**文章编号:** 1000−8152(2010)03−0391−04

# 基于投影寻踪的非线性鲁棒偏最小二乘法及应用

贾润达[1], 毛志忠[1,2], 常玉清[1,2], 周俊武[3]

(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110004; 2. 东北大学 流程工业综合自动化教育部重点实验室, 辽宁 沈阳 110004;
3. 北京矿冶研究总院 信息技术与自动化研究设计所, 北京 100044)

**摘要**: 来自工业现场的数据往往具有非线性特性且包含离群点, 利用非线性偏最小二乘法(partial least squares, PLS)建模易受离群点的影响. 针对这一问题, 结合径向基函数(radial basis function, RBF)网络, 本文提出了一种基于投影寻踪的非线性鲁棒PLS方法. 该方法首先利用RBF变换将自变量与因变间的非线性关系转化为线性关系; 然后利用投影寻踪算法提取变换后自变量的鲁棒偏最小二乘法成分; 最后建立鲁棒PLS成分与因变量之间的鲁棒线性回归模型. 将该方法应用于湿法冶金萃余液pH值软测量建模问题, 结果验证了其有效性.

**关键词**: 径向基函数; 投影寻踪; 偏最小二乘法; 鲁棒性; 非线性

**中图分类号**: TP273　　**文献标识码**: A

# Nonlinear robust partial least squares based on projection pursuit and its application

JIA Run-da[1], MAO Zhi-zhong[1,2], CHANG Yu-qing[1,2], ZHOU Jun-wu[3]

(1. School of Information Science & Engineering, Northeastern University, Shenyang Liaoning 110004, China;
2. Key Laboratory of Integrated Automation of Process Industry, Northeastern University,
Ministry of Education, Shenyang Liaoning 110004, China;
3. Sub-Institute for IT & Automation, Beijing General Research Institute of Mining & Metallurgy, Beijing 100044, China)

**Abstract:** Data from industrial field usually possess nonlinear feature and contain outliers; modeling with nonlinear partial least squares(PLS) method may suffer from these outliers. For this case, combining with radial basis function(RBF) networks, we present a nonlinear robust PLS method based on the projection pursuit. First, the nonlinear relationship between independent and dependent variables is changed into a linear one by RBF transformation. Then, projection pursuit algorithm is employed to extract the robust PLS components of transformed independent variables. Finally a robust linear regression model is established between robust PLS components and the dependent variable. Applying the method to the soft-sensor modeling for pH value of raffinate solution in hydrometallurgy, we validate the effectiveness by the results.

**Key words:** radial basis function; projection pursuit; partial least squares; robustness; nonlinear

## 1 引言(Introduction)

PLS是一种多元线性回归方法, 用以解决自变量间的多重相关性问题, 辨识数据中的信息与噪声. 然而来自实际工业现场的数据, 往往具有较强的非线性特性, 而PLS本质上是一种线性方法, 难以精确描述过程的非线性特性, 因此各种非线性PLS方法相继出现[1~4]. 但仅仅引入非线性PLS并不足以解决实际过程数据的建模问题, 因为来自工业现场的数据往往包含离群点, 由于经典的PLS算法[5]易受离群点影响而出现过拟合现象, 因此上述基于PLS算法的非线性方法也会由于离群点的出现而失去其应有的泛化能力.

为此, 本文提出了一种基于投影寻踪的非线性鲁棒PLS算法. 该方法首先通过RBF变换将低维空间的非线性关系转变为高维空间的线性关系; 接下来利用投影寻踪算法提取鲁棒的PLS成分; 为了克服因变量中离群点的影响, 利用迭代再加权最小二乘法(IRLS)计算鲁棒PLS成分与因变量之间的回归系数. 本文还通过萃余液pH值软测量建模中与PLS[6], PRM[7]以及RBF-PLS[3]的比较, 进一步验证其优越性.

## 2 模型结构(Architecture of model)

RBF网络通过RBF变换, 将低维空间的非线性问题转化为高维空间的线性问题[8], 因此可以利

收稿日期: 2008−09−17; 收修改稿日期: 2009−04−11.
基金项目: 国家“863”高技术研究发展计划资助项目(2006AA060201).

