Adaptive Kernel Density-based Anomaly Detection for Nonlinear Systems

Liangwei Zhang*, Jing Lin, Ramin Karim
Division of Operation and Maintenance Engineering, Luleå University of Technology, SE-971 87, Luleå, Sweden
* (E-mail: liangwei.zhang@ltu.se)

ABSTRACT

This paper presents an unsupervised, density-based approach to anomaly detection. The purpose is to define a smooth yet effective measure of outlierness that can be used to detect anomalies in nonlinear systems. The approach assigns each sample a local outlier score indicating how much one sample deviates from others in its locality. Specifically, the local outlier score is defined as a relative measure of local density between a sample and a set of its neighboring samples. To achieve smoothness in the measure, we adopt the Gaussian kernel function. Further, to enhance its discriminating power, we use adaptive kernel width: in high-density regions, we apply wide kernel widths to smooth out the discrepancy between normal samples; in low-density regions, we use narrow kernel widths to intensify the abnormality of potentially anomalous samples. The approach is extended to an online mode with the purpose of detecting anomalies in stationary data streams. To validate the proposed approach, we compare it with several alternatives using synthetic datasets; the approach is found superior in terms of smoothness, effectiveness and robustness. A further experiment on a real-world dataset demonstrated the applicability of the proposed approach in fault detection tasks.

Keywords: maintenance modelling, fault detection, unsupervised learning, nonlinear data, kernel density

1. INTRODUCTION

Anomaly detection, also called outlier detection, intends to detect observations which deviate so much from others that they are suspected of being generated by nonconforming mechanisms [1]. In industry, the process is known as fault detection and aims to identify defective states of industrial systems, subsystems and components. Early detection of such states can help to rectify system behavior and, consequently, to prevent unplanned breakdowns and ensure system safety [2]. Fault detection constitutes a vital component of Condition-Based Maintenance (CBM) and Prognostics and Health Management (PHM). Modern industrial systems tend to be complex, so field reliability data (incl. System Operating/ Environmental data, or SOE data) are often highly nonlinear. This presents significant challenges to anomaly detection applications.

Nonlinear modelling has been considered as one of the main challenges wherein reliability meets Big Data [3]. Nonlinearity is an inherent phenomenon in nature. It is very often approximated by linear (or piecewise linear) relationships between features in practice; see [4] for an example. But for complex systems, linear approximation may easily underfit the problem. In light of this, many nonlinear models have been proposed to directly depict the interactions between system inputs, states and outputs for better anomaly detection. As a result, model-based approaches constitute a significant type of anomaly detection [5]. However, the first principle of the system must be known for these models to work well, and this is hard, especially in modern complex systems. Another type of anomaly detection attempts to acquire hidden knowledge from empirical data. This technique, the knowledge-based data-driven approach, is now receiving more attention [5]. Knowledge-based anomaly detection can be further divided into supervised and unsupervised approaches, depending on whether the raw data are labelled or not. The former method needs plentiful positive (anomalous) and negative (normal) data to learn the underlying generating mechanisms of different classes of data. Although anomalous data are easily obtained in laboratory experiments, they are generally insufficient in real-world applications [6]–[8]. Moreover, the generalization capability of supervised approaches to situations that have never occurred ("unhappened" anomalies) before is poor [9], [10]. In this paper, we only consider unsupervised, knowledge-based, data-driven anomaly detection techniques.

In the unsupervised regime, many existing anomaly detection techniques can deal with nonlinearity to a different extent. First, statistical methods detect anomalies based on the low probability of sample generation. Parametric ones typically require extensive a priori knowledge on the application to make strong assumptions on the data distribution; an example is the Gaussian Mixture Model (GMM) [11]. Non-parametric methods, such as the Parzen window estimator, estimate the probability density of data distribution using some smooth functions and then set a threshold to single out anomalies [12], [13]. Although they make no assumptions on the data distribution, they may perform badly when different density regions exist in the data. Second, density-based approaches (in a spatial sense) are another type of nonlinear technique in anomaly detection; of these, the Local Outlier Factor (LOF) approach is the best known. LOF is free of assumptions on the data distributions and has many desired properties, such as computational simplicity [14]. However, the metric local outlier factor is discontinuous and highly dependent on its input parameter. Third, an Artificial Neural Network (ANN) can handle nonlinearity because of its nonlinear activation function and multi-layer architecture. Self-Organizing Map (SOM) is a typical unsupervised ANN; it learns to cluster groups of similar input patterns onto low-dimensional output spaces (most commonly a two-dimensional discrete lattice). Even though SOM has been used in anomaly detection applications [15], its original purpose was dimensionality reduction or clustering, not anomaly detection. Last but not least, in the machine learning field, the kernel method is a common trick to deal with nonlinearity. In the kernel method, nonlinear transformations are conducted from the original input space to a high-dimensional (possibly infinite) feature space. Traditional linear approaches applied in the feature space can then tackle nonlinear problems in the original input space. Examples in the context of anomaly detection include Support Vector Data Description (SVDD) and Kernel Principal Component Analysis (KPCA), and so on [16], [17]. The main problem with this type of learning is the lack of interpretability and the difficulty of tuning input parameters in an unsupervised fashion. An inappropriate setting of input parameters may easily lead to underfitting or overfitting.

In this paper, we propose an adaptive kernel density-based anomaly detection (Adaptive-KD for simplicity) approach with the purpose of detecting anomalies in nonlinear systems. The approach is instance-based and assigns a degree of being an anomaly to each sample, i.e., a local outlier score. Specifically, the local outlier score is a relative measure of local density between a point and a set of its reference points. Here, the reference set is simply defined as geometrically neighboring points that are presumed to resemble similar data generating mechanisms. The measure local density is defined via a smooth kernel function. The main novelty is that when computing local density, the kernel width parameter is adaptively set depending on the average distance from one candidate to its neighboring points: the larger the distance, the narrower the width, and vice versa. The method allows the contrast between potentially anomalous and normal points to be highlighted and the discrepancy between normal points to be smoothed out, something desired in anomaly detection applications. We extend the approach to an online mode to conduct anomaly detection from stationary data streams. To evaluate the proposed approach, we compare it with several alternatives using both synthetic and real-world datasets. The results demonstrate the efficacy of our approach in terms of smoothness, effectiveness, and robustness.

The rest of the paper proceeds as follows. In Section 2, we introduce two density-based anomaly detection approaches that are closely related to this research. The discussion of the strengths and weaknesses of these two approaches leads to our explanation of the original motivation for this study. We present the Adaptive-KD approach in Section 3; in this section, we focus on the computation of local density using adaptive kernel width. The approach is then consolidated to an integrated algorithm and extended to an online mode to detect anomalies in a stationary data stream. In Section 4, we compare the smoothness, effectiveness, and robustness of our approach with several alternatives, including LOF, SVDD and KPCA, using

synthetic datasets. The verification of the approach using a real-world dataset is also presented. Finally, in Section 5, we offer a conclusion.

2. Density-based anomaly detection approaches

Density is often interpreted from the perspective of probability as the mass of likelihood a random variable can take on a given value or interval. It is naturally connected to the degree of belongingness of one sample to a certain class (e.g. normal or abnormal). Non-parametric density estimation can be achieved through either the kernel approach or the k nearest neighbor approach. The former uses information on the number of samples falling into a region of fixed size, while the latter considers the size of the region containing a fixed number of samples [13]. Corresponding to these two main types, in this section we briefly introduce two density-based anomaly detection approaches, the Parzen window estimate for anomaly detection and the local outlier factor. The discussion clarifies the motivation for this study.

2.1 Parzen window estimate for anomaly detection

The Parzen window estimate, also called the Kernel Density Estimate (KDE), is a non-parametric method to estimate the probability density function of random variables. Low probability density may imply that the occurrence of a sample does not conform to an underlying data generating mechanism, hence indicating a possible anomaly, and vice versa. Let $\boldsymbol{X}$ (a $m \times n$ matrix) denote m independently and identically distributed samples $\{x_1,\ x_2, \ldots x_m\}$ drawn from some unknown probability density $p(x)$ in a n-dimensional Euclidean space. The kernel density estimator at x is given by:

$$\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m} h^{-n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where $K(\cdot)$ represents a kernel function, and h is the width parameter for controlling the smoothness of the estimator. The coefficients $1/m$ and h^{-n} normalize the density estimate such that it integrates to one in the domain of x. Commonly used kernel functions include Gaussian, Laplace, Epanechnikov, Uniform, Tri-cube and many others. To achieve smoothness in the density estimation, a smoothing kernel is required. A smoothing kernel is a function of an argument which satisfies these properties: $\int K(x)dx = 1$, $\int xK(x)dx = 0$, and $\int x^2K(x)dx > 0$ [9].

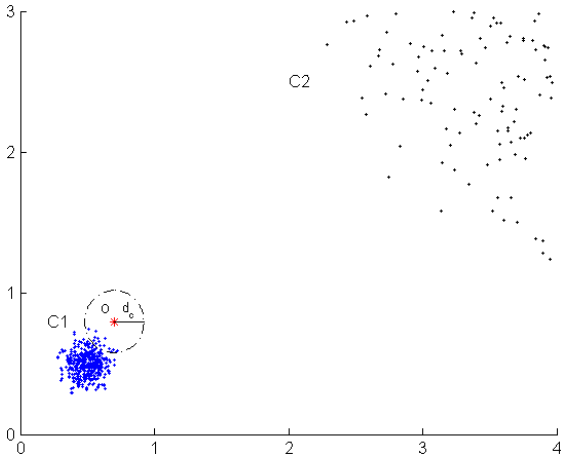

Figure 1: Parzen window estimator for anomaly detection; as a global measure of outlierness, it may fail to detect the outlying point *o* in the data

To detect anomalies from the given set $\boldsymbol{X}$, we can evaluate the density of all the samples using formula (1), and then set a threshold on this univariate density [18]. The samples with small density may be regarded as potential anomalies. In contrast to parametric approaches, the Parzen window estimate is free of assumptions on the data distribution and, hence, is of greater practical importance. That being said, however, it may perform badly in detecting anomalies in datasets containing several clusters with significant differences in their densities. This is shown by the example explained below and illustrated in Figure 1.

