Estimating the distribution density function using a DOG wavelet

V S Timofeev¹, E V Isaeva¹, ²

¹ Novosibirsk State Technical University, 20, Karl Marks ave., Novosibirsk, 630073, Russia

² E-mail: isaeva@corp.nstu.ru

Abstract. It is often necessary to obtain information on the distribution form of a random variable within the framework of statistical data analysis. There are several ways of estimating probability density distributions. Wavelet density estimators distribution of a random variable, built with DOG wavelet, would be considered. Three methods are proposed for determining the normalizing parameter, which provides a correction for the non-orthogonality of the DOG wavelet. In addition, studies have been carried out on the impact on quality wavelet estimates of parameters such as sample size, a number of coefficients of expansion, a function by series in the expression for wavelet density estimates. In the course of this process, it was noted that the quality of estimates on the smoothing parameter is substantial. Moreover, there is its best meaning. The conclusion is made that the estimate is based on DOG wavelet, provides quality recovery density functions.

1. Introduction and problem statement
Solving practical tasks, related to statistical analysis of data, usually presupposes the existence of the information about the form of the distribution signs of stochastic nature. So, to perform a regression analysis using maximum likelihood is impossible without knowledge distribution law random errors. Only very rarely this information is available a priori, but one can get information (from existing) statistical data available [1]. One of the characteristics that contains complete information about the distribution investigated signs is the density function. The estimate density function of the statistical data received in the absence of information about the form of the distribution can be calculated using non-parametric methods. Parsen-Rosenblatt's kernel estimation [2] gets best known among such methods, and it was also used extensively by the authors in the construction of regression dependencies. However, its quality of estimates on the smoothing parameter is substantial. Sub-optimal smoothing parameter choices can lead to sharp fluctuations in graphics, the density functions is restored. This is especially true for small samples. This circumstance forces researchers be cautious in using kernel estimation.

This particular work explores a different approach. It must also be treated as a nonparametric. It is based on the use of Wavelet theory. Wavelet theory has been widely used to approximate different functions. Its aim is to explore Wavelet estimation, based on DOG wavelet and estimation of recovery density function quality in a variety of contexts including the small samples size.
Let \( \{ x_j \mid j = 1, n \} \) be a sample of independent values of a random variable \( \xi \), defined (given on) in an arbitrary segment \([c, d]\), where \( c = \min x_j \), \( d = \max x_j \). Law and density functions distribution \( f(t) \), a random variable \( \xi \) are unknown. It is required to construct a wavelet estimation of the distribution density using sample data.

2. Probability density wavelet estimations

According to [3,4], wavelet estimation of density function \( \hat{f}_n(t) \) of the random variable is given by series expansion on some orthonormal basis functions \( \psi_i(t) \):

\[
\hat{f}_n(t) = \sum_{i=1}^{N} \hat{c}_i \psi_i(t),
\]

(1)

where \( N \) is the number of members of several smoothing parameters, \( \hat{c}_i \)- estimates of the decomposition coefficients in this basis, which are determined according to available statistics:

\[
\hat{c}_i = \frac{1}{n} \sum_{j=1}^{n} \psi_i(x_j).
\]

(2)

In view of (2), the density function wavelet estimator can be written as:

\[
\hat{f}_n(t) = \frac{1}{n} \sum_{j=1}^{N} W_N(t, x_j),
\]

(3)

where \( W_N(t, x_j) = \sum_{i=1}^{N} \psi_i(t) \psi_i(x_j) \).

According to [3,4], a orthonormal system is \( \psi_i(t) \). We consider a set of orthonormal basis functions using \([0,1]\):

\[
\psi_i(t) = 2^{k/2} \psi \left( 2^k t - (j-1) \right),
\]

(4)

where \( k \geq 0, 1 \leq j \leq 2^k \), such that \( i = 2^k + j \). \( \psi(t) \) is mother wavelet [3-6] or wavelet function. The expression (4) is true for \( i > 1 \), when \( i = 1 \) function \( \psi_1(t) = 1 \) for all \( t \in [0,1] \), \( \psi_1(t) = 0 \) otherwise. According to [4], relation (1) can be used if the system of functions (4) is not orthogonal.

It is noted that the orthonormal system functions \( \psi_i(t) \) described above are defined on the interval \([0,1]\) and do not necessarily coincide with actual range \( \xi \). So to calculate \( \hat{f}_n(t) \) transition is required to orthonormal system functions \( \tilde{\psi}_i(t) \), defined on the same segment \([c, d]\) as a value of random variable. According to [7] we get basis functions transformation during the transition from one system functions to another:

\[
\tilde{\psi}_i(t) = \frac{1}{\sqrt{d-c}} \psi_i \left( \frac{t-c}{d-c} \right).
\]

(5)

The final wavelet estimation of density function \( \hat{f}_n(t) \) random variable is expressed by a decomposition (3) on arbitrary segments \([c, d]\) where instead of the basis functions \( \psi_i(t) \), \( \tilde{\psi}_i(t) \) are used.

