# 癌の高次元遺伝子解析の諸問題（1）

成蹊大学　名誉教授
新村秀一

## 1. 癌の高次元遺伝子解析を取り巻く諸問題について

1999 年から 2004 年まで、米国の 6 研究グループが医学誌に論文を掲載し、研究に用いた microarray データをインターネット上に公開した。ハーバード大学医学部の Golub ら（1999）はサイエンスに論文を発表し、「顕微鏡などで細胞や遺伝子の生物学的な変異から癌遺伝子を見つける限界に言及し、1970 年から microarray による体系的な研究が必要であり研究しているが、大きな成果を未だ得ていない」と真摯に述べている。彼らは独自の手法を開発し、それらを Self-organizing maps(SOM)や LOO や生存時間解析等の統計手法と組み合わせている。これらのデータは、例えば癌と健常者の 2 群 100 例を、1 万個の遺伝子の発現量で判別する問題であり、最も判別に適した遺伝子をがん遺伝子と特定する研究である。しかし、幾つかの論文では SVM の利用は述べられているが明確な結果の紹介はない。そして残念なことに NIH が乳がんを除いて、この種の研究に疑問を呈したレポートを出し、医学的な研究が終わったようだ。

## 2. 新しい判別理論と Matryoshka Feature Selection Method(新理論 2)による癌の遺伝子解析の成功

筆者は 2015 年 10 月 25 日に富山市で開催された科研費シンポジュームで石井博士の発表で上記の 6 種のデータが公開されていることを知った。彼女から 28 日にメールで microarray を入手できる HP を知り入手した。そして 12 月 22 日までの 56 日間で癌の遺伝子解析を完成させた。

Fact3：6 種の microarray すべてが LSD すなわち最小誤分類数 MNM=0 である。

Fact4：全ての microarray は、多数の線形分離可能な Small Matryoshuka(SM)とそれ以外の雑音空間に分割できる。最近まで、2 群が各 SM の部分空間で完全に分かれて識別できるので、SM に含まれる遺伝子が癌遺伝子であり、癌の Basic Gene Set(BGS)と考えていた。そしてその和集合と各 BGS は信号空間であると定義した。すなわち、

1) 分散共分散行列に基づく判別分析は LSD を正しく判別できないので、癌の遺伝子解析に全く役に立たない。
2) 改定 IP-OLDF（RIP）と改定 LP-OLDF とハードマージン最大化 SVM（H-SVM）だけが microarrays が LSD であることがわかる。しかしなぜ多くの研究者は SVM を利用しているのに、この重要な信号を見過ごしたかという問題が起きる。
3) RIP と改定 LP-OLDF だけが、microarray を多数の SM と雑音空間に分割できる。これが今回の発表のテーマである。

そして、SM は小標本であるので、簡単に分析できる。しかし、2 群が各 SM で完全に分かれているにもかかわらず、主成分分析、クラスター分析、一元配置による分散分析などで線形分離可能な事実を示さない。ロジスティック回帰は全ての SM が NM=0 であるが、分散共分散行列による LDF や QDF は NM=0 にならないものが多い。一方、RIP、改定 LP-OLDF と H-SVM は、2 つの SV で-1 以下に class1、1 以上に class2 の症例を正しく判別し、判別スコア(Discriminant Score, DS)の範囲に対する比率 RatioSV が 5%以上になるものが多い。そこで遺伝子の変わりに各 SM に含まれる遺伝子の総合特性値である RIP、改定 LP-OLDF と H-SVM の DS である RipDSs、LpDSs と HsvmDSs を変数とするデータを作成した。これらを PCA やクラスター分析で分析すると、2 群は完全に分かれる。

以上から SM で、癌と健常者の 2 群は完全に分かれているが、通常の統計手法ではそれがわからない。しかし、DS で作成したデータでは LSD の事実が簡単にわかる。すなわち RIP、改定 LP-OLDF と H-SVM の判別スコアが、高次元 microarray の信号でないかと考えるに至った。これに関する詳細は、広島のシンポジュームで報告を予定している。

## 3. 統計的判別分析の問題

Golub 以前にも他の遺伝子データが公開されていて、統計研究者やパターン認識などの工学者研究者は、質が高く無償で提供されたこの高次元データ(small n large p)を、格好の研究テーマとしてとらえて研究を行ってきた。多くの研究論文には”Feature Selection Methods（統計的にはモデル選択とか変数選択）”とか”Filtering”という用語を含んだものが多い。そして医学論文には見られない、次の 3 つの困難を指摘する論文もある。

