北京科技大学
University of Science and Technology Beijing

自动化工程设计报告

题　　目：基于 PCA 及其改进方法的 TE 工业过程故障诊断

作　　者：袁浩天

学　　号：41918055

学　　院：高等工程师学院

班　　级：自 E191

指导教师：董洁

成　　绩：

2022 年　10 月

任务书

一、学生姓名：袁浩天　　学号：41918055

二、题目: 基于 PCA 及其改进方法的 TE 工业过程故障诊断

三、主要内容：

1. 通过查阅文献，了解故障诊断的发展概况及主要完成的任务，故障诊断方法分类，对比各种方法的应用背景，明确多元统计方法的优势；

2. 学习主成分分析（PCA）算法的基本原理，了解故障检测技术及基本框架；

3. 在理解田纳西-伊斯曼(Tennessee Eastman, TE)化工过程的运行机理的基础上，了解各变量及多种故障的含义，掌握过程数据训练集和测试集的使用方法；

4. 实现数据的主成分特征提取，建立基于 PCA 方法的过程监测模型，设计统计量与控制限，完成对 TE 过程典型故障的检测；

5. 针对 TE 过程的非线性特性，建立基于 PCA 改进方法的过程监测模型，设计统计量与控制限，完成对 TE 过程典型故障的检测；

6. 编程实现基于 PCA 及其改进方法的 TE 过程典型故障的检测并进行对比分析，给出结论。

四、主要（技术）要求：

1. 了解故障检测技术流程及基本框架；

2. 推导 PCA 及改进方法的算法数学公式，应用 PCA 方法对数据降维并选取最优主元、基于 PCA 的方法设计 $\boldsymbol{T}^2$ 和 $\boldsymbol{SPE}$ 监控统计量，建立故障检测模型，实现程序仿真；

3. 利用多种故障数据测试检测模型，对比分析结果，给出结论。

五、日程安排：

第 2 周　明确任务，了解过程监控系统的研究背景意义和设计需求；

第 3 周　学习 MATLAB 仿真软件，查阅文献，掌握 PCA 及改进方法的算法原理，了解 TE 过程的工作原理及数据集构成；

第 4-5 周　建立基于 PCA 及改进算法的过程监测模型，并初步编程实现算法用于 TE 过程故障检测；

第 6-7 周　算法完善、软件测试和系统仿真；

第 8 周　撰写课程设计报告，准备答辩。

六、主要参考文献和书目：

[1] 彭开香, 马亮, 张凯. 复杂工业过程质量相关的故障检测与诊断技术综述[J]. 自动化学报. 2017，43(3)：349-365.

[2]姚羽曼, 罗文嘉, 戴一阳. 数据驱动方法在化工过程故障诊断中的研究进展[J].化工进展. 2021, 355(04):1755-1764.

[3] Jiang Q, Yan X, Huang B. Performance-Driven Distributed PCA Process Monitoring Based on Fault-Relevant Variable Selection and Bayesian Inference[J]. IEEE Trans. Industrial Electronics, 2016, 63(1): 377-386.

[4] 张凯林. 基于主元分析和偏最小二乘的 TE 过程监测方法的研究[D].天津理工大学, 2015.

[5] 胡静. 基于多元统计分析的故障诊断与质量监测研究[D]. 浙江大学, 2015.

[6] 陈永禄. 基于多元统计方法的田纳西化工过程故障诊断[D]. 东北大学, 2011..

指导教师签字：　　年　月　日

学 生 签 字：　　年　月　日

目　　录

目　　录 I
1 文献综述 1
1.1 引言 1
1.2 故障诊断概述 1
1.2.1 基于解析模型的方法 2
1.2.2 基于知识的方法 3
1.2.3 基于信号驱动的方法 3
1.3 国内外研究现状 3
1.4 研究内容及结构安排 4
2 PCA 及 KPCA 算法原理 6
2.1 PCA 算法原理 6
2.2 KPCA 算法原理 7
2.3 T^2统计量及控制限的确定 8
2.4 SPE 统计量及控制限的确定 8
3 仿真实验与对比分析 9
3.1 TE 过程简介 9
3.2 检测结果分析 12
3.2.1 基于 PCA 的检测结果 12
3.2.2 基于 KPCA 的检测结果 14
3.3 对比分析 15
3.4 对 KPCA 检测结果进行 hampel 数据滤波改进 16
3.4.1 hampel 数据滤波原理 16
3.4.2 应用 hampel 数据滤波对 TE 过程实现预处理 16

4 总结..... 18

参 考 文 献..... 19

附录..... 20

指导教师意见..... 26

1 文献综述

1.1 引言

对于复杂多变的化工生产流程，生产装置的复杂化、模块化，以及原材料易燃易爆炸、有毒有害的特点，使得在化工生产流程中存在众多危险因素，这些危险因素在一定条件下可能转变为重大故障，进而引发严重的生产事故，不仅破坏正常的生产流程，甚至会危及人们的生命安全，并且带来严重的环境污染，造成重大的经济损失。由此可见，对化工生产过程进行过程监测，对系统中出现的故障进行诊断，提高系统安全系数，降低事故风险，对化工生产流程十分重要。

主元分析法（principal component analysis）即 PCA，一种利用统计原理建立描述系统的低维模型的方法，经过十多年的研究与发展，成功地应用于过程的监测与分析。然而基本的 PCA 处理是线性的、静态的，随着进一步的发展动态和非线性的改进方法也已经被提出，比如核主元分析法 KPCA（kernel principal component analysis）是一种很有效的非线性过程故障诊断方法。

TE（Tennessee Eastman）过程是由美国 Eastman 化学公司的 Downs 和 Vogel 提出的用来开发、研究和评价过程技术和监控方法的现实化工模型。许多国内外学者、专家均引用它作为数据源，以进行控制、优化和故障诊断等研究[1]。本文通过 TE 过程的应用实例，比较 PCA 及其改进方法应用于 TE 过程的诊断效果。

此次研究主要是学习 PCA 及其改进方法的基本原理，利用 MATLAB 建立基于 PCA 及其改进方法的过程检测模型，了解 TE 过程的工作原理及数据集构成，完成对 TE 过程的故障诊断。因此，本设计具有理论研究意义，并且对现代化工过程具有指导意义。

1.2 故障诊断概述

所谓故障诊断，是指由计算机利用特定方法，完成工况分析，对生产是否正常什么原因引起故障、故障的程度有多大等问题进行分析、判断，得出结论的过程。具体包括故障检测、故障分离、故障评价、故障决策四个方面。

评价一个故障诊断系统的性能指标主要有:故障检测的及时性，故障检测的灵敏度，故障的误报率和漏报率，故障定位和评价的准确性，故障决策的正确性和及时性，故障诊断系统的鲁棒性。图 1 为故障诊断的过程示意图:

