研究论文

双线性约束过程的鲁棒自适应数据校正方法

高 倩，阎威武，邵惠鹤
（上海交通大学电子信息及电器工程学院，上海 200240）

摘要：采用污染正态分布模型进行数据校正，相对于传统的最小二乘方法具有较好的鲁棒性，然而参数估计结果的精确度依赖于误差发生概率和方差比值两个先验模型参数的选取，这在实际生产中难以获得，采用固定的方差比也不符合实际，因而其应用受到了限制。本文针对污染正态分布模型的不足，提出了一种鲁棒自适应误差分布模型，该模型具有与标准正态分布模型相似的分布密度函数，不同之处在于采用鲁棒自适应可变权重因子调节误差方差，通过放大显著误差方差，减小其对参数估计的影响。将该模型用于双线性约束数据校正问题，并采用 Lagrange 乘子法得到鲁棒自适应最小二乘分析解，同时还对鲁棒自适应数据校正中的测量数据相关性问题进行了研究。仿真结果证实了该方法的有效性。

关键词：污染正态分布；鲁棒；自适应；显著误差；相关性；数据校正
中图分类号：TQ 015.9 **文献标识码**：A **文章编号**：0438－1157（2007）12－3108－09

Robust adaptive data rectification method for bilinear constraint process

GAO Qian, YAN Weiwu, SHAO Huihe
(*School of Electric Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*)

Abstract: A contaminated Gaussian distribution based method is robust for data rectification for its ability of taking probability distributions of random errors and gross errors into account simultaneously. But its application is limited because the precision of estimation depends on the selection of priori model parameters, which is difficult to obtain in practice. To avoid providing these parameters, a robust adaptive data rectification approach is proposed in this paper. First, a robust adaptive probability distribution model of errors is constructed. Adaptive factors, obtained from observations, are added to adjust the variances of the outlying observations. Then, Lagrange method is used to obtain the iterative algebraic solution. The correlation of measurements is also considered in this paper. Application to bilinear constraints process shows that the least square estimation based on the new approach can compensate the effect of gross errors effectively and give a robust estimation.

Key words: contaminated normal distribution; robust; adaptive; gross error; correlation; data rectification

2006－11－03 收到初稿，2007－03－07 收到修改稿。

联系人：邵惠鹤。**第一作者**：高倩（1972—），女，博士研究生。

基金项目：国家自然科学基金项目（60504033）；工业控制技术国家重点实验室开放课题基金项目（0708004）。

Received date: 2006－11－03.

Corresponding author: Prof. SHAO Huihe. **E－mail**: hhshao@sjtu.edu.cn

Foundation item: supported by the National Natural Science Foundation of China (60504033) and Open Fund of National Lab. of Industrial Control Technology of China (0708004).

引　言

数据校正的任务是根据过程内在物理化学规律，如物料及能量平衡方程，剔除数据中的随机误差和显著误差，使之更接近于过程的真实状态[1-3]。经典数据协调的一个基础假设是测量噪声服从正态分布，然而实际过程数据往往含有显著误差，不能满足这个假设，因而在进行数据协调之前应剔除显著误差[4]。有的学者偏向于保留原经典最小二乘估计法，采用工程近似的说法解释这个问题，Crowe[5]指出组分浓度和总流率的乘积作为一个变量参加协调时，虽然不严格服从正态假设，但当它们的标准差都不大时，可以认为其乘积近似服从正态分布。Bagajewicz[6]研究了开根号等运算对分布的影响，指出直接测量得到的温度、压力等量，可以认为是正态分布的，但间接测量得到的浓度等量的测量噪声多为非正态分布。事实上最小二乘估计的效率会因为变量偏离正态分布而变得非常低。为了克服这个缺点，许多学者提出了数据校正与显著误差侦破同步进行的方法，如 Tjoa 等[7]提出的污染正态分布模型校正法，Albuquerque 等[8-12]提出的鲁棒函数校正法。

污染正态分布模型校正法将误差污染分布模型作为极大似然目标函数，通过求解带约束的最优化问题进行参数估计[7]。Johnston 等[13]在此模型基础上引入先验历史数据，验证了鲁棒估计的可行性及其在含有显著误差数据校正中的良好性能。Ragot 等[14-15]采用此模型推导出线性及双线性数据校正的鲁棒最小二乘解。然而，所有基于污染正态分布模型的数据校正方法都需要事先给定显著误差的污染率和分布密度（即显著误差方差），这在实际应用中难以办到，在显著误差模型中采用固定的方差比也不符合实际生产过程。

本文针对污染正态分布模型的不足，提出一种鲁棒自适应分布模型，该模型具有与标准正态分布相似的密度函数形式，不同之处在于采用鲁棒自适应可变权重因子调节误差方差，通过放大显著误差方差，降低其对参数估计的影响。自适应调节因子通常采用鲁棒权函数，因而模型具有良好的鲁棒性。将该模型用于双线性约束数据校正问题，并通过 Lagrange 乘子法得到鲁棒自适应最小二乘分析解。

1　双线性约束问题的鲁棒最小二乘校正

双线性数据校正[14]的模型约束方程可以表述为

$$\boldsymbol{A}x = 0 \quad \boldsymbol{A} \in \mathbb{R}^{m\times n} \tag{1}$$

$$\boldsymbol{A}(x \otimes y) = 0 \tag{2}$$

式中　$\boldsymbol{A}$ 为模型系数矩阵，其元素依据物流流入、流出或与设备单元无关联分别取＋1、－1 或 0；x 为广度性质变量，如物料流量；y 为强度性质变量，如组成、温度等；⊗为 Kronecker 算子，定义如下：

