doi：10.12452/j.fxcsxb.26013103

基于混合高斯分解的近红外光谱奇异样本识别策略

朱远哲[1]，赵忠盖[1*]，李由然[2]，刘　飞[1]

（1. 江南大学　轻工过程先进控制教育部重点实验室，江苏　无锡　214122；2. 江南大学　粮食发酵与食品生物制造国家工程研究中心，江苏　无锡　214122）

摘　要：近红外光谱在定量分析中易受奇异样本影响。光谱由含氢基团吸收峰叠加形成，传统方法将整条光谱整体分析，仅依赖全局特征，难以识别特定峰处的细微奇异特征，且其固定阈值也缺乏自适应性。为此，该文提出基于混合高斯分解的奇异样本识别策略，在关键吸收峰内建立混合高斯模型，用多个高斯分量解析重叠谱峰，再对各分量幅值构建基于四分位距准则的正常范围，将超出范围的样本判定为奇异样本。将所提方法应用于柠檬酸发酵生产原料段混液的总糖浓度预测，结果表明该方法可有效识别人为构建和实际光谱中的奇异样本。与原始光谱相比，剔除奇异样本后构建的PLS模型，RMSE降低34.69%，R^2提高26.28%。该方法通过解析吸收峰组成结构，实现了对局部奇异特征的自适应检测，为近红外光谱质量提升提供了一种结构清晰、可解释性强的途径。

关键词：近红外光谱；奇异样本识别；混合高斯模型；重叠峰分解；柠檬酸发酵

中图分类号：O657.3；TQ921.1　　**文献标识码：**A　　**文章编号：**1004-4957(2026)06-0001-10

Gaussian Mixture Model-based Outlier Detection for Near-infrared Spectra

ZHU Yuan-zhe[1]，ZHAO Zhong-gai[1*]，LI You-ran[2]，LIU Fei[1]

（1. Key Laboratory for Advanced Process Control of Light Industry of Ministry of Education，Jiangnan University，Wuxi　214122，China；2. National Engineering Research Center of Cereal Fermentation and Food Biomanufacturing，Jiangnan University，Wuxi　214122，China）

Abstract：Near-infrared(NIR) spectroscopy is susceptible to interference from outliers in quantitative analysis. Since NIR spectra are formed by the superposition of absorption peaks from hydrogen-containing groups，traditional methods analyze the entire spectrum as a whole，relying solely on global features. This makes it difficult to detect subtle outlier features within specific peaks，and the use of fixed thresholds lacks adaptability. To address this，this paper proposed an outlier detection strategy based on Gaussian mixture decomposition. Specifically，a Gaussian mixture model was established within key absorption peak regions to resolve overlapping spectral peaks using multiple Gaussian components. A normal range was then constructed for the amplitude of each component based on the interquartile range criterion，and samples exceeding this range were identified as outliers. The proposed method was applied to the prediction of total sugar concentration in the raw-material mixture of citric acid fermentation. The results showed that the method cauld effectively identify both artificially introduced and naturally occurring spectral outliers. Compared with the original spectra，the PLS model built after removing the outliers achieves a 34.69% reduction in RMSE and a 26.28% increase in the coefficient of determination. By analyzing the compositional structure of absorption peaks，this method enables adaptive detection of local outlier features，providing a structured and highly interpretable approach for enhancing NIR spectral data quality.

Key words：near-infrared spectroscopy；outlier detection；Gaussian mixture model；overlapping peak resolution；citric acid fermentation

近红外光谱位于可见光与中红外之间，波长范围为780~2 526 nm[1]。该谱段与有机物中含氢基团（如O—H、N—H、C—H）振动的合频及倍频吸收区高度吻合[2]，因此样品的近红外光谱可直接反映其

收稿日期：2026-01-31；**修回日期：**2026-04-03
基金项目：国家自然科学基金项目(62473175)
*通讯作者：赵忠盖，博士，教授，研究方向：间歇过程统计监控、软测量与状态估计，E-mail：gaizihao@jiangnan.edu.cn

含氢官能团的组成与结构信息，是一种高效、无损的间接分析手段[3]。近红外光谱分析具有操作简便、快速及适用于原位检测等优势，广泛应用于农产品品质评价[4-5]、食品成分分析[6-7]、工业生产过程监控[8-9]等领域的定性与定量分析。

近红外光谱定量分析依赖由大量样本建立的校正模型，其性能直接受光谱数据质量影响。然而，实际检测中奇异样本难以避免，主要来源包括仪器端的光源不稳、探头污染[10]，环境端的温湿度骤变、杂散光干扰，以及样本端的制样不均、粒径异常或标签标注错误等[11]。这些奇异样本会通过高杠杆效应扭曲回归方向，其吸收峰异常、基线漂移等伪特征也会干扰有效信息提取，并且容易导致模型局部过拟合，严重降低模型的泛化能力与预测稳定性[12]。

根据其原理与实现机制，现有奇异样本检测方法可分为以下几类。第一类方法依据光谱的空间分布特征进行判别，主要利用样本在光谱空间中的位置或相似度关系。马氏距离(MD)通过考虑变量间的相关性和数据离散程度，衡量样本光谱与光谱整体分布的偏离程度[13]；光谱残差分析(SR)通过构建反映光谱“正常趋势”的模型，计算原始光谱与该模型的差异，以识别奇异形态[14]。然而，该类方法在高维光谱中易受维度影响，其判别阈值通常依赖经验设定，自适应性差，制约了其在复杂光谱数据中的稳健应用。

第二类方法在已构建的预测模型的基础上，识别对模型稳定性造成过度影响的样本。常用的Cook距离(Cook′s D)，通过计算剔除样本所引起的模型参数变化量，评估其对模型结构稳定性的影响[15]；基于交叉验证的检验方法通过系统评估样本在不同训练子集下的预测残差稳定性，直接识别导致模型泛化性显著下降的样本[16-17]。该类方法高度依赖所建模型的准确性与稳健性，若模型本身存在偏误、设定不当或过拟合，易导致奇异样本判别的误报或漏检。

第三类方法借助机器学习，从数据结构与密度分布的角度实现奇异检测。局部离群因子(LOF)通过比较样本与邻域的密度差异识别低密度区域的离群点[18]；孤立森林(iForest)通过随机构建二叉树并度量样本被“隔离”的路径长度，识别分布稀疏的奇异点[19]；也有研究采用自编码器等深度学习方法，通过学习光谱的正常分布模式检测奇异[20]。此类方法虽能数据驱动地捕捉复杂非线性奇异特征，但是对参数设置敏感，且识别出的奇异模式多为抽象数学表征，难以直接关联样品的物理化学状态，在需结合光谱化学先验知识的实际应用中存在解释瓶颈。

