**MKT 382 Marketing Analytics II**

**Assignment 5**

**Due: April 29th, 11:59pm**

**Principal Component and Factor Analysis**

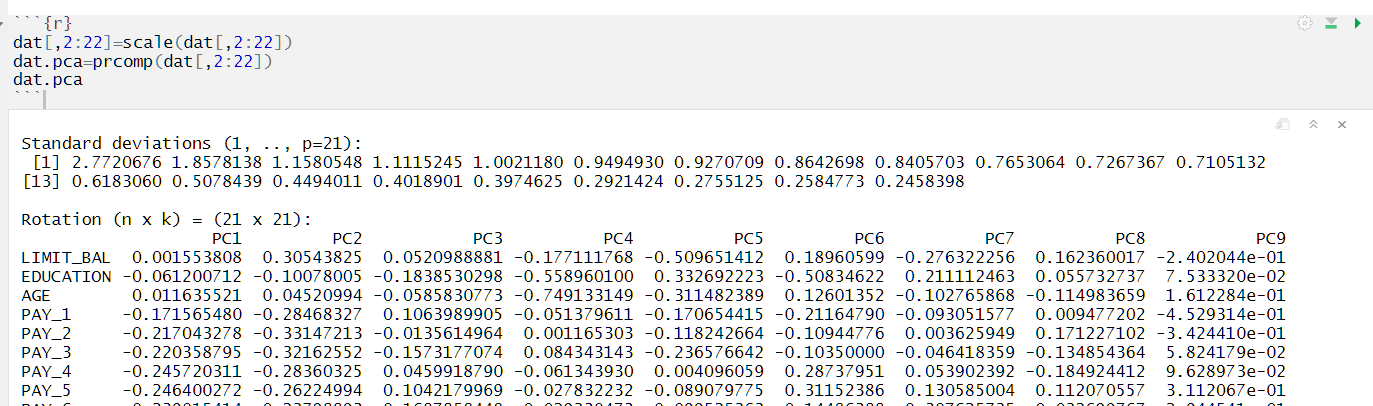
In this assignment, we will apply principal component analysis and factor analysis to multivariate data. Please download the data file "Credit\_clients.csv" from Canvas. This is a sample from a dataset which has been used to predict credit card clients' default payments. There are demographic variables and behavior-history variables. Many of the behavioral variables are highly correlated, so not all of them can be used in a regression for prediction. Multivariate techniques help us understand the relationship among many variables in order to reduce the dimensionality of the data. The variables in current sample dataset are as follows.

|  |  |
| --- | --- |
| CUSTOMERID | Customer ID |
| LIMIT\_BAL | Credit limit |
| Education | Education level |
| Age | Customer age |
| PAY\_1 to PAY\_6 | 6 months of payment history: -2 = 2 weeks before due, -1 = one week before due, 0 = in the due week, +1 = one week after due, etc. |
| BILL\_AMT1 to BILL\_AMT6 | 6 months of billed amount |
| PAY\_AMT1 to PAY\_AMT6 | 6 months of paid amount |

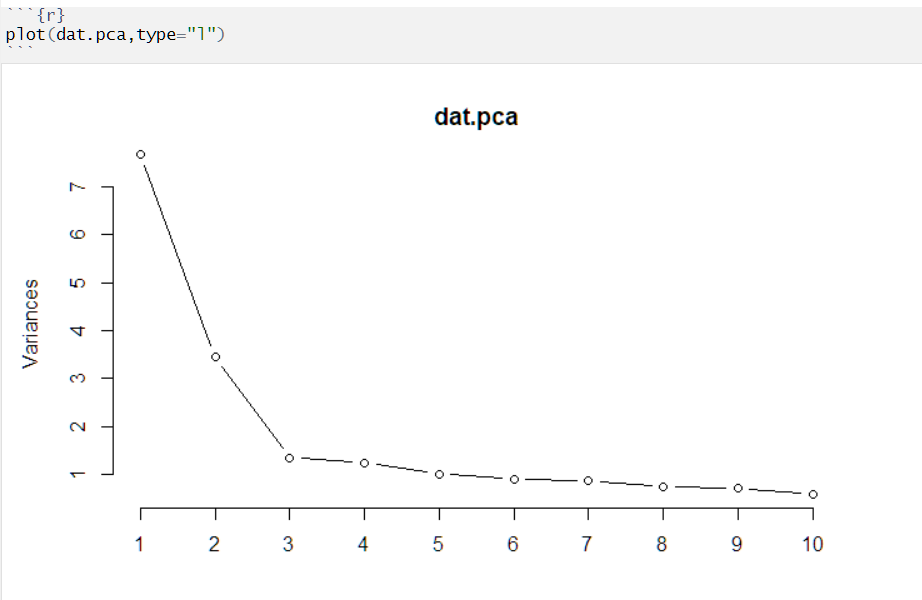
1). Use the log-transformation for LIMIT\_BAL, BILL\_AMT1 to BILL\_AMT6, PAT\_AMT1 to PAY\_AMT6 (Note: log(0) is –Inf, so please add a small value such as 1 to the original data before you take log).



