Feature Selection (FS) workflow report

May 18, 2017

Introduction

The report summarizing the Feature Selection pipeline results.

Feature Selection workflow

Univariate canonical correlation (X2) with Recursive Feature Elimination (RFE) wrapped with Random Forest (RF).

Dataset

Expression data from normal and prostate tumor tissues (GSE6919_GPL92).

Summary stats from training phase

Table 1: Best model metrics from 10-folds cross-validation resampling.

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4888	0.2549	0.1035	0.1245
2	0.5735	0.3705	0.08549	0.1258
3	0.6191	0.4298	0.1252	0.1952
4	0.6182	0.4323	0.07686	0.1278
5	0.5977	0.3931	0.1226	0.2001
6	0.6576	0.4842	0.1562	0.2457
7	0.6583	0.4858	0.1434	0.2326
8	0.6218	0.4295	0.1498	0.2417
9	0.6309	0.443	0.1464	0.2353
10	0.5953	0.392	0.18	0.2697
15	0.6582	0.4883	0.1529	0.2343
20	0.6417	0.4585	0.1413	0.2166
25	0.6417	0.4614	0.1496	0.2271
30	0.6052	0.4059	0.1705	0.2588
35	0.615	0.4199	0.1458	0.2235
40	0.6159	0.4207	0.1411	0.2096
45	0.6233	0.4309	0.1648	0.252
50	0.6498	0.4705	0.1777	0.2715
60	0.6773	0.517	0.1438	0.2165
70	0.6773	0.5128	0.1232	0.1877
80	0.6591	0.4859	0.1792	0.2674
90	0.6689	0.5005	0.1539	0.2325
100	0.6667	0.4978	0.1877	0.2829
416	0.6424	0.4624	0.1714	0.2525

Summary stats from testing phase

Table 2: Classification metrics from twenty class-balanced and randomized runs.

run	Variables	Accuracy	Kappa	AccuracyPValue
1	**60**	**0.8**	**0.7036**	**1.599e-09**
2	416	0.6182	0.4128	0.0008835
3	100	0.6727	0.5206	3.979e-05
4	40	0.7455	0.6173	2.106e-07
5	416	0.6	0.4039	0.002147
6	100	0.6	0.3823	0.002147
7	60	0.7636	0.6505	4.565e-08
8	50	0.6364	0.4486	0.0003385
9	50	0.6364	0.4519	0.0003385
10	80	0.6545	0.4687	0.0001205
11	416	0.7818	0.6715	8.988e-09
12	100	0.6545	0.487	0.0001205
13	416	0.6364	0.4458	0.0003385
14	416	0.7091	0.5626	3.424e-06
15	100	0.6545	0.4759	0.0001205
16	100	0.7091	0.5663	3.424e-06
17	25	0.6364	0.4444	0.0003385
18	20	0.7455	0.6123	2.106e-07
19	416	0.6545	0.4785	0.0001205
20	416	0.7636	0.646	4.565e-08

