计算机与智能教育技术

2026年10月 第1卷第3期

# 基于高斯过程回归的深紫外拉曼光谱反演硫酸钠溶液浓度算法

张丁方，吴 祺，陈 爽，孟 鑫*

江苏师范大学物理与光电学院，江苏徐州，221116

**摘　要** | 针对硫酸钠溶液浓度快速无损检测的需求，提出一种基于深紫外拉曼光谱与高斯过程回归（GPR）的硫酸钠溶液浓度反演方法。采用课题组研制的261nm深紫外拉曼光谱检测系统，采集12个浓度梯度（0～2.0 mol·$L^{-1}$）纯硫酸钠溶液共120条光谱。数据处理方面：提出以$SO_4^{2-}$ $\nu_1$特征峰为内部标准的波长自校准方法，修正了数据波数坐标约+0.3%的系统比例偏差，自校准后$SO_4^{2-}$各振动峰位与文献值偏差小于5 $cm^{-1}$；系统比较了静默区中值扣除与自适应迭代重加权惩罚最小二乘（airPLS）两种基线校正方法，发现airPLS在强特征峰区存在基线过拟合现象，导致峰区信息严重失真。建模方面：构建“全谱主成分分析数据驱动特征＋$\nu_1$峰物理特征”的9维混合特征，采用Matérn 5/2核GPR建立定量模型。留一浓度交叉验证结果表明：所提方法RMSE为0.0622 mol·$L^{-1}$，决定系数$R^2$为0.9914，均优于单变量线性回归（0.0699 mol·$L^{-1}$）、偏最小二乘回归（0.1015 mol·$L^{-1}$）和支持向量回归（0.2295 mol·$L^{-1}$）；同时GPR可输出预测不确定度，对测量结果置信度评估具有实用价值。

**关键词** | 拉曼光谱；浓度反演；高斯过程回归；基线校正；波长自校准；硫酸钠

Copyright © 2026 by the author(s).

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

https://creativecommons.org/licenses/by-nc/4.0/

## 1 引言

硫酸钠（$Na_2SO_4$）是重要的基础无机化工原料，广泛应用于洗涤剂、玻璃、造纸、纺织印染等行业，其水溶液浓度是产品质量控制与工艺优化的关键参数之一。溶液中$Na_2SO_4$浓度的常规测定方法包括硫酸钡沉淀重量法、滴定法等化学分析方法，以及折射率法、电导率法等物性方法：化学分析法准确度高，但操作繁琐、耗时较长；物性方法响应快，但测量的是溶液的整体物理性质，对溶质组分缺乏化学选择性，在多组分共存的混合溶液中难以给出单一组分的准确浓度。因此，发展一种快速、准确且具有化学选择性的硫酸钠溶液浓度检测方法具有实际意义。

拉曼光谱基于分子振动能级的非弹性散射，提

基金项目：徐州市产业前瞻创新产业集群研发方向项目（KC25132）。

作者简介：张丁方，江苏师范大学，学生，研究方向为拉曼光谱检测与分析；吴祺，江苏师范大学，学生，研究方向为LIBS 光谱分析；陈爽，江苏师范大学，学生，研究方向为拉曼光谱检测与分析；孟鑫（*通讯作者），江苏师范大学副教授，研究方向为激光光谱测试技术。

文章引用：张丁方, 吴祺, 陈爽, 孟鑫. 基于高斯过程回归的深紫外拉曼光谱反演硫酸钠溶液浓度算法[J]. 计算机与智能教育技术, 2026, 1(3)：102-112.

https://doi.org/10.68070/a7pqjd92

供分子水平的“化学指纹”信息，具有天然的化学特异性。$SO_4^{2-}$离子在约981 $cm^{-1}$处具有特征的对称伸缩振动峰（$\nu_1$），其强度与浓度直接相关，可作为定量分析的理想探针[1-2]。在激发波长选择上，拉曼散射强度与激发频率的四次方成正比，短波长激发具有信号强度优势；更重要的是，采用261 nm深紫外激发时，斯托克斯拉曼光谱整体位于常见有机污染物荧光发射带之外，可从源头上避开荧光干扰，有利于复杂实际样品的检测[3]。此外，拉曼测量无需取样和试剂，不破坏样品，适合溶液浓度的快速无损检测。

在光谱定量建模方面，传统方法以单变量回归（基于特征峰强度或峰面积）和偏最小二乘回归（PLSR）等多元校正方法为主[4-5]。近年来，机器学习方法在光谱定量领域得到广泛应用。其中，高斯过程回归（Gaussian Process Regression, GPR）作为一种非参数贝叶斯方法，能够灵活建模非线性响应，天然输出预测不确定度，且通过核函数引入光滑性先验，在小样本条件下具有良好的泛化性能，与光谱定量标定中“标定样本数量有限”的实际情况相契合[6]。

在硫酸钠溶液浓度检测的工业应用场景中，该方法可用于化工结晶过程监控、湿法冶金卤水分析、工业废水硫酸钠排放监测等环节，相比传统化学滴定方法可将单样品分析时间从数十分钟缩短至数秒，显著提升在线过程分析的效率。目前深紫外拉曼定量研究中，多数工作采用随机划分训练集与测试集的验证方式，容易出现模型对已知浓度水平过拟合、对外推浓度预测能力不足的问题；同时，传统线性校正模型在高浓度段因离子缔合效应出现非线性响应时，预测精度往往明显下降。

值得注意的是，光谱数据处理环节（波长坐标校准、基线校正、特征提取）对定量模型性能具有决定性影响，但相关研究往往对预处理方法的适用性讨论不足。例如，经典的自适应迭代重加权惩罚最小二乘（airPLS）基线校正方法在荧光背景复杂的紫外拉曼光谱上的适用性有待检验[7]。

作为智能计算技术在光谱定量分析领域的典型应用，本文构建的高斯过程回归反演算法框架，也可进一步封装为智能实验教学模块，用于分析化学、仪器分析课程中「光谱预处理-机器学习建模」环节的虚拟仿真教学，帮助学生直观理解化学计量学方法的原理与流程，为智能教育场景下的实验数据分析提供可复用的算法支撑。

