基于直方图分层映射的近红外光谱预处理算法

王丽杰[1,2], 杨羽翼[1,2], 代 敏[1,2], 高 玮[3]

[1]哈尔滨理工大学测控技术与通信工程学院, 黑龙江 哈尔滨 150080;

[2]哈尔滨理工大学测控技术与仪器黑龙江省高校重点实验室, 黑龙江 哈尔滨 150080;

[3]哈尔滨理工大学应用科学学院, 黑龙江 哈尔滨 150080

摘要 为解决近红外光谱快速检测乳品成分及含量时光谱数据的预处理问题,提出一种基于直方图分层映射技术的近红外光谱主成分得分重置(SR)预处理方法。以葡萄糖氯化钠水溶液三组分样品中的葡萄糖含量、鲜牛奶样品中的乳糖含量为定量检测目标,进行散射光谱主成分得分累计贡献率的分层分段规定化映射预处理,利用偏最小二乘(PLS)回归分析建模手段,对相应近红外光谱中的糖含量信息进行测试及分析。结果表明,经过 SR 预处理后,牛奶中乳糖含量 PLS 模型的校正集样品交互验证预测偏差降低 23.9%,实际预测偏差降低 27.8%;验证集实际预测偏差降低 16.7%。该 SR 光谱预处理方法兼顾光谱、参考值及组分相关性等多尺度信息,以实现光谱信息增强去噪,能避免有用信息误删,防止不充分拟合及过拟合。

关键词 光谱学; 光谱分析; 光谱预处理; 主成分得分重置; 直方图分层映射

中图分类号 O433.4 **文献标识码** A

doi: 10.3788/LOP54.093001

Near Infrared Spectral Pre-Processing Algorithm Based on Histogram Layering Mapping

Wang Lijie[1,2], Yang Yuyi[1,2], Dai Min[1,2], Gao Wei[3]

[1] *School of Measurement-Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin, Heilongjiang* 150080, *China*;

[2] *Higher Educational Key Laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, Harbin, Heilongjiang* 150080, *China*;

[3] *School of Applied Sciences, Harbin University of Science and Technology, Harbin, Heilongjiang* 150080, *China*

Abstract In order to solve the pre-processing problem of spectral data in near infrared rapid detection analysis of content of milk components, a pre-processing algorithm for score resetting (SR) of principal components (PC) in the near infrared spectrum on the basis of histogram layering mapping technology is proposed. With glucose content in the three components samples consisting of glucose, NaCl and water, and lactose content in the fresh milk samples, as the detecting objects, cumulative contribution rates of the near infrared scattering spectral PC scores are pre-processed by means of mapping by layer and by piece. Furthermore, partial least squares (PLS) regression analysis method is used for modeling, thereby test and analysis of sugar content information in corresponding near infrared spectra are completed. The results show that after SR pretreatment, the predicted deviation of the calibration curve of the milk lactose content PLS model is reduced by 23.9%, the actual prediction deviation is reduced by 27.8%, and the actual prediction deviation of the verification set is reduced by 16.7%. This SR spectral preprocessing method takes into account multi-scale information such as spectra, reference content value and component correlation to realize spectral information denoising enhancement. Therefore, false deletion of useful information can be avoided, and inadequate fitting and overfitting can be prevented.

Key words spectroscopy; spectrum analysis; spectral pre-processing; principal component score resetting; histogram layering mapping

OCIS codes 300.6340; 300.6170; 070.4790

收稿日期: 2017-01-19; **收到修改稿日期**: 2017-03-21

基金项目: 国家自然科学基金(11574065)

作者简介: 王丽杰(1971—),女,博士,教授,主要从事精密仪器及机械方面的研究。E-mail: wlj@hrbust.edu.cn

1 引 言

基于分子振动理论、化学计量学建模方法和近红外光谱采集仪器，采用近红外光谱分析手段对乳品等含氢基团的物质进行快速测定已成为大分子物质成分分析领域的一个研究方向[1-3]。然而，分子振动理论虽然比较成熟[4]，但由于被测物质不同，其光谱结构不同，相应倍频及合频振动的光谱特性也不同，加上近红外区域分子振动的倍频吸收和合频吸收强度弱、谱峰重叠，共存成分之间相关影响等等，导致光谱背景复杂、有用样品信息提取特别困难，如何通过适合的光谱预处理手段和化学计量学方法实现快速检测，一直是近红外光谱分析领域众多研究者多年来致力于解决的问题。

针对光谱数据预处理方法的探讨，近年来研究较多的是正交信号校正(OSC)法[5-6]及其类似算法[7]，但由于存在过拟合问题，虽然校正结果非常好，但预测结果较差，相比于多元散射校正(MSC)、平滑等常用预处理方法[8]，OSC 法虽提出了十多年却并未得到广泛的实际应用[9]。本文在正交法研究的基础上[10]，通过重置光谱主成分得分的累计贡献率来提高有用信息的主成分得分(权重)，采用直方图规定化的数学统计思想，实现近红外光谱有用主成分信息得分的映射增强处理，利用光谱主成分分层映射方式，实现光谱信息的增强去噪及避免过拟合或不充分拟合。