In Figure 1, point o (the red asterisk) is an anomaly adjacent to the dense cluster C1 (the blue points) and far away from the scattered cluster C2 (the black points). Suppose L_2 norm is chosen as the distance measure, and the uniform kernel with width d_c is adopted. If we ignore the normalization constant, $\hat{p}(o)$ (the density of point o) computed using formula (1) can be intuitively interpreted as the number of points falling in the d_c-ball (the dashed circle). Given the magnitude of d_c in Figure 1, $\hat{p}(o)$ may be higher than the density of many points in cluster C2. A threshold set for the density estimate that is large enough to capture point o may also lead to a high Type I error, i.e., false alarm rate. This is mainly because the density estimate here is a global measure of outlierness. It represents a lack of power in discriminating the outlying point o from those points in a less dense cluster, C2. Apart from this, a fixed kernel width in formula (1) is not advisable in segregating potential anomalies from normal samples, as will be discussed in Section 3.

2.2 Local outlier factor

Although the k nearest neighbor density estimator converges to the underlying probability density as the number of samples goes to infinity, the model produced by the k nearest neighbors approach is not a valid probability density model because its integral over all space diverges [13]. Thus, the nearest neighbor density estimator is rarely used in density estimation problems. However, the underlying idea remains instructive in many other problems. For example, the Local Outlier Factor (LOF) approach defines density based on the size of the region containing k nearest neighbors. In LOF, the so-called "local reachability density" of the ith point is defined as follows:

$$lrd(x_i) = 1/\left[\frac{\sum_{j\in kNN(x_i)} \text{reach-dist}_k(x_i, x_j)}{k}\right] \tag{2}$$

where $kNN(x_i)$ denotes the index set of the ith point's k nearest neighbors, and $\text{reach-dist}_k(x_i, x_j)$ is called the reachability distance of point x_i with respect to x_j in the set $kNN(x_i)$, as defined in the following:

$$\text{reach-dist}_k(x_i, x_j) = \max\left[d(x_i, x_j), k\text{-dist}(x_j)\right] \tag{3}$$

In formula (3), $d(x_i, x_j)$ is a measure of distance (e.g. L_2 norm.) from point x_i to x_j, and k-dist(x_j) is the distance from point x_j to its kth nearest neighbor (the kth element in $kNN(x_j)$ after sorting the distance in ascending order). The purpose of introducing reachability distance is to reduce statistical fluctuation in the distance measure.

Intuitively, local reachability density is a measure that can reflect the size of the region containing a point's k nearest neighbors. The smaller the local reachability density, the more confident we should be about the outlierness of a point, and vice versa. However, local reachability density is not necessarily a measure of local outlierness. It may suffer from the problem encountered in Figure 1 with the Parzen window estimator for anomaly detection. To resolve this problem, LOF defines a secondary metric, a local outlier factor, to measure local outlierness. The local outlier factor of the ith point is defined as follows:

$$LOF(x_i) = \frac{\frac{1}{k}\sum_{j \in kNN(x_i)} lrd(x_j)}{lrd(x_i)} \quad (4)$$

This is a relative measure computing the quotient between the average local reachability densities of a point's k nearest neighbors and the point's own local reachability density. Typically, points with a local outlier factor around (or less than) one should be considered normal, as their densities are roughly the same as (or larger than) the average density of their neighbouring points. A point with a local outlier factor remarkably larger than one is more likely to be an anomaly.

The key to defining a local outlierness measure (e.g., a local outlier score) is to compare the primary metric (e.g., local reachability density) of a point with those of its reference points (e.g., k nearest neighbors). Based on the LOF approach and many of its variants, a recent study has pointed out that the importance of defining the outlierness measure in a local sense is that a local outlierness measure is relatively more invariant to the fluctuations in the density estimate and, hence, is more comparable over a dataset with varying densities [19].

Despite its extensive applications in the real world, the LOF approach has two drawbacks: First, the primary metric (local reachability density) is not smooth, and this may cause discontinuities in the measure of the local outlier factor, as will be shown in Section 4. Second, its accuracy is very sensitive to the input parameter, namely, the number of nearest neighbors. A bad selection of this parameter can easily conceal the structure in the data and lead to a failure to detect potential anomalies; see Figure 6 (1.d) for an example.

With the aim of fostering the strengths of and circumventing the weaknesses in the above two approaches, this study combines them to get a smooth local outlierness measure that can detect anomalies from nonlinear data. The LOF approach provides a basic scheme for defining local outlierness, while the idea of using kernel functions in the Parzen window estimate approach is helpful in deriving a smooth density estimate. To enhance the discriminating power of the local outlierness measure, we explore the use of flexible kernel widths, as has been done in some "adaptive" kernel density estimation approaches.

3. Adaptive kernel density-based anomaly detection approach

Anomaly detection aims to identify observations which deviate so much from others that they are suspected of being generated by nonconforming mechanisms. A desirable anomaly detection approach should not only produce a binary output (abnormal or normal) but also assign a degree of being an anomaly to each observation. Based upon the two approaches introduced above, this section suggests using an adaptive kernel density-based approach to measure this degree of deviation. We start with the general idea of the approach; we then introduce the computation into local density and local outlier scores. We consolidate the parts in an integrated algorithm and extend it to an online mode. Finally, we discuss the time complexity.

3.1 General idea of the approach

The main purpose of the Adaptive-KD approach is to compute the degree of deviation of data points in a local sense. The significance of measuring the outlierness of a point locally has been highlighted in Section 2. To maintain a uniform notation, we follow the definition in Section 2 and let $\boldsymbol{X}$ be a given dataset containing m data points in R^n. The Adaptive-KD approach attempts to define a function f mapping from $\boldsymbol{X}$ to a real valued vector LOS in R^m; i.e., $f: \boldsymbol{X} \rightarrow LOS$, where $LOS(x_i)$ represents the i-th point's local outlier score.

To obtain a local measure of outlierness, the Adaptive-KD approach follows the basic steps of the LOF approach: defining the reference set, deriving the primary metric (local density), and then computing the secondary metric (local outlierness) based on the primary metric and the reference set. The main difference lies in the second step – how to compute samples' local density. To achieve smoothness in the final local outlierness measure, we adopt the idea of the Parzen window estimate to define the primary metric using a smooth kernel function. To enhance the ability to discriminate anomalous samples from normal ones, we use adaptive kernel width. The general idea of using adaptive kernel width to define local density is elucidated below.

In a classical density estimation problem using the Parzen window estimate, the width parameter h is fixed for all points. However, in regions of high density, a large width may lead to over-smoothing and a washing out of structure that might otherwise be learned from the data, while a small width may result in noisy estimates in regions of low density. Thus, the optimal choice for the width may be dependent on concrete locations within the data space. A natural solution to this problem is to apply large h in high-density regions and small h in low-density regions. But acquiring the information about high-density and low-density regions requires knowing the density, which is precisely the purpose of density estimation. Earlier studies tackled this paradox by using adaptive kernel density estimation, an example of which is Silverman's rule [20]. This rule uses the information on the average distance from a point to its k nearest neighbors as a rough estimate to the density of the point and defines the kernel width h_i as follows:

$$h_i = \frac{c}{k} \sum_{j \in kNN(x_i)} d(x_i, x_j) \tag{5}$$

where c is a user-defined parameter controlling the overall smoothing effect. The density estimate is then given by:

$$\hat{p}(x) = \frac{1}{m} \sum_{i=1}^{m} {h_i}^{-n} K(\frac{x - x_i}{h_i}) \tag{6}$$

In the context of anomaly detection, the favored settings for the kernel width are exactly the opposite of those in density estimation problems. In other words, a large width is preferred in high-density regions, and a small width is preferred in low-density regions. First, in high-density regions, although there may be some interesting structures, they are typically not of interest to us because they are non-informative in attempts to distinguish anomalies from normal samples. Moreover, an over-smoothing density estimate in high-density regions may reduce the variance of the local outlierness measure of the normal samples, which is helpful to single out anomalies. Second, in low-density regions, a narrow width will lead to smaller density estimates because the contribution from the "long tail" of a kernel is likely to be greatly reduced. This can make anomalous points stand out and enhance the sensitivity of the approach to anomalies.

3.2 Computation of local density using adaptive kernel width

To distinguish our approach from the above-described adaptive kernel density estimation approach, we use r_i and $\rho(x_i)$ to denote the kernel width and the local density of the ith point respectively. It is worth noting that the metric local density in our approach does not need to be a probability density; hence, the normalization constant in formula (6) can be ignored. Nor do we need to define local density for the whole data space; a metric defined on each data point in a given set is sufficient. By applying the Gaussian kernel, also known as the Radial Basis Function (RBF), the ith point's local density is given as follows:

$$\rho(x_i) = \frac{1}{m-1} \sum_{j \in \{1,2,\ldots,m\} \setminus \{i\}} \exp\left\{ -\left(\frac{x_i - x_j}{r_i} \right)^2 \right\} \tag{7}$$

The right-hand side of formula (7) excludes the contribution from the ith point itself (i.e., $\exp\{-(x_i - x_i)^2/{r_i}^2\} = 1$) in the summation. The purpose is to highlight the relative difference in density between different points (e.g., the quantity $0.1/0.3$ is much less than the quantity $1.1/1.3$). In addition, the subscript of the kernel width in formula (7) is different from the one in formula (6). It is only associated with the point of our concern, leading to a simple explanation of one point's local density as the following: the average contribution from the remaining points in the Gaussian kernel with a locality dependent width. A more intuitive interpretation is illustrated by a one-dimensional example containing five points $\{x_1, x_2, x_3, x_4, x_5\}$ in Figure 2 (a). The point x_2's local density is the average height of the blue, solid, vertical lines underneath the Gaussian kernel evaluation. From Figure 2 (a), it is also evident that local density reflects the extent to which one point is supported by others. The more neighboring points close to the point of concern, the larger its local density, and vice versa.

