doi: 10.12011/1000-6788(2018)09-2416-08 中图法分类号: TP202.2 文献标志码: A

基于 t 分布噪声的鲁棒 PPLS 回归模型

李庆华, 陈家益, 潘 丰, 赵忠盖

(江南大学 轻工过程先进控制教育部重点实验室, 无锡 214122)

摘 要 概率 PLS(PPLS) 模型中, 数据源 (主元) 和噪声满足正态分布, 容易受离群点的影响. 鲁棒 PPLS 算法采用拖尾更长的 t 分布描述数据源和噪声, 提高了模型的鲁棒性. 但是, 实际工业过程中, 离群点由测量噪声导致, 而不是由产生过程变量和质量变量的数据源产生. 基于此, 提出一种基于 t 分布噪声的鲁棒 PPLS 模型. 该模型采用 t 分布拟合测量噪声的分布, 而主元依然用标准正态分布描述, 更符合实际测量状况. 考虑到潜在变量的存在, 采用极大似然方法结合 EM 算法对模型的参数进行了估计, 并将该模型用于对过程变量和质量变量的回归估计. 最后, 通过仿真实例进行了验证.

关键词 概率偏最小二乘算法; t 分布噪声; 回归模型; 参数估计

Robust PPLS regression modeling subject to t-distributed noise

LI Qinghua, CHEN Jiayi, PAN Feng, ZHAO Zhonggai

(Ministry of Education Key Laboratory of Advanced Process Control for Light Industry, Jiangnan University, Wuxi 214122, China)

Abstract In the probabilistic partial least squares (PPLS) model, both the data source (scores) and noise satisfy Gaussian distribution, which makes it sensible to the outlier. In the existing robust PPLS model, to improve the model robustness, a t-distribution with a longer tail rather than a Gaussian distribution with a shorter tail is employed to describe the distribution of noise and scores. However, in the real industrial processes, it is the measurement noise rather than the data source which results in the outliers. To accurately extract the information hidden in the polluted data by noise, a robust PPLS model is proposed based on the t-distributed noise assumed, where similar to the PPLS model the latent variables are normally distributed to capture the character of measurements available. Due to latent variables involved in the model, the maximum likelihood method along with the EM algorithm are employed to estimate the model parameter. Furthermore, the proposed model is used for the estimation of process variables and quality variables. Finally, a simulation case is employed to verify the proposed model.

Keywords probabilistic partial least squares algorithm; t-distributed noise; regression model; parameter estimation

1 引言

在工业过程生产中, 人们对数据重要性的认识越来越清晰. 随着大数据技术的普及和推广, 如何提取数据中的有效信息成为工业界和学术界的关注焦点 [1]. 主元分析 (PCA) 通过将数据投影到主元轴, 实现了主元信息与噪声信息的分离 [2]. 而偏最小二乘 (PLS) 则在保证过程变量的主元和质量变量的主元相关性最大的前提下, 建立了质量变量和过程变量之间的关系模型 [3]. 上述方法中, 主元变量和残差为确定向量, 而实

收稿日期: 2017-04-05

作者简介: 李庆华 (1976–), 女, 湖北襄阳人, 博士研究生, 研究方向: 工业过程监控与诊断.

基金项目: 国家自然科学基金 (61273131)
Foundation item: National Natural Science Foundation of China (61273131)
中文引用格式: 李庆华, 陈家益, 潘丰, 等. 基于 t 分布噪声的鲁棒 PPLS 回归模型 [J]. 系统工程理论与实践, 2018, 38(9): 2416–2423.
英文引用格式: Li Q H, Chen J Y, Pan F, et al. Robust PPLS regression modeling subject to t-distributed noise[J]. Systems Engineering — Theory & Practice, 2018, 38(9): 2416–2423.

际工业过程中, 由于随机噪声的存在, 用随机分布描述样本数据更符合统计规律 [4]. 与常规的投影方法不同, 概率数据模型采用概率分布方式描述主元和噪声, 具有更大的适应性, 并且经典概率推理的成果, 比如贝叶斯等方法, 能够直接应用于数据的建模中. 概率 PCA(PPCA) 和概率 PLS(PPLS) 分别是 PCA 和 PLS 的概率表达形式 [5,6]. 与 PCA 和 PLS 不同, 在概率形式下, 测量数据的解释方式发生了变化: 前者中, 主元由测量数据通过去除噪声后压缩提取, 主元是由测量数据产生; 后者中, 主元被认为是数据源, 分布已知, 变量则由多个数据源混合加上测量噪声产生 [7]. 显然, 后者的解释能够更好地描述测量数据的产生本质.

在实际生产过程中, 受测量环境的影响, 变量的测量存在较大的干扰, 离群点现象普遍存在, 给概率模型的建立带来了困难: 高斯分布拖尾短, 离群点会直接影响高斯分布的均值和方差, 导致概率模型鲁棒性较差, 对离群点较为敏感; 而忽略毛刺点, 则可能会丢失数据中部分有用信息. 相比于正态分布, t 分布具有较长的拖尾, 并且能够根据离群点的大小和数量, 改变拖尾长度, 从而在基本不改变均值和方差的情况下, 实现对离群点的拟合. 基于此, t 分布被引入到概率模型中, 提高了模型的鲁棒性 [8,9]. 鲁棒 PPCA 模型和鲁棒 PPLS 分别在 PPCA 和 PPLS 模型框架下 [10,11], 采用 t 分布代替高斯分布用于对主元和噪声的描述, 能够较好地克服离群点的影响. 但是, 离群点通常由测量误差导致, 而不是由数据源产生, 即数据源本身没有发生变化. 另一方面, 概率模型是一种生成模型, 测量数据由已知分布的主元信息主导生成, 在模型辨识过程中, 未知参数最终由噪声分布和回归参数组成, 而如果采用 t 分布描述主元信息, 则增加了模型中的未知参数, 导致模型复杂. Sadeghian 等提出了一种噪声满足两个正态分布加权和的鲁棒 PPLS[12], 其中, 主元采用正态分布描述. 但是这种方法只是针对噪声分布的特殊情形, 适应性较弱, 且计算量很大.

