

Competency Assessment Assignments

Juin 2, 2022

Hao ZHANG

Task 1 _a

Choose of approaches

- Because the original cancer datasets do not contain annotations, the effect of joint dimensionality reduction approaches on factors with clinical annotations or biological annotations does not need to be considered when processing such datasets. Just need to consider the effect of factors on survival.
- In general, MCIA, RGCCA, and JIVE achieved the best performances, finding factors significantly associated with survival in seven out of ten cancer types.

Task 1 _a

Choose of datasets

- Between the five multi-omics cancer datasets, I'd live to choose BIC (dataset of breast), LIHC (dataset of the liver) and SKCM (dataset of melanoma). Because MCIA performed the best for melanoma; JIVE performed the best in liver cancer ; RGCCA performed the best in breast cancer.

Task 1 _a

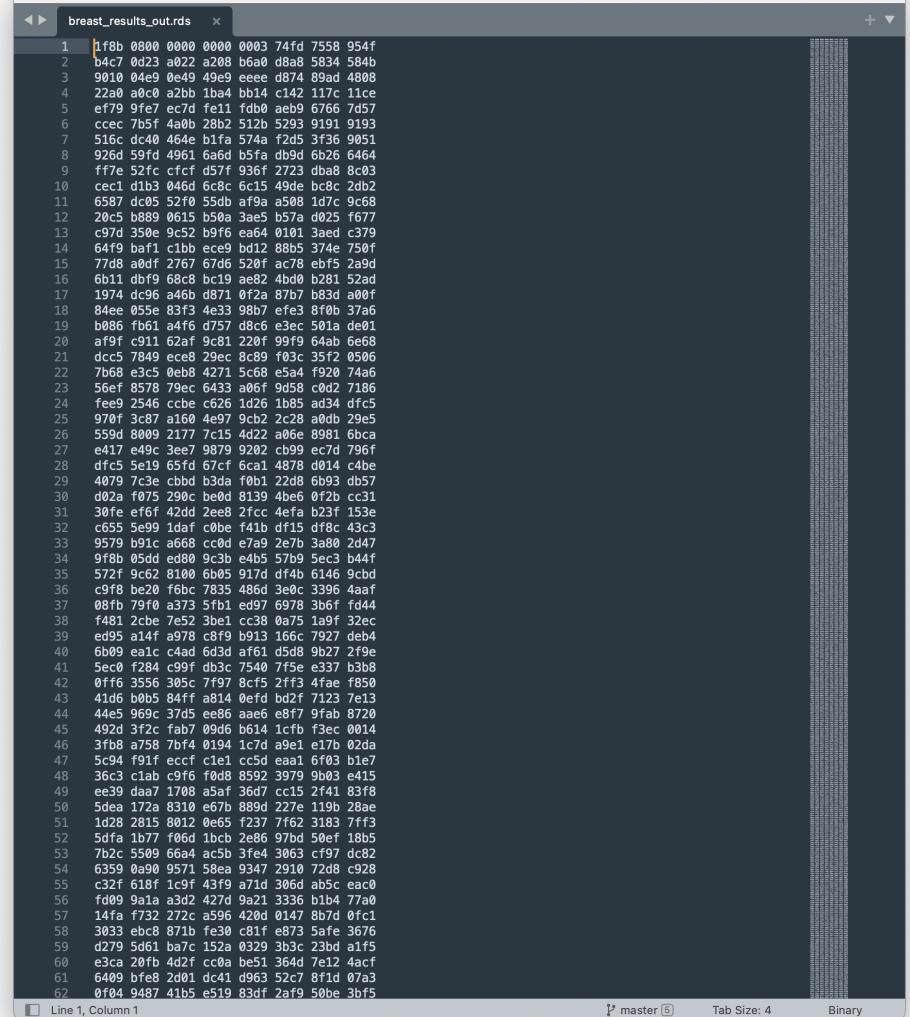
Description of datasets

- The Cancer Genome Atlas (TCGA), has profiled thousands of tumor samples for multiple molecular assays, including mRNA, microRNAs, DNA methylation, and proteomics. The cancer TCGA data were downloaded from http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. These data are composed of three omics layers: gene expression, DNA methylation, and miRNA expression. The number of samples ranges from 170 for Acute Myeloid Leukemia (AML) to 621 for Breast cancer.

