

DOI: 10.19650/j.cnki.cjsi.J2007158

# 基于多模态数据全信息的概率主成分 分析故障检测研究\*

李元<sup>1</sup>, 张昊展<sup>1</sup>, 唐晓初<sup>2</sup>

(1. 沈阳化工大学信息工程学院 沈阳 110142; 2. 沈阳航空航天大学自动化学院 沈阳 110136)

**摘要:**针对工业过程复杂的数据分布特性,本文提出了一种基于局部近邻标准化(LNS)的概率主成分分析(PPCA)故障检测方法(LNSPPCA)来解决由于过程数据的多模态特性和不确定性所引起的故障检测效果不理想问题。首先,通过LNS解决数据多模态问题,使标准化后数据尽可能的服从单一高斯分布,然后,使用PPCA方法从概率的角度对数据进行分析,能够考虑到数据的随机性,从而更真实的描述数据,提取更加全面有价值的信息,有效的在复杂的数据分布过程中对故障进行检测。因此,LNSPPCA方法可以有效提高多模态过程复杂数据分布的工业过程故障检测能力。利用数值例子和TE过程进行应用实验,并将测试结果与主成分分析法(PCA)、PPCA方法进行对比,验证了LNSPPCA方法的有效性。

**关键词:**多模态;全信息;局部近邻标准化;概率主成分分析;故障检测

**中图分类号:** TP277 TH165+.3 **文献标识码:** A **国家标准学科分类代码:** 510.80

## Study on probabilistic principal component analysis fault detection based on full information of multimodal data

Li Yuan<sup>1</sup>, Zhang Haozhan<sup>1</sup>, Tang Xiaochu<sup>2</sup>

(1. School of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China;

2. School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

**Abstract:** Aiming at the complex data distribution characteristics of industrial processes, this paper proposes a probabilistic principal component analysis fault detection method based on local neighborhood standardization (LNSPPCA) to solve the problem of unsatisfactory fault detection effect caused by multi-modal characteristics and uncertainty of process data. Firstly, LNS is used to solve the data multi-modal problem, so that the standardized data obey a single Gaussian distribution as much as possible. Then, the PPCA method is used to analyze the data from the perspective of probability, which can take into account the randomness of the data, so as to describe the data more realistically, extract more comprehensive and valuable information, and effectively detect faults in the complex data distribution process. Therefore, the LNSPPCA method can effectively improve the industrial process fault detection capability in multi-modal process complex data distribution. Numerical examples and TE process were used to conduct application experiments, and the test results are compared with those of principal component analysis (PCA) and PPCA methods, which verifies the effectiveness of the LNSPPCA method.

**Keywords:** multimodal; full information; local neighborhood standardization; probabilistic principal component analysis; fault detection

## 0 引言

随着现代科学技术的进步与发展,工业生产过程的

现代化水平和复杂性日益提高,在这些工业生产过程中如果发生故障,将会造成严重的财产损失和人员伤亡。因此,故障检测技术在现代化工业生产中的地位变的尤为重要<sup>[1]</sup>。

收稿日期:2020-11-23 Received Date: 2020-11-23

\* 基金项目:国家自然科学基金(61673279)资助项目

主元分析 (principal component analysis, PCA) 作为多元统计过程监控的方法, 已经被现代工业过程故障检测所广泛应用<sup>[2]</sup>。PCA 在工业过程检测时通常将原始空间划分为主元子空间 (principal component subspace, PCS) 和残差子空间 (residual subspace, RS), 然后分别使用 Hotelling's  $T^2$  和平方预测误差 (square prediction error, SPE) 作为统计量来对工业过程进行监控<sup>[3]</sup>。然而, PCA 假设所有数据都服从标准正态分布, 把多维原始数据直接投影到低维空间中, 没有考虑到数据的随机性, 忽略了噪声的影响, 当面对复杂的大规模数据时, 不能从数据中提取更全面有价值的信息, 造成监控效果存在误差<sup>[4]</sup>。为了解决上述问题, Tipping 等<sup>[5]</sup>提出了概率主成分分析 (probabilistic principal component analysis, PPCA), PPCA 在概率框架下推导 PCA, 对噪声也进行同步监控, 建立过程的概率生成模型, 并用贝叶斯理论公式来估计潜在变量的后验概率, 利用最大期望 (expectation-maximization algorithm, EM) 方法迭代估计参数。工业过程中的所有数据都包含着噪声, 其在根本上有很大的不确定性, 这样通过概率的方法可以更好的处理这些随机性问题, 更真实的描述数据, 从而可以从数据中获得更全面有用的信息<sup>[6]</sup>。但是多数工业过程中数据存在多模态特性, 不服从单一的高斯分布, 然而, PPCA 要求数据服从单一的高斯分布, 才能更好的考虑随机性从而进行故障检测。但是当原始数据为多模态数据时, PPCA 中使用的 Z-score 标准化方法在故障检测中的效果受到了限制。因为 Z-score 对数据进行标准化处理时利用的是总体样本的均值和标准差, 所以在处理多模态数据时, 并不能很好实现消除多模态特征的目的。针对数据多模态问题, He 等<sup>[7]</sup>提出了一种基于  $K$  近邻的故障检测方法。该方法的故障检测指标为局部距离之和  $D^2$ , 有效降低了多模态的影响。但当多模态数据较多且方差差异较大时, 该方法的计算量变大且检测能力降低。Ma 等<sup>[8]</sup>提出了一种基于局部近邻标准化的故障检测方法, 该方法基于局部近邻标准化很好的处理多模态问题, 使数据尽可能的服从单一高斯分布。然而, 后续基于 PCA 方法进行故障检测, 对数据的随机性具有一定的局限性。最近, 一些基于概率的方法被提出来解决数据的随机性问题。Yu 等<sup>[9]</sup>提出了一种基于贝叶斯的有限高斯混合模型的方法来进行多模态故障检测。Yang 等<sup>[10]</sup>提出了一种基于整合混合概率主成分分析的方法来解决过程数据不确定性问题, Zhou 等<sup>[11]</sup>提出了一种基于切换自回归动态潜变量模型的方法来进行动态故障检测。上述这些方法都建立了概率混合模型, 不仅考虑了数据的随机性, 而且很好的处理了多模态问题, 但是在这些方法中都是使用概率混合模型并用 EM 算法迭代估计多个混合参数, 增加了计算的复杂度。

本文针对工业多模态过程复杂的数据分布特性, 提出了一种基于局部近邻标准化的概率主成分分析故障检测方法。首先, 在训练样本中寻找训练样本点的  $K$  近邻, 应用局部近邻标准化对当前样本进行处理并构建 PPCA 模型, 然后, 根据贝叶斯理论公式计算出潜在变量的后验概率分布, 并用 EM 法迭代计算参数, 最后, 计算出统计量和控制限进行故障检测。LNSPPCA 集成了局部近邻标准化 (local neighborhood standardization, LNS) 和 PPCA 处理多模态和数据分布复杂的优势, 提取更全面有用的信息, 可以很好的对多模态数据分布复杂的工业过程进行故障检测, 通过应用实验进一步证明了本文提出的方法的有效性。