As argued above, the width r_i should be locality dependent. A large r_i is preferred in high-density regions and a small one in low-density regions. This is intuitively demonstrated in Figure 2 (b) where the kernel evaluations of three different points are plotted. The leftmost bell curve (in red) corresponding to the outlying point x_1 has the narrowest shape. The middle bell curve (in black) associated with x_2 has the widest shape because the point is near the center. The rightmost bell curve (in green) is associated with x_5 and has an intermediate width. As expected, this locality dependent width will lead to two results: points that are far away from others will be more isolated, and the discrepancy between normal points will be blurred.

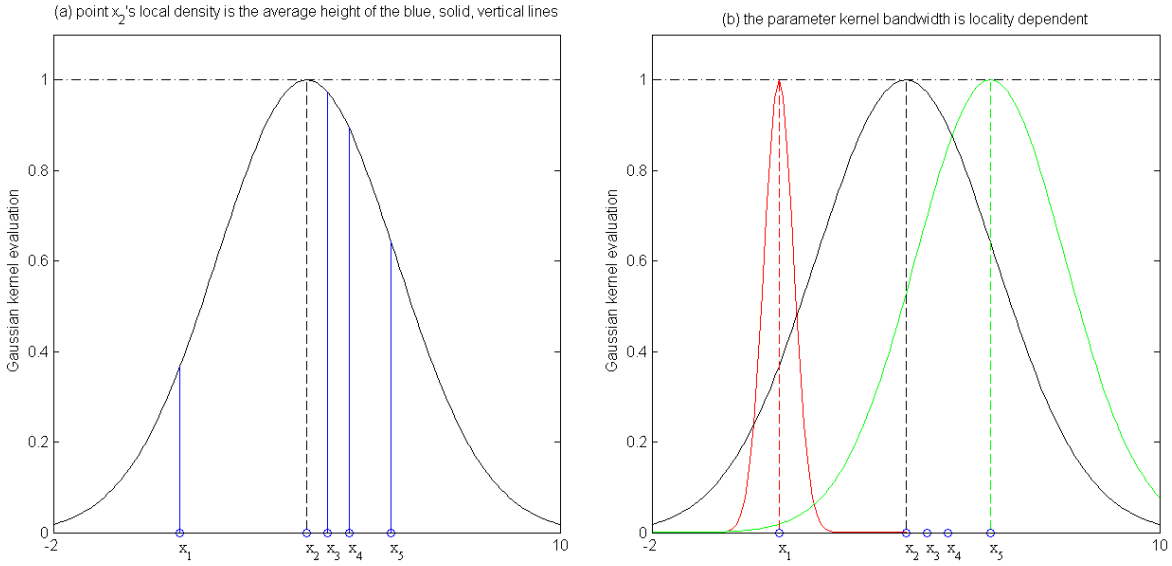

Figure 2: Illustration of (a) definition of local density, and (b) locality dependent width

Now, we discuss how to adaptively set the width parameter r_i in formula (7). Given the role of kernel width, we restrict it to be strictly positive. For the ith point, we let $d_k(x_i)$ denote the average distance to its k nearest neighbors; i.e., $d_k(x_i) = (1/k)\sum_{j \in kNN\{x_i\}} d(x_i, x_j)$. Further, we let $d_{k\text{-max}}$ and $d_{k\text{-min}}$, respectively, be the largest and the smallest quantity in the set $\{d_k(x_i) | \ i = 1,2, \dots, m\}$. Similar to Silverman's rule, we can first use $d_k(x)$ as a rough estimate of points' density and then construct a negative correlation between the width r and $d_k(x)$. Given these requirements, we define the ith point's width r_i as follows:

$$r_i = c[d_{k\text{-max}} + d_{k\text{-min}} + \varepsilon - d_k(x_i)] \tag{8}$$

where c ($c > 0$) is again the scaling factor controlling the overall smoothing effect, and ε is a significantly small positive quantity (e.g., 10^{-5}) ensuring that the width is non-zero ($d_{k\text{-min}}$ could be zero in some exceptional cases). We have two reasons for bringing in the term $d_{k\text{-max}} + d_{k\text{-min}}$. First, the width satisfies the requirement of being positive. Second, even without the scaling factor c, the width and the numerator in the exponent of formula (7) will be on the same scale. Some heuristic ways for selecting parameter c can now be applied. Silverman's rule of thumb suggests c should be from 0.5 to 1 in density estimation problems; this applies in our case.

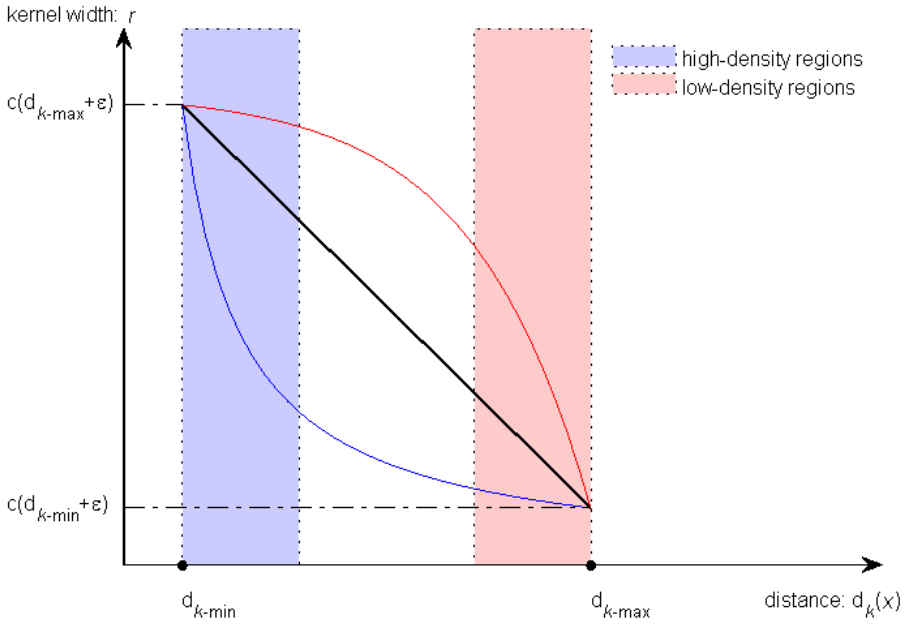

Figure 3: Illustration of the adaptive setting of kernel width

Note that the kernel width r in formula (8) has a linearly negative correlation with the quantity $d_k(x)$. This is shown by the black solid line in Figure 3 where kernel width r is plotted against the quantity $d_k(x)$. In general, as long as the above-described requirements are satisfied, the relationship between these two quantities can be of any form. Two other examples of these are given by the blue (with positive curvature) and the red (with negative curvature) solid curves in Figure 3. Informally, we assume points with small $d_k(x)$ are in high-density regions, and points with large $d_k(x)$ are in low-density regions. In the case of the blue curve, the kernel width of points in high-density regions drops rapidly as $d_k(x)$ increases but has a much slower decay rate in low-density regions. We can obtain the opposite results when formula (8) has a form resembling the red curve. Of course, piecewise functions can be applied here to establish the relationship between r and $d_k(x)$, but the form of the function should depend on the data structure of the problem and may be chosen differently depending on the application.

3.3 Computation of local outlier score

The name "local density" does not imply a local measure of outlierness. Rather, it serves as the primary metric in defining a relative measure of local outlierness, as in the LOF approach. The local outlier score for the ith point is defined as:

$$LOS(x_i) = \log\left[\frac{\frac{1}{k}\sum_{j\in kNN(x_i)}\rho(x_j)}{\rho(x_i)}\right] = \log\left[\sum_{j\in kNN(x_i)}\rho(x_j)\right] - \log[k\rho(x_i)] \quad (9)$$

An intuitive interpretation of the above quantity is that it is a relative comparison of the average local densities of one point's nearest neighbors and its own local density. The higher the local outlier score, the more we are confident in classifying the point as an anomaly, and vice versa. Here, the notion of locality is not only reflected by the selection of reference set (k nearest neighbors), but also by the definition of local density using adaptive kernel width. By introducing the monotonic logarithm function, we can use the "log-sum-exp" trick to prevent numerical underflow or overflow problems. Note that it requires some work to apply the trick to the first term of the second row in formula (9), a "log-sum-sum-exp" operation. For illustrative purposes, the definitions of local density and local outlier score are discussed separately; in practice, they should always be considered together to prevent numerical problems.

3.4 Model integration and its online extension

In the preceding sections, we have described the general idea and the main steps of the Adaptive-KD approach, notably the procedure for calculating local density using adaptive kernel width. Figure 4 streamlines the steps and consolidates them in an integrated algorithm. Most contents of the pseudo code in Figure 4 have already been covered, with the exception of feature normalization. Feature normalization is an important technique to standardize the numeric ranges of different features. It avoids having features in greater numeric ranges dominate those in smaller ranges in later calculations. In anomaly detection applications, we recommend the use of the Z-score normalization rather than the Min-Max scaling because the latter may suppress the effect of anomalies. The Z-score method normalizes the given matrix $\boldsymbol{X}$ to a dimensionless matrix $\boldsymbol{X}^*$. The i-th point x_i can be normalized as follows: $x_i^* = (x_i - \bar{x})/\sigma$, where $\bar{x}$ and σ are the column-wise mean vector and standard deviation vector of $\boldsymbol{X}$.