针对上述问题，本文以纯$Na_2SO_4$溶液为对象，开展深紫外拉曼光谱浓度反演研究：1）针对数据波数坐标存在系统偏差的问题，提出以$SO_4^{2-}$ $\nu_1$峰为内部标准的波长自校准方法；2）系统比较静默区中值扣除与airPLS两种基线校正方法对定量精度的影响；3）构建融合全谱主成分（PCA）数据驱动特征与$\nu_1$峰物理特征的混合特征集，建立Matérn 5/2核GPR定量模型，并通过留一浓度交叉验证（LOCO-CV）与单变量回归、PLSR、支持向量回归（SVR）等方法进行全面对比。

## 2 实验系统与数据采集

### 2.1 深紫外拉曼光谱系统

实验采用的深紫外拉曼光谱检测系统如图1所示，与文献[11]所报道的系统一致，主要由261 nm全固态紫外激光器、石英比色皿、信号收集光路与拉曼光谱仪组成。激发光源为二极管泵浦全固态紫外激光器：以808 nm激光二极管泵浦Nd:YLF晶体，经腔内偏振片选择1047 nm基频振荡，采用声光调Q技术获得高峰值功率脉冲，再经两级LBO非线性倍频晶体完成二倍频与四倍频转换，输出261 nm紫外脉冲激光（波长261±0.1 nm，输出功率≥100 mW，脉冲重复频率1 kHz）。激光经带孔反射镜与聚焦透镜入射石英比色皿中的样品；拉曼散射光由收集透镜准直，穿过带孔反射镜并经滤光片滤除瑞利散射后，由透镜耦合至光纤(SMA905接口，纤芯600 μm)，导入光谱仪。

光谱仪采用Littrow光路构型，以单一离轴抛物面镜同时完成准直与聚焦，结构紧凑、光能利用率高；配合3820线/mm高刻线密度平面光栅与50 μm入射狭缝，探测波段264～295.6 nm，光谱分辨率优于0.07 nm，自校准后覆盖拉曼位移约420～4390 $cm^{-1}$，完整包含$SO_4^{2-}$各振动峰（$\nu_1/\nu_2/\nu_4$）与水OH伸缩振动带，满足溶液组分定量分析的需求。系统主要参数见表1。探测器采用背照式深紫外增强CCD，工作制冷温度为−20 ℃，在260~300 nm波段量子效率优于60%，有效抑制深紫外波段的暗噪声。测量过程中样品室保持室温25 ℃、相对湿度40%~60%，避免环境温湿度波动对激光输出功率和光谱仪响应稳定性的影响。石英比色皿光程为10 mm，样品液面距激光聚焦点10 mm，采用前向收集方式减少溶液自吸收对散射信号的衰减。


https://doi.org/10.68070/a7pqjd92

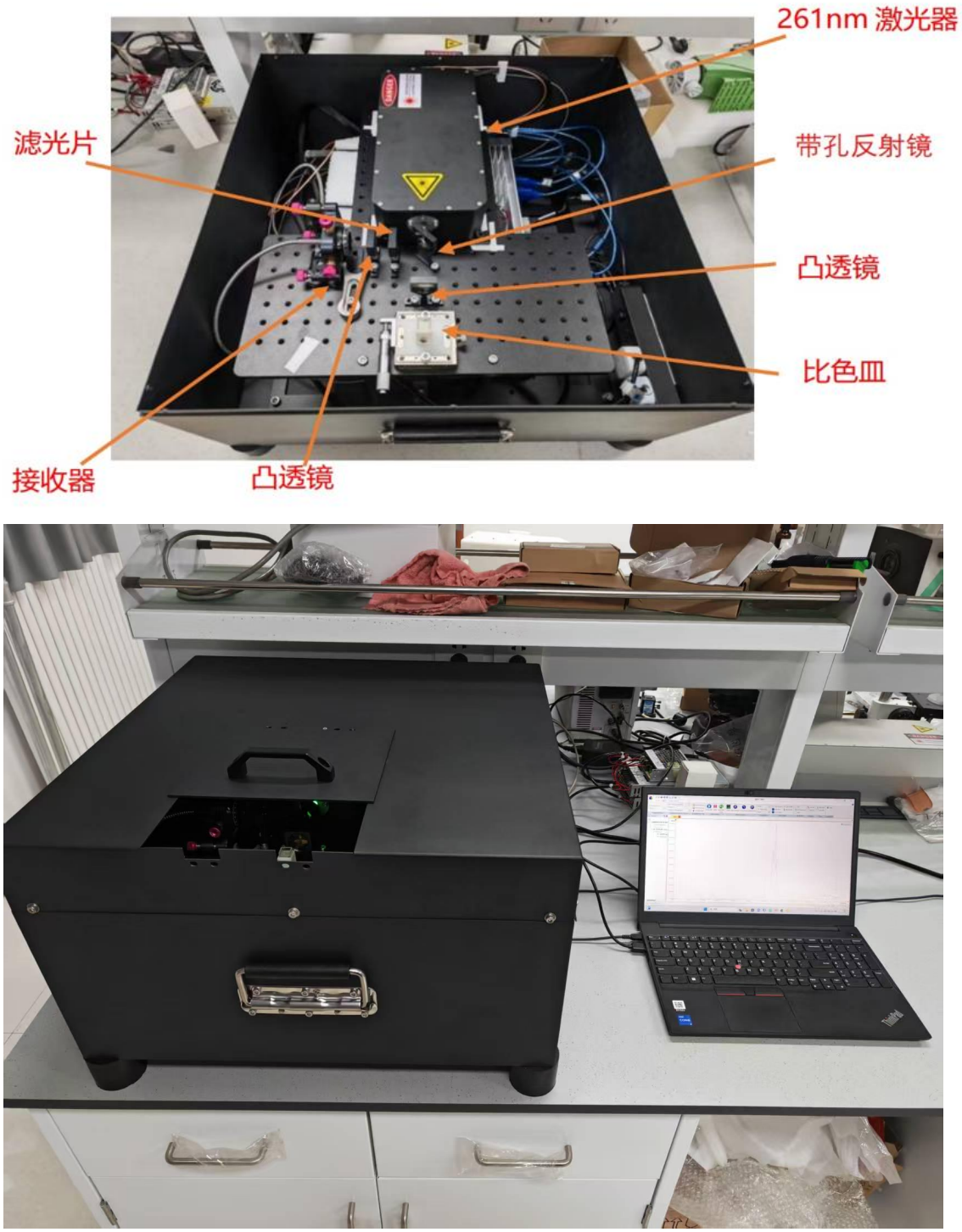



图 1 深紫外拉曼光谱检测系统：(a) 系统内部光路实物；(b) 整机系统实物

表 1 拉曼光谱检测系统主要参数

| 参数 | 参数值 |
|---|---|
| 激发光源 | 全固态紫外激光器（Nd:YLF 晶体，808 nm LD 泵浦，声光调 Q，LBO 四倍频） |
| 激光波长 | 261±0.1 nm |
| 输出功率 | ≥100 mW |
| 脉冲重复频率 | 1 kHz |

