Shuichi Shinmura

High-dimensional Microarray Data Analysis

Cancer Gene Diagnosis and Malignancy Indexes by Microarray

Springer

High-dimensional Microarray Data Analysis

Shuichi Shinmura

High-dimensional Microarray Data Analysis

Cancer Gene Diagnosis and Malignancy Indexes by Microarray

Springer

Shuichi Shinmura
Emeritus Professor
Seikei University
Musashino, Tokyo, Japan

ISBN 978-981-13-5997-2 ISBN 978-981-13-5998-9 (eBook)
https://doi.org/10.1007/978-981-13-5998-9

Library of Congress Control Number: 2018966861

© Springer Nature Singapore Pte Ltd. 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This book extends the possibility of a cancer gene diagnosis using many results. Medical researchers tried to identify oncogenes from genetic data such as microarrays since 1970, but they did not obtain precise results because the statistical discriminant analysis was useless for their research. In 2017, we explained our surprising results to Japanese genetic expert. He told us as follows: "After NIH reports microarrays are useless for cancer gene diagnosis, many researchers believe that this theme has ended. Therefore, you terminate your research." I am regretful to start the study from 2015. If we could show our results before NIH's report, we believe that microarray genetic diagnosis has contributed to cancer control at this time. Some statisticians focused on this research theme as a new field of "big or high-dimensional data analysis" which is different from a small sample (small n and small p data). However, they pointed out three excuses for the difficulty of research. Although it was easy to use highly reliable data collected by physicians, they did not obtain a definite result. The discriminant analysis is the most useful method to classify the two groups of cancer and normal patients or two different cancers. However, since the statistical discriminant functions are utterly useless, medical researchers use cluster analysis such as "self-organizing map" (SOM) and so forth. They seemed to have used discriminant analysis in the early stages of the study, but they probably judged it to be utterly useless.

In this book, as a successful application example of "high-dimensional data analysis" using microarray, it concretely shows that new discriminant theory (**Theory**) is most suitable for cancer gene analysis and diagnosis in addition to the small samples (small n and small p). The fatal problem of conventional studies is that they do not know that the two classes are entirely separable in the high-dimensional gene space (**Fact3**). There was no research of linearly separable data (**LSD**) discrimination except for our research. Most researchers did not understand that only H-SVM and Revised IP-Optimal Linear Discriminant Function (Revised IP-OLDF, **RIP**) can find Fact3, and other LDFs including LASSO cannot discriminate microarray correctly. This fact indicates only mathematical programming (MP)-based LDFs can find Fact3. Statistical discriminant functions are useless for cancer gene analysis. Therefore, they could not define "signal" in high-dimensional genetic space clearly. They cannot select cancer genes from microarray or filter oncogenes from noise without being based on the correct signal. We call the linearly separable gene space and subspaces as Matryoshka. Microarrays (big Matryoshka) include many small Matryoshkas in it. Moreover, RIP and the Matryoshka feature selection method (**Method2**) can decompose microarray into many Small Matryoshkas (SMs) and noise gene subspace (**Fact4**). Because the quadratic programming (QP) defines SVMs, those cannot decompose into many SMs. QP finds only one optimal H-SVM on the whole region. In order to find optimal subspace (**SM**), H-HVM surveys all possible models. It is NP-hard. If we call the smallest Matryoshka as cancer basic gene set (BGS), LINGO Program3 can find many SMs and LINGO Program4 can find many BGSs. At first, because each SM (or BGS) consists of few genes, we expected statistical methods analyzed those small samples and obtained many useful results for cancer gene diagnosis. However, although all NMs of logistic regression are zero for all SMs of six microarrays, other methods do not show the linear separable facts (**Problem6**). After many trials, we produce the signal data made by RIP discriminant scores of all SMs instead of genes included in SM or BGS. LINGO Program3 decomposes Alon's microarray into 64 SMs, and LINGO Program4 decomposes it into 130 BGSs. The RatioSVs of 64 SMs and 130 BGSs are [2.33%, 26.76%] and [0.00%, 0.9%], respectively. Because 64 RatioSVs of all SMs are over 2.33%, we judge SMs are useful for cancer gene diagnosis. On the other hand, BGSs are useless for cancer gene diagnosis because 130 RatioSVs of all BGSs are 0.9% or less. We expect BGS is important for cancer gene research as same as Yamanaka's four genes in iPS research. That is, when a normal patient becomes cancer, RIP discriminates two classes clearly. However, statistical discriminant functions cannot discriminate two classes (Problem6) because of two reasons. First reason is those cannot discriminate LSD theoretically. Second reason is all RatioSVs of BGS are tiny. When other genes are added to BGS and become 64 SMs, SV can separate two classes very easy. This result seems that SM is more suitable for the cancer gene diagnosis than BGS. As a future task, we must clarify of the classification and roles of many SMs and BGS (Problem7).

This book proposes the cancer gene diagnosis and malignancy indexes analyzing all SMs obtained from six microarrays. However, the malignancy indexes need to be verified by medical professionals. Therefore, we disclose LINGO programs and explain many statistical results used for verification in this book. These results offer benefits for statistical researchers and statistical education because many persons can easily participate in this field, using our successful examples of the "high-dimensional data analysis." Also, due to maximum use of our statistical knowledge, this book can be used for the excellent guidebook of the data analysis. Moreover, seven problems and four facts that no one has pointed out in statistics will undoubtedly be useful to improve your actual data analysis abilities. We expect many persons such as medical researchers, statisticians, and statistical users contribute to the cancer gene diagnosis, in order to produce useful results. However, although many engineers such as pattern recognition and machine-learning tried

Problem5, they did not succeed also. It was very strange because they were free from the restriction of normal distribution.

Chapter 1 introduces a novel theory of discriminant analysis and its application to the genetic analysis of cancer with a new perspective (*New Theory of Discriminant Analysis After R. Fisher*, Springer 2016). I graduated from the university in 1971 and participated in the development project of "Electrocardiogram Automatic Diagnosis System" at the Osaka Prefectural Adult Disease Center. Dr. Nomura, the leader of project, given us the theme of diagnostic logic to separate normal symptom and several abnormal symptoms by discriminant analysis. Four years the discriminant study was inferior to empirical branching logic developed by doctor Nomura at all. The reason is that the statistical discriminant theory is useless because many data used for medical diagnosis are not a normal distribution. This failure was motivated to research new discriminant theory. Then, based on many empirical studies such as medical data until 2015, I established a new discriminant theory. I first showed the relationship between number of misclassification (NM) and discriminant coefficient (Fact1). From this fact, we could explain many defects of NM (Problem1). We have developed IP-OLDF and Revised IP-OLDF (RIP) based on minimum NM (MNM) criterion instead of NM. I found a monotonic decrease of MNM (Fact2). Also, for Swiss banknote data with six variables, MNM = 0 for two variables (X4, X6). In other words, we can ultimately distinguish between genuine and counterfeit notes. With MNM monotonic decreasing nature, the 16 models containing these two variables are MNM = 0, and 47 out of the remaining MNMs are more than one. This fact is a first discriminant study on LSD that is essential for the genetic analysis of cancer (Problem2). There are other two problems such as deficiencies of generalized inverse matrices (Problem3) and discriminant theory that is not inference statistics (Problem4). Because both problems have little relation with cancer gene analysis, we do not explain in this book precisely. The six research groups in the USA published papers on the genetic diagnosis of cancer using microarrays during the period from 1999 to 2004. They released the microarrays on the Internet. When RIP discriminates the microarrays in 54 days from 25th October to 20th December 2015, we found that the six MNMs are zero (Problem5). No researchers could solve this problem since 1970 because the existing discriminant theory was useless. That is, cancer and normal patients are entirely separable in the high-dimensional genetic space, which is the fact that it is LSD (Fact3). Based on Fact2, we found that the gene space is a Matryoshka structure containing many SMs in which MNM = 0. We developed a Matryoshka feature selection method (Method2). RIP and Method2 could decompose microarrays into many SMs (or BGS) (Fact4). Because of completing the research theme since 1971, we published "New Theory of Discriminant Analysis After R. Fisher" from Springer (2016). In Chap. 1, Method2 decomposes Swiss banknote data and Japanese car data into several SMs. In other words, Method2 is a general-purpose method for high-dimensional data and common data. Furthermore, it shows how RIP and Revised LP-OLDF can easily produce many SMs. The reason why H-SVM using QP cannot obtain SM can be understood by the common sense of MP. That is, the cancer gene analysis cannot be done with a statistical discriminant function based on normal distribution. And the cancer gene analysis is easy for MP-based LDFs. Using LINGO Program3 introduced in Chap. 10, we can divide arbitrary microarray and ordinary data into SM. We analyze this SM by statistical method and propose genetic diagnosis of cancer in Chap. 2 and below.

Chapter 2 introduces the cancer gene diagnosis using SMs (*From Cancer Gene Analysis to Cancer Gene Diagnosis*. 2017). In order to evaluate many SMs found in Method2, we created a statistic called RatioSV. Like MNM, this is an essential statistic of LSD-discrimination. In Alon's dataset (*Proc.Natl.Acad. Sci. USA* 96: 6745–6750, 1999), RIP found 130 pairs of BGS in addition to 64 pairs of SM. The 130 SVs of BGS separated cancer and normal patients at less than 1%. The 64 SVs of SM separated the two groups from 2.4% to 26.8%. Although these results indicate the discrimination of SM is easy, no researchers could not succeed from 1970. BGS is vital for the study of oncogene combinations, but we judged that it was not useful for cancer gene diagnosis. Because SM is a small sample (small n and small p), we considered the standard statistical methods are useful for the analysis of SM. However, only logistic regression was found to be NM = 0 for all SMs. Two groups often overlapped by other statistical methods (Problem6). Therefore, we created new data with RIP discriminant score (RipDS) as a variable and showed this signal data is a true signal in microarrays. By this breakthrough, the analysis was carried out by standard statistical methods using signal data. Especially, PCA and cluster analysis separate the two groups completely. It was also found that the first principal component of PCA represents the malignancy index of cancer the same as the DS of each SM. Because we need to verify these results medically, we published the book from Amazon to call for cooperation among the six research groups. However, there were no answers as following reasons: (1) Six projects may have ended after 2004, (2) they did not access this book and our papers because we are medically unknown, and (3) the Kindle version is not an academic journal. In Chap. 2, we outline the results of cluster analysis and PCA obtained by using six microarrays. After Chap. 3, we examine our claim about the signal by many approaches.

Chapter 3 explains the cancer gene diagnosis of Alon dataset to compare 39 SMs by Revised LP-OLDF and 56 SMs by RIP. In 2017, only RIP and Revised LP-OLDF were convinced that the datasets could be decomposed into different combinations of SMs. Therefore, if 39 pairs of SM obtained by Revised LP-OLDF with a short calculation time are useful for genetic analysis of cancer, it is more useful than using 56 sets of SM obtained by RIP. Therefore, they were analyzed by RatioSV and various statistical methods, compared and evaluated. In conclusion, almost the same results were obtained in any analysis.

In Chap. 4, we try that we have not done so far. One is the evaluation of the signal and noise separated by the RIP and Revised LP-OLDF. For this reason, we analyze Alon's microarray (2000 genes). RIP finds 62 SMs (1968 genes) and noise subspace (32 genes). Revised LP-OLDF finds 32 SMs (1005 genes) and noise subspace (995 genes). Although we have analyzed individual SMs so far, we have not evaluated a signal subspace and noise subspace. When we discriminate the signal and noise subspaces by RIP, it is certainly confirmed that the MNM of the signal subspace is 0 and the noise subspace is more than one. In addition, many normal cases locate on SV = −1, and many cancer cases were on SV = 1. This shows that many cases are concentrated on two points in a high-dimensional signal subspace. The Revised LP-OLDF decomposes a signal subspace with 1005 genes into 32 SMs and a noise subspace with 995 genes. Both the signal and noise subspaces are NM = 0, which indicates that Revised LP-OLDF cannot separate SM from the noise subspace. This is the reason why the Revised LP-OLDF cannot make NM = 0 for all of the linearly separable subspaces (Fact1). We examine the correlation of the genes contained in the signal subspace, and it was found that they are all fairly high-positive correlations. Moreover, we explain the reason why the statistical methods cannot find Fact3.

From Chaps. 5 to 9, we introduce the cancer gene diagnosis of other five datasets. Those datasets are Golub dataset (*Science* 286(5439): 531–537. 1999), Shipp dataset (*Nature Medicine* 8(1.1): 68–74. 2002), Chiaretti dataset (*Blood* 103: 2771–2778. 2004), Singh dataset (*Cancer Cell* 1(1.1): 203–209. 2002), and Tian dataset (*The New England Journal of Medicine*, 349: 2483–2494. 2003). Each chapter shows different verification results to explain Problem6 and Problem7.

In Chap. 10, we will discuss three LINGO programs. The first model is the LINGO sample model developed by Schrage, which is explained by common data such as Swiss banknote data, Japanese automobile data, and iris data. Since the high-dimensional gene datasets are unfamiliar for a statistical user, the threshold is high for statistical users. By explaining genetic diagnosis with common data, familiarity is born even for general statistical users. In particular, Swiss banknote data and Japanese automobile data are LSD, but RatioSV is very small, less than 0 0.1% as same as BGSs. This contrasts with the genetic diagnosis. With these programs, not only microarrays but also other data can easily be decomposed by RIP. This will be useful for research on marketing and exam questions and product characteristics as a new research theme to classify many variables. We are released from the curse of high-dimensional data and prove the theory can solve six problems of discriminant analysis.

Research Gate: https://www.researchgate.net/profile/Shuichi_Shinmura

Research Map (Japanese Researchers DB): https://researchmap.jp/read0049917/

Economic Department HP: http://sun.econ.seikei.ac.jp/~shinmura/

Please refer to Research Gate for the update of this book.

Musashino, Japan

Shuichi Shinmura
Emeritus Professor of Seikei University

Acknowledgements

We can achieve our research by the dominant software such as LINGO supported by LINDO Systems Inc. and JMP backed by SAS Institute Japan Ltd. JMP Japan Division.

I wish to acknowledge the following researchers who contributed to my research for this book.

Linus Schrage, Kevin Cunningham, Mark Wiley (LINDO Systems Inc.); Hitoshi Ichikawa (LINDO Japan); John Sall, Noriki Inoue, Kyoko Takenaka (JMP Division of SAS Institute Inc.); Takaichirou Suzuki, Akira Ooshima, Yutaka Nomura (Osaka International Cancer Institute, Center for Adult Diseases); Naoji Tsuda (SCS); Hiromi Wada (Kyoto University); Toshio Fukuzumi (Gene Science); Aki Ishii, Masahiro Mizuta, Mika Sato-Ilic, Atsuhiro Hayashi, Ian B Jeffery, Kazunori Yamaguchi, Michiko Watanabe.

I also wish to thank my family: Reiko, Makiko, Hideki, Kana, and Yasuhiro Watanabe. Moreover, I am grateful for the legacy of my late father, Otojirou Shinmura, who supported the research.

Contents

Abbreviations

Cancer gene	In our cancer gene analysis, we use cancer genes instead of oncogene.
Cancer gene analysis	RIP, Revised LP-OLDF and H-SVM find six microarrays are LSD (Fact3). Moreover, Method2 decompose each microarray into many SMs and noise subspace (Fact3) by RIP and Revised LP-OLDF, not H-SVM. We call these analysis as cancer gene analysis. It is important statistical analysis are useless
Cancer gene diagnosis	If we make three types of signal data by Rip DSs (RipDSs), LpDSs and HsvmDSs, statistical methods can analyze these signal data and find many malignancy indexes those open the new frontier of cancer gene diagnosis
Common data	The iris data, the student data, the CPD data, the Swiss banknote data, the Japanese automobile data, and the pass/fail determination using examination data
CP	A convex polyhedron on discriminant coefficient space
HsvmDSs	H-SVM discriminant scores
LDF	Linear discriminant functions such as Fisher's LDF, logistic regression, four OLDFs, and three SVMs
LOO	Leave-one-out method
LpDSs	Revised LP-OLDF discriminant scores
LSD	A linearly separable data, MNM of which is zero

Matryoshka	All linear separable spaces and subspaces
Matryoshka structure	The microarray is a big Matryoshka that includes small Matryoshka in it. MNM monotonic decrease is the same idea as Matryoshka structure
Method1	The 100-fold cross-validation for small sample
Method2	Matryoshka feature selection method that can discriminate the common data and the microarrays. It can find SM and decompose LSD into many SMs
OCP	An optimal CP, NM of which is MNM
Oncogenes	This word is used for cancer genes found by physicians
PCA	Principal component analysis
Prin1	The first principal component
QDF	A quadratic discriminant function
RDA	A regularized discriminant analysis
RipDSs	RIP discriminant scores
Signal data	Made by RIP, Revised LP-OLDF, and H-SVM
SOM	Self-organizing map
Standard statistical methods	One-way ANOVA with t-test, correlation analysis, univariate analysis, hierarchical cluster analysis, principal component analysis (PCA), QDF, Fisher's LDF, logistic regression
Statistical discriminant functions	Fisher's LDF, QDF, RDA and LASSO including logistic regression. However, only logistic regression can discriminate all SMs correctly. Other discriminant functions are fatal in determining the LSD and are useless

Symbols

Our Research Theme: Discrimination of two classes (n Cases and p Variables) by Eight LDFs and QDF

LDF	Linear Discriminant Function $f(\mathbf{x}_i) = b_1*x_1 + ,\ldots, + b_p*x_p + c$
DS	Discriminant score $f(\mathbf{x}_i)$ for ith case $\mathbf{x}_i$ for $i = 1,\ldots,n$
Extended DS	$y_i* f(\mathbf{x}_i)$ for $y_i = -1$ for class1 and $y_i = 1$ for class2
RipDS	RIP discriminant score
LpDSs	Revised LP-OLDF discriminant score
HsvmDSs	H-SVM discriminant score

Book0 Optimum Linear Discriminant Functions, 2010, JUSE Press, Ltd.

Book1 New Theory of Discriminant Analysis After R. Fisher: advanced research by the feature selection method for microarray data, 2016, Springer.

The theory consisted of two facts, two methods, and four optimal linear discriminant functions (OLDF) and solved five problems.
Four OLDFs and three SVMs are solved by LINGO Program1.
Method1 is solved by LINGO Program2.
Method2 is solved by LINGO Program3.

Book2 From Cancer Gene Analysis to Cancer Gene Diagnosis, 2017, Amazon Kindle version.

Book3 High-dimensional Microarray Data Analysis—Cancer Gene Diagnosis and Malignancy Indexes by Microarray, 2019, Springer.

Two Facts of Theory

Fact1 The relation of the number of misclassification (NM) and discriminant coefficient.

Fact2 Minimum NM (MNM) decreases monotonously (MNM$_k$> = MNM$_{(k+1)}$)

Six Mathematical Programming (MP)-Based LDFs by LINGO

IP-OLDF Integer programming (IP) defines IP-OLDF based on a minimum number of misclassifications (MNM).
The definition of IP-OLDF found Fact1 and Fact2.

Revised IP-OLDF (RIP) RIP can solve Problem1, Problem2, and Problem5.

Revised LP-OLDF Linear programming (LP) defines Revised LP-OLDF that minimizes the summation of distance from support vector (SV) only for misclassified cases.

Revised IPLP-OLDF A mixture of Revised LP-OLDF and Revised IP-OLDF.

H-SVM Quadratic programming (QP) defines SVM.
Hard-margin SVM solves Problem2 theoretically.

S-SVM A soft-margin SVM.

SVM4 S-SVM for penalty c = 10000.

SVM1 S-SVM for penalty c = 1.

Two Methods

Method1 The 100-fold cross-validation for small sample method solves Problem4.

M1 The mean error rate in the training sample.

M2 The mean error rate in the validation sample.

The Best Model The model with minimum M2 instead of leave-one-out by Method1

Method2 The Matryoshka feature selection method solves Problem5 and finds Fact3

Important Statistics

MNM	Minimum number of misclassifications
NM	A number of misclassifications. There are many defects of NMs.
M1	The minimum mean of error rate in the training sample
M2	The minimum mean of error rate in the validation sample
Best model	The model having the minimum M2 among all possible models. To choose the best model is simple feature selection method instead of LOO methods.
Malignancy indexes	When RIP, Revised LP-OLDF, and H-SVM discriminate all SMs, we obtain many DSs. Because two classes are separable by DSs,
RatioSV	RatioSV is the distance of SV/the range of discriminant score*100. RatioSV evaluates the malignancy indexes.
RatioSV of PCA	We calculate RatioSV on the Prin1 axis of PCA.
RatioS	The ratio of number of genes contained in all SMs/total number of genes.

LINGO Programs:

LINGO	MP solver that can solve LP, IP, QP, NLP, and stochastic programming.
LINGO Program1	Program1 of RIP and H-SVM finds microarrays are LSD (Fact3).
LINGO Program2	Program2 solves six MP-based LDFs by Method1.
LINGO Program3	Program3 of Method2 can decompose microarray into many SMs and noise subspace (Fact4).
LINGO Program4	Program4 finds BGS.

Discriminant Functions by JMP

JMP	Statistical software supported by the JMP division of SAS Institute Japan.
Fisher's LDF	LDF under Fisher's assumption.
LDF1	The prior probability is set 1:1.
LDF2	The prior probability is proportional to the numbers of two classes.
Logist	Logistic regression.
QDF	Quadratic discriminant function.
RDA	Regularized discriminant analysis.

Seven Problems and Four Facts

Five Problems of Discriminant Analysis

Book1 introduces the theory using six different types of common data. Method1 evaluates eight LDFs such as three OLDFs, three SVMs, logistic regression, and Fisher's LDF. Moreover, Method2 firstly succeeds in the cancer gene analysis using six microarrays as follows:

(1) RIP and Fact1 solve Problem1.
(2) RIP and H-SVM solve Problem2 by Fact2.
(3) Method1 solves Problem4. We compare six MP-based LDFs, Fisher's LDF, and logistic regression by the best models of six different common data. All M2s of RIP are better than those of other seven LDFs. These facts show LDFs based on MNM criterion do not overestimate the validation samples.
(4) Method2 and RIP solve Problem5.

Problem1 All LDFs cannot discriminate the cases on the discriminant hyperplane. This is one of the defects of NM.

Problem2 All LDFs, except H-SVM and RIP, cannot recognize linearly separable data (LSD) theoretically. Error rates of discriminant functions based on variance–covariance matrices are very high.

Problem3 The defect of the generalized inverse matrix technique and quadratic discriminant function (QDF) misclassifies all cases as other classes for a particular case. Adding small random noise to the constant values solves Problem3.

Problem4 Fisher never formulated an equation for the standard error of the error rate and discriminant coefficient. Method1 offers a 95% confidence interval (CI) for the error rate and coefficient. Because M1 and M2 are more useful than 95% CI of the error rate, there are explanations for the 95% CI of the error rate.

Book3 discusses two problems and two facts.

Matryoshka We call all linear separable gene space, and subspaces are Matryoshka.

SM Method2 finds small Matryoshka, genes of which are less than or equal to n. LINGO Program3 of RIP finds all SMs correctly.

BGS The minimum-dimensional SM. LINGO Program4 finds BGS.

Problem5 From 1970, many researchers could not succeed in the cancer gene diagnosis. Three OLDFs and H-SVM find the microarrays are LSD (**Fact3**). Moreover, only RIP and Revised LP-OLDF can decompose the microarrays into many SMs and noise subspace. These facts show RIP and Revised LP-OLDF make feature selection and separate signal gene subspace from noise subspace naturally.

Method2 of RIP finds a surprising structure of the microarrays those are the exclusive unions of many SMs. Therefore, we can analyze all SMs for the cancer gene diagnosis.

However, Method2 of H-SVM cannot decompose the microarrays into SM because QP finds only one H-SVM on the whole domain. H-SVM needs to compute all possible models to find SM.

Problem6 Although the microarrays and all SMs are LSD, statistical methods cannot find the linear separable facts. Book3 explains this reason.

Problem7 We must survey the categories of many SMs and explain the relation of SM and BGS. This theme will be explained by the next book (Book4).

Fact3 Because the six microarrays are LSDs and the two classes are completely separated in the high-dimensional gene space, LSD is an important signal for cancer gene diagnosis.

Fact4 Only RIP and Revised LP-OLDF can decompose six microarrays into several SMs (signal, MNM = 0) and noise subspace (MNM> = 1). H-SVM cannot find SM.

Fact5 All SMs are small samples, but not all statistical methods can show linearly separable signs for all SMs. Only logistic regression can correctly discriminate all SMs, and all NMs are empirically zero because it is free from Fisher's assumption.

Problem6 RatioSV of many RipDSs is large and easy to discriminate two classes correctly, but statistical methods other than logistic regression are utterly useless. We discuss this reason in Chap. 4. RipDS data gives a hint in this chapter. It seems that the signal found by the RIP and logistic regression may have small variations that are hidden by massive variations of noise. Problem6 is the second reason why many researchers could not find useful meaning in microarrays

Chapter 1
New Theory of Discriminant Analysis and Cancer Gene Analysis

Abstract This chapter explains the “New Theory of Discriminant Analysis after R. Fisher (Theory)” and the first success of cancer gene analysis as its application (Problem 5). The theory consists of four Optimal Linear Discriminant Functions (Optimal LDFs, OLDFs), two facts of discriminant analysis, two methods, and two statistics such as MNM and RatioSV. Section 1.1 summarises the theory and explains new results. Section 1.2 explains two facts as follows: (1) the relation of NM and LDF coefficient that solves Problem 1 (the defect of NM). (2) MNM monotonic decrease that is important for Problem5. Furthermore, we explain the reason why statisticians and machine learning researchers could not solve the cancer gene analysis since 1970. Only RIP and Revised LP-OLDF can decompose microarrays into many SMs. This fact is vital for cancer gene diagnosis. Section 1.3 introduces five severe problems of discriminant analysis. Section 1.4 introduces four OLDFs and three SVMs in addition to statistical discriminant functions. Section 1.5 explains the Matryoshka feature selection method (Method2) that solves Problem5 completely. Section 1.6 describes how to validate Method2 by two common data such as Swiss banknote data and Japanese car data those are LSD. Thus, this section indicates Method2 is useful for LSD including the common data and microarrays. Section 1.7 is the conclusion. We can explain the reason why only RIP and Revised LP-OLDF can decompose the microarray into many SMs. This reason is the answer why statisticians and machine learning researchers could not solve the cancer gene analysis since 1970.

Keywords Microarrays · Cancer gene analysis · Matryoshka feature selection method (Method2) · Small Matryoshka (SM) · Revised IP-OLDF (RIP) · Minimum number of misclassifications (MNM) · Relation of NM and LDF (Fact1) · Monotonic Decrease of MNM (Fact2)

1.1 Introduction

We found five serious problems of discriminant analysis (Shinmura 2014a, 2015c, d) through our discriminant analysis research after 1973 (Shimizu et al. 1975; Shinmura et al. 1973, 1974, 1983, 1987; Nomura and Shinmura 1978; Shinmura and

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_1

Miyake 1979; Shinmura 1984, 2001). We developed the new theory of discriminant analysis (Theory; Shinmura 2016d) that consists of four optimal linear discriminant functions (optimal LDFs, OLDFs) using mathematical programming (MP), two methods, and two statistics such as the minimum number of misclassifications (minimum NM, MNM) and RatioSV after 1997. Theory solved five problems of discriminant analysis completely. Especially, IP-OLDF and Revised IP-OLDF (**RIP**) defined by integer programming (IP) are very important LDFs based on **MNM** criterion (Shinmura 1998, 2000a, b, 2003, 2004, 2005, 2007a, b, 2011a; Shinmura and Tarumi 2000). All LDFs, except for IP-OLDF and RIP, use the **NM** that has many defects (Problem1). On the other hand, MNM is the best statistic in the discriminant analysis instead of NM. Let us consider the two-class discrimination of data with n cases and p variables. Because the formulation of IP-OLDF is shown on the p-discriminant coefficient space by fixing discriminant intercept = 1, it can reveal the relation of NM and LDF coefficient clearly (Fact1) and introduce MNM that is important statistics for LSD (Fact2). However, IP-OLDF cannot find a right vertex of an optimal convex polyhedron (optimal CP, OCP) if data does not satisfy the general position (Shinmura 2000a). Thus, we developed RIP that looked for the interior point of OCP. Only RIP can solve Problem1. Because Revised LP-OLDF is weak for Problem1, it cannot find all SMs from the microarray explained in Chap. 4. Moreover, other NMs and error rates of LDFs may not be correct. Only RIP and hard-margin SVM (H-SVM; Vapnik 1995) can discriminate LSD theoretically. Thus, statistical discriminant functions based on the variance–covariance matrix are useless for LSD-discrimination, especially cancer gene analysis using microarrays (Problem2). This is the reason why researchers could not solve the cancer gene analysis since 1970. Although the generalized inverse matrix has a fatal defect (Problem3), Problem3 is not important for Problem5. Because Fisher never defined the standard error of discriminant coefficient and error rate (Problem4; Miyake and Shinmura 1976), we developed the 100-fold cross-validation for a small sample (Method1; Shinmura 2013; 2014c; 2015a). Although most cancer gene researchers validated their results by the leave-one-out (LOO) procedure (Lachenbruch and Mickey 1968), we do not validate our results by Method1 because two classes are completely separable in the microarrays and all SMs. RIP and two methods solved five problems. Especially, RIP and the Matryoshka feature selection method (Method2) solved cancer gene analysis as its application in 2015 (**Problem5**).

Section 1.2 explains two new facts of discriminant analysis as follows:

(1) The relation of NM and LDF coefficient that solves Problem1 (the defect of NM).
(2) MNM monotonic decrease that is important for Problem5.

We explain the reason why statisticians and machine learning researchers could not solve the cancer gene analysis since 1970. Furthermore, only RIP and Revised LP-OLDF can decompose microarrays into many SMs.

Section 1.3 summarizes five severe problems of discriminant analysis and three difficulties of Problem5. Section 1.4 explains four OLDFs and three SVMs in addition to statistical discriminant functions.

Section 1.5 explains Method2 that solves Problem5 entirely and introduces the RatioSV. Section 1.6 describes how to validate Method2 by two common data such as Swiss banknote data (Flury and Riedwyl 1988) and Japanese car data (Shinmura 2016c) those are LSD. Thus, this section indicates Method2 is useful for LSD including the common data and microarrays. Section 1.7 is the conclusion.

In this chapter, we can explain the reason why only RIP and Revised LP-OLDF can decompose the microarray into many SMs. This result is the answer why researchers could not solve the cancer gene analysis since 1970 (Shinmura 2018a, b).

After all, gene analysis of microarray was possible by discriminant analysis defined by mathematical programming (MP). Since Chap. 2 and later, it was possible to divide the microarray into many small subspaces (SMs) that are easy to handle, so we propose various approaches for genetic diagnosis of cancer by statistical analysis.

1.2 Fundamental of Theory

1.2.1 The Motivation of Our Research

Although we developed a diagnostic logic of Electrocardiogram (ECG) data by Fisher's LDF and the quadratic discriminant function (QDF) from 1971 to 1974, our research was inferior to the decision tree logic developed by the medical doctor (Shinmura et al. 1973, 1974). After this experience, we concluded these discriminant functions are not adequate for the discrimination of the normal and abnormal subjects because of two main reasons as follows:

(1) There are many patients (cases) nearby the discriminant hyperplane. The doctor is trying to discriminate the case (patient) near the discriminant hyperplane. Exam scores and rating data have the same characteristic. Most statisticians do not understand our claim because they are not interested in real data analysis.

(2) If the value of some variable increases or decreases continuously, the probability of belonging to abnormal disease increases from 0 to 1. Fisher's LDF assumes the typical abnormal patients are the average of the abnormal class. However, typical cases of patients are far from healthy subjects. We proposed "Earth Model" in medical diagnosis (Shinmura 1984).

(3) The normal group is the land, and the abnormal group is the mountain range. The discriminant hyperplane is the horizontal line. While many patients locate near the horizon, a typical patient is at the summit. Taguchi method (Taguchi and Jugular 2002) was one of multi-class discrimination by Mahalanobis distance based on the variance–covariance matrices. Our claim is the same perception as Taguchi theory. Although the pass/fail determinations using exam score data are LSD, we observed several error rates are over 20%. These results are caused by many pass students nearby the discriminant hyperplane obtained by Fisher's assumption. Since these data do not satisfy the Fisher hypothesis, the hyperplane obtained based on the normal distribution does not coincide with the actual

distribution, and it misclassifies many successful applicants. Although this book illustrates the two classes are separable in SMs, the error rates of SM by statistical LDF are very high.

(4) If some independent variable of logistic regression increases or decreases, the probability "p" belonging to class1 (normal symptom) increases from 0 (class1) to 1 (class2). This way is suitable for medical diagnosis and is the same as our claim. Moreover, the maximum likelihood method developed by Fisher solves logistic regression coefficients (Cox 1958). It finds the coefficients that fit the real data and can almost discriminate LSD correctly. Thus, most Japanese medical researchers use logistic instead of Fisher's LDF empirically. Because JMP does not support logistic regression for high-dimensional microarrays, we cannot discriminate the microarrays by logistic regression. Even if the logistic regression could discriminate the microarray into two classes, most of the coefficients are not zero like H-SVM, as inferred from common data discrimination results.

(5) Many statisticians focus on RDA (Friedman 1989) and LASSO (Simon et al. 2013) based on the variance–covariance matrix, but these discriminant functions are not suitable for medical diagnosis, especially cancer gene analysis. With these methods, we cannot distinguish between two classes like Fisher's LDF. Cox extends the new frontier of second-generation discriminant analysis, and logistic regression is the best way of statistical LDF. Also, Vapnik opened a new boundary for MP-based LDF after the first generation summarized by Stam (1997). He disseminated SVM in the field of pattern recognition and avoided the research area on statistics and operations research (OR). He was able to avoid statisticians and OR researchers from ignoring SVM in the same way as us and refusing. H-SVM and kernel SVM apply to many kinds of data. However, as far as we know, there are few comparisons with other LDFs, and H-SVM clearly defines LSD-discrimination, but there is no research of LSD-discrimination.

1.2.2 IP-OLDF Based on MNM Criterion and Two Facts

Above ECG failure was our motivation to develop theory. After many experiences of the discriminant analysis, we developed IP-OLDF based on MNM criterion in Eq. (1.1) (Shinmura 1998). Because we fix the intercept of IP-OLDF to one, we can define it in the p-dimensional coefficient space instead of (p + 1)-dimensional space. Although $\left({}^{t}\mathbf{x}_{i}\mathbf{b} + 1\right)$ is a discriminant score (DS), we use the extended DS such as $y_{i} * \left({}^{t}\mathbf{x}_{i}\mathbf{b} + 1\right)$. It is a linear hyperplane and divides discriminant space into two half-planes such as plus half-plane $\left(y_{i} * \left({}^{t}\mathbf{x}_{i}\mathbf{b} + 1\right) > 0\right)$ and minus half-plane $\left(y_{i} * \left({}^{t}\mathbf{x}_{i}\mathbf{b} + 1\right) < 0\right)$. If we choose $\mathbf{b}_{k}$ in plus hyperplane as LDF coefficient, LDF such as $y_{i} * \left({}^{t}\mathbf{b}_{k}\mathbf{x}_{i} + 1\right)$ discriminates $\mathbf{x}_{i}$ correctly because of $y_{i} * \left({}^{t}\mathbf{b}_{k}\mathbf{x}_{i} + 1\right) = y_{i} * \left({}^{t}\mathbf{x}_{i}\mathbf{b}_{k} + 1\right) > 0$. On the other hand, if we choose $\mathbf{b}_{k}$ in minus hyperplane, LDF misclassifies $\mathbf{x}_{i}$ because of $y_{i} * \left({}^{t}\mathbf{b}_{k}\mathbf{x}_{i} + 1\right) = y_{i} * \left({}^{t}\mathbf{x}_{i}\mathbf{b}_{k} + 1\right) < 0$. However, we must solve the other two models such as the intercept $= -1$ and 0. Setting the intercept as arbitrary positive real value is the similar result obtained by intercept $= 1$.

It looks for the right vertex of an OCP if data is a general position. There are only p-cases on the discriminant hyperplane, and it becomes the vertex of correct OCP. On the other hand, if data is not general position, it may not look for the right vertex of OCP because there are over (p + 1) cases on the discriminant hyperplane, and we cannot correctly discriminate these cases (Problem1). Thus, we developed RIP that looks for the interior point of right OCP in Eq. (1.4). Equation (1.1) defines IP-OLDF based on MNM criterion after the heuristic approach (Miyake and Shinmura 1980). The e_i is a 1/0-binary integer, and IP defines IP-OLDF. If the extended DS classifies the case $\mathbf{x}_i$ correctly, the e_i becomes zero. Otherwise, if the extended DS misclassifies the case $\mathbf{x}_i$, it becomes one. Thus, the minimum value of the objective function is the minimum NM. If data is LSD, NM becomes MNM = 0. It looks for the vertex of a correct OCP if data is a general position. There are only p-cases on the discriminant hyperplane, and it becomes the exact vertex of OCP. However, if data is not the general position and the vertex consist of over (p + 1) cases, the vertex may not be the correct vertex of OCP. Because all the LDFs correctly discriminate just the same case, these LDFs are equivalent. There are only a finite equivalent LDFs corresponding to a limited number of CPs. Thus, all interior points of each CP correspond to each equivalent LDF.

$$\begin{aligned} &\mathrm{MIN} = \Sigma\, e_i; \\ &y_i * \left({}^t\mathbf{x}_i\mathbf{b} + 1\right) >= -\ M * e_i; \end{aligned} \tag{1.1}$$

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.
$\mathbf{b}_i$ discriminant coefficients.
M big M constant such as 1,000.

If we exchange $\mathbf{x}_i$ and $\mathbf{b}_i$, we understand IP-OLDF on data space. This model indicates two relevant facts as follows.

(1) Fact1: We explain the notation of IP-OLDF by the Golub microarray. 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) correctly. The 72 linear hyperplanes, the 7,129 coefficients of those are values of each case, divide the 7,129-dimensional 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. Thus, because the internal points of each CP have unique NM, we can define the OCP with MNM. Many examinations show the best models (Shinmura 2016b) of RIP are better than other seven LDFs by Method1.

(2) Fact2: Let MNM_k be MNM in the k-dimensional subspace. MNM decreases monotonously ($\mathrm{MNM}_k >= \mathrm{MNM}_{(k+1)}$). If $\mathrm{MNM}_k = 0$, all MNMs including these k variables (genes) are zero. We call the minimum-dimensional SM as Basic Gene Sets (BGS) that is as same as the Yamanaka's four genes in iPS research. Yamanaka's three genes do not produce iPS cell. If we drop one gene

from BGS, MNM is greater than one. We consider "MNM = 0" equals to produce iPS cell. This fact tells us BGS can completely describe the Matryoshka structure of gene space by monotonic decreases of MNM.[1] "MNM monotonic decrease" is the same idea as the Matryoshka structure of microarray. However, iPS cell does not have the characteristics of "MNM monotonic decrease." When we discriminate all possible models of Swiss banknote data [Shinmura 2016d (Chap. 6)], IP-OLDF finds a two-variable model, such as (X4, X6), is unique BGS. By the monotonic decrease of MNN, 16 MNMs including BGS are zero among 63 models ($=2^6 - 1 = 63$). Other 47 MNMs are greater than one. We can define 16 models as the signal, and 47 models are noise. However, there was no precise definition of signal and noise until now because researchers did not know microarrays are LSD and have Matryoshka structure. This book claims that the 16 models are signals, and the other 47 models are noise in LSD-discrimination. Because we had already LSD-discrimination study by common data, we could solve the cancer gene analysis in 54 days in 2015.

1.2.3 Simple Example

IP-OLDF can explain the relation of NMs and discriminant coefficients clearly (Fact1). Let us consider the discrimination using three cases and two variables ($n = 3$, $p = 2$) as follows:

Class1 ($y_1 = 1$): case1 = ($-1/18$, $-1/12$)
Class2 ($y_2 = -1$, $y_3 = -1$): case2 = (-1, $1/2$), case3 = ($1/9$, $-1/3$).

Equation (1.2) defines the model of IP-OLDF (Shinmura 2000a, b). We need to be aware that $y_2 = y_3 = -1$. To multiply y_2 and y_3 changes the signs of case2, case3, and constant. This role of y_i aligns the inequality signs with case1 (Class1). However, Schrage (2006) proposes the LDF that does not change the sign of data in Chap. 10.

$$\begin{aligned}
&\mathrm{MIN} = \Sigma\, e_i;\\
&y_1 * \{-(1/18) * b_1 - (1/12) * b_2 + 1\} >= -\ e_1;\\
&y_2 * \{-b_1 + (1/2) * b_2 + 1\} >= -\ e_2;\\
&y_3 * \{(1/9) * b_1 - (1/3) * b_2 + 1\} >= -\ e_3;
\end{aligned} \tag{1.2}$$

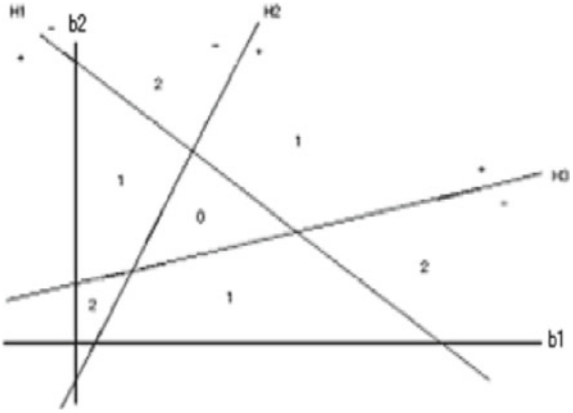
We consider three linear equations in Eq. (1.3) from the three constraints of Eq. (1.2).

$$\begin{aligned}
&H_1 = -(1/18) * b_1 - (1/12) * b_2 + 1 = 0,\\
&H_2 = b_1 - (1/2) * b_2 - 1 = 0,\\
&H_3 = -(1/9) * b_1 + (1/3) * b_2 - 1 = 0
\end{aligned} \tag{1.3}$$

[1] Chapter 6 (Shinmura 2016d) explains this problem in detail.

Three linear equations divide the two-dimensional coefficient space into seven CPs in Fig. 1.1. The number of CP is NM of each LDF that equals the number of minus half-planes of $H_i(\mathbf{b})$ that surrounds CP. The interior point in the triangle locates in three-plus hyperplanes, NM of which is zero and MNM. This triangle becomes a feasible region and optimal CP (OCP) for IP-OLDF. This OCP has a special feature because all interior points have the same values such as MNM = 0. Thus, all points of OCP become the optimal solutions. Because two linear equations make three vertexes of OCP, this data is general position and free from Ploblem1. Because n = 3 and p = 2, the two linear equalities (p = 2) chosen from three cases (n = 3) make three intersections of the feasible region. In the common data (n > p), the dimension of a vertex is less than p. In the high-dimensional data such as microarrays ($n \ll p$), this graph explains the dimension of a vertex is less than n. This fact indicated RIP and Revised LP-OLDF defined by LP can find SM having less than n genes at the first iteration (Shinmura 2018a, b). NMs of three opposite CPs of OCP are one. Namely, NMs of adjacent CPs differ by 1. Although we fix the constant to one, we must solve three models as follows: the constant = 1, the constant = −1, and the constant = 0, because we cannot decide the sign of discriminant score (DS) a priori. When we fix the constant = 2, Fig. 1.1 is similarly enlarged to twice.

Fig. 1.1 Relation of NM and discriminant coefficient



1.2.4 Ordinary LP Solution

We explain the ordinary MP model by Eq. (1.3′). Change the objective function to "MIN = Σ 2 * b_1 + 3 * b_2;" Fix the three binary variables to zeros such as $e_i = 0$ (i = 1, 2, 3). This model consists of n = 3 and p = 2 (n >= p). The feasible region of b_1 and b_2 is triangles in Fig. 1.1.

$$\begin{aligned}\mathrm{MIN} &= 2 * \mathrm{b}_1 + 3 * \mathrm{b}_2;\\ &(1/18) * \mathrm{b}_1 + (1/12) * \mathrm{b}_2 + 1 >= 0;\\ &-\mathrm{b}_1 + (1/2) * \mathrm{b}_2 + 1 =< 0;\\ &(1/9) * \mathrm{b}_1 - (1/3) * \mathrm{b}_2 + 1 =< 0;\end{aligned} \tag{1.3$'$}$$

The optimal solution is $(b_1, b_2) = (3, 4)$ at the intersection with the second and third constraints. The minimum value is 18. In this way, the LP solution usually selects one of the vertexes of the OCP obtained by selecting p constraints from n constraints. On the other hand, because the gene data is $(p \gg n)$, the intersection becomes at most n simultaneous equations by setting (p−n) genes to zero. If we list up all the candidates and assign them to the objective function, the brute force method obtains the minimum solution, also. However, the simplex method of LP algorithm can do this efficiently. In the IP model, the executable area of the LP model is limited to the integer variable specified. However, because the e_i do not affect the feasible region, RIP and IP-OLDF have the same feasible region as Revised LP-OLDF. Thus, RIP and Revised LP-OLDF can find the optimal subspace having less than same n genes at the first step of Method2 and decompose microarrays into many SMs (new Fact4). On the other hand, three SVMs find the only one optimal SVMs on the whole domain. These facts are the reason why no statisticians and machine learning researchers could not solve the cancer gene analysis since 1970 (Shinmura 2018a, b).

1.3 Five Serious Problems and Three Excuses

We found five severe problems and two facts of discriminant analysis (Shinmura 2016d). Moreover, theory solved five problems introduced in Chap. 1. After Chap. 2, we discuss two new problems and two new facts of cancer gene diagnosis.

1.3.1 Four Problems

Problem1[2]

The discriminant rule is straightforward. However, most researchers believe in the wrong law. Even now, they think that the following law is correct.

(1) If $f(\mathbf{x}_i) >= 0$ and $\mathbf{x}_i$ belongs to class1, $\mathbf{x}_i$ belongs to class1 correctly (TP). Otherwise, if $f(\mathbf{x}_i) >=0$ and $\mathbf{x}_i$ belongs to class2, $\mathbf{x}_i$ is misclassified to class1 (FP).

[2] Chapter 1 (Shinmura 2016d) explains this problem in detail.

(2) If f($\mathbf{x}_i$) < 0 and $\mathbf{x}_i$ belongs to class2, $\mathbf{x}_i$ is classified to class2 correctly (TN). Otherwise, if f($\mathbf{x}_i$) < 0 and $\mathbf{x}_i$ belongs to class1, $\mathbf{x}_i$ is misclassified to class2 (FN).
(3) NM is defined by (FP + FN).

They misunderstand as following two points:

(1) Generally speaking, it is not possible to determine to which class the case on the discriminant hyperplane belongs.
(2) If f ($\mathbf{x}_i$) >= 0 by statistical LDF, we cannot decide that $\mathbf{x}_i$ belongs to class1 a priori.

The values of y_i are 1 for class1 with a symbol □ and −1 for class2 with symbol × in the graphs of this book. The y_i is the objective variable of the regression model also. Let f($\mathbf{x}$) be LDF and y_i * f($\mathbf{x}_i$) be an extended discriminant score (DS) for $\mathbf{x}_i$. In the MP-based LDF, the following rule is correct.

(1) If y_i * f($\mathbf{x}_i$) > 0, $\mathbf{x}_i$ is classified to class1/class2 correctly.
(2) If y_i * f($\mathbf{x}_i$) < 0, $\mathbf{x}_i$ is misclassified to class1/class2 correctly.
(3) We cannot properly discriminate $\mathbf{x}_i$ on the discriminant hyperplane (f($\mathbf{x}_i$) = 0). Many researchers ignored this unsolved problem until now. Fact1 found by IP-OLDF solved this Problem1 completely.

Only RIP can treat this Problem1 appropriately. Indeed, except for RIP, no LDFs can count the NMs correctly. These LDFs should count the number of cases where f(x_i) = 0 and display this figure h alongside the NM in the output. The correct NM may increase up to h.

Student data[3] (Shinmura 2010a) tells us the defect of IP-OLDF caused by Problem1. Thus, we develop RIP. RIP looks for the interior point of OCP. Only the internal points of CP avoid the cases on the discriminant hyperplane explained by Fact1. The vertex and edge of CP have over p-cases on the discriminant hyperplane. If another LDF corresponds to the vertex of CP, it cannot avoid Problem1. Indeed, except for RIP, no LDFs can count the NMs correctly because these LDFs may choose the vertex or edge of CP.

Problem2[4]
Only H-SVM and RIP can recognize LSD theoretically. Although many statisticians and users use NM, without doubt, LSD-discrimination reveals NM is not reliable because each LDFs have different NM and we get different NM by the change of discriminant hyperplane. Experimentally, Revised LP-OLDF discriminates LSD correctly. Nevertheless, it tends to collect cases on the discriminant hyperplane (**Problem1**). If we discriminate exam scores by two testlets score such as T1 and T2, and the pass mark is 50 point, we can obtain a trivial LDF such as f = T1 + T2 − 50 [Shinmura 2015b, 2016d (Chap. 5)]. Although these data are LSD, NMs

[3] Chapter 4 (Shinmura 2016d) explains this problem in detail.

[4] Chapters 4–8 (Shinmura 2016d) explains this problem in detail.

or error rates are very high. Furthermore, seven LDFs,[5] except for Fisher's LDF, become the same trivial LDFs if we divide all coefficients by the intercept value and fix the intercept = 1. We can judge the student pass the exam if $\mathrm{f}(\mathbf{x}_\mathrm{i} : \mathrm{T1}, \mathrm{T2}) >= 0$[6] and fail the exam if $\mathrm{f}(\mathbf{x}_\mathrm{i}) < 0$. However, error rates of Fisher's LDF and QDF based on variance–covariance matrices are very high because exam scores do not satisfy Fisher's assumption (Shinmura 2016c). Thus, these LDFs are useless for important applications such as cancer gene analysis in addition to medical diagnosis, pattern recognition, and different rating.

Problem3[7]

Problem3 is the defect of the generalized inverse matrix technique. When we discriminated math exam scores by QDF and RDA, all successful students were misclassified in the failed class because all successful students correctly answered a few questions and the answers of failed students are scattered. In this case, if we add slight random noise to the constant values, we can solve Problem3 completely. Moreover, discriminant functions based on the variance–covariance matrix cannot correctly discriminate LSD, such as many SMs in addition to Swiss banknote data, Japanese car data, pass/fail determination using examination scores, and student linearly separable data.

Problem4[8]

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 the LOO procedure in the discriminant analysis, we propose Method1. It offers the 95% CIs of error rates and discriminant coefficients. Moreover, it provides a simple and powerful model selection procedure such as the best model. We confirmed the best models of RIP were better than Fisher's LDF, logistic regression, and five MP-based LDFs using six common data. JMP script of Fisher's LDF and logistic regression discriminates these data. SAS Institute Japan Ltd. JMP Japan Division supported us to develop the script of Method2. LINDO Systems Inc (Schrage 1991, 2006) supported six MP-based LDFs by LINGO Program2.[9] We can establish the theory by JMP and LINGO in 2015.

[5] We compare eight LDFs such as three OLDFs, three SVMs, logistic regression, and Fisher's LDF in addition to QDF.

[6] Since the discrimination rule is defined with two scores, the path class contains an equal sign.

[7] Chapter 7 (Shinmura 2016d) explains this problem in detail.

[8] Chapters 2–7 (Shinmura 2010a, 2016d) explains this problem in detail.

[9] Chapter 9 (Shinmura 2016d) explains this problem in detail.

1.3.2 Problem5[10]

Since 1970, many researchers were struggling to select cancer gene features from the microarrays. However, when we discriminate six microarrays by three Revised OLDFs, MNM, and two NMs are zero, and most of the coefficients become zero (Fact3). This fact implies three OLDFs can select cancer genes naturally without feature selection method. Moreover, we established Method2 within 54 days and found all SMs of six microarrays (Fact4). On the other hand, we spent three years to solve Problem3. This comparison tells us that the theory is most suitable for cancer gene analysis because we study LSD-discrimination by common data. It finds the surprising fact the microarrays consist of the disjoint unions of many SMs and noise gene subspace (MNM > 0). Until now, many researchers analyze high-dimensional gene space with noise directly by standard statistical methods and could not obtain the meaningful results. Everybody can analyze each SM by these methods now. Those are one-way ANOVA, t-test, correlation analysis, hierarchical cluster analysis, principal component analysis (PCA), QDF, logistic regression, and Fisher's LDF (Fisher 1936, 1956)

1.3.3 Three Excuses of Cancer Gene Analysis

Since 1970 (Golub et al. 1999), many researchers are struggling to select a cancer gene (**Problem5**). They pointed out the three difficulties (or excuses) about cancer gene analysis as follows:

(1) Small n and large p data (Diao and Vidyashankar 2013, Buhlmann and Geer 2011)

To estimate the variance–covariance matrices for small n and large p was difficult for statistical discriminant functions based on the variance–covariance matrices. On the other hand, MP-based LDFs are free from this difficulty. Sall (1981) announced Fisher's LDF for high-dimensional microarrays by the singular value decomposition (SVD) at the Discovery Summit in Tokyo, November 2015. However, when Fisher's LDF discriminated the microarrays, six NMs were not zero in Table 1.1. This fact is the defect of discriminant analysis based on variance–covariance matrices. In addition, discriminant functions by maximizing the correlation ratio could not discriminate LSD correctly for many data. Only two standards such as the maximization of SV distance by H-SVM and RIP based on MNM criterion can discriminate LSD theoretically. Moreover, six MP-based LDFs discriminate the small n and large p data compared with the large n and small p data easily.

[10] Chapter 8 (Shinmura 2016d) explains this problem in detail.

(2) Statistical feature selection is NP-hard (Charikar et al. 2000)

In general speaking, it is challenging to select proper cancer genes for large p variables by statistical discriminant functions. Many statisticians do not understand that "feature selection" is finding one of the optimal solutions of subspaces from p-dimensional space. In general, the statistical discriminant function finds one LDF on the p-dimensional domain. To see the optimum LDF of the subspaces needs a stepwise variable selection procedure or all possible models (Goodnight 1978) to search all optimal subspaces such as SMs. If p is gigantic, it will surely be NP-hard. However, the LDF formulated with LP and IP can find one of the optimum OLDFs of subspaces as explained in Sect. 1.2.1. However, because the SVM expressed by QP finds the only SVM's coefficient to maximize/minimize the objective function in the p-dimensional space, it needs feature selection as same as the statistical discriminant functions (NP-hard).

Because the microarrays are LSD, six MP-based LDFs can discriminate the microarrays within 10 s. B and B algorithm of LINGO IP solver has the same algorithm to compute all possible models. Thus, RIP can find SM among the massive number of MNM = 0 on gene subspaces. Moreover, Revised LP-OLDF defined by LP finds the vertex of the feasible region made by n constraints in the first step (Shinmura 2018a, b). The vertex is an intersection point of n constraint equations and corresponds to a subspace of p-dimension ($p \gg n$).

On the other hand, SVM can find only one optimal SVM for high-dimensional gene space (whole domain) and cannot find SM that is one of the small gene subspaces. Because QP defines SVM, it can see only one minimum or maximum optimal solution of the quadratic objective function. Thus, SVM must compute all possible models for large p genes. This computation is NP-hard because there is no efficient algorithm such as the B and B algorithm. Revised LP-OLDF can decompose the microarrays into another combination of SMs as same as RIP. However, it cannot find all SMs from the microarray introduced in Chap. 4.

(3) The signal is buried in noise (Brahim and Lima 2014)

This difficulty is unclear because the definition of signal and noise is not defined explicitly until now. Alternatively, it may only refer to the signal/noise ratio used in the paper by Golub et al. We claim the gene sets included in each SM or BGS is cancer gene set and signal because two classes are entirely separable in SM. Until now, because no researchers knew this critical Fact3, they could not define oncogenes of microarrays clearly. We can decompose signal subspace into many SMs (Fact4). Thus, RIP can separate the microarrays into signal and noise naturally. Also, RIP decomposes signal subspace into many SMs that are small signals. Although to measure microarray is expensive, to measure SM saves the expense for cancer gene diagnosis if we can decide the best SM for cancer diagnosis (Shinmura 2017a, b, c) that is the future work (Problem7). In this book, we solve the following Problem6 explained after Chap. 3.

Problem6: Why can no researchers find the linear separable facts in SM since 1970?

We could solve five problems and found two facts. In this book, we discuss two new problems and two new facts. Only RIP, Revised LP-OLDF, and H-SVM find the microarrays are LSD (Fact3), and RIP and Revised LP-OLDF can decompose the microarrays into many SMs (Fact4). At first, because all SMs are small samples, the standard statistical methods can easily find the linear separable facts that two classes are entirely separable in all SMs. However, only RIP, Revised LP-OLDF, and logistic regression can find the linear separable facts. From Chaps. 3–9 examine several approaches and explain the several reasons for Problem6.

1.4 Four OLDFs and MNM Instead of NM

We developed four OLDFs, two facts, two methods, and two statistics such as MNM and RatioSV. Those LDFs are IP-OLDF, RIP, Revised LP-OLDF, and Revised IPLP-OLDF (Shinmura 2010b, 2014b) that is the mixture model of Revised LP-OLDF and RIP. Thus, we do not focus on Revised IPLP-ODF in this book. IP-OLDF found two facts about LDF. Those are (1) the relation of NM and the LDF discriminant coefficients, and (2) MNM monotonous decreases. Two methods are Method1 and Method2. Six MP-based LDFs by LINGO Program2 and two statistical LDFs by JMP discriminate all possible models of 100 training samples of six different types of common data and compute the minimum means of 100 error rates (M1) in the 100 training samples. Because M1 decreases monotonously as same as MNM, M1s of the full model are always the minimum value. Thus, M1 is not proper for model selection statistic. Obtained LDFs are applied for the 100 validation samples and choose the best model with the minimum means of 100 error rates in the validation samples (M2). We confirmed the M2s of RIP are less than those of the other seven LDFs. Although some referees of Japanese statistical journal rejected heuristic OLDF based on MNM criterion 38 years ago because MNM criterion overfitted for the training samples, JSMEBE (Miyake and Shinmura 1980) accepted our paper later. Our best model results prove the former journal referees were wrong. However, we do not discuss Method1 in this book. We need not validate our findings by Method1 because two classes are entirely separable in SM. RIP and two methods solved five problems of discriminant analysis, and we established the theory in 2015. Notably, our theory is the most suitable for cancer gene analysis as its application. On the other hand, because all LDFs, except for RIP and H-SVM, cannot discriminate LSD theoretically, these LDFs are useless for cancer gene analysis of microarrays.

1.4.1 Revised IP-OLDF and the Defects of Number of Misclassifications

On the other hand, if data is not general position and there are over (p + 1) cases on the discriminant hyperplane, it may not look for the vertex of correct OCP and cannot discriminate these cases correctly. Thus, we developed RIP that looks for the interior point of right OCP in Eq. (1.4) directly.

$$\begin{aligned}&\mathrm{MIN} = \Sigma \mathrm{e_i};\\ &\mathrm{y_i} * \left({}^{t}\mathbf{x}_{i}\mathbf{b} + \mathrm{b_0}\right) >= 1 - \mathrm{M} * \mathrm{e_i};\end{aligned} \tag{1.4}$$

b_0 free decision variables.
M 10,000 (Big M constant).
e_i 1/0 binary integer.

Because b_0 is the free variable, RIP is defined in (p + 1)-dimensional coefficient space,[11] and we cannot understand the relation of NM and LDF found by IP-OLDF directly. 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 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. The binary decision variable chooses two alternatively. Thus, we expect DSs of misclassified cases are less than −1, and there are no cases within SV. SV is a window with length two that separate two classes completely for LSD. If M is a small constant, it does not work correctly (Shinmura 2010a). 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 in p-dimension space. All LDFs, except for RIP, cannot solve **Problem1** theoretically. Thus, these LDFs must check the number of cases (h) on the discriminant hyperplane. Correct NM may increase up to h. **Problem1** suggests us the severe defects of NM that is not a reliable statistic for the discriminant analysis as follows:

(1) Above fact shows NM is not a reliable statistic.
(2) Seven NMs, except for RIP, are often different. Moreover, MNM becomes the lower limit of all NMs. If the data satisfies Fisher's assumption, NM of Fisher's LDF decreases to MNM. Because there is no proper test statistic for Fisher's hypothesis, we can validate whether data satisfies it. If both values are similar, we can judge data fills it.
(3) We need to select one of the prior probabilities. The first option is proportional to 1:1 (Fisher's LDF1). The second option is equivalent to the case numbers (Fisher's LDF2). Both NMs are often different. In statistical meaning, the first option is better. However, because we must evaluate two statistical LDFs and six MP-based LDFs, we choose the latter Fisher's LDF2.
(4) Although LDF decides the discriminant hyperplane theoretically, we often choose better result by changing the discriminant hyperplane. In the logistic regression, we accept the minimum NM by changing the discriminant hyperplane on the receiver operating characteristic curve (ROC). If NM of logistic regression is zero and MNM = 0 confirmed by RIP, we judge logistic regression can discriminate LSD correctly. However, some statisticians and users do not trust logistic regression for LSD because of the defect pointed out by Firth (1993).

[11] In pattern recognition, (p + 1) dimensional space with the intercept defines LDF.

On the other hand, MNM is better than NM.

(1) Because MNM decreases monotonously, MNM of full model is always the minimum value. Moreover, M1 is always the minimum value. We propose the best model with M2 among all possible models. We examined the best models of RIP are better than other best models of seven LDFs using five different types of common data. Only the iris data (Anderson 1945) is almost the same as MNM.

(2) For iris data, the best model of Fisher's LDF is almost the same as that of RIP. Fisher has chosen the better test data to validate Fisher's LDF. In the discriminant analysis, many researchers evaluate their results using this data. However, it is not adequate for test data because it does not show the severe differences and consists of only four variables. In Sect. 10.2.5, other OLDF based on MNM criterion developed by Linus analyze this data. For a while, Japanese academic journals asked to create and verify training and verification samples with the normal random numbers. This request is an inaccurate request indicating the goodness of the method made by assuming a normal distribution.

(3) For Swiss banknote data, the two-variable model such as (X4, X6) is BGS. However, the best model is a five-variable model such as (X1, X3–X6) [Shinmura 2016d (Chap. 6)]. This truth suggests us BGS is not proper for cancer gene diagnosis explained in Chap. 2. We discuss this theme in Chap. 3. In Sect. 10.2.1, Linus OLDF analyzes this data.

(4) Japanese statistical referee rejected our paper about a heuristic OLDF based on MNM criterion. He claimed MNM was the foolish discriminant criterion and overestimated the training samples. However, the medical journal published our paper (Miyake and Shinmura 1980) because the referees knew the real data examination. The results of the best model proved the first claim was wrong after 38 years later. Another referee rejected our paper in 2015. He claimed the purpose of the discriminant analysis is to discriminate the overlapping data, not LSD. However, he could not distinguish whether data is LSD or overlap because he could not judge by "MNM = 0 or MNM >=1." The reason why many researchers could not be successful in cancer gene analysis is the lack of knowledge of MNM.

(5) Moreover, we showed several error rates of Fisher's LDF were very high for LSD-discrimination. If medical researchers abandoned their research because of high error rates, they have better reviewed their studies with RIP because they can obtain a smaller error rate by RIP. RIP can solve Problem1 and Problem2. Moreover, because it can naturally select features for common data and the microarrays, it can explain Problem5. However, we develop more powerful model selection procedure such as the best model. Thus, we had ignored the natural feature selection for common data before Method2.

1.4.2 Revised LP-OLDF and Revised IPLP-OLDF

If we change the 1/0 binary integer variable e_i to a nonnegative real variable, RIP changes to Revised LP-OLDF in Eq. (1.5). LP solves this model that is faster than RIP. Because it tends to collect several cases on the discriminant hyperplane, we recommend not to use it for the overlapping data because NM is not often correct (Problem1). However, it can decompose the microarrays into another different combination of SMs. We examine these SMs compared with SMs obtained by a RIP from Chaps. 3–9.

$$\begin{aligned} &\mathrm{MIN} = \Sigma e_i; \\ &y_i * ({}^t\mathbf{x_i b} + b_0) >= 1 - M * e_i; \end{aligned} \tag{1.5}$$

e_i nonnegative real values

Revised IPLP-OLDF is a mixture model of Revised LP-OLDF and RIP. In the first step, Revised LP-OLDF discriminates all cases. In the second stage, RIP discriminates the restricted cases fixing $e_i = 0$ for classified cases in the first step. In this book, we do not focus on Revised IPLP-OLDF precisely.

1.4.3 Hard-Margin SVM (H-SVM)

Vapnik proposed three different SVM models. H-SVM indicates the discrimination of LSD clearly. IP-OLDF confirms that Swiss banknote data is LSD and realizes the importance of Problem2. H-SVM adapted the maximization of the SV distance to obtain an excellent k-variable model with good generalization ability, which is similar to “not overestimating the validation data” in statistics. It is redefined to minimize (1/distance of SV) in Eq. (1.6). H-SVM can discriminate the only LSD, not overlapping data. This restriction might ignore the research of LSD-discrimination. Some statisticians erroneously believe that LSD-discrimination is easy. In statistics, there was no technical term for LSD before H-SVM. However, the condition “MNM = 0” is the same as being linearly separable. Note that “NM = 0” does not imply that the data is linearly separable. Also, because the correct NM may be higher than the obtained NM, NM is not a reliable statistic. It is unfortunate that there has been no research on LSD-discrimination for Problem5. Thus, many researchers cannot select cancer genes naturally. We guess LASSO cannot discriminate LSD correctly as same as Fisher’s LDF, also. Although H-SVM can discriminate the microarrays accurately, it cannot find SM because of QP.

$$\mathrm{MIN} = ||\mathbf{b}||^2/2; y_i \times \left({}^t\mathbf{x_i b} + b_0\right) >= 1; \tag{1.6}$$

b: p-discriminant coefficients. b_0: H-SVM constant

1.4.4 Soft-Margin SVM (S-SVM)

Because real data is rarely LSD, most users use S-SVM defined in Eq. (1.7). S-SVM permits certain cases that are not discriminated by SV ($\mathrm{y_i} \times ({}^t\mathbf{x_i b} + \mathrm{b_0}) < 1$). The second objective is to minimize the summation of distances of misclassified cases ($\Sigma \mathrm{e_i}$) from SV. The penalty c combines two objects. Revised LP-OLDF minimizes the summation of misclassified distance from the discriminant hyperplane as same as the second objective function in Eq. (1.7). This fact is crucial for cancer gene analysis because Revised LP-OLDF can select SM. On the other hand, H-SVM and SVM4 cannot select SM. The Markowitz portfolio model (Markowitz 1959) that minimizes risk and maximizes return is the same as S-SVM. Moreover, because all NMs of SVM1 are often not zero, SVM1 is not used for cancer gene analysis. However, the return is a constraint, and the objective function minimizes the only risk. The decision maker selects a solution on the efficient frontier. On the contrary, S-SVM does not have the rule to determine a proper c as same as RDA; nevertheless, an optimization solver solves it. Thus, we compare two S-SVMs, such as SVM4 (c = 10,000) and SVM1 (c = 1). In many trials, NMs of SVM4 are less than NMs of SVM1. We claim the methods with tuning parameters such as S-SVM and RDA are useless for statistical users because they must pay their efforts to select the best parameters for each data. On the other hand, although RIP must set the big M constant, we confirmed M = 10,000 (or 1000) causes good results using six different types of common data and all possible models. We surveyed and investigated to change the value M from c = 0.1, 1, 10, 100, 10^3, 10^4 and 10^6 (Shinmura 2010a).

$$\begin{aligned} \mathrm{MIN} &= ||\mathbf{b}||^2/2 + \mathrm{c} \times \Sigma \mathrm{e_i}; \\ \mathrm{y_i} \times \left({}^t\mathbf{x}_\mathrm{i}\mathbf{b} + \mathrm{b_0}\right) &>= 1 - \mathrm{M} * \mathrm{e_i} \end{aligned} \tag{1.7}$$

c penalty c for combining two objectives.
$\mathrm{e_i}$ nonnegative real value.
M big M constant.

1.4.5 Statisticians Claim for MP-Based LDFs

Some statisticians claimed we did not describe the algorism of four OLDFs. Although the notations of four OLDFs and three SVMs are similar, IP solver solves IP-OLDF and RIP, LP solves Revised LP-OLDF, and QP solves three SVMs. Thus, IP, LP, and QP solvers are the algorism of MP-based LDFs and conclude completely different results. First, we must be aware of the optimization criteria. IP-OLDF and RIP use MNM criterion. Revised LP-OLDF uses to minimize the summation of misclassified distance from the SVs that is the same as the second object of S-SVM. H-SVM maximizes the SV distance that is the same as the first object of S-SVM. Because three SVMs cannot select feature naturally, their standard may cause these results, and QP solver prevents to find SMs typically. Second, we must be aware the all points of the feasible region are the optimal solutions. Moreover, three SVMs and Fisher's LDF accept each one optimal LDF on the whole gene space, not on the subspace. We must understand the stepwise and all possible model methods are methods to find the better model in the subspaces. On the other hand, the B and B algorithm of the IP solver outputs many optimal LDFs in the whole gene space and many subspaces.

1.5 Matryoshka Feature Selection Method (Method2) and RatioSV

1.5.1 Method2

Many statistical researchers have raised the above three excuses for reasons why they could not succeed in Problem5. However, when we discriminated against six microarrays downloaded from Higgins HP (Jeffry et al. 2006), the following surprising results were obtained.

(1) The microarrays are LSD (Fact3). To the best of our knowledge, there is no research about LSD-discrimination. MNM decreases monotonously (MNM_k >= $\text{MNM}_{(\text{k}+1)}$). If $\text{MNM}_\text{k} = 0$, all MNMs including these k variables are zero (Fact2). This fact is essential for cancer gene analysis. We call all linearly separable microarray and subspaces as Matryoshkas in gene analysis. The full model having p-variable is a big Matryoshka that includes all smaller Matryoshkas in it. When RIP discriminates the microarrays, most coefficients of those are zero, and few coefficients are not zero. RIP can find smaller Matryoshka naturally, gene number of which is less than the case number n. When again discriminating Matryoshka, RIP found smaller Matryoshka than the previous Matryoshka. If we cannot see smaller Matryoshka anymore, we call it the first SM1. Next, RIP discriminates the reduced microarray removed SM1 again. RIP finds the second SM2. Moreover, RIP finds many SMs, MNM of those are zero. Thus, we develop Method2 within 54 days from October 28, 2015, to December 20, 2015. On the other hand, we spent three years to solve Problem3 because we approached by wrong trials from the multivariate analysis. We found the reason of Problem3 by checking all variables by one-way ANOVA. Many statisticians think Problem5 using microarrays may be impossible because they could not solve it from 1970. The actual reason is that the statistical discriminant functions are useless for Problem5 (Shinmura 2018b). This suggests that RIP is the best LDF for cancer gene analysis[12] for the following reasons. (a) RIP and H-SVM can theoretically discriminate LSD. (b) RIP can find one of the SMs from the many optimal solutions which make MNM = 0, but H-SVM finds only the optimal

[12] Because our results are true from the viewpoint of statistical analysis and are not confirmed by medical research, we use cancer gene analysis instead of oncogene analysis.

SVM coefficient for maximizing SV's distance. Thus, H-SVM cannot separate the microarrays into signal and noise because of NP-hard. Furthermore, the statistical discriminant functions based on variance–covariance matrices cannot discriminate SMs theoretically. On the other hand, logistic regression, three OLDFs, and H-SVM can discriminate SMs correctly.

(2) Method2 finds the microarray consists of disjoint unions of many SMs and another noise subspace (MNM >= 1). We think SMs are signals, and another gene subspace is noise in cancer gene analysis because we can discriminate two classes entirely by genes set of each SM and misclassify two classes by noise subspace. However, we could not find the linear separable fact of all SMs by the standard statistical methods. The "linear separable fact" means that two classes are separable in each SM.

(3) Because NM of Fisher's LDF is often large for LSD-discrimination, it is useless for cancer gene analysis in addition to medical diagnosis, pattern recognition, rating, and so forth. JMP (Sall, Creighton and Leman 2004) does not support logistic regression to analyze the microarrays now. Even if logistic regression could discriminate high-dimensional microarray and its NM may be zero, most of the coefficients are not zero like H-SVM. Thus, H-SVM and logistic regression must compute all possible models to find SM from the microarrays.

(4) Because six NMs of H-SVM are zero and most coefficients are not zero, H-SVM is useless for cancer gene analysis. H-SVM must compute all possible models to find SMs. Thus, NP-hard is true for H-SVM as same as statistical discriminant functions. The maximization distance of two SVs that is the objective function of H-SVM in Eq. (1.6) and the first objective of S-SVM in Eq. (1.7) causes this defect because Revised LP-OLDF can select gene feature naturally. Primarily, the object function of Revised LP-OLDF is the same as the second object function of S-SVM. Moreover, it is an essential fact that all LDFs, except for RIP and Revised LP-OLDF, cannot find one of the several optimal LDFs.

Shinmura (2015e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, 2016a) finds all SMs of six microarrays. Thus, we recommend LASSO researchers will evaluate and compare their results with our results using the microarrays. At first, they must check whether their methods can discriminate LSD correctly. We consider the LASSO cannot discriminate LSD successfully. Next, it cannot separate the microarrays into many small signals and noise as same as Fisher's LDF. If our claim is wrong, please show the results in papers or books.

1.5.2 *RatioSV: Measurement of the Degree of Linear Separability*

To evaluate LSD-discrimination by RIP, Revised LP-OLDF, and H-SVM, we introduce a RatioSV that is the ratio of support vector (SV) distance and DS range defined by Eq. (1.8).

$$\mathrm{RatioSV} = \mathrm{SV\,distance} * 100/\mathrm{discriminant\ score(DS)} \tag{1.8}$$

Because the maximum RatioSV range of six microarrays by RIPs is [11.67, 38.98%] explained in Chap. 2, we expect the RatioSV may be useful as cancer gene malignancy indicators. Although Golub et al. validated their results by LOO method, we need not validate our results by Method1 because two classes are separated in SM entirely. After many trials, we make new data having n cases (subjects) and all **RipDS**s as variables. When we analyze new data by PCA and cluster analysis, these results show the clear linear separable facts. Thus, we conclude our three new data made by RipDSs, LpDSs, and HsvmDSs are the true signals instead of SMs. We show many truths about our claim after Chap. 3.

1.5.3 Six Famous Microarrays

We developed two new methods such as Method1 and **Method2**. Because Method1 solved Problem4 completely, we misunderstood to establish the theory in 2015. On October 25, 2015, we presented our theory at the Japanese statistical conference held in our native Toyama city. Next day, because doctor course student presents her research using the microarrays (Ishi et al. 2014), we realized we did not solve Problem5. On October 28, we could download the following microarrays from Higgins HP.

1. Alon’s microarray (1999) consists of 62 cases and 2,000 genes. Two classes are the 22 normal cases (Normal, Class1) and the 40 tumor cases (Tumor, Class2). File volume is 21.12 Kb.
2. Golub microarray (1999) consists of 72 cases and 7,129 genes. Two classes are the 25 Acute Myeloid Leukemia cases (AML, Class1) and the 47 Acute Lymphoblastic Leukemia cases (ALL, Class2). File volume is 6,190 Kb.
3. Shipp microarray (2002) consists of 77 cases and 7,129 genes. Two classes are the 19 Follicular Lymphoma cases (FL, Class1) and 58 DLBCL cases (DLBCL, Class2). File volume is 9,344 Kb.
4. Chiaretti microarray (2004) consists of 128 cases and 7,129 genes. Two classes are the 95 patients (B-cell, Class1) and 33 patients (T-cell, Class2). File volume is 27,409 Kb.
5. Singh’s microarray (2002) consists of 102 cases and 12,625 genes. Two classes are the 50 normal subjects (Normal, Class1) and the 52 tumor prostate patients (Tumor, Class2). File volume is 21,888 Kb.
6. Tian’s microarray (2003) consists of 173 cases and 12,625 genes. Two classes are the 36 false cases (FALSE, Class1) and the 137 true cases (TRUE, Class2). File volume is 37,000 Kb.

We could develop the first three microarrays on Excel 32 bit version. Later, after we bought Office 64 bit version and replaced LINGO 32 bit version to LINGO 64 bit version, we could discriminate other three microarrays.

1.5.4 *How to Develop Method2 (a Surprising 54-Day Research Diary)*

Between October 28 and December 28, 2015, we discriminated against six microarrays. The microarrays were LSDs (Fact3), and those could be very easily decomposed into many SMs (Fact4). Since 1970, many researchers tried to analyze high-dimensional data such as a microarray. Because there was no success in cancer gene analysis, there were three excuses. However, our theory solved cancer gene analysis entirely without being influenced by three excuses.

1. On October 28, 2015, we discriminated against the Shipp microarray. RIP's CPU time is less than 1 s. The MNM of RIP was zero. Also, 32 coefficients of the 7,129 genes were not zero, and the other 7,097 coefficients were zero. All the discriminant coefficients on the first sheet were 0, but 32 nonzero coefficients were found by scrolling the sheet. Thus, 32 genes with nonzero coefficients were judged to be cancer genes in a statistical sense, as they can ultimately distinguish between the two classes. In this book, we use "cancer genes" in place of the technical term "oncology" discovered by medical research. Because cancer gene analysis is very important for humans, we decided to upload the results to Research Gate (Shinmura 2015e) as a position paper.
2. On November 1, 2015, six MP-based LDFs such as three OLDFs and three SVMs discriminated Alon, Golub, and Shipp microarrays. We find three microarrays are LSD. Although the nonzero coefficients of the three OLDFs are less than 62, zero coefficients of three SVMs are few. Although about 1,000 coefficients of H-SVM and SVM1 are zero for Golub microarray, these results are not used for cancer gene analysis because NMs of these models with nonzero coefficients are not zero. We claim the models chosen by LASSO are not used for cancer gene analysis because NMs of these models with nonzero coefficients may be not zero. Because we count these numbers by eyesight, there are mistakes in the values. We conclude only three OLDFs find SMs and three SVMs cannot find SMs (Shinmura 2015f).
3. On November 3, 2015, we considered three steps of feature selection methods to find smaller gene set using Shipp microarray. In step 1, RIP reduces 72 cases with 7,129 genes to 72 cases with 72 genes. In step 2, the stepwise forward method finds 72 cases with six genes data is LSD because of NM of logistic regression = 0. In step 3, all possible models of six variables find two three-gene models are a minimum number of SMs. Later, we called these SMs as BGSs.
4. On November 5, 2015, we confirmed the above feature selection method for Alon's microarray. In step 1, RIP reduces 77 cases with 2,000 genes to 77 cases with 63 genes. In step 2, the stepwise forward method finds 77 cases with six genes data is LSD because of NM of "logistic regression = 0." In step 3, all possible models of six variables find BGSs.
5. On November 9, 2015, we confirmed the above feature selection method for Golub microarray. In step 1, RIP reduces 72 cases with 7,129 genes to 72 cases with 72 genes. In step 2, the stepwise forward method finds 77 cases with six

genes data is LSD because of NM of "logistic regression = 0." In step 3, all possible models of six variables find BGSs.

6. At Discovery Summit held in Tokyo, Doctor Sall announced JMP version 12 that can support Fisher's LDF for high-dimensional data. We borrow JMP ver.12 from JMP Japan division and discriminate above three microarrays by Fisher's LDF. On November 11, 2015, we compared MNM of RIP and two NMs of Revised LP-OLDF and Fisher's LDF using three microarrays. NMs of Alon, Shipp, and Golub are zero, three, and eight, respectively. We report this result to JMP technical staff. Later, NM of Alon's microarray becomes to five (Shinmura 2015j).
7. On November 18, 2015, although RIP reduces Golub microarray (72 cases and 7,129 genes) to 72 genes, we get a smaller SM with 46 genes when RIP discriminates 72 genes again. By three trials of discrimination, we obtained the following Matryoshka process: Matryoshka7129 -> Matryoshka72 -> Matryoshka46 -> Matryoshka36. If RIP cannot find a smaller Matryoshka anymore, we stop and call it SM1 (Shinmura 2015k).
8. On November 22, 2015, we defined the Matryoshka Trap of Feature Selection Method that was confirmed by six MP-based LDFs and JMP using three microarrays. Moreover, we find another new truth. When we remove SM1 from microarray and discriminate the reduced microarray, we find the second SM2. We realized it was difficult for us to find all SMs manual work (Shinmura 2015). Caution: Because above eight position papers have several mistypes, nobody had better read.
9. We realize we cannot find all SMs by manual work. Thus, we develop LINGO Program3 of Method2 on December 4. Shinmura lists up all SMs of three microarrays of Shipp, Golub, and Alon (Shinmura 2015m, n, o).
10. To develop Singh, Tian, and Chisretti microarrays on Excel files, we bought the 64-bit version of Excel, PC, and OS. Shinmura (2015p, q, r) lists up all SMs of Singh, Tian, and Chiaretti microarrays by LINGO Program3.

After we recognize Problem5 on October 28, we completely solve it on December 20, 2015 with 54 days.[13] Although we misunderstand the discrimination of microarrays requests colossal CPU time, Fisher's LDF by JMP ver.12 (JMP12) and other MP-based LDFs coded by LINGO can solve microarrays less than 20 s because the microarrays are LSD. Although many researchers have complained that Problem5 is NP-hard, LSD-discrimination is easy. Moreover, MP-based LDFs are free from the small n and large p problem because these LDFs need not construct the variance—covariance matrices. Besides, because three OLDFs can decompose the microarrays into many SMs and noise gene subspace, we get signals naturally. Although there are many types of research of the filtering systems and feature selection methods, we need not use these methods. From December 11th to 16th, I presented my paper at the CMStatistics conference held at the University of London. Other presentations

[13] I presented at CMStatistics held at London University from December 11 to 16, 2015. My presentation was related to Method1, but since other researchers' presentations were a genetic analysis of cancer by LASSO, I switched the presentation to cancer gene analysis.

were to analyze microarrays by Lasso in the near future. I did not understand why they did not analyze at once.

1.5.5 *Results of Six Microarrays*

Table 1.1 shows the summary of all SMs found by December 20, 2015. The “Description” line is the details of the two classes. The row “size” is the number of patients and the number of genes. “SM: Gene” is the number of SMs and the total number of genes included in all SMs. The parentheses are reference papers listing them. Six papers (Shinmura 2015m, n, o, p, q, r) include full gene names of SM. “Min, Mean, Max” rows are the minimum, mean, and maximum values of genes included in all SMs of each microarray. Rows JMP12 are two by two contingency tables of the discrimination by Fisher’s LDF. Six NMs (=False Positive + False Negative) are 5, 3, 8, 10, 3 and 29. Rows “% and error rate” are the percentages of (Maximum value/number of patients) and error rates of Fisher’s LDF. Maximum percent is 88% of Tien’s microarray. Minimum percent is 45% of Singh’s microarray. The maximum error rate is 17% of Tian’s microarray, and the minimum error rate is 2% of Chiaretti microarray.

Table 1.1 Summary of six microarrays (December 2015)

Data	**Alone et al.** (1999)	**Chiaretti et al.** (2004)	**Golub et al.** (1999)
Description	Normal (22) : tumor cancer (40)	B-cell (95) : T-cell (33)	All (47) : AML (25)
Size	62 * 2000	128 * 12,625	72 * 7129
(SM: Gene)	66:1131(Shinmura 2015o)	269:5220 (Shinmura 2015r)	67:1203 (Shinmura 2015n)
Min, mean, max	11, 17.1, 32	9, 19.4, 71	10, 19.4, 41
JMP12	20:2/3:37	94:1/2:31	20:5/3:44
% and error rate	52, 8%	55, 2%	57, 11%
Data	**Singh et al.** (2002)	**Shipp et al.** (2002)	**Tian et al.** (2003)
Description	Normal (50) : tumor prostate (52)	Follicular lymphoma (19) : DLBCL (58)	False (36) : true (137)
Size	102 * 12,626	77 * 7129	173 * 12,625
(SM: Gene)	178: 3984 (Shinmura 2015p)	214: 3040 (Shinmura 2015m)	159: 7221 (Shinmura 2015q)
Min, mean, max	13, 22.4, 46	7, 14.2, 39	28, 48, 152
JMP12	46:4/6:46	17:2/1:51	16:20/9:128
% and error rate (%)	45, 10%	51, 4%	88, 17%

(SM: Gene): Five results of SM: Gene, except for Tian et al. is replaced in new results of Table 1.2

BGSs may be unique in each microarray. On the other hand, there are defects of SMs as follows:

(1) Revised LP-OLDF and Revised IPLP-OLDF (Shinmura 2009) find another different SMs. These differences are caused by different optimization criteria and MP solvers such as IP and LP. From Chaps. 2–9 describe the results of new SMs obtained by the different versions of LINGO from 2016 to 2018.
(2) LINGO Program3 need to select the iteration number of RIP as an option. In Table 1.1, because we do not know the appropriate value of it, we used 10 or 15, and so forth. Because we are not skilled in programming techniques, we chose a simple program structure to specify the iteration number.
(3) In addition to the iteration number, a yearly version up of LINGO may cause a different combination of SMs, especially for a RIP. RIP outputs one of the SMs among many optimal solutions.

In Chap. 3, we consider the method of determining the appropriate number of repetitions, and we are conducting analyzes on the SM obtained by this method. Seven chapters from Chaps. 3–9 introduce the cancer gene diagnosis using new SMs and several different themes.

1.5.6 The Reason for Natural Feature Selection

After finding all the SMs of six microarrays, the following questions arose.

(1) Why couldn’t statisticians find the microarrays are LSD?

If some researchers discriminate the microarray with H-SVM or RIP, they can find an essential fact that the six microarrays are LSDs (Fact3). Although some papers used SVM, there were no explanations which they used H-SVM or S-SVM. Probably, we understand they used S-SVM because H-SVM does not work correctly for the overlapping data. When they discriminated the microarray with H-SVM or RIP, they found essential clues for cancer gene analysis. Several papers have pointed out that the NMs of SVM were zero with certain small gene combinations selected by the medical judgments. However, we found LSD has Matryoshka structure by Swiss banknote data, but no researchers derived essential facts of the Matryoshka structure of microarrays. In summary, most researchers do not recognize the importance of LSD. Moreover, it is crucial to find that microarray is LSD and think that systematic understanding could not be obtained even if the NM of S-SVM was 0 with arbitrarily selected genes.

(2) Why couldn't Fisher's LDF find LSD?

Fisher opened the new frontier of the discriminant analysis and developed the maximum likelihood method. He defined Fisher's LDF based on the variance–covariance matrices under Fisher's assumption. Two classes belong to two same normal distributions with the same variance and different averages such as $\mathbf{m}_1$ and $\mathbf{m}_2$ ($\mathbf{m}_1$ not $= \mathbf{m}_2$). It maximizes the correlation ratio. Because he had no computer power, we guess he used the character of the exponential function in Eq. (1.9).

$$\begin{aligned}\log\{\mathrm{f}_1(\mathrm{x}:\mathbf{s},\mathbf{m}_1)/\mathrm{f}_2(\mathrm{x}:\mathbf{s},\mathbf{m}_2)\} &= \log[\mathrm{e}^{\{(\mathrm{x}-\mathbf{m}_2)^2-(\mathrm{x}-\mathbf{m}_1)^2\}}/(2*\mathbf{s}^2)] \\ &= (\mathbf{m}_1-\mathbf{m}_2)/\mathbf{s}^2*\mathrm{x}+(\mathbf{m}_2^2-\mathbf{m}_1^2)/\left(2*\mathbf{s}^2\right) \quad (1.9)\end{aligned}$$

If we consider the discriminant hyperplane as $\mathrm{f}_1(\mathrm{x}:\mathbf{s},\mathbf{m}_1)=\mathrm{f}_2(\mathrm{x}:\mathbf{s},\mathbf{m}_2)$, we obtain the discriminant hyperplane in Eq. (1.10). In this case, NM becomes MNM because data satisfies Fisher's assumption. If the data does not satisfy Fisher's assumption, NM is greater than equal MNM.

$$(\mathbf{m}_1-\mathbf{m}_2)/\mathbf{s}^2*\mathrm{x}+(\mathbf{m}_2^2-\mathbf{m}_1^2)/\left(2*\mathbf{s}^2\right)=0 \quad (1.10)$$

We consider Fisher's LDF by $\mathrm{F}(\mathbf{x}) = \mathbf{b} * \mathbf{x} + \mathrm{b}_0$. Later, he or another statistician introduced the maximization criterion of correlation ratio. Most statisticians believe this standard is essential for the discrimination. They do not doubt the defect of this standard that cannot discriminate LSD correctly and solve Problem5. We had already confirmed QDF and RDA were very weak for LSD-discrimination, also. We believe the LASSO cannot solve cancer gene analysis as same as Fisher's LDF, QDF, and RDA. If the maximum likelihood method solves Fisher's LDF as same as logistic regression, we believe it can often discriminate LSD correctly for a small sample. In summary, we believe that posterity researchers did not enhance discriminant theory to solve real essential problems.

(3) Why couldn't LASSO find SM?

Some statisticians misunderstand LASSO can solve the cancer gene analysis because it can make several coefficients zero. Because it does not adopt the MNM criterion, the found subspaces are rare to be LSD. We already explained in Fig. 1.1. First, because it cannot discriminate LSD theoretically, we think it cannot solve the cancer gene analysis. Next, only coefficients related to BGS or SM must be the nonzero coefficients. Because all LDFs, except the three OLDFs, find one optimal solution in the whole domain, it does not find one of the optimal solutions from subspaces. Thus, it must compute all possible models to find SMs. We must realize the feature selection methods including all possible models are to find the optimal solution of subspaces. In summary, we suggested that LASSO researchers examine the above matters after 2016. It is also worthwhile to announce failed research results.

(4) Why couldn't H-SVM find SM?

H-SVM can correctly identify LSD, but it cannot find SM because it can see only one optimal H-SVM coefficients on the whole domain. This is because QP defines SVM based on the maximization of SV's distance. QP minimizes or maximizes the quadratic objective function on the entire domain. To summarize, if researchers understand that microarray is LSD and LSD has Matryoshka structure, they may have found SM by overcoming the difficulty of NP-hard.

(5) Why could only OLDFs find SM?

Because Revised LP-OLDF finds a vertex of the feasible region which is a solution of simultaneous equations obtained from n or fewer constraints as an optimal solution, it can discover SM having less than same n genes easily and quickly. Furthermore, the feasible region has a unique feature that all MNMs of the feasible region are 0. The reason why RIP can decompose the microarrays is simple. B and B algorithm of IP solver is the efficient algorithm of all possible models that search all subspaces. Moreover, the IP model is a restricted LP model that the decision variables are the binary integers. Thus, it can output one of SM.

On the other hand, in summary, RIP based on the MNM criterion found SMs at first. However, Revised LP-OLDF (and Revised IPLP-OLDF) can discriminate against six microarrays correctly and decompose six microarrays into many different SMs, also. Because Revised LP-OLDF has the defect of Problem1 and its NMs may not be correct showed in Chap. 3, we recommend not to use it for the overlap data. For six microarrays, although six MNMs are zero, Revised LP-OLDF cannot find all SMs from the microarrays. From Chaps. 4–7 introduce this truth. However, because three microarrays such as Singh, Tian, and Chiaretti consist of 12,625 genes and find over 150 SMs, we show the evaluation results using SMs found by Revised LP-OLDF from Chaps. 7–9, and almost the same results as Alon, Golub, and Shipp. In other words, when there are more than 200 SMs, even if there is a loss, analysis by SM obtained by Revised LP-OLDF is also conceivable.

1.5.7 Two New Facts

(1) Two Known Facts

This book discusses the two new facts in addition to two known facts such as:

(1) IP-OLDF and Fig. 1.1 can explain the relation of NMs and LDF coefficients on the p-discriminant hyperplane clearly. This known fact proves the defect of NM that is not reliable. Correct NM may be higher than obtained NM. Although microarrays are LSD (Fact3), six NMs of Fisher's LDF are not zero. This truth is one of the reasons why researchers could not solve the cancer gene analysis since 1970, also. The error rate of Fisher's LDF using Tian's microarray is 17% in Table 1.1. In the pass/ fail judgment of the test, the error rate was high.

However, some researchers said that the pass/fail decision of the examination was meaningless because the results are not applied for the next examination. The same result was shown even with cancer determination using microarray.

(2) "MNM monotonic decrease" explains the Matryoshka structure of LSD. We had already found the Swiss banknote data, Japanese car data, and the pass/fail determination by exam scores (Shinmura 2011b, 2015b) were LSD. We can easily understand the new idea about Matryoshka structure and SMs of the microarrays by the results of the above common data. Section 1.6 shows the Matryoshka structure using the Swiss banknote data and the Japanese car data.

(2) The New Facts in This Book

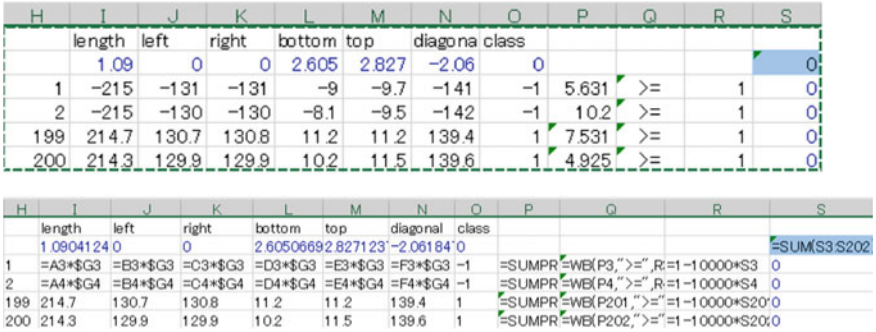
We introduce several truths to explain the reason why all LDFs, except for three OLDFs, cannot find SMs (Fact4). Swiss banknote data illustrates the reason for Fact4. At first, we found MNM of the two-variable model (X4, X6) is zero after we examined all possible models (Goodnight 1978). This model is the minimum-dimensional SM. We call it as basic gene set (BGS) among 63 models. The BGS is as same as Yamanaka's four genes. If we drop one gene from BGS, MNMs of (X4) and (X6) are higher than zero. Next, we found the MNM monotonic decrease. Third, if $MNM_k = 0$, all MNMs including these k-variable are zero. Thus, 16 MNMs including (X4, X6) are zero and are signals. Other 47 MNMs those do not include (X4, X6) are higher than 0. The 47 models are noise. IP-OLDF and RIP could separate signals and noise naturally and are free from three excuses explained in Sect. 1.3.3.

1.6 Validation of Method2 by Common Data

1.6.1 Matryoshka Structure of Swiss Banknote Data

Although several discriminant coefficients of Swiss banknote data and Japanese car data became zero by RIP, we do not use this fact for feature selection because we developed the best model instead of feature selection method for six common data. Furthermore, we had found the several coefficients of LP-OLDF and IP-OLDF are zero by the Iris data that is not LSD (Shinmura 2000b).[14] We ignored the fact that a few coefficients become zero. We are happy to avoid a wrong approach because the nonzero variable model is not valuable. Probably, even if we omit the variables with zero coefficients, the model differs from SM. Figure 1.2 shows What's Best!, add-in solver of Excel. Seven coefficients are output on "I2: O2." Two hundred e_i are from S3 to S202 cells. Cell S2 defines the objective function. P column stores 200 discriminant scores and R column stores 1 or −9999 of 200 e_i. If we choose "IT=4", the five-variable model (X1, X3–X6) in Fig. 1.2.

[14] This dissertation can be downloaded from Research Gate as same as all English papers related to OLDFs. Even if LASSO could make some discriminate coefficients of 0, it would be of no use to cancer research at all because its MNM is not zero.



H	I	J	K	L	M	N	O	P	Q	R	S
	length	left	right	bottom	top	diagona	class				
	1.09	0	0	2.605	2.827	−2.06	0				0
1	−215	−131	−131	−9	−9.7	−141	−1	5.631	>=	1	0
2	−215	−130	−130	−8.1	−9.5	−142	−1	10.2	>=	1	0
199	214.7	130.7	130.8	11.2	11.2	139.4	1	7.531	>=	1	0
200	214.3	129.9	129.9	10.2	11.5	139.6	1	4.925	>=	1	0

H	I	J	K	L	M	N	O	P	Q	R	S
	length	left	right	bottom	top	diagonal	class				
	1.0904124	0	0	2.6050669	2.827123	−2.06184	0				=SUM(S3:S202
1	=A3*$G3	=B3*$G3	=C3*$G3	=D3*$G3	=E3*$G3	=F3*$G3	−1	=SUMPR	=WB(P3,">=",R	=1−10000*S3	0
2	=A4*$G4	=B4*$G4	=C4*$G4	=D4*$G4	=E4*$G4	=F4*$G4	−1	=SUMPR	=WB(P4,">=",R	=1−10000*S4	0
199	214.7	130.7	130.8	11.2	11.2	139.4	1	=SUMPR	=WB(P201,">="	=1−10000*S20	0
200	214.3	129.9	129.9	10.2	11.5	139.6	1	=SUMPR	=WB(P202,">="	=1−10000*S20	0

Fig. 1.2 Swiss banknote data

1.6.2 Validation of LINGO Program3 Results

We validate the results of Table 1.1 by other approaches as follows:

(1) Although we counted the number of nonzero coefficients on the Excel files by our eyesight in Table 1.1, we compute those by JMP in Table 1.2.

(2) We explain the logic of Program3 in Sect. 1.6.2.2. If we choose the different iteration number (option), we may obtain the various combination of SMs. On the other hand, each microarray has a unique disjoint union of BGSs. Thus, we targeted to look for all BGSs. However, we realize the analysis of BGSs is useless for cancer gene diagnosis because the 130 RasioSVs of Alon's BGS are too small compared with those of Alon's SVs in Chap. 2.

1.6.2.1 Validation of Discriminant Coefficients by JMP

We are regretful not to count the number of zero coefficients by JMP. LINGO Program1 discriminates the microarrays by six LDFs and outputs the discriminant coefficients in array Vark100 of LINGO Program1 and Excel array in Chap. 10 of Shinmura (2016d). After JMP replaces zero coefficient as 0 and another nonzero coefficient as 1, JMP counts the number of 0/1 in Table 1.2. Alon's microarray has 2,000 genes. The 1,938 coefficients of RIP are zero, and only 62 coefficients are not zero. Bold figures indicate that the intercept becomes zero. Thus, the bold figures 1938 mean that 1938 coefficients and the constant are zero. Because 40 coefficients of Revised IPLP-OLDF and Revised LP-OLDF are not zero, RIP gene subspace is 22 greater than Revised IPLP-OLDF and Revised LP-OLDF. All coefficients including the intercept of three SVMs are not zero. Golub microarray has 7,129 genes. The 903 coefficients of H-SVM and 904 coefficients of SVM1 are zero. H-SVM and SVM1 can select features of Golub microarray. However, we consider these SVMs cannot find SM and BGS, because number of non-zero coefficients are large.

To summarize these results are as follows:

(1) H-SVM and SVM1 can select features of Golub microarray, and 903 and 904 coefficients of two SVMs are zero. H-SVM and SVM1 cannot reduce the subspace to smaller SM again. Thus, these LDFs, like LASSO, cannot find SM and BGS fewer than case numbers because of QP prevents it. We hope Golub give us valuable information about this fact.

(2) However, because Revised LP-OLDF is faster than RIP, Revised LP-OLDF is another choice to survey about SMs.

Table 1.2 Validation of discriminant coefficients

LDF	Level	Alon	Chiaretti	Golub	Shipp	Singh	Tian
RIP	0	**1938**	12,498	**7057**	7065	12,534	**12,452**
	1	62	**127**	72	**64**	**91**	173
	Total	2000	12,625	7129	7129	12,625	12,625
IPLP	0	**1960**	**12,587**	**2102**	7108	**12,550**	**12,507**
	1	40	38	27	**21**	75	118
	Total	2000	12,625	7129	7129	12,625	12,625
LP	0	**1960**	**12,587**	2103	7108	**12,550**	**12,486**
	1	40	38	**26**	**21**	75	139
	Total	2000	12,625	7129	7129	12,625	12,625
HSVM	0	0	0	903	0	0	0
	1	**2000**	**12,625**	**6226**	**7129**	**12,625**	**12,625**
	Total	2000	12,625	7129	7129	12,625	12,625
SVM4	0	0	0	0	0	0	0
	1	**2000**	**12,625**	**7129**	**7129**	**12,625**	**12,625**
	Total	2000	12,625	7129	7129	12,625	12,625
SVM1	0	0	0	904	0	0	0
	1	**2000**	**12,625**	**6225**	**7129**	**12,625**	**12,625**
	Total	2000	12,625	7129	7129	12,625	12,625

1.6.2.2 Detail of the Matryoshka Feature Selection Method

We explain Method2 briefly. Table 1.3 is the output of Golub microarray by LINGO Program3. Two columns LOOP1 and LOOP2 are the sequence number of big and small loops of Method2. RIP discriminates the microarray with 7,129 genes in the LOOP1 = 1 and LOOP2 = 1, and only 34 coefficients of RIP 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 Matryoshka with 34 genes again, and only 11 coefficients are not zero. Thus, we get the Matryoshka sequence such as Matryoshka7,129 → Matryoshka34 → Matryoshka11 drastically. We stop at LOOP2 = 4 because we cannot find the smaller Matryoshka. We call Matryoshka 11 as the

SM1 because RIP cannot locate the smaller Matryoshka anymore. We exclude the first SM1 with 11 genes from the big Matryoshka with 7,129 genes and make the second big Matryoshka with 7,118 genes. In the second big loop at LOOP1 = 2, we get the second SM2 with 16 genes. We can continue this loop until it cannot naturally select features and find small subspace with MNM >= 1.

Table 1.3 Outlook for theory2

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 Program3 finds 69 SMs in Table 1.4, it stops the big loop when MNM is higher than one at LOOP1 = 70. However, we can continue this loop by changing the option and list up all small subspaces with MNM >= 1. Thus, Method2 can apply for other gene data that are not LSD. However, it is difficult to find valid meanings in non-LSD subspaces. Because Golub microarray consists of 69 SMs that are LSD, it is vital for us to analyze all SMs for cancer gene diagnosis.

Table 1.4 All SMs of Golub et al. microarray

Loop1	Loop2	Gene	n	MNM
1	11	7129	11	0
2	11	7118	16	0
3	11	7102	11	0
–	–	–	–	–
32	11	6683	19	0
33	11	6664	16	0
34	11	6648	18	0
35	11	6630	17	0
36	11	6613	19	0
37	11	6594	12	0
38	11	6582	16	0
–	–	–	–	–
67	11	5976	23	0
68	11	5953	31	0
69	11	5922	31	0

1.6.3 Validation of Method2 by Japanese 44 Cars Data

1.6.3.1 Japanese 44 Cars Data

Japanese car data consists of the 29 regular cars and 15 small cars. The six independent variables are the emission rate (X1), the price (X2), the number of seats (X3), CO2(X4), fuel (X4), and sales (X6). Because the emission (X1) and capacity (X3) of small cars are less than those of regular cars, two MNMs of these one-variable models are zero, and those are BGSs. Figure 1.3 shows the Box-whisker plots of X1 and X3. These graphs tell us that X1 and X3 are linearly separable models. Thus, 48 MNMs including (X1) or (X3) are zero. Other 15 MNMs are not zero.

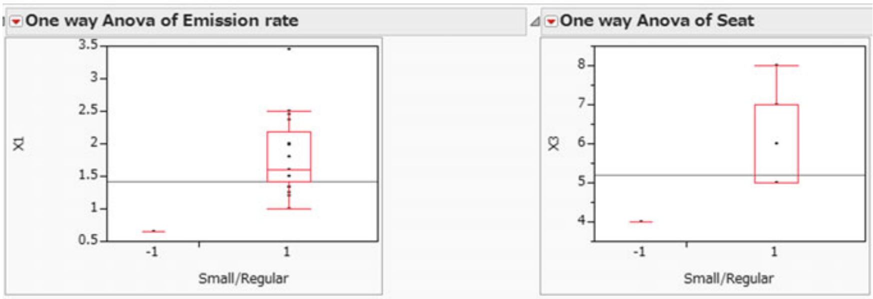


Fig. 1.3 Box-whisker plots of emission and capacity (−1: small car, 1: regular car)

Table 1.5 shows the result of a stepwise forward method that chooses X1, X2, X3, X4, X5, and X6 in this order. "MNM" column is MNMs of RIP. Because X1 is BGS, all models including X1 are zero by the MNM monotonic decrease. On the other hand, LDF column (Fisher's LDF) shows four NMs of (X1), (X1, X2), (X1, X2, X3), and (X1, X2, X3, X4) are not zero. Although two NMs of (X1) and (X1, X2) by QDF are zero, QDF misclassifies all 29 regular cars to the small cars because the seat numbers of 15 small cars are four and those numbers of 29 regular cars vary from five to eight. If we add a small random number the constant value, we can solve Problem3. If we set two parameters such as "$\lambda = \gamma = 0.8$," RDA's NMs are over two. By the grid search of two parameters such as "$\lambda = \gamma = 0.1$," all NMs of six models change zero. Because there is no rule to choose the best parameter values, we must survey the better parameters by try and error. This is the reason why we do not recommend RDA and S-SVM.

Table 1.5 Comparison of MNM and NMs

p	Var.	MNM	LDF	QDF	$\lambda = \gamma =$ 0.8	0.1
1	Emission (X1)	0	2	0	2	0
2	Price (X2)	0	1	0	4	0
3	Capacity (X3)	0	1	29	3	0
4	CO_2 (X4)	0	1	29	4	0
5	Fuel (X5)	0	0	29	5	0
6	Sales (X6)	0	0	29	5	0

1.6.3.2 Validation of Method2 by Japanese Car Data

When LINGO Program3 discriminates against Japanese car data, we obtain the result in Table 1.6. “SM” column is the sequential number of SM found by Program3. “IT” column shows the iteration of LOOP2 until three steps introduced in Table 1.3. In the three steps, Program3 finds the first SM1. From the fifth column to the tenth column in the third row shows the value of “Choice.” Because six values are 1 s, Program3 discriminates against six-variable model at first. “SUM” column shows the number of selected variables. The last column “c” means that the constant is always included in the model. Although the constant sometimes becomes zero, Method2 fixes the constant to 1. The first discrimination, MNM = 0. Because only the coefficient of X1 is not zero, the other five values from X2 to X6 become to 0 s in the second step. When Program3 discriminates one-variable model again, there is no change. Because of choosing “IT = 3,” Program3 discriminates one-variable model again and stop the first big loop. We obtain SM1 including X1 that is the first BGS1. In the second big loop, Program3 drops X1 and discriminates five-variable model in the first step. Moreover, the only third coefficient is not zero. In the second and third steps, Program3 discriminates against this model and stops the second big loop. Thus, Program3 finds the second SM2. In the third big loop, it discriminates the four-variable model, and two coefficients of X2 and X5 are not zero. Because of “MNM = 4,” this is not SM. However, we call it SM3 in this section. In the fourth step, it finds SM4 that consists of the two-variable model such as (X4, X6). Because we terminate big loop under the condition “NM >=15,” Program3 terminates in the fifth big loop and output “NM = 15.” The first row indicates Program3 finds four SMs as follows: SM1 = (X1), SM2 = (X3), SM3 = (X2, X5), and SM4 = (X4). Four NMs of SM1, SM2, SM3, and SM4 are 0, 0, 4 and 9, respectively.

Table 1.6 Results by RIP

			Matryoshka	1	3	2	4	3	4	0
SM	IT	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	1	0	6	1	1	1	1	1	1	1
1	2	0	1	1	0	0	0	0	0	1
1	**3**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**1**
2	1	0	5	0	1	1	1	1	1	1
2	2	0	1	0	0	1	0	0	0	1
2	**3**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**
3	1	4	4	0	1	0	1	1	1	1
3	2	4	2	0	1	0	0	1	0	1
3	**3**	**4**	**2**	**0**	**1**	**0**	**0**	**1**	**0**	**1**
4	1	9	2	0	0	0	1	0	1	1
4	2	9	2	0	0	0	1	0	1	1
4	**3**	**9**	**2**	**0**	**0**	**0**	**1**	**0**	**1**	**1**
5	1	15	0	0	0	0	0	0	0	1
5	2	15	0	0	0	0	0	0	0	1
5	**3**	**15**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**

Table 1.7 is the result of Revised LP-OLDF. Program3 finds four SMs as follows: SM1 = (X1), SM2 = (X3), SM3 = (X2, X4, X6), and SM4 = (X5). Because NM of (X5) is over than 15, it terminates in the fourth big loop and output "NM = 15."

Table 1.7 Result by Revised LP-OLDF

SM	IT	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	1	0	6	1	1	1	1	1	1	1
1	2	0	1	1	0	0	0	0	0	1
1	**3**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**1**
2	1	0	5	0	1	1	1	1	1	1
2	2	0	4	0	1	1	1	0	1	1
2	**3**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**
3	1	6	4	0	1	0	1	1	1	1
3	2	6	3	0	1	0	1	0	1	1
3	**3**	**6**	**3**	**0**	**1**	**0**	1	**0**	**1**	**1**
4	1	15	1	0	0	0	0	1	0	1
4	2	15	0	0	0	0	0	0	0	1
4	**3**	**15**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**

Table 1.8 is the result of H-SVM, SVM4, and SVM1. Because six coefficients of three SVMs are not zero, Program3 terminates in the first big loop.

Table 1.8 Result by H-SVM, SVM4, and SVM1

SM	IT	NM	SUM	X1	X2	X3	X4	X5	X6	c
1	1	–	6	1	1	1	1	1	1	1
1	2	–	6	1	1	1	1	1	1	1
1	3	–	6	1	1	1	1	1	1	1

1.6.3.3 Six Coefficients of Six MP-Based LDFs

Table 1.9 shows six MP-based LDF coefficients by Program1. We set the "absolute value <= 10ˆ−9" is showed as zero. In three OLDFs of step1, both X3 and constant are zero. Three columns such as "0< , =0, >0" show the case number of which discriminant score $y_i * f(x_i)$ satisfy the condition. All 44 cars are classified correctly. Seven "1/0" values of "Choice" row indicate which variables are included in the model. In step 2, we drop (X2, X3) in the models. Six LDFs are in Eq. (1.11). Thus, three OLDFs choose X1 as BGS correctly. The emission rate of small and regular cars ranges from [0.657, 0.658] to [0.996, 3.456], respectively. We can discriminate the data by X1 = (0.658 + 0.996)/2 = 0.827. Because X1 = 4.89/5.9172 = 0.825, the hyperplane of the threshold are almost the same.

$$\begin{aligned}
&\text{Three OLDFs} : 5.9172 * X1 - 4.89 \\
&\text{HSVM} : 5.9172 * X1 + 1E - 08 * X4 + 7E - 08 * X5 - 4.89 \\
&\text{SVM4} : 5.9175 * X1 - 0 * X4 - 0.02 * X5 + 8E - 07 * X6 - 3.97 \\
&\text{SVM1} : 2.9806 * X1 + 4E - 06 * X4 + 1E - 05 * X5 - 2.96
\end{aligned} \tag{1.11}$$

In step 3, we drop (X1) in the models. Six LDFs are in Eq. (1.12). Five LDFs except for SVM4 find the second BGS X3 correctly. The seats of small and regular cars are 4 and [5, 8], respectively. We can discriminate the data by X3 = (4 + 5)/ 2 = 4.5 as the discriminant hyperplane is X3 = 4.5. This means the average of small cars seat numbers and the minimum seat number of regular cars.

$$\begin{aligned}
&\mathrm{RIP} : 2 * \mathrm{X3} - 9\\
&\mathrm{IPLP} : 2 * \mathrm{X3} - 9\\
&\mathrm{LP} : 2 * \mathrm{X3} - 9\\
&\mathrm{HSVM} : 2 * \mathrm{X3} - 9\\
&\mathrm{SVM4} : 9\mathrm{E} - 09 * \mathrm{X2}\ + 2.005 * \mathrm{X3}\ -\ 0 * \mathrm{X4}\ -\ 4\mathrm{E} - 08 * \mathrm{X5}\ - 8.99\\
&\mathrm{SVM1} : 2 * \mathrm{X3}\ - 9
\end{aligned} \tag{1.12}$$

In step 4, we drop X1 and X3 in the models. NMs of RIP, IPLP, LP, H-SVM, SVM4, and SVM1 are 3, 3, 4, 4, 4, and 4, respectively. Thus, LINGO Program1 can simulate Program3 by step-by-step discrimination and conclude as follows:

(1) Three OLDFs can select two BGSs correctly.
(2) H-SVM and SVM1 can scarcely select features.

Table 1.9 Six MP-based LDF coefficients

Step	0<	=0	>0	LDFs	X1	X2	X3	X4	X5	X6	c
1	0	0	44	RIP	5.9999	−3E−08	0	−0.02	−0.13	−2E−05	0
	0	0	44	IPLP	5.9999	−3E−08	0	−0.02	−0.13	−2E−05	0
	0	0	44	LP	5.9999	−3E−08	0	−0.02	−0.13	−2E−05	0
	0	0	44	HSVM	0.6268	−2E−07	1.777	−0.01	−0.06	4E−06	−6.08
	0	0	44	SVM4	0.6509	−1E−07	1.779	−0.01	−0.05	3E−06	−6.5
	0	0	44	SVM1	0.6805	−6E−08	1.622	−0	−0.01	−8E−07	−7.37
	Choice				1	1	1	1	1	1	1
2	0	0	44	RIP	5.9172			0	0	0	−4.89
	0	0	44	IPLP	5.9172			0	0	0	−4.89
	0	0	44	LP	5.9172			0	0	0	−4.89
	0	0	44	HSVM	5.9172			1E−08	7E−08	0	−4.89
	0	0	44	SVM4	5.9175			−0	−0.02	8E−07	−3.97
	0	0	44	SVM1	2.9806			4E−06	1E−05	0	−2.96
	Choice				1	0	0	1	1	1	1
3	0	0	44	RIP		0	2	0	0	0	−9
	0	0	44	IPLP		0	2	0	0	0	−9
	0	0	44	LP		0	2	0	0	0	−9
	0	0	44	HSVM		0	2	0	0	0	−9
	0	0	44	SVM4		9E−09	2.005	−0	−0	−4E−08	−8.99
	0	0	44	SVM1		0	2	0	0	0	−9
	Choice				0	1	1	1	1	1	1

(continued)

Table 1.9 (continued)

Step	0<	=0	>0	LDFs	X1	X2	X3	X4	X5	X6	c
4	3	0	41	RIP		0.0033		−46.4	−199	−0.036	5343
	3	0	41	IPLP		0.0033		−46.4	−199	−0.036	5343
	4	0	40	LP		6E−06		−0.15	−0.8	−2E−05	25.93
	4	0	40	HSVM		6E−06		−0.15	−0.8	−2E−05	25.93
	4	0	40	SVM4		6E−06		−0.15	−0.8	−2E−05	25.93
	4	0	40	SVM1		6E−06		−0.15	−0.79	−3E−05	25.05
	Choice				0	1	0	1	1	1	1

1.6.4 Examination of Duplicate Data

We copy the data twice because there may be redundant pairs of genes. Six original variable names replace to C1–C6. First and second copy variable names change to c11–c16 and c21–c26. Program3 discriminates this data by ten small loops in Table 1.10. Program3 can decompose three SMs and three BGS. Three SMs include c21, c11 and C1. Three BGSs include one of c23, C3 and c13. These results show that three copies of X1 do not become BGSs because X1 fluctuates. On the other hand, because X3 is the constant, Program3 can find three BGSs.

$$\begin{aligned}&\text{Japanese Car Data with 18 variables} = \text{S1} \cup \text{S2} \cup \text{S3} \cup \text{S4} \cup \text{S5} \cup \text{S6}\\&= (\text{C2, C6, c21, c24, c25}) \cup (\text{c11, c14, c15, c16, c22}) \cup (\text{C1, C4, C5, c12, c26})\\&\cup(\text{c23}) \cup (\text{C3}) \cup (\text{c13})\end{aligned} \tag{1.13}$$

Program3 cannot find three BGSs such as (C1), (c11), and (c21). However, it finds three SMs such as (C2, C6, c21, c24, c25) ∪ (c11, c14, c15, c16, c22) ∪ (C1, C4, C5, c12, c26) before three BGSs such as (c23) ∪ (C3) ∪ (c13). Therefore, we expect Program3 can decompose the redundant gene pairs.

Table 1.10 Decomposition of duplicate data

SM	IT	T	NM	SUM	C1	C2	C3	C4	C5	C6	c11	c12	c13	c14	c15	c16	c21	c22	c23	c24	c25	c26	c
1	1	18	0	18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	**10**	**5**	**0**	**5**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**
2	1	13	0	13	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1
2	**10**	**5**	**0**	**5**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
3	1	8	0	8	1	0	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0	1	1
3	**10**	**5**	**0**	**5**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**
4	1	3	0	3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
4	**10**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**1**
5	1	2	0	2	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
5	**10**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
6	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
6	**10**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
7	1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	**10**	**0**	**15**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**

Table 1.11 shows six RIPs corresponding to Table 1.10. Three SMs and three BGSs have the same coefficients.

Table 1.11 Six RIPs corresponding to Table 1.10

IT	**C2**	**C6**	**c21**	**c24**	**c25**	**c**
1	−3E−08	−2E−05	5.9999	−0.019	−0.132	0
10	−3E−08	−2E−05	5.9999	−0.019	−0.132	0
	c11	**c14**	**c15**	**c16**	**c22**	**c**
1	5.9999	−0.019	−0.132	−2E−05	−3E−08	0
10	5.9999	−0.019	−0.132	−2E−05	−3E−08	0
	C1	**C4**	**C5**	**c12**	**c26**	**c**
1	5.9999	−0.019	−0.132	−3E−08	−2E−05	0
10	5.9999	−0.019	−0.132	−3E−08	−2E−05	0
	1	**2**	**c23**	**4**	**5**	**c**
1			2			−9
10			2			−9
	1	**2**	**C3**	**4**	**5**	**c**
1			2			−9
10			2			−9
	1	**2**	**c13**	**4**	**5**	**c**
1			2			−9
10			2			−9

1.7 Conclusion

From 1999 to 2004, six major research groups published papers on oncogene analysis using six microarrays and released their microarrays to the Internet. Golub et al. published an article in Science 1999 and summarized their research as follows. “Although cancer classification has improved over the past 30 years, we identify new cancer classes (class findings) or assign tumors to known classes (class predictions). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to acute human leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.” Therefore, we can see that this kind of research started at least around 1970.

Because these microarrays are high-dimensional data characterized as “small n large p,” statisticians targeted this theme as a new frontier of statistics. However, they offered no noticeable result. They summarized the difficulties with the following three excuses.

(1) Small n large p
(2) NP-hard
(3) It is difficult to separate signal and noise in high-dimensional gene space.

These excuses are mainly for statistical discriminant functions based on variance–covariance matrices because these discriminant functions could not discriminate LSD correctly. This fact is the reason why researchers could not solve the cancer gene analysis since 1970.

On the other hand, we developed the theory (Shinmura 2016d) that solved this theme only 54 days in 2015 as follows.

(1) Six microarrays are LSD and have the Matryoshka structure (Fact3).
(2) Method2 finds the microarrays consist of many SMs and other noise subspace very easy (Fact4). That is, three excuses are right only for statistical discriminant functions based on variance–covariance matrices. Why did not statistical researchers know that the microarray was LSD? This answer is that they did not know the essential definition of the signal. There was no exact definition of the signal until now. Furthermore, they could not solve Problem5 by the statistical discriminant functions and decompose the microarrays into many SMs. Chapter 1 explains these two new facts clearly. After Chap. 3, we explain the reasons why the statistical discriminant functions cannot discriminate the microarrays and all SMs from the viewpoints of many examinations.