## 1 PPCA 算法

PPCA 算法将原始观测样本分为主元和噪声两个数据集, 对主元子空间和噪声子空间同时进行监控, 更加全面的提取有价值的信息。PPCA 模型可以由以下方式构建<sup>[12]</sup>:

$$\mathbf{x} = \mathbf{P}t + e \quad (1)$$

式中:  $\mathbf{x} \in \mathbf{R}^n$  为过程观测变量;  $t \in \mathbf{R}^k$  为过程潜在变量;  $\mathbf{P} \in \mathbf{R}^{n \times k}$  ( $n > k$ ) 为负载矩阵;  $e \in \mathbf{R}^n$  代表方差为  $\sigma$  的测量噪声<sup>[13]</sup>。假设  $t$  和  $e$  服从高斯分布, 即  $t \sim N(0, \mathbf{I})$ ;  $e \sim N(0, \sigma^2 \mathbf{I})$ , 则潜在变量  $t$  的先验概率分布为:

$$p(t) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2} \mathbf{x}^T \mathbf{x}\right\} \quad (2)$$

观测变量  $x$  关于潜在变量  $t$  空间上的概率分布为:

$$p(x | t) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{P}t\|^2\right\} \quad (3)$$

在上式概率分布上对潜在变量  $t$  求边际分布, 则观测变量  $\mathbf{x}$  的概率分布为:

$$p(\mathbf{x}) = \int p(\mathbf{x} | t) p(t) dt = (2\pi)^{-n/2} |\mathbf{C}|^{-1/2} \exp\left\{-\frac{1}{2} \mathbf{x}^T \mathbf{C}^{-1} \mathbf{x}\right\} \quad (4)$$

其中,  $\mathbf{C} = \mathbf{P}\mathbf{P}^T + \sigma^2 \mathbf{I}$  为观测变量的方差。根据贝叶斯理论概率公式:  $p(t | \mathbf{x}) = p(t)p(\mathbf{x} | t)/p(\mathbf{x})$ , 潜在变量  $t$  的后验概率分布为:

$$p(t | \mathbf{x}) = (2\pi)^{-k/2} |(\sigma^2)^{-1} \mathbf{M}|^{-1/2} \times \exp\left\{-\frac{1}{2} [t - \mathbf{M}^{-1} \mathbf{P}^T \mathbf{x}^T]^T \cdot [(\sigma^2)^{-1} \mathbf{M}] [t - \mathbf{M}^{-1} \mathbf{P}^T \mathbf{x}^T]\right\} \quad (5)$$

其中,  $\mathbf{M} = \mathbf{P}\mathbf{P}^T + \sigma^2 \mathbf{I}$ 。为了计算出参数  $\mathbf{P}$  和  $\sigma^2$ , 可将潜在变量  $t$  当作遗失数据, 则观测数据  $\mathbf{x}$  和潜在数据  $t$  的联合分布的概率为:

$$p(x_i, t_i) = p(x_i | t_i) p(t_i) =$$

$$(2\pi)^{-n/2} \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{P}t_i\|^2}{2\sigma^2}\right\} (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right\} \quad (6)$$

采用 EM 算法迭代计算模型参数  $\mathbf{P}$  和  $\sigma^2$  的最大似然估计,其计算步骤分为求期望 ( $E$  步) 和最大化阶段 ( $M$  步),基本过程为:

$E$  步求解关于  $p(t_m | x_m, \mathbf{P}, \sigma^2)$  分布的期望:

$$\langle L_c \rangle = - \sum_{i=1}^n \left\{ \frac{n}{2} \ln \sigma^2 + \frac{1}{2} \text{tr}(\langle x_i x_i^T \rangle) + \frac{1}{2} x_i^T x_i - \frac{1}{\sigma^2} \langle x_i \rangle^T \mathbf{P}^T x_i + \frac{1}{2\sigma^2} \text{tr}(\mathbf{P}^T \mathbf{P} \langle x_i x_i^T \rangle) \right\} \quad (7)$$

其中,  $\langle x_i \rangle = \mathbf{M}^{-1} \mathbf{P}^T t_i \langle x_i x_i^T \rangle = \sigma^2 \mathbf{M}^{-1} + \langle x_i \rangle \langle x_i \rangle^T M$  步通过计算最大化目标函数期望来获得新参数值  $\mathbf{P}_{new}$  和  $\sigma_{new}^2$ 。

$$\mathbf{P}_{new} = \mathbf{S} \mathbf{P} (\sigma^2 \mathbf{I} + \mathbf{M}^{-1} \mathbf{P}^T \mathbf{S} \mathbf{P})^{-1} \quad (8)$$

$$\sigma_{new}^2 = \frac{1}{n} (\mathbf{S} - \mathbf{S} \mathbf{P} \mathbf{M}^{-1} \mathbf{P}^T) \quad (9)$$

其中,  $\mathbf{S} = \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T$ , 表示观测数据的协方差矩阵。

通过迭代计算  $E$  步和  $M$  步来估计 PPCA 模型的参数值,直到目标函数的期望值收敛,从而获得参数  $\mathbf{P}$  和  $\sigma^2$  的值<sup>[14]</sup>。

由于 PPCA 仅使用马氏范数这一种量度对主元和噪声进行检测,所以可以直接对观测数据的白化值范数进行监控就能表示主元和噪声监控的综合结果。Kim 等<sup>[15]</sup> 和 Yang 等<sup>[16]</sup> 提出基于 PPCA 的综合统计量  $ST$  来监控过程是否发生异常,统计量  $ST$  的控制限  $ST_{cl}$  可以由核密度估计法求得。 $ST$  统计量的计算方法为:

$$ST = \left\| (\mathbf{P} \mathbf{P}^T + \sigma^2 \mathbf{I})^{-0.5} \mathbf{x}_i \right\|^2 \quad (10)$$

## 2 LNSPPCA 算法

局部近邻标准化的核心思想是寻找样本集中的每个样本点  $x_i$  的  $K$  近邻,并组成其近邻集  $n(x_i)$ ,下一步计算出该近邻集的均值和标准差,分别为  $m(n(x_i))$  和  $s(n(x_i))$ ,最后使用近邻集中的均值和标准差根据式(11)对当前样本  $x_i$  进行标准化处理<sup>[17]</sup>。

$$\mathbf{x}_i = \frac{\mathbf{x}_i - m(n(\mathbf{x}_i))}{s(n(\mathbf{x}_i))} \quad (11)$$

由式(1)可看出,局部近邻标准化方法能够有效的降低多模态数据中心漂移和方差差异显著的影响,并且可以让数据进行标准化后尽可能的服从单峰高斯分布。为了提高 PPCA 方法在多模态工业数据下的故障检测能力,本文提出 LNSPPCA 故障检测方法,主要包括离线建模和在线检测两个阶段,其故障检测步骤如图 1 所示。

