PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **16** (2004), 55-64

PLISKA
STUDIA MATHEMATICA
BULGARICA

MONTE CARLO METHOD FOR RECONSTRUCTION OF DENSITIES

Sofiya Ivanovska

The present paper considers the problem how to construct the unknown density having N realizations of the random variable using B-splines approximation, least squares method and Monte Carlo method. It is shown that B-splines are appropriate for density modeling. The results from approximation of an unknown density distribution for the considered algorithm are compared with some non-parametric statistical methods such as histogram and kernel density estimation. A large number of numerical experiments are made using Matlab 6.

1. Introduction

Numerical methods of Monte Carlo type are important area in applied mathematics. Monte Carlo methods are a powerful tool for solving many problems in the field of mathematics, physics and engineering. It is known that they provide statistical estimations for a functional of the solution using sample of a certain random variable whose mathematical expectation is equal to the given functional ([2, 5]). Monte Carlo methods become more popular because of their high rate of parallelism.

The density function modeling is a very important task when solving many real problems in ecology, probabilistic theory and physics. For example the problem of developing efficient algorithms for density function modeling is of significant interest in the problem of air pollution transport.

2000 *Mathematics Subject Classification:* 65C05
Key words: Monte Carlo algorithms

The problems of this type arising when it is necessary to find approximately the unknown density of a random variable using a given number of its realizations.

2. The Monte Carlo Method

Consider the problem of approximation of the unknown density function $p(x)$ defined in [a,b] by given N realizations $\{\xi_i\}_{i=1}^N \in [a,b]$ of the random variable ξ with density $p(x)$ ([1]).

Suppose $p(x) \in C^k[a,b]$ where $k \geq 0$ is an integer number. It divides the interval $[a,b]$ into the m subintervals and introduce following set points:

$$\omega_m = \{a = x_0 < x_1 < \ldots < x_m = b\},$$

with step $h = \dfrac{b-a}{m}$. $2k$ new nodes are added to the set ω_m and the result is the set:

$$T_n = \{t_1 < t_2 < \ldots < t_{k+1} = x_0 < \ldots < x_m = t_{n-k} < t_{n-k+1} < \ldots < t_n\}.$$

Consequently $n = 2k + m + 1$.
It holds approximation of $p(x)$ with B-splines of k-th degree:

$$p(x) = \sum_{i=1}^{L} c_i B_{i,k}(x), \quad x \in [a,b], \quad L = n - k - 1$$

with an error $O(h^k)$ where $c_i, i = 1, \ldots, L$ are the approximate coefficients.
The I-th B-spline of k-th degree with nodes $t_i, \ldots, t_{i+k+1}$ is defined as the divided difference of truncated power function $(t-x)_+^k$ with respect to t in the points $t_i, \ldots, t_{i+k+1}$:

$$B_{i,k}(x) = (\cdot - x)_+^k[t_i, \ldots, t_{i+k+1}], \quad i = 1, \ldots, L,$$

where x is fixed point and

$$(t-x)_+^k = \begin{cases} (t-x)^k, & t > x \\ 0, & t \leq x. \end{cases}$$

Each spline is represented as a linear combination of truncated power functions. Therefore $B_{i,k}(x)$ could be represented as follows:

$$B_{i,k}(x) = \sum_{s=i}^{i+k+1} \frac{(t_s - x)_+^k}{\omega'_{i,k}(t_s)},$$

where $\omega_{i,k}(t) = (t - t_i) \dots (t - t_{i+k+1})$.

The advantage when we apply splines for approximation of unknown function is the usage of algebraic polynomials of low degree because the computations with polynomials of high degree are difficult. Consequently the increasing of the accuracy in the approximation of the function could be achieved only by the division into small intervals.

In order to obtain the coefficients $c_i (i = 1, \dots, L)$, we apply the least squares method. The point of the method consists in choosing the coefficients in such way ensures the minimization of the integral value of the least squares error.

$$U = \int_a^b \left(p(x) - \sum_{i=1}^{L} c_i B_{i,k}(x) \right)^2 \mathrm{d}x$$

The function U is a function of L variables:

$$U = U(c_1, \dots, c_L) = \int_a^b \left(p(x) - \varphi(x; c_1, \dots, c_L) \right)^2 \mathrm{d}x$$

$$U = \int_a^b p^2(x)\,\mathrm{d}x - 2\sum_{i=1}^{L} c_i (p, B_{i,k}(x)) + \int_a^b \left(\sum_{i=1}^{L} c_i B_{i,k}(x) \right)^2 \mathrm{d}x.$$

All the partial derivatives of $U = U(c_1, \dots, c_L)$ are continuous. The necessary condition the function $U = U(c_1, \dots, c_L)$ to possess a minimum is expressed with the system of linear algebraic equations:

$$\frac{\partial U}{\partial c_1} = 0, \quad \frac{\partial U}{\partial c_2} = 0, \dots, \quad \frac{\partial U}{\partial c_L} = 0.$$

