

The NIR corn data set

Hongwei PENG

Supervisor: Prof. Tom Fearn
Department of statistical science
University College London

September 3, 2019

- One of the most easily accessible public data sets
- high-dimensional NIR spectroscopic data
- 80 samples
- 3 spectra : m5, mp5 and mp6 spectrum
- 700 spectral points for every spectrum
- 4 constituents: moisture, oil, protein and starch
- corn dataset is available in the website: <http://www.eigenvector.com/data/Corn/index.html>

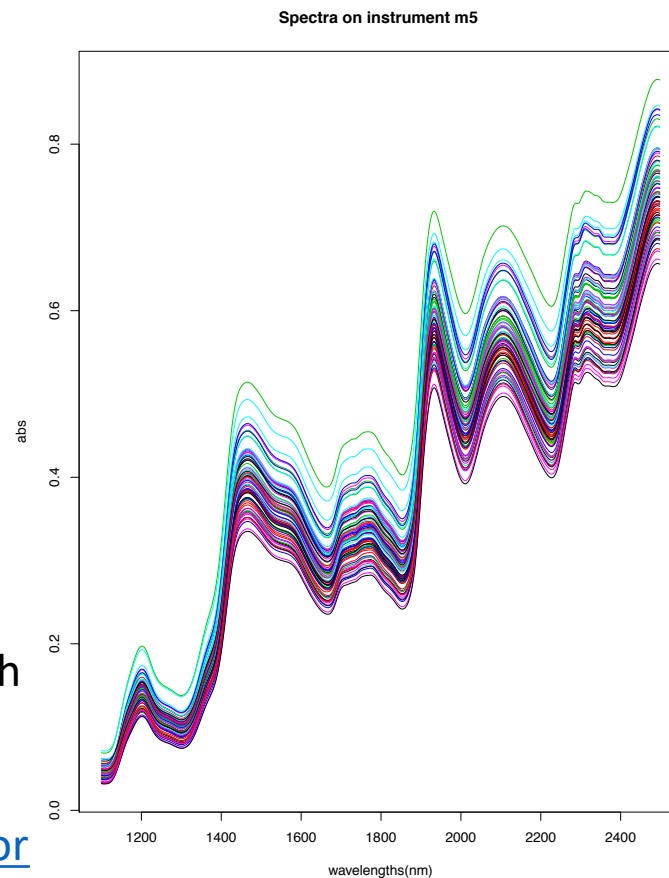
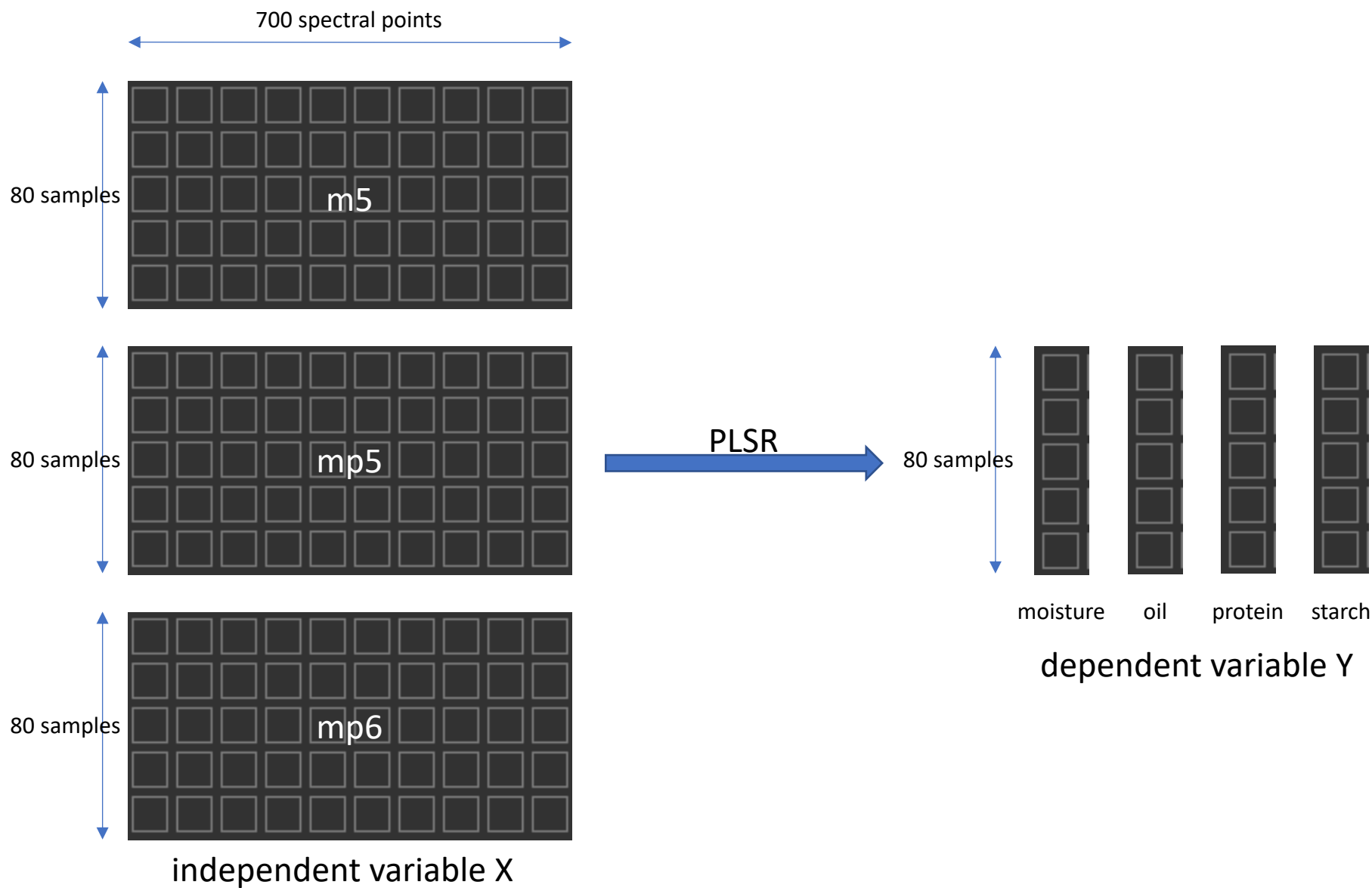


Figure: m5 spectrum



Parallel computing and HPC

HPC: high performance computing system

```

Every 2.0s: qstat
job-ID prior  name      user      state submit/start at   queue                slots ja-task-ID
-----
1590476 3.49966 clonealign uc2lhpe  r    08/27/2019 03:37:04 Bagpuss@node-h00a-034      9
1590477 2.44266 clonealign uc2lhpe  qw   08/27/2019 03:40:53                               36

(idp3) [uc2lhpe@login12 Scratch]$ module unload compilers mpi
(idp3) [uc2lhpe@login12 Scratch]$ module load r/recommended
R Version 3.5.1 setup.
NOTE: Rmpi in R 3.5.1 has major changes!
(idp3) [uc2lhpe@login12 Scratch]$ R
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]
>