令 $\boldsymbol{A}=[a_{ij}]$ 是 $n\times m$ 阶矩阵，$\boldsymbol{B}=[b_{ij}]$ 是 $k\times l$阶矩阵，则 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的 Kronecker 矩阵积定义为 $(n\cdot k)\times(m\cdot l)$ 阶矩阵，记作 $\boldsymbol{A}\otimes\boldsymbol{B}$

$$\boldsymbol{A}\otimes\boldsymbol{B}=[a_{ij}\cdot\boldsymbol{B}] \tag{3}$$

如果过程变量部分已测，那么已测变量可以表示为

$$x_{\mathrm{m}} = \boldsymbol{H}_x x + \varepsilon_x \quad \varepsilon_x \sim N(0,\boldsymbol{\Sigma}_x) \quad \boldsymbol{H}_x \in \mathbb{R}^{p\times n} \quad x_m \in \mathbb{R}^{p\times 1} \tag{4}$$

$$y_{\mathrm{m}} = \boldsymbol{H}_y y + \varepsilon_y \quad \varepsilon_y \sim N(0,\boldsymbol{\Sigma}_y) \quad \boldsymbol{H}_y \in \mathbb{R}^{q\times n} \quad y_m \in \mathbb{R}^{q\times 1} \tag{5}$$

其中，$\boldsymbol{H}_x$和$\boldsymbol{H}_y$定义为变量 x 和 y 的选择矩阵，其元素由 0 和 1 组成。$\boldsymbol{\Sigma}_x$ 和 $\boldsymbol{\Sigma}_y$ 为测量误差协方差矩阵。

由于测量数据不可避免地含有显著误差，采用正态分布密度进行描述存在较大偏差，Tjoa 等[7]提出的污染正态分布数据校正模型同时考虑了随机误差和显著误差的分布情况，更符合实际，其测量误差分布模型为

$$P_x = \gamma p_{x,1} + (1-\gamma)p_{x,2} \tag{6}$$

$$P_y = \gamma p_{y,1} + (1-\gamma)p_{y,2} \tag{7}$$

$$p_{x,j} = \frac{1}{\sqrt{2\pi\det(\boldsymbol{\Sigma}_{x,j})}}\exp\left[-\frac{1}{2}(\boldsymbol{H}_x x - x_{\mathrm{m}})^{\mathrm{T}}\boldsymbol{\Sigma}_{x,j}^{-1}(\boldsymbol{H}_x x - x_{\mathrm{m}})\right], \quad j=1,2 \tag{8}$$

$$p_{y,j} = \frac{1}{\sqrt{2\pi\det(\boldsymbol{\Sigma}_{y,j})}}\exp\left[-\frac{1}{2}(\boldsymbol{H}_y y - y_{\mathrm{m}})^{\mathrm{T}}\boldsymbol{\Sigma}_{y,j}^{-1}(\boldsymbol{H}_y y - y_{\mathrm{m}})\right], \quad j=1,2 \tag{9}$$

式中　γ 为显著误差污染率，$p_{x(y),j}$（$j=1$，2）表示随机误差和显著误差的分布密度模型，$\boldsymbol{\Sigma}_{x(y),j}$（$j=1$，2）为随机误差和显著误差的协方差矩阵，当测量误差不相关时

$$\boldsymbol{\Sigma}_{x,1} = \begin{bmatrix}\sigma_{x,1,1}^2 & & \\ & \ddots & \\ & & \sigma_{x,1,n}^2\end{bmatrix} \tag{10}$$

$$\boldsymbol{\Sigma}_{x,2}=b_x\boldsymbol{\Sigma}_{x,1}=b_x\begin{bmatrix}\sigma^2_{x,1,1}&&\\&\ddots&\\&&\sigma^2_{x,1,n}\end{bmatrix} \tag{11}$$

$$\boldsymbol{\Sigma}_{y,1}=\begin{bmatrix}\sigma^2_{y,1,1}&&\\&\ddots&\\&&\sigma^2_{y,1,n}\end{bmatrix} \tag{12}$$

$$\boldsymbol{\Sigma}_{y,2}=b_y\boldsymbol{\Sigma}_{y,1}=b_y\begin{bmatrix}\sigma^2_{y,1,1}&&\\&\ddots&\\&&\sigma^2_{y,1,n}\end{bmatrix} \tag{13}$$

其中，b_x 和 b_y 为显著误差和随机误差的方差比值。如果 x 和 y 相互独立，数据校正表述为如下约束优化问题

$$\min\boldsymbol{\Phi}=-\ln[\gamma p_{x,1}+(1-\gamma)p_{x,2}][\gamma p_{y,1}+(1-\gamma)p_{y,2}]$$
$$\text{s.t.}\quad \boldsymbol{A}x=0$$
$$\boldsymbol{A}(x\otimes y)=0 \tag{14}$$

其 Lagrange 函数为

$$\boldsymbol{\Gamma}=\boldsymbol{\Phi}+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}x+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}(x\otimes y) \tag{15}$$

式中 $\boldsymbol{\lambda}$ 和 $\boldsymbol{\mu}$ 为 Lagrange 乘子。将 $\boldsymbol{\Gamma}$ 分别对 x、y、λ 和 μ 求导，并令导数为零，得到

$$\boldsymbol{H}_x^{\mathrm{T}}\boldsymbol{W}_x^{-1}(H_xx-x_{\mathrm{m}})+\boldsymbol{A}^{\mathrm{T}}\lambda+(\boldsymbol{A}\otimes y)^{\mathrm{T}}\mu=0 \tag{16}$$

$$\boldsymbol{H}_y^{\mathrm{T}}\boldsymbol{W}_y^{-1}(\boldsymbol{H}_yy-y_{\mathrm{m}})+(\boldsymbol{A}\otimes x)^{\mathrm{T}}\mu=0 \tag{17}$$

$$\boldsymbol{A}x=0 \tag{18}$$

$$\boldsymbol{A}(x\otimes y)=0 \tag{19}$$

其中

$$\boldsymbol{W}_x^{-1}=\frac{\gamma\boldsymbol{P}_{x,1}\boldsymbol{\Sigma}_{x,1}^{-1}+(1-\gamma)\boldsymbol{P}_{x,2}\boldsymbol{\Sigma}_{x,2}^{-1}}{\gamma\boldsymbol{P}_{x,1}+(1-\gamma)\boldsymbol{P}_{x,2}} \tag{20}$$

$$\boldsymbol{W}_y^{-1}=\frac{\gamma\boldsymbol{P}_{y,1}\boldsymbol{\Sigma}_{y,1}^{-1}+(1-\gamma)\boldsymbol{P}_{y,2}\boldsymbol{\Sigma}_{y,2}^{-1}}{\gamma\boldsymbol{P}_{y,1}+(1-\gamma)\boldsymbol{P}_{y,2}} \tag{21}$$

