entropy

Article

Fault Detection Based on Multi-Dimensional KDE and Jensen–Shannon Divergence

Juhui Wei [1], Zhangming He [1,2,*], Jiongqi Wang [1], Dayi Wang [2] and Xuanying Zhou [1]

[1] College of Liberal Arts and Sciences, National University of Defense Technology, Changsha 410073, China; weijuhui_nudt@nudt.edu.cn (J.W.); wangjq@nudt.edu.cn (J.W.); zhouxy@nudt.edu.cn (X.Z.)
[2] Beijing Institute of Spacecraft System Engineering, China Academy of Space Technology, Beijing 100094, China; dayiwang@163.com
* Correspondence: hezhangming@nudt.edu.cn; Tel.: +86-1893-248-7430

Abstract: Weak fault signals, high coupling data, and unknown faults commonly exist in fault diagnosis systems, causing low detection and identification performance of fault diagnosis methods based on T^2 statistics or cross entropy. This paper proposes a new fault diagnosis method based on optimal bandwidth kernel density estimation (KDE) and Jensen–Shannon (JS) divergence distribution for improved fault detection performance. KDE addresses weak signal and coupling fault detection, and JS divergence addresses unknown fault detection. Firstly, the formula and algorithm of the optimal bandwidth of multidimensional KDE are presented, and the convergence of the algorithm is proved. Secondly, the difference in JS divergence between the data is obtained based on the optimal KDE and used for fault detection. Finally, the fault diagnosis experiment based on the bearing data from Case Western Reserve University Bearing Data Center is conducted. The results show that for known faults, the proposed method has 10% and 2% higher detection rate than T^2 statistics and the cross entropy method, respectively. For unknown faults, T^2statistics cannot effectively detect faults, and the proposed method has approximately 15% higher detection rate than the cross entropy method. Thus, the proposed method can effectively improve the fault detection rate.

Keywords: fault detection; optimal bandwidth; kernel density estimation; JS divergence; bearing

check for updates

Citation: Wei, J.; He, Z.; Wang, J.; Wang, D.; Zhou, X. Fault Detection Based on Multi-Dimensional KDE and Jensen–Shannon Divergence. *Entropy* **2021**, *23*, 266. https://doi.org/10.3390/e23030266

Academic Editors: Yongbo Li, Fengshou Gu and Xihui (Larry) Liang

Received: 30 January 2021
Accepted: 22 February 2021
Published: 24 February 2021

1. Introduction

The development of industrial informatization has given rise to a large amount of data in various fields. This has led to data processing becoming a difficult problem in the industry, especially for fault diagnosis. The explosive growth of data provides more information, and therefore, typical data analysis theories often fail in achieving the necessary results. The main reason for this failure can be attributed to the typical data analysis theory that often sets the data distribution type through prior information and performs analyses based on this assumption. Once the distribution type is set, the subsequent analysis can perform the estimation and parametric analysis based on only that distribution type; however, with the growth of data, more information is provided, and thus, the type of data distribution will need to be modified. As a nonparametric estimation method, kernel density estimation (KDE) is the most suitable method for the massive amount of the current data. KDE does not employ a priori assumption for the overall data distribution, and it directly starts from the sample data. When the sample size is sufficient, the KDE can approximate different distributions. Furthermore, Sheather and Jones [1] provides the optimal bandwidth estimation formula for a one-dimensional KDE and proves that the kernel function is asymptotically unbiased and consistent in the density estimation. However, with the growth of the dimension, the multidimensional KDE becomes more complex, and its optimal bandwidth formula is not provided. The distribution of multidimensional data has been described to a certain extent by estimating the kernel density of the reduced data in different dimensions Muir [2], Laurent [3]. In fact, the optimal KDE of multidimensional data is a problem that needs to be studied further.

In the field of fault diagnosis, an essential problem is measuring the difference between samples. A frequency histogram has been used to indicate the distribution difference between two samples Sugumaran and Ramachandran [4], Scott [5]; however, there are three shortcomings to this method: (1) the large number of discrete operations require a higher amount of time; (2) the process depends on the selection of the interval, which is more subjective; (3) there is no intuitive index to reflect this difference. In fact, based on KDE, the JS divergence can be used to measure the difference in data distribution, which can overcome the above shortcomings to a certain extent. For example, the failure of a rolling bearing, which is a key component of mechanical equipment, will have a serious effect on the safe and stable operation of the equipment, and the incipient fault detection of rolling bearings can help avoid equipment running with faults and avoid causing serious safety accidents and economic losses, which has important practical and engineering significance.

In Saruhan et al. [6], vibration analysis of rolling element bearings (REBs) defects is studied. The REBs are the most widely used mechanical parts in rotating machinery under high load and high rotational speeds. In addition, characteristics of bearing faults are analyzed in detail in references Razavi-Far et al. [7], Harmouche et al. [8]. Compared with traditional fault diagnosis, the fault diagnosis of rolling bearings is more complex:

- The fault signal is weak: Bearing data is a type of high-frequency data, and the fault signal is often covered by these high-frequency signals, thereby leading to the failure of traditional fault diagnosis methods. KDE is highly accurate in describing data distribution, so it can identify weak signals.
- Data is highly coupled: Bearing data is reflected in the form of a vibration signal, and there is strong coupling in different dimension signals, thereby making fault diagnosis difficult. Multi-dimensional KDE plays an important role in depicting the correlation of data, which can characterize the relationship between different dimensions of data.
- Incomplete data set: Most bearings work under normal conditions, and the fault data collected are often fewer, which makes the data incomplete, thereby resulting in the imperfection of the fault data set and increasing the difficulty of fault detection. The fault detection method constructed by JS divergence can deal with unknown faults and incomplete data sets without using additional data sets.

To overcome these problems, in-depth research has been conducted on this topic. Fault detection technology based on trend elimination and noise reduction has been proposed previously He et al. [9], Demetriou and Polycarpou [10]. The signal trend ratio is enhanced by eliminating the trend, and the signal–noise ratio is enhanced by noise reduction, and therefore, the fault detection effect is improved. However, this method uses the traditional detection method and cannot effectively solve the problem of data coupling. In reference Zhang et al. [11], Fu et al. [12], a fault detection method based on PCA dimension reduction and modal decomposition feature extraction is proposed. For multidimensional data, PCA dimension reduction is performed to reduce data dimensions and eliminate correlation between different dimensions. Then, the modal decomposition method is used to extract features among dimensions for fault detection. This method can effectively solve the strong coupling between data; however, it will lose some information in the process of PCA dimension reduction, and it leads to a reduction in the fault detection effect. In reference Itani et al. [13], Kong et al. [14], Jones and Sheather [15], Desforges et al. [16], a bearing fault detection method based on KDE is proposed. These studies analyzed the feasibility of KDE method in fault detection, and combined different classification methods for experiments. However, these methods only use one-dimensional KDE, and cannot directly describe high-dimensional data.

The data distribution is reconstructed by KDE and the cross-entropy function is constructed to measure the distribution difference for improving the fault detection results.

However, this method cannot reflect the correlation between different dimensions, and the cross-entropy function is not precise in the description of density distribution, which leads to a reduction in the fault detection effect, especially for unknown fault detection, which is not included in the fault set.

In this study, the KDE method is extended to multidimensional data to avoid information loss caused by the KDE for each dimension, and to better describe the density probability distribution of the data. Meanwhile, this study improves the traditional method using the cross-entropy function as the measurement of density distribution difference, and it uses JS divergence as the measurement of density distribution difference, thereby avoiding the relativity caused by the cross-entropy function. Most fault identification methods are based only on distance measurement; however, only relying on distance measurement cannot effectively detect unknown faults. Based on JS divergence, distribution characteristics of JS divergence between the sample density distribution and population density distribution are derived using the sliding window principle. Thus, the detection threshold of fault identification is assigned to realize the identification of unknown faults.

This paper is based on the following structure. In Section 2, the trend elimination method and detection method are introduced, and the intrinsic and extrinsic signals in the observation data are separated. Then, the fault detection threshold is constructed via statistics. In Section 3, the KDE method is extended to multidimensional data, and the optimal bandwidth is derived. Then, JS divergence is employed to measure the difference between probability distributions of different densities. In Section 4, the sliding window principle is used to sample the training data to obtain the distribution characteristics of JS divergence between the sample density distribution and the overall density distribution, and the detection threshold of fault identification is obtained using the KDE method. In Section 5, the normal data, two known faults, and one unknown fault are identified using the bearing data of the Case Western Reserve University Bearing Data Center as the fault diagnosis data. The experimental results show that the method can identify all types of faults well.

2. T^2 Statistics Fault Detection

In the operation process of the complex equipment or systems, the common observation state can be divided into intrinsic and extrinsic parts. In general, the intrinsic part represents the main working state of the system, which has a certain trend, monotony, and periodicity. The extrinsic part represents system noise, which has a certain zero mean value, high frequency vibration, and statistical stability. For the intrinsic part, the state equation of the system can be used to describe the law. When a fault occurs in the intrinsic part, the symptoms are relatively significant, and the corresponding fault detection methods are relatively mature. However, for high-frequency vibration signals, the incipient fault is often hidden in the extrinsic part, which is easily covered by noise. Therefore, it is necessary to analyze the observed data in depth.

2.1. Signal Decomposition

In the initial operation stage of the equipment, the unstable operation of the system causes large data fluctuations, which will not only have a great effect on the system trend, but also affect the statistical characteristics of the data. Therefore, it is necessary to truncate the data to remove unstable signals [9]. The corresponding time of the time series after removing the nonstationary period data is $t_1, t_2, \cdots, t_m$, and the following m observation data are obtained:

$$\boldsymbol{Y} = [\boldsymbol{y}(t_1), \boldsymbol{y}(t_2), \cdots, \boldsymbol{y}(t_m)]. \tag{1}$$

Each sampling $\boldsymbol{y}(t_i)$ contains n features, which are expressed as components in the form of

$$\boldsymbol{y}(t_i) = [y_1(t_i), y_2(t_i), \cdots, y_n(t_i)]^{\mathrm{T}}, i = 1, 2, \cdots, m. \tag{2}$$

Then, the data $\boldsymbol{Y}$ can be decomposed into

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{R}, \tag{3}$$

where $\hat{\boldsymbol{Y}}$ denotes the intrinsic part, which is composed of trend, and $\boldsymbol{R}$ denotes the extrinsic part, which is composed of observation noise and fault data.

