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

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.

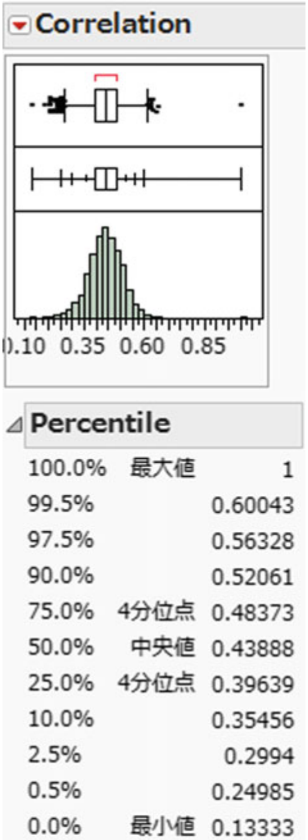
8.2 Examination of Revised LP-OLDF Discriminant Scores and SMs

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



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.

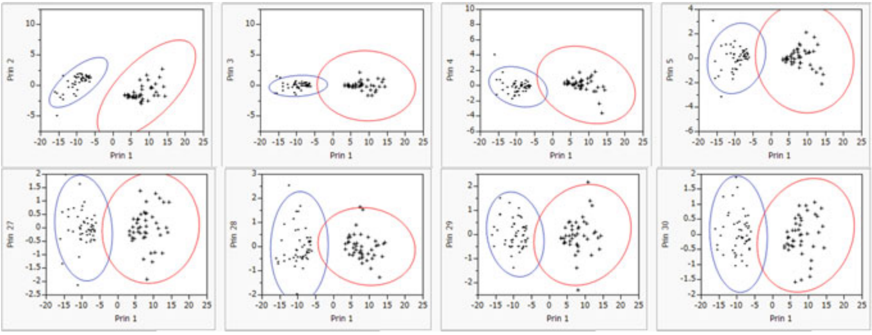


Fig. 8.2 Eight scatter plots (x-axis: Prin1; upper y-axes: From Prin2 to Prin5; lower y-axes: from Prin27 to Prin30)

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.

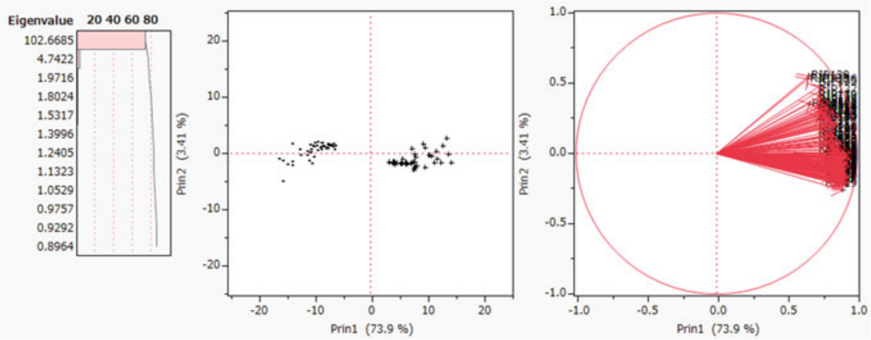


Fig. 8.3 PCA output of the 104 LpDSs

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.

