Stat 437 Project 2

Ling Jin (student ID 011880184)

General rule and information

You must show your work in order to get points. Please prepare your report according to the rubrics on projects that are given in the syllabus. If a project report contains only codes and their outputs and the project has a total of 100 points, a maximum of 25 points can be taken off. Please note that your need to submit codes that would have been used for your data analysis. Your report can be in .doc, .docx, .html or .pdf format.

The project will assess your skills in support vector machines and dimension reduction, for which visualization techniques you have learnt will be used to illustrate your findings. This project gives you more freedom to use your knowledge and skills in data analysis.

Task A: Analysis of gene expression data

For this task, you need to use PCA and Sparse PCA.

Data set and its description

Please download the data set "TCGA-PANCAN-HiSeq-801x20531.tar.gz" from the website https://archive.ics.uci.edu/ml/machine-learning-databases/00401/. A brief description of the data set is given at https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq. Please read the description carefully, and you may need to read a bit more on gene expression data to help you complete this project.

You need to decompress the data file since it is a .tar.gz file. Once uncompressed, the data files are "labels.csv" that contains the cancer type for each sample, and "data.csv" that contains the "gene expression profile" (i.e., expression measurements of a set of genes) for each sample. Here each sample is for a subject and is stored in a row of "data.csv". In fact, the data set contains the gene expression profiles for 801 subjects, each with a cancer type, where each gene expression profile contains the gene expressions for the same set of 20531 genes. The cancer types are: "BRCA", "COAD", "KIRC", "LUAD" and "PRAD". In both files "labels.csv" and "data.csv", each row name records which sample a label or observation is for.

Data processing

Please use `set.seed(123)` for random sampling via the command `sample`.

- Filter out genes (from "data.csv") whose expressions are zero for at least 300 subjects, and save the filtered data as R object "gexp2".

- Use the command `sample` to randomly select 1000 genes and their expressions from "gexp2", and save the resulting data as R object "gexp3".
- Use the command `scale` to standardize the gene expressions for each gene in "gexp3". Save the standardized data as R object "stdgexpProj2".

You will analyze the standardized data.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stats)

data <- read.csv("TCGA-PANCAN-HiSeq-801x20531/data.csv", row.names = 1)
labels <- read.csv("TCGA-PANCAN-HiSeq-801x20531/labels.csv",
    row.names = 1)

set.seed(123)

# Filter genes whose expressions are zero for >=300
# subjects
gexp2 <- data[, colSums(data == 0) < 300]

# Randomly select 1000 genes
genes_selected <- sample(colnames(gexp2), 1000)
gexp3 <- gexp2[, genes_selected]

# Standardize
gexp3_scaled <- scale(gexp3)
stdgexpProj2 <- gexp3_scaled
```

Interpretations

Genes with zero expression in at least 300 subjects were removed to reduce noise and retain informative features. From the remaining genes, 1,000 were randomly selected to ensure computational efficiency while preserving variability. The selected gene expression values were then standardized to give each gene equal weight in subsequent PCA and Sparse PCA analyses.

Questions to answer when doing data analysis

Please also investigate and address the following when doing data analysis:

(1.a) Are there genes for which linear combinations of their expressions explain a significant proportion of the variation of gene expressions in the data set? Note that each gene corresponds to a feature, and a principal component based on data version is a linear combination of the expression measurements for several genes.

```
# Perform PCA
pca_result <- prcomp(stdgexpProj2, center = TRUE, scale. = TRUE)

# Proportion of variance explained
summary(pca_result)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     11.134 10.3429 9.20906 7.35728 6.59935 5.80690 5.27376
## Proportion of Variance  0.124  0.1070 0.08481 0.05413 0.04355 0.03372 0.02781
## Cumulative Proportion   0.124  0.2309 0.31575 0.36988 0.41343 0.44715 0.47496
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     4.91642 4.32899 3.7013 3.31466 3.19596 3.13543 2.78237
## Proportion of Variance 0.02417 0.01874 0.0137 0.01099 0.01021 0.00983 0.00774
## Cumulative Proportion  0.49913 0.51787 0.5316 0.54256 0.55277 0.56261 0.57035
##                           PC15    PC16    PC17   PC18   PC19    PC20    PC21
## Standard deviation     2.73661 2.65783 2.62237 2.5700 2.5293 2.40321 2.35799
## Proportion of Variance 0.00749 0.00706 0.00688 0.0066 0.0064 0.00578 0.00556
## Cumulative Proportion  0.57784 0.58490 0.59178 0.5984 0.6048 0.61055 0.61611
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     2.3012 2.27818 2.26151 2.21045 2.16269 2.13407 2.12899
## Proportion of Variance 0.0053 0.00519 0.00511 0.00489 0.00468 0.00455 0.00453
## Cumulative Proportion  0.6214 0.62660 0.63171 0.63660 0.64128 0.64583 0.65036
##                           PC29    PC30  PC31    PC32    PC33    PC34    PC35
## Standard deviation     2.10788 2.05141 1.999 1.97712 1.93339 1.91726 1.90229
## Proportion of Variance 0.00444 0.00421 0.004 0.00391 0.00374 0.00368 0.00362
## Cumulative Proportion  0.65481 0.65902 0.663 0.66692 0.67066 0.67433 0.67795
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     1.87910 1.84069 1.80857 1.80368 1.78526 1.75660 1.74012
## Proportion of Variance 0.00353 0.00339 0.00327 0.00325 0.00319 0.00309 0.00303
## Cumulative Proportion  0.68148 0.68487 0.68814 0.69140 0.69458 0.69767 0.70070
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     1.71516 1.69481 1.67677 1.65689 1.63174 1.62177 1.60336
## Proportion of Variance 0.00294 0.00287 0.00281 0.00275 0.00266 0.00263 0.00257
## Cumulative Proportion  0.70364 0.70651 0.70932 0.71207 0.71473 0.71736 0.71993
##                           PC50    PC51    PC52    PC53   PC54    PC55   PC56
## Standard deviation     1.58959 1.57698 1.57044 1.56285 1.5489 1.53361 1.5172
## Proportion of Variance 0.00253 0.00249 0.00247 0.00244 0.0024 0.00235 0.0023
## Cumulative Proportion  0.72246 0.72495 0.72741 0.72985 0.7322 0.73461 0.7369
##                           PC57    PC58    PC59    PC60    PC61    PC62    PC63
```

```
## Standard deviation     1.49902 1.48972 1.48047 1.47124 1.45517 1.45365 1.43389
## Proportion of Variance 0.00225 0.00222 0.00219 0.00216 0.00212 0.00211 0.00206
## Cumulative Proportion  0.73915 0.74137 0.74357 0.74573 0.74785 0.74996 0.75202
##                           PC64    PC65    PC66    PC67    PC68    PC69    PC70
## Standard deviation     1.42922 1.40999 1.40288 1.40221 1.38921 1.38433 1.37411
## Proportion of Variance 0.00204 0.00199 0.00197 0.00197 0.00193 0.00192 0.00189
## Cumulative Proportion  0.75406 0.75605 0.75802 0.75998 0.76191 0.76383 0.76572
##                           PC71    PC72    PC73    PC74    PC75    PC76   PC77
## Standard deviation     1.36250 1.35906 1.35520 1.33807 1.33292 1.32019 1.3044
## Proportion of Variance 0.00186 0.00185 0.00184 0.00179 0.00178 0.00174 0.0017
## Cumulative Proportion  0.76757 0.76942 0.77126 0.77305 0.77482 0.77657 0.7783
##                           PC78    PC79    PC80    PC81    PC82    PC83   PC84
## Standard deviation     1.3026 1.29426 1.29077 1.28767 1.27858 1.26885 1.2657
## Proportion of Variance 0.0017 0.00168 0.00167 0.00166 0.00163 0.00161 0.0016
## Cumulative Proportion  0.7800 0.78164 0.78331 0.78496 0.78660 0.78821 0.7898
##                           PC85    PC86    PC87    PC88    PC89    PC90    PC91
## Standard deviation     1.24823 1.23912 1.23251 1.23098 1.22197 1.21495 1.21092
## Proportion of Variance 0.00156 0.00154 0.00152 0.00152 0.00149 0.00148 0.00147
## Cumulative Proportion  0.79137 0.79290 0.79442 0.79594 0.79743 0.79891 0.80037
##                           PC92    PC93    PC94   PC95    PC96    PC97    PC98
## Standard deviation     1.20458 1.19711 1.19391 1.1840 1.17525 1.16906 1.16635
## Proportion of Variance 0.00145 0.00143 0.00143 0.0014 0.00138 0.00137 0.00136
## Cumulative Proportion  0.80183 0.80326 0.80468 0.8061 0.80747 0.80883 0.81019
##                           PC99   PC100   PC101   PC102  PC103   PC104   PC105
## Standard deviation     1.15751 1.15613 1.15128 1.14669 1.1424 1.13301 1.12801
## Proportion of Variance 0.00134 0.00134 0.00133 0.00131 0.0013 0.00128 0.00127
## Cumulative Proportion  0.81153 0.81287 0.81420 0.81551 0.8168 0.81810 0.81937
##                          PC106   PC107   PC108   PC109   PC110   PC111   PC112
## Standard deviation     1.12537 1.11758 1.11617 1.10742 1.10258 1.09933 1.09065
## Proportion of Variance 0.00127 0.00125 0.00125 0.00123 0.00122 0.00121 0.00119
## Cumulative Proportion  0.82064 0.82189 0.82313 0.82436 0.82557 0.82678 0.82797
##                          PC113   PC114   PC115   PC116   PC117   PC118   PC119
## Standard deviation     1.08195 1.07984 1.07575 1.07229 1.06528 1.06176 1.06050
## Proportion of Variance 0.00117 0.00117 0.00116 0.00115 0.00113 0.00113 0.00112
## Cumulative Proportion  0.82914 0.83031 0.83147 0.83262 0.83375 0.83488 0.83600
##                          PC120   PC121  PC122   PC123   PC124   PC125   PC126
## Standard deviation     1.05610 1.05215 1.0488 1.04474 1.03986 1.03453 1.02609
## Proportion of Variance 0.00112 0.00111 0.0011 0.00109 0.00108 0.00107 0.00105
## Cumulative Proportion  0.83712 0.83823 0.8393 0.84042 0.84150 0.84257 0.84362
##                          PC127   PC128   PC129   PC130   PC131   PC132   PC133
## Standard deviation     1.02479 1.01855 1.01672 1.01460 1.00836 1.00741 1.00262
## Proportion of Variance 0.00105 0.00104 0.00103 0.00103 0.00102 0.00101 0.00101
## Cumulative Proportion  0.84467 0.84571 0.84674 0.84777 0.84879 0.84980 0.85081
##                         PC134   PC135   PC136   PC137   PC138   PC139   PC140
## Standard deviation     1.0000 0.99635 0.99350 0.98770 0.98300 0.97443 0.97030
## Proportion of Variance 0.0010 0.00099 0.00099 0.00098 0.00097 0.00095 0.00094
## Cumulative Proportion  0.8518 0.85280 0.85379 0.85476 0.85573 0.85668 0.85762
##                          PC141   PC142   PC143   PC144   PC145  PC146  PC147
```

