基于多变量检测限的模型变量筛选方法研究*

彭严芳[1]，史新元[1]，李　洋[1]，周璐薇[1]，裴艳玲[1]，华国栋[2]，
吴志生[1**]，乔延江[1**]

（1. 北京中医药大学中药信息工程研究中心　北京　100102；2. 北京中医药大学东方医院　北京　100102）

摘　要：目的：优选清开灵注射液中黄芩苷含量的偏最小二乘模型的变量筛选方法。方法：以近红外透射模式，采集清开灵注射液近红外光谱，分别采用间隔偏最最小二乘算法（iPLS）、向后间隔偏最小二乘算法（BiPLS）、移动窗口间隔偏最小二乘算法（mwPLS）对光谱进行变量筛选，建立清开灵中黄芩苷定量模型，与全谱 PLS 模型进行比较，优选出最佳变量筛选方法；计算不同变量筛选方法下的多变量检测限（MDL），以 MDL 值优选出最佳变量筛选方法，验证经典化学指示参数得出的结果。结果：经不同变量筛选方法处理后所建 PLS 定量模型的预测性能不同，其中 iPLS 算法所建模型预测能力最好，预测集决定系数（R_{pre}^2）和校正均方根误差（SEP）分别为 0.996 5 和 602.3 μg·mL^{-1}；iPLS 变量筛选所得 MDL 值最低（1.19 mg·mL^{-1}），MDL 值计算结果与经典化学指示参数结果一致。结论：本文以 MDL 值作为最佳变量筛选方法评价指标，多变量检测限同时考虑了待测成分校正集样本相应误差，测试集样本相应误差以及未知样品的杠杆率，作为一个综合量化指标，能够用于中药分析体系变量筛选方法的优选。

关键词：清开灵注射液　间隔偏最小二乘　向后间隔偏最小二乘　移动窗口间隔偏最小二乘　多变量检测限

doi: 10.11842/wst.2014.05.003　　中图分类号：O657.3　　文献标识码：A

近红外光谱法（Near Infrared Spectroscopy，NIRS）具有分析速度快、无污染、低消耗、非破坏性，可以实现多组分同时测定等优点[1]，在中药领域的应用也越来越多，主要包括了定性鉴别[2~4]、含量测定[5, 6]、生产过程质量监控[7~10]等。由于近红外光谱由含氢基团伸缩振动的倍频和组合频吸收产生，易受体系中其它含氢基团吸收的影响，且光谱吸收强度弱，谱峰重叠严重，这在一定程度上阻碍了 NIRS 在中药中的应用。在模型建立的过程中，筛选有效变量并剔除不相关和非线性信息能够简化模型并提高模型预测能力[11]。但不同的变量筛选方法有其自身特点，目前的定量模型筛选以经典化学指示参数，如预测均方根误差（RMSEP）、交叉验证均方根误差（RMSECV）和决定系数（R^2）等，利用预测值与参考值之间的误差来考察模型的预测性能，进而优选最佳变量筛选方法，忽略了模型的定量检测性能，即检测限。中药有效成分含量多低于 1%，所建立模型能否准确检测所含物质的含量至关重要。本文以清开灵注射液为载体，采用间隔偏最小二乘法（Interval Partial Least Squares，iPLS）、向后间隔偏最小二乘算法（Backward Interval Partial Least Squares，BiPLS）、移动窗口偏最小二乘算法（Moving Window

收稿日期：2013-08-14
修回日期：2013-08-29

*　科学技术部国家“十一五”重大新药创制专项（2010ZX09502-002）：清开灵注射液安全性关键技术研究，负责人：乔延江；科学技术部国家“十一五”重大新药创制专项（2011ZX09201-201-24）：安宫牛黄丸生产过程优化及在线质量控制研究，负责人：马群；国家自然科学基金委青年基金项目（81303218）：多源信息融合的中药近红外模型适用性评价方法研究，负责人：吴志生；北京中医药大学校级基金项目（2013-X-043）：黄酮类成分的近红外 MDL 与 MQL 规律研究，负责人：吴志生。

