文章编号: 1001-0920(2008)04-0415-05

基于鲁棒规范变量分析的故障诊断方法

邓晓刚, 田学民
(中国石油大学 信息与控制工程学院, 山东 东营 257061)

摘 要: 针对工业过程的建模数据中含有离群点的情况,提出一种基于鲁棒规范变量分析(CVA)的故障诊断方法. 该方法使用相关系数的鲁棒估计代替传统的相关系数,通过基于粒子群算法的投影寻踪技术计算最大化鲁棒相关系数的规范变量,从而建立统计模型并监控统计量检测过程的变化. 连续搅拌反应器(CSTR)系统的仿真结果说明,鲁棒规范变量分析方法能在含离群点数据的基础上建立准确的统计模型,比规范变量分析更有效地监控过程变化.

关键词: 故障诊断; 离群点; 鲁棒规范变量分析; 粒子群优化算法; 投影寻踪技术
中图分类号: TP273 **文献标识码**: A

Fault diagnosis method based on robust canonical variate analysis

DENG Xiao-gang, TIAN Xue-min
(College of Information and Control Engineering, China University of Petroleum, Dongying 257061, China. Correspondent: TIAN Xue-min, E-mail: tianxm@hdpu.edu.cn)

Abstract: A fault diagnosis method based on robust canonical variate analysis is presented to analyze model data with outliers in industrial processes. This method replaces the traditional correlation coefficient with its robust estimator, and applies projection pursuit technique based on particle swarm optimization algorithm to calculate the canonical variate that maximizes the robust correlation coefficient. Then statistical model is built and monitoring statistics are constructed to detect process faults. The simulation results on a continuous stirred tank reactor(CSTR) system show that robust canonical variate analysis can built accurate statistical model from process data with outliers and monitor process changes more effectively than canonical variate analysis.

Key words: Fault diagnosis; Outlier; Robust canonical variate analysis; Particle swarm optimization algorithm; Projection pursuit technique

1 引 言

基于数据分析的方法是故障诊断和过程监控领域的一个热点问题,引起了学术界和工程人员的广泛关注. 该类方法不需要严格的机理模型,使用多元统计理论分析正常的过程历史数据并建立统计模型,适合于复杂工业过程的在线监控和故障诊断. 规范变量分析(CVA)方法是一种基于数据分析的故障诊断方法[1],现已引起人们的广泛研究. 文献[2]分析了 CVA 的基本原理并用于过程的故障诊断; [3,4]将 CVA 方法分别与主元分析法、动态主元分析法和偏最小二乘法进行比较; [5]将 CVA 方法与核函数技术相结合,提出一种非线性 CVA 故障诊断方法; [6]分析了基于 CVA 的变量重构技术,并应用于故障变量的分离.

CVA 方法用于故障诊断,对建模数据要求较为严格,数据噪声需要服从正态分布的假设,这样才能使用均值和标准差进行数据标准化处理,得到相关阵的无偏估计. 在实际生产过程中,由于传感器故障和过程偶然波动等原因,系统记录的数据中时常存在离群数据,这时建立的统计模型缺乏准确性,会使故障诊断系统的误报率和漏报率上升. 一种简单的处理方法是对数据进行剔除预处理,但并不能完全解决问题. 这是因为单变量消除离群点的方法并不适用于多变量相关的情况,而且有时还会破坏数据的相关性.

本文提出一种改进的鲁棒 CVA 算法,用以增强 CVA 方法对建模数据中离群点的适应能力. 改进的方法使用一种相关系数的鲁棒估计器代替传统

收稿日期: 2006-12-14; **修回日期**: 2007-04-19.
基金项目: 国家 863 计划项目(2004AA412050).
作者简介: 邓晓刚(1981—),男,山东广饶人,博士生,从事过程故障诊断的研究; 田学民(1955—),男,山东文登人,教授,博士生导师,从事过程动态模拟、先进控制与优化等研究.