用RBF网络来描述系统的非线性特性.

设有$n\times m$维输入数据矩阵$X$(已进行标准化处理), 其中$n$代表样本数据的个数, $m$代表样本数据的维数; $n\times 1$维输出数据向量$\boldsymbol{y}$. 若选用高斯径向基函数, 将输入数据矩阵$X$转化为激活矩阵$A$, 则$A$中的元素可以利用下式进行计算:

$$a_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{c}_j\|^2/\sigma_j^2),\ i,j=1,\cdots,n. \tag{1}$$

其中: $\boldsymbol{x}_i$是第$i$个数据样本的输入向量, $a_{ij}$是$A$第$i$行, 第$j$列的元素, $\boldsymbol{c}_j$和$\sigma_j$分别是高斯函数的中心和宽度参数. 可以选取每个输入数据样本作为RBF的中心, 即

$$\boldsymbol{c}_j=\boldsymbol{x}_j. \tag{2}$$

而宽度参数$\sigma_j$可通过计算每个输入数据样本距离其它输入数据样本欧氏距离的均值确定, 即

$$\sigma_j=\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|. \tag{3}$$

因此矩阵$A$是一个对角元为1的$n\times n$维方阵.

经过RBF变换后, 模型结构可以利用线性回归的形式进行表述:

$$\boldsymbol{y}=A\boldsymbol{b}+\boldsymbol{\varepsilon}. \tag{4}$$

其中: $\boldsymbol{b}$是回归系数向量, $\boldsymbol{\varepsilon}$是残差向量.

由于经典的PLS算法对于离群点缺乏鲁棒性, 因此直接利用PLS算法进行回归将使预测结果不够可靠. 为此, 本文提出了一种基于投影寻踪的鲁棒PLS算法用以求解式(4)中的回归系数.

## 3 非线性鲁棒PLS算法(Nonlinear robust PLS algorithm)

### 3.1 PLS算法原理(Theory of PLS algorithm)

PLS是一种多元线性回归方法. 为了计算式(4)中线性回归模型的回归系数$\boldsymbol{b}$, 可通过提取成分的方式代替直接求解, 则有

$$\boldsymbol{y}=T\boldsymbol{q}+\boldsymbol{\varepsilon}. \tag{5}$$

其中: $T$是由$h$个PLS成分组成的$n\times h$维得分矩阵; $\boldsymbol{q}$是相应的回归系数. 而得分矩阵$T$可通过下式进行计算:

$$T=A[W(P^{\mathrm{T}}W)^{-1}]. \tag{6}$$

其中$P=[\boldsymbol{p}_1,\cdots,\boldsymbol{p}_h]$代表$n\times h$维载荷矩阵, 而$W=[\boldsymbol{w}_1,\cdots,\boldsymbol{w}_h]$代表$n\times h$维权值矩阵, 权值向量$\boldsymbol{w}_k$可利用如下准则顺序计算:

$$\boldsymbol{w}_k=\arg\max_{\boldsymbol{w}\in W_{\mathrm{P}}}\mathrm{cov}(A\boldsymbol{w},\boldsymbol{y}), \tag{7}$$

且满足如下约束条件:

$$\|\boldsymbol{w}_k\|=1. \tag{8}$$

对于$1\leqslant l\leqslant k$有

$$\mathrm{cov}(A\boldsymbol{w}_k,A\boldsymbol{w}_l)=0. \tag{9}$$

其中$k=1,\cdots,h$, $W_{\mathrm{P}}$是权值向量的集合. 在求得得分矩阵$T$以后, $\boldsymbol{q}$可以利用下式计算