现有奇异样本检测方法虽有效，但仍面临共性挑战：多数方法侧重基于整体光谱指标，对源自分子基团振动变化的局部特征响应不足，易漏检由局部污染或组分微变引起的异常。近红外光谱本质上是多个吸收子峰叠加形成的复合谱带，奇异样本往往首先表现为局部子峰结构的差异，而非整体光谱均匀变化[21]。此外，依赖固定阈值或经验参数的方式，难以兼顾不同光谱数据的判别敏感性与稳健性。因此，发展一种能够结合光谱形成机理、专注子峰奇异特征且具有自适应判别能力的检测方法，对提升近红外光谱模型性能具有重要意义。

为实现对子峰结构的精确刻画，需要采用能够解析重叠谱峰并提取结构参数的建模方法。混合高斯模型(GMM)源于有限混合分布理论，其系统理论由Geoffrey McLachlan等学者完善[22]。该模型假设观测数据由多个高斯分量加权叠加生成，每个分量对应独立的峰位、幅值和峰宽参数，适用于多峰或重叠信号的建模分析，并已广泛应用于复杂信号处理与分类识别领域[23-24]。近红外光谱是由多个子峰叠加形成的复合信号，采用分量化建模可以解析每个子峰的幅值和形态特征，从而更精细地捕捉样本间局部差异。混合高斯模型的分量化思想与这一特性高度契合：各分量用于描述谱带的局部结构特征，通过分析分量幅值及形态差异，可揭示样本间局部子峰结构变化，为奇异样本识别提供结构依据。相比直接基于高维光谱的全局统计分析，分量参数表示不仅可降低数据维度，还增强了异常识别的可解释性。因此，引入混合高斯分解能够刻画近红外谱带的局部子峰特征，弥补传统全局方法对局部异常不敏感的不足，为提高定量模型稳健性提供理论基础。

基于上述理论背景，本文提出一种基于混合高斯分解的光谱奇异样本识别方法。该方法通过混合高斯模型解析重叠吸收峰，得到对应特定基团的子峰分量；进而对各分量幅值计算四分位距并设定IQR准则下的正常波动区间，若某样本在任一高斯分量的幅值超出该区间，则判为奇异样本并予以剔除，基于此完成奇异样本的高效识别与数据质量的精准控制。该方法基于光谱叠加机制，从子峰特征

出发进行判定，可有效克服全局方法对局部奇异特征不敏感的问题，为提升近红外定量模型的稳健性与预测精度提供了数据质量保障。

1　近红外光谱结构特性与奇异样本识别问题分析

近红外光谱主要来源于样品中含氢基团振动的倍频与合频吸收，其光谱通常表现为多个吸收子峰叠加形成的宽而平滑的复合谱带。这种复合结构一方面源于样品成分复杂，同一含氢基团在不同化学微环境下的振动频率存在差异；另一方面，由于倍频与合频谱带本身较宽且吸收位置接近，不同基团的子峰容易发生重叠与融合，如图 1 中的实线所示。在图 1 中当局部峰强度发生扰动时（蓝色和绿色虚线表示局部峰变化示意），相关子峰的吸收特征会产生明显差异，而叠加后的整体复合光谱仅表现为较小幅度的变化（黑色虚线表示变化后的整体光谱）。因此，当样品发生异常变化时，其影响往往体现在特定波数区间内吸收峰的形变、位移或强度变化等局部结构差异，而整体统计特征的变化可能并不显著。

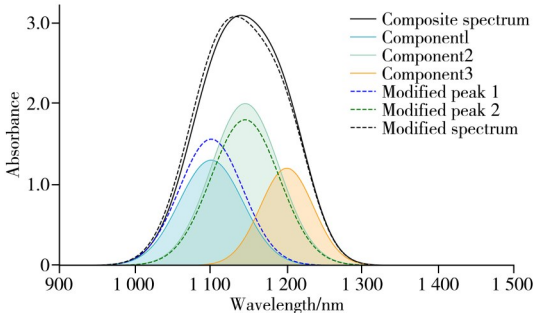

图 1　宽近红外吸收峰的分解示意图

Fig. 1　Decomposition of broad near-infrared absorption bands

然而，现有奇异样本识别方法多基于整体光谱特征构建，侧重于样本在全谱空间中的分布偏离，其判别边界通常建立在全谱特征空间分布假设之上。当异常仅表现为局部子峰的微小形变、位移或强度波动时，这类全局统计量对局部结构扰动的响应易被其他稳定波段所稀释，从而降低检测灵敏度，甚至导致漏检。同时，基于协方差结构的距离度量易受异常样本影响，使整体分布估计产生偏移，削弱判别结果的稳定性与可靠性。因此，在峰形高度重叠且局部变化与整体偏移并存的复杂近红外光谱数据中，仅依赖全谱统计量难以准确反映谱峰结构层面的细微差异。基于光谱形成机理，从子峰结构稳定性出发开展局部解析，有助于构建更加符合光谱机理特性的异常识别框架。

2　基于混合高斯分解的奇异样本识别方法

2.1　理论基础

近红外光谱源于分子振动的倍频与合频跃迁，其吸收谱带是大量微观振动跃迁在凝聚态环境（如液体或固体体系）中叠加后的宏观表现。在凝聚态体系中，由局域微环境差异引起的非均匀展宽通常占主导[25]，使跃迁频率在统计意义上呈现近似正态分布特征，因此近红外吸收峰整体上可近似表示为高斯型函数[26]。

然而，在实际测量中，吸收峰线型可能受到多种实验因素干扰而偏离理想高斯分布：光谱仪有限的分辨率会引入仪器函数卷积效应，使子峰发生畸变；样品中的悬浮颗粒（如发酵液中的微生物菌体）会引起显著的米氏散射，导致基线抬升并引发谱峰非对称或拖尾；此外，基质效应通过增强局域微环境差异而扩大振动频率分布的离散性，使谱线可能出现展宽增强或非对称等复杂形态。

