ICS 71.040.99
N 30

中华人民共和国国家标准

GB/T 37969—2019

近红外光谱定性分析通则

Standard guidelines for near infrared qualitative analysis

2019-08-30 发布　　2020-03-01 实施

国家市场监督管理总局
中国国家标准化管理委员会　发布

目　　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准由中华人民共和国科学技术部提出。

本标准由全国仪器分析测试标准化技术委员会(SAC/TC 481)归口。

本标准起草单位:上海烟草集团北京卷烟厂、北京化工大学、南开大学、石油化工科学研究院、中国食品药品检定研究院、中国检验检疫科学研究院、军事科学院评估论证研究中心、北京市农林科学院、中国农业大学、中国计量科学研究院、中检国研(北京)科技有限公司、云南中烟工业有限责任公司、云南同创检测技术股份有限公司、西派特(北京)科技有限公司。

本标准起草人:马雁军、袁洪福、王家俊、周骏、邵学广、褚小立、杜国荣、王纪华、尹利辉、田高友、马莉、李军会、宋春风、侯英、邹明强、袁天军、温亚东、许育鹏、陶鹰、宋德伟、胡爱琴、杨玉清、李伟、杨盼盼、王明锋、齐小花、王冬、王建平。

近红外光谱定性分析通则

1 范围

本标准规定了近红外光谱定性分析的基本原理和方法、使用软件、仪器设备、光谱测量、样品、定性分析试验步骤、试验数据处理、试验报告等内容的通用要求。

本标准适用于吸收范围为 12 820 cm^{-1}～4 000 cm^{-1}(即 780 nm～2 500 nm)近红外光谱定性分析。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 8322 分子吸收光谱法 术语

GB/T 29858—2013 分子光谱多元校正定量分析通则

3 术语和定义

GB/T 8322 界定的以及下列术语和定义适用于本文件。

3.1

训练样品 training samples

学习样品 learning samples

参考样品 reference samples

属性、特征或组成已知的样品。

3.2

训练集 training set

学习集 learning set

参考集 reference set

训练样品的集合。

3.3

验证样品 validation samples

用于验证类模型判别能力的样品,其属性、特征或组成已知。

3.4

验证集 validation set

验证样品的集合。

3.5

训练 training

学习 learning

找出光谱与样品属性、特征或组成之间关系的过程,即建立类模型的过程。

3.6

类　class

对特征、组成或性质赋予样品的一种属性。

3.7

识别率　recognition rate

采用类模型判别出来的正确样品数占本类训练样品总数计算的百分比，在定性分析中用来评价类模型的理论判别能力。

3.8

判别正确率　discriminant rate

采用类模型判别出来的正确样品数占本类验证样品总数计算的百分比，在定性分析中用来评价类模型的实际判别能力。

3.9

模式识别方法　pattern recognition method

计算机学习(亦称训练)的相关方法。

3.10

有监督的模式识别方法　supervised pattern recognition method

在已知样本的监督下进行的模式识别方法。

3.11

无监督的模式识别方法　unsupervised pattern recognition method

在没有先验知识下进行的模式识别方法。

3.12

质量控制样品　quality control sample

具有一种或多种的物理或化学特征值，且均匀稳定的物质或材料，用于检查和校正在用测量系统的精密度和稳定性。

4　原理与方法

4.1　基本原理

样品的近红外光谱与物质本身的化学组成及含量相关，包含丰富的化学组成与结构信息，样品的化学组成、物质结构及含量决定样品的属性、特征，将样品的近红外光谱作为变量，首先采用适合的模式识别方法，建立样品类属与样品近红外光谱之间的对应关系(即类模型)，然后将类模型应用于待测样品的近红外光谱，通过计算确定该样品的类属或特征。

4.2　模式识别方法

4.2.1　模式识别方法常分为有监督的模式识别方法和无监督的模式识别方法。

4.2.2　有监督的模式识别方法包含族类的独立软模式，即 SIMCA (soft independent modeling of class analog)、偏最小二乘判别分析(partial least squares discriminant analysis，PLS-DA)、人工神经网络(artificial neural networks，ANN)和支持向量机(support vector machine，SVM)等。其共同特征为使用适量类别或特征已知的样品光谱作为训练集，应用计算机软件向训练集光谱学习，通过学习过程获得光谱与样品类属或特征之间的对应关系，即建立类模型。SIMCA 和偏最小二乘判别分析(PLS-DA)建

立类模型过程，参见附录 A。

4.2.3 无监督的模式识别方法包括主成分分析（principal component analysis，PCA）方法和系统聚类（或称分层聚类）分析法（hierarchical cluster analysis，HCA）等。其共同特征为不知道样品分类的情况下，对样品光谱无需训练过程的分类方法。模式识别方法的使用不限于上述例举几种方法。

5 化学计量学软件

应使用与近红外光谱仪匹配的化学计量学软件，且具备以下基本的数据处理方法及功能：

a) 样品相关信息和光谱数据的录入、存取、数据格式转换与编辑；

b) 均值中心化（mean centering）、标准化（auto scaling）、MSC（multiplicative scatter correction）、SNV（standard normal variate）、微分和平滑等数据预处理及处理结果的浏览、列表和可视化；

c) SIMCA 和 PLS-DA 分类方法及分析结果浏览、列表和可视化；

d) 异常样品的统计识别和删除；

e) 模型自动交互验证及验证统计结果的浏览、列表和可视化；

f) 测定结果的浏览、列表和输出。

6 仪器设备

使用仪器设备应符合 GB/T 29858—2013 中第 5 章规定的仪器设备要求。

7 光谱测量

光谱测量应按 GB/T 29858—2013 中第 6 章规定的光谱测量要求进行。

8 样品

8.1 总则

应按照样品所属行业的国家标准或行业标准等方法取制、保存和使用。

8.2 训练样品的选择

8.2.1 建立一个类模型，应明确训练样品的类别属性、特征或组成，训练集应包含使用该类模型判别待测样品中可能存在的特征范围。

8.2.2 用 SIMCA 和 PLS-DA 分类法建立类模型，如果使用 A 个（>3）主成分建立类模型，剔除异常样品后，在训练集中的每一类至少应含有 $6A$ 个样品；如果建模光谱数据进行均值中心化处理，剔除异常样品后，在训练集中的每一类至少应含有 $6(A+1)$ 个样品；满足实际应用要求的样本数宜更多为好，每一类最少应含有 24 个独立样品。

8.2.3 用基于 PCA 为基础的分类法建立类模型，在实际训练过程中，应根据样品的复杂性和建模所需要的主成分数，确定建立类模型所需的训练样品数量，参照经验统计规则，剔除异常样品后，适宜的训练样品数量应不低于 $10A$ 个。

8.3 验证样品的选择

8.3.1 验证集样品的类别属性、特征或组成已知，应包含使用类模型分析待测样品中可能存在的特征范围，验证样品也应包含与被验证类相近的其他类样品。

8.3.2 所需验证样品的数量取决于类模型的复杂性，如果类模型使用的主成分数 $A \leqslant 5$，需要的验证样品数不能少于 20；如果类模型使用的主成分数 $A \geqslant 5$，则所需要的验证样品数应不少于 $4A$。

8.3.3 异常样品为类模型不包含的样品。在验证样品中置入适量实际应用中可能遇到的异常样品，用于检验类模型的判别异常样品能力。

9 近红外光谱定性分析试验步骤

9.1 近红外光谱定性分析试验步骤主要包括：

a) 收集适量具有代表性的类别或特征已知的样品作为训练样品和验证样品；

b) 测量前仪器应诊断校验，校验通过后使用仪器测量训练样品和验证样品的近红外光谱，每个样品平行测量两次；

c) 应将训练样品近红外光谱进行预处理，选择合适的模式识别方法，使用计算机软件向预处理后的训练集光谱数据学习，通过学习过程建立类模型，然后使用验证样品的近红外光谱验证类模型，统计判别正确率，评价类模型的判别能力和有效性，决定类模型能否适用；