续 表 1

| 参数 | 参数值 |
|---|---|
| 光谱仪光路 | Littrow 构型（单一离轴抛物面镜集成准直与聚焦） |
| 探测波段 | 264～295.6 nm |
| 衍射光栅 | 3820 线/mm 平面反射光栅 |
| 入射狭缝 | 50 μm |
| 抛物面镜焦距 | 200 mm |
| 光谱分辨率 | 优于0.07 nm |
| 探测器 | 背照式 CCD（Hamamatsu S10420） |
| 瑞利滤除 | 261 nm 滤光片 |
| 信号收集 | 石英比色皿，透镜收集，光纤耦合（SMA905，纤芯600 μm） |
| 积分时间 | 8 s |
| 拉曼位移范围 | 约420～4390 $cm^{-1}$（波长自校准后） |

### 2.2 样品配制与光谱采集

以分析纯$Na_2SO_4$与超纯水配制12个浓度梯度的纯$Na_2SO_4$溶液：0、0.05、0.1、0.2、0.4、0.6、0.8、1.0、1.2、1.5、1.8、2.0 $mol·L^{-1}$，覆盖从稀溶液到接近并超过溶解度极限（20 ℃下$Na_2SO_4$溶解度约1.35 $mol·L^{-1}$）的宽浓度范围。每个浓度溶液独立采集10条光谱（积分时间8 s），共获得120条光谱，构成定量建模数据集。图2(a)给出了不同浓度溶液的原始拉曼光谱，可见$SO_4^{2-}$特征峰随浓度增强、水OH伸缩振动带（约3000～3600 $cm^{-1}$）随浓度发生系统性变化。配制过程中1.8、2.0 $mol·L^{-1}$高浓度样品处于室温过饱和状态，经目视检查无明显晶体沉淀，摇匀后立即完成光谱采集，保证测量过程中样品均一稳定。

## 3 数据处理与算法模型

### 3.1 基于$\nu_1$内部标准的波长自校准

拉曼位移按式(1)由光谱波长换算：

$$\Delta\nu = (1/\lambda_0 - 1/\lambda)\times 10^7 \quad (1)$$

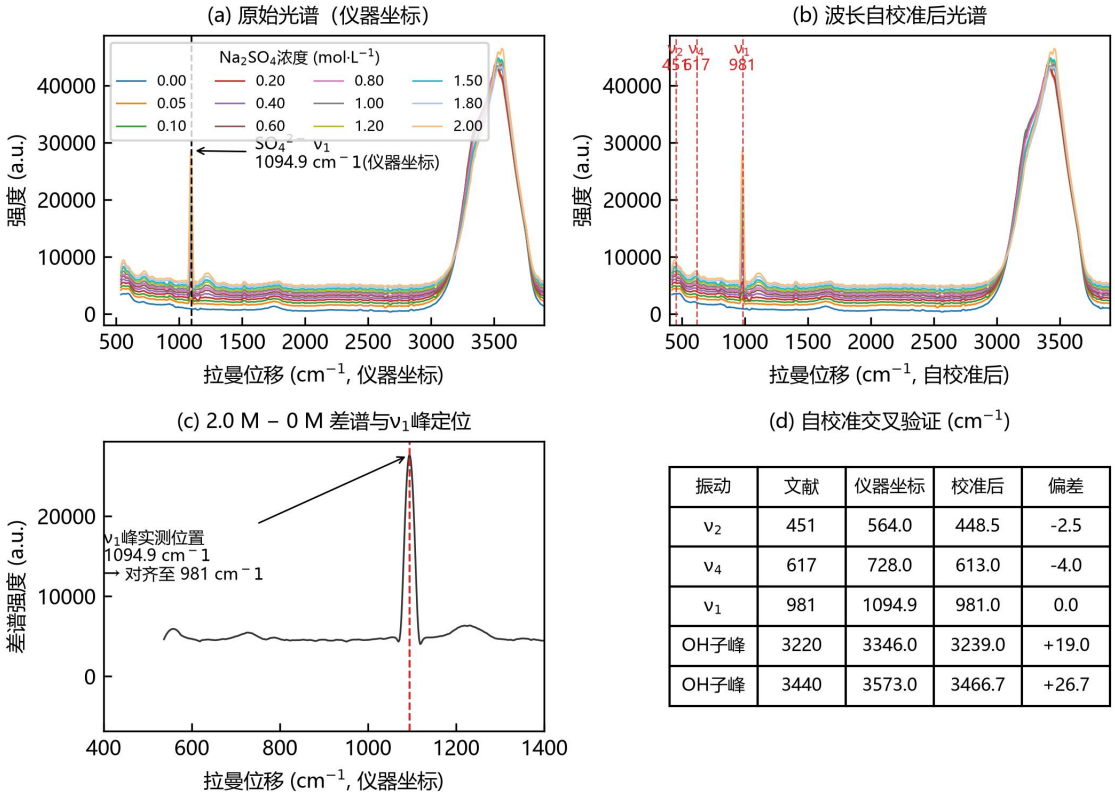
式中：$\lambda_0$为激发波长（261 nm），$\lambda$为散射光波长（nm），$\Delta\nu$为拉曼位移（$cm^{-1}$）。

在数据分析中发现，样本数据的波数坐标存在系统性偏差：$SO_4^{2-}$ $\nu_1$峰在仪器坐标中出现在约1094.9 $cm^{-1}$处，与文献值981 $cm^{-1}$偏离约114 $cm^{-1}$。进一步分析表明，该偏差为波长坐标的比例误差（约+0.3%），而非平移误差：以同一比例系数换算，$\nu_2$（文献值451 $cm^{-1}$）、$\nu_4$（617 $cm^{-1}$）及水OH伸缩振动子峰（约3220、3440 $cm^{-1}$[9]）的观测位置均与文献值一致（图2(c)(d)）。据此，提出以$\nu_1$峰为内部标准的波长自校准方法：

$$\lambda_{true} = \lambda_{file}/s,\ s = \lambda_{file}(\nu_1)/\lambda_{true}(\nu_1) \quad (2)$$

