The NIR corn data set

Hongwei PENG

Supervisor: Prof. Tom Fearn

Department of statistical science

University College London

September 3, 2019

Datasets

- One of the most easily accessible public data sets
- high-dimensional NIR spectroscopic data
- corn dataset is available in the website: http://www.eigenvector.com/data/Corn/index.html

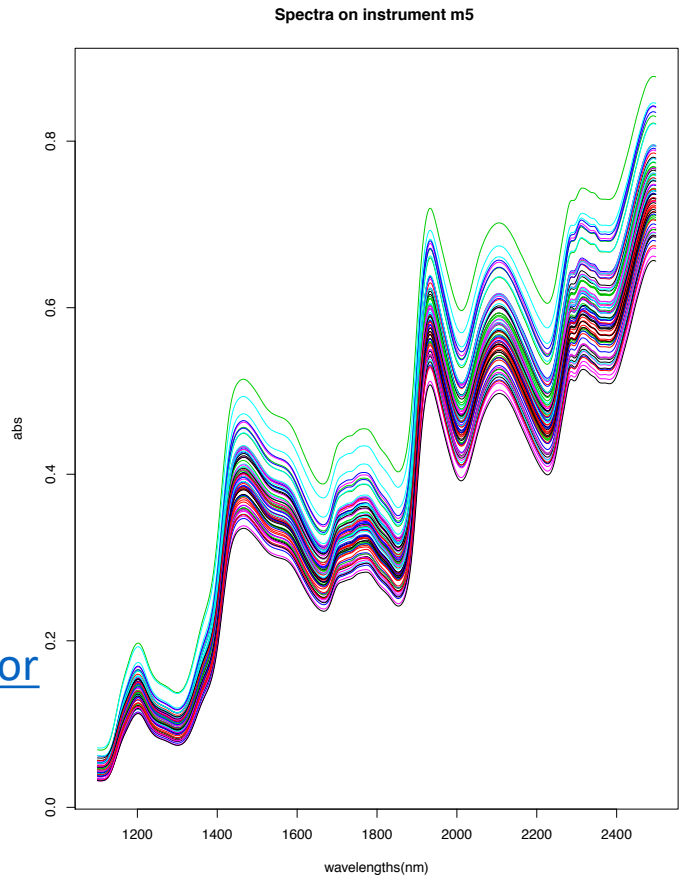
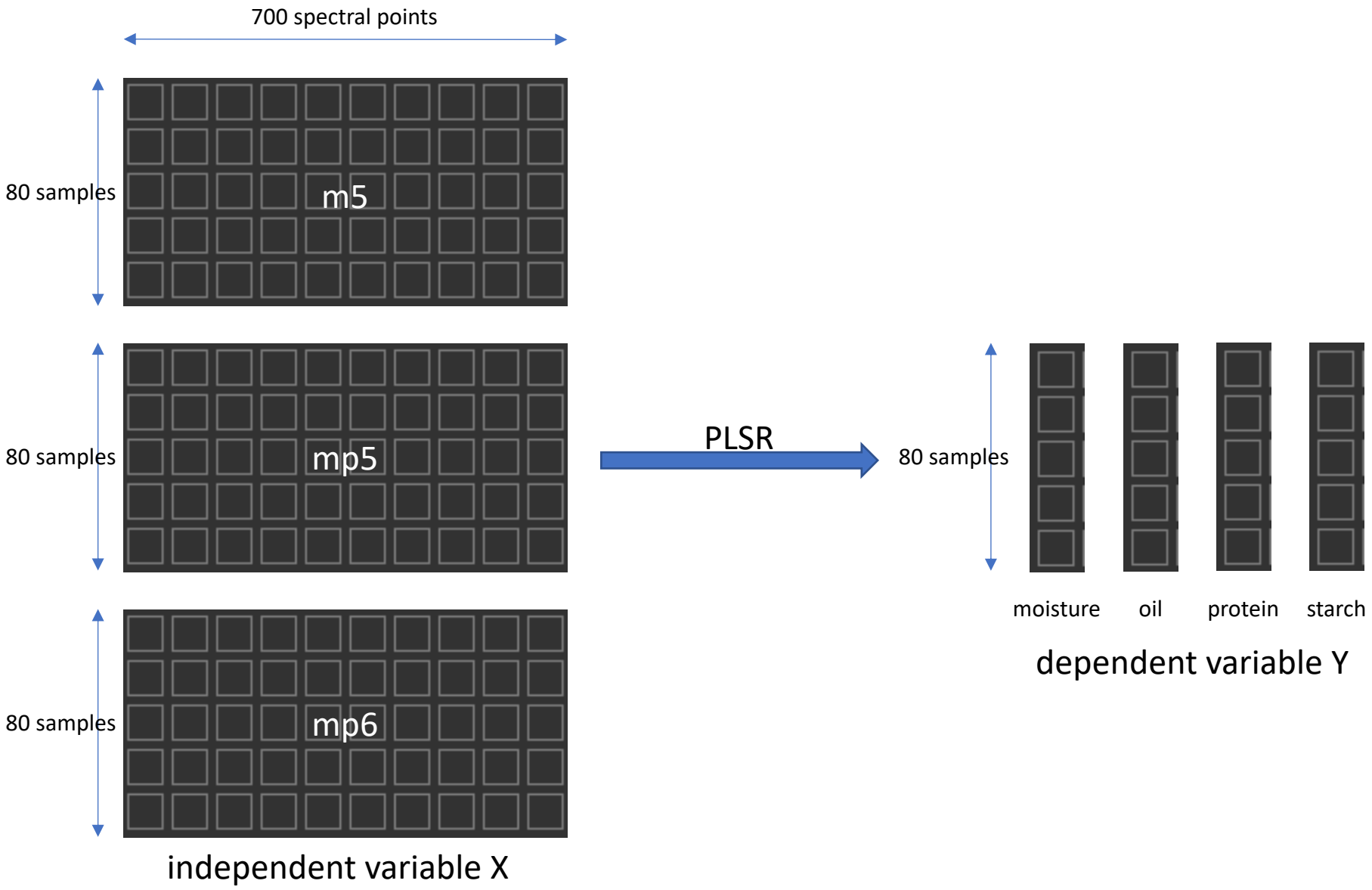


