IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 28, 2020, accepted October 20, 2020, date of publication October 28, 2020,
date of current version November 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034550

Monitoring Statistics and Tuning of Kernel Principal Component Analysis With Radial Basis Function Kernels

RUOMU TAN[1,3], JAMES R. OTTEWILL[2], (Member, IEEE), AND NINA F. THORNHILL[1], (Senior Member, IEEE)

[1]Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, U.K.
[2]Hitachi ABB Power Grids Research, 31154 Kraków, Poland
[3]ABB Corporate Research Center, 68526 Ladenburg, Germany

Corresponding author: Ruomu Tan (r.tan@imperial.ac.uk)

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant 675215-PRONTO-H2020-MSCA-ITN-2015.

ABSTRACT Kernel Principal Component Analysis (KPCA) using Radial Basis Function (RBF) kernels can capture data nonlinearity by projecting the original variable space to a high-dimensional kernel feature space and obtaining the kernel principal components. This article examines the tuning of the kernel width when using RBF kernels in KPCA, showing that inappropriate kernel widths result in RBF-KPCA being unable to capture nonlinearity present in data. The paper also considers the choice of monitoring statistics when RBF-KPCA is applied to anomaly detection. Linear PCA requires two monitoring statistics. The Hotelling's T^2 monitoring statistic detects when a sample exceeds the healthy operating range, while the Squared Prediction Error (SPE) monitoring statistic detects the case when the sample does not follow the model of the training data. The analysis in this article shows that SPE for RBF-KPCA can detect both cases. Moreover, unlike the case of linear PCA, the T^2 monitoring statistic for RBF-KPCA is non-monotonic with respect to the magnitude of the anomaly, making it not optimal as a monitoring statistic. The paper presents examples to illustrate these points. The paper also provides a detailed mathematical analysis which explains the observations from a theoretical perspective. Tuning strategies are proposed for setting the kernel width and the detection threshold of the monitoring statistic. The performance of optimally tuned RBF-KPCA for anomaly detection is demonstrated via numerical simulation and a benchmark dataset from an industrial-scale facility.

INDEX TERMS Anomaly detection, asymptotic analysis, fault detection, kernel principal component analysis, monitoring statistic, multivariate statistics.

I. INTRODUCTION

When monitoring an industrial process, anomalous data can be an indicator of faults which may cause performance degradation or may even lead to failures and unplanned shut downs. Data-driven anomaly detection determines if a data sample is anomalous when compared to healthy data available for training. However, an inaccurate description of the healthy data may result in increased false or missed alarm rates.

Various sources of nonlinearity may be present in a process for example due to valve characteristics, multimode operation or specific mass balance relationships. Kernel-based methods can account for nonlinearity [1]. They achieve a more accurate description of the healthy data and thereby improve the anomaly detection performance.

Many researchers have adopted Kernel Principal Component Analysis (KPCA) for data-driven anomaly detection [2]–[6]. Most use the Radial Basis Function (RBF) kernel. Both the kernel width of the RBF kernel and the monitoring statistics need to be specified when using RBF-KPCA. The kernel width is an important adjustable parameter which can determine how accurately the healthy training data is described, while the monitoring statistics are important for the detection of anomalies. These aspects are not independent, because monitoring statistics are functions of the kernel principal components obtained using specific kernel widths.

The associate editor coordinating the review of this manuscript and approving it for publication was Fanbiao Li.

This article gives new insights into the tuning of RBF-KPCA applied to anomaly detection. It gives recommendations for correct tuning of the kernel-width parameter. Specifically, it will be shown that RBF-KPCA leads to increased false alarms when the kernel width is exceeding small, while exceedingly large kernel widths will lead to incorrect models and missed alarms.

We will also prove that the Hotelling's T^2 and SPE (squared prediction error) monitoring statistics that are widely used in linear principal component analysis do not have the same interpretation or behaviour when used with RBF-KPCA. The SPE for PCA is sensitive only to anomalies in the residual sub-space, whereas the SPE for RBF-KPCA is sensitive to all anomalies. Moreover, the paper will demonstrate that in RBF-KPCA the T^2 statistic is not monotonic with respect to the magnitude of the anomaly. Therefore, use of the T^2 monitoring statistic in RBF-KPCA can lead to false and missed detection of anomalies. This observation may explain difficulties that other researchers have had in applying T^2 for RBF-KPCA.

Based on these findings, we will propose novel strategies for tuning RBF-KPCA and for setting the thresholds for the SPE monitoring statistic for anomaly detection in nonlinear systems. These strategies will be demonstrated for anomaly detection in experimental data from a multiphase flow facility.

The next section discusses previous work. Section III of the paper reviews the KPCA formulation using RBF kernels and compares the SPE and T^2 as monitoring statistics. It also shows the influence of the kernel width on anomaly detection. Sections IV and V derive and discuss the asymptotic behaviour of RBF kernels with large and small kernel widths and investigates the behaviour of the SPE and T^2 monitoring statistics for RBF-KPCA. Section V shows that SPE increases monotonically as the magnitude of anomalies increases, whereas T^2 does not. These findings form the basis of a strategy for tuning the RBF kernel under the RBF-KPCA framework for anomaly detection, which is given in Section VI. Section VII illustrates the issues identified in this article and applies the strategy using both synthetic and experimental data. The performance of the proposed strategy is also demonstrated in that section.

II. BACKGROUND AND CONTEXT

The kernel width of the RBF kernels is a tuning parameter for an RBF-KPCA. Kernel width is usually specified according to empirical values [7], empirical equations [8], cross-validation [9] and optimization with respect to the correct detection performance [10], [11]. Some works have compared the performance of kernel-based methods with various kernel widths empirically [12], [13]. To date, it has been shown that the RBF kernel will approach to the linear kernel when the kernel width is large [14], [15]. The influence of the kernel width on the performance of kernel-based anomaly detection methods may also be considered from a theoretical perspective. For example, [15] investigated the influence of the tuning of the RBF kernel width on the anomaly detection performance of support vector machines.

The standard Hotelling's T^2 is used as a monitoring statistic in many works [3], [4], [13], [16], [17]. Recent works have defined the SPE for KPCA in various different ways [13], [17], [18]. Recently, [13] also demonstrated that the value of the RBF kernel will approach zero when the test sample moves sufficiently far away from the training data, leading to the monitoring statistic T^2 approaching a constant value. Indicators based on a combination of T^2 and SPE [16], [19], as well as other statistics [4], [20], have also been proposed to improve the anomaly detection performance. The underlying problem motivating these developments is that in RBF-KPCA, the value of T^2 does not increase monotonically with respect to the magnitude of the anomaly. Usually, an upper control limit is adopted following the practice in linear PCA, such that an anomaly is detected when the monitoring statistic is larger than its control limit. A few works have considered both upper and lower control limits for T^2 due to the non-monotonic behaviour [9].

To summarize, the kernel width tuning and the monitoring statistic selection influence the anomaly detection performance of RBF-KPCA. This article will address these issues by analysing the behaviour of RBF-KPCA and associated monitoring statistics with respect to the kernel width. Based on the behaviour, we will make a recommendation for the monitoring statistic selection and a tuning strategy for RBF-KPCA.

III. KERNEL PCA AND RBF KERNELS

A. KERNEL PCA WITH RBF KERNELS FOR MONITORING

As formulated in [21], KPCA first projects the original variables to a new feature space, then conducts dimension reduction in the new feature space to obtain the kernel Principal Components (PCs). Assuming the normalized training dataset $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \in \mathbb{R}^{m\times n}$ includes n data samples of m variables, under the KPCA framework, X is first projected to another feature space Φ. The feature space Φ has infinite dimensions when the RBF kernel is applied. This means that a vector of measurements $\boldsymbol{x}$ undergoes a mapping to the feature space $\Phi(\boldsymbol{x})$, $\boldsymbol{x} \mapsto \Phi(\boldsymbol{x})$, where $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]$ and $\Phi(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_\infty(\boldsymbol{x})]$. The kernel features $\Phi(\boldsymbol{x})$ are the projected variables in the new feature space. Although $\Phi(\boldsymbol{x})$ cannot be calculated directly, $K \in \mathbb{R}^{n\times n}$, the matrix of dot products of $\Phi(\boldsymbol{x})$ can be obtained using the kernel function. The RBF kernel function defines the entries of the kernel matrix K:

$$\begin{aligned} K_{i,j} &= \Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j) \\ &= \phi_1(\boldsymbol{x}_i)\phi_1(\boldsymbol{x}_j) + \cdots + \phi_\infty(\boldsymbol{x}_i)\phi_\infty(\boldsymbol{x}_j) \\ &= \exp\left(-\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\sigma^2}\right) \end{aligned} \tag{1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are two data samples. σ is the kernel width of RBF kernels. The kernel function $K_{i,j}$ is the dot product of $\Phi(\boldsymbol{x}_i)$ and $\Phi(\boldsymbol{x}_j)$.

It is not guaranteed that $\Phi(X)$, the projections of X to the kernel feature space, are centered. Therefore the K matrix is centered such that $\Phi(X)$ is also centered:

$$\begin{aligned}\tilde{K} &= (\Phi(X) - \bar{\Phi})^\top \cdot (\Phi(X) - \bar{\Phi}) = \tilde{\Phi}(X)^\top \cdot \tilde{\Phi}(X) \\ &= K - \frac{1}{n}\mathbf{1}_{n\times n}K - \frac{1}{n}K\mathbf{1}_{n\times n} + \frac{1}{n^2}\mathbf{1}_{n\times n}K\mathbf{1}_{n\times n} \end{aligned} \quad (2)$$

where $\bar{\Phi} = 1/n\sum_{i=1}^{n}\Phi(\boldsymbol{x}_i)$ is the center of $\Phi(X)$. $\tilde{\Phi}(X)$ is the centered result of $\Phi(X)$. $\mathbf{1}_{n\times n}$ is an $n \times n$ matrix with all entries having value 1.

