Stat159 Fall2016 Project2

Jamie Stankiewiz, Philhoon Oh

October 30, 2016

Abstract

In this project, we will be testing alternate fit models to a data set aspiring to generate a response variable by many predictor variables that would not be feasible by regular least squares fitting.
These alternative fit models are different from linear models, like least squares, because we are now dealing with a lot more predictor variables. It is difficult to predict relationships as well as fitting because we may now deal with high colinearity, multi-colinearity, and weighting variables that correlate more or less to the response variable. The goal of this project is to explore alternative methods of model fitting.
We will be using the data set credit.csv from the book *An Introduction to Statistical Learning* by James et al. The methods, models and code used are also derived from this book.

Introduction

The goal of this project is to perform alternate regression analyses to find the best model that produces the best response variable (balance).

In this project, we will look at our data set credit and perform different regression analyses. We will first pre-process the data set to set up variables to our likings. We then shall standardize the data set and then explore credit to look for correlations between variables. From that, we can determine which variables are closely related and which are not related at all to each other and to our response variable, balance.
We then shall look at which categorical variable has a significant effect on the response variable.

Once we have thoroughly analyzed the data set properties and variables, we can go ahead and run our 5 picked alternative regression models: OLS(Ordinary Least Squares), RR(ridge regression), LR(lasso regression), PCR(principle components regression) and PLSR(partial least squares regression).

By looking at the MSE(mean squared error) of which each model produced, we can determine which model generated the best regression, and the best fit for our response variable.
We will explore all of these concepts in our project.

Data

We first start out with the original data set: credit.csv from the book *An Introduction to Statistical Learning* by James et al. We stored this data set in R as the variable, credit. Credit initally contains 12 columns(categories or predictors). However, we remove the first column because it just contains the row numbers of each row, which is irrelevant to our analysis, so we will remember to subset that part out of the data set in R before doing analysis.
There are two important steps in premodeling data proces.

1. convert factors into dummy variables
2. mean centering and standardization

First, we need to factors into dummy variables. We pre-process the data set to make sure that it is in the format that we want. Notable changes are changing categorical variables (such as: M/F, Yes/No,

Caucasian/Asian) into 0/1 slots.
Second, we need to standardize each predictor variables to apply 5 regression linear methods.

The first 6 results of the standardized data set are shown in Table 1 in the Tables and Figures Sections.

With this data, we are going to apply five regression methods: OLS(Ordinary Least Squares), RR(ridge regression), LR(lasso regression), PCR(principle components regression) and PLSR(partial least squares regression).

Method

For the project, we will analyze the given data with five different approaches: OLS(Ordinary Least Squares), RR(ridge regression), LR(lasso regression), PCR(principle components regression) and PLSR(partial least squares regression).

1) Ordinary Least Square Method(OLS)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p$$

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

Ordinary least squares(OLS) is estimataing parameters in a linear regression model by minimizing the sum of squares of the distance between observed responses and predicted responses. This approach is perfomed under certain assumptions: errors are finite, homoscedastic, uncorrelated, and normally distributed. Under these assumption OLS provides the minimum-variance and unbiased mean estimators.

Shrinkage Methods

Shrinkage methods starts with fitting a modeling with all p predictors. However, it forces the estimated coefficients to be close to zero or to be zero. Also, it requires the data to be standardized before applying the method. There are two methods of shrinkage methods: ridge regression method, and lasso regression method. Depending on the method, some of the cofficients forced to be estimated to be exactly zero, which is the case of lasso regression. Thus, shrinkage method does variable selection.

2) Ridge Regression(RR)

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$minimize_\beta \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \} \quad subject\ to \quad \lambda \sum_{j=1}^{p} \beta_j^2 \leq s$$

where every value of λ *there is a corresponding* s

Equations above are two different representations of Ridge Regression. The first equation explains the similarity between OLS regression and Ridge Regression. Ridge regression calculates the coefficients by

minimizing $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$. If λ, the tunning parameter, is equal to zero, the estimated coefficients will be same as those of OLS regression.

