High Dimensional Regression

Chenghui Li, Haoxiang Wei, Mufang Ying, Qintao Ying

May 1, 2019

Group 4

01

Problem Description

Gene Prediction

Now, suppose we want to predict micro -organism survival time Y from a set of gene expression measurements from DNA $X_1, X_2, \ldots, X_p$. Now we have N individuals.

Because DNA is so complicated that it may contain numerous measurements. Here we can assume p $\gg N$.

High Dimensional Situation

Now the number of predictors greatly exceeds the number of observations , conventional regression techniques may produce unsatisfactory results.

High Dimensional Situation

This is a common situation for regression analysis. In order to deal with it generally, here we tried four methods:

- PCA
- PLS
- LASSO
- SPCA

02

Data Simulation

Data Simulation

First, we simulated two dataset. The first one is:

$$X_{ij} = \begin{cases} 3 + \varepsilon_{ij} & i \leq 50, j \leq 50 \\ 4 + \varepsilon_{ij} & i \leq 50, j > 50 \\ 3.5 + \varepsilon_{ij} & i > 50, j > 50 \end{cases}, \qquad y_j = \frac{\sum_{i=1}^{50} X_{ij}}{25} + \varepsilon_{ij}, i \leq 5000, j \leq 100$$

Let X_{ij} denotes the i-th gene of the j-th micro-organism, where the ε_{ij} s are independent normal random variables with mean 0 and standard deviation 1.5.

Data Simulation

The second one is:

$$X_{ij} = \begin{cases} 3 + \varepsilon_{ij} & i \leq 50, j \leq 50 \\ 4 + \varepsilon_{ij} & i \leq 50, j > 50 \\ 3.5 + 1.5I\left(u_{1j} < 0.4\right) + \varepsilon_{ij} & 50 < i \leq 100 \\ 3.5 + 0.5I\left(u_{2j} < 0.7\right) + \varepsilon_{ij} & 100 < i \leq 200 \\ 3.5 - 1.5I\left(u_{1j} < 0.3\right) + \varepsilon_{ij} & 200 < i \leq 300 \\ 3.5 + \varepsilon_{ij} & i > 300 \end{cases}, i \leq 5000, j \leq 100$$

Here the u_{ij} are uniform random variables on (0,1) and $I(X)$ is an indicator function.

03

PCA

PCA

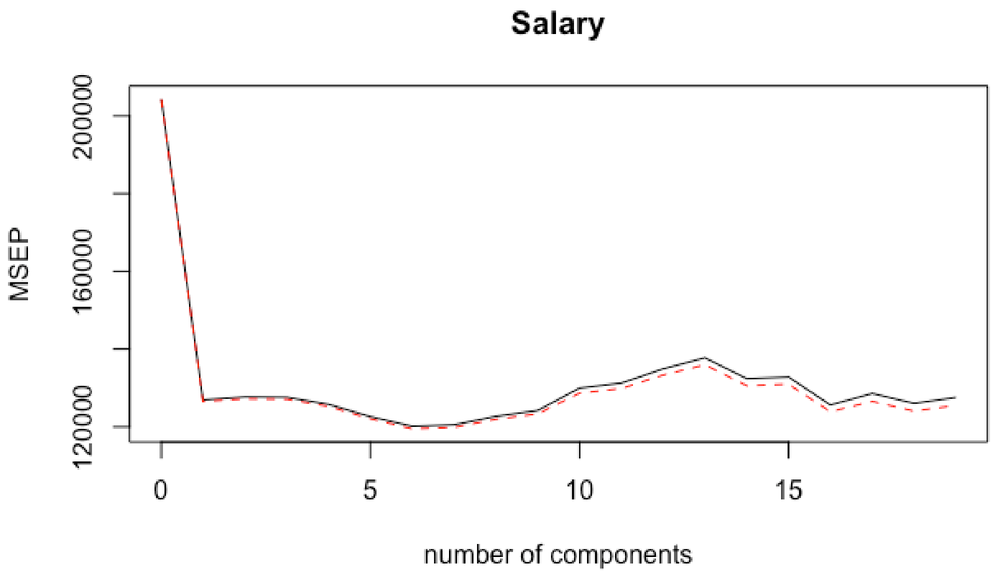
Procedure of picking the number of components

