Cancer Gene Analysis using Small Matryoshka (SM) Found by Matryoshka Feature Selection Method

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

***Abstract*— We established a new theory of discriminant analysis after R. Fisher. We developed two methods and four Optimal Linear Discriminant Functions (OLDFs) in order to solve five serious problems of discriminant analysis. Over than 30 years, many researchers could not select cancer genes from gene datasets such as microarray datasets (Problem 5). The Matryoshka feature selection method (Method 2) could separate the dataset to several linearly separable subspaces (Small Matryoshka, SM) and non-linearly separable subspace definitely. We consider genes including each SM are cancer genes because they can discriminate cancer patients versus normal patients completely. On the other hand, other genes cannot discriminate two classes correctly and are noise. Therefore, Method 2 can separate the dataset to several signal SMs and noise subspace naturally. In this study, we introduce how to analyze all SMs of six datasets using common statistical methods and propose malignancy indexes for cancer gene diagnosis. Because our standard statistical approach obtains almost the same successful results, we explain the results by Alon et al. dataset. Researchers can analyze their dataset by our approach.**

Keywords—Cancer Gene Analysis; Cancer Gene Diagnosis; Microarray Dataset; Matryoshka Feature Selection Method; Small Matryoshka (SM); Basic Gene Set (BGS); NP-hard; large p small n.

I. INTRODUCTION

We established a new theory, which in this paper will be referred to as Theory, of discriminant analysis [28] after Fisher [8] and solved five serious problems of discriminant analysis [16][19]. We developed four Optimal Linear Discriminant Functions (OLDFs) and two new methods. Integer Programming (IP) defines IP-OLDF and Revised IP-OLDF (RIP) based on Minimum NM (MNM) instead of Number of Misclassifications (NM). Linear Programming (LP) defines Revised LP-OLDF. Revised IPLP-OLDF is a mixture model of Revised LP-OLDF and RIP. IP-OLDF found two new facts about discriminant analysis such as 1) the relation of NM and LDF, 2) MNM monotonic decrease ($MNM_k >= MNM_{(k+1)}$).

In Section 2, we explain the Matryoshka feature selection method achieved by LINGO [14] Program 3, which we call Method 2. Program 3 found all SMs of six datasets in Table 1. In Section 3, we analyze all 64 SMs of Alon et al. dataset [1] by common statistical methods. Those are one-way ANOVA with t-test, Ward cluster analysis, Principal Component Analysis (PCA), logistic regression [5][7], Fisher's Linear Discriminant Function (LDF) and a Quadratic Discriminant Functions (QDF). In Section 4, we analyze 64 RIP discriminant score data of 64 SMs; we get straightforward and surprising results. In Section 5, we discuss the reason why statistical methods cannot analyze the datasets with noise because of the large p small n problem [6] and NP-hard [3]. In Section 6, we conclude that RIP discriminant scores data is very useful because we can propose malignancy indexes for cancer gene diagnosis.

II. NEW THEORY OF DISCRIMINANT ANALYSIS

In this section, we introduce Theory outlook and all SMs of six datasets found by LINGO Program 3 [28].

A. *Five Serious Problems of Discriminant Analysis*

The five problems that we address in this paper are the following. Only RIP can discriminate the cases on the discriminant hyperplane correctly (we will refer to this as Problem 1). Because other LDFs cannot discriminate those cases correctly, their NMs may increase. Other LDFs, except for RIP and a Hard margin SVM (H-SVM) [35], cannot discriminate linearly separable data (LSD) (we will refer to this as Problem 2). Moreover, error rates of Fisher's LDF are very high for LSD discrimination. Problem 3 is the defect of the generalized inverse. QDF misclassifies all cases in class 1 to class 2 if some variable including class 1 is constant and its values including class 2 vary. If we add slight random numbers to the variable, we can solve Problem 3. The discriminant analysis is not traditional inference statistics that offer standard error (SE) equation derived from the normal distribution (we will refer to this as Problem 4). The 100-fold cross-validation method (Method 1) offers the 95% confidence intervals (CIs) of error rates and discriminant coefficients by the computer-intensive approach [15]. Over more than 30 years [10], many medical and statistical researchers were struggling to select cancer genes from the high-dimensional datasets with noise (we will refer to this as Problem 5). We call all linearly separable gene space as Matryoshka. We downloaded six microarray datasets from Higgins HP [12] on October 28th, 2015. When we discriminate these datasets, all NMs of three Revised OLDFs are zero and few coefficients less than case number "n" are not zero. We will refer to this subspaces as first Small Matryoshka (SM1). Next, when we discriminate