为尽量避免上述干扰并保证观测谱线接近高斯分布，本研究在仪器选择、样品处理及数据预处理方面采取了严格控制措施。实验使用高分辨率傅里叶变换近红外（FT-NIR）光谱仪，其较高的信噪比及优化的切趾函数处理显著降低了背景噪声影响，并可减小仪器函数引起的线型卷积效应[27]，使其在建模中可忽略。在样品处理上，充分搅拌样品以保证体系均质，控制温度恒定以减少氢键网络波动引起的峰位漂移，并统一光程条件以降低光学路径差异对谱线的影响。同时，通过多元散射校正（MSC）预处理，削弱悬浮颗粒散射及基质差异引起的加性与乘性偏差[28]，从而确保观测谱线主要反映化学组分的本征吸收特性。

本研究关注的近红外光谱波长范围为 900~1 700 nm，且样品为液体体系，谱线主要体现非均匀展宽效应，因此观测到的谱线能够保持接近高斯形态。基于这一机制，混合高斯模型能够合理表征子峰峰形态，其参数在统计意义上可保持稳定，从而提高模型对谱带结构表征的稳定性与一致性。

2.2 光谱混合高斯模型拟合

针对显著吸收峰区间的谱线结构，采用混合高斯模型进行分量化建模，其数学表达式如式(1)所示。

$$f(x,\theta)=\sum_{k=1}^{K}\frac{w_k}{\sqrt{2\pi{\sigma_k}^2}}\exp\left(-\frac{\left(x-\mu_k\right)^2}{2{\sigma_k}^2}\right) \tag{1}$$

其中，x为数据采集点；μ_k为第k个分量的均值，表示高斯分量的中心位置；${\sigma_k}^2$为第k个分量的方差，表示高斯分量的宽度；w_k为第k个分量的混合权重，决定第k个高斯分量在整体拟合曲线中的相对贡献大小；K为模型中高斯分量的总个数；θ为包含各高斯分量权重、均值和标准差的参数集合，即$\theta=\{w_k,\ \mu_k,\ \sigma_k|k=1,\ 2,\ \cdots,\ K\}$。

近红外光谱中，同一官能团的吸收峰位置与峰宽由分子振动固有属性决定。在同一检测过程中，由于样本的化学组成基本一致，同一官能团所处的化学微环境可视为相对固定，因此其吸收峰位置与峰宽在不同样本间保持稳定。基于这一前提，表征同一化学键的高斯分量应具有一致的均值与方差，以维持物理解释的一致性。各高斯分量的权重w_k根据其在光谱中的实际吸光度贡献进行设定，以直接反映不同样本间吸收峰的强度差异，从而在保持光谱结构稳定的同时，更精确地表征样本的定量变化特征。

混合高斯模型拟合曲线的效果采用均方误差(MSE)来衡量，MSE值越小，代表拟合曲线与原始曲线之间的偏差越小，吻合度越高。其数学表达式如式(2)所示。

$$\mathrm{MSE}=\frac{1}{n}\sum_{j=1}^{n}\left(y_j-f_j(x,\theta)\right)^2,\theta\in\Omega \tag{2}$$

其中，y_j为第j个波长处的真实吸光度值，$f_j(x,\ \theta)$为第j个波长处的模型拟合吸光度值，n为波长总个数，Ω为参数的取值范围。将全部m个样本的MSE累加，构建如式(3)所示的目标函数。

$$\mathrm{Minimize}\ J(\theta)=\sum_{i=1}^{m}\mathrm{MSE}_i \tag{3}$$

模型参数估计以最小化目标函数实现最优拟合。针对近红外光谱数据的高维特点，为避免传统优化方法陷入局部最优的问题，本研究采用带边界约束的有限内存BFGS算法(L-BFGS-B)进行稳健参数估计。该算法作为有限记忆拟牛顿法的扩展形式，适用于含边界约束的大规模优化问题。它通过迭代更新近似Hessian逆矩阵，并利用投影梯度与投影线搜索确保迭代点始终处于可行域内，具有内存占用低、收敛效率高及边界处理稳定的优势，尤其适用于光谱高维参数空间的优化求解。其迭代更新过程如式(4)所示。

$$\theta_{p+1}=P\left[\theta_p-\alpha_pH_p\nabla J\left(\theta_p\right)\right] \tag{4}$$

其中，θ_p为第p步参数向量，$\nabla J\left(\theta_p\right)$为目标函数梯度，$H_p$为逆Hessian近似矩阵，$\alpha_p$为经投影线搜索确定的步长，$P[\cdot]$表示投影至可行域内的算子。式(4)所示的更新机制从两方面保障了算法的性能：首先，投影算子$P[\cdot]$将每次迭代的参数θ_{p+1}约束在可行域内，避免因步长过大导致参数越界或目标函数发散，从而确保收敛稳定性；其次，算法在迭代过程中仅通过存储少量历史梯度差向量来隐式构造搜索方向$H_p\nabla J\left(\theta_p\right)$，无需显式计算和存储完整的Hessian矩阵，显著降低了高维光谱数据下的计算开销，进而提升求解效率。

混合高斯模型中高斯分量的数量通过贝叶斯信息准则(BIC)确定。BIC在最大似然估计基础上引入与参数个数成比例的惩罚项，从而在模型复杂度和拟合效果间取得平衡，其计算公式如式(5)所示。

$$\mathrm{BIC}=k\cdot\ln(n)-2\ln\left(\hat{L}\right) \tag{5}$$

其中，k为模型参数个数，n为区间内的波长数，$\hat{L}$为模型似然函数的最大值。在实际应用中，对于混合高斯模型，通过比较不同分量数k对应的BIC值，选取使BIC最小的k作为最优分量数，从而在模型拟合精度与模型复杂度之间取得平衡。高斯分量的数量并不对应样品中的实际化学组分数，而是用于描述对光谱曲线影响较为显著的若干子峰结构。若分量数过少，模型可能无法充分刻画谱带的局部结

构特征；而分量数过多则会增加参数数量并带来过拟合风险。因此，基于BIC的模型选择获得合理的谱带分解结果，可为后续奇异样本识别提供稳定的结构参数基础。

2.3　奇异样本识别策略

完成混合高斯拟合后，为识别由样本特性、测量噪声或吸收机制变化引起的局部奇异特征，本方法基于各高斯分量幅值在样本间的分布，通过量化其偏移程度来判定奇异状态，实现样本质量评估。