式中：$\lambda_{file}$为数据文件波长坐标，$\lambda_{true}$为自校准后的波长，s为比例系数，本数据集中s=1.00306。自校准后，$\nu_2$峰实测448.5 $cm^{-1}$（偏差−2.5 $cm^{-1}$），$\nu_4$峰实测613.0 $cm^{-1}$（偏差−4.0 $cm^{-1}$），验证了自校准方法的有效性。需要指出，该偏差源于仪器装调与数据采集间的光路几何变化，波长自校准可作为仪器维护周期内保证波数精度的有效手段。


https://doi.org/10.68070/a7pqjd92




| 振动 | 文献 | 仪器坐标 | 校准后 | 偏差 |
|---|---|---|---|---|
| $\nu_2$ | 451 | 564.0 | 448.5 | -2.5 |
| $\nu_4$ | 617 | 728.0 | 613.0 | -4.0 |
| $\nu_1$ | 981 | 1094.9 | 981.0 | 0.0 |
| OH子峰 | 3220 | 3346.0 | 3239.0 | +19.0 |
| OH子峰 | 3440 | 3573.0 | 3466.7 | +26.7 |

图 2 光谱总览与波长自校准：(a) 原始光谱（仪器坐标）；(b) 自校准后光谱；(c) 2.0 mol·L$^{-1}$与0 mol·L$^{-1}$差谱及$\nu_1$峰定位；(d) 自校准交叉验证

### 3.2 光谱预处理与基线校正方法比较

光谱预处理依次包括：Savitzky-Golay（SG）平滑（3阶多项式、21点窗口）以抑制随机噪声[8]；基线校正以扣除宽带背景；以及后续的L2矢量归一化以消除激发强度波动的影响。

宽带背景是深紫外拉曼光谱定量分析的主要干扰因素之一。本数据集的光谱背景呈现两个特点：其一，1800～2800 cm$^{-1}$区间（静默区）无任何拉曼特征，其强度完全来自背景；其二，背景强度随$Na_2SO_4$浓度升高而系统性增强（可能与溶液散射及CCD杂散响应有关），若不加处理将引入虚假的浓度相关性。本文比较了两种基线校正方法：

（1）静默区中值扣除法。以静默区强度的中值作为背景估计，逐条光谱整体扣除。该方法实现简单、物理意义明确（静默区即为“背景探针”），且不引入峰形失真。

（2）airPLS法。采用经典的自适应迭代重加权惩罚最小二乘基线估计[7]，平滑参数λ=1×10$^6$，最大迭代50次。

两种方法的处理结果对比如图3所示。airPLS法在全谱范围内拟合出贴近光谱下包络的平滑基线；然而在$\nu_1$峰区，其拟合基线明显穿透光谱本身：1.0 mol·L$^{-1}$样品校正后$\nu$1区强度低至−11357，2.0 mol·L$^{-1}$样品低至−21353（负值无物理意义）。其原因在于：高浓度下$\nu_1$峰叠加在强度急剧增强的宽带背景上，airPLS的权重迭代机制将峰区内被“过度下压”的光谱点误判为背景，收敛到穿透峰区的平滑曲线。与之相比，静默区中值扣除法完整保留了$\nu_1$峰形，峰高随浓度单调增长。因此，本文采用静默区中值扣除作为基线校正方法，airPLS的定量后果将在4.4节进一步讨论。

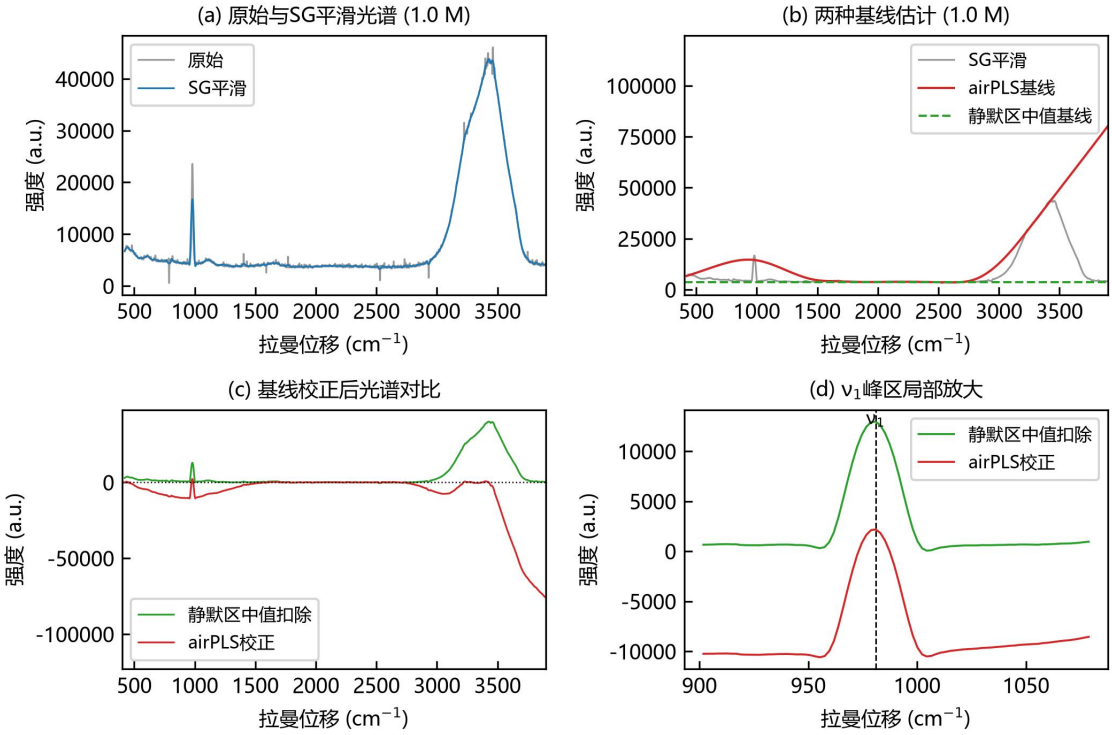



图 3 光谱预处理流程（1.0 mol·L⁻¹样品）：(a) 原始与 SG 平滑光谱；(b) 两种基线估计对比；(c) 基线校正后光谱对比；(d) $\nu_1$峰区局部放大