按照国际故障诊断权威德国的 P. M. Frank 教授的观点，故障诊断方法可以划分为基于解析模型的方法、基于知识的方法、基于信号处理的方法 3 种[8]。

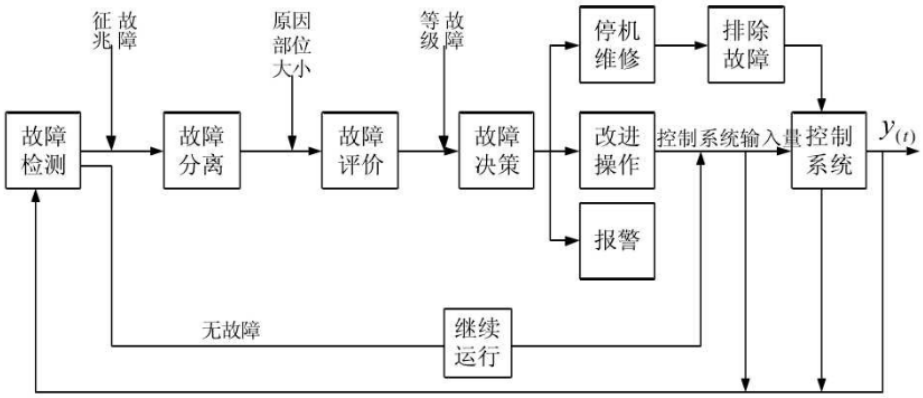


图 1 故障诊断过程示意图

1.2.1 基于解析模型的方法

基于解析模型的故障诊断方法又称为基于深层知识的诊断方法，它是指使用系统的结构、行为和功能等方面的知识对系统进行诊断推理，这就需要建立系统结构、行为和功能模型。简单的说，此法一般利用构造出来的观测器估计预测出系统的输出值或过程变量的估计值，再将估计值与实际值比较产生残差。在系统正常运行时，此残差应该是零值或接近零值的数，当有故障发生时残差量将明显偏离零值，超出容许范围。这种方法具体应用有三种方式:基于状态估计的故障诊断方法、基于参数估计的故障诊断方法和等价空间故障诊断方法。

(1)基于状态估计的故障诊断方法是利用系统的解析模型和可测信息，设计检测观测器，重构系统的某个可测变量，然后由滤波器的输出与真实系统的输出构造残差，再对残差进行分析处理，以实现系统的故障诊断。在能获得系统的精确数学模型的情况下，状态估计方法是最直接有效的故障诊断方法，然而在实际操作中，控制系统对象模型很难获得，所以目前的状态估计方法用于故障诊断的研究主要集中于线性系统，对非线性系统的研究成果还比较少。

(2)等价空间方法的基本思想就是通过系统的输入、输出（或部分输出）的实际值检测被诊断对象数学关系的等价一致性，从而达到检测和分离故障的目的。基于系统的动态方程产生具有方向性残差的方法，用动态等价方程产生残差序列，再利用等价方程中参数留下的自由度进行重新设计，使得残差序列对故障具有特定的方向性，因此此法更利于故障的分离。

(3)参数估计方法根据模型参数及相应的物理参数的变化来检测和分离故障。其基本思想是许多被诊断对象的故障可以看作是其过程系数的变化，而这些过程系数的变化又往往导致系统参数的变化。因此，可以根据系统参数及相应的过程系数变化来检测故障。基于系统参数估计的故障诊断方法主要有滤波器方法和最小二乘法。与状态估计法相比，参数估计法更利于故障的分离。

1.2.2 基于知识的方法

(1)基于神经网络的故障诊断方法

将神经网络技术应用于故障诊断，一方面依赖于新型高效的神经网络和学习算法以及神经网络硬件实现的发展，另一方面如何将系统知识和诊断知识融入到神经网络的设计和诊断算法中也是当前一个明显的研究趋势。

(2)基于模糊逻辑的故障诊断方法

基于模糊信息处理的方法应用到控制系统故障诊断中的优点体现在:模糊逻辑在概念上易于理解，在表达上接近人的自然思维，从而使人的故障诊断知识能很容易地通过模糊逻辑的方式表达和应用;具有 T-S 形式的模糊模型和神经网络一样具有对任意非线性的逼近能力，而且其后件为线性模型，这为非线性问题的解决提供了一条将定量和定性知识集成在一起的方式。

(3)专家系统法

专家系统应用到故障诊断领域一般是使用专家知识由推理机直接根据故障征兆推理诊断出故障原因等结果。专家系统通常由三个部分构成:1),数据库:它是专家系统的主要的数据结构，存储与求解问题有关的已知的或推导出的数据;2),知识库:存储与求解问题有关的特殊专家知识;3),推理机:它的任务是选择最合适的控制或推理步骤，从而实现问题的求解过程。专家系统能够不依赖于数学模型,以模拟专家思维的方式，进行过去只有专家才能完成的高级任务。

(4)基于定性模型的方法

基于定性模型的故障诊断方法近年来在欧洲受到了高度重视，得到了迅猛发展。定性仿真是基于定性模型的故障诊断方法的重要部分，它用表示系统物理参数的定性变量和表示各参数间相互关系的定性微分方程构成约束模型，描述并模仿系统的结构，以确定从给定的初始状态出发得到的系统状态。

1.2.3 基于信号驱动的方法

基于信号驱动的方法通过利用信号模型，如相关函数、频谱、自回归滑动平均等，直接分析测量信号，提取诸如方差、幅值、频率等特征值，从而检测故障的发生。一般的有基于信息融合，基于小波变换和基于多变量统计分析模型三大类方法。其中，基于多变量统计分析模型的故障诊断是一类特别有效的方法，主要有主元分析 PCA(Principle Component Analysis)、部分最小二乘法 PLS (Partial Least Squares)、Fisher 判别分析法 FDA (Fisher Discriminant Analysis)。其中，PCA 及 PLS 都是处理高维相关数据的有效手段，用于对含有噪声的和高度相关的测量数据进行分析，并将高维信息压缩到低维子空间，而保留了主要的过程信息。

1.3 国内外研究现状

目前，应用于 TE 过程的故障诊断方法越来越成熟，主要集中在多变量统计

分析、神经网络、支持向量机等方法。而核学习方法也得到了广泛应用，比如核主元分析方法，另外还有核部分最小二乘、核费舍尔判别分析法、核规范变量分析法。

主元分析法（PCA）已经广泛应用到了 TE 过程中，它是由 Peatson 于 1901 年提出的一种线性方法，1933 年 Harold Hotelling 使主元分析更加完善。

何菲[3]等人提出了一种主元分析（PCA）和支持向量机（SVM）的故障诊断给方法。首先用 PCA 对 TE 过程数据进行降维和故障检测,然后用 SVM 方法对处理后的数据进行分类和识别。结果表明，该方法要比 PCA-KNN 方法和 C-SVM 方法简单且易实现，并且有较高的多分类准确率。

缪素云[4]等人在论文“基于概率神经网络的 TE 过程故障诊断”中提出了一种基于主元分析的故障检测方法，并结合概率神经网络对化工过程进行故障诊断。首先用 PCA 对数据进行了降维，然后将处理过的数据作为网络的输入，进行故障诊断。

但是主元分析是一种线性降维技术，并且主元分析方法是首先假设数据是服从正态分布的，而一般实际的过程都是非线性的，不符合这一假设。对于 TE 过程来说，也是一个非线性的过程，不符合这一要求，所以用主元分析的效果不是十分的好，这导致在实际应用中的效率降低。所以，一些学者对主元分析进行了改进。

schölkopf 等学者提出了核主元分析（KPCA）的概念，核主元分析核一些改进的核主元分析也被广泛的应用到了 TE 过程中。

赵小强[5]等人在论文“基于 TE 的化工过程故障诊断算法研究”中提出了一种基于贡献图的故障诊断方法,并且结合小波变换对其进行数据处理。首先用小波变换对数据去噪，对处理后的数据用核主元分析贡献图进行故障诊断，识别出是哪发生了故障。