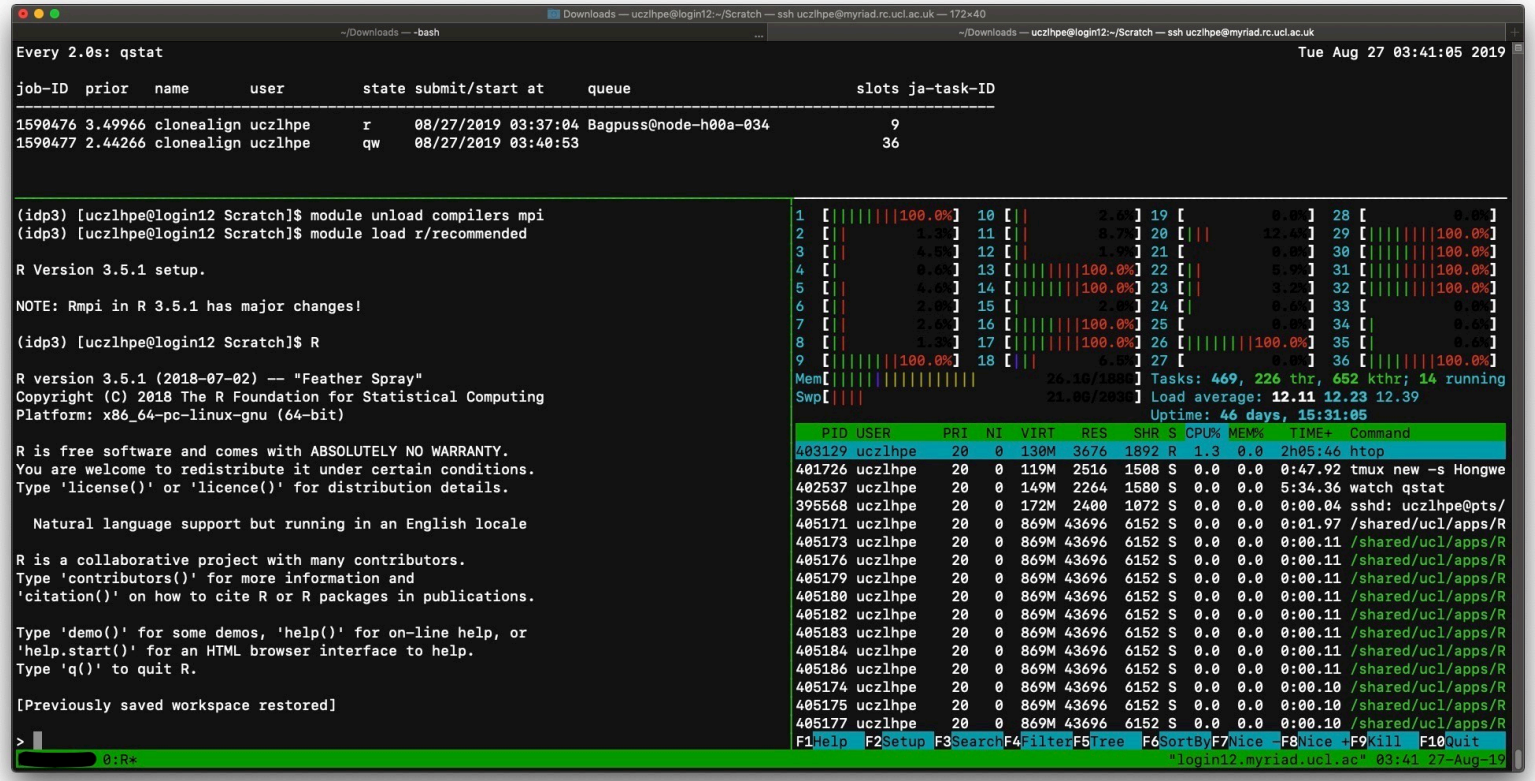
Figure: m5 spectrum

Datasets


700 spectral points
80 samples
m5
80 samples
mp5
80 samples
mp6
independent variable X
PLSR
80 samples
moisture
oil
protein
starch
dependent variable Y

Parallel computing and HPC

HPC: high performance computing system



Number of core	1	9	18	27	36
Running time (s)	97.42	68.84	58.56	55.77	54.97

PLS algorithm

- The partial least squares regression (PLS) is a multi-regression technique proposed by Wold et al. (1984).

- It is also the most common statistical method in the research of near-infrared spectroscopy (NIR).

PLS algorithm

the root mean square error of cross-validation (RMSECV):

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where : n is the number of samples;
y_i is the experimental value of the i-th sample;
$\hat{y}_i$ is the predicted value of the i-th sample by cross-validation which includes removing the set of i-th sample from the calibration set, building a model with the remaining samples, and applying the model to i-th sample.

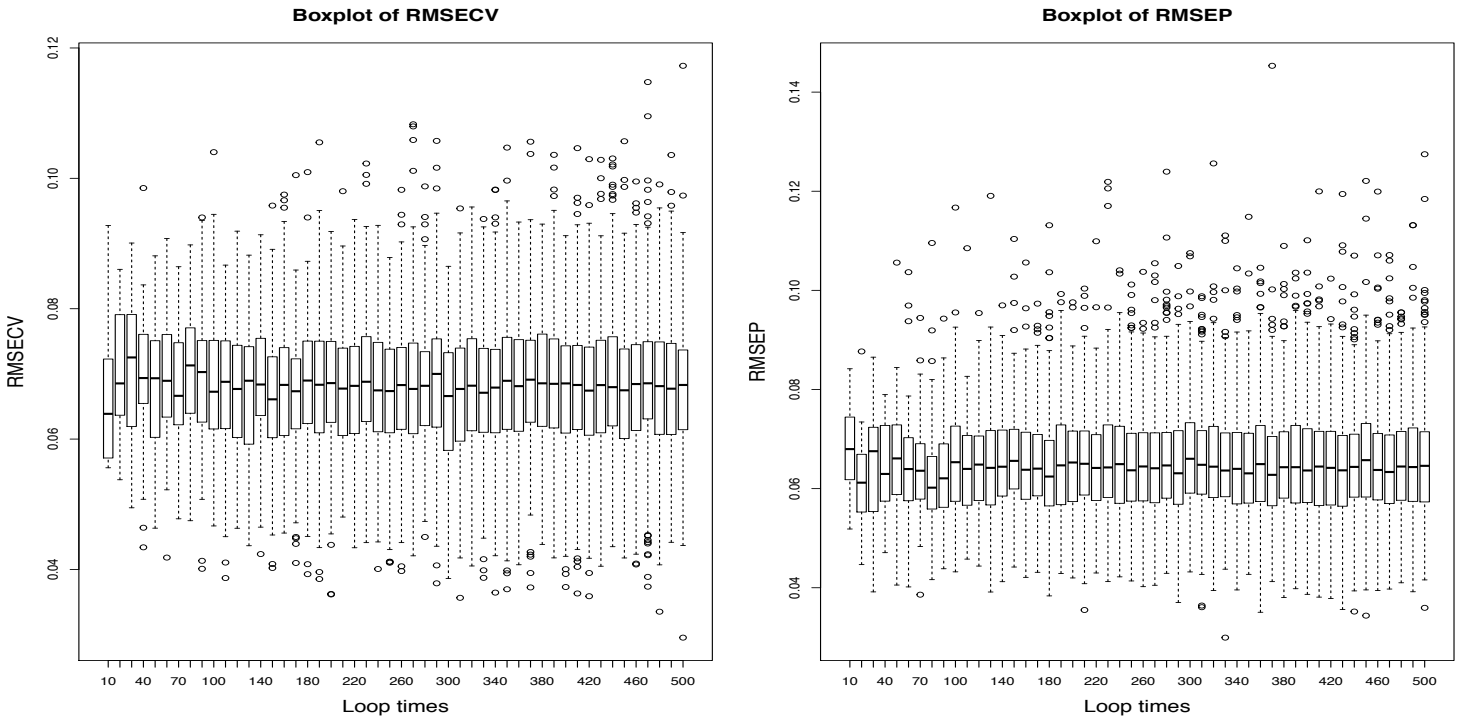
PLS algorithm

the root mean square error of prediction (RMSEP):

