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

Triple-Negative Breast Cancer (TNBC) proteome. Label-free deep proteome analysis of 44 (samples and technical repliques) human breast specimens.

Summary stats from training phase

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

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.4083	0.3029	0.4129	0.4471
2	0.525	0.3418	0.3094	0.4151
3	0.5333	0.3579	0.3562	0.4732
4	0.575	0.4007	0.3631	0.4965
5	0.7083	0.62	0.2998	0.3846
6	0.7667	0.6956	0.2772	0.3638
7	0.7417	0.6623	0.3129	0.4127
8	0.7667	0.6885	0.2772	0.3672
9	0.7417	0.65	0.3129	0.4191
10	0.7917	0.7218	0.2613	0.3459
15	0.8583	0.8123	0.1927	0.2519
20	0.8917	0.8551	0.1845	0.2427
25	0.8833	0.843	0.1532	0.209
30	0.825	0.774	0.2203	0.2837
35	0.85	0.8073	0.225	0.2892
40	0.8583	0.8071	0.1524	0.2095
45	0.825	0.7766	0.2498	0.3178
50	0.8	0.7476	0.2428	0.3037
60	0.8833	0.85	0.1933	0.2439
70	0.825	0.7667	0.2498	0.3459
80	0.7833	0.6762	0.2582	0.4083
90	0.8667	0.7859	0.1851	0.3329
100	0.825	0.7667	0.2498	0.3459
757	0.8333	0.743	0.1884	0.3298

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	50	1	1	2.216e-07
2	20	0.7692	0.7	0.000816
3	**20**	**1**	**1**	**2.216e-07**
4	80	1	1	2.216e-07
5	80	1	1	2.216e-07
6	70	0.8462	0.7969	9.42e-05
7	30	0.9231	0.9	6.703e-06
8	757	0.8462	0.7969	9.42e-05
9	30	0.9231	0.8992	6.703e-06
10	757	0.8462	0.803	9.42e-05
11	35	1	1	2.216e-07
12	50	1	1	2.216e-07
13	80	0.9231	0.9	6.703e-06
14	757	0.9231	0.9008	6.703e-06
15	40	0.6923	0.5938	0.004876
16	100	1	1	2.216e-07
17	45	1	1	2.216e-07
18	15	0.8462	0.8	9.42e-05
19	100	0.8462	0.803	9.42e-05
20	25	1	1	2.216e-07