基于概率模型的基本特性, 本文改进 RPPLS 模型, 提出一种基于 t 分布噪声的鲁棒 PPLS 模型. 该模型引入 t 分布用于描述噪声, 而主元信息与 PPLS 模型一样满足标准正态分布, 使离群点对模型的影响局限于噪声, 提高了模型的鲁棒性. 本文的主要结构如下: 首先简要介绍 PPLS 模型; 第 3 节则分析 RPPLS 模型; 基于 t 分布噪声的鲁棒 PPLS 模型在第 4 节提出; 最后是仿真验证.

2 PPLS 模型

2.1 PPLS 模型

假设 $M_{\boldsymbol{x}}$ 个过程变量和 $M_{\boldsymbol{y}}$ 个输出变量的 N 组测量数据分别为: $\boldsymbol{X} = \{\boldsymbol{x}_n|\boldsymbol{x}_n \in R^{M_x}\}_{n=1}^N$, $\boldsymbol{Y} = \{\boldsymbol{y}_n|\boldsymbol{y}_n \in R^{M_y}\}_{n=1}^N$. 假设主元个数为 $q < M_x$, 过程变量和质量变量的均值和模型参数分别为 μ_x, μ_y, $\boldsymbol{P} \in R^{M_x\times q}$, $\boldsymbol{C} \in R^{M_y\times q}$, 则 PPLS 模型如下 [13]:

$$\boldsymbol{x}_n = \boldsymbol{P}\boldsymbol{t}_n + \mu_x + \xi_n, \tag{1}$$

$$\boldsymbol{y}_n = \boldsymbol{C}\boldsymbol{t}_n + \mu_y + \varepsilon_n. \tag{2}$$

其中, 过程变量误差 $\xi_n \sim N(0, \sigma_x^2\boldsymbol{I}_{M_x})$, 质量变量噪声 $\varepsilon_n \sim N(0, \sigma_y^2\boldsymbol{I}_{M_y})$, 主元 $\boldsymbol{t}_n \sim N(0, \boldsymbol{I}_q)$, $\boldsymbol{I}_\triangle$ 表示 $\triangle$ 维的单位矩阵, 为简略计, 后续都采用 $\boldsymbol{I}$ 表示.

从 PPLS 模型可以看出, 主元满足标准正态分布, 测量数据在一定均值基础上, 由主元组合并伴随着未知的测量噪声生成. PPLS 模型的未知参数为 $\Theta = (\boldsymbol{P}, \boldsymbol{C}, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$. 基于 PPLS 模型结构, 将主元作为隐变量, 可采用极大似然方法结合 EM 算法对参数进行估计, 具体可参见文献 [13].

2.2 离群点对 PPLS 模型的影响

不失一般性, 以离群点对一维正态分布数据的影响进行说明. 假设数据的均值和方差分别为 μ 和 σ^2. 根据正态分布的特性, 数据在范围 $[\mu - 3\sigma, \mu + 3\sigma]$ 之外的概率为 0.26%. 但是, 实际生产过程中, 检测数据离群点出现的概率远大于 0.26%. 为了在概率上保证离群点被包含于正态分布, 则范围 $[\mu - 3\sigma, \mu + 3\sigma]$ 必然要扩大, 最终使正态分布的均值和方差改变. 反映在 PPLS 模型上, 由于主元满足标准正态分布, 为了使离群点在概率上满足正态分布, 模型的参数会发生变化, 最终使得测量变量的分布范围很大. 因此, PPLS 模型对离群点较为敏感, 鲁棒性较差.

3 RPPLS 模型

3.1 RPPLS 模型

RPPLS 模型结构与 PPLS 相同, 如式 (1) 和式 (2) 所示. 模型中主元和噪声的分布都满足 t 分布, 即 $\xi_n \sim t(0,\sigma_x^2\boldsymbol{I},v)$, $\varepsilon_n \sim t(0,\sigma_y^2\boldsymbol{I},v)$, $\boldsymbol{t}_n \sim t(0,\boldsymbol{I},v)$, 其中, 自由度, 均值和系统矩阵分别为 v, μ 和 Σ 的 t 分布 $t(\mu,\Sigma,v)$ 表示为:

$$p(\boldsymbol{x};\mu,\Sigma,v)=\frac{\Gamma((M_x+v)/2)|\Sigma|^{-1/2}}{\Gamma(v/2)(v\pi)^{M_x/2}}[1+(\boldsymbol{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\mu)/v]^{-(M_x+v)/2}. \tag{3}$$

其中, $\Gamma(\cdot)$ 表示 Gamma 函数, 其概率密度函数定义为: $\Gamma(x)=\int_0^\infty z^{x-1}\mathrm{e}^{-z}\mathrm{d}z$. 当 $v>2$ 时, x 的协方差矩阵等于 $v\Sigma/(v-2)$. 因此, t 分布可在不改变分布的均值, 很少改变协方差矩阵的情况下, 通过调整自由度 v 的大小, 改变拖尾宽度, 从而使分布能够涵盖离群点, 相比于正态分布, 具有极好的鲁棒性.

3.2 问题分析

PPLS 模型中主元和噪声都满足正态分布, 过程变量和质量变量由主元和噪声混合得到. PPLS 在过程监控、质量变量的预测中得到了较好的应用. 特别地, PPLS 模型中, 有一个正态分布数据源, 符合很多数据的产生特性.