In the second step, PCA is implemented in the Φ space by applying eigenvalue decomposition to the centered kernel matrix $\tilde{K}$:

$$\tilde{K} = \boldsymbol{\alpha}^\top \Lambda^{-1}\boldsymbol{\alpha} \quad (3)$$

where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^{(1)}, \cdots, \boldsymbol{\alpha}^{(n)}\}$. The l-th eigenvector is $\boldsymbol{\alpha}^{(l)} = \{\alpha_1^{(l)}, \cdots, \alpha_n^{(l)}\} \in \mathbb{R}^{n\times 1}$ is the l-th eigenvector. $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_n\}$ where λ_l is the l-th eigenvalue. Assuming the first L kernel PCs are retained, the value of the l-th kernel PC for $\boldsymbol{x}_j$ is $y_j^{(l)}$, given by:

$$y_j^{(l)} = \boldsymbol{V}^{(l)} \cdot \tilde{\Phi}(\boldsymbol{x}_j) = \sum_{i=1}^{n}\tilde{\alpha}_i^{(l)}\tilde{K}_{i,j} \quad (4)$$

where $\tilde{\alpha}_i^{(l)}$ is the i-th entry of a normalized version of $\boldsymbol{\alpha}^{(l)}$ such that $\|\tilde{\boldsymbol{\alpha}}^{(l)}\|_2 = 1/\lambda_l$ for $l = 1, \ldots, L$ and $\boldsymbol{V}^{(l)}$ is the l-th row of the projection matrix $\boldsymbol{V}$. Although $\Phi(X)$ and $\boldsymbol{V}$ cannot be calculated in RBF-KPCA, the kernel PCs, which are the principal components of $\Phi(\boldsymbol{x})$, can be calculated explicitly using Eqn (4) with $\tilde{\boldsymbol{\alpha}}$ and $\tilde{K}$. Throughout the paper, L is chosen such that 99% of the variability in the centered kernel features $\tilde{\Phi}(\boldsymbol{x})$ is explained. The centered kernel features can be reconstructed as

$$\hat{\tilde{\Phi}}(\boldsymbol{x}) = \boldsymbol{V}^\top \cdot \boldsymbol{V} \cdot \tilde{\Phi}(\boldsymbol{x}). \quad (5)$$

The difference between $\tilde{\Phi}(\boldsymbol{x})$ and $\hat{\tilde{\Phi}}(\boldsymbol{x})$ is because of the infinite number of kernel PCs that are not retained when $\boldsymbol{V}$ is used. A significant difference between $\tilde{\Phi}(\boldsymbol{x})$ and $\hat{\tilde{\Phi}}(\boldsymbol{x})$ can indicate that the KPCA model does not adequately describe the data.

A monitoring statistic is usually defined as a function of the retained kernel PCs. For example, the T^2 statistic of the j-th sample $\boldsymbol{x}_j$ is:

$$T_j^2 = \boldsymbol{y}_j^\top \Lambda_L^{-1}\boldsymbol{y}_j = \sum_{l=1}^{L}\lambda_l^{-1}\left(\sum_{i=1}^{n}\tilde{\alpha}_i^{(l)}\tilde{K}_{i,j}\right)^2 \quad (6)$$

where $\Lambda_L = \text{diag}\{\lambda_1, \ldots, \lambda_L\}$ is a diagonal matrix with the first L eigenvalues that correspond to the first L kernel PCs. For a test sample $\boldsymbol{x}_{\text{test}}$, T^2 is calculated using the corresponding kernel PCs $\boldsymbol{y}_{\text{test}}$. The value of $\boldsymbol{y}_{\text{test}}$ is small when $\boldsymbol{x}_{\text{test}}$ is close to the center of healthy data because in Eqn (2), the projected $\Phi(\boldsymbol{x}_{\text{test}})$ will be close to $\bar{\Phi}$ and the $\tilde{K}$ will be close to zero. Hence T^2 will be small. The value of $\boldsymbol{y}_{\text{test}}$ is also small when $\boldsymbol{x}_{\text{test}}$ is anomalous, i.e. located far away from the healthy data, because the projection of $\boldsymbol{x}_{\text{test}}$ to the retained L kernel PCs will be small. T^2 will also be small in this case. The fact that T^2 is small for both cases means that it is difficult to distinguish anomalies from healthy data when using this particular monitoring statistic. This comment applies even if n kernel PCs are retained, because there will always be an infinite number of kernel PCs that are not retained in $\boldsymbol{y}_{\text{test}}$. The non-monotonic behaviour makes T^2 suboptimal as a monitoring statistic when applying RBF-KPCA.

The SPE for KPCA-based monitoring has been defined in various ways [3], [4], [9]. Here we define the SPE of $\boldsymbol{x}_j$ as the second order norm of the difference between $\tilde{\Phi}(\tilde{\boldsymbol{x}}_j)$, the centered projection of a normalized sample $\tilde{\boldsymbol{x}}_j$ in the kernel feature space, and $\hat{\tilde{\Phi}}(\tilde{\boldsymbol{x}}_j)$, the reconstruction of $\tilde{\Phi}(\tilde{\boldsymbol{x}}_j)$ using kernel PCs obtained by KPCA [2]:

$$\begin{aligned}\text{SPE}_j &= \|\tilde{\Phi}(\tilde{\boldsymbol{x}}_j) - \hat{\tilde{\Phi}}(\tilde{\boldsymbol{x}}_j)\|_2 \\ &= \|\tilde{\Phi}(\tilde{\boldsymbol{x}}_j) - \boldsymbol{V}^\top \cdot \boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j)\|_2 \\ &= \|\tilde{\Phi}(\tilde{\boldsymbol{x}}_j)\|_2 - 2(\tilde{\Phi}^\top(\tilde{\boldsymbol{x}}_j) \cdot \boldsymbol{V}^\top) \cdot (\boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j)) \\ &\quad + (\boldsymbol{V}^\top \cdot \boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j))^\top \cdot \boldsymbol{V}^\top \cdot \boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j) \\ &= \|\tilde{\Phi}(\tilde{\boldsymbol{x}}_j)\|_2 - (\boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j))^\top \cdot (\boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j)) \end{aligned} \quad (7)$$

where, as in Eqn (4), $\boldsymbol{V}$ is the projection matrix from the feature space Φ to the kernel PC space such that $\boldsymbol{V} \cdot \boldsymbol{V}^\top = I$. $\Phi_0 = 1/n\sum_{i=1}^{n}\Phi(\tilde{\boldsymbol{x}}_i)$ is the center of the projections of the training samples in the kernel feature space. Although $\tilde{\Phi}(\tilde{\boldsymbol{x}})$ cannot be obtained directly, its second order norm is:

$$\begin{aligned}\|\tilde{\Phi}(\tilde{\boldsymbol{x}}_j)\|_2 &= \Phi^\top(\tilde{\boldsymbol{x}}_j) \cdot \Phi(\tilde{\boldsymbol{x}}_j) - 2\Phi^\top(\tilde{\boldsymbol{x}}_j) \cdot \Phi_0 + \Phi_0^\top\Phi_0 \\ &= k(\boldsymbol{x}_j, \boldsymbol{x}_j) - \frac{2}{n}\sum_{i=1}^{n}k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{n^2}\sum_{i,j=1}^{n}k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= 1 - 2\bar{K}_j + \bar{K}. \end{aligned} \quad (8)$$

where $\bar{K}_j$ is the mean of the j-th row in K and $\bar{K}$ is the mean of all entries in K. The second term of Eqn (7) is the second order norm of the kernel PCs $\boldsymbol{y}_j$ of $\boldsymbol{x}_j$:

$$(\boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j))^\top \cdot (\boldsymbol{V} \cdot \tilde{\Phi}(\tilde{\boldsymbol{x}}_j)) = \boldsymbol{y}_j^\top\boldsymbol{y}_j. \quad (9)$$

Therefore, SPE for KPCA can be written explicitly:

$$\text{SPE}_j = 1 - 2\bar{K}_j + \bar{K} - \boldsymbol{y}_j^\top\boldsymbol{y}_j \quad (10)$$

For a test sample $\boldsymbol{x}_{\text{test}}$, the SPE is calculated by using the mean of the kernel vector $1/n\sum_{i=1}^{n}k(\boldsymbol{x}_i, \boldsymbol{x}_{\text{test}})$ and $\boldsymbol{y}_{\text{test}}$ as $\bar{K}_j$ and $\boldsymbol{y}_j$ in Eqn (10), respectively.

B. ILLUSTRATIVE EXAMPLES

1) THE BEHAVIOUR OF MONITORING STATISTICS

The following illustrative example compares the performance of T^2 and SPE obtained by PCA and RBF-KPCA for various types of anomalies. The example demonstrates that the roles of T^2 and SPE, respectively, in RBF-KPCA are not the same as their roles in PCA. The data sets used for training and testing are plotted in Fig. 1. The training data are generated from a linear algebraic model with white Gaussian disturbances.

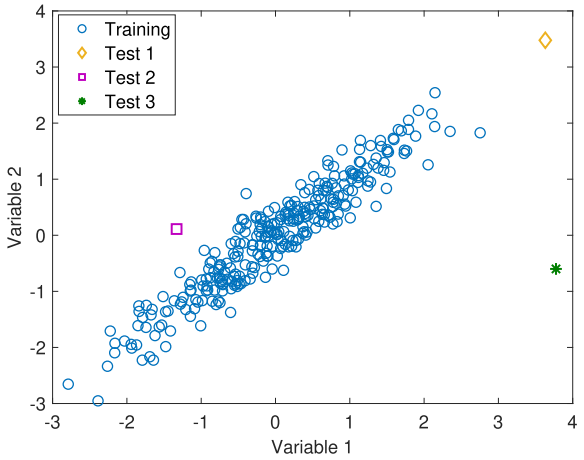


FIGURE 1. Data plot for the illustrative example.

Three test samples represent three anomalous cases. Test 1 represents the case where the linear relationship between the variables still holds, but the values exceed the healthy range. Test 2 is the case where the measurements of variable 1 and variable 2 each fall within the same range of values as the healthy case, but the relationship between the variables is not the same as in the healthy case. Test 3 combines both cases where the variables exceed the range and follow a different relationship. PCA and RBF-KPCA are applied to these data. In this example, PCA obtains two PCs. The first PC with the largest variance is retained to calculate T^2 and the second PC is used for SPE. In KPCA, kernel PCs are retained such that the percentage of the accumulated variance explained by the kernel PCs is over 99%.

TABLE 1. Monitoring statistics for illustrative example of Fig. 1.

	PCA		RBF-KPCA	
	T^2	SPE	T^2	SPE
Upper control limit	3.9116	0.1895	0.3228	0.3561
Test 1	12.9739	0.0119	0.0827	1.0455
Test 2	0.3822	1.0397	0.1397	0.9708
Test 3	2.5886	9.5703	0.0310	1.0931

Table 1 compares the T^2 and the SPE for PCA and KPCA, respectively. The upper control limit for each monitoring statistic is defined as the 95th percentile of the monitoring statistic values obtained in the training set. An anomalous sample is detected by a monitoring statistic if the value of this statistic obtained for the sample exceeds the control limit. The results demonstrate that:

1) In PCA, T^2 detects the case where the variables exceed the healthy operating range. SPE detects the case where the sample does not follow the model of the training data. PCA needs both T^2 and SPE to detect all the three anomalies.
2) T^2 for RBF-KPCA cannot detect the anomalies when using the upper control limit because the value of T^2 in Eqn (6) approaches zero when the anomaly is large.
3) In contrast, SPE for RBF-KPCA can detect all three types of anomalies.

This example has demonstrated that the T^2 and the SPE for KPCA behave differently when compared with the T^2 and the SPE for PCA. In particular, SPE as formulated in Eqn (10) is sensitive both to changes in the model of the relationships between variables and is also sensitive to samples falling outside of the range of the data from healthy operating conditions. These findings are explored mathematically and explained in Section V.