The intrinsic part is composed of multiple signals. Selecting the appropriate basis function $\boldsymbol{f}(\mathrm{t}) = [f_1(t), f_2(t), \cdots, f_s(t)]^{\mathrm{T}}$ can help describe the intrinsic part. By traversing m data to model the nonlinear data $\boldsymbol{Y}$,

$$[\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_m] = \begin{bmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \cdots & \beta_s^{(1)} \\ \beta_1^{(2)} & \beta_2^{(2)} & \cdots & \beta_s^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_1^{(n)} & \beta_2^{(n)} & \cdots & \beta_s^{(n)} \end{bmatrix} \begin{bmatrix} f_0(t_1) & f_0(t_2) & \cdots & f_0(t_m) \\ f_1(t_1) & f_1(t_2) & \cdots & f_1(t_m) \\ \vdots & \vdots & \ddots & \vdots \\ f_s(t_1) & f_s(t_2) & \cdots & f_s(t_m) \end{bmatrix}. \tag{4}$$

Note that

$$\boldsymbol{F} \triangleq \begin{bmatrix} f_0(t_1) & f_0(t_2) & \cdots & f_0(t_m) \\ f_1(t_1) & f_1(t_2) & \cdots & f_1(t_m) \\ \vdots & \vdots & \ddots & \vdots \\ f_s(t_1) & f_s(t_2) & \cdots & f_s(t_m) \end{bmatrix}, \boldsymbol{\beta} \triangleq \begin{bmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \cdots & \beta_s^{(1)} \\ \beta_1^{(2)} & \beta_2^{(2)} & \cdots & \beta_s^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_1^{(n)} & \beta_2^{(n)} & \cdots & \beta_s^{(n)} \end{bmatrix} \tag{5}$$

Then, Equation (4) can be expressed as

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{F}. \tag{6}$$

Thus, the efficient estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{Y}\boldsymbol{F}^{\mathrm{T}}\left(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\right)^{-1}. \tag{7}$$

Using Equations (3) and (7), the signal can be decomposed into

$$\begin{cases} \hat{\boldsymbol{Y}} = \hat{\boldsymbol{\beta}}\boldsymbol{F} = \boldsymbol{Y}\boldsymbol{F}^{\mathrm{T}}\left(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\right)^{-1}\boldsymbol{F} \\ \boldsymbol{R} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y}\left(\boldsymbol{I} - \boldsymbol{F}^{\mathrm{T}}\left(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\right)^{-1}\boldsymbol{F}\right) \end{cases} \tag{8}$$

Usually, the choice of the basis function is a problem worthy of discussion, and it depends on prior knowledge of practical application scenarios; however, this is not the focus of this paper, and is therefore not covered here.

Remark 1. *For the bearing data, the data is generally stable and periodic. Therefore, Fourier transform is usually used to extract periodic features instead of more complex basis functions, such as a polynomial basis function and wavelet basis function.*

2.2. T^2 Statistics Detection

For simplicity, remember $\boldsymbol{r}_i = \boldsymbol{r}(t_i), i = 1, 2, \cdots, m$. According to Equation (8), the training data after signal decomposition are $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m]$, which is generally considered a normal random vector with expectation of $\boldsymbol{0}$, so that

$$\boldsymbol{r}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{9}$$

where $\mathbf{\Sigma}$ denotes the total covariance matrix. When the covariance matrix $\mathbf{\Sigma}$ is unknown, the unbiased estimation is given by

$$\hat{\mathbf{\Sigma}} = \frac{\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}}}{m-1}. \tag{10}$$

Let $\boldsymbol{Z} = \left[z_1, z_2, \ldots, z_p\right]$ be the data in the test window to be tested; the sample mean value $\bar{z}$ is

$$\bar{z} = \frac{1}{p}\sum_{i=1}^{p} z_i. \tag{11}$$

Then, $\bar{z}$ is still normal distributed and

$$\bar{z} \sim N\left(\mathbf{0}, \frac{1}{p}\mathbf{\Sigma}\right). \tag{12}$$

The T^2 statistics can be constructed as

$$T^2 = p\bar{z}^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{-1}\bar{z}. \tag{13}$$

Reference Solomons and Hotelling [17] reports that the distribution of the T^2 statistic satisfies

$$\frac{m-n}{n(m-1)}T^2 = \frac{p(m-n)}{n(m-1)}\bar{z}^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{-1}\bar{z} \sim F(n, m-n). \tag{14}$$

Therefore, if the significance level is α, we can get that

$$\frac{m-n}{n(m-1)}T^2 = \frac{l(m-n)}{n(m-1)}\bar{z}^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{-1}\bar{z} < F_{\alpha}(n, m-n). \tag{15}$$

The testing data $\boldsymbol{Z}$ and the training data $\boldsymbol{R}$ both come from the same mode; otherwise, they are considered different. The error rate of this criterion is α.

3. Optimal Kernel Density Estimation

Section 2 introduces the fault detection method based on T^2 statistics, including the signal decomposition technology and fault detection method based on the T^2 statistics. However, the fault detection method based on the T^2 statistics assumes that data satisfies the normal distribution, while the actual observation data may not meet the hypothesis, which can lead the discriminant performance of the T^2 statistics to not satisfy the design requirements. In addition, the statistics test the data from the angle of the intrinsic part $\hat{\boldsymbol{Y}}$ and covariance matrix $\hat{\mathbf{\Sigma}}$. These two attributes are not sufficient to describe all statistical characteristics of the system. When the incipient fault is submerged by data noise, it is easy to miss the detection. In this study, a KDE method for multidimensional data is constructed to describe the probability and statistical characteristics of the data more accurately.

3.1. Optimal Bandwidth Theorem

For the observed data, the frequency histogram can be used to show its statistical characteristics directly. However, in the actual application process, the frequency histogram is a discrete statistical method, the interval number of the histogram is difficult to divide, and more importantly, the discretization operation inconveniences the subsequent data processing. To overcome these limitations, the KDE method is proposed. This method is a nonparametric estimation method that estimates the population probability density distribution directly by sampling data.

For any point $\boldsymbol{x} \in \mathbb{R}^n$, assuming that the probability density of a certain mode is $f(\boldsymbol{x})$, the kernel density of $f(\boldsymbol{x})$ is estimated based on the sampling data $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m]$ in Section 2.1. As reported in reference Rao [18], the estimation formula is

$$\hat{f}_K(\boldsymbol{x}, h_m) = \frac{1}{m h_m^n} \sum_{i=1}^{m} K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right), \tag{16}$$

where m, n, $K(\cdot)$, and h_m denote the number of sampling data, dimension of sampling data, kernel function, and bandwidth, respectively.

For the sake of convenience in the following discussions, in the case of no doubt,

$$\begin{cases} \hat{f}_K(\boldsymbol{x}) \triangleq \hat{f}_K(\boldsymbol{x}, h_m) \\ \int g(\boldsymbol{x}) d\boldsymbol{x} \triangleq \int_{\boldsymbol{x} \in \mathbb{R}^n} g(\boldsymbol{x}) d\boldsymbol{x} \end{cases} \tag{17}$$

The kernel function $K(\cdot)$ satisfies $\int K(\boldsymbol{x}) d\boldsymbol{x} = 1$; therefore, $\int K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) d\boldsymbol{x} = h_m^n$, that is, $\int \hat{f}_K(\boldsymbol{x}) d\boldsymbol{x} = 1$. Thus, $\hat{f}_K(\boldsymbol{x})$ satisfies both positive definiteness, continuity, and normality. Therefore, it is reasonable to use it as the KDE. The Gaussian kernel function is a good choice as given by

$$K(\boldsymbol{x}) = (2\pi)^{-n/2} e^{-(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})/2} \tag{18}$$

In this study, the performance of the kernel density estimator is characterized by the mean integral square error (MISE).

$$\mathrm{MISE}\left(\hat{f}_K(\boldsymbol{x})\right) = \int \mathrm{E}\left[\hat{f}_K(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2 d\boldsymbol{x} \tag{19}$$

Reference Rao [18] shows that the estimation result $\hat{f}_K(\boldsymbol{x})$ is not sensitive to the selection of the kernel function $K(\cdot)$; that is, the MISE of the estimation results obtained using different kernel functions is almost the same, which is reflected in the subsequent derivation process. In addition, the MISE depends on the selection of the bandwidth h_m. If h_m is too small, the density estimation $\hat{f}_K(\boldsymbol{x})$ shows an irregular shape because of the increase in the randomness. While h_m is too large, density estimation $\hat{f}_K(\boldsymbol{x})$ is too averaged to show sufficient detail.

The optimal bandwidth formula is provided in the following theorem, and it is one of the key theoretical results of this study.

Theorem 1. *For any dimensional probability density function $f(\cdot)$ and any kernel function $K(\cdot)$ with a symmetric form, if $\hat{f}_K(\cdot)$ in Equation (16) is used to estimate $f(\cdot)$, and if the function* $\mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right)$ *with respect to $\boldsymbol{x}$ is integrable when the* $\mathrm{MISE}\left(\hat{f}_K(\cdot)\right)$ *in Equation (19) is the minimum, the bandwidth h_m satisfies*

$$h_m = \left(\frac{m d_K^2}{n^3 c_K} \int tr\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right)^2 d\boldsymbol{x}\right)^{-1/(n+4)}, \tag{20}$$

where c_K and d_K are two constant values given by

$$\begin{cases} c_K = \int K^2(\boldsymbol{x}) d\boldsymbol{x} \\ d_K = \int \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} K^2(\boldsymbol{x}) d\boldsymbol{x} \end{cases} \tag{21}$$

Equation (20) is called the optimal bandwidth formula and h_m denotes the optimal bandwidth.

A detailed proof of this theorem is given below.

Proof. It can be proved that the following two equations hold

$$\begin{cases} \mathrm{E}\left[\hat{f}_K(\boldsymbol{x})\right] = \int K(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u} \\ \mathrm{E}\left[\hat{f}_K^2(\boldsymbol{x})\right] = \dfrac{\int K^2(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u}}{mh_m^n} + \dfrac{(m-1)(\int K(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u})^2}{m} \end{cases} \tag{22}$$

In fact,

$$\begin{aligned} \mathrm{E}\left[\hat{f}_K(\boldsymbol{x})\right] &= \int \cdots \int \prod_{i=1}^{m} f(\boldsymbol{r}_i) \frac{1}{mh_m^n} \sum_{i=1}^{m} K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) d\boldsymbol{r}_m \cdots d\boldsymbol{r}_1 \\ &= \frac{1}{mh_m^n} \sum_{i=1}^{m} \int f(\boldsymbol{r}) K\left(\frac{\boldsymbol{r}-\boldsymbol{x}}{h_m}\right) d\boldsymbol{r} \\ &= \int f(\boldsymbol{x}+h_m\boldsymbol{u})K(\boldsymbol{u})d\boldsymbol{u}. \end{aligned} \tag{23}$$