d) 采用验证通过的类模型对待测样品的近红外光谱进行定性判别分析，确定待测样品的归属类别或特征。

9.2 近红外光谱定性分析试验的主要步骤及相应要求，详见流程示意图 1。

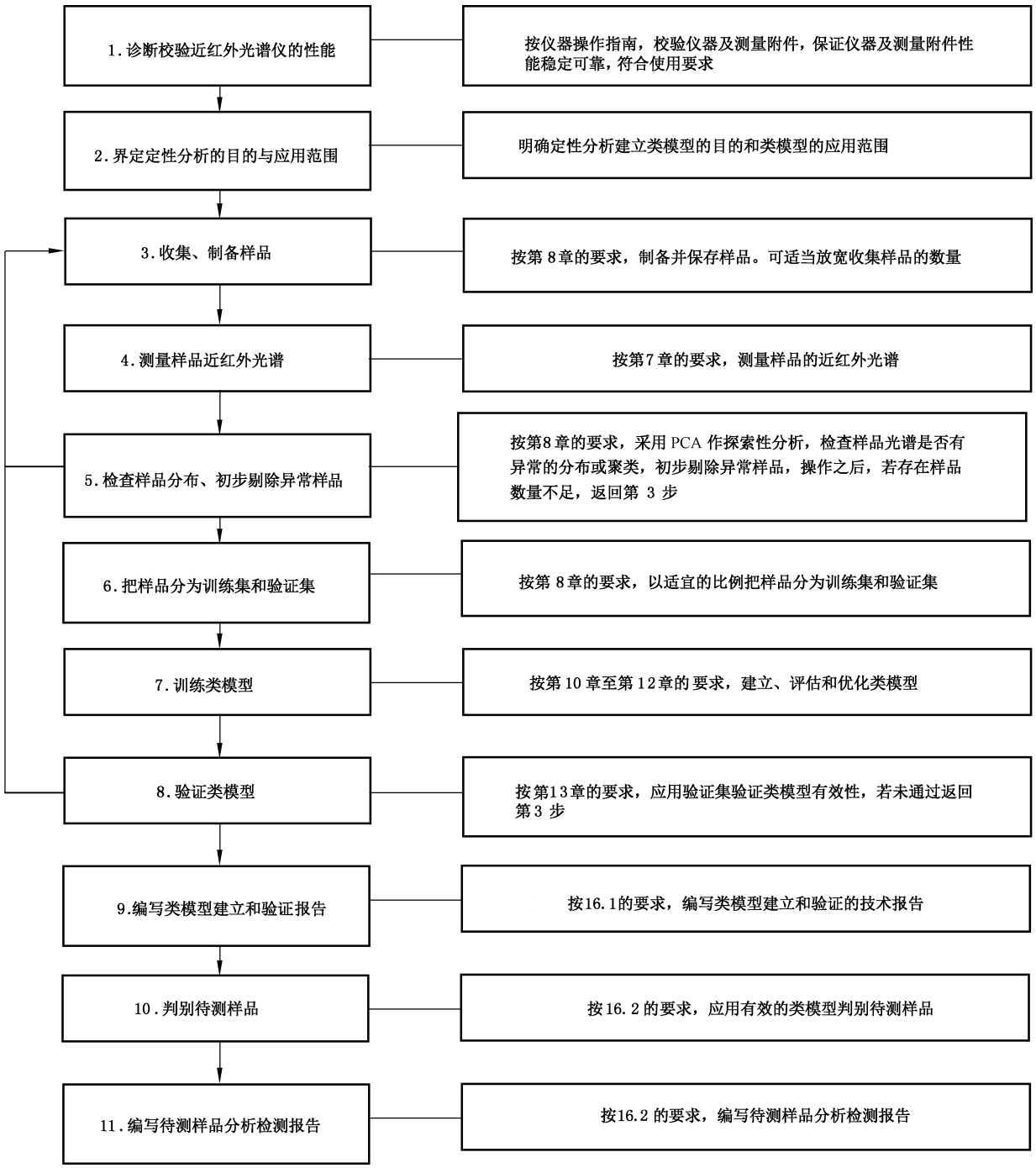

图1　近红外光谱定性分析流程示意图

10　光谱数据预处理

10.1　在实际建模过程中，可独立或联合应用不同的数据预处理方法（中心化、归一化等）和降噪处理方法处理光谱数据，增强光谱之间的可比性和降低噪声干扰，以达到建模效果。

10.2　降噪处理方法包括，但不限于：平滑、微分、MSC和SNV等。微分可降低光谱基线变化的不利影响，但同时也会降低光谱的信噪比；平滑可提高信噪比，但也会降低分辨率；MSC或SNV可降低固体颗

粒散射作用的影响。应根据样品近红外光谱的实际情况，可通过化学计量学软件上的窗口选择添加功能来实现，对比应用不同的数据预处理和降噪处理方法后的效果，挑选出较优结果的光谱数据预处理方法，为建立类模型中使用。

10.3 处理训练样品光谱、验证样品光谱及待测样品光谱时，应采用包含参数在内皆相同的数据预处理方法。

11 光谱特征变量选择

11.1 对样品的原始光谱或预处理后的光谱进行相关性统计分析，选择特征性强、模型化能力强，对分类贡献大的光谱特征变量（即光谱波段或波长/波数）建模。

11.2 选择适宜的光谱特征变量，通常采用具备光谱数据的统计分析功能或选择光谱特征变量的方差比较法、Fisher 比率法等，在训练类模型的过程中，可使用这些统计功能或方法得出来效果的比较，来进行光谱特征变量选择。对于熟悉样品化学成分的光谱特征归属，可以人工选择光谱区间，但不宜选择过窄的光谱波段或过少的波长数目。光谱特征变量选择方法不限于以上几种方法。

11.3 建模时针对训练集光谱所采用的选择光谱特征变量的方法，在使用验证集光谱和待测样品光谱时，也应采用相同的方法选择光谱特征变量。

12 类模型的建立

12.1 代表性样品筛选和分集

收集足够数量的代表性样品，采集样品光谱，并用主成分分析对每一类样品光谱进行探索性分析，参见附录 A；采用 F 检验，计算其临界值，可以查附录 B 中的 F 分布临界值表，初步判断剔除异常样品。选择合适数量的训练样品和验证样品分别组成训练集和验证集。

12.2 分类方法的选择

SIMCA 或 PLS-DA 等分类方法进行定性分析，具体可根据训练样品的复杂程度，选择适宜的分类方法建立类模型，参见附录 A。

12.3 类模型的建立

12.3.1 选择合适数据预处理方法对光谱进行预处理，应选择模型化能力强的光谱特征变量及适宜主成分数等建模条件，建立类模型。

12.3.2 为优化类模型，提高类模型的判别能力和稳健性，在训练类模型的过程中应进行异常样品的统计与识别，并选择适宜的主成分数建模，参见附录 A。

12.3.3 如果类模型的判别正确率满足使用者预期要求，则类模型建立完毕；如果类模型的判别正确率不能满足使用者预期要求，类模型的有效性可疑，则需要检查训练类模型过程中的每个步骤，选择其他建模方法或建模条件，重新训练类模型，直到类模型满足要求。

13 类模型的有效性验证

13.1 将 8.3 选择的验证样品光谱输入类模型计算，判别验证样品的归属类别或特征，并计算类模型的判别正确率，然后根据判别正确率评估类模型的有效性。

13.2 在训练类模型的过程中，采用识别率初步评估类模型的判别能力；在验证类模型的过程中，采用判别正确率衡量类模型的实际判别能力。通常，类模型的识别率大于判别正确率。识别率越高，类模型

的理论判别能力就越强。在验证类模型时，宜把识别率与判别正确率结合，用于综合评估类模型的判别能力。使用者应根据实际需要，设定一个合适的判别正确率，然后通过验证集来验证类模型，如果类模型判别验证样品，得到的判别正确率不低于预期，则类模型通过有效性验证。

13.3 可采用混淆矩阵(confusion matrix)来进一步分析判别类别与实际类别的对应关系，从而对类模型的判别能力进行评价。

13.4 编写类模型建立与验证报告，见16.1。

14 类模型的应用

14.1 在测量待测样品的光谱时，应在相同环境，采用相同仪器和试验条件下，使用与测量训练样品光谱完全一样的步骤测量待测样品的光谱。

14.2 调用经有效性验证的类模型计算待测样品光谱，通过 F 检验判别待测样品的归属类别，不同风险水平的 F 分布临界值，参见附录B。

14.3 编写试验分析报告，见16.2。

15 类模型的维护

15.1 在类模型投入使用之后，应使用质量控制样品对类模型的有效性进行连续的监测。如果是类模型的性能变差，导致类模型判别正确率下降，应适量增删训练集的样品，并按本标准要求优化或重新建立类模型，以保证类模型的适应性和有效性。

15.2 应采用仪器自带或者相关标准的性能测试方法对仪器的性能进行监测。若仪器性能下降，导致类模型判别正确率低于预期要求，则应对仪器性能进行检测评价，校正检测出的仪器问题。若仪器进行了维修，如更换光学元器件、检测器等，应当重新标定仪器，检验仪器硬件的一致性，并使用质量控制样品对类模型的有效性进行评估。如果维修后仪器硬件的一致性改变，导致类模型判别正确率低于预期要求，则应进行类模型传递或重新建立类模型。

注：模型传递的有关方法及其应用超出本文件范围，参见GB/T 29858—2013附录或有关化学计量学文献。

15.3 在测量质量控制样品的光谱时，应使用与测量训练样品光谱完全相同的步骤及参数测量质量控制样品的光谱。

15.4 编写类模型维护报告，见16.3。

16 试验报告

16.1 类模型建立与验证报告

类模型建立与验证报告内容包括但不限于：

a) 类模型建立人员和建立时间；
b) 类模型建立的实验室，包括实验室的名称、地址、联系方式等；
c) 样品选择情况，包括训练集和验证集样品的属性、数量、采集时间、制备、保存方法等；
d) 环境条件，包括温度、湿度等；
e) 仪器名称、型号、测量条件以及诊断仪器性能的测试结果等；
f) 类模型建立，包括建立类的模型名称、采用化学计量学软件的名称和版本号、光谱预处理的方法、使用的定性分类法名称、异常样品的处理方法、验证统计结果及可视化图示等；
g) 建立的类模型适用范围，包括适用样品类型和条件、测量方法、测量条件等；

h） 遵守本标准规定的程度，包括对分析结果可能有影响而本标准未包括的操作或者任选的操作；
i） 测定中观察到的异常现象。

16.2 待测样品试验分析报告

待测样品试验分析报告内容包括但不限于：
a） 待测样品名称、编号、类型、送样时间；
b） 分析人员、分析时间；
c） 测试实验室：包括名称、地址、联系方式等；
d） 环境条件，包括温度、湿度等；
e） 仪器名称、型号、测量条件以及诊断仪器性能的测试结果等；
f） 分析判别结果；
g） 遵守本标准规定的程度，包括对分析结果可能有影响而本标准未包括的操作或者任选的操作；
h） 测定中观察到的异常现象。