2 算 法

直方图规定化是为了增强低对比度而广泛采用的数字图像预处理技术[11]。近红外光谱谱峰重叠以及吸收强度弱，实质上就是光谱信息与干扰信息的对比度低的问题。大数据时代，无论是图像信息还是光谱信息，其本质和形式上都属于二维数字矩阵，所以可以采用相同的信息增强手段提高对比度、滤除干扰、提取有用信息，即光谱与数字图像的预处理方法具备同一性，只是由于应用领域不同而导致称谓、实现手段、评价指标等在形式上有不同[12-13]。因此，从这一层面来理解，无论是数字图像的二维矩阵，还是近红外光谱的二维矩阵，为了信息增强及去噪，都可以应用直方图增强的数学统计手段。所以，研究基于直方图规定化的近红外光谱得分重置(SR)预处理方法，借助于数字图像处理领域中目前较为成熟的直方图增强手段来实现光谱净化去噪。

2.1 理论依据及数学基础

直方图规定化的数学思想源于直方图统计学，它是通过有目的地增强某个灰度级分布范围内的图像，从而实现按照预先设定的形状调整图像直方图，达到增强对比度的目的。如果一幅图像在[0,1]区间内的灰度级是随机变量，每一个瞬间它们都是连续的，那么可以用 $P_r(r)$ 和 $P_z(z)$ 分别表示原始输入图像和希望得到的规定输出增强图像的概率密度函数，将灰度直方图从 $P_r(r)$ 变换到 $P_z(z)$ 的过程，就是直方图规定化的过程。

如果针对的是二维数字图像，则信息从一维连续信号转变为二维离散信号，这种情况下，原图像的像素值 r 的直方图描述为

$$h(r)=\sum_{p\in I}\delta(I_p-r), \tag{1}$$

式中 δ 为单位冲激函数，P 为像素位置，I 为像素邻域。

当像素值 $I_p=r$ 时，$\delta(I_p-r)=1$，当 $I_p\neq r$ 时，$\delta(I_p-r)=0$，这是输入图像的全局直方图的离散化表示形式。

把直方图看作一个一维列向量 $\boldsymbol{h}$，第 r 个元素 $h(r)$ 表示像素值 r 对应的像素数目，那么第 r 个元素的概率密度函数 $p_r(r)$ 可通过除以图像像素总数获得。所以概率密度函数 $p_r(r)$ 为

$$p_r(r)=\frac{h(r)}{\mathbf{1}^{\mathrm{t}}h}, \tag{2}$$

式中 $\mathbf{1}$ 表示列向量，其所有元素都是 1。

于是可得像素值 r 对应的累积概率分布函数 $T(r)$ 为

$$T(r)=\sum_{i=0}^{r}p_r(i)。 \tag{3}$$

数字图像增强中，通常取累积概率分布函数作为从输入图像到输出图像的灰度映射函数：

$$s_k=T(r_k)=\sum_{i=0}^{k}P_r(r_i)=\sum_{i=0}^{k}\frac{n_i}{n},\ r_k\geqslant 0, s_k\leqslant 1, k=0,1,\cdots,M-1, \tag{4}$$

式中 M 为原图像的灰度级数，r_k 为原图像中的灰度级，s_k 为规定图像中的灰度级，n_i 为相应灰度级出现的次数，n 为灰度级总数。

对规定图像也进行上述处理，即将每个灰度级 z_l 对应的累积概率分布函数作为其灰度映射变换函数：

$$u_l = G(z_l) = \sum_{j=0}^{l} P_z(z_j) = \sum_{j=0}^{l} \frac{n_j}{n},\ u_l \geqslant 0, z_l \leqslant 1, l = 0,1,\cdots,N-1, \tag{5}$$

式中 G 为规定图像的灰度映射函数，N 为规定图像的灰度级数，z_l 为规定图像中的灰度级，u_l 为对规定图像进行映射处理后图像的灰度级。

直方图规定化过程中，存在一个将输入图像的直方图转换为规定的输出图像直方图的映射关系，该映射函数的选择至关重要，其实质是一个度量优化问题：

$$s = T(r) = \min_{s=T(r)} D(R,Z), \tag{6}$$

式中 T 是所寻求的最优单调变换，R 与 Z 分别是结果直方图和目标直方图，D 是某种相似性度量。

采用单映射规则通过 s_k 和 u_l 的值建立 r_k 与 z_l 的映射关系，将原始直方图灰度级映射到规定的直方图上，由此求出规定化后的结果：

$$z(r) = \arg\min_{z} \left| T(r) - G(z) \right|,\ \forall r \in F, \tag{7}$$

式中 F 表示输入灰度级的范围[0,1]。

实际工程中，灰度映射函数通常采用曲线拟合的方法实现。上述直方图规定化数学统计思想构建了 SR 预处理方法的理论基础[14]。

2.2 方法研究及算法设计

充分借鉴目前国内外基于直方图规定化算法的新图像增强思想[15-16]，采用主成分(PC)分层分段映射法实现近红外光谱的 SR 和光谱重组。其中的关键技术就是目标规定直方图的确定和映射准则的选择，为了形象说明，绘制出其原理示意图，如图 1 所示，其中，PC_i 表示第 i 个主成分，其值定义为 c_{PC_i}，S_{PC_i} 表示第 i 个主成分得分的累计贡献率。