2) THE INFLUENCE OF TUNING

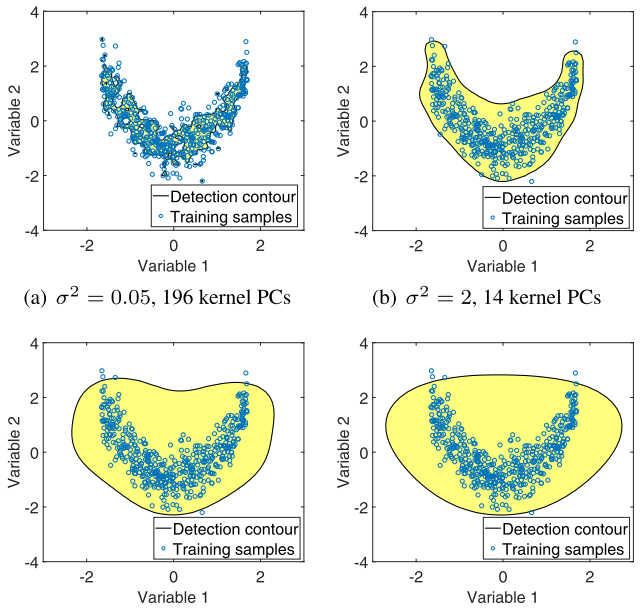
In the RBF kernel function shown in Eqn (1), the kernel width σ regulates the behaviour of the kernel function. Such behaviour will further influence the KPCA-based modelling and anomaly detection. The following nonlinear algebraic example is used to demonstrate the influence of σ:

$$\begin{aligned} x^{(1)} &\sim U[-1.5, 1.5], \\ x^{(2)} &= x^{(1)2} - 1 + e, \\ \text{where} \quad e &\sim N(0, 0.3) \end{aligned} \tag{11}$$

The first variable $x^{(1)}$ follows a uniform distribution in order to represent a process variable that is bounded. The relationship between $x^{(1)}$ and $x^{(2)}$ is quadratic, resulting in a nonlinear dataset. The second variable $x^{(2)}$ has Gaussian noise that represents the measurement noise that is often found in real-life process data. The training dataset contains 500 samples randomly generated using Eqn (11). Various σ values are used to train the KPCA model on the dataset. When applied to anomaly detection, the SPE is calculated for the samples. The upper control limit of SPE is used such that a data sample with the SPE value exceeding the control limit is detected as an anomaly. For this two-dimensional problem, it is possible to visualize the detection contours obtained by selecting the control limit of the SPE as the 99th percentile of the SPE values obtained on the training data and connecting the points at which the SPE reaches its control limit for each KPCA model.

The shaded area in each figure shows the healthy range and samples in the white area are anomalous. Fig. 2(a) shows that a small σ value yields an over-fitted model, where 'over-fitted' means that the contour also captures the noise existing in the training data. New data generated by the same model in Eqn (11) will be detected as anomalies due to the existence of noise. Larger σ values will result in relaxed detection contours. However, when σ is too large, Fig. 2(d) shows that the under-fitted detection contour loses its ability to capture the nonlinear profile of the data.

To summarize, the tuning of σ influences the performance of RBF-KPCA and the SPE for anomaly detection. In the following sections we will investigate the influence of σ and the behaviour of monitoring statistics through



(a) $\sigma^2 = 0.05$, 196 kernel PCs (b) $\sigma^2 = 2$, 14 kernel PCs

(c) $\sigma^2 = 20$, 5 kernel PCs (d) $\sigma^2 = 20000$, 2 kernel PCs

FIGURE 2. SPE detection contours of RBF-KPCA with various kernel widths. Yellow-shaded areas: range of values classified as healthy.

asymptotic analysis. Moreover, we will analyse and explain why SPE (Eqn (10)) and not T^2 (Eqn (6)) should be used as the monitoring statistic for RBF-KPCA applications.

IV. ASYMPTOTIC BEHAVIOUR OF RBF KERNELS

This section investigates the asymptotic behaviour of the RBF kernels, both when the σ value tends to an exceedingly large value and conversely when it tends to an exceedingly small value. We assume the kernel width $\sigma > 0$ throughout this article.

A. EXCEEDINGLY LARGE KERNEL WIDTH ($\sigma \to \infty$)

This section shows that when σ is large, the centered RBF kernel matrix is a scaled version of the centered kernel matrix obtained by a linear kernel. The reason for comparing centered kernel matrices is that, as mentioned in Section III-A, the PCA step is applied to the centered kernel matrix $\tilde{K}$ in KPCA.

1) TRAINING THE MONITORING MODEL

The i,j-th entry of the kernel matrix K is:

$$\begin{aligned} K_{i,j} &= \lim_{\sigma^2\to\infty} \exp\left(-\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\sigma^2}\right) \\ &= \lim_{\sigma^2\to\infty}\left[1-\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\sigma^2} \right. \\ &\quad \left. + o\left(\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\sigma^2}\right)\right] \\ &\approx 1-\frac{1}{\sigma^2}\left(\boldsymbol{x}_i^\top\boldsymbol{x}_i+\boldsymbol{x}_j^\top\boldsymbol{x}_j-2\boldsymbol{x}_i^\top\boldsymbol{x}_j\right) \end{aligned} \tag{12}$$

In this article, σ is considered exceedingly large when $o\left((\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)/\sigma^2\right) \approx 0$ for all $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the training dataset. Eqn (12) keeps the second order terms after the Taylor expansion because the result in Eqn (12) is a more accurate estimation of $K_{i,j}$ than assuming $K_{i,j}$ as a constant. Moreover, the kernel width σ cannot be infinity in practice. A test sample $\boldsymbol{x}_{\text{test}}$ may deviate significantly from the training data, making the magnitude of $\boldsymbol{x}_{\text{test}}$ comparable with σ. In this case, the second order terms cannot be neglected. Section V investigates the case when $\boldsymbol{x}_{\text{test}}$ as well as σ approaches infinity.

The kernel matrix K will be centered using Eqn (2). The i,j-th entry of $\tilde{K}$ is therefore:

$$\tilde{K}_{i,j} = K_{i,j} - \bar{K}_{i,\text{row}} - \bar{K}_{\text{col},j} + \bar{K} \tag{13}$$

In Eqn (12), $\bar{K}_{i,\text{row}}$ and $\bar{K}_{\text{col},j}$ are the means of i-th row and j-th column of K, respectively.

$$\begin{aligned} \bar{K}_{i,\text{row}} &= \frac{1}{n}\sum_{j=1}^{n} K_{i,j} \\ &= \frac{1}{n}\sum_{j=1}^{n}\left[1-\frac{1}{\sigma^2}\left(\boldsymbol{x}_i^\top\boldsymbol{x}_i+\boldsymbol{x}_j^\top\boldsymbol{x}_j-2\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)\right] \\ &= 1-\frac{1}{\sigma^2}\boldsymbol{x}_i^\top\boldsymbol{x}_i-\frac{1}{n\sigma^2}\sum_{j=1}^{n}\boldsymbol{x}_j^\top\boldsymbol{x}_j+\frac{2}{n\sigma^2}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j \end{aligned} \tag{14}$$

$$\bar{K}_{\text{col},j} = 1-\frac{1}{\sigma^2}\boldsymbol{x}_j^\top\boldsymbol{x}_j-\frac{1}{n\sigma^2}\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_i+\frac{2}{n\sigma^2}\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j \tag{15}$$

$\bar{K}$ is the mean of all entries of K:

$$\bar{K} = 1-\frac{2}{n\sigma^2}\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_i+\frac{2}{n^2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j. \tag{16}$$

Hence the centered kernel matrix $\tilde{K}$ has the following entry:

$$\begin{aligned} \tilde{K}_{i,j} = \frac{2}{\sigma^2}\left[\boldsymbol{x}_i^\top\boldsymbol{x}_j-\frac{1}{n}\left(\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j+\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)\right. \\ \left. +\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j\right]. \end{aligned} \tag{17}$$

As shown in [22], KPCA with a linear kernel, defined as $K_{\text{lin}}(i,j) = \boldsymbol{x}_i^\top\boldsymbol{x}_j$, will reduce to the ordinary linear PCA. The centered linear kernel matrix is defined as:

$$\begin{aligned} \tilde{K}_{\text{lin},i,j} = \boldsymbol{x}_i^\top\boldsymbol{x}_j-\frac{1}{n}\left(\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j+\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j\right) \\ +\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j. \end{aligned} \tag{18}$$

A comparison of Eqns (17) and (18) shows that $\tilde{K}_{i,j} = 2\sigma^{-2}\tilde{K}_{\text{lin},i,j}$. Thus, when σ is large, the RBF kernel will generate a centered kernel matrix whose entries are proportional to the entries of the centered kernel matrix obtained by the linear kernel. Hence the eigenvectors and eigenvalues of the

kernel matrix obtained by the RBF kernel will be proportional to those of the linear kernel matrix. Moreover, when σ is exceedingly large, the number of kernel PCs of the illustrative example in Fig. 2 has already reduced to two because $L = 2$ given the criterion of 99% of variability, which is the same as the linear PCA result. This explains the behaviour in Fig. 2(d).

2) FOR A TEST SAMPLE x_{test}

The test kernel vector $K_{i,\text{test}}$ of $\boldsymbol{x}_{\text{test}}$ is:

$$K_{i,\text{test}} = \lim_{\sigma^2\to\infty} \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_{\text{test}})^\top(\boldsymbol{x}_i - \boldsymbol{x}_{\text{test}})}{\sigma^2}\right) \approx 1 - \frac{1}{\sigma^2}\left(\boldsymbol{x}_i^\top\boldsymbol{x}_i + \boldsymbol{x}_{\text{test}}^\top\boldsymbol{x}_{\text{test}} - 2\boldsymbol{x}_i^\top\boldsymbol{x}_{\text{test}}\right). \quad (19)$$

The centered value $\tilde{K}_{i,\text{test}}$ is:

$$\tilde{K}_{i,\text{test}} = \frac{2}{\sigma^2}\left(\boldsymbol{x}_i^\top\boldsymbol{x}_{\text{test}} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^\top x_{\text{test}} - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j^\top + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j\right). \quad (20)$$

For the linear kernel,

$$K_{\text{lin},i,\text{test}} = \boldsymbol{x}_i^\top\boldsymbol{x}_{\text{test}}. \quad (21)$$

The centered kernel vector $\tilde{K}_{\text{lin},i,\text{test}}$ is:

$$\tilde{K}_{\text{lin},i,\text{test}} = \boldsymbol{x}_i^\top\boldsymbol{x}_{\text{test}} - \frac{1}{n}\left(\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_{\text{test}} + \sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{x}_j. \quad (22)$$

Therefore, $\tilde{K}_{i,\text{test}} = 2\sigma^{-2}\tilde{K}_{\text{lin},i,\text{test}}$. To conclude, the RBF kernel will result in centered kernel matrices and kernel vectors that are proportional to the equivalent centered kernel matrix and the equivalent kernel vector obtained by a linear kernel when σ is exceedingly large relative to the training dataset X_{train} and the test sample $\boldsymbol{x}_{\text{test}}$. The behaviour of RBF-KPCA when $\boldsymbol{x}_{\text{test}}$ also approaches infinity will be investigated later.