After obtaining the local outlier score of each point, we may want to classify which points are anomalous; we may even report alarms accordingly. Unfortunately, there is no deterministic way to map these continuous local outlier scores to binary labels, i.e., normal samples or anomalies. One simple way is to treat the top-most points with largest local outlier scores as anomalies, with the number of anomalies pre-determined by the user. Another way is to set a threshold and consider those objects with larger local outlier scores than the threshold as anomalies. We may also employ a reject option, refusing to classify some points to a given class because of lack of confidence. The objective of introducing yet another parameter (threshold) is to achieve higher precision and recall, in other words, to reduce the probability of committing both type I (false positive) and type II (false negative) error.

The Adaptive-KD approach introduced above is only able to detect anomalies from a given dataset. From a computational perspective, the algorithm needs to be executed in a batch-mode fashion. We can easily extend the approach to an online mode to detect anomalies from streaming data. This is of special interest in applications where real-time monitoring is of great importance. For example, timeliness is a significant factor in designing industrial fault detection applications. Typically, an online anomaly detection task has two phases: offline model training and online testing, as shown in Figure 5. In an unsupervised setting, the first phase tries to learn the normal behavior of the monitored system, and the second phase compares newly generated samples against the learned normal pattern upon their arrival. At testing time, the degree of

deviation of a sample from the normal pattern is used as evidence to discriminate anomalies from normal samples. This type of scheme is also known as one-class anomaly detection in machine learning, as it requires the training set to be restricted to negative samples (i.e., normal samples). Assuming the training set and testing set are already preprocessed, we explain the online extension of the Adaptive-KD approach in the following.

Algorithm 1 Adaptive-KD$(\boldsymbol{X}, k, c)$

BEGIN
Initialize LOS;
Conduct feature normalization on $\boldsymbol{X}$, and save it to matrix $\boldsymbol{X}^*$;
Compute pairwise distance $d(x_i^*, x_j^*)$ for all $i, j \in \{1,2,\dots,m\}, i \neq j$;
FOREACH $x_i^* \in \boldsymbol{X}^*$
Derive the reference set: k nearest neighbors $kNN(x_i^*)$ by sorting the above distances;
Calculate the average distance to its k nearest neighbors $d_k(x_i^*)$;
END
Obtain $d_{k\text{-min}}$ and $d_{k\text{-max}}$ from all the quantities $d_k(x_i^*)$ where $i \in \{1,2,\dots,m\}$;
FOREACH $x_i^* \in \boldsymbol{X}^*$
Compute the kernel width of the ith point r_i using formula (8);
Compute the local density of the ith point $\rho(x_i^*)$ using formula (7);
END
FOREACH $x_i^* \in \boldsymbol{X}^*$
Compute the local outlier score $LOS(x_i^*)$ using formula (9);
END
RETURN LOS;
END

Figure 4: Adaptive kernel density based anomaly detection (Adaptive-KD) algorithm

The offline model training phase, as shown in Figure 5, basically follows the procedure of the Adaptive-KD algorithm in Figure 4. Since this phase intends to learn the pattern of system normal behavior, it is worthwhile to meticulously select anomalous-free samples to construct the training set. The existence of anomalies may reduce the local outlier score of samples at testing time and could possibly lead to missed detections. To solve this, we add a data refinement procedure to exclude those samples with remarkably high local outlier scores from the training set and then retrain the model. Although the condition as to when data refinement is needed is somewhat subjective, it gives us a way to select representative training sets. This is often not possible in one-class anomaly detection approaches, such as the SVDD approach.

The normal pattern learned in the first phase is yielded as model parameters which are used in the second phase. Since the Adaptive-KD approach is an instance-based approach, the model parameters consist of all samples in the training set (possibly refined in the first phase) and their local densities. Other intermediate parameters that can be reused in the testing phase should also be included. For example, parameters $\bar{x}$ and σ are required to rescale online samples, and $d_{k\text{-min}}$ and $d_{k\text{-max}}$ are necessary for computing kernel width of samples at testing time. Notably, our model's parameters are fixed once trained. The fundamental assumption of this online extension is that the normal behavior of the system does not evolve as time goes on (no concept drift in the data stream). In other words, the monitored system is presumed to be stationary, or the change in the system normal behavior is negligible in the monitoring period. We can also retrain the model regularly to absorb normal changes in the system.

The online testing phase takes in real-time samples and computes their local outlier scores sequentially. A single testing sample goes through a routine similar to that of the first phase. Model parameters learned in the first phase provide necessary

information throughout the process, from feature normalization to the computation of local outlier score (dashed arrows). In the testing phase, the average distance of the previously unseen online samples to their k nearest neighbors could be extremely large. This may lead to a negative kernel width when applying formula (8), violating the positivity requirement. Thus, we redefine the kernel width using the following rectified linear function. Without incurring ambiguity, we still use x_i to denote the ith point irrespective of where it comes from (training set or testing set).

$$r_i = \begin{cases} c[d_{k\text{-min}} + \varepsilon], & d_k(x_i) > d_{k\text{-max}} \\ c[d_{k\text{-max}} + d_{k\text{-min}} + \varepsilon - d_k(x_i)], & \text{otherwise} \end{cases} \tag{10}$$

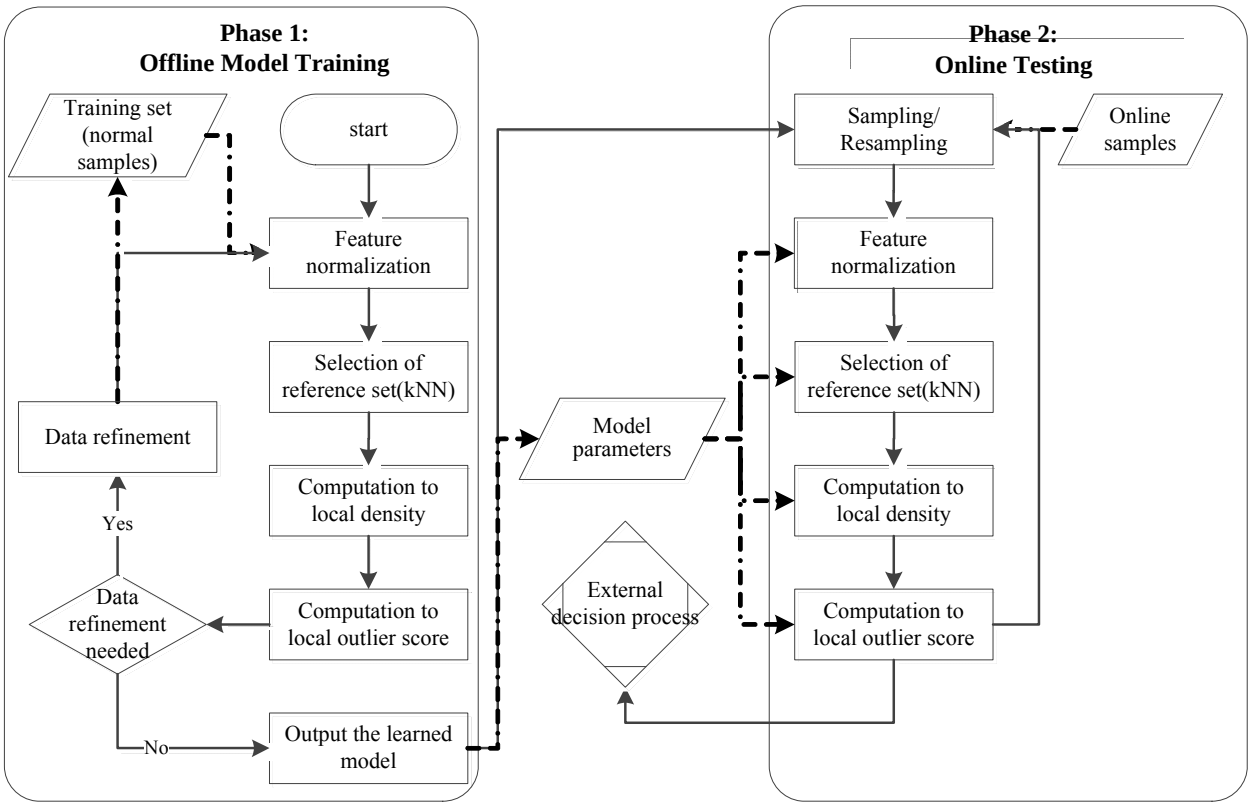

Figure 5: Online extension of Adaptive-KD algorithm for monitoring stationary systems

Apart from the above difference, the remaining computations in the testing phase follow exactly the same procedure as given in Figure 4. Notably, the reference set of any testing samples originates from the training set. After the local outlier score of an online sample computed, it is outputted to an external process to decide whether or not there is an anomaly occurs. As mentioned earlier, it is nontrivial to specify a threshold for singling out anomalous points with large local outlier scores. In cases where we have labeled data, especially anomalous samples, cross validation can be adopted to suggest the threshold, as is frequently done in supervised learning.

3.5 Time complexity analysis

Now we discuss the time complexity of the Adaptive-KD algorithm and its online extension. The most computationally intensive steps in the algorithm are the derivation of k nearest neighbors and the computation of local density, both of which

take the time complexity of $O(m^2 \cdot \max(n, k))$. Thus, the overall time complexity for the primitive Adaptive-KD algorithm and the offline model training phase (assuming there are m data points in the training set) of its extension are $O(m^2 \cdot \max(n, k))$. It is possible to reduce the computational cost by applying the following considerations to the above two steps.