1) 主成分与得分

主成分分析(PCA)实质是数据降维，通过转换原变量，使得数目较少的新变量变成原变量线性组合，且新变量携带原变量数据结构特征及其信息。对于二维光谱矩阵，对其进行中心化得 $\boldsymbol{X}_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n1} & \cdots & x_{nm} \end{bmatrix}$，假定 $\boldsymbol{X}_{n\times m}$ 的秩为 $r[r < \min(n,m)]$，将 $\boldsymbol{X}$ 写成 q 个秩为 1 的矩阵之和：

$$\boldsymbol{X} = \boldsymbol{z}_1 + \boldsymbol{z}_2 + \cdots + \boldsymbol{z}_i + \cdots + \boldsymbol{z}_q, \tag{8}$$

式中秩为 1 的矩阵 $\boldsymbol{z}_i$ 可以表示为两个向量之间的外积运算，即：$\boldsymbol{z}_i = \boldsymbol{t}_i \boldsymbol{p}_i'$，其中 $\boldsymbol{t}_i$ 即为得分，维数为 n，而 $\boldsymbol{p}_i$ 即为载荷，维数为 m。

$\boldsymbol{z}_i$ 维数为 $n \times m$，与上面 $\boldsymbol{X}$ 的维数相同，变换(8)式得：

$$\boldsymbol{X} = \boldsymbol{t}_1 \boldsymbol{p}'_1 + \boldsymbol{t}_2 \boldsymbol{p}'_2 + \cdots + \boldsymbol{t}_i \boldsymbol{p}'_i + \cdots + \boldsymbol{t}_q \boldsymbol{p}'_q。 \tag{9}$$

进一步用矩阵形式表示，如图 1(a)所示为

$$\boldsymbol{X}_{n\times m} = \boldsymbol{T}_{n\times q} \boldsymbol{P}'_{m\times q}, \tag{10}$$

式中 $\boldsymbol{T}$ 代表得分矩阵，$\boldsymbol{P}$ 代表载荷矩阵，n 是光谱矩阵中被测样品总数，m 是光谱矩阵中波长个数，q 是主成分总数。

剖析光谱因素，包括成分以及成分间相互作用、光谱采集过程中光电传感器噪声等的影响、测量环境的影响、光路杂散光的影响等。理想情况下，研究人员或者测量者希望校正集样品光谱的最大变化都是由于被测样品中待测成分的含量变化所引起的。但实际光谱变化则是由上述各种光谱因素的综合影响产生的，每种光谱都可看作是用一个权重(即所谓得分)去乘以对应的“纯光谱”而得到的。换句话说，将上述所有的这些“纯光谱”乘上其相对应的权重后再进行相加运算，那么，就能够得到重新建立后的重组光谱[4]，如(9)式向

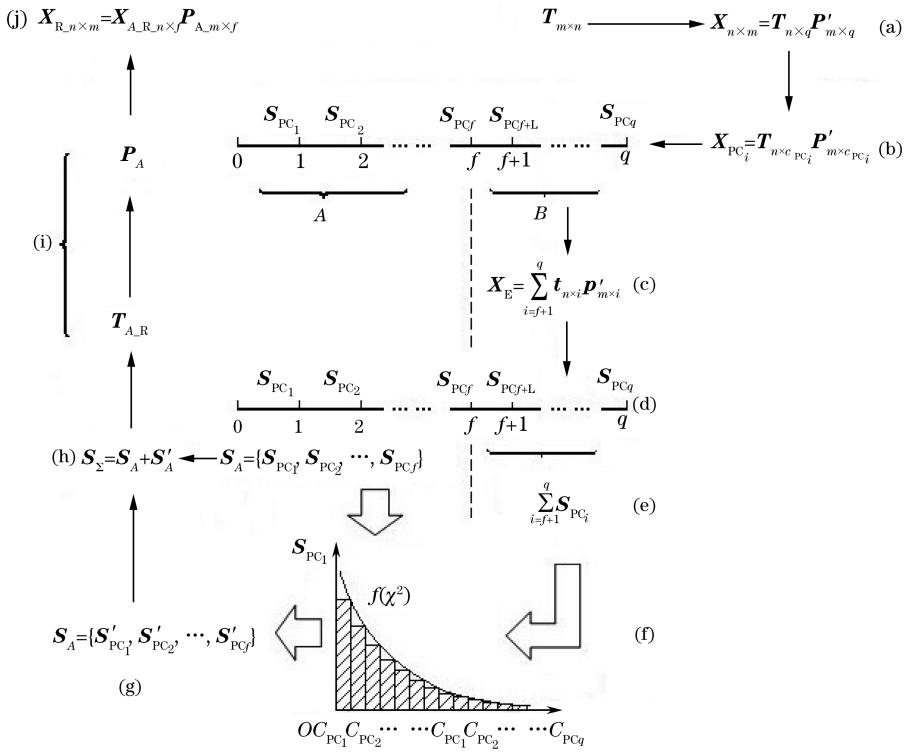

图 1 SR 法示意图。(a)主成分分析;(b)分段映射;(c)分层光谱合并;(d)计算各子层光谱得分的累计贡献率;(e) B 段子层光谱得分累计贡献率求和;(f)确定灰度映射规则;(g)灰度映射结果;(h)重新计算 A 段子层光谱得分累计贡献率;(i)光谱重置结果;(j)重组光谱