B. EXCEEDINGLY SMALL KERNEL WIDTH ($\sigma^2 \to 0$)

1) TRAINING THE MONITORING MODEL

On the other hand, the value of σ is considered exceedingly small when the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ will reduce to the Kronecker Delta (δ) function:

$$K_{i,j} = \lim_{\sigma^2\to 0}\exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\sigma^2}\right) = \begin{cases}1, & i = j\\ 0, & i \neq j.\end{cases} \quad (23)$$

This results in the kernel matrix K being an $n \times n$ identity matrix $I_{n\times n}$. The centered kernel matrix $\tilde{K}$ then becomes:

$$\tilde{K} = \begin{bmatrix} 1-1/n & -1/n & -1/n & \dots & -1/n \\ -1/n & 1-1/n & -1/n & \dots & -1/n \\ \dots\dots & & & & \\ -1/n & -1/n & -1/n & \dots & 1-1/n \end{bmatrix}. \quad (24)$$

which has $n-1$ eigenvalues $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = 1$ and one eigenvalue $\lambda_n = 0$. The first $n-1$ normalized eigenvectors satisfy the following condition:

$$\sum_{i=1}^{n}\tilde{\alpha}_i^{(l)} = 0 \quad \text{for } l = [1, 2, \cdots, n-1]. \quad (25)$$

2) FOR A NEW SAMPLE x_{test}

According to Eqn (23), $K_{i,\text{test}} = 0$ if $\boldsymbol{x}_{\text{test}} \notin X_{\text{train}}$ for all i. The centered value $\tilde{K}_{i,\text{test}}$ is:

$$\tilde{K}_{i,\text{test}} = K_{i,\text{test}} - \bar{K}_{i,\text{test}} - \bar{K}_{i,\text{row}} + \bar{K} = 0 \quad (26)$$

where $\bar{K}_{i,\text{row}} = \bar{K} = 1/n$. Therefore, both the kernel vector $K_{i,\text{test}}$ and centered kernel vector $\tilde{K}_{i,\text{test}}$ are zero vectors. Fig. 2(a) is an example of an over-fitted model caused by σ being set too small. In the extreme case of over-fitting, the detection contour will shrink into a Dirac measure of the training set X_{train} in the variable space. In other words, any test sample that is not identical to a sample in the training set will be detected as an anomaly.

V. BEHAVIOUR OF MONITORING STATISTICS IN RBF-KPCA

This section demonstrates why the SPE defined by Eqn (10) is a good choice for a general-purpose single monitoring statistic for RBF-KPCA. The calculation of SPE given by Eqns (7)-(10) is a quadratic measure of the mismatch between the infinite number of features and their reconstructed results after applying PCA to the feature space. As discussed in Section III-A, a large reconstruction error between $\tilde{\Phi}(\tilde{\boldsymbol{x}})$ and $\hat{\tilde{\Phi}}(\tilde{\boldsymbol{x}})$ indicates that the KPCA model no longer applies. SPE increases monotonically as the reconstruction error increases.

Other than for the over-fitted case when σ is too small, the SPE for RBF-KPCA increases monotonically with respect to the magnitude of anomalies. Therefore, the anomalies can be detected using the SPE for RBF-KPCA and its upper control limit. Nevertheless, T^2 has been widely used in the literature as a monitoring statistic in RBF-KPCA. Section V-B will analyse and explain the properties of T^2. Its non-monotonic behaviour explains the unsatisfactory detection performance of T^2 for KPCA. Moreover, some of the adjustments that previous authors have made to adapt T^2 as a monitoring statistic for RBF-KPCA, such as the need for both upper and lower control limits [9], can also be explained.

A. BEHAVIOUR OF SPE FOR RBF-KPCA

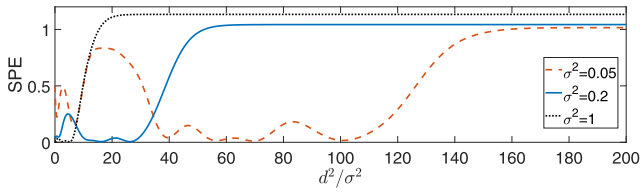
As previously given in Eqn (7), we define the SPE as the difference between the kernel features $\tilde{\Phi}(\tilde{\boldsymbol{x}})$ and the

reconstructed $\hat{\tilde{\Phi}}(\tilde{\boldsymbol{x}})$ after applying PCA in the Φ space. When using the RBF kernel, the following limit of SPE exists when $\boldsymbol{x}_{\text{test}} \to \infty$:

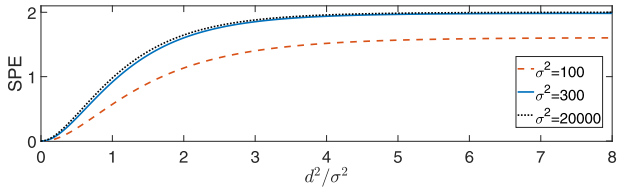
$$\begin{aligned} \text{SPE}_{\text{test,lim}} &= 1 - 2\bar{K}_{i,\text{test}}^{(\text{lim})} + \bar{K} - \boldsymbol{y}_{\text{test}}^{(\text{lim})\top}\boldsymbol{y}_{\text{test}}^{(\text{lim})} \\ &= 1 + \bar{K} - \boldsymbol{y}_{\text{test}}^{(\text{lim})\top}\boldsymbol{y}_{\text{test}}^{(\text{lim})} \\ \text{s.t.} \quad \boldsymbol{y}_{\text{test}}^{(\text{lim})\top}\boldsymbol{y}_{\text{test}}^{(\text{lim})} &= \sum_{l=1}^{L}\left(\sum_{i=1}^{n}\tilde{\alpha}_i^{(l)}\left[\bar{K} - \bar{K}_{i,\text{row}}\right]\right)^2 \end{aligned} \tag{27}$$

where $\bar{K}_{i,\text{test}}^{(\text{lim})} = 0$. Since SPE converges to a non-zero finite value when $\boldsymbol{x}_{\text{test}}$ approaches infinity, SPE cannot be χ^2 distributed because a χ^2-distributed random variable ranges from zero to infinity. Therefore, unlike in ordinary PCA, the control limits for T^2 and SPE should not be set according to the χ^2 distribution.

The illustrative example presented in Fig. 1 and Table 1 has demonstrated that the SPE in RBF-KPCA can detect model mismatch and violation of the healthy range. The reason is that, when the RBF kernel is used, the higher-dimensional kernel PCs are supposed to be a comprehensive description of the training data in the original variable space since these kernel PCs are obtained such that the reconstruction error is minimized for the training data. Therefore, the process model and the feasible range are learned simultaneously.



(a) Small σ values



(b) Large σ values

FIGURE 3. SPE with respect to d^2/σ^2. $\sigma^2 = 0.05$: over-fitted model. $\sigma^2 = 20000$: linear model.

The SPE can distinguish between the case where the test sample is located close to the center of the training data and the case where the test sample is located far away from the training data. To demonstrate the behaviour, Fig. 3 shows the trends of SPE with respect to d^2/σ^2 given σ values, where $d = \sqrt{\|\boldsymbol{x}\|_2}$ is the Euclidean distance between a data sample and the origin in the variable space. The quantity d represents the distance between a data sample and the normalized training dataset. When σ is extremely small, the RBF-KPCA model is over-fitted to the data in the training set, as shown in Fig. 2(a). In this situation, as may be expected, the SPE has non-monotonic behaviour because any new data point that is in between the training samples is considered as anomalous. For larger values of σ, the SPE increases monotonically as d increases, indicating the sample $\boldsymbol{x}$ deviates from the training data. It is also necessary to notice that, although the SPE for KPCA increases monotonically when σ is extremely large, e.g. in Fig. 3(b), the σ values may lead to under-fitted models that cannot capture the data nonlinearity and, as a result, such σ values should be avoided.

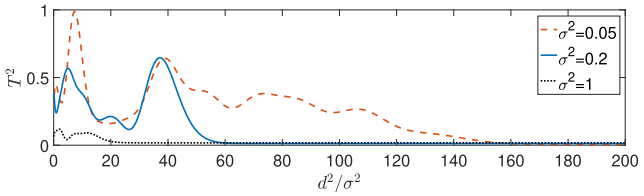
The criterion for anomaly detection using SPE is:

$$\text{SPE}_{\text{test}} > \text{SPE}_{\text{UCL}} \tag{28}$$

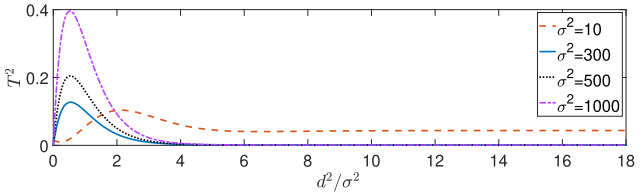
where SPE_{UCL} is the upper control limit of SPE. This value may be set according to the training data.

B. BEHAVIOUR OF T^2

The T^2 statistic defined by Eqn (6) is suitable for PCA-based anomaly detection because it increases as a test data sample moves away from the training set and an anomaly is detected if T^2 exceeds its upper control limit. However, in RBF-KPCA T^2 cannot be monotonic, as highlighted in Section III-A. In this section we investigate the behaviour of T^2 with respect to both σ and $\boldsymbol{x}_{\text{test}}$. Fig. 4 shows the trends of T^2 with respect to d^2/σ^2 given σ values for the illustrative example of Fig. 2. It is evident that the T^2 statistics do not increase monotonically as d increases for any choice of σ.



(a) Small σ values



(b) Large σ values

FIGURE 4. T^2 with respect to d^2/σ^2. $\sigma^2 = 0.05$: over-fitted model. $\sigma^2 = 10000$: linear model.

It is necessary for a statistic to have different values for healthy samples than for anomalous samples. However, in the case of RBF-KPCA, the non-monotonicity of T^2 exists in all curves in Fig. 4. In all curves in Fig. 4, T^2 has a relatively small value when d^2/σ^2 is small, i.e. when the test sample is located among the healthy data. However, as d increases, T^2 becomes small again when the test sample is far away from the healthy data and the anomaly is significant. Therefore, T^2 is non-monotonic no matter if the monitoring model is over-fitted, is under-fitted, or is properly fitted. Such non-monotonic behaviour makes it impossible to establish if the

test sample is healthy or anomalous on the sole basis of a small value of T^2. Such non-monotonic behaviour can be demonstrated by mathematical analysis.

1) UPPER BOUND

For an arbitrary test sample $\boldsymbol{x}_{\text{test}} \in \mathbb{R}^{r\times 1}$, the monitoring statistic T^2 is calculated as follows using the centered kernel vector $\tilde{K}_{\text{test}}$ and the eigenvectors obtained from the training data.

$$\begin{aligned} T^2_{\text{test}} &= \sum_{l=1}^{L} \lambda_l^{-1} \left(\sum_{i=1}^{n} \tilde{\alpha}_i^{(l)} \tilde{K}_{i,\text{test}} \right)^2 \\ &\le \sum_{l=1}^{L} \lambda_l^{-1} \left[\sum_{i=1}^{n} \tilde{\alpha}_i^{(l)2} \tilde{K}_{i,\text{test}}^2 + \sum_{i\neq j}^{n} \left| \tilde{\alpha}_i^{(l)} \tilde{\alpha}_j^{(l)} \right| \left| \tilde{K}_{i,\text{test}} \tilde{K}_{j,\text{test}} \right| \right] \end{aligned} \tag{29}$$

where $\tilde{K}_{i,\text{test}}$ and $\tilde{K}_{j,\text{test}}$ are the i-th and j-th entry of the centered kernel vector $\tilde{K}_{\text{test}}$, respectively. $\tilde{\alpha}_i^{(l)}$ and $\tilde{\alpha}_i^{(l)}$ are i-th and j-th entry of the l-th normalized eigenvector $\tilde{\boldsymbol{\alpha}}^{(l)}$, respectively.