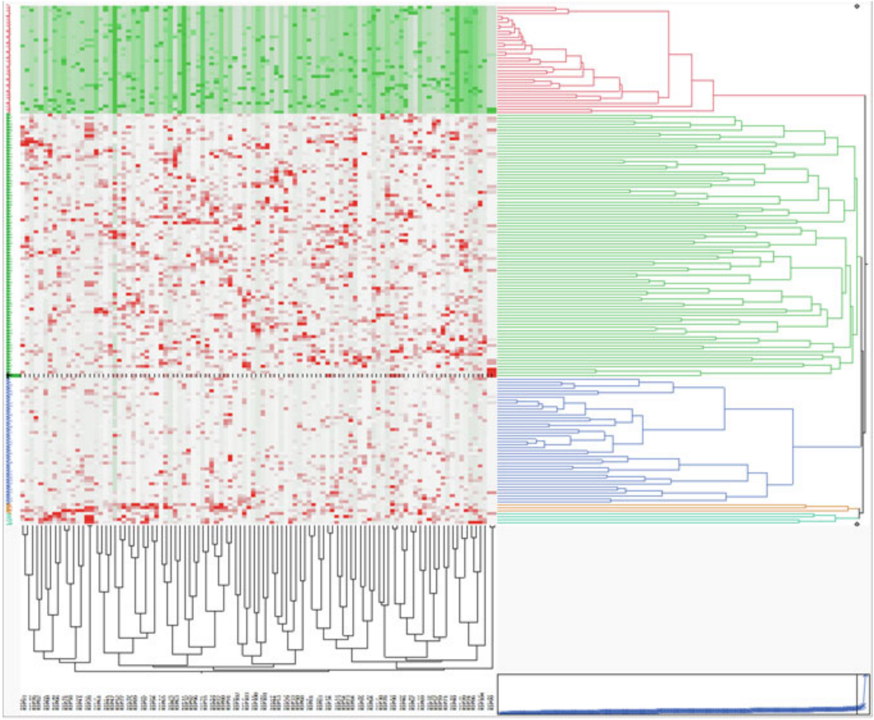


Fig. 8.4 Ward cluster analysis of RipDSs signal data

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.

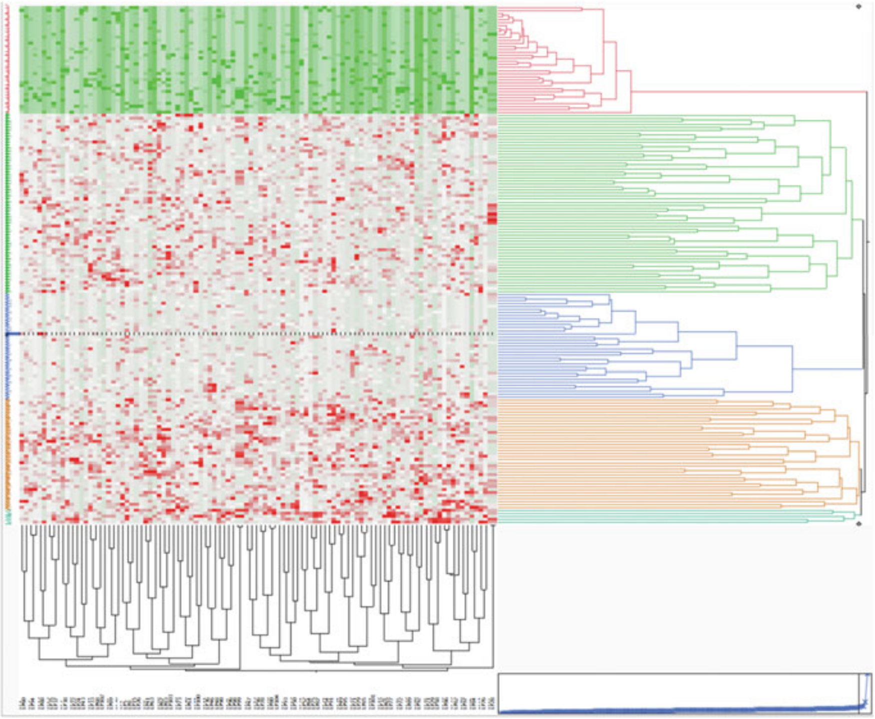


Fig. 8.5 Ward cluster analysis of LpDSs signal data

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.