### 3.3 特征工程

定量建模特征由数据驱动特征与物理特征两部分构成，共9维，见表2。

（1）数据驱动特征（4维）。对基线校正后的光谱进行L2归一化，再进行全谱PCA，取累计方差贡献率达99%的前4个主成分得分（PC1～PC4）作为特征。主成分编码了全谱的整体变化模式，无需先验知识即可捕获与浓度相关的光谱信息。需要说明的是，为严格避免数据泄漏，PCA模型的均值与载荷向量在每个外层训练折内仅由该折的110条训练光谱独立拟合确定，主成分个数固定为前4个（各训练折内前4个主成分的累计方差贡献率均达99%以上）；测试折光谱不参与PCA拟合，仅利用训练折统计量完成投影变换。

（2）物理特征（5维）。围绕$\nu_1$峰提取峰高、峰面积、峰位（强度重心）、半高全宽（FWHM）与静默区均值5个物理量。其中，峰高与峰面积直接反映$SO_4^{2-}$浓度（拉曼强度正比于散射体数目）；峰位与FWHM反映离子对形成、溶剂化壳层变化等微观环境信息[2]；静默区均值监测背景水平，为基线校正质量的补充指示。

表 2 定量建模特征集（9维混合特征）

| 特征 | 符号 | 定义 | 物理意义 |
|---|---|---|---|
| 全谱 PCA | PC1～PC4 | L2归一化光谱前4个主成分得分 | 全谱数据驱动变化模式 |
| $\nu_1$峰高 | h | 940～1040 cm⁻¹窗口内最大值 | $SO_4^{2-}$散射强度 |
| $\nu_1$峰面积 | A | 940～1040 cm⁻¹窗口内积分面积 | $SO_4^{2-}$总散射截面 |
| $\nu_1$峰位 | $\nu_p$ | 窗口内强度重心位置 | 离子对/溶剂化环境 |
| $\nu_1$半高宽 | w | 半峰高处的全宽 | 峰形展宽，微观环境 |
| 静默区均值 | b | 1800～2800 cm⁻¹强度均值 | 背景/基线漂移监测 |

### 3.4 高斯过程回归模型

GPR是一种非参数贝叶斯回归方法[6]。设训练集$D=\{X, y\}$，假定观测值$y=f(X)+\varepsilon$，其中函数$f$服从高斯过程先验$f\sim GP(0, k(x, x'))$，$\varepsilon$为均值为0、方差为$\sigma_n^2$的高斯噪声。本文采用Matérn 5/2核函数:


https://doi.org/10.68070/a7pqjd92

$$k(r) = \sigma_f^2(1 + \sqrt{5}r/\ell + 5r^2/3\ell^2)\exp(-\sqrt{5}r/\ell) \quad (3)$$

式中：$r = \|x - x'\|$ 为输入间欧氏距离，$\sigma_f$为信号幅度，$\sigma_f^2$为信号方差，$\ell$为长度尺度。相比常用的平方指数核，Matérn 5/2核对应二次可微函数类，对特征噪声具有更强的鲁棒性，在本文的有限样本条件下避免了过拟合。核超参数$\theta=\{\sigma_f, \ell, \sigma_n\}$通过最大化边缘似然估计。对测试输入$x^*$，GPR给出预测均值与预测方差：

$$\mu(x^*) = k^{*T}(K + \sigma_n^2 I)^{-1} y \quad (4)$$

$$\sigma^2(x^*) = k(x^*, x^*) - k^{*T}(K + \sigma_n^2 I)^{-1} k^* \quad (5)$$

式中：$K$为训练集协方差矩阵，$k^*$为测试点与训练集的协方差向量。预测方差可作为浓度测量不确定度的贝叶斯估计，这是GPR区别于确定性回归方法的独特优势，对测量结果的可靠性评估具有直接价值。

模型验证采用留一浓度交叉验证（LOCO-CV）：以浓度为分组单位，每次留出某一浓度的全部10条光谱作为测试集，其余11个浓度共110条光谱用于训练；训练集内特征按列做z-score标准化，测试集按训练集统计量变换；特征提取（含PCA拟合）与标准化均严格限定在各训练折内完成，测试折仅应用训练折统计量。LOCO-CV模拟了“对训练中未见浓度水平的外推预测”场景，比随机划分训练/测试集更为严格，能更真实地反映模型的泛化能力[5]。以预测浓度与名义浓度的均方根误差（RMSE）、决定系数（$R^2$）与平均绝对误差（MAE）作为评价指标。整体算法流程如图4所示

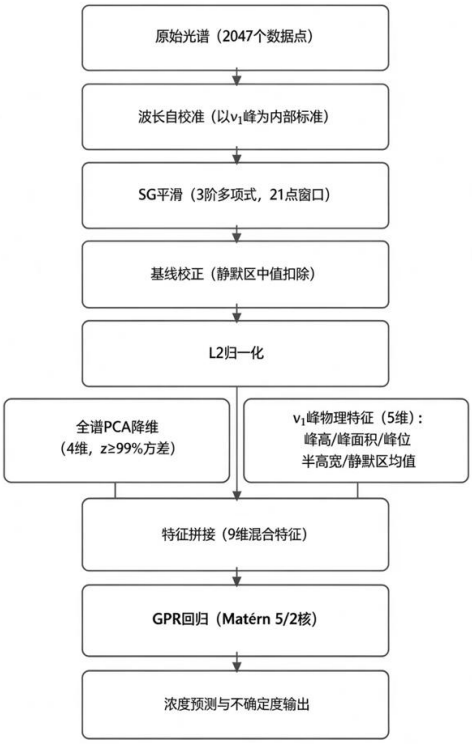



图 4　浓度反演算法整体流程

## 4 结果与讨论

### 4.1 $\nu_1$峰浓度响应分析

基线校正后，$\nu_1$峰区光谱随浓度的演化如图5(a)所示：峰高从0.05 mol·L$^{-1}$时的约1250（计数）单调增长至2.0 mol·L$^{-1}$时的约23500，峰位与峰宽亦随浓度发生细微变化。图5(b)给出了峰高—浓度标定关系，全样本线性拟合决定系数$R^2$=0.9915，表明在0～2.0 mol·L$^{-1}$范围内$\nu_1$峰高与浓度呈良好的线性关系，单变量标定即可获得较高精度，也为后续机器学习建模提供了物理基础。

