

可见-近红外光谱测定某品种苹果的甜度

传感器实验报告

韩超越 电信 1902 2019012703

2021/6/23

一、实验综述

1、实验目的

利用已知的 120 个某品种苹果样本的不同甜度的可见-近红外光谱数据建立甜度模型，期望可以使用该含糖量预测模型对未知甜度的样本光谱曲线进行预测。

2、实验仪器、设备或软件

Win10 电脑 、软件 Matlab2020a

3、实验思路

本实验是利用近红外光谱技术对某品种苹果的甜度进行无损检测。通过 550.087~950.235nm 近红外光谱采集了 120 个苹果样本的光谱信息。分别采用多元散射校正(MSC)、标准正态变量交换(SNV)、归一化(Normalize)、数据中心化(Mean centering)、标准化(Autoscales)、移动窗口平滑、Savitzky-Golay 卷积平滑法、一阶导数(FD)、二阶导数(SD)等方法对光谱进行预处理，并采用 PCA 主成分分析 (principal component analysis)结合马氏距离法对近红外校正样品集中的异常样品进行剔除。剔除样本后使用偏最小二乘法(PLS)建立模型对样本进行定量分析。比较各种光谱预处理的方法以及剔除异常值的权重阈值的选取，获取最佳 PLS 模型便可对待测样本进行含糖量预测。

二、实验过程（实验步骤、记录、数据、分析）

1、光谱数据预处理方法

近红外光谱往往包含一些与待测样品性质无关的因素带来的干扰，导致了近红外光谱的基线漂移和光谱的不重复，为消除基线漂移、散射、高频随机噪声、样本不均匀等的影响，加强被分析信号权重，提高有用信号比例，因此须对原始光谱进行预处理。

以下为采用不同光谱预处理方法之后的某品种苹果光谱图。

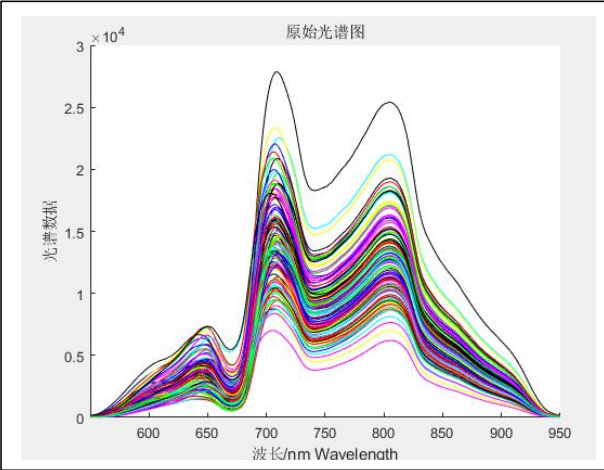


Fig1-1 原始光谱

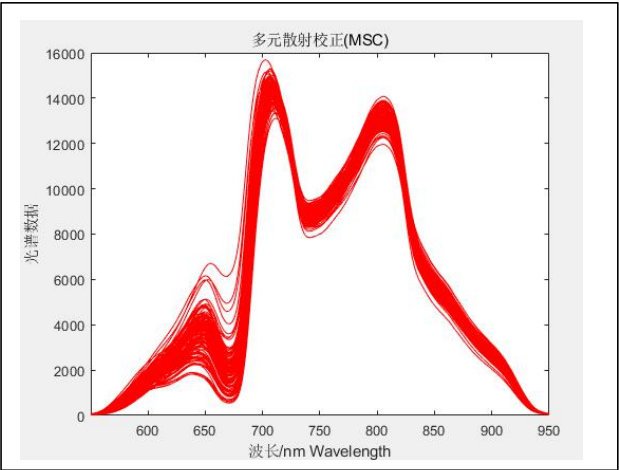


Fig1-2 多元散射校正 (MSC)

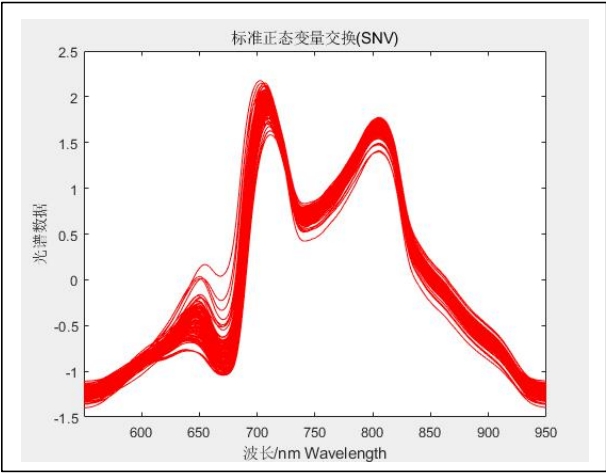


Fig1-3 标准正态变量交换(SNV)

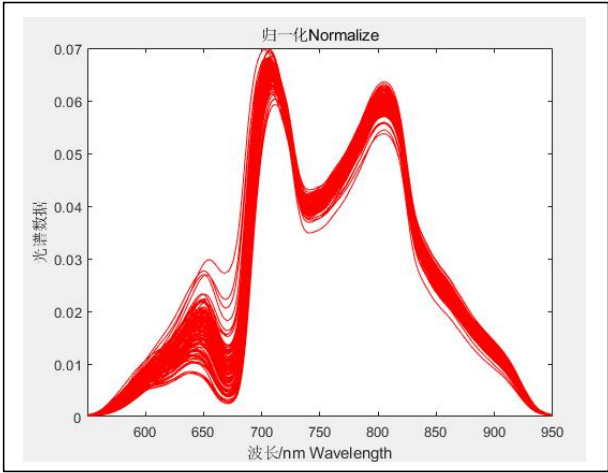


Fig1-4 归一化(Normalize)