16.3 类模型维护报告

类模型维护报告内容包括但不限于：
a） 因类模型性能下降，通过增删训练样品提高类模型判别能力的有关记录；
b） 因仪器及相关附件维修维护或更换，对类模型进行传递、重建的有关记录；
c） 维修维护或更换仪器及其相关附件（含仪器控制软件）前的仪器性能测试报告；
d） 维修维护或更换仪器及其相关附件（含仪器控制软件）后的仪器性能测试报告；
e） 所更换的仪器相关附件（含仪器控制软件）的制造商、名称和型号（版本）的有关记录；
f） 建立类模型的化学计量学软件以及相关数据分析处理软件升级的有关记录。

17 试验质量保证要求

近红外光谱定性分析试验质量保证要求包括但不限于：
a） 具备规范的光谱实试验室和试验环境，满足近红外光谱仪正常运行；
b） 类模型建立人员应具备参照本标准中要求建立类模型、定期验证及维护类模型的能力；
c） 操作分析人员应按照检测分析机构相关技术要求或规范，具备从事仪器测试试验的能力；
d） 具有按照检测分析机构有关技术要求规范，执行诊断校验近红外光谱仪的能力；
e） 使用质量控制样品，对测量全过程进行跟踪，实施全过程质量保证。

18 常见的误差类型、来源及解决途径

在近红外光谱定性分析中，常见的误差类型、来源及解决途径，参见附录C。

19 类模型建立与验证示范实例

19.1 基于近红外光谱PLS-DA法判别三类性质相近药品实例，参见附录D。
19.2 近红外透射光谱对汽油质量等级分类SIMCA法实例，参见附录E。

附 录 A
（资料性附录）
近红外光谱模式识别常用方法简介

A.1 SIMCA 分类方法

A.1.1 SIMCA 分类方法简要解释

SIMCA 以基于一种主成分分析的有监督的模式识别方法进行的分类，它利用先验分类知识，对每一种类别建立一个 PCA 类模型，然后利用该类模型判别待测样品的类别归属。

A.1.2 主成分分析与主成分类模型的建立

选择适量的已知类别样品光谱作为训练集。在训练集中，对任意一个由 n 个波长点（即光谱变量）构成的样品光谱 x_i 可用矢量表示，见式（A.1）：

$$x_i = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{in}] \qquad \cdots\cdots (A.1)$$

那么，对 m 个样品构成的样品集，就可表达为 $m \times n$ 阶的光谱数据矩阵，见式（A.2）：

$$\boldsymbol{X}_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \qquad \cdots\cdots (A.2)$$

对 $\boldsymbol{X}_{m\times n}$ 进行主成分分析，是通过正交变换将标准化处理后的光谱矩阵 $\boldsymbol{X}_{m\times n}$ 中可能存在线性相关的光谱变量组合为一组线性不相关的综合变量 $F_1, F_2, \cdots, F_A$（$a=1,2,3,\cdots,A$，且 $A<n$），称之为主成分，记 F_1 为第一主成分，F_2 为第二主成分，依次类推。这 A 个主成分可以在信息损失最少的原则下，对高维光谱变量降维，且最多地解释样品光谱矩阵 $\boldsymbol{X}_{m\times n}$ 中的信息。通常，在样品光谱矩阵 $\boldsymbol{X}_{m\times n}$ 中的信息可以用全部光谱变量方差的总和来衡量，方差越大，可视为 $\boldsymbol{X}_{m\times n}$ 中包含的信息越多。在主成分分析的结果中，若以方差 $\mathrm{Var}(F_a)$ 来衡量第 a 个主成分 F_a 所提取的信息，则当抽取 A 个主成分时，这 A 个主成分所携带的信息总和等于特征值 λ_a（也称本征值，eigenvalues）总和，可按式（A.3）进行计算：

$$\sum_{a=1}^{A}\mathrm{Var}(F_a) = \sum_{a=1}^{A}\lambda_a \qquad \cdots\cdots (A.3)$$

其中：

$$\mathrm{Var}(F_1) \geqslant \mathrm{Var}(F_2) \geqslant \cdots \geqslant \mathrm{Var}(F_a) < 0, \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_a < 0。$$

即第一主成分 F_1 提取的信息量最大，第二主成分 F_2 次之，依次减小。也就是说方差 $\mathrm{Var}(F_a)$ 越大、特征值 λ_a 就越大，主成分 F_a 所提取的信息量就越多。根据以上对应关系，第 a 个主成分 F_a 所提取的信息也可使用本征值来评估。通常，以 A 个主成分所携带的方差总和 $[\sum_{a=1}^{A}\mathrm{Var}(F_a)]$ 占原数据总方差（$\sum_{j=1}^{n}Q_j^2$）的比重，即主成分的累计方差贡献率（cumulative percent variance，CPV，也称累计贡献率）来表示 A 个主成分 $F_1, F_2, \cdots\cdots, F_A$ 概括原样品光谱矩阵中包含信息的精度，CPV_A 按式（A.4）进行计算：

$$\mathrm{CPV}_A = \frac{\sum_{a=1}^{A}\mathrm{Var}(F_a)}{\sum_{j=1}^{n}Q_j^2} = \frac{\sum_{a=1}^{A}\lambda_a}{\sum_{j=1}^{n}Q_j^2} \qquad \cdots\cdots (A.4)$$

如果 A 个主成分的累计贡献率不低于某个比率(如 85%),那么,A 个主成分 F_1,F_2,…,F_A 可以不低于 85%的精度来概括原样品光谱矩阵中包含的信息,所以,累计方差贡献率 CPV 常用作选择适宜主成分数的一个统计指标。

通过主成分分析,就可建立各个类(如第 q 类)的主成分类模型为:

$$\boldsymbol{X}_q = \overline{\boldsymbol{X}}_q + \boldsymbol{T}_q \boldsymbol{P}_q^{\boldsymbol{T}} + \boldsymbol{E}_q \qquad \cdots\cdots (A.5)$$

式中:

$\overline{\boldsymbol{X}}_q$——第 q 类经中心化处理后所得的矩阵均值,每一行均相同,且等于 $\boldsymbol{X}_q$ 所有行的平均值;

$\boldsymbol{T}_q$——第 q 类的的得分矩阵;

$\boldsymbol{P}_q$——第 q 类的荷载矩阵;

$\boldsymbol{E}_q$——第 q 类的残差矩阵。

如果提取第 q 类光谱矩阵 $\boldsymbol{X}_q$ 中的每一个量测值 x_{ij}^q 的主成分分析结果 $\hat{x}_{ij}^q$,那么,x_{ij}^q 的主成分类模型可按式(A.6)表示:

$$x_{ij}^q = \overline{x}_j^q + \hat{x}_{ij}^q + e_{ij}^q = \overline{x}_j^q + \sum_{a=1}^{A_q} t_{ia}^q p_{aj}^q + e_{ij}^q \qquad \cdots\cdots (A.6)$$

式中:

$\overline{x}_j^q$——第 q 类变量 j 的均值;

$\hat{x}_{ij}^q$——类模型对量测值 x_{ij}^q 解释的结果,等于 $\sum_{a=1}^{A_q} t_{ia}^q p_{aj}^q$;

A_q——第 q 类中的适宜主成分数;

t_{ia}^q——第 q 类中样品 i 在第 a 个主成分上的得分值;

p_{aj}^q——第 q 类中变量 j 在第 a 个主成分上的荷载值;

e_{ij}^q——第 q 类中样品 i 的变量 j 的残差值,等于($x_{ij} - \hat{x}_{ij}$)。

在主成分分析中,得分矩阵隐含着训练样品的分布,即样品与样品之间的亲疏关系,在主成分空间中,通过互相正交的得分矢量的投影可反映出这类样品之间的关系;荷载矩阵反映了主成分与原始光谱变量之间的相互关联程度,包含光谱变量之间的相互影响。残差矩阵包含样品光谱信息被类模型拟合后留下的、没有被类模型解释的噪音,也称残余方差。残余方差越小,表示类模型提取的光谱信息越多,遗留下的光谱噪音越小,也就是说,类模型的解释能力越强。那么,如何衡量类模型的解释能力,如何评估样品之间亲疏程度,统计识别异常样品,优化主成分数,选择适宜主成分数构建理想的类模型,这是应用 SIMCA 分类法要解决的问题。