实际工业过程中, 过程运行故障主要表现在数据源之间的相关性发生变化或者数据源本身发生变化. 不同于过程运行故障, 离群点主要由传感器故障、人工操作错误、实验分析错误等原因造成, 本质上是测量出现了较大的噪声, 导致正常正态分布数据源和噪声的混合值过大, 偏离正常数据范围, 而过程的数据源本身没有发生变化. 在鲁棒 PPLS 中, 为提高模型的鲁棒特性, 假定过程变量和质量变量由满足 t 分布的噪声和主元产生. 显然, 对主元的分布进行歪曲, 不符合主元的真实分布.

基于以上分析, 本文提出一种基于 t 分布噪声的鲁棒 PPLS 方法, 其中主元 (数据源) 与 PPLS 方法一样, 服从正态分布, 而噪声用于拟合测量噪声, 假定为服从 t 分布.

4 tRPPLS 模型

4.1 tRPPLS 模型

模型结构与 RPPLS 和 PPLS 模型相同, 只是主元和噪声的分布分别为 $\boldsymbol{t}_n \sim N(0,\boldsymbol{I})$, $\xi_n \sim t(0,\sigma_x^2\boldsymbol{I},v)$, $\varepsilon_n \sim t(0,\sigma_y^2\boldsymbol{I},v)$. 由于引入了 t 分布, 很难直接求解过程变量、质量变量和主元的概率分布. 根据 $\boldsymbol{x} \sim t(\mu,\Sigma,v)$ 的特性 [14], 如果任意随机变量 $\boldsymbol{u}_n$ 满足伽马分布 $p(\boldsymbol{u}_n)=G_a(\boldsymbol{u}_n;v/2,v/2)$, 则 $\boldsymbol{x}$ 关于 $\boldsymbol{u}_n$ 的条件分布 $\boldsymbol{x}|\boldsymbol{u}_n$ 服从正态分布. 因此, 可以通过引入中间随机变量 $\boldsymbol{u}_n$, 将过程变量、质量变量和主元的分布转化成正态分布计算, 简化了计算过程, 具体分布如下:

$$p(\xi_n|\boldsymbol{u}_n)=N(0,\sigma_x^2\boldsymbol{I}/v), \tag{4}$$

$$p(\varepsilon_n|\boldsymbol{u}_n)=N(0,\sigma_y^2\boldsymbol{I}/v), \tag{5}$$

$$p(\boldsymbol{t}_n)=N(0,\boldsymbol{I}). \tag{6}$$

则 $\boldsymbol{x}_n$, $\boldsymbol{y}_n$ 关于 $\boldsymbol{t}_n$, $\boldsymbol{u}_n$ 的条件分布和联合分布如下:

$$p(\boldsymbol{x}_n|\boldsymbol{t}_n,\boldsymbol{u}_n)=N(\boldsymbol{P}\boldsymbol{t}_n+\mu_x,\sigma_x^2\boldsymbol{I}/v), \tag{7}$$

$$p(\boldsymbol{y}_n|\boldsymbol{t}_n,\boldsymbol{u}_n)=N(\boldsymbol{C}\boldsymbol{t}_n+\mu_y,\sigma_y^2\boldsymbol{I}/v), \tag{8}$$

$$p(\boldsymbol{z}_n|\boldsymbol{t}_n,\boldsymbol{u}_n)=N(\boldsymbol{W}\boldsymbol{t}_n+\mu,\Phi/v). \tag{9}$$

其中: $\boldsymbol{z}_n=\begin{pmatrix}\boldsymbol{x}_n\\ \boldsymbol{y}_n\end{pmatrix}$, $\boldsymbol{W}=\begin{pmatrix}\boldsymbol{P}\\ \boldsymbol{C}\end{pmatrix}$, $\mu=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}$, $\Phi=\begin{pmatrix}\sigma_x^2\boldsymbol{I} & 0\\ 0 & \sigma_y^2\boldsymbol{I}\end{pmatrix}$. 记 $\Theta=(\boldsymbol{P},\boldsymbol{C},\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)$ 为待估计的参数. 对于隐含变量 $p(\boldsymbol{t}_n,\boldsymbol{u}_n|\boldsymbol{z}_n;\Theta)$ 期望值的计算分成如下两个部分求解:

$$p(\boldsymbol{t}_n,\boldsymbol{u}_n|\boldsymbol{z}_n;\Theta)=p(\boldsymbol{u}_n|\boldsymbol{z}_n;\Theta)p(\boldsymbol{t}_n|\boldsymbol{z}_n;\Theta). \tag{10}$$

第一部分运用贝叶斯公式可以得到 $\boldsymbol{u}_n$ 关于 $\boldsymbol{z}_n$ 的后验分布 [15]:

$$p(\boldsymbol{u}_n|\boldsymbol{z}_n;\Theta)\propto p(\boldsymbol{z}_n|\boldsymbol{u}_n;\Theta)p(\boldsymbol{u}_n;\Theta)=G_a\left(\boldsymbol{u}_n;\frac{D_x+D_y+v}{2},\frac{(\boldsymbol{z}_n-\mu)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{z}_n-\mu)+v}{2}\right). \tag{11}$$

第二部分可以直接得到 $\boldsymbol{t}_n$ 关于 $\boldsymbol{z}_n$, $\boldsymbol{u}_n$ 的后验分布:

$$p(\boldsymbol{t}_n|\boldsymbol{z}_n;\Theta)=\frac{p(\boldsymbol{z}_n|\boldsymbol{t}_n;\Theta)p(\boldsymbol{t}_n)}{p(\boldsymbol{z}_n)}=N(\boldsymbol{B}\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}(\boldsymbol{z}_n-\mu),\boldsymbol{B}). \tag{12}$$

则 t_n 的期望如下:

$$\langle\boldsymbol{t}_n\rangle=\boldsymbol{B}\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}(\boldsymbol{z}_n-\mu). \tag{13}$$

其中, $\boldsymbol{A}=(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\Phi)^{-1}$, $\boldsymbol{B}=(\boldsymbol{I}+\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}\boldsymbol{W})^{-1}$.