的相关系数,从而使优化指标具有鲁棒性;以基于粒子群优化(PSO)算法的投影寻踪技术取代传统的分解算法,计算满足鲁棒优化指标的规范变量;最后建立统计模型,进行过程监控. 文中使用 CSTR 仿真系统来验证算法的有效性.

2 鲁棒 CVA 算法

2.1 CVA 算法

CVA 在统计分析中称为典型相关变量分析(CCA)[7],是一种线性降维技术. 它使两个变量集的相关性最大化.

对于两个随机向量 $x \in R^m$ 和 $y \in R^n$,记 $x = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$, $y = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$,如果 x 和 y 已经分别进行标准化处理,则可对 x 和 y 进行线性变换,得到两个新的变量

$$u = a^{\mathrm{T}}x = \sum_i^m a_i x_i,\ v = b^{\mathrm{T}}y = \sum_i^n b_i y_i. \quad (1)$$

CVA 方法的目的是找出使 u 与 v 之间的相关系数 $\rho(u,v)$ 最大的 a 和 b. 此时 u 和 v 是第 1 组规范变量,依次可找出第 2 组、第 3 组、…,各组规范变量之间互不相关. 多数情况下,只使用 k 组(k 不大于 m 和 n)规范变量便可反映 x 与 y 之间的相关情形. 考虑 u 和 v 方差为 1 的情形,CVA 问题可用数学语言描述如下:

$$\max \rho(u,v) = \max \rho(a^{\mathrm{T}}x, b^{\mathrm{T}}y), \quad (2)$$

$$\text{s.t. } \mathrm{Var}(u) = \mathrm{Var}(a^{\mathrm{T}}x) = 1,$$

$$\mathrm{Var}(v) = \mathrm{Var}(b^{\mathrm{T}}y) = 1. \quad (3)$$

如果变量的观测数据中不存在异常点,则假设数据满足正态分布. 常用的样本相关系数计算方式是 Pearson 相关系数[8],如下式所示:

$$\rho(u,v) = S_{uv}/(\sqrt{S_{uu}}\sqrt{S_{vv}}). \quad (4)$$

其中:S_{uv} 表示变量 u 和 v 观测数据的协方差,S_{uu} 和 S_{vv} 表示 u 和 v 观测数据的方差,标准化后为 1.

记数据矩阵 $X \in R^{m\times N}$, $Y \in R^{n\times N}$,分别表示由 N 次 x 和 y 的观测值形成的数据矩阵,则式(2)和(3)的描述变为

$$\max \rho(a^{\mathrm{T}}x, b^{\mathrm{T}}y) = \max\left\{\frac{1}{N-1}a^{\mathrm{T}}XY^{\mathrm{T}}b\right\}, \quad (5)$$

$$\text{s.t. } \frac{1}{N-1}a^{\mathrm{T}}XX^{\mathrm{T}}a = 1,$$

$$\frac{1}{N-1}b^{\mathrm{T}}YY^{\mathrm{T}}b = 1. \quad (6)$$

上述优化问题最终可归结为一个奇异值分解问题[1]或广义特征值分解问题[5-7].

2.2 基于投影寻踪技术的鲁棒 CVA 算法

Pearson 相关系数对数据中的离群点非常敏感,不具有鲁棒性,因此需要使用一种相关系数的鲁棒估计器来代替 Pearson 相关系数. 对于标准化的随机变量,Pearson 相关系数可通过标准差来构造[9,10]. 本文使用 Q_n 估计作为样本标准差的鲁棒估计,得到相关系数的鲁棒估计器. 其表达式为

$$\rho(u,v) = \frac{Q_n^2(u+v) - Q_n^2(u-v)}{Q_n^2(u+v) + Q_n^2(u-v)}. \quad (7)$$

对于任意随机变量 x, Q_n 估计定义为所有样本之间距离对的上 4 分位数[11]. 具体表达式为

$$Q_n(x) = 2.221\,9c_n\{|x_i - x_j|, i < j\}_{(k)}. \quad (8)$$

其中

$$k = \left[\frac{h}{2}\right] \approx \frac{1}{4}\left[\frac{n}{2}\right],$$

c_n 为小样本校正因子,当 $n \to \infty$ 时趋于 1.

