# 红参提取物总皂苷近红外定量分析建模中的变量筛选

安思宇[1,2]，张 磊[1]，尚献召[1]，岳洪水[1]，柳文媛[2*]，鞠爱春[1*]

1. 天津天士力之骄药业有公司，天津市中药注射剂安全性评价企业重点实验室，天津 300402
2. 中国药科大学药物质量与安全预警教育部重点实验室，江苏 南京 210009

**摘 要** 注射用益气复脉(冻干)是由红参、麦冬、五味子3种药材制成的新型冻干粉针制剂。红参提取物总皂苷是注射用益气复脉(冻干)生产过程的重要质控指标，传统分析方法分析结果具有滞后性，无法快速反馈生产过程质量信息。近红外光谱(NIR)作为一种快速无损的过程监控工具已经广泛应用于中药质量控制领域。中药成分复杂且近红外光谱吸收强度弱、谱区重叠严重，如何从干扰严重的复杂光谱中提取有效信息是提高测量准确度的关键。模型集群分析(MPA)通过随机采样，最大限度地提取了样本信息，打破了传统一次性建模思路，为变量筛选方法提供了新的思想。采集了55批红参提取物近红外光谱数据，运用多元散射校正(MSC)进行光谱数据预处理，并将MPA衍生的随机蛙跳法(RF)、竞争自适应重加权(CARS)、变量组合集群分析法(VCPA)、VCPA联合迭代保留信息变量(IRIV)方法与OPUS软件自带的变量筛选方法分别用于总皂苷含量偏最小二乘(PLS)定量分析模型的建立中。结果表明，OPUS软件、CARS-PLS与RF-PLS所建模型校正集相关系数($R_c$)仅为0.601 3，0.565 3与0.644 0，拟合效果不理想。VCPA-PLS法所建模型的$R_c$为0.951 2，是几种变量筛选方法中最高的，但是其预测性能不佳，模型稳健性不理想。VCPA-IRIV-PLS模型具有最好的预测效果，$R_c$为0.928，RSEP%为7.99%。

**关键词** 近红外光谱；注射用益气复脉(冻干)；红参提取物总皂苷；偏最小二乘法；变量筛选；变量组合集群分析；迭代保留信息变量

**中图分类号**：R91 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)01-0206-04

## 引 言

近红外光谱技术(near infrared spectroscopy，NIR)具有分析速度快、分析效率高、分析成本低、可偶联光纤进行远距离操作、操作技术要求低等优势，已经成为过程分析技术的重要组成部分。通过建立中药生产过程关键质量指标的NIR定量分析模型，可以实现活性成分的快速无损测定，控制中药产品生产过程质量稳定性。注射用益气复脉(冻干)是基于传统中药古方生脉散发展起来的一种新型冻干粉针制剂，由红参、麦冬、五味子3种药材组成，具有益气复脉，养阴生津的功效。总皂苷含量是注射用益气复脉(冻干)红参提取过程的质量指标，因此需要建立该指标的监测方法控制注射用益气复脉(冻干)生产过程质量。

NIR存在吸收强度弱且谱区信息重叠严重等瓶颈问题[1-2]，进行光谱预处理虽然可以消除一些影响因素带来的干扰信息，但是有效的波长仍然掩盖在整条光谱中，因此选择适当的变量筛选方法是提升模型精度的重要手段。目前主要的变量筛选方法有：连续投影方法(successive projections algorithm，SPA)[3]、间隔偏最小二乘法(interval partial least squares，IPLS)[4]、移动窗口偏最小二乘法(moving window partial least squares，MWPLS)[5]、无信息变量消除法(uninformative variable elimination，UVE)[6]和遗传算法(genetic algorithm，GA)[7]以及OPUS、TQ Analyst等光谱分析软件自带的变量筛选方法，它们都是一次性变量筛选方法且没有考虑到变量之间的交互作用。

模型集群分析(model population analysis，MPA)的思想打破了传统的一次性建模思路[8]，随机蛙跳法(random frog，RF)[9-10]、竞争自适应重加权(competitive adaptive reweighted sampling，CARS)[11-12]、变量组合集群分析法(variable combination population analysis，VCPA)[13-14]以及迭代保留信息变量(iteratively retaining informative variables，