Task 1_a

Results sample

The jDR factorization will decompose the P omics matrices into a product of a single factor matrix F and multiple weight/projection matrices Ai, i=1...P.



```
breast_results_out.rds
1 1f8b 0800 0000 0000 0003 74fd 7558 954f
2 b4c7 0d23 a022 a28b b6a0 d8a8 5834 584b
3 9010 04e9 0e49 49e9 eeee d874 89ad 4808
4 22a0 a0c0 a2bb 1ba1 bb14 c142 117c 11ce
5 ef79 9fe7 ec7d fe11 fdb0 ae9b 6766 7d57
6 ccec 7b5f 4a0b 28b2 512b 5293 9191 9193
7 516c dc40 464e b1fa 574a f2d5 3f36 9051
8 926d 59f0 4961 6a6d b5fa db9d 6b26 6464
9 fff7e 52fc cf1c d57f 936f 2723 dbab 8c03
10 cec1 db13 046d 6c8b 6c15 49de bc8c 2db2
11 6587 dc05 52f0 55d0 af9a a508 1d7c 9c68
12 20c5 b889 0615 b58a 3a5e b57a d025 f677
13 c97d 350e 9c52 b9f6 ea64 0101 3aed c379
14 64f9 baf1 c1bb ece9 bd12 88b5 374e 750f
15 77d8 a0df 2767 6746 520f ac78 ebf5 2a9d
16 6b11 dbf9 68c8 bc19 ae82 4b80 b281 52ad
17 1974 dc96 a46b d871 0f2a 87f7 b63d 00f1
18 84ee 055e 83f3 4e33 98b7 ef63 870b 37a6
19 b086 fb61 a4f6 d757 6dc6 e3ec 501a de01
20 af9f c911 62af 9c81 220f 99f9 64ab 6e68
21 ddc5 7849 ece8 29e9 8c89 f03c 35f2 0506
22 7b68 e3c5 0eb8 4271 5c68 e5a4 f920 74a6
23 56ef 8578 79ec 6433 a0f6f 9d58 c9d2 7186
24 fee9 2546 cbce c626 1d26 1b85 ad34 dfc5
25 970f 3c87 a169 4e97 9cb2 2c28 a0db 29e5
26 559d 8009 2177 7c15 4d22 a06e 8981 6bc4
27 e417 e49c 3e67 9879 9202 cb99 ec7d 796f
28 dfc5 5e19 65fd 67cf 6ca1 4878 d014 c4b6
29 4079 7c3e cbcd b3d0 f0b1 22d8 6b93 db57
30 d02a f075 29c0 be09 8139 4be6 0f2b cc31
31 30fe ef6f 42dd 2ee8 2fcc 4ef6 b23f 153e
32 c655 5e99 1daf c0b8 f41b df15 df8c 43c3
33 9579 b91c a668 cc9d e7a9 2e7b 3a80 2d47
34 9f8b 05dd ed80 9c3b e4b5 57b9 5ec3 b44f
35 572f 9c62 8100 6b85 917d df4b 6146 9bcd
36 c9f8 be20 f6bc 7835 486d 3e0c 3396 4aa5
37 08fb 79f0 a373 5fb1 ed97 6978 3b6f d44
38 f481 2cbe 7e52 3be1 cc38 0a75 1a9f 32ec
39 ed95 a14f a978 c8f9 b913 166c 7927 deb4
40 6b09 ea1c c4ad 6d3d af61 d5d8 9b27 2f9e
41 5ec0 f284 c9f9 db3c 7540 7f5e e337 h3b8
42 0ff6 3556 305c 7f97 8cf5 2ff3 4fae f850
43 41d6 b0b5 84ff a814 0eff bd2f 7123 7e13
44 44e5 969c 37d5 ee86 aae6 e8f7 9fab 8720
45 492d 3f2c fab7 09d0 b614 1cfb f3ec 0014
46 3fb8 a758 7bf4 0194 1c7d a9e1 e17b 02da
47 5c94 f91f ecfc c1e1 cc5d ea11 6f03 b1e7
48 36c3 c1ab c9f9 f0d8 8592 3799 9b03 e415
49 ee39 daa7 1708 a5af 36d7 cc15 2f41 83f8
50 5dea 172a 8310 e67b 889d 227e 119b 28ae
51 1d28 2815 8012 0e65 f237 7ff2 3183 7ff3
52 5dfa 1b77 f06d 1bcb 2886 97bd 50ef 18b5
53 7b2c 5509 664a ac5d 3f4e 3063 cf97 dc82
54 6359 0a99 9571 58e9 9347 2910 72db c928
55 c32f 618f 1c9f 43f9 a71d 306d ab5c eac0
56 fd09 9a1a a3d2 427d 9a21 3336 b1b4 77a0
57 14fa f732 272c a596 420d 0147 8b7d 0fc1
58 3033 ebc8 871b fe30 c81f e873 5afe 3676
59 d279 5d61 ba7c 152a 0329 3b3c 23bd a1f5
60 e3ca 20fb 4d2f cc0a be51 364d 7e12 4acf
61 6469 bfef 8201 dc41 d963 52c7 8f1d 07a3
62 0f04 9487 41b5 e519 83df 2af9 50be 3bf5
```

