第 39 卷 第 10 期
2020 年 10 月
分析测试学报
FENXI CESHI XUEBAO(Journal of Instrumental Analysis)
Vol. 39 No. 10
1231 ~ 1238

研究报告

doi：10.3969/j.issn.1004 - 4957.2020.10.008

偏最小二乘近红外光谱模型中潜变量个数对模型传递性能的影响

李永琪[1]，洪士军[1]，黄　雯[2]，张立国[1]，葛　炯[2*]，栾绍嵘[1]，倪力军[1*]

（1. 华东理工大学　化学与分子工程学院，上海　200237；2. 上海烟草集团有限责任公司
技术中心理化实验室，上海　200082）

摘　要：以玉米中水分、蛋白质、脂肪和淀粉 4 种主要成分含量以及烟叶总植物碱的偏最小二乘近红外光谱(PLS - NIRs)模型传递为例，考察了模型中潜变量个数(nLVs)对模型传递误差的影响。研究发现，根据累积贡献率大于 99.9% 确定的玉米、烟叶样品 PLS - NIRs 模型的 nLVs 分别为 1 和 13，nLVs = 1 时建立的玉米模型对两台从机样品 4 个成分的预测值和主机预测值的重现性指标均满足国标要求；nLVs = 13 时建立的烟叶总植物碱模型经分段直接校正(PDS)后，可使 4 台从机样品的平均相对预测误差(MRE)小于 6% 。采用留一交叉验证或四折交叉验证确定的玉米、烟叶 PLS - NIRs 模型的 nLVs 分别为 5 ~ 10，16 与 19，在这些 nLVs 下建立的玉米 PLS - NIRs 模型对从机样品的预测误差显著增大，超过许可的误差范围，且模型即使经 PDS 校正后，从机样品预测值与主机样品预测值的重现性指标大多不满足国标要求；nLVs > 13 时所建烟叶总植物碱 PLS - NIRs 模型的转移误差随 nLVs 增大而增大，且 PDS 校正后不能保证模型对所有从机样品的 MRE 小于 6% 。根据累积贡献率大于 99.9% 或接近 99.9% 为准则选取 nLVs，可有效避免过拟合，提高 NIRs 模型的传递性能。

关键词：近红外光谱模型传递；偏最小二乘；潜变量个数；玉米；烟叶
中图分类号：O657.3　　**文献标识码**：A　　**文章编号**：1004 - 4957(2020)10 - 1231 - 08

Effect of Number of Latent Variables for Partial Least Square Model Based on Near Infrared Spectroscopy on Models Transfer Performance

LI Yong-qi[1]，HONG Shi-jun[1]，HUANG Wen[2]，ZHANG Li-guo[1]，GE Jiong[2*]，
LUAN Shao-rong[1]，NI Li-jun[1*]

（1. College of Chemistry and Molecular Engineering，East China University of Science and Technology，Shanghai　200237，China；2. Technology Center Psychological Laboratory，Shanghai Tobacco Group Co.，Ltd.，Shanghai　200082，China）

Abstract：Using the calibration model transfer of PLS - NIRs models for predicting contents of moisture，protein，fat and starch in corn，as well as total alkaloids in tobacco leaves as an example，effect of number of latent variables(nLVs) on the transfer errors of the models were investigated in this paper. It was found that the nLVs in PLS - NIRs models for corn and tobacco leaves selected by cumulative contribution rate greater than 99.9% were 1 and 13，respectively. The prediction reproducibilities for the four ingredients in corn between master and slave samples predicted by the PLS - NIRs models with one latent variable all satisfied the requirements of national standards. When the PLS - NIRs model predicting total alkaloids content built on the master with 13 latent variables was transferred to four slaves，mean of relative prediction errors(MRE) of tobacco leaves tested on the four slaves were all lower than 6% after piecewise direct standardization(PDS) correction. While the nLVs in PLS - NIRs models for corn and tobacco leaves determined by leaving one sample in turn as cross validation set or fourth-fold cross validation method were 5 - 10，16 and 19，respectively. The

收稿日期：2020 - 07 - 02；**修回日期**：2020 - 08 - 07
基金项目：国家烟草专卖局卷烟烟气重点实验室开放性课题(K2018 - 156P)
* **通讯作者**：倪力军，博士，教授，研究方向：分子光谱技术及其应用，E - mail：nljfyt@163.com
葛　炯，工程师，研究方向：烟草化学和光谱技术，E - mail：gej@sh.tobacco.com.cn

prediction errors for the slave corn samples derived from the models with nLVs greater than 5 were significantly increased and exceeded the allowable error level. Even after being corrected by PDS method, most indices of prediction reproducibility for the four ingredients in corn between master and slave samples given by these models could not satisfy the requirements of national standards. The transfer errors of PLS – NIRs models for total alkaloids in tobacco leaves by selecting nLVs greater than 13 increased with the increase of nLVs, while PDS correction cannot guarantee the MRE for all slave instruments given by these models lower than 6%. Results indicated that selecting nLVs for PLS – NIRs models based on the principle of accumulative contribution rate greater than 99.9% or near to 99.9% could effectively avoid over-fitting and improve the transfer performance of the models.

