文章编号:1671-4598(2022)01-0229-05 DOI:10.16526/j.cnki.11—4762/tp.2022.01.035 中图分类号:TP391.9 文献标识码:A

基于 KOPLS 的多组分物质光谱分析方法

皮世威[1,2]，林 朝[2]，黄哲学[1]
(1. 深圳大学 计算机与软件学院，广东 深圳 518060；
2. 深圳市理邦精密仪器股份有限公司，广东 深圳 518067)

摘要：为了提升多组分物质浓度的光学测量精度，针对光谱与被测组分的非线性模型，提出了一种基于 KOPLS 算法的多组分物质光谱分析方法；该方法采用核矩阵将正交无关项转换至高维空间，通过迭代计算与剔除，建立了光谱信号与浓度矩阵之间的非线性回归模型，在保证算法高计算效率的同时解决了传统算法对非线性项分析准确度较低的问题，实现了对多组分物质光谱的高精度分析；通过实验对比了不同算法下的全血样本浓度预测值，实验结果表明 KOPLS 算法大幅提升了多组分物质浓度计算准确度，实验证明该方法在多组分检测仪器中具有很强的工程应用价值。

关键词：KOPLS；多组分物质；光谱分析；非线性回归

KOPLS—Based Spectral Analysis Method for Multi—Component Substances

PI Shiwei[1,2]，LIN Chao[2]，HUANG Zhexue[1]
(1. Shenzhen University，College of Computer Science & Software Engineering，
Shenzhen 518060，China；2. EDAN Instruments，Inc，Shenzhen 518067，China)

Abstract：In order to improve the optical measurement accuracy，a KOPLS—based spectrum analysis method for multi—component substances is proposed for the nonlinear model between the spectrum and the measured component. This method uses a kernel matrix to convert orthogonal irrelevant items into a high—dimensional space. Through iterative calculation and elimination，a nonlinear regression model between the spectral signal and the concentration matrix is established. It ensures the high computational efficiency of the algorithm and solves the problem of low non—linear regression accuracy in traditional algorithms. This method realizes high—precision spectral analysis of multi—component. Through experiments，the predicted concentration values of whole blood samples are compared under different algorithms. The experimental results show that the KOPLS algorithm greatly improves the prediction accuracy of the concentration of multi—component substances. The experiment proves that this method has strong engineering application value in multi—component measurement instruments.

Keywords：KOPLS；multi—component；spectral analysis；nonlinear regression

0 引言

在化学测量领域中，多组分物质的浓度光学测量是研究经典问题，其分析方法被广泛应用于临床检验中，对于医学判定和辅助诊断具有重要的作用。由于生物样本的复杂特性，化学计量的信号具有高噪声[1]、多重叠[2]、特征变量多的特点[3]，分析过程中包含大量的非线性拟合过程[4]，通常需要应用大量的回归分析来建立因变量与自变量之间的非线性关系。当前化学计量最常应用的回归方法有普通最小二乘(OLS)[5]，偏最小二乘法 (PLS)，正交偏最小二乘法 (OPLS)[6]等，其中普通最小二乘是经典的拟合

收稿日期:2021-11-16； **修回日期**:2021-11-24。
基金项目:深圳市战略性新兴产业发展专项(深发改[2018]1499 号)。
作者简介:皮世威(1988-)，男，湖南长沙人，博士，主要从事精密医疗检测器械方向的研究。
林 朝(1961-)，男，福建福州人，博士，教授，主要从事化学计量与传感器方向的研究。
黄哲学(1959-)，男，黑龙江大兴安岭人，博士，教授，主要从事大数据系统计算技术与应用方向的研究。
引用格式:皮世威，林 朝，黄哲学. 基于 KOPLS 的多组分物质光谱分析方法[J]. 计算机测量与控制，2022，30(1)：229-233，265.