Fig. 1 Diagram of SR method. (a) PCA; (b) mapping by piece; (c) merging spectra by layer; (d) calculating the cumulative contribution rate of score for each sub spectrum; (e) sum of cumulative contribution rates of score for each sub spectrum within B piece; (f) determining gray scale mapping rules; (g) gray scale mapping results; (h) recalculating the cumulative contribution rate of score for each sub spectrum within A piece; (i) spectral resetting results; (j) rebuild the spectrum

量形式和(10)式矩阵形式所示。综上所述,将主成分分析与直方图规定化相互关联,取其共性数学基础,将光谱信息中主成分分析后得到的主成分(即纯光谱或载荷)与直方图中像素的灰度级值相对应,将得分(权重)的累计贡献率与直方图中各灰度级出现的概率相对应,通过采用直方图规定化的数学思想的进行光谱有用信息的得分贡献率规定化处理。

2) 光谱分层处理

用数学上的向量形式进行解析,"纯光谱"$\boldsymbol{p}_i$ 就是特征向量,即载荷,就是所谓的主成分;而前面所述的权重 $\boldsymbol{t}_i$ 就是得分。如果用矩阵形式说明,载荷矩阵就是 $\boldsymbol{T}_{n\times q}$,得分矩阵就是 $\boldsymbol{P}_{m\times q}$。

上述主成分彼此独立正交。第一主成分,也就是 PC_1,反映的是对应的原变量最大方差变化;PC_2,即第二主成分,次之;PC_3,即第三主成分,更次之;以此类推。即前边的主成分包含了最原始的光谱矩阵的绝大部分有用信息,而后边的主成分则往往与噪声或者干扰影响等因素有关。所以,基于主成分分析中的每个主成分分别代表的是原始光谱中包含的不同因素的贡献,研究中利用直方图分层映射方法的数学思想,进行近红外光谱主成分的分层处理。依据各个主成分及其得分,对原光谱执行分层计算,得到和第一主成分 PC_1、第二主成分 PC_2,…,第 q 个主成分 PC_q 分别对应的多个子层光谱,如图 1(b)所示:

$$\boldsymbol{X}_{PC_i}=\boldsymbol{T}_{n\times c_{PC_i}}\boldsymbol{P}'_{m\times c_{PC_i}},i=1,2,\cdots,q。\tag{11}$$

3) 得分分段规定化映射

将上述所有子层光谱划分为两段区域,即区间集合 A 和区间集合 B,利用分段直方图规定化的方法[14],执行分层后光谱得分 A、B 两段的分段规定化处理。其中 $A=\{c_{PC_1},c_{PC_2},\cdots,c_{PC_f}\}$,$B=\{c_{PC_{f+1}},c_{PC_{f+2}},\cdots\}$,如图 1(b)所示。

研究中,分段阈值取最佳主成分数 f。预先设定主成分总数为 q,采用偏最小二乘(PLS)回归分析方法对被测样品中各成分的含量建模。为防止不充分拟合或过度拟合,仅选择前边少数与被测成分含量有关的主成分参加建模,以排除噪声干扰,利用交互验证法求解确定最佳主成分数 f[4]。

①计算累积概率 $\sum_{i=f+1}^{q}\boldsymbol{S}_{\mathrm{PC}_i}$。

割掉 $\{c_{\mathrm{PC}_i}\in B\,|\,f+1\leqslant i\leqslant q\}B$ 段内子层光谱,进行分层光谱合并,得割舍光谱 $\boldsymbol{X}_{\mathrm{E}}$,如图 1(c)所示:

$$\boldsymbol{X}_{\mathrm{E}}=\sum_{i=f+1}^{q}\boldsymbol{t}_{n\times i}\boldsymbol{p}'_{m\times i}。\tag{12}$$

计算各子层光谱得分的累计贡献率,如图 1(d)所示,得到 A 段内 $\{c_{\mathrm{PC}_i}\in A\,|\,1\leqslant i\leqslant f\}$ 集合:

$$\boldsymbol{S}_A=\{\boldsymbol{S}_{\mathrm{PC}_1},\boldsymbol{S}_{\mathrm{PC}_2},\cdots,\boldsymbol{S}_{\mathrm{PC}_f}\}。\tag{13}$$

进一步地,将 B 段内子层光谱得分累计贡献率的值相加,得到 $\sum_{i=f+1}^{q}\boldsymbol{S}_{\mathrm{PC}_i}$,如图 1(e)所示。

②采用曲线拟合的方法进行灰度映射。

考虑到近红外光谱本身的复杂性,研究中根据各子层光谱得分累计贡献率的实际情况,采用散点图绘制和曲线拟合的方法统计,推断各子层光谱得分的累计贡献率的分布类型,确定灰度映射规则,如图 1(f)所示。

