УДК 519.24

Software Implementation of Numerical Operations on Random Variables

Boris S. Dobronets*
Institute of Space and Information Technology,
Siberian Federal University,
Kirenskogo, 26, Krasnoyarsk, 660074
Russia
Artem M. Krantsevich†
Nikolay M. Krantsevich‡
Institute of Mathematics and Computer Science,
Siberian Federal University,
Svobodny, 79, Krasnoyarsk, 660041
Russia

Received 23.11.2012, received in revised form 26.12.2012, accepted 26.01.2013

We consider a software implementation of numerical operations on different types of random variables and introduce algorithms for arithmetic operations on random variables represented by their probability densities. We also estimate the accuracy of these algorithms and compare the accuracy of histogram arithmetic and the Monte Carlo methods.

Keywords: numerical operations on random variables, histogram arithmetic, second order histogram, Monte Carlo method, interval mathematics.

Introduction

Statistical methods are being increasingly used in a wide range of applications. The presence of uncertainties in the input data of many practical problems motivates the need for methods that take them into account.

Analytical methods of probabilistic analysis are limited and can not be used for most applications. Monte Carlo methods is a powerful and versatile approach, which is widely used for stochastic modeling. Despite its strengths, it has some shortcomings; one of the most serious is its slow convergence.

An alternative approach, may be, in some cases, interval analysis [1], approaches relying on numerical operations on random variables [2–5], numerical methods of probabilistic analysis [6,7].

The packages that numerically implement histogram arithmetic [2,3,5] have some drawbacks, because arithmetic operations on random variables largely employ the Cartesian product of subintervals, which significantly affects the accuracy of the results.

1. The probability density function types

In this section we list different types of probability density functions of random variables, on which we consider arithmetic operations.

*BDobronets@sfu-kras.ru
†akrantsevich@gmail.com
‡krantsevich@gmail.com
© Siberian Federal University. All rights reserved

Discrete random variables. A discrete random variable ξ assumes values $x_1, x_2, ..., x_n$, each with probability $p(x_i)$. The function $p(x)$ is sometimes called the probability function or the probability density.

Histograms. A histogram is a random variable whose density function is piecewise constant. The histogram P is defined by the grid $\{x_i | i = 0, ..., n\}$; on the interval $[x_{i-1}, x_i]$, $i = 1, \ldots, n$ the histogram takes constant value p_i.

Interval histograms. In most applications it is impossible to obtain an accurate probability density function, then we estimate it from below and above. Such estimates are usually approximated by intervals. A random variable is called an interval histogram if its probability density function $P(x)$ is a piecewise-interval function.

Second order histograms. In the case of epistemic uncertainty *second order histograms* are also used, along with interval histograms. The probability density function $P(x)$ of a second order histogram is a piecewise-histogram function, i.e., a histogram such that each column of it is again a histogram [8].

Piecewise linear functions. Piecewise linear functions also can be considered as a tool for approximation of the density function of a random variable. A piecewise linear function is a continuous function that is linear on each segment $[x_{i-1}, x_i]$, $i = 1, \ldots, n$.

Splines. A spline is a sufficiently smooth piecewise-polynomial function. We consider random variables whose probability density functions are approximated by splines.

Analytically given probability density. Random variables with probability density given analytically.

2. Operations on probability densities of random variables

Here we consider the arithmetic operations on probability density functions of different kinds.

Operations on discrete values. Let $* \in \{+, -, \cdot, /, \uparrow\}$ be an arithmetic operation of two independent discrete random variables ξ and η. ξ and η take values x_i and y_i with probabilities p_i and q_i, respectively. The result of this arithmetic operation is a random variable ψ that assumes values $x_i * y_j$ with probability $p_i q_j$.

Because "the combinatorial explosion" is possible, it is necessary to transform discrete random variables to other types, e.g., a histogram. For this purpose, the algorithms of [9], for example, can be used.

