一种基于变量稳定性和可信度的紫外-可见特征波长选择方法

孙 涛，阳春华，朱红求*，李勇刚，陈俊名

中南大学自动化学院，湖南 长沙 410083

摘 要 针对多组分金属离子混合溶液的紫外-可见吸收光谱(UV-Vis)重叠严重、难以分离的问题，提出了一种基于稳定性和可信度偏最小二乘法(SCPLS)的特征波长选择方法。在SCPLS中，引入指数衰减函数(EDF)以迭代的方式对波长变量进行选择。在每次迭代中对蒙特卡罗采样所得到的数据集建模，计算各波长变量的稳定性和可信度指标，并通过EDF选择具有较高稳定性和可信度的变量，选择的变量作为新的变量集进入下一次变量选择迭代。迭代全部完成后，计算每一次迭代所选的变量集建模的交叉验证均方根误差(RMSECV)，选择RMSECV最小的变量集作为波长变量选择的结果。利用Zn(Ⅱ)，Cu(Ⅱ)和Co(Ⅱ)混合溶液的紫外-可见光谱数据集和Zn(Ⅱ)和Co(Ⅱ)混合溶液的紫外-可见光谱数据集对所提方法性能进行了验证，并与全波段偏最小二乘、移动窗口偏最小二乘法(MWPLS)、蒙特卡罗无信息变量消除方法(MC-UVE)、竞争性自适应加权算法(CARS)和稳定性竞争自适应加权算法(SCARS)进行了比较分析。结果表明：该方法不仅能降低波长选择的复杂度，还能在保证波长选择过程稳定的情况下，选出对模型重要的波长变量，较之其他方法所提出的方法选取的变量建立的模型RMSECV最小，对于Zn(Ⅱ)，Cu(Ⅱ)和Co(Ⅱ)数据集，使用SCPLS方法得到的Zn(Ⅱ)，Cu(Ⅱ)和Co(Ⅱ)的RMSECV值分别比全光谱PLS下降60.5%，40.2%和31.8%，与SCARS相比分别下降29.8%，26.1%和0.8%，Zn(Ⅱ)，Cu(Ⅱ)和Co(Ⅱ)平均相对误差分别为2.14%，1.25%和0.74%，其中Zn(Ⅱ)的最大相对误差为4.67%，Cu(Ⅱ)的最大相对误差为3.99%，Co(Ⅱ)的最大相对误差为3.12%；对于Zn(Ⅱ)和Co(Ⅱ)数据集，使用SCPLS方法得到的Zn(Ⅱ)和Co(Ⅱ)的RMSECV值分别比全光谱PLS下降39.4%和24.9%，与SCARS相比分别下降35.3%和13.3%，Zn(Ⅱ)和Co(Ⅱ)平均相对误差分别为1.23%，1.10%，其中Zn(Ⅱ)的最大相对误差为4.45%，Co(Ⅱ)的最大相对误差为4.57%，有效提高光谱建模精度。

关键词 波长选择；稳定性；可信度；紫外-可见光谱

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)11-3438-08

引 言

紫外可见分光光度法(UV-Vis)[1]是一种广泛应用于金属离子浓度检测的方法，具有操作简单、准确度高、重现性好、测量方便等特点。该方法通常联合多元分析方法对全波段数据进行建模，分离计算多金属离子的浓度，实现复杂混合样品中多金属离子浓度同时检测[2]。然而，当各金属离子吸收光谱互相干扰和重叠严重时，传统的全波段多元分析建模方法存在很大的误差及大量冗余信息，导致模型精度低且实时性差。因此，如何选择有效的特征波长变量参与建模具有重要意义。

目前，大量国内外学者对波长选择方法进行了相关研究，提出了更可靠、更精确的模型。其中一些方法基于模型性能的统计量来评价变量，如区间偏最小二乘法(IPLS)[3]和移动窗口偏最小二乘法(MWPLS)[4]。其他的方法是根据变量的统计特性，如相关系数和信噪比等，这种方法包括无信息变量消除(UVE)[5]，蒙特卡罗无信息变量消除方法(MC-UVE)[6]和竞争性自适应加权算法(CARS)[7]等。蒙特卡罗无信息变量消除方法(MC-UVE)将蒙特卡罗采样应用于UVE，以降低过拟合的风险，从而获得更好的结果。竞争性自适应加权算法(CARS)以回归系数的绝对值大小作为衡量指标对光谱数据进行变量筛选。基于CARS的稳定性竞争自适应加权算法(SCARS)[8]以变量的稳定性作为衡量指标，延

收稿日期：2018-09-19，**修订日期**：2019-01-25

基金项目：国家自然科学基金重点项目(61533021)，中南大学中央高校基本科研业务费专项资金项目(2018zzts556)资助

作者简介：孙 涛，1994年生，中南大学自动化学院硕士研究生 e-mail：ssunttao@csu.edu.cn

*通讯联系人 e-mail：hqcsu@csu.edu.cn

续了 CARS 方法的变量选择流程。但在光谱重叠严重的情况下，前一类方法是对变量区间进行选择的，并未针对性地选择特征变量，选择过程中通常会出现特征波长变量多选或漏选的情况[9]；后一类方法单独对每一个波长进行抽样[10]选择，但抽样过程随机性大，导致变量指标计算不准确，影响特征波长变量的选择结果。

