10/08/2020

MVA ASSIGNMENT 4

MEMBER INFORMATION

Ruixin Yang (RUID: 197000459)

Aishwarya Senthilvel (RUID: 199001269)

PRINCIPAL COMPONENT ANALYSIS

The primary aim of PCA being for reducing the dimensionality of large datasets and increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximizing variance.

Here we first apply the principal component analysis after we replace categorical variables with corresponding dummy variables. And the result is following. For simplicity, we only choose the first three components whose proportions of the total variance is larger than 0.05.

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.77377	1.63864	1.5323	1.40758	1.29325	1.20117	1.18361	1.1299	1.11828	1.10338
Proportion of Variance	0.07491	0.06393	0.0559	0.04717	0.03982	0.03435	0.03336	0.0304	0.02977	0.02899
Cumulative Proportion	0.07491	0.13884	0.1947	0.24192	0.28174	0.31609	0.34945	0.3799	0.40962	0.43861

	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
Standard deviation	1.07454	1.06949	1.06780	1.05786	1.03523	1.02101	1.01577	1.01403	1.00901	0.99625
Proportion of Variance	0.02749	0.02723	0.02715	0.02664	0.02552	0.02482	0.02457	0.02448	0.02424	0.02363
Cumulative Proportion	0.46610	0.49333	0.52048	0.54712	0.57264	0.59746	0.62203	0.64651	0.67075	0.69438

	PC21	PC22	PC23	PC24	PC25	PC26	PC27	PC28	PC29	PC30
Standard deviation	0.9892	0.98442	0.96933	0.95055	0.93983	0.92810	0.91295	0.9051	0.88047	0.87715
Proportion of Variance	0.0233	0.02307	0.02237	0.02151	0.02103	0.02051	0.01984	0.0195	0.01846	0.01832
Cumulative Proportion	0.7177	0.74076	0.76313	0.78464	0.80567	0.82618	0.84602	0.8655	0.88399	0.90230

	PC31	PC32	PC33	PC34	PC35	PC36	PC37	PC38	PC39	PC40
Standard deviation	0.79078	0.77263	0.74649	0.70252	0.6768	0.66809	0.51164	0.46249	0.41573	0.33635
Proportion of Variance	0.01489	0.01421	0.01327	0.01175	0.0109	0.01063	0.00623	0.00509	0.00412	0.00269
Cumulative Proportion	0.91719	0.93141	0.94467	0.95643	0.9673	0.97796	0.98419	0.98928	0.99340	0.99609

	PC41	PC42
Standard deviation	0.30714	0.26424
Proportion of Variance	0.00225	0.00166
Cumulative Proportion	0.99834	1.00000

❖ EIGEN VALUES

Principal Components are associated with the eigenvectors of either the covariance or correlation matrix of the data. The i^{th} principal component (PC) is the line that follows the eigenvector associated with the i^{th} largest eigenvalue. Here the sum of eigenvalues is 42.

Below is the outcome of eigen values:

```
> eigen_bank
       PC1        PC2        PC3        PC4        PC5        PC6        PC7        PC8        PC9 
3.14625561 2.68514013 2.34786758 1.98128892 1.67248914 1.44280203 1.40092360 1.27677434 1.25054166 
      PC10       PC11       PC12       PC13       PC14       PC15       PC16       PC17       PC18 
1.21744494 1.15463484 1.14380981 1.14019897 1.11906935 1.07170466 1.04245266 1.03178244 1.02825789 
      PC19       PC20       PC21       PC22       PC23       PC24       PC25       PC26       PC27 
1.01809444 0.99250696 0.97859660 0.96908810 0.93959342 0.90354629 0.88328120 0.86137010 0.83346915 
      PC28       PC29       PC30       PC31       PC32       PC33       PC34       PC35       PC36 
0.81917619 0.77521876 0.76939165 0.62533861 0.59695081 0.55725320 0.49353626 0.45800530 0.44634380 
      PC37       PC38       PC39       PC40       PC41       PC42 
0.26177840 0.21390024 0.17283339 0.11312971 0.09433686 0.06982199 
```

❖ PROPORTION OF VARIANCE

The proportion of variance of the dataset is found by dividing the sum of squares of the columns of Λ^Λ^ (the eigenvalues of sum of squared) by the sum of the eigenvalues of SS.

```
> propvar
        PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8 
0.074910848 0.063931908 0.055901609 0.047173546 0.039821170 0.034352429 0.033355324 0.030399389 
        PC9        PC10        PC11        PC12        PC13        PC14        PC15        PC16 
0.029774801 0.028986784 0.027491306 0.027233567 0.027147594 0.026644508 0.025516778 0.024820302 
       PC17        PC18        PC19        PC20        PC21        PC22        PC23        PC24 
0.024566249 0.024482331 0.024240344 0.023631118 0.023299919 0.023073526 0.022371272 0.021513007 
       PC25        PC26        PC27        PC28        PC29        PC30        PC31        PC32 
0.021030505 0.020508812 0.019844504 0.019504195 0.018457589 0.018318849 0.014889014 0.014213114 
       PC33        PC34        PC35        PC36        PC37        PC38        PC39        PC40 
0.013267933 0.011750863 0.010904888 0.010627233 0.006232819 0.005092863 0.004115081 0.002693565 
       PC41        PC42 
0.002246116 0.001662428 
```