Task 1 _b

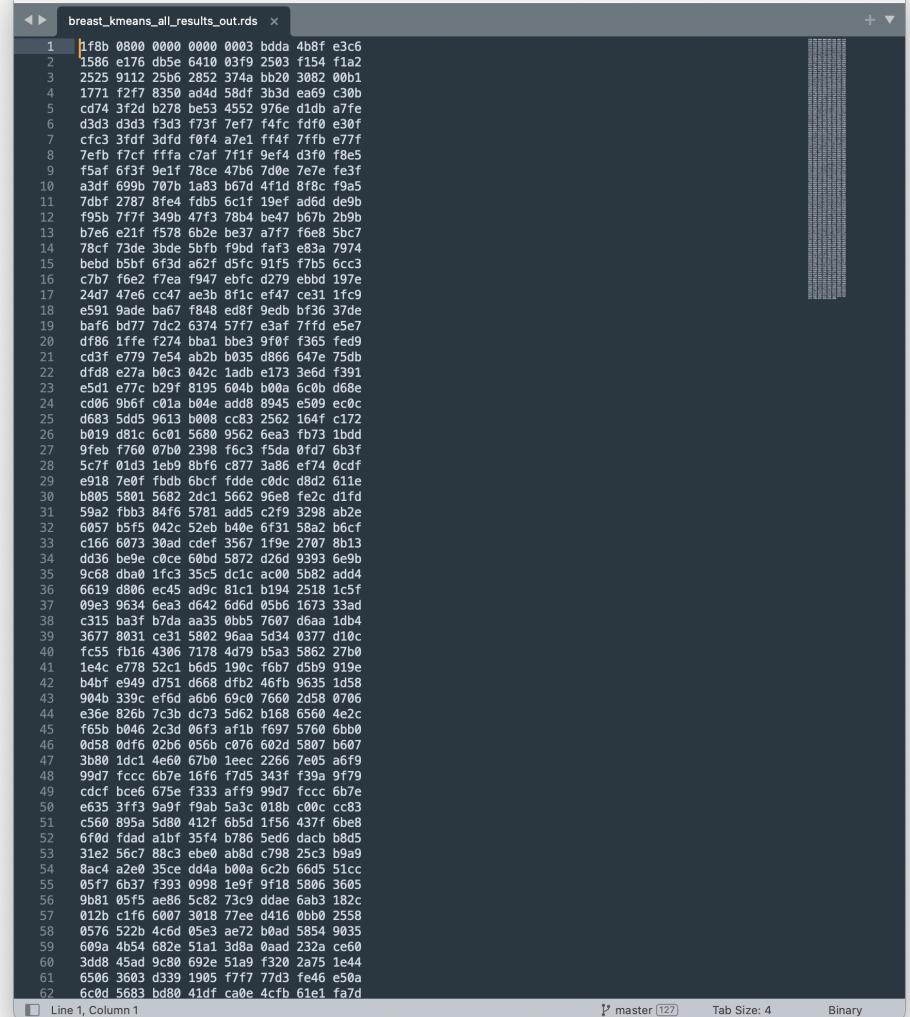
Choose of approaches

- No algorithm consistently outperformed all others in either differential survival or enriched clinical parameters. With respect to survival, MCCA had the best prognostic value, while MultiNMF was second and LRACluster third.
- However, limited by the installation of the package and the program provided, I can only choose K-means, Spectral and SNF for clustering.

Task 1_b (1)

Results sample

Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters.