4.2 模型求解

极大似然方法是求解概率模型参数的有效方法, 但是在 tRPPLS 模型中 $\{\boldsymbol{u}_n,\boldsymbol{t}_n\}$ 与似然函数直接相关, 且未知, 因此将其作为缺失数据, 引入 EM 算法来求解参数. 似然函数如下:

$$L(\Theta)=\ln\prod_{n=1}^{N}p(\boldsymbol{z}_n,\boldsymbol{t}_n,\boldsymbol{u}_n;\Theta)=\sum_{n=1}^{N}\ln\{p(\boldsymbol{z}_n,\boldsymbol{t}_n,\boldsymbol{u}_n;\Theta)\}=\sum_{n=1}^{N}\ln\{p(\boldsymbol{z}_n|\boldsymbol{t}_n,\boldsymbol{u}_n;\Theta)p(\boldsymbol{t}_n|\boldsymbol{u}_n;\Theta)p(\boldsymbol{u}_n;\Theta)\}. \tag{14}$$

EM 算法的求解步骤分成两步, E 步和 M 步, 具体如下:

在 E 步, 根据 $p(\boldsymbol{t}_n,\boldsymbol{u}_n|\boldsymbol{z}_n;\Theta)$ 的期望求 $L(\Theta)$ 期望 $\langle L(\Theta)\rangle$, 如下:

$$\begin{aligned}\langle L(\Theta)\rangle=-\sum_{n=1}^{N}\Bigg\{&\frac{D_x}{2}\ln\sigma_x^2+\frac{\langle\boldsymbol{u}_n\rangle}{2\sigma_x^2}(\boldsymbol{x}_n-\mu_x)^{\mathrm{T}}(\boldsymbol{x}_n-\mu_x)-\frac{1}{\sigma_x^2}\langle\boldsymbol{u}_n\boldsymbol{t}_n\rangle^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}(\boldsymbol{x}_n-\mu_x)+\\&\frac{1}{2\sigma_x^2}\mathrm{tr}\{\langle\boldsymbol{u}_n\boldsymbol{t}_n\boldsymbol{t}_n^{\mathrm{T}}\rangle\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\}+\frac{D_y}{2}\ln\sigma_y^2+\frac{\langle\boldsymbol{u}_n\rangle}{2\sigma_y^2}(\boldsymbol{y}_n-\mu_y)^{\mathrm{T}}(\boldsymbol{y}_n-\mu_y)-\\&\frac{1}{\sigma_y^2}\langle\boldsymbol{u}_n\boldsymbol{t}_n\rangle^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}(\boldsymbol{y}_n-\mu_y)+\frac{1}{2\sigma_y^2}\mathrm{tr}\{\langle\boldsymbol{u}_n\boldsymbol{t}_n\boldsymbol{t}_n^{\mathrm{T}}\rangle\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\}+\frac{v}{2}\ln\frac{v}{2}+\\&\left(\frac{v}{2}-1\right)\langle\ln\boldsymbol{u}_n\rangle-\ln\Gamma\left(\frac{v}{2}\right)-\frac{v}{2}\langle\boldsymbol{u}_n\rangle\Bigg\}.\end{aligned} \tag{15}$$

其中,

$$\langle\boldsymbol{u}_n\rangle=\frac{v+D_x+D_y}{(\boldsymbol{z}_n-\mu)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{z}_n-\mu)+v}, \tag{16}$$

$$\langle\ln\boldsymbol{u}_n\rangle=\Psi\left(\frac{v+\boldsymbol{D}_x+\boldsymbol{D}_y}{2}\right)-\ln\left(\frac{(\boldsymbol{z}_n-\mu)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{z}_n-\mu)+v}{2}\right), \tag{17}$$

$$\langle\boldsymbol{t}_n\rangle=\boldsymbol{B}\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}(\boldsymbol{z}_n-\mu), \tag{18}$$

$$\langle\boldsymbol{u}_n\boldsymbol{t}_n\rangle=\langle\boldsymbol{u}_n\rangle\langle\boldsymbol{t}_n\rangle, \tag{19}$$

$$\langle\boldsymbol{u}_n\boldsymbol{t}_n\boldsymbol{t}_n^{\mathrm{T}}\rangle=\langle\boldsymbol{u}_n\rangle\boldsymbol{B}+\langle\boldsymbol{u}_n\rangle\langle\boldsymbol{t}_n\rangle\langle\boldsymbol{t}_n\rangle^{\mathrm{T}}. \tag{20}$$