The following inequalities hold when using the RBF kernel:

$$\begin{aligned} 0 \le K_{i,\text{test}} \le 1, \quad 0 \le \bar{K}_{i,\text{row}} \le 1, \\ 0 \le \bar{K}_{i,\text{test}} \le 1, \quad 0 \le \bar{K} \le 1. \end{aligned}$$

As a result, the range of $\left|\tilde{K}_{i,\text{test}}\right|$ may be given as:

$$0 \le \left|\tilde{K}_{i,\text{test}}\right| = \left|K_{i,\text{test}} - \bar{K}_{i,\text{row}} - \bar{K}_{i,\text{test}} + \bar{K}\right| \le 2. \tag{30}$$

The upper bound of T^2_{test} is:

$$\begin{aligned} T^2_{\text{test}} &\le \sum_{l=1}^{L} \lambda_l^{-1} \left[\sum_{i=1}^{n} 4\tilde{\alpha}_i^{(l)2} + \sum_{i\neq k}^{n} 4 \left| \tilde{\alpha}_i^{(l)} \tilde{\alpha}_k^{(l)} \right| \right] \\ &= 4 \sum_{l=1}^{L} \lambda_l^{-1} \left[\sum_{i=1}^{n} \sum_{k=1}^{n} \left| \tilde{\alpha}_i^{(l)} \tilde{\alpha}_i^{(k)} \right| \right] \end{aligned} \tag{31}$$

which is dependent only on the $\tilde{\boldsymbol{\alpha}}$s and λs obtained in the training procedure. Eqn (31) shows that the monitoring statistic T^2 of all possible samples has an upper bound when the training data and the kernel width are both fixed.

2) LARGE x_{test}

This section examines the extreme case of anomalies, i.e. the $\boldsymbol{x}_{\text{test}}$ deviates significantly from the training data. Assuming a test sample $\boldsymbol{x}_{\text{test}}^{(\text{lim})}$ has sufficiently large distances from all training samples such that:

$$\begin{aligned} K_{i,\text{test}}^{(\text{lim})} &= k(\boldsymbol{x}_{\text{test}}^{(\text{lim})}, \boldsymbol{x}_i) \\ &= \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_{\text{test}}^{(\text{lim})})^\top (\boldsymbol{x}_i - \boldsymbol{x}_{\text{test}}^{(\text{lim})})}{\sigma^2} \right) = 0 \\ &\text{for } i = [1, 2, \ldots, n], \end{aligned} \tag{32}$$

the centered kernel vector of this test sample is:

$$\begin{aligned} \tilde{K}_{i,\text{test}}^{(\text{lim})} &= K_{i,\text{test}}^{(\text{lim})} - \bar{K}_{i,\text{tow}}^{(\text{lim})} - \bar{K}_{i,\text{test}} + \bar{K} \\ &= \bar{K} - \bar{K}_{i,\text{row}} \end{aligned} \tag{33}$$

where $K_{i,\text{test}}^{(\text{lim})} = \bar{K}_{i,\text{test}}^{(\text{lim})} = 0$.

The T^2 statistic in this case becomes:

$$\begin{aligned} T^2_{\text{test,lim}} &= \sum_{l=1}^{L} \lambda_l^{-1} \left[\sum_{i=1}^{n} \tilde{\alpha}_i^{(l)} \tilde{K}_{i,\text{test}}^{(\text{lim})} \right]^2 \\ &= \sum_{l=1}^{L} \lambda_l^{-1} \left(\sum_{i=1}^{n} \tilde{\alpha}_i^{(l)} \left[\bar{K} - \bar{K}_{i,\text{row}} \right] \right)^2 \end{aligned} \tag{34}$$

which is a constant when the kernel matrix K of the training data is known. It can be seen that the monitoring statistic T^2 will converge to this constant value when the test sample deviates significantly from the training samples.

3) LARGE σ AND LARGE x_{test}

Fig. 5 shows the contour plot of T^2 when zooming out Fig. 2(c) to a larger scale. It shows that T^2 is non-monotonic in all directions as d increases. Fig. 4(b) further suggests that a common turning point of T^2 exists when the σ is exceedingly large.

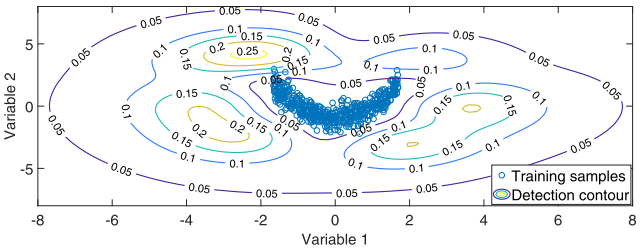


FIGURE 5. T^2 contour for the illustrative example when $\sigma^2 = 1$.

For a given large σ value, the common turning point of

$$\hat{d}_{\text{test}} = \sigma/\sqrt{2} \tag{35}$$

exists for T^2 statistics when the nonlinear part in RBF kernels becomes dominant. A detailed derivation of the turning point is given in the Appendix for the case where $\boldsymbol{x}_{\text{test}}$ is one-dimensional and $d_{\text{test}} = |x_{\text{test}}|$. The turning point is in accordance with the observation in Fig. 4(b): for σ^2 =300, 500 and 1000, the turning points of T^2 are all at $d^2/\sigma^2 = 0.5$.

4) LIMITATION OF T^2 AS A MONITORING STATISTIC

The main issue of T^2 as a monitoring statistic in RBF-KPCA is that it is non-monotonic. When the σ value is too large, T^2 will firstly increase then will decrease as $\boldsymbol{x}_{\text{test}}$ moves away from the training data according to Fig. 4(b). The non-monotonic behaviour indicates that, if the upper control limit of T^2 is selected such that a fault is detected if $T^2 > T^2_{\text{UCL}}$, larger anomalies may be missed. A value of σ may exist such that T^2 is monotonically decreasing when the test sample moves away from the training data. A lower control limit of

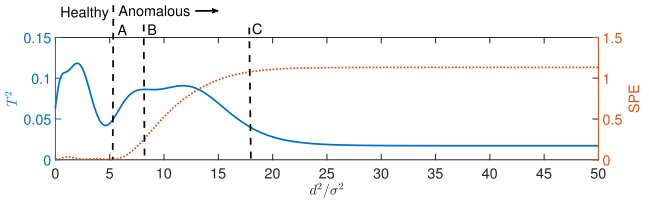


FIGURE 6. Asymptotic behaviour of SPE and T^2 when $\sigma^2 = 1$. A: the boundary of healthy behaviour. B: SPE detects an anomaly using an upper control limit of 0.25. C: T^2 detects an anomaly using a lower control limit of 0.04.

T^2 could be used and a fault is detected when $T^2 < T^2_{\text{LCL}}$. However, since T^2 can also be small when the test samples are close to the center of the training data (left side of the curves in Fig. 4(a)), the lower control limit will identify these test samples as anomalies while they are within the training dataset, leading to the small detection contours in Fig. 4(b) and increased false alarms. In addition to having clear practical implications, such false alarms can lead to misleading results when tuning the kernel width σ using empirical approaches.

Fig. 6 compares the T^2 and the SPE with respect to d^2/σ^2 when $\sigma^2 = 1$. The training data are in the range $d^2/\sigma^2 \leq 5.6$, shown by line A. SPE is close to zero when the test sample is within the range of the training data and rises as the test sample moves away from the training data. An SPE control limit of 0.25 detects an anomaly when $d^2/\sigma^2 \geq 8.2$, as shown by line B. The value of T^2 is not monotonic, as is also evident in Fig. 4. As discussed above, T^2 could be used with a lower control limit. A lower control limit of 0.04 would detect a test sample as anomalous when $d^2/\sigma^2 \geq 18.6$, as shown by line C. If the lower control limit for T^2 were higher, for example 0.075, then anomalies could be detected earlier, but then the healthy data within the range of the training data would also be detected as anomalous.

It may be observed that T^2 has low values both when a test sample is located close to the center of the training data ($d^2 \to 0$) and when a test sample is located far from the training data ($d^2 \to \infty$). The value of T^2 in the latter case can be calculated using Eqn (34). In contrast the SPE is low when the test sample is close to the training data ($d^2 \to 0$) and rises as the test sample moves away from the training data ($d^2 \to \infty$).

VI. TUNING STRATEGY FOR RBF-KPCA

Previous sections have demonstrated that the tuning of the kernel width σ influences the performance of RBF-KPCA. When training the RBF-KPCA model for anomaly detection, the dataset used for training is usually assumed to be from healthy operations, containing no samples that may be considered as anomalous. A cross-validation approach for tuning σ divides the data from healthy operations into training and validation sets. The RBF-KPCA model with various initial guesses of σ are trained on the training set and the σ which achieves a low false alarm rate on the cross-validation set is chosen as the appropriate σ_{opt}. However, this approach may not be sufficient for RBF-KPCA. In Fig. 2(d), the small number of training samples lying outside the detection contour indicates that, even when σ is inappropriate, the number of alarms triggered on the original dataset because of the mismatch between model and data can still be small. Hence the cross-validation approach may not tune the σ correctly if the initial guesses of σ are in an incorrect range. Therefore, this section proposes the strategy for tuning the kernel width σ in RBF-KPCA which combines the estimation of σ based on the previous analysis and the cross-validation approach.

A. MAXIMUM VALUE OF σ

It is important to avoid a too large σ value as large σ values may impact the ability of RBF-KPCA to capture data nonlinearity (e.g. Fig. 2(d)). Therefore, an upper bound of σ is important. According to Eqn (53) given in the Appendix, it is possible to estimate the maximum value of σ by the following empirical equation:

$$\sigma_{\max} = \sqrt{2} d_{\text{train,max}} \tag{36}$$

where $d_{\text{train,max}}$ is the maximum distance defined from the training set, i.e. $d_{\text{train,max}} = \max \sqrt{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2}$ for $\boldsymbol{x}_i, \boldsymbol{x}_j \in X_{\text{train}}$. A criterion for a maximum value of σ is required such that the RBF kernels in Eqn (1) are sufficiently localized without being over-fitted. This can be achieved if the values of $\boldsymbol{y}_{\text{test}}$ from Eqn (4) decrease monotonically when the test sample $\boldsymbol{x}_{\text{test}}$ is located outside the training data set. This is achieved for the same value of d/σ as in Eqn (35). Setting $\sigma_{\max}$ such that the largest distance between the training samples ($d_{\text{train,max}}$) can be accounted for leads to Eqn (36).

