基于近红外光谱的金振口服液固含量预测模型优化及维护研究

刘乐乐[1, 2]，徐芳芳[1, 3*]，张永超[1, 3]，赵媛媛[1, 2]，刘佳丽[1, 3]，李秀梅[1, 3]，候化蕊[1, 3]，张　欣[1, 3*]

1. 中药制药过程控制与智能制造技术全国重点实验室（江苏康缘药业股份有限公司/南京中医药大学），江苏 南京 211112
2. 南京中医药大学康缘中药学院，江苏 南京　210023
3. 江苏康缘药业股份有限公司，江苏 连云港　222001

摘　要：目的　针对近红外光谱（near-infrared spectroscopy，NIRS）建模过程依赖“地毯式”试错、缺乏理论指导的问题，以金振口服液（Jinzhen Oral Liquid，JOL）固含量预测模型为研究对象，从光谱信息质量角度揭示最佳模型的优化方向，并验证其在模型维护中的应用价值。**方法**　采集 380 个样本的 NIRS 和固含量数据。经 9 种预处理方法处理后，分别使用偏最小二乘法（partial least squares，PLS）和支持向量回归（support vector regression，SVR）建立固含量的 NIRS 预测模型。创新性地引入香农熵、主成分分析（principal component analysis，PCA）以及自编码器，构建一个从信息丰富度、线性结构集中度和非线性结构可捕获性 3 个维度量化光谱信息质量的评价框架。最后通过系统分析光谱信息特性与模型性能间的关联，揭示最佳预处理方法的方向，并将此关联规律应用于新增 294 个样本时的模型维护，以筛选最佳光谱数据集。**结果**　对于谱峰宽泛重叠的 NIRS，其信息密度与信息保留率与 PLS 模型性能呈负相关。基于此关联规律，成功预测出模型维护时的最佳数据集，其建模效果（R_p^2＝0.990 9）显著优于其他数据集。**结论**　研究发现的关联规律能够有效解释预处理对模型性能的影响，为光谱模型的优化与维护提供了理论依据和指导工具，实现了从“盲目试错”到“主动改善”的转变，为建立标准化、智能化的近红外光谱模型构建与维护流程提供了新思路。

关键词：固含量预测模型；模型维护；近红外光谱；预处理；香农熵；主成分分析；自编码器

中图分类号：R283.6　　**文献标志码：**A　　**文章编号：**0253 - 2670(2026)08 - 3051 - 10

DOI: 10.7501/j.issn.0253-2670.2026.08.018

Optimization and maintenance of prediction model for solid content in Jinzhen Oral Liquid based on near-infrared spectroscopy

LIU Lele[1, 2], XU Fangfang[1, 3], ZHANG Yongchao[1, 3], ZHAO Yuanyuan[1, 2], LIU Jiali[1, 3], LI Xiumei[1, 3], HOU Huarui[1, 3], ZHANG Xin[1, 3]

1. State Key Laboratory of Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture (Jiangsu Kanion Pharmaceutical Co., Ltd., & Nanjing University of Chinese Medicine), Nanjing 211112, China
2. Kanion School of Chinese Material Medica, Nanjing University of Chinese Medicine, Nanjing 210023, China
3. Jiangsu Kangyuan Pharmaceutical Co., Ltd., Lianyungang 222001, China

Abstract: Objective　To address the issues of reliance on exhaustive “blind” trial-and-error and the lack of theoretical guidance in the modeling process of near-infrared spectroscopy (NIRS), this study used the solid content prediction model of Jinzhen Oral Liquid (JOL, 金振口服液) as a case study. It aimed to reveal the direction for optimizing the best model from the perspective of spectral information quality and verify its application value in model maintenance. **Methods**　The NIRS and solid content data of 380 samples were collected. After being processed by nine preprocessing methods, prediction models for solid content were established using partial least squares (PLS) and support vector regression (SVR), respectively. An evaluation framework was innovatively constructed by

收稿日期：2025-11-04

基金项目：国家工信部产业基础再造和制造业高质量发展专项（TC2308068）；中药制药过程控制与智能制造技术全国重点实验室开放基金课题（SKL2023D02003）

作者简介：刘乐乐（2000—），女，硕士研究生，研究方向为中药新药研发及应用研究。E-mail: lll1223390646@163.com

***通信作者：**张　欣，博士，研究方向为中药制药过程新技术。E-mail: zxtcm@126.com

徐芳芳，副主任药师，硕士生导师，从事中药智能制造研究。E-mail: 879164331@qq.com

introducing Shannon entropy, principal component analysis (PCA), and autoencoders to quantify spectral information quality from three dimensions: information richness, linear structure concentration, and non-linear structure capturability. Finally, by systematically analyzing the correlation between spectral information characteristics and model performance, the direction for the optimal preprocessing method was revealed. This correlation rule was then applied to model maintenance involving 294 newly added samples to screen for the optimal spectral dataset. **Results** It was found that for NIRS characterized by broad and overlapping peaks, both information density and information retention rate were negatively correlated with PLS model performance. Based on this correlation rule, the optimal dataset for model maintenance was successfully predicted, achieving a modeling performance (R_p^2 = 0.990 9) significantly superior to that of other datasets. **Conclusion** The correlation rules identified in this study effectively explain the impact of preprocessing on model performance. They provide a theoretical basis and guiding tools for the optimization and maintenance of spectral models, facilitating a shift from "blind trial-and-error" to "active improvement". This offers new insights for establishing a standardized and intelligent workflow for NIRS model construction and maintenance.

