DOI: 10.19650/j.cnki.cjsi.J1905728

# 一种基于频率与回归系数相结合的自举柔性收缩变量选择方法*

张 峰,汤晓君,仝昂鑫,王 斌,王经纬
(西安交通大学 电力设备电气绝缘国家重点实验室 西安 710049)

**摘 要**:针对傅里叶变换红外光谱仪获取的谱线数量庞大,直接选用全部谱线进行多元线性回归易导致过拟合、稳定性差、分析周期长等问题,提出了一种基于频率与回归系数相结合的自举柔性收缩变量选择方法。该算法以变量的权重作为选择的依据,在每次迭代过程中,根据变量的回归系数与频率计算变量的权重,通过加权自举采样技术实现对变量的柔性收缩。应用玉米红外光谱集对该方法进行了验证,在玉米油数据集中,其预测均方根误差(RMSEP)与相关系数($R_p$)分别为 0.020 2 和 0.976 5,变量数目由原始的 700 个减少到 13 个;在玉米蛋白质数据集中,*RMSEP* 与 $R_p$ 分别为 0.027 9 和 0.996 8,变量数目由原始的 700 个减少到 16 个。结果表明,提出的变量选择算法选择的变量少而精,具有实际的应用价值。

**关键词**: 波长选择;加权自举采样;近红外光谱;偏最小二乘
**中图分类号**: TH741 **文献标识码**: A **国家标准学科分类代码**: 150.25

# Bootstrapping soft shrinkage variable selection method based on the combination of frequency and regression coefficient

Zhang Feng, Tang Xiaojun, Tong Angxin, Wang Bin, Wang Jingwei
(*State Key Laboratory of Electrical Insulation & Power Equipment*, *Xi'an Jiaotong University*, *Xi'an 710049*, *China*)

**Abstract**: Aiming at the problems that the spectral lines obtained using Fourier transform infrared spectrometer are enormous, and directly using all the spectral lines to perform multiple linear regression easily leads to over-fitting, poor stability and long analysis period. In this paper, a bootstrap soft shrinkage variable selection method based on the combination of frequency and regression coefficient is proposed. This method selects the variables based on the weight of the variables; in each iterative process, the new weight of the variable is calculated according to the regression coefficient and frequency of the variable, and the soft shrinkage of the variables is realized through weighted bootstrap sampling technology. The method was verified using the infrared spectrum datasets of corn. On the corn oil dataset, the root mean square error of prediction (RMSEP) and correlation coefficients ($R_p$) are 0.020 2 and 0.976 5, respectively, the number of variables is reduced from the original 700 to 13. On the corn protein dataset, the *RMSEP* and $R_p$ are 0.027 9 and 0.996 8, respectively, the number of variables is reduced from the original 700 to 16. The result shows that the proposed variable selection algorithm can select fewer and more precise variables, and has practical application value.

**Keywords**: wavelength selection; weighted bootstrap sampling; near infrared spectroscopy; partial least square

## 0 引 言

傅里叶变换红外光谱仪因分析速度快,灵敏度高等特点,被广泛应用于环境保护、石油天然气勘探、煤矿灾害预警等领域[1-4]。通常,红外光谱仪获取的谱线有成百上千个,这些谱线之间存在严重的共线问题,并且不可避免地会包含干扰谱线(未知组分吸收谱线)以及无用信息谱线(无吸收区域谱线)。如果将全部谱线变量用来建模分析,不仅会增加模型的复杂程度,甚至还会降低模型的预测性能[5-8]。因此,在建立分析模型之前对谱线变量进行提取具有重要的意义。

