近红外光谱谱线特性对物质浓度分析误差影响的研究

赵 喆[1,2,3]，王 慧[1]，王慧泉[1,2,3*]，何鑫伟[1]，缪竟鸿[1,2]，王金海[1,2*]

1. 天津工业大学电子与信息工程学院，天津 300387
2. 天津市光电检测技术与系统重点实验室，天津 300387
3. 天津大学精密仪器与光电子工程学院，天津 300072

摘 要 为解决近红外光谱法分析物质浓度过程中缺乏可测度分析而导致测量过程存在一定盲目性问题，研究在已知测量条件、样品种类、被测组分以及建模分析方法的条件下，利用近红外光谱谱线特性作为参数，在大量样品近红外光谱采集和标准法测得浓度数据等工作前，对被测物质浓度的分析误差做大致估算。经过大量尝试和试验提出等效信噪比(ESNR)和谱线重叠系数(OC)两个重要参数，其中ESNR反映待测组分吸光度占总吸光度的比重，而OC则反映待测组分近红外光谱曲线间的重叠程度。通过理论仿真得到光谱分析中用经典的偏最小二乘回归建立定量分析模型时谱线特性与物质浓度分析误差的关系，分别计算ESNR和OC与被测组分浓度分析误差(RMSE)的关系，并且研究两个谱线参数的独立性。利用理论分析得到结果对浓度为8%～12%乙醇水溶液进行可测度分析，并与近红外光谱法分析的实际结果进行比较。研究通过理论仿真得到使用光谱分析中经典的偏最小二乘回归建立定量分析模型时谱线特性与物质浓度分析误差的关系，其中ESNR与RMSE成反比关系，而OC与被测组分分析误差成非线性的单调关系，并且验证了ESNR和OC两个参数的独立性。通过理论计算和乙醇水溶液近红外光谱检测实验对等效信噪比和谱线重叠系数与光谱分析浓度误差的定量关系进行讨论，通过理论分析得到的乙醇浓度RMSE预估值为0.30%，近红外光谱分析实际RMSE为0.32%，相对误差6.67%，二者结果相符。实现了在测量条件、样品种类、被测组分以及建模分析方法已知的条件下基于近红外光谱分析的待测组分含量理论误差的定量计算和实验验证。该研究明确了对近红外光谱法分析物质浓度有明确定量关系的两个谱线参数，给出了使用光谱分析中经典的偏最小二乘回归建立定量分析模型时的分析误差经验曲线，以及利用曲线进行近红外光谱法待测组分浓度可测度分析方法。结果表明所提出的ESNR和OC两个谱线特性参数的有效性，以及分析误差预估方法的有效性。为近红外光谱法待测组分浓度定量分析提供了有效、快捷的预估方法，完善了近红外光谱法成分含量可测度分析理论，对近红外光谱法物质浓度定量分析研究具有一定指导意义。

关键词 光谱重叠系数；等效信噪比；近红外光谱；可测度分析

中图分类号：O443.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)04-1070-05

引 言

近红外光谱分析方法被广泛的应用于食品、医学、环境、化工等领域的检验和监测[1-4]，具有无损、环保、操作便捷等优势，但存在无特征吸收峰并且谱线重叠严重问题，需要通过化学计量学方法建立模型。减小分析误差是光谱分析研究领域的首要问题，现有研究主要集中在：(1) 研制更高性能的光谱采集仪器，获得更佳的光谱数据[5-6]；(2) 使用适当的光谱预处理方法，提高光谱数据质量[7-8]；(3) 采用合适的算法，建立性能更优更稳定的定量分析模型[9-10]。对于光谱分析误差的影响因素研究，主要集中在：(1) 光谱采集条件的控制和优化[11]；(2) 光谱数据处理和建模算法研究[12-14]。然而，对于样品中待测组分本身的光谱特性对分析误差的影响目前仍然停留在定性分析。研究者们普遍认同：光谱数据的信噪比越高，分析误差越小；待测组分所占比重越大，分析误差越小；待测组分谱线与背景谱线重叠程度越小，分析误差越小。但是，这些谱线特性与分析误差的定量