Key words: solid content prediction model; model maintenance; near-infrared spectroscopy; pretreatment; Shannon entropy; principal component analysis; autoencoder

近红外光谱（near-infrared spectroscopy，NIRS）凭借其快速、无损、绿色等优势，已成为中药质量过程分析与控制的重要工具[1-5]。在中药液体制剂的质量评价体系中，固含量常作为一项关键质量属性（critical quality attribute，CQA），不仅能有效反映制剂中有效成分与杂质的整体概况[6]，也被相关指导原则与产品标准列为核心控制指标[7-8]，对保障药物的安全、有效与稳定具有决定性意义。因此，发展固含量的快速、准确检测方法至关重要。

对于金振口服液（Jinzhen Oral Liquid，JOL）而言，目前基于 NIRS 的固含量预测模型已有较好研究基础[9]。然而，该技术在实际应用，尤其是面向工业生产时，仍面临一个根本性挑战：模型构建过程本身在很大程度上依赖于“地毯式”试错，缺乏理论指导[10]。具体而言，分析人员需对大量光谱预处理方法及其参数组合进行盲目尝试，并仅以预测误差最小化为准则选择最终模型[11-12]。首先，建模过程效率低，耗费大量计算与时间资源；其次，模型成功与否缺乏可解释性，知识难以积累与迁移运用；最关键的是，当生产过程出现光谱漂移时，由于不理解模型的内在规律，模型维护工作再次沦为盲目试错，严重制约了 NIRS 模型的长期稳定性与工业化应用潜力。因此，揭示最优预测模型的优化方向，并将其用于指导模型的后期维护，是本研究的核心目标。

针对最佳预测模型的优化研究，即模型可解释性的研究是当下乃至未来研究的重点方向[13]，但现有研究仍存在部分局限。一部分研究侧重于模型的事后解释，如采用局部可解释模型无关解释（local interpretable model-agnostic explanations，LIME）、沙普利加性解释（SHapley Additive ex Planations，SHAP）等方法，解释已建好模型的预测结果，但这无法反向指导建模前期的预处理策略选择[14]；另一部分研究，则尝试利用自编码器或信息熵等单一指标评价光谱信息，但未能将其与建模算法的性能关联[15]。目前尚缺乏一个从光谱数据信息质量本身出发贯穿预处理与建模全过程的系统性可解释性分析框架。

为弥补上述空白，本研究优选了香农熵、主成分分析（principal component analysis，PCA）和自编码器这 3 种算法，构建了涵盖信息丰富度、线性结构集中度及非线性结构可捕获性这 3 个维度的信息评价框架，对光谱信息进行量化表征。通过系统分析这些信息特征与多种建模算法性能之间的关联规律，揭示预处理影响模型性能的原因，为光谱模型的优化提供理论依据和可解释性框架。最终，本研究将该规律应用于指导工业场景下的模型迭代与增量维护，并以筛选最佳光谱数据集为例进行验证，为实现 NIRS 模型的“建-维-用”一体化提供新范式。

1 材料与仪器

1.1 材料

JOL 生产过程中第 1 次调碱前、第 2 次调碱前、第 2 次调碱后、灭菌后 4 个工序的 95 批样品，批号分别为 2410305～2410312、2410314、2411302～2411316、2411318～2411339、2411341～2411345、2412301～2412345，共 380 个 2024 年的样本。此外还包括 294 个 2023 年的 JOL 样本，具体批号为 230403、230407、230411、230415、230501、230504、230513、230514、230516、230521、230524、230530、

230532、230533、230536、230539、230542、230544、230548、230556、230560、230564、230568、230572、230576、230581、230584、230588、230592、230601、230605、230613、230617、230641、230645、230647、230649、230653、230657、230675、230677、230680～230697、230510、230811～230826。以上 JOL 样品均由江苏康缘药业股份有限公司提供。

1.2 仪器

Mettler Toledo ME104E 型电子天平，梅特勒-托利多仪器（上海）有限公司；KQ-500DE 型数控超声波清洗器，昆山市超声仪器有限公司；Antaris II 型傅里叶近红外变换光谱仪，赛默飞世尔科技（中国）有限公司；DHG-9145A 型电热鼓风干燥箱，上海一恒科学仪器有限公司。

2 方法与结果

2.1 技术路线图

为方便理解全文框架，绘制技术路线图，结果如图 1 所示。

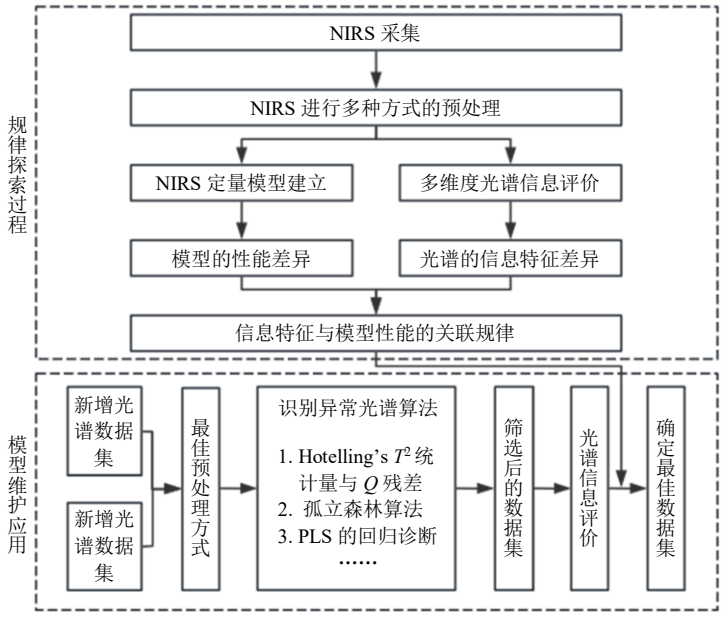


图 1 技术路线图

Fig. 1 Technology roadmap

2.2 固含量的测定

参照文献方法[9]，精密称取 5 g 样品溶液，置已干燥至恒定质量（X_0）的蒸发皿中，称定质量（X_1），置水浴上蒸干后，在 105 ℃条件下烘至恒定质量，移置干燥器中，冷却 30 min，迅速称定质量（X_2），计算固含量。

