文章编号: 1001-0920(2006)07-0745-05

基于改进混合概率主元分析模型的过程监控

赵忠盖, 刘 飞, 徐保国
(江南大学 自动化研究所, 江苏 无锡 214122)

摘 要: 基于混合概率主元分析(MPPCA)的监控方法, 存在要求各子模型中主元个数相同、监控指标不一致、监控表格过多等缺陷. 为此对MPPCA 算法进行改进, 分两步建立模型: 首先求出混合高斯模型(GMM), 然后利用概率主元分析(PPCA)建立每个子模型的主元模型. 改进方法中各子模型主元的选取兼顾了主元的解释率及其变化趋势, 并引进基于 PPCA 的监控方法, 保证了监控指标的一致性, 减少了过程监控图.

关键词: 混合概率主元分析模型; 过程监控; EM 算法
中图分类号: TP206.3 **文献标识码**: A

Process Monitoring Based on Improved Mixture Probabilistic Principal Component Analysis Model

ZHAO Zhong-gai, LIU Fei, XU Bao-guo
(Institute of Automation, Southern Yangtze University, Wuxi 214122, China. Correspondent: LIU Fei, E-mail: dr-fliu@hotmail.com)

Abstract: Mixture probabilistic principal component analysis (MPPCA) model is dealt with. To the disadvantages in MPPCA-based methods which the MPPCA model needs all the sub-models have the same number of principal components, its two monitoring indices often can not consist with each other, and the monitoring charts are too many, an improved MPPCA model is presented in two steps. Firstly, a mixture Gaussian model is built. Then, by using probabilistic principle component analysis (PPCA), the principal component model of each sub-model is developed. Compared with MPPCA, the improved MPPCA selects the principal components based on not only the explanation of process variables in different sub-models but also its trend of changing. By introducing PPCA, the consistence of monitoring indices is guaranteed and the number of charts used by monitor is reduced.

Key words: Mixture probabilistic principal component analysis model; Process monitoring; EM algorithm

1 引 言

基于主元分析(PCA)或概率主元分析(PPCA)的统计过程监控理论认为, 系统正常运行时, 过程变量满足正态分布; 当过程出现异常时, 过程变量不符合正态分布[1,2]. 在实际的生产中, 工业过程存在动态和非线性, 并且许多工业系统往往具有多个稳定的工作状态或操作状态, 因此系统正常运行时, 过程变量一般不满足正态分布. 如果使用传统的统计监控方法, 则会出现严重的误报或漏报.

为克服这一问题, 混合 PCA 模型被引入过程监控[3]. 该方法首先将采样数据聚类, 每个类中的采样数据满足高斯分布, 然后对每个类内的数据使用 PCA 进行监控. 这种方法硬性地将每个采样数据划分到某个模型中, 属于"硬性"聚类方法, 没有考虑采样数据的总体分布情况. 混合概率主元分析(MPPCA)模型采用"软性"聚类方法[4]. 它考虑了数据划分到各个聚类的概率, 在每个聚类内使用 PPCA 模型, 通过期望最大化(EM)算法估计出各子模型的参数. 基于MPPCA 模型的监控则分别计算过程变量在每个子模型中的统计指标 SPE 和 T^2

收稿日期: 2005-05-18; **修回日期**: 2005-08-25.
基金项目: 国家十五攻关计划课题(2004BA204B08); 教育部科学技术研究重点项目(1105088).
作者简介: 赵忠盖(1976—), 男, 湖北荆州人, 博士生, 从事工业系统监控与诊断等研究; 刘飞(1965—), 男, 安徽宣城人, 教授, 博士生导师, 从事先进控制理论与应用、工业系统监控与诊断等研究.

值, 如果至少有一个子模型的 SPE 和 T^2 值在控制限内, 则过程处于正常运行状态; 否则, 过程可能出现异常.

