correlations

Marat Habibullin

27.12.2014

```
library(e1071)
library(MASS)

samples_num = 50
predictors_num = 10000 + 1 # + 1 for response

# let's get samples from normal distribution
mat = matrix(rnorm(samples_num * predictors_num), samples_num, predictors_num)
data_frame = data.frame(mat)

# let's extract most correlated predictors
data_abs_cors = abs(cor(data_frame))
sorted_df = data_frame[, order(data_abs_cors[1, ], decreasing = TRUE)]
most_cor_df = sorted_df[, 1:21]
tune(lm, X1 ~ ., data = most_cor_df, tunecontrol = tune.control(cross = nrow(most_cor_df)))
```

```
##
## Error estimation of 'lm' using leave-one-out: 0.3782806
```

```
# as we can see, the error is pretty good
# let's try to predict response for test samples using obtained predictors
mat = matrix(rnorm(samples_num * predictors_num), samples_num, predictors_num)
data_frame = data.frame(mat)
test_df = data_frame[, names(most_cor_df)]
tune(lm, X1 ~ ., data = test_df, tunecontrol = tune.control(cross = nrow(test_df)))
```

```
##
## Error estimation of 'lm' using leave-one-out: 2.226828
```

```
# here the error is high so we have overfitting
# let's look at predictors
model = lm(X1 ~ ., data = test_df)
summary(model)
```

```
##
## Call:
## lm(formula = X1 ~ ., data = test_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9816 -0.4861  0.1635  0.5937  1.7225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08584    0.18337  -0.468    0.643
```

```
## X2840       -0.17931    0.19707  -0.910    0.370
## X785        -0.18484    0.21566  -0.857    0.398
## X7639        0.02055    0.17028   0.121    0.905
## X8263        0.18146    0.24540   0.739    0.466
## X9944       -0.05951    0.18945  -0.314    0.756
## X2138       -0.12755    0.16887  -0.755    0.456
## X1390       -0.06079    0.20492  -0.297    0.769
## X3834        0.08816    0.19780   0.446    0.659
## X8460       -0.01154    0.19294  -0.060    0.953
## X3619        0.00380    0.18929   0.020    0.984
## X5616       -0.21466    0.23407  -0.917    0.367
## X3538        0.03185    0.20038   0.159    0.875
## X3130       -0.07121    0.18223  -0.391    0.699
## X3390        0.02354    0.17227   0.137    0.892
## X636         0.12150    0.26506   0.458    0.650
## X5230       -0.34606    0.20657  -1.675    0.105
## X7876       -0.19833    0.21434  -0.925    0.362
## X2320       -0.03552    0.28351  -0.125    0.901
## X4305       -0.03110    0.24158  -0.129    0.898
## X5416       -0.12303    0.21401  -0.575    0.570
##
## Residual standard error: 1.098 on 29 degrees of freedom
## Multiple R-squared:  0.2549, Adjusted R-squared:  -0.259
## F-statistic: 0.496 on 20 and 29 DF,  p-value: 0.9466
```