 ISBN: 978-1-61208-560-9

the modified dataset with SM1 removed, we find second SM (SM2). If we cannot find other SM anymore, we stop this iteration. This process is Method 2. Therefore, Program 3 finds the microarray dataset is disjoint unions of several SMs and non-linearly separable subspace. Because six microarray datasets are LSD, we believed other microarray datasets are LSD and have the Matryoshka structure, also. Moreover, "MNM monotonic decrease" explains the Matryoshka structure of LSD. Until now, there are no clear explanations about cancer genes. Most researchers do not know the microarray dataset is LSD. We consider genes including each SM as "cancer genes" because MNM using these genes is zero and we can discriminate cancer patients versus normal patients. Therefore, Program 3 can separate the microarray dataset into multiple signals and noise by only discriminating the dataset. Researchers at LASSO [31] [34] tried to zero some discrimination coefficients, and there were many filtering methods [2], but Method 2 has both abilities of LASSO and filtering method. Statistical methods could not be successful by analyzing noise containing microarray datasets, but these SMs are small samples and can be analyzed very easy. In this study, we introduce how to analyze all SMs by the standard statistical approach, and propose several malignancy indexes for cancer gene diagnosis.

B. Six MP-based LDFs and Two Statistical LDFs

Although we developed a diagnostic logic of Electrocardiogram data by Fisher's LDF and QDF, our research was inferior to the decision tree logic developed by the medical doctor. This experience is our motivation to develop Theory. After many experiences of the discriminant analysis, we developed IP-OLDF expressed in (1). Because we fix the intercept of IP-OLDF to 1, IP-OLDF is defined in the p-dimensional coefficient space omitting the intercept. 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 $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1)$ discriminates $\mathbf{x}_i$ 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 misclassifies $\mathbf{x}_i$ because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) < 0$.

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

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 IP-OLDF has a defect if data is not a general position, we developed RIP that looks for the interior point of true Optimal Convex Polyhedron (OCP) defined in (2) directly, NM of which is MNM. Because b_0 is free variable, RIP 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 to be less than -1; there are no cases within SV. Because there are no cases on the discriminant hyperplane, we must understand that the optimal solution is an interior point of OCP defined by IP-OLDF [15], NM of which is MNM. Because all LDFs except for RIP cannot solve Problem 1 theoretically, these LDFs must check the number of cases (h) on the discriminant hyperplane. Correct NM may increase (NM + h). If e_i is non-negative real variable, equation (2) changes Revised LP-OLDF. Revised IPLP-OLDF is a mixture model of Revised LP-OLDF in the first phase and RIP in the second phase.

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

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

When we discriminate Swiss banknote data with six variables [29], IP-OLDF finds that two variables models, such as (X4, X6), are 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 1. This fact is important for gene analysis because (X4, X6) can list all Matryoshkas. Therefore, (X4, X6) is called cancer Basic Gene Set (BGS).

C. Theory Outlook

We established Theory and solved five serious problems of discriminant analysis. Let us consider two-class discrimination with n-cases and p-variables. IP-OLDF is defined on p-discriminant coefficient space. N-linear hyperplanes made by n-cases values as coefficients divide this space into finite Convex Polyhedron (CP). All LDFs corresponding to interior points of CP have unique NM and misclassify the same NM cases. Therefore, the relation between NM and LDF coefficient was first clarified. If we consider that all LDFs corresponding to the same interior points are equivalent, there are finite LDFs as same as finite CPs. There is an OCP. MNM decreases monotonously ($MNM_k >= MNM_{(k+1)}$) because k-coefficient space is sub-space of (k+1)-coefficient spaces. If $MNM_k = 0$ and k is a minimum number of variables, all MNMs including these k-variable are zero. This fact means LSD has Matryoshka structure. Therefore, MNM is a critical statistic of LSD discrimination. We call the k-variable model a cancer BGS, that is the smallest SM in gene analysis because it can explain the Matryoshka structure of high-dimensional dataset by BGS. IP-OLDF finds the right vertex of OCP if data is a general position. However, it may not find the right vertex of OCP if data is not general position. Therefore, we develop RIP that looks for interior point of OCP directly. All LDFs except for RIP cannot discriminate the cases on discriminant hyperplane correctly if they choose the vertex or edge of CP. Therefore, NM of these LDFs may not be correct (Problem 1). Only H-SVM and RIP can discriminate LSD correctly (Problem 2). Although Revised LP-OLDF tends to collect the cases on the discriminant hyperplane, it can discriminate LSD correctly. Logistic regression almost discriminate LSD.

 ISBN: 978-1-61208-560-9

However, Fisher's LDF and QDF often cannot discriminate LSD. QDF and Regularized Discriminant Analysis (RDA) [9] misclassify all cases to another class because of the defect of a generalized inverse (Problem 3). Fisher proposed Fisher's LDF based on Fisher's assumption and established the discriminant theory based on variance-covariance matrices. Although many data do not satisfy Fisher's assumption and there is no real test statistics for it, Fisher's LDF solves many applications. Because there is no equation of SE of LDFs, the discriminant analysis is not the inferential statistical method (Problem 4). Therefore, we developed Method 1 [17] and LINGO Program 2 that can offer the 95% CIs of discriminant coefficients and error rates [18]. Moreover, we developed useful and simple model selection method such as the best model with a minimum mean of error rate in the validation samples (M2) [29]. After examining all the models, we concluded that the best models of RIP were almost better than other seven LDFs such as two OLDFs, three SVMs, logistic regression and Fisher's LDF. About Problem 5, we discuss in section D.