许洁[6]等人在论文“基于 KPCA 和 MKL-SVM 的非线性过程监控与故障诊断”提出了一种基于核主元分析(KPCA)和多重核学习支持向量机(MKL-SVM)的非线性故障故障诊断方法。首先用核主元分析(KPCA)对数据进行处理，若发生故障，则将核主元作为多重学习支持向量机的输入，进行故障类型识别。结果表明，该方法不但有效地辨识出故障，而且提高了故障诊断的速度。

Deng[7]等人将深度学习与 KPCA 算法融合，提出了深度主成分分析法（DePCA)，以增强 PCA 方法的非线性特征提取能力并改善计算复杂度。

1.4 研究内容及结构安排

本设计针对复杂化工过程，基于 PCA 算法及其改进算法 KPCA，在 MATLAB 中建立基于 PCA 及其改进方法的过程检测模型，了解 TE 过程的工作原理及数

据集构成，完成对 TE 过程的故障诊断，并将两种算法检测结果进行对比，得出结论。

本文的具体结构安排如下：

第一章简述故障检测与诊断的课题背景和研究意义，以及国内外利用 PCA 和 KPCA 在故障检测方面的发展情况。

第二章详细介绍 PCA 及 KPCA 算法的基本原理及其应用于故障检测时统计量和置信限的确定。

第三章以 TE 过程为仿真对象，分析说明 PCA 应用于故障检测的整个流程，并与 KPCA 方法进行对比，给出结论。

第四章对本设计内容进行总结，提出研究中存在的问题及未来研究展望。

2 PCA 及 KPCA 算法原理

基于 PCA 及其改进方法的过程监测和故障诊断方法是利用过程变量间的相关关系，建立正常工况下的主元模型，通过检验新的数据样本相对于主元模型的背离程度，从而发现异常和故障。本文正是针对这种方法进行研究的，将在以下章节详细讨论说明 PCA 及 KPCA 的算法基本原理及统计量和控制限的确定。

2.1 PCA 算法原理

PCA 的基本方法是构造原变量的线性组合，产生一系列线性无关的新变量，使它们含有原始数据尽可能多的信息，并达到线性降维的目的。

首先我们给定一个数据集，假设它含有 m 个观测值、n 个过程变量，写成一个(m×n)维矩阵如下：

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{1}$$

对原始数据做标准化处理：即对每一个指标分量进行标准化处理：

$$X_{ij} = \frac{A_{ij} - \bar{A}_j}{S_j} \tag{2}$$

其中样本均值为：

$$\bar{A}_j = \frac{1}{m}\sum_{i=1}^{m} A_{ij} \tag{3}$$

样本标准差为：

$$S_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(A_{ij} - \bar{A}_j)^2} \tag{4}$$

从而得到标准化后的样本矩阵：

$$X = (x_{ij})_{m\times n} \tag{5}$$

随后计算样本矩阵的相关系数矩阵：

$$R = \frac{1}{m-1}X^T \cdot X = (r_{ij})_{n\times n} \tag{6}$$

运用 Jacobi 迭代方法计算 R 的特征值$\lambda_1,\ldots,\lambda_n$，即对应的特征向量$v_1,\ldots,v_n$。并通过特征值按降序排序得$\lambda_1' > \ldots > \lambda_n'$并对特征向量进行相应调整得$v_1',\ldots,v_n'$。随后通过施密特正交化方法单位正交化特征向量，得到$\alpha_1,\ldots,\alpha_n$。

计算特征值的累积贡献率$B_1,\ldots,B_n$，根据给定的提取效率p，如果$B_t \geq p$,则

提取t个主成分$\alpha_1,\ldots,\alpha_t$。

2.2 KPCA 算法原理

从具体操作过程上看，核方法首先采用非线性映射将原始数据由数据空间映射到特征空间，进而在特征空间进行对应的线性操作，如图 2 所示

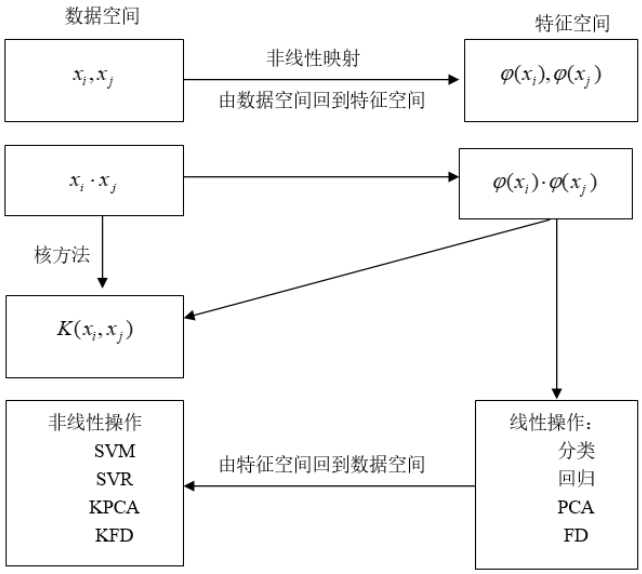


图 2　核方法框架示意图

由于采用了非线性映射，且这种非线性映射往往是比较复杂的，从而大大增强了非线性数据的处理能力。

从本质上讲，核方法实现了数据空间、特征空间、和类别空间之间的非线性变换。设x_i和x_j为数据空间中的样本点，数据空间到特征空间的映射函数为Φ，核函数的基础是实现向量的内积变换

$$(x_i, x_j) \to K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{7}$$

KPCA 处理数据的整个流程如下：

首先与 PCA 由式（1）得到数据矩阵 A。

选定核函数,本文选用高斯镜像基函数（RBF）核函数:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\sigma^2}\right) \tag{8}$$

计算核函数矩阵

$$K_{\mu\nu} := \left(\Phi(x_\mu) \cdot \Phi(x_\nu)\right) \tag{9}$$

并修正核函数矩阵得到 KL：

$$KL \to K_{\mu\nu} - \frac{1}{M}\left(\sum_{w=1}^{M} K_{\mu w} + \sum_{w=1}^{M} K_{w\nu}\right) + \frac{1}{M^2}\sum_{w,\tau=1}^{M} K_{w\tau} \tag{10}$$

其余步骤与 PCA 得到特征向量相同。

2.3 T^2统计量及控制限的确定

T^2统计量反映的是各个变量在主元子空间中的变化量，是变量在主元子空间内的投影大小，其数学表达式为：

$$T^2 = x^T P \Lambda^{-1} P^T x \le T_\alpha^2 \tag{11}$$

式中$\Lambda = diag\{\lambda_1 \lambda_2 \ldots\ldots \lambda_A\}$，$T_\alpha^2$表示在置信度为$\alpha$时的$T^2$控制限，其数据样本服从正态分布，表达式为：

$$T_\alpha^2 = \frac{A(n^2-1)}{n(n-A)} F_{(A,n-A;\alpha)} \tag{12}$$

式中，$F_{(A,n-A;\alpha)}$表示带有 A 和 n-A 个自由度，置信度为 α 的 F 分布的临界值。

2.4 SPE 统计量及控制限的确定

SPE （squared prediction error，平方预测误差）统计量（也称 Q 统计量)可体现数据样本向量在残差子空间内投影的变化，即

$$SPE = ||(I - PP^T)x||^2 \le \delta_\alpha^2 \tag{13}$$

式中，δ_α^2表示的是置信水平为α时的控制限。正常生产工况是，SPE 处于控制限内，否则将超出控制限。控制限δ_α^2为：

$$\delta_\alpha^2 = \left(\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 \sqrt{(h_0-1)}}{\theta_1^2} \right)^{\frac{1}{h_0}} \tag{14}$$