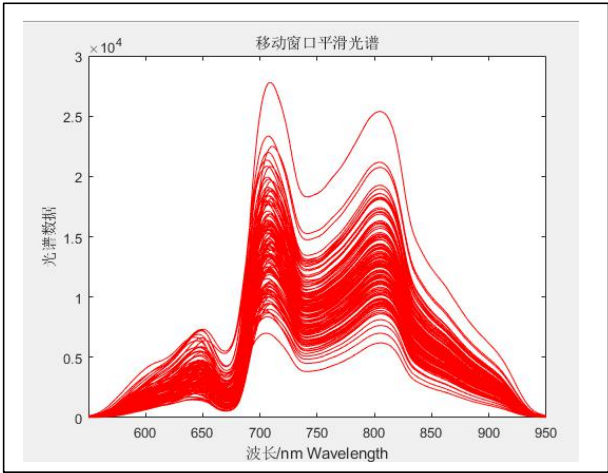
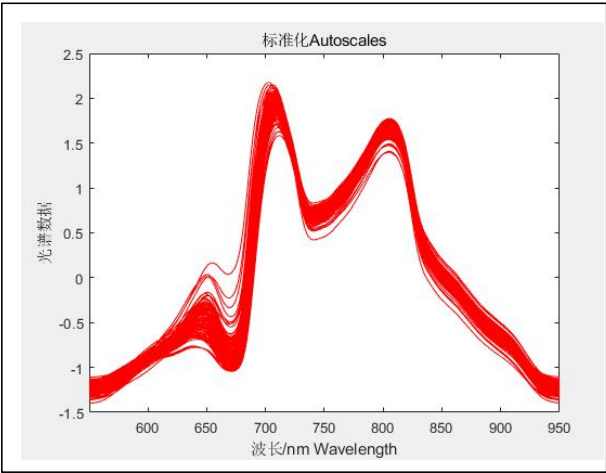


Fig1-5 标准化(Autoscales)

Fig1-6 移动窗口平滑

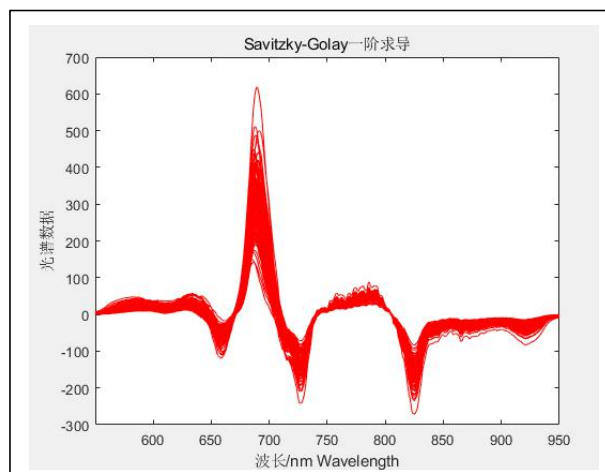
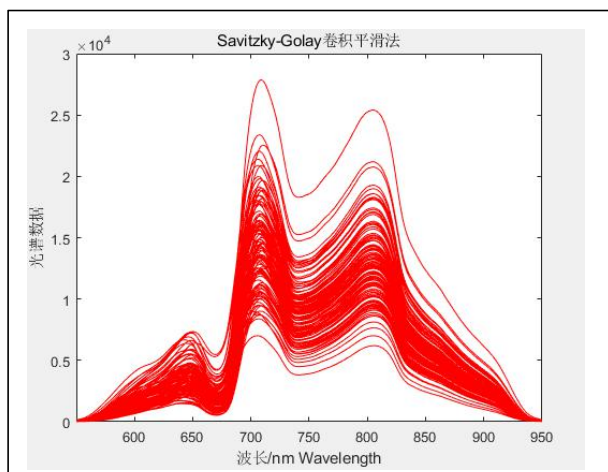


Fig1-7 Savitzky-Golay 卷积平滑法

Fig1-8 Savitzky-Golay 一阶求导

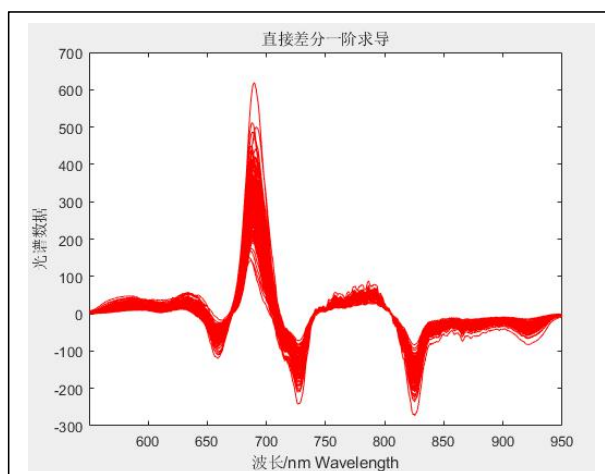
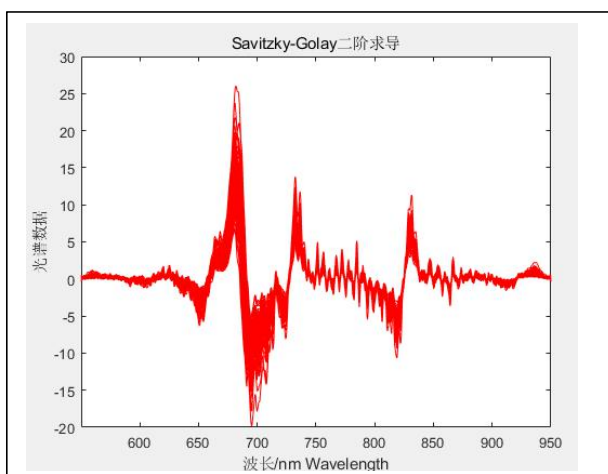