D. Problem 5 and Matryoshka Feature Selection Method

First of all, we developed Program 3 for RIP. However, Program 3 can discriminate the dataset with Revised LP-OLDF, H-SVM, and two Soft-margin SVMs (S-SVMs). Two S-SVMs are SVM 4 (penalty c = 10000) and SVM 1 (penalty c = 1). Revised LP-OLDF and Revised IPLP-OLDF can find SM1 and stop the iteration. Although NMs of three SVMs are zero, most coefficients are not zero. Therefore, three SVMs are not helpful for cancer gene selection because we must survey all possible models [11] to find SM and it is NM-hard. Because NMs of Fisher's LDF by JMP are not zero, Fisher's LDF is not useful for gene analysis. Therefore, only three Revised OLDFs could find that the dataset consisted disjoint unions of several SMs and another noise subspace. To the best of our knowledge, there was no absolute definition of cancer gene, and the purpose of cancer gene analysis was not clear until now. Now, we can define cancer gene set that can discriminate cancer patients versus normal patients or just two different types of cancer. Moreover, because Program 3 separated several signal SMs and noise subspace, Program 3 is a good filtering system that removes noise subspace from the dataset. Program 3 could find all SMs of six datasets in Table 1 [26]. Moreover, we confirmed and validated Program 3 using Swiss banknote data and Japanese automobile data [27] also. Therefore, we claim cancer gene analysis is very easy and exciting theme. When we discriminated Shipp et al. dataset [30] on Oct. 28, 2015, RIP could select 32 genes among 7129 genes. We thought the discrimination having 7129 variables needed huge CPU time by NP-hard. However, Fisher's LDF [13] and six MP-based LDFs [14] can solve the datasets in less than 20 seconds because the datasets are LSD. Generally speaking, MP-based six LDFs are difficult by NM-hard. If datasets are LSD, these LDFs are free from NP-hard. However, most coefficients of three SVMs are not zero. Therefore, SVMs are not helpful for feature selection for gene analysis in addition to common data. Because Revised LP-OLDF minimizes the summation of misclassified case distance from the discriminant hyperplane, which is the second objective function of S-SVM, it can discriminate LSD correctly because this standard is the same as MNM standard only for LSD. However, this standard tends to gather cases on discrimination planes (Problem 1). The SV distance maximization criterion solved by the Quadratic Programming (QP) seems to be preventing the SVM discriminant coefficient from becoming zero. Table 1 shows the summary of six datasets found until December 20th, 2015. Rows "Size" are the case number by the gene number. Rows "SM: Gene" are the number of SM: the total number of genes including in all SMs. Six papers [20] - [25] include full gene name including each SM. Rows "JMP12" are two by two tables of the discrimination by Fisher's LDF. Six NMs are 5, 3, 8, 3, 10 and 29 and error rates are very high. If BGS has k genes, the dataset with p variables includes many smaller Matryoshka from (p - 1) variables to k variables. Program 3 finds that 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 3 can find all list of BGSs, we can understand the Matryoshka structure of the dataset by these BGSs completely. Because we can analyze each SM using common statistical methods, we expect to obtain new facts of gene analysis and hope many researchers try to analyze these SMs. By our breakthrough, the cancer gene analysis becomes an interesting theme.

TABLE I. SUMMARY OF SIX MICROARRAY DATASETS

Data	**Alone et al. [1]**	**Chiaretti et al. [4]**
Size	62 *2000	128*12625
SM: Gene	**64** [22]:**1999**	269 [25]:5220
JMP12	20:2 / 3:37	94:1 / 2:31
Data	**Golub et al. [10]**	**Shipp et al. [30]**
Size	72*7129	77 *7130
SM: Gene	67 [21]:1203	214 [20]:3040
JMP12	20:5 / 3:44	17:2 / 1:51
Data	**Singh et al. [32]**	**Tian et al. [33]**
Size	102 *12626	173 *12625
SM: Gene	178 [23]:3984	159 [24]:7221
JMP12	46:4/6:46	16:20/9:128

III. ANALYSIS OF 64 SMS

In this section, we analyze 64 SMs by common statistical methods because all SMs are small samples.

A. 64 SMs of Alon et al. Dataset found by Method 2

We discriminate Alon et al. [1] dataset by LINGO Program 3. Table 2 tells us the dataset consists of disjoint unions of 64 SMs from SM=1 to SM=64 and one gene. Therefore, Alon et al. were successful in separating signals and noise although they may drop some signals. "Gene"