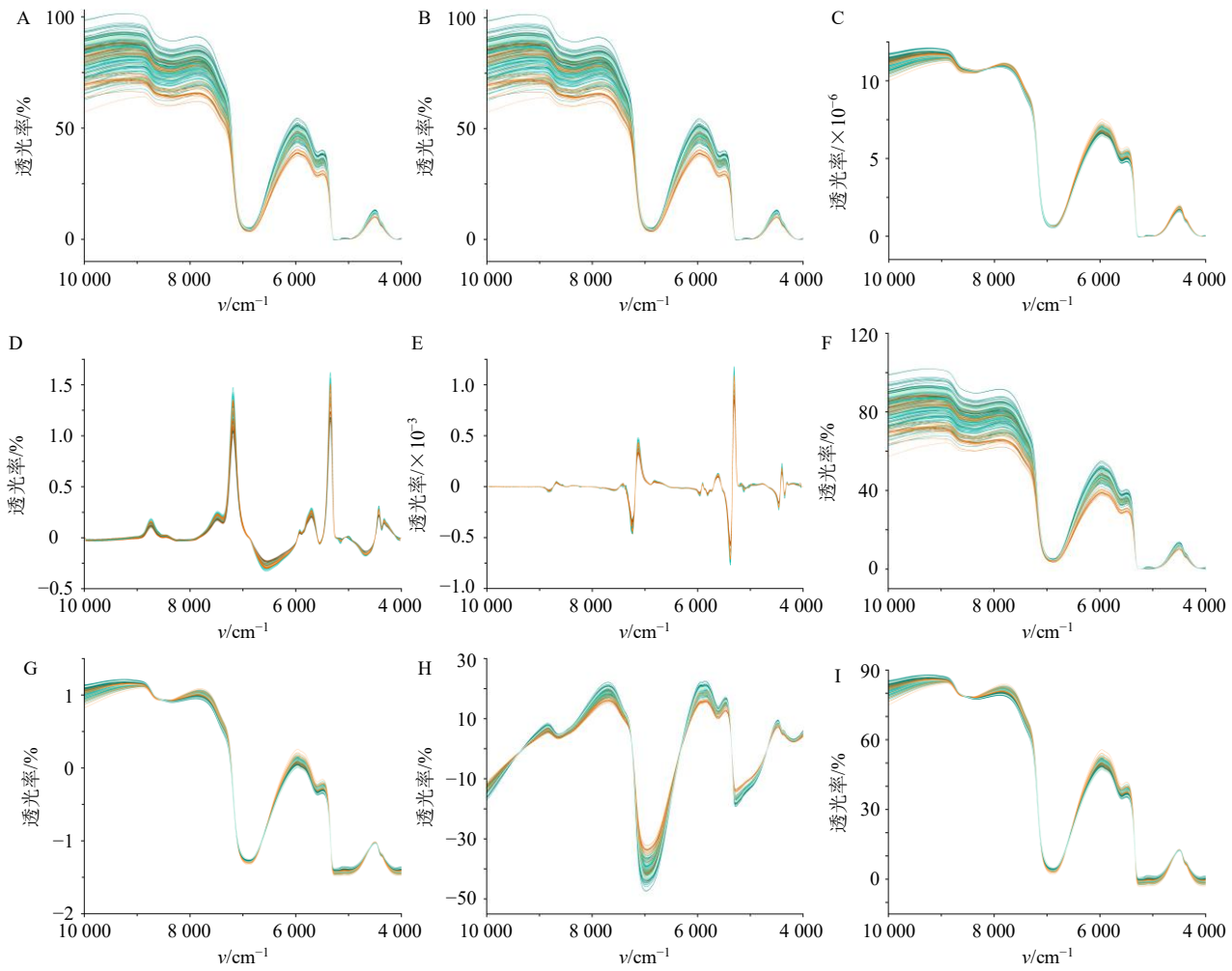
$$固含量=(X_2-X_0)/(X_1-X_0) \quad (1)$$

2.3 NIRS 的采集

取适量样品置于 1 mm 比色皿内，以空气为背景，扫描范围 4 000～10 000 cm^{-1}，扫描次数 32 次，分辨率 8 cm^{-1}，衰减器为 C 模块，增益为 1 倍。

2.4 NIRS 的预处理

NIRS 的采集会受到外界环境的影响，预处理可消除其他因素对数据信息的影响，将 NIRS 转化成最佳拟合条件，从而提高模型的性能[10,16]。本实验采用 SG 平滑、归一化、SG 一阶导数（Savitzky-Golay first derivative，SG 1st-Der）、SG 二阶导数（Savitzky-Golay second derivative，SG 2nd-Der）、基线校正、标准正态变量变换（standard normal variate，SNV）、去趋势、多元散射校正（multiplicative scatter correction，MSC）对 NIRS 进行预处理，结果见图 2。图 2 直观展示了 9 种预处理方法对 NIRS 的形态影响。如图 2-A（无预处理）所示，未经过预处理的原始 NIRS，在纵轴方向上呈现出显著的离散度，表明样本间存在较大的光程差异，且 NIRS 包含大量由颗粒散射和仪器响应引起的基线漂移与噪声干扰。对比观察发现，图 2-B（SG 平滑）、图 2-F（基线校正）保留了原始光谱的整体轮廓。其中，SG 平滑主要用于去除高频随机噪声，提升信噪比，但未改变光谱的波形特征；基线校正虽然在一定程度上将光谱拉回统一基准，修正了低透光率区域的偏差，但样本间的纵向分布差异依然存在。相比之下，图 2-C（归一化）、图 2-G（标准正态变量变换）、图 2-H（去趋势）和图 2-I（多元散射校正）的处理效



A-无预处理；B-SG 平滑；C-归一化；D-SG 1st-Der；E-SG 2nd-Der；F-基线校正；G-SNV；H-去趋势；I-MSC。

A-no pretreatment; B-SG smoothing; C-normalization; D-SG 1st derivative; E-SG 2nd derivative; F-baseline correction; G-SNV; H-detrending; I-MSC.

