Prob4.2_Universities_PCA.R

cmrump

Wed Feb 20 15:58:26 2019

```
#Shmueli 4.2
setwd("~/My Courses/Data Mining/Datasets/DMBA-3eR-datasets")
universities.df<-read.csv("Universities.csv")
nrow(universities.df)
```

```
## [1] 1302
```

```
#turn off scientific notation
options(scipen = 999)
options(digits=4) #limit to 5 decimal places

#Remove missing data and first two (text) columns
universities.complete.df<-na.omit(universities.df[,-c(1,2)])
nrow(universities.complete.df)
```

```
## [1] 471
```

```
head(universities.complete.df)
```

```
##    Public..1...Private..2. X..appli..rec.d X..appl..accepted
## 1                        2             193               146
## 3                        1             146               117
## 10                       2             805               588
## 12                       2             608               520
## 22                       2            4414              1500
## 26                       1            1797              1260
##    X..new.stud..enrolled X..new.stud..from.top.10.
## 1                     55                        16
## 3                     89                         4
## 10                   287                        67
## 12                   127                        26
## 22                   335                        30
## 26                   938                        24
##    X..new.stud..from.top.25. X..FT.undergrad X..PT.undergrad
## 1                         44             249             869
## 3                         24             492            1849
## 10                        88            1376             207
## 12                        47             538             126
## 22                        60             908             119
## 26                        35            6960            4698
##    in.state.tuition out.of.state.tuition room board add..fees
## 1              7560                 7560 1620  2500       130
## 3              1742                 5226 2514  2250        34
## 10            11660                11660 2050  2430       120
## 12             8080                 8080 1380  2540       100
## 22             5666                 5666 1424  1540       418
## 26             2220                 4440 1935  3240       291
##    estim..book.costs estim..personal.. X..fac..w.PHD stud..fac..ratio
## 1                800              1500            76             11.9
## 3                500              1162            39              9.5
## 10               400               900            74             14.0
## 12               500              1100            63             11.4
## 22              1000              1400            56             15.5
## 26               750              2200            96              6.7
##    Graduation.rate
## 1               15
## 3               39
## 10              72
## 12              44
## 22              46
## 26              33
```

```
#Summary statistics
summary(universities.complete.df)
```

```
##  Public..1...Private..2. X..appli..rec.d X..appl..accepted
##  Min.   :1.00            Min.   :   77   Min.   :   61
##  1st Qu.:1.00            1st Qu.:  802   1st Qu.:  636
##  Median :2.00            Median : 1646   Median : 1227
##  Mean   :1.73            Mean   : 3147   Mean   : 2063
##  3rd Qu.:2.00            3rd Qu.: 3862   3rd Qu.: 2456
##  Max.   :2.00            Max.   :48094   Max.   :26330
##  X..new.stud..enrolled X..new.stud..from.top.10. X..new.stud..from.top.25.
##  Min.   :  27          Min.   : 1                Min.   :  9.0
##  1st Qu.: 264          1st Qu.:15                1st Qu.: 40.0
##  Median : 443          Median :23                Median : 54.0
##  Mean   : 781          Mean   :28                Mean   : 55.6
##  3rd Qu.: 896          3rd Qu.:36                3rd Qu.: 69.0
##  Max.   :6392          Max.   :96                Max.   :100.0
##  X..FT.undergrad X..PT.undergrad in.state.tuition out.of.state.tuition
##  Min.   :  249   Min.   :    1   Min.   :  608    Min.   : 1044
##  1st Qu.: 1018   1st Qu.:   82   1st Qu.: 3650    1st Qu.: 7290
##  Median : 1715   Median :  299   Median : 9858    Median :10100
##  Mean   : 3563   Mean   :  797   Mean   : 9407    Mean   :10575
##  3rd Qu.: 4056   3rd Qu.:  869   3rd Qu.:13246    3rd Qu.:13286
##  Max.   :31643   Max.   :21836   Max.   :20100    Max.   :20100
##       room          board          add..fees     estim..book.costs
##  Min.   : 640   Min.   : 531   Min.   :  10   Min.   :  90
##  1st Qu.:1740   1st Qu.:1750   1st Qu.: 138   1st Qu.: 500
##  Median :2090   Median :2082   Median : 280   Median : 500
##  Mean   :2221   Mean   :2122   Mean   : 379   Mean   : 549
##  3rd Qu.:2663   3rd Qu.:2420   3rd Qu.: 486   3rd Qu.: 600
##  Max.   :4816   Max.   :4541   Max.   :3247   Max.   :2340
##  estim..personal.. X..fac..w.PHD   stud..fac..ratio Graduation.rate
##  Min.   : 250      Min.   :  8.0   Min.   : 2.9     Min.   : 15.0
##  1st Qu.: 850      1st Qu.: 63.0   1st Qu.:11.3     1st Qu.: 53.0
##  Median :1200      Median : 76.0   Median :13.4     Median : 66.0
##  Mean   :1312      Mean   : 73.2   Mean   :14.0     Mean   : 65.6
##  3rd Qu.:1600      3rd Qu.: 87.0   3rd Qu.:16.4     3rd Qu.: 79.0
##  Max.   :6800      Max.   :103.0   Max.   :28.8     Max.   :118.0
```

```
#Correlation
universities.cor <- cor(universities.complete.df)
universities.cor[18,] #show correlations to last variable: Graduation rate
```

```
##   Public..1...Private..2.           X..appli..rec.d
##                   0.33673                   0.18206
##         X..appl..accepted     X..new.stud..enrolled
##                   0.09835                   0.01274
## X..new.stud..from.top.10. X..new.stud..from.top.25.
##                   0.55819                   0.57538
##           X..FT.undergrad           X..PT.undergrad
##                  -0.04221                  -0.23527
##          in.state.tuition      out.of.state.tuition
##                   0.57968                   0.62133
##                      room                     board
##                   0.36956                   0.41115
##                 add..fees         estim..book.costs
##                   0.04618                   0.05028
##         estim..personal..             X..fac..w.PHD
##                  -0.23946                   0.41313
##          stud..fac..ratio           Graduation.rate
##                  -0.31899                   1.00000
```

```
#Principal Component Analysis on covariance matrix (not normalized)
pcs.cov=prcomp(universities.complete.df, scale=FALSE)
summary(pcs.cov)
```

```
## Importance of components:
##                             PC1      PC2      PC3       PC4       PC5
## Standard deviation     7430.914 5987.989 1854.641 1192.5293 967.42792
## Proportion of Variance    0.561    0.365    0.035    0.0145   0.00951
## Cumulative Proportion     0.561    0.926    0.961    0.9753   0.98484
##                             PC6       PC7       PC8       PC9      PC10
## Standard deviation     679.6527 596.97612 580.62990 417.61364 318.12719
## Proportion of Variance   0.0047   0.00362   0.00343   0.00177   0.00103
## Cumulative Proportion    0.9895   0.99316   0.99658   0.99836   0.99938
##                             PC11      PC12 PC13 PC14 PC15 PC16 PC17  PC18
## Standard deviation     188.86761 155.60617   19 12.5   11 5.33 2.91 0.169
## Proportion of Variance   0.00036   0.00025    0  0.0    0 0.00 0.00 0.000
## Cumulative Proportion    0.99975   0.99999    1  1.0    1 1.00 1.00 1.000
```

```
#Principal Componenent rotations (weights) - first 3
pcs.cov$rot[,c(1:3)]
```

```
##                                   PC1          PC2         PC3
## Public..1...Private..2.    0.00004787 -0.000000725  0.00000591
## X..appli..rec.d           -0.27188262 -0.551183388  0.66445794
## X..appl..accepted         -0.19410703 -0.321299373  0.19095677
## X..new.stud..enrolled     -0.08472979 -0.101589931 -0.08745130
## X..new.stud..from.top.10.  0.00089847 -0.001732235  0.00013613
## X..new.stud..from.top.25.  0.00081134 -0.001924733  0.00004003
## X..FT.undergrad           -0.45812113 -0.492263413 -0.63530316
## X..PT.undergrad           -0.10825320 -0.073409535 -0.28535277
## in.state.tuition           0.67018731 -0.382489131 -0.08278654
## out.of.state.tuition       0.45453453 -0.428685058 -0.12940964
## room                       0.03342006 -0.055583985  0.04011290
## board                      0.03423588 -0.040897364 -0.00823166
## add..fees                 -0.01320940 -0.008746080  0.03286783
## estim..book.costs          0.00005792 -0.003290568  0.00031583
## estim..personal..         -0.03755717 -0.001185110 -0.05465889
## X..fac..w.PHD              0.00020469 -0.001564059 -0.00099533
## stud..fac..ratio          -0.00029544  0.000158708  0.00002522
## Graduation.rate            0.00107232 -0.001397446  0.00092015
```