收稿日期：2018-02-04，**修订日期**：2018-07-29

基金项目：国家自然科学基金项目(61705164)，中国博士后科学基金项目第61批资助

作者简介：赵 喆，1986年生，天津工业大学讲师 e-mail：zhaozhe@tjpu.edu.cn

*通讯联系人 e-mail：huiquan85@126.com

关系研究鲜有报道。

现有近红外光谱分析研究存在一定的盲目性，缺乏研究中的可测度分析环节，通常是先采集大量数据，然后尝试各种建模方法。因此，我们对谱线特性在近红外光谱分析误差的影响进行系统的分析，提出等效信噪比和谱线重叠系数两个参数，通过理论仿真给出光谱分析中经典的偏最小二乘回归定量分析模型的误差经验曲线，并进行乙醇水溶液浓度分析验证。

1　理论分析

1.1　等效信噪比

在实际检测中通常对待测组分的相对分析误差进行评估，因此引入了等效信噪比这一参数。若仪器获得光谱数据的信噪比 SNR，组分 k 的等效信噪比 ESNR，有

$$\mathrm{ESNR}=\frac{A_k}{A}\mathrm{SNR} \tag{1}$$

其中 A 为样品总吸收光谱，A_k 为组分 k 的吸收光谱。进而研究在相同数据处理和分析方法条件下，定量分析均方根误差 RMSE(建模集均方根误差 RMSEC 和预测集均方根误差 RMSEP 的均值作为 RMSE)与等效信噪比 ESNR 的关系。

1.2　重叠系数

经过多种尝试和计算，提出重叠系数(overlapping coefficient，OC)，此参数与 RMSE 成单调关系。重叠系数 OC 的定义如下：

若把样品看作由待测组分 k 和非待测组分的背景 B 这两种组分构成，将两组分消光系数曲线 $\varepsilon_k(\lambda)$ 和 $\varepsilon_B(\lambda)$ 面积归一化，即令 $\varepsilon'_k(\lambda)=\dfrac{\varepsilon_k(\lambda)}{\sum_{\lambda=1}^{N}\varepsilon_k(\lambda)}$，$\varepsilon'_B(\lambda)=\dfrac{\varepsilon_B(\lambda)}{\sum_{\lambda=1}^{N}\varepsilon_B(\lambda)}$，使得 $\sum_{\lambda=1}^{N}\varepsilon'_k(\lambda)=\sum_{\lambda=1}^{N}\varepsilon'_B(\lambda)=1$，其中 N 为光谱波长数。待测组分 k 的重叠系数 OC 定义为

$$OC=1-\frac{1}{2}\sum_{\lambda=1}^{N}|\varepsilon'_k(\lambda)-\varepsilon'_B(\lambda)| \tag{2}$$

其几何意义为：面积归一化后，待测组分 k 与背景 B 的消光系数曲线重叠部分面积与背景组分曲线所围面积的比值。由定义可知，重叠系数 OC 的取值范围为[0，1]，其值越接近 1，说明待测组分 k 与背景 B 的消光系数曲线重叠程度越大。

2　实验部分

在 ESNR 和 OC 两个光谱曲线参数固定时，正弦波、三角波、高斯曲线等不同谱线形状对分析结果无显著影响。考虑到近红外光谱形状与多高斯曲线叠加的形状相似，仅列出通过改变峰宽和峰距的高斯型光谱曲线获得不同谱线特性参数 OC 对近红外光谱分析误差影响的实验、结果及分析。

2.1　等效信噪比对近红外光谱分析误差影响研究

首先在谱线重叠系数 OC 及其他参数固定的情况下，构造 ESNR 的单变量实验。共 512 个样本，设定吸收光谱数据长度为 512 个波长，信噪比为 1 000，组成光谱矩阵 $\boldsymbol{S}$。各样品中待测组分浓度 c_k 为均值为 1% 的 0.8%～1.2%均匀分布随机数，背景组分浓度为 100%－c_k。构造三组消光系数曲线，第 1 组和第 2 组幅值相同而形状不同，第 3 组幅值为前两组的 1/10，计算得到待测组分 k 的等效信噪比 ESNR 为：6.3～221.2。