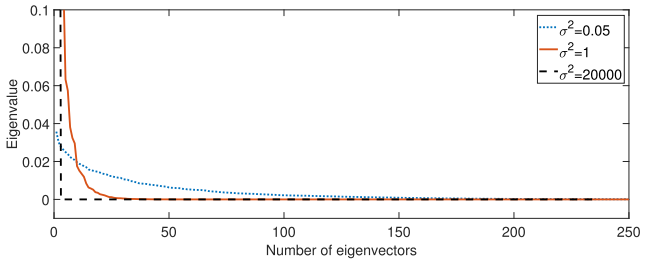
B. SUMMARY OF THE TUNING STRATEGY

After $\sigma_{\max}$ is estimated, the appropriate σ value will be determined by the cross-validation performance. Eqn (27) shows that the kernel PCs converge to finite values, indicating that the kernel PCs cannot be Gaussian distributed. Thus the analytic form for the distribution of SPE for RBF-KPCA is not known. Therefore, it is recommended to use a percentile of the SPEs of the training data as the control limit of SPE with a certain confidence level.

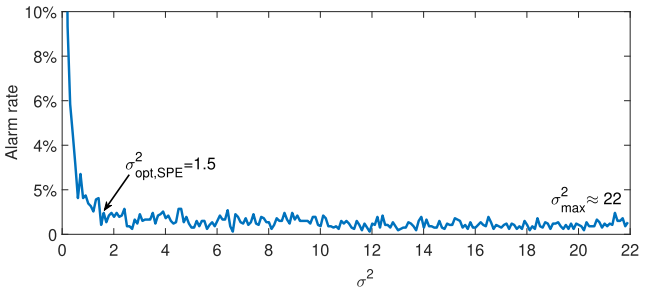
The strategy for tuning the kernel width in RBF-KPCA is summarized as:

1. Estimate the upper limit of σ by $\sigma_{\max} = \sqrt{2} d_{\text{train,max}}$.
2. Enumerate between $\sigma = 0$ and $\sigma = \sigma_{\max}$ to get the alarm rates on the cross-validation set. In this step a provisional control limit of a monitoring statistic is set so that the alarm rate is expected to be minimized (usually the maximum value of the monitoring statistic of the training set).
3. Set the smallest σ that leads to an acceptable level of alarm rates on the cross-validation set as σ_{opt}.
4. Specify the final control limit of the monitoring statistic for anomaly detection.

For the illustrative example in Fig. 2, Fig 7(a) compares the eigenspectra obtained when σ^2 is equal to 0.05, 1, and 20000, respectively. It may be observed that when σ is selected



(a) Eigenspectra of the illustrative example



(b) Alarm rates and $\sigma^2_{\text{opt,SPE}}$

FIGURE 7. Eigenspectra and alarm rates of the illustrative example.

properly, such as $\sigma^2 = 1$, the eigenvalues have a rapid trend of decreasing, resulting in a small number of kernel PCs. Fig. 7(b) shows the alarm rates with respect to σ^2 and the optimal σ value estimated when using the SPE as the monitoring statistic.

VII. EXAMPLES

A. NUMERICAL SIMULATION

The first example is based on the illustrative example described in Section III-B. In addition to the qualitative detection contours presented in Fig. 2, the anomaly detection performance of RBF-KPCA with various σ values will be quantitatively compared. Set 1 is the healthy set with 500 samples in the illustrative example. It is randomly divided into training and cross-validation sets with 250 samples in each set. The σ value is tuned using the training and the cross-validation sets by the strategy proposed in Section VI-B, as shown in Fig. 7(b). The confidence level of control limits is set as 1% of the monitoring statistic values obtained on the training set. The RBF-KPCA anomaly detection model is trained accordingly using Set 1. Set 2 comprises another 500 healthy samples generated using Eqn (11). The performance of the RBF-KPCA model on Set 2 is used to evaluate the robustness of the RBF-KPCA approach. Set 3 is an anomalous data set of 394 samples used for validating the fault detection performance. The blue circles and the red crosses in Fig. 8 represent Set 2 and Set 3, respectively. An anomaly detection approach should be able to identify the samples in Set 2 as healthy data and detect the samples in Set 3 as anomalies.

In order to demonstrate the influence of kernel widths, a variety of σ values are used. We select $\sigma^2 = 0.2, 1.5, 5, 10$ and 100 for the RBF-KPCA and the SPE as the monitoring statistic. The detection contours generated by the upper control limits of SPE are compared in Fig. 8. The contour obtained by PCA is also visualized (denoted as "Linear" in Fig. 7(b)). Fig. 8 shows that $\sigma^2_{\text{opt,SPE}} = 1.5$ can generate a good detection contour while the contour is over-fitted when σ is smaller than the optimal value ($\sigma^2 = 0.2$), and the contour becomes loose when σ increases ($\sigma^2 = 10$).

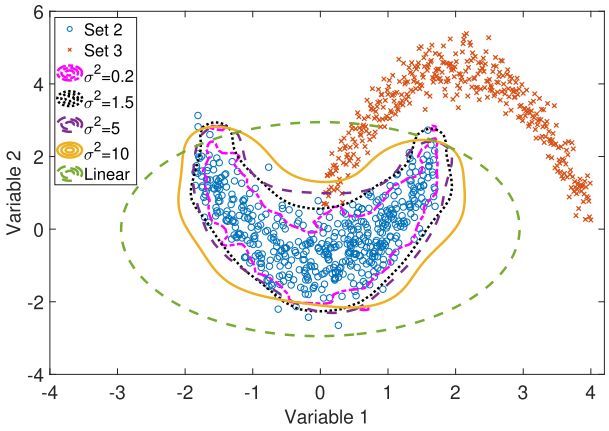


FIGURE 8. Detection contours obtained by RBF-KPCA and PCA.

For quantitative comparison, the False Alarm (FA) rate for Set 2 and the Missed Alarm (MA) rate for Set 3 are defined as:

$$\text{FA} = \frac{N_{\text{AD}}}{N_{\text{Set2}}}, \quad \text{MA} = \frac{N_{\text{ND}}}{N_{\text{Set3}}} \tag{37}$$

where N_{AD} is the number of anomalies detected in Set 2 and N_{Set2} is the number of samples in Set 2. N_{ND} denotes the number of samples which are detected as healthy samples in Set 3 and N_{Set3} is the number of samples in Set 3. The FA rate represents the robustness of the monitoring model to random variations in the healthy data. Since the confidence level of control limits is set as 1%, the FA rate should be close to 1%. The MA rate represents the sensitivity of the monitoring model to anomalies. By inspecting Set 2 and Set 3 in Fig. 8, a good monitoring model should have no missed alarms, i.e. zero MA rate, because the two sets do not overlap.

TABLE 2. Quantitative performance.

	SPE					Linear[1]
σ^2	0.2	1.5	5	10	100	N/A
FA (%)	17.4	2.2	2	1.4	1	0.8
MA (%)	0	0	1.52	3.3	10.91	16.24

Table 2 compares the quantitative performance, i.e. the FA rate on Set 2 and the MA rate on Set 3, of the RBF-KPCA approach with various σ values and the linear PCA approach. It can be observed that, relative to other

[1] T^2 is used as the monitoring statistic in linear PCA because both two PCs are retained. A data sample can always be reconstructed by the PCA model with two PCs in this example.

combinations, the SPE with $\sigma^2_{\text{opt,SPE}} = 1.5$ estimated in the previous section (Fig. 7(b)) can achieve a MA rate equal to zero with a FA rate close to 1%. The σ value smaller than the $\sigma_{\text{opt,SPE}}$ results in an over-fitted model which also achieves a MA rate of zero, but with a high FA rate. This indicates that the model is not robust to the randomness in the healthy data. Moreover, larger σ values (e.g. $\sigma^2 = 100$ and 20000) may also achieve low FA rates as the detection contours become relaxed. However, since the contours achieved by these σ values do not match the profile of the healthy data well, the monitoring model cannot differentiate properly between the healthy data and the anomalous data. Thus the MA rate increases as σ increases. An extreme case occurs when linear PCA is applied. The FA rate is low while the MA rate is high since the contour in Fig. 8 achieved by linear PCA is different from the profile of the healthy data. This further indicates that, although various σ values may achieve similar FA rates in cross-validation, their performance in anomaly detection can be different. A cross-validation strategy that purely minimizes false alarms is insufficient for tuning the σ.

B. EXPERIMENTAL BENCHMARK DATASET

1) DATASET DESCRIPTION

This example uses a real-life, experimental benchmark dataset obtained from a pilot-scale multiphase flow facility. It provides a demonstration of the influence of σ and compares the performance of SPE and T^2 for KPCA on higher dimensional real-life data with varying magnitudes of faults. Eight process variables including flow rate, pressure, and water level, are used for model training and validation. For a full description of these variables and the experimental facility, one may refer to [23].

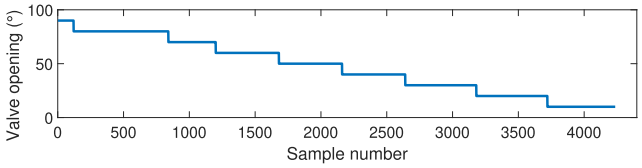


FIGURE 9. Operating sequence for the air blockage fault.

This example uses the data from an operating mode with 120 $m^3\,h^{-1}$ inlet air and 0.1 $kg\,s^{-1}$ inlet water. The data used for training were recorded when the facility was operating in healthy conditions. A fault was manually induced in the inlet air line by reducing the valve opening in a sequence of stepwise increments to simulate a developing blockage fault in the pipeline. This blockage fault influences the flow regime in the facility, therefore changes the relationship of flow rates and pressure measurements. The valve opening sequence for introducing this fault is shown in Fig. 9. The real-life process data collected when the fault is induced are plotted in Fig. 10 and RBF-KPCA is applied to the faulty data in Fig. 10. It can be observed that the deviation of process measurements becomes visible as the fault magnitude increases.

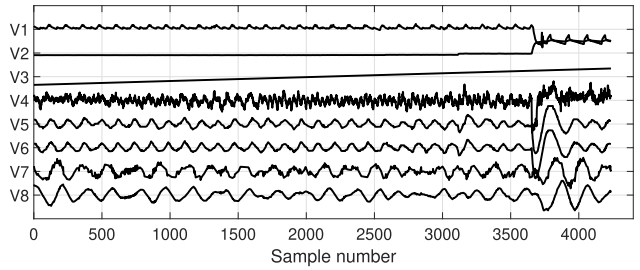


FIGURE 10. High density plot of the measurements from the air blockage fault.

Therefore, this faulty dataset includes both an anomaly with small and large magnitudes, making it suitable for demonstrating the performance of anomaly detection.

2) RESULTS AND DISCUSSIONS

When applying the RBF-KPCA approach to this dataset, the σ is tuned based on the strategy proposed in Section VI-B. The results for the optimum value of σ^2 are shown in Fig. 11. A fault is detected when the monitoring statistic exceeds its control limit for a continuous sequence of 50 samples.

