*mathematics*

MDPI

*Article*

# An Improved Model for Kernel Density Estimation Based on Quadtree and Quasi-Interpolation

**Jiecheng Wang [1], Yantong Liu [2,*] and Jincai Chang [3]**

[1] Department of Statistics, School of Economics, Anhui University, Hefei 230039, China; i20101008@stu.ahu.edu.cn
[2] School of Management, North China University of Science and Technology, Tangshan 063210, China
[3] School of Science, North China University of Science and Technology, Tangshan 063210, China; jincai@ncst.edu.cn
* Correspondence: lytok1114@126.com

**Abstract:** There are three main problems for classical kernel density estimation in its application: boundary problem, over-smoothing problem of high (low)-density region and low-efficiency problem of large samples. A new improved model of multivariate adaptive binned quasi-interpolation density estimation based on a quadtree algorithm and quasi-interpolation is proposed, which can avoid the deficiency in the classical kernel density estimation model and improve the precision of the model. The model is constructed in three steps. Firstly, the binned threshold is set from the three dimensions of sample number, width of bins and kurtosis, and the bounded domain is adaptively partitioned into several non-intersecting bins (intervals) by using the iteration idea from the quadtree algorithm. Then, based on the good properties of the quasi-interpolation, the kernel functions of the density estimation model are constructed by introducing the theory of quasi-interpolation. Finally, the binned coefficients of the density estimation model are constructed by using the idea of frequency replacing probability. Simulation of the Monte Carlo method shows that the proposed non-parametric model can effectively solve the three shortcomings of the classical kernel density estimation model and significantly improve the prediction accuracy and calculation efficiency of the density function for large samples.

**Keywords:** large samples; kernel density estimation; quasi-interpolation; quadtree algorithm; adaptive

**MSC:** 62G07; 41A63

check for updates

**Citation:** Wang, J.; Liu, Y.; Chang, J. An Improved Model for Kernel Density Estimation Based on Quadtree and Quasi-Interpolation. *Mathematics* **2022**, *10*, 2402. https://doi.org/10.3390/math10142402

Academic Editor: Christophe Chesneau

Received: 19 June 2022
Accepted: 7 July 2022
Published: 8 July 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Density estimates are a common technique in modern data analysis. They are usually used to analyze statistical characteristics, such as skewness and multimodality of samples, and quantify uncertainties. They have been widely used in engineering, economics, medicine, geography and other fields. The methods of density estimation contain the parametric method and nonparametric method. The parametric method requires strong assumptions for the prior model to restrict the probability density function drawn from a given parametric family of distribution, and then calculates the corresponding parameter estimates from the samples. The main problem of the parametric method is that inaccurate setting of the prior parameter model may lead to wrong conclusions. Moreover, in the process of testing the posterior model, there is a common situation that multiple assumptions of prior models can pass a posterior test, which greatly affects the accuracy and efficiency of data analysis. Therefore, to avoid the defects in the parametric method, Fix and Hodges [1] first eliminate the strong assumptions of the parametric method by introducing the idea of discriminant analysis, which is also the fundamental thought source of the nonparametric method. The simplest histogram method is an intuitive embodiment of this idea. Nonparametric methods do not require any prior assumptions and can predict

the real distribution of samples only by a data-driven method. Because of their simplicity and efficiency, nonparametric methods have attracted wide attention. As a classical nonparametric method, the kernel density estimator was first proposed by Rosenblatt [2] and Parzen [3]. It requires only a finite linear combination of a given kernel centered on the samples, while it has optimal convergence rate, optimality and regularization properties and, thus, it is the most widely studied and applied nonparametric format at present. However, since actual samples usually come from the bounded region, the prediction results of the classical kernel density estimator near boundary points are often poor, called boundary problems. In addition, the selection of bandwidth for the kernel density estimator is very important. It will lead to an over-smoothing phenomenon in high (low)-density regions and lack local adaptivity.

In order to eliminate the boundary problem of the classical kernel density estimator, Schuster [4] proposed a data reflection method to obtain a complete dataset through reflection processing of data near the endpoints of bounded intervals, thus, avoiding the boundary problem. However, the data reflection method only applies to the case where the derivative of true density function is zero at the end of the bounded interval, because it actually corrects the inconsistency of classical kernel estimation on bounded intervals. Compared with the limitation of the data reflection method on the boundary derivative of the true density function, the boundary kernel method was first proposed by Gasser and Muller [5] as another more general solution and has been further extended. It can adapt to the density function of any shape. A disadvantage of the boundary kernel method is that the estimated value obtained may be negative. Jones and Foster [6] proposed a non-negative boundary kernel estimator to solve this defect by combining the advantages of the local linear estimator and re-normalized kernel estimator. Another problem with the boundary kernel method increases the variance in the estimator. In order to improve this problem, there are some methods proposed, such as transformation method, pseudo data method, generalized reflection method and so on. In particular, Chen [7,8] attempted to provide a simpler method to eliminate the boundary problem without sacrificing the nonnegativity of density estimation. He further proposed a Beta kernel estimator on the support $[0, 1]$ and a Gamma kernel estimator on the support $[0, \infty]$, while he denoted that they have smaller variances. The proposal of a Beta kernel estimator and Gamma kernel estimator has attracted wide attention. Markovich [9] discussed the good characteristics of the Gamma kernel estimator and extended it to the estimation of multivariate density functions and their partial derivatives. Based on the Gamma estimator, Lin [10] discussed the relationship between temperature and number of potential novel coronavirus infections in China. However, Zhang [11,12] showed that the Beta kernel estimator and Gamma kernel estimator have serious boundary problems and perform worse than the well-known boundary kernel estimation method, when the true density function does not meet the shoulder condition (that is, the first derivative of the density function at the boundary is 0). Cherfaoui [13] et al. further discussed the properties of Gamma kernels in the case that the density function does not satisfy the shoulder condition. Therefore, the boundary problem of kernel density estimation still has a lot of room for improvement. Moreover, the above density estimates method to eliminate the boundary problem generally faces the complexity of kernel function construction or requires some certain applicable conditions, while they are low efficiency in dealing with large samples.