MPPCA 模型应用于过程监控, 还存在诸多缺陷: 1) 它采用 EM 算法一次性估计出 PPCA 模型中的负荷向量和噪声方差值, 故要求各模型中的主元个数相同[5], 而对于实际数据, 因为主元对各子模型中过程变量的解释率不同, 所以各子模型的主元个数可能不一样; 2) 在一般MPPCA 模型中, 监控指标 SPE 和 T^2 使用的是不同的量度, 因此在监控中会存在不一致的情况[6]; 3) 当MPPCA 子模型很多时, 该方法表格过多, 导致监控任务繁重.

本文对MPPCA 方法进行改进, 分两步建立过程监控模型: 首先建立高斯混合模型(GMM); 然后根据 GMM 中每个高斯模型的协方差矩阵, 采用 PPCA 算法估计出每个模型的主元个数, 由此建立MPPCA 模型. 在每个子模型内部, 使用基于 PPCA 的监控指标. 本文将MPPCA 模型引入过程监控, 结合实际化工分离过程进行研究.

2 改进的MPPCA 模型

2.1 一般MPPCA 模型[4]

假设满足任意概率密度函数 $p(\mathrm{X}|\theta)$ 的采样数据为 $\mathrm{X}=\{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_N\} \in R^{d\times N}$. 其中 d 为变量个数, N 为采样数, θ 为概率密度函数的参数. 将 $p(\mathrm{X}|\theta)$ 表示为 K 个高斯密度函数的加权和, 即将 X 表示为含 K 个高斯子模型的 GMM, 而在每个高斯子模型中, 采样数据满足生成模型 $\mathrm{X}=\mathrm{WT}+\mu+\epsilon$. 其中 W 为负荷向量矩阵, μ 为均值, ϵ 为噪声, 且 $\epsilon \sim N(0, \Psi)$, T 为主元矩阵, 且 $\mathrm{T} \sim N(0, \mathrm{I})$. 每个高斯子模型可表示成一个 PPCA 模型, 有

$$p(\mathrm{X}|\theta)=\sum_{i=1}^{K}\beta_i p(\mathrm{X}|i). \tag{1}$$

其中: β_i 为数据由混合模型第 i 个聚类产生的概率, 满足 $\sum_{i=1}^{K}\beta_i=1, 0\le\beta_i\le 1$; $p(\mathrm{X}|i)$ 为用第 i 个PPCA 模型表示的概率密度函数, 假设其参数为 θ_i, 包括均值向量、噪声方差、负荷向量矩阵、协方差矩阵等. 则 MPPCA 模型参数 $\theta=\{\beta_1, \beta_2, \cdots, \beta_k, \theta_1, \theta_2, \cdots, \theta_k\}$. X 的 log 概率为

$$\log p(\mathrm{X}|\theta)=\log\prod_{n=1}^{N}p(\mathrm{x}_n|\theta).$$

采用极大似然算法, 则 θ 的最优估计值为 $\theta=\arg\max_{\theta}\{\log p(\mathrm{X}|\theta)\}$. 根据Bayes 定理, 第 i 个模型产生 X 的后验分布为

$$p(i|\mathrm{X})=p(\mathrm{x}|i)\beta_i\Big/\sum_{j=1}^{K}p(\mathrm{x}|j)\beta_j,$$

第 n 个采样时刻的后验分布为 $R_{ni}=p(i|\mathrm{x}_n, \theta)$. 由期望最大化(EM) 算法可得

$$\tilde{\beta}_i=\frac{1}{N}\sum_{n=1}^{N}R_{ni}, \tag{2}$$

$$\tilde{\mu}_i=\sum_{n=1}^{N}R_{ni}(\mathrm{x}_{ni}-\tilde{\mathrm{W}}_i\langle\mathrm{T}_{ni}\rangle)\Big/\sum_{n=1}^{N}R_{ni}, \tag{3}$$

