The Reliability of Numerical Modeling in Remote Sensing Data Analysis

Boris Dobronets[1,*] and *Olga* Popova[1,**]

[1]Institute of Space and Information Technology, Siberian Federal University, Kirenskogo 26, Krasnoyarsk, 660074 Russia

Abstract. The article deals with the problem of calculating reliable estimates of empirical distribution functions under conditions of small sample and data uncertainty. To study these issues, we develope computational probabilistic analysis as a new area in computational statistics. We propose a new approach based on random interpolation polynomials and order statistics. Arithmetic operations on probability density functions and procedures for constructing the probabilistic extensions are used.

1 Introduction

The presence of uncertainties in the remote sensing data requires the development of numerical methods that take into account these uncertainties. Thus, interval uncertainty lead to interval methods. The interval approach is actually one of the most important, but far from the only means of obtaining reliable results in mathematical computations.

Reliability can also be based on other approaches, both purely mathematical and related to computer tools for solving mathematical problems. To improve the reliability of calculations we propose to use the computational probabilistic analysis.

The paper discusses reliable estimates of remote sensing data analysis. The approach is based on the use of random interpolation polynomials and order statistics. One of the known approaches is the use of Kolmogorov-Smirnov confidence limits. Similar methods are used to construct the interval boundaries of empirical distribution functions or so called P-Boxes [1].

Information availability on the probability density function makes it possible to take into account the influence of data uncertainty in the calculations and to obtain results in the form of random variables with a constructed probability density. One approach to accounting for the random nature of the input data is Monte Carlo method [2].

With all its positive qualities, this method has several disadvantages. One of the most significant drawbacks is the low convergence rate. It is important that many practical tasks with random inputs require faster methods. Computational probabilistic analysis is one of these approaches. The main idea of computational probabilistic analysis is to use numerical operations and relations over probability densities.

In computational probabilistic analysis, various types of representations of the density function of random variables are used piecewise polynomial function. Piecewise polynomial

[*]e-mail: BDobronets@yandex.ru
[**]e-mail: OlgaArc@yandex.ru

functions are determined by grids of dimension m and the values of the functions at the grid nodes. Histograms, frequency polygons, splines, etc. are examples of such functions [4–6].

The paper considers probabilistic extensions of interpolation polynomials and splines in the case when the input data are random variables given by their probability densities.

The probability density functions of random variables x, y, z will be denoted by bold font $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$. Let us identify through $\boldsymbol{R}$ the set of all probability density functions.

The interpolation problem can be formulated as follows [3]. Let f be some function, its values $f_i = f(x_i)$ at the points $a = x_0 < x_1 < x_2 ... < x_n = b$ are random variables f_i with joint probability density function $\boldsymbol{p}(f_0, f_1, \dots, f_n)$. The problem arises of approximate recovery of all realizations of the function f at an arbitrary point x.

Further, this problem will be solved be using computational probabilistic analysis and applying the concept of probabilistic extension. For these purposes, we will construct probabilistic extensions of Lagrange interpolation polynomials, piecewise linear functions and cubic splines.

2 Elements of computational probabilistic analysis

Let $(x_1, x_2, \dots, x_n)$ be a system of continuous random variables with joint probability density functions $\boldsymbol{p}(x_1, x_2, \dots, x_n)$. Let the random variable z be a function

$$z = f(x_1, x_2, \dots, x_n),$$

where $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 1 *We say that the random function $\boldsymbol{f} : \boldsymbol{R}^n \to \boldsymbol{R}$ is a probabilistic continuation of the deterministic function $f : \mathbb{R}^n \to \mathbb{R}$ on the set $D \subset \mathbb{R}^n$, if $\boldsymbol{f}(x) = f(x)$ for all arguments $x \in D$.*

Definition 2 *The random function $\boldsymbol{f} : \boldsymbol{R}^n \to \boldsymbol{R}$ is called the probabilistic extension of the deterministic function $f : \mathbb{R}^n \to \mathbb{R}$ on the set $D \subset \mathbb{R}^n$, if*

(i) it is probabilistic continuation of f on D,

(ii) the probability density function $\boldsymbol{f}$ coincides with the probability density function $\boldsymbol{z}$ of the random variable $z = f(x_1, x_2, \dots, x_n)$, where $(x_1, x_2, \dots, x_n)$ is a system of continuous random variables with joint probability density functions $\boldsymbol{p}(x_1, x_2, \dots, x_n)$.

