Feature Selection (FS) workflow report

May 18, 2017

Introduction

The report summarizing the Feature Selection pipeline results.

Feature Selection workflow

Recursive Feature Elimination (RFE) wrapped with Random Forest (RF).

Dataset

Expression data from normal and prostate tumor tissues (GSE6919_GPL93).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4626	0.2304	0.1533	0.2444
2	0.5083	0.2827	0.1206	0.1814
3	0.5165	0.2892	0.1111	0.1796
4	0.5074	0.2901	0.09441	0.1371
5	0.5082	0.2877	0.09491	0.1407
6	0.5573	0.354	0.129	0.1905
7	0.5739	0.3795	0.1243	0.1711
8	0.6088	0.4287	0.1141	0.1661
9	0.6264	0.4536	0.1269	0.1836
10	0.6056	0.4224	0.1003	0.1466
15	0.6255	0.4476	0.1254	0.1806
20	0.6089	0.4215	0.1513	0.2148
25	0.5838	0.3866	0.1929	0.2735
30	0.5814	0.3811	0.1248	0.1768
35	0.5929	0.3962	0.149	0.205
40	0.6012	0.4082	0.1567	0.2183
45	0.5738	0.3615	0.1368	0.1948
50	0.5995	0.3986	0.1632	0.2385
60	0.6071	0.4057	0.1499	0.221
70	0.6064	0.4073	0.1405	0.2072
80	0.6321	0.4424	0.1518	0.2266
90	0.6255	0.4343	0.1578	0.231
100	0.6147	0.4173	0.1468	0.217
12579	0.5973	0.387	0.1346	0.1996

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	70	0.7547	0.6376	1.238e-07
2	100	0.6604	0.5016	9.108e-05
3	70	0.6792	0.524	2.884e-05
4	45	0.6038	0.4064	0.001802
5	45	0.6981	0.5508	8.413e-06
6	25	0.717	0.5717	2.255e-06
7	40	0.6792	0.5177	2.884e-05
8	**80**	**0.7925**	**0.6899**	**4.606e-09**
9	100	0.717	0.5735	2.255e-06
10	90	0.6981	0.5527	8.413e-06
11	80	0.6792	0.5195	2.884e-05
12	90	0.6981	0.548	8.413e-06
13	60	0.6604	0.4942	9.108e-05
14	60	0.5849	0.3827	0.004204
15	50	0.6038	0.4095	0.001802
16	80	0.6604	0.5044	9.108e-05
17	60	0.717	0.5717	2.255e-06
18	50	0.6604	0.4917	9.108e-05
19	12579	0.6038	0.3984	0.001802
20	60	0.6415	0.4635	0.0002658