The advantage of ridge regression over OLS regression is bias-variance trade-off. As λ increases, the estimated coefficients forced to be close to zero, leading bias estimators and decreasing the variance. As λ decreases, the estimated coefficients do not have to be close to zero, leading less - bias estimators and increasing the variance.

3) Lasso Regression(LR)

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

$$minimize_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\} \quad subject\ to \quad \lambda \sum_{j=1}^{p}|\beta_j| \leq s$$

where every value of λ there is a corresponding s

Equations above are two different representations of lasso regression. The first equation explains the similarity between OLS regression, ridge regression and lasso regression. Lasso regression calculates the coefficients by minimizing $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$. If λ, the tunning parameter, is equal to zero, the estimated coefficients will be same as those of OLS regression.

Lasso regression has the same advantage over OLS regression as Ridge regression does: bias-variance trade-off. However, as λ increases, the lasso regression forces esimates to be exactly equal to zero. Thus, lasso shrinkage method performs the variable selection and a model generated by lasso regression is easier to interpret than a model generated by ridge regression. Lasso regression involves only a subset of the variables referred as sparse-model.

Dimension Reduction Methods

$$Z_m = \sum_{j=1}^{p}\phi_{jm}X_j$$

$$where \quad \sum_{j=1}^{p}\phi_{jm}^2 \ = \ 1$$

$$y_i = \theta_0 + \sum_{m=1}^{M}\theta_m z_{im} + \varepsilon_i \quad i \ = \ 1, ..., n$$

$$\sum_{m=1}^{M}\theta_m z_{im} \ = \ \sum_{m=1}^{M}\theta_m \sum_{j=1}^{p}\phi_{jm}x_{ij} \ = \ \sum_{j=1}^{p}\sum_{m=1}^{M}\theta_m\phi_{jm}x_{ij} \ = \ \sum_{j=1}^{p}\beta_j x_{ij}$$

$$where \quad \beta_j \ = \ \sum_{m=1}^{M}\theta_m\phi_{jm}$$

Dimension reduction method is transforming the predictors,
$Z_m = \sum_{j=1}^{p}\phi_{jm}X_j$ and fit a least square models using transformed variables,
$y_i = \theta_0 + \sum_{m=1}^{M}\theta_m z_{im} + \varepsilon_i \quad i \ = \ 1, ..., n$. This approach is used when there is a large number of predictor variables. By computing M different linear combinations, transforming the predictors, we can project p

predictors into a M-dimensional subspace where M is less than or equal to p. Then M projections can be used as predictors to fit a linear model by ordinary least squares.

If M is equal to p, there is no dimension reduction occurs, yielding the same result equivalent to performing ordinary least squares on the p predictors.

There are two major dimension reduction methods: principal components regression and partial least squares regression. Those methods work in two steps. First step is transforming the predictors,
$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$. Second steop is selecting ϕ_{jm}.

4) Principle Components Regression(PCR)

Principle Components Regression is based on the principal component analysis. Principal components analysis extracts the most variation of datasets to reduce the dimension of the data.

Find the first principal component Z_1

$$Z_1 \ = \ \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{p1} X_p$$

Z_1 is the first factor maximizing the dispersion of obervations.

Then find the second principal component Z_2

$$Z_2 \ = \ \phi_{12} X_1 + \phi_{22} X_2 + ... + \phi_{p2} X_p$$

Z_2 is the second factor maximizing the dispersion of observations as well as orthogonal to the first factor.

By doing this process iteratively up to desired M, it gives us the Principal Components Regression on a given M space.

Principle Components Regression shows a linear regression between the data on the factors that are correlated. It enables us to solve the colinearity between predictor variables and dimensionality of the given data. That is, the number of samples, n, does not necessarily bigger than the number of predictor variables.

5) Partial Least Squares Regression(PLSR)

Like Principle Component Regression, Partial Least Square Regression calculuates $Z_1, .., Z_m$ and fit a linear model on those transformed predictors. However, Partial Least Square Regression takes into account the structure of both the explanatory and dependent variable unlike Principle Component Regression. That being said, it uses response variable Y to find directions that help explaining both the response and the predictors.

