# 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

Analysis of breast cancer tumor samples using 2-color cDNA microarrays (GSE5325).

## Summary stats from training phase

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

| Variables | Accuracy | Kappa | AccuracySD | KappaSD |
|---|---|---|---|---|
| 1 | 0.6 | 0.1724 | 0.1127 | 0.2126 |
| 2 | 0.7 | 0.3915 | 0.1421 | 0.2688 |
| 3 | 0.7571 | 0.5113 | 0.1788 | 0.3498 |
| 4 | 0.7571 | 0.5155 | 0.1656 | 0.3182 |
| 5 | 0.7857 | 0.5741 | 0.1934 | 0.3738 |
| 6 | 0.7714 | 0.536 | 0.1536 | 0.2961 |
| 7 | 0.7714 | 0.5433 | 0.1536 | 0.2898 |
| 8 | 0.8 | 0.607 | 0.1677 | 0.3147 |
| 9 | 0.7857 | 0.5765 | 0.1543 | 0.2859 |
| 10 | 0.7857 | 0.5765 | 0.1543 | 0.2859 |
| 15 | 0.7857 | 0.5769 | 0.1934 | 0.3699 |
| 20 | 0.7857 | 0.5702 | 0.1543 | 0.3004 |
| 25 | 0.8286 | 0.6587 | 0.1475 | 0.2865 |
| 30 | 0.8143 | 0.6284 | 0.1513 | 0.2924 |
| 35 | 0.8 | 0.6005 | 0.1536 | 0.2985 |
| 40 | 0.8 | 0.5958 | 0.138 | 0.271 |
| 45 | 0.7857 | 0.5701 | 0.1388 | 0.2671 |
| 50 | 0.7857 | 0.5654 | 0.1214 | 0.2353 |
| 60 | 0.8143 | 0.6285 | 0.1656 | 0.3231 |
| 70 | 0.8143 | 0.6262 | 0.1355 | 0.2656 |
| 80 | 0.8429 | 0.68 | 0.1421 | 0.2826 |
| 90 | 0.8286 | 0.652 | 0.1313 | 0.2603 |
| 100 | 0.8429 | 0.6848 | 0.171 | 0.3369 |
| 1697 | 0.7857 | 0.5701 | 0.1388 | 0.2671 |

## Summary stats from testing phase

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

| run | Variables | Accuracy | Kappa | AccuracyPValue |
|---:|---:|---:|---:|---:|
| 1 | 25 | 0.8571 | 0.7107 | 0.0003014 |
| 2 | 5 | 0.8286 | 0.65 | 0.001202 |
| 3 | 35 | 0.8 | 0.595 | 0.003999 |
| 4 | 45 | 0.8857 | 0.7667 | 6.136e-05 |
| 5 | 1697 | 0.8571 | 0.7009 | 0.0003014 |
| 6 | 80 | 0.9143 | 0.8264 | 9.733e-06 |
| 7 | 6 | 0.8857 | 0.7667 | 6.136e-05 |
| 8 | 35 | 0.8286 | 0.65 | 0.001202 |
| 9 | 10 | 0.9143 | 0.8235 | 9.733e-06 |
| 10 | 1 | 0.6857 | 0.3937 | 0.115 |
| 11 | 9 | 0.8857 | 0.7627 | 6.136e-05 |
| 12 | 30 | 0.8857 | 0.7667 | 6.136e-05 |
| 13 | 25 | 0.8 | 0.5882 | 0.003999 |
| 14 | 1697 | 0.8286 | 0.6379 | 0.001202 |
| 15 | 35 | 0.7714 | 0.5333 | 0.01134 |
| 16 | 25 | 0.8857 | 0.7705 | 6.136e-05 |
| 17 | 20 | 0.8571 | 0.7107 | 0.0003014 |
| **18** | **80** | **0.9429** | **0.8814** | **1.128e-06** |
| 19 | 70 | 0.8 | 0.608 | 0.003999 |
| 20 | 15 | 0.8286 | 0.65 | 0.001202 |

| Accuracy_Mean | Accuracy_SD | Accuracy_Max |
|:---:|:---:|:---:|
| 0.8471 | 0.05883 | 0.9429 |