**　通讯作者：吴志生，讲师，主要研究方向：中药质量控制；乔延江，本刊编委，教授，主要研究方向：中药质量控制。

Partial Least Squares，mwPLS）等多种变量筛选方法建立黄芩苷定量模型，比较不同变量筛选方法下模型的预测性能。并以基于两类误差理论的近红外光谱技术多变量检测限（Multivariate Detection Limit，MDL）为指标，评价不同变量筛选方法下 NIR 模型的多变量检测限估计值，以指导近红外最佳变量筛选方法的优选。

1 材料与方法

1.1 仪器与试剂

全息光栅型近红外光谱仪（美国 FOSS 公司），1100 型高效液相色谱仪包括四元泵、真空脱气机、自动进样器、柱温箱、二极管阵列检测器（DAD）及 HP 数据处理工作站（美国 Agilent 公司）。色谱级甲醇（美国 Tedia 公司），磷酸（天津大学试剂厂，分析级），纯净水（杭州娃哈哈集团有限公司）。黄芩苷对照品由中国食品药品检定研究院提供（批号：110777-201005），6 批清开灵注射液由指定药厂提供。

1.2 样品的制备

取 6 批清开灵注射液，其黄芩苷含量按 2010 年版《中华人民共和国药典》清开灵注射液项下高效液相色谱法（High Performance Liquid Chromatography，HPLC）进行测定，每批样品均用纯净水稀释成系列浓度，取其中 3 批作为校正集，另 3 批作为验证集。

1.3 光谱采集条件

采用近红外透射模式，光谱分析软件采集光谱，以仪器内部的空气为背景，分辨率为 0.5 nm，扫描范围 400~2 500 nm，扫描次数 32 次，每个样品平行测定 3 次，取平均光谱。

1.4 数据处理

运用 Unscrambler 7.0 软件（挪威 CAMO 软件公司）对光谱数据进行预处理和模型计算。iPLS 法、BiPLS 法、mwPLS 法工具包由 Nørgaard 等人提供的网络共享*，其余各计算程序均自行编写，采用 MATLAB 软件工具（美国 Mathwork 公司）计算。

1.5 多变量检测限原理

根据国际纯粹与应用化学联合会（International Union of Pure and Applied Chemistry，IUPAC）的定义，检测限为某特定方法在给定的置信度内可从样品中检出待测物质的最小浓度或量。而采用基于两类误差理论和误差传递理论得到的逆矩阵模型 MDL 简略计算公式（1），同时考虑了一类误差（假阳性）和二类误差（假阴性），能够全面反映分析方法在一定置信限下能被检出的最低浓度[12~16]。

$$MDL = \Delta p,q\left[(1+h)MSEC-\sigma_c^2\right] \quad (1)$$

其中，$\Delta p,q$ 为在自由度 V 下，非中心 t 分布在置信限为 p,q 时的分位数。h 为未知样本的杠杆值，可通过偏最小二乘（Partial Least Squares，PLS）、主成分分析（Principal Component Analysis，PCA）等模型求得。MSEC 为校正集均方差，σ_c^2 为被分析物参考值测量误差。

2 结果与讨论

2.1 光谱预处理方法优筛

图 1 为清开灵注射液近红外原始光谱。由图 1 可知，光谱在 1 540 nm 处存在一个强的水分吸收峰，主要由 O-H 伸缩振动一级倍频产生。整个原始光谱重叠严重，并伴有随机噪声、基线漂移等干扰因素。需运用合理的光谱预处理方法，消除各种噪声和干扰因素，以提取近红外光谱中待测对象的特征信息。本文比较了一阶导数（1D）、二阶导数（2D）、标准正则变换（SNV）、S.G.平滑及一系列组合预处理方法对模型的影响（液体体系中 1 900~2 500 nm 波段内光谱波动较大，故剔除此处波段）。