$$\tilde{S}_i=\frac{1}{\tilde{\beta}_i N}\sum_{n=1}^{N}R_{ni}(\mathrm{x}_n-\tilde{\mu}_i)(\mathrm{x}_n-\tilde{\mu}_i)^{\mathrm{T}}, \tag{4}$$

$$\tilde{\mathrm{W}}_i=\tilde{S}_i\mathrm{W}_i(\sigma_i^2\mathrm{I}+\mathrm{M}_i^{-1}\mathrm{W}_i^{\mathrm{T}}\tilde{S}_i\mathrm{W}_i)^{-1}, \tag{5}$$

$$\tilde{\sigma}_i^2=\frac{1}{d}\mathrm{tr}(\tilde{S}_i-\tilde{S}_i\mathrm{W}_i\mathrm{M}_i^{-1}\tilde{\mathrm{W}}_i^{\mathrm{T}}). \tag{6}$$

其中: $\mathrm{M}_i^{-1}=(\sigma_i^2\mathrm{I}+\mathrm{W}_i^{\mathrm{T}}\mathrm{W}_i)^{-1}$, $\langle\mathrm{T}_{ni}\rangle=\mathrm{M}_i^{-1}\mathrm{W}_i^{\mathrm{T}}(\mathrm{x}_n-\mu_i)$, μ_i, σ_i^2, W_i, S_i 分别表示第 i 个 PPCA 模型的均值向量、噪声方差值、负荷向量矩阵和协方差矩阵; 向量或矩阵上加符号~ 表示新的迭代值.

式(1)~ (6) 构成了 EM 算法. 反复迭代直到收敛, 即可得到每个 PPCA 模型的 β_i, μ_i, W_i, σ_i^2, S_i.

2.2 改进的MPPCA 模型

由式(1)~ (6) 可知, W_i ($i=1, 2, \cdots, K$) 维数相同, 即MPPCA 模型中每个子模型主元个数一样. 改进的MPPCA 模型将建模过程分为两部分: 第一部分利用 GMM 表示采样数据的概率函数, 并通过 EM 算法估计出参数; 第二部分将 GMM 中每个子模型的协方差矩阵, 由 PPCA 算法分解成负荷向量的协方差矩阵和噪声矩阵, 而每个子模型的主元个数则由主元对每个过程变量的解释率来决定.

将概率密度函数 $p(\mathrm{X}|\theta)$ 用 GMM 表示, 则第 i 个模型产生 X 的后验分布为

$$p(i|\mathrm{X})=G_x[\mu_i, \Sigma_i]\beta_i\Big/\sum_{j=1}^{K}G_x[\mu_j, \Sigma_j]\beta_j, \tag{7}$$

其中 $G_x[\mu_i, \Sigma_i]$ 表示均值为 μ_i, 方差矩阵为 Σ_i 的 X 的高斯密度函数. 由 EM 算法可得

$$\mu_i^{\mathrm{new}}=\sum_{n=1}^{N}R_{ni}\mathrm{x}_n\Big/\sum_{n=1}^{N}R_{ni}, \tag{8}$$

$$\Sigma_i^{\mathrm{new}}=\frac{\sum_{n=1}^{N}R_{ni}\mathrm{x}_n\mathrm{x}_n^{\mathrm{T}}}{\sum_{n=1}^{N}R_{ni}}-(\mu_i^{\mathrm{new}})(\mu_i^{\mathrm{new}})^{\mathrm{T}}, \tag{9}$$

$$\beta_i^{\mathrm{new}}=\frac{1}{N}\sum_{n=1}^{N}R_{ni}. \tag{10}$$

其中 μ_i^{new}, Σ_i^{new} 和 β_i^{new} 分别是均值向量、协方差矩阵和第 i 个高斯模型产生采样数的先验概率的新迭代值. 在得到上面各参数后, 对每个模型利用PPCA 算法估计出主元的个数和噪声的方差值.