1) 高次元データ(small n large p；n<<p)の困難さ：これは、例えば100症例から1万次元の分散共分散行列を構築することが端的な事例である。2000 年以前に、国際会議で発表もあったがいつのまにかなくなった。2015年11月に六本木で行われたJMPのDiscovery SummitでJMPの創業者のSall博士が特異値分解を用いた横長データのFisherのLDFを発表した。無償で1か月借り受け、6種のmicroarrayを分析したが、誤分類数（NM）は高かった。すなわち、分散共分散に基づくLDFはLSDを正しく判別できず、癌の遺伝子解析に役に立たない点である。また、数理計画法では全く問題にならずむしろlarge n small pの方が、制約式が増えて計算時間がかかる。すなわち、この困難は2変数の相関でもって1万変数が関係づけられる統計に限定された困難である。数理計画法は変数間の関係が小さい。
2) NP-hard:1万変数の判別分析で、適切な部分モデルを選ぶことは困難である。この困難の真の意味を考えていない側面がある。即ち統計的判別関数や2次計画法で定式化されたH-SVMは、定義域で唯一の最適な判別関数が求まる。部分空間にも最適解がある場合、モデル選択で部分モデルの中から最適解を見つける必要がある。さらに問題なのはLSDでは、最小次元のBGSを含むすべてのモデルが最適解になることである。Feature Selection でMNM=0という基準で最適化モデルを探さない限り、これ等の研究は無意味である。
3) 信号と雑音の分離：この問題は的を得ているが、「信号」の定義がはっきりしない。癌のMicroarrayデータにおける信号はMNM=0である。2017年1月の金沢における科研費シンポジュームで、私が初めて癌の遺伝子解析に成功したという説明に、青嶋氏より我々の方が先行している旨の意見があった。よく考えてみると、2015年の富山で石井氏がmicroarrayデータでPCAを行うと、第1固有値だけがスパイク上に大きな固有値になり、一般的な常識で考えられない結果になるという話を聞いた。しかし、私はmicroarrayが簡単に入手でき、これまで判別分析で4つの問題を解決してきたが、「未解決の5番目の癌の遺伝子解析」を解決していないことに気づいた。癌と健常あるいは異なった2種の癌が、遺伝子空間で完全に分かれて2つの球に布置しているというのが青嶋と矢田らの結論である。このことは、私の「2群はmicroarrayでMNM=0であり、それが多くのMNM=0であるSMとMNMが1以上の雑音空間に分割できる」という驚く結果を統計的に検証した研究であることに気づいた。

## 4. 癌の遺伝子解析から癌の遺伝子診断

**癌の遺伝子解析**は、判別分析の4つの問題と応用問題として6種のmicroarrayで全てのSMを求めることができSpringerの本で解決した問題5(Shinmura, 2016d)を指す。豊富な実証研究の成果である。

2016年になって、SMはn個以下の遺伝子で構成された小標本であるので、統計手法で簡単に分析できると考えた。しかしロジスティック回帰だけがNM=0になり、PCAやクラスター分析では２つのclassが線形分離可能な事実を示さなかった。それも仕方がないと当初は考えたが、RIP、改定LP-OLDFとH-SVMは、MNMあるいはNMが0である。そこで判別スコアの範囲に対して2つのSV間の距離2の比をRatioSVとして求めると6種の最大値は[11.67%, 38.93%]と異常に大きい。これに反して、Alonの130個のBGSは1%未満である。そして、RIPの判別スコアを遺伝子の代わりに変数として用いたデータを作成した（RipDSsデータ)。これをPCAで分析すると健常症例が第1主成分で負のある値以下に、癌症例が正のある値以上に布置することが分かった。そして各SMで求めたRipDSとPCAで求めた総合化されたRipDSを**癌の悪性度指標**と呼び、医学の素人ながら**癌の遺伝子診断**の突破口を開いたのではないかと考えた。それらの成果をまとめて、2017年にAmazonからKindle版として出版した。予約注文で600部以上がダウンロードされ幸先が良いと思ったがその後が続かない。きっとResearch Gateで出版案内したので、癌の遺伝子解析に関連した研究者がNIHの敗北宣言後も600人程度は細々といると考えられる。9月時点でRGのRead数が11万を超えたが癌の遺伝子解析関連のRead数は少ない。また7月末にラスベガスで開催されたBiocomp18で8月3日（金）に開催ホテルのLuxorで朝3時に目が覚め空港に行く7時まで,初めて約140にいたFollowingとFollowerの所属を調べた。帰国後RGがダウンし復旧後に1391人に増えた。ひょっとしてBiocomp18で遺伝子関連の専門家に注目されたかと調べてみたが、40人程度しかいないようである。また、彼らが癌の遺伝子関連のDraftを特に読んでいる事実も得られなかった。癌の遺伝子診断の成果は、専門家に検証してもらわなければ意味がないので、袋小路に入っている。

今後の課題として、青嶋・矢田らの結果に対して、筆者の研究アプローチで具体的な幾つかの事実で同じことを示していることを示す予定でいる。

[1] Aoshima M, Yata K (2017) Two-sample tests for high-dimension, strongly spiked eigenvalue models, Statistica Sinica Preprint No.ss-2016-0063R2:1-31

以下は、Biocomp18（2018/7/30~8/2, Las Vegas, Luxor Hotel）で開催されたConference PaperでなぜRIPとRevised LP-OLDFが簡単にmicroarrayをSMと雑音空間に分割できるかを、BCD2018、IEEE Conference（7/9~7/11, Yonago）とともに示している。

# First Success of Cancer Genetic Analysis by Microarrays

**Shuichi Shinmura**
Seikei University, Kichijouji, Tokyo, Japan

**Abstract** - *Specification of cancer genes using microarrays has been done since 1970. Six prominent US projects published papers and released their microarrays. Statisticians considered cancer gene analysis as a new research theme because microarrays are high-quality and high-dimensional data. However, they could not succeed because the discriminant analysis was not helpful. We found that six microarrays are linearly separable data by Revised IP-OLDF. In addition, microarrays could easily be decomposed from 64 pairs to 179 pairs of small genes less than n patients. Furthermore, we found many malignancy indexes and opened the possibility of the genetic diagnosis of cancer. Many researchers believe that useful information cannot be obtained from microarrays. Genetic analysis of cancer was unsuccessful because the statistical discriminant analysis was useless. On the other hand, our new theory of discriminant analysis could solve this theme completely. In this paper, we explain these reasons by the many empirical studies.*