方法。偏最小二乘回归线性拟合对于独立自变量多于因变量的案例具有较好的回归应用效果，可以有效剔除高维度变量中的无关成分[7]。正交偏最小二乘法(OPLS)[8]是在偏最小二乘的基础上通过建立自变量的正交映射，能够更为高效的去除变量中的无关信息，在迭代求解残差项中的正交量过程中，快速提取有效信息，对于复杂背景噪声下的变量具有明显的主成分提取作用，正交偏最小二乘法在代谢组以及化学分析领域中具有广泛的应用[9-10]，但该方法对于多变量应用下易发生过拟合现象，导致预测精度降低。文献［11］在 OPLS 拟合的基础上，通过引入 Kernel 核变量将原空间中的正交成分转化到特征空间，在高维空间完成正交无关项的预测与分离，该方法能够提升变量的非线性拟合精度，提高多变量的预测能力。Kernel 矩阵对于非线性拟合具有较高的拟合精度[12]，并且对于分析过程中的分类方式具有较好的可视化特性。KOPLS（核正交偏最小二乘法）的非线性拟合优势被应用于代谢组[13-14]和多组分浓度光谱分析中。尤其在多组分物质浓度测量中，由于被测样本成分复杂，光谱信号往往包含非线性噪声和干扰[15]，高精度的光谱分析建模方法成为预测准确度控制的重要因素。

1 建模方法

在多组分物质光谱分析应用中算法的核心思想是建立原数据矩阵 $\boldsymbol{X}$ 和浓度矩阵 $\boldsymbol{Y}$ 之间的映射模型。在保证模型准的预测准确度的前提下，同时还需要具有较快的计算速度。常用的建模算法有偏最小二乘(PLS)、正交投影映射（OPLS）等。

1.1 PLS 算法

偏最小二乘法 PLS 的核心思想是最大化自变量与因变量数据之间的协方差来解析自变量中的正交得分向量，首先对原数据矩阵 $\boldsymbol{X}$ 和浓度矩阵 $\boldsymbol{Y}$ 进行主成分分析：

$$\begin{cases}\boldsymbol{X}=TP+E\\ \boldsymbol{Y}=UQ+F\end{cases} \tag{1}$$

主成分个数由权重项 w 决定，分解得到各自对应的得分矩阵 $\boldsymbol{T}$、$\boldsymbol{U}$ 和载荷矩阵 $\boldsymbol{P}$、$\boldsymbol{Q}$，残差项分别为 E、F。通过主成分分析 $T=XP$，可以将原数据矩阵和浓度矩阵分别降至相应的低纬度空间，并保留原矩阵中的大部分有效信息。再对建立各自得分矩阵之间的线性回归方程：

$$\boldsymbol{U}=\boldsymbol{TB} \tag{2}$$

$\boldsymbol{B}$ 为回归系数矩阵。

$$\boldsymbol{B}=\boldsymbol{TU}(\boldsymbol{T}^T\boldsymbol{T})^{-1} \tag{3}$$

最后 $\boldsymbol{X}$ 相对与 $\boldsymbol{Y}$ 的线性回归可转变为 $\boldsymbol{X}$ 得分矩阵相对于 $\boldsymbol{Y}$ 的线性回归：

$$\boldsymbol{Y}=\boldsymbol{TBQ} \tag{4}$$

步骤 1～6 为标准的 NIPALS PLS 法

1. $w=\max<eig(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X})>$；
2. $w=w/\|w\|$；
3. $t=\boldsymbol{X}w/(w^Tw)$；
4. $q^T=t^T\boldsymbol{Y}/(t^Tt)$；
5. $u=\boldsymbol{Y}c/(c^Tc)$；
6. $p^T=t^T\boldsymbol{X}/(t^Tt)$；

1.2 OPLS 方法

OPLS 是在 PLS 的基础上建立的，通过筛选原数据矩阵 $\boldsymbol{X}$ 与浓度矩阵 $\boldsymbol{Y}$ 的不相关信息，使分类信息快速集中在主成分中，从而搭建简洁的 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的线性关系，这种建模方法适用于多元数据统计，OPLS 建模的具体步骤如下。首先和 PLS 一样，通过主成分分析法建立 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的线性组合：

$$\boldsymbol{X}=t_1\,p_1^T+t_2\,p_2^T+\cdots t_n\,p_n{}^T+E \tag{5}$$

$$\boldsymbol{Y}=u_1\,q_1^T+u_2\,q_2^T+\cdots u_n\,q_n^T+F \tag{6}$$

其中：t，u 可由权系数 w，c 求得：

$$t=\boldsymbol{X}w/\|w\| \tag{7}$$

$$u=\boldsymbol{Y}c/\|c\| \tag{8}$$

为使 t、u 之间的相关性最大，可以目标化 t、u 的协方差为最大，即：

$$\mathrm{Max}:Cov(t,u) \tag{9}$$

采用拉格朗日方法求解极值问题，w 为 $\boldsymbol{XTYYTX}$ 矩阵的最大特征值对应的特征向量，c 为 $\boldsymbol{YTXXTY}$ 矩阵的最大特征值对应的特征向量。随之即可求相应的得分向量 t，u。这样 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的载荷矩阵可通过关系式求得：