对于每个子模型, 由生成模型有 $\Sigma_i=\mathrm{W}\mathrm{W}_i^{\mathrm{T}}+\sigma_i^2\mathrm{I}$. 根据 PPCA 算法, 有

$$\tilde{W}_i = \Sigma_i W_i (\sigma_i^2 I + M_i^{-1} W_i^T \Sigma_i W_i)^{-1}, \quad (11)$$

$$\tilde{\sigma}_i^2 = \frac{1}{M} \mathrm{tr}(\Sigma_i - \Sigma_i W_i M_i^{-1} \tilde{W}_i^T). \quad (12)$$

反复迭代式(11) 和(12), 直到收敛得到第 i 个子模型的负荷向量 W_i 和误差方差值 σ_i^2. 在改进的MPPCA 模型中, 噪声方差和负荷向量都是根据先前估计的协方差矩阵, 由PPCA 算法得到的, 因此可根据需要选择不同的主元个数 σ_i^2 反映了过程原因外的干扰 在过程正常运行的情况下, 主元个数越多, σ_i^2 越小; 当主元个数达到一定值后, 此定值即为过程潜在的主元个数; 主元个数如果继续增加, 则 σ_i^2 变化不大(收敛), 即主元对过程变量的解释率随主元个数的继续增加而基本不变 在模型 i 中, 主元对过程变量 l 的解释率为

$$r_{il} = \mathrm{diag}(W_i W_i^T)_l / [\mathrm{diag}(W_i W_i^T)_l + \sigma_i^2].$$

其中: $\mathrm{diag}(W_i W_i^T)_l$ 表示 $W_i W_i^T$ 对角线上第 l 个元素, r_{il} 是关于主元个数的函数 计算出 $r_{il}(q_l + 1) - r_{il}(q_l) < \epsilon$ 后(ϵ为阈值), 则模型 i 包含的潜在主元个数为 $g_i = \max(q_1, q_2, \cdots, q_M)$. 同理 可计算出所有模型的主元个数[6]. 根据主元个数, 重新选定或计算各子模型的 W_i 和 σ_i^2 值, 建立含不同主元个数子模型的MPPCA 模型

2.3　算法实现

在上述算法中, 高斯模型个数选取的多少直接影响到过程监控的效果 模型选取过多, 则参数太多, 模型不易收敛, 且误报增加; 模型选取过少, 则模型收敛快, 但漏报增加 利用确定性方法确定模型的个数[4,7]. 设模型个数 $\hat{v}$ 在 $v_{\min}$ 与 $v_{\max}$ 之间, 则

$$\hat{v} = \arg\min_v \{C(\hat{\theta}(v), v), v = v_{\min}, \cdots, v_{\max}\}.$$

其中: $\hat{\theta}(v)$ 为高斯模型为 v 时估计出的模型参数, $C(\hat{\theta}(v), v)$ 为模型选择标准函数, 表示为

$$C(\hat{\theta}(v), v) = -\log p(x|\hat{\theta}(v)) + P(v).$$

式中 $P(v)$ 是关于 v 的惩罚函数, v 越大, $P(v)$ 值也越大 在计算 v 由 $v_{\max}$ 变化到 $v_{\min}$ 的 $C(\hat{\theta}(v), v)$ 值时, 采用检测指标

$$J(i,j) = [R_{1i}, R_{2i}, \cdots, R_{Ni}][R_{1j}, R_{2j}, \cdots, R_{Nj}]^T.$$

$J(i,j)$ 越大, 表示第 i 个模型和第 j 个模型越相近 选择使 $J(i,j)$ 最大的两个模型, 将其合并, 令 $v = v - 1$, 重新计算 $C(\hat{\theta}(v), v)$ 值 最终使 $C(\hat{\theta}(v), v)$ 值最小的 v 即为最佳的模型个数

当过程变量很多时, 数据分散且参数多, 使得EM 算法收敛很慢, 甚至不收敛, 因此很难得到合适的GMM. 而 PCA 能在保证数据信息量损失最小的情况下, 对数据进行压缩 为减少需要估计的模型参数, 采用的方法是先由 PCA 将原过程变量数据压缩, 即将数据投影到模型子空间和残差空间, 再对模型子空间的数据实现 EM 算法, 从而得到 GMM[8].