❖ CUMULATIVE PROPORTION

This is simply the accumulated amount of explained variance, i.e. if we used the first 10 components, we would be able to account for >95% of total variance in the data

```
> cumvar_bank
       PC1        PC2        PC3        PC4        PC5        PC6        PC7        PC8        PC9 
0.07491085 0.13884276 0.19474436 0.24191791 0.28173908 0.31609151 0.34944683 0.37984622 0.40962102 
      PC10       PC11       PC12       PC13       PC14       PC15       PC16       PC17       PC18 
0.43860781 0.46609911 0.49333268 0.52048028 0.54712478 0.57264156 0.59746186 0.62202811 0.64651044 
      PC19       PC20       PC21       PC22       PC23       PC24       PC25       PC26       PC27 
0.67075079 0.69438190 0.71768182 0.74075535 0.76312662 0.78463963 0.80567013 0.82617894 0.84602345 
      PC28       PC29       PC30       PC31       PC32       PC33       PC34       PC35       PC36 
0.86552764 0.88398523 0.90230408 0.91719310 0.93140621 0.94467414 0.95642501 0.96732990 0.97795713 
      PC37       PC38       PC39       PC40       PC41       PC42 
0.98418995 0.98928281 0.99339789 0.99609146 0.99833757 1.00000000 
```

❖ T-TEST

Standard deviations of scores for all the PC's are classified by term deposit subscription status. Attached below is the computed scores of term deposit status along with standard deviation and mean values.

means.xlsx
sds.xlsx
score.xlsx

```
> t.test(PC1~bank_2$y,data=bank_pca)

        Welch Two Sample t-test

data:  PC1 by bank_2$y
t = -12.363, df = 603.04, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.409213 -1.022866
sample estimates:
 mean in group no mean in group yes
       -0.1401364         1.0759030

> t.test(PC2~bank_2$y,data=bank_pca)

        Welch Two Sample t-test

data:  PC2 by bank_2$y
t = -5.365, df = 679.32, p-value = 1.111e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5358779 -0.2487300
sample estimates:
 mean in group no mean in group yes
       -0.0452091         0.3470948

> t.test(PC3~bank_2$y,data=bank_pca)

        Welch Two Sample t-test

data:  PC3 by bank_2$y
t = 1.8821, df = 609.23, p-value = 0.06029
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006929547  0.326094104
sample estimates:
 mean in group no mean in group yes
       0.01839026       -0.14119202
```

❖ F-TEST

As you can see there are many components from out dataset, we further use F test to test if there is any difference on variances of principal components between customers who bought the product and people who do not. And it turns out that we only fail to reject the null hypothesis in the second component under the level of significance of 0.1. That means there are significant differences on the first and the third components. But the difference on the second component is insignificant.

```
        F test to compare two variances

data:  PC1 by bank$y
F = 0.59357, num df = 3999, denom df = 520, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5198603 0.6733164
sample estimates:
ratio of variances
         0.5935662
```

```
        F test to compare two variances

data:  PC2 by bank$y
F = 1.1096, num df = 3999, denom df = 520, p-value = 0.1237
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9718051 1.2586696
sample estimates:
ratio of variances
          1.109588
```

```
        F test to compare two variances

data:  PC3 by bank$y
F = 0.63637, num df = 3999, denom df = 520, p-value = 3.383e-13
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5573521 0.7218753
sample estimates:
ratio of variances
         0.6363736
```

Furthermore, we applied the Levene’s test to see if the difference mentioned above. Under the level of significance of 0.01, the outcome is the same as that of F tests. So, the outcome above can be trusted.

```
> (LTPC1 <- leveneTest(PC1~bank$y,data=bank_2))
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  95.313 < 2.2e-16 ***
      4519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
```

```
> (LTPC1 <- leveneTest(PC2~bank$y,data=bank_2))
Levene's Test for Homogeneity of Variance (center = median)
        Df F value  Pr(>F)
group    1   4.288 0.03844 *
      4519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
```

```
> (LTPC1 <- leveneTest(PC3~bank$y,data=bank_2))
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  55.051 1.395e-13 ***
      4519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
```

Finally, we replicate the original values from principal components with the following code.

d1= data.frame(drop(scale(bank_new,center=center, scale=scale)%*%bank_pca$rotation[,1]))

d2= data.frame(drop(scale(bank_new,center=center, scale=scale)%*%bank_pca$rotation[,2]))

d3= data.frame(drop(scale(bank_new,center=center, scale=scale)%*%bank_pca$rotation[,3]))

And replicated variables look the same as those of the original dataset.

And we later use the predicting function to predict the first three components. The summary of descriptive statistics for the first three components is the following, respectively.

```
> summary (c)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.8357 -1.1311 -0.5924  0.0000  0.1895  8.8741
```

```
> summary (c)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-4.76248 -1.29067 -0.06197  0.00000  1.16311  3.93291
```

```
> summary (c)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.7193 -0.9479 -0.1709  0.0000  1.1107  4.1420
```