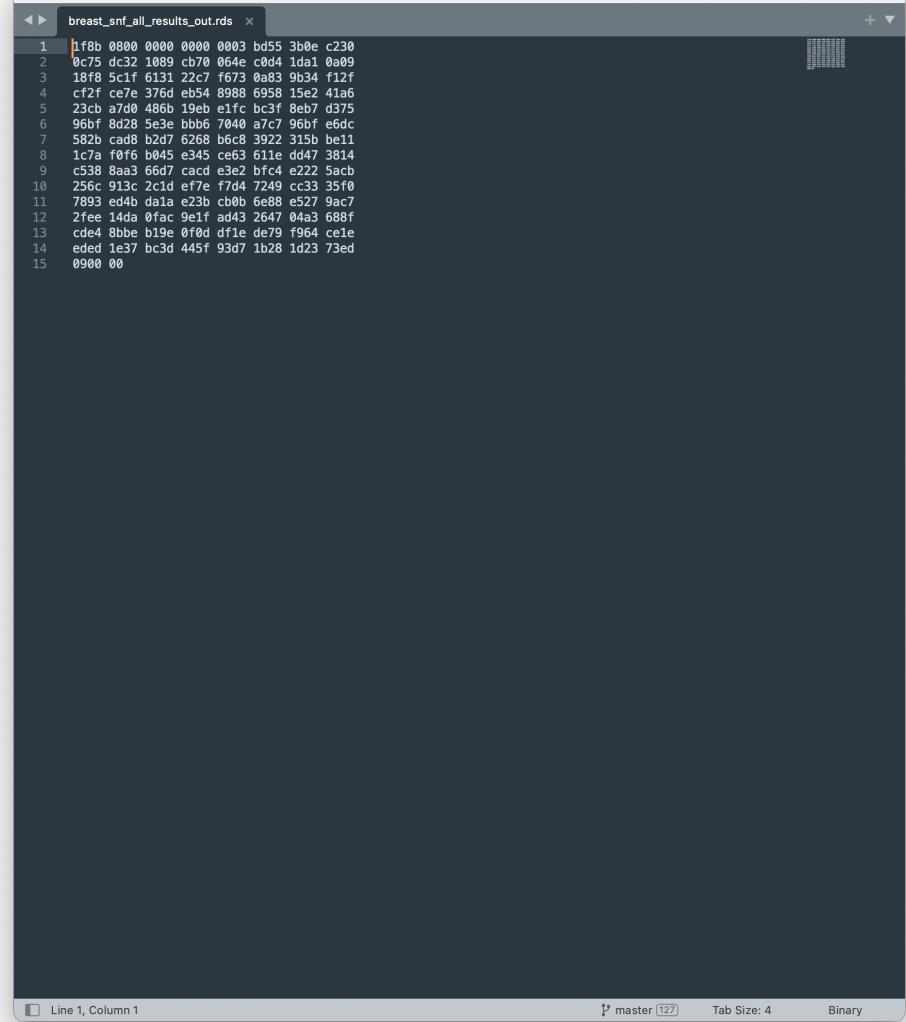


```
breast_kmeans_all_results_out.rds
1 lf8b 0800 0000 0000 0003 bdda 4b8f e3c6
2 1586 e17e db5e 641e 03f9 25d3 f154 f1a2
3 2525 9112 25b6 2852 374a bb20 3082 00b1
4 1771 f2f7 8358 ad44 58df 3b3d ea69 c30b
5 cd74 3f2a b278 be53 4552 976e d1db a7fe
6 d3d3 d3d3 f3d3 f73f 7ef7 f4fc fdff e30f
7 cfc3 3fdf 3dfd f0f4 a7e1 ffaf 7ffb c77f
8 7efb f7cf fffa c7af 711f 9ef4 d2f0 f8e5
9 f5af 6f3f 9e1f 78c4 47b6 7d0e 7e7e fe3f
10 a3df 699b 707b 1a83 b67d 4f1d 878c 9a5
11 7dbf 2787 8fe4 fdb5 6c1f 19ef ad6d de9b
12 f95b 77f7 349b 47f3 78b4 be47 b67b 2b9b
13 b7e6 e21f f578 6b2b be37 a7ff f6e8 5bc7
14 78cf 73de 3bde 5bf5 f9bd faf3 e83a 7974
15 bebd b5bf 6f3d a62f d5fc 91f5 f7b5 6cc3
16 c7b7 f6e2 f7ea f947 ebfc d279 ebcd 197e
17 24d7 47e6 cc47 ae3b 8f1c ef47 ce31 1fc9
18 e591 9ade ba67 f848 ed8f 9ed8 bf36 37de
19 ba6f bd77 7dc2 6374 57f7 e3af 7ff7 e5e7
20 df86 1ffe f274 bba1 bbe3 9f0f f365 fed9
21 cd3f e779 7e54 ab2b b035 0866 647e 75db
22 dfd8 e27a b6c3 042c 1adb e173 3e6d f391
23 e5d1 e77c b29f 8195 604b b00a 6c0b d68e
24 cd06 9b6f c01a b04e add8 8945 e509 ec0c
25 d683 5dd5 9613 b008 cc83 2562 164f c172
26 b019 d81c 6c01 5680 9562 6ea3 fb73 1bdd
27 9feb f761 07b0 2398 f6c3 f5da 0fd7 6b3f
28 5c7f 01d3 1eb9 8bf6 c877 3a86 e174 0cdf
29 e918 7e0f fdbb 6bcf fdde c0dc dd82 611e
30 b805 5801 5682 2dc1 5662 96e8 fe2c d1fd
31 59a2 fbb3 84f6 5781 add5 c2f9 3298 ab2e
32 6057 b5f5 042c 52eb b40e 6f31 58a2 b6cf
33 c166 6073 30ad cdef 3567 1f9e 2707 8b13
34 dd36 be9e c0ce 60hd 5872 d26d 9393 6e9b
35 9c68 dba0 1fc3 35c5 dc1c ac00 5882 add4
36 6619 d806 ec45 ad91 81c1 b194 2518 1c5f
37 e0e3 9634 6ea3 d642 6d6d 05b6 1673 33ad
38 c315 ba3f b7da aa35 0bb5 7607 dgaa 1db4
39 3677 8031 ce31 5802 96aa 5d34 0377 d10c
40 fc55 fb16 4306 7178 4d79 b5a3 5862 27b0
41 1e4c e778 52c1 b6d5 190c f6b7 d5b9 919e
42 b4bf e949 d751 d668 dfb2 46fb 9635 1d58
43 904b 339c ef6d a6b1 69c0 7660 2d58 0706
44 e36e 8261 7c3b dc73 5d62 b168 6560 4ec2
45 f65b bb46 2c3d 0613 a1b1 f697 5760 6b00
46 0d58 0dff 02b6 0561 c076 602d 5807 b607
47 3bb0 1dc1 4e60 67b4 1eec 2266 7e05 a6f9
48 99d7 fccc 6b7e 16f1 f7d5 343f f39a 9f79
49 cdcf bcef 675e f333 aff9 99d7 fccc 6b7e
50 e635 3ff3 9a9f 9fa8 5a3c 018b c00c cc83
51 c560 895a 5d80 412f 6b5d 1f56 437f 6be8
52 6f0d fdad a1bf 35f4 b786 5ed6 dacb b8d5
53 31e2 56c7 88c3 eb0e ab8d c798 25c3 b9a9
54 8ac4 a2e0 35ce dd44 b00a 6c2b 66d5 51cc
55 05f7 6b37 f393 0994 1e9f 9f18 5006 3605
56 9b81 05f5 ae86 5c82 73c9 idae 6ab3 182c
57 012b c1f6 6087 3018 77ee d416 0bb0 2558
58 0576 522b 4c6d 05e3 ae72 b0ad 5854 9035
59 609a 4b54 682e 51a1 3d8a 0aad 232a ce60
60 3dd8 45ad 9c80 692e 51a9 f320 2a75 1e44
61 6506 3603 d339 1985 f7f7 77d3 fe46 e50a
62 6c0d 5683 bd80 41df ca0e 4cfc 6iel fa7d
```