式中$\theta_1 = \sum_{j=A+1}^{M} \lambda_j^i (i = 1,2,3), h_0 = 1 - 2\theta_1\theta_3/3\theta_1^2$，$\lambda_i$表示数据样本协方差矩阵的第 i 个特征值。

3 仿真实验与对比分析

本文使用 MATLAB 编写调试代码，搭建基于 PCA 和 KPCA 的故障检测模型，详细代码见附录。采用 TE 平台提供的数据集，对模型进行训练。然后，对测试集进行过程监测，并比较它们的故障检测效果，得出结论。

3.1 TE 过程简介

田纳西-伊斯曼过程(Tennessee-Eastman Process，TEP)是一个现实的工业过程[2]，为的是评价过程控制中的故障检测和诊断方法。TE 过程包括五个主要单元:反应器、冷凝器、压缩机、分离器和汽提塔,如图 1 所示.气体成分 A、C、D 和 E 以及惰性组分 B 被喂入反应器,液态产物 G 和 H 在反应器中形成.反应器的产品流通过冷凝器冷却,然后送入到汽液分离器.从分离器出来的蒸汽通过压缩机再循环送入反应器，为了防止过程中惰性组分和反应副产品的积聚,必须排放一部分再循环流。来自分离器的冷凝成分被泵送到汽提塔.剩余反应物与再循环流结合.从汽提塔底部出来的产品 G 和 H 被送到下游过程.副产品主要在汽液分离器中以气体的形式从系统中排空。

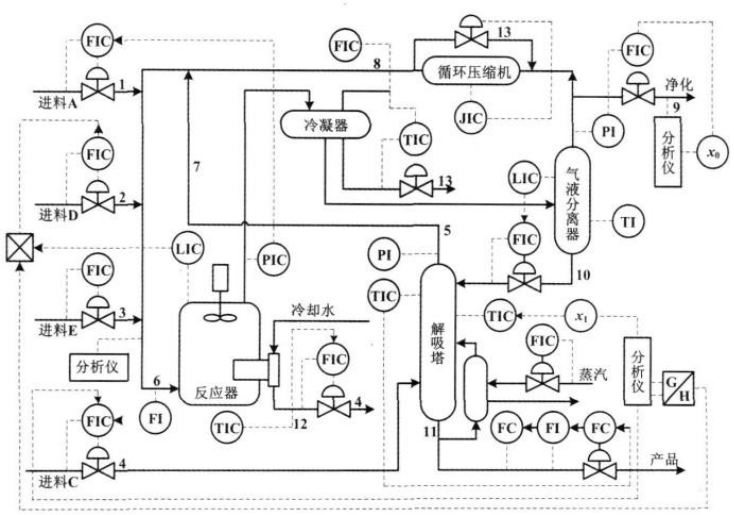


图 3　TE 过程流程图

在反应器中喂入气体成分 ACDE 以及惰性组分 B，则在反应器中形成了液态产物 G 和 H。反应器的各种反应是：

$$A(g) + C(g) + D(g) \rightarrow G(liq), \quad (16)$$

$$A(g) + C(g) + E(g) \rightarrow H(liq), \quad (17)$$

$$A(g) + E(g) \rightarrow F(liq), \quad (18)$$

$$3D(g) \rightarrow 2F(liq) \quad (19)$$

其中，物质 F 是反应的副产品。反应是不可逆的、放热的，并且对反应物

浓度来说，近似为一阶的。反应速度是温度的 Arrhenius 函数，其中生成 G 的反应要比生成 H 的反应有更高的活化能，致使对温度具有更高的灵敏度。

TE 过程是一个大样本的复杂非线性化工系统，它包括 21 种预先设定好的故障，代表阶跃、随机变化、慢漂移、粘滞和恒定位置等故障类型，如表 1 所示。

表 1　TE 过程故障

故障编号	故障描述	类型
1	A/C 进料流量比变化,组分 B 含量保持不变	阶跃
2	组分 B 含量发生变化，A/C 进料流量比不变	阶跃
3	物料 D 的温度发生变化	阶跃
4	反应器冷却水入口温度发生变化	阶跃
5	冷凝器冷却水入口温度发生变化	阶跃
6	物料 A 损失	阶跃
7	物料 C 压力损失	阶跃
8	物料 A、B、C 的组成发生变化	随机
9	物料 D 的温度发生变化	随机
10	物料 C 的温度发生变化	随机
11	反应器冷却水入口温度发生变化	随机
12	冷凝器冷却水入口温度发生变化	随机
13	反应动力学特性发生变化	慢漂移
14	反应器冷却水阀门	粘滞
15	冷凝器冷却水阀门	粘滞
16~20	未知	未知
21	阀门固定在稳态位置	恒定位置

TE 数据集由训练集和测试集构成，两种数据集都包括 1 种正常状态和 21 种故障状态的监测值。带有故障的训练集包括 480 个样本，测试集包括 960 个样本，前 160 个样本为正常数据，后 800 个样本为引入故障之后的数据。

TE 过程的所有变量，包括 11 个操纵变量和 41 个测量变量，其中 11 个操纵变量如下表 2 所示：

表 2　TE 中的 11 种操纵变量

编号	变量名称	单位
1	D 的入流量	kg/h
2	E 的入流量	kg/h
3	A 的入流量	kscmh
4	A 和 C 的入流量	kscmh

5	压缩机循环阀	%
6	净化阀	%
7	分离池液体流量	m^3/h
8	汽提塔液体流量	m^3/h
9	汽提塔蒸汽阀	%
10	反应池冷却水流量	m^3/h
11	压缩机冷却水流量	m^3/h

41 个测量变量如下表 3 所示：

表 3　TE 中的 41 种操纵变量

编号	变量名称	单位
1	A 蒸汽流量	Km^3/h
2	D 蒸汽流量	kg/h
3	E 蒸汽流量	kg/h
4	A 和 C 流量	km^3/h
5	回收循环流量	km^3/h
6	反应池入料速率	km^3/h
7	反应池压力	kPa
8	反应池液位	%
9	反应池温度	℃
10	净化速率	km3/h
11	分离池温度	℃
12	分离池液位	%
13	分离池压力	kPa
14	分离池潜流量	km^3/h
15	汽提塔液位	%
16	汽提塔压力	kPa
17	汽提塔潜流量	km^3/h
18	汽提塔温度	℃
19	汽提塔蒸汽流量	m^3/h
20	压缩机功率	kw
21	反应池水温度	℃
22	分离池水温度	℃
23	反应池 A 的含量	mo1%
24	反应池 B 的含量	mo1%
25	反应池 C 的含量	mo1%
26	反应池 D 的含量	mo1%
27	反应池 E 的含量	mo1%
28	反应池 F 的含量	mo1%
29	净化气体中 A 的含量	mo1%
30	净化气体中 B 的含量	mo1%
31	净化气体中 C 的含量	mo1%
32	净化气体中 D 的含量	mo1%
33	净化气体中 E 的含量	mo1%
34	净化气体中 F 的含量	mo1%
35	净化气体中 G 的含量	mo1%
36	净化气体中 H 的含量	mo1%
37	产物中 D 的含量	mo1%
38	产物中 E 的含量	mo1%
39	产物中 F 的含量	mo1%
40	产物中 G 的含量	mo1%
41	产物中 H 的含量	mo1%