另外, 各个高斯模型的初始均值和协方差矩阵的选取, 对算法的速度和收敛性都有很大的影响 根据文献[8], 本文先用 K-聚类的方法得到每个模型的初始均值, 再通过每个聚类的数据估计出模型的初始协方差矩阵

3　基于改进MPPCA 模型的监控方法

在离线建立改进的MPPCA 模型后, 将当前采样数据代入MPPCA 模型的各子模型, 如果监控指标值处于某个子模型的控制限以内, 则过程运行正常, 否则过程可能出现异常 现有文献中基于MPPCA 的监控方法, 在每个子模型内部使用基于PCA 的过程监控, 即 SPE 和 T^2 统计量 这两个监控指标反映的不是同一个量度, SPE 为 Euclidian 范数, T^2 为Mahalanobis 范数 因此在实际应用中可能出现不一致的情况[6]. 另外, 一个子模型需要两张监控表, 当子模型个数过多时监控表格过多, 从而造成监控程序复杂化 本文在子模型中使用基于 PPCA 的监控方法 类似于 PCA, 基于 PPCA 的监控方法监控主元空间和误差空间, 采用如下指标:

$$\| M_i W_i^T x_n \|^2 \le \chi^2_{(1-\alpha, q_i)}, \quad (13)$$

$$\| \sigma_i^{-1} (I - W_i M_i W_i^T) x_n \|^2 \le \chi^2_{(1-\alpha, d)}. \quad (14)$$

其中 q_i 为第 i 个子模型中的主元个数 式(13) 监控的是主元, 不等号前面的值相当于 PCA 中的 T^2 检验值; 式(14) 检测的是噪声变量, 相当于 PCA 中的 SPE 值, 不等式后面的值即为控制限

PPCA 的主元检测和噪声变量监控使用的是同一种量度(都是Mahalanobis 范数), 所以对采样数据白化值范数的监控反映了主元和噪声变量监控的综合结果 对采样数据白化值范数进行监控的表达式为

$$\| (W_i W_i^T + \sigma_i^2 I)^{-0.5} x_n \|^2 \le \chi^2_{(1-\alpha, q)}. \quad (15)$$

只需对式(15) 进行监控即可, 从而大大减少了监控的工作量

需要指出的是, 若进一步利用过程出现每种故障时的数据建立故障的MPPCA 模型, 当监控算法检测到过程出现故障时, 通过判断故障数据符合何种故障模型, 便可使基于MPPCA 模型的过程监控技术扩展到故障诊断

4　应用实例

某芳烃联合装置中的分离过程共有 15 个可检测过程变量, 过程的详细描述参见文献[9]. 取正常运行情况下的 400 组采样数据进行建模, 然后选出另有故障的 200 组作为试验数据进行监控

首先选择MPPCA 模型中 PPCA 模型的个数

对过程数据进行 PCA 压缩, 为使数据信息量的损失最小, 将主元对过程数据的解释率定为 99%. 这样由 8 个线性主元变量即可代替 15 个过程变量, 原始数据的信息基本不损失. 选取模型最大个数和最小个数分别为 $v_{max}=10$, $v_{min}=1$, 计算出模型个数从 v_{max} 变化到 v_{min} 的 $C(\hat{\theta}(v),v)$ 值(见图 1). 由图 1 可知, 当模型个数选为6时, $C(\hat{\theta}(v),v)$ 值最小, 因此模型个数选为 6.