Fig1-9 Savitzky-Golay 二阶求导

Fig1-10 直接差分一阶求导

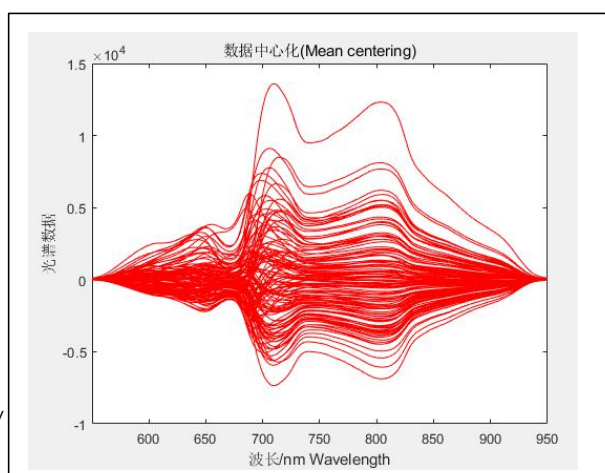
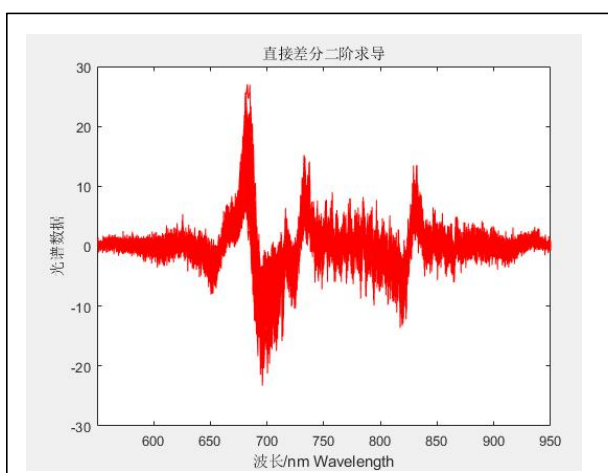


Fig1-11 直接差分二阶求导

Fig1-12 数据中心化(Mean centering)

2、主成分分析(PCA)结合马氏距离剔除异常点

(1) 主成分分析

主成分分析应用线性变换,在不丢失主要光谱信息的前提下选择维数较少的新变量来代替原来较多变量,是实现压缩光谱数据维数的一种有效方法,可以得出各个变量对各主成分的贡献,同时获取样本光谱的主成分得分。

(2) 马氏距离

在定性分析中,一般把光谱异常对应的样本定义为异常样本,认为对建模有较大影响的样本点。结合以上主成分分析(PCA)所得样本的主成分得分,引入马氏距离鉴定校正集中的异常样本。通过主成分分析求得样本的主成分得分矩阵,计算各个样本的马氏距离(D_i),并设置离群阈值(D_{th})实现异常样本的鉴定和剔除。

计算各样本的马氏距离方法如下:

$$D_i^2 = (t_i - \bar{T})M^{-1}(t_i - \bar{T})'$$

$$\bar{T} = (\sum_1^m t_i)/m$$

其中 M 为校正集光谱主成分得分矩阵的协方差阵; t_i 为样本 i 的主成分得分向量; \bar{T} 为 m 个校正集样本的平均得分矩阵; D_i 为校正集样本 i 的马氏距离。

检验校正集中的异常样本存在的阈值计算公式如下

$$D_{th} = D_m + e \cdot \sigma_d$$

给定阈值调整权重系数 e , D_m 和 σ_d 分别为 m 个样本马氏距离的平均值和标准差。凡满足 $D_i \geq D_{th}$,认为校正集中第 i 个样本是异常样本,予以剔除;反之 $D_i < D_{th}$,认为校正集中样本 i 的光谱在主成分空间中相似。

针对以上不同的 e 值所选取的阈值范围,分别使用PLS建模回归预测,来进行阈值范围的选取。

3、PLS 模型预测

偏最小二乘法PLS算法在分析高度共线性的数据集表现出了良好的效果,适

用于独立变量数大于样本数的数据集分析,利用交互有效性原则来确定最佳主成分数。其主要思想是利用预测光谱残差平方和(PRESS)确定主成分数,进而评价模型的预测能力。

三、结果与分析

1、样本划分

先把 120 份某品种苹果样本顺序随机打乱,目的是在划分样本时使含糖量数据在测试集和校正集中分散均匀,然后按照 3:1 的比例把样本数据分成校正集和预测集。其统计结果如表 3.1 所示。

表 3.1 甜度数据统计

Table3.1 Statistics of moisture contents	
	样本数
校正集	90
预测集	30
总数	120

2、不同光谱预处理的主成分分析

为了比较不同光谱预处理方法下的主成分聚类效果,分别对原始光谱进行多元散射校正(MSC)、标准正态变量交换(SNV)、归一化(Normalize)、数据中心化(Mean centering)、标准化(Autoscales)、移动窗口平滑、Savitzky-Golay 卷积平滑法、一阶导数(FD)、二阶导数(SD)等方法对光谱进行预处理,然后求得主成分,比较前 15 个主成分特征值贡献率,见图 3-1.