```

1 [|||||100.0%] 10 [|| 2.6%] 19 [|| 0.0%] 28 [|| 0.0%]
 2 [|| 1.3%] 11 [|| 8.7%] 20 [|| 12.5%] 29 [|||||100.0%]
 3 [|| 4.8%] 12 [|| 1.9%] 21 [|| 0.0%] 30 [|||||100.0%]
 4 [|| 0.4%] 13 [|||||100.0%] 22 [|| 0.0%] 31 [|||||100.0%]
 5 [|| 4.4%] 14 [|||||100.0%] 23 [|| 0.0%] 32 [|||||100.0%]
 6 [|| 2.0%] 15 [|| 2.0%] 24 [|| 0.0%] 33 [|| 0.0%]
 7 [|| 2.0%] 16 [|||||100.0%] 25 [|| 0.0%] 34 [|| 0.0%]
 8 [|| 1.3%] 17 [|||||100.0%] 26 [|||||100.0%] 35 [|| 0.0%]
 9 [|||||100.0%] 18 [|| 5.5%] 27 [|| 0.0%] 36 [|||||100.0%]
 Mem[|||||] 26.16/1800 Tasks: 469, 226 thr, 652 kthr; 14 running
 Swp[|||||] 21.00/2036 Load average: 12.11 12.23 12.39
 Uptime: 46 days, 15:31:05

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
403129	uc2lhpe	20	0	130M	3676	1892	R	1.3	0.0	2h05:46	htop
401726	uc2lhpe	20	0	119M	2516	1508	S	0.0	0.0	0:47.92	tmux new -s Hongwe
402537	uc2lhpe	20	0	149M	2264	1580	S	0.0	0.0	5:34.36	watch qstat
395568	uc2lhpe	20	0	172M	2400	1072	S	0.0	0.0	0:00.04	sshd: uc2lhpe@pts/
405171	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:01.97	/shared/uc1/apps/R
405173	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405176	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405179	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405180	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405182	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405183	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405184	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405186	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.11	/shared/uc1/apps/R
405174	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.10	/shared/uc1/apps/R
405175	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.10	/shared/uc1/apps/R
405177	uc2lhpe	20	0	869M	43696	6152	S	0.0	0.0	0:00.10	/shared/uc1/apps/R

F1Help F2Setup F3SearchF4FilterF5Free F6SortByF7Nice F8Kill F9Quit
 "login12.myriad.ucl.ac" 03:41 27-Aug-19

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

- 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).

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.

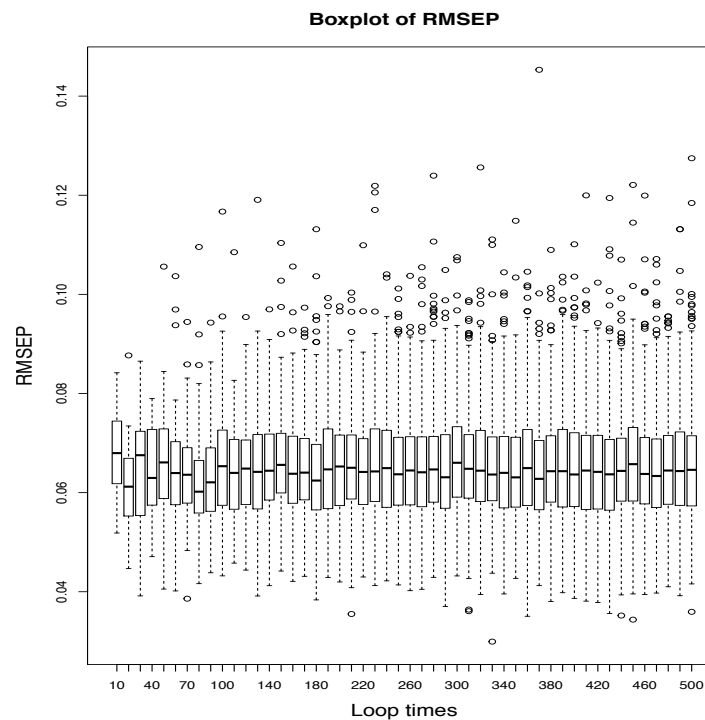
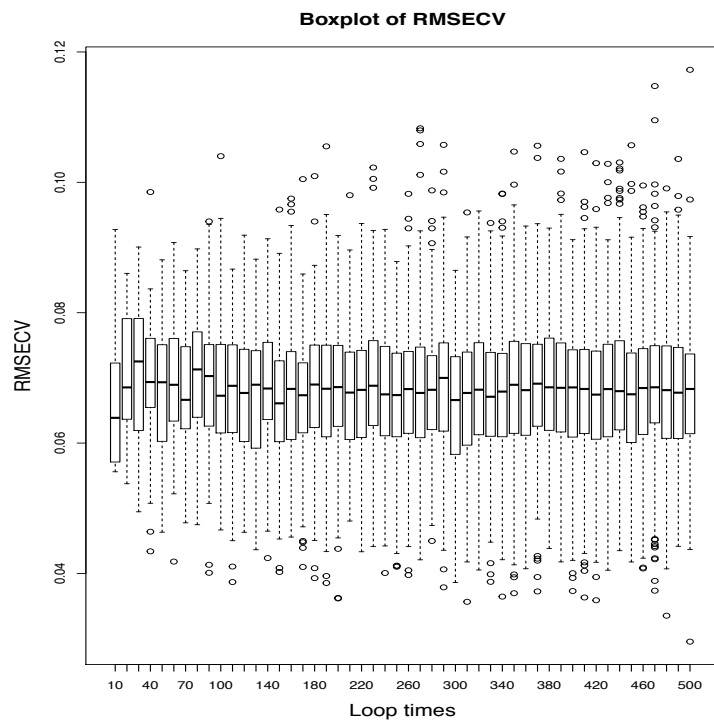
the root mean square error of prediction (RMSEP):

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

Where : m is the number of prediction sets;