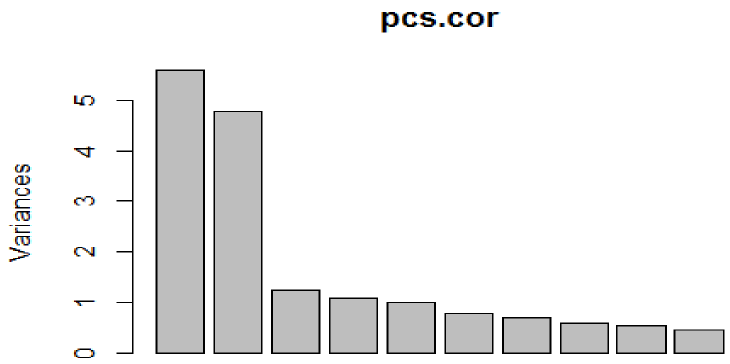
```
#Principal Component Analysis on correlation matrix (normalized)
pcs.cor=prcomp(universities.complete.df, scale=TRUE)
summary(pcs.cor)
```

```
## Importance of components:
##                          PC1   PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation     2.365 2.188 1.1102 1.0328 0.9908 0.8738 0.8347
## Proportion of Variance 0.311 0.266 0.0685 0.0593 0.0545 0.0424 0.0387
## Cumulative Proportion  0.311 0.577 0.6452 0.7045 0.7590 0.8014 0.8401
##                           PC8    PC9   PC10  PC11  PC12   PC13   PC14
## Standard deviation     0.7728 0.7339 0.6627 0.630 0.585 0.4585 0.4377
## Proportion of Variance 0.0332 0.0299 0.0244 0.022 0.019 0.0117 0.0106
## Cumulative Proportion  0.8733 0.9032 0.9276 0.950 0.969 0.9804 0.9910
##                           PC15    PC16   PC17   PC18
## Standard deviation     0.30051 0.18897 0.1472 0.1198
## Proportion of Variance 0.00502 0.00198 0.0012 0.0008
## Cumulative Proportion  0.99602 0.99800 0.9992 1.0000
```

```
#Principal Componenent rotations (weights) - first 5
pcs.cor$rot[,c(1:5)]
```