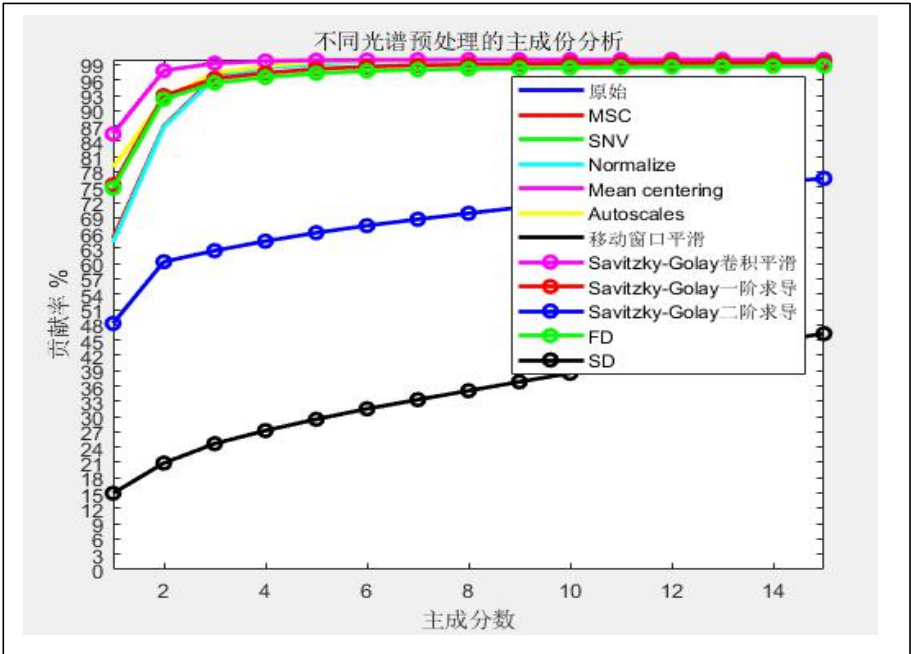


Fig3-1 不同光谱预处理的前 15 个主成分累计贡献率

具体累计贡献率数据如下：

percent_explained											
89x10 double											
	1	2	3	4	5	6	7	8	9	10	11
1	85.4197	64.9488	78.9321	64.1526	85.4197	78.9321	85.4609	85.4199	75.4289	48.2683	
2	97.8267	86.9826	92.3556	86.7484	97.8267	92.3556	97.8425	97.8268	92.9347	60.3688	
3	99.2900	96.6766	97.3766	96.3919	99.2900	97.3766	99.2997	99.2901	96.2784	62.5326	
4	99.7119	98.6215	98.7471	98.3885	99.7119	98.7471	99.7164	99.7120	97.3770	64.3937	
5	99.8630	99.2461	99.5897	99.0859	99.8630	99.5897	99.8656	99.8631	98.1739	66.0545	
6	99.9158	99.6728	99.7935	99.6400	99.9158	99.7935	99.9178	99.9160	98.6268	67.4399	
7	99.9642	99.8067	99.8836	99.7743	99.9642	99.8836	99.9649	99.9643	98.9013	68.6877	
8	99.9792	99.8791	99.9326	99.8497	99.9792	99.9326	99.9798	99.9793	99.0345	69.8575	
9	99.9865	99.9123	99.9547	99.8980	99.9865	99.9547	99.9870	99.9866	99.1588	70.9933	
10	99.9915	99.9408	99.9698	99.9301	99.9915	99.9698	99.9918	99.9916	99.2362	72.0519	
11	99.9941	99.9615	99.9810	99.9544	99.9941	99.9810	99.9944	99.9942	99.2971	73.0857	
12	99.9959	99.9728	99.9873	99.9655	99.9959	99.9873	99.9962	99.9960	99.3519	74.0455	
13	99.9976	99.9817	99.9912	99.9747	99.9976	99.9912	99.9977	99.9977	99.4043	74.9499	
14	99.9983	99.9884	99.9931	99.9825	99.9983	99.9931	99.9985	99.9984	99.4530	75.8413	
15	99.9987	99.9903	99.9947	99.9889	99.9987	99.9947	99.9989	99.9988	99.4913	76.7122	
16	99.9990	99.9920	99.9957	99.9907	99.9990	99.9957	99.9991	99.9991	99.5243	77.5753	
17	99.9992	99.9931	99.9963	99.9923	99.9992	99.9963	99.9993	99.9993	99.5545	78.3844	
18	99.9994	99.9939	99.9967	99.9934	99.9994	99.9967	99.9995	99.9995	99.5818	79.1606	
19	99.9995	99.9947	99.9971	99.9942	99.9995	99.9971	99.9996	99.9996	99.6082	79.9082	
20	99.9996	99.9952	99.9973	99.9949	99.9996	99.9973	99.9997	99.9997	99.6326	80.6208	
21	99.9997	99.9957	99.9976	99.9954	99.9997	99.9976	99.9998	99.9997	99.6551	81.3265	
22	99.9997	99.9961	99.9978	99.9958	99.9997	99.9978	99.9998	99.9998	99.6745	82.0131	
23	99.9997	99.9964	99.9980	99.9962	99.9997	99.9980	99.9998	99.9998	99.6923	82.6728	
24	99.9998	99.9967	99.9981	99.9966	99.9998	99.9981	99.9999	99.9998	99.7085	83.3001	
25	99.9998	99.9970	99.9983	99.9968	99.9998	99.9983	99.9999	99.9999	99.7242	83.9079	
26	99.9999	99.9972	99.9984	99.9971	99.9999	99.9984	99.9999	99.9999	99.7397	84.5074	

Fig3-2 具体累计贡献率数据

从图 3-1 和图 3-2 中可以看出，多元散射校正(MSC)和归一化(Normalize)的第一主成分贡献率达到 64.9488%和 64.1526%，前 2 个累积贡献率大于 86.5%。结合主成分聚类效果，多元散射校正(MSC)和归一化(Normalize)效果不错，之后选择了这两种光谱预处理的方法的前 2 个主成分得分进行马氏距离计算。

3、剔除异常品对预测效果的研究

为了在后续偏最小二乘法建立的判别模型精度更高，应用马氏距离法鉴别异常样品(对于该光谱作业而言，我们并不知道具体样本的情况，为了判别的准确率，需要考虑剔除异常值)。

本实验设置阈值调整权重系数 $e = 0.1:0.1:3$;校正集中的90个样本的主成分

马氏距离如图3-3所示。选择不同的权重系数 e ，得到不同的阈值，马氏距离大于设定阈值时，对应的样本予以剔除。随着 e 由小变大，剔除样品的个数由多变少。剔除样品后，采取PLS分别建立模型，其主成分数的选择采用交互验证(cross validation)方法来选取。在不同参数 e 下建立的PLS模型效果，选择RMSEC最小时所对应的参数 e 和其剔除异常值后的校正集所建立的PLS作为最终确定的定量校正模型。