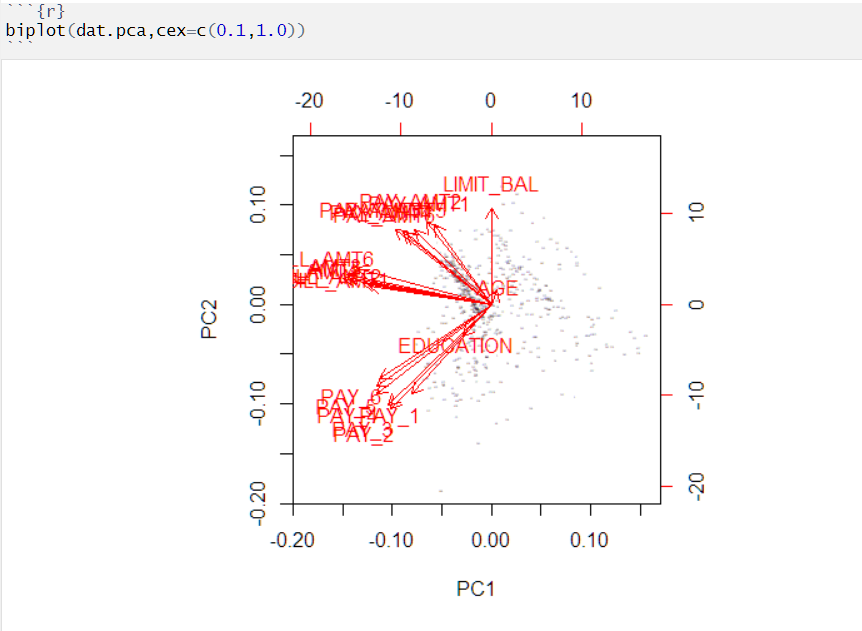
2). Please use the scale( ) function to scale the log-transformed LIMIT\_BAL, BILL\_AMT1 to BILL\_AMT6 and PAT\_AMT1 to PAY\_AMT6 (for example, scale(LIMIT\_BAL)). Then use the function prcomp( ) to run principal component analysis. Copy and paste the PCA results here.



3).Please use the plot( ) functions to create a scree plot of the principal component variances and post the figure here. How many variances (eigen values) are greater than 1? Use biplot( ) (you can set cex=c(0.2, 1.0) in biplot to create a clearer plot) to plot the PCA and paste the figure here. How do you interpret the plot?



From the graph above, we can tell that the variances of the first two PCAs are obviously greater than 1. Using standard deviation we got from prcomp(), we can tell that the variances of pc3, pc4 and pc5 are also slightly greater than 1.