经不同预处理方法获得 PLS 模型的决定系数（R^2）和交叉验证均方根误差（RMSECV）值见表 1。由表 1 可知，清开灵注射液光谱 S.G.平滑预处理后

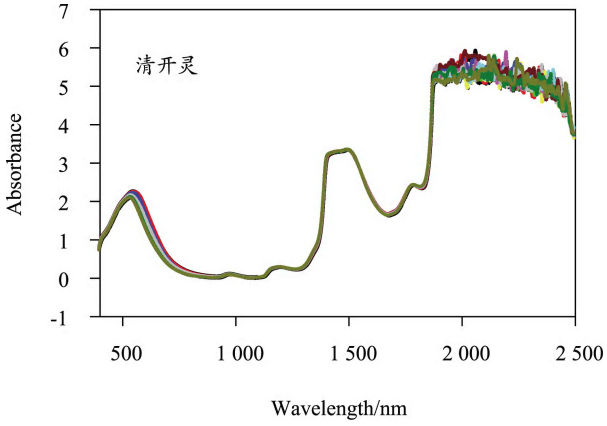

图 1 近红外原始光谱图

* http://www.models.kvl.dk/source/iToolbox/

表 1　清开灵注射液不同光谱预处理方法 PLS 模型预测结果 (μg·mL^{-1})

Pretreatment	Latent factors	Calibration set		Validation set	
		R^2	RMSEC	R^2	RMSECV
Raw	2	0.986 6	1 519	0.985 5	1 600
1D	4	0.983 5	1 689	0.970 9	2 310
2D	3	0.978 4	1 933	0.896 0	4 273
2D+SG	4	0.983 2	1 701	0.968 1	2 355
SG	2	0.986 6	1 519	0.987 1	1 552
SNV	2	0.985 8	1 566	0.985 0	1 632
1D+SG	4	0.983 2	1 701	0.968 1	2 355

的 R^2 最大，RMSECV 值最小，选取该光谱预处理方法建立 PLS 模型。

2.2　PLS 模型的建立

采用 PLS 法，将光谱数据与样品的 HPLC 分析结果相关联建立校正模型。模型预测值与 HPLC 测定值的相关关系见图 2。由图 2 可知，PLS 模型具有良好的预测性能，模型的预测集决定系数 (R_{pre}^2) 和预测均方根误差 (SEP) 分别为 0.987 1 和 1 552 μg·mL^{-1}。

2.3　变量筛选

2.3.1　iPLS 法

iPLS 法的基本思想是将近红外光谱等分为若干个光谱区间，对每个区间建立偏最小二乘模型，并对潜变量因子数进行优化，最终找出最优建模区间[17]。本文比较了不同区间划分方法对模型预测性能的影响，结果见表 2。由表 2 可知，清开灵注射液全谱模型中潜变量因子为 5，分别比较不同区间数下最优间隔数的 RMSECV 值，iPLS 模型确定最佳区间数是 22，最佳间隔数为 19。模型的 R_{pre}^2 和 SEP 分别为 0.996 5 和 602.3 μg·mL^{-1}。与 PLS 模型相比，iPLS 模型性能显著提高，SEP 值由原来 1 552 μg·mL^{-1} 降到 602.3 μg·mL^{-1}，说明采用合理的变量筛选方法能够提高模型的预测性能。

2.3.2　BiPLS 法

采用 BiPLS 法筛选有效变量，整个光谱被分为若干个等宽的子区间，子区间的个数对模型的预测能力有影响。BiPLS 每次排除一个子区间，以剩下的区间建模，计算 RMSECV 值，使剩下的区间具有最小 RMSECV 值的区间就是第一个排除的区间，以此类推，计算直到剩下一个区间为止[11]。通过比较模

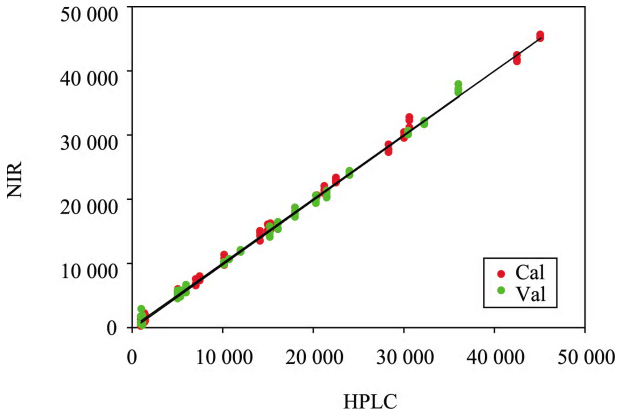

图 2　清开灵注射液近红外预测值与 HPLC 参考值相关关系图

表 2　不同区间数下 iPLS 模型的预测性能 (清开灵/μg·mL^{-1})