In addition,

$$\begin{aligned} &\mathrm{E}\left[\hat{f}_K^2(\boldsymbol{x})\right] = \int \cdots \int \prod_{i=1}^{m} f(\boldsymbol{r}_i) \left((mh_m^n)^{-1} \sum_{i=1}^{m} f(\boldsymbol{r}_i) K\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right)\right)^2 d\boldsymbol{r}_1 \cdots d\boldsymbol{r}_m \\ &= (mh_m^n)^{-2} \int \cdots \int \prod_{i=1}^{m} f(\boldsymbol{r}_i) \left(\sum_{i=1}^{m} f(\boldsymbol{r}_i) K\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right)\right)^2 d\boldsymbol{r}_1 \cdots d\boldsymbol{r}_m \\ &= (mh_m^n)^{-2} \int \cdots \int \prod_{i=1}^{m} f(\boldsymbol{r}_i) \left(\sum_{i=1}^{m} K^2\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) + \sum_{i \neq j}^{m} K\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) K\left(\tfrac{\boldsymbol{r}_j - \boldsymbol{x}}{h_m}\right)\right) d\boldsymbol{r}_1 \cdots d\boldsymbol{r}_m \\ &= (mh_m^n)^{-2} \int \cdots \int \left(\prod_{i=1}^{m} f(\boldsymbol{r}_i) \sum_{i=1}^{m} K^2\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) + \prod_{i=1}^{m} f(\boldsymbol{r}_i) \sum_{i \neq j}^{m} K\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) K\left(\tfrac{\boldsymbol{r}_j - \boldsymbol{x}}{h_m}\right)\right) d\boldsymbol{r}_1 \cdots d\boldsymbol{r}_m \\ &= (mh_m^n)^{-2} \left(\sum_{i=1}^{m} \int f(\boldsymbol{r}_i) K^2\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) d\boldsymbol{r} + \sum_{i \neq j}^{m} \int \int \left(f(\boldsymbol{r}_i) K\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) f(\boldsymbol{r}_j) K\left(\tfrac{\boldsymbol{r}_j - \boldsymbol{x}}{h_m}\right)\right) d\boldsymbol{r}_i d\boldsymbol{r}_j \right) \\ &= (mh_m^n)^{-2} \left(m \int f(\boldsymbol{r}) K^2\left(\tfrac{\boldsymbol{r} - \boldsymbol{x}}{h_m}\right) d\boldsymbol{r} + m(m-1) \left(\int f(\boldsymbol{r}) K\left(\tfrac{\boldsymbol{r} - \boldsymbol{x}}{h_m}\right) d\boldsymbol{r}\right)^2 \right) \\ &= (mh_m^n)^{-2} \left(mh_m^n \int K^2(\boldsymbol{u}) f(\boldsymbol{x}+h_m\boldsymbol{u}) d\boldsymbol{u} + m(m-1)(h_m^n \int f(\boldsymbol{x}+h_m\boldsymbol{u}) K(\boldsymbol{u}) d\boldsymbol{u})^2 \right). \end{aligned} \tag{24}$$

From Equation (23),

$$E\left[\hat{f}_K(\boldsymbol{x})\right] - f(\boldsymbol{x}) = \frac{h_m^2}{2} \int \boldsymbol{u}^{\mathrm{T}} \left(\frac{\partial^2 f(\boldsymbol{x}+\theta h_m \boldsymbol{u})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right) \boldsymbol{u} K(\boldsymbol{u}) d\boldsymbol{u}, \tag{25}$$

where θ represents a constant value between 0 and 1. According to Equations (23) and (24),

$$E\left[\hat{f}_K^2(\boldsymbol{x})\right] - \left(E\left[\hat{f}_K(\boldsymbol{x})\right]\right)^2 = \frac{\int K^2(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u}}{mh_m^n} - \frac{(\int K(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u})^2}{m}. \tag{26}$$

According to the Equations (25) and (26), the following equation holds.

$$\begin{aligned} E\left[\hat{f}_K(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2 &= E\left[\hat{f}_K^2(\boldsymbol{x})\right] - \left(E\left[\hat{f}_K(\boldsymbol{x})\right]\right)^2 + \left(E\left[\hat{f}_K(\boldsymbol{x})\right] - f(\boldsymbol{x})\right)^2 \\ &= \frac{\int K^2(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u}}{mh_m^n} - \frac{(\int K(\boldsymbol{u})f(\boldsymbol{x}+h_m\boldsymbol{u})d\boldsymbol{u})^2}{m} \\ &+ \left(\frac{1}{2} h_m^2 \int \boldsymbol{u}^{\mathrm{T}} \left(\frac{\partial^2 f(\boldsymbol{x}+\theta h_m \boldsymbol{u})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right) \boldsymbol{u} K(\boldsymbol{u}) d\boldsymbol{u} \right)^2 \end{aligned} \tag{27}$$

To facilitate the subsequent reasoning, the following theorem is given.

Theorem 2. *For any matrix* $\boldsymbol{\Phi}$, $K(\cdot)$ *is a kernel density function with symmetric form; then,*

$$\int \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} = \frac{tr(\boldsymbol{\Phi})}{n}\int \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x}. \tag{28}$$

Proof. If the odd function $g(x)$ is integrable on $\mathbb{R}$, there must be $\int_{-\infty}^{\infty} g(x)dx = 0$. Similarly, it can be verified that the kernel function $K(\cdot)$ with a symmetric form satisfies

$$\int \cdots \int \sum_{i \neq j} \boldsymbol{\Phi}_{ij} x_i x_j K(\boldsymbol{x}) dx_1 \cdots dx_n = 0. \tag{29}$$

Then,

$$\begin{aligned}
\int \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} &= \int \cdots \int \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{x}K(\boldsymbol{x})dx_1 \cdots dx_n \\
&= \int \cdots \int \sum_{i} \boldsymbol{\Phi}_{ii} x_i^2 K(\boldsymbol{x}) dx_1 \cdots dx_n + \int \cdots \int \sum_{i \neq j} \boldsymbol{\Phi}_{ij} x_i x_j K(\boldsymbol{x}) dx_1 \cdots dx_n \\
&= \mathrm{tr}(\boldsymbol{\Phi}) \int \cdots \int x_1^2 K(\boldsymbol{x}) dx_1 \cdots dx_n \\
&= \frac{tr(\boldsymbol{\Phi})}{n} \int \cdots \int \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}K(\boldsymbol{x}) dx_1 \cdots dx_n \\
&= \frac{tr(\boldsymbol{\Phi})}{n} \int \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}K(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned} \tag{30}$$

Thus, the Theorem 2 is proved. □

For any unit length vector $\boldsymbol{u} \in \mathbb{R}^n$, the Taylor expansion can be used to obtain

$$\begin{cases} f(\boldsymbol{x} + h_m\boldsymbol{u}) = f(\boldsymbol{x}) + h_m\boldsymbol{u}^{\mathrm{T}}\nabla(f(\boldsymbol{x})) + o(h_m) \\ \dfrac{\partial^2 f(\boldsymbol{x} + \theta h_m \boldsymbol{u})}{\partial x_i \partial x_j} = \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} + \theta h_m \boldsymbol{u}^{\mathrm{T}} \nabla\left(\dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right) + o(h_m) \end{cases} \tag{31}$$

If the bandwidth h_m satisfies the condition

$$\begin{cases} \lim\limits_{m \to \infty} (h_m) = 0, \\ \lim\limits_{m \to \infty} \left(\dfrac{1}{m h_m^n}\right) = 0, \end{cases} \tag{32}$$

Then, from Equations (22)–(32), we get that

$$\mathrm{E}\left[\hat{f}_K(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2 = \frac{c_K f(\boldsymbol{x})}{m h_m^n} + o\left(\frac{1}{m h_m^n}\right) + \frac{h_m^4 d_K^2}{4n^2}\left(\mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right)\right)^2 + o\left(h_m^4\right). \tag{33}$$

In fact,

$$\begin{aligned}
\mathrm{E}\left[\hat{f}_K(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2 &= \frac{\int K^2(\boldsymbol{u}) f(\boldsymbol{x} + h_m \boldsymbol{u}) d\boldsymbol{u}}{m h_m^n} - \frac{\left(\int K(\boldsymbol{u}) f(\boldsymbol{x} + h_m \boldsymbol{u}) d\boldsymbol{u}\right)^2}{m} \\
&\quad + \left(\frac{h_m^2}{2} \int \boldsymbol{u}^{\mathrm{T}} \left(\frac{\partial^2 f(\boldsymbol{x} + \theta h_m \boldsymbol{u})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right) \boldsymbol{u} K(\boldsymbol{u}) d\boldsymbol{u}\right)^2 \\
&= \frac{c_K f(\boldsymbol{x})}{m h_m^n} + o\left(\frac{1}{m h_m^n}\right) - \frac{f(\boldsymbol{x})^2}{m} + o\left(\frac{1}{m}\right) + \left(\frac{h_m^2}{2n} \mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x} + \theta h_m \boldsymbol{u})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right) \int \boldsymbol{u}^{\mathrm{T}} \boldsymbol{u} K(\boldsymbol{u}) d\boldsymbol{u}\right)^2 \\
&= \frac{c_K f(\boldsymbol{x})}{m h_m^n} + o\left(\frac{1}{m h_m^n}\right) + \left(\frac{h_m^2}{2n} \mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right) d_K + o\left(h_m^2\right)\right)^2 \\
&= \frac{c_K f(\boldsymbol{x})}{m h_m^n} + o\left(\frac{1}{m h_m^n}\right) + \frac{h_m^4 d_K^2}{4n^2}\left(\mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right)\right)^2 + o\left(h_m^4\right).
\end{aligned} \tag{34}$$

Based on Equation (33), if $\mathrm{tr}\left(\frac{\partial^2 f(x)}{\partial x \partial x^{\mathrm{T}}}\right)$ is integrable, there is

$$\begin{aligned} \mathrm{MISE}\left(\hat{f}_K(\boldsymbol{x})\right) &= \int \left(\frac{c_K f(\boldsymbol{x})}{m h_m^n} + \frac{h_m^4}{4n^2}\left(d_K \mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right) \right)^2 \right) d\boldsymbol{x} + o\left(\frac{1}{m h_m^n}\right) + o(h_m) \\ &= \frac{c_K}{m h_m^n} + \frac{1}{4n^2} h_m^4 d_K^2 \int \mathrm{tr}\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right)^2 d\boldsymbol{x} + o\left(\frac{1}{m h_m^n}\right) + o(h_m). \end{aligned} \tag{35}$$

When $\mathrm{MISE}\left(\hat{f}_K(\cdot)\right)$ is the smallest, the derivative of Equation (35) with respect to h_m is 0, which means

$$\frac{\partial MISE\left(\hat{f}_K(\boldsymbol{x})\right)}{\partial h_m} = 0. \tag{36}$$

Thus, the optimal bandwidth h_m in Theorem 1 is obtained as

$$h_m = \left(\frac{m d_K^2}{n^3 c_K} \int tr\left(\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right)^2 d\boldsymbol{x} \right)^{-1/(n+4)}. \tag{37}$$

□

Remark 2. *When the number of samples m is determined, the appropriate bandwidth h_m can be selected using Equation (20) to construct the KDE, which can better fit the sample distribution. In Equation (20), the influence of the kernel function on bandwidth selection is on c_K and d_K, which are almost the same under different kernel function selection, and they have a slight effect on the final bandwidth selection.*

3.2. Optimal Bandwidth Algorithm

The optimal bandwidth formula is given by Equation (20). However, $f(\boldsymbol{x})$ is unknown in Equation (20), and therefore, $\int \mathrm{tr}\left(\frac{\partial^2 f(x)}{\partial x \partial x^{\mathrm{T}}}\right) d\boldsymbol{x}$ is also unknown. An approximate value of the bandwidth parameter h_m can be obtained by replacing $f(\boldsymbol{x})$ with $\hat{f}_K(\boldsymbol{x})$ in Equation (16). Furthermore, an iterative algorithm can be used to calculate a more accurate bandwidth parameter. Theorem 3 shows that the algorithm is convergent.

Theorem 3. *For any n-dimensional probability density function $f(\cdot)$ and Gaussian kernel function $K(\cdot)$, if $\hat{f}_K(\cdot)$ in Equation (16) is used to estimate $f(\cdot)$, then the iterative calculation formula of h_m is obtained as*

$$h_{m,k+1} = \left(\frac{m d_K^2}{n^3 c_K} \int \mathrm{tr}\left(\frac{\partial^2 \hat{f}_K(\boldsymbol{x}, h_{m,k})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right)^2 d\boldsymbol{x} \right)^{-1/(n+4)} \tag{38}$$

and it is convergent, where $h_{m,k}$ is the value of h_m during the $k-\mathrm{th}$ iteration.