```
## Standard deviation     0.96797 0.96394 0.96107 0.95774 0.95203 0.9508 0.9464
## Proportion of Variance 0.00094 0.00093 0.00092 0.00092 0.00091 0.0009 0.0009
## Cumulative Proportion  0.85856 0.85949 0.86041 0.86133 0.86224 0.8631 0.8640
##                           PC148   PC149   PC150   PC151   PC152   PC153   PC154
## Standard deviation     0.94248 0.93951 0.93677 0.93369 0.92969 0.92552 0.92430
## Proportion of Variance 0.00089 0.00088 0.00088 0.00087 0.00086 0.00086 0.00085
## Cumulative Proportion  0.86492 0.86581 0.86668 0.86756 0.86842 0.86928 0.87013
##                           PC155   PC156   PC157   PC158   PC159   PC160   PC161
## Standard deviation     0.92032 0.91543 0.91283 0.91173 0.91092 0.90622 0.90074
## Proportion of Variance 0.00085 0.00084 0.00083 0.00083 0.00083 0.00082 0.00081
## Cumulative Proportion  0.87098 0.87182 0.87265 0.87348 0.87431 0.87513 0.87594
##                           PC162  PC163   PC164   PC165   PC166   PC167   PC168
## Standard deviation     0.89871 0.8922 0.89088 0.88976 0.88536 0.88174 0.87772
## Proportion of Variance 0.00081 0.0008 0.00079 0.00079 0.00078 0.00078 0.00077
## Cumulative Proportion  0.87675 0.8776 0.87834 0.87913 0.87992 0.88069 0.88146
##                           PC169   PC170   PC171   PC172   PC173   PC174   PC175
## Standard deviation     0.87530 0.87173 0.87010 0.86721 0.86347 0.86104 0.86036
## Proportion of Variance 0.00077 0.00076 0.00076 0.00075 0.00075 0.00074 0.00074
## Cumulative Proportion  0.88223 0.88299 0.88375 0.88450 0.88524 0.88599 0.88673
##                           PC176   PC177   PC178   PC179   PC180   PC181  PC182
## Standard deviation     0.85552 0.85152 0.84857 0.84631 0.84476 0.83981 0.8385
## Proportion of Variance 0.00073 0.00073 0.00072 0.00072 0.00071 0.00071 0.0007
## Cumulative Proportion  0.88746 0.88818 0.88890 0.88962 0.89033 0.89104 0.8917
##                           PC183   PC184   PC185   PC186   PC187   PC188   PC189
## Standard deviation     0.83311 0.83168 0.83062 0.82503 0.82316 0.82082 0.81806
## Proportion of Variance 0.00069 0.00069 0.00069 0.00068 0.00068 0.00067 0.00067
## Cumulative Proportion  0.89243 0.89313 0.89382 0.89450 0.89517 0.89585 0.89652
##                           PC190   PC191   PC192   PC193   PC194   PC195   PC196
## Standard deviation     0.81591 0.81366 0.81050 0.80671 0.80527 0.80405 0.80128
## Proportion of Variance 0.00067 0.00066 0.00066 0.00065 0.00065 0.00065 0.00064
## Cumulative Proportion  0.89718 0.89785 0.89850 0.89915 0.89980 0.90045 0.90109
##                           PC197   PC198   PC199   PC200   PC201   PC202   PC203
## Standard deviation     0.79829 0.79510 0.79196 0.79116 0.78834 0.78560 0.78223
## Proportion of Variance 0.00064 0.00063 0.00063 0.00063 0.00062 0.00062 0.00061
## Cumulative Proportion  0.90173 0.90236 0.90299 0.90361 0.90423 0.90485 0.90546
##                           PC204   PC205  PC206  PC207   PC208   PC209   PC210
## Standard deviation     0.78021 0.77856 0.7753 0.7739 0.77071 0.76953 0.76656
## Proportion of Variance 0.00061 0.00061 0.0006 0.0006 0.00059 0.00059 0.00059
## Cumulative Proportion  0.90607 0.90668 0.9073 0.9079 0.90847 0.90906 0.90965
##                           PC211   PC212   PC213   PC214   PC215   PC216   PC217
## Standard deviation     0.76339 0.76006 0.75634 0.75560 0.75309 0.75167 0.74758
## Proportion of Variance 0.00058 0.00058 0.00057 0.00057 0.00057 0.00057 0.00056
## Cumulative Proportion  0.91023 0.91081 0.91138 0.91196 0.91252 0.91309 0.91365
##                           PC218   PC219   PC220   PC221   PC222   PC223   PC224
## Standard deviation     0.74670 0.74345 0.74168 0.73948 0.73708 0.73508 0.73291
## Proportion of Variance 0.00056 0.00055 0.00055 0.00055 0.00054 0.00054 0.00054
## Cumulative Proportion  0.91420 0.91476 0.91531 0.91585 0.91640 0.91694 0.91747
##                           PC225   PC226   PC227   PC228   PC229   PC230   PC231
```

```
## Standard deviation     0.72967 0.72946 0.72568 0.72397 0.72069 0.71979 0.71679
## Proportion of Variance 0.00053 0.00053 0.00053 0.00052 0.00052 0.00052 0.00051
## Cumulative Proportion  0.91801 0.91854 0.91907 0.91959 0.92011 0.92063 0.92114
##                           PC232   PC233   PC234  PC235  PC236  PC237  PC238
## Standard deviation     0.71507 0.71347 0.71122 0.7105 0.7073 0.7052 0.7039
## Proportion of Variance 0.00051 0.00051 0.00051 0.0005 0.0005 0.0005 0.0005
## Cumulative Proportion  0.92165 0.92216 0.92267 0.9232 0.9237 0.9242 0.9247
##                           PC239   PC240   PC241   PC242   PC243   PC244   PC245
## Standard deviation     0.70208 0.70118 0.69632 0.69502 0.69107 0.68924 0.68750
## Proportion of Variance 0.00049 0.00049 0.00048 0.00048 0.00048 0.00048 0.00047
## Cumulative Proportion  0.92516 0.92565 0.92613 0.92662 0.92710 0.92757 0.92804
##                           PC246   PC247   PC248   PC249   PC250   PC251   PC252
## Standard deviation     0.68403 0.68188 0.67985 0.67595 0.67520 0.67370 0.67317
## Proportion of Variance 0.00047 0.00046 0.00046 0.00046 0.00046 0.00045 0.00045
## Cumulative Proportion  0.92851 0.92898 0.92944 0.92989 0.93035 0.93080 0.93126
##                           PC253   PC254   PC255   PC256   PC257   PC258   PC259
## Standard deviation     0.67125 0.66933 0.66809 0.66583 0.66287 0.65983 0.65830
## Proportion of Variance 0.00045 0.00045 0.00045 0.00044 0.00044 0.00044 0.00043
## Cumulative Proportion  0.93171 0.93216 0.93260 0.93305 0.93349 0.93392 0.93435
##                           PC260   PC261   PC262   PC263   PC264   PC265   PC266
## Standard deviation     0.65658 0.65384 0.65294 0.65005 0.64790 0.64688 0.64482
## Proportion of Variance 0.00043 0.00043 0.00043 0.00042 0.00042 0.00042 0.00042
## Cumulative Proportion  0.93479 0.93521 0.93564 0.93606 0.93648 0.93690 0.93732
##                           PC267   PC268   PC269   PC270  PC271  PC272  PC273
## Standard deviation     0.64191 0.63907 0.63786 0.63657 0.6339 0.6317 0.6307
## Proportion of Variance 0.00041 0.00041 0.00041 0.00041 0.0004 0.0004 0.0004
## Cumulative Proportion  0.93773 0.93814 0.93854 0.93895 0.9394 0.9397 0.9402
##                           PC274   PC275   PC276   PC277   PC278   PC279   PC280
## Standard deviation     0.62792 0.62660 0.62432 0.62236 0.62069 0.61744 0.61509
## Proportion of Variance 0.00039 0.00039 0.00039 0.00039 0.00039 0.00038 0.00038
## Cumulative Proportion  0.94054 0.94093 0.94132 0.94171 0.94210 0.94248 0.94286
##                           PC281   PC282   PC283   PC284   PC285   PC286   PC287
## Standard deviation     0.61435 0.61309 0.61062 0.60944 0.60705 0.60576 0.60468
## Proportion of Variance 0.00038 0.00038 0.00037 0.00037 0.00037 0.00037 0.00037
## Cumulative Proportion  0.94323 0.94361 0.94398 0.94435 0.94472 0.94509 0.94545
##                           PC288   PC289   PC290   PC291   PC292   PC293   PC294
## Standard deviation     0.60320 0.60095 0.59947 0.59746 0.59632 0.59321 0.59083
## Proportion of Variance 0.00036 0.00036 0.00036 0.00036 0.00036 0.00035 0.00035
## Cumulative Proportion  0.94582 0.94618 0.94654 0.94690 0.94725 0.94760 0.94795
##                           PC295   PC296   PC297   PC298   PC299   PC300   PC301
## Standard deviation     0.58973 0.58869 0.58745 0.58590 0.58456 0.58229 0.58172
## Proportion of Variance 0.00035 0.00035 0.00035 0.00034 0.00034 0.00034 0.00034
## Cumulative Proportion  0.94830 0.94865 0.94899 0.94933 0.94968 0.95002 0.95035
##                           PC302   PC303   PC304   PC305   PC306   PC307   PC308
## Standard deviation     0.57976 0.57860 0.57561 0.57238 0.57094 0.57088 0.56908
## Proportion of Variance 0.00034 0.00033 0.00033 0.00033 0.00033 0.00033 0.00032
## Cumulative Proportion  0.95069 0.95102 0.95136 0.95168 0.95201 0.95234 0.95266
##                           PC309   PC310   PC311   PC312   PC313   PC314   PC315
```

```
## Standard deviation     0.56650 0.56595 0.56345 0.56202 0.55856 0.55638 0.55550
## Proportion of Variance 0.00032 0.00032 0.00032 0.00032 0.00031 0.00031 0.00031
## Cumulative Proportion  0.95298 0.95330 0.95362 0.95393 0.95425 0.95456 0.95486
##                           PC316   PC317  PC318  PC319  PC320  PC321  PC322
## Standard deviation     0.55503 0.55401 0.5503 0.5495 0.5471 0.5460 0.5446
## Proportion of Variance 0.00031 0.00031 0.0003 0.0003 0.0003 0.0003 0.0003
## Cumulative Proportion  0.95517 0.95548 0.9558 0.9561 0.9564 0.9567 0.9570
##                           PC323   PC324   PC325   PC326   PC327   PC328   PC329
## Standard deviation     0.54247 0.53963 0.53889 0.53772 0.53511 0.53476 0.53184
## Proportion of Variance 0.00029 0.00029 0.00029 0.00029 0.00029 0.00029 0.00028
## Cumulative Proportion  0.95727 0.95756 0.95785 0.95814 0.95843 0.95872 0.95900
##                           PC330   PC331   PC332   PC333   PC334   PC335   PC336
## Standard deviation     0.53083 0.52939 0.52837 0.52733 0.52663 0.52433 0.52241
## Proportion of Variance 0.00028 0.00028 0.00028 0.00028 0.00028 0.00027 0.00027
## Cumulative Proportion  0.95928 0.95956 0.95984 0.96012 0.96039 0.96067 0.96094
##                           PC337   PC338   PC339   PC340   PC341   PC342   PC343
## Standard deviation     0.52036 0.51942 0.51794 0.51781 0.51380 0.51317 0.51134
## Proportion of Variance 0.00027 0.00027 0.00027 0.00027 0.00026 0.00026 0.00026
## Cumulative Proportion  0.96121 0.96148 0.96175 0.96202 0.96228 0.96255 0.96281
##                           PC344   PC345   PC346   PC347   PC348   PC349   PC350
## Standard deviation     0.50959 0.50891 0.50788 0.50546 0.50389 0.50247 0.50183
## Proportion of Variance 0.00026 0.00026 0.00026 0.00026 0.00025 0.00025 0.00025
## Cumulative Proportion  0.96307 0.96333 0.96359 0.96384 0.96409 0.96435 0.96460
##                           PC351   PC352   PC353   PC354   PC355   PC356   PC357
## Standard deviation     0.50023 0.49888 0.49769 0.49499 0.49322 0.49227 0.49126
## Proportion of Variance 0.00025 0.00025 0.00025 0.00025 0.00024 0.00024 0.00024
## Cumulative Proportion  0.96485 0.96510 0.96535 0.96559 0.96583 0.96608 0.96632
##                           PC358   PC359   PC360   PC361   PC362   PC363   PC364
## Standard deviation     0.48884 0.48738 0.48557 0.48424 0.48271 0.48144 0.48030
## Proportion of Variance 0.00024 0.00024 0.00024 0.00023 0.00023 0.00023 0.00023
## Cumulative Proportion  0.96656 0.96679 0.96703 0.96726 0.96750 0.96773 0.96796
##                           PC365   PC366   PC367   PC368   PC369   PC370   PC371
## Standard deviation     0.47803 0.47753 0.47676 0.47487 0.47378 0.47192 0.47131
## Proportion of Variance 0.00023 0.00023 0.00023 0.00023 0.00022 0.00022 0.00022
## Cumulative Proportion  0.96819 0.96842 0.96864 0.96887 0.96909 0.96932 0.96954
##                           PC372   PC373   PC374   PC375   PC376   PC377   PC378
## Standard deviation     0.46926 0.46846 0.46750 0.46588 0.46410 0.46245 0.46114
## Proportion of Variance 0.00022 0.00022 0.00022 0.00022 0.00022 0.00021 0.00021
## Cumulative Proportion  0.96976 0.96998 0.97020 0.97041 0.97063 0.97084 0.97106
##                           PC379   PC380   PC381   PC382   PC383   PC384  PC385
## Standard deviation     0.45991 0.45883 0.45729 0.45680 0.45613 0.45334 0.4521
## Proportion of Variance 0.00021 0.00021 0.00021 0.00021 0.00021 0.00021 0.0002
## Cumulative Proportion  0.97127 0.97148 0.97169 0.97190 0.97210 0.97231 0.9725
##                          PC386  PC387  PC388  PC389  PC390  PC391  PC392   PC393
## Standard deviation     0.4516 0.4503 0.4489 0.4477 0.4456 0.4442 0.4433 0.44143
## Proportion of Variance 0.0002 0.0002 0.0002 0.0002 0.0002 0.0002 0.0002 0.00019
## Cumulative Proportion  0.9727 0.9729 0.9731 0.9733 0.9735 0.9737 0.9739 0.97411
##                           PC394   PC395   PC396   PC397   PC398   PC399   PC400
```