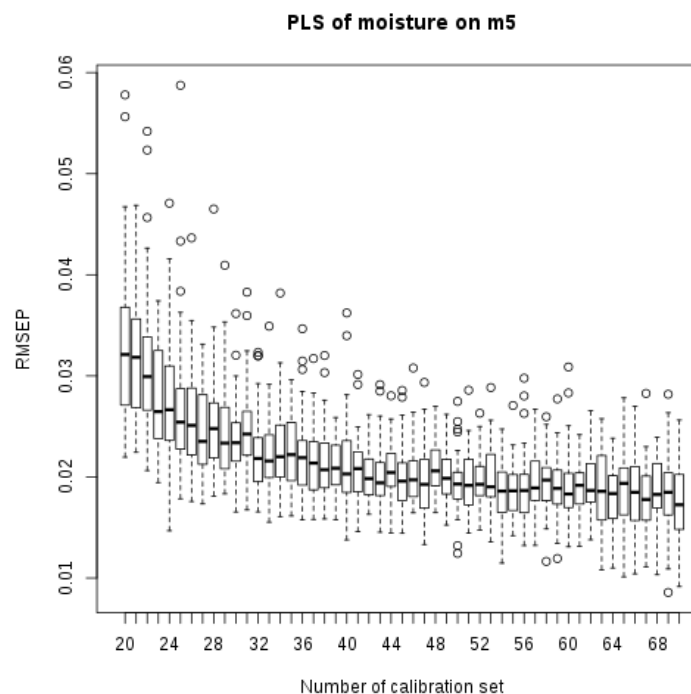
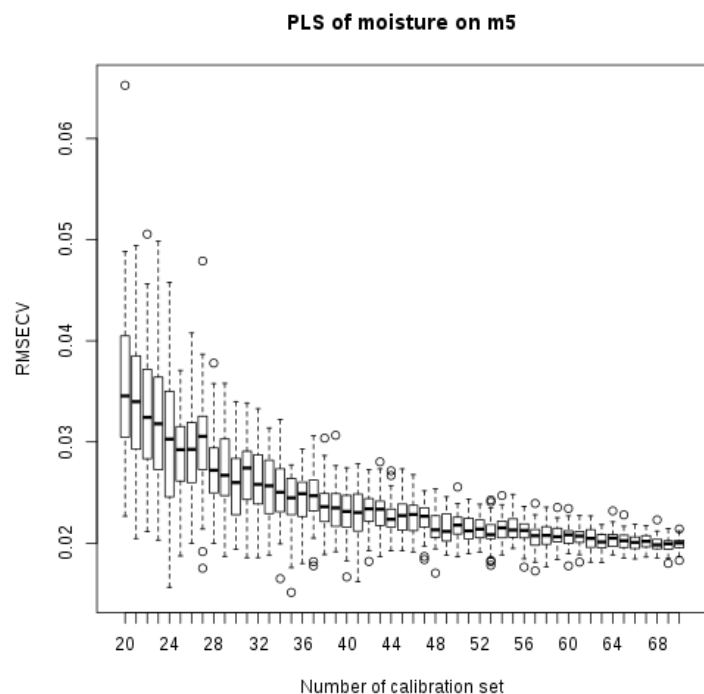
y_i is the experimental value of the i -th sample in the prediction set;

\hat{y}_i is the prediction value of model for the i -th sample.



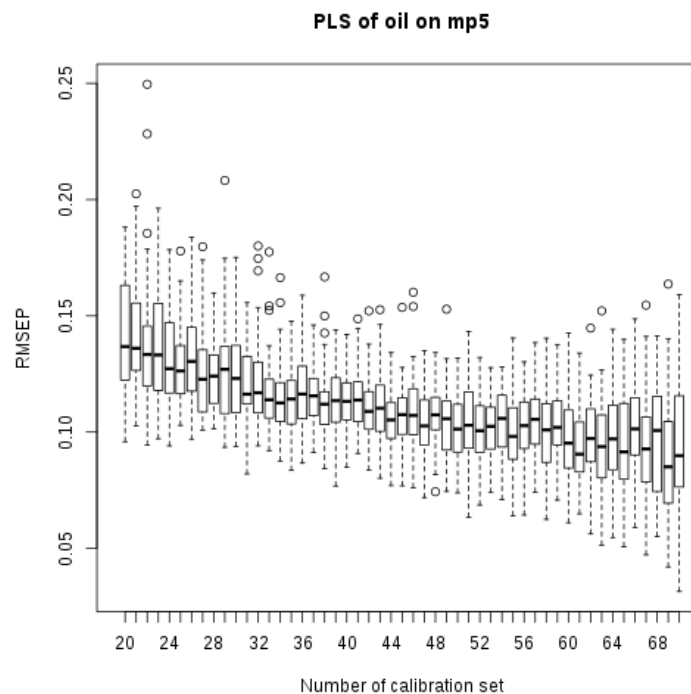
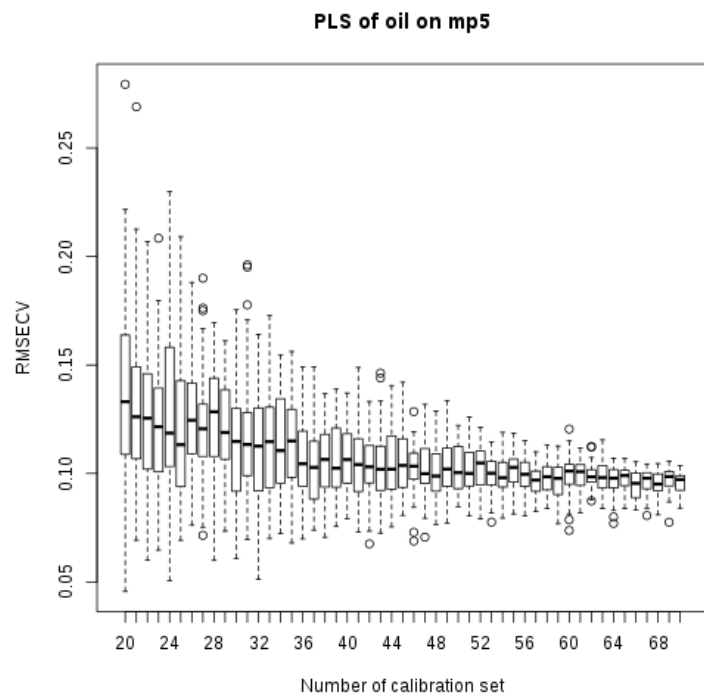
Boxplot of RMSECV and RMSEP under different loop times