The next step in working, estimate (3), which is based on a mother's DOG wavelet, will be fully considered and investigated.
3. Estimation of the density distribution using a DOG wavelet

We consider DOG wavelet [3, 4, 6], given by the ratio:

$$\psi(t) = e^{-t^2/2} - \frac{1}{2} e^{-t^2/8}$$

for which orthonormal system functions (4) on segments \([0,1]\) will be as follows:

$$\psi_i(t) = 2^{k/2} \left( e^{-\frac{1}{2} \left(2^k t - (j-1)\right)^2} - \frac{1}{2} e^{-\frac{1}{8} \left(2^k t - (j-1)\right)^2} \right),$$

where \(i, k, j\) are the same as (4). It can be shown that the DOG wavelet is not orthonormalized as a consequence of the system of functions (7).

Since the mother wavelet is not orthonormal, the quality of the restoration of the density function may be worse than that when using orthonormal wavelets [3-5, 7]. To improve the quality of the wavelet estimate \(\hat{f}_n(t)\) in such cases, a number of authors introduce a normalization factor of \(z\) for the mother wavelets, for example, for the Mexican hat wavelet \(z = \frac{1.031}{\sqrt{2}}\) and for the wavelet \(z = \frac{-1.786}{\sqrt{2}}\) [6]. In our case, such factor of \(z\) makes sense to introduce in relation (7):

$$\psi_i(t) = 2^{k/2} z \left( e^{-\frac{1}{2} \left(2^k t - (j-1)\right)^2} - \frac{1}{2} e^{-\frac{1}{8} \left(2^k t - (j-1)\right)^2} \right),$$

where values \(i, k, j\) are the same as (4). There are several ways to select option \(z\). The first is to use one of the main properties of the density function:

$$\int_{-\infty}^{+\infty} \hat{f}_n(t) = 1.$$  

(9)

The second method is to analyze the degree of closeness of \(\hat{f}_n(t)\) and \(f(t)\) by minimizing the value of statistics \(\chi^2_{\text{emp}}\) [8]:

$$\chi^2_{\text{emp}} = \sum_{i=1}^{T} \frac{\left(\hat{f}_n(t_i) - f(t_i)\right)^2}{f(t_i)},$$

(10)

where \(T\) is the number of sampling intervals obtained by the Sturges formula [8], \(t_i\) is the middle of the \(i\)-th interval. But in this case, the initial distribution density function should be known a priori, which is not always feasible in practice. Therefore, \(f(t)\) can be replaced by its empirical estimate obtained from the sample. This can actually be considered to determine \(z\). In this paper, option two is used for simplicity.

In view (5), the transition from system functions (8) to orthonormal system \(\tilde{\psi}_i(t)\) on arbitrary segments \([c,d]\) gives results:

$$\tilde{\psi}_i(t) = \sqrt{\frac{z^2}{d-c}} 2^{k/2} \left( e^{-\tau^2/2} - \frac{1}{2} e^{-\tau^2/8} \right).$$

(11)
where \( \tau = \frac{2^k}{d-c} (t-c) - (j-1) \), and values \( i, k, j \) are the same as (4). \( z \) is determined by one of the methods proposed above.

The result is that the wavelet estimation of the density function \( \hat{f}_n(t) \) of the random variable on arbitrary segments \([c,d]\) using DOG wavelet is given expansion (3) using orthonormal basis functions (11).

4. Study of a wavelet estimation of the density function

Several computational experiments were undertaken for comparison of the quality of the restored density function using a DOG wavelet. The studies were conducted in different sampling conditions and different meanings of the smoothing parameter.

We show how the choice of the value of parameter \( z \) affects the quality of the approximation of the density function. For this, we take a sample of \( n=1000 \), composed of independent values of random variables \( \xi \) distributed according to the normal law with parameters \((0,1)\). We construct from the sample a wavelet estimate of the density function at each value \( N \), starting with 5 to 50, as well as for each value \( z \), starting with 0.5 to 1 using a DOG wavelet. An analysis of the quantitative proximity of \( \hat{f}_n(t) \) and \( f(t) \) according to criterion \( \chi^2 \). The results for \( N=34;200 \) and \( z=0.5;0.7;1 \) are given in table 1 from which it is clear that \( \chi^2_{\text{emp}} \) – the minimum value corresponds to \( z=0.7 \).

| \( n \) | \( T \) | \( z=0.5 \) | \( z=0.7 \) | \( z=1 \) | \( z=0.5 \) | \( z=0.7 \) | \( z=1 \) |
|------|------|----------|----------|----------|----------|----------|----------|
| 100  | 8    | 0.29     | 0.12     | 1.14     | 0.624    | 0.38     | 2.87     |
| 500  | 10   | 0.37     | 0.07     | 1.13     | 0.49     | 0.135    | 1.85     |
| 1000 | 11   | 0.41     | 0.1      | 0.018    | 0.42     | 0.22     | 0.28     |

The results of recovery of \( f(t) \) at \( N=34 \) and \( z=0.5;0.7;1 \) are shown in Fig. 1. Together with the obtained estimates, the figure shows a graph of the true distribution density of the standard normal law \( f(t) \).