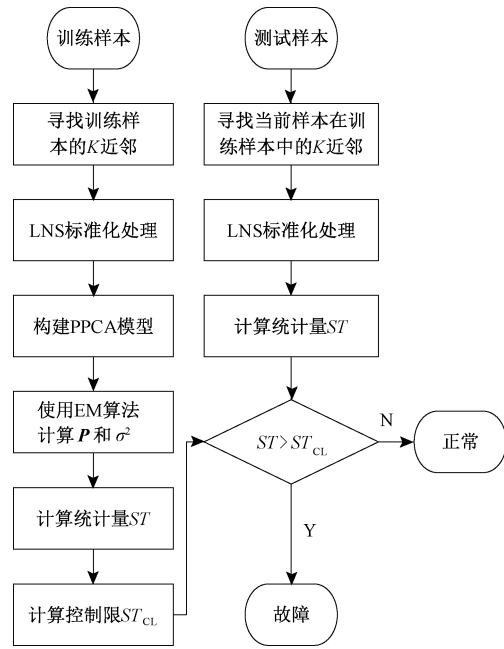


图 1 LNSPPCA 故障检测步骤

Fig. 1 LNSPPCA fault detection procedure

### 2.1 离线建模

使用正常过程样本数据经过处理后,构建起 PPCA 模型并计算统计量和控制限,其具体步骤如下:

- 1) 采集正常过程中样本数据  $\mathbf{X}$ 。
- 2) 在样本数据中,查找每个样本  $x_i$  的  $K$  近邻集  $n(x_i)$ ,并使用式(11)对每一个样本进行 LNS 标准化处理。
- 3) 构建 PPCA 模型并使用 EM 算法迭代计算参数  $\mathbf{P}$  和  $\sigma^2$  的值。
- 4) 根据式(10)计算统计量  $ST$ 。
- 5) 使用核密度估计法计算出  $ST$  统计量的控制限  $ST_{cl}$ 。

### 2.2 在线检测

假设测试样本  $x_*$  经过处理后,使用已构建的 PPCA 模型对其进行故障检测,并判断是否为故障样本,其具体步骤如下:

- 1) 在训练数据  $\mathbf{X}$  中找到  $x_*$  的  $K$  近邻集  $n(x_*)$  并使用式(11)进行标准化处理。
- 2) 通过已构建的 PPCA 模型对测试样本进行故障检测,并求出统计量  $ST$ 。
- 3) 将步骤 2) 计算出的统计量  $ST$  与离线建模步骤中的控制限  $ST_{cl}$  进行比较,若  $ST > ST_{cl}$ ,则认为该样本为故障样本,反之为正常样本。

### 3 数值例子

本文利用一个简单的多元线性系统来验证所提出的 LNSPPCA 故障检测方法的有效性。仿真模拟数据来自以下多元线性系统。

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0.5195 & 0.4292 \\ 0.4131 & 0.0323 \\ 0.7121 & 0.5143 \\ 0.7261 & 0.6264 \\ 0.2019 & 0.5079 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} \quad (12)$$

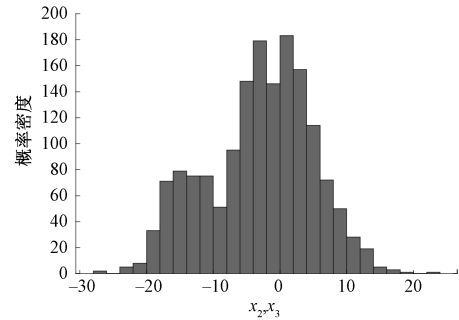
其中,  $e_1 \sim e_5$  为服从  $N(0, 0.01)$  的噪声;  $s_1$  和  $s_2$  为控制系统的主要变量, 在系统仿真过程中通过改变  $s_1$  和  $s_2$  可以改变系统的运行模式, 从而产生多模态数据。两种模态数据的产生方式如式(13)、(14)所示。

$$\text{Model1:} \begin{cases} s_1 \sim N(3, 6) \\ s_2 \sim N(6, 4) \end{cases} \quad (13)$$

$$\text{Model2:} \begin{cases} s_1 \sim N(10, 5) \\ s_2 \sim N(9, 2) \end{cases} \quad (14)$$

过程运行共产生 800 个正常样本组成训练集, 其中包含两个不同模态, 每个模态各包含 400 个样本。为了验证本文提出的方法在多模态下的有效性, 在过程正常运行时, 从 201~400 时刻在变量  $x_2$  上添加斜坡故障 0.1 ( $k-200$ ), 从 601~800 时刻在变量  $x_5$  上, 添加幅值为 20% 的阶跃故障。本文使用 PCA、PPCA 和 LNSPPCA 3 种方法对上述数值例子进行故障检测分析, 通过 90% 累积贡献率确定 PCA 中的主元个数为 2, PPCA 和 LNSPPCA 中的主元个数设为 2, LNSPPCA 中的局部近邻  $K$  值设为 40。

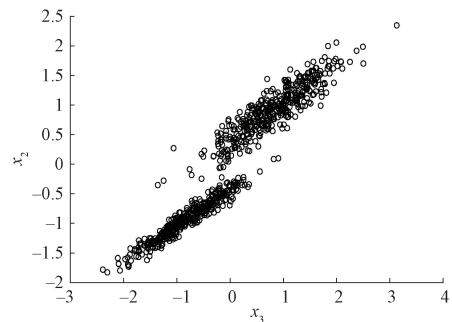
图 2~4 分别为原始数据、经过 Z-score 处理后和经过 LNS 处理后数据的散点及高斯分布图。如图 3 所示, 原始数据多模态特性在经过 Z-score 处理后两个模态数据方差差异明显, 数据仍然具有多模态特性, 这会影响到传统的 PPCA 的故障检测性能。如图 4 所示, 数据经过 LNS 处理后, 原始样本的多模态数据被缩放至各个模态的中心, 并且数据被近似融合为服从单峰高斯分布的单一模态。



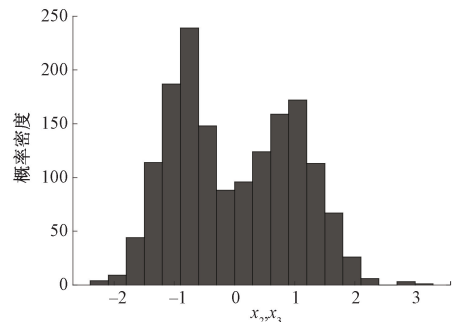
(b) 原始数据高斯分布  
(b) Raw data Gaussian distribution