收稿日期:2019-10-15 Received Date:2019-10-15
*基金项目:国家重点研发计划(2016YFF0102805)项目资助

为了提高模型的预测能力与效率,国内外学者们提出了一系列的变量选择方法,这些方法可以归为区间选择和个体选择两类。其中区间选择方法主要包括间隔偏最小二乘法(interval partial least square, iPLS)[9]、移动窗口偏最小二乘法(moving window partial least-squares, MWPLS)[10]、可变窗口移动偏最小二乘法(changeable size moving window partial least squares, CSMWPLS)[11],向后间隔偏最小二乘法(backward interval partial least square, BiPLS)[12-13]、组合区间偏最小二乘法(synergy interval partial least squares, SiPLS)[14]和区间组合优化算法(interval combination optimization, ICO)[15]等。基于波长区间的选择算法没能考虑到区间内谱线变量之间的共线性问题,通常选择的变量集中在几个区间内,这些区间的变量还可以做进一步的提取。个体变量选择方法具有代表性的包括蒙特卡洛无信息变量消除法(Monte Carlo non-information variable elimination, MCUVE)[16]、竞争性自适应重加权采样法(competitive adaptive reweighted sampling, CARS)[17]、遗传算法(genetic algorithm, GA)[18]、连续投影算法(successive projections algorithm, SPA)[19]和自举柔性收缩算法(bootstrapping soft shrinkage, BOSS)[20]等。其中 MCUVE 是根据变量稳定性来选择有效变量的,当稳定性大于事先设置的阈值时,认为该变量为有用的信息变量。该方法选择的变量数目通常会较多,且阈值的设置范围对结果影响很大。CARS 算法在进行变量选择时,采用指数递减函数来强制删除变量回归系数绝对值较低的变量。然而,当采用不同样本建立模型时,回归系数可能发生变化,这导致了算法的稳定性较差。应用 GA 算法来进行变量选择时,当变量个数大于 200 个时,GA 算法存在很高的过拟合风险,容易产生局部最优解,导致计算效率低、预测性能差[21]。SPA 算法可以最大程度的降低变量之间的共线性问题,然而有效变量之间的投影距离并不一定最大,筛选出来的变量中会包含无用变量甚至是干扰变量。此外,由于优化过程中每个变量都要作为起点,进行连续投影排序,得到一组变量子集,往往导致该算法的计算量偏大。BOSS 是一种基于加权二进制采样的变量选择方法,通过加权自举采样使得变量随机组合在一起,产生大量的变量子模型,计算变量在预测效果较好部分中的回归系数,根据回归系数绝对值来更新权重值,对于回归系数较大的变量在下次迭代过程中该变量有更大机会被选中。但是 BOSS 算法只考虑了回归系数这个特征,而忽略了频率这个重要特征,选择的变量未必是最优的。

针对上述波长选择方法存在的问题,本文提出了一种频率与回归系数相结合的柔性收缩变量选择方法(frequency and regression coefficient, FRC)。该方法继承了 BOSS 方法中加权自举采样的优点,并以变量的回归系数与频率作为评价指标,实现对变量的优选。将 FRC 与 MCUVE、CARS、BOSS 4 种方法应用于玉米数据集中,建立玉米中油与蛋白质含量的 PLS 预测模型。结果表明,本文提出的 FRC 算法选择的变量最少,预测效果优于其余 3 种方法,是一种有效的变量选择方法。

## 1　光谱变量选择原理分析

### 1.1　加权自举采样

加权自举采样(weighted bootstrap sampling, WBS)是一种基于统计学的有放回的随机采样方法,在每次采样过程中,从全部变量空间中根据变量权重选择指定数目的变量,在实际运行中,变量的权重向量会进行单位化处理,这样保证了每个变量的权重在 0 与 1 之间,每个变量被选中的概率为:

$$p_i = \left(1 - \frac{w_i}{\sum_1^n w_i}\right)^R \tag{1}$$

式中:$n$ 为变量的个数;$w_i$ 为第 $i$ 个变量的权重;$R$ 为上一代迭代保留的变量数目。从式(1)中可以看出,假如某个变量的权重很大,该变量在采样过程中有更大概率被选中。还可以推算出,每次迭代后,获取到的变量个数约为 $R$ 的 0.632 倍。这样保证了每次迭代过程中剩余的变量都会逐次减少,实现对变量空间的柔性收缩。

### 1.2　变量权重更新方法

假设红外光谱矩阵为 $\boldsymbol{X}_{n\times p}$,其中 $n$ 为样本数,$p$ 为变量的个数;$\boldsymbol{y}_{n\times 1}$ 为所对应的浓度信息向量。根据比尔定律,可以得出光谱矩阵与浓度向量的关系可以表示为:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2}$$

