Modeling Under Uncertainty: a Comparison of Approaches

B S Dobronets¹ and O A Popova¹

¹Institute of Space and Information Technology, Siberian Federal University, Kirenskogo 26, Krasnoyarsk, 660074 Russia
E-mail: BDobronets@yandex.ru

Abstract. A new approach to data processing and uncertainty modeling based on the use of computational probabilistic analysis is considered. The basis of computational probabilistic analysis is numerical operations on probability density functions represented by piecewise polynomial functions. The problems of predicting the time series of distributions and estimating the probability densities of solutions of boundary value problems and systems of nonlinear equations with random coefficients are considered. Interval analysis, functional data analysis, symbolic data analysis and Monte Carlo method are currently used to study such data. A comparison of these approaches is given. Numerical examples show the effectiveness of the proposed approaches.

1. Introduction

The development, production and operation of complex systems under the condition of uncertainty is the driving force behind the making of new numerical simulation algorithms. The presence of various types of uncertainty in the data significantly affects the results of numerical modeling and causes problems associated with the accuracy of the results.

The need for fast and reliable computing requires the development of new approaches for processing incoming information. These approaches are based not only on increasing the speed of computers, but also on new software. Therefore, the problem of numerical modeling under conditions of uncertainty becomes very relevant.

There are several approaches implemented in the framework of numerical modeling and analysis of complex systems with input parameters containing uncertainties of various types. One uncertainty type is interval uncertainty. Interval uncertainty contributed to the development of interval analysis. Random uncertainty deals with random variables and processes. Knowledge of the distribution laws of random variables allows us to estimate the parameters of stochastic systems using the Monte Carlo method. Fuzzy set theory is widely used to model systems and make decisions.

At present, computational probabilistic analysis is used for problems under stochastic uncertainty [1, 2, 3, 4]. In some cases, it successfully replaces the Monte Carlo method, having a significantly higher rate of convergence. Unlike the Monte Carlo method, it allows one to successfully evaluate not only the parameters of systems, such as the mathematical expectation or variance of the output variables, but also to construct the probability densities of the output variables. This significantly increases the reliability of the obtained numerical solutions.
Interval analysis focuses on the calculation of guaranteed bounds of solution sets. Further generalizations of interval analysis are aimed at obtaining more detailed characteristics of the resulting solution sets [4].

Information availability of 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 the Monte Carlo method [5].

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 function.

In computational probabilistic analysis, piecewise polynomial functions are used to represent various probability density functions. Piecewise polynomial 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 [1, 2, 3].

The probability density functions of random variables $x, y, z$ will be denoted by bold type $\mathbf{x}, \mathbf{y}, \mathbf{z}$.

Let us identify through $\mathbf{R}$ the set of all probability density functions.

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

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

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

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

**Definition 2** The random function $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f$ coincides with the probability density function $z$ of the random variable $z = f(x_1, x_2, \ldots, x_n)$, where $(x_1, x_2, \ldots, x_n)$ is a system of continuous random variables with the joint probability density functions $p(x_1, x_2, \ldots, x_n)$.

Consequently, we can write

$$z = f(x_1, \ldots, x_n).$$

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

$$z(\xi) = f(x_1, \ldots, x_n)(\xi).$$

Let $f(x_1, \ldots, x_n)$ be a rational function. We can obtained probabilistic extension $f$ of real rational functions $f$ by replacing (i) the real variables $x_1, x_2, \ldots, x_n$ with an probability density functions $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and (ii) the real arithmetic operations with corresponding probabilistic operations. The result $f$ is called a natural probabilistic extension [6].

**Theorem 1.1** ([6]) Let $x_1, \ldots, x_n$ be independent random variables. If $f(t_1, \ldots, t_n)$ is a rational expression where each variable $t_i$ occurs not more than once, then the natural probabilistic extension coincides with the probabilistic extension.
Theorem 1.2 ([7]) Let \( x_1, \ldots, x_n \) be independent random variables, \( f(x_1, x_2, \ldots, x_n) \) be a probabilistic extension of a function \( f(x_1, x_2, \ldots, x_n) \) and for all real \( t \) a function \( f(t, x_2, \ldots, x_n) \) be probabilistic extension of the function \( f(t, x_2, \ldots, x_n) \). Then

\[
f(x_1, x_2, \ldots, x_n)(\xi) = \int_{\text{supp}(x_1)} x_1(t)f(t, x_2, \ldots, x_n)(\xi)dt. \tag{1}
\]

Corollary 1 ([7]) Theorem 1.2 implies the possibility of recursive computations for the general form of probability extensions and, eventually, reduction to the calculation of the one-dimensional case.