```r
# all predictors are insignificant!

model.aic = stepAIC(model)
```

```
## Start:  AIC=24.11
## X1 ~ X2840 + X785 + X7639 + X8263 + X9944 + X2138 + X1390 + X3834 +
##     X8460 + X3619 + X5616 + X3538 + X3130 + X3390 + X636 + X5230 +
##     X7876 + X2320 + X4305 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X3619  1    0.0005 34.964 22.114
## - X8460  1    0.0043 34.967 22.120
## - X7639  1    0.0176 34.981 22.139
## - X2320  1    0.0189 34.982 22.140
## - X4305  1    0.0200 34.983 22.142
## - X3390  1    0.0225 34.986 22.146
## - X3538  1    0.0305 34.993 22.157
## - X1390  1    0.1061 35.069 22.265
## - X9944  1    0.1190 35.082 22.283
## - X3130  1    0.1841 35.147 22.376
## - X3834  1    0.2395 35.203 22.455
## - X636   1    0.2533 35.216 22.474
## - X5416  1    0.3984 35.361 22.680
## - X8263  1    0.6592 35.622 23.047
## - X2138  1    0.6879 35.651 23.087
## - X785   1    0.8857 35.849 23.364
## - X2840  1    0.9981 35.961 23.521
## - X5616  1    1.0139 35.977 23.543
## - X7876  1    1.0323 35.995 23.568
```

```
## <none>               34.963 24.113
## - X5230  1    3.3835 38.347 26.732
##
## Step:  AIC=22.11
## X1 ~ X2840 + X785 + X7639 + X8263 + X9944 + X2138 + X1390 + X3834 +
##     X8460 + X5616 + X3538 + X3130 + X3390 + X636 + X5230 + X7876 +
##     X2320 + X4305 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X8460  1    0.0049 34.968 20.121
## - X7639  1    0.0180 34.981 20.140
## - X4305  1    0.0195 34.983 20.142
## - X2320  1    0.0216 34.985 20.145
## - X3390  1    0.0228 34.986 20.147
## - X3538  1    0.0303 34.994 20.157
## - X9944  1    0.1211 35.085 20.287
## - X1390  1    0.1232 35.087 20.290
## - X3130  1    0.1955 35.159 20.393
## - X3834  1    0.2390 35.203 20.455
## - X636   1    0.2607 35.224 20.485
## - X5416  1    0.4100 35.373 20.697
## - X2138  1    0.6878 35.651 21.088
## - X8263  1    0.7037 35.667 21.110
## - X785   1    0.8958 35.859 21.379
## - X2840  1    1.0019 35.965 21.527
## - X7876  1    1.0558 36.019 21.602
## - X5616  1    1.1054 36.069 21.670
## <none>               34.964 22.114
## - X5230  1    3.5625 38.526 24.965
##
## Step:  AIC=20.12
## X1 ~ X2840 + X785 + X7639 + X8263 + X9944 + X2138 + X1390 + X3834 +
##     X5616 + X3538 + X3130 + X3390 + X636 + X5230 + X7876 + X2320 +
##     X4305 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X2320  1    0.0181 34.986 18.147
## - X4305  1    0.0199 34.988 18.150
## - X7639  1    0.0201 34.989 18.150
## - X3390  1    0.0243 34.993 18.156
## - X3538  1    0.0336 35.002 18.169
## - X9944  1    0.1188 35.087 18.291
## - X1390  1    0.1202 35.089 18.293
## - X3130  1    0.2069 35.175 18.416
## - X3834  1    0.2344 35.203 18.455
## - X636   1    0.3117 35.280 18.565
## - X5416  1    0.4051 35.373 18.697
## - X8263  1    0.6989 35.667 19.111
## - X2138  1    0.7509 35.719 19.183
## - X785   1    0.9325 35.901 19.437
## - X2840  1    1.0214 35.990 19.561
## - X7876  1    1.1016 36.070 19.672
## - X5616  1    1.1796 36.148 19.780
## <none>               34.968 20.121
```

```