Number of samples



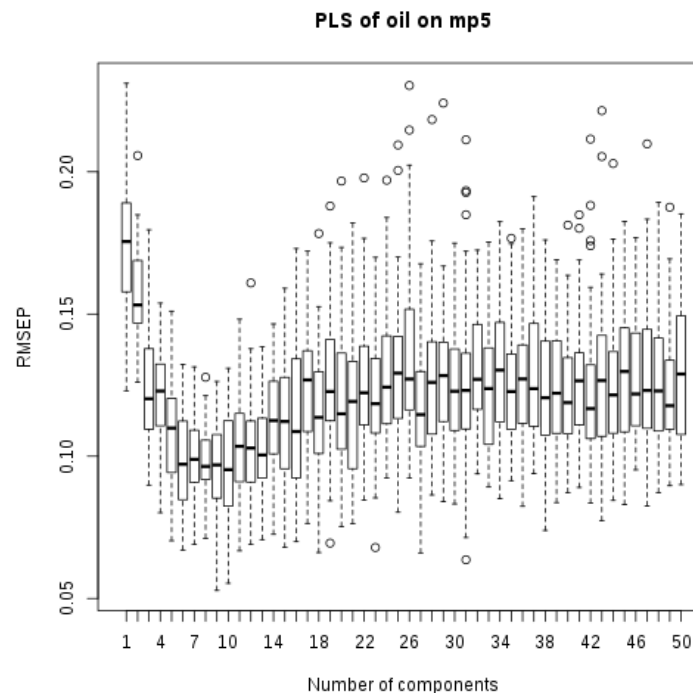
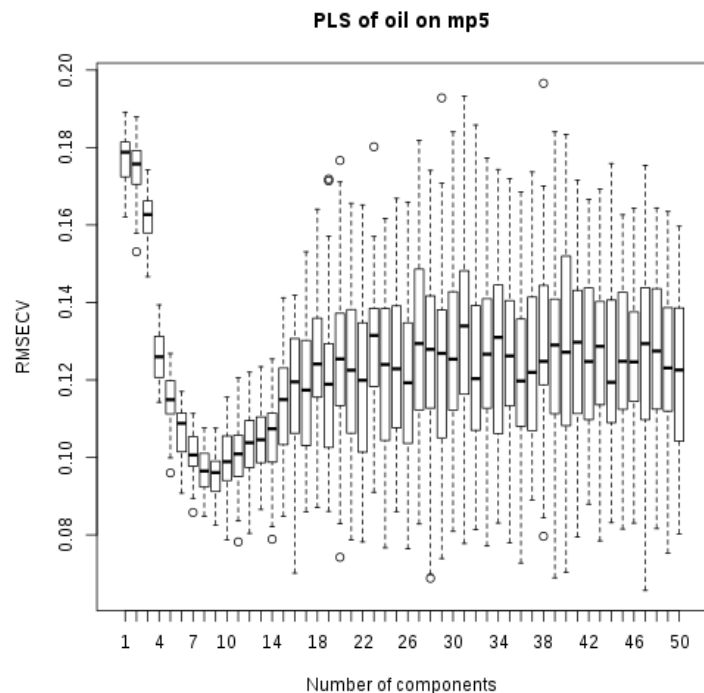
Boxplot of RMSECV and RMSEP under different number of calibration set

Number of samples



Boxplot of RMSECV and RMSEP under different number of calibration set

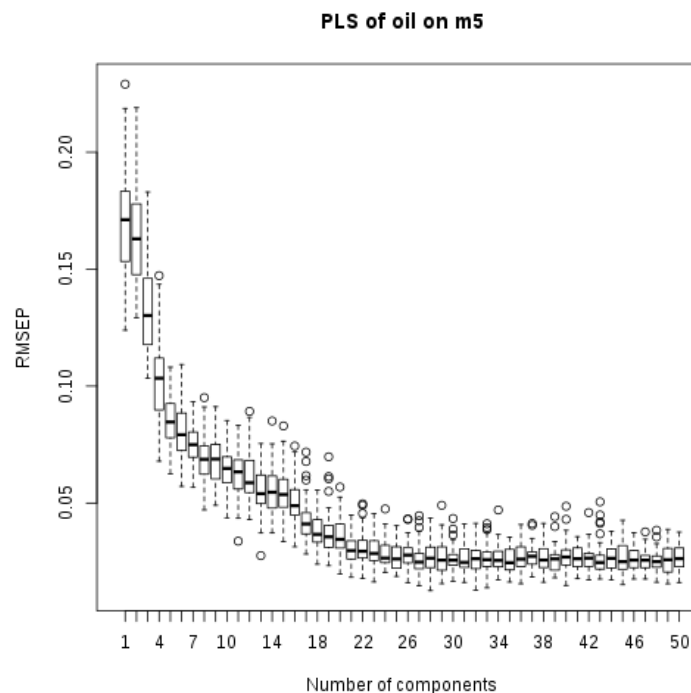
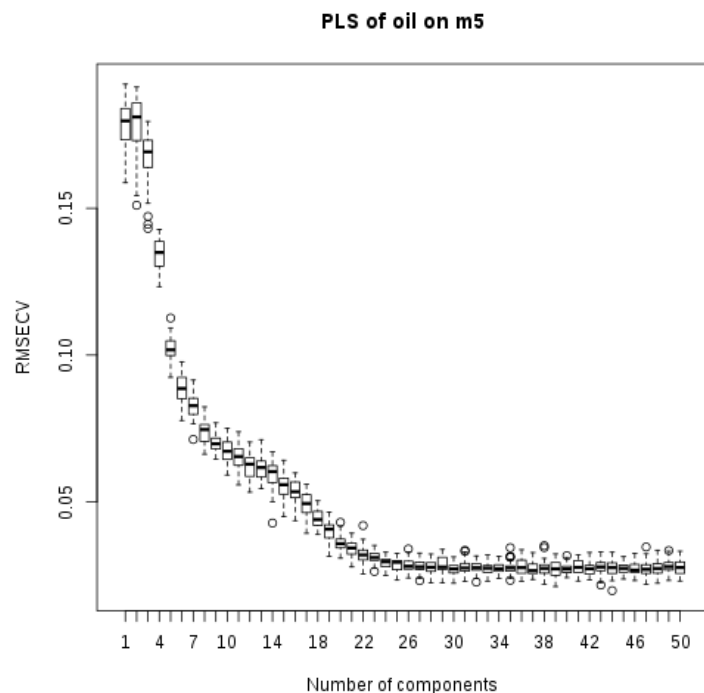
Number of components



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

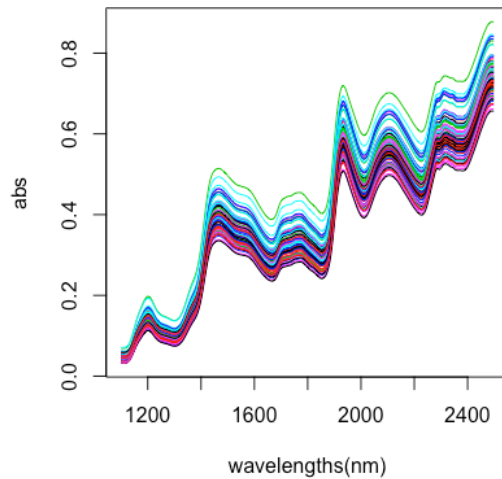
But ...