Proof. For a particular Gaussian kernel function

$$K(\boldsymbol{u}) = (2\pi)^{-n/2} e^{-\left(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}\right)/2} \tag{39}$$

d_K is a χ^2 distribution with degree of freedom n, and the expectation is equal to the degree of freedom.

$$d_K = \int \boldsymbol{u}^{\mathrm{T}} \boldsymbol{u} K(\boldsymbol{u}) d\boldsymbol{u} = n \tag{40}$$

In addition,

$$c_K = \int K^2(\boldsymbol{u}) d\boldsymbol{u} = \int (2\pi)^{-n} e^{-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}} d\boldsymbol{u} = \left(2\sqrt{\pi}\right)^{-n}. \tag{41}$$

Substituting Equations (39)–(40) into Equation (20) and substituting $\hat{f}_K(x)$ in Equation (16) for $f(x)$, the iterative form of calculating h_m is obtained as

$$\begin{aligned} h_{m,k+1} &= \left(\tfrac{n}{m}\right)^{1/(n+4)} \left(2\sqrt{\pi}\right)^{-n/(n+4)} \left(\int \mathrm{tr}\left(\tfrac{\partial^2 \hat{f}_K(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right)^2 d\boldsymbol{x} \right)^{-1/(n+4)} \\ &= \left(mnh_{m,k}^{2n} \right)^{1/(n+4)} \left(2\sqrt{\pi}\right)^{-n/(n+4)} \left(\int \mathrm{tr}\left(\tfrac{\partial^2}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \left(\sum_{i=1}^{m} K\left(\tfrac{\boldsymbol{r}_i - \boldsymbol{x}}{h_{m,k}} \right) \right) \right)^2 d\boldsymbol{x} \right)^{-1/(n+4)} \end{aligned} \tag{42}$$

To facilitate the subsequent reasoning, the following lemma is given as

Lemma 1. *For any function $f_1, f_2, \cdots, f_n$, inequality*

$$\int (f_1 + f_2 + \cdots + f_n)^2 d\boldsymbol{x} \leq \int n\left(f_1^2 + f_2^2 + \cdots + f_n^2 \right) d\boldsymbol{x}. \tag{43}$$

If and only if $f_1(x) = f_2(x) = \cdots = f_n(x)$ holds almost everywhere.

Proof. In fact, for any function $f_1, f_2, \cdots, f_n$, there are

$$0 \leq (f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_n(\boldsymbol{x}))^2 \leq n\left(f_1(\boldsymbol{x})^2 + f_2(\boldsymbol{x})^2 + \cdots + f_n(\boldsymbol{x})^2 \right). \tag{44}$$

Thus, the two sides of Equation (44) are integrated as

$$\int (f_1 + f_2 + \cdots + f_n)^2 d\boldsymbol{x} \leq \int n\left(f_1^2 + f_2^2 + \cdots + f_n^2 \right) d\boldsymbol{x}. \tag{45}$$

It is obvious that the sign of Equation (43) holds the condition that $f_1(x) = f_2(x) = \cdots = f_n(x)$ is almost everywhere. □

Because the second derivative of Equation (39) with respect to x_i is

$$\frac{\partial^2}{\partial x_i \partial x_i} K(\boldsymbol{x}) = (2\pi)^{-n/2} e^{-\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)/2} \left(x_i^2 - 1 \right). \tag{46}$$

In addition,

$$\begin{aligned} \int \left(\frac{\partial^2}{\partial x_j \partial x_j} \left(K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_{m,k}} \right) \right) \right)^2 d\boldsymbol{x} &= \int \left(\frac{\partial^2}{\partial x_j \partial x_j} \left((2\pi)^{-n/2} e^{-(\boldsymbol{r}_i - \boldsymbol{x})^{\mathrm{T}} (\boldsymbol{r}_i - \boldsymbol{x}) \big/ 2h_{m,k}^2} \right) \right)^2 d\boldsymbol{x} \\ &= \frac{3}{4} \left(2\sqrt{\pi}\right)^{-n} h_{m,k}^{n-4}. \end{aligned} \tag{47}$$

From Lemma 1 and Equation (47)

$$\begin{aligned} \int tr\left(\frac{\partial^2}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \left(\sum_{i=1}^{m} K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_{m,k}} \right) \right) \right)^2 d\boldsymbol{x} &\leq \int nm \sum_{i,j} \left(\frac{\partial^2}{\partial x_j \partial x_j} \left(K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_{m,k}} \right) \right) \right)^2 d\boldsymbol{x} \\ &= \frac{3}{4} (nm)^2 \left(2\sqrt{\pi}\right)^{-n} h_{m,k}^{n-4}. \end{aligned} \tag{48}$$

When $h_{m,k}$ is sufficiently large, we can assume that $K\left(\frac{r_i - x}{h_{m,k}} \right)$ is almost the same everywhere, i.e., the equal sign in Equation (48) is tenable.

$$\begin{aligned} h_{m,k+1} &= \left(\frac{mnh_{m,k}^{2n}}{2\sqrt{\pi}} \right)^{1/(n+4)} \left(\frac{3}{4} (nm)^2 \left(2\sqrt{\pi}\right)^{-n} h_{m,k}^{n-4} \right)^{-1/(n+4)} \\ &= h_{m,k} \left(\frac{3}{4} nm \right)^{-1/(n+4)} < h_{m,k} \end{aligned} \tag{49}$$

When $h_{m,k}$ is large, the iterative process decreases. Because $h_{m,k}$ has a lower bound, the algorithm converges. □

In summary, the KDE method based on optimal bandwidth is given (see Algorithm 1), and the flowchart of the KDE method is shown in Figure 1.

Algorithm 1: Kernel density estimation (KDE) method based on optimal bandwidth

```
Input: Training set: $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m]$; Given the estimation accuracy: $\varepsilon$;
       Maximum number of iterations: $k_{max}$.
Output: Optimal bandwidth: $h_m$; Optimal KDE: $\hat{f}_K(\boldsymbol{x})$.
1  Select the initial iteration $h_{m,1} = h_{m,0}$;
2  for $k = 1, 2, \cdots, k_{max}$ do
3      Calculate the KDE $\hat{f}_K(\boldsymbol{x})$ using Equation (16)
4      Update the optimal bandwidth $h_{m,k}$ by Equation (38)
5      if $k < k_{max} \& |h_{m,k} - h_{m,k-1}| > \varepsilon$ then
6          $k = k + 1$, return step3;
7      else if $k = k_{max}$ then
8          Iteration times overrun, jump out;
9      else if $|h_{m,k} - h_{m,k-1}| < \varepsilon$ then
10         Obtain the optimal bandwidth.
11     end
12 end
```

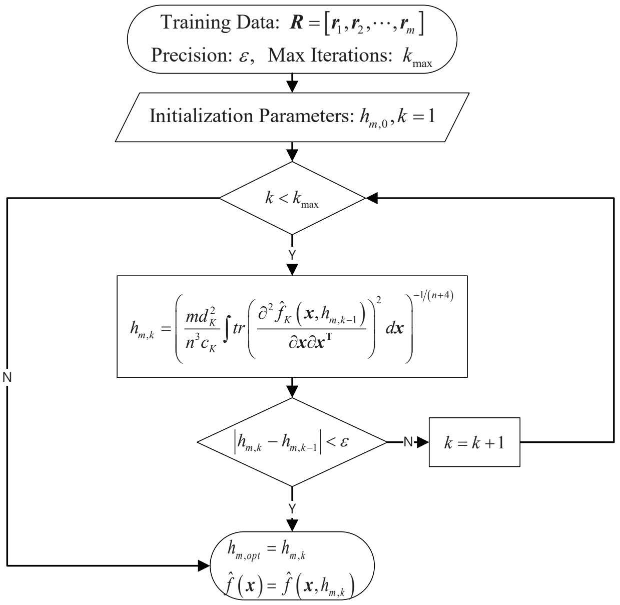

Training Data: $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m]$
Precision: ε, Max Iterations: $k_{\max}$
Initialization Parameters: $h_{m,0}, k = 1$
$k < k_{\max}$
Y
$h_{m,k} = \left(\frac{md_K^2}{n^3 c_K} \int tr \left(\frac{\partial^2 \hat{f}_K(\boldsymbol{x}, h_{m,k-1})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \right)^2 d\boldsymbol{x} \right)^{-1/(n+4)}$
N
$|h_{m,k} - h_{m,k-1}| < \varepsilon$
N
$k = k + 1$
Y
$h_{m,opt} = h_{m,k}$
$\hat{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}, h_{m,k})$

Figure 1. Flowchart of KDE method based on optimal bandwidth.

4. Fault Detection Method Based on JS Divergence Distribution

In Section 3, we construct a multidimensional KDE method based on the optimal bandwidth; this method can accurately describe the density distribution of multidimensional data. JS divergence is used to measure the distribution difference, and thus, it can highlight the difference in the statistical characteristics of different mode data.

4.1. Mode Difference Index

In Section 3, the probability density estimation of multidimensional data is obtained using the kernel function method, and the optimal bandwidth formula is derived. When the system fails, the state of the system will inevitably change, and the statistical characteristics of the system output will also change, thereby leading to significant changes in the density distribution of the observed data. For two groups of the sample window data $\boldsymbol{R}$ and $\boldsymbol{Z}$, the cross entropy $H(\boldsymbol{R}, \boldsymbol{Z})$ can be used to measure the distribution difference of $\boldsymbol{R}$ and $\boldsymbol{Z}$.

$$H(\boldsymbol{R}, \boldsymbol{Z}) = \int -\hat{f}_{K,Z}(\boldsymbol{x}) \log\left(\hat{f}_{K,R}(\boldsymbol{x})\right) d\boldsymbol{x}, \tag{50}$$

where $\hat{f}_{K,R}, \hat{f}_{K,Z}$ represents the optimal KDE of $\boldsymbol{R}$ and $\boldsymbol{Z}$ calculated using Equation (16).

$H(\boldsymbol{R}, \boldsymbol{Z})$ does not satisfy the definition of distance because $H(\boldsymbol{R}, \boldsymbol{Z})$ does not necessarily satisfy positive definiteness and symmetry; that is, $H(\boldsymbol{R}, \boldsymbol{Z}) < 0$ or $H(\boldsymbol{R}, \boldsymbol{Z}) \neq H(\boldsymbol{R}, \boldsymbol{Z})$.

- The smaller the difference of distribution, the smaller is $H(\boldsymbol{R}, \boldsymbol{Z})$, which means that even $H(\boldsymbol{R}, \boldsymbol{Z}) < 0$, and therefore, it is reasonable to use $H(\boldsymbol{R}, \boldsymbol{Z})$ to measure the distribution difference of $\boldsymbol{R}$ and $\boldsymbol{Z}$.
- However, the quantitative description of distribution difference must satisfy symmetry; otherwise, the exchange position and distribution difference will be different, which is difficult to accept.