Key words: near infrared spectroscopy model transfer; partial least square; number of latent variables; corn; tobacco

近红外光谱(NIRs)技术作为一种快速、无损的绿色检测技术，在各行各业的定量与定性分析中得到了广泛应用[1]。该技术以一些具有代表性的定标样品的定量指标或定性指标为因变量，其近红外光谱信息为自变量，通过多元统计方法建立相关指标的近红外光谱定量模型或样品的定性模型，根据模型实现对未知样品的定量或定性分析[2]。建立一个良好的近红外光谱模型需要积累大量样品的光谱和待测性质数据，并优化模型中的相关参数，模型建立和维护的工作量较大。通常希望在一台机器上建立的光谱模型(该机器通常称为主机)能够转移到其他仪器上(简称为从机)继续使用[3]，简称为模型传递或模型共享[4-6]。但由于主、从机光谱在不同区域存在或大或小的差异，通常光谱模型传递到从机后误差会增大，因而出现了各种降低模型对从机样品预测误差的模型传递方法[7]。分段直接校正(Piecewise direct standardization, PDS)方法是最经典常用的模型传递方法，该方法以主、从机均测试的转移集样品为基础，通过对从机光谱分段校正后再应用主机模型预测从机样品[8]。

近红外光谱定量模型通常采用偏最小二乘(Partial least squares, PLS)方法建立样品光谱信息与待测物质信息间的数学模型[9]。PLS 模型建立过程中需要确定潜变量的个数(nLVs)，一般采用留一交叉验证或四折(三折)交叉验证的方法确定 nLVs[10]或是选取内部检验集样品预测误差最小时对应的潜变量个数作为最佳值。本课题组研究发现，采用这种原则确定的近红外光谱 PLS 模型通常能够对单台仪器给出不错的结果，但这样选取的 nLVs 往往个数偏多，会引入噪声和无效信息，导致模型传递时预测误差显著增大，使得模型不能在从机直接应用。本文以网上公开发布的玉米数据及烟草企业多台近红外仪器所测烟叶样品数据为例，探究 nLVs 的选取对主、从机模型误差的影响，为建立稳健、可共享的近红外光谱模型提供依据和支持。

1 实验与方法

1.1 样品与数据集

玉米样品数据集来自 http://www.eigenvector.com/data/Corn/corn.mat。包含 M5、MP5、MP6 3 台近红外仪上测得的 80 个玉米样品的近红外光谱及这些样品中主要营养成分的含量数据。玉米样品中水分的质量分数在 9.38%~10.99% 之间，均值为 10.23%；蛋白质的质量分数在 7.65%~9.71% 之间，均值为 8.67%；脂肪的质量分数在 3.09%~3.83% 之间，均值为 3.50%；淀粉的质量分数在 62.84%~66.47% 之间，均值为 64.69%。烟叶样品有 2 套数据集，Set A 由 78 个烟叶样本分别在主机 M(Master)、4 台从机 S1、S2、S3 和 S4 上测得的近红外光谱组成，5 台近红外仪均为 Antaris Ⅱ 近红外仪器(赛默飞世尔科技有限公司)，生产年份不尽相同；Set B 则由 1 070 个在主机 M 上测得的烟叶样本光谱组成。Set A、Set B 中各烟叶样品的总植物碱采用 YC/T 160 – 2002[11]测定，其含量在 0.55%~6.30% 之间。

1.2 模型建立与评价

根据课题组前期研究结果，采用标准正态变换(SNV)结合一阶导数进行 31 点平滑对样品的近红外光谱进行预处理可消除因散射和背景漂移引起的光谱误差，基于该预处理光谱所建模型与其他预处理光谱(多元散射校正、一阶导数、原始光谱等)模型的效果相当[12-13]。由于该法不需要使用其他样品

的光谱信息，故本文采用 SNV + 一阶导数光谱建立玉米中主要营养成分及烟叶总植物碱的近红外光谱定量模型。采用蒙特卡洛采样(Monte - Carlo Sampling，MCS)方法剔除异常点[14]。采用综合考虑光谱与待测性质信息来筛选代表性样品的 SPXY(Sample set partitioning based on joint $x-y$ distance)方法[15]挑选主机建模样本，剩余样品作为内部验证集。一般情况下采用建模集均方根残差(RMSEC)来评价模型的拟合性能，验证集的均方根残差(RMSEP)来评价模型的预测性能[2]。考虑到 RMSEP 相当于绝对误差，难以根据该指标判断模型误差的相对大小，本文增加检验集或从机样本模型预测值与实测值相对误差的绝对值均值(简称为平均相对误差，MRE)来评估模型对主、从机样本的预测性能。另外，为与国标[16-18]要求的评估指标相对应，本文还采用验证样品组分的近红外模型值扣除系统偏差后与其标准值(实测值)之间的校准标准差(SEP)来评估主机模型调整后的准确度。相关评价指标的计算公式如下：