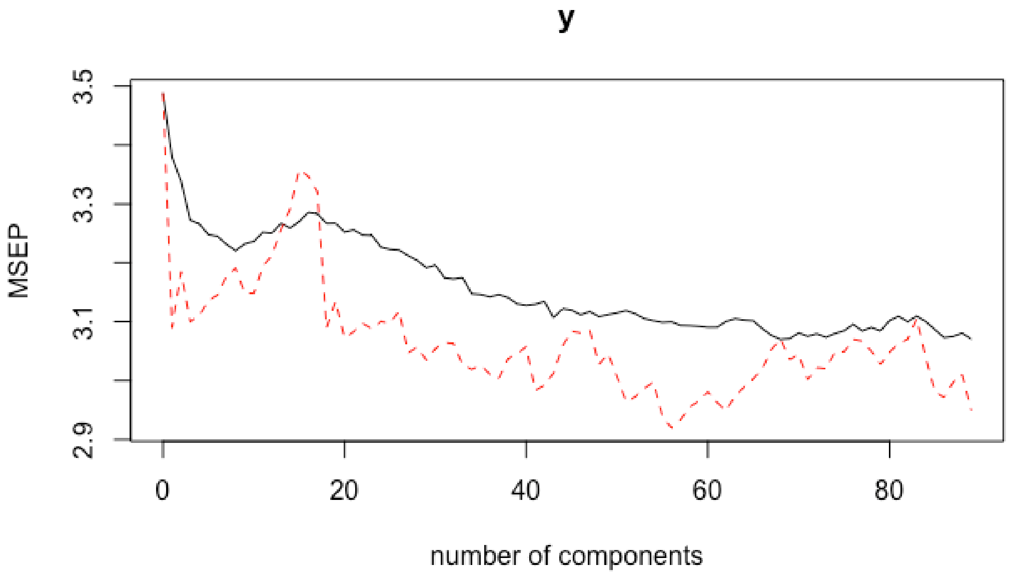
1.Plot the MSEP vs number of components

2.Check the variances explained by the components

3.Determine the number of components

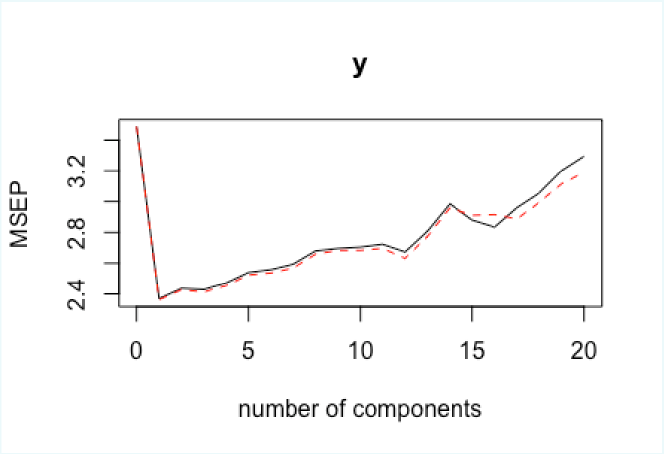
The optimal number of components here is not so obvious.