将 B 段内的累积概率 $\sum_{i=f+1}^{q}\boldsymbol{S}_{\mathrm{PC}_i}$ 按照上述分布规律在 A 段内 $\{c_{\mathrm{PC}_i}\in A\,|\,1\leqslant i\leqslant f\}$ 进行数值分配,得到光谱得分的附加的累计贡献率,如图 1(g)所示:

$$\boldsymbol{S}'_A=\{\boldsymbol{S}'_{\mathrm{PC}_1},\boldsymbol{S}'_{\mathrm{PC}_2},\cdots,\boldsymbol{S}'_{\mathrm{PC}_f}\}。\tag{14}$$

③SR。

将 A 段内子层光谱相应得分的原始的累计贡献率与附加的累计贡献率分别相加,得到重置后的得分累计贡献率:

$$\boldsymbol{S}_{\sum}=\boldsymbol{S}_A+\boldsymbol{S}'_A,\tag{15}$$

$\boldsymbol{S}_{\sum}$ 即为处理后按照上述映射关系得到的规定“直方图”,如图 1(h)所示。

④光谱重组。

将割舍光谱 $\boldsymbol{X}_{\mathrm{E}}$ 信息按照(12)式所示附加贡献率比例进行重置分配,分别加和于 A 段内各子层光谱信息中,得到 $\boldsymbol{T}_{A\_\mathrm{R}}$,如图 1(i)所示。

⑤重组光谱 $\boldsymbol{X}_{\mathrm{R}}$。

将 $\boldsymbol{T}_{A\_\mathrm{R}}$ 和 A 段内各主成分 $\boldsymbol{P}_A$ 重新计算,得到重组光谱 $\boldsymbol{X}_{\mathrm{R}}$,如图 1(j)所示:

$$\boldsymbol{X}_{\mathrm{R}\_n\times m}=\boldsymbol{T}_{A\_\mathrm{R}\_n\times f}\boldsymbol{P}'_{A\_m\times f}。\tag{16}$$

3 实 验

实验样品采用以下两种:一种是自行配制葡萄糖氯化钠水溶液三组分样品总计 92 个,其中校正集样品 82 个,验证集样品 10 个,葡萄糖质量浓度的参考值采用配样时的计算值;葡萄糖的质量浓度和氯化钠的质量浓度分布梯度如图 2 所示。另一种是由某牛奶中心提供的鲜牛奶样品总计 120 个,其中校正集样品 86 个,验证集样品 27 个,选用样品覆盖了多区域多头奶牛在哺乳产奶期间的多个典型品种(图 3),乳糖质量浓度参考值采用 FOSS 傅里叶红外全谱扫描乳品成分快速分析仪 MilkoScan FT120 测定;乳糖成分的质量浓度变化范围如图 4 所示。

样品光谱的采集利用实验室声光可调谐滤光器型(AOTF)近红外光谱测试系统测试,波数范围为 10000~5800 cm^{-1},根据 AOTF 频率设置对应选择波长位置 98 个,采用符号 W1~W98 表示。

4 分析与讨论

原始光谱均经中心化处理,光谱分析分两种情况进行:一种是针对三组分样品,另一种是多组分鲜奶样品。光谱分析中,光谱 SR 预处理算法采用 MATLAB 软件编程实现,主成分分析和 PLS 回归分析建模过程

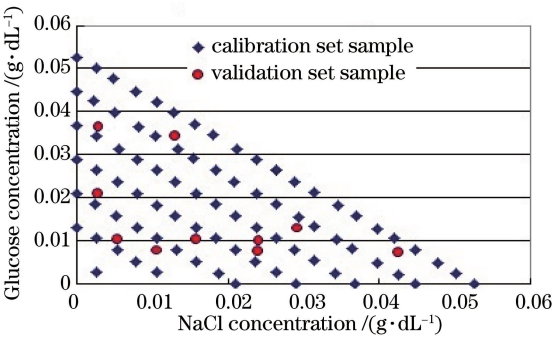

图 2 三组份样品浓度梯度分布

Fig. 2 Distribution of concentration gradient for three components samples

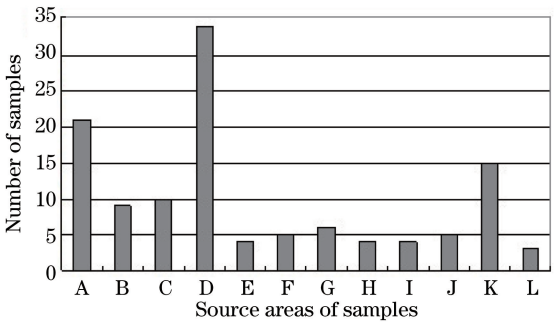

图 3 牛奶样品种类分布区域示意图

Fig. 3 Diagram of distribution areas for the types of milk samples

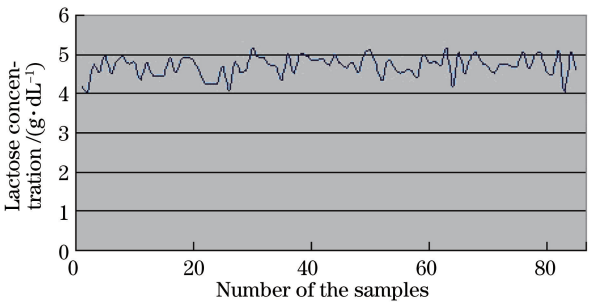

图 4 牛奶样品乳糖含量范围

Fig. 4 Lactose concentration range of milk samples

采用 CAMO 公司的 The Unscrambler 多元数据分析软件实现。相应 PLS 模型预测结果的评价指标，对校正集样品而言，采用测量值与参考值之间的交互验证均方差(RMSEC)，实际预测均方差(RMSEP)分别进行评价；对验证集样品，实际预测均方差用 RMSEP 评价。稳健性越高的模型，均方差越低。

