This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



# Nonstationary Discrete Convolution Kernel for Multimodal Process Monitoring

Ruomu Tan, James R. Ottewill, *Member, IEEE*, and Nina F. Thornhill, *Senior Member, IEEE*

***Abstract*— Data-driven process monitoring has benefited from the development and application of kernel transformations, especially when various types of nonlinearity exist in the data. However, when dealing with the multimodality behavior that is frequently observed in the process operations, the most widely used radial basis function (RBF) kernel has limitations in describing process data collected from multiple normal operating modes. In this article, we highlight this limitation via a synthesized example. In order to account for the multimodality behavior and improve the fault detection performance accordingly, we propose a novel nonstationary discrete convolution kernel, which derives from the convolution kernel structure, as an alternative to the RBF kernel. By assuming the training samples to be the support of the discrete convolution, this new kernel can properly address these training samples from different operating modes with diverse properties and, therefore, can improve the data description and fault detection performance. Its performance is compared with RBF kernels under a standard kernel principal component analysis framework and with other methods proposed for multimode process monitoring via numerical examples. Moreover, a benchmark data set collected from a pilot-scale multiphase flow facility is used to demonstrate the advantages of the new kernel when applied to an experimental data set.**

***Index Terms*— Fault detection, kernel-based learning methods, multivariate statistics, nonstationary kernels, process monitoring.**

## I. Introduction

KERNEL transformation in multivariate statistical process monitoring (MSPM) has been popular due to its ability to handle nonlinearities existing in the process data and its compatibility with various dimension reduction algorithms, such as principal component analysis (PCA) [1], partial least squares (PLS) [2], and independent component analysis (ICA) [3]. The ability of MSPM in identifying new operating modes, some of which may reflect faults in the process, has been improved substantially by adopting the aforementioned kernel-based approaches. However, nonlinearity in the process data may be caused by a different process behavior and can take various forms. For example, varying loading conditions or demands of production can mean that a process may run in multiple different modes even during the course of a typical, healthy operation. Data recorded from such a process will itself be multimodal in nature. It is important to be able to account for this multimodality so that anomalous behavior may be distinguished from normal operations accurately and robustly.

Though a variety of kernel structures have been proposed and reviewed in kernel-based learning in general, the radial basis function (RBF) kernel has been the most widely used. While its advantages in one-class classification problems have been previously discussed [4], Li and Yang [5] also highlighted that a single selected RBF kernel is usually not the most effective for detecting various faults. Moreover, the separability of data could be even worse after kernel projection when an inappropriate kernel is used [6]. Recently, several efforts have been made to apply kernel-based approaches to multimodal process monitoring [7]–[12]. Though still using RBF kernels, these approaches adopt localization factors, such as just-in-time and nearest neighbors, to calculate a revised kernel matrix, indicating that on their own, RBF kernels may not be able to fully capture the covariance existing between samples due to their inadequacy in considering varying data behavior caused by different operating modes. It will be demonstrated later in this article that a single RBF kernel is not sufficient to describe the normal data set if the data are collected from multiple normal operating modes. Similar issues that exist in data explanation and modeling based on RBF kernels, such as spatially varying length scale and inhomogeneity of covariances, have been reported in the areas of geostatistics [13], terrain surface estimation [14], and natural language modeling [15].

The reason that stationary kernels, such as the RBF kernel, are insufficient is that the multimodality issue may lead to the covariance structure of process variables varying between operating modes. The nonstationary kernel is proposed to cope with this issue. Though a generic formulation of such nonstationary kernels based on convolution is available in [16], in practice, a parameter tuning step is necessary [14], [15]. Amari and Wu [17] proposed a revised kernel structure that included a data-dependent weighting function to the original RBF kernel. This structure has been applied to MSPM [18], [19]. However, it is also necessary to determine the data-dependent weighting function via a separate optimization step [6], [20]. Other kernel formulations have been explored for various aspects of the kernel-based approach. For example,

Manuscript received June 22, 2018; revised January 28, 2019 and June 7, 2019; accepted September 24, 2019. This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement 675215-PRONTO-H2020-MSCA-ITN-2015. *(Corresponding author: Ruomu Tan.)*

R. Tan and N. F. Thornhill are with the Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: r.tan@imperial.ac.uk; n.thornhill@imperial.ac.uk).

J. R. Ottewill is with the ABB Corporate Research Center, 31-038 Kraków, Poland (e-mail: james.ottewill@pl.abb.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2945847

2162-237X © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



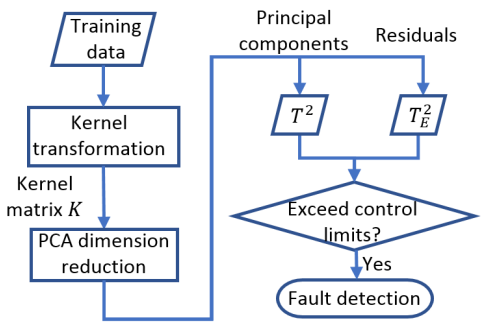


Fig. 1. Flowchart of the kernel PCA.

spectral mixture kernels may improve extrapolation in the Gaussian process modeling [21]. Multiple kernel learning approaches that combine existing kernels are reported in [22].

This article proposes a novel formulation of nonstationary and data-dependent kernel functions that can account for the varying covariance structure caused by multimodality without extra parameterization. Since the convolution formulation can generate both the stationary RBF kernel and nonstationary kernels, we define the nonstationary discrete convolution (NSDC) kernel as the covariance function of the outputs of the nonlinear regression with a finite number of basis functions, yielding the convolution on discrete, finite support. By using the samples from normal operation as the centers of the basis functions, the NSDC kernel can improve the accuracy of kernel-based models in kernel PCA. In contrast to other nonstationary formulations, this new kernel does not introduce additional parameters other than the kernel width used in RBF kernels. Therefore, the NSDC kernel can improve fault detection performance without causing overfitting issues.

The rest of this article proceeds as follows. First, we briefly review the algorithm structure of kernel PCA and demonstrate the limitation of the RBF kernel in accounting for multimodal data in MSPM using kernel PCA via an illustrative example. To retain the advantages of the RBF kernel while also making better use of the multimodal training set, Section III proposes the NSDC kernel and discusses the selection of the associated kernel widths. Numerical simulations are used to compare the anomaly detection performance of the NSDC-kernel PCA with the RBF-kernel PCA and other approaches proposed for monitoring multimodal processes. Section IV also compares the performance of the NSDC kernel and the RBF kernel in the process monitoring under the kernel PCA framework using the PRONTO benchmark data set. This article ends with a discussion of the qualitative comparison with other methods and the implementation considerations of the NSDC kernel and the conclusions.

## II. Limitation of RBF Kernel in MSPM Using Multimodal Data

### A. Mathematical Preliminary

The general structure of an MSPM algorithm based on kernel PCA is shown in Fig. 1. For simplicity, the measurement $x$ is taken as a scalar input to the algorithm.

Kernel PCA first projects the measurements $x$ to a higher dimensional nonlinear variable space $\Psi$.[1] Instead of assuming functional structures with respect to $x$ for these nonlinear variables, kernel functions are defined for $x$ in order to obtain $K$, the covariance matrix of unknown nonlinear variables, directly. For example, $K$ is defined by the RBF kernel function for $i$th and $j$th samples of $x$ in the following equation:

$$K_{i,j} = k(x(i), x(j)) = \exp\left(-\frac{(x(i) - x(j))^2}{l^2}\right) \tag{1}$$

where $l^2$ is the kernel width parameter. In practice, it is common to assume $l^2 = \delta\sigma_x^2$, where $\sigma_x^2$ is the sample variance of $x$ and $\delta$ is a scaling factor.

It has been proven that PCA can be applied to $K$ and feature extraction may be realized in the nonlinear variable space by solving the following eigenvalue problem [24]. Assume that $n$ samples of $x$ are available for training

$$\begin{aligned} n\lambda\boldsymbol{\alpha} &= K\boldsymbol{\alpha} \\ \boldsymbol{z} &= \sum_{i=1}^{n} \alpha_i \boldsymbol{k}_i \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}$ and $\boldsymbol{k}_i = \{K_{1,i}, K_{2,i}, \ldots, K_{n,i}\}$.