图 2 原始数据散点及高斯分布

Fig. 2 Raw data scatter and Gaussian distribution



(a) Z-score处理后数据散点图  
(a) Data scatter diagram after Z-score processing

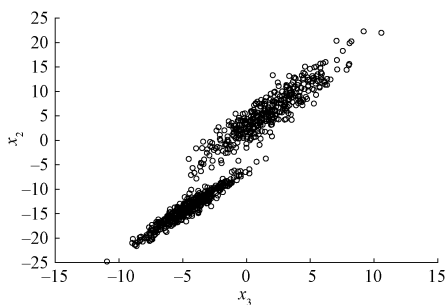


(b) Z-score处理后数据高斯分布  
(b) Data Gauss distribution after Z-score processing

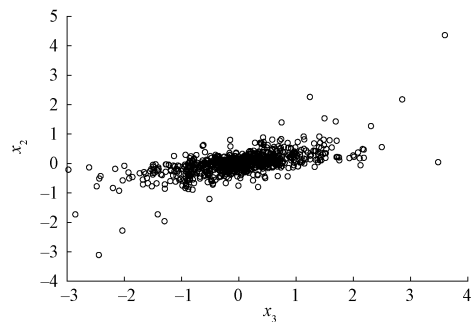
图 3 Z-score 处理后数据散点及高斯分布

Fig. 3 Data scatter and Gaussian distribution

after Z-score processing



(a) 原始数据散点图  
(a) Raw data scatter diagram



(a) LNS处理后数据散点图  
(a) Data scatter diagram after LNS processing

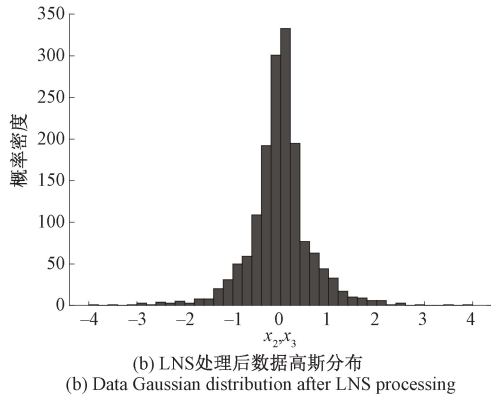
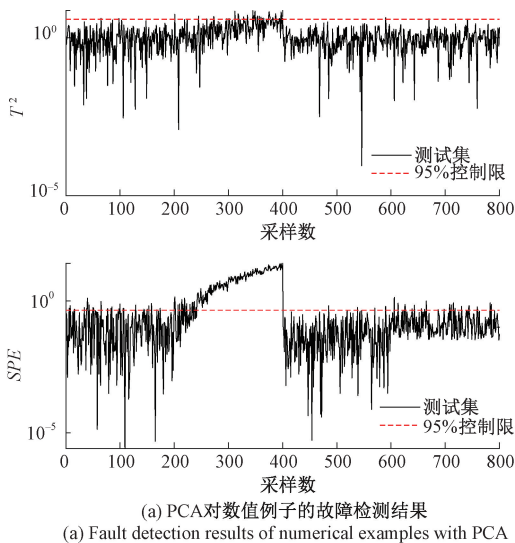


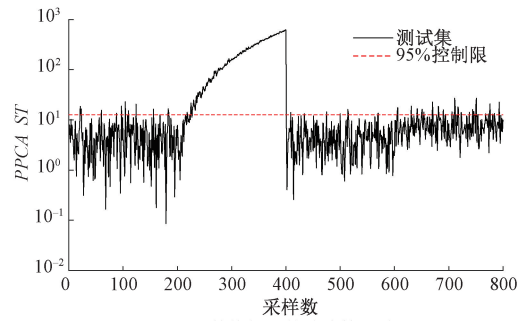
图 4 LNS 处理后数据散点及高斯分布

Fig. 4 Data scatter and Gaussian distribution after LNS processing

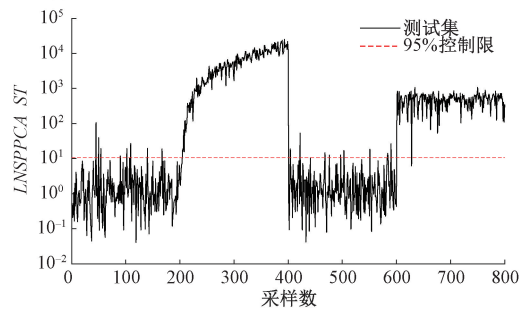
图 5 给出了 PCA,PPCA 和 LNSPPCA 3 种方法对此数值例子的故障检测效果。对比 PCA 和 PPCA 两种方法,从图 5(a)、(b)中可以看出,在前 400 时刻,PPCA 有更好的检测效果,从第 201 时刻开始,故障可以被准确的检测出来,而 PCA 对故障并不敏感,检测的准确率也比较低。PPCA 从概率的角度建立模型并分析数据,减小了过程数据的随机性,考虑了噪声变量中存在更多有用的信息,所以 PPCA 的故障检测效果优于 PCA。但是两种方法在第 601 时刻开始,对故障的检测率降低,说明 PPCA 只适合单一模态的故障检测,对于多模态数据并不能很好的进行检测。本文提出的 LNSPPCA 故障检测方法结果如图 5(c) 所示,LNSPPCA 使多模态数据处理后近似服从单一高斯分布,并且可以提取更加全面有价值的信息,使用基于 LNSPPCA 算法进行故障检测后,故障在第 201~400 采样点和第 601~800 采样点可以被准确的检测出来,具有很好的检测效果。因此 LNSPPCA 具有较高的检测率。



(a) PCA对数值例子的故障检测结果  
(a) Fault detection results of numerical examples with PCA



(b) PPCA对数值例子的故障检测结果  
(b) Fault detection results of numerical examples with PPCA



(c) LNSPPCA对数值例子的故障检测结果  
(c) Fault detection results of numerical examples with LNSPPCA

图 5 3 种方法对数值例子的故障检测结果

Fig. 5 Fault detection results of numerical examples with three methods

#### 4 LNSPPCA 在 TE 工业过程的应用

伊斯曼化学品公司创建了田纳西-伊斯曼过程(TE 过程),TE 过程是一个典型的复杂多变量化工过程,近年来 TE 过程已经在故障检测研究领域得到了广泛的应用<sup>[18]</sup>。其工艺流程如图 6 所示,本节应用实验中的所有数据均来源于 TE 过程。

TE 过程一共有 21 个故障,每个数据集都包括 52 个观测变量,其 21 个故障分为 4 种类型<sup>[19]</sup>。具体故障数据类型的描述如表 1 所示。

本文应用实验使用的为 TE 过程生产模式 1 和 3,不同模式的生产过程需要使用不同的生产物 G/H 的比例,具体如表 2 所示,将两个模态的数据混合使用,即训练数据为 961 个,用于检测的样本数为 1 922 个,其中包含各个模态的正常数据 160 个和对应故障的 800 个数据样本<sup>[20]</sup>。通过 90% 累积贡献率确定 PCA 中的主元个数为 28,PPCA 和 LNSPPCA 中的主元个数设为 15,LNSPPCA 中的局部近邻 K 值设为 50,监控统计量的控制限均设置为 95%。