Compared with the kernel density estimator calculated directly, the binned kernel density estimator is beneficial to save a lot of calculation. By pre-grouping samples on isometric grids and applying an appropriate kernel density estimator to the sample data after pre-classification, the calculation is greatly reduced. Moreover, some researchers have shown that a large binned number will not affect the mean integrated squared error of the kernel density estimator [14]. Hall [15] studied the accuracy of density estimation of binned kernel under general binned rules. He provided the minimum mesh size required to obtain a given accuracy level by discussing the influence for accuracy from binned rules and the smoothness of the kernel function. Luo [16] improved the accuracy of the kernel

density estimator method based on the resampling strategy of a multi-point grid. Harel [17] discussed the asymptotic normality of a binned kernel density estimator for non-stationary random variables. Peherstorfer [18] proposed a density estimation based on sparse grid, which can be viewed as improved binned rules. It used a sparse grid instead of full grid to reduce the bins. Although the binned kernel density estimator improves the processing efficiency of large sample data through the binned strategy, it still faces the boundary problem of the kernel density estimator in essence. In addition, there are some other methodologies to apply kernel density estimation to large datasets. Cheng [19] proposed a quick multivariate kernel density estimation for massive datasets by viewing the estimator as a two-step procedure: first, kernel density estimator in sub-interval and then function approximation based on pseudo data via the Nedaraya–Watson estimator. However, the research of Gao [20] demonstrated that the generalized rational form estimators provide a low convergence rate. Moreover, the computation of pseudo data using a kernel density estimator brings more computation than the above binned rule and does not consider the boundary problem of the kernel density estimator. Zheng [21] focused on the choice method of samples from large data to produce a proxy for the true data with a prescribed accuracy, which is more complex than the direct binned rule. Moreover, the research does not pay much attention to the discussion of the kernel density estimator. Therefore, the binned method is very simple and clear. Recently, we proposed a kernel density estimator based on quasi-interpolation and proved its theoretical statistical properties, but the research does not provide a solution for the over-smoothing phenomenon [22].

Another problem (over-smoothing phenomenon) for kernel density estimators is caused by the improper selection of bandwidth, and different scholars have adopted different methods to reduce the occurrence of this phenomenon. The most classical method to choose the bandwidth is the thumb rule, which calculates the optimal bandwidth by the standard deviation and dimension of the samples. Due to the simplicity of this method, it is regarded as a common tool in most application studies of kernel density estimators. However, the actual samples are usually random and uneven, and the optimal bandwidth obtained by the thumb rule is fixed. It only provides a calculation criterion of an optimal bandwidth in a sense and has a very limited improvement effect on the over-smoothing phenomenon. An adaptive bandwidth approach is used to ameliorate this phenomenon viewed as a correction to the thumb rule, which consists of two steps. Firstly, the evaluated function is calculated with a fixed bandwidth and the quantitative relationship between the pointwise function value of samples and the geometric mean value of the samples is established. Then, according to the quantitative relationship, the pointwise correction coefficient is determined to modify bandwidth. The final kernel density estimator can be obtained based on these modified bandwidth. The adaptive bandwidth method improves the accuracy of kernel density estimators for a fixed bandwidth, but it is difficult to apply to large samples because each sample will affect the determination of the correction coefficient and the computational efficiency is low. Barreiro Ures [23] proposed a bandwidth selection method for large samples via using subagging. The subagging can be viewed as an improvement on the cross-validation method. Therefore, it is difficult to capture local changes in samples. Moreover, the research does not consider the boundary problem.

In conclusion, a classical kernel density estimator is a convenient vehicle and it is widely used in many branches of science and technology. However, the majority of research usually did not consider the constraints of the kernel density estimator model itself. These limitations and deficiencies for the kernel density estimator need to be further considered. In addition, previous methods of kernel density estimators are not synthetically considered among the boundary problem, smooth problem and large sample computation efficiency. Therefore, in view of the insufficiency of the classical kernel density estimator, this paper proposes a new modeling process of multivariate adaptive binned kernel density estimators based on the quadtree algorithm and quasi-interpolation, which significantly improves the prediction accuracy of the estimation density function. Research works in this paper are summarized as follows:

(1) Aiming at the boundary problem of the classical kernel density estimator defined over bounded region a new set of asymmetric kernel functions is introduced based on the quasi-interpolation theory to avoid the boundary problem.

(2) To improve the computational efficiency of the classical kernel density estimator for large samples, the idea of binned kernel density estimation is introduced. The coefficient explicit expression of the density estimator under the binned rule of data is derived, which greatly reduces the computation and improves the computational efficiency of the model.

(3) To alleviate the over-smoothing phenomenon of classical kernel density estimators, this paper proposes an adaptive strategy based on the segmentation thought of quadtree algorithm. We set the segmentation thresholds from sample size, bin width and kurtosis to achieve adaptive computation for the amount of bin and bin width. It can effectively avoid the over-smoothing phenomenon in the high (low)-density area and increase local adaptability in the model for samples and further improve the accuracy in the model.

(4) We extend the univariate model based on the quadtree algorithm to the multivariate model. The numerical simulation based on Monte Carlo shows that the constructed models in this paper perform well in the boundary problem, large samples and over-smoothing phenomenon, which are significantly better than the current widespread use of kernel density estimation methods.

## 2. Univariate Adaptive Quasi-Interpolation Density Estimator

### *2.1. Univariate Quasi-Interpolation Density Estimator*

Let $X_1, X_2, \cdots, X_n$ be a set of random samples subject to the probability distribution of an unknown probability density function $f(x)$. The classical non-parametric kernel density estimator is defined as:

$$f_{1,n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where $h$ denotes bandwidth and $K(x)$ denotes kernel function or weight function. There are some common symmetric kernel functions shown in Table 1:

**Table 1.** Common kernel functions.

| Type of Kernel Function | Expression of Kernel Function |
|---|---|
| Gaussian kernel | $K(x) = \frac{1}{\sqrt{2\pi}} exp(-x^2/2)$ |
| Epanechnikov kernel | $K(x) = \frac{3}{4}(1 - x^2),\ \|x\| \le 1$ |
| Triangular kernel | $K(x) = 1 - \|x\|,\ \|x\| \le 1$ |
| Uniform kernel | $K(x) = \frac{1}{2},\ \|x\| \le 1$ |
| Exponential kernel | $K(x) = \frac{1}{2}\lambda exp(-\lambda\|x\|)$ |

According to Equation (1), the classical kernel density estimator requires one to calculate the distance between the predicted point and each sampling point to allot weight function. It means that the computation increases rapidly with the increase in sample size. We can note that the prediction points are mainly influenced by the samples in the limited bandwidth domain, while the samples outside the bandwidth domain have very little influence. The pointwise calculation of large samples outside the bandwidth domain greatly reduces the computational efficiency. Therefore, the binned kernel density estimator was proposed:

$$g_{1,n}(x) = \frac{1}{nh}\sum_{j \in Z} n_j K\left(\frac{x - t_j}{h}\right). \tag{2}$$

where $t_j$ denotes the centers of the $j$-th bin and $n_j$ denotes the number of samples dropped in the $j$-th bin, satisfying $\sum n_j = n$. For clarity, we remind readers: $X_i$ denotes random sample and $t_j$ represents center of bin. According to Equation (2), it can be found that the binned kernel density estimator transforms the pointwise calculation of the classical kernel density estimator into the calculation for bin centers. Its essential idea is to treat the samples in a small region as a whole and the central points of each region as the core samples. Therefore, it can ignore the bandwidth difference between each individual sample and the central point of the region. In this way, unnecessary detailed calculation in the classical kernel density estimator can be reduced and the computational efficiency can be improved on the premise of ensuring accuracy.