Locality dependent kernel width is better than choosing a uniformly constant kernel width. However, this increases the computational complexity of performing local density evaluation, as it requires finding k nearest neighbors before figuring out the kernel width of points. A typical way to reduce the time complexity of finding k nearest neighbors is to employ an indexing structure, such as k-d tree or R* tree. The time complexity can be reduced to $O(m \cdot \log(m) \cdot \max(n, k))$ at the expense of additional memory space. Another improvement, random projection, can alleviate the high computational cost of finding k nearest neighbors when the dimensionality is high. This is supported by the Johnson-Lindenstrauss theorem claiming that a set of m points in a high-dimensional Euclidean space can be embedded into a $O(\log(m/\epsilon^2))$ dimensional Euclidean space such that any pairwise distance changes only by a factor of $(1 \pm \epsilon)$ [21].

The complication of local density computation lies in the Gaussian kernel evaluation, mainly because the Gaussian kernel has an unbounded support. In other words, the Gaussian kernel function needs to be evaluated for each point with respect to all remaining points. While the shape of the kernel function may important in theoretical research, from a practical perspective, it matters far less than the width parameter. Thus, other kernel functions with compact support, such as the Epanechnikov or the Tri-cube kernel, can be adopted. However, they require introducing additional parameters to determine the size of their support. Typically, only those points with a distance less than a given threshold to the point of interest will be evaluated using the chosen kernel function.

The online testing phase of the algorithm's extension continuously processes new samples upon their arrival. The time complexity of this phase is much more important in the sense that it decides whether the algorithm can give real-time or near real-time responses to a fast-flowing data stream. It is necessary to maintain those model parameters yielded from the training phase to avoid repetitive computations at testing time. This is where the concept of trading space for time applies. As in the offline model training phase, the most computationally demanding steps in the online testing phase are the derivation of k nearest neighbors and the computation of local density, both of which have a time complexity of $O(m \cdot \max(n, k))$. With the same considerations as discussed before, the computational cost can be vastly reduced.

4. Numerical illustration

This section evaluates the proposed approach using synthetic datasets and a real-world dataset. Concretely, we contrast the online extension of our approach with the LOF online extension, SVDD and KPCA using synthetic datasets. Then we compare the Adaptive-KD algorithm with LOF and Parzen window estimate approach using a dataset from the railway industry. Before diving into the numerical examples, we briefly introduce some uncovered approaches which are chosen here for comparison.

Building on the idea introduced in Subsection 3.4, the LOF approach can be extended to an online mode, the application of which is explained by [22]. The SVDD approach applies the "kernel trick" to implicitly conduct nonlinear mapping from the original input space to a high-dimensional feature space. It tries to find a minimum volume hyper-sphere that can enclose normal samples in the feature space [23]. For any testing sample, the outlierness measure is the difference between the distance from the testing sample to the hyper-sphere center and the radius of the hyper-sphere. The larger the measure, the more likely the sample is to be anomalous. The hyper-sphere can be obtained by minimizing an objective function containing

two terms: the first measures the volume of the hyper-sphere; the second penalizes larger distances from samples to the hyper-sphere center. An input parameter λ is needed to address the trade-off between the two. In the following experiments, we use the Gaussian kernel with an input parameter σ_{rbf} as the kernel width.

The KPCA approach is based on the spectral theory, which assumes normal samples and anomalies appear as significant discrepancies in a lower-dimensional subspace embedding. Similar to the SVDD approach, KPCA applies the "kernel trick" to extend Principle Component Analysis (PCA) to nonlinear cases. It learns the normal pattern from a training set by retaining most of the variance in the principal components. Then, the reconstruction error of the testing samples is used to depict their degree of outlierness [24]. The higher the reconstruction error, the more a testing sample disagrees with the learned pattern and the more likely it is to be an anomaly. In the following experiments, we use the Gaussian kernel with width parameter σ_{rbf}. Further, we let τ denote the proportion of variance retained in subspace.

4.1 Smoothness test on the "aggregation" dataset

In previous sections, we claimed our approach defines a smooth local outlierness measure. To justify this claim, we apply the online extension of the approach to the "aggregation" dataset and compare it with other alternatives. As shown in Figure 6 (1.a), the "aggregation" dataset contains 788 samples forming seven different clusters. The purpose is not to detect anomalies in this dataset. Instead, these samples constitute the training set, and they are considered normal. The testing set is obtained by discretizing the horizontal axis (from 0 to 40) and the vertical axis (from 0 to 30) using a step size 0.2. This leads to a two-dimensional grid with 30351 (151×201) intersecting points, i.e., the testing set. Training sets consisting of multiple clusters are common in reality. Each cluster represents a normal behavior of the monitored system running in a particular operational mode.

For all the anomaly detection approaches introduced so far, each testing sample can be assigned a degree of outlierness. For comparative purposes, all the outlierness measures are standardized to a range from 0 to 1. The larger the measure is, the more likely a testing sample is to be anomalous. In Figure 6, from subplot (1.b) to (1.h), each testing sample is marked using a colored point in the coordinate system. As indicated by the color bar, the degree of outlierness increases as the color evolves from dark blue to dark red. Each subplot from (1.b) to (1.h) corresponds to a particular approach under a specific parameter setting. The influence of parameters c and k to our approach will be explained later. Here, we simply present the result of our approach when $c = 1$ and $k = 40$. To illustrate how the LOF approach is affected by parameter k, we try two different settings: $k = 20$ and $k = 40$. As suggested in the original paper on the SVDD approach, the trade-off parameter λ should take value 1 when the training set is noiseless. Thus, we only vary the width parameter σ_{rbf} in the experiment. We fix parameter τ at 0.9 and vary the kernel width in the KPCA approach. The corresponding contour curves of the degree of outlierness are given in subplots (2.b) to (2.h).

An ideal approach should be able to detect the nonlinear shape of the clusters. Samples are also expected to have a low degree of outlierness when they fall inside the clusters, and a large degree when they are away from the clusters. Moreover, the transition in the outlierness measure from cluster cores to cluster halos should be smooth. As subplots (1.b) and (2.b) suggest, our approach can correctly detect the shape of the clusters and give a very smooth local outlierness measure. In addition, the results are fairly robust to the change of parameter k in this example. Another example of the contour plot when parameter $k = 20$ is presented in subplot (2.a). Notice that in the cluster cores, the local outlierness scores are almost identical. This is caused by the smoothing effect of large kernel width in high-density regions.

Although the LOF approach can detect the shape of the clusters when k is small, as shown in (1.c), it ruins the structure at the bottom-left two clusters when k takes a relatively large value, as shown in (1.d). Besides, as shown in subplots (2.c) and (2.d), the contour curve of the local outlier factor ripples in a wiggly line from cluster core to cluster halo because the local reachability density, from which the LOF measure is derived, is not a smooth metric. As shown in (1.e), the SVDD approach tends to underfit and fails to detect the shape of the clusters in the dataset when the kernel width is small. When σ_{rbf} is large, the approach can capture the overall shape of different clusters but, again, the measure of outlierness is not smooth, as indicated by the light blue hollows inside the clusters in (1.f). As opposed to the SVDD approach, the KPCA approach tends to underfit when σ_{rbf} is relatively large. Although the KPCA approach successfully identifies the shape of the clusters when σ_{rbf} is small, as shown in (1.g), its measure of outlierness is not as smooth as the local outlier scores produced using our approach.