Therefore, representative features $\boldsymbol{z}$ are extracted from the nonlinear variable space using $K$ in kernel PCA. Similar to PCA-based MSPM, these features are further divided into principal components (PCs) for monitoring systematic errors and residuals for model-data mismatch according to their eigenvalues. The PCs and residuals are used for calculating monitoring statistics $T^2$ and $T_E^2$, respectively [25]. When implementing this MSPM algorithm, the monitoring model is trained offline using kernel PCA and the control limits of the monitoring statistics are set. In online monitoring, monitoring statistics of a new sample are calculated using the same kernel PCA model and are compared with their control limits for fault detection. A detailed description of the MSPM algorithm based on kernel PCA can be found in [26].

### B. RBF Kernel Performance in Multimodal Data

The following bivariate model of $\boldsymbol{x} = [x_1, x_2]$ with four operating modes is considered, and 100 samples are drawn randomly from each to formulate the training set, as shown in Fig. 2.

*Mode 1:*

$$\begin{aligned} x_1 &= e_{11} \\ x_2 &= 1.5x_1 + e_{12} \end{aligned} \tag{3}$$

where $e_{11} \sim N(0, 1)$ and $e_{12} \sim N(0, 9)$.

*Mode 2:*

$$\begin{aligned} x_1 &= e_{21} + 8 \\ x_2 &= -0.2x_1 + 5 + e_{22} \end{aligned} \tag{4}$$

where $e_{21} \sim N(0, 2.25)$ and $e_{22} \sim N(0, 0.25)$.

[1]This projection is different from the kernel mapping that generates spatial maps representing local image characteristics [23].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



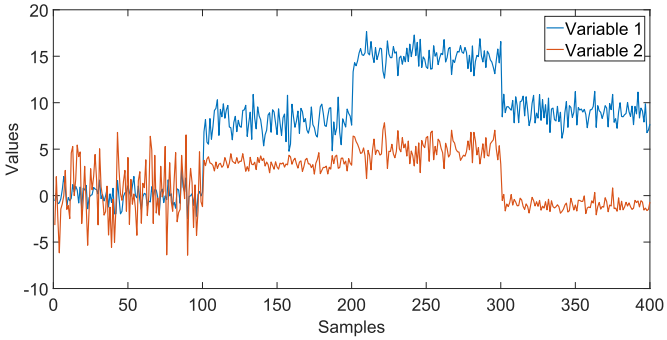


Fig. 2. Trend plot of training set 1.

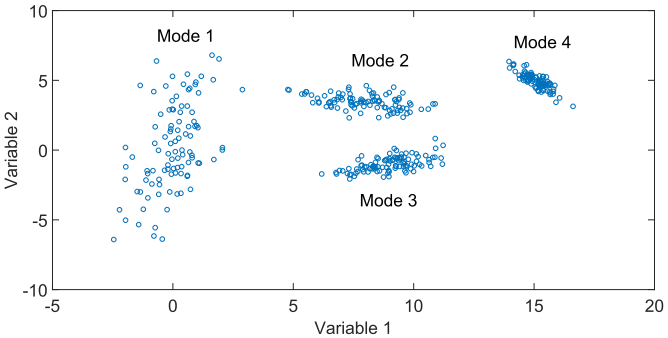


Fig. 3. Scatter plot of training set 1.

*Mode 3:*

$$\begin{aligned} x_1 &= e_{31} + 15 \\ x_2 &= -x_1 + 20 + e_{32} \end{aligned} \tag{5}$$

where $e_{31} \sim N(0, 0.25)$ and $e_{32} \sim N(0, 0.09)$.

*Mode 4:*

$$\begin{aligned} x_1 &= e_{41} + 9 \\ x_2 &= \frac{1}{3}x_1 - 4 + e_{42} \end{aligned} \tag{6}$$

where $e_{41} \sim N(0, 1)$ and $e_{42} \sim N(0, 0.25)$.

It may be observed from the trend plot in Fig. 2 and from the scatter plot in Fig. 3 that the variance of the data set from Mode 1 is larger than the variances of the other three sets. In practice, such differences in variance may exist due to the nonlinearity in the process variables or measurement instruments. For example, a flow measurement might have higher measurement variability if the air is entrained in the process fluid.

The limitation of RBF kernels will be demonstrated via kernel PCA models trained from this data set. The $T^2$ statistic defined after kernel PCA projection represents the Mahanalobis distance defined in the feature space [1]. After obtaining representative features $\boldsymbol{z}$, the first $r$ features in $\boldsymbol{z}$ that hold more than 99% of the overall variability in the eigenvalues are selected as PCs and are used for $T^2$ calculation. In order to show the boundaries for anomaly detection, 95% control limits of $T^2$ obtained after kernel PCA with different scaling factors $\delta$ are visualized alongside the original samples.

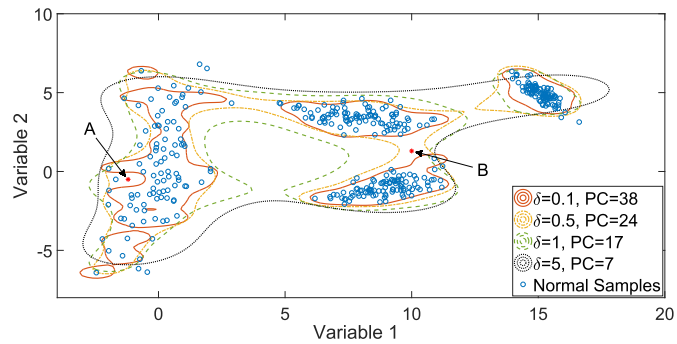


Fig. 4. Control limits of $T^2$ obtained by RBF kernels for training set 1.

When inspecting Fig. 4, sample A may be considered as belonging to Mode 1 and sample B may be considered to be an anomaly because it cannot be clearly associated with any existing mode. While sample B can only be detected when $\delta = 0.1$, sample A will be identified as anomalous due to the overfitting issue of the same model. Therefore, monitoring models obtained by kernel PCA with RBF kernels have limitations when applied to this data set.

It can be observed that large $\delta$ values will result in overly relaxed detection boundaries that cannot detect the transitions and/or deviations from normal operating modes ($\delta = 1, 5$); on the other hand, the number of PCs retained after kernel PCA model increases rapidly and will soon result in overfit models as $\delta$ reduces ($\delta = 0.5, 0.1$). In summary, the model accuracy for the multimodal data set may be limited if trained using a single RBF kernel.

## III. NSDC Kernel

The major notations used in this section are as follows:

1) $\boldsymbol{x} \in \mathbb{R}^m = \{x_1, x_2, \ldots, x_m\}$: a vector with $m$ measured process variables;
2) $C \in \mathbb{R}^{m \times n}$: the data set from normal operations with $n$ samples of $\boldsymbol{x}$ used for training the model;
3) $y \in \mathbb{R}$: a variable defined in the nonlinear space $\Psi(\boldsymbol{x})$;
4) $\phi(\boldsymbol{x})$ : a function of $\boldsymbol{x}$ that is used as a basis function to reconstruct $y$;
5) $l^2$: kernel widths for RBF and NSDC kernels.

### A. NSDC Kernel as a Covariance Function

Equation (7) gives a regression model using $p$ basis functions $\phi_i(\boldsymbol{x})$, where $i = 1, 2, \ldots, p$

$$y = \sum_{i=1}^{p} w_i \phi_i(\boldsymbol{x}) \tag{7}$$

where $\boldsymbol{x} \in \mathbb{R}^m$ is the model input and $y \in \mathbb{R}$ is the model output. $w_i \sim N(0, \sigma_w^2)$ are the regression coefficients with the i.i.d. Gaussian distribution corresponding to basis functions $\phi_i(\boldsymbol{x})$ for $i = 1, 2, \ldots, p$.

The covariance of two new outputs, $y$ and $y^*$, can be calculated as a function of input samples, namely, $\boldsymbol{x}$ and $\boldsymbol{x}^*$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



This is known as the kernel function

$$\begin{aligned}\mathrm{cov}(y, y^*) &= k(\boldsymbol{x}, \boldsymbol{x}^*)\\ &= \mathbb{E}\left[\sum_{i=1}^{p} w_i\phi_i(\boldsymbol{x})\sum_{i=1}^{p} w_i\phi_i(\boldsymbol{x}^*)\right]\\ &= \Phi^T(\boldsymbol{x})\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]\Phi(\boldsymbol{x}^*)\end{aligned} \tag{8}$$

where $\Phi(\boldsymbol{x}) = \{\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_p(\boldsymbol{x})\}$. $\boldsymbol{w} = \{w_1, w_2, \ldots, w_p\}$ is the coefficient vector, and $\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$ denotes the expectation of $\boldsymbol{w}^T\boldsymbol{w}$. Since $w_i$ are the i.i.d. Gaussian distributed with variance $\sigma_w^2$, (8) can be written as

$$\mathrm{cov}(y, y^*) = \sigma_w^2\sum_{i=1}^{p}\phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x}^*). \tag{9}$$

When the RBF is adopted in (7) with $\boldsymbol{c}^{(i)}$ as its center, the covariance of $y$ and $y^*$ is

$$\mathrm{cov}(y, y^*) = \sigma_w^2\sum_{i=1}^{p}\left[\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}-\boldsymbol{c}^{(i)})}{l^2}\right)\times\exp\left(-\frac{(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})}{l^2}\right)\right] \tag{10}$$

where $\boldsymbol{c}^{(i)}$ is the center of the $i$th RBF.

