Probability density function analysis based on logistic regression model

Lingling Fang*, Yunxia Zhang,
Department of Science Education, Jiangxi University of Technology,
Nanchang, 330098
China

Abstract—The data fitting level in probability density function analysis has great influence on the analysis results, so it is of great significance to improve the data fitting level. Therefore, a probability density function analysis method based on logistic regression model is proposed. The logistic regression model with kernel function is established, and the optimal window width and mean square integral error are selected to limit the solution accuracy of probability density function. Using the real probability density function, the probability density function with the smallest error is obtained. The estimated probability density function is analyzed from two aspects of consistency and convergence speed. The experimental results show that compared with the traditional probability density function analysis method, the probability density function analysis method based on logistics regression model has a higher fitting level, which is more suitable for practical research projects.

Keywords—Consistency, kernel function, logistic regression analysis, probability density function.

I. Introduction

Probability density function estimation, which means estimating the probability density function of continuous random variables from the known observation values, is the basis to solve many machine learning problems and pattern recognition problems. Many literatures have elaborated the application of the probability density function estimation in neural network, data analysis and statistical learning theory [1]. At the same time, probability density function estimation plays an important role in many fields such as archaeology, banking, climatology, economics, genetics, hydrography and physiology. The probability density function analysis is beneficial for the estimation of probability density function. The commonly used methods of probability density function analysis include parametric probability density function analysis method, nonparametric probability density function analysis method and semi parametric probability density function analysis method [2]. Parametric estimation is conducted under the condition that probability density function model of the observation sample is known but parameters unknown. It is necessary to estimate the unknown parameters from the known samples, so as to determine the specific form of the probability density function [3]. Nonparametric analysis refers to the analysis of the unknown probability density function by certain strategy on the premise that the model information of the observed sample probability density function is unknown. Semi parametric analysis is a method which combines nonparametric estimation and parametric estimation. It uses the weighted average of the parametric estimation and nonparametric estimation to analyze the real probability density function [1]-[3].

In statistics, the commonly used means to define or describe a statistical model in a distributed way are mostly characterized by "small at both ends, large in the middle, left and right symmetries" [4]. However, the probability density function image can generally show the characteristics of the distribution function, and the method of probability density function estimation is the most suitable and convenient method in the practical application. Firstly, take samples from the total sample capacity to estimate the probability density function. If the mathematical form of probability density function is known, then parameter estimation method is used; if the mathematical form of probability density function is unknown, then nonparametric estimation method is used [4]. However, in many practical application problems, people do not know any information of probability density function, so probability density function is generally required to have a mathematical form, so the probability density function analysis is nonparametric.

With the development of data mining technology, database has been widely used by the public. At the same time, people have paid more and more attention to the problem of probability density estimation [5]. For example, it can be used as an analysis tool to query approximate results for large databases, and can also be applied to various data classification and pattern recognition. In addition, the main research direction of

econometrics is probability density function. As far as the current research forms are concerned, it is difficult to judge whether the parameters between economic variables are non-linear. Therefore, in practical application, econometric models are usually not consistent, which cannot meet the needs of management and economic research. The best way to solve this problem is to use kernel function to estimate probability density function, and apply it to actual projects [6]. The main reason why this method is widely used and deeply studied is that it can give the display form of density function and facilitate the analysis and research of probability density function.

Many scholars have put forward research methods for the analysis of probability density function. Some scholars constructed the maximum absolute value-state quantity joint vector process by introducing extended state vector, thus the maximum process without Markov property was transformed into vector random process with Markov property [7]. On this basis, the transition probability density function of the joint vector process was established based on the relationship between the maximum absolute value and state quantity. Furthermore, combining Chapman-Kolmogorov equation and path integration method, this paper put forward a numerical method for solving the probability density function with maximum absolute value. Some scholars put forward that for one-dimensional continuous random variables [8], when the discontinuous derivative points of the distribution function are finite in number, the requirements can be met by adding a proper definition to the probability density. For two-dimensional continuous random variables, when the second-order mixed partial derivatives of the distribution function are discontinuous on a finite smooth curve, the requirements can be met only by adding a proper definition to the probability density. However, the research methods mentioned above are limited by the inability in obtaining high precision results.

In this method, it is important to select the kernel function and determine the width of the window. The fitting degree of the actual data is mainly affected by the kernel function. Therefore, the selection of the kernel function is crucial in the analysis of probability density function. The determination of the window width can affect the fitting degree of the actual data, and window width is a variable that needs to be emphatically assumed on the basis of theory. Therefore, the window width determination is important. For traditional probability density function analysis method, which is affected by the window width, the fitting degree of actual data is poor and the data distribution is sparse. Therefore, logistic regression analysis is cited to achieve the purpose of analyzing probability density function. The predicted results obtained by logistic regression analysis are very close to the actual results. The kernel function is obtained by logistic regression model. The probability density function is estimated by the kernel function, and the analysis of probability density function is completed from the two aspects of consistency and convergence speed. Through verification, it can be found that the designed method has compact data distribution, high fitting level and good application performance, which provides certain reference for the analysis of probability density function.

