A Multivariate Quality Assurance Approach for Credit Card Customers and Some Features

Mian Arif Shams Adnan
Bowling Green State University, OH 43403

Abstract
Credit card is an important tool for the transaction purposes (eg. buying goods online, shopping in Walmart, paying bill in restaurants) to every individual (person or institution). Customer's age, gender, identity, income category, education level, months inactive, credit limit, total revolving balance, average opening balance to buy, total transaction amount, average utilization ratio, etc are the important features of a credit card company. Attempts have made here to demonstrate several multivariate features of credit card customers and their transactions. Some multivariate quality assurance/maintenance approaches have also been developed. These methodologies can also be applied to propose the proper derivatives for the credit card customers.

Key Words: Eigen Value, PCA, Scree Plot, Control Chart.

1. Introduction

Financial organizations like banks want to know the transaction behaviors for implementing its short run or long run derivatives. Theoretical foundations and practical knowledge of univariate and/or multivariate analyses and prophesies of the transitions of their socio-economic and financial evolution may present the entire infrastructure of the various behaviors of the interdependent socio-eco financial systems.

But cash outflow, truncation, inventories, liabilities, change of interest rate or value of money, volatility, momentums, etc. that vary over groups, companies, localities, make the overall analyses much more challenging and interesting.

As such the aim of this paper is to inaugurate a quick way of exploring a quick way of detecting a set of few components and factors that contribute the highest variation in transactions. The financial institutions want to predict these components and factors.

2. Method and Methodology

Apply the independent PCA(s) to get the regression model free from multi-collinearity and dimension reduced covariance matrix and check whether the fitted model gives a close value for the fitted coefficients of the PCA's for each new customer to asses how worthy it will be to give a new credit card. As for example, I can fit

```
m1=multinom(G~PCA1,data=D),
m2=multinom(G~PCA1+PCA2,data=D)
```

models. The dimension of the overall covariance matrix of the data from a 23 ordered square matrix is reduced to 2 ordered square matrix is as below such that V(PCA1) = 2.86=1st Eigen Value of the Correlation Matrix, V(PCA2) = 2.56 =2nd Eigen Value of the Correlation Matrix, and Cov(PCA1, PCA2) = 0.

We can reduce the dimension of the variance covariance matrix for least number of factors. Communalities and specific or unspecific variances can be obtained. Factors will help us find the least number of observations to ensure the criterion of a stratified random sample.

Generally, no distributional assumption is required to find the Principal Components and Factor Analysis of a multivariate data set. However, under large sample approximation 95% Confidence Region can also be found to check whether new customer along with the associated features fall within the elliptical region.

3. Results

Data of Credit Customer's identity, age, gender, identity, income category, education level, months inactive, credit limit, total revolving balance, average opening balance to buy, total transaction amount, average utilization ratio, etc have been collected from the public data hub (addressed as https://www.kaggle.com/sakshigoyal7/credit-card-customers).

There are **10,127 customers** and **23 variables**. So, each customer has 23 readings. Here, n = 10127, p = 23.

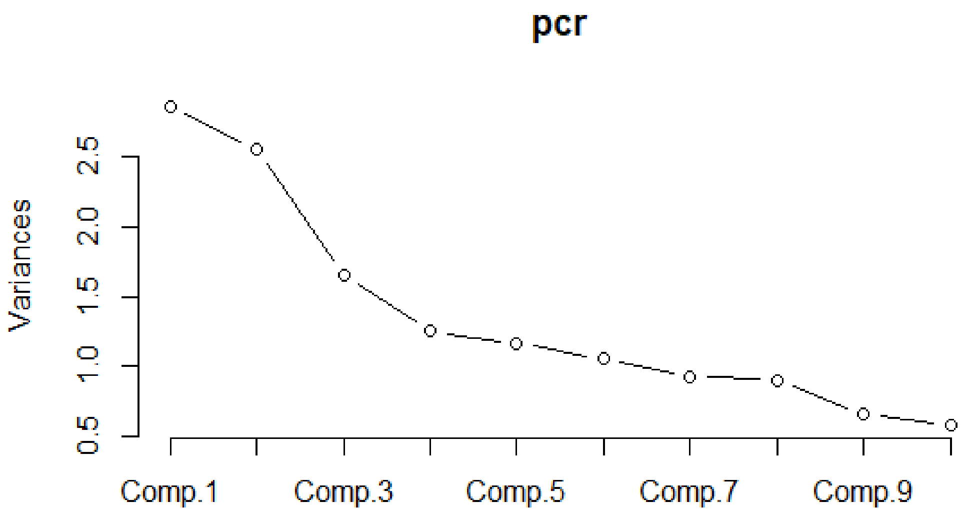

Figure 1. Scree Plot of PCAs from Sample Correlation Matrix.