由式(7)可将式(2)中的优化目标函数表示为

$$\max \rho(a^{\mathrm{T}}x, b^{\mathrm{T}}y) =$$

$$\max \frac{Q_n^2(a^{\mathrm{T}}x + b^{\mathrm{T}}y) - Q_n^2(a^{\mathrm{T}}x - b^{\mathrm{T}}y)}{Q_n^2(a^{\mathrm{T}}x + b^{\mathrm{T}}y) + Q_n^2(a^{\mathrm{T}}x - b^{\mathrm{T}}y)}. \quad (9)$$

约束条件同样需要进行鲁棒处理,如下式所示:

$$a^{\mathrm{T}}\Psi_{xx}a = 1,\ b^{\mathrm{T}}\Psi_{xx}b = 1. \quad (10)$$

其中 Ψ_{xx} 和 Ψ_{yy} 表示随机向量 x 和 y 的鲁棒相关阵,可通过鲁棒主元分析等方法计算获得[12].

式(7)在估计相关系数时使用了 Q_n 估计,而 Q_n 估计没有显式表达形式,因此 CVA 的传统解法不再适用于问题的求解,必须找出一种新的迭代算法. 本文使用投影寻踪算法来解鲁棒 CVA 问题.

投影寻踪技术是一种分析和处理高维数据尤其是非正态分布高维数据的探索性分析方法. 其主要思想是通过极小化(或极大化)某个投影指标,将高维数据投影到低维子空间,从而寻找反映高维数据结构或特征的投影,在低维空间对数据结构进行分析[13]. 主元分析法和规范变量分析法可看作是投影寻踪技术的特例.

为了迭代求解所有的规范化变量,使用投影寻踪算法之前,必须对鲁棒 CVA 算法的结构进行必要的变换. 对式(10)的 Ψ_{xx} 和 Ψ_{yy} 分别进行谱分解,得到

$$\Psi_{xx} = KD_xK^{\mathrm{T}},\ \Psi_{yy} = LD_yL^{\mathrm{T}}. \quad (11)$$

定义 $a^{*\mathrm{T}} = a^{\mathrm{T}}KD_x^{1/2}$, $b^{*\mathrm{T}} = b^{\mathrm{T}}LD_y^{1/2}$,则由式(10)和(11)可将约束条件转换为

$$a^{*\mathrm{T}}a^* = 1,\ b^{*\mathrm{T}}b^* = 1. \quad (12)$$

定义 $x^* = D_x^{-1/2}K^{\mathrm{T}}x$, $y^* = D_y^{-1/2}L^{\mathrm{T}}y$,则优化目标变为

$$\max \rho(a^{\mathrm{T}}x, b^{\mathrm{T}}y) = \max \rho(a^{*\mathrm{T}}x^*, b^{*\mathrm{T}}y^*). \quad (13)$$

至此,式(12)和(13)描述的鲁棒 CVA 算法优化问题便可用投影寻踪技术进行求解. 投影寻踪算

法首先找出满足鲁棒优化指标的第 1 对投影方向 a^* 和 b^*,然后在它们的正交补空间寻找满足优化指标的第 2 对投影方向. 依次类推,可以找出所有的投影方向. 在投影寻踪的过程中选择合适的优化算法,可以加快寻优的过程,更好地找出满足优化指标的投影方向.

本文选用粒子群优化算法(PSO),它是一种演化计算技术[14]. 该算法中的个体在学习自身经历的同时彼此相互作用,并且种群成员逐渐移入问题空间的更好区域,算法的实现过程简单,不受问题维数的影响,适合解决上述优化问题.