Operation on histograms. Let $p(x, y)$ be the joint probability density function of two random variables x and y, and p_z be the histogram that approximates the probability density of the arithmetic operation on two random variables $x * y$, where $* \in \{+, -, \cdot, /, \uparrow\}$. Then the probability of z being in the interval $[z_i, z_{i+1}]$ is determined by the formula from [10]

$$P(z_k < z < z_{k+1}) = \int_{\Omega_k} p(x, y) dx dy, \tag{1}$$

where $\Omega_k = \{(x, y) | z_k \leqslant x * y \leqslant z_{k+1}\}$.

The numerical implementation of this method follows. Let the histogram variable x be given by the grid $\{a_i\}$ and probabilities $\{p_i\}$, and the variable y by $\{b_i\}$ and $\{q_i\}$, respectively. Let $[a_0, a_n]$ and $[b_0, b_n]$ be the supports of the probability densities of these variables, and the rectangle $[a_0, a_n] \times [b_0, b_n]$ be the support of the joint probability density $p(x_1, x_2)$. We divide the rectangle $[a_0, a_n] \times [b_0, b_n]$ into n^2 rectangles $[a_i, a_{i+1}] \times [b_j, b_{j+1}]$, and the probability of getting into such is a constant $p_i q_j$ for independent random variables, and p_{ij} for dependent ones.

To compute the required histogram p_z, we walk through all the rectangles $[a_i, a_{i+1}] \times [b_j, b_{j+1}]$; and for each of them we calculate its contribution into each segment $[z_k, z_{k+1}]$ of the resultant histogram. To this end, we consider the region Ω'_k:

$$\Omega'_k = \Omega_k \cap ([a_i, a_{i+1}] \times [b_j, b_{j+1}]),$$

and compute the integral over Ω'_k

$$p_{zk} = \int_{\Omega'_k} p(x,y)dxdy. \tag{2}$$

Note that for each $[a_i, a_{i+1}] \times [b_j, b_{j+1}]$ the joint probability density $p(x,y)$ is constant, therefore this integral equals the ratio of the area of Ω'_k to the area of $[a_i, a_{i+1}] \times [b_j, b_{j+1}]$. Having walked through all the boxes, we compute the histogram p_z. These computations require $O(n^2)$ arithmetic operations.

Operations on a histogram and a discrete random variable. Consider operations of the form $x * c$, where $* \in \{+, -, \cdot, /\}$, c is a constant and x is a random variable with the probability density f_x. If $* \in \{+, -\}$ then the probability density function f_z of the random variable $z = x * c$ can be easily expressed as follows: $f_z(\xi * c) = f_x(\xi)$, $\xi \in R$. Let * be multiplication and $c \neq 0$, then $f_z(\xi) = f_x(\xi/c)/c$. If $c = 0$ then the random variable z takes only one value 0 with probability 1. The operation of division by $c \neq 0$ is performed analogously: $f_z(\xi) = f_x(\xi \cdot c) \cdot c$, $\xi \in R$.

In the case of operations on a discrete and a histogram random variable we consider the segments $A_i \times [b_j, b_{j+1}]$ instead of the rectangles $[a_i, a_{i+1}] \times [b_j, b_{j+1}]$. We proceed analogously to the previous case, walking through all these segments and computing the contribution of each of them into the resultant histogram. The only difference in the numerical implementation is that we compute the ratio of the segments' lengths, not areas. However, if the the number of values assumed by the discrete random variable is large, these calculations can cause a significant increase of time cost. In such a situation, a discrete random variable is represented by a histogram, and then the operation produces a histogram, as described in the preceding subsection.

Operation on a histogram and an analytically given density function. In this case the computation is similar to the case with two histograms. But the joint probability density is not constant, and we need to compute the integrals of the form (2). The result is then a histogram that approximates the density distribution of the resulting random variable.