II. Probability Density Function Analysis Based On Logistic Regression Model

A. Establishment of Logistic Regression Model with Kernel Function

The logistic regression model with kernel function is deduced by using logistic regression model [7]. Assuming that the response variable $Q_1, Q_2, \ldots, Q_n$ is independent and $Q_n \cdot Bernoull\,\mathrm{n}(f)$ and $Bernoull$ are family of exponential distributions [8]. Supposing f satisfies the following equation:

$$\log\left(\frac{fn}{1-fn}\right) = a + bx \tag{1}$$

where, $a+bx$ represents a one-dimensional stochastic equation and x represents a random variable. The relationship between fn and random variable x is established by the above equation. The left side of (1) is the logarithm of probability distribution of response variable Q_n. Logistic regression model assumes that logarithm is a linear function of random variable x [9]. Then the $Bernoull$ probability density function can be written in the following exponential form:

$$\begin{aligned} f^n(1-f)^{1-q} &= (1-f)\exp\left[\log\left(\frac{f}{1-f}\right)^q\right] \\ &= (1-f)\exp\left[q\cdot\log\left(\frac{f}{1-f}\right)\right] \end{aligned} \tag{2}$$

where exp represents an exponential function based on the natural constant e. For (1), it can be rewritten as:

$$f(x) = \frac{e^{a+bx}}{1+e^{a+bx}} \tag{3}$$

where $f(x)$ is the kernel function, and e represents the bottom number of logarithms. Because:

$$\begin{aligned} f(x) &= \frac{e^{a+bx}\mathrm{n}b\mathrm{n}\left(1+e^{a+bx}\right) - e^{a+bx}\mathrm{n}b\mathrm{n}e^{a+bx}}{(1+e^{a+bx})^2} \\ &= \frac{e^{a+bx}\mathrm{n}b}{1+e^{a+bx}}\mathrm{n}\frac{1}{1+e^{a+bx}} \\ &= b(x)\mathrm{n}f(x)\mathrm{n}(1-f(x)) \end{aligned} \tag{4}$$

When b is positive, $f(x)$ is a strictly increasing function; when b is negative, $f(x)$ is a strictly decreasing function; in particular, when b =0, $f(x) = \frac{e^a}{1+e^a}$, it is a simple linear regression model [10]. In addition, in the logistic regression model $f(x) = \frac{e^a}{1+e^a}$, there exists $f\left(-\frac{a}{b}\right) = \frac{1}{2}$, that is, the

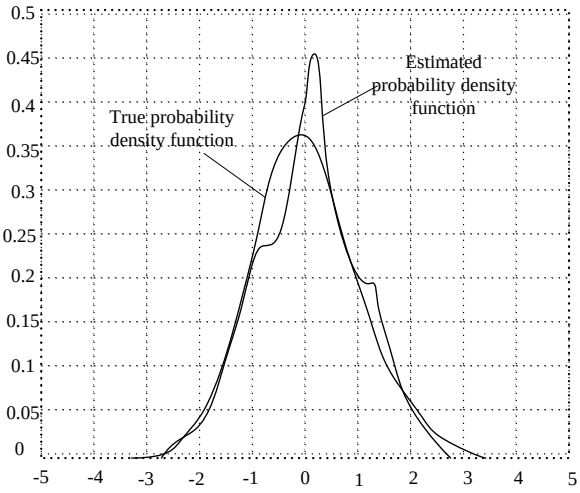
model is symmetric and the symmetry axis is $x=-\frac{a}{b}$. We can get $f\left(-\frac{a}{b}+c\right)=1-f\left(-\frac{a}{b}-c\right)$. a,b,c are the natural constants in the above equations. It can be seen from the above process that the kernel function derived from logistic regression model conforms to the basic properties of the kernel function and can be used for probability density function estimation.

B. Estimation of Probability Density Function

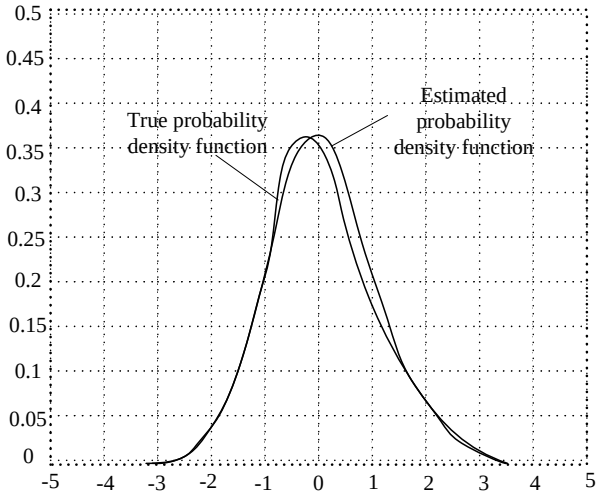
The logistic regression model with kernel function is established to determine the kernel function used for estimating the probability density function $f(x)$, and the unknown probability density function is estimated using a parametric-free method. Then we have:

$$G(x)=\frac{\sum_{i=1}^{n} f\left(\frac{x-x_i}{l}\right)}{nl} \tag{5}$$

where, $G(x)$ denotes an unknown probability density function, $f(x)$ denotes a kernel function, and l denotes window width. The selection of kernel functions is of great importance on the final estimated probability density function [11], as shown in Fig. 1.



(a) Comparison of probability density function with window width of 0.25



(b) Comparison of probability density function with window width of 0.45

Fig. 1 Effect of different window width on estimation of probability density function

As shown in Fig. 1(a), when the window width $l=0.25$, the estimated probability density function differs greatly from the true probability density function, with the biggest difference reaching 0.1, which is a very serious error in the operation and output of the results. By analyzing 1(b), when $l=0.45$, the estimated probability density function fits basically with the true probability estimation density function. The maximum difference is kept within 0.01, which is far less than the result when the window width is 0.25 [12].

Therefore, the concept of mean square integral error and the optimal window width after i are introduced to ensure the correctness and reliability of the estimation results of probability density function.