对式（16）～式（19）进行求解，得到

$$x=[\boldsymbol{G}_x-\boldsymbol{G}_x\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{G}_x\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\boldsymbol{G}_x]\times$$
$$[\boldsymbol{H}_x^{\mathrm{T}}\boldsymbol{W}_x^{-1}x_{\mathrm{m}}-\boldsymbol{A}_y^{\mathrm{T}}(\boldsymbol{A}_x\boldsymbol{G}_y\boldsymbol{A}_x^{\mathrm{T}})^{-1}\boldsymbol{A}_x\boldsymbol{G}_y\boldsymbol{H}_y^{\mathrm{T}}\boldsymbol{W}_y^{-1}y_{\mathrm{m}}] \tag{22}$$

$$y=[\boldsymbol{G}_y-\boldsymbol{G}_y\boldsymbol{A}_x^{\mathrm{T}}(\boldsymbol{A}_x\boldsymbol{G}_y\boldsymbol{A}_x^{\mathrm{T}})^{-1}\boldsymbol{A}_x\boldsymbol{G}_y]\boldsymbol{H}_y^{\mathrm{T}}\boldsymbol{W}_y^{-1}y_{\mathrm{m}} \tag{23}$$

$$\boldsymbol{G}_x=(\boldsymbol{H}_x^{\mathrm{T}}\boldsymbol{W}_x^{-1}\boldsymbol{H}_x+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1} \tag{24}$$

$$\boldsymbol{G}_y=(\boldsymbol{H}_y^{\mathrm{T}}\boldsymbol{W}_y^{-1}\boldsymbol{H}_y+\boldsymbol{A}_x^{\mathrm{T}}\boldsymbol{A}_x)^{-1} \tag{25}$$

$$\boldsymbol{A}_x=\boldsymbol{A}\cdot\mathrm{diag}(x) \tag{26}$$

$$\boldsymbol{A}_y=\boldsymbol{A}\cdot\mathrm{diag}(y) \tag{27}$$

如果考虑变量的上下限约束

$$x_{\mathrm{lb}}\leqslant x\leqslant x_{\mathrm{ub}} \tag{28}$$

可以通过 Bayesian 基本估计式加入到数据校正问题中。Bayesian 估计为

$$p(x\mid x_{\mathrm{m}})=\frac{p(x_{\mathrm{m}}\mid x)p(x)}{p(x_{\mathrm{m}})} \tag{29}$$

式中 $p(x\mid x_{\mathrm{m}})$ 为 x 的验后概率密度，$p(x_{\mathrm{m}}\mid x)$为 x 的验前概率密度，$p(x)$ 为状态变量的验前概率密度，$p(x_{\mathrm{m}})$ 为测量变量的概率密度，$p(x_{\mathrm{m}})$ 通常为一常数。式（28）可以表述为

$$p(x)=\frac{1}{2}\left[\tanh\left(\frac{x-x_{\mathrm{lb}}}{r}\right)-\tanh\left(\frac{x-x_{\mathrm{ub}}}{r}\right)\right] \tag{30}$$

式中 r 为可调参数，其值越小，逼近效果越好；x 的误差分布函数为

$$P(x\mid x_{\mathrm{m}})=[\gamma p_{x,1}+(1-\gamma)p_{x,2}]p(x) \tag{31}$$

$$p_{x,j}=\frac{1}{\sqrt{2\pi\det(\boldsymbol{\Sigma}_{x,j})}}\exp\left[-\frac{1}{2}(\boldsymbol{H}_xx-x_{\mathrm{m}})^{\mathrm{T}}\boldsymbol{\Sigma}_{x,j}^{-1}(\boldsymbol{H}_xx-x_{\mathrm{m}})\right]$$
$$j=1,2 \tag{32}$$

数据校正的 Lagrange 表达式为

$$\boldsymbol{\Gamma}=-\ln p(x_{\mathrm{m}}\mid x)+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}x \tag{33}$$

对 x 求导得到

$$\frac{\partial p(x_{\mathrm{m}}\mid x)}{\partial x}=\boldsymbol{W}_x^{-1}[x-(x_{\mathrm{m}}+\boldsymbol{W}_xh_x)]+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} \tag{34}$$

其中

$$\boldsymbol{W}_x^{-1}=\frac{\gamma p_{x,1}\boldsymbol{\Sigma}_{x,1}^{-1}+(1-\gamma)p_{x,2}\boldsymbol{\Sigma}_{x,2}^{-1}}{\gamma p_{x,1}+(1-\gamma)p_{x,2}} \tag{35}$$

$$h_x=\frac{1}{r}\left[\tanh\left(\frac{x-x_{\mathrm{lb}}}{r}\right)-\tanh\left(\frac{x-x_{\mathrm{ub}}}{r}\right)\right] \tag{36}$$

这样，将 $x_{\mathrm{m}}+W_xh_x$ 代替 x_{m}，式（22）和式（23）的结果仍然适用。如果考虑变量 y 的上下限约束，推导类似。

2 鲁棒自适应数据校正方法

2.1 误差分布模型

基于污染正态分布模型的数据校正方法需要知道显著误差的先验概率 γ 和显著误差先验方差 $\sigma_{x,2}^2$，这在实际应用中很难获得，而且采用固定的显著误差方差 $\sigma_{x,2}^2$也不符合实际，因而具有一定的局限性。本文针对污染正态分布模型的不足，提出鲁棒自适应误差分布模型

$$P_i=\frac{1}{\sqrt{2\pi}k_i\sigma_i}\exp\left(-\frac{(x_i-x_{\mathrm{m},i})^2}{2k_i^2\sigma_i^2}\right) \tag{37}$$

式中 k_i为误差方差的鲁棒自适应调节因子，当某变量标定为含有显著误差时，其方差按误差幅度进行放大，以减小其对参数估计的影响。鲁棒自适应放大因子可以根据鲁棒函数进行计算。自适应误差分布模型和正态分布模型的概率密度如图 1 所示。