However, since actual samples are usually sampled from the bounded domain, the above two classes of the kernel density estimator face the same problem; that is, the boundary problem will occur when a fixed symmetric kernel function is used to predict the true density function defined over a bounded domain. The main reason for the boundary problem is that the weight is allotted outside the density support when smoothing near the boundary point by using a fixed symmetric kernel function. A natural strategy is to use a kernel that has no weight allotted outside the support. Therefore, under the framework of numerical approximation, combining with the theory of quasi-interpolation and the binned idea to improve the above models, this paper proposes a new binned quasi-interpolation density estimator, which can not only improve the computation efficiency of large samples, but also eliminate the boundary problem.

Let us start with some definitions and lemmas:

**Definition 1.** *Let $[a, b]$ be a bounded interval, and $a, b$ be known, $a = t_0 < t_1 < \cdots < t_n = b$ be a set of scattered centers on the interval $[a, b]$, $f(t_j)_{j=0}^n$ be the discrete function values corresponding to scattered centers. Let $c$ be the positive shape parameter, $\phi_j(x) = \sqrt{c^2 + (x - t_j)^2}$ be the MQ function first constructed by Hardy* [24]*, then we have a quasi-interpolation ($L_D$ operator).*

$$L_D f(x) = \sum_{j=0}^{n} f(t_j)\psi_j(x),$$

*where $\{\psi_j\}_{j=0}^n$ are the asymmetric MQ kernels*

$$\begin{aligned}
\psi_0(x) &= 0.5 + \frac{\phi_1(x) - (x - t_0)}{2(t_1 - t_0)}, \\
\psi_1(x) &= \frac{\phi_2(x) - \phi_1(x)}{2(t_2 - t_1)} - \frac{\phi_1(x) - (x - t_0)}{2(t_1 - t_0)}, \\
\psi_j(x) &= \frac{\phi_{j+1}(x) - \phi_j(x)}{2(t_{j+1} - t_j)} - \frac{\phi_j(x) - \phi_{j-1}(x)}{2(t_j - t_{j-1})},\ j = 2, 3, \cdots, n-2, \\
\psi_{n-1}(x) &= \frac{(t_n - x) - \phi_{n-1}(x)}{2(t_n - t_{n-1})} - \frac{\phi_{n-1}(x) - \phi_{n-2}(x)}{2(t_{n-1} - t_{n-2})}, \\
\psi_n(x) &= 0.5 + \frac{\phi_{n-1}(x) - (t_n - x)}{2(t_n - t_{n-1})}.
\end{aligned} \tag{3}$$

Here, these kernels satisfy $0 < \psi_j(x) < 1$ and $\sum_{j=0}^{n} \psi_j(x) = 1$.

In addition, we obtain the following error estimates, which can be found in Wu and Schaback [25].

**Lemma 1.** *For any function $f \in C^2[a,b]$, let $h = \max\{t_{j+1} - t_j\}_{j=0}^{n-1}$, $c$ be a positive shape parameter, there exists some constant $K_1$, $K_2$, $K_3$ independent of $h$ and $c$, such that the following inequality*

$$||f - L_D f||_\infty \leq K_1 h^2 + K_2 ch + K_3 c^2 |\log h|$$

*holds.*

According to lemma 1, for any shape parameter $c$ satisfying $0 \leq c \leq O\left(h/\sqrt{|\log h|}\right)$, the convergence rate $O(h^2)$ for the whole bounded interval can be provided by quasi-interpolation $L_D$. Furthermore, the research of Ling [26] shows that the multivariate $L_D$ operator by the tensor product technique(dimension-splitting) can provide the same convergence rate as the univariate case. Inspired by the convergence characteristics of quasi-interpolation and the idea of the binned kernel density estimator, we construct a univariate adaptive quasi-interpolation density estimator based on the quadtree algorithm, which consists of three steps. Suppose that $X$ is a random variable, $\{X_k\}_{k=1}^n$ are the $n$ independent samples in the random variable $X$. There is an unknown density function $f(x)$ on the bounded interval. The first step is to divide the interval $[a,b]$ into $N$ bins $\left\{[t_j, t_{j+1})\right\}_{j=0}^{N-1}$. Let $n_j$ denote the number of samples $\{X_k\}_{k=1}^n$ dropping into the corresponding bin $[t_j, t_{j+1})$. In the second step, we construct a new univariate binned density estimator as follows:

$$Q_{1,D} f(x) = \sum_{j=0}^{N} \alpha_j(f) \psi_j(x), \ x \in [a,b]. \tag{4}$$

Here, $\{\psi_j\}_{j=0}^N$ denote the asymmetric MQ kernels defined by Equation (3), and the coefficients $\{\alpha_j(f)\}_{j=0}^N$ are defined as

$$\begin{aligned}
\alpha_0(f) &= \frac{t_2 + t_1 - 2t_0}{(t_2 - t_0)(t_1 - t_0)} \frac{n_0}{n} + \frac{t_0 - t_1}{(t_2 - t_0)(t_2 - t_1)} \frac{n_1}{n}, \\
\alpha_j(f) &= \frac{t_j - t_{j-1}}{(t_{j+1} - t_j)(t_{j+1} - t_{j-1})} \frac{n_j}{n} + \frac{t_{j+1} - t_j}{(t_j - t_{j-1})(t_{j+1} - t_{j-1})} \frac{n_{j-1}}{n}, \ j = 1, 2, \cdots, N-1, \\
\alpha_N(f) &= \frac{2t_N - t_{N-1} - t_{N-2}}{(t_N - t_{N-1})(t_N - t_{N-2})} \frac{n_{N-1}}{n} + \frac{t_{N-1} - t_N}{(t_N - t_{N-2})(t_{N-1} - t_{N-2})} \frac{n_{N-2}}{n}.
\end{aligned} \tag{5}$$

According to Equation (4) and lemma 1, we can note that the introduction of asymmetric MQ kernels can avoid the boundary problem caused by the weight allotted outside the support when the traditional kernel function smooths near the boundary points. Moreover, Equation (5) shows that $n_j/n$ represents the frequency of samples falling into the corresponding bin $[t_j, t_{j+1})$. Through the linear combination of frequencies between adjacent bins, the explicit expression of coefficients of the estimator under the binned rule is given, which can effectively improve the calculation efficiency in the model. Thirdly, the over-smoothing phenomenon in the kernel density estimator is considered. In the above two steps, we built a univariate binned quasi-interpolation density estimator. Based on the known samples and interval, the interval was divided into a certain number of bins, and then the estimated density function could be calculated by the endpoint position of the bin and the number of samples in the bins. If the number of bins is too few, the predicted result is over-smoothing, which differs greatly from the actual scenario. If the number of bins is too great, the calculation efficiency will be greatly reduced. How to determine the number and width of bins is the key to both model accuracy and calculation efficiency. The most common method is the thumb rule, which takes the idea of a fixed bandwidth and calculates the bandwidth as follows:

$$h = \left(\frac{4}{d+2}\right)^{1/(d+4)} \sigma n^{-1/(d+4)}. \tag{6}$$

Here, $d$ denotes the dimension and σ denotes standard deviation of samples. In particular, to maintain notational clarity, we remind readers: $d$ denotes dimension and $D$ is a mark of $L_D$ operator. The number of bins is calculated by ceiling $(b-a)/h$. This method uses equal bandwidth, and similar equal bandwidth methods include the unbiased cross-validation method and insertion method, etc. However, due to the strong randomness and uneven distribution of actual samples, the equal bandwidth method generally has the problem of insufficient description of details for the high-density area, which causes the over-smoothing phenomenon. Therefore, it is expected that the bandwidth can be adjusted adaptively with the density of samples. The bandwidth should be smaller in high-density areas to enhance local characterization and improve accuracy. In addition, the bandwidth should be larger in the gentle area to avoid excessive calculation and improve calculation efficiency. A common adaptive method determines the number of bins according to the thumb rule and obtains the estimated value of bin centers. Then, the ratio between each estimated value and the geometric mean of each estimated value is taken as the correction coefficient of bandwidth, so as to achieve the purpose of taking smaller bandwidth in the intensive area and larger bandwidth in the sparse area. This adaptive method is simple and easy to operate, but it also has three disadvantages: First, this method is based on the estimation of thumb rule, and the adaptive process does not change the number of bins, which can be regarded as the optimal configuration of bandwidth in essence. Second, the degree of adaptive refinement is insufficient and the determination of bandwidth correction coefficient is too rough, which is susceptible to extreme values. Moreover, it is difficult to distinguish sharp peaks from wide peaks. Third, the adaptive effect of multi-peak distribution is poor. In addition, the density near the boundary is usually small, and increasing the width of the bin easily aggravates the boundary problem. Therefore, this paper proposes a new adaptive binned method.

### *2.2. Adaptive Binned Method Based on Quadtree Algorithm*

The quadtree algorithm, as a space partition index technology, is widely used in the image processing field [27]. The key idea is an iterated segmentation of data space. The number of iterations depends on the number of samples in bins and bin-width threshold. Therefore, the density of samples can be characterized by the number and width of bins. The area with dense samples has more iterated segmentation and the area with sparse samples has less iterated segmentation. Therefore, according to the idea of quadtree segmentation, we can adaptively adjust the bin number and bin width in the quasi-interpolation density estimator via a data-driven method. The high-density area is divided into more bins to obtain a smaller bin width, which can more keenly capture the distribution details of the area, while the gentle area is divided into fewer bins to save the cost of calculation, so as to achieve a reasonable distribution of bins and improve the accuracy in the model. The adaptive binned method based on the quadtree algorithm is shown in Figure 1:

First of all, the sample space is divided into four bins and the number of samples $\{n_i\}_{i=1}^4$ in each bin and the bin widths $\{L_i\}_{i=1}^4$ are recorded. Secondly, we set the threshold of sample number $n_{max}$ and bin width $L_{max}$. The setting of sample number threshold $n_{max}$ captures distribution details in the high-density area with more bins and improves computing efficiency in the gentle area with less bins. It not only solves the over-smoothing problem but also takes into account computing efficiency. The setting of the bin-width threshold ensures the segmentation level of the whole domain and avoids an insufficient number of bins, which leads to the large estimation error or boundary problem. Following the thumb rule, we set the bin-width threshold to $1.06\sigma n^{-1/5}$. In addition, we set a kurtosis threshold to identify the peak distribution of samples and improve the accuracy. Finally, the number of samples and bin width in each bin are compared with the number of sample number threshold $n_{max}$, bin-width threshold $L_{max}$ and kurtosis threshold. The segmentation is finished when all of these conditions are met.

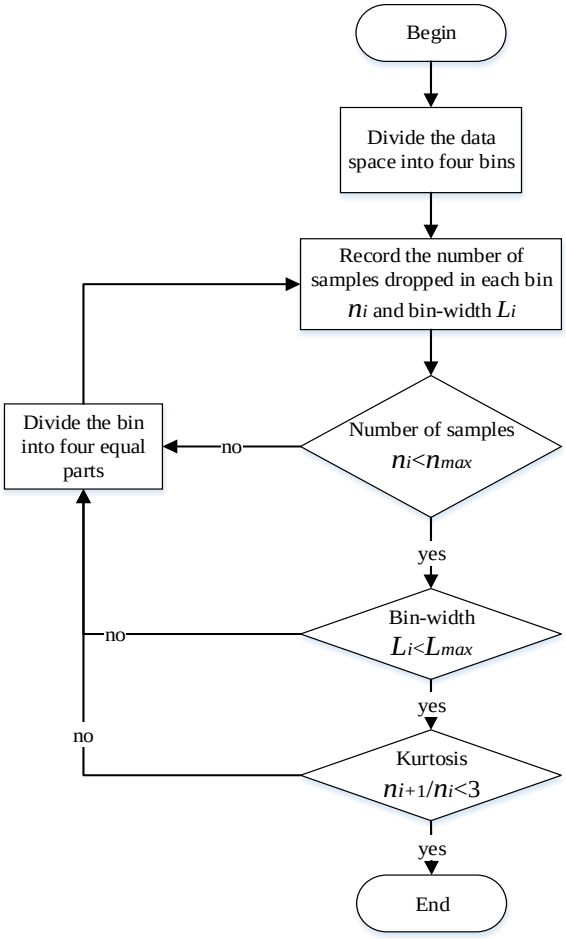


**Figure 1.** Univariate adaptive binned method based on quadtree algorithm.

## 3. Multivariate Adaptive Binned Quasi-Interpolation Density Estimator

Based on the idea of the above univariate adaptive binned quasi-interpolation density estimator, we extend it to the multivariate model. Following the above process, we first construct the multivariate binned density estimator. The classical multivariate kernel density estimator and multivariate binned density estimator are extended from the univariate model via the tensor product technique. They are defined as follows:

$$f_{d,n}(\mathbf{x}) = \frac{1}{n\mathbf{H}} \sum_{i=1}^{n} K^d(\mathbf{x} - \mathbf{X_i}),$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_d)$, $\mathbf{H} = \prod_{j=1}^{d} h_j$, $K^d(\mathbf{x} - \mathbf{X_i}) = K\big((x_1 - X_{i_1})/h_1\big)K\big((x_2 - X_{i_2})/h_2\big)$ $\cdots K\big((x_d - X_{i_d})/h_d\big)$.

$$g_{d,n}(\mathbf{x}) = \frac{1}{n\mathbf{H}} \sum_{j_1 \in Z} \sum_{j_2 \in Z} \cdots \sum_{j_d \in Z} n_{j_1, j_2, \cdots, j_d} K\left(\frac{x_1 - t_{j_1}}{h_1}\right) \cdots K\left(\frac{x_d - t_{j_d}}{h_d}\right), \tag{7}$$

where $\sum \cdots \sum n_{j_1, j_2, \cdots, j_d} = n$. Based on the above univariate binned quasi-interpolation density estimator, we also extended it to the multivariate binned quasi-interpolation density estimator via tensor product technique:

Let $X$ be a $d$-dimension random variable with an unknown density function $f$ defined on a bounded hyperrectangle $\mathrm{I}^d = \prod_{i=1}^{d} [a_i, b_i]$, $a_i < b_i \in R$. Let $\{X_k\}_{k=1}^{n}$ be the $n$ independent

samples of random variable $X$. Inspired by the idea of histogram, $\mathrm{I}^d$ is divided into **N** sub-intervals $\left\{\prod_{i=1}^{d}\left[t_{i,\,j_i}, t_{i,\,j_i+1}\right)\right\}$, $a_i = t_{i,0} < t_{i,1} < \cdots < t_{i,N_i} = b_i$, $1 \le i \le d$. Here, $\{N_i\}_{i=1}^{d}$ are positive integers, $0 \le j_i \le N_i - 1$, $1 \le i \le d$. Then, the frequency of $\{X_k\}_{k=1}^{n}$ dropping into each bin $\prod_{i=1}^{d}\left[t_{i,\,j_i}, t_{i,\,j_i+1}\right)$ can be calculated by $n_{j_1,\,j_2,\,\cdots,\,j_d}/n$. A multivariate (bivariate) quasi-interpolation density estimator via the tensor product technique is as follows:

$$Q_{2,D}f(\mathbf{x}) = \sum_{j_1=0}^{N_1}\sum_{j_2=0}^{N_2}\alpha_{j_1,\,j_2}(f)\psi_{j_2}(x_2)\psi_{j_1}(x_1), \tag{8}$$

For $1 \le j_i \le N_i - 1$, $i = 1,\ 2,\ \cdots,\ d$, the coefficient $\alpha_{j_1,\,j_2}(f)$ is

$$\begin{aligned}\alpha_{j_1,\,j_2}(f) = {}& \beta_{j_1,\,j_2}\frac{n_{j_1-1,\,j_2-1}}{n\left(t_{1,\,j_1} - t_{1,\,j_1-1}\right)\left(t_{2,\,j_2} - t_{2,\,j_2-1}\right)}\\ &+\beta_{j_1+1,\,j_2}\frac{n_{j_1-1,\,j_2-1} + n_{j_1,\,j_2-1}}{n\left(t_{1,\,j_1+1} - t_{1,\,j_1-1}\right)\left(t_{2,\,j_2} - t_{2,\,j_2-1}\right)}\\ &+\beta_{j_1,\,j_2+1}\frac{n_{j_1-1,\,j_2-1} + n_{j_1-1,\,j_2}}{n\left(t_{2,\,j_2+1} - t_{2,\,j_2-1}\right)\left(t_{1,\,j_1} - t_{1,\,j_1-1}\right)},\end{aligned}$$

where

$$\begin{aligned}\beta_{j_1,\,j_2} &= \frac{\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right)\left(t_{1,\,j_1+1} + t_{1,\,j_1-1} - 2t_{1,\,j_1}\right) - \left(t_{1,\,j_1+1} - t_{1,\,j_1}\right)\left(t_{2,\,j_2} - t_{2,\,j_2-1}\right)}{\left(t_{1,\,j_1+1} - t_{1,\,j_1}\right)\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right)},\\ \beta_{j_1+1,\,j_2} &= \frac{t_{1,\,j_1} - t_{1,\,j_1-1}}{t_{1,\,j_1+1} - t_{1,\,j_1}},\\ \beta_{j_1,\,j_2+1} &= \frac{t_{2,\,j_2} - t_{2,\,j_2-1}}{t_{2,\,j_2+1} - t_{2,\,j_2}}.\end{aligned}$$

For $j_i \in \{0, N_i\}$, $i = 1,\ 2,\ \cdots,\ d$, the coefficient $\alpha_{j_1,\,j_2}(f)$ is

$$\begin{aligned}\alpha_{j_1,\,j_2}(f) = {}& \gamma_{j_1,\,j_2}\frac{n_{j_1,\,j_2}}{n\left(t_{1,\,j_1+1} - t_{1,\,j_1}\right)\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right)}\\ &+\gamma_{j_1+1,\,j_2}\frac{n_{j_1,\,j_2} + n_{j_1+1,\,j_2}}{n\left(t_{1,\,j_1+2} - t_{1,\,j_1}\right)\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right)}\\ &+\gamma_{j_1,\,j_2+1}\frac{n_{j_1,\,j_2} + n_{j_1,\,j_2+1}}{n\left(t_{1,\,j_1+1} - t_{1,\,j_1}\right)\left(t_{2,\,j_2+2} - t_{2,\,j_2}\right)},\end{aligned} \tag{9}$$

where

$$\begin{aligned}\gamma_{j_1,\,j_2} &= \frac{t_{1,\,j_1+1}\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right) - t_{1,\,j_1+1}\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right) + t_{1,\,j_1}\left(t_{2,\,j_2+1} - t_{2,\,j_2+2}\right)}{\left(t_{1,\,j_1+1} - t_{1,\,j_1+2}\right)\left(t_{2,\,j_2+1} - t_{2,\,j_2}\right)},\\ \gamma_{j_1+1,\,j_2} &= \frac{t_{1,\,j_1+1} - t_{1,\,j_1}}{t_{1,\,j_1+1} - t_{1,\,j_1+2}},\\ \gamma_{j_1,\,j_2+1} &= \frac{t_{2,\,j_2+1} - t_{2,\,j_2}}{t_{2,\,j_2+1} - t_{2,\,j_2+2}}.\end{aligned}$$

In Equation (9), for $i = 1,\ 2,\ \cdots,\ d$, there are $N_i + 1 := N_i - 1$ and $N_i + 2 := N_i - 2$. To avoid the over-smoothing phenomenon, we use the advantage of the tensor product to transform the multivariate adaptive binned problem into a univariate problem, and the adaptive process is shown in Figure 2.