为了克服上述波长选择方法的不足，提出了一种新的波长选择方法，即基于稳定性和可信度偏最小二乘法(stability and credibility partial least squares，SCPLS)。首先根据稳定性选取贡献较大的变量，然后应用可信度指标从高稳定性变量中选择更可信的变量(对模型性能影响较大的变量)。SCPLS 应用 EDF 以迭代的方式筛选变量，以避免信息不丰富的变量产生误导性结果。通过交叉验证[11]评价用子集建立模型的性能。以最小的 RMSECV 值的变量子集被认为是最佳变量子集。为了测试 SCPLS 的性能，将该方法应用于两个 UV-Vis 数据集，即来自 Zn(Ⅱ)，Cu(Ⅱ) 和 Co(Ⅱ)混合溶液的数据集和 Zn(Ⅱ)和 Co(Ⅱ)混合溶液的数据集。与 MWPLS，MCUVE，CARS 和 SCAR 方法相比，用 SCPLS 方法选取的变量建立的模型达到了最小 RMSECV。

1　实验部分

矩阵 $\boldsymbol{X}_{n\times p}$ 为所测样本的光谱吸光度矩阵，n 为混合溶液样本数，p 为波长变量数；矩阵 $\boldsymbol{y}_{n\times m}$ 表示浓度矩阵，n 为混合溶液样本数，m 为组分数。在建模过程中，$\boldsymbol{X}_{n\times p}$ 和 $\boldsymbol{y}_{n\times m}$ 都是以均值为中心，在 PLS 模型中，$\boldsymbol{\beta}_{p\times m}$ 和 $\boldsymbol{E}_{n\times m}$ 分别定义为回归系数矩阵和误差矩阵，浓度矩阵 $\boldsymbol{y}_{n\times m}$ 可以描述为

$$\boldsymbol{y}_{n\times m} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times m} + \boldsymbol{E}_{n\times m} \tag{1}$$

以稳定性和可信度作为评价变量重要性的指标，先用稳定性来选择对 $\boldsymbol{y}$ 贡献较大的变量，然后根据可信度选取对模型性能影响较大的变量。通过交叉验证来评价模型的性能，降低了过拟合风险。

1.1　基于蒙特卡罗采样的变量稳定性定义

蒙特卡罗采样是从多个角度对数据集进行评估，因此可以减少过拟合的风险。采用蒙特卡罗采样法从数据集中随机抽取 k 个样本(通常占数据集的 80%～90%)建立 PLS 回归模型并计算相应的回归系数矩阵，经过 M 次采样后可得到一个回归系数矩阵 $\boldsymbol{B}_{M\times p}$([$\boldsymbol{b}_1$，$\boldsymbol{b}_2$，…，$\boldsymbol{b}_j$，…，$\boldsymbol{b}_p$])，并计算第 j 个波长变量的稳定性值为

$$s_j = \frac{|\bar{b}_j|}{\sqrt{\dfrac{1}{M}\sum_{i=1}^{M}(b_{ij}-\bar{b}_j)^2}} \tag{2}$$

其中，s_j 表示 M 次采样后第 j 个变量的稳定性值，b_{ij} 是第 i 次蒙特卡罗采样中第 j 个变量的回归系数，$\bar{b}_j$ 为 M 次采样后第 j 个变量回归系数的平均值。从式(2)可以看出，$|\bar{b}_j|$ 值越大，标准偏差越小，表明该变量的稳定性值越大，重要性越强。

1.2　基于后向选择的可信度定义

计算当前变量集剔除某个变量后模型性能的变化，分析每个变量对模型性能的可信程度，则采用一种后向选择方法。该方法每执行一次都会从初始数据集中删除一个变量，形成一个新的数据集，并使用该数据集生成一个新模型。以第 j 个变量为例，分别计算当前变量集的校正均方根误差(root mean square error of calibration，RMSEC)RMSEC_0 和当前变量集剔除第 j 个变量后的校正均方根误差 RMSEC_j。与 RMSEC_0 相比，RMSEC_j 变小表明新模型的性能有所提高，意味着第 j 个变量对模型性能有负面影响。第 j 个变量的可信度定义为 r_j

$$r_j = \mathrm{RMSEC}_j - \mathrm{RMSEC}_0 \tag{3}$$

1.3　基于 EDF 的变量保留率

变量的性能相互影响，尚未淘汰的冗余变量可能会误导变量的选择。因此通过迭代选择变量更为可靠。模型随变量的选择而变化，变量的稳定性和可信度也随每次迭代发生变化。每一次迭代变量消除率并不相同。最初，变量集包含许多信息不丰富和不重要的变量，这些波长点将被迅速消除，这是一个“粗略选择”阶段。然后，随着信息不足和不重要变量的减少，消除速度会减慢，因为如果波长点仍然被迅速消除，关键变量可能会被错误地消除。这一阶段称为“精选”。为了实现这两阶段变量的选择，使用指数衰减函数(EDF)强制消除变量。R_i 被定义为第 i 次迭代时的变量保留率。其中，a 和 k 为第 1 次和第 N 次循环时样本集中建模数目，为遍历所有变量，第 N 次设为 2 个变量，所以 $R_1=1$；$R_N=2/p$，在以上条件下，R_i 可以表示为