式中:$\boldsymbol{e}$ 为随机误差向量;$\boldsymbol{\beta}$ 为回归系数向量。采用加权自举采样方法产生 $N$ 个变量组合空间,对这些变量空间分别建立偏最小二乘(partial least square, PLS)模型,这样每次迭代过程中可以获得回归系数矩阵 $\boldsymbol{\beta}_{N\times P}$ 与 $N$ 个模型的交互验证均方根误差(root mean squared error of cross validation, RMSECV)值,从 $N$ 个模型中选择 $N\delta$($\delta$ 可以设置为 0.1)个最优模型,将 $N\delta$ 的值记作 $k$。此时,从 $k$ 个模型中可以统计出每个变量出现的频率,对其进行单位化处理得到每个变量的频率向量 $\boldsymbol{f}_{1\times p}$。同时,将选择的 $k$ 个模型的回归系数矩阵 $\boldsymbol{\beta}_{k\times P}$ 取绝对值,并进行求和运算,最后对求和后的回归系数向量进行单位化处理,得到 $\boldsymbol{\beta}_{1\times P}$,该向量可由式(3)计算。

$$\boldsymbol{\beta}_{1\times p} = \frac{\mathrm{sum}(\mathrm{abs}(\boldsymbol{\beta}_{k\times p}))}{\mathrm{norm}(\mathrm{sum}(\mathrm{abs}(\boldsymbol{\beta}_{k\times p})))} \tag{3}$$

式中:sum(·)表示求和运算;norm(·)为求模运算;

abs(·)是绝对值运算。在获得了变量的频率系数 $\boldsymbol{f}$ 与回归系数 $\boldsymbol{\beta}$ 后,变量的权重 $\boldsymbol{w}$ 可以由式(4)进行更新。

$$\boldsymbol{w} = \alpha \times \boldsymbol{f} + (1 - \alpha) \times \boldsymbol{\beta} \tag{4}$$

式中:$\alpha$ 为频率与回归系数的融合系数,取值范围为 0~1,当取值为 0 时,权重 $\boldsymbol{w}$ 由回归系数决定,对融合后的权重向量再进行单位化处理。通过上述变换后,每个变量的权重融合了频率与回归系数两个重要指标。

## 2 数据来源与实验方案

### 2.1 实验数据集

研究所用的近红外光谱数据集来源于网址 http://www.eigenvector.com/data/SWRI/index.html。该数据集常被用来检验新方法的性能。数据集是由同一批 80 个玉米样本分别在编号为 M5、MP5、MP6 3 台红外光谱仪上采样获得的。数据集中给出了 4 种物质的含量,分别为水分、油、蛋白质和淀粉。3 台仪器的波长扫描范围均为 1 100~2 498 nm,波长点采样间隔为 2 nm。本文选用 M5 光谱仪扫描的玉米光谱数据进行建模,油与蛋白质含量作为评价指标。在进行建模之前,应用联合 $x$-$y$ 距离样本划分(sample set partitioning based on joint $x-y$ distances,SPXY)方法[22]将 80 个玉米样本分为训练集(60 个样本,用来建立模型)与测试集(20 个样本,用来检验模型)。

### 2.2 FRC 变量选择方法

FRC 算法以变量的回归系数与频率作为评价指标,当两者融合后的权重越大时,表明在下次迭代中该变量被选择的可能性越大。当剩余的变量为 1 时,迭代停止。具体的实现步骤如下:

1)根据变量的权重 $w$,利用加权自举采样方法生成 $N$ 个变量空间子集。需要指出的是,每个变量的初始权重值相等,这样保证了在迭代开始时每个变量都有相同的机会被选择;

2)应用 PLS 算法计算 $N$ 个变量子集的 RMSECV;

3)从 $N$ 个子集中选择 $N\delta$ 个最优模型,记录每次迭代过程中获取到的交互验证均方根误差的最小值(minimum value of root mean squared error of cross validation, minRMSECV);

4)利用 1.2 节中提出的权重融合算法,计算每次迭代过程中变量的权重 $w$;

5)求取每次迭代过程中剩余变量的个数 $p1$,当 $p1>1$ 时,返回步骤 1),进入下一次迭代,否则,执行步骤 6);

6)选择最小的 minRMSECV 对应的变量组合作为最终选择的变量。

### 2.3 FRC 参数确定

FRC 算法中需要确定的参数有:1)迭代过程中加权自举采样方法生成子模型的个数 $N$;2)选择的最优模型占全部子模型的比例 $\delta$;3)频率与回归系数的融合系数 $\alpha$。可以分两步来进行参数确定。