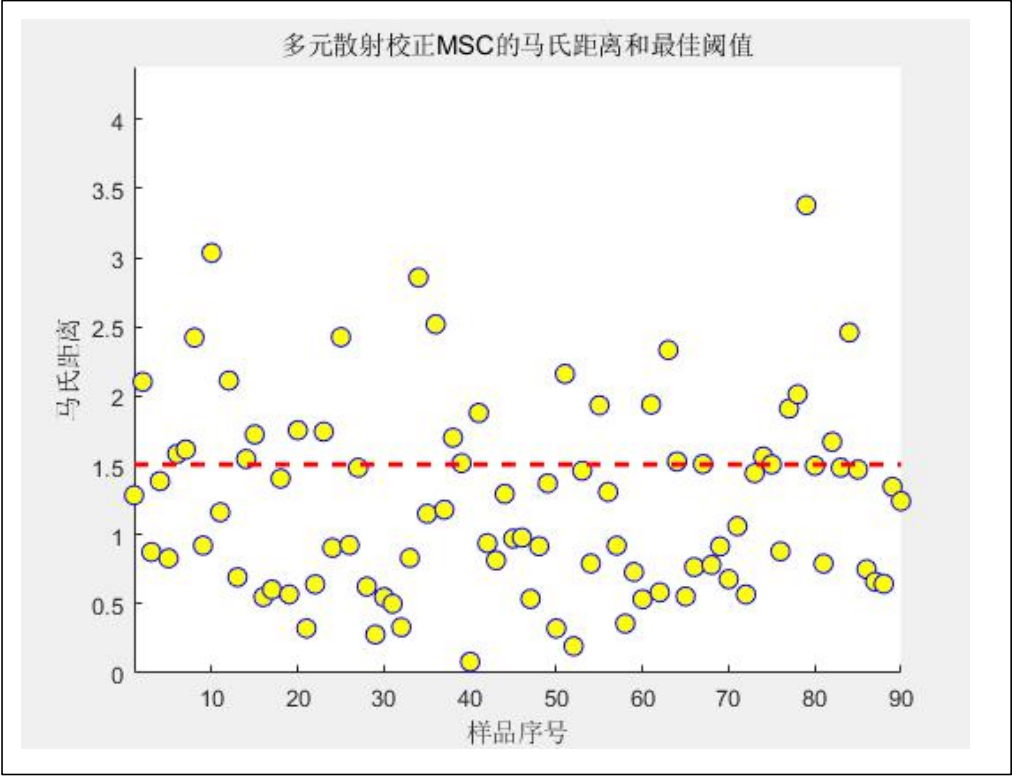


Fig3-3 多元散射校正(MSC)光谱预处理的马氏距离和其最佳阈值($e=1.60$)

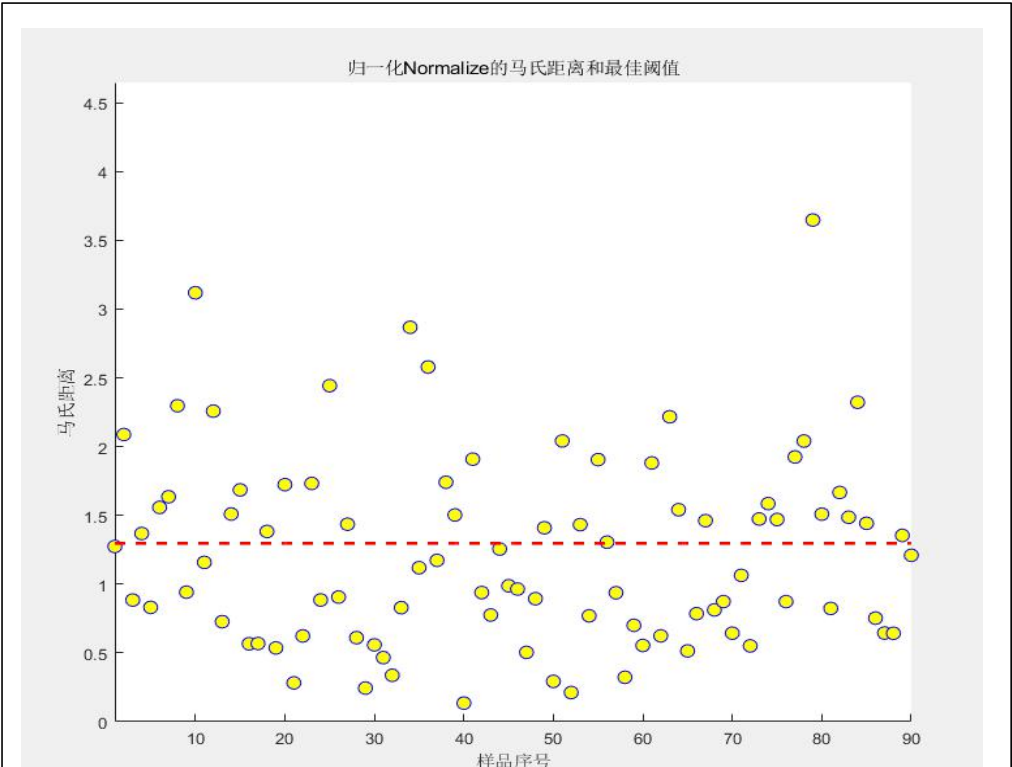


Fig3-4 归一化(Normalize)光谱预处理的马氏距离和其最佳阈值($e=0.10$)

4、建模预测效果比较

以校正集样品的甜度与预测甜度的相关系数 R_{train} ，预测集样品的甜度和预测甜度的相关系数 R_{test} ，校正集样本均方根误差(root mean squared error of calibration, RMSEC) 及预测集样本均方根误差(root mean squared error of prediction, RMSEP) 作为评价PLS模型优劣的指标。相关系数越高、均方根误差越低且主因子数越小，则PLS模型的预测效果越好。

按照实验步骤应当逐渐测试出最佳建模参数进行PLS模型建立，由于时间关系仅测试了有限的几个值。有限测试选择建模参数下建立的PLS模型效果如表3.2所示。

表 3.2 最佳建模参数下建立的 PLS 模型效果

预处理方法	权重阈值 e	PLS 主因子个数	校正集(90 个)		预测集(30 个)	
			R_{train}	RMSEC	R_{test}	RMSEP
MSC	1.60	11	0.7159	0.6256	0.6634	0.6160
Normalize	0.10	7	0.6349	0.7112	0.7494	0.5141