$$p^T=t^T\boldsymbol{X}/(t^Tt) \tag{10}$$

$$q^T=u^T\boldsymbol{Y}/(u^Tu) \tag{11}$$

计算 $\boldsymbol{X}$ 的正交权重向量 w_{orth}：

$$w_{\mathrm{orth}}=p-[w^Tp\ /(w^Tw]w \tag{12}$$

$$w_{\mathrm{orth}}=w_{\mathrm{orth}}/\|w_{\mathrm{orth}}\| \tag{13}$$

那么 $\boldsymbol{X}$ 正交矩阵的得分向量：

$$t_{\mathrm{orth}}=\boldsymbol{X}w_{\mathrm{orth}}/(w_{\mathrm{orth}}^Tw_{\mathrm{orth}}) \tag{14}$$

$\boldsymbol{X}$ 正交矩阵的载荷：

$$p_{\text{orth}}^T = t_{\text{orth}}^T \boldsymbol{X}/(t_{\text{orth}}^T t_{\text{orth}}) \tag{15}$$

求解正交残差项 EOPLS、FOPLS：

$$E_{\text{OPLS}} = \boldsymbol{X} - t_{\text{orth}} p_{\text{orth}}^T \tag{16}$$

将 $\boldsymbol{X}$ 替换为 EOPLS 因此则有：

$$\boldsymbol{X}_{\text{orth}} = T_{\text{orth}} P_{\text{orth}}^T \tag{17}$$

再对 $\boldsymbol{X}_{\text{orth}}$进行主成分分析：

$$\boldsymbol{X}_{\text{orth}} = T_{\text{o-pca}} P_{\text{o-pca}}^T + E_{\text{o-pca}} \tag{18}$$

由式（4）可将 $\boldsymbol{X}$ 相对与 $\boldsymbol{Y}$ 的线性回归可转变为得分矩阵 $\boldsymbol{X}_{\text{orth}}$ 相对于 Y 的线性回归。相对于 PLS，OPLS 能够对系统变量进行单独分析，通过去除正交无关量，可以降低过拟合发生的现象。但是当变量之间差异性较小，非线性耦合程度较高时，差异变量无法有效的被剔除，此时 OPLS 模型的计算准确度会降低。

1.3　基于 K－OPLS 的多组分物质分析方法

KOPLS 在 OPLS 算法的基础上保留了正交无关项的理念，并对建模方法做了进一步的改进，通过引入 Kernel 核矩阵来对数据中的非线性结构进行建模，同时仍可像 OPLS 一样对数据中的无关项进行筛选。在 KOPLS 算法中，通过对预测成分 T_p 和正交成分 Y_{orth}的建模来有效提取数据中的相关成分，这使得 KOPLS 模型的预测精度与基于核矩阵的偏最小二乘法（KPLS）[16]以及支持向量机模型[17]的预测精度保持一致，但 Kernel 矩阵通过在高维度的特征空间内对信号中的噪声进行转换建模，能够有效的消除原数据中由于外界因素带来的异常信息，例如测量仪器的信号漂移、样本中的生物耦合变量等。因此 KOPLS 对于非线性因素影响较大的生物化学多组分析具有较高的估算精度。

在 KOPLS 具体算法中，Kernel 矩阵的引入将原矩阵 $\boldsymbol{X}$ 中的变量转化为高维特征空间内的点积（$\boldsymbol{X}\boldsymbol{X}^T$），接着通过将 $\boldsymbol{X}\boldsymbol{X}^T$ 替换为 Kernel 矩阵 $\boldsymbol{K}$，Kernel 矩阵 $\boldsymbol{K}$ 中的元素 $K_{i,j}$由 X 的第 i 和第 j 行向量组成，Kernel 变换通过简洁的计算方式在将 X 映射到了高维空间中。因此 KOPLS 算法的第一步是选择合适的核函数，常用的核函数有线性、多项式和高斯核函数，其表达式分别为：

