable. We confirmed this feature selection of Revised IP-OLDF by Swiss banknote data and Japanese automobile data, also. Therefore, we need not struggle with high-dimension genes space. In this paper, we develop a LINGO program to find all SMs and confirm that the dataset consists of disjoint union of SMs and high-dimension subspace that is not linearly separable. Because it is very easy for us to analyze these SMs that are small samples, we may be able to find new facts of gene analysis. Lasso researchers will have better results compared with our results.**

Keywords- Minimum Number of Misclassifications (MNM); Revised IP-OLDF; SVM; Fisher's LDF; Gene Analysis; Small Matroska (SM); Basic Gene Subspase (BGS); Lasso.

I. INTRODCTION

Fisher [6] [7] developed a Linear Discriminant Function (Fisher's LDF) under Fisher's assumption and established the theory of discriminant analysis. Because Fisher's assumption was too strict for the real data, a Quadratic Discriminant Function (QDF) was developed. In addition to two discriminant functions, logistic regression [4] and a Regulalized Discriminant Analysis (RDA) [9] were proposed as the statistical discriminant functions. These statistical discriminant functions apply for many applications, and statistical software packages became essential tools for the science and industries. On the other hand, it is well known that Mathematical Programming (MP) can define the discriminant models [16]. Linear Programming (LP) sets out Least Absolute Deviation (LAD) discriminant function. Quadratic Programming (QP) defines an L2-norm discriminant function (Least square method). Nonlinear Programming (NLP) defines Lp-norm discriminant functions. Before 1997, there were many papers of MP-based discriminant functions summarized by Stam [57]. We think the first generation research ended in 1997 because these researches lacked examination of real data and comparison with statistical discriminant functions. Vapnik [61] proposed three Support Vector Machines (SVMs) such as a Hard-margin SVM (H-SVM), Soft-margin SVM (S-SVM) and kernel SVM in 1995. H-SVM clearly defined a Linearly Separable Data (LSD) and generalization ability. However, because most real data are not LSD, and H-SVM can be used only for LSD, we use S-SVM for actual data. QP defines these SVMs. Although kernel SVM is one of nonlinear discriminant function and provides an attractive idea, we do not discuss it in this research because our concern is a comparison of LDFs. Many researchers use SVMs because there are many examinations of real data compared with the first generation research of MP-based discriminant theory. From 1971 to 1974, we became a member of the project to develop a computer system for an Electrocardiogram (ECG) data. Project leader, Doc. Nomura gave us a theme to develop a diagnostic logic using Fisher's LDF. Our research was inferior to Nomura's experimental decision tree algorithm. At first, we thought this failure was caused by our poor experience and knowledge of statistics. However, we considered the discriminant functions based on the variance-covariance matrices were not suitable for the medical diagnosis discussed in Section Ⅱ. Moreover, we found all LDFs cannot correctly discriminate the cases on the discriminant hyperplane (Problem 1).

In Section Ⅱ, although Fisher established discriminant analysis based on variance-covariance matrices, we explain a new theory of MP-based discriminant analysis (Theory) [53]. At first, we developed an Optimal LDF based on a Minimum Number of Misclassifications (minimum NM, MNM) criterion (IP-OLDF) in (1) [19] - [21]. It reveals two important facts of discriminant analysis. Those are 1) the relation of NM and LDF in the discriminant coefficient space, 2) monotonic decrease of MNM that is very crucial for gene analysis. It shows the good result by comparison with Fisher's LDF and QDF using Fisher's iris data [2] and Cephalo Pelvic Disproportion (CPD) data [14]. It finds Swiss banknote data is LSD [8]. All LDFs except for H-SVM and Revised IP-OLDF in (2) cannot discriminate LSD theoretically (Problem 2). Experimentally, Revised LP-OLDF in (2), one of L1-norm LDF using LP, can discriminate LSD. Nevertheless, it tends to gather cases on the discriminant hyperplane (Problem 1). Student data [24] reveals the defect of IP-OLDF caused by Problem 1. Therefore, Revised IP-OLDF is developed. It is only LDF to solve Problem 1. The pass/fail determination using exam scores [28] shows the defect of QDF and RDA caused by the defect of generalized inverse of variance-covariance matrices (Problem 3). If we add random noise to constant values of some particular variable, we can solve Problem 3. Japanese automobile data [35] explain Problem 3, also. Because Fisher never formulate the equation of Standard Error (SE) of error rate and discriminant coefficient, discriminant analysis is not traditional inferential statistics based on normal distribution

 ISBN: 978-1-61208-488-6

(Problem 4). A 100-fold cross-validation for small sample method (Method 1) offers the 95% Confidence Interval (CI) of error rate and discriminant coefficient [23] [25] - [27]. Moreover, because the best model with minimum mean of error rate in the validation samples is powerful model selection method and we can explain the meaning of discriminant coefficient [51][52], we understand to establish Theory. However, we know many researchers have been struggling in the gene analysis for more than ten years (Problem 5) [12].