设经高斯混合模型拟合后，第i个样本获得第k个分量的权重为w_{ik}、均值为μ_k，方差为σ_k^2，其对应的高斯分量幅值高度可表达为式(6)，该幅值高度反映了吸收分量的相对强度。

$$a_{ik}=\frac{w_{ik}}{\sqrt{2\pi\sigma_k^2}} \tag{6}$$

对于全体样本的第k个分量，提取其对应的幅值高度序列$\boldsymbol{a}_k=\left[a_{1k},\ a_{2k},\ \cdots,\ a_{mk}\right]^T$，经降序排列后得到该序列的第一四分位数$Q_{1k}$和第三四分位数$Q_{3k}$，使用式(7)计算出该高斯分量处的四分位距$\mathrm{IQR}_k$。

$$\mathrm{IQR}_k=Q_{3k}-Q_{1k} \tag{7}$$

设定正常幅值区间为$[Q_{1k}-3\mathrm{IQR}_k,\ Q_{3k}+3\mathrm{IQR}_k]$，超出该区间的即为奇异值，通过式(8)计算出第$i$个样本在第$k$个分量处的偏离量。

$$d_{ik}=\begin{cases}\dfrac{\left(Q_{1k}-3\mathrm{IQR}_k\right)-a_{ik}}{\mathrm{IQR}_k+\varepsilon}, & a_{ik}<Q_{1k}-3\mathrm{IQR}_k\\[2ex] \dfrac{a_{ik}-\left(Q_{3k}+3\mathrm{IQR}_k\right)}{\mathrm{IQR}_k+\varepsilon}, & a_{ik}>Q_{3k}+3\mathrm{IQR}_k\\[2ex] 0, & \text{else}\end{cases} \tag{8}$$

其中，$\varepsilon=10^{-9}$为数值稳定项。

由于每个高斯分量均对应特定基团的吸收响应，任一吸收特征的显著偏离意味着该样本在对应组成和状态上已偏离正常范畴，破坏了光谱与性质之间应有的物理关联；同时，由于光谱特征间存在较强的共线性，局部光谱奇异特征极易通过特征间的相关结构对整体模型参数产生干扰，进而影响模型的预测稳健性与泛化性能。因此，为同时保障模型的可解释性与可靠性，若某样本在任一高斯分量上的偏离量大于0，即$d_{ik}>0$，则将该样本视为奇异样本并予以剔除。

3　结果与讨论

3.1　数据来源与预处理

实验选用采集自某工厂的柠檬酸发酵生产原料段混液样本，涵盖不同生产日期与批次。为降低人为操作及环境因素带来的随机误差，每个样品均进行两次重复取样，并取其均值作为最终数据，经过上述操作共得到123份样本数据。样品的关键指标总糖(TS)含量通过高效液相色谱仪进行测定。近红外光谱数据采用德国Bruker公司生产的MPA Ⅱ型近红外光谱仪采集，仪器分辨率为16 $\mathrm{cm^{-1}}$，扫描次数为64次，采样频率为10 kHz，光谱波长范围为900~1 700 nm。原始光谱数据通过OPUS 8.0软件读取与导出，后续处理及分析均在Python 3.12环境中完成。原始近红外光谱如图2所示。

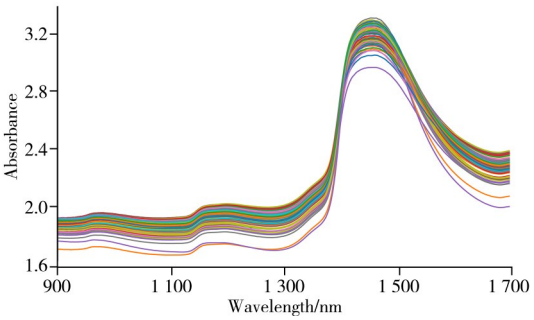

图2　柠檬酸发酵混液的原始光谱

Fig. 2　Original near-infrared spectra of the citric acid fermentation broth

为降低样品散射及基质效应并提高信噪比，本研究采用Savitzky-Golay平滑结合多元散射校正(MSC)对图2中的原始近红外光谱进行预处理。SG平滑可削弱高频噪声，MSC则校正因样品粒度、光程或散射引起的基线偏移，使光谱主要反映化学组分吸收特征。处理后光谱如图3所示，在不改变光谱形状的前提下，基线干扰明显降低，曲线平滑度提高。

为聚焦信息密度高且化学意义明确的谱段，选取特征吸收峰区域作为关键分析区间。通过局部极

值分析确定了3个吸收峰，峰位分别为972、1 195、1 455 nm，见图中黑色竖线；结合吸收峰的半高宽计算左右波长边界，最终确定特征吸收峰实际覆盖区间为938~1 022 nm、1 149~1 250 nm和1 389~1 591 nm，具体划分如图3中红色虚线框所示。这些波段的选择不仅基于数据驱动方法，同时考虑了总糖分子振动特性。总糖分子通常包含羟基(—OH)、甲基/亚甲基(—CH_3/–CH_2—)、糖苷键(—C—O—C—)以及环氧/羰基邻近C—H等基团。本文检测到的3个吸收峰区间分别对应总糖中羟基(—OH)的第一泛音振动、羟基(—OH)与甲基/亚甲基(—CH_3/—CH_2—)的组合振动，以及羟基(—OH)的基本伸缩振动。

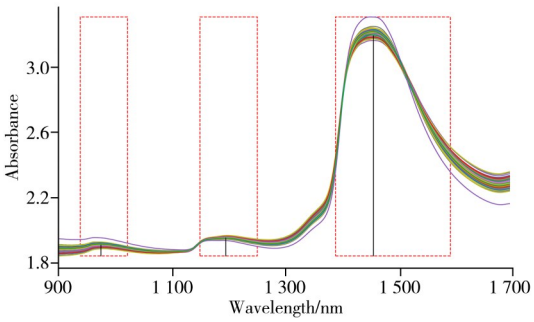

图3 预处理后的柠檬酸发酵混液光谱及特征吸收峰区间

Fig. 3 Preprocessed spectra of citric acid fermentation broth and characteristic absorption peak regions

3.2 奇异样本构建

为验证所提出方法对局部谱带异常的识别能力，需先构建基准光谱集合。为筛选具有代表性的样本，本研究采用主成分分析(PCA)进行降维，并利用前两个主成分构建得分空间。以得分空间中心为参考，计算各样本到中心的欧氏距离并进行排序，从中选取距离中心最近的60条光谱作为谱线最为集中的基准光谱集合。图4展示了所有光谱在前两个主成分得分空间中的分布，其中黄色点对应的光谱即为该基准集合。