从图 1 可以看到，随着 k 值增大，分布曲线的尾部越长，可以用以描述含有显著误差的分布模式，当 $k=1$ 时，自适应误差分布退化为正态分布。如果各测量变量相互独立，数据校正问题可以表述为

$$\max\sum_i\ln P_i=\sum_i\ln\left\{\frac{1}{\sqrt{2\pi}k_i\sigma_i}\exp\left[-\frac{(x_i-x_{\mathrm{m},i})^2}{2k_i^2\sigma_i^2}\right]\right\}$$
$$\text{s.t.}\quad f(x)=0 \tag{38}$$

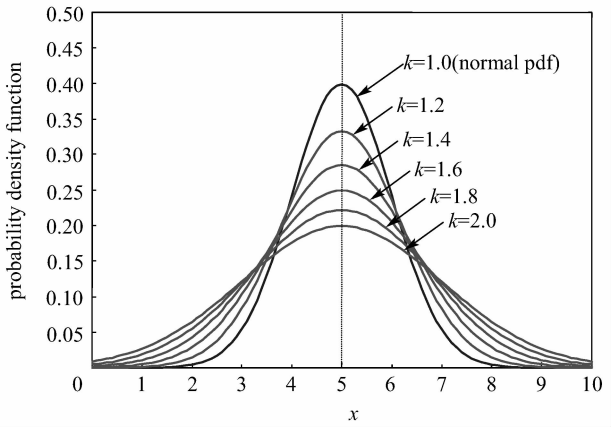

图 1　自适应误差分布和正态分布模型的概率密度示意图

Fig. 1　Probability density function of adaptive error distribution and normal distribution

鲁棒自适应因子可以表示为鲁棒估计权函数的倒数，$k_i=1/w_i$，w_i 为 Huber 估计、Hampel 截尾估计以及 Tukey 估计等鲁棒估计函数的权函数。本文采用 Huber 估计权函数

$$w_i(\varepsilon_i)=\begin{cases}1 & |\varepsilon_i|\leqslant c\\ \dfrac{c\,\mathrm{sgn}(\varepsilon_i)}{\varepsilon_i} & |\varepsilon_i|>c\end{cases} \tag{39}$$

式中　$\varepsilon_i=(x_i-x_{\mathrm{m},i})/\sigma$ 为测量标准残差，常数 c 为 Huber 模型可调参数，根据实际数据中的显著误差污染率 γ 来确定

$$\frac{1}{1-\gamma}=2\boldsymbol{\Phi}(c)-1+2\,\frac{\phi(c)}{c} \tag{40}$$

式中　$\boldsymbol{\Phi}(c)$ 和 $\phi(c)$ 分别表示标准正态分布在 c 点处的分布函数和概率密度值。实际计算中污染率 γ 的准确值事先并不知道，常遇到的显著误差所占比例大约在 1%～10%之间，因此 c 值一般在 1～2 之间，最常用的 c 值在 1.5 左右。Huber 估计是一种具有凸目标函数的 M 估计，其计算精度受显著误差的影响不大，也不受初值好坏的影响，而且具有一致收敛的优点。

2.2　双线性数据校正

对于如式（4）和式（5）所示的双线性过程，测量变量的鲁棒自适应误差分布模型表示为

$$p_x=\frac{1}{\sqrt{2\pi\det(\tilde{\boldsymbol{\Sigma}}_x)}}\exp\left[-\frac{1}{2}(\boldsymbol{H}_x x-x_{\mathrm{m}})^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_x^{-1}(\boldsymbol{H}_x x-x_{\mathrm{m}})\right] \tag{41}$$

$$p_y=\frac{1}{\sqrt{2\pi\det(\tilde{\boldsymbol{\Sigma}}_y)}}\exp[-\frac{1}{2}(\boldsymbol{H}_y y-y_{\mathrm{m}})^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\boldsymbol{H}_y y-y_{\mathrm{m}})] \tag{42}$$

其中

$$\tilde{\boldsymbol{\Sigma}}_x=\mathrm{diag}(k_{x,i}^2\sigma_{x,i}^2)=\mathrm{diag}\left(\frac{\sigma_{x,i}^2}{w_{x,i}^2}\right) \tag{43}$$

$$\tilde{\boldsymbol{\Sigma}}_y=\mathrm{diag}(k_{y,i}^2\sigma_{y,i}^2)=\mathrm{diag}\left(\frac{\sigma_{y,i}^2}{w_{y,i}^2}\right) \tag{44}$$

数据校正问题描述为

$$\begin{aligned}\min -\ln P_x P_y=&\frac{1}{2}[(\boldsymbol{H}_x x-x_{\mathrm{m}})^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_x^{-1}(\boldsymbol{H}_x x-x_{\mathrm{m}})+\\&(\boldsymbol{H}_y y-y_{\mathrm{m}})^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_y^{-1}(\boldsymbol{H}_y y-y_{\mathrm{m}})]-\ln\left(2\pi\sqrt{\det(\tilde{\boldsymbol{\Sigma}}_x)\det(\tilde{\boldsymbol{\Sigma}}_y)}\right)\\ \text{s.t.}\quad&\boldsymbol{A}x=0\\&\boldsymbol{A}(x\otimes y)=0\end{aligned} \tag{45}$$

采用 Lagrange 乘子法求得的优化解与式（22）、式（23）非常相似，不同之处在于权重矩阵 $\boldsymbol{W}_x$ 和 $\boldsymbol{W}_y$ 的计算及含义不同

$$\boldsymbol{W}_x^{-1}=\tilde{\boldsymbol{\Sigma}}_x^{-1}=\mathrm{diag}\left[\frac{w_{x,i}^2}{\sigma_{x,i}^2}\right] \tag{46}$$

$$\boldsymbol{W}_y^{-1}=\tilde{\boldsymbol{\Sigma}}_y^{-1}=\mathrm{diag}\left[\frac{w_{y,i}^2}{\sigma_{y,i}^2}\right] \tag{47}$$