Task 1_b (2)

Results sample

Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters.



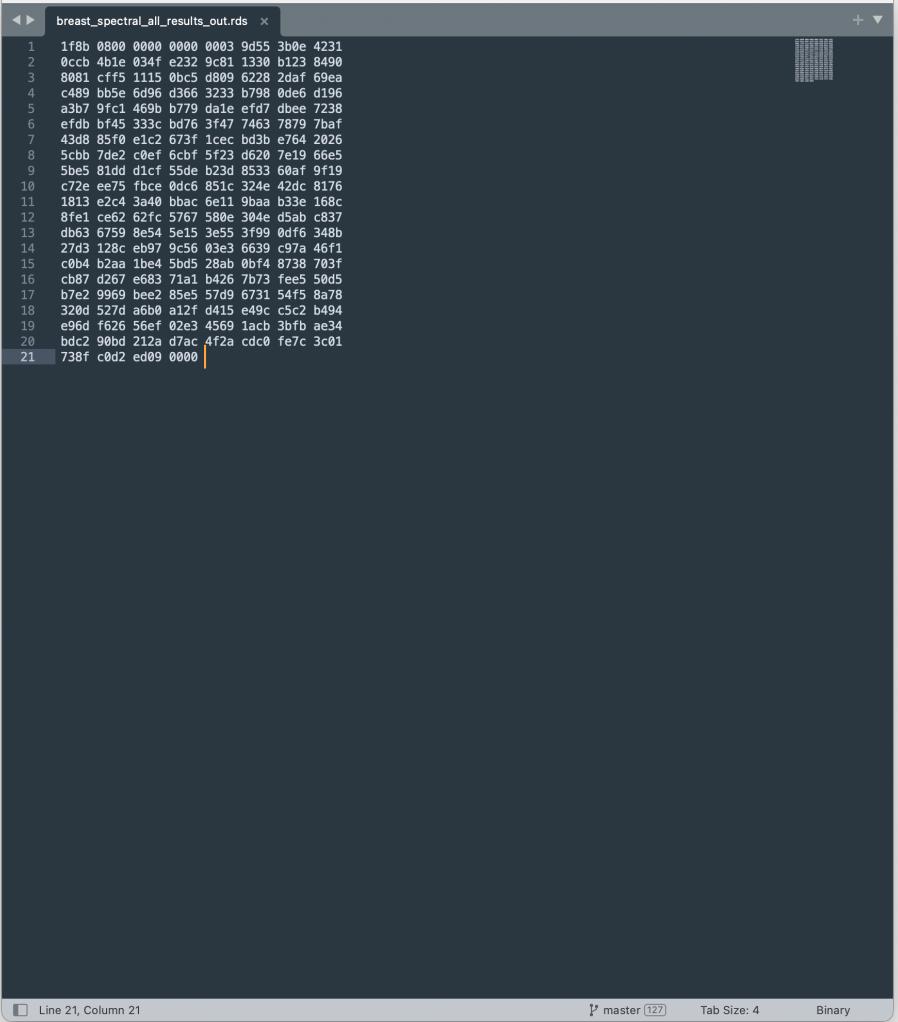
```
breast_snf_all_results_out.rds
1 1f8b 0800 0000 0000 0003 bd55 3b0e c230
2 0c75 dc32 1089 cb70 064c cd41 1da1 0009
3 18ff 5c1f 6131 22c7 f673 0a83 9b34 f12f
4 cf2f ce7e 3761 eb54 8988 6958 15e2 41a6
5 23cb a7d0 486b 19eb e1fc bc3f 8eb7 d375
6 96bf 8d28 5e3b bbb6 7040 a7c7 96bf e5dc
7 582b cad8 b2d7 6268 b6c8 3922 315b be11
8 1c7a f0f6 b045 e345 ce63 611e dd47 3814
9 c538 8aa3 66d7 cacc e3e2 bfcc e222 5acb
10 256c 913c 2c1d ef7e f7d4 7249 cc33 35f0
11 7893 ed4b da1a e23b cb0b 6e88 e527 9ac7
12 2fee 14da 0fac 9e1f ad43 2647 04a3 688f
13 cde4 8bbe b19e 0f0d df1e de79 f964 ce1e
14 eded 1e37 bc3d 445f 93d7 1b28 1d23 73ed
15 0900 00
```

Line 1, Column 1 master [127] Tab Size: 4 Binary

Task 1_b (3)

Results sample

Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters.