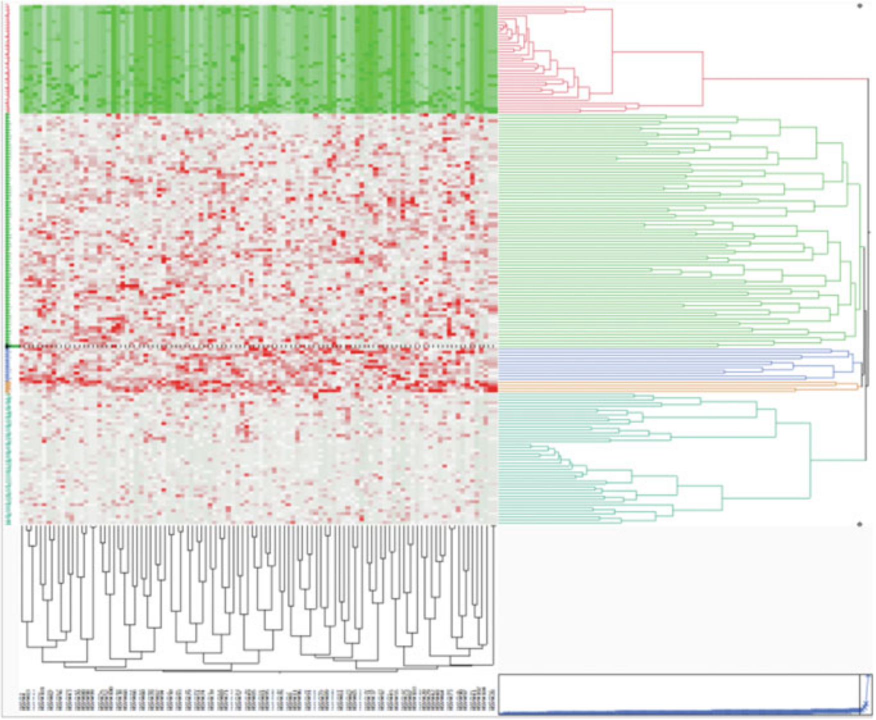


Fig. 8.6 Ward cluster analysis of HsvmDSs signal data

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.

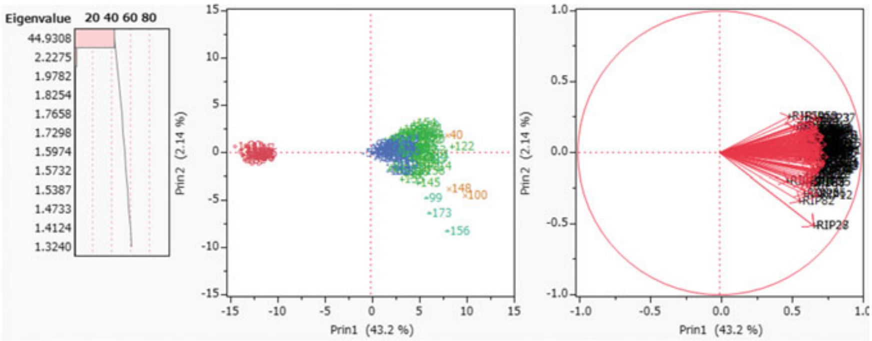


Fig. 8.7 Three plots of PCA (RipDS signal data)

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).

$$\text{RatioSV of PCA} = (11.48{-}0.83)/(14.28 + 10.02) * 100 = 1065/24.3 = 43.82716\%. \qquad (8.1)$$

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).

$$\begin{aligned}\text{RatioSV of PCA by LpDSs} &= (11.24-2.13)\,/\,(14.06+10.63) * 100\\ &= 9.11 * 100/24.69 = 36.89753\,\% \qquad (8.2)\end{aligned}$$

Because the maximum RatioSV of LpDSs is 21.81, RatioSV of PCA becomes a malignancy index.

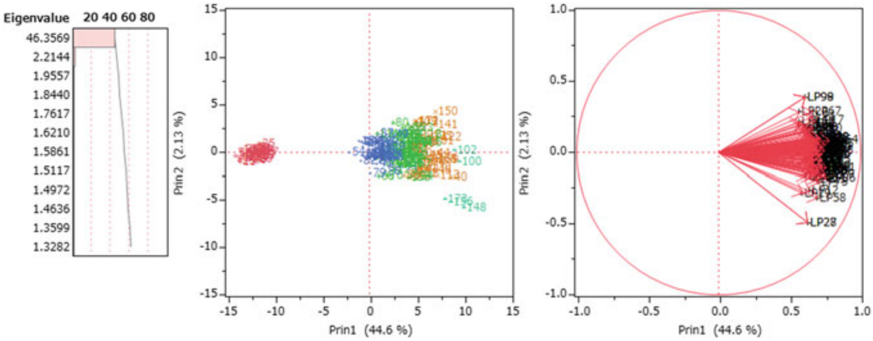


Fig. 8.8 Three plots of PCA (LpDS signal data)

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).

$$\text{RatioSV of PCA by HsvmDSs} = (13.95-1.88)/(16.93+13.75) * 100 = 39.34159\,\% \qquad (8.3)$$

Because the maximum RatioSV of HsvmDSs is 28.74, RatioSV of PCA is helpful as a malignancy index.

