The NIR Corn Data Set

Hongwei PENG

Supervisor : Prof Tom Fearn
Department of Statistical Science
University College London

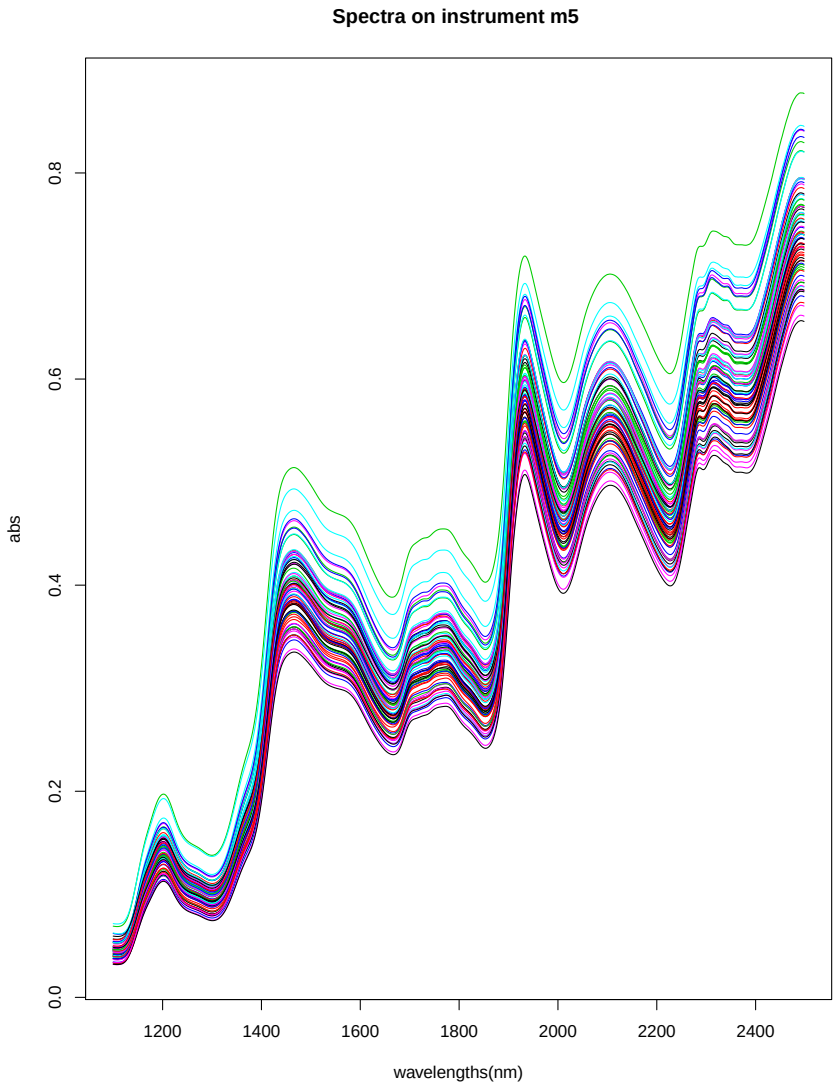
Department of Statistical Science

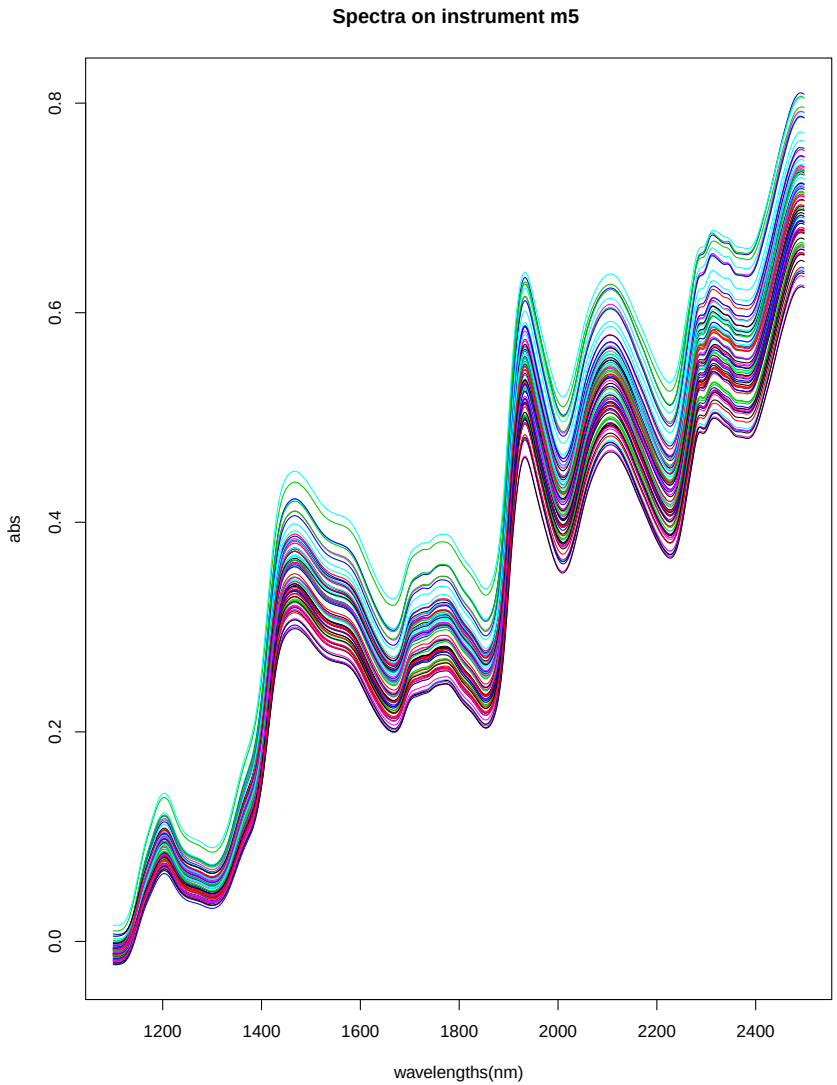
September 2, 2019

Datasets

Corn data is the most readily available high-dimensional Near-infrared spectral data. This data was published on the Internet (http://www.eigenvector.com/data/Corn/index.html) by Eigenvector Research, Inc. in 2005. The data consisted of 80 corn samples measured on three different NIR spectrometers named m5, mp5 and mp6. The spectral wavelength range is 1100$\sim$2498nm with an interval of 2nm. Hence there are 700 channels for each spectrum in each sample. Figure is the plot of spectra on m5 and mp5.

Datasets


Spectra on instrument m5
abs
0.0
0.2
0.4
0.6
0.8
1200
1400
1600
1800
2000
2200
2400
wavelengths(nm)


Spectra on instrument m5
abs
0.0
0.2
0.4
0.6
0.8
1200
1400
1600
1800
2000
2200
2400
wavelengths(nm)

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

Compare with papers

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

Table 1: F-test on regression of moisture.

Compare with papers

In section, although we calculated the difference between PLS and the developed method, there is no evaluation index to evaluate whether there is a significant difference between the two method. Because RMSEP does not obey the common distribution, which is why it is difficult to test, this section gives an approximate test method. This method is true when the model's bias is much smaller than the variance. According to section, RMSEP is calculated as follows, it can be divided into two parts, variance and bias.

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

Here, if the model's bias is much smaller than the variance, then the BIOS can be ignored. Thus, RMSEP obeys the χ^2 distribution.

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