$$\mathrm{RMSEP}=\sqrt{\frac{\sum_{i=1}^{m}(y_{i,actual}-y_{i,predicted})^2}{m-1}} \tag{1}$$

$$\mathrm{MRE}=\frac{\sum_{i=1}^{m}\left(\frac{|y_{i,predicted}-y_{i,actual}|}{y_{i,actual}}\right)}{m} \tag{2}$$

$$\mathrm{SEP}=\sqrt{\frac{\sum_{i=1}^{m}(y_{i,predicted}-y_{i,actual}-\mathrm{biasm})^2}{m-1}} \tag{3}$$

$$\mathrm{biasm}=\frac{1}{m}\sum_{i=1}^{m}(y_{i,predicted}-y_{i,actual}) \tag{4}$$

式(1)~(4)中 $y_{i,actual}$为第 i 个样品的实测值，$y_{i,predicted}$为第 i 个样品的模型预测值，m 为检验集样品数目。biasm 是系统偏差，即检验集样品 i 的近红外测定值与标准值(实测值)之差的均值。如果不考虑系统偏差校正，式(3)的 SEP 即为式(1)的 RMSEP。

PLS 回归分析时前 n 个潜变量(主因子)的方差之和占所有潜变量方差之和的百分比 η 称为累积贡献率，其计算公式如下：

$$\eta=\sum_{i=1}^{n}\lambda_i/\sum_{i=1}^{p}\lambda_i \tag{5}$$

式(5)中 λ_i 为第 i 个潜变量的方差，p 为所有方差不为零的潜变量个数，$p\leqslant\min\{$样本数，波长个数$\}$。

对于从机，采用 RMSEP、MRE 评价模型转移后的准确度，采用重现性指标 SR 评价从机近红外测定结果与主机近红外测定结果的一致性。国标[16]定义玉米水分、蛋白质近红外模型测定结果再现性指标 SR 的计算公式如下：

$$\mathrm{SR}=\sqrt{\frac{\sum_{i=1}^{m}(y_{i,slave}-y_{i,master}-\mathrm{biast})^2}{m-1}} \tag{6}$$

$$\mathrm{biast}=\frac{1}{m}\sum_{i=1}^{m}(y_{i,slave}-y_{i,master}) \tag{7}$$

式(6)与(7)中的 $y_{i,slave}$与 $y_{i,master}$分别表示样品 i 的从机近红外测定值和主机近红外测定值；biast 为验证样品 i 的从机近红外测定值与主机近红外测定值之差的均值，m 为检验集(预测集)样本个数。

对于玉米中的脂肪与淀粉，国标要求在不同实验室，由不同操作人员使用同一型号不同设备，按相同测试方法，对相同的玉米样品的两个脂肪独立实验结果之间的绝对差值应不大于 0.3%[17]，对相同的玉米样品的两个淀粉独立实验结果之间差值应不大于其算术平均值的 15%[18]。参照国标的上述描述，本文定义玉米中脂肪、淀粉的再现性评价指标 SRo 与 SRs 如下：

$$\mathrm{SRo}=\frac{\sum_{i=1}^{m}|y_{i,slave}-y_{i,master}|}{m} \tag{8}$$

$$\mathrm{SRs}=\frac{\sum_{i=1}^{m}\frac{|y_{i,slave}-y_{i,master}|}{y_{i,m}}}{m} \tag{9}$$

式(9)中的 $y_{i,m}$为样品 i 的主机近红外测定值 $y_{i,master}$与从机近红外测定值 $y_{i,slave}$的均值。表 1 列出了国标规定的玉米中 4 种主要成分近红外模型相关评价指标的范围(上限)。

本文所有算法在 MATLAB 平台完成。

表 1 粮油近红外分析仪性能基本要求中玉米主要成分的近红外模型评价标准[16-18]

Table 1 Near infrared model evaluation standards for the main components of corn based on the basic performance requirements of near infrared analyzers for determining grain and oil contents[16-18]

Components	SEP≤	SR(SRo, SRs)≤	Index and formula
Moisture	0.25%	0.14%	SR, formula(7)
Protein	0.30%	0.15%	SR, formula(7)
Oil	0.4%	0.3%	SRo, formula(8)
Starch	1%	0.15	SRs, formula(9)

2 结果与讨论

2.1 玉米中主要成分的 PLS-NIRs 模型对主机样品的预测误差随 nLVs 的变化

3 台仪器上测定的玉米样品的平均光谱如图 1 所示，由该图可看出 M5 与 MP5、MP6 的原始平均光谱有明显差异，经 SNV + 一阶导数预处理后 3 台仪器上样品的平均光谱差异减小，但在某些波峰、波谷区域仍有肉眼可见的差异，MP6 与 MP5 的平均光谱很相近。故选取 M5 作为主机，MP5、MP6 两台光谱仪为从机。MCS 方法未发现异常样本。根据 SPXY 方法从 M5 测试的 80 个玉米样品中选取前 60 个样品作为校正集，剩余 20 个样品作为内部检验集。