```
## Standard deviation     0.44040 0.43936 0.43658 0.43605 0.43350 0.43213 0.43077
## Proportion of Variance 0.00019 0.00019 0.00019 0.00019 0.00019 0.00019 0.00019
## Cumulative Proportion  0.97430 0.97450 0.97469 0.97488 0.97506 0.97525 0.97544
##                           PC401   PC402   PC403   PC404   PC405   PC406   PC407
## Standard deviation     0.43001 0.42844 0.42755 0.42635 0.42599 0.42305 0.42176
## Proportion of Variance 0.00018 0.00018 0.00018 0.00018 0.00018 0.00018 0.00018
## Cumulative Proportion  0.97562 0.97581 0.97599 0.97617 0.97635 0.97653 0.97671
##                           PC408   PC409   PC410   PC411   PC412   PC413   PC414
## Standard deviation     0.42091 0.41948 0.41731 0.41601 0.41503 0.41322 0.41191
## Proportion of Variance 0.00018 0.00018 0.00017 0.00017 0.00017 0.00017 0.00017
## Cumulative Proportion  0.97689 0.97706 0.97724 0.97741 0.97758 0.97775 0.97792
##                           PC415   PC416   PC417   PC418   PC419   PC420   PC421
## Standard deviation     0.41122 0.41031 0.40840 0.40772 0.40645 0.40582 0.40493
## Proportion of Variance 0.00017 0.00017 0.00017 0.00017 0.00017 0.00016 0.00016
## Cumulative Proportion  0.97809 0.97826 0.97843 0.97859 0.97876 0.97892 0.97909
##                           PC422   PC423   PC424   PC425   PC426   PC427   PC428
## Standard deviation     0.40410 0.40225 0.40016 0.39924 0.39810 0.39701 0.39497
## Proportion of Variance 0.00016 0.00016 0.00016 0.00016 0.00016 0.00016 0.00016
## Cumulative Proportion  0.97925 0.97941 0.97957 0.97973 0.97989 0.98005 0.98020
##                           PC429   PC430   PC431   PC432   PC433   PC434   PC435
## Standard deviation     0.39414 0.39361 0.39233 0.39042 0.38943 0.38923 0.38743
## Proportion of Variance 0.00016 0.00015 0.00015 0.00015 0.00015 0.00015 0.00015
## Cumulative Proportion  0.98036 0.98051 0.98067 0.98082 0.98097 0.98112 0.98127
##                           PC436   PC437   PC438   PC439   PC440   PC441   PC442
## Standard deviation     0.38539 0.38425 0.38310 0.38216 0.38002 0.37910 0.37856
## Proportion of Variance 0.00015 0.00015 0.00015 0.00015 0.00014 0.00014 0.00014
## Cumulative Proportion  0.98142 0.98157 0.98172 0.98186 0.98201 0.98215 0.98229
##                           PC443   PC444   PC445   PC446   PC447   PC448   PC449
## Standard deviation     0.37751 0.37576 0.37394 0.37304 0.37140 0.37029 0.36971
## Proportion of Variance 0.00014 0.00014 0.00014 0.00014 0.00014 0.00014 0.00014
## Cumulative Proportion  0.98244 0.98258 0.98272 0.98286 0.98299 0.98313 0.98327
##                           PC450   PC451   PC452   PC453   PC454   PC455   PC456
## Standard deviation     0.36837 0.36791 0.36597 0.36470 0.36349 0.36290 0.36198
## Proportion of Variance 0.00014 0.00014 0.00013 0.00013 0.00013 0.00013 0.00013
## Cumulative Proportion  0.98340 0.98354 0.98367 0.98381 0.98394 0.98407 0.98420
##                           PC457   PC458   PC459   PC460   PC461   PC462   PC463
## Standard deviation     0.36059 0.35982 0.35811 0.35652 0.35530 0.35481 0.35292
## Proportion of Variance 0.00013 0.00013 0.00013 0.00013 0.00013 0.00013 0.00012
## Cumulative Proportion  0.98433 0.98446 0.98459 0.98472 0.98484 0.98497 0.98509
##                           PC464   PC465   PC466   PC467   PC468   PC469   PC470
## Standard deviation     0.35191 0.35094 0.34945 0.34919 0.34882 0.34760 0.34668
## Proportion of Variance 0.00012 0.00012 0.00012 0.00012 0.00012 0.00012 0.00012
## Cumulative Proportion  0.98522 0.98534 0.98546 0.98558 0.98570 0.98583 0.98595
##                           PC471   PC472   PC473   PC474   PC475   PC476   PC477
## Standard deviation     0.34442 0.34389 0.34271 0.34157 0.34148 0.33981 0.33784
## Proportion of Variance 0.00012 0.00012 0.00012 0.00012 0.00012 0.00012 0.00011
## Cumulative Proportion  0.98606 0.98618 0.98630 0.98642 0.98653 0.98665 0.98676
##                           PC478   PC479   PC480   PC481   PC482   PC483   PC484
```

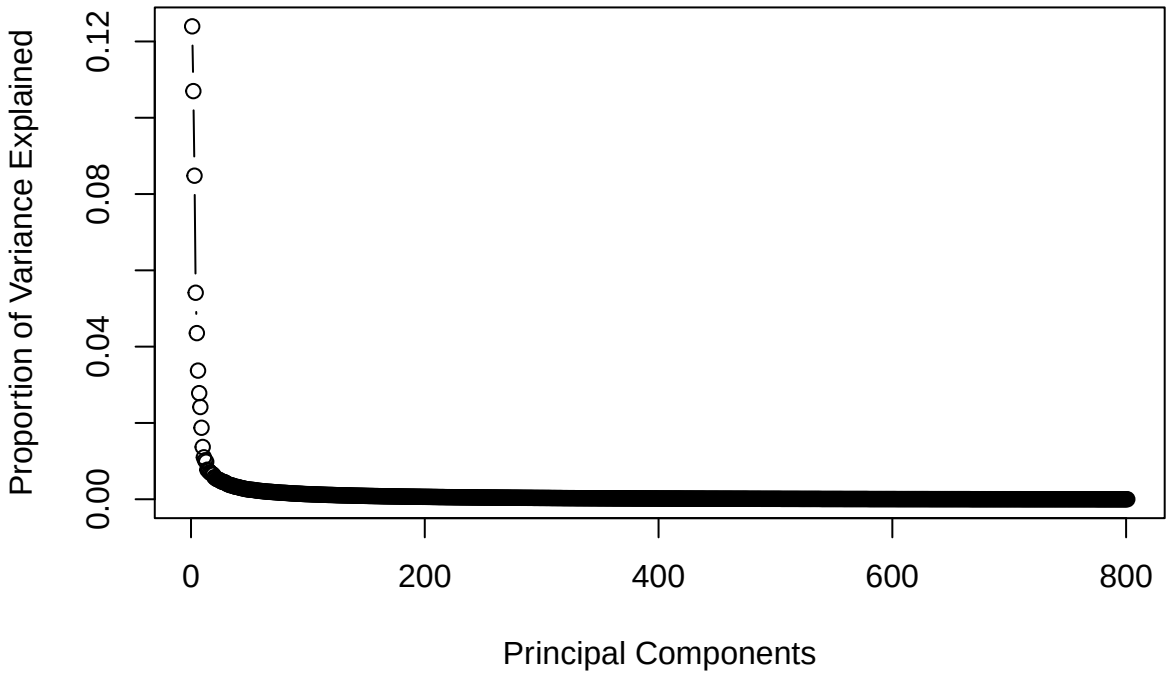
```
## Standard deviation     0.33687 0.33530 0.33397 0.33271 0.33121 0.33031 0.32827
## Proportion of Variance 0.00011 0.00011 0.00011 0.00011 0.00011 0.00011 0.00011
## Cumulative Proportion  0.98688 0.98699 0.98710 0.98721 0.98732 0.98743 0.98754
##                           PC485   PC486   PC487  PC488  PC489  PC490  PC491
## Standard deviation     0.32758 0.32676 0.32504 0.3237 0.3229 0.3218 0.3211
## Proportion of Variance 0.00011 0.00011 0.00011 0.0001 0.0001 0.0001 0.0001
## Cumulative Proportion  0.98764 0.98775 0.98786 0.9880 0.9881 0.9882 0.9883
##                           PC492  PC493  PC494  PC495  PC496  PC497  PC498  PC499
## Standard deviation     0.3198 0.3192 0.3183 0.3177 0.3170 0.3158 0.3154 0.3138
## Proportion of Variance 0.0001 0.0001 0.0001 0.0001 0.0001 0.0001 0.0001 0.0001
## Cumulative Proportion  0.9884 0.9885 0.9886 0.9887 0.9888 0.9889 0.9890 0.9891
##                           PC500  PC501  PC502  PC503  PC504   PC505   PC506
## Standard deviation     0.3124 0.3117 0.3108 0.3084 0.3083 0.30692 0.30643
## Proportion of Variance 0.0001 0.0001 0.0001 0.0001 0.0001 0.00009 0.00009
## Cumulative Proportion  0.9892 0.9893 0.9894 0.9895 0.9896 0.98965 0.98975
##                           PC507   PC508   PC509   PC510   PC511   PC512   PC513
## Standard deviation     0.30538 0.30462 0.30365 0.30267 0.30258 0.30026 0.29933
## Proportion of Variance 0.00009 0.00009 0.00009 0.00009 0.00009 0.00009 0.00009
## Cumulative Proportion  0.98984 0.98993 0.99003 0.99012 0.99021 0.99030 0.99039
##                           PC514   PC515   PC516   PC517   PC518   PC519   PC520
## Standard deviation     0.29863 0.29710 0.29597 0.29475 0.29381 0.29257 0.29158
## Proportion of Variance 0.00009 0.00009 0.00009 0.00009 0.00009 0.00009 0.00009
## Cumulative Proportion  0.99048 0.99057 0.99065 0.99074 0.99083 0.99091 0.99100
##                           PC521   PC522   PC523   PC524   PC525   PC526   PC527
## Standard deviation     0.29069 0.29000 0.28850 0.28777 0.28692 0.28613 0.28457
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00008 0.00008 0.00008
## Cumulative Proportion  0.99108 0.99117 0.99125 0.99133 0.99141 0.99150 0.99158
##                           PC528   PC529   PC530   PC531   PC532   PC533   PC534
## Standard deviation     0.28402 0.28217 0.28183 0.28130 0.28014 0.27915 0.27742
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00008 0.00008 0.00008
## Cumulative Proportion  0.99166 0.99174 0.99182 0.99190 0.99197 0.99205 0.99213
##                           PC535   PC536   PC537   PC538   PC539   PC540   PC541
## Standard deviation     0.27713 0.27597 0.27496 0.27407 0.27323 0.27189 0.27183
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99221 0.99228 0.99236 0.99243 0.99251 0.99258 0.99266
##                           PC542   PC543   PC544   PC545   PC546   PC547   PC548
## Standard deviation     0.27001 0.26882 0.26729 0.26656 0.26505 0.26470 0.26369
## Proportion of Variance 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99273 0.99280 0.99287 0.99294 0.99301 0.99308 0.99315
##                           PC549   PC550   PC551   PC552   PC553   PC554   PC555
## Standard deviation     0.26273 0.26216 0.26121 0.26026 0.25992 0.25874 0.25710
## Proportion of Variance 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99322 0.99329 0.99336 0.99343 0.99349 0.99356 0.99363
##                           PC556   PC557   PC558   PC559   PC560   PC561   PC562
## Standard deviation     0.25671 0.25644 0.25479 0.25467 0.25292 0.25191 0.25108
## Proportion of Variance 0.00007 0.00007 0.00006 0.00006 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99369 0.99376 0.99382 0.99389 0.99395 0.99402 0.99408
##                           PC563   PC564   PC565   PC566   PC567   PC568   PC569
```