建模集和预测集按比例 7∶3 在样品集中随机选取。使用偏最小二乘回归方法建立 c_k 定量分析模型，主因子数由逐一交叉验证法确定为 2，记录三组消光系数所对应的待测组分的均方根误差 RMSE。

2.2　重叠系数对近红外光谱分析误差影响研究

在 ESNR 固定的条件下，通过设定被测组分和背景组分消光系数曲线吸收峰的位置、高度和宽度等来改变两曲线的重叠程度，从而探究 OC 对定量分析误差的影响。

用高斯曲线构造 1 000 个等效信噪比为 500 的吸收光谱作为样品。根据朗伯比尔定律，设定光程长为 10 mm，随机抽取 70%的样品作为建模集，其余为预测集。使用偏最小二乘法建立光谱预测模型，经过逐一交叉认证选取主因子数为 2。通过改变两个高斯曲线的峰距比较 16 个 OC 时的 RMSE。

2.3　等效信噪比与重叠系数对近红外光谱分析误差影响研究

为研究 ESNR 与 OC 对近红外光谱分析误差共同影响，研究在 2.2 中实验基础上加入 ESNR 为 50 和 5 000 两组实验，进行偏最小二乘建模分析，研究 OC 对分析误差的影响。若 RMSE 按比例变化，则可将 ESNR 和 OC 视为两个独立变量。

2.4　乙醇水溶液近红外光谱分析误差影响研究

为验证通过 ESNR 和 OC 对近红外光谱分析误差影响研究的实用性，对乙醇水溶液样品进行近红外光谱采集及分析。使用 Ocean Optics 公司 NIRQUEST512 近红外光谱仪对乙醇和纯水的近红外光谱进行测量，积分时间 8 ms，样品池厚度 10 mm，得到待测组分乙醇和蒸馏水的面积归一化消光系数曲线如图 1 所示，可由式(2)计算得到它们的重叠系数 OC。

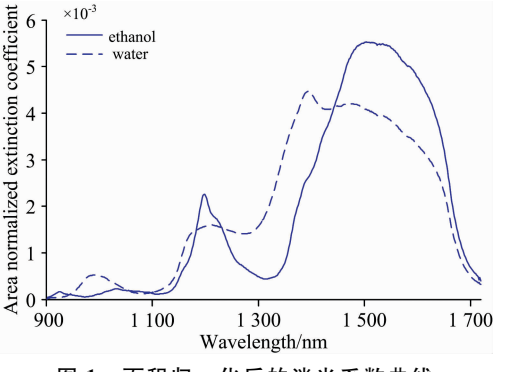

图 1　面积归一化后的消光系数曲线

Fig. 1　The extinction coefficient curve after area normalization

配制浓度为 8%至 12%的乙醇水溶液样品进行近红外光谱数据采集和分析，以浓度间隔为 0.2%配制 21 个样品构成建模集，以 0.5%为浓度间隔 9 个样品构成预测集。用偏最

小二乘法建立乙醇浓度分析模型。

3 结果与讨论

3.1 等效信噪比对近红外光谱分析误差影响研究

在不同 ESNR 下，使用偏最小二乘回归方法进行光谱定量分析，得到待测组分 k 的等效信噪比 ESNR 与定量分析均方根误差 RMSE 的关系如图 2 所示。从图 2 中可以看出，RMSE 与 ESNR 成反比关系。

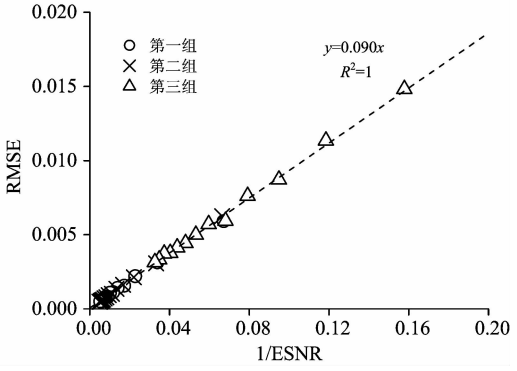

图 2 等效信噪比对浓度预测误差的影响

Fig. 2 The effect of equivalent SNR on concentration prediction error

3.2 谱线重叠系数对近红外光谱分析误差影响研究

ESNR 为 500 时不同重叠系数情况下建模和预测均方误差由图 3 中第一组实验数据表示。当信噪比一定时，待测组分浓度的 RMSE 随两组分 OC 的变化显著增长，可以看出 OC 是决定待测组分近红外定量分析误差的重要因素。