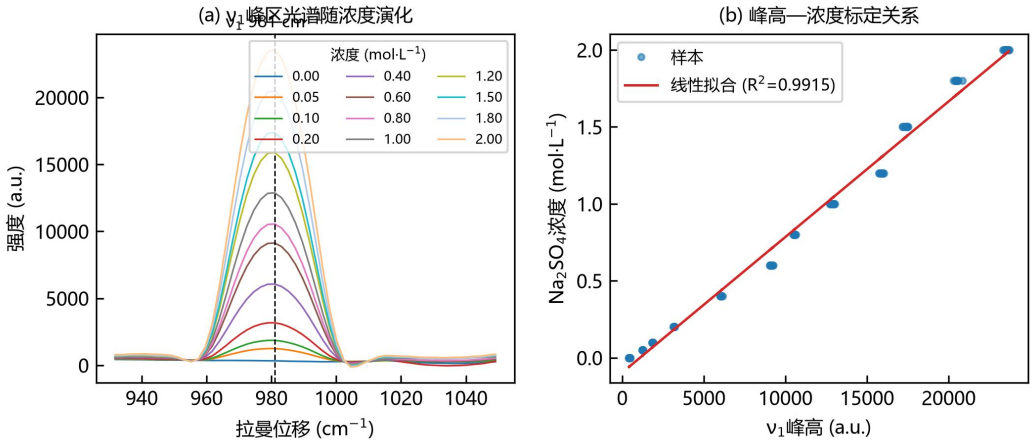



图 5　$\nu_1$峰浓度响应：(a) 不同浓度下$\nu_1$峰区光谱；(b) 峰高—浓度标定关系及线性拟合

### 4.2 多模型性能对比

在统一的LOCO-CV框架下，比较了8种建模方案的定量性能，结果见表3与图6。

表 3 不同建模方案 LOCO-CV 性能对比

| 模型 | 特征/输入 | RMSE/(mol·L$^{-1}$) | $R^2$ | MAE/(mol·L$^{-1}$) |
|---|---|---|---|---|
| M1 单变量线性回归 | $\nu_1$峰高 | 0.0699 | 0.9891 | 0.0576 |
| M2 单变量二次回归 | $\nu_1$峰高 | 0.0729 | 0.9882 | 0.0497 |
| M3 PLSR | 全谱(2047点) | 0.1015 | 0.9770 | 0.0792 |
| M4 SVR | 9维混合特征 | 0.2295 | 0.8826 | 0.1233 |
| M5 GPR | 4维 PCA | 0.0611 | 0.9917 | 0.0407 |
| M6 GPR | 5维物理特征 | 0.0778 | 0.9865 | 0.0573 |
| M7 GPR（本文方法） | 9维混合特征 | 0.0622 | 0.9914 | 0.0452 |
| M8 GPR | airPLS 基线+9维特征 | 0.1175 | 0.9692 | 0.0969 |

注：M3～M8除标注外均采用静默区中值扣除基线；全部模型在同一 LOCO-CV 划分下评估。

各模型超参数均采用统一网格搜索与内部留一交叉验证确定最优值，避免超参数选择对模型对比公平性的干扰：PLSR模型的主成分数通过内部留一交叉验证选取，最终保留4个主成分（累计方差解释率98.7%）；SVR模型采用径向基核函数，惩罚系数C在[0.1,100]范围、核宽度gamma在[0.001,1]范围进行网格搜索，最终取C=10、gamma=0.01；GPR模型的超参数通过L-BFGS算法从10组随机初始值出发最大化边缘似然，避免局部最优。

由表3可知：1）单变量线性回归（M1）已达到RMSE=0.0699 mol·L$^{-1}$的较高精度，印证了$\nu_1$峰高与浓度间稳定的线性关系；2）本文方法（M7，GPR＋9维混合特征）RMSE=0.0622 mol·L$^{-1}$，较单变量回归降低约11%，较PLSR（M3）降低约39%，较SVR（M4）降低约73%；3）纯数据驱动特征（M5，仅4维PCA输入）RMSE为0.0611 mol·L$^{-1}$，与混合特征模型M7的0.0622 mol·L$^{-1}$性能接近，二者差异在测量不确定度范围内，这表明全谱主成分已能捕获与浓度相关的主要光谱信息。需要指出的是，纯PCA特征属于黑箱数据驱动特征，缺乏明确的物理意义；而混合特征中引入的峰高、峰位、半高宽等物理量可直接对应拉曼光谱的经典定量参数，在保持相当预测精度的同时，显著提升了模型结果的可解释性，也便于实际测量中对异常点进行物理层面的溯源分析；4）SVR在小样本、高特征噪声条件下泛化能力较弱，说明核方法与超参数优化策略的选择对建模成败至关重要。

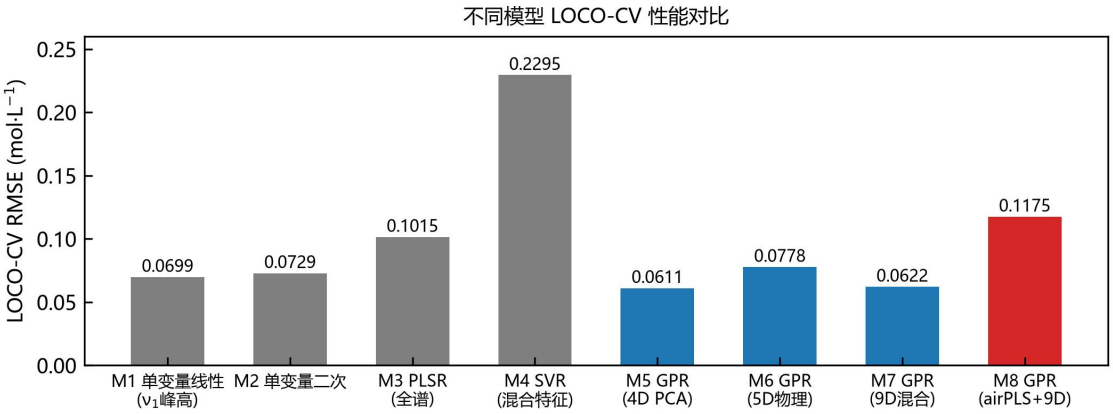



图 6 不同建模方案 LOCO-CV RMSE 对比

### 4.3 本文方法详细结果与误差分析

图7给出了本文方法（M7）的预测结果。图7(a)为预测值—名义浓度散点图及95%置信区间（±1.96σ，σ为GPR预测标准差）：绝大多数样本落在对角线附近，且低浓度段置信区间较窄，高浓度段（1.2～1.5 mol·L$^{-1}$附近）置信区间明显变宽，与图7(b)残差分布一致。


