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 Multivariate Correlation filter (MC) follow by 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	0.5055	0.3003	0.1167	0.1974
2	0.5327	0.3086	0.1059	0.1543
3	0.6298	0.4504	0.109	0.1578
4	0.6282	0.45	0.07632	0.1016
5	0.683	0.5307	0.1166	0.1582
6	0.648	0.4827	0.1269	0.1706
7	0.6779	0.527	0.1806	0.2479
8	0.6764	0.5214	0.1609	0.2228
9	0.6574	0.4951	0.1641	0.2288
10	0.6748	0.5249	0.1665	0.2278
15	0.6738	0.5175	0.1492	0.2082
20	0.692	0.5366	0.09322	0.1352
25	0.6994	0.5518	0.09171	0.1255
30	0.6629	0.4957	0.1039	0.1496
35	0.6636	0.4962	0.1235	0.1783
40	0.6738	0.5119	0.08651	0.1169
45	0.7027	0.5571	0.1182	0.1566
50	0.7027	0.5599	0.1115	0.14
60	0.6936	0.5428	0.113	0.1527
70	0.6853	0.5307	0.1543	0.2219
80	0.702	0.5524	0.09934	0.1365
90	0.6838	0.5261	0.1102	0.1486
100	0.6573	0.4886	0.1423	0.1921
327	0.6952	0.5449	0.1619	0.2312

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	100	0.7273	0.5846	8.87e-07
2	40	0.6182	0.4231	0.0008835
3	100	0.7455	0.6119	2.106e-07
4	60	0.6727	0.5015	3.979e-05
5	80	0.6182	0.4116	0.0008835
6	50	0.7273	0.5881	8.87e-07
7	60	0.7273	0.5821	8.87e-07
8	100	0.6545	0.4746	0.0001205
9	327	0.7091	0.5535	3.424e-06
10	327	0.7455	0.6158	2.106e-07
11	30	0.6727	0.503	3.979e-05
12	327	0.7455	0.6146	2.106e-07
13	**45**	**0.7818**	**0.6697**	**8.988e-09**
14	25	0.6364	0.4514	0.0003385
15	327	0.6182	0.419	0.0008835
16	100	0.6	0.3913	0.002147
17	50	0.7818	0.6713	8.988e-09
18	327	0.7455	0.6101	2.106e-07
19	45	0.6909	0.536	1.215e-05
20	327	0.6545	0.4703	0.0001205