图 2　NIRS 的预处理

Fig. 2　NIRS pretreatment

果更为显著。经这些方法处理后，原本离散的光谱曲线高度收敛，样本间的重叠度大幅增加。这说明上述方法有效消除了由于样本颗粒不均匀和光散射效应引起的物理干扰，使光谱差异更能反映样本化学成分的真实变化。形态变化最为剧烈的是图 2-D（SG 一阶导数）和图 2-E（SG 二阶导数），导数处理彻底改变了光谱的原始形态，将原本宽大重叠的吸收峰转化为尖锐的峰谷，且使基线回归至零刻度附近。这一过程不仅消除了与波长无关的基线平移和线性倾斜，更实现了重叠峰的有效分离，显著提高了光谱的分辨率，突出了微弱的特征信息。

2.5　数据处理

采用 Unscrambler（挪威 Camo Analytics 公司）软件进行光谱预处理、建立 PLS 和 SVR 模型；采用 Python 软件进行样本集划分以及香农熵、PCA 和自编码器 3 种算法评价；采用 Origin 学生版进行所有绘图。

2.6　光谱的信息评价框架

为系统性、多维度地量化不同预处理方法对光谱信息特性的影响，本研究优选了香农熵、PCA 和自编码器这 3 种算法，创新性地构建了一个包含信息丰富度、线性结构与非线性冗余度 3 个维度的信息评价框架。该框架整合了信息论、线性代数与非线性神经网络中的经典算法，旨在全方位表征光谱的信息质量。

2.6.1　香农熵算法　香农熵源于信息论，是衡量系统不确定性或信息量的基本指标[17]。在本研究中，它被用作一个模型无关的宏观指标，以评估光谱数据集整体的信息丰富度与复杂度。本研究将每个波长下所有样品的吸光度（A）视作离散变量，基于直

方图法估算分布概率，据式（2）～（4）求取各波长熵值并取平均，反映整体信息丰富度。

$$H(X)=-\sum_{i=1}^{n} p(x_i)/\log_2 p(x_i) \quad (2)$$

$$p(x_i)=f(x_i)\times d(x_i) \quad (3)$$

$$n=1+3.322\times \lg N \quad (4)$$

$H(X)$是离散化后随机变量 X 的熵，$p(x_i)$是随机变量 X 在第 i 个区间的概率，其值为概率密度与区间宽度的乘积；$f(x_i)$为概率密度；$d(x_i)$为区间 i 的宽度；n 是离散化后区间的总数，依据斯特格斯规则（Sturges' rule），样本量 N 为 380 时，n 取值为 10

熵值越高，表明光谱信号的分布越分散、包含的变化和细节越多，意味着信息更丰富，但也可能暗示存在离群样本，需要结合另外 2 个指标综合分析。香农熵提供了一个全局性的光谱信息复杂度基准，帮助评估预处理对光谱整体信息量的影响。具体而言，该指标为理解预处理是在“增强细节”（如导数处理，通常会增加熵）还是在“平滑去噪”（如平滑处理，通常会降低熵）提供了直接的量化依据。

2.6.2 PCA 算法 PCA 是一种经典的线性降维技术，通过最大化方差来寻找全局的主成分方向，捕捉整个数据集的线性分布结构，常用来实现数据可视化及特征提取[18]。然而，本研究的创新之处在于，并非将其用于降维，而是基于其结果独创性地提出了信息密度指标，用以评估光谱信息在线性空间中的集中程度。

本研究通过对经 Z-score 标准化处理后的数据矩阵进行奇异值分解来执行 PCA，统计各预处理方法下累积解释 95%方差所需的主成分数，计算信息密度，其计算公式如下。

$$\text{信息密度}=1-95\%\text{方差所需主成分数}/\text{总主成分数} \quad (5)$$

高信息密度意味着 NIRS 的主要变异越能被少数几个线性主成分所概括，数据的共线性结构越强，数据的主要变异是线性的、信息集中度高，这为判断数据是否适合 PLS 这类线性模型提供了关键指标。

2.6.3 自编码器算法 自编码器属于一种无监督神经网络，通过非线性变换将数据压缩后再重建，来捕获数据的内在结构和模式[19]。本研究构建了对称、基于多层感知机（multilayer perceptron，MLP）的网络拓扑结构，输入层和输出层的节点数为光谱的总波长点数（M），2 个隐藏层的节点数为 2/3 M，瓶颈层的节点数为 1/3 M，激活函数采用 ReLU，优化器为 Adam。本研究通过平均重建误差（average reconstruction error，ARE）衡量原始与重建数据误差，信息保留率（information retention rate，IRR）来评估光谱信号的非线性结构强度和可捕获性，揭示了 PCA 无法有效衡量的非线性结构特征。

$$\text{ARE}=\sum_{i=1}^{N}[\sum_{j=1}^{D}(X_{i,j}-\bar{X}_{i,j})^2/D]/N \quad (6)$$

$$\text{IRR}=1-\text{ARE} \quad (7)$$

N 是样本数量，D 是特征维度，$X_{i,j}$ 是第 i 个样本的第 j 个特征的原始值，$\bar{X}_{i,j}$ 是第 i 个样本的第 j 个特征的重建值

与传统应用不同，本研究的创新点在于采用了一个结构简单的浅层自编码器。其目的并非追求完美的特征提取，而是利用其有限的学习能力来探测数据的内在结构复杂程度。当原始 NIRS 数据结构相对简单，主要由平滑背景组成，简单的自编码器模型就能轻易地学习并重建它，从而信息保留率较高；当原始 NIRS 数据经过预处理后（如导数运算），其内在的非线性结构变得更加复杂、细节更丰富时，简单自编码器模型难以完全学习并重构数据，从而得到较低的信息保留率。

2.7 信息评价结果

针对 NIRS 信息特征，采用 3 种信息评价算法分析不同方法预处理后的 NIRS，具体评价结果见表 1。从预处理后 NIRS 信息评价的结果来看，数据集的平均熵值大多在 2.3～2.4，经导数处理后提升至 2.5～2.6，NIRS 的信息丰富度提升，表明经导数处理后放大了 NIRS 的变化趋势，数据中包含的细节变化变多；多数 NIRS 信息密度＞0.98，信息在低维空间高度集中，仅需 3～4 个主成分即可解释 95%的方差，说明数据适合 PLS 这类线性模型；信息保留率普遍高于 95%，意味着神经网络能够高效