Accuracy__Mean	Accuracy__SD	Accuracy__Max
0.6755	0.05252	0.7925

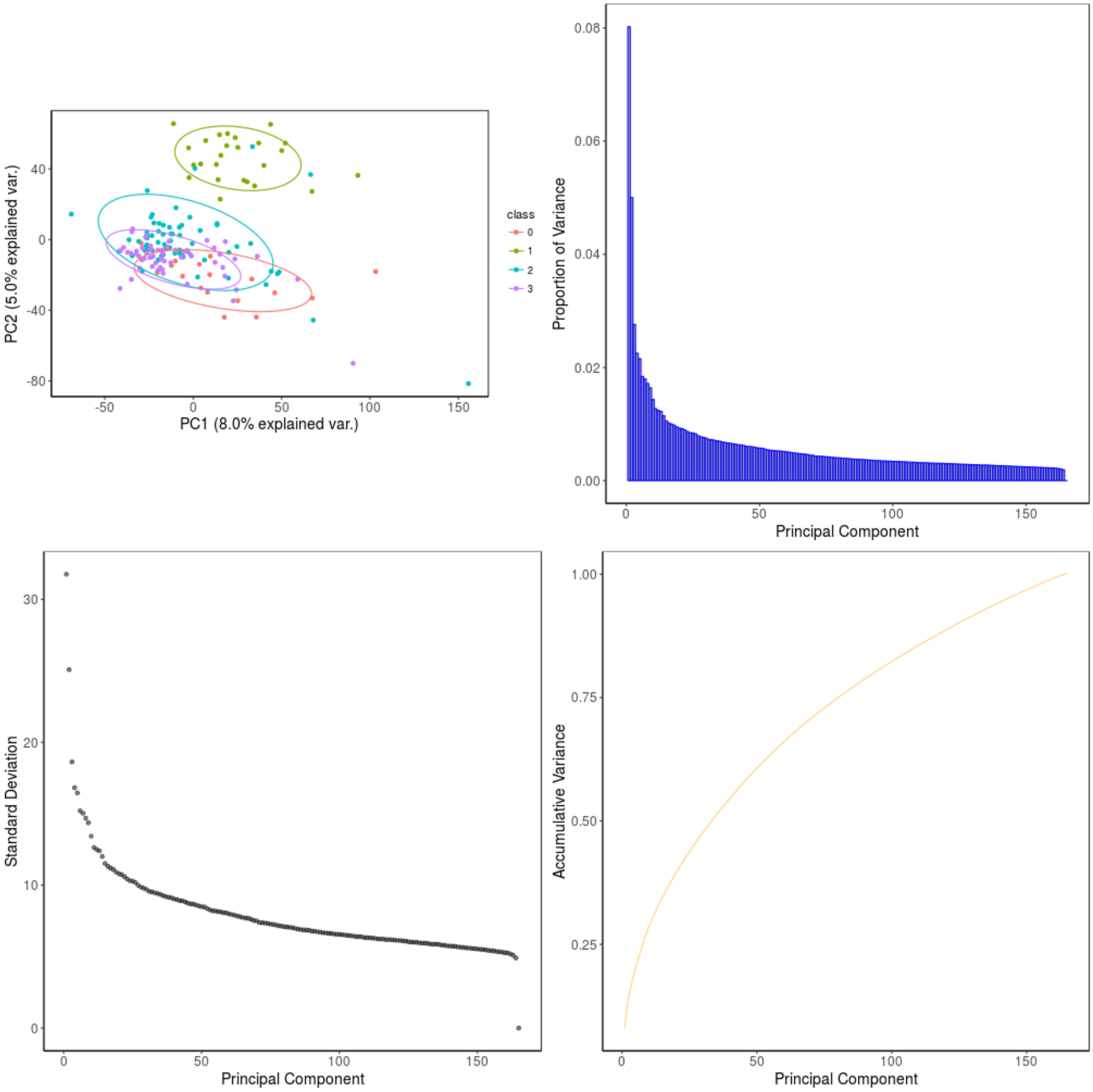
Workflow runtime

85.014 minutes

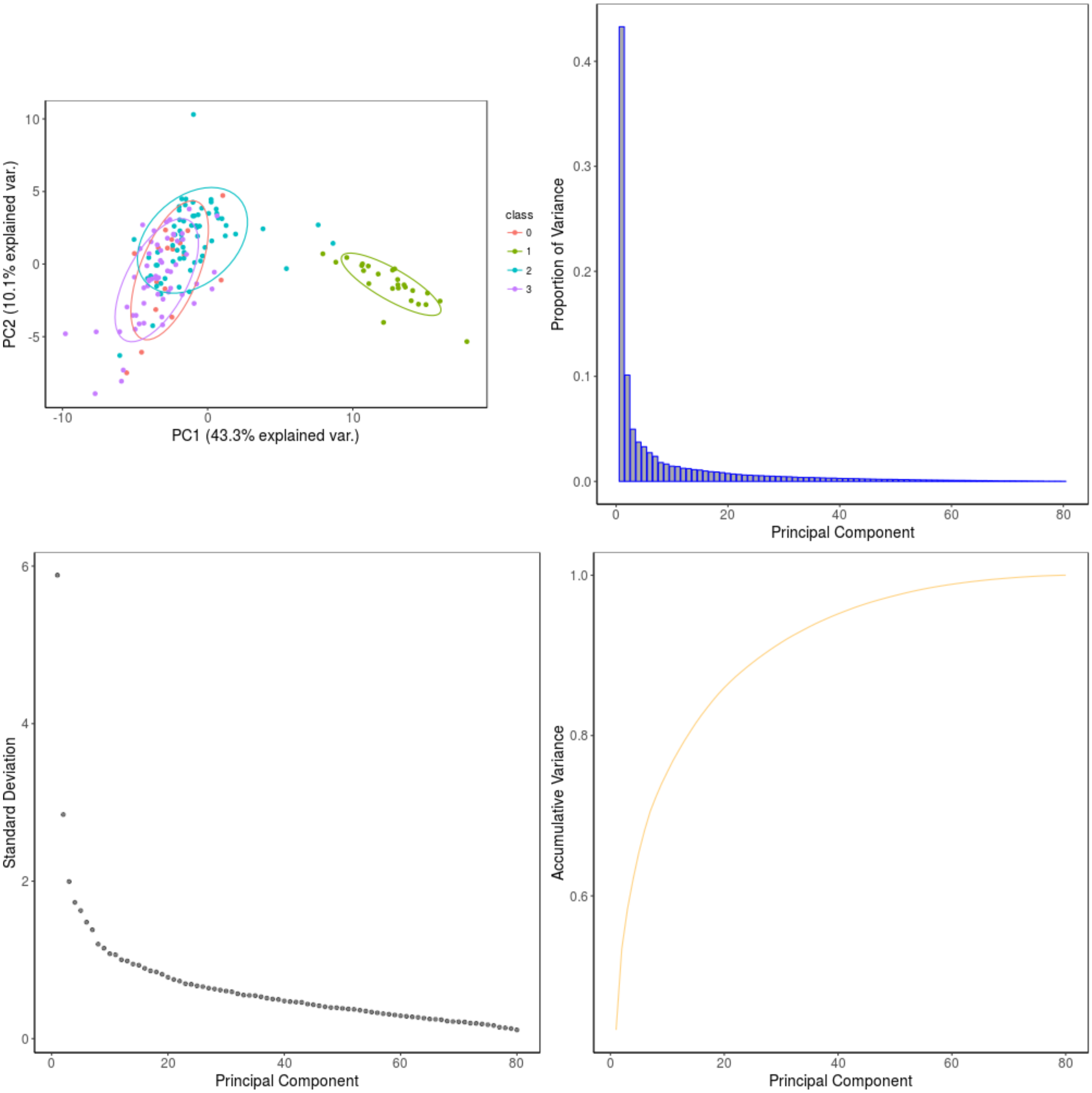
Plots

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 (10.1% explained var.)
PC1 (43.3% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component