3.2 检测结果分析

引入故障检测率(FDR)和故障误报率(FAR)来分析检测结果，是检测结果更加直观。

故障检测率 FDR 的定义为：

$$\mathrm{FDR}=\frac{\mathrm{n_1}}{\mathrm{N_1}}\times 100\% \tag{20}$$

式中，n_1是测试集中实际故障数据被检测为故障数据的数目，$\mathrm{N_1}$为是测试集中实际故障数据的数目。

故障误报率 FAR 的定义为：

$$\mathrm{FAR}=\frac{\mathrm{n_0}}{\mathrm{N_0}}\times 100\% \tag{21}$$

式中，$\mathrm{n_1}$是测试集中正常数据被检测为故障数据的数目，$\mathrm{N_1}$为是测试集中正常运行数据的数目。

3.2.1 基于 PCA 的检测结果

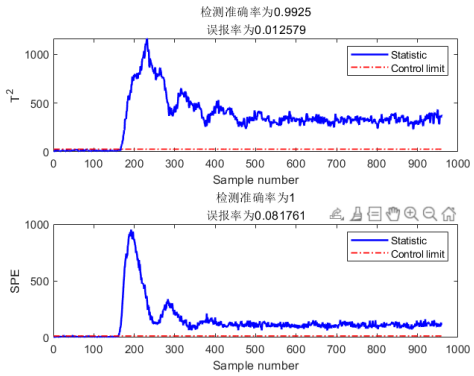
在基于 PCA 的故障检测中，根据一般工业过程的现实依据设置累计贡献率设为 85%来确定主元个数，设置 T^2 统计量的置信度为 0.99。

PCA 的故障检测结果如下表 2 所示：

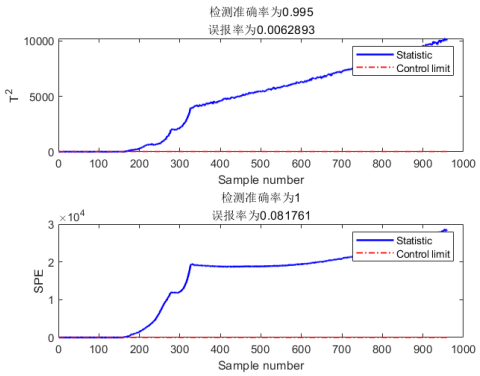
表 4　PCA 故障检测结果

故障编号	检测率	误报率	故障编号	检测率	误报率
1	100%	8.18%	12	98.75%	9.43%
2	98.38%	7.55%	13	95.63%	8.18%
3	17.38%	13.69%	14	100%	6.92%
4	100%	11.32%	15	20.25%	13.84%
5	29.63%	11.32%	16	61%	25.16%
6	100%	8.18%	17	97.13%	16.98%
7	100%	8.18%	18	92%	18.23%
8	97.28%	6.29%	19	7.88%	10.69%
9	18.13%	16.35%	20	69%	10.06%
10	65.38%	11.95%	21	67.38%	19.5%
11	86.38%	11.32%			

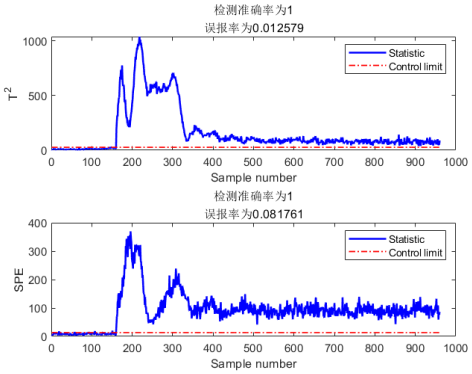
由结果可知：PCA 对故障 1、6、7、14 的检测效果较好，检测率达到 100%，并且误报率相对较低，它们的故障检测结果图如下图 4 所示：



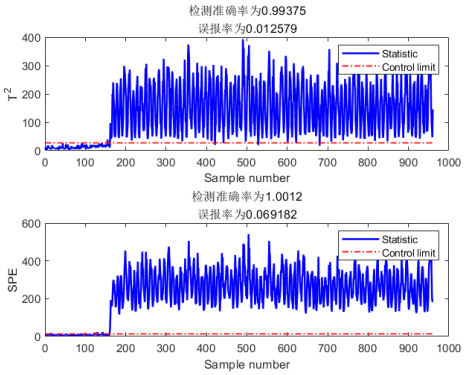
A 故障 1 的检测结果



B 故障 6 的检测结果



C 故障 7 的检测结果



D 故障 14 的检测结果

图 4　PCA 对故障 1、6、7、14 的检测结果

对 PCA 的诊断结果分析可以得出：

（1） 该模型对故障 1（A/C 进料流量比变化，组分 B 含量保持不变（流 4））、故障 2（组分 B 含量发生变化，A/C 进料流量比不变（流 4））、故障 6（物料 A 损失（流 1））、故障 7（物料 C 压力损失（流 4））、故障 8（物料 A,B,C 的组成发生变化（流 4））、故障 12（冷凝器冷却水入口温度发生变化）、故障 13（反应动力学特性发生变化）、故障 14（反应器冷却水阀门）的检测率都在 90%以上，且误报率都在 10%以下，检测效果较好。

（2） 该模型对故障 4（反应器冷却水入口温度发生变化）、故障 17、故障 18 的检测率都在 90%以上，但是误报率都在 10%以上，检测效果一般。

（3） 该模型对故障 10（物料 C 的温度发生变化（流 2））、故障 11（反应器冷却水入口温度发生变化）、故障 16、故障 20、故障 21（流 4 的阀门固定在稳态位置）的检测率在 60%~90%之间，检测效果不好。

（4） 该模型对故障 3（物料 D 的温度发生变化（流 2））、故障 5（冷凝器冷却水入口温度发生变化）、故障 9（物料 D 的温度发生变化（流 2））、故障 15（冷凝器冷却水阀门粘滞）、故障 19 的检测率在 60%以下，检测效果很差。

3.2.2 基于 KPCA 的检测结果

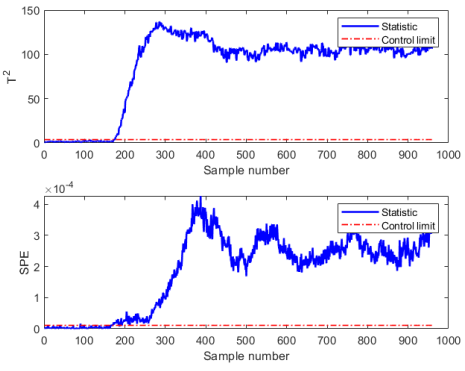
KPCA 作为 PCA 的改进算法，解决了 PCA 对非线性数据处理不佳的问题。在基于 KPCA 的故障检测中，根据一般工业过程的现实依据设置累计贡献率设为 95%来确定主元个数，设置核函数的核宽为 10000，设置 T^2 统计量的置信度为 0.99。