A.1.3 主成分数的优化

选取适宜的主成分数 A 用于建立类模型是关键的一个步骤。没有一个严格快捷的规则用于主成分数的选择。通常,如果所用主成分数过少,则可能未充分利用信息,模型欠拟合,导致类模型判别能力下降。如果使用主成分数过多,则可能会引入一些多余的噪声,模型过拟合,导致类模型不稳健,微小的噪音也可能会导致类模型判别能力产生显著性变化。为了实现主成分空间与残差空间的最优分离,宜使用交互验证(cross-validation)方法来评估和选择建立类模型的主成分数。在交互验证过程中,从训练集中剔除一个样品(也称去一法或留一法)后,使用剩余的样品建立类模型并预测被剔除样品的残差。将剔除的样品重新放回训练集,然后再从训练集中剔除另外一个样品,重复循环以上过程直至所有训练样品都被剔除过一次。那么,交互验证的训练集样品的预测残差平方和 PRESS(predictive residual error sum of squares) 按式(A.7)进行计算:

$$\mathrm{PRESS} = \sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} (x_{ij} - \hat{x}_{ij})^2 \qquad \cdots\cdots (A.7)$$

在确定适宜主成分数的过程中,通常从使用第一个主成分数开始,从训练集剔除一个样品后,使用

训练集中剩余的样品建立类模型并预测被剔除样品的残差。然后将被剔除的样品放回训练集，对其余样品重复迭代此过程，直到所有的训练样品被剔除过一遍，计算 PRESS 值。依次增加主成分数建模，重复以上过程，计算使用不同的主成分数建模的 PRESS 值，直到预设的主成分数才停止。把 PRESS 值作为主成分数的因变量进行计算作图，并通过 F 检验来判断 PRESS_A 与 $\mathrm{PRESS}_{(A+1)}$ 是否存在显著性差异，引入的 F 检验统计量按式(A.8)计算：

$$F=\frac{n}{n-A-1}\left(\frac{\mathrm{PRESS}_A-\mathrm{PRESS}_{(A+1)}}{\mathrm{PRESS}_{(A+1)}}\right) \quad \cdots\cdots(\mathrm{A.8})$$

F 检验的风险水平可设为 5%，临界值 $F_{0.05}[n,(n-A-1)]$ 在自由度 $[n,(n-A-1)]$ 下，从 F 分布临界值表中查出。如果 F 值小于临界值，说明 PRESS_A 与 $\mathrm{PRESS}_{(A+1)}$ 两者无显著性差异，增加主成分数对提升模型预测精度的贡献不显著。那么，适宜的主成分数就为 A。使用适宜主成分数建立类模型，训练集样品的光谱信息可被类模型最大程度拟合。

如果训练集中存在异常样品，会造成 PRESS 明显波动，导致误选主成分数。所以，在交互验证过程中，应识别并剔除异常样品。

在一般情况下，前三个主成分对应于最大特征值的得分矢量所包含样品光谱的信息量最大，所以，常用前三个主成分的得分可视化表示样品之间的亲疏程度。

通常可采用交互验证有效性 CV^2 来衡量类模型的预测精度，设某一类光谱 $\boldsymbol{X}$ 的每一个变量 x_k，交互验证有效性 CV^2 和累计交互验证有效性 $\mathrm{CV}^2_{\mathrm{cum}}$ 分别按式(A.9)和式(A.10)进行计算：

$$\mathrm{CV}^2=\left(1.0-\frac{\mathrm{PRESS}_A}{\mathrm{PRESS}_{(A-1)}}\right)_k \quad \cdots\cdots(\mathrm{A.9})$$

$$\mathrm{CV}^2_{\mathrm{cum}}=\left[1.0-\prod\left(\frac{\mathrm{PRESS}_A}{\mathrm{PRESS}_{(A-1)}}\right)_{ka}\right] \quad [a=1,\cdots,A] \quad \cdots\cdots(\mathrm{A.10})$$

一般 $\dfrac{\mathrm{PRESS}_A}{\mathrm{PRESS}_{(A-1)}}$ 的值小于 1.0，且越小，CV^2 值就越大，模型的预测精度就越高。尚若类模型中存在异常样品，PRESS 值就会偏高，影响类模型的拟合精度，所以，为了提升主成分类模型的拟合精度，在交互验证过程中，经常伴随着应用合适的统计手段识别和剔除异常样品。

A.1.4 异常样品的统计与识别

异常样品是由自身的组成结构特殊或仪器工作状态异常导致其光谱不具代表性，远离训练集整体平均水平的样品，对适宜主成分数的选择以及对模型稳健性有强烈的干扰，且具有较强的相互掩蔽性。在主成分分析中，常用 Hotelling's T^2 检验、F 检验从主成分空间和残差空间中统计识别异常样品。

Hotelling's T^2，简称 T^2，是一个常见的统计量，在多元假设检验中有重要作用。对某类 m 个样品光谱进行主成分分析，如果主成分数为 A，从得分矩阵中提取第 i 个样品光谱的主成分得分，则 T^2 按式(A.11)进行计算：

$$T_i^2=\sum_{a=1}^{A}\frac{t_{ia}^2}{s_{ta}^2} \quad \cdots\cdots(\mathrm{A.11})$$

式中：

t_{ia}——第 i 个样品光谱在第 a 个主成分上的得分。

s_{ta}^2——第 i 个样品光谱在第 a 主成分上得分向量 t 的方差。

通过 Hotelling's T^2 检验，就可以识别第 i 个样品在主成分空间中是否远离总体平均水平，为了便于使用，基于 T^2，引入 F 检验统计量，F 按式(A.12)进行计算：

$$F=\frac{m(m-A)}{A(m^2-1)}T_i^2 \quad \cdots\cdots(\mathrm{A.12})$$

在一般情况下，F 检验的风险水平设为 5%(或 1%)，临界值 $F_\alpha(A,m-A)$ 在自由度 $(A,m-A)$

下，从 F 分布临界值表中查出。如果

$$T_i^2 > \frac{A(m^2-1)}{m(m-A)}F_a(A,m-A)(\alpha=5\%,\text{或}\ 1\%) \quad \cdots\cdots(\text{A.13})$$

则第 i 个样品光谱在 95%(或 99%)的置信范围之外，该样品不具代表性，可视为异常样品剔除。

如果从残差矩阵中提取第 i 个样品的残余方差，记为 Q_i^2，则 Q_i^2 按式(A.14)进行计算：

$$Q_i^2 = \sum_{j=1}^{n}\frac{e_{ij}^2}{n-A_q} \quad \cdots\cdots(\text{A.14})$$

式中：

e_{ij}^q——第 q 类样品 i 的变量 j 的残差值。

A_q——第 q 类的主成分数。

那么，对于 m 个训练样品构成的整个 q 类，其平均总残余方差，记为 Q^2，则 Q^2 按式(A.15)进行计算：

$$Q^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{e_{ij}^2}{(m-A_q-1)(n-A_q)} \quad \cdots\cdots(\text{A.15})$$

总残余方差可用于评估整个 q 类训练样品在模式空间中的聚散程度，平均总残余方差越小，训练样品的聚类程度越高，类模型所拟合的训练样品的特征的集中度也就越高。

于是，可采用 F 检验判断样品 i 的残余方差 Q_i^2 与整个类的总残余方差 Q^2 的差异性来识别样品是否为异常样品。结合式(A.14)和式(A.15)得式(A.16)：

$$F = \frac{Q_i^2}{Q^2} \quad \cdots\cdots(\text{A.16})$$

在一般情况下，F 检验的风险水平可设为 5%(或 1%)。临界值 $F_a[(n-A),(m-A-1)(n-A)]$ 在自由度$[(n-A),(m-A-1)(n-A)]$下，从 F 分布临界值表中查出。如果

$$Q_i^2 > Q^2F_a[(n-A),(m-A-1)(n-A)][\alpha=5\%(\text{或}\ 1\%)] \quad \cdots\cdots(\text{A.17})$$

则第 i 个样品光谱在 95%(或 99%)的置信范围之外，该样品 i 远离类模型，样品 i 不归属该类，可视为异常样品剔除。

要注意的是，实因不同的软件在设计上的差异，用于统计识别异常样品光谱的统计量也会存在差异，在应用 F 检验时，自由度的选取也就不同。在主成分分析中，很多软件常把主成分得分可视化，展示在主成分空间中样品与样品之间的亲疏程度，选择具有得分、残差可视化相结合功能的化学计量学软件，有利于从 T^2-Q^2 分布分析样品交叉情况，便于识别和剔除异常样品。尽管对异常值的统计与识别方法研究得比较广泛，但尚无通用的方法可以遵循，如果利用其他或更新的方法能有效的统计和识别异常样品，则这些方法均可使用。

A.1.5 主成分类模型的评估

A.1.5.1 主成分类模型的测定系数

类模型测定系数，记为 R^2X，是指类模型可解释的变异占总变异的比重，是评估类模型解释能力的一项统计指标。如果类模型对量测值 x_{ij} 的解释结果为 $\hat{x}_{ij}$，m 个样品的 $\hat{x}_{1j}$，$\hat{x}_{2j}$，$\hat{x}_{3j}$，…，$\hat{x}_{mj}$ 的均值为 $\overline{x}_j$，则 R^2X 按式(A.18)进行计算，累计解释能力 R^2X_{cum} 按式(A.19)进行计算：

$$R^2X = 1-\frac{\text{RSS}}{\text{SSX}} = 1-\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\hat{x}_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\overline{x}_j)^2} \quad \cdots\cdots(\text{A.18})$$

$$R^2X_{\text{cum}} = \sum_{a=1}^{A}(R^2X)_a \quad \cdots\cdots(\text{A.19})$$

式中：

SSX——光谱变量的总离差平方和，即中心化处理后光谱变量的总变异平方和，等于 $\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\overline{x}_j)^2$；