$$R_i = a\mathrm{e}^{-ki} \tag{4}$$

参数 a 和 k 表示为

$$a = \left(\frac{p}{2}\right)^{\frac{1}{N-1}}$$

$$k = \frac{\ln(p/2)}{N-1} \tag{5}$$

1.4　SCPLS 波长选择方法

SCPLS 以稳定性和可信度作为评价变量重要性的指标，该方法通过交叉验证来评价模型的性能，降低了过拟合的风险。SCPLS 算法具体步骤如下：

Step1：设定循环次数初始值 $i=1$；

Step2：对数据集进行蒙特卡罗采样建立 M 个模型；

Step3：计算第 i 次迭代的变量保留率 R_i；

Step4：依据回归系数计算各波长变量的稳定性指标，剔除稳定性低的变量，选择对浓度矩阵 $\boldsymbol{y}$ 贡献较大的变量；

Step5：计算每个波长变量的可信度，并利用 EDF 选择可信度高的变量，把所选定的变量作为下一次迭代的新子集；循环次数 $i=i+1$；

Step6：若 $i\leqslant N$，依次执行 Step2，Step3，Step4，Step5；若 $i=N+1$，执行 Step7；

Step7：经过 N 次循环，获得 N 个变量子集，分别用这 N 个变量子集建立偏最小二乘(PLS)模型，计算各模型的 RMSECV，并选取 RMSECV 最小的变量集作为最优变量集。

1.5　数据集

在两个真实数据集上进行测试 Zn(Ⅱ)，Cu(Ⅱ) 和 Co(Ⅱ)混合溶液的紫外-可见光谱数据集和 Zn(Ⅱ)和 Co(Ⅱ)混

合溶液的紫外-可见光谱数据集。

1.5.1 UV-Vis 数据集 1

UV-Vis 数据集 1 使用北京普析 T9 紫外可见分光光度仪获得。该数据集含有 27 个 Zn(Ⅱ)，Cu(Ⅱ)和 Co(Ⅱ)混合溶液样品的紫外-可见光谱数据，在 400～700 nm 范围内，间隔 1.0 nm 测量并打印各点的吸光度。混合溶液中 Zn(Ⅱ)，Cu(Ⅱ) 和 Co(Ⅱ)的浓度范围是 0.1～1.0 mg·L^{-1}，用 2-(5-溴-2-吡啶偶氮)-5-二乙氨基苯酚(5-Br-PADAP)溶液作为显色剂。吸光度矩阵 $\boldsymbol{X}$ 包含 27 个样品在 301 个波长点(400～700 nm)的吸光度。浓度矩阵分别为 $\boldsymbol{y}_{Zn}$，$\boldsymbol{y}_{Cu}$ 和 $\boldsymbol{y}_{Co}$。该数据集的原始紫外-可见光谱如图 1 所示。

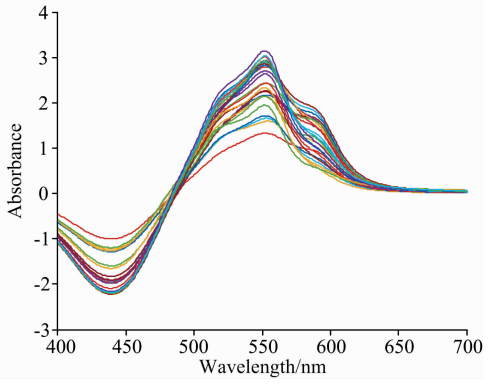

图 1 数据集 1 的原始紫外-可见光谱

Fig. 1 The original UV-Vis spectra of dataset 1

1.5.2 UV-Vis 数据集 2

UV-Vis 数据集 2 采用与 UV-Vis 数据集 1 相同的方法获得。该数据集含有 80 个 Zn(Ⅱ)和 Co(Ⅱ)混合溶液样品的紫外-可见光谱数据。

混合溶液中 Zn(Ⅱ)和 Co(Ⅱ)的浓度范围分别为 0.5～4.0 和 0.25～2.50 mg·L^{-1}，用二甲酚橙溶液作为显色剂。吸光度矩阵 $\boldsymbol{X}$ 包含 80 个样品在 301 个波长点(400～700 nm)的吸光度。浓度矩阵分别为 $\boldsymbol{y}_{Zn}$ 和 $\boldsymbol{y}_{Co}$。该数据集的原始紫外-可见光谱如图 2 所示。

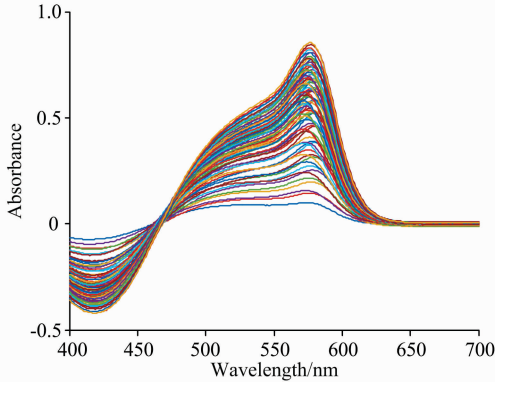

图 2 数据集 2 的原始紫外-可见光谱

Fig. 2 The original UV-Vis spectra of dataset 2

2 结果与讨论

2.1 参数的影响

SCPLS 方法的性能受以下 3 个参数的影响：蒙特卡罗采样率，蒙特卡罗采样数和迭代次数，分别表示为 R，M 和 N。为了分析 R，M 和 N 对 SCPLS 性能的影响，采用了一系列不同的 R，M，N 的值对 SCPLS 的性能进行了测试：M 范围设置为 50～300，间隔为 50 并且 R=0.9，N=100；R 的范围设置为 0.8～0.9，间隔为 0.025 并且 M=100，N=100；N 设置为 10～200，并且 R=0.9，M=100。UV-Vis 数据集 1 中 Zn(Ⅱ)的三个参数的箱形图如图 3 所示。