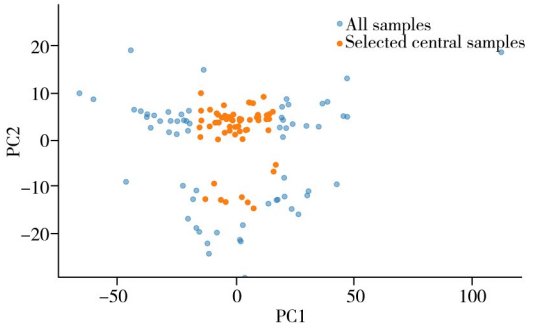

图4 所有光谱的PCA得分及基准光谱集合

Fig. 4 PCA scores of all spectra and the reference spectral set

在获得基准光谱集合后，基于混合高斯模型分解得到的子峰参数，通过人为施加扰动构建奇异样本。选取8、13、25、36、42、50号样本，设计峰位偏移、峰宽变化和峰强变化3种扰动类型，并分别通过调整高斯峰的中心位置、方差参数和权重系数实现，图5展示了3种异常类型的示意图，其中红色虚线表示原始高斯分量，蓝色虚线表示施加异常后的高斯分量，黑色实线表示原始光谱线，蓝色实线表示施加异常后的光谱线。为覆盖不同程度的情况，每类奇异样本设置两种强度，6条光谱的具体参数设置列于表1。

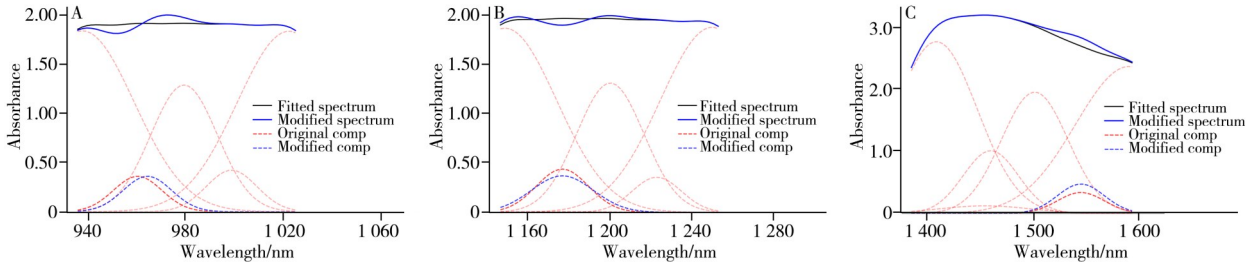

图5 构建的3类奇异样本光谱及高斯分量变化

Fig. 5 Spectra and Gaussian component variations of three types of constructed outliers

A：peak shift；B：width change；C：amplitude change

表1 构建的奇异样本参数设置

Table 1 Parameters of artificially generated outliers

Sample index	Band index	Comp index	Outlier type	Outlier magnitude
8	0	0	Peak shift	Shifted left by 2 sampling points
13	0	1	Peak shift	Shifted right by 5 sampling points
25	1	2	Width change	Decreased by 5%
36	1	3	Width change	Increased by 20%
42	2	4	Amplitude change	Increased by 5%
50	2	5	Amplitude change	Decreased by 20%

3.3　方法有效性验证

在图3所示的3个吸收峰区间内，建立混合高斯模型进行分解。各区间中高斯分量的个数依据BIC确定，图6A、C、E给出了不同分量数对应的BIC变化结果，B、D、F给出了最优模型的拟合效果。根据BIC最小原则，在938~1 022 nm区间选择5个高斯分量，在1 149~1 250 nm区间选择6个高斯分量，在1 389~1 591 nm区间选择6个高斯分量，由此得到的混合高斯模型能够较好地描述各吸收峰区间的谱带结构。