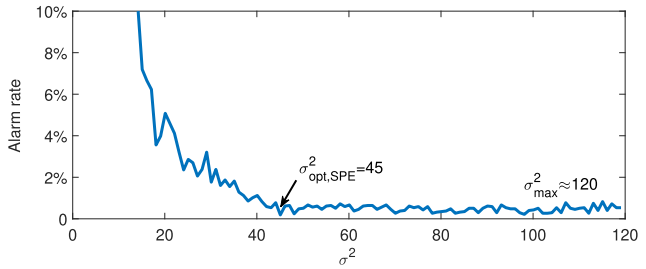


FIGURE 11. Alarm rates and $\sigma^2_{\text{opt,SPE}}$ obtained by SPE for KPCA.

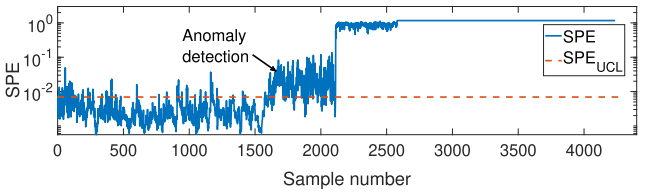


FIGURE 12. Trend plot of SPE for RBF-KPCA with $\sigma^2_{\text{opt}} = 45$.

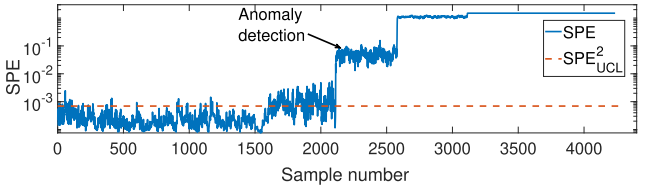


FIGURE 13. Trend plot of SPE for RBF-KPCA with $\sigma^2 = 500$.

Figs 12 to 14 show the performance of SPE. The SPE obtained by RBF-KPCA with $\sigma^2_{\text{opt,SPE}} = 45$ (Fig. 12) can

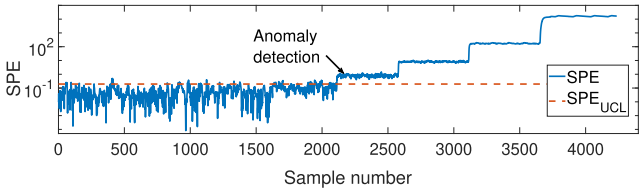


FIGURE 14. Trend plot of SPE for PCA.

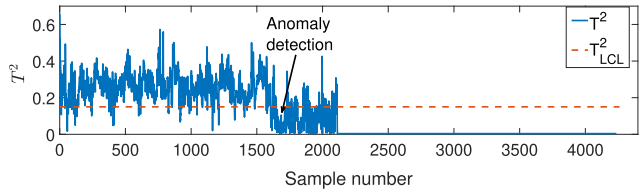


FIGURE 15. Trend plot of T^2 for RBF-KPCA with $\sigma^2 = 2$.

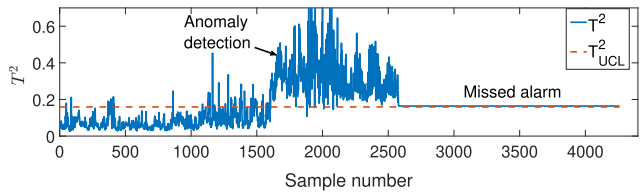


FIGURE 16. Trend plot of T^2 for RBF-KPCA with $\sigma^2 = 45$.

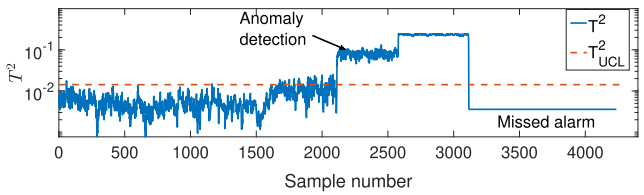


FIGURE 17. Trend plot of T^2 for RBF-KPCA with $\sigma^2 = 1500$.

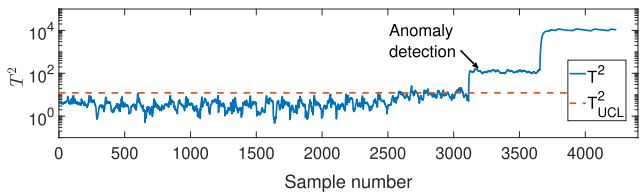


FIGURE 18. Trend plot of T^2 for PCA.

detect the blockage fault earlier than the cases when σ is exceedingly large (Fig. 13) and when the SPE is calculated from linear PCA (Fig. 14).

Figs 15 to 18 compare the performance of T^2 obtained by RBF-KPCA with $\sigma^2 = 2$, 45 and 1500 and by PCA, respectively. When σ is small, Fig. 15 shows that T^2 decreases when a fault occurs. This observation explains the decision of some authors [9] to use a lower control limit for T^2. In Fig. 16, where $\sigma^2 = 45$, T^2 first increases then reduces with respect to the fault development. When σ is inappropriately tuned (Fig. 17) and the upper control limit is used, T^2 can detect the fault when its magnitude is small while faults with larger magnitudes will be missed due to the non-monotonicity of T^2. The non-monotonicity issue of T^2 does not exist when PCA is applied (Fig. 18). However, linear PCA with T^2 has a later detection when compared with the result in Fig. 12 because it cannot capture data nonlinearity.

This example shows that the behaviour of T^2 in RBF-KPCA when σ is exceedingly large can be misleading. In such a situation T^2 may be increasing when the fault magnitude is small and the upper control limit can detect the fault. However, since T^2 is non-monotonic, it drops back below the upper control limit when the fault magnitude increases, leading to the RBF-KPCA approach failing to detect the fault with large magnitudes. On the other hand, the performance of SPE for RBF-KPCA is not influenced by the non-monotonicity issue. RBF-KPCA with properly tuned σ values and SPE as the monitoring statistic can detect faults with both large and small magnitudes. Compared to the numerical example, this real-life example with eight variables also shows that the findings on the SPE and T^2 for RBF-KPCA can be generalized to a higher dimensional, real-life dataset with a real fault.

VIII. COMMENT ON OUTLIERS AND ANOMALIES

The assumption of the paper is that the training data are healthy with no outliers or other issues of data quality. Hawkins defined an outlier as an "Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism" [24]. Detection and removal of outliers from the training data set falls under the task of data cleaning and there are many method for data cleaning, including various tests for outlier detection outlined in the monograph by Hawkins.

The analysis in this article is formulated based on the assumption that the training set has been cleaned of outliers and contains only samples that arise during healthy operation. Any data sample that is far away from the training data will therefore be classified as anomalous. The findings of the paper including the influence of the kernel width, the behaviour of the monitoring statistics, and the tuning strategy, are valid under this assumption.

A future research direction is to consider the behaviour of RBF-KPCA and the monitoring statistics when outliers exist in the training data. The formulation of KPCA needs to be adapted in order to account for outliers in training data. The authors of [25]–[27] reported several ways to make KPCA robust to outliers. The authors of [26] gave an example of analysing the influence of outliers on the outcome of KPCA. Also, [28], [29] presented alternative solutions to the problem of anomaly detection with outliers. It will be interesting to investigate the influence of the kernel width and the behaviour of monitoring statistics in these robust methods.

IX. CONCLUSION

The selection of monitoring statistics and the tuning of the RBF-KPCA for anomaly detection was investigated in this article. The asymptotic analysis with respect to the kernel width highlighted that inappropriate tuning of the kernel width in RBF-KPCA may impact the performance. When the kernel width is too small, the anomaly detection model will be over-fitted and the false alarm rate will be high. When the kernel width is too large, the model cannot capture nonlinearity.

The behaviour of SPE and T^2 as monitoring statistics in RBF-KPCA is proven to be different from the behaviour of SPE and T^2 in linear methods. It is shown that T^2 is non-monotonic with respect to the magnitude of the anomaly, making it not optimal as a monitoring statistic. Under the RBF-KPCA framework, the SPE is a better monitoring statistic because it can detect both anomalies that exceed the healthy range of variables and anomalies which do not follow the model for the healthy data. The SPE as formulated for RBF-KPCA can detect anomalies that would require both T^2 and SPE in a linear method. Moreover, the SPE for RBF-KPCA increases monotonically as the magnitude of the anomalies gets larger, making it possible to set an upper control limit for anomaly detection, which cannot be adopted for T^2 due to its non-monotonicity. A tuning strategy for the kernel width was proposed. Numerical and real-life case studies showed the effectiveness of the tuning strategy for the kernel width and for the SPE as a single monitoring statistic.

APPENDIX

Fig. 4(b) has shown that a common turning point at a specific value of d^2/σ^2 exists for T^2 when using large σ values, resulting in a non-monotonic T^2. We explore the non-monotonicity behaviour of T^2 and the common turning point in this Appendix. To find the turning point, the derivative of T^2_{test} with respect to x_{test} is investigated. Since T^2_{test} is a function of $\tilde{K}_{i,\text{test}}$, we try to find the local optimum for T^2_{test} by finding a local optimum that applies to all $\tilde{K}_{i,\text{test}}$.

It is clear that $K_{i,\text{test}}$ is monotonically decreasing with respect to the Euclidean distance between $\boldsymbol{x}_{\text{test}}$ and $\boldsymbol{x}_i$, i.e. $d_{i,\text{test}} = \sqrt{\|\boldsymbol{x}_{\text{test}} - \boldsymbol{x}_i\|_2}$, because of the RBF function:

$$K_{i,\text{test}} = \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_{\text{test}})^\top(\boldsymbol{x}_i - \boldsymbol{x}_{\text{test}})}{\sigma^2}\right). \tag{38}$$

However, $\tilde{K}_{i,\text{test}}$ may not be monotonic after centering (Eqn (2)). The derivative of $\tilde{K}_{i,\text{test}}$ with respect to $\boldsymbol{x}_{\text{test}}$ is investigated to check its monotonicity:

$$\frac{\partial \tilde{K}_{i,\text{test}}}{\partial \boldsymbol{x}_{\text{test}}} = \frac{\partial\left[K_{i,\text{test}} - \bar{K}_{i,\text{test}} + C_i\right]}{\partial \boldsymbol{x}_{\text{test}}} \tag{39}$$

where $C_i = 1/n^2 \sum\sum(K_{.,.}) - 1/n \sum_i (K_{i,.})$ is constant with respect to $\boldsymbol{x}_{\text{test}}$. For simplicity, we consider the case where x_{test} is a scalar.