$$\boldsymbol{q}=(T^{\mathrm{T}}T)^{-1}T^{\mathrm{T}}\boldsymbol{y}. \tag{10}$$

将式(6)代入式(5), 即得最终的回归系数

$$\boldsymbol{b}=W(P^{\mathrm{T}}W)^{-1}\boldsymbol{q}. \tag{11}$$

### 3.2 基于投影寻踪的鲁棒PLS算法(Robust PLS algorithm based on projection pursuit)

在经典的PLS算法中, 准则(7)中的协方差并不具有鲁棒性, 而离群点的出现会改变所提取的PLS成分与因变量间的相关性, 因此可以利用$\alpha$截尾协方差对其进行估计, 忽略离群点对协方差的影响, 其中$0<\alpha<0.5$, 代表过程数据中被污染的百分比(本文均选为0.1). 设

$$\boldsymbol{z}_k=A\boldsymbol{w}_k, \tag{12}$$

则$\alpha$截尾协方差可根据下式进行计算:

$$\mathrm{cov}_\alpha(\boldsymbol{z}_k,\boldsymbol{y})=\frac{1}{n-2p}\sum_{i=p+1}^{n-p}\alpha_{(i)}, \tag{13}$$

且

$$\alpha_i=(z_i-\overline{z}_\alpha)(y_i-\overline{y}_\alpha). \tag{14}$$

其中: $p=[n\alpha]+1$, $[n\alpha]$代表小于$n\alpha$的整数中最大的一个; $z_i$和$y_i$分别代表$\boldsymbol{z}$和$\boldsymbol{y}$的第$i$个元素; $\overline{z}_\alpha$和$\overline{y}_\alpha$分别代表$\boldsymbol{z}$和$\boldsymbol{y}$的$\alpha$截尾均值[9]; 而$\alpha_{(1)}\leqslant\cdots\leqslant\alpha_{(n)}$代表将$\alpha_i$从小到大进行排列.

利用$\alpha$截尾协方差代替标准的协方差以后, 无法再利用经典的PLS算法提取相应的PLS成分, 因此选择投影寻踪算法来解决上述问题. 该算法不仅可以用于鲁棒PLS成分的提取, 而且可以在提取任意数量的成分后停止计算, 降低运算量. 设有单位鲁棒权值向量$\widetilde{\boldsymbol{w}}_k$表示的一个投影方向, 则变换后的输入数据在该方向上的投影可以用$A\widetilde{\boldsymbol{w}}_k$表示.

投影寻踪通过寻找所有可能的投影方向, 在满足约束条件式(8)(9)的同时, 使$\alpha$截尾协方差取得极大值[10]. 在实际应用过程中, 通常将投影方向限定在一定的范围之内. 由于取得极大值的投影方向几乎不可能出现在RBF变换后$n$维空间中没有数据点的区域, 因此我们将投影方向限定在$n$个变换后数据所确定的$n$个投影方向. 对于由$\alpha$截尾协方差所确定的寻踪目标, 确定的数据样本会确定唯一的寻踪目标; 因此对于有限个数据样本所确定的有限个投影方向, 算法在遍历所有的投影方向以后, 必然会稳定收敛于寻踪目标的极大值方向.

又由于因变量中的离群点会严重影响回归系数, 因此式(5)中的回归系数仍需利用鲁棒回归进行计

算. 这里选择IRLS[11]对其进行估计, 该方法通过反复迭代计算, 自适应地为样本分配0到1之间连续的权值, 从而克服因变量中离群点对回归模型的影响, 且具有较快的运算速度.

综上所述, 基于投影寻踪的非线性鲁棒PLS算法流程图如图1所示. 由于算术平均值不具有鲁棒性, 而离群点的出现可能歪曲变换后数据的中心, 因此流程图中选用$L1$中位值[12]对激活矩阵进行中心化处理, 图1中记$A$的鲁棒中心为$\mu_{L1}(A)$, $\mathbf{1}_n$是具有$n$个元素1的列向量; $\boldsymbol{a}_i^{\mathrm{C}}$代表$A^{\mathrm{C}}$的第$i$行数据.