Interval numbers	Selected interval	Latent factors	RMSECV
40	34	4	446.8
38	32	3	385.9
36	31	5	518.1
34	29	4	451.5
32	27	4	388.5
30	26	3	466.5
28	24	5	441.1
26	22	4	395.8
24	21	5	419.8
22	19	5	352.9
20	17	5	388.6

表 3 不同区间数下 BiPLS 模型的预测性能(清开灵注射液/μg·mL^{-1})

Number intervals	Selected interval	RMSECV	Number of variables
40	36、34、26、25、14、13、12、8、7、6、5	214.565 8	825
38	33、32、20、14、6、5	221.174 8	553
36	33、31、29、21、3	283.086 6	500
34	29、22、6、5	245.210 7	442
32	29、28、27、21、17、16、13、11、9、8、7、6、5、4	231.249 6	1 407
30	26、25、18、4	238.978 8	500
28	24、18、6、4	221.646 6	536
26	23、22、6、4、3	262.104 6	693
24	21、20、15、10、9、8、7、6、4、3	276.230 1	1 375
22	19、14、13、9、8、6、5、3	239.987 0	1 228
20	17、8	241.223 8	450

型的 RMSECV 大小筛选出最佳子区间的组合。不同区间数下 BiPLS 模型的预测性能见表 3。BiPLS 模型确定最佳区间数为 40，最优的区间组合为 13、6、26、14、7、36、8、12、5、25、34。模型的 R_{pre}^2 和 SEP 分别为 0.774 0 和 3 362.7 μg·mL^{-1}，与全谱 PLS 模型相比预测性能未见提高。

2.3.3 mwPLS 法

mwPLS 法筛选特征波段的基本思想是采用固定的窗口大小在全谱上移动，然后对每个移动窗口截取的波段建立 PLS 模型[11]。本文采用不同窗口大小筛选最佳波段区间，并对其进行了 PLS 建模，结果见表 4。当窗口大小为 19，所建模型的 RMSECV 值最小，因此使用优选波段建立 PLS 模型。模型的 R_{pre}^2 和 SEP 分别为 0.996 1 和 625.99 μg·mL^{-1}，模型性能与全谱模型性能相比显著提高，但与 iPLS 模型性能相比略有下降。

以上结果说明清开灵注射液采用不同的变量筛选方法所建模型具有不同的预测性能，采用 BiPLS 法筛选变量后模型预测性能显著下降，而 iPLS 和 mwPLS 法变量筛选后所建模型预测性能显著提高，iPLS 略好于 mwPLS 法。提示在实际应用中应根据载体的不同进行多种变量筛选方法的考察，通过对筛选结果的比较，获得最优的变量筛选方法。

表 4 最优窗口大小 mwPLS 模型的预测性能
(清开灵中间体银黄液体系/μg·mL^{-1})

Interval numbers	Selected interval	Latent factors	RMSECV
31	2 425~2 551	5	452.0
29	2 425~2 550	5	450.0
27	2 425~2 550	5	448.5
25	2 428~2 550	5	448.1
23	2 430~2 538	5	442.3
21	2 428~2 548	5	428.3
19	2 431~2 535	5	427.1
17	2 433~2 534	5	434.8

2.3.4 不同变量筛选方法的多变量检测限研究

根据两类误差分析理论[12]，采用公式 (1) 计算不同变量筛选方法下的近红外多变量检测限 (式中参考值的采样误差忽略)，结果见表 5。由表 5 可知，对于不同的误差类型，基于模型的近红外多变量检测限估计值不同；同一误差类型下，不同变量筛选方法 MDL 值不同。PLS 和 BiPLS 算法所得 MDL 值较大，而 iPLS 和 mwPLS 算法所得 MDL 值较小。其中采用 iPLS 算法能够获得最低的近红外多变量检

表 5　基于不同$\triangle_{p,q}$黄芩苷含量全谱近红外多变量检测限估计值(清开灵/mg·mL^{-1})