KPCA 的故障检测结果如下表 3 所示：

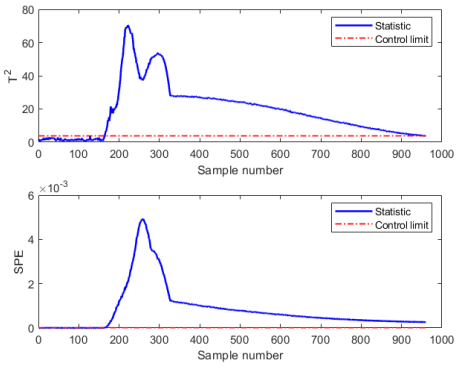
表 5　KPCA 故障检测结果

故障编号	检测率	误报率	故障编号	检测率	误报率
1	100%	6.88%	12	99.75%	25%
2	99.63%	5.63%	13	96%	3.13%
3	32.63%	3%	14	100%	6.25%
4	99.75%	6.25%	15	30.75%	4.38%
5	45%	6.23%	16	59.13%	52.5%
6	99.75%	3.13%	17	98.5%	14.38%
7	100%	6.25%	18	92.5%	11.88%
8	99.5%	3.75%	19	33.63%	11.88%
9	27.75%	11.25%	20	69.13%	4.38%
10	63.63%	8.13%	21	73.38%	24.38%
11	77.13%	14.38%			

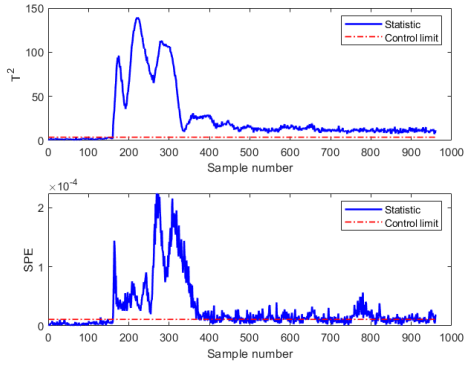
由结果可知：KPCA 对故障 2、4、7、14 的检测效果较好，它们的故障检测结果图如下图 5 所示：



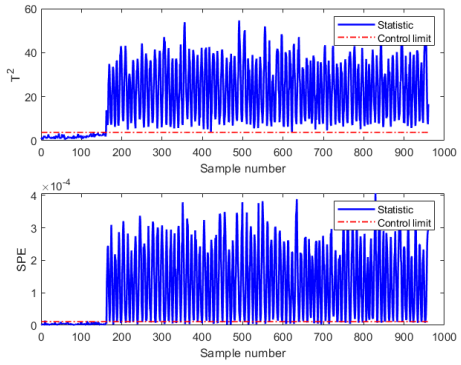
A 故障 2 的检测结果



B 故障 4 的检测结果



C 故障 7 的检测结果



D 故障 14 的检测结果

图 5　KPCA 对故障 2、4、7、14 的检测结果

对 KPCA 的诊断结果分析可以得出：

（1） 该模型对故障 1（A/C 进料流量比变化，组分 B 含量保持不变（流 4））、故障 2（组分 B 含量发生变化，A/C 进料流量比不变（流 4））、故障 4（反应器冷却水入口温度发生变化）、故障 6（物料 A 损失（流 1））、故障 7（物料 C 压力损失（流 4））、故障 8（物料 A,B,C 的组成发生变化（流 4））、故障 13（反应动力学特性发生变化）、故障 14（反应器冷却水阀门）的检测率都在 90%以上，且误报率都在 10%以下，检测效果较好。

（2） 该模型对故障 12（冷凝器冷却水入口温度发生变化）、故障 17、故障 18 的检测率都在 90%以上，但是误报率都在 10%以上，检测效果一般。

（3） 该模型对故障 10（物料 C 的温度发生变化（流 2））、故障 11（反应器冷却水入口温度发生变化）、故障 20、故障 21（流 4 的阀门固定在稳态位置）的检测率在 60%~90%之间，检测效果不好。

（4） 该模型对故障 3（物料 D 的温度发生变化（流 2））、故障 5（冷凝器冷却水入口温度发生变化）、故障 9（物料 D 的温度发生变化（流 2））、故障 15（冷凝器冷却水阀门粘滞）、故障 16、故障 19 的检测率在 60%以下，检测效果很差。

3.3 对比分析

计算得到 PCA 模型对 TE 过程 21 种故障的平均检测率为 72.46%；KPCA 模型对 TE 过程 21 种故障的平均检测率为 76.07%，KPCA 模型的故障检测率略高于 PCA 模型。

PCA 模型对 TE 过程 21 种故障的平均误报率为 11.92%；KPCA 模型对 TE 过程 21 种故障的平均检测率为 12.38%，KPCA 模型的故障检测率略也高于 PCA 模型。

对于一些难检测的故障，KPCA 模型明显优于 PCA 模型，举例如下表：

表 6　PCA 与 KPCA 对于难检测故障的诊断对比

故障号	3	5	9	15	19
PCA	17.38%	29.63%	18.13%	20.25%	7.88%
KPCA	32.63%	45%	27.75%	30.75%	33.63%

3.4 对 KPCA 检测结果进行 hampel 数据滤波改进

由上述结果可知：虽然 KPCA 模型对 TE 过程的平均检测率比 PCA 模型的要高，但是 KPCA 模型的平均误报率也比 PCA 模型的要高一点。

通过查阅资料，了解到可以对 TE 过程数据集做数据预处理，从而降低误报率[9] 。在本文中，选择 hampel 滤波对 TE 过程测试集进行数据预处理。

3.4.1 hampel 数据滤波原理

Hampel 滤波器是一种基于决策的滤波器，通过该滤波器可以找出数据序列中的异常数据点，并以更有代表性的数值替换，如滤波器移动窗口中短序列的中值。Hampel 滤波法能够在不获取数据完整趋势的情况下判断粗差，具有良好的实时性和识别效果，其定义如下:

对于数据序列$\alpha_1,\alpha_2.\alpha_3..,\alpha_n$，每个样本两边的样本数和为 s，滑动窗口长为 2s+1，则窗口内样本的中值可表示为

$$\bar{\alpha_i} = \mathrm{median}(\alpha_{i-l},\alpha_{i-l+1},\cdots,\alpha_i,\cdots,\alpha_{i+l-1},\alpha_{i+l}) \tag{22}$$

中值绝对偏差的尺度估计：

$$e_i = 1.4826\mathrm{median}(|\alpha_{i-l}-\bar{\alpha_i}|,\cdots,|\alpha_{i+l}-\bar{\alpha_i}|) \tag{23}$$

数据序列可表示为：

$$b_i\begin{cases}\alpha_1 & |\alpha_i-\bar{\alpha_i}|\le 3e_i\\ \bar{\alpha_i} & |\alpha_i-\bar{\alpha_i}|>3e_i\end{cases} \tag{24}$$