Acknowledgements We can achieve our research by the dominant software such as LINGO supported by LINDO Systems Inc. and JMP backed by SAS Institute Japan Ltd. JMP Japan Division.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA, 96(12):6745–6750

Anderson E (1945) The irises of the Gaspe Peninsula. Bull Am Iris Soc 59:2–5

Brahim AB, Lima M (2014) Hybrid instance based feature selection algorithms for cancer diagnosis. Pattern Recogn Lett 8

Buhlmann P, Geer AB (2011) Statistics for high-dimensional data-method, theory, and applications. Springer, Berlin

Charikar M, Gurus V, Kumar R, Rajagopalan S, Saha A (2000) Combinatorial feature selection problems. IEEE Xplore, pp 631–640

Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood 103/7: 2771–2778

Cox DR (1958) The regression analysis of binary sequences (with discussion). J Roy Stat Soc B 20:215–242

Diao G, Vidyashankar AN (2013) Assessing genome-wide statistical significance for large p small n problems. Genetics 194:781–783

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27–39

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7:179–188

Fisher RA (1956) Statistical methods and statistical inference. Hafner Publishing Co., New Zealand

Flury B, Riedwyl H (1988) Multivariate statistics: a practical approach. Cambridge University Press, New York

Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):165–175

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Goodnight JH (1978) SAS technical report—the sweep operator: its importance in statistical computing—R(100). SAS Institute Inc, USA

Ishi A, Yata K, Aoshima M (2014) Asymptotic distribution of the largest eigenvalue via geometric representations of high-dimensional, low-sample-size data. Sri Lankan J Appl Statist, Special issue: modern statistical methodologies in the cutting edge of science (ed. Mukhopadhyay N): 81–94

Jeffery IB, Higgins DG, Culhane C (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinf 7(1):359 https://doi.org/10.1186/1471-2105-7-359

Lachenbruch PA, Mickey MR (1968) Estimation of error rates in discriminant analysis. Technometrics 10(1):11

Markowitz HM (1959) Portfolio selection, efficient diversification of investment. Wiley, USA

Miyake A, Shinmura S (1976) Error rate of linear discriminant function. In: Dombal FT, Gremy F (ed) North-Holland Publishing Company, The Netherland, pp 435–445

Miyake A, Shinmura S (1980) An algorithm for the optimal linear discriminant function and its application. Japan Soc Med Electron Biol Eng 1815:452–454

Nomura Y, Shinmura S (1978) Computer-assisted prognosis of acute myocardial infarction. MEDINFO 77, In: Shires W (ed) IFIP, North-Holland Publishing Company, pp 517–521

Sall JP (1981) SAS regression applications. SAS Institute Inc. USA (Shinmura S. translate Japanese version)

Sall JP, Creighton L, Lehman A (2004) JMP start statistics, 3rd edn. SAS Institute Inc. USA (Shinmura S. edits Japanese version)

Schrage L (1991) LINDO—an optimization modeling systems. The Scientific Press, USA (Shinmura S, Takamori H translate Japanese version)

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shimizu T, Tsunetoshi Y, Kono H, Shinmura S (1975). Classification of subjective symptoms of junior high school students affected by photochemical air pollution. J Jpn Soc Atmos Environ 9(4):734–741. Translated for NERC Library, EPA, from the original Japanese by LEO Cancer Associates, P.O. Box 5187 Redwood City, California 94063, Nov 1975 (TR 76-213)

Shinmura S, Kitagawa M, Takagi Y, Nomura Y (1973) The spectrum diagnosis by a two-stage weighting. In: The 12th conference of BME, pp107–108

Shinmura S, Kitagawa M, Nomura Y (1974) The spectrum diagnosis (Part 2). In: The 13th conference of BME, pp 414–415

Shinmura S, Miyake A (1979) Optimal linear discriminant functions and their application. COMPSAC 79:167–172

Shinmura S, Suzuki T, Koyama H, Nakanishi K (1983) Standardization of medical data analysis using various discriminant methods on a theme of breast diseases. MEDINFO 83, In Vann Bemmel JH, Ball MJ, Wigertz O (ed) North-Holland Publishing Company, pp 349–352

Shinmura S (1984) Medical data analysis, model, and OR. Oper Res 29(7):415–421

Shinmura S, Iida K, Maruyama C (1987) Estimation of the effectiveness of cancer treatment by SSM using a null hypothesis model. Inf Health Soc Care 7(3):263–275. https://doi.org/10.3109/14639238709010088

Shinmura S (1998) Optimal linear discriminant functions using mathematical programming. J Japanese Soc Comput Stat 11(2):89–101

Shinmura S, Tarumi T (2000) Evaluation of the optimal linear discriminant functions using integer programming (IP-OLDF) for the normal random data. J Japanese Soc Comput Stat 12(2):107–123

Shinmura S (2000a) A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics 5:133–142

Shinmura S (2000b) Optimal linear discriminant function using mathematical programming. Dissertation, Okayama University, Japan, p 101, March 2000

Shinmura S (2001) Analysis of effect of SSM on 152,989 cancer patient. ISI2001.1–2. https://doi.org/10.13140/rg.2.1.30779281

Shinmura S (2003) Enhanced algorithm of IP-OLDF. ISI2003 CD-ROM, pp 428–429

Shinmura S (2004) New algorithm of discriminant analysis using integer programming. IPSI 2004 Pescara VIP Conference CD-ROM, pp 1–18

Shinmura S (2005) New age of discriminant analysis by IP-OLDF—beyond fisher's linear discriminant function. ISI2005, pp 1–2

Shinmura S (2007a) Overviews of discriminant function by mathematical programming. J Japanese Soc Comput Stat 20(12):59–94

Shinmura S (2007b) Comparison of revised IP-OLDF and SVM. ISI2009, pp 1–4

Shinmura S (2009) Practical discriminant analysis by IP-OLDF and IPLP-OLDF. In: IPSI 2009, Belgrade VIPSI Conference, CD-ROM, pp 1–17

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

Shinmura S (2010b) Improvement of CPU time of revised IP-OLDF using linear programming. J Japanese Soc Comput Stat 22(1):39–57

Shinmura S (2011a) Beyond fisher's linear discriminant analysis—new world of the discriminant analysis. ISI2011 CD-ROM, pp 1–6

Shinmura S (2011b) Problems of discriminant analysis by mark sense test data. Japanese Soc Appl Stat 4012:157–172

Shinmura S (2013) Evaluation of optimal linear discriminant function by 100-fold cross-validation. ISI2013 CD-ROM, pp 1–6

Shinmura S (2014a) End of discriminant functions based on variance-covariance matrices. ICORE2014, pp 5–16

Shinmura S (2014b) Improvement of CPU time of linear discriminant functions based on MNM criterion by IP. Stat Optim Inf Comput 2:114–129

Shinmura S (2014c) Comparison of linear discriminant functions by K-fold cross-validation. Data Anal 2014:1–6

Shinmura S (2015a) The 95% confidence intervals of error rates and discriminant coefficients. Stat Optim Inf Comput 2:66–78

Shinmura S (2015b) A trivial linear discriminant function. Stat Optim Inf Comput 3:322–335. https://doi.org/10.19139/soic.20151202

Shinmura S (2015c) Four serious problems and new facts of the discriminant analysis. In: Pinson E, Valente F, Vitoriano B (ed) Operations research and enterprise systems. Springer, Berlin, pp 15–30. ISSN 1865-0929, ISBN 978-3-319-17508-9, https://doi.org/10.1007/978-3-319-17509-6)

Shinmura S (2015d) Four problems of the discriminant analysis. ISI 2015:1–6

Shinmura S (2015e) The discrimination of microarray data (Ver. 1). Res Gate 1–4

Shinmura S (2015f) Feature selection of three microarray data. Res Gate 1–7

Shinmura S (2015g) Feature selection of microarray data—Shipp et al microarray data. Res Gate 1–11

Shinmura S (2015h) Validation of feature selection—Alon et al microarray data. Res Gate 1–11

Shinmura S (2015i) Repeated feature selection method for microarray data. Res Gate 1–12

Shinmura S (2015j) Comparison fisher's LDF by JMP and revised IP-OLDF by LINGO for microarray data. Res Gate 1–10

Shinmura S (2015k) Matryoshka trap of feature selection method—Golub et al microarray data. Res Gate 1–14

Shinmura S (2015l) Minimum sets of genes of Golub et al. Microarray Data. Research Gate: 1.12

Shinmura S (2015m) Complete lists of small matryoshka in Shipp et al. microarray data. Res Gate 1–81

Shinmura S (2015n) Sixty-nine small matryoshka in Golub et al. microarray data (9). Res Gate 1–58

Shinmura S (2015o) Simple structure of Alon et al. microarray data. Res Gate (10):1–34

Shinmura S (2015p) Feature selection of Singh et al. microarray data. Res Gate (11):1–89

Shinmura S (2015q) Final list of small matryoshka in Tian et al. microarray data. Research Gate (12):1–60

Shinmura S (2015r) Final list of small matryoshka in Chiaretti et al. microarray data. Research Gate (13):1–16

Shinmura S (2015s) Matryoshka feature selection method for microarray data. Research Gate (14):1–16

Shinmura S (2016a) Matryoshka feature selection method for microarray data. Biotechnol 2016:1–8 (Best Paper Award)

Shinmura S (2016b) The best model of Swiss bank note data. Stat Optim Inf Comput 4:118–131. https://doi.org/10.19139/soic.v4i2.178, ISSN 2310-5070 (online), ISSN 2311-004X (print)

Shinmura S (2016c) Discriminant analysis of the linearly separable data—Japanese 44 cars. J Stat Sci Appl 4(7–8):165–178. https://doi.org/10.17265/2328-224x/2016.0708.001

Shinmura S (2016d) New theory of discriminant analysis after R. Fisher. Springer. ISBN 978-981.10-2163-3, ISBN 978-981.10-2164-0 (eBook), https://doi.org/10.1007/978-981.10-2164-0

Shinmura S (2016e) The 100-fold cross-validation for small sample. Data Anal 2016:1–8

Shinmura S (2017a) From cancer gene to cancer gene diagnosis. Amazon

Shinmura S (2017b) Examination of 64 small matryoshka (SM) of Alon et al. microarray microarray. Biotechno2017 1–8

Shinmura S (2017c) Cancer gene analysis by Singh et al. microarray data. ISI2017, pp 1–6

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD'18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. In: Biocomp'18, pp 1–7

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1.1):68–74. https://doi.org/10.1038/nm0102-6

Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. J Comput Graph Stat 22:231–245

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Lada M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(2):203–209

Stam A (1997) Non-traditional approaches to statistical classification: some perspectives on Lp-norm methods. Ann Oper Res 74:1–36

Taguchi G, Jugular R (2002) The Mahalanobis-Taguchi strategy—a pattern technology system. Wiley

Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD (2003) The Role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. N Engl J Med 349(26):2483–2494

Vapnik V (1995) The nature of statistical learning theory. Springer

Chapter 2
Overview of Cancer Gene Diagnosis

Abstract This chapter explains the cancer gene diagnosis using all Small Matryoshkas (SMs) of six microarrays found in 2016. Section 2.2 explains the different role of cancer gene analysis and cancer gene diagnosis because these technical terms are our original ones. Section 2.3 shows the analysis of 64 SMs obtained by RIP using Alon's microarray. Section 2.4 shows the usefulness of 64RIP discriminant scores (RipDSs) and new data made by 64 RipDSs instead of 2,000 genes. Thus, we consider RipDSs new data is signal instead of 64 SM. Section 2.5 shows the same analysis of 130 BGSs of Alon's microarray found by LINGO Program4 in 2016. BGS is as same as the Yamanaka's four genes in iPS research. Section 2.6 shows the cancer gene diagnosis of other five microarrays those are analyzed in the same way as Alon. Section 2.7 is the conclusion. Alon and Singh's microarrays consist of cancer and normal classes. Other four microarrays consist of two different types of cancer classes. It is vital for us that six results are almost the same. Thus, we expect another microarray's result is as same as our results if medical researchers control two classes strictly.

Keywords Gene diagnosis · Malignant indicators · Small Matryoshka (SM) · Basic gene subspace (BGS) · Discriminant scores (DSs) · RatioSV of RIP and PCA

2.1 Introduction

We developed the new theory of discriminant analysis (theory) and solved five problems of discriminant analysis by 2015 (Shinmura 2016). Since 1970, many statisticians and engineers failed to identify oncogenes from microarrays (Problem5) because statistical discriminant functions were useless for cancer gene analysis. Mainly, we could completely solve the cancer gene analysis (Problem5) as an applied problem of our theory. Six medical projects published their articles in prominent medical journals from 1999 to 2004 and released the microarrays on the Internet (Golub et al. 1999; Alon et al. 1999; Shipp et al. 2002; Singh et al. 2002; Tian et al. 2003; Chiaretti et al. 2004). When Revised IP-OLDF (RIP) and the Matryoshka feature selection method (Method2) analyzed six microarrays, our cancer gene analysis

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_2

quickly completed in 54 days of 2015. RIP and Method2 could decompose microarrays into many SMs. Chapter 1 explains these results. Our cancer gene analysis was the first successful example of “big data analysis or high-dimensional data analysis” which was the latest leading theme in statistical analysis. We found the five problems and two facts of the discriminant analysis and solved five problems entirely.

This chapter explains the cancer gene diagnosis by statistical analysis of all SMs found in 2016 (Shinmura 2017). Section 2.2 explains the different role of cancer gene analysis and cancer gene diagnosis because these technical terms are our original ones. Although medical research authorized about 100 “oncogenes,” we define the gene combination included in each SM found by Method2 as “cancer genes.” Thus, we claim we can first succeed in cancer gene analysis. However, physicians need to examine and validate our malignant indicators are useful for cancer gene diagnosis. Section 2.3 shows the analysis of 64 SMs obtained by RIP using Alon’s microarray. The standard statistical methods analyze the 64 SMs. We choose SM8 because its RatioSV is large. One-way ANOVA with t-test indicates there are three types of t-values such as the positive, almost zero, and negative values. These results tell us a proper combination of thee-types genes can separate two classes. If we omit genes that have t-values near zero, MNM will exceed 1. We claimed we could first succeed in cancer gene analysis. However, we expect that physicians examine and validate our malignant indicators that are useful for cancer gene diagnosis because we are not the specialist in this area. If they confirm our results, we will be able to open the new frontier of cancer gene diagnosis shortly. We confirmed MNM using 47 oncogenes was not zero by data gathered at Japanese cancer blood testing center. However, our all results showed two classes were separable in the proper gene combinations such as SM or BGS. Section 2.4 shows the usefulness of 64 RIP discriminant scores (RipDSs) and new data made by 64 RipDSs. PCA and Ward cluster analyze new data and obtain the linear separable fact, and PCA shows several outliers. If we analyze the transpose data of new data, we obtain many outliers those are expected to be new classes of cancer pointed out by Golub et al. Sect. 2.5 shows the same analysis of 130 BGSs of Alon’s microarray found by LINGO Program4 in 2016. BGS is as same as the Yamanaka’s four genes in iPS research. If we drop one gene from BGS, MNM of which is greater than 1 and noise. Because 130 RatioSVs of BGSs are less than 1%, we think BGS is not useful for cancer gene diagnosis. We must validate the results of BGS by Method1. Probably, BGS is essential to survey the role of the cancer gene. We examine BGS128 by standard statistical methods because its RatioSV is the maximum value 0.9%. Section 2.6 shows the cancer gene diagnosis of other five microarrays analyzed in the same way as Alon. Section 2.7 is the conclusion. Alon and Singh’s microarrays consist of the cancer and normal classes. Other four microarrays consist of two different types of cancer classes. It is vital for us that six results are almost the same. Thus, we expect another microarray’s result is as same as our results if medical research strictly controls two classes.

Chapter 1 introduced theory about five problems and two facts. After this chapter, we discuss the cancer gene diagnosis using cancer malignancy indicators.

2.2 Cancer Gene Diagnosis

In this book, all chapter, except Chaps. 1 and 10, discuss cancer gene diagnosis analyzing SMs found by the RIP and Revised LP-OLDF.

(1) Why could not the standard statistical methods find the linearly separable fact in SMs?

At first, we expect the standard statistical methods can find the linear separable fact that two classes are separable in each SM. Although we tried to consider the meaning of PCA and cluster analysis, we could not find useful results. Thus, we recognized our trials are meaningless. On the other hand, three OLDFs and H-SVM can discriminate all SMs entirely, and there are many RatioSVs over 5%. We find many facts that explain the reason why the standard statistical methods cannot find the linearly separable fact in SMs. However, because RIP, Revised LP-OLDF, and H-SVM can discriminate all SMs completely, we make several new data made by RIP discriminant scores (RipDSs), Revised LP-OLDF DSs (LpDSs) and H-SVM DSs (HsvmDSs) as variables instead of genes. Although there are many malignancy indicators by RipDSs, LpDSs, and HsvmDSs, Prin1 is another malignancy indicator. Many outliers found by PCA may be the new subclasses of cancer pointed out by Golub et al. (1999). They developed new methods of cancer gene analysis and cancer gene diagnosis. However, their methods are difficult for researchers without the background of medical bits of knowledge. On the other hand, our methods are the standard statistical methods and offer many pieces of information for both medical experts and medical non-experts.

(2) This Book Conclusion

This book shows that only RIP and Revised LP-OLDF can decompose microarrays into many SMs. Although the standard statistical methods cannot find the linear separable fact, RIP, Revised LP-OLDF, and H-SVM can separate both SMs found by the RIP and Revised LP-OLDF into two classes entirely. Thus, we obtain many malignancy indicators made by three LDFs using both different types of SMs. We make six types of new data by RipDSs, LpDSs, and HsvmDSs as variables. However, five chapters from Chaps. 2 to Chap. 6 introduce the results of new data made by the combination of RIP using SM found by RIP. Chapters 7–9 introduce the new data made by LpDSs. Our examinations show that all malignancy indicators by RipDSs, LpDSs, and HsvmDSs are useful for cancer gene diagnosis. Moreover, PCA and several hierarchical cluster analyses show almost the same results. It seems that for the data managed for research, microarray data may give almost the same results as shown in this book. Outliers found by PCA may be useful to find new subclasses of cancer.

2.3 Analysis of 64 SMs Obtained by Alon's Microarray

In this section, LINGO Program3 (Schrage 2006) of RIP can discriminate Alon's microarray correctly and separate the microarray into 64 SMs (1,999 genes) and noise subspace (one gene). We omit this one gene from our analysis. Alon et al. analyzed 6,500 genes by SOM and identified 2,000 genes as oncogenes. Because both ways got nearly the same result that 2,000 genes are oncogenes and cancer genes[1] included in 64 SMs, it shows the validity of each other's method. Although Revised LP-OLDF can select cancer genes naturally as same as RIP, we do not discuss these SMs because we wish to accomplish our analysis of SM as soon as possible in Alon's microarray. At first, standard statistical methods analyze 64 SMs that consist of 62 subjects (62 cases) and 1,999 genes (1,999 variables). Because all NMs of logistic regression are zero, logistic regression confirms 64 SMs as signals. However, other standard statistical methods do not show two classes are separable in most SMs. On the other hand, when analyzing new data consisting of 62 subjects and 64 RipDSs (64 variables), surprising results are found that two classes are entirely separable in new data by PCA and cluster analysis. These two different results reveal the reason why cancer gene analysis and cancer gene diagnosis are difficult until now (Problem6). Now, only three OLDFs can decompose microarray into signals and noise. H-SVM and statistical discriminant functions are useless for cancer gene analysis.

2.3.1 Analysis of 64 SMs

2.3.1.1 NMs of 64 SMs by Four Statistical Discriminant Functions

Table 2.1 shows 64 SMs from SM = 1 to SM = 64 that is the order found by RIP. This table is sorted in descending order by RatioSV. "Gene" column is the number of genes included in each SM. The range of gene number is [21, 42] and 64 SMs include 1,999 genes. LP and IP select less 62 nonzero coefficients from 2,000 genes. The maximum number of actual nonzero coefficients is 42. Because the ratio of signal (RatioS) is 99.9% (=1,999/2,000 * 100%), only this microarray has high RatioS compared with other five microarrays. It was worthy of praise that SOM could select 2,000 oncogenes from 6,500 genes. Although anyone cannot achieve this achievement, we consider it is not a contribution of SOM but a result of medical knowledge. From a statistical point of view, it is surprising that cluster analysis is so useful for oncogenes research. Furthermore, in a different approach according to Method2, it decomposes 2000 genes into nearly 64 SMs and 130 BGSs, with low noise. These results indirectly indicate that their approach and Method2 are appropriate. However, the usage of Method2 is easy for many researchers who do not have medical knowledge.

[1] Cancer gene means a set of genes included in SM. Those genes separate two classes entirely.

Table 2.1 NMs of 64 SMs

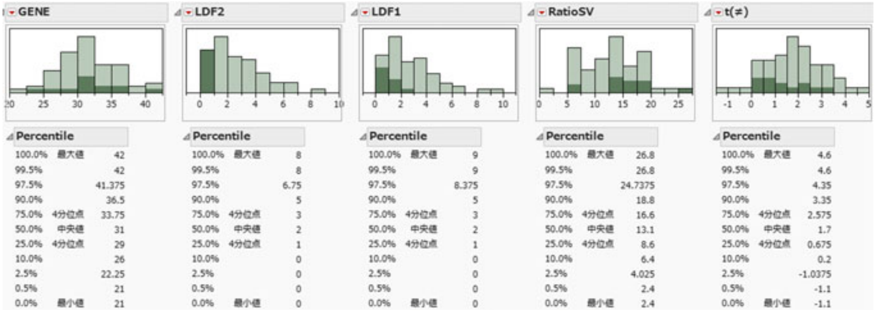
SM	GENE	Logistic	QDF	LDF2	LDF1	DS	RatioSV	t (≠)
8	31	0	0	0	0	7.5	26.8	0.3
35	30	0	0	1	1	8.5	23.5	4.6
11	25	0	0	2	2	9.7	20.7	3.3
53	29	0	0	3	3	10.1	19.8	2.4
27	28	0	0	1	1	10.2	19.6	2.6
–	–	–	–	–	–	–	–	–
63	40	0	0	6	8	32.4	6.2	3.5
7	41	0	0	0	1	32.2	6.2	3
59	36	0	0	6	6	37.2	5.4	1.3
14	26	0	0	0	0	39.8	5	0.5
64	42	0	0	8	9	84.9	2.4	3.5
Max	42	0	0	8	9	84.9	26.8	4.6
Mean	31.23	0	0	2.1	2.17	19.04	12.84	1.72
Min	21	0	0	0	0	7.5	2.4	−1.1
Sum	1999	0	0	134	139	1218	821.8	110.3
NM = 0		64	64	13	12			

All NMs of logistic regression are zero. Because all NMs of QDF are 0 also, the distance of two-class averages may be larger than the condition of the difference of two variance–covariance matrices (Aoshima and Yata 2017; Yata and Aoshima 2010). "LDF2 and LDF1" are NMs of two different prior probability options of Fisher's LDFs. The prior probability of LDF2 is proportional to the case number of 22:40. That of LDF1 is "1:1" that is default in much statistical software. However, we used the proportional prior probability because we wish to compare NMs of six MP-based LDFs in our research. The NM's ranges of LDF2 and LDF1 are [0, 8] and [0, 9], respectively. The 13 NMs of LDF2 and 12 NMs of LDF1 are zero. Because three LSD ratios of QDF, LDF1, and LDF2 are 64/64 = 100%, 13/64 = 20.3%, and 12/64 = 18.7%, these results are better than the other five microarrays. Thus, we forecast that two classes of Alon are fairly separable in each SM. "DS" is the range of RipDSs. The range of 64 RipDSs is [7.5, 84.9]. "RatioSV" is the value calculated by "2/DS * 100 (%)" that indicates the ratio of the SV width and the RipDSs width. The range of 64 RatioSVs is [2.4, 26.8]. The eighth RipDF (RipDS8) has the maximum RatioSV among 64 RipDSs. RipDS64 has the minimum RatioSV less than 5%. The RatioSV recommends RipDS8 because it is the maximum value of 64 RipDSs. We claim RatioSV is the best index for the LSD-discrimination of two classes and is helpful for cancer gene diagnosis. "t" column is t-value to test the mean's difference of the RipDSs. The ranges of t-values are [−1.1, 4.6]. Furthermore, we surveyed all t-values of genes included in each RipDSs. Those values are either of minus, almost zero, and positive values, not only positive values. On the other hand, some papers claimed high positive t-values or Welch values are oncogenes. However, our results of t-test showed that a t-test or Welch's test was not helpful to find cancer genes.

Our Claim: The t-test and Welch test may be useless for cancer gene diagnosis.

2.3.1.2 Histogram and Correlation

Figure 2.1 is a histogram of gene, LDF2, LDF1, RatioSV, and t-values. If we select the case "NM of LDF 2 equals 0," these cases will be dark green in other variables also. The dark green cases of "gene, RatioSV, and t (≠)" spread throughout the range. On the other hand, dark green cases of LDF1 are less than 3. This fact indicates one of the NM's defects that cannot find the linear separable fact at all. In general, examining the histogram is essential. However, it is more critical whether MNM = 0 or not, so the results that do not contribute to this are not meaning at all. We obtained the results of various studies and conducted wasteful investigations.



GENE

Percentile		
100.0%	最大値	42
99.5%		42
97.5%		41.375
90.0%		36.5
75.0%	4分位点	33.75
50.0%	中央値	31
25.0%	4分位点	29
10.0%		26
2.5%		22.25
0.5%		21
0.0%	最小値	21

LDF2

Percentile		
100.0%	最大値	8
99.5%		8
97.5%		6.75
90.0%		5
75.0%	4分位点	3
50.0%	中央値	2
25.0%	4分位点	1
10.0%		0
2.5%		0
0.5%		0
0.0%	最小値	0

LDF1

Percentile		
100.0%	最大値	9
99.5%		9
97.5%		8.375
90.0%		5
75.0%	4分位点	3
50.0%	中央値	2
25.0%	4分位点	1
10.0%		0
2.5%		0
0.5%		0
0.0%	最小値	0

RatioSV

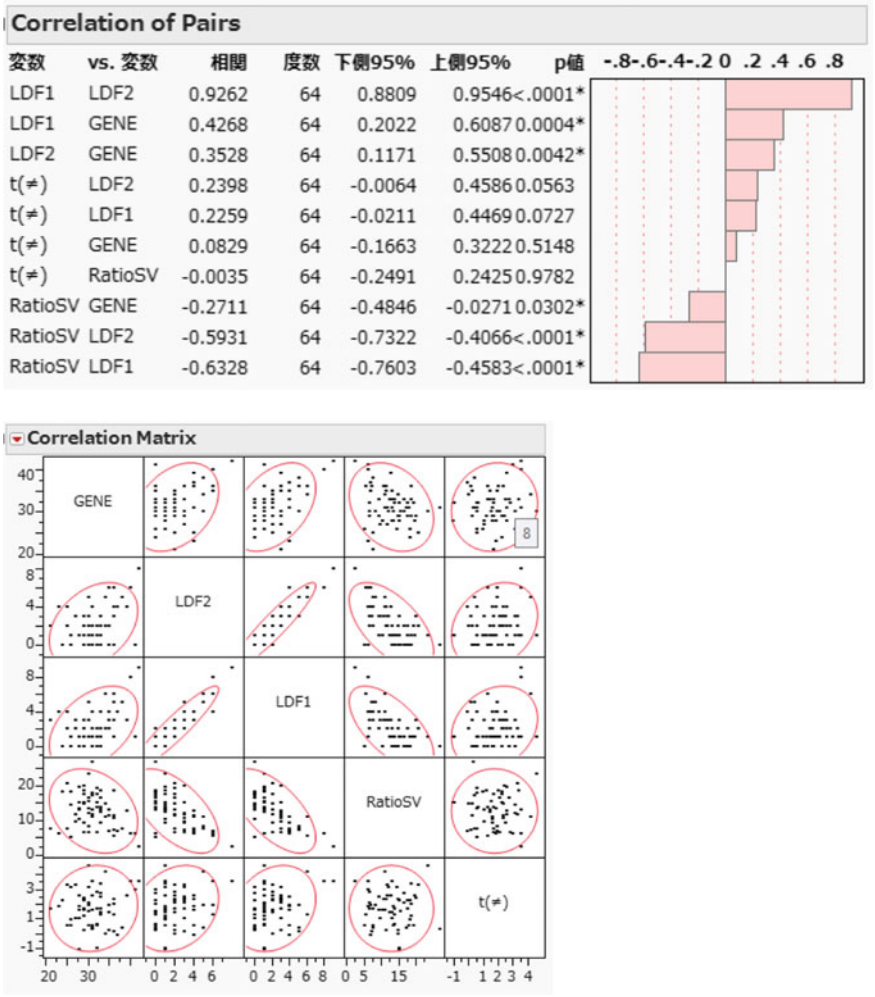
Percentile		
100.0%	最大値	26.8
99.5%		26.8
97.5%		24.7375
90.0%		18.8
75.0%	4分位点	16.6
50.0%	中央値	13.1
25.0%	4分位点	8.6
10.0%		6.4
2.5%		4.025
0.5%		2.4
0.0%	最小値	2.4

t(≠)

Percentile		
100.0%	最大値	4.6
99.5%		4.6
97.5%		4.35
90.0%		3.35
75.0%	4分位点	2.575
50.0%	中央値	1.7
25.0%	4分位点	0.675
10.0%		0.2
2.5%		-1.0375
0.5%		-1.1
0.0%	最小値	-1.1

Fig. 2.1 Histograms of gene, QDF, LDF2, and LDF1, RatioSV and t (≠)

Figure 2.2 is the matrix correlation of five variables. The upper figure shows the correlation of ten pairs. We focus on the three positive correlations of (gene, LDF1, and LDF2). The positive correlations indicate that the larger the number of genes, the higher the number of misclassifications. That is, SM containing many genes tends to have a large NM. Three correlations between RatioSV and (gene, LDF1, and LDF2) are negative correlations. These correlations indicate that the smaller the number of genes and two NMs, the larger the value of RatioSV. Four correlations between the t-test with the other four variables are almost zero. These correlations show that the t-tests are uncorrelated with the number of genes and two NMs. Again, we emphasize that using the t-test is useless for cancer gene diagnosis. The figure below is a scatter plot matrix of five variables.



Correlation of Pairs

変数	vs. 変数	相関	度数	下側95%	上側95%	p値
LDF1	LDF2	0.9262	64	0.8809	0.9546	<.0001*
LDF1	GENE	0.4268	64	0.2022	0.6087	0.0004*
LDF2	GENE	0.3528	64	0.1171	0.5508	0.0042*
t(≠)	LDF2	0.2398	64	-0.0064	0.4586	0.0563
t(≠)	LDF1	0.2259	64	-0.0211	0.4469	0.0727
t(≠)	GENE	0.0829	64	-0.1663	0.3222	0.5148
t(≠)	RatioSV	-0.0035	64	-0.2491	0.2425	0.9782
RatioSV	GENE	-0.2711	64	-0.4846	-0.0271	0.0302*
RatioSV	LDF2	-0.5931	64	-0.7322	-0.4066	<.0001*
RatioSV	LDF1	-0.6328	64	-0.7603	-0.4583	<.0001*

Fig. 2.2 Matrix correlation of five variables

2.3.2 Analysis of RipDS8 by Standard Statistical Methods

2.3.2.1 One-Way ANOVA with T-Test

We analyze RipDS8 by standard statistical methods because the RatioSV of RipDS8 has the maximum value among 64 SMs. Four NMs of logistic regression, QDF, and both LDFs are zero. Although this fact shows that QDF and Fisher's LDF can discriminate RipDS8 correctly in addition to logistic regression, this is a rare case. However, it is crucial other statistical methods cannot show the linear separable fact. This fact indicates the discriminant analysis is the best methods for LSD-discrimination and cancer gene analysis. QDF and Fisher's LDF often cannot discriminate SM correctly for other five microarrays. Table 2.2 is the result of one-way ANOVA of RipDS8 that consists of 31 genes. Two columns "Min and Max" are the range of class1 (normal, 22 cases). "MIN and MAX" are the range of class2 (tumor, 40 cases). All two classes overlap. However, RipDS8 is linearly separable that two classes are entirely separable in RipDS8. Standard statistical methods cannot show the linear separable fact of SM8 (Problem6). Two t-tests are done to test the difference of means on each gene. "t (≠)" is the t-test value under two variances which are not equal, and "t (=)" is the t-test value in the case of two variances which are equal. The table is sorted in descending order by the value of "t (≠)." The range of t-values is [−6.5, 4.27]. Some papers expected t-value was useful for cancer gene analysis and claimed that some genes with sizeable positive t-value were oncogenes. We doubt their claims because two results in Tables 2.1 and 2.2 denied their claim.

Table 2.2 Result by one-way ANOVA of RIPDS8

Gene	Min	Max	MIN	MAX	t (≠)	t (=)
X1473	4.74	8.12	4.89	9.55	4.20	4.27
X698	6.37	8.53	6.50	10.46	3.95	3.43
X1896	3.91	6.60	4.17	6.75	2.27	2.51
X1859	5.11	7.17	4.55	7.66	2.22	2.20
X1485	5.89	10.59	4.87	10.58	2.04	1.92
X1662	5.44	7.15	5.37	7.86	1.70	1.70
X1961	4.44	7.08	4.39	7.98	1.60	1.53
X1464	6.62	11.02	5.16	11.62	1.21	1.11
X1024	6.16	11.50	5.72	11.09	1.20	1.27
S8X6	10.99	13.07	10.28	13.07	0.84	0.87
X1606	4.61	9.72	3.76	10.29	0.60	0.65
X1706	4.69	7.84	4.76	7.96	0.30	0.31
X1077	4.39	9.50	4.30	10.16	0.19	0.20
X1571	3.76	7.05	5.18	7.57	0.17	0.18
X1883	5.02	7.38	4.35	7.33	0.01	0.01
X1207	5.11	7.23	5.29	7.39	−0.01	−0.01
X1177	5.83	9.23	6.30	9.44	−0.34	−0.34
X1228	5.75	7.62	5.52	7.87	−0.51	−0.51
X1958	5.16	7.49	3.50	7.52	−0.91	−0.81
X1433	5.33	9.74	4.47	10.05	−1.03	−0.99
X1295	4.85	7.53	5.10	7.55	−1.23	−1.16
X1448	4.93	7.53	4.39	6.86	−1.42	−1.36
X1182	5.80	9.10	5.86	8.83	−1.54	−1.56
X404	7.30	9.15	6.65	10.24	−1.74	−1.58
X1842	5.50	7.65	4.47	7.75	−2.46	−2.33

(continued)

Table 2.2 (continued)

Gene	Min	Max	MIN	MAX	t (≠)	t (=)
X1348	7.08	10.05	4.94	10.07	−2.94	−2.64
X1642	6.22	8.54	4.35	8.48	−3.06	−2.73
X119	10.22	13.07	7.50	12.66	−3.46	−3.07
X1387	6.21	11.52	4.76	11.43	−4.25	−4.14
X14	11.18	12.84	10.18	12.84	−5.24	−5.12
X1423	6.03	10.75	3.50	9.84	−6.68	−6.50

Figure 2.3 is the box–whisker plots of two classes. Three t-values of X1423, X14, and X1387 are −6.68, −5.24, and −4.25, respectively. Because three averages of class2 (tumor subjects) are less than those of class1 (normal subjects), these genes may prevent cancer disease (suppressor of cancer). Although these three genes are high negative values, two classes are not linearly separable. We investigate 64 SMs and get the same results. Although some cancer gene researchers expect there is one oncogene that can discriminate two classes completely, our examination by one-way ANOVA suggests their expectations are wrong. Our study shows that the appropriate set of genes included in SM can correctly distinguish between cancer and normal classes.

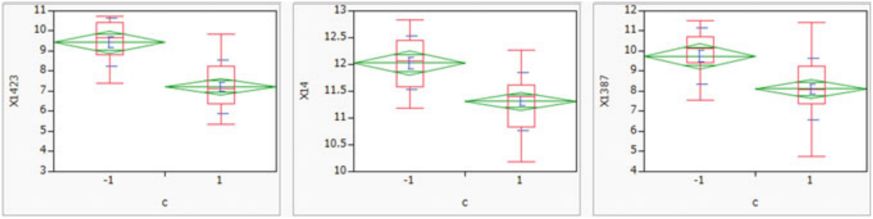


Fig. 2.3 Box–whisker plots of two classes

2.3.2.2 Ward Cluster Analysis of RipDS8

We analyze RipDS8 in Ward cluster analysis. Figure 2.4 is a heat map of RipDS8 (64 cases and 31 genes). The right side is the dendrogram of 22 normal cases with marks □ and 40 tumor subjects with marks ×. Although it is difficult to decide the proper number of clusters, we accept four clusters. However, four clusters include normal and tumor subjects in each cluster. In the bottom dendrogram of the gene, we cannot classify genes into clear clusters. In this case, it is meaningless to analyze the variable dendrogram displayed below. However, sixth and seventh genes from the right firstly join at a small distance. These are thought to be mutually substitutable. Nevertheless, four discriminant functions can discriminate RipDS8 completely in Table 2.1. This fact indicates that the expression level of genes cannot classify two classes well (Problem6). Moreover, we understand the discriminant functions are the best methods to discriminate between two classes. However, we concluded that it would be better not to analyze the result of each SM in detail. That is, we consider the gene contained in SM as an oncogene. However, standard statistical methods cannot find linearly separable fact. Therefore, RIP, Revised LP-OLDF, and H-SVM DSs are considered to be a signal.

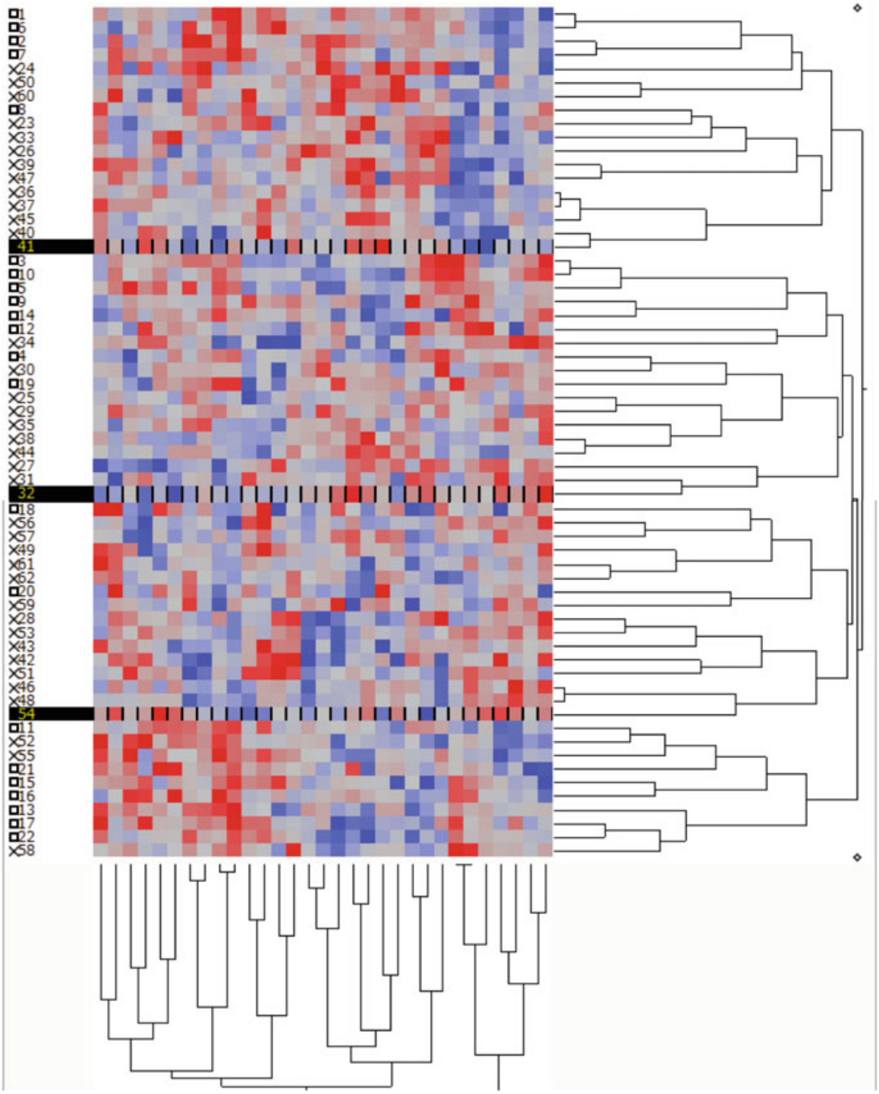


Fig. 2.4 Heat map of RipDS8 with 31 genes by the ward method

2.3.2.3 PCA and Canonical Plot of LDF2

Figure 2.5 is PCA outputs. Left plot is an eigenvalue. Ten eigenvalues from Prin1 to Prin10 are greater than one. The central plot is a scatter plot. Two classes overlap. The right plot is the factor loading plot. The 31 genes are located in four quadrants. Most factor loading of 64 SMs have the same tendencies. PCA cannot explain the meaning of two classes. In the end, we concluded that it would be better not to analyze this result in detail.

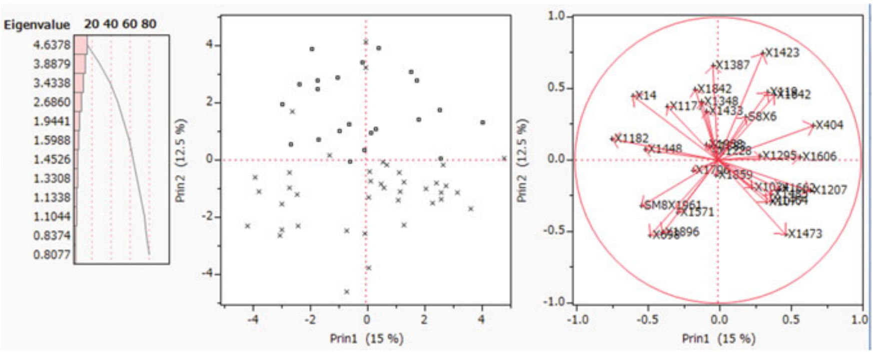


Fig. 2.5 PCA figures (eigenvalue, scatter plot, and factor loading plot)

Figure 2.6 is the canonical plot of LDF2. Because three discriminant function's NMs are zero, two classes are separable entirely. This figure is quite different from Fig. 2.5 of PCA. Thus, this result indicates the discriminant analysis is better than other statistical methods from the viewpoint of classifying two classes. On the other hand, because statistical discriminant functions are useless for discriminating microarrays, most medical researchers did not use discriminant functions after 1999.

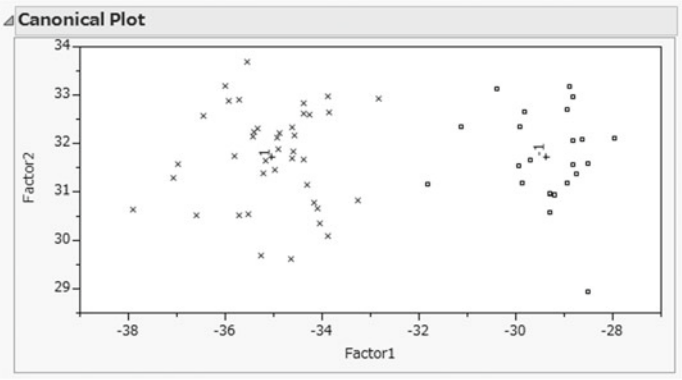


Fig. 2.6 Canonical plot of QDF

2.3.2.4 Our Conclusion of Standard Statistical Methods

At first, we expected to obtain a good result by analyzing all SMs using standard statistical methods. Although we analyzed all SMs by these methods, only logistic regression was meaningful. In this section, NMs of QDF and Fisher's LDF occasionally became zero. However, most analyses of SMs showed no useful results that two classes were completely separable. Thus, we recommended researchers did not expect the useful results except for definitive result by logistic regression.

Strange fact of cancer gene analysis (Problem5): Although two classes are entirely separable in high-dimensional microarrays and SMs, we could not observe the linear separability of two classes by the standard statistical methods. Moreover, logistic regression can show the linear separability for SMs. This fact implies the noise entirely includes the signal found by RIP. We forecast the variance of the signal is smaller than those of noise. This book shows facts by many examinations.

2.4 Analysis of 64 RipDSs Data

We claimed standard statistical methods could analyze SM very easily because each SM is a small sample (small n and small p_i (p_i <= n)). However, we cannot obtain useful results analyzing 64 SMs in Sect. 2.3. Next, when the standard statistical methods analyze 64 RipDSs new data with 62 cases and 64 RipDSs (64 variables), we get the next surprising success explained in this section.

2.4.1 Examination of 64 RipDSs and RatioSV of RIP

Table 2.3 is two ranges of two classes, the range of DS, RatioSV (=200/DS), and two t-values. Three columns such as DS, RatioSV, and t ($\neq$ 0) are the same as Table 2.1. The range of 22 cases in class1 (Min and Max columns) is less than equal −1, and the range of 40 cases in class2 (MIN and MAX columns) is greater than equal 1. SV separates two classes of 64 SMs. We consider "RatioSV of RIP" is the most important statistics for cancer gene analysis because it shows the ease of classification by two classes. The table is sorted in descending order of the value of "RatioSV of RIP." Although the SV distance is 2, it is 26.27% for the range of RipDS8. SV becomes a wide window and separate two classes completely. The last three rows are the maximum, mean, and minimum of eight variables. The range of DS, RatioSV, and two t-values are [7.47, 84.94], [2.35, 26.76], [4.22, 15.5], and [3.12, 14.76], respectively. Table 2.3 shows 64 discriminations by RIP are very easy. However, standard statistical methods are difficult to obtain the linear separable fact (Problem6). This fact implies the difficulties of cancer gene analysis until now and answers why many researchers could not succeed cancer gene analysis from 1970 because these methods are useless for microarrays and those SMs. We must choose proper methods for cancer gene diagnosis as same as cancer gene analysis.

Table 2.3 Sixty-four RipDSs, range of DS, RatioSV (=200/DS) and t-values

RipDS	Min	Max	MIN	MAX	DS	RatioSV	t (≠)	t (=)
RipDS8	−3.35	−1	1	4.12	7.47	**26.76**	15.50	14.76
RipDS35	−2.58	−1	1	5.92	8.51	**23.52**	13.02	9.94
RipDS11	−4.15	−1	1	5.52	9.67	**20.68**	12.71	11.17
RipDS53	−3.72	−1	1	6.40	10.13	**19.75**	11.98	10.20
RipDS27	−3.75	−1	1	6.45	10.20	**19.62**	13.19	11.58
RipDS46	−5.04	−1	1	5.44	10.48	**19.09**	10.73	11.04
RipDS30	−3.92	−1	1	6.86	10.79	**18.54**	11.32	9.67
RipDS33	−5.34	−1	1	5.50	10.85	**18.44**	11.46	10.77
RipDS3	−4.74	−1	1	6.14	10.88	**18.39**	11.80	10.01
RipDS25	−5.55	−1	1	5.39	10.94	**18.29**	11.66	11.45
RipDS17	−4.05	−1	1	7.01	11.06	**18.08**	12.99	11.14
RipDS15	−5.70	−1	1	5.47	11.17	**17.90**	10.52	11.01
RipDS51	−3.98	−1	1	7.35	11.34	**17.64**	11.68	9.95
RipDS42	−6.28	−1	1	5.15	11.43	**17.50**	10.47	10.38
RipDS19	−6.38	−1	1	5.32	11.70	**17.09**	10.00	10.42
RipDS9	−3.35	−1	1	8.70	12.04	**16.60**	12.13	10.00
RipDS22	−4.14	−1	1	7.90	12.05	**16.60**	10.79	9.25
RipDS6	−4.30	−1	1	8.49	12.78	**15.64**	11.21	9.93
RipDS23	−7.26	−1	1	5.93	13.19	**15.16**	9.65	10.75
RipDS10	−6.33	−1	1	6.92	13.25	**15.09**	9.66	10.26
RipDS16	−5.25	−1	1	8.10	13.35	**14.98**	10.51	9.29
RipDS48	−6.76	−1	1	6.67	13.43	**14.89**	10.03	10.22
RipDS31	−5.06	−1	1	8.43	13.50	**14.82**	10.17	9.52
RipDS1	−5.35	−1	1	8.18	13.53	**14.78**	11.02	9.80
RipDS24	−4.60	−1	1	9.18	13.78	**14.52**	10.96	9.12
RipDS21	−4.79	−1	1	9.82	14.62	**13.68**	10.05	8.70
RipDS18	−5.88	−1	1	8.77	14.66	**13.64**	10.32	9.46
RipDS2	−5.30	−1	1	9.38	14.68	**13.62**	9.25	8.07
RipDS32	−7.66	−1	1	7.19	14.85	**13.47**	9.05	9.32
RipDS34	−5.27	−1	1	9.69	14.96	**13.37**	9.73	8.29
RipDS36	−8.23	−1	1	6.97	15.20	**13.16**	9.73	10.00
RipDS5	−6.94	−1	1	8.29	15.22	**13.14**	9.42	9.04

(continued)

Table 2.3 (continued)

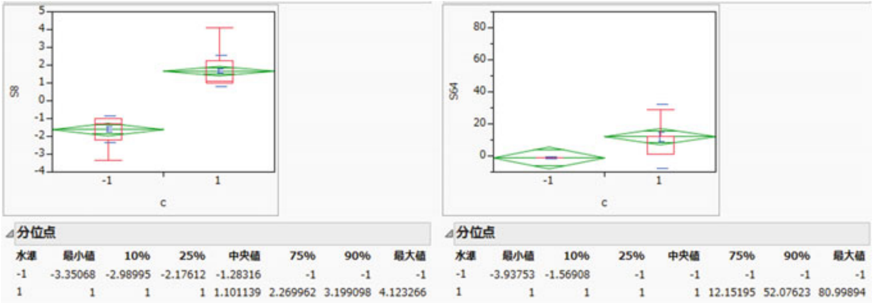
RipDS	Min	Max	MIN	MAX	DS	RatioSV	t (≠)	t (=)
RipDS20	−6.06	−1	1	9.26	15.32	**13.06**	9.12	8.42
RipDS49	−4.20	−1	1	11.23	15.43	**12.96**	9.84	7.67
RipDS56	−5.28	−1	1	10.61	15.89	**12.59**	8.02	6.44
RipDS26	−6.71	−1	1	9.26	15.96	**12.53**	9.39	8.53
RipDS54	−5.22	−1	1	11.47	16.69	**11.99**	9.19	7.34
RipDS38	−7.40	−1	1	9.67	17.07	**11.72**	9.05	7.94
RipDS29	−6.46	−1	1	10.82	17.28	**11.57**	10.05	8.55
RipDS55	−9.86	−1	1	7.82	17.68	**11.31**	8.28	8.52
RipDS58	−5.67	−1	1	12.14	17.81	**11.23**	8.73	7.26
RipDS60	−4.60	−1	1	13.44	18.05	**11.08**	9.71	7.52
RipDS52	−5.98	−1	1	12.23	18.21	**10.98**	8.95	7.61
RipDS40	−7.12	−1	1	11.69	18.81	**10.63**	8.18	6.83
RipDS43	−8.34	−1	1	10.87	19.21	**10.41**	9.21	8.41
RipDS4	−7.90	−1	1	12.87	20.77	**9.63**	9.02	8.53
RipDS45	−12.49	−1	1	9.24	21.73	**9.20**	7.58	7.66
RipDS44	−8.02	−1	1	13.82	21.84	**9.16**	8.26	7.16
RipDS37	−10.52	−1	1	13.33	23.84	**8.39**	7.88	7.07
RipDS41	−10.67	−1	1	15.54	26.21	**7.63**	8.98	8.18
RipDS39	−14.35	−1	1	11.99	26.35	**7.59**	8.20	8.43
RipDS61	−12.37	−1	1	14.67	27.04	**7.40**	6.68	6.50
RipDS62	−8.83	−1	1	19.24	28.07	**7.13**	6.65	6.12
RipDS57	−12.95	−1	1	15.70	28.66	**6.98**	6.99	6.48
RipDS12	−16.52	−1	1	13.00	29.52	**6.78**	6.06	6.46
RipDS47	−12.97	−1	1	17.10	30.07	**6.65**	7.19	6.55
RipDS50	−11.52	−1	1	18.78	30.30	**6.60**	7.59	6.95
RipDS13	−10.77	−1	1	19.77	30.54	**6.55**	8.04	6.94
RipDS28	−15.22	−1	1	16.77	31.99	**6.25**	7.69	7.58
RipDS7	−7.94	−1	1	24.25	32.19	**6.21**	6.46	5.06
RipDS63	−8.38	−1	1	24.00	32.38	**6.18**	6.22	4.72
RipDS59	−15.25	−1	1	21.91	37.17	**5.38**	6.17	5.53
RipDS14	−21.94	−1	1	17.85	39.79	**5.03**	7.16	7.42
RipDS64	−3.94	−1	1	81.00	84.94	2.35	4.22	3.12
MAX	−2.58	−1	1	81.00	84.94	26.76	15.50	14.76
MEAN	−7.35	−1	1	11.69	19.04	12.84	9.49	8.63
MIN	−21.94	−1	1	4.12	7.47	2.35	4.22	3.12

Figure 2.7 is two box–whisker plots of RipDS8 and RipDS64. Left plot is RipDS8 with maximum RatioSV, and the right plot is RipDS64 with minimum RatioSV. Two ranges are [−3.35, 4.12] and [−3.94, 81], respectively. Both ranges of DS are 7.47 and 84.94. RatioSVs of RipDS8 and RipDS64 opens windows of 26.76 and 2.35% for each DS. If we will get the validation samples, RipDS8 discriminates validation sample in two classes very easy and completely. Thus, we judged not to validate RipDS8 by Method1. However, RipDS64 may not be able to discriminate validation samples correctly. In other words, six projects are verifying using LOO, but we think there is no need to verify the malignancy indicator for the case of large RatioSV. This threshold is future research, but we think that it is enough if it is 5% or more.

Our Claim: If we use the malignancy indicator with "RatioSV >= 5%," we do not need to verify it with Method1.

On the other hand, if we remove the outlier value 81 of RipDS64, RatioSV of RipDS64 becomes large, and we obtain the different result. In this book, we do not consider the effects of the outliers of cancer gene data. The treatment of outliers is very vital and the future important research theme.

Future Research: We must consider the effect of the outliers in the cancer gene diagnosis.



分位点

水準	最小値	10%	25%	中央値	75%	90%	最大値
-1	-3.35068	-2.98995	-2.17612	-1.28316	-1	-1	-1
1	1	1	1	1.101139	2.269962	3.199098	4.123266

分位点

水準	最小値	10%	25%	中央値	75%	90%	最大値
-1	-3.93753	-1.56908	-1	-1	-1	-1	-1
1	1	1	1	1	12.15195	52.07623	80.99894

Fig. 2.7 Two box–whisker plots of RipDS8 and RipDS64

2.4.2 *Ward Cluster Analysis of RipDSs New Data*

Ward cluster analyzes the RipDSs new data (62 cases and 64 variables). Figure 2.8 is the result of two clusters of heat map and dendrogram (22 healthy cases and 40 cancer cases). Two classes become two clear clusters. The top blue cluster is 22 healthy cases in class1, and the second red cluster is 40 cancer cases in class2. The medical specialist may be able to explain over ten clusters by dendrogram of 62 subjects. However, 64 RipDSs variable dendrogram shown in under heatmap have more complex clustering. Over five pairs of two RipDSs become one cluster in the early stage of clustering. This fact may show two RipDSs in each pair can be exchanged with each other. These pairs show redundancy. Moreover, if it is possible to exchange with each other, there is a possibility that BGSs will not be unique.

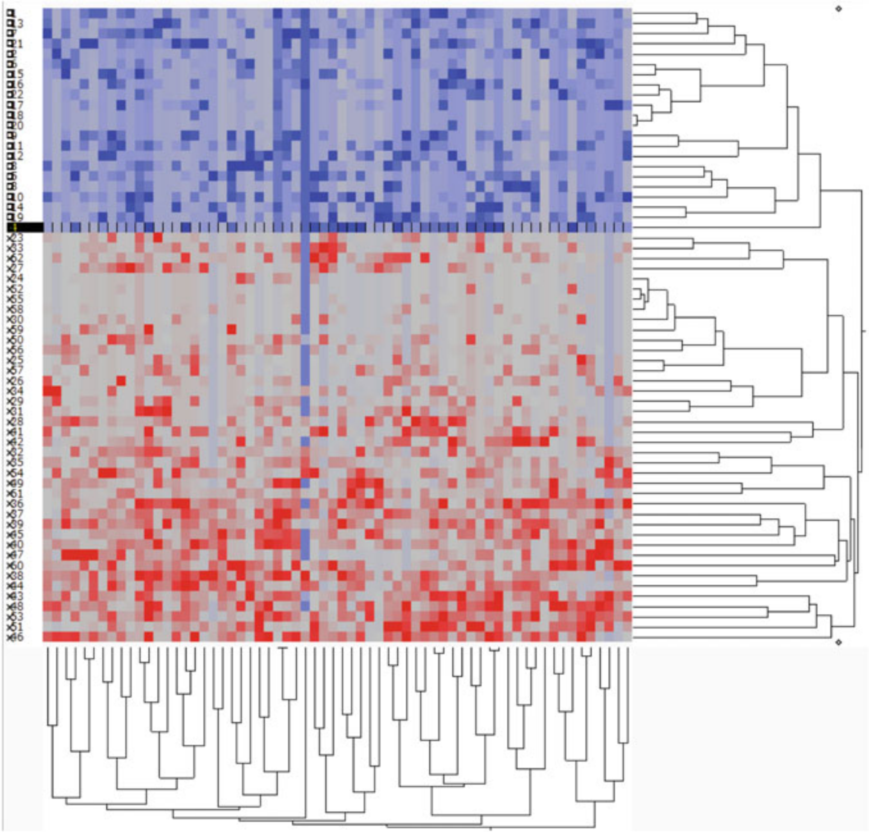


Fig. 2.8 Heat map and dendrogram of two classes

2.4.3 PCA Results of New Data

2.4.3.1 PCA Using Correlation Matrix

Figure 2.9 is three plots of PCA. Left plot is an eigenvalue. The first eigenvalue of the Prin1 is 39.36 and enormous because two classes are entirely separable on the Prin1. Its cumulative contribution ratio is 61.508%. The following three reasons may cause this result:

(1) Two classes are almost on Prin1. This fact means 64 RipDSs have almost the same axes in the high-dimensional microarray.

(2) Because two classes are entirely separable on Prin1, the first eigenvalue is enormous. Because the eigenvalues of the other 63 principal components are smaller than Prin1, the discrimination axes of 64 RipDSs are almost the same.

(3) The dispersions of the other 63 principal components are 38.5%. Thus, in the space of the high-dimensional microarray, we assume that the 64 axes produced by 64 RipDSs are in almost the same direction, the variation of 64 RipDSs is small, and the noise includes it. We confirm this claim in the later chapter.

The 22 normal cases overlap with the almost negative segment of the Prin1 axis. The 40 cancer cases scatter on the first and fourth quadrants. Right factor loading plot explains the correlation of Prin1 and Prin2 with 64 RipDSs. The correlation of the Prin1 with 64 RipDSs are almost from 0.5 to 1, and the correlation of Prin2 with 64 RipDSs is almost from −0.4 to 0.5. These plots imply us the normal subjects have a small variance on other 63 principal components, and severe cancer subjects have a slightly large variance on other 63 principal components. Severe cancer subjects locate a wide range compared with healthy subjects. These characteristic meets our common knowledge about cancer. Because two classes of Singh's microarray are healthy and cancer subjects in Fig. 2.23, its scatter plot is almost the same. On the other hand, because other microarrays are two different types of cancers, both classes do not locate on the segment of the Prin1 such as a healthy class in Fig. 2.9. Thus, we conclude the 63 variances of the healthy subjects are very small compared with cancer subjects. Factor loading plot locates on the first and fourth quadrants. If we obtain the validation cases, we can confirm the scatter plot is useful for cancer gene diagnosis as same as 64 individual RipDS. Thus, Prin1 indicates the cancer malignancy indicator in addition to 64 RipDSs.

Alon and Singh consist of two classes of cancer and healthy subjects. Even though the four microarrays except Alon and Singh are two different types of cancer, we obtained almost the same results as Alon and Singh. From this, we think that our analysis results are generally as follows. Because six eigenvalues of the Prin1 in Sect. 2.6 are large, this fact shows two classes of six microarrays are entirely separable

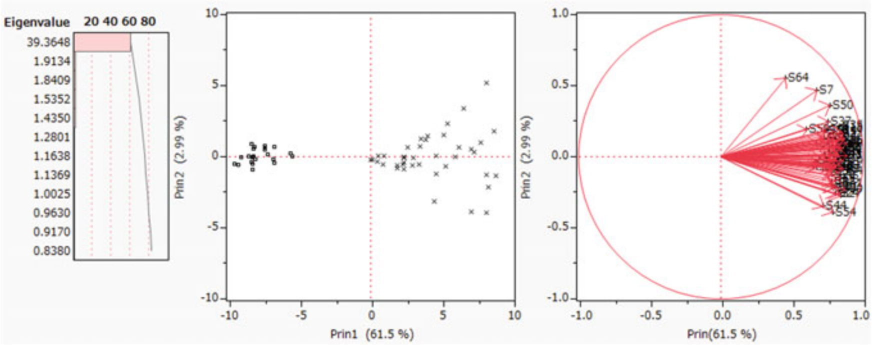


Fig. 2.9 Three plots of PCA

in high-dimensional microarray gene space. It is critical that our results may be almost the same by other microarrays and other types of gene data. We think another microarrays show the same results.

Figure 2.10 is 63 scatter plots of new data. The x-axis is Prin1. Y-axes are from Prin2 to Prin5 and from Prin61 to Prin64. Because other 55 scatter plots are almost the same, we omit those plots. Although PCA using genes of SMs cannot separate two classes, it can easily separate two classes by RipDSs new data. The 63 scatter plots of new data show two classes are separable in 64-dimensional space. Most statistical users misunderstand PCA can grasp the relevant information by Prin1 and Prin2 which represent large data variations. However, many statistical methods as same as PCA cannot find the linear separable fact of the signal having small variations.

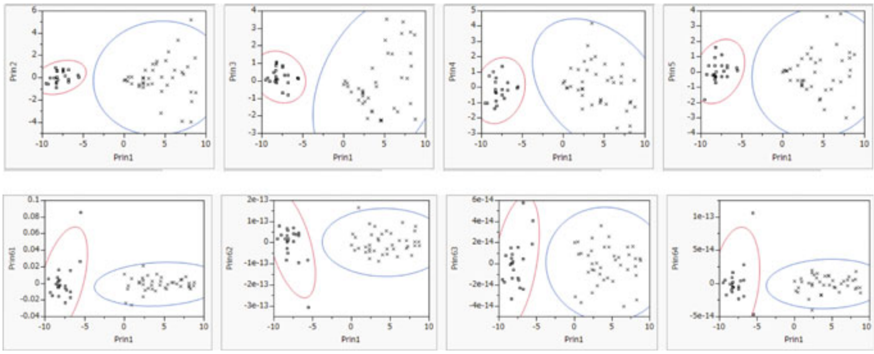


Fig. 2.10 Sixty-three scatter plots of new data (x-axis: Prin1, y-axis: from Prin2 to Prin5 and from Prin61 to Prin64)

Table 2.4 is the values of four principal components corresponding to Fig. 2.10. "ID" is the sequential number from 1 to 62. "Prin1" is the values of Prin1 sorted in ascending order. We consider "Prin1" is the cancer malignancy indicators. Both ranges of the 22 cases in class1 and 40 cases in class2 are [−9.57, −5.55] and [0.02, 8.77], respectively. There is 5.57 (=0.02 + 5.55) window width between two classes. RatioSV of PCA is 30.4% (=(0.02 + 5.55) * 100/(8.77 + 9.57) = 557/18.34 = 30.4). Because RatioSV of RipDS8 is 26.76%, RatioSV of PCA is 3.64% greater than RatioSV of RipDS8 because RatioSV of PCA is the total characteristic value of 64 RIPs. If the doctor confirms that the order of magnitude of the DS and the general severity of the subject are almost the same, Prin1 is available as the total characteristic value of the cancer malignancy indicator in addition to 64 RipDSs.

Table 2.4 Ranking of four principal components

ID	Prin1	R1	Prin2	R2	Prin3	R3	Prin4	R4
4	**−9.57**	**1**	−0.52	19	0.02	34	−1.04	12
10	**−9.33**	**2**	−0.62	12	0.15	40	−1.05	11
21	**−9.12**	**3**	−0.07	28	−0.05	31	0.72	49
9	−8.58	4	−0.02	32	0.46	45	0.99	51
19	−8.43	5	0.87	53	−0.05	32	−1.39	6
7	−8.37	6	−0.42	21	0.57	46	−0.29	23
5	−8.34	7	−0.92	8	0.23	42	−0.14	29
11	−8.33	8	−0.55	16	0.11	37	−0.28	24
1	−8.32	9	−0.03	31	0.85	50	0.34	41
13	−8.29	10	−0.01	33	1.06	53	0.58	46
3	−8.27	11	0.54	45	0.99	52	−0.37	22
12	−8.23	12	0.65	47	−0.13	28	−0.92	15
14	−8.15	13	−0.21	25	0.08	35	−1.22	8
16	−7.48	14	0.48	43	−0.72	17	−0.51	20
2	−7.47	15	0.61	46	0.76	48	1.34	55
8	−7.46	16	0.75	52	0.87	51	−0.17	27
15	−7.28	17	0.14	40	0.29	43	−0.73	19
6	−6.84	18	−0.18	27	0.31	44	0.30	39
22	−6.81	19	0.66	48	−0.12	29	−0.19	26
17	**−6.80**	**20**	−0.43	20	−0.81	16	−0.19	25
20	**−5.64**	**21**	0.22	41	0.13	38	−0.05	32
18	**−5.55**	**22**	0.01	34	0.10	36	0.03	35
55	**0.02**	**23**	−0.23	24	−0.20	26	0.29	38
52	**0.14**	**24**	−0.19	26	0.00	33	0.43	43
58	**0.43**	**25**	−0.31	23	−0.14	27	−0.02	33
24	0.52	26	0.08	39	−0.32	24	−0.01	34
30	0.90	27	−0.54	17	−0.63	21	1.27	54
59	0.92	28	0.04	37	−0.63	20	0.06	36
23	1.86	29	−0.62	13	−1.08	13	1.20	53
25	1.86	30	−0.77	10	−0.44	23	−0.14	28
57	2.29	31	−0.52	18	−0.68	19	0.39	42
34	2.32	32	0.02	36	−0.54	22	−1.40	5
29	2.35	33	−0.39	22	−1.27	11	−0.50	21
26	2.36	34	−0.83	9	−0.85	15	−0.11	30
56	2.37	35	−0.04	30	−0.72	18	−0.07	31
50	2.59	36	0.69	50	−1.17	12	1.72	58

(continued)

Table 2.4 (continued)

ID	Prin1	R1	Prin2	R2	Prin3	R3	Prin4	R4
31	2.96	37	−0.59	14	−2.13	5	1.13	52
33	3.01	38	−0.07	29	−0.96	14	2.75	61
62	3.40	39	−0.55	15	−1.70	8	1.85	59
27	3.50	40	0.71	51	−1.75	7	4.19	62
35	3.53	41	1.25	56	0.72	47	−1.14	9
61	3.90	42	1.21	55	−0.08	30	2.04	60
32	4.03	43	1.48	57	−0.31	25	0.86	50
42	4.46	44	−3.15	3	−2.28	1	−0.99	14
28	4.57	45	0.04	38	−2.22	3	−0.87	17
41	4.62	46	−1.19	7	−2.26	2	−0.78	18
49	5.11	47	1.56	58	2.34	57	1.52	56
47	5.25	48	−0.67	11	3.53	62	−1.07	10
54	5.43	49	2.34	60	1.59	55	0.47	44
45	5.91	50	0.01	35	1.54	54	0.63	47
37	6.17	51	0.67	49	−1.55	9	0.31	40
44	6.56	52	3.39	61	−1.76	6	−2.63	2
40	7.05	53	0.51	44	2.73	59	0.72	48
51	7.10	54	−3.90	2	0.21	41	0.28	37
60	7.24	55	0.30	42	3.39	61	1.55	57
39	7.74	56	1.01	54	0.14	39	−1.53	4
38	8.13	57	5.19	62	−1.29	10	−2.52	3
43	8.14	58	−1.26	6	1.59	56	−2.94	1
48	8.16	59	−3.96	1	2.78	60	−1.01	13
53	**8.29**	**60**	−2.12	4	0.83	49	−0.91	16
36	**8.69**	**61**	1.80	59	2.62	58	0.47	45
46	**8.77**	**62**	−1.35	5	−2.17	4	−1.29	7

2.4.3.2 Comparison of Correlation Matrix and Variance–Covariance Matrix

(1) Analysis of New Data by Correlation Matrix

In general, PCA uses a correlation matrix. Correlation matrix makes it possible to avoid the influence of variables having different units. Because all the variables are the same units of the gene expression level, it is meaningful to analyze microarray by the variance-covariance matrix. Attempts with the variance–covariance matrix, the results were surprisingly different. Examination of the result is future research.

(2) Analysis of New Data by Variance–Covariance Matrix

If we analyze new data made by the variance–covariance matrix, we obtain the different result in Fig. 2.11. The first eigenvalue is 612.23, and its cumulative contribution ratio is 51.848%. The second eigenvalue is 208.002, and its cumulative contribution ratio is 69.4%. Scatter plot has two tendencies. Many subjects are located on a line about 40° relative to Prin1. Over 13 cancer subjects widely scatter under this line at an angle of −45° with Prin1. Factor loading plot explains this meaning.

RipDS64 and RipDS7 are in the fourth quadrant. These two RipDSs are different from other 62 RipDSs. Especially, RipDS64 may relate over 13 patients. This example shows the merit of our approach because general knowledge of statistics can interpret unique cases.

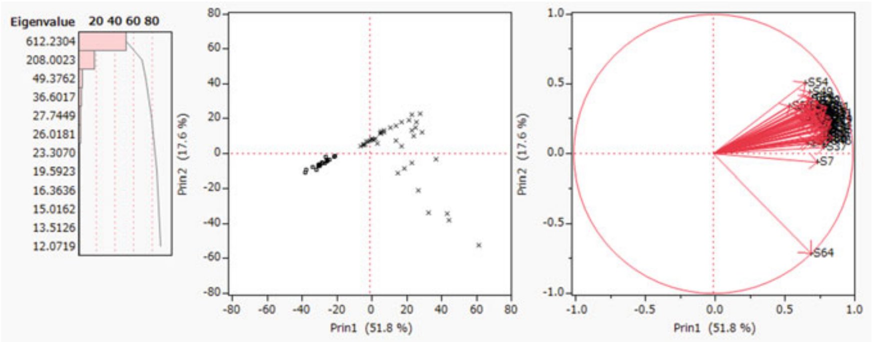


Fig. 2.11 Three plots of PCA (variance–covariance matrix)

(3) The 62 RipDSs new data using the correlation matrix

After we drop two variables such as RipDS64 and RipDS7, we analyze the 62 RipDSs new data with 62 subjects and 62 RipDSs. Figure 2.12 is the three plots using the correlation matrix. The Prin1 eigenvalue is 38.719, and its cumulative contribution ratio is 62.45%. Scatter plot and factor loading plot are similar to Fig. 2.9. However, RipDS63 is different from other 61 RipDSs those are two groups in the first and fourth quadrants. These two groups have different meanings in Prin2.

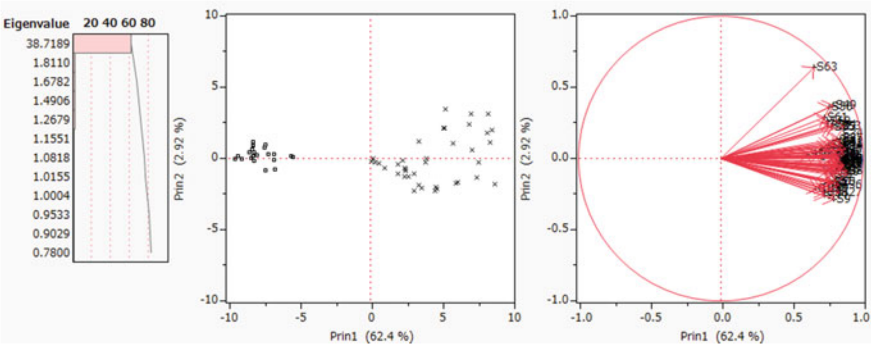


Fig. 2.12 Three plots of PCA without two RipDSs (correlation matrix)

(4) The 62 RipDSs New Data using the Variance–Covariance Matrix

Figure 2.13 shows three plots of PCA using variance–covariance matrix instead of the correlation matrix in Fig. 2.12. The Prin1 eigenvalue is 500.687, and its cumulative contribution ratio is 58.515%. Although two eigenvalues of Prin1 and Prin2 in Fig. 2.11 are 612.2 and 208, those of Prin1 and Prin2 in Fig. 2.13 are 500.7 and 49.4. We guess to remove RIP64 reduce the variance of Prin2 as Fig. 2.13.

If we remove RIP63 in next, RIP13 or RIP44 may become outliers. In future research, the surveys of PCA using correlation matrix and variance–covariance matrix may suggest more useful knowledge about cancer gene diagnosis.

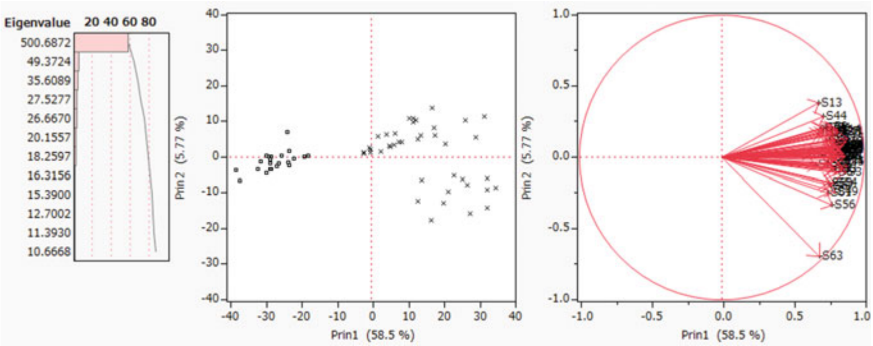


Fig. 2.13 Three plots of PCA without two RipDSs (variance–covariance matrix)

2.4.3.3 Analysis of Transpose Data

We transpose the new data with 62 subjects and 64 RipDSs and analyze the transpose data that consists of 64 RipDSs (64 cases) and 62 subjects (62 variables). Figure 2.14 is three plots of PCA. Eleven eigenvalues are over one. Scatter plot shows most RipDSs locate on around the origin. The 13th, 14th, and 28th RipDSs are an outlier in the fourth quadrant. The 64th DS (64) is an outlier in the first quadrant. These four RIPs may become different malignancy indicators from other 60 RipDSs. Factor loading plot shows most of the healthy subjects in class1 having variable name prefix "N" located in the second quadrant. The tumor cases variable name prefix "C" locate in the fourth and first quadrants.

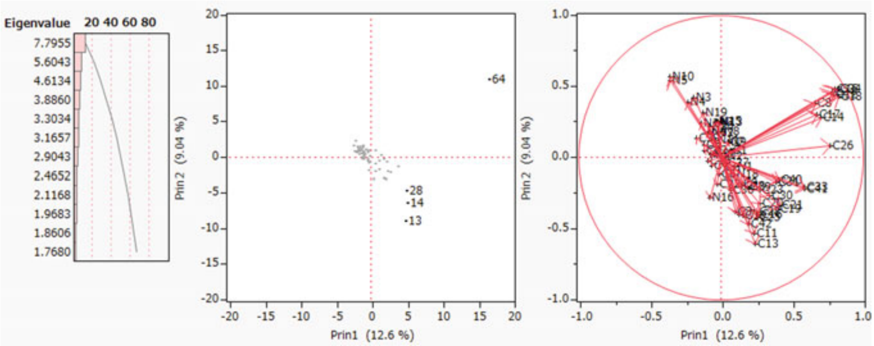


Fig. 2.14 Three plots of PCA

Figure 2.15 is four scatter plots of PCA. The x-axis is the Prin1. Y-axes are Prin2, Prin3, Prin4, and Prin5. We find several outliers. Both outliers of new data and transposed new data may be the new subclass of cancers pointed out by Golub et al. This is the future theme.

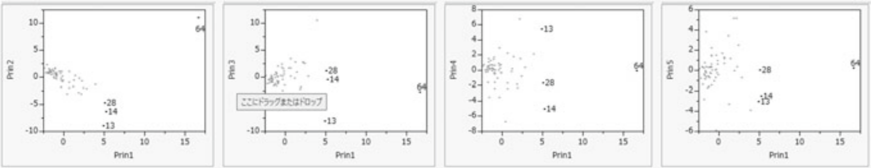


Fig. 2.15 Three scatter plots of transposed data

2.4.3.4 Summary

At first, we claimed standard statistical methods could analyze SM very easily because these subspaces were small samples with small n and small p (Shinmura 2016). However, our examination shows it is difficult for us to obtain good results using the 64 SMs. However, we can get surprising results from the RipDSs new data. Primarily, the 62 subjects are ranking on the Prin1. Thus, we can rank the malignancy of cancer by the value of Prin1 in Fig. 2.9. Moreover, the Ward cluster analysis can identify two clusters entirely in Fig. 2.8. Usually, cluster analysis cannot cluster two classes. However, we can separate two classes completely by RipDS new data. We need cooperation with an expert on gene diagnosis. Especially, we expect seven research members of Alon et al. They can validate our results and confirm our claim. Alon et al. can prove their research was right. If they offer new validation samples, we analyze those samples and feedback the results to them.

2.5 The 130 BGSs of Alon's Microarray

2.5.1 Results by Standard Statistical Methods

After LINGO Program3 found 64 SMs in 2016, LINGO Program4 separates 130 BGSs and noise subspace with five genes in 2016. In this section, we analyze 130 BGSs by standard statistical methods. We claimed standard statistical methods could analyze BGS very easily because each BGS was a small sample. However, we cannot obtain useful results of 130 BGS showed in this section. Next, we analyze 130 RipDSs new data with 62 cases and 130 RipDSs (130 variables) by standard statistical methods and get the surprising success in Sect. 2.5.3. However, RatioSVs show BGSs are not helpful for cancer gene diagnosis because 130 RatioSVs using 130 BGSs are less than 1%.

2.5.1.1 Validation of 130 BGSs by Three Statistical Discriminant Functions

Table 2.5 is 130 BGSs from SN = 1 to SN = 130. "BGS" column shows 130 BGSs sorted by descending order of RatioSV values. "Gene" column is the number of genes in each BGS. Because all NMs of logistic regression are zero, we can confirm 130 BGSs are linearly separable. The 60 NMs of QDF are zero. "LDF2 and LDF1" are NMs of two different prior probability options of Fisher's LDFs. The prior probability of LDF2 is proportional to the case number of 22:40. The prior probability of LDF1 is "1:1" that is default in much statistical software. However, we use the proportional prior probability because we wish to compare NMs of six MP-based LDFs. We omit LDF1 from Table 2.5. The RatioSV of RIP using BGS in Table 2.5 recommends

RipDS128 because it is the maximum value of 130 BGSs. We claim RatioSV is the best index for the LSD-discrimination of two classes. Last three rows are the maximum, mean, minimum values of 130 BGSs. The range of gene is [9, 25]. The ranges of QDF, LDF2 and LDF1 are [0, 7], [3, 14] and [0, 13], respectively. Because 60 NMs of QDF are zero and maximum NM is almost half of NMs of LDF2, QDF is better than LDF2. Because the range of 130 RatioSVs is [0.00, 0.9], 130 BGSs are not helpful for cancer gene diagnosis. However, in the case of Swiss banknote data, two-variable (X4, X6) is BGS, and its RatioSV is 0.524% in Fig. 10.2. Thus, 0.9% is not such a bad value for BGS. On the other hand, RatioSV of SM is abnormally large beyond our knowledge.

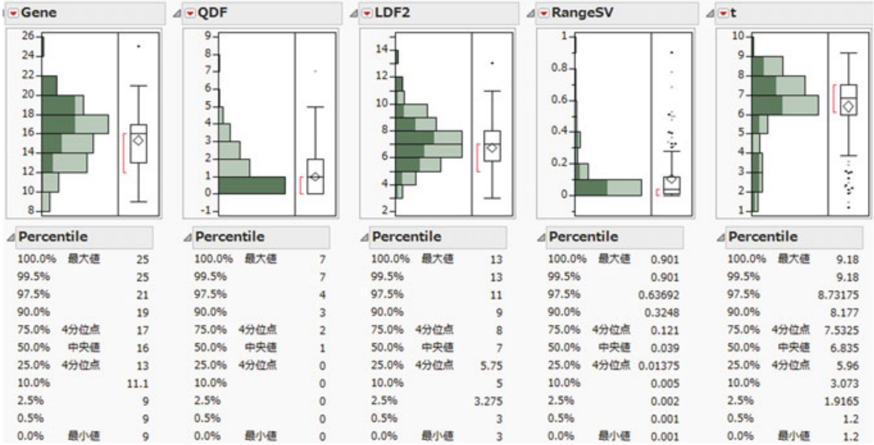
Future Research (Problem7): We must survey and compare the relation of BGSs and SMs.

Table 2.5 Hundred and thirty NMs of BGSs by three discriminant functions and RatioSV

SN	BGS	Gene	Logistic	QDF	LDF2	DS	RatioSV
128	**BGS128**	**11**	**0**	**0**	**7**	**222**	**0.90**
93	BGS93	12	0	3	7	258	0.77
56	BGS56	9	0	2	7	295	0.68
129	BGS129	12	0	3	5	380	0.53
–	–	–	–	–	–	–	–
1	BGS1	20	0	**0**	9	395	0.51
23	BGS23	14	0	2	6	84,975	0.00
32	BGS32	15	0	**0**	10	82,982	0.00
127	BGS127	14	0	2	14	113,129	0.00
64	BGS64	12	0	4	8	300,470	0.00
83	**BGS83**	**18**	**0**	**0**	**9**	**176,727**	**0.00**
	MAX	25	0	7	14	300,470	0.90
	MEAN	15.35	0	1.14	7.36	16,093	0.11
	MIN	9	0	0	3	222	0.00
	SUM	1995	0	148	957	2,092,140	13.93

2.5.1.2 Histogram and Correlation

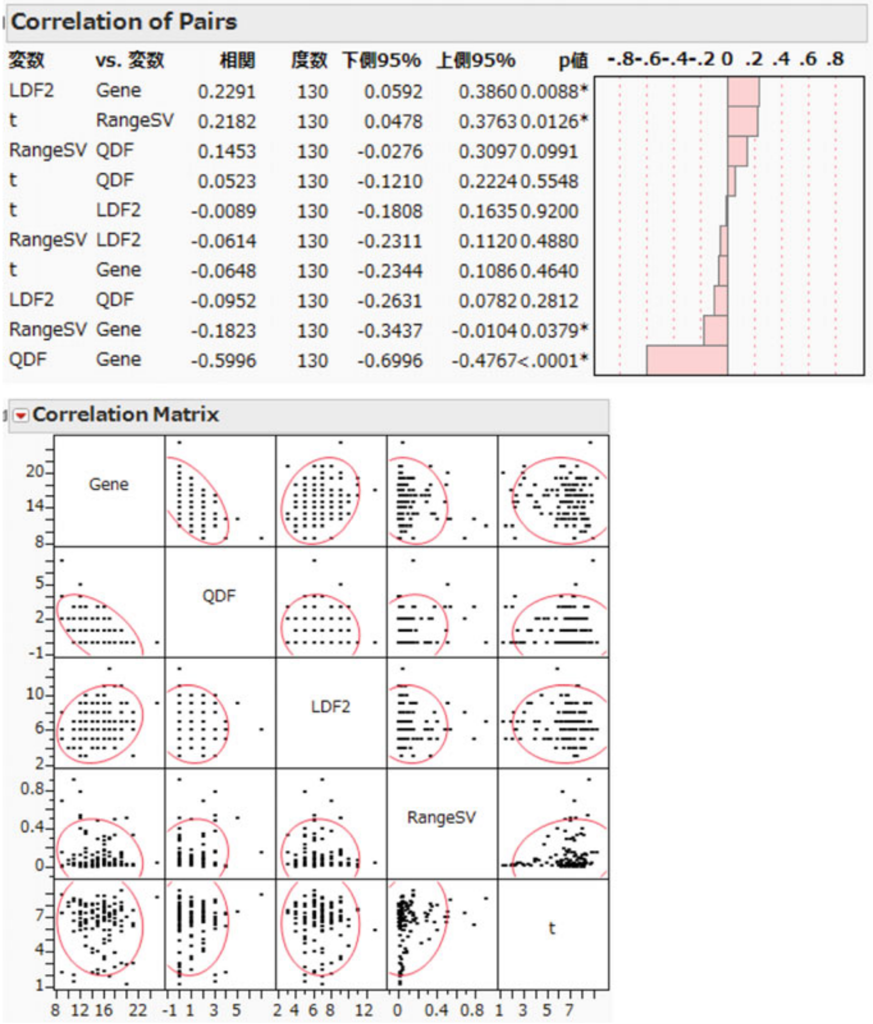
Figure 2.16 is the histograms of gene, QDF, LDF2, RangeSV, and t-value. If we select the cases with “NM of QDF = 0,” those cases become dark green. Dark green cases of the other four variables have wide ranges.



Percentile		Gene	QDF	LDF2	RangeSV	t
100.0%	最大値	25	7	13	0.901	9.18
99.5%		25	7	13	0.901	9.18
97.5%		21	4	11	0.63692	8.73175
90.0%		19	3	9	0.3248	8.177
75.0%	4分位点	17	2	8	0.121	7.5325
50.0%	中央値	16	1	7	0.039	6.835
25.0%	4分位点	13	0	5.75	0.01375	5.96
10.0%		11.1	0	5	0.005	3.073
2.5%		9	0	3.275	0.002	1.9165
0.5%		9	0	3	0.001	1.2
0.0%	最小値	9	0	3	0.001	1.2

Fig. 2.16 Histograms of gene, QDF, LDF2, RatioSV and t-value

Figure 2.17 is the matrix correlation of five variables. Two correlations of (gene, LDF2) and (RangeSV, t) are positive correlations such as 0.229 and 0.218. These are weak positive correlations compared with SMs. Two correlations of (gene, RangeSV) and (gene, QDF) are negative correlations such as −0.182 and −0.599. Generally speaking, as the number of genes increases, NM of QDF becomes smaller, so it is not very useful information. That is, this analysis is useless.



Correlation of Pairs

変数	vs. 変数	相関	度数	下側95%	上側95%	p値
LDF2	Gene	0.2291	130	0.0592	0.3860	0.0088*
t	RangeSV	0.2182	130	0.0478	0.3763	0.0126*
RangeSV	QDF	0.1453	130	-0.0276	0.3097	0.0991
t	QDF	0.0523	130	-0.1210	0.2224	0.5548
t	LDF2	-0.0089	130	-0.1808	0.1635	0.9200
RangeSV	LDF2	-0.0614	130	-0.2311	0.1120	0.4880
t	Gene	-0.0648	130	-0.2344	0.1086	0.4640
LDF2	QDF	-0.0952	130	-0.2631	0.0782	0.2812
RangeSV	Gene	-0.1823	130	-0.3437	-0.0104	0.0379*
QDF	Gene	-0.5996	130	-0.6996	-0.4767	<.0001*

Fig. 2.17 Matrix correlation of five variables (BGS)

2.5.1.3 Analysis of Eleven Genes of BGS128 by Standard Statistical Methods

We analyze eleven genes included in BGS128 by standard statistical methods because the RatioSV of RipDS128 has the maximum value among 130 BGSs. Table 2.6 is the result of one-way ANOVA of BGS128 that consists of 11 genes. Two columns "Min and Max" are the range of class1 (normal, 22 cases). "MIN and MAX" are the range of class2 (tumor, 40 cases). "t ($\neq$)" is the t-test value under two variances which are not equal, and "t (=)" is the t-test value under two variances which are equal. The most important fact is that there are seven positive values and four negative values. Moreover, these absolute values are not so large. This fact indicates it is entirely wrong that genes with significant positive values are oncogenes.

Table 2.6 Result by one-way ANOVA of BGS128

Gene	Min	Max	MIN	MAX	t (≠)	t (=)
B128X1963	5.18	8.48	3.21	10.53	2.49	2.23
X1964	3.91	6.92	3.21	7.69	−1.13	−1.09
X1965	4.30	6.57	3.76	6.94	3.46	3.29
X1966	4.85	7.14	3.50	7.77	−1.00	−0.89
X1967	7.28	10.95	3.50	10.52	−4.79	−4.17
X1968	4.39	7.13	5.20	7.57	0.35	0.37
X1969	3.76	9.36	4.66	9.18	1.73	1.72
X1970	5.17	7.23	3.21	7.62	−2.45	−2.08
X1971	4.35	8.94	6.27	9.29	1.68	1.89
X1976	4.24	10.43	4.44	9.16	0.23	0.24
X1978	4.61	7.70	3.50	8.08	1.28	1.17

Figure 2.18 is three plots of PCA. Left plot is an eigenvalue. Four principal components from Prin1 to Prin4 are larger than one. The central plot is a scatter plot. Two classes overlap. The right plot is the factor loading plot. The 11 genes are located in three quadrants except for the third quadrant. Because two classes overlap, these results are not valuable. The presence or absence of the linear separable fact is vital for the cancer gene diagnosis. This indicator will be an excellent guide to mitigating sizeable genetic analysis work.

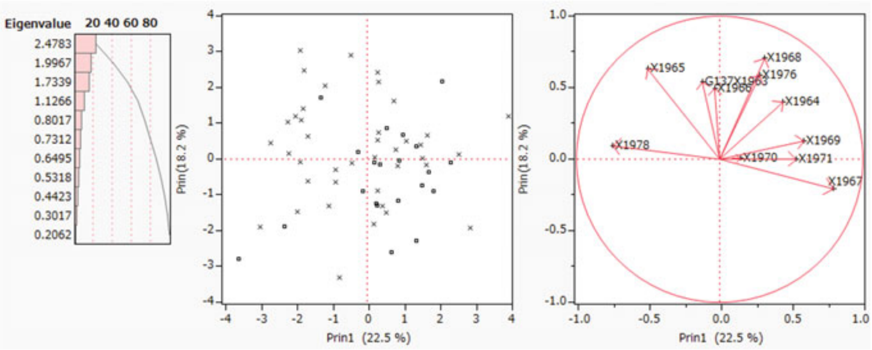


Fig. 2.18 PCA plots (eigenvalue, scatter plot, and factor loading)

2.5.2 *Examination of RipDSs of 130 BGSs*

2.5.2.1 Validation of RipDSs and RatioSV of 130 BGSs

Table 2.7 is the summary of 130 BGSs. "BGS" columns correspond "SN" columns in Table 2.5. The table is sorted in descending order by RatioSV value. "BGS = 128" is RipDS of BGS128. "Min and Max" columns are the range of class1. The range of 22 normal subjects with BGS128 is [−74, −1]. "MIN and MAX" columns are the range of class2. The range of 40 tumor subjects of BGS128 is [1, 148]. Thus, RipDSs range of BGS128 has 222 widths in "RipDS" column. Because maximum RatioSV of BGS128 is 0.901% (=200/222), BGS may be useless for cancer gene diagnosis. However, because BGS explains the structure of Alon's microarray by the monotonic decrease of MNM, it is essential for the study of the gene's role in cancer. Furthermore, RipDS of BGS may be a valid signal instead of RipDS of SM (Problem7).

Table 2.7 Summary of RipDSs and RatioSV of 130 BGSs

BGS	Min	Max	MIN	MAX	RipDS	RatioSV
128	−74	−1	1	148	222	**0.901**
93	−69	−1	1	189	258	0.774
57	−84	−1	1	211	295	0.679
129	−167	−1	1	213	380	0.526
1	−149	−1	1	246	395	0.507
–	–	–	–	–	–	–
23	−46,017	−1	1	38,958	84,975	0.002
32	−35,120	−1	1	47,862	82,982	0.002
127	−46,881	−1	1	66,247	113,129	0.002
64	−118,569	−1	1	181,901	300,470	**0.001**
83	−72,581	−1	1	104,146	176,727	0.001
Max	−69	−1	1	**181,901**	**300,470**	**0.901**
Mean	−5979	−1	1	9947	15,926	0.107
Min	**−118,569**	−1	1	148	**222**	0.001
Range	5910	0	0	171,954	284,544	0.794

We misunderstood BGS was more critical than SM because BGS could explain the Matryoshka structure of microarrays entirely. However, we think BGS may not discriminate validation samples correctly. This fact recollects us two facts.

1. Two-variable (X4, X6) is a unique BGS of Swiss banknote data (Flury and Riedwyl 1988). However, the best model is five variable (X1, X3, X4, X5, X6) that has the minimum M2 among 63 models. We explain this fact in Table 6.5

of the Book1 (Shinmura, 2016). Because the M2 of (X4, X6) is larger than (X1, X3–X6), the two-variable model is not the best model.

2. If we remove one gene from BGS, the removed gene subspace is not LSD and is not used for cancer gene analysis anymore. For these reasons, we change the research theme and analyze the SM instead of BGS in the seven chapters after this chapter. We think that the four genes of Yamanaka's iPS cells are BGS. However, it is different in that there is no feature corresponding to the monotonous decrease of MNM.

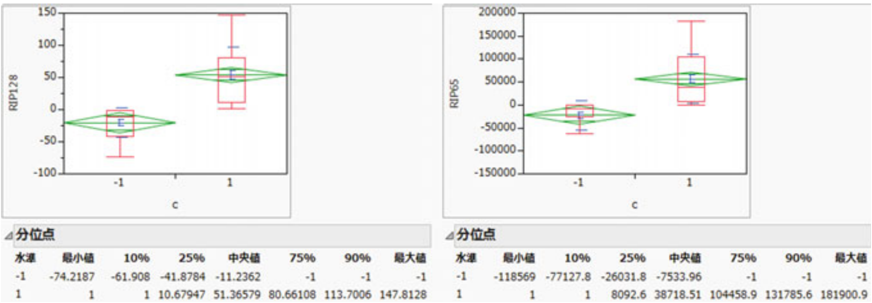
Table 2.7 is the result of 130 RipDSs using 130 BGSs. "Min and Max" columns show the range of 22 normal cases ($y_i = -1$), and "MIN and MAX" columns show the range of 40 tumor cancer cases ($y_i = 1$). Because all 130 pairs of "Max and MIN" are -1 and 1, this fact tells us 130 BGSs are linearly separable gene subspaces. "RipDS" is the range of DS of RIP such as [Min, MAX]. "RatioSV of BGS" is the value calculated by 200/DS that is the ratio (%) of the SV distance and DS. The RatioSV of BGS128 is 0.901% and maximum value. Although the distance of SV is two, it is 0.901% of DS. The RatioSV of BGS83 is 0.001% and minimum value. These ratios tell us that the degree of linear separability is very tiny. Thus, it is hard for us to find linear separable fact by standard statistical methods for 130 BGSs. Last three rows are a summary of 130 BGSs. The range of RipDSs is [222, 300470] that is abnormally large. Because the range of RatioSV is [0.001, 0.901], 130 RIPs scarcely discriminate two classes in 130 BGSs.

2.5.2.2 Box–Whisker Plots of BGS128 and BGS65

Figure 2.19 is two box–whisker plots of BGS128 with the maximum RatioSV and BGS65 with the minimum RatioSV. Because SV separates both two classes completely, many researchers usually willingly accept these results. Until now, there is no research on LSD-discrimination except for us. MNM is critical statistics, and RatioSV is the second crucial statistics in addition to MNM. Three SVMs and three OLDFs discriminate the microarrays by SV that divide the data space into three subspaces such as $y_i * f(\mathbf{x}_i) >= 1$, $1 > y_i * f(\mathbf{x}_i) > -1$ and $-1 >= y_i * f(\mathbf{x}_i)$. Vapnik defined LSD as follows:

(1) There are no cases in "$1 > y_i * f(\mathbf{x}_i) > -1$."
(2) Two classes can be assigned to either one of "$y_i * f(\mathbf{x}_i) >= 1$" or "$-1 >= y_i * f(\mathbf{x}_i)$."

In this book, we assign the class1 in "$-1 >= y_i * f(\mathbf{x}_i)$" and the class2 in "$y_i * f(\mathbf{x}_i) >= 1$." MNM is a statistic introduced by Shinmura. This statistic explains the Matryoshka structure of the microarrays that is the same idea of "MNM monotonic decrease." RatioSVs of BGS128 and BGS65 are 0.901% (=2/222 * 100) and 0.001% (=2/176727 * 100). The distances of SV are 0.901% and 0.001% for the width of RipDSs. These values are too small. On the other hand, all RatioSVs of SMs are larger than RatioSV of BGS. Although BGS can explain Matryoshka structure of the microarrays completely, we judge those are useless for cancer gene diagnosis. Yamanaka's four genes are a kind of BGS. His team made a new stem cell to become gospel to human beings, iPS cell. We consider stem cell formation as the same phenomenon as LSD. His team replaced c-myc with L-myc to prevent cancer. The c-myc and L-myc may be compatible, and our research foresaw that there are other Yamanaka's genes that pair with L-myc.



分位点

水準	最小値	10%	25%	中央値	75%	90%	最大値
-1	-74.2187	-61.908	-41.8784	-11.2362	-1	-1	-1
1	1	1	10.67947	51.36579	80.66108	113.7006	147.8128

分位点

水準	最小値	10%	25%	中央値	75%	90%	最大値
-1	-118569	-77127.8	-26031.8	-7533.96	-1	-1	-1
1	1	1	8092.6	38718.51	104458.9	131785.6	181900.9

Fig. 2.19 Two box–whisker plots of BGS128 and BGS65

2.5.3 *Examination of RipDSs New Data by PCA and Cluster Analysis*

2.5.3.1 PCA

We examine 130 RipDSs new data by PCA. The data consist of 62 cases and 130 variables. Figure 2.20 is three plots of PCA. Left plot is an eigenvalue. The first eigenvalue is 71.322, and contribution is 54%. The central plot is a scatter plot. The range of 22 normal cases with □ on Prin1 is [−13.642, −5.007], and the range of 40 cancer cases with × is [−3.133, 13.005]. In many analyses, all healthy cases have negative scores, and all cancer patients have positive scores. How to evaluate cancer patients having negative scores as in this example is a future study. In particular, it is important whether or not medical meaning can be found. Thus, RatioSV is 7.294%. Because all subjects are almost nearby Prin1 axis, we can rank the malignancy indicator by the value of the Prin1. Because maximum RatioSV of 130 RipDSs is 0.901%, RatioSV on Prin1 is eight times wider than those of 130 RipDSs. The right plot is the factor loading plot. All RipDSs are in the first and fourth quadrants. The 130 correlations of Prin1 with RipDSs are from 0.620 to 0.863, and the 130 correlations of Prin2 with RipDSs are from −0.402 to 0.394.

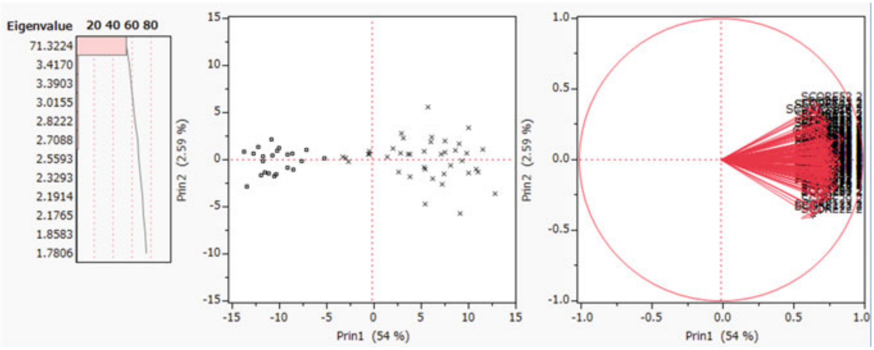


Fig. 2.20 Three plots of PCA

Figure 2.21 is four scatter plots. The x-axis is the Prin1, and y-axes are the Prin2, Prin3, Prin4, and Prin5 from the left plot to right plot. A small 95% confidence ellipse in the left is the normal class, and the right large 95% confidence ellipse is the cancer class. The negative DS corresponds to a healthy subject group or a mild cancer group, which is a feature shared by six microarrays. There are outliers, and two classes are separable visually.

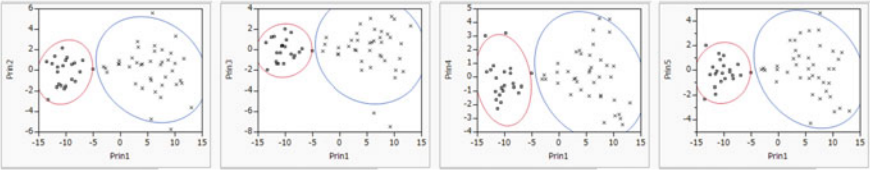


Fig. 2.21 Four scatters plots

2.5.3.2 Cluster Analysis

We analyze the new data by Ward cluster analysis. Figure 2.22 is the heat map and case dendrogram of 62 cases. We categorize two clusters. The upper cluster includes 22 normal cases marked by □. The lower cluster includes 40 cancer cases marked by ×. Although many researchers approached the gene analysis using cluster analysis, they could not obtain clear results such as this figure because the cluster analysis cannot cluster the microarrays. However, by creating signal data, we can easily capture the linearly separable facts in cluster analysis and PCA.

The bottom dendrogram is the variable dendrogram of 130 RipDSs. More than ten clusters can be classified. About seven distances of pairs are very close. We think that these seven pairs are redundant and interchangeable. That is, the orders of the discrimination scores of 62 cases are considered to be almost the same.

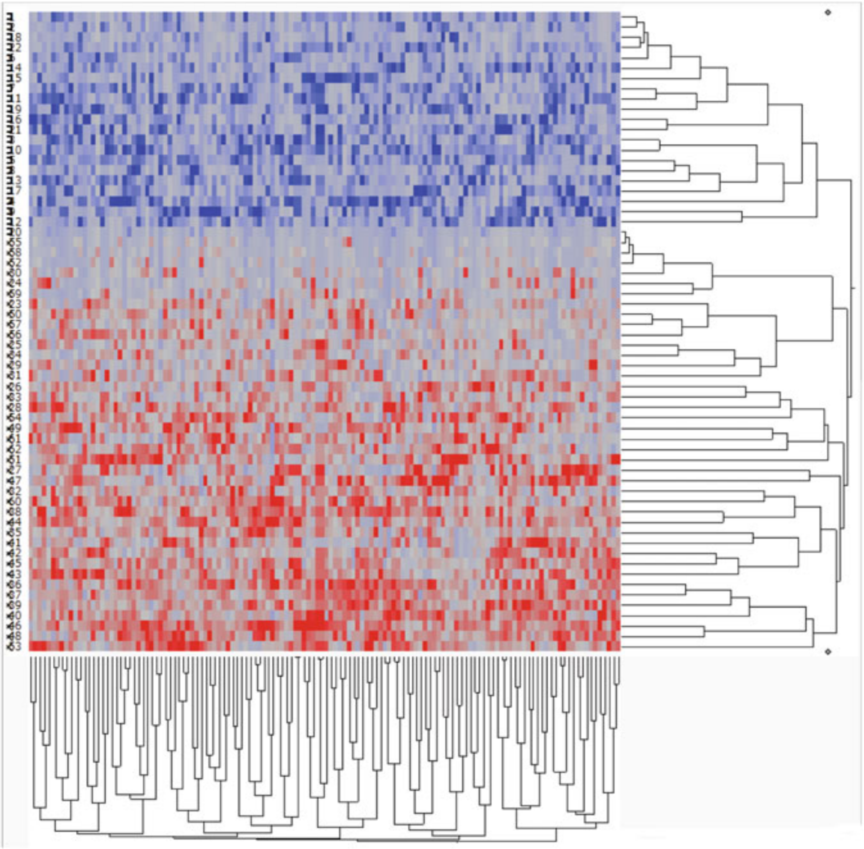


Fig. 2.22 Heat map and case dendrogram of 62 cases by Ward method. (upper: normal 22 cases; lower: cancer 40 cases)

2.5.4 *Summary*

We claimed standard statistical methods could analyze BGS very easily because these subspaces were small samples as same as SM. However, our examination shows it is difficult for us to obtain good results from BGS. On the other hand, we can get useful results from the RipDSs new data as same as SM. Notably, the 64 subjects are ranking on the Prin1. Thus, we can rank the cancer malignancy indicator by the value of Prin1 in Table 2.7. Moreover, the Ward cluster analysis can identify two clusters entirely.

2.6 Other Five Microarrays

We introduce cancer gene diagnosis of other five microarrays. Because all results are almost the same as 64 SMs of Alon’s microarray, we focus on the results of malignancy indicators and outliers of transposed data.

2.6.1 Singh’s Microarray

2.6.1.1 Outlook

We analyze Singh’s microarray that consists of two classes, such as 50 normal subjects (class1) and 52 tumor prostate subjects (class2) with 12,625 genes. LINGO Program3 finds 179 SMs that include 1,238 genes. Because other 11,387 gene subspace is not linearly separable, we omit this subspace from our analysis. In other words, we choose only the signal by removing noise. Ratio of signal (RatioS) is 9.01% (=1,238/12,625 * 100%). The portion of noise is 91%. Because all researchers analyze microarrays with noise and do not understand the Matryoshka structure, they could not find significant facts from 1970. At first, we analyze 179 SMs by standard statistical methods and cannot obtain useful results. Because 179 NMs of logistic regression and 26 NMs of QDF are zero, only logistic regression can find that all SMs are linearly separable. Other standard statistical methods cannot locate linear separable fact. Because we are not the gene specialists, we cannot conclude the useful meaning of these results without linear separable fact. However, if we analyze new data with 102 subjects and 179 RipDSs, we find surprising results for cancer gene analysis. Ward cluster analysis and PCA can separate two classes correctly. The range of 179 RatioSVs is [0.28, 11.67%]. If we suppose RatioSV with over 5% is useful for cancer gene analysis, 38 RIPs among 179 SMs become different cancer malignancy indicators. Moreover, both ranges of the 50 subjects in class1 and 52 subjects in class2 on the Prin1 are [–17.89, –4.81] and [0.99, 22.53], respectively. There is 5.8 (=4.81 + 0.99) window width between two classes. Thus, RatioSV of PCA is 14.35% (=(4.81 + 0.99) * 100/(17.89 + 22.53) = 580/40.42). Because RatioSV of RIP2 is 11.67%, RatioSV of PCA is 2.68% greater than RatioSV of RIP2 because RatioSV of PCA is the total characteristic value of 179 RIPs. If cancer gene specialists validate and confirm our results, we can open a new frontier of cancer gene diagnosis by 38 RIPs and Prin1 malignancy indicators. These results will be helpful for cancer gene diagnosis. We expect Singh et al. researchers validate our results and confirm our claim to open a new frontier of cancer gene diagnosis.

At first, we plan to obtain all BGSs and analyze all BGSs. However, because the range of 130 BGSs RatioSVs of Alon et al. microarray is [0.001, 0.9%], we judge RIPs of BGSs are not helpful for cancer gene diagnosis.

2.6.1.2 Malignancy Indexes

Figure 2.23 is three plots of PCA. The first eigenvalue is 113.749 and contribution ratio is 63.5%. Scatter plot shows two classes are entirely separable. The 50 cases in class1 locate on negative first principal axis (Prin1). The 52 cases in class2 scatter on the first and fourth quadrants. Factor loading plot locates on the first and fourth quadrants.

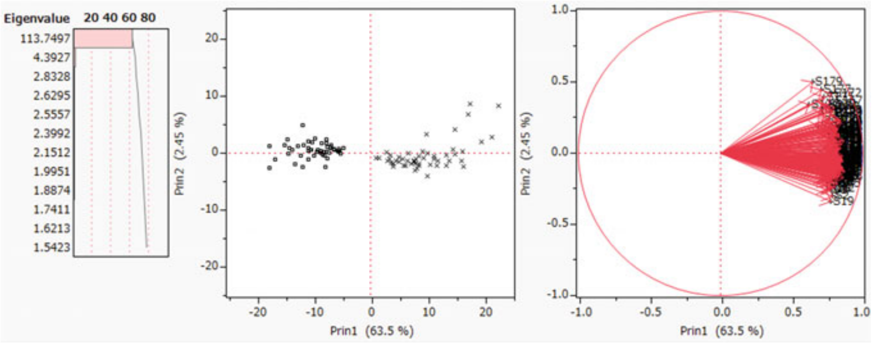


Fig. 2.23 Three plots of PCA

Figure 2.24 is three scatter plots. The x-axis is Prin1. Y-axes are Prin2, Prin3, and Prin4 from left, central, and right plots, respectively. The two ellipses are 99% confidence ellipses. A small one in the left is class1, and right large one is class2. The 42nd, 54th, 57th, and 100th cases may be outliers. Cancer subjects scatter to a great area.

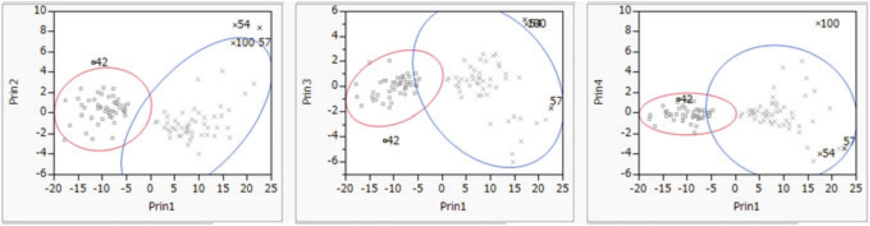


Fig. 2.24 Three scatter plots

2.6.1.3 Analysis of Transpose New data

We transpose the new data with 102 subjects and 179 RipDSs and analyze the transposed new data. Figure 2.25 is three plots of PCA. Factor loading plot shows 50 normal subjects in class1 locate in the second and third quadrants, and 52 cancer cases (variables) locate in the first and fourth quadrants. The variable name with the suffix "N" shows the normal subjects in class1, and other variable names with "C" show the cancer subjects in class2. Scatter plot shows most RipDSs are on the line of 45° with Prin1. The 179th, 178th, 175th, and other several RipDSs are outliers those may be different types of malignancy indicators.

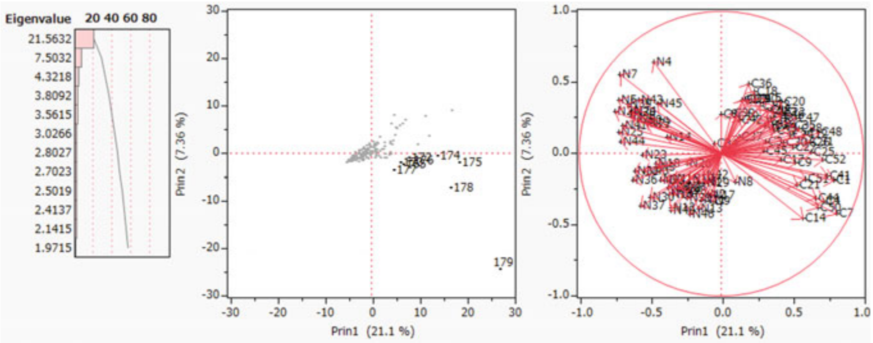


Fig. 2.25 Three plots of PCA

Figure 2.26 is three scatter plots of PCA. The x-axis is the Prin1. Y-axes are Prin2, Prin3, and Prin4. There are more outliers than other microarrays. Especially, there is one large cluster of outliers. This consideration is the future work. If Singh et al. validate our results and confirm our claim, they can prove their research was right.

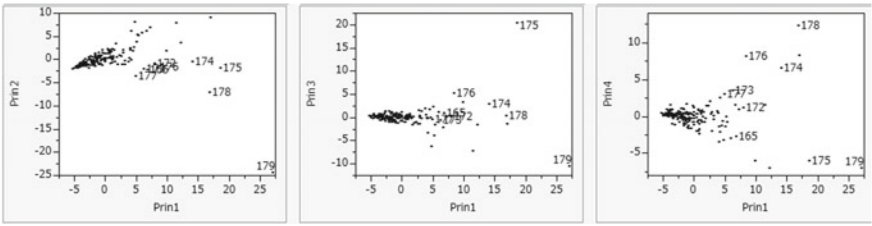


Fig. 2.26 Three scatter plots of PCA

2.6.2 Golub Microarray

2.6.2.1 Outlook

We analyze the Golub microarray that consists of two classes, such as 25 acute myeloid leukemia (AML, class1) and 47 acute lymphoblastic leukemia (ALL, class2) with 7,129 genes. LINGO Program3 finds 69 SMs that include 1,238 genes. Because other 5,891 gene subspaces are not linearly separable and noise, we omit these 5,891 genes from our analysis. In other words, we choose the only signal. The ratio of signal (RatioS) is 17.36% (=1,238/7,129 * 100), and the portion of noise is 82.64%. At first, we analyze 69 SMs by standard statistical methods. Because 69 NMs of logistic regression are zero, only logistic regression can find that 64 SMs are linearly separable. The 16 NMs of QDF and one NM of both Fisher's LDFs are zero. Other standard statistical methods cannot show linear separable fact. Because we are not the specialists of the cancer gene, we cannot conclude the useful meaning of these results without linear separable fact. However, if we analyze 69 RipDSs new data, we find surprising results. Ward cluster analysis and PCA can separate two classes correctly. The range of 69 RatioSVs of RIP is [0.004, 15.69%]. Ninth RIP (RIP9) of SM9 has the maximum value of 15.69%. Thus, RIP9 can discriminate new validation samples very easy. If we supposed RatioSV with over 5% is useful for cancer gene diagnosis, 28 RIPs become different cancer malignancy indicators. Moreover, both ranges of the 25 cases in class1 and 47 cases in class2 on the Prin1 are [−11.72, −4.66] and [−1.66, 23.16], respectively. There is 3 (=4.66 − 1.66) window width between two classes. RatioSV of PCA is 8.6% (=3 * 100/(11.72 + 23.16) = 300/34.88 = 30.4). Because RatioSV of RIP9 is 15.69%, RatioSV of Prin1 is 7.69% less than RatioSV of RIP9. This result is different from Alon and Singh results that are two-class discriminations between cancer subjects versus normal subjects. Although RatioSV on Prin1 is the total characteristic value of 69RIPs, Prin1 does not reflect the merit of PCA. By analyzing 69 RipDSs, we find surprising results for cancer gene diagnosis. If medical experts will validate our results and confirm our claim, we can open a new frontier of cancer gene diagnosis.

2.6.2.2 Malignancy Indexes

Figure 2.27 is three plots of PCA. Left eigenvalue shows the eigenvalue of Prin1 is 45.02 and the contribution ratio is 65.246%. Because two classes are two different types of cancer, both classes do not locate on Prin1 as same as Alon and Singh's microarrays. The first eigenvalue is very large. Scatter plot shows two classes are completely separable. The 25 AML cases locate on negative Prin1. The 47 ALL cases scatter on the first and fourth quadrants. Factor loading plot scatters on the first and fourth quadrants. However, some RipDSs have small correlations with Prin1 and Prin2.

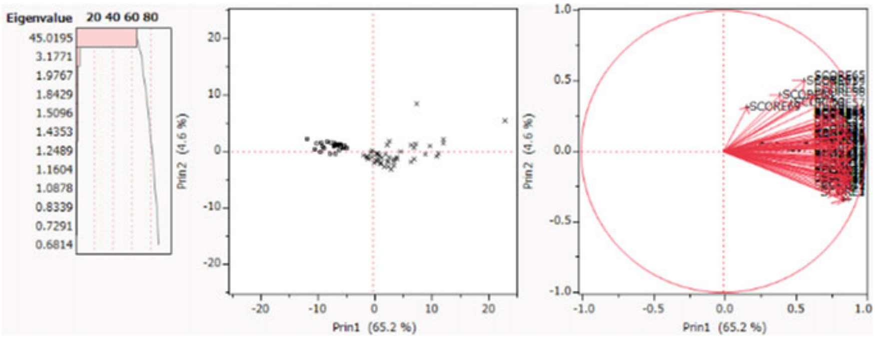


Fig. 2.27 Three plots of PCA

Figure 2.28 is three scatter plots. Left small 95% ellipse is AML. Because the right 95% ellipse is larger than AML ellipse, we guess ALL is more variance than AML and may be severe cancer. The 30th, 62nd, and 65th cases in class2 are outliers. We expect some member of Golub et al. explain the reason why these ALL subjects are outliers.

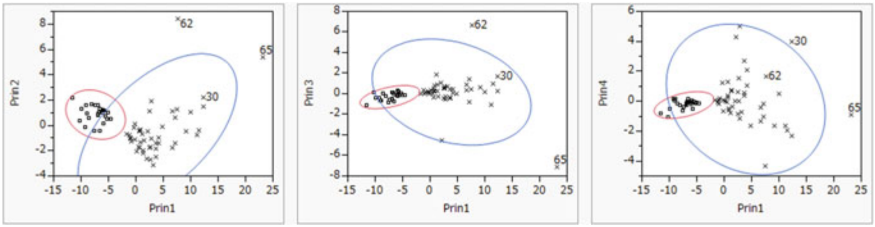


Fig. 2.28 Three scatter plots (x-axis: Prin1, y-axis: Prin2, Prin3, Prin4)

2.6.2.3 Analysis of Transpose New data

Figure 2.29 is three plots of PCA using transpose new data with 69 RipDSs and 72 subjects. Factor loading plot shows 25 AML subjects with prefix N located in the second and third quadrant. However, this class may be categorized into three groups. Those are two high negative correlations with Prin1 or Prin2, and low negative correlations with Prin1 and Prin2. The 47 ALL subjects with prefix C located in the first and fourth quadrants. However, this class may be categorized into three groups. Those are two high positive correlations with Prin1 or Prin2, and low positive correlations with Prin1 and Prin2. Scatter plot shows most RipDSs are nearby the origin. Two large outliers discriminate two ALL groups having two different types of high positive correlations.

Furthermore, although PCA result of Fig. 2.27 is almost the same as the other five results, Fig. 2.29 is quite different from others. The reason why the results of the transposed matrix greatly differ is the future research topic.

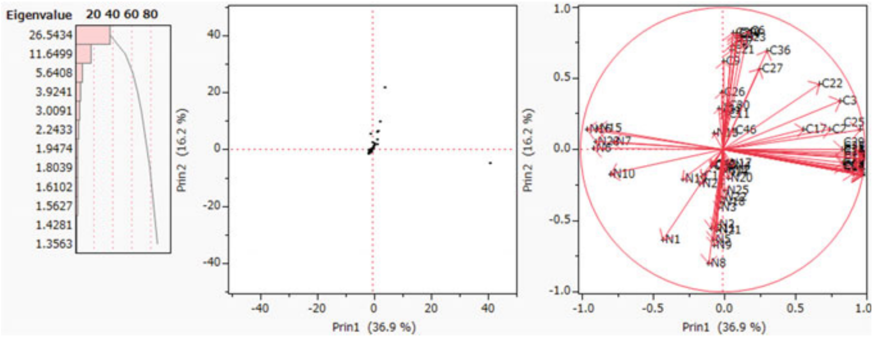


Fig. 2.29 Three plots of PCA using transpose new data

Figure 2.30 is three plots of PCA. The x-axis is Prin1. Y-axes are the Prin2, Prin3, and Prin4. The RIP30, RIP62, RIP65, and other RipDSs may be outliers. If Golub et al. validate our results and confirm our claim, they can prove their research was right.

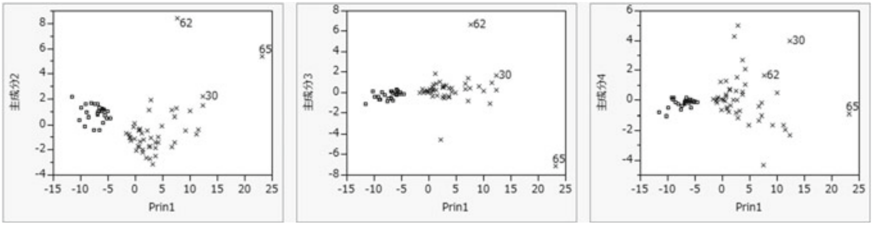


Fig. 2.30 Three plots of PCA

2.6.3 *Tian's Microarray*

2.6.3.1 Outlook

Tian's microarray consists of two classes, such as 36 cases (false, class1) and 137 cases (true, class2) with 12,625 genes. LINGO Program3 finds 159 SMs that include 7,221 genes and other 5,404 genes are noise. We omit these noises from our analysis. The ratio of signal (RatioS) is 57.2%, and the portion of noise is 42.8%. At first, we analyze 159 SMs by standard statistical methods and cannot obtain success. Because 159 NMs of logistic regression are zero, 159 SMs are linearly separable.

Because 158 NMs of QDF are zero, two classes may be fairly separated. Other standard statistical methods cannot show linear separable fact. Because we are not the specialists of the cancer gene, we cannot conclude the useful meaning of these results without linear separable fact. However, if we analyze 159 RipDSs new data, we find surprising results for cancer gene analysis. Ward cluster analysis and PCA can separate two classes correctly. The range of RatioSVs is [0.63, 19.13%]. The 21st RIP (RIP21) has the maximum value of 19.13%. Thus, RIP21 can discriminate new validation samples very easily and may indicate a cancer malignancy indicator for cancer gene diagnosis. If we supposed RatioSV with over 5% is useful for cancer gene analysis, 27 RIPs become different cancer malignancy indicators. Moreover, both ranges of the 36 cases in class1 and 137 cases in class2 on the Prin1 are [−17.68, −11.94] and [−4.09, 15.07], respectively. There is 7.85 (=11.94 – 4.09) window width between two classes. RatioSV of PCA is 24% (=7.85 * 100/(17.68 + 15.07) = 785/32.75). Because RatioSV of RIP21 is 19.13%, RatioSV of PCA is 4.87% greater than RatioSV of RIP21. If medical doctors validate and confirm our results, we can open a new frontier of cancer gene diagnosis by 63 RIPs and Prin1.

2.6.3.2 Malignancy Indexes

Figure 2.31 is three plots of PCA. The eigenvalue of Prin1 is 62.73 and larger than other eigenvalues. Its contribution ratio is 39.339%. Scatter plot shows two classes are completely separable. Factor loading plot locates on the first and fourth quadrants. However, 159 correlations of Prin1 with 159 RipDSs are smaller than other microarrays. We cannot explain this reason now.

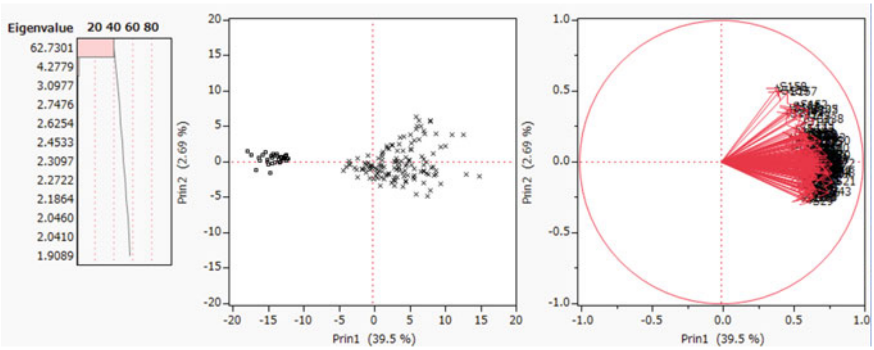


Fig. 2.31 Three plots of PCA

2.6.3.3 Analysis of Transpose New Data

We transpose the new data and analyze it that consists of 159 RipDSs and 173 subjects. Figure 2.32 is three plots of PCA. Factor loading plot shows FALSE class1 cases locate in the second and third quadrants, and TRUE class2 cases locate in the first and fourth quadrants. Scatter plot indicates that it fluctuates largely in a trumpet or fan shape as it goes from left to right. Approximately ten RIPs widely spread from the first quadrant to fourth quadrants like the top of a fan. These outliers are thought to represent different information with other RIP and cancer diagnosis.

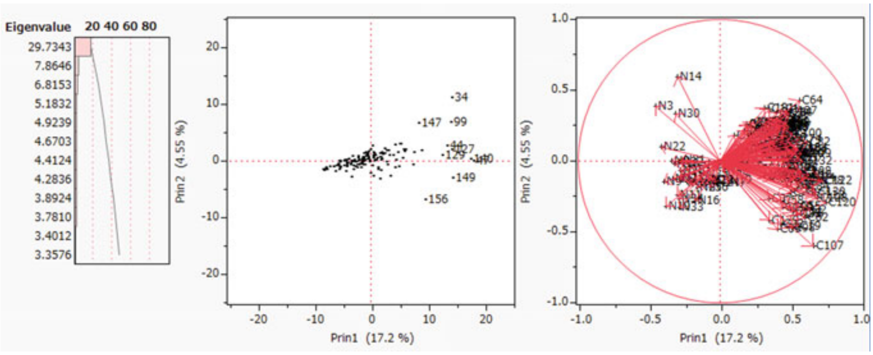


Fig. 2.32 Three Plots of PCA

Figure 2.33 is three scatter plots of PCA. The x-axis is Prin1. Y-axes are Prin2, Prin3, and Prin4. We indicate about ten outliers. This consideration is the future work. If Tian et al. validate our results and confirm our claim, they can prove their research was right.

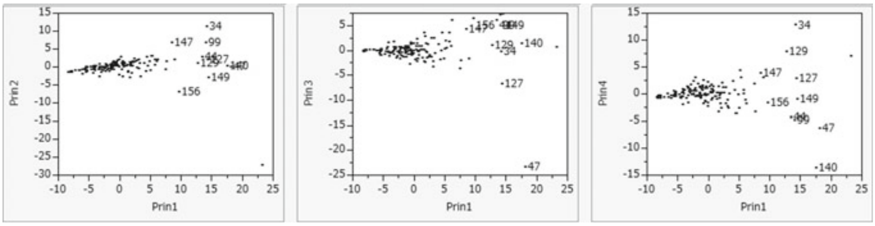


Fig. 2.33 Three scatter plots of transposed new data

2.6.4 Chiaretti Microarray

2.6.4.1 Outlook

Chiaretti microarray consists of two classes, such as 95 subjects (B-cell, class1) and 33 subjects (T-cell, class2) with 7,129 genes. LINGO Program3 finds 95 SMs with 5,163 genes, and other 1,956 genes are noise. We omit these genes from our analysis.

Ratio of signal is 72.36% (=5,160/7,129 * 100%). At first, we analyze 95 SMs by standard statistical methods and cannot obtain success. Because 95 NMs of logistic regression and QDF are zero, logistic regression and QDF confirms all SMs are linearly separable. The 92 NMs of LDF2 and 94 NMs of LDF1 are zero. Thus, two classes are the most separated among six microarrays. Because we are not the cancer gene specialists, we cannot find the useful meaning of these results without linear separable fact. However, if we analyze 95 RipDSs new data, we find surprising results for cancer gene analysis. Ward cluster analysis and PCA can separate two classes correctly. The range of RatioSVs is [10.73, 38.93%]. The 23rd RIP (RIP23) of SM23 has the maximum value of 38.93%. RIP23 can discriminate new validation samples very easily and may indicate the cancer malignancy indicator for cancer gene diagnosis. If we supposed RatioSV with over 5% is useful for cancer gene analysis, 95 RIPs become different cancer malignancy indicators. Moreover, both ranges of the 95 cases in class1 and 33 cases in class2 on the Prin1 are [−11.4, −1.71] and [12.73, 16.66], respectively. There is 14.44 (=12.73 + 1.71) window width between two classes. RatioSV of PCA is 51.46% (=(12.73 + 1.71) * 100/(16.66 + 11.4) = 14.44/28.06). Because RatioSV of RIP23 is 38.98%, RatioSV of PCA is 12.48% greater than RatioSV of RIP23. If medical gene specialists validate and confirm our results, we can open a new frontier of cancer gene diagnosis by 95 RIPs and Prin1.

2.6.4.2 Malignancy Indexes

Figure 2.34 is three plots of PCA. Three eigenvalues are greater than one. The eigenvalue of Prin1 is 72.243, and contribution ratio is 76.046%. Scatter plot shows two classes are completely separable. The 95 subjects in class1 locate on negative first principal axis. The 33 subjects in class2 scatter on the positive first axis. Factor loading plot locates on the first and fourth quadrants. The 95 correlations of Prin1 with 95 RipDSs are approximately over 0.8.

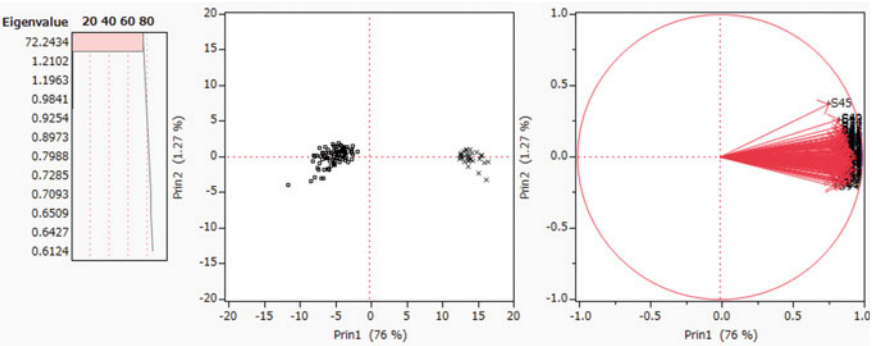


Fig. 2.34 Three plots of PCA

Figure 2.35 is three scatter plots. The x-axis is Prin1. Y-axes are Prin2, Prin3, and Prin4, respectively. Left ones are B-cell classes. Right ones are T-cell classes.

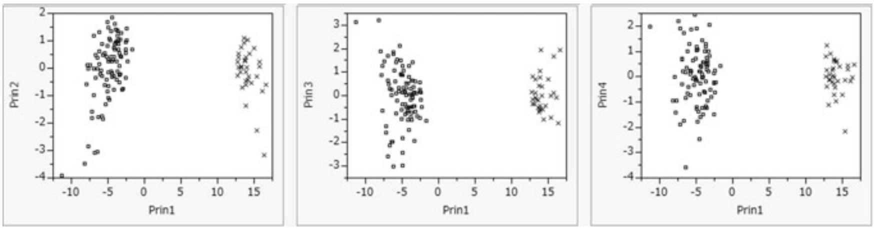


Fig. 2.35 Three scatter plots

2.6.4.3 Analysis of Transpose New Data

We transpose the new data and analyze it that consists of 95 RipDSs and 128 subjects. Figure 2.36 is three plots of PCA. Factor loading plot shows T-cell in class2 locates in the second and third quadrants and B-cell in class1 locates in the first and fourth quadrants. Scatter plot shows most RipDSs are nearby the origin. The 3rd, 4th, 41st, 46th, and 95th RipDSs are outliers. These five RIPs are considered to show different diagnostic results from the other 90 RIPs.