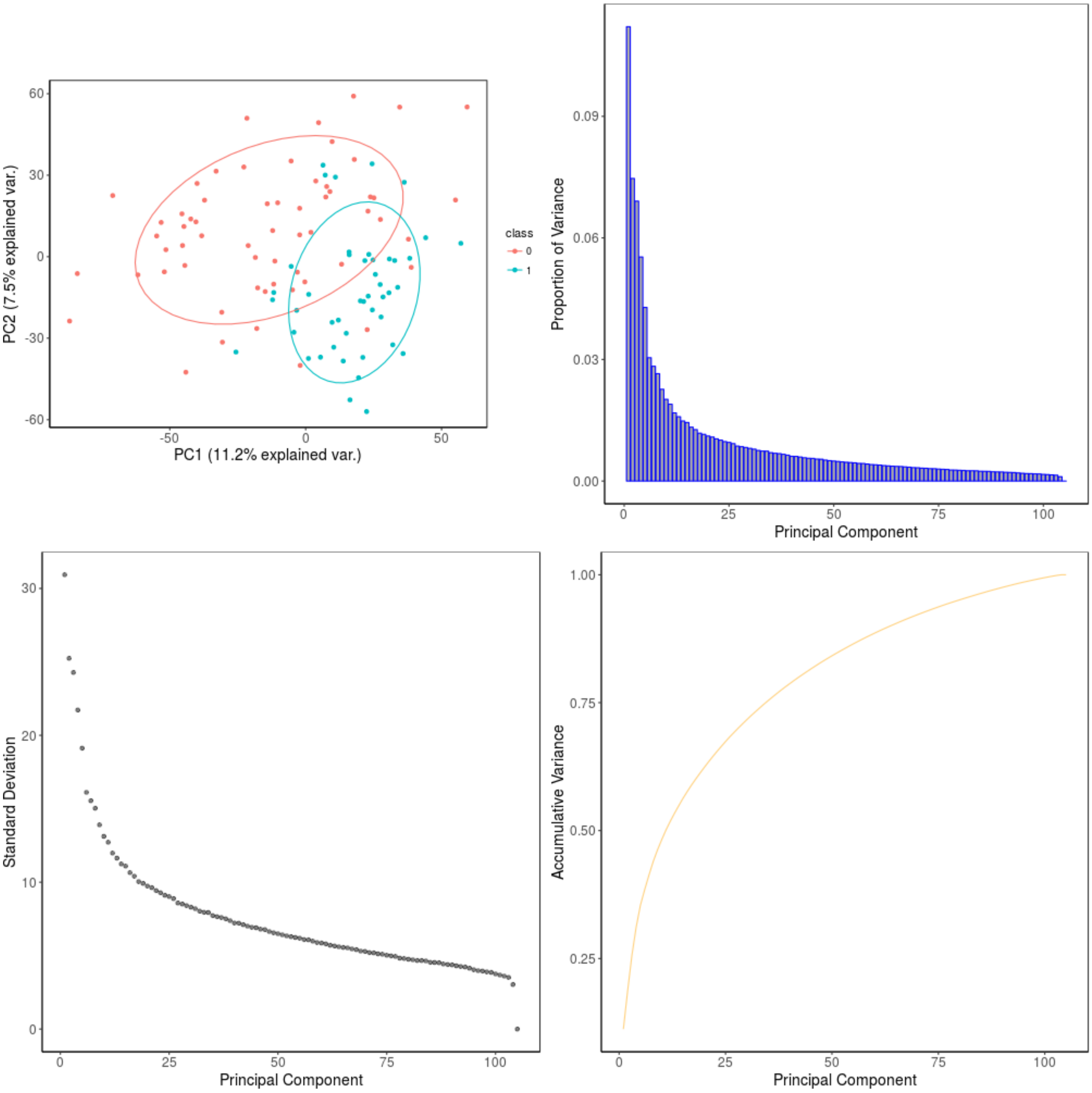
## Workflow runtime

8.178 minutes

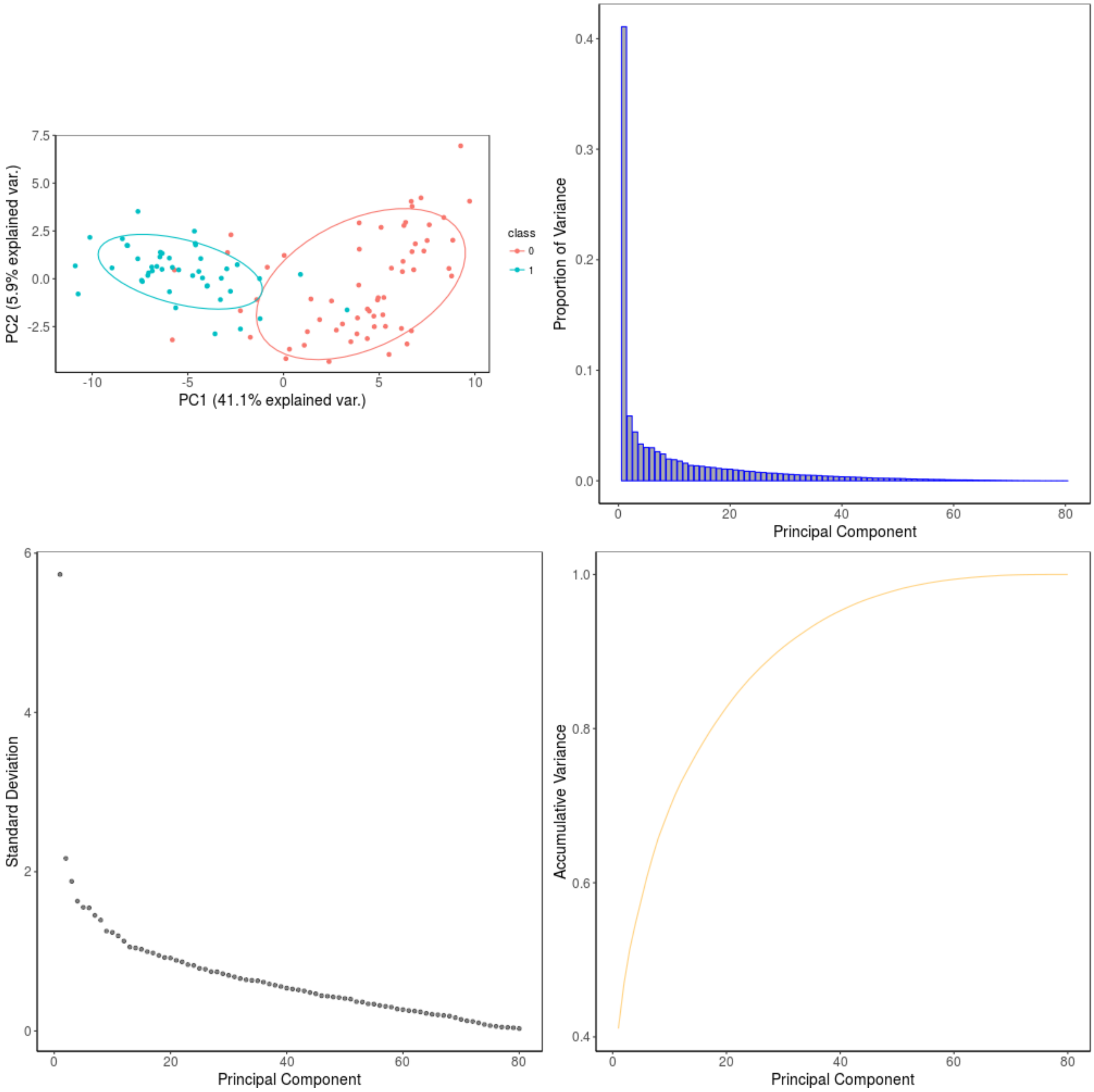
# Plots

## 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 (5.9% explained var.)
PC1 (41.1% explained var.)
class
Proportion of Variance
Principal Component
Standard Deviation
Principal Component
Accumulative Variance
Principal Component