然后建立MPPCA 模型. 用 K-聚类法将数据聚类, 估计出初始均值和协方差矩阵, 通过 EM 算法估计出GMM 中6个高斯模型的参数. 在6个高斯模型中, 利用PPCA 计算出主元个数1～8时主元对过程变量的解释率(见图 2). 由图 2 可知, 对不同的子模型取相同的主元个数, 主元对数据的解释率明显不同. 在传统的基于MPPCA 模型的监控中, 各子模型中选择的主元个数一样, 这样会造成: 1) 主元个数过多, 主元模型过拟合, 从而包含噪声, 影响了监控效果; 2) 主元个数过少, 主元模型不能完全包含过程变量的信息, 引起信息丢失. 利用改进的MPPCA 模型方法, 主元数的选取兼顾两个方面: 一是保证主元对每个变量的解释率都大于 60%; 二是看解释率随主元个数的增大是否趋于稳定. 据此在 6 个模型中选取的主元个数见表 1. 根据每个模型内部选取的主元个数, 通过 EM 算法估计出负荷向量和噪声方差值, 从而建立MPPCA 模型.

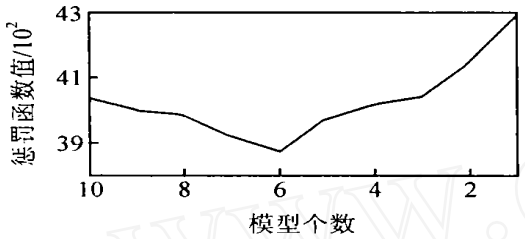
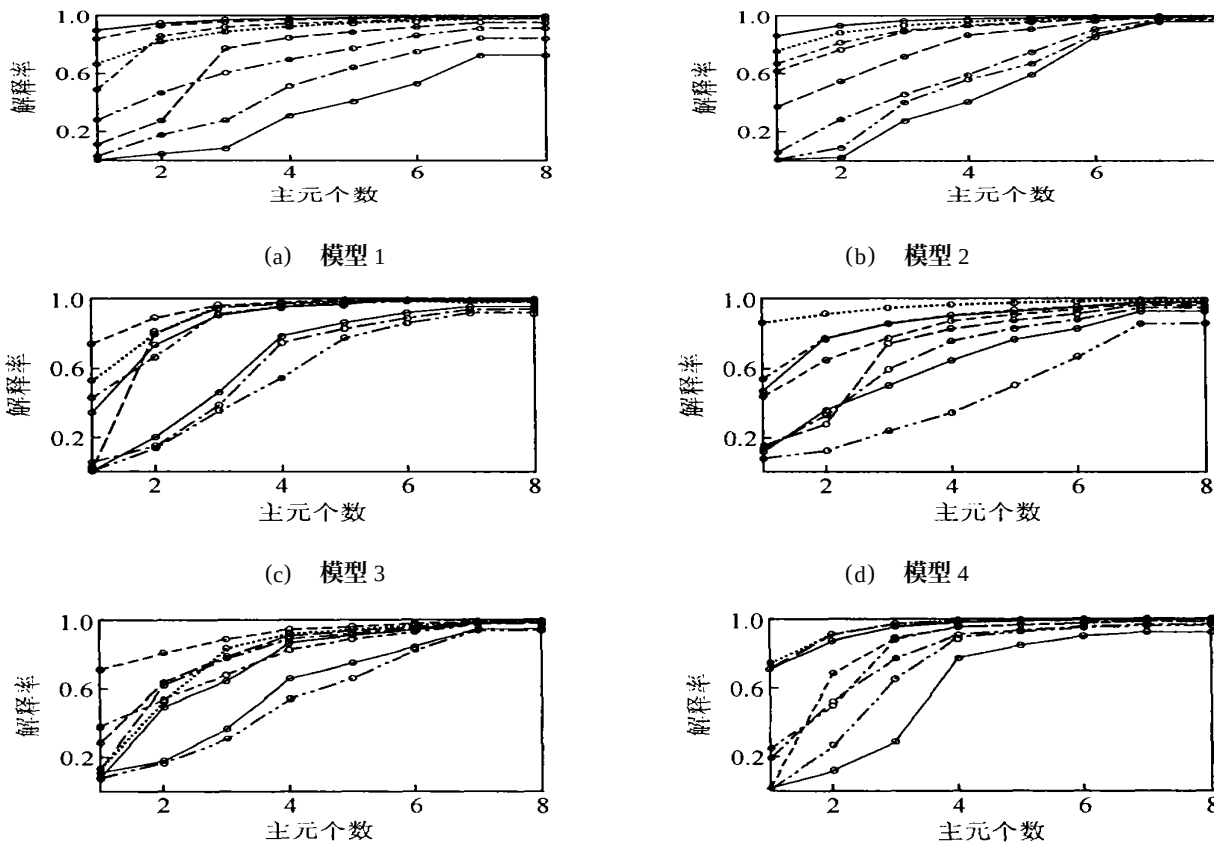