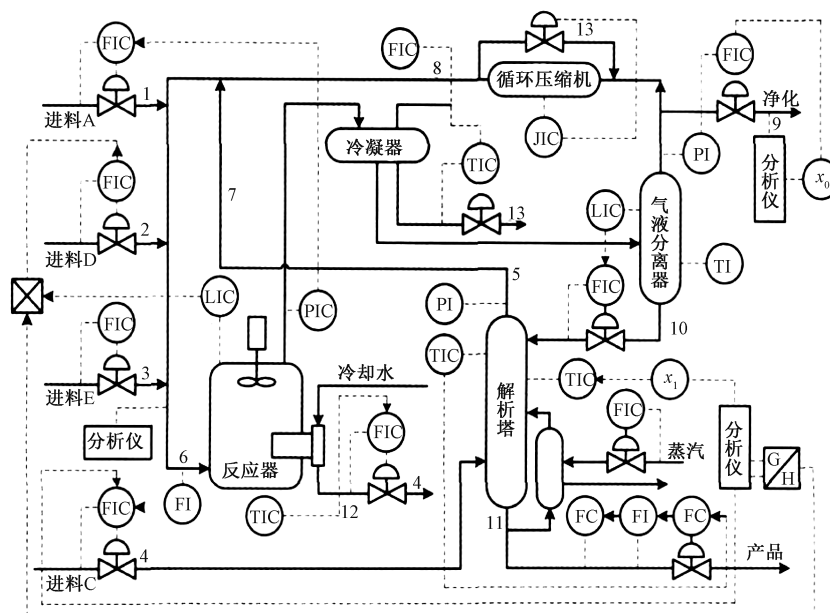


图6 田纳西-伊斯曼过程工艺流程图

Fig. 6 Tennessee Eastman (TE) process technological flow chart

表1 故障类型  
Table 1 Fault types

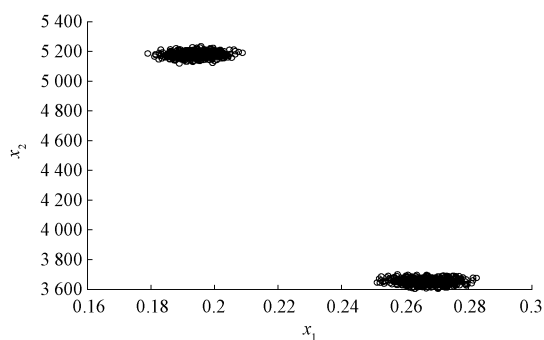
故障编号	性质描述	变化类型
1	物料 A/C 进料比改变, 物料 B 含量不变	阶跃
2	物料 A/C 进料比不变, 物料 B 含量改变	阶跃
3	物料 D 进料温度改变	阶跃
4	反应器冷却入口温度改变	阶跃
5	冷凝器冷却入口温度改变	阶跃
6	物料 A 进料损失	阶跃
7	物料 C 压力损失	阶跃
8	物料 A、B、C 的组成比例改变	随机变量
9	物料 D 进料温度改变	随机变量
10	物料 C 进料温度改变	随机变量
11	反应器冷却水入口温度改变	随机变量
12	冷凝器冷却水入口温度改变	随机变量
13	反应动力学参数改变	慢偏移
14	反应器冷却阀门	粘住
15	冷凝器冷却阀门	粘住
16	未知	未知
17	未知	未知
18	未知	未知
19	未知	未知
20	未知	未知
21	物流 4 阀门固定在恒定位置	恒定位置

表2 TE 过程生产模式

Table 2 TE process production mode

生产模式	G/H 比率	产品生产率
1	50/50	7 038 kgh <sup>-1</sup> G 和 7 038 kgh <sup>-1</sup> H
3	90/10	1 000 kgh <sup>-1</sup> G 和 1 111 kgh <sup>-1</sup> H

如图 7~9 所示,任取两个变量方向分别绘制原始数据、Z-score 标准化处理后和 LNS 标准化处理后的数据散点及高斯分布图。由图 8 看出,经过 Z-score 标准化处理后数据变量间的方差差异依旧明显,多模态结构特性显著,经过 LNS 方法对数据进行处理后,将多模态数据近似的转换为服从单峰高斯分布的单一模态数据,其结果如图 9 所示。



(a) 多模态TE过程变量散点图

(a) Variable scatter diagram of multi modal TE process

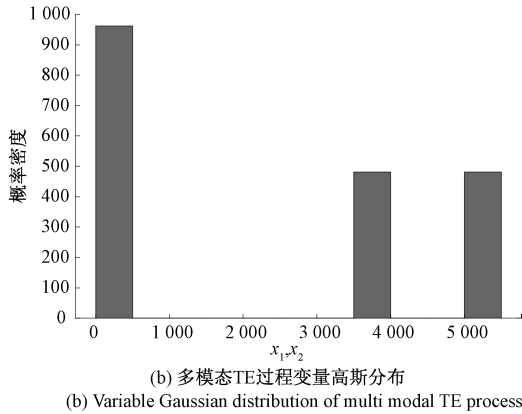


图 7 多模态 TE 过程变量散点及高斯分布  
Fig. 7 The variable scatter and Gaussian distribution of the multi-modal TE process

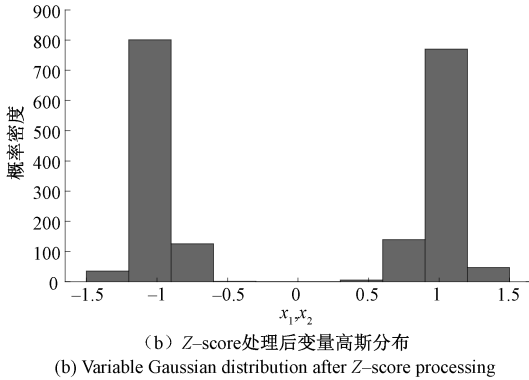
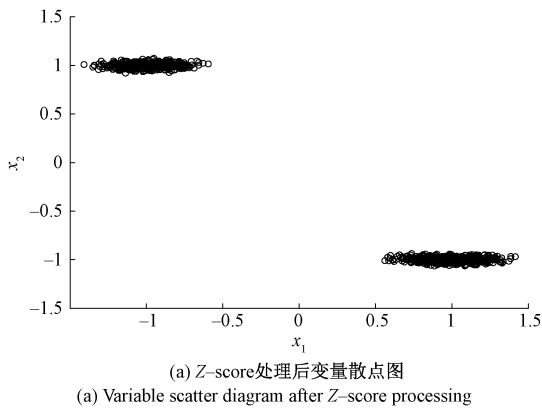


图 8 Z-score 处理后变量散点及高斯分布  
Fig. 8 Variable scatter and Gaussian distribution after Z-score processing