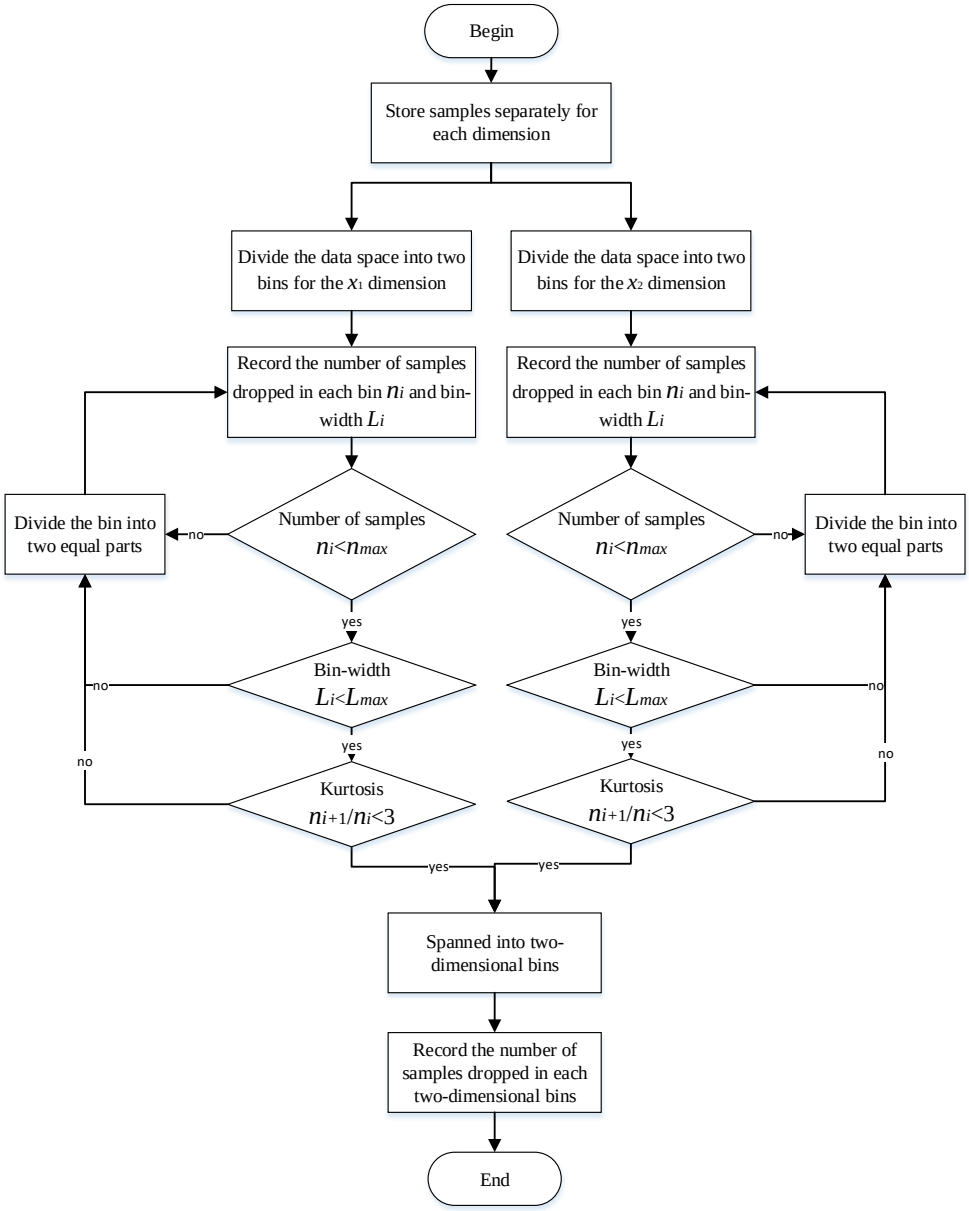


**Figure 2.** Bivariate adaptive binned method based on quadtree algorithm.

First, we divide the domain into two bins for each dimension and record the number of samples and the bin width in each bin from the univariate dimension. Secondly, they are compared with the threshold of sample number, bin width and kurtosis to achieve iterative segmentation. Finally, these bins in each dimension are spanned into some two-dimensional bins via the tensor product technique, and the number of samples falling in each two-dimensional bin is recorded.

## 4. Numerical Simulation

In order to verify the performance of the model proposed in this paper, the Monte Carlo method is used for numerical simulation in this section. Maximal Mean Squared Error (MMSE)

$$\mathrm{MMSE}(f) = \sup_{x\in[a,b]} \mathbb{E}\Big(Q_{(*)}f(x) - f(x)\Big)^2$$

and Mean Integrated Squared Error (MISE)

$$\mathrm{MISE}(f) = \mathbb{E}\sum\Big(Q_{(*)}f(x) - f(x)\Big)^2$$

are used to quantify the difference between the estimated density function and the true density function. Here, $\mathbb{E}$ denotes the expectation value, $Q_{(*)}f(x)$ denotes estimated density function and $f(x)$ denotes true density function. MMSE and MISE error are used to measure the local and overall accuracy in the model, respectively.

### *4.1. Univariate Test*

As the first example, we test the prediction accuracy of the univariate model by using the following test function:

$$f(x) = \frac{1}{2}\mathbb{N}\left(\frac{1}{2}, \frac{1}{6}\right) + \sum_{l=-2}^{2} \frac{2^{1-l}}{31}\mathbb{N}\left(\frac{2(l+3)+1}{12}, \frac{2^{-l}}{60}\right),\ x \in [0,1].$$

Here, $\mathbb{N}(\mu, \sigma)$ denotes a normal distribution with an expectation $\mu$ and variance $\sigma$. The test function (called asymmetric claw distribution) is a combination of the five different parameters' normal distribution, which has five peaks and troughs of different heights on the considered interval [0, 1]. Next, the comparison of the quasi-interpolation density estimator (QIDE), univariate adaptive binned quasi-interpolation density estimator (AQIDE) based on the quadtree algorithm, classical kernel density estimator (KDE) and binned kernel density estimator (BKDE) are shown in Figure 3.

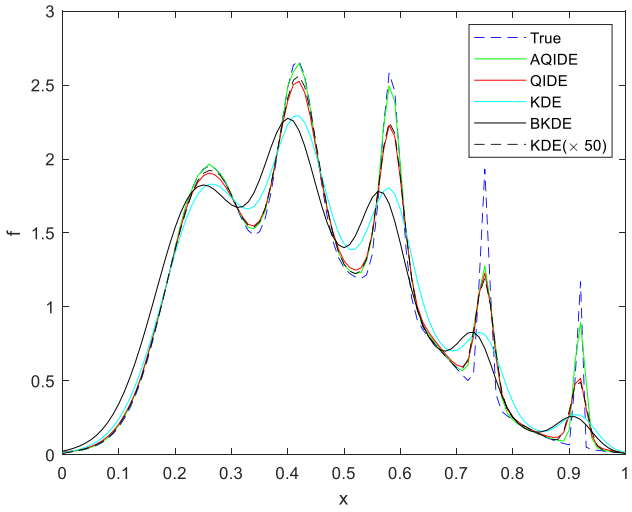


**Figure 3.** Sketches of asymmetrical claw distribution.

Figure 3 shows the sketches of different density estimators when the sample number is $n = 12,400$ and the number of simulation experiments is 100. Furthermore, we provide a comparison sketch of KDE under the larger sample number, $n = 12,400 \times 50 = 620,000$. In these simulation experiments, the bandwidth selection for the KDE model and bin-width selection of the BKDE model both adopt the thumb rule of Equation (6). The bin number and bin width in the AQIDE model proposed in this paper are adaptively obtained by the univariate quadtree segmentation algorithm designed in Section 3. The shape parameter is selected as $c = \min(L_i)$. The threshold of bin width $L_{max}$ is determined by using the thumb rule $L_{max} = 1.06\sigma n^{-0.2}$ from Equation (6). The threshold of bin number $n_{max}$ is determined by $n_{max} = nL_{max}$ based on the thumb rule, and the kurtosis threshold is determined as 3. In addition, to compare the performance of the quasi-interpolation model after an adaptive

processing proposed in this paper, the same bin number and shape parameters are selected for the QIDE model and AQIDE model.