图 1 模型选择标准 $C(\hat{\theta}(v),v)$ 值

表 1 各子模型选取的主元个数

模型序号	1	2	3	4	5	6
主元个数	7	6	5	7	6	4

最后对过程实施监控. 将当前采样数据投影后的 8 个线性主元值作为过程变量代入MPPCA 模型, 计算其在每个模型中的白化值. 如果至少有一个模型中白化值的范数满足自由度为 8 的正态分布, 即白化值在控制限以内, 则过程运行正常; 否则, 表明过程出现异常. 基于MPPCA 模型的监控如图 3 所示.



(a) 模型 1 (b) 模型 2

(c) 模型 3 (d) 模型 4

(e) 模型 5 (f) 模型 6

图 2 主元对过程变量的解释率

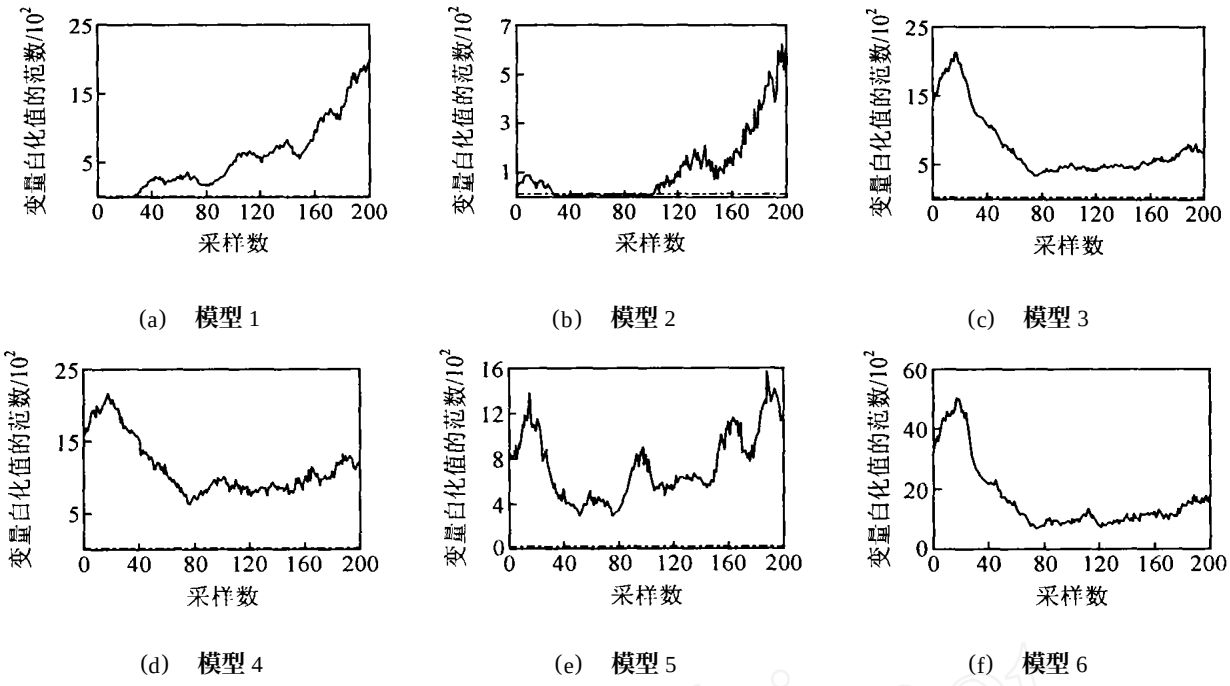


图 3　基于M PPCA 模型的监控

从图 3 可以看出, 1～ 25 间采样数据的白化值范数在第 1 个 PPCA 模型中没有超出控制限, 而 26～ 100 间采样数据的白化值范数在第 2 个 PPCA 模型中也处于控制限内, 因此可以判断过程在 1～ 100 时刻运行正常. 101～ 200 间采样数据的白化值在 6 个模型中都超出了控制限, 因此可以判断出过程在 101～ 200 时刻可能运行异常. 如果使用传统的 M PPCA 模型, 在每个子模型内部使用基于 PCA 的监控方法, 则每个子模型中需要 2 幅监控图(即 SPE 和 T^2 图), 总的监控图达到 12 幅, 增加了监控的工作量, 且监控指标不一致.

5　结　　论

本文对 M PPCA 模型算法进行改进, 将 M PPCA 模型的建立由传统的一步改为两步: 第 1 步建立 GMM 模型, 第 2 步由 GMM 建立M PPCA 模型. 文中考虑了各模型内部主元对过程变量解释率的不同, 克服了M PPCA 模型内部各模型主元相同而与实际不符等缺点. 在M PPCA 模型子模型内使用基于 PPCA 的监控方法, 代替传统的基于 PCA 的监控方法, 克服了监控指标不一致和监控图过多的缺点.

参考文献(References)

[1] M ichael E Tipping, ChristopherM Bishop. M ixtures of Principal Component Analyzers[A]. *Proc of the IEE Fifth Int Conf on Artificial Neural Networks* [C]. London: Cambridge, 1997: 13-18.

[2] Tipping M E. Probabilistic Principle Component Analysis[J]. *J of the Royal Statistical Society*, 1999, 3 (1): 71-86.