Find the first direction Z_1

$$Z_1 \ = \ \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{p1} X_p$$

$$\textit{by setting } \phi_{j1} \ = \ \beta_j \quad \textit{where } \beta j \textit{ is from simple linear regression } Y \textit{ onto } X_j \textit{ and } j = 1, ..., p$$

The first direction Z_1 consequently places the weights on each predictor variables proportional to its correlation with the resoponse variable.

To find the second direction Z_2, we need to regressing each variable on Z_1 and take the residuals. Then compute the Z_2 with the residuals that are orthogonal to the original data by applying the same procedure to compute the first directon Z_1

By doing this process iteratively up to desired M, it gives us the Partial Least Squares Regression on a given M space.

Analysis

In analysis part, we will see the results of the each methods we applied to a given data set.

1) Ordinary Least Square Method(OLS)

There are several steps in applying the Ordinary Least Square method.

STEP1. Using the result from F-statistics, ruling out the qualitative variables that is not significant in identifying the response variable. As you can see in the anova Tables 2-5, only student variable is signifant in evaluating the response variable balance. Thus, we are ruling out ethnicity, gender, and marriage variables.

STEP2. Apply the OLS on our training dataset on selected predictor variables: income, limit, rating, cards, ages, eduction, student as shown in Table 5. This also enables us to identify the optimal predictors to predict the balance: income, limit, cards, students in Table 6.

STEP3. Using the result from step2, apply linear regression model on opitmal predictor variables to estimate the balance: income, limit, cards, student shown in Table 7. Figure 1 shows the residual plot of OLS on test dataset.

STEP4. To identify the best OLS model. we are going to compare the adjusted r-squared values of both linear models. Model1, using selected predictor yields the 0.95371. Model2 using optimal predictor yields 0.95366. Therefore, the second model using optimal predictors is the best OLS model because it uses few variables but produces almost same adjusted r-squared value as the first model as shown in Table 8. Lastly, using the best model we found, apply it to full data set to the "official coefficients" as shown in the Table 9.

2) Ridge Regression(RR)

First, using ten-fold cross-validation, ridge regression selects the best model when $\lambda = 0.01$ and cross-validation plot is shown in Figure 2.

After finding out the best model, applying the best model to our test set to compute the test mean squared error (MSE), which is 0.0479811, later use this figure to compare different method in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients" as shown in the Table 9.

3) Lasso Regression(LR)

Lasso regression does exactly same procedure as ridge regression. It selects the best model when $\lambda = 0.01$ and cross-validation plot is shown in Figure 3.

After finding out the best model, applying the best model to our test set to compute the test mean squared error, which is 0.0515181, later use this figure to compare different method in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients" as shown in the Table 9.

4) Principle Components Regression(PCR)

Principle Component Regression selecst the best model when $M = 11$ where M represents the compression, reduced dimension and cross-validation plot is shown in Figure 4.

After finding out the best model, applying the best model to our test set to compute the test mean squared error, which is 0.0474722, later use this figure to compare different methods in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients" as shown in the Table 9.

5) Partial Least Squares Regression(PLSR)

Principle Component Regression selects the best model when $M = 8$ where M represents the compression, reduced dimension and cross-validation plot is shown in Figure 5.

After finding out the best model, applying the best model to our test set to compute the test mean squared error, which is 0.0464641, later use this figure to compare different methods in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients" as shown in the Table 9.

Results

From analyzing each regression model, we were able to come up with the TEST MSE value. By comparing them, we can determine which method produced the smallest TEST MSE value, and therefore the best regression model.

Table of TEST MSE values of each regression is shown in Table 10.

According to the Table 10, PLSR has the smallest TEST MSE is 0.0464641.

Therefore, we can conclude that among all the regression methods we considered, PLSR yields the smallest TEST MSE and thus is the best model for our given dataset.

