文章编号: 1001-0920(2005)07-0823-04

鲁棒 PLS 在间歇生产过程监控中的应用

董胜利, 王树青, 谢 磊
(工业控制技术国家重点实验室, 浙江大学 先进控制技术研究所, 杭州 310027)

摘 要: 间歇和半间歇过程在化学工业中占有重要地位, 如何对其进行监控一直是过程控制领域研究的热点之一. 现实过程中, 数据大都存在离群点, 易使多向部分最小二乘(MPLS)模型造成误差. 针对MPLS统计监控受离群点影响的问题, 提出一种基于鲁棒MPLS的统计监控分析和相应鲁棒监控统计量的计算方法. 相对于普通MPLS, 鲁棒MPLS在建模数据中存在离群点时仍能给出正确的统计监控模型, 降低了建模过程对数据的要求.

关键词: 间歇过程监控; 部分最小二乘, 多向部分最小二乘; 鲁棒部分最小二乘
中图分类号: TP277 **文献标识码**: A

Robust PLS and Its Application to Batch Process Monitoring

DONG Sheng-li, WANG Shu-qing, XIE Lei
(National Key Lab of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, China. Correspondent: DONG Sheng-li, E-mail: sldong@iipc.zju.edu.cn)

Abstract: Batch and semi-batch processes play an important role in chemical industry. In order to reduce the variations of the product quality, multivariate statistical process control methods based on multi-way partial least squares (MPLS) is proposed for on-line batch process monitoring. However outliers always exist in the data, traditional MPLS methods are strongly affected by outlying observations. A batch process monitoring method based on robust MPLS is proposed. The robust normal operating condition model and robust control limits are discussed in detail. The results show that the robust MPLS is resistant to possible outliers.

Key words: Batch process monitoring; PLS; Multi-way PLS; Robust PLS

1 引 言

间歇和半间歇过程生产模式因其灵活性高、柔性好, 非常适于小批量、多品种的生产, 已成为精细化工、食品、制药等工业的重要生产方式. 但由于过程本身的复杂性以及干扰的影响, 实际操作条件往往会与生产方案有所区别, 并最终导致产品质量下降. 因此, 间歇过程的监控和故障诊断一直是过程控制领域研究的热点之一. 实施间歇过程监控的目的在于及时发现生产过程的非正常情况, 提高生产安全性, 降低生产成本, 及时发现并排除生产故障, 提高产品质量的一致性. 目前间歇过程的监控比较常用的方法[1~4]有多向主元分析(MPCA)和多向部分最小二乘(MPLS)等.

统计过程监控基于历史正常生产数据, 使用数理统计的理论建立正常工况下操作条件的统计监控模型, 并用于实际生产的在线监控. 作为数据驱动的监控方法, 统计过程监控不需要任何过程的机理模型, 因此适用十分广泛. 多向部分最小二乘方法(MPLS)是一种目前比较成熟的用于间歇生产统计过程监控的方法, 实际的工业应用结果也证明了这种方法的有效性. 但传统的MPLS方法当建模历史数据存在离群点(Outliers)或建模数据中存在非正常批次时, 可能会给出误差较大的监控模型, 出现故障误报或漏报现象. 同时由于间歇过程的复杂性, 建模前很难保证将所有离群点全部剔除. 对此, 本文提出一种基于鲁棒PLS监控的模型建立方法, 用于处

收稿日期: 2004-07-14; **修回日期**: 2004-09-06.
基金项目: 国家 863 计划项目(2001AA413110).
作者简介: 董胜利(1981—), 男, 河南漯河人, 硕士生, 从事生化过程建模与过程监控等研究; 王树青(1939—), 男, 浙江仙居人, 教授, 博士生导师, 从事工业生产过程模型化与优化控制、先进控制等研究.

理建模数据中存在离群点的问题

2 传统部分最小二乘法[5~7]

部分最小二乘法(PLS)由Wold提出[8], 它是多元线性回归算法的扩展, 对主元成分进行线性回归, 建立主元成分之间的线性回归模型. PLS有两大主流算法: NIPALS和SMPLS. 与NIPALS相比, SMPLS算法应用更为普遍. SMPLS采用x-变量和y-变量之间的互协方差进行分析得到主元, 然后主元通过线性最小二乘回归方法回归因变量.