```
## Standard deviation     0.25065 0.24915 0.24820 0.24746 0.24720 0.24554 0.24394
## Proportion of Variance 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99414 0.99420 0.99427 0.99433 0.99439 0.99445 0.99451
##                           PC570   PC571   PC572   PC573   PC574   PC575   PC576
## Standard deviation     0.24376 0.24255 0.24052 0.24015 0.23905 0.23717 0.23677
## Proportion of Variance 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99457 0.99463 0.99468 0.99474 0.99480 0.99485 0.99491
##                           PC577   PC578   PC579   PC580   PC581   PC582   PC583
## Standard deviation     0.23632 0.23549 0.23465 0.23362 0.23242 0.23182 0.23031
## Proportion of Variance 0.00006 0.00006 0.00006 0.00005 0.00005 0.00005 0.00005
## Cumulative Proportion  0.99497 0.99502 0.99508 0.99513 0.99519 0.99524 0.99529
##                           PC584   PC585   PC586   PC587   PC588   PC589   PC590
## Standard deviation     0.22927 0.22833 0.22777 0.22600 0.22520 0.22451 0.22375
## Proportion of Variance 0.00005 0.00005 0.00005 0.00005 0.00005 0.00005 0.00005
## Cumulative Proportion  0.99535 0.99540 0.99545 0.99550 0.99555 0.99560 0.99565
##                           PC591   PC592   PC593   PC594   PC595   PC596   PC597
## Standard deviation     0.22345 0.22232 0.22173 0.22121 0.21937 0.21926 0.21798
## Proportion of Variance 0.00005 0.00005 0.00005 0.00005 0.00005 0.00005 0.00005
## Cumulative Proportion  0.99570 0.99575 0.99580 0.99585 0.99590 0.99595 0.99599
##                           PC598   PC599   PC600   PC601   PC602   PC603   PC604
## Standard deviation     0.21701 0.21643 0.21494 0.21379 0.21332 0.21215 0.21192
## Proportion of Variance 0.00005 0.00005 0.00005 0.00005 0.00005 0.00005 0.00004
## Cumulative Proportion  0.99604 0.99609 0.99613 0.99618 0.99622 0.99627 0.99631
##                           PC605   PC606   PC607   PC608   PC609   PC610   PC611
## Standard deviation     0.21172 0.20907 0.20793 0.20691 0.20664 0.20611 0.20542
## Proportion of Variance 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004
## Cumulative Proportion  0.99636 0.99640 0.99645 0.99649 0.99653 0.99657 0.99662
##                           PC612   PC613   PC614   PC615   PC616   PC617   PC618
## Standard deviation     0.20453 0.20383 0.20271 0.20257 0.20150 0.20020 0.19978
## Proportion of Variance 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004
## Cumulative Proportion  0.99666 0.99670 0.99674 0.99678 0.99682 0.99686 0.99690
##                           PC619   PC620   PC621   PC622   PC623   PC624   PC625
## Standard deviation     0.19859 0.19771 0.19734 0.19538 0.19478 0.19360 0.19270
## Proportion of Variance 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004
## Cumulative Proportion  0.99694 0.99698 0.99702 0.99706 0.99710 0.99713 0.99717
##                           PC626   PC627   PC628   PC629   PC630   PC631   PC632
## Standard deviation     0.19215 0.19162 0.19064 0.18845 0.18781 0.18760 0.18599
## Proportion of Variance 0.00004 0.00004 0.00004 0.00004 0.00004 0.00004 0.00003
## Cumulative Proportion  0.99721 0.99724 0.99728 0.99732 0.99735 0.99739 0.99742
##                           PC633   PC634   PC635   PC636   PC637   PC638   PC639
## Standard deviation     0.18515 0.18446 0.18361 0.18318 0.18180 0.18131 0.18065
## Proportion of Variance 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003
## Cumulative Proportion  0.99746 0.99749 0.99752 0.99756 0.99759 0.99762 0.99765
##                           PC640   PC641   PC642   PC643   PC644   PC645   PC646
## Standard deviation     0.17943 0.17858 0.17836 0.17743 0.17643 0.17592 0.17501
## Proportion of Variance 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003
## Cumulative Proportion  0.99769 0.99772 0.99775 0.99778 0.99781 0.99784 0.99788
##                           PC647   PC648   PC649   PC650   PC651   PC652   PC653
```

```
## Standard deviation     0.17389 0.17312 0.17209 0.17054 0.16993 0.16942 0.16890
## Proportion of Variance 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003
## Cumulative Proportion  0.99791 0.99794 0.99796 0.99799 0.99802 0.99805 0.99808
##                           PC654   PC655   PC656   PC657   PC658   PC659   PC660
## Standard deviation     0.16799 0.16748 0.16608 0.16557 0.16476 0.16435 0.16348
## Proportion of Variance 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003
## Cumulative Proportion  0.99811 0.99814 0.99816 0.99819 0.99822 0.99825 0.99827
##                           PC661   PC662   PC663   PC664   PC665   PC666   PC667
## Standard deviation     0.16246 0.16166 0.16108 0.16074 0.15890 0.15880 0.15760
## Proportion of Variance 0.00003 0.00003 0.00003 0.00003 0.00003 0.00003 0.00002
## Cumulative Proportion  0.99830 0.99832 0.99835 0.99838 0.99840 0.99843 0.99845
##                           PC668   PC669   PC670   PC671   PC672   PC673   PC674
## Standard deviation     0.15697 0.15664 0.15589 0.15528 0.15474 0.15419 0.15326
## Proportion of Variance 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99848 0.99850 0.99853 0.99855 0.99857 0.99860 0.99862
##                           PC675   PC676   PC677   PC678   PC679   PC680   PC681
## Standard deviation     0.15145 0.14967 0.14911 0.14850 0.14797 0.14676 0.14607
## Proportion of Variance 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99864 0.99867 0.99869 0.99871 0.99873 0.99875 0.99877
##                           PC682   PC683   PC684   PC685   PC686   PC687   PC688
## Standard deviation     0.14561 0.14516 0.14402 0.14299 0.14153 0.14081 0.14067
## Proportion of Variance 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99880 0.99882 0.99884 0.99886 0.99888 0.99890 0.99892
##                           PC689   PC690   PC691   PC692   PC693   PC694   PC695
## Standard deviation     0.13919 0.13872 0.13764 0.13722 0.13651 0.13602 0.13464
## Proportion of Variance 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99894 0.99896 0.99898 0.99899 0.99901 0.99903 0.99905
##                           PC696   PC697   PC698   PC699   PC700   PC701   PC702
## Standard deviation     0.13416 0.13403 0.13272 0.13175 0.13038 0.12882 0.12817
## Proportion of Variance 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99907 0.99909 0.99910 0.99912 0.99914 0.99915 0.99917
##                           PC703   PC704   PC705   PC706   PC707   PC708   PC709
## Standard deviation     0.12803 0.12728 0.12641 0.12548 0.12486 0.12353 0.12340
## Proportion of Variance 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99919 0.99920 0.99922 0.99924 0.99925 0.99927 0.99928
##                           PC710   PC711   PC712   PC713   PC714   PC715   PC716
## Standard deviation     0.12289 0.12174 0.12082 0.12000 0.11875 0.11844 0.11776
## Proportion of Variance 0.00002 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99930 0.99931 0.99933 0.99934 0.99935 0.99937 0.99938
##                           PC717   PC718   PC719   PC720   PC721   PC722   PC723
## Standard deviation     0.11756 0.11710 0.11511 0.11409 0.11392 0.11259 0.11213
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99940 0.99941 0.99942 0.99944 0.99945 0.99946 0.99947
##                           PC724   PC725   PC726   PC727   PC728   PC729   PC730
## Standard deviation     0.11165 0.11083 0.11035 0.10879 0.10822 0.10697 0.10639
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99949 0.99950 0.99951 0.99952 0.99953 0.99955 0.99956
##                           PC731   PC732   PC733   PC734   PC735   PC736   PC737
```