When an infinite number of basis functions are considered, i.e., $p \to \infty$, all $\boldsymbol{c}^{(i)}$ are allocated evenly from $-\infty$ to $\infty$ and $\mathrm{d}\boldsymbol{c} = \boldsymbol{c}^{(i+1)} - \boldsymbol{c}^{(i)}$. Equation (10) becomes

$$\begin{aligned}\mathrm{cov}(y, y^*) &= \sigma_w^2\sum_{i=1}^{\infty}\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x})\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x}^*)\\ &= \sigma_0^2\int_{-\infty}^{\infty}[\phi_{\boldsymbol{c}}(\boldsymbol{x})\phi_{\boldsymbol{c}}(\boldsymbol{x}^*)]d\boldsymbol{c}\\ &= \sqrt{\pi}l\sigma_0^2\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*)}{l^2}\right)\end{aligned} \tag{11}$$

where $\sigma_w^2$ is selected as $(\sigma_0^2/p)$ to avoid the covariance value approaching to infinity (see the Appendix). Equation (11) is the convolution of two Gaussian functions. This convolution formulation leads to a scaled and multivariate formulation of the RBF kernel function presented in (1). The covariance matrix $K$ in kernel-based methods can be calculated accordingly.

Like the RBF kernel [see (11)], the NSDC kernel also derives from (10). Instead of having an infinite number of centers allocated from $-\infty$ to $\infty$, the data dependence in the kernel function can be reflected by selecting only training samples as the centers of the basis functions. Assuming that $P$ clusters of normal training samples obtained from $P$ operating modes exist in $C \in \mathbb{R}^{m\times n}$, i.e., $C = C_1 \bigcup \cdots \bigcup C_P$, the kernel function can be defined with $C$ as its support

$$k_{\mathrm{NSDC}}(\boldsymbol{x}, \boldsymbol{x}^*) = \mathrm{cov}(y, y^*) = \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x})\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x}^*) \tag{12}$$

where $\boldsymbol{c}^{(i)} \in C$. It is important to note that the number of basis functions is also equal to $n$, the number of samples from normal operating modes. The univariate and multivariate solutions to the discrete convolution structure in (12) will yield the new NSDC kernel function. For conciseness, we denote $k_{\mathrm{NSDC}}$ as $k$ from now on.

### B. Univariate Formulation

For simplicity, we first assume $\boldsymbol{x}$ to be univariate. By using RBFs in (12), the NSDC kernel can be derived as

$$\begin{aligned}k(x, x^*) &= \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\left[\exp\left(-\frac{(x-c^{(i)})^2}{l^2}\right)\times\exp\left(-\frac{(x^*-c^{(i)})^2}{l^2}\right)\right]\\ &= \sigma_0^2\exp\left(-\frac{d^2}{2l^2}\right)\times\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\left(c^{(i)}-\frac{1}{2}(x+x^*)\right)^2}{\frac{1}{2}l^2}\right)\end{aligned} \tag{13}$$

where $d = x - x^*$ is the distance between $x$ and $x^*$. This new kernel has a similar formulation to the RBF kernel. However, given $c_i \in C$, the weighting coefficient

$$\frac{1}{n}\sum_{i=1}^{n}\left[\exp\left(-\frac{\left(c^{(i)}-\frac{1}{2}(x+x^*)\right)^2}{\frac{1}{2}l^2}\right)\right]$$

is proportional to the conditional likelihood of $(x + x^*)/2$ given the training set $C$, $P_{\mathrm{KDE}}((x + x^*)/2|C)$, using kernel density estimation. Therefore, the extra weighting coefficient makes this new kernel dependent on the training set $C$. In addition, this kernel is nonstationary as it is dependent not only on the distance $d$ between two input samples but also on the locations of these samples.

Moreover, when considering the autocovariance of a single sample $x^*$, $d = 0$ and its mean is $x^*$

$$k(x^*, x^*) = \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\exp\left(-\frac{(c^{(i)}-x^*)^2}{\frac{1}{2}l^2}\right). \tag{14}$$

The autocovariance of $x^*$ is, therefore, proportional to the conditional likelihood of $x^*$, given the training set $C$.

### C. Multivariate Extension

For the multidimensional case, $\boldsymbol{x}$ is taken to be an $m$-dimensional vector. The NSDC kernel is, thus, given as

$$k(\boldsymbol{x}, \boldsymbol{x}^*) = \frac{\sigma_0^2}{n}\exp\left(-\frac{\boldsymbol{d}^T\boldsymbol{d}}{2l^2}\right)\sum_{i=1}^{n}\exp\times\left(-\frac{\left(\boldsymbol{c}^{(i)}-\frac{1}{2}(\boldsymbol{x}+\boldsymbol{x}^*)\right)^T\left(\boldsymbol{c}^{(i)}-\frac{1}{2}(\boldsymbol{x}+\boldsymbol{x}^*)\right)}{\frac{1}{2}l^2}\right). \tag{15}$$

Similar to the univariate case, this revised kernel function is the product of the RBF kernel with respect to the distance $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}^*$ and the likelihood of the mean $(x + x^*)/2$, given $C$ using kernel density estimation.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



### D. Specification of Kernel Width

Due to the multimodal nature of the training set, the ideal kernel width, which represents the rate of covariance of two projected samples decreasing with respect to the distance between the original samples, may vary because of: 1) different underlying mechanisms of each operating mode and 2) different variances of individual variables in the same operating mode. Therefore, it is necessary to specify the kernel widths properly.

*1) Kernel Widths for Individual Variables:* RBF kernels based on the Mahanalobis distance have been investigated as an approach for estimating the kernel width of variables and avoiding the optimization of kernel width parameters [27], [28]. Similarly, in the NSDC kernel, the kernel width can be estimated by the covariance matrix of process variables and a global scaling factor. By introducing the $m \times m$ covariance matrix $\Lambda = \mathrm{cov}(\boldsymbol{x})$, the basis function $\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x})$ will be revised

$$\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x}) = \exp\left(-\frac{1}{\delta}(\boldsymbol{x}-\boldsymbol{c}^{(i)})^T \Lambda^{-1}(\boldsymbol{x}-\boldsymbol{c}^{(i)})\right). \quad (16)$$

The NSDC kernel is derived accordingly

$$k(\boldsymbol{x},\boldsymbol{x}^*) = \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\left[\exp\left(-\frac{1}{2\delta}\boldsymbol{d}^T\Lambda^{-1}\boldsymbol{d}\right) \times \exp\left(-\frac{2}{\delta}\left(\boldsymbol{c}^{(i)}-\frac{\boldsymbol{x}+\boldsymbol{x}^*}{2}\right)^T \times \Lambda^{-1}\left(\boldsymbol{c}^{(i)}-\frac{\boldsymbol{x}+\boldsymbol{x}^*}{2}\right)\right)\right]. \quad (17)$$

The covariance matrix $\Lambda$ can be estimated using the sample covariance of the training set. The scaling factor $\delta$ is the only parameter to be specified.

*2) Kernel Widths for Operating Modes:* Kernels based on the Mahanalobis distance have been widely studied and applied to different areas in the literature. However, multimodal behavior in the process data is less studied because the training clusters are not represented explicitly in the kernel function. When *a priori* information about data clusters with respect to operating modes and transition periods is available, it is possible to assign an individual covariance matrix $\Lambda_p$ for the $p$th cluster $C_p$ with $n_p$ samples in order to represent different operating modes. The basis function for the $p$th cluster is

$$\phi_{\boldsymbol{c}_p^{(i)}}(\boldsymbol{x}) = \exp\left(-\frac{1}{\delta}\left(\boldsymbol{x}-\boldsymbol{c}_p^{(i)}\right)^T\Lambda_p^{-1}\left(\boldsymbol{x}-\boldsymbol{c}_p^{(i)}\right)\right) \quad (18)$$

where $\boldsymbol{c}_p^{(i)} \in C_p$.