可以看到，基于鲁棒自适应最小二乘的数据校正方法采用可变方差描述含有显著误差的数据采样分布，物理意义明确，计算更加合理。由于 x 和 y 的计算与权重因子 $w_{x,i}$ 和 $w_{y,i}$ 的计算相互关联，因而需要迭代求解，计算步骤如下。

（1）初始化。令迭代次数 $k=0$，$x^{(k)}=x_{\mathrm{m}}$，$y^{(k)}=y_{\mathrm{m}}$，根据显著误差的先验发生概率估算 Huber模型参数 c。

（2）迭代估计。由式（39）计算 $w_x^{(k)}$ 和 $w_y^{(k)}$，式（46）和式（47）计算 $W_x^{(k)-1}$ 和 $W_y^{(k)-1}$，式（26）和式（27）计算 A_x^k 和 A_y^k，式（24）和式（25）计算 $\boldsymbol{G}_x$ 和 $\boldsymbol{G}_y$，于是有

$$\begin{aligned}x^{(k+1)}=&[\boldsymbol{G}_x^{(k)}-\boldsymbol{G}_x^{(k)}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{G}_x^{(k)}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\boldsymbol{G}_x^{(k)}][\boldsymbol{H}_x^{\mathrm{T}}\boldsymbol{W}_x^{(k)-1}x_{\mathrm{m}}-\\&\boldsymbol{A}_y^{\mathrm{T}}(\boldsymbol{A}_x\boldsymbol{G}_y^{(k)}\boldsymbol{A}_x^{\mathrm{T}})^{-1}\boldsymbol{A}_x\boldsymbol{G}_y^{(k)}\boldsymbol{H}_y^{\mathrm{T}}\boldsymbol{W}_y^{(k)-1}y_{\mathrm{m}}]\end{aligned} \tag{48}$$

$$y^{(k+1)}=[\boldsymbol{G}_y^{(k)}-\boldsymbol{G}_y^{(k)}\boldsymbol{A}_x^{\mathrm{T}}(\boldsymbol{A}_x\boldsymbol{G}_y^{(k)}\boldsymbol{A}_x^{\mathrm{T}})^{-1}\boldsymbol{A}_x\boldsymbol{G}_y^{(k)}]\boldsymbol{H}_y^{\mathrm{T}}\boldsymbol{W}_y^{(k)-1}y_{\mathrm{m}} \tag{49}$$

（3）收敛。令 $\eta_x^{(k+1)}=\|\hat{x}^{(k+1)}-\hat{x}^{(k)}\|$，$\eta_y^{(k+1)}=\|\hat{y}^{(k+1)}-\hat{y}^{(k)}\|$，当 $|\eta^{(k+1)}-\eta^{(k)}|\leqslant e$ 时，收敛，e 为误差限，否则 $k=k+1$，继续循环，直至收敛。

2.3　相关性

对于多维分布，变量 x 的鲁棒自适应分布模型为

$$p\propto\exp\left[-\frac{1}{2}(\boldsymbol{H}_x x-x_{\mathrm{m}})^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_x^{-1}(\boldsymbol{H}_x x-x_{\mathrm{m}})\right] \tag{50}$$

$$\tilde{\boldsymbol{\Sigma}}_x=\boldsymbol{B}_x\boldsymbol{\Sigma}_x \tag{51}$$

$$\boldsymbol{B}_x = \mathrm{diag}(1/w_{x,i}^2) \tag{52}$$

若变量相关，测量误差协方差矩阵为 $\boldsymbol{\Sigma}_x$，一般情况下，$\boldsymbol{\Sigma}_x$ 为正定矩阵，将其 Cauchy 分解

$$\boldsymbol{\Sigma}_x = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} \tag{53}$$

将式（53）代入式（50）得

$$P \propto \exp\left[\frac{1}{2}(\boldsymbol{H}_x x - x_{\mathrm{m}})^{\mathrm{T}}(\boldsymbol{C}^{-1})^{\mathrm{T}}\boldsymbol{B}_x^{-1}\boldsymbol{C}^{-1}(\boldsymbol{H}_x x - x_{\mathrm{m}})\right] \tag{54}$$

令

$$z - z_{\mathrm{m}} = \boldsymbol{C}^{-1}(\boldsymbol{H}_x x - x_{\mathrm{m}}) \text{ 且 } \boldsymbol{\Sigma}_z = \boldsymbol{I} \tag{55}$$

于是

$$P \propto \exp\left[\frac{1}{2}(z - z_{\mathrm{m}})^{\mathrm{T}}\boldsymbol{B}_x^{-1}(z - z_{\mathrm{m}})\right] \tag{56}$$

可以看到，相关测量协方差矩阵经 Cauchy 分解后，转化为不相关问题，前述方法仍然适用。

3 计算实例

本文采用文献［14］的例子，该过程包括 9 个模块 16 个流股，每个流股含有两种组分，流程如图 2 所示，x 代表流率，y 代表关键组分组成，测量数据由变量真值加上随机误差和显著误差产生。流率和组成的随机误差方差为 $\sigma_x = 0.1x$，$\sigma_y = 0.1y$，变量 x_3、x_7、x_{16} 和 y_1、y_9、y_{12} 含有显著误差，其幅度分别为 10、8、5、3、3、3。变量真值及测量数据如表 1 所示。

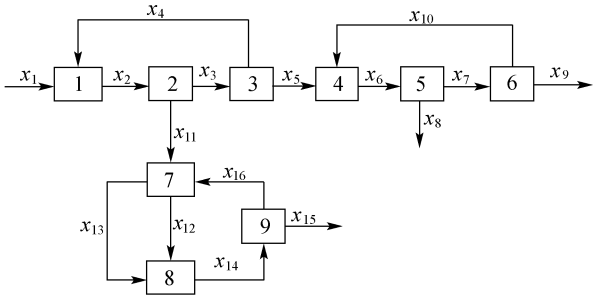

图 2 过程流程图

Fig. 2 Process flowsheet

为了比较鲁棒最小二乘与鲁棒自适应最小二乘的计算性能，首先对污染正态分布模型和 Huber 模型的参数进行整定，使其具有相近的估计效率（95.5%），估计效率通常定义为基于理想误差分布估计方差的倒数。本文通过仿真产生 2000 组数据，对模型参数进行整定，获得 Huber 模型参数 $c = 1.4$，污染正态分布模型参数 $\gamma = 0.235$，$b = 10$。本文采用的性能评价指标为偏差平方和（SSE），计算结果如表 1 所示。SSE 定义如下

$$\mathrm{SSE} = \sum(x - x_{\mathrm{true}})^2 + \sum(u - u_{\mathrm{true}})^2 \tag{57}$$