In Section Ⅲ, we propose a Matroska feature selection method for gene analysis (Method 2). When we discriminate six microarray datasets (the datasets) [12], our three OLDFs can naturally select features [37] - [44]. However, three SVMs cannot select features. Moreover, Fisher's LDF cannot discriminate six datasets correctly because six NMs are not zero. In [42] [43] we explained in detail the results of Fisher's LDF. Revised IP-OLDF by Method 2 reveals the dataset consists of disjoint union of small linearly separable subspaces (SMs) and high-dimensional subspace that is not linearly separable (MNM >=1). This perception is essential for gene analysis.

In Section Ⅳ, we explain how to analyze each SM and find a Basic Gene Subspace (BGS) in each SM by ordinary statistical methods. We can analyze each SM very easy because all SMs are small samples. Moreover, we can understand the structure of dataset by BGSs because of monotonic decrease of MNM.

II. NEW THEORY OF DISCRIMINANT ANALYSIS

We develop four OLDF including IP-OLDF that find two new facts and solve four problems. Moreover, we confirm the best models of Revised IP-OLDF are better than other seven LDFs by six ordinary data introduced in Section Ⅰ.

A. Four Problems of Discriminant Analysis

There are four problems with the discriminant analysis [31][35] [36].

Problem 1: The discriminant rule is very simple. Let $f(\mathbf{x})$ be LDF and $y_i*f(\mathbf{x}_i)$ be a discriminant score for $\mathbf{x}_i$. 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. We cannot properly discriminate $\mathbf{x}_i$ on the discriminant hyperplane ($f(\mathbf{x}_i) = 0$). Many researchers ignore this unresolved problem until now. They consider a discriminant rule as follows: If $f(\mathbf{x}_i) >= 0$, $\mathbf{x}_i$ is classified to class 1 correctly. Otherwise, if $f(\mathbf{x}_i) < 0$, $\mathbf{x}_i$ is classified to class 2 correctly. Their discriminant rule is not logical. Only Revised IP-OLDF can treat this problem appropriately. Indeed, except for Revised IP-OLDF, no LDFs can count the NMs correctly. These LDFs should count the number of cases where $f(\mathbf{x}_i) = 0$, and display this figure alongside the NM in the output. Student data tells us the defect of IP-OLDF. Therefore, we develop Revised IP-OLDF.

Problem 2: Only H-SVM and Revised IP-OLDF can recognize LSD theoretically. Experimentally, Revised LP-OLDF discriminates LSD correctly. Nevertheless, it tends to collect cases on the discriminant hyperplane (Problem 1). If we discriminate exam scores by four testlets score, and the pass mark is 50 point, we can obtain a trivial LDF such as f = T1 + T2+ T3+ T4 -50 [36]. We can judge the student pass the exam if $f(\mathbf{x}_i) >= 0$ and fail the exam if $f(\mathbf{x}_i) < 0$. However, error rates of Fisher's LDF and QDF are very high [35] because exam scores do not satisfy Fisher's assumption. Therefore, these LDFs should not be used in important applications such as medical diagnosis, pattern recognition, and rating.

Problem 3: Problem 3 is the defect of generalized inverse. When we discriminated math exam scores by QDF and RDA, all pass students were misclassified in the failed class because all pass students answered some item scores correctly, and scores of failed student vary. In this case, if we add random noise to the constant values, we can solve this problem.

Problem 4: Fisher never formulated the equation of SE of discriminant coefficients and error rates based on the normal distribution. Because there is no model selection procedure instead of a leave-one-out (LOO) procedure [13], we propose Method 1. It offers the 95% CI of error rates and discriminant coefficients. Moreover, it offers simple and powerful model selection procedure such as the best model with a minimum mean of error rate in the validation samples. We confirmed the best models of Revised IP-OLDF were better than Fisher's LDF, logistic regression and five MP-based LDFs using six ordinary data [29] [30] [33] [34]. Fisher's LDF and logistic regression discriminate these data by JMP script [15]. JMP division of SAS Institute Inc. Japan supports us to develop it. Six MP-based LDFs are Revised IP-OLDF, Revised LP-OLDF, Revised IPLP-OLDF, H-SVM and two S-SVMs such as SVM4 (penalty c = 10000) and SVM1 (penalty c = 1) by LINGO program that is supported by LINDO Systems Inc [17]. We can establish Theory by JMP and LINGO.

B. MP-based LDFs

Although we developed a diagnostic logic of ECG data by Fisher's LDF, our research was inferior to the decision tree logic developed by the medical doctor. After this experience, we concluded it is not adequate for the discrimination of the normal and abnormal diseases because of two main reasons [18].

1) There are many cases nearby the discriminant hyperplane. Medical doctors are striving to discriminate the cases nearby the discriminant hyperplane.

2) If the value of some variable increases or decreases, the probability belonging to abnormal disease increases from 0 to 1. Fisher's LDF assumes the typical abnormal patients are the average of the abnormal classes. However, the typical patients are far from the normal patients. Taguchi et al. [58] method was one of multi-class discrimination by Mahalanobis-distance based on the variance-covariance matrices. The authors claim that the cases belonging to abnormal states are far from the normal state. Their claim is the same perception as our claim. If some independent variable of logistic regression increases or decreases, the

probability 'p' belonging to class1 (abnormal symptom) increases from 0 (class2) to 1 (class1). Therefore, most medical users use logistic regression instead of Fisher's LDF. However, because JMP does not support logistic regression for the datasets, we never discuss logistic regression in Section III.