column is the number of genes in each subspace. All NMs of logistic regression and QDF are zero; we omit two columns from the table. "LDF2 and LDF1" are NMs of two different prior probability options of Fisher's LDFs. The prior probabilities of LDF2 are proportional to the 22 cases in class 1 and 40 cases in class 2. Those of LDF1 are the same and are the default setting in most statistical software. However, we use the former prior probability because we wish to compare NMs of six MP-based LDFs. Ten NMs of LDF2 are greater than NMs of LDF1, and 13 NMs of LDF2 are less than NMs of LDF1. "SR" is the range of RIP discriminant scores. "RatioSV" is the value calculated by 2/SR*100 % that indicates the ratio of the distance of SV for the range of RIP discriminant scores. "t" column is t-value of the discriminant score. Last four rows are maximum, mean, minimum and summation of 64 SMs. The range of Gene is [21, 42], and 64 SMs include 1999 genes. The ranges of LDF2 and LDF1 are [0, 8] and [0, 9], respectively. The 13 NMs of LDF2 are zero, and 12 NMs of LDF1 are zero. The ranges of SR and RatioSV are [7.5, 84.9] and [2.4%, 26.8%], respectively. The maximum RatioSV 26.8% means SV width is 26.8% of RIP discriminant scores and separates two classes. Moreover, the 23 values of RatioSVs are over 5%. We must survey the proper threshold for malignancy indexes by validation in the near future. In this paper, we assume the threshold is 5% or 10%. The range of t-values is [-1.1, 4.6]. Nevertheless, 64 SMs are linearly separable; t-test is not helpful to find linearly separable signs.

TABLE II. SUMMARY OF 64 SMS (OMITTED 54 SMS)

SM	GENE	LDF2	LDF1	SR	RatioSV	t
8	31	0	0	7.5	**26.8**	**0.3**
1	29	2	2	13.5	14.8	2.3
2	33	1	1	14.7	13.6	1.7
60	34	5	6	18	11.1	4.2
4	27	3	3	20.8	9.6	1
44	33	3	4	21.8	9.2	2.5
45	29	1	2	21.7	9.2	1.6
37	30	5	5	23.8	8.4	1.9
14	26	0	0	39.8	**5**	0.5
64	42	8	9	84.9	**2.4**	3.5
Max	42	8	9	84.9	26.8	4.6
Mean	30.8	2.1	2.2	19	12.8	1.7
Min	21	0	0	7.5	2.4	-1.1
SUM	1999	134	139	1218.5	821.6	109.8

B. *Histogram and Correlation*

Because 64 SMs of logistic regression and QDF are zero, these discriminant functions confirm 64 SMs are linearly separable and the quality of Program 3.

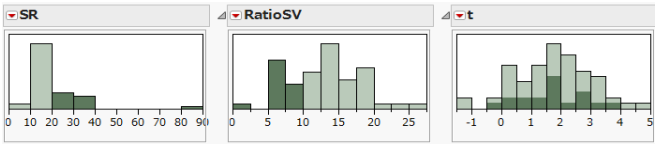

Fig. 1. The Histograms of Gene, LDF2, SR, RatioSV and t.

Figure 1 shows the histograms of SR, RatioSV, and t-value. We select 19 SMs with "RatioSV < 10% and show the portion of 19 SMs by the dark green. The "SR >= 20" are the same as 19 SMs with "RatioSV < 10%. These RIPs using 19 SMs probably need the validation by Method 1 because 19 RIPs may misclassify some patients. Dark green t-value almost cover the ranges. Although most researchers use the t-test, we think t-test is not helpful for gene analysis because we cannot find the useful meaning of all genes included in 64 SMs by t-tests.

C. *Ward Cluster Analysis*

Because four NMs of logistic regression, QDF, LDF2, and LDF1 are zero, we focus on this SM9. We analyze the dataset by Ward cluster analysis. Figure 2 is the heat map of SM9 with 33 genes. Right dendrogram shows cases, and lower dendrogram shows variables. It tells us Ward cluster cannot classify two classes by two clusters clearly. We categorize two clusters. We omit upper cluster including 33 cases that consist of 13 normal cases marked by □ and 20 cancer patients marked by ×. The lower cluster includes 29 cases that include nine normal and 20 cancer cases. This result shows the cluster analysis is not helpful for gene analysis.

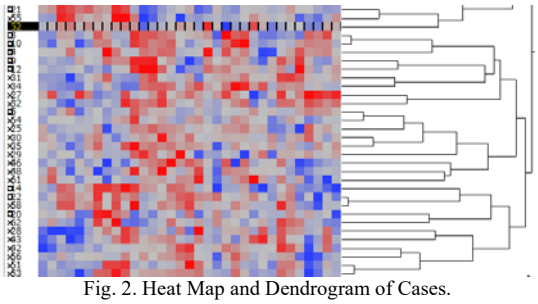

Fig. 2. Heat Map and Dendrogram of Cases.

Figure 3 is the dendrogram of 33 genes. We expect the specialist explains this result. Due to the small distance between X471 and X201, these genes may be able to replace other genes with each other.

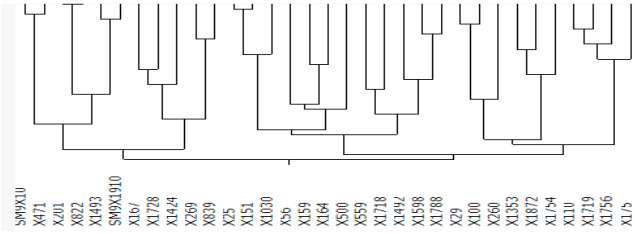

Fig. 3. Dendrogram of Genes.