首先,固定融合系数 $\alpha$ 的值,设置为 0.5,来确定子模型的个数 $N$ 与最优模型比例 $\delta$。在不影响运算效率的情况下,$N$ 的值越大越好,因为 $N$ 值越大,表示随机生成的子模型数量越多,每个变量都有机会分配到子模型中,这样在统计学中表示更合理。但是随着 $N$ 值的增加,计算时间几乎与 $N$ 值的大小成正比。$\delta$ 的值应尽量小,这样更有利于获得 $N$ 个 PLS 模型中预测效果好的子集。以玉米中蛋白质数据集为例,设置 $N$ 的初始值为 500,以 100 为间隔从 500 到 2 500 取值,$\delta$ 的初始值为 0.05,以 0.05 为间隔从 0.05 到 0.5 取值。这样经过 200 次 FRC 计算后,可以获得到 210 个模型的 RMSECV 值,选择 RMSECV 最小的值所对应的 $N$ 与 $\delta$ 作为最终选择的参数值。图 1 所示为 RMSECV 的值随 $N$ 与 $\delta$ 变化趋势。根据图 1 可知,当 $N=2\ 300$,$\delta=0.05$ 时,获得的 RMSEV 值最小。

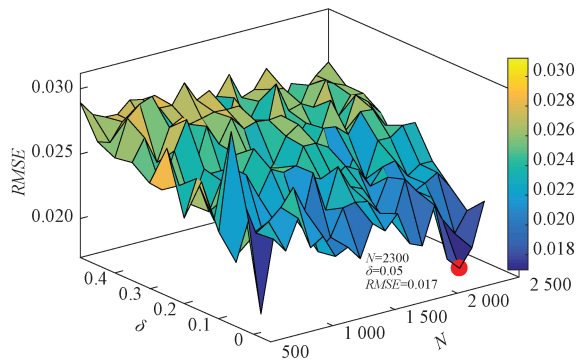



图 1 FRC 算法中参数 $N$ 与 $\delta$ 的优化选择

Fig.1 The optimization and selection of the parameters $N$ and $\delta$ in the FRC algorithm

然后,将 $N$、$\delta$ 两个变量作为已知量,进行融合系数 $\alpha$ 的确定。同样地,采用遍历搜索法,设置 $\alpha$ 的取值范围为 0.1~0.9,间隔为 0.1,$\alpha$ 每次取值时,FRC 算法将重复运行 30 次,求取 30 次运算获得 RMSECV 的平均值,共经历 270 次 FRC 运算后,可以获得 9 组 $\alpha$ 不同取值时对应的平均交互验证均方根误差(mean root mean squared error of cross validation, mRMSEV),选择 mRMSEV 最小值时所对应的 $\alpha$ 值作为最终的融合系数。图 2 所示为融合系数 $\alpha$ 与 RMSECV 关系。由图 2 可知,当 $\alpha=0.4$ 时,对应的平均 RMSECV 值最小。因此,最终确定的 $N$、$\delta$ 与 $\alpha$ 的值分别为 2 300、0.05、0.4。

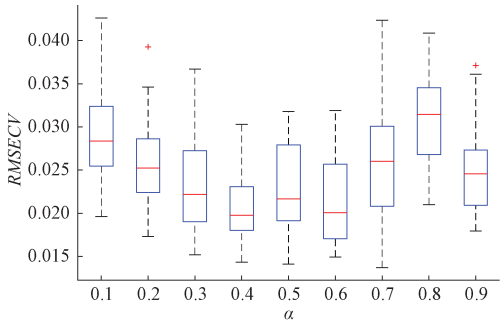
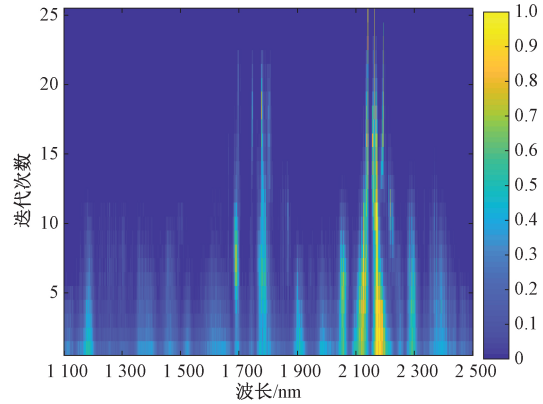