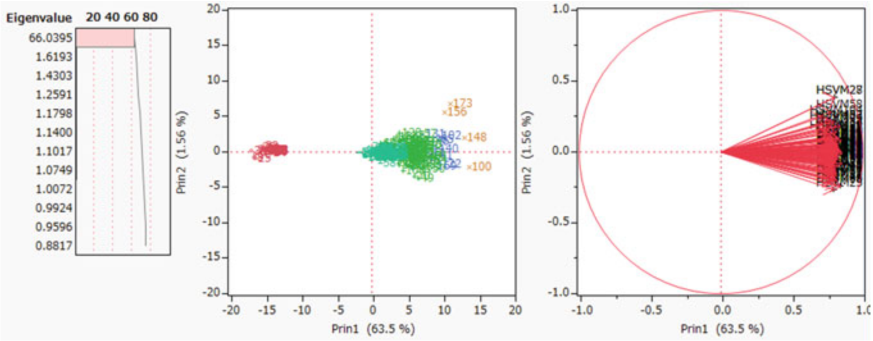


Fig. 8.9 Three plots of PCA (HsvmDS signal data)

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.

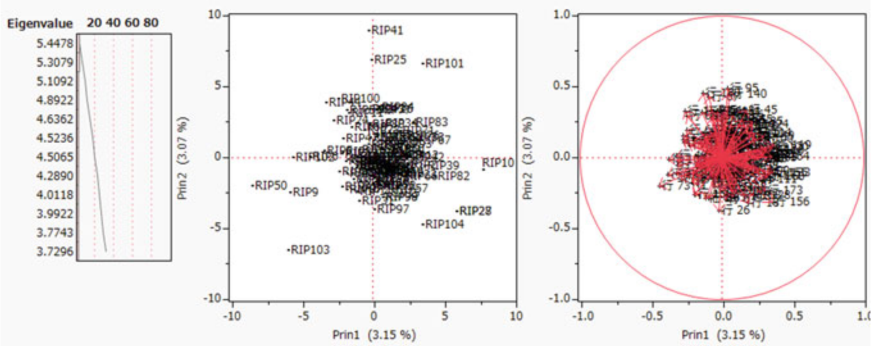


Fig. 8.10 Three plots of PCA (RipDS data)

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.

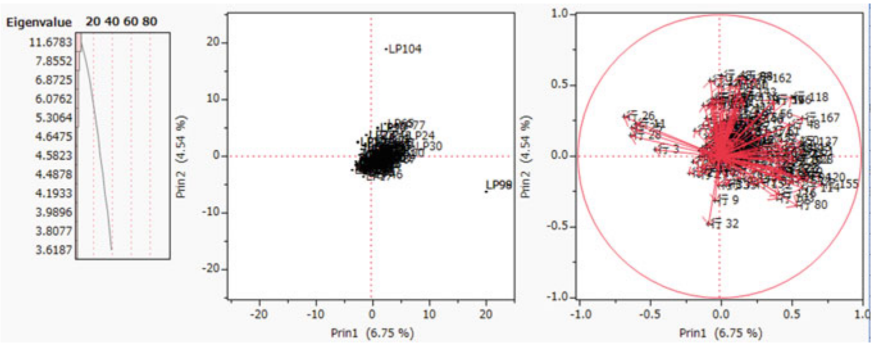


Fig. 8.11 Three plots of PCA (LpDS data)

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.

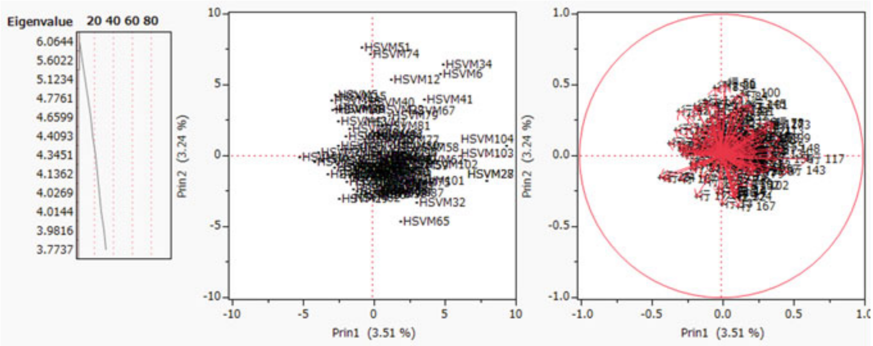


Fig. 8.12 Three plots of PCA (HsvmDS data)

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