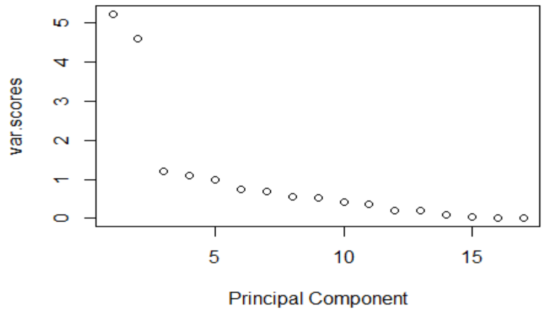
```
##                                PC1      PC2       PC3       PC4      PC5
## Public..1...Private..2.   -0.31659 -0.14748  0.171296 -0.032228  0.19892
## X..appli..rec.d            0.08825  0.40572  0.001790  0.063741  0.07163
## X..appl..accepted          0.13920  0.39331  0.014698  0.104870  0.15252
## X..new.stud..enrolled      0.19078  0.38112  0.005916 -0.042852  0.12227
## X..new.stud..from.top.10. -0.26938  0.23999 -0.139189 -0.366049 -0.18097
## X..new.stud..from.top.25. -0.24877  0.25581 -0.157072 -0.380724 -0.17068
## X..FT.undergrad            0.20969  0.37111  0.035582 -0.065777  0.12282
## X..PT.undergrad            0.19663  0.20594  0.299609  0.047936  0.29568
## in.state.tuition          -0.39658  0.02763  0.091330  0.036630  0.16487
## out.of.state.tuition      -0.37125  0.13174  0.044006  0.075838  0.13409
## room                      -0.21511  0.16760  0.207258  0.467781 -0.04503
## board                     -0.25264  0.13408  0.241015  0.304649  0.15474
## add..fees                  0.08259  0.14938 -0.295487  0.460392 -0.55344
## estim..book.costs         -0.03027  0.07431  0.592505 -0.001139 -0.58704
## estim..personal..          0.18045  0.07282  0.414603 -0.388779 -0.14930
## X..fac..w.PHD             -0.13624  0.29189 -0.220135 -0.071846 -0.03634
## stud..fac..ratio           0.28359 -0.03072 -0.170716  0.096338  0.04013
## Graduation.rate           -0.27320  0.16449 -0.190082 -0.014842 -0.07363
```

```
#Z scores (to replace original X data)
pcs.scores.cor <- pcs.cor$x
head(pcs.scores.cor)

##          PC1     PC2     PC3     PC4     PC5     PC6      PC7      PC8
## 1   0.60194 -2.0604  1.9975 -0.3163 -0.3898 -0.4883  1.14804  0.78503
## 3   2.04406 -2.5561  1.2103  0.9154  0.5347 -0.7890  0.46192  1.15893
## 10 -2.14970 -0.1710 -1.1086 -1.4803  0.2973  0.2222  0.52778 -0.08734
## 12 -0.06876 -2.0200  0.3034 -0.4909  0.5690 -0.3004  0.04514  0.35269
## 22  1.10941 -1.2119  1.3319 -0.9285 -2.1491  1.5154 -0.28724  0.97351
## 26  2.02237  0.8082  2.5997 -0.0851 -0.1848 -2.5392  2.54612  0.31877
##         PC9   PC10    PC11    PC12     PC13     PC14     PC15     PC16
## 1   0.6678 1.1036 -1.5778 -0.6505  0.67262  0.26639 -0.08204 -0.02878
## 3  -0.6834 0.7984  0.1983  2.1807 -0.60103  0.25501 -0.06858  0.14340
## 10  0.3455 1.2687  0.7837  0.1804  0.08845 -0.03488  0.30999  0.28645
## 12  0.8037 1.0124 -0.8787  0.4494  0.34545  0.19075  0.14474  0.08151
## 22  0.8578 0.3687 -0.2520 -0.6333  0.57488  0.68882  0.06925 -0.10468
## 26  0.8644 0.4206 -1.2215  1.7081  0.48088  0.08747  0.44237  0.03356
##         PC17      PC18
## 1   0.05361 -0.027375
## 3   0.03237  0.043043
## 10 -0.26510 -0.001884
## 12 -0.11128  0.008375
## 22  0.08351 -0.044022
## 26 -0.16211  0.364728

#scree plot of summary variances which sum to #predictors, p
plot(pcs.cor)
```



```
#diagonal shows the variance captured by each principal component
var.scores <- diag(cov(pcs.scores.cor))
var.scores

##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## 5.59224 4.78918 1.23245 1.06670 0.98175 0.76347 0.69674 0.59726 0.53857
##    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
## 0.43921 0.39672 0.34183 0.21027 0.19160 0.09031 0.03571 0.02166 0.01435
```

```
plot(var.scores, xlab = "Principal Component")
```