由式（24）可知，如果窗口内某个值大于 3 倍的中值绝对偏差，则认为是离群点，并用窗口均值代替。

3.4.2 应用 hampel 数据滤波对 TE 过程实现预处理

根据 MATLAB 关于 hampel 函数的用法，选取各边三个数据计算窗口中值，样本对中值的标准差超过 1 时将此样本值用窗口中值替换。

由此得到的 hampel 数据滤波后进行的 KPCA 模型对 TE 过程的检测结果如下表 7 所示：

表 7　hampel 滤波后的 KPCA 故障检测结果

故障编号	检测率	误报率	故障编号	检测率	误报率
1	**100%**	**8.13%**	**12**	**99.63%**	**21.25%**

2	**99.63%**	**1.88%**	**13**	**96.25%**	**4.38%**
3	**26%**	**17.5%**	**14**	**99.63%**	**0**
4	**100%**	**1.88%**	**15**	**27%**	**1.88%**
5	**42.13%**	**1.88%**	**16**	**55.13%**	**38.13%**
6	**99.63%**	**0.63%**	**17**	**98.38%**	**5.63%**
7	**100%**	**7.5%**	**18**	**91.38%**	**0**
8	**99.38%**	**5.63%**	**19**	**35.75%**	**7.5%**
9	**24.5%**	**8.13%**	**20**	**70.25%**	**1.88%**
10	**62%**	**4.38%**	**21**	**72.75%**	**25%**
11	**77.63%**	**12.5%**			

此改进方法得到的对 TE 过程故障检测率为 75.1%，故障误报率为 8.37%，相比于原始 KPCA76.07%的故障检测率和 12.38%的故障误报率；经过 hampel 数据预处理的 KPCA 模型在保证故障检测率较高的水平下，大大降低了故障漏报率，提高了 KPCA 模型对 TE 过程的故障检测精度。

由此，经过 hampel 数据预处理的 KPCA 模型不仅故障检测率高于 PCA 模型，其故障漏报率也要低于 PCA 模型。

4 总结

本文建立了基于 PCA 和 KPCA 的故障检测模型，实现对 TE 过程的故障检测，并通过加入 hampel 滤波对数据进行预处理，得到了改进后的 KPCA 故障检测模型。

首先，本文叙述了该设计的研究意义及故障诊断的发展流程及国内外发展现状，此后研究 PCA 及 KPCA 的算法基本原理，并根据各自原理建立基于 PCA 和 KPCA 的故障检测模型，并通过构建 T^2 和 SPE 统计量及其控制限，仿真得到对 TE 过程的古战检测，计算得到各自的故障检测率与故障误报率。

通过对 PCA 和 KPCA 模型的故障检测结果进行对比分析，可以得到以下结论：

（1） 基于 PCA 和 KPCA 的故障检测模型均能很好的完成 TE 过程的故障诊断。相比之下，KPCA 能够适用于非线性系统的故障诊断。

（2） 相比而言，基于 KPCA 的故障检测模型的故障检测率要比基于 PCA 的故障检测模型的故障检测率高一些，但是其故障误报程度要严重一些。

（3） 对于难诊断的 TE 过程故障，基于 KPCA 的故障检测模型相比于基于 PCA 的故障检测模型的故障检测率高，其他故障的诊断效果两者基本相同。

（4） 通过 hampel 数据滤波改进后的 KPCA 模型对于 TE 过程的故障检测率比 PCA 模型的高，故障误报率比 PCA 模型的要低，总体故障检测效果比 PCA 模型的要好。

然而本文仍存在一些不足：

（1） 可以选用更多的核函数来改进 PCA 模型，以达到更好的故障检测效果。

（2） 可以运用更多的例如 EMD 数据预处理方法，对 TE 过程的数据进行预处理，从而得到更为理想的故障诊断效果。

参 考 文 献

[1]. 薄翠梅,张湜,张广明,等.基于特征样本核主元分析的 TE 过程快速故障辨别方法[J].化工学报，2008，59（7）：1783-1789.

[2]. Down JJ，Vogel EF.A plant-wide industrial process control problem. Computers & Chemical Engineering，1993.17(3):245-255

[3]. 何菲，杜文莉等.PCA_SVM 的多故障分类方法在 TE 过程中的应用.计算机与应用化学，2010.27(10):1321-1324

[4]. 缪素云，张峰等.基于概率神经网络的 TE 过程故障诊断.仪器仪表与检测技术，2011.30(5):78-86

[5]. 赵小强，王新明.基于改进核主元分析的 TE 过程故障诊断.工业仪表与自动化装置，2010. (3):7-11

[6]. 许洁，胡寿松.基于 KPCA 和 MKL-SVM 的非线性过程监控与故障诊断.仪器仪表学报，2010.31(11): 2428-2433

[7]. Deng X G，Tian X M，Chen S，et al. Deep Principal Component Analysis Based on Layerwise Feature Extraction and Its Application to Nonlinear Process Monitoring[J]. IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY，2019.27(06):2526-2540.

[8]. 周东华，王桂增．故障诊断技术综述[J]．化工自动化及仪表，1998.

[9]. 陈奥，基于改进多元统计方法的故障诊断技术研究[D].哈尔滨工业大学，2018.01

[10]. Jiang Q, Yan X, Huang B. Performance-Driven Distributed PCA Process Monitoring Based on Fault-Relevant Variable Selection and Bayesian Inference[J]. IEEE Trans. Industrial Electronics, 2016, 63(1): 377-386.