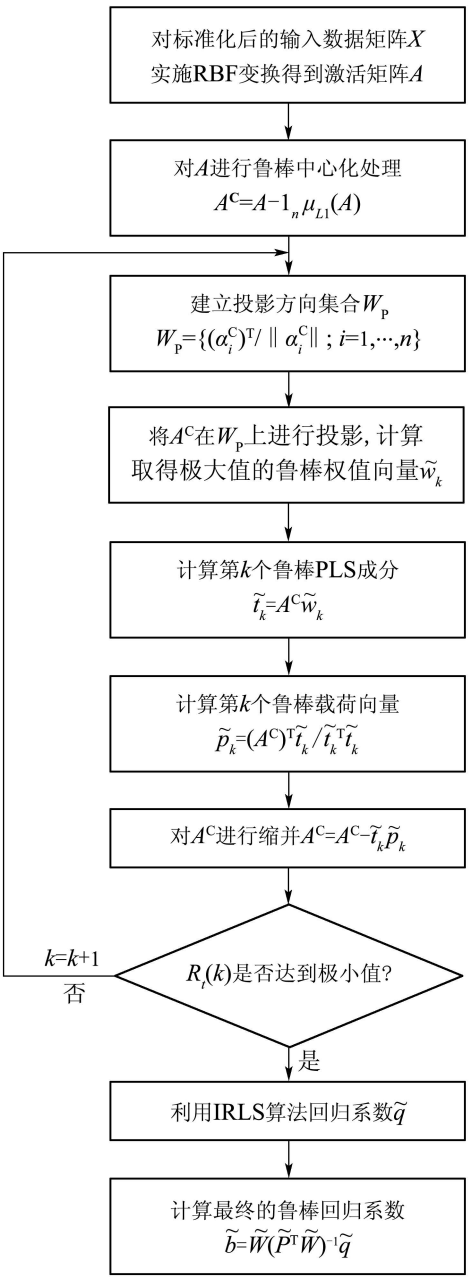



图 1　基于投影寻踪的非线性鲁棒PLS算法流程图

Fig. 1　The flowchart of nonlinear robust PLS algorithm based on projection pursuit

### 3.3　成分提取个数的确定(Confirming the number of components)

成分提取个数对于非线性鲁棒PLS模型的建立至关重要, 通常PLS采用留一交叉检验的方式确定成分的提取个数. 但当样本数据中存在离群点的时候, 鲁棒模型对于这些离群点的预测值会与其测量值之间存在较大偏差, 因此鲁棒交叉检验[13]是一种更好的成分个数确定方法. 该方法仅计算部分正常数据样本的交叉检验均方根误差, 而忽略离群点对交叉检验的影响, 则具有$k$个鲁棒PLS成分的截尾交叉检验均方根误差$R_t^{\mathrm{CV}}(k)$可通过下式进行计算:

$$R_t^{\mathrm{CV}}(k)=\sqrt{\frac{1}{n-p}\sum_{i=1}^{n-p}e_{(i)}^k},\tag{15}$$

且

$$e_i^k=(y_i-\hat{y}_{-i}^k)^2.\tag{16}$$

其中: $\hat{y}_{-i}^k$代表去掉第$i$个样本, 利用剩余的$n-1$个样本建模, 提取$k$个鲁棒PLS成分, 模型对第$i$个样本的预测值; 而$e_{(1)}^k\leqslant\cdots<e_{(n)}^k$则代表将$e_i^k$从小到大进行排列.

## 4　萃余液pH值软测量(Soft sensing for pH value of raffinate solution)

利用一个湿法冶金萃余液pH值软测量建模实例来比较本文所提出的非线性鲁棒PLS (NRPLS)与PLS, PRM以及RBF-PLS的预测效果, 萃取是湿法冶金重要的除杂手段, 通常通过控制萃余液的pH值来实现萃取剂对金属杂质的选择性萃取[14,15]. 然而料液会对pH计造成严重腐蚀, 要实现萃余液pH值的在线检测会大大提高生产成本, 因此可以通过软测量来解决这一难题. 通过分析, 影响萃取过程萃余液pH值的主要因素有: 有机相的流量, 料液的流量, 洗涤液的流量, 实测皂化率, 料液的pH值及温度. 将以上6个影响因素作为软测量模型的自变量, 将萃余液pH值作为因变量, 采集到132组稳态数据. 选择其中90组用于建模, 余下的42组数据用于验证.