多元散射校正光谱预处理以及最佳阈值($e=1.60$)时的 PLS 预测结果如下：

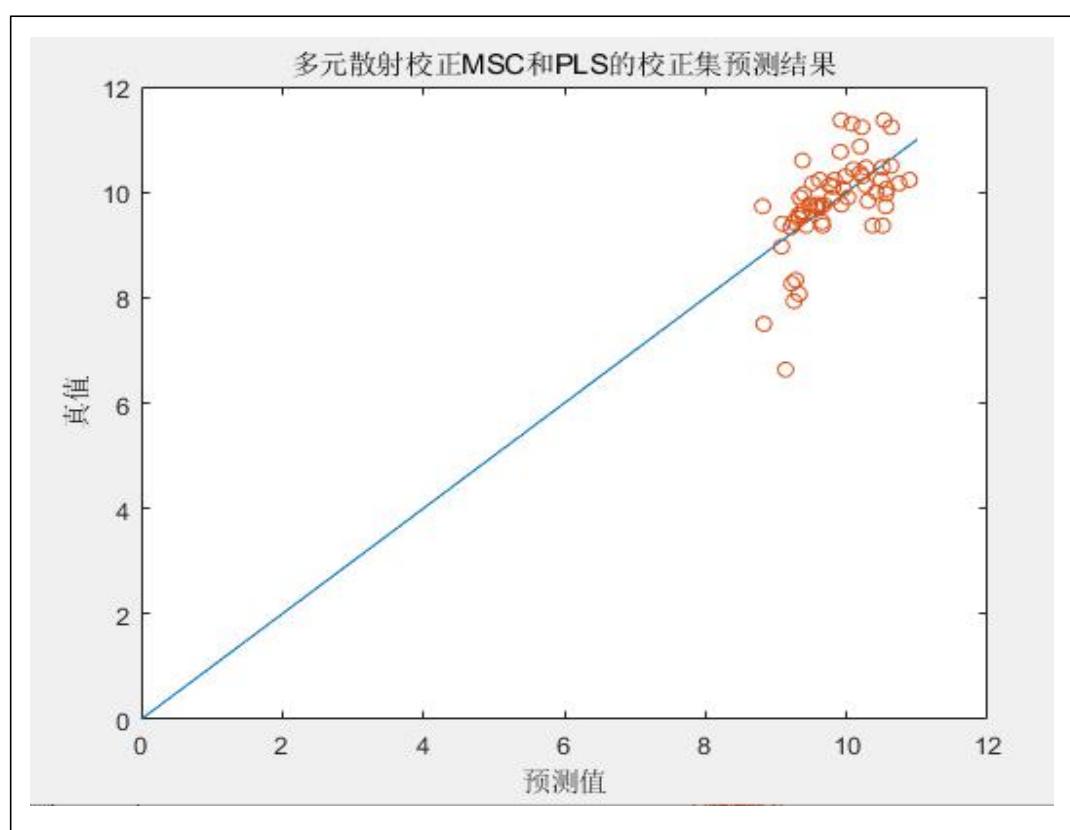


Fig3-5 多元散射校正(MSC)剔除异常点后校正集的预测值和真值之间的关系

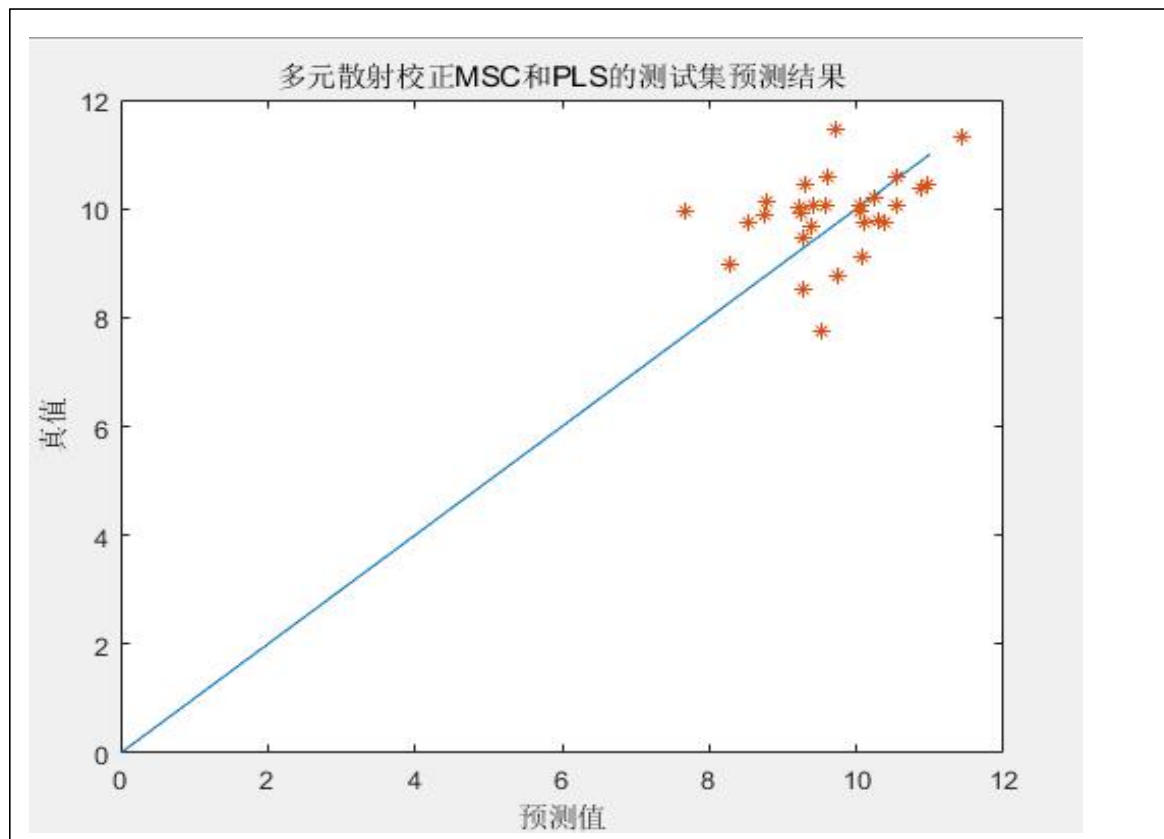


Fig3-6 多元散射校正(MSC)测试集的预测值和真值之间的关系

归一化(Normalize)光谱预处理以及最佳阈值($e=0.10$)时的 PLS 预测结果如下:

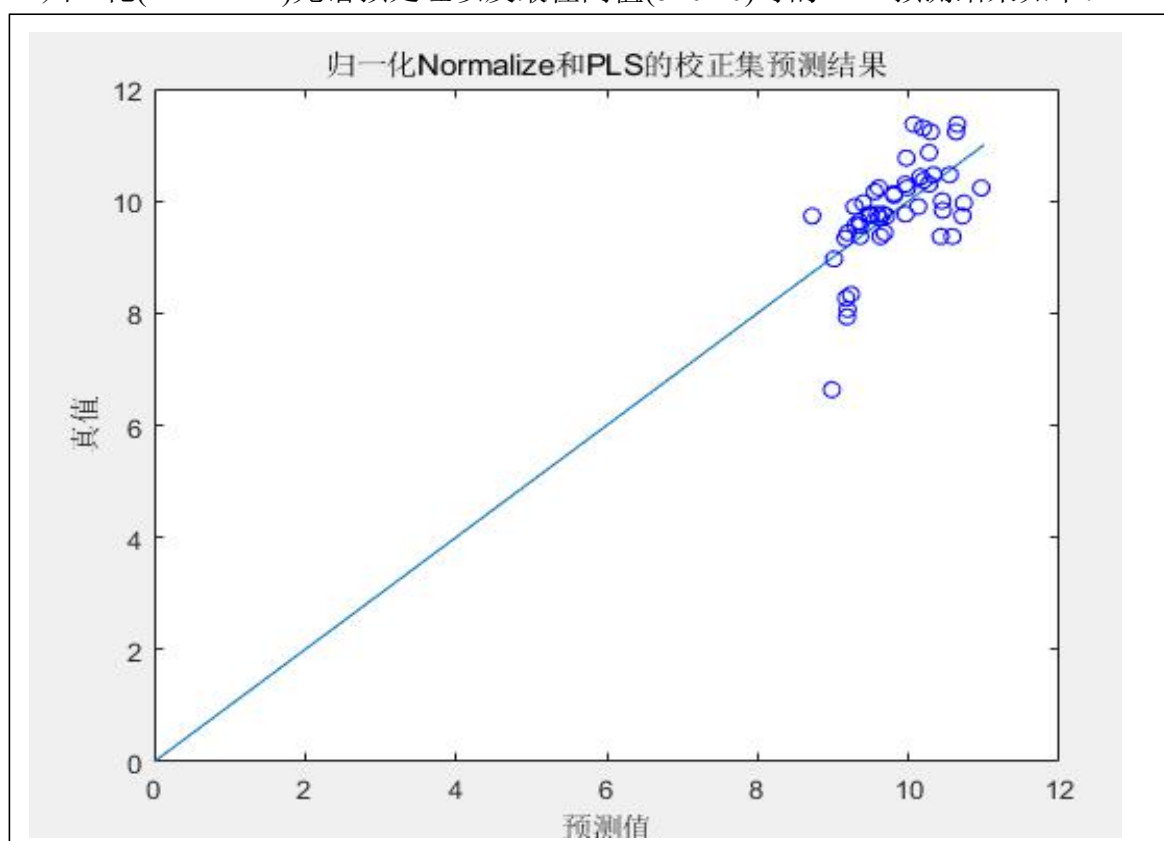


Fig3-7 归一化(Normalize)剔除异常点后校正集的预测值和真值之间的关系

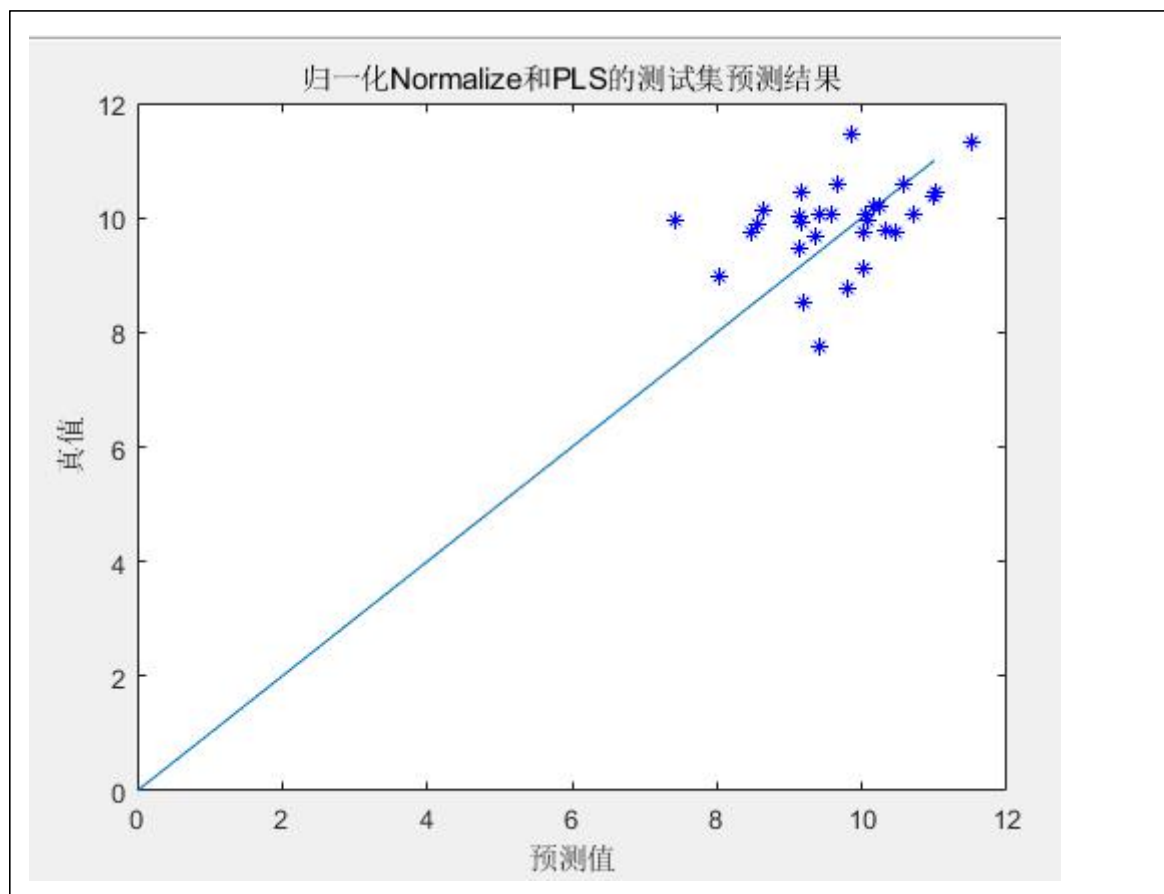


Fig3-8 归一化(Normalize)测试集的预测值和真值之间的关系

通过对表 3.2 以及多元散射校正(MSC)和归一化(Normalize)光谱预处理定值与预测值的散点图综合分析，使用多元散射校正(MSC)预处理方法建立了 PLS 模型。

可以看出 PLS 模型建立效果不是很好，后续需要继续改进建模参数，增加数据集。

三、结论

1、实验结果































我们获得的模型：利用 PCA 对校正集分析，多元散射校正 (MSC) 光谱预处理的前两个主成分累计贡献率达 86%，利用样本的主成分得分结合马氏距离法，选择了阈值调整权重系数 $e = 1.60$ ，剔除了异常样本。剩余的样本作为定标建模样本，交互验证法验证后建立 PLS 模型，对测试集预测。PLS 模型的预测结果显示预测值和实际值相关度不高，需要后续改进模型。

四、参考文献

- [1] 李丹,何建国,刘贵珊等.基于高光谱成像技术的小黄瓜含糖量无损检测[J]. 红外与激光工程, 2014,43(7): 2393-2397.
- [2] 吕萍,薛龙,何秀文等. 生姜甜度的可见-近红外光谱检测[J].江西农业大学学报.2011, 33(3) :0602-0607.
- [3] 屠振华,冯霖,孙丽娟等. 近红外光谱测定蜂蜜中甜度特征波长选择和分析研究[C]. 科学仪器服务民生学术大会论文集,2011.
- [4] 王海青,姬长英,陈坤杰.基于光谱分析技术的黄瓜与茎叶识别研究[J].光谱学与光谱分析.2011,31(10):2834-2838.
- [5] 陈斌,邹贤勇,朱文静.PCA 结合马氏距离剔除近红外异常样品[J].江苏大学学报.2008,29(4):0277-0281.
- [6] 褚莹,丁武,齐强强.基于近红外光谱技术实现掺假山羊奶的定性和定量检测[J].西北农业学报.2011,20(2):192-196.
- [7] 沈林峰,沈掌泉.应用遗传算法和PLS的近红外光谱预测玉米中淀粉含量的研究[J].分析测试技术与仪器.2008,14(4):214-217.
- [8] 李晓丽,何勇,袁正军.一种基于可见-近红外光谱快速鉴别茶叶品种的新方法[J].光谱学与光谱分析.2007,27(2):279-282.
- [9] 赵羚志.短波近红外光谱技术结合人工神经网络用于药物无损定量分析的研究[D].吉林大学,2009.
- [10] 汤真,刘福强,苟玉慧. 粉末药品安体舒通的无损定量-分析人工神经网络-近红外光谱法的应用[J].分析测试学报.2001,20(3):0062-0065.

思路来源于论文，代码来源于csdn，数据对模型的适用性不够好，需要后续的改进。

文件夹：数据及代码预览

 data	2021/6/22 18:19	文件夹	
 预处理工具箱	2021/6/22 18:25	文件夹	