There were four common pre-treatments:

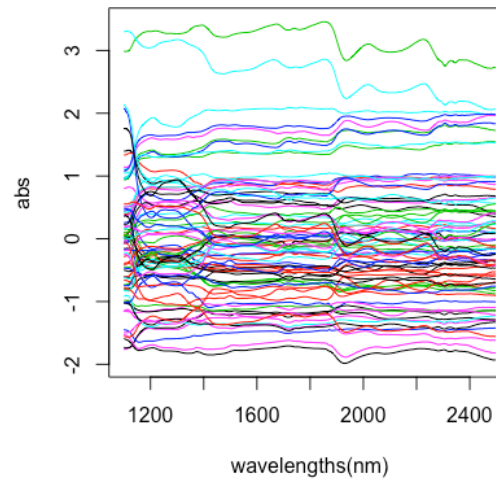
- Nothing to deal with (Su et al., 2006).
- Scale the data (Ergon, 2006).
- Savitzky-Golay filter processing on the data (Galvão et al., 2007).
- Delete the outliers. For example Ji et al. (2015) take out the 75th and 77th corn spectrum from dataset as outliers.

Spectra on instrument m5



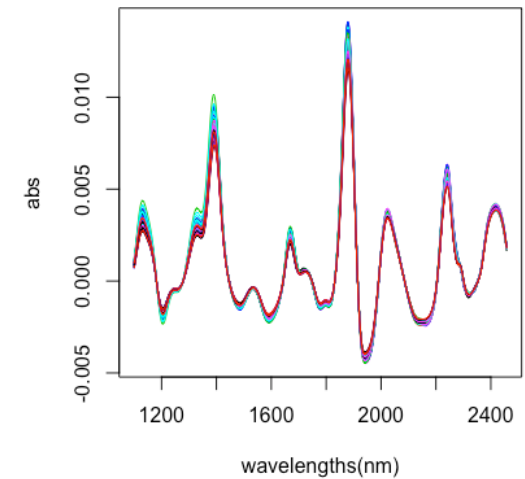
Raw data

Scaled spectra on instrument m5

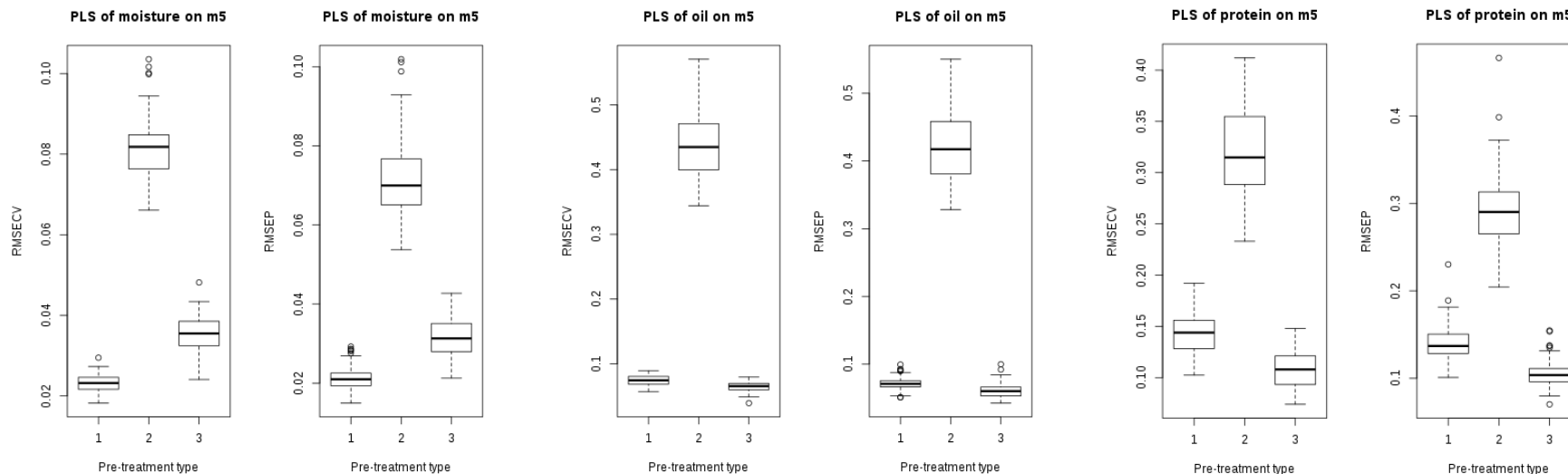


Standardization

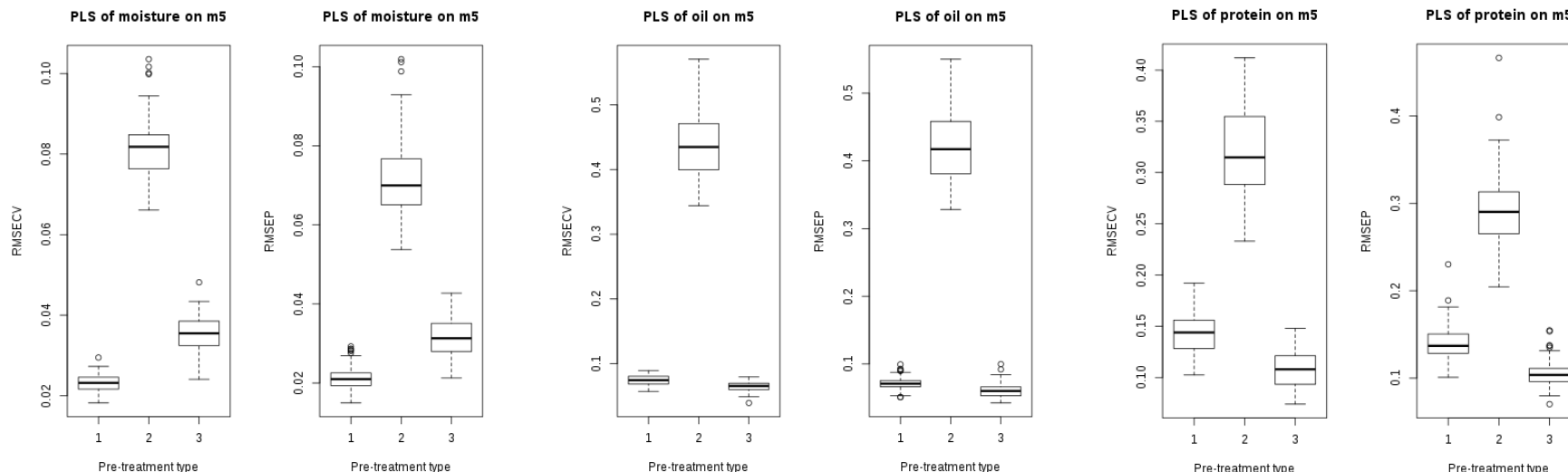
SavitzkyGolay filter spectra on m5



Savitzky-Golay filter

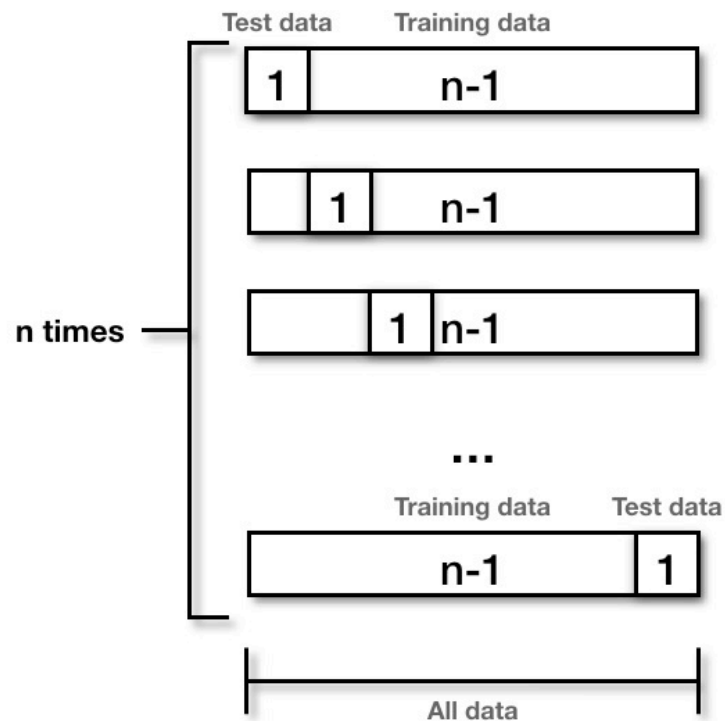