The JS divergence $JS(\boldsymbol{R}, \boldsymbol{Z})$ was used as a measure of the distribution difference between $\boldsymbol{R}$ and $\boldsymbol{Z}$ in reference Zhang et al. [19], Bruni et al. [20] as follows:

$$JS(\boldsymbol{R}, \boldsymbol{Z}) = \int \begin{array}{ll} \hat{f}_{K,R} \log\left(\hat{f}_{K,R}\right) & +\hat{f}_{K,Z} \log\left(\hat{f}_{K,Z}\right) \\ & -\left(\hat{f}_{K,R} + \hat{f}_{K,Z}\right) \log\left(\left(\hat{f}_{K,R} + \hat{f}_{K,Z}\right)/2\right) \end{array} d\boldsymbol{x}. \tag{51}$$

It is easy to get that

$$\begin{cases} JS(\boldsymbol{R}, \boldsymbol{Z}) \geq 0 \\ JS(\boldsymbol{R}, \boldsymbol{Z}) = JS(\boldsymbol{Z}, \boldsymbol{R}) \end{cases} \tag{52}$$

In this paper, Equation (52) is used to measure the distribution difference between testing data $\boldsymbol{Z}$ and training data $\boldsymbol{R}$ for realizing fault detection and isolation.

4.2. Mode Discrimination Method

If the training data has q patterns $\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_q\}$, the JS divergence set

$$\{JS(\boldsymbol{Z}, \boldsymbol{R}_1), JS(\boldsymbol{Z}, \boldsymbol{R}_2), \cdots, JS(\boldsymbol{Z}, \boldsymbol{R}_q)\}$$

between the testing data $\boldsymbol{Z}$ and different modes $\boldsymbol{R}$ can be calculated using Equation (51).

If i_0 is the schema tag corresponding to the minimum JS divergence, it means that

$$i_0 = \arg\min\{JS(\boldsymbol{Z}, \boldsymbol{R}_1), JS(\boldsymbol{Z}, \boldsymbol{R}_2), \cdots, JS(\boldsymbol{Z}, \boldsymbol{R}_q)\}. \tag{53}$$

It is reasonable to assume that testing data $\boldsymbol{Z}$ and training data $\boldsymbol{R}_{i_0}$ belong to the same mode. However, for a new failure mode that may be unknown in the application, Equation (50) evaluates the testing data $\boldsymbol{Z}$ as the known failure mode of type i_0, which is obviously unreasonable.

If $JS(\boldsymbol{Z}, \boldsymbol{R}_{i_0})$ is too large, we believe that testing data $\boldsymbol{Z}$ comes from an unknown new failure mode; its label is $q+1$. However, the method to obtain the threshold JS_{high} of $JS(\boldsymbol{Z}, \boldsymbol{R}_{i_0})$ is a problem that should be investigated. A method to determine JS_{high} is provided below.

For the training data $\boldsymbol{R}_{i_0} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m]$ of the i_0 mode, the density estimation of the data set can be obtained using Equation (16).

$$\hat{f}_{K,R}(\boldsymbol{x}) = \frac{1}{m(h_m)^n} \sum_{i=1}^{m} K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_m}\right) \tag{54}$$

In addition, if the length of the sampling window is fixed as $p(p < m)$, the new sampling data is $\boldsymbol{R}^{(j)} = \left[\boldsymbol{r}_j, \boldsymbol{r}_{j+1}, \cdots, \boldsymbol{r}_{j+p}\right] \subset \boldsymbol{R}_{i_0}, j = 1, 2, \cdots, m-p$ by sliding the sampling window. For each $\boldsymbol{R}^{(j)}$, the density of the dataset can be estimated as

$$\hat{f}_{K,R^{(j)}}(\boldsymbol{x}) = \frac{1}{p(h_p)^n} \sum_{i=j}^{j+p} K\left(\frac{\boldsymbol{r}_i - \boldsymbol{x}}{h_p}\right). \tag{55}$$

Using Equation (52), the divergence between the training data $\boldsymbol{R}$ and the sample data $\boldsymbol{R}^{(j)}$ can be obtained as

$$\begin{aligned} JS_j &= JS\left(\boldsymbol{R}, \boldsymbol{R}^{(j)}\right) \\ &= H\left(\left(\hat{f}_{K,R} + \hat{f}_{K,R^{(j)}}\right), \left(\hat{f}_{K,R} + \hat{f}_{K,R^{(j)}}\right)/2\right) - H\left(\hat{f}_{K,R}\right) - H\left(\hat{f}_{K,R^{(j)}}\right). \end{aligned} \tag{56}$$

Using Equation (55), we can obtain a series of JS divergence calculation value sets

$$\boldsymbol{JS} = \left\{JS_1, JS_2, \cdots, JS_{m-p}\right\}.$$

We use this set to provide the estimation formula $\hat{f}_{JS}(x)$ of the density function $f_{JS}(x)$ of the JS divergence as

$$\hat{f}_{JS}(x) = \frac{1}{(m-p)(h_{m-p})^n} \sum_{j=1}^{m-p} K\left(\frac{JS_j - x}{h_{m-p}}\right). \tag{57}$$

If the significance level is α, the probability of $\hat{f}_{JS}(x)$ that exceeds the threshold JS_{high} is

$$P\left\{\int_0^{JS_{\text{high}}} \hat{f}_{JS}(x)dx.\right\} < \alpha \tag{58}$$

Because the distribution type of JS divergence is not a common random distribution, the quantile cannot be obtained by looking up the table; instead, it can only be obtained by numerical integration. If h is the step size, and

$$\int_{h*(i-1)}^{+\infty} \hat{f}_{JS}(x)dx \le \alpha \le \int_{h*i}^{+\infty} \hat{f}_{JS}(x)dx, \tag{59}$$

it is reasonable to deduce that

$$JS_{\text{high}} = h * i. \tag{60}$$

The following fault detection and isolation criteria are constructed by Equation (58).

Criterion 1. *Suppose i_0 is the pattern label corresponding to the minimum JS divergence—see Equation (38)—the training data $\mathbf{R}_{i_0} = [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_m]$ corresponding to the i_0 mode and the upper bound of JS divergence is JS_{high}—see Equation (56). If the testing data $\mathbf{Z} = [z_1, z_2, \ldots, z_l]$ meet the requirements,*

$$JS(\mathbf{Z}, \mathbf{R}_{i_0}) \le JS_{\text{high}}. \tag{61}$$

The testing data $\mathbf{Z}$ and training data $\mathbf{R}_{i_0}$ belong to the same failure mode; otherwise, the testing data $\mathbf{Z}$ are considered to originate from the unknown new failure mode, and their label is marked as $q+1$.

In conclusion, the fault diagnosis method based on optimal bandwidth is provided (See Algorithm 2), and the corresponding fault diagnosis method flowchart is shown in Figure 2.

Algorithm 2: Fault Diagnosis Method Based on Optimal KDE

```
Input: Training data: ${R_1, R_2, \cdots, R_p}$; Significance level: $\alpha$; Testing data:
       $Z = [z_1, z_2, \ldots, z_l]$.
Output: Pattern classification labels for testing data $Z$.
Calculate the optimal KDE $JS = \{JS_1, JS_2, \cdots, JS_{m-l}\}$ of $R$ by Algorithm 1;
Calculate the optimal KDE $\hat{f}_{K,Z}(x)$ of $Z$ by Algorithm 1;
Calculate the JS divergence set $\{JS(Z, R_1), JS(Z, R_2), \cdots, JS(Z, R_p)\}$ of $R$ and $Z$
  using Equation (51);
Calculate the minimum JS divergence label $i_0$ using Equation (53), and the
  corresponding training data were $R_{i_0} = [r_1, r_2, \cdots, r_m] \in R$;
for $j = 1, 2, \cdots, m - l$ do
    Get the training data $R^{(j)} = [r_j, r_{j+1}, \cdots, r_{j+l}] \subset R_{i_0}$ by sliding the windows;
    Calculate $\hat{f}_K(x)$ based on $h_{m,i}$, kernel function $K(\cdot)$, and Equation (16);
    Update the optimal bandwidth $h_{m,i}$ by Equation (37);
    Calculate the optimal KDE of $R^{(j)}$ using Algorithm 1 and Equation (55);
    Calculate $JS\left(\hat{f}_{K,R}, \hat{f}_{K,R^{(j)}}\right)$ according to Equation (56)
end
Calculate the density function of the JS divergence according to Equation (57)
  $f_{JS}(x)$;
Calculate the upper bound $JS_{\text{high}}$ of the JS divergence according to Equation (58)
  and 60;
Assess the pattern of testing data $Z$ according to Criterion 1,
```