由表 3 可知,纵观比较 3 种方法对 21 个故障的检测率, LNSPPCA 的检测率皆高于其余两种方法。LNSPPCA、PPCA 和 PCA 对故障 1,6,8,13,20 的检测率较高,均为 80% 以上,而对于故障 10,11,14,18,PPCA 和 PCA 的检测率都没有 LNSPPCA 好,也均在 80% 以下,其中在这 4 个故障中,PPCA 对故障的检测能力优于 PCA。

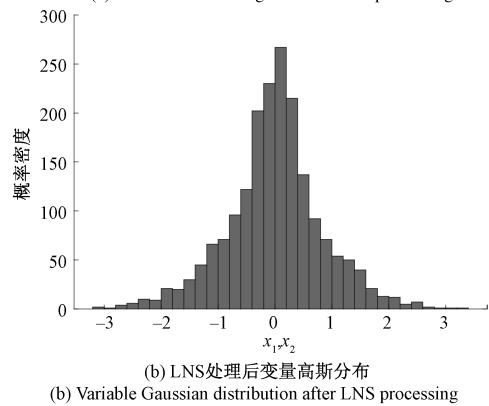
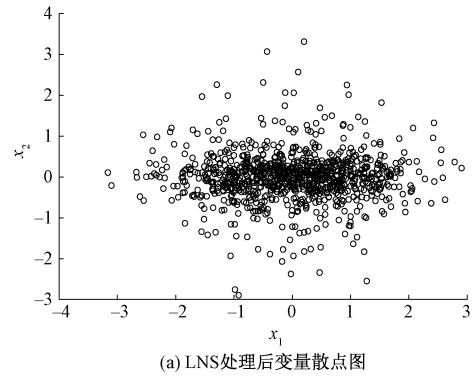


图 9 LNS 处理后变量散点及高斯分布  
Fig. 9 Variable scatter and Gaussian distribution after LNS processing

表 3 3 种方法检测率汇总表

Table 3 Summary of the detection rates for three methods

故障编号	PCAT <sup>2</sup>	PCASPE	PPCAST	LNSPPCAST
1	44.69	99.19	93.44	99.69
2	37.50	69.50	85.31	98.81
3	3.63	6.50	5.44	7.07
4	4.50	5.81	6.56	6.81
5	46.75	51.06	50.38	51.31
6	95.88	98.13	98.13	98.13
7	7.69	10.88	11.19	25.19
8	81.81	94.81	93.81	98.12
9	7.00	11.56	10.88	16.19
10	5.75	18.81	67.75	89.25
11	28.38	37.44	70.69	91.63
12	49.81	56.75	58.81	67.06
13	84.94	89.75	93.88	95.19
14	4.69	10.25	69.06	98.63
15	3.63	6.13	4.94	11.13
16	4.00	5.88	5.69	6.31
17	12.75	75.06	83.63	93.63
18	47.25	51.31	59.44	84.50
19	40.38	46.81	59.38	61.94
20	77.19	82.19	90.13	94.13
21	3.44	5.81	5.56	5.94

在故障 2 中,物料 A/C 进料比不变,物料 B 含量产生了一个阶跃性改变,导致总进料量偏离正常运行状态,扰乱了系统的正常运行,从而发生故障,在此故障中, LNSPPCA 的检测率高于 PCA 和 PPCA 两种检测方法,分别提高了 13.5% 和 29.31%,其中,PPCA 相比于 PCA 对于数据的描述更加真实,提取的信息更加全面,所以检测率比 PCA 高。3 种方法对于故障 2 的检测结果如图 10 所示,在图 10(c)中,使用 LNSPPCA 对故障 2 检测后,正常数据和故障数据区分明显,故障数据分别从 161 个采样点后和 1 122 个采样点后超出 95% 控制限,对比

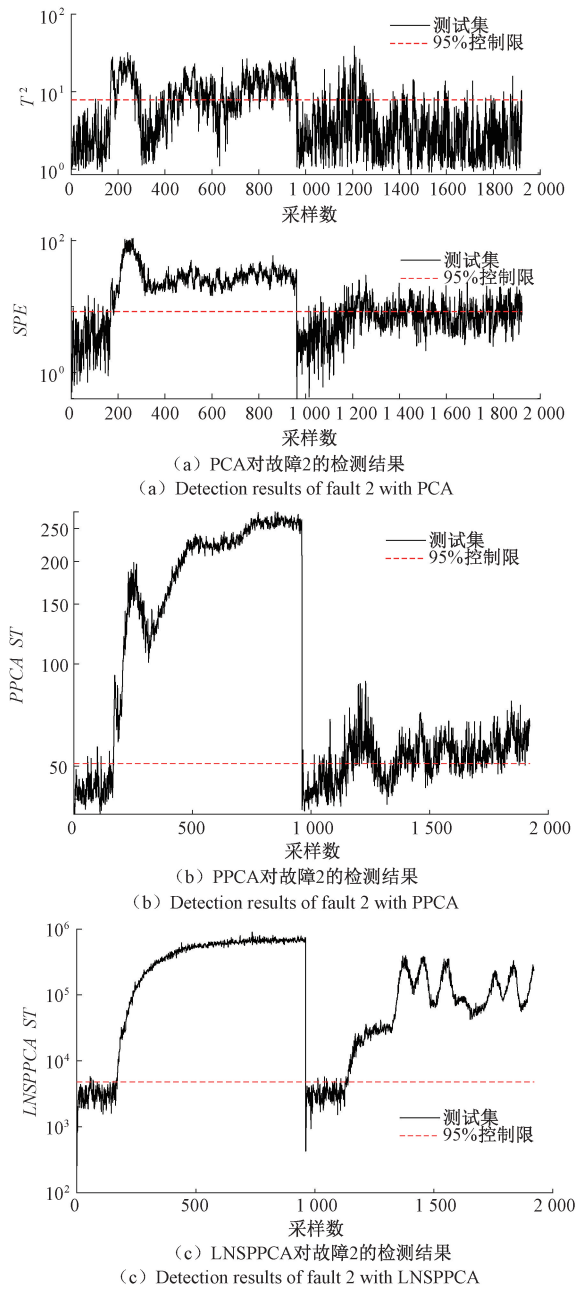


图 10 3 种方法对故障 2 的检测结果  
Fig. 10 Detection results of fault 2 with three methods

图 10(a)、(b) 可得知,使用 LNSPPCA 对故障进行检测可以有有效的处理多模态特性和提取更全面信息,比 PCA 和 PPCA 对故障检测更加准确。

故障 10 为物料 C 进料温度改变的一个随机干扰,此故障会引起物料 C 的进料温度的波动,如图 11(a)、(b) 所示,PPCA 从概率的角度对数据进行分析并建立概率模型,提取了更加全面有用的信息,其对故障的检测率优于 PCA。但两种方法均未准确的检测到故障的发生,在两种方法对数据采用 Z-score 处理后,多模态数据的中心不一致和方差差异特征并没有被很好的解决,使得这两