$$RMSEP = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

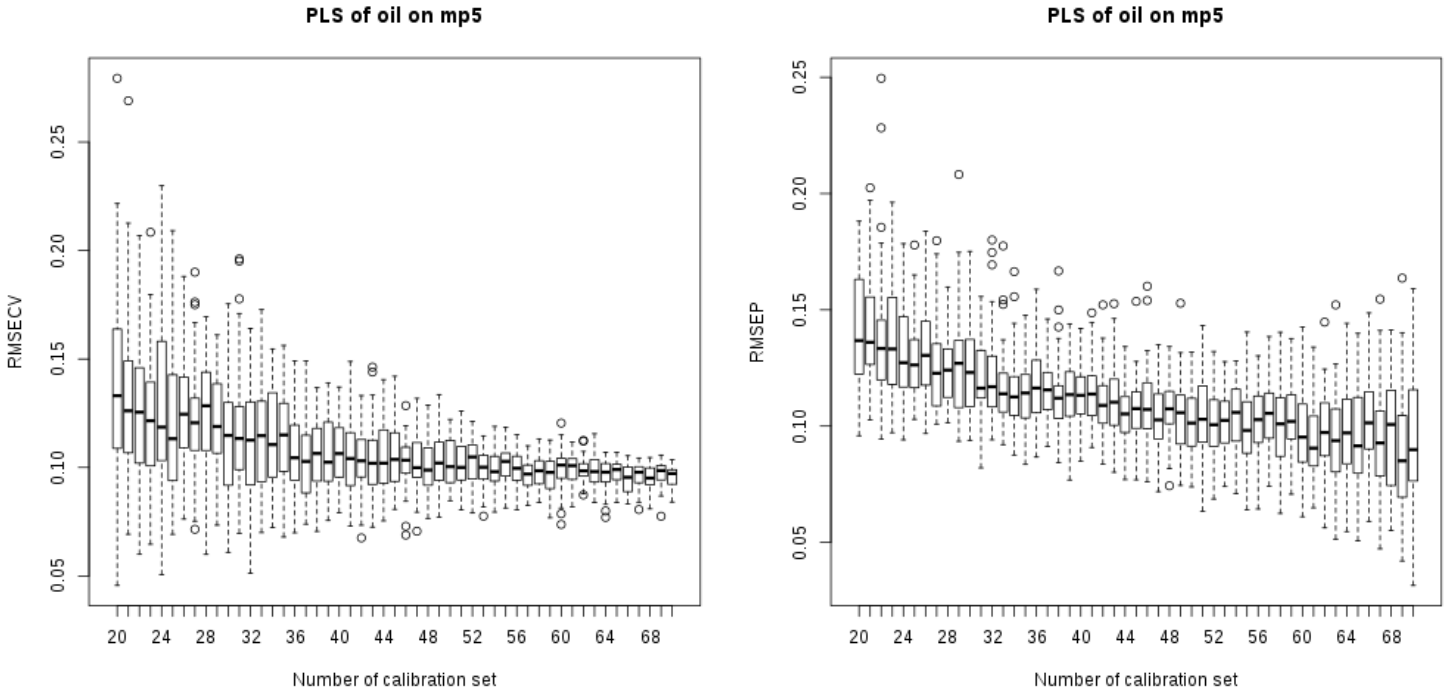
Where : m is the number of prediction sets;
y_i is the experimental value of the i-th sample in the prediction set;
$\hat{y}_i$ is the prediction value of model for the i-th sample.

Loop times



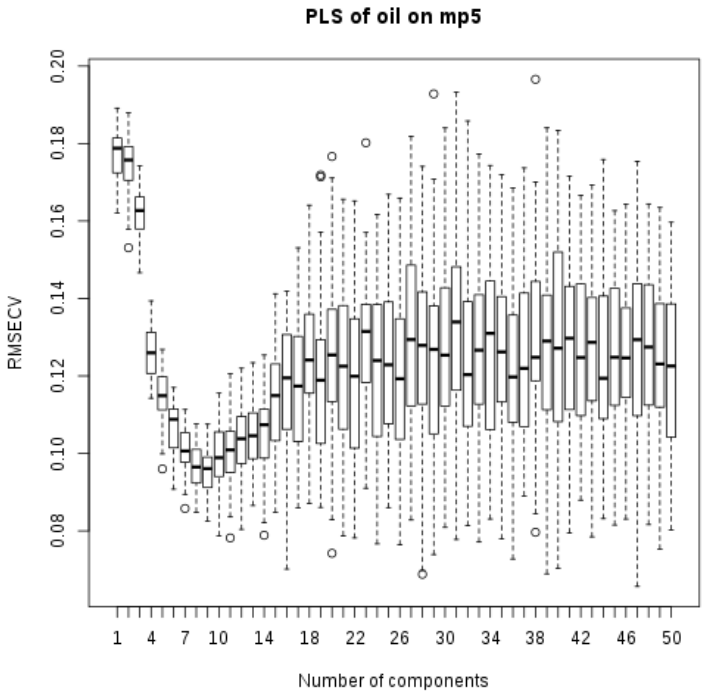
Boxplot of RMSECV and RMSEP under different loop times