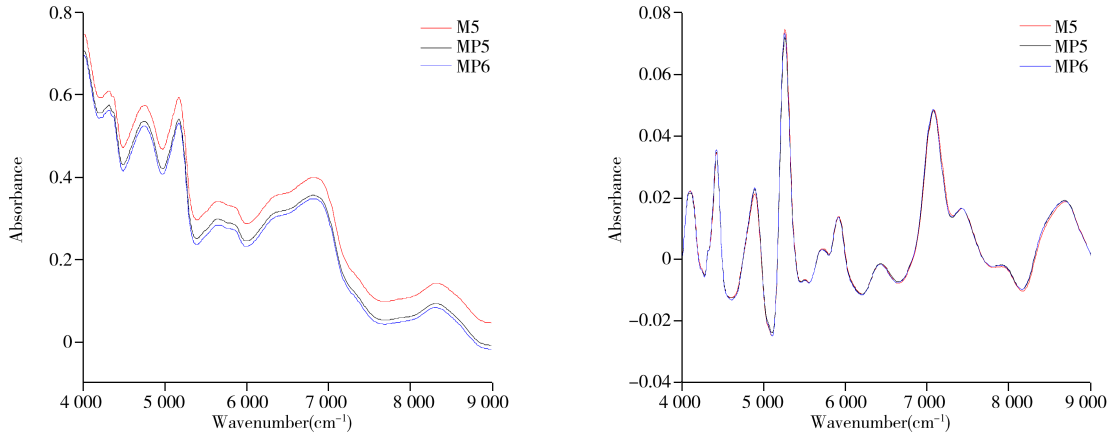

图 1 3 台近红外光谱仪所测玉米样品的平均光谱

Fig. 1 Average spectra of corn samples measured on three NIRs instruments

图 2 为主机 M5 检验集样品各主要成分的平均相对误差随 nLVs 的变化。由该图可知，nLVs = 1 时，各成分的 MRE 已经小于 3%，淀粉的 MRE 在 nLVs = 1 时甚至低于 1%。蛋白质、水分、脂肪含量的 MRE 均呈现在 nLVs < 10 范围逐步降低到一个相对低点后有所升高，nLVs > 10 后又逐步降低的趋势。一般选取预测误差第一次达到相对最小时对应的 nLVs 作为最佳潜变量个数。根据该原则，脂肪和淀粉模型可选 nLVs = 6；蛋白质和水分模型可选 nLVs = 4。

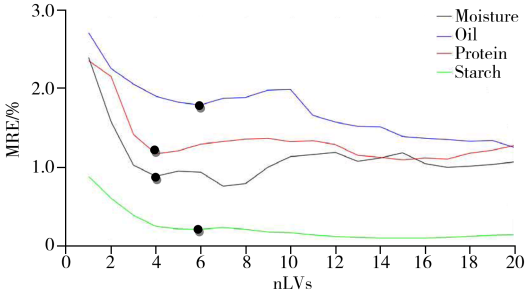

图 2 玉米中 4 种成分含量的 PLS-NIRs 模型对主机检验集样品的平均相对误差(MRE)随 nLVs 的变化

Fig. 2 The average relative error(MRE) of the PLS-NIRs model for the content of the four components in corn of the samples of the host test set varies with nLVs

采用留一交叉验证、四折交叉验证确定的玉米各营养成分的 PLS 模型中 nLVs 一般在 5~10 之间。以水分含量的 PLS-NIRs 模型为例，模型的前 5 个潜变量(LV)对应的方差分别为：0.999 39、0.000 44、0.000 08、0.000 05、0.000 01。第一个潜变量的方差非常之大，占据了所有潜变量方差之和的 99.9% 以上。玉米中另外 3 个成分脂肪、蛋白质及淀粉含量 PLS-NIRs 模型的第一个潜变量对应的累积贡献率也大于 99.9%。因此，如果根据前 nLVs 个潜变量累积贡献率大于 99.9% 选取潜变量个

数，玉米样品近红外光谱模型的 nLVs = 1，该值大大小于常规方法确定的潜变量个数。

2.2　潜变量个数对玉米中主要成分 PLS – NIRs 模型及模型转移结果的影响

表 2 ~ 4 给出了不同潜变量个数下所建立的主机模型对主机样品和从机样品中水分、脂肪和淀粉含量的预测结果，以及经过 PDS 校正后模型对从机样品的预测结果。根据文献建议值及经验，本文选择 PDS 校正方法中转移因子数为 2，转移集数目为 12 个，窗口宽度为 5，容忍度为 0.01[19]。