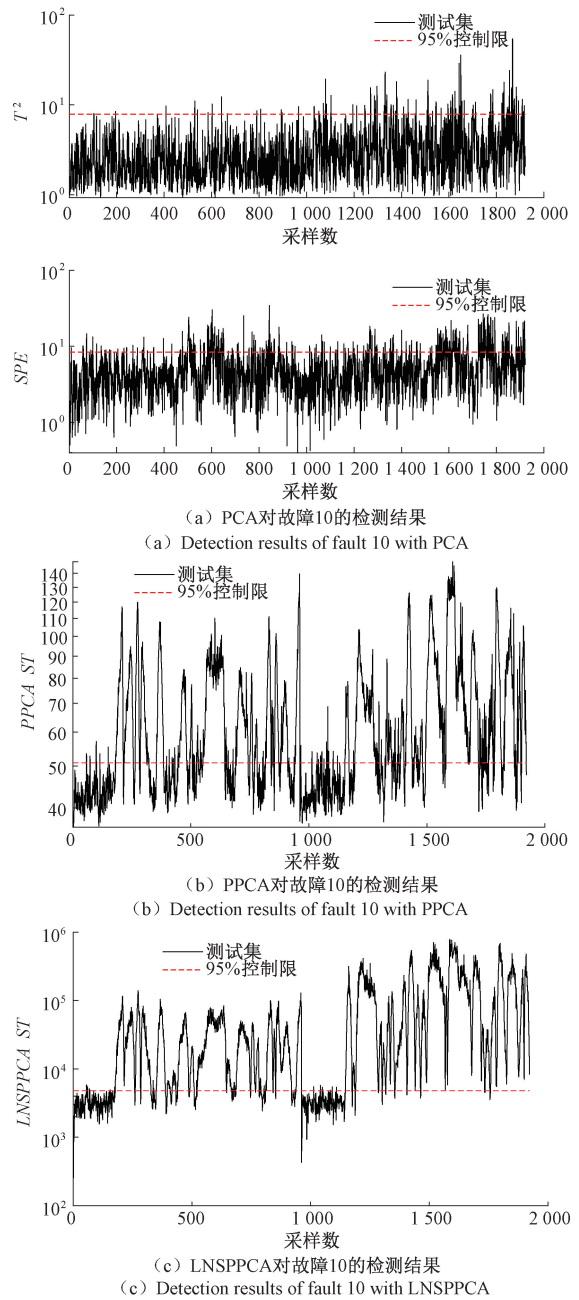


图 11 3 种方法对故障 10 的检测结果  
Fig. 11 Detection results of fault 10 with three methods



个模型描述的正常区域范围较预期偏大,因此 PCA 和 PPCA 的检测效果没有 LNSPPCA 方法效果明显, LNSPPCA 对故障 10 的检测结果如图 11(c) 所示。

故障 18 是一种未知故障,3 种方法对此故障的检测率如图 12 所示,由于 PCA 和 PPCA 使用 Z-score 方法处理的数据构建的控制限受到数据多模态特性的影响,如图 12(a)、(b) 所示,导致模态 1 中的故障都没有能够很好的被检测出来。而在 LNSPPCA 方法中,样本经过 LNS 的处理后,建立起概率模型,对数据的描述更加真实,提取有价值的信息更加全面,故障在第 161~961 采样点和

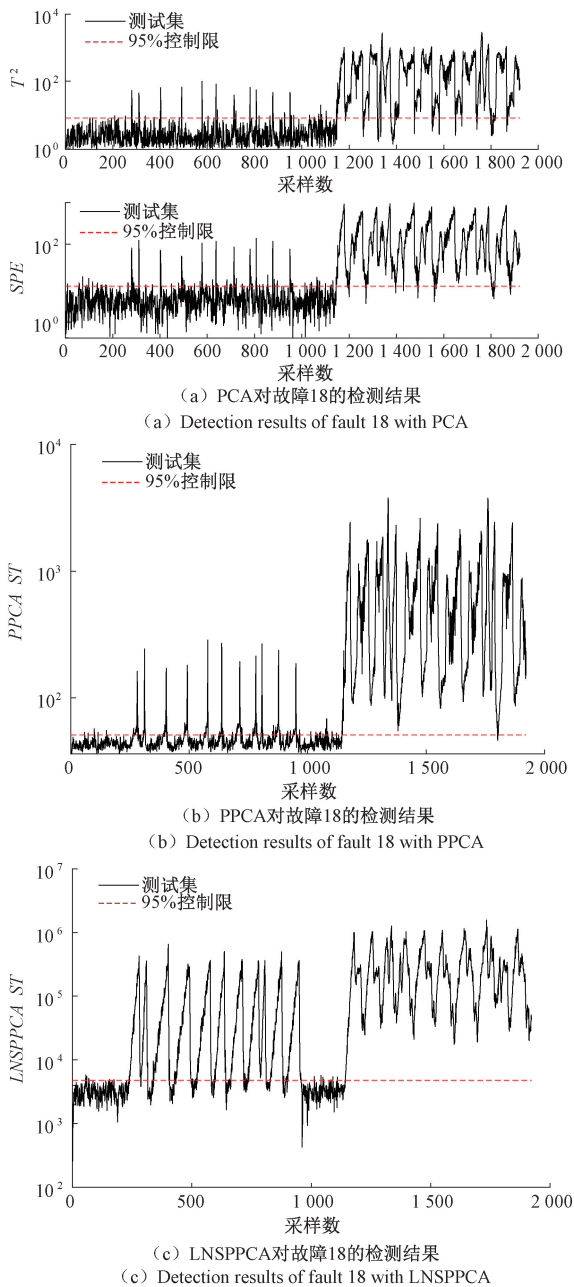


图 12 3 种方法对故障 18 的检测结果

Fig. 12 Detection results of fault 18 with three methods

第 1 122~1 922 采样点可以被准确的检测出来,因此 LNSPPCA 方法在上述故障中的检测率最高, LNSPPCA 对故障 18 的检测结果如图 12(c) 所示。

## 5 结 论

本文针对工业过程的数据分布复杂和多模态特性,提出了一种基于 LNSPPCA 故障检测方法。使用 LNS 对原始数据标准化处理,再使用 PPCA 从概率的角度对数据进行分析,相比于 PCA 和 PPCA 方法,该方法处理了多模态问题并考虑了过程数据的随机性,提取了更全面有用的信息用于故障检测,提高了故障检测能力。最后通过数值例子和 TE 多模态过程证明了基于 LNSPPCA 方法的故障检测的有效性与优越性。在本文研究的基础上,过程的动态特性可以被进一步研究来提高故障检测性能。

## 参考文献

- [1] 宋浏阳,李石,王芃鑫,等. 基于动态统计滤波与深度学习的智能故障诊断方法[J]. 仪器仪表学报,2019,40(7):39-46.  
SONG L Y, LI SH, WANG P X, et al. Intelligent fault diagnosis method based on dynamic statistical filtering and deep learning [J]. Chinese Journal of Scientific Instrument, 2019,40(7):39-46.
- [2] 周卫庆,司风琪,徐治皋,等. 基于 KPCA 残差方向梯度的故障检测方法及应用[J]. 仪器仪表学报,2017,38(10):2518-2524.  
ZHOU W Q, SI F Q, XU ZH G, et al. Fault detection method and application based on KPCA residual direction gradient [J]. Chinese Journal of Scientific Instrument, 2017,38(10):2518-2524.
- [3] GUO J, ZHONG L, LI Y. Fault detection based on difference locality preserving projections for the semiconductor process [J]. Journal of Chemometrics, 2018,32(1-3):e3035.
- [4] ZHANG J, CHEN H, CHEN S, et al. An improved mixture of probabilistic PCA for nonlinear data-driven process monitoring [J]. IEEE Transactions on Cybernetics, 2018,49(1):198-210.
- [5] TIPPING M E, BISHOP C M. Probabilistic principal component analysis [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999,61, DOI: 10.1111/1467-9868.00196.