图 2　FRC 算法中参数 α 的优化选择

Fig.2　The optimization and selection of the parameter α in the FRC algorithm




(c) 变量的权重与迭代次数关系

(c) The relationship between the weight of variable and number of iterations

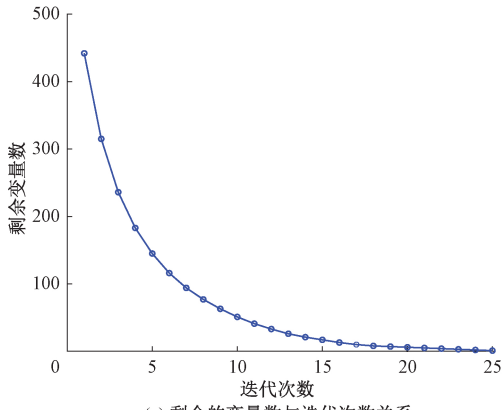
图 3　FRC 算法变量选择过程

Fig.3　The variable selection process of the FRC algorithm

## 3　光谱变量选择方法应用实例

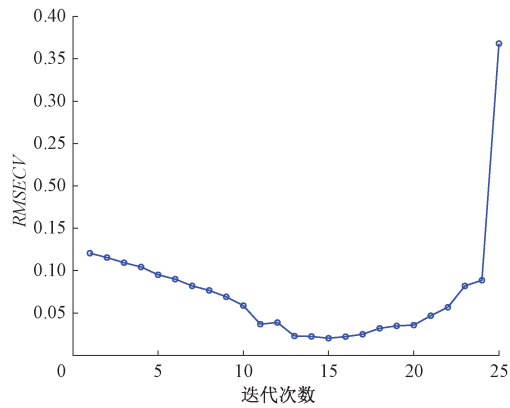
### 3.1　FRC 变量选择过程

根据 2.3 节所确定的参数,以玉米中蛋白质数据集为例,采用 FRC 算法对玉米中 700 个谱线进行筛选,筛选过程如图 3 所示。




(a) 剩余的变量数与迭代次数关系

(a) The relationship between the number of remaining variables and number of iterations




(b) 交互验证均方根误差与迭代次数关系

(b) The relationship between RMSECV and number of iterations

图 3(a)所示为变量选择过程中剩余变量的个数的变化趋势,可以看出,随着迭代次数的逐步增加,剩余的变量个数逐渐减小,并且减小趋势是逐步降低的,体现出了 FRC 算法筛选变量时粗选与精选两个过程,经过 25 次迭代后,剩余的变量为 1。

图 3(b)所示为变量筛选过程中 RMSECV 值的变化趋势,可以看出,RMSECV 的值呈先减小后增大的规律,这是因为在前 15 次迭代过程中剔除了干扰变量或者无用变量,在 15 次迭代后,剔除的变量包含重要的信息变量,RMSECV 值在迭代 15 次时最低,最终选择第 15 次迭代所对应的变量子集。图 3(c)所示为变量的权重随着迭代次数的变化趋势。FRC 算法首先将每个变量的权值设置相同,随后在每次迭代过程中进行权重值更新,将权重值进行单位化缩放处理。可以看出,随着迭代的进行,一些变量的权重值逐渐减小,直至为 0,这部分变量将被逐渐剔除,另外一些变量,尤其是在 1 800 与 2 200 nm 附近的变量,始终保持较大的权重,这些变量最终将被选择。

### 3.2　变量选择方法预测结果比较

为了验证提出波长选择方法的性能,将 FRC 与 MCUVE、CARS、BOSS 4 种方法来进行变量筛选,对筛选后的变量分别建立 PLS 模型来预测玉米数据集中油与蛋白质的含量。4 种方法均重复运行 50 次,对 50 次结果取平均值,预测结果如图 4,表 1、2 所示。

从图 4 和表 1、2 中可以看出,4 种变量提取方法无论是在校正集还是在测试集上,预测性能均比直接采用 PLS 方法好,进一步体现出变量选择的必要性。与 PLS 方法相比,FRC 算法的预测性能得到显著地提高,对于玉米油数据集,测试集的可决系数 $R_p$ 从 0.719 5 提高到 0.976 5;对于蛋白质数据集,$R_p$ 从 0.943 7 提高到 0.996 8,与其他3 种变量选择方法对比,FRC 算法预测效