Consequently, we can write

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n).$$

If at some point ξ it is necessary to directly indicate the value of the probability density function $\boldsymbol{f}$, we will use the notation

$$\boldsymbol{z}(\xi) = \boldsymbol{f}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)(\xi).$$

The support of a continuous function f is the closure of the set $\{x \mid f(x) \neq 0\}$ and it is denoted by the symbol supp(f).

Assuming $f(x_1, x_2) = x_1 * x_2$, where $* \in \{+, -, \cdot, /\}$, we can obtain analytic formulas for determining the probability densities of the arithmetic operation results for random variables.

Let $f(x_1, \dots, x_n)$ be a rational function. We can obtained probabilistic extension $\boldsymbol{f}$ of real rational functions f by replacing (i) the real variables x_1, x_2, …, x_n with an probability density functions $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_n$ and (ii) the real arithmetic operations with corresponding probabilistic operations. The result $\boldsymbol{f}$ is called a *natural probabilistic extension* [8].

Theorem 1 ([8]) *Let $x_1, \dots, x_n$ be independent random variables. If $f(t_1, \dots, t_n)$ is a rational expression where each variable t_i occurs not more than once, then the natural probabilistic extension approximates a probabilistic extension.*

Theorem 2 ([9]) *Let $\boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$ be probabilistic extension of function $f(x_1, x_2, \dots, x_n)$ and for all real t function $\boldsymbol{f}(t, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$ be probabilistic extension of the function $f(t, x_2, \dots, x_n)$. Then*

$$\boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)(\xi) = \int_{\text{supp}(\boldsymbol{x}_1)} \boldsymbol{x}_1(t)\boldsymbol{f}(t, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)(\xi)dt. \tag{1}$$

Corollary 1 ([9]) *Theorem 2 implies the possibility of recursive computations for the general form of probability extensions and reduction to the calculation of the one-dimensional case.*

Consider example

$$\boldsymbol{z} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + ... + a_n\boldsymbol{x}_n. \tag{2}$$

If the random variables are independent, we can calculate the value of (2) using numerical probabilistic arithmetic sequentially calculating piecewise polynomial approximations. To calculate one addition, we need Cm^2 polynomial calculations. Accordingly, the total number of calculations will be equal to Cnm^2.

Thus, the article [7] compares the number of generate random variables of Monte Carlo methods and numerical operations using probabilistic arithmetic to achieve the same accuracy. Thus, the accuracy of addition of n uniform random variables ($m = 10$) is achieved by Monte Carlo methods with the number of generate uniform random variables equal to $n \cdot 10^6$. In the case of dependent random variables, according to 1 and [9], the number of operations increases as m^n.

In the general case, Monte Carlo methods are used [2]. To overcome the shortcomings associated with the Monte Carlo methods, we will use a new approach, computational probabilistic analysis. This allows in some cases to calculate integrals of the form (1) with the required accuracy.

Next we will use random functions in the form of linear combinations

$$\boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i(x), \quad \text{where} \quad g_i \in C^m[a, b].$$

For this type of random functions, we introduce the following concepts. Then the formal derivative of $\boldsymbol{f}(x)$ is defined as follows:

$$\partial^k \boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i^{(k)}(x), \quad k = 0, ..., m,$$

Integral of $\boldsymbol{f}(x)$ is

$$\int_a^b \boldsymbol{f}(x)dx = \sum_{i=1}^{n} \boldsymbol{a}_i \int_a^b g_i(x)dx.$$

Function

$$f(x) = \sum_{i=1}^{n} a_i g_i(x), \quad \text{where} \quad a_i \in \text{supp}(\boldsymbol{a}_i)$$

we will call *the constriction of the function* $\boldsymbol{f}$ by constants $\boldsymbol{a}_i$.

3 Interpolation problems

The interpolation problem is formulated as follows. Let the probability densities $\boldsymbol{f}_i$ be known at the points $a = x_0 < x_1 < x_2 ... < x_n = b$ and their joint probability function $\boldsymbol{p}(f_0, f_1, \dots, f_n)$ is given. We need to build a random polynomial interpolation $\boldsymbol{l}_n(x)$: $\boldsymbol{l}_n(x_i) = \boldsymbol{f}_i$.