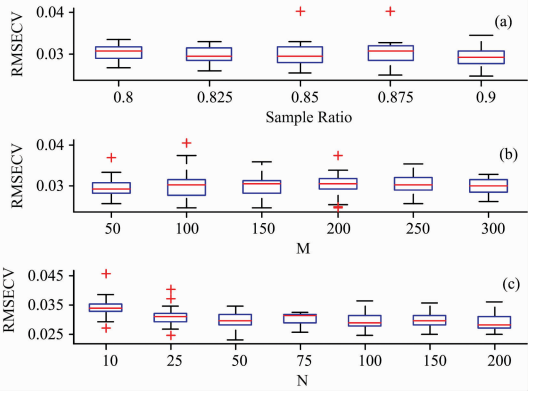

图 3 UV-Vis 数据集 1 中 Zn(Ⅱ)的 R (a)，M (b)和 N (c)的箱形图

Fig. 3 The box-plots of R (a), M (b) and N (c) for Zn(Ⅱ) in UV-Vis dataset 1

从图 3 中可以看出，在 UV-Vis 数据集 1 中，Zn(Ⅱ)的最佳蒙特卡罗采样率 R 和采样数 M 分别为 0.825 和 50，并且迭代次数 N 的最佳范围为 100～200。

2.2 UV-Vis 数据集 1

分析 Zn(Ⅱ)的波长变量选择过程，并采用留一交叉验证方法对模型性能进行评价。RMSECV 的变化趋势和所选变量的数量如图 4 所示。

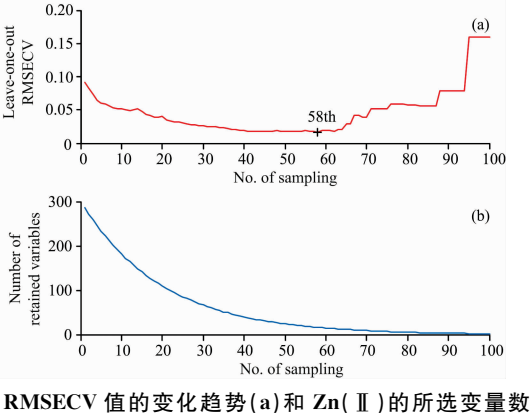

图 4 RMSECV 值的变化趋势(a)和 Zn(Ⅱ)的所选变量数(b)

Fig. 4 The change trend of the RMSECV values (a) and number of selected variables for Zn(Ⅱ) (b)

图 4(a)表明 RMSECV 的值在开始时(迭代 1～35 次)明显下降，是因为消除了最不重要的变量。然后，RMSECV 值的下降(迭代 35～58 次)是由于消除率较低和被淘汰变量的重要性增加所致。在 RMSECV 到达最低点(迭代 58 次)后，继续迭代则会消除关键变量，因此 RMSECV 值开始增加(迭代 58～100)。在所有迭代完成之后，选择具有最小 RMSECV 值的子集作为最优子集。图 4(b)表明变量的数量先是迅速减少，然后随着迭代次数的增加而减慢，这两个阶段被认为是“粗略选择”和“细化选择”阶段。

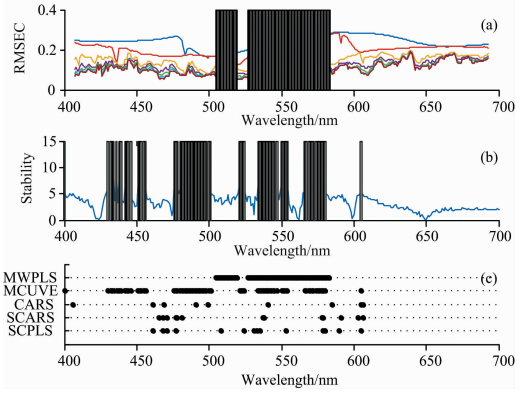

图 5　5 种不同方法对 UV-Vis 数据集 1 中 Zn(Ⅱ)的波长选择性能的比较

(a)：RMSEC 较小；(b)：MCUVE 法稳定性较高；(c)：各方法比较

Fig. 5　Comparison of wavelengths selected by five different methods for Zn(Ⅱ) of UV-Vis dataset 1

(a)：MWPLS method for smaller RMSEC;

(b)：MCUVE method for more slability;

(c)：Comparision results of wavelength selection of each method

为了进一步评估 SCPLS 方法的性能，在同一数据集上应用了几种常用的变量选择方法，即 MWPLS，MCUVE，CARS 和 SCARS，所选波长结果如图 5(a,b,c)所示。

图 5(a)表明用 MWPLS(窗口大小为 15)筛选出较小 RMSEC 的波长区域为 505～519 和 527～583 nm，图 5(b)表明用 MCUVE 筛选出较高稳定性的波长区域为 430～456，476～501，521～524，535～554 和 556～580 nm。在图 5(c)中可以看到与 MWPLS 和 MCUVE 相比，CARS，SCARS 和 SCPLS 会优先选择离散波长变量。尽管这 5 种波长选择方法筛选出一些相同的波长点，但是它们仍有许多不同之处。MCUVE 没有选择稳定性的局部峰 461～471 nm，相反，使用 CARS，SCARS 和 SCPLS 从该区域中选择了几个波长点，表明这些方法以迭代方式可以更好地选择潜在变量。并且通过 CARS，SCARS 和 SCPLS 选择的波长点也不相同。MWPLS 选中波长区域 505～510 和 527～533 nm，SCPLS 在该区域选中 3 个波长点，但 MCUVE，CARS 和 SCARS 在该区域没有选中波长点。这是因为 MWPLS 和 SCPLS 都以变量对模型性能的影响为标准进行选择的。