Lastly, we are going to compare the each official coefficients of regression methods. By plotting trend lines of the official coefficients. We can easily see that limit has the most influential aspect on balance on all regression method. Trend line plot is shown in Figure 6

Conclusion

Among the 5 regression methods, it turns out that Partial Least Squares Regression(PLSR) yields the best result for predicting reponse variable. The advantage of PLSR over PCR is that it takes account into the direction of the response variable as well as reducing the dimension. Thus, we actually need the subset of transformed predictor variables. This enables us to find the best model where number of observations is smaller than predictor variables. Again, below Table 11 shows the official coefficients of PLSR when the compression is equal to 8 and Test MSE is equal to 0.0464641.

Tables and Figures

income	limit	rating	cards	age
-0.8605	-0.4894	-0.4650	-0.6983	-1.2561
1.7253	0.8272	0.8277	0.0310	1.5265
1.6846	1.0135	1.0280	0.7602	0.8889
2.9425	2.0659	2.1074	0.0310	-1.1402
0.3025	0.0699	0.0133	-0.6983	0.7149
0.9920	1.4346	1.3835	0.7602	1.2367

education	gender	student	marriage	asian	caucasian	balance
-0.7839	-1.0343	-0.3329	0.7944	-0.5843	1.0038	-0.4068
0.4960	0.9644	2.9962	0.7944	1.7071	-0.9938	0.8330
-0.7839	-1.0343	-0.3329	-1.2557	1.7071	-0.9938	0.1305
-0.7839	0.9644	-0.3329	-1.2557	1.7071	-0.9938	0.9657
0.8159	-1.0343	-0.3329	0.7944	-0.5843	1.0038	-0.4111
-1.1039	-1.0343	-0.3329	-1.2557	-0.5843	1.0038	1.3724

Table 1: First 6 results of Premodeling data processing results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ethnicity	2.00	18454.20	9227.10	0.04	0.96
Residuals	397.00	84321457.71	212396.62		

Table 2: Anova analysis of ethnicity and Balance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	1.00	38891.91	38891.91	0.18	0.67
Residuals	398.00	84301020.00	211811.61		

Table 3: Anova analysis of gender and Balance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
marriage	1.00	2714.77	2714.77	0.01	0.91
Residuals	398.00	84337197.14	211902.51		

Table 4: Anova analysis of marriage and Balance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
student	1.00	5658372.27	5658372.27	28.62	0.00
Residuals	398.00	78681539.64	197692.31		

Table 5: Anova analysis of student and Balance

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0021	0.0124	-0.1680	0.8667
income	-0.5940	0.0211	-28.1295	0.0000
limit	1.0702	0.1919	5.5779	0.0000
rating	0.2676	0.1923	1.3917	0.1651
cards	0.0538	0.0150	3.5954	0.0004
age	-0.0078	0.0127	-0.6174	0.5374
education	-0.0115	0.0127	-0.9051	0.3662
student	0.2764	0.0129	21.4850	0.0000

Table 6: MODEL1 using selected predictor variables: income, limit, rating, cards, ages, eduction, student

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0021	0.0124	-0.1676	0.8670
income	-0.5946	0.0210	-28.3296	0.0000
limit	1.3361	0.0203	65.9319	0.0000
cards	0.0666	0.0122	5.4648	0.0000
student	0.2764	0.0128	21.5288	0.0000

Table 7: MODEL2 using Optimal predictor variables: income, limitm cards, student

	Adjusted.R.Square	Value
1	Model1	0.9537
2	Model2	0.9537

Table 8: Adjusted R Square of Model1 and Model2

	OLS	RR	LASSO	PCR	PLSR
(Intercept)	0.0000	0.0000	0.0000	0.0000	0.0000
income	-0.6009	-0.5687	-0.5517	-0.5982	-0.5983
limit	1.3387	0.7187	0.9250	0.9584	0.8525
rating	0.0000	0.5931	0.3679	0.3825	0.4874
cards	0.0000	0.0443	0.0450	0.0529	0.0478
age	0.0000	-0.0254	-0.0167	-0.0230	-0.0192
education	0.0000	-0.0059	0.0000	-0.0075	-0.0131
gender	0.0000	-0.0107	0.0000	-0.0116	-0.0108
student	0.2807	0.2732	0.2668	0.2782	0.2820
marriage	0.0000	-0.0110	0.0000	-0.0091	-0.0058
asian	0.0000	0.0164	0.0000	0.0160	0.0148
caucasian	0.0000	0.0110	0.0000	0.0110	0.0095