Accuracy__Mean	Accuracy__SD	Accuracy__Max
0.6936	0.05758	0.7818

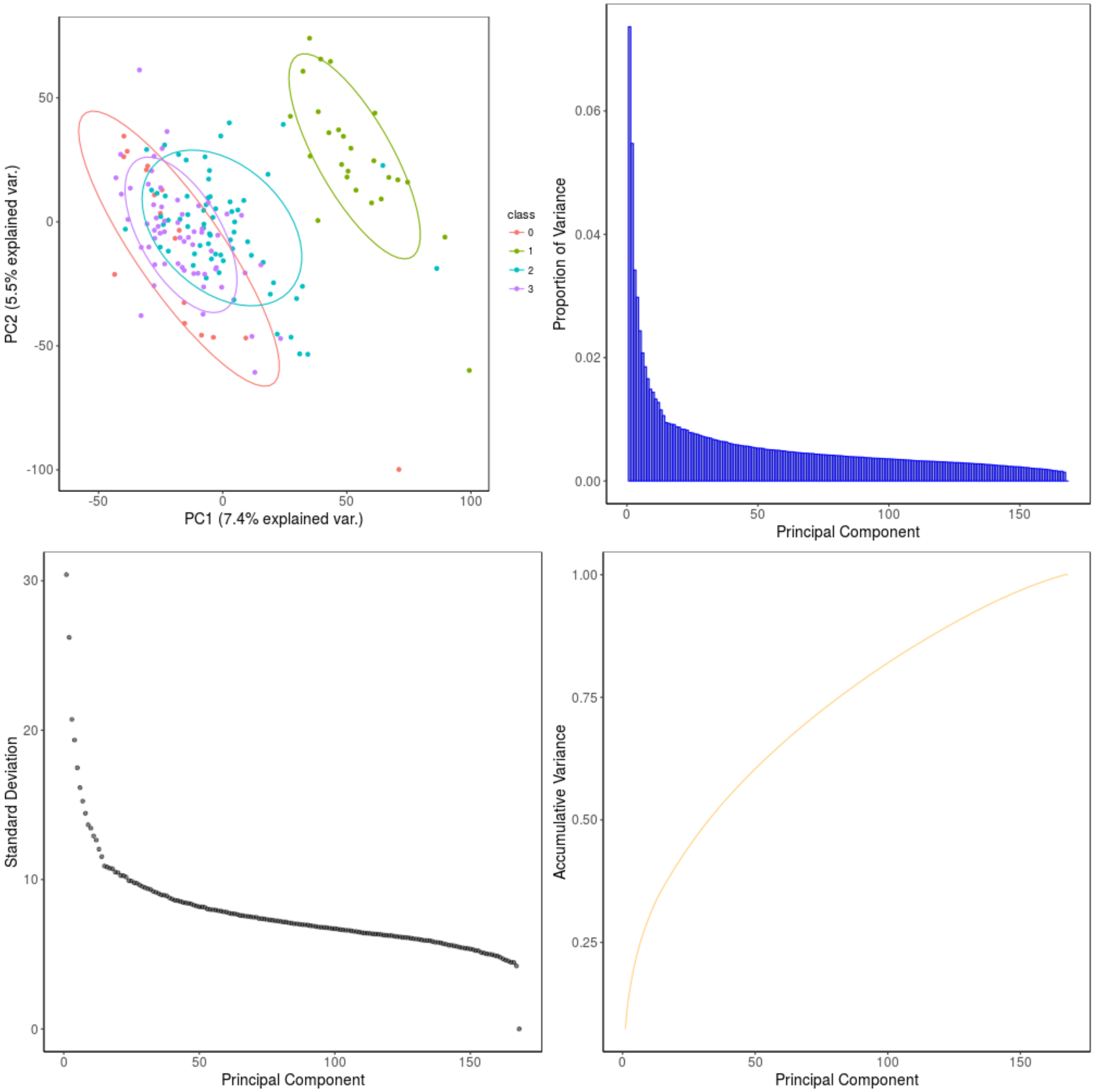
Workflow runtime

8.189 minutes

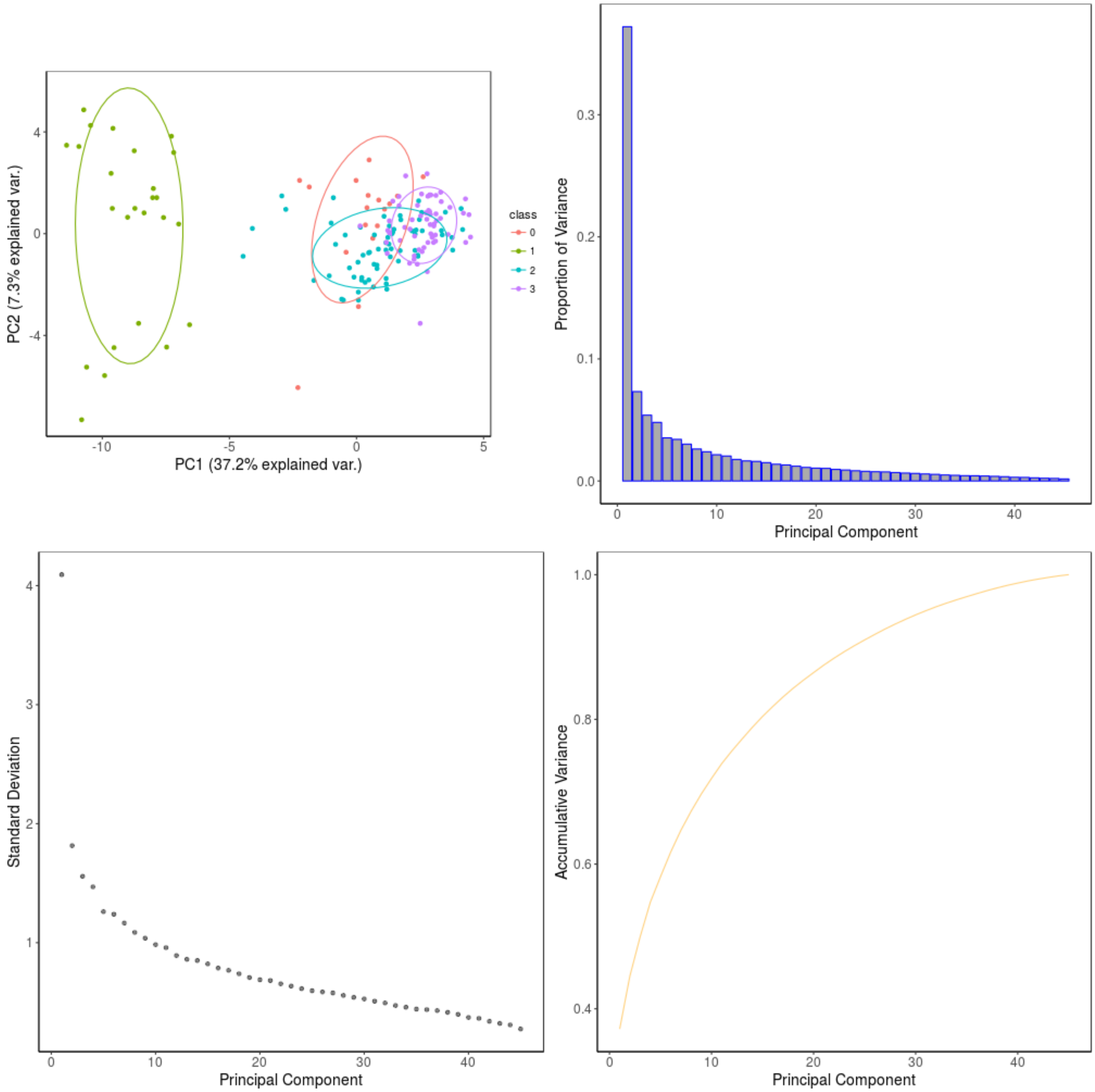
Plots

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 (7.3% explained var.)
PC1 (37.2% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component