Figure 4 is PCA output. Left figure is an eigenvalue. Nine principal component eigenvalues are greater than 1. The middle figure is a scatter plot. The right figure is factor loading plot. Although MNM of two classes is zero, scatter plot shows two classes overlap. We confirm 64 SMs scatter plots overlap and cannot find the linearly separable sign. We conclude it is difficult to find the useful meaning of these results. Comparison with Figure 6 confirms our claim.

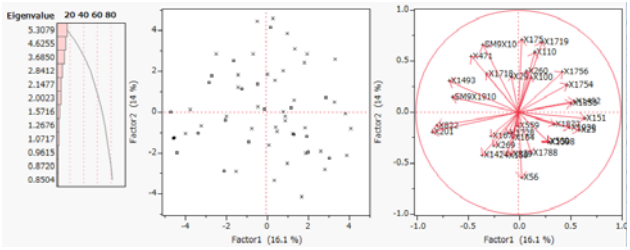

Fig. 4. Three Figures of PCA.

IV. Analysis of 64 RIP Discriminant Scores

We claimed we could analyze SM by common statistical methods. However, we cannot obtain useful results. In this section, we analyze 64 RIP discriminant scores data by common statistical methods and obtain the surprising results. The data consists 62 patients (cases) with 64 discriminant scores (variables).

A. *Examination of 64 RIP Discriminant Scores and RarioSV*

Table 3 shows the summary of 64 RIP discriminant scores. Min and Max columns are the range of normal class 1. MIN and MAX are the range of tumor class 2. SR and RatioSV are the same in Table 2. The 22 normal cases are less than or equal to -1, and the 40 cancer cases are greater than or equal to 1. RIP SV separate two classes of 64 SMs completely. RatioSVs of SM8, SM14 and SM64 are 26.76%, 5.03% and 2.35%, respectively. We guess that the 63 RIPs whose RatioSVs are over 5% may be good malignancy indexes for cancer gene diagnosis.

TABLE III. Summary of 64 discriminant scores omit 59 SM)

SM	Min	Max	MIN	MAX	SR	RatioSV
8	-3.35	-1	1	4.12	7.47	**26.76**
35	-2.58	-1	1	5.92	8.51	23.52
11	-4.15	-1	1	5.52	9.67	20.68
59	-15.25	-1	1	21.91	37.17	5.38
14	-21.94	-1	1	17.85	39.79	**5.03**
64	-3.94	-1	1	81	84.94	**2.35**
MAX	**-2.58**	-1	1	**81**	**84.94**	**26.76**
MIN	**-21.94**	-1	1	**4.12**	**7.47**	**2.35**

B. *Ward Cluster Analysis and PCA*

Ward cluster analyzes the discriminant scores data that consists of 62 patients (cases) with 64 RIP discriminant scores (variables); see Figure 5. The upper cluster is 22 normal cases, and the lower cluster is 40 cancer cases. Ward cluster separates two classes. However, it cannot separate two classes in Figure 2. The 62 cases dendrogram becomes over ten clusters. The 64 discriminant scores dendrogram has more complex clustering.

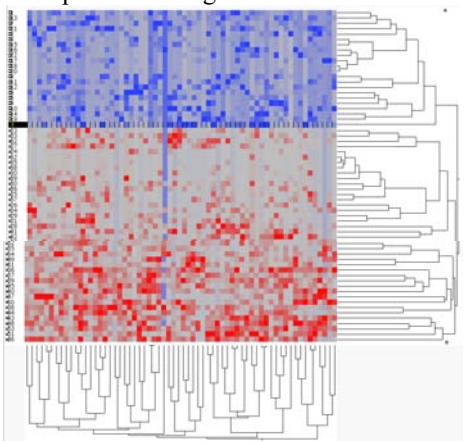

Fig. 5. Ward Cluster Analysis.

Figure 6 is three plots of PCA. The eigenvalue of the first principal component (Prin1) is enormous. Scatter plot shows two classes are completely separable. The 22 normal cases are on a negative axis of Prin1. The 40 cancer cases scatter on the first and fourth quadrants that look like a fan, same as factor loading plot. If we obtain the validation cases, we can judge whether the scatter plot is useful for malignancy index. Although we could not separate two classes in Figure 4, we can separate two classes in Figure 6 very clearly.

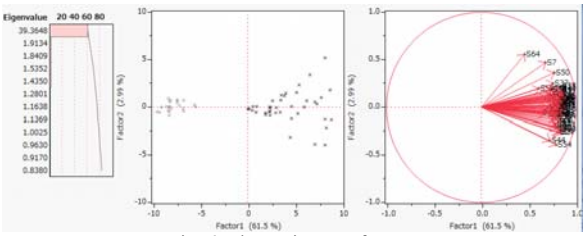

Fig. 6. Three Figures of PCA.