## - X5230  1    3.5847 38.553 23.001
##
## Step:  AIC=18.15
## X1 ~ X2840 + X785 + X7639 + X8263 + X9944 + X2138 + X1390 + X3834 +
##     X5616 + X3538 + X3130 + X3390 + X636 + X5230 + X7876 + X4305 +
##     X5416
##
##         Df Sum of Sq    RSS    AIC
## - X3390  1    0.0246 35.011 16.182
## - X4305  1    0.0278 35.014 16.187
## - X7639  1    0.0297 35.016 16.189
## - X3538  1    0.0439 35.030 16.210
## - X9944  1    0.1093 35.096 16.303
## - X1390  1    0.1113 35.098 16.306
## - X3130  1    0.1907 35.177 16.419
## - X3834  1    0.2208 35.207 16.461
## - X5416  1    0.3920 35.378 16.704
## - X636   1    0.4026 35.389 16.719
## - X8263  1    0.7110 35.697 17.153
## - X2138  1    0.7810 35.767 17.251
## - X785   1    0.9167 35.903 17.440
## - X2840  1    1.0034 35.990 17.561
## - X7876  1    1.1204 36.107 17.723
## <none>               34.986 18.147
## - X5616  1    1.6671 36.654 18.474
## - X5230  1    3.6170 38.603 21.066
##
## Step:  AIC=16.18
## X1 ~ X2840 + X785 + X7639 + X8263 + X9944 + X2138 + X1390 + X3834 +
##     X5616 + X3538 + X3130 + X636 + X5230 + X7876 + X4305 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X4305  1    0.0196 35.031 14.210
## - X7639  1    0.0368 35.048 14.235
## - X3538  1    0.0504 35.061 14.254
## - X1390  1    0.1217 35.133 14.356
## - X3130  1    0.1660 35.177 14.419
## - X9944  1    0.1675 35.179 14.421
## - X3834  1    0.2556 35.267 14.546
## - X5416  1    0.3757 35.387 14.716
## - X636   1    0.4832 35.494 14.867
## - X2138  1    0.7594 35.770 15.255
## - X8263  1    0.7953 35.806 15.305
## - X785   1    0.8960 35.907 15.445
## - X2840  1    0.9823 35.993 15.566
## - X7876  1    1.1348 36.146 15.777
## <none>               35.011 16.182
## - X5616  1    1.6680 36.679 16.509
## - X5230  1    3.6324 38.643 19.118
##
## Step:  AIC=14.21
## X1 ~ X2840 + X785 + X7639 + X8263 + X9944 + X2138 + X1390 + X3834 +
##     X5616 + X3538 + X3130 + X636 + X5230 + X7876 + X5416
##
```

```
##         Df Sum of Sq    RSS    AIC
## - X7639  1    0.0372 35.068 12.263
## - X3538  1    0.0517 35.082 12.284
## - X1390  1    0.1022 35.133 12.356
## - X9944  1    0.1529 35.184 12.428
## - X3130  1    0.1631 35.194 12.442
## - X3834  1    0.2800 35.311 12.608
## - X636   1    0.5063 35.537 12.928
## - X5416  1    0.5396 35.570 12.974
## - X2138  1    0.7428 35.773 13.259
## - X2840  1    0.9829 36.014 13.594
## - X785   1    1.0133 36.044 13.636
## - X8263  1    1.0244 36.055 13.651
## - X7876  1    1.1851 36.216 13.874
## <none>               35.031 14.210
## - X5616  1    1.6509 36.682 14.512
## - X5230  1    3.6185 38.649 17.125
##
## Step:  AIC=12.26
## X1 ~ X2840 + X785 + X8263 + X9944 + X2138 + X1390 + X3834 + X5616 +
##     X3538 + X3130 + X636 + X5230 + X7876 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X3538  1    0.0512 35.119 10.336
## - X1390  1    0.0743 35.142 10.369
## - X9944  1    0.1697 35.238 10.505
## - X3130  1    0.2309 35.299 10.591
## - X3834  1    0.2510 35.319 10.620
## - X636   1    0.5312 35.599 11.015
## - X5416  1    0.5527 35.621 11.045
## - X2138  1    0.7283 35.796 11.291
## - X2840  1    0.9571 36.025 11.610
## - X785   1    0.9854 36.053 11.649
## - X8263  1    0.9942 36.062 11.661
## - X7876  1    1.1553 36.223 11.884
## <none>               35.068 12.263
## - X5616  1    1.6374 36.705 12.545
## - X5230  1    3.8279 38.896 15.443
##
## Step:  AIC=10.34
## X1 ~ X2840 + X785 + X8263 + X9944 + X2138 + X1390 + X3834 + X5616 +
##     X3130 + X636 + X5230 + X7876 + X5416
##
##         Df Sum of Sq    RSS     AIC
## - X1390  1    0.0933 35.212  8.4688
## - X9944  1    0.1542 35.273  8.5552
## - X3130  1    0.2175 35.337  8.6448
## - X3834  1    0.2463 35.365  8.6855
## - X636   1    0.4970 35.616  9.0388
## - X5416  1    0.6582 35.777  9.2645
## - X2138  1    0.7656 35.885  9.4144
## - X8263  1    1.0573 36.176  9.8192
## - X2840  1    1.1021 36.221  9.8811
## - X785   1    1.1681 36.287  9.9721
```