```
## Standard deviation     0.10536 0.10432 0.10354 0.10277 0.10219 0.10175 0.10026
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99957 0.99958 0.99959 0.99960 0.99961 0.99962 0.99963
##                           PC738   PC739   PC740   PC741   PC742   PC743   PC744
## Standard deviation     0.09963 0.09893 0.09848 0.09790 0.09704 0.09662 0.09596
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99964 0.99965 0.99966 0.99967 0.99968 0.99969 0.99970
##                           PC745   PC746   PC747   PC748   PC749   PC750   PC751
## Standard deviation     0.09507 0.09441 0.09364 0.09216 0.09140 0.09103 0.09039
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99971 0.99972 0.99973 0.99973 0.99974 0.99975 0.99976
##                           PC752   PC753   PC754   PC755   PC756   PC757   PC758
## Standard deviation     0.08922 0.08920 0.08759 0.08725 0.08646 0.08525 0.08495
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99977 0.99977 0.99978 0.99979 0.99980 0.99980 0.99981
##                           PC759   PC760   PC761   PC762   PC763   PC764   PC765
## Standard deviation     0.08367 0.08325 0.08305 0.08171 0.08019 0.07961 0.07858
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99982 0.99983 0.99983 0.99984 0.99985 0.99985 0.99986
##                           PC766   PC767   PC768   PC769   PC770   PC771   PC772
## Standard deviation     0.07839 0.07761 0.07636 0.07538 0.07498 0.07413 0.07307
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99986 0.99987 0.99988 0.99988 0.99989 0.99989 0.99990
##                           PC773   PC774   PC775   PC776   PC777   PC778   PC779
## Standard deviation     0.07248 0.07204 0.07067 0.06995 0.06839 0.06816 0.06788
## Proportion of Variance 0.00001 0.00001 0.00000 0.00000 0.00000 0.00000 0.00000
## Cumulative Proportion  0.99990 0.99991 0.99991 0.99992 0.99992 0.99993 0.99993
##                           PC780   PC781   PC782   PC783   PC784   PC785   PC786
## Standard deviation     0.06654 0.06645 0.06544 0.06375 0.06304 0.06243 0.06075
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
## Cumulative Proportion  0.99994 0.99994 0.99995 0.99995 0.99995 0.99996 0.99996
##                           PC787   PC788  PC789   PC790  PC791   PC792   PC793
## Standard deviation     0.0604 0.05965 0.0584 0.05819 0.0543 0.05374 0.05298
## Proportion of Variance 0.0000 0.00000 0.0000 0.00000 0.0000 0.00000 0.00000
## Cumulative Proportion  1.0000 0.99997 1.0000 0.99998 1.0000 0.99998 0.99998
##                           PC794   PC795   PC796   PC797   PC798   PC799   PC800
## Standard deviation     0.05167 0.05071 0.04986 0.04874 0.04786 0.04503 0.04267
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
## Cumulative Proportion  0.99999 0.99999 0.99999 0.99999 1.00000 1.00000 1.00000
##                           PC801
## Standard deviation     4.944e-15
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```r
# Scree plot to visualize
plot(summary(pca_result)$importance[2, ], type = "b", xlab = "Principal Components",
    ylab = "Proportion of Variance Explained", main = "Variance Explained by Principal Componer
```

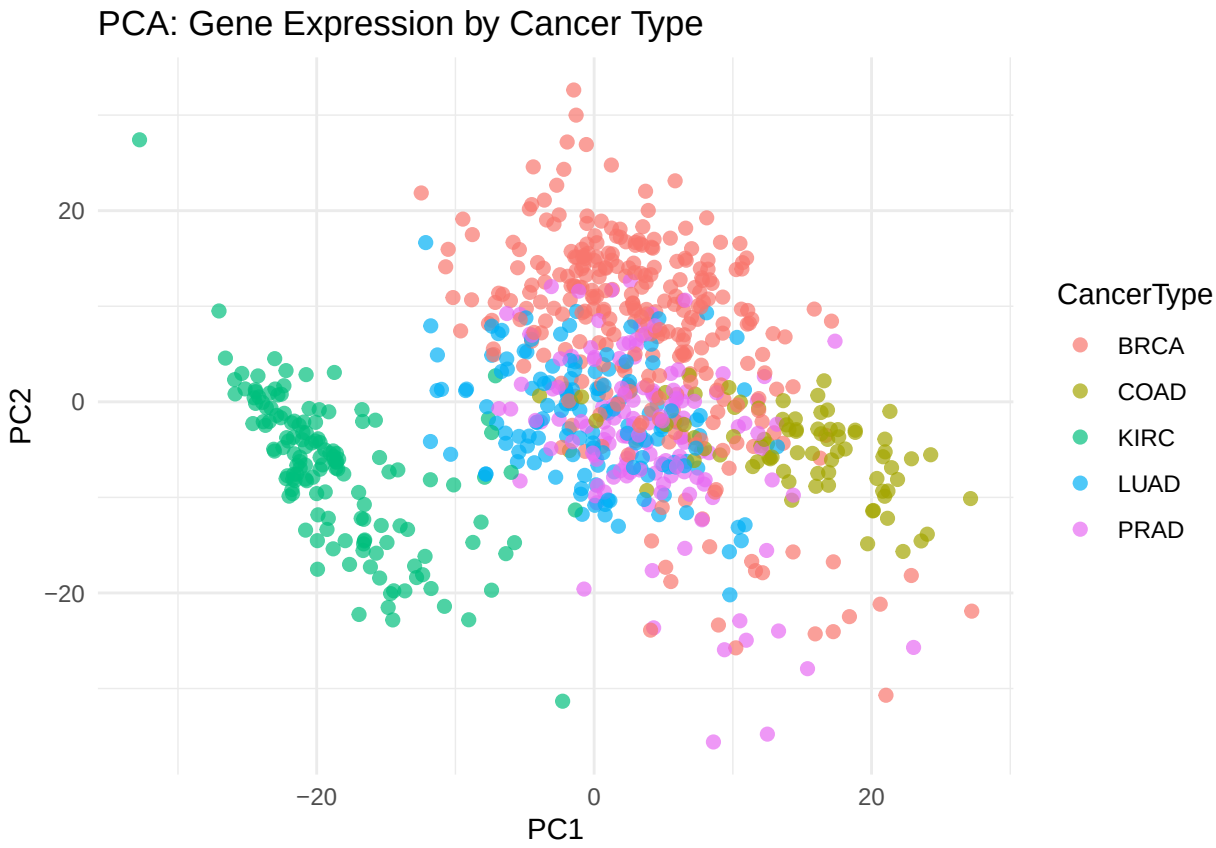


##

Interpretations The scree plot and PCA summary indicate that the first few principal components explain a substantial proportion of the total variance in the gene expression data, with the variance dropping off sharply after the initial components. This suggests that linear combinations of gene expression values, captured by these leading principal components, effectively summarize the main sources of variability in the dataset. Therefore, a small number of components can be used to represent the high-dimensional gene expression data in a lower-dimensional space without substantial information loss.

(1.b) Ideally, a type of cancer should have its "signature", i.e., a pattern in the gene expressions that is specific to this cancer type. From the "labels.csv", you will know which expression measurements belong to which cancer type. Identify the signature of each cancer type (if any) and visualize it. For this, you need to be creative and should try both PCA and Sparse PCA.

```
# PCA scores (first 2 PCs)
pca_scores <- as.data.frame(pca_result$x[, 1:2])
pca_scores$CancerType <- as.factor(labels[, 1])

# PCA Plot
library(ggplot2)
ggplot(pca_scores, aes(x = PC1, y = PC2, color = CancerType)) +
    geom_point(size = 2, alpha = 0.7) + labs(title = "PCA: Gene Expression by Cancer Type") +
    theme_minimal()
```

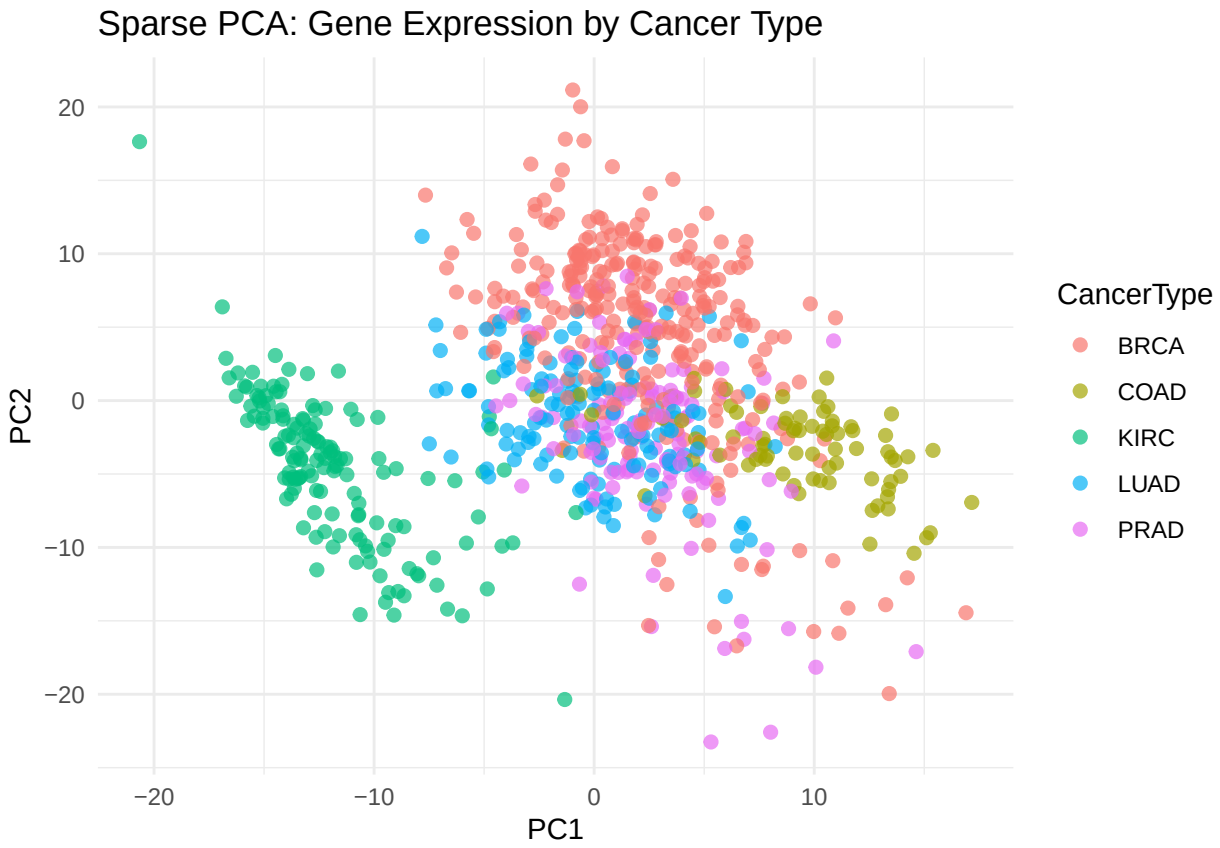



```
library(elasticnet)
```

```
## Loading required package: lars
```

```
## Loaded lars 1.3
```

```
spca_result <- spca(stdgexpProj2, K = 5, para = rep(50, 5))

# Compute scores manually
sparse_scores <- stdgexpProj2 %*% spca_result$loadings

# Convert to data frame and label
spca_scores <- as.data.frame(sparse_scores[, 1:2])
colnames(spca_scores)[1:2] <- c("PC1", "PC2")
spca_scores$CancerType <- as.factor(labels[, 1])

# Plot
ggplot(spca_scores, aes(x = PC1, y = PC2, color = CancerType)) +
    geom_point(size = 2, alpha = 0.7) + labs(title = "Sparse PCA: Gene Expression by Cancer Ty
    theme_minimal()
```



Interpretations

The PCA and Sparse PCA plots both show that gene expression patterns vary by cancer type, with noticeable clustering in the transformed space. In the PCA plot, KIRC samples are clearly separated from other types, while BRCA, COAD, LUAD, and PRAD show moderate overlap. In contrast, the Sparse PCA plot shows tighter and more distinct groupings, especially for BRCA and KIRC, suggesting that a smaller subset of genes effectively distinguishes the cancer types. These results support the presence of cancer-specific gene expression signatures, and demonstrate that Sparse PCA enhances interpretability by focusing on the most informative features.

(1.c) There are 5 cancer types. Would 5 principal components, obtained either from PCA or Sparse PCA, explain a dominant proportion of variability in the data set, and serve as the signatures of the 5 cancer types? Note that the same set of genes were measured for each cancer type.

```
# Proportion of variance explained by the first 5 PCA
# components
pve <- summary(pca_result)$importance[2, 1:5]
cumulative_pve <- cumsum(pve)
print(pve)
```

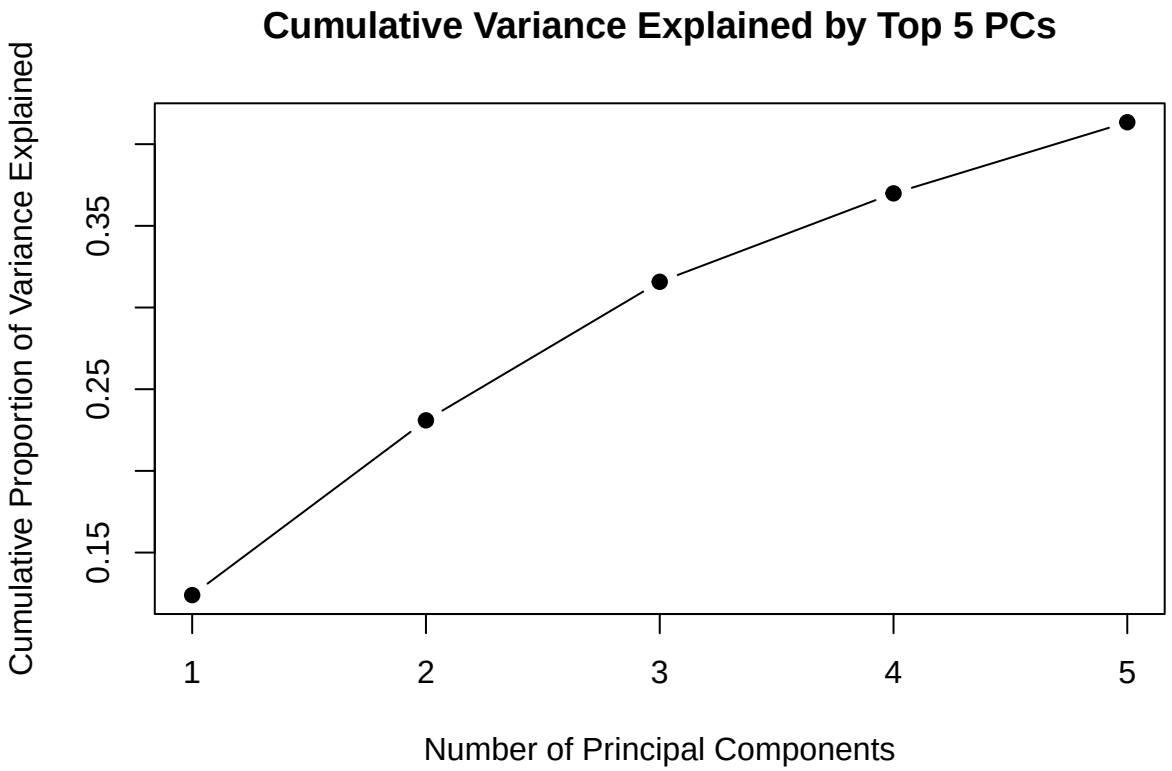
```
##     PC1     PC2     PC3     PC4     PC5
## 0.12396 0.10698 0.08481 0.05413 0.04355
```