蛋白质预测结果与表 1 相似，限于篇幅，该结果省略。表 2 ~ 4 中斜体数据表明对应的指标满足表 1 的要求。由这 3 个表可知，不同潜变量个数所建模型中，nLVs = 1 时所建立的 PLS – NIRs 模型直接转移到从机后，对从机样品各成分含量的预测误差 RMSEP 及 MRE 最小，且模型预测从机样品的误差与主机样品预测误差相差不多。模型对主机验证集样品的 SEP 以及从机的再现性评价指标均满足表 1 所列的国标要求。PDS 校正对 nLVs = 1 下所建模型的传递效果的改进很有限，且 PDS 校正后模型对从机样品脂肪、淀粉含量的预测误差高于模型直接传递的预测误差（见表 3、表 4 中 * 标注的数据）。说明模型直接传递误差不大时，没必要采用 PDS 方法进行模型传递。

由留一交叉验证和四折交叉验证选取的 nLVs 均大于 4，在此原则下建立的玉米各营养成分 PLS – NIRs 模型对主机样品的预测误差 RMSEP、MRE 随 nLVs 的增大而不同程度地降低，但各模型对从机样品的 RMSEP 及 MRE 显著增大，是主机样品对应误差指标的几倍到十几倍，其误差水平超出许可范围。经 PDS 校正从机光谱后，模型对从机样品的预测误差降低到与主机相当的水平。nLVs > 1 时建立的玉米营养成分的 PLS – NIRs 模型给出的主、从机预测值的重现性较 nLVs = 1 时所建模型的重现性高一个量级，nLVs > 4 时所建模型对从机样品中各成分含量的预测值大多不满足表 1 所列的重现性指标要求。说明从第二个潜变量开始，仪器间光谱信息的一致性变差，导致 nLVs > 1 时各模型主、从机间近红外测试值的重现性变差。虽然 nLVs 增大可改进模型对主机样品的预测准确度，但会导致模型传递误差变大，使得模型无法直接转移到从机。

表 2　玉米水分 PLS – NIRs 模型直接传递及 PDS 校正后的传递结果

Table 2　Direct transfer results and transfer results after PDS correction of the PLS – NIRs model for predicting moisture content in corn

Calibration/validation spectra	NLVs * *	RMSEC (%)	RMSEP (%)	MRE(%)	SEP(%)	SR(%)
M5/M5	1	0.343 5	0.244 9	1.97	*0.22 < 0.25*	
	2	0.210 5	0.161 9	1.27	*0.16 < 0.25*	
	5	0.124 7	0.111 5	0.90	*0.11 < 0.25*	
	7	0.113 6	0.108 5	0.76	*0.10 < 0.25*	
M5/MP5	1		0.330 0/0.325 3	2.55/2.51		*0.03 < 0.14/0.04 < 0.14*
	2		0.753 2/0.228 1	6.93/1.73		*0.11 < 0.14/0.12 < 0.14*
	5		1.822 2/0.255 7	17.45/2.04		0.23/0.20
	7		2.040 1/0.269 2	19.57/2.16		0.26/0.22
M5/MP6	1		0.326 9/0.325 5	2.51/2.51		*0.03 < 0.14/0.04 < 0.14*
	2		0.751 7/0.238 8	6.90/1.81		*0.14 ≤ 0.14*/0.16
	5		1.866 0/0.2967	17.82/2.39		0.27/0.25
	7		2.150 9/0.318 8	20.59/2.60		0.31/0.28

* * nLVs = 1，7，5 are the number of latent variables of the PLS – NIRs model for corn water determined respectively according to the cumulative contribution rate greater than 99.9%，four – fold cross – validation，and leave – one – out cross – validation. The number above "/" is the result of direct model transfer，and the number below "/" is the model transfer result after PDS correction（nLVs = 1、7、5 是分别根据累计贡献率大于 99.9% 、四折交叉验证、留一交叉验证确定的玉米水分 PLS – NIRs 模型的潜变量个数。"/" 之上数字为模型直接传递结果，"/" 之下数字为 PDS 校正后的模型传递结果）

表 3　玉米脂肪 PLS – NIRs 模型直接传递及 PDS 校正后的传递结果

Table 3　Direct transfer results and transfer results after PDS correction of the PLS – NIRs model for predicting oil content in corn

Calibration/validation spectra	nLVs * *	RMSEC (%)	RMSEP(%)	MRE(%)	SEP(%)	SRo(%)
M5/M5	1	0.149 9	0.135 6	3.15	*0.13 < 0.4*	