https://doi.org/10.68070/a7pqjd92

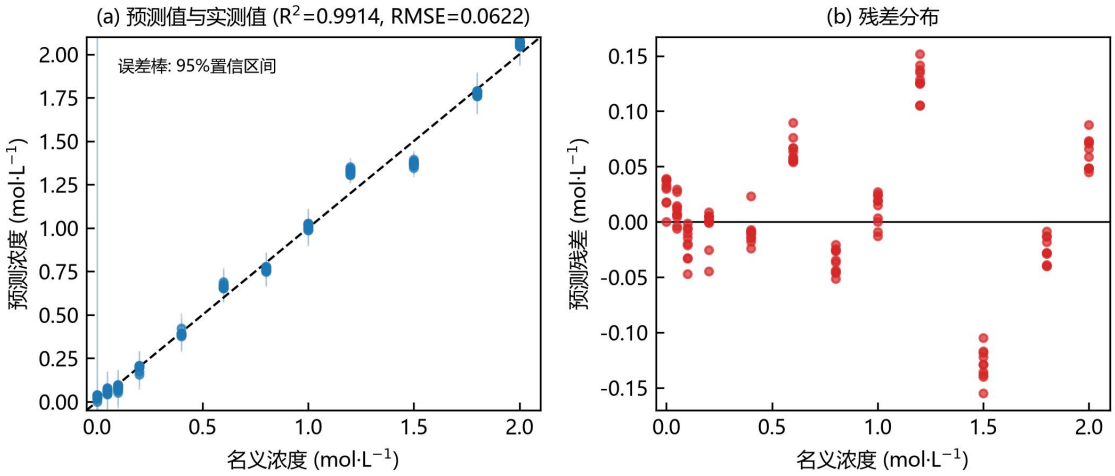



图 7 本文方法 LOCO-CV 预测结果：(a) 预测值与名义浓度对比（误差棒为95%置信区间）；(b) 预测残差分布

分浓度统计结果见表4。在0.05～1.0 mol·L$^{-1}$区间内，各浓度RMSE均不超过0.066 mol·L$^{-1}$，其中0.05、0.2、0.4 mol·L$^{-1}$处RMSE低于0.02 mol·L$^{-1}$；1.2与1.5 mol·L$^{-1}$处误差最大（RMSE约0.129 mol·L$^{-1}$），对应图7中置信区间最宽的区间。分析认为：该浓度区间$\nu_1$峰形发生明显变化（峰位偏移与峰宽增加，与$SO_4^{2-}$离子缔合及溶剂化壳层重组有关[2]），峰形信息与浓度的关系出现局部非线性，跨浓度外推时误差增大；GPR给出的预测不确定度在该区间同步增大，正确地反映了模型置信度的下降。2.0 mol·L$^{-1}$处RMSE回落至0.063 mol·L$^{-1}$，表明模型对高浓度端仍具较好预测能力。上述结果说明，GPR的不确定度输出可为测量异常值的甄别与测量可靠性评估提供定量依据。

表 4 本文方法分浓度 LOCO-CV 误差统计

| 浓度/(mol·L$^{-1}$) | RMSE/(mol·L$^{-1}$) | MAE/(mol·L$^{-1}$) | 平均偏差/(mol·L$^{-1}$) |
|---|---|---|---|
| 0 | 0.0308 | 0.0283 | +0.0283 |
| 0.05 | 0.0150 | 0.0122 | +0.0102 |
| 0.10 | 0.0237 | 0.0192 | −0.0192 |
| 0.20 | 0.0167 | 0.0096 | −0.0046 |
| 0.40 | 0.0150 | 0.0138 | −0.0092 |
| 0.60 | 0.0652 | 0.0643 | +0.0643 |
| 0.80 | 0.0370 | 0.0355 | −0.0355 |
| 1.00 | 0.0176 | 0.0154 | +0.0110 |
| 1.20 | 0.1288 | 0.1280 | +0.1280 |
| 1.50 | 0.1297 | 0.1290 | −0.1290 |
| 1.80 | 0.0280 | 0.0256 | −0.0256 |
| 2.00 | 0.0634 | 0.0619 | +0.0619 |

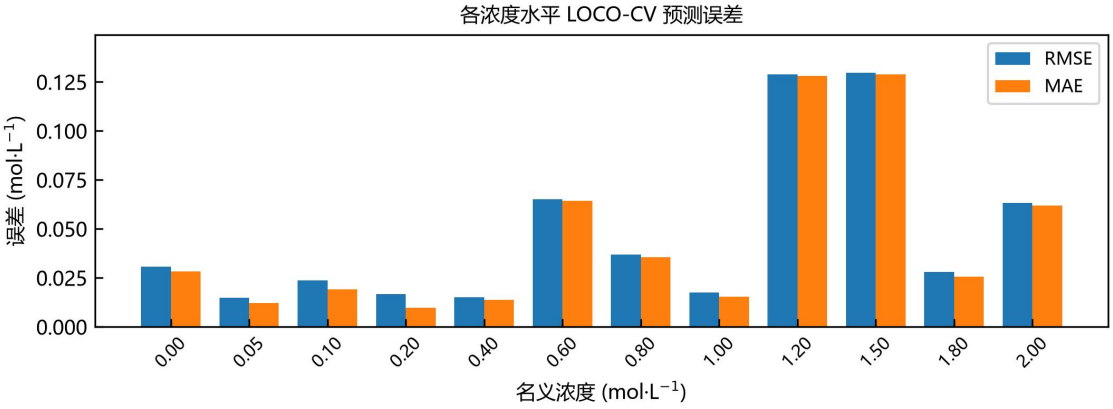



图 8 本文方法各浓度水平预测误差

### 4.4 基线校正方法对定量精度的影响

为量化基线校正方法选择的影响，在其他条件完全一致（9维混合特征、GPR、LOCO-CV）的情况下，将基线由静默区中值扣除替换为airPLS（模型 M8），RMSE 由 0.0622 mol·L$^{-1}$ 恶化至 0.1175 mol·L$^{-1}$，相对退化约89%（以本文方法RMSE为基准）。分浓度结果显示，M8在0.2～2.0 mol·L$^{-1}$全段均出现显著偏差（如2.0 mol·L$^{-1}$处偏差−0.22 mol·L$^{-1}$），与3.2节中观察到的airPLS在强峰区基线过拟合现象相互印证：$\nu_1$峰信息被基线校正破坏后，模型只能依赖水OH带等次级特征，定量精度显著下降。这一结果提示：基线校正方法的选择必须结合光谱形态加以验证，对强特征峰叠加宽带背景的紫外拉曼光谱，简单稳健的静默区扣除法优于追求“自适应”的复杂算法。