**收稿日期**：2019-12-06，**修订日期**：2020-04-19

**基金项目**：国家自然科学基金项目(81573557)，天津市科技计划项目(18YFCZZC00430)资助

**作者简介**：安思宇，女，1996年生，中国药科大学药物质量与安全预警教育部重点实验室硕士研究生 e-mail：15295772410@163.com

*通讯作者 e-mail：juach@tasly.com；liuwenyuan@cpu.edu.cn

IRIV)[15-16]等都是在 MPA 思想下衍生出的变量筛选方法。其中 RF[9-10]、CARS[11-12]已经广泛用于近红外光谱信息变量筛选中以提高模型性能，而对于 VCPA 与 IRIV 的应用研究较少，目前未有在中药生产过程质量检测的应用实例。本研究采用 VCPA 迭代 IRIV[17]变量筛选算法建立红参总皂苷偏最小二乘(partial least squares，PLS)定量分析模型，对预测集总皂苷含量进行预测，并与其他变量筛选条件下建立的模型性能进行比较。

## 1　实验部分

### 1.1　仪器与试药

MATRIX-F 型傅里叶变换近红外光谱仪(德国 Bruker)，配有 OPUS 数据处理软件(版本 7.5)；SHIMADZU UV-2600 紫外-可见分光光度计(日本 Shimadzu)；MS204TS 型电子分析天平(瑞士 Mettler Toledo)；MATLAB 数学软件(美国 MathWorks)；红参提取过程终产物(天津天士力之骄药业有限公司提供，共 55 批，批号 A1—A55)；人参皂苷 Re 对照品(中国药品生物制品检定研究院)。

### 1.2　紫外-可见分光光度法测定红参提取物中总皂苷含量

#### 1.2.1　供试品溶液的制备

精密称定 0.1 g 红参提取终产物，用 10 mL 0.5 mol·$L^{-1}$氢氧化钠溶液溶解，上预先处理好的 AB-8 树脂柱，分别用 0.5 mol·$L^{-1}$的氢氧化钠的 20%甲醇溶液与 20%甲醇溶液洗脱杂质，最终用甲醇洗脱待测物，过滤膜，取续滤液作为供试品溶液。

#### 1.2.2　对照品溶液制备

精密称取人参皂苷 Re 对照品，加甲醇制成每 1 mL 含 2 mg 的溶液，摇匀，即得。

#### 1.2.3　测定条件

精密量取对照品溶液 20，40，60，80 和 100 μL，及供试品溶液 40 μL，分别置于 10 mL 具塞试管中。置水浴中挥尽溶剂后取出，放冷，精密加新配制含 5%香草醛的冰醋酸溶液和高氯酸混和液(2∶8) 1 mL，摇匀。置 60 ℃水浴中加热 15 min，取出，立即置冰浴中冷却 2 min。精密加冰醋酸 5 mL，摇匀，在室温下放置 5 min。以相应试剂为空白，在 550 nm 处测定吸收度，计算，即得。

### 1.3　近红外光谱采集

称取 2.5 g 红参提取终产物，使用纯化水定容至 25 mL，混匀，使提取物充分溶解。将所得溶液转移至离心管，使用德国 Bruker 公司 MATRIX-F 型近红外光谱仪采集近红外光谱。以内部空气作为参比，光谱采集模式为透射，采集方式为在线探头采集。NIR 采集参数为：光程为 2 mm，分辨率为 2 $cm^{-1}$，光谱扫描范围 4 000～12 000 $cm^{-1}$，扫描 32 次。收集样本的原始光谱如图 1。

### 1.4　校正集与验证集的选择

采用 1.2 项中的紫外可见分光光度法测定样本中的总皂苷值，利用联合 x-y 距离的样本集划分(sample set partitioning based on joint x-y distance，SPXY)方法将 55 批红参样本分为 40 个校正样本与 15 批外部检验样本。

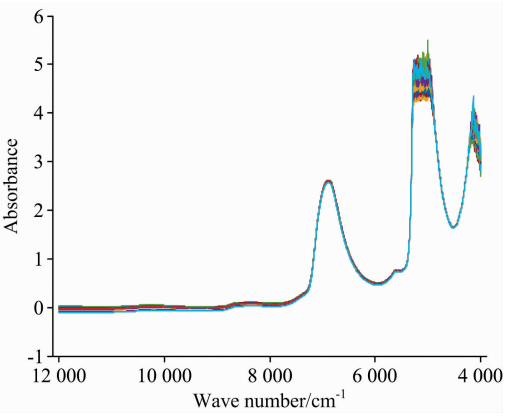


**图 1　红参提取物原始近红外光谱**

**Fig. 1　Original near infrared (NIR) spectra of red ginseng**

**表 1　红参提取物校正集和验证集的划分结果**

**Table 1　Statistical characteristics of total saponins value**

| 样本类型 | 样本数 | 最小值/(mg·$mL^{-1}$) | 最大值/(mg·$mL^{-1}$) | 均值/(mg·$mL^{-1}$) | 标准差 |
|---|---|---|---|---|---|
| 校正集 | 40 | 8.08 | 14.72 | 10.587 | 1.34 |
| 检验集 | 15 | 9.52 | 12.93 | 10.685 | 0.938 |

