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_GPL93).

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4463	0.1884	0.1126	0.1428
2	0.5005	0.2461	0.1049	0.1733
3	0.5888	0.3956	0.1349	0.2021
4	0.576	0.3648	0.1786	0.2794
5	0.5691	0.3598	0.1618	0.2487
6	0.5586	0.3419	0.1671	0.2494
7	0.5691	0.3559	0.1728	0.2645
8	0.6047	0.4036	0.1665	0.2607
9	0.5865	0.3832	0.1496	0.2308
10	0.5879	0.3864	0.1258	0.1987
15	0.612	0.4207	0.1253	0.1881
20	0.637	0.4603	0.1531	0.2283
25	0.6461	0.4738	0.1369	0.2073
30	0.6483	0.4751	0.141	0.2158
35	0.6664	0.5026	0.1445	0.2201
40	0.6673	0.5015	0.127	0.1895
45	0.6673	0.5014	0.127	0.1902
50	0.6673	0.5027	0.1116	0.167
60	0.6483	0.4699	0.1199	0.1876
70	0.6293	0.4454	0.1316	0.196
80	0.6484	0.4767	0.112	0.161
90	0.6193	0.4295	0.1456	0.2222
100	0.6475	0.47	0.1241	0.1899
522	0.6823	0.5245	0.1471	0.2227

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	**522**	**0.7925**	**0.6956**	**4.606e-09**
2	522	0.6604	0.4863	9.108e-05
3	60	0.6792	0.5177	2.884e-05
4	50	0.6038	0.4089	0.001802
5	522	0.6981	0.5485	8.413e-06
6	80	0.717	0.5809	2.255e-06
7	522	0.7736	0.6558	2.513e-08
8	522	0.6038	0.4035	0.001802
9	15	0.5849	0.3974	0.004204
10	90	0.6604	0.4915	9.108e-05
11	90	0.717	0.5811	2.255e-06
12	522	0.7358	0.5932	5.533e-07
13	90	0.6604	0.486	9.108e-05
14	8	0.5849	0.3775	0.004204
15	25	0.6604	0.4868	9.108e-05
16	522	0.6792	0.5182	2.884e-05
17	50	0.6604	0.4979	9.108e-05
18	30	0.7547	0.6296	1.238e-07
19	60	0.7547	0.636	1.238e-07
20	70	0.7358	0.6111	5.533e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.6858	0.06191	0.7925