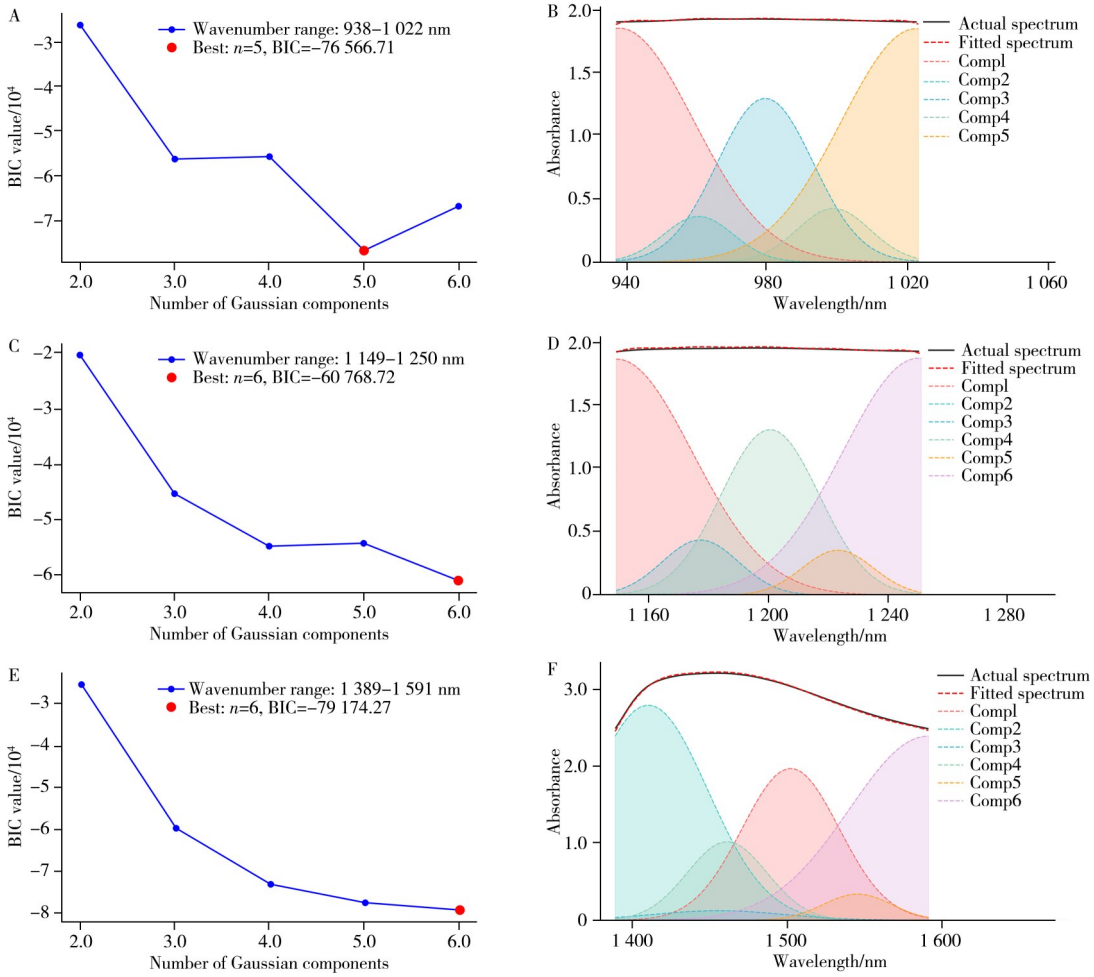

图6　吸收峰区间拟合结果

Fig. 6　Fitting results of the absorption peak region

left：Gaussian component selection results；right：Corresponding fitting results

基于模型拟合结果，对各高斯分量的幅值分布进行奇异分析。采用IQR准则对各样本在不同分量处的偏移值进行检验，其分布如图7箱线图所示，其中横坐标对应3个吸收峰区间内所有高斯分量按波长从低到高排列的序号，共17个。从图中可见，共有12个高斯分量的样本数据超出IQR触须范围，这些样本点被识别为潜在奇异样本。

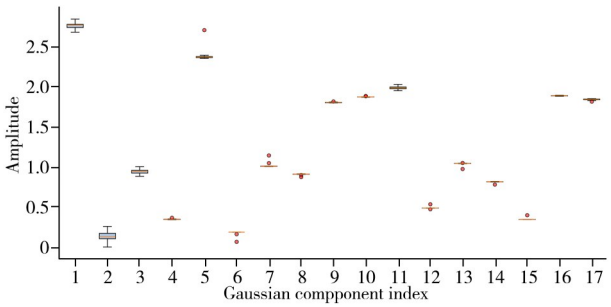

图7　各高斯分量的IQR分布箱线图

Fig. 7　Boxplot of IQR distribution for Gaussian components

在此基础上，进一步统计各样本的累计偏移量。如图8所示，累计偏移量大于0的样本共有6个。根据统计结果，判定奇异样本为8、13、25、36、42和50，该识别结果与实验中人为构建的奇异样本完全一致。其中，13、36和50号样本的累计偏移量明显高于设置为相同奇异类型的其他样本，这也对应了给予这三者更大异常幅度的设计，说明

所提出的方法能够准确有效地识别光谱中的异常样本。

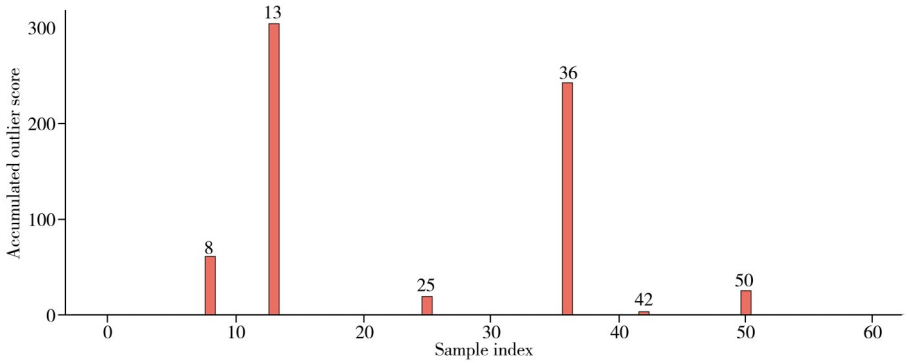

图8 各样本累计偏移量结果图

Fig. 8 Cumulative deviation results for each sample

3.4 实际光谱异常检测

为验证所提出基于混合高斯分解的奇异样本检测方法在实际应用中的有效性，本研究将采集的全部123份柠檬酸发酵混液光谱数据纳入分析。尽管实验过程中已通过样品处理与光谱预处理尽量减少仪器误差与散射干扰，但在实际采集条件下仍可能受到样品混合均匀性、微生物悬浮颗粒分布及环境因素等影响，从而导致个别样本在局部吸收峰区域出现峰位或吸收强度异常。在实际光谱分析中，选取“3.3”确定的3个特征吸收峰区间(938~1 022 nm、1 149~1 250 nm和1 389~1 591 nm)，对应的最优高斯分量数分别为5、6和6，基于拟合得到的混合高斯模型进行分析。随后计算各分量参数在样本间的偏移量，结合IQR准则与累计偏移量识别潜在异常样本，实现对自然光谱异常的检测。

表2汇总了在各高斯分量上幅值超出正常范围的样本情况。图9展示了所有样本的累计偏移量，其中累计偏移量大于0的样本共7个，据此判定的奇异样本索引为0、1、10、14、16、80、104。

表2 各高斯分量超限样本列表

Table 2 List of outliers for each Gaussian component

Wavelength range	Gaussian component	Outlier
938~1 022 nm	1，4，5，6	104
	2，3	0，1，10，104
1 389~1 591 nm	2，3	14，16，104
	5，6	80，104

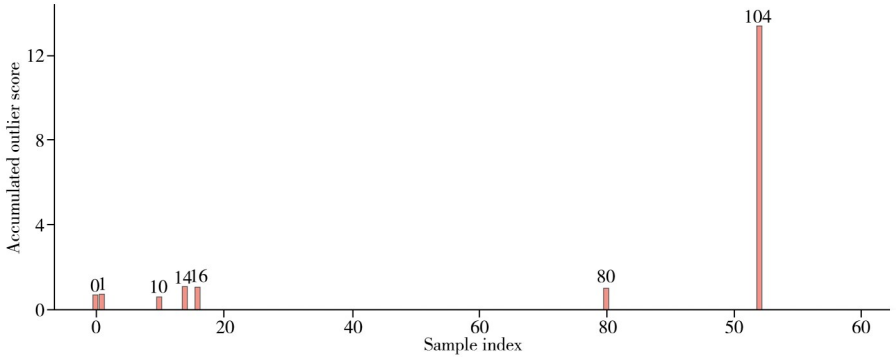

图9 各样本奇异偏移量累计结果图

Fig. 9 Cumulative outlier deviation results for each sample

剔除这些奇异样本后的光谱如图10所示，其中红色谱线表示识别出的奇异样本，蓝色谱线表示剩余样本。观察发现，剔除之后剩余样本光谱形态更趋一致，被识别出的奇异样本在特征吸收峰区域表现出不同程度的谱带偏移：样本104在吸收强度上异常明显，而其余样本在整体光谱形态上与正常样本接近，难以通过直接观察判别，但在高斯分量参数层面仍显示显著偏离。该结果表明，所提出的方法不仅能够识别明显异常样本，还能有效发现与正常谱线混合的潜在结构异常，提高光谱数据质量与分析可靠性。

