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_GPL8300).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.409	0.1511	0.1432	0.1773
2	0.4972	0.2599	0.09301	0.1394
3	0.488	0.2454	0.08514	0.1403
4	0.5544	0.3377	0.148	0.2401
5	0.5715	0.375	0.09287	0.1536
6	0.5813	0.3829	0.1148	0.1904
7	0.5813	0.3829	0.1161	0.1939
8	0.5873	0.3898	0.09826	0.1638
9	0.5996	0.4163	0.1011	0.1629
10	0.6064	0.42	0.1016	0.1658
15	0.6534	0.4971	0.09695	0.1335
20	0.6541	0.4919	0.1053	0.1602
25	0.6429	0.4745	0.09305	0.143
30	0.6162	0.4309	0.1215	0.19
35	0.6255	0.4482	0.1114	0.1732
40	0.6409	0.4676	0.1202	0.1898
45	0.6689	0.5077	0.07844	0.1303
50	0.6407	0.4638	0.1378	0.2189
60	0.6318	0.4535	0.1445	0.219
70	0.635	0.4587	0.1351	0.2051
80	0.6348	0.464	0.1309	0.196
90	0.6218	0.4397	0.1339	0.2102
100	0.6446	0.4711	0.1566	0.2419
479	0.6543	0.4831	0.1428	0.2219

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	70	0.7143	0.5894	2.677e-07
2	90	0.7679	0.6573	2.45e-09
3	479	0.6964	0.5607	1.078e-06
4	50	0.7321	0.612	6.116e-08
5	90	0.6964	0.5603	1.078e-06
6	70	0.6607	0.503	1.375e-05
7	70	0.8036	0.714	6.683e-11
8	479	0.6786	0.5279	4.002e-06
9	479	0.75	0.6375	1.281e-08
10	60	0.8036	0.712	6.683e-11
11	90	0.5714	0.3841	0.002204
12	479	0.6964	0.5522	1.078e-06
13	80	0.75	0.6312	1.281e-08
14	**45**	**0.8036**	**0.7174**	**6.683e-11**
15	20	0.7143	0.589	2.677e-07
16	50	0.7679	0.6664	2.45e-09
17	25	0.7143	0.5803	2.677e-07
18	479	0.75	0.6425	1.281e-08
19	100	0.6786	0.5254	4.002e-06
20	479	0.6607	0.4991	1.375e-05