由于近红外光谱分析中谱线重叠程度较为严重，一般情况下 OC 值较大，而第一组实验中 OC 为 0.8～1 区间的各点距离较大，故加入第二组实验，在保持 ESNR 等实验条件和建模分析方法一致的情况下，通过改变谱线峰距和峰宽，增加了 48 个 OC 较大时的数据点，分析结果如图 3 所示。

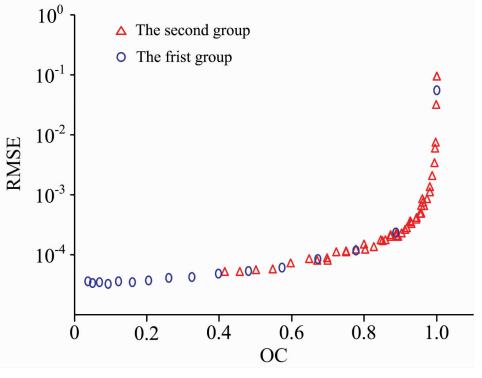

图 3 重叠系数对分析误差的影响

Fig. 3 The effect of OC on analytical error

由此，通过上述实验及分析，得到近红外光谱分析中，被测组分和背景组分的重叠程度可以用光谱重叠系数 OC 进行表述，并且图 3 即为 OC 与建模分析均方根误差 RMSE 的关系。该关系曲线不能用简单函数进行拟合，实际使用时可使用分段插值等方式进行计算。

3.3 等效信噪比与重叠系数对近红外光谱分析误差影响研究

ESNR 比为 50，500 和 5 000 的样品数据进行建模分析后 RMSE 按 ESNR 比倍数进行数值统一，分别乘以 100，10 和 1。所得结果如图 4 所示。

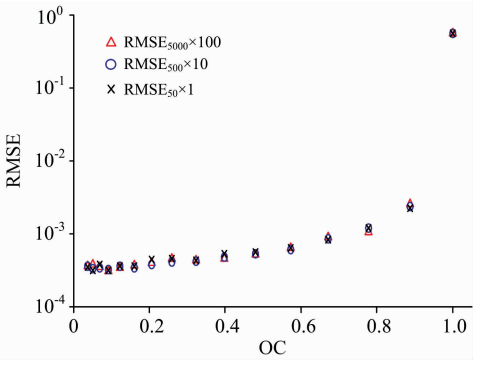

图 4 不同 ESNR 情况下 OC 对分析误差的影响

Fig. 4 The effect of OC on analytical error in different ESNR

通过对不同 ESNR 的分析结果进行对比，可以得到在不同 ESNR 条件下下，被测组分浓度 RMSE 随 OC 单调变化且变化趋势是一致的，且各组 RMSE 的数值上相差的倍数与 ESNR 的倍数一致。由此，通过研究得到 OC 与 ESNR 比对近红外光谱定量分析误差的影响是相对独立的。均方根误差 RMSE 可以表示为 ESNR 和 OC 的函数，即

$$\mathrm{RNSE} = f(\mathrm{ESNR}, \mathrm{OC}) \tag{5}$$

3.4 乙醇水溶液近红外光谱分析误差影响研究

首先通过 ESNR 和 OC 对乙醇水溶液样品近红外光谱分析误差进行预估，使用 NIRQUEST512 光谱仪得到近红外光谱的信噪比约为 230。依据式(1)由乙醇和纯水的消光系数曲线计算得到当被测样品中乙醇浓度均值为 10%时，被测样品中乙醇的 ESNR 约为 20。依据式(2)计算得到乙醇与纯水的 OC 为 0.81，代入图 5 曲线可得，当 ESNR 为 500 时，RMSE 为 1.20×10^{-4}。按照 ESNR 的比值进行换算，本实验条件下 10%的乙醇水溶液的浓度分析理论上的 RMSE=1.20×10^{-4}