RSS——光谱变量的残差平方和，等于 $\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\hat{x}_{ij})^2$。

由 R^2X 定义可知，RSS 越小，R^2X 就越大，类模型解释光谱信息能力越强，模型拟合精度也就越高。如果采用 A 个主成分数建模，那么，最终类模型的累计解释能力来自于不同主成分类模型解释能力的贡献。若使用主成分数过多，导致模型过拟合，虽然 R^2X 会增大，但类模型不稳健，微小的噪音会导致类模型判别能力产生显著性变化。所以，在 SIMCA 分类法中，采用适宜的主成分数建立稳健的主成分类模型是比较关键的一个步骤。

A.1.5.2 类间距

在训练集存在多个类别（如 p 类与 q 类）的模式识别中，类与类之间的距离，也称类间距，是衡量类与类分离程度的统计量。p 类与 q 类之间的距离 D_{pq}（或 D_{qp}）按式（A.20）进行计算：

$$D_{pq}=D_{qp}=\sqrt{\frac{Q_{pq}^2+Q_{qp}^2}{Q_p^2+Q_q^2}}-1 \qquad \cdots\cdots(A.20)$$

式中：

Q_{pq}^2——q 类的模型拟合 p 类中各样品得到的残余方差；

Q_{qp}^2——p 类的模型拟合 q 类中各样品得到的残余方差；

Q_p^2——p 类样品的总残余方差；

Q_q^2——q 类样品的总残余方差。

类间距越大，类与类的差异性越明显，类与类的分离程度就越高，所建立的类模型的判别能力就越强。在模式识别中类与类要获得理想的分离程度，类间距的经验值应不低于 3。

A.1.5.3 识别率与判别正确率

在训练类模型的过程中，常用主成分数作为自变量对识别率（因变量）作图来估计适宜主成分数的选择，但这仅是初步评估类模型理论判别能力的一个方面，通常，识别率大于判别正确率，判别正确率是衡量类模型的实际判别能力，判别正确率难以到达 100%的状态。所以，一种相对合理的评估办法是使用者首先根据实际需要，设定一个合适的预期的判别正确率作为评估类模型有效性的重要指标，然后结合识别率来综合评估类模型判别能力。

当然，也可应用一种比较系统的方法——混淆矩阵（confusion matrix）来进一步分析类模型理论判别类别与实际类别的对应关系，从而评价类模型的判别能力。

A.1.6 类模型的验证

就是通过验证集来验证类模型，如果类模型判别验证样品，得到的判别正确率不低于使用者预期设定的判别正确率，则类模型通过有效性验证，类模型有效。反之，重新训练类型。

A.1.7 应用

当 SIMCA 分类法的类模型通过有效性验证完毕之后，就可应用于预测待测样品的归属。但要注意的是，在应用 F 检验时，风险水平一般设为 5%或 1%，使用者也可根据自己的实际需要，在训练类模型时预先自定义合适的风险水平。

A.1.8 应用 SIMCA 分类法的流程

应用 SIMCA 分类方法的一般流程为：

a） 对训练集样品光谱数据进行预处理，如中心化、求导和平滑等；

b） 选取某一类样品进行 PCA 建模；

c） 识别并剔除异常样品，重复运算、交互验证，确定 PCA 类模型的适宜主成分数；

d） 验证类模型的有效性；

e） 重复 b)～d)，以建立其他类的类模型。

A.2 PLS-DA 分类方法

A.2.1 PLS-DA 分类方法简要解释

PLS-DA 以一种基于偏最小二乘回归分析的有监督的模式识别方法进行分类，它利用先验分类知识，将类别作为分类变量（因变量）量化，本标准推荐本类取值为 1，非本类取值为 0 的量化方法。然后将光谱变量与分类变量进行 PLS 校正，建立 PLS-DA 类模型，最后利用类模型判别待测样品的归属。

A.2.2 PLS-DA 类模型的建立

假设在 m 个样品构成的光谱数据矩阵 $\boldsymbol{X}_{m\times n}$ 中，包含 A，B，C，…，N 个类别，每个类别又分别包含若干个个样品，将 m 个样品的类别设定为分类变量，并进行量化，本类取值为 1，非本类取值为 0，那么，由 1 和 0 构成的分类变量矩阵 $\boldsymbol{Y}_{m\times N}$ 表示，见式（A.21）：

$$\boldsymbol{Y}_{m\times N}=\begin{bmatrix} A\text{ 类} & B\text{ 类} & & & N\text{ 类} \\ 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}\begin{matrix} \\ \left.\begin{matrix}\\ \\ \\ \\ \end{matrix}\right\} A\text{ 类} \\ \left.\begin{matrix}\\ \\ \\ \\ \end{matrix}\right\} B\text{ 类} \\ \\ \left.\begin{matrix}\\ \\ \\ \\ \end{matrix}\right\} N\text{ 类} \end{matrix} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots (\text{A}.21)$$

将 $\boldsymbol{X}_{m\times n}$ 与 $\boldsymbol{Y}_{m\times N}$ 通过 PLS 进行校正，在校正过程中，不但对光谱矩阵进行正交分解，同时也对分类变量矩阵进行正交分解，并求出模型系数矩阵，见式（A.22）～（A.24）：

$$\boldsymbol{X}=\boldsymbol{TP}+\boldsymbol{E}_{\boldsymbol{X}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (\text{A}.22)$$

$$\boldsymbol{Y}=\boldsymbol{UQ}+\boldsymbol{E}_{\boldsymbol{Y}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (\text{A}.23)$$

$$\boldsymbol{U}=\boldsymbol{TB} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (\text{A}.24)$$

式中：

$\boldsymbol{T}$ ——$\boldsymbol{X}_{m\times n}$ 主成分分析的得分矩阵；

$\boldsymbol{P}$ ——荷载矩阵；

$\boldsymbol{E}_{\boldsymbol{X}}$ ——残差矩阵；

$\boldsymbol{U}$ ——$\boldsymbol{Y}_{m\times N}$ 主成分分析的得分矩阵；

$\boldsymbol{Q}$ ——荷载矩阵；

$\boldsymbol{E}_Y$ ——残差矩阵；

$\boldsymbol{B}$ ——相关系数矩阵。

PLS有不同算法，如非线性迭代算法、基于矩阵运算的SIMPLS算法等，不同计算方法的结果相同。

若判断某一类待测样品 $\boldsymbol{X}_{un}$ 的类别时，将 $\boldsymbol{X}_{un}$ 进行主成分分解得得分矩阵 $\boldsymbol{T}_{un}$，按式(A.25)就可计算出该类别的分类变量值：

$$\boldsymbol{Y}_{un}=\boldsymbol{T}_{un}\boldsymbol{B}\boldsymbol{Q} \qquad \cdots\cdots(A.25)$$

已设定量化分类变量本类为1，非本类为0，那么，分类变量值是否接近1，就是一个衡量类模型分类效果的重要指标。对每一个待测样品的分类变量值(记为 y)取整之后，根据其大小，可以准确判别待测样品的类别是否属于本类，判别规则为：①当 $y>0.5$，且偏差<0.5 时，判定样品属于本类；②当 $y<0.5$，偏差<0.5 时，判定样品不属于本类；③当 y 的偏差$\geqslant 0.5$，判定不稳定。

计算中也可以按照类别将分类矩阵 $\boldsymbol{Y}_{m\times N}$ 拆分成 N 个分类向量，依据式(A.22)～式(A.24)分别建立类模型，并根据公式(A.25)计算每类的分类变量值，得出所有类别的分类变量值。不同的化学计量学软件在算法设计、分类变量取值(有用数字0，1，2，…，N 表示类别的量化方法)上会存在一些差异，使用者在应用PLS-DA分类方法时，首先要了解软件的设置说明。

在构建PLS-DA类模型过程中，主成分分析仍然是核心，如何评估类模型的判别能力，如何诊断异常样品，选择适宜主成分数建立判别能力强的类模型，这仍然是应用PLS-DA分类法要解决的重要问题。

A.2.3 主成分数的优化

预测残差平方和PRESS是优化主成分数常用的统计量，由于化学计量学软件在设计上的差异，功能全面的软件一般具备了既可根据光谱残差、又可按应变量残差计算PRESS值来选择主成分数训练模型。训练PLS-DA类模型，本质上是一个光谱变量与分类变量的PLS校正过程。宜以分类变量为主计算PRESS值，或是计算其他用于评估类模型的统计量，如模型决定系数(采用光谱变量计算称测定系数)、交互验证有效性 CV^2 等。应用PRESS统计量优化选择适宜主成分数的过程与确定PCA类模型适宜主成分数的过程基本一致，参见A.1.2。