在 M 步, 求期望 $\langle L(\Theta)\rangle$ 的极大值, 从而得到参数的更新估计值 $\tilde{\Theta}$. 对 $\langle L(\Theta)\rangle$ 关于各参数分别求偏导, 并使之等于 0, 得更新值如下:

$$\tilde{\mu}=\begin{pmatrix}\tilde{\mu}_x\\\tilde{\mu}_y\end{pmatrix}=\frac{\sum_{n=1}^{N}\langle\boldsymbol{u}_n\rangle(\boldsymbol{z}_n-\boldsymbol{W}\langle\boldsymbol{t}_n\rangle)}{\sum_{n=1}^{N}\langle\boldsymbol{u}_n\rangle}, \tag{21}$$

$$\tilde{\boldsymbol{W}}=\begin{pmatrix}\tilde{\boldsymbol{P}}\\\tilde{\boldsymbol{C}}\end{pmatrix}=\left(\sum_{n=1}^{N}(\boldsymbol{z}_n-\tilde{\mu})\langle\boldsymbol{u}_n\boldsymbol{t}_n\rangle^{\mathrm{T}}\right)\left(\sum_{n=1}^{N}\langle\boldsymbol{u}_n\boldsymbol{t}_n\boldsymbol{t}_n^{\mathrm{T}}\rangle\right)^{-1}, \tag{22}$$

$$\tilde{\sigma_x^2}=\frac{1}{ND_x}\sum_{n=1}^{N}\{\langle\boldsymbol{u}_n\rangle\|\boldsymbol{x}_n-\tilde{\mu}_x\|^2-2\langle\boldsymbol{u}_n\boldsymbol{t}_n\rangle^{\mathrm{T}}\tilde{\boldsymbol{P}}^{\mathrm{T}}(\boldsymbol{x}_n-\tilde{\mu}_x)+\mathrm{tr}\{\langle\boldsymbol{u}_n\boldsymbol{t}_n\boldsymbol{t}_n^{\mathrm{T}}\rangle\tilde{\boldsymbol{P}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\}\}, \tag{23}$$

$$\tilde{\sigma_y^2}=\frac{1}{ND_y}\sum_{n=1}^{N}\{\langle\boldsymbol{u}_n\rangle\|\boldsymbol{y}_n-\tilde{\mu}_y\|^2-2\langle\boldsymbol{u}_n\boldsymbol{t}_n\rangle^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}(\boldsymbol{y}_n-\tilde{\mu}_y)+\mathrm{tr}\{\langle\boldsymbol{u}_n\boldsymbol{t}_n\boldsymbol{t}_n^{\mathrm{T}}\rangle\tilde{\boldsymbol{C}}^{\mathrm{T}}\tilde{\boldsymbol{C}}\}\}, \tag{24}$$

$$1+\ln\left(\frac{\tilde{v}}{2}\right)-\Psi\left(\frac{\tilde{v}}{2}\right)+\frac{1}{N}\sum_{n=1}^{N}(\langle\ln\boldsymbol{u}_n\rangle-\langle\boldsymbol{u}_n\rangle)=0. \tag{25}$$

其中, 函数 $\Psi(x)=\mathrm{d}\ln\Gamma(x)/\mathrm{d}\boldsymbol{x}$, 式 (25) 没有给出 v 的解析表达式, 可采用数值求解方法求解该式得到 v 的估计值. 反复迭代 E 步和 M 步, 直到收敛, tRPPLS 模型建立完成. EM 算法可以保证收敛到一个稳定点, 关于 EM 算法的收敛性, 文献 [16] 进行了证明, 本文不再赘述. 主元个数的确定方法同 PPLS 法, 假定主元个数已经确定, 则具体算法步骤如表 1.

表 1 tRPPLS 模型的 EM 求解

1. 初始化: $\mu_x = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$, $\mu_y = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{y}_n$, $\sigma_x = \sigma_y = 1$, $k = 0$, $\boldsymbol{P} = \boldsymbol{X}(:, 1:q)$, $\boldsymbol{C} = \boldsymbol{Y}(:, 1:q)$. 确定 Θ_k.
2. E 步: 计算 Θ_k 下的似然函数的期望: $\boldsymbol{Q}(\Theta, \Theta_k) = E_{p(\boldsymbol{t}_n, \boldsymbol{u}_n|\boldsymbol{z}_n;\Theta)}\log(p(\boldsymbol{z}_n|\boldsymbol{t}_n, \boldsymbol{u}_n))$.
3. M 步: 求 $\boldsymbol{Q}(\Theta, \Theta_k)$ 极值, 得到 Θ_{k+1} 为 $\Theta_{k+1} = \mathrm{argmax}Q(\Theta, \Theta_k)$.
4. 判断是否收敛, 如果不收敛, 则令 $k = k + 1$, 跳转到第 2 步; 如果收敛, 则 Θ_{k+1} 即为所求参数.

4.3 基于 tRPPLS 模型的回归

tRPPLS 模型提供了质量变量和过程变量分别与数据源之间的相关关系. 根据建立的 tRPPLS 模型, 则关于过程变量和质量变量的估计为:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{P}\boldsymbol{t}_n + \hat{\mu}_x, \tag{26}$$

$$\hat{\boldsymbol{y}}_n = \boldsymbol{C}\boldsymbol{t}_n + \hat{\mu}_y. \tag{27}$$

其中, 先验知识下, 主元 $\boldsymbol{t}_n$ 为满足标准正态分布的随机量. 但是, 当采样时刻 n 下, 过程变量和质量变量的测量值已知, 则可对 $\boldsymbol{t}_n$ 的分布进行实时更新, 得到后验概率如式 (12) 所示, 即 $\boldsymbol{t}_n \sim N(\boldsymbol{B}\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}(\boldsymbol{z}_n - \mu), \boldsymbol{B})$. 根据参考文献 [17], 采用 $\boldsymbol{t}_n$ 的分布的均值 $\bar{t}_n$ 代替 $\boldsymbol{t}_n$, 则质量变量的估计值为:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{P}\boldsymbol{B}\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}(\boldsymbol{z}_n - \mu) + \hat{\mu}_x, \tag{28}$$

$$\hat{\boldsymbol{y}}_n = \boldsymbol{C}\boldsymbol{B}\boldsymbol{W}^{\mathrm{T}}\Phi^{-1}(\boldsymbol{z}_n - \mu) + \hat{\mu}_y. \tag{29}$$