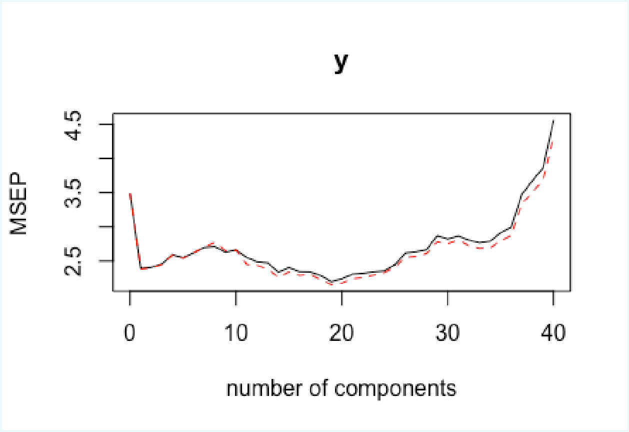


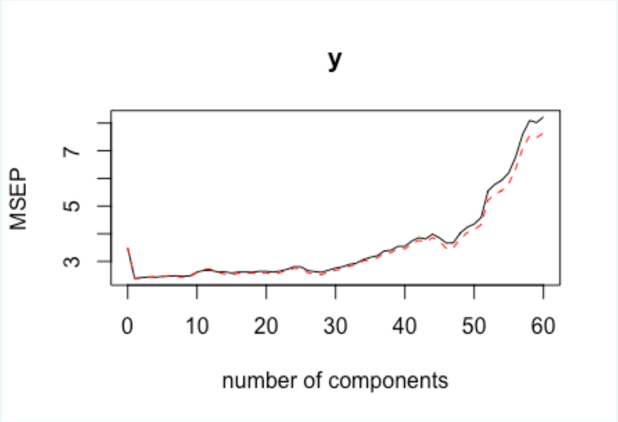


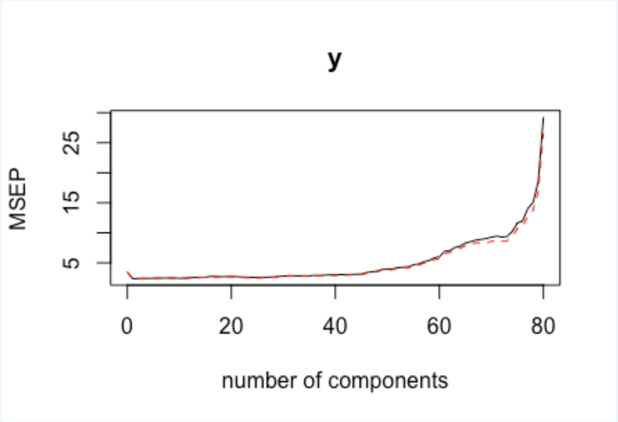
Experiment

I use the same dataset but only include 20,40,60,80 columns respectively.









Main drawback

(1)

The principle components are not related to the response variable.

(2)

Sometimes hard to identify the optimal number of components when dataset is complicatded.

Result

CV Error

One component: 337.8
8 components: 322.2
40 components: 312.9
81 components: 310.8

Test Error

One component: 304.4
8 components: 304.1
40 components: 286.1
81 components: 298.6

04

PLS Analysis

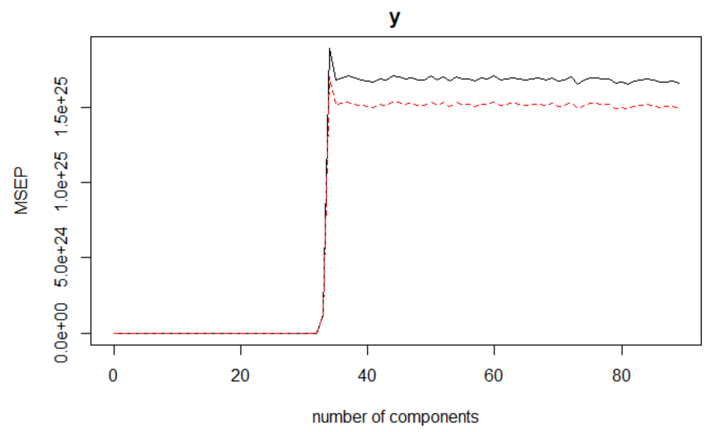
PLS (Partial Least Square)

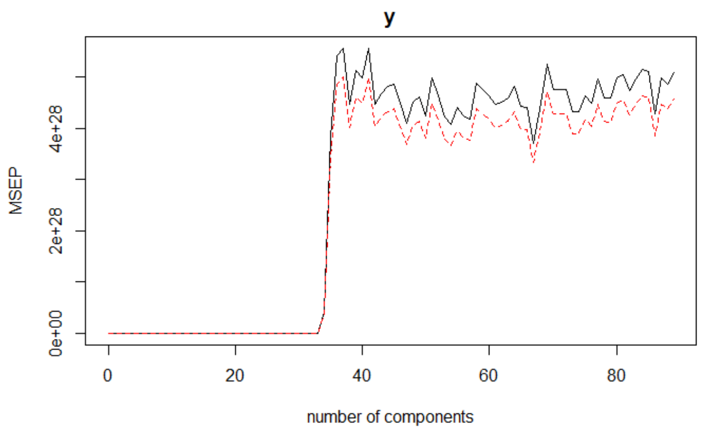
PLS working process:

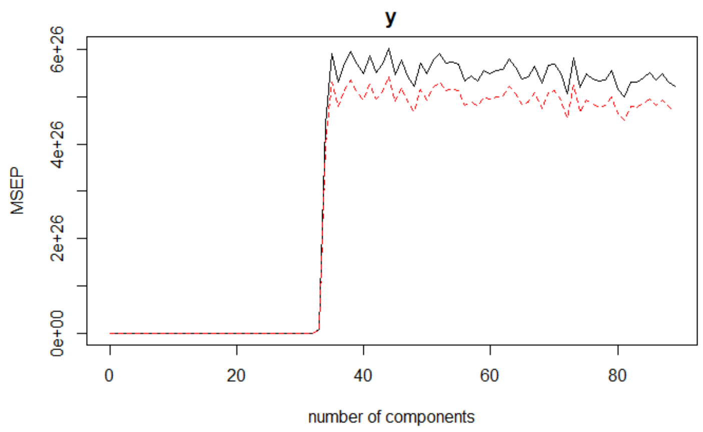
1.Standardize each of the variables to have mean 0 and unit norm, and compute the univariate regression coefficients $\mathbf{w} = \mathbf{X}^T\mathbf{y}.$