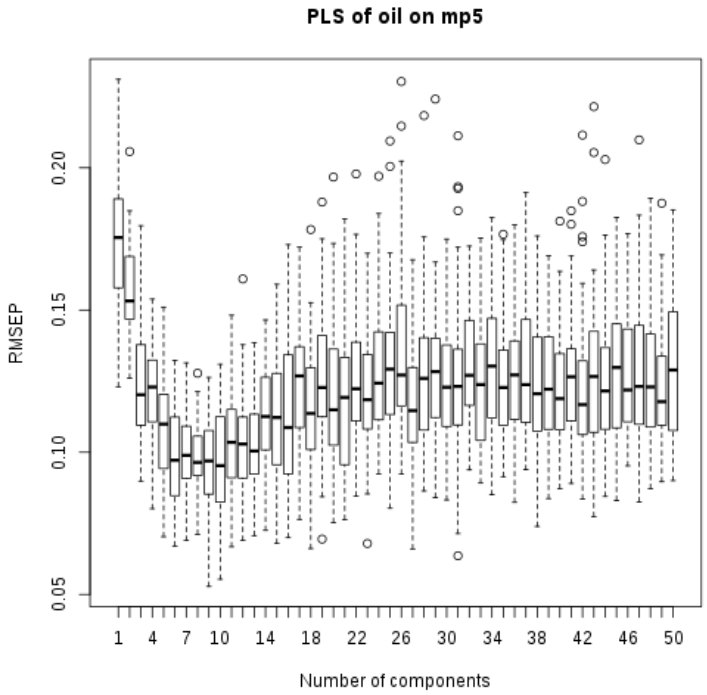
Number of samples



Boxplot of RMSECV and RMSEP under different number of calibration set

Number of components

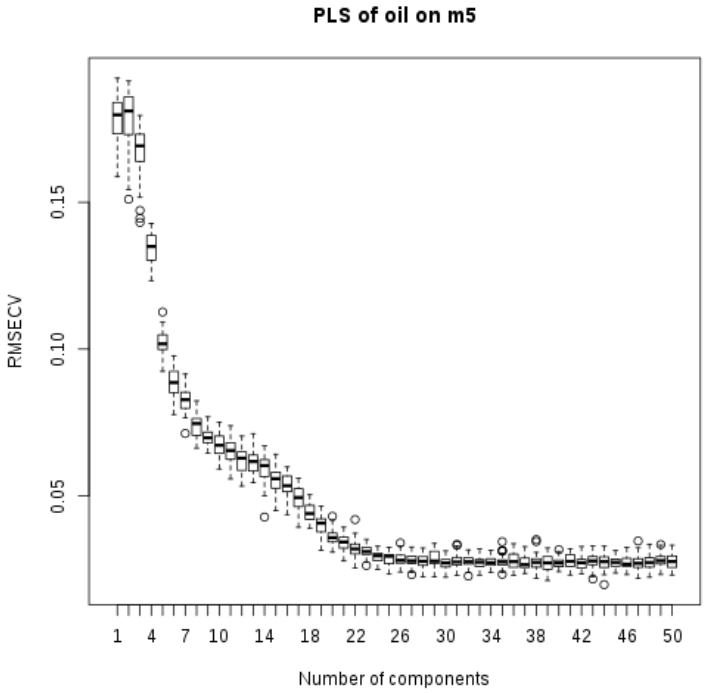


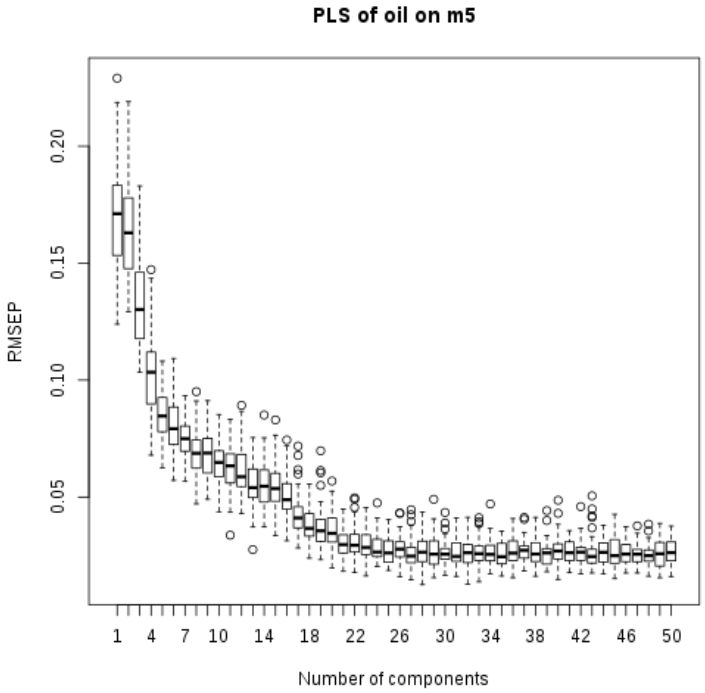


As the number of components increases, the regression performer of models should be getting better and better. However, when the number of components is too large and over the threshold, there will be a situation of overfitting. The model's RMSECV and RMSEP will be reduced to a minimum and then rise again, and the variance will gradually increase during the rise process.

Number of components

But …



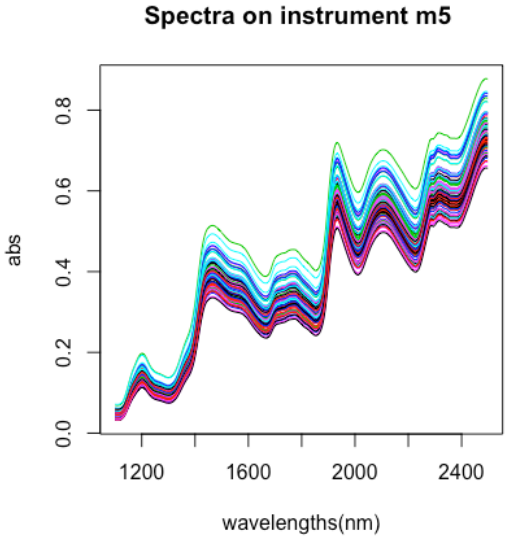


Pre-treatment

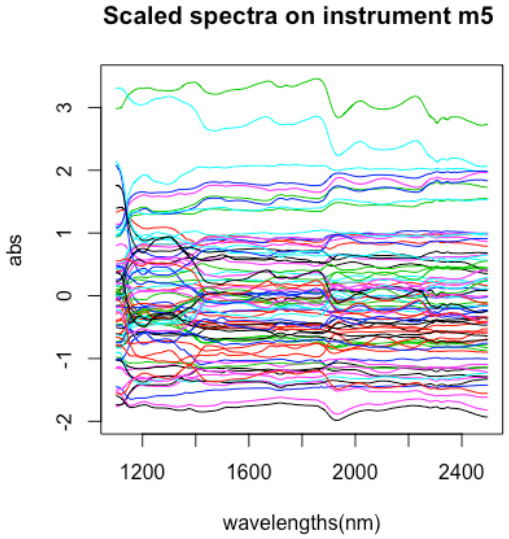
There were four common pre-treatments:

- Nothing to deal with (Su et al., 2006).
- Scale the data (Ergon, 2006).
- Savitzky-Golay filter processing on the data (Galvão et al., 2007).
- Delete the outliers. For example Ji et al. (2015) take out the 75th and 77th corn spectrum from dataset as outliers.