Table 9: Official Coefficients of all Regression Methods

RegressionMethods	MSE
OLS	0.0501
Ridge	0.0480
Lasso	0.0515
PCR	0.0475
PLSR	0.0465

Table 10: Test MSE of all regression methods

(Intercept)	income	limit	rating	cards
0.0000	-0.5983	0.8525	0.4874	0.0478

age	education	gender	student	marriage	asian	caucasian
-0.0192	-0.0131	-0.0108	0.2820	-0.0058	0.0148	0.0095

Table 11: The official coefficients of best model(PLSR Regression)

Figure 1: Residuals plot of OLS on the test dataset

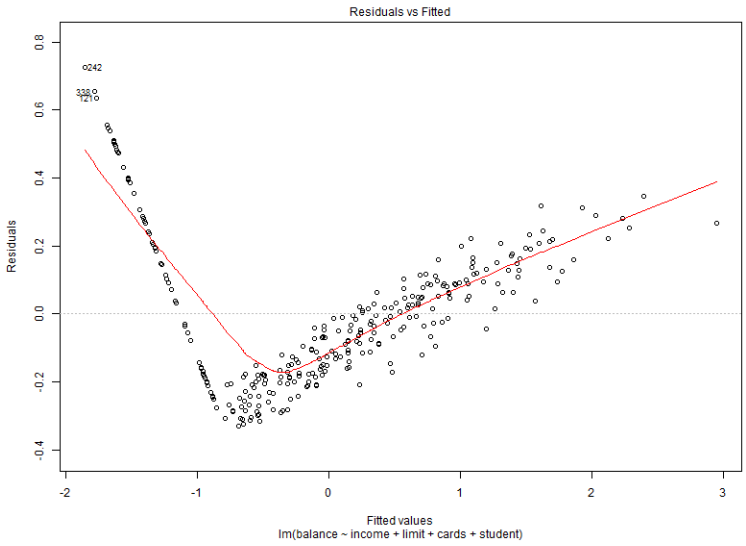