SMPLS算法假设x变量和y变量通过一个双线性的模型发生关联, 这就意味着两步算法: 首先构建k维向量$\tilde{T}_{n,k} = (\tilde{t}_1, \cdots, \tilde{t}_n)' = \tilde{X}_{n,p}R_{p,k} = \tilde{X}_{x,p}(r_1, \cdots, r_k)$, 记为主元; 然后通过迭代运算获得$k$个主元成分, 并用这$k$个主元来回归因变量. 回归模型为

$$y_i = \alpha_0 + A'_{q,k}\tilde{t}_i + f_i. \tag{1}$$

引入标准PLS权向量r_a和q_a, 多重线性回归(MLR)对参数估计如下:

$$\begin{aligned}&\hat{A}_{k,q} = (S_t)^{-1}S_{ty} = (R'_{k,p}S_xR_{p,k})^{-1}R'_{k,p}S_{xy},\\ &\hat{\alpha}_0 = \bar{y} - \hat{A}'_{q,k}\bar{\tilde{t}},\\ &S_f = S_y - \hat{A}'_{q,k}S_x\hat{A}_{k,q}.\end{aligned} \tag{2}$$

多向部分最小二乘方法[9]是部分最小二乘法应用于三维数据阵的扩展. 对于间歇生产批报而言, 每批数据都可看作一个二维数据阵, 多批数据则构成了三维数据阵$\underline{X}$. 针对三维数据, 一个自然想法就是将其进行重新排列, 如图1所示, 沿着时间轴方向进行切分, 然后将切分得到的数据时间片依次向右水平排列, 如此构成了一个新的二维数据阵X, 然后利用PLS进行分析.

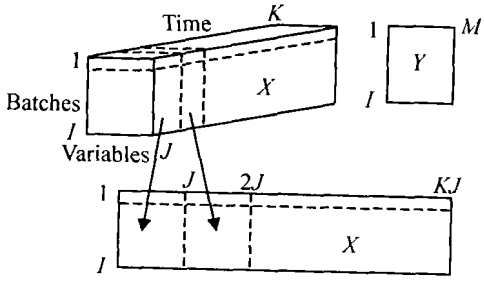

图1 MPLS分解示意

可以看出, 同MPCA一样,MPLS将每一批完整的数据看作间歇处理过程的一次采样, 多批数据构成样本集合, 并在此样本集合上进行PLS分析. 以上特点决定了MPLS在应用于实际监控时会出现采样数据不完整的问题. 因为在间歇过程进行中, 只有当前时刻及以前的数据是已知的, 这些数据不足以构成对间歇过程的一次完整采样. Nomikos[3]提出了解决该问题的几种方法, 基本思想是设法预测过程变量的未来输出. 常用的方法包括: 1) 补充数据为全0, 即认为以后的数据不偏离平均轨迹; 2) 补充数据为当前归一化采样值, 即认为以后的数据偏离平均轨迹的程度与当前时刻相同; 3) 缺损数据补充法. 在实际应用中常采用第3种方法, 这种方法利用了PLS处理缺损数据的能力, 在不知道将来时刻数据的情况下通过将已知数据投影到一个子空间来完成.

3 鲁棒算法RSMPLS及其统计量[6, 10]

传统PLS分析的一个主要不足是鲁棒性不够好, 当采样数据中存在离群点时, 可能无法得到正确的统计监控模型. 鲁棒部分最小二乘分析方法正是克服传统PLS的不足而提出的.

鲁棒PLS算法是对普通算法的改进. 为了获得鲁棒主元, 首先对数据$Z_{n,m} = (X_{n,p}, Y_{n,q})$进行鲁棒主元分析(ROBPCA), 得到$Z$的估计$\hat{\mu}_Z = (\hat{\mu}'_x, \hat{\mu}'_y)'$和$\hat{\Sigma}_Z$. 分解$\hat{\Sigma}_Z$得到$\hat{\Sigma}_{xy}$, 代替$S_{xy}$进行迭代计算, 在每一步中鲁棒主元可表示为

$$t_{ia} = \tilde{x}'_i r_a = (x_i - \hat{\mu}_x)'r_a. \tag{3}$$

获得鲁棒主元后, 便可进行鲁棒线性回归. 回归模型与式(1)一样, 不同的是现在是基于鲁棒主元t_i, 参数估计如式(2).

