Project 2

Jeffrey Zhuohui Liang

3/24/2021

```
## Warning: Missing column names filled in: 'X560' [560]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   cell_name = col_character(),
##   X560 = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
## Warning: 716 parsing failures.
## row col    expected      actual     file
##   1  -- 560 columns 559 columns 'ss.csv'
##   2  -- 560 columns 559 columns 'ss.csv'
##   3  -- 560 columns 559 columns 'ss.csv'
##   4  -- 560 columns 559 columns 'ss.csv'
##   5  -- 560 columns 559 columns 'ss.csv'
## ... ... ........... ........... ........
## See problems(...) for more details.
```

Method

Data Preparation

```
# drop gene that 90% is zero
drop_gene =
  sngcll %>%
  summarise(across(everything(),~sum(.x==0)/n()<0.9)) %>%
  slice(1) %>%
  unlist() %>%
  as.vector()

sngcll_dp0 =
  sngcll[,drop_gene]
```

```r
# PCA with scale
sngcll_pca =
  #predict(preProcess(sngcll_dp0,c("center","scale","pca")),sngcll_dp0)
  prcomp( ~ .,
          data = sngcll_dp0,
          tol = sqrt(.Machine$double.eps),
          center = T,
          scale = T)

summary(sngcll_pca)$importance %>%
  t() %>%
  .[seq(1,230,8),] %>%
  as_tibble(rownames = NA) %>%
  knitr::kable(digits = 2)
```

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	5.12	0.11	0.11
PC9	2.16	0.02	0.35
PC17	1.62	0.01	0.47
PC25	1.41	0.01	0.54
PC33	1.28	0.01	0.60
PC41	1.16	0.01	0.66
PC49	1.10	0.01	0.70
PC57	1.01	0.00	0.74
PC65	0.95	0.00	0.77
PC73	0.90	0.00	0.80
PC81	0.85	0.00	0.83
PC89	0.80	0.00	0.85
PC97	0.76	0.00	0.87
PC105	0.72	0.00	0.89
PC113	0.67	0.00	0.91
PC121	0.63	0.00	0.92
PC129	0.59	0.00	0.93
PC137	0.56	0.00	0.94
PC145	0.53	0.00	0.95
PC153	0.49	0.00	0.96
PC161	0.45	0.00	0.97
PC169	0.42	0.00	0.98
PC177	0.39	0.00	0.98
PC185	0.35	0.00	0.99
PC193	0.32	0.00	0.99
PC201	0.28	0.00	0.99
PC209	0.25	0.00	1.00
PC217	0.22	0.00	1.00
PC225	0.16	0.00	1.00

Using 90% as threshold for Principle component selection, we see that up to 110 can explain 90% of the standard deviation in data.

```r
sngcll_pca =
  predict(preProcess(
    sngcll_dp0,
    c("center", "scale", "pca"),
    pcaComp = sum(summary(sngcll_pca)$importance[3, ]<.9)
  ), sngcll_dp0)
```