在训练PLS-DA类模型的过程中，如何检验随主成分数递增PRESS值发生变化的显著性，除了使用 F 检验来判断 $PRESS_A$ 与 $PRESS_{(A+1)}$ 是否存在显著性差异来选择主成分数之外，常定义，当 $\dfrac{PRESS_A}{PRESS_{(A-1)}}\leqslant 0.95^2$ 时，带入式(A.9)计算得，$CV^2\geqslant 0.097\ 5$，增加主成分数，对提升类模型预测精度有贡献；反之，当 $CV^2<0.097\ 5$ 时，增加主成分数，对提升类模型预测精度无明显贡献，可参考这一经验规则，来辅助选择适宜主成分数。

A.2.4 异常样品的统计与识别

除了使用A.1.3的有关统计识别方法诊断、剔除异常样品之外，在训练PLS-DA类模型的交互验证过程中，当发现某一类训练样品的分类变量值不满足第一条判别规则(当 $y>0.5$，且偏差<0.5 时)时，均可视为异常样品剔除。

随着化学计量学的不断发展，研究PLS校正过程中异常值的统计识别的方法越来越多，如果采用其他方法能有效的诊断异常样品，则这些方法均可使用。

A.2.5 PLS-DA类模型的评估

A.2.5.1 PLS-DA类模型的决定系数

决定系数(R^2，determination coefficient)是评价校正模型拟合程度的一项指标，是指模型可解释的

变异占总变异的比重。如果 $\hat{y}_i$ 是某类训练集(包含 N 个样品)的预测值,$\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_N$ 的均值为 $\overline{y}$,y_i 为对应的分类变量(设定为 1),则决定系数按式(A.26)进行计算:

$$R^2=\frac{S_{SSR}}{S_{SST}}=\frac{S_{SST}-S_{SSE}}{S_{SST}}=\frac{\sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2}=1-\frac{\sum_{i=1}^{N}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2} \quad \cdots\cdots(A.26)$$

式中:

S_{SST}——分类变量的总离差平方和,等于 $\sum_{i=1}^{N}(y_i-\overline{y})^2$;

S_{SSR}——分类变量的回归平方和,等于 $\sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2$;

S_{SSE}——分类变量的残差平方和,等于 $\sum_{i=1}^{N}(\hat{y}_i-y_i)^2$。

其中,$S_{SST}=S_{SSR}+S_{SSE}$。

由 R^2 定义可知,S_{SSR} 和 S_{SSE} 越小,R^2 就越大,模型拟合精度也就越高。R^2 取值范围为 0~1。如果采用 A 个主成分数建模,那么,最终模型的累计解释能力来自于不同主成分的模型解释能力的贡献,最终的决定系数按式(A.27)进行计算:

$$R^2=R^2Y_{cum}=\sum_{a=1}^{A}(R^2Y)_a \quad \cdots\cdots(A.27)$$

若采用的主成分数越大,R^2 也随之增大,但这并不意味着分类变量就越接近设定值 1。通常,如果所用的主成分数过少,则未充分利用信息,模型欠拟合,R^2 会比较小;如果使用主成分数过多,虽然 R^2 接近 1,则可能会引入一些多余的噪声,导致模型过拟合。所以,采用适宜的主成分数建立稳健的类模型是比较重要的一个步骤。

A.2.5.2 预测残差平方和

在训练 PLS-DA 类模型的交互验证过程中,预测残差平方和(PRESS)在用于优化模型主成分数的同时,也是衡量类模型解释分类变量能力的重要统计量。按已设定的,已知某类别训练样品的分类变量均为 1,且训练集包含 N 个训练样品,每个分类变量的预测值以 $\hat{y}_{cvi}$ 表示,那么,交互验证的预测残差平方和 PRESS 按式(A.28)计算:

$$PRESS=\sum_{i=1}^{N}(\hat{y}_{cvi}-1)^2 \quad \cdots\cdots(A.28)$$

显然,在交互验证的过程中,若每一个训练样品分类变量的预测值越接近于 1,PRESS 也就越小,说明类模型对训练样品的解释能力也就越强。

若类模型中存在异常样品,会造成 PRESS 明显的波动。所以,在使用 PRESS 评估类模型之前,应识别并剔除异常样品。

A.2.5.3 识别率与判别正确率

在训练 PLS-DA 类模型的过程中,因为不满足第一条判别规则(当 $y>0.5$,且偏差<0.5 时)的训练样品均被视为异常样品剔除,所以,PLS-DA 类模型的识别率通常为 100%。而判别正确率一般难以到达 100%,使用者可首先根据实际需要,设定一个预期的判别正确率作为评估类模型有效性的重要指标,然后使用验证集检验类模型,计算判别正确率来评估类模型的实际判别能力是否达到使用者的预期。

A.2.6 类模型的验证

要求参见 A.1.6。

A.2.7 应用

当 PLS-DA 类模型通过有效性验证完毕之后，结合 A.2.2 中所述的判别规则，就可应用于预测待测样品的归属。

A.2.8 应用 PLS-DA 分类法的流程

应用 PLS-DA 分类方法的一般流程为：

a) 对训练集样品光谱数据进行光谱预处理，如中心化、求导和平滑等；

b) 应用 PLS-DA 训练类模型；

c) 统计识别并剔除异常样品，重复运算、交互验证，确定 PLS-DA 类模型的适宜主成分数；

d) 验证 PLS-DA 类模型的有效性。

A.2.9 应用 SIMCA 和 PLS-DA 分类方法的一般经验

因为 SIMCA 和 PLS-DA 这两种方法都是基于 PCA 降维的投影分类方法，选择适宜的主成分数建模至关重要，没有一个严格的规则用于选择适宜的主成分数，不同的数据分析处理软件，所采用的统计指标会存在一些差异，经验的操作是把 PRESS，CPV，本征值（特征值）等这些统计指标协同使用，综合分析评估后来选择主成分数。通常，当主成分数为 3 时，CPV 一般不低于 85%，当主成分数大于 15 时，模型过拟合风险增大，适宜的主成分数控制在 15 以下，定性分析的样品类别控制在 2～3 类比较合适。

从实践经验来看，对定性分析比较复杂的对象，若在使用者认定为合格的同一个类别中，存在样品质量波动范围大的情况，宜收集足够多的代表性样品，采用 SIMCA 分类法训练类模型。对于此种情况，如果采用 PLS-DA 分类方法则效果比较差，甚至不能建立 PLS-DA 类模型，PLS-DA 分类方法适用于类内样品质量波动幅度小、质量相对比较均匀的定性分析。

近红外定性分析过程是一个对复杂多维数据循环往复统计处理的协同过程，涉及的各种统计量、统计参数会做反复比较，数据的可视化功能起到重要的辅助作用，宜选用二维/三维（2D/3D）数据可视化功能丰富的软件，有利于分析处理数据。

附 录 B
（资料性附录）
F 分布临界值表

F 分布临界值表见表 B.1 和表 B.2。

表 B.1 F 分布临界值表 $P[F(n_1,n_2)>F_\alpha(n_1,n_2)]=\alpha=0.05$

n_2	n_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

表 B.2　F 分布临界值表 $P[F(n_1,n_2)>F_\alpha(n_1,n_2)]=\alpha=0.01$

n_2	n_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4 052	4 999	5 403	5 625	5 764	5 859	5 928	5 981	6 022	6 056	6 106	6 157	6 209	6 235	6 261	6 287	6 313	6 339	6 366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

附 录 C
（资料性附录）
常见的误差类型、来源和解决途径

从训练类模型到应用类模型和维护类模型是一个循环往复、不断优化的过程。在每个环节，应谨慎操作，减少或降低有可能存在的误差，使类模型应用获得理想的效果。常见的误差类型、来源和解决途径，见表 C.1。

表 C.1 常见的误差类型、来源和解决途径

误差类型	误差来源	解决方法（途径）
光谱误差	仪器性能差	定期测试仪器性能，监测仪器性能的变化；测定质量控制样品，监测仪器性能变化是否会影响判别结果
	吸收超出检测器响应范围	确认仪器响应范围；选择合适光程，使吸收值在线性响应范围内
	光谱采集条件不稳定	改进采集光谱的方法，或者使用光学特性一致性优良的光谱测量附件
	光谱仪部件被污染	检查光路窗口等，并做必要的清洁和仪器保养
样品误差	固体样品不均匀	改善混合方法和粉碎工序，提高混合均匀程度；将样品磨成粒径均匀的颗粒，粒径不低于 40 目。使用旋转样品池重复装样测量光谱，并将测量光谱进行平均。 对不能粉碎的样品，训练集增加质量波动在允差范围内的样品的比例
	样品的理化性质随时间改变	样品制备后，立即测定分析或对样品干冻储藏；建模时，剔除快速变化的光谱区域
	样品的物理状态不一致	测量训练样品、验证样品、待测样品的光谱时，控制样品粒径、环境温度的一致性
	液体样品中有气泡	对单相样品检查所需压力；检查比色池的流体特性
建模误差	光谱对类模型特征不灵敏	尝试其他光谱区域
	训练集中样品代表性差	依据训练集的选择方法，筛选具有代表性的训练样品
	训练集中有异常样品	依据异常样品诊断统计规则，剔除异常样品
	对基线漂移敏感	对数据进行预处理，降低基线漂移影响
	数据录入错误	重新核对数据
判别误差	类模型性能差	使用有代表性的验证集，检验已经建立的类模型
	仪器性能差	诊断仪器性能；用质量控制样品检查仪器性能
	模型传递差	验证模型传递和仪器标定过程