通过交叉检验、累计方差百分比等方法确定主元数目k后, 对数据进行鲁棒PLS分析, 可得到以下统计量:

$$\{X \quad Y\} \xrightarrow{\text{鲁棒PLS}} [T, W, P, B, Q]. \tag{4}$$

1) Hotelling统计量[11]:

$$T_f^2 = t^{\mathrm{T}}\Lambda^{-1}t \sim \frac{k(m^2-1)}{m(m-k)}F_{k,m-k}, \tag{5}$$

其中: m为样本容量, Λ^{-1}为对角矩阵, 与主元方向相关.

2) Q统计量, 有如下结论[11]:

$$q = (x - \tilde{x})^{\mathrm{T}}(x - \tilde{x}) = \sum e^{\mathrm{T}}e \sim g\chi^2_h, \tag{6}$$

其中: $\tilde{x} = xW(P^{\mathrm{T}}W)^{-1}P^{\mathrm{T}}$为通过PLS模型获得的$x$的估计值, $W = [r_1, \cdots, r_k]$,

$$gh = \mathrm{mean}(q),\ 2g^2h = \mathrm{var}(q).$$

4 基于鲁棒MPLS的间歇生产统计过程监控

在鲁棒部分最小二乘分析方法的基础上, 提出一种基于鲁棒MPLS的间歇生产统计过程监控方法. 与传统MPLS相同的是, 该方法也是用Hotelling统计量和Q统计量对生产过程进行监控; 不同的是, 它应用了鲁棒PLS方法, 当建模历史数据存在离群点或建模数据中存在非正常批次时, 仍能给出正确的监控模型.

基于鲁棒MPLS的间歇生产统计过程监控包括3个阶段: 1) 数据收集与预处理; 2) 建立统计监控模型; 3) 在线监控

4.1 数据收集与预处理

1) 收集历史正常工况数据, 构成历史数据阵 $\underline{X}$, 按照图 1 展开.

2) 数据标准化, 包括零均值化以及方差归一化

$$\tilde{X} = \frac{X - \mathrm{mean}(X)}{\mathrm{std}(X)},$$

$$\tilde{Y} = \frac{Y - \mathrm{mean}(Y)}{\mathrm{std}(Y)}, \tag{7}$$

其中: $\mathrm{mean}(X)$ 和 $\mathrm{mean}(Y)$ 为均值, $\mathrm{std}(X)$ 和 $\mathrm{std}(Y)$ 为方差

4.2 建立统计监控模型

1) 确定主元数目 k. 确定主元数目的方法有多种, 包括交叉检验法、累计方差百分比等. 考虑到抗差性, 这里选用累计方差百分比作为选取 k 的标准. 前 l 个主元的累计方差百分比定义如下:

$$\mathrm{CPV}(l) = \frac{\sum_{i=1}^{l} s_i}{\sum_{i=1}^{\min(n,m)} s_i} \times 100\% = \frac{\sum_{i=1}^{l} s_i}{n} \times 100\%, \tag{8}$$

其中: s_i 为对角阵 Λ^{-1} 的元素; 主元数目 k 一般选取满足 $\mathrm{CPV}(k)$ 大于特定限, 如 85% 的最小 l 值

2) 对标准化后的数据进行鲁棒MPLS处理, 以获得相应的得分向量和负载向量等, 如式(4).

3) 根据式(5) 和式(6) 确定相关统计量的控制限, 保存相关结果以备在线监控时使用.

4.3 在线监控

在新批次运行过程中, 通过Nomikos提出的数据填充方法补充缺损的数据, 进而计算得到MPLS主元的得分 t 以及Q 统计量, 结合离线建模阶段得到的统计控制限实施监控并改进生产.