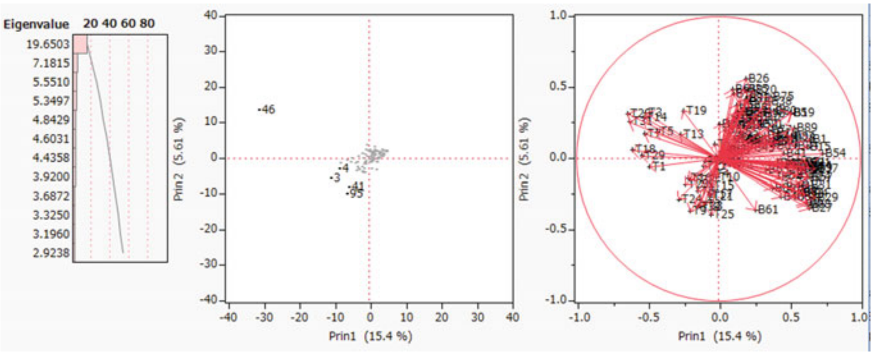


Fig. 2.36 Three plots of PCA

Figure 2.37 is three scatter plots of PCA. The x-axis is Prin1. Y-axes are Prin2, Prin3, and Prin4. We indicate the outliers. This consideration is the future work. The 3rd, 4th, 41th, 46th, and 95th RipDSs may be outliers. Thus, these outliers indicate the different roles from other RIP. If Chieretti et al. validate our results and confirm our claim, they can prove their research was right.

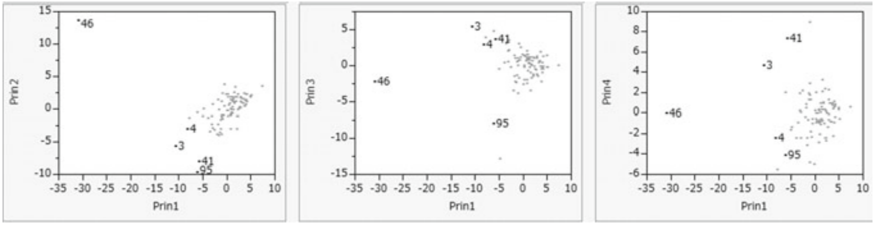


Fig. 2.37 Three scatter plots of PCA

2.6.5 *Shipp Microarray*

2.6.5.1 Outlook

Shipp microarray consists of two classes, such as 19 cases (follicular lymphoma, class1) and 58 cases (DLBCL, class2) with 7,129 genes. LINGO Program3 finds 239 SMs with 4,716 genes. Because it is hard work to analyze 239 SMs manually, we focus on 130 SMs with 3,827 genes in this book. In other words, we choose only signal by removing noise and 109 SMs. Because all researchers analyze microarray with noise, they could not find the significant facts from 1970. At first, we analyze 130 SMs by standard statistical methods and cannot obtain success. Because 130 NMs of logistic regression are zero, only logistic regression can find 130 SMs are linearly separable. QDF, LDF2, and LDF1 can discriminate 121, 46, and 53 SMs correctly. Other standard statistical methods cannot show linear separable fact. Because we are not the cancer gene specialists, we cannot find the useful meaning of these results without linear separable fact. However, if we analyze 130 RipDSs, we find surprising results for cancer gene analysis. Ward cluster analysis and PCA can separate two classes correctly. The range of 130 RatioSVs is [4.99, 30.67%]. The 11th RIP (RIP11) of SM11 discriminates two classes by SV completely. Although SV distance is two, the ratio of this distance is 30.67% of the RIP11 DS range. Thus, RIP11 can discriminate new validation samples very easily and may indicate the cancer malignancy indicator for cancer gene diagnosis. If we suppose RatioSV with over 5% is useful for cancer gene diagnosis, 129 RIPs become different cancer malignancy indicators. Moreover, both ranges of the class1 and class2 on the Prin1 are [−18.98, −12.56] and [−2.23, 13.62], respectively. There is 10.33 (=12.56 − 2.23) window width between the two classes. RatioSV of PCA is 31.69% (=10.33 * 100/(13.62 + 18.98) = 1033/32.6 = 31.69). Because RatioSV of RIP11 is 30.67%, RatioSV of PCA is 1.02% greater than RatioSV of RIP11 because RatioSV of PCA is the total characteristic value of 130 RIPs. If medical doctors validate and confirm our results, we can open a new frontier of cancer gene diagnosis by 129 RIPs and Prin1.

2.6.5.2 Malignancy Indexes

Figure 2.38 is three PCA plots using a correlation matrix. Left plot is an eigenvalue. Twelve principal components from Prin1 to Prin12 are greater than one. The cumulative contribution ratio is 60.82%. The central plot is a scatter plot. Two classes are separable. The 19 subjects in class1 distribute on negative Prin1. The 58 subjects in class2 distribute the first and fourth quadrants that separate three groups. The first group consists of ID = 14, 40, and 54 DLBCL subjects in the first quadrant. The second group distributes on positive Prin1. The third group consists of over 13 subjects such as ID = 4, 27, 32, 39, 44, 55–58, and other subjects in the fourth quadrants. The right figure is the factor loading plot. Moreover, both ranges of the 19 cases in class1 and 58 cases in class2 on Prin1 are [−18.98, −12.56] and [−2.23, 13.62], respectively. There is 10.33 (=12.56 − 2.23) window width between the two classes. RatioSV of PCA is 31.69% (=10.33 * 100/(13.62 + 18.98) = 1033/32.6 = 31.69). Because RatioSV of RIP11 is 30.67%, RatioSV of PCA is 1.02% greater than RatioSV of RIP11 because RatioSV of PCA is the total characteristic value of 130 RIPs. If medical doctors validate and confirm our results, we can open a new frontier of cancer gene diagnosis by 129 RIPs and Prin1.

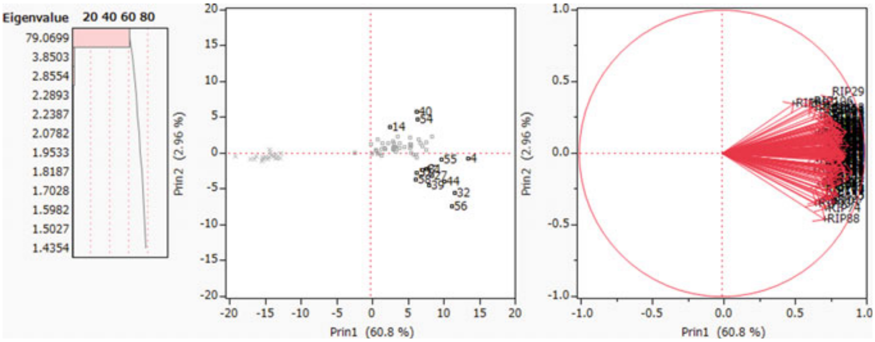


Fig. 2.38 PCA plots (eigenvalue, scatter plot, and factor loading)

Figure 2.39 is three scatter plots. The x-axis is Prin1. The y-axes are Prin2, Prin3, and Pin4. The left small cluster is 58 class1 cases. The right large cluster is 58 DLBCL class2 those include many outliers.

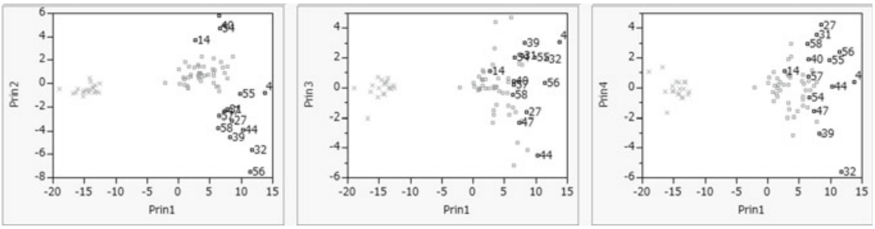


Fig. 2.39 Three scatter plots

2.6.5.3 Analysis of Transpose New Data

We transpose the new data and analyze the transpose new data that consists of 130 RipDSs and 77 subjects (77 variables). Figure 2.40 is three plots of PCA using a correlation matrix. Factor loading plot shows all patients locate only in the first and fourth quadrants. Scatter plot shows most RipDSs are nearby the origin. Outliers are located in the first and fourth quadrants.

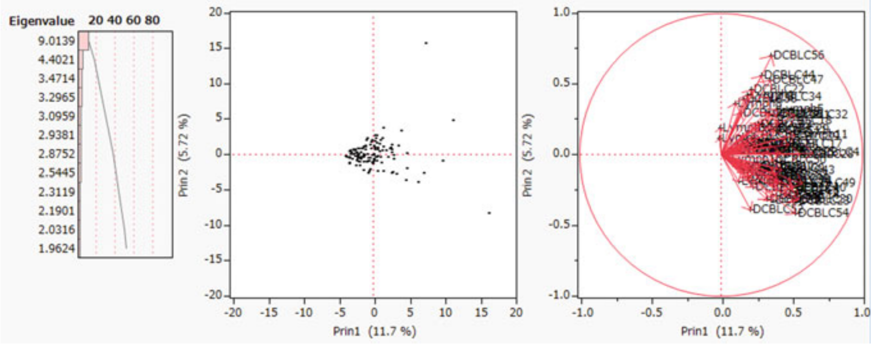


Fig. 2.40 Three plots of PCA

Figure 2.41 is three scatter plots of PCA using variance–covariance matrixes. The x-axis is the Prin1. Y-axes are Prin2, Prin3, and Prin4. We indicate the outliers. This consideration is the future work. There are many outliers those indicate the different roles from other RIPs. If Shipp et al. validate our results and confirm our claim, they can prove their research was right.

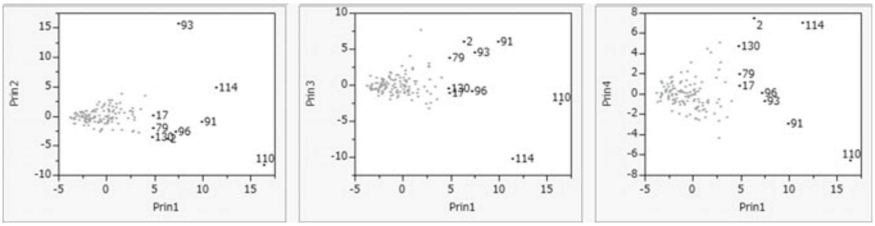


Fig. 2.41 Three scatter plots of PCA

2.7 Conclusion

After establishing theory and solving cancer gene analysis, our next research theme was to obtain all BGSs of microarrays at first. However, because LINGO Program4 to find all BGSs needed much computational time, we changed to find each BGS step by step. Because finding 130 BGSs of Alon's microarray took over one week, we compared 130 BGSs with 64 SMs of Alon's microarray. RatioSV of 63 SMs among 64 SMs is over 5%. On the other hand, 130 RatioSVs of 130 BGSs are less than 0.9%. Thus, we judged BGS is useless for cancer gene diagnosis. Moreover, we changed to analyze all SMs of microarrays. Even though two classes are completely separable in all SMs, all standard statistical methods except for logistic regression cannot show linear separable fact. Thus, we made new data having RipDSs as variables. Cluster analysis and PCA separate two classes completely. Especially, the Prin1 of PCA illustrates malignancy indicators very well. Next, the scatter plot of transposed new data shows many RipDSs become outliers. These outliers may be expected unknown subclass of cancer pointed out by Golub et al. We confirmed the other five microarrays are almost the same results. Table 2.8 is the summary of this chapter. RatioS is the ratio of (the number of genes included in all SMs/total genes). " >=5%" is the number of SMs, RatioSVs of those are over than 5%. The ratio of RatioSV over 5% are 98.4%, 18.8%, 99.2%, 21.2%, 16.9%, and 100%, respectively. Alon, Shipp, and Chiaretti microarrays are 98.4% over. RatioSV of PCA is slightly different about this trend. Last three columns are the number of linearly separable SMs by QDF, LDF1, and LDF2. These numbers indicate two classes are well separable in all SMs as same as the trend of " >=5%."

So far, we think the genes included in SM or BGS are oncogenes, and they are signals. However, we could not obtain the right results by standard statistical methods. That is, the discriminant score obtained by the genes included in SM and BGS may be as a signal. In 2017, we obtain two kinds of SMs from the RIP and Revised LP-OLDF. In Chaps. 4–9, we compared two results of new data and transposed new data made by RipDSs, LpDSs, and HsvmDSs. Furthermore, by comparing the signal subspace that is the union of all SMs, we explain the reason why the standard statistical methods cannot find the linear separable fact in SM.

Table 2.8 Summary of this chapter

	n * p	SM:gene	RatioS	RatioSV	>= 5%	RatioSV of PCA	QDF = 0	LDF1[a]	LDF2[b]
Alon	62 * 2,000	64:1999	99.950	[2.4, 26.8]	63 (98.4%)	30.40%	64	13	12
Golub	72 * 7,129	69:1,238	17.366	[0.004, 15.69]	13 (18.8%)	34.88%	16	1	1
Shipp	77 * 7,129	239:4,716	66.152						
		130:3,827	53.682	[4.99, 30.67]	129 (99.2%)	31.69%	121	53	46
Singh	102 * 12,626	179:1238	9.805	[0.28, 11.67]	38 (21.2%)	14.35%	26	0	0
Tian	173 * 12,625	159:7222	57.204	[0.63, 19.13]	27 (16.9%)	24%	158	1	0
Chiaretti	128 * 7,129	95:5,162	72.422	[10.73, 38.93]	95 (100%)	51.46%	95	94	92
	n * p	BGS:gene	RatioS	RatioSV	>= 5%	of PCA	QDF = 0	LDF1	LDF2
Alon	62 * 2,000	130:1995	99.75	[0.001, 0.901]	0	0.90%	0	0	0

[a] Prior probabilities are 0.5 versus 0.5

[b] Prior probabilities are proportional to the case number

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad Sci USA, 96(1.1): 6745–6750

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

Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood 103(7): 2771–2778

Flury B, Riedwyl H (1988) Multivariate statistics: a practical approach. Cambridge University Press, New York

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2016) New theory of discriminant analysis after R. Fisher. Springer

Shinmura S (2017) From cancer gene analysis to cancer gene diagnosis. Amazon

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1.1):68–74. (https://doi.org/10.1038/nm0102-6)

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Lada M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(1.1):203–209

Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD (2003) The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. N Engl J Med 349(26):2483–2494

Yata K, Aoshima M (2010) Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. J Multivar Anal 101(2010):2060–2077. https://doi.org/10.1016/j.jmva.2010.04.006

Chapter 3
Cancer Gene Diagnosis of Alon's microarray by RIP and Revised LP-OLDF

Abstract This chapter discusses the following three points. (1) We have introduced only SMs obtained with the RIP in Chap. 2. RIP analyzed SMs by Program3' arbitrary iteration number. In 2017, we increase the number of iterations successively from 1 and select the iteration number that the number of SM obtained is constant. Moreover, we compare two types of SMs obtained by the RIP and Revised LP-OLDF and evaluate the eight LDFs and QDF by RatioSV and the number of misclassifications (NMs). (2) The microarrays are linearly separable data (LSD). However, because the statistical discriminant functions cannot discriminate LSD theoretically, many researchers could not solve the cancer gene analysis completely from 1970 (Problem5). Moreover, the Matryoshka feature selection method (Method2) and LINGO Program3 can decompose the microarray into many SMs those are LSD. Although all SMs are small samples, many statistical methods cannot find the linear separable facts. However, RIP, Revised LP-OLDF, and H-SVM can discriminate all SMs correctly. We realized the three data made by three LDFs are signal data and reduce the high-dimensional microarray to low-dimensional signal data. (3) We propose the standard procedure for how to analyze all SMs. Specialists of gene analysis can solve the cancer gene analysis and approach the cancer gene diagnosis from the new aspect. On the other hand, statisticians recognize the difficulties of cancer gene analysis and understand the easiness of the cancer gene diagnosis by statistical methods. Statistical users can analyze many SMs those are a gift from high-dimensional data and skill-up their statistical ability to solve practical applications.

Keywords Cancer gene diagnosis · Malignancy indicators · Small Matryoshka (SM) · Revised IP-OLDF (RIP) · Revised LP-OLDF · Hard-margin SVM (H-SVM) · Signal data made by discriminant scores

3.1 Introduction

Chapter 1 outlined the new theory of discriminant analysis after R. A. Fisher and explained the successful example of cancer gene analysis (Theory). Chapter 2 described the cancer gene diagnosis and malignancy indicators. LINGO Program3

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_3

and LINGO Program4 of Revised IP-OLDF (RIP) found 64 Small Matryoshkas (SMs) and 130 basic gene sets (BGSs) of Alon's microarray. Although 63 RatioSVs among 64 SMs are higher than 5%, all RatioSVs of 130 BGSs were less than 1%. These facts showed us BGSs were useless for the cancer gene diagnosis. Thus, we decided not to look for BGSs of other five microarrays anymore. After many trials, we realized RIP discriminant scores (RipDSs) become proper malignancy indicators and contains much information. Thus, we made the signal data made by RipDSs. With this breakthrough, we can propose the cancer gene diagnosis by malignancy indicators.

This chapter discusses the following three points.

(1) We have introduced only SMs obtained with the RIP in Chap. 2. RIP analyzed SMs by Program3' arbitrary iteration number. In 2017, we increase the number of iterations successively from 1 and select the iteration number that the number of SM obtained is constant. Moreover, we compare two types of SMs obtained by the RIP and Revised LP-OLDF and evaluate the eight LDFs and QDF by RatioSV and the number of misclassifications (NMs).

(2) The microarrays are linearly separable data (LSD). However, because the statistical discriminant functions cannot discriminate LSD theoretically, many researchers could not solve the cancer gene analysis completely from 1970 (Problem5). Moreover, the Matryoshka feature selection method (Method2) and LINGO Program3 can decompose the microarray into many SMs those are LSD. Although all SMs are small samples, many statistical methods cannot find the linear separable facts easily. However, RIP, Revised LP-OLDF, and H-SVM can discriminate all SMs correctly. We realized the three data made by three LDFs are signal data and reduce the high-dimensional microarray to low-dimensional signal data.

(3) We propose the standard procedure for how to analyze all SMs. Specialists of gene analysis can solve the cancer gene analysis and approach the cancer gene diagnosis from the new aspect. On the other hand, statisticians recognize the difficulties of cancer gene analysis and understand the easiness of the cancer gene diagnosis by statistical methods. Statistical users can analyze plenty of SMs that is a gift from high-dimensional data and skill-up their statistical ability to solve practical applications.

This chapter considers the following two points because we have only studied SM obtained with the RIP in Chap. 2. (1) In this chapter, we compare two types of SMs obtained by the RIP and Revised LP-OLDF. (2) So far, we have chosen SMs by Program3' arbitrary iteration number. For this time, we decided to increase the number of iterations successively from 1 and use the number of times that the number of SM obtained is constant. In Chaps. 3 to 9, we examined with the two sets of SMs obtained by the RIP and Revised LP-OLDF. Section 3.2 examines the iteration option of LINGO Program3 and chooses the 39 SMs by Revised LP-OLDF and the 56 SMs by the RIP. Also, Revised LP-OLDF chooses 36 SMs by fixing "IT = 3." We compare three different SMs by RatioSVs and NMs. Sections 3.4 and 3.5 introduce the cancer gene diagnosis of 39 SMs and 56 SMs. Section 3.6 is the conclusion.

3.2 Outlook of This Chapter

Alon's microarray consists of the 62 cases (the 22 Normal subjects and the 40 Tumor cancer patients) with 2,000 genes. This chapter introduces the third time discrimination of Alon's microarray in 2017 and evaluates the results from a new point by LINGO ver.17. RIP finds the 56 SMs (1,999 genes), and Revised LP-OLDF find 39 SMs (992 genes). Until now, although we used an arbitrary iteration option value of LINGO Program3, we decide a proper option by changing the value this time. RIP and Revised LP-OLDF choose the different combinations of SMs. Because Revised LP-OLDF are faster than RIP and choose a small number of SMs, it is convenient for many researchers to use Revised LP-OLDF's SMs if there are no problems with SMs found by Revised LP-OLDF. Thus, we validate two different combinations of SMs by several points.

3.2.1 Alon's microarray

Alon et al. (1999)[1] published a paper entitled "Broad patterns of gene expression revealed by clustering analysis of Tumor and Normal colon tissues probed by oligonucleotide arrays." They explained their research purpose in their abstract as follows: "Oligonucleotide arrays can provide a broad picture of the state of the cell, by monitoring the expression level of thousands of genes at the same time. It is of interest to develop techniques for extracting useful information from the resulting microarrays. Here we report the application of a two-way clustering method for analyzing a microarray consisting of the expression patterns of different cell types. Gene expression in 40 tumors and 22 normal colon tissue samples were analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. An efficient two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Coregulated families of genes clustered together, as demonstrated for the ribosomal proteins. Clustering also separated cancerous from non-cancerous tissue and cell lines from in vivo tissues by subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups."

When RIP discriminated Alon's microarray by LINGO Program3, it found 66 SMs (1,131 genes) in 2015 by LINGO ver.15 (Schrage 2006). Shinmura (2015) showed all the SMs gene name. In 2016, RIP found 64 SMs (1,999 genes) by the different iteration option value by LINGO ver.16. Moreover, RIP found 130 BGSs (1,995 genes) by LINGO Program4 and manual cooperation. We are stuck in the defect of SM which selects different combinations of SM by different iteration option value and LINGO yearly version up. Although many papers said that high-dimensional noise embeds the signal, Alon's microarray has few noises. In position book (Shinmura 2017) and Chap. 2, we evaluate the 64 SMs and 130 BGSs by RatioSVs. The range

[1] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ.

of 64 RatioSVs and 130 BGSs are [2.35%, 26.76%] and [0.00%, 0.9%], respectively. Thus, we concluded BGSs were useless for cancer gene diagnosis and stopped to search other five BGSs. Although they specified 2,000 genes by two-way clustering (SOM: Self-Organizing Map) from the knowledge of cancer gene, our different approaches decomposed 2,000 genes to almost the same 64 SMs and 130 BGSs. Alon's research confirms our results. Several statistical papers pointed out it was difficult for statisticians to separate signal and noise (one of the difficulties). However, Alon's microarray is considered to be almost the signal. We cannot judge whether this fact is by chance or SOM is helpful for gene analysis in addition to RIP. Because we have no experience for SOM and it requests the proper clustering number k, we use the hierarchical cluster analysis in our research. Furthermore, because two SMs in 2015 and 2016 are different, we discuss deciding the proper iteration number in this chapter.

3.2.2 Examination of the Iteration Option of LINGO Program3

Table 3.1 shows how to determine the proper iteration values for the Revised LP-OLDF and RIP.

If we find the same number of SMs consecutively, choose the first IT value. "CPU" is the computation time (minute: second). "SM" is the number of SMs found by the specified IT. "Gene" is total number included in SMs. "Gene/SM" is the average gene number per SM. Revised LP-OLDF chooses "IT = 5" because six IT values from five to ten choose 39 SMs. Thus, we evaluate 39 SMs of Revised LP-OLDF in this chapter. Until now, we ignored the SMs except for RIP and chosen the IT by an arbitrary constant value. Revised LP-OLDF chooses the signal (992 gens) and separates 1,008 genes as the noise. Moreover, it decomposes the signal to 39 SMs. RIP chooses "IT = 3" because three IT values from three to five choose 56 SMs. Thus, we evaluate 56 SMs of RIP in this chapter. RIP chooses the signal (1,999 gens) and separates one gene as the noise. Moreover, it decomposes the signal to 56 SMs. Nevertheless, RIP found 64 SMs (1,999 genes) in 2016.

Table 3.1 Proper iteration values of revised LP-OLDF and RIP

IT	LP				RIP			
	CPU	SM	Gene	Gene/SM	CPU	SM	Gene	Gene/SM
1	10	24	1063	44	44	36	1989	55
2	**29**	**36**	**1042**	**29**	1:19	55	1965	36
3	36	37	1003	27	**1:48**	**56**	**1999**	**36**
4	47	38	988	26	2:16	56	1999	36
5	**56**	**39**	**992**	**25**	2:47	56	1999	36
6	1:08	39	992	25				
7	1:47	39	992	25				
10	1:47	39	992	25				

3.3 Comparison of 39 SMs by Revised LP-OLDF and 56 SMs by RIP

In Sect. 3.3, we compare three different SMs found by Revised LP-OLDF and RIP. We evaluate these three different SMs by six MP-based LDFs such as RIP, Revised LP-OLDF (LP), Revised IPLP-OLDF (IPLP), hard-margin SVM (H-SVM), SVM4 (penalty c = 10^4) and SVM1 (penalty c = 1) using RatioSV.

3.3.1 Result of 39 SMs by Revised LP-OLDF

3.3.1.1 Pre-survey of 36 SMs by "IT = 2"

Before examination of 39 SMs obtained by "IT = 5", we survey 36 SMs of "IT = 2" in Table 3.2. "SM" is the sequential number of 36 SMs. "Gene" is the total gene numbers included in each SM. Six columns are RatioSVs of six MP-based LDFs. "Max and Min" are maximum and minimum values of four RatioSVs except for SVM4 and SVM1. Last four columns are four NMs of SVM4, SVM1, and LDF2[2] and QDF. Although RatioSV is the proper statistic for LSD-discrimination, its values become large for "NM >= 1" and are not reliable. Last five rows are five elementary statistics. "SUM" is the total number of genes included in 36 SMs. "Max RatioSV" is the total number of SMs having a maximum value of four RatioSVs except for SVM4 and SVM1. Although the values of three OLDFs are ten, 16 RatioSVs of H-SVM are the maximum values among the four MP-based LDFs. We think the maximization of SV criterion causes this good result for H-SVM. Four ranges of RatioSV by "RIP, LP, IPLP, and HSVM" are [0.19, 36.06], [0.19, 37.44], [0.19, 37.44] and [**0.19, 37.98**], respectively. Thus, RatioSV of H-SVM is better than three OLDFs. Last four ranges of NMs are [0, 1], [0, 6], [0, 8] and [0, 0], respectively. QDF is better than SVM4, SVM1, and LDF2. In Alon's microarray, only logistic regression and QDF can discriminate 36 SMs correctly in addition to three OLDFs and H-SVM. Although four NMs are fairly small NMs, other five microarrays have larger NMs than Alon. However, because we focus on the linear separable facts, we count the number of SM with NM = 0. Those are 35, 22, 19, and 36, respectively. The only QDF discriminate 36 SMs correctly. These facts conclude SVM4, SVM1, (LDF1), and LDF2 are useless for cancer gene diagnosis.

[2] Prior probability is proportional to the number of patients.

Table 3.2 Result of 36 SMs

SM	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM4	SVM1	LDF2	QDF
1	24	25.92	15.99	15.99	**28.50**	28.48	28.50	28.50	15.99	0	0	0	0
2	24	31.69	32.88	32.88	**34.00**	34.00	34.00	34.00	31.69	0	0	0	0
3	25	36.06	**37.44**	**37.44**	36.04	36.04	36.04	37.44	36.04	0	0	0	0
4	29	35.69	33.82	33.82	**37.98**	38.00	37.98	38.00	33.82	0	0	0	0
5	25	28.41	27.48	27.48	**32.83**	32.81	32.83	32.83	27.48	0	0	0	0
6	27	30.74	**30.75**	**30.75**	27.06	27.06	27.06	30.75	27.06	0	0	0	0
7	30	27.78	**36.39**	**36.39**	32.20	32.21	32.21	36.39	27.78	0	0	0	0
8	29	26.10	25.82	25.82	**29.05**	29.05	29.05	29.05	25.82	0	0	0	0
9	29	20.30	24.58	24.58	**31.37**	31.37	31.37	31.37	20.30	0	0	0	0
10	27	31.27	34.32	34.32	**36.41**	36.41	36.42	36.42	31.27	0	0	0	0
11	31	**34.75**	30.52	30.52	28.66	28.65	28.66	34.75	28.65	0	0	0	0
12	30	27.40	**34.93**	**34.93**	32.12	32.16	32.12	34.93	27.40	0	0	0	0
13	26	31.31	**33.92**	**33.92**	32.46	32.46	33.86	33.92	31.31	0	0	0	0
14	26	**23.39**	22.21	22.21	21.34	21.34	24.06	24.06	21.34	0	0	0	0
15	25	17.31	**22.29**	**22.29**	20.79	20.79	22.76	22.76	17.31	0	0	0	0
16	30	19.71	22.71	22.71	**23.59**	23.59	26.64	26.64	19.71	0	0	1	0
17	30	26.64	**36.66**	**36.66**	31.85	31.85	32.49	36.66	26.64	0	0	0	0
18	25	19.57	22.50	22.50	**23.04**	23.04	25.83	25.83	19.57	0	0	1	0
19	26	**19.31**	19.23	19.23	18.21	18.21	22.29	22.29	18.21	0	0	3	0
20	28	18.92	18.64	18.64	**21.18**	21.18	25.29	25.29	18.64	0	0	1	0
21	28	21.56	18.79	18.79	**22.75**	22.74	24.21	24.21	18.79	0	0	1	0
22	27	25.15	24.53	24.53	**28.38**	28.38	31.20	31.20	24.53	0	0	0	0
23	29	25.00	**30.19**	**30.19**	24.94	24.94	~~37.75~~	37.75	24.94	0	**1**	1	0
24	28	24.43	**27.08**	**27.08**	25.74	25.74	~~38.37~~	38.37	24.43	0	**2**	0	0
25	28	17.32	20.18	20.18	**21.17**	21.17	~~32.56~~	32.56	17.32	0	**3**	0	0
26	29	**21.95**	17.10	17.10	20.86	20.86	~~39.66~~	39.66	17.10	0	**4**	2	0
27	29	**16.01**	14.54	14.54	15.68	15.68	~~30.50~~	30.50	14.54	0	**4**	2	0
28	28	**16.83**	14.41	14.41	16.58	16.58	~~29.16~~	29.16	14.41	0	**3**	5	0
29	32	13.71	16.81	16.81	**17.08**	17.08	~~38.72~~	38.72	13.71	0	**5**	3	0
30	30	7.05	7.43	7.43	**8.24**	8.24	~~27.27~~	27.27	7.05	0	**5**	6	0
31	31	**13.97**	7.89	7.89	13.69	13.69	~~32.34~~	32.34	7.89	0	**5**	4	0
32	29	9.43	9.30	9.30	**10.16**	10.16	~~43.98~~	43.98	9.30	0	**5**	4	0
33	36	**18.67**	9.54	9.54	15.64	15.64	~~35.94~~	35.94	9.54	0	**5**	3	0
34	33	**7.33**	6.28	6.28	7.29	7.29	~~49.24~~	49.24	6.28	0	**5**	6	0
35	35	**6.19**	5.94	5.94	**6.19**	6.19	~~39.80~~	39.80	5.94	0	**6**	8	0
36	44	**0.19**	**0.19**	**0.19**	**0.19**	~~2.55~~	~~41.06~~	41.06	**0.19**	**1**	**5**	6	0
Max	RatioSV	10	10	10	16								
MAX	44	36.06	37.44	37.44	**37.98**	38.00	49.24	49.24	36.04	1	6	8	0
MIN	24	0.19	0.19	0.19	**0.19**	2.55	22.29	22.29	0.19	0	0	0	0
Mean	28.9	21.585	22.036	22.036	23.146	23.213	32.533	33.157	20.056	0	2	2	0
SUM	1042												

3.3.1.2 Examination of 39 SMs

Table 3.3 shows the result of 39 SMs. Last three columns are three NMs of SVM1, LDF2, and QDF. Because all NMs of logistic regression and SVM4 are zero, those are omitted from the table. "Max RatioSV" row is the total number of SMs having a maximum value of four RatioSVs. Four values of three OLDFs and H-SVM are 13, 12, 10, and 17. The maximization of SV causes this good result for H-SVM. Five ranges of "gene, RIP, LP, IPLP, and HSVM" are [18, 35], [0.78, 40.59], [**0.74, 45.14**], [**0.74, 45.14**], and [0.78, 43.95], respectively. Thus, Revised LP-OLDF and Revised IPLP-OLDF can separate two classes more than RIP and H-SVM. Last three ranges are [0, 9], [0, 8], and [0, 0], respectively. QDF, SVM4, and logistic regression can discriminate 39 SMs correctly. SVM1 and LDF2 cannot discriminate 17 SMs and 19 SMs. There are 1,042 and 1,002 genes in 34 SMs and 39 SMs, but 39 SMs contain genes smaller than 40 genes. By increasing repetition, we obtain many SMs with fewer genes, so choose 39 SMs. These facts indicate we choose the proper iteration value at this time. However, because we focus on the linear separable facts, we count the number of SM with NM = 0. The SVM1 and LDF2 discriminate 22 and 20 SMs correctly. The SVM4 and QDF discriminate all SMs correctly. These facts conclude SVM1, (LDF1), and LDF2 are useless for cancer gene diagnosis.

Table 3.3 Result of 39 SMs

SM	IT	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM1	LDF2	QDF
1	6	20	**28.01**	26.83	26.83	27.21	27.20	27.21	28.01	26.83	0	0	0
2	6	24	36.05	26.98	26.98	**37.17**	37.18	37.17	37.18	26.98	0	0	0
3	6	19	36.04	**41.06**	41.06	38.37	38.36	38.37	41.06	36.04	0	0	0
4	6	23	31.52	**36.85**	36.85	35.31	35.31	35.31	36.85	31.52	0	0	0
5	6	28	29.83	30.32	30.32	**38.13**	38.14	38.13	38.14	29.83	0	0	0
6	6	24	27.26	27.61	27.61	**28.60**	28.59	28.60	28.60	27.26	0	0	0
7	6	24	40.59	36.92	36.92	**43.95**	43.95	43.95	43.95	36.92	0	1	0
8	6	28	35.98	**45.14**	**45.14**	41.83	41.83	41.83	45.14	35.98	0	0	0
9	6	28	29.45	**37.50**	**37.50**	35.04	35.04	35.04	37.50	29.45	0	0	0
10	6	25	21.40	23.27	23.27	**32.63**	32.63	32.63	32.63	21.40	0	0	0
11	6	29	23.30	23.76	23.76	**28.65**	28.64	29.02	29.02	23.30	0	0	0
12	6	25	26.24	**29.76**	**29.76**	28.92	28.92	28.92	29.76	26.24	0	0	0
13	6	27	**24.83**	22.39	22.39	22.58	22.58	25.11	25.11	22.39	0	0	0
14	6	18	**24.83**	22.39	22.39	22.58	22.58	25.11	25.11	22.39	0	0	0
15	6	27	**35.93**	34.43	34.43	34.97	34.95	35.20	35.93	34.43	0	0	0
16	6	26	23.34	22.12	22.12	**26.95**	26.96	28.38	28.38	22.12	0	1	0
17	6	23	**19.54**	18.38	18.38	19.40	19.40	21.31	21.31	18.38	0	0	0
18	6	27	24.83	28.60	28.60	**29.54**	29.54	30.50	30.50	24.82	0	0	0

(continued)

Table 3.3 (continued)

SM	IT	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM1	LDF2	QDF
19	6	28	23.09	**26.45**	**26.45**	23.72	23.72	27.14	27.14	23.09	0	1	0
20	6	25	**33.11**	23.68	23.68	32.76	32.74	37.87	37.87	23.68	0	0	0
21	6	28	17.76	**21.79**	**21.79**	21.64	21.64	26.20	26.20	17.76	0	0	0
22	6	26	25.31	**28.27**	**28.27**	27.60	27.60	29.16	29.16	25.31	0	0	0
23	6	23	**21.39**	19.04	19.04	20.86	20.86	~~31.35~~	31.35	19.04	**2**	1	0
24	6	23	**19.26**	16.90	16.90	18.87	18.87	~~27.76~~	27.76	16.90	**2**	1	0
25	6	26	**20.59**	19.01	19.01	20.51	20.51	~~23.90~~	23.90	19.01	**3**	0	0
26	6	25	**19.20**	18.32	18.32	18.95	18.95	~~26.99~~	26.99	18.31	**2**	3	0
27	6	21	19.23	19.74	19.74	**19.81**	19.81	~~37.19~~	37.19	19.23	**4**	1	0
28	6	26	11.66	13.10	13.10	**14.00**	14.00	~~26.36~~	26.36	11.66	**6**	5	0
29	6	23	9.63	12.61	12.61	**13.23**	13.22	~~25.43~~	25.43	9.63	**3**	4	0
30	6	21	9.19	**10.71**	**10.71**	**10.71**	10.71	~~35.90~~	35.90	9.19	**6**	3	0
31	6	28	20.77	**20.78**	**20.78**	21.29	21.29	~~37.17~~	37.17	20.77	**5**	1	0
32	6	27	10.16	11.22	11.22	**11.38**	11.38	~~31.63~~	31.63	10.16	**5**	3	0
33	6	26	7.70	8.07	8.07	**9.98**	9.98	~~30.55~~	30.55	7.70	**6**	3	0
34	6	27	7.88	8.85	8.85	**9.34**	9.34	~~35.62~~	35.62	7.88	**6**	4	0
35	6	28	**7.43**	**7.43**	**7.43**	7.26	7.26	~~30.59~~	30.59	7.26	**4**	8	0
36	6	32	8.11	**10.15**	**10.15**	9.08	9.08	~~34.03~~	34.03	8.11	**5**	5	0
37	6	26	3.03	3.17	3.17	**3.57**	3.57	~~37.16~~	37.16	**3.03**	**9**	6	0
38	6	33	**3.03**	2.64	2.64	3.01	3.01	~~37.11~~	37.11	**2.64**	**5**	6	0
39	6	25	**0.78**	0.74	0.74	**0.78**	0.78	~~34.68~~	34.68	**0.74**	**5**	6	0
Max RatioSV			13	12	10	17							
Max		35	40.59	45.14	45.14	**43.95**	43.95	43.95	45.14	36.92	9	8	0
Min		18	0.78	0.74	0.74	**0.78**	0.78	21.31	21.31	0.74	0	0	0
Mean		26	20.96	21.46	21.46	22.82	22.82	31.94	32.26	19.93	2	2	0
Sum		992											

3.3.2 Result of 56 SMs by RIP

Table 3.4 shows the result of 56 SMs found by RIP. Last three columns are three NMs of SVM1, LDF2, and QDF. Because NMs of logistic regression and SVM4 are zero, we omit two columns. RatioSV is the proper statistic for LSD-Discrimination. However, its values become large for "NM >= 1" and are not reliable. Last five rows are five elementary statistics. "Max RatioSV" is the total number of SMs having a maximum value of four RatioSVs. Four values of MP-based LDFs are 7, 6, 6, and 43. The maximization of SV causes this good result for H-SVM. Five ranges of "gene, RIP, LP, IPLP, and HSVM" are [29, 45], [2.98, 25.25], [4.72, 25.94], [4.72, 25.94], and [**5.76, 32.58**], respectively. Thus, H-SVM can discriminate two classes better than other LDFs. Last three ranges are [0, 5], [0, 7], and [0, 0], respectively. QDF, SVM4, and logistic regression discriminate two classes correctly. Because 56 SMs include 1,999 genes, 56 SMs include lager genes than two SMs by Revised LP-OLDF. The SVM4 and QDF discriminate 56 SMs correctly. Although SVM1 can discriminate 31 SMs and LDF2 can discriminate 23 SMs correctly, these discriminant functions are useless for cancer gene diagnosis.

Table 3.4 Result of 56 SMs by RIP

SM	IT	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM1	LDF2	QDF
1	6	35	7.19	12.77	12.77	**19.98**	19.98	19.98	19.98	7.19	0	**1**	0
2	6	34	11.06	12.11	12.11	**16.90**	16.90	16.90	16.90	11.06	0	**1**	0
3	6	29	14.42	7.52	7.52	**19.46**	19.46	19.46	19.46	7.52	0	**2**	0
4	6	44	18.01	14.76	14.76	**23.36**	23.36	23.36	23.36	14.76	0	0	0
5	6	35	21.55	25.94	25.94	**32.58**	32.58	32.58	32.58	21.55	0	0	0
6	6	36	17.71	23.41	23.41	**23.47**	23.47	23.47	23.47	17.71	0	2	0
7	6	37	11.57	12.32	12.32	**19.29**	19.29	19.29	19.29	11.57	0	1	0
8	6	37	**25.25**	24.14	24.14	23.54	23.55	23.54	25.25	23.54	0	0	0
9	6	35	14.20	12.03	12.03	**21.82**	21.81	21.82	21.82	12.03	0	**2**	0
10	6	31	18.18	14.14	14.14	**18.36**	18.36	18.36	18.36	14.14	0	**1**	0
11	6	35	14.69	12.04	12.04	**19.64**	19.64	19.64	19.64	12.04	0	0	0
12	6	35	15.16	23.16	23.16	**26.05**	26.05	26.05	26.05	15.16	0	0	0
13	6	34	22.56	19.45	19.45	**25.84**	25.84	25.84	25.84	19.45	0	0	0
14	6	35	22.96	22.65	22.65	**25.98**	25.98	25.98	25.98	22.65	0	0	0
15	6	34	18.33	20.93	20.93	**25.08**	25.08	25.08	25.08	18.33	0	**1**	0
16	6	36	22.08	16.18	16.18	**28.84**	28.84	28.84	28.84	16.18	0	**1**	0
17	6	33	15.45	17.47	17.47	**19.76**	19.76	19.76	19.76	15.45	0	**2**	0
18	6	36	19.53	17.17	17.17	**24.76**	24.75	~~24.76~~	24.76	17.17	0	0	0
19	6	32	17.35	16.96	16.96	**21.13**	21.13	~~23.78~~	23.78	16.96	1	0	0
20	6	41	17.31	18.37	18.37	**22.79**	22.79	28.25	28.25	17.31	1	0	0
21	6	32	12.38	**17.68**	**17.68**	17.13	17.13	18.74	18.74	12.38	0	**2**	0
22	6	34	18.39	**20.16**	**20.16**	18.96	18.96	18.96	20.16	18.39	0	0	0
23	6	34	15.73	12.36	12.36	**19.45**	19.45	21.53	21.53	12.36	0	**1**	0
24	6	33	**17.26**	13.93	13.93	17.19	17.19	19.08	19.08	13.93	0	0	0
25	6	37	10.90	15.90	15.90	**19.66**	19.66	19.66	19.66	10.90	0	**2**	0
26	6	31	11.25	14.45	14.45	**17.69**	17.69	21.13	21.13	11.25	0	**1**	0
27	6	32	10.50	4.94	4.94	**15.27**	15.28	~~30.28~~	30.28	**4.94**	1	**1**	0
28	6	33	12.82	14.68	14.68	**19.18**	19.18	20.81	20.81	12.82	0	**2**	0
29	6	36	12.38	10.83	10.83	**18.47**	18.47	~~22.20~~	22.20	10.83	1	**1**	0
30	6	36	17.69	**21.65**	**21.65**	21.58	21.58	22.55	22.55	17.69	0	0	0
31	6	32	**17.41**	10.79	10.79	16.81	16.81	~~18.93~~	18.93	10.79	1	1	0
32	6	34	23.04	16.92	16.92	**24.63**	24.63	24.63	24.63	16.92	0	0	0
33	6	39	17.78	15.20	15.20	**18.52**	18.52	~~26.27~~	26.27	15.20	2	2	0
34	6	34	15.71	15.62	15.62	**20.82**	20.81	22.76	22.76	15.62	0	2	0
35	6	38	**18.95**	15.65	15.65	15.23	15.23	~~23.59~~	23.59	15.23	2	2	0
36	6	32	13.48	11.56	11.56	**17.89**	17.89	~~32.52~~	32.52	11.56	2	2	0

(continued)

Table 3.4 (continued)

SM	IT	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM1	LDF2	QDF
37	6	42	20.12	20.82	20.82	**28.65**	28.65	28.65	28.65	20.12	0	0	0
38	6	37	14.95	8.33	8.33	**17.69**	17.69	36.37	36.37	8.33	3	2	0
39	6	31	13.24	9.65	9.65	**18.63**	18.63	20.65	20.65	9.65	0	1	0
40	6	36	10.23	11.10	11.10	**15.12**	15.12	~~25.93~~	25.93	10.23	3	4	0
41	6	35	12.77	15.46	15.46	**21.12**	21.12	25.38	25.38	12.77	0	0	0
42	6	38	13.32	9.96	9.96	**18.05**	18.05	~~23.64~~	23.64	9.96	1	**2**	0
43	6	34	**13.63**	9.65	9.65	12.10	12.10	~~23.59~~	23.59	9.65	2	**2**	0
44	6	35	5.89	11.42	11.42	**12.90**	12.90	~~20.71~~	20.71	5.89	4	**5**	0
45	6	37	**13.40**	9.43	9.43	12.43	12.43	~~23.93~~	23.93	9.43	4	**2**	0
46	6	35	9.48	8.32	8.32	**13.10**	13.10	~~24.00~~	24.00	8.32	2	**2**	0
47	6	35	**9.38**	8.01	8.01	8.77	8.77	~~25.72~~	25.72	8.01	3	**2**	0
48	6	36	8.30	6.83	6.83	**14.77**	14.77	~~24.81~~	24.81	6.83	5	**3**	0
49	6	43	12.12	12.74	12.74	**18.11**	18.11	~~25.74~~	25.74	12.12	3	**2**	0
50	6	33	5.92	**8.81**	**8.81**	8.34	8.34	~~28.74~~	28.74	5.92	4	**4**	0
51	6	42	12.30	10.12	10.12	**15.71**	15.70	~~32.80~~	32.80	10.12	5	**3**	0
52	6	32	9.29	**10.46**	**10.46**	9.65	9.65	~~37.33~~	37.33	9.29	4	**5**	0
53	6	39	8.48	8.12	8.12	**10.48**	10.48	~~31.30~~	31.30	8.12	4	**7**	0
54	6	41	11.79	6.75	6.75	**17.19**	17.19	~~36.85~~	36.85	6.75	4	**3**	0
55	6	42	2.98	4.72	4.72	**5.76**	5.76	~~36.18~~	36.18	**2.98**	5	**5**	0
56	6	45	9.29	**10.46**	**10.46**	9.65	9.65	~~37.33~~	37.33	9.29	4	**6**	0
Max RatioSV			7	6	6	43							
Max		45	25.25	25.94	25.94	**32.58**	32.58	37.33	37.33	23.54	5	7	0
Min		29	2.98	4.72	4.72	**5.76**	5.76	16.90	16.90	2.98	0	0	0
Mean		36	14.41	13.88	13.88	18.67	18.67	25.00	25.05	12.61	1	2	0
Sum		1999											

3.3.3 Comparison of Three Results

Table 3.5 summarizes the three results. “SM” column shows three different SMs. “Gene” column shows the total genes included in all SMs. The 56 SMs by RIP include 1,999 genes and is double of others. Three “Best Range of RatioSV (By LDF column)” are the ranges of H-SVM. In the 39 SMs by LP, the number of genes involved is small, and the maximum value of RatioSV is the largest. On the other hand, the 56 SMs by RIP have a sizeable minimum value of RatioSV over 5.78. From these results, the result of 56 SMs may be better than 36 SMs and 39 SMs.

Table 3.5 Summary of three results

SM	Gene	Best range of RatioSV	By LDF
36 SMs by LP	1,042	[0.19, 37.98]	H-SVM
39 SMs by LP	992	[0.78, 43.95]	H-SVM
56 SMs by RIP	1,999	[5.78, 32.58]	H-SVM

3.4 Three Signal Data Using 39 SMs Found by Revised LP-OLDF

Although standard statistical methods analyze all 39 SMs found by Revised LP-OLDF, the results are almost the same useless results explained in Chap. 2. Thus, we omit these results. Section 3.4 introduces only results of cluster analysis and PCA. RIP, Revised LP-OLDF, and H-SVM discriminated all 39 SMs obtained by Revised LP-OLDF and made three signal data by RipDSs, Revised LP-OLDF DSs (LpDSs) and H-SVM DSs (HsvmDSs). Ward cluster and PCA explain three results. Because JMP supports six hierarchical cluster methods, we get the 18 different analyses and survey the various aspects of SMs. However, it is difficult for non-specialists of oncogenes to study the cluster analysis. Thus, we propose a useful procedure on how to use the cluster analysis. On the other hand, non-specialists can understand the results of PCA and the Prin1 becomes important malignancy indicators in addition to 39 individual RipDSs, LpDSs, and HsvmDSs. Moreover, we expect several outliers are candidates of the new subclass of cancer pointed out by Golub et al. We expect our approach by cluster analysis and PCA will assist many researchers. We sincerely hope some researchers validate our claim and will write a paper.

3.4.1 Signal Data Made by 39 RipDSs Using 39 SMs Found by Revised LP-OLDF

(1) The combination of cluster analysis and PCA

Figure 3.1 is the result of RipDS signal data using 39 SMs found by Revised LP-OLDF. Although Ward cluster cannot separate 39 SMs into two classes clearly, it can divide the signal data into two clusters.

We introduce the output of cluster analysis. The left part is the case number with symbols. Next large square is the color (or heat) map that consists of 62 cases (rows) and 39 RipDSs (columns). Each row corresponds to case that includes 39 variables (39 RipDSs). Upper green part is 22 normal cases, and lower white, and red part is 40 tumor patients. The tone of green → white → red corresponds to the magnitude of the value corresponding to (case and variable). A red pixel indicates that the value of the (case and variable) is large. Thus, the color map shows the 62 cases are entirely separable into two clusters such as the 22 normal class and 40 tumor cases. Right plot is the dendrogram of 22 normal subjects and 40 tumor patients. Move the upper right diamond left or right to divide 62 cases into the desired number of clusters. We can interactively obtain different clusters by a simple operation. On the other hand, SOM and k-means is generally difficult to use because it is necessary to determine the number of clusters in advance. We have designated five clusters here, and Ward cluster divides the tumor into four clusters. Five clusters can be identified by color and left symbol.

Below figures of the color plot are the dendrogram of variable and the scree plot. If we divide four clusters, those consist of 22, 9, 5, and 3 RipDSs, respectively. Because RIP9 and RIP10 become the first cluster, the correlation of RipDS9 and RipDS10 may be close to 1 and compatible. RIP32 and RIP33 become the second cluster.

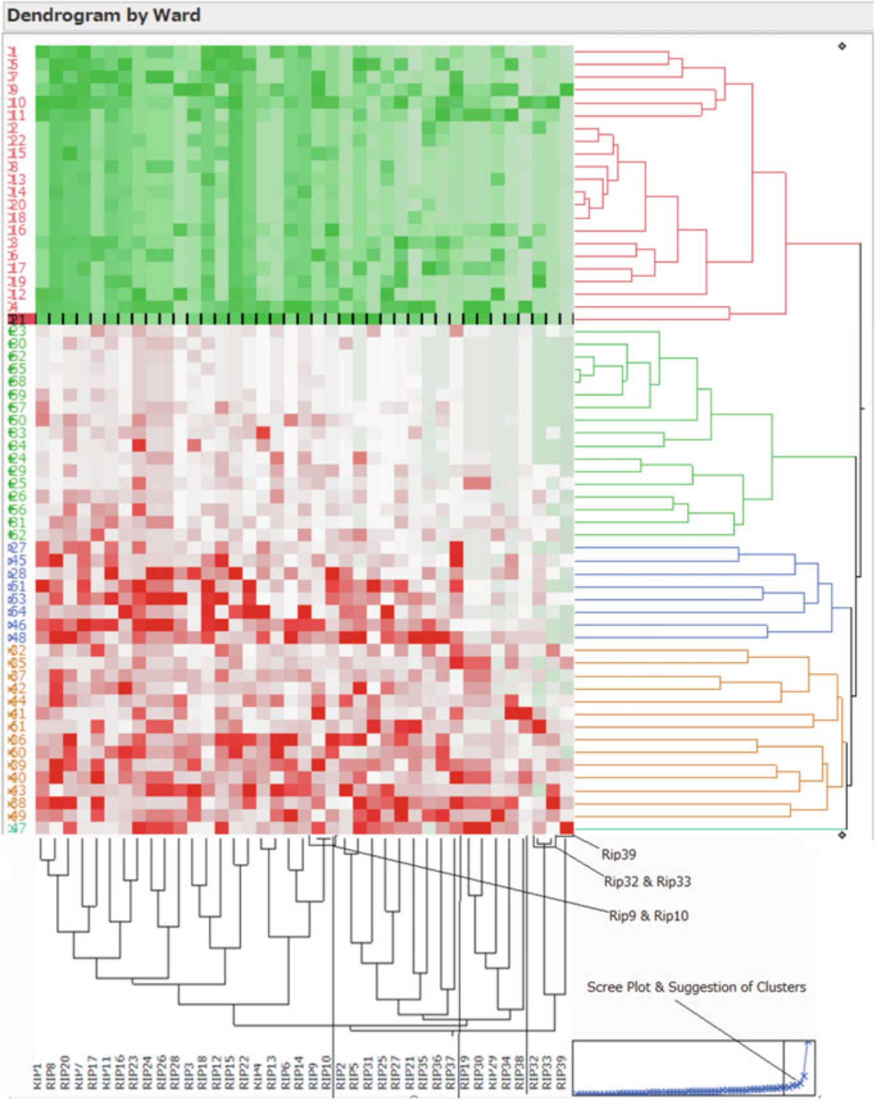
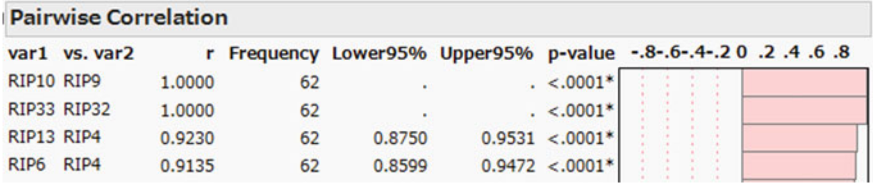


Fig. 3.1 Result of RipDSs signal data using 39 SMs of Revised LP-OLDF

Figure 3.2 shows the above two-pair correlations are one. That is, the correlation is often 1 in pairs that are clustered quickly in cluster analysis.



Pairwise Correlation

var1	vs. var2	r	Frequency	Lower95%	Upper95%	p-value	-.8-.6-.4-.2 0 .2 .4 .6 .8
RIP10	RIP9	1.0000	62	.	.	<.0001*	
RIP33	RIP32	1.0000	62	.	.	<.0001*	
RIP13	RIP4	0.9230	62	0.8750	0.9531	<.0001*	
RIP6	RIP4	0.9135	62	0.8599	0.9472	<.0001*	

Fig. 3.2 Pairs correlations

As we move the diamond to the left, the clustered distance is small, and it becomes 62 clusters. As we slide to the right, the distance to be clustered increases, the cases are sequentially assembled and finally become one cluster. If we regard the scree plot as a cliff, rolling stone from above will stop at the flat hem. So, we decide the number of clusters. In this figure, we choose five clusters. We chose the five clusters of the case by diamond mark and scree plot. However, we choose the four clusters of the variable by our judgment.

After cluster analysis with mark and color, we recommend PCA in Fig. 3.3. Cluster analysis is difficult without data expertise. Furthermore, it is not generally definitive. However, because the discriminant analysis is not useful at all in medical gene research, medical researchers use cluster analysis. We think this is because it is easy to find out what represents the results of medical care among many cluster results. Non-experts like us advance combinations with PCA. The analysis results of the five clusters are represented by colors and marks in the middle scatter plot. Especially, 20 normal cases located on the negative Prin1 axis are less than −4, and 42 tumor patients look like a fan that consists of four clusters. The green cluster is a mild tumor. Brown and blue are divided into nearly the first and fourth quadrants and are severe. These results correspond to the RipDSs in the first and the fourth quadrants of the right factor loading amount. The 47th tumor case is an outlier, corresponding to RIP39 from factor loading. This result corresponds to the red pixel in the lower right corner of Fig. 3.1. Furthermore, it is meaningful to examine whether 38 RipDSs without RIP39 become two groups in the first and fourth quadrants. Golub et al. narrow down genetic candidates by signal/noise ratio, further select by weighted vote method, and predict new subclass of cancer with SOM. This operation involves much work as the definition of the signal is incorrect. However, if we admit signal data is signal space, the analysis is easy. However, our result needs to be verified and confirmed by the doctors.

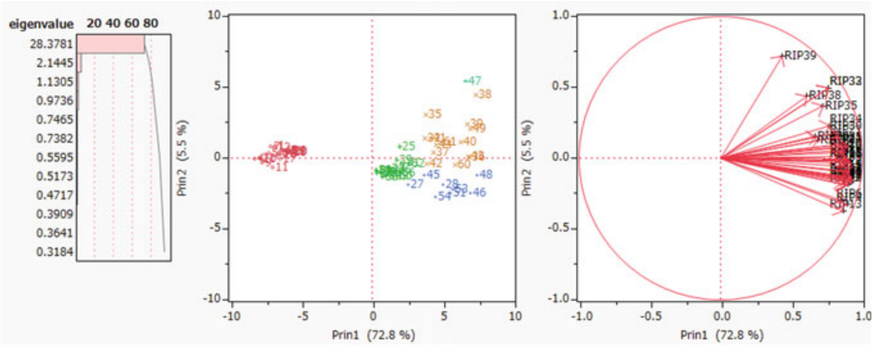


Fig. 3.3 PCA reflecting the results of cluster analysis

(2) How to validate PCA result

The first and second columns of Table 3.6 correspond to a malignancy indicator on the Prin1 in Fig. 3.3. The following remaining four columns are for Figs. 3.8 and 3.13. The numbers below the third row in the first column are the identification number (SM) for normal and cancer patients. The second column is the value of Prin1, which is sorted in ascending order from a small value. Because the minimum value of normal subjects is −8.24 (SN = 4) and the maximum value is −5.67 (SN = 20), its range is [−8.24, −5.67]. On the other hand, the range of 40 cancer patients is [0.28, 7.38]. Thus, the SV separates two classes by large SV window (−5.67, 0.28) completely. Because the range of RipDS is [−8.24, 7.38] on Prin1, RatioSV of PCA is 38.1027% (=(5.67 + 0.28)/(8.24 + 7.38) * 100). Because this is a comprehensive of RatioSV of 39 RIPs, it becomes a large value. That is, Prin1 can be considered as a malignancy indicator and illustrates an idea of malignancy indicator very well. Although we cannot understand the relation of 39 RipDSs, PCA shows most RipDSs have almost the same axes. This scatter plot shows that 992 genes included in 39 SMs reduce to the 39-dimensional signal space. A large window of 38.10% opens. Assuming that it is about 40%, it means that normal and cancer subjects share in the remaining 60% of RipDS range. Such easy discrimination is not in past research. However, it is as a new research field of statistics, and it was impossible to solve until now. We solved this new problem (Problem6). Only green and blue clusters of Fig. 3.3 correspond to G and B in Table 3.6. RIP62 has the maximum discriminant score of 2.59 in the green cluster. RIP27 has the minimum discrimination score of 2.56 in the blue cluster, and RIP48 has the maximum discriminant score of 7.38 in the blue cluster. RIP47 has the discriminant score of 6.63 and is the outlier.

Table 3.6 Three malignancy indices

RatioSV	38.10	35.52		36.22	
SN	RIPPrin1	SN	LPPrin1	SN	HSVMPrin1
4	**−8.24**	4	−8.52	4	−8.90
21	−7.96	10	−8.02	10	−8.32
10	−7.90	21	−7.96	21	−8.04
9	−7.71	9	−7.90	9	−7.68
5	−7.23	11	−7.32	11	−7.31
7	−7.16	5	−6.99	7	−7.05
11	−7.02	7	−6.90	3	−7.03
12	−6.84	12	−6.89	12	−6.90
1	−6.78	3	−6.77	5	−6.89
3	−6.68	1	−6.73	1	−6.77
6	−6.54	19	−6.43	19	−6.50
17	−6.49	6	−6.41	2	−6.48
16	−6.44	17	−6.28	13	−6.37
19	−6.27	13	−6.28	17	−6.35
2	−5.95	16	−6.17	15	−6.29
15	−5.94	15	−6.17	6	−6.26
8	−5.92	2	−6.06	16	−6.23
22	−5.91	22	−6.06	14	−6.21
13	−5.86	14	−6.04	22	−6.08
14	−5.82	8	−5.92	8	−6.05
18	−5.72	20	−5.67	20	−5.86
20	**−5.67**	**18**	**−5.65**	**18**	**−5.85**
58 G (green)	**0.28**	**52**	**0.17**	**55**	**0.27**
55 G (green)	0.30	55	0.20	52	0.28
52 G (green)	0.36	59	0.28	58	0.32
59 G (green)	0.60	58	0.32	59	0.33
30 G (green)	0.70	30	0.53	30	0.58
34 G (green)	0.97	57	0.72	34	0.88
57 G (green)	1.11	34	0.92	57	1.00
23 G (green)	1.19	23	1.28	24	1.32
24 G (green)	1.26	50	1.32	23	1.32
33 G (green)	1.49	24	1.49	56	1.48
26 G (green)	1.68	29	1.53	50	1.49
29 G (green)	1.74	33	1.61	29	1.58
50 G (green)	1.74	56	1.74	33	1.78
25 G (green)	1.91	26	2.09	25	2.17
56 G (green)	1.96	25	2.14	26	2.25

(continued)

Table 3.6 (continued)

RatioSV	38.10	35.52		36.22	
SN	RIPPrin1	SN	LPPrin1	SN	HSVMPrin1
31 G (green)	2.25	31	2.31	62	2.40
27 B (blue)	**2.56**	62	2.39	27	2.43
62 G (green)	**2.59**	27	2.57	31	2.51
45 B (blue)	3.70	32	3.55	32	3.68
32	3.70	45	3.67	45	3.82
35	3.77	42	3.84	42	3.99
42	3.84	35	4.09	35	4.22
41	4.14	41	4.27	41	4.29
37	4.35	44	4.39	37	4.73
54 B (blue)	4.47	37	4.64	61	4.81
44	4.52	54	4.68	54	4.97
61	4.88	61	5.23	44	5.02
28 B (blue)	5.00	28	5.43	28	5.27
51 B (blue)	5.54	51	5.57	60	5.42
53 B (blue)	5.68	60	5.63	51	5.84
60	5.86	53	5.76	53	5.90
40	6.29	40	5.97	40	5.97
47 (outlier)	**6.63**	49	6.66	49	6.28
39	6.73	47	6.71	47	6.50
43	6.80	46	6.75	43	6.68
36	6.86	43	6.92	46	6.96
49	6.90	39	7.04	39	7.33
46 B (blue)	6.95	36	7.21	36	7.36
38	7.36	48	7.66	38	7.99
48 B (blue)	**7.38**	38	7.87	48	8.00

(3) Analysis of Transposed Data of 39 RipDSs using 39 SMs

We transpose signal data made by 39 RipDSs using 39 SMs and analyze it by Ward cluster in Fig. 3.4. We choose the nine clusters of 39 RIPs. Upper 27 RipDSs become one cluster. Other 12 RipDSs become eight clusters such as two green RipDSs (RIP21, RIP28), four blue RipDSs (RIP32, RIP33, RIP35, RIP36), and six one clusters such as (RIP29), (RIP34), (RIP30), (RIP37), (RIP38) and (RIP39). Other 12 RIPs may relate to the outliers. All transposed analyses indicate many outliers that offer many candidates of new subclasses of cancer.

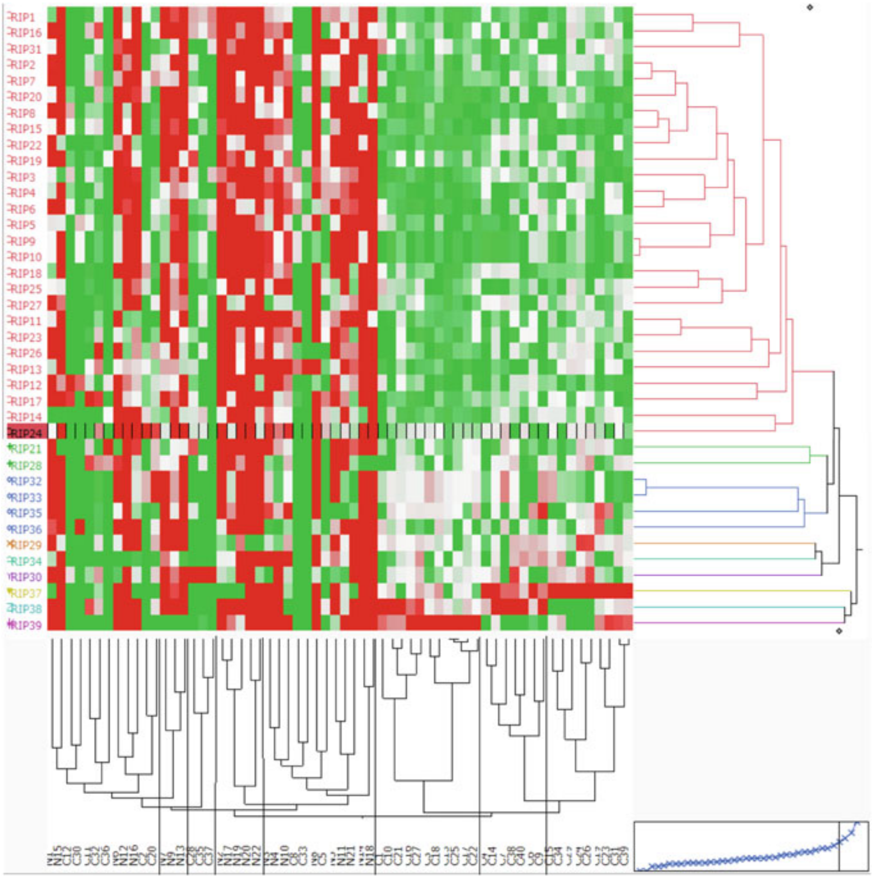


Fig. 3.4 Ward cluster of transpose signal data made by 39 RipDSs using 39 SMs

The variable dendrogram consists of eight clusters. The first cluster includes five normal cases and seven tumor cases. The second cluster includes three normal cases. The third cluster consists of three tumor cases. The fourth cluster consists of five normal cases. The fifth cluster consists of nine normal cases and three tumor cases. The sixth cluster includes 11 tumor cases. The seventh cluster consists of seven tumor cases. The eighth cluster consists of nine tumor cases.

Figure 3.5 is the result of transposed data of 39 RipDSs signal data by PCA. Scatter plot shows six RipDSs of Fig. 3.4 are six one cluster outliers. Factor loading plot indicates two features. Tumor cases are in the first and fourth quadrants, and normal cases are in the second and third quadrants. The second feature is divided into patients centered at the origin and patients group having a radius close to 1. The principles of interpretation of PCA's score plot and factor loading plot are the same. In each quadrant, we can judge that clusters making clumps have the same properties. Therefore, if those are making lumps across quadrants, we should consider them separately. Since we cannot consider in the high dimensional quadrant, we will limit it to Prin1 and Prin2 as a simple method. The factor loading plot in Fig. 3.5 is difficult to understand. Therefore, as shown in Fig. 3.6, a scatter diagram of a factor load amount is separately made. There are cases surrounding the origin of radius 0.25. Think of them as four clusters, or because they are close to the origin, ignoring the quadrant difference, it can be considered to be a cluster of relatively mild cancer patients and healthy cases close to cancer patients. For cases scattered in the other four quadrants, we can consider two cancer patient clusters in the first and fourth quadrants and two clusters of healthy cases in the second and third quadrants. Cancer patients in first quadrant are divided into two more clusters, but whether it is meaningful or not is what the medical expert should judge. In this way, clusters that can be classified by PCA provide different information from discriminant analysis and cluster analysis. Although Golub et al. tried to find a new subclass of cancer by different methods, our approach is more straightforward because we analyze the signal subspace. These verifications are areas of specialists.

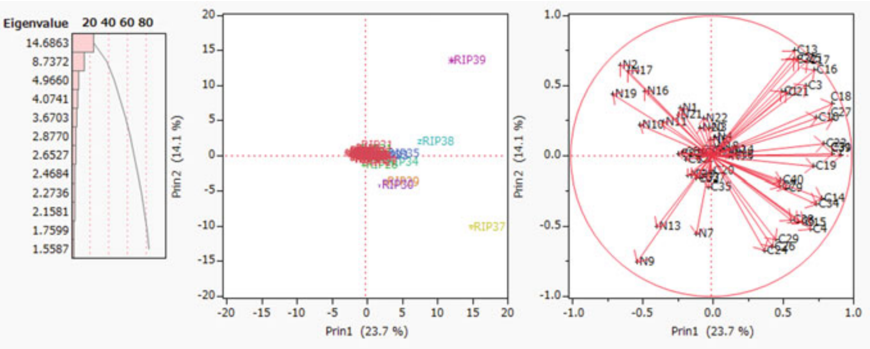


Fig. 3.5 Transposed data of RipDSs signal data by PCA

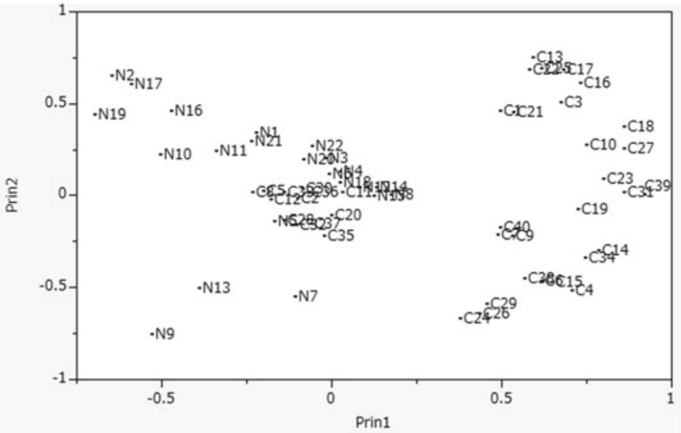


Fig. 3.6 Factor loading plot of transposed data of RipDSs using 39 SMs

3.4.2 Signal Data Made by 39 LpDSs Using 39 SMs Found by Revised LP-OLDF

(1) Ward Cluster

Figure 3.7 is the result of LpDS signal data using 39 SMs found by Revised LP-OLDF. Although Ward cluster cannot separate 39 SMs into two classes clearly, it can divide the signal data into two clusters.

Upper green part is 22 normal cases, and lower white and red part is 40 tumor patients. We have designated five clusters here, and Ward cluster divides the tumor into four clusters. The color and left symbol identify five clusters.

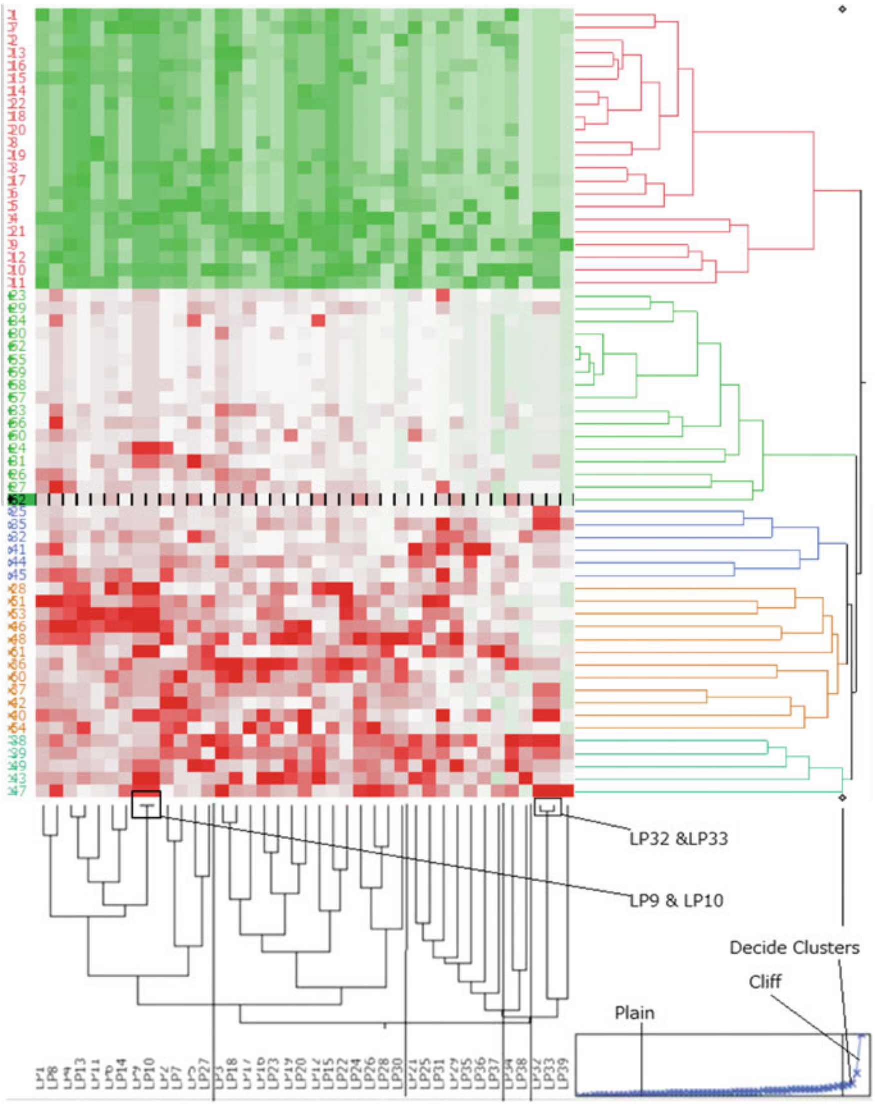


Fig. 3.7 Ward cluster result of signal data made by 39 LpDSs using SMs

If we divide the variable dendrogram into five clusters, those consist of 13, 14, 7, 2, and 3 LpDSs, respectively. Because LP9 and LIP10 become the first cluster and its correlation is 1, those are compatible. The correlation of LP32 and LP33 is one, also.

(2) PCA

Figure 3.8 is the result of PCA with colors and marks. The scatter plot shows 20 normal cases locate on the negative Prin1 axis, and 42 tumor patients look like a fan that consists of four clusters. The green cluster is a mild tumor. Brown patients are in the first and fourth quadrants. Pale green patients in the first quadrant are outliers and severe cancer. The factor loading plot shows the LP30 is a clear outlier. The third and fourth columns of Table 3.6 are the identification number (SN) and the Prin1 value. We sincerely hope researchers of Alon project validate whether Table 3.6 corresponds to severity of cancer patients.

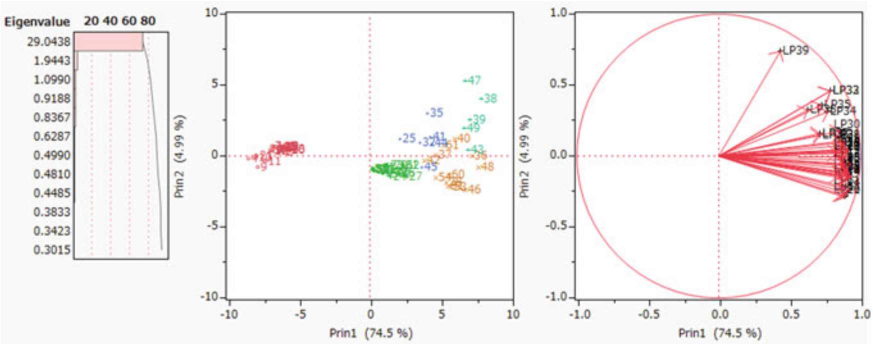


Fig. 3.8 Result of PCA with colors and marks

(3) Analysis of Transposed Signal data

Figure 3.9 is the Ward cluster of transposed signal data made by 39 LpDSs using 39 SMs. If we choose five clusters, the 35 LpDSs become one large cluster and other four one clusters those are LP34, LP37, LP38, and LP39.

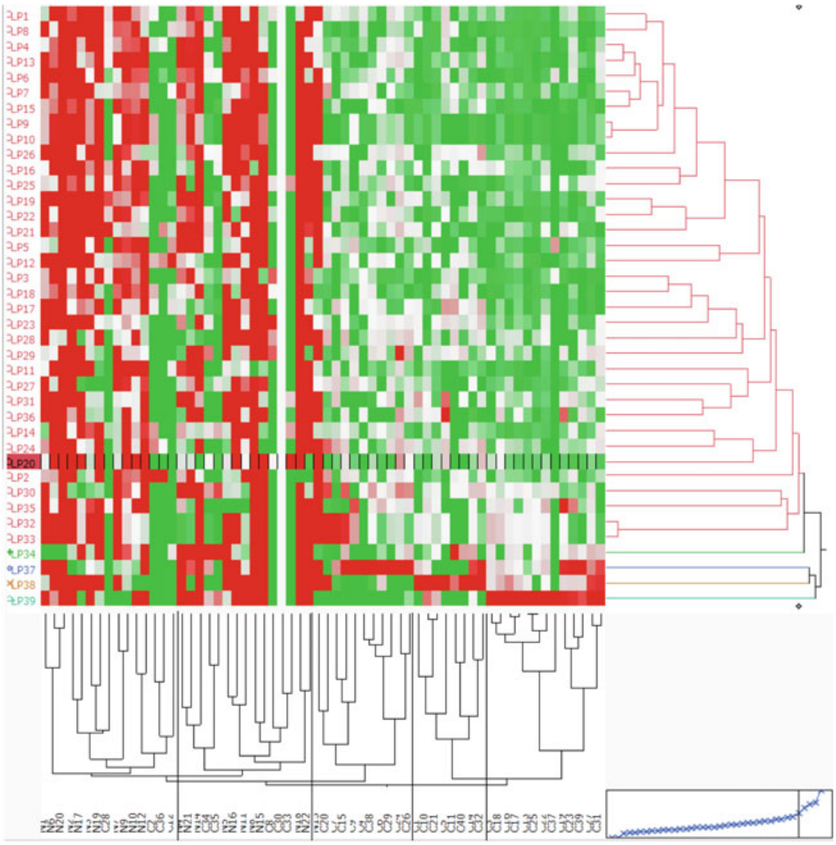


Fig. 3.9 Ward cluster of transposed signal data made by 39 LpDSs using 39 SMs

Figure 3.10 is the PCA result. Scatter plot shows the four outliers among 39 LPs. The factor loading plot shows that the 40 tumor cases are in the first and fourth quadrants, and the 22 normal cases are in the second and third quadrants.

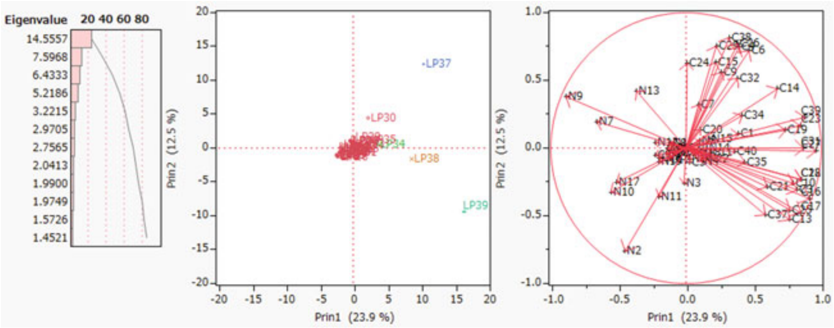


Fig. 3.10 PCA result of signal data made by LpDSs using 39 SMs

Figure 3.11 is the factor loading plot of transposed data of LpDSs signal data. However, several cancer patients overlapped with normal subjects in the left area. This plot shows the unclear feature compared with Fig. 3.6. Although we analyze the 39 SMs found by Revised LP-OLDF, we conclude the signal data made by 39 LpDSs is not better than the signal data by RipDSs.

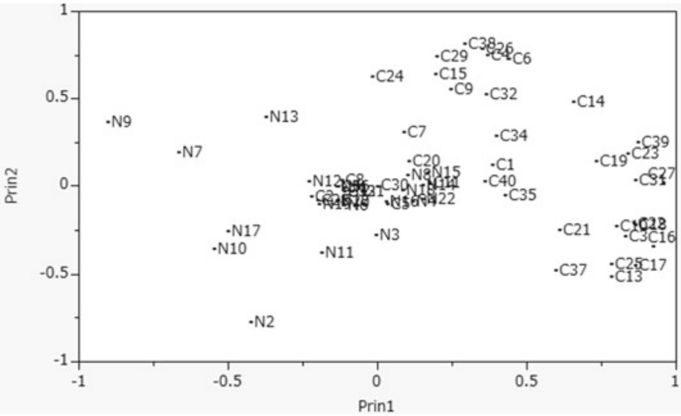


Fig. 3.11 Factor loading plot of transposed data of LpDSs

3.4.3 *Signal Data Made by 39 HsvmDSs Using 39 SMs Found by Revised LP-OLDF*

(1) Ward Cluster

Figure 3.12 is the result of HsvmDSs signal data using 39 SMs found by Revised LP-OLDF. Although Ward cluster cannot separate 39 SMs into two classes clearly, H-SVM can divide the signal data into two clusters. The upper green part is 22 normal cases, and lower white and red part is 40 tumor patients. We have designated five clusters here, and Ward cluster divides the tumor into four clusters. Five clusters can be identified by color and left symbol.

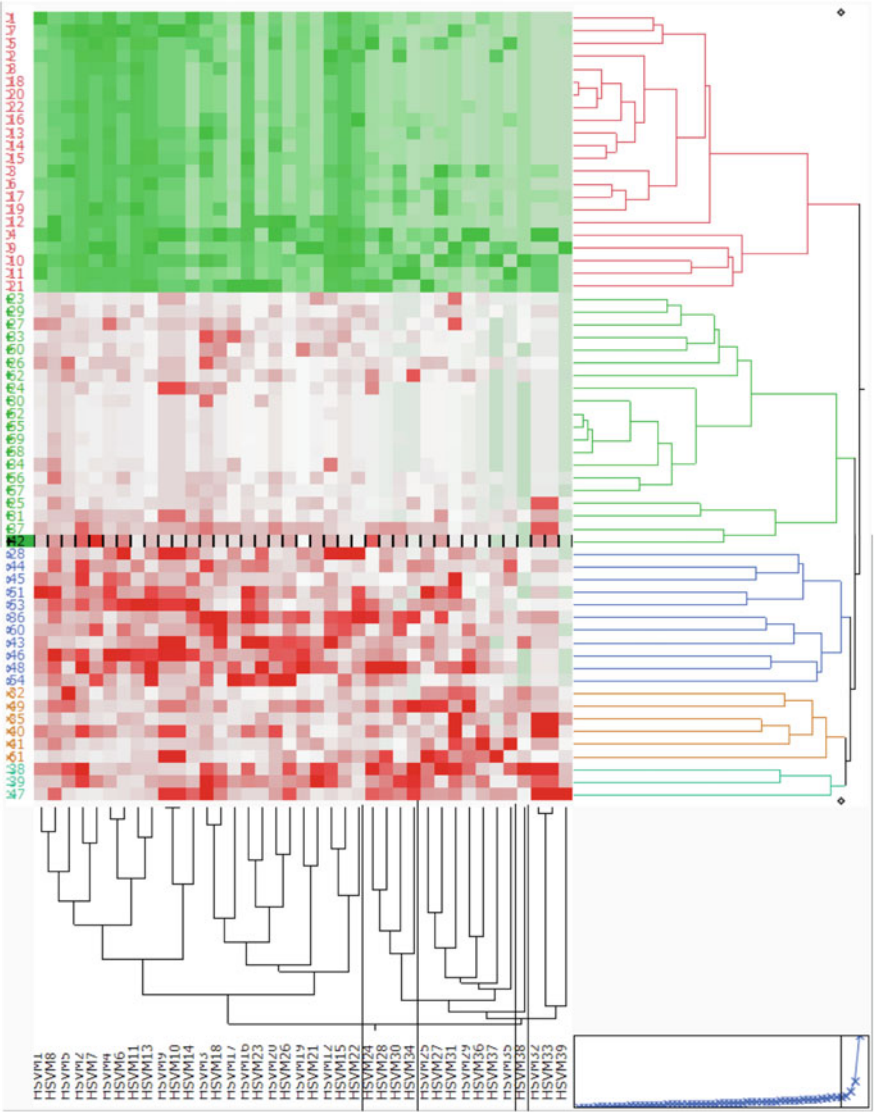


Fig. 3.12 Ward cluster result of signal data made by 39 HsvmDSs using SMs

If we divide the variable dendrogram into five clusters, those consist of 24, 4, 7, 1 and 3 HsvmDSs, respectively. Because one cluster consists of 24 HsvmDSs, the valance of five clusters is terrible. Because LP9 and LIP10 become the first cluster and its correlationis 1, those are compatible. The correlation of LP32 and LP33 is one, also.

(2) PCA Results

Figure 3.13 is the result of PCA with colors and marks. The scatter plot shows 20 normal cases locate on the negative Prin1 axis, and 42 tumor patients look like a fan that consists of four clusters in the first and fourth quadrants. The green cluster is a mild tumor. Blue patients are in the first and fourth quadrants. Pale green patients in the first quadrant are outliers and severe tumor. The factor loading plot shows the HSVM39 is a clear outlier. The fifth and sixth columns of Table 3.6 are the identification number (SN) and the Prin1 value.

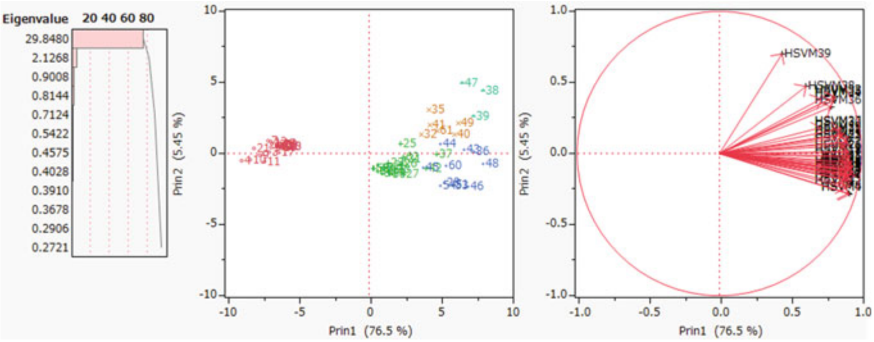


Fig. 3.13 Result of PCA with colors and marks

(3) Analysis of Transposed Signal data

Figure 3.14 is the Ward cluster of transposed signal data made by 39 HsvmDSs using 39 SMs. If we choose five clusters, the 28 HsvmDSs become one large cluster, and other four clusters those consist of eight HSVMs and three one cluster such as HSVM37, HSVM38, and HSVM39. Five clusters are slightly unbalance as same as 39 LpDSs.

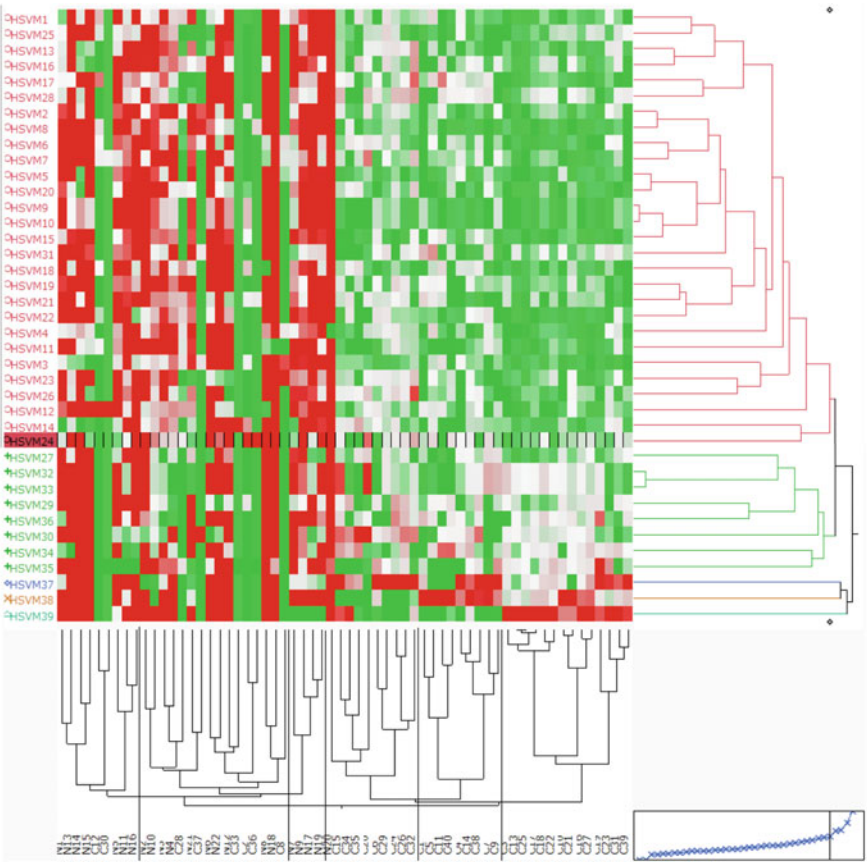


Fig. 3.14 Ward cluster of transposed signal data made by 39 HsvmDSs using 39 SMs

Figure 3.15 is the PCA result of transposed data. Scatter plot shows the four outliers among 39 HSVMs. Factor loading plot shows the 40 tumor cases are in the first and fourth quadrants, and the 22 normal cases are in the second and third quadrants.

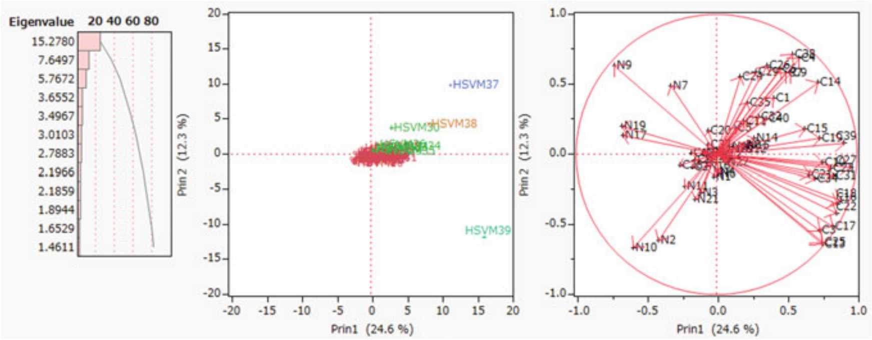


Fig. 3.15 PCA plots of transposed data

Figure 3.16 is the factor loading plot of transposed data of HsvmDSs signal data. However, several cancer patients overlapped with normal subjects in the center area including the origin. This plot shows the unclear feature compared with Fig. 3.6. Although we analyze the 39 SMs found by Revised LP-OLDF, we conclude the signal data made by 39 HsvmDSs is not better than the signal data by RipDSs.

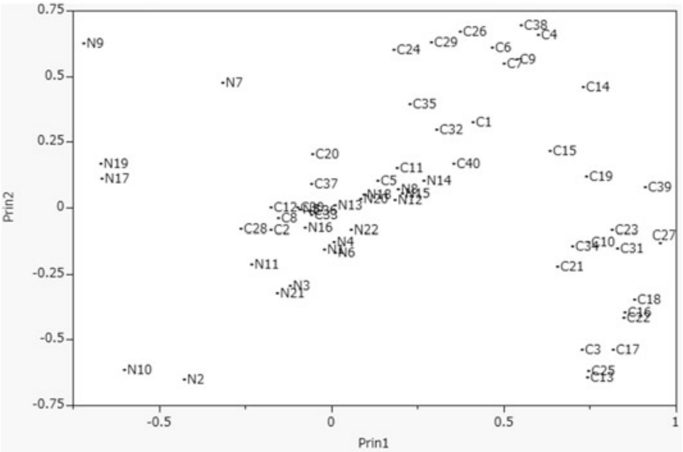


Fig. 3.16 Factor loading plot of transposed data of HsvmDSs

3.5 Analysis of Three Signal Data Using 56 SMs Found by RIP

Although standard statistical methods analyze all 56 SMs of RIP, the results are almost the same results explained in Chap. 2. Thus, we omit these results, also. In this Section, RIP, Revised LP-OLDF, and H-SVM discriminate all 56 SMs found by the RIP and made three signal data such as RipDSs, LpDSs, and HsvmDSs signal data. Ward cluster and PCA explain three results. Especially, Prin1 becomes proper malignancy indicators in addition to 56 DSs by RIP, Revised LP-OLDF, and H-SVM.

3.5.1 Signal Data Made by 56 RipDSs Using 56 SMs Found by RIP

(1) Ward Cluster

Figure 3.17 is the result of RipDSs signal data using 56 SMs discovered by RIP. Although Ward cluster cannot separate 56 SMs into two classes clearly, RIP can divide the signal data into two clusters. The upper green part is 22 normal cases, and lower white and red part is 40 tumor patients. We have designated five clusters here, and Ward cluster divides the tumor into four clusters. Five clusters can be identified by color and left symbol. However, two clusters consist of only two tumor patients such as (46, 48) and (51, 53). These clusters may be outliers.

If we divide the variable dendrogram into eight clusters, those consist of 5, 10, 4, 3, 17, 7, 2, and 9 RipDSs, respectively. It is necessary to categorize a large number of found SMs according to some criteria and to study their roles in the future. The result of this variable dendrogram is likely to be useful for future research. Because RIP26 and RIP28 become the first cluster, its correlation is 0.828 and the maximum value. Thus, there are no pairs with r = 1.

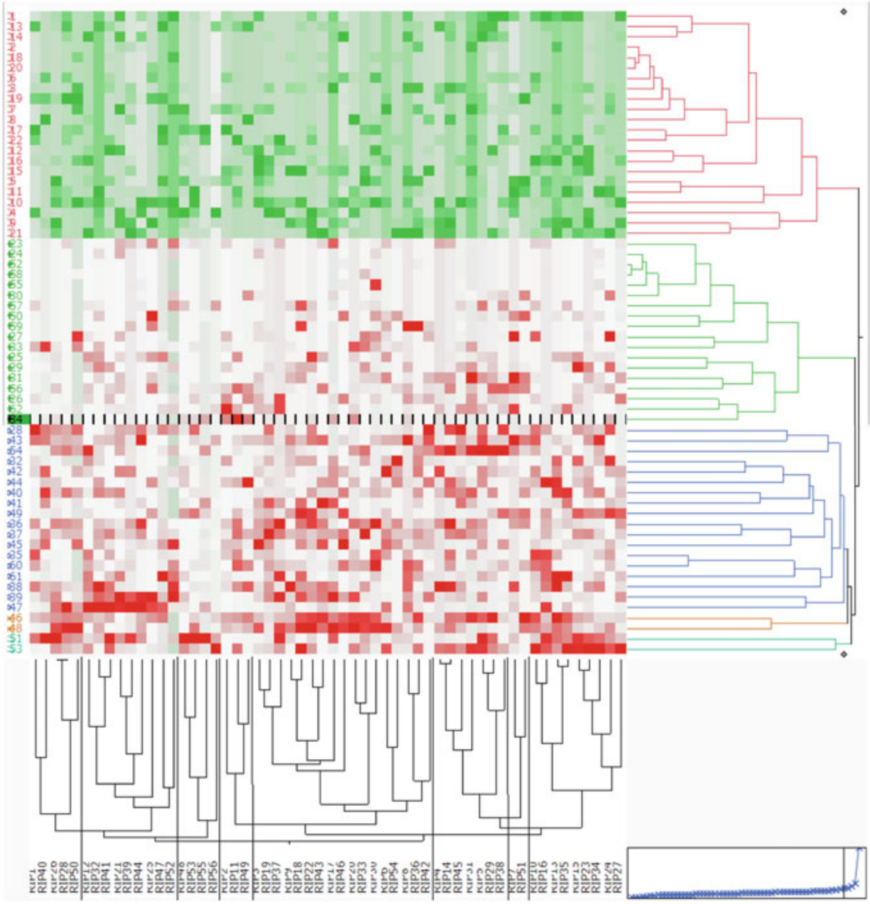


Fig. 3.17 Ward cluster result of signal data made by 56 RipDSs using SMs

(2) PCA Results

Figure 3.18 is the result of PCA with colors and marks. The scatter plot shows 22 normal cases locate on the negative Prin1 axis, and 40 tumor patients look like a fan that consists of four clusters in the first and fourth quadrants. The green cluster is a mild tumor. Blue patients are in the first and fourth quadrants. Two pale green patients are in the first quadrant, and two tawny patients are in the fourth quadrant. They are the outliers in Fig. 3.17. The factor loading plot shows the three RIPs are clear outlier and effect for C29 and C31.

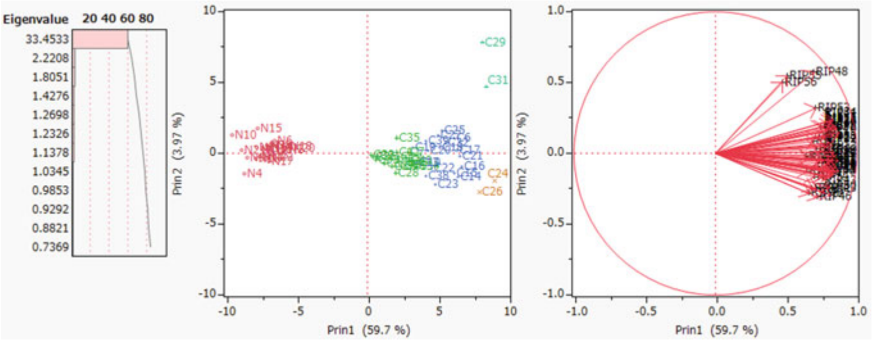


Fig. 3.18 Result of PCA with colors and marks

The first and second columns of Table 3.7 correspond to a malignancy indicator in Fig. 3.18. The numbers below three rows in the first column are the SN values for healthy and cancer patients. The second column is the value of Prin1, which is sorted in ascending order from a small value. Because the minimum value of normal subjects is −9.54 (N10) and the maximum value is −5.39 (N20), its range is [−9.542, −5.391]. On the other hand, the range of 40 cancer patients is [0.269, 8.979]. Thus, the SV separates two classes by large SV window (−5.391, 0.269) completely. Because the range of RipDS is [−9.542, 8.979] on Prin1, RatioSV of PCA is 30.557% (=(0.269 + 5.39)/(9.542 + 8.979) * 100). Because this is a comprehensive of RatioSV of 64 RIPs, it becomes a large value. That is, Prin1 can be considered as a malignancy indicator and illustrates an idea of malignancy indicator very well. A large window of 30.557% opens. Assuming that it is about 31%, it means that health and cancer subjects are placed in the remaining 69% of DS.

Table 3.7 Three malignancy indices

RatioSV	30.557	26.683		31.390	
Class	RIPScore	Class	LPScore	Class	HSVMScore
N10	**−9.542**	N4	−10.183	N4	−10.369
N21	−8.881	N10	−8.902	N9	−9.952
N4	−8.634	N9	−8.809	N10	−9.440
N9	−8.496	N21	−8.554	N21	−9.159
N1	−7.890	N11	−8.299	N7	−8.766
N5	−7.769	N17	−7.993	N11	−8.603
N15	−7.735	N13	−7.966	N3	−8.554
N11	−7.581	N5	−7.947	N5	−8.471
N16	−7.415	N7	−7.667	N17	−8.238
N12	−7.333	N12	−7.549	N1	−8.142
N13	−7.286	N1	−7.507	N16	−7.921
N7	−7.107	N14	−7.484	N13	−7.850
N3	−7.055	N16	−7.352	N19	−7.850
N17	−6.964	N3	−7.321	N14	−7.823
N22	−6.941	N15	−7.164	N15	−7.722
N19	−6.918	N19	−7.135	N12	−7.693
N14	−6.918	N6	−7.030	N6	−7.194
N8	−6.589	N22	−6.412	N8	−6.984
N6	−6.530	N8	−6.093	N22	−6.961
N2	−5.738	N2	−5.722	N2	−6.688
N18	−5.570	N18	−5.477	N18	−6.415
N20	**−5.391**	**N20**	**−5.377**	**N20**	**−6.318**
C30	**0.269**	**C36**	**−0.305**	**C30**	**−0.070**
C36	0.389	C30	−0.125	C33	−0.070
C33	0.599	C33	−0.088	C36	−0.045
C2	0.717	C2	0.275	C2	0.595
C8	1.177	C12	1.011	C37	0.652
C37	1.701	C37	1.090	C8	0.714
C28	1.982	C8	1.340	C12	1.439
C12	2.079	C1	1.390	C35	1.629
C35	2.089	C35	1.862	C1	1.949
C11	2.168	C3	2.404	C3	2.346
C1	2.211	C34	2.432	C28	2.404
C5	2.388	C28	2.902	C34	2.488
C3	2.724	C11	2.908	C11	2.627
C7	2.856	C7	2.916	C7	2.699

(continued)

Table 3.7 (continued)

RatioSV	30.557	26.683		31.390	
Class	RIPScore	Class	LPScore	Class	HSVMScore
C4	2.918	C39	3.002	C9	3.201
C40	2.941	C9	3.181	C39	3.391
C9	3.062	C13	3.193	C13	3.536
C19	3.202	C5	3.310	C40	3.602
C13	3.405	C4	3.343	C4	3.668
C34	3.489	C10	3.390	C5	3.797
C10	3.530	C19	3.469	C10	4.113
C39	4.109	C40	4.068	C19	4.594
C38	4.135	C27	4.430	C32	4.933
C20	4.282	C20	4.655	C20	5.433
C22	4.550	C6	5.056	C22	5.600
C23	4.824	C25	5.445	C23	5.671
C27	4.979	C38	5.451	C27	5.874
C18	5.111	C32	5.655	C38	5.910
C25	5.254	C22	5.735	C6	5.982
C32	5.503	C23	5.800	C25	6.047
C6	6.121	C15	6.194	C29	6.779
C15	6.136	C29	6.884	C15	7.049
C17	6.358	C21	6.936	C18	7.702
C14	6.430	C18	7.172	C21	7.811
C21	6.526	C14	7.565	C16	8.190
C16	6.645	C31	7.655	C31	8.244
C26	7.922	C16	7.724	C26	8.943
C29	8.149	C26	7.883	C14	9.017
C31	8.372	C17	7.911	C17	9.133
C24	**8.979**	C24	8.825	C24	9.535

(3) Analysis of Transposed Signal data

Figure 3.19 is the Ward cluster of transposed signal data made by 56 RipDSs using 56 SMs. If we choose five clusters, the 52 RipDSs become one large cluster, and other four one clusters consist of (RIP52), (RIP55), (RIP50), (RIP56). Five clusters are slightly unbalanced.

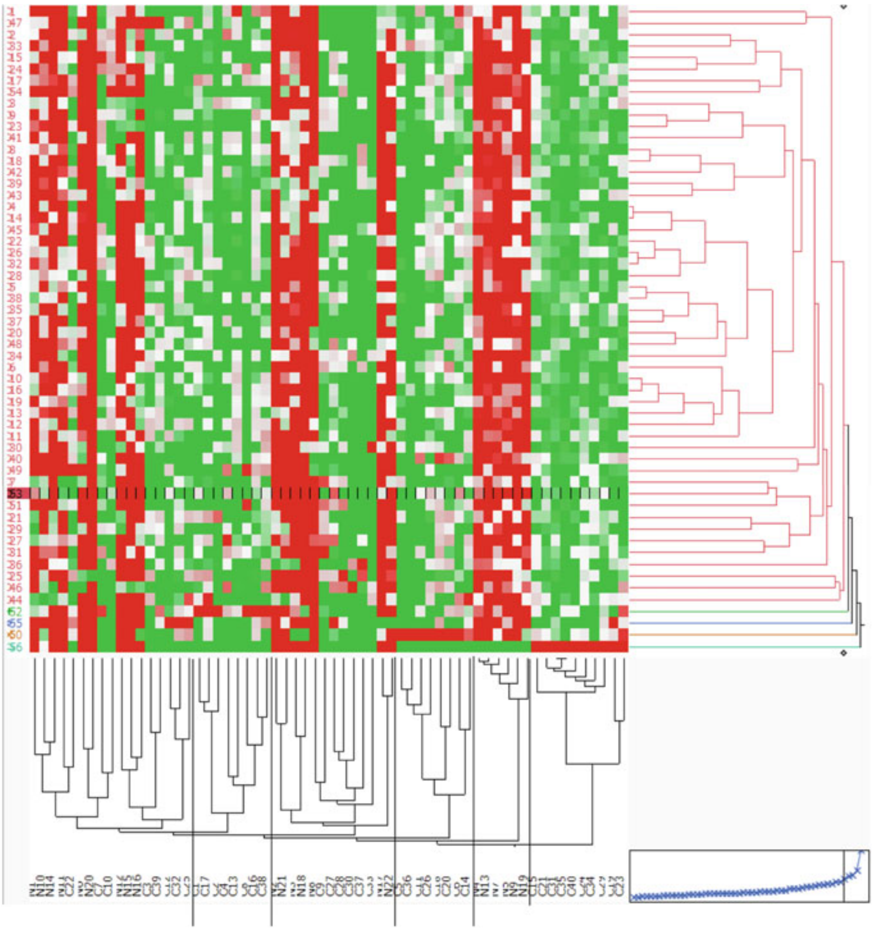


Fig. 3.19 Ward cluster of transposed signal data made by 56 RipDSs using 56 SMs

Figure 3.20 is the PCA result of transposed data. Scatter plot shows the four outliers among 56 RIPs. The factor loading plot shows that the 40 tumor cases are almost in the first quadrant, and the 22 normal cases are almost in the third quadrant.

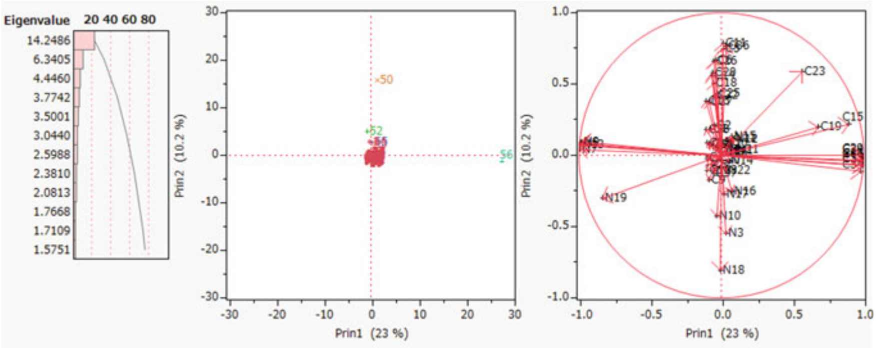


Fig. 3.20 PCA plots of Transposed data

Table 3.8 shows factor loading plots. The row shows 22 normal subjects from N1 to N22 and 40 cancer patients from C1 to C40. Prin1 and Prin2 are the first and second factor loading values which are correlations of 62 patients with Prin1 and Prin2. “Mark” is a marker corresponding to five groups of 62 patients corresponding to Fig. 3.21.

Table 3.8 Factor loading plot

Row	Prin1	Prin2	Mark
N1	0.097	0.031	0
N2	0.048	0.022	0
N3	0.032	−0.549	−2
N4	−0.994	0.070	−1
N5	−0.989	0.085	−1
N6	0.060	0.019	0
N7	−0.995	0.040	−1
N8	0.042	0.017	0
N9	−0.981	0.097	−1
N10	−0.032	−0.428	−2
N11	0.071	0.120	0
N12	0.086	0.115	0
N13	−0.994	0.076	−1
N14	0.054	−0.037	0
N15	0.076	0.132	0
N16	0.073	−0.250	−2
N17	0.018	−0.272	−2

(continued)

Table 3.8 (continued)

Row	Prin1	Prin2	Mark
N18	−0.007	−0.804	−2
N19	−0.836	−0.303	
N20	0.016	0.078	0
N21	0.091	0.041	0
N22	0.032	−0.108	0
C1	−0.073	0.188	0
C2	−0.048	0.205	0
C3	−0.070	−0.065	0
C4	−0.103	−0.027	0
C5	0.010	0.749	2
C6	−0.041	0.669	2
C7	−0.098	−0.011	0
C8	−0.052	0.068	0
C9	−0.090	−0.172	0
C10	−0.101	−0.107	0
C11	0.014	0.789	2
C12	−0.028	0.076	0
C13	−0.089	0.078	0
C14	−0.068	0.559	2
C15	0.892	0.217	
C16	−0.107	0.380	2
C17	−0.085	0.371	2
C18	−0.055	0.507	2
C19	0.677	0.197	
C20	−0.062	0.581	2
C21	0.995	−0.034	1
C22	−0.100	0.085	0
C23	0.570	0.588	
C24	0.991	−0.003	1
C25	−0.034	0.437	2
C26	−0.052	0.657	2
C27	−0.095	−0.070	0
C28	−0.053	−0.104	0

(continued)

Table 3.8 (continued)

Row	Prin1	Prin2	Mark
C29	0.981	0.002	1
C30	−0.034	−0.032	0
C31	0.995	−0.072	1
C32	−0.043	0.421	2
C33	−0.042	0.040	0
C34	0.982	−0.111	1
C35	0.994	−0.067	1
C36	0.036	0.766	2
C37	−0.053	−0.128	0
C38	−0.106	0.182	0
C39	−0.082	0.071	0
C40	0.991	−0.041	1

Mark "1" (Prin1> 0.97, −0.2< Prin2< 0) corresponds to the right rectangle containing seven cancer patients such as c21 c24, c29, c31, c34, c35, and c40. Those correspond to RIP56 in scatter plot. RIP56 is considered to be the outliers associated with these seven cancer patients. On the other hand, "−1" (Prin1< −0.97, |Prin2|< 0.2) corresponds to a left rectangle containing five normal subjects such as N4, N5, N7, N9, and N13. These five patients are probably patients to contrast with seven cancer patients corresponding to RIP56 in scatter plot. Mark "2" (|Prin1| < 0.2, 0.35 <Prin2) corresponds to the upper rectangle containing 12 cancer patients such as c5, c6, c11, c14, c16–c18, c20, c25, c26, c32, and c36. These 12 cancer patients correspond to RIP 50 and RIP52 in the scatter plot. Both RIPs are considered to be outliers in these 12 cancer patients. On the other hand, "−2" (0 < Prin1 < 0.2, Prin2 < − 0.2) corresponds to a bottom rectangle containing five normal subjects such as N3, N10, N16–N18. They are probably patients to contrast with 12 cancer patients corresponding to RIP50 and RIP52 in scatter plot. Mark "0" (|Prin1| < 0.2, |Prin2| < 0.2) corresponds to the mid-rectangle containing 11 normal subjects and 17 cancer patients. Although Prin2 is less variance than Prin1, Prin2 explains more patients than Prin1. Thus, we expect medical specialists will evaluate these results because our results suggest there are two different types of outliers. Although Golub et al. found new subclasses of cancer by SOM, they changed the number of clusters from two and looked for the proper number of clusters. Because they did not know that microarray is LSD, they seem to have made a great effort. If we do not know the proper number of clusters, we must try several trials. We think DSs obtained by all SMs include much information. Moreover, our procedure is straightforward and offers precise results. Thus, we expect medical specialists to challenge our approaches in addition to their approach.

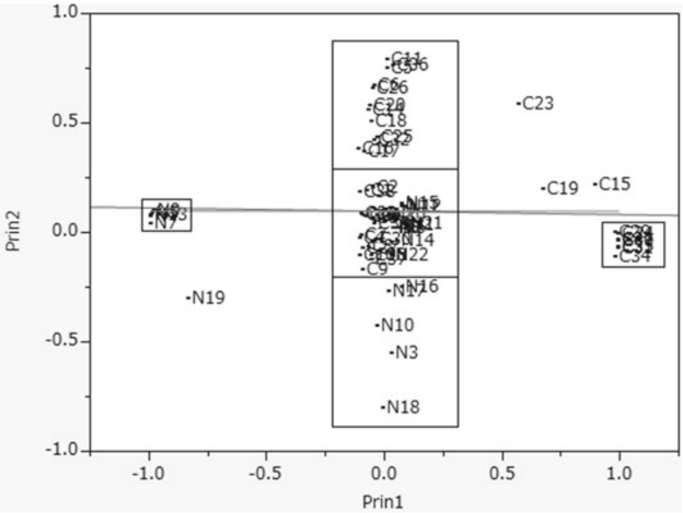


Fig. 3.21 Five groups of 62 patients

3.5.2 *Signal Data Made by 56 LpDSs Using 56 SMs Found by RIP*

(1) Ward Cluster

Figure 3.22 is the result of LpDSs signal data using 56 SMs found by RIP. Although Ward cluster cannot separate 56 SMs into two classes clearly, Revised LP-OLDF can separate the signal data into two clusters. The upper green part is 22 normal cases, and lower white and red part is 40 tumor patients. We have designated five clusters here, and Ward cluster divides the tumor into four clusters. The color and left symbol identify five clusters. The fourth and fifth clusters consist of six and two LpDSs may be the different outliers.

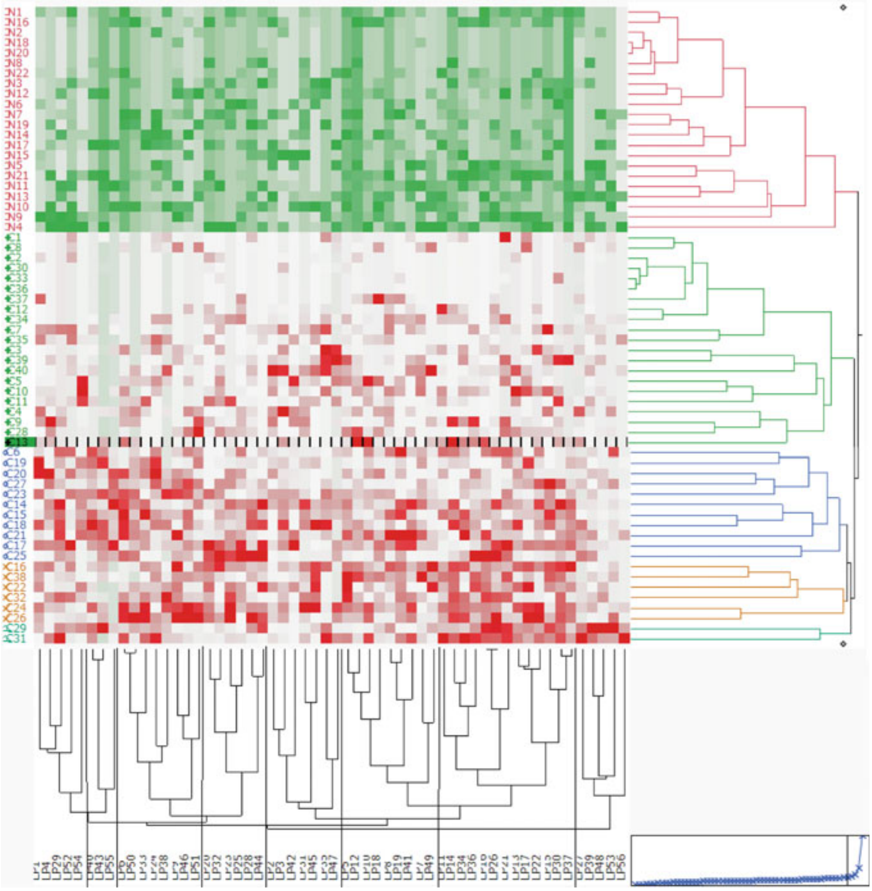


Fig. 3.22 Ward cluster result of signal data made by 56 LpDs using 56 SMs

If we divide the variable dendrogram into eight clusters, those consist of 5, 3, 8, 6, 7, 9, 13, and 5 LpDSs, respectively. These segmentations will be helpful for the researchers to divide 56 pairs of genes into compatible groups. LP16 and LIP26 become the first cluster and its correlation 0.8542. Thus, there is no correlation with r = 1.

(2) PCA Results

Figure 3.23 is the result of subjects with colors and marks. The scatter plot shows 22 normal cases locate on the negative Prin1 axis, and 40 tumor patients look like a fan that consists of four clusters in the first and fourth quadrants. The green patients are in the first and fourth quadrants and mild tumor. Tawny and blue patients overlap and are severe tumor patients. The factor loading plot shows the LP56 is a clear outlier. The third and fourth columns of Table 3.7 are the identification number (Class) and the Prin1 value.

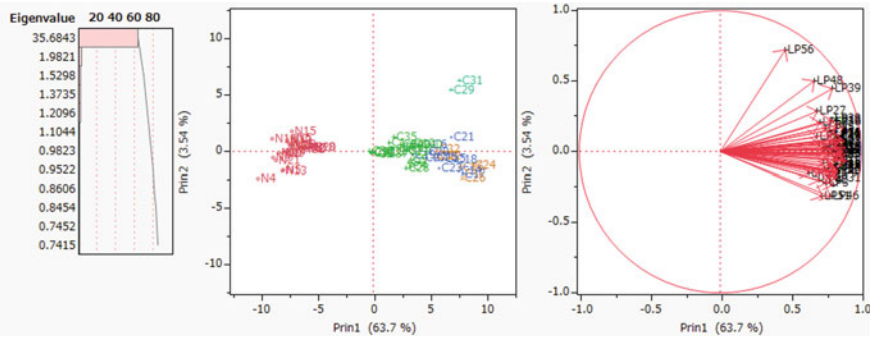


Fig. 3.23 Result of subjects with colors and marks

(3) Analysis of Transposed Signal data

Figure 3.24 is the Ward cluster of transposed signal data made by 56 LpDSs using 56 SMs. If we choose six clusters, the 44 LpDSs become one large cluster, and other five clusters consist of eight LPs and four one cluster such as LP69, LP63, LP11, and LP12. It is difficult for us to judge whether six clusters are unbalancing or other four clusters are clear outliers.

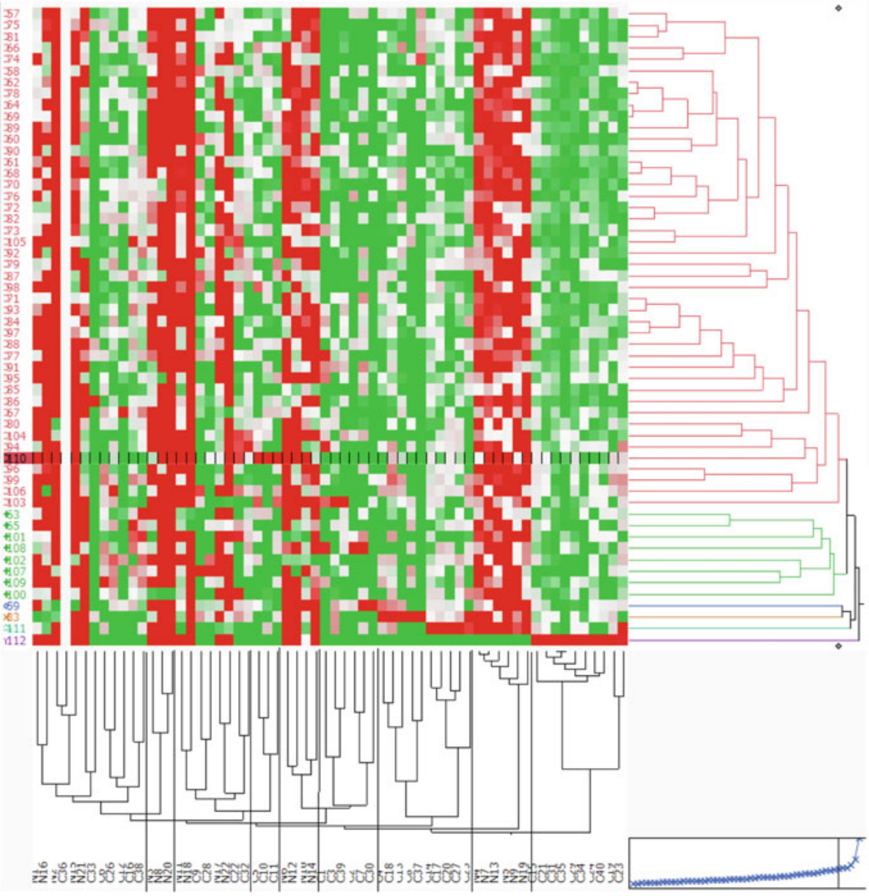


Fig. 3.24 Ward cluster of transposed signal data made by 56 LpDSs using 56 SMs

Figure 3.25 is the PCA result of transposed data. Scatter plot shows the four outliers among 56 LPs. The 40 tumor cases are almost in the first quadrant, and the 22 normal cases are in the third quadrant.

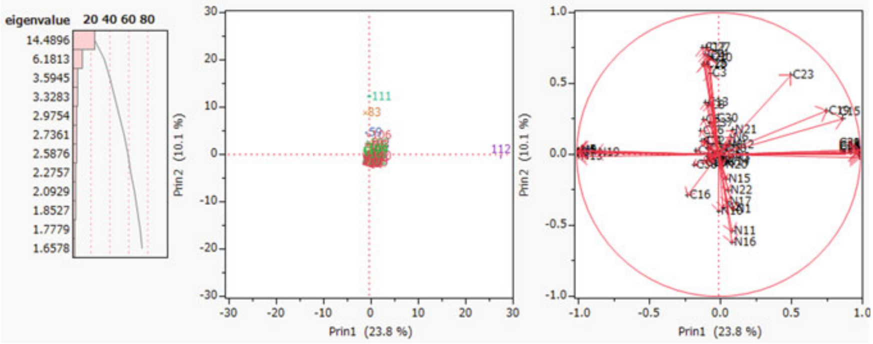


Fig. 3.25 PCA plots of transposed data

Table 3.9 shows the factor loading plot. The "Row" shows 22 normal subjects from N1 to N22 and 40 cancer patients from C1 to C40. "Prin1 and Prin2" are the first and second factor loading values which are correlations of 62 patients with Prin1 and Prin2. "Mark" is a marker corresponding to five groups of 62 patients corresponding to Fig. 3.26.

Table 3.9 Factor loading plot

Row	Prin1	Prin2	Mark
N1	0.104	−0.388	−2
N2	0.030	−0.372	−2
N3	0.103	−0.017	0
N4	−0.995	0.024	−1
N5	−0.988	0.031	−1
N6	0.084	0.122	0
N7	−0.994	0.010	−1
N8	0.050	0.065	0
N9	−0.978	0.027	−1
N10	0.001	−0.397	−2
N11	0.094	−0.543	−2
N12	0.078	0.069	0
N13	−0.990	−0.014	−1
N14	0.044	−0.052	0
N15	0.054	−0.173	−2
N16	0.094	−0.622	−2
N17	0.061	−0.335	−2

(continued)

Table 3.9 (continued)

Row	Prin1	Prin2	Mark
N18	0.028	0.043	0
N19	−0.864	0.013	−1
N20	0.031	−0.076	0
N21	0.097	0.170	0
N22	0.068	−0.250	−2
C1	−0.061	−0.021	0
C2	−0.066	−0.004	0
C3	−0.060	0.573	2
C4	−0.074	0.693	2
C5	−0.071	0.038	0
C6	−0.106	0.247	0
C7	−0.108	0.095	0
C8	−0.074	0.347	2
C9	−0.098	0.096	0
C10	−0.078	−0.014	0
C11	−0.081	0.047	0
C12	−0.058	−0.036	0
C13	−0.091	0.365	2
C14	−0.093	0.710	2
C15	0.875	0.252	
C16	−0.219	−0.287	
C17	−0.114	0.753	2
C18	−0.104	0.634	2
C19	0.762	0.310	
C20	−0.068	0.682	2
C21	0.995	0.030	1
C22	−0.157	0.027	0
C23	0.509	0.560	
C24	0.984	0.016	1
C25	−0.100	0.641	2
C26	−0.125	0.169	0
C27	−0.080	0.752	2
C28	−0.104	−0.061	0

(continued)

Table 3.9 (continued)

Row	Prin1	Prin2	Mark
C29	0.990	0.032	1
C30	−0.024	0.251	0
C31	0.996	−0.004	1
C32	−0.123	0.094	0
C33	−0.052	−0.052	0
C34	0.990	0.006	1
C35	0.996	−0.021	1
C36	0.000	0.000	0
C37	−0.063	0.218	0
C38	−0.176	−0.072	0
C39	−0.078	−0.048	0
C40	0.983	0.013	

Mark “1” (Prin1> 0.98, |Prin2|< 0.03) corresponds to the right rectangle containing six cancer patients such as c21 c24, c29, c31, c34, and c35. They correspond to LP56 in scatter plot. LP56 is considered to be outliers in these six cancer patients. On the other hand, “−1” (Prin1< −0.97, |Prin2|< 0.03) corresponds to a left rectangle containing six normal subjects such as N4, N5, N7, N9, N13, and N19. They are probably patients to contrast with six cancer patients corresponding to LP56 in scatter plot. Mark “2” (|Prin1| < 0.2, Prin2 > 0.3) corresponds to the upper rectangle containing 12 cancer patients such as c3, c4, c8, c13, c14, c17, c18, c20, c25, c27, c30, and c37. Those correspond to LP55 and Lp27 in scatter plot. Both LpDSs are considered to be outliers in these 12 cancer patients. On the other hand, “−2” (|Prin1| < 0.2, Prin2 < − 0.2) corresponds to a bottom rectangle containing eight normal subjects such as N1, N2, N10, N11, N15, N16, N17, and N22. They are probably patients to contrast with 12 cancer patients corresponding to both LpDSs in scatter plot. Mark “0” (|Prin1| < 0.2, |Prin2| < 0.6) corresponds to the mid-rectangle containing eight normal subjects and 19 cancer patients.