After many experiences of the discriminant analysis [14] [22], we developed IP-OLDF in (1). Because we fix the intercept of IP-OLDF to one, it is in the p-dimensional coefficient space. Although $y_i*({}^t\mathbf{x}_i\mathbf{b}+1)$ is discriminant scores, $y_i*({}^t\mathbf{x}_i\mathbf{b}+1) = 0$ is a linear hyperplane and divides discriminant space to two half planes such as plus half plane ($y_i*({}^t\mathbf{x}_i\mathbf{b}+1) > 0$) and minus half plane ($y_i*({}^t\mathbf{x}_i\mathbf{b}+1) < 0$). If we choose $\mathbf{b}_k$ in plus hyperplane as LDF, LDF such as $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1)$ discriminate $\mathbf{x}_i$ correctly because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) > 0$. On the other hand, if we choose $\mathbf{b}_k$ in minus hyperplane, LDF misclassify $\mathbf{x}_i$ because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) < 0$. However, we must solve other two models such as the intercept = -1 and 0. It looks for the right vertex of an Optimal Convex Polyhedron (optimal CP, OCP) if data is a general position. There are only p-cases on the discriminant hyperplane, and it becomes the vertex of OCP. On the other hand, if data is not general position, it may not look for the correct vertex of OCP because there are over (p+1) cases on the discriminant hyperplane, and we cannot correctly discriminate these cases. Therefore, we developed Revised IP-OLDF that looks for the interior point of true OCP in (2) directly. Because b_0 is free variable, it is defined in (p+1)-dimensional coefficient space. If it discriminates $\mathbf{x}_i$ correctly, $e_i = 0$ and $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) >= 1$. If it cannot discriminate $\mathbf{x}_i$ correctly, $e_i = 1$ and $y_i* ({}^t\mathbf{x}_i\mathbf{b}+b_0) >= -9999$. Although support vector (SV) for classified cases are $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) = 1$, SV for misclassified cases are $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) = -9999$. Therefore, we expect a discriminant score of misclassified cases are less than -1, and there are no cases within two SVs. Therefore, if M is small constant, it does not work correctly [27]. Because there are no cases on the discriminant hyperplane, we can understand the optimal solution is an interior point of OCP defined by IP-OLDF. All LDFs except for Revised IP-OLDF cannot solve Problem 1 theoretically. Therefore, these LDFs must check the number of cases (h) on the discriminant hyperplane. Correct NM may increase (NM + h).

$$\text{MIN} = \Sigma e_i;\ y_i*({}^t\mathbf{x}_i\mathbf{b} + 1) >= - e_i ; \quad (1)$$

e_i: 0/1 integer variable corresponding to classified/misclassified cases.
y_i: 1/-1 for class1/class2 or object variable.
$\mathbf{x}_i$: p-independent variables.
$\mathbf{b}$: discriminant coefficients.

Because we can consider IP-OLDF in (1) on the data and discriminant coefficients spaces, we find two relevant facts as follows.

1) We explain the notation of IP-OLDF by the Golub et al. dataset [10]. It consists of two classes such as "All (47 cases)" and "AML (25 cases)" with 7,129 genes. Our primary concern is to discriminate two classes by 7,129 variables (genes). The 72 linear hyperplane, the 7,129 coefficients of those are values of each case, divide the discriminant coefficient space into finite CP. The interior points of each CP correspond to the discriminant coefficient of LDF that discriminates the same cases correctly and misclassifies another same case. Therefore, because the interior points of each CP have unique NM, we can define the OCP with MNM. Many examinations show the best models of Revised IP-OLDF are better than other seven LDFs.

2) Let MNM_k be MNM in the k-dimensional subspace. MNM decreases monotonously ($MNM_k >= MNM_{(k+1)}$). If $MNM_k = 0$, all MNMs including these k-variables (genes) are zero. This fact tells us the smallest Matroska (Basic Gene Subspace, BGS) can completely describe the structure of gene space by monotonic decreases of MNM.

When we discriminate Swiss banknote data with six variables, IP-OLDF finds two-variables models, such as (X4, X6), is linearly separable. By the monotonic decrease of MNN, 16 MNMs including these two variables are zero among 63 models ($= 2^6-1 = 63$). Other 47 MNMs are greater than one. Revised IP-OLDF in (2) can naturally select features for ordinary data and six datasets. However, we develop more powerful model selection procedure such as the best model by Method 1. Therefore, we had ignored the natural feature selection for ordinary data before Method 2.

$$\text{MIN} = \Sigma e_i ;\ \ y_i* ({}^t\mathbf{x}_i\mathbf{b} + b_0) >= 1 - M* e_i ; \quad (2)$$

b_0: free decision variables.
M: 10,000 (Big M constant).

If e_i is non-negative real variable, equation (2) changes Revised LP-OLDF. Revised IPLP-OLDF [32] is a mixture model of Revised LP-OLDF in the first phase and Revised IP-OLDF in the second phase. The equation (3) is S-SVM. If we set $c=10^4$ or c=1, it becomes SVM4 or SVM1. If we omit "$c* \Sigma e_i$" and "$- e_i$", it becomes H-SVM. QP solves both SVMs.

$$\text{MIN} = \|\mathbf{b}\|^2/2 + c*\Sigma e_i ;\ \ y_i* ({}^t\mathbf{x}_i\mathbf{b} + b_0) >= 1 - e_i ; \quad (3)$$

c: penalty c to combine two objectives.
e_i: non-negative real value.