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6836	0.06309	0.8

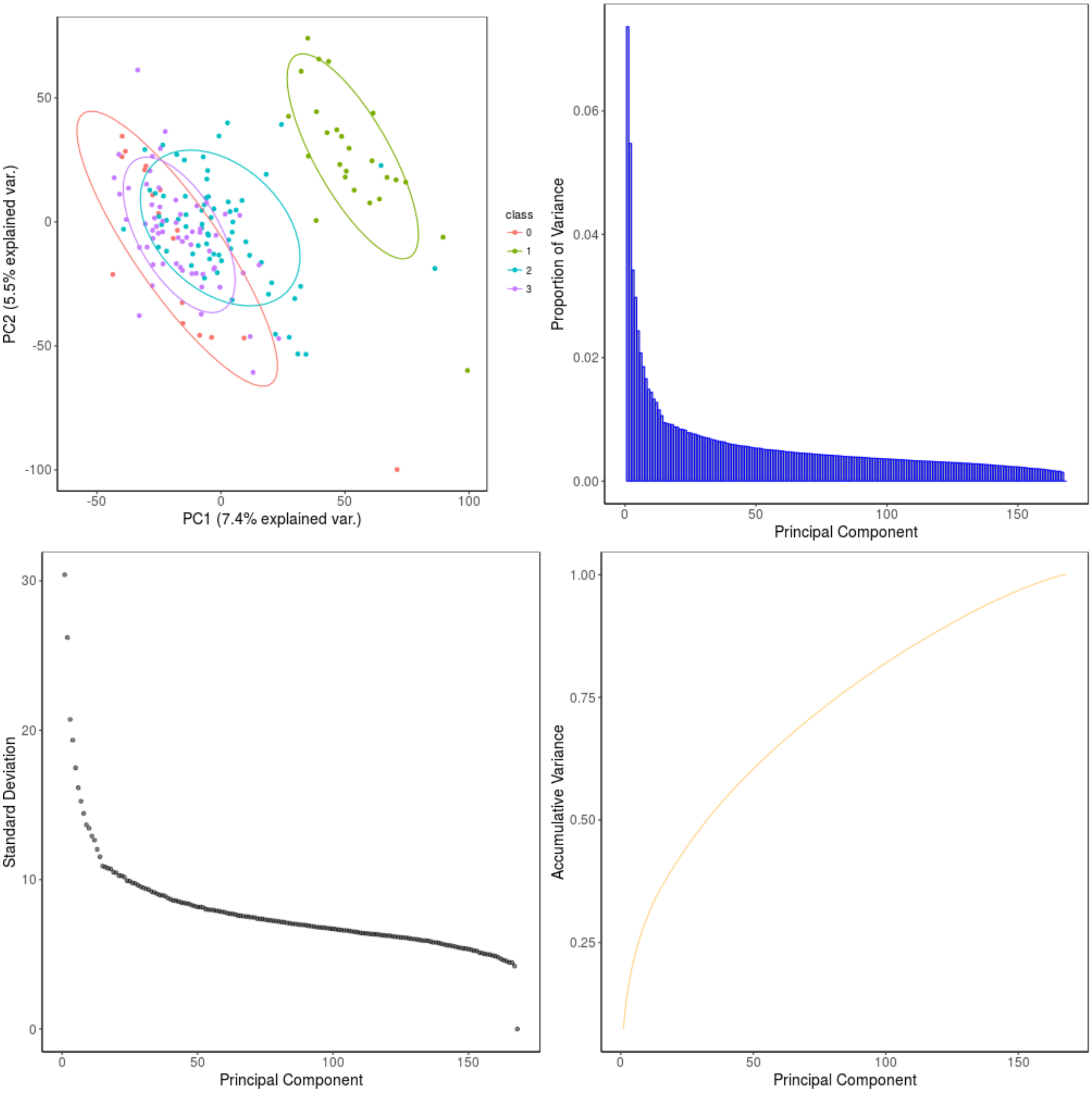
Workflow runtime

9.127 minutes

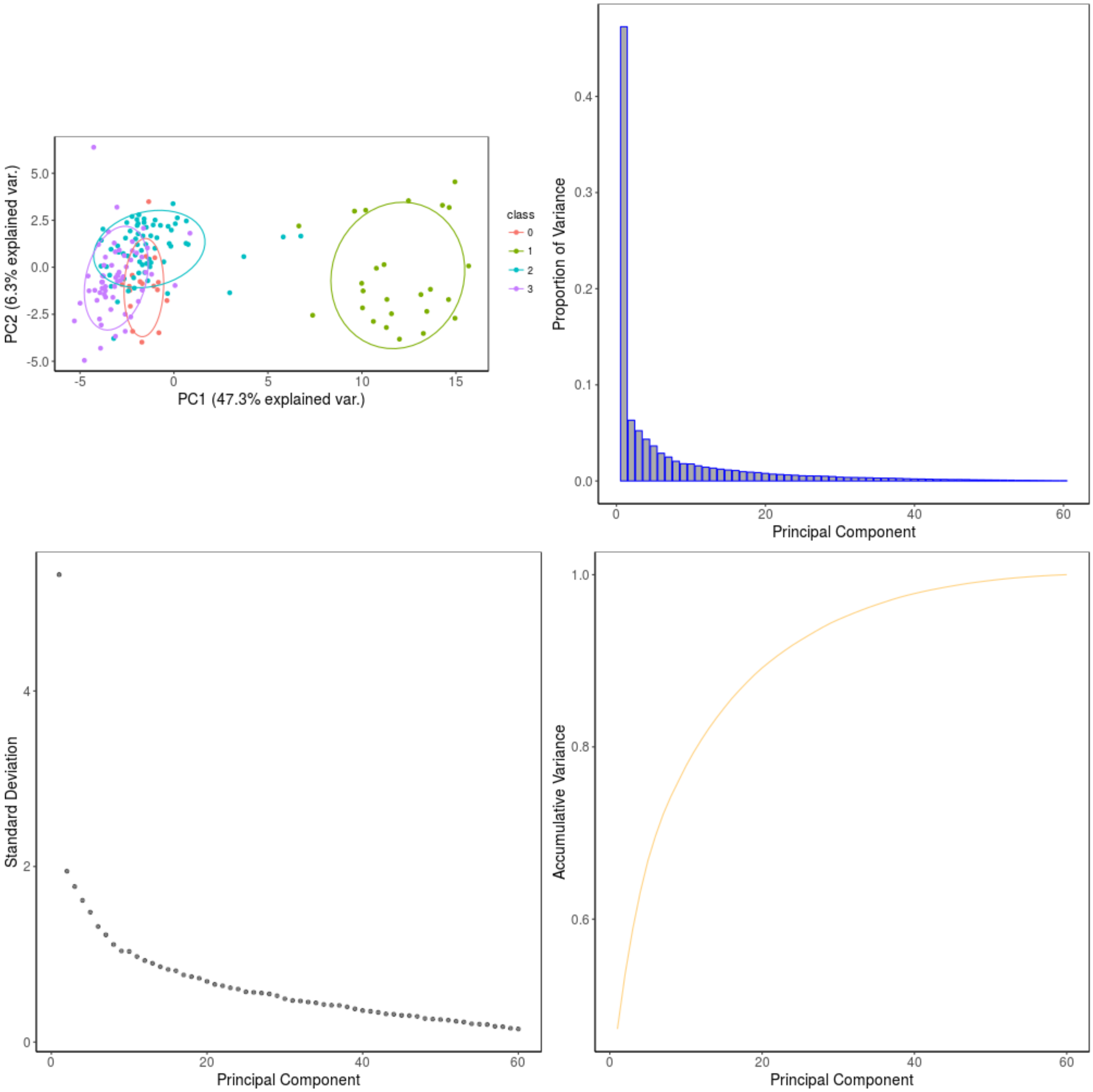
Plots

Visualization of the classification using PCA

- Groups distribution on the first two Principal Components (PC1 and PC2) from the original data (without apply any FS method).



- Groups distribution on the first two Principal Components (PC1 and PC2) after to apply the FS workflow.


PC2 (6.3% explained var.)
PC1 (47.3% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component