Thus, we expect medical specialists will evaluate these results because our results suggest there are two different types of outliers very easy. Now, it is difficult for us to explain the role of c15, c19, c23, and c16.

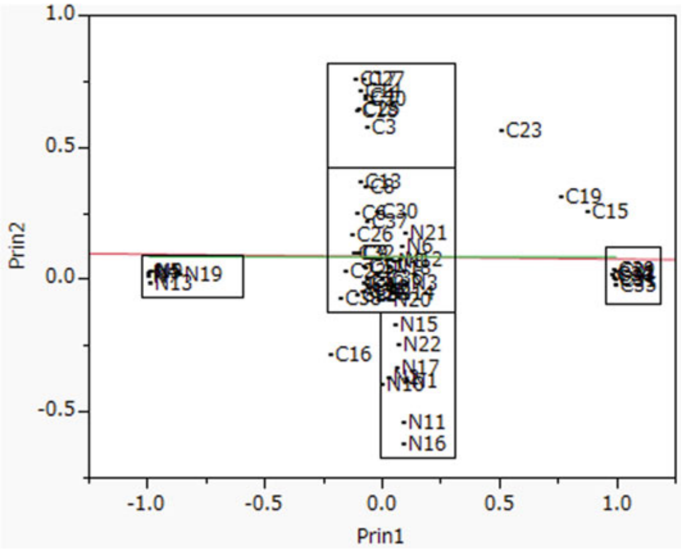


Fig. 3.26 Five groups of 62 patients

3.5.3 *Signal Data Made by 56 HsvmDSs Using 56 SMs Found by RIP*

(1) Ward Cluster

Figure 3.27 is the result of HsvmDSs signal data using 56 SMs discovered by RIP. Although Ward cluster cannot separate 56 SMs into two classes clearly, H-SVM can divide the signal data into two clusters. The upper green part is 22 normal cases, and lower white and red part is 40 tumor patients. We have designated five clusters here, and Ward cluster divides the tumor into four clusters. The color and symbol separate five clusters.

If we divide the variable dendrogram into nine clusters, those consist of 11, 12, 13, 2, 1, 8, 1, 7, and 1 HsvmDSs, respectively. Because HSVM22 and HSVM23 become the first cluster, its correlation is r = 0.9035.

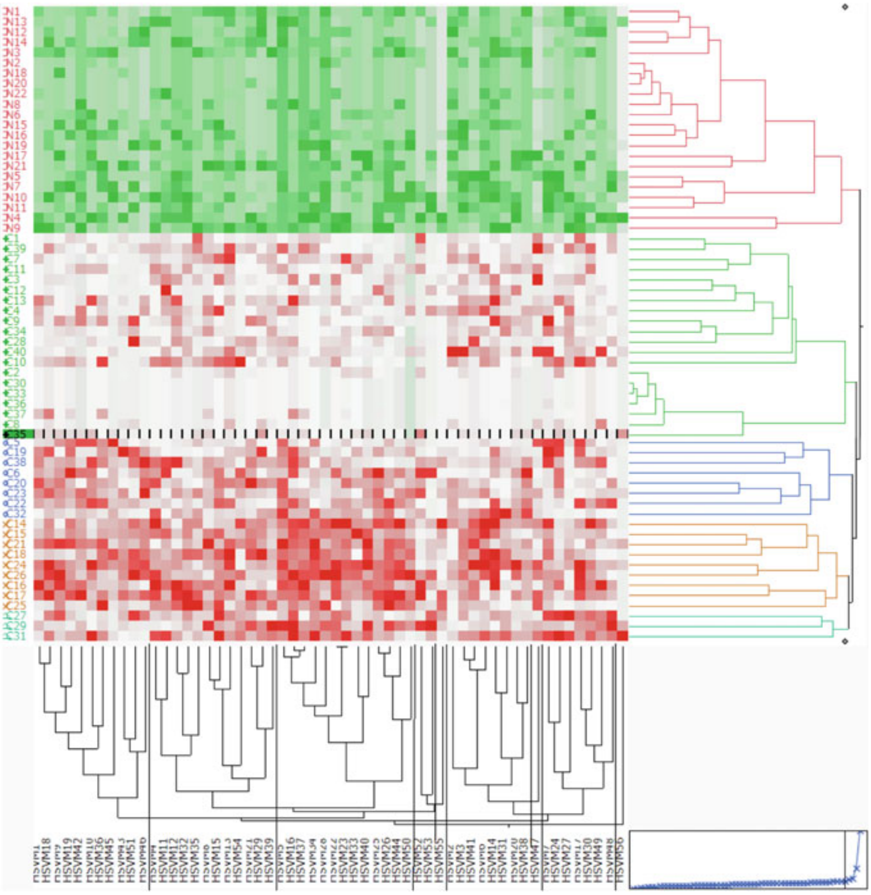


Fig. 3.27 Ward cluster result of signal data made by 56 HsvmDSs using SMs

(2) PCA Results

Figure 3.28 is the result of subjects with colors and marks. The scatter plot shows 22 normal cases locate on the negative Prin1 axis, and 40 tumor patients look like a fan that consists of four clusters in the first quadrant. The green cluster is a mild tumor. Blue patients are in the first and fourth quadrants. Pale green patients in the first quadrant are the severe tumor. The factor loading plot shows the HSVM56 is a clear outlier. The fifth and sixth columns of Table 3.7 are the identification number (Class) and the Prin1 value.

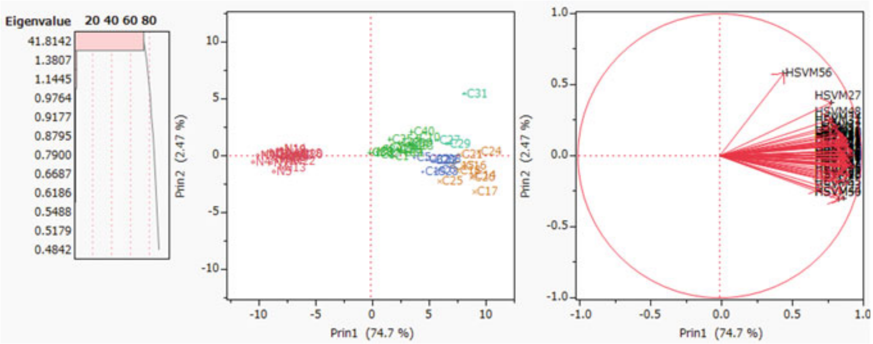


Fig. 3.28 Result of PCA with colors and marks

(3) Analysis of Transposed Signal data

Figure 3.29 is the Ward cluster of transposed signal data made by 56 HsvmDSs using 56 SMs. If we choose six clusters, the 47 HSVMs are one large cluster. Next, after five HSVMs become one cluster, it becomes one cluster with the first 45 HSVMs. Finally, the four HSVMs of HSVM15, HSVM18, HSVM20, and HSVM 24 are merged sequentially into one cluster. Regardless of the Ward method, it has the same characteristics as the nearest neighbor method.

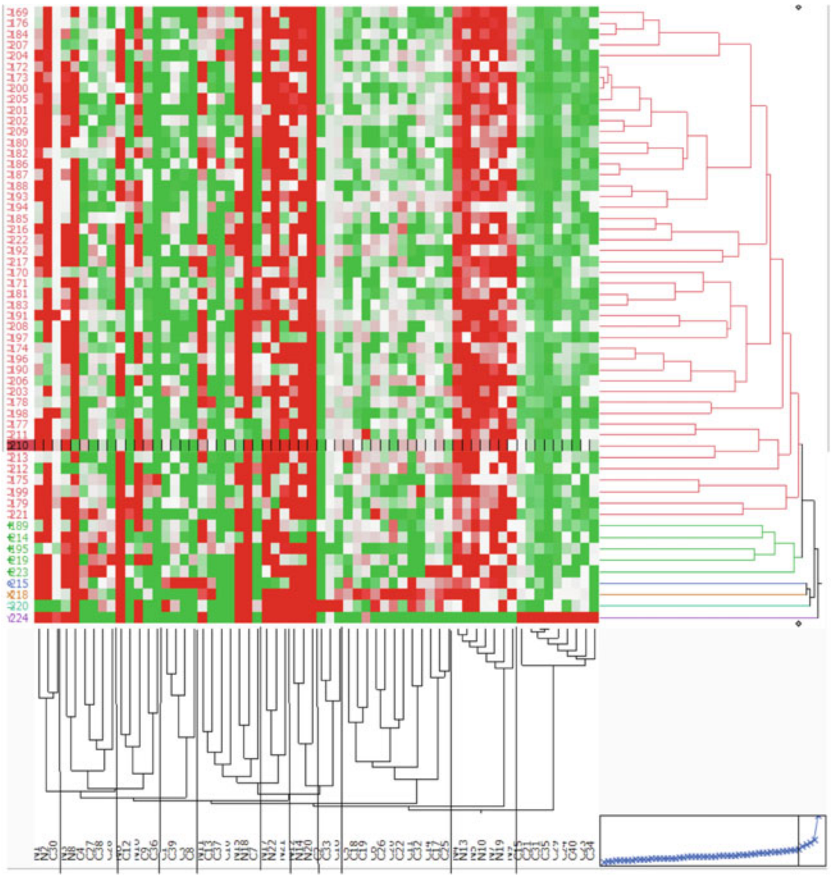


Fig. 3.29 Ward cluster of transposed signal data made by 56 HsvmDSs using 56 SMs

Figure 3.30 is the PCA result of transposed data. Scatter plot shows the four outliers among 56 HSVMs. The factor loading plot shows that the 40 tumor cases are in the first and fourth quadrants, and the 22 normal cases are in the second and third quadrants.

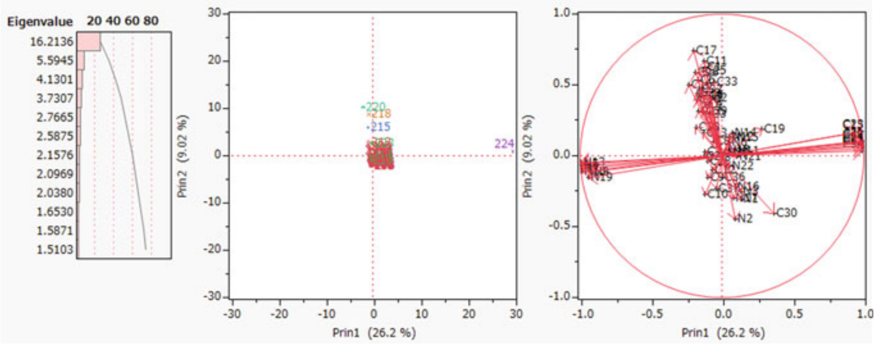


Fig. 3.30 PCA plots of transposed data

Table 3.10 shows the factor loading plot. The "row" shows 22 normal subjects from N1 to N22 and 40 cancer patients from C1 to C40. "Prin1 and Prin2" are the first- and second-factor loading values which are correlations of 62 patients with Prin1 and Prin2. "Mark" is a marker corresponding to five groups of 62 patients corresponding to Fig. 3.31.

Table 3.10 Factor loading plot

Row	Prin1	Prin2	Mark
N1	0.133	−0.309	−2
N2	0.093	−0.444	−2
N3	0.135	−0.251	−2
N4	−0.994	−0.075	−1
N5	−0.993	−0.057	−1
N6	0.055	0.094	0
N7	−0.987	−0.092	−1
N8	0.058	0.025	0
N9	−0.923	−0.113	−1
N10	−0.986	−0.111	−1
N11	0.092	0.033	0
N12	0.056	0.135	0
N13	−0.996	−0.048	−1
N14	0.071	0.159	0
N15	0.085	0.125	0
N16	0.096	−0.221	−2
N17	0.081	−0.301	−2

(continued)

Table 3.10 (continued)

Row	Prin1	Prin2	Mark
N18	0.026	0.040	0
N19	−0.937	−0.153	−1
N20	0.022	0.018	0
N21	0.106	−0.006	0
N22	0.056	−0.072	0
C1	−0.118	0.626	2
C2	−0.083	0.407	2
C3	−0.088	0.292	2
C4	−0.114	0.359	2
C5	−0.109	0.428	2
C6	−0.169	0.533	2
C7	−0.119	0.028	0
C8	−0.058	0.100	0
C9	−0.102	−0.151	−2
C10	−0.119	−0.274	−2
C11	−0.125	0.665	2
C12	−0.067	−0.099	0
C13	−0.128	0.157	0
C14	−0.163	0.418	2
C15	0.938	0.164	1
C16	−0.234	0.500	2
C17	−0.199	0.739	2
C18	−0.170	0.395	2
C19	0.283	0.187	0
C20	−0.176	0.440	2
C21	0.995	0.074	1
C22	−0.164	0.311	2
C23	0.980	0.172	1
C24	0.991	0.090	1
C25	−0.132	0.599	2
C26	−0.184	0.590	2
C27	−0.125	−0.006	0
C28	−0.102	−0.038	0
C29	0.990	0.105	1
C30	0.365	−0.407	−2
C31	0.997	0.053	1
C32	−0.157	0.467	2

(continued)

Table 3.10 (continued)

Row	Prin1	Prin2	Mark
C33	−0.056	0.520	2
C34	0.988	0.036	1
C35	0.995	0.069	1
C36	−0.001	−0.150	−2
C37	−0.039	−0.232	−2
C38	−0.183	0.203	0
C39	−0.125	0.312	2
C40	0.986	0.099	1

Mark "1" (Prin1> 0.92, 0< Prin2< 0.02) corresponds to the right rectangle containing nine cancer patients such as c15, c21, c23 c24, c29, c31, c34, c35, and c40. They correspond to HSVM224 (HsvmDS224) in Fig. 3.30. It affects these nine cancer patients. On the other hand, "−1" (Prin1< −0.92, −0.12< Prin2 < −0.04) corresponds to a left rectangle containing seven normal subjects such as N4, N5, N7, N9, N10, N13, and N19. They are probably patients to contrast with nine cancer patients corresponding to HsvmDS224 in Fig. 3.30. Mark "2" (−0.25 < Prin1 < −0.05, Prin2 > 0.3) corresponds to the upper rectangle containing 18 cancer patients such as c1–c6, c11, c14, c16-c18, c20, c22, c25, c26, c32, c33 and c39. HSVM220, HSVM218 and HSVM215 may affect these 18 patients. On the other hand, "−2" (|Prin1| < 0.36, −0.45 < Prin2 < −0.1) corresponds to a bottom rectangle containing six normal subjects such as N1–N3, N16, N17 and N22 in addition to seven cancer patients such as c9, c10, c12, c30, c36, and c37. They are probably patients to contrast with 18 cancer patients corresponding to three HsvmDSs in scatter plot. Mark "0" (|Prin1| < 0.3, |Prin2| < 0.2) corresponds to the mid-rectangle containing nine normal subjects and seven cancer patients. Thus, we expect a medical expert will evaluate these results because our results suggest there are two different types of outliers.

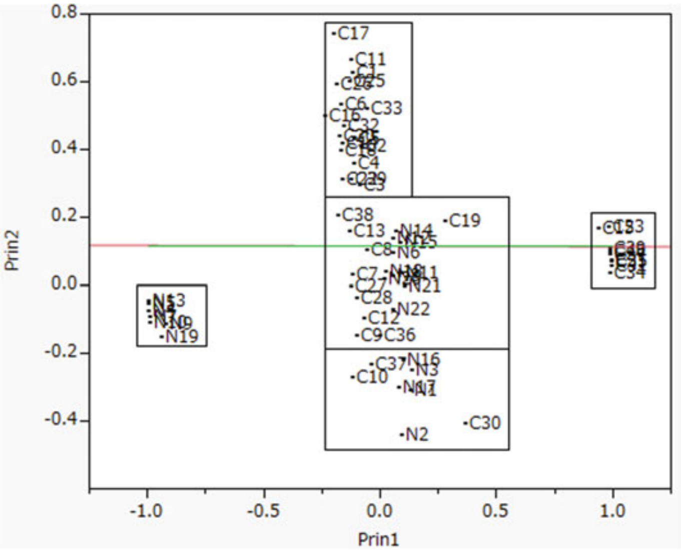


Fig. 3.31 Five groups of 62 patients

3.6 Conclusion

In this chapter, we try innovations beyond Chaps. 1 and 2.

Early 2016, we found Revised LP-OLDF of LINGO Program3 could decompose six microarrays into many SMs as same as RIP, and three SVMs could not decompose the microarrays into SMs. However, we did not survey the cancer gene analysis and diagnosis of Revised LP-OLDF until now for the following reasons:

(1) Revised LP-OLDF is weak for Problem1 and tends to collect several cases on the discriminant hyperplane. We do not trust the NM of Revised LP-OLDF as same as other discriminant functions.
(2) Now, we find Revised LP-OLDF can discover many SMs because LP finds one of the endpoints (vertexes) made by at most n constraints of feasible region of microarrays. Thus, we evaluate 39 SMs obtained by Revised LP-OLDF as same as 56 SMs obtained by a RIP in this Chapter.

Next, we propose how to choose the proper numbers of SM and obtain different combinations of SMs found by the RIP and Revised LP-OLDF.

Third, because two classes are separable in each SM, we misunderstand genes included in each SM is the cancer genes and signal. However, statistical methods could not find the linear separable facts at all, except for three LDFs such as RIP, Revised LP-OLDF, and H-SVM. Thus, we reconsider signal data made by three LDFs are signal.

Fourth, we develop how to analyze all signal data by the PCA and the hierarchical cluster analysis and propose the cancer gene diagnosis.

Although the medical specialists must evaluate these results, our research is a milestone to open the cancer gene diagnosis using microarrays. The statistical discriminant functions are useless for cancer gene analysis. On the other hand, the proper discriminant functions easily open the new frontier for a human being.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad Sci USA 96(12):6745–6750

Schrage L (2006) Optimization Modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2015) Simple Structure of Alon et al. et al. microarray data. Res Gate (10):1–34

Shinmura S (2017) From cancer gene analysis to cancer gene diagnosis. Amazon

Chapter 4
Further Examinations of SMs—Defect of Revised LP-OLDF and Correlations of Genes

Abstract In this chapter, we analyze Alon's microarray in 2018 and obtain two SMs from the RIP and Revised LP-OLDF. In Sect. 4.2, RIP separates the microarray into a union of 62 SMs (1,968 genes). Six MP-based LDFs find this subspace is LSD and a noise subspace (32 genes) is not LSD. In Sect. 4.3, Revised LP-OLDF separates the microarray into a union of 32 SMs (1,005 genes) and a noise subspace (995 genes). Six MP-based LDFs find both subspaces are LSD. This fact suggests us that a noise subspace includes other SMs in it. We find Revised LP-OLDF cannot find all SMs from the microarray correctly. We guess Problem1 causes the defect of Revised LP-OLDF. Namely, Revised LP-OLDF cannot find other SMs from noise subspace. Section 4.4 analyzes 62 SMs found by the RIP and evaluates 62 SMs by RatioSV and NM. Moreover, the 1,891 correlations of 62 RIP discriminant scores (RipDSs) are computed. At first, we consider each gene set included in SM is cancer genes and a signal subspace. However, standard statistical methods cannot show the linear separable facts. Thus, we conclude that the gene sets included in all SMs are not signals. We recognize the data made by RipDSs is signal data. Two signal data of SM13 with maximum RatioSV and SM62 with minimum RatioSV are validated. Section 4.5 analyzes two signal data made by RipDSs and HsvmDSs obtained by 62 SMs of the RIP. The results are almost the same in Chaps. 2 and 3. However, these findings can open a new field of cancer gene diagnosis only after verification of the subjects used in the study of Alon et al. (Proc Natl Acad Sci USA, 96(1.1): 6745–6750 1999). Section 4.6 explains the reason why standard statistical methods could not find the linear separable facts. Section 4.9 is the conclusion.

Keywords Alon's microarray · Small matryoshka (SM) · Defect of SMs by Revised LP-OLDF · RIP discriminant scores (RipDSs) · Signal data · T-test of two classes · Correlations of genes · Structure of cancer genes

4.1 Introduction

Chapter 1 outlined the new theory of discriminant analysis after R. Fisher (the Theory) and explained the first success of cancer gene analysis (Shinmura 2016). Also, we explained why Revised IP-OLDF (RIP) and Revised LP-OLDF solved

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_4

unresolved cancer gene analysis (Problem5). We explain the reason why only the linear programming (LP) and the integer programming (IP) could decompose microarrays into many SMs (signal). However, the H-SVM could not find Small Matryoshkas (SMs) because the quadratic programming (QP) find only one optimal solution for the whole region. Moreover, we explain the reason why many researchers could not solve Problem5 because other statistical discriminant functions such as Fisher's LDF (Fisher 1956) and the quadratic discriminant function (QDF) were useless for cancer gene analysis at all from 1970. Chapter 2 outlined the cancer gene diagnosis using all SMs of six microarrays found by the RIP in 2016. Chapter 3 outlined that RIP and Revised LP-OLDF discriminate Alon's microarray by changing the iteration number of LINGO Program3 from one to 20. Revised LP-OLDF chooses 39 SMs (992 genes) and RIP chooses 56 SMs (1,999 genes) in 2017. We evaluate both SMs and describe how to analyze all SMs by the combination of the cluster analysis and PCA.

In this chapter, we analyze Alon's microarray in 2018 and obtain two different sets of SMs from the RIP and Revised LP-OLDF. In Sect. 4.2, RIP finds a union of 62 SMs (1,968 genes). Six MP-based LDFs find this subspace is LSD and a noise subspace (32 genes) is not LSD. H-SVM causes a computational error because a noise subspace is not LSD. Other five NMs of the noise subspace are over one. In Sect. 4.3, Revised LP-OLDF separates the microarray into a union of 32 SMs (1,005 genes) and a noise subspace (995 genes). Six MP-based LDFs find both subspaces are LSD. This fact suggests to us that a noise subspace is LSD and includes other SMs in it. We guess Problem1 causes the defect of Revised LP-OLDF. Namely, Revised LP-OLDF cannot find other SMs from the noise subspace. Section 4.4 analyzes 62 SMs found by the RIP and evaluates 62 SMs by RatioSV and NM. Moreover, the 1,891 correlations of 62 RIP discriminant scores (RipDSs) are computed. At first, we consider each gene set included in SM is cancer genes and a signal subspace. However, standard statistical methods cannot show the linear separable facts. Thus, we conclude that the new data made by RipDSs are signals themselves. Two signal data of SM13 with maximum RatioSV and SM62 with minimum RatioSV are validated. Section 4.5 analyzes two signal data made by RipDSs and HsvmDSs using 62 SMs of the RIP. The results are almost the same in Chaps. 2 and 3. However, these findings can open a new field of cancer gene diagnosis only after verification of the subjects used in the study of Alon et al. Sect. 4.6 explains the reason why standard statistical methods could not find the linear separable facts. Because the fluctuation of RipDS is embedded in the scatter plot made by the Prin1 and Prin2, we can understand our claim visually. Section 4.9 is the conclusion.

4.2 Detail Survey of Signal and Noise Subspaces Found by RIP

Method2 and RIP can decompose the Alon's microarray (1999) into many SM and noise subspace. In addition to RIP, Revised LP-OLDF and H-SVM, logistic regression confirmed that all SMs were LSD. However, we ignored to examine the noise subspace and did not confirm that the MNM of the noise subspace is greater than zero. In this section, MNM and RatioSV evaluate the signal subspace (the union of all SMs) and the noise subspace.

4.2.1 Confirmation of Signal and Noise Subspaces Found by RIP

LINDO Systems Inc. releases LINGO ver.18 in 2018 that enhances the algorithm of IP (Schrage 2006). RIP of LINGO Program3 (ver.18) separates the microarray into a signal subspace (1,968 genes) and noise subspace (32 genes). Table 4.1 shows the discriminant results of six MP-based LDFs in the microarray dataset, signal, and noise subspaces. Three values are as follows: (1) "$y_i * f(\mathbf{x}_i) > 0$" is the number of correctly classified subjects. (2) "$y_i * f(\mathbf{x}_i) = 0$" is the number of subjects on the discriminant hyperplane. (3) "$y_i * f(\mathbf{x}_i) < 0$" is the number of misclassifications (NM). Because the microarray and signal spaces are LSD (MNM = 0), six MP-based LDFs can discriminate two classes correctly. On the other hand, because noise subspace is not LSD, H-SVM cannot discriminate noise subspace because of computation error. Moreover, five NMs of each LDF are over one shown in the column "$y_i * f(\mathbf{x}_i) < 0$." Until now, because we believed LINGO Program3 worked correctly, we never confirmed these facts. We must be aware of the NMs of noise subspace with 32 genes. The NM of SVM1 is 7. The NMs of Revised IPLP-OLDF and SVM4 are 4. The NMs of RIP and Revised IPLP-OLDF are 1.

Table 4.1 Three numbers of six LDF in the microarray, signal and noise space

	Alon's microarray			Signal (62 SMs, 1968 genes)			Noise (32 genes)		
	$y_i * f(\mathbf{x}_i) < 0$	$y_i * f(\mathbf{x}_i) = 0$	$y_i * f(\mathbf{x}_i) > 0$	$y_i * f(\mathbf{x}_i) < 0$	$y_i * f(\mathbf{x}_i) = 0$	$y_i * f(\mathbf{x}_i) > 0$	$y_i * f(\mathbf{x}_i) < 0$	$y_i * f(\mathbf{x}_i) = 0$	$y_i * f(\mathbf{x}_i) > 0$
RIP	0	0	62	0	0	62	1	0	61
LP	0	0	62	0	0	62	4	0	58
IPLP	0	0	62	0	0	62	1	0	61
HSVM	0	0	62	0	0	62	–	–	–
SVM4	0	0	62	0	0	62	4	0	58
SVM1	0	0	62	0	0	62	7	0	55

4.2.2 Detail Survey of Signal and Noise Subspaces Found by RIP

By analyzing 62 SMs by RIP, we propose the 62 malignant indicators by RipDSs and make the new data that consists of 62 subjects (cases) and 62 RipDSs (variables). In this section, we survey the detail of six DSs of the microarrays, signal and noise subspaces. Table 4.2 is a survey using six MP-based LDFs. The first column shows the sequential number of subjects. The 22 normal subjects are from 1 to 22, and 40 cancer subjects are from 23 to 62. First six columns from the second column to the seventh column are six DSs (y_i * f($\mathbf{x}_i$)) for the microarray. The second six columns from the eighth column to 13th column are six DSs (y_i * f($\mathbf{x}_i$)) for a signal subspace. The last five columns from 14th column to 18th column are five DSs (y_i * f($\mathbf{x}_i$)) for noise subspace. The third row is 17 RatioSVs of six LDFs for the microarray, signal and noise subspaces. In the fourth row, we propose the new statistics (Out/NM) as follows. The subjects on the SVs are closest to each other with a distance of 2 and are the central existence of the two classes. On the other hand, we think other subjects are outliers and have different personality features. In other words, in the high-dimensional gene space, the cores of the two groups that have only small variations are fixed to two SVs, and the typical cases are considered to be outliers. If physicians examine the difference, we expect they find the new facts. In row "Out/NM," the first 12 columns show the 12 outliers and the last five columns show 5 NMs of noise subspace. We explain the meaning of RIP's outlier "0/1" by the signal columns. That implies the 22 normal subjects lie on SV = −1 and 39 cancer subjects lie on SV = 1. Thus, the 32nd cancer patient is an outlier. Its RipDS is 3.2. Signal fluctuation is much smaller than noise. In the signal space, the value of 3.2 seems to be large, but it is a small change in the fluctuation of the microarray. PCA or cluster analysis cannot successfully detect small differences in signal space. When the RIP discriminates signal subspace, the 20 normal subjects take the values −1, 39 tumor patients take the values 1 and one tumor patient is the value 3.2. The range becomes [−1, 3.2]. In the row "Out/NM" of the signal, RIP column shows "0/1." It seems that the "0/1" of RipDS is abnormal if we consider RIP's result alone. However, the other five outliers are 5/11, 6/15, 3/20, 5/37 and 4/24. To catch several cases on two SVs cause these different results. In the five columns of the noise subspace, the status "1 and 0" of RIP means the only 18th normal subject is misclassified, and all cancer subjects are classified. Thus, five NMs are 1 (0 and 1), 4 (1 and 3), 1 (1 and 0), 4 (1 and 3) and 7 (3 and 4), respectively.

Future Works: By examination of the outlier in the signal subspace, physicians may find new facts. Moreover, this table shows the reason why standard statistical methods cannot show the linear separable facts. Only six LDFs can separate two classes. However, because the ranges of DSs are a minute variation compared to the variation of the data in the microarray, PCA and cluster analysis cannot find the linear separable fact.

Table 4.2 Survey of 18 DSs

	The microarray with 2,000 genes						Signal with 1,968						Noise with 32 genes				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
Ratio	23.3	17.9	11.8	59.3	57.3	59.1	47.8	17.9	11.8	59.3	57	59.1	5.22	22.2	4.41	22.2	46.5
Out/NM	1/4	5/11	6/15	3/20	4/37	4/24	0/1	5/11	6/15	3/20	5/37	4/24	**1 and 0**	**1 and 3**	**1 and 0**	**1 and 3**	**3 and 4**
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	**0.67**
2	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−2.8	−1	−4.2	−1	−0.8
3	−1	−4.7	−1	−1	−1	−1	−1	−4.7	−1	−1	−1.1	−1	−2.5	−1	−3.1	−1	−1
4	− 1	−1	−1	−1.1	− 1.2	−1.1	−1	−1	−1	−1.1	−1.2	−1.1	−1	−1	−1	−1	−1.5
5	−1	−1	−5.1	−1	−1	−1	−1	−1	−5.1	−1	−1	−1	−17	−2.6	−19	−2.6	−1
6	−1	−1	−3.9	−1	−1	−1	−1	−1	−3.9	−1	−1	−1	−2	−1	−1.9	−1	−1
7	−1	−4	−1	−1	−1	−1	−1	−4	−1	−1	−1	−1	−7.4	−1	−5.2	−1	−1.3
8	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1.9	−1	−0.4
9	−1	−1	−1	−1.3	−1.4	−1.3	−1	−1	−1	**−1.3**	**−1.4**	**−1.3**	−2	−1.7	−1	−1.7	−0.9
10	−1	−1.7	−1	−1.1	−1.1	−1.1	−1	−1.7	−1	−1.1	−1.1	−1.1	−7.1	−1	−7.3	−1	−0.2
11	−1	−1.7	−8.2	−1	−1	−1	−1	−1.7	**−8.2**	−1	−1	−1	−1	−1	−1	−1	−0.6
12	−1	−1	−6.9	−1	−1	−1	−1	−1	−6.9	−1	−1	−1	−1	−1.6	−1	−1.6	−1
13	−1	−1	−1.5	−1	−1	−1	−1	−1	−1.5	−1	−1	−1	−1	−1	−1	−1	−0.5
14	−1	−5.5	−1.3	−1	−1	−1	−1	**−5.5**	−1.3	−1	−1	−1	−1	−1	−2.2	−1	−1
15	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−0.5
16	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1.9	−1	−1.9	−1.2
17	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−0.2

(continued)

Table 4.2 (continued)

	The microarray with 2,000 genes						Signal with 1,968						Noise with 32 genes				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
18	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	**31.5**	4.34	31.8	4.34	2.55
19	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1.2	−1	−1	−1	−0.3
20	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	**1.15**
21	−1	−1	−1	−1	−1.1	−1.1	−1	−1	−1	−1	−1.1	−1.1	−6.2	−1	−2	−1	−1
22	**−6**	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−0.3
23	1	1	1	1	1.02	1	1	1	1	1	1.03	1	1	1.3	1	1.3	1
24	1	1	1	1	1.01	1	1	1	1	1	1.01	1	2.76	1	3.98	1	0.21
25	1	1	7.29	1	1.01	1	1	1	7.29	1	1.02	1	1	2.51	3.15	2.51	1
26	1	1	1	1	1.09	1.02	1	1	1	1	1.11	1.01	1	5.66	1	5.66	1.59
27	1	1	7.06	1	1.01	1	1	1	7.06	1	1.02	1	1.01	1	1	1	1.78
28	**1.2**	3.17	1.29	1.3	1.38	1.3	1	3.17	1.29	1.3	1.41	1.31	1	3.93	1	3.93	1
29	1	1	1	1	1.01	1	1	1	1	1	1.02	1	1	1	1	1	0.04
30	1	3.88	3.24	1	1.01	1	1	3.88	3.24	1	1.01	1	1	1	1	1	0.79
31	1	2.91	1	1	1.05	1	1	2.91	1	1	1.07	1	1	1	1	1	0.44
32	1	1	1	1.21	1.27	1.22	**3.2**	1	1	1.21	1.29	1.22	1	1.85	1	1.85	1
33	1	1	1	1	1.01	1	1	1	1	1	1.02	1	1	1	1	1	**−0.4**
34	1	1	1	1	1.01	1	1	1	1	1	1.01	1	1	1	1	1	1
35	1	1	1	1	1.15	1.03	1	1	1	1	1.17	1.03	1.41	2.91	3.87	2.91	1.1
36	1	1	1	1.84	1.97	1.85	1	1	1	1.84	1.99	1.85	11.3	3.16	12.4	3.16	1
37	1	2.35	1	1.56	1.66	1.57	1	2.35	1	1.56	1.68	1.57	8.64	1	5.75	1	1

(continued)

Table 4.2 (continued)

	The microarray with 2,000 genes						Signal with 1,968						Noise with 32 genes				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
38	1	1	1	1.83	1.96	1.84	1	1	1	1.83	1.98	1.84	15.9	2.46	18.2	2.46	2.13
39	1	1	1	1.77	1.89	1.78	1	1	1	1.77	1.91	1.78	11.9	1.74	14	1.74	1
40	1	1	1	1.5	1.61	1.51	1	1	1	1.5	1.64	1.51	3.58	3.21	1	3.21	1.71
41	1	1	1	1.23	1.33	1.24	1	1	1	1.23	1.34	1.23	1	1	1	1	0.77
42	**1.5**	1	4.61	1.34	1.44	1.35	1	1	4.61	1.35	1.46	1.36	1	1.52	1	1.52	1
43	1	1	1	1.64	1.77	1.66	1	1	1	1.65	1.8	1.66	1.77	2.99	2.86	2.99	1.02
44	**3**	1.53	1	1.43	1.54	1.44	1	1.53	1	1.43	1.57	1.44	5.39	3.2	4.39	3.2	2.04
45	1	3.88	3.86	1.37	1.46	1.38	1	3.88	3.86	1.38	1.48	1.38	1	4.1	2.9	4.1	1.84
46	1	4.28	5.85	2	2.12	2.01	1	4.28	5.85	2.01	**2.15**	2.02	1	4.63	1.11	4.63	2.79
47	1	1	8.14	1.26	1.39	1.27	1	1	8.14	1.26	1.41	1.27	19	6.45	26.2	6.45	1.27
48	1	3.88	8.73	1.88	2.01	1.9	1	3.88	**8.73**	1.89	2.04	1.9	1.56	2.85	5.16	2.85	1.84
49	1	1	2.57	1.11	1.24	1.12	1	1	2.57	1.11	1.26	1.12	12.2	5.24	11.6	5.24	1.66
50	1	1.68	1.58	1	1.01	1	1	1.68	1.58	1	1.02	1	1	1	1	1	0.48
51	1	1	2.56	1.41	1.51	1.42	1	1	2.56	1.41	1.54	1.42	1	1	1	1	1
52	1	1	1	1	1	1	1	1	1	1	1	1	1	**−3.9**	1	**−3.9**	**−1.7**
53	1	1	1	2.05	2.14	2.06	1	1	1	**2.06**	**2.15**	**2.06**	4.25	1	5.95	1	1
54	1	1	4.37	1.54	1.62	1.56	1	1	4.37	1.55	1.64	1.56	1	1	1	1	1
55	**2.98**	1	1	1	1	1	1	1	1	1	1	1	1	**−1.3**	1	**−1.3**	**−1.4**
56	1	2.78	2.5	1	1.01	1	1	2.78	2.5	1	1.01	1	1	1	1.22	1	1
57	1	5.66	1	1	1.01	1	1	**5.66**	1	1	1.01	1	5	3.93	9.34	3.93	1.3
58	1	1	1	1	1	1	1	1	1	1	1	1	1	**−4.2**	1	**−4.2**	**−0.1**

(continued)

Table 4.2 (continued)

	The microarray with 2,000 genes						Signal with 1,968						Noise with 32 genes				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
59	1	1	1	1	1.01	1	1	1	1	1	1.01	1	1	4.36	1	4.36	1
60	1	1	1	1.54	1.63	1.55	1	1	1	1.53	1.65	1.54	20.9	1	22.4	1	1.23
61	1	1	4.26	1	1.1	1.02	1	1	4.26	1	1.12	1.02	1.11	1	1	1	1
62	1	1	1	1	1.09	1.01	1	1	1	1	1.11	1.01	10.3	4.7	5.99	4.7	1.77

4.2.3 Basic Structure of Six Microarrays

Table 4.2 tells us the basic structure of six microarrays. Subjects of two classes are separate in the microarray (2,000 genes) and a signal subspace (1,986 genes) entirely. RIP gathers all normal subjects on SV = −1 and 39 cancer subjects on SV = 1. Aoshima and Yata (2017) pointed out two classes locate on two balls in high-dimensional gene space, also. We guess their balls correspond to two SVs. If so, two different approaches found the same results that microarrays are LSD. Moreover, the first eigenvalue is large because two classes are almost on the Prin1 and two balls are on the Prin1. Because they found their results using high-dimensional PCA, they cannot find that one cancer subject is an outlier. Moreover, their approach does not decompose the microarrays into many SMs and the noise subspace. SM is more valuable than the microarrays for cancer gene diagnosis.

4.3 Detail Survey of Signal and Noise Subspaces Found by Revised LP-OLDF

RIP can decompose the microarray into many SMs and noise subspace. Revised LP-OLDF can decompose the microarray into signal and noise subspaces as same as RIP, also. However, because Revised LP-OLDF often gathered subjects on the discriminant hyperplane by the discrimination of overlapping data, and NM is not reliable (Problem1), we have analyzed only SM of RIP until now. In this chapter, we investigate the possibility of SMs obtained by Revised LP-OLDF in addition to RIP's SMs.

4.3.1 Confirmation of Signal and Noise Subspaces Found by Revised LP-OLDF

Revised LP-OLDF of LINGO Program3 can separate Alon's microarray into the union of 32 SMs (1,005 genes) and noise subspace (995 genes). Table 4.3 shows three results of six MP-based LDFs in the microarray, signal and noise subspaces. Because the microarray and signal subspaces are LSD, six LDFs can discriminate those correctly. However, noise subspace is LSD, also. These results indicate Revised LP-OLDF cannot find other SMs included in the noise subspace. We guess Problem1 causes this defect. Because H-SVM discriminates the noise subspace correctly, this fact shows the noise subspace is LSD, also.

Table 4.3 Three numbers of six MP-based LDF's DSs of the microarray, signal, and noise spaces

	Alon's microarray			Signal (32 SMs with 1005 genes)			Noise with 995 genes		
	$y_i * f(\mathbf{x}_i) < 0$	$y_i * f(\mathbf{x}_i) = 0$	$y_i * f(\mathbf{x}_i) > 0$	$y_i * f(\mathbf{x}_i) < 0$	$y_i * f(\mathbf{x}_i) = 0$	$y_i * f(\mathbf{x}_i) > 0$	$y_i * f(\mathbf{x}_i) < 0$	$y_i * f(\mathbf{x}_i) = 0$	$y_i * f(\mathbf{x}_i) > 0$
RIP	0	0	62	0	0	62	0	0	62
LP	0	0	62	0	0	62	0	0	62
IPLP	0	0	62	0	0	62	0	0	62
HSVM	0	0	62	0	0	62	**0**	**0**	**62**
SVM4	0	0	62	0	0	62	0	0	62
SVM1	0	0	62	0	0	62	0	0	62

4.3.2 Detail Survey of Signal and Noise Subspaces Separated by Revised LP-OLDF

Table 4.4 shows results using six LDFs. The first column shows the sequential number of 62 subjects. The third row is 18 RatioSVs by six LDFs for the microarray, signal and noise subspaces. Row "Outlier" shows the status of 18 outliers because the noise subspace is LSD. The "RatioSV and Outlier" of RIP in the signal subspace are 14.1% and 8/27. RIP becomes almost the same results of other five LDFs. On the other hand, those in Table 4.2 are 47.8% and 0/1. Thus, two signals of RIP and Revised LP-OLDF may have different characteristics.

Table 4.4 Survey of the microarray, signal, and noise subspaces by revised LP-OLDF (The 22 rows from SN = 1 to SN = 22 are class1. All values are negative values)

	The Microarray						Signal with 1,005 genes						Noise with 995 genes					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
Ratio	**15.8**	27.6	20.1	59.3	58.6	59.2	**14.1**	14.3	23.6	56.2	55.1	56.2	**3.41**	7.73	14.7	55.4	53.3	55.3
Outlier	**2/3**	2/12	8/25	4/20	11/32	4/24	**8/27**	3/13	2/13	3/23	7/29	3/23	**5/13**	6/12	2/15	5/23	13/32	6/24
1	1	1	1	1	1.01	1	**3.18**	2.35	1.52	1	1.01	1	1	1	1	1	1.01	1
2	**8.09**	1	1	1	1	1	1	1	1	1	1	1	3.43	1	1	1	1	1
3	1	2.28	1.3	1	1.03	1	**1.01**	1	1	1	1.01	1	1	9.57	1	1	1.1	1.01
4	1	1	1	1.11	1.16	1.11	1	1	1	1.15	1.21	1.15	15	1	1	1.06	1.15	1.07
5	**3.54**	1	2.14	1	1.01	1	1	1	1	1	1	1	1	1.55	1	1	1.02	1
6	1	1	1	1	1.01	1	**1.23**	1	3.2	1	1	1	1	1	1	1	1.01	1
7	1	3.33	1	1	1.02	1	1	1	1	1	1.02	1	1	2.36	3.93	1	1.02	1
8	1	1	3.78	1	1	1	1	1	1	1	1	1	1	1	1	1	1.01	1
9	1	1	2.17	1.32	1.32	1.32	1	1	1	1.16	1.18	1.16	10.8	2.51	1	1.59	1.61	1.59
10	1	1	1.82	1.06	1.11	1.07	1	1	1	1.22	1.25	1.22	1	2.99	1	1.01	1.09	1.02
11	1	1	2.43	1	1.01	1	1	5.83	1	1	1	1	1	1	3.77	1.04	1.09	1.05
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	2.48	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	3	1	1	1.01	1	**1.26**	1	1	1	1	1	1	3.3	1	1	1	1
16	1	1	1.46	1	1	1	**2.25**	1	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	**7.12**	1	1	1	1	1	1	1	1	1	1.01	1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1	8.26	1	1	1	1.01	1

(continued)

Table 4.4 (continued)

	The Microarray						Signal with 1,005 genes						Noise with 995 genes					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	1	1	2.26	1.04	1.1	1.05	**2.67**	1	1	1	1.01	1	23.8	1	1	1.14	1.17	1.14
22	1	1	1	1	1	1	**5.79**	1	1	1	1	1	1	1	1	1	1	1
23	1	1	3.15	1	1.02	1	1	1	1	1	1	1	16.7	4.98	2.11	1.07	1.1	1.07
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	1	2.04	1	1.01	1	1	1	3.69	1	1	1	11.4	1	1.86	1	1.03	1
26	1	3.1	2.22	1	1.07	1.01	1	8.15	1	1.22	1.25	1.22	1	2.03	1	1	1.02	1
27	1	1.52	1.61	1	1.01	1	**4.4**	3.01	2.47	1	1.01	1	1	1	7.5	1	1.01	1
28	1	1	1.22	1.3	1.33	1.3	**2.12**	1	2.13	1.25	1.3	1.25	1	1	1	1.33	1.39	1.33
29	1	1	1	1	1.01	1	1	1	1.38	1	1.01	1	5.35	1	1	1	1.02	1
30	1	1	1	1	1	1	**5.11**	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1.04	1	**1.99**	1	1	1.1	1.14	1.1	1	1	1	1	1.01	1
32	1	1	3.34	1.21	1.25	1.22	**5.1**	1	1	1.01	1.08	1.02	6.4	12.6	9.67	1.41	1.46	1.41
33	1	1	1	1	1.01	1	1	4.23	3.27	1	1.02	1	7.6	1	1	1	1.01	1
34	1	1.84	1	1	1	1	**4.25**	1	1	1	1	1	1	1	1	1	1.01	1
35	1	1	2.92	1	1.1	1.03	**6.48**	1	1	1.06	1.14	1.06	1	12.5	3.6	1.09	1.19	1.09
36	1	3.39	2.33	1.84	1.9	1.85	**6.18**	1	1	1.88	1.95	1.88	1	1	1	1.89	2	1.9
37	1	2.78	1.45	1.56	1.61	1.57	**2.82**	1	1	1.84	1.88	1.84	6.81	1	1	1.42	1.5	1.42
38	1	2.86	2.62	1.83	1.89	1.84	**1.94**	1	1.43	1.81	1.87	1.81	3.58	9.05	2.13	2.02	2.14	2.03
39	**2.79**	1	1	1.77	1.82	1.78	**3.33**	1	1	1.76	1.83	1.76	1	9.93	2.3	1.92	2	1.92
40	1	1	1	1.5	1.55	1.51	**6.22**	5.58	1	1.8	1.84	1.8	1	1	1	1.39	1.46	1.39
41	1	1	3.33	1.23	1.28	1.24	**2.31**	1	1	1.27	1.32	1.27	1	1	3.85	1.12	1.2	1.12

(continued)

Table 4.4 (continued)

	The Microarray						Signal with 1,005 genes						Noise with 995 genes					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
42	1	1.51	2.51	1.34	1.38	1.35	**3.46**	1.58	1	1.54	1.58	1.54	1	1	1	1.18	1.25	1.18
43	1	1.57	3.27	1.64	1.71	1.66	1	1.83	1	1.66	1.71	1.66	20.3	1	1	1.68	1.79	1.69
44	1	1	2.13	1.43	1.49	1.44	**3.35**	1.78	2.99	1.57	1.62	1.57	19.7	4.79	3.79	1.38	1.48	1.39
45	1	1	2.38	1.37	1.41	1.38	**4.61**	3.22	1	1.58	1.63	1.58	6.51	2.93	1	1.24	1.33	1.25
46	**2.92**	1.57	4.48	2	2.05	2.01	**3.75**	5.7	5.26	2.34	2.38	2.34	34.9	3.24	1	1.7	1.8	1.71
47	1	1	2.14	1.26	1.33	1.27	1	1	1	1.15	1.23	1.15	1	16.3	1.24	1.43	1.52	1.44
48	**4.54**	2.45	4.43	1.88	1.94	1.89	**3.72**	1	1	2.03	2.08	2.03	1	1	3.98	1.77	1.9	1.78
49	1	1	1	1.11	1.19	1.12	**1.21**	1	1	1.26	1.32	1.26	1	1	3.44	1.09	1.19	1.09
50	1	1	3.64	1	1.01	1	**5.46**	2.51	1.46	1	1.05	1	1	1	1	1	1.01	1
51	1	1	2.02	1.41	1.45	1.42	1	5.71	1	1.85	1.89	1.85	1	1	1	1	1.12	1.01
52	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
53	1	1	3.25	2.05	2.09	2.06	**6.6**	2.67	1.51	2.18	2.21	2.18	1	1	5.45	1.9	1.98	1.91
54	1	1	3.95	1.54	1.58	1.55	**7.04**	1	4.42	1.76	1.81	1.76	1	1	1	1.31	1.37	1.31
55	1	1	1	1	1	1	1	1	1	1	1	1	10.6	1	1	1	1	1
56	1	1	1	1	1	1	**2.09**	1	1	1	1	1	1	6.38	1	1	1	1
57	1	3.91	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
58	1	1	1	1	1	1	**2.5**	1	1	1	1	1	1	1	1	1	1	1
59	1	1	2.51	1	1.01	1	1	2.3	1	1	1	1	1	1	1	1	1	1
60	1	2.98	6.17	1.54	1.59	1.54	**5.32**	1	1.71	1.77	1.82	1.77	1	1	4.41	1.31	1.4	1.31
61	1	1	3.9	1	1.07	1.01	**2.23**	1	1	1	1.01	1	1	1	1	1.18	1.26	1.18
62	1	1	1	1	1.06	1.01	**5.32**	1	4.88	1	1.02	1	5.02	3.19	4.39	1.04	1.14	1.05

4.4 Analysis of 62 SMs Found by RIP

4.4.1 Examination of RatioSV and NM

Table 4.5 shows the result of 62 SMs. "SM and Gene" columns are the sequential number of 62 SMs and the number of genes included in each SM. Last four rows are elementary statistics. The largest SM contains 44 genes, and minimum SM contains 24 genes. The average is about 32 genes. It is an important fact that the maximum gene is less than the case number (62) because Method2 find the number of nonzero coefficients less than 62 at the first step. That is, RIP and Revised LP-OLDF are irrelevant to the curse of dimensionality. Moreover, those can separate the signal and noise subspaces naturally. Next six columns are RatioSVs of six MP-based LDFs. RatioSV is the proper statistic for LSD-discrimination, and its values become large for overlapping data and are not reliable. In this book, we calculate the RatioSVs of SVM4 and SVM1. There are 36 underlined RatioSVs of SVM1 those are over the maximum values because those SVM1 sometimes cannot discriminate these SMs correctly. "Max and Min" columns are the maximum and minimum RatioSV of RIP, LP, IPLP, and H-SVM because SVM4 and SVM1 are often overlapping. Because H-SVM maximizes the SV distance, it takes the maximum value 47 among six LDFs. "Max and Min" columns are the maximum and minimum values of four LDFs. The ranges of "Max and Min" columns are [3.81, 31.22] and [3.3, 21.74], respectively. Notably, 61 maximum values are over 8.01%. This truth means the discrimination of the two classes is apparent in 61 SMs. Last two columns are two NMs of SVM1 and Fisher's LDF (LDF2). SVM1 discriminates 25 SMs, and LDF2 discriminates 22 SMs correctly. Because all NMs of logistic regression, SVM4 and QDF are zero, we omit three columns from the table. These truths are very critical. RIP and H-SVM discriminate six microarrays and all SMs theoretically. Revised LP-OLDF and Revised IPLP-OLDF discriminate six microarrays and all SMs empirically. Logistic regression discriminates all SMs empirically. SVM4 and QDF often discriminate many SMs empirically. However, cluster analysis, PCA, one-way ANOVA and t-tests cannot show the linearly separable facts of all SMs. Thus, although statistical discriminant functions are useless for Problem5, those are better than other statistical methods. Furthermore, only RIP can decompose microarrays and other ordinary data into SMs and BGSs correctly. Although Revised LP-OLDF decompose microarrays into SMs, it cannot find all SMs correctly. SVMs cannot decompose microarrays into SMs. However, statistical methods are useful for cancer gene diagnosis if those analyze signal data such as RipDSs, LpDSs and HsvmDS using SMs found by RIP.

Table 4.5 RatioSVs and NMs of 62 SMs

SM	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM1	LDF2
1	28	22.53	22.24	20.99	**26.1**	25.92	26.1	26.1	20.99	0	0
2	31	**27.08**	23.05	14.93	25.42	24.13	25.42	27.08	14.93	0	1
3	30	17.13	19.23	17.15	**20.89**	20.81	20.89	20.89	17.13	0	0
4	25	11.85	14.33	16.09	**18.87**	18.88	~~23.98~~	18.87	11.85	1	2
5	29	24.68	25.87	17.05	**27.47**	26.63	27.47	27.47	17.05	0	0
6	35	**20.52**	11.46	12.26	18.91	18.58	18.91	20.52	11.46	0	1
7	31	21.74	22.66	23.06	**25.33**	24.75	25.33	25.33	**21.74**	0	0
8	31	17.36	**20.1**	10.48	18.93	18.88	18.93	20.1	10.48	0	0
9	31	14.61	16.29	7.18	**17.37**	17.4	19.65	17.37	7.18	0	0
10	29	22.82	**27.38**	12.02	23.7	23.64	23.7	27.38	12.02	0	0
11	31	16.78	20.06	17.62	**23.9**	23.78	23.9	23.9	16.78	0	0
12	26	16.2	15.28	15.03	**22.77**	22.72	22.77	22.77	15.03	0	0
13	37	29.03	21.98	21.15	**31.22**	30.03	31.22	**31.22**	21.15	0	0
14	27	11.47	12.31	9.16	**12.98**	12.88	~~21.55~~	12.98	9.16	1	3
15	27	**27.63**	11.91	15.78	17.32	17.2	~~22.82~~	27.63	11.91	1	0
16	25	14.16	14.1	15	**16.59**	16.59	~~16.56~~	16.59	14.1	1	1
17	35	**23.98**	21.21	13.63	23.78	23.77	23.78	23.98	13.63	0	0
18	24	9.45	15.95	15.2	**16.8**	16.74	~~21.53~~	16.8	9.45	1	2
19	32	12.5	13.14	15.74	**17.12**	17.11	~~20.47~~	17.12	12.5	1	2
20	37	24.68	21.33	13.9	**28.85**	28.43	28.85	28.85	13.9	0	0
21	27	14.71	**15.45**	12.27	15.04	14.98	18.94	15.45	12.27	0	2
22	31	18.57	16.57	13.12	**19.93**	19.91	26.45	19.93	13.12	0	2
23	29	18.56	16.53	**21.01**	18.29	18.31	22.27	21.01	16.53	0	1
24	26	**19.1**	18.52	15.96	18.55	18.55	20.46	19.1	15.96	0	2
25	31	21.75	20.81	21.68	**24.58**	24.7	25.71	24.58	20.81	0	0
26	25	13.67	13.83	12.97	**14.79**	14.76	~~24.18~~	14.79	12.97	1	2
27	29	22.02	10.74	16.02	**26.04**	26.1	32.88	26.04	10.74	0	0
28	33	18.61	17.28	14.62	**21.88**	21.7	29.19	21.88	14.62	1	0
29	29	23.29	17.43	25.38	**26.98**	26.58	26.86	26.98	17.43	0	0
30	36	22.07	18.24	20.58	**25.98**	26.27	~~25.99~~	25.98	18.24	1	0
31	31	12.21	12.98	13.29	**17.5**	17.57	~~23.18~~	17.5	12.21	1	2
32	30	12.12	**17.34**	12.41	15.31	15.26	~~23.24~~	17.34	12.12	2	3
33	31	15.7	15.48	6.3	**17.4**	17.4	~~24.05~~	17.4	6.3	1	2
34	38	24.37	21.26	22.46	**29.6**	29.6	29.83	29.6	21.26	0	0

(continued)

Table 4.5 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	SVM4	SVM1	Max	Min	SVM1	LDF2
35	31	**22.04**	19.33	21.67	20.25	20.19	~~24.23~~	22.04	19.33	1	0
36	31	16.95	10.7	10.75	**17.01**	16.98	20.04	17.01	10.7	0	2
37	31	11.52	9.52	12.61	**15.2**	15.32	~~29.3~~	15.2	9.52	2	2
38	28	12.44	15.44	9.29	**20.38**	20.25	~~27~~	20.38	9.29	3	1
39	29	11.66	14.87	15.76	**20.67**	20.68	~~28~~	20.67	11.66	3	3
40	27	17.52	16.38	16.07	**19.69**	19.68	~~25.68~~	19.69	16.07	1	1
41	32	11.93	9.77	10.57	**14.03**	14.03	~~22.25~~	14.03	9.77	2	5
42	32	18.98	19.06	13.19	**25.31**	25.26	~~29.2~~	25.31	13.19	1	0
43	32	14.24	17.42	12.93	**18.56**	18.48	~~19.77~~	18.56	12.93	1	1
44	28	18.76	16.13	9.48	**19.06**	19.01	~~21.45~~	19.06	9.48	1	1
45	32	9.48	11.03	8.27	**12.16**	12.12	~~19.91~~	12.16	8.27	3	3
46	32	13.89	12.48	11.64	**18.65**	18.63	~~26.57~~	18.65	11.64	2	2
47	30	14.83	14.78	10.17	**19.63**	19.52	24.94	19.63	10.17	0	2
48	32	11.78	10.76	9.66	**13.46**	13.47	~~18.76~~	13.46	9.66	2	2
49	40	20.95	19.57	15.64	**23.73**	23.51	24.35	23.73	15.64	0	0
50	38	18.32	11.09	14.26	**21.47**	21.47	~~28.3~~	21.47	11.09	2	0
51	38	12.27	6.51	11.08	**13.08**	13.08	~~26.69~~	13.08	6.51	4	3
52	33	11.36	**12.45**	11.59	11.85	11.84	~~21.72~~	12.45	11.36	3	1
53	31	6.74	6.31	7.53	**8.35**	8.35	~~25.05~~	8.35	6.31	4	6
54	36	**17.1**	15.57	15.53	16.44	16.47	~~35.18~~	17.1	15.53	4	2
55	38	13.02	11.2	12.98	**14.58**	14.55	~~28.17~~	14.58	11.2	3	4
56	31	4.43	3.99	3.53	**4.66**	4.66	~~25.29~~	4.66	3.53	5	6
57	35	**6.78**	6.76	5.52	6.34	6.34	~~32.29~~	6.78	5.52	5	4
58	36	7.03	7.26	4.87	**9.66**	9.67	~~34.6~~	9.66	4.87	5	6
59	37	6.41	5.38	6.69	**6.83**	6.83	~~38.48~~	6.83	5.38	4	4
60	39	6.83	**9.76**	7.5	8.55	8.53	~~36.14~~	9.76	6.83	6	5
61	44	6.93	4.5	7.87	**8.01**	8.01	~~44.51~~	8.01	4.5	6	4
62	38	3.3	3.62	3.44	**3.81**	3.81	~~36.92~~	3.81	3.3	5	5
MAX	44	29.03	27.38	25.38	**31.22**	30.03	44.51	**31.22**	21.74	6	6
MIN	24	3.3	3.62	3.44	3.81	3.81	16.56	3.81	3.3	0	0
Mean	31.74	15.97	14.90	13.41	18.35	18.25	25.68	18.82	12.26	1.47	1.66
SUM	1968										

4.4.2 Correlations of 62 RipDSs

Table 4.6 shows 1,891 pairs of correlations (abbreviate R) of 62 RipDSs sorted in descending order by R. The top ten pairs are high correlations, and the lower ten pairs are lower correlations. Of the ten pairs taking values of R = 0.841 to R = 0.865, five pairs of RIP13 and (RIP30, RIP17, RIP33, RIP7, RIP34) are 0.845 or more. Because these five pairs are expected to be the core of the discrimination between two classes, we analyze mainly these pairs in this study. On the other hand, the ten lower pairs have correlations of 0.332~0.389, and the six pairs of RIP61 and (RIP21, RIP48, RIP28, RIP62, RIP55, RIP60) have low correlation with each other.

Golub et al. (1999) said they were studying class prediction to discover new cancer classes and assign tumors to known classes from 1970. Because it is difficult to rely on traditional biological insights to classify cancers in a non-systematic and biased approach, they tried to classify cancers by co-expression levels of thousands of genes using microarrays. They developed a more systematic approach to cancer and began to discover cancer variants. After the weighted voting method selects genes as candidates, SOM divides two classes, and LOO evaluates the predicted results. However, by examining six pairs with 62 RipDSs with correlations as small as 0.389 or less, it is possible to expect the possibility of finding subspecies in addition to outliers found by PCA analysis of new data.

Table 4.6 1,891 pairs of correlations of 62 RIPs DSs (RIP13 is an abbreviation of 13th RipDS)

	Variable	Versus variable	Correlation	Frequency	Lower 95%	Upper 95%	p-value
1	RIP7	RIP2	0.865	62	0.784	0.917	1.34E−19
2	**RIP30**	**RIP13**	0.853	62	0.767	0.909	1.31E−18
3	**RIP17**	**RIP13**	0.850	62	0.762	0.907	2.52E−18
4	RIP26	RIP23	0.847	62	0.758	0.906	3.81E−18
5	**RIP33**	**RIP13**	0.845	62	0.755	0.904	5.54E−18
6	**RIP13**	**RIP7**	0.845	62	0.755	0.904	5.59E−18
7	RIP29	RIP4	0.845	62	0.755	0.904	5.67E−18
8	**RIP34**	**RIP13**	0.845	62	0.754	0.904	5.94E−18
9	RIP15	RIP1	0.844	62	0.753	0.903	6.79E−18
10	RIP38	RIP23	0.841	62	0.748	0.901	1.26E−17
–	–	–	–	–	–	–	–
1882	**RIP61**	**RIP21**	0.389	62	0.154	0.582	0.001785

(continued)

Table 4.6 (continued)

	Variable	Versus variable	Correlation	Frequency	Lower 95%	Upper 95%	p-value
1883	**RIP61**	**RIP48**	0.387	62	0.152	0.581	0.001893
1884	RIP60	RIP43	0.375	62	0.138	0.571	0.002689
1885	**RIP61**	**RIP28**	0.375	62	0.138	0.571	0.002691
1886	RIP57	RIP55	0.374	62	0.137	0.571	0.002725
1887	RIP62	RIP11	0.362	62	0.124	0.561	0.003796
1888	**RIP62**	**RIP61**	0.347	62	0.106	0.549	0.005784
1889	**RIP61**	**RIP55**	0.340	62	0.099	0.544	0.00678
1890	**RIP61**	**RIP60**	0.334	62	0.092	0.539	0.007998
1891	RIP62	RIP56	0.332	62	0.089	0.537	0.008464

Figure 4.1 shows the correlation matrix of the six RipDSs. RIP13 has a high correlation with (RIP30, RIP17, RIP61) and has a low correlation with (RIP55, RIP60). All discriminant scores are SV ≤ -1 or SV ≥ 1. The scatter plot does not explain the useful facts.

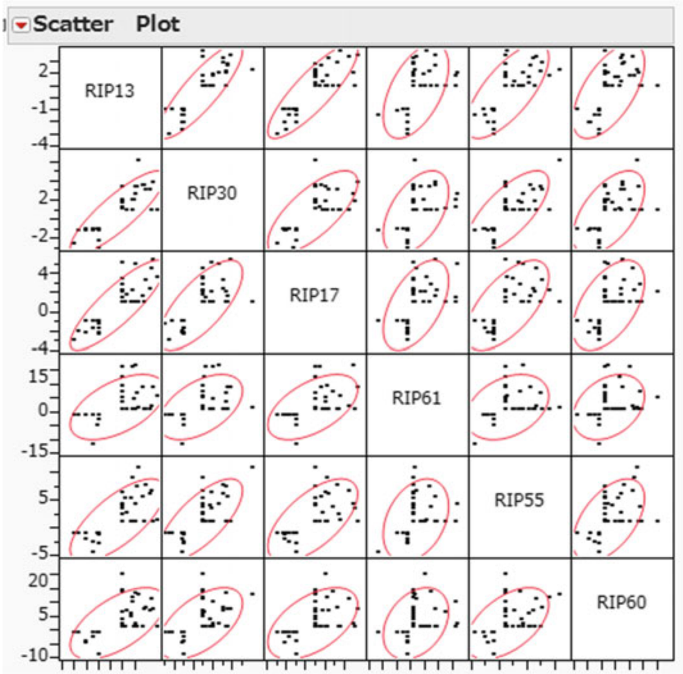
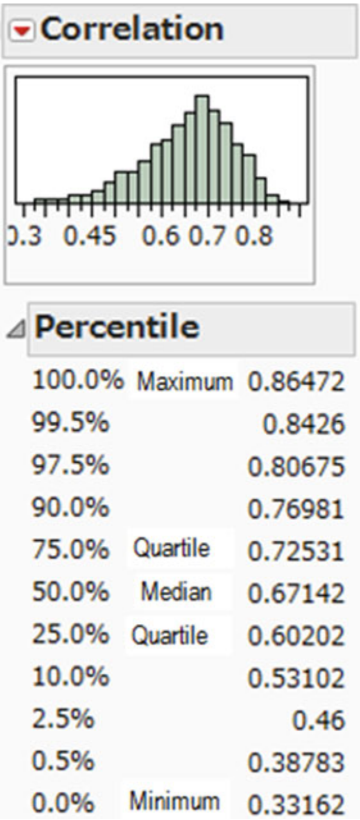


Fig. 4.1 Correlation matrix of 62 RipDSs (RIP13 is an abbreviation for 13th RipDS)

Figure 4.2 shows the distribution of 1,891 correlations of 62 RipDSs. The correlations widely vary from 0.331 to 0.864 and are a monomodal distribution with the lower long skirt. Due to the diversity of cancer, the method of determining which SM group is complementary to each other or a group representing different cancer variants is a future research topic. From this distribution, all correlations are positive, and the 75% point is r = 0.72531. That is, the correlations of 75% are 0.725 or more, and there is no negative correlation. On the other hand, correlations of genes included in SM take the positive, almost zero and negative values. Thus, we conclude the signal data is true signal and SM is not signal. From these results, the particular data structure where positive correlations occupy a majority may have much influence on cancer gene analysis.



Percentile

100.0%	Maximum	0.86472
99.5%		0.8426
97.5%		0.80675
90.0%		0.76981
75.0%	Quartile	0.72531
50.0%	Median	0.67142
25.0%	Quartile	0.60202
10.0%		0.53102
2.5%		0.46
0.5%		0.38783
0.0%	Minimum	0.33162

Fig. 4.2 Distribution of 1,891 correlations of 62 RipDSs

4.5 Validation of SM13 and SM62

4.5.1 RatioSVs and Outliers of SM13 and SM62

Table 4.7 focuses on SM13 and SM62 as a representative of all 62 SMs because two SMs take the maximum and minimum RatioSVs. We compare the RatioSVs and the number of outliers by the six LDFs. In the signal subspace of Table 4.2, there are many cases on the SV. However, the proportion of cases on SV dramatically decreases in SMs those are ordinary small samples. **That is, this comparison indicates the difference of discrimination in high-dimensional space and small samples in our research.**

The RatioSVs of the H-SVM take the maximum value among the four LDFs. However, H-SVMs have many outliers and those fluctuate within a narrow range of H-SVM. Aoshima and Yata indicate two classes distribute balls in high-dimensional gene space. We consider two balls correspond to two SVs. Moreover, Method2 decomposes the microarrays into 62 SMs. All 62 SMs show genetic diversity. Furthermore, it is consistent with common sense that the tumor subjects change mainly than normal subjects. On the other hand, in SM62, five RatioSVs except for SVM1 are smaller than SM13. Although SVM1 discriminate SM13 correctly, SVM1 misclassifies five cases of SM62. Thus, RatioSV of SVM1 becomes 36.92 in SM62. The reason why nobody recognizes such apparent results is as follows.

(1) Many researchers cannot find the fact that the microarray is LSD (Fact3) because the statistical discriminant functions cannot discriminate LSD (or MNM = 0) correctly. In the microarrays, MNM of the signal subspace is zero and MNM of the noise subspace is over one. This fact is totally unexpected and innovative.

(2) RIP and Revised LP-OLDF can decompose the microarray into many SMs and H-SVM cannot find SMs (Fact4). Although statistical methods cannot separate two classes included in SM, RIP and Revised LP-OLDF can separate two classes. Although many methods have been proposed to find oncogenes in many studies, the evidence that they are oncogenes is ambiguous compared to our results.

(3) The maximum value of RatioSVs for six microarrays exceeds 30%. The reason why nobody finds such an obvious fact is that when normal subjects become cancer, it completely separates from the normal class, but its variation is too small. **Because statistics assume that large variance is more meaningful than the small variance, standard statistical methods could not detect small variation**. Furthermore, microarrays have a particular structure having all positive correlations. We explain our claim in later.

Check Subjects on SVs by RIP and H-SVM for SM13: RIP and H-SVM find the eight normal subjects (1, 4, 6, 10, 15, 18, 20, 22) and 10 cancer subjects (24, 31, 33, 37, 47, 52, 55, 56, 57, 61) locate on SVs. If physicians examine the difference between SV's subjects and outliers, they may be able to find new facts for cancer gene diagnosis.

Table 4.7 Comparison of SM13 with the largest RatioSV and SM62 with the smallest RatioSV

	SM13						SM62					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
Ratio	29.03	21.98	21.15	**31.22**	30.03	31.22	3.3	3.62	3.44	**3.81**	3.81	**36.92**
Outlier	7/18	8/24	12/21	13/29	18/40	13/29	7/17	4/20	5/21	6/21	7/23	5
1	−1	−1	−1	−1	−1.002	−1	−6.28	−1	−1	−5.76	−5.76	−0.33
2	−2	−1	−1	−1.006	−1.128	−1.006	−1	−1	−1	−1	−1	−0.79
3	−1	−1.312	−2.344	−1.888	−1.92	−1.888	−1	−1	−1	−1	−1	−0.28
4	−1	−1	−1.685	−1	−1.024	−1	−32.2	−21.4	−1	−22.8	−22.8	−1
5	−1	−1	−2.307	−1.251	−1.354	−1.251	−4.26	−1.05	−10.5	−3.48	−3.47	−1.35
6	−1	−2.333	−1	−1	−1.005	−1	−1	−1	−1	−1	−1	−1
7	−1.467	−2.313	−1.182	−1.32	−1.611	−1.32	−1.96	−1.94	−1	−2.17	−2.17	−1
8	−1	−1	−1	−1.246	−1.435	−1.246	−1	−1	−6.61	−1	−1	−0.39
9	−1.4965	−1	−4.037	−3.283	−3.415	−3.283	−1	−1	−3.58	−1.41	−1.41	−1.06
10	−1	−1	−2.238	−1	−1.005	−1	−10.8	−1	−17.9	−1	−1	−1
11	−2.0802	−1	−3.717	−1	−1.007	−1	−1	−1	−1	−1	−1	−1
12	−1	−1	−1	−1.066	−1.148	−1.066	−1	−1	−1	−1	−1	−1
13	−1	−1	−1	−1.277	−1.404	−1.277	−1	−1	−1	−1	−1	−0.88
14	−2.0794	−1.173	−1.674	−2.333	−2.41	−2.333	−1	−1	−1	−1	−1	−1.25
15	−1	−1	−1	−1	−1.022	−1	−1	−1	−1	−1	−1	−1
16	−2.6304	−1.683	−1.131	−2.344	−2.307	−2.344	−4.58	−1	−1	−1	−1.01	−1.05
17	−1.1033	−3.243	−1.201	−1.682	−1.727	−1.682	−1	−5.89	−1	−1	−1	−1

(continued)

Table 4.7 (continued)

	SM13						SM62					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
18	−1	−1	−1	−1	−1.001	−1	−1	−1	−1	−1	−1	**0.621**
19	−1	−1	−1.378	−1.364	−1.358	−1.364	−1	−1	−1	−1	−1	−0.31
20	−1	−1	−1	−1	−1.007	−1	−1	−1	−1	−1	−1	**1.661**
21	−3.0419	−2.581	−1.005	−2.242	−2.387	−2.242	−8.17	−1	−5.57	−3.15	−3.15	−1
22	−1	−1.344	−1.847	−1	−1.008	−1	−1	−1	−1	−1	−1	−0.59
23	1.2501	1	1	1.9264	2.0603	1.9264	1	1	1	1	1	1
24	1	1	1	1	1.0025	1	1	1	1	1	1	0.739
25	1	1	1	1.3299	1.3916	1.3299	1	1	1	1	1	1
26	2.01441	4.5581	1.8048	1.1421	1.1893	1.1421	1	1	1	1	1	1
27	1	2.4909	1	2.9234	3.0668	2.9234	1.816	6.851	1.838	3.561	3.563	1
28	1	4.9665	3.9234	2.467	2.7198	2.467	1	1	1	1	1	1.025
29	1	3.9011	3.9192	1.716	1.7736	1.716	1	1	18.84	2.936	2.931	1.348
30	1.75679	1	1	1.521	1.451	1.521	1	1	1	1	1	0.698
31	1	1	1	1	1.0075	1	1	1	1	1	1	1
32	1.75897	3.2636	2.0724	1	1.07	1	1	1	1	9.375	9.374	2.081
33	1	1	1	1	1.0098	1	1	1	1	1	1	1
34	1.73337	2.7123	4.7237	2.2163	2.2187	2.2163	1	1	5.484	1	1	1
35	2.60771	5.1892	3.8988	2.5108	2.47	2.5108	3.811	4.8	17.76	8.51	8.516	1
36	1.28087	3.8579	3.9685	1.6623	1.8069	1.6623	28.42	33.91	40.14	29.65	29.65	4.07
37	1	1	1	1.212	1.3324	1.212	14.75	14.4	23.58	10.92	10.92	1.363

(continued)

Table 4.7 (continued)

	SM13						SM62					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
38	1	4.3202	1	1.7353	1.889	1.7353	6.153	9.102	18.97	12.71	12.7	1.511
39	2.86202	5.2042	3.6142	2.423	2.6232	2.423	20.69	14.34	32.03	18.41	18.4	2.372
40	3.31821	3.558	3.6471	2.3447	2.457	2.3447	18.71	23.22	24.95	19.51	19.51	1.892
41	1	1	1	1.9342	1.9685	1.9342	1	5.804	1	1	1.003	1.89
42	2.20623	1.538	2.3625	2.7576	2.8684	2.7576	1	1	1	1	1.005	1
43	3.84821	2.759	3.6544	2.7497	2.7975	2.7497	1	1	1	1	1	1.16
44	2.05459	4.9973	4.7246	2.8829	3.0048	2.8829	15.49	15.05	26.91	19.42	19.42	1
45	1.97218	1.9421	2.2446	2.0758	2.2204	2.0758	1.672	12.41	1.269	6.083	6.079	1.418
46	3.50017	1	4.6155	3.1236	3.2458	3.1236	11.87	14.62	10.05	8.567	8.564	1
47	1	3.6877	1	1	1.0535	1	8.084	6.944	12.97	11.38	11.38	2.674
48	3.37486	4.1485	4.7643	2.6852	2.8174	2.6852	5.883	12.59	7.316	10.88	10.88	2.256
49	1.85755	4.2427	4.2051	2.4029	2.4766	2.4029	1	1	10.27	3.697	3.707	1.4
50	1	3.0241	1	1.7589	1.8397	1.7589	1	1	1	1	1	0.257
51	2.3213	5.5432	5.4189	2.4294	2.474	2.4294	7.984	13.07	22.39	9.687	9.686	1.662
52	1	1	1	1	1.0545	1	1	1	1	1	1	**−1.27**
53	1	3.4922	3.5965	1.713	1.8054	1.713	7.132	18.29	14.66	3.91	3.906	2.061
54	2.77146	5.8568	3.7804	2.0148	2.1303	2.0148	3.632	9.426	7.023	5.085	5.084	1
55	1	1	1	1	1.0016	1	1	1	1	1	1	**−2.22**

(continued)

Table 4.7 (continued)

	SM13						SM62					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
56	1	1	1	1	1.0037	1	1	1	1	1	1	1
57	1	1	1	1	1.0076	1	1	8.496	6.481	7.979	7.979	1.642
58	1	1	2.2241	1.3484	1.3856	1.3484	1	1	1	1	1	**−1.39**
59	1	1	1	1.3924	1.5478	1.3924	1	1	1	1	1	1
60	1	1.3973	1.0506	1.0999	1.3636	1.0999	3.412	1.546	2.117	6.065	6.063	1.988
61	1	1	1	1	1.0113	1	7.538	11.41	23.73	14.96	14.96	1.338
62	1	3.0812	1	1.474	1.5096	1.474	1	5.844	1	1	1	1

4.5.2 *T-Tests of Mean's Difference Between the Tumor and Normal Subjects in SM13 and SM62*

Table 4.8 shows t-tests of the difference of two means between the tumor and normal subjects in SM13 and SM63. Some studies were looking for genes with larger t-values as oncogenes. However, their approaches are mistakes. In SM13, we confirm the 5% significance level if the t-value is 2.04 or the p-value is 0.0459. If we categorize 37 genes by t-values, we judge that the cancer averages of the top six genes are more significant than the normal averages, and the cancer averages of the bottom seven genes are smaller than normal averages. Both means of middle 24 genes are the same by a 5% significance level. In SM62, the cancer averages of the top three genes are more significant than the normal average, and the cancer averages of the bottom two genes are smaller than normal. Both means of middle 33 genes are the same by a 5% significance level. Each several genes of the upper six and three genes with positive t-values are oncogenes. Each several genes of lower three and two genes with negative t-values may be suppressor genes. Many researchers consider the middle genes are irrelevant to the oncogenes because of no difference between the two group's mean. However, if we drop these genes, the remaining genes cannot separate two classes. Thus, we conclude the following fact.

Fact5: The three categories of genes are necessary for cancer gene diagnosis, and the t-test or Welch test is not useful for identifying oncogenes. Many SMs show almost the same results.

Table 4.8 T-test of 37 genes of SM13 and 38 genes of SMS63

SM13 (37genes)			SM62 (38gene)		
Y	Difference	t-value	Y	Difference	t-value
X1060	0.893	5.375	X447	0.386	2.816
X1473	1.177	4.200	X397	0.509	2.422
X1808	0.819	3.224	X940	0.399	2.252
s13X23	0.409	3.143	X1156	0.383	1.951
X1048	0.720	2.803	X635	0.210	1.442
X329	0.519	2.482	X433	0.267	1.422
X733	0.373	1.863	X1195	0.173	1.375
X1123	0.330	1.709	X431	0.154	1.280
X252	0.241	1.453	X729	0.096	0.833
X223	0.189	0.846	X594	0.096	0.700
X171	0.089	0.608	X755	0.066	0.554
X263	0.058	0.199	X623	0.036	0.303
X1318	0.016	0.055	X891	0.023	0.192
X152	−0.003	−0.016	X598	0.013	0.130
X1755	−0.005	−0.021	X162	0.013	0.112

(continued)

Table 4.8 (continued)

SM13 (37genes)			SM62 (38gene)		
Y	Difference	t-value	Y	Difference	t-value
X1749	−0.033	−0.184	X333	0.013	0.102
X497	−0.045	−0.233	X750	0.010	0.081
X966	−0.046	−0.325	X455	0.011	0.075
X1434	−0.068	−0.335	X335	0.000	−0.004
X1214	−0.060	−0.393	X646	−0.015	−0.091
X439	−0.087	−0.479	X1045	−0.019	−0.116
X757	−0.167	−0.686	X637	−0.019	−0.182
X933	−0.181	−0.794	X1019	−0.033	−0.225
X1818	−0.167	−0.808	X605	−0.054	−0.329
X1291	−0.184	−0.954	X675	−0.049	−0.356
X1185	−0.202	−1.082	X673	−0.115	−0.512
X217	−0.197	−1.169	X530	−0.107	−0.529
X1819	−0.337	−1.463	X944	−0.103	−0.699
X404	−0.251	−1.738	X720	−0.111	−0.781
X883	−0.389	−1.910	X1037	−0.160	−0.966
X202	−0.359	−2.048	X641	−0.134	−1.074
X1246	−0.336	−2.105	X1595	−0.181	−1.223
X44	−0.324	−2.141	X445	−0.265	−1.286
X293	−0.423	−3.019	X811	−0.174	−1.295
X1394	−0.468	−3.065	X569	−0.198	−1.500
X617	−0.728	−3.078	X608	−0.186	−1.598
X1258	−1.208	−4.040	X645	−0.305	−2.335
			X772	−0.272	−2.406

Figure 4.3 is a four box–whisker plot of the four genes of SM13. “−1” of X2001 is normal and “1” corresponds to the tumor. The mean of the tumor class is larger than the normal class in x23. Next, the mean of the tumor class is smaller than the normal class in x44. Last two plots show the averages of the two genes are almost the same. Although not shown here, it is difficult to interpret the t-value of the two classes when there are significant outliers. We had better checked the Box-whisker plot for t-test.

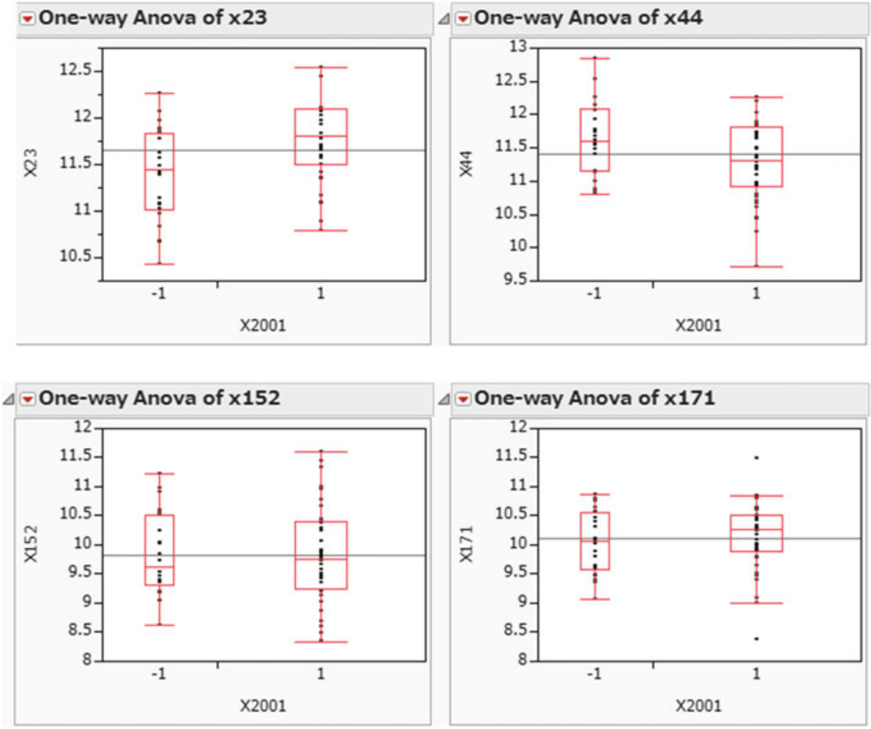


Fig. 4.3 Four box–whisker plots of the four genes of SM13 (X2001 = −1: Normal, X2001 = 1: Cancer)

4.5.3 *PCA and Cluster Analysis of SM13 and SM62*

4.5.3.1 PCA of SM13 and SM62

Figure 4.4 shows eigenvalue, scatter plot, and factor loading plots of PCA in SM13. “0” are 22 normal subjects and “1” are 40 cancer subjects. Although the two groups are separable entirely in SM13 and RatioSV of RipDS is about 30%, two classes in the SM13 overlap. Although the Prin1 and Prin2 can express large variations of data, it cannot indicate linear separable fact because two groups may have less data fluctuation. At the time of 2015, we thought that if the specialist considered the result of PCA, we expected they found something useful knowledge, but now we conclude that they cannot get anything.

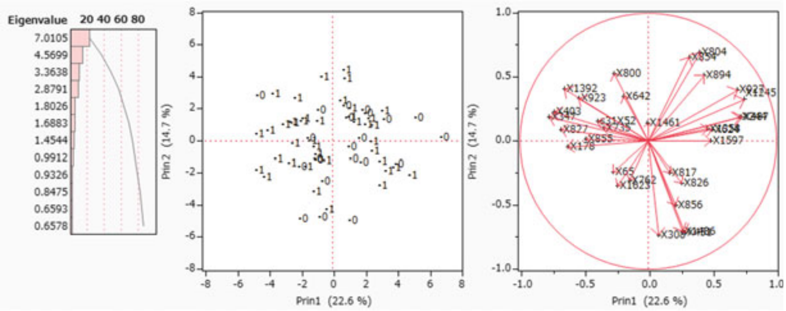


Fig. 4.4 Eigenvalue, scatter plot, and factor loading plots of SM13

Caution: In cancer gene diagnosis, never examine the results that do not show the linear separable facts.

Figure 4.5 shows three plots for SM62. It is almost the same as SM13, and every plot of all SMs shows the same results.

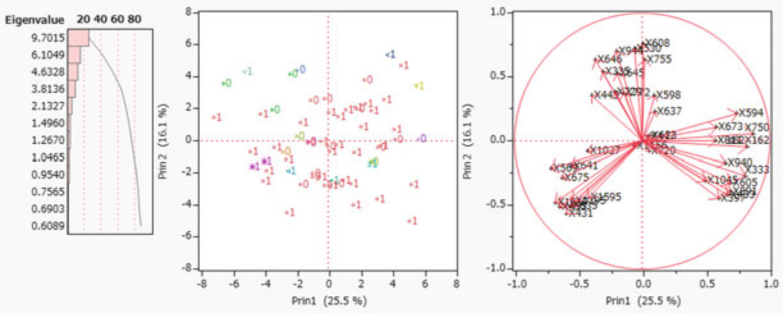


Fig. 4.5 Eigenvalues, scatter plots, and factor loadings

4.5.3.2 Hierarchical Cluster Analysis of SM13

Six research groups rarely used discriminant analysis, and used cluster analysis typified by SOM. Figure 4.6 shows the results of SM13 by the nearest neighbor method used in several studies of gene analysis. The cases painted black on the left are 22 normal cases, and the other cases are 40 tumor cases. The two classes become 22 groups from top to bottom alternately. The bottom left cell corresponds the 59th patient and x23. This cell becomes one cluster on both dendrograms of the subject and gene finally. This figure tells us the cluster analysis of SMs is useless as same as PCA. However, the 11 normal groups and the 11 tumor groups may be helpful to discover new subclasses of cancer. The square color map in the middle is the mesh of cases and genes. The red-colored mesh indicates that the relationship between the case and the gene is strong, and the blue color is weak in the color hierarchy. The right is the dendrogram of the case. The nearest neighbor method is clustered sequentially from the closest distance cases. Thus, we do not recommend this method in general. The bottom dendrogram is a variable dendrogram of 37 genes. The X23 at the left end is clustered at the last after the other 36 genes are sequentially clustered. The 59th patient in the bottom is classified at the last after the other 61 cases are sequentially clusters. In general, although we do not recommend this method, some Japanese gene researchers used this method.

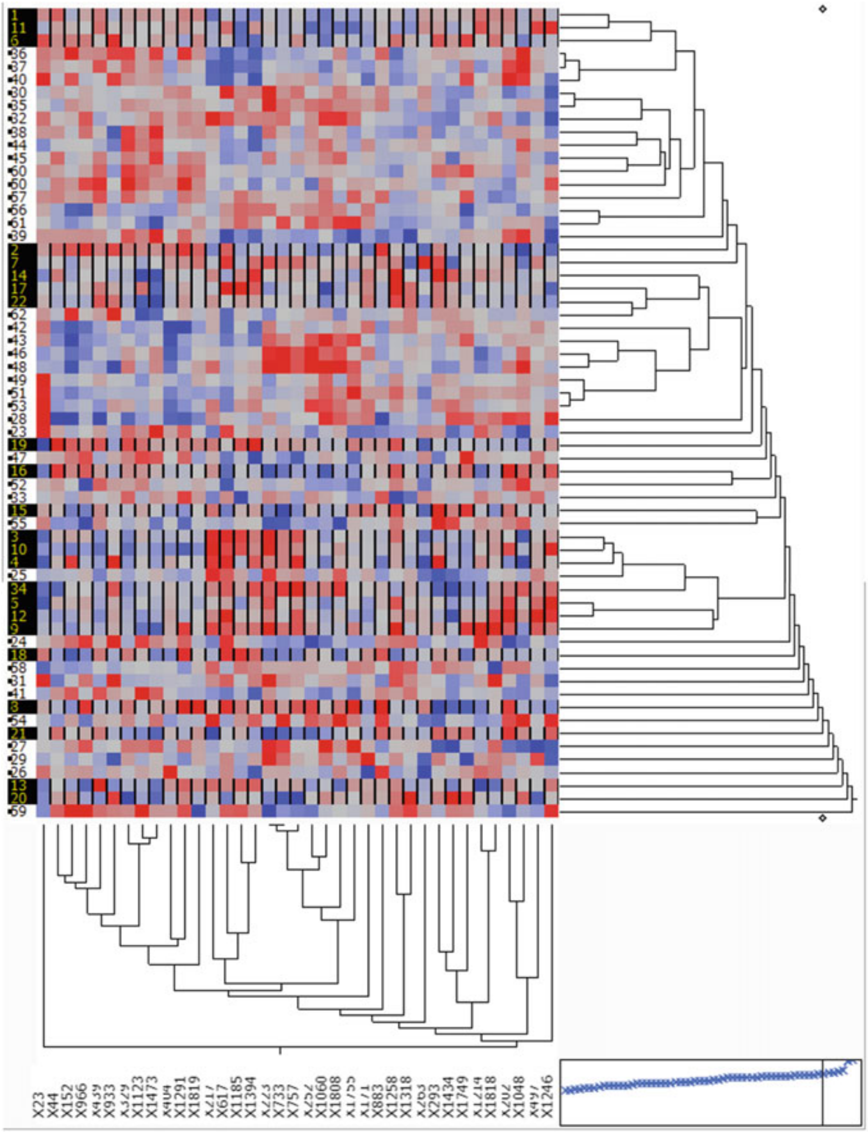


Fig. 4.6 Nearest neighbor method of SM13

Figure 4.7 is Ward cluster analysis of SM13. There are the 13 normal groups and the 13 tumor groups as same as the nearest neighbor method, but the tendency of both dendrograms are different. In the recent gene analysis, many kinds of research use the cluster analysis. Although statistical discriminant functions are useless at all, the cluster analysis offers various results by the combination of different cluster methods and distances used for clustering. If medical researchers can explain their medical conclusions well as a result of cluster analysis, there will be no problem.

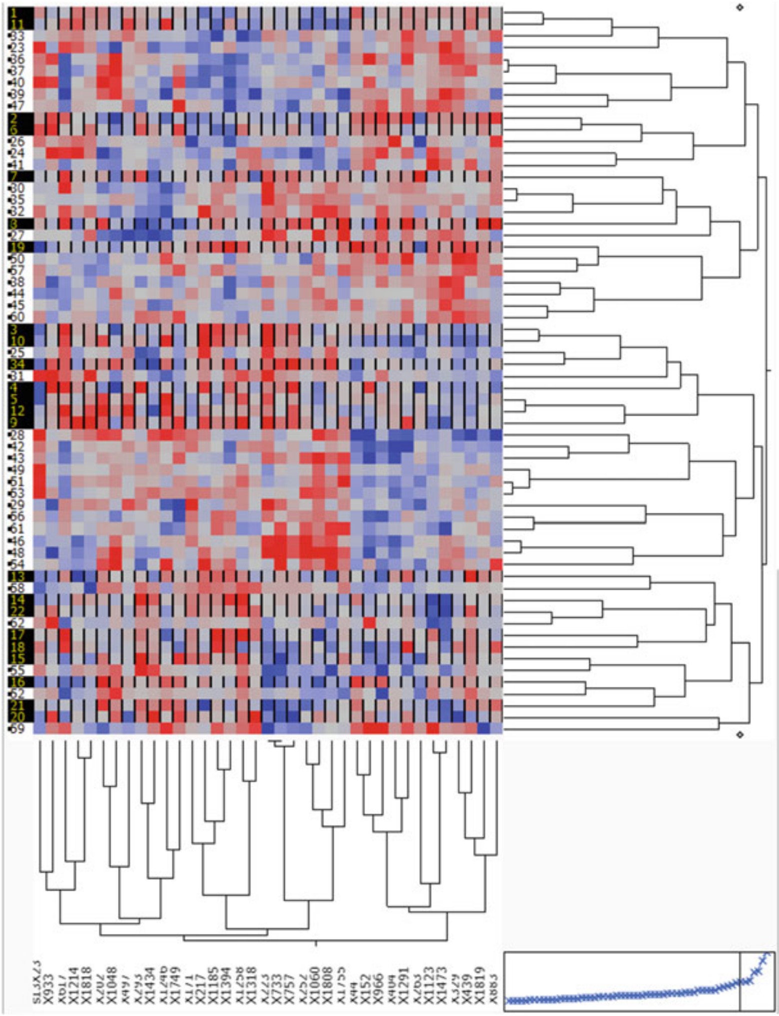


Fig. 4.7 Ward cluster of SM13

In particular, Alon et al. examined 6,500 genes by SOC and judged 2,000 genes as cancer genes. Method2 and RIP decompose the 2,000 genes into 62 SMs containing 1,968 genes using the Method2 by us, although we are not specialists for gene analysis of cancer at all. The number of genes included in the noise is only 32 genes. Thus, because two different research approaches are almost the same, we consider both approaches validate each other. Moreover, LINGO Program4 decomposes the 2,000 genes into 130 BGS, and we omit only five genes as noise in Chap. 2. Because LINGO Program4 finds 130 BGSs by the combination with manual work, these five genes may be the operation miss. The critical point is that 2,000 genes medically discovered could be divided into 130 BGSs. We think that this result can be considered to guarantee to some extent the validity of both approaches by Alon and us in 2017.

4.5.4 Examination of Correlation of 37 Genes Included in SM13

We examine the 666 correlations of 37 genes included in SM13. Other 61 SMs are almost the same results. Table 4.9 shows the 666 correlation coefficients were sorted by descending order. There were 83 positive correlations from SN1 = 1 to SN1 = 83 at 1% significant level. There were 53 negative correlations from SN1 = 614 to SN1 = 666 at 1% significant level. The 530 correlations are uncorrelated. That is, unlike the correlation between RipDSs and HsvmDSs, about 80% of the genes contained in SM are uncorrelated, and the rest 20% are positive and negative correlations. This result is one of the facts that RipDSs and HsvmDSs are valid signals. In other words, it does not become a signal with the gene alone, but when RIP, Revised LP-OLDF, and H-SVM discriminate these genes, it becomes a signal.

Table 4.9 666 correlations of 37 genes included in SM13

SN1	SN2	Var1	Versus Var2	Corr.	2.5%	97.5%	p-value
1	1	X733	X223	0.757	0.626	0.847	0.000
2	2	X757	X223	0.717	0.569	0.820	0.000
3	3	X757	X733	0.714	0.565	0.818	0.000
4	4	**X1473**	X1123	0.681	0.520	0.796	0.000
5	5	**X1473**	X329	0.629	0.450	0.759	0.000
6	6	X1818	X1214	0.622	0.441	0.754	0.000
7	7	**X1060**	X252	0.608	0.422	0.745	0.000
8	8	X1394	X1185	0.598	0.409	0.737	0.000

(continued)

Table 4.9 (continued)

SN1	SN2	Var1	Versus Var2	Corr.	2.5%	97.5%	p-value
9	9	X733	X252	0.597	0.408	0.737	0.000
10	10	X152	X44	0.585	0.392	0.728	0.000
11	11	X966	X152	0.565	0.367	0.714	0.000
12	12	X439	X44	0.563	0.365	0.713	0.000
13	13	X1318	X1258	0.563	0.364	0.713	0.000
14	14	**X1048**	X202	0.557	0.357	0.708	0.000
15	15	**X1808**	X252	0.557	0.356	0.708	0.000
16	16	X933	X152	0.554	0.354	0.706	0.000
17	17	**X1808**	X733	0.551	0.349	0.704	0.000
18	18	X1755	X252	0.548	0.345	0.702	0.000
19	19	X439	X329	0.547	0.345	0.701	0.000
20	20	X1394	X617	0.536	0.331	0.693	0.000
21	21	X1291	X404	0.532	0.325	0.690	0.000
22	22	**X1808**	**X1060**	0.527	0.319	0.687	0.000
23	23	X1291	X44	0.521	0.312	0.682	0.000
24	24	X1434	X293	0.521	0.312	0.682	0.000
25	25	X1291	X966	0.520	0.311	0.681	0.000
26	26	**X1808**	X223	0.514	0.303	0.677	0.000
27	27	X1291	**X329**	0.501	0.287	0.667	0.000
28	28	X966	X44	0.491	0.275	0.660	0.000
29	29	X1291	X439	0.490	0.274	0.659	0.000
30	30	X1185	X217	0.482	0.264	0.653	0.000
31	31	X1123	**X329**	0.478	0.259	0.650	0.000
32	32	X1394	X217	0.478	0.259	0.650	0.000
33	33	X1755	**X1060**	0.478	0.259	0.650	0.000
34	34	X966	X404	0.478	0.259	0.650	0.000
35	35	X404	X44	0.464	0.242	0.640	0.000
36	36	X1123	X152	0.453	0.230	0.632	0.000
37	37	X1123	X44	0.444	0.218	0.624	0.000
38	38	X1819	**X329**	0.443	0.218	0.624	0.000
39	39	X1749	X1434	0.440	0.214	0.622	0.000
40	40	X404	X152	0.440	0.214	0.621	0.000
41	41	**X1473**	X966	0.439	0.213	0.621	0.000
42	42	X252	X223	0.437	0.210	0.619	0.000
43	43	**X1060**	X733	0.433	0.206	0.616	0.000
44	44	X497	X202	0.427	0.198	0.612	0.001

(continued)

Table 4.9 (continued)

SN1	SN2	Var1	Versus Var2	Corr.	2.5%	97.5%	p-value
45	45	X1755	X1185	0.427	0.198	0.611	0.001
46	46	**X329**	X44	0.423	0.193	0.608	0.001
47	47	X1291	X152	0.418	0.187	0.604	0.001
48	48	X1819	X1291	0.404	0.172	0.594	0.001
49	49	X617	X404	0.402	0.169	0.592	0.001
50	50	X1185	X171	0.394	0.160	0.586	0.002
51	51	X757	X617	0.385	0.150	0.579	0.002
52	52	X1819	X439	0.385	0.149	0.579	0.002
53	53	X757	X217	0.382	0.146	0.577	0.002
54	54	X1755	X1394	0.380	0.144	0.575	0.002
55	55	X252	X217	0.379	0.143	0.575	0.002
56	56	**X1060**	X171	0.378	0.142	0.574	0.002
57	57	X217	X171	0.368	0.130	0.566	0.003
58	58	X1214	X293	0.366	0.128	0.565	0.003
59	59	X1123	X966	0.366	0.128	0.564	0.003
60	60	**X1394**	X757	0.363	0.125	0.562	0.004
61	61	X1819	X883	0.362	0.123	0.561	0.004
62	62	X1258	X1185	0.360	0.121	0.559	0.004
63	63	X757	X252	0.359	0.120	0.559	0.004
64	64	X1394	X1258	0.359	0.120	0.559	0.004
65	65	X883	X404	0.354	0.114	0.554	0.005
66	66	X966	**X329**	0.350	0.109	0.551	0.005
67	67	X733	X217	0.349	0.109	0.551	0.005
68	68	X223	X217	0.348	0.107	0.550	0.006
69	69	X1291	X883	0.344	0.103	0.547	0.006
70	70	**X1394**	X171	0.343	0.102	0.546	0.006
71	71	X1258	X217	0.341	0.099	0.544	0.007
72	72	X1291	X263	0.339	0.098	0.543	0.007
73	73	X1258	X293	0.339	0.097	0.543	0.007
74	74	X1755	X733	0.339	0.097	0.543	0.007
75	75	X1434	X1318	0.336	0.094	0.541	0.008
76	76	X1749	X1048	0.336	0.094	0.540	0.008
77	77	X1185	X617	0.332	0.090	0.537	0.008
78	78	**X1060**	X223	0.332	0.089	0.537	0.008
79	79	**X1808**	X1755	0.331	0.089	0.537	0.009
80	80	X1755	X757	0.328	0.085	0.534	0.009

(continued)

Table 4.9 (continued)

SN1	SN2	Var1	Versus Var2	Corr.	2.5%	97.5%	p-value
81	81	X1818	X617	0.325	0.082	0.532	0.010
82	82	X1258	X171	0.324	0.080	0.531	0.010
83	83	X1123	X439	0.323	0.080	0.530	0.010
84	84	X883	X439	0.320	0.076	0.527	0.011
–	–	–	–	–	–	–	–
613	169	X1818	X1048	0.322	0.529	0.078	0.011
614	170	X1048	X404	– 0.323	– 0.530	– 0.080	0.010
615	171	X329	X217	– 0.331	– 0.536	– 0.088	0.009
616	172	X1123	X252	– 0.333	– 0.538	– 0.091	0.008
617	173	X1749	X223	– 0.333	– 0.538	– 0.091	0.008
618	174	X966	X202	– 0.338	– 0.542	– 0.096	0.007
619	175	X1246	**s13X23**	– 0.339	– 0.543	– 0.098	0.007
620	176	X1060	X404	– 0.339	– 0.543	– 0.098	0.007
621	177	X1394	X1123	– 0.343	– 0.546	– 0.103	0.006
622	178	X966	X171	– 0.348	– 0.550	– 0.107	0.006
623	179	X1749	X757	– 0.349	– 0.551	– 0.109	0.005
624	180	X1060	X152	– 0.351	– 0.552	– 0.111	0.005
625	181	X171	X152	– 0.351	– 0.553	– 0.111	0.005
626	182	X1819	X1246	– 0.354	– 0.555	– 0.114	0.005
627	183	X1394	**s13X23**	– 0.357	– 0.557	– 0.118	0.004
628	184	**X1060**	X439	– 0.362	– 0.561	– 0.124	0.004
629	185	**X1473**	X1434	– 0.363	– 0.562	– 0.125	0.004
630	186	X1214	**X1048**	– 0.367	– 0.565	– 0.129	0.003
631	187	X1755	**X44**	– 0.367	– 0.565	– 0.129	0.003
632	188	**X1060**	X933	– 0.368	– 0.566	– 0.131	0.003
633	189	X202	X152	– 0.369	– 0.566	– 0.131	0.003
634	190	X1318	X497	– 0.370	– 0.567	– 0.133	0.003
635	191	X439	X252	– 0.371	– 0.568	– 0.134	0.003
636	192	X1749	X1394	– 0.375	– 0.571	– 0.138	0.003
637	193	X1318	**X329**	– 0.377	– 0.573	– 0.141	0.002
638	194	X404	X252	– 0.387	– 0.581	– 0.152	0.002
639	195	X217	**s13X23**	– 0.392	– 0.584	– 0.158	0.002
640	196	X293	X223	– 0.394	– 0.586	– 0.160	0.002
641	197	X1749	X733	– 0.398	– 0.589	– 0.164	0.001
642	198	X1123	X202	– 0.399	– 0.590	– 0.166	0.001
643	199	X1291	X202	– 0.399	– 0.590	– 0.166	0.001
644	200	**X1060**	X293	– 0.400	– 0.591	– 0.167	0.001

(continued)

Table 4.9 (continued)

SN1	SN2	Var1	Versus Var2	Corr.	2.5%	97.5%	p-value
645	201	**X1473**	X293	– 0.404	– 0.594	– 0.172	0.001
646	202	X252	X152	– 0.409	– 0.598	– 0.177	0.001
647	203	X1434	X617	– 0.414	– 0.602	– 0.184	0.001
648	204	**X1060**	X44	– 0.419	– 0.605	– 0.189	0.001
649	205	X1258	**X329**	– 0.430	– 0.613	– 0.201	0.000
650	206	X1818	X439	– 0.430	– 0.613	– 0.201	0.000
651	207	**X1473**	X1258	– 0.431	– 0.615	– 0.203	0.000
652	208	X439	X217	– 0.436	– 0.618	– 0.209	0.000
653	209	X733	X293	– 0.444	– 0.624	– 0.218	0.000
654	210	**X1473**	X1185	– 0.446	– 0.626	– 0.221	0.000
655	211	X1819	X1318	– 0.457	– 0.634	– 0.234	0.000
656	212	X252	X44	– 0.465	– 0.640	– 0.243	0.000
657	213	X439	X223	– 0.470	– 0.644	– 0.249	0.000
658	214	**X1473**	X1394	– 0.475	– 0.648	– 0.256	0.000
659	215	**X1808**	X293	– 0.516	– 0.679	– 0.306	0.000
660	216	X1749	X617	– 0.519	– 0.681	– 0.310	0.000
661	217	**X1808**	X439	– 0.524	– 0.684	– 0.315	0.000
662	218	X404	X171	– 0.529	– 0.688	– 0.321	0.000
663	219	X1434	X757	– 0.531	– 0.690	– 0.325	0.000
664	220	X1434	X733	– 0.565	– 0.714	– 0.367	0.000
665	221	X1434	X223	– 0.593	– 0.734	– 0.403	0.000
666	222	**X1048**	X617	– 0.647	– 0.772	– 0.474	0.000

4.6 The Reason Why Standard Statistical Methods Could not Find Fact6

Because Alon's microarray consists of only 2,000 genes and the smallest among six microarrays, we analyze it by every version up of LINGO. In 2016, RIP finds 64 SMs (1,999 genes), and Chap. 2 introduces the results. In 2017, RIP finds 56 SMs (1,999 genes), and Chap. 3 introduces the results. In 2018, RIP finds 62 SMs (1,968 genes), and this chapter introduces the results by the different approaches. Thus, the RatioSV of PCA summarizes the malignant indicators of the 62 RatioSVs. In this section, because statistical methods could not obtain useful results from SM analysis, we make three signal data made by RipDSs, LpDSs and HsvmDSs those consist of 62 cases (subjects) and 62 DSs as variables. After that, Ward cluster and PCA analyze the signal data and those transposed data. We obtain almost the same results as Chap. 3. Thus, we discuss the reason why standard statistical methods could not find Fact6 and statistical discriminant functions could not find Fact3.

Fact6: Statistical methods cannot find the linear separable facts by all SMs. However, when we created new data with DS discriminated by RIP, Revised LP-OLDF and H-SVM, we obtained a remarkably good result, so we decided to call this data as signal data. We need to discuss the big problem of cancer gene analysis. Statistical methods, except for logistic regression, could not find the linear separable facts. We already had the clues of above reason using the common data. PCA analysis of the modified linear separable CPD data using 19 variables shows that the scatter plot on the Prin1 and the Prin2 cannot separate the two classes. The t-value of the Prin8 was the maximum value that means two classes have the most massive difference between the average. This fact means "Prin1 and Prin2 show a large variance of data, but irrelevant to linearly separable fact."

Figure 4.8 is two scatter plots of SM18 on such as (Prin1 vs. Prin2) and (Prin1 vs. Prin3). Both plots show two classes overlap. The line segment represents the RipDS13. Among the data variations, the variation in RipDS18 is small. PCA cannot find the critical information of LSD because its fluctuation is small. The DS that divides the two groups offers more useful information than PCA. Four figures from Figs. 4.4. 4.5, 4.6 and 4.7 show the obscure reason. Figure 4.8 is the direct explanation for this problem. Statisticians and statistical users should abandon the expectation that the variation of data can catch the LSD phenomena. If we analyze with high-dimensional PCA, the line segment of DS will be smaller than Fig. 4.8.

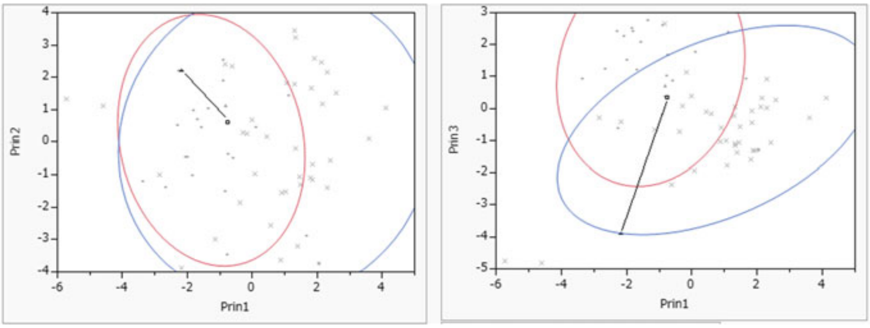


Fig. 4.8 Two scatter plots of SM such as (Prin1 vs. Prin2) and (Prin1 vs. Prin3)

4.7 Another Problem Suggested by Linus Schrage

(1) **An Important Advice by Linus Schrage about Fact3**

When we told Linus Schrage about Fact3, he gave us advice on high-dimensional data (small n and large p). He is the Emeritus professor of Chicago Booth and the founder of LINDO Systems Inc. Most MP-researchers know the high-dimensional linear hyperplane discriminates two classes of the random number data (small n and large p data) instead of the microarray. Although the probability is few, we must be aware of this fact. About this problem, his mail is as follow:

> With 2,000 explanatory variables but only 62 observations, there should be many separating hyperplanes. We have not derived the relevant probabilities, but if we have a lot more variables than observations, then there should be many separating hyperplanes, even if the explanatory variables are simply random numbers with no relationship to the dependent variable (e.g., have cancer or do not have cancer). So, a hyperplane that uses only one or two variables and misclassifies one or two observations is more likely to be the correct conclusion than a hyperplane that uses more than 62 variables and has no misclassifications, i.e., completely separates the data.

This advice is general knowledge of MP society. Keep this advice in mind; we have been looking for many facts to clear up his doubt. This section introduces our following findings:

(1) RIP, Revised LP-OLDF and H-SVM discriminate two classes by SVs. These SVs fix many cases on these SVs shown in Tables 4.2 and 4.4. However, when divided into many SMs, the proportion of cases on SV became small, and the cases spread widely in DS. We conclude these results show the specific feature of the microarrays. That is, it is not possible to divide the random number data into meaningful SMs and obtain the same result.

(2) Even if we can distinguish two groups with high-dimensional random number data, we cannot create random number data with a unique structure similar to the microarray described in (1). That is, even if the high-dimensional random number data is MNM = 0, it cannot split into many SMs of n dimension or less. Moreover, they cannot show almost the same result.

(2) **First Advice for OLDF by Linus Schrage**

When we started the research of three OLDFs in 1997, Linus sent us a list of papers about MP-based discriminant functions because we did not survey previous research. We were shocked by Stam (1997). He summarized about 200 papers and confessed honestly "Why have statisticians rarely used Lp-norm methods?" in his paper. We realized Linus suggested us Stam's paper declared the end of our research theme. "Lp-norm" is an attractive study that comprehensively thinks various regression and discriminant models. However, there are five problems in discriminant analysis, and NM is useless. To solve five problems and develop MNM instead of NM is more important than other models proposed by 200 types of research. Thus, RIP and Revised LP-OLDF can find SMs that opens the new frontier of cancer gene diagnosis. Moreover, in cancer gene analysis (Problem5), it is most important that the microarrays are LSD (Fact3). However, Japanese famous gene researchers told us that NIH decided to terminate the oncogene research using microarray after publishing six papers and advised us we had better terminated our research in 2016. Thus, the medical specialists in Japan do not pay their attention to our research.

(3) **Second Advice**
In 2012, I presented my paper entitled "Beyond Fisher's Discriminant Analysis" at the Informs held in Chicago. In the presentation, I said "I found a new continent like Columbus," but no one laughed. After the presentation, he advised me not to use such the radical title.
Linus, thank you for many pieces of advice.

4.8 Comparison of Our Research with iPS Cell Research and Problem6

In this book, we do not consider the relationship between SM and BGS in detail. In Chap. 2, we evaluated 64 SMs and 130 BGSs by RatioSV. Moreover, the maximum RatioSV of BGS was less than 1%, and we concluded that BGSs were useless for cancer gene diagnosis. It is necessary to confirm this fact with the remaining five microarrays. Although Method2 finds many SMs and BGS, we must classify these into medically meaningful groups and think about ways useful for cancer gene diagnosis. We want to collaborate on these issues with medical researchers in the future.

In this section, we compare our research and iPS research. Professor Yamanaka selected 100 genes among 30,000 human genes that are activated only in mouse ES cells. Also, he narrowed it to 24 genes. Moreover, he instructed Dr. Takahashi to find a set of genes to make iPS cells. Unlike common sense in genetics, Dr. Takahashi made iPS cells using 24 genes. Dr. Yamanaka thought that his experiment was wrong because he graduated from the Faculty of Engineering and was lacking a common biological sense. However, it turned out that there was no mistake. Next, he narrowed down beneficial genes from 24. He also proposed a method beyond common sense. It is a method that performs a backward selection method from 24 genes. If he cannot make iPS cells by 23 genes without one gene, that gene is necessary to generate iPS cells. After several cell cultures, he found four genes, and Yamanaka team opened the door to iPS cell research.

(1) His approach is an application of the backward variable selection method, which is the same as our method of finding BGS (LINGO Program4). The stepwise selection methods use the increase/decrease of the deviation sum of squares. In our study, we have 1/0 criteria that gene subspace is LSD or not. The iPS cell research depends on 1/0 criterion whether iPS cell masses are formed or not.

(2) Three genes dropping one gene from 4 genes will not create iPS cells. In our research, we call it the smallest SM (basic gene set, BGS), 24 genes are one of the SMs because they make iPS cells.

(3) The four genes contain the oncogene C-Myc, and many researchers surprised. However, it was found to be useful for the proliferation of iPS cells. Dr. Nakagawa struggled to replace it with L-Myc. The existence of L-Myc suggests that there is a possibility that there is another BGS that consists of other genes with L-Myc.

(4) In our study, there are several SM and BGS. Remove four genes containing C-Myc from 24 genes. With these 20 genes, for example, other iPS genes containing or not containing L-Myc may be found.

A doctor at Harvard University was able to find the human iPS gene first by adding one gene to four genes including C-Myc. He found SM to produce iPS cell. Even if we add an arbitrary gene to the BGS, our research guarantees the new SM is LSD. However, it is self-evident that even if we add any genes to 24 genes, not all can make iPS cells. There is no doubt that there are limit that iPS cells can be made, perhaps between 24 and 100. This point is different from our research.

4.9 Conclusion

This chapter performs the following innovative verifications, and good results were obtained.

(1) The defect of SM found by Revised LP-OLDF
RIP finds microarrays are LSD and can decompose microarray into many SMs. Then, we developed Method2 and created LINGO Program3. We extend Program3 for other five kinds of LDFs. Because Revised LP-OLDF finds SM like RIP and computing time is shorter than IP, it is useful for many researchers. Moreover, we evaluate the signal subspaces of the union of all SMs and the noise subspaces found by the RIP and Revised LP-OLDF. However, we find that Revised LP-OLDF left SM in the noise subspace in Table 4.3. We think that this is an influence of Problem1.

(2) Subjects on SVs and Outliers
Table 4.2 shows the all 22 normal subjects on SVs and 39 cancer subjects on SVs. Only one cancer patient' DS is 3.2. We denote the "Outlier = 0/1" in Table 4.2. When RIP discriminates two classes in 1968 genes included in 62 SMs, all normal subjects locate on SV = −1, and the 39 cancer patients locate on SV = 1. Only one cancer subject becomes an outlier. We tried to find answers to Linus's advice. Two outliers of SM13 with 37 genes and SM62 with 38 genes are "7/18 and 7/17," respectively. These facts show: (1) For SM13, the 15 normal subjects locate on SV = −1 and the 22 cancer patients locate on SV = 1. (2) For SM62, the 15 normal subjects locate on SV = −1, and the 23 cancer patients locate on SV = 1. Moreover, MP-based LDFs and logistic

regression can discriminate 62 SMs correctly. On the other hand, these LDFs cannot discriminate between two classes made by random numbers correctly. Many statisticians have studied microarray data as a new field of "big data analysis or high-dimensional data analysis." Aoshima and Yata found that microarrays analyzed by us distributed on two different spheres in gene space. Each distribution is a chi-square distribution from the center of the sphere. Their results are probably the same as our results in Table 4.2. We think that the mode of the chi-square distribution corresponds to the two SVs, and the outlier corresponds to the tail of the distribution.

(3) Table 4.8 shows t-tests of mean's difference between the tumor and normal subjects in SM13 and SM63. Because the gene sets included in 63 SMs can discriminate two classes correctly, these 63 gene sets are the oncogenes. We showed both t-value ranges of SM13 and SM63 are from negative values, almost zero and positive values. The genes with negative values are the suppressor of cancer. The genes having values of almost zero are thought to be involved in canceration of patients in combination with other genes. Medically about 100 representative oncogenes have been found, but these serve as the core role of any of the 63 SMs, but they alone cannot distinguish the two groups correctly. That is, SMs that do not contain 100 oncogenes at all may be an oncogene that has not been discussed yet, for example, may be related to metabolism. This argument is a future research teme which should be considered in BGS.

Future Research: We must examine the different roles of all BGSs that are one of Problem6.

(4) Table 4.6 shows 1,891 pairs of correlations of 62 RipDSs. The range of correlations is [0.332, 0.865]. The fact that all correlations are positive is a feature of signal space.

(5) Why could not researchers solve the cancer gene analysis from 1970? There are many reasons as follows.

- Because statistical discriminant functions cannot discriminate LSD theoretically, these discriminant functions cannot find microarrays are LSD.
- Although only H-SVM and RIP can discriminate LSD theoretically, H-SVM cannot decompose microarrays into many SMs.
- The fluctuation of two classes is too small compared with the variation of microarray data that is noise.

Future Work: Above fact explains the reason why researchers could not find microarrays were LSD since 1970.

Caution: In cancer gene diagnosis, never examine the results that do not show the linear separable facts.

New Fact: Although we could not find meaningful results by analyzing SM, the signal data created by RDSs, LpDSs, and HsvmDSs as variables brings the surprising results. The transposed data indicates many outliers that may be the new subclasses of cancer pointed by Golub.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA, 96(12):6745–6750

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

Fisher RA (1956) Statistical methods and statistical inference. Hafner Publishing Co., New Zealand

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2016) New theory of discriminant analysis after R. Springer, Fisher

Stam A (1997) Non-traditional approaches to statistical classification: some perspectives on Lp-norm methods. Ann Oper Res 74:1–36

Chapter 5
Cancer Gene Diagnosis of Golub et al. Microarray

Abstract Golub microarray consists of 72 patients and 7,129 genes. They analyzed the microarray by various statistical methods. For example, they analyzed "marker" genes having the highest correlation with the target class-by-class separation statistics (signal-to-noise ratio), weighted votes, and SOM. Mainly, discriminant analysis is the most proper method to identify oncogenes. However, because the statistical discriminant analysis was useless at all, medical researchers had developed many methods. Our theory shows that six microarrays are LSD (MNM = 0). Method2 can decompose the microarray into many Small Matryoshka (SM) those are LSD. Then, by analyzing SM, we achieved cancer gene diagnosis by malignancy indexes. If Golub et al. validate our results, cancer gene diagnosis will be more improved. Method2 already obtained the different sets of SM in Chap. 2. In 2018, we change the number of iterations of RIP and Revised LP-OLDF in Method2 and decided the proper number of iterations as same as Alon's microarray in Chap. 4. We obtained SM by those iteration numbers. We examined the signal data made by RIP discriminant scores (RipDSs). We confirm the Revised LP-OLDF cannot find all SMs as same as Alon's microarray. Thus, we analyze only 179 SMs obtained by the RIP and examine the correlation coefficient of 179 RipDSs. We compare RatioSV of six MP-based LDFs and NM of statistical discriminant function. Then, the cluster analysis and PCA analyze signal data made by RIP and H-SVM. We propose the possibility of cancer gene diagnosis such as malignancy indexes. We propose how to find new subclasses of cancer pointed out by Golub et al. (Science 286(5439): 531–537, 1999).

Keywords Golub microarray · Cancer gene diagnosis · Malignancy indicators · Small Matryoshka (SM) · RatioSV · RipDSs · LpDSs · HsvmDSs · Signal data

Thanks to Golub et al.

Golub microarray (7,129 genes) consists of 25 acute myeloid leukemia (AML) patients (class1) and 47 acute lymphoblastic leukemia (ALL) patients (class2). Golub et al. (1999) analyzed 72 cases by various statistical methods. They analyzed "marker" genes with the highest correlation with the target class-by-class separation statistics (signal-to-noise ratio), weighted votes, and self-organizing map (SOM). Shipp et al. (2002) used these methods, also. Shipp used SVM (Vapnik 1995). Mainly,

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_5

discriminant analysis is the most appropriate method to identify oncogenes from the microarray. However, because the statistical discriminant analysis was useless at all, medical researchers had no choice but to develop many methods for cancer gene analysis. Our theory shows that six microarrays are the linearly separable data (LSD) and the minimum number of misclassifications (MNMs) = 0. The Matryoshka feature selection method (Method2) can decompose microarray into many sets of Small Matryoshka (SM). Then, by analyzing SM, cancer gene diagnosis typified by malignancy indicators was proposed. If Golub and other researchers validate our results, cancer gene diagnosis will be more improved. They used leave-one-out (LOO) which was developed in the age of poor computing environment to verify their outcome (Lachenbruch and Mickey 1968). Shinmura (2010) proposed the 100-fold cross-validation for the small sample (Method1). If Golub and others used RIP or the hard margin SVM (H-SVM), they could find their microarray was LSD, and the number of misclassifications (NM) is 0. It is extraordinary why there are no researches that the two classes are completely separable in the high-dimensional microarray. If researchers manage two classes well, probably the other microarrays may be LSD, also. This fact (MNM = 0) is the most important in the cancer gene analysis (Fact3). In our research, we consider that gene subspace with MNM = 0 defines signal subspace and gene subspace with MNM >=1 defines noise subspace at first. However, they evaluate various methods with a 2 * 2 contingency table. For example, in weighted voting, six patients out of ALL patients with moderate malignancy are incorrectly identified as low-grade AML. Also, they are examining the survival rate of Kaplan–Meier. From these facts, historically, because the discriminant functions based on the variance–covariance matrix were useless for cancer gene analysis, we believe that they originally developed the several methods. However, if they discriminate the microarray by RIP or H-SVM, they find it is LSD and obtain other simple results.

We thank **Golub** for providing excellent data. Below, we will quote their abstract:

> Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction) Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between AML and ALL without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

5.1 Introduction

Chapter 1 outlined the New Theory of Discriminant Analysis After R. Fisher (the theory) and explained the success of cancer gene analysis (Shinmura 2016a, b). Also, we explained why Revised IP-OLDF (RIP) and Revised LP-OLDF solved unresolved cancer gene analysis (Problem5). Chapter 2 outlined the cancer gene diagnosis using all SMs of six microarrays found by the RIP in 2016. Chapter 3 outlined RIP and Revised LP-OLDF discriminant Alon's microarray (1999) by changing the iteration number of LINGO Program3 (Schrage 2006). Chapter 4 discriminates Alon's microarray in 2018 and obtains two SMs by the RIP and Revised LP-OLDF. We challenge three new research themes.

This chapter confirms three research themes proposed in Chap. 4 using Golub microarray. Section "Thanks to Golub et al." introduces Golub et al. research and microarray. Section 5.1 introduces the overviews from Chap. 1 to Chap. 5. Section 5.2 validates two different types of SMs found by the RIP and Revised LP-OLDF. These SMs are obtained by searching the proper number of iterations of both OLDFs. We discriminate the signal and noise subspaces obtained by Revised LP-OLDF and RIP. We confirm the same defect of Revised LP-OLDF that cannot choose all SMs as signal subspace. Section 5.3 analyzes 179 SMs by RIP. After we evaluate 179 SMs by six MP-based LDFs and discriminant functions, we examine the 15,931 correlations of 179 RipDSs. Section 5.4 verifies SM3 with the maximum RatioSV and SM179 with the minimum RatioSV. Moreover, SM3 and SM179 are evaluated by RatioSV, NMs, and t-test. Section 5.4.3 introduces the relation of BGS and Yamanaka's Four Genes of iPS research. Section 5.5 examines the signal data made by 179 RipDSs and discusses the reason why standard statistical methods could not find the linear separable facts proposed in Chap. 4. LINGO (Schrage 2006) decomposes Golub microarray into many SMs and opens a new frontier of cancer gene analysis. JMP (Sall et al. 2004) analyzes all SMs and offers cancer gene diagnosis. Shinmura (2016a, 2017, 2018a, b) relate to this Chapter.

5.2 Validation of SM Found by the RIP and Revised LP-OLDF

In Chap. 3, we increase the number of iterations from 1 and decide to select the one that can obtain the same number of SMs consecutively. After Chap. 5 to Chap. 9, we choose all SMs by this approach and confirm that the results are almost the same as the results in Chap. 2. Although we develop LINGO Program3 for a RIP at first, Revised LP-OLDF can decompose microarray into smaller SMs and the calculation time is faster than RIP. Thus, we consider replacing the SMs obtained by RIP to those of Revised LP-OLDF. However, we find the defect of Revised LP-OLDF that cannot find all SMs from the microarray in Chap. 4. Thus, we introduce the results of SMs found by the RIP in Chaps. 5, 6, and 7. In Chaps. 8 and 9, we introduce the results of SMs found by Revised LP-OLDF because RIP finds many SMs than Revised LP-OLDF.

5.2.1 Verification of the Number of Iterations of Revised LP-OLDF and RIP

Table 5.1 shows the result of verification by changing the number of iterations from 1 to 7 of the repeat option (IT) of LINGO Program3 in Chap. 1. In the Revised LP-OLDF (abbreviated as LP in the table), “IT = 5” chooses 46 SMs that contain 1,134 genes, and an average is 24.7 genes. When it is “IT = 5 or more,” it becomes a steady state including 1,134 in 46 SMs with 24.7 genes on average. Thus, we choose the 46 SMs by Revised LP-OLDF. In contrast, RIP chooses 5,990 genes in 179 SMs with IT = 3, the average is 33.46, and the noise subspace contains 1,139 genes. We evaluate the signal and noise subspaces by six MP-LDFs in Sects. 5.2.2 and 5.2.3.