C. New Theory of Discriminant Analysis (Theory)

We explain the outlook of Theory. There are four serious problems with the discriminant analysis. We developed four MP-based OLDFs. IP-OLDF finds two new facts of discriminant analysis. Revised IP-OLDF solves Problem 1 and Problem 2 related to this paper. Because Method 1 solves Problem 4 and four problems are solved completely, we misunderstand to establish Theory. In 2015, when we discriminated six datasets by MP-based LDFs and Fisher's LDF, only Revised IP-OLDF could naturally select features because coefficients less than 173 are not zero and other coefficients become zeroes [37]. After we recognize this fifth problem, we completely solve Problem 5 by Method 2 in Dec.

2015. Although we had observed the feature selection of Revised IP-OLDF by Swiss banknote data and Japanese automobile data that are LSD, we ignore this fact because the best model is an excellent model selection procedure for ordinary six data. In gene analysis, if we call all linearly separable models as Matroskas that are linearly separable gene subspaces, Revised IP-OLDF reduces the high-dimension gene space, the big Matroska, to small subspace (SM) drastically. After we remove genes in the first SM1 from the big Matroska, Revised IP-OLDF discriminates the new gene space (the second big Matroska), again. It can find the second different SM2. We repeat this process and locate the dataset that consists of the disjoint union of SMs and high-dimension gene subspace (MNM>=1). Therefore, we develop Method 2. We make a program of Method 2 by LINGO and can list up all SMs of the six datasets very easy. Although many researchers have been struggling to analyze the high-dimension gene datasets by a statistical approach [55] [60], we can analyze each SM very easy because it is a small sample. In Section Ⅳ, we show how to find BGSs by manual operation and analyze one of each SM by the ordinary statistical approach. In Section Ⅴ, we discuss the use and application of our results.

D. Short Story of Feature Selection

At the end of October 2015, we presented our Theory at Japanese statistical conference and knew six datasets presented by another researcher presentation. After the conference, we discriminated six datasets by seven LDFs. Because error rates of Fisher's LDF were very high for eighteen exam scores [35], it is self-evident we cannot obtain better results in the gene datasets. Therefore, users never use it for gene analysis. Although NMs of three SVMs are zero, all coefficients are not zero. Therefore, three SVMs are not helpful for the feature selection. Several coefficients of Revised IP-OLDF are not zero, and most of the coefficients are zero. It can naturally select features of the datasets within few seconds and reduce high-dimension genes spaces to the smaller subspace that is one of the Matroska. Next, when we discriminate the Matroska again, we can find smaller Matroska. Therefore, the dataset has the structure of Matroska. When we cannot locate the smaller Matroska again, we call the last subspace as the Small Matroska (SM1). Moreover, after we exclude the first SM1 from the dataset, we find the second different SM2. At last, we can list up all SMs by a LINGO program of Method 2 and conclude the dataset consists of the disjoint union of SMs and another high-dimension gene subspace that is not linearly separable. Six studies [45] - [50] include full genes lists of the SMs about six datasets. If we analyze all SMs, we may be able to obtain new facts of gene analysis. Although some researchers try to discriminate the dataset by LASSO based on variance-covariance matrices, our Theory showed only H-SVM and Revised IP-OLDF can discriminate LSD theoretically, and revealed the structure of datasets. If LASSO researchers compare their results with our results using our two ordinary data and six datasets, it is expected to improve the research of feature selection method more deeply.

III. MATROSKA FEATURE SELECTION METHOD

In this section, we introduce Method 2.

A. Outlook of Method 2

When we discriminate Shipp et al. data [54] on Oct. 28, 2015, only Revised IP-OLDF can select thirty-two genes among 7129 genes [37]. Although we misunderstand the discrimination having 7129 variables requests huge CPU time, Fisher's LDF by JMP ver.12 (JMP12) and other MP-based LDFs coded by LINGO can solve the datasets less than 20 seconds because the datasets are LSD. However, most coefficients of these LDFs except for Revised IP-OLDF are not zero. Therefore, these LDFs are not helpful for feature selection for gene analysis in addition to ordinary data. In this research, we call the smallest Matroska as the BGS with k-variables. The biggest Matroska with 7129 variables includes many smaller Matroskas from 7128 (= 7129 - 1) variables to k variables. LINGO program found the datasets are the disjoint union of SMs with h-variables (p > h >= k) and another high-dimension gene subspace with "MNM >= 1." Now, we must survey the BGSs from SM by manual operation. If Revised LINGO program can find all list of BGSs, we can understand the structure of the dataset by these BGSs completely. Because we can analyze each SM using ordinary statistical methods, we expect to obtain new facts of gene analysis and hope many researchers try to analyze these SMs. By our breakthrough, the feature selection becomes exciting theme.

We guess the reason why Revised IP-OLDF can naturally select features as follows.

1) MNM criterion works well for the feature selection. This expectation will be right if LASSO cannot list up all SMs or BGSs correctly as same as Revised IP-OLDF because it does not use MNM criterion. We consider the discrimination of LSD requests MNM criterion or maximization of two SVs.