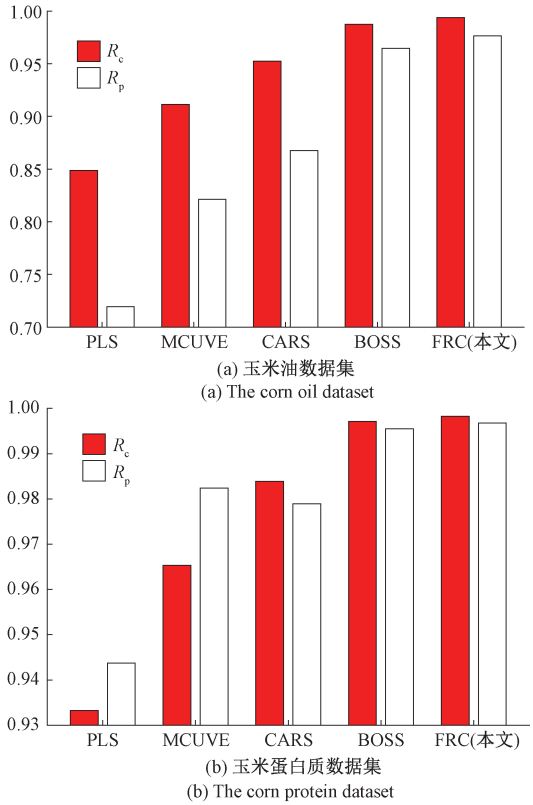



图 4 5 种模型的可决系数 $R_c$ 与 $R_p$ 在玉米油与蛋白质数据集上的值

Fig.4 The coefficients of determination $R_c$ and $R_p$ of five models on the corn oil and protein datasets

**表 1 5 种模型在玉米油数据集的预测结果**

**Table 1 The prediction results of five models on the corn oil dataset**

| 模型 | 变量数 | 校正集 | | 预测集 | |
|---|---|---|---|---|---|
| | | $R_c$ | *RMSECV* | $R_p$ | *RMSEP* |
| PLS | 700 | 0.848 7 | 0.072 1 | 0.719 5 | 0.070 2 |
| MCUVE-PLS | 88.3 | 0.911 3 | 0.054 9 | 0.821 4 | 0.055 8 |
| CARS-PLS | 20.1 | 0.952 3 | 0.039 3 | 0.867 7 | 0.046 3 |
| BOSS-PLS | 15.2 | 0.987 5 | 0.020 7 | 0.964 5 | 0.024 7 |
| FRC-PLS(本文) | 13.4 | 0.994 0 | 0.014 4 | 0.976 5 | 0.020 2 |

**表 2 5 种模型在玉米蛋白质数据集的预测结果**

**Table 2 The prediction results of five models on the corn protein dataset**

| 模型 | 变量数 | 校正集 | | 预测集 | |
|---|---|---|---|---|---|
| | | $R_c$ | *RMSECV* | $R_p$ | *RMSEP* |
| PLS | 700 | 0.933 2 | 0.127 9 | 0.943 7 | 0.117 8 |
| MCUVE-PLS | 110 | 0.965 3 | 0.090 6 | 0.982 4 | 0.064 9 |
| CARS-PLS | 21.9 | 0.983 9 | 0.062 2 | 0.978 9 | 0.071 0 |
| BOSS-PLS | 18.1 | 0.997 1 | 0.026 5 | 0.995 5 | 0.032 8 |
| FRC-PLS(本文) | 16.6 | 0.998 2 | 0.020 8 | 0.996 8 | 0.027 9 |

果最好。在玉米油数据集中,FRC 获得的 $R_c$ 与 $R_p$ 值分别为 0.994 0、0.976 5,预测均方根 *RMSEP*=0.020 2。而其余 3 种方法 $R_c$ 与 $R_p$ 值分别为 0.911 3、0.952 3、0.987 5与 0.821 4、0.867 7、0.964 5,*RMSEP* 分别为 0.055 8、0.046 3、0.024 7;在玉米蛋白质数据集中,FRC 获得的 $R_c$ 与 $R_p$ 值分别为 0.998 2、0.996 8,预测均方根 *RMSEP*=0.027 9,而其余 3 种方法 $R_c$ 与 $R_p$ 值分别为 0.965 3、0.983 9、0.997 1 与 0.982 4、0.978 9、0.995 5,*RMSEP* 分别为 0.064 9、0.071 0、0.032 8。

从表 1、2 中还可以看出,无论是在油还是蛋白质数据集上,MCUVE 选择的变量是最多的,而 FRC 选择的变量是最少的。进一步表明了利用 FRC 方法建立的模型效率更高、预测性能更好。