在粒子群算法的寻优过程中,问题的解被看作 m 维寻优空间的一个没有质量和体积的微粒 $x_i = [x_{i1}, x_{i2}, \cdots, x_{im}]$,优化指标被看作该微粒的适应值. 当一组粒子群初始化后,便开始在寻优空间以一定的速度 $v_i = [v_{i1}, v_{i2}, \cdots, v_{im}]$ 运动,运动速度的大小和方向受到微粒个体最优历史位置 p_{id} 和群体最优历史位置 p_{gd} 的影响. 微粒的运动方程如下:

$$v_{id} = wv_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}), \tag{14}$$

$$x_{id} = x_{id} + v_{id}. \tag{15}$$

其中:w 为惯性权重,c_1 和 c_2 为加速度常数,r_1 和 r_2 为 $[0,1]$ 之间的随机数.

2.3　模型比较

为验证鲁棒 CVA 方法的有效性,需要确定一个指标来比较两个统计模型的相似性.

投影方向 a 和 b 是 CVA 统计模型的转换变量. 如果两个 CVA 模型的转换变量完全相同或相近,则可认为两个 CVA 模型描述了同一模型. 投影方向的相似性比较通常用夹角的余弦来描述[15],例如投影方向 a_{I} 和 a_{II} 的相似性描述为

$$|\cos(\theta)| = \frac{|a_{\mathrm{I}}^{\mathrm{T}} a_{\mathrm{II}}|}{\|a_{\mathrm{I}}\| \, \|a_{\mathrm{II}}\|}. \tag{16}$$

如果夹角余弦的绝对值趋于 1,则两个方向为同一投影方向;否则,两个投影方向趋于正交.

3　基于鲁棒 CVA 算法的故障检测

设过程系统的输入输出为 $u_t \in R^{m_u}$, $y_t \in R^{m_y}$,则系统的线性状态空间表达式为

$$\begin{cases} x_{t+1} = \Phi x_t + Gu_t + w_t, \\ y_t = Hx_t + Au_t + Bw_t + v_t. \end{cases} \tag{17}$$

在给定的时刻 t,含有过去信息的向量记为

$$p_t = [y_{t-1}^{\mathrm{T}}, u_{y-1}^{\mathrm{T}}, \cdots, y_{t-h}^{\mathrm{T}}, u_{t-h}^{\mathrm{T}}]^{\mathrm{T}}, \tag{18}$$

含有未来输出信息的向量记为

$$f_t = [y_t^{\mathrm{T}}, \cdots, y_{t+l-1}^{\mathrm{T}}]^{\mathrm{T}}, \tag{19}$$

其中 h 和 l 的确定可参考文献[1]. 本文重点分析鲁棒 CVA 方法的思路,对于具体参数的确定方法不作详述.

对于输入数据矩阵,过去和未来的信息向量集分别为

$$P = [p_1, \cdots, p_t, \cdots], \quad F = [f_1, \cdots, f_t, \cdots]. \tag{20}$$

首先用式(20)中的 P 和 F 代替式(5)和(6)中的 X 和 Y,然后按照鲁棒 CVA 算法的思路建立统计模型,最后求出投影方向矩阵. 与 CVA 监控算法相似,鲁棒 CVA 算法同样可构造 3 个统计量进行过程监控,如下式所示:

$$\begin{cases} T_s^2 = p_t^{\mathrm{T}} A_s^{\mathrm{T}} A_s p_t, \quad T_r^2 = p_t^{\mathrm{T}} A_r^{\mathrm{T}} A_r p_t, \\ Q = p_t^{\mathrm{T}} (I - A_s A_s^{\mathrm{T}}) (I - A_s^{\mathrm{T}} A_s) p_t. \end{cases} \tag{21}$$

3 个统计量的置信限计算方式如下[1]:

$$\begin{cases} T_{s,\alpha}^2 = \dfrac{s(n^2-1)}{n(n-s)} F_\alpha(s, n-s), \\ T_{r,\alpha}^2 = \dfrac{r(n^2-1)}{n(n-r)} F_\alpha(r, n-r), \\ Q_\alpha = g\chi_q^2. \end{cases} \tag{22}$$

其中:s 为选取的 CVA 状态空间模型阶数,n 为建模样本数目,$r = h(m_u + m_y) - s$,g 和 q 为由 Q 统计量的鲁棒位置估计和鲁棒散度估计确定的参数.