表 1 NIRS 的信息评价结果

Table 1 Information evaluation results of NIRS

预处理方法	香农熵平均熵值	PCA		自编码器	
		主成分数	信息密度	平均重建误差	信息保留率/%
无预处理	2.363 0	4	0.989 5	0.027 4	97.26
SG 平滑	2.362 5	4	0.989 5	0.025 1	97.49
归一化	2.499 6	4	0.989 5	0.012 6	98.74
SG 1st-Der	2.545 1	4	0.989 5	0.042 2	95.74
SG 2nd-Der	2.671 1	22	0.942 1	0.031 9	96.77
基线校正	2.333 4	4	0.989 5	0.014 3	98.57
SNV	2.420 8	3	0.992 1	0.005 4	99.46
去趋势	2.494 8	3	0.992 1	0.005 2	99.48
MSC	2.414 7	3	0.992 1	0.005 4	99.46

还原有效信息，数据的内在结构简单，可能含有较多的冗余信息。

2.8 固含量预测模型的建立

2.8.1 校正集与验证集的划分 对于采集到的 4 个工序共 380 个样本，分别采用光谱-理化值共生（sample set partitioning based on joint *x*-*y* distance，SPXY）算法，将样本按照 4∶1 的比例，划分为校正集和验证集。使用 SPXY 算法兼顾参考值与光谱距离，确保划分后的样本集的光谱和参考值都覆盖较大范围且均匀分布[20]。其结果见表 2。

2.8.2 PLS 与 SVR 模型的建立 本研究以固含量为因变量，以不同预处理后 NIRS 为自变量，分别采用 PLS 和 SVR 建立固含量预测模型。PLS 通过最大化自变量与因变量的协方差，提取潜在变量来构建线性模型[21]；SVR 通过核函数将数据映射到高维空间，建立非线性模型[22]。本研究中 PLS 选择前 10 个主成分中预测均方根误差（root mean square error of prediction，RMSEP）最低的主成分数为最佳的潜变量数（latent variables，LVs）；SVR 选择径向基函数作为核函数，以交叉验证均方根误差（cross-validated root mean squared error，RMSECV）最小化为优化目标，通过网格搜索确定惩罚系数 C 和核函数参数 γ。本研究选择校正集决定系数（R_c^2）、验证集决定系数（R_p^2）、校正均方根误差（root mean square error of calibration，RMSEC）、RMSEP、预测标准差（standard error of prediction，SEP）、以及预测相对误差（relative standard error of prediction，RSEP）作为模型准确度的评价指标。R_c^2 和 R_p^2 越接近于 1，RMSEC、RMSEP、SEP 和 RSEP 越小，表明模型的预测准确度越高。由表 3 可知，在 PLS 模型中，导数处理后的模型性能最好，SVR 模型中无预处理的模型性能最好。整体来看，PLS 模型的 R_p^2 全都大于 0.9，SVR 模型的 R_p^2 则更容易受到预处理方法的影响。

表 2 NIRS 样本集划分结果

Table 2 NIRS sample set partitioning results

样本集	样本数	固含量/%			
		最大值	最小值	平均值	中位数
校正集	304	24.69	18.43	20.52	19.18
验证集	76	24.48	18.71	19.14	19.07

表 3 不同预处理方法对 NIRS 的 PLS、SVR 模型性能的影响

Table 3 Effects of different pretreatment methods on performance of PLS, SVR models for NIRS

预处理方法	PLS 模型						SVR 模型					
	LVs	RMSEC/%	RMSEP/%	SEP/%	R_c^2	R_p^2	RSEP/%	RMSEC/%	RMSEP/%	SEP/%	R_c^2	R_p^2	RSEP/%
无预处理	8	0.002 7	0.001 7	0.001 7	0.955 9	0.918 3	0.90	0.002 0	0.001 5	0.001 5	0.991 9	0.948 0	0.76
SG 平滑	8	0.002 7	0.001 7	0.001 7	0.985 3	0.927 8	0.89	0.002 0	0.001 5	0.001 5	0.991 8	0.947 7	0.76
归一化	7	0.002 6	0.001 8	0.001 8	0.986 4	0.921 7	0.93	0.002 1	0.001 7	0.001 7	0.991 1	0.930 6	0.87
SG 1st-Der	9	0.002 4	0.001 6	0.001 6	0.988 5	0.938 1	0.83	0.002 2	0.001 7	0.001 5	0.990 1	0.929 5	0.88
SG 2nd-Der	9	0.002 3	0.001 6	0.001 6	0.989 3	0.933 5	0.86	0.002 3	0.002 3	0.002 0	0.989 2	0.870 0	1.20
基线校正	9	0.002 6	0.001 7	0.001 7	0.986 2	0.926 2	0.90	0.002 1	0.001 8	0.001 8	0.991 0	0.916 1	0.96
SNV	8	0.002 5	0.001 8	0.001 8	0.987 2	0.916 4	0.96	0.002 1	0.001 8	0.001 8	0.990 8	0.919 7	0.94
去趋势	8	0.002 7	0.001 9	0.001 9	0.985 7	0.915 3	0.97	0.002 2	0.002 6	0.002 2	0.990 1	0.835 3	1.35
MSC	8	0.002 6	0.001 9	0.001 9	0.987 0	0.909 2	1.00	0.002 2	0.001 8	0.001 8	0.990 8	0.918 3	0.95

2.9 NIRS 的信息特性与模型性能的关联规律探究

以无预处理的数据为基准，计算 9 种光谱预处理后的光谱信息评价结果和建模结果的相对偏移量。光谱信息评价结果选取平均熵值、信息密度和信息保留率 3 个为指标，建模结果选取 RMSEP、SEP、R_p^2、RSEP 这 4 个指标。通过计算相对偏移量，消除不同参数间的量纲差距，同一参数的相对大小关系保持不变。根据相对偏移量采用 Spearman 相关系数绘制热力图，其结果如图 3 所示。由相关性分析图可知，仅有 PLS 模型与信息评价指标之间存在显著相关性。在 NIRS 的 PLS 建模中，信息密度、信息保留率与 RMSEP 呈正相关，即信息密度越高，保留的主成分数越少，信息保留率越高，模型的预测误差越大。