表 1 过程测量值与估计值

Table 1 Measurements and estimations

Stream No.	True value x	Meas. x_{m}	RALS① $\hat{x}_1$	RLS② $\hat{x}_2$	LS③ $\hat{x}_3$	True value y	Meas. y_m	RALS $\hat{y}_1$	RLS $\hat{y}_2$	LS $\hat{y}_3$
1	57.72	57.74	58.322	58.077	56.688	6.23	**8.96**	**6.822**	**6.9868**	**7.0926**
2	65.71	67.05	66.358	71.458	69.479	6.89	—	7.5111	7.5858	7.9431
3	52.98	**63.91**	**53.376**	**58.558**	**60.758**	6.37	6.60	7.1941	7.5072	7.6903
4	7.99	7.94	8.0364	13.381	12.791	11.65	11.84	11.665	11.773	11.712
5	44.98	—	45.339	45.177	47.967	5.43	5.45	6.3131	6.8607	6.6179
6	55.71	55.89	55.985	55.893	58.248	6.40	6.26	6.9125	7.1928	7.2055
7	32.13	**39.49**	**32.059**	**32.209**	**36.197**	7.24	7.10	8.0594	8.2529	8.436
8	23.58	23.44	23.926	23.684	22.051	5.26	5.42	5.1162	5.3162	5.1857
9	21.40	21.35	21.414	21.493	25.916	5.62	**8.78**	**7.387**	**7.8985**	**7.8365**
10	10.73	10.45	10.645	10.716	10.281	10.47	10.31	9.7192	10.125	9.9473
11	12.73	13.03	12.982	12.9	8.7209	9.07	9.25	9.5332	9.1128	9.7041
12	17.05	16.56	16.698	16.556	16.62	8.80	**11.85**	**9.8378**	**11.963**	**9.7369**
13	2.42	2.38	2.5205	2.3797	2.7997	22.02	22.81	22.254	22.806	22.625
14	19.47	18.90	19.219	18.936	19.42	10.45	10.23	11.396	11.186	11.595
15	12.73	13.07	12.982	12.9	8.7209	9.07	9.13	9.5332	9.1128	9.7041
16	6.74	**11.63**	**6.2366**	**6.0358**	**10.699**	13.05	12.46	12.997	11.768	13.136
SSE	—	—	1.854	94.52	201.82	—	—	9.295×10^{-4}	2.42×10^{-3}	1.56×10^{-3}

① RALS：Robust adaptive least square.
② RLS：Robust least square.
③ LS：least square.
Note：The values of y variable are multiplied by 100.

从表中结果可以看到，鲁棒自适应最小二乘与鲁棒最小二乘数据校正算法都可以有效减小显著误差对校正结果的影响，得到较可靠的数据校正结果，具有良好的鲁棒性，而传统最小二乘算法将误差扩散到其他变量上，鲁棒性差。另外，表中 SSE 计算结果还表明，当污染正态模型和 Huber 模型具有相近的估计效率时，Huber 估计具有更好的鲁棒性，其 SSE 更小，估计值更接近过程真实值。实际计算中，由于数据的污染率和显著误差方差难以准确获知，基于污染正态分布的鲁棒最小二乘算法采用固定的显著误差方差显然不符合实际情况，基于 Huber 模型的鲁棒自适应最小二乘采用可变方差描述误差的分布特征，更符合实际，模型参数 k 只受数据污染率的影响，便于调节。

为了更好地说明文章所提出方法的有效性，采用 Monte Carlo 仿真实验产生 1000 组模拟数据，变量 x_3、x_7、x_{16}和 y_1、y_9、y_{12}含有显著误差，其幅度分别为 20、18、8、0.05、0.08、0.05，根据测量值与估计值的差值进行误差探测，取 4 和 0.02 为变量 x 和 y 误差侦破限，得到每个变量含有显著误差的概率，并与传统的最小二乘方法进行比较，其结果如图 3 所示。从图中清楚地看到，采用鲁棒自适应最小二乘数据校正方法探测到 x_3、x_7、x_{16}和 y_1、y_9、y_{12}含有显著误差的几率较大，而最小二乘算法无法准确探测到显著误差的具体位置。

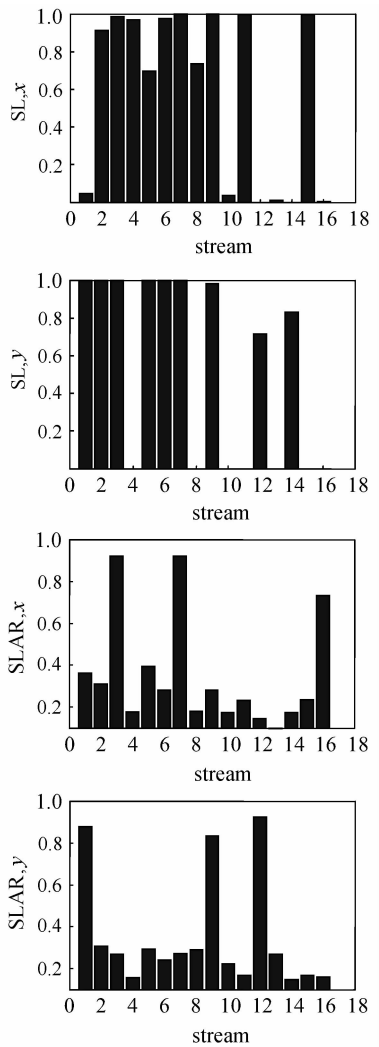

图 3　RALS 和 LS 方法显著误差探测性能比较

Fig. 3　Percentage of gross error detection for RALS and LS

3　工业应用

某焦化分厂的过程物流示意图如图 4 所示，该过程包括 15 个节点，30 个流股变量，每个流股含有 5 个组分：CO、H_2、CO_2、N_2、CH_4，未测流量的流股 13 个，未测组成的流股 9 个，如表 2 所示，流股 25 和 26 为人为加入的虚拟流股，表示反应增加的物质流量。测量值误差的方差-协方差矩阵根据实时测量数据进行计算，Huber 模型参数 $c_x=c_y=1.5$。各流股流量和组成的测量值和估计值如表 3 所示，流量测量的显著误差根据测量残差进行确定，如图 5 所示，从图中可以看到流股 3 的流量测量值具有最大的测量残差，为可疑显著误差数据，经确认该流股流量计量仪表的确存在测量问题。