The mean square integral error (MISE) is used to measure the difference between the estimated probability density function $G(x)$ and the true probability density function $G'(x)$, which is expressed as:

$$MISE(l)=\delta\left[\int\left\{\left|G(x)-G'(x)\right|\right\}^2 dx\right] \tag{6}$$

where δ denotes the mean. MISE is used as a standard to measure the degree of difference between the true probability density function and the estimated probability density function. Other metrics include mean absolute difference, mean variance and integral variance [13], of which the equations are as follow:

$$MAE(l)=\delta\left[\left|G(x)-G'(x)\right|\right] \tag{7}$$

$$MSE(l)=\delta\left[\left|G(x)-G'(x)\right|^2\right] \tag{8}$$

$$ISE(l)=\int\left\{G(x)-G'(x)\right\}^2 dx \tag{9}$$

where, equation (7) represents mean absolute value difference, equation (8) represents mean variance, and equation 9 represents integral variance. Compared with ISE and MISE standards, MAE and MSE standards are relatively easy to be calculated mathematically. However, from (7) and (8), MAE

and MSE standard forms are relatively more simple. Solutions to MAE and MSE standards have been included in the process of calculating ISE standards, and ISE standards make full use of the information given in the process of calculation [14]. The disadvantage of ISE standard is that it only considers the information of observed values given, without considering the information of unknown samples obeying the distribution $G^{'}(x)$, and ISE standard does not give the degree of difference between $G(x)$ and $G^{'}(x)$ globally. Therefore, we use the MISE standard as the error measure of $G(x)$ and $G^{'}(x)$, which is more suitable for measuring the error of $G(x)$ and $G^{'}(x)$ on the whole of the known and unknown sample sets, and the calculation process of the MISE standard includes the calculation of MAE and MSE standards [15].

The calculation equation of MISE standard is deduced, and the calculation equation of optimal window width is given accordingly:

$$\begin{aligned} MISE(l) &= \delta\left[\int\left\{\left|G(x)-G^{'}(x)\right|\right\}^2 dx\right] = \int\delta\left[\left\{\left|G(x)-G^{'}(x)\right|\right\}^2\right]dx \\ &= \int\delta\left\{G(x)-\delta\left[G(x)\right]+\delta\left[G(x)\right]-G^{'}(x)\right\}^2 dx \\ &= \int\left[\delta\left\{G(x)-\delta\left[G(x)\right]\right\}^2+\left\{\delta\left[G(x)\right]-G^{'}(x)\right\}^2\right]dx \\ &= \int\left[\mathrm{var}\left(G(x)\right)+bias^2\left(G(x)\right)\right]dx \\ &= \int\mathrm{var}\left(G(x)\right)dx+\int bias^2\left(G(x)\right)dx \end{aligned} \tag{10}$$

where, var denotes variance and bias denotes deviation. Since $\delta(x)=\int_{-\infty}^{+\infty}xG^{'}(x)dx$ and $\dfrac{\sum_{i=1}^{n}f\left(\dfrac{x-x_i}{l}\right)}{nl}$, we have:

$$\begin{aligned} \delta\left[G(x)\right] &= \delta\left[\frac{1}{nl}\sum_{i=1}^{n}f\left(\frac{x-x_i}{l}\right)\right] = \delta\left[\frac{1}{l}f\left(\frac{x-y}{l}\right)\right] \\ &= \int\left[\frac{1}{l}f\left(\frac{x-y}{l}\right)G(y)\right]dy \end{aligned} \tag{11}$$

where, *dy* denotes a differential. The terms $\mathrm{var}\left(G(x)\right)$ and $bias^2\left(G(x)\right)$ in (10) are calculated by using the above equations:

$$\begin{aligned} &\mathrm{var}\left(G(x)\right) = \delta\left\{G(x)-\delta\left[G(x)\right]\right\}^2 \\ &= \delta\left\{\left[G(x)\right]^2-2G(x)\delta\left[G(x)\right]+\delta\left[G(x)\right]\right\}^2 \\ &= \delta\left[G(x)\right]^2-\left\{\delta\left[G(x)\right]\right\}^2 \\ &= \delta\left[\frac{1}{nl}\sum_{i=1}^{n}f\left(\frac{x-x_i}{l}\right)\right]^2-\left\{\delta\left[\frac{1}{nl}\sum_{i=1}^{n}f\left(\frac{x-x_i}{l}\right)\right]\right\}^2 \\ &= \frac{1}{n}\int\left[\frac{1}{l^2}f\left(\frac{x-y}{l}\right)G(y)\right]dy-\frac{1}{n}\left\{\int\left[\frac{1}{l^2}f\left(\frac{x-y}{l}\right)G(y)\right]dy\right\}^2 \end{aligned} \tag{12}$$

Let the intermediate quantity $z=\dfrac{x-y}{l}$, then $y=x-zl$, which is substituted into (12) to obtain:

$$\begin{aligned} &\mathrm{var}\left(G(x)\right) \\ &= \frac{1}{nl}\int\left[f(z)G(x-zl)\right]dz-\frac{1}{nl}\left\{\int\left[f(z)G(x-zl)\right]dz\right\}^2 \end{aligned} \tag{13}$$

Using Taylor expansion to expand $G(x-zl)$ term in (13), we have:

$$G(x-zl)=G(x)-lzG^{'}(x)+\frac{1}{2}l^2z^2G^{''}(x) \tag{14}$$

The equation of $bias^2\left(G(x)\right)$ is now calculated:

$$\begin{aligned} bias^2\left(G(x)\right) &= \delta\left[G(x)\right]-G^{'}(x) = \int\left[\frac{1}{l}f\left(\frac{x-y}{l}\right)G(y)\right]dy-G^{'}(x) \\ &= \int\left[f(z)G(x-lz)\right]dz-G^{'}(x) \\ &= \int\left\{f(z)\left[G(x-lz)-G^{'}(x)\right]\right\}dz \end{aligned} \tag{15}$$

By substituting (14) into (15), we have:

$$\begin{aligned} bias^2\left(G(x)\right) &= \int\left\{f(z)\left[G^{'}(x)-lzG^{''}(x)+\frac{1}{2}l^2z^2G^{'''}(x)\right]\right\}dz \\ &= -lG^{''}(x)\int zf(z)dz+\frac{1}{2}l^2G^{'''}(x)\int z^2f(z)dz \end{aligned} \tag{16}$$

where $\int zf(z)dz=0$ and $\int f(z)dz=1$, equation (16) is transformed into:

$$bias^2\left(G(x)\right)=\frac{1}{2}l^2G^{'''}(x)\int z^2f(z)dz \tag{17}$$

According to (13):

$$\begin{aligned} &\mathrm{var}\left(G(x)\right) \\ &= \frac{1}{nl}\int\left[f(z)^2G(x-zl)\right]dz-\frac{1}{n}\left\{G^{'}(x)+bias\left(G(x)\right)\right\}^2 \end{aligned} \tag{18}$$