2) The algorithm of LINGO IP solve uses the branch and bound. We believe Revised IP-OLDF coded by another IP algorithm cannot naturally select features. On the other hand, we cannot control the flow of the branch and bound. When IP solver finds the model with MNM=0 at first, LINGO program output it and end. This treatment is the reason why LINGO program may not be able to find BGS directly. This research is our future theme.

B. Results of Six Microarray Data

Table Ⅰ shows the summary of six datasets. Rows "Description" show two classes. Rows "Size" are the case number by the gene number. Rows "SM: Gene" are the number of SM [with reference number]: the total number of genes including in all SMs. Six lists of full gene name are in the references. Rows "Min, Mean, Max" are the minimum, mean and maximum values of genes including in all SMs. Rows "JMP12" are 2 by 2 tables of the discrimination by Fisher's LDF. Six NMs are 5, 3, 8, 3, 10 and 29. Rows "% and error rate" are the percentages of (Maximum value/case number) and error rates of JMP12. Maximum percent is 63%

by Alon et al. dataset. Minimum percent is 43% by Golub et al. dataset. Maximum error rate is 17% by Tian et al. dataset and minimum error rate is 1% by Chiaretti et al. dataset

TABLE I. SUMMARY OF SIX MICROARRAY DATASETS

Data	Alone et al. [1]	Chiaretti et al. [2]
Description	Normal (22) vs. tumor cancer (40)	B-cell (95) vs. T-cell (33)
Size	62 *2000	128*12625
SM: Gene	64 [47]:1152	270 [50]:5385
Min/Mean/Max	11/18/39	9/19/62
JMP Ver.12	20:2/3:37	94:1/2:31
% and error rate	63%, 8%	49%, 1%
Data	Golub et al. [10]	Shipp et al. [54]
Description	All (47) vs. AML (25)	Follicular lymphoma (19) vs. DLBCL (58)
Size	72*7129	77 *7130
SM: Gene	69 [46]:1238	213 [45]:3032
Min/Mean/Max	10/18/31	7/14/43
JMP12	20:5/3:44	17:2/1:57
% and error rate	43%, 11%	56%, 4%
Data	Singh et al. [56]	Tian et al. [59]
Description	Normal (50) vs. tumor prostate (50)	False (36) vs. True (137)
Size	102 *12626	173 *12625
SM: Gene	179 [48]:3990	159 [49]:7221
Min/Mean/Max	13/22/47	28/45/104
JMP Ver.12	46:4/6:46	16:20/9:128
% and error rate	46%, 10%	60%, 17%

C. Detail of the Matroska Feature Selection Method

We explain Method 2 briefly. Table Ⅱ is the output of Golub et al. dataset by LINGO program. Two columns "LOOP1 and LOOP2" are the sequence number of big and small loops of Method 2. Revised IP-OLDF discriminate the dataset with 7129 genes in the LOOP1=1 and LOOP2=1, and only 34 coefficients of Revised IP-OLDF are not zero. In general, this number is less than the case number such as 72. In the second small loop (LOOP1=1, LOOP2=2), we discriminate the smaller Matroska with 34 genes again, and only 11 coefficients are not zero. Therefore, we get the Matroska sequence such as Matroska7129 → Matroska34 → Matroska11. We stop at LOOP2=4 because we cannot find the smaller Matroska. We call Matroska11 as the SM1 because Revised IP-OLDF cannot locate the smaller Matroska. We exclude the first SM1 with 11 genes from the big Matroska with 7129 genes and make the second big Matroska with 7118 genes. In the second big loop at LOOP1 = 2, we get the second SM2 with 16 genes.

TABLE II. THE OUTLOOK OF THE THEORY 2

SN	LOOP1	LOOP2	Gene	MNM
1	1	1	7129	0
2	1	2	34	0
3	1	3	11	0
4	1	4	11	0
16	2	1	7118	0
17	2	2	36	0
18	2	3	18	0
19	2	4	16	0
20	2	5	16	0

After LINGO program finds sixty-nine SMs in Table Ⅲ, it stops the big loop when we find MNM is greater than one at LOOP1=70. However, we can continue this loop until it cannot naturally select features and list up all small subspaces with "MNM >= 1." Therefore, Method 2 can discriminate other types of genes datasets that are not LSD. Because Golub et al. dataset consists 69 SMs that are linearly separable models or subspaces, it is very easy for us to analyze all SMs because the 68th and 69th SMs are the biggest samples with 72 cases by 31 genes.

TABLE III. ALL SMALL MATROSKA OF GOLUB ET AL. DATA

LOOP1	LOOP2	Gene	n	MNM	35	11	6630	17	0
1	11	7129	11	0	36	11	6613	19	0
2	11	7118	16	0	37	11	6594	12	0
3	11	7102	11	0	38	11	6582	16	0
-	-	-	-	-	-	-	-	-	-
32	11	6683	19	0	67	11	5976	23	0
33	11	6664	16	0	68	11	5953	31	0
34	11	6648	18	0	69	11	5922	31	0

IV. BGS AND STATISTICAL ANALYSIS

In this section, we introduce how to find BGS and analyze it.