方法	$\triangle_{0.1,0.1}$	$\triangle_{0.1,0.05}$	$\triangle_{0.1,0.01}$	$\triangle_{0.05,0.1}$	$\triangle_{0.05,0.05}$	$\triangle_{0.05,0.01}$	$\triangle_{0.01,0.1}$	$\triangle_{0.01,0.05}$	$\triangle_{0.01,0.01}$
PLS	6.79	7.75	9.55	7.78	8.74	10.56	9.69	10.66	12.49
iPLS	0.92	1.05	1.30	1.05	1.19	1.43	1.31	1.45	1.70
BiPLS	6.34	7.24	8.93	7.27	8.18	9.87	9.06	9.97	11.68
mwPLS	1.05	1.20	1.48	1.20	1.35	1.63	1.50	1.65	1.93

测限估计值，如考虑 I 类误差 5%和 II 类误差 5%，清开灵注射液中黄芩苷含量的近红外多变量检测限估计值 1.19 mg·mL^{-1}。结果与各变量筛选方法下模型的经典指示参数结果一致，表明多变量检测限可以作为近红外定量模型变量筛选方法优选的评价指标。

以上结果可以看出，采用不同的变量筛选方法对同一样本进行变量筛选，存在近红外检测限估计值明显差别，说明采用近红外技术作为分析方法时应该对不同变量筛方法进行对比，以保证待测成分能够被准确检出。而待测成分的近红外检测限是建立在多变量模型的基础上，具有特殊性，需经过严格的实验才能得知。本文采用的多变量检测限分析方法能够用于 NIRS 的研究。

此外，对于近红外模型筛选，本文提出一种基于两类误差分析理论的多变量检测限参数指标，用于筛选最佳模型。模型变量筛选采用的多变量检测限参数与经典化学指示参数 (RMSEP、RMSECV 和 R^2) 相比，更为全面地考虑了待测成分校正集样本相应误差，测试集样本相应误差以及未知样品的杠杆率。因此，采用多变量检测限参数能够更加客观地反映模型的性能。

3　结论

本文以清开灵注射液中间体银黄液为载体，探讨了多种变量筛选方法对研究体系的适用性问题。结果表明，不同变量筛选方法所建模型的近红外多变量检测限差异较大，如考虑 I 类误差 5%和 II 类误差 5%时，多变量检测限变化从 8.74 mg·mL^{-1} 到 1.19 mg·mL^{-1}。对于不同的变量筛选方法，近红外多变量检测限具有特殊性，需经过建立数学模型后对比才能得知；对于本文载体而言，采用近红外多变量检测限参数作为评价指标，可以看出 iPLS 更加适于作为变量筛选方法，这一结果可从经典化学指示参数所得结论得到验证。综上，多变量检测限同时考虑了假阳性和假阴性误差，作为一个综合量化指标，能够用于中药分析体系变量筛选方法的优选。

参考文献

1　Zhao Z, Liang Z. Application and advantage of near infrared spectroscopy technology in authentication of Chinese materia medica. *China J Chin Mater Med*, 2012, 37(8):1062~1065.

2　卜海博,聂黎行,王丹,等.近红外光谱法无损识别林下山参及其生长年限.光谱学与光谱分析,2012,32(7):1801~1805.

3　Wu Z, Tao O, Cheng W, *et al*. Visualizing excipient composition and homogeneity of compound liquorice tablets by near-infrared chemical imaging. *Spectrochim Acta Part A*, 2012, 86:631~636.

4　Lu H, Wang S, Cai R, *et al*. Rapid discrimination and quantification of alkaloids in Corydalis Tuber by near-infrared spectroscopy. *J Pharmaceut Biomed*, 2012, 59(2):44~49.

5　Li W, Cheng Z, Wang Y, *et al*. A study on the use of near-infrared spectroscopy for the rapid quantification of major compounds in *Tanreqing* injection. *Spectrochim Acta A*, 2013, 101(1):1~7.

6　陈雪怡,杜伟锋,蔡宝昌,等.近红外光谱法快速测定浙麦冬中麦冬总皂苷的含量.中药新药与临床药理,2013,24(1):85~88.

7　Wu Z, Tao O, Dai X, *et al*. Monitoring of a pharmaceutical blending process using near infrared chemical imaging. *Vib Spectrosc*, 2012, 63(11):371~379.