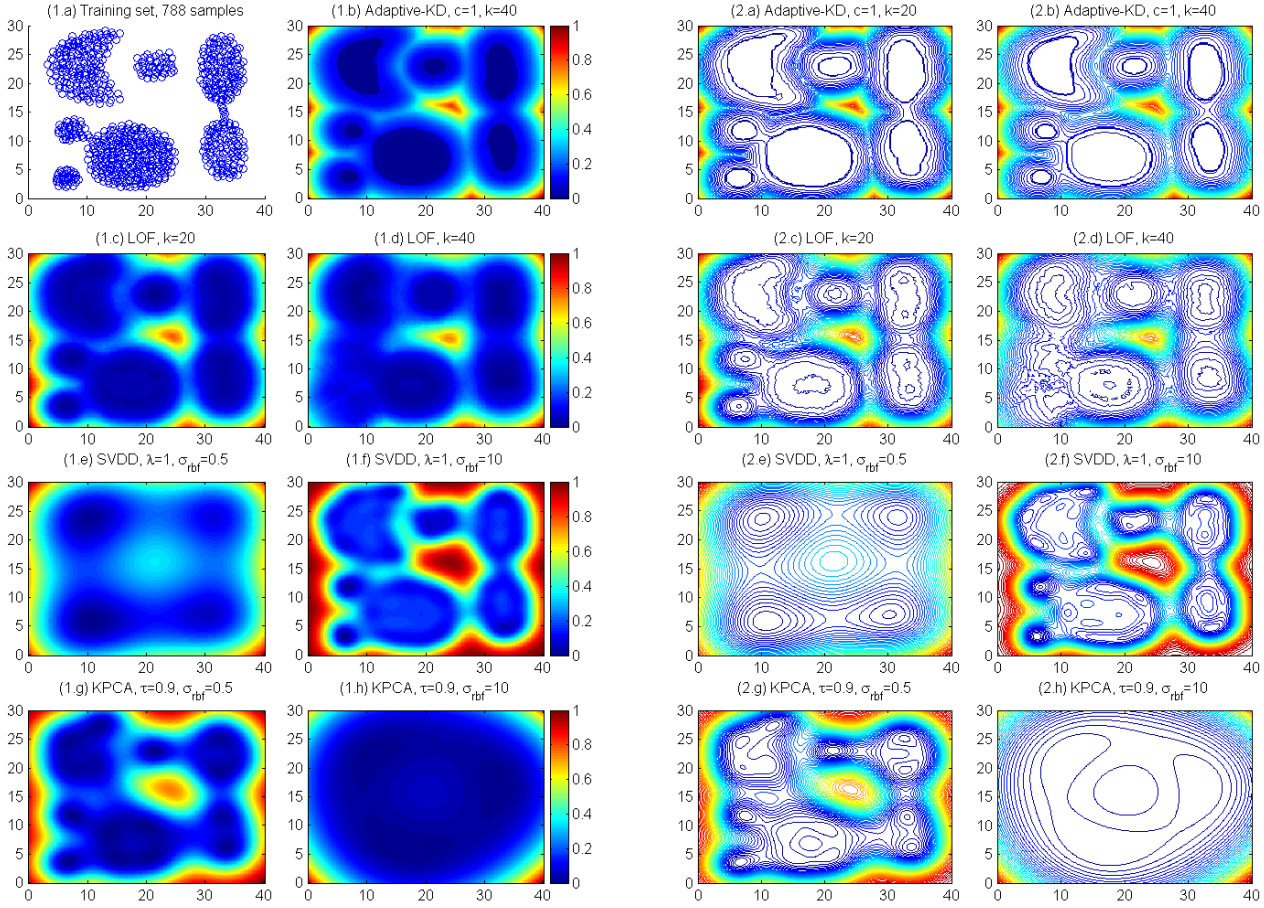

Figure 6: Smoothness test on the "aggregation" dataset

4.2 Effectiveness test on a highly nonlinear dataset: a two-dimensional toroidal helix

With a setup similar to the one used in the above example, we apply these approaches to a highly nonlinear dataset and compare the results in this section. The training set is a two-dimensional toroidal helix containing 1000 samples, as shown in Figure 7 (1.a). It is clear that our approach can effectively detect the shape of the data and the contour plot ripples smoothly

towards both outside and inside hollows, as shown in Figure 7 (1.b) and (2.b). Again, the LOF approach can somewhat recognize the shape of the data. But the contour plot is rather uneven, and the discontinuities in the measure of local outlierness is significant, especially when k takes a large value. The SVDD approach detects the shape when the kernel width is large, while the KPCA approach works when the width parameter is small. It seems SVDD performs better than KPCA in the interior of the toroidal helix. However, the outlierness measure of all three alternatives is not as smooth as we expected.

As we vary parameter k while fixing c in our approach, the results could appear to be over-smoothing or under-smoothing. This is mainly because the kernel width defined in formula (8) is also affected by parameter k. In general, a small k will lead to a small $d_k(x)$ and r, thereby decreasing the overall smoothing effect. The phenomenon can be compensated for by choosing a larger c. In Figure 7 (2.a), we present another comparable result; in this example, $c = 0.8$ and $k = 10$. The effect of over-smoothing and under-smoothing is elaborated in detail in the next subsection.

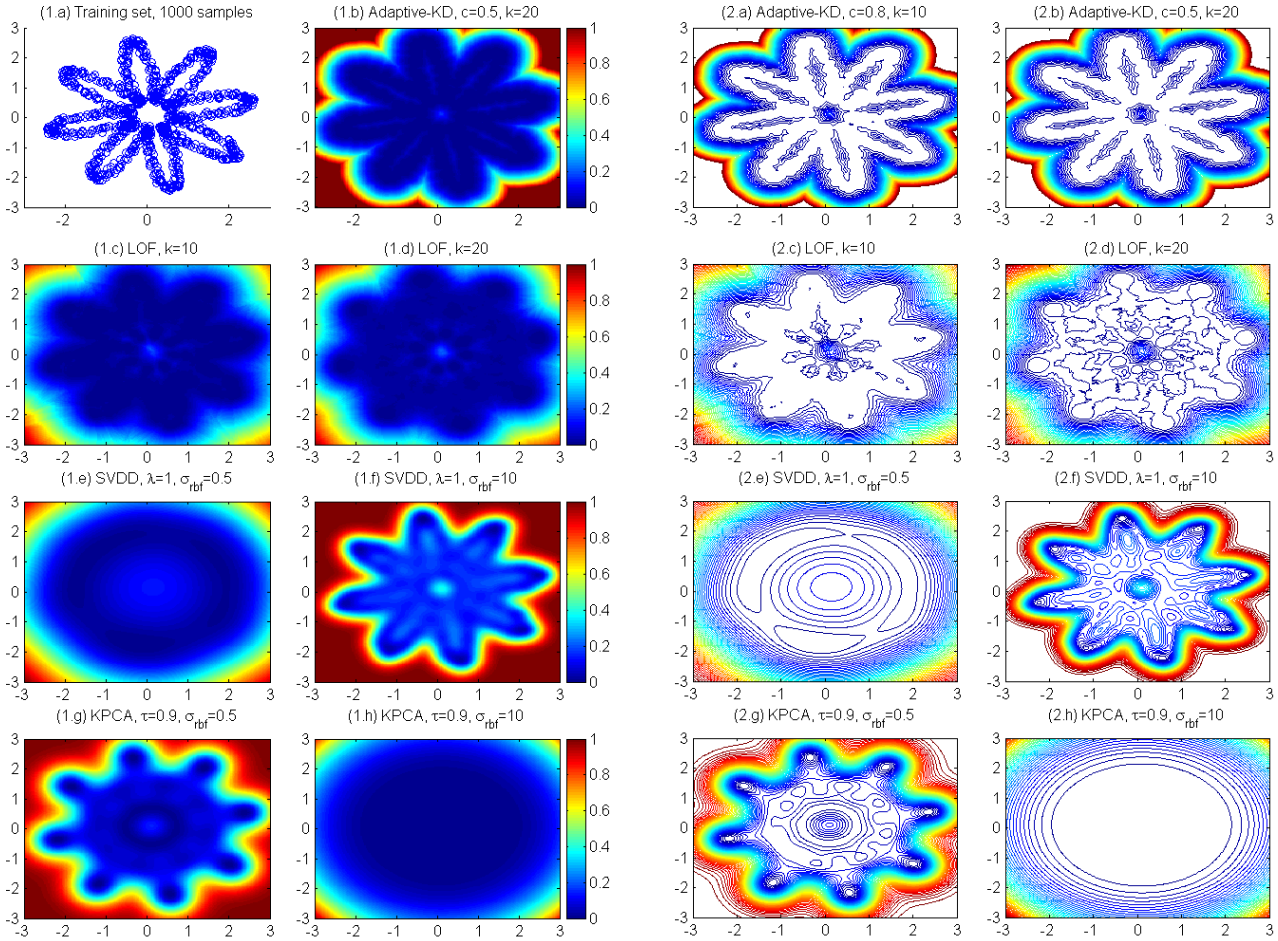

Figure 7: Effectiveness test on a two-dimensional toroidal helix dataset

In the above two examples, the purpose is to compare our approach with the selected alternatives. Even though a global measure of outlierness derived from a well-tuned kernel density estimator can achieve comparable smoothness in these examples, it may fail in a dataset where clusters have significant differences in their densities, as we argued in Subsection 2.1.

4.3 Robustness test on the "flame" dataset

In the following, we use the "flame" dataset to show how the existence of anomalies in the training set affects these approaches. We also discuss the robustness of our approach to the perturbation of input parameters. The "flame" dataset is shown in Figure 8 (1.a); the top-left-most two points are considered anomalies. The remaining sub-graphs in Figure 8 agree with our assessment of the smoothness and effectiveness of these approaches in the previous two examples. They also demonstrate that all approaches are affected by the two anomalies, albeit to a different extent. As described earlier, the Adaptive-KD approach naturally has the ability to assign a local outlier score to any sample in the training set. Thus, the data refinement step in the offline training stage should be able to capture and discard these two anomalies and then retrain a model on the refined set. The LOF approach can recognize the two anomalies with the same routine. However, it is non-trivial for the SVDD and KPCA approach to mitigate the effect exerted by anomalies in the training set.