A. How to find BGSs

Because we cannot control the flow of branch and bound algorithm, there may be several BGSs in the SM. We propose how to find BGSs by manual operation as follows:

1) To find the smaller linear separable model in SM

We analyze the first SM1 with 11 genes by the forward stepwise procedure and obtain the five columns from 'Step' to 'BIC' in Table Ⅳ. The last column is NM of logistic

regression. Although there is no theoretical guarantee that logistic regression can discriminate LSD correctly [5], we judge it discriminates LSD correctly under the condition of "MNM=0 and NM=0". Therefore, we can judge BGS exists among all combination models in four genes subspace [11]. We know the four-variable model is linearly separable. Cp, AIC and BIC recommend this model. Usually, these three statistics recommend the different models by our many trials.

TABLE IV. FORWARD STEPWISE AND LOGISTIC REGRESSION.

Step	Gene	Cp	AIC	BIC	logistic
1	M11722_at	72.56	137.78	144.26	5
2	X59871_at	38.42	118.62	127.13	2
3	U05259_rna1_at	9.92	96.07	106.54	2
4	D21063_at	3.88	90.15	102.52	0
5	M22919_rna2_at	3.80	90.30	104.49	0
6	M21624_at	4.27	91.09	107.02	0
7	M25280_at	4.63	91.79	109.38	0
8	L13210_at	6.15	93.93	113.09	0
9	X82240_rna1_at	8.02	96.56	117.21	0
10	HG3039-HT3200_at	10.01	99.44	121.47	0
11	L76159_at	12.00	102.41	125.73	0

TABLE V. FIFTEEN MODEL BY FOUR GENES

p	X1	X2	X3	X4	c	MNM	ZERO
4	1	1	1	1	1	0	0
3	1	1	1	0	1	1	0
3	1	1	0	1	1	1	0
3	1	0	1	1	1	3	0
3	0	1	1	1	1	2	0
2	1	1	0	0	1	2	0
2	1	0	1	0	1	4	0
2	0	1	1	0	1	3	0
2	1	0	0	1	1	4	0
2	0	1	0	1	1	13	0
2	0	0	1	1	1	6	0
1	1	0	0	0	1	5	0
1	0	1	0	0	1	25	0
1	0	0	1	0	1	10	0
1	0	0	0	1	1	17	0

2) Search BGSs by all possible combination models

We search BGSs by all possible combination models using Revised IP-OLDF. Table V is the 15 models by four genes that are four combinations of 0/1 values from the second column to the fifth column. Column "c" is the intercept of Revised IP-OLDF. The column "p" is the number of independent variables from four-variable model (p=4) to four one-variable models (p=1). The binary values, such as 1/0, mean each model include or not include four variable in the model. Column "MNM" is MNM of 15 models. Column "ZERO" is the number of cases on the discriminant hyperplane. Only full model is linearly separable. Therefore, we find one BGS in the first SM, such as (X1: M11722_at, X2: X59871_at, X3: U05259_rna1_at, X4: D21063_at). All MNMs including these four genes are linearly separable in Golib et al. dataset. Therefore, although there are numerous Matroskas in the dataset, we can understand the structure of Matroska by BGS because of the monotonic decrease of MNM. The big Matroska with 7129 genes includes numerous smaller Matroska from 7128 genes to four genes. Although there are 7129 subspaces with 7128 genes, there are 7125 smaller Matroska with 7128 genes and four subspaces with 7128 genes that are not Matroska. By monotonic decrease of MNM, we can completely understand the structure of Matroska. It is hard for us to analyze the dataset by the ordinary statistical methods without knowledge of this fact.

B. How to analyze each SM

Figure 1 is the output of principal component analysis (PCA). Left figure is the eigenvalues. Two eigenvalues are greater than one and contribution ratio is about 0.75. The middle figure is the scatter plot. The symbol + are "AMLs" that are in the third quadrant. Forty-seven cases of "ALL" are situated in the fourth, first and second quadrant. The right plot is the factor loading plot. "M11722_at" is overlapped on the first component and "X59871_at" is overlap on the second component. It is very important for specialists of gene analysis to consider the reason why two groups are orthogonal. We expect specialists of gene analysis to examine the meaning of statistical outputs of SMs.

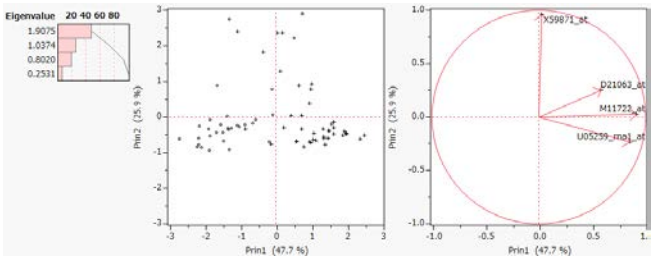
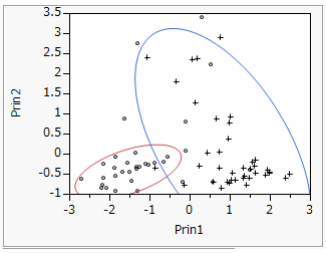


Figure 1. The principal component analysis.