In Figure 3, the blue dashed line denotes the true density function, while the turquoise, black, red and green lines represent the results by the KDE, BKDE, QIDE and AQIDE models, respectively. The black dashed line denotes the result of KDE for larger samples. We can note that the ability of classical KDE to catch the last two high peaks is poor. It performs nearly as well as our QIDE only when the sample number is increased to 620,000. The MMSE error and MISE error corresponding to each model are shown in Table 2. According to Figure 3 and Table 2, the binned technique does not affect the fitting accuracy. Moreover, the KDE and BKDE models both have a serious over-smoothing phenomenon, and the prediction effect of peaks and troughs is poor. The QIDE and AQIDE models in this paper can alleviate the problem. The fitting effect of peaks and troughs performs significantly better than the KDE and BKDE models. In addition, according to the adaptive algorithm proposed in this paper, we calculate the bin number, and then we provide the results of the equidistant QIDE and AQIDE model under the same bin number. These results show that the AQIDE model performs better than the QIDE model when the bin number is the same. It means that the proposed adaptive method based on the quadtree algorithm can better capture the distribution details than the case of equidistance bin width and improve the fitting accuracy of the model by increasing or reducing adaptive bins in the high-density or gentle area.

**Table 2.** Accuracy of univariate model.

| Model | MMSE | MISE |
|---|---|---|
| KDE | 1.1197 | 0.0559 |
| BKDE | 1.1333 | 0.0587 |
| QIDE | 0.7065 | 0.0166 |
| AQIDE | 0.6601 | 0.0105 |

### *4.2. Bivariate Test*

In order to further test the performance of the multivariate model proposed in this paper, we choose the following modified bivariate density function as the test function:

$$f(x_1, x_2) = G \cdot \left( \frac{3}{4} e^{-((9x_1-2)^2+(9x_2-2)^2)/4} + \frac{3}{4} e^{-((9x_1+1)^2/49-(9x_2+1)/10)} + \frac{1}{2} e^{-((9x_1-7)^2+(9x_2-3)^2)/4} - \frac{1}{5} e^{-((9x_1-4)^2+(9x_2-7)^2)} \right), \; x_1, x_2 \in [0,1].$$

The function originates from the classic Franke function, which is difficult to approximate due to two Gaussian peaks of different heights and a small dip. Therefore, it is widely used as a test function in numerical analysis. In the test function, a constant $G$ is introduced to ensure that the final test function $f$ is the density function defined over the domain $[0,1]^2$. A comparison of the adaptive multivariate binned quasi-interpolation density estimator (AMQIDE), multivariate binned quasi-interpolation density estimator (MQIDE), classical multivariate kernel density estimator (MKDE) and multivariate binned kernel density estimator (MBKDE) is shown in Figure 4.

Figure 4 shows the sketches of different multivariate density estimators under the samples $N = 300,000$ and the number of simulation experiments is 50. In these simulation experiments, the bandwidth of the MKDE model and the bin width of the MBKDE model both adopt the thumb rule from Equation (6). The bin number and bin width in the AQIDE model are calculated by the multivariate adaptive quadtree algorithm. The shape parameter is chosen as $c = h$ and the threshold of the bin width $L_{max}$ is given by the thumb rule $L_{max} = \sigma n^{-1/6}$ from Equation (6). The threshold of the sample number $n_{max}$ is determined by $n_{max} = nL_{max}$ based on the thumb rule, and the kurtosis threshold is determined as 3.

In addition, the QIDE model chooses the same bin number and shape parameter as the AQIDE model.

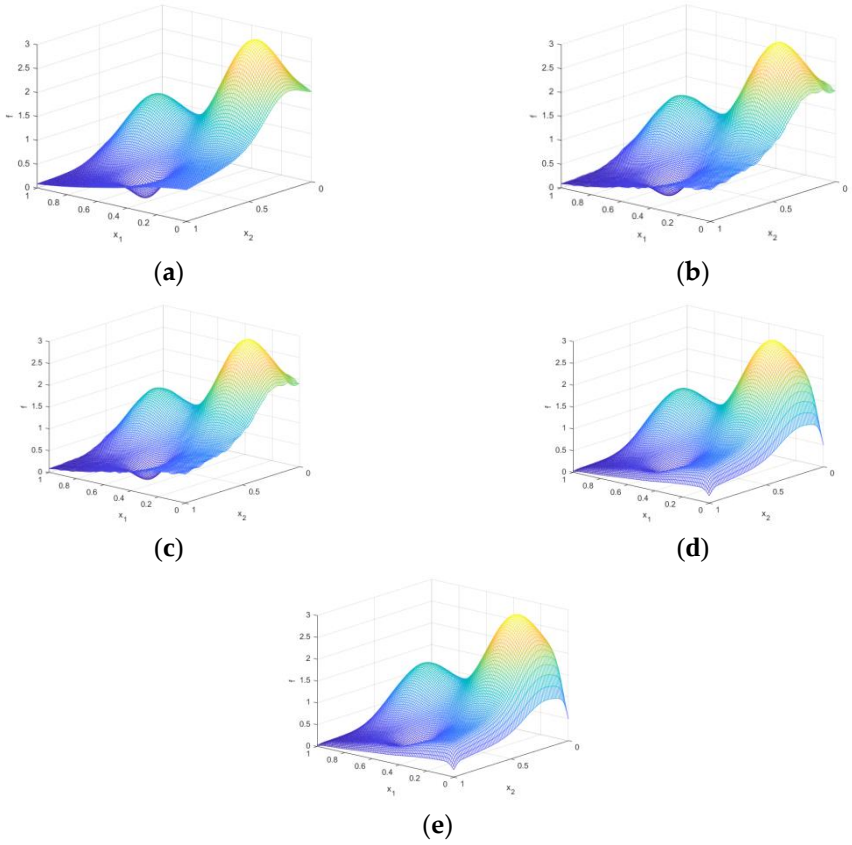

**Figure 4.** Sketches of Franke function. (**a**) True density function. (**b**) AMQIDE. (**c**) MQIDE. (**d**) MKDE. (**e**) MBKDE.