4　仿真研究

本节对 CSTR 系统进行仿真,用于验证算法的有效性. 带有控制系统的 CSTR 系统如图 1 所示. 在系统反应过程中,物料 A 进入反应器发生一级不可逆反应,生成物质 B;同时放出大量的热,冷却剂通过夹套把热量带走.

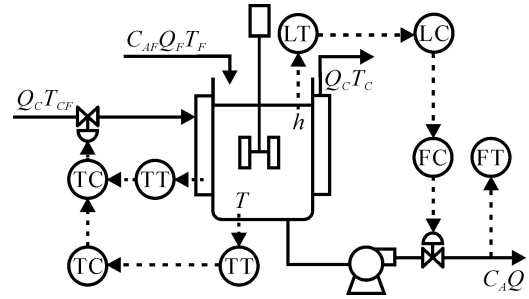


图 1　CSTR 系统

在程序仿真过程中,采集 CSTR 系统的 10 个变量(即图 1 标出的 10 个变量)数据作为过程数据,并将正态分布的高斯噪声添加到所有测量变量上. 仿真数据包括正常工况和 10 种故障情形(见表 1),每种情形的数据集中有 1 000 个采样点. 在正常工况数据的基础上,随机抽取 5 % 的数据,在原测量值上增加或减少 5 %,作为离群数据.

对于正常工况数据,滞后阶次 h 和 l 取值为 2,建立过去和未来信息向量集. 首先利用 CVA 方法得到统计模型,该模型是标准模型,记为 model Ⅰ;然

表 1 故障类型

故障	故障描述
F_1	进料量突然发生变化
F_2	进料温度逐渐发生变化
F_3	进料浓度逐渐发生变化
F_4	冷水散热能力降低
F_5	催化剂失活
F_6	冷却水进入温度变化
F_7	反应器内温度设定值变化
F_8	进料温度传感器出现故障
F_9	反应温度传感器出现故障
F_{10}	冷却水调节阀出现故障

后分别使用本文的鲁棒 CVA 方法和传统 CVA 方法,分析含有离群点的正常工况数据,建立模型 model Ⅱ和 model Ⅲ

3 个模型前 5 对规范变量的相关系数分布如表 2 所示. model Ⅰ是对正常工况数据进行 CVA 模型分析的描述,model Ⅱ中各对规范变量描述的相关系数与 model Ⅰ的情况非常相近. 从 model Ⅲ可以看出,前 3 对规范变量描述的数据相关性明显降低. 这是因为直接对含有离群点的数据进行 CVA 分析,相关变量构造受到离群点的影响,各对相关变量所解释的数据相关性偏离了真实模型.