```
print(cumulative_pve)
```

```
##     PC1     PC2     PC3     PC4     PC5
```

```
## 0.12396 0.23094 0.31575 0.36988 0.41343
```

```
plot(cumulative_pve, type = "b", pch = 19, xlab = "Number of Principal Components",
    ylab = "Cumulative Proportion of Variance Explained", main = "Cumulative Variance Explained
```



Interpretations

The first five principal components explain approximately 41.3% of the total variance in the gene expression data, as shown in the cumulative variance plot. While this captures a meaningful portion of the variability, it does not represent a dominant share of the total information in the dataset. Given the complexity and high dimensionality of gene expression data, more than five components are likely needed to fully capture cancer-specific patterns. Therefore, while the first five components may contribute to identifying cancer type signatures, they are not sufficient on their own to serve as complete representations of all five cancer types.

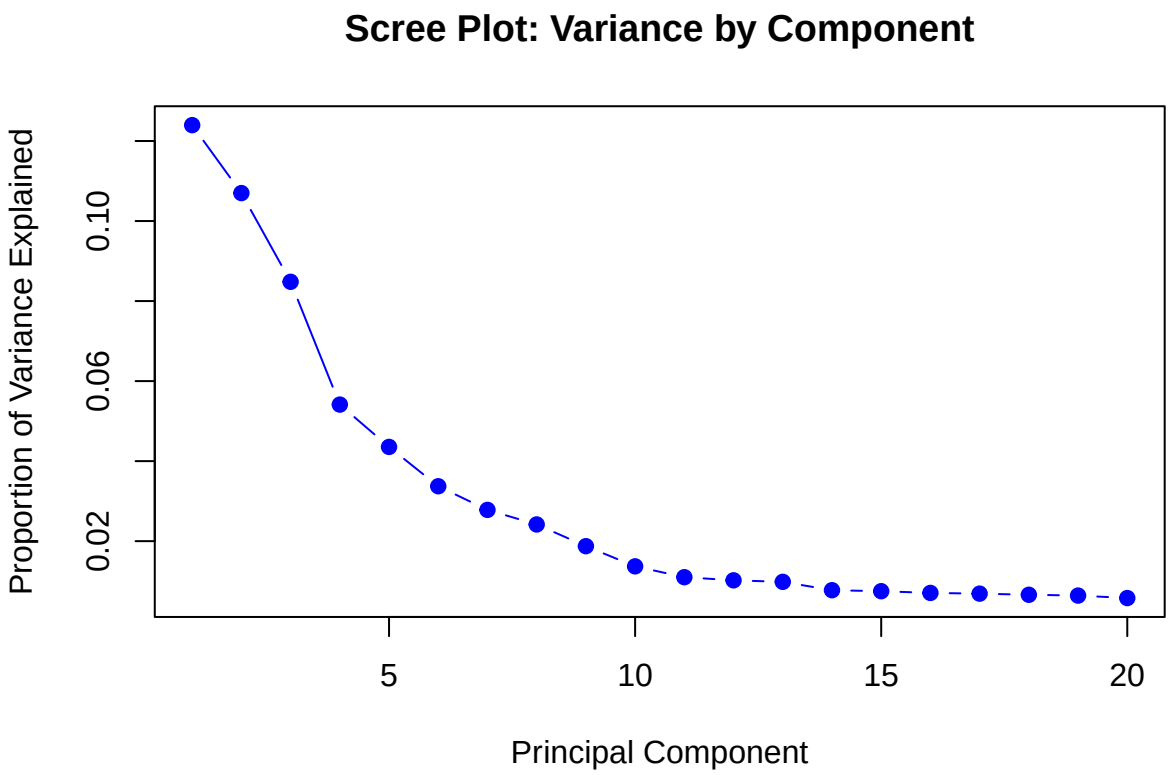
Identify patterns and low-dimensional structures

Please implement the following:

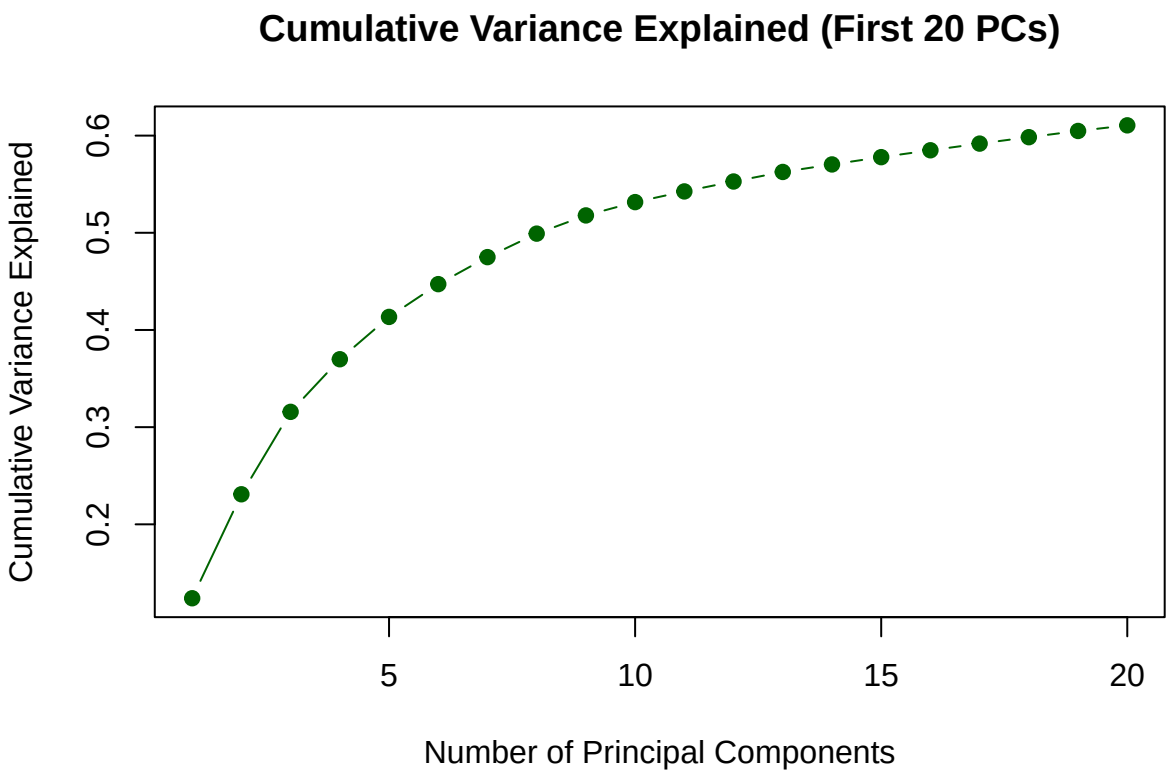
(2.a) Apply PCA, determine the number of principal components, provide visualizations of low-dimensional structures, and report your findings. Note that you need to use "labels.csv" for the task of discoverying patterns such as if different cancer types have distinct transformed gene expressions (that are represented by principal components). For PCA or Sparse PCA, low-dimensional structures are usually represented by the linear space spanned by some principal components.

```
# Variance explained by components
pve <- summary(pca_result)$importance[2, ]
cumulative_pve <- cumsum(pve)
```

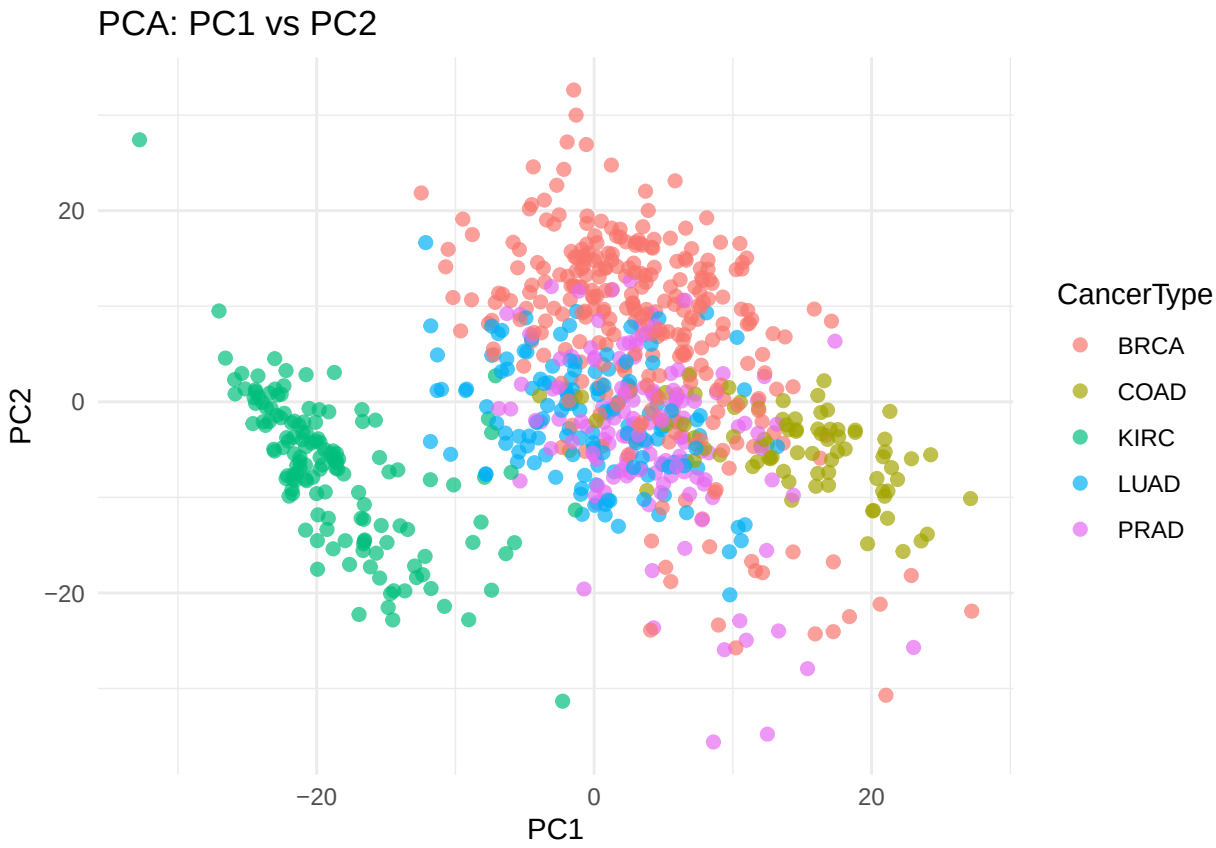
```r
# Scree plot
plot(pve[1:20], type = "b", pch = 19, col = "blue", xlab = "Principal Component",
    ylab = "Proportion of Variance Explained", main = "Scree Plot: Variance by Component")
```



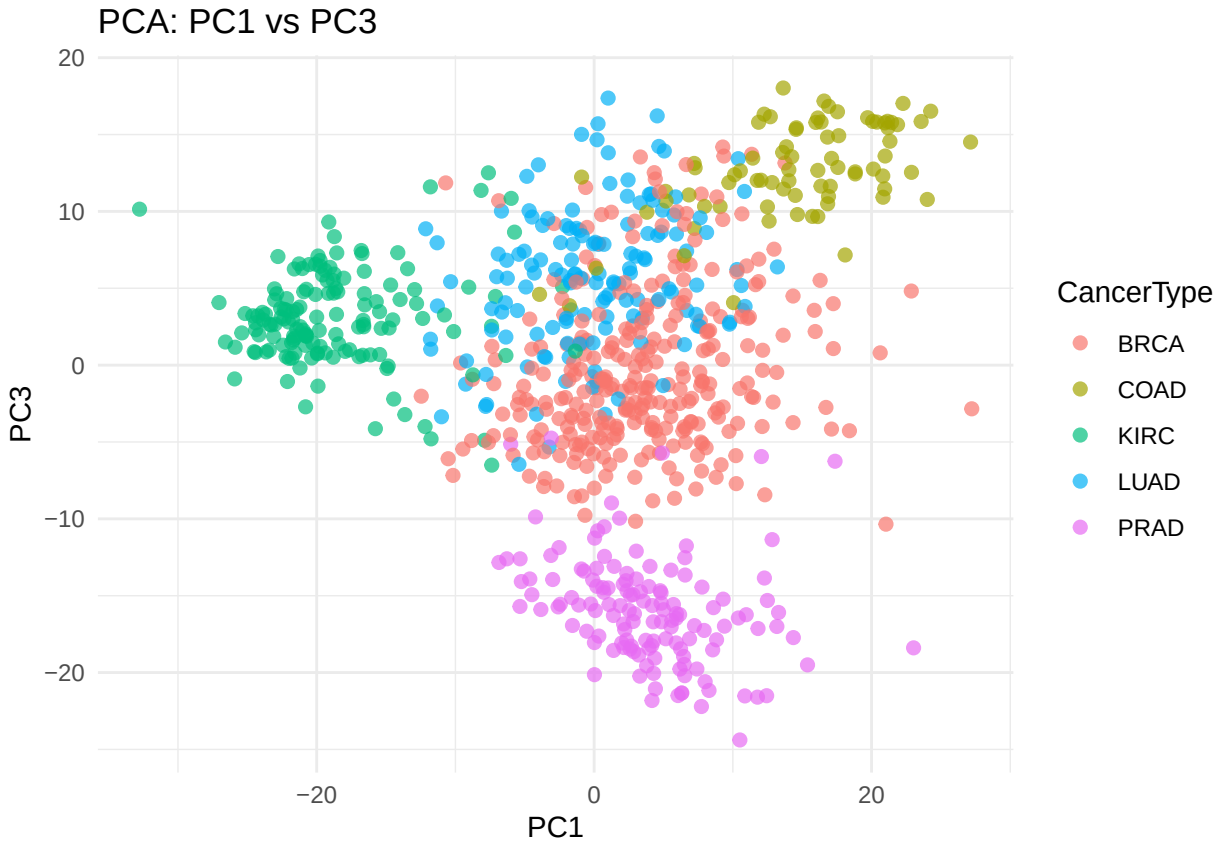