同样，在 UV-Vis 数据集 1 中，SCPLS 对 Cu(Ⅱ)和 Co(Ⅱ)的波长选择性能也与 MWPLS，MCUVE，CARS 和 SCARS 进行了比较，结果如图 6(a—f)所示。

从图 6(c)中可以看出 CARS，SCARS 和 SCPLS 比 MWPLS 和 MCUVE 选择的变量少。SCPLS 在 515～530 和 570～580 nm 波段的波长选择与 MWPLS，MCUVE，CARS 或 SCARS 方法的选择相似，但是 SCPLS 选择的波长比 CARS 和 SCARS 所选择的波长要少。在图 6(f)中，CARS，SCARS 和 SCPLS 的波长选择差异更大。SCPLS 在 RMSEC 的局部槽或稳定性的局部峰中选择了许多其他方法没有选择

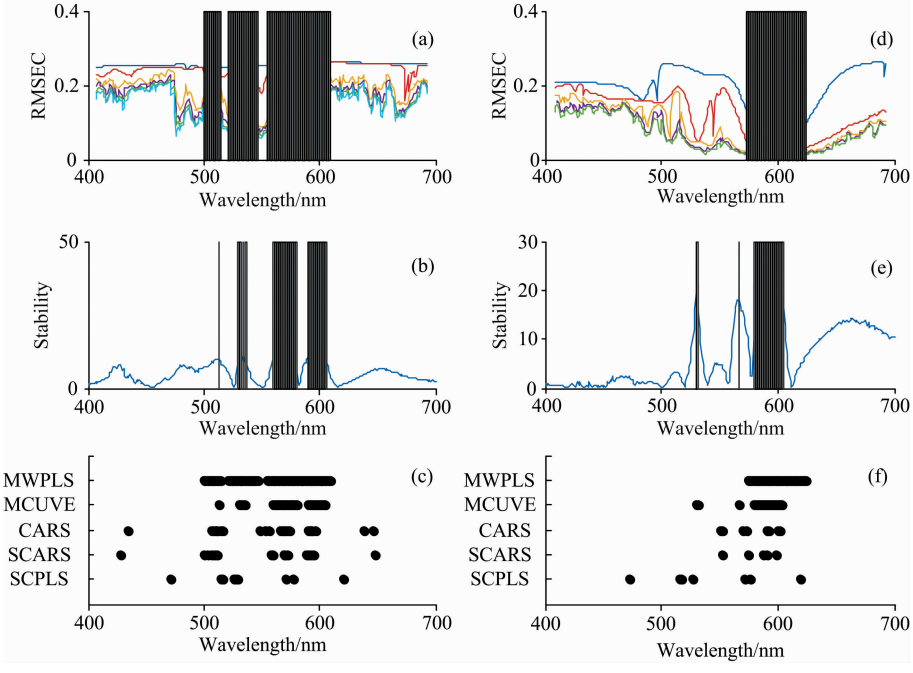

图 6　5 种不同方法对 UV-Vis 数据集 1 中 Cu(Ⅱ)(a—c)和 Co(Ⅱ)(d—f)波长选择性能的比较

Fig. 6　Comparison of wavelengths selected by five different methods for Cu(Ⅱ) (a—c) and Co(Ⅱ) (d—f) of UV-Vis dataset 1

的波长点，例如 400～475 和 515～520 nm 波段。这个结果反映了 SCPLS 具有更好地选择潜在变量的能力。

表 1 展示了使用全光谱 PLS，MWPLS，MCUVE，CARS，SCARS 和 SCPLS 的 RMSECV 值、潜在变量数以及选定变量数。与全光谱 PLS 相比，波长选择方法选择较少的变量，并获得较小的 RMSECV 值。在所有方法中，SCPLS 获得最小的 RMSECV 值，并且使用 SCPLS 得到的潜在变量数和选择的变量数与 CARS 和 SCARS 相似。此外，使用 SCPLS 方法得到的 Zn(Ⅱ)，Cu(Ⅱ)和 Co(Ⅱ)的 RMSECV 值分别比全光谱 PLS 下降 60.5%，40.2% 和 31.8%，与 SCARS 相比分别下降 29.8%，26.1%和 0.8%。

表 1 UV-Vis 数据集 1 的 6 种方法的性能结果

Table 1 Results of six methods of variable selection for UV-Vis dataset 1

Method	Zn			Cu			Co		
	RMSECV	nLVs[a]	nVAR[b]	RMSECV	nLVs[a]	nVAR[b]	RMSECV	nLVs[a]	nVAR[b]
PLS[c]	0.071 31	7	301	0.044 44	6	301	0.017 63	5	301
MWPLS	0.065 63	7	72	0.038 90	6	96	0.013 92	5	51
MCUVE	0.060 18	7	77	0.037 42	6	43	0.012 95	5	28
CARS	0.046 06	6	9	0.036 94	5	26	0.012 13	5	8
SCARS	0.040 15	6	13	0.035 98	5	20	0.012 17	5	5
SCPLS	0.028 20	7	14	0.026 58	6	13	0.012 07	5	8

注：nLVs[a]表示 PLS 的潜在变量的数量；nVAR[b]表示选定变量的数量；PLS[c]采用全光谱(400～700 nm)PLS 建模

UV-Vis 数据集 1 经 SCPLS 建模后样本浓度预测值和实际值之间的散点图如图 7 所示，Zn(Ⅱ)，Cu(Ⅱ)和 Co(Ⅱ)平均相对误差分别为 2.14%，1.25%和 0.74%，其中 Zn(Ⅱ)的最大相对误差为 4.67%，Cu(Ⅱ)的最大相对误差为 3.99%，Co(Ⅱ)的最大相对误差为 3.12%，该方法检测精度较高，效果较理想。