![Figure 1. Wavelet estimate of the density function based on DOG wavelet.](image-url)
We note that estimates \( \hat{f}_n(t) \) obtained at \( z=0.5 \) and \( z=1 \) do not allow us to consider them as distribution densities, since property (9) is not satisfied. This confirms the fact that the non-optimal choice of \( z \) may lead to an unsatisfactory result. Optimal choice \( z \) provides a high-quality restoration of the density function. It should also be noted that the value of \( \chi^2_{emp} \) is minimal at \( z=0.7 \) for any value of the smoothing parameter \( N \).

Further, taking into account the chosen factor of \( z \), the quality of the restoration of the density function of the standard normal distribution of \( \hat{f}_n(t) \) from the number of members of the \( N \) row (3) was investigated. For the sample simulated above, for each value of \( N \), starting from 5 and up to 200, wavelet estimates of the density function (3) were constructed using the DOG wavelet. Figure 2 presented the results of the recovery \( f(t) \) for \( N=5;40;200 \). The resulting estimates are presented together with a graph of the true distribution density of the normal law \( f(t) \) in fig. 2.

![Figure 2. Wavelet estimate of the density function based on DOG wavelet.](image)

It should be noted that type of estimation \( \hat{f}_n(t) \) depends heavily on the selection of the smoothing parameter. It can be seen that value \( N \) influences values of the magnitude of the deviation of the wavelet estimation from density distribution law. The deviation of the estimate \( \hat{f}_n(t) \) from the true distribution density function is significant when \( N \) is small. There is additional extreme for estimations of the density function that worsens the estimation quality when \( N \) is big. As a result estimation \( \hat{f}_n(t) \) gave the misleading impression about behavior of the true function of distributing \( f(t) \). The study concluded that the best smoothing parameter value is \( N=34 \) for DOG wavelet. Studies were also conducted on the dependence of the quality of estimates on values of the smoothing parameter \( N \) with smaller sample sizes \( n=50;100;500 \). The results qualitatively coincided with the above mentioned ones.

The best found value of the parameter \( N \) was used to study the dependence of the quality of estimation of the density function \( \hat{f}_n(t) \) on the sample size \( n \). For this, samples of independent random values \( \xi \) were modelled, distributed according to the normal law with parameters \((0,1)\) with size \( n=100;500;1000 \). Wavelet density function (3) estimation is built using DOG wavelet for each of them. Fig. 2 presented the results; the schedule of estimates is presented together with a graph of the true distribution density of the normal law \( f(t) \).
The results point to the initial sample size impact on the quality of wavelet estimate. There is a deviation of the estimate \( \hat{f}_n(t) \) from the true distribution density function, when \( N \) is small. It is presented in figure 3.

In a similar way, the study of choice was carried out, taking into account various laws of distribution. In particular, the actual law with parametric \( \lambda = 2 \) gamma distribution was considered. The results obtained qualitatively coincided with those given above.

5. Conclusion
In this work we examined the estimation of the distribution density function of a random variable using a DOG wavelet. Three methods are proposed for determining the normalizing parameter, which provides a correction for the non-orthogonality of the DOG wavelet. For this wavelet sample size the impact on quality of wavelet estimates was identified. There is a deviation of the estimate \( \hat{f}_n(t) \) from true distribution density function \( f(t) \) when \( N \) is small. The quality of estimates on the smoothing parameter is substantial. And there is its best meaning, which was obtained through the computational experiment.

References
[1] Timofeev V S and Khailenko E A 2010 Adaptive evaluation of parameters of regression models using generalized lambda – distribution Reports of the Academy of Sciences of Sciences of Higher School of the Russian Federation 2 (15) 25-36
[2] Fan J and Gijbels I 1996 Local Polynomial Modelling and Its Application (London: Chapman and Hall)
[3] Dobeshi I 2001 Ten Lectures on Weevletam (Izhevsk: Regular and Chaotic Dynamics)
[4] Chui C 2001 Introduction to wavelets (Moscow: Mir)
[5] Lepik Ü and Hein H 2014 Haar Wavelets: with applications (Switzerland: Springer International Publishing)
[6] Nagornov O.V. 2010 Weevlet analysis in examples (Moscow: National Research Nuclear University MEPhI)
[7] Timofeev V S and Ioseva E V 2019 On the estimation of the distribution density function of a random variable using wavelets. Nauchnyi vestnik Novosibirskogo gosudarstvenogo tekhnicheskogo universiteta – Science bulletin of Novosibirsk state technical university 4 (77) 71–84
[8] Gevorkyan P S 2016 Probability Theory and Mathematical Statistics (Moscow: Fizmatlit)