3.5　方法对比

为验证本文所提方法的有效性，选取马氏距离、Cook 距离、蒙特卡洛交叉验证(MCCV)与孤立森林 4 种主流奇异检测方法进行对比。原始光谱统一预处理后，分别采用不同异常样本识别方法进行样本筛选，并基于筛选后的数据建立偏最小二乘(PLS)定量模型。由于样本量为 123 个，为充分利用数据并减少单次划分带来的偏差，采用 5 折交叉验证评估模型性能：每一折中约 4/5 样本用于建模，1/5 样本用于验证。以交叉验证预测结果的均方根误差(RMSE)与决定系数(R^2)作为评价指标，比较不同方法对模型预测性能的影响，结果见表 3。

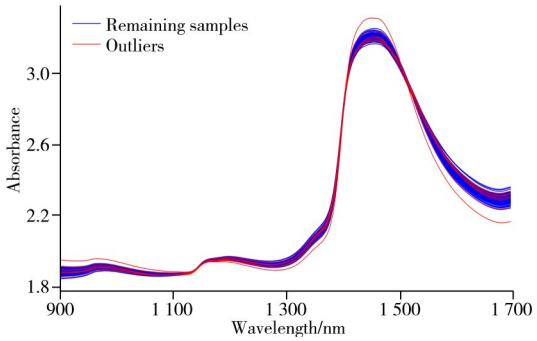

图 10　剔除奇异样本后的柠檬酸发酵混液光谱

Fig. 10　Spectra of citric acid fermentation broth after outlier removal

从表 3 可知，相比现有方法，本文提出的基于混合高斯分解的奇异识别策略有效提升了预测精度。采用该方法筛选后的样本构建 PLS 模型，其 RMSE 由 0. 576 3 降至 0. 376 4，相对降幅为 34. 69%；R^2 由 0. 661 1 提升至 0. 834 8，相对增幅达 26. 28%。图 11A~F 直观对比了不同奇异识别方法处理后 PLS 模型的预测效果：与原始光谱建模相比，剔除奇异样本均能提升模型性能，表现为预测值与实际值的相关性增强、泛化能力提高，其中基于混合高斯分解的策略所带来的提升最为显著。

表 3　不同奇异样本剔除方法效果对比

Table 3　Performance comparison of different outlier detection methods

Outlier detection method	Number of anomalous samples	Number of remaining samples	RMSE	R^2
Original	0	123	0. 576 3	0. 661 1
MD	5	118	0. 464 1	0. 775 9
Cook′s D	4	119	0. 483 9	0. 750 8
MCCV	11	112	0. 392 9	0. 808 3
iForest	7	116	0. 478 0	0. 766 6
GMM	7	116	0. 376 4	0. 834 8

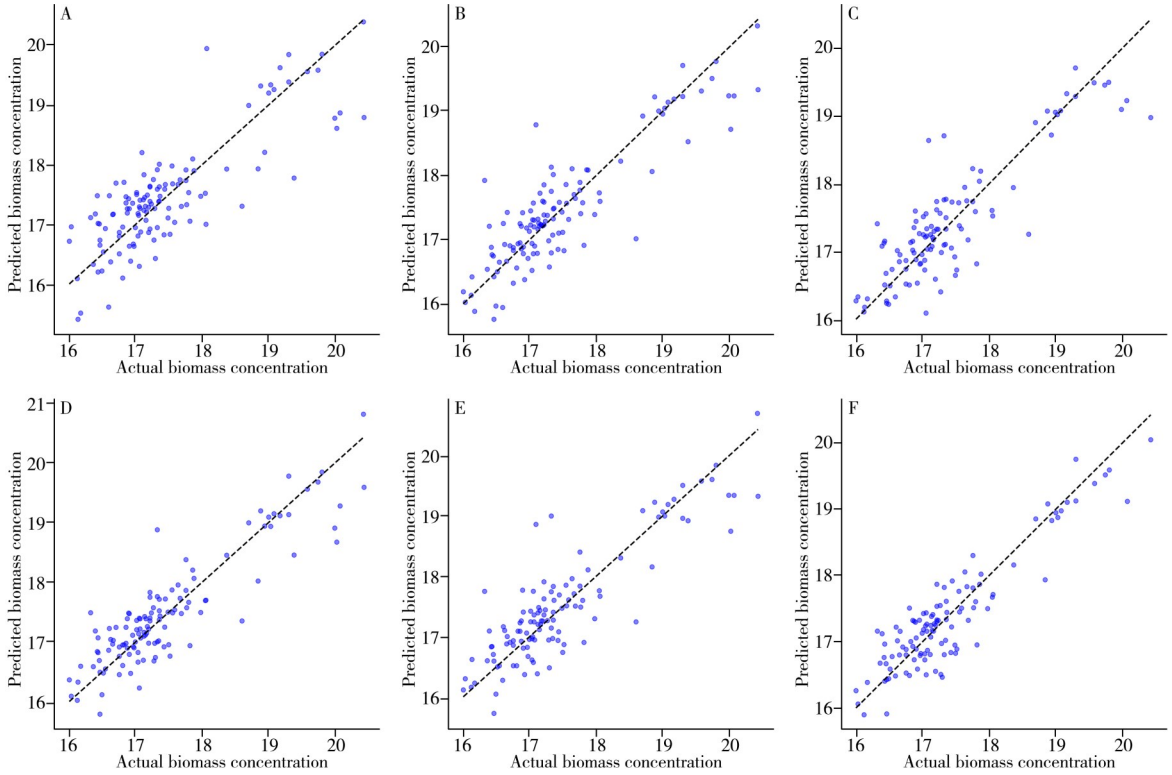

图 11　不同奇异样本检测方法的 PLS 模型预测结果图

Fig. 11　Prediction results of PLS models using different outlier detection methods