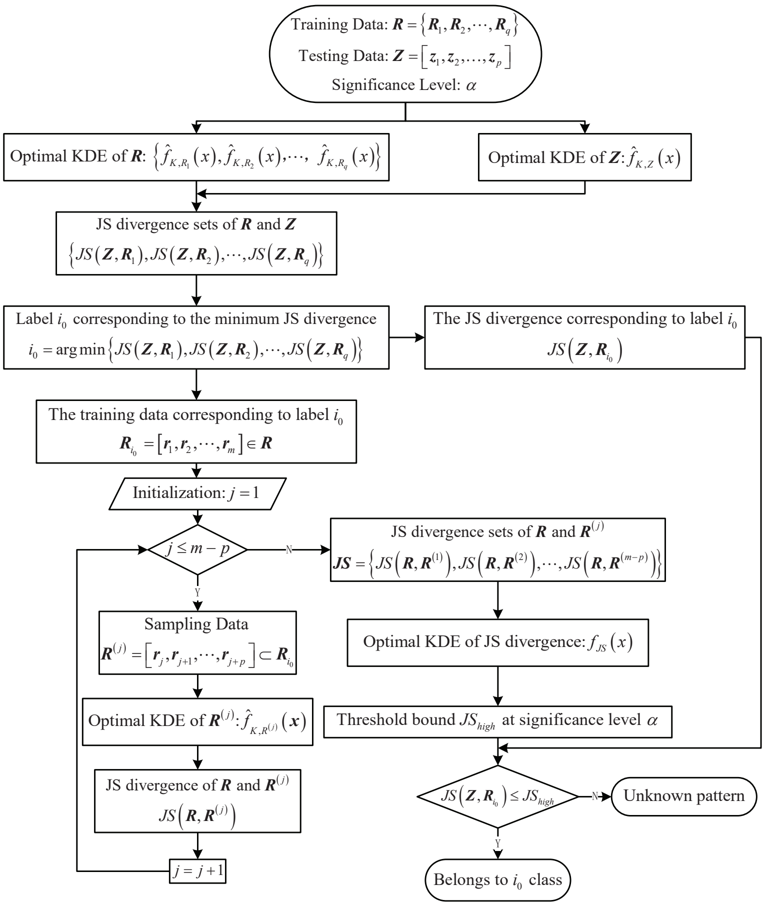

Training Data: $\boldsymbol{R} = \{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_q\}$
Testing Data: $\boldsymbol{Z} = [z_1, z_2, \ldots, z_p]$
Significance Level: α
Optimal KDE of $\boldsymbol{R}$: $\{\hat{f}_{K,R_1}(x), \hat{f}_{K,R_2}(x), \cdots, \hat{f}_{K,R_q}(x)\}$
Optimal KDE of $\boldsymbol{Z}$: $\hat{f}_{K,Z}(x)$
JS divergence sets of $\boldsymbol{R}$ and $\boldsymbol{Z}$
$\{JS(\boldsymbol{Z},\boldsymbol{R}_1), JS(\boldsymbol{Z},\boldsymbol{R}_2), \cdots, JS(\boldsymbol{Z},\boldsymbol{R}_q)\}$
Label i_0 corresponding to the minimum JS divergence
$i_0 = \arg\min\{JS(\boldsymbol{Z},\boldsymbol{R}_1), JS(\boldsymbol{Z},\boldsymbol{R}_2), \cdots, JS(\boldsymbol{Z},\boldsymbol{R}_q)\}$
The JS divergence corresponding to label i_0
$JS(\boldsymbol{Z},\boldsymbol{R}_{i_0})$
The training data corresponding to label i_0
$\boldsymbol{R}_{i_0} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m] \in \boldsymbol{R}$
Initialization: $j = 1$
$j \le m - p$
N
Y
JS divergence sets of $\boldsymbol{R}$ and $\boldsymbol{R}^{(j)}$
$\boldsymbol{JS} = \{JS(\boldsymbol{R},\boldsymbol{R}^{(1)}), JS(\boldsymbol{R},\boldsymbol{R}^{(2)}), \cdots, JS(\boldsymbol{R},\boldsymbol{R}^{(m-p)})\}$
Sampling Data
$\boldsymbol{R}^{(j)} = [\boldsymbol{r}_j, \boldsymbol{r}_{j+1}, \cdots, \boldsymbol{r}_{j+p}] \subset \boldsymbol{R}_{i_0}$
Optimal KDE of JS divergence: $f_{JS}(x)$
Optimal KDE of $\boldsymbol{R}^{(j)}$: $\hat{f}_{K,R^{(j)}}(\boldsymbol{x})$
Threshold bound JS_{high} at significance level α
JS divergence of $\boldsymbol{R}$ and $\boldsymbol{R}^{(j)}$
$JS(\boldsymbol{R},\boldsymbol{R}^{(j)})$
$JS(\boldsymbol{Z},\boldsymbol{R}_{i_0}) \le JS_{high}$
N
Unknown pattern
Y
$j = j + 1$
Belongs to i_0 class

Figure 2. Flowchart of fault diagnosis method based on optimal KDE.

Remark 3. *Equations (54) and (55) show that the calculation result of JS divergence is directly related to the length of sampling data. Indeed, with the increase in the sampling data length, the density estimation obtained by Equation (54) can describe the distribution characteristics of samples more effectively, thereby significantly improving the accuracy of fault detection.*

5. Numerical Simulation

The bearing data from Case Western Reserve University Bearing Data Center were used as the diagnosis research object, and they have been considered as a case for many fault diagnosis, such as in references Smith and Randall [21], Lou and Loparo [22], Rai and Mohanty [23].

The sampling frequency of the motor data was 12 kHz, and 12 kHz is the default sampling frequency for Case Western Reserve University Bearing Data Center. The dataset contains four groups of sample data: normal data (f_0), 0.007 inch inner raceway fault data (f_1), 0.014 inch inner raceway fault data (f_2), and 0.014 inch outer raceway fault data(f_3). Each group of data had two dimensions: the acceleration data of the drive end ($f_i - DE$) and the acceleration data of the fan end ($f_i - FE$). All the experiments were conducted on an Lenovo Ryzen 3700X CPU with 3.60 GHz processor, 16 GB RAM.

5.1. Data Preprocessing

The observed data in the process of the bearing operation show obvious periodicity, which needs to be eliminated. Taking normal data f_0 as an example, the main frequency in the observed signal can be obtained by fast Fourier transform (FFT), and the result of the FFT is shown in Figure 3.

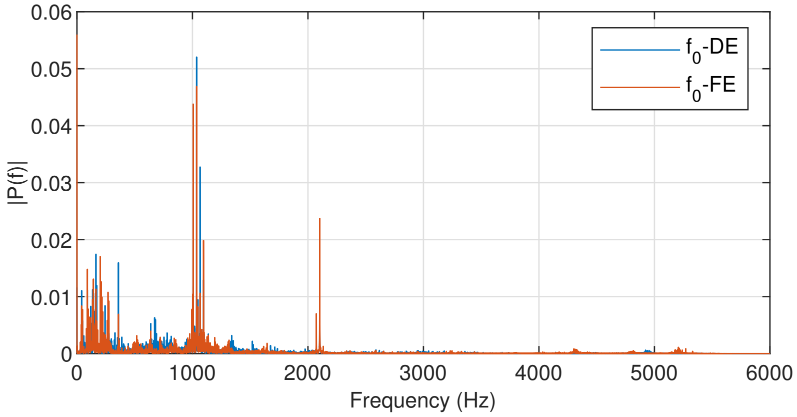


Figure 3. Single-sided amplitude spectrum of f_0.

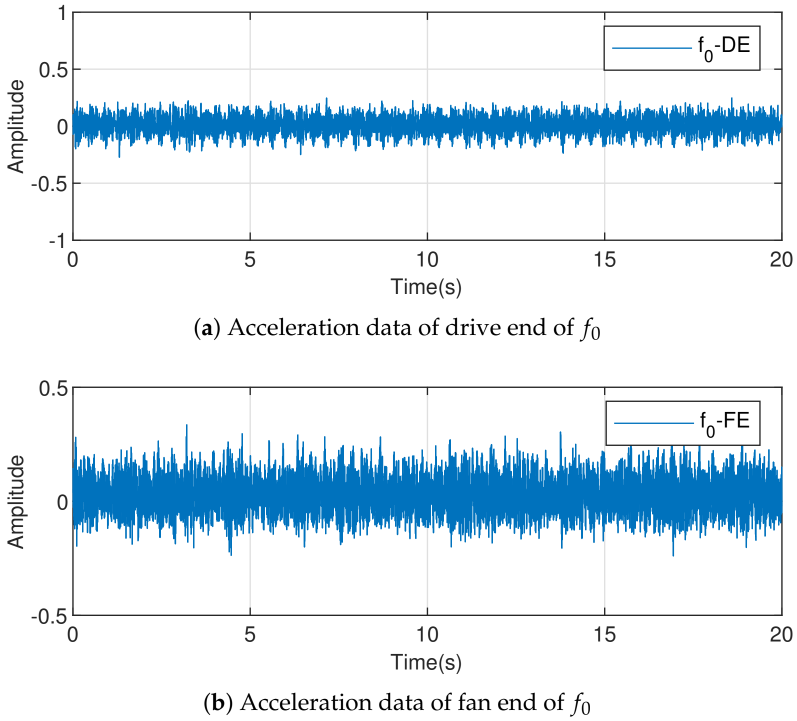
Figure 3 indicates that the main frequency is approximately 1036 Hz, and thus, the basis function is constructed as

$$\boldsymbol{f}(t) = \begin{bmatrix} 1 & \sin(1036 \times 2\pi t) & \cos(1036 \times 2\pi t) \end{bmatrix}^{\mathrm{T}}.$$

The estimation of $\boldsymbol{\beta}$ calculated using Equation (7) is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.0116 & -0.0158 & 0.0548 \\ 0.0280 & 0.0326 & -0.0396 \end{bmatrix}.$$

Thus, the data after removing the intrinsic signal are shown in Figure 4, where Figure 4a represents the acceleration data of the drive end and Figure 4b represents the acceleration data of the fan end.



(**a**) Acceleration data of drive end of f_0

(**b**) Acceleration data of fan end of f_0

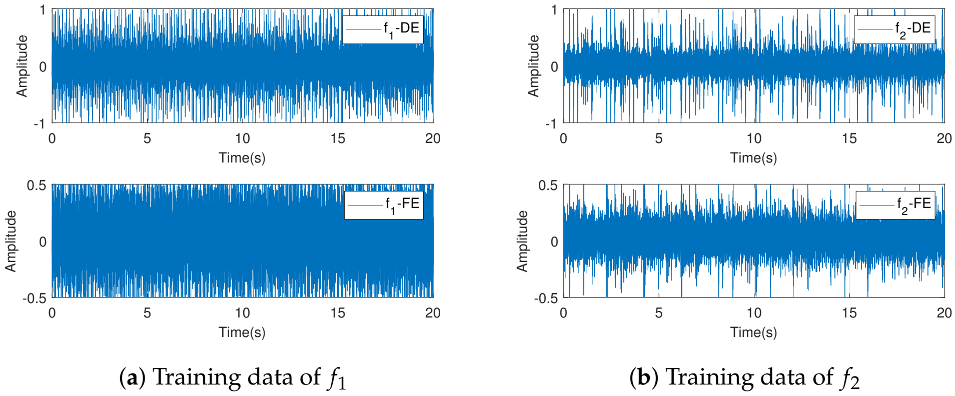
Figure 4. Preprocessed data to remove trends by fast Fourier transform (FFT).

In the later fault detection process, the data of all modes are similar to the above operation, and the results are recorded as f_i.

5.2. Fault Detection Effect

5.2.1. Norm Data and Known Fault

For the norm data f_0 and the known fault f_1, f_2, the first 20,480 sample points are selected as the training set, which are recorded as $f_{i-\text{train}}$. The last 81,920 sample points are taken as the testing set, which are recorded as $f_{i-\text{test}}$. A total of 128 sample points are used as detection objects in each test. The training set data are shown in Figure 5, where Figure 5a,b represent data $f_{i-\text{train}}, i = 1, 2$ of the two dimensions, respectively.



(**a**) Training data of f_1

(**b**) Training data of f_2

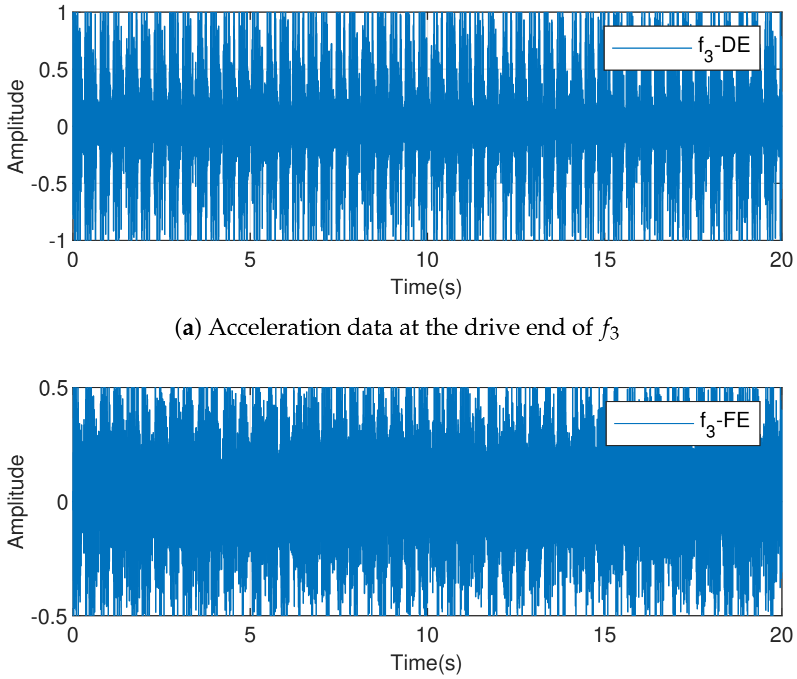
Figure 5. Training data f_1, f_2 after being preprocessed.

Figure 5 shows that the bearing data have high frequency, and the fault does not change the observed mean value; however, it changes the dispersion characteristics or the correlation of data.