5 仿真实例

给定一个多变量过程, 含 7 个过程变量, 3 个质量变量, 分别有 4 个隐含变量线性组合而成, 模型如下:

$$\begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \\ \boldsymbol{x}_5 \\ \boldsymbol{x}_6 \\ \boldsymbol{x}_7 \end{pmatrix} = \begin{pmatrix} 1.9458 & -0.2848 & 0.7887 & 0.1087 \\ 0.1169 & -0.6409 & -0.3228 & -2.1190 \\ 0.1788 & -0.1272 & -0.5463 & -0.1249 \\ 0.7870 & -0.9782 & -1.3781 & -0.3241 \\ 0.7470 & -1.5114 & -2.0372 & -0.2204 \\ -1.6617 & -0.9843 & 1.9540 & -0.3603 \\ -1.3525 & -1.2072 & 0.3674 & -0.4347 \end{pmatrix} \begin{pmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \\ \boldsymbol{t}_3 \\ \boldsymbol{t}_4 \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \boldsymbol{e}_3 \\ \boldsymbol{e}_4 \\ \boldsymbol{e}_5 \\ \boldsymbol{e}_6 \\ \boldsymbol{e}_7 \end{pmatrix}. \tag{30}$$

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \end{pmatrix} = \begin{pmatrix} -1.0690 & 1.7710 & 0.9190 & -1.3784 \\ -1.0907 & 0.3172 & 0.3385 & -0.7937 \\ 0.0284 & 0.7830 & 0.3432 & 0.3267 \end{pmatrix} \begin{pmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \\ \boldsymbol{t}_3 \\ \boldsymbol{t}_4 \end{pmatrix} + \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \boldsymbol{f}_3 \end{pmatrix}. \tag{31}$$

其中: 隐含变量 $\boldsymbol{t}_i (i = 1, 2, 3, 4)$ 服从标准正态分布, 噪声变量 $\boldsymbol{e}_i (i = 1, \cdots, 7)$ 和 $\boldsymbol{f}_i (i = 1, 2, 3)$ 均服从均值为 0, 方差为 10^{-4} 的正态分布. 根据模型生成 200 组数据. 为了模拟实际工业过程, 在过程变量和质量变量中同时引入 3% 的离群值, 即随机选取 3% 的样本点, 然后用离群值替换掉原来的值.

在过程变量 $\boldsymbol{x}_1$ 上随机加入 3% 的离群值, 同时在质量变量 $\boldsymbol{y}_3$ 上也随机加入 3% 的离群值, 正常数据和加入离群值数据如图 1~4 所示.

根据过程变量和质量变量的数据, 分别建立 PPLS, RPPLS, tRPPLS 模型. 主元个数的选择参照文献 [17], 三个模型都选择 4 个主元, 本文不再赘述. 分别对 $\boldsymbol{x}_1$ 和 $\boldsymbol{y}_3$ 进行回归 (估计), 效果如图 5, 6 和 7 所示.