A：original；B：MD；C：Cook′s D；D：MCCV；E：iForest；F：GMM

4 结 论

本文针对近红外光谱分析中谱峰重叠及局部形变导致传统方法难以识别局部奇异特征的问题，提出一种基于混合高斯分解的奇异样本识别方法。该方法根据近红外光谱的形成机理，在特征吸收峰区间建立混合高斯模型，将宽复合谱带解析为对应不同化学基团的多个高斯分量。同时依据各分量幅值在样本间的分布，采用数据驱动的四分位距准则设定各分量的正常波动区间，实现对局部谱峰奇异特征的定量评估。若样本在任一高斯分量上的幅值超出该区间，即判定为奇异样本并予以剔除。

以柠檬酸发酵生产原料段混液中总糖浓度预测为应用对象，实验结果表明，该方法能够有效识别人为构建的奇异样本及实际光谱中的局部吸收峰形变、位移或强度异常；将剔除奇异样本后的数据用于构建PLS定量模型，预测性能有所提升，相较于多种主流方法表现出优势。本研究不仅提供了一种不依赖固定阈值、基于近红外光谱结构的局部检测工具，而且为处理近红外光谱局部奇异特征提供了更精细、可追溯的分析框架，对提升定量模型的稳健性与预测可靠性具有实用价值。

参考文献：

[1] Zhang W, Kasun L C, Wang Q J, Zheng Y, Lin Z. *Sensors*, **2022**, 22(24): 9764.
[2] Beć K B, Grabska J, Huck C W. *Molecules*, **2020**, 25(12): 2948.
[3] Wu T H, Tung I C, Hsu H C, Kuo C C, Chang J H, Chen S, Tsai C Y, Chuang Y K. *Sensors*, **2020**, 20(19): 5451.
[4] Wang Y, Liu Q, Hou H D, Rho S, Gupta B, Mu Y X, Shen W Z. *J. Comput. Sci.*, **2018**, 26: 178-189.
[5] Chen W, Chen Z G, Liu S, Liu J M, Wang H. *J. Instrum. Anal.*(陈雯, 陈争光, 刘烁, 刘金明, 王河. 分析测试学报), **2025**, 44(10): 2079-2086.
[6] Zareef M, Chen Q, Hassan M M, Arslan M, Hashim M M, Ahmad W, Kutsanedzie F Y H, Agyekum A A. *Food Eng. Rev.*, **2020**, 12(2): 173-190.
[7] Ozturk S, Bowler A, Rady A, Watson N J. *J. Food Eng.*, **2023**, 341: 111339.
[8] Zheng J H, Du Y J, Li W X, Liu Z D, Wang H P. *J. Instrum. Anal.*(郑佳辉, 杜宇君, 李文霞, 刘正东, 王华平. 分析测试学报), **2020**, 39(11): 1365-1370.
[9] Pu Y Y, O'Donnell C, Tobin J T, O'Shea N. *Int. Dairy J.*, **2020**, 103: 104623.
[10] Li X X, Xiao J F, Zhang H M, Lü B, Yin X H, Zhao M, Ma F, Fu J, Hu Y, Li Z H, Wang F D, Shen Y C, Dai S Y. *Spectrosc. Spectral Anal.*(李晓星, 肖金凤, 张洪明, 吕波, 尹相辉, 赵明, 马飞, 符佳, 胡艳, 李志豪, 王福地, 沈永才, 戴舒宇. 光谱学与光谱分析), **2025**, 45(6): 1566-1577.
[11] Min S G, Li N, Zhang M X. *Spectrosc. Spectral Anal.*(闵顺耕, 李宁, 张明祥. 光谱学与光谱分析), **2004**, (10): 1205-1209.
[12] Xiao X, Yuan J, Li J, Wang W N, Nie Y, Jiao F, Yang J. *Food Ferment. Ind.*(肖徐, 袁进, 李静, 王薇娜, 聂叶, 焦富, 杨洁. 食品与发酵工业), **2025**, 51(24): 368-374.
[13] Pan S, Zhang X, Xu W, Yin J, Gu H, Yu X. *Spectrochim. Acta A*, **2022**, 271: 120936.
[14] Xiang J K, Huang Y, Guan S H, Shang Y Q, Bao L W, Yan X J, Hassan M, Xu L J, Zhao C. *Sustainability*, **2023**, 15(16): 12423.
[15] Alfryyan N, Saqib M, Ali S, Mubashir T, Tahir M H, Alrowaili Z A, Al-Buriahi M S. *Mater. Today Commun.*, **2023**, 36: 106556.
[16] Liu Z C, Cai W S, Shao X G. *Sci. China Ser. B*(刘智超, 蔡文生, 邵学广. 中国科学: B辑), **2008**, 38(4): 316-323.
[17] Bian X, Cai W, Shao X, Chen D. *Analyst*, **2010**, 135(11): 2841-2847.
[18] Alghushairy O, Alsini R, Soule T, Ma X. *Big Data Cogn. Comput.*, **2021**, 5(1): 1.
[19] Mouret F, Albughdadi M, Duthoit S, Kouamé D, Rieu G, Tourneret J Y. *Remote Sens.*, **2021**, 13(5): 956.
[20] Vasafi P S, Paquet-Durand O, Brettschneider K, Hinrichs J, Hitzmann B. *J. Food Eng.*, **2021**, 299: 110510.
[21] Lu W Z. *Modern Near Infrared Spectroscopy Analysis Technology*. 2nd ed. Beijing: China Petrochemical Press(陆婉珍. 现代近红外光谱分析技术. 2版. 北京: 中国石化出版社), **2010**: 7-9.
[22] McLachlan G J, Lee S X, Rathnayake S I. *Annu. Rev. Stat. Appl.*, **2019**, 6(1): 355-378.
[23] Gupta P, Moorthy A K, Soundararajan R, Bovik A C. *Signal Process. Image Commun.*, **2018**, 66: 87-94.
[24] Lu W, Ding D, Wu F, Yuan G. *Knowl.-Based Syst.*, **2025**, 310: 112942.
[25] Makarov A A, Petrova I Y, Ryabov E A, Letokhov V S. *J. Phys. Chem. A*, **1998**, 102(9): 1438-1449.
[26] Miyamoto M, Arai T, Komatsu M, Yamamoto A, Mikouchi T. *Polar Sci.*, **2009**, 3(2): 110-116.
[27] Armstrong P R, Maghirang E B, Xie F, Dowell F E. *Appl. Eng. Agric.*, **2006**, 22(3): 453-457.
[28] Silalahi D D, Midi H, Arasan J, Mustafa M S, Caliman J P. *Vib. Spectrosc.*, **2018**, 97: 55-65.

(责任编辑：盛文彦)