- [6] 周乐. 基于概率的工业过程数据建模与故障检测[D]. 杭州:浙江大学, 2015.  
ZHOU L. Industrial process data modeling and fault detection based on probability [D]. Hangzhou: Zhejiang University, 2015.
- [7] HE Q P, WANG J. Fault detection using the K-nearest neighbor rule for semiconductor manufacturing processes[J]. IEEE Transactions on Semiconductor Manufacturing, 2007, 20(4): 345-354.
- [8] MA H, HU Y, SHI H. A novel local neighborhood standardization strategy and its application in fault detection of multimode processes[J]. Chemometrics & Intelligent Laboratory Systems, 2012, 118: 287-300.
- [9] YU J, QIN S J. Multimode process monitoring with Bayesian inference based finite Gaussian mixture models[J]. Aiche Journal, 2008, 54(7): 1811-1829.
- [10] YANG Y W, MA Y X, SONG B, et al. Multimodal chemical process fault detection based on coordinated hybrid probabilistic principal component analysis [J]. Chinese Journal of Chemical Engineering, 2015, 23(8): 1357-1363.
- [11] ZHOU L, ZHENG J, GE Z, et al. Multimode process monitoring based on switching autoregressive dynamic latent variable model [J]. IEEE Transactions on Industrial Electronics, 2018, 8184-8194.
- [12] 郭小萍, 杨猛, 李元. 基于改进重构贡献图的故障定位方法[J]. 仪器仪表学报, 2015, 36(5): 1193-1200.  
GUO X P, YANG M, LI Y. Fault location method based on improved reconstruction contribution graph [J]. Chinese Journal of Scientific Instrument, 2015, 36(5): 1193-1200.
- [13] ZENG J, LIU K, HUANG W, et al. Sparse probabilistic principal component analysis model for plant-wide process monitoring[J]. Korean Journal of Chemical Engineering, 2017, 34(8): 2135-2146.
- [14] LING J, LIU G J, LI J L, et al. Fault prediction method for nuclear power machinery based on Bayesian PPCA recurrent neural network model[J]. Nuclear Science and Techniques, 2020, 31(8): 15-25.
- [15] KIM D, LEE I B. Process monitoring based on probabilistic PCA [J]. Chemometrics & Intelligent Laboratory Systems 2003, 67(2): 109-123.
- [16] 杨沛武, 刘飞. PCA 和 PPCA 在化工分离过程监控中的应用比较[J]. 化工自动化及仪表, 2007(6): 7-11.  
YANG P W, LIU F. Application comparison of PCA and PPCA in chemical separation process monitoring [J]. Chemical Automation and Instrumentation, 2007(6): 7-11.
- [17] 李元, 杨东昇, 冯立伟, 等. 基于局部近邻标准化和主多项式算法的故障检测[J]. 山东科技大学学报: 自然科学版, 2020(3): 76-83.  
LI Y, YANG D SH, FENG L W, et al. Fault detection based on local nearest neighbor normalization and principal polynomial algorithm [J]. Journal of Shandong University of Science and Technology: Natural Science Edition, 2020(3): 76-83.
- [18] 肖应旺. 基于 WTPCA-MSVMs 过程监控方法[J]. 仪器仪表学报, 2010, 31(3): 558-564.  
XIAO Y W. Process monitoring method based on WTPCA-MSVMs [J]. Chinese Journal of Scientific Instrument, 2010, 31(3): 558-564.
- [19] 石怀涛, 周乾, 王雨桐, 等. 基于相对变换的 ICA 故障检测方法[J]. 电子测量与仪器学报, 2017, 31(7): 1040-1046.  
SHI H T, ZHOU Q, WANG Y T, et al. ICA fault detection method based on relative transformation [J]. Journal of Electronic Measurement and Instrument, 2017, 31(7): 1040-1046.
- [20] 郭小萍, 刘诗洋, 李元. 基于稀疏残差距离的多工况过程故障检测方法研究[J]. 自动化学报, 2019, 45(3): 175-183.  
GUO X P, LIU SH Y, LI Y. Research on fault detection method of multi working conditions based on sparse residual distance [J]. Journal of Automation, 2019, 45(3): 175-183.

### 作者简介



李元(通信作者), 1986年和2001年于沈阳化工大学分别获得学士学位和硕士学位, 2005年于东北大学获得博士学位, 现为沈阳化工大学信息工程学院教授, 主要研究方向为基于数据驱动技术的复杂过程故障检测与诊断。

E-mail: li-yuan@mail.tsinghua.edu.cn

LI Yuan (Corresponding author) received her B. Sc. and

M. Sc. degree both from Shenyang University of Chemical Technology in 1986 and 2001, respectively, received her Ph. D. degree in 2005 from Northeast University. Now, she is a professor in Institute of Information Engineering, Shenyang University of Chemical Technology. Her main research interest is fault detection and diagnosis of complex process based on data-driven technology.



**张昊展**, 2019 年于安阳工学院获得学士学位, 现为沈阳化工大学信息工程学院硕士研究生, 主要研究方向为基于概率统计模型的多模态故障检测。

E-mail: zhanghz0519@163.com

**Zhang Haozhan** received his B. Sc. degree in 2019 from Anyang Institute of Technology. Now, he is a master student in Institute of Information Engineering, Shenyang University of

Chemical Technology. His main research interest is multimodal fault detection based on probabilistic statistical model.



**唐晓初**, 2006 年和 2009 年于沈阳化工大学分别获得学士学位和硕士学位, 2015 年于东北大学获得博士学位, 现为沈阳航空航天大学自动化学院讲师, 主要研究方向为统计过程监控和复杂工业过程的质量预测。

E-mail: tangxc0420@126.com

**Tang Xiaochu** received her B. Sc. and M. Sc. degrees both from Shenyang University of Chemical Technology in 2006 and 2009, respectively, received her Ph. D. degree in 2015 from Northeast University. Now, she is a lecturer in Institute of Automation, Shenyang Aerospace University. Her main research interest includes statistical process monitoring and quality prediction of complex industrial processes.