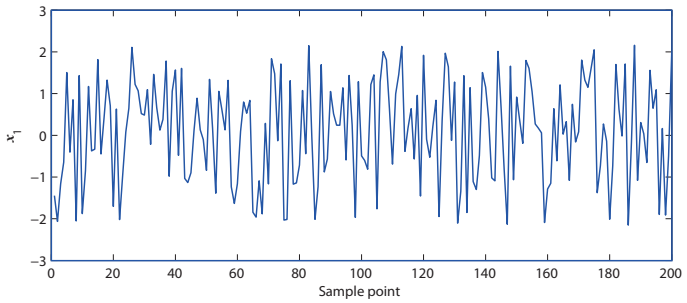

图 1 过程变量 $\boldsymbol{x}_1$ 真实值

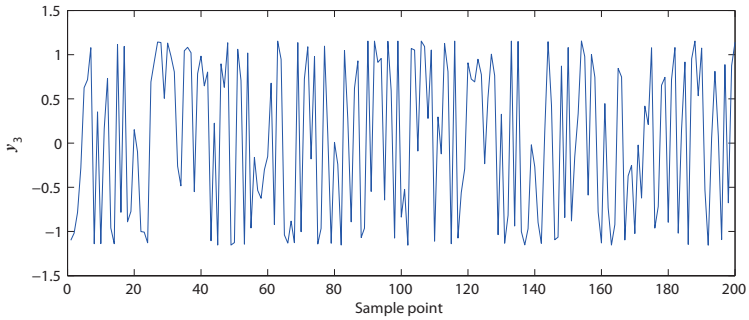

图 2　加入 3% 离群值的变量过程变量 x_1

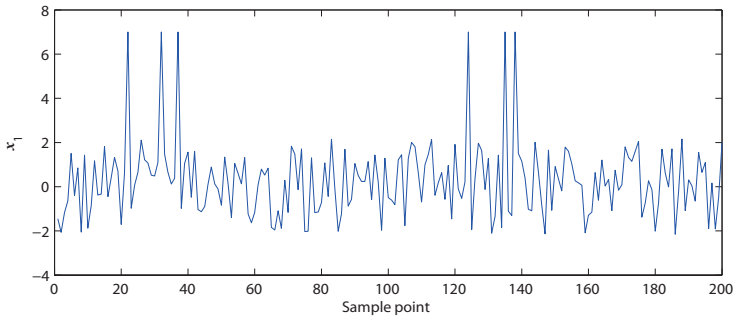

图 3　过程变量 y_3 真实值

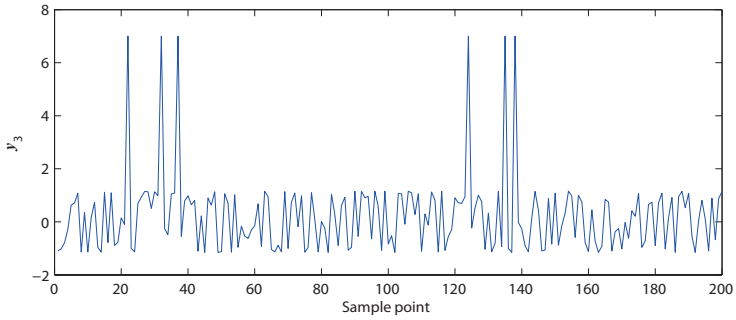

图 4　加入 3% 离群值的变量过程变量 y_3

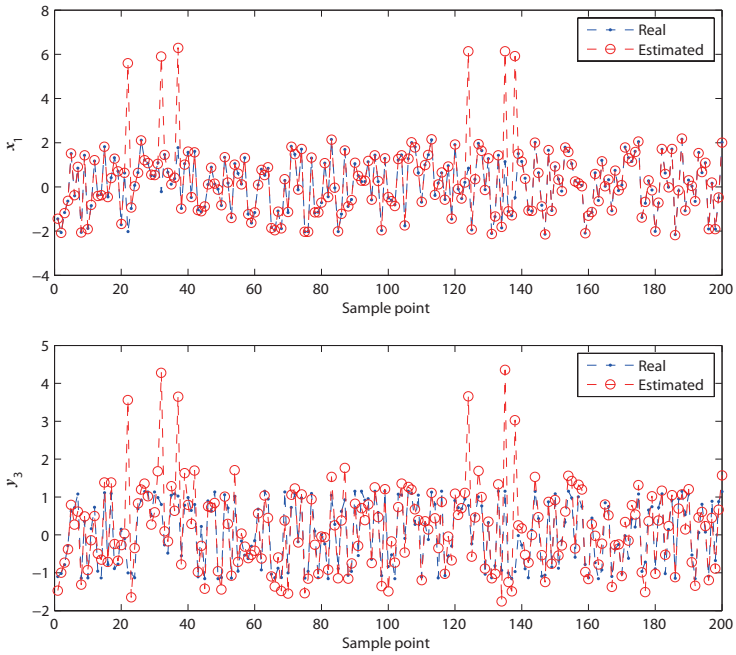

图 5　基于 PPLS 模型的回归效果

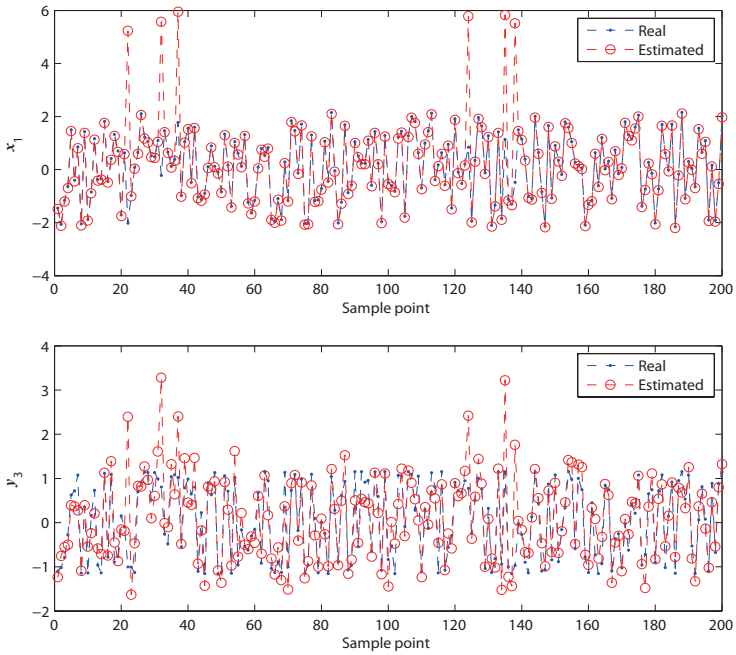

图 6 基于 RPPLS 模型的回归效果

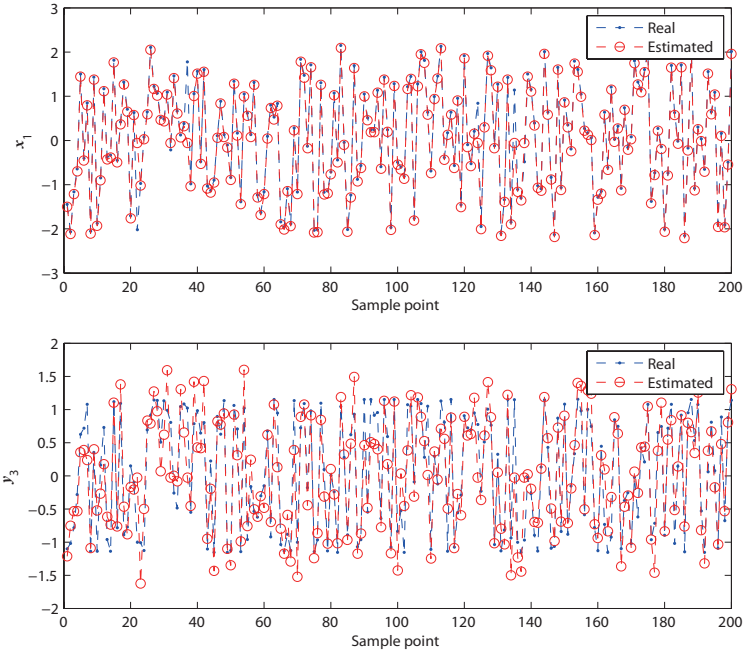

图 7 基于 tRPPLS 模型的回归效果

表 2 每个变量的估计均方误差

变量	PPLS	RPPLS	tRPPLS
$\boldsymbol{x}_1$	0.3871	0.3649	0.0854
$\boldsymbol{x}_2$	0.0743	0.0680	0.0607
$\boldsymbol{x}_3$	0.0713	0.0647	0.0261
$\boldsymbol{x}_4$	0.0467	0.0211	0.0423
$\boldsymbol{x}_5$	0.0641	0.0669	0.0695
$\boldsymbol{x}_6$	0.0459	0.0133	0.0794
$\boldsymbol{x}_7$	0.1101	0.1339	0.0431
$\boldsymbol{y}_1$	0.2585	0.2494	0.2301
$\boldsymbol{y}_2$	0.2361	0.2062	0.1848
$\boldsymbol{y}_3$	0.4023	0.3025	0.2108