The kernel function can be constructed accordingly

$$k(\boldsymbol{x},\boldsymbol{x}^*) = \frac{\sigma_0^2}{n}\sum_{p=1}^{P}\left[\exp\left(-\frac{1}{2\delta}\boldsymbol{d}^T\Lambda_p^{-1}\boldsymbol{d}\right) \times \sum_{i=1}^{n_p}\exp\left(-\frac{2}{\delta}\left(\boldsymbol{c}_p^{(i)}-\frac{\boldsymbol{x}+\boldsymbol{x}^*}{2}\right)^T \times \Lambda_p^{-1}\left(\boldsymbol{c}_p^{(i)}-\frac{\boldsymbol{x}+\boldsymbol{x}^*}{2}\right)\right)\right] \quad (19)$$

where $\Lambda_p = \mathrm{cov}(\boldsymbol{x})$ such that $\boldsymbol{x} \in C_p$ can be estimated by the sample covariance of the $p$th data cluster and the scaling factor $\delta$. In particular, when the data clustering information is not available, the NSDC kernel can still be implemented by assuming $P = 1$, yielding (17).

To summarize, the new NSDC kernel adopts the sample covariance matrices of each data cluster in its formulation. In this formulation, the scaling factor $\delta$ regulates the overall behavior of the NSDC kernel, while the nonstationary covariance structure is captured by the varying sample covariances. Consequently, compared to the RBF kernel, the NSDC kernel can handle the nonstationary behavior caused by multiple operating modes without introducing additional parameters.

### E. Monitoring Statistics

Under the kernel PCA framework presented in Section II-A, we apply PCA to the kernel matrix $K$ obtained by the NSDC kernel for feature extraction. Assume that the training data set is $[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}] \in \mathbb{R}^{m\times n}$; $q$ features, namely, $z^{(1)}, z^{(1)}, \ldots, z^{(q)}$, are obtained by applying PCA to its kernel matrix $K$ such that $K_{ij} = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. Following the algorithm structure in Fig. 1, the sum of squares of Mahanalobis distances in the PC space and the sum of squares of residuals are used for quantifying the systematic error and model-data mismatch. Equations (20) and (21) define these two statistics with respect to the features

$$T^2 = z_R^T D_R^{-1} z_R^T \quad (20)$$

$$T_E^2 = z_E^T z_E^T \quad (21)$$

where $z_R = \{z^{(1)}, z^{(2)}, \ldots, z^{(r)}\}$ are the first $r$ PCs that explain the majority of variability in the feature space and $z_E = \{z^{(r+1)}, z^{(d+2)}, \ldots, z^{(q)}\}$ are considered as the residual vectors with minimal variability; $D_R$ is an $r \times r$ diagonal matrix with the first $r$ eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ corresponding to $z_R$ in the descending order. By introducing $T^2$ and $T_E^2$, both systematic and model-data mismatch faults can be detected. In order to ensure potential model-data mismatch behaviors being captured by $T_E^2$, $q$ should be sufficiently large. For fault detection, the lower control limit of $T^2$ and the upper control limit of $T_E^2$ with a certain confidence level are defined by applying kernel density estimation to $T^2$ and $T_E^2$ values on the training set. The reason for using lower control limits of $T^2$ is that due to the multimodality in the data, it is not appropriate to center the data using their mean. Therefore, if the data set and the kernel matrix are not centered, the zero-mean assumption of the PCs is no longer valid. In summary, a sample is detected as faulty if the following holds for its monitoring statistics:

$$T^2 < T_{\mathrm{LCL}}^2 \quad \text{or} \quad T_E^2 > T_{E,\mathrm{UCL}}^2. \quad (22)$$

### F. Parameter Tuning

It has been shown in Section II-B that selections of the scaling factor $\delta$ will result in MSPM models varying significantly. Recent literature on kernel-based MSPM still adopts various empirical values of kernel width [29], [30] or empirical equations [31]. Reference [1] pointed out that the parameter

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



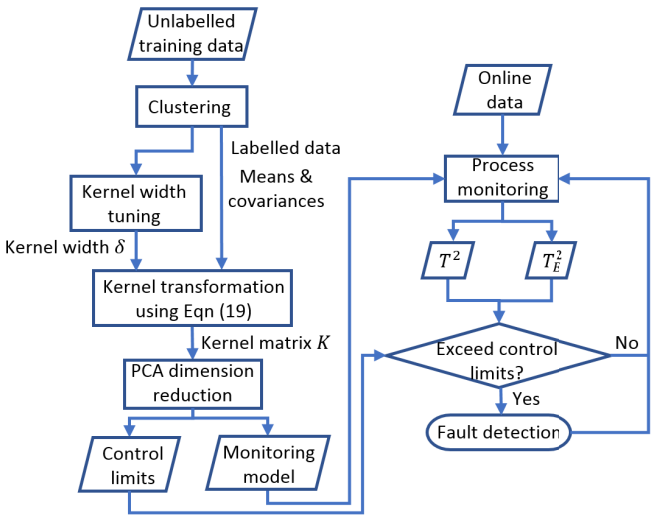


Fig. 5. Implementing NSDC-kernel PCA for process monitoring.

tuning of KPCA may lead to improved model robustness, whereas model sensitivity decreases. In this work, we will determine the scaling factor for RBF and NSDC kernels by cross validation using the data set collected from normal operation; however, instead of minimizing the alarm rate on the cross-validation set, this strategy aims at a balance between the sensitivity and the robustness of monitoring models.

We propose to select the parameter that maintains a reasonable alarm level on the cross-validation set. For example, 95% control limits of $T^2$ statistics, which assumes 5% of the data in the training set should indicate an anomaly, might be selected as the control limit. Given that the cross-validation set is comprised of random samples from the same data as the training set, ideally, 5% of the cross-validation data should also be indicated as anomalous. If more than 5% of data from the cross-validation set are indicated to be anomalous, this would indicate that the parameters of the model are inappropriate such that, in practice, the model would be prone to false alarms, e.g., due to overfitting of the model. On the other hand, if less than 5% of data from the cross-validation set are indicated as anomalies, this would indicate that the parameters of the model are such that the model would be prone to missed alarms, e.g., due to underfitting. To further increase the confidence in the obtained anomaly detection rates, results may be averaged over multiple Monte Carlo simulations. By repeating this analysis for multiple selections of the scaling factor $\delta$, it is possible to identify the $\delta$ that optimally describes the training data.

### G. NSDC-Kernel PCA in Process Monitoring

The flowchart in Fig. 5 summarizes the procedure of using the NSDC-kernel PCA for process monitoring. In comparison with Fig. 1, the NSDC-kernel PCA model is trained using the clustered multimode training data, and the kernel width $\delta$ is tuned using the strategy introduced previously. When deployed for online monitoring, the NSDC-kernel PCA model can be used without additional clustering or tuning.

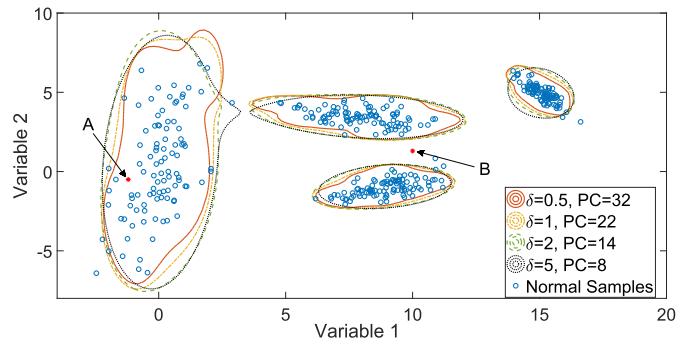


Fig. 6. Control limits obtained by NSDC kernels.

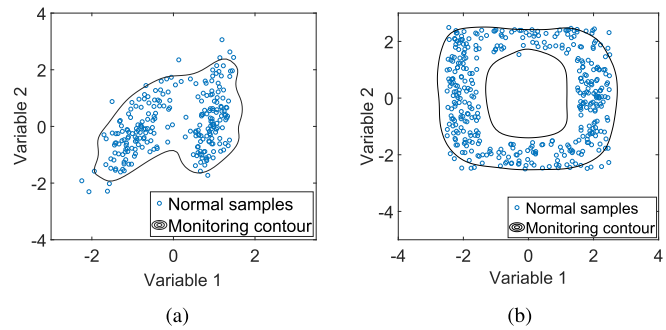


Fig. 7. Control limits obtained by the NSDC kernel for other nonlinear examples. (a) Example 2. (b) Example 3.