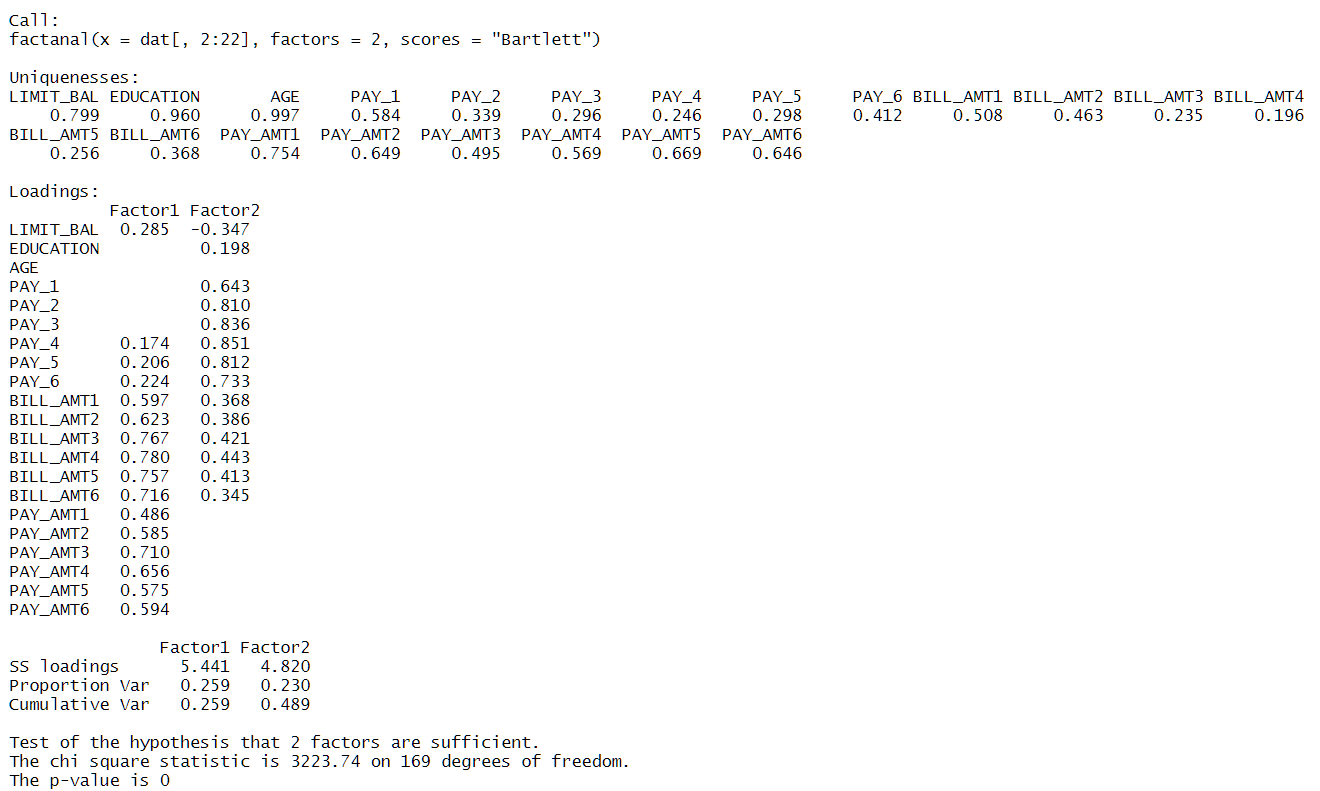
As shown in the biplot above:

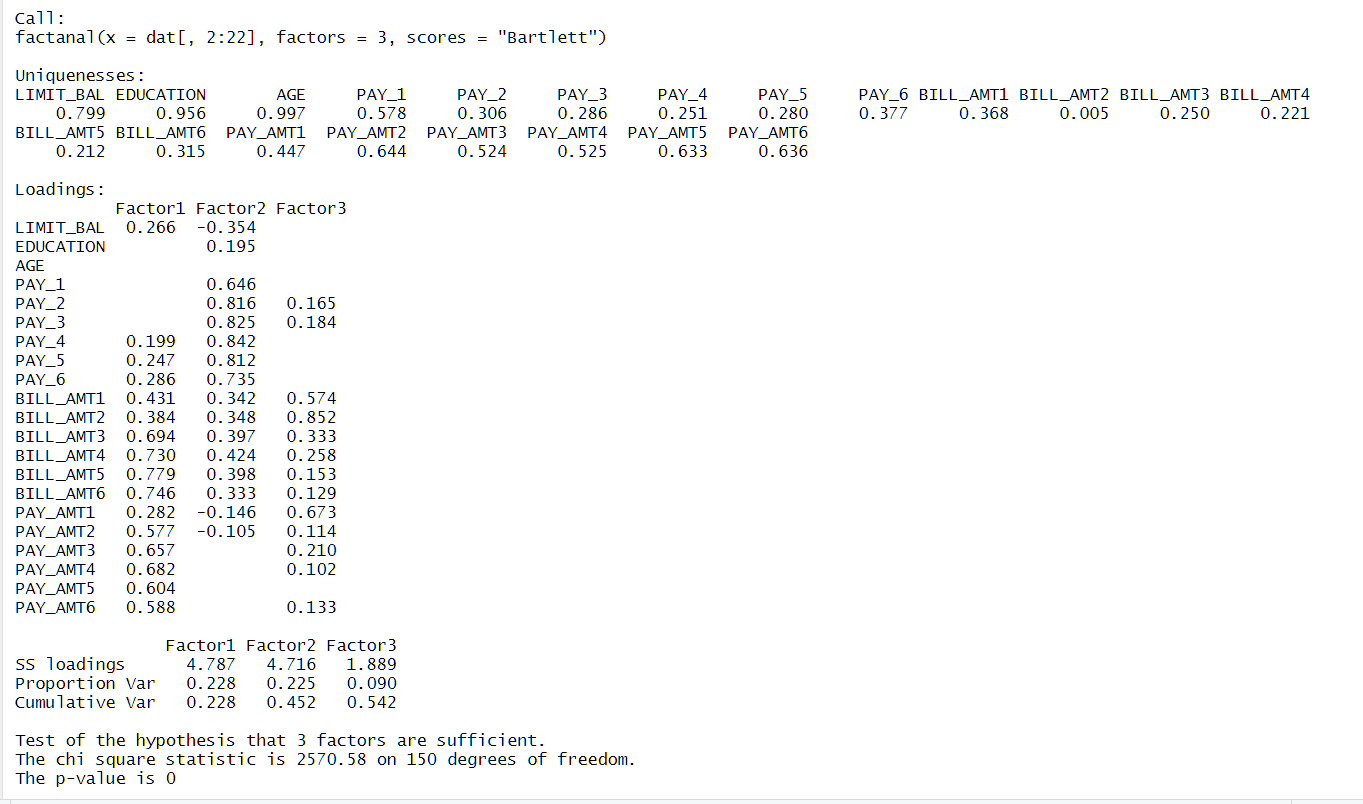
* the 6 variables related to payment amount are highly correlated with each other
* the 6 variables related to payment date are highly correlated with each other
* the 6 variables related to bill amount are highly correlated with each other
* payment amount and payment date are independent
* payment amount and bill amount are related since they are in similar direction, but they are not highly correlated