具体的在线监控方法如下:

1) 对于 k 时刻新的间歇过程数据 $\underline{X}_{\mathrm{new}}(K \times J)$ 按图 1 展开为 $X_{\mathrm{new}}^{\mathrm{T}}(1 \times KJ)$, Y_{new} 不变, 然后分别进行标准化处理

2) 首先用Nomikos提出的 3 种数据补充方法进行处理, 通常采用第 3 种方法. 然后计算 t_{new} 和 e_{new}. 在每一个采样时刻 k, $r = 1:R$,

$$\hat{t}(1, r) = X_{\mathrm{new}} W(1:kJ, r) / (P(1:kJ, r)^{\mathrm{T}} W(1:kJ, r)),$$

$$e_{\mathrm{new},k} = X_{\mathrm{new},k} - \hat{t}(1, r) P(1:kJ, r)^{\mathrm{T}}. \tag{9}$$

3) 通过下式计算 Hotelling 统计量和Q 统计量:

$$T^2 = t_{\mathrm{new}}^{\mathrm{T}} \Lambda^{-1} t_{\mathrm{new}}, \tag{10}$$

$$\mathrm{SPE}_K = \sum_{c=(K-1)J+1}^{JK} e_{\mathrm{new}}(c)^2. \tag{11}$$

4) 判断 T^2 和 SPE_k 统计量是否超过控制限, 进行故障诊断

5 应用示例

SBR 数据(Nomikos, MacGregor) 是对苯乙烯-丁二烯橡胶乳化聚合半间歇过程的模拟, 它假设 50 个批次, 每个批次每次采样可得到 9 个变量, 共有 200 个采样点, 因此 $\underline{X}$ 数据是三维矩阵(50 × 9 × 200). 产品质量、反应器属性、乳汁属性等质量变量 Y 为(50 × 5) 矩阵. 其中不正常的批次(Bad batch) 是在间歇过程操作中间时刻出错.

为比较传统MPLS和鲁棒MPLS监控的效果, 本文用 SBR 数据分别建立了传统MPLS 和鲁棒MPLS 模型.

图 2～ 图 5 为采用不同建模方法对异常批次的 Q 统计量监控结果, 虚线对应的均为 95% 统计控制限.

图 2 和图 3 显示的是全部采用 50 个正常批次建立的MPLS 和鲁棒MPLS 模型的Q 统计量监控结果, 两者都可以检测出异常工况. 图 4 和图 5 显示的是采用含有离群点批次建立的MPLS 模型的Q 统计量监控结果.

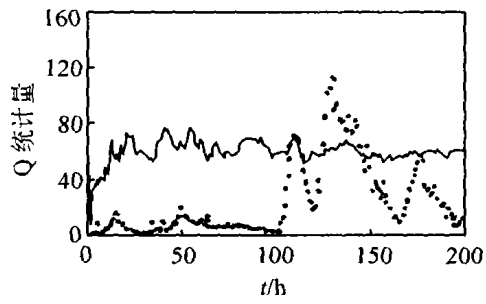

图 2　MPLS 监控结果
(建模数据无离群点)

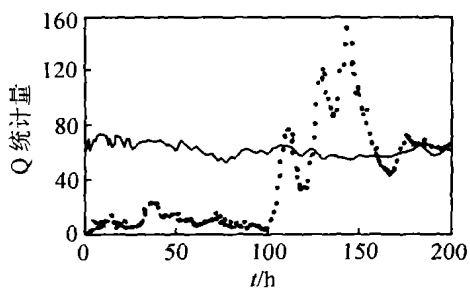

图 3　鲁棒MPLS 监控结果
(建模数据无离群点)

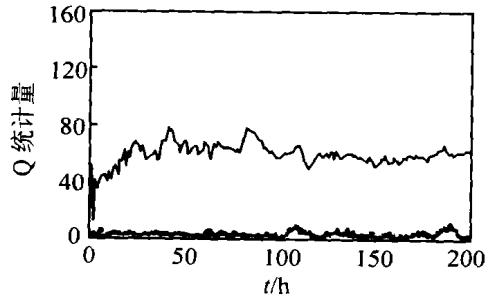

图 4 MPLS 监控结果
(建模数据有离群点)