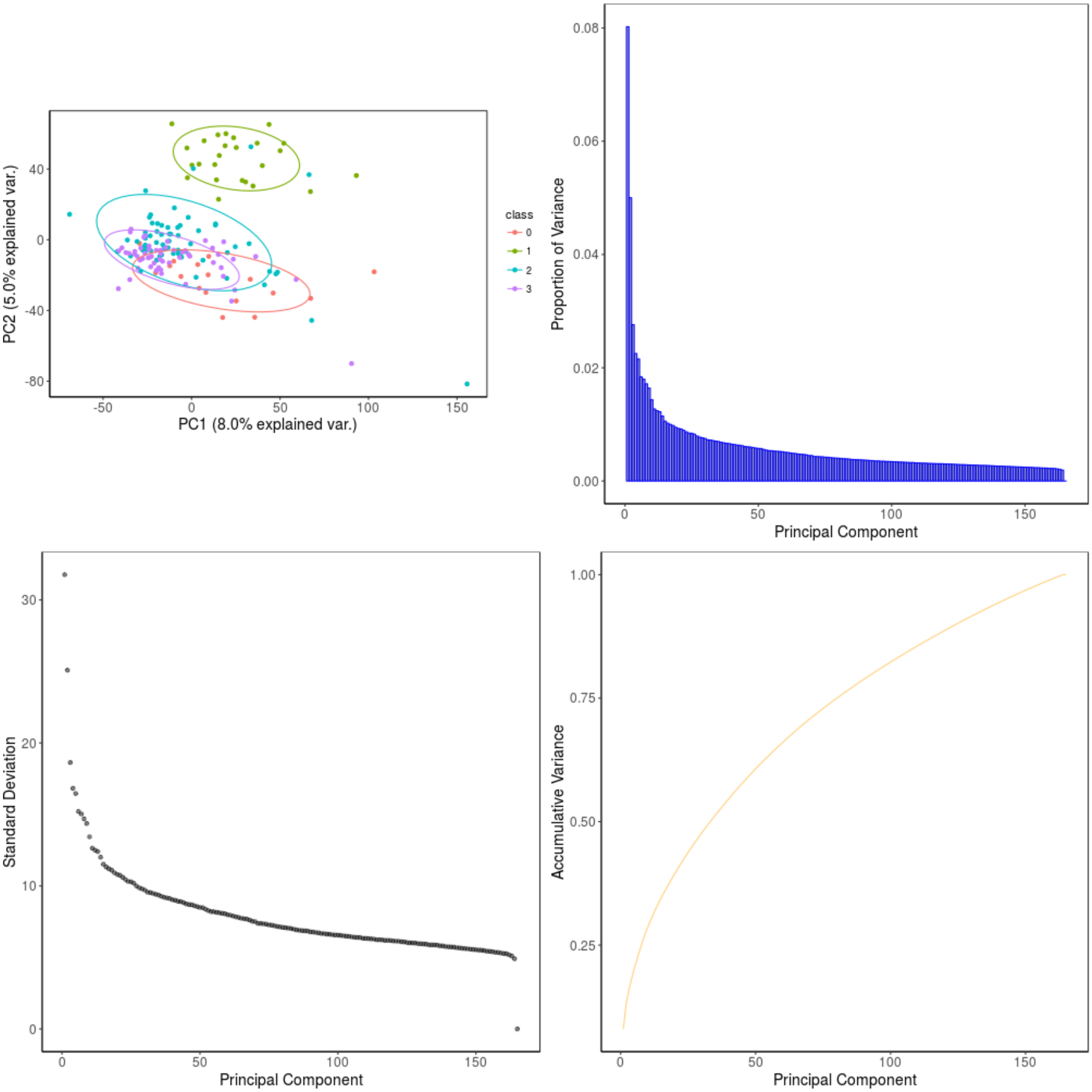
Workflow runtime

14.338 minutes

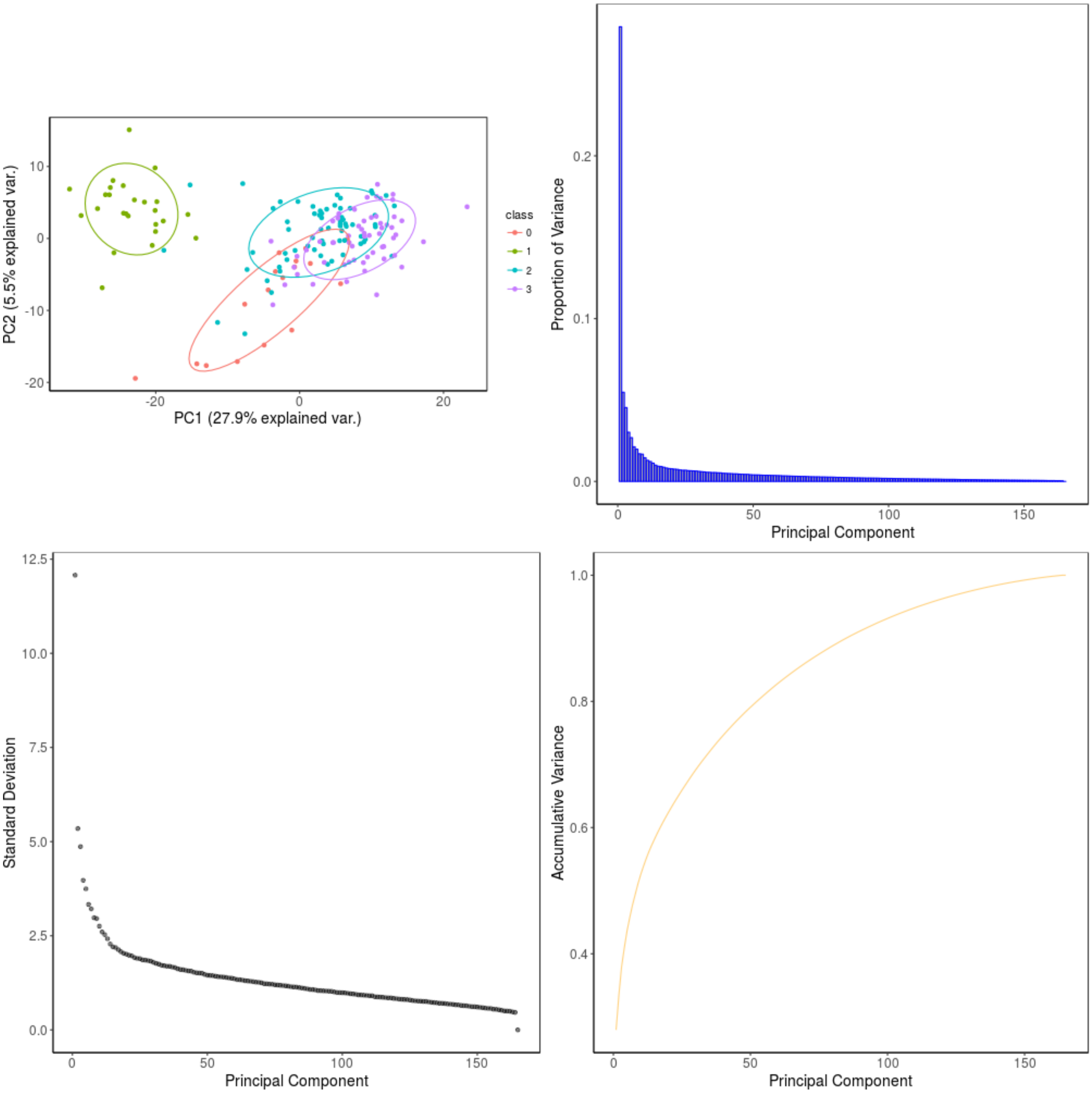
Plots

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 (5.5% explained var.)
PC1 (27.9% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component