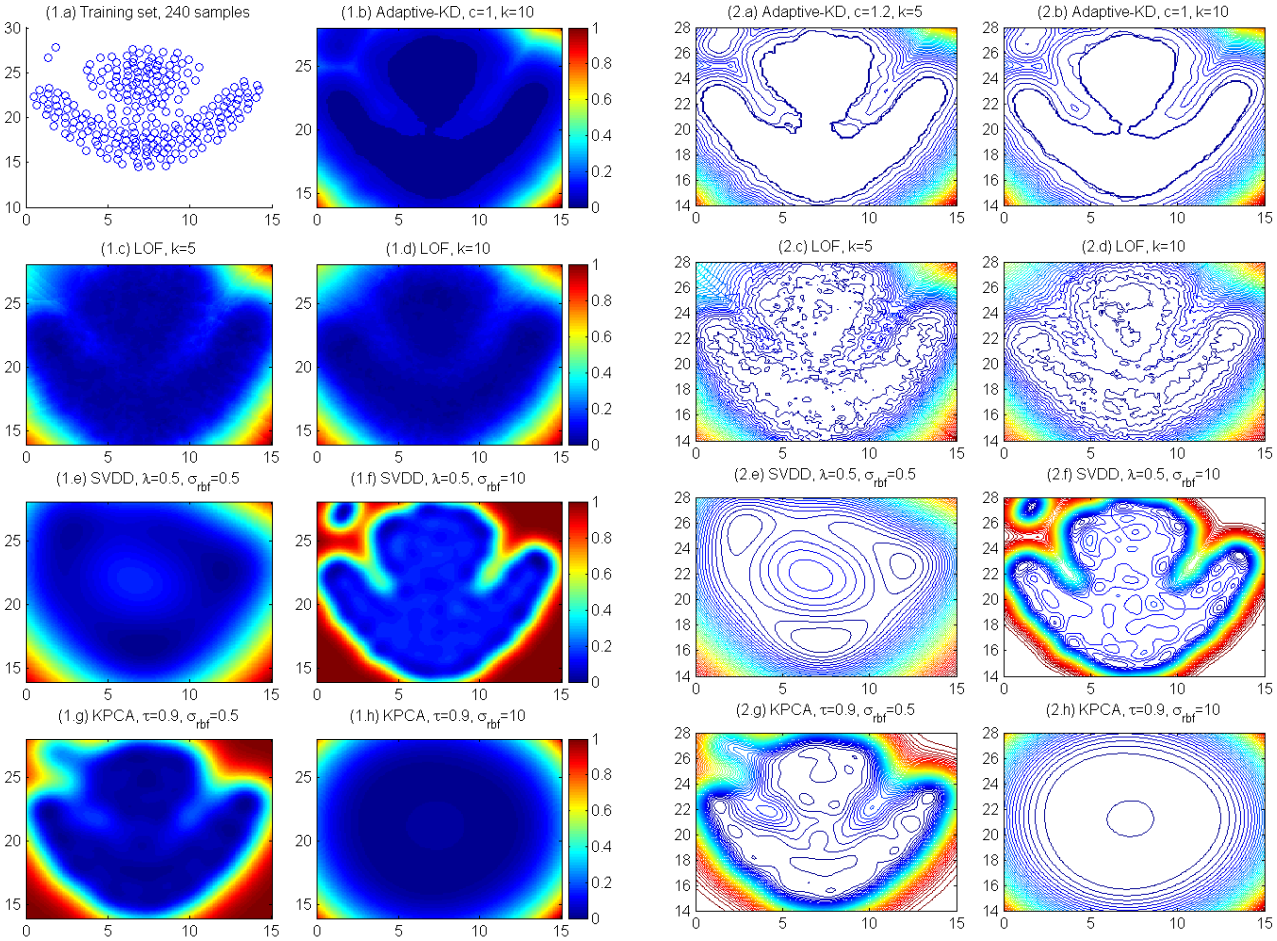

Figure 8: Robustness test on the existence of anomalies in the training set

The impacts of perturbing input parameters on our approach are presented in Figure 9. First, we vary parameter c while fixing k; the results are shown in (1.a) and (1.b), the corresponding contour plots of which are given in (2.a) and (2.b). As expected, parameter c directly controls the overall smoothing effect. A small c may cause the fine details in the data to be

enhanced, leading to overfitting, whereas a large one may lead to over-smoothing and underfitting. Note that when a large c is chosen, the influence of anomalies in the training set can be somewhat counteracted because the local information at the two anomalies is smoothed out. Second, we vary parameter k while fixing c; the results are shown in (1.c) and (1.d), the corresponding contour plots of which are given in (2.c) and (2.d). Unsurprisingly, since parameter k has an indirect influence on the scale of kernel width, it can affect the smoothing effect in a manner similar to c. The main difference is that k also decides the number of reference sets and consequently affects the local outlierness measure. This explains why the contour plot shown in (2.c) has a very wiggly interior when k takes a small value.

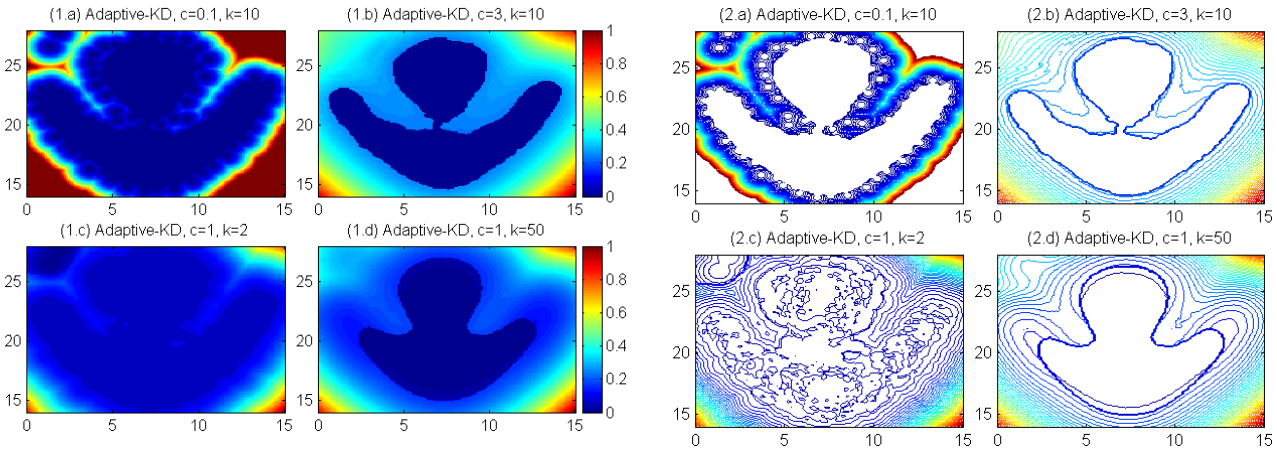

Figure 9: Robustness test on the perturbation of input parameters

As with other unsupervised learning approaches, the Adaptive-KD approach relies on the similarity (or dissimilarity) measure between points. Specifically, the measure LOS computes how similar one point's local density is to the densities of its k nearest neighbors. In an extreme case, when k takes the value of the size of the training set, the measure LOS recovers to a global measure of outlierness because the nominator in formula (9) is identical for every point, and the rank in the outlierness measure is simply the rank in the metric local density in reverse order. If k takes a very small value, however, the local densities of the very few reference points may dominate the calculation of the point's local outlier score, thereby leading to discontinuities in the outlierness measure, as shown in Figure 9 (2.c). According to our experiments in the above three examples, the results are fairly robust to changes in parameter k as long as it does not fall into a too large or too small range. Thus, we recommend setting k to a reasonably small value to capture the notion of locality and then adjusting c accordingly. Although the purpose of anomaly detection differs from that of density estimation, some heuristic methods (such as minimizing the frequentist risk) in density estimation applications can be employed to make a preliminary selection of parameter c.

4.4 Verification using a real-world dataset

In the railway industry, rolling stock wheel-set is one of the most important subsystems and is essential to service. Its service life can be significantly reduced by failure or damage, as both lead to accelerated deterioration and excessive costs [25], [26]. To monitor the health state of rolling stock wheel-sets and initiate maintenance actions accordingly, the Swedish railway industry continuously measures the dynamic forces of wheel-sets in their operation. These measurements may be indicative of the faults in the wheel-sets, such as surface defects (incl., cracks.), subsurface defects (incl., residual stress.),

polygonization (incl., discrete defects, roughness.), wheel profile defects (incl., wheel diameter irregularity), and so forth. How to effectively detect these faults from the measurements is crucial to the system reliability and safety.

High nonlinearity is observed in the sensor measurements, as can be seen in Figure 10, where vertical force on the right wheel of a wheel-set is plotted against its vertical transient force. In the graph, different clusters with various densities exist in the data, which may correspond to different loading weights, operational modes, etc. As we argued in Subsection 2.1, a global measure of outlierness (such as the Parzen window estimate approach) in this case may not easily detect faulty samples which are adjacent to some dense clusters. On the other hand, a too simple linear method might not be able to capture the nonlinear structure in the data. Notably, this high nonlinearity also appears in other features in the dataset, which further rationalizes the need of a model with sufficiently expressive power.

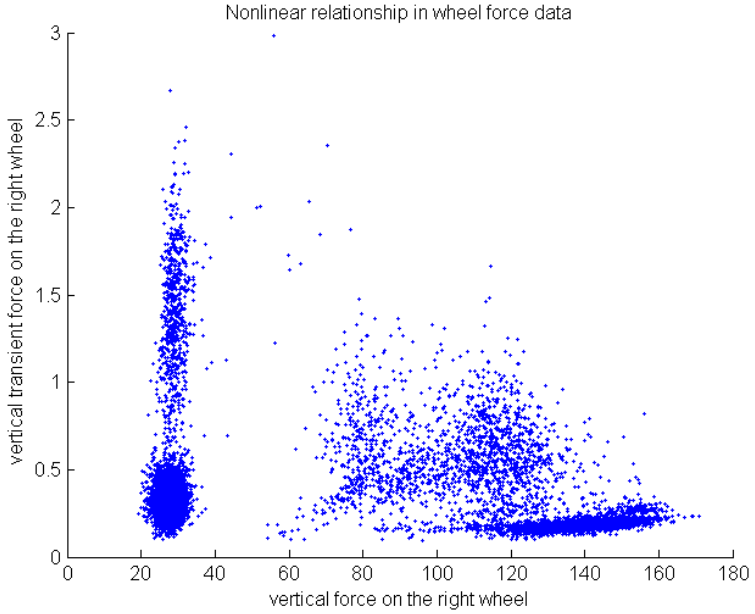

Figure 10: High nonlinearity exists in a real-world dataset

The dataset for verification is constructed via the following procedure: (i) We randomly select 10000 samples from the wheel-sets force data pertaining to normal operating conditions, and the time of measurement is in the range from September to December in 2015. (ii) We then apply the Adaptive-KD algorithm on the dataset and filter out those samples with significantly large local outlier scores. In this experiment, 9940 samples that are considered representative of the normal behavior of the wheel-sets are remained. (iii) We add another 30 samples that are considered abnormal to the dataset. These samples are obtained by tracing historical failure data, and re-profiling parameters that are regularly measured at wagon inspection workshop. Finally, a dataset with 9970 samples, of which 30 samples are anomalies, is constructed. The dimension of the dataset is 8, including, vertical forces on the wheel of both sides, lateral forces on the wheel of both sides, vertical forces on the axle, angle of attack, and vertical transient forces on the wheel of both sides.