Table 5.1 Result of the number of iterations by LINGO Program3

	LP				RIP			
IT	CPU	SM	Gene	Gene/SM	CPU	SM	Gene	Gene/SM
1	55	26	1142	43.9	2m46	89	6176	69.39
2	**2:05**	**42**	**1173**	27.9	7m28	166	6079	36.62
3	2:44	40	1070	26.8	**9m47**	**179**	**5990**	**33.46**
4	3:36	43	1087	25.3	11m53	179	5990	33.46
5	**4:30**	**46**	1134	24.7	**12m50**	179	5990	33.46
6	5:15	46	1134	24.7	14m38	179	5990	33.46
7	6:02	46	1134	24.7	16m33	179	5990	33.46

5.2.2 Analysis of Signal Subspace and Noise Subspace Obtained by Revised LP-OLDF

We develop LINGO Program3 of Method2 for the RIP at first. However, we confirm Revised LP-OLDF, and Revised IPLP-OLDF can decompose six microarrays into many SMs also. Because the calculation time of Revised LP-OLDF is faster than RIP, and the number of SMs obtained is smaller than RIP, we consider replacing the analysis of SM obtained by RIP to Revised LP-OLDF. However, we did not evaluate signal and noise subspaces obtained by the RIP and Revised LP-OLDF before 2017. Thus, we evaluate the microarray, signal subspace of union of all SMS, and the noise subspace as same as Alon’s microarray in Chap. 4. Because the result of the microarray is as same as signal subspace, we omit it from Table 5.2. Up to now, the Revised LP-OLDF seems to be preferable because it separates it into the small size of signal subspace and the large size of noise subspace with many genes, so the total analysis work becomes shorter than RIP.

The left six columns of Table 5.2 show the signal subspace with 1,134 genes included in 46 SMs, and the right six columns show the results of the noise subspace with 5,995 genes. The first column under the fourth row shows the sequential number (SN). SNs from 1 to 25 are 25 AML patients, and SNs from 26 to 72 are 47 ALL patients. Second row "6 LDFs" shows the six MP-based LDFs. Third row "RatioSV" indicates RatioSVs of six LDFs for the signal subspace and noise subspace. The DSs of 25 cases (class1) and 47 cases (class2) are shown in the fifth line or less. In the signal subspace, six RatioSVs are 100%, 16.06%, 11.97%, 34.71%, 34.36%, and 34.61%, respectively. Because 25 class1 patients are on "SV = −1" and 47 class2 patients are on "SV = 1" RatioSV of RIP is 100%. In this study, we define that outlier patients belonging to class1 are less than −1 and outlier patients belonging to class2 are greater than 1. That is, in the signal space, we found that all patients are on two SVs of the RIP. We summarize six RatioSVs as follows:

(1) RIP discovered a relationship where 25 and 47 patients lie on two parallel hyperplanes represented by two 1,134-dimensional SVs. The surprising result is accepted by three facts as follows: (a) Because the other five RatioSVs are 35% or less, this result shows RIP finds the real relation in the high-dimensional gene space. That is, MNM is an optimization criterion that finds hidden features of data. For this reason, the range of RipDS is [−2, 2], whereas the range of LpDS is largely [−6.75, 9.95]. Even with the same data, the difference in subjects of binding to SV causes about four times range's difference. Because the SV distance is fixed to 2, we do not think that the actual distance of six ranges is very different. Namely, data fluctuation is considered to be small. In the future, if we can visualize two groups in a high-dimensional space, it will be clear. (b) RIP decomposes 1,134 signal subspaces into 179 SMs, and the range of 179 RatioSVs in Table 5.7 is [0.52, 28.8]. This fact is the first case showing the specific examples of the SV in a high-dimensional signal subspaces such as "RatioSV=100" and the 179 SVs of 179 SMs. (c) Because the RatioSV of Alon's signal subspace is 14.1% in Table 4.4, this value indicates that Golub data structure causes the surprising figure of Golub signal subspace.

(2) Because three SVMs maximize the distance between the SVs, and two SVs fix some patients on two 1,134-dimensional hyperplanes, those do not show an extreme value like the RIP. Three RatioSVs are about 34%.

(3) RatioSV of Revised LP-OLDF seems to be smaller than SVMs because it does not fix subjects to SVs.

Because the noise subspace is LSD, H-SVM works correctly. Row "Outlier/NM" shows the number of outliers (Outlier) for a signal subspace, and six NMs for the noise subspace. These facts tell us the noise subspace is LSD and includes several SMs in it.

Table 5.2 Evaluation of the signal subspace and noise subspace (Revised LP-OLDF)

	Signal (1,134 genes included in 46 SMs)						Noise (5,995 genes)					
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
RatioSV	**100**	16.063	11.979	34.714	34.362	34.618	100	15.514	4.371	**62.926**	52.340	61.767
Outlier/NM	0/0	8/23	9/30	**10/30**	25/47	15/39	0	0	0	**0**	0	0
1	−1	−1	−1	−1.37854	−1.39664	−1.38521	−1	−1.40846	−15.234	−1	−1.17384	−1.01281
2	−1	−1	−1	−1	−1.00173	−1.00016	−1	−1	−1	−1	−1.14392	−1.01096
3	−1	−5.3022	−4.9262	−1.12396	−1.15761	−1.13279	−1	−3.72337	−3.08507	−1.11782	−1.3182	−1.13783
4	−1	−1	−1	−1.02744	−1.07101	−1.04178	−1	−1.37129	−13.8004	−1.05085	−1.27472	−1.07941
5	−1	−1	−1	−1	−1.00204	−1.00019	−1	−3.71541	−1	−1	−1.05082	−1.00135
6	−1	−2.98542	−1	−2.02789	−2.04961	−2.03417	−1	−2.91759	−1	−1	−1.01782	−1.00079
7	−1	−1	−1	−1.34486	−1.35652	−1.34743	−1	−5.16253	−11.3328	−1	−1.13007	−1.00001
8	−1	−1	−1	−1.15215	−1.16346	−1.15458	−1	−3.10441	−1	−1	−1.01997	−1.00086
9	−1	−1	−1	−1	−1.0238	−1.00311	−1	−1	−1	−1	−1.01638	−1.00058
10	−1	−3.53048	−4.07094	−1	−1.00813	−1.00115	−1	−5.86375	−1	−1	−1.02768	−1.00127
11	−1	−1	−1	−1	−1.00181	−1.00017	−1	−1	−1	−1	−1.00746	−1.00027
12	−1	−2.26011	−1	−1	−1.00384	−1.00036	−1	−1	−19.7522	−1.13936	−1.33229	−1.15643
13	−1	−1	−1	−1	−1.00294	−1.00029	−1	−4.23861	−1	−1	−1.08108	−1.00173
14	−1	−1	−5.25481	−1	−1.00298	−1.00029	−1	−4.65075	−1	−1	−1.07044	−1.00053
15	−1	−2.79503	−1	−1.20758	−1.21931	−1.21092	−1	−6.17355	−1	−1.19521	−1.30626	−1.20029
16	−1	−5.9045	−5.02382	−1.1979	−1.21654	−1.2032	−1	−1	−1	−1	−1.0129	−1.00047
17	−1	−1	−1.11629	−1	−1.01193	−1.00111	−1	−4.29977	−9.66058	−1.31823	−1.5851	−1.33866
18	−1	−1	−1	−1	−1.03761	−1.00796	−1	−1	−1	−1.31539	−1.62557	−1.34888

(continued)

Table 5.2 (continued)

	Signal (1,134 genes included in 46 SMs)						Noise (5,995 genes)					
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
19	−1	−1	−2.57776	−1.02436	−1.07632	−1.04248	−1	−1	−1	−1	−1.0277	−1.00085
20	−1	−1	−1	−1	−1.00182	−1.00016	−1	−1	−16.7241	−1	−1.02728	−1.00072
21	−1	−3.46904	−1	−1.17955	−1.21142	−1.18981	−1	−4.04779	−7.71447	−1	−1.07649	−1.00024
22	−1	−2.2106	−1.64184	−1	−1.00592	−1.00069	−1	−3.09869	−1	−1.26638	−1.5425	−1.29029
23	−1	−1	−6.74716	−1	−1.00458	−1.00046	−1	−4.18437	−1	−1	−1.12347	−1.0086
24	−1	−1	−1	−1	−1.00112	−1.00009	−1	−1	−1	−1	−1.01311	−1.0005
25	**−1**	**−1**	**−3.95499**	**−1**	**−1.00331**	**−1.00033**	**−1**	**−5.71552**	**−12.3177**	**−1.25219**	**−1.52406**	**−1.27866**
26	**1**	**2.491398**	**5.238529**	**1.292216**	**1.311987**	**1.298164**	**1**	**4.585687**	**1**	**1.00155**	**1.249569**	**1.048054**
27	1	1	9.93047	1	1.002901	1.000277	1	1	6.962039	1	1.028885	1.001269
28	1	1	1	1.000001	1.014983	1.00247	1	1	1	1	1.027717	1.000932
29	1	1.370267	3.23441	1	1.008816	1.000715	1	1.772793	1	1	1.031435	1.001114
30	1	4.912292	1.235503	2.815479	2.850709	2.82586	1	3.14804	19.2334	1.458813	1.721019	1.480145
31	1	4.453237	1	1.730579	1.760386	1.742519	1	1	1	1.000001	1.07973	1.000264
32	1	1.507499	3.327226	1.165981	1.188517	1.171163	1	1.369551	20.07756	1.410734	1.663227	1.429892
33	1	1.649354	1	1.418023	1.438416	1.423811	1	1	19.33288	1	1.019998	1.000789
34	1	4.817817	6.489941	1.544361	1.567658	1.551247	1	2.431901	16.17909	1.478039	1.748126	1.499954
35	1	2.979798	7.433853	1.481756	1.493085	1.484593	1	2.092595	13.13297	1.312959	1.514176	1.327646
36	1	3.3027	5.682091	1.437031	1.456469	1.442624	1	2.654957	13.27088	1.269748	1.472541	1.285915
37	1	1	1	1	1.007927	1.001438	1	1.642259	9.116508	1	1.080763	1.00018
38	1	1.434975	1	1	1.010861	1.001887	1	2.627415	4.595891	1.108843	1.30267	1.131823

(continued)

Table 5.2 (continued)

	Signal (1,134 genes included in 46 SMs)						Noise (5,995 genes)					
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
39	1	3.056575	3.148292	1.352878	1.378683	1.360609	1	1.100464	7.493038	1.04172	1.279791	1.079453
40	1	2.772783	3.233322	2.486947	2.51386	2.493969	1	3.452423	1	1.384866	1.605843	1.402606
41	1	4.957842	7.000641	2.783151	2.820587	2.793041	1	1	8.897861	1.160199	1.392542	1.186289
42	1	1	1	1	1.001315	1.000113	1	1	5.821985	1	1.013194	1.000435
43	1	1	1	1	1.001947	1.000185	1	1	1	1	1.012251	1.000553
44	1	1	3.085397	1.250948	1.264342	1.25492	1	2.181227	10.395	1.000001	1.11961	1.001088
45	1	1.202215	5.532655	1.325616	1.342854	1.331123	1	2.05818	1	1.000001	1.073887	1.000813
46	1	1.649484	6.417837	2.196442	2.228017	2.206011	1	4.701262	1	1.13147	1.36415	1.156271
47	1	1	1	1.038899	1.082402	1.051974	1	1	1	1	1.029376	1.00111
48	1	1	3.645618	1.867523	1.889888	1.872951	1	4.385619	3.795768	1.029316	1.266243	1.062548
49	1	1	2.131231	1.854938	1.871984	1.860044	1	2.536775	2.822555	1.210832	1.40411	1.224736
50	1	2.45513	1	1.959737	1.979467	1.965404	1	1.644538	1	1.021654	1.246336	1.056416
51	1	1	5.043033	1.124682	1.140854	1.1276	1	2.213915	1	1.08178	1.259449	1.09867
52	1	1.789315	1	1	1.004871	1.000457	1	5.11758	1	1	1.077528	1.000226
53	1	1	7.407957	1.164375	1.192524	1.172051	1	3.787116	1	1.000001	1.084819	1.000277
54	1	1.638509	5.284846	2.281986	2.318899	2.290664	1	3.044317	7.984743	1.000001	1.143644	1.003435
55	1	1	3.248883	1	1.003981	1.000364	1	1	4.480029	1	1.060112	1.000613
56	1	1	1	1.000004	1.025714	1.003915	1	6.717484	1	1.223062	1.47356	1.246992
57	1	1	1	1	1.00077	1.000065	1	1	1	1	1.014257	1.000611
58	1	4.296339	4.938716	2.619918	2.654301	2.630033	1	2.20841	1	1.860068	2.195568	1.889068

(continued)

Table 5.2 (continued)

	Signal (1,134 genes included in 46 SMs)						Noise (5,995 genes)					
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
59	1	1	6.060859	1	1.009395	1.001209	1	1	8.030451	1.000001	1.127303	1.00005
60	1	5.071377	3.843741	2.948089	2.982729	2.958292	1	3.6437	20.60081	1.660826	1.975526	1.68926
61	1	6.545783	3.3396	2.265499	2.279105	2.268956	1	1	7.248019	1.323783	1.565281	1.343793
62	1	1	2.265593	1	1.008574	1.000841	1	1	1	1	1.027206	1.001208
63	1	1	1	1	1.002231	1.000207	1	1	1	1	1.05908	1.000907
64	1	1	4.581196	1.78592	1.797574	1.789604	1	2.41646	1	1.231606	1.402576	1.242951
65	1	1	9.947427	3.733327	3.770672	3.743099	1	1	26.00105	1.586423	1.874218	1.609178
66	1	1	1	1	1.030722	1.002773	1	2.712617	12.28112	1.443421	1.728761	1.468991
67	1	1	1	1	1.004351	1.0004	1	3.596926	1	1.078373	1.285704	1.104776
68	1	1	1.264801	1	1.007262	1.00064	1	3.063185	7.818267	1.138851	1.361948	1.159524
69	1	1.534902	1	2.023226	2.04277	2.027213	1	4.322948	1	1.289801	1.576181	1.32022
70	1	1	5.838869	1.215004	1.232713	1.220045	1	3.683702	1	1	1.036475	1.001191
71	1	3.954635	7.336682	2.152787	2.173759	2.159257	1	1	1	1.160783	1.371441	1.182424
72	1	1	1	1.023456	1.070588	1.040903	1	1	1	1.085104	1.247162	1.097498

Table 5.3 shows the discriminant results by six MP-LDFs. If each LDF correctly discriminates patients, the " >0" column shows the classified patient number. The "0" column means the number of patients on the discriminant hyperplane, and " <0" column means the number of misclassified patients. As a result, six MP-LDFs classified all patients correctly for the microarray, the signal, and noise spaces. This fact indicates the noise subspace is LSD. Thus, we conclude we do not analyze 46 SMs obtained by Revised LP-OLDF in this chapter.

Table 5.3 Discriminant results by six MP-LDFs of 46 SMs obtained by Revised LP-OLDF

	Golub (7,129 genes)			Signal (1,134 genes)			Noise (5,995 genes)		
	<0	0	>0	<0	0	>0	<0	0	>0
RIP	0	0	77	0	0	77	0	0	77
LP	0	0	77	0	0	77	0	0	77
IPLP	0	0	77	0	0	77	0	0	77
HSVM	0	0	77	0	0	77	0	0	77
SVM4	0	0	77	0	0	77	0	0	77
SVM1	0	0	77	0	0	77	0	0	77

5.2.3 *Analysis of Signal Subspace and Noise Subspace Obtained by RIP*

Table 5.4 shows the discriminant results of 179 SMs by six MP-LDFs. Six MP-LDFs can discriminate signal subspace correctly. H-SVM cannot discriminate the noise subspace. Other five NMs are 1, 1, 1, 1, and 15, respectively. Moreover, SVM1 is worse than the other four LDFs. We recommended the penalty c = 1000 or 10000 instead of c = 1. Thus, we can confirm RIP can separate the microarray into signal and noise subspaces correctly.

Table 5.4 Discriminant results by six MP-LDFs of 179 SMs obtained by RIP

	Signal (5,999 genes, 179 SMs)			Noise (1,139 genes)		
	<0	0	>0	<0	0	>0
RIP	0	0	72	1	0	71
IPLP	0	0	72	1	0	71
LP	0	0	72	1	0	71
HSVM	0	0	72	–	–	–
SVM4	0	0	72	1	0	71
SVM1	0	0	72	15	0	57

Table 5.5 is the analysis of signal subspace and noise subspaces obtained by RIP. The left six columns show the signal subspace with 5,990 genes included in 179 SMs, and the right six columns show the results of the noise subspace. Because the microarray is as same as the signal subspace, we omit it from the table. In the signal subspace, six RatioSVs are 100%, 18.59%, 0.362%, 61.88%, 55.71%, and 59.67%, respectively. Because 25 class1 patients are on "SV = −1" and 47 class2 patients are on "SV = 1," RatioSV of RIP is 100%. However, the RatioSV of IPLP is 0.362%. Revised IPLP-OLDF is a mixed model of Revised LP-OLDF and RIP. One subject of AML class is outlier −161.5. Of the 47 subjects of ALL, only 8 are outliers, and the maximum value of DS is 391.1. Following the RIP, there are as many as 63 subjects on the SV. However, nine subjects have large outliers, and RatioSV is the smallest. This point is a disadvantage of RatioSV, which indicates that we must interpret carefully. Although the magnitude of the actual fluctuation is small, the DSs may become extremely different due to the influence of outliers. Because the noise subspace is not LSD, H-SVM outputs the error. We consider that the RatioSVs are useless for the overlapping data. Row "Outlier/NM" shows the numbers of outliers are 0/0, 15/28, 1/8, 8/26, 25/47, and 12/47, respectively. Five NMs for the noise subspace are 1, 1, 1, 1, and 15, respectively. The 15 positive bold figures of 25 AML patients show the misclassified patients.

Table 5.5 Evaluation of the signal subspace and noise subspace (RIP)

	Signal (5,990 genes, 179 SMs)						Noise (1,139 genes)				
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
Ratio	100	18.597	0.3619	61.875	55.706	59.674	0.0051	2.3346	2.185	16.674	81.577
Outlier/NM	0/0	15/28	1/8	8/26	25/47	12/47	1	1	1	1	15
1	−1	−3.857	−1	−1	−1.115	−1.035	−10324	−1	−1	−1.03	**0.071**
2	−1	−1	−1	−1	−1.06	−1.008	−13600	−1	−1	−1.007	**0.877**
3	−1	−1.643	−1	−1.085	−1.208	−1.124	−1	−1	−6.135	−1.022	−1
4	−1	−4.173	−1	−1.035	−1.177	−1.086	−1	−1	−1	−1.036	−1
5	−1	−1.992	−1	−1	−1.021	−1.003	−1	−1	−1	−1.3	−1
6	−1	−1.77	−1	−1	−1.016	−1.005	−1	−1	−1	−1.004	**0.745**
7	−1	−4.649	−1	−1	−1.098	−1.028	−1	−1	−1	−1.205	−1
8	−1	−1	−1	−1	−1.009	−1.002	−4377	−1	−5.867	−1.031	**0.292**
9	−1	−1	−1	−1	−1.008	−1.002	−1	−1	−1	−1.004	**0.74**
10	−1	−3.324	−1	−1	−1.013	−1.003	−1	−1.695	−12.94	−1.416	**0.404**
11	−1	−1	−1	−1	−1.004	−1.001	−1	−1	−1	−1.002	**0.963**
12	−1	−1.338	−1	−1.104	−1.227	−1.15	−1	−1	−1	−1.01	**0.331**
13	−1	−2.252	**−161.5**	−1	−1.031	−1.005	−1	−1	−1	−1.12	−1
14	−1	−1	−1	−1	−1.012	−1.005	−1	−1	−1	−1.428	−1
15	−1	−1	−1	−1.138	−1.205	−1.156	−8500	−1	−19.73	−1.004	**0.817**
16	−1	−1	−1	−1	−1.007	−1.002	−7087	−1	−11.96	−1.002	**0.288**
17	−1	−2.409	−1	−1.281	−1.437	−1.33	−1	−1	−1	−4.274	−1
18	−1	−2.522	−1	−1.274	−1.461	−1.339	−1	−1	−1	−1.014	−0.271

(continued)

Table 5.5 (continued)

	Signal (5,990 genes, 179 SMs)						Noise (1,139 genes)				
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
19	−1	−1	−1	−1	−1.023	−1.004	−1	−1	−1	−1.062	−1
20	−1	−1	−1	−1	−1.01	−1.002	−2793	−1	−1	−1.004	**0.64**
21	−1	−2.261	−1	−1	−1.042	−1.003	−1	−1	−11.27	−1.001	**0.687**
22	−1	−1.769	−1	−1.223	−1.398	−1.281	−1	−1	−1	−1.033	**0.252**
23	−1	−2.189	−1	−1	−1.065	−1.013	−1	−1	−1	−1	**0.866**
24	**−1**	**−1**	**−1**	**−1**	**−1.006**	**−1.001**	**−1**	**−1**	**−1**	**−1.003**	**0.626**
25	**−1**	**−1.537**	**−1**	**−1.111**	**−1.284**	**−1.173**	**−1**	**−1**	**−1**	**−1.074**	**−1**
26	1	1.2886	155.59	1.0422	1.228	1.1139	1	1	1	1.2922	1
27	1	3.8333	1	1	1.0093	1.0025	1	1	1	1.0206	1
28	1	1	1	1	1.0121	1.0031	1	1	1	1.0288	1
29	1	1	1	1	1.0099	1.0029	1	5.744	1	2.6811	1
30	1	3.8152	21.149	1.6048	1.7522	1.6539	1	14.357	1	1.175	1
31	1	1	1	1	1.1045	1.0245	1	1	1	1.0341	1
32	1	4.5978	1	1.346	1.4803	1.3907	104.43	1	1	1.596	1
33	1	1	1	1	1.012	1.0036	1	1	1	1.0289	1
34	1	2.6291	117.48	1.4582	1.6398	1.5199	10417	1	1	1.0116	0.5963
35	1	5.2432	1	1.3294	1.449	1.3643	1	1	71.801	1.1218	1
36	1	6.1055	1	1.2857	1.4186	1.327	1	18.82	1	1.7282	1
37	1	1	1	1	1.0719	1.0065	1	4.7375	1	1.0491	1
38	1	3.3499	1	1.1104	1.2306	1.1504	2964.3	1	1.6598	1.0119	0.849

(continued)

Table 5.5 (continued)

	Signal (5,990 genes, 179 SMs)						Noise (1,139 genes)				
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
39	1	1	1	1.0985	1.2611	1.1583	1	1	1	1.0388	1
40	1	3.8941	1	1.5194	1.6232	1.552	1	1	1	1.2592	1
41	1	3.6921	1	1.4958	1.6233	1.5333	25289	1	11.617	1.1686	1
42	1	1	1	1	1.0058	1.001	1	1	1	1.5676	1
43	1	1	1	1	1.004	1.0011	1	1	1	1.0035	0.7522
44	1	1	1	1	1.1263	1.0429	1	1	1	1.3553	1
45	1	1	103.79	1	1.117	1.0448	1	1	1	2.5309	1
46	1	2.7304	1	1.3252	1.4874	1.3785	5095.2	1	1	1.151	1
47	1	1	1	1	1.0148	1.0031	5525.3	1	1	1.0102	1
48	1	4.4722	1	1.0641	1.2289	1.1296	1	1	1	1.0438	1
49	1	3.2	140	1.2813	1.3893	1.3162	1	1	3.7053	1.0611	1
50	1	1	1	1.1362	1.2566	1.1768	1	1	1	1.5223	1
51	1	1.7989	1	1.0044	1.1447	1.0644	6008.7	1.0546	1	1.1596	1
52	1	4.0736	1	1	1.0111	1.0069	1	1	23.751	1.004	0.2205
53	1	4.7732	1	1	1.0569	1.0033	1	1	1	1.6654	1
54	1	1	1	1.1131	1.241	1.1592	1	1	1	1.031	1
55	1	1	1	1	1.012	1.0054	1	1	1	1	0.8589
56	1	5.8773	1	1.0882	1.2494	1.1515	4022	1	3.8791	1.0086	0.91
57	1	2.3041	1	1	1.0067	1.0014	1	1	1	1	0.8961
58	1	3.6704	**391.1**	1.9512	2.1293	2.013	1	1	1	1.0192	1.0922

(continued)

Table 5.5 (continued)

	Signal (5,990 genes, 179 SMs)						Noise (1,139 genes)				
Six LDFs	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
59	1	4.3343	1	1	1.0372	1.0034	1	1	1	1.006	1
60	1	2.4667	1	1.8338	2.0098	1.8922	1	2.1506	4.6145	1.0321	1
61	1	1	1	1.4983	1.6397	1.5413	1	42.9	1	1.0093	1.1162
62	1	1	1	1	1.0088	1.0026	1	1	1	1.0469	1
63	1	1	1	1	1.0108	1.0036	1	1	1	1.0994	0.989
64	1	3.4369	1	1.3592	1.4459	1.3836	1	1	1	1.0022	0.268
65	1	1	1	1.8858	2.0519	1.9354	13634	20.192	1	7.7209	1.4517
66	1	2.3893	1	1.3155	1.4754	1.37	1	83.974	1	1.8167	1
67	1	4.3987	186.09	1.0046	1.1738	1.0746	1	1	1	1.008	0.3595
68	1	3.9818	1	1	1.1411	1.0496	1	1.3272	1	1.0151	1
69	1	2.5488	1	1.402	1.5513	1.4538	18668	1	1.9032	1.0479	1
70	1	1.7608	1	1	1.0311	1.0041	2982.2	1.0714	1	1.0067	0.8972
71	1	3.3115	339.2	1.355	1.4944	1.398	1046	1	1	1.0017	1
72	1	1	1	1	1.131	1.051	1	1	1	1.0603	1

Table 5.6 shows the coefficients of six MP-based LDFs for the signal subspace in Table 5.5. Column “SN1” shows the sequential number of 7,129 genes corresponding to the last column “Gene” that indicates gene name downloaded from Higgins HP. Six figures of the second row show the nonzero number of six MP-based LDFs. The 71 RIP coefficients are nonzero coefficients and displayed in the table. The 1,134 coefficients of three SVMs are not zero, and other 5,995 coefficients are zero naturally. Column “SN2” shows only 71 nonzero coefficients of the RIP. We hide 7,058 rows with zero coefficients of the RIP. This column indicates the exciting information as follows:

(1) Although nonzero coefficients of Revised IPLP-OLDF and Revised LP-OLDF are 41 and 33, most nonzero coefficients are different from those of RIP. Because LP defines Revised LP-OLDF and Revised IPLP-OLDF, both OLDFs can find another vertex of the feasible region, the dimension of which is less than or equal to the patient number 72.

(2) The 5,995 coefficients of three SVMs become zero naturally. This fact suggests us we need not construct the complex theory such as LASSO. Even though SVMs make many coefficients to zero, these SVMs are useless for the cancer gene diagnosis because those cannot decompose Golub microarray into many SMs.

Table 5.6 Coefficients of six MP-based LDFs for the signal subspace

SN1	SN2	RIP	IPLP	LP	HSVM	SVM4	SVM1	Gene
		71	41	33	1134	1134	1134	
127	1	– 0.188	– 0.279	0	0.002	0.003	0.002	X1868
134	2	– 0.116	0	0	0.002	0.002	0.002	X4505
149	3	0.183	0	0	0.002	0.002	0.002	X3214
153	4	0.865	0	0	0.004	0.005	0.004	X3617
199	5	0.008	0	0	0.000	0.000	0.000	X3868
204	6	0.911	0	0	– 0.001	– 0.001	– 0.001	X5259
214	7	– 0.475	0	0	0.008	0.008	0.008	X490
228	8	0.124	0	0	– 0.001	– 0.001	– 0.001	X5711
235	9	0.359	0	0	0.010	0.010	0.010	X668
246	10	1.227	0	0	0.012	0.012	0.012	X4399
252	11	0.268	0	0	0.007	0.007	0.007	X6340
254	12	0.630	0	0	0.010	0.010	0.010	X7073
258	13	– 0.348	0	0	0.004	0.004	0.004	X1350
285	14	0.472	0	0	0.005	0.005	0.005	X2942
286	15	– 0.324	0	0	0.004	0.004	0.004	X3070

(continued)

Table 5.6 (continued)

SN1	SN2	RIP	IPLP	LP	HSVM	SVM4	SVM1	Gene
289	16	0.000	0	0	– 0.002	– 0.002	– 0.002	X3661
290	17	– 0.144	0	0	0.000	0.001	0.000	X4463
296	18	0.320	0	0	0.004	0.004	0.004	X5602
302	19	– 0.035	0	0	0.001	0.002	0.001	X469
310	20	– 0.388	0	0	0.001	0.002	0.001	X1694
313	21	– 0.586	0	0	0.005	0.005	0.005	X3824
329	22	0.392	0	0	0.004	0.004	0.004	X2045
332	23	0.505	0	0	0.006	0.006	0.006	X2241
341	24	0.892	0	0	0.002	0.002	0.002	X4621
343	25	0.069	0	0	0.002	0.002	0.002	X5006
347	26	0.010	0.2012	0	– 0.003	– 0.004	– 0.004	X6734
357	27	0.690	0	0	0.005	0.005	0.005	X1520
368	28	0.156	0	0	0.007	0.007	0.007	X5300
369	29	– 0.525	0	0	0.004	0.004	0.004	X5460
375	30	0.279	0	0	0.000	0.000	0.000	X695
380	31	0.139	0	0	0.003	0.003	0.003	X2685
382	32	– 0.643	0	0	0.005	0.005	0.005	X2995
390	33	0.080	0	0	– 0.004	– 0.003	– 0.004	X5088
391	34	– 0.132	0	0	0.000	0.000	0.000	X5356
401	35	0.106	0	0	0.008	0.008	0.008	X3494
412	36	0.174	0	0	0.005	0.005	0.005	X5335
413	37	– 0.466	0	0	0.003	0.002	0.002	X5739
424	38	– 0.046	0	0	0.004	0.003	0.004	X1354
425	39	– 0.149	0	0	0.006	0.005	0.006	X1365
427	40	– 0.356	0	0	0.004	0.004	0.004	X2855
432	41	0.192	0	0	0.006	0.006	0.006	X4307
440	42	0.125	0	0	0.000	0.001	0.001	X54
447	43	– 0.028	0	0	– 0.002	– 0.001	– 0.002	X1120
476	44	– 0.440	0	0	0.003	0.003	0.003	X3344
482	45	2.046	0	0	0.003	0.003	0.003	X5390
485	46	– 0.849	0	0	0.005	0.005	0.005	X6725
507	47	0.186	0	0	0.005	0.005	0.005	X6744
509	48	– 0.877	0	0	– 0.001	– 0.001	– 0.001	X815
521	49	– 0.084	0	0	– 0.004	– 0.004	– 0.004	X4104
568	50	0.137	0	0	0.001	0.002	0.001	X3688
642	51	– 0.099	0	0	– 0.004	– 0.004	– 0.004	X2463
670	52	1.434	0	0	0.002	0.002	0.002	X2278

(continued)

Table 5.6 (continued)

SN1	SN2	RIP	IPLP	LP	HSVM	SVM4	SVM1	Gene
672	53	– 0.918	0	0	– 0.001	0.000	– 0.001	X2372
683	54	0.821	0	0	0.002	0.002	0.002	X6935
694	55	0.055	0	0	– 0.008	– 0.008	– 0.008	X2379
707	56	– 0.122	0	0	0.002	0.003	0.002	X6801
713	57	– 0.899	0	0	– 0.003	– 0.003	– 0.003	X1277
720	58	– 0.131	0	0	– 0.003	– 0.003	– 0.003	X2761
726	59	– 0.580	0	0	– 0.005	– 0.006	– 0.006	X5465
729	60	– 0.416	0	0	0.000	0.000	0.000	X5639
802	61	0.377	0	0	– 0.009	– 0.008	– 0.009	X5715
844	62	– 0.134	0	0	0.002	0.002	0.002	X2715
879	63	0.335	0	0	– 0.007	– 0.008	– 0.007	X4812
944	64	– 0.377	0	0	– 0.003	– 0.002	– 0.002	X6629
962	65	– 0.434	0	0	– 0.005	– 0.005	– 0.005	X3969
974	66	– 0.156	0	0	– 0.010	– 0.009	– 0.010	X6839
976	67	0.029	0.0208	0	– 0.015	– 0.014	– 0.014	X618
981	68	0.102	0	0	– 0.003	– 0.003	– 0.003	X1828
983	69	– 0.257	0	0	– 0.004	– 0.004	– 0.004	X2164
1036	70	– 0.239	– 0.111	0	– 0.002	– 0.002	– 0.002	X5506
1082	71	0.477	0	0	– 0.001	– 0.001	– 0.001	X990
C	6	– 19.869	0	0	– 23.693	– 23.900	– 23.751	X7130

5.3 Analysis of 179 SMs of Golub et al. Microarray (2018)

In 2015, LINGO Program3 decomposed the microarray into 67 SMs with 1,203 genes. However, when RIP of LINGO Program3 decomposes the microarray again in 2018, it finds 179 SMs with 5,990 genes. We obtain more SMs and genes in 2018. A yearly update of LINGO or the different iteration numbers cause these differences. We consider 179 SMs are signals, and 1,134 gene subspaces are noise. Program3 can separate signals and noise very quickly. We need not develop a filtering method. Although we can analyze 179 SMs by the standard statistical methods, we cannot find linear separable facts that two classes are entirely separable in each SM. Only logistic regression can discriminate all SMs correctly (Cox 1958; Firth 1993). Because the 179 MNMs of SMs are zero, we can specify 179 pairs of genes included in 179 SMs as the cancer genes. We hope the medical specialists examine these SMs as the cancer genes. However, there is also the idea that SM is not a signal representing cancer genes. In other words, we recognize the signal data created by DS as the true signal instead of genes included in SM.

5.3.1 *Validation of 179 SMs by Six MP-Based LDFs and Discriminant Functions*

Table 5.7 shows the 179 SMs from SM = 1 to SM = 179 found by RIP. Although Revised LP-OLDF can decompose the microarray into 46 SMs, we explain those results in Sect. 5.2. On the other hand, H-SVM can discriminate microarrays correctly. However, H-SVM could not decompose microarrays explained in Chap. 1. The "Gene" column is the number of genes of each SM. The range of genes included in the 179 SMs is [11, 54]. The average is 33.5. Row "Total" indicates 179 SMs contain 5,990 genes.

Three RatioSVs of 179 SMs are shown from RIP column to H-SVM column. Three ranges of RatioSV are [0.52, 28.80], [0.72, 25.1], and [1, 33.34], respectively. Three averages of RatioSVs are 13.66%, 13.05%, and 16.68%, respectively. Row "Max" indicates the number of the maximum RatioSVs among 179 SMs those are 28, 18, and 133, respectively. To summarize these results, the range, average, and maximum number of H-SVM are better than RIP because the maximization SV of H-SVM works well for LSD. Moreover, two RatioSVs of LP and IPLP are bigger than those of the signal in Table 5.5. Two columns "Max and Min" are the maximum and minimum values of three LDFs except for IPLP, SVM4 and SVM1.

Because all NMs of logistic regression and SVM4 are zero and 179 SMs are linearly separable, we omit these columns from the table. Three columns "SVM1, LDF2, and QDF" show the NM (Sall et al. 2004). Three ranges are [0, 23], [0, 9], and [0, 30], respectively. The averages are 2.12, 1.53, and 2.54, respectively. Three numbers of misclassified SMs are 68, 112, and 59, respectively. SVM1 cannot discriminate 68 SMs correctly, and NMs may increase according to decreased RatioSV. SVM1, LDF2, and QDF cannot discriminate 68 SMs, 112 SMs, and 59 SMs correctly.

Table 5.7 Evaluation of 179 SMs by RatioSVs and NMs

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
1	18	15.91	14.58	**18.03**	18.03	14.58	0	2	0
2	30	17.38	**17.58**	17.33	17.58	17.33	0	0	0
3	17	4.74	**5.19**	5.18	5.19	4.74	2	5	21
4	26	15.32	**15.65**	14.95	15.65	14.95	0	2	0
5	19	19.38	19.86	**20.36**	20.36	19.38	0	1	1
6	29	18.53	13.25	**25.37**	25.37	13.25	0	0	0
7	22	21.19	21.94	**22.37**	22.37	21.19	0	0	0
8	20	8.73	21.27	**22.22**	22.22	8.73	0	0	0
9	26	17.17	19.24	**19.99**	19.99	17.17	0	3	11
10	25	23.09	25.10	**26.38**	26.38	23.09	0	0	0
11	29	17.78	20.20	**21.83**	21.83	17.78	0	0	0

(continued)

Table 5.7 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
12	28	14.72	17.69	**18.94**	18.94	14.72	0	1	0
13	16	**19.01**	15.94	12.69	19.01	12.69	0	1	0
14	32	16.63	17.92	**20.68**	20.68	16.63	0	1	0
15	21	15.84	17.28	**21.94**	21.94	15.84	0	2	0
16	12	9.82	9.99	**10.09**	10.09	9.82	0	3	0
17	28	18.14	20.21	**27.61**	27.61	18.14	0	0	0
18	27	**25.28**	22.16	23.12	25.28	22.16	0	1	0
19	16	13.10	14.02	**17.44**	17.44	13.10	0	2	0
20	37	14.46	12.93	**21.22**	21.22	12.93	0	1	0
21	20	**19.65**	18.89	19.14	19.65	18.89	0	0	0
22	28	**19.62**	16.21	18.48	19.62	16.21	0	0	5
23	48	**27.21**	13.54	23.80	27.21	13.54	0	0	0
24	37	23.46	22.57	**30.64**	30.64	22.57	0	0	0
25	45	21.71	17.11	**23.11**	23.11	17.11	0	0	0
26	39	**23.23**	16.02	22.37	23.23	16.02	0	0	0
27	38	22.46	19.09	**25.04**	25.04	19.09	0	0	0
28	45	9.15	12.48	**24.68**	24.68	9.15	0	0	0
29	36	**27.18**	24.47	26.48	27.18	24.47	0	0	0
30	26	20.61	**22.69**	21.56	22.69	20.61	0	0	0
31	26	18.76	**18.91**	16.94	18.91	16.94	0	0	0
32	29	**28.80**	16.86	19.82	28.80	16.86	0	0	0
33	27	20.79	17.83	**24.32**	24.32	17.83	0	0	0
34	20	15.02	14.52	**15.49**	15.49	14.52	0	2	0
35	24	8.27	12.46	**13.01**	13.01	8.27	0	1	0
36	22	23.46	22.72	**24.64**	24.64	22.72	0	0	0
37	16	9.61	8.00	**10.34**	10.34	8.00	0	4	3
38	18	8.55	**10.45**	9.61	10.45	8.55	0	4	19
39	28	10.76	**20.81**	17.82	20.81	10.76	0	2	0
40	31	10.13	17.46	**20.90**	20.90	10.13	0	0	3
41	26	**19.77**	13.18	15.21	19.77	13.18	0	0	0
42	29	18.83	24.32	**29.34**	29.34	18.83	0	1	0
43	17	9.76	**11.97**	11.48	11.97	9.76	0	1	0
44	54	13.91	18.17	**30.04**	30.04	13.91	0	1	0
45	28	20.46	19.23	**20.66**	20.66	19.23	0	0	0
46	39	13.25	9.69	**14.77**	14.77	9.69	0	1	0
47	28	22.99	21.28	**23.38**	23.38	21.28	0	0	0
48	34	13.07	11.16	**14.40**	14.40	11.16	0	2	0
49	20	10.72	16.58	**16.59**	16.59	10.72	0	0	0
50	34	12.74	13.98	**16.25**	16.25	12.74	0	2	0

(continued)

Table 5.7 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
51	26	11.71	**14.14**	13.20	14.14	11.71	0	2	0
52	13	10.79	11.02	**11.11**	11.11	10.79	0	1	11
53	46	17.86	16.48	**20.23**	20.23	16.48	0	1	0
54	22	14.80	**18.74**	17.20	18.74	14.80	0	1	0
55	34	11.31	19.32	**23.16**	23.16	11.31	0	1	1
56	43	15.71	11.81	**21.76**	21.76	11.81	0	0	0
57	26	15.70	**16.83**	15.50	16.83	15.50	0	1	0
58	33	20.78	24.13	**25.36**	25.36	20.78	0	0	0
59	31	20.59	13.03	**27.08**	27.08	13.03	0	1	0
60	32	12.45	14.30	**19.03**	19.03	12.45	0	0	0
61	21	**13.30**	12.32	12.89	13.30	12.32	0	4	4
62	37	**22.05**	18.98	18.61	22.05	18.61	0	1	0
63	37	22.96	17.02	**27.83**	27.83	17.02	0	0	0
64	24	14.27	15.72	**17.87**	17.87	14.27	0	0	0
65	29	15.12	10.87	**20.11**	20.11	10.87	0	2	0
66	23	18.65	**22.06**	20.70	22.06	18.65	0	1	0
67	23	7.15	**13.88**	12.37	13.88	7.15	0	2	0
68	34	11.95	13.82	**16.50**	16.50	11.95	0	2	0
69	28	16.55	18.64	**21.50**	21.50	16.55	0	0	0
70	30	14.55	16.74	**23.61**	23.61	14.55	0	0	0
71	44	23.47	20.37	**29.53**	29.53	20.37	0	0	0
72	25	20.89	19.75	**21.63**	21.63	19.75	0	1	0
73	28	19.44	13.61	**19.86**	19.86	13.61	1	1	9
74	44	19.45	18.30	**20.55**	20.55	18.30	0	0	1
75	11	4.27	4.23	**4.83**	4.83	4.23	1	1	0
76	42	13.54	15.29	**20.64**	20.64	13.54	0	1	0
77	43	15.10	11.68	**22.68**	22.68	11.68	0	0	0
78	23	4.17	14.44	**16.54**	16.54	4.17	1	1	0
79	35	11.86	13.77	**18.55**	18.55	11.86	0	0	0
80	21	**17.78**	17.23	17.58	17.78	17.23	0	1	0
81	34	14.84	13.12	**16.29**	16.29	13.12	0	0	3
82	39	16.03	10.00	**17.75**	17.75	10.00	0	2	0
83	40	**27.29**	20.35	23.38	27.29	20.35	0	0	0
84	41	15.05	14.22	**25.70**	25.70	14.22	0	0	0
85	30	7.03	11.73	**13.91**	13.91	7.03	0	3	0
86	44	12.66	11.17	**18.40**	18.40	11.17	0	1	0
87	39	12.73	13.49	**17.95**	17.95	12.73	0	1	2
88	32	21.98	**22.81**	20.25	22.81	20.25	0	0	0

(continued)

Table 5.7 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
89	34	16.30	16.18	**24.38**	24.38	16.18	0	0	1
90	23	13.81	14.49	**17.90**	17.90	13.81	0	1	0
91	**47**	20.65	14.02	**33.34**	33.34	14.02	0	0	0
92	30	20.28	18.79	**21.45**	21.45	18.79	0	0	0
93	40	17.80	18.55	**23.04**	23.04	17.80	0	0	0
94	32	16.59	13.42	**20.66**	20.66	13.42	0	0	0
95	17	15.73	16.71	**19.43**	19.43	15.73	0	0	0
96	38	11.86	9.94	**15.43**	15.43	9.94	0	1	0
97	30	15.04	13.43	**15.41**	15.41	13.43	0	0	0
98	29	14.27	13.15	**14.82**	14.82	13.15	0	2	0
99	43	**11.97**	10.48	11.95	11.97	10.48	0	1	0
100	31	15.21	11.63	**16.66**	16.66	11.63	0	0	0
101	38	11.93	11.28	**15.59**	15.59	11.28	2	1	1
102	36	8.40	9.44	**11.33**	11.33	8.40	5	5	3
103	32	10.18	9.75	**13.79**	13.79	9.75	1	2	0
104	44	8.43	12.56	**19.91**	19.91	8.43	0	0	0
105	25	**15.39**	12.76	13.71	15.39	12.76	0	2	0
106	35	17.41	17.05	**17.53**	17.53	17.05	2	1	4
107	39	12.39	11.36	**16.71**	16.71	11.36	0	1	3
108	35	5.12	7.65	**13.92**	13.92	5.12	0	2	1
109	34	**26.38**	23.82	25.91	26.38	23.82	0	0	0
110	33	10.05	12.82	**14.09**	14.09	10.05	0	1	1
111	27	9.34	11.48	**14.27**	14.27	9.34	1	3	0
112	36	20.27	11.46	**22.38**	22.38	11.46	0	0	0
113	34	**15.07**	10.56	13.83	15.07	10.56	1	4	0
114	38	15.62	18.36	**19.85**	19.85	15.62	0	1	3
115	39	14.58	10.53	**17.25**	17.25	10.53	2	1	0
116	37	12.03	7.21	**19.45**	19.45	7.21	1	0	0
117	42	15.02	16.49	**16.78**	16.78	15.02	0	1	0
118	37	**15.49**	11.37	13.89	15.49	11.37	3	3	0
119	41	13.61	15.25	**21.32**	21.32	13.61	0	0	0
120	34	14.39	9.24	**15.25**	15.25	9.24	0	3	3
121	32	13.06	11.51	**15.86**	15.86	11.51	0	2	0
122	31	8.95	**20.25**	18.87	20.25	8.95	1	1	0
123	43	14.31	15.34	**27.17**	27.17	14.31	0	0	0
124	28	9.66	8.63	**9.78**	9.78	8.63	2	3	0
125	35	**12.41**	9.49	11.44	12.41	9.49	2	1	2
126	45	10.55	**15.99**	14.61	15.99	10.55	3	0	0
127	54	**24.40**	14.33	23.86	24.40	14.33	0	1	0
128	31	7.82	**11.15**	10.44	11.15	7.82	2	2	0

(continued)

Table 5.7 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
129	30	**6.31**	5.65	5.91	6.31	5.65	2	8	2
130	44	11.67	10.51	**13.64**	13.64	10.51	2	3	1
131	41	14.64	10.65	**17.32**	17.32	10.65	1	0	0
132	46	**25.52**	14.96	25.44	25.52	14.96	0	0	0
133	28	9.64	8.90	**10.42**	10.42	8.90	2	5	14
134	34	11.63	11.17	**13.85**	13.85	11.17	1	2	0
135	36	10.60	9.60	**11.53**	11.53	9.60	2	4	0
136	33	6.97	8.82	**8.90**	8.90	6.97	1	3	4
137	34	7.54	**11.92**	11.31	11.92	7.54	5	2	0
138	38	8.68	11.18	**16.75**	16.75	8.68	2	1	8
139	33	13.42	11.23	**16.90**	16.90	11.23	1	0	0
140	25	10.35	10.21	**11.42**	11.42	10.21	1	0	0
141	29	10.60	13.24	**15.12**	15.12	10.60	2	0	13
142	34	13.36	14.49	**18.19**	18.19	13.36	2	0	13
143	32	14.51	11.05	**17.75**	17.75	11.05	2	0	13
144	40	6.13	7.62	**13.63**	13.63	6.13	7	0	2
145	39	11.81	7.39	**12.83**	12.83	7.39	3	0	13
146	36	17.81	13.31	**22.72**	22.72	13.31	1	0	10
147	37	4.03	6.34	**9.71**	9.71	4.03	5	2	16
148	31	13.57	13.64	**13.91**	13.91	13.57	4	1	2
149	35	**16.91**	11.52	14.79	16.91	11.52	1	1	2
150	32	6.85	6.33	**7.99**	7.99	6.33	5	3	1
151	35	4.60	4.19	**8.79**	8.79	4.19	7	3	22
152	40	5.26	3.90	**7.84**	7.84	3.90	5	4	0
153	48	15.12	14.24	**26.13**	26.13	14.24	1	1	6
154	35	12.28	10.77	**13.11**	13.11	10.77	6	2	1
155	52	9.19	6.72	**13.93**	13.93	6.72	6	0	0
156	27	**12.09**	9.68	11.18	12.09	9.68	4	3	1
157	42	3.97	4.93	**5.35**	5.35	3.97	5	5	9
158	36	9.16	8.27	**14.58**	14.58	8.27	4	1	6
159	37	7.27	5.72	**12.26**	12.26	5.72	6	4	0
160	37	**9.98**	7.73	9.55	9.98	7.73	6	5	0
161	48	**9.46**	3.26	7.21	9.46	3.26	8	5	8
162	32	6.80	4.98	**7.29**	7.29	4.98	6	8	2
163	49	5.22	5.71	**8.17**	8.17	5.22	6	3	22
164	46	7.40	3.98	**10.86**	10.86	3.98	6	5	0
165	49	7.62	8.21	**15.05**	15.05	7.62	5	3	7

(continued)

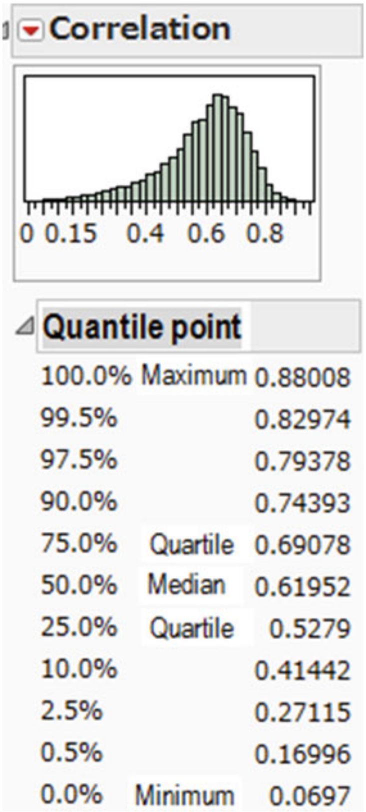
Table 5.7 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
166	44	9.79	9.50	**11.86**	11.86	9.50	5	2	1
167	45	4.25	2.51	**6.66**	6.66	2.51	12	6	6
168	49	9.08	1.84	**9.56**	9.56	1.84	7	3	6
169	46	6.81	2.75	**7.90**	7.90	2.75	10	2	0
170	49	2.76	3.57	**4.17**	4.17	2.76	12	7	0
171	43	**5.32**	3.03	4.22	5.32	3.03	17	3	0
172	40	3.99	4.01	**5.53**	5.53	3.99	16	5	13
173	41	2.65	1.53	**3.17**	3.17	1.53	11	9	0
174	45	7.46	4.84	**9.12**	9.12	4.84	15	4	10
175	43	4.34	1.52	**4.82**	4.82	1.52	21	3	4
176	42	**6.71**	2.79	5.26	6.71	2.79	23	3	30
177	39	0.63	0.72	**1.00**	1.00	0.63	23	9	28
178	32	0.52	1.52	**2.30**	2.30	0.52	21	4	17
179	46	5.41	5.08	**5.83**	5.83	5.08	21	2	25
Max	54	28.80	25.10	33.34	33.34	24.47	23	9	30
Min	11	0.52	0.72	1.00	1.00	0.52	0	0	0
Average	33.4637	13.66	13.05	16.68	17.09	11.89	2.1229	1.5307	2.5363
Max SMs		28	18	133	Misclassified		68	112	59
Total	5,990				SMs				

5.3.2 *Correlation Coefficient of Discriminant Score of 179 RIPs*

Figure 5.1 shows the distribution of correlation coefficients of 179 RipDSs. The correlation coefficient varies widely in the range of 0.07–0.88, and it is a unimodal distribution in which the skirt spreads to the lower value side. Due to the diversity of cancer, it is necessary to determine which RipDS group having high correlation coefficients are complementary to each other, or which RipDS having small correlations represent different kinds of cancer; it is a future research topic.

Fig. 5.1 Distribution of 179 RipDSs correlations



Correlation

Quantile point

100.0%	Maximum	0.88008
99.5%		0.82974
97.5%		0.79378
90.0%		0.74393
75.0%	Quartile	0.69078
50.0%	Median	0.61952
25.0%	Quartile	0.5279
10.0%		0.41442
2.5%		0.27115
0.5%		0.16996
0.0%	Minimum	0.0697

Table 5.8 shows correlation coefficient pairs of 179 RipDSs arranged in descending order by correlation coefficient values. The table shows only the top ten sets with high correlation coefficient and the ten lower groups with lower correlation coefficient. Of the ten pairs taking values from 0.86 to 0.88, five sets of RIP21 and (RIP54, RIP36, RIP1, RIP18, RIP45) have a correlation coefficient of RIP21 and 0.86 or more. These five pairs are expected to be the core of discrimination between AML and ALL. On the other hand, the lower ten sets have correlation coefficients of 0.07–0.1, and the eight pairs of RIP178 and (RIP177, RIP56, RIP133, RIP163, RIP128, RIP150, RIP172, RIP170) have low correlation with each other. Golub et al. have said they are studying class prediction to discover new cancer classes and assign tumors to known classes from 1970. Because the conventional biological insights are difficult to classify cancer with a systematic and unbiased approach, they used microarray to classify cancer by the co-expression of thousands of genes. They said that they developed a more systematic approach to cancer and started to discover variants of cancer. Then, SOC is divided into two classes, and they evaluate two clusters by the weighted voting method with LOO. However, our analysis indicates ten pairs of DS with a small correlation coefficient of 0.1 or less are different DSs

with no correlation with each other. We think that it is meaningful to study these medically. Chapter 4 indicates the correlations of genes included in each SM take the positive, almost zero, and negative values, and all correlations of 179 RipDSs are positive values. Thus, we have considered signal data is a reliable signal.

Table 5.8 Correlation coefficient pairs of 179 RipDSs

	Var.	Versus var.	Correlation	Frequency	Lower 95%	Upper 95%	p-value
1	***RIP54***	***RIP21***	0.88	72	0.81	0.92	0.00
2	***RIP36***	***RIP21***	0.88	72	0.81	0.92	0.00
3	*RIP21*	*RIP1*	0.87	72	0.80	0.92	0.00
4	*RIP21*	*RIP18*	0.87	72	0.80	0.92	0.00
5	RIP24	RIP5	0.87	72	0.80	0.92	0.00
6	RIP95	RIP23	0.87	72	0.80	0.92	0.00
7	RIP36	RIP1	0.87	72	0.80	0.92	0.00
8	**RIP45**	**RIP21**	0.86	72	0.79	0.91	0.00
9	RIP58	RIP18	0.86	72	0.79	0.91	0.00
10	RIP36	RIP31	0.86	72	0.79	0.91	0.00
–	–	–	–	–	–	–	–
15922	RIP178	RIP177	0.10	72	−0.14	0.32	0.41
15923	RIP170	RIP102	0.10	72	−0.14	0.32	0.42
15924	RIP178	RIP56	0.09	72	−0.14	0.32	0.43
15925	RIP178	RIP133	0.09	72	−0.14	0.32	0.44
15926	RIP178	RIP163	0.09	72	−0.14	0.32	0.44
15927	RIP178	RIP128	0.09	72	−0.15	0.31	0.46
15928	RIP178	RIP150	0.08	72	−0.16	0.31	0.51
15929	RIP177	RIP167	0.08	72	−0.16	0.30	0.53
15930	**RIP178**	**RIP172**	0.07	72	−0.16	0.30	0.55
15931	**RIP178**	**RIP170**	0.07	72	−0.16	0.30	0.56

Figure 5.2 is a correlation matrix of pairs of RIP36 and RIP54 highly correlated with RIP21 and a combination of RIP170 and RIP172 having a low correlation with RIP178. There are no correlations between scatter plots of low correlation pairs. Moreover, there are many patients on SV $= -1$ or 1. Scatter plots between (RIP21, RIP36, RIP54) and (RIP170, RIP172, RIP178) indicate the surprising fact. If we ignore the outliers of (RIP170, RIP172, RIP178), the variations of these three variables are smaller than (RIP21, RIP36, RIP54). If this is a general trend, in addition to choosing RipDSs with large RatioSV values, it is conceivable to use the high correlations pairs of RipDSs.

Future research topic: We must confirm the above fact in the future theme.

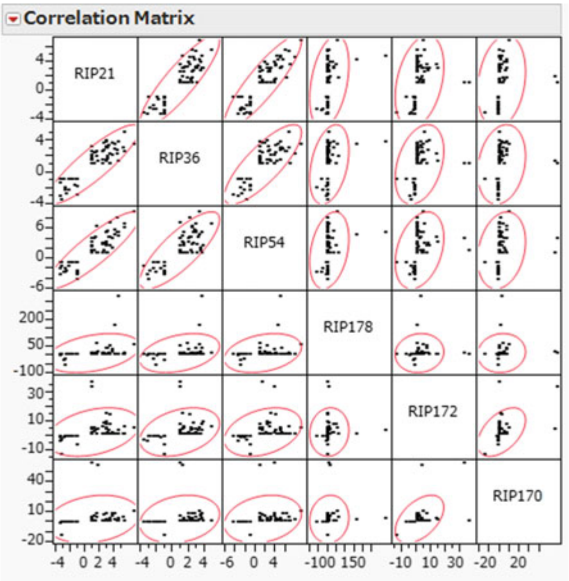


Fig. 5.2 Correlation matrix of six variables

5.4 Verification of SM3 and SM179

5.4.1 RatioSV of SM3 and SM179

Table 5.9 is the comparison of SM3 and SM179 because all results of 179 SMs take many pages. Although RIP chooses SM3 early, RatioSV of RIP is small as 4.745%. Although SM179 was the last selected, RatioSV of RIP is fairly large 5.41%. The numbers of RIP's outliers are larger than Table 5.5. This indicates RIP find the special data structure of RatioSV = 100%. Importantly, when judging the signal space of 179 SMs in Table 5.5, all patients locate on SV. However, if microarray is divided into 179 SMs, the "RatioSV and Outlier" of RipDSs are the same as for other LDFs. We want to explain this fact more clearly in the future.

Table 5.9 Comparison of SM3 and SM179

	SM3						SM179					
	RIP	LP	IPLP	H-SVM	SVM4	SVM1	RIP	LP	IPLP	H-SVM	SVM4	SVM1
Ratio	**4.745**	5.194	5.194	**5.179**	5.179	10.48	**5.41**	5.077	5.091	**5.834**	5.829	80.66
Outlier	**18/37**	18/37	18/36	**19/37**	19/37	NM = 2	**10/16**	12/15	11/18	**11/18**	11/19	NM = 21
−1	−1	−1	−1	−1.61	−1.61	−1	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−9	−13.3	−13.3	−10	−10	−5.56	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−9.53	−13.9	−13.9	−11.7	−11.7	−6.23	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−1	−5.64	−5.64	−3.92	−3.92	−1.81	−1	−1	−1	−1	−1	−1
−1	−6.83	−5.12	−5.12	−4.83	−4.83	−1.81	−1	−1	−1	−1	−1	−1
−1	−7.11	−9.41	−9.41	−7.15	−7.15	−3.04	−1	−1	−1	−1	−1	**0.97**
−1	−1	−1	−1	−1	−1	−1.02	−1	−1	−1	−1	−1	−0.66
−1	−13	−9.5	−9.5	−8.81	−8.81	−3.88	−1.7136	−1.37	−1.42	−1.25	−1.25	**0.97**
−1	−3.29	−1	−1	−2.14	−2.14	−2.32	−1	−1	−1	−1	−1	**0.84**
−1	−3.4	−5.53	−5.53	−7.64	−7.64	−3.89	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−1	−1	−1	−1	−1	−0.02	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−3.75	−3.6	−3.6	−3.32	−3.32	−2.42	−1	−1	−1	−1	−1	**0.89**
−1	−1	−2.69	−2.69	−2.77	−2.77	−1.09	−1	−1	−1	−1	−1	−1
−1	−1	−1	−1	−1	−1	**0.15**	−1	−1	−1	−1	−1	**0.02**
−1	−1.69	−4.79	−4.79	−4.88	−4.88	−2.46	−1	−1	−1	−1	−1	**0.93**
−1	−3.82	−2.4	−2.4	−3.46	−3.46	−1	−1	−2.21	−1	−1	−1	**0.94**
−1	−8.55	−11.9	−11.9	−10.3	−10.3	−4.73	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−3.83	−6.69	−6.69	−5.24	−5.24	−2.96	−1	−1.17	−1.22	−1.05	−1.06	**0.85**

(continued)

Table 5.9 (continued)

	SM3						SM179					
	RIP	LP	IPLP	H-SVM	SVM4	SVM1	RIP	LP	IPLP	H-SVM	SVM4	SVM1
−1	−1	−1	−1	−1	−1	−1	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−2.12	−1.54	−1.54	−1	−1	−1	−1	−1	−1	−1	−1	**1.01**
−1	−17	−13.6	−13.6	−15.6	−15.6	−7.56	−1	−1	−1	−1	−1	**0.92**
−1	−9.62	−11.1	−11.1	−9.15	−9.15	−4.75	−1	−1	−1	−1	−1	**0.97**
−1	−8.75	−6.02	−6.02	−5.65	−5.65	−2.79	−1.2239	−1.17	−1.22	−1.05	−1.05	**0.97**
−1	−2.7	−1	−1	−1	−1	−0.97	−1	−1	−1	−1	−1	**0.9**
−1	**−2.42**	**−3.05**	**−3.05**	**−3.18**	**−3.18**	**−2.5**	**−2.862**	**−4.19**	**−2.96**	**−2.95**	**−2.95**	**0.63**
1	**6.953**	**7.396**	**7.396**	**5.301**	**5.301**	**3.427**	**1**	**1**	**1**	**1**	**1.001**	**1**
1	5.83	3.115	3.115	7.477	7.477	4.029	1	1	1	1.424	1.423	0.912
1	1	1	1	1	1	1	1	1	1	1	1	1
1	8.874	6.984	6.984	8.971	8.971	4.351	1	1	1	1	1	1.052
1	15.74	13.99	13.99	15.22	15.22	8.197	4.15996	5.405	4.249	2.968	2.97	1.1
1	3.538	3.818	3.818	2.364	2.364	2.06	6.40324	7.675	6.401	6.229	6.229	0.894
1	15.22	10.77	10.77	11.87	11.87	6.22	3.94724	1	1	1	1	0.974
1	1	5.41	5.41	1	1	1	1	1	1	1	1	1
1	9.497	15.21	15.21	9.347	9.347	4.951	1	1	1	1	1	0.979
1	12.25	13.41	13.41	12.05	12.05	6.474	6.10114	1	1	1	1	1
1	12.99	10.04	10.04	12.68	12.68	6.27	5.55466	5.851	5.897	5.03	5.03	1
1	1	1	1	1	1	1	1.06825	1	1.097	1	1	0.985
1	1.63	3.546	3.546	2.687	2.687	1.165	1	1	1	1	1	0.996
1	4.66	6.528	6.528	4.279	4.279	2.779	4.99587	4.541	4.979	3.367	3.352	1.078

(continued)

Table 5.9 (continued)

	SM3						SM179					
	RIP	LP	IPLP	H-SVM	SVM4	SVM1	RIP	LP	IPLP	H-SVM	SVM4	SVM1
1	17.85	18.02	18.02	16.97	16.97	8.805	4.54821	4.94	4.539	3.602	3.592	1.138
1	7.853	9.501	9.501	9.764	9.764	5.371	4.55515	6.15	4.708	5.455	5.447	1
1	1	1	1	1	1	**−0.5**	1	1	1	1	1	1.02
1	1	1	1	1	1	1	1	1	1.062	1	1	0.847
1	1	1.011	1.011	1	1	1.117	1	1	1	1	1	1
1	7.835	3.495	3.495	5.685	5.685	3.363	13.2176	9.419	24.93	8.272	8.272	1
1	6.664	7.331	7.331	8.383	8.383	4.456	1	1	1	1	1	1
1	4.769	1	1	1	1	1	7.40609	6.883	7.387	5.578	5.578	0.953
1	10.97	6.597	6.597	7.953	7.953	5.381	1	1	1	1	1	0.914
1	11.74	9.183	9.183	13.06	13.06	6.5	1	1	1	1	1	1
1	5.926	7.679	7.679	7.631	7.631	4.088	2.19083	2.364	2.228	1.462	1.462	0.94
1	19.01	11.72	11.72	12.76	12.76	7.927	1	1	1	1	1	1
1	1	3.842	3.842	2.761	2.761	1.986	1	1	1	1	1	0.932
1	16.86	11.21	11.21	12.45	12.45	7.056	1	4.836	6.502	1.865	1.864	1.077
1	15.18	7.555	7.555	9.055	9.055	5.944	1	1	1	1	1	1.08
1	20.08	16.26	16.26	15.58	15.58	8.461	1	1	1	1	1	1
1	18.19	12.69	12.69	13.86	13.86	7.741	1	1	1	1	1	0.963
1	2.374	1	1	2.339	2.339	1	1	1	1	1	1	0.966
1	13.71	10.58	10.58	10.87	10.87	6.3	1	1	1	1.039	1.053	1
1	8.298	4.309	4.309	6.146	6.146	4.431	20.1173	8.239	8.302	8.092	8.099	1
1	11.11	12.38	12.38	12.69	12.69	6.687	1	1	7.27	1	1	1.024

(continued)

Table 5.9 (continued)

	SM3						SM179					
	RIP	LP	IPLP	H-SVM	SVM4	SVM1	RIP	LP	IPLP	H-SVM	SVM4	SVM1
1	7.359	9.789	9.789	8.421	8.421	4.807	1	3.904	1	11.42	11.42	1
1	1	1	1	1	1	1	1	1	1	10.1	10.09	1
1	2.326	1	1	1.545	1.545	1.528	1	1	1	1	1	1
1	3.66	5.005	5.005	5.766	5.766	3.509	1	1	1	1	1	1
1	25.18	24.58	24.58	22.98	22.98	11.51	33.4101	35.2	36.32	31.34	31.36	1.479
1	4.884	7.046	7.046	7.761	7.761	3.135	34.0999	22.37	23.7	11.72	11.72	1
1	9.063	4.514	4.514	6.159	6.159	3.922	1	1	1.533	1	1	0.979
1	7.164	4.73	4.73	4.178	4.178	3.381	1	1	1	1	1	0.991
1	13.24	15.39	15.39	14.38	14.38	7.35	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1.024
1	12.37	11.99	11.99	11.02	11.02	6.051	1.67956	1.585	1.676	1.348	1.348	0.997
1	1	1	1	1	1	1	1	1	1	1	1	1

5.4.2 T-Test of RIP3 and RIP179

Table 5.10 is a t-test in which the variance of the difference between the average values of ALL and AML of 17 genes included in SM3 and 46 genes included in SM179 is different. In some studies, there are things like looking for ones with large t-values as oncogenes. However, this is a mistake. In the SM3, there are six genes having t-values higher than 1 and four genes for t-values less than −2. Other seven genes are from −0.743 to 1. In SM179, twelve t-values are higher than 1, and two t-values are less than −1. Other 32 genes are from −1 to 1. In both cases, the genes that are thought to have no difference in the average value of the expression levels of AML and ALL are the most frequent. Although these genes may not be distinguished by themselves, we consider that these genes are essential for discrimination in combination with other genes. Moreover, those that become positive are probably cancer genes, while those that become negative are suppression genes. However, those genes that have no differences in mean on t-test are considered to be necessary for the discrimination, also. That is, it is impossible to identify cancer genes by t-test.

Table 5.10 T-test with 17 genes in SM3 and 46 genes in SM179

RIP3				RIP179			
X	Y	Difference	T-value	X	Y	Difference	T-value
X7130	X4375	1.625	4.977	X7130	X3186	0.034	2.156
X7130	**X4621**	0.978	2.971	X7130	X2709	0.041	1.978
X7130	**X4677**	0.731	2.839	X7130	X6503	0.059	1.944
X7130	**X5088**	1.175	2.677	X7130	X3173	0.025	1.768
X7130	**X4625**	0.498	1.855	X7130	X1754	0.019	1.654
X7130	X4566	0.212	1.381	X7130	X5203	0.015	1.415
X7130	X7127	0.004	1.000	X7130	X6235	0.025	1.318
X7130	X4045	0.191	0.969	X7130	X1556	0.018	1.235
X7130	s3X3762	0.123	0.560	X7130	X1237	0.008	1.122
X7130	**X5081**	0.051	0.114	X7130	X2535	0.031	1.050
X7130	X4062	−0.066	−0.176	X7130	X3589	0.029	1.046
X7130	X4619	−0.253	−0.593	X7130	X5000	0.025	1.038
X7130	**X6171**	−0.161	−0.743	X7130	X6085	0.003	1.000
X7130	**X6196**	−0.945	−2.394	X7130	X2747	0.005	1.000
X7130	**X5987**	−0.792	−2.461	X7130	X4080	0.005	1.000
X7130	**X6218**	−2.966	−5.867	X7130	s179X5	0.004	1.000
X7130	**X6201**	−3.537	−6.240	X7130	X3001	0.007	1.000

(continued)

Table 5.10 (continued)

RIP3				RIP179			
X	Y	Difference	T-value	X	Y	Difference	T-value
				X7130	X2466	0.009	1.000
				X7130	X888	0.013	1.000
				X7130	X1148	0.018	1.000
				X7130	X320	0.014	0.869
				X7130	X2465	0.009	0.838
				X7130	X3470	0.031	0.674
				X7130	X3610	0.008	0.388
				X7130	X1841	0.004	0.276
				X7130	X5616	0.002	0.024
				X7130	X5961	−0.002	−0.067
				X7130	X4009	−0.002	−0.108
				X7130	X2790	−0.005	−0.151
				X7130	X1489	−0.005	−0.246
				X7130	X3212	−0.023	−0.538
				X7130	X1388	−0.019	−0.555
				X7130	X5724	−0.018	−0.628
				X7130	X5222	−0.050	−0.733
				X7130	X2192	−0.095	−0.786
				X7130	X6718	−0.055	−0.827
				X7130	X1637	−0.002	−0.857
				X7130	X3143	−0.014	−0.875
				X7130	X4603	−0.083	−0.921
				X7130	X5038	−0.013	−1.000
				X7130	X1783	−0.038	−1.000
				X7130	X3976	−0.042	−1.000
				X7130	X6541	−0.082	−1.000
				X7130	X4338	−0.089	−1.000
				X7130	X2310	−0.038	−1.016
				X7130	X1707	−0.086	−1.289

5.4.3 BGS and Yamanaka's Four Genes of IPS Research

LINGO Program4 is a program that directly finds BGS. We will explain it with Alon's microarray having 2,000 genes. Six microarrays are LSD (Fact3). If we omit the first gene and a set having 1,999 genes is LSD, Program4 omits the first gene. If the set having 1,999 genes is not LSD, the first gene is the necessary gene to be LSD, so restore it. Program4 repeats this simple operation until the 2,000th gene. In this way, Program4 obtains the first BGS (BGS1). This procedure is the same as the way Dr. Yamanaka's group found four genes from 24 genes in iPS research. That is, Yamanaka's four genes are the same as BGS. In our study, we have searched for BGS1 in 2,000 genes instead of 24 genes, so it took about 30 min to compute. We stopped here and omitted BGS1 from 2,000 genes. Next, we found the second BGS (BGS2). By repeating this operation, we found 130 pairs of BGS in about one week. We regret not to notice the relationship with iPS research sooner. The four genes contain carcinogenic c-myc. Instead, they took many efforts to find a L-myc without carcinogens. If our research is applicable, we omit the first four genes from 24 genes and find something equivalent to BGS2 that may include L-myc; we believe we could contribute to iPS cell research. However, even if we add another gene to the gene of LSD, it is LSD. On the other hand, in iPS cell research, it is decisively different in that adding other genes to the set of 24 genes has a limit not to become iPS cells.

If Program4 analyzes SM3 having 17 genes, we can see that a set of 10 bold genes in Table 5.10 is BGS in a short time. If we omit either of X5081 or X6171, those NMs are three.

In addition, when we discriminate 5668 subjects (1864 cancer and 3804 healthy) by 47 oncogenes found medically, the error rate is 22.6182%. Although it is clinical center data including many patients who have received treatment such as surgery and have a poor prognosis, we believe we cannot discriminate even strictly managed data by oncogenes alone. Only SMs are useful for cancer gene diagnosis.

This claim realizes the dream of Golub et al. That is, since oncogenes found by conventional medical approach is insufficient for medical diagnosis, they wished to establish a method to systematically analyze information possessed by microarray. Perhaps we think our method is useful for achieving their purpose after they will examine and validate our results.

5.4.4 PCA and Cluster Analysis

Figure 5.3 shows eigenvalues, scatter plot, and factor loading plot of SM3. Although the three RatioSVs of three OLDFs are about 5%, they overlap perfectly. The figures displayed in the scatter plot are SNs. The 25 AML patients are SN = 1,…, 25, and the 47 ALL patients are SN = 26, …, 72. We can recognize that both classes overlap. We cannot find the linear separable facts of AML and ALL on the scatter plot of Prin1 and the Prin2 that can show the large data variations. This fact indicates that to find the LSD is irrelevant to large data variations. There is a possibility that it may be possible to capture signs of linear separation with high-order principal components having less data fluctuation. However, three OLDFs can easily catch the linear separability. At the time of 2015, we thought that the standard statistical methods offer the critical meaning of cancer gene analysis. However, we conclude that those methods, except for logistic regression, are utterly meaningless now. Therefore, initially we considered SM as a signal, but we reconsidered the signal data as a signal.

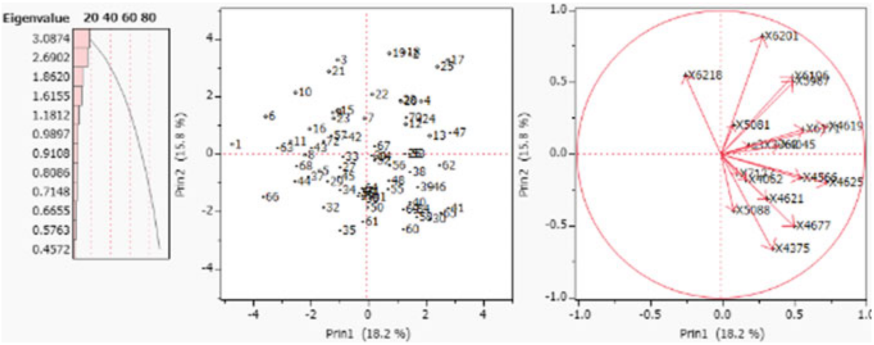


Fig. 5.3 Eigenvalues, scatter plot, and factor loading plot of SM3

Figure 5.4 is Ward cluster analysis of SM3. The left black cases are 25 AML patients, and the white color cases are 47 ALL patients. Although we tried to interpret of this result, it came to think that it had no meaning as same as PCA in Fig. 5.3.

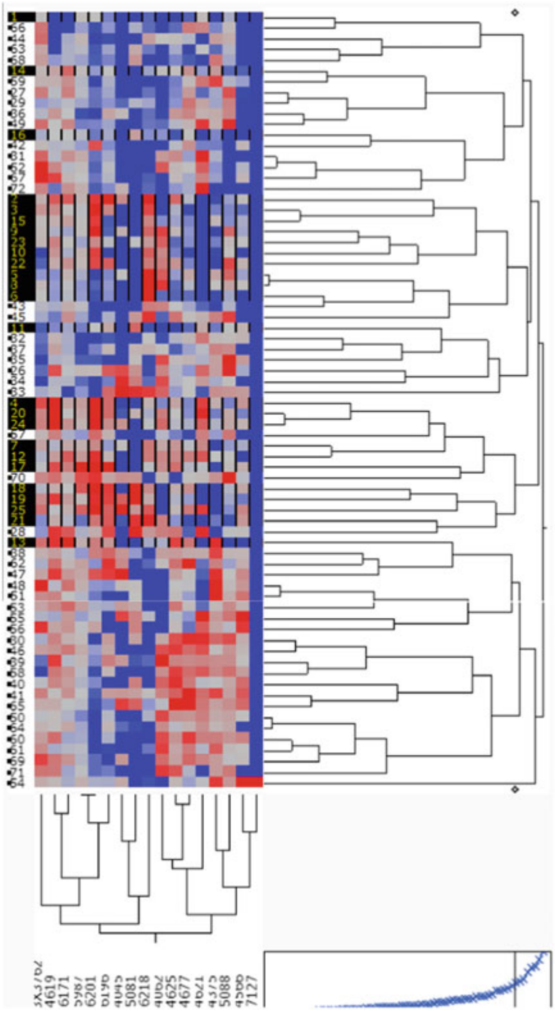


Fig. 5.4 Ward cluster analysis of SM3

5.5 Analysis of Signal Data Made by 179 RipDSs

When logistic regression discriminates 179 SMs, all NMs become 0. As Table 5.7 shows, SVM1, LDF2, and QDF cannot discriminate 23, 9, and 30 cases, respectively. Also, all NMs of SVM4 are 0, but many NMs of SVM1 are not 0. There are no signs of linear separation in PCA or cluster analysis. Thus, cluster analysis and PCA analyze the signal data made by RIP, LP, HSVM. The 64 SMs obtained by LP get almost the same result, but we omit the results.

5.5.1 *Cluster Analysis and PCA of RipDSs Signal Data*

(1) Ward Cluster

Figure 5.5 is a cluster analysis of RipDS signal data. The upper blue part is 25 cases of AML, and the red part is 47 cases of ALL. If we choose five clusters, AML becomes one cluster and ALL consists of four clusters. RIP66 (Brawn) and RIP65 (Pale green) of bottom two DSs become two different clusters.

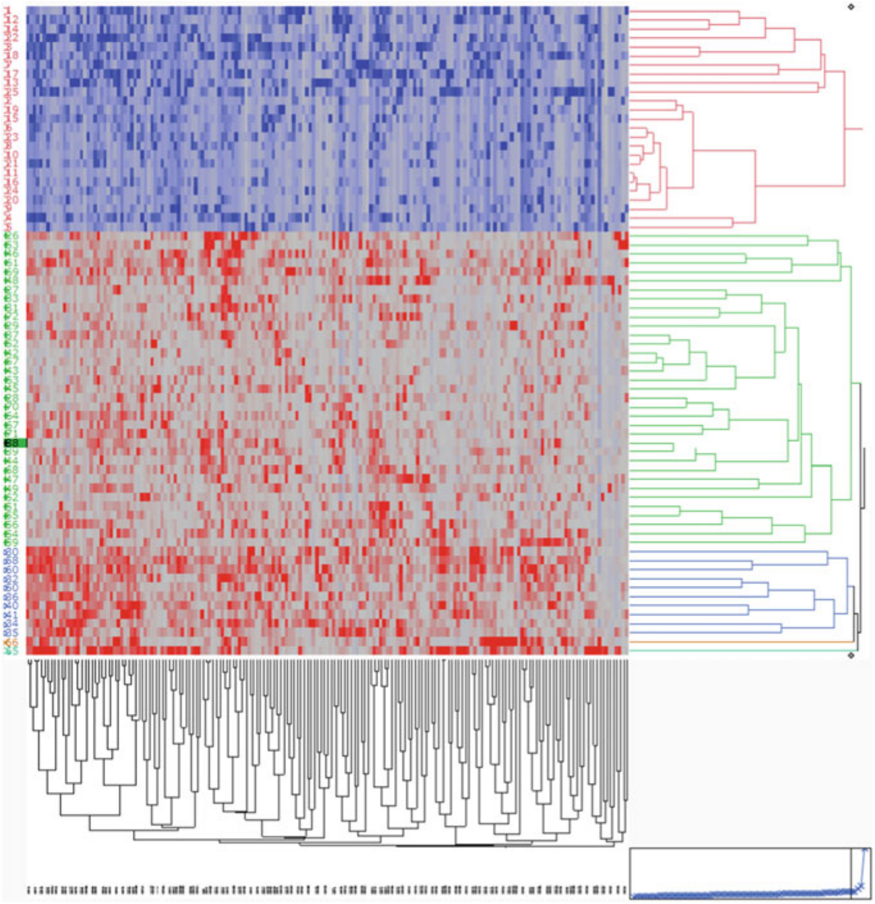


Fig. 5.5 Cluster analysis of RipDSs signal data

(2) PCA

Figure 5.6 shows the result of RipDS signal data by PCA. From the left eigenvalue, the eigenvalue of the Prin1 is larger than the others. The first eigenvalue is 109.327, and the contribution rate is 61.077%. The second eigenvalue is 4.756, the contribution rate is 2.65%, and the cumulative contribution rate is 63.734%. That is, the Prin1 presents 179 RipDS signal data. From the scatter plot in the middle, because the second eigenvalue is a small value and the variance is small, AML cases are almost placed on the Prin1 axis of −7 or less. It is as same as the healthy subjects in Alon and Singh's microarrays. The range of ALL is [2.556, 16.713], and the variation of the Prin2 becomes slightly large as it departs from AML. This result is similar to the result of Alon and Singh between cancer and healthy subjects. The case of ALL with SN = 65 is a clear outlier. The Prin1 becomes a malignancy indicator as well as individual RipDS.

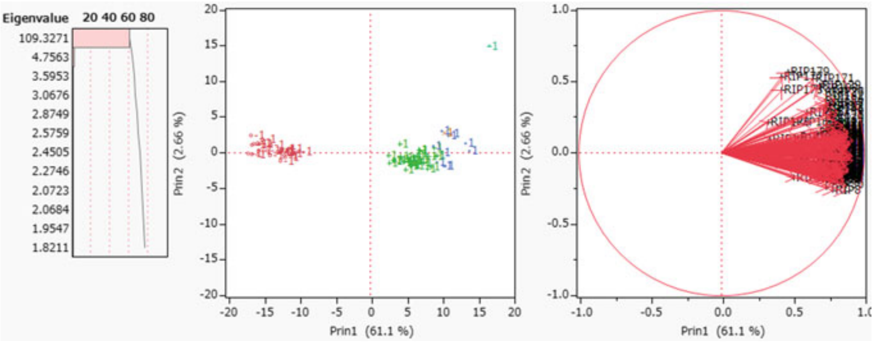


Fig. 5.6 Result of RipDS signal data by PCA

(3) RIP, LP, and H-SVM Prin1 Values and RatioSV

Third rows or less of the first column of Table 5.11 indicates the value of SN. The second column is the value of the Prin1, and we sort this value in ascending order from a small value. In Fig. 5.6, the left end is the SN = 25 belonging to AML, and the value of the Prin1 is −17.071. The SN = 11 of AML has a value of −9.986, which is closest to ALL class, and the range of AML is [−17.071, −9.986].

On the other hand, the SN = 42 of ALL class is closest to AML and is 2.556. The SN = 65 of ALL class is 16.713, and the range of ALL is [2.556, 16.713]. A window is open in the interval (−9.986, 2.556), and when RatioSV of PCA is calculated for the range [−17.071, 16.713] on the Prin1, a large window of 37.122% opens. AML and ALL locate on two separate ranges those are in 63% remaining. Because it is comprehensive of 179 RipDSs, it is 8.3% larger than the maximum value of 179 RipDSs, which is 28.799%. That is, the Prin1 axis of PCA becomes a malignancy indicator.

Table 5.11 RIP, LP, and HSVM Prin1 values and RatioSV

RatioSV	37.122	30.806		31.100	
SN	RIPprin1	SN	LPprin1	SN	HSVMprin1
25	**−17.071**	**25**	**−17.518**	**17**	**−19.364**
18	−16.640	17	−17.160	25	−18.822
22	−16.577	18	−16.923	18	−18.272
17	−16.504	22	−16.547	22	−17.949
3	−16.223	1	−15.767	1	−17.217
1	−15.709	4	−15.660	3	−16.681
12	−15.618	3	−15.564	12	−16.590
13	−15.403	12	−15.058	4	−16.430
4	−15.119	13	−14.376	14	−15.039
7	−14.461	15	−14.273	7	−14.997
14	−13.985	7	−14.222	13	−14.969
15	−13.822	14	−14.147	15	−14.958
23	−12.949	2	−13.934	2	−14.220
21	−12.580	19	−13.217	23	−13.976
5	−12.563	23	−13.016	5	−13.754
8	−12.542	6	−12.759	19	−13.683
2	−12.438	5	−12.726	21	−13.575
9	−12.380	21	−12.659	6	−13.086
6	−12.361	10	−12.225	8	−12.982
19	−12.242	8	−12.177	10	−12.807
16	−11.985	16	−11.888	20	−12.608
20	−11.927	20	−11.862	16	−12.429
24	−11.395	9	−11.459	9	−12.010
10	−11.392	24	−11.231	24	−11.770
11	**−9.986**	**11**	**−9.826**	**11**	**−10.381**
42	**2.556**	**57**	**1.236**	**57**	**1.884**
57	2.724	42	1.820	42	1.918
27	3.261	27	2.849	43	2.957
43	3.277	43	2.857	27	3.326
62	4.121	63	3.683	70	4.350
70	4.235	70	4.071	47	4.563
52	4.405	37	4.431	63	4.751
63	4.720	52	4.457	52	4.820
55	4.743	55	4.576	28	4.875
28	4.802	28	4.631	55	4.901
33	4.950	67	4.769	62	5.130
51	5.374	47	4.934	37	5.230

(continued)

Table 5.11 (continued)

RatioSV	37.122	30.806		31.100	
SN	RIPprin1	SN	LPprin1	SN	HSVMprin1
47	5.399	38	5.048	33	5.267
37	5.421	72	5.155	38	5.423
44	5.636	51	5.392	72	5.586
67	5.670	44	5.736	29	5.606
72	5.725	29	5.780	67	5.814
38	5.927	33	5.781	44	6.006
31	5.961	62	5.916	51	6.252
29	5.971	53	6.209	45	6.363
45	6.090	31	6.286	31	6.368
53	6.378	59	6.561	53	6.563
39	6.496	45	6.596	59	7.030
49	6.766	26	6.727	68	7.102
64	6.808	64	6.828	39	7.269
59	6.976	39	6.853	56	7.527
71	7.023	71	7.015	49	7.611
68	7.189	68	7.064	71	7.719
56	7.507	49	7.193	26	7.782
26	7.995	36	7.225	64	7.797
54	8.185	56	7.298	36	9.090
46	8.598	35	8.262	54	9.128
50	8.748	50	9.062	48	9.430
48	8.990	66	9.253	46	9.502
69	9.090	61	9.503	35	9.842
36	9.318	54	9.596	61	10.098
61	9.394	46	9.735	50	10.307
35	9.947	32	9.845	32	10.501
41	9.998	48	9.869	69	10.595
32	10.298	34	10.245	41	11.142
66	10.337	41	10.636	34	11.333
34	10.572	69	10.931	66	11.426
40	10.951	40	11.034	40	12.174
30	11.335	30	14.306	30	14.108
58	13.363	60	14.670	60	15.586
60	13.930	58	15.876	58	16.443
65	**16.713**	**65**	**18.391**	**65**	**20.074**

5.5.2 Cluster Analysis and PCA of HsvmDSs Signal Data

(1) Ward Cluster

Figure 5.7 is a cluster analysis of HsvmDS signal data. The upper blue part is 25 cases of AML, and the red part is 47 cases of ALL. If we choose five clusters, AML becomes two clusters and ALL consists of three clusters. HSVM65 (Pale green) becomes one cluster. In Sect. 5.5.1, AML is mild cancer compared with ALL, showing a tendency similar to the normal class of Alon and Singh. However, in the signal data made by H-SVM, it was divided into two clusters. It is easy to obtain different clusters by changing the number of clusters in three signal data by five hierarchical cluster analyses. That is, it may be suitable for exploring new cancer subclasses pointed by Golub et al.

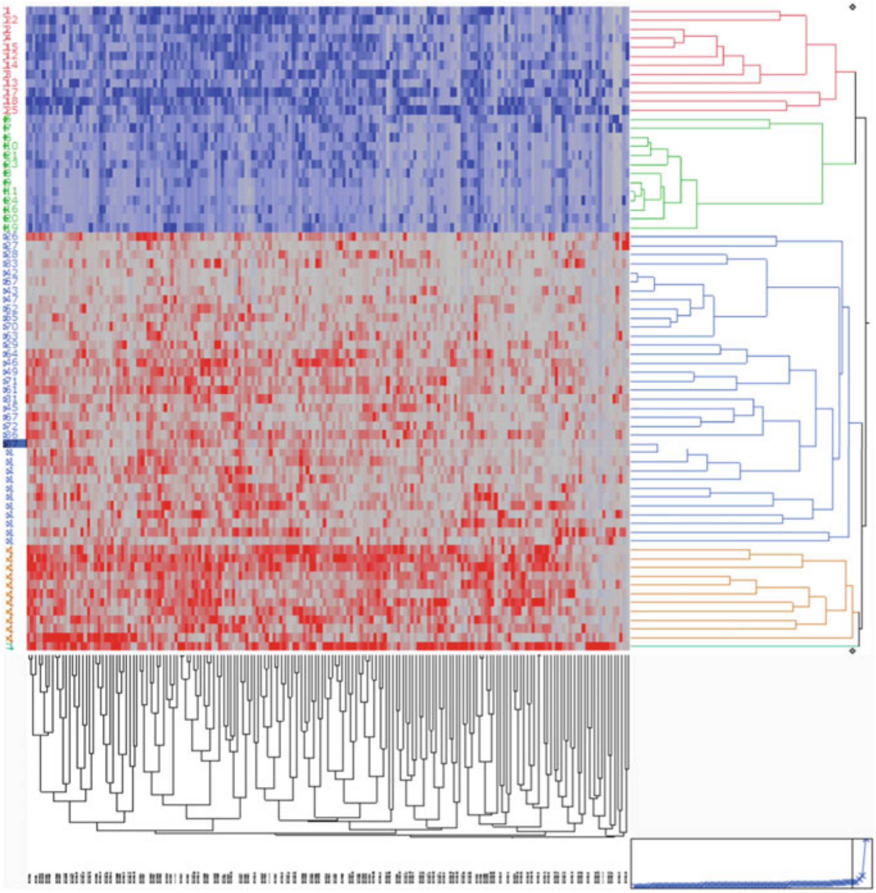


Fig. 5.7 Ward cluster analysis of HsvmDSs signal data

Figure 5.8 shows the result of HsvmDS signal data by PCA. The eigenvalues of the Prin1 are larger than others. The first eigenvalue is 128.23, and the contribution rate is 71.64%. The second eigenvalue is 4.496, the contribution rate is 2.51%, and the cumulative contribution rate is 74.15%. That is, the Prin1 presents 179 HsvmDSs signal data. From the scatter plot, because the second eigenvalue is small and the variance is small, AML cases are almost placed on the Prin1 axis of −10 or less. The ALL patients are in the range [1.88, 20.07], and the variation of the Prin2 becomes large as it departs from AML class. In other words, the Prin1 becomes an indicator of malignancy indicator as well as individual HsvmDS. The fifth column in Table 5.11 is SN value, and the sixth column is the value of the Prin1 of the HsvmDS signal data, which was sorted in ascending order. The leftmost point is the SN = 17, and the value of the Prin1 is −19.36. The SN = 11 takes a value of −10.381, closest to ALL. The range of AML is [−19.36, −10.38]. On the other hand, the SN = 57 of ALL class is closest to AML and is 1.884. Moreover, the SN = 65 of ALL class is 20.074. The range of ALL class is [1.88, 20.07]. The window (−10.38, 1.88) opens, and when RatioSV of PCA is calculated for the range [−19.36, 20.07], a window of 31.1% open, which is smaller than 33.34% of the maximum value of RatioSV of HsvmDSs. In other words, rather than considering the Prin1 axis of PCA as a malignancy indicator, individual HsvmDSs are more useful for malignancy indicators. However, RatioSV of PCA is superior to visual explanation.

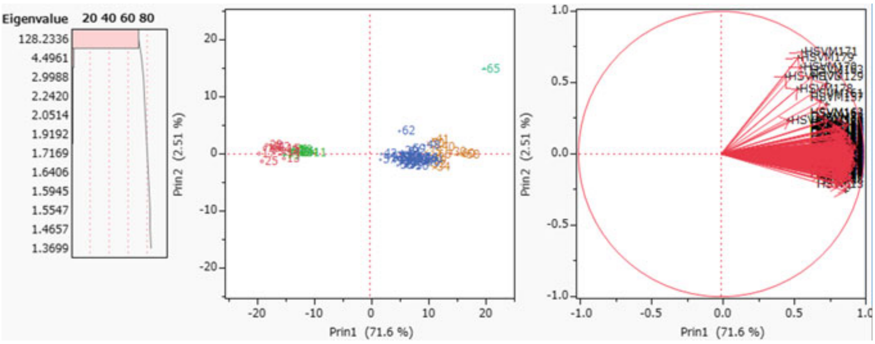


Fig. 5.8 HsvmDS signal data by PCA

5.5.3 Analysis of Transposed Data

Figure 5.9 shows the result of transposed signal data. RIP151, RIP3, RIP75, RIP78, and RIP178 are large outliers. These DSs probably indicate cancer diversity. We hope specialists verify this result. We can identify more outliers by the transposed data.

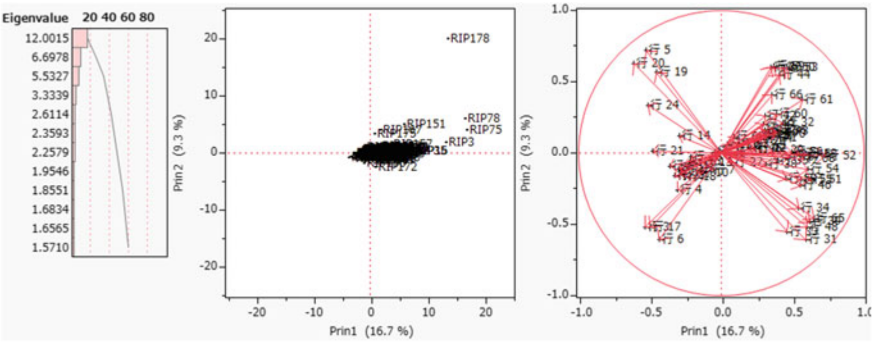


Fig. 5.9 Transposed data of RipDS signal data

Figure 5.10 shows the result obtained by transposed signal data of HsvmDS signal data. HSVM178, HSVM3, HSVM75, and HSVM177 have large outliers. These DSs are presumably indicative of cancer diversity. These validations are areas of specialists.

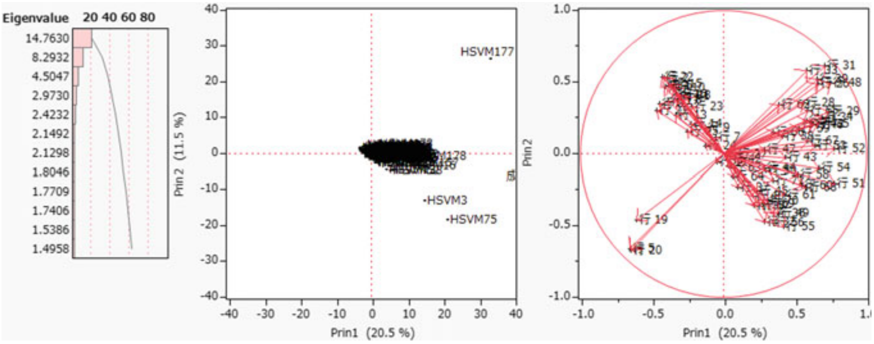


Fig. 5.10 Transposed data of HsvmDSs

5.6 Conclusions

Golub and colleagues used a non-hierarchical cluster analysis called SOP as a new class finding of cancers and verified candidates for newly discovered cancer variants with the weighted voting method and LOO. Many studies widely used cluster analysis because the statistical discriminant analysis was useless at all. The main reason is the following:

(1) Statistical discriminant functions assume that the two groups are normal distributions based on the variance–covariance matrix, and maximize the correlation ratio. For this reason, the medical researchers never knew that the two groups were LSDs in high-dimensional gene space.

(2) If researchers used RIP or H-SVM, they could find that six microarrays were NM = 0, but there is no paper pointed out this fact. Some studies use S-SVM. However, because NM may or may not become 0 by a set of genes obtained by medical consideration, it seems that the essential structure that gene space is LSD was not known. We do not know why H-SVM could not find the microarrays are LSD. Instead of the discriminant analysis, many medical researchers trusted the cluster analysis. Thus, they could not obtain the definitive results.

However, only three OLDFs could find six microarrays are LSD and many SMs. Because we had already established LSD-discrimination using common data, it was easy to obtain the same results. However, we found that the LSD has a Matryoshka structure and could decompose microarrays into many SMs and BGSs. Although all SMs and BGSs are small samples, the standard statistical methods, except for logistic regression, could not find the linear separable facts. Thus, we made signal data by RipDSs, LpDSs, and HsvmDSs. PCA and cluster analysis showed clear indications of linear separability. Moreover, we find new facts as follows:

(1) We confirm Revised LP-OLDF cannot find the correct sets of SMs as same as Alon's microarray in Chap. 4.

(2) Although the RatioSV of RIP using microarray and signal subspace is 100%, we can explain the reason and find the new fact of two classes in signal subspace. This fact answers the reason why no researchers could not find that microarrays and SMs were LSD from 1970. Although two classes are entirely separable in microarrays and SMs, the ranges of DSs are too small compared with microarray data variation. Thus, only three OLDFs can find these tiny variations. PCA and cluster analysis could not find the linear separable facts.

(3) The RatioSV value of the SM found earlier in Method2 is generally larger than the RatioSV of the SM found later, so SM with a large RatioSV may be useful for cancer gene diagnosis. Moreover, RipDSs having high correlations may offer useful clues of cancer gene diagnosis.

References

Cox DR (1958) The regression analysis of binary sequences (with discussion). J Roy Stat Soc B 20:215–242

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27–39

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Lachenbruch PA, Mickey MR (1968) Estimation of error rates in discriminant analysis. Technometrics 10(1):11

Sall JP, Creighton L, Lehman A (2004) JMP start statistics, 3rd edn. SAS Institute Inc. USA (Shinmura S. edits Japanese version)

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2010) Saiteki senkei hanbetsu kansu (The optimal linearly discriminant function). JUSE Press, Tokyo, Japan. ISBN 978-4-8171-9364-3

Shinmura S (2016a) New theory of discriminant analysis after R. Fisher. Springer, Tokyo

Shinmura S (2016b) The 100-fold cross-validation for small sample. Data Anal 2016:1–8

Shinmura S (2017) Cancer gene analysis by Singh et al. Microarray data. ISI2017:1–6

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD', vol 18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. Biocomp', vol 18, pp 1–7

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1): 68–74. (https://doi.org/10.1038/nm0102-6)

Vapnik V (1995) The nature of statistical learning theory. Springer

Chapter 6
Cancer Gene Diagnosis of Shipp et al. Microarray

Abstract Shipp microarray consists of 77 patients and 7,129 genes. They analyzed the microarray by various statistical methods and uploaded a supplemental document with 67 pages. They used almost the same methods as Golub et al. except for SVM and nearest neighbor cluster. Mainly, discriminant analysis is the most appropriate method to identify oncogenes from the microarray. However, because the statistical discriminant analysis was useless at all, medical researchers had developed many methods for cancer gene analysis. Our theory shows that six microarrays are LSD (Fact3). Method2 decomposes the microarrays into many SMs (Fact4). Then, by analyzing SM, we propose cancer gene diagnosis and malignancy indexes. If Shipp et al. validate our research results, we will improve cancer gene diagnosis. Method2 already obtained SM twice in Chap. 2. In this research, we change the number of iterations of RIP and Revised LP-OLDF in Method2 and decided the proper number of iterations. We obtain SMs by those iteration numbers in 2018. We examined the signal subspace made by all SMs and the noise space. However, Revised LP-OLDF cannot correctly find all SMs from Shipp microarray as same as Chap. 4. Thus, we analyze only 237 SMs obtained by the RIP and examine the correlation coefficient of RipDSs. RatioSVs evaluate RIP, Revised LP-OLDF, and H-SVM. Then, we analyze two signal data and transposed data made by RIP and H-SVM. By the hierarchical cluster analysis and PCA, we can propose the possibility of cancer gene diagnoses such as malignancy indexes.

Keywords Shipp microarray · Cancer gene diagnosis · Malignancy indexes · Defect of signal found by Revised LP-OLDF · Small Matryoshka (SM) · RatioSV of PCA · RipDSs · LpDSs · HsvmDSs

Thanks to Shipp et al.

We thank Shipp et al.[1] (2002) for providing excellent data and supplemental document. Below, we will quote “Abstract” for the reader. The downloaded microarray is slightly different from the abstract.

[1] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C.Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval1, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster & T. R. Golub.

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_6

"Diffuse large B-cell lymphoma (DLBCL), the most common lymphoid malignancy in adults, is curable in less than 50% of patients. Prognostic models based on pre-treatment characteristics, such as the International Prognostic Index (IPI), are currently used to predict outcome in DLBCL. However, clinical outcome models identify neither the molecular basis of clinical heterogeneity nor specific therapeutic targets. We analyzed the expression of 6,817 genes in diagnostic tumor specimens from DLBCL patients who received cyclophosphamide, adriamycin, vincristine, and prednisone (CHOP)-based chemotherapy, and applied a supervised learning prediction method to identify cured versus fatal or refractory disease. The algorithm classified two categories of patients with very different five-year overall survival rates (70% vs. 12%). The model also effectively delineated patients within specific IPI risk categories who were likely to be cured or to die of their disease. Genes implicated in DLBCL outcome included some that regulate responses to B-cell-receptor signaling, critical serine/threonine phosphorylation pathways, and apoptosis. Our data indicate that supervised learning classification techniques can predict outcome in DLBCL and identify rational targets for intervention."

Shipp microarray consists of 58 diffuse large B-cell lymphomas (DLBCL) patients (class1) and 19 follicular lymphomas (FL) patients (class2). They analyzed 77 cases by various statistical methods using 7,129 genes. Those methods are almost the same as Golub et al. (1999), and they announced a supplemental document with 67 pages. For example, they analyzed "marker" genes with the highest correlation with the target class-by-class separation statistics (signal-to-noise ratio), weighted votes, nearest neighbor cluster, and SVM (Vapnik 1995). These methods are used in Golub et al. except for SVM. Mainly, discriminant analysis is the most appropriate method to identify oncogenes from the microarray. However, because the statistical discriminant analysis was not useful at all, medical researchers had no choice but to develop many methods for cancer gene analysis. On the other hand, our theory shows that six microarrays are LSD (Fact3) and Method2 can decompose the microarrays into many SMs (Fact4). Then, by analyzing SM, cancer gene diagnosis and malignancy indexes are proposed. If Shipp and others validate our research results, cancer gene diagnosis will be more improved. They used leave-one-out (LOO), which was developed in the age of poor computing environment to verify their outcome (Lachenbruch and Mickey 1968). On the other hand, Shinmura (2010) proposed 100-fold cross-validation for the small sample (Method1). Although they used the hard-margin SVM (H-SVM), they could not find their microarray was LSD, and the number of misclassifications (NM) was 0. It is extraordinary why there are no researches that the two classes are completely separable in the high-dimensional microarray. If physicians manage two classes severe, probably the other microarrays are considered to be completely separable also. This fact is the most important in the cancer gene analysis. In our research, genes subspace with MNM = 0 defines signal and genes subspace with MNM >= 1 defines noise at first. On the other hand, they evaluate their results by various methods such as a 2 * 2 contingency table. For example, they say, in weighted voting, six DLBCL patients with moderate malignancy are incorrectly identified as low-grade FL patients. Also, they are examining the survival rate of Kaplan–Meier. From these facts, historically, because the discriminant functions based on the variance–covariance matrix were useless for cancer gene analysis, they develop several methods. They used SVM and nearest neighbor cluster, but there was no discrimination using 7,129 genes.

6.1 Introduction

Chapter 1 outlined the New Theory of Discriminant Analysis After R. Fisher (Theory) and explained a success of cancer gene analysis (Shinmura 2016a). Chapter 2 outlined the cancer gene diagnosis using all SMs of six microarrays found by the RIP in 2016. Chapter 3 outlined that RIP and Revised LP-OLDF discriminate Alon's microarray by changing the iteration number of LINGO Program 3. Chapter 4 explains about two SMs obtained by the RIP and Revised LP-OLDF using Alon's microarray in 2018. We challenge three new research themes as follows: (1) We evaluate two signal subspaces by the RIP and Revised LP-OLDF and find Revised LP-OLDF cannot find all SMs from Alon's microarray correctly. (2) After RatioSV and NM evaluate the 62 SMs found by RIP, we compute the 1,891 correlations of 62 RipDSs and find that all correlations are positive correlations. (3) Because standard statistical methods cannot find the linear separable facts from all SMs, we reconsider that RipDSs are signals themselves or the signal data made by RipDSs and HsvmDSs are signal instead of SMs themselves. (4) We try to explain the reason why statistical discriminant functions could not discriminate six microarrays and all SMs (Problem 6). Because the fluctuation of RipDS is small, we cannot observe the linear separable fact on the scatter plot made by the Prin1 and Prin2. In Chap. 6, we confirm the above research themes proposed in Chap. 4. LINGO (Schrage 2006) decomposes Shipp microarray into 237 SMs and opens a new frontier of cancer gene analysis. JMP (Sall et al. 2004) analyzes all SMs and offers cancer gene diagnosis. Shinmura (2016, 2017, 2018a, b) relate to this Chapter.

6.2 Validation of SM Found by the RIP and Revised LP-OLDF

In Chap. 6, we increase the number of iterations from 1 to 11 and choose SMs by the RIP and Revised LP-OLDF. However, we find the defect of Revised LP-OLDF that cannot find all SMs from the microarray in Chap. 4. Thus, we introduce the results of SMs found by the RIP in Chaps. 5 and 6.

6.2.1 Verification of the Number of Iterations of Revised LP-OLDF and RIP

Table 6.1 shows the results by changing the iteration number (IT) from 1 to 11 of LINGO Program 3 (Schrage 2006). In the Revised LP-OLDF (abbreviated as LP in the table), "IT = 3" chooses 89 SMs (2,784 genes), and an average is 31.28 genes. The selection of "IT = 5 or more" becomes a steady state that includes 96 SM (2,789 genes) with 29.05 genes on average. Thus, we decide that Revised LP-OLDF chooses 96 SM. On the other hand, RIP chooses 237 SMs (7,109 genes and 30 average) with IT = 10. Moreover, we evaluate the microarray, signal subspace (the union of SMs), and noise subspace with six MP-LDFs as same as Golub microarray. In the Revised LP-OLD, we consider the union of 96 SMs as signal subspace and the remaining 4,340 genes as noise subspace. On the other hand, in RIP, the union of 237 SMs is considered as signal subspace, and the remaining 20 genes are considered noise subspace. In this study, we compare Shipp microarray (omitted from the table), signal subspace, and noise subspace with RatioSV, MNM, or NM.

Table 6.1 Result of the number of iterations by LINGO Program 3

	LP				RIP			
IT	CPU	SM	Gene	Gene/SM	CPU	SM	Gene	Gene/SM
3	3:55	89	2784	31.28	4:56	223	7123	31.94
4	5:09	95	2816	29.64				
5	**6:12**	**96**	**2789**	**29.05**	21:05	243	7118	29.29
8	8:59	96	2789	29.05	25:21	238	7089	29.79
9	9:54	96	2789	29.05				
10	11:26	96	2789	29.05	**29:33**	**237**	**7109**	**30.00**
11					33:35	237	7109	30.00

6.2.2 *Analysis of Signal and Noise Subspaces Obtained by Revised LP-OLDF*

Until now, because the Revised LP-OLDF divides microarray into a signal subspace of fewer SMs and genes, we consider analyzing SMs found by the Revised LP-OLDF instead of RIP. Moreover, calculation time is shorter than RIP. The left six columns of Table 6.2 show the signal subspace, and the right six columns show the noise subspace. Row "Ratio" indicates six RatioSVs of MP-LDFs. The DSs of 58 cases (class1) and 19 cases (class2) are in the fourth line or less. Row "Outlier" shows the number of cases not on SVs of two classes. For example, "35/12 of RIP" indicates 35 DLBCL patients are higher than 1, and 23 patients are on SV = 1. On the other hand, 12 FL patients are less than −1, and seven patients are on SV = −1. Because the RatioSV of RIP is 2.83%, it is the minimum value among the six MP-based LDFs. This result is very different from the result of Golub in Table 5.2, the RatioSV of which is 100%. In a high-dimensional space, unless we study these results compared to the visual representation of two classes subjects, it is difficult for us to understand this situation clearly.

Table 6.2 Evaluation of the signal subspace and noise subspace (Revised LP-OLDF)

	Signal (2,789 genes)						Noise (4,340 genes)					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
Ratio	2.83	14.66	5.71	58.33	58.11	58.11	3.53	14.61	3.69	47.65	46.28	47.88
Outlier	**35/12**	40/6	16/3	34/5	34/5	31/5	22/8	18	2	36/3	38/4	46/6
1	36.1	4.5	8.78	1.41	1.41	1.41	8.34	2.35	1	1.03	1.11	1.04
2	15.21	1	1	1.05	1.12	1.11	1	1	1	2.21	2.29	2.2
3	6.42	1	1	1	1	1	1	1	24.34	1	1.01	1
4	1	4.62	1	1.78	1.8	1.8	23.18	1	1	2.43	2.51	2.41
5	28.52	2.3	1	1.41	1.42	1.42	20.73	1	15.3	1.37	1.46	1.36
6	1	7.88	1	1.41	1.41	1.41	1	1.94	1	1.14	1.22	1.12
7	4.55	1	1	1	1	1	1	1	28.69	1	1.01	1
8	16.41	1.85	1	1.28	1.27	1.28	1	3.16	18.83	1.45	1.52	1.45
9	1	1	1	1	1	1	1	1	1	1	1	1
10	30.16	3.73	7.85	1.47	1.47	1.47	15.6	3.55	27.87	1.32	1.36	1.32
11	1	3.46	1	1	1	1	1	1.27	1	1	1.02	1
12	1	1	1	1.01	1.09	1.08	1	1	1	2.45	2.55	2.44
13	17.52	3.3	1	1	1	1	16.44	1	1	1	1	1
14	19.49	1	1	1	1	1	19.91	1	3.33	1	1.03	1
15	1	1	1	1.1	1.12	1.12	1	4.1	11.67	1	1.03	1
16	1	9	1	1.64	1.64	1.65	1	3.93	1	1.74	1.8	1.73
17	21.2	1	1	1	1	1	1	1	1	1	1.01	1
18	3.12	7.91	1	1.08	1.1	1.1	7.44	1	9.38	1.67	1.72	1.66
19	1	1.58	1	1	1	1	19.54	1	1	1	1.01	1

(continued)

Table 6.2 (continued)

	Signal (2,789 genes)						Noise (4,340 genes)					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
20	32.81	1	1	1	1	1	1	1	1	1	1.01	1
21	25.14	4.04	1	1.48	1.48	1.48	20.91	7.73	1	1.36	1.42	1.35
22	1	1	1	1	1	1	6.85	1	1	1	1	1
23	1	5.05	21.3	1.58	1.6	1.6	1	1	12.77	1.57	1.63	1.56
24	33.68	1.26	9.54	1.65	1.65	1.65	1	2.33	1	1.02	1.09	1.03
25	18.77	4.09	1	1.42	1.43	1.43	17.9	2.92	1	1.62	1.69	1.62
26	21.92	3.72	1	2.01	2.02	2.02	4.1	1	8.1	1.34	1.4	1.34
27	1	1.58	1	1	1	1	10.05	1	35.56	1.84	1.92	1.83
28	8.76	5.58	1	1	1	1	18.12	1	1	1.05	1.12	1.04
29	1	1	1	1	1	1	1	1	1	1	1	1
30	1	4.3	1	1	1	1	27.4	3.94	1	1.24	1.34	1.22
31	27.3	6.89	1	1.66	1.66	1.67	1	1	1	1.82	1.87	1.8
32	1	3.22	1	2.03	2.05	2.05	1	6.14	29.7	2.4	2.5	2.38
33	1	7.64	1	1.01	1.04	1.03	6.15	1	5.42	1	1.01	1
34	10.58	5.51	1	1.3	1.32	1.32	1	1	1	1.26	1.33	1.24
35	41.39	7.64	1	1.23	1.23	1.23	11.63	5.39	1	1.37	1.41	1.37
36	26.44	1.05	1	1	1	1	11.16	3.49	1	1	1.01	1
37	27.86	2.41	1	1.07	1.08	1.08	35.09	5.65	1	1.65	1.71	1.64
38	31.01	1	4.16	1	1	1	7.12	1	1	1	1.06	1.01
39	1	5.18	10.85	1.83	1.86	1.86	27.74	5.32	1	2.37	2.46	2.35
40	31.45	6.37	1	1.45	1.45	1.45	1	1	25.21	1.56	1.59	1.56
41	1	5.96	4.86	1	1	1	10.99	1	1	1	1.02	1

(continued)

Table 6.2 (continued)

	Signal (2,789 genes)						Noise (4,340 genes)					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
42	19.67	1	2.26	1	1	1	14.33	2.39	22.22	1.18	1.24	1.17
43	24.61	1	17.95	1.88	1.88	1.88	30.28	4.83	1	1.69	1.77	1.68
44	17.68	5.9	1	1.71	1.72	1.72	15.2	1	1	2.15	2.2	2.14
45	17	5.42	10.94	1.94	1.95	1.95	1	1	1	1.88	1.94	1.87
46	1	4.18	29.81	1	1	1	1	1	1	1	1.01	1
47	1	5.34	1	1.08	1.11	1.11	1	1.82	1	1.37	1.45	1.35
48	1	4.83	12.6	1	1	1	29.14	1	1	1	1.09	1.01
49	1	2.89	1	1.54	1.54	1.54	1	1	1	1.9	1.96	1.89
50	19.7	1	1	1	1	1	1	1	1	1	1.03	1
51	1	2.99	10.08	1	1	1	21.88	1	1	1	1	1
52	12.33	1.11	2.05	1	1	1	1	1	1	1	1	1
53	18.32	3.46	1	1	1	1	1	1	1	1	1.01	1
54	20.39	5.13	5.44	1.71	1.72	1.72	12.16	1.87	17.04	1.55	1.6	1.54
55	6.66	4.29	11.56	1.71	1.74	1.74	16.9	6.06	1	2.23	2.32	2.21
56	1	1	1	1.81	1.84	1.84	8.78	1	10.52	2.37	2.43	2.35
57	7.94	1	1	1.64	1.65	1.65	1	1	32.85	1.47	1.53	1.45
58	**14.5**	**1**	**1**	**1.26**	**1.27**	**1.27**	**12.96**	**1.16**	**1**	**1.8**	**1.86**	**1.8**
59	**−1**	**−1**	**−1**	**−1**	**−1**	**−1**	**−9.74**	**−1**	**−1**	**−1**	**−1**	**−1**
60	−1	−1	−1	−1	−1	−1	−1	−1	−8.32	−1	−1.01	−1
61	−7.99	−1.06	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
62	−10.45	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
63	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1.01	−1

(continued)

Table 6.2 (continued)

	Signal (2,789 genes)						Noise (4,340 genes)					
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	HSVM	SVM4	SVM1
64	−18.47	−4.65	−1	−1	−1	−1	−4.77	−1	−1	−1	−1	−1
65	−14.01	−1	−1	−1	−1	−1	−1	−2.32	−1	−1	−1.01	−1
66	−21.59	−1	−1	−1	−1	−1	−3.29	−2.78	−1	−1	−1	−1
67	−9.43	−1.6	−1	−1	−1	−1	−1	−1	−1	−1	−1.01	−1
68	−29.21	−1	−1	−1	−1	−1	−1	−1.63	−1	−1	−1.01	−1
69	−12.79	−3.28	−1	−1.4	−1.4	−1.4	−8.19	−5.95	−1	−1.75	−1.77	−1.74
70	−1	−2.69	−1	−1.16	−1.17	−1.17	−21.58	−1	−1	−1	−1.04	−1.01
71	−1	−1	−1	−1.3	−1.31	−1.31	−1	−4.52	−1	−1.35	−1.38	−1.34
72	−1	−1	−5.21	−1	−1	−1	−1	−1.92	−1	−1	−1.01	−1
73	−1	−1	−1	−1.04	−1.06	−1.05	−1	−1.72	−18.66	−1	−1.01	−1
74	−13.57	−1	−1	−1	−1	−1	−1	−1.89	−1	−1	−1.02	−1
75	−16.55	−1	−5.18	−1	−1	−1	−1	−1	−1	−1	−1.01	−1
76	−12.82	−4.33	−1	−1.38	−1.39	−1.39	−1	−1	−1	−1.17	−1.2	−1.16
77	−10.55	−1	−4.2	−1	−1	−1	−1	−1	−1	−1	−1	−1

Table 6.3 shows the discriminant results by six MP-LDFs. If each LDF correctly discriminates patients, “>0” column shows it. “0” column means the number of patients on the discriminant hyperplane, and “<0” column means the number of misclassified patients. As a result, all NMs are 0 in 6 MP-LDFs in the noise space, so H-SVM in the eleventh column of Table 6.2 discriminates all patients correctly. This fact means the noise subspace found by Revised LP-OLDF is LSD. Thus, we do not analyze SMs found by Revised LP-OLDF in this chapter.

Table 6.3 The discriminant results by six MP-LDFs of 96 SMs obtained by Revised LP-OLDF

	Shipp			Signal			Noise		
	<0	0	>0	<0	0	>0	<0	0	>0
RIP	0	0	77	0	0	77	0	0	77
LP	0	0	77	0	0	77	0	0	77
IPLP	0	0	77	0	0	77	0	0	77
HSVM	0	0	77	0	0	77	0	0	77
SVM4	0	0	77	0	0	77	0	0	77
SVM1	0	0	77	0	0	77	0	0	77

6.2.3 Analysis of 237 SMs and Noise Spaces Obtained by RIP

Table 6.4 shows the discriminant results by six MP-LDFs. Unlike Table 6.3, H-SVM is in error in the noise subspace, and the five NMs are 6–19 for the rest. Because RIP finds all SMs correctly, we analyze only 237 SMs obtained by the RIP.

Table 6.4 The discriminant results by six MP-LDFs of 237 SMs obtained by RIP in 2018

	Shipp			Signal			Noise		
	<0	0	>0	<0	0	>0	<0	0	>0
RIP	0	0	77	0	0	77	6	0	71
LP	0	0	77	0	0	77	12	0	65
IPLP	0	0	77	0	0	77	6	0	71
HSVM	0	0	77	0	0	77	–	–	–
SVM4	0	0	77	0	0	77	12	0	65
SVM1	0	0	77	0	0	77	19	0	58

Table 6.5 shows the evaluation of the signal and noise subspaces obtained by RIP. Because the H-SVM cannot discriminate the noise subspace, we omit it from the table. Row "Outlier/NM" means the figures of noise subspace show NM instead of outliers. Because the difference between microarray and the signal is only 20 genes, both results are almost the same. The RatioSV of H-SVM is the maximum value of 48.41. Because the "RatioSV and Outlier" of RIP are 5.66% and "26/11," this result is very different from the result of Golub in Table 5.2. In the Golub microarray, RIP finds a unique data structure that all cases of two classes are on the two parallel hyperplanes.

Table 6.5 The evaluation of total gene space, the signal subspace, and noise subspace obtained by RIP

	Signal						Noise				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
Ratio	5.66	15.93	16.39	**48.41**	46.52	48.40	0.76	54.54	0.01	54.54	202.51
Outlier/NM	**26/11**	35/5	7/2	**37/4**	57/14	37/5	**6**	**8**	**6**	**12**	**19**
1	9.03	3.66	1	1.21	1.27	1.21	30.95	1.06	3550.60	1.06	0.93
2	2.29	1	1	1.87	2.00	1.87	41.07	1.37	6007.51	1.37	1
3	1	5.56	1	1	1.01	1	1	1	1	1	1.03
4	1	1	1	2.35	2.47	2.35	105.05	1	10782.11	1	1.23
5	8.48	3.83	1	1.43	1.55	1.43	120.88	1.41	9783.49	1.41	1.12
6	13.10	1	1	1.47	1.55	1.47	33.62	1	3817.79	1	0.89
7	1	1	1	1	1.01	1	1	1.78	1	1.78	1.08
8	16.80	4.68	1	1.45	1.52	1.45	30.85	1.97	2966.27	1.97	1.10
9	1	1	1	1	1.01	1	38.79	1.96	3023.69	1.96	1.24
10	3.62	6.96	1	1.43	1.48	1.43	22.91	1.36	1	1.36	1.22
11	1	1	1	1	1.04	1	1	1	588.94	1	0.99
12	1	1	1	2.03	2.18	2.03	105.28	2.55	13801.01	2.55	1
13	5.05	4.95	1	1	1.01	1	1	1	1	1	0.99
14	5.99	1	1	1	1.01	1	11.98	0.42	1532.66	0.42	0.87
15	1	1	1	1.11	1.21	1.11	42.88	1	3075.82	1	1.11
16	9.60	6.98	1	1.71	1.78	1.71	1	1.40	1	1.40	1
17	9.53	1	1	1	1.03	1	107.75	0.84	9205.56	0.84	0.99
18	8.37	2.88	1	1.54	1.63	1.54	79.47	1	5323.91	1	1.06
19	1	1	5.43	1	1.01	1	1	1	1	1	1

(continued)

Table 6.5 (continued)

	Signal						Noise				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
20	1	1	1	1	1.02	1	13.43	1.17	2331.78	1.17	1
21	1	2.57	1	1.52	1.59	1.52	91.29	2.67	146.99	2.67	1.21
22	1	1	1	1	1	1	1	0.05	1	0.05	0.76
23	1	2.84	1	1.59	1.69	1.59	1	0.81	1	0.81	1.01
24	1	1	4.15	1.35	1.42	1.36	36.30	1	4492.60	1	0.99
25	13.71	4.36	1	1.60	1.69	1.60	9.29	0.31	550.67	0.31	0.97
26	5.38	3.86	1	1.69	1.77	1.69	35.72	1.25	2543.28	1.25	1
27	6.32	7.98	1	1.53	1.62	1.53	57.09	1.13	3388.56	1.13	1.09
28	14.75	4.45	1	1.25	1.36	1.25	1	1.25	1	1.25	1.09
29	1	2.07	1	1	1.01	1	1	0.21	173.11	0.21	0.78
30	1.94	3.93	1	1.33	1.45	1.33	27.46	1	4861.42	1	1.05
31	2.68	6.06	1.40	1.73	1.80	1.73	128.95	2.41	11797.25	2.41	1.28
32	1	1.69	1	2.45	2.58	2.45	102.06	1	10763.86	1	1
33	1	1	1	1	1.03	1	1	0.44	1	0.44	0.81
34	10.17	3.22	1	1.40	1.51	1.40	1	1	879.03	1	1
35	1	1	4.89	1.32	1.36	1.32	25.39	2.05	3103.38	2.05	1.17
36	6.84	3.39	1	1	1.01	1	32.00	1	1	1	1
37	9.56	1	1	1.50	1.60	1.50	105.25	2.18	7597.50	2.18	1.24
38	1	1	1	1	1.04	1	4.33	0.80	1	0.80	0.94
39	1	5.38	1	2.16	2.27	2.16	82.37	2.30	10504.03	2.30	1.05
40	1	5.78	1	1.43	1.47	1.43	72.62	2.29	7095.68	2.29	1.14

(continued)

Table 6.5 (continued)

	Signal						Noise				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
41	3.28	5.69	1	1	1.07	1	91.38	1.61	9041.31	1.61	1
42	1	1	1	1.10	1.18	1.11	1	1.10	1	1.10	1
43	3.74	1	1	1.70	1.78	1.70	62.51	2.24	4326.40	2.24	1.18
44	20.86	3.18	4.50	2.14	2.24	2.14	65.86	1	2755.04	1	1.18
45	1	1	1	1.79	1.86	1.79	54.94	2.44	6820.85	2.44	1.06
46	1	5.55	1	1	1.02	1	1	1	1	1	1
47	19.77	3.39	1	1.38	1.50	1.39	39.16	0.76	4196.34	0.76	1.02
48	1	1	1	1	1.03	1	145.07	2.02	15433.07	2.02	1.02
49	1	1	1	1.91	1.99	1.91	14.66	1.11	845.04	1.11	1.06
50	4.47	1	1	1	1.06	1	3.84	0.92	1064.51	0.92	0.98
51	13.35	1.53	1	1	1.01	1	18.56	0.33	223.07	0.33	0.89
52	1	1	1	1	1	1	9.07	1.49	855.94	1.49	1.06
53	1	6.38	3.33	1	1.06	1	28.02	1.42	2236.89	1.42	1.10
54	1	1.65	1	1.61	1.70	1.61	87.59	1.50	6995.42	1.50	1.11
55	1	3.75	1	2.15	2.29	2.15	105.70	1.98	10663.02	1.98	1.06
56	1	6.59	1	2.14	2.25	2.14	71.57	1	8598.32	1	1
57	1	5.77	1	1.64	1.73	1.64	1.19	1	1	1	1
58	**1**	**1**	**4.49**	**1.67**	**1.73**	**1.67**	**151.53**	**1.37**	**12400.41**	**1.37**	**1.20**
59	**2.83**	**1**	**1**	**1**	**1**	**1**	**111.84**	**1**	**8763.78**	**1**	**−0.65**
60	1	1	1	1	1	1	**−48.25**	**−0.39**	**−4526.42**	**−0.39**	**−0.92**
61	3.40	2.93	1	1	1	1	2.25	**−0.67**	1	**−0.67**	**−0.83**

(continued)

Table 6.5 (continued)

	Signal						Noise				
	RIP	LP	IPLP	HSVM	SVM4	SVM1	RIP	LP	IPLP	SVM4	SVM1
62	10.28	1	1	1	1.01	1	1	**−1.32**	1	**−1.32**	**−1.03**
63	1	1	1	1	1.01	1	1	**−1.16**	309.26	**−1.16**	**−1.05**
64	8.49	2.07	1	1	1	1	33.17	1	4718.72	1	**−0.72**
65	8.40	1	1	1	1.01	1	7.19	0.16	1075.18	0.16	**−0.79**
66	1	1	1	1	1.02	1	1	1	1	1	**−0.29**
67	4.71	1	1	1	1.01	1	**−1.24**	**−1.75**	**−8240.90**	**−1.75**	**−1.10**
68	1	1	1	1	1.02	1	**−99.49**	**−1.05**	**−7717.57**	**−1.05**	**−1.02**
69	7.96	4.57	1	1.68	1.72	1.68	1	1	1	1	**−0.73**
70	14.45	1	1	1.12	1.16	1.12	**−68.32**	**−1.94**	**−7269.91**	**−1.94**	**−1.15**
71	1	1	1	1.29	1.36	1.29	38.20	1	2766.11	1	**−0.50**
72	1	1	1	1	1.03	1	1.10	**−1.18**	923.97	**−1.18**	**−1.14**
73	1	1	6.77	1	1.07	1.01	**−98.68**	**−1.19**	**−9999.00**	**−1.19**	**−0.94**
74	8.02	1	1	1	1.02	1	1	**−1.70**	1664.81	**−1.70**	**−1.23**
75	1.43	1	1	1	1.02	1	1	**−1.50**	93.69	**−1.50**	**−1.08**
76	1	1.43	3.66	1.21	1.27	1.21	1	0.80	1	0.80	**−0.76**
77	2.56	1.08	1	1	1	1	**−80.04**	**−1.74**	**−9022.54**	**−1.74**	**−1.08**

6.3 Analysis of 237 SMs of Shipp et al. Microarray (2018)

In 2015, RIP of LINGO Program3 decomposed Shipp microarray into 214 SMs (3,040 genes) and noise gene subspace (4,089 genes). However, when LINGO Program3 decomposes the microarray again in 2018, 237 SMs (7,109 genes) are found. We obtain more SMs and genes in 2018. A yearly update of LINGO and choosing different iteration number cause these differences. We consider 237 SMs are signals, and 20 gene subspace is noise at first. However, we cannot find linear separable facts of 237 SMs by the standard statistical methods (Problem 6). Moreover, statistical discriminant functions cannot find six microarrays are LSD (Fact3). We believe the two reasons are the same. Although the actual separation distance between the two classes is small compared to the variation of the microarray, many RatioSVs seem to be large because the MP-based LDF fixes the SV width to 2. It is noted that this value is the ratio of signal space with small variations. RIP, Revised LP-OLDF, and H-SVM can find all SMs are LSD. Because we find LSD is the true signal, we claim that signal data created from the DSs made by RIP, Revised LP-OLDF, and H-SVM is a true signal.