The Figure 4a is the real density function. Figure 4b,c are the estimated density functions obtained by the AMQIDE model and MQIDE model, while Figure 4d,e are the estimated density functions obtained by the MKDE model and MBKDE model. In addition, corresponding MMSE errors and MISE errors in the four models are provided in Table 3. From Figure 4 and Table 3, it can be noted that the kurtosis in the Franke density function is small, and the estimated results of the MQIDE model and AMQIDE model are consistent, meaning that our adaptive method can effectively identify high-density areas. The results of the MKDE model and MBKDE model are similar to the univariate situation, which perform poorly with a serious boundary problem. The performance is much lower than the MQIDE and AMQIDE models proposed in this paper.

**Table 3.** Accuracy of bivariate model.

| Model | MMSE | MISE |
|---|---|---|
| MKDE | 1.3906 | 0.0211 |
| MBKDE | 1.3891 | 0.0204 |
| MQIDE | 0.0680 | $2.3017 \times 10^{-4}$ |
| AMQIDE | 0.0680 | $2.3017 \times 10^{-4}$ |

## 5. Conclusions

This paper proposes a multivariate adaptive quasi-interpolation density estimation model based on the quadtree algorithm. The key goal to achieve the adaptive segmentation for samples via the quadtree algorithm and obtain the proper binned number and

bin width. The method can be adjusted adaptively according to the distribution of the samples. It not only identifies details of distribution in the high-density area, but also avoids the inefficiency of large bins, which can effectively avoid the over-smoothing phenomenon. Moreover, based on the good properties of quasi-interpolation, the theory of quasi-interpolation is introduced to construct the kernel function for the density estimator, which can avoid the boundary problem of the classical kernel density estimator. Finally, the idea of frequency approximation probability is used to construct the coefficient of the binned density estimator, which can handle large samples and improve computational efficiency. The simulation of Monte Carlo shows that the proposed nonparametric model has strong robustness and can estimate the density function with high performance.

**Author Contributions:** Conceptualization, J.W.; methodology, J.W.; software, J.W. and Y.L.; validation, Y.L. and J.C.; investigation, J.W.; data curation, J.W.; writing—original draft preparation, J.W.; writing—review and editing, Y.L. and J.C.; visualization, J.W.; supervision, J.C. 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 (No. 11871074).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Fix, E.; Hodges, J.L. Discriminatory analysis, non-parametric discrimination: Consistency properties. *Int. Stat. Rev.* **1951**, *57*, 238–247. [CrossRef]
2. Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **1956**, *27*, 832–837. [CrossRef]
3. Parzen, E. On estimation of probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
4. Schuster, E.F. Incorporating support constraints into nonparametric estimators of densities. *Commun. Stat. Theory Methods* **1985**, *14*, 1123–1136. [CrossRef]
5. Gasser, T.; Müller, H.G. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*; Springer: Berlin/Heidelberg, Germany, 1979; Volume 1979, pp. 23–68.
6. Jones, J.C.; Foster, P.J. A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sin.* **1996**, *6*, 1005–1013.
7. Chen, S.X. Beta kernel estimators for density functions. *Comput. Stat. Data Anal.* **1999**, *31*, 131–145. [CrossRef]
8. Chen, S.X. Probability density function estimation using gamma kernels. *Ann. Inst. Stat. Math.* **2000**, *52*, 471–480. [CrossRef]
9. Markovich, L.A. Nonparametric estimation of multivariate density and its derivative by dependent data using Gamma kernels. *J. Math. Sci.* **2021**, *254*, 550–573. [CrossRef]
10. Lin, W.; He, Q. The influence of potential infection on the relationship between temperature and confirmed cases of COVID-19 in China. *Sustainability* **2021**, *13*, 8504. [CrossRef]
11. Zhang, S.; Karunamuni, R.J. Boundary performance of the beta kernel estimators. *Nonparametr. Stat.* **2010**, *22*, 81–104. [CrossRef]
12. Zhang, S. A note on the performance of the gamma kernel estimators at the boundary. *Stat. Probab. Lett.* **2010**, *80*, 548–557. [CrossRef]
13. Cherfaoui, M.; Boualem, M.; Aïssani, D. Influence of the density pole on the performances of its gamma-kernel estimator. *Afr. Stat.* **2017**, *12*, 1235–1251. [CrossRef]
14. Scott, D.W.; Sheather, S.J. Kernel density estimation with binned data. *Commun. Stat. Theory Methods* **1985**, *14*, 1353–1359. [CrossRef]
15. Hall, P.; Wand, M.P. On the accuracy of binned kernel density estimators. *J. Multivar. Anal.* **1996**, *56*, 165–184. [CrossRef]
16. Luo, J. Improving the accuracy of binned kernel density estimators. *J. Comput. Inf. Syst.* **2014**, *10*, 7477–7488.
17. Harel, M.; Lenain, J.F.; Ngatchou-Wandji, J. Asymptotic normality of binned kernel density estimators for non-stationary dependent random variables. In *Mathematical Statistics and Limit Theorems*; Springer International Publishing: Cham, Switzerland, 2015.
18. Peherstorfer, B.; Pfluger, D.; Bungartz, H.J. Density estimation with adaptive sparse grids for large data sets. In Proceedings of the 2014 SIAM International Conference on Data Mining (SDM), Philadelphia, PA, USA, 24–26 April 2014; Volume 2017, pp. 443–451.
19. Cheng, K.F.; Chu, C.K.; Lin, D.K.J. Quick multivariate kernel density estimation for massive datasets. *Appl. Stoch. Models Bus. Ind.* **2006**, *22*, 533–546. [CrossRef]
20. Gao, W.W.; Fasshauer, G.E.; Sun, X.P.; Zhou, X. Optimality and regularization properties of quasi-interpolation: Both deterministic and stochastic perspectives. *SIAM J. Numer. Anal.* **2020**, *58*, 2059–2078. [CrossRef]

21. Zheng, Y.; Jestes, J.; Philips, J.M.; Li, F. Quality and efficiency in kernel density esimates for large data. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 22–27 June 2013; pp. 1–12.
22. Gao, W.W.; Wang, J.C.; Zhang, R. Quasi-interpolation for multivariate density estimation on bounded domain. *MATCOM* **2022**. *submitted*.
23. Barreiro Ures, D. Nonparametric Density and Regression Estimation for Samples of Very Large Size. Ph.D. Thesis, Universidade da Coruna, Corunha, Spain, 2021.
24. Hardy, R. Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **1971**, *76*, 1905–1915. [CrossRef]
25. Wu, Z.M.; Schaback, R. Shape preserving properties and convergence of univariate multiquadric quasi-interpolation. *Acta Math. Appl. Sin.* **1994**, *10*, 441–446. [CrossRef]
26. Ling, L. Multivariate quasi-interpolation schemes for dimension-splitting multiquadric. *Appl. Math. Comput.* **2005**, *161*, 195–209. [CrossRef]
27. Arroyuelo, D.; Navarro, G.; Reutter, J.L. Optimal joins using compressed quadtrees. *ACM Trans. Database Syst.* **2022**, *47*, 1–53. [CrossRef]