We consider Lagrange interpolation polynomials for the case of linear interpolation. Let f_1, f_2 be known values of some function f at the points x_1, x_2. In the case of linear interpolation we obtain the exact equality

$$f(x) = l_1(x) + \frac{(x - x_1)(x - x_2)}{2} f''(\xi), \ \xi \in [x_1, x_2],$$

where l_1 is the first-degree Lagrange polynomial

$$l_1(x) = f_1 \frac{x_2 - x}{x_2 - x_1} + f_2 \frac{x - x_1}{x_2 - x_1}.$$

If the values of $f_1 \in \mathrm{supp}(\boldsymbol{f}_1)$, $f_2 \in \mathrm{supp}(\boldsymbol{f}_2)$ are not exactly known, it is necessary to construct a linear random function $\boldsymbol{l}(x)$ satisfying the interpolation conditions $\boldsymbol{l}(x_1) = \boldsymbol{f}_1$ and $\boldsymbol{l}(x_2) = \boldsymbol{f}_2$. Thus, using natural probabilistic extensions, we construct random Lagrange polynomials of the first degree

$$\boldsymbol{l}(x) = \boldsymbol{f}_1 \frac{x_2 - x}{x_2 - x_1} + \boldsymbol{f}_2 \frac{x - x_1}{x_2 - x_1}.$$

The interpolation function $\boldsymbol{l}(x)$ is equal to the given values at the interpolation nodes. It is important that the constriction of a random linear function with respect to the constants $\boldsymbol{f}_i$ is a real linear function. Further, if it is necessary to construct a random function $\boldsymbol{l}$ satisfying the inclusion $f \in \mathrm{supp}(\boldsymbol{l})$, for all $x \in [x_1, x_2]$ then you will need to have a priori information about the probability density of $f'' \in \mathrm{supp}(\boldsymbol{f}'')$ on the interval $[x_1, x_2]$. Then we can get an estimate

$$f(x) \in \mathrm{supp}\left(\boldsymbol{l}(x) + \frac{(x - x_1)(x - x_2)}{2} \boldsymbol{f}''\right).$$

Next, we consider the general case for the Lagrange interpolations polynomial. We have

$$\boldsymbol{l}_n(x) = \sum_{i=0}^{n} \boldsymbol{f}_i \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)}.$$

Thus the calculation of Lagrange interpolation polynomial at an arbitrary point reduces to calculating the sum of $\boldsymbol{f}_i$ with weights. If the random variables $\boldsymbol{f}_i$ are independent the calculations are simple because they fall under the conditions of Theorem 1.

For the number of nodes $n \gtrapprox 5$, the application of the Lagrange interpolation polynomials is not effective. In this case you can use piecewise linear interpolation defined by the first degree Lagrange polynomial on each interval $[x_i, x_{i+1}]$.

$$\boldsymbol{l}_1(x) = \boldsymbol{f}_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + \boldsymbol{f}_{i+1} \frac{x - x_i}{x_{i+1} - x_i}. \tag{3}$$

Let us estimate the mathematical expectation of the Lagrange interpolation polynomials. In accordance with the linearity property, the expectation of the interpolation polynomial will be a linear combination of the expectations of the function values. It will coincide with the Lagrange interpolation polynomial constructed from the expectation values of the function:

$$\mathsf{E}[\boldsymbol{l}_1(x)] = \mathsf{E}[\boldsymbol{f}_i] \frac{x_{i+1} - x}{x_{i+1} - x_i} + \mathsf{E}[\boldsymbol{f}_{i+1}] \frac{x - x_i}{x_{i+1} - x_i}.$$

If, for the mathematical expectation of the random function $\boldsymbol{f}$, we are estimates of the second derivative $\max_{x \in [a,b]} |\mathsf{E}[\boldsymbol{f}^{(2)}]|$, then following theorem takes place:

Theorem 3 *[[3]] Let $\boldsymbol{l}_1$ be a piecewise linear interpolation of the random function $\boldsymbol{f}$. Then we have the estimate*

$$|\mathsf{E}[\boldsymbol{l}_1] - \mathsf{E}[\boldsymbol{f}]| \leq Kh^2 \max_{x \in [a,b]} |\mathsf{E}[\boldsymbol{f}^{(2)}]|,$$

where K is a constant independent of h.

Consider the properties of a variance for piecewise linear interpolation $\boldsymbol{l}_1$ of a random function $\boldsymbol{f}$.

Theorem 4 ([3]) *The variance of piecewise linear interpolation $\boldsymbol{l}_1$ satisfies following estimate*

$$\mathsf{Var}[\boldsymbol{L}_1] \leq \max_{0 \leq i \leq n-1} \{\max\{\mathsf{Var}[\boldsymbol{F}_i], \mathsf{Var}[\boldsymbol{F}_{i+1}], K_{i,i+1}\}\}.$$

4 Reliable approximation of the distribution function

The section discusses the construction of reliable approximation of the empirical distribution function.