To sum up:

$$\mathrm{var}\left(G(x)\right)=\frac{1}{nl}G(x)\int f(z)^2dz, n\to\infty \tag{19}$$

$$bias^2\left(G(x)\right)=\frac{1}{2}l^2G^{'''}(x)\int z^2f(z)dz, n\to\infty \tag{20}$$

Substituting (19) and (20) into (10) yields:

$$\begin{aligned} MISE(l) &= \int\mathrm{var}\left(G(x)\right)dx+\int bias^2\left(G(x)\right)dx \\ &= \int\left[\frac{1}{nl}G(x)\int f(z)^2dz\right]dx+\int\left[\frac{1}{2}l^2G^{'''}(x)\int z^2f(z)dz\right]^2 \\ &= \frac{1}{nl}\left[\int f(z)^2dz\right]\left[\int G^{'}(x)dx\right]+\frac{1}{4}l^4\left[\int z^2f(z)dz\right]^2\left\{\int\left[G^{'''}(x)\right]^2dx\right\} \end{aligned} \tag{21}$$

Where $\int G(z)dz=1$, then we finally get:

$$MISE(l)=\frac{1}{nl}\left[\int f(z)^2dz\right]+\frac{1}{4}l^4\left[\int z^2f(z)dz\right]\left\{\int\left[G^{'''}(x)\right]^2dx\right\} \tag{22}$$

Through the above transformation process, the final $MISE(l)$ equation is obtained, and equation (22) is called the error of asymptotic mean square integral. In summary, the criteria for judging the error between the estimated probability density function and the true probability density function are finally obtained, and equation (22) is simplified into:

$$MISE(l)=\frac{1}{nl}W(f)+\frac{1}{4}l^4\left[r_2(f)\right]^2W\left(G^{'''}\right) \tag{23}$$

where $W(f)=\int f(z)^2\,dz$, $r_2(f)=\int z^2 f(z)dz$, $W\left(G^{'}{}''\right)=\int\left[G^{'}{}''(x)\right]^2 dx$, it can be seen from the equation that $MISE(l)$ is an equation about window width l. The optimal window width l_{ex} can make the value of the equation $MISE(l)$ reach the minimum window width value, then the optimal window width value l_{ex} is calculated by deriving the function $MISE(l)$.

Let $\frac{d}{dl}\left[MISE(l)\right]=0$, then the optimal calculation equation of window value is:

$$\frac{d}{dl}\left\{\frac{1}{nl}W(f)+\frac{1}{4}l^4\left[r_2(f)\right]W\left(G^{'}{}''\right)\right\} =-\frac{1}{nl^2}W(f)+l^3\left[r_2(f)\right]^2 W\left(G^{'}{}''\right)=0 \tag{24}$$

To solve this equation, we have:

$$l_{ex}=\left\{\frac{W(f)}{\left[r_2(f)\right]^2 W\left(G^{'}{}''\right)}\right\}^{\frac{1}{5}} \tag{25}$$

The minimum value of $MISE(l)$ can be obtained by introducing l_{ex} into (23). At this time, the error between the estimated probability density function $G(x)$ and the true probability density function $G^{'}(x)$ is minimized. The equation of the minimum value of $MISE(l)$ is:

$$\underset{\min}{MISE(l)}=\frac{5}{4}\left\{\left[r_2(f)\right]^2\left[W(f)\right]^4 W\left(G^{'}{}''\right)\right\}^{\frac{1}{5}} n^{-\frac{4}{5}} \tag{26}$$

By observing (25), it can be found that the values of function $W(f)$ and $r_2(f)$ are directly related to the kernel function selected. When given a specific kernel function, the values of function $W(f)$ and $r_2(f)$ can be determined by calculation, and then the probability density function can be determined. Then the probability density function is analyzed from the aspects of consistency and convergence rate.

C. Consistency Analysis of Probability Density Function

Consistency is a basic requirement for the estimated probability density function. If an estimated probability density function cannot estimate the estimated parameters in the probability density function to arbitrary accuracy while the sample size is constantly increasing, then the estimation is not credible. Therefore, the consistency analysis of probability density function is very important.

Assuming that both the kernel function $f(x')$ and the probability density function $G(x)$ are functions defined on $(-\infty,+\infty)$, and both of them satisfy the following conditions:

(1) Kernel functions are bounded on $(-\infty,+\infty)$;

(2) The derivation of the absolute value of the kernel function in the interval $(-\infty,+\infty)$ is less than that the bound on $(-\infty,+\infty)$;

(3) The value of the absolute value of the nuclear function after calculating the limit is a natural number;

(4) The difference of the absolute value of the probability density function in the interval $(-\infty,+\infty)$ is less than that the bound on $(-\infty,+\infty)$;

When the window width l is a constant that is infinitely close to 0 but greater than 0, we have:

$$\lim_{l\to 0}\frac{1}{l}\int_{-\infty}^{+\infty} f\left(\frac{x'}{l}\right)G(x-x')dt=G(x)\int_{-\infty}^{+\infty} f(x')dt \tag{27}$$

The above is a probability density function analysis theorem, which needs to be proved according to the kernel function obtained in the above process and the estimated probability density function. For the convenience of calculation, the kernel function is expressed as $f(t_i)$ and the calculation can be as follow:

$$G(x)-\delta\left[G(x)\right]=\frac{1}{nl}\sum_{i=1}^{n}a(t_i) \tag{28}$$

where $a(t_i)=f\left(\frac{x-t_i}{l}\right)-\delta\left[f\left(\frac{x-t_i}{l}\right)\right]$, for any $a\ge 1$, there is:

$$\begin{aligned}&\delta\left|a(t_i)\right|=\delta\left|f\left(\frac{x-t_i}{l}\right)-\delta\left[f\left(\frac{x-t_i}{l}\right)\right]\right|^a\\&\le 2^{a-1}\left\{\delta\left|f\left(\frac{x-t_i}{l}\right)\right|^a+\left|\delta\left[f\left(\frac{x-t_i}{l}\right)\right]\right|^a\right\}\\&=2^a\int_{-\infty}^{+\infty}\left|\delta\left(\frac{x-t_i}{l}\right)\right|^a G(t_i)\,dt_i\\&=2\int_{-\infty}^{+\infty}\left|f\left(\frac{u}{l}\right)\right|G(x-u)\,du\end{aligned} \tag{29}$$