## IV. Performance

### A. Numerical Simulation

The data are both generated and used in batch for model training in this case study. The same data set generated in Section II-B is used for performance comparison of the NSDC kernel and the RBF kernel. The first $r$ PCs with 99% accumulated variability are chosen as $z_R$, i.e., $r$ is selected such that

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{q} \lambda_i} \geq 99\% \tag{23}$$

where $\lambda_i$ is the $i$th element of $q$ eigenvalues corresponding to $z$ in the descending order.

The 95% control limits of NSDC-kernel PCA are visualized in Fig. 6. These control limits identify sample A as a normal sample and detect sample B as an anomaly. By comparing Figs. 4 and 6, we can conclude that NSDC-kernel PCA gives better descriptions of the multimodal training data set and will significantly improve the anomaly detection performance.

Fig. 7 shows the monitoring contours generated by the NSDC kernel for two further examples. In these samples, the data are not clustered in advance. By setting the cluster number $P = 1$, one can implement the NSDC kernel defined by (17) and obtain proper monitoring contours. The contours also demonstrate that the NSDC kernel can cope with other types of nonlinearity without considering the varying covariance structures of each data cluster.

These results indicate that the NSDC kernel will yield a kernel PCA model that generates a better control limit

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



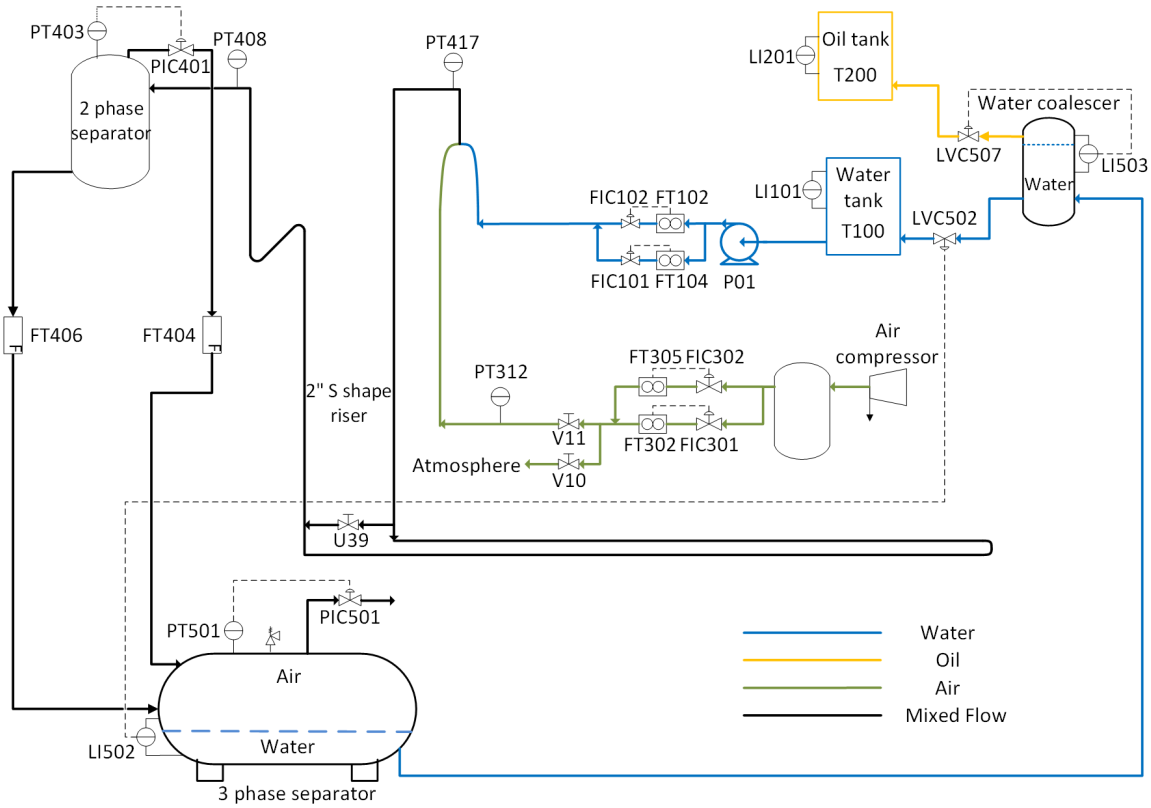


Fig. 8. Schematic of the Multiphase flow facility.

than the RBF kernel for anomaly detection of multimodal data. The NSDC kernel also suffers less from the issues associated with overfitting or underfitting. Even when there is no data clustering information available, the performance of the NSDC kernel will not be significantly compromised. It also indicates that the NSDC kernel can handle other types of data nonlinearity in addition to the multiple operating modes.

### *B. Case Study in Process Monitoring*

*1) Process Description:* We compare the fault detection performance of the RBF-kernel PCA and the NSDC-kernel PCA using the PRONTO benchmark data set that is established for algorithm design and validation for process monitoring [32]. This data set was collected sequentially during an experiment on the multiphase flow facility located at Cranfield University. Being a fully automated industrial-scale pilot plant, this facility implements mixing, transportation, and separation of multiphase flows, such as oil, water, and air. Fig. 8 presents the layout of this facility. One may refer to [32] for further details about the facility and the benchmark case study.

Table I summarizes the process variables used in this test. The multimodal behavior is realized by specifying inlet water and airflow rates according to Table II. A high-density plot [33] shown in Fig. 9 visualizes the normalized time trends of process variables measured in normal operating modes.

As shown in Table III, three faults have been seeded individually by manually opening (for air leakage and diverted flow) or closing (for air blockage) corresponding valves in both operating Modes A and B. Starting from the normal operating condition, the valve opening is changed gradually in order to simulate the development of incipient faults in real-life process operations. As an example, Fig. 10 presents the high-density plot of process variables when Fault 2 was seeded in operating mode B.

TABLE I
PROCESS VARIABLES

| No. | Tag | Description |
|---|---|---|
| 1 | FT305/302 | Inlet air flow rate |
| 2 | PT312 | Air delivery pressure |
| 3 | FT102/104 | Inlet water flow rate |
| 4 | PT417 | Pressure in the mixing zone |
| 5 | PT408 | Pressure at the riser top |
| 6 | PT403 | Pressure in the 2-phase separator |
| 7 | FT404 | 2-phase separator output air flow rate |
| 8 | FT406 | 2-phase separator output water flow rate |
| 9 | PT501 | Pressure in the 3-phase separator |
| 10 | PIC501 | Air outlet valve 3-phase separator |
| 11 | LI502 | Water level 3-phase separator |
| 12 | LI503 | Water coalescer level |
| 13 | LVC502 | Water coalescer outlet valve |
| 14 | LI101 | Water tank level |

TABLE II
NORMAL OPERATING CONDITIONS

| Mode | Input air flow rate | Input water flow rate | Data length |
|---|---|---|---|
| A | 120 $m^3$/h | 0.1 kg/s | 501 |
| B | 150 $m^3$/h | 0.5 kg/s | 561 |

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE III
MANUALLY SEEDED FAULTS

| No. | Fault | Location | Data length | |
|---|---|---|---|---|
| | | | Mode A | Mode B |
| 1 | Input air leakage | V10 | 920 | 840 |
| 2 | Input air blockage | V11 | 1360 | 1260 |
| 3 | Riser base diverted flow | U39 | 1395 | 1100 |

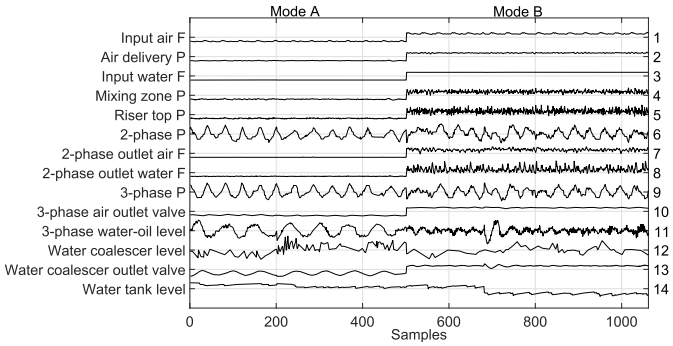


Fig. 9. High-density plot of process variables from normal operations.

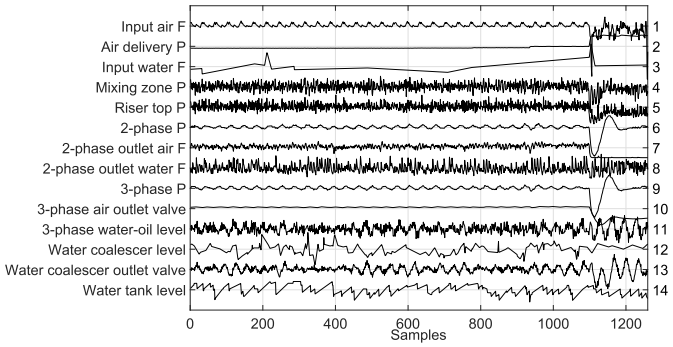


Fig. 10. High-density plot of process variables from Fault 2, Mode B.