6.3.1 Validation of 237 SMs by Six MP-Based LDFs and Discriminant Functions

Table 6.6 shows the 237 SMs from SM = 1 to SM = 237. Although Revised LP-OLDF finds 96 SMs, we omit those results because of the defect of Revised LP-OLDF. Program3 determines the order of SMs from SM = 1 to SM = 237. The "gene" column is the number of genes of each SM. The range of genes included in the 237 SMs is [15, 62]. The average is 30 genes. Row "SUM" indicates 237 SMs contain 7,109 genes. Columns from RIP to H-SVM show four RatioSVs of 237 SMs by RIP, Revised LP-OLDF, Revised IPLP-OLDF, and H-SVM. Four ranges of RatioSVs are [0.52, 29.8], [0.72, 25.1], [0.72, 25.1], and [1, 33.34], respectively. Four averages of RatioSVs are 13.66%, 13.05%, 13.05%, and 16.68%, respectively. Row "SUM/Max" indicates 237 SMs includes 7,109 genes. RIP, LP, IPLP, and HSVM have 59, 30, 16, and 132 SMs having the maximum RatioSVs. "183, 146 and 10" indicate the number of SMs those are not LSD by SVM1, LDF2, and QDF. To summarize these results, the range, average, and maximum number of H-SVM are slightly better than RIP because the maximization SV of H-SVM works well. Two columns "Max and Min" are the maximum and minimum RatioSV values of four LDFs. Because all NMs of logistic regression and SVM4 are zero, we omit these columns from the table. Three columns "SVM1, LDF2, and QDF" show the NM. The ranges are [0, 18], [0, 9], and [0, 1], respectively. The averages are 2.12, 1.53, and 2.54, respectively. Moreover, SVM1, LDF2, and QDF cannot discriminate 182 SMs, 156 SMs, and 12 SMs correctly.

Table 6.6 Summary of RatioSVs of four MP-based LDFs and NMs of other discriminant functions

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
1	28	22.19	**23.52**	10.22	20.9	23.52	10.22	0	0	0
2	28	**31.97**	18.77	7.5	26.66	31.97	7.5	0	0	0
3	27	19.68	18.4	18.54	**23.74**	23.74	18.4	0	0	0
4	27	**27.85**	20.69	17.42	20.72	27.85	17.42	0	0	0
5	23	18.59	**22.26**	15.46	17.04	22.26	15.46	1	1	0
6	31	21.88	17.39	19.42	**23.47**	23.47	17.39	0	0	0
7	26	**23.41**	15.85	21.02	21.8	23.41	15.85	0	0	0
8	21	16.92	15.06	14.28	**17.74**	17.74	14.28	2	1	0
9	15	**9.23**	8.26	7.84	9.4	9.4	7.84	2	2	0
10	25	15.75	18.66	15.55	**21.17**	21.17	15.55	0	0	0
11	26	**21.85**	12.15	21.95	18.73	21.95	12.15	0	0	0
12	26	19.28	**21.65**	17.26	16.89	21.65	16.89	1	0	0
13	29	17.72	23.37	**27.48**	17.75	27.48	17.72	1	0	0
14	27	14.29	10.15	17.91	**21.06**	21.06	10.15	2	0	0
15	15	16	14.4	15.78	**16.02**	16.02	14.4	1	1	0
16	25	18.45	**18.94**	18.86	18.25	18.94	18.25	0	0	0
17	34	24.3	14.2	25.51	**29.94**	29.94	14.2	0	0	0
18	27	**22.84**	22.07	16.25	21.2	22.84	16.25	1	1	0
19	23	12.69	8.14	11.79	**14.01**	14.01	8.14	1	1	0
20	24	**21.22**	20.49	9.85	19.56	21.22	9.85	0	0	0
21	31	25.25	20.28	15.65	**29.86**	29.86	15.65	0	0	1
22	31	**24.9**	13.72	21.22	23.13	24.9	13.72	1	0	0
23	21	9.55	10.42	9.91	**11.76**	11.76	9.55	1	1	0
24	20	14.48	10.43	10.13	**15.82**	15.82	10.13	0	0	0
25	25	**18.08**	13.6	14.94	18.06	18.08	13.6	1	0	0
26	30	16.21	11.35	13.66	**19.12**	19.12	11.35	1	1	0
27	24	**29.22**	29.22	11.29	27.37	29.22	11.29	1	0	0
28	23	18.82	14.29	12.84	**19.33**	19.33	12.84	0	0	0
29	21	9.54	13.32	**17.54**	15.83	17.54	9.54	0	1	0
30	30	18.45	25.01	18.27	**25.32**	25.32	18.27	1	0	0
31	29	**27.77**	14.58	17.22	26.28	27.77	14.58	0	0	0
32	23	**21.86**	18.69	16.96	20.64	21.86	16.96	1	0	0
33	21	24.31	19.42	12.76	**25.63**	25.63	12.76	0	0	0
34	37	**25.27**	14.59	18.03	22.95	25.27	14.59	0	0	0
35	26	15.01	18.99	15.84	**19.4**	19.4	15.01	0	1	0

(continued)

Table 6.6 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
36	21	9.28	9.5	8.53	**11.86**	11.86	8.53	1	1	0
37	25	**23.08**	15.85	19.34	18.56	23.08	15.85	1	1	0
38	20	19.75	16.66	17.81	**21.61**	21.61	16.66	0	0	0
39	17	**18.78**	16.12	16.98	15.2	18.78	15.2	0	0	0
40	23	17.29	17.99	19.58	**23.4**	23.4	17.29	0	0	0
41	30	17.29	17.99	19.58	**23.4**	23.4	17.29	1	0	0
42	29	**23.59**	15.6	12.65	21.92	23.59	12.65	0	0	0
43	28	**20.27**	13.34	19.97	16.12	20.27	13.34	2	1	0
44	25	15.05	17.96	11.73	**19.84**	19.84	11.73	0	1	0
45	25	22.99	**23.99**	15.82	22.96	23.99	15.82	0	0	0
46	20	19.81	18.49	17.88	**21.23**	21.23	17.88	0	0	0
47	25	18.91	19.89	19.43	**22.02**	22.02	18.91	0	0	0
48	32	18.42	11.14	**22.26**	19.74	22.26	11.14	0	1	0
49	32	16.61	15.86	15.78	**19.07**	19.07	15.78	0	0	0
50	28	**25.74**	17.64	14.88	22.25	25.74	14.88	1	0	0
51	28	**22.66**	21.01	16.02	21.61	22.66	16.02	0	0	0
52	28	19.91	16.28	14.5	**21.21**	21.21	14.5	2	1	0
53	30	10.78	**18.71**	14.93	16.62	18.71	10.78	1	1	0
54	21	8.88	9.78	**12.33**	11.54	12.33	8.88	0	2	0
55	32	21.43	**22.93**	18.17	22.35	22.93	18.17	3	1	0
56	25	**22.22**	13.41	16.69	22.11	22.22	13.41	1	1	0
57	35	27.71	16.35	17.62	**30.43**	30.43	16.35	0	0	0
58	21	13.76	**15.71**	13.84	14.33	15.71	13.76	2	2	0
59	26	9.47	11.5	**11.97**	11.52	11.97	9.47	2	3	0
60	25	18.25	16.45	13.42	**18.54**	18.54	13.42	0	1	0
61	27	16.81	16.6	13.4	**20.72**	20.72	13.4	1	0	0
62	28	**24.08**	19.69	17.74	22.39	24.08	17.74	1	0	0
63	29	17.53	16.79	12.29	**17.77**	17.77	12.29	2	2	0
64	21	16.59	16.63	**16.97**	16.9	16.97	16.59	1	0	1
65	25	**16.81**	15.02	13.35	15.33	16.81	13.35	1	1	0
66	31	**16.93**	9.8	8.03	16.16	16.93	8.03	1	0	0
67	23	**17.34**	8.82	12.8	14.64	17.34	8.82	2	1	0
68	19	19.67	**20.71**	16.12	18.56	20.71	16.12	1	1	0
69	27	**25.87**	15.28	16.81	24.29	25.87	15.28	2	0	0
70	19	15.8	**18.99**	16.21	16.34	18.99	15.8	0	3	0
71	23	**23.61**	21.01	20.21	22.42	23.61	20.21	0	0	0
72	30	**29.77**	23.6	16.12	28.25	29.77	16.12	0	0	0

(continued)

Table 6.6 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
73	31	16.97	22.86	15.02	**24.32**	24.32	15.02	1	0	0
74	28	17.91	15.44	11.15	**19.21**	19.21	11.15	1	1	0
75	28	15.25	11.28	13.84	**18.47**	18.47	11.28	1	1	0
76	28	15.51	11.75	7.83	**17.88**	17.88	7.83	2	1	0
77	26	12.25	10.57	10.05	**12.9**	12.9	10.05	3	1	0
78	28	17.55	**19.61**	16.8	18.34	19.61	16.8	2	1	0
79	21	18.63	8.8	14.02	**18.67**	18.67	8.8	1	1	0
80	24	11.8	12.93	14.38	**14.89**	14.89	11.8	4	2	0
81	27	19.4	15.55	16.24	**20.01**	20.01	15.55	0	0	1
82	24	20.01	**22.38**	17.99	21.73	22.38	17.99	1	0	0
83	27	17.47	**18.66**	16.46	17.05	18.66	16.46	1	1	0
84	24	15.88	12.44	**18.33**	16.53	18.33	12.44	2	1	0
85	27	8.25	9.82	5.65	**11.43**	11.43	5.65	2	2	0
86	25	15.95	21.45	19.91	**24.06**	24.06	15.95	0	0	0
87	25	22.69	**23.98**	15.55	18.47	23.98	15.55	0	0	0
88	27	12.62	11.88	8.01	**15.69**	15.69	8.01	0	1	0
89	23	**15.54**	15.38	11.87	15.29	15.54	11.87	0	1	0
90	25	16.01	13.32	8.72	**17.6**	17.6	8.72	1	0	0
91	30	12.95	12.94	8.74	**15.96**	15.96	8.74	1	1	0
92	27	10	**12.81**	10.68	12.44	12.81	10	0	2	0
93	28	8.4	14.66	16.7	**18.43**	18.43	8.4	0	2	0
94	23	12.3	**14.05**	11.78	12.63	14.05	11.78	2	2	0
95	25	15.54	17.85	13.89	**18.65**	18.65	13.89	1	1	0
96	24	22.57	22.87	22.66	**24.37**	24.37	22.57	0	1	0
97	31	11.55	**12.73**	7.11	11.34	12.73	7.11	3	4	0
98	27	17.62	**17.66**	12.35	15.77	17.66	12.35	1	2	0
99	22	14.61	**16.28**	11.47	15.4	16.28	11.47	3	0	0
1	18	5.64	7.78	7.2	**8.75**	8.75	5.64	3	2	0
101	30	14.23	11.79	7.52	**15.33**	15.33	7.52	6	1	0
102	33	17.99	15.75	13.5	**19.48**	19.48	13.5	1	2	0
103	20	11.11	8.13	9.61	**12.47**	12.47	8.13	1	3	0
104	24	9.25	16.86	16.68	**21.33**	21.33	9.25	1	3	0
105	34	14.35	10.58	18.46	**19.62**	19.62	10.58	1	2	0
106	29	15.75	11.35	14.76	**18.15**	18.15	11.35	1	0	0
107	33	18.3	14.36	18.75	**22.91**	22.91	14.36	2	0	1
108	33	21.02	19.58	7.29	**21.41**	21.41	7.29	2	0	0
109	33	**22.68**	6.94	11.46	21.1	22.68	6.94	0	1	0

(continued)

Table 6.6 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
110	24	12.07	13.42	13.47	**16.92**	16.92	12.07	3	1	0
111	34	**26.72**	15.7	14.53	25.11	26.72	14.53	0	0	0
112	24	10.65	**11.52**	7.45	11.42	11.52	7.45	4	1	0
113	33	14.53	9.76	11.97	**14.79**	14.79	9.76	3	2	0
114	26	12.82	10.89	6.66	**16.11**	16.11	6.66	2	2	0
115	26	**17.97**	15	16.02	15.89	17.97	15	1	0	0
116	28	19.93	20.46	8.28	**21.58**	21.58	8.28	2	1	0
117	30	16.84	**18.37**	13.3	16.14	18.37	13.3	0	0	0
118	20	12.06	13.57	12.63	**17.91**	17.91	12.06	3	1	0
119	23	23.84	22.16	20.8	**25.72**	25.72	20.8	0	0	0
120	18	16.25	16.21	15.92	**22.35**	22.35	15.92	1	1	0
121	30	22.28	**25.79**	19.98	24.94	25.79	19.98	2	0	0
122	26	13.99	11.83	12.64	**15.5**	15.5	11.83	1	2	0
123	36	**13.46**	8.38	9.72	13.35	13.46	8.38	5	1	0
124	28	17.15	16.11	14.18	**17.56**	17.56	14.18	1	2	0
125	19	5.95	9.8	**10.51**	9.91	10.51	5.95	3	3	0
126	35	17.13	**20.16**	12.52	19.16	20.16	12.52	2	1	0
127	25	13.55	9.16	12.99	**13.88**	13.88	9.16	1	0	0
128	25	13.99	9.87	16.71	**20.58**	20.58	9.87	1	0	0
129	28	**19.53**	14.53	10.84	18.4	19.53	10.84	3	2	0
130	30	18.91	10.07	12.74	**21.94**	21.94	10.07	0	0	0
131	30	22.05	24	11.21	**24.3**	24.3	11.21	1	1	0
132	35	11.49	10.3	11.31	**16.24**	16.24	10.3	2	2	0
133	32	22.31	15.92	12.77	**25.02**	25.02	12.77	0	1	0
134	33	20.99	15.76	10.62	**23.88**	23.88	10.62	2	0	0
135	23	**16.59**	12.34	11.71	16.09	16.59	11.71	1	0	0
136	26	11.89	15.87	11.86	**19.55**	19.55	11.86	1	0	0
137	33	30.92	24.97	18.74	**31.12**	31.12	18.74	0	0	0
138	28	23.44	20.14	15.92	**26.91**	26.91	15.92	1	0	0
139	28	17.44	**18.02**	14.91	17.92	18.02	14.91	7	2	0
140	29	10.16	12.93	10.16	**21.56**	21.56	10.16	3	0	0
141	33	12.48	9.41	13.13	**17.42**	17.42	9.41	5	0	0
142	25	13.69	17.81	12.99	**21.28**	21.28	12.99	0	0	0
143	26	**16.99**	10.56	15.66	16.74	16.99	10.56	1	0	0
144	29	7.12	10.75	8.43	**12.49**	12.49	7.12	3	1	0
145	33	10.33	12.46	11.53	**14.86**	14.86	10.33	2	1	0
146	27	10.72	**13.03**	11.07	11.95	13.03	10.72	3	3	0

(continued)

Table 6.6 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
147	30	16.38	10.42	10.61	**16.74**	16.74	10.42	1	1	1
148	24	16.97	20.37	7.96	**23.71**	23.71	7.96	0	0	0
149	35	13.88	6.53	9.51	**20.77**	20.77	6.53	0	0	0
150	30	12.22	23	11.83	**23.22**	23.22	11.83	1	0	0
151	31	22.09	17.27	14.72	**23.65**	23.65	14.72	2	1	0
152	29	6.91	7.27	8.56	**10.86**	10.86	6.91	5	3	0
153	33	20.2	17.42	16.37	**20.65**	20.65	16.37	5	1	0
154	28	**16.16**	7.8	9.52	14.82	16.16	7.8	1	1	0
155	33	9.42	10.09	10.29	**16.85**	16.85	9.42	5	2	0
156	36	15.02	17.41	13.28	**18.43**	18.43	13.28	3	1	0
157	27	19.36	19.14	23.24	**25.18**	25.18	19.14	0	0	0
158	25	**20.24**	19.76	20.04	18.45	20.24	18.45	2	0	0
159	27	11.5	9.72	**12.22**	12.04	12.22	9.72	4	3	1
160	37	**17.33**	15.1	12.21	15.75	17.33	12.21	1	1	0
161	29	13.69	13.43	**14.39**	12.19	14.39	12.19	4	3	0
162	35	13.44	12.85	9.66	**14.95**	14.95	9.66	3	1	0
163	30	7.76	5.86	6.53	**9.14**	9.14	5.86	10	2	0
164	28	10.62	7.48	10.79	**12.58**	12.58	7.48	5	2	0
165	39	17.19	18.83	13.53	**21.06**	21.06	13.53	4	2	0
166	33	**28**	21.48	21.32	25.36	28	21.32	3	0	0
167	38	8.23	5.69	4.36	**9.34**	9.34	4.36	8	6	0
168	40	28.04	14.96	19.22	**28.98**	28.98	14.96	4	0	0
169	32	12.25	**15.55**	10.22	14.7	15.55	10.22	3	2	0
170	35	**16.51**	9.06	10.68	13.46	16.51	9.06	4	3	0
171	46	12.93	11.78	7.67	**21.64**	21.64	7.67	3	0	0
172	26	5.75	7.32	7.22	**8.06**	8.06	5.75	5	4	0
173	24	6.09	5.49	**6.86**	6.78	6.86	5.49	4	4	0
174	45	**22.95**	18.74	5.87	15.23	22.95	5.87	4	2	1
175	34	**23.07**	12.85	13.81	20.22	23.07	12.85	4	0	1
176	34	**16.34**	11.16	11.6	15.67	16.34	11.16	4	3	0
177	38	11.67	12.2	10.55	**18.24**	18.24	10.55	5	1	0
178	37	13.29	9.9	13.41	**13.6**	13.6	9.9	5	2	0
179	29	19.71	12.87	14.72	**20.2**	20.2	12.87	4	0	0
180	28	12.01	11.22	13.76	**15.45**	15.45	11.22	3	1	0
181	**33**	2.23	3.34	**4.75**	1.23	4.75	1.23	5	1	0
182	29	10.04	6.94	7.14	**11**	11	6.94	3	4	0
183	31	19.55	22.98	7.2	**23.65**	23.65	7.2	3	1	1

(continued)

Table 6.6 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
184	25	15.17	11.26	9.52	**23.05**	23.05	9.52	4	1	0
185	28	10.05	7.68	8.42	**10.29**	10.29	7.68	5	3	1
186	24	7.69	**9.42**	9.38	9.28	9.42	7.69	6	5	0
187	37	**12.04**	11.44	7.11	11.14	12.04	7.11	11	3	0
188	38	13.59	14.42	10.23	**16.56**	16.56	10.23	6	2	0
189	37	15.23	12.79	17.6	**19.82**	19.82	12.79	8	0	0
190	38	**17.2**	13.94	7.61	14.83	17.2	7.61	4	0	0
191	30	6.85	9.9	6.8	**10.72**	10.72	6.8	5	3	0
192	35	7.08	7.52	5.26	**11.16**	11.16	5.26	10	5	0
193	33	**22.28**	17.3	9.87	22.09	22.28	9.87	5	2	0
194	29	15.88	9.53	11.1	**15.94**	15.94	9.53	7	2	0
195	30	3.28	6.67	**7.97**	1.55	7.97	1.55	7	0	0
196	28	9.6	9.22	3.91	**9.85**	9.85	3.91	6	4	0
197	36	13.95	13.3	10.71	**14.06**	14.06	10.71	8	0	0
198	28	10.95	11.12	9.27	**12.13**	12.13	9.27	7	2	0
199	38	**11.75**	10.61	8.65	11.71	11.75	8.65	7	3	1
200	31	13.34	9.69	10.15	**15.5**	15.5	9.69	5	1	0
201	38	16.67	7.41	16.32	**20.36**	20.36	7.41	13	2	0
202	30	**9.17**	7.62	8.58	8.97	9.17	7.62	10	7	0
203	38	**12.72**	11.02	8.43	11.86	12.72	8.43	10	2	0
204	31	**14.16**	12.58	13.18	13.86	14.16	12.58	5	3	0
205	30	18.03	17.03	16.05	**19.91**	19.91	16.05	5	1	0
206	32	16.59	12.78	17.28	**23.04**	23.04	12.78	1	4	0
207	37	16.51	16.68	13.56	**19.62**	19.62	13.56	3	3	0
208	27	**17.6**	13.08	16.6	16.2	17.6	13.08	7	0	0
209	34	17.03	18.86	**22.59**	17.11	22.59	17.03	9	2	0
210	31	**17.6**	13.08	16.6	16.2	17.6	13.08	7	0	1
211	37	17.03	18.86	**22.59**	17.11	22.59	17.03	9	2	0
212	36	11.43	8.86	10.14	**14.2**	14.2	8.86	7	3	0
213	36	13.43	13.35	12.39	**14.02**	14.02	12.39	12	4	0
214	38	13.42	11.39	9.61	**18.35**	18.35	9.61	3	0	0
215	38	17.71	17.75	11.49	**18.49**	18.49	11.49	10	1	0
216	40	16.21	17.53	14.64	**26.76**	26.76	14.64	9	0	0
217	33	7.88	11.62	11.74	**13.1**	13.1	7.88	12	2	0
218	32	**16.08**	13.24	13.77	13.72	16.08	13.24	9	2	0
219	32	**18**	7.25	15.11	14.29	18	7.25	11	0	0
220	41	14.13	17.82	11.78	**18.9**	18.9	11.78	9	1	0

(continued)

Table 6.6 (continued)

SM	Gene	RIP	LP	IPLP	HSVM	Max	Min	SVM1	LDF2	QDF
221	34	3.28	6.67	**7.97**	1.55	7.97	1.55	10	2	0
222	37	**14.13**	7.05	7.83	12.42	14.13	7.05	14	5	0
223	41	8.23	**12.78**	4.9	10.49	12.78	4.9	12	6	0
224	47	15.99	**29.93**	15.82	24.72	29.93	15.82	12	1	0
225	47	9.21	15.09	11.52	**20.82**	20.82	9.21	11	0	0
226	43	7.21	6.45	7.27	**8.45**	8.45	6.45	11	4	0
227	41	**11.23**	8.89	7.96	10.28	11.23	7.96	16	6	0
228	43	**13.45**	10.44	5.37	10.3	13.45	5.37	13	6	0
229	42	**13**	10.31	10.47	11.36	13	10.31	12	5	0
230	50	7.6	10.51	8.11	**10.78**	10.78	7.6	11	6	0
231	42	8.11	9.12	5.75	**14.04**	14.04	5.75	17	3	0
232	46	11.97	10.09	11.24	**15.68**	15.68	10.09	14	3	0
233	52	**16.45**	5.79	15.37	15.65	16.45	5.79	18	5	0
234	56	5.45	5.99	4.12	**6.74**	6.74	4.12	15	7	0
235	54	7.34	9.33	6.28	**11.63**	11.63	6.28	18	6	0
236	62	5.26	5.02	4.83	**9.89**	9.89	4.83	16	8	0
237	56	6.08	**7.21**	6.14	6.55	7.21	6.08	15	9	0
Max	**62**	28.8	25.1	25.1	33.34	33.34	24.47	18	**9**	**1**
Min	**15**	0.52	0.72	0.72	1	1	0.52	0	0	0
Average	**30**	13.66	13.05	13.05	16.68	17.09	11.89	2.12	1.53	2.54
SUM/Max	**7109**	59	30	16	132			183	146	10

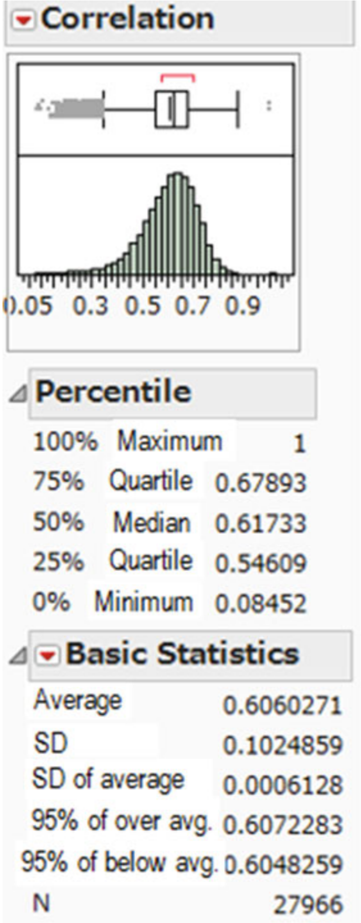
6.3.2 Correlation Coefficients of 237 RipDSs

Although Method2 and RIP find 237 SMs which are LSD, statistical methods, except for logistic regression, cannot find the linear separable facts. However, RIP, Revised LP-OLDF, Revised IPLP-OLDF, and H-SVM can separate two classes in each SM entirely. Thus, we consider RIP discriminant score (RipDS) as malignant indicator and signal. We make a signal data made by 237 RipDSs instead of 7,129 genes. The hierarchical cluster methods divide two classes into several definitive clusters which may be new subclasses of cancer. The Prin1 of PCA shows the linear separability of two classes and becomes comprehensible malignancy indicator to summarize 237 indicators. Several definitive clusters suggest us new subclasses of cancer, also. The transposed data gives us other useful information. However, we cannot categorize the 237 RipDSs and explain the role for cancer gene diagnosis (Problem 7). Thus, we try to analyze 237 RipDSs by correlation analysis.

Problem 7 From the standpoint of genetic diagnosis of cancer, it is necessary to divide SM and BGS into several categories and clarify their roles.

Figure 6.1 shows the histogram of the 27,966 correlation coefficients (correlations, abbreviate r) obtained with 237 RIP discriminant scores (RipDSs) by JMP (Sall et al. 2004). The correlations differ from 0.085 to 1, but all correlations are positive values. The 237 RipDSs are malignancy indicators which separate two groups completely. If they are about the same role, the correlation should be high. Small correlations mean the diversity of cancer. From Q1 and Q3, a beard with a length 1.5 times of the interquartile range 0.132 is drawn, many outliers below 0.346, and three outliers (R = 1) above 0.878. The average value is 0.606, and the median value is 0.617, which is a unimodal distribution. SMs with r = 1 are considered to be complementary to each other. Whether SMs with small correlations represent subclasses of different cancers is a future research topic. However, because medical verification is necessary for our claim, we will be willing to provide necessary information if requested by researchers who can access patients of Shipp et al. Because the median value is 0.617, the half of correlations are over 0.617. We consider that this fact indicates a unique feature of microarrays. That is, the main feature of the microarray used for cancer gene analysis is that all correlations are positive values and half of the total correlates are very high. Due to this feature, we consider that the signal is not a gene contained in 237 SMs but DS which separates the two classes. That is, in the statistics, the discriminant axis has not been considered important so far, but it is an optimal means to represent LSD.

Fig. 6.1 The 27,966 correlations by 237 SMs



Percentile

100%	Maximum	1
75%	Quartile	0.67893
50%	Median	0.61733
25%	Quartile	0.54609
0%	Minimum	0.08452

Basic Statistics

Average	0.6060271
SD	0.1024859
SD of average	0.0006128
95% of over avg.	0.6072283
95% of below avg.	0.6048259
N	27966

Table 6.7 is “27,966 correlations of RIPi and RIPj (i < j).” The RIPi denotes the individual RipDS. Those are classified by sorting from RIP1 to RIP236 in ascending order with the smaller suffix of RIPi as the sort key. The “n” column indicates the number of RIPj paired with each RIPi. It decreases from 236 to 1. The 100%, 50%, and 0% are the maximum, median, and minimum values. Row “RIP = 1” means the 236 correlations of RIP1 and RIPj (j = 2, …, 237). Three percentiles are r = 0.811, 0.669, and 0.371. Last row “RIP = 236” means the one correlation of RIP236 and RIP237 is r = 0.088. The last four rows are the maximum, the minimum, the average, and the median of 236 three percentiles. The range of the 236 maximum values is [0.088, 1], the median values are [0.088, 0.723], and the minimum values is [0.085, 0.391]. In the last row “50%”, three columns such as “100, 50, and 0%” are 0.772, 0.599, and 0.257. The 118 correlations of maximum, median, and minimum

values are over r = 0.772, 0.599, and 0.257. That is, half of the maximum, median, and minimum values are correlated 0.772, 0.599, and 0.257 over. Although Table 6.7 shows only the correlation between the two RipDSs, they expect to be a foothold for the solution of Problem7 in the future. Three correlations of (RIP195 and 196), (RIP = 208 and 210) and (RIP = 209 and 211) are one, and three pairs are sure to be mutually compatible.

Table 6.7 Maximum, median, and minimum of 27,966 correlation coefficients

RIP	100%	50%	0%	n
1	0.811	0.669	0.371	236
2	0.873	0.723	0.272	235
3	0.776	0.641	0.283	234
4	0.857	0.702	0.323	233
5	0.822	0.670	0.197	232
6	0.818	0.660	0.260	231
7	0.810	0.667	0.220	230
8	0.821	0.648	0.356	229
9	0.791	0.618	0.249	228
10	0.825	0.652	0.297	227
11	0.789	0.665	0.334	226
12	0.784	0.575	0.232	225
13	0.750	0.592	0.235	224
14	0.794	0.614	0.301	223
15	0.840	0.661	0.247	222
16	0.814	0.682	0.285	221
17	0.799	0.615	0.352	220
18	0.835	0.647	0.289	219
19	0.833	0.605	0.321	218
20	0.847	0.684	0.182	217
21	0.844	0.663	0.261	216
22	0.831	0.679	0.334	215
23	0.776	0.632	0.308	214
24	0.797	0.655	0.283	213
25	0.816	0.660	0.274	212
26	0.785	0.651	0.225	211
27	0.823	0.662	0.278	210
28	0.834	0.654	0.253	209
29	0.800	0.644	0.193	208

(continued)

Table 6.7 (continued)

RIP	100%	50%	0%	n
30	0.775	0.635	0.215	207
31	0.820	0.677	0.264	206
32	0.861	0.701	0.243	205
33	0.842	0.680	0.389	204
34	0.790	0.648	0.276	203
35	0.783	0.608	0.209	202
36	0.823	0.608	0.224	201
37	0.811	0.638	0.340	200
38	0.835	0.651	0.208	199
39	0.863	0.673	0.299	198
40	0.792	0.645	0.236	197
41	0.801	0.646	0.223	196
42	0.856	0.689	0.298	195
43	0.768	0.634	0.229	194
44	0.826	0.616	0.353	193
45	0.827	0.654	0.322	192
46	0.842	0.665	0.314	191
47	0.764	0.608	0.269	190
48	0.780	0.608	0.259	189
49	0.792	0.637	0.281	188
50	0.807	0.662	0.261	187
51	0.864	0.691	0.284	186
52	0.841	0.659	0.364	185
53	0.701	0.551	0.302	184
54	0.789	0.628	0.284	183
55	0.777	0.617	0.315	182
56	0.812	0.636	0.163	181
57	0.840	0.668	0.334	180
58	0.782	0.638	0.284	179
59	0.729	0.548	0.328	178
60	0.794	0.648	0.299	177
61	0.801	0.633	0.304	176
62	0.824	0.664	0.292	175
63	0.815	0.669	0.309	174
64	0.763	0.605	0.245	173
65	0.809	0.598	0.270	172
66	0.760	0.594	0.276	171
67	0.804	0.641	0.298	170

(continued)

Table 6.7 (continued)

RIP	100%	50%	0%	n
68	0.862	0.680	0.292	169
69	0.816	0.644	0.306	168
70	0.823	0.648	0.213	167
71	0.832	0.689	0.352	166
72	0.811	0.666	0.303	165
73	0.807	0.639	0.306	164
74	0.805	0.645	0.345	163
75	0.751	0.596	0.272	162
76	0.780	0.613	0.320	161
77	0.747	0.582	0.186	160
78	0.778	0.604	0.215	159
79	0.817	0.649	0.320	158
80	0.738	0.564	0.234	157
81	0.774	0.627	0.215	156
82	0.843	0.654	0.314	155
83	0.797	0.623	0.344	154
84	0.782	0.594	0.260	153
85	0.703	0.508	0.150	152
86	0.752	0.616	0.279	151
87	0.834	0.670	0.300	150
88	0.810	0.613	0.239	149
89	0.831	0.644	0.318	148
90	0.789	0.631	0.255	147
91	0.766	0.573	0.233	146
92	0.717	0.558	0.252	145
93	0.664	0.498	0.238	144
94	0.744	0.532	0.204	143
95	0.760	0.620	0.322	142
96	0.872	0.682	0.302	141
97	0.760	0.567	0.232	140
98	0.776	0.647	0.307	139
99	0.759	0.618	0.201	138
100	0.828	0.614	0.323	137
101	0.735	0.552	0.181	136
102	0.840	0.599	0.206	135
103	0.758	0.608	0.180	134
104	0.748	0.585	0.237	133
105	0.704	0.512	0.216	132
106	0.811	0.630	0.291	131

(continued)

Table 6.7 (continued)

RIP	100%	50%	0%	n
107	0.814	0.639	0.321	130
108	0.740	0.547	0.236	129
109	0.813	0.611	0.221	128
110	0.831	0.634	0.242	127
111	0.785	0.623	0.203	126
112	0.733	0.575	0.214	125
113	0.799	0.606	0.201	124
114	0.747	0.609	0.234	123
115	0.795	0.596	0.251	122
116	0.825	0.629	0.208	121
117	0.824	0.652	0.255	120
118	0.765	0.611	0.311	119
119	0.789	0.584	0.205	118
120	0.791	0.608	0.240	117
121	0.750	0.609	0.335	116
122	0.803	0.616	0.261	115
123	0.703	0.550	0.217	114
124	0.794	0.595	0.222	113
125	0.781	0.588	0.330	112
126	0.709	0.508	0.223	111
127	0.779	0.568	0.250	110
128	0.793	0.582	0.294	109
129	0.800	0.617	0.221	108
130	0.825	0.626	0.346	107
131	0.809	0.644	0.353	106
132	0.746	0.537	0.184	105
133	0.741	0.591	0.264	104
134	0.821	0.613	0.265	103
135	0.861	0.636	0.283	102
136	0.750	0.563	0.221	101
137	0.818	0.675	0.270	1
138	0.806	0.641	0.384	99
139	0.771	0.617	0.342	98
140	0.655	0.521	0.243	97
141	0.703	0.511	0.292	96
142	0.724	0.596	0.313	95
143	0.786	0.602	0.232	94
144	0.606	0.438	0.182	93
145	0.684	0.522	0.280	92

(continued)

Table 6.7 (continued)

RIP	100%	50%	0%	n
146	0.723	0.525	0.240	91
147	0.803	0.649	0.222	90
148	0.759	0.567	0.247	89
149	0.689	0.525	0.202	88
150	0.758	0.593	0.263	87
151	0.782	0.609	0.242	86
152	0.758	0.532	0.199	85
153	0.724	0.600	0.287	84
154	0.769	0.596	0.263	83
155	0.704	0.522	0.185	82
156	0.723	0.573	0.231	81
157	0.731	0.559	0.263	80
158	0.780	0.611	0.316	79
159	0.757	0.570	0.177	78
160	0.717	0.560	0.138	77
161	0.781	0.612	0.290	76
162	0.683	0.561	0.298	75
163	0.708	0.546	0.317	74
164	0.623	0.472	0.217	73
165	0.713	0.547	0.261	72
166	0.772	0.640	0.238	71
167	0.749	0.459	0.106	70
168	0.720	0.579	0.272	69
169	0.735	0.565	0.260	68
170	0.727	0.535	0.300	67
171	0.742	0.532	0.233	66
172	0.649	0.484	0.163	65
173	0.733	0.507	0.217	64
174	0.672	0.520	0.181	63
175	0.736	0.611	0.260	62
176	0.709	0.601	0.186	61
177	0.649	0.495	0.303	60
178	0.691	0.519	0.234	59
179	0.785	0.608	0.317	58
180	0.688	0.510	0.287	57
181	0.764	0.535	0.228	56
182	0.729	0.576	0.302	55
183	0.728	0.564	0.357	54
184	0.775	0.607	0.321	53

(continued)

Table 6.7 (continued)

RIP	100%	50%	0%	n
185	0.697	0.513	0.127	52
186	0.716	0.597	0.168	51
187	0.772	0.548	0.286	50
188	0.725	0.530	0.235	49
189	0.705	0.498	0.148	48
190	0.683	0.582	0.303	47
191	0.696	0.519	0.171	46
192	0.604	0.481	0.257	45
193	0.741	0.545	0.288	44
194	0.723	0.608	0.350	43
195	1.000	0.558	0.231	42
196	0.715	0.556	0.231	41
197	0.705	0.516	0.263	40
198	0.730	0.515	0.147	39
199	0.796	0.524	0.243	38
200	0.721	0.566	0.374	37
201	0.625	0.456	0.148	36
202	0.711	0.489	0.232	35
203	0.658	0.458	0.254	34
204	0.686	0.567	0.360	33
205	0.772	0.594	0.292	32
206	0.691	0.519	0.217	31
207	0.678	0.486	0.350	30
208	**1.000**	0.493	0.233	29
209	**1.000**	0.530	0.218	28
210	0.632	0.486	0.233	27
211	0.735	0.524	0.218	26
212	0.553	0.407	0.226	25
213	0.663	0.554	0.203	24
214	0.765	0.548	0.229	23
215	0.655	0.519	0.268	22
216	0.628	0.453	0.203	21
217	0.652	0.499	0.172	20
218	0.702	0.581	0.391	19
219	0.624	0.514	0.275	18
220	0.586	0.416	0.253	17

(continued)

Table 6.7 (continued)

RIP	100%	50%	0%	n
221	0.638	0.471	0.260	16
222	0.591	0.496	0.248	15
223	0.635	0.421	0.251	14
224	0.585	0.421	0.257	13
225	0.492	0.420	0.197	12
226	0.650	0.380	0.161	11
227	0.663	0.432	0.290	10
228	0.498	0.386	0.243	9
229	0.618	0.458	0.190	8
230	0.498	0.321	0.138	7
231	0.584	0.369	0.233	6
232	0.533	0.429	0.216	5
233	0.469	0.287	0.165	4
234	**0.175**	**0.096**	**0.085**	3
235	0.448	0.352	0.256	2
236	**0.088**	**0.088**	**0.088**	1
Max	1.0	0.723	0.391	
Min	0.088	0.088	0.085	
Mean	0.750	0.578	0.258	
50%	0.772	0.599	0.257	

Table 6.8 shows the ratio at which the correlation R becomes 0.6 or more in 236 groups in Table 6.7. “RIPi” represents a set of correlations of RIPi and RIPj (j > i), which is the stratified group name. As in Table 6.7, it decreases from 236 pairs to one pair such as (RIP236, RIP237). Within the stratum, column “RIPj” shows the suffix of RIPj having the maximum r. “Ratio” is the ratio of correlations over 0.6. For example, the RIP1 row is the summary of 236 correlations from RIP2 to RIP237, but when it is arranged in descending order by correlation, the correlation 0.811 with RIP4 becomes the maximum. After that, “r > = 0.6” column indicates that there are 189 pairs having “r >= 0.6” and that the ratio is 0.8 with respect to 236 pairs.

The three ratios of (RIP195, RIP196), (RIP208, RIP210), and (RIP209, RIP211) are 0.21, 0.17, and 0.29, respectively. Because these three pairs have “r = 1,” we focus on three pairs in this chapter.

Table 6.8 Ratio at which the R is 0.6 or more in 236 groups of Table 6.7

RIPi	R	RIPj	R >= 0.6	Ratio
1	**0.811**	**4**	**189**	**0.801**
2	0.873	22	210	0.894
3	0.776	79	150	0.641
4	0.857	73	205	0.880
5	0.822	166	179	0.772
6	0.818	9	174	0.753
7	0.810	31	108	0.470
8	0.821	49	167	0.729
9	0.791	45	140	0.614
10	0.825	44	165	0.727
11	0.789	30	191	0.845
12	0.784	87	99	0.440
13	0.750	184	108	0.482
14	0.794	25	136	0.610
15	0.840	22	174	0.784
16	0.814	154	185	0.837
17	0.799	22	138	0.627
18	0.835	65	156	0.712
19	0.833	60	119	0.546
20	0.847	32	183	0.843
21	0.844	28	162	0.750
22	0.831	130	165	0.767
23	0.776	143	148	0.692
24	0.797	71	153	0.718
25	0.816	45	159	0.750
26	0.785	109	170	0.806
27	0.823	72	166	0.790
28	0.834	45	148	0.708
29	0.800	32	145	0.697
30	0.775	117	144	0.696
31	0.820	41	160	0.777
32	0.861	72	184	0.898
33	0.842	72	162	0.794
34	0.790	41	150	0.739
35	0.783	71	121	0.599

(continued)

Table 6.8 (continued)

RIPi	R	RIPj	R >= 0.6	Ratio
36	0.823	142	112	0.557
37	0.811	72	145	0.725
38	0.835	179	149	0.749
39	0.863	46	153	0.773
40	0.792	51	139	0.706
41	0.801	82	140	0.714
42	0.856	57	162	0.831
43	0.768	168	133	0.686
44	0.826	72	114	0.591
45	0.827	82	143	0.745
46	0.842	72	147	0.770
47	0.764	137	107	0.563
48	0.780	127	105	0.556
49	0.792	72	130	0.691
50	0.807	137	152	0.813
51	0.864	137	160	0.860
52	0.841	117	144	0.778
53	0.701	83	56	0.304
54	0.789	68	117	0.639
55	0.777	89	109	0.599
56	0.812	87	124	0.685
57	0.840	179	143	0.794
58	0.782	71	131	0.732
59	0.729	60	53	0.298
60	0.794	166	138	0.780
61	0.801	138	128	0.727
62	0.824	138	143	0.817
63	0.815	147	135	0.776
64	0.763	71	95	0.549
65	0.809	68	92	0.535
66	0.760	71	84	0.491
67	0.804	166	113	0.665
68	0.862	107	139	0.822
69	0.816	71	129	0.768
70	0.823	122	124	0.743
71	0.832	87	143	0.861
72	0.811	82	134	0.812

(continued)

Table 6.8 (continued)

RIPi	R	RIPj	R >= 0.6	Ratio
73	0.807	99	107	0.652
74	0.805	98	124	0.761
75	0.751	89	82	0.506
76	0.780	138	90	0.559
77	0.747	219	69	0.431
78	0.778	143	91	0.572
79	0.817	135	115	0.728
80	0.738	166	55	0.350
81	0.774	96	106	**0.679**
82	0.843	117	114	0.735
83	0.797	98	104	0.675
84	0.782	138	76	0.497
85	0.703	127	13	0.086
86	0.752	87	91	0.603
87	0.834	117	123	0.820
88	0.810	137	88	0.591
89	0.831	111	101	0.682
90	0.789	129	104	0.707
91	0.766	109	58	0.397
92	0.717	114	47	0.324
93	0.664	138	9	0.063
94	0.744	204	36	0.252
95	0.760	142	88	0.620
96	0.872	120	118	0.837
97	0.760	109	50	0.357
98	0.776	138	97	0.698
99	0.759	114	78	0.565
100	0.828	159	79	0.577
101	0.735	168	29	0.213
102	0.840	196	70	0.519
103	0.758	147	78	0.582
104	0.748	165	59	0.444
105	0.704	164	24	0.182
106	0.811	120	89	0.679
107	0.814	147	90	0.692
108	0.740	157	39	0.302
109	0.813	166	74	0.578

(continued)

Table 6.8 (continued)

RIPi	R	RIPj	R >= 0.6	Ratio
110	0.831	134	82	0.646
111	0.785	137	81	0.643
112	0.733	169	55	0.440
113	0.799	179	76	0.613
114	0.747	147	70	0.569
115	0.795	138	61	0.500
116	0.825	151	83	0.686
117	0.824	138	86	0.717
118	0.765	130	70	0.588
119	0.789	157	51	0.432
120	0.791	168	64	0.547
121	0.750	184	73	0.629
122	0.803	137	67	0.583
123	0.703	187	32	0.281
124	0.794	205	56	0.496
125	0.781	138	51	0.455
126	0.709	134	18	0.162
127	0.779	161	47	0.427
128	0.793	138	46	0.422
129	0.800	135	63	0.583
130	0.825	183	67	0.626
131	0.809	138	76	0.717
132	0.746	153	29	0.276
133	0.741	204	49	0.471
134	0.821	166	61	0.592
135	0.861	153	71	0.696
136	0.750	168	34	0.337
137	0.818	150	82	0.820
138	0.806	184	67	0.677
139	0.771	179	61	0.622
140	0.655	168	12	0.124
141	0.703	158	18	0.188
142	0.724	165	48	0.505
143	0.786	205	50	0.532
144	0.606	175	2	0.022
145	0.684	154	13	0.141
146	0.723	157	17	0.187

(continued)

Table 6.8 (continued)

RIPi	R	RIPj	R >= 0.6	Ratio
147	0.803	191	66	0.733
148	0.759	179	39	0.438
149	0.689	177	24	0.273
150	0.758	198	42	0.483
151	0.782	166	46	0.535
152	0.758	191	24	0.282
153	0.724	182	45	0.536
154	0.769	168	42	0.506
155	0.704	220	14	0.171
156	0.723	198	34	0.420
157	0.731	181	29	0.363
158	0.780	207	46	0.582
159	0.757	160	33	0.423
160	0.717	183	25	0.325
161	0.781	179	46	**0.605**
162	0.683	211	22	0.293
163	0.708	166	17	0.230
164	0.623	166	3	0.041
165	0.713	178	18	0.250
166	0.772	205	53	0.746
167	0.749	196	8	0.114
168	0.720	210	29	0.420
169	0.735	176	23	0.338
170	0.727	193	15	0.224
171	0.742	211	16	0.242
172	0.649	194	10	0.154
173	0.733	204	18	0.281
174	0.672	216	12	0.190
175	0.736	179	37	0.597
176	0.709	179	36	0.590
177	0.649	183	8	0.133
178	0.691	219	10	0.169
179	0.785	207	32	0.552
180	0.688	204	12	0.211
181	0.764	193	16	0.286
182	0.729	191	26	0.473
183	0.728	204	23	0.426

(continued)

Table 6.8 (continued)

RIPi	R	RIPj	R >= 0.6	Ratio
184	0.775	200	30	0.566
185	0.697	194	12	0.231
186	0.716	200	26	0.510
187	0.772	204	16	0.320
188	0.725	209	12	0.245
189	0.705	213	6	0.125
190	0.683	205	19	0.404
191	0.696	204	16	0.348
192	0.604	218	1	0.022
193	0.741	201	15	0.341
194	0.723	221	23	0.535
195	**1.000**	**196**	9	0.214
196	0.715	200	8	0.195
197	0.705	213	12	0.300
198	0.730	219	10	0.256
199	0.796	205	12	0.316
200	0.721	218	13	0.351
201	0.625	206	1	0.028
202	0.711	209	4	0.114
203	0.658	206	5	0.147
204	0.686	218	16	0.485
205	0.772	213	13	0.406
206	0.691	218	6	0.194
207	0.678	215	7	0.233
208	**1.000**	**210**	5	0.172
209	**1.000**	**211**	8	0.286
210	0.632	216	3	0.111
211	0.735	213	7	0.269
212	0.553	227	0	0.000
213	0.663	225	6	0.250
214	0.765	232	7	0.304
215	0.655	223	3	0.136
216	0.628	220	1	0.048
217	0.652	219	3	0.150
218	0.702	219	6	0.316
219	0.624	220	2	0.111
220	0.586	229	0	0.000

(continued)

Table 6.8 (continued)

RIPi	R	RIPj	R >= 0.6	Ratio
221	0.638	228	1	0.063
222	0.591	224	0	0.000
223	0.635	227	1	0.071
224	0.585	235	0	0.000
225	0.492	226	1	0.083
226	0.650	232	1	0.091
227	0.663	231	1	0.100
228	0.498	235	0	0.000
229	0.618	232	1	0.125
230	0.498	233	0	0.000
231	0.584	237	0	0.000
232	0.533	237	0	0.000
233	0.469	237	0	0.000
234	**0.175**	235	0	0.000
235	0.448	237	0	0.000
236	**0.088**	237	0	0.000
		Max	**210**	**0.900**
		Min	**0**	0.000
		Sum	**16380**	111.320
		Mean	**69.41**	0.470

6.3.3 Examination of Three RipDSs with a Correlation of 1

We examine three pairs of (RIP195, RIP196), (RIP208, RIP210), and (RIP209, RIP211) having r = 1 in Table 6.8. Figure 6.2 shows PCA plots of these six variables. The left plot is the eigenvalue. The first eigenvalue is 4.24, and the contribution rate is large at 70.67. The second eigenvalue is 0.8978, and the contribution rate is 14.96%, and the cumulative rate is 85.635%. The score plot in the middle also shows the same characteristics described in other microarrays. The plot on the right is a factor loading plot. The six RipDSs locate in the first quadrant and fourth quadrants.

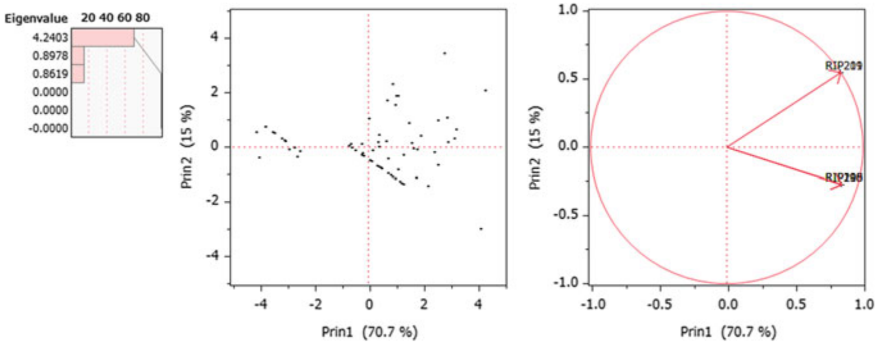


Fig. 6.2 Three plots of PCA

Table 6.9 shows the details of the factor loading plot in Fig. 6.2. The row directions from the Prin1 to the Prin6 are the factor loading amount of six variables. The factor loading amounts of Prin1 and Prin2 correspond to the right-side plot of Fig. 6.2. Because two correlations between (RIP195 vs. RIP196) and (RIP208 vs. RIP210) are 1, and the factor loading amounts of the two principal components are the same, these four RipDSs are the same in the fourth quadrant of the factor loading plot. The factor loading amounts show that the correlations between four RipDSs and Prin1 are 0.843, and the correlations with Prin2 are −0.274. Thus, these four RipDSs have the same role. RIP209 and RIP211 have a correlation coefficient of 1, and the factor loading amount is in the first quadrant. These two RipDSs are considered to replace each other at all. Because they have high correlations with other RIPs, we are expecting that they correspond to the new class pointed by Golub et al. that is a subclass of cancer. Verification by a specialist is necessary.

Table 6.9 Details of factor loading plots up to the six principal component

PCA/Var.	RIP195	RIP196	RIP208	RIP210	RIP209	RIP211
Prin1	0.843	0.843	0.843	0.843	0.836	0.836
Prin2	−0.274	−0.274	−0.270	−0.270	0.549	0.549
Prin3	0.463	0.463	−0.465	−0.465	0.002	0.002
Prin4	0.166	0.166	0.144	0.144	0.874	0.874
Prin5	−0.619	−0.619	−0.974	−0.974	−0.729	−0.729
Prin6	0.482	0.482	−0.330	−0.330	0.413	0.413

6.4 Analysis of 30 RipDSs of 30 SMs and 18 HsvmDSs of 18 SMs

In this section, we select 30 SMs and 18 SMs from 237 SMs and discriminate those by RIP and H-SVM. RIPi indicates the DS of SMi discriminated by RIP. The 30 RIPs consist from RIP1 to RIP18, six RIPs having R = 1, and last six RIPs from RIP232 to RIP237. On the other hand, the 18 HSVMs consist six HSVMs from HSVM1 to HSVM6, six HSVMs having R = 1, and last six HSVMs from HSVM232 to HSVM237. By examining these two sets of SMs, we aimed to validate no difference of results. RIP and H-SVM discriminate between two types of SMs, and we make two signal data. The cluster analysis and PCA show two classes are entirely separable as same as 234 RIPs. After that, we analyze the transposed new data by cluster analysis and PCA. Most results are the same as Alon and Golub et al.

6.4.1 Cluster Analysis

Figure 6.3 is Ward cluster analysis of 30 RipDSs new data of 30 SMs. It consists of 77 cases (patients) and 30 variables (RIPs). The color map with 77 rows and 30 columns is entirely separated in the upper 58 rows of DLBCL and lower 19 rows of FL. Furthermore, 58 DLSCL becomes seven clusters. The lower dendrogram is a dendrogram of these 30 RIPs. At first, RIP195 and RIP196 become the first cluster. Next, RIP208 and RIP210 become the second cluster. Third, RIP209 and RIP211 become the third cluster. The correlations of these three pairs are 1.

We had better specified the cluster number by moving right top marker diamond shape. Now, we choose eight clusters and identify by the same color and symbols. Thus, the case dendrogram on the right consists seven clusters of 58 DLBCL cases showed by seven colors. Below that, 19 FL cases become one green cluster. On the left side, the different markers identify the corresponding eight colors. The top red ○ mark is the first cluster of 29 DLBCL cases. Thereafter, there are the second cluster of six cases of green + signs, the third cluster of three cases of blue ◇ marks, the fourth cluster of 12 cases of orange × marks, the fifth cluster of two cases of light blue Δ marks, the sixth cluster of four cases of purple marks, and the seventh yellow cluster of two cases. There is the one cluster of 19 cases of green FL cases after that. It is essential that DCBCL and FL are divided into two classes cleanly. The same result can be obtained by using 237 DSs. It seems that seven clusters of DCBCL also have medical implications. In the middle color map, the values of each case and variable are hierarchized with green → white → red. Because the green pixel is the smallest and the red is large, it can be understood that the FL class is a mild cancer class compared to DLBCL without medical knowledge. Also, the seventh yellow cluster having two cases is strongly related to three RipDSs (RIP195, RIP196, RIP235) because the six cells made by two cases and three RipDSs are red. On the other hand, all FL cases seem to have a weak relationship with RipDSs because most of the cells are green.

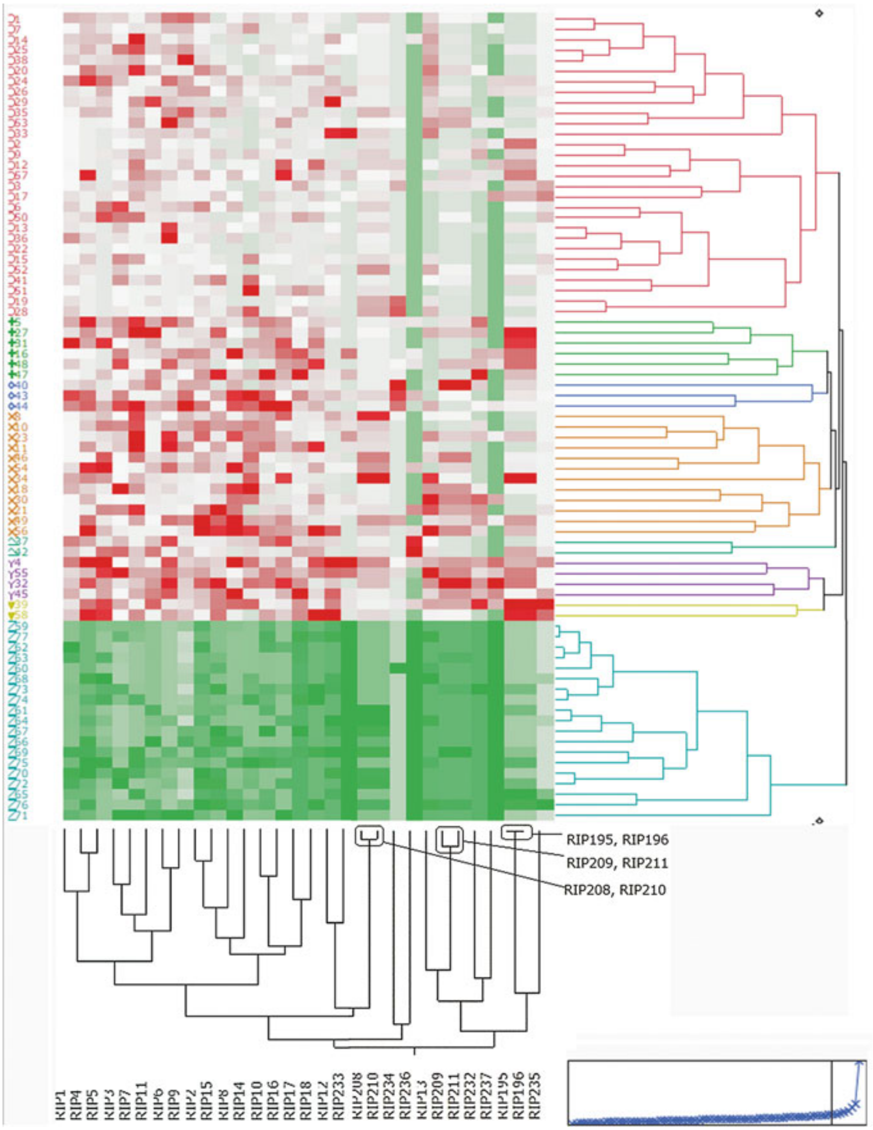


Fig. 6.3 Ward cluster analysis of 30 RipDSs data

Figure 6.4 is a cluster analysis of 18 HsvmDSs signal data. The DLBCL has seven clusters as follows: 1: 29 red ◯ marks, 2: 4 green + marks, 3: 8 blue ◇ marks, 4: 8 orange × marks, 5: 5 green △ marks, 6: 3 purple Y marks 7: one yellow ▼ marks. The 19 FL cases are one cluster. Because we cannot discuss the medical meaning of cluster analysis, we expect the medical specialists to examine our cluster analysis. However, FL has few cases, but because DLBCL is divided into seven clusters, it can be judged that it is mild compared to DLBCL without medical knowledge. This proves that signal data correctly represents signal information. Furthermore, we must be aware cluster analysis does not show the information of R = 1 well. Fig. 6.3 and Fig. 6.4 tells us this remark because the merged distances of three pairs are different.

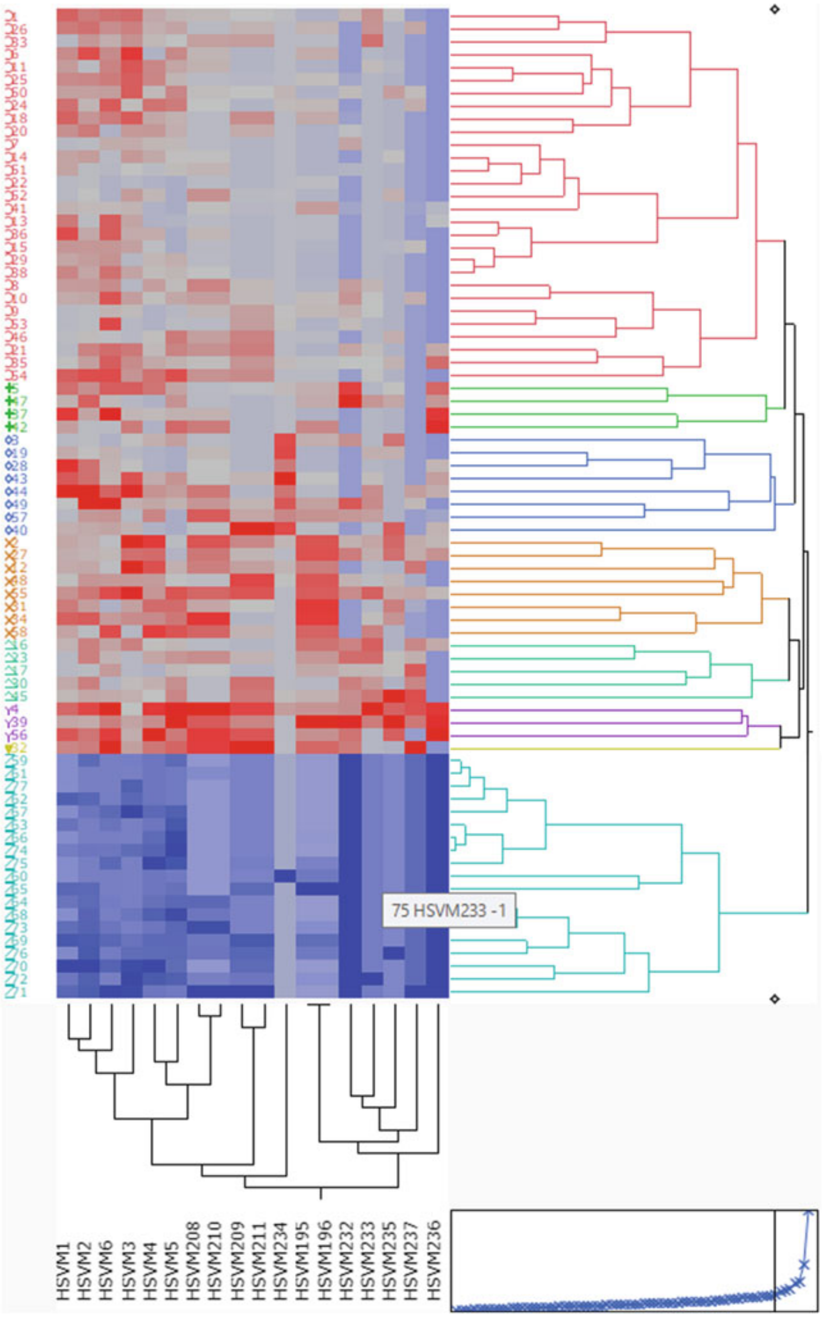


Fig. 6.4 Ward cluster analysis of 18HsvmDSs data

6.4.2 Principal Component Analysis (PCA)

Figure 6.5 is the PCA plots of RipDSs signal data. From the left eigenvalue, the first eigenvalue is 17.97, and the contribution rate is 59.9%. The second eigenvalue is 1.588, the contribution rate is 5.295%, and the cumulative contribution rate is 65.195%. That is, the Prin1 almost represents 30 RipDSs. From the score plot in the middle, because the second eigenvalue is small and the fluctuation is small, it can be seen that the FL case is almost placed on the axis of −4.47 or less of the Prin1. DLBCL cases are in the range of approximately −1 to 7. As the distance from FL increases, the dispersion of the Prin2 becomes large. This result is relatively similar to the results of Alon and Singh which consist of cancer and healthy patients. However, it is a slightly different point that FL cases are in the fan shape. In Fig. 6.3, 29 cases of red ○ are apart from FL, but in PCA it is found that it is closest to the FL case. We must be aware of this fact that the cluster analysis does not represent the spatial location information of each patient well. Six cases of green + signs are adjacent to it. Four cases of purple Y signs are outliers. Shipp et al. can verify whether these are subspecies of DLBCL.

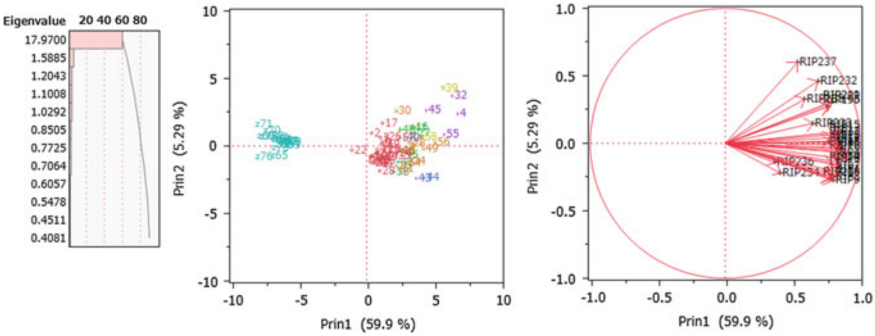


Fig. 6.5 PCA three plots of 18 RipDSs

The first two columns of Table 6.10 are the details of RipDSs corresponding to Fig. 6.5. The number of three lines or less in the first column is SN. The second column is the value of the Prin1, and it is rearranged in descending order from a large value. The leftmost one in Fig. 6.5 is the FL patient, which is −6.209 of SN = 76. In FL patients closest to DLBCL, SN = 77 is −4.377, and FL range is [−6.209, −4.377]. On the other hand, the DLBCL range is [−0.744, 6.176]. SV opens the window of (−4.377, −0.744). RatioSV is 29.328% for the range [−6.209, 6.176] on the Prin1. Assuming that it is about 30%, FL and DLBCL are separated and placed in the remaining 70%. Because this is a comprehensive of 30 RIPs, it is 0.5% larger than 28.80% of the maximum value of 237 RatioSVs discussed in Table 6.6. With this degree of improvement, it is necessary to examine many genes and cannot be used for cancer gene diagnosis. It can be thought of simply as a conceptual diagram representing a malignancy indicator. For the new class discovery pointed out by Golub, they use the SOP of the K-means method and clusters that are considered to be a new class and are identified by fixing the number of clusters to two at first. After that, it is verified by the class prediction method such as the voting method. In the second step, they survey the three clusters and so on. On the other hand, in this research, we propose to find the outliers of RipDSs signal data by cluster analysis and PCA without fixing the assumed number of clusters. This is possible because the signal data omit the noise. However, even if SM is clustered merely, such an obvious result cannot be obtained, and RIP, Revised LP-OLDF, and H-SVM are considered to have a significant effect. Because we use a DS that best discriminates between FL and DLBCL, we believe that outlier cases are more likely to be new classes.

Table 6.10 RIP and HSVM Prin1 values sort by RatioSV

RatioSV	29.328	RatioSV	24.535
SN	Prin1 by RIP	SN	Prin1 by HSVM
4	6.176	4	7.48
32	5.443	39	6.64
39	5.143	32	5.60
55	4.665	56	5.49
49	3.297	55	5.21
34	3.140	2	4.22
58	2.971	27	3.45
56	2.936	34	3.40
45	2.765	45	3.12
40	2.762	44	3.00
54	2.693	49	2.98
23	2.552	58	2.95
48	2.528	12	2.77
43	2.359	31	2.45
27	2.307	48	2.41
46	2.269	54	2.36
16	2.261	23	2.32
5	2.209	57	2.28
21	2.143	5	2.24
57	2.041	16	2.20
44	2.026	40	2.04
31	1.923	21	1.84
42	1.918	42	1.80
24	1.817	43	1.74

(continued)

Table 6.10 (continued)

RatioSV	29.328	RatioSV	24.535
SN	Prin1 by RIP	SN	Prin1 by HSVM
30	1.676	30	1.74
47	1.640	47	1.62
35	1.625	35	1.54
18	1.527	26	1.40
8	1.383	18	1.38
17	1.357	10	1.33
6	1.343	1	1.20
10	1.296	33	1.05
26	1.175	37	1.05
9	1.131	24	1.05
1	1.079	6	1.05
3	1.031	8	1.05
37	0.956	50	0.92
53	0.910	46	0.83
50	0.893	3	0.82
12	0.888	25	0.66
33	0.834	17	0.58
2	0.826	11	0.53
52	0.766	20	0.35
20	0.740	9	0.32
19	0.717	36	0.29
11	0.675	28	0.22
13	0.581	19	0.05
28	0.564	38	0.01
25	0.561	41	0.00
41	0.490	52	−0.05
36	0.438	53	−0.08
15	0.348	14	−0.14
38	0.034	15	−0.21
14	−0.019	13	−0.22
29	−0.058	29	−0.30
7	−0.131	7	−0.38
51	−0.499	51	−0.48
22	**−0.744**	**22**	**−1.02**
77	**−4.377**	**61**	**−4.47**
60	−4.401	77	−4.48
59	−4.413	59	−4.50
66	−4.554	66	−4.62

(continued)

Table 6.10 (continued)

RatioSV	29.328	RatioSV	24.535
SN	Prin1 by RIP	SN	Prin1 by HSVM
74	−4.585	63	−4.64
67	−4.624	62	−4.75
63	−4.644	74	−4.85
62	−4.686	67	−4.88
68	−4.973	60	−4.90
61	−4.995	75	−5.05
73	−5.003	68	−5.09
64	−5.142	64	−5.17
71	−5.295	73	−5.57
65	−5.314	65	−5.59
75	−5.407	72	−5.64
72	−5.829	76	−5.67
70	−5.953	70	−5.76
69	−5.970	69	−5.97
76	−6.209	71	−6.59

Figure 6.6 shows the result of 18 HsvmDSs signal data by PCA. From the eigenvalues, it can be seen that the eigenvalue of the Prin1 is larger than others. The first eigenvalue is 11.382, and the contribution rate is 63.2%. The second eigenvalue is 1.299, the contribution rate is 7.22%, and the cumulative contribution rate is 70.42% which is the same as RIP. That is, the Prin1 almost represents 18 HsvmDSs signal data. From the middle score plot, the second eigenvalue is small, and the fluctuation is small, so it can be seen that the FL cases are almost on the Prin1 axis of −4.47 or less. The DLBCL cases are in the range of approximately −1.02 to 7.48, and the variation of the Prin2 becomes large as it leaves the FL. That is, the Prin1 can be used as a conceptual diagram of cancer malignancy index as well as individual RipDS. If the priority of the DS used for diagnosing the malignancy of cancer is established, the diagnosis can be carried out by inspecting fewer genes. Therefore, the use of PCA decreases.

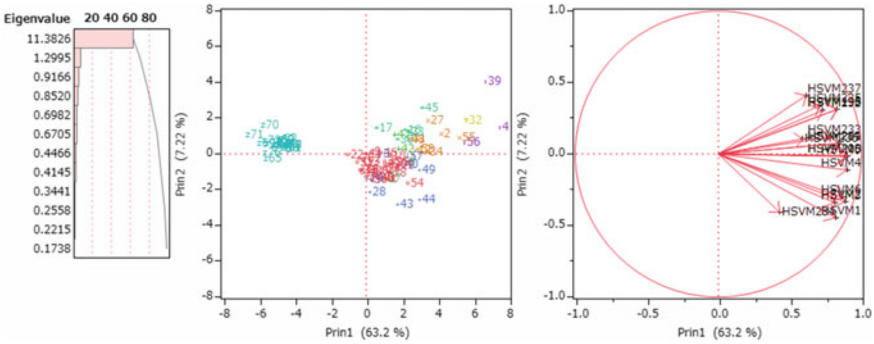


Fig. 6.6 Analysis result of discriminant score data of HSVM by PCA

The numbers of three rows or less of the third and fourth columns in Table 6.10 are the values of the HsvmDSs data. The fourth column is the value of the Prin1 by the PCA, and it is rearranged in descending order. The range of FL is [−6.59, −4.47]. The range of DLBCL is [−1.02, 7.48]. Because the range of SV is (−4.47, −1.02) and the range of HsvmDSs is [−6.59, 7.48], RatioSV is 24.535%. Assuming that it is about 25%, FL and DLBCL patients are separated and placed in the remaining 75%. Although this is a comprehensive of 18 RIPs, it is 9% smaller than the 33.34% of the maximum value of 237 RatioSVs discussed in Table 6.6. RatioSV has drawbacks depending on the minimum and maximum values, but we do not know if there is another reason why RatioSV of PCA is small.

6.5 Analysis of Transposed Data

Figure 6.7 shows the results of cluster analysis by transposed data of RipDSs signal data. The row corresponds to 237 RipDSs, and the variable in the column direction corresponds to the 58 cases of DLBCL and the red 19 cases of FL. Because it is a transposed data, the variable dendrogram separates both patients with DLBCL and FL into two clusters. On the other hand, the case dendrogram categorizes 237 RipDSs into the 20 clusters such as the 92 red RipDSs, the 91 green RipDSs, and the miscellaneous RipDSs. It seems better to divide it into three DSs. Unlike the previous two clusters, the last 54 RIPs become the clustering by the large distances. It is different from the previous 183 RIPs. The 54 RIPs are considered to be affected by specific patients. We think that a new class finding should be studied by contrasting these 54 with the previous 183. However, exploratory cluster analysis often results in different results when changing the method of analysis, the definition of distance, changing the number of cases, and variables analyzed. Which results to adopt should be done with medical knowledge. Here, we will focus on 54 cases.

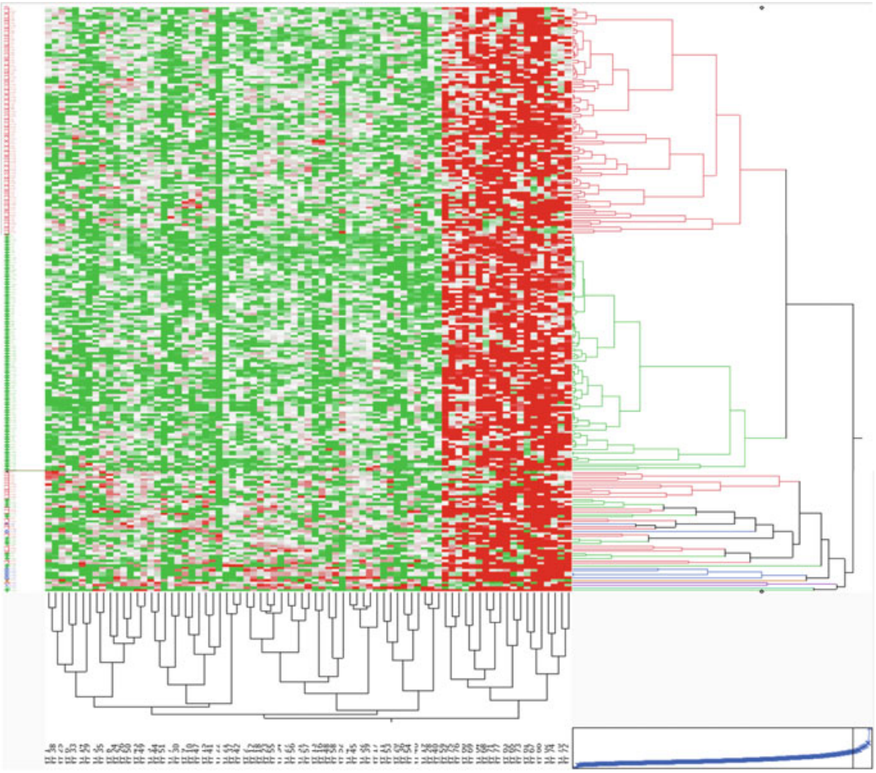


Fig. 6.7 Cluster analysis of transposed signal data by 237 RipDSs

In Fig. 6.8, cluster analysis was performed using the 54 RipDSs explained in Fig. 6.7. By omitting 183 cases, the shape of the dendrogram changed, but this should be judged based on whether the medically useful explanation is possible or not. The 24 ◯ marks are clusters that are away from each other, and the 19 green marks with shorter distances are sequentially clustering. The last 11 cases become four clusters of seven cases, one case, one case, and two cases at vast distances. We need to compare this result by case studies of patients. The cluster analysis is ambiguous and cumbersome because statistician obtain various results. However, it provides a variety of different perspectives, so physicians can carefully choose what suits their specific knowledge.

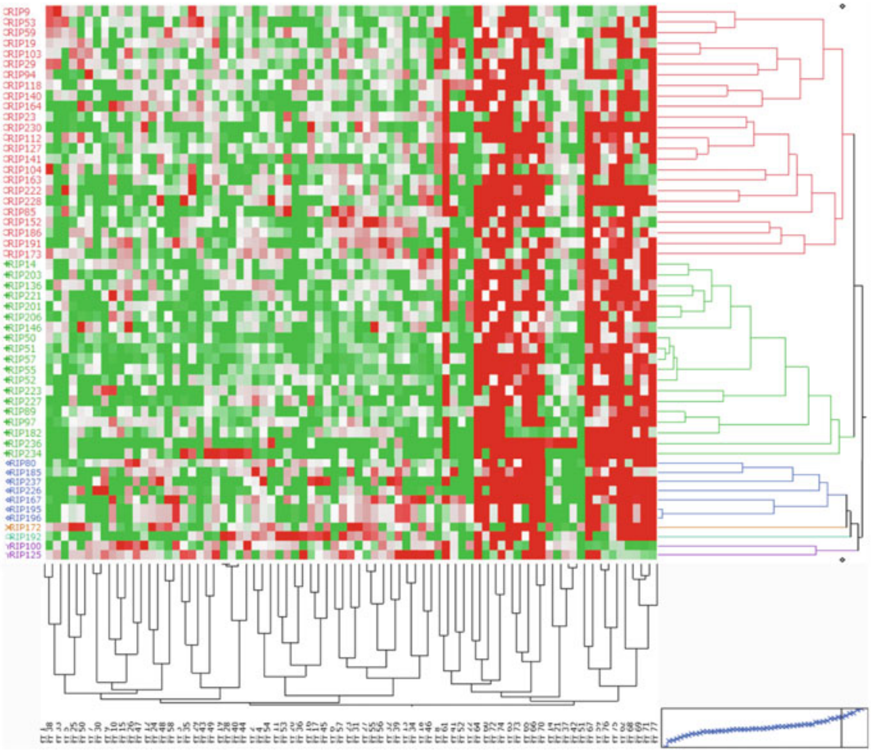


Fig. 6.8 Cluster analysis using the 54 RipDSs signal data

Figure 6.9 shows the results of PCA corresponding to Fig. 6.8. In cluster analysis, the positioning in the data space of each case is ambiguous. However, even if we lack medical knowledge, we do not make a big mistake to interpret the results of PCA. Because 19 green RipDSs include the origin, these cases are our understanding base. Moreover, the 24 red cases (◯) in the first and fourth quadrants surround the green cases. Cases in first quadrant of the factor loading plot are closely related to RipDSs in first quadrant of the score plot. That is, the principle that the association between (RipDS and case) in each quadrant of (score plot and factor plot) is large is essential for PCA interpretation. Moreover, 11 cases are outliers, but because they are separated by a considerable distance, respectively, it turns out that interpretation of these clusters by cluster analysis is difficult. However, score plot indicates these are outliers in the first and fourth quadrants. This shows the appropriateness of the method of interpreting the cluster indicated by cluster analysis with PCA. In the research of gene analysis, it is regretful that cluster analysis was taken and PCA was not used.

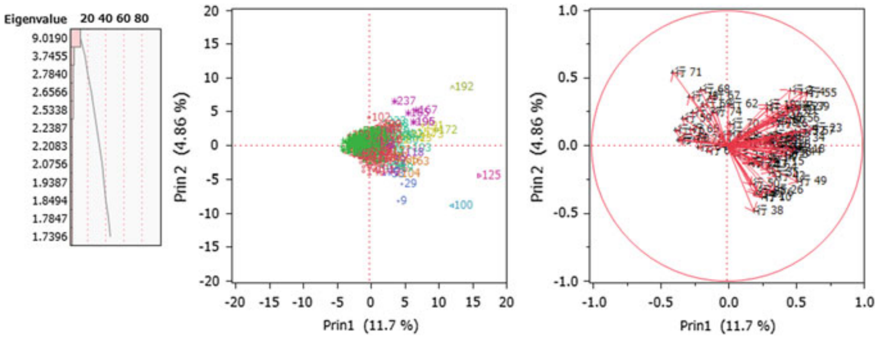


Fig. 6.9 PCA result of 54 RipDSs signal data

Figure 6.10 shows the result of transposed data of the 237 HsvmDSs signal data. Green HSVM234 separates upper red cluster and lower blue cluster. Under them, five clusters appear. However, in the variable dendrogram, only one red FL case of SN = 62 is separated from the remaining 18 FL cases and shows no sign of linear separation of two classes. Physicians should consider why this patient is clustered to DLBCL patients.

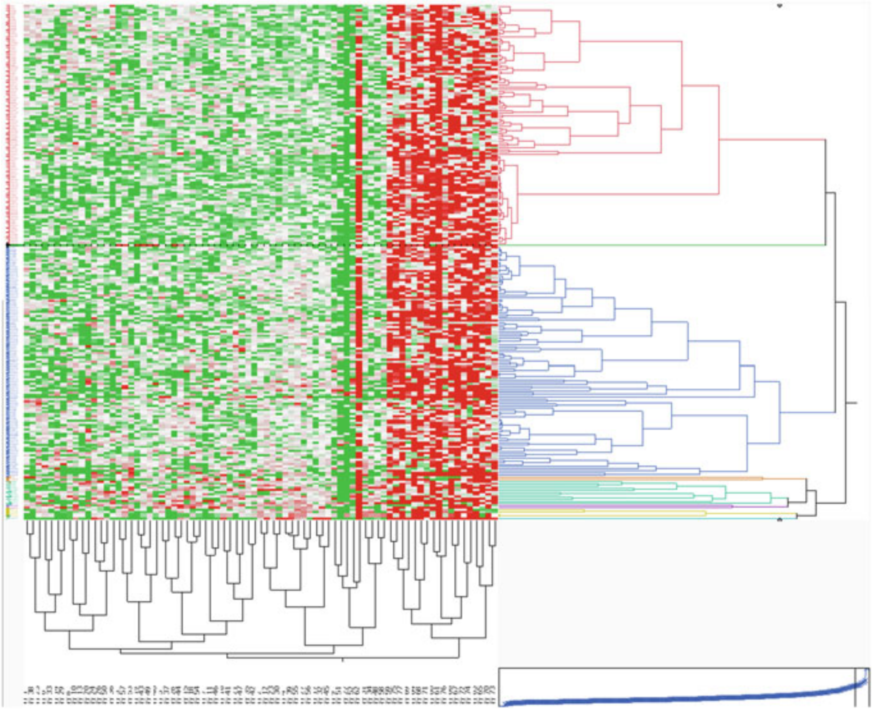


Fig. 6.10 Cluster analysis transposed signal data by 237 HsvmDSs

Figure 6.11 shows the result of transposed signal data by PCA. The red cluster in Fig. 6.10 contains the origin and is the core of our study. Moreover, blue clusters are predominantly in the first quadrant; other clusters are in the first and fourth quadrants, overlapping blue. Several HsvmDSs are outliers in the first and second quadrants. Verification of the medical meaning of 237 HsvmDSs should take priority on what is included in the red cluster, but it may be easier to verify specific outliers. The FL patient with SN = 62 in Fig. 6.10 is close to 0.3 on Prin2 axis and is separate from other FL patients. Moreover, it is clear that it is close to the patients with DLBCL in the first quadrant. Clear outliers indicate the possibility of new subspecies, but because there is no medical knowledge, we point out only the possibilities.

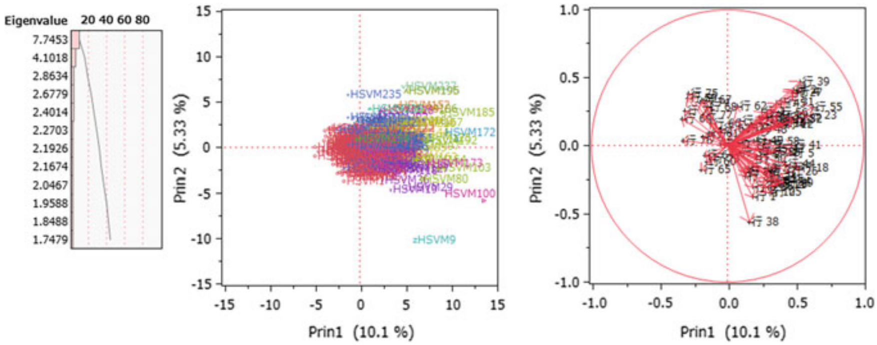


Fig. 6.11 PCA result of transposed data of 237 HsvmDSs

6.6 Conclusions

In this chapter unlike previous analyses of Alon and Golub, we focus on 30 RipDSs and 18 HsvmDSs chosen from 237 SMs. We analyzed in detail only two sets of different SMs. This comparison allowed us to examine cluster analysis and PCA results in detail. The most significant result will be the method of discovering new subclasses of cancer proposed by Golub et al. Because RIP and HSVM can predict cancer classes, both LDFs offer a simple approach. We hope that this approach will contribute to the diagnosis of cancer malignancy.

References

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Lachenbruch PA, Mickey MR (1968) Estimation of error rates in discriminant analysis. Technometrics 10(1):11

Sall JP, Creighton L, Lehman A (2004) JMP start statistics, 3rd edn. SAS Institute Inc, USA (Shinmura S edits Japanese version)

Schrage L (2006) Optimization Modeling with LINGO. LINDO Systems Inc (Shinmura S translates Japanese version)

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

Shinmura S (2016) New theory of discriminant analysis after R. Springer, Fisher

Shinmura S (2016a) New theory of discriminant analysis after R. Fisher, Springer, Tokyo

Shinmura S (2017) From cancer gene analysis to cancer gene diagnosis. Amazon

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD'18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. In: Biocomp'18, pp 1–7

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1.1):68–74. (https://doi.org/10.1038/nm0102-6)

Vapnik V (1995) The nature of statistical learning theory. Springer

Chapter 7
Cancer Gene Diagnosis of Singh et al. Microarray

Abstract Chapter 1 explained the new theory of discriminant analysis after R. A. Fisher (Theory). The theory solved five problems completely. Especially, Revised IP-OLDF (RIP) and Method2 firstly succeeded in the cancer gene analysis. RIP could find six microarrays were LSD (Fact3). LINGO Program3 of Method2 could decompose the microarray into many SMs and another noise subspace (Fact4). In Chap. 2, we make signal data made by RIP discriminant scores (RipDSs). Our breakthrough opens the new frontier of cancer gene diagnosis and malignancy indexes. We find the new problem (Problem6): "Why could no researchers find the linear separable facts in microarrays and SM from 1970?" In this book, we explain the several answers of Problem6. In this chapter, we survey how to make different RipDSs from many SMs. It explains why microarray consists of many SMs and the different RipDSs. By these results, we wish to classify SMs into several categories of malignancy indexes in the future.

Keywords Singh's microarray · Cancer gene diagnosis · Malignancy indexes and RatioSV · RIP discriminant scores (RipDSs) · Signal data made by RipDSs · Correlations of RipDSs

Thanks to Singh et al.
We thank Singh et al.[1] (2002) for providing their microarray that consists of 102 subjects (50 normal subjects and 52 tumor prostate patients) and 12,625 genes. We will quote their "Abstract" for the reader.

> Prostate tumors are among the most heterogeneous of cancers, both histologically and clinically. Microarray expression analysis was used to determine whether global biological differences underlie common pathological features of prostate cancer and to identify genes that might anticipate the clinical behavior of this disease. While no expression correlates of age, serum prostate-specific antigen (PSA), and measures of the local invasion were found, a set of genes was identified that strongly correlated with the state of tumor differentiation as measured by Gleason score. Moreover, a model using gene expression data alone accurately

[1] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, and William R. Sellers.

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_7

predicted patient outcome following prostatectomy. These results support the notion that the clinical behavior of prostate cancer is linked to underlying gene expression differences that are detectable at the time of diagnosis.

7.1 Introduction

Chapter 1 explained the new theory of discriminant analysis after R. A. Fisher (Theory) that solves five problems completely. Especially, Revised IP-OLDF (RIP) based on the MNM criterion and Matryoshka feature selection method (Method2) firstly succeeded in the cancer gene analysis. RIP finds six microarrays are LSD and MNM = 0 (Fact3). LINGO Program3 of RIP can decompose the microarray into the many SMs (MNMs = 0) and another noise gene subspace (MNM >= 1) that is Fact4. In Chap. 2, although all SMs are small samples, the standard statistical methods cannot find the linear separable facts. Thus, we make signal data made by RIP discriminant scores (RipDSs) in 2016. Our breakthrough opens the new frontier of cancer gene diagnosis and many malignancy indexes using all SMs of six microarrays in 2016. Furthermore, the 64 SMs and 130 BGSs of Alon are compared by RatioSVs in 2017. We judge that BGSs are useless for cancer gene diagnosis because all RatioSVs are less than 1%. In Chap. 3, we develop the method of choosing the proper SMs, and we compare two sets of SMs chosen by the RIP and Revised LP-OLDF. In Chap. 4, we find two critical facts. The first fact is the defect of Revised LP-OLDF that cannot find all SMs from the microarray. The second fact is the reason why researchers cannot find the linear separable facts in microarray and SM from 1970 (Problem6). After Chap. 5, the above themes are intensely surveyed using the other five microarrays. In this chapter, we survey how to make different RipDSs from many SMs. It explains why microarray consists of many SMs and different RipDSs using SMs. By these results, we classify SMs into several categories of malignancy indexes for cancer gene diagnosis in the future (Problem7). LINGO (Schrage 2006) decomposes Singh's microarray into 139 SMs and opens a new frontier of cancer gene analysis. JMP (Sall et al. 2004) analyzes all SMs and offers cancer gene diagnosis. Shinmura (2016, 2017, 2018a, b) relate to this Chapter.

7.2 Problem6 of Cancer Gene Analysis

RIP and Method2 decompose Singh's microarray into 139 SMs (4,046 genes). We analyze 139 SMs by the standard statistical methods such as one-way ANOVA, t-test, Ward cluster analysis, principal component analysis (PCA), logistic regression, Fisher's LDF, and quadratic discriminant analysis (QDA). Although we expect those methods are useful for cancer gene diagnosis, only logistic regression can discriminate all SMs correctly, and other methods do not show the linear separable facts. On the other hand, because RIP discriminates 139 SMs, and the range of 139 RatioSVs is [1.04, 29.11%], we make signal data that consists of 102 subjects and 139 RipDSs instead of 12,625 genes. By this breakthrough, cluster analysis can separate two classes as two clusters. In addition to this result, the Prin1 of PCA indicates malignancy indexes of three signal data as same as 139 malignancy indexes by RIP, Revised LP-OLDF, and H-SVM. Moreover, we expect that several outliers by the scatter plot of the transposed signal data may be new subclasses of cancer pointed out by Golub et al. (1999). We hope to cooperate with Singh and other medical experts to validate our results of their microarray. If they examine and confirm our malignancy indexes, we will open the new frontier of cancer gene diagnosis by microarrays through our cooperation. They will be able to accomplish their research completely.

On the other hand, we explain the reason why researchers cannot find the linear separable facts in microarrays and SMs from 1970 (Problem6) in Chap. 4. Although two SVs can separate two classes of microarray or each SM, the variation of the two classes is tiny, and the signal is buried in noise. In this chapter, we explain how to make 139 RipDSs.

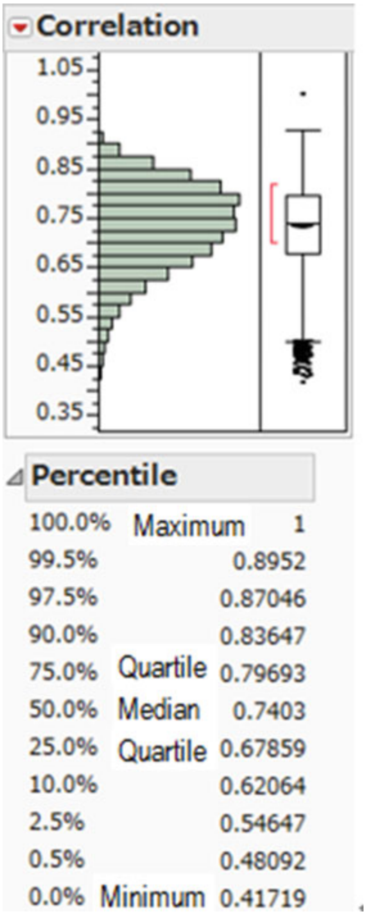
7.3 Examination of RipDSs and SMs

At first, we considered all SMs (or BGSs) were signals. Now, we consider the signal data made by RipDSs, LpDSs, and HsvmDSs are valid signals. The correlation analysis and PCA analyze the signal data and show several results.

7.3.1 Correlations of 139 RipDSs

Figure 7.1 is the histogram of 9,591 correlations (abbreviated R) by 139 RipDSs. The range of correlations is [0.417, 1]. R = 1 is an outlier, and nine Rs are over 0.912. The range of Shipp is [0.0697, 1], also. In the cancer gene diagnosis, we expect the two SMs with R = 1 are complementary. We think that medical experts shall validate this claim.

Fig. 7.1 Histogram of 9,591 correlations of 139 RipDSs



Percentile

100.0%	Maximum	1
99.5%		0.8952
97.5%		0.87046
90.0%		0.83647
75.0%	Quartile	0.79693
50.0%	Median	0.7403
25.0%	Quartile	0.67859
10.0%		0.62064
2.5%		0.54647
0.5%		0.48092
0.0%	Minimum	0.41719

Table 7.1 is the list of 9,591 correlations sorted by descending order of R. The [2.5, 97.5%] is the 95% confidence interval of R. Because all p-values are 0.000 ($p < 0.0005$), 9,591 correlations have the positive values. We claim this fact indicates the 139 RipDSs are signals instead of 139 SMs. Thus, everybody can quickly analyze the 139-dimensional signal subspace that consists of 102 subjects and 139 RipDSs. That is, RipDSs with high correlation have the same effect, and small ones have different effects for diagnosis.

Table 7.1 List of 9,591 correlations sorted by descending order of R

Var1	Versus Var2	R	n	2.5%	97.5%	p-value
RIP101	RIP100	1.000	102			0.000
RIP3	RIP2	0.928	102	0.895	0.951	0.000
RIP5	RIP1	0.928	102	0.895	0.951	0.000
RIP15	RIP2	0.925	102	0.890	0.949	0.000
RIP59	RIP35	0.922	102	0.887	0.947	0.000
RIP33	RIP22	0.914	102	0.875	0.941	0.000
RIP12	RIP2	0.913	102	0.873	0.940	0.000
RIP15	RIP3	0.913	102	0.873	0.940	0.000
–	–	–	–	–	–	–
RIP139	RIP49	0.439	102	0.267	0.583	0.000
RIP138	RIP54	0.438	102	0.266	0.583	0.000
RIP139	RIP68	0.436	102	0.263	0.581	0.000
RIP134	RIP114	0.433	102	0.261	0.579	0.000
RIP138	RIP57	0.433	102	0.260	0.579	0.000
RIP139	RIP19	0.431	102	0.258	0.577	0.000
RIP139	RIP27	0.427	102	0.253	0.574	0.000
RIP139	RIP83	0.417	102	0.242	0.566	0.000

The correlation between RIP100 and RIP101 is 1. Although SM100 has 22 genes and SM101 has 27 genes, we guess two basic gene sets (BGSs) included in SM100 and SM101 have the same effects for cancer gene diagnosis. This survey is in future research (Problem7). Figure 7.2 shows the three scatter plots. The x-axis is RIP100, and three y-axes are RIP101, RIP98, and RIP139. Three straight lines are the simple regression lines. Three plots suggest us 100th RipDS (Rip100) and 101th RipDS (RIP101) have the same role in cancer gene diagnosis.

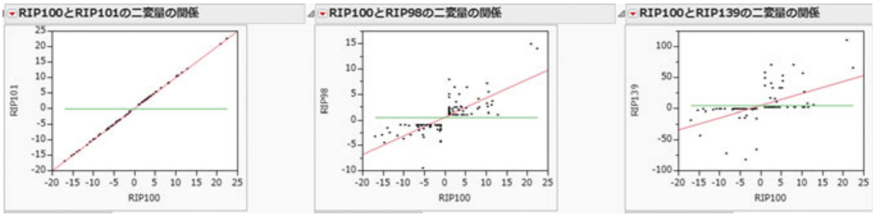


Fig. 7.2 Three scatter plots

Table 7.2 is four RipDSs such as RipDS1 (abbreviated RIP1), RIP83, RIP94, and RIP139 corresponding to SM1, SM83, SM94, and SM139. The second column (SM1) shows the case number sorted by the third column score (RIP1 DS). The 40th case and 51st case have the minimum DS (−3.253) and maximum DS (5.705). The range and RatioSV are 8.958 and 22.326. The 102 subjects are sorted in the ascending order. The minimum DS is −3.253. This table means that the 40th normal subject is the minimum value and all 50 normal subjects are less than equal −1 in SM1. The ten subjects are on the SV = −1, and the 40 subjects are outliers. The range of the normal class is [−3.253, −1]. On the other hand, the range of tumor class is [1, 5.705]. The 13 tumor patients are on SV = 1, and the 39 tumor patients are outliers. The 51st patient of tumor class has the maximum value. The range of 102 subjects is [−3.253, 5.705] and its width is 8.958. Thus, RatioSV of RIP1 is 22.326%. We choose the intermediate RIP 83 and RIP 94, RatioSVs of those are 6.356% and 4.889%. Although the RIP139 has few outliers, it has the minimum RatioSV (1.042%) because the range 191.944 is large. The ranges of normal and tumor classes are [−81.925, −1] and [1, 110.02], respectively. For the normal class, the 14 subjects are outliers, and the 36 subjects are on SV = −1. For tumor class, the 21 patients are outliers, and the 31 patients lie on SV = 1. This table shows that RIP fixes many subjects on two SVs and finds the linear separable facts of two classes. In general, SM with many subjects on SV enlarges the RatioSVs shown in Table 5.2, and SM with many outliers may reduce RatioSV.

Table 7.2 Four RipDSs

	SM1	RIP1	SM83	RIP83	SM94	RIP94	SM139	RIP139
Min/Range	**40**	**8.958**	39	31.467	49	40.981	25	191.944
Max/Ratio	**51**	**22.326**	57	6.356	54	4.880	57	1.042
1	**40**	**−3.253**	**39**	**−14.514**	**49**	**−8.936**	**25**	**−81.925**
2	34	−3.005	20	−11.274	12	−7.508	36	−72.163
3	24	−2.760	38	−9.089	24	−7.313	35	−66.073
4	44	−2.739	8	−9.042	25	−7.290	12	−44.026
5	1	−2.651	16	−8.609	9	−7.037	41	−18.500
6	37	−2.570	26	−8.181	15	−5.503	37	−16.822
7	49	−2.567	6	−7.817	36	−5.299	5	−15.080
8	41	−2.473	43	−7.752	30	−5.297	44	−14.581
9	39	−2.425	19	−6.868	14	−4.573	31	−11.841
10	2	−2.348	23	−6.543	44	−4.264	39	−4.911
11	35	−2.240	40	−6.233	8	−3.943	43	−3.629
12	25	−2.221	33	−6.116	26	−3.853	24	−2.881
13	31	−2.123	46	−5.737	31	−3.765	50	−2.502

(continued)

Table 7.2 (continued)

	SM1	RIP1	SM83	RIP83	SM94	RIP94	SM139	RIP139
14	50	−2.108	41	−4.956	21	−3.575	18	−2.161
15	12	−2.067	24	−4.842	3	−3.555	1	−1
16	13	−2.035	27	−4.800	20	−3.477	2	−1
17	43	−2.024	2	−4.044	22	−3.468	3	−1
18	23	−1.979	22	−3.561	40	−3.138	4	−1
19	5	−1.767	37	−3.075	37	−2.894	6	−1
20	22	−1.743	21	−2.845	39	−2.105	7	−1
21	30	−1.733	12	−2.804	43	−2.097	8	−1
22	21	−1.672	4	−2.476	35	−2.043	9	−1
23	9	−1.667	15	−2.384	17	−1.981	10	−1
24	38	−1.666	5	−2.301	28	−1.966	11	−1
25	33	−1.654	25	−2.273	11	−1.884	13	−1
26	15	−1.623	50	−1.861	32	−1.861	14	−1
27	17	−1.610	17	−1.811	41	−1.808	15	−1
28	11	−1.591	36	−1.540	7	−1.750	16	−1
29	46	−1.579	45	−1.536	27	−1.703	17	−1
30	16	−1.530	30	−1.496	19	−1.691	19	−1
31	42	−1.520	47	−1.473	23	−1.351	20	−1
32	4	−1.510	31	−1.419	2	−1.281	21	−1
33	36	−1.461	1	−1.191	6	−1.208	22	−1
34	20	−1.449	18	−1.156	1	−1.171	23	−1
35	26	−1.360	42	−1.137	48	−1.008	26	−1
36	48	−1.348	7	−1.082	4	−1	27	−1
37	7	−1.235	3	−1	5	−1	28	−1
38	6	−1.218	9	−1	10	−1	29	−1
39	3	−1.194	10	−1	13	−1	30	−1
40	28	−1.117	11	−1	16	−1	32	−1
41	8	−1	13	−1	18	−1	33	−1
42	10	−1	14	−1	29	−1	34	−1
43	14	−1	28	−1	33	−1	38	−1
44	18	−1	29	−1	34	−1	40	−1
45	19	−1	32	−1	38	−1	42	−1
46	27	−1	34	−1	42	−1	45	−1
47	29	−1	35	−1	45	−1	46	−1
48	32	−1	44	−1	46	−1	47	−1

(continued)

Table 7.2 (continued)

	SM1	RIP1	SM83	RIP83	SM94	RIP94	SM139	RIP139
49	45	−1	48	−1	47	−1	48	−1
50	**47**	**−1**	**49**	**−1**	**50**	**−1**	**49**	**−1**
51	**52**	**1**	**54**	**1**	**56**	**1**	**51**	**1**
52	53	1	73	1	58	1	56	1
53	60	1	82	1	67	1	58	1
54	61	1	84	1	71	1	60	1
55	67	1	86	1	74	1	61	1
56	69	1	87	1	75	1	63	1
57	80	1	89	1	84	1	65	1
58	84	1	92	1	85	1	66	1
59	92	1	95	1	87	1	67	1
60	95	1	61	1.104	89	1	68	1
61	96	1	74	1.109	91	1	69	1
62	98	1	71	1.453	92	1	71	1
63	99	1	69	1.640	99	1	72	1
64	66	1.043	101	2.620	68	1.068	73	1
65	56	1.045	81	2.693	60	1.254	74	1
66	58	1.059	68	2.742	90	1.421	79	1
67	94	1.064	58	3.403	73	1.432	80	1
68	54	1.167	56	3.484	79	1.577	81	1
69	93	1.225	70	3.571	95	1.618	83	1
70	86	1.314	64	3.693	80	1.867	84	1
71	64	1.345	91	3.741	102	1.885	86	1
72	81	1.513	66	4.470	72	1.966	87	1
73	63	1.522	97	4.481	76	1.985	89	1
74	68	1.644	62	4.521	63	2.248	90	1
75	78	1.650	79	4.760	78	2.380	92	1
76	71	1.751	72	4.812	66	2.422	93	1
77	87	1.841	80	4.830	62	2.449	94	1
78	89	1.845	98	5.055	70	2.457	95	1
79	65	1.888	100	5.230	86	2.943	96	1
80	79	1.934	93	5.595	83	2.966	99	1
81	100	1.967	76	5.627	77	3.400	101	1
82	73	1.985	65	5.733	53	3.555	85	1.810
83	59	2.055	55	6.796	69	3.675	88	5.064

(continued)

Table 7.2 (continued)

	SM1	RIP1	SM83	RIP83	SM94	RIP94	SM139	RIP139
84	72	2.062	60	6.809	101	3.719	52	8.045
85	82	2.306	99	6.949	65	3.733	77	9.389
86	83	2.579	53	7.114	51	3.861	55	14.985
87	90	2.626	96	7.384	93	3.874	64	15.859
88	77	2.766	83	8.308	81	3.924	78	25.962
89	76	2.861	67	8.334	52	3.986	53	32.307
90	74	3.071	59	8.583	88	5.024	82	32.360
91	88	3.167	75	9.176	82	5.530	75	32.496
92	75	3.199	90	9.349	59	5.825	97	36.619
93	55	3.307	52	10.120	97	5.848	62	40.128
94	102	3.313	85	10.527	64	6.249	59	43.835
95	62	3.636	102	11.118	61	6.308	76	48.871
96	91	3.923	77	11.929	96	7.126	102	52.551
97	85	4.037	78	12.152	55	8.764	98	56.262
98	70	4.108	51	13.223	98	8.961	91	57.771
99	97	4.153	88	13.635	94	9.111	100	65.514
100	101	4.280	63	13.891	100	9.123	70	69.813
101	57	5.383	94	14.906	57	16.879	54	69.938
102	**51**	**5.705**	**57**	**16.953**	**54**	**32.045**	**57**	**110.020**

7.3.2 PCA of Signal Data Made by 139 RipDSs

PCA analyzes the signal data made by 139 RipDSs and outputs the 30 principal components showed in Table 7.3. The eigenvalue of Prin1 is 102.668, and the contribution rate is 73.862%. This fact shows that two separable classes are almost explained by the Prin1. The eigenvalue of Prin2 is 4.742, and the contribution rate is 3.412%. Thus, two principal components explain the 77.274% of total variance.

Table 7.3 PCA of 139 RipDSs

Prin	Eigenvalue	Contribution	Cumulative
1	102.668	73.862	73.862
2	4.742	3.412	77.274
3	1.972	1.418	78.692
4	1.802	1.297	79.989
5	1.532	1.102	81.091
6	1.400	1.007	82.098

(continued)

Table 7.3 (continued)

Prin	Eigenvalue	Contribution	Cumulative
7	1.240	0.892	82.990
8	1.132	0.815	83.805
9	1.053	0.757	84.562
10	0.976	0.702	85.264
11	0.929	0.668	85.933
12	0.896	0.645	86.578
13	0.888	0.639	87.216
14	0.832	0.598	87.815
15	0.775	0.557	88.372
16	0.704	0.506	88.878
17	0.693	0.499	89.377
18	0.683	0.491	89.868
19	0.648	0.466	90.335
20	0.607	0.437	90.771
21	0.581	0.418	91.190
22	0.555	0.399	91.589
23	0.531	0.382	91.971
24	0.500	0.360	92.330
25	0.488	0.351	92.682
26	0.471	0.339	93.020
27	0.440	0.316	93.337
28	0.419	0.301	93.638
29	0.403	0.290	93.928
30	0.392	0.282	94.210

Figure 7.3 is eight scatter plots. All x-axes are Prin1. Upper y-axes are from Prin2 to Prin5, and lower y-axes are from Prin27 to Prin30. A left ellipse is the 99% confidence ellipse of the normal class, and the right ellipse is the 99% confidence ellipse of the tumor class. We confirm the 138 scatter diagrams are almost the same results as Prin1 in Fig. 7.4 that two classes are entirely separated on the Prin1. Individual RipDS divides the two groups with SV = 1 and −1. However, because PCA analyzes 139 RipDSs at the same time, the values of two SVs are not at 1 and −1 in Fig. 7.4. The critical fact here shows that two classes are entirely separable in all scatter plots of signal data. In this fact, we recognize that a signal data is an actual signal instead of the genes included in SM or BGS.

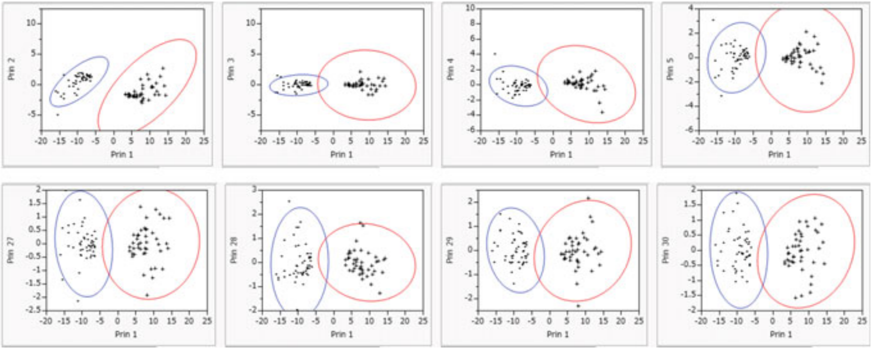


Fig. 7.3 Eight scatter plots (x-axis: Prin1; upper y-axis: from Prin2 to Prin5; lower y-axes: from Prin27 to Prin30)

Figure 7.4 is PCA output of the 139 RipDSs. The eigenvalue of Prin1 is 102.669, and the contribution rate is 73.9%. The eigenvalue of Prin2 is 4.742, and the contribution rate is 3.41%. The cumulative rate is 77.37%. Thus, we consider the Prin1 is the malignancy index of PCA. Because 139 correlations of 139 RipDSs and Prin1 are higher than 0.7, Prin1 represents all RipDSs well. On the other hand, because 139 correlations of 139 RipDSs and Prin2 are range from −0.25 to 0.5, Prin2 shows that 139 RipDSs belong to two groups with positive and negative correlation.

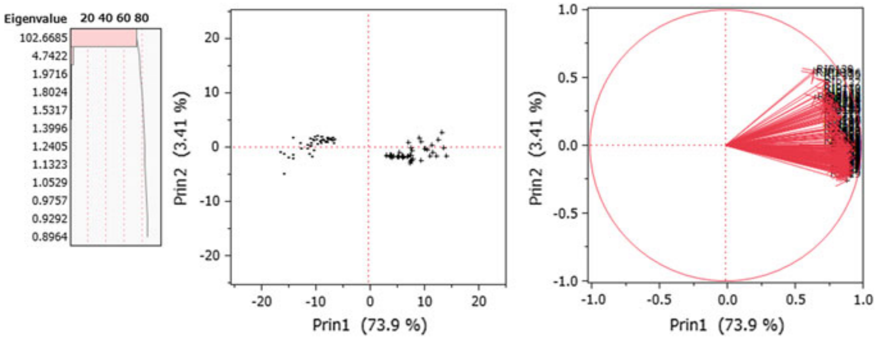


Fig. 7.4 PCA output of the signal data of 139 RipDSs

7.3.3 How to Categorize 139 RipDSs

Our claim is the 139 RipDSs (139 malignancy indexes) open the door of cancer gene diagnosis. We believe that these malignant indicators will belong to several categories having different roles, but are currently unresolved. Therefore, we examined the mechanism that many RipDSs appear. We survey how to build 139 RipDSs in this section. Table 7.4 shows 139 rows from SM1 to SM139. Each row shows the minimum and maximum identification number of 102 cases in the Normal and Tumor columns. Range and RatioSV columns are the range and RatioSV of each RipDS. In SM1, the 40th normal subject takes the minimum value among 50 Normal subjects, and 51st tumor subject takes the maximum value among 52 Tumor patients. The selected two subjects are considered to be in the most normal state and the worst cancer state. The RatioSV is 22.326%.

Table 7.4 Minimum and maximum subject's SM and its RatioSV

SM	Normal	Tumor	Range	RatioSV
SM1	40	51	8.958	22.326
SM2	41	94	7.333	27.274
SM3	34	91	6.951	28.775
SM4	41	91	9.163	21.828
SM5	36	101	7.423	26.943
SM6	18	54	7.277	27.484
SM7	42	91	8.631	23.173
SM8	41	51	10.635	18.806
SM9	41	91	11.799	16.951
SM10	1	94	8.469	23.616
SM11	37	100	8.354	23.941
SM12	13	51	7.454	26.833
SM13	37	55	9.071	22.049
SM14	37	62	6.830	29.284
SM15	37	70	6.965	28.714
SM16	41	51	10.872	18.396
SM17	2	62	8.977	22.280
SM18	1	91	9.162	21.829
SM19	29	62	9.715	20.586
SM20	37	100	13.459	14.860
SM21	39	51	9.288	21.534
SM22	20	94	10.256	19.501
SM23	37	91	10.863	18.411

(continued)

Table 7.4 (continued)

SM	Normal	Tumor	Range	RatioSV
SM24	37	97	9.508	21.034
SM25	20	77	11.089	18.035
SM26	36	85	9.986	20.028
SM27	37	86	14.248	14.038
SM28	24	62	8.352	23.947
SM29	20	51	9.959	20.083
SM30	16	51	10.645	18.789
SM31	34	101	11.384	17.569
SM32	1	85	10.936	18.288
SM33	29	62	8.435	23.711
SM34	37	89	12.306	16.252
SM35	20	57	10.273	19.468
SM36	37	100	14.953	13.376
SM37	34	51	12.051	16.596
SM38	37	94	11.286	17.720
SM39	25	100	16.496	12.124
SM40	16	62	12.945	15.450
SM41	30	102	12.948	15.446
SM42	30	51	9.974	20.052
SM43	20	51	12.146	16.466
SM44	45	100	9.234	21.659
SM45	31	81	9.802	20.405
SM46	25	62	13.749	14.547
SM47	16	94	14.992	13.341
SM48	14	88	9.847	20.310
SM49	20	94	18.766	10.658
SM50	20	62	14.668	13.635
SM51	16	102	11.810	16.935
SM52	20	91	9.601	20.831
SM53	20	62	12.122	16.499
SM54	21	91	11.750	17.021
SM55	18	51	23.322	8.575
SM56	14	70	11.533	17.341
SM57	16	88	17.090	11.703
SM58	25	94	12.931	15.467

(continued)

Table 7.4 (continued)

SM	Normal	Tumor	Range	RatioSV
SM59	20	57	10.105	19.792
SM60	19	52	8.717	22.944
SM61	16	91	17.734	11.278
SM62	33	97	12.742	15.697
SM63	25	57	20.723	9.651
SM64	34	70	12.041	16.611
SM65	37	51	16.339	12.240
SM66	29	54	10.960	18.248
SM67	37	97	10.792	18.532
SM68	42	51	11.766	16.998
SM69	14	54	17.347	11.530
SM70	16	62	16.450	12.158
SM71	25	51	12.841	15.575
SM72	16	94	25.266	7.916
SM73	25	94	15.480	12.920
SM74	16	102	13.553	14.757
SM75	25	62	23.707	8.436
SM76	37	52	15.506	12.898
SM77	37	51	13.805	14.488
SM78	20	85	16.204	12.342
SM79	30	85	17.944	11.146
SM80	36	51	24.447	8.181
SM81	20	57	14.796	13.517
SM82	44	100	14.990	13.342
SM83	39	57	31.467	6.356
SM84	16	51	21.265	9.405
SM85	41	70	13.738	14.558
SM86	21	64	16.417	12.183
SM87	37	62	20.704	9.660
SM88	44	51	26.679	7.497
SM89	20	57	27.688	7.223
SM90	37	51	25.639	7.800
SM91	20	97	15.673	12.761
SM92	20	62	20.835	9.599
SM93	41	59	20.661	9.680

(continued)

Table 7.4 (continued)

SM	Normal	Tumor	Range	RatioSV
SM94	49	54	40.981	4.880
SM95	20	57	31.692	6.311
SM96	37	85	21.997	9.092
SM97	16	55	13.090	15.279
SM98	20	57	24.523	8.156
SM99	41	91	51.249	3.902
SM100	41	100	39.350	5.083
SM101	41	100	39.350	5.083
SM102	20	51	55.426	3.608
SM103	41	85	25.523	7.836
SM104	36	57	31.126	6.426
SM105	20	62	29.604	6.756
SM106	37	54	48.447	4.128
SM107	16	97	32.532	6.148
SM108	36	100	43.070	4.644
SM109	44	51	30.312	6.598
SM110	41	57	47.119	4.245
SM111	25	51	50.451	3.964
SM112	36	57	60.278	3.318
SM113	36	54	36.942	5.414
SM114	25	70	28.812	6.941
SM115	36	100	35.950	5.563
SM116	41	57	33.926	5.895
SM117	18	57	47.592	4.202
SM118	44	57	28.357	7.053
SM119	41	57	25.266	7.916
SM120	36	57	29.406	6.801
SM121	37	57	34.512	5.795
SM122	5	57	70.984	2.818
SM123	20	54	37.973	5.267
SM124	36	88	30.050	6.656
SM125	36	54	43.175	4.632
SM126	37	100	49.543	4.037

(continued)

Table 7.4 (continued)

SM	Normal	Tumor	Range	RatioSV
SM127	41	57	54.947	3.640
SM128	34	57	30.673	6.520
SM1209	37	94	37.315	5.360
SM130	37	57	61.048	3.276
SM131	41	57	36.844	5.428
SM132	20	91	64.680	3.092
SM133	37	100	53.502	3.738
SM134	35	100	76.348	2.620
SM135	36	57	130.324	1.535
SM136	37	100	77.227	2.590
SM137	36	91	53.635	3.729
SM138	5	100	85.988	2.326
SM139	25	57	191.944	1.042

Table 7.5 is sorted in descending order by the second column value (Normal) that takes the minimum value in Table 7.4. In SM32, the first subject of the normal class takes the minimum value, and the 85th subject takes the maximum value. Because there is no other pair of the same patients, the pair column is blank. In SM30 and SM84, because the 16th subject takes the minimum value and the 51st patient takes the maximum values, the pair number is two, and the correlation of RIP30 and RIP84 is 0.822. In this way, we focus on the RipDSs having the pair. Because it is difficult to examine the similarity of 139 RipDSs with 9,591 correlation coefficients, we consider as one possibility. The 20 two pairs choose the same minimum and maximum subjects; 3 three pairs choose the same three minimum and maximum subjects, 2 four pairs choose the four same minimum and maximum subjects, 1 five pairs chooses the five same minimum and maximum subjects, and 2 six pairs choose the six minimum and maximum subjects. We expect these segmentations are useful to categorize SMs and RipDSs. The other 65 SMs select the different minimum patient and maximum patient and may indicate cancer diversity. However, if we check the rank correlations of 139 RipDSs, we may obtain better results because those evaluate all subjects. The last three columns are the correlation coefficients of the two RipDSs. The range of correlations is [0.572, 0.922]. Whether SMs of each pair have the same role as each other is a future research theme. Although SM100 and SM101 contain different gene sets, RIP 100 and RIP 101 are considered to provide the same information as malignant tumor indices because R = 1.

Table 7.5 Sorted in descending order of the second column (Normal1) and the seventh column (Tumor2)

SM	Normal1	Tumor1	Pair1	Correlation		
SM32	1	85				
SM18	1	91				
SM10	1	94				
SM17	2	62				
SM122	5	57				
SM138	5	100				
SM12	13	51				
SM69	14	54				
SM56	14	70				
SM48	14	88				
SM30	16	51	2	0.822		
SM84	16	51				
SM97	16	55				
SM40	16	62	2	0.832		
SM70	16	62				
SM57	16	88				
SM61	16	91				
SM47	16	94	2	0.739		
SM72	16	94				
SM107	16	97				
SM51	16	102	2	0.824		
SM74	16	102				
SM55	18	51				
SM6	18	54				
SM117	18	57				
SM60	19	52				
SM29	20	51	3	0.834		
SM43	20	51		0.737		
SM102	20	51		0.659		
SM123	20	54				
SM35	20	57	6	0.922	0.849	0.697
SM59	20	57		0.860	0.778	0.839
SM81	20	57		0.808	0.790	0.674
SM89	20	57		0.771	0.775	
SM95	20	57		0.808	0.754	
SM98	20	57		0.800	0.737	
SM50	20	62	4	0.812	0.756	

(continued)

Table 7.5 (continued)

SM	Normal1	Tumor1	Pair1	Correlation		
SM53	20	62		0.766	0.718	
SM92	20	62		0.737		
SM105	20	62		0.838		
SM25	20	77				
SM78	20	85				
SM52	20	91	2	0.714		
SM132	20	91				
SM22	20	94	2	0.828		
SM49	20	94				
SM91	20	97				
SM86	21	64				
SM54	21	91				
SM28	24	62				
SM71	25	51	2	0.785		
SM111	25	51				
SM63	25	57	2	0.584		
SM139	25	57				
SM46	25	62	2	0.801		
SM75	25	62				
SM114	25	70				
SM58	25	94	2	0.879		
SM73	25	94				
SM39	25	100				
SM66	29	54				
SM19	29	62	2	0.872		
SM33	29	62				
SM42	30	51				
SM79	30	85				
SM41	30	102				
SM45	31	81				
SM62	33	97				
SM37	34	51				
SM128	34	57				
SM64	34	70				
SM3	34	91				
SM31	34	101				
SM134	35	100				

(continued)

Table 7.5 (continued)

SM	Normal1	Tumor1	Pair1	Correlation		
SM80	36	51				
SM113	36	54	2	0.754		
SM125	36	54				
SM104	36	57	4	0.714	0.769	
SM112	36	57		0.759	0.805	
SM120	36	57		0.728		
SM135	36	57		0.749		
SM26	36	85				
SM124	36	88				
SM137	36	91				
SM108	36	100	2	0.710		
SM115	36	100				
SM5	36	101				
SM65	37	51	3	0.831		
SM77	37	51		0.791		
SM90	37	51		0.808		
SM76	37	52				
SM106	37	54				
SM13	37	55				
SM121	37	57	2	0.778		
SM130	37	57				
SM14	37	62	2	0.822		
SM87	37	62				
SM15	37	70				
SM96	37	85				
SM27	37	86				
SM34	37	89				
SM23	37	91				
SM38	37	94	2	0.733		
SM129	37	94				
SM24	37	97	2	0.848		
SM67	37	97				
SM11	37	100	6	0.859	0.746	0.72781
SM20	37	100		0.835	0.727	0.77163
SM36	37	100		0.687	0.676	0.67226
SM126	37	100		0.722	0.688	
SM133	37	100		0.601	0.664	

(continued)

Table 7.5 (continued)

SM	Normal1	Tumor1	Pair1	Correlation		
SM136	37	100		0.835	0.572	
SM21	39	51				
SM83	39	57				
SM1	40	51				
SM8	41	51	2	0.877		
SM16	41	51				
SM110	41	57	5	0.731	0.737	
SM116	41	57		0.752	0.693	
SM119	41	57		0.715	0.732	
SM127	41	57		0.715	0.713	
SM131	41	57		0.788	0.721	
SM93	41	59				
SM85	41	70				
SM103	41	85				
SM4	41	91	3	0.852		
SM9	41	91		0.721		
SM99	41	91		0.781		
SM2	41	94				
SM100	41	100	2	1		
SM101	41	100				
SM68	42	51				
SM7	42	91				
SM88	44	51	2	0.761		
SM109	44	51				
SM118	44	57				
SM82	44	100				
SM44	45	100				
SM94	49	54				

7.4 Analysis of 139 SMs of Singh et al. Microarray (2018)

LINGO Program3 found the Singh's microarray consists of 179 SMs (1,238 genes) in 2015. However, we obtain 139 SMs (4,046 genes) in 2018. We obtain fewer SMs and more genes in 2018. A yearly update of LINGO may cause these differences. We analyze the signal data and obtain almost the same results introduced in Chap. 2. We think that these facts are suitable for considering signal data as a signal rather than SM.

7.4.1 Validation of 139 SMs by Six MP-Based LDFs and Discriminant Functions

Table 7.6 shows the 139 SMs from SM = 1 to SM = 139. Program3 determines this order of SM. The "Gene" column is the number of genes of each SM. The range of genes included in the 139 SMs is [20, 66]. The average is 29.1. Row "SUM" indicates 139 SMs which contain 4,046 genes. From RIP column to H-SVM column show three RatioSVs of 139 SMs. Three ranges of RatioSV are [1.04, 29.11], [1.06, 32], and [1.07, 29.48], respectively. Three averages of RatioSVs are 13.2%, 13%, and 13.32%, respectively. Row "SUM" indicates the number of the maximum RatioSVs of 139 SMs those are 37, 33, and 55, respectively. To summarize these results, the range, average, and maximum number of H-SVM are slightly better than RIP because the maximization SV of H-SVM work well. Two columns "MAX and MIN" are the maximum and minimum values of six LDFs including Revised IPLP-OLDF, SVM4 and SVM1. Because all NMs of logistic regression and SVM4 are zero and 139 SMs are linearly separable, we omit two columns from the table. Three columns "SVM1, LDF2, and QDF" show the NM. SVM1 cannot discriminate 113 SMs correctly. LDF2 cannot discriminate 92 SMs correctly. The 26 NMs of QDF are not zero.