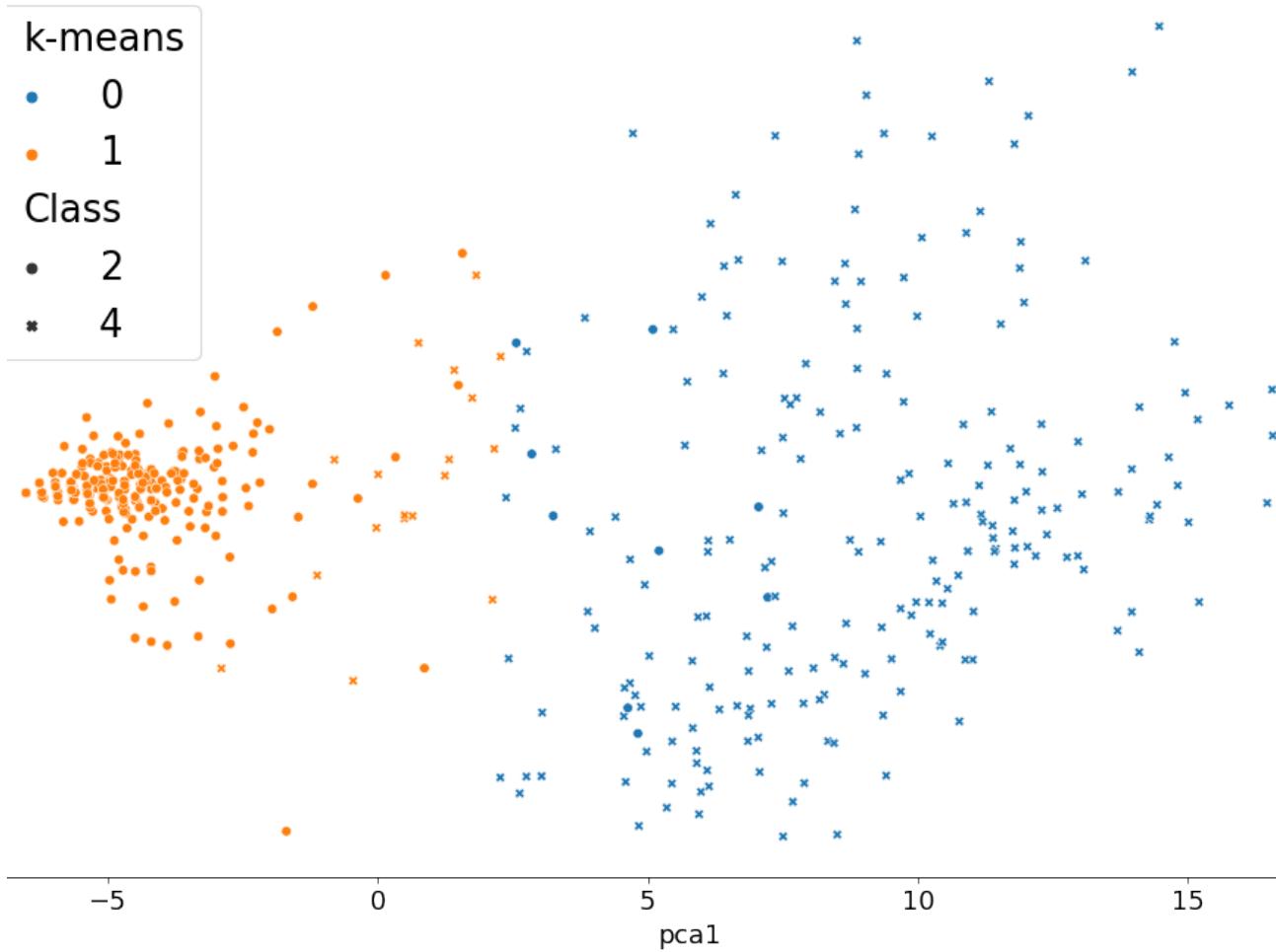
```
breast_spectral_all_results_out.rds
1 1f8b 0800 0000 0000 0003 9d55 3b0e 4231
2 0cc0 4b1e 034f e232 9c8b 1330 b123 8490
3 8081 cff5 1115 0bc5 d808 6228 2daf 69ea
4 c489 bb5e 6d99 d366 3233 b798 0de6 d196
5 a3b7 9fc1 469b b779 da1a ef7d7 dbee 7238
6 efd1 bf45 333c bd76 3f47 7463 7879 7baf
7 43d8 85f0 e1c2 673f 1cec bd3b c764 2026
8 5cbb 7de2 c0ef 6cbf 5f23 d620 7e19 66e5
9 5be5 81dd d1cf 55de b23d 8533 60af 9f19
10 c72e ee75 fbc0 0dc6 851c 324e 42dc 8176
11 1813 e2c4 3a40 bbac 6e11 9baa b33e 168c
12 8fe1 ce62 62fc 5767 5808 304e d5ab c837
13 db63 6759 8e54 5e15 3e55 3f99 0df6 348b
14 27d3 128c eb97 9c56 03e3 6639 c97a 46f1
15 c0b4 b2aa 1be4 5bds 28ab 0bf4 8738 703f
16 cb87 d267 e683 71a1 b426 7b73 fee5 50d5
17 b7e2 9969 bee2 85e5 57d9 6731 54f5 8778
18 320d 527d a6b0 a12f d415 e49c c5c2 b494
19 e96d f626 56ef 02e3 4569 1acb 3bfb ae34
20 bdc2 90bd 212a d7ac 4f2a cdc0 fe7c 3c01
21 738f c0d2 ed09 0000 |
```