### 1.5　光谱预处理

常见的光谱预处理方法有均值中心化、多元散射校正、卷积平滑法和小波变换等。本研究的预处理方法均为多元散射校正(multiplicative scatter correction，MSC)，用来修正各样品近红外光谱间的相对基线平移和偏移现象。

### 1.6　基于不同方法的特征变量提取过程

为了比较不同变量筛选方法对红参提取物总皂苷近红外定量模型的影响，将经过 MSC 预处理的红参总皂苷光谱数据分别利用 VCPA-IRIV，VCPA，CARS，RF 以及 OPUS 软件自带的变量筛选方法进行变量筛选。

其中 VCPA-IRIV 运用二进制矩阵采样法(BMS)从红参提取物近红外光谱变量中采样 1 000 次，得到 1 000 组不同的变量组合，运用 PLS 方法分别对这 1 000 组变量组合进行光谱建模，计算交叉验证均方根误差(cross validation root mean square error，RMSECV)最小的前 15%的变量组合中各光谱变量出现的次数，以及 RMSECV 最大的前 5%的变量组合中光谱变量出现的次数，两者相减即为相应光谱变量的贡献值。运用指数递减(EDF)函数迭代运行 40 次，删除贡献小的变量，最终剩下 100 个变量。迭代结束后每个变量被选择的频率如图 2。接着联用 IRIV 方法，BMS 采样 200 次，逐个波长变量计算包含和不包含该变量时的 RMSECV 平均值，得到两者之差 DMEAN(difference of mean values)和非参数检验方法曼-惠特尼 U 检验的 $P$ 值，按表 2 所示变量筛选规则去除无信息与干扰信息变量，保留强信息与弱信息波长变量，经多次迭代循环直至无信息和干扰信息变量全部消除，最终筛选出 18 个变量。接下来为了体现 VCPA 与 IRIV 联用的优越性，单独使用 VCPA 方法进行变量筛选。

CARS 模仿进化论中的“适者生存”法则，采用蒙特卡洛采样法抽取 80%的样本为校正集，建立 PLS 回归模型，计算每个变量回归系数，回归系数绝对值越大则贡献值越大，利用 EDF 函数去除贡献值小的波长点，此过程迭代 500 次，得到 500 组不同的变量子集，最后建立每个子集的 PLS 回归模型，其中 RMSECV 最小的变量子集即为最优子集。

RF 是一种类似可逆跳转的马尔科夫链蒙特卡洛算法，通过在模型空间模拟一条正态分布的马尔可夫链来计算每个变量被选择的概率，进而实现变量选择，迭代 1 000 次后，被选择概率前 10 的波数变量为最后的特征变量。以上 4 种变量筛选算法均在 MATLAB 软件中使用。

光谱采集用的 MATRIX-F 型傅里叶变换近红外光谱仪，附带数据处理分析软件 OPUS，该软件自带的变量筛选方法操作简便、应用广泛，考察该软件所建 PLS 模型的性能，可以论证 MPA 下衍生的变量筛选方法是否有推广应用的价值。

**表 2 变量筛选规则**

**Table 2 Variable classification rules**

| 波数变量类型 | 分类规则 |
|---|---|
| 强信息变量 | DEMEAN<0，$p$<0.05 |
| 弱信息变量 | DEMEAN<0，$p$>0.05 |
| 无信息变量 | DEMEAN>0，$p$>0.05 |
| 干扰信息变量 | DEMEAN>0，$p$<0.05 |