***Keywords:*** *Linearly Separable Data (LSD); Matryoshka feature selection method (Method2); Small Matryoshka (SM); Revised IP-OLDF (RIP); Gene Analysis; Gene Diagnosis.*

## *1* Introduction

Although we developed a diagnostic logic of ECG data by Fisher’s linear discriminant function (LDF) [8] and quadratic discriminant function (QDF), our research was inferior to the decision tree logic developed by the medical doctor in 1974 because ECG data did not satisfy Fisher’s assumption. This is our motivation to develop the new theory of discriminant analysis. After many experiences of the discriminant analysis, we found two facts and five severe problems on discriminant analysis [18-20] [23]. We developed four optimal LDFs (OLDFs) [17] and the 100-fold cross-validation for small samples (Method1) [21] that solved five problems completely. Section 2 explains the new theory of discriminant analysis after R. Fisher (Theory) [35]. Section 3 explains the cancer gene analysis (Problem5) and the Matryoshka feature selection method (Method2) [32] [36]. Although many medical and statistical researchers studied to specify cancer genes from microarrays, they could not succeed. However, Revised IP-OLDF (RIP) based on the minimum number of misclassifications (minimum NM, MNM) and Method2 could decompose six microarrays into small plural subspaces (Small Matryoshkas, SMs) and the noise subspace. All MNMs of SMs are zero and signals. MNM of noise subspace is over than one. Thus, we can define the definition of signal and noise clearly. Section 4 explains the cancer gene diagnosis by malignancy indexes and RatioSV [37]. Section 5 is the conclusion.

## *2* New Theory of Discriminant Analysis

### 2.1 Fisher’s LDF

Fisher defined Fisher's LDF by Fisher’s assumption and developed the theory of discriminant analysis. The discriminant analysis becomes an important statistical method as same as the regression analysis. However, since there is no proper test for Fisher’s assumption, Fisher’s LDF is applied for many applications that do not satisfy Fisher's assumption. Moreover, statisticians ignored many problems of the discriminant analysis and developed many discriminant functions based on the variance-covariance matrices those were useless for linearly separable data (LSD). The fact that these discriminate functions could not discriminate LSD correctly was the serious problem (Problem2). Considering the two groups as a Gaussian distribution of $f_i$ =e ^ (- (**x** - **m**) $^2$ / 2s$^2$) / (SQRT (2 π *s$^2$) for i=1,2, the logarithm of these ratio becomes the following linear equation (1). We think that Fisher defined (1) by the feature of the exponential function.

$$\log(f_1/f_2) = \log [\ e\text{^}\{-(x-m_1)^2/2s^2+(x-m_2)^2/2\ s^2\}]$$
$$= (m_1 - m_2)/s^2{*}x + (m_2^2-m_1^2)/(2{*}s^2) \qquad (1)$$

Nowadays, most researchers misunderstand that Fisher’s LDF was obtained by maximizing the correlation ratio and partial differential obtains this optimal solution. Statistical researchers and users are the most distant from mathematical programming (MP). Fisher easily constructed the discriminant theory in the era without the computational environment by avoiding the optimum solution obtained by partial differential according to the actual data. Fisher also developed the maximum likelihood estimation method that was used for the logistic regression [5] [7]. In addition, he or the same generation of researchers developed QDF. If actual data does not satisfy Fisher's assumption, they recommended using QDF. We must study the flexible correspondence and wisdom of the predecessor. He never found Fisher's LDF that matches the data by maximum likelihood estimation. When assuming a multidimensional normal distribution, statistical LDFs can be easily obtained simply by obtaining the variance-covariance matrices of the p variable. For these reasons, both statisticians and statistical users were relieved from the troublesomeness without knowing the difference between the maximum/ minimum values and the local maximum/local minimum values and enjoyed the advantage compared with MP theory. Six microarrays are LSD, NM of those are MNM=0. However, the maximization criterion of the correlation ratio cannot correctly distinguish the LSD. We had already show we could not determine the pass or failure of the examinations using exam scores because the error rates were very high (Problem2). Therefore, it is quite useless for genetic analysis of cancer. This is because discriminant theory did not study discrimination of LSD at all. Thus, biostatisticians and gene specialists could not solve Problem5 from 1970.

## 2.2 Summary of New Theory

### 2.2.1 Two New Facts found by IP-OLDF and MNM

We established the Theory in 2015 that consists four OLDFs and two methods such as the Method1 and Method2. Although there are five severe problems of discriminant analysis, Theory can solve five problems completely. In 1997, the definition of IP-OLDF using integer programming (IP) found two new facts of discriminant analysis as follows:

1) The definition of IP-OLDF reveals the relation of NM and LDF on the discriminant coefficient space. This fact explained the defect of NM clearly (Problem1). Moreover, only RIP can find correct NM, NM of which is MNM. All NMs of other LDFs may not be right and increase.

2) The MNM decreases monotonously ($MNM_k >= MNM_{(k+1)}$). If data are LSD and $MNM_k = 0$, all MNMs of models including these k-variables are zero [33]. This fact means that “MNM monotonic decrease” means the Matryoshka structure of LSD and microarrays. LSD includes many small subspaces (SMs), MNMs of those are zero. We call all linearly separable space and its subspaces as Matryoshka. Swiss banknote data [9] [34] consist of six variables and 200 banknote bills (n > p). When we discriminate all possible models [12], we found MNM of the two-variable model (X4, X6) was zero. This is the smallest SM (Basic Gene Set, BGS) in gene analysis. Thus, we can find 16 SMs in this data by the monotonic decrease of MNM. Other 47 models are not Matryoshkas, MNMs of which are over one. Since IP defines RIP and linear programming (LP) defines Revised LP-OLDF, both OLDFs find the vertex of a convex polyhedron (feasible region) made by p-constraints out of n-constraints made by p-variables.