*2) Results:* The case study in this work uses both normal and faulty data from the experiment. Normal data from Modes A and B are randomly partitioned into training, cross validation, and test sets with an equivalent amount of samples. The training and cross-validation sets are used to train the monitoring model. The monitoring model is then applied to the test data set and the faulty data for fault detection. The data are used in batch to train the monitoring model and detect the faults. Since the temporal correlation is not considered in the kernel-based approaches, fault detection can also be conducted using sequential data.

The same strategy as in (23) is adopted for grouping the features obtained by kernel PCA into PCs and residuals. The confidence level of control limits for $T^2$ and $T_E^2$ is set as 95%. To evaluate the performance, the false alarm rate (FAR) defined in (24) on the cross-validation set is compared over the scaling factor $\delta$ values in Fig. 11

$$\mathrm{FAR} = \frac{n_{\mathrm{FA}}}{n_{\mathrm{norm}}} \tag{24}$$

where an anomaly is detected if the monitoring statistics of a test sample fulfill (22); otherwise, the test sample is labeled as normal. $n_{\mathrm{FA}}$ denotes the number of normal samples being detected as anomalies, and $n_{\mathrm{norm}}$ denotes the number of all normal samples. When applied to monitoring, $\delta$ is set to be 60 for the RBF kernel and 100 for the NSDC kernel according to the strategy proposed in Section III-F. Fig. 11 shows that reduced $\delta$ values will lead to larger FARs on cross-validation sets, of which the samples are supposed to be normal, indicating the model being overfit. On the other hand, since the confidence level of control limits is selected as 95%, FARs below 5% imply that the monitoring model might have a higher missed detection rate (MDR) and, hence, be less sensitive to faults when the $\delta$ value is large.

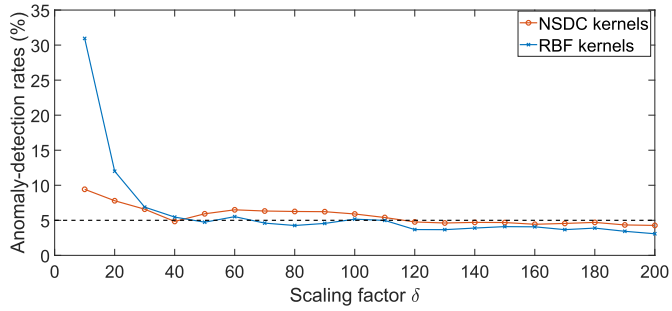


Fig. 11. Alarm rates on cross-validation sets of multiphase flow data.

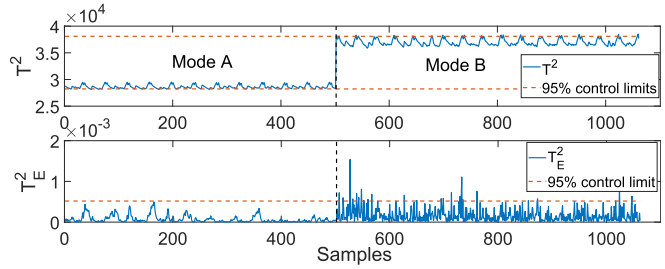


Fig. 12. Monitoring result for normal data: linear PCA.

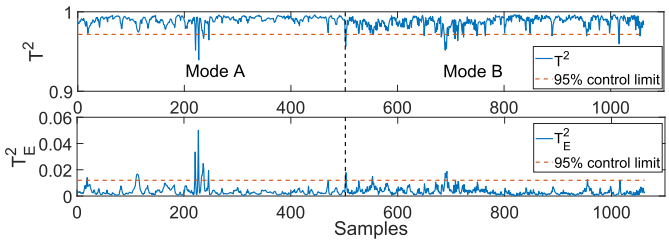


Fig. 13. Monitoring result for normal data: RBF kernel.

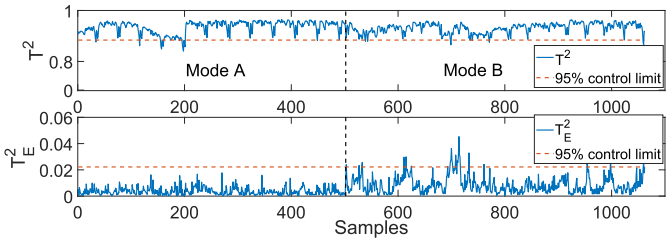


Fig. 14. Monitoring result for normal data: NSDC kernel.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE IV
MONITORING PERFORMANCE

| Kernel | Normal | Fault scenario | | | | | |
|---|---|---|---|---|---|---|---|
| | FAR (%) | Mode A: DT | | | Mode B: DT | | |
| | | F1 | F2 | F3 | F1 | F2 | F3 |
| RBF | 5.48 | 232 | 1020 | OD[1] | 159 | 799 | 383 |
| NSDC | 4.69 | 230 | 336 | OD | 159 | 655 | 43 |

[1] A detection occurs from the first sample.

Monitoring statistics of the data set in normal operations obtained by Linear PCA, RBF-kernel PCA, and NSDC-kernel PCA are shown in Figs. 12–14. For the clarity of visualization, $T^2$ statistics are plotted in the logarithmic scale in Figs. 13 and 14. In Fig. 12, the $T^2$ statistic obtained by the linear PCA has a large increase, and the $T_E^2$ statistic has a larger variance when the process switched from Mode A to Mode B while the statistics obtained by the RBF-kernel PCA and the NSDC-kernel PCA do not. Therefore, it is clear that the influence of multimodality in the training set can be reduced by applying kernel approaches. Moreover, as shown in Fig. 13, the $T_E^2$ statistics obtained by the RBF-kernel PCA has a peak around sample 200 that lasts for approximately 30 samples. This may cause the detection threshold of $T_E^2$ to be overly relaxed.

The second column in Table IV compares the FARs of the NSDC and the RBF kernels on the test sets that comprise normal samples that are not used for model training and parameter tuning. Since the confidence level is set to be 95%, the tuning strategy in Section III-F ensures that both the RBF and the NSDC kernels have an FAR that is close to 5% on the test set.

The fault detection time (DT), defined by (25), provides a measure of the sensitivity of the monitoring model in this example. Since the fault severity increased gradually in the experiment (e.g., the valve opening sequence in Fig. 15) and the variation in the process variables may not be visible in the early stage (see Fig. 10), it is difficult to define the faulty period clearly. Therefore, the DT is used instead of the MDR. Due to the persistent existence of the fault, we define that the fault detection occurs when a consecutive sequence of 20 samples exceeds the control limits. This reduces the influence of noise in the process measurements

$$\mathrm{DT} = t(\text{fault detection}) - t(\text{fault occurrence}). \tag{25}$$

Table IV shows that the process monitoring model based on the NSDC kernels is capable of detecting the faults earlier when air blockage (F2) and diverted flow (F3) occur. Fig. 15 visualizes the sequence of valve opening adopted for seeding Fault 2 in operating mode B. In Figs. 16 and 17, the monitoring performance in mode B is compared against this sequence. In particular, $T_E^2$ statistics can have an earlier detection when the blockage is less severe. It indicates that at the early stage of an incipient fault and when the deviation in PCs is not significant, small model-data mismatches existing in the process measurements caused by the fault can be captured by the monitoring model using the NSDC kernel. As a result, the incipient fault can be detected and dealt with before severe performance degradation occurs in the process.

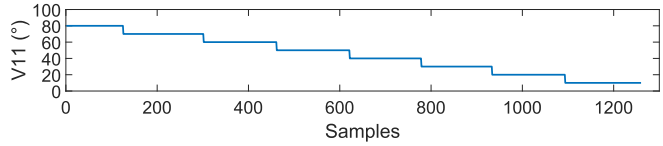


Fig. 15. Valve opening sequence for Fault 2, Mode B.

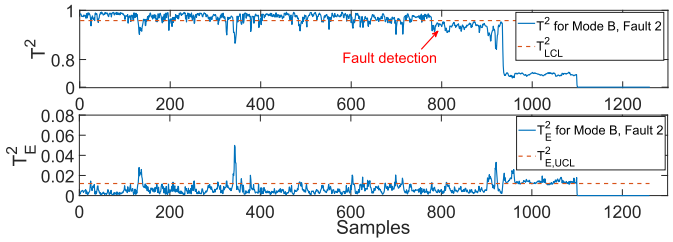


Fig. 16. RBF Monitoring performance for Fault 2, Mode B.