The boxplot of RMSECV and RMSEP from 100 loop times PLSR under different pre-treatment. “1” stand for none pre-treatment; “2” stand for scale X; “3” stand for Savitzky-Golay filler with 1 differentiation order, 2 polynomial order, 21 window size. And 40 samples as calibration set take leave-one-out as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

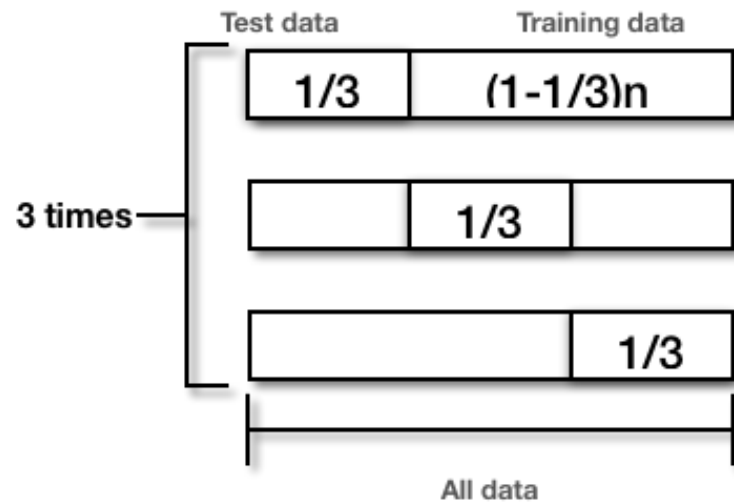


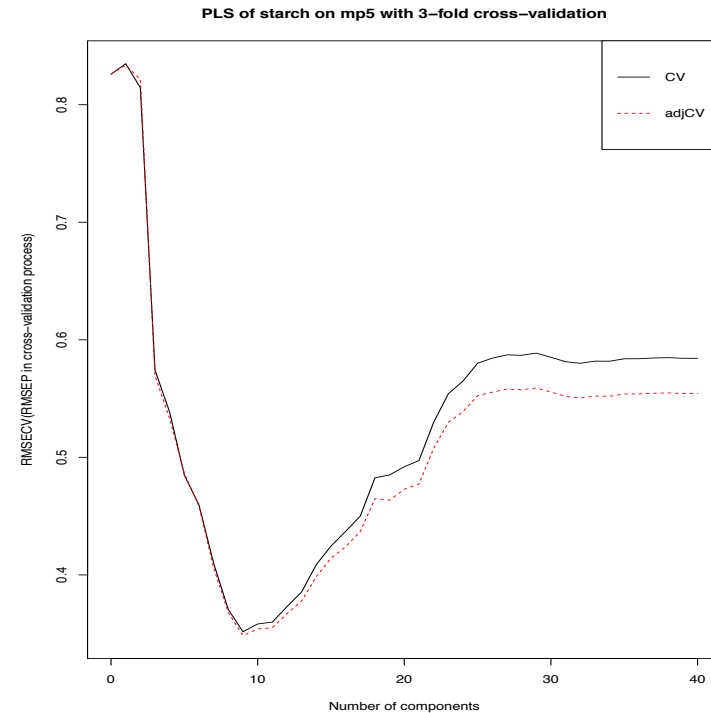
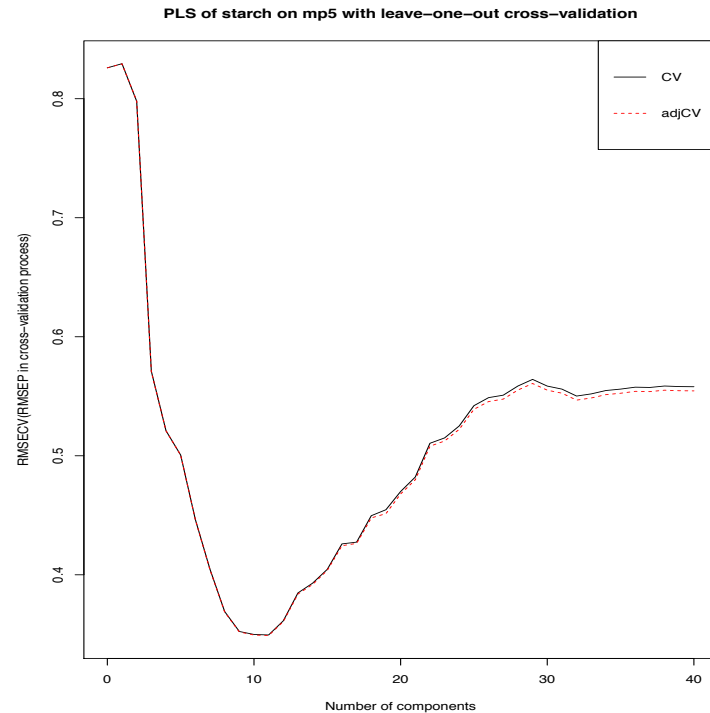
- Standardization has the worst performance.
 - Because Standardization takes meaningless data into PLSR.
- Savitzky-Golay filter has a better performance in most situations.
 - Because Savitzky-Golay filter smoothes data, and fixes data's displacement.
 - SavitzkyGolay filter makes model more robust.