（续表3）

Calibration/validation spectra	nLVs**	RMSEC (%)	RMSEP(%)	MRE(%)	SEP(%)	SRo(%)
	2	0.096 2	0.099 4	2.40	*0.10 < 0.4*	
	5	0.059 6	0.096 6	2.08	*0.10 < 0.4*	
	10	0.029 0	0.074 7	1.63	*0.07 < 0.4*	
M5/MP5	1		0.152 3/0.160 9*	3.52/3.70*		*0.05 < 0.3*/*0.01 < 0.3*
	2		0.157 9/0.105 0	3.80/2.47		*0.16 < 0.3*/*0.05 < 0.3*
	5		0.305 0/0.095 9	8.26/2.18		*0.28 < 0.3*/*0.05 < 0.3*
	10		0.848 7/0.083 8	24.10/1.88		0.83/*0.04 < 0.3*
M5/MP6	1		0.152 8/0.161 6*	3.52/3.71*		*0.05 < 0.3*/*0.01 < 0.3*
	2		0.255 9/0.104 1	6.77/2.40		*0.22 < 0.3*/*0.05 < 0.3*
	5		0.443 1/0.094 9	12.37/2.15		0.43/*0.04 < 0.3*
	10		0.877 6/0.086 7	24.93/1.94		0.87/*0.05 < 0.3*

** nLVs = 1, 5 and 10 are the number of latent variables of the PLS – NIRs model for corn fat determined respectively according to the cumulative contribution rate greater than 99.9%, four – fold cross – validation, and leave – one – out cross validation. The number above "/" is the result of direct model transfer, and the number below "/" is the model transfer result after PDS correction(nLVs = 1、5、10 是分别根据累计贡献率大于99.9%、四折交叉验证、留一交叉验证确定的玉米脂肪 PLS – NIRs 模型的潜变量个数，"/"之上数字为模型直接传递结果，"/"之下数字为 PDS 校正后的模型传递结果)

表4 玉米淀粉 PLS – NIRs 模型直接传递及 PDS 校正后的模型传递结果

Table 4 Direct transfer results and transfer results after PDS correction of the PLS – NIRs model for predicting starch content in corn

Calibration/validation spectra	nLVs**	RMSEC (%)	RMSEP (%)	MRE(%)	SEP(%)	SRs
M5/M5	1	0.826 2	0.666 0	0.80	*0.65 < 1*	
	2	0.594 6	0.581 2	0.74	*0.57 < 1*	
	5	0.216 6	0.183 4	0.24	*0.18 < 1*	
	10	0.117 1	0.138 7	0.17	*0.14 < 1*	
M5/MP5	1		0.826 2/0.769 3	0.81/0.95*		*0.08 < 0.15*/0.22
	2		0.680 6/0.594 2	0.85/0.75		0.35/0.39
	5		2.080 2/0.376 2	3.16/0.46		3.27/0.34
	10		1.721 9/0.351 3	2.61/0.44		2.75/0.32
M5/MP6	1		0.847 7/0.773 3	1.10/0.96		0.29/0.21
	2		0.894 7/0.610 3	1.11/0.76		0.72/0.40
	5		1.826 5/0.371 7	2.76/0.46		2.89/0.33
	10		1.784 2/0.359 7	2.70/0.45		2.84/0.31

** nLVs = 1, 5 and 10 are the number of latent variables ofthe PLS – NIRs model for corn starch determined respectively according to the cumulative contribution rate greater than 99.9%, four – fold cross – validation, and leave – one – out cross validation. The number above "/" is the result of direct model transfer, and the number below "/" is the model transfer result after PDS correction(nLVs = 1、5、10 是分别根据累计贡献率大于99.9%、四折交叉验证、留一交叉验证确定的玉米淀粉 PLS – NIRs 模型的潜变量个数，"/"之上数字为模型直接传递结果，"/"之下数字为 PDS 校正后的模型传递结果)

2.3 潜变量个数对烟叶总植物碱 PLS – NIRs 模型及模型转移结果的影响

以烟叶数据集中 Set B 作为建模集，Set A 中主机的78个样品光谱为外部验证集，建立烟叶总植物碱的 PLS – NIRs 模型。通过 MCS 方法发现两个异常点，最终取 Set B 中的1 068个样本建立模型。根据累积贡献率大于99.9%选取的 nLVs = 13，四折和留一交叉验证选取的 nLVs 分别为16和19。表5给出了分别取13、16、19个潜变量时得到的烟叶总植物碱的 PLS – NIRs 模型结果，以及经过 PDS 校正后模型对从机样品的预测结果。表中斜体数据表明对应的指标满足小于6%的企业内控要求。取 nLVs = 13 所建立的烟叶总植物碱 PLS – NIRs 模型直接转移到从机后，对 S1 从机的 MRE 小于6%，但对其他3台从机样品的 MRE 均大于6%；经 PDS 校正后，nLVs = 13 下所建模型对4台从机的预测误差均小于6%。而潜变量个数大于13时所建立的烟叶总植物碱的 PLS – NIRs 模型对主机样品的预测误改进很有限，且模型直接转移到从机后，除 nLVs = 16 模型对 S1 样品的 MRE 小于6%外，对其他从机样品的 MRE 均大于6%，即使经过 PDS 校正也不能保证这些模型对所有从机样品的 MRE 满足企业的内控要求。