5.2.2. Unknown Fault

The training data does not necessarily contain all types of patterns, and the detection of unknown faults is always a difficult problem.f_3 is used as an unknown fault for fault detection; the training set sample does not contain any information about f_3. The unknown fault data are shown in Figure 6, where in Figure 6a represents the acceleration data at the driving end and Figure 6b represents the acceleration data at the fan end.

Figure 6 shows that the data of unknown faults is close to the other two types of fault data. If the fault detection method is not sensitive, the detection rate will be reduced significantly.



(**a**) Acceleration data at the drive end of f_3

(**b**) Acceleration data at the fan end of f_3

Figure 6. Training data f_3 after preprocessed.

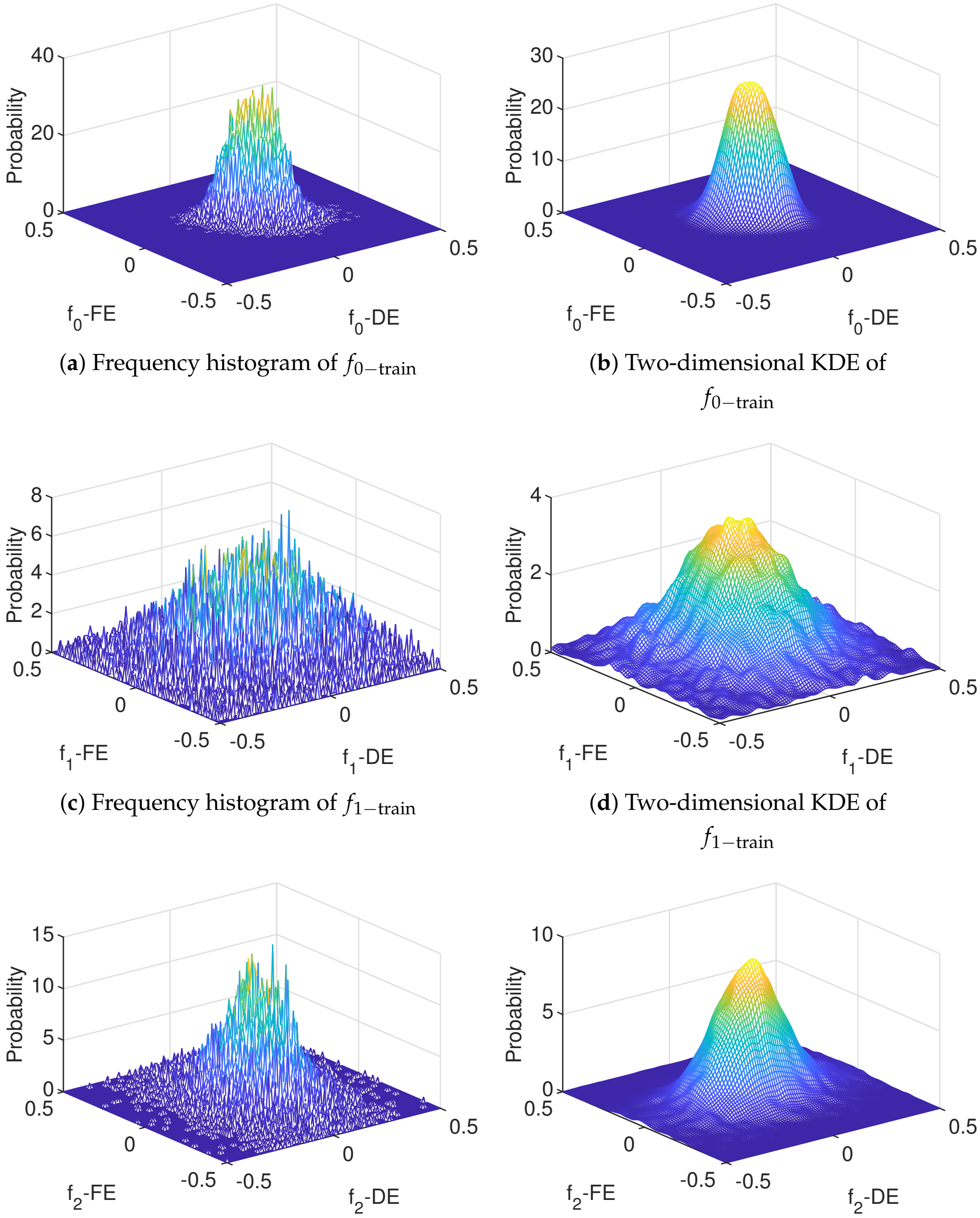
5.2.3. Detection Effect

The characteristics of bearing data make bearing fault detection extremely challenging. The input of the training set is $f_{0-\text{train}}$, the estimation accuracy is $\varepsilon = 10^{-4}$, and the maximum number of iterations is $k_{max} = 100$, according to Algorithm 1, the optimal bandwidth is

$$h_m = 0.0445.$$

The KDE of the training set is obtained by Equation (15), and the results are shown in Figure 7, where Figure 7a,c,e represent the two-dimensional frequency histograms of the training data $f_{i-\text{train}}, i = 0, 1, 2$, and Figure 7b,d,f represent the two-dimensional KDE of the training data $f_{i-\text{train}}, i = 0, 1, 2$.

Figure 7 further shows that the bearing fault changes the dispersion characteristics and data correlation. Meanwhile, Figure 7 shows that the KDE of the training data obtained by Equation (15) is in good agreement with the data distribution of the training data, and therefore, this method can really describe the distribution of multidimensional data.



(**a**) Frequency histogram of $f_{0-\text{train}}$

(**b**) Two-dimensional KDE of $f_{0-\text{train}}$

(**c**) Frequency histogram of $f_{1-\text{train}}$

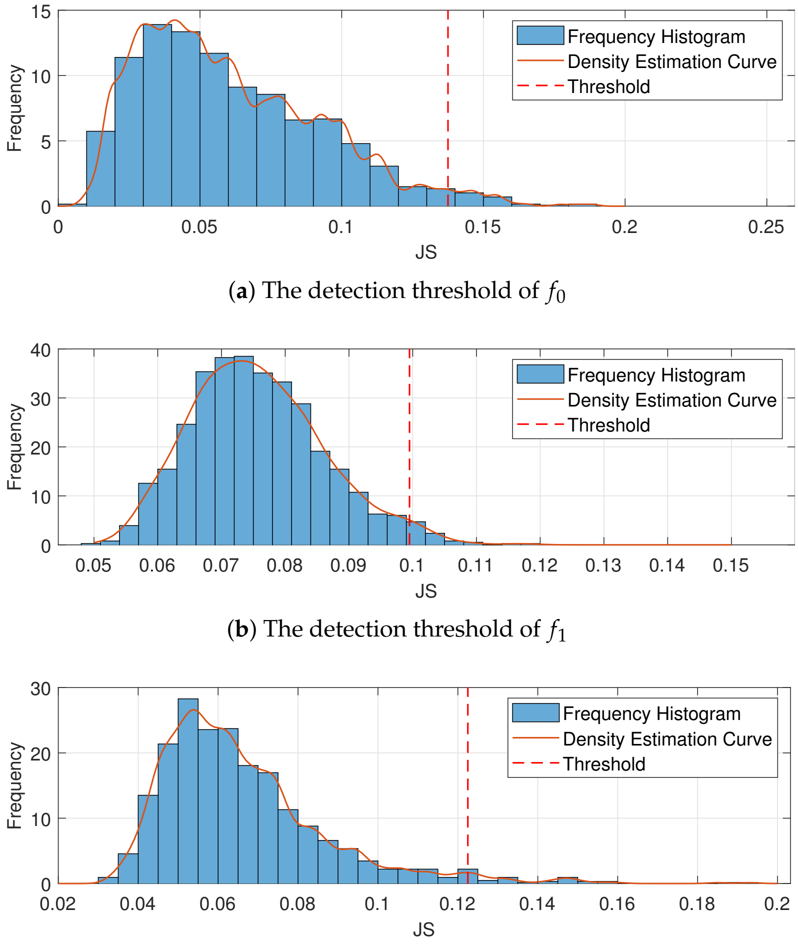
(**d**) Two-dimensional KDE of $f_{1-\text{train}}$

(**e**) Frequency histogram of $f_{2-\text{train}}$

(**f**) Two-dimensional KDE of $f_{2-\text{train}}$

Figure 7. Training data after being preprocessed.

The JS divergence of the training data and KDE of the distribution are obtained by Equations (51) and (58); the results are shown in Figure 8.



(**a**) The detection threshold of f_0

(**b**) The detection threshold of f_1

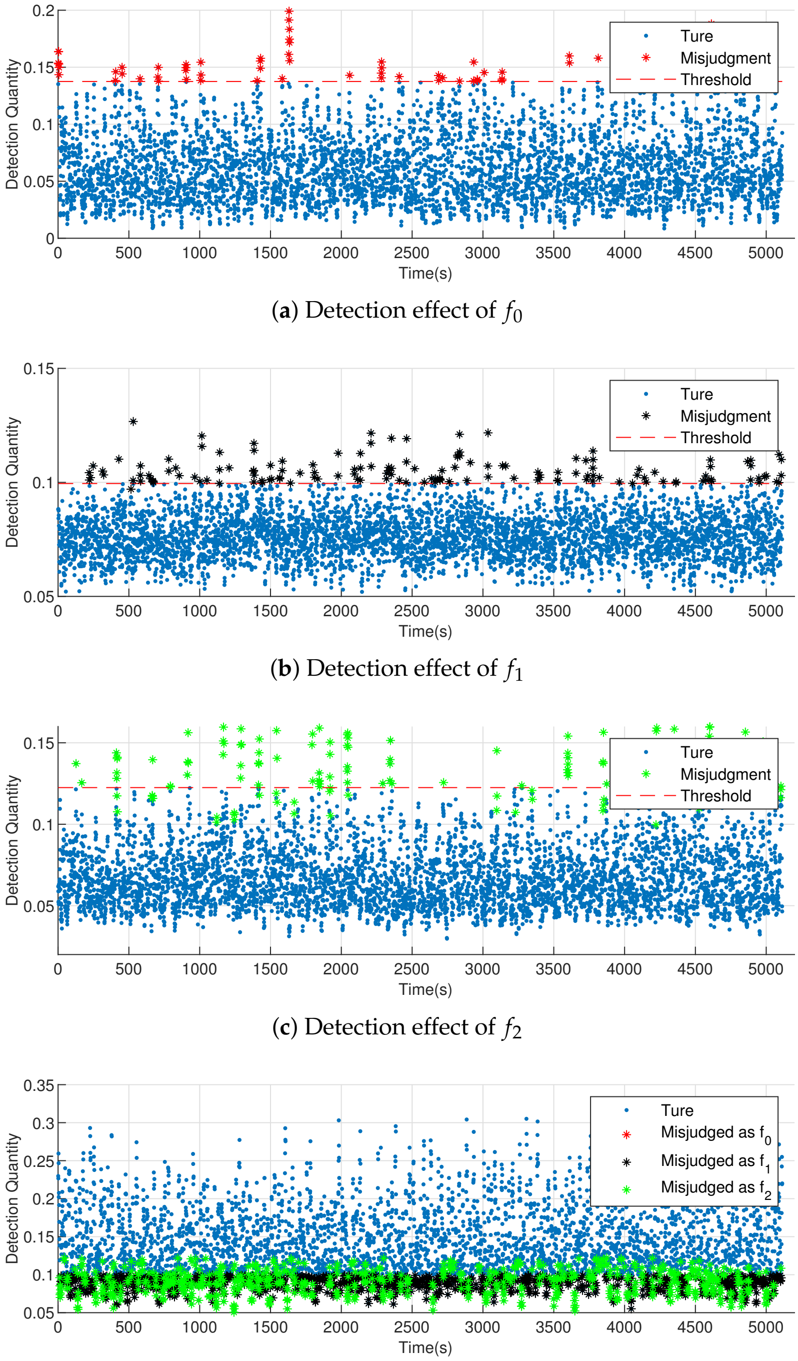
(**c**) The detection threshold of f_2

Figure 8. The results of detection threshold.