8　Wu Z, Xu B, Du M, *et al*. Validation of a NIR quantification method for the determination of chlorogenic acid in *Lonicera japonica* solution in ethanol precipitation process. *J Pharmaceut Biomed*, 2012, 62(3):1~6.

9　杨辉华,郭拓,马晋芳,等.一种近红外光谱在线监测新方法及其在中药柱层析过程中的应用.光谱学与光谱分析,2012,32(5):1247~1250.

10　耿炤,胡浩武,李胜华,等.近红外光谱技术用于川红活血胶囊提取过程的研究.中国药业,2012,21(11):14~15.

11　刘冰,毕开顺,孙立新,等.近红外光谱结合不同偏最小二乘法测

定乳块消片醇沉液中丹参素和橙皮苷含量.世界科学技术-中医药现代化,2009,11(3):388~396.

12 Alcalà M, León J, Ropero J, *et al*. Analysis of low content drug tablets by transmission near infrared spectroscopy: selection of calibration ranges according to multivariate detection and quantitation limits of PLS models. *J Pharmaceut Sci*, 2008, 97(12):5318~5327.

13 Blanco M, Castillo M, Peinado A, *et al*. Determination of low analyte concentrations by near-infrared spectroscopy: Effect of spectral pretreatments and estimation of multivariate detection limits. *Anal Chim Acta*, 2007, 581(2):318~323.

14 Boqué R, Larrechi M S, Rius F X. Multivariate detection limits with fixed probabilities of error. *Chemomet Intell Lab Sys*, 1999, 45(1):397~408.

15 Clayton C A, Hines J W, Elkins P D. Detection limits with specified assurance probabilities. *Anal Chem*, 1987, 59(20):2506~2514.

16 Wu Z, Sui C, Xu B, *et al*. Multivariate detection limits of on-line NIR model for extraction process of chlorogenic acid from *Lonicera japonica*. *J Pharmaceut Biomed*, 2013, 77(4):16~20.

17 Norgaard L, Saudland A, Wagner J, *et al*. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectrosc*, 2000, 54(3):413~419.

Optimization of Near Infrared Variable Selection Method Based on Multivariate Detection Limit

Peng Yanfang[1], Shi Xinyuan[1], Li Yang[1], Zhou Luwei[1], Pei Yanling[1], Hua Guodong[2], Wu Zhisheng[1], Qiao Yanjiang[1]

(1. Research Center of Traditional Chinese Medicine Information Engineering, Beijing University of Chinese Medicine, Beijing 100102, China;

2. Dongfang Hospital, Beijing University of Chinese Medicine, Beijing 100102, China)

Abstract: This study was aimed to optimize the near infrared (NIR) variable selection method based on multivariate detection limit (MDL). Using *Qing-Kai-Ling* (QKL) injection as object, three variable selection methods (interval partial least-squares, iPLS; backward interval partial least squares, BiPLS; moving window interval partial least squares, mwPLS) were used to establish the PLS models of baicalin in QKL injection, respectively. The prediction ability of different variable selection method was compared. MDL of all models were calculated in contrast to the MDL value of full spectra PLS model, to select optimal variable selection method. The results showed that different variable selection methods had different prediction ability. Among them, iPLS had the best performance which determination coefficient of prediction (R_{pre}^2) and the root mean square errors of prediction (SEP) were 0.996 5 and 602.3 $\mu g \cdot mL^{-1}$, respectively. All MDLs of different variable selection methods were reduced compared with the full spectra PLS model. The value of iPLS was the lowest comes to be 1.19 $\mu g \cdot mL^{-1}$. The results above indicated that the best variable selection method for baicalin in QKL injection was iPLS. MDL theory took the error of calibration and validation set and the leverage of external sample into account, which can comprehensively evaluate model detection performance compared to the classic chemical indicator parameters. This method was particularly suitable for the variable selection method optimization of NIR quantitative model of low concentration sample such as Chinese herbal medicine.

Keywords: *Qing-Kai-Ling* injection, iPLS, BiPLS, mwPLS, multivariate detection limit

(责任编辑:李沙沙 张志华,责任译审:王 晶)