Operations on the second order histograms. Let X, Y be two second order histograms defined by the grids $\{ v_i, i = 0,1, \ldots n \}$ and $\{ w_i, i = 0,1, \ldots n \}$, and the sets of histograms $\{ Px_i \}$, $\{ Py_i \}$. Let $Z = X * Y$ and $* \in \{+, -, \cdot, /, \uparrow\}$. We compute Z as a second order histogram. Let $\{z_i, i = 0, 1, \ldots n\}$ be a grid, then following (1) the histogram Pz_i on the interval $[z_k, z_{k+1}]$ is determined by the formula

$$Pz_k = \iint_{\Omega_k} X(\xi)Y(\eta)d\xi d\eta/(z_{k+1} - z_k),$$

where $\Omega_k = \{(\xi, \eta) | z_k \leqslant \xi * \eta \leqslant z_{k+1}\}$.

Note that the function $X(\xi)Y(\eta)$ on each rectangle $[v_{i-1}, v_i] \times [w_{j-1}, w_j]$ is a constant histogram $Px_i \cdot Py_j$. The integral of a constant histogram over a region is the value of the histogram multiplied by the area of the region.

3. Procedures

The designed package includes the following modules: **addition, subtraction, addition of a random variable and a number, multiplication, multiplication of a random variable by a number, division, rational exponents of a random variable, the computation of the mean and the dispersion, normalization**.

4. Test. Comparison with Monte Carlo

In order to test the numerical operations on histogram variables, we consider the addition of four random variables uniformly distributed on $[0, 1]$.

Note that the probability density of the sum of n uniformly distributed variables is

$$p_n(x) = \frac{1}{(n-1)!}(x^{n-1} - C_n^1(x-1)^{n-1} + C_n^2(x-2)^{n-1} - ...) \tag{3}$$

where C_n^k are binomial coefficients, and for each fixed value of the argument x the sum in brackets comprises only those terms for which the value of $(x-k), k = 1, 2, ...$ is nonnegative [11].

Thus, when $n = 4$ we have:

$$p(x) = \begin{cases} \dfrac{1}{6}x^3, & \text{if } 0 \leqslant x \leqslant 1; \\ -\dfrac{1}{2}x^3 + 2x^2 - 2x + \dfrac{2}{3}, & \text{if } 1 \leqslant x \leqslant 2; \\ \dfrac{1}{2}x^3 - 4x^2 + 10x - \dfrac{22}{3}, & \text{if } 2 \leqslant x \leqslant 3; \\ -\dfrac{1}{6}x^3 + 2x^2 - 8x + \dfrac{32}{3}, & \text{if } 3 \leqslant x \leqslant 4. \end{cases}$$

Table 1. The errors of histogram arithmetic and Monte Carlo Methods

n	$N = 10^4$	$N = 10^5$	$N = 10^6$	$\|\|H_n - P_n\|\|_2$.
10	0.0059	0.00168	0.00037	4.16e-3
20	0.0055	0.00198	0.00041	5.39e-4
50	0.0026	0.00103	0.00026	3.47e-5
100	0.0023	0.00062	0.00018	4.35e-6
150	0.0016	0.00055	0.00016	1.28e-6
200	0.0014	0.00044	0.00014	5.44e-7

Let N be the number of repetitions, n be the mesh of the grid, H_n the histogram probabilistic extension of p for n (the exact histogram), P_n the natural histogram extension of p for n obtained by performing arithmetic operations, and $MC_{n,N}$ the histogram approximation of p obtained by Monte Carlo method for n, N

The table presents the approximation errors $||H_n - P_n||_2$ and $||H_n - MC_{n,N}||_2$ in ℓ_2 norm for the sum of four uniformly distributed random variables. We note that for a fixed n the error of Monte Carlo method decreases as $\approx 1/\sqrt{N}$, while the rate of convergence for the natural histogram extension is $\alpha \approx 3.5$ [12]. Moreover, the number of operations in histogram arithmetic is $O(n^2)$, and the number of operations for Monte Carlo method is $O(N)$.