When the significance level is $\alpha = 95\%$, the detection thresholds of the training set, which are calculated using Equation (58), are

$$\begin{cases} f_0 : JS_{\text{high}} < 0.1375 \\ f_1 : JS_{\text{high}} < 0.0995 \\ f_2 : JS_{\text{high}} < 0.1225 \end{cases}$$

Thus, the detection results of using JS divergence methods on the testing data are shown in Figure 9. If the detection points fall within the threshold, the data set to be detected is in the same pattern; otherwise, the data have different patterns.



(**a**) Detection effect of f_0

(**b**) Detection effect of f_1

(**c**) Detection effect of f_2

(**d**) Detection effect of f_3

Figure 9. Fault detection effect using JS divergence as index.

Furthermore, detection rates using different methods are shown in Table 1.

Table 1. Detection rate of normal and different failure modes using different methods.

Method	T^2 Statistics Detection	Cross Entropy	JS Divergence
Normal mode f_0	95.80%	96.95%	97.03%
Known fault f_1	83.47%	94.41%	95.81%
Known fault f_2	78.11%	94.19%	95.36%
Unknown fault f_3	\	53.16%	69.49%

For the known fault, Table 1 indicates that the bearing fault identification based on multidimensional KDE and JS divergence achieves better results compared to those obtained using the T^2 statistics detection methods in the testing data. The detection rate of normal data f_0 increases from 95.08% to 97.03%, the detection rate of fault data f_1 increases from 81.33% to 95.81%, and the detection rate of fault data f_2 increases from 70.69% to 95.36%. Meanwhile, compared with the cross-entropy methods, the detection rate of normal data f_0 increased from 96.95% to 97.03%; of fault data f_1 increased from 94.41% to 95.81%; and of fault data f_2 increased from 94.19% to 95.36%.

For the unknown fault f_3, Table 1 shows that the T^2 statistics detection method cannot detect the unknown faults. The method using cross entropy as a measure can only detect unknown faults with a detection rate of 53.16%, which is not obvious. The JS divergence method constructed in this study can identify the unknown fault accurately, and the detection rate reaches 69.49%. This is because JS divergence is more accurate at measuring the differences between distributions.

5.3. Influence of Window Width on Fault Diagnosis

The fault diagnosis effect is related to the data window width; therefore, the fault diagnosis effect under different window widths is investigated. The results are shown in Figure 10.

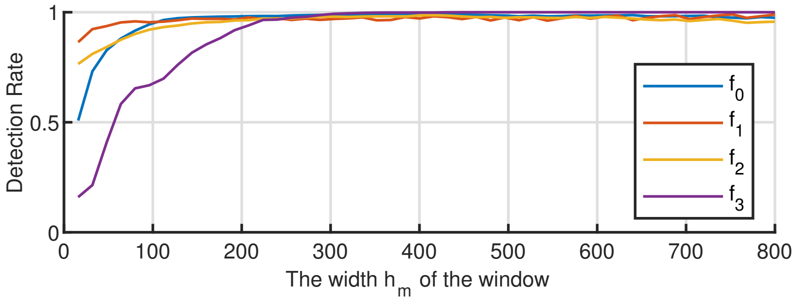


Figure 10. Fault diagnosis effect under different window width h_m.

Figure 10 indicates that, with the increase in the detection window, the detection performance of the proposed method for the known fault detection first rises, and then, it tends to be stable. This is because when the length of the detection window increases to a certain extent, the data to be detected already contains sufficient information. Meanwhile, if the detection window continues to increase, the contribution rate to the improvement of the fault detection rate is not large. Meanwhile, for unknown faults, the detection rate increases rapidly with the length of the detection window because the longer the detection window, the higher the amount of information contained in the data to be detected, and the better is the difference characterized between the fault and the known fault.

6. Conclusions

In this study, a method of bearing fault detection and identification was constructed using multidimensional KDE and JS divergence. The distribution characteristics of JS divergence between the sample density distribution and population density distribution were derived using the sliding sampling window method. Thus, the threshold of fault detection was provided, and therefore, different faults, especially unknown faults, could be identified. The theory showed that the multidimensional KDE method could reduce information loss caused by processing each dimension; the JS divergence is more accurate than the traditional cross entropy to measure the difference in density distribution. The experimental results verified the above conclusions.

For a known fault, the detection effect of this method was obviously better than that of the traditional method, and it also had a certain degree of improvement compared with the cross-entropy method. Second, for unknown faults, the traditional method could not detect the distribution difference accurately, while the detection effect of the proposed method was significantly improved.

Furthermore, the detection effect of this method depends on the window width. The detection effect improved with a growth in the detection window. In this paper, under the condition of a given window width, the estimation formula for the optimal bandwidth of a multidimensional KDE was provided. The experimental results showed that the formula was applicable to any mode of data, and therefore, it had a certain universality.

However, this study has certain limitations. Firstly, although the calculation formula of multidimensional KDE is given in this study, the computational complexity will increase when the dimension is large, which may restrict the further application of the method. Secondly, the calculation of JS divergence is time consuming, which is not conducive to rapid fault diagnosis.

In future research, we can try to use the PCA dimension reduction method to solve the computational complexity caused by very large dimension, and optimize the algorithm flow of JS divergence to expedite the calculation. In the latest study Ginzarly et al. [24], prognosis of the vehicle's electrical machine is treated using a hidden Markov model after modeling the electrical machine using the finite element method. Therefore, we will try to combine this method in future work and apply it to the fault detection of other systems.

Author Contributions: Conceptualization and methodology, J.W. (Juhui Wei); formal analysis and visualization, Z.H.; validation and data curation, J.W. (Jiongqi Wang); resources, D.W.; writing—review and editing X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant number 61903366, 61903086, 61773021, Natural Science Foundation of Hunan Province grant number 2019JJ50745, 2019JJ20018, 2020JJ4280 and Foundation of Beijing Institute of Control Engineering grant number HTKJ2019KL502007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors express their appreciation to the Associate Editor and anonymous reviewers for their helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KDE	kernel density estimation
JS	Jensen–Shannon
PCA	principal component analysis
MISE	mean integral square error

References

1. Sheather, S.J.; Jones, M.C. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc.* **1991**, *53*, 683–690. [CrossRef]
2. Muir, D. Multidimensional Kernel Density Estimates over Periodic Domains. Circular Statistics. 2017. Available online: https://www.mathworks.com/matlabcentral/fileexchange/44129-multi-dimensional-kernel-density-estimates-over-periodic-domains (accessed on 21 February 2021).
3. Laurent, B. Efficient estimation of integral functionals of a density. *Ann. Stat.* **1996**, *24*, 659–681. [CrossRef]
4. Sugumaran, V.; Ramachandran, K.I. Fault diagnosis of roller bearing using fuzzy classifier and histogram features with focus on automatic rule learning. *Expert Syst. Appl.* **2011**, *38*, 4901–4907. [CrossRef]
5. Scott, D.W. Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *Ann. Stat.* **1985**, *13*, 1024–1040. [CrossRef]
6. Saruhan, H.; Sardemir, S.; Iek, A.; Uygur, L. Vibration analysis of rolling element bearings defects. *J. Appl. Res. Technol.* **2014**, *12*, 384–395. [CrossRef]

7. Razavi-Far, R.; Farajzadeh-Zanjani, M.; Saif, M. An integrated class-imbalanced learning scheme for diagnosing bearing defects in induction motors. *IEEE Trans. Ind. Inform.* **2017**, *13*, 2758–2769. [CrossRef]
8. Harmouche, J.; Delpha, C.; Diallo, D. Incipient fault amplitude estimation using kl divergence with a probabilistic approach. *Signal Process.* **2016**, *120*, 1–7. [CrossRef]
9. He, Z.; Shardt, Y.A.W.; Wang, D.; Hou, B.; Zhou, H.; Wang, J. An incipient fault detection approach via detrending and denoising. *Control Eng. Pract.* **2018**, *74*, 1–12. [CrossRef]
10. Demetriou, M.A.; Polycarpou, M.M. Incipient fault diagnosis of dynamical systems using online approximators. *IEEE Trans. Autom. Control* **1998**, *43*, 1612–1617. [CrossRef]
11. Zhang, X.; Polycarpou, M.M.; Parisini, T. A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems. *IIEEE Trans. Autom. Control* **2002**, *47*, 576–593.
12. Fu, F.; Wang, D.; Li, W.; Li, F. Data-driven fault identifiability analysis for discrete-time dynamic systems. *Int. J. Syst. Sci.* **2020**, *51*, 404–412. [CrossRef]
13. Itani, S.; Lecron, F.; Fortemps, P. A one-class classification decision tree based on kernel density estimation. *Appl. Soft Comput.* **2020**, *91*, 106250. [CrossRef]
14. Kong, Y.; Li, D.; Fan, Y.; Lv, J. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Ann. Stat.* **2017**, *45*, 897–922. [CrossRef]
15. Jones, M.C.; Sheather, S.J. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Stat. Probab. Lett.* **1991**, *11*, 511–514. [CrossRef]
16. Desforges, M.J.; Jacob, P.J.; Ball, A.D. Fault detection in rotating machinery using kernel-based probability density estimation. *Int. J. Syst. Sci.* **2000**, *31*, 1411–1426. [CrossRef]
17. Solomons, L.M.; Hotelling, H. The limits of a measure of skewness. *Ann. Math. Stat.* **1932**, *3*, 141–142.
18. Rao, P. *Nonparametric Functional Estimation*; Elsevier: Amsterdam, The Netherlands, 1983.
19. Zhang, X.; Delpha, C.; Diallo, D. Incipient fault detection and estimation based on Jensen–Shannon divergence in a data-driven approach. *Signal Process.* **2019**, *169*, 107410. [CrossRef]
20. Bruni, V.; Rossi, E.; Vitulano, D. On the equivalence between Jensen–Shannon divergence and Michelson contrast. *IEEE Trans. Inf. Theory* **2012**, *58*, 4278–4288. [CrossRef]
21. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64–65*, 100–131. [CrossRef]
22. Lou, X.; Loparo, K.A. Bearing fault diagnosis based on wavelet transform and fuzzy inference. *Mech. Syst. Signal Process.* **2004**, *18*, 1077–1095. [CrossRef]
23. Rai, V.K.; Mohanty, A.R. Bearing fault diagnosis using fft of intrinsic mode functions in Hilbert–Huang transform. *Mech. Syst. Signal Process.* **2007**, *21*, 2607–2615. [CrossRef]
24. Ginzarly, R.; Hoblos, G.; Moubayed, N. From modeling to failure prognosis of permanent magnet synchronous machine. *Appl. Sci.* **2020**, *10*, 691. [CrossRef]