### 2.2.2 MP-based LDFs

IP defines RIP in (2). If $e_i$ is non-negative real variable, Eq. (2) changes Revised LP-OLDF that is solved by LP. Revised IPLP-OLDF is a mixture model of Revised LP-OLDF in the first phase and RIP in the second phase. Feasible region is defined by constraints and is the convex polyhedron. LP optimal solution is one of the endpoints of the feasible region. In the case of small samples (n>p), it is a solution of at most p constraints selected from n constraints. In the case of microarrays (n<<p), it is a solution of at most n constraints by setting (p – n) coefficients zeros.

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

$b_0$: free decision variable.

$\mathbf{b}$: p-coefficients.

$e_i$ : 0/1 integer variable

$y_i$ : -1 for class1, 1 for class2

M: 10,000 (Big M constant)

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

$e_i$: non-negative real value.

$$MIN=||\mathbf{b}||^2/2+c*\Sigma e_i; \tag{4}$$

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

c: penalty c to combine two objectives.

The equation (3) is a hard margin SVM (H-SVM) [43] that explains LSD-discrimination firstly. The quadratic programming (QP) defines SVMs. Before H-SVM, nobody can define whether data is LSD or overlap. Moreover, the most researcher believes LSD-discrimination is easy. Now, LSD is defined by “MNM=0,” and overlap data is “MNM>=1” clearly. Thus, no researchers could define LSD clearly before H-SVM and MNM. However, H-SVM causes the computation error for the overlap data. This fact may be the reason why nobody discriminates microarrays or study LSD-discrimination.

The equation (4) is soft-margin SVM (S-SVM). If we set $c=10^4$ or c=1, it becomes SVM4 or SVM1. We compare SVM4 and SVM1 because there is no research to choose the proper c. By the results of best models [34], the best models of SVM4 are almost better than SVM1. If we omit “$||\mathbf{b}||^2/2$ and c=1”, it becomes Revised LP-OLDF that can find SMs. However, S-SVM cannot find SM.

Even though Revised LP-OLDF can find SMs, three SVMs cannot select SM. This difference is caused by QP that looks for the one minimum solution on the gene space and cannot find one of the minimum solutions on the gene subspaces. This fact indicates QP prevent to find SM.

### 2.2.3 Five Problems of Discriminant Analysis

The only RIP based on the MNM criterion can discriminate the cases on the discriminant hyper-plane theoretically. Because other LDFs may not be able to discriminate these cases correctly, pure NMs of these LDFs may increase (Problem1). Although NM is the vital statistic of the discriminant analysis, no statisticians recognize the defect of NM. Thus, even though we developed MNM instead of NM, some journal rejected our paper for the reason that MNM criterion was a foolish idea. Since Fisher never proposed the standard errors of error rate and discriminant coefficients, the discriminant analysis was not the traditional inferential statistics (Problem4). Thus, we proposed the Method1 that offered the 95% confidence interval of coefficients and error rates [22]. Moreover, we proposed the model selection method such as the best model with a minimum mean of error rate in the validation samples (M2) instead of a leave-one-out method [14]. The best model is the M2 obtained by the 100 validation samples among all possible models. The best models of RIP almost have minimum M2s among eight LDFs using six different types of common data such as Swiss banknote data, Fisher’s iris data, student data, Cephalo Pelvic Disproportion data, many pass/fail determinations of exam scores, Japanese 44 cars data. Seven LDFs are two OLDFs, three SVMs, logistic regression and Fisher’s LDF. The best models of Fisher’s LDF were worst, except for Fisher’s iris data. This fact indicates MNM is robust statistics instead of NM. RIP and H-SVM can discriminate LSD theoretically (Problem2). Although the pass/fail determination using examination scores are LSD, error rates of Fisher’s LDF and QDF are very high [24]. This fact indicates the statistical discriminant functions based on the variance-covariance matrices are useless for LSD-discrimination such as microarrays. However, only logistic regression can discriminate all SMs empirically because it is solved by the maximum likelihood. Therefore, we consider Cox models and logistic regression open the new second frontier of the discriminant analysis. We found the defect of generalized inverse matrices of variance-covariance matrices (Problem3). At first, JMP [15] QDF misclassified all students of the passed class to the failed class if some variable of the passed class is constant. This is a disadvantage of traditional statistics that all data seems to be different. We solved this problem to add a small random number to the constant variable. We spent three years to solve Problem3 because our approach was wrong as same as the cancer gene analysis that could not solve from 1970. Although many researchers were struggling for Problem5, we solved it within 54 days in 2015 by MP-based LDFs instead of statistical discriminant analysis.