Table 4 is the ranking of four principal components. “R1” is the ranking of Prin1 of 62 cases in descending order. We expect “Prin1” to be the malignant degree of cancer. Because the Prin1 values of tumor class are greater than 0.02 and those of normal class are less than -5.55, RatioSV of Prin1 is 30.37% (= (0.02+5.55) / (8.77+9.57) * 100). We believe cancer patients of ID = 46, 36 and 53 are malignant patients. Moreover, normal patients of ID =18, 20, 17 may become cancer. If medical doctors confirm these patients are serious patients and normal patients with risk (possibility to

become cancer), we can use PCA as a cancer diagnosis. If these indexes misclassify cancer patients cured with some treatment to the normal class, the doctor can decide that the patient is cured ~~earlier than the diagnosis by~~ before 5-years ~~survival rate~~ because their measurements belong to the range of normal class.

TABLE IV. MALIGNANCY INDEX OF PCA

ID	Prin1	R1	Prin2	R2	Prin3	R3	Prin4	R4
46	**8.77**	**1**	-1.35	58	-2.17	59	-1.29	56
36	**8.69**	**2**	1.8	4	2.62	5	0.47	18
53	**8.29**	**3**	-2.12	59	0.83	14	-0.91	47
58	0.43	38	-0.31	40	-0.14	36	-0.02	30
52	0.14	39	-0.19	37	0	30	0.43	20
55	**0.02**	40	-0.23	39	-0.2	37	0.29	25
18	**-5.55**	**41**	0.01	29	0.1	27	0.03	28
20	**-5.64**	**42**	0.22	22	0.13	25	-0.05	31
17	**-6.8**	**43**	-0.43	43	-0.81	47	-0.19	38
21	-9.12	60	-0.07	35	-0.05	32	0.72	14
10	-9.33	61	-0.62	51	0.15	23	-1.05	52
4	-9.57	62	-0.52	44	0.02	29	-1.04	51

C. *Analysis of Transpose Data*

We examine three figures of PCA using transpose data with 64 discriminant scores (cases) by 62 cases (variables) in Figure 7. Factor loading plot shows 22 normal cases (variable) locate in the 2nd quadrant, and 40 cancer cases (variables) locate in the first and fourth quadrants. Scatter plot shows most discriminant scores are placed on the line of -45 degrees with the Prin1. The 13th, 14th, 28th, 63th and 64th discriminant scores are outliers. The first eigenvalue is not large compared with Figure 9.

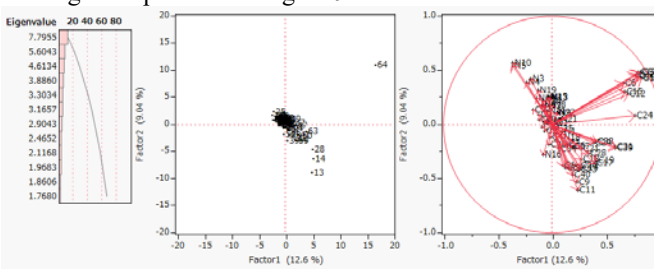

Fig. 7. Three Figures of PCA.

Figure 8 is four scatter plots of PCA. The x-axis is the Prin1. Y-axes are Prin2 and Prin3.

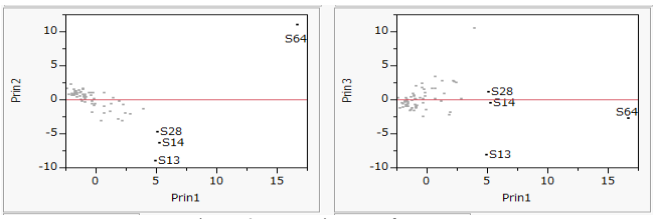

Figure 8. Four Figures of PCA.

There are four RIPs different from other RIPs. These RIPs may show different malignancy indexes. This remains to be validated by medical specialists in future work

D. *Summary of Six Datasets*

Table 5 is the overview of six datasets. Alon et al. and Singh et al. are normal class versus cancer class discrimination. Other four datasets are just two different types of cancer discrimination. However, we obtained almost the same results as 64SMs of Alon et al. dataset. "SM" column shows the number of SMs. Because we find 130 BGSs of Alon et al. dataset, we analyze 130 BGSs in the first row. The range of 130 RIP RatioSVs and 64 RIP SMs are [0.00%, 0.9%] and [2.4%, 26.8%], respectively. Therefore, 130 RatioSVs of BGS are not helpful for malignancy indexes. However, we expect medical specialists examine the gene combination including each BGS and find useful results. ">=5%" column shows the number of SMs with RatioSV over 5%. All 95 RatioSVs of Chiaretti et al. are greater than 5%. Most RatioSVs of Alon et al. and Shipp et al. are greater than 5%, also. Therefore, we conclude Alon et al., Chiaretti et al. and Shipp et al. discriminate two classes easier than other three datasets. "PCA" column shows the RatioSVs of Prin1 that are greater than maximum RatioSVs of RIPs. All NMs of logistic regression are zero. "QDF and LDF2" columns are the number of NM=0 of QDF and LDF2. Although logistic regression and QDF discriminate 159 and 158 SMs by NM=0, Tian et al. dataset has 27 SMs with ">=5%". Although logistic regression and QDF discriminate 159 and 158 SMs by NM=0, Tian et al. [33] has 27 SMs with ">=5%". Tien et al. dataset is easier to discriminate two classes by logistic regression and QDF. Nevertheless, Fisher's LDF are not useful for the datasets except for Chiaretti et al.; statistical discriminant functions can sometimes discriminate SMs correctly.