$$k(x,y) = x \times y \tag{19}$$

$$k(x,y) = (x^T y + 1)^P \tag{20}$$

$$k(x,y) = \exp(- \| x - y \|^2 / 2\sigma^2) \tag{21}$$

接下来的步骤是将原数据矩阵替换为核矩阵具体计算步骤如下。

①Kernel 矩阵中心化：

KOPLS 算法往往用于处理维度较大的数据，因此首先需要对 Kernel 矩阵进行中心化处理，中心化计算方法：

$$K = \left(I_n - \frac{1}{n} E_n E_n^T\right) K \left(I_n - \frac{1}{n} E_n E_n^T\right) \tag{22}$$

式中，En 为 $n \times 1$ 的向量，元素等于 1。

②求解权重向量：

建立 Kernel 矩阵 $\boldsymbol{K}$ 后，需确定数据中的正交成分个数 N。这样 $\boldsymbol{K}$ 表示为被剔除第 N 个正交成分后所组成的矩阵。接着通过对 YTKY 特征值分解求得权重向量 Cp 和 $\sum p$。

③求 $\boldsymbol{Y}$ 预测得分矩阵：

通过将 $\boldsymbol{Y}$ 映射到 Cp 上可求得 $\boldsymbol{Y}$ 的预测得分矩阵：

$$U_p = \boldsymbol{Y} C_p$$

④求 $\boldsymbol{X}$ 预测得分矩阵：

$\boldsymbol{X}$ 的预测得分矩阵：

$$\boldsymbol{T}_p = \boldsymbol{K}^T U_p \sum\nolimits_p^{(-1/2)}$$

⑤在正交成分个数 1 到 N 内，迭代循环：

对 $TpTQiTp$ 特征值分解，求得 $\boldsymbol{Y}$ 正交载荷向量 C_{orth}；

计算 $\boldsymbol{Y}$ 正交得分向量 $t_{(\text{orth}-i)} = \boldsymbol{Q}_i T_p C_{\text{orth}}$。

⑥对 t_{orth}抽取 Ki，得到 $Ki+1$；

此时预测得分矩阵：

$$\boldsymbol{T}_p = \boldsymbol{K}(i+1) U_p \sum\nolimits_p^{(-1/2)}$$

⑦最后建立回归方程：

$$\boldsymbol{B}_t = (\boldsymbol{T}_p^T \boldsymbol{T}_p)^{-1} \boldsymbol{T}_p^T \boldsymbol{U}_p$$

KOPLS 算法在预测项远大于测量项的应用下，具有较好的预测准确度，因此在非线性回归和分类应用较多的组分学中，KOPLS 的优势较为明显。例如对于使用光谱吸光度信号对样本中的多组分物质浓度分析时，样本本身复杂的络合状态往往伴随较大的非线性信号结构，KOPLS 中特有的将 $\boldsymbol{Y}$ 预测成分与正交无关成飞在特征空间中分离步骤，相对于 OPLS 和 PLS 都具有更好的预测执行能力。

KOPLS 在代谢组学研究中已有成功的应用。而对于多组分物质光谱分析应用场景，需要通过算法拟

合建立光谱与各组分物质浓度之间的非线性关系，由于需要从单个光谱中解析的多组分物质较多，而各组分之间特征耦合度高，且存在较多的非线性关系，KOPLS 的特点正好适用于此类场景的建模，模型拟合效果相对于 OPLS 和 PLS 更好。

2 实验结果

为了对比各算法在多组分物质光谱分析中的应用优劣，通过对多组分物质的浓度分析实验来评估。实验用多组分物质采用血液中的血红蛋白及其衍生物为样本。其测量方法基于分光光度法[18]，即由于多组分物质中各物质的吸收波长各不相同，因此根据朗博比尔定律：

$$A = k_1 c_1 l + k_2 c_2 l + \cdots + k_n c_n l \qquad (23)$$

其中：A 为总吸光度系数；k_i 为各组分物质的吸光度系数；c_i 为各组分物质的浓度；l 为测量光学量程。

由此可知对于不同浓度下的衍生物，其总吸光度也各不同，通过建立吸光度曲线与各组分物质浓度之间的关系，即可实现通过测量吸光度来完成对多组分物质的浓度检测，如图 1 所示。