附 录 D
（资料性附录）
基于近红外光谱 PLS-DA 法判别三类性质相近药品实例

D.1 近红外光谱仪性能的诊断与校验

本例采用傅里叶变换近红外光谱仪进行试验。在试验之前，按仪器操作手册规定的要求，对仪器的性能进行诊断检验，保证光谱仪运行正常。

D.2 模式分类的目的与适用范围

应用近红外光谱对市面上常见药品阿奇霉素、琥乙红霉素、罗红霉素三类性质相近药品进行类别判别，有利于药品市场流通领域的规范监督。

D.3 样品准备

收集到国内市面常见厂家三类药品 217 个，其中阿奇霉素样品 85 个、琥乙红霉素样品 70 个、罗红霉素 62 个，样品用铝塑包装，在常温避光保存。

D.4 仪器工作参数设置和样品的近红外光谱采集

本实例采用带 InGaAs 检测器的傅里叶变换近红外光谱仪，包括光纤探头采样附件，以及仪器配套的采样、建模软件。采用漫反射模式，仪器的主要工作参数为：光谱扫描范围 12 000 cm^{-1}～4 000 cm^{-1}，分辨率 8 cm^{-1}，扫描次数 64 次。

在常温条件下，开机预热光谱仪 2 h 至稳定，仪器自校后调用编写好的工作流程，先采集背景光谱，然后将样品从铝塑包装中取出，使用光纤探头直接采集样品的傅里叶变换漫反射近红外光谱，每批测定 6 片，每片测一次，所得到的近红外光谱图，见图 D.1。

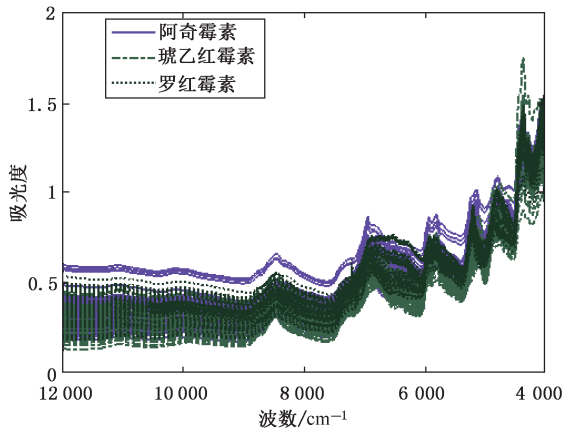

图 D.1 三类药品的近红外光谱示意图

D.5 训练样品和验证样品的选择

采用马氏距离法对 217 个样品进行计算，发现 2 个异常样品。剔除异常样品后，采用 kennard-stone 方法对样品光谱进行分组，训练集、验证集比例约为 2∶1，详细信息见表 D.1。

表 D.1 训练集和验证集样品的选择

类别	样品数	异常样品数数	训练集样品数	验证集样品数
阿奇霉素	85	2	55	28
琥乙红霉素	70	0	46	24
罗红霉素	62	0	41	21

D.6 数据预处理

由于样品颜色、形态、仪器性能等影响，近红外光谱往往伴随着噪声，对样品定性分析具有很大的影响。本实例采用 SG 一阶导数和平滑对光谱进行预处理，预处理后光谱图，见图 D.2。

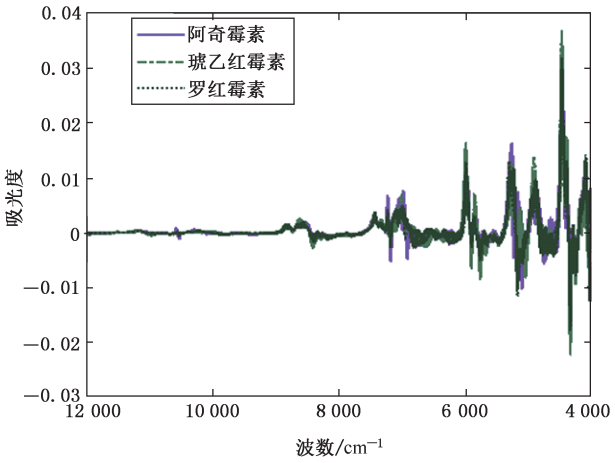

图 D.2 预处理后的样品近红外光谱示意图

D.7 类模型的建立

建模时，待建立类别对应样品的分类变量值设定为 1，非本类样品的分类变量值设为 0。在建模前对数据进行中心化处理。应用 PLS-DA 建模，图 D.3 显示了模型训练集样品交叉验证识别率随主成分数增加的变化图。3 类样品的识别率都随 PLS-DA 主成分数的增加而增加，阿奇霉素、琥乙红霉素、罗红霉素的主成分数分别在 3、3、4 以后稳定。本实例阿奇霉素、琥乙红霉素、罗红霉素模型的主成分数设定为 3、3、4。

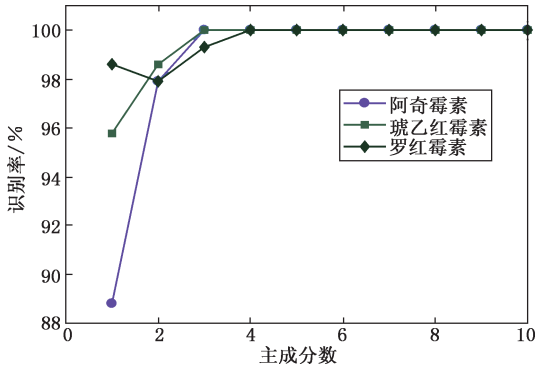

图 D.3　识别率随主成分数变化示意图

D.8　类模型的验证

使用验证集样品检验模型。根据本标准 PLS-DA 的判别原则，预测值大于 0.5 且偏差小于 0.5 定为本类样品；预测值小于 0.5，且偏差小于 0.5，定为非本类样品。图 D.4A～C 分别列出了类型分析模型证集样品的预测结果。可以直观地看出，验证集样品的判别正确率均为 100%。表 D.2 列出了阿奇霉素、琥乙红霉素、罗红霉素模型验证集样品预测值的最小值、最大值、平均值、判别正确率。

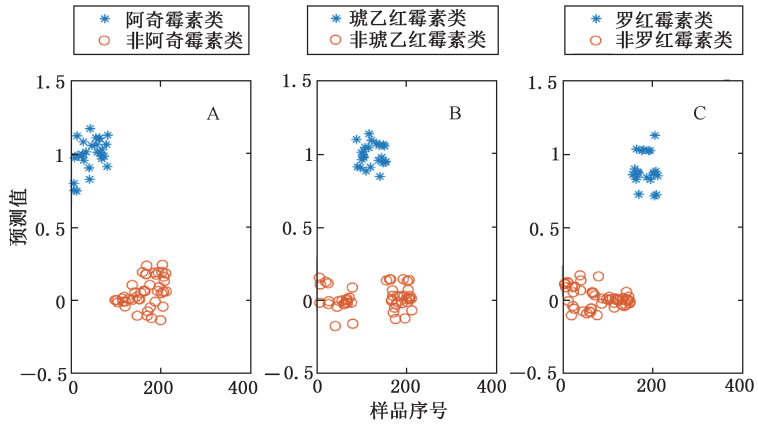

图 D.4　PLS-DA 类模型的预测结果示意图

表 D.2　验证集样品预测值统计情况

类别	样品数	最小值	最大值	均值	判别正确率/%
阿奇霉素	28	0.745	1.170	0.992	100
非阿奇霉素	44	−0.138	0.240	0.052	
琥乙红霉素	24	0.844	1.135	0.991	100
非琥乙红霉素	49	−0.178	0.151	0.012	
罗红霉素	21	0.711	1.124	0.889	100
非罗红霉素	51	−0.105	0.168	0.012	

本方法采用近红外光谱和 PLS-DA 方法建立的定性判别类模型能准确判别阿奇霉素、琥乙红霉素、罗红霉素等三类药品。该方法适用于药品质量监管部门、药品流通领域及第三方检测机构进行类别判属。

附 录 E
（资料性附录）
近红外透射光谱对汽油质量等级分类 SIMCA 法实例

E.1 近红外光谱仪性能的诊断与校验

本例采用傅里叶变换近红外光谱仪进行试验。在试验之前，按仪器操作手册要求，对仪器的性能进行诊断检验，保证整个光谱仪运行正常。

E.2 模式分类的目的与适用范围

应用傅里叶变换近红外透射光谱法对汽油进行质量等级分类。

E.3 样品准备

汽油样品包括从某炼厂采集的 92＃汽油和 95＃汽油。

E.4 仪器及参数设置

本实例采用带 InGaAs 检测器的傅里叶变换近红外光谱仪，包括能够适应样品瓶的采样附件以及仪器配套的采样、建模软件。采用仪器自带软件设置仪器工作参数，同时编写采集光谱数据的工作流程。采用透射模式，波数范围为 6 000 cm^{-1}～9 000 cm^{-1}，分辨率 8 cm^{-1}，扫描次数 64 次。