Consider example

\[
z = a_1x_1 + a_2x_2 + \ldots + a_nx_n. \tag{2}
\]

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 \( Cm^3 \) polynomial calculations. Accordingly, the total number of calculations will be equal to \( Cnm^2 \).

The article [8] compares the number of generated 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 generated uniform random variables equal to \( n \cdot 10^6 \).

In the case of dependent random variables, according to Corollary 1 and [7], the number of operations increases as \( m^n \). In the general case, Monte Carlo methods are used [5]. 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.

2. Distributional Time Series

Distributional Time Series (DTS) are covered in Functional Data analysis [9] and Symbolic Data Analysis [10]. The DTS analysis in functional data analysis is based on the use of functional principal components. In symbolic data analysis, DTS are presented as histogram-valued Time Series [11]. Distributional time series are used to analyze processes such as: individual earnings, global temperatures, household income, household expenditures, stock returns [12, 13].

Consider the time series of the distributions \( y_i \), \( i = 1, 2, \ldots, n \), where the probability density functions \( y_i \) are piecewise polynomial functions with supports \([a_i, b_i]\). We will build the grids based on the features of the probability density function: we will take into account the boundaries of carriers, points of local extrema, etc

\[
\omega_i = \{x_i^0 = a_i, \ldots, x_i^m = b_i\}.
\]

On the sets \( X_j = \{x_j^i, i = 1, 2, \ldots, n\} \), we construct the regression curves \( r_j(t) \). Further, let \( S(t) \) be the transformation of the set \( \{r_j(t), j = 0, 1, \ldots, m\} \) into \( \omega = \{x_0 = a, \ldots, x_m = b\} \).

Piecewise polynomial functions \( y_i \) and grids \( \omega_i \) we transform using \( S(t) \) in the function \( s_i \) on the grid \( \omega \). Consider, for example, the use of the fifth degree Hermitian splines. Then for each node \( x_l \in \omega \) the values \( s_{li}, s'_{li}, s''_{li} \) are known and we can construct the regression curves \( s_l(t), s'_l(t), s''_l(t) \). Using these regression curves, it is possible to construct piecewise polynomial functions \( s(t) \) on the grid \( \omega \) for any \( t \).

Then, using the inverse transformation \( S^{-1}(t) \), construct the value \( \tilde{y}(t) \) for any \( t \). Note that in this case it is necessary to check the conditions

\[
\tilde{y}(t)(\xi) \geq 0, \quad \xi \in [r_0(t), r_m(t)],
\]
\[ \int_{r_0(t)}^{r_m(t)} \bar{y}(t)(\xi) d\xi = 1. \]

This example reviews the data of temperature in the city of Krasnoyarsk over the past 70 years. Then we build a regression of temperature densities by days. In order to build the temperature distribution on April 15, we sampled a temperature for all 70 years on April 15. Next, using kernel estimators and quintic Hermite spline, we constructed a piecewise polynomial approximation of the temperature distribution. Figure 6 shows the distributional time series of temperature by days from 15.04 to 15.06.

Temperature data in the city of Krasnoyarsk has been aggregated by days for the past 70 years. Lines 1 are piecewise polynomial approximations of temperature probability density functions 15.04, 1.05, 15.05, 1.06, 15.06. Using empirical probability densities for April 15, May 1 and 15, June 1, a distributional time series represented by lines 2 was constructed. Lines 3 are the smoothed boundaries of supports and maxima of the probability density functions. Regression of distributions is predicted for June 7 and 15.