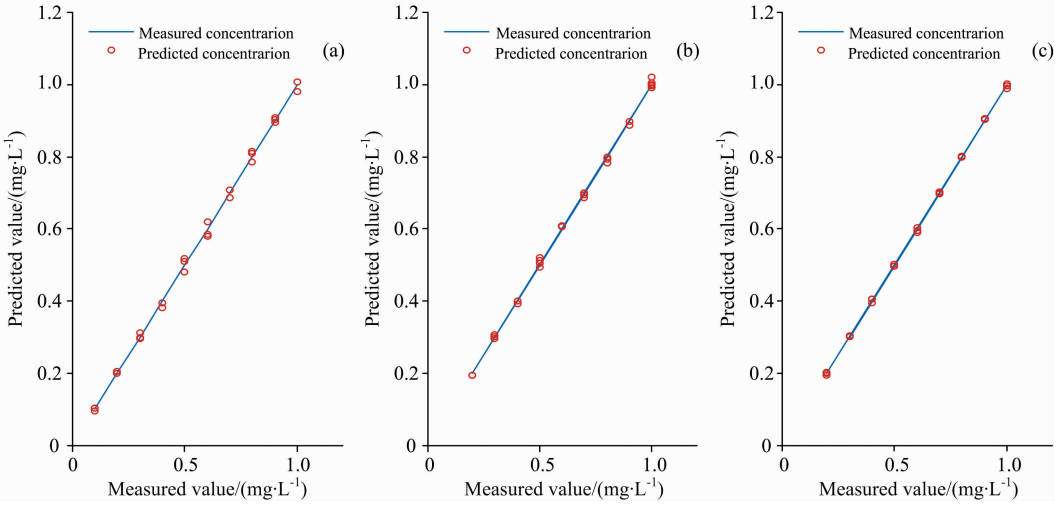

图 7 实际值与预测值的散点图

(a)：Zn(Ⅱ)；(b)：Cu(Ⅱ)；(c)：Co(Ⅱ)

Fig. 7 Scatter diagram of actual value and predicted

(a)：Zn(Ⅱ)；(b)：Cu(Ⅱ)；(c)：Co(Ⅱ)

2.3 UV-Vis 数据集 2

将 SCPLS 方法应用于 UV-Vis 数据集 2，与 UV-Vis 数据集 1 相似，也对三个参数进行了优化。对于 UV-Vis 数据集 2 中的 Zn(Ⅱ)，R，M 和 N 分别设为 0.9，50 和 150；对于 UV-Vis 数据集 2 中的 Co(Ⅱ)，R，M 和 N 分别设为 0.9，250 和 100。采用十折交叉验证 RMSECV 对模型性能进行了评价。

图 8 显示了五种方法(MWPLS，MCUVE，CARS，SCARS 和 SCPLS)对 UV-Vis 数据集 2 的波长选择情况。许多变量既被 SCPLS 选择，也被 CARS 或 SCARS 选择，但是仍有许多不同。例如，在图 8(c)中，SCPLS 在 420～450 nm 波段范围内选择了三个波长点，其中一个在稳定性的局部峰和一个在 RMSEC 的局部槽，而其他方法则没有选择。

表 2 展示了使用六种方法的 RMSECV 值、潜在变量数以及选定变量数。波长选择方法的 RMSECV 值都小于全光谱 PLS 方法的 RMSECV 值。与 UV-Vis 数据集 1 的结果相比，使用 MWPLS 的结果要好于使用 MCUVE，CARS 和 SCARS方法的结果，这表明基于模型统计性能评价变量的

表 2　UV-Vis 数据集 2 的 6 种方法的性能结果

Table 2　The results of six methods of variable selection for UV-Vis dataset 2

Method	Zn			Co		
	RMSECV	nLVs[a]	nVAR[b]	RMSECV	nLVs[a]	nVAR[b]
PLS[c]	0.137 59	5	301	0.080 07	4	301
MWPLS	0.110 16	5	57	0.067 48	4	58
MCUVE	0.126 11	5	35	0.072 39	4	142
CARS	0.122 12	5	7	0.069 40	3	5
SCARS	0.128 79	4	5	0.069 34	3	5
SCPLS	0.083 39	5	10	0.060 12	4	5

注：nLVs[a]表示 PLS 的潜在变量的数量；nVAR[b]表示选定变量的数量；PLS[c]采用全光谱(400～700 nm)PLS 建模

选择方法要比基于此数据集中的变量属性评价变量的选择方法有更好的性能。SCPLS 是基于变量稳定性和可信度来评价变量的，该方法得到了所有方法的最小 RMSECV 值。并且使用 SCPLS 方法得到的 Zn(Ⅱ)和 Co(Ⅱ)的 RMSECV 值分别比全光谱 PLS 下降 39.4%和 24.9%，与 SCARS 相比分别下降 35.3%和 13.3%。

UV-Vis 数据集 2 经 SCPLS 建模后样本浓度预测值和实际值之间的散点图如图 9 所示，Zn(Ⅱ)和 Co(Ⅱ)平均相对误差分别为 1.23%，1.10%，其中 Zn(Ⅱ)的最大相对误差为 4.45%，Co(Ⅱ)的最大相对误差为 4.57%，该方法检测精度较高，效果较理想。