2.10 基于关联规律的模型长期维护应用

本研究发掘的 NIRS 信息评价与模型性能的关联规律，除了能为光谱预处理方法的选择与优化提供理论依据和可解释性框架，还可以指导实验员在模型维护过程中确定最佳光谱数据集，避免异常样本剔除再次陷入盲目试错。“垃圾进垃圾出”，数据

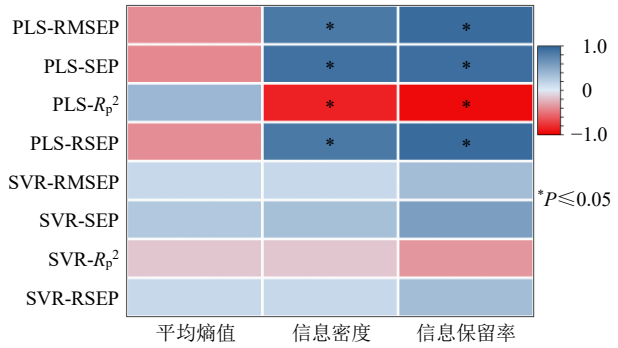


图 3　NIRS 算法评价结果与建模结果的相关性分析

Fig. 3　Correlation analysis between algorithm evaluation results and modeling results of NIRS

质量会直接影响模型结果，不加挑选直接接融合数据可能会引入噪声和降低模型稳健性[23]。以最常用的 NIRS 为例，本研究在现有 380 个样品的基础上，融合 294 个 2023 年 JOL 样品的固含量以及 NIRS 数据，进行模型维护，流程如下。

（1）整合数据：按照现有模型的最佳预处理方法，对 674 条光谱进行预处理。

（2）对预处理后的光谱分别用 3 种常见的方法进行异常光谱筛选：方法 1 基于 PCA 技术，采用 Hotelling's T^2 统计量与 Q 残差（squared prediction error）进行联合筛选[24]，置信区间为 95%；方法 2 采用孤立森林（isolation forest）算法[25]，异常率设为 10%；方法 3 基于 PLS 的回归诊断，分别杠杆值和学生化残差进行异常光谱识别[26]，杠杆值的阈值为 $3(A+1)/n$，其中 A 为主成分数，n 为样本数，学生化残差的置信区间为 95%。

（3）对异常光谱筛选后的 3 种光谱数据进行光谱信息评价，结果见表 4。

表 4　3 种异常光谱筛选后 NIRS 的信息评价结果

Table 4　Evaluation results of NIRS information after screening three kinds of abnormal spectra

数据集	样本量	平均熵值	主成分数	信息密度	信息保留率/%
方法 1	616	2.700 4	6	0.990 3	98.57
方法 2	606	2.801 3	6	0.990 1	97.92
方法 3	642	2.582 1	6	0.990 7	87.82

（4）基于上文的光谱信息评价与模型性能的关联规律和表 4 对 3 种异常光谱筛选后的建模结果进行预测。

（5）建立模型验证预测结果，3 个数据集样本的划分结果见表 5，模型验证的结果见表 6，校正集与验证集的实测值-预测值相关图见图 4。

表 5　3 个数据集的样本划分结果

Table 5　Sample partitioning results for three datasets

数据集	样本集	样本数	固含量/%			
			最大值	最小值	平均值	中位数
方法 1	校正集	493	0.336 0	0.131 3	0.216 6	0.227 1
	验证集	123	0.245 8	0.185 9	0.199 7	0.191 1
方法 2	校正集	485	0.336 0	0.131 3	0.215 5	0.224 3
	验证集	121	0.245 8	0.185 9	0.200 4	0.191 1
方法 3	校正集	514	0.273 6	0.165 0	0.216 2	0.227 2
	验证集	128	0.245 8	0.185 9	0.199 4	0.191 1

表 6　3 种异常光谱筛选后 NIRS 的建模结果

Table 6　Modeling results of NIRS after screening of three abnormal spectra

数据集	LVs	RMSEC/%	R_c^2	RMSEP/%	R_p^2
方法 1	9	0.009 848	0.863 7	0.002 298	0.984 7
方法 2	8	0.009 958	0.863 5	0.002 015	0.989 0
方法 3	9	0.002 880	0.986 8	0.001 754	0.990 9

由表 4 可知，3 个数据集的信息密度相差不大，主要是平均熵值与信息保留率的区别。根据“2.9”项发现的信息保留率，与模型性能呈负相关这一核心规律，预测信息保留率最低的数据集（方法 3）将产生最佳模型性能。

方法 1 和 2 数据集的平均熵值超过了 NIRS 常规范围（2.33～2.67），表明数据集的信息变异性较高，但值得注意的是这 2 个数据集的信息保留率较高，表明数据结构简单，数据容易被捕获重构，与数据存在较多的变化信息相违背，推测这 2 个数据集中存在少量离群样本，造成高变异性高熵值，但同时整体光谱结构相对简单，较易被学习重建。

由表 6 可知，建模结果验证了该预测，方法 3 数据集的模型性能显著优于其他两者，成功证明了 NIRS 的信息特性与建模效果的关联规律，可以应用于模型维护时最佳光谱数据集的确定，指导模型优化的方向。此外从异常样本预测结果来看，方法 1 和 2 的数据集，出现了校正集 R^2 小于验证集的反常现象，结合图 4 中方法 1 和方法 2 数据集的实测值-预测值相关图，图中存在几个明显偏离对角线的离群点，证明了这 2 个数据集的校正集中存在异常样本。