# *3* Cancer Gene Analysis by Method2

## 3.1 All SMs of Six Microarrays

Jeffery, Higgins, and Culhane upload six microarrays used by six prominent US medical researchers and propose ten feature selection methods [13]. To the best of our knowledge, there are no papers to point out six microarrays are LSD definitely. LSD has the Matryoshka structure that includes SMs in it. Table 1 shows six microarrays used in six papers published from 1999 to 2004. It shows the summary obtained by LINGO Program3 in 2015 [16]. “Description” shows two classes. Singh et al. microarray [38] [41] consist the 50 healthy subjects (class 1) and the 52 tumor patients (class 2). “Size” are the number of case and gene. “SM: Gene” are the number of SM and the total number of genes included in all SMs. “JMP12” are NM of Fisher’s LDF. Six NMs are 5, 3, 8, 3, 10 and 29. Error rates in the parenthesis are 8, 2, 11, 4, 10 and 17%, respectively. Especially, Tian error rate is very large. Although JMP enhances Fisher’s LDF for microarrays (JMP ver12, JMP12) [15], this fact indicates that discriminant functions based on variance-covariance matrices are useless for Problem5. Whether or not LSD can be accurately discriminated is the first step in cancer gene analysis. Whether the discriminant coefficient can be 0 or not is the second barrier important for the genetic diagnosis of cancer. If researchers discriminated microarrays by H-SVM, they could find microarrays were LSD. However, there was no research that microarrays were LSD definitely. It is unbelievable why researchers could not find this important fact.

Table 1. Summary of six Microarrays by Method2

| **Data** | **Description** | **Size** | **SM:Gene** | **JMP12** |
|---|---|---|---|---|
| Alone<br>et al. [1] | Normal (22) vs.<br>tumor cancer (40) | 62 *<br>2000 | 64 :<br>1152 [28] | 5<br>(8) |
| Chiaretti<br>et al.[4] | B-cell (95) vs.<br>T-cell (33) | 128*<br>12625 | 270:<br>5385 [31] | 3<br>(2) |
| Golub<br>et al.[11] | All (47) vs.<br>AML (25) | 72*<br>7129 | 69:<br>1238 [27] | 8<br>(11) |
| Shipp<br>et al[40] | Follicular lymphoma(19)<br>vs. DLBCL (58) | 77*<br>7129 | 213:<br>3032 [26] | 3<br>(4) |
| Singh<br>et al[41] | Normal (50) vs.<br>tumor prostate (52) | 102 *<br>12625 | 179:<br>11387 [38] | 10<br>(10) |
| Tian<br>et al[42] | False (36) vs.<br>True (137) | 173 *<br>12625 | 159:<br>7221 [30] | *29 (17)* |

When we discriminated Shipp et al. microarray [25] on Oct. 28, 2015, only 32 RIP coefficients were not zero. Since MNM of 32 genes is zero, these genes are oncogenes. Those discriminated two classes completely. We misunderstand the discrimination having 7,129 variables requests huge CPU time. However, Fisher’s LDF by JMP12 and other six MP-based LDFs coded by LINGO can solve microarrays less than 20 seconds because those are LSD. However, most coefficients of SVMs are not zero. Thus, SVMs are useless for feature selection of gene analysis. If BGS has k-variables, the biggest Matryoshka with 7,129 variables includes much smaller Matryoshka from 7,128 (= 7,129 - 1) variables to k variables. LINGO Program3 can decompose microarrays into plural SMs with $h_i$-variables ($p > h_i >= k$) and another high-dimension noise gene subspace with “MNM >= 1.” If LINGO Program4 can find all list of BGSs quickly, we can understand the Matryoshka structure of microarrays by these BGSs completely. Because we can analyze each SM using standard statistical methods, we expect to obtain new facts of gene diagnosis and hope many researchers try to analyze these SMs. By our breakthrough, the cancer gene analysis becomes an interesting theme.

## 3.2 Three Difficulties or Excuses

From 1970 [11], many statisticians could not succeed to specify oncogenes from microarrays (Problem5). They claimed three difficulties or excuses. These difficulties are merely excuses caused by a narrow world of statistics. They could not understand that only discriminant functions suffered these difficulties. MP-based LDFs are free from these difficulties. Fisher's LDF explains why.

1) It was difficult to obtain the variance-covariance matrix for small n large p data [6]. However, with singular value decomposition JMP developed Fisher's LDF which can distinguish microarray. However, six NMs are not zero. Since the correlation ratio maximization criterion cannot correctly distinguish LSD, it is entirely useless for gene analysis. For MP-based LDFs, "Small n large p" is easier to analyze than "large n small p" from the computation time.

2) NP-hard to select gene feature [3]. Since statistical discriminant functions and SVMs find only one optimal functions on the whole domain, these functions must compute all possible models to find SMs.

3) It is difficult to separate signal and noise. Because there is no precise definition of the signal, signal and noise cannot be appropriately separated. In our study, we defined the set of genes with MNM = 0 as the signal.

Fisher's LDF is useless for gene analysis because NMs of six microarrays are not zero. Fisher's assumption, variance-covariance matrix and correlation ratio maximization cannot theoretically discriminate LSD. Although regularized discriminant analysis [10] and LASSO [2] are mainstream of discriminant analysis after R. Fisher, those cannot discriminate LSD theoretically. This is because they disregarded Fisher's consideration, ignored reality data, and used normal distribution as the starting point of the theory not based on MNM criterion. "Lotus eating" brings unfortunate results.

### 3.3 The reason why LP and IP can find SMs

Microarrays are high dimensional data that is called as small n and large p data (n<<p). In this case, RIP and Revised LP-OLDF find the vertex of a convex polyhedron (feasible region) made by n-constraints having p-variables. One of the apexes of the feasible region is a solution of n simultaneous equations, and it is obtained by setting (p - n) genes to 0. Thus, LP and IP can find one of the subspaces (SMs) as the optimal solution. This means only $p_1$ ($p_1$ <= n) discriminant coefficients of both OLDFs are not zero and other coefficients become zero. Since six microarrays are LSD, RIP and Revised LP-OLDF can find SM with less than n genes because of n<<p. This fact is the reason why RIP could solve Problem5 54 days from October 28 to December 20 in 2015.