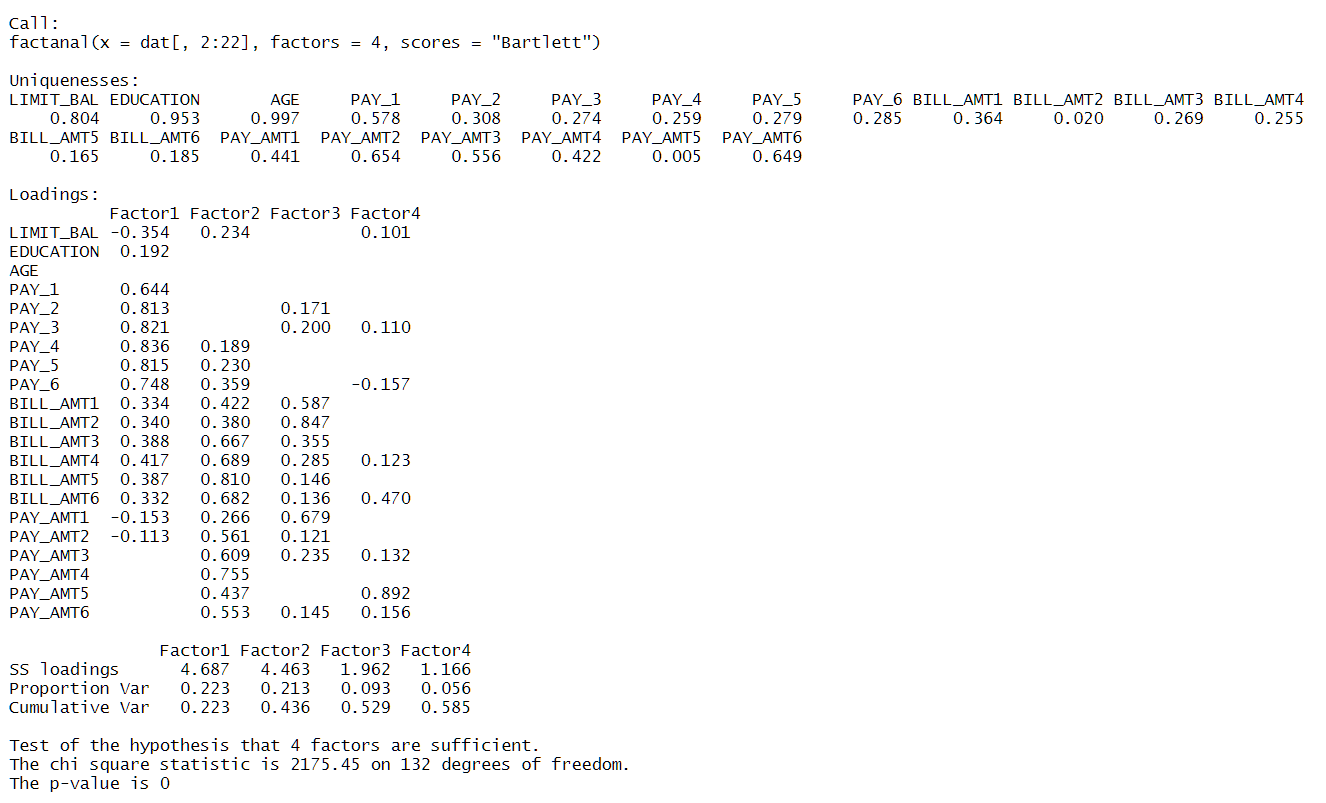
4). Next, we will conduct factor analysis. Use the R function factanal( ) to run orthogonal factor analysis with 2, 3 and 4 factors. Copy and paste the results here.

Which of the 2, 3, and 4 factor models do you find most interpretable? How would you interpret the factors from the factor loading matrix?

As shown in the FA with 2, 3 or 4 factors are all sufficient, so actually we can use any one of them. For me, I think the 2 factors model is the most interpretable, where factor 1 represents bill amount and pay amount, and factor 2 represents pay date.







5). To plot the results of factor analysis, you need to create a function (copy and paste the following code to R)

biplot.fa = function (fa,...) {

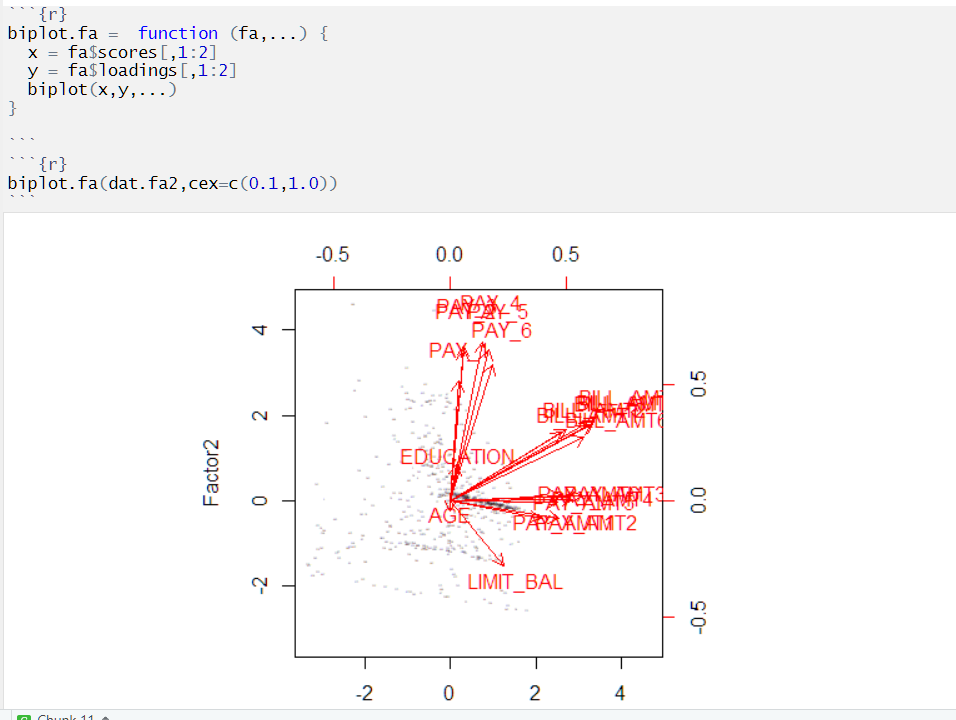
x = fa$scores[,1:2]