4.1 三组分样品

针对三组分样品的散射光谱进行 SR 预处理，采用 PLS 方法建模预测分析。在葡萄糖氯化钠水溶液中，葡萄糖是红外活性分子，易溶于水[4]。因此，近红外光入射到样品溶液后，出射光能够携带葡萄糖分子的振动信息出来；氯化钠不含氢键，对近红外光没有吸收，所以添加氯化钠的目的就是将其作为干扰信息的一种模拟进行方法验证。

对三光谱进行主成分分析和 SR 预处理，并采用交互验证方式建立葡萄糖 PLS 模型。表 1 列出光谱主成分载荷和主成分得分的累计贡献率以及 SR 处理后主成分得分的累计贡献率。

表 1 主成分载荷和主成分得分的累计贡献率

Table 1 Cumulative contribution rates of PC loadings and PC scores %

PC	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
Cumulative contribution rates of PC loadings	39	53	5	1	1	0	0	0	0	0
Cumulative contribution rates of PC scores	45	15	23	10	2	1	1	2	1	0
Cumulative contribution rates of PC scores after SR	47	16	24	10	2	1	0	0	0	0

分析表 1 的主成分载荷累计贡献率的计算和统计分布结果，可以看出：PC_1 和 PC_2 载荷的累计贡献率显著高于其他主成分载荷。考虑到葡萄糖氯化钠混合水溶液样品中，氯化钠属于典型的非红外活性物质，三组分样品中能对红外光产生吸收的只有两种红外活性分子，一种是葡萄糖分子，另一种是水分子。所以，对应的三组分样品的光谱中，PC_1 和 PC_2 与葡萄糖和水的关联较大，选择最佳主成分数应该是 2。

为验证上述猜测，利用文献[17]中方法测试葡萄糖和水的吸光系数，并与主成分载荷累计贡献率进行相关分析，推出光谱各主成分所表征信息的归属，结果表明：第一主成分对应葡萄糖和水的吸光特征信息，其中占比重较大的是葡萄糖信息；第二主成分在一定程度上对应水和葡萄糖吸光特征信息，其中占比重较大的是

水信息;氯化钠组分本身属非红外活性分子,但由于其水溶液所体现的浓度梯度变化与水直接相关,所以,光谱的具体信息表征中,氯化钠组分在一定相对意义的程度上伪呈现出一些红外活性分子的表象[17]。

主成分对应的是载荷,得分对应的是权重,上述解析结果以及表 1 中 PC_1 和 PC_2 载荷的较高贡献率(39%和 53%),都在一定程度上表明三组分光谱的前三主成分是光谱结构中最重要的贡献信息,直接关联到被测组分含量信息。相对应的,表 1 中前三主成分对应的主成分得分累计贡献率(45%、15%和 23%)也明显高于其他主成分。

表 2 给出了三组分光谱的葡萄糖 PLS 模型校正集和验证集的预测结果。

表 2 葡萄糖 PLS 模型预测结果

Table 2 Predicted results of PLS model for glucose

Optimal PC number	Calibration set		Validation set
	RMSEC	RMSEP	RMSEP
6	0.128	0.140	0.354

从表 2 可以看出,最佳主成分数 $f=6$,校正集样品的 RMSEC 和 RMSEP 相对较小,结果比较理想。但是验证集的 RMSEP(0.354)则相对较大,说明建模过程中,一些与葡萄糖含量无关的主成分信息被过多地叠加到模型回归分析过程中,使得模型的实际预测能力有所降低,在一定程度上造成了过拟合,校正结果虽然较好,但实际预测结果不够理想。

为了降低过拟合,对中心化后的三组分样品光谱进行 SR 预处理。根据最佳主成分数确定 A 段为 $A=\{c_{PC_1},c_{PC_2},\cdots,c_{PC_6}\}$,$B$ 段为 $B=\{c_{PC_7},c_{PC_8},c_{PC_9}\}$。根据表 1 中主成分得分的累计贡献率数值,采用一元非线性回归分析方法拟合其数学模型,统计推断结果为自由度 $\nu=2$ 的 χ^2 分布。根据表 1 中主成分得分累计贡献率重置前数据,计算得 A 段内贡献率加和为 96%,B 段为 4%。割舍 B 段主成分信息,利用基于主成分得分的分层映射函数 $\chi^2(2)$ 分布,将其映射给 A 段各主成分,由此完成 SR 和光谱重组。重组光谱的 SR-PLS建模预测结果如表 3 所示。