To verify the proposed approach, we apply the Adaptive-KD algorithm on the wheel-set force dataset and compare it with the LOF and the Parzen window estimate (for anomaly detection) approach using the Receiver Operating Characteristic (ROC) curve. The ROC curve is a well-established graphical tool that can display the accuracy of a binary classifier. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, and hence it is threshold independent.

The larger the area under the curve (AUC), the better accuracy a classifier can achieve. In this experiment, the parameter k for both the LOF approach and our approach is set at 40, the parameter c in our approach is set at 0.5, and the kernel width (the Gaussian kernel is used) for the Parzen window estimate approach is set such that points' average number of neighbors is 2% of the sample size in the dataset. As shown in Figure 11, the Adaptive-KD approach outperforms the other two in terms of the accuracy. The AUC values of these approaches are 0.9974, 0.9828, and 0.9762, respectively. Though, seemingly, the three AUC values differ slightly, they can make a huge difference in reducing potential production losses and maintenance costs in practice.

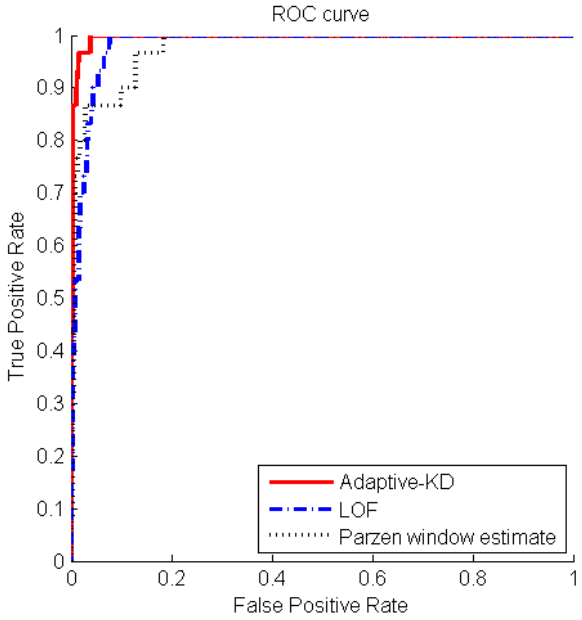

Figure 11: ROC curve comparison to different approaches on the wheel force data

After identifying a faulty sample using our approach, one may want to further investigate the reason of declaring the abnormality of the point. This can be informative to the ensuing procedure of fault diagnosis, which intends to probe into the type, source and severity of the underlying faults. In our approach, we can trace back to all the calculations to the point's k nearest neighbors, kernel width, local density, and its local outlier score. Then, a preliminary explanation for the abnormal behavior of the recognized anomalous sample may be given. Notably, it is nontrivial to analyze the results of approaches which implicitly conduct nonlinear transformations, such as the SVDD approach. This shows another merit of our approach – interpretability – over some of the kernel methods.

5. Conclusion

This paper presents an unsupervised, density-based approach to anomaly detection from nonlinear systems. Like many other unsupervised learning approaches, it uses the similarity measure between different points and assigns each point a degree of being an anomaly, namely, a local outlier score (LOS). LOS is defined here as a relative measure of local density between a point and a set of its neighboring points, and local density is the similarity measure evaluating how similar one point is to its neighboring points. To achieve smoothness in the measure, we adopt the Gaussian kernel function. To enhance the measure's discriminating power, we use locality dependent kernel width: wide kernel widths are applied in high-density regions, while

narrow ones are used in low-density regions. By doing so, we can blur the discrepancy between normal samples and intensify the abnormality of potentially anomalous samples. When Silverman's rule is adopted, the recognition of regions of different density simply becomes a rough estimate of density, i.e., the average distance from one point to its k nearest neighbors (in a negative correlation).

Based on the numerical illustration, we conclude the following: (i) The approach is able to recognize nonlinear structures in the data. (ii) The proposed local outlier score is a smooth measure. Further, local outlier scores of points in cluster cores are nearly identical and those in cluster halos are significantly larger. This indicates that locality dependent kernel width can enhance the power to discriminate in anomaly detection tasks. (iii) With the data refinement step, the online extension of the approach is more robust to the existence of anomalies in the training set. The approach is also more robust to the change of parameter k than is the LOF approach. (iv) The interpretability of the approach is much greater than other kernel methods which implicitly conduct nonlinear transformations from the input space to a feature space. (v) The experiment on the industrial dataset shows the applicability of the algorithm in real-world applications.

The following considerations are left to future work: (i) Our approach can be extended to detect faults in non-stationary data streams in a temporal context, using, for example, the sliding window strategy. (ii) The computation can be speeded up by using other smoothing kernel functions with compact support, but the impact of using another kernel function needs to be fully investigated.

Reference

[1] D. M. Hawkins, *Identification of outliers*. London: Chapman and Hall, 1980.

[2] L. Zhang, J. Lin, and R. Karim, "An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection," *Reliab. Eng. Syst. Saf.*, vol. 142, pp. 482–497, 2015.

[3] R. Göb, "Discussion of 'Reliability Meets Big Data: Opportunities and Challenges,'" *Qual. Eng.*, vol. 26, no. 1, pp. 121–126, 2013.

[4] C. Alippi, M. Roveri, and F. Trova, "A self-building and cluster-based cognitive fault diagnosis system for sensor networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 25, no. 6, pp. 1021–1032, 2014.

[5] X. Dai and Z. Gao, "From model, signal to knowledge: a data-driven perspective of fault detection and diagnosis," *IEEE Trans. Ind. Informatics*, vol. 9, no. 4, pp. 2226–2238, 2013.

[6] M. J. Gómez, C. Castejón, and J. C. García-Prada, "Automatic condition monitoring system for crack detection in rotating machinery," *Reliab. Eng. Syst. Saf.*, vol. 152, pp. 239–247, 2016.

[7] B. Cai, Y. Zhao, H. Liu, and M. Xie, "A Data-Driven Fault Diagnosis Methodology in Three-Phase Inverters for PMSM Drive Systems," *IEEE Trans. Power Electron.*, no. doi: 10.1109/TPEL.2016.2608842, 2016.

[8] B. Cai, Y. Liu, Q. Fan, Y. Zhang, Z. Liu, S. Yu, and R. Ji, "Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network," *Appl. Energy*, vol. 114, pp. 1–9, 2014.

[9] K. P. Murphy, *Machine learning: a probabilistic perspective*, 1st ed. MIT Press, 2012.

[10] L. Zhang, J. Lin, and R. Karim, "Sliding Window-Based Fault Detection From High-Dimensional Data Streams," *IEEE Trans. Syst. Man, Cybern. Syst.*, no. doi: 10.1109/TSMC.2016.2585566, 2016.

[11] J. Yu, "A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes," *Chem. Eng. Sci.*, vol. 68, no. 1, pp. 506–519, 2012.

[12] J. Kim and C. D. Scott, "Robust Kernel Density Estimation," *J. Mach. Learn. Res.*, vol. 13, no. Sep, pp. 2529–2565, 2012.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., vol. 1. Springer-Verlag New York, 2006.

[14] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF : Identifying Density-Based Local Outliers," *ACM Sigmod Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[15] H. Yu, F. Khan, and V. Garaniya, "Risk-based fault detection using Self-Organizing Map," *Reliab. Eng. Syst. Saf.*, vol. 139, pp. 82–96, 2015.

[16] C. M. Rocco S. and E. Zio, "A support vector machine integrated system for the classification of operation anomalies in nuclear components and systems," *Reliab. Eng. Syst. Saf.*, vol. 92, no. 5, pp. 593–600, 2007.

[17] C. Sun, Z. He, H. Cao, Z. Zhang, X. Chen, and M. J. Zuo, "A non-probabilistic metric derived from condition information for operational reliability assessment of aero-engines," *IEEE Trans. Reliab.*, vol. 64, no. 1, pp. 167–181, 2015.

[18] M. Markou and S. Singh, "Novelty detection: A review - Part 1: Statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[19] E. Schubert, A. Zimek, and H. P. Kriegel, "Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection," *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 190–237, 2014.

[20] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. CRC press, 1986.

[21] S. Dasgupta and A. Gupta, "An Elementary Proof of a Theorem of Johnson and Lindenstrauss," *Random Struct. Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[22] J. Lee, B. Kang, and S.-H. Kang, "Integrating independent component analysis and local outlier factor for plant-wide process monitoring," *J. Process Control*, vol. 21, no. 7, pp. 1011–1021, Aug. 2011.

[23] D. M. J. Tax and R. P. W. Duin, "Support Vector Data Description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.

[24] A. Nowicki, M. Grochowski, and K. Duzinkiewicz, "Data-driven models for fault detection using kernel PCA: A water distribution system case study," *Int. J. Appl. Math. Comput. Sci.*, vol. 22, no. 4, pp. 939–949, 2012.

[25] J. Lin, M. Asplunda, and A. Paridaa, "Reliability analysis for degradation of locomotive wheels using parametric bayesian approach," *Qual. Reliab. Eng. Int.*, vol. 30, no. 5, pp. 657–667, 2014.

[26] J. Lin, J. Pulido, and M. Asplund, "Reliability analysis for preventive maintenance based on classical and Bayesian semi-parametric degradation approaches using locomotive wheel-sets as a case study," *Reliab. Eng. Syst. Saf.*, vol. 134, pp. 143–156, 2015.