```
## - X7876  1    1.1721 36.291  9.9776
## <none>               35.119 10.3361
## - X5616  1    1.6017 36.721 10.5660
## - X5230  1    3.8303 38.949 13.5121
##
## Step:  AIC=8.47
## X1 ~ X2840 + X785 + X8263 + X9944 + X2138 + X3834 + X5616 + X3130 +
##     X636 + X5230 + X7876 + X5416
##
##         Df Sum of Sq    RSS     AIC
## - X9944  1    0.1612 35.374  6.6971
## - X3130  1    0.2171 35.430  6.7761
## - X3834  1    0.2732 35.486  6.8551
## - X636   1    0.6013 35.814  7.3153
## - X5416  1    0.6929 35.905  7.4431
## - X2138  1    0.7147 35.927  7.4735
## - X2840  1    1.0535 36.266  7.9428
## - X8263  1    1.0900 36.302  7.9930
## - X7876  1    1.0919 36.304  7.9957
## - X785   1    1.2397 36.452  8.1988
## <none>               35.212  8.4688
## - X5616  1    1.7074 36.920  8.8362
## - X5230  1    3.7911 39.003 11.5814
##
## Step:  AIC=6.7
## X1 ~ X2840 + X785 + X8263 + X2138 + X3834 + X5616 + X3130 + X636 +
##     X5230 + X7876 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X3834  1    0.2506 35.624 5.0501
## - X3130  1    0.2639 35.638 5.0688
## - X636   1    0.5225 35.896 5.4303
## - X2138  1    0.6597 36.033 5.6210
## - X5416  1    0.7476 36.121 5.7428
## - X7876  1    1.0446 36.418 6.1522
## - X8263  1    1.0522 36.426 6.1628
## - X2840  1    1.1088 36.482 6.2403
## <none>               35.374 6.6971
## - X785   1    1.6237 36.997 6.9410
## - X5616  1    1.7514 37.125 7.1134
## - X5230  1    3.8124 39.186 9.8148
##
## Step:  AIC=5.05
## X1 ~ X2840 + X785 + X8263 + X2138 + X5616 + X3130 + X636 + X5230 +
##     X7876 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X3130  1    0.2762 35.900 3.4363
## - X636   1    0.3963 36.021 3.6033
## - X2138  1    0.6558 36.280 3.9622
## - X5416  1    0.7628 36.387 4.1095
## - X7876  1    0.8608 36.485 4.2439
## - X2840  1    0.8735 36.498 4.2613
## - X8263  1    1.0912 36.715 4.5586
```

```
## <none>               35.624 5.0501
## - X5616  1    1.8369 37.461 5.5640
## - X785   1    2.2588 37.883 6.1240
## - X5230  1    3.7927 39.417 8.1087
##
## Step:  AIC=3.44
## X1 ~ X2840 + X785 + X8263 + X2138 + X5616 + X636 + X5230 + X7876 +
##     X5416
##
##         Df Sum of Sq    RSS    AIC
## - X636   1    0.4332 36.334 2.0360
## - X2138  1    0.5317 36.432 2.1714
## - X2840  1    0.7346 36.635 2.4490
## - X7876  1    0.7674 36.668 2.4938
## - X5416  1    0.8467 36.747 2.6017
## - X8263  1    1.2533 37.154 3.1521
## <none>               35.900 3.4363
## - X5616  1    1.9149 37.815 4.0345
## - X785   1    2.1759 38.076 4.3784
## - X5230  1    3.5998 39.500 6.2141
##
## Step:  AIC=2.04
## X1 ~ X2840 + X785 + X8263 + X2138 + X5616 + X5230 + X7876 + X5416
##
##         Df Sum of Sq    RSS    AIC
## - X2138  1    0.6376 36.971 0.9058
## - X7876  1    0.7315 37.065 1.0326
## - X8263  1    1.0416 37.375 1.4492
## - X2840  1    1.0693 37.403 1.4862
## - X5416  1    1.1685 37.502 1.6187
## <none>               36.334 2.0360
## - X5616  1    1.9268 38.260 2.6195
## - X785   1    2.0978 38.431 2.8425
## - X5230  1    3.3794 39.713 4.4828
##
## Step:  AIC=0.91
## X1 ~ X2840 + X785 + X8263 + X5616 + X5230 + X7876 + X5416
##
##         Df Sum of Sq    RSS     AIC
## - X7876  1    0.4582 37.429 -0.4784
## - X8263  1    0.8044 37.776 -0.0179
## - X2840  1    1.0231 37.994  0.2707
## - X5416  1    1.4746 38.446  0.8613
## <none>               36.971  0.9058
## - X785   1    1.8296 38.801  1.3209
## - X5616  1    1.9408 38.912  1.4640
## - X5230  1    3.4553 40.426  3.3731
##
## Step:  AIC=-0.48
## X1 ~ X2840 + X785 + X8263 + X5616 + X5230 + X5416
##
##         Df Sum of Sq    RSS      AIC
## - X8263  1    0.5699 37.999 -1.72278
## - X2840  1    0.7663 38.196 -1.46506
```

```
## - X5416  1    1.5200 38.949 -0.48805
## <none>               37.429 -0.47838
## - X785   1    1.6377 39.067 -0.33717
## - X5616  1    2.4114 39.841  0.64343
## - X5230  1    3.4502 40.880  1.93030
##
## Step:  AIC=-1.72
## X1 ~ X2840 + X785 + X5616 + X5230 + X5416
##
##         Df Sum of Sq    RSS      AIC
## - X2840  1    0.6348 38.634 -2.89444
## - X5416  1    1.1117 39.111 -2.28101
## - X785   1    1.3031 39.302 -2.03685
## <none>               37.999 -1.72278
## - X5616  1    2.4615 40.461 -0.58453
## - X5230  1    3.3276 41.327  0.47447
##
## Step:  AIC=-2.89
## X1 ~ X785 + X5616 + X5230 + X5416
##
##         Df Sum of Sq    RSS     AIC
## - X5416  1    0.7696 39.404 -3.9082
## <none>               38.634 -2.8944
## - X785   1    1.6843 40.318 -2.7609
## - X5616  1    2.0074 40.641 -2.3617
## - X5230  1    3.4032 42.037 -0.6734
##
## Step:  AIC=-3.91
## X1 ~ X785 + X5616 + X5230
##
##         Df Sum of Sq    RSS     AIC
## - X785   1    1.5514 40.955 -3.9773
## <none>               39.404 -3.9082
## - X5616  1    1.8024 41.206 -3.6719
## - X5230  1    3.0257 42.429 -2.2091
##
## Step:  AIC=-3.98
## X1 ~ X5616 + X5230
##
##         Df Sum of Sq    RSS     AIC
## <none>               40.955 -3.9773
## - X5616  1    2.3117 43.267 -3.2319
## - X5230  1    3.1381 44.093 -2.2859
```

```
summary(model.aic)
```

```
##
## Call:
## lm(formula = X1 ~ X5616 + X5230, data = test_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8793 -0.5432  0.1664  0.7639  1.4376
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03807    0.13428  -0.284   0.7780
## X5616       -0.23984    0.14725  -1.629   0.1100
## X5230       -0.27040    0.14249  -1.898   0.0639 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9335 on 47 degrees of freedom
## Multiple R-squared:  0.1272, Adjusted R-squared:  0.09003
## F-statistic: 3.424 on 2 and 47 DF,  p-value: 0.04091
```

```r
# here the situations is the same

# now let's try crossvalidations instead of test train
# we need an appropriate learner function

special_lm <- function(formula, data, subset) {
  train = data[subset, ]
  train_cor_abs = abs(cor(train))
  train_selected = train[, order(train_cor_abs[1, ], decreasing = TRUE)[1:21]]
  return(lm(X1 ~ ., data = train_selected))
}

# I have reduced the predictors size here because with 10000 my computer
# hangs (high computational complexity)
samples_num = 50
predictors_num = 7500 + 1 # + 1 for response

# let's get samples from normal distribution
mat = matrix(rnorm(samples_num * predictors_num), samples_num, predictors_num)
data_frame = data.frame(mat)

tune(special_lm, X1 ~ ., data = data_frame, tunecontrol = tune.control(sampling = "cross"))
```

```
##
## Error estimation of 'special_lm' using 10-fold cross validation: 1.50498
```