Figure 2 is two score plots. X-axis is the first component. Y-axis of left and right score plots correspond the second component and the third component. Because PCA cannot separate two classes, ordinary statistical analysis such as one-way ANOVA, cluster analysis, and PCA cannot conclude clear results for the datasets directly. Jeffery et al. compared the efficiency of the ten feature selection methods using conventional statistical approaches. It tells us the limitation of conventional statistical methods.



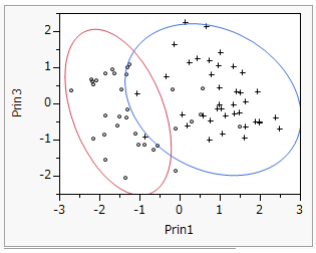


Figure 2. Two score plots.

Table VI is the correlation matrix. The absolute correlations of "X59871_at" with other three genes are less than 0.088 that are the same result as the factor loading plot.

TABLE VI. CORRELATION MATRIX.

Var.	X1	X2	X3	X4
M11722_at	1	0.076	0.713	0.371
X59871_at	0.076	1	-0.088	0.052
U05259_rna1_at	0.713	-0.088	1	0.220
D21063_at	0.371	0.052	0.220	1

V. CONCLUSION

We developed Theory, Method 1 and Method 2. Revised IP-OLDF solves Problem 1, Problem 2 and Problem 5. Moreover, the best models of Revised IP-OLD are better than another seven LDFs. Although H-SVM discriminate LSD correctly, it cannot naturally select features for six datasets. Because Problem 3 is the defect of the generalized inverse and error rates of Fisher's LDF and QDF are very high for LSD, we guess the discriminant analysis and regression analysis based on variance-covariance matrices may not be helpful for gene analysis. Although the discriminant analysis is not the traditionaly inferential statistical method, Method 1 offers the 95% CI of error rate and discriminant coefficient and the validation of Revised IP-OLDF by six ordinary data. In this paper, we do not discuss the validation of six microarray datasets. However, because Method 1 validated already six ordinary data, we will validate the results of six microarray datasets in near future. Because the best model is powerful model selection procedure for ordinary data, we ignore some parameters of Revised IP-OLDF are zeroes in ordinary data. Because other LDFs cannot naturally select features, they may be difficult for gene datasets. If we can develop Revised LINGO program that can find all BGSs, it will be more useful in gene analysis. LINGO program is useful for other gene dataset, such as RNA-Seq., in addition to the six datasets. Although we surveyed to clarify the long-term survivors of the Maruyama vaccine (SSM) administration patients, our trial failed [22]. If we compare two lists of cancer genes, (normal and cancer patient data) vs. (normal and SSM Administration patient data), and find the differences between two gene lists, it may show the proof of the effectiveness of SSM. Now, we plan this new theme and have proposed a joint research with the inspection agency of microarray in Japan.

We would like to propose a joint research with medical doctors in the world.

REFERENCES

[1] A. Alon et al., "Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. Natl. Acad. Sci. USA, 96, pp. 6745-6750, 1999.

[2] E. Anderson, "The irises of the Gaspe Peninsula," Bulletin of the American Iris Society vol. 59, pp. 2-5, 1945.

[3] S. Chiaretti et al., "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," Blood. April 1, 2004, 103/7, pp. 2771-2778, 2004.

[4] D. R. Cox, "The regression analysis of binary sequences (with discussion)," J Roy Stat Soc B 20, pp. 215-242, 1958.

[5] D. Firth, "Bias reduction of maximum likelihood estimates," Biometrika, vol. 80, pp. 27-39, 1993.

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

[7] R. A. Fisher, Statistical methods and statistical inference. Hafner Publishing Co. 1956.

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

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

[10] T. R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science. 1999 Oct 15, 286/5439, pp. 531-537, 1999.

[11] J. H. Goodnight, SAS Technical Report – The Sweep Operator: Its Importance in Statistical Computing – (R-100). SAS Institute Inc. 1978.

[12] I. B. Jeffery, D. G. Higgins, and C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," BMC Bioinformatics. Jul 26 7:359, pp.1-16, Jul. 2006. doi: 10.1186/1471-2105-7-359. (Accessed Oct. 28, 2015).

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

[14] A. Miyake, S. Shinmura, **"**An Algorithm for the Optimal Linear Discriminant Function and its Application," Japan Society of Medical Electronics and Biological Engineering, 18/6, pp. 452-454, 1980.

[15] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics. Third Edition. SAS Institute Inc. 2004. (S. Shinmura, supervise Japanese version)

[16] L. Schrage, LINDO - An Optimization Modeling Systems - . The Scientific Press. 1991. (S. Shinmura & H. Takamori, translate Japanese version)

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

[18] S. Shinmura, "Medical Data Analysis, Model, and OR," Operations Research, 29/7, pp. 415-421, 1984.

[19] S. Shinmura, "Optimal Linear Discriminant Functions using Mathematical Programming," Journal of JSCS, 11 / 2, pp. 89-101, 1998.

 ISBN: 978-1-61208-488-6

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

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

[22] S. Shinmura, "Analysis of Effect of SSM on 152,989 Cancer Patient," ISI2001, pp. 1-2. 2001.

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

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

[25] S. Shinmura, "Overviews of Discriminant Function by Mathematical Programming," Journal of JSCS, 20/1-2, pp. 59-94, 2007.