2.4　讨论与分析

玉米样品中主要成分的 PLS – NIRs 模型潜变量个数取 1 时，模型传递误差最小且 4 个成分的 PLS – NIRs 模型对主、从机样品预测值的重现性均满足国标要求。由于第一潜变量的方差已经占据所有潜变量方差总和的 99.9% 以上，说明第一潜变量之后的潜变量所包含的有效信息加起来不足 0.1% ，引入这些有效信息很少的潜变量，易导致模型过拟合：即对建模样品或主机样品模型的误差很小(小于潜变量个数为 1 的模型误差)，但对从机样品的误差过大。

表 5　烟叶总植物碱 PLS – NIRs 模型直接传递及 PDS 校正后的传递结果

Table 5　Direct transfer results and transfer results after PDS correction of the PLS – NIRs model for predicting total alkaloid contents in tobacco leaves

Calibration/validation spectra	nLVs**	RMSEC(%)	RMSEP(%)	MRE(%)
M/M	13	0.110 9	0.124 3	*3.52 <6%*
	16	0.098 0	0.119 9	*3.53 <6%*
	19	0.091 2	0.117 6	*3.41 <6%*
M/S1	13		0.188 0/0.178 9	*5.60 <6% /4.87 <6%*
	16		0.205 7/0.251 9	*5.79 <6% /7.29*
	19		0.220 6/0.266 8	6.65/7.94
M/S2	13		0.295 2/0.213 8	10.15/*5.89 <6%*
	16		0.231 4/0.205 7	7.96/*5.78 <6%*
	19		0.269 8/0.220 6	9.86/6.65
M/S3	13		0.256 5/0.161 6	8.77/*4.69 <6%*
	16		0.220 5/0.153 5	7.57/*4.29 <6%*
	19		0.239 3/0.144 0	8.73/*4.04 <6%*
M/S4	13		0.218 8/0.153 9	8.29/*5.22 <6%*
	16		0.262 8/0.205 3	10.37/7.70
	19		0.318 5/0.220 7	13.05/8.57

** nLVs = 13，16，19 are the number of latent variables of the PLS – NIRs model for total alkaloids in tobacco leaves determined respectively according to the cumulative contribution rate greater than 99.9% ，four – fold cross – validation，and leave – one – out cross – validation. The number above "/" is the result of direct model transfer，and the number below "/" is the model transfer result after PDS correction(nLVs = 13、16、19 是分别根据累计贡献率大于 99.9% 、四折交叉验证、留一交叉验证确定的烟叶总植物碱 PLS – NIRs 模型的潜变量个数。"/" 之上数字为模型直接传递结果，"/" 之下数字为 PDS 校正后的模型传递结果) =

图 3 给出了玉米中水分 PLS – NIRs 模型的第一载荷轴及 M5、MP5 样品光谱的差谱绝对值的标准方差光谱(简称 SDDSI1)。由图 3 可看出，第一载荷轴的峰值位于 SDDSI1 很小或较小的区域，而 SDDSI1 的峰值所对应第一载荷取值均在 0 附近，说明第一潜变量中对模型贡献大的波长点有效避开了仪器间光谱差异波动大的区域，因此当玉米 PLS 模型的潜变量个数 nLVs 取 1 时，对从机样品的预测误差与主机相当。其次，该模型摒弃了方差小于 0.1% 、有效信息含量很低的潜变量，大大提高了模型的稳健性，使得模型传递到从机后误差无明显变化。

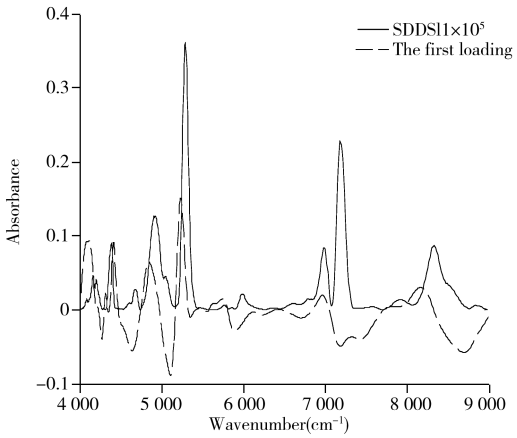

图 3　玉米水分 PLS – NIRs 模型的第一载荷轴与 M5、MP5 差谱绝对值的标准方差谱(SDDSI1)

Fig. 3　The first loading of PLS – NIRs model for predicting corn moisture and the standard deviance spectrum of absolute difference spectra between M5 and MP5

3　结　论

PLS – NIRs 模型中潜变量个数 nLVs 的选取对模型的稳健性、传递性能有重要影响。nLVs 够用即可，过高的 nLVs 容易造成过拟合，影响模型的稳健性，使得模型转移时误差过大。根据累积贡献率大于 99.9% 选取 nLVs 建立的 PLS – NIRs 模型稳健性最好，易于获得好的模型传递结果。而根据留一交叉验证及四折交叉验证或单台(主机)仪器验证集预测误差最小等原则选取的 nLVs 个数均高于根据