4 种方法所选择的变量如图 5 所示。在玉米油数据集中,BOSS 与 FRC 方法均选择了 1 700 与 2 300 nm 附近的变量,这部分区域对应了 C-H 键的伸缩振动吸收区域。此外,CARS 还选择了 2 400 nm 附近的变量,然而这些变量由于模型的预测效果差被认为是干扰变量或无用变量,MCUVE 没有选择 1 700 nm 附近的变量,而这些变量由于预测效果好被证明是有用信息变量;对于蛋白质数据集,4 种方法均选择了 1 760 与 2 180 nm 附近的变量,这些区域对应为 C-H 基团与 N-H 基团的吸收谱带。除此之外,MCUVE 还选择了 1 620 与 2 000 nm 附近的变量,FRC 与 BOSS 选择的变量相对集中,CARS 选择的变量相对分散。

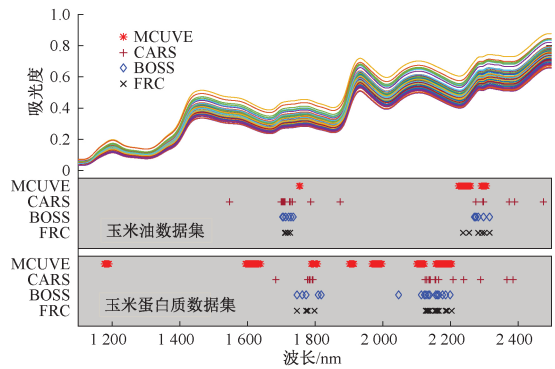



图 5 4 种方法选择的变量分布

Fig.5 The distribution diagram of the variables selected with four methods

## 4 结 论

傅里叶变换红外光谱仪获得的谱线变量多,这些变量无可避免的会含有一些干扰变量与无用变量,需要对这些变量进行筛选。本文分析了近年来常用的波长选择方法,提出了 FRC 变量选择方法,利用玉米近红外光谱

数据对该方法进行了验证,并将该方法与 MCUVE、CARS、BOSS 波长选择方法进行了对比,结果表明本文提出的波长变量选择方法具有更好的预测性能,在波长选择数量上最少,是一种有效的谱线变量选择方法。

## 参考文献


[1] 王智宏, 陈琛, 千承辉, 等.基于粒子群寻优的光谱仪波长误差修正方法[J].仪器仪表学报, 2017, 38(10): 2430-2436.
WANG ZH H, CHEN CH, QIAN CH H, et al. Spectrometer wavelength error correction method based on particle swarm optimization [J]. Chinese Journal of Scientific Instrument, 2017, 38(10): 2430-2436.

[2] TANG X J, LI Y J, ZHU L J, et al. On-line multi-component alkane mixture quantitative analysis using Fourier transform infrared spectrometer [J]. Chemometrics and Intelligent Laboratory Systems, 2015, 146:371-377.

[3] LIU L Y, ZHENG F, ZHANG G Y, et al. Development of solar spectroradiometer for meteorological observation[J]. Instrumentation, 2017, 4(1):1-8.

[4] 韩建, 李雪昭, 曹志敏, 等.原油含水率红外光谱测量的超稀疏表示方法.[J].仪器仪表学报, 2019, 40(6): 78-85.
HAN J, LI X ZH, CAO ZH M, et al. Ultra-sparse representation method for measuring crude oil water content using infrared spectroscopy technique [J]. Chinese Journal of Scientific Instrument, 2019, 40(6): 78-85.

[5] 宦克为, 刘小溪,郑峰,等.基于蒙特卡罗特征投影法的小麦蛋白质近红外光谱测量变量选择[J].农业工程学报, 2013, 29(4): 266-271.
HUAN K W, LIU X X, ZHENG F, et al. Selection of variables for wheat protein near infrared spectroscopy based on Monte Carlo characteristic projection [J]. Journal of Agricultural Engineering, 2013, 29(4): 266-271.

[6] YUN Y H, WANG W T, DENG B C, et al. Using variable combination population analysis for variable selection in multivariate calibration [J]. Analytica Chimica Acta, 2015, 862: 14-23.

[7] CHEN J, YANG C, ZHU H, et al. A novel variable selection method based on stability and variable permutation for multivariate calibration[J]. Chemometrics and Intelligent Laboratory Systems, 2018,182:188-201.

[8] 洪明坚, 温志渝,张小洪.基于稀疏优化的近红外光谱波长选择方法[J].仪器仪表学报, 2011, 32(5): 1114-1118.
HONG M J, WEN ZH Y, ZHANG X H. New wavelength selection algorithm based on sparse optimization [J]. Chinese Journal of Scientific Instrument, 2011, 32(5): 1114-1118.