On the other hand, NMs of H-SVM and Soft-margin SVM (S-SVM) are zero, and most coefficients are not zeros. Thus, these SVMs are useless for gene diagnosis. QP defines three SVMs and finds only one optimal solution on the whole region as same as statistical discriminant functions. In order to find SMs, these SVMs need to compute all possible models. This computation is NP-hard. This claim is our final conclusion [39].

## 4. Cancer Gene Diagnosis by RatioSV

### 4.1 Analysis of all SMs

Since all SMs were small samples with $n_i$ subjects and $k_j$ genes and all $k_j$ are less than $n_i$, we expected the standard statistical methods analyzed all SMs and could show good results for cancer gene diagnosis. Those statistical methods are one-way ANOVA, t-test, cluster analysis, principal component analysis (PCA), logistic regression, Fisher’s LDF and QDF. Since all NMs of logistic regression were zero, logistic regression confirmed all SMs were LSD. However, Fisher’s LDF and QDF could not discriminate all SMs correctly. Moreover, other methods did not show the linearly separable signs that two classes were utterly separable in each SM. At first, we expected “medical specialists will be able to find useful meanings from these results.” However, we concluded these results had no useful meanings at all.

### 4.2 RIP discriminant scores and RatioSV

We could not obtain useful results of all SMs by standard statistical methods, except for logistic regression. Next, we discriminate all SMs by the RIP and obtain RIP discriminant scores (RipDSs). Since Singh et al. microarray is decomposed 179 SMs, we get 179 RipDSs from 179 SMs. **Table 2** is the summary of 179 RipDSs that is sorted in descending order of RatioSV in (5). RatioSV is the second important statistic for LSD in addition to MNM.

RatioSV = SV distance *100/ the range of RipDS
= 200 / RDS (%) (5)

Table 2. 179 RipDSs of Singh Microarray

| RIP | Min | Max | MIN | MAX | RDS | RatioSV | t (≠) |
|---|---|---|---|---|---|---|---|
| RIP2 | -8.58 | -1 | 1 | 8.56 | 17.14 | *11.67* | 14.57 |
| RIP179 | -266.57 | -1 | 1 | 440.43 | 706.99 | 0.28 | 5.78 |
| MAX | -8.22 | -1 | 1 | 440.43 | *706.99* | *11.67* | 15.5 |
| MEAN | -33.94 | -1 | 1 | 47.47 | 81.41 | 3.59 | 10.85 |
| MIN | -266.57 | -1 | 1 | 8.56 | 17.14 | *0.28* | 5.78 |

The “Min and Max” columns are the range of the 50 healthy subjects. “MIN and MAX” columns are the range of the 52 tumor patients. The 50 healthy subjects are less than equal -1, and the 52 tumor patients are higher than equal 1. SV separates 102 subjects correctly. The sixth column is the range of DS (RDS). The seventh column is RatioSV. Because the distance of SV is two, this statistic is the ratio of SV’s width to RDS (%). We expect this statistic indicates the degree of separation of the two classes and malignancy index of a cancer gene diagnosis. The last column is the t-values under the condition that both variances are not equal. Since this t-test checks the difference between two averages on DS, all values are positive. However, if we check all t-values of each gene included in each SM, those values are either of negative, almost zero and positive values. Therefore, although some studies claimed genes with large positive t-values were the oncogene, those claims were not right. Although we cannot explain the meaning of genes with almost zero, these genes are needed for diagnosis.

RatioSV is good statistics for LSD-discrimination because it gives us the degree that SV separate two classes. We can understand RatioSV of SM2 by RIP (RIP2) can discriminate two classes very easy, and SV of RIP179 scarcely separates two classes because its RDS is 706.99 and very large. The last three rows are the maximum, mean and minimum of seven variables. The range of RDS, RatioSV and t-value are [17.14, 706.99], [0.28%, 11.67%] and [5.78, 15.5], respectively. “RatioSV” recommends RIP2 because it is the maximum value among 179 RIPs. The range of RIP2 is [-8.58, 8.56] and its width is 17.14 (RDS). We focus on RIP2 of SM2. We think RatioSV is vital statistics for the LSD-discrimination. In SM2, SV of RIP2 can divide two classes completely by 11.67% width against RDS. On the other hand, SM179 has a minimum value of RatioSV; that is 0.28%. Therefore, the RIP179 may not discriminate the validation samples correctly. Until now, there is no research on LSD-discrimination. MNM is the first important statistics because it defines LSD by MNM=0 and overlapping data by "MNM>=1" clearly. Some statisticians claim the purpose of discrimination is to discriminate the overlapping data, not LSD. However, they cannot define the overlapping data definitely because they did not have a technical term such as MNM. NM cannot judge the data are overlapping or not. Since RatioSV shows the ease of classification of the two classes, it is another important statistic of cancer gene analysis and diagnosis.

## 4.3 New Data made by RipDSs

Since we could not obtain the useful results of SMs by standard statistical methods, we make new data with $n_i$ subjects and all RipDSs as variables instead of genes. By this breakthrough, six new data made by microarrays have almost the same marvelous results [37].

### 4.3.1 Ward Cluster Analysis

Many researchers analyzed microarrays by cluster analysis. However, if we analyze 179 RipDSs data of Singh, the Ward cluster separates two classes as two clusters clearly, and both dendrograms of case and variable may be meaningful in **Figure 1**. The 102 subjects become six clusters. Upper three clusters belong to class1 and lower three clusters belong to class2 clearly. We expect the medical gene specialist will explain the medical meaning of our statistical results. We are willing to offer the more precise results. It is critical that both cluster analysis and PCA of other five microarrays have almost the same results as Singh et al.