```r
# Cumulative variance plot
plot(cumulative_pve[1:20], type = "b", pch = 19, col = "darkgreen",
    xlab = "Number of Principal Components", ylab = "Cumulative Variance Explained",
    main = "Cumulative Variance Explained (First 20 PCs)")
```




```r
# Visualize low-dimensional structure (PC1 vs PC2, PC1 vs
# PC3)
library(ggplot2)

pca_scores <- as.data.frame(pca_result$x)
pca_scores$CancerType <- as.factor(labels[, 1])

# PC1 vs PC2
ggplot(pca_scores, aes(x = PC1, y = PC2, color = CancerType)) +
    geom_point(size = 2, alpha = 0.7) + labs(title = "PCA: PC1 vs PC2",
    x = "PC1", y = "PC2") + theme_minimal()
```




```
# PC1 vs PC3
ggplot(pca_scores, aes(x = PC1, y = PC3, color = CancerType)) +
    geom_point(size = 2, alpha = 0.7) + labs(title = "PCA: PC1 vs PC3",
    x = "PC1", y = "PC3") + theme_minimal()
```



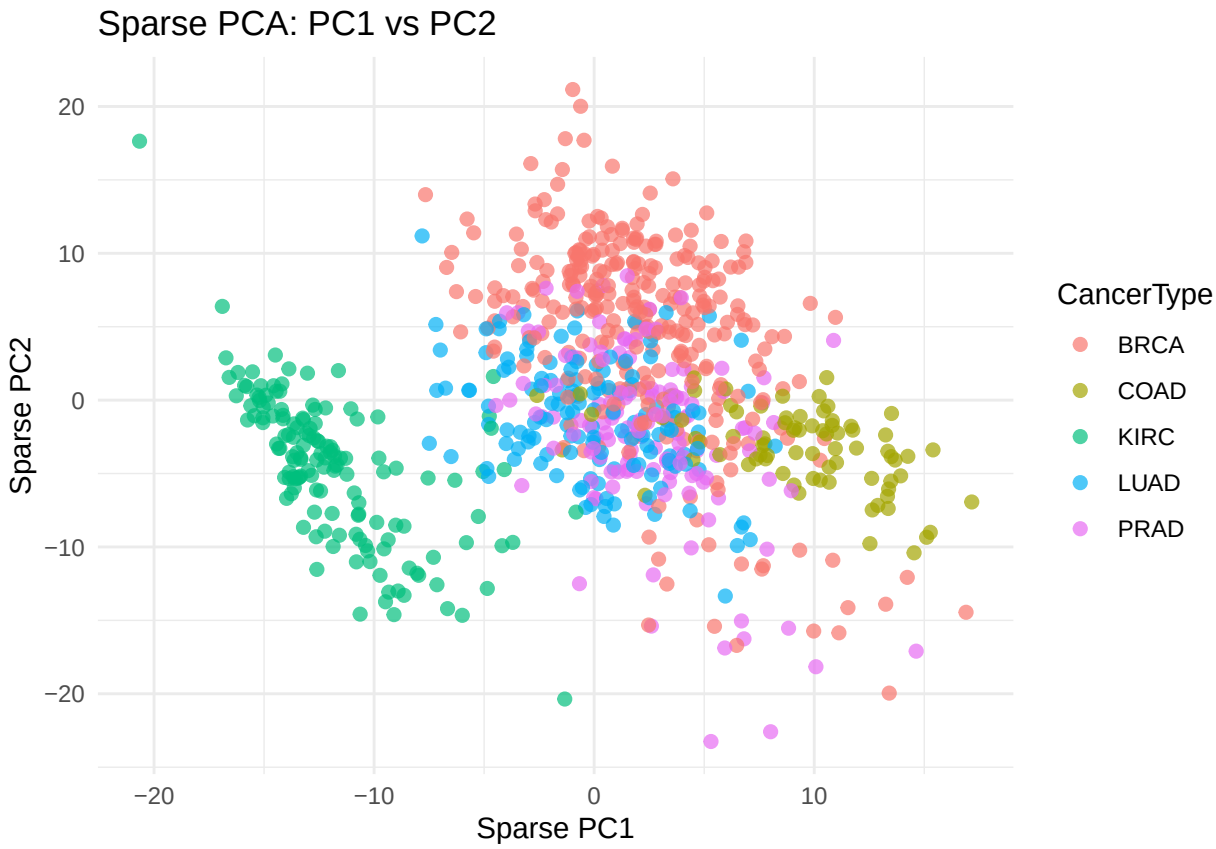
Interpretations

The scree plot and cumulative variance curve show that the first few principal components capture a substantial portion of the variance in the data, with approximately 60% explained by the first 20 components. Visualizations of the data projected onto PC1 vs PC2 and PC1 vs PC3 demonstrate clear low-dimensional structure. Specific cancer types, such as KIRC and PRAD, exhibit distinct clustering in the principal component space, while others like BRCA and LUAD show moderate overlap. These patterns suggest that PCA successfully reduces the dimensionality of the gene expression data while preserving meaningful variation, and that linear subspaces defined by leading principal components reflect biologically relevant groupings among cancer types.

(2.b) Apply Sparse PCA, provide visualizations of low-dimensional structures, and report your findings. Note that you need to use "labels.csv" for the task of discoverying patterns. Your laptop may not have sufficient computational power to implement Sparse PCA with many principal components. So, please pick a value for the sparsity controlling parameter and a value for the number of principal components to be computed that suit your computational capabilities.

```
library(elasticnet)
spca_result <- spca(stdgexpProj2, K = 5, para = rep(50, 5))

# Manually compute sparse PCA scores
sparse_scores <- stdgexpProj2 %*% spca_result$loadings

# Convert to data frame and label
spca_scores <- as.data.frame(sparse_scores[, 1:2])
```

```r
colnames(spca_scores) <- c("PC1", "PC2")
spca_scores$CancerType <- as.factor(labels[, 1])

# Sparse PCA Plot: PC1 vs PC2
library(ggplot2)
ggplot(spca_scores, aes(x = PC1, y = PC2, color = CancerType)) +
    geom_point(size = 2, alpha = 0.7) + labs(title = "Sparse PCA: PC1 vs PC2",
    x = "Sparse PC1", y = "Sparse PC2") + theme_minimal()
```



Interpretations

The Sparse PCA projection onto the first two sparse principal components shows well-defined clustering among several cancer types, particularly KIRC, COAD, and PRAD. Compared to standard PCA, the separation between groups appears more distinct and compact, with reduced overlap in the low-dimensional space. This suggests that Sparse PCA effectively captures biologically meaningful variation using fewer genes, enhancing both interpretability and visual clarity. By enforcing sparsity in the principal components, the method highlights gene subsets most relevant to distinguishing cancer types, revealing strong low-dimensional structure within the expression data.

(2.c) Do PCA and Sparse PCA reveal different low-dimensional structures for the gene expressions for different cancer types?

Interpretations

PCA and Sparse PCA reveal similar overall structure in the gene expression data, with both methods identifying major sources of variation across cancer types. However, Sparse PCA shows sharper separation between groups and less within-group dispersion. This is because Sparse PCA emphasizes only the most relevant genes for each principal component, reducing noise from uninformative features. In contrast, standard PCA incorporates all genes, which can dilute group-specific signals. As a result, Sparse PCA provides a more refined view of cancer-specific patterns, revealing clearer low-dimensional structures that may be more biologically interpretable.

Task B: analysis of SPAM emails data set

For this task, you need to use PCA and SVM.

Dataset and its description

The spam data set "SPAM.csv'' is attached and also can be downloaded from https://web.stanford .edu/~hastie/CASI_files/DATA/SPAM.html. More information on this data set can be found at: https://archive.ics.uci.edu/ml/datasets/Spambase. The column "testid" in "SPAM.csv" was used to train a model when the data set was used by other analysts and hence should not be used as a feature or the response, the column "spam" contains the true status for each email, and the rest contain measurements of features. Here each email is represented by a row of features in the .csv file, and a "feature" can be regarded as a "predictor". Also note that the first 1813 rows, i.e., observations, of the data set are for spam emails, and that the rest for non-spam emails.

Data processing

Please do the following:

- Remove rows that have missing values. For a .csv file, usually a blank cell is treated as a missing value.

```
library(readr)
library(dplyr)

spam <- read.csv("SPAM.csv")
names(spam)
```

```
##  [1] "spam"       "testid"     "make"       "address"    "all"
##  [6] "X3d"        "our"        "over"       "remove"     "internet"
## [11] "order"      "mail"       "receive"    "will"       "people"
## [16] "report"     "addresses"  "free"       "business"   "email"
## [21] "you"        "credit"     "your"       "font"       "X000"
## [26] "money"      "hp"         "hpl"        "george"     "X650"
## [31] "lab"        "labs"       "telnet"     "X857"       "data"
## [36] "X415"       "X85"        "technology" "X1999"      "parts"
## [41] "pm"         "direct"     "cs"         "meeting"    "original"
## [46] "project"    "re"         "edu"        "table"      "conference"
## [51] "ch."        "ch..1"      "ch..2"      "ch..3"      "ch..4"
## [56] "ch..5"      "crl.ave"    "crl.long"   "crl.tot"
```

```r
spam_clean <- na.omit(spam)
dim(spam_clean)
```

```
## [1] 4601   59
```

- Check for highly correlated features using the absolute value of sample correlation. Think about if you should include all or some of highly correlated features into an SVM model. For example, "crl.ave" (average length of uninterrupted sequences of capital letters), "crl.long" (length of longest uninterrupted sequence of capital letters) and "crl.tot" (total number of capital letters in the e-mail) may be highly correlated. Whethere you choose to remove some highly correlated features from subsequent analysis or not, you need to provide a justification for your choice.

Note that each feature is stored in a column of the original data set and each observation in a row. You will analyze the processed data set.