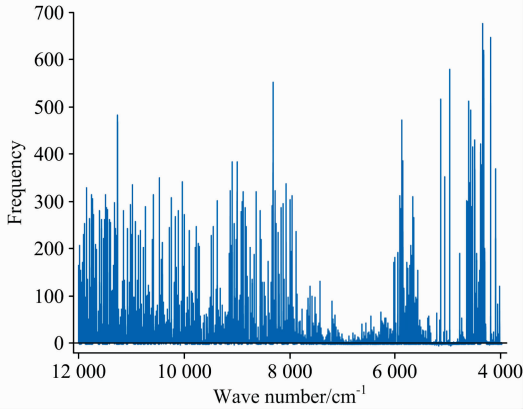


**图 2 VCPA-IRIV 变量筛选过程中变量被选择的次数**

**Fig. 2 The frequency of variables selected by VCPA-IRIV**

## 2 结果与讨论

用不同方法筛选出的变量建立 PLS 定量分析模型，并对建模结果进行统计，结果如表 3 所示。

**表 3 不同变量筛选方法的模型结果**

**Table 3 Comparison on modeling results by different variables screening methods**

| 建模方法 | 筛选变量数 | 主成分数 | 校正集 | | 验证集 | |
|---|---|---|---|---|---|---|
| | | | $R_c$ | RMSEC | RMSEP | RSEP/% |
| FULL-PLS | 2 074 | 1 | 0.560 1 | 1.110 5 | 1.231 | 11.48 |
| OPUS-PLS | 93 | 3 | 0.601 3 | 1.086 4 | 1.018 | 9.49 |
| CARS-PLS | 51 | 2 | 0.565 3 | 1.105 7 | 1.130 3 | 10.54 |
| RF-PLS | 10 | 3 | 0.644 | 1.025 5 | 1.095 0 | 10.21 |
| VCPA-PLS | 12 | 10 | 0.951 2 | 0.412 3 | 1.067 2 | 9.95 |
| VCPA-IRIV-PLS | 18 | 10 | 0.928 | 0.494 1 | 0.856 7 | 7.99 |

结果表明 OPUS 软件自带变量筛选方法以及 CARS、RF 虽然减少了冗余信息，但是其建模效果并不理想，$R_c$ 分别为 0.601 3，0.565 3 与 0.644 0，与全波长光谱建模效果相当，并没有成功筛选出有效变量。这是因为 OPUS 软件采用一次性采样方法建模，且只能筛选成段的变量，CARS 利用蒙特卡洛采样法在样本空间进行重复取样，RF 利用蒙特卡洛采样法在变量空间进行重复取样，却都没有考虑到变量之间的组合效应，不适用于提取红参提取物中的有效变量。

VCPA-PLS 模型的 $R_c$ 为 0.951 2，是所有方法中最高的，但是其 RSEP(%)为 9.95%，预测效果没有 VCPA-IRIV-PLS 好，且其 RMSEC 与 RMSEP 之间差值最大，过拟合现象最严重。这表明虽然 VCPA 方法可以通过 BMS 采样得到 1 000 组不同的变量组合，很好地考虑了变量组合效应，但是当变量数远大于样本数时，VCPA 无法实现所有变量的组合，容易受到无关信息与干扰信息的影响，导致模型的过度拟合，而 IRIV 可以更好地去除无关信息与干扰信息，两者联用时可以很好地弥补自身的缺点，提高模型预测效果。因此 VCPA-IRIV 更适用于红参提取物总皂苷的模型建立，可以用来检测本公司注射用益气复脉(冻干)生产过程，红参提取过程终产物的总皂苷值。

VCPA-IRIV 在考虑变量组合效应的同时，很好地排除无关信息与干扰信息变量，不仅解决了红参提取物中的变量筛选问题，也为变量之间存在相关关系的近红外定量分析模型的拟合带来思路，该方法的推广有利于近红外光谱技术在中药制剂生产过程的质量控制中的应用。将该方法应用于其他分析对象时，可通过调整相关控制系数，以达到最佳的模型拟合效果，除了 IRIV，还可以在 VCPA 的基础上叠加使用其他变量选择方法，以提升相关质量指标分析模型的预测性能。

## References

[1] Li W, Prasad S, Fowler J E. IEEE Transactions on Geoscience & Remote Sensing, 2012, 50(4): 1185.

[2] HUAN Ke-wei, LIU Xiao-xi, ZHENG Feng(宦克为，刘小溪，郑 峰). Transactions of the Chinese Society of Agricultural Engineering (农业工程学报), 2013, 29(4): 266.

[3] Duan F, Fu X, Jiang J J, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2018, 143: 12.

[4] Nespeca M G, Pavini W D, de Oliveira J E. Vibrational Spectroscopy, 2019, 102: 97.