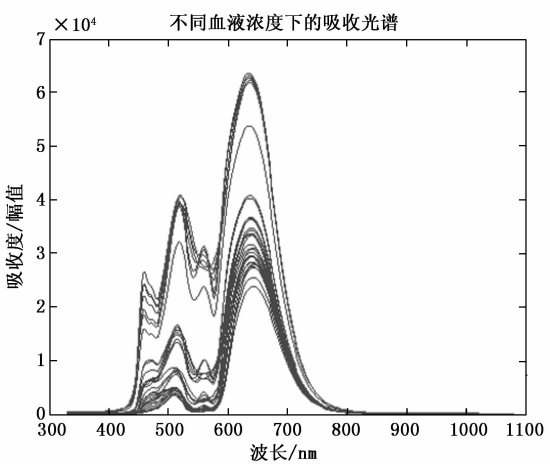


图 1 不同血液样本浓度下的吸收度光谱曲线分布图

实验首先通过 200 组数据进行建模，为保证原始数据的准确性，每组样本的光谱信号采用海洋光学光谱仪（QE65 Pro）进行采集，对样本数据搭建基于 PLS，KOPLS 模型的吸光度－多组分物质浓度的分析模型。再通过 111 组样本和靶标带入模型中进行计算，通过对比各自算法下的预测值与靶标值的差异，如图 2 所示。

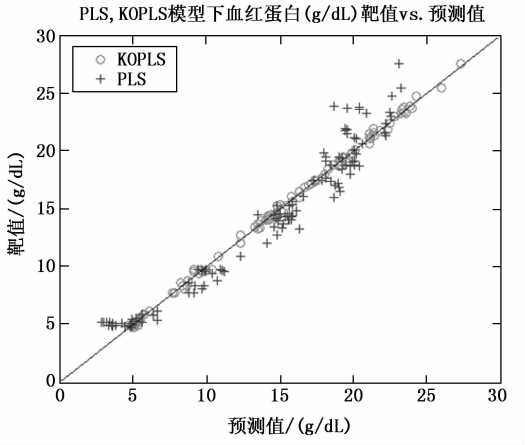


图 2 PLS，KOPLS 模型下的血红蛋白浓度预测结果对比

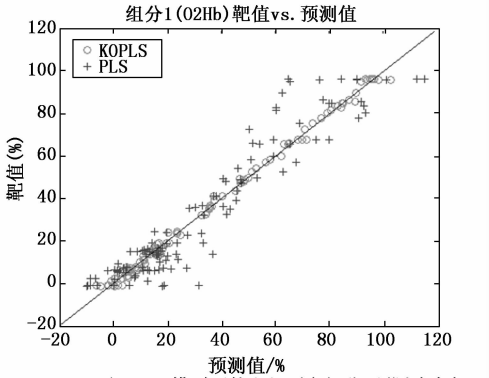
从计算结果图 2 可以看出，当采用 PLS 算法进行建模时，血红蛋白浓度预测值相对靶值的准确度达到±5.2 g/dL，在高浓度区间时，预测结果的离散度增大，这是由于被测样本为血液，其物质组成复杂，光谱分析过程包含过多的非线性因素。当采用 KOPLS 算法建模时，设定迭代剔除正交无关项 20 次，预测值相对靶值的准确度达到±0.6 g/dL，准确度得到较大的提升。对比结果如表 1 所示，实验结果说明 KOPLS 对于非线性因素较多的血液样本单一物质浓度分析具有较高的预测精度。

表 1 PLS，KOPLS 模型下的血红蛋白浓度预测准确度

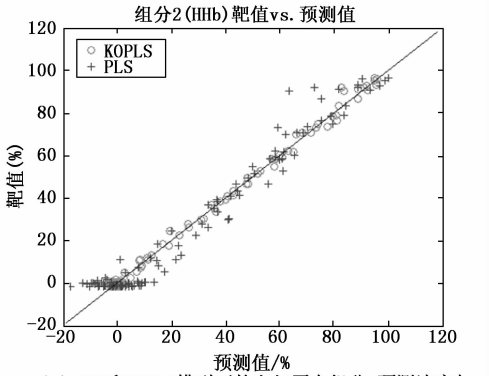
模型	准确度(g/dL)
PLS	5.2
KOPLS	0.6

对于血红蛋白中多组分物质浓度测量，光谱数据中各组分对应的特征量存在部分重叠。在建模过程中需要在不同血红蛋白浓度下配置不同梯度的多组分衍生物浓度。建模过程采用 PLS 模型和 KOPLS 模型，在两种模型下分别对 111 组样本进行预测，预测精度对比如图 3 所示。