Table 7.6 Summary of four RatioSVs of four LDFs and NMs of other discriminant functions

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
1	22	22.96	25.22	22.88	25.22	21.62	0	0	0
2	24	**29.11**	28.58	26.42	29.11	26.42	0	0	0
3	24	**28.77**	18.75	26.66	28.77	18.75	0	0	0
4	27	**22.36**	20.37	21.06	22.36	20.37	0	0	0
5	30	**26.94**	24.84	24.80	26.94	19.05	0	0	0
6	28	27.48	**32.00**	29.48	32.00	27.48	0	0	0
7	29	**26.33**	25.02	25.04	26.33	24.94	0	0	0
8	22	19.82	19.14	**20.42**	20.42	19.14	1	1	0
9	26	17.31	**20.18**	18.57	20.18	17.31	1	0	0
10	30	**23.62**	216.10	23.23	23.62	18.29	0	0	0
11	26	24.81	**26.50**	24.17	26.50	24.17	0	0	0
12	28	26.83	26.89	26.74	27.47	26.74	0	0	0
13	25	**22.94**	22.19	21.33	22.94	20.76	1	0	0
14	26	26.88	25.79	**27.44**	27.44	25.79	0	0	0
15	32	25.67	**27.03**	26.56	27.03	25.67	0	0	0
16	27	21.61	21.12	21.92	21.93	21.12	0	0	0
17	27	**23.28**	21.88	23.27	23.28	20.83	0	0	0
18	24	21.83	21.45	21.69	22.58	21.45	0	0	0

(continued)

Table 7.6 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
19	21	20.59	**22.69**	21.50	22.69	18.81	3	0	1
20	24	16.01	15.81	**16.15**	16.15	15.81	4	0	1
21	26	22.83	19.67	**22.98**	22.98	19.67	0	0	0
22	30	19.50	19.20	**203.2**	203.2	18.99	0	1	0
23	23	**21.35**	18.10	18.95	21.35	18.10	1	1	0
24	24	**21.65**	21.00	20.92	21.65	19.33	0	1	0
25	25	18.04	17.96	18.01	19.70	17.96	0	0	0
26	27	20.03	**20.85**	20.51	20.85	20.03	1	1	0
27	20	**16.88**	11.89	16.07	16.88	11.89	2	0	0
28	31	**23.95**	21.98	21.58	23.95	19.06	0	0	0
29	29	20.08	**24.10**	23.03	24.10	20.08	0	0	0
30	27	**18.79**	17.71	16.33	18.79	16.33	0	0	0
31	32	17.57	16.57	**18.60**	18.60	16.57	1	0	0
32	25	18.29	**19.06**	18.38	19.06	17.93	1	0	0
33	27	23.71	23.27	23.02	24.63	23.02	0	0	0
34	28	**19.77**	16.60	19.56	19.77	16.60	5	0	0
35	28	19.47	213.2	**21.47**	21.47	18.57	0	0	0
36	20	13.61	13.62	**13.69**	13.69	13.61	2	3	0
37	28	16.60	15.58	**17.32**	17.32	15.58	3	0	0
38	24	**18.48**	15.95	18.28	18.48	15.95	3	0	0
39	26	12.12	13.77	**13.87**	13.87	12.12	0	0	1
40	25	**15.45**	14.83	15.19	15.45	14.18	2	0	0
41	25	15.45	14.98	**17.07**	17.07	14.82	2	1	0
42	28	20.05	**21.20**	21.02	21.20	223	2	1	0
43	27	15.96	15.48	16.10	16.32	15.48	4	1	1
44	26	21.92	20.73	**22.04**	22.04	20.44	4	0	0
45	29	20.40	20.66	21.01	21.73	20.40	4	1	0
46	23	14.55	14.16	**14.63**	14.63	14.08	1	1	0
47	25	13.34	**14.60**	13.92	14.60	13.34	4	2	0
48	24	**20.31**	19.97	19.20	20.31	18.48	3	0	0
49	23	10.66	10.42	**11.14**	11.14	10.31	5	4	1
50	29	13.64	225	**16.92**	16.92	13.64	4	1	0
51	25	227	**16.63**	15.72	16.63	15.72	2	6	0
52	28	20.83	20.89	21.54	21.54	20.83	0	0	0
53	28	16.50	14.52	**16.52**	16.52	14.52	5	2	0
54	30	17.02	13.36	**17.57**	17.57	13.36	1	1	1
55	23	8.58	8.42	**8.59**	8.59	8.42	7	4	0

(continued)

Table 7.6 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
56	28	17.34	15.81	**17.73**	17.73	13.66	4	2	1
57	23	9.76	9.30	**10.23**	10.23	9.30	6	2	0
58	27	15.47	**16.72**	14.88	16.72	14.88	2	0	0
59	28	19.79	20.66	**21.71**	21.71	19.79	4	0	0
60	28	22.94	**23.67**	23.29	23.67	20.92	1	0	0
61	28	11.28	10.86	11.14	11.73	10.86	7	2	0
62	26	15.70	15.26	**16.50**	16.50	15.26	7	0	0
63	23	9.65	9.54	**9.73**	9.73	9.12	5	3	2
64	24	**16.61**	16.13	15.83	16.61	15.83	4	1	0
65	22	14.19	14.23	**14.24**	14.24	14.19	8	3	0
66	31	**18.25**	17.16	17.62	18.25	17.16	3	1	1
67	32	18.56	18.02	18.80	20.01	18.02	4	0	0
68	31	18.59	18.49	18.18	18.68	18.18	7	3	0
69	28	11.53	13.26	12.68	12.71	11.53	8	0	0
70	27	6.16	12.96	**15.04**	15.04	11.38	7	0	0
71	27	15.58	15.89	**15.98**	15.98	15.57	6	1	0
72	24	7.92	8.12	8.10	8.12	7.92	7	4	0
73	27	12.92	12.56	**13.33**	13.33	12.56	4	1	0
74	30	14.76	**14.94**	14.68	14.94	12.05	2	2	0
75	25	8.44	83.2	**8.63**	8.63	83.2	10	4	0
76	33	12.50	12.84	**13.09**	13.09	12.50	4	2	0
77	26	**15.74**	15.21	15.66	15.74	15.21	4	2	0
78	29	12.34	13.00	13.28	13.48	12.34	5	2	0
79	24	11.15	10.81	12.09	12.48	10.81	10	2	0
80	25	8.18	**8.35**	**8.35**	8.35	8.18	10	3	0
81	27	13.52	13.59	**14.68**	14.68	13.52	9	0	0
82	31	14.23	13.07	**15.95**	15.95	13.07	5	1	0
83	22	**7.09**	6.95	6.77	7.09	6.77	8	6	2
84	25	9.41	**10.66**	10.65	10.66	9.41	15	2	1
85	27	15.52	14.96	**15.71**	15.71	14.96	9	0	0
86	29	12.18	12.38	**12.68**	12.68	12.18	11	2	0
87	30	10.57	10.77	**12.75**	12.75	10.57	8	1	0
88	25	7.66	**8.04**	8.02	8.04	7.66	12	5	0
89	27	**73.2**	7.18	7.18	73.2	7.18	8	7	1
90	28	9.55	9.43	9.83	9.84	9.43	11	5	0
91	27	12.76	**13.09**	12.37	13.09	12.37	10	1	0
92	33	**9.60**	8.33	9.38	9.60	8.33	10	5	0
93	28	**10.00**	9.38	9.87	10.00	8.86	11	4	1

(continued)

Table 7.6 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
94	27	5.06	4.64	**5.16**	5.16	4.64	14	8	0
95	30	6.31	**7.36**	6.97	7.36	5.59	15	4	1
96	25	9.13	9.25	**9.30**	9.30	9.00	13	4	1
97	34	15.28	13.36	**15.39**	15.39	13.36	10	1	0
98	33	8.16	7.76	8.01	8.20	7.76	7	5	0
99	26	4.19	**4.41**	4.30	4.41	4.19	12	11	0
100	22	4.19	**4.41**	4.30	4.41	4.19	12	6	4
101	27	**5.37**	5.24	5.28	5.37	5.24	14	5	0
102	28	**3.61**	3.44	3.60	3.61	3.31	16	9	1
103	27	**10.43**	10.38	10.25	10.43	8.89	12	4	0
104	24	6.43	**6.44**	6.37	6.44	6.37	15	6	0
105	29	**6.76**	6.55	6.64	6.76	6.55	12	4	0
106	26	4.48	**4.50**	**4.50**	4.50	4.47	17	8	0
107	26	6.15	7.27	7.09	7.47	6.15	13	6	1
108	23	4.64	4.57	**4.85**	4.85	4.27	15	9	2
109	27	7.40	7.43	**7.65**	7.65	7.40	12	7	1
110	28	4.35	4.35	4.35	4.43	4.35	15	8	1
111	25	3.96	3.95	**4.05**	4.05	3.95	11	11	0
112	29	3.32	3.33	3.33	3.34	3.32	16	11	2
113	36	5.41	5.43	5.37	5.44	5.37	13	6	0
114	31	6.94	7.00	**7.31**	7.31	6.94	19	7	0
115	32	**5.56**	5.50	**5.56**	5.56	5.47	15	10	1
116	32	6.21	6.01	**6.23**	6.23	6.01	18	7	0
117	27	**4.20**	**4.20**	**4.20**	4.20	4.11	14	8	1
118	33	7.55	**7.65**	7.21	7.65	7.21	18	8	0
119	42	**8.60**	7.43	7.18	8.60	7.14	11	6	0
120	35	6.80	6.83	**6.85**	6.85	6.57	15	6	0
121	29	**5.45**	5.14	5.45	5.45	5.14	17	7	2
122	29	**2.82**	**2.82**	2.81	2.82	2.81	11	12	3
123	36	5.27	**5.59**	5.50	5.59	5.25	11	10	0
124	42	6.66	**6.91**	6.61	6.91	6.61	16	10	0
125	29	**4.63**	4.56	**4.63**	4.63	4.56	16	9	0
126	36	4.18	**4.19**	4.18	4.19	4.17	17	16	0
127	29	3.68	**3.71**	**3.71**	3.71	3.68	13	12	0
128	48	**6.52**	6.47	6.12	6.52	6.12	16	13	0

(continued)

Table 7.6 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2	QDF
129	42	**6.20**	**6.20**	6.19	6.20	6.19	14	11	0
130	37	3.52	**3.57**	3.56	3.57	3.41	15	11	0
131	51	5.03	**5.08**	4.98	5.08	4.98	19	15	0
132	40	3.09	3.08	**3.17**	3.17	3.08	21	16	0
133	36	4.45	4.47	**4.49**	4.49	4.45	15	11	0
134	44	**2.62**	2.58	2.60	2.62	2.58	20	13	0
135	38	1.53	1.53	**1.54**	1.54	1.53	19	15	0
136	51	3.09	3.09	**3.14**	3.14	3.04	15	14	0
137	48	3.73	3.72	3.66	3.81	3.66	15	17	0
138	56	2.33	2.14	2.38	2.52	2.14	16	13	0
139	66	1.04	1.06	**1.07**	1.07	1.04	19	13	0
Max	66	29.11	32.00	29.48	32.00	27.48	21	17	4
Min	20	1.04	1.06	1.07	1.07	1.04	0	0	0
Mean	29.1	13.20	13.00	13.32	13.73	12.35	7.37	3.85	0.26
Sum	4046	37	33	**55**			113	93	26

7.4.2 Analysis of Signal Data Using 139 SMs Found by RIP

Because we cannot obtain useful results of 139 SMs, we analyze three signal data made by 139 RipDSs, 139 LpDSs, and 139 HsvmDSs using 139 SMs found by RIP. We get the following surprising success as same as the other five microarrays.

7.4.2.1 Ward Cluster Analysis and PCA of Signal Data Made by RipDSs

Figure 7.5 is a Ward cluster analysis of RipDSs signal data. If Ward cluster analyzes 139 SMs individually, they cannot show the clear clusters. However, the upper green part is 50 normal subjects, and the lower red part is 52 cancer patients. We consider the remarkable effects of RipDSs cause this surprising result. In the dendrogram of 50 normal subjects shown on the right side, those become two clusters with 44 and 6 cases from the top. Singh's microarray shows that normal class has two clusters. Surely, there will be some medical implication. The 52 prostate cancers consist of three clusters of 10, 3, and 39 patients. We expect that these clusters will have medical implications easily. Variable dendrogram may be analyzed by relating the results of case dendrogram with mosaic charts. Our approach is more simple and reliable than other approaches, as it is an analysis of the signal data.

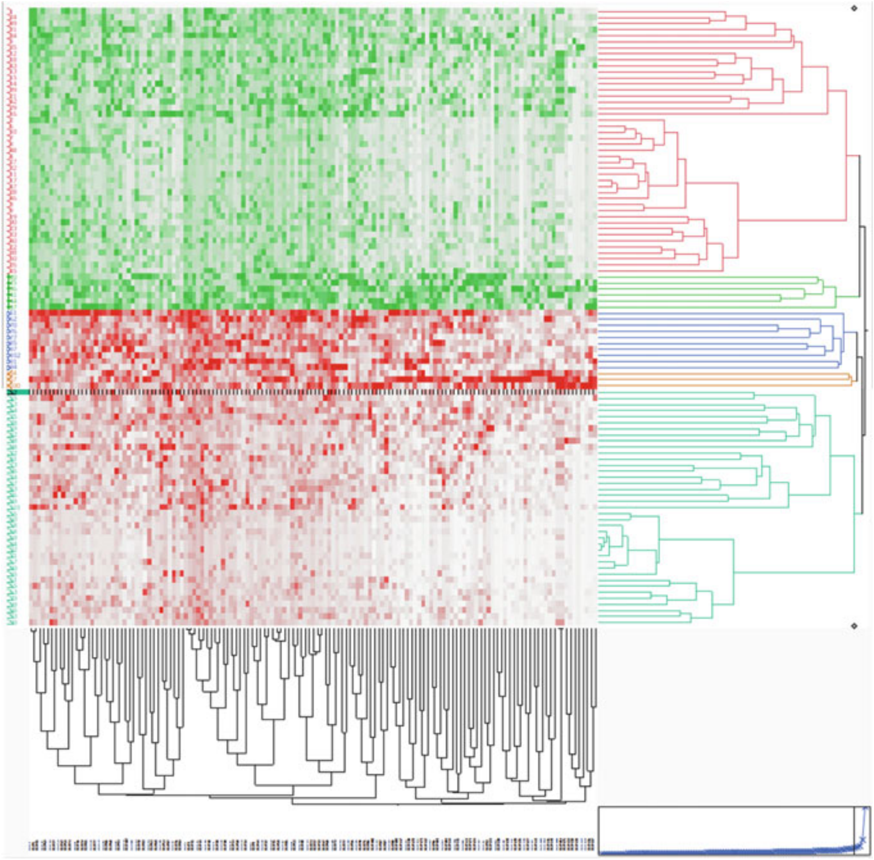


Fig. 7.5 Cluster analysis of signal data by RipDSs

Figure 7.6 shows the result of RipDSs signal data by PCA. The first eigenvalue is 102.668, and the contribution ratio is 73.862%. The second eigenvalue is 4.742, the contribution ratio is 3.412, and the cumulative contribution ratio is 77.274%. That is, the Prin1 almost represents 102 subjects. From the scatter plot, because the second eigenvalue is small and the variation is small, it is understood that the normal subjects are almost placed on the Prin1 axis of −6.32 or less in Table 7.7. Cluster analysis shows that six green normal subjects are adjacent to 10 blue tumor patients. However, PCA shows that between these clusters, there are 44 normal subjects and 39 cancer patients. Thus, caution is required for cluster analysis. The range of cancer patients is [3.28, 22]. Three brown patients such as 57th, 100th, and 54th cases are significant outliers. We recommend physicians survey the green and brown cases.

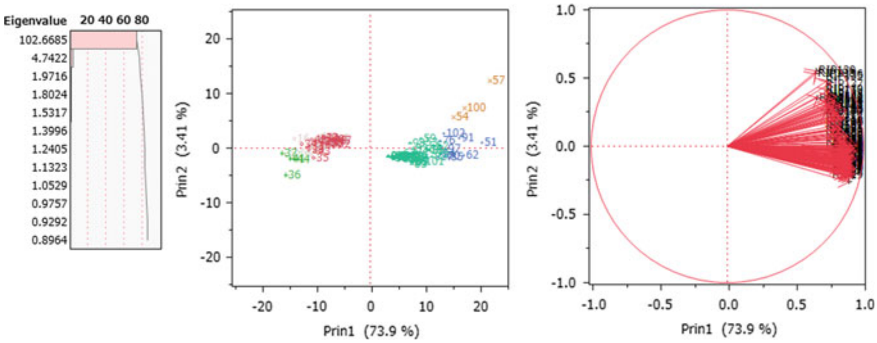


Fig. 7.6 Three plots of PCA (RipDS signal data)

The first columns and second columns of Table 7.7 show the case number corresponding "SM" in Table 7.2 and its value of Prin1 axis. Upper 50 rows are corresponding to the normal class, and lower 52 rows are corresponding to cancer class in Fig. 7.6. The range of normal class is [−16.10, −6.32], and the range of cancer class is [3.28, 22]. SV opens the window having the width (−6.32, 3.28). The RatioSV for PCA in (7.1) is 25.2%.

$$\text{RatioSV of PCA} = (6.32 + 3.28)/(16.10 + 22) * 100 = 960/38.1 = 25.2\% \tag{7.1}$$

Although this is the overall characteristic value of RatioSV of 139 RIP, it is smaller than the maximum value of RatioSV of 139 RIPs 29.11. In later, we conclude the same results of both RaioSV of PCA by Revised LP-OLDF and HSVM. Because we cannot explain this reason, it is future work.

Table 7.7 Prin1 values of three LDFs sorted by each Prin1

RIP	Prin1	LP	Prin1	HSVM	Prin1
37	−16.10	37	−16.00	37	−16.09
36	−15.48	20	−15.61	36	−15.60
20	−15.46	36	−15.49	20	−15.19
41	−14.63	41	−14.75	41	−14.97
16	−13.83	25	−14.03	16	−14.09
25	−13.78	16	−13.70	25	−14.08
44	−13.77	44	−13.59	44	−13.71
34	−1.16	34	−12.47	12	−12.56
12	−12.30	12	−12.31	34	−12.40
31	−11.14	39	−11.10	42	−11.41

(continued)

Table 7.7 (continued)

RIP	Prin1	LP	Prin1	HSVM	Prin1
39	−11.12	42	−10.94	39	−11.23
5	−10.77	31	−10.92	31	−10.96
42	−10.65	5	−10.75	5	−10.81
18	−10.63	18	−10.49	21	−10.61
29	−10.51	29	−10.45	18	−10.58
21	−10.47	24	−10.37	29	−10.56
24	−10.45	1	−10.35	24	−10.45
35	−10.34	21	−10.31	1	−10.33
1	−10.21	35	−10.28	35	−10.16
49	−10.03	43	−9.91	49	−10.07
43	−9.95	49	−9.88	43	−9.94
13	−9.84	13	−9.82	13	−9.93
15	−9.54	26	−9.54	26	−9.49
14	−9.36	15	−9.53	15	−9.46
26	−9.34	14	−9.33	14	−9.32
30	−8.79	30	−8.75	30	−8.89
40	−8.74	40	−8.66	45	−8.72
19	−8.66	33	−8.46	40	−8.60
45	−8.61	45	−8.46	23	−8.45
33	−8.47	23	−8.44	33	−8.44
23	−8.38	19	−8.28	19	−8.25
22	−8.16	22	−8.21	22	−8.20
50	−7.87	2	−7.96	2	−8.05
3	−7.78	3	−7.84	3	−7.93
38	−7.69	50	−7.72	46	−7.85
46	−7.51	46	−7.68	38	−7.79
7	−7.44	38	−7.63	50	−7.61
2	−7.44	7	−7.52	7	−7.38
9	−7.31	27	−7.40	8	−7.37
27	−7.23	8	−7.37	27	−7.30
8	−7.18	9	−7.08	9	−7.18
48	−6.94	48	−6.91	11	−6.88
11	−6.88	10	−6.88	10	−6.88
10	−6.80	11	−6.71	48	−6.87
6	−6.70	6	−6.62	6	−6.72
17	−6.57	4	−6.53	4	−6.45

(continued)

Table 7.7 (continued)

RIP	Prin1	LP	Prin1	HSVM	Prin1
28	−6.49	28	−6.34	32	−6.44
4	−6.48	17	−6.24	17	−6.40
32	−6.41	47	−6.16	28	−6.25
47	**−6.32**	**32**	**−6.15**	**47**	**−6.16**
84	**3.28**	**92**	**3.15**	**84**	**3.20**
92	3.31	84	3.24	92	3.27
69	3.97	58	3.78	69	3.88
58	4.08	69	3.90	58	3.90
80	4.10	80	4.12	80	4.29
67	4.39	71	4.24	67	4.30
71	4.40	67	4.25	95	4.30
95	4.41	95	4.32	71	4.38
68	4.65	68	4.61	68	4.65
60	5.26	60	5.14	60	5.11
61	5.54	63	5.39	63	5.25
63	5.60	61	5.42	61	5.51
56	5.62	99	5.55	56	5.76
99	5.70	56	5.65	99	5.77
79	6.20	65	6.01	65	5.95
65	6.26	93	6.29	93	6.16
90	6.28	79	6.30	90	6.21
93	6.43	90	6.62	79	6.25
66	7.21	66	7.21	66	7.36
73	7.25	98	7.28	73	7.40
98	7.36	73	7.34	96	7.55
86	7.61	72	7.67	89	7.65
96	7.63	96	7.69	86	7.65
83	7.69	86	7.69	98	7.75
74	7.78	89	7.78	87	7.87
89	7.85	82	7.79	83	7.88
82	7.91	83	7.88	81	7.96
72	7.94	81	8.04	82	8.05
87	7.95	74	8.05	72	8.07
81	8.11	87	8.16	74	8.16
59	9.27	59	9.02	59	9.16

(continued)

Table 7.7 (continued)

RIP	Prin1	LP	Prin1	HSVM	Prin1
53	9.64	101	9.74	53	9.60
101	9.71	53	9.75	64	10.14
64	10.37	64	10.16	101	10.29
78	10.50	78	10.36	78	10.36
55	10.82	55	10.89	55	11.14
88	11.34	88	11.27	88	11.50
52	122	52	11.74	52	122
75	11.90	75	11.75	75	11.99
77	12.42	77	12.61	77	12.70
76	12.83	76	12.84	94	13.00
94	12.91	94	13.06	76	13.10
102	13.49	70	13.46	70	13.83
70	13.66	102	13.87	102	13.96
97	13.88	97	14.12	97	14.25
85	14.34	85	14.62	85	14.52
54	15.28	54	14.82	54	14.68
91	16.36	91	16.02	91	16.06
62	17.11	62	16.98	100	17.03
100	17.27	100	17.20	62	173.2
51	20.48	51	20.64	51	20.78
57	22.00	57	22.681	57	22.68

7.4.2.2 Ward Cluster Analysis and PCA of Signal Data Made by LpDSs

Figure 7.7 is a Ward cluster analysis of LpDSs signal data. The upper green part is 50 normal subjects, and the lower red part is 52 cancer patients. We consider the great effects of LpDSs cause this surprising result. In the dendrogram of 50 normal subjects shown on the right side, those become two clusters with 44 and 6 cases from the top. Singh's microarray shows that normal class has two clusters. Surely, there will be some medical implication. The 52 prostate cancers consist of three clusters of 11, 3, and 38 patients. We expect that these clusters will have medical implications easily. Variable and case dendrograms can be analyzed via mosaic charts. Our approach is more simple and reliable than other approaches, as it is an analysis of the signal data.

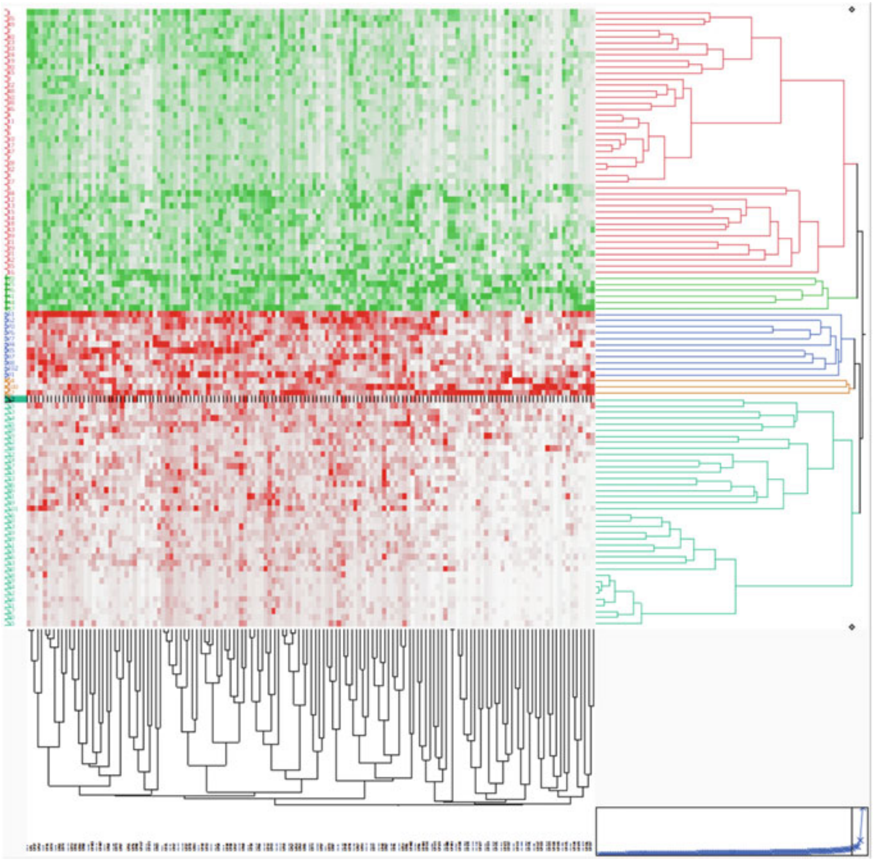


Fig. 7.7 Cluster analysis of signal data of LpDSs

Figure 7.8 shows the result of LpDSs signal data by PCA. The first eigenvalue is 102.128, and the contribution ratio is 74%. The second eigenvalue is 4.744, the contribution ratio is 3.24, and the cumulative contribution ratio is 77.24%. That is, the Prin1 almost represents 102 subjects. Cluster analysis shows that six green normal subjects are adjacent to 11 blue tumor patients. However, PCA shows that between these clusters, there are 44 normal subjects and 38 cancer patients. Thus, caution is required for cluster analysis. Three brown patients such as 57th, 100th, and 54th cases are significant outliers. We recommend physicians survey the green and brown cases.

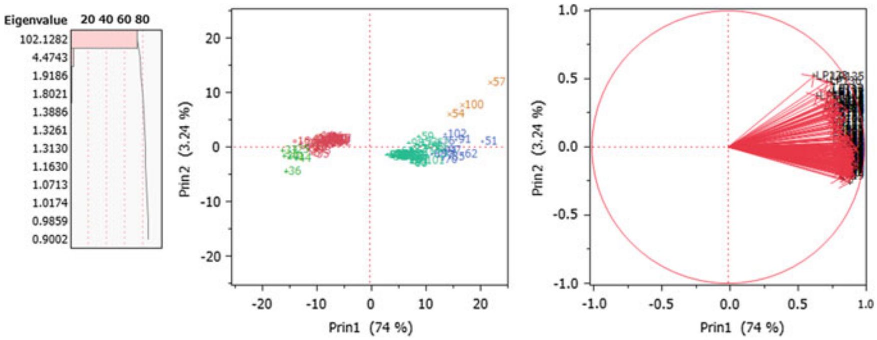


Fig. 7.8 Three plots of PCA (LpDS data)

The ranges of normal class and cancer class are [−16, −6.15] and [3.15, 22.681], respectively. SV opens the window that is the interval (−6.15, 3.15). RatioSV of PCA by LpDSs is (7.2).

$$\text{RatioSV of PCA by LpDSs} = (6.15 + 3.15)/(16 + 22.681) * 100 = 24.04281\% \tag{7.2}$$

Because the maximum RatioSV of HsvmDSs is 32, RatioSV of PCA is useless as a malignancy index. However, there are several outliers as same as RipDSs. This fact is the merit to analyze the signal data by PCA.

7.4.2.3 Ward Cluster Analysis and PCA of Signal Data Made By HsvmDSs

Figure 7.9 is a Ward cluster analysis of HsvmDSs signal data. The upper green part is 50 normal subjects, and the lower red part is 52 cancer patients. We consider the great effects of HsvmDSs cause this surprising result. In the dendrogram of 50 normal subjects shown on the right side, those become two clusters with 44 and 6 cases from the top. Singh's microarray shows that normal class has two clusters. Surely, there will be some medical implication. The 52 prostate cancers consist of three clusters of 11, 3, and 38 patients. We expect that these clusters will have medical implications easily. Variable dendrogram may be analyzed by relating the results of case dendrogram with mosaic charts. Our approach is easier and more reliable than other approaches, as it is an analysis of the signal data.

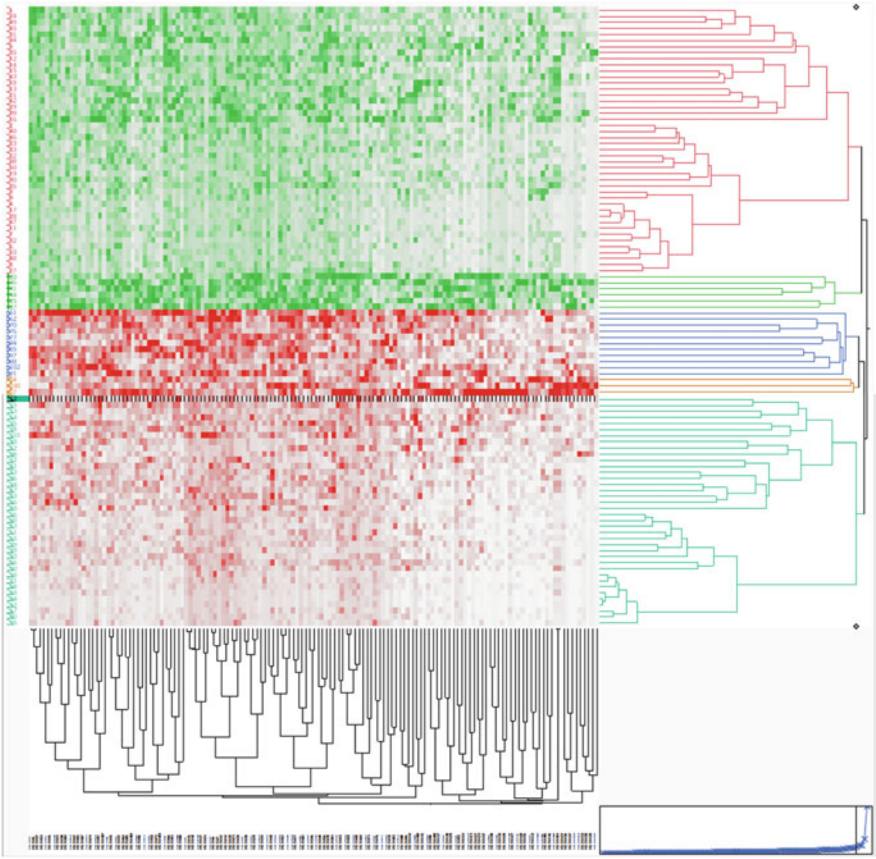


Fig. 7.9 Cluster analysis of signal data of HsvmDSs

Figure 7.10 shows the result of HsvmDSs signal data by PCA. The first eigenvalue is 103.908, and the contribution ratio is 74.8%. The second eigenvalue is 4.789, the contribution ratio is 3.45%, and the cumulative contribution ratio is 79.25%. The Prin1 represents 102 subjects. From the scatter plot, we confirmed several outliers as same as Fig. 7.8. Although the second eigenvalue is small, the dispersion of cancer patient class is large on the Prin2. In other words, the Prin1 becomes an indicator of cancer malignancy as same as individual RipDSs. The fifth and sixth columns of the ranges of normal and cancer classes are [−16.09, −6.16] and [3.2, 22.68]. SV opens the window that is the interval (−6.16, 3.2). RatioSV of PCA by HsvmDSs is (7.3).

$$\text{RatioSV of PCA by LpDSs} = (6.16 + 3.2)/(16.09 + 22.68) * 100 = 24.14238\% \tag{7.3}$$

Because the maximum RatioSV of HsvmDSs is 29.48, RatioSV of PCA is useless as a malignancy index. However, scatter plot tells us several outliers. This fact is the merit to analyze the signal data by PCA.

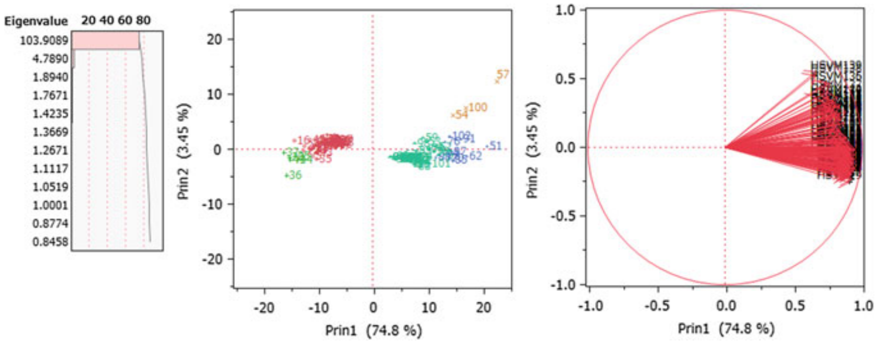


Fig. 7.10 Three plots of PCA (HsvmDS data)

7.4.3 *Transposed Data of RipDSs Using 139 SMs Found by RIP*

Figure 7.11 is the Ward cluster analysis of transposed data of RipDSs using 139 SMs found by RIP. Analyzing the transposed matrix gives different results. 139 RipDSs are divided into five clusters having 115 (Red), 18 (Green), two (Blue), three (Brown) and one (RIP139) subjects. Whether these five clusters of malignancy indicators play the same role medically, it is a future research subject. The dendrogram of the variable (102 subjects) tells us that the two green cancers separate 50 normal cases into two clusters. We hope physicians examine this result. Until now, medical researchers have developed and used their methods because conventional statistical methods are largely useless. However, if they create signal data, they only need to medically examine the abundant results analyzed by general statistical methods.

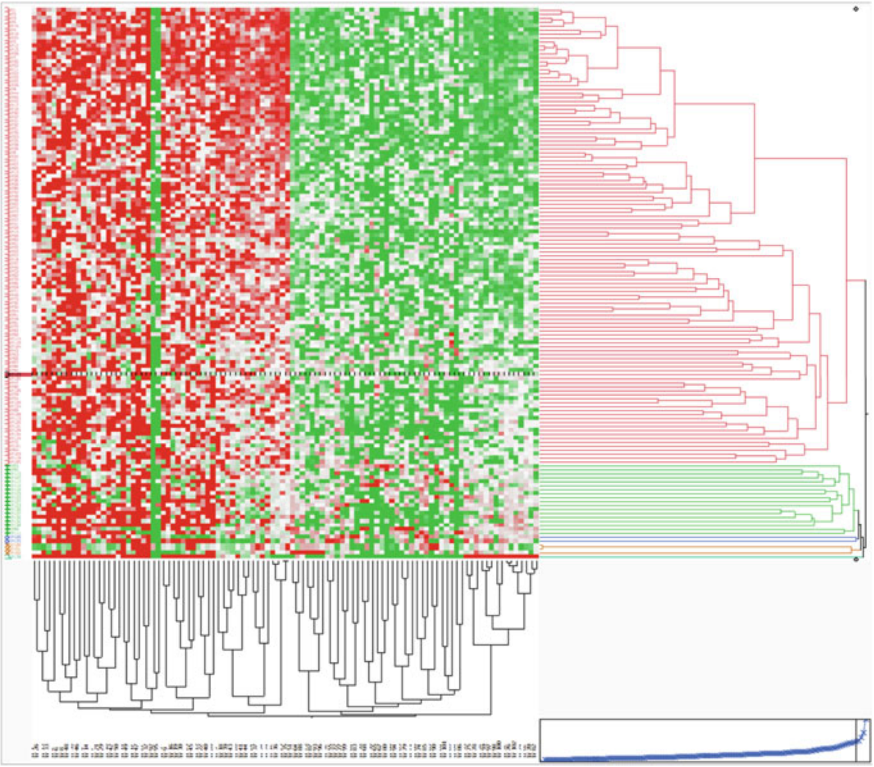


Fig. 7.11 Cluster analysis of signal data of RipDS using 139 SMs found by RIP

Figure 7.12 shows three plots of PCA. Because the 115 red RipDSs include the origin of the scatter plot, to some extent, those can be indicators of general malignancy. However, roles are slightly different if those belong to different quadrants. The 18 green RipDSs surround the red cluster. The two brown RipDSs are in the fourth quadrant. Two blue RipDSs are in the first and fourth quadrant. One cluster of RIP139 has small RatioSV, and we cannot find it. Both scatter plot and factor loading plot have meanings in each quadrant. The contribution ratio of Prin1 and Prin2 are only 20.5% and 9.31%, and the cumulative contribution rate is not sufficient as 29.81%. Based on medical knowledge, the results will be useful for the categorization of many SM and BGS.

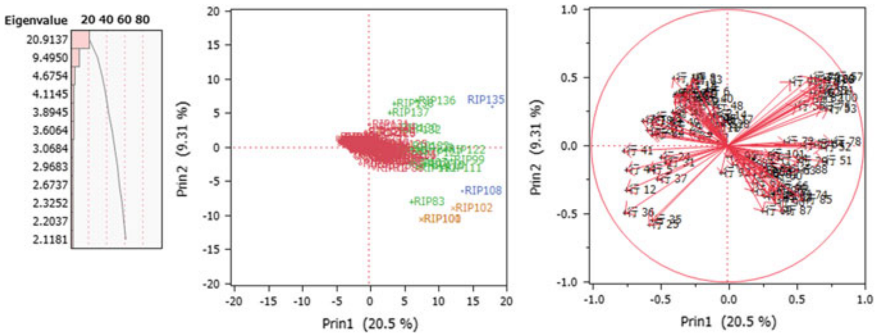


Fig. 7.12 PCA of transposed data of 139 RipDSs

7.5 Conclusions

Only Alon and Singh consist of healthy subjects and cancer patients. Other four microarrays consist of two groups of different cancers. However, it is essential that the results of all SMs obtained by three signal data are almost the same. If microarrays are well managed for research purposes, we believe that the two classes are LSDs and obtain almost the same results as this book. In other words, microarray does not include cancer patients receiving treatment for cancer and healthy subjects suspected of cancer. The range of 9,591 correlations of 139 RipDSs is [0.417, 1], and we indicate two SMs with R = 1 may be redundant to each other. Moreover, we show how to make the 139 different RipDSs and categorize those into several groups. Singh's microarray consists of 139 RipDSs, but we must categorize medically into several groups. We proposed to classify those by statistical methods, but physicians need to validate those categories.

References

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Sall JP, Creighton L, Lehman A (2004) JMP start statistics (3rd edn). SAS Institute Inc. USA (Shinmura S. edits Japanese version)

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2016) New theory of discriminant analysis after R. Fisher. Springer, Tokyo

Shinmura S (2017) Cancer gene analysis by Singh et al. microarray data. ISI2017, pp 1–6

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD'18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. In: Biocomp'18, pp 1–7

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1):68–74. (https://doi.org/10.1038/nm0102-6)

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Lada M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(2):203–209

Chapter 8
Cancer Gene Diagnosis of Tian et al. Microarray

Abstract We developed the New Theory of Discriminant Analysis after R. A. Fisher (theory). Although there are five severe problems of discriminant analysis, theory solves five problems completely. Especially, Revised IP-OLDF (RIP) based on MNM and Method2 firstly succeed in the cancer gene analysis (Problem5) from 1970. RIP decomposes six microarrays into the many SMs those are signals (MNM = 0) explained in Chap. 1. Although Revised LP-OLDF decomposes the microarray into many SMs as same as RIP, we find the defect of Revised LP-OLDF that cannot find all SMs from the microarray in Chap. 4. However, Revised LP-OLDF can find many SMs faster than RIP. It may be convenient for many researchers to analyze SMs found by Revised LP-OLDF. Tian's microarray consists of 173 subjects (36 False subjects and 137 True patients) and 12,625 genes. In this chapter, Revised LP-OLDF decomposes Tian's microarray into the 104 SMs. We analyze 104 SMs by the standard statistical method such as one-way ANOVA, t-test, Ward cluster analysis, PCA, logistic regression, and Fisher's LDF. Although we expected standard statistical methods were useful for cancer gene diagnosis, only logistic regression could discriminate 104 SMs correctly, and other methods did not show the linear separable facts. Because Revised LP-OLDF discriminates 104 SMs, and the range of 104 RatioSVs is [8.34%, 22.79%], we make signal data by 104 Revised LP-OLDF discriminant scores (LpDSs) instead of 12,625 genes. By this breakthrough, hierarchical cluster methods can separate two classes as two clusters entirely. In addition to these results, the Prin1 axis of PCA indicates proper malignancy indexes as same as 104 malignancy indexes. Thus, we reconsider the signal data is the signal. Moreover, we examine the characteristic of 104 LpDSs precisely as same as Chap. 7 using the correlation analysis.

Keywords Cancer gene diagnosis · Malignancy indexes · Revised LP-OLDF discriminant scores (LpDSs) · Correlation analysis · Small Matryoshka (SM) · RatioSV of PCA · Ward cluster · PCA

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_8

Thanks to Tian et al.

We appreciate Tian et al. (2003)[1] for providing excellent data. Below, we will quote their "summary" for the reader.

Background

Myeloma cells may secrete factors that affect the function of osteoblasts, osteoclasts, or both.

Methods

We subjected purified plasma cells from the bone marrow of patients with newly diagnosed multiple myeloma and control subjects to oligonucleotide microarray profiling and biochemical and immunohistochemical analyses to identify molecular determinants of osteolytic lesions.

Results

We studied 45 control subjects, 36 patients with multiple myeloma in whom focal lesions of bone could not be detected by magnetic resonance imaging (MRI), and 137 patients in whom MRI detected such lesions. **Different patterns of expression of 57 of approximately 10,000 genes** from purified myeloma cells could be used to distinguish the two groups of patients ($P < 0.001$). Permutation analysis, which adjusts the significance level to account for multiple comparisons in the datasets, showed that 4 of these 57 genes were significantly overexpressed by plasma cells from patients with focal lesions. One of these genes, dickkopf1 (DKK1), and its corresponding protein (DKK1) were studied in detail because DKK1 is a secreted factor that has been linked to the function of osteoblasts. Immunohistochemical analysis of bone marrow–biopsy specimens showed that only myeloma cells contained detectable DKK1. Elevated DKK1 levels in bone marrow plasma and peripheral blood from patients with multiple myeloma correlated with the gene-expression patterns of DKK1 and were associated with the presence of focal bone lesions. Recombinant human DKK1 or bone marrow serum containing an elevated level of DKK1 inhibited the differentiation of osteoblast precursor cells in vitro.

Conclusion

The production of DKK1, an inhibitor of osteoblast differentiation, by myeloma cells is associated with the presence of lytic bone lesions in patients with multiple myeloma."

8.1 Introduction

We developed the New Theory of Discriminant Analysis after R. A. Fisher (theory) (Shinmura 2016). Although there are five severe problems of discriminant analysis (Shinmura 2016), theory solves five problems completely. Especially, Revised IP-OLDF (RIP) based on MNM and Method2 firstly succeed in the cancer gene analysis

[1] Erming Tian, Fenghuang Zhan, Ronald Walker, Erik Rasmussen, Yupo Ma, Bart Barlogie, and John D. Shaughnessy.

(Problem5) since 1970. RIP decomposes six microarrays into the many SMs those are signals and linearly separable gene subspaces (MNM = 0) explained in Chap. 1 (Schrage 2006). Although Revised LP-OLDF decomposes the microarray into many SMs as same as RIP, we find the defect of Revised LP-OLDF that cannot find all SMs from the microarray in Chap. 4. However, Revised LP-OLDF can find many SMs faster than RIP. It may be convenient for many researchers to analyze SMs found by Revised LP-OLDF. Tian's microarray consists of 173 subjects (36 False subjects and 137 True patients) and 12,625 genes. In this chapter, Revised LP-OLDF decomposes Tian's microarray into the 104 SMs. We analyze 104 SMs by six MP-based LDFs. Because the ranges of 104 RatioSVs by the RIP and Revised LP-OLDF are [8.34, 22.79%] and [4.2, 21.8%], this chapter introduces the result of Revised LP-OLDF. We make signal data that consists of 173 subjects and 104 Revised LP-OLDF discriminant scores (LpDSs) instead of 12,625 genes. By this breakthrough, Ward cluster analysis can separate two classes as two clusters, and the Prin1 axis of PCA indicates proper malignancy index as same as 104 malignancy indexes. Moreover, we examine the characteristic of 104 LpDSs precisely as same as Chap. 7. Furthermore, we examine the Problem6 of cancer gene analysis using 104 SMs and LpDSs as follows:

Problem6: Why can no researchers find the linear separable facts in SM since 1970?

We had already obtained the hint of Problem6 in Chaps. 4 and 5. The hint is as follows: Although two SVs can separate two classes of microarray, the variation of the two classes is tiny, and the signal is buried in the noise. This fact is already pointed out as one of three difficulties discussed by the statisticians. In this chapter, we explain the reason by clear information about LpDSs and SMs using the correlation analysis. This book concept is as follows. LINGO (Schrage 2006) decomposes Tian's microarray into 104 SMs and opens a new frontier of cancer gene analysis. JMP (Sall et al. 2004) analyzes all SMs and offers cancer gene diagnosis. Shinmura (2016, 2017, 2018a, b) relate to this Chapter.

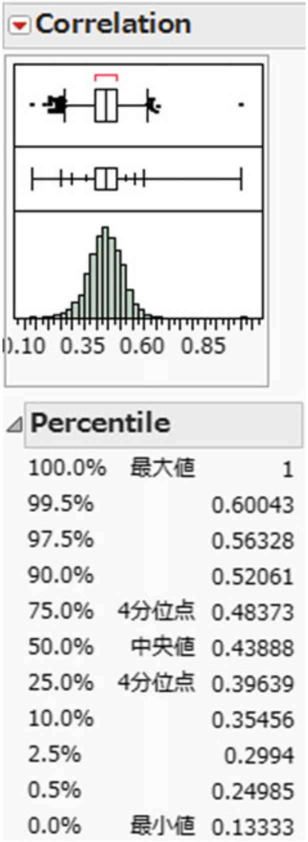
8.2 Examination of Revised LP-OLDF Discriminant Scores and SMs

Because we obtain almost the same results by the RIP and Revised LP-OLDF, we answer the Problem6 from the examination of 104 LpDSs and SMs.

8.2.1 Correlation of 104 LpDSs

Figure 8.1 is the histogram of 5,356 correlations (abbreviated R) of 104 LpDSs analyzed by JMP. The range of correlations is [0.133, 1]. We believe that two LpDSs with a correlation of 1 will play the same role in oncogenic diagnosis. The correlation analysis finds four important SMs such as (SM27, SM28) and (SM98, SM99) in Table 8.1. We will deeply survey four SMs for solving Problem6 in future research. If we omit the four SMs, the range of R is [0.133, 0.600]. Tian's 100 LpDSs seem to be relatively low correlated.

Fig. 8.1 Histogram of 5,356 correlations by 104 LpDSs



Percentile

100.0%	最大値	1
99.5%		0.60043
97.5%		0.56328
90.0%		0.52061
75.0%	4分位点	0.48373
50.0%	中央値	0.43888
25.0%	4分位点	0.39639
10.0%		0.35456
2.5%		0.2994
0.5%		0.24985
0.0%	最小値	0.13333

Table 8.1 is the list of 5,356 correlations sorted by descending order of R. The [2.5, 97.5%] is the 95% confidence interval of each R. Because 5,354 p-values are 0.01, these correlations are positive. However, we cannot explain the reason why there are no high correlations of 0.658 to less than 1. On the other hand, we expect four LpDSs having correlation1 may be useful medically.

Table 8.1 List of 5,356 correlations sorted by descending order of R

Var1	Versus Var2	Correlation	n	2.5%	97.5%	p-value
LP28	LP27	1	173	1	1	0.000
LP99	LP98	1	173	1	1	0.000
LP80	LP70	0.658	173	0.564	0.735	0.000
LP86	LP39	0.651	173	0.556	0.729	0.000
LP78	LP56	0.636	173	0.538	0.717	0.000
LP85	LP79	0.634	173	0.536	0.716	0.000
LP49	LP34	0.626	173	0.526	0.709	0.000
LP95	LP49	0.625	173	0.525	0.709	0.000
LP56	LP23	0.624	173	0.524	0.707	0.000
LP53	LP39	0.617	173	0.515	0.702	0.000
–	–	–	–	–	–	–
LP99	LP50	0.226	173	0.079	0.363	0.003
LP96	LP24	0.215	173	0.068	0.353	0.004
LP104	LP98	0.208	173	0.060	0.346	0.006
LP104	LP99	0.208	173	0.060	0.346	0.006
LP104	LP72	0.205	173	0.057	0.343	0.007
LP104	LP17	0.204	173	0.057	0.343	0.007
LP104	LP40	0.204	173	0.057	0.343	0.007
LP104	LP30	0.204	173	0.056	0.343	0.007
LP104	LP102	0.186	173	0.038	0.326	**0.014**
LP104	LP46	0.133	173	−0.016	0.277	**0.080**

8.2.2 *PCA of 104 LpDSs*

We analyze the 104 LpDSs by PCA and output the 30 principal components showed in Table 8.2. The eigenvalue of Prin1 is 102.668, and the contribution rate is 73.862%. The eigenvalue of Prin2 is 4.742, and the contribution rate is 3.412%. Thus, two principal components explain the 77.274% of total variance and 30 principal components explain the 94.21% of total variance. Because two classes are completely separated in the signal data, the first eigenvalue is very large.

Table 8.2 PCA of 104 LpDSs

Prin	Eigenvalue	Contribution	Cumulative
1	102.668	73.862	73.862
2	4.742	3.412	77.274
3	1.972	1.418	78.692
4	1.802	1.297	79.989
5	1.532	1.102	81.091
6	1.400	1.007	82.098
7	1.240	0.892	82.990
8	1.132	0.815	83.805
9	1.053	0.757	84.562
10	0.976	0.702	85.264
11	0.929	0.668	85.933
12	0.896	0.645	86.578
13	0.888	0.639	87.216
14	0.832	0.598	87.815
15	0.775	0.557	88.372
16	0.704	0.506	88.878
17	0.693	0.499	89.377
18	0.683	0.491	89.868
19	0.648	0.466	90.335
20	0.607	0.437	90.771
21	0.581	0.418	91.190
22	0.555	0.399	91.589
23	0.531	0.382	91.971
24	0.500	0.360	92.330
25	0.488	0.351	92.682
26	0.471	0.339	93.020
27	0.440	0.316	93.337
28	0.419	0.301	93.638
29	0.403	0.290	93.928
30	0.392	0.282	94.210

Figure 8.2 is eight scatter plots. All x-axes are Prin1. The y-axes in the upper plots are from Prin2 to Prin5, and the y-axes in lower plots are from Prin27 to Prin30. Left circles are the 99% confidence ellipse of the False class, and right circles are the 99% confidence ellipse of the True class. The 29 scatter diagrams shows two classes are separable on Prin1 entirely. Thus, the Prin1 of PCA becomes the malignancy index to summarize 104 LpDSs.

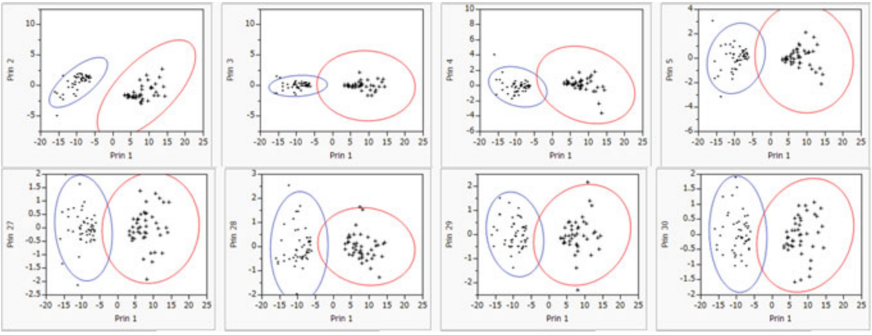


Fig. 8.2 Eight scatter plots (x-axis: Prin1; upper y-axes: From Prin2 to Prin5; lower y-axes: from Prin27 to Prin30)

Figure 8.3 is PCA output of the 104 LpDSs. The scatter plot is the same as the left upper scatter plot in Fig. 8.2. If we look for the 29 scatter plots from Prin2 to Prin30, False's 99% confidence ellipse becomes large sequentially, approaching the same size as True's ellipse. Because the eigenvalues of Prin2 and higher are small, Prin1 is considered to be a malignant index representing two classes.

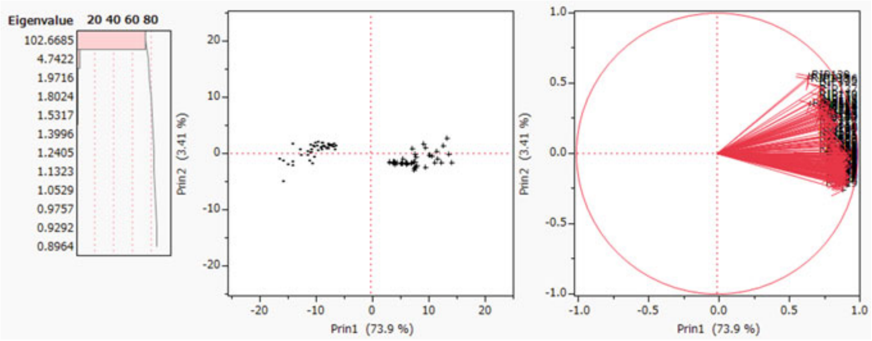


Fig. 8.3 PCA output of the 104 LpDSs

8.2.3 How to Categorize Many 104 LpDSs

RIP and Revised LP-OLDF can decompose the microarrays into many SMs (Fact4). Because RIP, Revised LP-OLDF, and H-SVM can discriminate two classes of all SMs entirely, we consider the genes included in each SM as cancer genes and signals. However, other statistical discriminant functions cannot discriminate between two classes completely. On the other hands, because six signal data made by RIP, Revised

LP-OLDF, and H-SVM using two kinds of SMs found by RIP and Revised LP-OLDF show the linear separable facts by other statistical methods, we consider six signal data are signals. These facts indicate that only three LDFs can discriminate two classes entirely and other methods cannot find the linear separable facts.

By the breakthrough of signal data made by 104 LpDSs, we can succeed to obtain the 104 malignancy indexes and open the door of cancer gene diagnosis. Thus, we survey how to build 104 LpDSs in this section. The second and third columns of Table 8.3 show the minimum and maximum subjects of LpDS included in each SM from SM1 to SM104. Because we choose the minimum number of each LpDS from the 36 False classes, the selected subject is considered to be fairely better. The maximum number of LpDS among the 137 True classes is that the degree of True is the worst. The fifth column is the range of LpDS (abbreviated LPi), and the last column is RatioSV of each LPi. The range of 104 LpDSs is [4.2%, 21.8%]. The maximum value 21.8% is small compared with other microarrays.

Table 8.3 Minimum and maximum subject's SM and its RatioSV

SM	Min	Max	LpDS	Range	RatioSV
SM1	6	150	LP1	15.3	13.1
SM2	23	93	LP2	11.3	17.7
SM3	1	52	LP3	14.0	14.2
SM4	19	157	LP4	10.8	18.5
SM5	34	92	LP5	15.4	13.0
SM6	6	173	LP6	16.3	12.3
SM7	23	107	LP7	13.5	14.9
SM8	23	38	LP8	15.6	12.8
SM9	8	70	LP9	15.6	12.9
SM10	3	145	LP10	24.7	8.1
SM11	33	55	LP11	14.5	13.8
SM12	16	148	LP12	16.4	12.2
SM13	30	154	LP13	12.8	15.7
SM14	29	157	LP14	13.2	15.1
SM15	23	170	LP15	13.5	14.8
SM16	23	37	LP16	15.7	12.7
SM17	26	150	LP17	10.3	19.5
SM18	34	37	LP18	14.2	14.1
SM19	9	51	LP19	15.3	13.0
SM20	3	150	LP20	16.8	11.9
SM21	10	143	LP21	11.3	17.8

(continued)

Table 8.3 (continued)

SM	Min	Max	LpDS	Range	RatioSV
SM22	11	145	LP22	19.0	10.5
SM23	25	169	LP23	12.1	16.6
SM24	6	65	LP24	23.6	8.5
SM25	16	82	LP25	14.0	14.3
SM26	23	101	LP26	23.0	8.7
SM27	35	68	LP27	17.4	11.5
SM28	35	68	LP28	17.4	11.5
SM29	14	173	LP29	24.1	8.3
SM30	25	48	LP30	23.8	8.4
SM31	24	73	LP31	14.9	13.4
SM32	7	102	LP32	11.7	17.1
SM33	4	75	LP33	14.2	14.1
SM34	3	84	LP34	14.4	13.9
SM35	10	169	LP35	11.5	17.4
SM36	19	103	LP36	12.3	16.2
SM37	18	46	LP37	13.6	14.7
SM38	22	129	LP38	16.8	11.9
SM39	5	100	LP39	12.7	15.7
SM40	3	44	LP40	15.5	12.9
SM41	8	136	LP41	15.4	13.0
SM42	8	164	LP42	16.9	11.8
SM43	29	84	LP43	20.7	9.6
SM44	32	85	LP44	17.8	11.2
SM45	31	71	LP45	22.5	8.9
SM46	31	100	LP46	16.1	12.4
SM47	5	166	LP47	16.7	12.0
SM48	5	153	LP48	12.8	15.6
SM49	24	46	LP49	10.6	18.9
SM50	16	164	LP50	17.1	11.7
SM51	1	110	LP51	24.8	8.1
SM52	1	63	LP52	19.5	10.3
SM53	31	122	LP53	10.2	19.5
SM54	21	63	LP54	14.6	13.7
SM55	8	169	LP55	13.7	14.5
SM56	33	112	LP56	10.3	19.3
SM57	20	148	LP57	21.0	9.5
SM58	8	102	LP58	12.0	16.6

(continued)

Table 8.3 (continued)

SM	Min	Max	LpDS	Range	RatioSV
SM59	10	120	LP59	12.0	16.7
SM60	30	44	LP60	12.1	16.6
SM61	25	37	LP61	18.5	10.8
SM62	12	164	LP62	13.3	15.0
SM63	26	144	LP63	11.7	17.0
SM64	7	114	LP64	9.2	21.8
SM65	33	102	LP65	17.2	11.6
SM66	30	156	LP66	11.7	17.1
SM67	15	71	LP67	19.6	10.2
SM68	31	141	LP68	14.7	13.6
SM69	2	129	LP69	16.9	11.9
SM70	10	65	LP70	12.9	15.5
SM71	12	61	LP71	13.2	15.2
SM72	20	70	LP72	15.1	13.3
SM73	6	100	LP73	14.7	13.6
SM74	22	173	LP74	21.2	9.4
SM75	10	79	LP75	14.0	14.3
SM76	13	102	LP76	13.8	14.5
SM77	3	102	LP77	21.2	9.4
SM78	19	65	LP78	10.4	19.2
SM79	25	44	LP79	11.5	17.4
SM80	30	148	LP80	13.9	14.4
SM81	6	43	LP81	13.9	14.4
SM82	14	107	LP82	15.3	13.0
SM83	9	38	LP83	15.6	12.8
SM84	8	164	LP84	16.6	12.0
SM85	24	40	LP85	12.5	16.0
SM86	3	40	LP86	13.7	14.6
SM87	6	130	LP87	17.9	11.2
SM88	9	153	LP88	15.5	12.9
SM89	10	124	LP89	14.7	13.6
SM90	11	102	LP90	17.7	11.3
SM91	9	173	LP91	9.4	21.3
SM92	13	43	LP92	12.6	15.9
SM93	28	107	LP93	25.3	7.9

(continued)

Table 8.3 (continued)

SM	Min	Max	LpDS	Range	RatioSV
SM94	28	70	LP94	12.5	16.0
SM95	34	44	LP95	17.7	11.3
SM96	12	43	LP96	17.8	11.3
SM97	1	37	LP97	15.2	13.2
SM98	11	102	LP98	25.7	7.8
SM99	11	102	LP99	25.7	7.8
SM100	32	105	LP100	20.8	9.6
SM101	8	103	LP101	16.7	12.0
SM102	8	147	LP102	18.8	10.7
SM103	13	100	LP103	22.1	9.1
SM104	32	43	LP104	48.0	4.2

We sort the second column of Table 8.3 in descending order. As also shown in Chap. 7, the left five columns of Table 8.4 are the first 52 results and the right five columns are the remaining 52 results. The Pair column is the number of SMs with the same minimum and maximum value. The correlation shows their correlation coefficient. There are two sets of two LpDSs having the same pair, and the correlation coefficients are 1 and 0.397. There are one set of three LpDSs having the same pair, and the correlation coefficients are 1, 0.457, and 0.457. It reflects that only two correlations are 1, and the rest are less than 0.6 and is entirely different from Singh's LpDSs. Because other 97 correlation coefficients are between 0.13 and 0.6, these LpDSs may be different malignancy indexes. Correlation analysis tells us the difference between LpDSs. In the abstract, Tian et al. introduce as follows: "Different patterns of expression of 57 of approximately 10,000 genes from purified myeloma cells could be used to distinguish the two groups of patients (P < 0.001)." We would like to compare 104 LpDSs with their patterns.

Table 8.4 Sorted in descending order of the second column (False) and the seventh column (False)

SM	False	True	Pair	Corr	SM	FALSE	TRUE	Pair	Corr
SM97	1	37			SM29	14	173		
SM3	1	52			SM67	15	71		
SM52	1	63			SM25	16	82		
SM51	1	110			SM12	16	148		
SM69	2	129			SM50	16	164		
SM86	3	40			SM37	18	46		

(continued)

Table 8.4 (continued)

SM	False	True	Pair	Corr	SM	FALSE	TRUE	Pair	Corr
SM40	3	44			SM78	19	65		
SM34	3	84			SM36	19	103		
SM77	3	102			SM4	19	157		
SM10	3	145			SM72	20	70		
SM20	3	150			SM57	20	148		
SM33	4	75			SM54	21	63		
SM39	5	100			SM38	22	129		
SM48	5	153			SM74	22	173		
SM47	5	166			SM16	23	37		
SM81	6	43			SM8	23	38		
SM24	6	65			SM2	23	93		
SM73	6	100			SM26	23	101		
SM87	6	130			SM7	23	107		
SM1	6	150			SM15	23	170		
SM6	6	173			SM85	24	40		
SM32	7	102			SM49	24	46		
SM64	7	114			SM31	24	73		
SM9	8	70			SM61	25	37		
SM58	8	102			SM79	25	44		
SM101	8	103			SM30	25	48		
SM41	8	136			SM23	25	169		
SM102	8	147			SM63	26	144		
SM42	8	164	2	0.397	SM17	26	150		
SM84	8	164			SM94	28	70		
SM55	8	169			SM93	28	107		
SM83	9	38			SM43	29	84		
SM19	9	51			SM14	29	157		
SM88	9	153			SM60	30	44		
SM91	9	173			SM80	30	148		
SM70	10	65			SM13	30	154		
SM75	10	79			SM66	30	156		
SM59	10	120			SM45	31	71		
SM89	10	124			SM46	31	100		
SM21	10	143			SM53	31	122		
SM35	10	169			SM68	31	141		
SM90	11	102	3	0.457	SM104	32	43		
SM98	11	102		0.457	SM44	32	85		

(continued)

Table 8.4 (continued)

SM	False	True	Pair	Corr	SM	FALSE	TRUE	Pair	Corr
SM99	11	102		1.000	SM100	32	105		
SM22	11	145			SM11	33	55		
SM96	12	43			SM65	33	102		
SM71	12	61			SM56	33	112		
SM62	12	164			SM18	34	37		
SM92	13	43			SM95	34	44		
SM103	13	100			SM5	34	92		
SM76	13	102			SM27	35	68	2	1.000
SM82	14	107			SM28	35	68		

8.3 Analysis of 104 SMs of Tian et al. Microarray (2018)

In 2018, RIP of LINGO Program3 decomposes Tian's microarray into 104 SMs (12,334 genes). At first, we consider 104 SMs are signals, and 291 gene subspaces are noise. This fact indicates signal subspace includes 12,334 genes and noise subspace includes only 291 genes. If this definition of the signal is valid, other statistical methods can find the linear separable facts easily. However, those methods cannot find the linear separable facts. Thus, we consider six signal data define the true definition of signal. If we accept this definition, we can explain two reasons: (1) why only three LDFs can separate two classes, and (2) why other statistical methods cannot find the linear separable fact (Shinmura 2018a, b).

Table 8.5 shows the 104 SMs from SM = 1 to SM = 104, which is SM found by RIP. Although Revised LP-OLDF can decompose microarrays into other types of SMs, we omit those results. Program3 determines this order of SM. The "gene" column is the number of genes of each SM. The range of genes included in the 104 SMs is [93,144]. The average is 118.6. Row "SUM" indicates 104 SMs contain 12,334 genes. LP and IP can find an optimal solution of a small gene subspace whose number of genes is n (173) subjects or less explained in Chap. 1. From RIP column to H-SVM column show three RatioSVs of 104 SMs by RIP, Revised LP-OLDF and H-SVM. Three ranges of RatioSV are [8.34, 22.79], [4.17, 21.81], and [14.65, 28.75], respectively. Three averages of RatioSVs are 14.18%, 13.39%, and 20.53%, respectively. Row "Max Ratio" indicates the number of the maximum RatioSVs of 104 SMs those are 5, 1, and 98, respectively. To summarize these results, the range, average, and maximum number of H-SVM are better than RIP because the maximization SV of H-SVM works well. Two columns "MAX and MIN" are the maximum and minimum values of three LDFs. Because all NMs of logistic regression, SVM4 and QDF are zero and 104 SMs are linearly separable, we omit these columns from the table. Two columns "SVM1 and LDF2" show the NMs. Although SVM4 can discriminate 104 SMs completely, SVM1 cannot discriminate four SMs correctly. The 71 NMs of LDF2 are not zero.

Table 8.5 Summary of six RatioSVs of six MP-based LDFs and NMs of other discriminant functions

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2
1	112	**17.48**	13.06	17.24	17.48	10.17	0	2
2	117	14.34	15.73	**21.96**	21.96	14.34	0	0
3	132	15.50	17.66	**20.98**	20.98	15.50	0	0
4	114	17.31	14.24	**24.53**	24.53	14.24	0	3
5	109	12.19	18.52	**19.62**	19.62	12.19	0	1
6	116	12.52	12.99	**16.64**	16.64	12.52	0	3
7	126	8.52	12.28	**19.74**	19.74	8.52	0	2
8	117	11.66	14.86	**21.29**	21.29	11.66	0	1
9	117	13.85	12.79	**18.03**	18.03	12.79	0	2
10	119	12.26	12.86	**21.79**	21.79	12.26	0	0
11	116	11.16	8.11	**17.65**	17.65	8.11	0	2
12	119	13.01	13.82	**21.07**	21.07	13.01	0	0
13	119	14.60	12.21	**18.16**	18.16	12.21	0	4
14	127	16.35	15.66	**26.74**	26.74	15.66	0	1
15	121	16.75	15.10	**19.19**	19.19	15.10	0	0
16	100	**18.76**	14.77	17.36	18.76	14.77	0	0
17	119	19.31	12.71	**23.76**	23.76	12.71	0	1
18	137	17.16	19.48	**25.00**	25.00	13.44	0	0
19	134	16.21	14.09	**27.25**	27.25	14.09	0	0
20	123	17.19	13.03	**20.75**	20.75	13.03	0	0
21	108	18.59	11.88	**19.08**	19.08	11.88	0	2
22	111	14.84	17.75	**21.22**	21.22	12.41	0	2
23	117	12.77	10.53	**20.36**	20.36	10.53	0	0
24	107	14.08	16.56	**17.54**	17.54	12.05	0	0
25	111	14.27	8.48	**15.42**	15.42	8.48	0	3
26	128	8.36	14.29	**17.36**	17.36	8.36	0	1
27	123	12.01	8.70	**18.67**	18.67	8.70	0	3
28	119	14.452	11.515	**17.15**	17.15	11.52	0	3
29	134	13.80	8.31	**18.87**	18.87	8.31	0	0
30	118	12.13	8.40	**18.70**	18.70	8.40	0	2
31	130	14.68	13.41	**26.88**	26.88	13.41	0	0
32	109	14.52	17.10	**19.25**	19.25	14.52	0	3
33	128	15.01	14.12	**25.35**	25.35	13.72	0	1
34	116	14.05	13.93	**16.57**	16.57	10.45	0	2
35	120	13.85	17.43	**22.95**	22.95	13.85	0	3
36	130	11.60	16.25	**19.02**	19.02	11.28	0	3

(continued)

Table 8.5 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2
37	128	21.32	14.73	**24.53**	24.53	12.35	0	0
38	125	15.59	11.90	**21.78**	21.78	11.90	0	2
39	114	16.59	15.70	**18.73**	18.73	15.70	0	0
40	121	**21.10**	12.93	17.90	21.10	11.15	0	1
41	113	9.08	13.01	**17.57**	17.57	9.08	0	1
42	125	17.22	11.81	**18.92**	18.92	11.81	0	1
43	106	15.78	9.65	**18.20**	18.20	9.65	0	0
44	123	10.54	11.20	**17.57**	17.57	10.54	0	2
45	123	12.73	8.89	**22.03**	22.03	8.89	0	1
46	116	15.49	12.41	**21.03**	21.03	12.41	0	0
47	120	**18.50**	11.95	18.30	18.50	11.95	0	3
48	117	14.73	15.62	**20.48**	20.48	14.73	0	0
49	110	15.84	18.91	**28.02**	28.02	15.84	0	0
50	134	9.45	11.69	**20.45**	20.45	9.45	0	2
51	120	12.76	8.06	**17.05**	17.05	8.06	0	2
52	115	12.69	10.28	**19.53**	19.53	10.28	0	0
53	118	12.23	19.53	**20.90**	20.90	12.23	0	1
54	124	14.55	13.74	**21.64**	21.64	13.74	0	1
55	117	16.16	14.55	**16.97**	16.97	11.28	0	1
56	118	16.40	19.34	**23.58**	23.58	16.40	0	0
57	126	13.70	9.53	**22.69**	22.69	9.53	0	0
58	116	11.84	**16.63**	14.87	16.63	11.16	0	4
59	115	12.26	16.73	**19.54**	19.54	12.26	0	1
60	123	11.59	16.55	**22.99**	22.99	11.59	0	1
61	105	9.31	10.82	**19.40**	19.40	9.31	0	1
62	104	10.18	14.99	**15.40**	15.40	10.18	0	3
63	115	14.75	17.03	**24.28**	24.28	14.75	0	1
64	111	16.40	21.81	**27.08**	27.08	16.40	0	0
65	99	15.71	11.61	**23.74**	23.74	11.61	0	1
66	112	13.38	17.08	**21.06**	21.06	13.38	0	1
67	110	9.46	10.20	**18.51**	18.51	9.46	0	1
68	112	16.70	13.56	**20.39**	20.39	13.56	0	2
69	122	12.69	11.85	**26.09**	26.09	11.85	0	0
70	119	18.62	15.55	**18.52**	18.62	14.43	0	1
71	109	13.63	15.17	**17.65**	17.65	13.63	0	1
72	118	16.94	13.27	**17.11**	17.11	13.27	0	1
73	104	16.31	13.60	**18.04**	18.04	13.60	0	1

(continued)

Table 8.5 (continued)

SM	Gene	RIP	LP	HSVM	Max	Min	SVM1	LDF2
74	108	16.76	9.44	**18.19**	18.19	9.44	0	2
75	112	15.32	14.33	**25.87**	25.87	14.33	0	2
76	127	13.63	14.46	**18.90**	18.90	13.63	0	2
77	93	13.749	9.437	**16.08**	16.08	9.44	0	3
78	116	11.96	19.23	**20.32**	20.32	11.96	0	5
79	109	16.578	17.44	**20.11**	20.11	16.58	0	1
80	102	12.946	14.356	**21.62**	21.62	12.95	0	0
81	112	14.409	14.35	**19.6**	19.60	14.35	0	1
82	139	9.033	13.043	**24.88**	24.88	9.03	0	0
83	103	12.749	12.83	**22.04**	22.04	12.75	0	2
84	109	19.28	12.04	**21.16**	21.16	12.04	0	1
85	112	13.91	16.02	**19.70**	19.70	13.91	0	0
86	95	16.41	14.59	**24.16**	24.16	14.59	0	1
87	117	17.43	11.20	**18.57**	18.57	11.20	0	5
88	115	13.32	12.90	**21.80**	21.80	12.90	0	1
89	132	18.47	13.62	**26.18**	26.18	13.62	0	0
90	99	14.19	11.30	**19.93**	19.93	11.30	0	1
91	117	**22.79**	21.28	22.78	22.79	21.28	0	0
92	142	13.26	15.87	**21.58**	21.58	13.26	0	0
93	100	15.21	7.91	**23.67**	23.67	7.91	0	0
94	140	17.942	15.977	**28.75**	28.75	15.98	0	0
95	137	13.65	11.28	**23.31**	23.31	11.28	0	1
96	112	13.51	11.25	**20.15**	20.15	11.25	0	3
97	133	11.08	13.16	**20.08**	20.08	11.08	1	0
98	137	11.14	7.78	**19.80**	19.80	7.78	0	0
99	119	11.14	7.78	**19.80**	19.80	7.78	0	4
100	131	12.10	9.60	**22.87**	22.87	9.60	1	2
101	132	8.34	11.97	**16.86**	16.86	8.34	0	1
102	138	9.14	10.66	**20.17**	20.17	9.14	4	2
103	142	9.08	9.06	**14.65**	14.65	8.06	8	5
104	144	8.97	4.17	**15.44**	15.44	4.17	0	4
MAX	144	22.79	21.81	28.75	28.75	21.28	8	5
MIN	93	8.34	4.17	14.65	14.65	4.17	0	0
Mean	118.60	14.18	13.39	20.53	20.59	11.95	0.13	1.32
Max Ratio		5	1	98				
SUM	12334							

8.4 Analysis of Three Signal Data Made by 104 DSs

We cannot obtain useful results of 104 SMs (173 cases and 12,334 genes) until now. Next, we analyze three signal data made by RipDSs, LpDSs, and HsvmDSs having 104 DSs instead of 12,334 genes. The cluster analysis and PCA get almost the same excellent results. Although we show the results of several cluster methods, we do not interpret detailed analysis results. Many medical researchers use SOMs, but the use of hierarchical cluster methods are easy. Although the results of hierarchical methods usually vary, it is critical that the result of this book is almost the same in each microarray. Interpretation of the case and variable dendrograms will undoubtedly yield results that will be useful for medical researchers. For PCA, healthy subjects place on the negative axis of Prin1. Many cancer patients are on the positive axis, but there is a common feature that it varies even at Prin2 when the malignancy becomes high. PCA can easily identify outliers, also.

Short Column

The work of Tien et al. (2003) is different from the other five.They approached their theme by logistic regression and statistical testing, and validated their medical diagnosis as follows:

They studied 45 control subjects, 36 patients with multiple myeloma in whom focal lesions of bone could not be detected by MRI (False), and 137 patients in whom MRI detected such lesions (True). Different patterns of expression of 57 of 12,625 genes could be used to distinguish the two groups of patients ($p<0.001$). Logistic regression was used to model bone disease in multiple myeloma. The signal for each probe set was log transformed on a base-2 scale before it was entered into the logistic regression model and subjected to permutation analysis, which adjusts the significance level to account for multiple comparisons in data sets with high dimensionality.

Significant differences in patients' characteristics according to their bone-disease status were evaluated with the use of either Fisher's exact test or the chi-square test. Spearman's correlation coefficient was used to measure the correlation between the level of gene expression and protein levels.

They analyzed 12,625 genes from two groups by logistic regression analysis and identified 57 genes that were expressed differently ($P<0.001$) in the two groups of patients.

Thus, they overcame the curse of higher dimension. Because of this, its NM will probably not be zero. And it is considered that 57 genes are divided and included in several SMs. That is, SMs containing 57 genes is a potential SM candidate for gene diagnosis.

8.4.1 Cluster Analysis of Three Signal Data

Figure 8.4 is a Ward cluster analysis of RipDSs signal data. Even if it analyzes 104 SMs individually, it cannot separate two classes, but the upper green part is 36 False subjects, and the lower red part is 137 True patients. We consider the marvelous effects of RipDSs cause this surprising result. The case dendrogram shows one cluster of the False class and four clusters of the True class. Four clusters consist of the 88 green patients, the 42 blue patients, the three orange patients, and the four green patients. Among the six research groups, Alon et al. succeeded in using a self-organizing map (SOM). Furthermore, if medical AI based on the cluster analysis will analyze SM, it may be able to find useful results among many clusters made by many clustering methods.

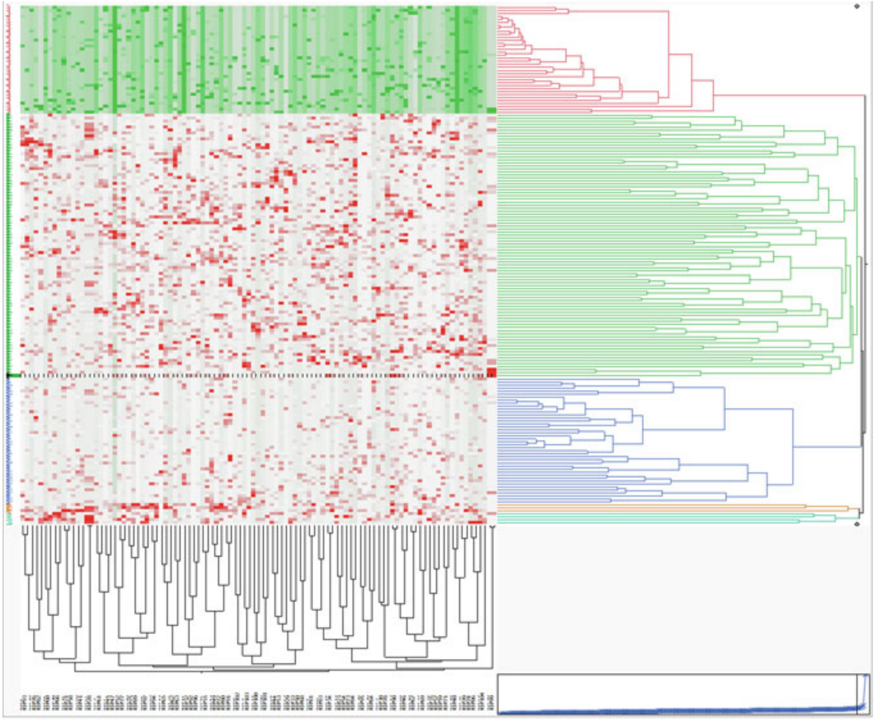


Fig. 8.4 Ward cluster analysis of RipDSs signal data

Figure 8.5 is a Ward cluster analysis of LpDSs signal data. The upper green part is 36 False subjects, and the lower red part is 137 True patients. The case dendrogram shows one cluster of the False class and four clusters of the True class. Four clusters consist of the 61 green patients, the 35 blue patients, the 36 orange patients, and the five green patients. Four clusters are slightly different from Fig. 8.4.

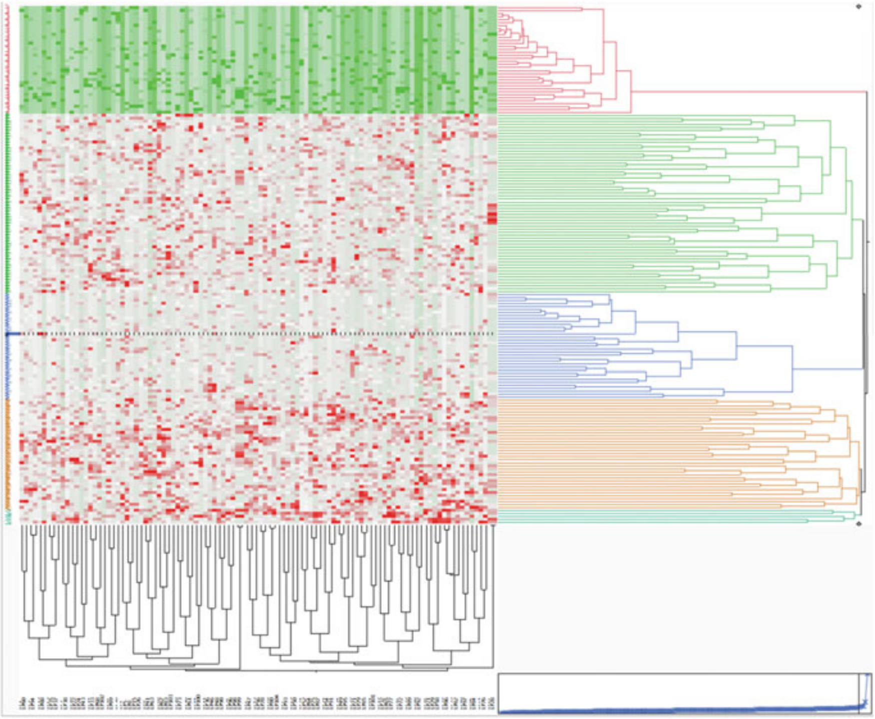


Fig. 8.5 Ward cluster analysis of LpDSs signal data

Figure 8.6 is a Ward cluster analysis of HsvmDSs signal data. The upper green part is 36 False subjects, and the lower red part is 137 True patients. The case dendrogram shows one cluster of the False class and four clusters of the True class. Four clusters consist of the 78 green patients, the 11 blue patients, the four orange patients, and the 44 pale green patients. Because four clusters by RipDSs, LpDSs, and HsvmDSs are entirely different, this is because two classes of Tian et al. have a different structure from the other five. This theme is a future research subject. Generally it is not desirable that the results differ depending on the method of cluster analysis. But if an expert can find a specific meaning in several clusters, it might be useful for genetic diagnosis of cancer.

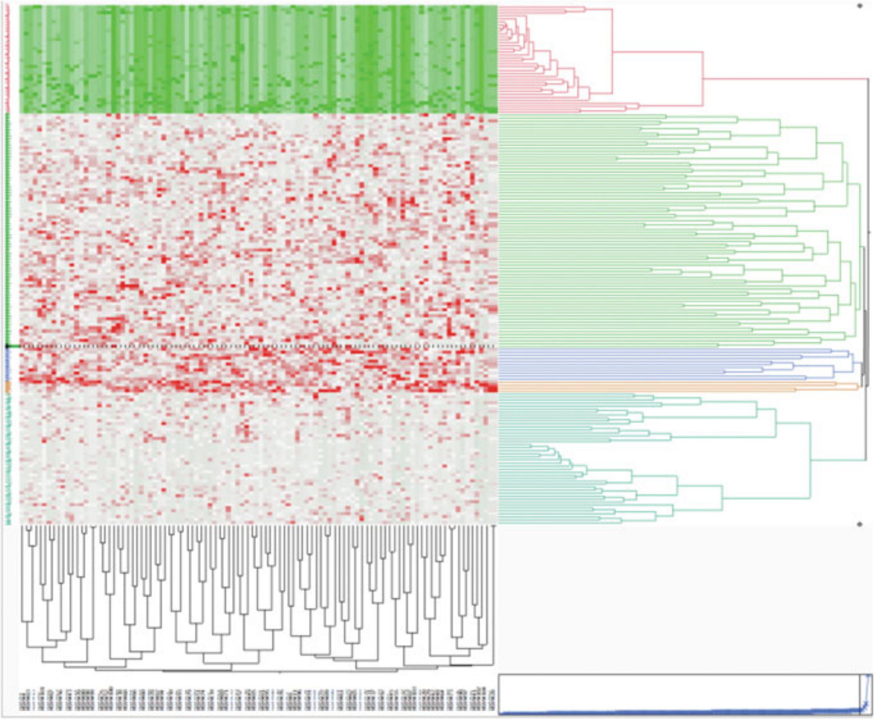


Fig. 8.6 Ward cluster analysis of HsvmDSs signal data

8.4.2 PCA of Three Signal Data

Figure 8.7 shows the result of RipDS signal data by PCA. Left eigenvalue shows that the eigenvalue of Prin1 is larger than the others. The first eigenvalue is 44.930, and the contribution ratio is 43.2%. The second eigenvalue is 2.227, the contribution ratio is 2.14, and the cumulative contribution ratio is 45.34%. That is, the Prin1 almost presents 172 subjects. The score plot shows the second eigenvalue is small and the variation is small. Although the False subjects are almost on the Prin1, its shape is the ellipse because they are not healthy subjects. True patients are in the range [−1.88, 13.75], and as an increasing distance from False subjects, the dispersion of the Prin2 is large. Especially 156th, 100th, 173th, 148th, 99th, 145th, 122th, and 40th patients are large outliers. That is, the Prin1 becomes cancer malignancy index as same as 104 RipDSs. The score plot shows the second eigenvalue is small and the variation is small.

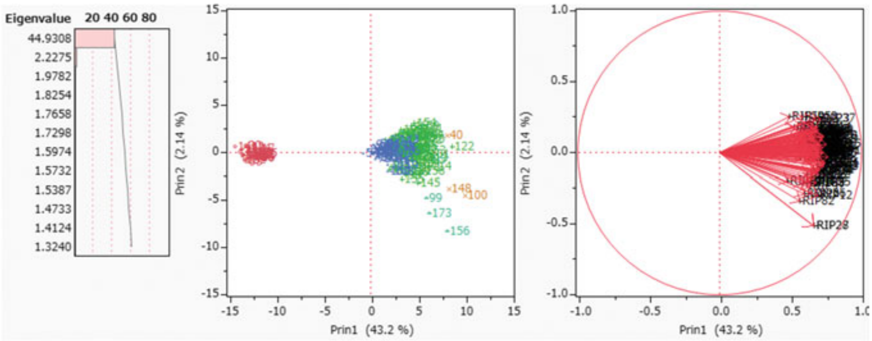


Fig. 8.7 Three plots of PCA (RipDS signal data)

The first columns and second columns of Table 8.6 show the case number corresponding RipDSs signal data and its value of Prin1 axis. The 173 rows have two parts. Upper 36 rows are corresponding to the False class, and lower 137 rows are corresponding to the True class in Fig. 8.7. These two columns are sorted in ascending order from a small value that corresponds from left to right of Prin1. In Fig. 8.7, the leftmost point is the 14th False subject, and the value of Prin1 is −14.28. The 35th False subject has a value of −11.48, which is closest to the True patient in the False case, and 36 cases of false cases are in the range [−14.28, −11.48]. On the other hand, the 54th patient is the nearest to False class, and the 100th patient is far from the False class. Its range is [−0.83, 10.02]. SV opens the window having the width (−11.48, −0.83).

Thus, we can define the RatioSV for PCA in Eq. (8.1).

$$\text{RatioSV of PCA} = (11.48{-}0.83)/(14.28 + 10.02) * 100 = 1065/24.3 = 43.82716\%. \quad (8.1)$$

Assuming that it is about 44%, SV separates two classes such as True patients and False subjects in the remaining 56% range. Because this is the overall characteristic value of RatioSV of 104 RIP, it is larger than the maximum value of RatioSV of 104 RIPs 22.79. In later, we conclude the same results of both RaioSV of PCA by Revised LP-OLDF and HSVM.

Table 8.6 Prin1 values of RIP and Revised LP-OLDF and HSVM sorted by each Prin1 values

RatioSV	43.83	36.90		39.34	
RIP	Prin1	LP	Prin1	HSVM	Prin1
14	**−14.28**	**3**	**−14.06**	**3**	**−16.93**
3	−13.79	31	−13.69	8	−16.59
28	−13.62	25	−13.69	10	−16.39
8	−13.48	10	−13.60	25	−16.34
34	−13.47	11	−13.59	14	−16.02
31	−13.36	8	−13.42	33	−15.75
25	−13.30	6	−13.41	6	−15.59
6	−13.25	33	−13.32	31	−15.43
30	−13.23	9	−13.26	34	−15.41
33	−13.12	23	−13.05	1	−15.36
12	−12.93	1	−12.81	22	−15.27
9	−12.92	14	−12.73	11	−15.25
10	−12.86	32	−12.70	28	−15.05
29	−12.82	30	−12.68	13	−15.05
1	−12.81	22	−12.66	9	−14.87
13	−12.78	34	−12.58	23	−14.78
11	−12.74	29	−12.53	19	−14.76
32	−12.71	13	−12.52	32	−14.67
26	−12.51	28	−12.49	5	−14.61
15	−12.41	5	−12.33	29	−14.51
19	−12.20	19	−12.24	4	−14.47
24	−12.16	12	−12.18	18	−14.45
4	−12.16	18	−12.16	12	−14.43
18	−12.10	20	−12.07	30	−14.38
5	−12.03	24	−11.99	24	−14.24
20	−12.01	26	−11.91	2	−14.23
22	−12.01	16	−11.83	26	−14.18
23	−11.94	4	−11.80	7	−14.17
2	−11.80	7	−11.75	16	−14.16
17	−11.80	2	−11.71	20	−14.12
7	−11.76	15	−11.67	21	−14.12
27	−11.76	36	−11.64	15	−14.04
36	−11.76	35	−11.63	35	−14.01

(continued)

Table 8.6 (continued)

RatioSV	43.83	36.90		39.34	
RIP	Prin1	LP	Prin1	HSVM	Prin1
16	−11.75	21	−11.62	17	−14.00
21	−11.70	27	−11.48	36	−13.99
35	**−11.48**	**17**	**−11.24**	**27**	**−13.95**
54	**−0.83**	**54**	**−2.13**	**54**	**−1.88**
82	−0.34	159	−1.10	82	−1.58
142	0.11	82	−0.90	94	−1.07
79	0.21	94	−0.79	90	−0.98
161	0.22	163	−0.77	108	−0.92
159	0.27	108	−0.77	161	−0.91
94	0.33	111	−0.66	159	−0.76
69	0.44	90	−0.57	142	−0.56
78	0.46	142	−0.51	79	−0.46
64	0.53	64	−0.26	111	−0.36
108	0.54	77	−0.25	77	−0.30
163	0.60	66	−0.08	69	−0.07
74	0.68	161	−0.04	64	0.19
58	0.73	69	0.04	163	0.19
105	0.75	165	0.09	66	0.20
77	0.77	79	0.14	160	0.38
116	0.92	104	0.30	88	0.39
50	1.09	109	0.36	72	0.45
72	1.30	74	0.48	116	0.48
135	1.36	78	0.64	87	0.53
67	1.37	146	0.70	165	0.66
95	1.47	160	0.76	109	0.68
111	1.47	116	0.80	104	0.72
109	1.48	81	0.80	78	0.83
81	1.50	41	0.82	67	0.86
115	1.53	58	0.83	81	0.89
90	1.58	60	0.86	95	0.93
66	1.63	87	0.88	76	1.06
147	1.69	96	0.94	50	1.07
76	1.71	68	0.97	147	1.09
87	1.77	135	1.13	74	1.12
165	1.77	72	1.14	68	1.14
160	1.78	95	1.17	58	1.15

(continued)

Table 8.6 (continued)

RatioSV	43.83	36.90		39.34	
RIP	Prin1	LP	Prin1	HSVM	Prin1
68	1.78	67	1.26	60	1.29
88	1.81	138	1.31	96	1.44
60	1.85	139	1.38	138	1.48
96	1.95	50	1.44	139	1.67
73	1.95	115	1.62	151	1.73
37	2.01	37	1.66	146	1.85
104	2.03	88	1.75	37	1.89
80	2.06	149	1.84	39	1.91
75	2.16	168	1.92	168	1.91
146	2.17	76	1.92	135	2.14
86	2.17	39	1.93	115	2.19
168	2.18	147	1.94	75	2.20
170	2.27	162	2.06	80	2.28
151	2.32	89	2.10	121	2.40
138	2.34	105	2.18	140	2.44
152	2.37	167	2.21	93	2.45
139	2.42	121	2.28	105	2.64
129	2.45	75	2.29	73	2.65
41	2.52	140	2.30	41	2.75
121	2.61	127	2.42	86	2.80
119	2.70	93	2.46	126	2.90
107	2.72	80	2.52	119	3.05
39	2.79	107	2.56	124	3.09
103	2.84	170	2.70	162	3.22
132	2.91	129	2.73	152	3.25
126	2.92	73	2.77	97	3.34
134	2.92	117	2.87	170	3.41
56	3.03	126	2.91	137	3.46
112	3.09	55	2.98	149	3.46
55	3.10	133	3.00	127	3.50
123	3.15	97	3.06	134	3.50
106	3.17	171	3.08	133	3.72
92	3.19	132	3.08	155	3.75
53	3.20	124	3.14	128	3.83
131	3.21	119	3.14	89	3.88
133	3.21	137	3.23	55	3.89

(continued)

Table 8.6 (continued)

RatioSV	43.83	36.90		39.34	
RIP	Prin1	LP	Prin1	HSVM	Prin1
162	3.26	151	3.29	56	3.89
155	3.29	106	3.29	129	3.97
127	3.31	152	3.33	106	4.00
124	3.33	56	3.44	171	4.07
47	3.39	112	3.44	110	4.14
167	3.39	91	3.48	123	4.18
140	3.39	155	3.50	154	4.20
93	3.43	83	3.61	167	4.29
43	3.56	52	3.66	117	4.30
97	3.59	86	3.67	132	4.32
171	3.61	84	3.69	103	4.40
62	3.68	128	3.70	99	4.43
98	3.73	47	3.75	107	4.47
137	3.74	38	3.89	112	4.51
83	3.79	43	4.02	43	4.66
128	3.83	92	4.08	53	4.72
149	3.89	158	4.13	84	4.86
61	3.97	110	4.17	92	4.87
110	3.98	157	4.22	172	4.91
172	3.98	154	4.24	52	4.92
120	4.00	45	4.25	120	4.93
118	4.01	144	4.33	157	4.93
49	4.02	123	4.36	83	4.98
117	4.13	134	4.40	45	5.26
89	4.14	120	4.40	118	5.36
144	4.27	53	4.44	62	5.38
166	4.29	61	4.48	145	5.54
46	4.34	62	4.64	91	5.56
84	4.42	42	4.65	153	5.60
91	4.44	57	4.69	38	5.67
65	4.44	85	4.71	51	5.69
157	4.44	136	4.81	143	5.70
42	4.46	103	4.85	61	5.73
52	4.49	131	4.85	144	5.77
136	4.52	114	4.85	47	5.79
154	4.62	153	4.87	125	5.82

(continued)

Table 8.6 (continued)

RatioSV	43.83	36.90		39.34	
RIP	Prin1	LP	Prin1	HSVM	Prin1
38	4.62	63	4.93	42	5.95
45	4.62	70	4.99	98	5.95
125	4.71	51	5.00	70	6.04
150	4.80	143	5.02	158	6.24
169	4.84	172	5.05	166	6.35
145	4.98	98	5.06	59	6.49
164	5.04	99	5.11	49	6.51
143	5.06	118	5.32	136	6.59
51	5.18	59	5.33	71	6.69
63	5.19	166	5.34	130	6.70
48	5.21	145	5.43	113	6.86
102	5.23	49	6.03	114	6.90
57	5.27	130	6.26	57	6.94
141	5.33	48	6.35	150	7.24
70	5.36	125	6.44	131	7.25
158	5.52	101	6.56	63	7.31
130	5.53	71	6.64	164	7.47
153	5.58	164	6.92	101	7.66
59	5.65	44	6.93	85	7.75
113	5.76	141	6.97	48	8.01
71	5.78	150	6.97	44	8.03
99	5.91	46	7.05	141	8.28
101	5.98	169	7.07	46	8.73
44	6.12	113	7.15	169	8.81
173	6.19	122	7.41	65	9.28
114	6.21	65	7.62	122	9.45
85	6.33	173	8.04	102	9.53
40	8.09	156	8.59	40	10.00
156	8.15	40	8.69	156	10.26
148	8.30	102	8.95	173	11.12
122	8.62	100	9.55	148	13.05
100	**10.02**	**148**	**10.03**	**100**	**13.75**

Figure 8.8 shows the result of LpDSs signal data by PCA. The first eigenvalue is 46.356, and the contribution ratio is 44.6%. The second eigenvalue is 2.214, the contribution ratio is 2.13%, and the cumulative contribution ratio is 46.73%. That is, the Prin1 almost presents 173 subjects. Although the score plot shows several outliers as same as in Fig. 8.7, the Prin1 becomes an indicator of cancer malignancy as same as 104 LpDSs. The third and fourth columns of Table 8.6 show the result of LpDSs. The ranges of False class and True class are [−14.06, −11.24] and [−2.13, 10.03]. SV opens the window that is the interval (−11.24, −2.13). RatioSV of PCA by LpDSs is Eq. (8.2).

$$\begin{aligned}\text{RatioSV of PCA by LpDSs} &= (11.24-2.13)\,/\,(14.06+10.63) * 100 \\ &= 9.11 * 100/24.69 = 36.89753\,\% \end{aligned} \tag{8.2}$$

Because the maximum RatioSV of LpDSs is 21.81, RatioSV of PCA becomes a malignancy index.

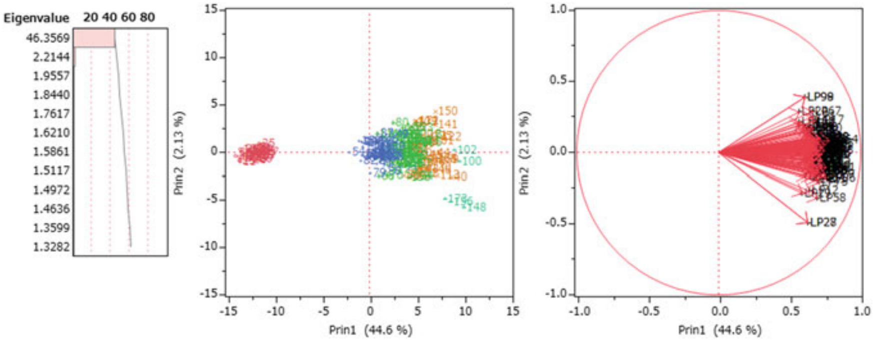


Fig. 8.8 Three plots of PCA (LpDS signal data)

Figure 8.9 shows the result of HsvmDSs signal data. The first eigenvalue is 66.039, and the contribution ratio is 63.5%. The second eigenvalue is 1.619, the contribution ratio is 1.56%, and the cumulative contribution ratio is 65.06%. That is, the Prin1 almost presents 173 subjects. The score plot shows several outliers as same as in Fig. 8.8. Because the second eigenvalue is small and the variation is small, the False subjects are on the axis of −13.95 or less of the Prin1. In other words, the Prin1 becomes a malignancy indicator as same as 104 HsvmDSs. The fifth and sixth columns of Table 8.6 show the result of HsvmDSs. The ranges of False class and True classes are [−16.93, −13.95] and [−1.88, 13.75], respectively. SV opens the window that is the interval (−13.95, −1.88). RatioSV of PCA by HsvmDSs is Eq. (8.3).

$$\text{RatioSV of PCA by HsvmDSs} = (13.95-1.88)/(16.93+13.75) * 100 = 39.34159\,\% \tag{8.3}$$

Because the maximum RatioSV of HsvmDSs is 28.74, RatioSV of PCA is helpful as a malignancy index.

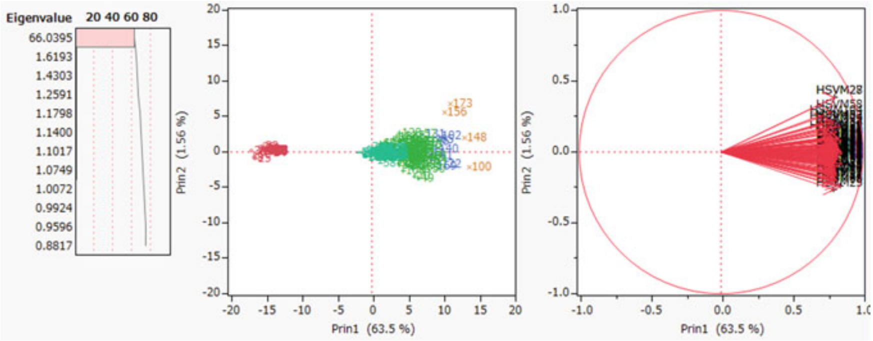


Fig. 8.9 Three plots of PCA (HsvmDS signal data)

8.4.3 *PCA of Transpose Signal Data*

We transpose the RipDSs signal data and analyze this transposed data with 104 RipDSs (104 cases) and 173 patients (173 variables). Figure 8.10 is three plots of PCA. Because the first eigenvalue is 5.447 and contribution ratio is 3.15%, Prin1 explains only 3.15% variance. This fact indicates us that 104 RipDSs play almost the same role in the transposed data. Thus, the factor loading plot shows all absolutes of correlation coefficients with Prin1 and Prin2 are less 0.5. We guess other absolute correlations with other principal components may be less 0.5 also. Scatter plot suggests us there are many outliers in the four quadrants. Although there are many outliers in scatter plots, these outliers are considered to represent a unique malignancy index independent from others.

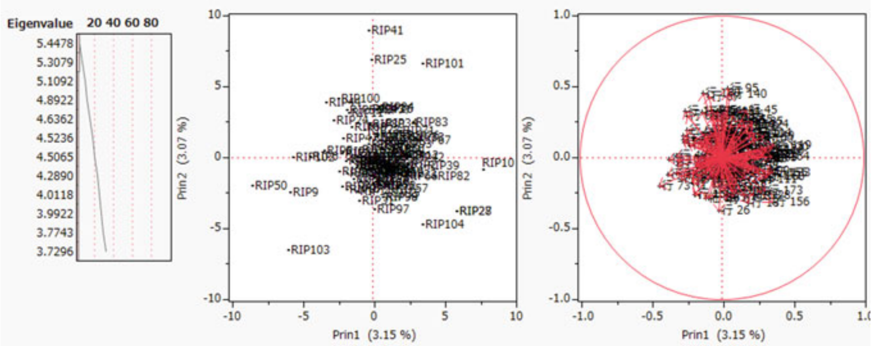


Fig. 8.10 Three plots of PCA (RipDS data)

We analyze transpose signal data made by 104 LpDSs. Figure 8.11 is three plots of PCA. Because the first eigenvalue is 11.678 and contribution ratio is 6.75%, Prin1 explains only 6.75% variance. This fact indicates us that 104 LpDSs play almost the same role. Thus, the factor loading plot shows all absolutes of correlation coefficients with Prin1 and Prin2 are less 0.8. We guess other absolute correlations with other principal components may be less 0.8 also. Scatter plot suggests us two different outliers such as (LP104) and (LP99). We expect two gene pairs included in (SM104) and (SM99) are the "new class of cancer subsets" pointed out by Golub et al.

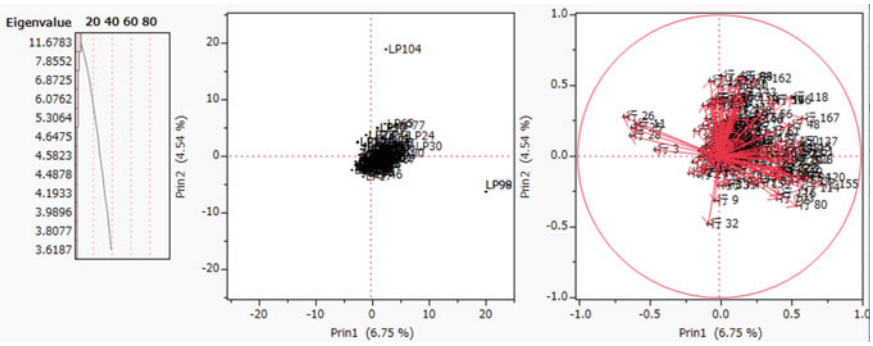


Fig. 8.11 Three plots of PCA (LpDS data)

We analyze the transpose data made by 104 HsvmDSs. Figure 8.12 is three plots of PCA. Because the first eigenvalue is 6.064 and contribution ratio is 3.51%, Prin1 explains only 3.51% variance. This fact indicates us that 104 HsvmDSs play almost the same role. Thus, the factor loading plot shows all absolutes of correlation coefficients with Prin1 and Prin2 are less 0.5. We guess other absolute correlations with other principal components may be less 0.5 also. Scatter plot suggests us there are many outliers belonging in the first and fourth quadrants such as (HSVM6, HSVM12, HSVM34, HSVM41, HSVM51, HSVM74, HSVM104) and (HSVM1, HSVM2, HSVM27, HSVM28, HSVM32, HSVM102). We expect seven and six gene pairs included in (SM6, SM12, SM34, SM41, SM51, SM74, SM104) and (SM1, SM2, SM27, SM28, SM32, SM102) are the same "new class of cancer subsets" pointed by Golub et al.

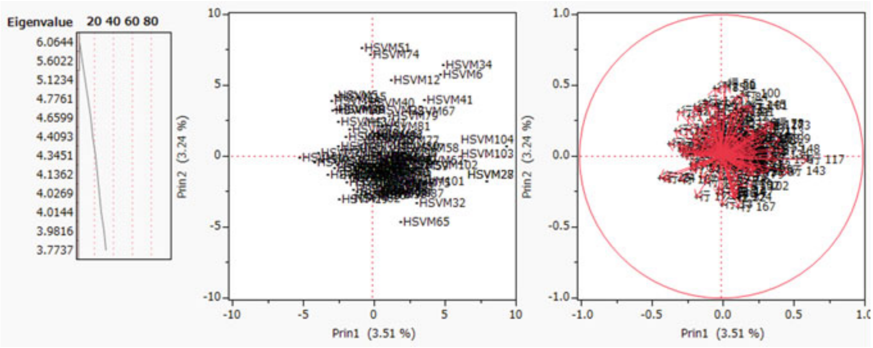


Fig. 8.12 Three plots of PCA (HsvmDS data)

8.5 Conclusions

In Chaps. 3 and 4, we examine Alon's microarray from the various angles of cancer gene diagnosis. After Chap. 5, we examine the other five microarrays from the viewpoints proposed in Chap. 4. Only two classes of Alon and Singh are the healthy subjects and cancer patients. The remaining four microarrays consist of different cancers. However, it is vital that the results of all SMs obtained by the RIP and Revised LP-OLDF are almost the same. Perhaps, if medical projects collect data for research purposes, we believe that the two classes in the microarray are LSDs (Fact3) and many SMs (Fact4) show almost the same results explained in this book. In other words, we believe that microarray provides useful information for cancer diagnosis. Furthermore, the LSD has a Matryoshka structure, and Method2 is valid even for general data. Our research is considered to be equally useful for data such as other high-dimensional data and common data. If researchers create multiple SMs with RIP and Revised LP-OLDF, they can quickly analyze by standard statistical analysis by creating signal data using these SMs. Because statistical discriminant methods were useless at all, Problem5 did not succeed. Moreover, the doctors had no choice but to develop analytical methods themselves. In addition to their methods, we believe that using a statistical method will open up a new world of cancer gene diagnosis.

In this chapter, although RIP and Revised LP-OLDF find two different SMs, we show the results of 104 SMs found by Revised LP-OLDF using the correlation analysis and explain the results of three signal data made by RIP, Revised LP-OLDF and H-SVM. Furthermore, cluster analysis and PCA analyze three signal data made by RipDSs, LpDSs, and HsvmDSs. We omit the many results of three signal data made by RipDSs, LpDSs, and HsvmDSs. The outline of these results is almost the same as other chapters. This fact means that six types of signal data are signals. Also, only RIP and Revised LP-OLDF can extract signals from noise. This is the gospel for researchers of cancer genetic diagnosis. A simple analysis method proposed in this book gives a large amount of information. Researchers can verify those results by real patients. We expect many people to contribute to cancer diagnosis.

References

Sall JP, Creighton L, Lehman A (2004) JMP start statistics (3rd edn). SAS Institute Inc. USA (Shinmura S. edits Japanese version)

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2016) New Theory of Discriminant Analysis after R. Fisher. Springer, Tokyo

Shinmura S (2017) From cancer gene analysis to cancer gene diagnosis. Amazon

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD'18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. In: Biocomp'18, pp 1–7

Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD (2003) The role of the Wnt-signaling Antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. N Engl J Med 349(26):2483–2494

Chapter 9
Cancer Gene Diagnosis of Chiaretti et al. Microarray

Abstract This chapter introduces the cancer gene diagnosis of Chiaretti microarray that consists of 128 patients and 12,625 genes. RIP finds 128 SMs, and Revised LP-OLDF finds 124 SMs. We confirm the defect of Revised LP-OLDF, also. Because both SMs are almost the same results, we introduce only the results of 124 SMs. In Sect. 9.2, we confirm the 7,626 correlations of 124 LpDSs are greater than 0.359 and standard statistical methods cannot find the linear separable facts of SMs. Thus, we conclude three signal data made by RIP, Revised LP-OLDF, and H-SVM are the better definition of the signal instead of SMs. Also, we explain how to build 124 LpDSs. In Sect. 9.3, the 124 SMs are evaluated by RatioSVs of six MP-based LDFs and NMs of statistical discriminant functions. In Sect. 9.4, five hierarchical cluster methods analyze three signal data of 124 RipDSs, LpDSs, and HsvmDSs. In Sects. 9.5 and 9.6, PCA analyzes signal data and transposed signal data. Section 9.7 concludes six microarrays have almost the same results. We believe that the consistency of these results confirms the reliability of cancer gene diagnosis.

Keywords Chiaretti microarray · Cancer gene diagnosis · Malignancy indexes · Small Matryoshka (SM) · RIP discriminant scores (RipDSs) · Signal data made by RipDSs · Hierarchical clustering · Principal component analysis (PCA)

Thanks to Chiaretti et al.

We thank Chiaretti et al. (2004) for providing excellent data. Below, we will quote "Abstract" for the reader.

> Gene expression profiles were examined in 33 adult patients with T-cell acute lymphocytic leukemia (T-ALL). Nonspecific filtering criteria identified 313 genes differentially expressed in the leukemic cells. Hierarchical clustering of samples identified two groups that reflected the degree of T-cell differentiation but was not associated with clinical outcome. Comparison between refractory patients and those who responded to induction chemotherapy identified a single gene, interleukin 8 (IL-8), that was highly expressed in refractory T-ALL cells and a set of 30 genes that were highly expressed in leukemic cells from patients who achieved complete remission. We next identified 19 genes that were differentially expressed in T-ALL cells from patients who either had a relapse or remained in continuous complete remission. A model based on the expression of 3 of these genes was predictive of duration of remission.

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_9

The 3-gene model was validated on a further set of T-ALL samples from 18 additional patients treated with the same clinical protocol. This study demonstrates that gene expression profiling can identify a limited number of genes that are predictive of response to induction therapy and remission duration in adult patients with T-ALL.

9.1 Introduction

Chapter 1 introduces the "New Theory of Discriminant Analysis after R. Fisher (Shinmura 2016)" and the cancer gene analysis as an applied problem of the Theory. Revised IP-OLDF (RIP) based on the minimum number of misclassifications (MNM) criterion can find six microarrays are linearly separable data (LSD; MNM = 0) that is new Fact3. Matryoshka feature selection method (Method2) and LINGO Program3 can decompose microarray into many Small Matryoshkas (SMs) and noise subspace (MNM > 0) that is new Fact4 (Schrage 2006). Chapter 2 introduces the cancer gene diagnosis and malignancy indexes using all SMs found by the RIP in 2016. Furthermore, we compare the RatioSVs of 64 SMs and 130 basic gene sets (BGSs) and conclude the malignancy indexes of SMs are better than those of BGSs. In Chap. 3, we propose how to choose the proper SMs and compare two different types of SMs found by the RIP and Revised LP-OLDF. Chapter 4 evaluates signal subspace made by the union of SMs and noise subspace. We find the defect of Revised LP-OLDF that cannot find all SMs from six microarrays. All correlations of RIP discriminant scores (RipDSs) are positive values, but correlations of genes included in SM take the positive values, almost the zero, and the negative values. Thus, we reconsider that signal data made by RipDSs, Revised LP-OLDF DSs (LpDSs), and H-SVM DSs (HsvmDSs) are true signal subspaces instead of SM itself (Shinmura 2018a, b). In Chaps. 5 and 6, we confirmed our results by SMs found by RIP using Golub and Shipp microarrays. In Chaps. 7 and 8, we confirmed the research results by SM found in Revised LP-OLDF using Singh et al. (2002) and Tian's microarrays.

In Chap. 9, we introduce the cancer gene diagnosis of Chiaretti microarray that consists of 128 patients (95 B-cell patients and 33 T-cell patients) and 12,625 genes. RIP finds 128 SMs (4,907 genes) and Revised LP-OLDF finds 124 SMs (4,552 genes). We confirm the defect of Revised LP-OLDF as same as other microarrays. Because both SMs are almost the same results, we introduce only the results of 124 SMs founded by Revised LP-OLDF. In Sect. 9.2, we confirm the 7,626 correlations of 124 LpDSs are greater than 0.359 and standard statistical methods cannot find the linear sparable facts of SMs. Thus, we conclude three signal data made by RipDSs, LpDSs, and HsvmDSs are the better definition of signal instead of SMs. Also, we explain how to build 124 LpDSs. In Sect. 9.3, the 124 SMs are evaluated by RatioSVs of six MP-based LDFs and NMs of statistical discriminant functions. In Sect. 9.4, five hierarchical clustering methods analyze three signal data of 124 RipDSs, LpDSs, and HsvmDSs. In Sects. 9.5 and 9.6, PCA analyzes signal data and transposed signal data. Section 9.7 concludes six microarrays have the almost same results. We believe that the consistency of these results confirms the reliability of cancer gene diagnosis. Researchers can obtain results with short research time by our approach. LINGO (Schrage 2006) decomposes microarray into many SMs and opens a new frontier of cancer gene analysis. JMP (Sall et al. 2004) analyzes all SMs and offers cancer gene diagnosis. Shinmura (2016, 2017, 2018a, b) relate to this Chapter.

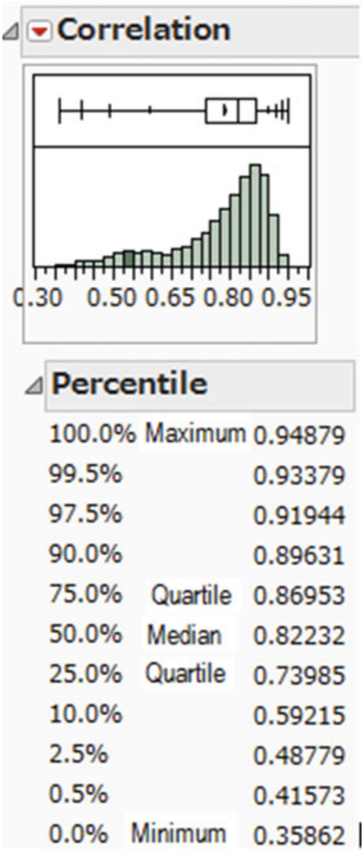
9.2 Examination of Discriminant Scores of 124 SMs Found by Revised LP-OLDF

Only RIP and Revised LP-OLDF can decompose microarrays into many SMs (Fact4). Although Revised LP-OLDF cannot find all SMs from microarrays, we use the 124 SMs. However, we discriminate the 124 SMs by RIP, Revised LP-OLDF, and H-SVM and make three signal data made by RipDSs, LpDSs, and HsvmDSs. We analyze three signal data by standard statistical methods and obtain the surprising results.

9.2.1 Correlation of 124 LpDSs

Figure 9.1 is the histogram of 7,626 correlations (abbreviated R) made by 124 LpDSs instead of 124 RipDSs because 124 RipDSs are almost the same results. JMP analyzes all statistical methods (Sall et al. 2004). The range of correlations is [0.358, 0.948]. Although correlations of Golub, Singh, and Tian have the “R = 1,” Chiaretti has not the “R = 1.” Thus, we cannot focus on the pairs of SMs with R = 1. Because the range of Golub, Singh, and Tian are [0.069, 0.880], [0.417, 0.895], and [0.133, 0.6] after omitting R = 1, the range of Chiaretti’s correlations is similar to Singh’s range. Although the microarray of Chiaretti contains many genes, it seems that there are no genes pairs to replace each other from the viewpoint of high correlation pairs (R = 1). Our claim is necessary to validate by medical specialists.

Fig. 9.1 Histogram of 7,626 correlations by 124 LpDSs



Percentile

100.0%	Maximum	0.94879
99.5%		0.93379
97.5%		0.91944
90.0%		0.89631
75.0%	Quartile	0.86953
50.0%	Median	0.82232
25.0%	Quartile	0.73985
10.0%		0.59215
2.5%		0.48779
0.5%		0.41573
0.0%	Minimum	0.35862

Table 9.1 is the list of 7,626 correlations sorted by descending order of R. The [2.5%, 97.5%] is the 95% confidence interval of each R. Because all p-values are less than equal to 0.000 ($p < 0.0005$), all correlations are positive. This fact is one of the reasons we consider the signal data is signal instead of SMs. On the other hand, the correlations of genes included in SM take the positive, almost zero, and negative values. Moreover, standard statistical methods cannot find the linear separable facts.

Table 9.1 List of 7,626 correlations sorted by descending order of r

Var1	Versus Var2	Correlation	n	2.5%	97.5%	p-value
LP30	LP24	0.949	128	0.928	0.964	0.000
LP20	LP19	0.948	128	0.927	0.963	0.000
LP2	LP1	0.947	128	0.925	0.962	0.000
LP28	LP24	0.942	128	0.919	0.959	0.000
LP18	LP1	0.942	128	0.919	0.959	0.000
LP14	LP1	0.941	128	0.917	0.958	0.000
LP18	LP14	0.940	128	0.917	0.958	0.000
LP19	LP10	0.940	128	0.916	0.958	0.000
LP25	LP10	0.940	128	0.916	0.957	0.000
LP41	LP19	0.939	128	0.914	0.957	0.000
–	–	–	–	–	–	–
LP124	LP44	0.390	128	0.233	0.528	0.000
LP123	LP3	0.390	128	0.232	0.528	0.000
LP124	LP4	0.379	128	0.220	0.518	0.000
LP124	LP57	0.376	128	0.216	0.516	0.000
LP123	LP12	0.374	128	0.215	0.514	0.000
LP123	LP13	0.374	128	0.214	0.514	0.000
LP124	LP1	0.372	128	0.213	0.513	0.000
LP124	LP87	0.372	128	0.212	0.512	0.000
LP124	LP12	0.371	128	0.211	0.512	0.000
LP124	LP55	0.359	128	0.197	0.501	0.000

9.2.2 *PCA Analysis of Signal Data Made by 124 LpDSs*

We analyze the signal data made by 124 LpDSs by PCA showed in Fig. 9.2. The eigenvalue of Prin1 is 98.578, and the contribution rate is 79.5%. The eigenvalue of Prin2 is 5.121, and the contribution rate is 4.13%. The cumulative rate is 83.65%. Thus, two principal components explain the 83.65% of total variance, and 30 principal components explain the 95.135% of total variance. We check 29 scatter plots. All x-axes are Prin1, and y-axes are from Prin2 to Prin30. All two classes are separate entirely. Thus, we consider the Prin1 is better malignancy index of PCA. The scatter plot shows the 91st patient in B-cell class and the 110th and 115th patients in T-cell class are the outliers. Although the factor loading plot shows that 124 correlations of 124 LpDSs and Prin1 are higher than 0.5, 124 correlations of 124 LpDSs and Prin2 are range from −0.25 to 0.7.

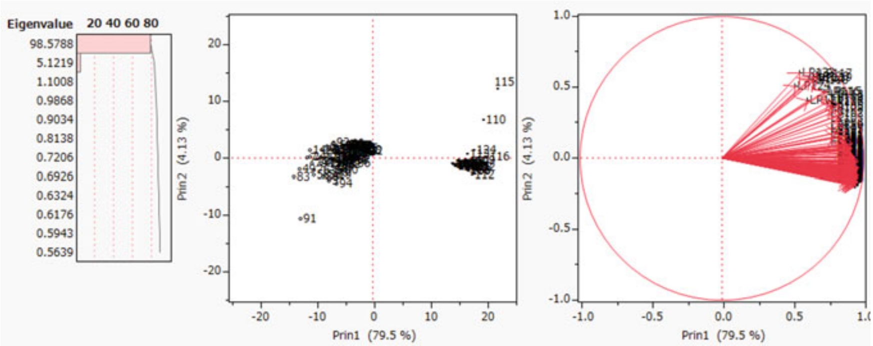


Fig. 9.2 PCA output of the 124 LpDSs

9.2.3 How to Categorize 124 LpDSs

We conclude only RIP and Revised LP-OLDF can decompose microarrays into many SMs (Fact4). Although we find the defect of Revised LP-OLDF that cannot find all SMs from the microarray, we examine the possibilities of SMs found by Revised LP-OLDF in Chaps. 7 and 8. In Chap. 9, we examine the possibilities of three signal data made by RipDSs, LpDSs, and HsvmDSs using 124 SMs found by Revised LP-OLDF. Our results indicate the proper discriminant functions such as the above three LDFs are the best methods for cancer gene diagnosis using two different types of SMs found by the RIP and Revised LP-OLDF. With the breakthrough of signal data, we can succeed to obtain the three different types of 124 malignancy indexes and open the door of cancer gene diagnosis. Section 9.2.3 examines how to construct 124 LpDSs. The second and third columns of Table 9.2 show the subject's SM taking the minimum and maximum values in each LpDS. Because the minimum number of LpDSs is chosen from the patient of B-cell class, the selected subject is considered to be in the typical one of B-cell. The maximum number of LpDSs is the typical patient of T-cell class. The fourth column is the LpDSs (abbreviated LPi in the figure). The fifth column is the range of LpDSi, and the last column is RatioSV of each LpDSi. We omit the 114 rows from SM6 to SM119.

Table 9.2 Minimum and maximum subject's SM and RatioSV

SM	B-cell	T-cell	LpDS	Range	RatioSV
SM1	91	122	LpDS1	6.658	30.040
SM2	49	99	LpDS2	8.346	23.964
SM3	74	106	LpDS3	7.057	28.342
SM4	33	106	LpDS4	5.856	34.154
SM5	84	122	LpDS5	14.373	13.915
Omitted	114 SMs				
SM120	91	115	LpDS120	41.585	4.809
SM121	89	113	LpDS121	53.404	3.745
SM122	56	115	LpDS122	59.840	3.342
SM123	58	115	LpDS123	112.889	1.772
SM124	60	115	LpDS124	1345.115	0.149

Table 9.3 shows how to categorize 124 LpDSs. Roughly speaking, we consider the combinations of the minimum and maximum LpDS's values specify 124 LpDSs. The left four columns (and right four columns) are sorted in ascending order by the second column value (B-cell) as the first sort key and the third T-cell value as the second sort key. The second column shows the B-cell's patient taking the minimum value of LpDS. The third column shows the T-cell patient taking the maximum value of LpDS. The first patient (second column) and the 128th patient (third column) included in both SM81 and SM88 (first column) take the same minimum value 1 and the maximum value 128 in two SMs such as SM81 and SM88. The fourth column shows the pair number. Because two SMs, such as SM81 and SM88, take the same minimum and maximum subjects, the pair number is two. Even though Table 9.4 shows SM81 includes 41 genes and SM88 includes 49 genes, LpDS81 and LpDS88 are almost the same malignancy indexes and may have the same effect in cancer gene diagnosis. We guess two different gene sets have the same role in cancer gene diagnosis and are redundant with each other. There is one group of seven SMs with the same pair. There are one set of four SMs that have the same pair. There are 11 sets in which two SMs have the same pair. As the other 91 SMs do not have pairs, they are considered to be useful for cancer gene diagnosis of one another.

Future Theme: Although RIP and Revised LP-OLDF find many SMs, these SMs will be classified meaningfully for cancer gene diagnosis.