表 3 葡萄糖 SR-PLS 模型预测结果

Table 3 Predicted results of SR-PLS model for glucose

Optimal PC number	Calibration set		Validation set
	RMSEC	RMSEP	RMSEP
6	0.126	0.142	0.347

表 3 的分析结果表明,经 SR 预处理后,SR-PLS 模型的校正集样品预测偏差比表 2 中 PLS 模型的略低,验证集 RMSEP 从原来的 0.354 降至 0.347,相对降低 2%。葡萄糖氯化钠水溶液自行配制三组分样品,成分单一,相对干扰较少,这也是 SR 光谱预处理具备优化效果但并不十分显著的原因。

光谱主成分即纯光谱信息决定了整个光谱结构,因此,光谱分析中任何主成分信息都不能无原则地随意增删。研究中,一方面仍然采用交互验证方法确定最佳主成分数为 6,以保证有足够的主成分纯光谱信息参与 PLS 建模预测,以避免有用信息误删、防止不充分拟合的发生;另一方面,在保持 $f=6$ 不变的前提下,进一步通过 SR 预处理,按照统计推断所得 χ^2 分布规律执行主成分得分累计贡献率值的分层映射重置处理,针对性地局部增强前述光谱有用信息 $PC_1\sim PC_3$ 得分(权重),使其累计贡献率从原来的 45%、15%、23%分别提升为 47%、16%、24%,而 $PC_4\sim PC_6$ 没有改变。SR 的上述处理手段,能够在不改变最佳主成分数的条件下,相对提升目标主成分 $PC_1\sim PC_3$ 的得分权重值、增强有用信息抑制干扰,降低过拟合风险。三组分样品光谱的实验和研究结果表明,无论是从理论上,还是从定性、定量分析角度,SR 光谱预处理思想、方法以及处理手段均具备可行性。

4.2 鲜牛奶样品

半透明悬浮类牛奶样品的组分较多,光进入样品后,同时会发生吸收、反射、漫反射等现象;光谱采集过程中,样品物化特性始终处于一种动态变化状态,因此,牛奶样品的光谱构成更加复杂。研究中,采用 SR 算法进一步对鲜牛奶样品散射光谱进行分析。SR 预处理前后,分别针对乳糖含量进行 PLS 建模,记作 PLS 模型和 SR-PLS 模型。采用平滑、MSC 等常用预处理手段以及目前用于牛奶近红外光谱预处理效果相对较

好的 OSC 方法，分别对中心化后的原始光谱进行处理并建模，记作 Smooth-PLS 模型、MSC-PLS 模型和 OSC-PLS 模型。对牛奶样品乳糖含量进行 SR 预处理和 PLS 建模，SR 预处理前后的乳糖含量，用不同 PLS 模型的预测结果如表 4 所示。

表 4 乳糖用不同 PLS 模型的预测结果

Table 4 Predicted results of different PLS models for lactose

Model	Optimal PC number	Calibration set		Validation set
		RMSEC	RMSEP	RMSEP
PLS	8	0.134	0.223	0.341
Smooth-PLS	8	0.161	0.227	0.339
MSC-PLS	8	0.106	0.198	0.326
OSC-PLS	3	0.098	0.145	0.298
SR-PLS	8	0.102	0.161	0.284

实验及数据分析结果表明：

1) 多组分奶样近红外光谱经过中心化处理后，相对于原始光谱，经平滑、MSC、OSC、SR 预处理后的 PLS 模型预测偏差的结果，在不同程度上均有所降低，说明上述预处理方法能够在不同程度上起到增强光谱信息、降低其他噪声影响的作用。与平滑、MSC 等成分分析中常用的预处理方法相比，SR-PLS 模型和 OSC-PLS 模型的预测偏差相对较低，OSC-PLS 最佳主成分数较少。SR-PLS 模型和 OSC-PLS 模型对校正集的预测偏差指标较接近，说明 SR-PLS 光谱分析方法具有可行性。SR-PLS 模型对验证集的预测偏差低于 OSC-PLS 的，说明 SR-PLS 方法可有效降低过拟合风险。

2) 相对于 PLS 模型，SR-PLS 模型校正集样品 RMSEC 从原来的 0.134 降低到 0.102，相对降低 23.9%，RMSEP 从原来的 0.223 降低到 0.161，相对降低 27.8%；验证集 RMSEP 从原来的 0.341 降低到 0.284，相对降低 16.7%。定量分析结果表明：SR 近红外光谱数据预处理方法既能保证足够多的主成分参与建模，避免有用信息误删，防止不充分拟合；又能在不改变最佳主成分数的前提下，相对提升目标主成分得分的权重，达到增强有用信息、抑制干扰、降低过拟合风险的预处理效果。SR 的这种基于主成分得分的分层映射规定性处理，正是直方图规定化统计思想在近红外光谱分析中的体现。

5 结 论

光谱数据预处理方法的探讨和研究，一直是乳品等含氢基团物质的成分浓度近红外光谱分析快速定量检测实现的技术瓶颈。基于直方图统计思想探讨近红外光谱 SR 预处理方法，通过光谱主成分得分累计贡献率的分层分段规定化映射，将直方图规定化方法从其常用数字图像处理领域平移至近红外光谱分析领域，探讨将直方图规定化的技术手段用于典型光谱定量分析(三组分样品和牛奶样品中糖类含量信息)的可行性，凭借光谱 SR 预处理方法的提出和研究，实现近红外光谱分析预处理方法的创新，对丰富近红外光谱分析手段具有重要的理论意义和实用价值。

参 考 文 献

[1] Wang Y W, Ding W, Kou L P, *et al*. A non-destructive method to assess freshness of raw bovine milk using FT-NIR spectroscopy[J]. Journal of Food Science and Technology, 2015, 52(8): 5305-5310.