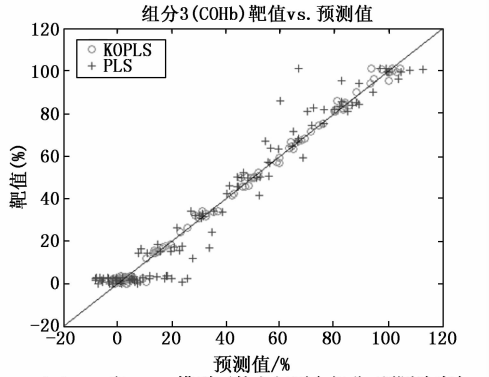
从计算结果图 3（a）以看出，当采用 PLS 算法进行建模时，血红蛋白组分 1 的浓度预测值相对靶值的准确度达到±32.6%，各梯度区域预测结果的离散度较大。当采用 KOPLS 算法建模时，预测值相对靶值的准确度达到±6.6%，准确度得到较大的提升，



(a) PLS和KOPLS模型下的血红蛋白组分1预测浓度与靶值浓度对比



(b) PLS和KOPLS模型下的血红蛋白组分2预测浓度与靶值浓度对比



(c) PLS和KOPLS模型下的血红蛋白组分2预测浓度与靶值浓度对比

图 3　PLS，KOPLS 算法建模下的多组分物质浓度预测结果对比

在不同梯度下的离散度也有了很大的降低。血红蛋白组分 2 的浓度预测值相对靶值的准确度达到±26.7%，结果如图 3（b）所示，各梯度区域预测结果的离散度同样较大。当采用 KOPLS 算法建模时，预测值相对靶值的准确度达到±9.3%，准确度得到了较大的提升，在不同梯度下组分 2 的离散度也有了大幅提升。血红蛋白组分 3 的浓度预测值相对靶值的准确度如图 3（c）所示达到±34.5%，各梯度区域预测结果的离散度较大。当采用 KOPLS 算法建模时，预测值相对靶值的准确度达到±9.48%，准确度和离散度同样得到较大的提升，对比结果如表 2 所示。对比结果说明 KOPLS 对于预测血液样本中的多组分物质浓度具有较高的预测精度，测量结果相对于 PLS 有明显的提升。

表 2　PLS，KOPLS 模型下的血液多组分物质浓度预测准确度

组分物质	预测准确度(单位：%)	
	PLS	KOPLS
组分 1(O2Hb)	32.6	6.6
组分 2(HHb)	26.7	9.3
组分 3(COHb)	34.5	9.48

3　结束语

本研究通过算法推导阐述了由 PLS 到 OPLS，再到 KOPLS 算法的演变过程。KOPLS 算法保留 OPLS 的正交映射思想，通过剔除正交无关量，快速提取原数据矩阵中的有效特征，建立原数据与变量之间的映射关系。同时通过 Kernel 变换，将原数据矩阵转化为高维特征空间的内积，建立原数据与变量之间的非线性关系。通过对血液样本的吸收光谱和多组分物质浓度进行 KOPLS 建模于预测计算，结果表明 KOPLS 对于具有大量非线性关系的血液多组分物质浓度分析具有明显的预测优势。这些特点可以在多组分物质浓度检测设备中得到应用。

参考文献：

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

[2] 许　禄，邵学广. 化学计量学方法 [M]. 北京：科学出版社，1995.

[3] 周　佳，周宗芳，刘雪琴. 影响血气分析检验结果可靠性的因素 [J]. 中华护理杂志，2001，36 (5)：159－160.

[4] 王惠文，吴载斌，孟　洁. 偏最小二乘回归的线性与非线性方法 [M]. (第 1 版). 北京：国防工业出版社，2006：267－275.

[5] 刘　明. 普通最小二乘法的几何分析 [J]. 统计与决策，2012 (4)：90－92.

[6] TRYGG J，WOLD S. Orthogonal projections to latent structures (O－PLS) [J]. Journal of Chemometrics，2010，16 (3) 119－128.

(下转第 265 页)