Leave-one-out cross validation:



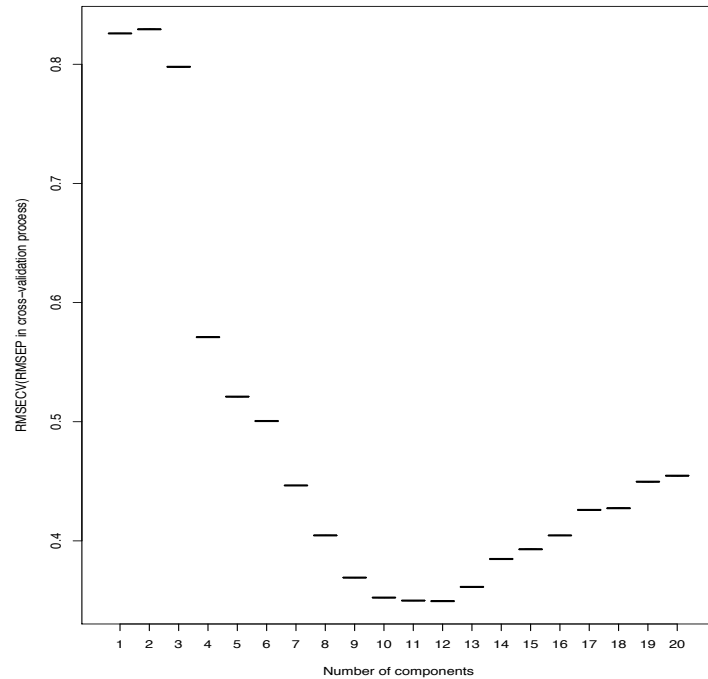
3-fold cross validation:



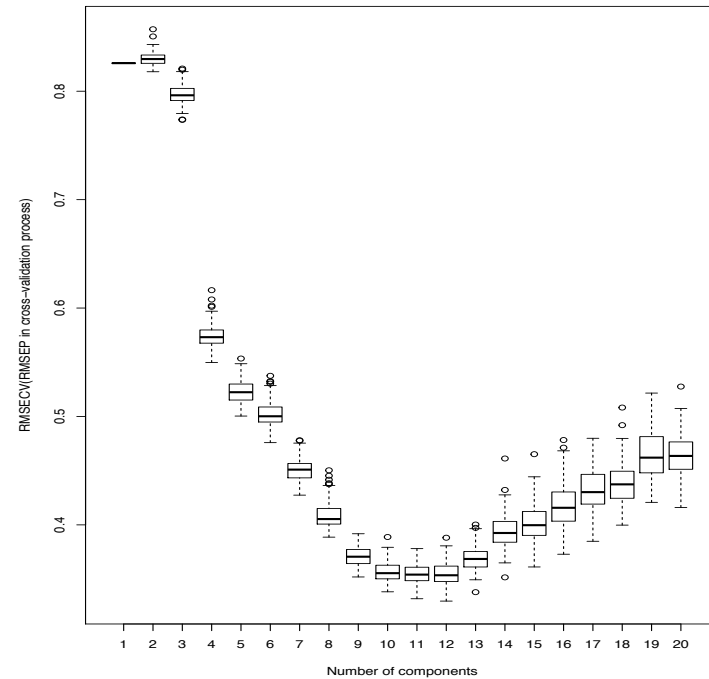


RMSEP curves under LOO and K-fold

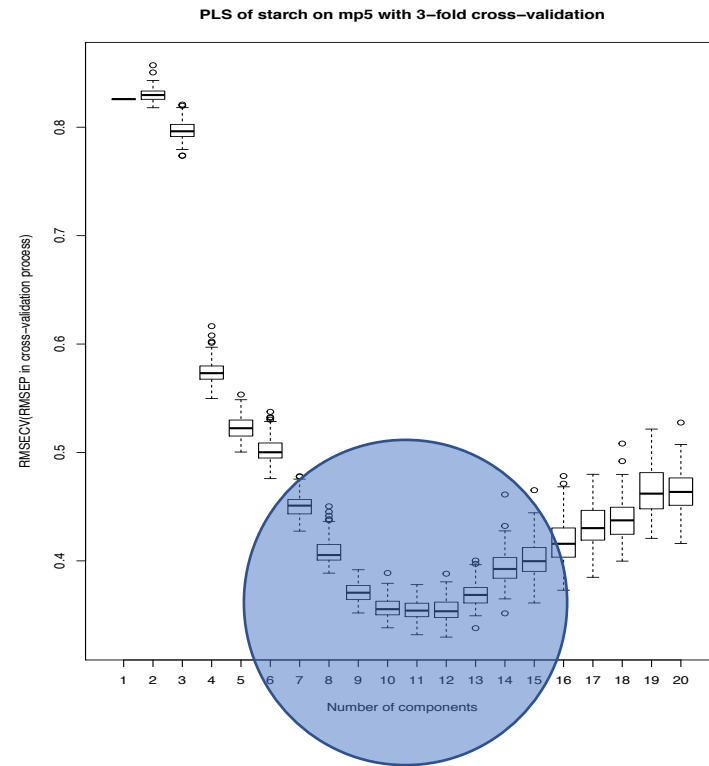
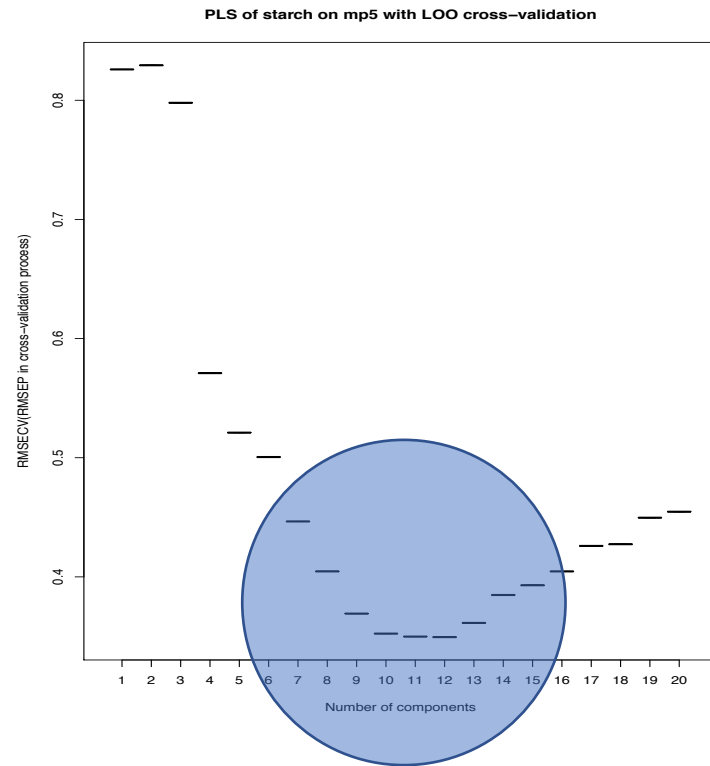
PLS of starch on mp5 with LOO cross-validation