[5] Chen H Z, Tang G Q, Song Q Q, et al. Analytical Letters, 2013, 46(13): 2060.
[6] Tan C, Wu T, Xu Z H, et al. Vibrational Spectroscopy, 2012, 58: 44.
[7] Theocharis John B, Tsakiridis Nikolaos L, Tziolas Nikolaos V, et al. European Journal of Soil Science, 2019, 70(3): 578.
[8] YUN Yong-huan, DENG Bai-chuan, LIANG Yi-ceng(云永欢, 邓百川, 梁逸曾). Chinese Journal of Analytical Chemistry(分析化学), 2015, 43(11): 1638.
[9] Yao X Q, Yang W, Li M Z, et al. IFAC-Papers OnLine, 2018, 51(17): 660.
[10] CHEN Li-dan, ZHAO Yan-ru(陈立旦, 赵艳茹). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2014, 30(8): 168.
[11] Jiang H, Xu W, Chen Q. Spectrochimica Acta-Part A: Molecular and Biomolecular Spectroscopy, 2019, 214: 366.
[12] Wang W, Jiang H, Liu G H, et al. Chinese Journal of Analytical Chemistry, 2017, 45(8): 1137.
[13] Yun Y H, Wang W T, Deng B C, et al. Analytica Chimica Acta, 2015, 862: 14.
[14] Zhao H, Huan K W, Shi X G. Chinese Journal of Analytical Chemistry, 2018, 46(1): 136.
[15] Yun Y H, Wang W T, Tan M L, et al. Analytica Chimica Acta, 2014, 807: 36.
[16] YU Lei, ZHANG Tao, ZHU Ya-xing, et al(于　雷, 章　涛, 朱亚星, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2018, 34(16): 148.
[17] Yun Y H, Bin J, Liu D L, et al. Analytica Chimica Acta, 2019, 1058: 58.

# Variable Selection Method in the NIR Quantitative Analysis Model of Total Saponins in Red Ginseng Extract

AN Si-yu[1,2], ZHANG Lei[1], SHANG Xian-zhao[1], YUE Hong-shui[1], LIU Wen-yuan[2*], JU Ai-chun[1*]

1. Tianjin Tasly Pride Pharmaceutical Co., Ltd., Tianjin Key Laboratory of Safety Evaluation Enterprise of TCM Injections, Tianjin　300402, China
2. Key Laboratory of Drug Quality Control and Pharmacovigilance, Ministry of Education, China Pharmaceutical University, Nanjing　210009, China

**Abstract**　Yiqi Fumai Lyophilized Injectionis a new type of freeze-dried powder injection made of red ginseng, ophiopogon japonicus and schisandrachinensis. The total saponin content of red ginseng extract is an important quality control index in the production process of Yiqi Fumai lyophilized injection. The results of traditional analysis methods lag far behind, which cannot feedback the quality information of the production process timely, it is necessary to establish a rapid method for the determination of total saponin. As a process monitoring tool, near-infrared spectroscopy (NIR) has been widely used in the quality control of traditional Chinese medicine. How to extract the effective information from the spectrum with weak absorption and serious overlapping of spectral regions is the key to improve the monitoring veracity. Model population analysis (MPA) provides a new idea for variable selection method. In this study, the near-infrared spectrum data of 55 batches red ginseng extracts were collected. Themulti scattering correction (MSC) method was used to preprocess the spectrum data, the variable screening methods derived from MPA, such as random frog (RF), competitive adaptive reweighted sampling (CARS), variable combination population analysis (VCPA), VCPA combined with IRIV (iterative retaining information variables) and the variable selection method of OPUS were respectively used in the establishment of PLS quantitative analysis model. The results showed that the $R_c$ of model established by OPUS, CARS-PLS and RF-PLS were only 0.601 3, 0.565 3 and 0.644 0, respectively. The $R_c$ of model established by VCPA-PLS was 0.951 2, which was the highest, but this model did not present good robustness. The model established by VCPA-IRIV-PLS had the best prediction effects; its $R_c$ was 0.928, RSEP% was 7.99%.

**Keywords**　Near-infrared spectroscopy; YiqiFumaiLyophilized Injection; Total saponins of red ginseng extract; Partial least squares; Variable selection; Variable combination cluster analysis; Iterative retention information variables

(Received Dec. 6, 2019; accepted Apr. 19, 2020)

* Corresponding authors