值得注意的是，以上算法在处理多组分化工数据校正问题时，由于没有考虑到模块的具体功能，

表 2　焦化碳一过程中的未测变量

Table 2　Unmeasured variables

Unmeasured variables	Streams
flow rate	1, 2, 17, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
composition	17, 19, 23, 25, 26, 27, 28, 29, 30

校正结果存在一定的误差，如节点 1、2、13 和 15 为分流器节点，出料流股与进料流股的组成应该相等，满足强度约束条件，从表 3 结果可以看到连接节点 13 的 14、21 和 24 流股的 CO 组成分别为 13.17，12.67 和 13.02，不满足组成相等的强度约束条件。另外，由于校正中没有考虑到化工多组分系统每一流股所含组分应满足组成归一条件，这也影响了组成数据的准确估计。但是由于该算法算式

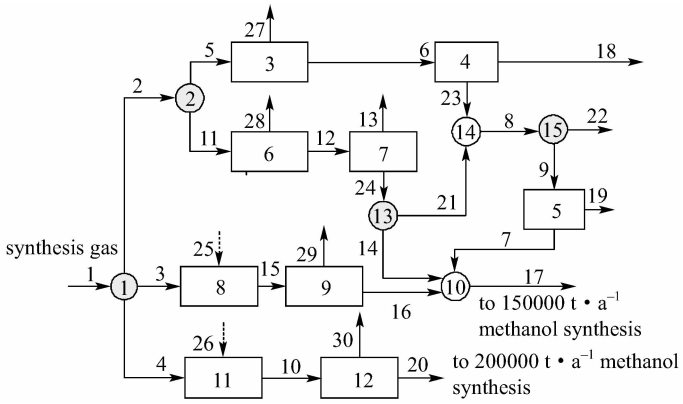

图 4 焦化数据校正简化节点拓扑图

Fig. 4 Process simplified flowsheet for data rectification

表 3 碳一流程某时刻的数据校正结果

Table 3 Measurements and estimates

Stream No.	Flow rate/m³ · h⁻¹		Composition/%(vol.)									
			CO		H_2		CO_2		N_2		CH_4	
	Meas.	Est.	Meas.	Est.	Meas.	Est.	Meas.	Est.	Meas.	Est.	Meas.	Est.
1	unm.	151222.1	46.84	46.84	35.39	35.48	17.33	17.33	0.31	0.31	0.13	0.13
2	unm.	34120.3	46.84	46.84	35.39	35.48	17.33	17.33	0.31	0.31	0.13	0.13
3	56576	45147.7	46.84	46.84	35.39	35.48	17.33	17.33	0.31	0.31	0.13	0.13
4	69783	70254.5	46.84	46.84	35.39	35.48	17.33	17.33	0.31	0.31	0.13	0.13
5	17930	17912.5	46.84	46.84	35.39	35.48	17.33	17.33	0.31	0.31	0.13	0.13
6	14109	14329.7	56.3	56.32	42.88	43.12	0.1	0	0.15	0.16	0.07	0.09
7	2885.4	2770.2	34.0	34.0	65.6	65.6	0	0	0.14	0.13	0	0
8	10387	10085.6	9.69	10.18	90.2	82.83	0	0	0.11	0.11	0	0
9	6233.6	6032.8	9.69	10.26	90.2	81.62	0	0	0.11	0.11	0	0
10	89288	86287.5	18.98	18.98	51.87	51.87	28.75	38.75	0.20	0.2	0.2	0.2
11	16190	16190.9	46.84	46.84	35.39	35.39	17.33	17.33	0.31	0.30	0.13	0.13
12	12903	13103.5	57.23	56.8	41.96	42.46	0.1	0.04	0.16	0.16	0.55	0.38
13	6698.6	6702.5	99.26	99.77	0.04	0.04	0	0	0.4	0.4	0.22	0.76
14	4717.3	4687.1	13	13.17	86	82.33	0	0.1	0.14	0.09	0.01	0
15	54376	54386.2	21.12	21.12	50.70	50.7	27.78	27.78	0.2	0.12	0.2	0.2
16	37903	38102.5	31.08	31.08	66.32	66.32	2.20	2.2	0.2	0.2	0.2	0.2
17	unm.	45145.5	unm.	29.41	unm.	67.93	unm.	1.84	unm.	0.43	unm.	0.17
18	6929.6	7029.5	99.02	99.43	0.04	0.03	0	0	0.56	0.56	0.21	0.19
19	3169.4	3262.6	unm.	0	unm.	96.2	unm.	0	unm.	0.01	unm.	0
20	54676	55676.2	33.50	33.5	63.97	63.97	0	0	0.2	0.18	0.2	0.2
21	unm.	2361.5	13	12.67	unm.	88.67	0	0.05	0.14	0.12	0.01	0
22	unm.	4152.8	9.69	10.05	90.20	84.74	0	0	0.11	0.11	0	0
23	unm.	7724.1	unm.	9.52	unm.	81.28	unm.	0	unm.	0	unm.	0
24	unm.	6834.3	10.6	13.02	86	84.27	0	0.09	0.14	0.13	0.01	0
25	unm.	0.0042	—	—	—	—	—	—	—	—	—	—
26	unm.	9288.5	—	—	—	—	—	—	—	—	—	—
27	unm.	14561.0	unm.	0.028	unm.	0	unm.	97.92	unm.	0	unm.	0.33
28	unm.	3234.2	unm.	0	unm.	0	unm.	100	unm.	0	unm.	0
29	unm.	2780.4	unm.	0	unm.	14.76	unm.	86.63	unm.	0	unm.	0.2
30	unm.	16235.0	unm.	0	unm.	32.76	unm.	74.17	unm.	0	unm.	0.2

Note: unm. is unmeasurement; — is dummy stream.