Line 21, Column 21 master [127] Tab Size: 4 Binary

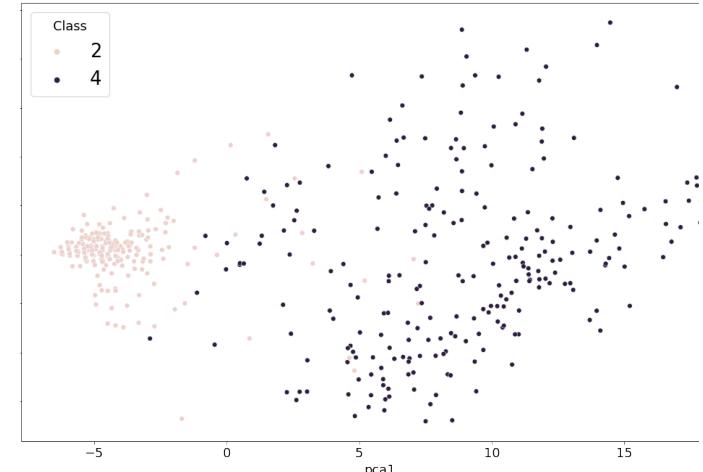
Task2

**Initial classification and clustering
effects**

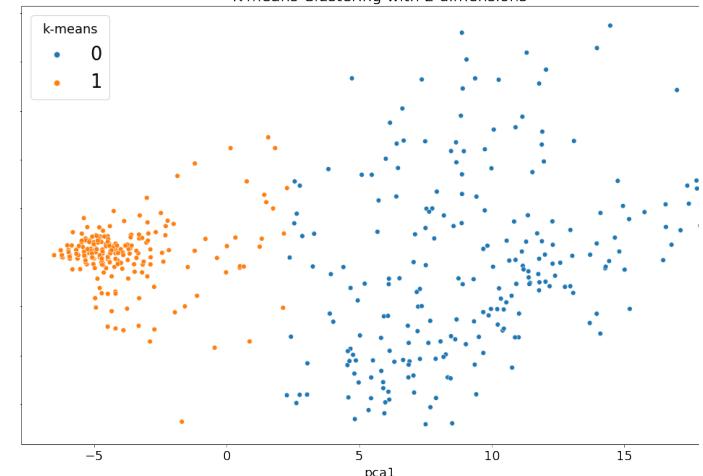
K-means Clustering and class labels with 2 dimensions