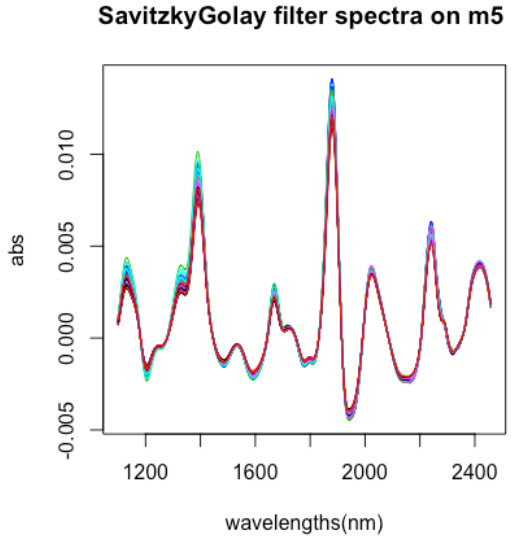
Pre-treatment



Raw data

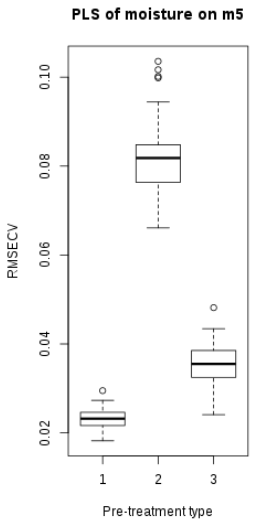


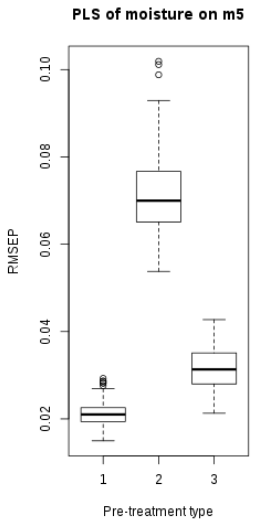
Standardisation

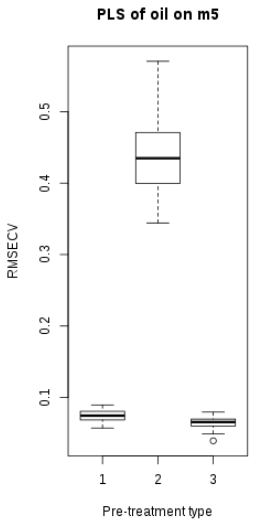


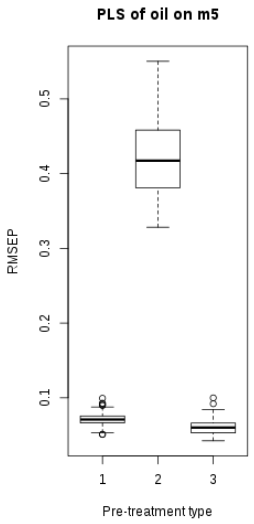
Savitzky-Golay filter

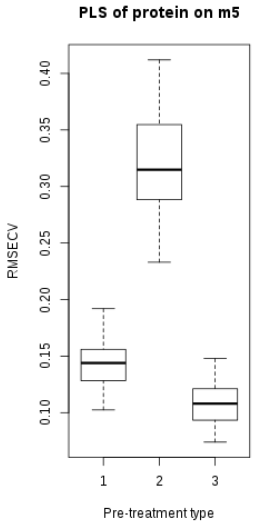
Pre-treatment

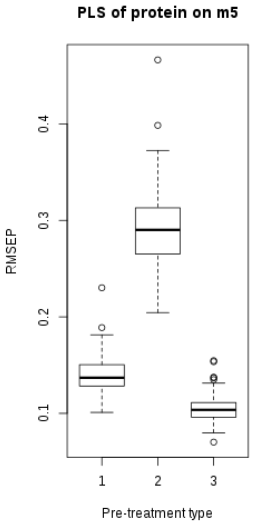






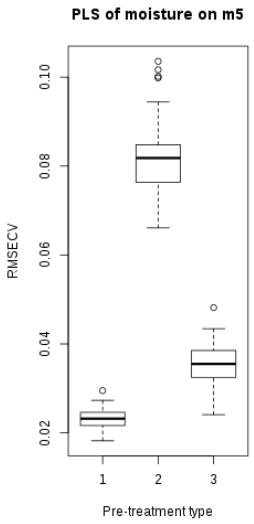


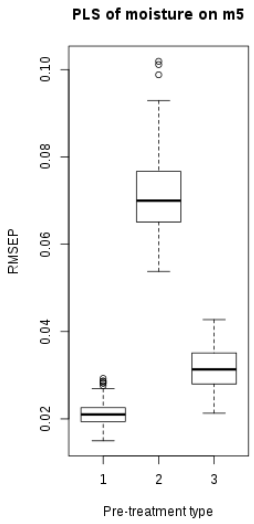


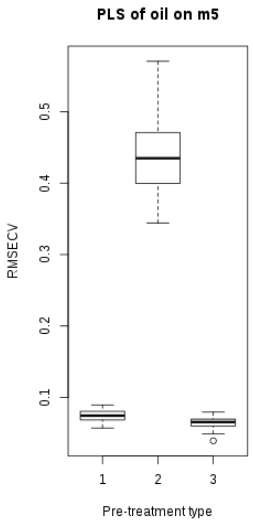


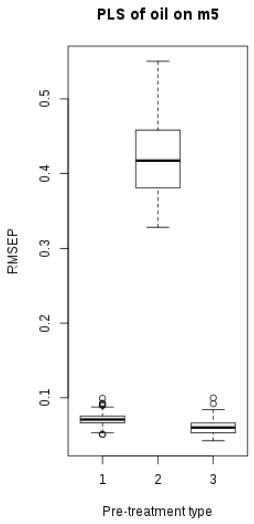
The boxplot of RMSECV and RMSEP from 100 loop times PLSR under different pre-treatment. “1” stand for none pre-treatment; “2” stand for scale X; “3” stand for Savitzky-Golay filler with 1 differentiation order, 2 polynomial order, 21 window size. And 40 samples as calibration set take leave-one-out as cross validation.

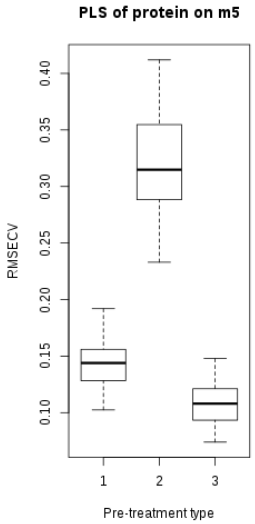
Pre-treatment

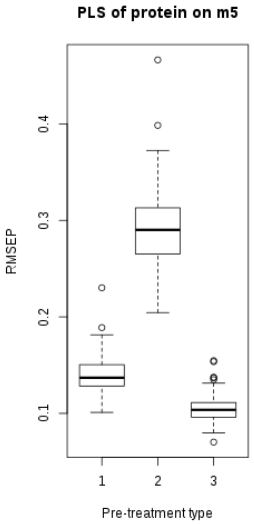










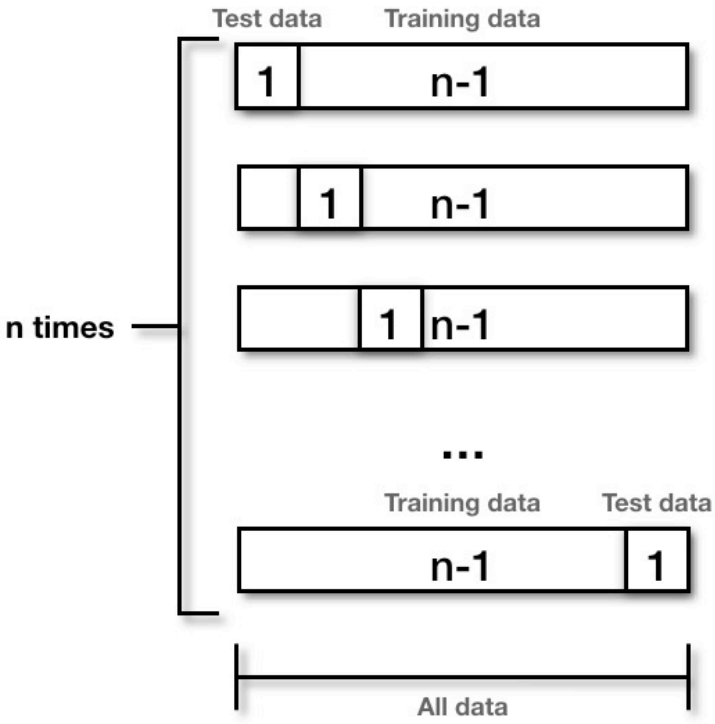


- Standardisation has the worst performance.
  - Because Standardisation takes meaningless data into PLSR.

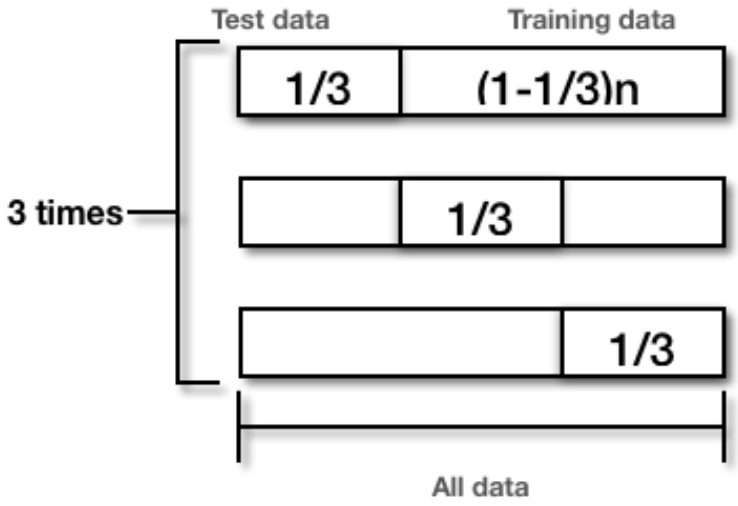
- Savitzky-Golay filter has a better performance in most situations.
  - Because Savitzky-Golay filter smoothes data, and fixes data's displacement.
  - Savitzky-Golay filter makes model more robust.