It is known that the kernel function $f(t_i)$ satisfies the conditions (1), (2), (3), so it is easy to verify that the $\left|a(t_i)\right|(a\ge 1)$ also satisfies the above conditions (1), (2), (3), so when $a\ge 1$, $\delta\left|G(x)-\delta\left[G(x)\right]\right|=\left(nl^2\right)$. At this point, it is only necessary to prove that the equation $\lim_{n\to\infty}\delta\left|G(x)-G^{'}(x)\right|=0$ holds, then the consistency of the estimated probability density function can be proved.

Through the above argument, when $\int_{-\infty}^{+\infty} f(t)dt=1$, and $\lim_{n\to\infty} l=0$, $\lim_{n\to\infty} nl^2=\infty$, there exists $\lim_{n\to\infty}\delta\left|G(x)-G^{'}(x)\right|=0$. Then we can see the consistency of the probability density function.

To calculate:

$$\delta\left|G(x)-G^{'}(x)\right|\le\left\{\delta\left|G(x)-\delta\left[G(x)\right]\right|+\left|\delta\left[G(x)\right]-G^{'}(x)\right|\right\} \tag{30}$$

According to the calculation of (30):

$$\left|\delta\left[G(x)\right]-G^{'}(x)\right|\to 0(n\to\infty) \tag{31}$$

It just proves:

$$\delta\left|G(x)-\delta\left[G(x)\right]\right|\to 0(n\to\infty) \tag{32}$$

Because:

$$\begin{aligned}\delta\left|G(x)-\delta\left[G(x)\right]\right| &\le \left\{\delta\left|G(x)-\delta\left[G(x)\right]\right|^{2n}\right\}^{\frac{1}{2n}} \\ &=\left[\left(nl^2\right)^{-n}\right]^{\frac{1}{2n}}\end{aligned} \tag{33}$$

So when $n\to\infty$, there is:

$$\delta\left|G(x)-\delta\left[G(x)\right]\right|\to 0 \tag{34}$$

In summary

$$\lim_{n\to\infty}\delta\left|G(x)-G^{'}(x)\right|=0 \tag{35}$$

Through the above proof process, we can verify the consistency between the estimated probability density function and the real probability density function, indicating the estimated probability density function with has strong consistency.

D. Convergence Speed Analysis of Probability Density Function

In the analysis of probability density function, the speed at which a convergent sequence approximates its limit is called convergence speed. Assuming that both the kernel function $f(x')$ and the probability density function $G(x)$ are functions defined on $(-\infty,+\infty)$, and both of them satisfy the following conditions:

(1) Kernel functions are bounded on $(-\infty,+\infty)$;

(2) The derivation of the absolute value of the kernel function in the interval $(-\infty,+\infty)$ is less than that bound on $(-\infty,+\infty)$;

(3) The value of the absolute value of the nuclear function after calculating the limit is a natural number;

If $G"(x)$ is bounded and continuous everywhere, $\lim_{n\to\infty} l=0$, $\lim_{n\to\infty} nl^2=\infty$, $\lim_{n\to\infty} nl^6\to\infty$, if there is a positive sequence $\{C_n\}$ and $C_n=\left(nl^2\right)^{\frac{1}{2}}$ is satisfied, then we have:

$$\lim_{n\to\infty}\delta\left|\hat{G}(x)-G(x)\right|=\frac{1}{C_n} \tag{36}$$

Then, the convergence speed of the probability density function can be verified only by proving $C_n\delta\left|\hat{G}(x)-G(x)\right|\to 0$.

Because:

$$C_n\delta\left|\hat{G}(x)-G(x)\right|\le C\delta\left|\hat{G}(x)-\delta\left[\hat{G}(x)\right]\right|C_n+C\left|\delta\left[\hat{G}(x)\right]-G(x)\right|C_n \tag{37}$$

And

$$\begin{aligned}\delta\left|\hat{G}(x)-\delta\left[\hat{G}(x)\right]\right|C_n &\le C_n\left[\left(nl^2\right)^{-n}\right]^{\frac{1}{2n}} \\ &=C_n\left[\frac{\left(nl^2\right)^{\frac{1}{2}}}{\left(nl^2\right)^{-n}}\right]^{\frac{1}{2n}}\left(nl^2\right)^{-n\frac{1}{2n}} \\ &=C_n\left(nl^2\right)^{-\frac{1}{2}}\to 0,(n\to\infty)\end{aligned} \tag{38}$$

From the derivation of (29), there is:

$$\begin{aligned}\delta\left[\hat{G}(x)\right]-G(x) &= \int_{-\infty}^{+\infty} f(u)\left[G(x-ul)-G(x)\right]du \\ &=\frac{1}{2}G"(x)l^2+l^2\end{aligned} \tag{39}$$

Therefore:

$$\begin{aligned}C_n\delta\left|\hat{G}(x)-G(x)\right| &= C_n\left\{\frac{1}{2}l^2\left|G"(x)\right|\right\}+\text{infinitesimal of higher order} \\ &=\left[\left(nl^2\right)^{\frac{1}{2}}\right]\mathrm{n}l^2 \\ &=\left[\left(nl^6\right)^{\frac{1}{2}}\right]\to 0\end{aligned} \tag{40}$$

Finally, we have:

$$C_n\delta\left|\hat{G}(x)-G(x)\right|\to 0,(n\to\infty) \tag{41}$$

In conclusion, the probability density function has a high convergence speed.