2.Define $\mathbf{u}_{\mathrm{PLS}} = \mathbf{X}\mathbf{w}$, and use it in a linear regression model with y.

3.Find **w** that $\max_{\|\mathbf{w}\|=1} \operatorname{corr}^2(\mathbf{y}, \mathbf{X}\mathbf{w}) \operatorname{var}(\mathbf{X}\mathbf{w}),$







Advantages and disadvantages of PLS

Advantages

(1)

PLS can explain the relationship between x&y and decrease multicollinearity simultaneously.

(2)

In theory, PLS performs better than PCA in high dimensional regression cases.

Disadvantages

(1)

Sometimes hard to find the optimal number of components.

(2)

Can be influenced by the noise in the unimportant features and maintain them in components.

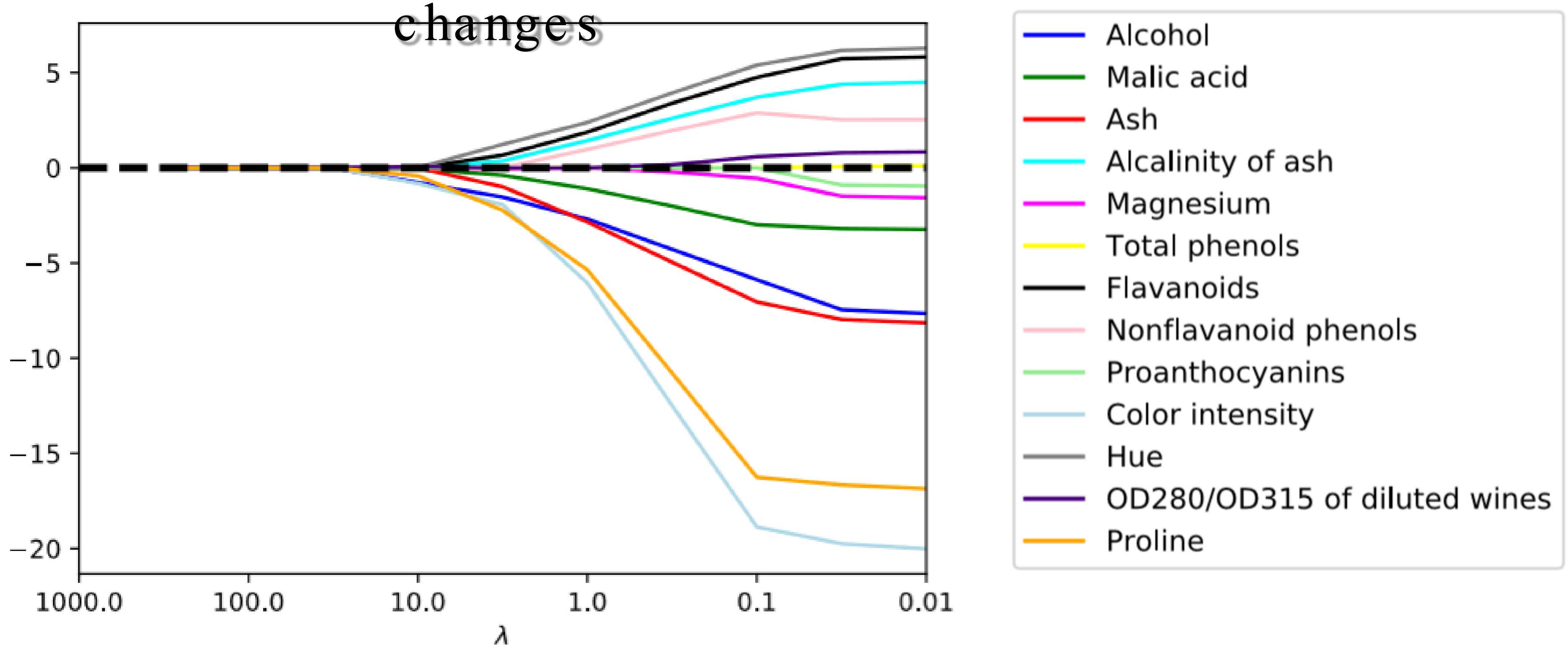
05

LASSO

LASSO

$$\min_{\beta} \|\mathbf{y} - \beta_0 - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

Coefficients change when λ changes



Advantages and disadvantages of LASSO

Advantages

(1)

LASSO decrease multicollinearity of the features.