Suppose that we want to achieve accuracy ε. The number of operations for Monte Carlo method is $O(\varepsilon^{-2})$ that should be compared to $O(\varepsilon^{-2/\alpha})$ required in histogram arithmetic. Thus, we conclude that the approach relying on the histogram operations is about $\varepsilon^{-2(1-1/\alpha)}$ times more efficient than Monte Carlo methods.

It follows immediately from Table 1 that histogram arithmetic is about 100–1000 times more efficient than Monte Carlo methods.

5. Increase in accuracy

More accurate results can be obtained if the sought probability density function is represented as a piecewise linear function or a spline. This can be achieved in two ways.

The first one is to smooth the resultant histogram. For example, connecting the middle point of the histogram columns we obtain a piecewise linear function that approximates the probability density.

Otherwise, one can determine the values of the probability density function of the operation on two random variables at specific points that are represented by curves on the graph of the joint probability density function of the random variables (these curves are lines for addition, subtraction, and division, and hyperbolas for multiplication). Computing the integral over these curves, we obtain the probability of getting into these specific points, after normalization of the result we construct a piecewise-linear function or a spline. Instead of an integral over a curve we can compute the probability of getting into a strip (as in the histogram case). The strip can be taken to be a sufficiently small neighborhood of the curve.

References

[1] B.S.Dobronets, Interval Mathematics, Krasnoyarsk, KSU, 2004, (in Russian).

[2] D.Berleant, Automatically verified reasoning with both intervals and probability density functions, *Interval Computations*, 1993, no. 2, 48–70

[3] W.Li, J.Hym, Computer arithmetic for probability distribution variables, *Reliability Engineering and System Safety*, **85**(2004).

[4] R.Williamson, T.Downs, Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds, *International Journal of Approximate Reasoning*, 1990, no. 4., 89–158.

[5] V.A.Gerasimov, B.S.Dobronets, M.Yu.Shustrov, Numerical operations of histogram arithmetic and their applications, *Automation and Remote Control*, **52**(1991), no. 2, 208–212.

[6] B.S.Dobronets, O.A.Popova, Numerical Operations on Random Variables and their Application, *Journal of Siberian Federal University. Mathematics & Physics*, **4**(2011), no. 2, 229–239 (in Russian).

[7] B.S.Dobronets, O.A.Popova, Numerical probabilistic analysis and probabilistic extension, Proceedings of the XV International EM'2011 Conference, Oleg Vorobyev, ed., Krasnoyarsk, SFU, RIFS, 2011, 67–69 (in Russian).

[8] B.S.Dobronets, O.A.Popova Histogram time series, Proceedings of the X International FAMES'2011 Conference, Oleg Vorobyev, ed., Krasnoyarsk, RIFS, SFU, KSTEI, 2011, 127–130 (in Russian).

[9] A.V.Kryanev, G.V.Lukin, Mathematical Methods for treatment undefined data, 2nd ed., Rev., Moscow, FIZMATLIT, 2006 (in Russian).

[10] B.V.Gnedenko, A course in the theory of probability, Moscow, Nauka, 1988 (in Russian).

[11] S.P.Shary, Interval analysis or Monte-Carlo methods? *Computational Technologies*, **12**(2007), no. 1, 103–112 (in Russian).

[12] B.S.Dobronets, O.A.Popova, Numerical probabilistic analysis under aleatory and epistemic uncertainty, 15th GAMM-IMACS Iternational Symposium SCAN'12, Book of Abstacts, Novosibirsk, Russia, Institute of Computational Technologies, 2012, 33–34.

Программная реализация операций над случайными величинами

Борис С. Добронец
Артем М. Кранцевич
Николай М. Кранцевич

В статье рассмотрена программная реализация операций над различными видами случайных величин. Представлены алгоритмы арифметических операций над случайными величинами, заданными своими функциями плотности вероятности. Рассмотрены задачи преобразования типов случайных величин. Приведены оценки точности построенных операций. Произведено сравнение точности реализованных операций с методом Монте-Карло.

Ключевые слова: численные операции над случайными величинами, гистограммная арифметика, гистограммы второго порядка, Монте-Карло, интервальная математика.