Class labels with 2 dimensions



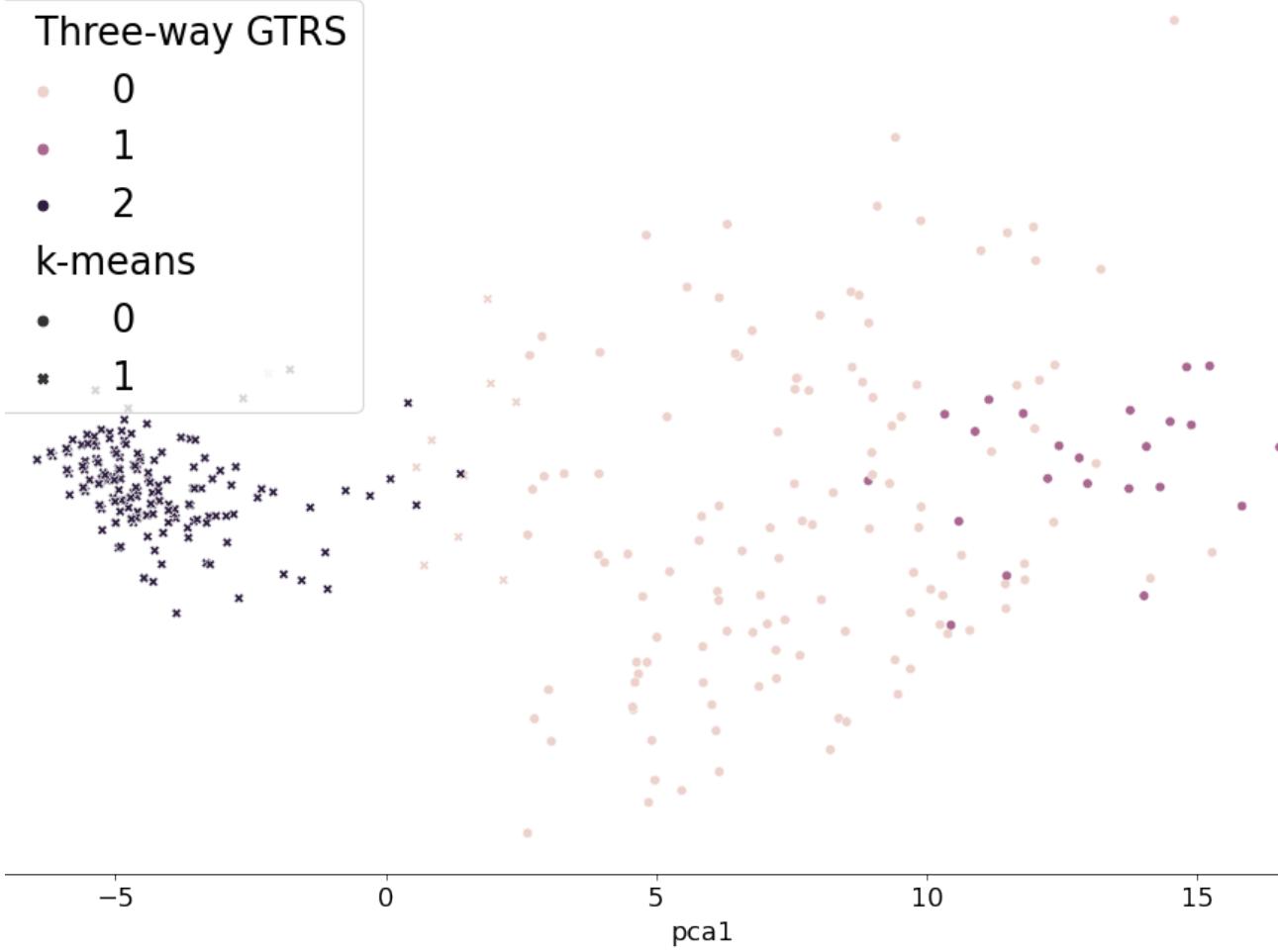
K-means Clustering with 2 dimensions



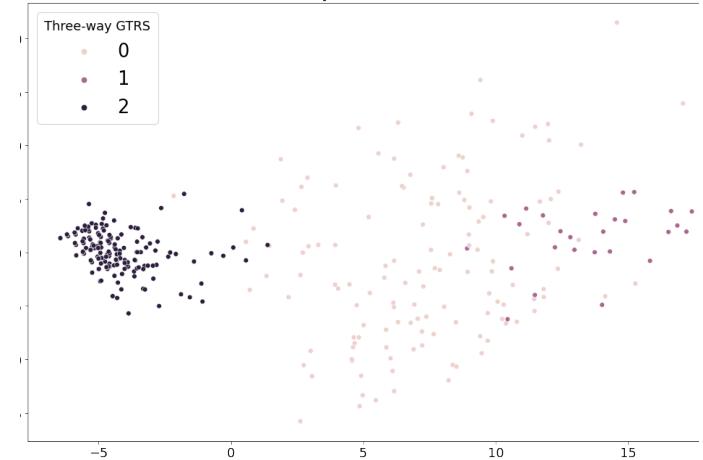
Task2

**Classification and clustering effects after
implementing the GTRS three-way clustering
algorithms**

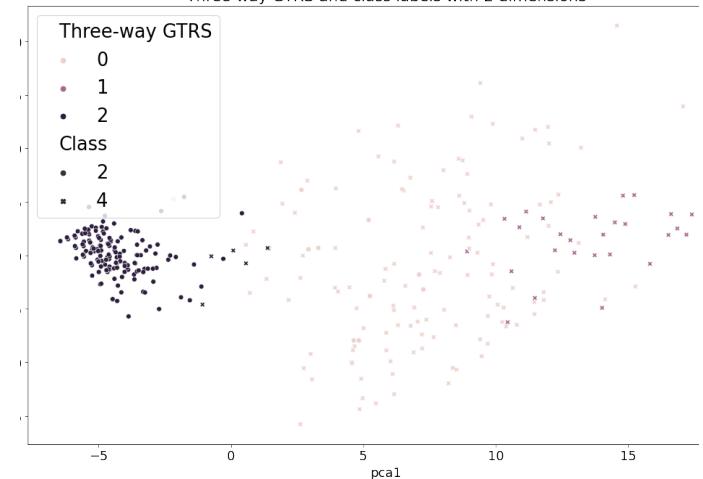
Three-way GTRS and K-means Clustering with 2 dimensions



Three-way GTRS with 2 dimensions



Three-way GTRS and class labels with 2 dimensions



Task 2

Discussion of the limitations of the algorithm

- The GTRS three-way clustering algorithm can effectively distinguish the data distributed in the two extreme parts (inside and outside), but once the algorithm is overfitting or underfitting in the process of use, it may cause the operation time to be too long or face the large amount of data that deals with partial parts cannot be differentiated.
- To enable GTRS three-way clustering algorithm to effectively cluster data and avoid overfitting or underfitting, it is necessary to fine-tune the parameters of the algorithm (such as α -, α --, β +, β ++).