[26] S. Shinmura, "Practical discriminant analysis by IP-OLDF and IPLP-OLDF," IPSI 2009, Belgrade VIPSI Conference, CD-ROM, pp. 1-17, 2009.

[27] S. Shinmura, The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing. 2010.

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

[29] S. Shinmura, "Beyond Fisher's Linear Discriminant Analysis - New World of Discriminant Analysis -," ISI2011 CD-ROM, pp. 1-6, 2011.

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

[31] S. Shinmura, "End of Discriminant Function based on Variance-Covariance Matrices," ICORES, pp. 5-14, 2014.

[32] S. Shinmura, "Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP," Statistics, Optimization and Information Computing, 2, pp. 14-129, 2014.

[33] S. Shinmura, "Comparison of Linear Discriminant Function by K-fold Cross-validation," Data Analytic 2014, pp. 1-6, 2014.

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

[35] 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, 2015. Springer (ISSN: 1865-0929, ISBN: 978-3-319-17508-9, DOI: 10.1007/978-3-319-17509-6).

[36] S. Shinmura, "A Trivial Linear Discriminant Function," Statistics, Optimization and Information Computing, 3, Dec. 2015, pp. 322-335, 2015.

[37] S. Shinmura, "The Discrimination of the microarray data (Ver. 1)," Research Gate (1), Oct. 28, 2015: pp. 1-4, 2015.

[38] S. Shinmura, "Feature Selection of three Microarray data," Research Gate (2), Nov.1, 2015: pp. 1-7, 2015.

[39] S. Shinmura, "Feature Selection of Microarray Data (3) – Shipp et al. Microarray Data," Research Gate (3), 2015: pp. 1-11, 2015.

[40] S. Shinmura, "Validation of Feature Selection (4) – Alon et al. Microarray Data," Research Gate (4), 2015: pp. 1-11, 2015.

[41] S. Shinmura, "Repeated Feature Selection Method for Microarray Data (5)," Research Gate (5), Nov. 9, 2015, pp. 1-12, 2015.

[42] S. Shinmura, "Comparison Fisher's LDF by JMP and Revised IP-OLDF by LINGO for Microarray Data (6)," Research Gate (6), Nov. 11, 2015, pp. 1-10, 2015.

[43] S. Shinmura, "Matroska Trap of Feature Selection Method (7) –Golub et al. Microarray Data," Research Gate (7), Nov. 18, 2015, pp. 1-14, 2015.

[44] S. Shinmura, "Minimum Sets of Genes of Golub et al. Microarray Data (8)," Research Gate (8), Nov. 22, 2015, pp. 1-12, 2015.

[45] S. Shinmura, "Complete Lists of Small Matroska in Shipp et al. Microarray Data (9)," Research Gate (9), Dec. 4, 2015, pp. 1-81, 2015.

[46] S. Shinmura, "Sixty-nine Small Matroska in Golub et al. Microarray Data (10)," Research Gate, Dec. 4, pp. 1-58, 2015.

[47] S. Shinmura, "Simple Structure of Alon et al. et al. Microarray Data (11)," Research Gate (11), Dec. 4, 2015, pp. 1-34, 2015.

[48] S. Shinmura, "Feature Selection of Singh et al. Microarray Data (12)," Research Gate (12), Dec. 6, 2015, pp. 1-89, 2015.

[49] S. Shinmura, "Final List of Small Matroska in Tian et al. Microarray Data," Research Gate (13), Dec. 7, pp. 1-160, 2015.

[50] S. Shinmura, "Final List of Small Matroska in Chiaretti et al. Microarray Data," Research Gate (14), Dec. 20, 2015, pp. 1-16, 2015.

[51] S. Shinmura, "The best model of the Swiss bank note data," Statistics, Optimization and Information Computing, 3, Spring. 2016, pp. 0-13, 2016. (unpublished).

[52] S. Shinmura, "Discriminant Analysis of the Linearly Separable Data," Journal of Statistical Science and Application, 2016. (unpublished).

[53] S. Shinmura, New Theory of Discriminant Analysis after R. Fisher, Springer, Dec. 2016. (unpublished).

[54] M. A. Shipp et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," Nature Medicine 8, pp. 68-74, 2002.

[55] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," J. Comput. Graph. Statist., 22. pp. 231-245, 2013.

[56] D. Singh et al., "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell: March 2002, Vol.1, pp. 203-209, 2002.

[57] A. Stam, "Non-traditional approaches to statistical classification: Some perspectives on Lp-norm methods," Annals of Operations Research, 74, pp. 1-36, 1997.

[58] G. Taguchi and R. Jugulum, The Mahalanobis-Taguchi Strategy-A Pattern Technology System. John Wiley & Sons. 2002.

[59] E. Tian et al., "The Role of the Wnt-Signaling Antagonist DKK1 in the Development of Osteolytic Lesions in Multiple Myeloma," The new England Journal of Medicine, Vol. 349, 26, pp. 2483-2494, 2003.

[60] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," J. R. Statist. Soc. B 58/1, pp. 267-288, 1996.

[61] V. Vapnik, The Nature of Statistical Learning Theory. Springer. 1995.

 ISBN: 978-1-61208-488-6