Figure 1 shows a comparison of the empirical density function and the forecast density function for June 15th.

\[ \text{Figure 1. Forecasting daily temperature time series distributions} \]

3. Boundary value problems with random coefficients

This section discusses the use of computational probabilistic analysis to solve boundary value problems of ordinary differential equations with random coefficients [14].

Consider the boundary value problem

\[ Lu = -pu'' + qu = f(x), x \in (0, 1), \quad \text{(3)} \]

with boundary conditions

\[ u(0) = 0, \quad u(1) = 0. \]

where \( p, q \) are independent random constants, \( p > 0, q \geq 0. \)

Suppose \( \omega_h = \{ x_i = ih, i = 1, 2, ..., N - 1, h = 1/N \} \) is a mesh and

\[ L^h u_i^h = -p \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + qu_i = f(x_i), i = 1, 2, ..., N - 1. \]

is a difference scheme. Further assume that \([\underline{p}, \overline{p}], [\underline{q}, \overline{q}]\]\ are supports \( p, q \). Finally construct the meshes \( \omega_p = \{ p_0 = \underline{p} < p_1 < ... < p_K = \overline{p} \} \), and \( \omega_q = \{ q_0 = \underline{q} < q_1 < ... < q_L = \overline{q} \} \).
Let us numerically solve $KL$ of the tasks

$$-p_k \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + qu_i = f(x_i), i = 1, 2, \ldots, N - 1.$$ 

Thus, we get an array of solutions $u_{ikl} = u_i(p_k, q_l)$. Consider the construction of the probability density function of the solution $u_i$. For these purposes, using the values $u_i(p_k, q_l)$, we construct Hermitian cubic splines $s_l(p_l)$, $l = 0, 1, \ldots, 10$. In figure 2 lines 1 represent Hermitian cubic splines $s_l$. In order to calculate the value of the probability density function of the solution $u_i$ at some point $\xi$ we find the roots of $p_l$

$$s_l(p_l) = \xi, l = 0, 1, \ldots, 10.$$ 

The values of $u_i(\xi)$ are calculated using numerical quadratures, for example, Simpson quadratures

$$u_i(\xi) = h \sum_{k=0}^{10} \gamma_k q_k p(p_k)/s'(p_k).$$

Figure 3 shows the integration line.

Note that the Simpson quadrature and cubic splines have accuracy $O(h^4)$. Thus, a piecewise polynomial approximation of $u_i$ can be constructed by calculating the values of the probability density function $u_i(\xi)$ for different values $\xi$.

Figure 2 presents a solution to a boundary value problem with random coefficients. Lines 1 are the probability density functions of the solutions of the boundary value problem at the grid nodes of the difference scheme, lines 2 show the boundaries of the support of the solution of the boundary value problem.

Numerical experiments at $K, L = 10$ showed good agreement with the Monte Carlo method with the number of casts $10^6$. In this case, CPA is about $10^4$ times more efficient than the Monte Carlo method.

As showed above, this example proves that the main computational costs include the array constructing $u_{ikl}$, which corresponds to the number of operations $\sim O(KLN)$.

To calculate the values of the probability density function $u_i(\xi)$ we need approximately $\sim O(m)$. Thus, by constructing the array $u_{ikl}$ one can relatively quickly find the probability density functions of the solution $u_i$ for various $p, q$. 

**Figure 2.** Probability density functions for solutions of a boundary value problem

Let us numerically solve $KL$ of the tasks
4. Systems of nonlinear equations

Consider the problem of determining the set of solutions of a system of nonlinear equations

\[ f_i(x, k) = 0, \ i = 1, \ldots, n, \]

where \( x \in \mathbb{R}^n \) is solution vector, \( k \in \mathbb{R}^m \) is parameter vector. Regarding \( k \in \mathbb{R}^m \), we assume that the probability density functions are known.

Consider the case when \( m = n \). Note that, using the results of theorem 1.2, the case of the strict inequality \( m > n \) can be reduced to the case of \( m = n \).

Consider an example of a system of nonlinear equations

\[
\begin{align*}
x^2 + y^2 - r^2 &= 0, \\
x y - c &= 0,
\end{align*}
\]

where \( r, c \) are uniform random variables whose probability densities have the supports \([1, 1.1], [0.4, 0.5]\).