Through the analysis of consistency and convergence speed of probability density function, the probability density function analysis based on logistic regression model is realized.

III. Simulation Experiment

Since China's Shanghai stock market started early, the data can fully represent the development of China's stock market. Moreover, broad market index integrates the impact of various aspects, so it can well reflect the relevant situation of China's stock market. The daily index of the KLCI is selected as our research object. We choose the closing price of the index from January 2, 2015 to June 30, 2018 as a sample. The closing price of the day is the opening price of the model's explanatory variable on the day and the closing and opening prices of the previous five days are the explanatory variables of the model. A rolling prediction method is implemented by inputting 1571 samples into the neural network quantile regression model, so as to determine the model structure, train and stabilize the neural network, and predict the acquisition from July 1, 2018 to July 2018. The closing price of the Shanghai Composite Index is during the 14th-200 consecutive conditional quantiles per day, and then substituted into the kernel density estimation model to determine the probability density curve of stock price changes on a certain day in the future. These 1571 samples are characterized by obvious "spike tail", and their descriptive statistics are shown in Table I.

Table I Shanghai Composite sample descriptive statistics

N	1571

Maximum	5497.9
Minimum	1706.7
Skewness	1.66188
kurtosis	7.51889

From Table I, we can conclude that the kurtosis value of this sample is 7.51889, which is obviously larger than the kurtosis value of the normal distribution. At the same time, the skewness value is 1.66188, which fully proves the "peak and thick tail" characteristics of the sample, instead of obeying the usual normal distribution. With traditional methods, it is difficult to establish the relationship between response variables and input variables. The quantile regression and kernel density estimation methods proposed in this paper not only overcome the defects of non-normal distribution of samples, but also obtain more useful information about future stock prices.

Therefore, we input the samples obtained by the rolling method into the neural network quantile regression network, train the neural network structure, and bring the 200 conditional quantiles of the Shanghai Composite Index into the kernel density estimation method every day. The complete probability density curve of the Shanghai Composite Index on a future day is shown in Fig. 2. As can be seen from Fig. 2, using the proposed method, we can first obtain a continuous probability density function graph of the stock price on a certain day in the future.

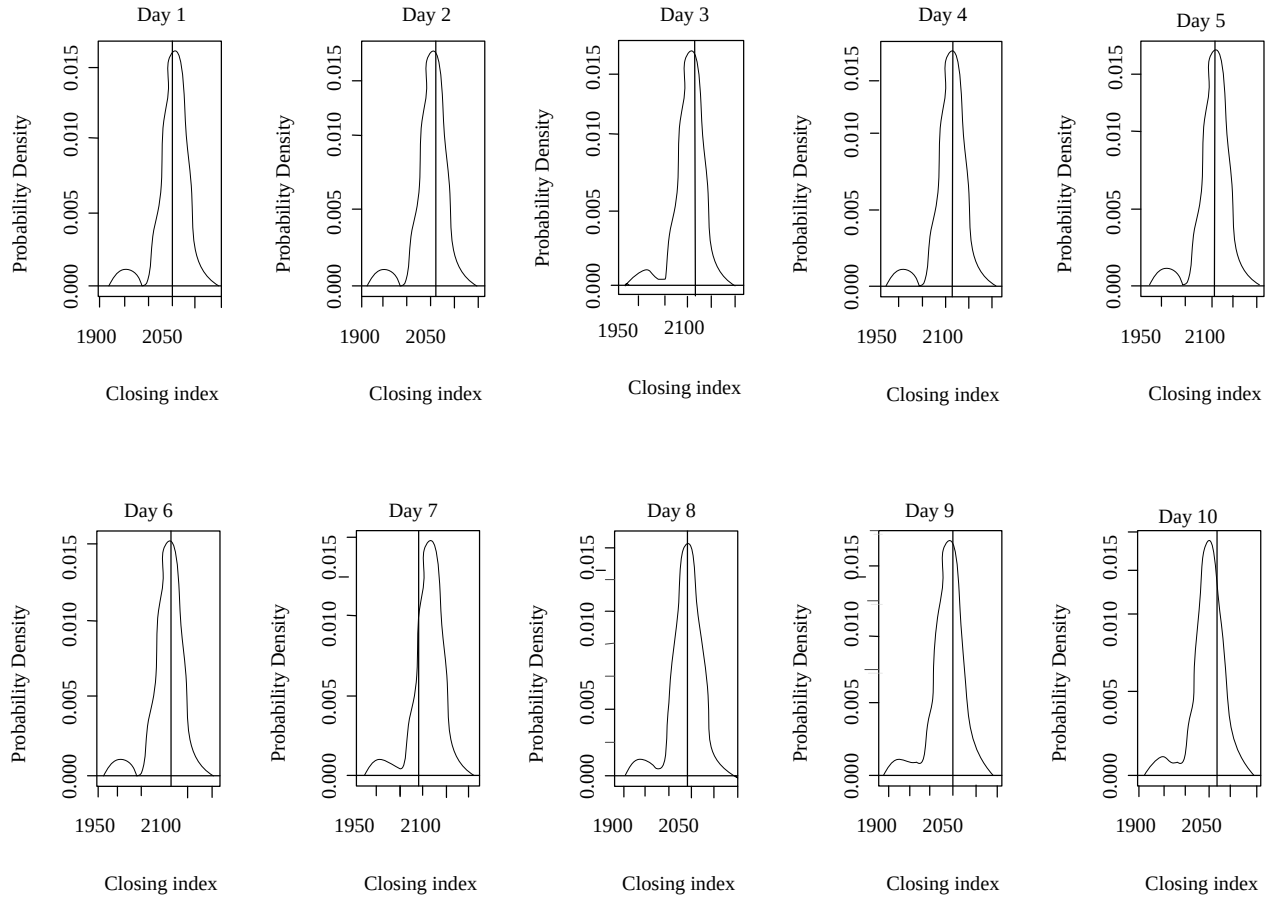


Fig. 2 Diagram of probability density curve of stock price

From the complete probability density curve, not only can the probability of each stock price appear, but also the probability situation of the stock price range can be obtained. Moreover, when it is sent to investors for investment, more investment decision basis is provided. Secondly, the closing prices are basically near the highest probability point on the probability density curve. A better prediction of the true value of the closing price provides investors with a stronger basis for decision-making. Finally, the closing price, the best price, and the lowest price of the stock all appear in the estimated probability. On the density curve, they all appear near the highest probability point. In this way, the probability density of the daily price fluctuation interval of the stock price can be obtained, which is beneficial for investors to predict the future stock price change interval. The proposed method can not only achieve more accurate point prediction values, but also provide more beneficial decision-making information for stock investors. Through analyzing the complete probability density curve of the stock price and calculating the mean, variance, and skewness of the stock price distribution characteristics, we can