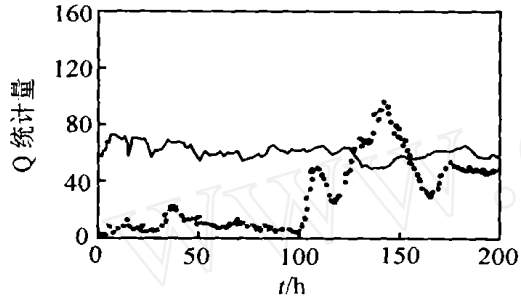

图 5 鲁棒MPLS 监控结果
(建模数据有离群点)

通过图 2 和图 4 对比可以看出, 离群点的存在导致MPLS 模型给出了错误的监控结果,Q 统计量不能及时发现过程的异常状况, 导致了漏报 原因在于离群点的存在使得MPLS 分析得到的正常工况区间发生偏离, 使其无法正确地进行过程监控

而鲁棒MPLS 因其具有消除离群点影响的能力, 弥补了MPLS 的不足 从图 3 和图 5 的对比可以看出, 建模数据中是否存在离群点对鲁棒MPLS 监控结果的影响并不大, 故障批次中的异常情况都能被发现, 为及时排除故障提供了可能

6 结 论

本文提出了一种基于鲁棒MPLS 的间歇生产过程统计监控方法 该方法克服了传统MPLS 鲁棒性不足, 即个别离群点便可能导致建立的统计监控模型失效的缺点 仿真示例 SBR 数据的应用表明, 鲁棒MPLS 能够有效克服离群点的影响, 保证统计模型不过分依赖建模数据, 有效减轻了模型对数据的要求

参考文献(References)

[1] Wentzell, Peter D, Vega Montoto, et al. Comparison of Principal Components Regression and Partial Least Squares Regression Through Generic Simulations of Complex Mixtures[J]. *Chemometrics Intell Lab*, 2003, 65(2): 257-279.

[2] Nomikos P, MacGregor J F. Monitoring of Batch Processes Using Multi-way Principal Components Analysis[J]. *American Institute of Chemical Engineers*, 1994, 40(8): 1361-1375.

[3] Nomikos P, MacGregor J F. Multivariate SPC Charts for Batch Processes[J]. *Technometrics*, 1995, 37(1): 41-59.

[4] Liang J, Qian X J. Multivariate Statistical Process Monitoring and Control: Recent Developments and Applications to Chemical Industry [J]. *Chinese J of Chemical Eng*, 2003, 11(2): 191-203.

[5] De Jong S. SIMPLS: An Alternative Approach to Partial Least Squares Regression[J]. *Chemometrics and Intelligent Laboratory Systems*, 1993, 55(2): 251-263.

[6] Hubert M, Vanden Branden K. Robust Methods for Partial Least Squares Regression [J]. *J of Chemometrics*, 2003, 17(10): 537-549.

[7] 张杰, 阳宪惠. *多变量统计过程控制*[M]. 北京: 化学工业出版社, 2000: 24-60.
(Zhang J, Yang X H. *Multi-variable Stat Process Control* [M]. Beijing: Chemistry Industry Publishing Company, 2000: 24-60.)

[8] Wold H. Estimation of Principal Components and Related Models By Iterative Least Squares [A]. *Multivariate Analysis*[C]. New York: Academic Press, 1966: 391-420.

[9] Nomikos P, MacGregor J F. Multi-way Partial Least Square in Monitoring Batch Processes [J]. *Chemometrics and Intelligent Laboratory Systems*, 1995, 30(1): 97-108.

[10] Hubert M, Rousseeuw P J, Verboven S. A Fast Method for Robust Principal Components with Applications to Chemometrics[J]. *Chemometrics and Intelligent Laboratory Systems*, 2002, 60(1/2): 101-111.

[11] MacGregor J F, Jaeckle C K, Costas K M. Process Monitoring and Diagnosis by Multiblock PLS Methods [J]. *American Institute of Chemical Engineers*, 1994, 40(5): 826-838.