By assuming $x_{\text{test}} > x_i$, Eqn (39) can be simplified as:

$$\begin{aligned}\frac{\partial \tilde{K}_{i,\text{test}}}{\partial x_{\text{test}}} &= \frac{\partial \exp\left(-\frac{(x_{\text{test}}-x_i)^2}{\sigma^2}\right)}{\partial x_{\text{test}}} - \frac{\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{(x_{\text{test}}-x_j)^2}{\sigma^2}\right)}{\partial x_{\text{test}}} \\ &= -\frac{2}{\sigma^2}(x_{\text{test}}-x_i)\exp\left(-\frac{(x_{\text{test}}-x_i)^2}{\sigma^2}\right) \\ &\quad + \frac{2}{n\sigma^2}\sum_{j=1}^{n}(x_{\text{test}}-x_j)\exp\left(-\frac{(x_{\text{test}}-x_j)^2}{\sigma^2}\right). \end{aligned} \tag{40}$$

When the stationary point occurs for $\tilde{K}_{i,\text{test}}$, Eqn (40) will be equal to zero. Denoting $a_i = 2/\sigma^2(x_{\text{test}} - x_i)\exp\left(-(x_{\text{test}} - x_i)^2/\sigma^2\right)$, Eqn (41) holds for the stationary point:

$$a_i - \frac{1}{n}\sum_{j=1}^{n} a_j = 0. \tag{41}$$

Since

$$T^2_{\text{test}} = \sum_{l=1}^{L}\lambda_l^{-1}\left(\sum_{i=1}^{n}\tilde{\alpha}_i^{(l)}\tilde{K}_{i,\text{test}}\right)^2 \tag{42}$$

and

$$\frac{\partial T^2_{\text{test}}}{\partial x_{\text{test}}} = 2\sum_{l=1}^{L}\lambda_l^{-1}\left(\sum_{i=1}^{n}\tilde{\alpha}_i^{(l)}\tilde{K}_{i,\text{test}}\right)\left(\sum_{i=1}^{n}\tilde{\alpha}_i^{(l)}\frac{\partial \tilde{K}_{i,\text{test}}}{\partial x_{\text{test}}}\right), \tag{43}$$

a sufficient but not necessary condition for T^2 having a maximum is that all $\tilde{K}_{i,\text{test}}$ have the same local maxima or minima. If there exists a common stationary point for all $\tilde{K}_{i,\text{test}}$, where $i = \{1, \ldots, n\}$, at x_{test}, the following matrix equation

$$\begin{bmatrix} 1/n-1 & 1/n & \ldots & 1/n \\ 1/n & 1/n-1 & \ldots & 1/n \\ \ldots & & & \\ 1/n & 1/n & \ldots & 1/n-1 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_n \end{bmatrix} = \mathbf{0} \tag{44}$$

is valid. The non-zero solution to Eqn (44) is any non-zero vector $\{a_1, a_2, \ldots, a_n\}$ that satisfies $a_1 = a_2 = \cdots = a_n$. However, since $x_{\text{test}} - x_i$ are different for different i values, the solution to Eqn (44) is infeasible. Instead, we consider the following minimization problem where all $\tilde{K}_{i,\text{test}}$ have their stationary points in a very small neighbourhood:

$$\begin{aligned}&\min_{x_{\text{test}}} \sum_{i=1}^{n}\left(a_i - \frac{1}{n}\sum_{j=1}^{n}a_j\right)^2 \\ &\text{s.t. } a_i = \frac{2}{\sigma^2}(x_{\text{test}} - x_i)\exp\left(-\frac{(x_{\text{test}} - x_i)^2}{\sigma^2}\right). \end{aligned} \tag{45}$$

To simplify this problem:

$$\begin{aligned}&\arg\min_{x_{\text{test}}} \sum_{i=1}^{n}\left(a_i - \frac{1}{n}\sum_{j=1}^{n}a_j\right)^2 \\ &= \arg\min_{x_{\text{test}}}\left[\frac{n-1}{n}\sum_{i=1}^{n}a_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}a_i a_j\right] \\ &= \arg\min_{x_{\text{test}}}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(a_i - a_j)^2 \end{aligned} \tag{46}$$

where

$$a_i - a_j = \frac{2}{\sigma^2}(x_{\text{test}} - x_i)\exp\left(-\frac{(x_{\text{test}} - x_i)^2}{\sigma^2}\right) - \frac{2}{\sigma^2}(x_{\text{test}} - x_j)\exp\left(-\frac{(x_{\text{test}} - x_j)^2}{\sigma^2}\right). \quad (47)$$

Now considering a function of x:

$$f(x) = \frac{2}{\sigma^2}x\exp\left(-\frac{x^2}{\sigma^2}\right), \quad (48)$$

Eqn (47) can be written as:

$$a_i - a_j = f(x_{\text{test}} - x_i) - f(x_{\text{test}} - x_j). \quad (49)$$

It is reasonable to assume that $|x_i|, |x_j| \ll x_{\text{test}}$. Eqn (49) becomes:

$$a_i - a_j = \left[f(x - x_i) - f(x - x_j)\right]\Big|_{x=x_{\text{test}}} \approx f'(x)\Big|_{x=x_{\text{test}}}(x_j - x_i). \quad (50)$$

Then the optimization problem in (46) becomes:

$$\arg\min_{x_{\text{test}}} \sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(f'(x)\Big|_{x=x_{\text{test}}}(x_j - x_i)\right)^2 = \arg\min_{x_{\text{test}}} f'(x)\Big|_{x=x_{\text{test}}} \sum_{i=1}^{n}\sum_{j\neq i}^{n}(x_i - x_j)^2. \quad (51)$$

Given that $\sum_{i=1}^{n}\sum_{j\neq i}^{n}(x_i - x_j)^2$ is constant when the training set is fixed, the solution to this optimization problem will be $\hat{x}_{\text{test}}$ such that:

$$f'(x)\Big|_{x=\hat{x}_{\text{test}}} = \frac{2}{\sigma^2}\left[(1 - \frac{2\hat{x}_{\text{test}}^2}{\sigma^2})\exp\left(-\frac{\hat{x}_{\text{test}}^2}{\sigma^2}\right)\right] = 0. \quad (52)$$

The solutions to this condition are:

$$\hat{x}_{\text{test},1} = \frac{\sigma}{\sqrt{2}} \quad \text{and} \quad \hat{x}_{\text{test},2} = \infty. \quad (53)$$

In this univariate case, the distance between x_{test} and the origin is $d_{\text{test}} = |x_{\text{test}}|$. Therefore, the tuning point of T^2_{test} exists such that $\hat{d}^2_{\text{test}} = \sigma^2/2$, which explains the behaviour in Fig. 4(b).

REFERENCES

[1] Z. Ge, Z. Song, and F. Gao, "Review of recent research on data-based process monitoring," *Ind. Eng. Chem. Res.*, vol. 52, no. 10, pp. 3543–3562, Mar. 2013.

[2] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognit.*, vol. 40, no. 3, pp. 863–874, Mar. 2007.

[3] J.-M. Lee, C. 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.

[4] Z. Ge, C. Yang, and Z. Song, "Improved kernel PCA-based monitoring approach for nonlinear processes," *Chem. Eng. Sci.*, vol. 64, no. 9, pp. 2245–2255, May 2009.

[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, Jan. 2015.

[6] Q. Jiang and X. Yan, "Parallel PCA–KPCA for nonlinear process monitoring," *Control Eng. Pract.*, vol. 80, pp. 17–25, Nov. 2018.

[7] 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.

[8] 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, May 2008.

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

[10] M. Jia, H. Xu, X. Liu, and N. Wang, "The optimization of the kind and parameters of kernel function in KPCA for process monitoring," *Comput. Chem. Eng.*, vol. 46, pp. 94–104, Nov. 2012.

[11] H. Lahdhiri, K. Ben Abdellafou, O. Taouali, M. Mansouri, and O. Korbaa, "New online kernel method with the tabu search algorithm for process monitoring," *Trans. Inst. Meas. Control*, vol. 41, no. 10, pp. 2687–2698, Jun. 2019.

[12] F. He, C. Wang, and S.-K.-S. Fan, "Nonlinear fault detection of batch processes based on functional kernel locality preserving projections," *Chemometric Intell. Lab. Syst.*, vol. 183, pp. 79–89, Dec. 2018.

[13] K. E. S. Pilario, Y. Cao, and M. Shafiee, "Mixed kernel canonical variate dissimilarity analysis for incipient fault monitoring in nonlinear dynamic processes," *Comput. Chem. Eng.*, vol. 123, pp. 143–154, Apr. 2019.

[14] C. J. Twining and C. J. Taylor, "The use of kernel principal component analysis to model data distributions," *Pattern Recognit.*, vol. 36, no. 1, pp. 217–227, Jan. 2003.

[15] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, Jul. 2003.

[16] C. F. Alcala and S. J. Qin, "Reconstruction-based contribution for process monitoring with kernel principal component analysis," *Ind. Eng. Chem. Res.*, vol. 49, no. 17, pp. 7849–7857, Sep. 2010.

[17] X. Deng, X. Tian, S. Chen, and C. J. Harris, "Nonlinear process fault diagnosis based on serial principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 560–572, Mar. 2018.

[18] C. Chakour, A. Benyounes, and M. Boudiaf, "Diagnosis of uncertain nonlinear systems using interval kernel principal components analysis: Application to a weather station," *ISA Trans.*, vol. 83, pp. 126–141, 2018.

[19] 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.

[20] C. Zhang, X. Gao, T. Xu, Y. Li, and Y. Pang, "Fault detection and diagnosis strategy based on a weighted and combined index in the residual subspace associated with PCA," *J. Chemometrics*, vol. 32, no. 11, Nov. 2018, Art. no. e2981.

[21] 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, Jul. 1998.

[22] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 1997, pp. 583–588.

[23] 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.

[24] D. M. Hawkins, *Identification of Outliers*, vol. 11. London, U.K.: Chapman & Hall, 1980.

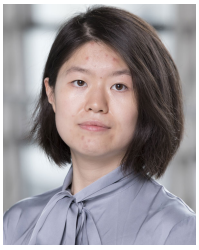
[25] N. Kwak, "Principal component analysis by L_p -Norm maximization," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 594–609, May 2014.

[26] S.-Y. Huang, Y.-R. Yeh, and S. Eguchi, "Robust kernel principal component analysis," *Neural Comput.*, vol. 21, no. 11, pp. 3179–3213, Nov. 2009.

[27] H.-H. Huang and Y.-R. Yeh, "An iterative algorithm for robust kernel principal component analysis," *Neurocomputing*, vol. 74, no. 18, pp. 3921–3930, Nov. 2011.

[28] N. Kumar, A. V. Rajwade, S. Chandran, and S. P. Awate, "Kernel generalized-gaussian mixture model for robust abnormality detection," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2017, pp. 21–29.

[29] N. Kumar, A. V. Rajwade, S. Chandran, and S. P. Awate, "Kernel generalized Gaussian and robust statistical learning for abnormality detection in medical images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4157–4161.



RUOMU TAN received the B.Eng. degree from Zhejiang University, China, in 2013, and the M.Sc. degree from the University of Alberta, Canada, in 2015. She is currently pursuing the Ph.D. degree with Imperial College London, London, U.K. She is currently a Research Scientist with the ABB Corporate Research Center, Ladenburg, Germany. Her research interests include data-driven nonlinear process monitoring, multivariate statistical analysis, and their application to process industries.

JAMES R. OTTEWILL (Member, IEEE) 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 Hitachi ABB Power Grids Research, Kraków, Poland, working in the field of applied analytics for condition monitoring applications. His main research interests include advanced physics-based and data-driven approaches for diagnostics and prognostics including dynamic testing, modeling and analysis of non-linear systems, signal processing, and information fusion.

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

NINA F. THORNHILL (Senior Member, IEEE) 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.

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