more intuitively understand the position of stock price changes and provide better decision support for investors.

Table II Stock price prediction results and relative errors

Date	Closing price	Highest price	Floor price	Forecast price	Relative error/%
2018-7-1	2050.38	2066.64	2041.94	2053.49	0.15
2018-7-2	2059.42	2060.60	2044.04	2048.44	-0.53
2018-7-3	2063.23	2066.64	2048.08	2053.08	-0.49
2018-7-4	2059.38	2065.08	2054.22	2064.42	0.20
2018-7-7	2059.93	2064.04	2050.89	2061.98	0.10
2018-7-8	2064.02	2064.43	2047.20	2059.37	-0.22
2018-7-9	2038.61	2062.47	2038.61	2063.14	1.20
2018-7-10	2038.34	2045.53	2034.96	2038.76	0.02
2018-7-11	2046.96	2051.24	2033.00	2038.09	-0.43
2018-7-14	2066.65	2067.34	2044.90	2048.94	-0.86

As shown in Table II, using the proposed method, the predicted maximum relative error of the stock price on a certain day in the future is 1.20%, and the predicted the minimum relative error is 0.02%, with an average absolute error of 0.42%. This result provides reference for investors when making investment decisions. This indicates that the proposed method can be well applied to the actual scene analysis, and provide stable operation results.

As shown in Fig. 3 and Table III, when the traditional Shangsi kernel is combined with the thumb principle for stock price prediction, although we can still obtain its complete probability density curve, the average absolute error obtained is 0.425%. These fully demonstrate that the neural network-based quantile regression and kernel density estimation method proposed in this paper achieve better prediction results and can more accurately reflect the real situation of the predicted object.

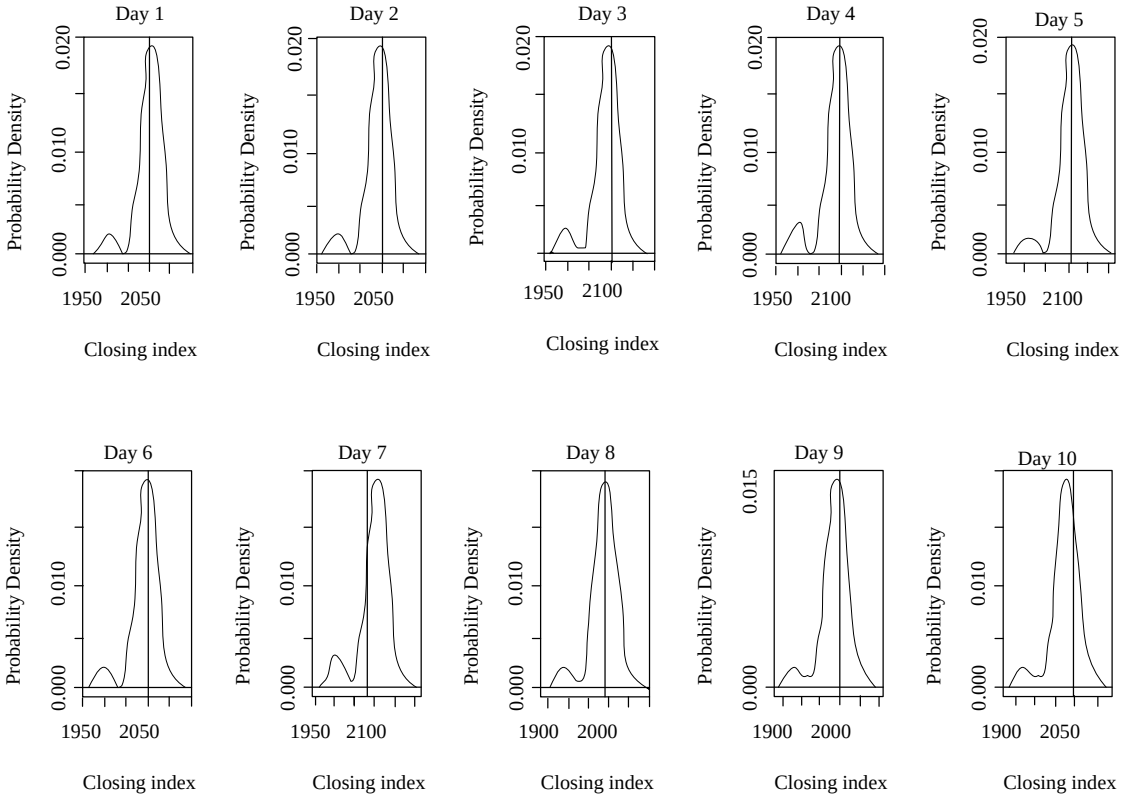


Fig. 3 Stock price odds density curve under Lithuanian nuclear

Table III Gaussian nuclear stock price prediction results and relative errors

Date	Closing price	Highest price	Floor price	Forecast price	Relative error /%
2018-7-1	2050.38	2066.64	2041.94	2053.43	0.09
2018-7-2	2059.42	2060.60	2044.04	2048.44	-0.52
2018-7-3	2063.23	2066.64	2048.08	2052.18	-0.54