[9] 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 [J]. Applied Spectroscopy, 2000, 54(3): 413-419.

[10] JIANG J H, BERRY R J, SIESLER H W, et al. Wavelength interval selection in multi-component spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data [J]. Analytical Chemistry, 2002, 74(14): 3555-3565.

[11] DU Y P, LIANG Y Z, JIANG J H, et al. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares [J]. Analytica Chimica Acta, 2004, 501(2): 183-191.

[12] LEARDI R, LARS N. Sequential application of backward interval PLS and genetic algorithms for the selection of relevant spectral regions [J]. Journal of Chemometrics, 2004, 18(11): 486-497.

[13] 屠振华, 冯霖, 孙丽娟, 等. 近红外光谱测定蜂蜜中水分含量特征波长选择和分析研究[J]. 仪器仪表学报, 2011, 32(增刊6): 276-281.
TU ZH H, FENG L, SUN L J, et al. Analysis and study of NIR characteristic wavelengths for honey water content[J]. Chinese Journal of Scientific Instrument, 2011, 32(Suppl.6): 276-281.

[14] 蒋薇薇, 鲁昌华, 张玉钧,等.基于 SiPLS 和 SPA 波长选择的玉米组分测量研究[J].电子测量与仪器学报, 2017, 31(12): 1960-1966.
JIANG W W, LU CH H, ZHANG Y J, et al. Research on maize component measurement of wavelength selection based on SiPLS and SPA [J]. Journal of Electronic Measurement and Instrument, 2017, 31(12): 1960-1966.

[15] SONG X Z, HUANG Y, YAN H, et al. A novel algorithm for spectral interval combination optimization[J]. Analytica Chimica Acta, 2016, 948: 19-29.

[16] HAN Q J, WU H L, CAI C B, et al. An ensemble of Monte Carlo uninformative variable elimination for wavelength selection [J]. Analytica Chimica Acta, 2008, 612: 121-125.

[17] LI H, LIANG Y, XU Q, et al. Key wavelengths

screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta, 2009, 648(1): 77-84.

[18] LEARDI R, AMPARO L G. Genetic algorithms applied to feature selection in PLS regression: how and when to use them[J]. Chemometrics and Intelligent Laboratory Systems, 1998, 41(2): 195-207.

[19] 成忠, 张立庆, 刘赫扬, 等.连续投影算法及其在小麦近红外光谱波长选择中的应用[J].光谱学与光谱分析, 2010, 30(4): 949-952.
CHEN ZH, ZHANG L Q, LIU H Y, et al. Successive projections algorithm and its application to selecting the wheat near-infrared spectral variables[J]. Spectroscopy and Spectral Analysis, 2010, 30(4): 949-952.

[20] DENG B C, YUN Y H, CAO D S, et al. A bootstrapping soft shrinkage approach for variable selection in chemical modeling[J]. Analytica Chimica Acta, 2016, 908: 63-74.

[21] YUN Y H, LI H D, DENG B C, et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra[J]. Trends in Analytical Chemistry, 2019, 113: 102-115.

[22] GALVOR K H, ARAUJO M C U, JOS G E, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4): 736-74.

## 作者简介

**张峰**,2012 年和 2015 年于西安工程大学分别获得学士学位和硕士学位,现为西安交通大学博士研究生,主要研究方向为光谱分析及智能传感器。
E-mail:774149296@qq.com

**Zhang Feng** received his B. Sc. degree and M. Sc. degree both from Xi′an Polytechnic University in 2012 and 2015, respectively. Now, he is a Ph. D. candidate in Xi′an Jiaotong University. His main research interest includes spectrum analysis and intelligent sensor.

**汤晓君**(通信作者),1998 年和 2001 年于西安理工大学分别获得学士学位和硕士学位,2004 年于西安交通大学获得博士学位,现为西安交通大学教授、博士生导师,主要研究方向为智能感器、智能控制、光谱分析、矿井录井及气测录井。
E-mail:xiaojun_tang@mail.xjtu.edu.cn

**Tang Xiaojun** (Corresponding author) received his B. Sc. degree iand M. Sc. degree both from Xi′an University of Technology in 1998 and 2001, respectively, received his Ph. D. degree in 2004 from Xi′an Jiaotong University. Now, he is a professor and doctoral supervisor at Xi′an Jiaotong University. His main research interests include intelligent sensor, intelligent control, spectrum analysis, mine logging and gas logging.