### 4.3.2 PCA

**Figure 2** is three plots of PCA. Left eigenvalue shows the first eigenvalue is 113.7 and its cumulative ratio is 63.5%. The second eigenvalue is 4.39, and its cumulative ratio is 2.45%. Moreover, other 178 eigenvalues share 36.5%, and 72 subjects almost vary on the first principal axis (Prin1). Middle scatter plot shows two classes are completely separable and scatter on the Prin1. Healthy subjects almost locate on minus Prin1. Tumor patients scatter on the first and fourth quadrants that look like a fan. Right factor loading plot locates on the first and fourth quadrants. The 179 correlations of Prin1 and 179 RipDSs are over 0.7. The 179 correlations range of Prin2 and 179 RipDSs are [-0.4, 0.5]. Therefore, Prin1 may be useful for the malignancy index of cancer. The ranges of tumor patients and healthy subjects are [0.99, 22.53] and [-17.89, -4.81], respectively. RDS is [-17.89, 22.53]. Thus, RatioSV of PCA = (0.99 + 4.81) * 100 / 40.42 = 14.3%. Because RatioSV of SM2 is 11.67 %, the Prin1 is more reliable than the discrimination of SM2 because the Prin1 is the total judgment result of 179 RIPs. Only RatioSV of PCA by Golub is smaller than its maximum RatioSV of individual RipDS.

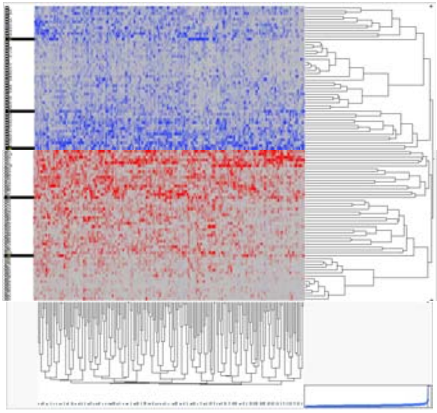

Figure 1. The Heat Map and Dendrogram of New Data

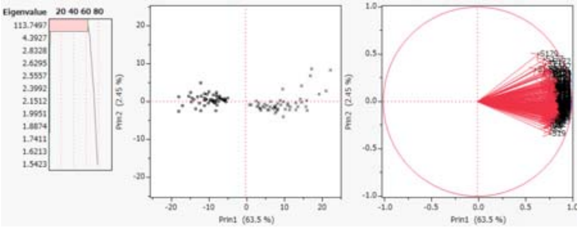

Figure 2. Three Plots of PCA

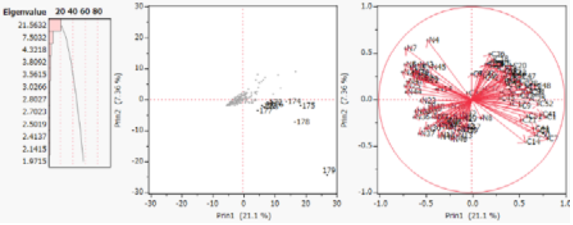

Figure 3. Three Plots of Transposed Data

We transpose the new data and analyze the transpose data with 179 RipDSs (179 cases) and 102 subjects (102 variables). **Figure 3** is three plots of PCA. Factor loading plot shows healthy subjects locate in the 2nd and third quadrants and tumor patients locate in the first and fourth quadrants. The scatter plot shows that the two classes are on two lines and roughly 45 degrees with Prin1. The 174th, 178th, 179th and other several RipDSs of tumor class are outliers those may indicate new subclasses pointed by Golub et al. If we can cooperate with medical specialists, we can understand the different role of 179 RIPs more precisely. If medical doctors confirm these RIPs shows the malignancy indexes of cancer, we can use 179 RIPs and Prin1 as a cancer diagnosis in addition to five-year survival rate. If so, it is the gospel to the patient.

### 4.3.3 Malignancy indexes by PCA

We analyze other five microarrays and publish the position book that proposes the cancer gene diagnosis [37]. In 2017, because we find 130 BGSs of Alon et al. microarray by LINGO Program4, **Table 3** is a summary of RatioSVs of 130 BGSs of Alon et al. in addition to all SMs of six microarrays. Although BGS is more critical than SM for the study of cancer gene research, BGS may be useless for cancer gene diagnosis because the ranges of BGS and SM of Alon are [0.001%, 0.9%] and [2.35%, 26.76%], respectively. We must investigate the threshold of RatioSV for cancer gene diagnosis in future work.

If RatioSVs over than 5% are useful for cancer gene diagnosis, 63 RIPs among 64 SMs are useful for malignancy indexes. If some cancer patients are cured by treatments and are misclassified into a healthy class by 63 RIP malignancy indexes, medical doctors may judge their patients are cured entirely before five years after treatments. This is our dream.