y = fa$loadings[,1:2]

biplot(x,y,...)

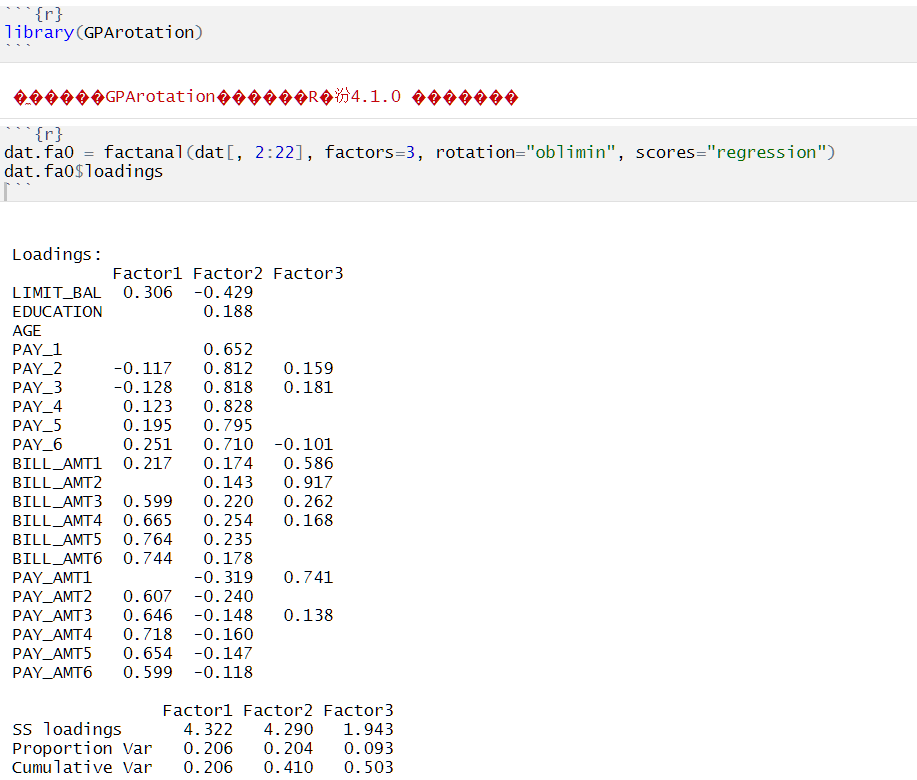
}

Please plot a two-factor model and paste the figure here. How do you interpret the plot? Is the plot similar to the PCA biplot?



The biplot and the interpretation of 2-factor FA are almost the same as the ones of PCA

6). Next, we will try the *oblimin* rotation for oblique EFA. Please install the package "GPArotation" for the oblimin rotation. Please create a 3-factor EFA model using the rotation *oblimin*. Copy and paste the factor loadings here. Are the results more interpretable? What is the correlation matrix between the factors?



The result with rotation is quite similar to the previous result without rotation, but I think this one is more interpretable since there are more large numbers (numbers close to 1) and less small numbers.

Correlation matrix here shows a more detailed overview of the correlations, for example, pay 5 is highly correlated with pay 4, suggesting a time factor here.