Suppose that for some values \( r_0, c_0 \) the solution of the system is \( x_0, y_0 \), then

\[
\begin{pmatrix}
2x_0 & 2y_0 \\
y_0 & x_0
\end{pmatrix}
\begin{pmatrix}
dx \\
dy
\end{pmatrix}
= \begin{pmatrix}
2r_0dr \\
dc
\end{pmatrix}.
\]

In this case, a rectangle \( S_0 \) with the sides \( dr, dc \) is passed on a quadrangle with the sides \( dx, dy \).

Having solved the system, we obtain the values \( dx, dy \).

The probability density of the set of solutions at the point \( (x_0, y_0) \) is proportional to the area ratio \( |S_0|/|S_1| \):

\[ p(x_0, y_0) = p_1(r_0, c_0)|S_0|/|S_1|. \]

Thus, to construct a joint density function of random variables \( x, y \), we construct grids in the region of supports of random variables \( r, c \): \( \{r_i, i := 0, \ldots, m\}, \{c_i, i := 0, \ldots, m\} \).

Solve \((m + 1)^2\) systems of nonlinear equations

\[ x^2 + y^2 - r_i^2 = 0, \]
The figure 4 shows the triangulation of the solution domain of a system of nonlinear equations. The vertices of the triangles are the solution points with the calculated values \( p(x_i, y_j) \) of the joint probability density function of the solution.

In this example, to construct a joint solution density function, it is necessary to solve a system of nonlinear equations only 25 times, the Monte Carlo method will require about \( 10^6 \) solutions. The general case of constructing probability densities can be implemented using the construction of probability extensions.

Thus, computational probabilistic analysis allows us to solve some modeling problems under uncertainty with a relatively small number of operations.

References

[1] Dobronets B and Popova O 2017 Improving the accuracy of the probability density function estimation Journal of Siberian Federal University, Mathematics and Physics 10(1) 16–21
[2] Dobronets B and Popova O 2018 Piecewise polynomial aggregation as preprocessing for data numerical modeling, IOP Conf. Series: Journal of Physics: Conf. Series 1015 032028
[3] Dobronets B and Popova O 2018 Improving reliability of aggregation, numerical simulation and analysis of complex systems by empirical data IOP Conf. Series: Materials Science and Engineering 354 012006
[4] Popova O 2017 Application of numerical probabilistic analysis in interpolation problems Computational Technologies 22 99–114 (In Russian)
[5] Mikhailov G A and Voytishek A V 2018 Statistical Modeling. Monte Carlo Methods (Moscow: Yurayt Publishing House) (In Russian)
[6] Dobronets B and Popova O 2014 Numerical probabilistic analysis under aleatory and epistemic uncertainty. Reliable Computing, 19 274–289
[7] Dobronets B and Popova O 2019 Computational aspects of probabilistic extensions. Tomsk State University Journal of Control and Computer Science, 47 41–48
[8] Dobronets B S, Krantsevich A M and Krantsevich N M 2013 Software implementation of numerical operations on random variables Siberian Federal University, Mathematics & Physics, 6(2) 168–173
[9] Ramsay J and Silverman B 2005 Functional Data Analysis (Springer 2nd ed., New York)
[10] Billard L and Diday E 2006 Symbolic Data Analysis: Conceptual Statistics and Data Mining (West Sussex PO19 8SQ, England : John Wiley & Sons)
[11] Arroyo J Gonzalez-Rivera G Maté C 2011 Smoothing Methods for Histogram-valued Time Series: An Application to Value-at-Risk Statistical Analysis and Data Mining, 4 216–228
[12] Chang Y Kim C and Park J 2016 Nonstationarity in time series of state densities Journal of Econometrics 192 152–167
[13] Popova O, Dobronets B and Merko A 2020 Distributional Time Series For Forecasting And Risk Assessment
Modeling and Analysis of Safety and Risk in Complex Systems: Proceedings of the Sixteenth International
Scientific Conference MASR — 2020 (Saint-Petersburg, Russia) pp 122–127

[14] Soong T 1973 Random Differential Equations in Science and Engineering (New York and London: Academic
Press)