3　讨论

3.1　固含量预测模型光谱信息-模型性能关联规律的阐释与意义

本研究最核心的发现是揭示了 NIRS 信息特性

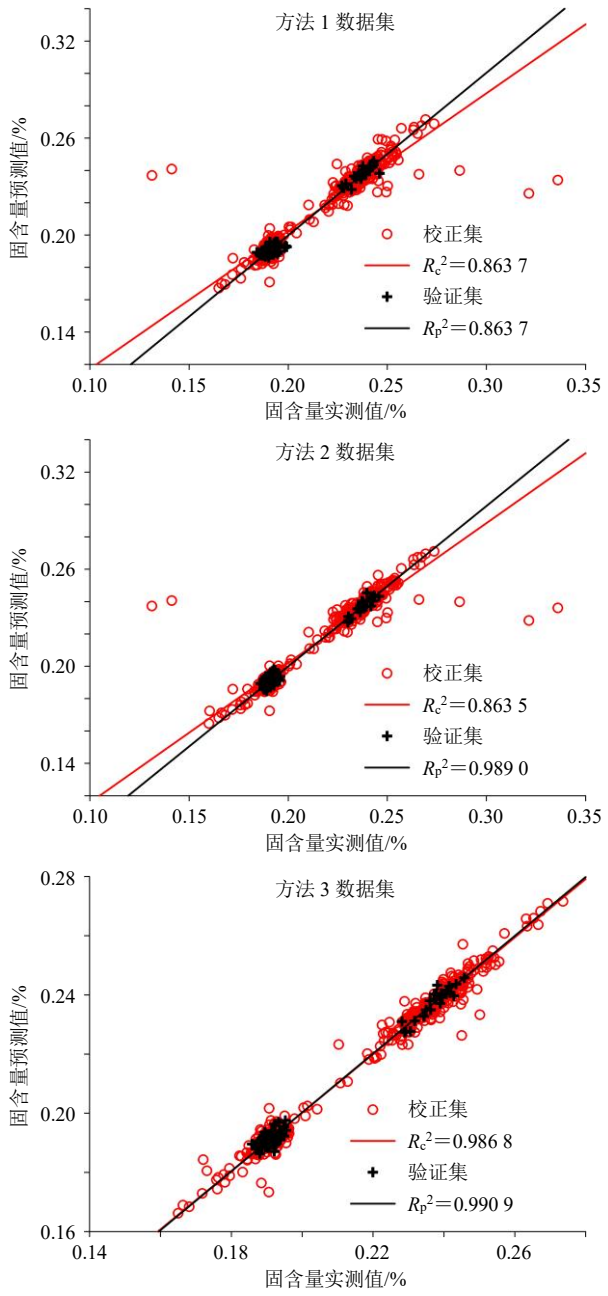


图 4　3 个数据集的实测值-预测值相关图

Fig. 4　Scatter plots of measured *vs* predicted values for three datasets

与 PLS 模型性能之间的关联规律。实验结果表明，信息密度和信息保留率与模型预测误差呈正相关。这一发现与直观认知可能相悖，却深刻揭示了 NIRS 建模的本质。NIRS 的有效信息一般表现在光谱的纵向差异上，同时由于 NIRS 是倍频和合频吸收，谱带非常宽且重叠严重，这些差异通常为样品中高含量成分所贡献。

当检测样品为液体时，宽大的水峰往往会掩盖微量溶质（如糖分、蛋白、活性成分）的精细特征峰[27]。中药口服液作为复杂体系，其背景吸收中的水峰和光散射（基线平移或倾斜）构成了主要的宏观变化。这些宏观变化在数据结构上较为简单，更容易被学习重建（高信息保留率），用更少的主成分数就能解释 95%的方差（高信息密度）。经过建模算法的挖掘后，仍有部分有效信息被掩藏在这种宏观变化之下，模型性能也因此受到影响，均方根误差较高。所以，需要对 NIRS 进行预处理，一是为了减少基线漂移；二是为了增强微弱峰或重叠峰的分辨率，发掘更多的有效信息。

对比其他预处理方式，导数处理在消除了加性或乘性基线漂移的同时，还增强微弱峰或重叠峰的分辨率，突出峰形变化[10,28]，使得与固含量化学键振动直接相关的微弱信号特征得以显现。这种处理虽然在统计学上增加了数据的复杂度和无序度，导致信息密度和信息保留率下降，但从化学计量学角度看，它有效提升了有效信息的占比，从而提升了模型性能。因此，本研究观察到的统计指标变化，本质上是光谱预处理剔除物理干扰、凸显化学特征的过程在信息论层面的数值投射。

本研究首次从光谱信息质量的角度解释了：为什么通过这种预处理方式建立的模型性能最佳。通过建立了“光谱信息指标”与“模型性能”的映射关系，那么在未来的模型维护或类似体系建模中，不论是采用何种预处理方式，或是同种预处理方式中多个参数的人工确定，还是不同的预处理方式的组合，都可以计算信息评价参数这种成本极低的无监督指标，即可快速预判哪种预处理方式或哪个子数据集（如文中筛选出的方法 3）最可能产生最优模型。从而实现了从“基于后验误差（RMSEP）的被动选择”向“基于光谱源头信息质量的主动筛选”的转变，从而显著降低了全流程的试错成本，尤其在模型后期大量新增样本进行维护时，建立模型的速度易受到样本量的限制。

3.2　关联规律在模型维护中的成功应用与价值

本研究并未止步于规律阐释，更进一步将其应用于工业实践中至关重要的模型维护环节。在整合新增 294 个数据时，传统异常样本剔除方法往往依赖统计阈值，缺乏明确导向，不能得到较为一致的结论。本研究基于信息保留率与模型性能呈反比的规律，成功预测并验证了“方法 3”数据集能产生最优模型（R_P^2=0.990 9）。方法 1 和 2 数据集虽信息变异性高，但高信息保留率暗示其数据结构简

单、可能存在未被有效识别的异常样本，这与它们建模时出现校正集 R^2 低于验证集的异常现象相互印证。

本研究的价值在于，它证实了从数据源头质量来指导建模的可行性，将模型维护从被动的、滞后的“定期验证-发现问题-重新试错”循环，转变为主动的、前瞻性的“基于规律-预测最优-精准验证”的新范式，显著提升了维护效率与模型长期稳健性。

3.3 研究框架的普适性、局限性

本研究以 JOL 为例验证了该框架的有效性，虽然不同检品的光谱特性存在差异，但本框架所采用的香农熵、信息密度和信息保留率 3 个评价维度，源于数据内在的统计学和结构特性，独立于具体样本的物理化学性质。因此，该评价体系有望推广至其他中药复杂体系的 NIRS 定量分析任务中，为不同应用场景下的模型维护提供理论参考。

值得注意的是，该框架与 SVR 模型性能未显显著相关。推测原因可能为，SVR 的模型性能高度依赖于通过核函数映射到高维特征空间后，少数位于回归管道边界的支持向量的局部几何分布。同时，其 ε-不敏感损失函数使其对管道内部的噪声具有天然的鲁棒性。然而，本框架的指标（如信息熵、PCA 主成分）均在原始空间对数据的全局结构进行评估，因此，难以预见数据在 SVR 模型中的最终表现。简言之，本框架衡量的光谱数据“全局性、宏观”统计特性，与 SVR 模型“局部化、基于边界”的学习机制存在根本性不匹配。这一局限性恰恰指明了未来研究方向：若要指导非线性模型优化，可能需要发展能评估数据局部几何结构的新指标，从而完善本框架。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 陈珊. 近红外光谱在中成药生产过程中质量控制的应用研究 [D]. 广州: 华南理工大学, 2022.