(2)

Interpretive. Select useful features in the model and exclude noise features.

Disadvantages

(1)

The choose of λ depends on data, and it is not computational efficient.

(2)

In high dimensional case ($n < p$), LASSO can at most select n features.

06

SPCA

Algorithm

1. Correlation

2. Shreshold

3. PCA Model Fitting

4. Prediction□

Example

#component	MSE_p	#variable
1	2.37	9
2	2.03	56
3	1.94	56

Latent model:

$$y_j = \frac{\sum_{i=1}^{50} x_{ij}}{25} + \epsilon_j, \epsilon_j \sim N(0,2.25)$$

Good
prediction

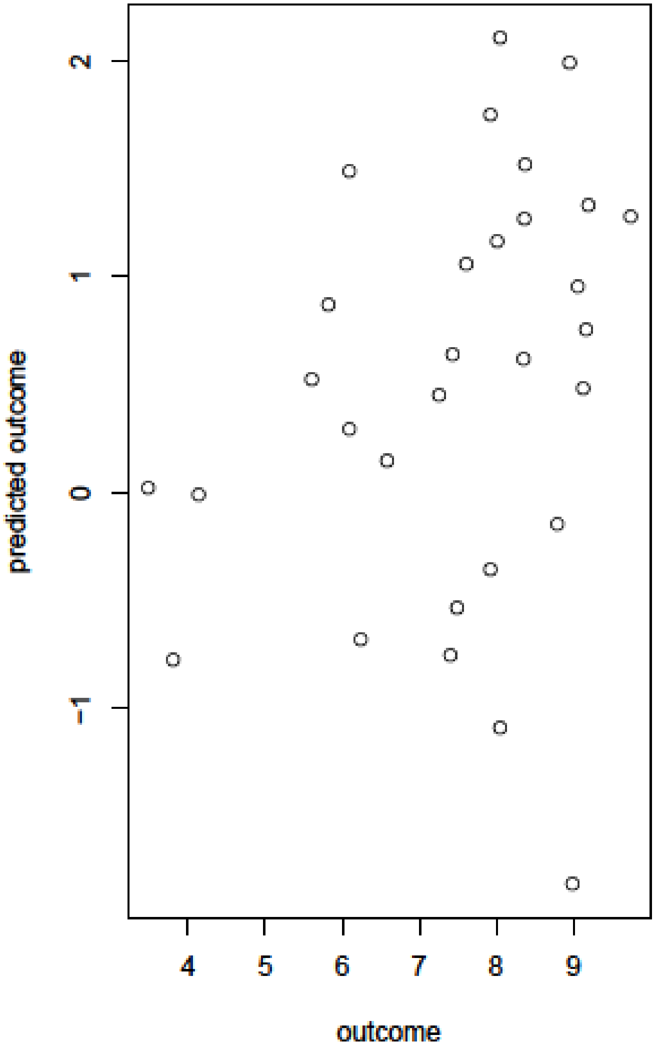
MSE
2.25

Variable and
Component

56
variables!