实际过程数据不可避免的存在离群点, 离群点产生的原因多种多样, 就萃取过程来说, 可能是传感器故障, 随机扰动, 也可能是工作点离开正常工作范围, 操作工错误的记录等等. 由于校验模型的数据也可能存在离群点, 而鲁棒方法对于上述离群点的预测可能存在很大偏差, 因此仅仅利用预测均方根误差($R^{\mathrm{P}}$)来检验模型的预测效果显然有失公平, 这里引入另外一种评判准则, 截尾预测均方根误差($R_t^{\mathrm{P}}$), 其值可通过下式进行计算:

$$R_t^{\mathrm{P}} = \sqrt{\frac{1}{n_t - p_{\mathrm{t}}} \sum_{t=1}^{n_t - p_t} r_{(t)}}, \tag{17}$$

且

$$r_t = (y_t - \widehat{y}_t)^2. \tag{18}$$

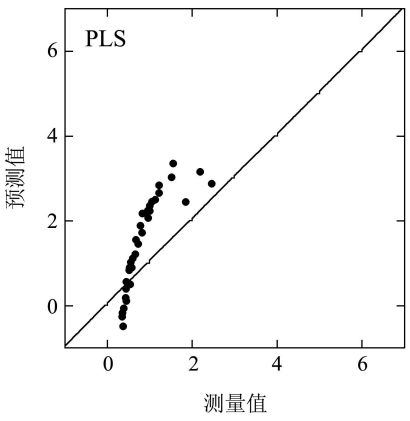
其中: $n_t$代表用于校验的样本数目, $p_{\mathrm{t}} = [n_t\alpha] + 1$, $\widehat{y}_t$代表模型对第$t$个校验样本的预测值; 而$r_{(1)\leqslant\cdots\leqslant r_{(n_t)}}$则代表将$r_t$从小到大进行排列; 且$t = 1, \cdots, n_t$.

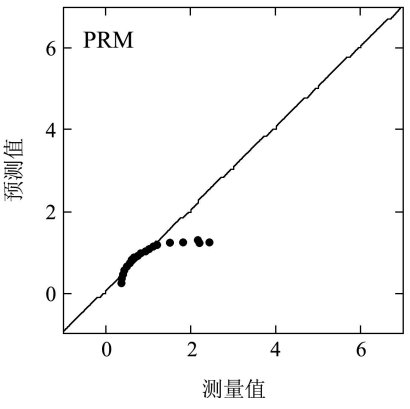
表 1 预测结果比较

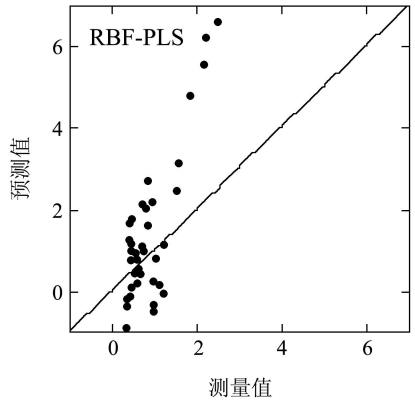
Table 1 Comparing prediction results

| | PLS[6] | PRM[7] | RBF-PLS[3] | NRPLS |
|---|---|---|---|---|
| $h$ | 3 | 4 | 19 | 15 |
| $R^{\mathrm{P}}$ | 0.8225 | 0.3311 | 1.3688 | 0.1629 |
| $R_t^{\mathrm{P}}$ | 0.7117 | 0.1197 | 0.8619 | 0.0444 |

表1中列出了利用PLS, PRM, RBF-PLS以及本文所提出的NRPLS的预测结果. RBF-PLS得到了最差的预测结果(无论是$R^{\mathrm{P}}$还是$R_t^{\mathrm{P}}$), 这说明建模数据中可能存在离群点, 使得该方法出现了严重的过拟合现象. 当采用PRM算法时, 预测精度明显有所提高, 但此时却忽视了过程的非线性特性. 利用本文提出的NRPLS方法进行建模, 获得了最高的预测精度.