Cross-validation

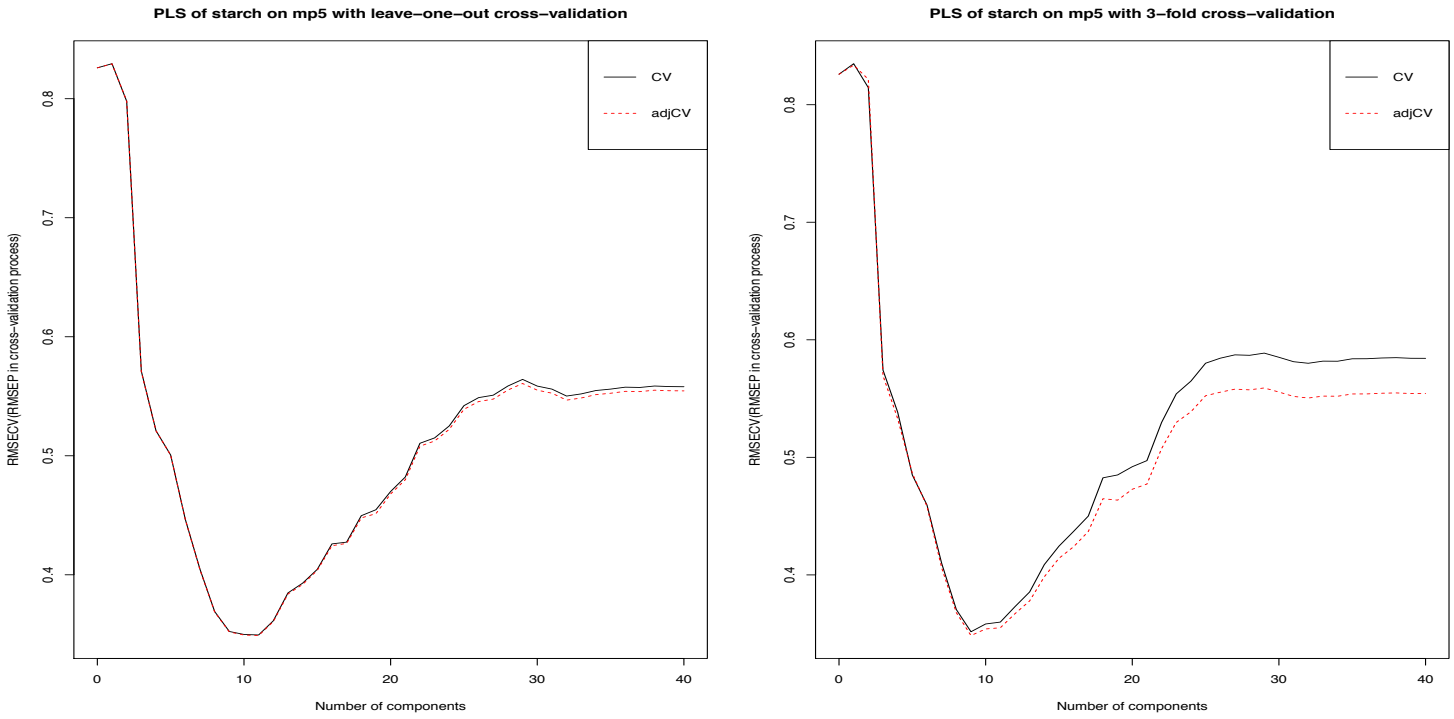
Leave-one-out cross validation:



3-fold cross validation:

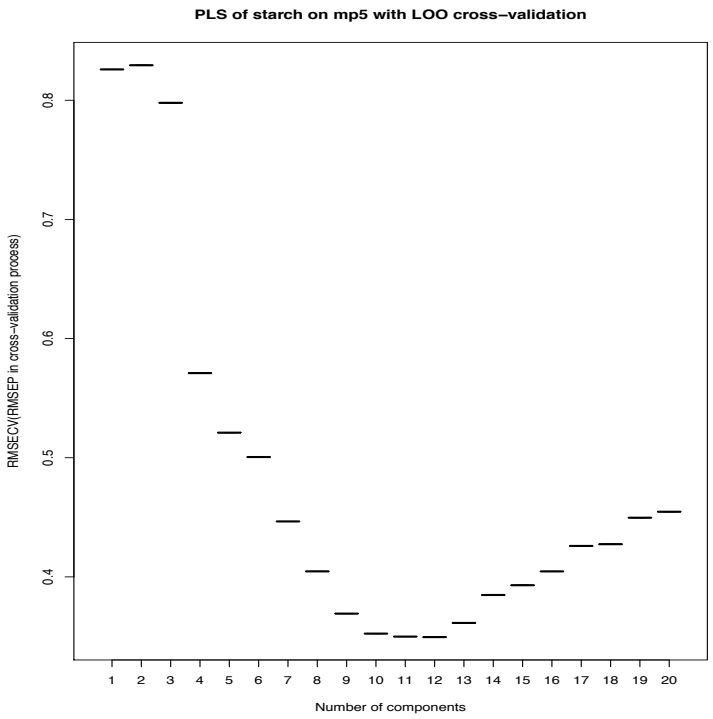


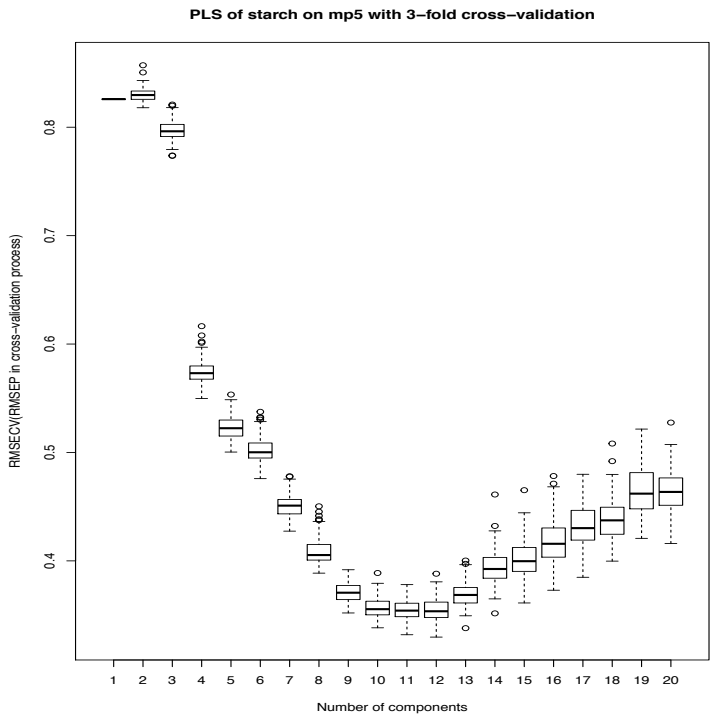
Cross-validation



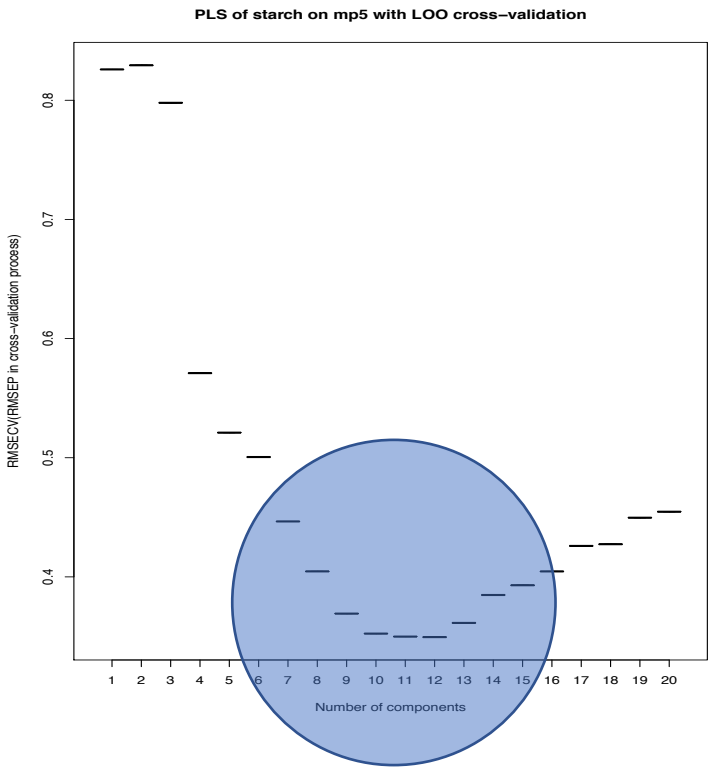
RMSEP curves under LOO and K-fold

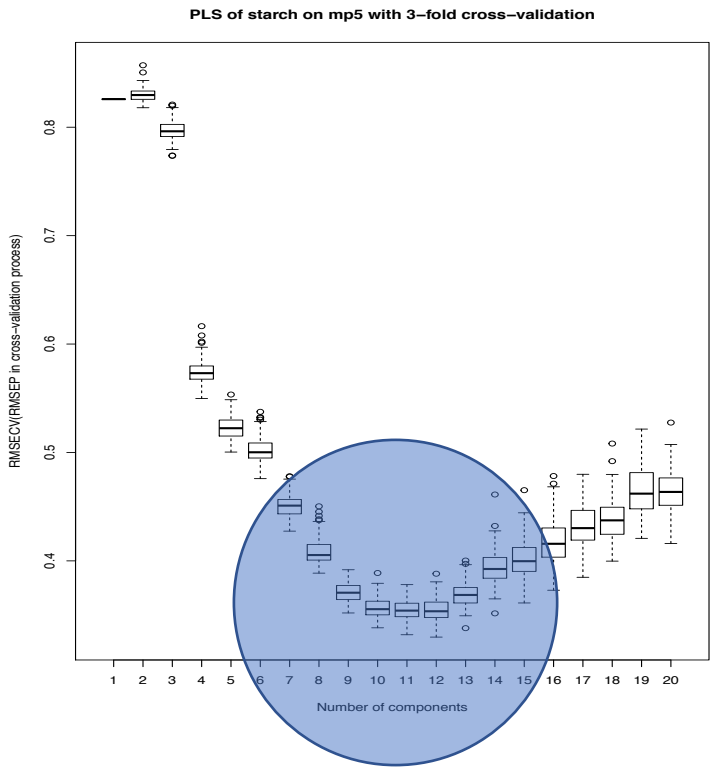
Cross-validation



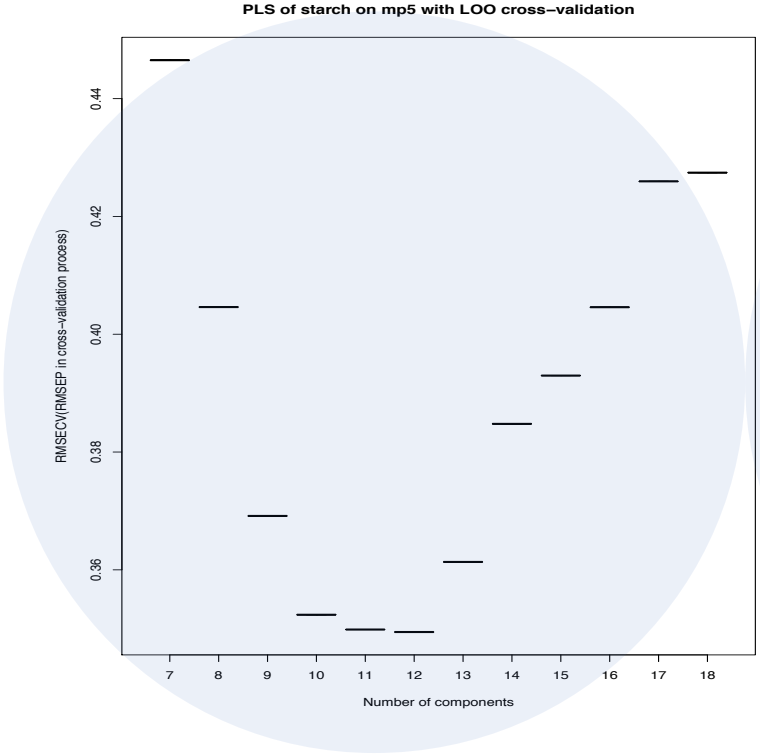


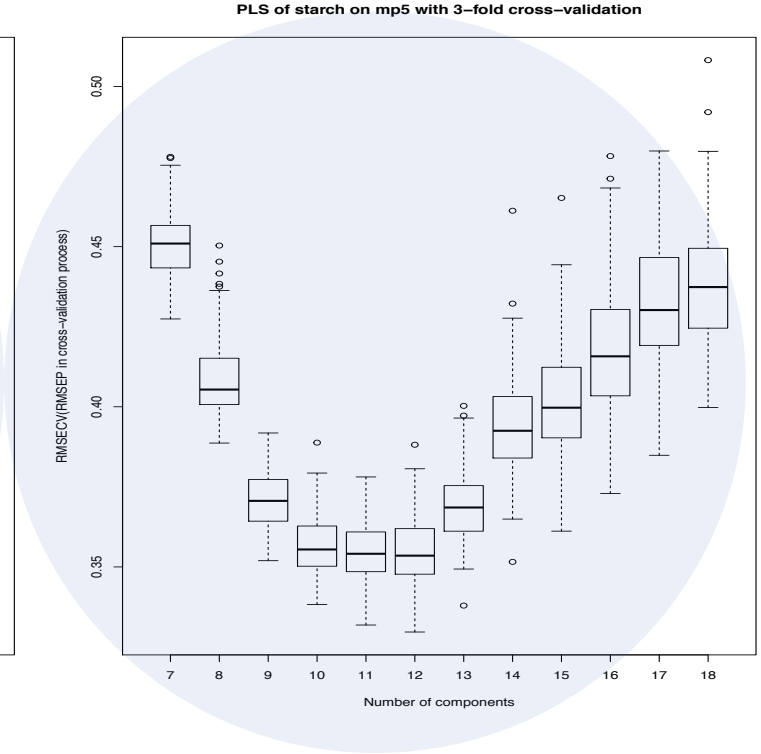
Cross-validation



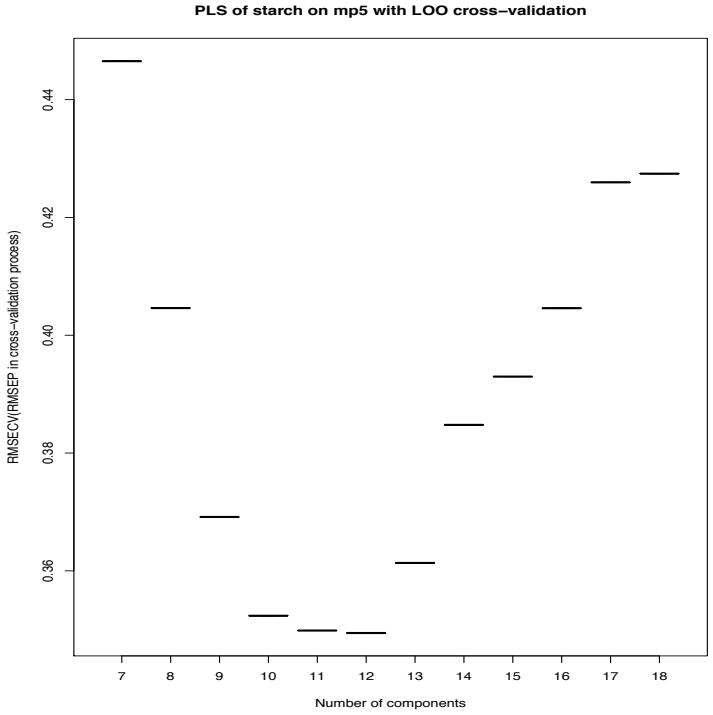


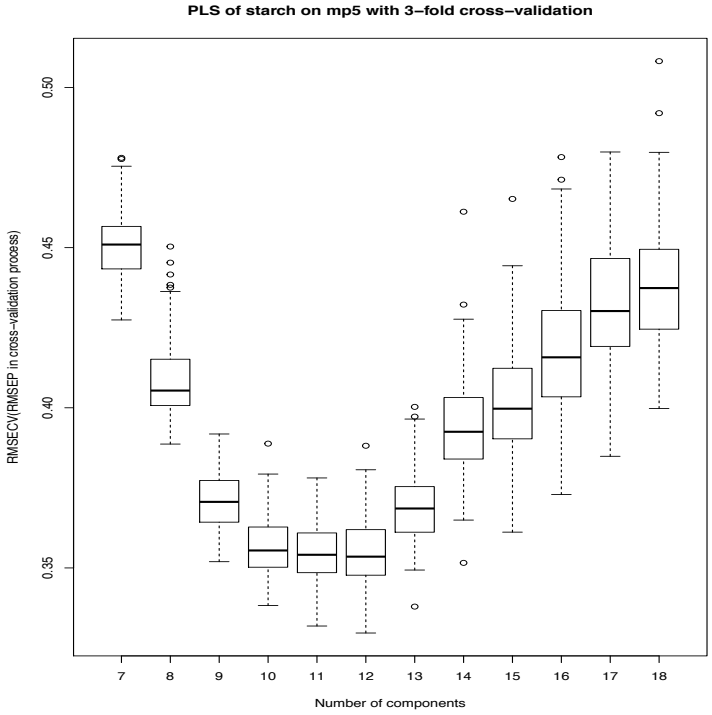
Cross-validation





Cross-validation





When the amount of sample data is small, the difference between LOO and K-fold is not obvious.

However, it should be noted that when the amount of data is too large, the LOO can easily cause the model over-fitting, which is not what we hope.

Compare with papers

Paper	Data set	Pre-treatment	Calibration set	Number of Components	Moisture		PLS in papers		Developed method	
					RMSECV	RMSEP	RMSECV	RMSEP	RMSECV	RMSEP
1	mp6	None	60(LOO)	10		0.148(0.0213)		0.159		0.139
2	m5	None	64(5-fold)	10	0.0152(0.000739)	0.0202(0.00319)	0.0149	0.0201	0.00026	0.00035
3	m5	Scale	40(LOO)	12		0.0231(0.00443)		0.3506		0.3485
3	mp5	Scale	40(LOO)	12		0.159(0.0178)		0.3506		0.3485
4	mp5	Scale	40(LOO)	10		0.405(0.0467)		0.357		0.265
5	m5	SG(1,2,13)*	60(3-fold)	5		0.0547(0.00942)		0.040		0.012
6	m5	SG(1,2,21)*	60(LOO)	6		0.0396(0.00625)		0.045		0.019
8	m5	Delete 75 , 77	52(LOO)	10	0.0221(0.0018)	0.0194(0.00298)	0.0124	0.0157	0.0047	0.0056

Paper	Data set	Pre-treatment	Calibration set	Number of Components	Oil		PLS in papers		Developed method	
					RMSECV	RMSEP	RMSECV	RMSEP	RMSECV	RMSEP
1	mp6	None	60(LOO)	10		0.0991(0.0161)		0.107		0.0948
3	m5	Scale	40(LOO)	14		0.396(0.0665)		0.6912		0.6902