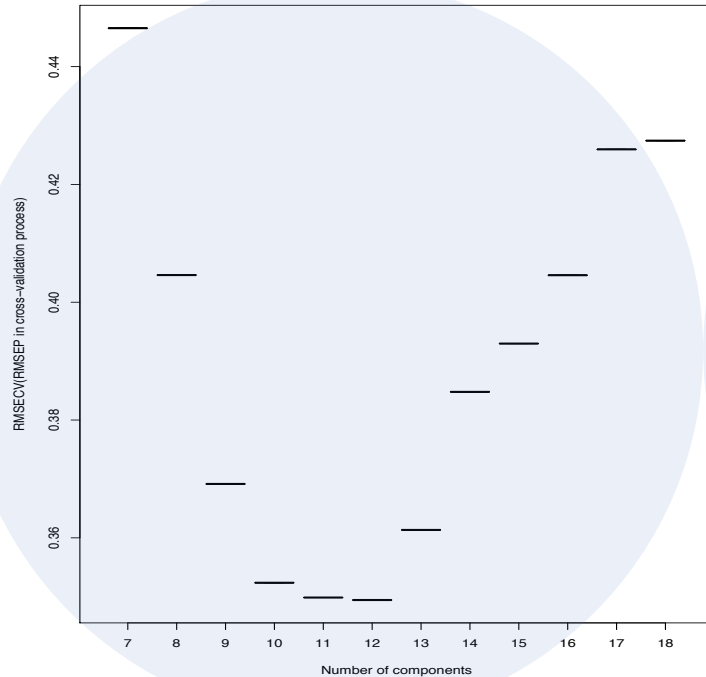
PLS of starch on mp5 with 3-fold cross-validation



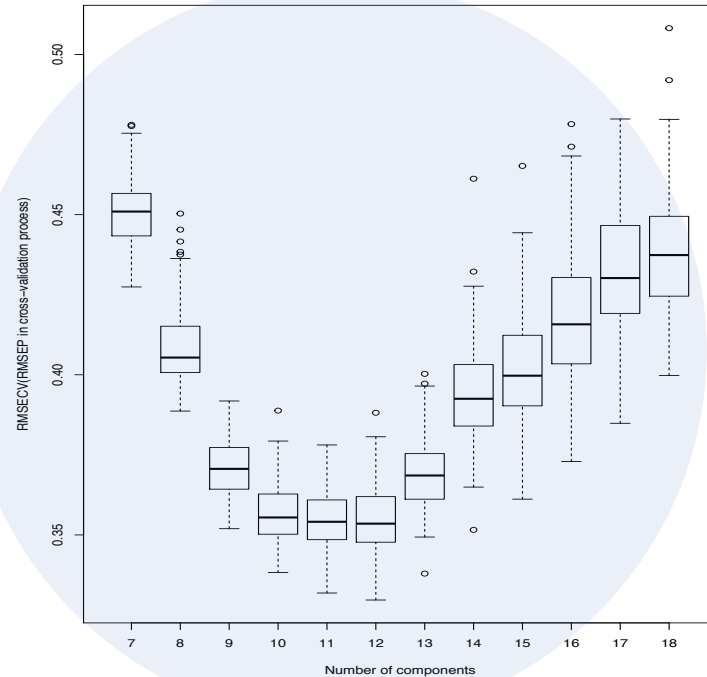
Cross-validation

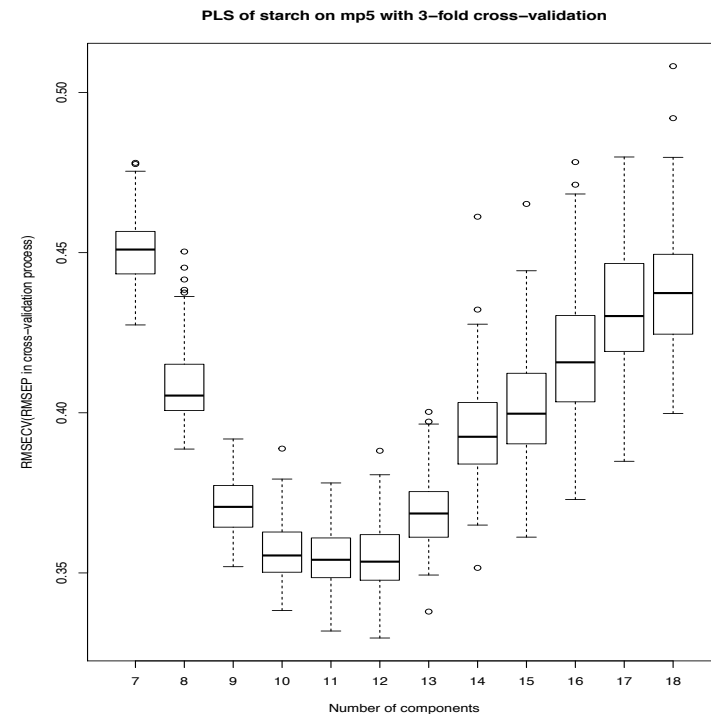
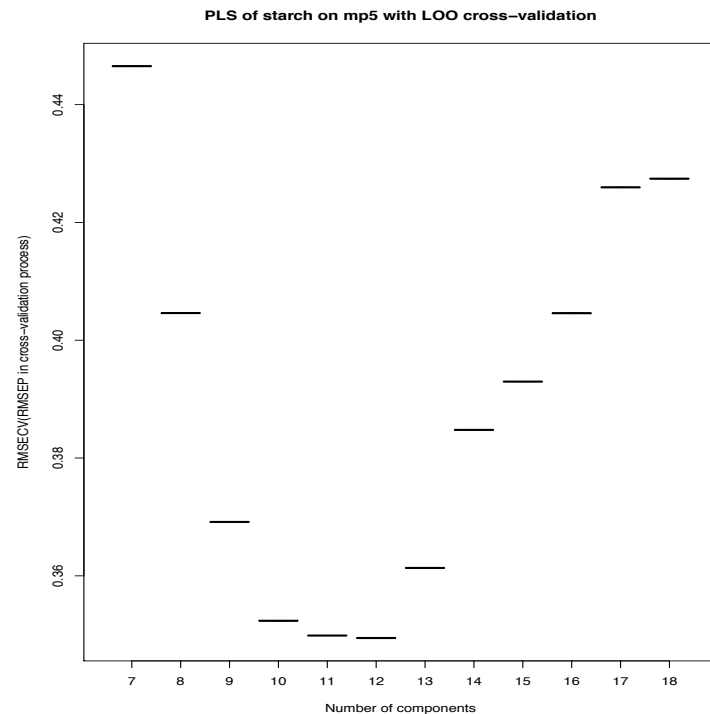


PLS of starch on mp5 with LOO cross-validation



PLS of starch on mp5 with 3-fold 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

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

Because:

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

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

×?W !"#\$%&'()*+,-./0123456789:

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

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.