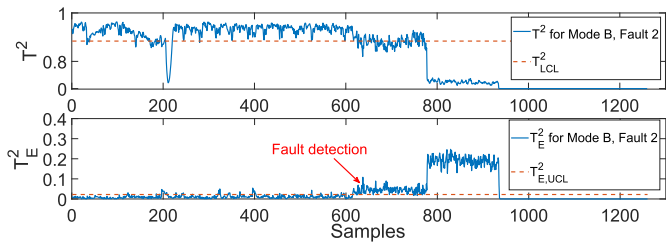


Fig. 17. NSDC Monitoring performance for Fault 2, Mode B.

## C. Comparison With Other Methods

Comparisons with other methods proposed for multimode process monitoring are challenging because it is difficult to implement and tune each method in a rigorous manner that ensures a fair comparison. In order to provide a fair evaluation of the performance of the NSDC kernel relative to the existing methods, the method is directly compared with the results reported in recently published approaches that adapt kernel-based methods for multimode process monitoring [10]–[12]. The simulated data sets described in each of these papers form the basis of the comparison. The FAR previously defined in (24) and the MDR defined in (26) are used to evaluate the anomaly detection performance

$$\mathrm{MDR} = \frac{n_{\mathrm{MD}}}{n_{\mathrm{anom}}} \tag{26}$$

where $n_{\mathrm{MD}}$ denotes the number of anomalous samples not being detected as anomalies and $n_{\mathrm{anom}}$ denotes the total number of anomalous samples. The MDR is the rate of missed detections in a test set that includes anomalous samples.

Table V presents the fault detection performance. The kernel widths of the RBF and the NSDC kernels and the control limits of $T^2$ and $T_E^2$ are tuned using the strategy proposed in Section III-F according to the confidence levels used in [10]–[12]. The confidence levels of monitoring statistics are used as the expected FARs for both the cross validation and the test sets. Therefore, the FAR obtained by NSDC and RBF

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE V
PERFORMANCE COMPARISON WITH THE EXISTING APPROACHES

| Tested scenarios | Confidence level (%) | FAR (%) | | | MDR (%) | | |
|---|---|---|---|---|---|---|---|
| | | NSDC | RBF | Other | NSDC | RBF | Other |
| Fault D2 from Section V.A of [10] | 5 | 4.13 | 4.85 | 5 | 8.08 | 11.58 | 14.2 |
| Case 4 from Section 4.1 of [11] | 5 | 3.7 | 4.42 | 4 | 1.15 | 5.68 | 18.2 |
| Case 2 from Section 5.1 of [12] | 1 | 0.83 | 1 | 0.73 | 0.27 | 0.3 | 0.46 |

TABLE VI
COMPARISON OF NSDC AND RBF KERNELS

| | NSDC | RBF |
|---|---|---|
| Convolution formulation? | $\surd$ | $\surd$ |
| Number of basis functions | Finite | Infinite |
| Centers of basis functions | Training samples | $(-\infty, \infty)$ |
| Parameters | Scaling factor $\delta$ | Scaling factor $\delta$ |
| Non-stationary? | $\surd$ | $\times$ |
| Optimization-free? | $\surd$ | $\surd$ |
| Prior knowledge needed? | $\surd$ | $\times$ |

kernels for the test sets should be similar to the confidence levels if the monitoring model can describe the normal data well. The column headed as *Other* shows the results reported in [10]–[12] for the respective multimode process monitoring methods described in each article. It can be seen that the NSDC-kernel PCA approach achieves lower MDRs than both the RBF-kernel PCA and the methods presented in [10]–[12]. For the second test case, the NSDC also achieves a smaller FAR because of the artificial outliers in the training data [11].

## V. DISCUSSION AND FUTURE WORK

### A. Comparison With the RBF Kernel and Other Methods

Table VI compares the new NSDC kernel with the RBF kernel. The main advantage of the NSDC kernel over the RBF kernel is that due to the new assumptions in the convolution kernel formulation, the NSDC kernel is nonstationary and data-dependent. Hence, the NSDC kernel can handle the nonstationary covariance caused by multimodality.

Recent works on multivariate approaches for multimodal process monitoring deal with the multimodality in an *ad hoc* way, including the locally weighted approach [11], [30], [34] and using the local statistics matrix [10] or the residuals obtained by kernel regression [12] instead of original measurements. For those methods using the data-dependent kernels, their parameters need to be optimized properly [20]. In this article, we propose a new and systematic way of formulating the nonstationary kernel function via convolution and derive the closed-form solution, i.e., the NSDC kernel. Since the NSDC kernel only requires one parameter, i.e., the kernel width to be tuned, it also avoids additional parameterization of the kernel function and makes the training and online monitoring procedure easier.

### B. Implementation Considerations

One can use the NSDC kernel for data-driven monitoring of processes with multiple operating modes, especially when the variations between modes are significant and there exist other types of nonlinearity in the processes. As shown in Fig. 5, the multimode training data need to be clustered when training the NSDC-kernel PCA model. This can be achieved either by incorporating the prior information of normal operating modes or by applying unsupervised clustering approaches, such as the $k$-means or the Dirichlet Process. Nevertheless, these clusters do not need to be precise if all data are from the normal operating conditions when applying the method for fault detection because the NSDC kernel can deal with other types of nonlinearity, as shown in Fig. 6. Moreover, the NSDC kernel can still be used when the clustering information is unavailable by setting the cluster number to 1.

To further optimize the scaling factor $\delta$, an optimization problem can be formulated based on the cross-validation approach proposed in Section III-F. On the other hand, the scaling factor in NSDC may also allow for domain knowledge that is usually available in practice. For instance, if one normal operating mode is considered to be critical such that any violation or unobserved behavior must be quickly identified as a fault at the cost of an increased number of false alarms, and then, a smaller scaling factor for this mode can be selected in order to improve the sensitivity in its neighborhood. Conversely, NSDC kernels can also be adjusted such that a variable with lower measurement reliability can be downweighted by increasing the scaling factor toward its direction in all operating modes.

## VI. CONCLUSION

This article has presented the NSDC kernel that is a novel type of the nonstationary data-dependent kernel function that is better suited for multimodal process monitoring. This NSDC kernel was defined as a covariance function by the discrete convolution on the normal data set only. The parameter specification of this NSDC kernel was also discussed. When compared to the RBF kernel under the kernel PCA framework, the NSDC kernel can yield a better monitoring model, which is robust to overfitting issues and more sensitive in fault detection. This approach directly benefits process monitoring by reducing false and missed alarm rates, as demonstrated in the industrial case study. Moreover, incipient faults seeded during operation of the industrial-scale multiphase flow facility were detected earlier using monitoring models based on the NSDC kernel. Being a data-dependent kernel that can account for process data from multiple operating modes, the NSDC kernel has only one parameter to be tuned, making it easier to apply the NSDC kernel to the data-driven process monitoring. The results in this article also suggest that use of the NSDC kernel may not be limited to multimodal process data. In general pattern recognition problems, if the training set is discrete with

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



obvious multiple operating modes or the covariance structure is nonstationary, the NSDC kernel can be combined with unsupervised clustering approaches in order to address the aforementioned issues.

## Appendix

Considering $\phi_{\boldsymbol{c}^{(i)}}(\boldsymbol{x}) = \exp(-((\boldsymbol{x}-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}-\boldsymbol{c}^{(i)}))/l^2)$ in (9), $\mathrm{cov}(y, y^*)$ will approach infinity as the number of base functions $p$ increases, given $\sigma^2$ is an arbitrary constant

$$\begin{aligned}\mathrm{cov}(y, y^*) &= \sigma^2 \lim_{p\to\infty} \sum_{i=1}^{p} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}-\boldsymbol{c}^{(i)})}{l^2}\right) \\ &\quad \times \exp\left(-\frac{(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})}{l^2}\right) \\ &\to \infty. \end{aligned} \tag{27}$$

Since

$$0 \le \exp\left(-\frac{(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})}{l^2}\right) \le 1 \tag{28}$$

the asymptotic behavior of $\mathrm{cov}(y, y^*)$ may be approximated using the triangular inequality

$$\begin{aligned} O(p) = \sum_{i=1}^{p} &\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}-\boldsymbol{c}^{(i)})}{l^2}\right) \\ &\times \exp\left(-\frac{(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})^T(\boldsymbol{x}^*-\boldsymbol{c}^{(i)})}{l^2}\right) \end{aligned} \tag{29}$$

where $O(p)$ denotes the group of functions with first-order infinity with respect to $p$. Therefore, the variance of $w_i$ is set to be $(\sigma_0^2/p)$ so as to avoid the divergence of $\mathrm{cov}(y, y^*)$.