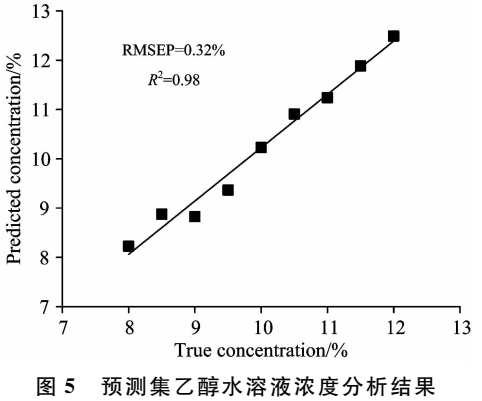

图 5 预测集乙醇水溶液浓度分析结果

Fig. 5 The result of the concentration analysis of the prediction set of ethanol aqueous solution

×500/20=0.30%。

通过对本实验中配制的 30 例乙醇浓度为 8 %至 12 %的乙醇水溶液样品(21 例建模，9 例预测)进行近红外光谱分析，得到预测集的乙醇的预测浓度和真实浓度的相关性分析结果如图 5 所示。

预测集乙醇水溶液的浓度均方根误差 RMSEP=0.32%，R^2=0.98。该建模分析结果与通过本研究中提出利用 ESNR 和 OC 计算得到的 RMSE 预估值相符，验证了本方法的正确性。

4　结　论

ESNR 和 OC 是在对多种谱线参数(如谱线形状、信噪比、相关系数、重叠面积等)的尝试和验证基础上提出两个有效参数。ESNR 不同于以往讨论的吸收光谱的信噪比，而是考虑待测组分的吸光度占总吸光度的比重，从而直接反映待测组分的信噪比。OC 这一参数的提出，使得近红外光谱曲线之间的重叠程度有了可度量的标准。

基于 ESNR 和 OC 对近红外光谱分析误差影响的系列实验，提出了对待测组分分析误差的预估方法。若预估值满足测量需要，即可进行大量数据采集建立分析模型；若不满足，可以通过提高光谱检测仪器信噪比、改进建模分析方法或者改用除近红外光谱分析外的其他生化检测方法对待测组分浓度进行定量分析。文中关系曲线是在使用光谱分析中经典的偏最小二乘回归建立定量分析模型条件下得到的，供本领域的研究者参考，若使用其他建模分析方法，需要重新计算关系曲线。

针对目前近红外光谱分析中缺乏可测度分析而普遍存在一定未知性和盲目性的问题，通过对谱线特性的系统分析，提出 ESNR 和 OC 两个参数，结合理论仿真和乙醇水溶液浓度分析实验并对两个参数对光谱分析误差的影响进行讨论。为近红外光谱定量分析的研究提供了有效快捷的预估分析误差方法，完善了近红外光谱法可测度分析理论，对近红外光谱法组分浓度定量分析的顺利进行有较高的实际意义。

References

[1] FAN Rui, SUN Xiao-kai, CHEN Jie, et al(范　睿，孙晓凯，陈　杰，等). Modern Food Science & Technology(现代食品科技), 2017, 33(11): 264.

[2] Han G, Han T, Xu K, et al. Journal of Biomedical Optics, 2017, 22(7): 77001.

[3] CHEN Hong-yan, ZHAO Geng-xing, ZHANG Xiao-hui, et al(陈红艳，赵庚星，张晓辉，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2014, 30(8): 91.

[4] Camara A B F, Carvalho L S D, Camilo L M M, et al. Fuel, 2017.

[5] Feng X M, Yu H X, Yi X Q, et al. Analytical Methods, 2017, 9: 2578.

[6] Zhao M, Downey G, O'Donnell C P. Food Control, 2016, 68: 260.

[7] Dotto A C, Dalmolin R S D, Caten A T, et al. Geoderma, 2018, 314: 262.

[8] Dotto A C, Dalmolin R S D, Grunwald S, et al. Soil & Tillage Research, 2017, 172: 59.

[9] WANG Li-jie, YANG Yu-yi(王丽杰，杨羽翼). Acta Optica Sinica(光学学报), 2017, 37(10): 350.