[3] Chen J H, Liu J L. Mixture Principal Component Analysis Models for Process Monitoring[J]. *Industrial Engineering Chemical Research*, 1999, 38 (4): 1478-1488.

[4] Zhang F. A Mixture Probabilistic PCA Model for Multivariate Processes Monitoring [A]. *Proc of American Control Conf* [C]. Boston, 2004: 3111-3115.

[5] M ichael E Tipping, ChristopherM Bishop. Mixtures of Principal Component Analyzers [J]. *Neural Computation*, 1999, 11(2): 443-482.

[6] Dongsoon Kim, In Beum Lee. Process Monitoring Based on Probabilistic PCA [J]. *Chemometrics and Intelligent Laboratory Systems*, 2003, 67(2): 109-123.

[7] Mario A T Figueiredo, Anil K Jain. Unsupervised Learning of Finite Mixture Models[J]. *IEEE Trans on PAM I*, 2002, 24(3): 381-396.

[8] Sang Wook Choi, Jin Hyun Park, In Beum Lee. Process Monitoring Using a Gaussian Mixture Model via Principal Component Analysis and Discriminant analysis [J]. *Computers and Chemical Engineering*, 2004, 28(8): 1377-1387.

[9] Liu F, Zhao Z G. Chemical Separation Process Monitoring Based on Nonlinear Principal Component Analysis[A]. *Advances in Neural Networks — ISNN* [C]. Dalian, 2004: 798-803.