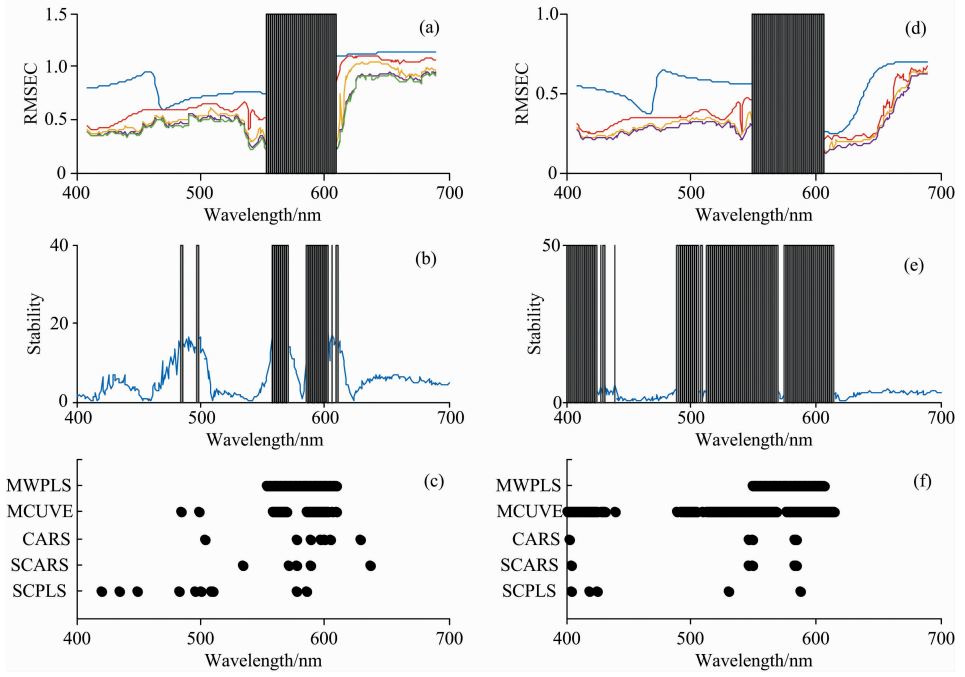

图 8　5 种不同方法对 UV-Vis 数据集 2 中 Zn(Ⅱ)(a—c)和 Co(Ⅱ)(d—f)波长选择性能的比较

Fig. 8　Comparison of wavelengths selected by five different methods for Zn(Ⅱ)(a—c) and Co(Ⅱ)(d—f) of UV-Vis dataset 2

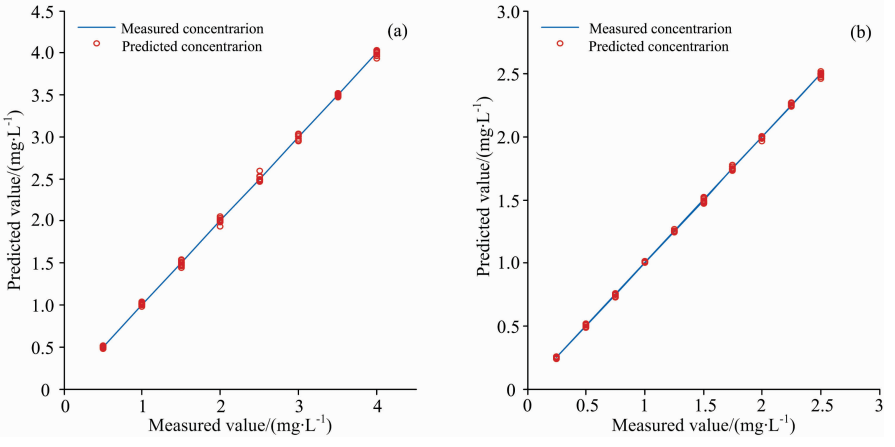

图 9　实际值与预测值的散点图

(a)：Zn(Ⅱ)；(b)：Co(Ⅱ)

Fig. 9　Scatter diagram of actual value and predicted

(a)：Zn(Ⅱ)；(b)：Co(Ⅱ)

3 结 论

针对多组分金属离子混合溶液的紫外-可见吸收光谱重叠难以解析分离的问题，提出了一种紫外-可见特征波长选择方法，该方法基于稳定性和可信度来选择贡献大、噪声低、对模型有积极影响的变量。以迭代的方式选择光谱波长；然后，选择的变量作为新的变量集进入下一次迭代进行变量选择。模型性能最好的子集(最小 RMSECV)被认为是最优子集。用两种紫外-可见光(UV-Vis)数据集对 SCPLS 的性能进行了测试，结果表明 SCPLS 选择的潜在变量数和波长数比 MWPLS 和 MCUVE 少，与使用 CARS 和 SCARS 获得的潜在变量数和波长数相近。SCPLS 有效增强了波长变量选择方法对变量重要性的评估能力，所选择的波长变量建立的模型性能得到有效提高，并得到了所有方法中最小的 RMSECV 值，为复杂光谱波长变量选择提供了一种新方法。