3	mp5	Scale	40(LOO)	14		0.694(0.095)		0.6912		0.6902
5	m5	SG(1,2,13)*	60(3-fold)	12		0.0329(0.00672)		0.029		0.022
6	m5	SG(1,2,21)*	60(LOO)	10		0.0505(0.0103)		0.028		0.030
7	m5	SG(0,2,13)*	64(5-fold)	7	0.0827(0.00419)	0.0716(0.0116)	0.0729	0.0855		0.0400
7	m5	SG(1,2,13)*	64(5-fold)	7	0.0639(0.00357)	0.0548(0.012)	0.0577	0.0682	0.0363	0.0400
7	m5	SG(2,2,13)*	64(5-fold)	7	0.0480(0.00312)	0.0368 (0.0088)	0.0370	0.0397	0.0363	0.0400
8	m5	Delete 75 , 77	52(LOO)	10	0.0651(0.00662)	0.0604(0.00876)	0.0613	0.0673	0.0483	0.0546

Developed method's F-test

$$
\begin{aligned}
& RMSEP^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \\
&= Var(\hat{y}) + Bias(\hat{y}, y)^2 \\
&\approx Var(\hat{y})
\end{aligned}
$$

Because:

$$\frac{df * Var(\hat{y})}{\sigma} \sim \chi^2_{df}$$

Hence:

$$\frac{RMSEP_1^2}{RMSEP_2^2} \sim F(df_1, df_2)$$

Where: $RMSEP_1$ *is RMSEP based on PLS*

$RMSEP_2$ *is RMSEP based on developed method*

Developed method's F-test

F-test on regression of moisture

Paper	Prediction set	RMSEP of PLS in papers	RMSEP of developed method	F value	Significant F statistic (0.05)
1	20	0.159	0.139	1.31	2.124155
2	16	0.0201	0.00035	3298.04	2.333484
3	40	0.3506	0.3485	1.01	1.692797
4	40	0.357	0.265	1.81	1.692797
5	20	0.040	0.012	11.11	2.124155
6	20	0.045	0.019	5.61	2.124155
8	26	0.0157	0.0056	7.86	1.929213

Developed method's F-test

There are two main problems with this model:

- The bias' distribution is non-central chi-square distribution. If the bias does not great less than the variance, this test will be invalid.

- $RMSEP_1$ and $RMSEP_2$ are not independent, because of the same prediction sets.