[2] Chu Xiaoli, Lu Wanzhen. Research and application progress of near infrared spectroscopy analytical technology in China in the past five years[J]. Spectroscopy and Spectral Analysis, 2014, 34(10): 2595-2605.
褚小立, 陆婉珍. 近五年我国近红外光谱分析技术研究与应用进展[J]. 光谱学与光谱分析, 2014, 34(10): 2595-2605.

[3] Fu Bo, Hu Yongxiang, Liu Rong, *et al*. Near-infrared measurement with medium concentration sample as reference [J]. Acta Optica Sinica, 2016, 36(2): 0230003.
傅 博, 胡永翔, 刘 蓉, 等. 基于中等浓度样品参考测量的近红外光谱检测方法[J]. 光学学报, 2016, 36(2): 0230003.

[4] Lu Wanzhen. Modern near infrared spectra analysis technology[M]. Beijing: China Petrochemical Press, 2000.
陆婉珍. 现代近红外光谱分析技术[M]. 北京: 中国石化出版社, 2000.

[5] Andersson C A. Direct orthogonalization[J]. Chemometrics and Intelligent Laboratory Systems, 1999, 47(1): 51-63.

[6] Fearn T. On orthogonal signal correction[J]. Chemometrics and Intelligent Laboratory Systems, 2000, 50(1): 47-52.

[7] Li Yaping, Zhang Guangjun, Li Qingbo. Application of O2-PLS in experimental study on non-invasive measurement of blood glucose[J]. Acta Optica Sinica, 2010, 30(3): 854-860.
李亚萍, 张广军, 李庆波. 基于 O2-PLS 方法的血糖无损检测实验研究[J]. 光学学报, 2010, 30(3): 854-860.

[8] Lu Wanzhen, Yuan Hongfu, Chu Xiaoli. Near infrared spectrometer[M]. Beijing: Chemical Industry Press, 2010: 27-30.
陆婉珍, 袁洪福, 褚小立. 近红外光谱仪器[M]. 北京: 化学工业出版社, 2010: 27-30.

[9] Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004,16(4): 528-542.
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.

[10] Wang L J, Zhou Z, Qin Y, *et al*. Analysis of data mining method for near infrared spectral data of dairy products based orthogonality[C]. International Conference of Electronic Engineering and Information Science, 2014, 981: 641-646.

[11] Gonzalez R C, Woods R E. Digital image processing[M]. Ruan Qiuqi, Transl. 2nd ed. Beijing: Publishing House of Electronics Industry, 2007: 70-84.
Gonzalez R C, Woods R E Eddins S L. 数字图像处理[M]. 阮秋琦, 译. 2 版. 北京: 电子工业出版社, 2007: 70-84.

[12] Lim J, Kim G, Mo C, *et al*. Detection of melamine in milk powders using near-infrared hyperspectral imaging combined with regression coefficient of partial least square regression model[J]. Talanta, 2016, 151: 183-191.

[13] Yu Shimiao, Lu Wei, Liang Kun, *et al*. Study on prediction of germination rate of rice seeds using hyperspectral imaging combined with PCA and GRNN[J]. Laser & Optoelectronics Progress, 2015, 52(11): 113001.
于施淼, 卢 伟, 梁 琨, 等. 基于高光谱成像技术结合 PCA-GRNN 的糙米发芽率检测方法研究[J]. 激光与光电子学进展, 2015, 52(11): 113001.

[14] Chen Wenfei. Research of image enhancement based on probability theory[D]. Wuhan: Wuhan University, 2011.
陈文飞. 基于概率统计方法的图像增强研究[D]. 武汉: 武汉大学, 2011.

[15] Nie Chao. Efficient image enhancement algorithm research based on histogram[D]. Hangzhou: Hangzhou Dianzi University, 2014: 27-29.
聂 超. 基于直方图的高效图像增强算法研究[D]. 杭州: 杭州电子科技大学, 2014: 27-29.

[16] Liu Bo, Hu Zhengping, Wang Chengru. Multi-level interactive image enhancement algorithm based on fuzzy relaxation iterative procedure[J]. Optical Technique, 2009, 35(1): 131-134.
刘 博, 胡正平, 王成儒. 基于模糊松弛迭代的分层图像增强算法[J]. 光学技术, 2009, 35(1): 131-134.

[17] Wang Lijie. Research on rapid measuring method and system for milk constituents detecting by near infared spectrum[D]. Harbin: Harbin University of Science and Technology, 2006: 80-82.
王丽杰. 快速检测牛奶成分的近红外光谱测量方法及系统研究[D]. 哈尔滨: 哈尔滨理工大学, 2006: 80-82.