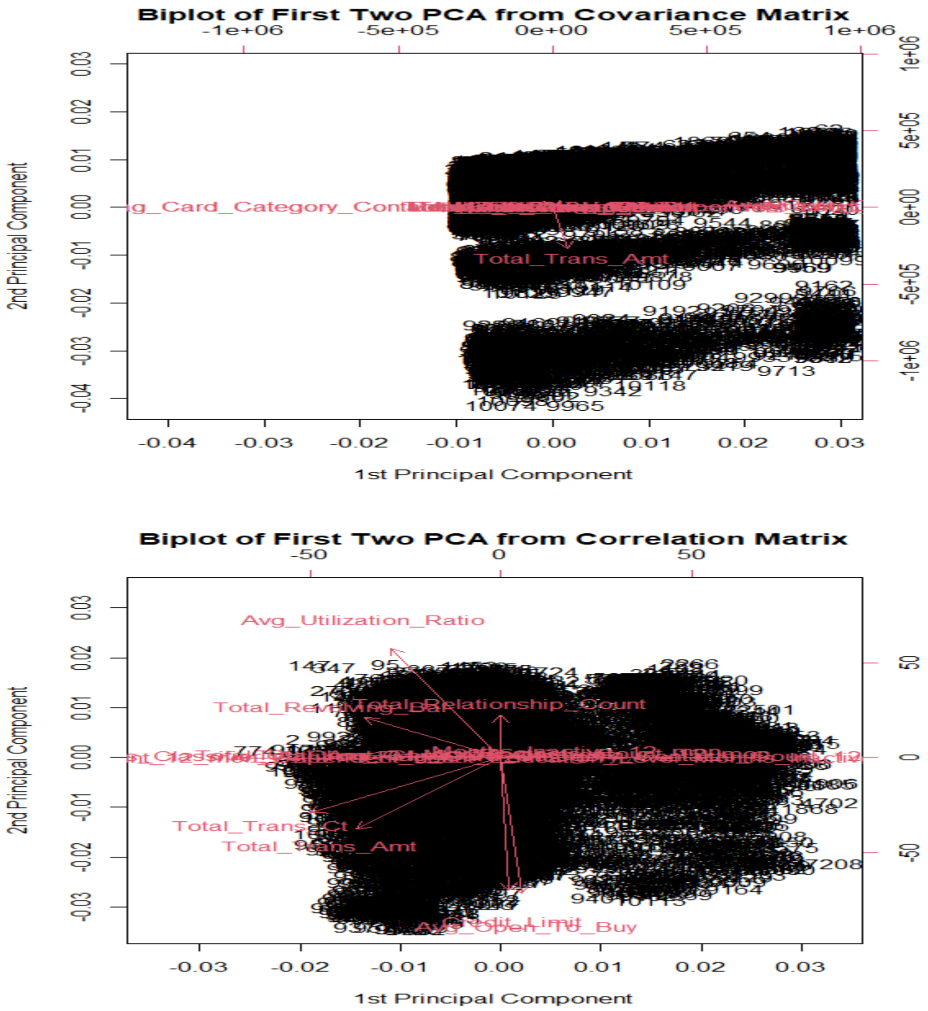

Figure 2. Biplot of PCAs.