E.5 近红外光谱采集

在常温条件下，开机预热光谱仪 2 h，调用编写好的工作流程，先采集背景光谱，然后将样品装入 5 mm内径样品瓶，采集样品的近红外光谱。226 个样品的近红外区光谱图，见图 E.1(A)。

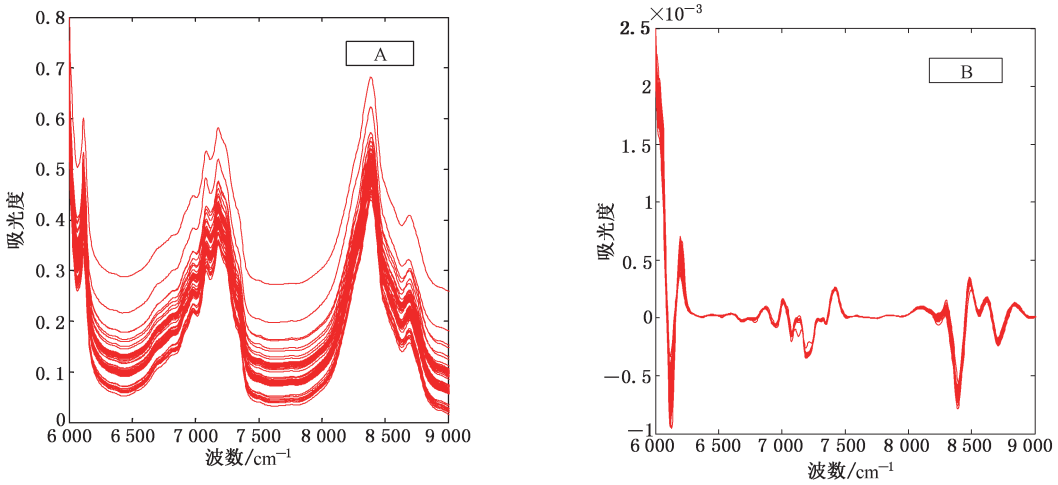

图 E.1 样品近红外光谱示意图(A)和预处理后的光谱示意图(B)

E.6 训练集和验证集样品选择

训练集和验证集样品选择见表 E.1。

表 E.1 训练集和验证集样品选择

样品类型	训练集样品数	验证集样品数
92#汽油	46	27
95#汽油	62	23

E.7 数据预处理

由于样品物理性质、仪器性能等影响，近红外光谱往往伴随着噪声，这些干扰对样品的定性分析具有很大的影响。本实例采用二阶微分对光谱进行了预处理，预处理后光谱图，见图 E.1(B)。使用全谱变量进行 SIMCA 分析可能会引入干扰变量，降低模型的预测能力。本实例选择了 6 000 cm^{-1}～9 000 cm^{-1}用于分析。

E.8 模型建立

针对两类汽油样品分别进行 SIMCA 建模，在建模前对数据进行中心化处理。应用 SIMCA 建模，主成分数是影响判别正确率大小的重要参数。图 E.2 显示了两种不同类型汽油样品训练集样品交叉验证的识别率随主成分数增加的变化图。从图中看出，两种汽油样品的识别率都随主成分数的增加而增加，在 10 以后，趋于稳定。本实例中对两类样品的主成分数均设定为 10，针对两类分别建立 SIMCA 模型。

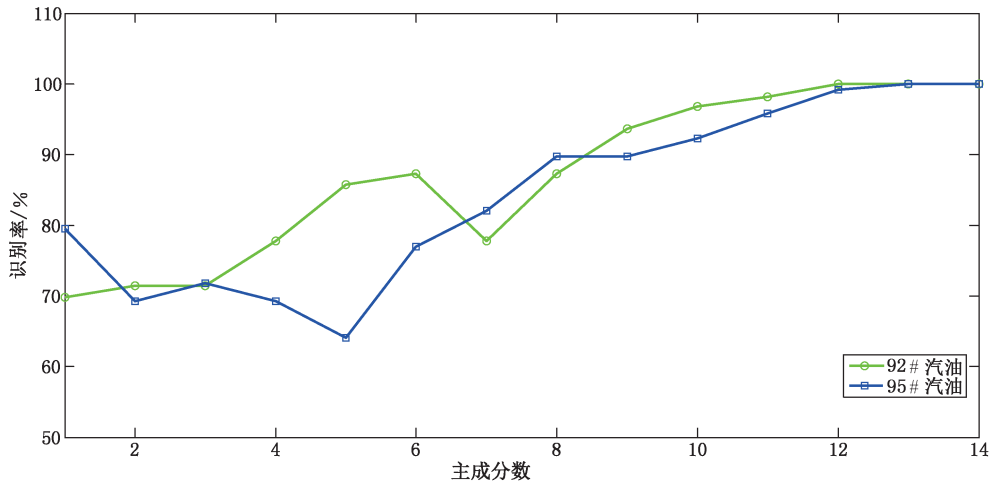

图 E.2 交叉验证误差随主成分数变化示意图

E.9 类间距评价

汽油样品训练集按两个类别建立模型进行分类，参照本标准附录 A 中的式(A.20)类间距计算公

式，得到两类之间类间距为：$D_{92\#汽油\text{-}95\#汽油}=138.9$，说明两类之间分类效果优良。

E.10 模型验证与优化

使用验证集样品检验模型。对于各个验证样品，F 值大于临界值的即为非本类样品。表 E.2 为92#汽油外部验证样品的识别结果。表 E.3 为 95#汽油外部验证样品的识别结果。相应的临界值 $F_{[0.05,(n-A),(n-A)\times(m-A-1)]}$（$n$，波长点，779；$A$，主成分数，10；$m$，样品数量，两种类型汽油样品数分别为46，62），查附录 B 中的表 B.1 F 分布临界值表（$\alpha=0.05$）得 1.00。

表 E.2 92#汽油外部验证样品的判别结果

样本序号	F	F_{cruit}	预测类别	实际类别	结果
1	0.07	1.00	92#汽油	92#汽油	正确
2	0.85	1.00	92#汽油	92#汽油	正确
3	0.01	1.00	92#汽油	92#汽油	正确
4	0.74	1.00	92#汽油	92#汽油	正确
5	0.03	1.00	92#汽油	92#汽油	正确
6	0.21	1.00	92#汽油	92#汽油	正确
7	0.56	1.00	92#汽油	92#汽油	正确
8	0.85	1.00	92#汽油	92#汽油	正确
9	0.08	1.00	92#汽油	92#汽油	正确
10	0.61	1.00	92#汽油	92#汽油	正确
11	0.66	1.00	92#汽油	92#汽油	正确
12	0.72	1.00	92#汽油	92#汽油	正确
13	0.22	1.00	92#汽油	92#汽油	正确
14	2.01	1.00	非 92#汽油	92#汽油	错误
15	0.32	1.00	92#汽油	92#汽油	正确
16	0.31	1.00	92#汽油	92#汽油	正确
17	0.62	1.00	92#汽油	92#汽油	正确
18	0.53	1.00	92#汽油	92#汽油	正确
19	0.45	1.00	92#汽油	92#汽油	正确
20	0.79	1.00	92#汽油	92#汽油	正确
21	0.47	1.00	92#汽油	92#汽油	正确
22	0.02	1.00	92#汽油	92#汽油	正确
23	0.17	1.00	92#汽油	92#汽油	正确
24	0.36	1.00	92#汽油	92#汽油	正确
25	0.12	1.00	92#汽油	92#汽油	正确
26	1.12	1.00	非 92#汽油	92#汽油	错误
27	0.10	1.00	92#汽油	92#汽油	正确

表 E.3　95#汽油外部验证样品的判别结果

样本序号	F	F_{cruit}	预测类别	实际类别	结果
1	0.10	1.00	95#汽油	95#汽油	正确
2	0.34	1.00	95#汽油	95#汽油	正确
3	0.77	1.00	95#汽油	95#汽油	正确
4	0.86	1.00	95#汽油	95#汽油	正确
5	1.39	1.00	非 95#汽油	95#汽油	错误
6	0.90	1.00	95#汽油	95#汽油	正确
7	0.11	1.00	95#汽油	95#汽油	正确
8	0.93	1.00	95#汽油	95#汽油	正确
9	0.05	1.00	95#汽油	95#汽油	正确
10	0.85	1.00	95#汽油	95#汽油	正确
11	0.88	1.00	95#汽油	95#汽油	正确
12	0.28	1.00	95#汽油	95#汽油	正确
13	0.54	1.00	95#汽油	95#汽油	正确
14	0.54	1.00	95#汽油	95#汽油	正确
15	1.07	1.00	非 95#汽油	95#汽油	错误
16	0.66	1.00	95#汽油	95#汽油	正确
17	0.08	1.00	95#汽油	95#汽油	正确
18	0.13	1.00	95#汽油	95#汽油	正确
19	0.06	1.00	95#汽油	95#汽油	正确
20	0.29	1.00	95#汽油	95#汽油	正确
21	0.10	1.00	95#汽油	95#汽油	正确
22	0.04	1.00	95#汽油	95#汽油	正确
23	0.04	1.00	95#汽油	95#汽油	正确

本方法采用傅里叶变换透射近红外光谱对不同等级汽油样品进行快速判别分析。在实际应用中可以用于不同等级汽油样品的快速甄别。