TABLE V. SUMMARY OF SIX MICROARRAY DATASETS

Data	SM	>=5%	PCA	QDF	LDF2
Alon et al. (BGS)	130	0	4.50%	60	0
Alon et al.	64	63	30.40%	64	13
Singh et al.	179	38	14.40%	26	0
Golub et al.	69	13	34.90%	16	1
Tien et al.	159	27	24.00%	158	1
Chiaretti et al.	**95**	95	51.50%	95	**92**
Shipp et al.	130	129	31.70%	121	46

Figure 9 is three plots of PCA using Chiaretti et al. dataset. Eigenvalue of Prin1 is gigantic compared with Fig. 7 because we guess two classes are apart from each other. Right factor loading plot locates on the first and fourth quadrants that look like a fan. Center scatter plot shows two

 ISBN: 978-1-61208-560-9

classes are completely separable. The 95 B-cell patients in class 1 locate on negative first principal axis (Prin1). The 33 T-cell patients locate on the positive Prin1. Figure 6 is a typical pattern of cancer prediction. On the other hand, Figure 9 may be a typical pattern of just two different types of cancer (cancer classification) introduced by Golub et al..

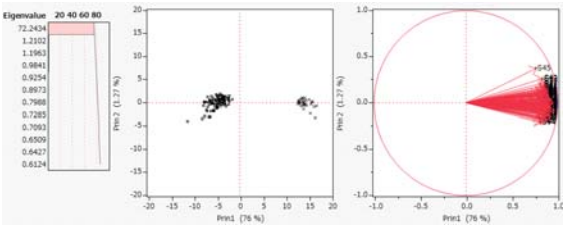

Fig. 9 Three plots of PCA (Chiaretti et al. dataset)

V. WHY IS CANCER GENE ANALYSIS DIFFICULT ?

Golub et al. said: "Although cancer classification has improved over the past 30 years, there has been no general approach for identifying cancer classes (class discovery) or for assigning tumors to known classes (class prediction)." We are unfamiliar with the history of cancer gene analysis and show "over ten years" in the book [28]. We knew some statisticians tried to discriminate gene data by Fisher's LDF about 20 years ago. Now, it is hard for us to find these studies because these trials were judged to be not useful for cancer gene analysis. We understood that most researchers are disappointed in the statistical discriminant analysis. At least, many scientists approached this theme by engineering methods [12] and restricted statistical methods, such as t-test, cluster analysis, and SVM. In this Section, we discuss the reason why cancer gene analysis was difficult.

Diao and Vidyashanker [6] explain cancer gene analysis "large p small n problem." In general, statistical methods treat small sample (small p small n problem) very easy. Now, statistical methods are difficult to analyze a big data (large p and large n) with noise. However, most statistical methods are easy to analyze the data with small p large n problem because of hardware and software ability improvement. Because one-way ANOVA with t-test analyzes each one variable, it is easy to analyze the datasets. On the other hand, some statistical methods are difficult to analyze the datasets with large p small n problem. Regression analysis and discriminant analysis based on variance-covariance matrices are difficult to the datasets with large p small n problem because it is hard to construct large p variance-covariance matrices using small n cases in addition to select feature. In Japan, although JMP released Fisher's LDF for large p small n problem in 2015, the error rate of six datasets are very high in Table 1 because Fisher's LDF cannot discriminate LSD correctly. Charikar et al. [3] introduced the problem called "combinatorial feature selection problems" is NP-hard. Their study gave a significant impact for gene feature selection researchers. In general, their claim is correct, especially for IP models such as RIP. However, six MP-based LDFs including RIP can discriminate each dataset in less than 20 seconds because the datasets are LSD. Until now, there was no study about LSD discrimination. RIP found six datasets are LSD. LSD has the Matryoshka structure because it includes small Matryoshkas in it. Alon et al. and Singh et al. discriminate two classes such as cancer and normal, and other datasets discriminate just two different types of cancer. Both types of dataset consist disjoint unions of several SMs (signals) and other subspace (noise) that is not linearly separable. Our last goal is to find BGS in each SM. BGSs can explain the Matryoshka structure of each dataset uniquely. Medical gene specialist will be able to explain the useful meaning of the genes combination including all SMs shortly. However, we can show the useful results using RIP discriminant scores of 64 SMs of Alon et al.. We already obtain the same results of other five datasets in Table 5, also.