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.9192	0.09161	1

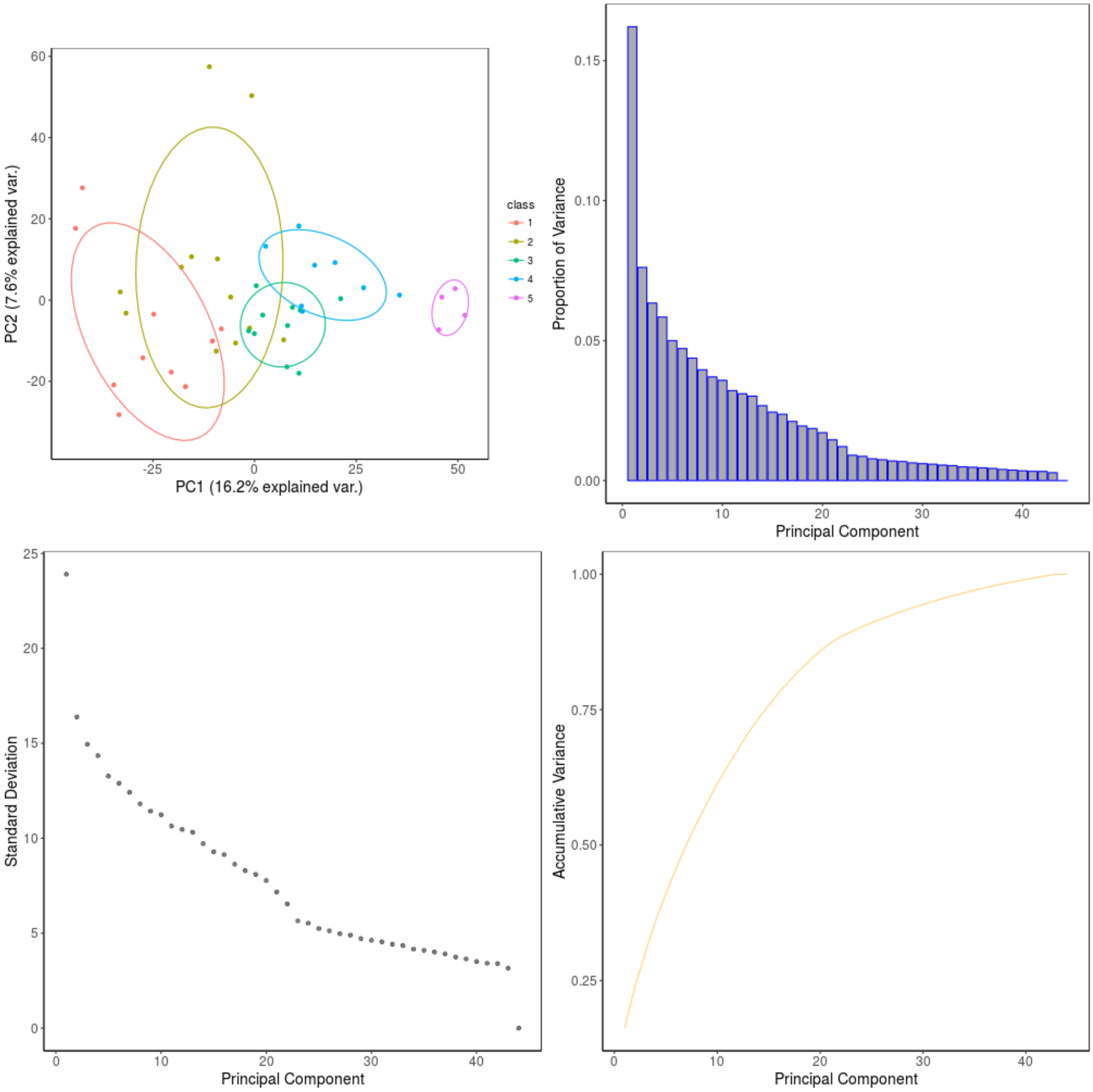
Workflow runtime

3.024 minutes

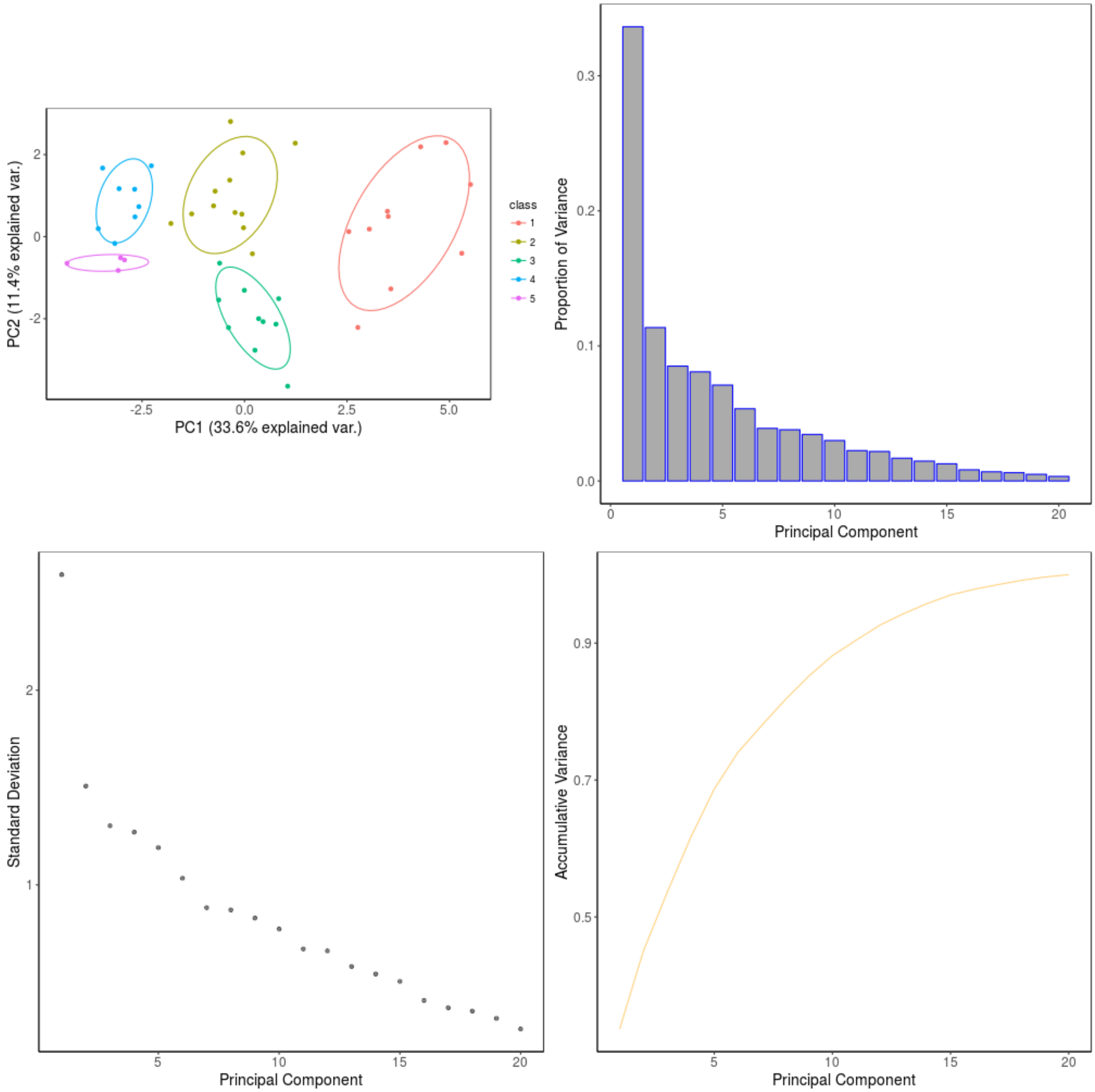
Plots

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 (11.4% explained var.)
PC1 (33.6% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component