Let $x_1, \dots, x_n$ be a sample of a random variable x with the distribution function $F(t), t \in [a, b]$. The empirical distribution function is defined as follows

$$F_n(t) = \frac{m_t}{n}, \tag{4}$$

where m_t is the number of $x_i < t$.

Consider $z_i = F(x_i), i = 1, \dots, n$. Notice, $z_i, i = 1, \dots, n$ are uniformly distributed random variables on [0, 1]. If $z_1 \leq z_2 \leq \dots \leq z_n$, then z_k is the kth order statistic and its expectation is equal $\mathsf{E}[z_k] = k/(n+1)$ [10].

Further, we will use the points $(x_i, i/(n+1))$ to construct an approximation of the distribution function $F(t)$. Suppose that $\omega = \{a = x_0 < x_1 < x_2 < \dots < x_n < b = x_{n+1}\}$ is a grid. Then we construct a piecewise linear function $l(t), t \in [a, b]$ and

$$l(x_i) = i/(n+1), \quad i = 1, \dots, n, \ l(a) = 0, \ l(b) = 1.$$

Note, if instead of mathematical expectations $i/(n+1)$ if used exact values z_i, then the error of the piecewise-linear function $l(t)$ with the step $h = \max_{0 \leq i \leq n-1}(x_{i+1} - x_i)$ would satisfy the estimate

$$\|F - l\|_\infty \leq Kh^2 \|F^{(2)}\|_\infty.$$

Hence, the constructed piecewise-linear function fairly well approximate the distribution function F even for relatively small n.

As for z_i, we are aware that they form the order statistics. It is known that the probability density of the kth order statistic is (see [10])

$$p_k(z) = \frac{n!}{(n-k)!(k-1)!} z^{k-1}(1-z)^{n-k}, \quad z \in [0, 1]. \tag{5}$$

The joint probability density for the vector (z_j, z_k) has the form (see also [10])

$$p_{j,k}(z_j, z_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} z_j^{j-1}(z_k - z_j)^{k-j-1}(1 - z_k)^{n-k}, \tag{6}$$

$$j < k, \quad 0 \leq z_j \leq z_k \leq 1.$$

For each random vector $(z_1, z_2, ..., z_n)$ we have the corresponding piecewise linear function l. Looking through all possible random vectors $(z_1, z_2, ..., z_n)$, we get the whole set of piecewise linear functions $\{l\}$. Note that $\{l\}$ contains the interpolant of the distribution function F. Hence, using the probability density of the kth order statistic for a node ξ_k, the set $\{l\}$ can be represented as a random piecewise linear function $\boldsymbol{l}$. Accordingly, $\boldsymbol{l}$ is a reliable approximation of the empirical distribution function.

5 Conclusion

We have discussed the problem of calculating reliable estimates for empirical distribution functions in a small sample. We have identified modeling and algorithmic advances necessary for success on calculation problems. For these purposes, random interpolation polynomials and order statistics are used.

We also have proposed the arithmetic operations on probability density functions and procedures for constructing probabilistic extensions as the basis of computational probabilistic analysis [3–9].

The application of the developed procedures allows us to get knowledge not only about the location of the solution area, but also to identify its probabilistic structure. There are other related issues that we can tackle: for example, the use of this approach to estimate the parameters of technical systems.

References

[1] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, L. Ginzburg, Tech. Rep. 2007-0939, Sandia National Laboratories (2007).

[2] G. S. Fishman, Monte Carlo: concepts, algorithms, and applications, Springer, New York, 1996.

[3] O. Popova, Computational Technologies 22 (2) (2017) 99–114. (In Russian)

[4] B. Dobronets, O. Popova, Journal of Siberian Federal University, Mathematics and Physics 10 (1) (2017) 16–21.

[5] B. Dobronets, O. Popova, IOP Conf. Series: Journal of Physics: Conf. Series 1015 (032028). doi:10.1088/1742-6596/1015/3/032028.

[6] B. Dobronets, O. Popova, IOP Conf. Series: Materials Science and Engineering 354 (012006). doi:10.1088/1757-899X/354/1/012006.

[7] B. Dobronets, A. Krantsevich, N. Krantsevich, Siberian Federal University. Mathematics and Physics 6 (2) (2013) 168–173.

[8] B. Dobronets, O. Popova, Reliable Computing 19 (2014) 274–289.

[9] B. Dobronets, O. Popova, Tomsk State University Journal of Control and Computer Science 47 (2019) 41–48.

[10] H. David, Order Statistics, John Wiley, New York, 1981.