Table 9.3 Categories of 124 LpDS

SM	B-cell1	T-cell1	Pair	SM	B-cell1	T-cell1	Pair
SM81	1	128	2	SM17	49	123	
SM88	1	128		SM14	49	124	2
SM90	6	100		SM68	49	124	
SM57	6	101		SM51	49	125	
SM102	6	110		SM52	49	127	
SM77	6	112		SM21	49	128	
SM43	6	113		SM33	54	117	
SM29	7	99	2	SM12	54	126	
SM37	7	99		SM63	56	105	
SM35	7	112	2	SM112	56	115	2
SM56	7	112		SM122	56	115	
SM42	7	116		SM123	58	115	
SM23	7	120		SM9	60	96	
SM87	7	124		SM124	60	115	
SM32	11	112		SM30	70	116	
SM73	11	121		SM101	71	115	
SM58	11	127		SM60	73	123	
SM13	14	99		SM3	74	106	
SM104	14	100		SM50	76	102	
SM19	14	109		SM46	76	112	
SM111	14	110		SM7	76	118	
SM44	14	112		SM20	83	100	
SM105	14	115		SM69	83	101	
SM40	14	116	2	SM74	83	103	
SM66	14	116		SM49	83	105	
SM28	14	121		SM53	83	106	
SM6	14	122		SM64	83	108	
SM91	14	123		SM18	83	109	2
SM55	14	125		SM98	83	109	
SM48	14	127		SM70	83	113	
SM26	15	97		SM38	83	115	4
SM22	15	116		SM95	83	115	
SM80	24	105		SM103	83	115	
SM16	24	113		SM115	83	115	
SM107	24	115		SM34	83	116	
SM10	25	108		SM83	83	122	
SM85	25	110		SM59	83	123	

(continued)

Table 9.3 (continued)

SM	B-cell1	T-cell1	Pair	SM	B-cell1	T-cell1	Pair
SM41	25	112		SM71	83	124	
SM27	29	112		SM15	84	112	
SM99	30	116		SM5	84	122	
SM4	33	106		SM8	85	120	
SM110	33	110		SM100	89	100	
SM84	37	101		SM11	89	108	
SM93	37	115		SM96	89	110	
SM65	39	101		SM121	89	113	
SM62	39	112		SM108	91	110	2
SM39	39	116		SM114	91	110	
SM54	45	112		SM106	91	115	7
SM78	48	115		SM109	91	115	
SM24	48	116		SM116	91	115	
SM2	49	99	2	SM117	91	115	
SM82	49	99		SM118	91	115	
SM61	49	101		SM119	91	115	
SM25	49	106		SM120	91	115	
SM72	49	110		SM75	91	119	
SM47	49	112	2	SM1	91	122	
SM67	49	112		SM113	91	124	
SM76	49	115		SM31	92	102	
SM79	49	116	2	SM45	93	116	
SM86	49	116		SM89	94	110	
SM92	49	121		SM97	94	120	
SM36	49	122		SM94	94	124	

9.3 Validation of 124 SMs by Six MP-Based LDFs and Discriminant Functions

Chiaretti microarray consists of 128 cases and 12,625 genes. RIP of LINGO Program3 found the 269 SMs (5,220 genes) and the noise gene subspace (7,405 genes) in 2015. Because the survey of 269 SMs requested huge research time, we did not analyze the 239 SMs until now. However, when Revised LP-OLDF decomposes Chiaretti's microarray again in 2018, 124 SMs (4,552 genes) are found. We obtain fewer SMs and genes in 2018. Thus, we analyze 124 SMs by standard statistical methods. However, we cannot find linear separable facts as same as other five microarrays. Next, we discriminate the 124 SMs by RIP, Revised LP-OLDF, and H-SVM. RatioSVs evaluate three signal data made by RipDSs, LpDSs, and

HsvmDSs. The last, we compare and examine the results of the cluster analysis and PCA.

Table 9.4 shows the 124 SMs from SM = 1 to SM = 124. The "gene" column is the number of genes included in each SM. The range of genes included in the 124 SMs is [6, 77]. The gene number of those is less than case number n (128). The average is 36.71. Row "SUM" indicates 124 SMs contain 4,552 genes. We compare six RatioSVs of six MP-based LDFs. From RIP column to H-SVM columns show three RatioSVs of 124 SMs by RIP, Revised LP-OLDF, and H-SVM. Three ranges of RatioSV are [0.15, 42.46], [0.17, 39.69], and [0.17, 45.81], respectively. Three averages of RatioSVs are 26.68%, 24.16%, and 26.7%, respectively. Row "MaxRatio" indicates the number of the maximum RatioSVs of 124 SMs those are 64, 16, and 27, respectively. The Revised IPLP-OLDF and the two S-SVMs take 17 maximum-values. Two columns "MAX and MIN" are the maximum and minimum values of six LDFs. In this chapter, we omit Revised IPLP-OLDF and two soft-margin SVM such as SVM4 (penalty c = 10,000) and SVM1 (penalty c = 1). To summarize these results, the range of H-SVM is slightly better than RIP because the maximization SV of H-SVM work well. On the other hand, the average of RIP are better than H-SVM.

Because all NMs of logistic regression are zero and 124 SMs are linearly separable, we omit this column from the table. Four columns "SVM4, SVM1, LDF2, and QDF" show the number of misclassifications (NM). SVM4 and QDF cannot discriminate one SM correctly. SVM1 and LDF2 cannot discriminate 23 and 21SMs correctly, respectively. The prior probability of LDF2 is proportional to the case number of 33:95. Because many NMs are 0, we can see that in the many SMs of Chiaretti, the two groups are somewhat separable. However, even in such SMs, cluster analysis and PCA cannot find linear separable facts. In other words, only RIP and Revised LP-OLDF are most suitable for analysis of microarray and SM. Moreover, H-SVM is suitable for SM. SVM4 and QDF are better than cluster analysis and PCA.

Table 9.4 Summary of six RatioSVs of six MP-based LDFs and NMs of other discriminant functions

SM	Gene	RIP	LP	HSVM	MAX	MIN	SVM4	SVM1	LDF2	QDF
1	6	29.85	26.58	**36.30**	36.30	26.58	0	0	0	0
2	10	33.66	25.89	38.59	40.44	25.89	0	0	0	0
3	10	29.07	28.84	28.40	30.62	28.40	0	0	0	0
4	11	26.02	**33.85**	31.22	33.85	26.02	0	0	0	0
5	8	27.90	15.62	**32.66**	32.66	15.62	0	0	0	0
6	13	**31.62**	23.07	24.29	31.62	22.00	0	0	0	0
7	14	33.19	31.12	31.74	35.02	31.12	0	0	0	1
8	18	39.73	37.20	**45.81**	45.81	36.71	0	0	0	0

(continued)

Table 9.4 (continued)

SM	Gene	RIP	LP	HSVM	MAX	MIN	SVM4	SVM1	LDF2	QDF
9	15	33.14	34.01	**34.72**	34.72	33.14	0	0	0	0
10	13	**31.00**	30.79	28.37	31.00	28.37	0	0	0	0
11	15	36.16	27.24	34.66	38.17	27.24	0	0	0	0
12	22	32.92	28.26	33.26	39.08	28.26	0	0	0	0
13	16	**34.95**	25.89	31.43	34.95	25.89	0	0	0	0
14	14	33.41	34.88	**36.48**	36.48	33.41	0	0	0	0
15	18	28.32	27.36	33.97	35.65	27.36	0	0	0	0
16	17	26.45	27.19	27.80	30.57	26.45	0	0	0	0
17	18	28.26	28.49	**34.54**	34.54	28.21	0	0	0	0
18	24	33.86	**38.56**	37.15	38.56	33.86	0	0	0	0
19	24	38.59	**39.69**	36.48	39.69	36.48	0	0	0	0
20	25	**41.95**	36.03	36.73	41.95	32.13	0	0	0	0
21	23	38.54	33.85	35.42	40.50	33.85	0	0	0	0
22	23	34.83	37.73	**40.45**	40.45	34.83	0	0	0	0
23	16	**25.53**	25.23	23.47	25.53	23.47	0	0	0	0
24	18	**35.24**	35.16	32.32	35.24	29.80	0	0	0	0
25	24	27.95	36.76	**38.89**	38.89	27.95	0	0	0	0
26	21	27.63	24.70	**32.58**	32.58	24.70	0	0	0	0
27	21	**26.30**	25.37	19.29	26.30	19.29	0	0	0	0
28	29	**39.01**	38.47	34.88	39.01	34.88	0	0	0	0
29	24	**35.07**	22.87	35.01	35.07	22.87	0	0	0	0
30	28	38.69	38.64	38.40	39.59	38.40	0	0	0	0
31	26	35.70	**36.71**	34.30	36.71	32.05	0	0	0	0
32	25	**30.04**	27.71	27.11	30.04	25.33	0	0	0	0
33	27	**32.65**	27.78	28.11	32.65	27.78	0	0	0	0
34	26	22.88	21.65	**29.37**	29.37	21.65	0	0	0	0
35	32	**28.29**	28.06	27.50	28.29	26.30	0	0	0	0
36	32	36.75	33.28	35.85	37.47	33.28	0	0	0	0
37	31	**34.97**	33.28	33.42	34.97	30.76	0	0	0	0
38	29	27.29	30.04	29.60	32.19	27.29	0	0	0	0
39	24	**25.89**	20.65	24.53	25.89	20.65	0	0	0	0
40	32	**42.46**	35.70	38.10	42.46	33.33	0	0	0	0
41	27	34.31	**36.84**	34.41	36.84	34.31	0	0	0	0
42	27	**32.64**	31.58	31.28	32.64	31.23	0	0	0	0
43	32	**39.88**	27.34	31.98	39.88	26.81	0	0	0	0
44	34	**33.82**	32.73	32.45	33.82	32.45	0	0	0	0
45	32	**41.73**	36.62	35.95	41.73	35.95	0	0	0	0

(continued)

Table 9.4 (continued)

SM	Gene	RIP	LP	HSVM	MAX	MIN	SVM4	SVM1	LDF2	QDF
46	33	26.18	28.91	**29.69**	29.69	24.63	0	0	0	0
47	31	**37.44**	34.68	34.30	37.44	24.84	0	0	0	0
48	36	31.93	**32.11**	30.48	32.11	30.48	0	0	0	0
49	31	**33.99**	25.12	30.10	33.99	25.12	0	0	0	0
50	30	**37.79**	31.19	33.24	37.79	25.44	0	0	0	0
51	36	27.35	**27.54**	24.78	27.54	24.78	0	0	0	0
52	33	33.14	33.32	33.00	35.76	33.00	0	0	0	0
53	33	**28.89**	22.24	22.90	28.89	22.24	0	0	0	0
54	34	**36.47**	31.40	34.83	36.47	23.12	0	0	0	0
55	36	**33.52**	27.67	32.82	33.52	27.67	0	0	0	0
56	29	**30.40**	28.86	28.20	30.40	26.22	0	0	0	0
57	28	**31.78**	30.24	28.18	31.78	22.51	0	0	0	0
58	36	**37.65**	27.92	27.86	37.65	27.86	0	0	0	0
59	31	23.03	25.12	**25.42**	25.42	23.03	0	0	0	0
60	35	**29.76**	28.40	26.85	29.76	18.56	0	0	0	0
61	33	33.16	25.23	26.57	35.35	25.23	0	0	0	0
62	34	**37.68**	29.95	27.44	37.68	27.44	0	0	0	0
63	34	**29.26**	22.61	22.07	29.26	21.82	0	0	0	0
64	39	**27.63**	20.30	27.33	27.63	20.30	0	0	0	0
65	31	**28.56**	17.36	21.46	28.56	17.36	0	0	0	0
66	40	**29.61**	25.27	25.66	29.61	21.54	0	0	0	0
67	43	**30.76**	23.11	25.03	30.76	23.11	0	0	0	0
68	37	27.18	**28.20**	27.00	28.20	27.00	0	0	0	0
69	40	**37.25**	30.54	32.43	37.25	30.54	0	0	0	0
70	38	**35.24**	22.25	28.37	35.24	22.25	0	0	0	0
71	39	**34.78**	28.30	30.60	34.78	28.30	0	0	0	0
72	35	19.82	21.67	**23.64**	23.64	19.82	0	0	0	0
73	36	**28.90**	24.23	26.01	28.90	24.23	0	0	0	0
74	41	**32.03**	28.01	30.52	32.03	27.67	0	0	0	0
75	43	**33.55**	27.11	29.86	33.55	27.11	0	0	0	0
76	38	**21.54**	16.83	16.55	21.54	16.55	0	0	0	0
77	40	**28.30**	23.09	27.24	28.30	23.09	0	0	0	0
78	46	29.44	**33.26**	32.98	33.26	29.44	0	0	0	0
79	39	**29.18**	25.93	28.50	29.18	25.87	0	0	0	0
80	40	23.85	23.43	**26.42**	26.42	23.43	0	0	0	0
81	41	**25.83**	21.48	23.71	25.83	21.48	0	0	0	0
82	43	23.95	**25.76**	23.76	25.76	22.66	0	0	0	0

(continued)

Table 9.4 (continued)

SM	Gene	RIP	LP	HSVM	MAX	MIN	SVM4	SVM1	LDF2	QDF
83	43	23.54	21.51	22.25	23.97	21.51	0	0	0	0
84	38	**24.29**	22.51	23.16	24.29	20.53	0	0	0	0
85	38	**28.68**	26.04	26.37	28.68	21.37	0	0	0	0
86	41	**27.77**	25.31	23.58	27.77	22.32	0	0	0	0
87	46	21.33	15.26	**24.89**	24.89	15.26	0	0	0	0
88	49	**25.22**	20.89	21.08	25.22	20.89	0	0	0	0
89	49	20.44	24.27	**25.23**	25.23	20.44	0	0	0	0
90	46	**25.85**	18.20	19.89	25.85	18.20	0	0	0	0
91	45	**24.38**	18.89	20.05	24.38	18.57	0	0	0	0
92	43	22.53	22.64	**24.47**	24.47	22.53	0	0	0	0
93	51	**26.04**	24.10	23.74	26.04	21.16	0	0	0	0
94	46	21.50	**22.73**	21.71	22.73	21.50	0	0	0	0
95	48	**25.40**	18.30	19.31	25.40	18.30	0	0	0	0
96	47	**25.20**	19.16	21.09	25.20	19.16	0	0	0	0
97	50	**23.68**	19.84	22.47	23.68	19.66	0	0	0	0
98	48	**25.75**	21.60	22.47	25.75	21.60	0	0	0	0
99	46	**26.30**	21.48	25.28	26.30	21.48	0	0	0	0
100	51	**25.25**	17.87	22.09	25.25	13.33	0	1	0	0
101	55	18.76	18.93	**19.31**	19.31	18.19	0	0	0	0
102	47	**15.72**	14.84	15.28	15.72	14.84	0	1	4	0
103	49	19.42	17.24	**21.68**	21.68	17.24	0	2	1	0
104	54	**22.42**	19.89	20.67	22.42	19.89	0	1	0	0
105	56	**16.69**	14.12	15.56	16.69	14.12	0	1	2	0
106	47	15.58	15.48	16.78	16.94	15.48	0	3	1	0
107	48	**21.43**	14.83	15.42	21.43	13.88	0	1	0	0
108	57	**15.46**	12.93	14.84	15.46	12.93	0	2	2	0
109	50	10.13	11.79	12.15	12.17	10.13	0	0	4	0
110	55	12.22	**12.68**	12.23	12.68	11.43	0	6	3	0
111	52	**13.94**	13.23	13.02	13.94	10.94	0	7	4	0
112	58	12.25	**14.18**	13.86	14.18	12.25	0	7	3	0
113	57	10.82	9.55	**11.12**	11.12	9.55	0	5	6	0
114	64	9.54	14.21	14.31	14.71	9.54	0	8	3	0
115	61	7.50	**9.55**	9.39	9.55	7.50	0	10	6	0
116	67	8.45	8.40	**10.24**	10.24	7.98	0	11	6	0
117	66	6.05	6.69	**6.77**	6.77	6.05	0	13	8	0
118	67	5.21	5.39	**6.24**	6.24	3.79	0	18	12	0
119	62	5.14	**7.25**	7.10	7.25	5.14	0	14	13	0
120	69	4.79	6.03	6.20	6.26	4.79	0	17	9	0

(continued)

Table 9.4 (continued)

SM	Gene	RIP	LP	HSVM	MAX	MIN	SVM4	SVM1	LDF2	QDF
121	62	3.84	3.97	**4.03**	4.03	3.84	0	14	16	0
122	71	3.07	3.86	**4.02**	4.02	3.07	0	11	11	0
123	77	1.92	1.91	**2.02**	2.02	1.91	0	15	13	0
124	60	0.15	**0.17**	**0.17**	0.17	0.15	1	15	14	0
MAX	77	42.46	39.69	45.81	45.81	38.40	1.00	18.00	16.00	1.00
MIN	6	0.15	0.17	0.17	0.17	0.15	0.00	0.00	0.00	0.00
MEAN	36.71	26.68	24.16	25.70	28.07	22.59	0.01	1.48	1.14	0.01
MaxRatio		64	16	27						
SUM	4552									

9.4 Analysis of Three Signal Data of 124 RipDSs, LpDSs, and HsvmDSs

We analyze 124 RipDSs signal data with 128 patients (128 cases) and 124 RipDSs (124 variables) by standard statistical methods and get the surprising success.

9.4.1 Hierarchical Clustering

Many cluster methods have two categories such as hierarchical and non-hierarchical methods. In the hierarchical methods, there are many methods such as the Ward method, group average method, center of gravity method, the shortest distance method, the longest distance method, and so forth. The representative of the non-hierarchical type is the k-means method such as SOM. Alon et al. succeed to find 2,000 genes by SOM. Here, we analyze the signal data using the five hierarchical clustering methods. Moreover, we show the various results. Which cluster is meaningful requires sufficient knowledge of the subject. We must choose the proper k (cluster number) to use the k-means methods or SOM. Because we can trial and error any number of clusters, the hierarchical clustering methods are very convenient compared with the k means. Probably, because medical researchers could not obtain good results using some gene subspaces by the hierarchical clustering before 2000, they analyzed their data by SOM. Now, although all researchers cannot obtain the right results by the hierarchical clustering, they can classify signal data and obtain the surprising results.

(1) Ward cluster of RipDSs signal data

Figure 9.3 is a Ward cluster analysis of RipDSs signal data. Even if it analyzes 124 SMs individually, it cannot separate two classes, but the upper blue part is 95 B-cell subjects, and the lower red part is 33 T-cell patients. We consider the marvelous effects of RipDSs cause this surprising result. The case dendrogram shows three clusters of the B-cell class and two clusters of the T-cell class. Three clusters consist of the 40 red patients, the 11 green patients, and the 44 blue patients. Two clusters consist of the 32 orange patients and the one green patient. Among the six research groups, Alon et al. succeeded in using SOM. Furthermore, if medical AI based on the cluster analysis will analyze SM, it may be able to find useful results among many clusters made by many clustering methods.

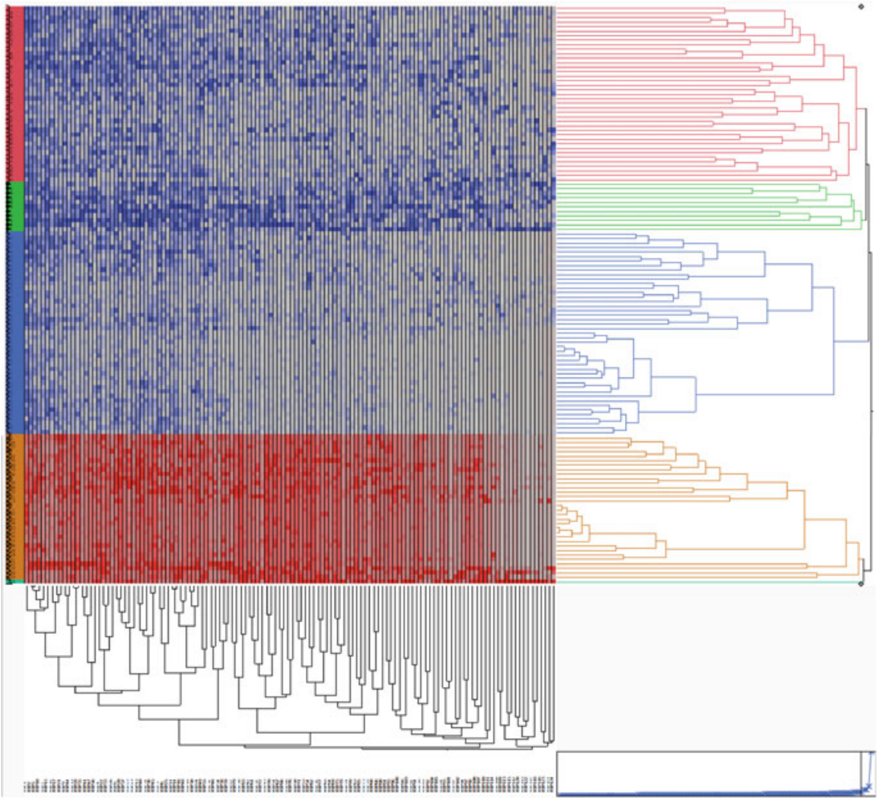


Fig. 9.3 Cluster analysis of RipDSs signal data

(2) The nearest neighbor (shortest distance) cluster analysis of RipDSs signal data

Figure 9.4 shows the cluster analysis of the RipDS signal data by the nearest neighbor method. The case dendrogram shows three clusters of the B-cell class and two clusters of the T-cell class. Three clusters consist of the 93 red patients, the one green patient, and the one blue patient. Two clusters consist of the 32 orange patients and the one green patient. The feature of the nearest neighbor method is that the other clusters are sequentially merged into the cluster with the shortest distance to become one cluster.

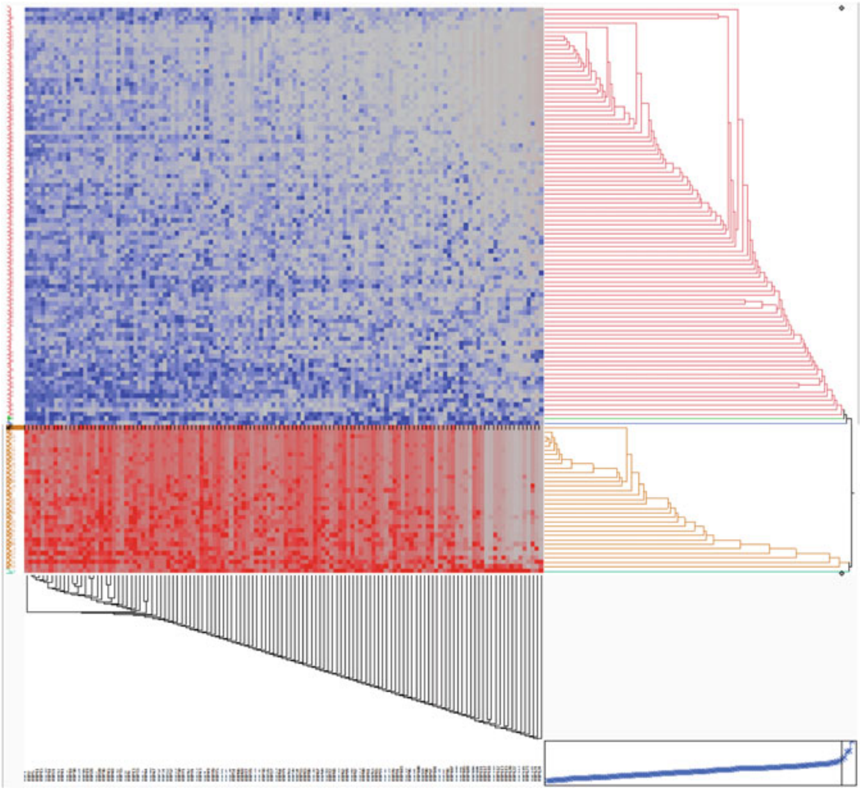


Fig. 9.4 Nearest neighbor cluster analysis of RipDSs signal data

(3) The longest distance cluster analysis of RipDSs signal data

Figure 9.5 shows the longest distance cluster analysis of signal data of RipDS. The case dendrogram shows three clusters of the B-cell class and two clusters of the T-cell class. Three clusters consist of the 92 red patients, the two green patients, and the one blue patient. Two clusters consist of the 31 orange patients and the two green patients. The 33 T-cell classes show a pattern like the nearest neighbor method.

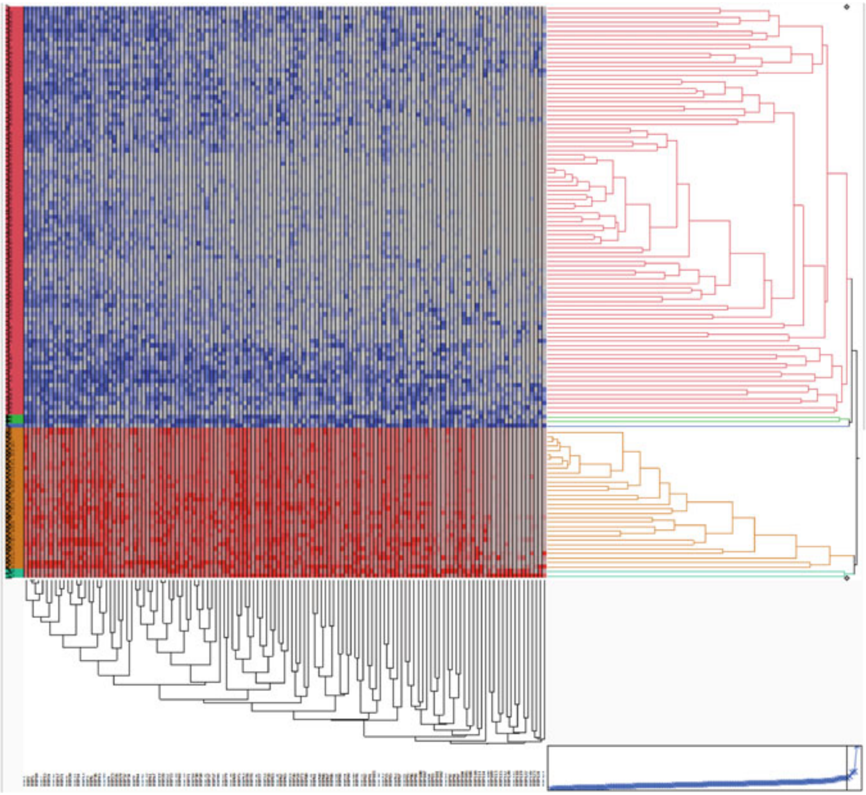


Fig. 9.5 Longest distance cluster analysis of RipDSs signal data

(4) The centroid cluster analysis of discriminant score data of RIP

Figure 9.6 is the centroid cluster analysis of RipDSs signal data. The case dendrogram shows three clusters of the B-cell class and two clusters of the T-cell class. Three clusters consist of the 93 red patients, the one green patient, and the one blue patient. Two clusters consist of the 32 orange patients and the one green patient. Both classes show patterns like the nearest neighbor method.

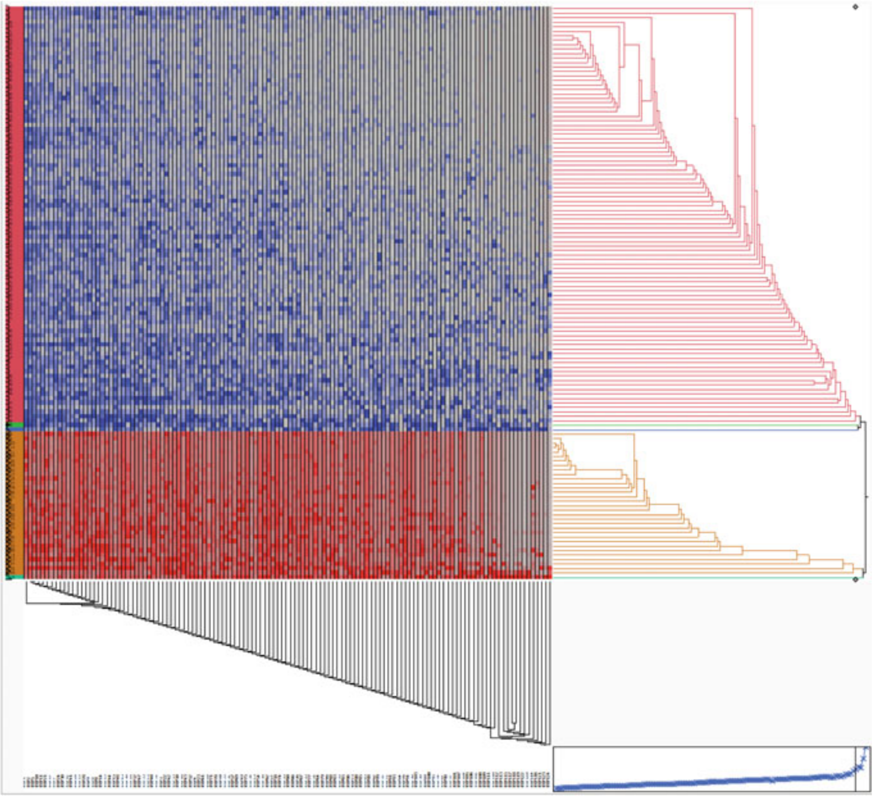


Fig. 9.6 Centroid cluster analysis of RipDSs signal data

(5) The group mean cluster analysis of discriminant score data of RIP

Figure 9.7 is the group mean cluster analysis of RipDSs signal data. The case dendrogram shows three clusters of the B-cell class and two clusters of the T-cell class. Three clusters consist of the 92 red patients, the two green patients, and the one blue patient. Two clusters consist of the 32 orange patients and the one green patient. Both classes show patterns like the nearest neighbor method.

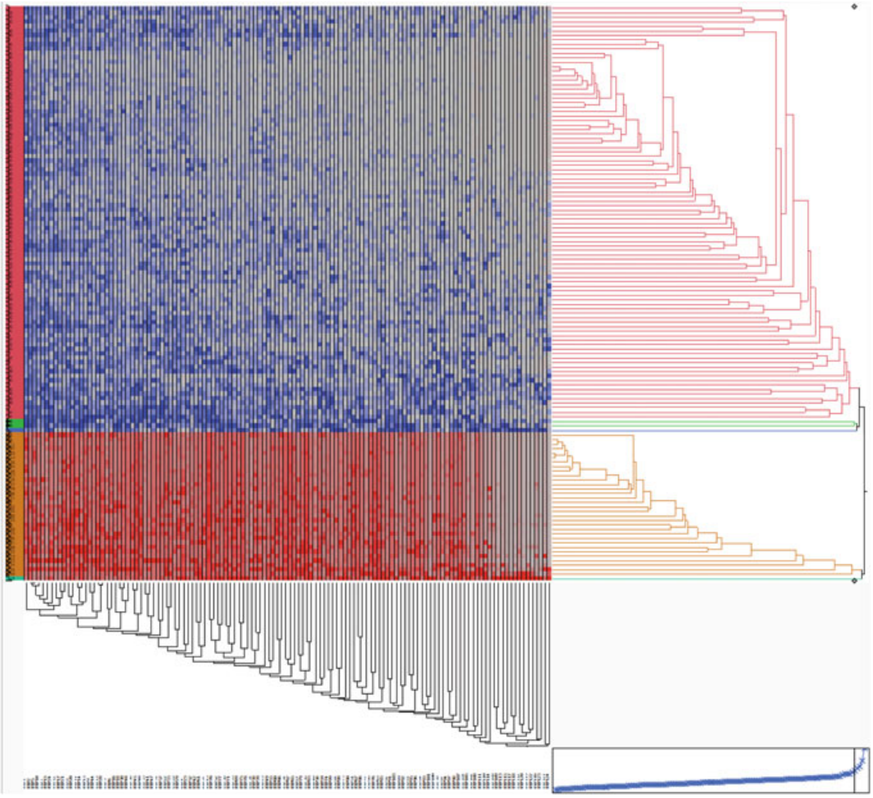


Fig. 9.7 Group mean cluster analysis of RipDSs signal data

9.4.2 Clustering of LpDSs and HsvmDSs Signal Data

(1) Ward cluster analysis of the LpDSs signal data

Figure 9.8 shows Ward cluster analysis of the LpDSs signal data. The case dendrogram shows three clusters of the B-cell class and two clusters of the T-cell class. Three clusters consist of the 43 red patients, the eight green patients, and the 44 blue patients. Two clusters consist of the 31 orange patients and the two green patients.

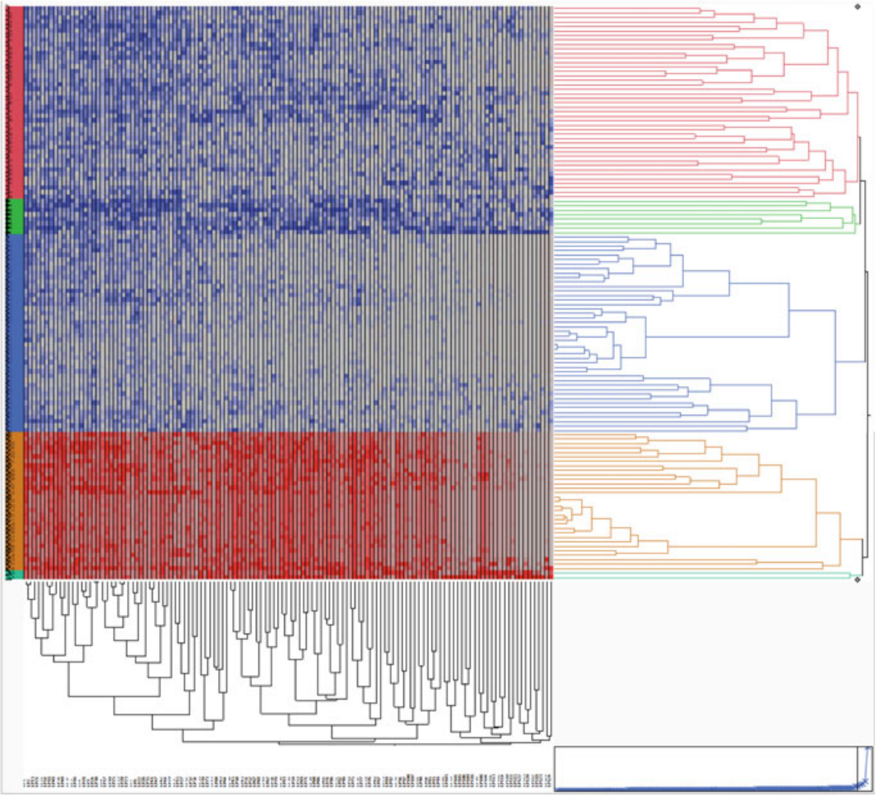


Fig. 9.8 Ward's cluster analysis of LpDSs signal data

(2) Ward cluster analysis of HsvmDS signal data

Figure 9.9 is Ward cluster analysis of HsvmDSs signal data. Because this cluster analysis is the most similar to Ward cluster analysis of RipDSs signal data, we choose the ten clusters. The case dendrogram shows seven clusters of the B-cell class and three clusters of the T-cell class. Seven clusters consist of the 19 red patients, the nine green patients, the six blue patients, the one orange patient, the 21 pale green patients, the 16 purple patients, and the 23 yellow patients. Three clusters consist of the 30 blue patients, the two purple patients, and the one yellow patient. For clarity of explanation, we have fixed the number of clusters to 5 and analyzed it. However, by changing to 10, we can see the abundant relationship of cases. If we prepare the signal data, we do not need to search the number of clusters from 2 to 3 and so on in SOM sequentially.

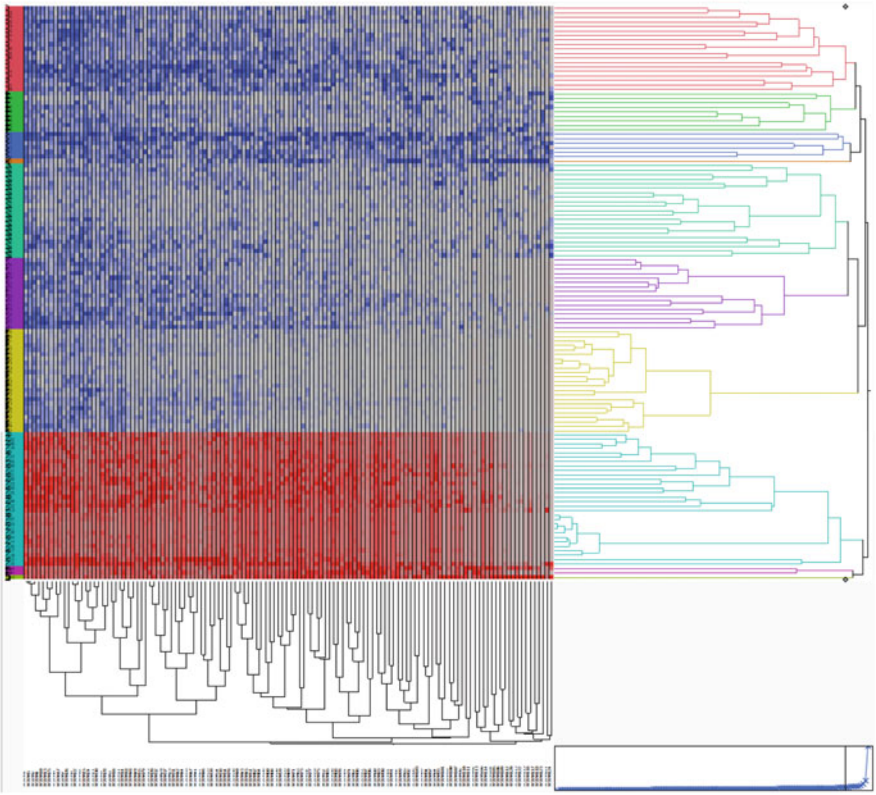


Fig. 9.9 Ward cluster analysis of HsvmDSs signal data

9.5 PCA Analysis of Signal Data

Figure 9.10 shows the result of RipDSs signal data by PCA. The first eigenvalue is 99.679, and the contribution ratio is 80.4%. The second eigenvalue is 4.8782, the contribution ratio is 3.93, and the cumulative contribution ratio is 84.33%. That is, the Prin1 represents the 128 patients. Although there are no healthy subjects, the two classes locate spherically on Prin1. The 95 B-cell patients are in the second and third quadrants. The 44 blue patients include the origin. The RipDS values decrease in 40 red cases and 11 green cases in that order. Moreover, several cases of green are outliers.

In 33 cases of T-cell, some of orange 32 cases are outliers, and one green case is an outlier.

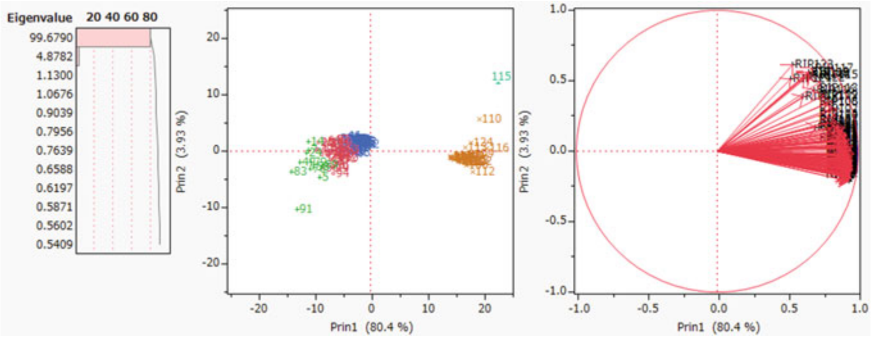


Fig. 9.10 Three plots of PCA (RipDS signal data)

The first columns and second columns of Table 9.5 show the case number corresponding "SM" in Table 9.4 and its value of Prin1 axis. The 124 rows have two parts. Upper 95 rows are a B-cell class, and lower 36 rows are a T-cell class in Fig. 9.10. These two columns are sorted in ascending order from a small value. The range of B-cell is [−14.040, −1.156]. The range of T-cell is [14.129, 22.638]. SVs open the window range of (−1.156, 14.129). Thus, we can define the RatioSV for PCA in Eq. (9.1).

$$\begin{aligned}\text{RatioSV of PCA} &= (1.156 + 14.129)/(14.040 + 22.638) * 100 \\ &= 1528.657/36.679 = 41.68\%. \end{aligned} \tag{9.1}$$

Although this is the overall characteristic value of RatioSV of 124 RIPs, it is smaller than the maximum value of RatioSV of 124 RIPs 42.46, so the malignancy index of PCA is useless. In later, we conclude the same results of both RatioSV of PCA by Revised LP-OLDF and HSVM.

Table 9.5 Prin1 values of RIP and Revised LP-OLDF and HSVM sorted by each Prin1

RIP	Prin1	LP	Prin1	HSVM	Prin1
83	−14.040	83	−13.751	83	−14.082
91	−12.970	49	−12.861	49	−13.556
49	−12.534	91	−12.591	91	−12.735
24	−11.272	7	−11.464	7	−11.492
7	−11.029	24	−11.013	14	−11.079
14	−10.980	14	−10.961	24	−10.956
76	−10.594	79	−10.668	58	−10.558
79	−10.481	76	−10.360	76	−10.476
58	−9.950	58	−10.310	79	−10.361
39	−9.330	39	−9.170	4	−10.101
4	−9.173	4	−9.060	39	−9.669
25	−9.132	42	−8.932	42	−9.149
5	−8.941	25	−8.820	25	−9.001
89	−8.513	89	−8.671	89	−8.812
42	−8.471	56	−8.270	40	−8.659
56	−8.362	71	−8.109	6	−8.550
40	−8.206	6	−8.062	71	−8.448
6	−8.067	40	−7.839	56	−8.201
54	−7.563	5	−7.701	5	−7.918
71	−7.500	2	−7.697	2	−7.863
11	−7.469	73	−7.254	11	−7.721
73	−7.299	11	−7.229	27	−7.545
1	−7.250	54	−7.147	73	−7.299
15	−7.020	90	−7.102	1	−7.251
27	−7.016	27	−7.014	15	−7.165
48	−6.806	15	−7.000	92	−7.141
50	−6.636	92	−6.874	54	−7.124
35	−6.634	1	−6.870	90	−7.117
75	−6.634	50	−6.643	57	−6.738
90	−6.554	33	−6.536	50	−6.632
19	−6.501	19	−6.443	19	−6.546
94	−6.490	35	−6.419	48	−6.535
51	−6.316	48	−6.394	33	−6.418
3	−6.316	57	−6.376	35	−6.345

(continued)

Table 9.5 (continued)

RIP	Prin1	LP	Prin1	HSVM	Prin1
2	−6.280	94	−6.337	52	−6.222
34	−6.264	75	−6.295	17	−6.189
92	−6.250	3	−6.256	34	−6.182
84	−6.228	17	−6.077	3	−6.159
33	−6.186	51	−6.069	29	−6.103
29	−6.061	84	−6.036	45	−5.968
8	−5.929	29	−5.928	75	−5.945
57	−5.890	30	−5.903	51	−5.944
17	−5.865	8	−5.889	20	−5.927
30	−5.702	34	−5.803	28	−5.812
59	−5.667	20	−5.664	84	−5.809
20	−5.626	45	−5.635	30	−5.794
37	−5.501	37	−5.498	94	−5.713
45	−5.425	52	−5.325	37	−5.584
67	−5.345	28	−5.308	8	−5.553
12	−5.234	60	−5.286	93	−5.318
52	−5.155	59	−5.270	60	−5.291
26	−4.926	26	−5.240	67	−5.283
28	−4.904	12	−5.040	12	−5.072
93	−4.854	67	−4.997	26	−4.967
60	−4.788	44	−4.910	44	−4.840
44	−4.621	78	−4.868	59	−4.821
13	−4.573	93	−4.786	78	−4.801
78	−4.539	74	−4.470	64	−4.623
46	−4.469	46	−4.324	46	−4.559
21	−4.390	64	−4.256	74	−4.429
64	−4.380	80	−4.227	13	−4.346
74	−4.366	13	−4.211	9	−4.259
36	−4.274	21	−4.077	21	−4.132
9	−4.154	9	−4.040	18	−4.130
80	−4.142	36	−4.031	80	−4.070
88	−4.111	65	−3.837	88	−3.969
65	−3.961	18	−3.830	23	−3.792
18	−3.805	47	−3.703	36	−3.692
10	−3.626	88	−3.541	22	−3.577
47	−3.604	10	−3.515	65	−3.492

(continued)

Table 9.5 (continued)

RIP	Prin1	LP	Prin1	HSVM	Prin1
23	−3.554	23	−3.374	10	−3.347
68	−3.406	69	−3.319	66	−3.337
22	−3.292	68	−3.279	47	−3.308
69	−3.182	82	−3.196	68	−3.090
62	−3.136	22	−3.152	77	−3.040
70	−3.125	86	−3.111	69	−3.020
77	−3.071	62	−3.051	82	−2.982
82	−3.003	77	−3.032	62	−2.914
86	−2.871	70	−3.028	70	−2.774
66	−2.781	66	−2.942	86	−2.774
41	−2.765	72	−2.719	41	−2.665
81	−2.671	16	−2.493	38	−2.528
16	−2.630	81	−2.475	72	−2.500
72	−2.605	85	−2.436	81	−2.396
63	−2.520	87	−2.427	16	−2.365
95	−2.452	95	−2.387	87	−2.199
31	−2.314	31	−2.248	31	−2.193
87	−2.312	41	−2.231	63	−2.189
61	−2.264	38	−2.141	95	−2.046
85	−2.233	63	−2.030	85	−2.031
38	−2.080	61	−1.990	61	−1.882
55	−1.916	55	−1.827	43	−1.748
43	−1.869	43	−1.761	55	−1.625
53	−1.712	53	−1.182	53	−1.393
32	**−1.156**	32	**−0.927**	32	**−0.926**
111	**14.129**	111	**14.037**	111	**13.896**
114	14.158	98	14.132	103	14.085
103	14.303	114	14.162	98	14.164
126	14.338	103	14.211	104	14.246
98	14.358	104	14.311	128	14.270
104	14.490	107	14.407	114	14.362
128	14.517	128	14.428	126	14.557
117	14.661	126	14.533	107	14.571
107	14.746	119	14.688	117	14.829
118	14.892	118	14.734	118	14.992
119	15.164	117	14.952	119	15.071

(continued)

Table 9.5 (continued)

RIP	Prin1	LP	Prin1	HSVM	Prin1
127	15.336	127	15.267	127	15.465
97	15.495	97	15.434	97	15.563
121	15.558	108	15.497	96	15.759
96	15.587	96	15.633	108	15.763
108	15.605	102	15.658	102	15.973
102	15.991	121	15.807	121	16.095
105	16.096	105	16.447	105	16.541
100	16.348	125	16.511	100	16.839
101	16.499	100	16.538	101	16.863
120	16.659	113	16.702	113	17.000
125	16.696	106	16.801	125	17.007
113	16.758	101	16.860	120	17.117
106	16.934	120	16.868	106	17.473
99	16.946	99	16.979	123	17.575
123	17.230	123	17.087	99	17.903
122	17.706	112	17.471	109	18.163
124	17.879	122	17.627	124	18.210
109	18.072	109	17.695	122	18.260
112	18.249	124	17.754	112	18.587
110	19.378	110	19.568	110	19.293
116	20.648	116	20.085	116	20.859
115	22.638	115	21.963	115	21.531

Figure 9.11 shows the result of LpDSs signal data by PCA. The first eigenvalue is 98.578, and the contribution ratio is 79.5%. The second eigenvalue is 5.121, the contribution ratio is 4.13%, and the cumulative contribution ratio is 83.63%. That is, the Prin1 almost represents 128 patients. The score plot is almost the same as Fig. 9.10. The third and fourth columns of Table 9.5 show the result of LpDSs. The ranges of B-cell class and T-cell class are [−13.751, −0.927] and [14.037, 21.963]. SV opens the window that is the interval (−0.927, 14.037). RatioSV of PCA by LpDSs is Eq. (9.2).

$$\text{RatioSV of PCA by LpDSs} = (0.9274 + 14.037)/(13.751 + 21.964) * 100 = 42.02\,\% \tag{9.2}$$

Because the maximum RatioSV of LpDSs is 39.69, RatioSV of PCA is helpful as a malignancy index.

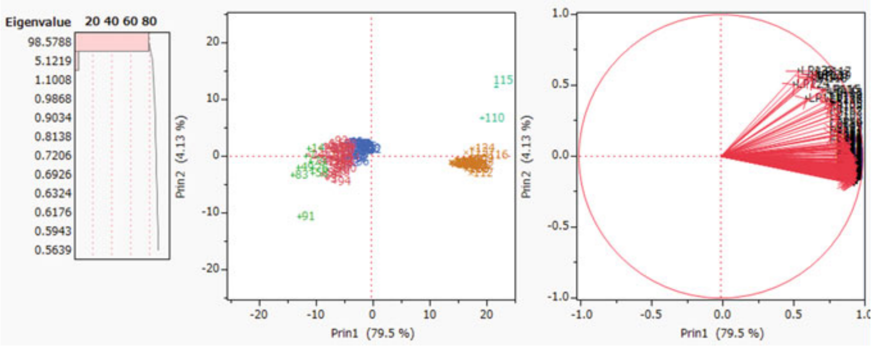


Fig. 9.11 Three plots of PCA (LpDS signal data)

Figure 9.12 shows the result of HsvmDSs signal data by PCA. The first eigenvalue is 101.901, and the contribution ratio is 82.2%. The second eigenvalue is 5.209, the contribution ratio is 4.2%, and the cumulative contribution ratio is 86.4%. That is, the Prin1 almost represents 128 patients. In the cluster analysis, we make 10 clusters, and score plot explains their positional relation. By changing the number of clusters, we believe it will be useful for medical case studies. The fifth and sixth columns of Table 9.5 show the result of HsvmDSs. The ranges of B-cell class and T-cell class are [−14.082, −0.926] and [13.896, 21.531]. SV opens the window that is the interval (−0.926, 13.897). RatioSV of PCA by HsvmDSs is Eq. (9.3).

$$\text{RatioSV of PCA by HsvmDSs} = (0.9262 + 13.89686)/(14.0821 + 21.53172) * 100 = 41.62\,\% \tag{9.3}$$

Because the maximum RatioSV of LpDSs is 45.81, RatioSV of PCA is useless as a malignancy index.

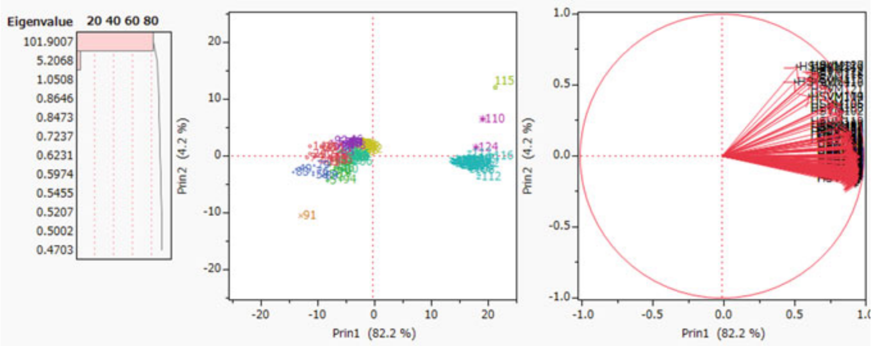


Fig. 9.12 Three plots of PCA (HsvmDS signal data)

9.6 PCA (Transposed Signal Data)

We transpose the RipDSs signal data and analyze this data with 124 RipDSs (124 cases) and 128 patients (128 variables). Figure 9.13 is three plots of PCA. Because the first eigenvalue is 60.09 and contribution ratio is 46.9%, Prin1 explains 46.9% variance. Score plot shows only R124 is the clear outlier among 124 RipDSs. Factor loading plot shows the surprising result. Like the cross, it almost overlaps the axes of Prin 1 and Prin 2. Now we cannot explain the meaning of the transposed data.

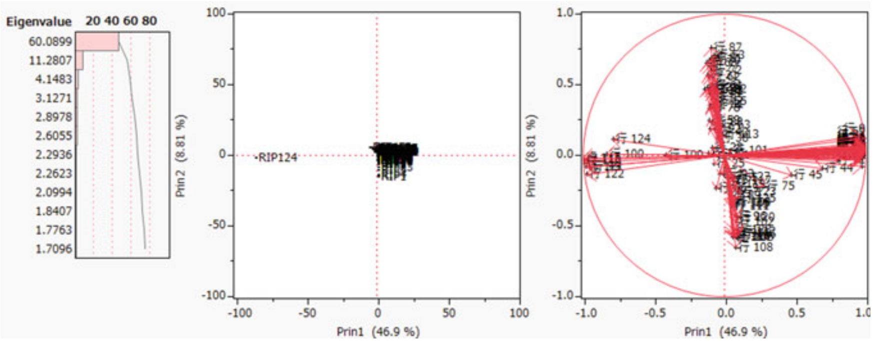


Fig. 9.13 Three plots of PCA (RipDS signal data)

Table 9.6 shows the five groups of factor loading plot. If Prin1 is greater than and equal to 0.2, "Group = 1." If Prin2 is greater than and equal to 0.2, "Group = 2." If Prin1 is less than and equal to −0.2, "Group = 3." If Prin2 is less than and equal to −0.2, "Group = 4." Otherwise, "Group = 0." Roughly speaking, the 95 B-cell patients belong to Group = 1 or 2, and the 33 T-cell patients belong to Group = 3 or 4. Other 15 patients belong to Group = 0. Although we cannot explain the medical meaning, our results of PCA and cluster analysis using two signal data indicate many different patient clusters.

Table 9.6 Group of factor loading plot

Row	Prin1	Prin2	Group	Prin3	Prin4
Row1	−0.036	−0.131	0	0.291	−0.018
Row2	0.985	0.062	1	−0.084	−0.012
Row3	0.997	0.036	1	−0.034	−0.015
Row4	0.975	−0.001	1	0.029	−0.018
Row5	0.997	0.033	1	−0.037	−0.004

(continued)

Table 9.6 (continued)

Row	Prin1	Prin2	Group	Prin3	Prin4
Row6	0.973	0.065	1	−0.122	−0.006
Row7	0.992	0.040	1	−0.061	0.008
Row8	0.997	0.030	1	−0.030	−0.001
Row9	0.960	0.140	1	−0.100	0.054
Row10	0.993	0.033	1	−0.035	−0.007
Row11	−0.111	0.013	0	−0.079	0.035
Row12	0.990	0.037	1	−0.029	0.004
Row13	0.994	0.058	1	−0.073	−0.016
Row14	0.989	0.044	1	−0.037	0.006
Row15	0.986	0.004	1	0.105	−0.003
Row16	0.954	0.027	1	−0.066	0.062
Row17	0.993	0.033	1	−0.009	0.004
Row18	−0.102	0.462	2	0.036	0.349
Row19	0.983	−0.004	1	0.053	0.029
Row20	−0.072	0.007	0	0.000	0.659
Row21	−0.123	0.569	2	−0.061	0.066
Row22	0.990	0.046	1	−0.069	−0.026
Row23	−0.122	0.477	2	−0.011	0.135
Row24	0.996	0.043	1	−0.059	−0.011
Row25	−0.074	−0.057	0	0.347	0.061
Row26	0.997	0.045	1	−0.010	−0.011
Row27	0.995	0.050	1	−0.051	−0.013
Row28	−0.086	0.056	0	−0.073	0.120
Row29	0.699	0.009	1	0.442	−0.167
Row30	0.996	0.051	1	−0.058	−0.017
Row31	−0.079	0.359	2	−0.009	−0.056
Row32	−0.060	0.470	2	0.107	0.058
Row33	0.944	0.071	1	−0.030	−0.035
Row34	0.987	0.019	1	0.018	−0.001
Row35	0.951	−0.017	1	0.134	−0.028
Row36	0.995	0.054	1	−0.060	−0.021
Row37	0.960	−0.035	1	0.124	−0.008
Row38	−0.044	0.120	0	0.060	−0.014
Row39	0.993	0.053	1	−0.076	−0.016
Row40	0.906	0.024	1	0.125	−0.111
Row41	−0.082	0.457	2	0.375	0.098

(continued)

Table 9.6 (continued)

Row	Prin1	Prin2	Group	Prin3	Prin4
Row42	0.986	0.000	1	0.003	0.029
Row43	−0.044	0.160	0	0.436	0.022
Row44	0.690	−0.092	1	−0.045	0.404
Row45	0.473	−0.139	1	0.531	−0.067
Row46	−0.090	0.480	2	0.152	−0.012
Row47	0.996	0.051	1	−0.055	−0.017
Row48	0.993	0.069	1	−0.069	−0.027
Row49	0.997	0.027	1	−0.002	−0.020
Row50	0.962	−0.029	1	0.073	0.043
Row51	−0.038	0.001	0	0.512	−0.251
Row52	0.995	0.055	1	−0.055	−0.001
Row53	−0.073	0.707	2	0.167	0.083
Row54	0.992	0.038	1	−0.004	0.019
Row55	−0.059	0.378	2	0.151	−0.076
Row56	0.993	0.013	1	−0.005	0.005
Row57	−0.137	0.463	2	0.034	−0.035
Row58	0.998	0.035	1	−0.030	−0.016
Row59	−0.105	0.249	2	0.141	−0.210
Row60	0.994	0.054	1	−0.075	−0.022
Row61	−0.087	0.491	2	−0.011	0.071
Row62	−0.107	0.556	2	−0.077	−0.129
Row63	−0.032	0.211	2	0.103	0.746
Row64	0.989	0.087	1	−0.084	−0.023
Row65	0.020	−0.159	0	0.441	0.385
Row66	−0.059	−0.236	3	−0.176	0.144
Row67	0.996	0.051	1	−0.063	−0.007
Row68	−0.125	0.651	2	−0.039	−0.080
Row69	0.984	0.122	1	−0.081	−0.026
Row70	−0.109	0.670	2	0.078	−0.128
Row71	0.989	0.005	1	0.077	−0.031
Row72	−0.095	0.600	2	0.065	0.009
Row73	0.997	0.033	1	−0.025	−0.016
Row74	−0.102	0.349	2	−0.041	0.063
Row75	0.282	−0.219	1	0.651	0.124
Row76	0.996	0.038	1	−0.035	−0.028
Row77	−0.107	0.632	2	0.060	0.103

(continued)

Table 9.6 (continued)

Row	Prin1	Prin2	Group	Prin3	Prin4
Row78	−0.108	0.334	2	0.040	0.337
Row79	0.998	0.035	1	−0.027	−0.011
Row80	0.997	0.048	1	−0.028	−0.012
Row81	−0.093	0.685	2	0.045	−0.027
Row82	0.995	0.064	1	−0.058	−0.011
Row83	0.982	−0.021	1	0.028	0.033
Row84	0.995	0.053	1	−0.056	−0.020
Row85	−0.088	0.403	2	−0.025	0.026
Row86	0.801	−0.040	1	0.150	0.029
Row87	−0.096	0.757	2	0.094	−0.013
Row88	−0.088	0.236	2	−0.052	0.305
Row89	0.042	−0.289	4	0.834	−0.070
Row90	0.963	−0.010	1	0.152	−0.001
Row91	0.972	−0.077	1	0.164	0.001
Row92	−0.102	0.187	0	−0.046	−0.027
Row93	−0.009	−0.133	0	0.507	−0.075
Row94	0.996	0.046	1	−0.069	−0.014
Row95	**−0.100**	**0.433**	**2**	**−0.038**	**0.044**
Row96	**0.080**	**−0.431**	**4**	**0.008**	**0.224**
Row97	0.073	−0.234	4	0.180	0.067
Row98	0.057	−0.552	4	−0.153	−0.046
Row99	−0.953	−0.081	3	−0.070	−0.005
Row100	−0.833	0.006	3	−0.118	−0.175
Row101	0.046	0.031	0	−0.090	0.087
Row102	0.089	−0.474	4	0.086	0.172
Row103	0.056	−0.346	4	−0.012	0.115
Row104	0.067	−0.332	4	0.003	0.046
Row105	0.089	−0.579	4	0.023	0.255
Row106	0.098	−0.562	4	0.054	0.093
Row107	0.075	−0.579	4	−0.116	−0.022
Row108	0.079	−0.651	4	−0.130	−0.013
Row109	−0.417	−0.003	3	−0.143	−0.331
Row110	−0.995	−0.031	3	0.056	0.014
Row111	0.054	−0.361	4	−0.087	−0.071
Row112	0.099	−0.542	4	0.077	0.127
Row113	−0.021	0.146	0	−0.721	0.269

(continued)

Table 9.6 (continued)

Row	Prin1	Prin2	Group	Prin3	Prin4
Row114	0.058	−0.580	4	−0.125	0.128
Row115	−0.998	−0.019	3	0.019	0.011
Row116	−0.997	−0.051	3	0.047	0.022
Row117	0.066	−0.177	0	0.033	−0.107
Row118	0.062	−0.572	4	−0.064	0.083
Row119	0.028	−0.209	4	−0.097	−0.751
Row120	0.093	−0.442	4	0.050	−0.149
Row121	−0.980	−0.082	3	0.085	−0.040
Row122	−0.973	−0.135	3	0.063	0.043
Row123	0.103	−0.257	4	0.114	0.093
Row124	−0.789	0.112	3	−0.227	−0.041
Row125	0.100	−0.305	4	0.094	0.089
Row126	0.056	−0.332	4	0.030	0.150
Row127	0.062	−0.155	0	−0.004	0.225
Row128	0.064	−0.528	4	−0.116	−0.060

We transpose the LpDSs signal data (128 patients and 124 LpDSs) and analyze this data with 124 LpDSs (124 cases) and 128 patients (128 variables). Figure 9.14 is three plots of PCA. Scatter plot shows three outliers. Factor loading plot shows almost the same five groups of 128 patients as Table 9.6.

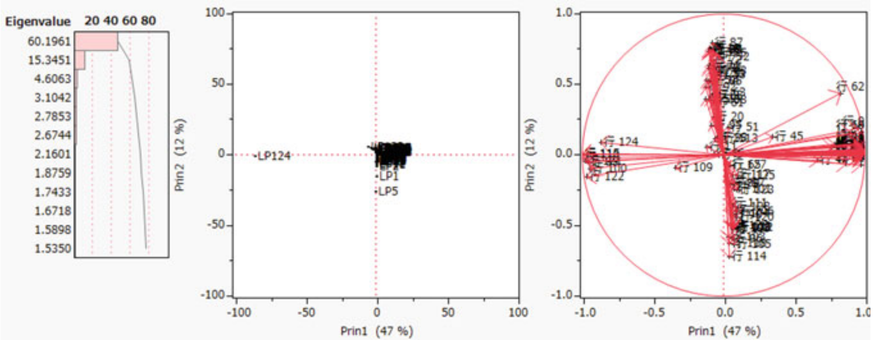


Fig. 9.14 Three plots of PCA (LpDS signal data)

We transpose the HsvmDSs signal data (128 patients and 124 HsvmDSs) and analyze this data with 124 HsvmDSs (124 cases) and 128 patients (128 variables). Figure 9.15 is three plots of PCA. Scatter plot shows one outlier. Factor loading plot shows almost the same five groups of 128 patients as Table 9.6. Although three signal data of Chiaretti show almost the same results as other microarrays, three transposed signal data of Chiaretti show the surprising different results. We cannot explain the reason now. This is the next theme of Book4.

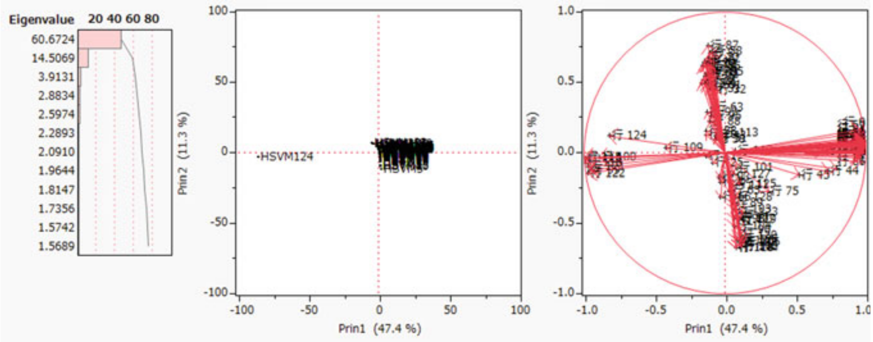


Fig. 9.15 Three plots of PCA (HsvmDSs signal data)

9.7 Conclusions

In this book, we introduce the cancer gene diagnosis to analyze all SMs of six microarrays. Only three Revised OLDFs and H-SVM find microarrays are LSD (Fact3). Only three Revised OLDFs can decompose microarrays into many SMs and noise subspace (Fact4). At first, we consider several genes included in the SMs and the union of SMs is signal subspaces because two classes are entirely separated. However, standard statistical methods, except for logistic regression, cannot find the linear separable facts (Problem6). On the other hand, many RatioSVs of SMs are over 5%. These facts indicate RipDSs, LpDSs, and HsvmDSs can discriminate two classes correctly. Moreover, PCA and cluster analysis find the linear separable facts of tree signal data made by RipDSs, LpDSs, and HsvmDSs. Thus, we consider three signal data are signal because RIP, Revised LP-OLDF, and H-SVM can find the linear separable facts in high-dimensional microarrays. Our claim is confirmed as follows:

(1) All correlations of DSs are the positive values. However, correlations of genes included in SMs take the positive, almost zero, and negative values.
(2) Statistical discriminant functions cannot discriminate between microarrays and all SMs correctly because they cannot discriminate LSD correctly. Moreover, although standard statistical methods cannot find the linear separable facts, those can analyze three signal data entirely and find many outliers those are the research themes of Golub et al.

Above facts invoke the Problem6 "Why cannot statistical discriminant functions discriminate microarrays and all SMs correctly, and standard statistical methods find the linearly separable signs of all SMs?" We show many facts that the fluctuation of the two classes is too small compared with the genes variation. Moreover, we propose to analyze SMs by Ward cluster analysis and make the five or ten clusters using signal data. If we analyze the signal data by PCA, the score plot reveals the relations of many clusters.

Acknowledgements We can achieve our research by the dominant software such as LINGO supported by LINDO Systems Inc. and JMP backed by SAS Institute Japan Ltd. JMP Japan Division.

References

Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood 103(7):2771–2778

Sall JP, Creighton L, Lehman A (2004) JMP start statistics, 3rd edn. SAS Institute Inc., USA (Shinmura S edits Japanese version)

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2016) New theory of discriminant analysis after R. Fisher. Springer, Tokyo

Shinmura S (2017) Cancer gene analysis by Singh et al. microarray data. ISI2017, pp 1–6

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD'18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. In: Biocomp'18, pp. 1–7

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Lada M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(2):203–209

Chapter 10
LINGO Programs of Cancer Gene Analysis

Abstract In "New Theory of Discriminant Analysis after R. Fisher" (2016), Shinmura had already explained LINGO Program1 in Chap. 2. LINGO Program1 defines six MP-based LDFs such as Revised IP-OLDF (RIP), Revised LP-OLDF, Revised IPLP-OLDF, H-SVM, two soft-margin SVMs such as SVM4 (penalty c = 10000) and SVM1 (penalty c = 1). Everyone can evaluate six MP-based LDFs in the training samples at once. If you can understand these models, you can develop your bespoken models by yourself. LINGO Program2 can discriminate a small training sample by the Method1 instead of LOO method. If you can understand LINGO Program2, you can build the complex MP models to control several optimizable models with many datasets as arrays. LINGO can control the complex optimization models. In this chapter, we explained the Matryoshka feature selection method (Method2). This chapter explains LINGO Program3 in addition to Linus's Linear Discriminant Function. Section 10.1 introduces the role of three LINGO programs. Section 10.2 introduces LINGO sample model (DiscrmSwiss.lng) that is a sample model downloaded from LINDO Systems Inc. HP (https://www.lindo.com/). Everybody can download many fine models, manuals, textbooks, and evaluation solvers such as LINGO, What's Best! (Excel add-in), and LINGO/API (c libraries to develop bespoken models and systems) in free. In order to simplify the program, we assume that class1 has a discriminant score (DS) of 1 or more and class2 becomes −1 or less as explained in Chap. 1. Then, it converts original data of class2 by multiplying by −1 and thinks that the extended DSs are judged correctly more than 1. The LDF introduced in this section can be used without converting a sign of class2 data. Section 10.3 introduces six MP-based LDFs. Section 10.4 introduces LINGO Program 3 of Method2. Section 10.5 introduces the validation of Method2 by LINGO Program1 using common data. Section 10.6 is conclusion.

Keywords Matryoshka feature selection method (Method2) · LINGO Program3 for Method2 · Small Matryoshka (SM) · Revised IP-OLDF(RIP) · Minimum number of misclassifications (MNM) · Revised LP-OLDF · Revised IPLP-OLDF · Support vector machine (SVM) · H-SVM · SVM4 · SVM1 · Number of misclassifications (NM) · Discriminant score (DS)

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9_10

10.1 Introduction

We developed many LINGO Programs for the MP-based LDFs (Schrage 2006). Those are three optimal linear discriminant functions (optimal LDFs, OLDFs) and three support vector machines (SVMs). Three OLDFs are Revised IP-OLDF (RIP) using integer programming (IP), Revised LP-OLDF using linear programming (LP) and Revised IPLP-OLDF using IP and LP. Three SVMs using quadratic programming (QP) are a hard-margin SVM (H-SVM), two soft-margin SVMs such as SVM4 (penalty c = 10000) and SVM1 (penalty c = 1). We summarized three standard LINGO Programs. Program1 discriminate the data by six MP-based LDFs at once. Section 2.3.3 of New Theory of Discriminant Analysis After Fisher (Shinmura 2016) explained it. We compare and evaluate six different types of common data by six MP-based LDFs and three statistical discriminant functions such as logistic regression, Fisher's LDF, and quadratic discriminant function (QDF). Six common data are used to explain the unique themes as follows:

Chapter 2: Iris data and Fisher's Assumption (Problem4)
Chapter 3: Cephalo-Pelvic Disproportion Data with Collinearities (Problem4)
Chapter 4: Student Data (Problem1 and Problem4)
Chapter 5: Pass/Fail Determination Using Examinations Scores (Problem2 and Problem3 and Problem4)
Chapter 6: Best Models for Swiss Banknote Data (Problem2 and Problem4)
Chapter 7: Japanese-Automobile Data (Problem2 and Problem3 and Problem4)
Chapter 8: Matryoshka Feature Selection Method for Microarray Dataset (Problem1 and Problem2 and Problem5)
Chapter 9: Explanation of LINGO Program2.
Because the statistical discriminant analysis is not the inferential statistic (Problem3), we develop the 100-fold cross-validation for small sample method (Method1) instead of LOO. We developed LINGO Program2 for Method1 supported by LINDO Systems Inc. Program2 offers the 100 error rates, and the discriminant coefficients of 100 validation samples explained in Chap. 9. Moreover, JMP script supports the 100 error rates and the discriminant coefficients of 100 validation samples for logistic regression and Fisher's LDF supported by the JMP division of SAS Institute Japan. We evaluate all common data by six MP-based LDFs and two statistical discriminant functions using Method1. We evaluated eight LDFs by comparing the minimum mean of error rate by the validation sample (M2). Six M2s of RIP are almost better than other seven LDFs. Although most statisticians believe MNM criterion overestimate the validation sample, this fact indicates MNM criterion is more robust and reliable statistics than NM.

In this chapter, we introduced LINGO Program3. Because only RIP and H-SVM can discriminate the linearly separable data (LSD) theoretically and other discriminant functions are useless for microarrays, many statisticians and machine learning researchers could not solve the cancer gene analysis from 1970 (Problem5). However, when we discriminate six microarrays at the end of 2015, RIP found six microarrays were LSD (Fact3), and RIP could decompose microarrays into many SMs (Fact4). Thus, we developed the Matryoshka feature selection method (Method2), and LINGO Program3 explained in this chapter. Moreover, we explain other topics.

10.2 LINGO Sample Model (DiscrmSwiss.lng)

In this section, we introduce a LINGO sample model named “DiscrmSwiss.lng” downloaded from LINDO Systems Inc. HP (https://www.LINDO.com). This HP offers three evaluation solvers such as LINGO, What’s Best!, LINDO/API in addition to manuals and textbooks. These materials are free. Professor Emeritus Linus Schrage developed this LDF based on the MNM criterion as same as RIP developed by us. We introduce it in this section. LINGO model consists of four sections such as SET, DATA, SUBMODEL, and CALC sections. The SET section defines a one-dimensional set with arrays. Moreover, the combinations of one-dimensional sets generate multi-dimensional sets with arrays. Data section defines the values on the array, and input/output the data with such as Excel. SUBMODEL section defines models. CALC section computes and controls the optimal complex models or systems.

10.2.1 Original DiscrmSwiss.lng

Program 10.1 is an original “DiscrmSwiss.lng” model. If everybody wishes to study how to use LINGO and many mathematical programming (MP) models, you download Linus textbook “Optimization Modeling with LINGO by Linus Schrage.” Everybody can master the theory of MP and how to make many MP-based models with many comments. “!..;” shows the comment. By these comments, you can understand the meaning of models.

The “SETS: … ENDSETS;” section defines one-dimensional sets and multi-dimensional sets. The combinations of one-dimensional sets define the multi-dimensional sets. “TEST:” is a one-dimensional set that defines two arrays such as “WGT and ZUSE.”

“DATA: … ENDDATA;” section defines the constant values and data sets. “TEST” has seven elements such as “Length, Left, Right, Bottom, Top, Diagonal, Good;” those are variable names. Set “OBS” has two arrays such as “DROP and SCORE.” Three scalar arrays such as “WGTSUSEDMX, WGTMX, DEPVAR” have three constants such as 2, 9999, and 7, respectively. “OBS, TSCR=” define two arrays. “OBS” is 200 elements from BN1 to BN200. “TSCR” is 200 rows and seven columns.

“SUBMODEL DISCRAMP: … ENDSUBMODEL” section defines submodel named “DISCRAMP.” If we use “DISCRAMP” in the Calc section, the defined character insert in that part.

In the CALC section, we can control the complex optimization model and draw graphs and print a customized report. By “@SOLVE(DISCRAMP)”, DiscrmSwiss.lng discriminates Swiss banknote data (Flury and Riedwyl 1988).

Program 10.1. Original DiscrmSwiss.lng

```
! Discriminant analysis by integer programming (DiscrmSwiss.lng);
! This is a form of categorical regression in which the dependent variable
is a categorical variable, e.g., Good or Bad.
 Basic idea:
   Given the values of various characteristics of an object, predict its
category, e.g.,
  Is a prospective customer good credit risk, or bad?
  Is a paper banknote good or counterfeit?
  Does a patient have a certain disease or not?
 We compute the weights in a scoring formula, so that
Score(i) >= 0 implies a good item, < 0 implies bad.
  There are various objectives one can use in finding an
optimal scoring function. Here we use the objective of
  Minimize the number of misclassifications;
! Keywords: Discriminant analysis, Classification, Clustering,
     Categorical regression, ChartScatter, Data Mining, Grouping,
     Scatter chart, Statistics;

SETS:
 TEST: WGT, ZUSE;
 OBS: DROP, SCORE;
 OXT(OBS, TEST): TSCR;
 OBS1(OBS): X1, Y1;
 OBS2(OBS): X2, Y2;
ENDSETS

DATA:
!  Genuine and counterfeit banknotes (100 Swiss Franks),
various measurements.
Banknotes BN1 to BN100 are genuine (Good=1),
all others are counterfeit (Good=0).
Dataset courtesy of H. Riedwyl, Bern, Switzerland;
 WGTSUSEDMX = 2;   ! Max # of weights to use;
 WGTMX = 99999;    ! Max absolute value of any weight;
 DEPVAR = 7;   ! Index of the dependent variable (Good);
 TEST= Length Left  Right   Bottom  Top  Diagonal  Good;
 OBS, TSCR=
BN1     214.8   131.0   131.1   9.0     9.7     141.0     1
BN2     214.6   129.7   129.7   8.1     9.5     141.7     1
..........................................................................
BN99    215.1   130.0   129.8   9.1     10.2    141.5     1
BN100   214.7   130.0   129.4   7.8     10.0    141.2     1
BN101   214.4   130.1   130.3   9.7     11.7    139.8     0
BN102   214.9   130.5   130.2   11.0    11.5    139.5     0
..........................................................................
BN199   214.7   130.7   130.8   11.2    11.2    139.4     0
BN200   214.3   129.9   129.9   10.2    11.5    139.6     0 ;
ENDDATA

SUBMODEL DISCRAMP:
! Minimize number of observations dropped to get a partition;
  MIN = OBJV;
    OBJV = @SUM( OBS( I): DROP( I));
! For bad observations, if DROP(I)=0, we want a strictly negative score;
@FOR( OBS(I)| TSCR( I, DEPVAR) #EQ# 0:
   SCORE( I) <= -1 + WGTMX*DROP( I);
    SCORE( I) >=    - WGTMX*(1- DROP(I));
    SCORE( I) =
    WGT0 + @SUM( TEST( J) | J #NE# DEPVAR: WGT( J)* TSCR(I,J));
    @FREE( SCORE(I));
     );
 ! For good observations, if DROP(I)=0, we want a strictly positive score;
 @FOR( OBS(I)| TSCR( I, DEPVAR) #EQ# 1:
    SCORE( I) >= 1 - WGTMX*DROP( I);
    SCORE( I) <= WGTMX*(1-DROP(I));
    SCORE( I) =
    WGT0 + @SUM( TEST( J) | J #NE# DEPVAR: WGT( J)* TSCR(I,J));
    @FREE( SCORE(I));
     );
  @FREE( WGT0);
  @FOR( TEST( J): @FREE( WGT( J));); ! The WFT(J) are unrestricted in sign;
  @FOR( OBS(I): @BIN( DROP(I))      ! The DROP(I) are 0 or 1;
      );
 ! Constraints limit number of nonzero weights;
  @FOR( TEST( K) | K #NE# DEPVAR:
     WGT( K) <= WGTMX*ZUSE( K);
    -WGT( K) <= WGTMX*ZUSE( K);
     @BIN( ZUSE( K));
      );
   @SUM( TEST( K) | K #NE# DEPVAR: ZUSE( K)) <= WGTSUSEDMX;
ENDSUBMODEL

CALC:
   @SOLVE( DISCRAMP);
 ! Get ready to plot a 2 dimensional subdimension;
 ! Set D1, D2 =  2 dimensions used;
   D1 = 0;
   @FOR( TEST( K) | ZUSE( k) #GT# 0.5:
     @IFC( D1 #EQ# 0:
         D1 = K;
        @ELSE
         D2 = K;
         );
       );
 ! Create set of the GOOD ones, with 2 dimensions in X1, Y1;
   @FOR( OBS(I) | TSCR( I, DEPVAR) #EQ# 1:
     @INSERT( OBS1, I);
     X1( I) = TSCR(I,D1);
     Y1( I) = TSCR(I,D2);
       );
 ! Create a set of BAD ones, with two dimensions in X2, Y2;
   @FOR( OBS(I) | TSCR( I, DEPVAR) #EQ# 0:
     @INSERT( OBS2, I);
     X2( I) = TSCR(I,D1);
     Y2( I) = TSCR(I,D2);
       );
  @WRITE( ' Measure      WGT', @NEWLINE(1));
  @WRITE( ' CONSTANT ', @FORMAT( WGT0, '10.3f'), @NEWLINE(1));
  @FOR( TEST( J) | J #NE# DEPVAR:
    @WRITE(  @FORMAT(  TEST(  J),'9s'),  @FORMAT(  WGT(  J),  '10.3f'),
 @NEWLINE(1));
      );
   @WRITE( @NEWLINE(1));
   @WRITE(' If CONSTANT + @SUM( TEST( j): WGT(j)*TSCR(i,j)) >= 0,',
 @NEWLINE(1));
   @WRITE('      Then predict as GOOD, else Predict as BAD.', @NEWLINE(1));
   @WRITE( @NEWLINE(1),'Number items incorrectly predicted= ', OBJV,
 @NEWLINE(1));

 ! Now do a scatter plot;
   @CHARTSCATTER( 'Swiss Bank Notes: Good vs. Counterfeit',!Chart title;
       @FORMAT(TEST(D1),"7s")+' MEASURE', !Legend for X axis;
       @FORMAT(TEST(D2),"7s")+' MEASURE', !Legend for Y axis;
       'Good', x1, y1,        !Point set 1;
       'Counterfeit', x2, y2); !Point set 2;
 ENDCALC
```

When we press the solve button in the LINGO menu bar, LINGO will output two windows. The Solution Report window is an optimization output described in Sect. 10.2.2. We introduce the result of Swiss banknote data that consists of six independent variables and two classes such as 100 genuine bills and 100 counterfeit bills identified by Good (1/0). Figure 10.1 is a LINGO chart window that is a scatter plot of two variables like the diagonal (y-axis) and bottom (x-axis). The two classes good (red) and counterfeit (green) appear to be almost separated.

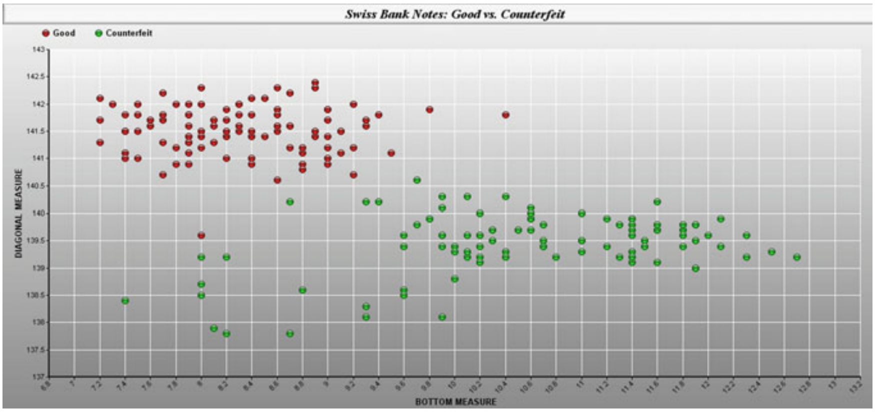


Fig. 10.1 LINGO chart window that is the graph of a scatter plot (diagonal by bottom)

10.2.2 Modified DiscrmSwiss.lng

Program 10.2 is a modified model for the use of cancer gene diagnosis. Mainly, we delete comments and graph output. In Data section, “OLE()” function inputs data by “TEST, OBS, TSCR=@OLE();” from values defined in Excel. LINGO and

Excel share the same array names such as TEST, OBS, and TSCR. Moreover, five underlined commands are essential.

1. First underlined command is crucial to find BGS. We can expect to control the number of genes in SM by changing this constant.
2. The second, third, and fourth underlined commands explain the meaning of DROP.
3. The fifth underlined commands are crucial to output the optimization results on Excel.

Program 10.2. A modified model for the use of cancer gene diagnosis.

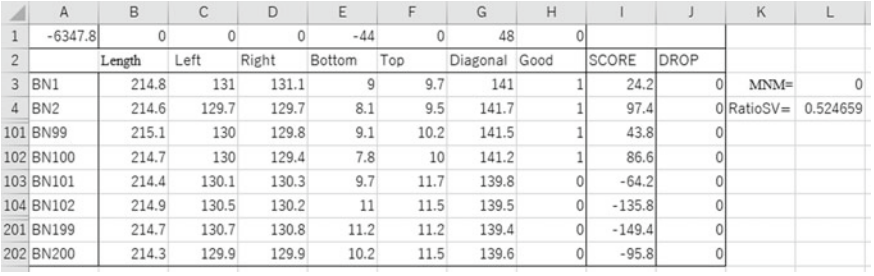
```
SETS:
 TEST: WGT, ZUSE;
 OBS: DROP, SCORE;
 OXT(OBS, TEST): TSCR;
ENDSETS

DATA:
 WGTSUSEDMX = 2; ! Max # of weights to use;
 WGTMX = 99999;  ! Max absolute value of any weight;
 DEPVAR = 7;    ! Index of the dependent variable (Good);
 TEST, OBS, TSCR=@OLE();
ENDDATA
SUBMODEL DISCRAMP:
! Minimize number of observations dropped to get a partition;
  MIN = OBJV;
    OBJV = @SUM( OBS( I): DROP( I));
! For bad observations, if DROP(I)=0, we want a strictly negative score;
@FOR( OBS(I)| TSCR( I, DEPVAR) #EQ# 0:
   SCORE( I) <= -1 + WGTMX*DROP( I);
   SCORE( I) >=    - WGTMX*(1- DROP(I));
   SCORE( I) =   WGT0 + @SUM( TEST( J) | J #NE# DEPVAR: WGT( J)* TSCR(I,J));
```

```
!DROP(I)=@IF(SCORE(I) #GE# 0,1,0);
   @FREE( SCORE(I));
    );
! For good observations, if DROP(I)=0, we want a strictly positive score;
@FOR( OBS(I)| TSCR( I, DEPVAR) #EQ# 1:
   SCORE( I) >= 1 - WGTMX*DROP( I);
   SCORE( I) <= WGTMX*(1-DROP(I));
   SCORE( I) = WGT0 + @SUM( TEST( J) | J #NE# DEPVAR: WGT( J)* TSCR(I,J));
!DROP(I)=@IF(SCORE(I) #LE# 0,1,0);
   @FREE( SCORE(I));
    );
 @FREE( WGT0);
 @FOR( TEST( J): @FREE( WGT( J));); ! The WFT(J) are unrestricted in sign;
 @FOR( OBS(I): @BIN( DROP(I))      ! The DROP(I) are 0 or 1; );
ENDSUBMODEL

CALC:
 @SOLVE( DISCRAMP);
 @OLE()=DROP, SCORE, WGT, WGT0;
ENDCALC
```

If we define three Excel range names such as OBS (A3: A202), TSCR (B3: H202) and TEST (B2: H2) in Fig. 10.2, “@OLE()” reads three range values from Excel and stores those values on three LINGO arrays such as TEST, OBS, and TSCR. Two one-dimensional sets such as OBS and TEST have 200 elements from BN1 to BN200 and seven elements that shows variable name. One-dimensional set of OBS and TEST defines two-dimensional set TSCR with 200 bills by seven variables. “@OLE()=DROP, SCORE, WGT, WGT0;” outputs four optimization results on Excel by “@OLE()=” function. WGT is LINGO array name and Excel cell range name (B1: H1). Although H1 is usually the intercept of Linus’s LDF, he uses this one as the label of good and counterfeit bills. We output the intercept value on cell A1. Thus, Linus LDF is : f = −44 * Bottom + 48 * Diagonal − 6347.8. The 200 DSs output on cell range (I3: I202). Because all bills are separable, 200 cells (J3: J202) have 0s. Thus, MNM = 0 by the formula (=SUM(J3: J202) that shows MNM. Cell L4 is RatioSV defined by “=100 * (MIN(I3:I102) − MAX(I103:I202))/(MAX(I3:I102) − MIN(I103:I202))”. Two variables model (Bottom, Diagonal) is SM and BGS having RatioSV = 0.524. Because 130 BGSs of Alon’s microarray are less than 1%, both results indicate the RatioSV of BGS may be tiny.



	A	B	C	D	E	F	G	H	I	J	K	L
1	-6347.8	0	0	0	-44	0	48	0				
2		Length	Left	Right	Bottom	Top	Diagonal	Good	SCORE	DROP		
3	BN1	214.8	131	131.1	9	9.7	141	1	24.2	0	MNM=	0
4	BN2	214.6	129.7	129.7	8.1	9.5	141.7	1	97.4	0	RatioSV=	0.524659
101	BN99	215.1	130	129.8	9.1	10.2	141.5	1	43.8	0		
102	BN100	214.7	130	129.4	7.8	10	141.2	1	86.6	0		
103	BN101	214.4	130.1	130.3	9.7	11.7	139.8	0	-64.2	0		
104	BN102	214.9	130.5	130.2	11	11.5	139.5	0	-135.8	0		
201	BN199	214.7	130.7	130.8	11.2	11.2	139.4	0	-149.4	0		
202	BN200	214.3	129.9	129.9	10.2	11.5	139.6	0	-95.8	0		

Fig. 10.2 Data input and solution output on Excel by @OLE()

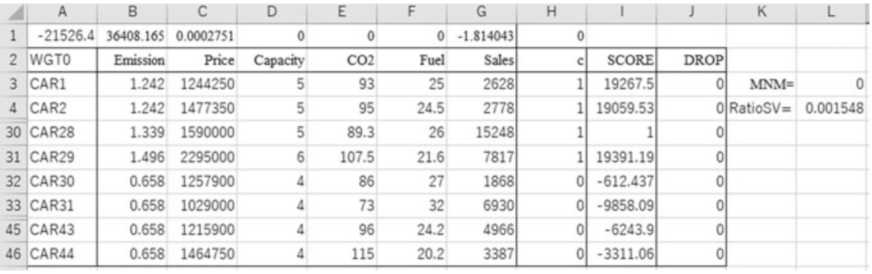
10.2.3 Japanese Cars Data

Japanese car data consists of 44 cars and six variables. We had already known that two one-variable models are BGSs (Shinmura 2016). In Program 10.3, we modify only two underlined statements of a modified model for Japanese car data. Because Swiss banknote data and Japanese cars data have six variables, there is no change of “DEPVAR = 7.” However, we must change the variable names such as “TEST = Emission Price Capacity CO2 Fuel Sales c;”. When we discriminate Alon’s microarray with 2,000 genes, we must change “DEPVAR=2001;” in addition to “TEST=X1–X2001;”.

Program 10.3. How to apply for Other Data

```
DEPVAR = 7;   ! Index of the dependent variable (Good);
 TEST= Emission Price Capacity CO2 Fuel Sales c;
```

Japanese car data consists of two classes such as the 29 regular cars (A2: A31) and the 15 small cars (A32: A46) in Fig. 10.3. Set "OBS" has a one-dimensional set with 44 labels from CAR1 to CAR44 (A2: A46). There are six independent variables (B2: G2) and the intercept (H2). Set "TEST" has a one-dimensional set with seven labels such as Emission, Price, Capacity, CO2, Fuel, Sales, and c (B2: H2). Two one-dimensional sets such as OBS and TEST define two-dimensional set OXT that defines two-dimensional array TSCR. Thus, TSCR defines two-class discriminant data with 44 cars and six variables. "1/0" of the intercept corresponds the regular or small car. After optimization, LINGO output six discriminant coefficients on cell range B1: G1. Because this model separates the intercept and six variables, the intercept is on A1. The modified model finds LDF as follows: f = 36408.165 * Emission + 0.0002751 * Price − 1.814043 * Sales − 21526.4. SCORE on I3: I46 are 44 discriminant scores. "1/0" on OMIT (J3: J46) shows the status wheather 44 cars are classified or misclassified into each class. Cell L3 contains the formula "=SUM(J3: J46)" and shows MNM. Because we know each variable of emission and capacity is BGS, three variables (emission, price, and sales) is SM.



	A	B	C	D	E	F	G	H	I	J	K	L
1	-21526.4	36408.165	0.0002751	0	0	0	-1.814043	0				
2	WGT0	Emission	Price	Capacity	CO2	Fuel	Sales	c	SCORE	DROP		
3	CAR1	1.242	1244250	5	93	25	2628	1	19267.5	0	MNM=	0
4	CAR2	1.242	1477350	5	95	24.5	2778	1	19059.53	0	RatioSV=	0.001548
30	CAR28	1.339	1590000	5	89.3	26	15248	1	1	0		
31	CAR29	1.496	2295000	6	107.5	21.6	7817	1	19391.19	0		
32	CAR30	0.658	1257900	4	86	27	1868	0	-612.437	0		
33	CAR31	0.658	1029000	4	73	32	6930	0	-9858.09	0		
45	CAR43	0.658	1215900	4	96	24.2	4966	0	-6243.9	0		
46	CAR44	0.658	1464750	4	115	20.2	3387	0	-3311.06	0		

Fig. 10.3 Optimization results in Excel (WGTSVSEDMX=5, 4, 3)

Table 10.1 shows the Japanese car data on cell range (A1: L46).

Table 10.1 Japanese car data

−21526.4	36408.165	0.0002751	0	0	0	−1.814043	0			
WGT0	Emission	Price	Capacity	CO_2	Fuel	Sales	c	SCORE	DROP	
CAR1	1.242	1,244,250	5	93	25	2628	1	19267.5	0	MNM = 0
CAR2	1.242	1,477,350	5	95	24.5	2778	1	19059.53	0	RatioSV = 0.001548
CAR3	1.599	1,711,500	5	132	17.6	2496	1	32633.22	0	
CAR4	2.457	2,362,500	5	166	14	1742	1	65418.31	0	
CAR5	2.362	4,428,350	7	137	17	2535	1	61089.35	0	
CAR6	1.797	2,090,000	7	142	16.4	2493	1	39951.61	0	
CAR7	1.329	1,350,000	5	106	21.8	7750	1	13172.58	0	
CAR8	1.986	2,330,000	7	171	13.6	4678	1	42935.1	0	
CAR9	2.362	3,950,000	7	137	17	3762	1	58731.92	0	
CAR10	2.362	2,960,000	7	187	12.4	3870	1	58263.64	0	
CAR11	2.493	3,040,000	5	99	23.4	1499	1	67356.21	0	
CAR12	1.496	2,440,000	5	116	20	4771	1	24956.65	0	
CAR13	3.456	5,400,000	5	166	14	3190	1	99999	0	
CAR14	1.496	1,655,000	7	135	17.2	2418	1	29009.13	0	
CAR15	1.986	2,520,000	7	171	13.6	2877	1	46254.46	0	
CAR16	0.996	1,000,000	5	112	20.8	4098	1	7577.255	0	
CAR17	1.797	2,170,000	5	71	32.6	22,998	1	2776.677	0	
CAR18	1.329	1,440,000	5	126	18.4	3277	1	21311.55	0	
CAR19	1.995	2,607,150	5	185	14.2	1828	1	48509.04	0	
CAR20	2.488	3,612,000	7	215	10.8	1517	1	67298.88	0	
CAR21	1.498	1,529,850	5	129	18	2969	1	28047.98	0	
CAR22	1.498	1,620,150	5	135	17.2	2642	1	28666.01	0	

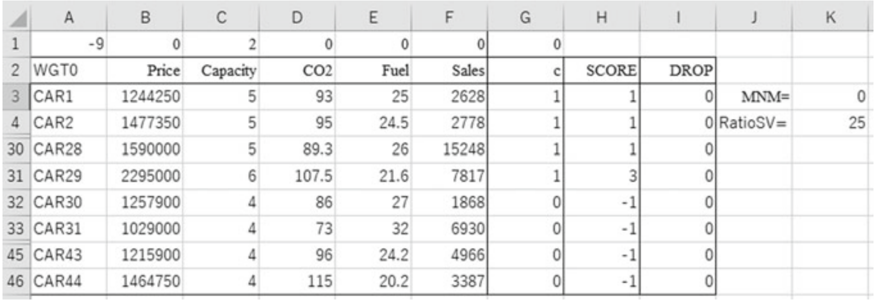
(continued)

Table 10.1 (continued)

−21526.4	36408.165	0.0002751	0	0	0	−1.814043	0			
WGT0	Emission	Price	Capacity	CO_2	Fuel	Sales	c	SCORE	DROP	
CAR23	1.997	2,499,000	8	159	14.6	6512	1	40055.13	0	
CAR24	1.498	1,596,000	5	116	18	2169	1	29517.41	0	
CAR25	1.498	1,339,800	5	129	18	3656	1	26749.45	0	
CAR26	1.198	103,6350	5	103	22.6	1727	1	19242.8	0	
CAR27	1.997	2,088,000	8	173.3	13.4	4396	1	43780.57	0	
CAR28	1.339	1,590,000	5	89.3	26	15,248	1	1	0	
CAR29	1.496	2,295,000	6	107.5	21.6	7817	1	19391.19	0	
CAR30	0.658	1,257,900	4	86	27	1868	0	−612.437	0	
CAR31	0.658	1,029,000	4	73	32	6930	0	−9858.09	0	
CAR32	0.658	1,344,000	4	153	15.2	1544	0	−1	0	
CAR33	0.658	1,207,500	4	103	22.5	3884	0	−4283.41	0	
CAR34	0.658	1,212,750	4	93	25	13,953	0	−22547.6	0	
CAR35	0.658	1,133,000	4	86	27	1512	0	−1	0	
CAR36	0.658	1,320,000	4	94	24.8	9289	0	−14057.4	0	
CAR37	0.658	950,000	4	95	24.5	17,603	0	−29241.1	0	
CAR38	0.658	1,220,000	4	86	27	10,147	0	−15641.3	0	
CAR39	0.658	1,120,000	4	91	25.5	1746	0	−429.062	0	
CAR40	0.658	1,240,000	4	104.6	22.2	2860	0	−2416.89	0	
CAR41	0.658	1,160,000	4	118.5	19.6	4290	0	−5032.98	0	
CAR42	0.657	982,000	4	110	21.2	2127	0	−1194.59	0	
CAR43	0.658	1,215,900	4	96	24.2	4966	0	−6243.9	0	
CAR44	0.658	1,464,750	4	115	20.2	3387	0	−3311.06	0	

10.2.4 Thank You for the Fabulous Model Creator Linus

We had already found two BGSs such as “emission and capacity” because these two variables can separate two classes. When we discriminate five-variable model without “Emission rate,” we obtain only “capacity” as BGS in Fig. 10.4. Linus LDF is f = 2 * Capacity − 9. This LDF indicates if “capacity” is higher than 4.5, the car belongs to regular car, otherwise small car. This result is as same as RIP explained in Chap. 7 of Shinmura (2016, 2018a, b). Although Capacity is BGS, RatioSV is 25%. This result is as same as RatioSVs of SMs.

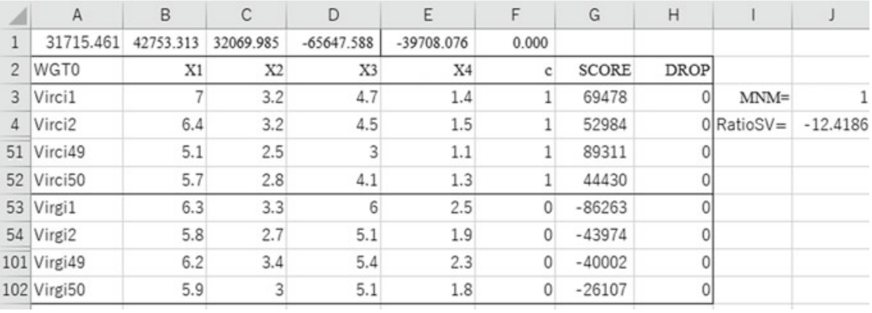


	A	B	C	D	E	F	G	H	I	J	K
1	-9	0	2	0	0	0	0				
2	WGT0	Price	Capacity	CO2	Fuel	Sales	c	SCORE	DROP		
3	CAR1	1244250	5	93	25	2628	1	1	0	MNM=	0
4	CAR2	1477350	5	95	24.5	2778	1	1	0	RatioSV=	25
30	CAR28	1590000	5	89.3	26	15248	1	1	0		
31	CAR29	2295000	6	107.5	21.6	7817	1	3	0		
32	CAR30	1257900	4	86	27	1868	0	-1	0		
33	CAR31	1029000	4	73	32	6930	0	-1	0		
45	CAR43	1215900	4	96	24.2	4966	0	-1	0		
46	CAR44	1464750	4	115	20.2	3387	0	-1	0		

Fig. 10.4 Capacity is BGS (DEPVAR = 6, WGTSUSEDMX = 5)

10.2.5 Iris Data

Iris data consists of four measurements of three species such as Setosa, Vircicle, and Virginica (Anderson 1945; Fisher 1956). Because Setosa and the other two species are LSD, we discriminate two classes such as Vircicle and Virginica. Although Anderson collected these data, this data is called Fisher’s iris data because Fisher used this data for validation of Fisher’s LDF. Because there are four independent variables, we set DEPVAL = 5. Figure 10.5 shows the result. Because Fisher’s LDF misclassifies the 34th versicolor, MNM = 1. Although RatioSV = −12.4186, the RatioSV is not used for overlapping data.



	A	B	C	D	E	F	G	H	I	J
1	31715.461	42753.313	32069.985	-65647.588	-39708.076	0.000				
2	WGT0	X1	X2	X3	X4	c	SCORE	DROP		
3	Virci1	7	3.2	4.7	1.4	1	69478	0	MNM=	1
4	Virci2	6.4	3.2	4.5	1.5	1	52984	0	RatioSV=	-12.4186
51	Virci49	5.1	2.5	3	1.1	1	89311	0		
52	Virci50	5.7	2.8	4.1	1.3	1	44430	0		
53	Virgi1	6.3	3.3	6	2.5	0	-86263	0		
54	Virgi2	5.8	2.7	5.1	1.9	0	-43974	0		
101	Virgi49	6.2	3.4	5.4	2.3	0	-40002	0		
102	Virgi50	5.9	3	5.1	1.8	0	-26107	0		

Fig. 10.5 Iris data (DEPVAR = 5)

10.3 Six MP-Based LDFs and LINGO Models

We explain six MP-based LDFs and LINGO models.

(1) RIP

RIP in Eq. (10.1) can find the interior point of OCP directly, all NMs of which become MNM. OCP is the feasible region, but all interior points have the same minimum values such as MNM. It can solve the Problem1 and Problem2 correctly. Because it can decompose common data and six microarrays into many SMs, it can solve Problem5 very easy. Because we developed the best model by Method1, we had ignored the natural feature selection obtained by common data before Method2. We suppose to discriminate the Golub microarray with 72 cases and 7,129 variables. Most researchers erroneously understand the gene analysis is the severe problem because of small n and large p problem. However, RIP, Revised LP-OLDF, and H-SVM can discriminate microarrays within 20 s and find those are LSD (Fact3).

$$MIN = \sum e_i;\ y_i^*\left({}^t\boldsymbol{x}_i\boldsymbol{b} + b_0\right) >= 1 - M * e_i; \tag{10.1}$$

b_0 free decision variables.
M 10000 (Big M constant)
e_i 1/0 binary integer.

LINGO can define MP models by natural expression and SET expression. The natural expression is effortless because it is almost the same as the arithmetic expression. However, it is hard to develop a large-scale MP model. The SET expression can build a complex MP optimization system that consists of the combinations of SET, DATA, SUBMODEL, and CALC sections. SUBMODEL section defines MP-based LDFs as follows. We explain how RIP can discriminate microarray by LINGO Program3. "SUBMODEL" section of LINGO defines RIP in Eq. (10.2). "@SUM and @FOR" are essential LINGO SET looping functions. "@" means the LINGO functions.

(1) The objective function (MIN=@SUM(N(I): E(I));) minimizes the summation of e_i those are 1/0 binary integers. Thus, it minimizes NM and finds the minimum NM using @SUM LINGO function.
The natural expression is "Min = $\sum_{i=1,\ldots,72} e_i$."
(2) The 72 constraint expressions limit the extended DS of each case to be higher than 1 of SV.
The natural expression of "@SUM(P1(J1): IS(I, J1) * VARK(J1) * CHOICE(J1)) > 1 − 10000 * E(I));" is as follows:

$$\sum\nolimits_{\mathrm{I}} \sum\nolimits_{\mathrm{J1}} \mathrm{IS(I, J1)} * \mathrm{VARK(J1)} * \mathrm{CHOICE(J1))} > 1 - 10000 * \mathrm{E(I)}.$$
$$\mathrm{J1} = 1, \ldots, 7130,$$
$$\mathrm{I} = 1, \ldots, 72. \tag{10.2}$$

Thus, Eq. (10.2) means the 72 constraints. The left side of the constraint expression is a summation of 7,130 elements of IS(I, J1) * VARK(J1) * CHOICE(J1). In the case of LSD, all e_i of the right side become zero.

"@FOR(N(I): @SUM …);" in Eq. (10.3) defines 72 constraints of 72 patients having 7,130 coefficients including the constant means as follows:

$$\sum\nolimits_{\mathrm{j1}=1,\ldots,7130} \mathrm{IS(I, J1)} * \mathrm{VARK(J1)} * \mathrm{CHOICE(J1))} > 1 - 10000 * \mathrm{E(I)}.$$
$$\text{For I} = 1, \ldots, 72. \tag{10.3}$$

Thus, Eq. (10.3) corresponds to the 72 constraints in Eq. (10.1). Although old-style MP solver was hard to construct a large-scale model, LINGO is very easy to define plural complex models as same as small-size models and control those models.

"@FOR (P1(J1): @FREE(VARK(J1)));" means as follows:

$$\text{@FREE(VARK(J1)); J1} = 1, \ldots, 7130. \tag{10.4}$$

Equation (10.4) means 7,130 discriminant coefficients such as VARK(J1) including the intercept are free variables. Because a decision variable is nonnegative real variable in default of MP world, it needs to set a free variable in MP. "@FOR(N(i): @BIN (E(i)));" defines 72 E(i) are 1/0 binary integers. Because all these decision variables (e_i) are not related to the executable area, all the optimal solutions are MNM = 0. In this sense, this model is a new model not existing in MP.

"SUBMODEL RIP:, …, ENDSUMMODEL" defines RIP by the LINGO IP model. The IP algorithm of LINDO products use the branch & bound (B&B). We are worried about another IP algorithm which cannot decompose the microarrays into SM. In the CALC section, we can program and control MP-based models. "@SOLVE(RIP);" discriminates the dataset by RIP. "!…,;" is a comment.

LINDO Systems Inc. offers free evaluation version, manual and Prof. Linus textbooks from http://www.lindo.jp/.

```
SUBMODEL RIP:                                          (10.1)
   MIN=ER;
   ER=@SUM(N(I): E(I));
     ! You can output array ER on Excel file;
   @FOR(N(I): @SUM(P1(J1): IS(I, J1)*VARK(J1)*
          CHOICE(J1)) > 1-10000*E(I));
      @FOR(P1(J1): @FREE(VARK(J1)));
      @FOR(N(I): @BIN(E(I)));
ENDSUBMODEL
  ……………….
CALC:
  ……………….
  @SOLVE(RIP);
  ………………..
ENDCALC
```

(2) Revised LP-OLDF

If e_i is nonnegative real variable, Eq. (10.1) changes Revised LP-OLDF in (10.5). Thus, the feasible region (OCP) is as same as RIP. However, LP finds one of the vertexes. Because the dimension of the vertex is less than equal n, the number of nonzero coefficients of Revised LP-OLDF and RIP is less than equal n in the first step of Method2.

$$\begin{aligned} \mathrm{MIN} &= \sum \mathrm{e_i}; \\ \mathrm{y_i} * \left({}^{t}\mathbf{x}_{\mathrm{i}}\mathbf{b} + \mathrm{b_0}\right) &>= 1 - \mathrm{M} * \mathrm{e_i}; \end{aligned} \tag{10.5}$$

b_0 free decision variables.
e_i nonnegative real variables.

Thus, if we drop "@FOR(N(I): @BIN(E(I)));" in Eq. (10.1), it becomes a Revised LP-OLDF. Many statisticians claim we do not explain the algorithm of MP-based LDFs in our papers. If we define the models of MP-based LDFs, LP, IP, QP, and NLP solvers of LINGO solve the many problems. LP, IP, QP, and NLP are the algorithm of our theory.

```
SUBMODEL LP:                                                    (10.5)
  MIN=ER;  ER=@SUM(N(I): E(I));
      @FOR(N(I): @SUM(P1(J1): IS(I, J1)*VARK(J1)*
       CHOICE(J1)) > 1-10000*E(I));
      @FOR(P1(J1): @FREE(VARK(J1)));
ENDSUBMODEL
```

(3) Revised IPLP-OLDF

Revised IPLP-OLDF is a mixture model of Revised LP-OLDF in the first step and RIP in the second step. Thus, SUBMODEL and CALC sections define Revised IPLP-OLDF in Eq. (10.6). At first, Revised LP-OLDF discriminates all cases and outputs 0/1 information to the array CONSTANT. If Revised LP-OLDF classifies the case $\mathbf{x}_i$ correctly, $e_i = 0$ and CONSTANT(I) = 0. Otherwise, $e_i = 1$ and CONSTANT(I) = 1.

"@FOR(N(I): @IFC (E(I) #EQ# 0: CONSTANT (I) = 0; @ELSE CONSTANT(I) = 1;));"

Next, RIP discriminates only cases misclassified by Revised LP-OLDF. "SUMMODEL IPLP" omits the classified cases by the following LINGO functions.

@FOR(N(I)| CONSTANT (I) #EQ# 0: E(I) = 0);

Thus, "@SOLVE(IPLP);" discriminates the restricted cases by RIP using the information of "CONSTANT(I) = 1."

```
SUBMODEL IPLP:                                                    (10.6)
  MIN = @SUM(N(I): E(I));
     @FOR(N(I): @SUM(P1(J1): IS(I, J1)*
     VARK(J1)*CHOICE(J1) )>=1-BIGM*E(I));
     @FOR(P1(J1): @FREE(VARK(J1)));
     @FOR(N(I)| CONSTANT(I) #EQ# 1: @BIN(E(I)));
     @FOR(N(I)| CONSTANT(I) #EQ# 0: E(I)=0);
ENDSUBMODEL

CALC:
......................
@SOLVE(LP);
    @FOR(N( I): @IFC( E( I) # EQ # 0: CONSTANT(I)=0; @ELSE CONSTANT(I)=1;));

NM=0; NMP=0; Z=0;
@SOLVE(IPLP); ! a restricted RIP for CONSTANT(i)=1;
.......................
ENDCALC
```

(4) Soft-margin SVM

Equation (10.7) is S-SVM. If we set c = 10^4 or c = 1 in the DATA section, it becomes SVM4 or SVM1 because there is no proper rule to decide penalty c. Our examinations tell us that SVM4 is almost better than SVM1 after many examinations. If we omit "c * Σe_i" and "$-e_{i,}$" it becomes H-SVM in Eq. (10.8). QP solves both SVMs. The second object function such as "c * Σe_i" is as same as Revised LP-OLDF. Thus, we can conclude that QP solver and the first object function such as "$||\mathbf{b}||^2/2$" prevent the natural feature selection by three SVMs. That QP solver can find only one minimum solution in the domain.

$$\begin{aligned} \mathrm{MIN} &= ||\mathbf{b}||^2/2 + \mathrm{c} * \sum \mathrm{e_i}; \\ \mathrm{y_i} * \left({}^{t}\mathbf{x}_i\mathbf{b} + \mathrm{b_0}\right) &>= 1 - \mathrm{e_i}; \end{aligned} \tag{10.7}$$

c penalty c to combine two objectives.
e_i nonnegative real value.

```
SUBMODEL SVM:                                              (10.7)
 MIN=@SUM(P1(J1): VARK(J1)^2)/2
           +Penalty*@SUM(N(I): E(I));
 @FOR(N(I): @SUM(P1(J): IS(I, J)*VARK(J)
              *CHOICE(J)) >= 1-E(I));
ENDSUBMODEL
.........................
CALC:
.......................
@SOLVE(SVM);
.......................
ENDCALC
```

(5) H-SVM

Equation (10.8) is an H-SVM that clearly defines LSD, and the generalization ability is defined by “minimization of $||\mathbf{b}||^2/2$”. However, most SVM researchers seem to have passed through the LSD-discrimination because H-SVM cannot solve the overlapping data. Moreover, SVM researchers and users pay their attention to the kernel SVM. We guess no researchers do not use H-SVM for the cancer gene analysis. They lost the chance that the signal of microarrays is LSD.

$$\mathrm{MIN} = ||\mathbf{b}||^2/2; \quad \mathrm{y_i} * \left({}^{\mathrm{t}}\mathbf{x}_{\mathrm{i}}\mathbf{b} + \mathrm{b_0}\right) >= 1 - \mathrm{e_i}; \tag{10.8}$$

e_i nonnegative real value.

```
SUBMODEL HSVM:                                              (10.8)
  MIN=@SUM(P1(J1): VARK(J1)^2)/2;
      @FOR(N(I): @SUM(P1(J): IS(I, J)*VARK(J)
       *CHOICE(J)) >= 1);
ENDSUBMODEL
……………………
CALC:
………………..
@SOLVE(HVM);
………………..
ENDCALC
```

10.4 LINGO Program3 of Method2

We introduce LINGO Program3 of Method2 that consists of four sections, such as SETS, DATA, SUBMODEL, and CALC sections. We explain to discriminate Golub dataset with 72 cases and 7129 genes by RIP. SETS section defines set with arrays in the form: “set-name: [array-names];.” “P, P1, P2, N, SN” are five one-dimensional sets. In the DATA section, “P = 1 … 7129;” means set “P” has 7129 elements. Without the definition of “P = 1 … 7129;”, LINGO estimates it by checking the array. Above five sets have 7129, 7130, 7131, 72, and 600 elements, respectively. Set “P” corresponds 7129 genes for Golub or Shipp datasets. Set “P1” has 7130 elements of “7129 genes + intercept” that has three arrays having 7130 elements such as “CHOICE, VARK, and MATRYOSHKA.” Set “N” has two arrays having 72 elements such as “E and DS” that correspond 1/0 binary decision variables and discriminant scores for 72 patients. Set “SN” has four arrays having 600 elements such as “SM, IT, T, NM” that store the results of SMs obtained by Program3. Because we cannot estimate the number of SMs, we set it 600. Set “DN (N, P1)” is two-dimensional set defined by two one-dimensional sets of “N and P1” that is 64 * 7130 elements. “IS” is Golub dataset. “SNCHOICE (SN, P1)” is a two-dimensional set made by two one-dimensional sets of “SN and P1” that is 600 × 7130 elements. Two arrays “CHOICE100 and VARK100” store 600 choice patterns and discriminant coefficients.

```
MODEL:                                                        (10.9)
SETS:
  P; P1: CHOICE, VARK, MATRYOSHKA; P2; N: E, DS;
 SN: SM, IT, T, NM;
 D (N, P1): IS;
 SNCHOICE (SN, P1): CHOICE100, VARK100;
ENDSETS
```

DATA section defines the element number of sets, the constant such as the penalty "c," and "IS=@OLE();." "IS" is the Excel cell array name that store Golub dataset. "@OLE ()" reads this data and LINGO stores this data as LINGO array name.

```
DATA:                                                          (10.10)
   P=1..7129; P1=1..7130; P2=1..7131; N=1..72; SN=1..600;
   IT1=600; IT2=5; PENALTY=10000;
  IS =@OLE();
ENDDATA
```

In the following SUBMODEL section, we define five MP-based LDFs. If we set "PENALTY = 10000;" in the DATA section, it becomes SVM4. If we change "PENALTY = 1;", it becomes SVM1. Although many researchers spend many research times to study MP-based LDFs by papers and books, it is effortless for them to discriminate their research datasets. We claim the right software improves our intelligence productivity.

```
SUBMODEL RIP:                                                  (10.11)
……….
ENDSUBMODEL

SUBMODEL LP:
……….
ENDSUBMODEL

SUBMODEL IPLP:
……….
ENDSUBMODEL

SUBMODEL HSVM:
……….
ENDSUBMODEL

SUBMODEL SVM:
……….
ENDSUBMODEL
```

```
SUBMODEL LP:                                                   (10.12)
 MIN=ER; ER=@SUM(N(I): E(I));
    @FOR(N(I): @SUM(P1(J1): IS(I, J1)*VARK(J1)*
       CHOICE(J1)) > 1-10000*E(I));
  @FOR(P1(J1): @FREE(VARK(J1)));
ENDSUBMODEL
```

```
SUBMODEL SVM:                                                  (10.13)
  MIN=ER;
 ER=@SUM(P1(J1): VARK(j1)^2)/2
     +Penalty*@SUM(N(I): E(I));
  @FOR(N(I): @SUM(P1(J): IS(I, J)*VARK(J)
      *CHOICE(J)) >= 1-E(I));
ENDSUBMODEL
```

```
SUBMODEL HSVM:                                                  (10.14)
  MIN=ER; ER=@SUM(P1(J1):VARK(J1)^2)/2;
      @FOR(N(I): @SUM(P1(J): IS(I, J)*VARK(J)
    *CHOICE(J)) >= 1);
ENDSUBMODEL
```

In the small loop of Method2, RIP repeatedly discriminates the dataset “IT2” times. First, when RIP discriminates the dataset, the 7130 discriminant coefficients with the intercept are stored on “VARK.” If the coefficient is zero, the value of “CHOICE” becomes zero, and this variable is removed from the second discrimination. If the coefficient is not zero, the value of “CHOICE” becomes 1, and the variable is included in the next discrimination. Second, RIP discriminates the dataset for the discriminant model with value 1 of “CHOICE.” It finds only 34 coefficients are not zero and 7095 coefficients are 0. Third, 34 coefficients decrease to 11 coefficients. Because fourth discrimination is as same as third discrimination, LINGO Program3 stop here. However, if “IT = 11”, we repeat to discriminate 11 times and end small loop (LOOP2). In big loop (LOOP1), selected 11 genes become MATRYOSHKA = 1 and Method2 omit the 11 genes from 7129 genes. In the second big loop, RIP discriminates the 7118 reduced gene space and finds the second SM2 (MATRYOSHKA=2).

```
CALC:                                                          (10.15)
@SET('DEFAULT'); @SET('TERSEO',2); (10.14)
PP=@SIZE(P);
K1=1; S=0;
@FOR (P1(J1): CHOICE(J1) = 1; MATRYOSHKA(J1) = 0; );
@WHILE (K1 #LE# IT1:
K2=1;
@WHILE(K2 #LE# IT2:
@FOR(P1( J1): VARK( J1) = 0; @RELEASE( VARK( J1)));
IC=0;
@SOLVE(HSVM);
@FOR(N(I):DS(I)=@SUM(P1(J1): IS(I, J1) * VARK(J1) * CHOICE(J1)));
@FOR(N(I): @IFC(DS(I) #LT# 0: IC=IC+1));
@FOR( P1(J1)| J1 #LE# PP:@IFC(VARK(J1) #EQ# 0: CHOICE(J1)=0; @ELSE
CHOICE(J1)=1;));
K2=K2+1);
S=S+1;
SM(S)=K1; IT(S)=K2; NM(S)=IC; T(S)=@SUM(P1(J1): CHOICE(J1)) - 1;
@FOR( P1( J1): CHOICE100(S, J1)=CHOICE(J1));
@FOR( P1(J1)| J1 #LE# PP: @IFC(VARK(J1) #NE# 0: MATRYOSHKA(J1)=k1 ));
@FOR( P1(J1)| J1 #LE# PP: @IFC(MATRYOSHKA(J1) #NE# 0: CHOICE(J1)=0; @ELSE
CHOICE(J1)=1;  ));
@OLE( )=MATRYOSHKA,CHOICE,VARK;
@IFC( IC #GE# 2: K1=IT1+1; @ELSE K1=K1+1 ));
ENDCALC
```

In the second DATA section, Program3 outputs the results of SMs on Excel cell array names such as “SM, IT, T, NM, CHOICE100, VARK100.”

```
DATA:                                                     (10.16)
  @OLE() = SM, IT, T, NM, CHOICE100, VARK100;
ENDDATA
END
```

10.5 Validation Method2 by LINGO Program1 Using Common Data

After publishing the Springer book (Shinmura 2016), we are afraid not to validate LINGO Program3. Thus, we develop Program1 (six LDFs’ version) that discriminates the data by six MP-based LDFs and outputs six discriminant coefficients. Because Swiss banknote and Japanese automobile data are useful as test datasets, Program1 can validate these datasets very easy. Next, we validate microarray using Program1 and simulate Program3 by manual operation.

```
MODEL:                                                 (10.17)
SETS:
   P; P2; P1: VARK, VARK0, CHOICE;
  N: E, CONSTANT, SCORE;
  MS: ; V2::
  D(N, P1):IS;
  G2: IC, NP, ZERO;
  VG(V2, P1):VARK100,VARK50;
ENDSETS

DATA:
  P=1..6; P1=1..7; P2=1..8;
  N=1..44; G2=1..6; MS=1; V2=1..6;  BIGM=10000;
  CHOICE = @OLE();
  IS = @OLE();
ENDDATA
```

```
Insert six MP-based LDFs;                                 (10.18)
```

```
CALC:                                                            (10.19)
@SET('DEFAULT'); @SET('TERSEO',2);
 G=1;
 NM=0; NMP=0; Z=0;
@FOR(P1(J1): VARK(J1) = 0; @RELEASE ( VARK ( J1)));
@SOLVE(RIP);
@FOR(P1(J1): VARK100(G, J1) = VARK(J1););
@FOR(N(I): SCORE(I)=@SUM(P1(J1): IS(I, J1)
                  *VARK(J1) * CHOICE(J1)));
 @FOR(N(I): @IFC(SCORE(I) #EQ# 0:  Z=Z+1));
 @FOR(N(I): @IFC(SCORE(I) #LT# 0:  NM=NM+1));
  @FOR(N(I): @IFC(SCORE(I) #GT# 0:  NMP=NMP+1));
 IC(G)=NM; ZERO(G)=Z; NP(G)=NMP;

 G=2;
 @SOLVE(LP);
  @FOR(N(I): @IFC(E(I) #EQ# 0: CONSTANT(I)=0; @ELSE CONSTANT(I)=1;));
 NM=0; NMP=0; Z=0;
 @SOLVE(IPLP);
  @FOR(P1(J1): VARK100(G, J1)  =  VARK(J1)
           *CHOICE(J1));
 @FOR(n(I): SCORE(I)=@SUM(P1(J1): IS(I, J1)
            *VARK(J1) * CHOICE(J1)));
  @FOR(n(I): @IFC(SCORE(I) #EQ# 0:  Z=Z+1));
  @FOR(n(I): @IFC(SCORE(I) #LT# 0:  NM=NM+1));
  @FOR(n(I): @IFC(SCORE(I) #GT# 0:  NMP=NMP+1));
 IC(G)=NM; ZERO(G)=Z; NP(G)=NMP;

G=3;
…………………………….
@SOLVE(LP);
…………………………….
G=4;
@SOLVE(HSVM);
…………………………….
G=5;
PENALTY=10000; NM=0; NMP=0; Z=0;
@SOLVE(SVM);
…………………………….
G=6;
PENALTY=1; NM=0; NMP=0; Z=0;
@SOLVE(SVM);
 …………………………….
ENDCALC
```

```
DATA:                                          (10.19)
 @OLE( )=VARK100, IC, ZERO, NP;
ENDDATA
END
```

10.6 Conclusion

Golub et al. began research to find oncogenes and subclasses of new cancers from microarray with a high will around 1970. However, the statistical discriminant function was not useful at all, so it could not be entirely completed (Problem5). Moreover, NIH reported that this kind of research is meaningless, medical researchers worldwide abandoned this research. Meanwhile, high-dimensional data with high quality is released free of charge, so many researchers of statistics, machine learning and pattern recognition continue research on high-dimensional data analysis as a new theme. However, research that non-expert finds cancer genes is medically meaningless. Perhaps they do not know the judgment of NIH. However, on October 28, 2015, we could analyze the six famous US microarrays and solved it in only 54 days. We have studied many themes of discriminant analysis so far and have solved four problems. However, this Problem5 could be solved in the shortest time. Because we have solved the valuable theme after retiring from the faculty of the university, we satisfy deeply as a researcher. Because the MP-based LDFs have solved four defects in the discriminant analysis, we can firstly succeed in the cancer gene analysis and diagnosis. That is, RIP based on MNM standard easily found that microarrays are MNM = 0 (Fact3). This fact shows that cancer and healthy subjects are LSDs in the high-dimensional microarray space. Moreover, using the common data we made the theory of LSD-discrimination, but Problem5 was just solved as applied research of this theory. Furthermore, LSD has a Matryoshka structure and includes many SMs (Fact4). Since SM is a small sample, we thought to analyze all SM easily with statistical software JMP and to provide the doctors with information useful for genetic diagnosis of cancer. However, statistical methods other than logistic regression did not find a fact that was linearly separable. Therefore, we prepare signal data using DS created by RIP and H-SVM instead of genes. Moreover, analyzing signal data instead of all the SMs, we found many facts that showed almost the same result. From this, we think that LINGO Program3 obtain good results with other high-dimensional data. Medical research failed to produce good results because the statistical discriminant function was not useful at all for LSD-discrimination. Medical researchers are not responsible for this failure. We statisticians must be responsible for this matter. The cause of the failure was not interested in the related MP theory, because most discriminant analysis researchers satisfy in a narrow world of a normal distribution. In this chapter, we explain how to find many SMs by analyzing high-dimensional microarrays data and other ordinary data with LINGO Program3. After that, if readers do the statistical analysis of the author proposed in this book, you are release easily from the curse of the high-dimensional Microarray data analysis. We are sure that NIH will recognize that microarrays are useful for cancer gene diagnosis and contribute to humanity. In the new project, in order to shorten the research time, the author wants to cooperate with your analysis. Also, we expect many researchers give students a large number of SMs as research and education issues and find facts that the author has not considered. If you download the lists of all SMs from Research Gate, those offers good educational and research data.

If LINGO Program4 find the other five BGSs, we will survey the relation of BGSs and SMs and solve the role of many SMs and BGSs. We will publish the fourth book in 2020.

References

Anderson E (1945) The irises of the Gaspe Peninsula. Bull Am Iris Soc 59:2–5

Fisher RA (1956) Statistical methods and statistical inference. Hafner Publishing Co., New Zealand

Flury B, Riedwyl H (1988) Multivariate statistics: a practical approach. Cambridge University Press, New York

Schrage L (2006) Optimization modeling with LINGO. LINDO Systems Inc. (Shinmura S translates Japanese version)

Shinmura S (2016) New theory of discriminant analysis after R. Springer, Fisher

Shinmura S (2018a) Cancer gene analysis of microarray data. In: 3rd IEEE/ACIS international conference on BCD'18, pp 1–6

Shinmura S (2018b) First success of cancer gene analysis by microarrays. In: Biocomp'18, pp 1–7

Index

© Springer Nature Singapore Pte Ltd. 2019

S. Shinmura, *High-dimensional Microarray Data Analysis*,
https://doi.org/10.1007/978-981-13-5998-9