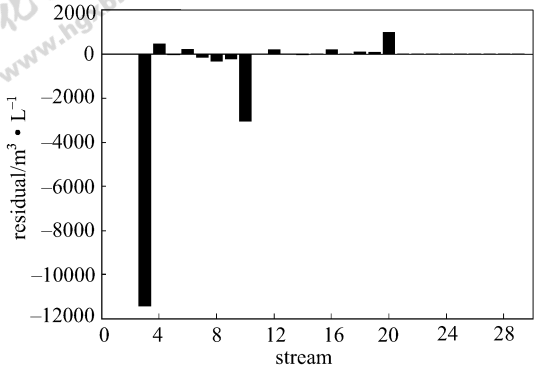

图 5　已测流股的测量残差示意图

Fig. 5　Measurements residual for measured streams

明晰，计算简单，非常适用于单元模块功能性不强以及组成部分测量的工艺场合。

4　结　论

基于污染正态分布的鲁棒最小二乘数据校正方法较传统基于正态分布的最小二乘方法具有更好的鲁棒性。然而，污染正态模型需要知道显著误差的污染程度及先验方差，这在实际中难以获得，采用固定的显著误差方差也不符合实际情况，因而其应用受到了限制。针对此方法的不足，本文提出了测量误差的鲁棒自适应分布密度模型，该模型从观测残差入手，通过鲁棒估计权函数计算自适应权重因子，将该因子用于调节误差方差，通过放大显著误差的方差，减小其对参数估计的影响。将该模型用于双线性约束数据校正问题，并采用 Lagrange 乘子法得到鲁棒自适应最小二乘分析解，同时，本文还对相关测量问题进行了研究。从仿真计算结果可以看到，该方法简单易行，具有更好的实用性，可以用于稳态及动态生产过程。

符　号　说　明

$\boldsymbol{A}$——模型系数矩阵

A_x，A_y——中间变量

b_x，b_y——分别为显著误差与随机误差方差比

c——Huber 模型参数

e——收敛精度

G_x，G_y——中间变量

$\boldsymbol{H}_x$，$\boldsymbol{H}_y$——分别为变量 x 和 y 的选择矩阵

k——鲁棒自适应调节因子

P_x，P_y——误差分布密度

p——概率密度

$\boldsymbol{W}_x$，$\boldsymbol{W}_y$——分别为变量 x 和 y 的权重矩阵

w——鲁棒权重

x——广度性质变量

y——强度性质变量

α——置信水平

$\boldsymbol{\Gamma}$——Lagrange 函数

γ——显著误差发生概率

ε——测量标准残差

η——中间变量

$\boldsymbol{\lambda}$——Lagrange 乘子

$\boldsymbol{\mu}$——Lagrange 乘子

$\boldsymbol{\Sigma}_x$，$\boldsymbol{\Sigma}_y$——随机误差协方差矩阵

$\hat{\boldsymbol{\Sigma}}_x$，$\hat{\boldsymbol{\Sigma}}_y$——误差协方差矩阵

σ_x，σ_y——方差

$\boldsymbol{\Phi}$——目标函数

下角标

i——观测数据个数

j——误差分类

k——迭代次数

lb——下限

m——测量值

true——真值

ub——上限

References

[1] Narasimhan S, Jordache C. Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data. Houston: Gulf Publishing Company, 2000

[2] José A Romagnoli, Mabel Cristina Sanchez. Data Processing and Reconciliation for Chemical Process Operations. London: Academic Press, 2000

[3] Yuan Yonggen (袁永根), Li Huasheng (李华生). Reconciliation Technology of Process Measurement (过程系统测量数据校正技术). Beijing: China Petrochemical Press, 1996

[4] Kuehn D R, Davidsonson H. Computer control (Ⅱ): Mathematics of control. *Chemical Engineering Progress*, 1961, **57** (6): 44-47

[5] Crowe C M. Reconciliation of process flow rates by matrix projection (Ⅱ): The nonlinear case. *AIChE J.*, 1986, **32**: 617-623

[6] Bagajewicz M J. On the probability distribution and reconciliation of process plant data. *Computers & Chemical Engineering*, 1996, **20** (6/7): 814-819

[7] Tjoa I B, Biegler L T. Simultaneously strategies for data reconciliation and gross error detection of nonlinear systems. *Computers & Chemical Engineering*, 1991, **15** (10): 679-690

[8] Albuquerque J S, Biegler L T. Data reconciliation and gross-error detection for dynamic systems. *AIChE J.*, 1996, **42**: 2841-2856

[9] Arora N, Biegler L T. Redescending estimators for data reconciliation and parameter estimation. *Computers & Chemical Engineering*, 2001, **25** (11/12): 1585-1599

[10] Ozyurt D B, Pike R W. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Computers & Chemical Engineering*, 2004, **28** (3): 381-402

[11] Wang D, Romagnoli J A. A framework for robust data reconciliation based on a generalized objective function. *Industrial & Engineering Chemistry Research*, 2003, **42** (13): 3075-3084

[12] Wang D, Romagnoli J A. Generalized T distribution and its applications to process data reconciliation and process monitoring. *Transactions of the Institute of Measurement and Control*, 2005, **27** (5): 367-390

[13] Johnston L P M, Kramer M A. Maximum likelihood data rectification: steady-state systems. *AIChE J.*, 1995, **41** (11): 2415-2426

[14] Ragot J, Chadli M, Maquin D. Mass balance equilibration: a robust approach using contaminated distribution. *AIChE J.*, 2005, **51** (5): 1569-1575

[15] Ragot J, Maquin D, Alhaj-Dibo A. Linear mass balance equilibration: a new approach for an old problem. *ISA Transactions*, 2005, **44** (1): 24-34