Feature Selection (FS) workflow report

Enrique Audain

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

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	0.3917	0.195	0.2005	0.2385
2	0.575	0.4405	0.3202	0.4173
3	0.625	0.4857	0.2972	0.4168
4	0.625	0.4857	0.2523	0.3438
5	0.675	0.5417	0.2734	0.3834
6	0.675	0.5488	0.2734	0.3832
7	0.6833	0.5726	0.225	0.3029
8	0.6917	0.5964	0.2547	0.3176
9	0.7833	0.7109	0.2194	0.2754
10	0.7833	0.7109	0.2194	0.2754
15	0.9083	0.8738	0.1493	0.207
20	0.8417	0.7917	0.2306	0.2867
25	0.8667	0.825	0.2331	0.2899
30	0.8833	0.8333	0.1532	0.2222
35	0.9417	0.9167	0.1245	0.18
40	0.9083	0.8738	0.1493	0.207
45	0.9417	0.9167	0.1245	0.18
50	0.9167	0.8833	0.1361	0.1933
60	0.9167	0.8833	0.1361	0.1933
70	0.9417	0.9167	0.1245	0.18
80	0.9417	0.9167	0.1245	0.18
90	0.9417	0.9167	0.1245	0.18
100	0.9417	0.9167	0.1245	0.18
3524	0.9417	0.9167	0.1245	0.18

Summary stats from testing phase

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

run	Variables	Accuracy	Kappa	AccuracyPValue
1	45	0.7692	0.6977	0.000816
2	60	1	1	2.216e-07
3	40	0.8462	0.8045	9.42e-05
4	**35**	**1**	**1**	**2.216e-07**
5	40	0.9231	0.9	6.703e-06
6	45	1	1	2.216e-07
7	70	0.8462	0.8	9.42e-05
8	80	1	1	2.216e-07
9	20	0.9231	0.9015	6.703e-06
10	90	1	1	2.216e-07
11	70	1	1	2.216e-07
12	3524	1	1	2.216e-07
13	60	0.8462	0.7984	9.42e-05
14	40	0.5385	0.3906	0.07065
15	45	0.9231	0.9	6.703e-06
16	3524	1	1	2.216e-07
17	100	0.9231	0.8992	6.703e-06
18	3524	0.8462	0.803	9.42e-05
19	30	0.8462	0.803	9.42e-05
20	3524	0.9231	0.9008	6.703e-06