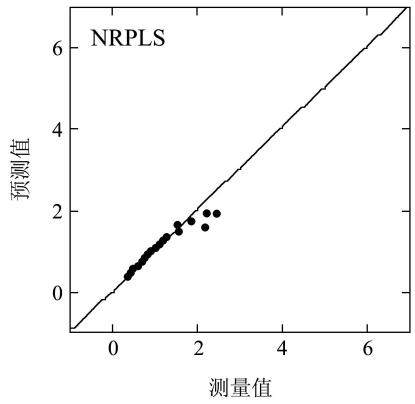




图 2 预测值与测量值的对比

Fig. 2 Prediction values versus measurement values

为了进一步比较4种算法的预测效果, 我们以测量值为横轴, 预测值为纵轴对42个校验样本作图, 结果如图2所示. 可以看出, 线性PLS模型难以进行精确的预测, 使数据分布几乎垂直于横轴. 对于RBF-PLS方法, 由于严重的过拟合现象, 使数据分布的杂乱无章. 而PRM方法, 由于其具有鲁棒性, 使得预测精度明显提高. 利用本文提出的NRPLS方法进行建模, 可以看到大部分数据很好的分布于对角线之上.

## 5 结论(Conclusion)

本文提出了一种基于投影寻踪的非线性鲁棒PLS算法用以解决过程数据建模问题. 该方法的优势在于, 可以有效解决具有非线性特性, 且包含离群点的过程数据建模问题. 而这两个特点, 在实际过程数据中是经常遇到的. 本文还通过萃余液pH值软测量建模问题验证了所提方法的有效性.

## 参考文献(References):

[1] QIN S J, MCAVOY T J. Nonlinear PLS modeling using neural networks[J]. *Computers & Chemical Engineering*, 1992, 16(4): 379 – 391.

[2] BAFFI G, MARTIN E B, MORRIS A J. Non-linear projection to latent structures revisited(the neural networks PLS algorithm[J]. *Computers & Chemical Engineering*, 1999, 23(9): 1293 – 1307.

(下转第399页)

change-point in a sequence of independent random variables[J]. *Annals of Statistics*, 1987, 3(6): 1321 – 1328.

[5] KOKOSZKA P, LEIPUS R. Change-point in the mean of dependent observations[J]. *Statistics and Probability Letters*, 1998, 40(4): 385 – 393.

[6] PITARAKIS J Y. Least squares estimation and tests of breaks in mean and variance under misspecification[J]. *Econometrics*, 2004, 7(1): 32 – 54.

[7] 刘元朋. 航空发动机管路测量数据分割方法[J]. 航空学报, 2008, 29(2): 285 – 290.
(LIU Yuanpeng. Segmentation method for point cloud of aeroengine pipelines[J]. *Acta Aeronautica et Astronautica Sinica*, 2008, 29(2): 285 – 290.)

[8] 张艳丽. 用SPC统计技术控制机车弯管质量[J]. 铁道技术监督, 2008, 36(1): 12 – 14.
(ZHANG Yanli. Quality control of locomotive air pipe in SPC statistics technology[J]. *Raiway Quality Control*, 2008, 36(1): 12 – 14.

[9] 方积乾, 陆盈. 现代医学统计学[M]. 北京: 人民卫生出版社, 2002.
(FANG Jiqian, LU Ying. *Modern Medical Statistics*[M]. Beijing: People's Medical Press, 2002.)

[10] 葛春蕾, 史晓平. 正态分布参数变点估计的强相合性[J]. 合肥工业大学学报, 2008, 31(2): 280 – 283.
(GE Chunlei, SHI Xiaoping. Consistency of the estimators for the change point of the parameters of a norm al distribution[J]. *Journal of Hefei University of Technology*, 2008, 31(2): 280 – 283.)