Problem 4.2

a. The categorical (text) variables in columns 1 & 2 are removed along with all but 471 of the original 1302 rows due to missing data. The first 6 rows of the complete data are shown on pp. 1-2. Summary statistics are shown on p. 2.

b. In the correlation analysis performed on p. 3, notice that graduation rates (the higher the better) are inversely correlated with student-faculty ratio, as expected: larger class sizes leads to poorer graduation outcomes. Graduation rates increase with better caliber of students (percentage in top 10%, 25% of high school class) and with more PhD faculty. (Surprised?!) Graduation rates also correlate higher amongst private colleges that charge higher tuition, room & board. We might suspect that these latter monetary variables are positively correlated with expensive private colleges and so it probably helps our modeling efforts to "wash away" this multi-collinearity by running a principal component analysis (PCA).

   A PCA on the raw data was performed first displaying the first three principal components (pp. 3-4). Notice that the student caliber percentages, PhD faculty percentages, and graduation rates aren't captured in these components since those relatively small numbers (on scale 0-100) are dominated by much larger (application and enrolled) student counts and tuition dollar figures which make up the bulk of the component weights. Same goes for important student-faculty ratios.

   A standardized PCA on was performed instead (p. 4). If we desire capturing 90% of the "information" in this dataset, that will require keeping the first 9 components; still that cuts the dataset by half of its original 18 numerical variables. Only the first five principal components are displayed on since a scree plot (bottom p. 5) shows only those components capture "a variable's worth" of variation. Notice now that all five capture the student caliber information and the first three capture information on graduation rates and PhD faculty to some degree. Student-faculty ratios appear in the first and third components.

   PC4 is somewhat interesting in that it primarily pits room, board, additional fees against student caliber and personal expenses, largely ignoring everything else (e.g., tuition and student counts).

   The head of the rescored data appears atop p. 5. This data could be used in building a predictive model. Perhaps though to predict graduation rates by regressing on other (explanatory) variables, we should start over and delete column 20 along with columns 1-2 at the start. Then append column 20 to the scored data frame.