Accuracy_Mean	Accuracy_SD	Accuracy_Max
0.9077	0.1133	1

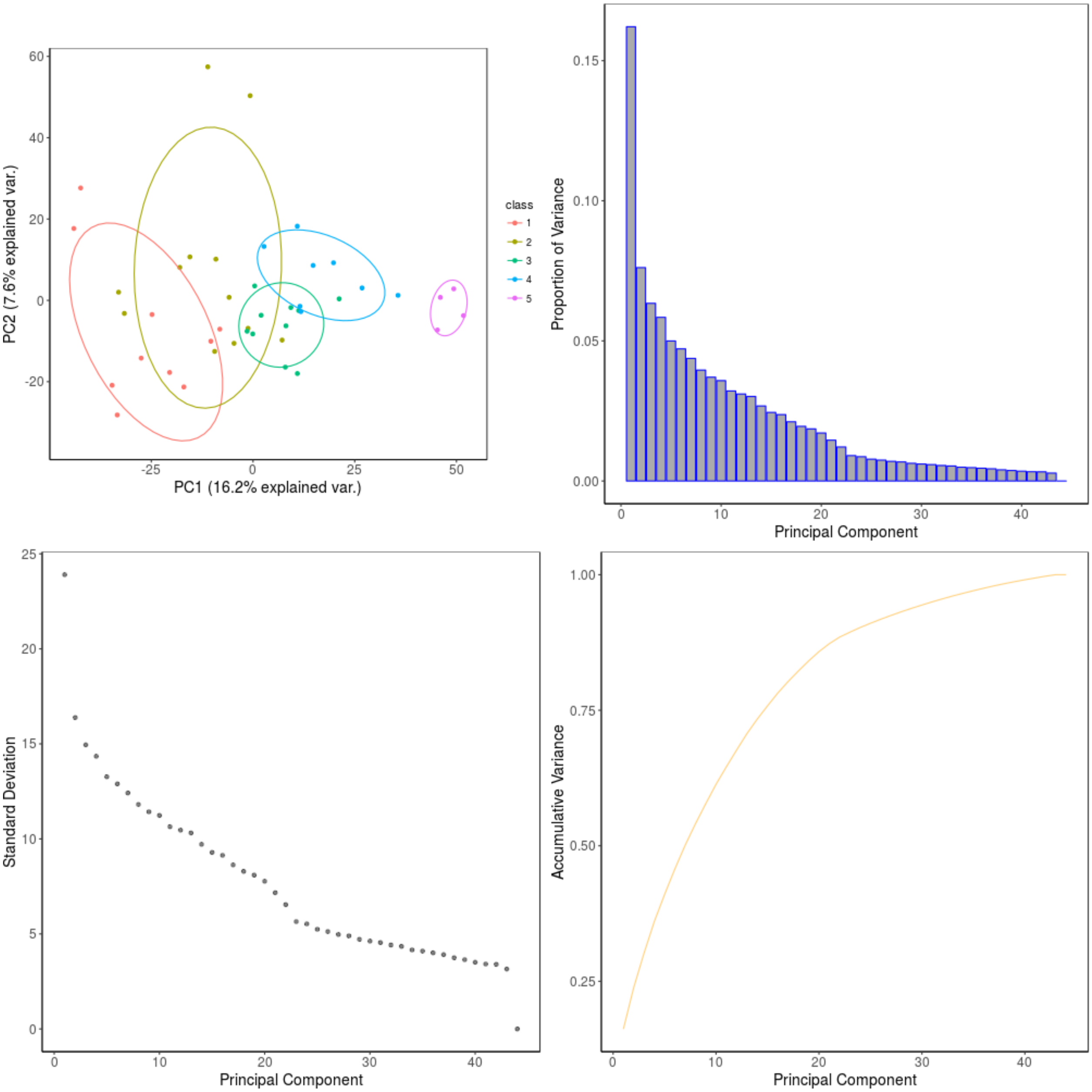
Workflow runtime

6.971 minutes

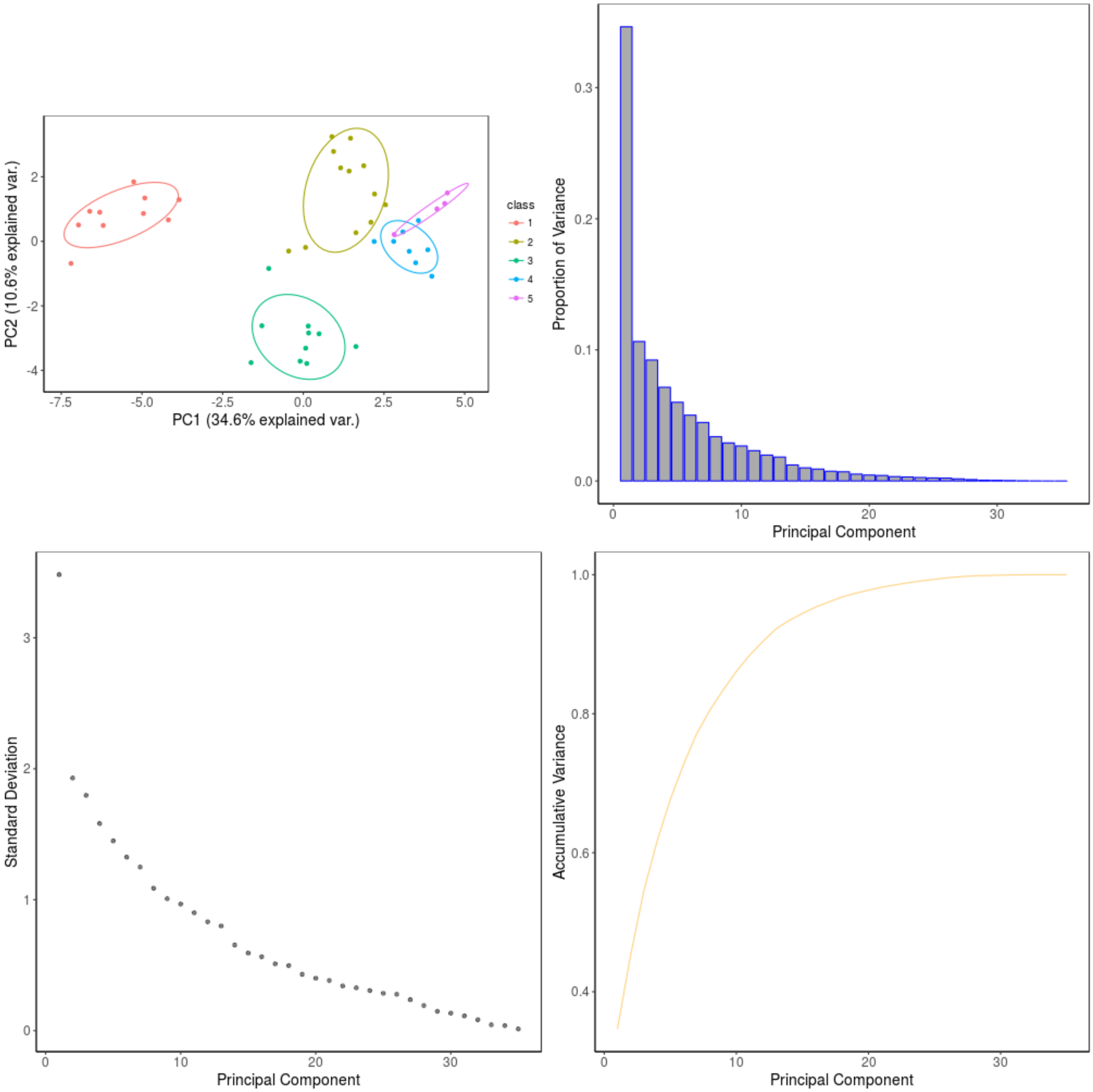
Plots

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 (10.6% explained var.)
PC1 (34.6% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component