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.7205	0.05743	0.8036

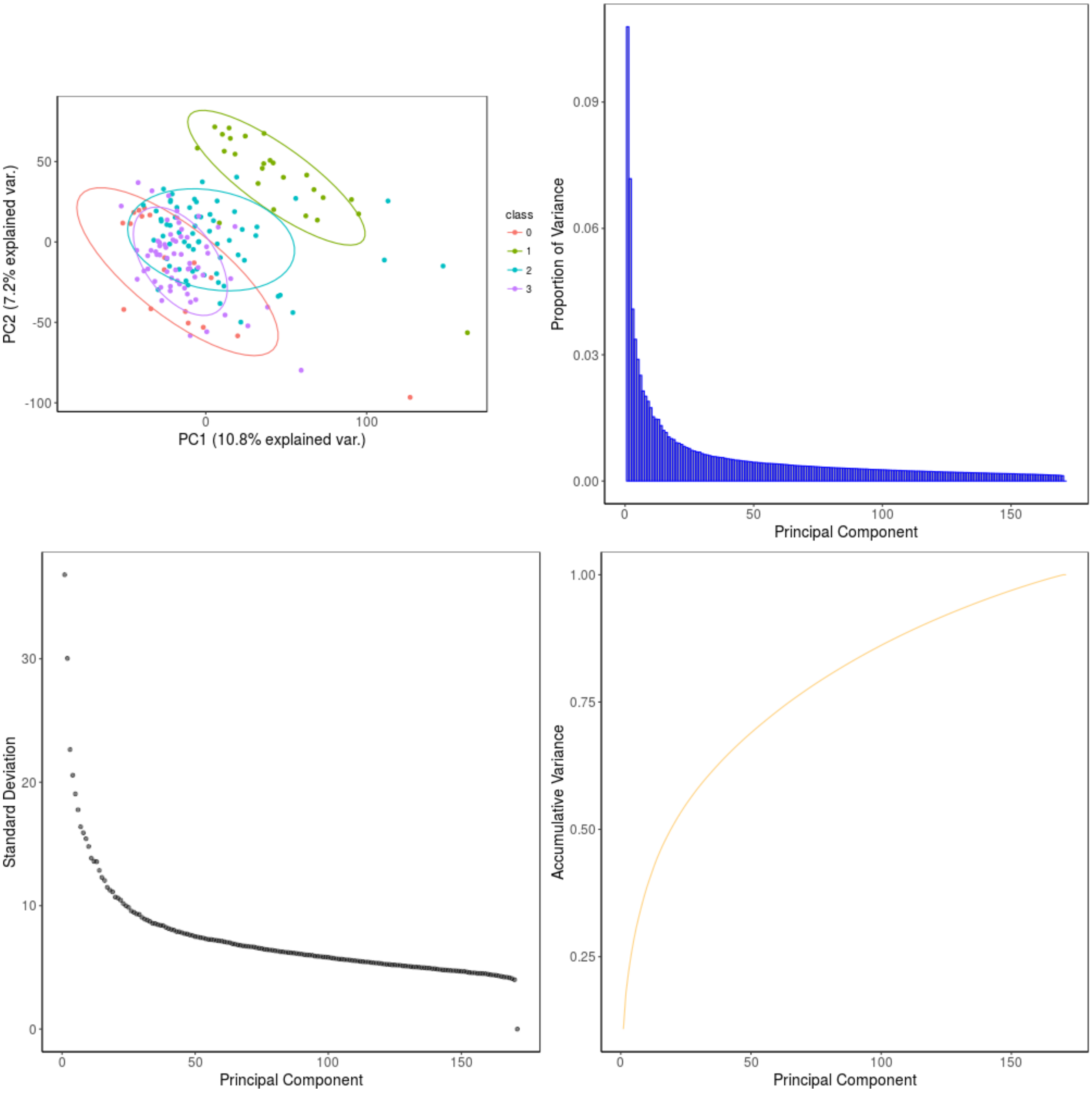
Workflow runtime

9.348 minutes

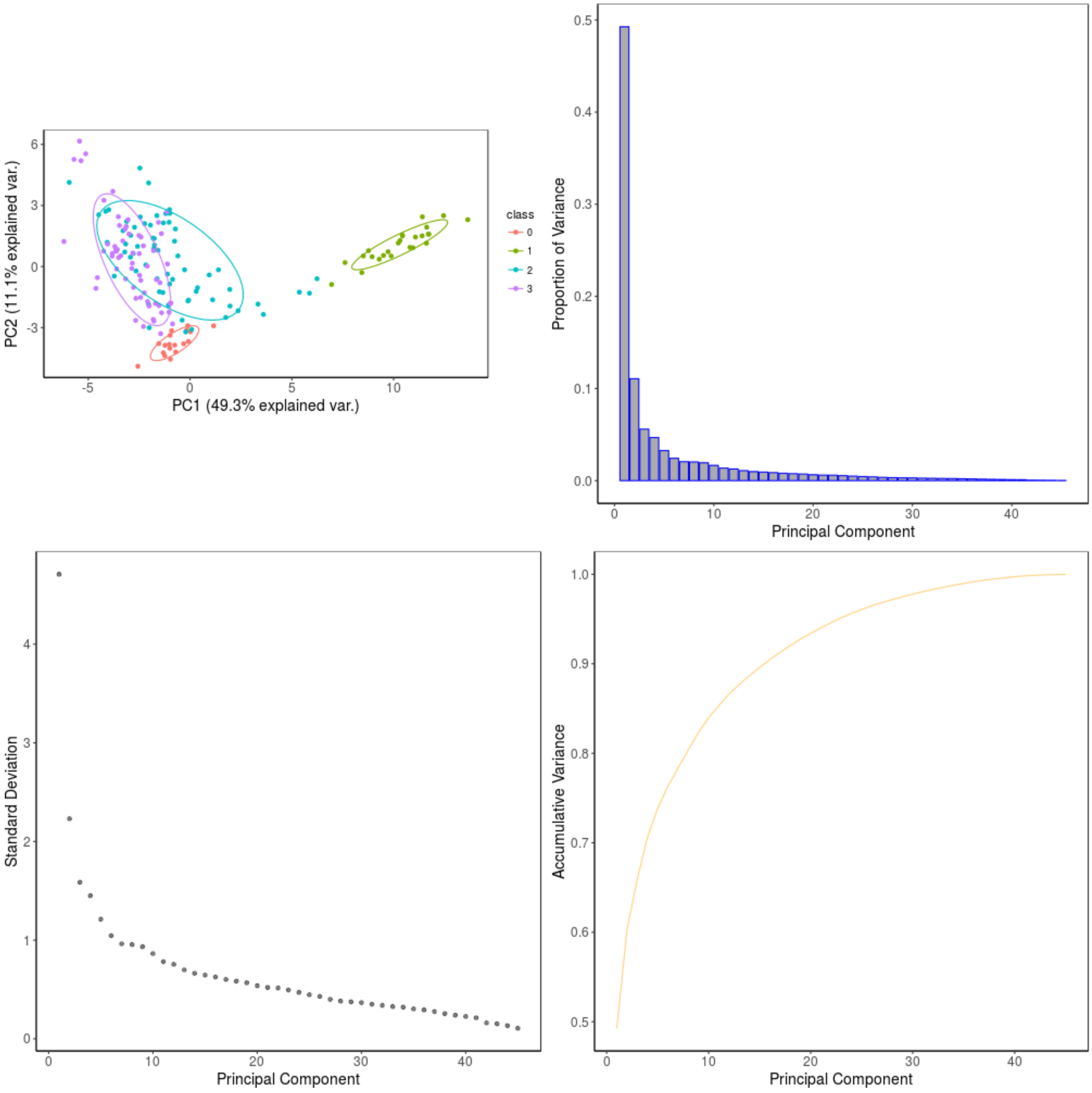
Plots

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 (11.1% explained var.)
PC1 (49.3% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component