附录

PCA 过程监测系统

```
clc
clear
close all
%% 产生训练数据
Train = load('C:\Users\admin\Desktop\data set\tennessee-eastman-profBraatz-master\d00.dat');
Train = Train';
xtrain=Train(1:480,[1:22,42:52]);
%% 产生测试数据
Test = load('C:\Users\admin\Desktop\data set\tennessee-eastman-profBraatz-master\d01_te.dat');
xtest =Test(1:960,[1:22,42:52]);
%标准化处理：
x_mean = mean(xtrain);
x_std = std(xtrain);
[x_row,x_col] = size(xtrain);
xtrain=(xtrain-repmat(x_mean,x_row,1))./repmat(x_std,x_row,1);
xtest_R = size(xtest,1);
xtest=(xtest-repmat(x_mean,xtest_R,1))./repmat(x_std,xtest_R,1);
%%
CM = (xtrain'*xtrain)/(x_row-1);
[T,lamda] = eig(CM);
E = flipud(diag(lamda));
num_P = 1;
while sum(E(1:num_P))/sum(E) < 0.85
num_P = num_P +1;
end
P = T(:,x_col-num_P+1:x_col);
TT=xtrain*P;
%%
%求 T2 和 Q 统计量
[P_row,P_row1] = size(P*P');
I = eye(P_row,P_row1);
for i = 1:xtest_R
    T2(i)=xtest(i,:)*P*pinv(lamda(x_col-num_P+1:x_col,x_col-num_P+1:x_col))*P'*xtest(i,:)';
    %T1(i)=hampel(T2(i),20,1);
    SPE(i) = xtest(i,:)*(I - P*P')*(I - P*P')'*xtest(i,:)';
    %SPE1(i) = hampel(SPE(i),20,1);
end

%for i = 1:xtest_R
%     T1(i)=hampel(T2(i),20,1);
%     SPE1(i) = hampel(SPE(i),20,1);
```

```
%end

%T1=hampel(T2,4,1);
%SPE1 = hampel(SPE,4,1);

JT=num_P*(x_row-1)*(x_row+1)*finv(0.99,num_P,x_row - num_P)/(x_row*(x_row - num_P));
for i = 1:3
      theta(i) = sum((E(num_P+1:x_col)).^i);
end
h0 = 1 - 2*theta(1)*theta(3)/(3*theta(2)^2);
ca = norminv(0.95,0,1);
JQ = theta(1)*(h0*ca*sqrt(2*theta(2))/theta(1) + 1 + theta(2)*h0*(h0 - 1)/theta(1)^2)^(1/h0);

%T2 故障检测准确率
   k1=0;
for i=160:960
    if T2(i)>JT
          k1=k1+1;
    end
end
FDRT=k1/(960-160)
%T2 误报率
k1=0;
for i=1:159
    if T2(i)>JT
          k1=k1+1;
    end
end
FART=k1/159
%SPE 故障检测准确率
k1=0;
for i=160:960
    if SPE(i)>JQ
          k1=k1+1;
    end
end
FDRSPE=k1/(960-160)
%SPE 误报率
k1=0;
for i=1:159
    if SPE(i)>JQ
          k1=k1+1;
    end
end
```

```
FARSPE=k1/159

%%
%给出仿真图
figure('color',[1 1 1]);
subplot(2,1,1)
plot(T2,'linewidth',1.5,'color','b');
hold on
plot(repmat(JT,1,xtest_R),'r-.','linewidth',1);
xlabel('Sample number');
ylabel('T^2');
legend('Statistic','Control limit');
title(['检测准确率为',num2str(FDRT)],['误报率为',num2str(FART)]);

subplot(2,1,2)
plot(SPE,'linewidth',1.5,'color','b');
hold on
plot(repmat(JQ,1,xtest_R),'r-.','linewidth',1);
xlabel('Sample number');
ylabel('SPE');
legend('Statistic','Control limit');
title(['检测准确率为',num2str(FDRSPE)],['误报率为',num2str(FARSPE)]);
```

KPCA 过程监测系统

```
tic
clc
clear
close all
%% 产生训练数据
Train = load('C:\Users\admin\Desktop\data set\tennessee-eastman-profBraatz-master\d00_te.dat');
%Train = Train';
Test = load('C:\Users\admin\Desktop\data set\tennessee-eastman-profBraatz-master\d16_te.dat');
xtrain=Train(1:480,[1:22,42:52]);
xtest =Test(1:960,[1:22,42:52]);
xtrain1=xtrain;
[xtrain_row,xtrain_col] = size(xtrain);
xtrain=(xtrain-repmat(mean(xtrain1),xtrain_row,1))./repmat(std(xtrain1),xtrain_row,1);
xtest_row = size(xtest,1);
xtest=(xtest-repmat(mean(xtrain1),xtest_row,1))./repmat(std(xtrain1),xtest_row,1);

%% 计算核矩阵
sita=10000;
%sita=xtrain_row*xtrain_col;
for i=1:xtrain_row
```

```
    for j=1:xtrain_row
   KX(i,j) = exp(-(xtrain(i,:)-xtrain(j,:))*(xtrain(i,:)-xtrain(j,:))'/sita);
    end
end
%% 中心化核矩阵
 n1= ones(xtrain_row,xtrain_row)/xtrain_row;
KX1=KX;
KX=KX-n1*KX-KX*n1+n1*KX*n1;

for i=1:xtest_row
    for j=1:xtrain_row
   KXnew(i,j) = exp(-(xtest(i,:)-xtrain(j,:))*(xtest(i,:)-xtrain(j,:))'/sita);
    end
end
%% 中心化处理
 q1= ones(xtest_row,xtrain_row)/xtrain_row;
KXnew=KXnew-q1*KX1-KXnew*n1+q1*KX1*n1;
%% 协方差矩阵的特征值分解
CM = KX/(xtrain_row-1);
[T,lamda] = eig(CM);
E = flipud(diag(lamda));
%% 确定主元个数
num_P = 1;
while sum(E(1:num_P))/sum(E) < 0.95
num_P = num_P +1;
end
P = T(:,xtrain_row-num_P+1:xtrain_row);
%% 计算 T2 及 SPE 以及控制限
temp1=KXnew*T; temp2=KXnew*T(:,xtrain_row-num_P+1:xtrain_row);
for i = 1:xtest_row
      T2(i)=KXnew(i,:)*P*pinv(lamda(xtrain_row-num_P+1:xtrain_row,xtrain_row-
num_P+1:xtrain_row))*P'*KXnew(i,:)';
      SPE(i) = temp1(i,:)*temp1(i,:)'-temp2(i,:)*temp2(i,:)';
end

%T1=T2;
%SPE1=SPE;
T1=hampel(T2,10,0.01);
SPE1 = hampel(SPE,10,0.01);
%T1=smoothts(T2);
%SPE1 = smoothts(SPE);

temp1_obs=KX*T; temp2_obs=KX*T(:,xtrain_row-num_P+1:xtrain_row);
for i = 1:xtrain_row
```

```
    T2_obs(i)=KX(i,:)*P*pinv(lamda(xtrain_row-num_P+1:xtrain_row,xtrain_row-
num_P+1:xtrain_row))*P'*KX(i,:)';
    SPE_obs(i) = temp1_obs(i,:)*temp1_obs(i,:)'-temp2_obs(i,:)*temp2_obs(i,:)';
end
JT=ksdensity(T2_obs,0.95,'function','icdf');
JQ=ksdensity(SPE_obs,0.95,'function','icdf');


%%
% % 计算故障检测率和故障误报率
%故障检测率
FDR1=0;
for i=160:960
    if JT < T1(i)
        FDR1 = FDR1+1;
    end
end
FDR1=FDR1/(960-160);

%故障误报率
FAR1=0;
for i=1:160
    if JT< T1(i)
        FAR1 = FAR1+1;
    end
end
FAR1=FAR1/160;


%故障检测率
FDR2=0;
for i=160:960
    if JQ < SPE1(i)
        FDR2 = FDR2+1;
    end
end
FDR2=FDR2/(960-160);

%故障误报率
FAR2=0;
for i=1:160
    if JQ< SPE1(i)
        FAR2 = FAR2+1;
    end
```

```
end
FAR2=FAR2/160;

%%
figure('color',[1 1 1]);
subplot(2,1,1)
plot(T2,'linewidth',1.5,'color','b');
hold on
plot(repmat(JT,1,xtest_row),'r-.','linewidth',1);
xlabel('Sample number');
ylabel('T^2');
legend('Statistic','Control limit');
title(['检测准确率为',num2str(FDR1)],['误报率为',num2str(FAR1)]);

subplot(2,1,2)
plot(SPE,'linewidth',1.5,'color','b');
hold on
plot(repmat(JQ,1,xtest_row),'r-.','linewidth',1);
xlabel('Sample number');
ylabel('SPE');
legend('Statistic','Control limit');
title(['检测准确率为',num2str(FDR2)],['误报率为',num2str(FAR2)]);

toc
```

指导教师意见

此页学生不用打印

指导教师意见请参见模板！ 指导教师写好意见后粘贴在此页

关于批改：至少三处批改痕迹，并作出文字点评，不能只画对号。