## References


[1] S. W. Choi, C. K. Lee, J.-M. Lee, J. H. Park, and I.-B. Lee, "Fault detection and identification of nonlinear processes based on kernel PCA," *Chemometrics Intell. Lab. Syst.*, vol. 75, no. 1, pp. 55–67, Jan. 2005.

[2] Y. Hu, H. Ma, and H. Shi, "Enhanced batch process monitoring using just-in-time-learning based kernel partial least squares," *Chemometrics Intell. Lab. Syst.*, vol. 123, pp. 15–27, Apr. 2013.

[3] C. Zhao, F. Gao, and F. Wang, "Nonlinear batch process monitoring using phase-based kernel-independent component analysis-principal component analysis (KICA-PCA)," *Ind. Eng. Chem. Res.*, vol. 48, no. 20, pp. 9163–9174, 2009.

[4] D. Tax, "One-class classification," Ph.D. dissertation, Dept. Elect. Eng., Math. Comput. Sci., Dept. Intell. Syst., Delft Univ. Technol., Delft, The Netherlands, 2001.

[5] N. Li and Y. Yang, "Ensemble kernel principal component analysis for improved nonlinear process monitoring," *Ind. Eng. Chem. Res*, vol. 54, no. 1, pp. 318–329, Dec. 2015.

[6] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 460–474, Mar. 2005.

[7] H. Ma, Y. Hu, and H. Shi, "A novel local neighborhood standardization strategy and its application in fault detection of multimode processes," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 287–300, Aug. 2012.

[8] C. Zhang, X. Gao, T. Xu, and Y. Li, "Nearest neighbor difference rule–based kernel principal component analysis for fault detection in semiconductor manufacturing processes," *J. Chemometrics*, vol. 31, no. 6, 2017, Art. no. e2888.

[9] X. Peng, Y. Tang, W. Du, and F. Qian, "Online performance monitoring and modeling paradigm based on just-in-time learning and extreme learning machine for a non-Gaussian chemical process," *Ind. Eng. Chem. Res.*, vol. 56, no. 23, pp. 6671–6684, 2017.

[10] X. Deng, N. Zhong, and L. Wang, "Nonlinear multimode industrial process fault detection using modified kernel principal component analysis," *IEEE Access*, vol. 5, pp. 23121–23132, 2017.

[11] H. Li, H. Wang, and W. Fan, "Multimode process fault detection based on local density ratio-weighted support vector data description," *Ind. Eng. Chem. Res.*, vol. 56, no. 9, pp. 2475–2491, 2017.

[12] J. Yu, J. Yoo, J. Jang, J. H. Park, and S. Kim, "A novel hybrid of auto-associative kernel regression and dynamic independent component analysis for fault detection in nonlinear multimode processes," *J. Process Control*, vol. 68, pp. 129–144, Aug. 2018.

[13] D. Higdon, "A process-convolution approach to modelling temperatures in the North Atlantic Ocean," *Environ. Ecological Statist.*, vol. 5, no. 2, pp. 173–190, Jun. 1998.

[14] M. K. Singh, K. S. Venkatesh, and A. Dutta, "Kernel based approach for accurate surface estimation," *Comput. Elect. Eng.*, vol. 56, pp. 763–772, Nov. 2016.

[15] S. Garg, G. V. Steeg, and A. Galstyan, "Stochastic learning of nonstationary kernels for natural language modeling," *CoRR*, vol. abs/1801.03911, 2018. [Online]. Available: http://arxiv.org/abs/1801.03911

[16] C. J. Paciorek, "Nonstationary Gaussian processes for regression and spatial modelling," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2003.

[17] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Netw.*, vol. 12, no. 6, pp. 783–789, 1999.

[18] J.-D. Shao, G. Rong, and J. M. Lee, "Learning a data-dependent kernel function for KPCA-based nonlinear process monitoring," *Chem. Eng. Res. Des.*, vol. 87, no. 11, pp. 1471–1480, 2009.

[19] L. Luo, S. Bao, J. Mao, and D. Tang, "Nonlinear process monitoring using data-dependent kernel global-local preserving projections," *Ind. Eng. Chem. Res.*, vol. 54, no. 44, pp. 11126–11138, 2015.

[20] B. Chen, H. Liu, and Z. Bao, "Optimizing the data-dependent kernel under a unified kernel optimization framework," *Pattern Recognit.*, vol. 41, no. 6, pp. 2107–2119, 2008.

[21] A. Wilson and R. Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1067–1075.

[22] M. Gönen and E. Alpaydin, "Localized multiple kernel learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 352–359.

[23] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2627–2635.

[24] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[25] J.-M. Lee, C. K. Yoo, S. W. Choi, P. A. Vanrolleghem, and I.-B. Lee, "Nonlinear process monitoring using kernel principal component analysis," *Chem. Eng. Sci.*, vol. 59, no. 1, pp. 223–234, Jan. 2004.

[26] S. W. Choi and I.-B. Lee, "Nonlinear dynamic process monitoring based on dynamic kernel PCA," *Chem. Eng. Sci.*, vol. 59, no. 24, pp. 5897–5908, Dec. 2004.

[27] S. Abe, "Training of support vector machines with Mahalanobis kernels," in *Proc. Int. Conf. Artif. Neural Netw.* Springer, 2005, pp. 571–576.

[28] D. Wang, D. S. Yeung, and E. C. C. Tsang, "Weighted Mahalanobis distance kernels for support vector machines," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1453–1462, Sep. 2007.

[29] M. Navi, N. Meskin, and M. Davoodi, "Sensor fault detection and isolation of an industrial gas turbine using partial adaptive KPCA," *J. Process Control*, vol. 64, pp. 37–48, Apr. 2018.

[30] X. Deng and L. Wang, "Modified kernel principal component analysis using double-weighted local outlier factor and its application to nonlinear process monitoring," *ISA Trans.*, vol. 72, pp. 218–228, Jan. 2018.

[31] J.-M. Lee, S. J. Qin, and I.-B. Lee, "Fault detection of non-linear processes using kernel independent component analysis," *Can. J. Chem. Eng.*, vol. 85, no. 4, pp. 526–536, Aug. 2007.

[32] A. Stief, R. Tan, Y. Cao, J. R. Ottewill, N. F. Thornhill, and J. Baranowski, "A heterogeneous benchmark dataset for data analytics: Multiphase flow facility case study," *J. Process Control*, vol. 79, pp. 41–55, Jul. 2019.

[33] N. F. Thornhill, H. Melbø, and J. Wiik, "Multidimensional visualization and clustering of historical process data," *Ind. Eng. Chem. Res.*, vol. 45, no. 17, pp. 5971–5985, 2006.

[34] B. Song and H. Shi, "Temporal-spatial global locality projections for multimode process monitoring," *IEEE Access*, vol. 6, pp. 9740–9749, 2018.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

**Ruomu Tan** received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2013, and the M.Sc. degree from the University of Alberta, Edmonton, AB, Canada, in 2015. She is currently pursuing the Ph.D. degree with Imperial College London, London, U.K.

She is also a Marie Curie Early Stage Researcher of the Horizon 2020 EID-ITN Project PRONTO. Her current research interests include data-driven nonlinear process monitoring, multivariate statistical analysis, and their application to process industries.

**James R. Ottewill** (M'18) was born in London, U.K. He received the B.Eng. (Hons.) and Ph.D. degrees in mechanical engineering from the University of Bristol, Bristol, U.K., in 2005 and 2009, respectively.

He is currently a Senior Principal Scientist with the ABB Corporate Research Center, Kraków, Poland, where he has been involved in the field of applied analytics for condition monitoring applications. His current research interests include advanced physics-based and data-driven approaches for diagnostics and prognostics, including dynamic testing, modeling, and analysis of nonlinear systems, signal processing, and information fusion.

Dr. Ottewill is also a member of the Institution of Mechanical Engineers, U.K. He is also a Chartered Engineer in U.K.

**Nina F. Thornhill** (SM'93) received the B.A. degree in physics from Oxford University, Oxford, U.K., in 1976, the M.Sc. degree from Imperial College London, London, U.K., and the Ph.D. degree from University College London, London.

She is currently a Professor with the Department of Chemical Engineering, Imperial College London, where she also holds the ABB Chair of Process Automation.