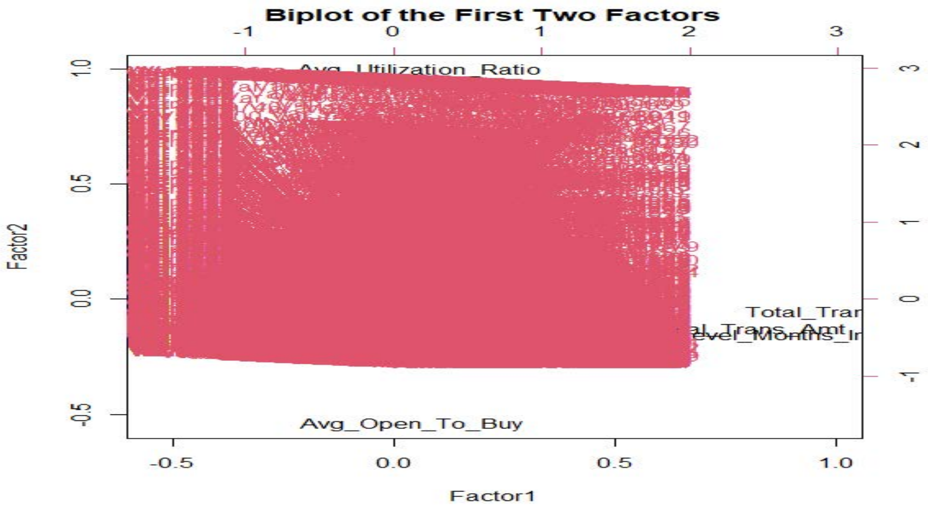

Figure 3. Biplot of Factors.

PCA1 = 0.0325**Z10** + 0.0002**Z11** + 0.118**Z12** + 0.1924**Z13** + 0.0167**Z14** + 0.263**Z15** + 0.0403**Z16** + 0.1508**Z17** – 0.2772**Z18** – 0.3674**Z19** + 0.2605**Z20** + 0.2107**Z21** + 0.5148**Z22** – 0.5149**Z23.**

Here, **Z14** = (Credit_Limit-Mean(Credit_Limit))/sd(Credit_Limit).

PCA2 = 0.0078**Z10** + 0.1729**Z11** + 0.0199**Z12** + 0.0118**Z13** -0.5424**Z14**+0.1622**Z15** – 0.5569**Z16** + 0.0069**Z17** – 0.2938**Z18** - 0.2266**Z19** + 0.0068**Z20** + 0.4484**Z21** – 0.0017**Z22** – 0.0017**Z23.**

We also have PCA3, PCA4, PCA5, PCA6, PCA7, PCA8 PCA9, PCA10, PCA11, PCA12, PCA13, PCA14. Here,

$$\mathrm{cov}(\mathrm{PCAi}, \mathrm{PCAj}) = 0.\ \forall\ i, j = 1,\ 2,\ \ldots, 14.$$

From the correlation matrix, the 1st PCA has the highest correlation (0.51 and -0.51) with the variables Naïve_Bayes_mon_1 and Naïve_Bayes_mon_2 meaning that these variables vary to the horizontal direction (direction of the 1st PC) in the right side and the left sides respectively.

The 2nd PCA has the higher negative correlation with Avg_Open_To_Buy (-0.56), Credit Limit (-0.54), Total_Trans_Amt, Total_Trans_CT, and positive correlation with Ave_Utilization_Ratio (0.45) and Total_Relationship_Count (0.17) meaning that these variables vary to the vertical direction (direction of the 2nd PC) in the right side and the left sides or the angular sides in between 1st PC and 2nd PC respectively.

From the Biplot it is evident that variables Total_Tarns_Amt and Total_Trans_Ct belong to the Factor 1(high factor loadings on horizontal Factor 1), while the Total_Revolving_Bal and Ang_Utilization_Ratio constitute vertical Factor 2. From the Biplot of figure A6, it is observed that variables of the first factor with high positive factor loadings with Total_Tarns_Amt and Total_Trans_Ct (0.810, 0.997, respectively) stay in the right side of the horizontal direction (direction of first factor).

Again, Total_Revolving_Bal variable with high positive loading (0.989) stay in the right side of the direction of 2nd factor (vertical direction). Since, Avg_Utilization_Ratio has a high positive loading (0.566) but higher negative loading in factor 3 (-0.717), it is staying in the left side of the factor 2 in the vertical direction.

Conclusion

Several other multivariate features can be obtained using PCA and Factors for the same data. R programming has been used.

References

Johnson, R. A. and Wichern, D. W. Applied Multivariate Statistical Analysis. Pearson Publisher.

B. Everitt and T. Hothorn, An Introduction to Applied Multivariate Analysis with R: Use R!,