模型结构主要体现在规范变量的转换向量 a 和 b 上. 以模型 model Ⅰ作为标准模型,分别计算它

表 2 3 个模型相关系数比较

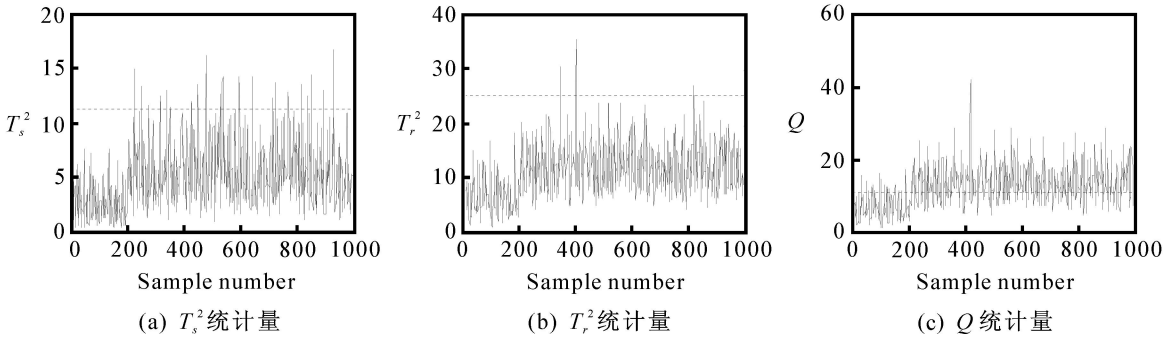
Order	model Ⅰ	model Ⅱ	model Ⅲ
1	0.881 1	0.856 9	0.581 8
2	0.614 5	0.564 5	0.254 5
3	0.530 0	0.518 3	0.233 8
4	0.245 0	0.279 8	0.208 2
5	0.214 0	0.270 6	0.195 4

与 model Ⅱ和 model Ⅲ各个转换变量角度余弦值的绝对值 $|\cos(\theta_i)|$ 和 $|\cos(\theta_i)|$. 前 5 对转换变量的比较如表 3 所示. 对于 model Ⅰ和 model Ⅱ,前 5 对转换变量之间夹角余弦值的绝对值有两对高达 0.95 以上,而 model Ⅰ和 model Ⅲ的转换向量之间夹角余弦值的绝对值均小于 0.90. 由此可见,model Ⅱ与 model Ⅰ的模型结构更为相近,而 model Ⅲ由于受离群点的影响,模型结构偏离了 CVA 统计模型 model Ⅰ.

表 3 CVA 模型结构比较

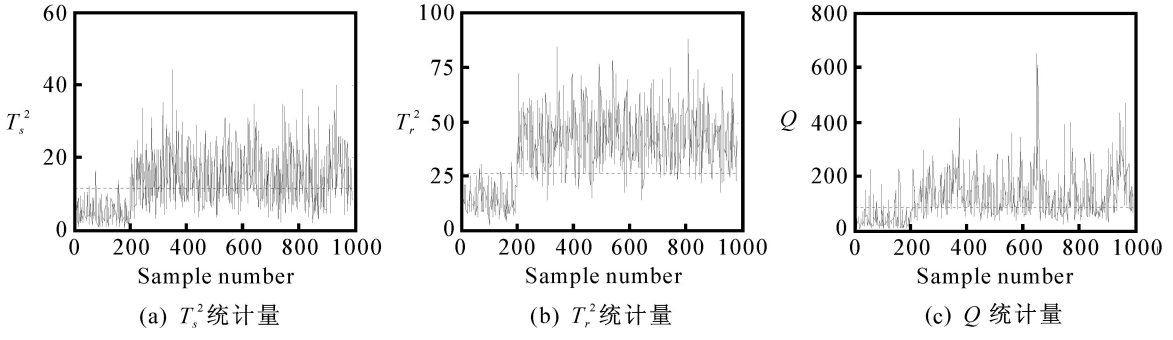
Order	$\lvert\cos(\theta_{\mathrm{Ⅰ,Ⅱ}})\rvert$		$\lvert\cos(\theta_{\mathrm{Ⅰ,Ⅲ}})\rvert$	
	a	b	a	b
1	0.979 3	0.983 4	0.791 1	0.108 7
2	0.707 0	0.858 9	0.779 7	0.845 8
3	0.984 0	0.970 4	0.544 1	0.359 2
4	0.739 3	0.426 1	0.497 2	0.062 5
5	0.271 0	0.564 5	0.244 2	0.031 5

分别使用 model Ⅱ和 model Ⅲ对所有仿真故



(a) T_s^2 统计量 (b) T_r^2 统计量 (c) Q 统计量

图 2 CVA 检测故障 F_8



(a) T_s^2 统计量 (b) T_r^2 统计量 (c) Q 统计量

图 3 鲁棒 CVA 检测故障 F_8

障进行检测,比较鲁棒 CVA 模型与传统 CVA 模型的检测效果,检测阈值为 95 %置信限. 这里仅列出故障 F_8 的检测结果.