Table 3 The Summary of RatioSVs of RIP and PCA

| Data | SM/BGS | Max Ratio | Min Ratio | PCA |
|---|---|---|---|---|
| *Alon et al.* | 130 | 0.90% | 0.001% | 4.50% |
| *Alon et al.* | 64 | 26.76% | 2.35% | 30.40% |
| *Singh et al.* | 179 | 11.67% | 0.28% | 14.35% |
| *Golub et al.* | 69 | 15.69% | 0.00% | 34.88% |
| *Tien et al.* | 159 | 19.13% | 0.63% | 24% |
| *Chiaretti et al.* | 95 | 38.98% | 10.73% | 51.46% |
| *Shipp et al.* | 130 | 30.67% | 4.99% | 31.70% |

# 5. Conclusions

We solve Problem5 within 53 days because Theory is most suitable for cancer gene analysis using microarrays. Many researchers could not solve Problem5 after 1970 because of the following reasons:

1) Statistical discriminant functions are useless for cancer gene analysis. These functions cannot discriminate LSD theoretically. The remaining two difficulties are unrelated excuses of these functions.
2) If some researchers discriminated microarrays by H-SVM, he or she found microarrays were LSD. Since we have already found "MNM monotonic decrease" before 2010, they could solve Problem5 around 2010.
3) When we explained the draft [37] to Japanese genetic specialist in 2017, he interrupted our explanation and suggested us as follows. “Because in the USA it has already been concluded that the microarrays were useless for genetic analysis at all, we had better terminate our research.” However, the microarrays used by the six US research groups included information useful for cancer gene diagnosis. It is much easier for them to verify our results, compared to the research they have done. We think that it was impossible for them to doubt that the statistical theory which seemed to be perfect was useless at all. We would like to propose they complete their research by verifying our results.
4) There are many reasons for failure. The development of the discrimination theory has been developed on a hypothesis of the normal distribution (Lotus eating), which was proposed in a period without a computer environment. Fisher verified his LDF by the actual data such as Fisher's iris data. QDF was recommended if data did not satisfy Fisher’s assumption. However, many posterity researchers have neglected empirical research based on real data and have developed a mathematical theory based on normal distribution. For these reasons, nobody found four problems of discriminant analysis. In particular, NM, which is the basis of the discriminant analysis, has many drawbacks (Problem1). Although there are problems with NM, even more, difficult statistics are proposed without actually considering whether it is useful or not. Moreover, although the discrimination result of LSD can be explicitly evaluated, it is a problem that this research is not done.
5) Logistic regression using the maximum likelihood estimation method confirms that it can empirically discriminate all SMs correctly. Many users today use logistic regression and SVM because they vaguely understand the problem of discriminant function based on the variance-covariance matrix.
6) We are the first success of cancer gene analysis that can decompose six microarrays into plural SMs and the noise subspace. Although the gene size included in all SMs were less than the number of subjects, the standard statistical methods could not analyze all SMs and obtain useful information. However, RatioSV indicates that several malignancy indexes are useful for cancer gene diagnosis. These results must be validated by medical gene specialists.
7) Notably, it is the most effective that members of six research groups will validate our results. We expect their cooperation will establish the cancer gene diagnosis using microarrays. It is clear that only the oncogenes already found medically cannot completely separate the two groups. We believe that the combination of newly discovered genes in microarrays will open up a new world of cancer gene diagnosis and contribute the human being.
8) If a research group with genetic data makes us a cooperative researcher, we will be able to complete the analysis shown in this research earlier than anyone and provide the results.

# 7. 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] P. Buhlmann, A. B. Geer, Statistics for High-dimensional Data - Method, Theory and Applications-. Springer, Berlin. 2011

[3] M. Charikar, V. Gurus, R. Kumar, S. Rajagopalan and A. Sahai, “Combinatorial feature selection problems,” IEEE Xplore, pp. 631-640. 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] B. Flury, H. Riedwyl, Multivariate Statistics: A Practical Approach. Cambridge University Press, New York. 1988.

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

[11] 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.

[12] J. H. Goodnight, SAS technical report – the sweep operator: its importance in statistical computing – R(100). SAS Institute Inc, USA. 1978

[13] 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.

[14] P. A. Lachenbruch. Discriminant Analysis. NY: Hafner. 1975.

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

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

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

[18] S. Shinmura, Optimal Linear Discriminant Function using Mathematical Programming. Dissertation, March 2000: 1-101, Okayama University, Japan. 2000b

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

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

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

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

[23] 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 (DOI: 10.1007/978-3-319-17509-6).

[24] S. Shinmura, “A Trivial Linear Discriminant Function. Statistics,” Optimization, and Information Computing, 3: 322-335 (DOI: 10.19139/ soic.20151202). 2015

[25] S. Shinmura, “The Discrimination of microarray data (Ver. 1),” Research Gate (1): 1-4, 28 Oct 2015.

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

[27] S. Shinmura, “Sixty-nine Small Matryoshka in Golub et al. Microarray Data (10),” Research Gate, Dec. 4, pp. 1-58, 2015.

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

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

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

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

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

[33] S. Shinmura, “Discriminant Analysis of the Linearly Separable Data,” Journal of Statistical Science and Application, 2016.

[34] S. Shinmura, “The Best Model of Swiss banknote data,” Statistics, Optimization and Information Computing, 4: 118-131, June 2016 (doi: 10.19139/ soic.v4i2.178 ISSN 2310-5070 (online) ISSN 2311-004X (print))

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

[36] S. Shinmura, “Cancer Gene Analysis using Small Matryoshka (SM) found by Matryoshka Feature Selection Method,” Biotechno 2017, pp.1-8 2017.

[37] S. Shinmura, From Cancer Gene Analysis to Cancer Gene Diagnosis. Amazon, June 2017.

[38] S. Shinmura, “Cancer Gene Analysis by Singh et al. Microarray Data,” ISI2017, pp.1-6, 2017.

[39] S. Shinmura, “Cancer Gene Analysis of Microarray Data,” 3rd IEEE/ACIS International Conference on BCD2018, pp.1-6, 2018.

[40] 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.

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

[42] 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.

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