 best_MSC_pls.mat	2021/6/22 18:17	MATLAB Data	8 KB
 best_Nor_pls.mat	2021/6/22 18:17	MATLAB Data	1 KB
 choose4.xls	2021/6/22 16:49	XLS 工作表	9 KB
 data_chuli001.m	2018/10/22 10:13	MATLAB Code	5 KB
 data_divide.m	2018/10/19 11:38	MATLAB Code	1 KB
 data_ys_divide.mat	2018/10/26 21:33	MATLAB Data	911 KB
 data_ys_divide08.mat	2021/6/22 18:17	MATLAB Data	493 KB
 determinand1.xlsx	2021/6/22 16:02	XLSX 工作表	40 KB
 Distance_maha.m	2018/10/18 13:40	MATLAB Code	2 KB
 erase_baddata001.m	2018/10/16 21:08	MATLAB Code	2 KB
 erase_baddata002.m	2018/10/20 14:53	MATLAB Code	9 KB
 EraseAndPLSpredict001.m	2021/6/22 17:10	MATLAB Code	5 KB
 main_001.m	2021/6/22 18:16	MATLAB Code	3 KB
 MSC_erase_PLS.m	2018/10/21 13:08	MATLAB Code	4 KB
 n3.xls	2021/6/22 16:53	XLS 工作表	11 KB
 Nor_erase_PLS.m	2018/10/29 9:35	MATLAB Code	4 KB
 norm_pca.m	2018/10/16 10:08	MATLAB Code	1 KB
 PLS_data.mat	2021/6/22 18:17	MATLAB Data	517 KB
 pls_plot.m	2018/10/19 9:09	MATLAB Code	1 KB
 PLS_regression.m	2018/10/20 10:05	MATLAB Code	4 KB
 Readme.doc	2018/10/29 15:58	DOC 文档	100 KB
 spectral_data&water_content1.xlsx	2021/6/22 16:01	XLSX 工作表	1,373 KB
 test1.xls	2021/6/22 16:47	XLS 工作表	251 KB
 tongji.m	2018/10/19 11:15	MATLAB Code	2 KB
 train2.xls	2021/6/22 16:49	XLS 工作表	739 KB
 water_content1.xlsx	2021/6/22 16:00	XLSX 工作表	10 KB
 wavelength1.xlsx	2021/6/22 16:00	XLSX 工作表	15 KB
 ys.png	2018/10/21 21:49	PNG 文件	96 KB