[10] Holland J K, Kemsley E K, Wilson R H. Journal of the Science of Food & Agriculture, 2015, 76(2): 263.

[11] SHEN Fei, YING Yi-bin, LI Bo-bin(沈　飞，应义斌，李博斌). Food Science(食品科学), 2014, 35(23): 25.

[12] WANG Hai-xia, SUO Tong-chuan, YU He-shui, et al(王海霞，所同川，余河水，等). China Journal of Chinese Materia Medica(中国中药杂志), 2016, 41(19): 3537.

[13] ZHANG Jin, CAI Wen-sheng, SHAO Xue-guang(张　进，蔡文生，邵学广). Progress in Chemistry(化学进展), 2017, 29(8): 902.

[14] XU Ling, LI Wei-hua, YANG Ying, et al(徐　玲，李卫华，杨　英，等). China Environmental Science(中国环境科学), 2016, 36(5): 1426.

Influence of Spectral Characteristics on the Accuracy of Concentration Quantitatively Analysis by NIR

ZHAO Zhe[1, 2, 3], WANG Hui[1], WANG Hui-quan[1, 2, 3*], HE Xin-wei[1], MIAO Jing-hong[1, 2], WANG Jin-hai[1, 2*]

1. School of Electronics and Information Engineering, Tianjin Polytechnic University, Tianjin 300387, China
2. Tianjin Key Laboratory of Optoelectronic Detection Technology and Systems, Tianjin 300387, China
3. School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin 300072, China

Abstract In order to solve the problem of measurement blindness caused by the lack of measurable analysis in the the near-infrared spectroscopy, we can roughly estimate the analytical error of the concentration of the tested substances using the spectral characteristics of near-infrared spectroscopy under the known conditions of measurement, sample types, components under analysis and modeling and analysis methods, before a large number of samples were collected by near-infrared spectroscopy and concentration data measured by standard method. In the research, two important parameters, ESNR and OC, were proposed and tested. ESNR reflects the proportion of the component absorbance to the total absorbance, while OC reflects the overlap degree between near-infrared spectral curves of the components. We got the relationship between spectral characteristics and concentration analysis error when using the classical partial least squares regression in spectral analysis to establish quantitative analysis model through theoretical simulation. The relationship between ESNR and OC and the concentration of analyte (RMSE) was calculated respectively, and the independence of the two spectral parameters was also studied. The results of theoretical analysis were used to measure the concentration of aqueous ethanol solution between 8% and 12%, and compared with the actual results of near infrared spectroscopy. The relationship between the spectral characteristics and the concentration analysis errors when using partial least squares regression to establish a quantitative analysis model was obtained through theoretical simulation. ESNR is inversely proportional to RMSE, and OC is in a non-linear monotonic relationship with the measured component analysis error, and the independence of ESNR and OC was verified. The quantitative relationship between ESNR and OC and spectral concentration error was discussed by theoretical calculations and near-infrared spectroscopy of ethanol aqueous solution. The RMSE of ethanol concentration was 0.3% which was estimated by theoretical analysis, and the RMSE of near infrared spectroscopy was 0.32%. The relative error was 6.67%. We have realized the quantitative calculation and experimental verification of the theoretical error of the content of the tested components based on near infrared spectroscopy under the conditions of the measurement conditions, the types of samples, the components to be measured, and the methods of modeling and analysis. This study identified two spectral parameters that have a clear and quantitative relationship with the concentration of the measured component in NIR spectroscopy. The analytical accuracy empirical curve was established when using the classical partial least-squares regression in spectral analysis. In addition, the analysis of the measurable degree of the concentration of the components could also be tested by near infrared spectroscopy. The results showed the effectiveness of the ESNR and OC in this paper, as well as the analytical method of error prediction. This study provided an effective and rapid prediction method for the quantitative analysis of near infrared spectroscopy, and optimized the theory of measurable analysis of near infrared spectroscopy, which has a good guidance for the quantitative analysis of the concentration of near infrared spectroscopy.

Keywords Overlapping coefficient; Noise-signal ratio; Spectrum analysis; Measurable analysis

(Received Feb. 4, 2018; accepted Jul. 29, 2018)

* Corresponding authors