对比图 5, 6 和 7 可知, 在出现离群点的时刻, PPLS 模型中, 对变量的估计无法克服离群点的影响, 因此导致估计值中也出现离群点的情况. 而 RPPLS 算法中, 由于对主元的分布进行了扭曲, 因此虽然相比 PPLS, 离群点对估计值的影响幅度减少, 但是估计值中还是出现了离群点. tRPPLS 算法则可很好地消除离群点的影响, 得到较好的估计值. 为衡量估计效果, 引入均方误差指标如下:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\boldsymbol{x}_{n,i} - \hat{\boldsymbol{x}}_{n,i})^2}{N}}. \tag{32}$$

三个模型对各个过程变量和质量变量的回归均方差值如表 2 所示. 很明显, 对质量变量, tRPPLS 算法的估计效果最好; 对过程变量, tRPPLS 算法在 4 个变量上的估计效果都是最好, 在另外 3 个变量上, 基本与其它两种方法的效果相当.

6 结论

针对工业过程中普遍存在的离群点问题, 提出一种基于 t 分布噪声的鲁棒 PPLS 算法. 该方法考虑到离群点主要由于测量噪声导致, 因此在不改变数据源分布的情况下, 采用 t 分布描述噪声, 从而使模型对离群点的鲁棒性大大提高. 更进一步, 将该模型应用到对过程变量和质量变量的回归估计中. 数值仿真例子验证了相对于常规 PPLS 和 RPPLS 算法, 该算法受离群点的影响较小, 过程变量和质量变量回归估计的效果最好.

参考文献

[1] Qin S J. Survey on data-driven industrial process monitoring and diagnosis[J]. Annual Reviews in Control, 2012, 36: 220–234.

[2] Chiang L H, Russell E L, Bratz R D. Fault detection and diagnosis in industrial systems[M]. London: Springer-Verlag London Limited, 2001: 99–106.

[3] Wold S, Sjostrom M, Eriksson L. PLS-regression: A basic tool of chemometrics[J]. Chemometrics and Intelligent Laboratory Systems, 2001, 58: 109–130.

[4] Zhao Z, Li Q, Huang B, et al. Process monitoring based on factor analysis: Probabilistic analysis of monitoring statistics in presence of both complete and incomplete measurements[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 142: 18–27.

[5] Sedghi S, Sadeghian A, Huang B. Mixture semisupervised probabilistic principal component regression model with missing inputs[J]. Computers & Chemical Engineering, 2017, 103: 176–187.

[6] Wang S, Yang J. A probabilistic model for latent least squares regression[J]. Neurocomputing, 2015, 149: 1155–1161.

[7] 赵忠盖, 刘飞. 因子分析及其在过程监控中的应用 [J]. 化工学报, 2007, 58(4): 970–974.
Zhao Z G, Liu F. Factor analysis and its application to process monitoring[J]. CIESC J (China), 2007, 58(4): 970–974.

[8] Yao W, Wei Y, Yu C. Robust mixture regression using the t-distribution[J]. Computational Statistics & Data Analysis, 2014, 71: 116–127.

[9] Chamroukhi F. Robust mixture of experts modeling using the t distribution[J]. Neural Networks, 2016, 79: 20–36.

[10] Sadeghian A, Wu O, Huang B. Robust probabilistic principal component analysis based process modeling: Dealing with simultaneous contamination of both input and output data[J]. Journal of Process Control, 2018, 67: 94–111.

[11] 陈家益, 赵忠盖, 刘飞. 鲁棒 PPLS 模型及其在过程监控中的应用 [J]. 化工学报, 2016, 67(6): 2907–2915.
Chen J Y, Zhao Z G, Liu F. Robust PPLS model and its applications in process monitoring[J]. CIESC J (China), 2016, 67(6): 2907–2915.

[12] Sadeghian A, Huang B. Robust probabilistic principal component analysis for process modeling subject to scaled mixture Gaussian noise[J]. Computers and Chemical Engineering, 2016, 90: 62–78.

[13] Li S, Gao J, Nyagilo J O, et al. Probabilistic partial least square regression: A robust model for quantitative analysis of Raman spectroscopy data[C]// IEEE International Conference on Bioinformatics and Biomedicine, 2011: 526–531.

[14] Liu C H, Rubin D B. ML-estimation of the t distribution using EM and its extensions, ECM and ECME[J]. Statistica Sinica, 1995, 5(1): 19–39.

[15] Chen T, Martin E, Montague G. Robust probabilistic PCA with missing data and contribution analysis for outlier detection[J]. Computational Statistics and Data Analysis, 2009, 53(10): 3706–3716.

[16] Wu C F J. On the convergence properties of the EM algorithm[J]. Annals of Statistics, 1983, 11(1): 95–103.

[17] Kim D, Lee I B. Process monitoring based on probabilistic PCA[J]. Chemometrics and Intelligent Laboratory Systems, 2003, 67(2): 109–123.