[2] 安思宇, 张磊, 岳洪水, 等. 基于近红外光谱的中药质量一致性控制研究进展 [J]. 中南药学, 2019, 17(9): 1439-1445.

[3] 陈方方, 厉奔, 王飞, 等. 近红外光谱用于药物生产中离子液体监测 [J]. 光学学报: 网络版, 2025, 2(15): 59-71.

[4] 段立鸣. 近红外光谱技术在我国中药研究中的应用现状 [J]. 实用医药杂志, 2008, 25(7): 874-875.

[5] 李国沼, 陈莘雨, 高建平, 等. 基于文献计量学的近红外光谱技术在中药质量控制领域的研究热点与趋势分析 [J]. 药物评价研究, 2025, 48(8): 2327-2338.

[6] 张永超, 刘佳丽, 李执栋, 等. 基于近红外光谱法和折光率法的热毒宁注射液金银花提取和浓缩工序中间体总固体量快速检测研究 [J]. 中草药, 2025, 56(5): 1587-1595.

[7] 王磊, 杨越, 李贞瑞, 等. 热毒宁注射液金银花提取浓缩工段过程性能指数研究 [J]. 中草药, 2017, 48(14): 2864-2869.

[8] 国家食品药品监督管理局. 国家食品药品监督管理局关于做好中药注射剂安全性再评价工作的通知 (国食药监办 [2009] 359 号) [EB/OL]. (2009-07-16) [2025-11-04]. https://law.pharmnet.com.cn/laws/detail_1973.html.

[9] 李秀梅, 徐芳芳, 张欣, 等. 基于近红外光谱和中红外光谱技术的金振口服液中间体含量预测模型研究 [J]. 中草药, 2023, 54(24): 8007-8017.

[10] 褚小立, 袁洪福, 陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用 [J]. 化学进展, 2004, 16(4): 528-542.

[11] 吴春艳, 杜文俊, 张伟东, 等. 基于近红外光谱技术的摩罗丹水提液浓缩过程的多指标快速检测 [J]. 中国现代应用药学, 2022, 39(1): 87-92.

[12] 童枫, 徐芳芳, 闫逸伦, 等. 热毒宁注射液金银花和青蒿 (金青) 萃取过程中固形物含量近红外光谱在线监测模型的建立及萃取终点判断研究 [J]. 中草药, 2024, 55(19): 6555-6565.

[13] 陈瀑, 杨健, 褚小立, 等. 近五年我国近红外光谱分析技术的研究与应用进展 [J]. 分析化学, 2024, 52(9): 1213-1224.

[14] Wang Y, Yao Q X, Zhang Q H, *et al*. Explainable radionuclide identification algorithm based on the convolutional neural network and class activation mapping [J]. *Nucl Eng Technol*, 2022, 54(12): 4684-4692.

[15] Tsimpouris E, Tsakiridis N L, Theocharis J B. Using autoencoders to compress soil VNIR-SWIR spectra for more robust prediction of soil properties [J]. *Geoderma*, 2021, 393: 114967.

[16] Morais C L M, Lima K M G, Singh M, *et al*. Tutorial: Multivariate classification for vibrational spectroscopy in biological samples [J]. *Nat Protoc*, 2020, 15(7): 2143-2162.

[17] 许可. 信息的度量问题概述 [J]. 硅谷, 2008(14): 44.

[18] 孙平安, 王备战. 机器学习中的 PCA 降维方法研究及其应用 [J]. 湖南工业大学学报, 2019, 33(1): 73-78.

[19] 袁非牛, 章琳, 史劲亭, 等. 自编码神经网络理论及应用综述 [J]. 计算机学报, 2019, 42(1): 203-230.

[20] 王世芳, 韩平, 崔广禄, 等. SPXY 算法的西瓜可溶性固形物近红外光谱检测 [J]. 光谱学与光谱分析, 2019, 39(3): 738-742.

[21] 唐敏, 李霄龙, 李嘉琪, 等. 基于近红外光谱和化学计量学的蒲黄炭快速判别和定量分析方法研究 [J]. 世界科学技术—中医药现代化, 2024, 26(9): 2385-2398.

[22] 褚小立, 许育鹏, 陆婉珍. 用于近红外光谱分析的化学计量学方法研究与应用进展 [J]. 分析化学, 2008, 36(5): 702-709.

[23] 韩志军. 浅谈建模与仿真过程中数据校验问题 [J]. 科学技术创新, 2018(25): 88-89.

[24] 肖枝洪, 冉小华. 运用主成分分析法的过程控制和诊断 [J]. 重庆理工大学学报: 自然科学, 2014, 28(1): 96-101.

[25] 徐东, 王岩俊, 孟宇龙, 等. 基于 Isolation Forest 改进的数据异常检测方法 [J]. 计算机科学, 2018, 45(10): 155-159.

[26] 李锋霞, 黄勇, 李强. 光谱检测哈密瓜品质中异常样本的综合分析 [J]. 中国瓜菜, 2023, 36(7): 18-23.

[27] Pasquini C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications [J]. *J Braz Chem Soc*, 2003, 14(2): 198-219.

[28] 严衍禄, 陈斌, 朱大洲, 等. 近红外光谱分析的原理、技术与应用 [M]. 北京: 中国轻工业出版社, 2013: 83-91.

[责任编辑 郑礼胜]