```r
spam_features <- spam_clean %>%
    select(-testid, -spam)
cor_matrix <- cor(spam_features)
high_cor <- which(abs(cor_matrix) > 0.9 & abs(cor_matrix) < 1,
    arr.ind = TRUE)
unique_pairs <- high_cor[high_cor[, 1] < high_cor[, 2], , drop = FALSE]

data.frame(Feature1 = rownames(cor_matrix)[unique_pairs[, 1]],
    Feature2 = colnames(cor_matrix)[unique_pairs[, 2]], Correlation = cor_matrix[unique_pairs])
```

```
##   Feature1 Feature2 Correlation
## 1     X857     X415   0.9960661
```

Interpretations

After removing rows with missing values, a correlation analysis was conducted on all numeric predictors in the SPAM dataset (excluding testid and spam). Using a threshold of 0.9 for high absolute correlation, no pairs of features exceeded this level. Therefore, no features were removed based on correlation. All features are retained for subsequent modeling since they are not significantly redundant, and removing any would not provide benefit in terms of reducing multicollinearity.

Classifiction via SVM

Please do the following:

(3.a) Use `set.seed(123)` wherever the command `sample` is used or cross-validation is implemented, randomly select without replacement 300 observations from the data set and save them as training set "train.RData", and then randomly select without replacement 100 observations from the remaining observations and save them as "test.RData". You need to check if the training set contains observations from both classes; otherwise, no model can be trained.

```r
library(dplyr)

set.seed(123)
```

```r
train_idx <- sample(1:nrow(spam_clean), size = 300, replace = FALSE)
train_set <- spam_clean[train_idx, ]
remaining <- spam_clean[-train_idx, ]

# Sample 100 from the remaining for the test set
test_idx <- sample(1:nrow(remaining), size = 100, replace = FALSE)
test_set <- remaining[test_idx, ]

# Check for both spam classes (0 and 1) in training data
table(train_set$spam)
```

```
##
## FALSE  TRUE
##   184   116
```

```r
# Save the datasets
save(train_set, file = "train.RData")
save(test_set, file = "test.RData")
```

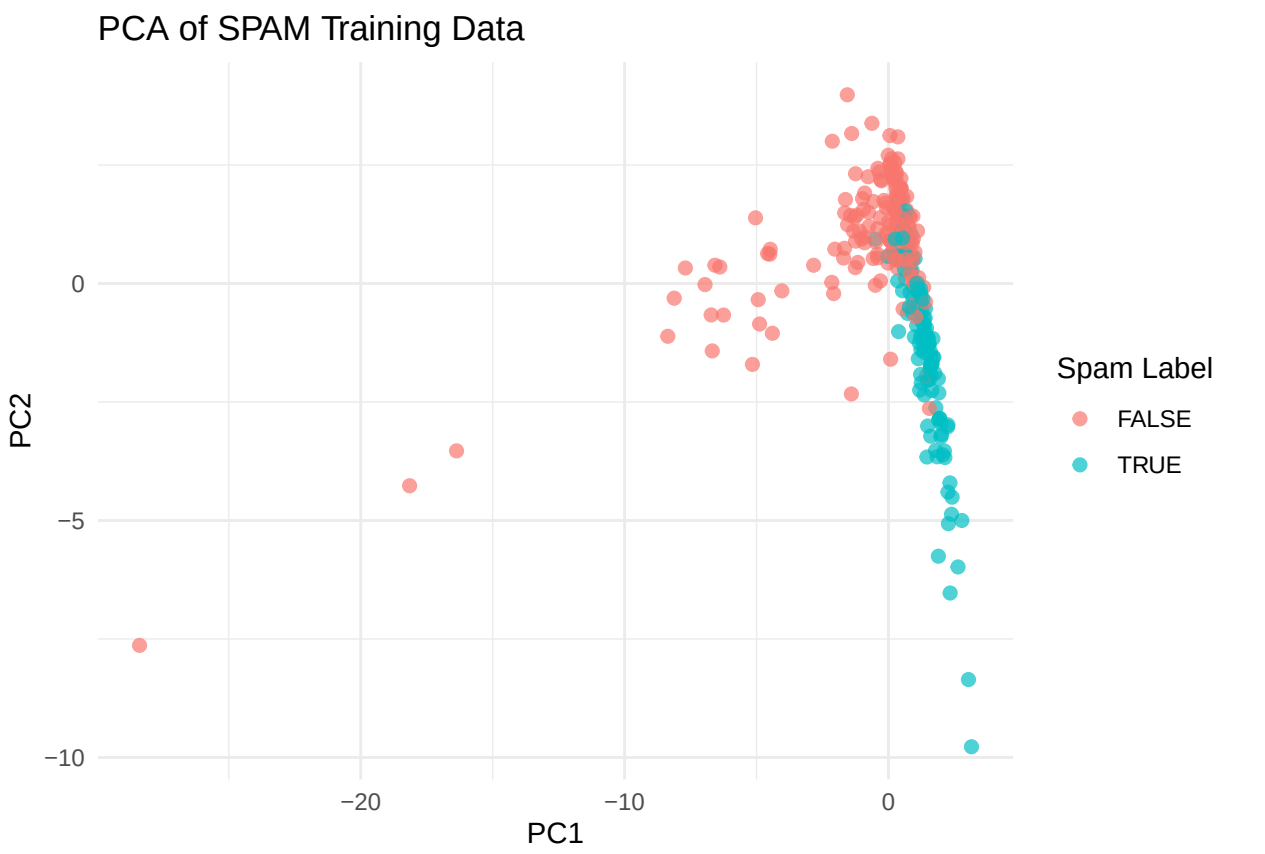
Interpretations

The data set was successfully partitioned into training and test sets using random sampling without replacement. The training set consists of 300 observations, and the test set contains 100 observations, both drawn from the cleaned data. To ensure the training data is appropriate for classification, the class distribution was checked. The training set includes 184 non-spam emails (label 0) and 116 spam emails (label 1), confirming that both classes are adequately represented. This balanced presence of both categories supports effective training of an SVM model, enabling it to learn to distinguish between spam and non-spam emails.

(3.b) Apply PCA to the training data "train.RData" and see if you find any pattern that can be used to approximately tell a spam email from a non-spam email.

```r
library(ggplot2)
load("train.RData")

train_features <- train_set[, !(names(train_set) %in% c("spam",
    "testid"))]
train_scaled <- scale(train_features)

# Apply PCA
pca_result <- prcomp(train_scaled)

# Create a data frame of the first two principal components
pca_df <- data.frame(PC1 = pca_result$x[, 1], PC2 = pca_result$x[,
    2], Spam = as.factor(train_set$spam))

# Plot the PCA results
ggplot(pca_df, aes(x = PC1, y = PC2, color = Spam)) + geom_point(alpha = 0.7,
    size = 2) + labs(title = "PCA of SPAM Training Data", color = "Spam Label") +
```

```
  theme_minimal()
```



Interpretations

The PCA plot of the SPAM training data reveals a distinguishable pattern between spam and non-spam emails when projected onto the first two principal components. Spam emails (labeled TRUE) tend to align more vertically along a narrow region with lower PC1 values, whereas non-spam emails (labeled FALSE) are spread more broadly and are concentrated in a different region of the PC1–PC2 space. This partial separation suggests that principal components derived from the data capture meaningful variance related to the spam classification and could be useful for downstream classification tasks. However, some overlap remains, indicating that PCA alone may not be sufficient for perfect classification but provides a helpful starting point.

(3.c) Use "train.RData" to build an SVM model with linear kernel, whose `cost` parameter is determined by 10-fold cross-validation, for which the features are predictors, the status of email is the response, and `cost` ranges in `c(0.01,0.1,1,5,10,50)`. Apply the obtained optimal model to "test.RData", and report via a 2-by-2 table on spams that are classified as spams or non-spams and on non-spams that are classified as non-spams or spams.

```
library(e1071)
library(caret)
```

```
## Loading required package: lattice
```

```r
load("train.RData")
load("test.RData")

train_x <- train_set[, !(names(train_set) %in% c("testid", "spam"))]
train_y <- as.factor(train_set$spam)

test_x <- test_set[, !(names(test_set) %in% c("testid", "spam"))]
test_y <- as.factor(test_set$spam)

set.seed(123)

tuned_model <- tune(svm, train.x = train_x, train.y = train_y,
    kernel = "linear", ranges = list(cost = c(0.01, 0.1, 1, 5,
        10, 50)), tunecontrol = tune.control(cross = 10))

# View the best model
best_svm <- tuned_model$best.model

# Predict on test data
pred_y <- predict(best_svm, test_x)

# Confusion matrix
conf_mat <- table(Predicted = pred_y, Actual = test_y)
print(conf_mat)
```

```
##           Actual
## Predicted FALSE TRUE
##     FALSE    55   12
##     TRUE      3   30
```

Interpretations

The confusion matrix indicates that the SVM model with a linear kernel classified the test emails with overall strong performance. Out of 100 test samples, 55 non-spam emails were correctly predicted as non-spam, and 30 spam emails were correctly identified as spam. However, the model misclassified 12 spam emails as non-spam (false negatives) and 3 non-spam emails as spam (false positives). While the model demonstrates good precision for the spam class, the number of false negatives suggests some spam messages may still go undetected. Nonetheless, the results reflect a well-balanced model with a relatively low error rate.

(3.d) Use "train.RData" to build an SVM model with radial kernel, whose "cost" parameter is determined by 10-fold cross-validation, for which the features are predictors, the status of email is the response, `cost` ranges in `c(0.01,0.1,1,5,10,50)`, and `gamma=c(0.5,1,2,3,4)`. Report the number of support vectors. Apply the obtained optimal model to "test.RData", and report via a 2-by-2 table on spams that are classified as spams or non-spams and on non-spams that are classified as non-spams or spams.

```r
library(e1071)
train_set$spam <- as.factor(train_set$spam)
```

```r
test_set$spam <- as.factor(test_set$spam)

# Tune SVM with radial kernel
set.seed(123)
tune_result_rbf <- tune(svm, spam ~ ., data = train_set, kernel = "radial",
    ranges = list(cost = c(0.01, 0.1, 1, 5, 10, 50), gamma = c(0.5,
        1, 2, 3, 4)), tunecontrol = tune.control(cross = 10))

# Best model
best_model_rbf <- tune_result_rbf$best.model

# Number of support vectors
num_support_vectors <- sum(best_model_rbf$nSV)
print(paste("Number of support vectors:", num_support_vectors))
```

```
## [1] "Number of support vectors: 295"
```

```r
# Predict on test data
pred_rbf <- predict(best_model_rbf, test_set)

# Confusion matrix
confusion_matrix_rbf <- table(Predicted = pred_rbf, Actual = test_set$spam)
print(confusion_matrix_rbf)
```

```
##           Actual
## Predicted FALSE TRUE
##     FALSE    58   38
##     TRUE      0    4
```

Interpretations

The SVM model with a radial kernel, tuned via 10-fold cross-validation, resulted in a total of 295 support vectors. When applied to the test set, the model demonstrated a strong ability to correctly classify non-spam emails (58 true negatives), but struggled significantly with spam detection. Of the 42 actual spam emails, only 4 were correctly identified (true positives), while 38 were misclassified as non-spam (false negatives). There were no false positives. This indicates that the radial kernel model is highly conservative, favoring the non-spam class and failing to capture the underlying structure of spam messages, leading to a high false negative rate.

(3.e) Compare and comment on the classification results obtained by (3.c) and (3.d).

Interpretations

Comparing the results from (3.c) and (3.d), the linear kernel SVM model clearly outperforms the radial kernel SVM in overall classification balance. In (3.c), the linear SVM achieved a better trade-off between sensitivity and specificity, correctly identifying 30 out of 42 spam emails and 55 out of 58 non-spam emails. In contrast, the radial kernel SVM in (3.d) only detected 4 spam emails while misclassifying 38 as non-spam, despite correctly identifying all non-spam emails. This suggests the linear model is more effective for this dataset, possibly due to the linear separability of the

features after preprocessing and PCA, whereas the radial model may have overfitted or failed to generalize well.