累积贡献率大于99.9%选取的nLVs，易导致模型过拟合。

建议根据累计贡献率大于99.9%或接近99.9%时对应的nLVs建立近红外光谱模型，虽然对于主机而言，模型误差比根据留一交叉验证或四折交叉验证选取nLVs建立的模型误差稍高，但模型传递误差小，易于实现模型共享，获得好的模型传递效果。本结论对玉米、烟叶之外的其他类型样品是否成立有待进一步验证。

参考文献：

[1] Chu X L, Shi Y Y, Chen P, Li J Y, Xu Y P. *J. Instrum. Anal.*(褚小立，史云颖，陈瀑，李敬岩，许育鹏. 分析测试学报), **2019**, 38(5): 603-611.

[2] Chu X L. *Practical Handbook of Near Infrared Spectroscopy*. Beijing: Machinery Industry Press(褚小立. 近红外光谱分析技术实用手册. 北京：机械工业出版社), **2016**: 127-128.

[3] Ni L J, Luan S R, Zhang L G. *China J. Chin. Mater. Med.*(倪力军，栾绍荣，张立国. 中国中药杂志), **2016**, 41(19): 3520-3527.

[4] Ni L J, Xiao L X, Zhang L G, Luan S R. *J. Instrum. Anal.*(倪力军，肖丽霞，张立国，栾绍嵘. 分析测试学报), **2018**, 37(5): 539-546.

[5] Ni L J, Han M Y, Zhang L G, Luan S R. *Chin. J. Anal. Chem.*(倪力军，韩明月，张立国，栾绍嵘. 分析化学), **2018**, 46(10): 1660-1668.

[6] Ni L J, Han M Y, Luan S R, Zhang L G. *Spectrochim. Acta A*, **2019**, 206: 350-358.

[7] Zhang J, Cai W S, Shao X G. *Prog. Chem.*(张进，蔡文生，邵学广. 化学进展), **2017**, 29(8): 902-910.

[8] Lin Z X, Wu B L, Wang H, Wang Q W. *J. Instrum. Anal.*(林振兴，邬蓓蕾，王豪，王群威. 分析测试学报), **2008**, 27(12): 1330-1333.

[9] Geladi P, Kowalski B R. *Anal. Chim. Acta*, **1986**, 185: 1-17.

[10] Ni L J, Zhang L G. *Basic Chemometrics and Its Applications*. Shanghai: East China University of Science and Technology Press(倪力军，张立国. 基础化学计量学及其应用. 上海：华东理工大学出版社), **2011**: 237.

[11] YC/T 160-2002. Tobacco and Tobacco Products—Determination of Total Alkaloids—Continuous Flow Method. Tobacco Industry Standards of the People's Republic of China(烟草及烟草制品 总植物碱的测定 连续流动法. 中华人民共和国烟草行业标准).

[12] Han M Y. *Building Robust Near Infrared Spectral Model Based on the Method of Screening Stable and Consistent Wavelengths*. Shanghai: East China University of Science and Technology(韩明月. 基于稳定、一致波长筛选方法建立稳健近红外光谱模型. 上海：华东理工大学), **2019**.

[13] Xiao L X. *Research on Several Problems in Global Robust Near Infrared Quantitative Model Based on "Crowdfunding" Mode*. Shanghai: East China University of Science and Technology(肖丽霞. 基于"众筹"模式的全局稳健近红外定量模型中若干问题的研究. 上海：华东理工大学), **2019**.

[14] Cao D S, Liang Y Z, Xu Q S, Li H D, Chen X. *J. Comput. Chem.*, **2010**, 31: 592-602.

[15] Galvão R K H, Araujo M C U, Jose G E, Pontes M J C, Silva E C, Saldanha T C B. *Talanta*, **2005**, 67: 736-740.

[16] GB/T 24895-2010. Grain and Oil Inspection Near Infrared Analysis Calibration Model Verification and General Rules for Network Management and Maintenance. National Standards of People's Republic of China(粮油检验 近红外分析定标模型验证和网络管理与维护通用规则. 中华人民共和国国家标准).

[17] GB/T 24902-2010. Inspection of Grain and Oil Determination of Crude Fat Content in Corn Near Infrared Method. National Standards of People's Republic of China(粮油检验 玉米粗脂肪含量测定 近红外法. 中华人民共和国国家标准).

[18] GB/T 25219-2010. Grain and Oil Inspection Corn Starch Content Determination Near Infrared Method. National Standards of People's Republic of China(粮油检验 玉米淀粉含量测定 近红外法. 中华人民共和国国家标准).

[19] Li Y Q. *Study on Near Infrared Model Transfer of Tobacco Total Plant Alkaloids*. Shanghai: East China University of Science and Technology(李永琪. 烟叶总植物碱近红外模型传递的研究. 上海：华东理工大学), **2020**.

(责任编辑：盛文彦)