故障 F_8 为在第 200 个采样时刻,进料温度传感器发生零点偏移故障. 鲁棒 CVA 方法和 CVA 方法的监控结果分别如图 2 和图 3 所示. 从图中可以看出,鲁棒 CVA 方法的 3 个统计量都能反映出故障的出现,而传统的 CVA 方法只有 Q 统计量曲线从第 200 个时刻起在阈值线附近波动, T_s^2 和 T_r^2 统计量虽然在趋势上有一定变化,但并没有明显超出阈值. 分析上述结果,由于离群点的存在,影响了数据预处理参数和 CVA 模型各参数,导致 CVA 方法过程监控效果变差.

5　结　　论

从仿真结果可以看出,在建模数据中存在离群点的情况下,基于投影寻踪技术的鲁棒 CVA 方法能建立准确的统计模型,及时检测到过程中出现的故障. 真实生产过程中的数据噪声并不满足正态分布,使用鲁棒 CVA 方法进行建模,有利于准确挖掘数据信息.

本文提出的方法主要针对离线建模阶段. 下一步工作将开展在线检测过程的方法研究,从而更有效地监控生产过程.

参考文献(References)

[1] Chiang L H, Russell E L, Braatz R D. Fault detection and diagnosis in industrial systems [M]. London: Springer-Verlag, 2001.

[2] Negiz A, Cinar A. Statistical monitoring of multivariate dynamic processes with state-space models[J]. AIChE J, 1997, 43(8): 2002-2020.

[3] Russell E L, Chiang L H, Braatz R D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis [J]. Chemometrics and Intelligent Laboratory Systems, 2000, 51(1): 81-93.

[4] Simoglou A, Martin E B, Morris A J. Statistical performance monitoring of dynamic multivariate process using state space modeling[J]. Computers and Chemical Engineering, 2002, 26(6): 909-920.

[5] 邓晓刚, 田学民. 基于核规范变量分析的非线性故障诊断方法[J]. 控制与决策, 2006, 21(10): 1109-1113.
(Deng Xiao-gang, Tian Xue-min. Nonlinear process fault diagnosis based on kernel canonical variate analysis [J]. Control and Decision, 2006, 21(10): 1109-1113.)

[6] Lee C, Choi S W, Lee I B. Variable reconstruction and sensor fault identification using canonical variate analysis [J]. J of Process Control, 2006, 16(7): 747-761.

[7] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 2003.
(Zhang Yao-ting, Fang Kai-tai. Introduction to multivairate statistical analysis [M]. Beijing: Science Press, 2003.)

[8] Abdullah M B. On a robust correlation coefficient[J]. The Statistician, 1990, 39(4): 455-460.

[9] Huber P J. Robust statistics[M]. New York: Wiley, 1981.

[10] Brunelli R, Messelodi S. Robust estimation of correlation with application to computer vision [J]. Pattern Recognition, 1995, 28(6): 833-841.

[11] Rousseuw P J, Croux C. Alternatives to the median absolute deviation [J]. J of American Statistic Association, 1993, 88(424): 1273-1283.

[12] Daszykowski M, Kaczmarek K, Heyden Y V, et al. Robust statistics in data analysis — A review: Basic concepts[J]. Chemometrics and Intelligent Laboratory Systems, 2007, 85(2): 203-219.

[13] Branco J A, Croux C, Filzmoser P, et al. Robust canonical correlations: A comparative study [J]. Computational Statistics, 2005, 20(2): 203-229.

[14] Kennedy J, Eberhart R. Particle swarm optimization [C]. Proc of IEEE Int Conf on Neural Networks. Piscataway: IEEE Service Center, 1995: 1942-1948.

[15] Raich A, Cinar A. Diagnosis of process disturbance by statistical distance and angle measures[J]. Computers and Chemical Engineering, 1997, 21(6): 661-673.