## 作者简介:

**袁　芳**　(1985—), 女, 硕士研究生, 主要研究方向为非线性时间序列分析与信息处理, E-mail: yuanfangyf2005@163.com;

**田　铮**　(1948—), 女, 教授, 博士生导师, 主要从事非线性时间序列分析、多尺度非线性随机模型、计算机视觉与图像处理等研究, E-mail: zhtian@nwpu.edu.cn;

**苏晓丽**　(1986—), 女, 硕士研究生, 主要研究方向为非线性时间序列分析与信息处理, E-mail: suxiaoli986@163.com;

**陈占寿**　(1982—), 男, 博士研究生, 主要研究方向为非线性时间序列分析与信息处理, E-mail: chenzhanshou@126.com.

(上接第394页)

[3] WALCZAK B, MASSART D L. The radial basis function-partial least squares approach as a flexible non-linear regression technique[J]. *Analytica Chimica Acta*, 1996, 331(3): 177 – 185.

[4] DURAND J F. Local polynomial additive regression through PLS and splines: PLSS[J]. *Chemometrics & Intelligent Laboratory Systems*, 2001, 58(2): 235 – 246.

[5] WOLD S, SJÖSTROM M, ERIKSSON L. PLS-regression: a basic tool of chemometrics[J]. *Chemometrics & Intelligent Laboratory Systems*, 2001, 58(2): 109 – 130.

[6] JONG S. SIMPLS: an alternative approach to partial least squares regression[J]. *Chemometrics & Intelligent Laboratory Systems*, 1993, 18(3): 251 – 263.

[7] SERNEELS S, CROUX C, FILZMOSER P, et al. Partial robust M-regression[J]. *Chemometrics & Intelligent Laboratory Systems*, 2005, 79(1/2): 55 – 64.

[8] WALCZAK B, MASSART D L. Local modelling with radial basis function networks[J]. *Chemometrics & Intelligent Laboratory Systems*, 2000, 50(2): 179 – 198.

[9] SERNEELS S, FILZMOSER P, CROUX C, et al. Robust continuum regression[J]. *Chemometrics & Intelligent Laboratory Systems*, 2005, 76(2): 197 – 204.

[10] CROUX C, FILZMOSER P, OLIVEIRA M R. Algorithms for projection-pursuit robust principal component analysis[J]. *Chemometrics & Intelligent Laboratory Systems*, 2007, 87(2): 218 – 225.

[11] KUTNER M H, NACHTSHEIM C J, NETER J. *Applied Linear Regression Models*[M]. Beijing: Higher Education Press, 2005: 439 – 441.

[12] DASZYKOWSKI M, KACZMAREK K, HEYDEN Y V, et al. Robust statistics in data analysis-a review: basic concepts[J]. *Chemometrics & Intelligent Laboratory Systems*, 2007, 85(2): 203 – 219.

[13] HUBERT M, BRANDEN K V. Robust methods for partial least squares regression[J]. *Journal of Chemometrics*, 2003, 17(10): 537 – 549.

[14] 王文忠, 郑平. Ni, Co萃取过程平衡pH值数学模型的建立[J]. 河北理工学院学报, 1999, 21(4): 19 – 22.
(WANG Wenzhong, ZHENG Ping. Establishing a mathematical model of equilibrium in the process of extraction Ni, Co[J]. *Journal of Hebei Institute of Technology*, 1999, 21(4): 19 – 22.)

[15] ANITHA M, SINGH H. Artificial neural network simulation of rare earths solvent extraction equilibrium data[J]. *Desalination*, 2008, 232(1/3): 59 – 70.

## 作者简介:

**贾润达**　(1981—), 男, 博士研究生, 主要研究方向为复杂工业系统建模与优化, E-mail: jiarunda@yahoo.com.cn;

**毛志忠**　(1961—), 男, 教授, 博士生导师, 主要研究方向为复杂工业系统建模、控制与优化, E-mail: maozhizhong@ise.neu.edu.cn;

**常玉清**　(1973—), 女, 副教授, 主要研究方向为软测量技术, E-mail: changyuqing@mail.neu.edu.cn;

**周俊武**　(1966—), 男, 博士, 研究员, 主要研究方向为复杂工业系统建模, E-mail: zhou_jw@bgrimm.com.