2018-7-4	2059.38	2065.08	2054.22	2062.66	0.16
2018-7-7	2059.93	2064.04	2050.89	2060.90	0.05
2018-7-8	2064.02	2064.43	2047.20	2058.35	-0.27
2018-7-9	2038.61	2062.47	2038.61	2062.47	1.17
2018-7-10	2038.34	2045.53	2034.96	2037.49	0.04
2018-7-11	2046.96	2051.24	2033.00	2036.34	-0.52
2018-7-14	2066.65	2067.34	2044.90	2048.96	-0.89

I. Conclusions

The analysis of PDF is of great help for solving machine learning problems and pattern recognition problems. Aiming at the disadvantages of the previous PDF analysis methods, the paper proposes the PDF analysis based on Logistic regression model, and take the advantages of logistic regression analysis to solve the problems in the traditional methods. Through comparative experiments, it is clear that the probability density function analysis method based on logistic regression model exhibits better performance, which provides basis and support for the future study of probability density function. The basic idea of this paper can be further extended to more complex stochastic dynamical systems, such as high-dimensional, non-white noise excitation, Markov processes, etc.

References

[1] J. Yao, and D. Liu, "Logistic regression analysis of risk factors for intracranial infection after multiple traumatic craniotomy and preventive measures," The Journal of Craniofacial Surgery, vol. 30, no. 7, pp. 1946-1948, 2019.

[2] A. Albasri, S. Prinjha, R. J. McManus, and J. P. Sheppard, "Hypertension referrals from community pharmacy to general practice: multivariate logistic regression analysis of 131 419 patients," The British Journal of General Practice: The Journal of the Royal College of General Practitioners, vol. 68, no. 673, pp. e541-e550, 2018.

[3] Y. Xie, W. Yi, L. Zhang, Y. Lu, and M. Li, "Evaluation of a Logistic regression model for predicting liver necroinflammation in hepatitis B e antigen-negative chronic hepatitis B patients with normal and minimally increased alanine aminotransferase levels," Journal of Viral Hepatitis, vol. 26, no. 1, pp. 42-49, 2019.

[4] J. Zetterqvist, K. Vermeulen, S. Vansteelandt, and A. Sjölander, "Doubly robust conditional Logistic regression," Statistics in Medicine, vol. 38, no. 23, pp. 4749-4760, 2019.

[5] X. Gong, J. L. Cui, Z. P. Jiang, L. J. Lu, and X. C. Li, "Risk factors for pedicled flap necrosis in hand soft tissue reconstruction: A multivariate Logistic regression analysis," ANZ Journal of Surgery, vol. 88, no. 3, pp. e127-e131, 2018.

[6] H. Byeon, "A laryngeal disorders prediction model based on cluster analysis and regression analysis," Medicine, vol. 98, no. 31, e16686, 2019.

[7] J. B. Chen, and M. Z. Lu, "A new method for solving the probability density of the maximum absolute value process of a class of Markov processes," Journal of Mechanics, vol. 51, no. 5, pp. 173-183, 2019.

[8] J. Chang, "Probability density and distribution function of continuous random variables," Science and Technology Information, vol. 17, no. 23, pp. 188-189, 2019.

[9] Z. S. Yu, G. H. Ren, Z. Y. Yu, C. H. N. Wei, and H. Y. Fan, "Time evolution of the wigner operator as a quasi-density operator in amplitude dessipative channel," International Journal of Theoretical Physics, vol. 57, no. 6, pp. 1888-1893, 2018.

[10] S. H. C. M. Veen, R. C. Kleef, W. P. M. M. Ven, and R. C. J. A. Vliet, "Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees," Health Economics, vol. 27, no. 2, pp. e1-e12, 2018.

[11] H. J. Lee, "Simple regression model for estimating reflectance reduction due to random surface roughness," International Journal of Thermophysics, vol. 40, no. 6, pp. 1-12, 2019.

[12] M. J. Lee, M. H. Rahbar, and H. Talebi, "A nonparametric method for assessment of interactions in a median regression model for analyzing right censored data," Statistical Methods in Medical Research, vol. 28, no. 4, pp. 1170-1187, 2019.

[13] H. R. Merrill, X. Y. Tang, and N. Bliznyuk, "Spatio-temporal additive regression model selection for urban water demand," Stochastic Environmental Research and Risk Assessment, vol. 33, no. 4-6, pp. 1075-1087, 2019.

[14] M. D. Iorio, N. Gallot, B. Valcarcel, and L. Wedderburn, "A Bayesian semiparametric Markov regression model for juvenile dermatomyositis," Statistics in Medicine, vol. 37, no. 10, pp. 1711-1731, 2018.

[15] S. V. Suryakala, and S. Prince, "Investigation of goodness of model data fit using PLSR and PCR regression models to determine informative wavelength band in NIR region for non-invasive blood glucose prediction," Optical and Quantum Electronics, vol. 51, no. 8, pp. 1-20, 2019.

Lingling Fang, female, was born in December 1979. Her title is associate professor. In 2003, she received a bachelor's degree in mathematics and applied mathematics from Jiangxi Normal

University. In 2008, she received a master's degree in basic mathematics from Nanjing Normal University. She is now working in Jiangxi University of technology. Her research fields include mathematics and applied mathematics education and mathematical education research. She has published eight academic papers and participated in three research projects.

Yunxia Zhang, female, was born in January 1982. Her title is lecturer. In 2005, she received a bachelor's degree in mathematics and applied mathematics from Harbin Normal University. In 2007, she received a master's degree in basic mathematics from Harbin Institute of Technology. She is now working in Jiangxi University of technology. Her research fields include probability theory and mathematical statistics, fuzzy mathematics and mathematical education research. She has published four academic papers and participated in two research projects.

Author Contributions:

Lingling Fang established a logistic regression model with kernel function. Yunxia Zhang analyzed the estimated probability density function. All authors conducted the experiments and analysed the results. All authors discussed the results and wrote the manuscript.