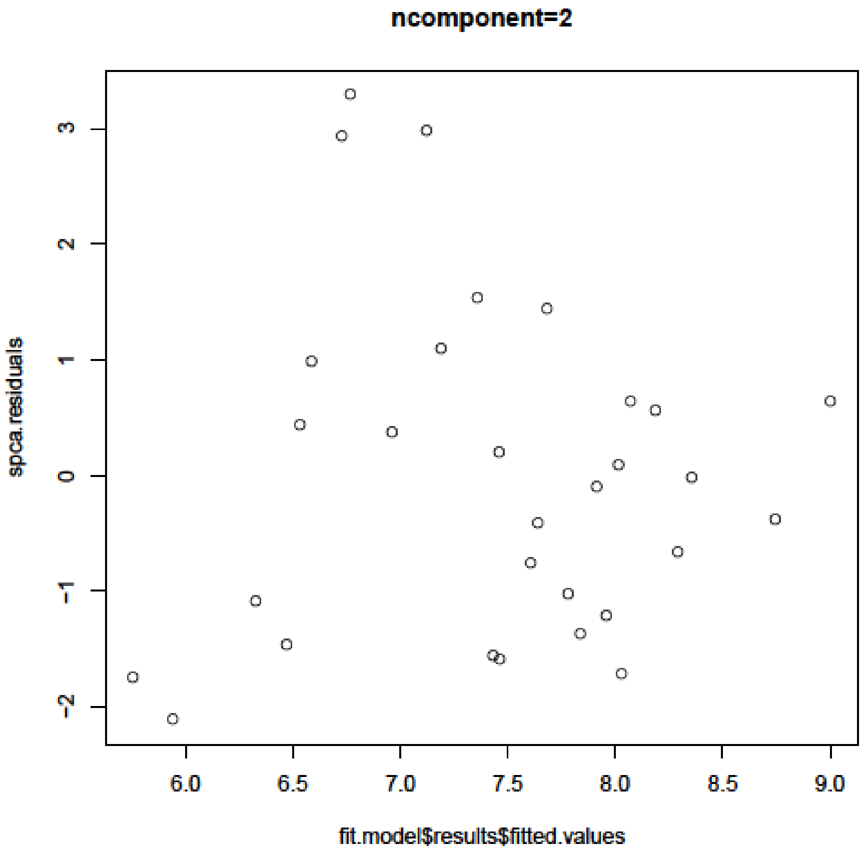
Residuals plot

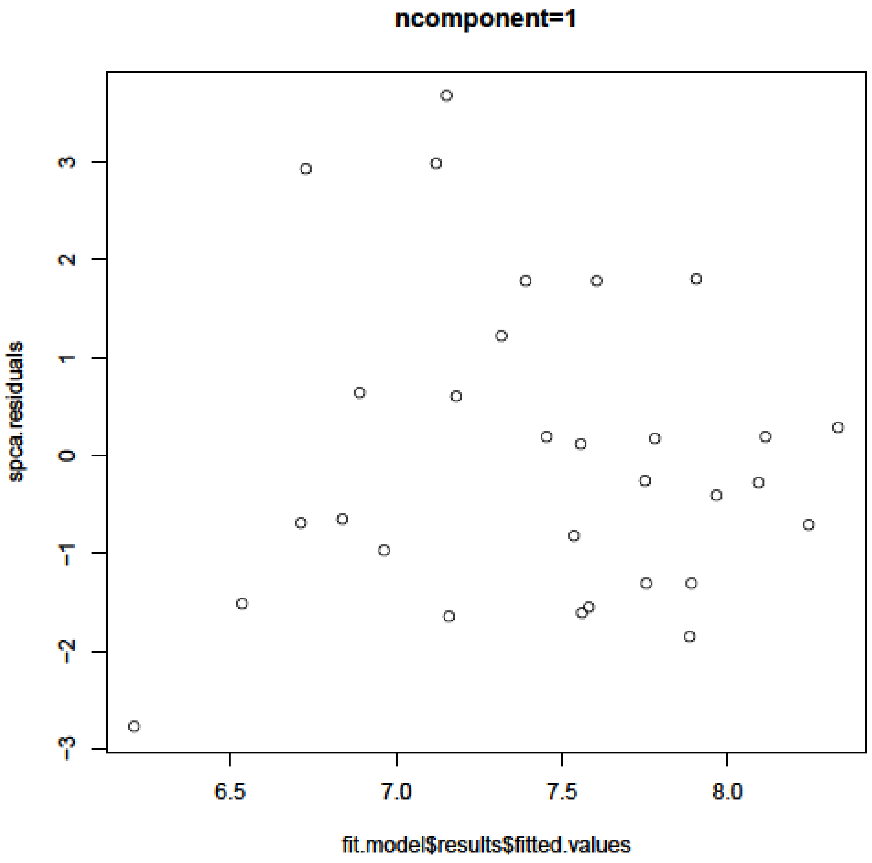
3 components



	Intcpt	Comp1	Comp2	Comp3
coef	7.191	1.249	−3.508	−1.827
se	0.308	1.174	1.497	1.689
T stat	23.376	1.064	−2.344	−1.081
pvalue	0	0.297	0.027	0.289
F−stat		2.38117		
pvalue		0.13403		

Residuals plot





07

Results

Algorithm comparison

Method	CV error	Test error
PCA	312.9356	286.1653
LASSO	169.8364	290.4857
PLS	340.0336	301.5314
SPCA	145.0775	237.2667

Thank you!