Why did many researchers spend to analyze the datasets over more than 30 years ? Our claims are as follows:

1) Many scientists analyze a dataset with noise by statistical methods. Because they could not obtain useful results, they trusted on engineering approaches, such as several filtering techniques. However, RIP can separate disjoint signals (SMs and BGSs) and noise subspace very easy.

2) If most gene data are LSD, the cancer gene analysis is very easy and straightforward. What is the cancer gene? The answer is very simple. It is the combination of genes including BGS. If we drop one gene from BGS, we can not discriminate this subspace correctly and misclassify some patients.

3) Because NMs of Fisher's LDF are not zero, it is not useful for cancer gene analysis. Because NMs of three SVMs are zero and most coefficients are not zero, SVM is not helpful for cancer gene analysis. If SVM computes all possible models, it can find BGS. However, this trial is NP-hard.

Our conclusion is as follows: Only RIP and Revised LP-OLDF can select gene feature naturally and find SMs. Now, we can find BGSs by manual operation after finding SMs. Every researcher can easily analyze SMs and BGs by standard statistical approaches. Moreover, we find several malignancy indexes. We must confirm our claim by the validation samples.

VI. CONCLUSION

We claimed common statistical methods could analyze SM easily because these subspaces were small samples [28]. However, our examination shows that it is difficult for us to obtain good results showed in Section 3. However, we can get clear results from the RIP discriminant scores in Section 4. Especially, the 63 RIPs with RatioSV over than 5% may be useful malignancy indexes for cancer gene diagnosis. If medical specialists examine and confirm our results and claims, our collaboration will open a new frontier of cancer gene diagnosis from cancer gene analysis. Moreover, the Ward cluster analysis can identify two clusters completely in

 ISBN: 978-1-61208-560-9

Figure 5. Usually, cluster analysis cannot cluster two classes clearly. Figure 6 and Table 3 of PCA show another reliable malignancy index. Healthy patients are located on the negative segment of the Prin1 axis, especially malignant tumor patients widely scatter in the first and fourth quadrants as in Figure 6. Furthermore, the scatter plot of Figure 8 shows the diversity of various tumor patients. We can find 64 SMs and analyze it successfully. However, these results are obtained using the statistical approach. We need co-operation with experts such as medical doctors as follows:

1) We already listed up all genes lists including SMs of six datasets. If the specialists check it, they may find the useful meaning of the combination of genes in each SM.
2) We expect the specialists examine the ranking on the Prin1 and confirm it shows the malignancy indexes.
3) If the specialists offer three types of validation samples such as normal cases, cancer patients and patients with cancer cured, we can discriminate those cases and expect the following result; 1) The normal cases are classified to the normal class. 2) The cancer patients are classified to the cancer class. 3) The patients with cancer cured are classified to the normal class. The misclassified number of the third group shows the degree of cure. If 63 RIPs misclassify the patients in the third group, they may be cured completely, and are relieved from the anxiety of five years.

ACKNOWLEDGMENT

We can achieve our research by the powerful software such as LINGO supported by LINDO Systems Inc. and JMP back up by SAS Institute Japan Ltd. JMP Japan Division. Moreover, reliable datasets are critical for our research. Jeffery et al. [12] upload six datasets and analyze six datasets by ten methods those are adequately conveys the overall picture of cancer gene analysis to us. Six research groups collected these datasets and opened to the Internet world. Although Chiaretti et al. dataset has four different object variables, we use only one variable.

REFERENCES

[1] U. Alon, et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of cancer and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proc. Natl. Acad. Sci. USA, 96, pp. 6745-6750. 1999.

[2] A. B. Brahim and M. Lima, "Hybrid Instance Based Feature Selection Algorithms for Cancer Diagnosis," Pattern Recognition Letters, pp.8. 2014.

[3] M. Charikar, et al., "Combinatorial feature selection problems," IEEE Xplore, pp. 631640. 2000.

[4] 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.

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

[6] G. Diao, and A. N. Vidyashankar. "Assessing Genome-Wide Statistical Significance for Large p Small n Problems," Genetics, 194, pp. 781–783, 2013.

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

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

[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.

[13] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics, Third Edition. SAS Institute Inc. 2004.

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

[15] S. Shinmura, The optimal linearly discriminant function. Union of Japanese Scientist and Engineer Publishing, Japan. 2010 (ISBN 978-4-8171-9364-3)

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

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

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

[19] 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).

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

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

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

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

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

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

[26] S. Shinmura S, "Matroska Feature Selection Method for Microarray Data," Biotechno 2016, pp.1-8 2016 (Best Paper Award)

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

[28] S. Shinmura, New Theory of Discriminant Analysis after R. Fisher. Springer, Dec. 2016.

 ISBN: 978-1-61208-560-9

[29] S. Shinmura, “The Best model of the Swiss banknote data,” Statistics, Optimization and Informatics Computing, 3 Spring 2016, pp. 1-24, 2016.

[30] 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.

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

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

[33] 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.

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

[35] V. Vapnik, The Nature of Statistical Learning Theory . Springer. 1999.

 ISBN: 978-1-61208-560-9