## 5 结论

本文面向硫酸钠溶液浓度快速无损检测需求，系统研究了基于深紫外拉曼光谱与高斯过程回归的纯硫酸钠溶液浓度反演方法，主要结论如下：

（1）提出了以$SO_4^{2-}$ $\nu_1$特征峰为内部标准的波长自校准方法，修正了数据波数坐标约+0.3%的系统比例偏差；自校准后$\nu_2$、$\nu_4$等特征峰位与文献值偏差小于5 cm$^{-1}$，为仪器维护周期内的波数精度保障提供了简便有效的手段。

（2）系统比较了静默区中值扣除与airPLS两种基线校正方法，发现airPLS在强特征峰叠加宽带背景的条件下存在基线过拟合，峰区信息严重失真；采用静默区中值扣除的定量模型RMSE较airPLS降低约47%，表明预处理方法的选择对深紫外拉曼定量精度具有决定性影响。

（3）构建了融合全谱PCA数据驱动特征与$\nu_1$峰物理特征的9维混合特征集，建立了Matérn 5/2核GPR定量模型。LOCO-CV结果表明：模型RMSE为0.0622 mol·L$^{-1}$（约为满量程的3.1%），$R^2$为0.9914，优于单变量回归、PLSR与SVR；GPR输出的预测不确定度能够正确反映不同浓度段的模型置信水平，为浓度测量结果的可靠性评估提供了依据。

后续工作将把本方法扩展至NaCl/$Na_2SO_4$等多组分混合盐体系，并探索将温度效应纳入特征空间的建模方案，以进一步拓宽方法的适用范围。

## 数据可用性声明

本研究采集的原始光谱数据与建模代码可通过通讯作者合理申请获取。

## 利益冲突声明

所有作者声明不存在利益冲突。

## 参考文献

[1] MCCREERY R L. Raman spectroscopy for chemical analysis[M]. New York: Wiley-Interscience, 2000.

[2] BEN MABROUK K, KAUFFMANN T H, AROUI H, et al. Raman study of cation effect on sulfate vibration modes in solid state and in aqueous solutions[J]. Journal of Raman Spectroscopy, 2013, 44(11): 1603-1608.

[3] ASHER S A, JOHNSON C R. Raman spectroscopy of a coal liquid shows that fluorescence interference is minimized with ultraviolet excitation[J]. Science, 1984, 225(4659): 311-313.

[4] WOLD S, SJÖSTRÖM M, ERIKSSON L. PLS-regression: a basic tool of chemometrics[J].


https://doi.org/10.68070/a7pqjd92

Chemometrics and Intelligent Laboratory Systems, 2001, 58(2): 109-130.
[5] SIMONE E, SALEEMI A N, NAGY Z K. Application of quantitative Raman spectroscopy for the monitoring of polymorphic transformation in crystallization processes using a good calibration practice procedure[J]. Chemical Engineering Research and Design, 2014, 92(4): 594-611.
[6] RASMUSSEN C E, WILLIAMS C K I. Gaussian processes for machine learning[M]. Cambridge: MIT Press, 2006.
[7] ZHANG Z M, CHEN S, LIANG Y Z. Baseline correction using adaptive iteratively reweighted penalized least squares[J]. Analyst, 2010, 135(5): 1138-1146.
[8] SAVITZKY A, GOLAY M J E. Smoothing and differentiation of data by simplified least squares procedures[J]. Analytical Chemistry, 1964, 36(8): 1627-1639.
[9] WALRAFEN G E. Raman spectral studies of the effects of temperature on water structure[J]. The Journal of Chemical Physics, 1967, 47(1): 114-126.
[10] SMOLA A J, SCHÖLKOPF B. A tutorial on support vector regression[J]. Statistics and Computing, 2004, 14(3): 199-222.
[11] 王凯. 261nm紫外拉曼光谱仪集成与醇类检测应用研究[D]. 徐州: 江苏师范大学, 2025.

# Concentration Inversion of Sodium Sulfate Solutions Based on Deep-Ultraviolet Raman Spectroscopy and Gaussian Process Regression

Zhang Dingfang, Wu Qi, Chen Shuang, Meng Xin

School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116

**Abstract:** To achieve rapid and non-destructive determination of sodium sulfate concentration in aqueous solutions, a concentration inversion method based on deep-ultraviolet (DUV) Raman spectroscopy and Gaussian process regression (GPR) is proposed. A 261 nm DUV Raman detection system developed by our group was employed to collect 120 spectra of pure $Na_2SO_4$ aqueous solutions at 12 concentration levels (0-2.0 mol·$L^{-1}$). A wavelength self - calibration method using the $SO_4^{2-}$ $\nu_1$ band as an internal standard was proposed, correcting a systematic scale error of about + 0.3% in the wavenumber axis; after calibration, the deviations of the sulfate vibrational bands from literature values are within 5 $cm^{-1}$. Two baseline correction methods were compared systematically: silent-region median subtraction and adaptive iteratively reweighted penalized least squares (airPLS). It was found that airPLS over-fits the baseline in strong band regions, severely distorting the band information. A 9-dimensional hybrid feature set combining data-driven principal components of the full spectrum and physically meaningful $\nu_1$ band descriptors was constructed, and GPR with a Matérn 5/2 kernel was applied. Leave-one-concentration-out cross-validation shows that the proposed method achieves an RMSE of 0.0622 mol·$L^{-1}$ and $R^2$ of 0.9914, outperforming univariate linear regression (0.0699 mol·$L^{-1}$), partial least squares regression (0.1015 mol·$L^{-1}$) and support vector regression (0.2295 mol·$L^{-1}$). Meanwhile, GPR provides predictive uncertainties, which is of practical value for confidence assessment.

**Keywords:** Raman spectroscopy; concentration inversion; Gaussian process regression; baseline correction; wavelength self-calibration; sodium sulfate