References

[1] TANG Bin, WEI Biao, MAO Ben-jiang, et al(汤 斌, 魏 彪, 毛本将, 等). Laser & Optoelectronics Progress, 2014, 51(4): 043002.
[2] Zhu Hongqiu, Wang Guowei, Yang Chunhua, et al. Transactions of Nonferrous Metals Society of China, 2013, 23(7): 2181.
[3] Suhandy D, Yulia M, Ogawa Y, et al. Engineering in Agriculture, Environment and Food, 2013, 6(3): 111.
[4] Chen H Z, Pan T, Chen J M, et al. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1): 139.
[5] Brusco Michael J. Computational Statistics & Data Analysis, 2014, 77: 38.
[6] Xu Deng, Fan Wei, Lv Huiying, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2014, 123: 430.
[7] Liu S S, Zhang J, Lin S H, et al. Laser & Optoelectronics Progress, 2018, 55(2): 023001.
[8] Rahman A, Kondo N, Ogawa Y, et al. Biosystems Engineering, 2016, 141: 12.
[9] WANG Yu-tian, YANG Zhe, HOU Pei-guo, et al(王玉田, 杨 哲, 侯培国, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(7): 2144.
[10] ZHU Hong-qiu, CHEN Jun-ming, YIN Dong-hang, et al(朱红求, 陈俊名, 尹冬航, 等). Journal of Chemical Industry and Engineering(化工学报), 2017, (3): 206.
[11] Zhang Bo, Sun Lanxiang, Yu Haibin, et al. Spectrochimica Acta Part B, 2015, 107(1): 32.

A Wavelength Selection Method of UV-Vis Based on Variable Stability and Credibility

SUN Tao, YANG Chun-hua, ZHU Hong-qiu*, LI Yong-gang, CHEN Jun-ming
School of Automation, Central South University, Changsha 410083, China

Abstract This paper proposes a wavelength selection method based on stability and credibility partial least squares (SCPLS), to solve the problem that the ultraviolet visible (UV-Vis) spectra of multi-metal ion mixture solution were seriously overlapped and difficult to separate. In SCPLS, an exponentially decreasing function (EDF) is applied to select the variables in an iterative manner. In each iteration, a series of models are built with the sub-datasets sampled using the Monte Carlo strategy. Then, the stability and credibility of each variable are calculated, and the variables with high stability and credibility are selected by the EDF. Subsequently, the selected variables are used to construct a new variable subset for the next iteration. After the selection iterations are terminated, the root mean square error of cross validation (RMSECV) of each subset is calculated. The variable subset with the minimum RMSECV value is considered to be the optimal variable subset. The performance of SCPLS is evaluated with UV-Vis Spectral data set of Zn(Ⅱ), Cu(Ⅱ) and Co(Ⅱ) mixture solution and UV-Vis Spectral data set of Zn(Ⅱ) and Co(Ⅱ) mixture solution, and compared with that of full spectrum partial least squares (PLS) modeling and the moving window PLS (MWPLS), Monte Carlo uninformative variable elimination (MC-UVE), competitive adaptive reweighted sampling (CARS) and stability competitive adaptive reweighted sampling (SCARS) methods. The results show that SCPLS can not only reduce the complexity of the wavelength selection, but also ensure the stability of the wavelength selection process. And it can select the subset with the minimum RMSECV value. Thus, the RMSECV of Zn(Ⅱ), Cu(Ⅱ) and Co(Ⅱ) models obtained by SCPLS are 60.5%, 40.2% and 31.8% respectively lower than that of full spectrum PLS, and 29.8%, 26.1% and 0.8% respectively lower than that of SCARS. The average relative error of Zn(Ⅱ), Cu(Ⅱ) and Co(Ⅱ) is 2.14%, 1.25% and 0.74% respectively, of which the maximum relative error of Zn(Ⅱ) is 4.67%, the maximum relative error of Cu(Ⅱ) is 3.99%, and the maximum relative error of Co(Ⅱ) is 3.12%. And the RMSECV of Zn(Ⅱ) and Co(Ⅱ) models obtained by SCPLS are 39.4% and 24.9% re-

spectively lower than that of full spectrum PLS, and 35.3% and 13.3% respectively lower than that of SCARS. The average relative error of Zn(Ⅱ) and Co(Ⅱ) are 1.23% and 1.10% respectively, of which the maximum relative error of Zn(Ⅱ) is 4.45% and the maximum relative error of Co(Ⅱ) is 4.57%. The proposed method can efficiently improve modeling accuracy.

Keywords Wavelength selection; Stability; Credibility; UV-Visible spectrophotometer

(Received Sep. 19, 2018; accepted Jan. 25, 2019)

* Corresponding author

《光谱学与光谱分析》对来稿英文摘要的要求

来稿英文摘要不符合下列要求者，本刊要求作者重写，这可能要推迟论文发表的时间。

1. 请用符合语法的英文，要求言简意明、确切地论述文章的主要内容，**突出创新之处**。

2. 应拥有与论文同等量的主要信息，包括四个要素，即研究目的、方法、结果、结论。其中后两个要素最重要。有时一个句子即可包含前两个要素，例如“用某种改进的 ICP-AES 测量了鱼池水样的痕量铅”。但有些情况下，英文摘要可包括研究工作的主要对象和范围，以及具有情报价值的其他重要信息。在结果部分最好有定量数据，如检测限、相对标准偏差等；结论部分最好指出方法或结果的优点和意义。

3. 句型力求简单，尽量采用被动式，建议经专业英语翻译机构润色，与中文摘要相对应。用 A4 复印纸单面打印。

4. 摘要不应有引言中出现的内容，换言之，摘要中必须写进的内容应尽量避免在引言中出现。摘要也不要对论文内容作解释和评论，不得简单重复题名中已有的信息；不用非公知公用的符号和术语；不用引文，除非该论文证实或否定了他人已发表的论文。缩略语、略称、代号，除相邻专业的读者也能清楚地理解外，在首次出现时必须加以说明，例如用括号写出全称。