Figure 2: Plot of the cross-validtion errors of Ridge Regression on Lambda

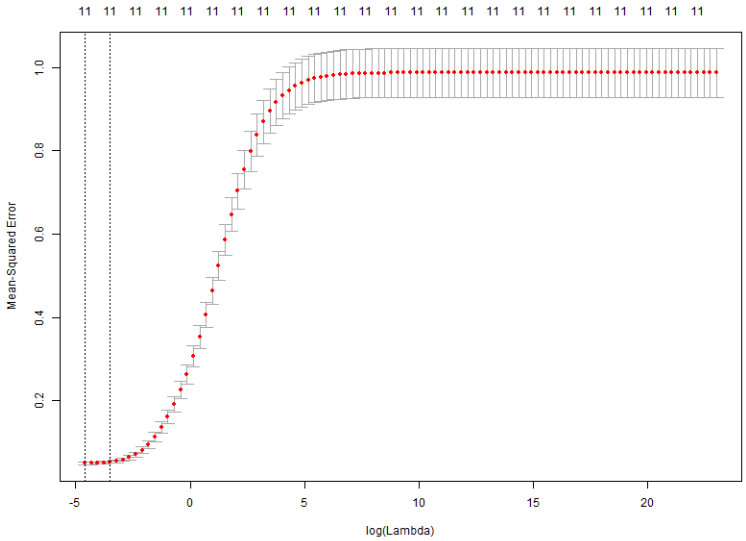

Figure 3: Plot of the cross-validtion errors of Lasso Regression on Lambda

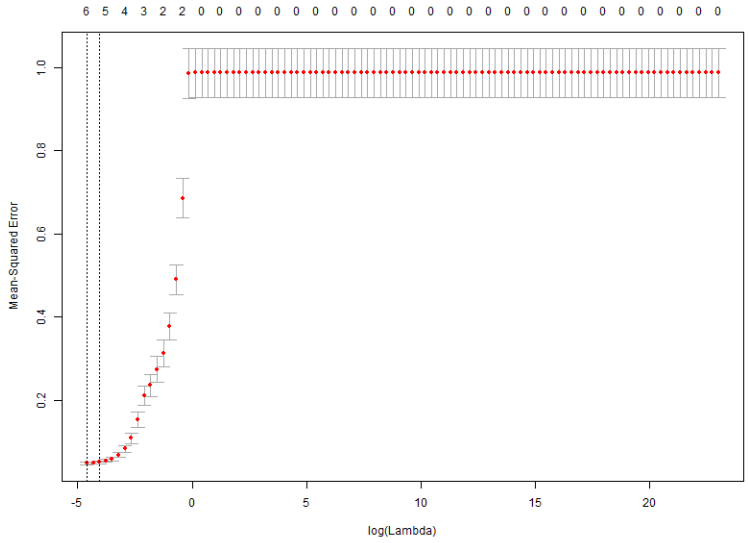

Figure 4: Plot of the cross-validtion errors of PCR Regression on M

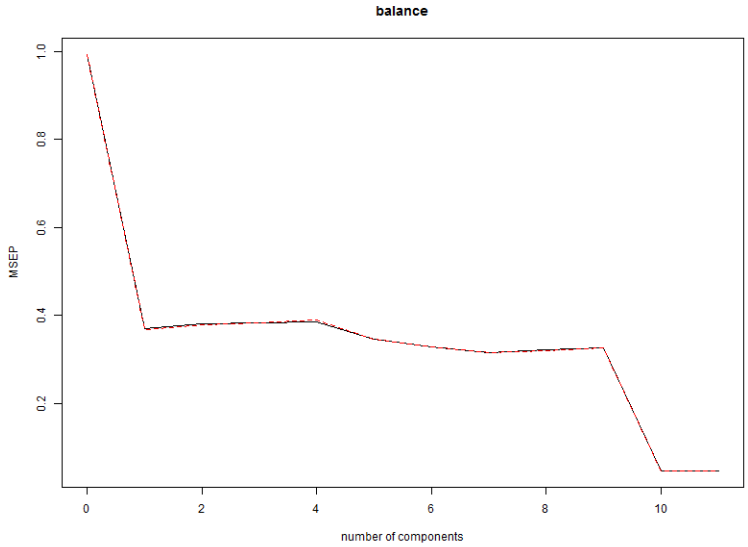

Figure 5: Plot of the cross-validtion errors of PLSR Regression on M

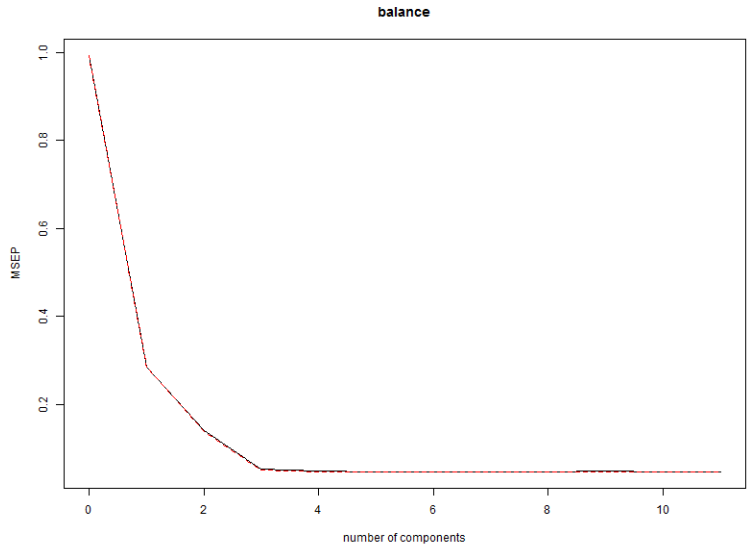

Figure 6: Plot of trend line of the coefficients on each methods