We obtain

$$\frac{\partial U}{\partial c_i} = -2(p, B_{i,k}(x)) + 2 \int_a^b \left(\sum_{j=1}^{L} c_j B_{j,k}(x) \right) B_{i,k}(x)\,\mathrm{d}x = 0$$

$$\sum_{j=1}^{L} (B_{i,k}(x), B_{j,k}(x)) c_j = (p(x), B_{i,k}(x)) \quad i = 1, \dots, n - k - 1,$$

where the inner product is defined by this means

$$(f, g) = \int_a^b f(x) g(x)\,\mathrm{d}x.$$

The unknown approximate coefficients c_j is obtained as a solution of the system of L equations. Obviously the functional $(p(x), B_{i,k}(x))$ is the mathematical expectation of $B_{i,k}(x)$ with a density $p(x)$ ([5]):

$$(p(x), B_{i,k}(x)) = \int_a^b p(x) B_{i,k}(x)\,\mathrm{d}x = \mathrm{E}B_{i,k}(\xi),$$

where the random number ξ has a density $p(x)$. The inner product $(p(x), B_{i,k}(x))$ is estimated using Monte Carlo method for calculation of the integrals:

$$(p(x), B_{i,k}(x)) \approx \frac{1}{N}\sum_{j=1}^{N} B(\xi_j) = \hat{\theta}_N.$$

where $\{\xi_i\}_{i=1}^N$ is the given sample. Each B-spline function of k-th degree is defined only in finite number of nodes - k+2. So, the first and the last nodes could be considered as limits of a subinterval corresponding to a subinterval which is obtained by splitting of the region using stratification method.

3. Non-parametric Techniques for Probability Density Estimation

In this section, some non-parametric techniques for probability density estimation are described (see [3, 4]). For these techniques few or no assumptions are made about what functional form the probability density takes. This is in contrast to a parametric method, where the density is estimated by assuming a distribution and then estimating the parameters. Two methods for probability density estimation are considered in the presented paper: histograms and kernel method.

3.1. Histograms

Histograms represent a graphical way of summarizing or describing a data set. A histogram visually conveys how a data set is distributed and provides information about relative frequencies of observations. Histograms are easy to create and are computationally feasible.

The histograms are the oldest and most widely used non-parametric density estimator. This is usually formed by dividing the real line into equally sized intervals, often called bins. The histogram is calculated using a random sample $\xi_1, \xi_2, \ldots, \xi_N$. Firstly, the origin x_0 for the bins and a bin width h have to be chosen. The bin width h is usually called a smoothing parameter since it controls the amount of "smoothing" being applied to the data. Our goal is to

estimate a probability density function, so we have to obtain a function $\hat{p}(x)$ that is nonnegative and satisfies the following condition:

$$\int_a^b \hat{p}(x)\,\mathrm{d}x = 1.$$

The one-dimensional histogram estimate $\hat{p}_{Hist}(x)$ at a point x is defined with the following expression:

$$\hat{p}_{Hist}(x) = \frac{\nu_i}{h\,N} = \frac{1}{h\,N}\sum_{i=1}^{N} I_{S_j}(\xi_i) \qquad \text{for} \quad x \in S_j,$$

where $S_j = [x_j, x_{j+1})$ is the j-th bin ($j = 0, \ldots, m-1$), ν_j is the number of observations in the j-th bin ($\sum_{j=0}^{m-1} \nu_j = N$), and $I_{S_j}(\xi_i)$ is the indicator function for the bin S_j.

3.2. Kernel Density Estimation

The histogram is informative but it is not smooth and not sensitive enough to local properties of the density $p(x)$. The kernel method overcomes this disadvantage of the histogram method.

Let $K(x)$ be a function that satisfies the conditions:

$$K(x) \geq 0, \qquad \int_a^b K(x)\mathrm{d}x = 1.$$

Then the kernel density estimator with kernel $K(x)$ is defined by

$$\hat{p}_{Ker}(x) = \frac{1}{\lambda\,N}\sum_{i=1}^{N} K\left(\frac{x-\xi_i}{\lambda}\right),$$

where λ is the bandwidth (smoothing parameter, window width).

4. Numerical Results

In this section the numerical results from the testing of described methods for probability density estimation are presented. The results are given as a function of the sample size N and the step of the mesh h. The sample is obtained using a pseudorandom generator. Absolute error in every point of mesh is computed.

The described algorithms are tested with different sample sizes for two different distribution-exponential and normal. The numerical results in all tables have been computed by using Matlab 6. In Tables 1 and 3 we summarize what is

known about the samples for considered distributions. Tables 2 and 4 show the error dependence of the sample size and the step of the mesh. When the number of the realizations increases the probable error decreases correspondingly. In view of the nature of the exponential distribution the step of the mesh increases with the increase of the number of realizations of the random variable but the Eu-

Table 1: Characteristics of the sample for exponential distribution.

Number of realizations	Min value in realization	Max value in realization	Probable error
100	0.00179	6.16620	0.00837
1 000	0.00098	6.33561	0.00279
10 000	3.39e-05	9.37242	0.00050
100 000	8.52e-06	13.14216	0.00009

Table 2: The Euclidean norm of the absolute error from exponential density reconstruction with Monte Carlo method and histogram method. The number of the intervals on the mesh is equal to 25.

Number of realizations	100	1 000	10 000	100 000
Step of the mesh	0.24657	0.25338	0.37489	0.52568
Monte Carlo method	0.37057	0.08860	0.09599	0.00696
Histogram method	0.31275	0.19913	0.22882	0.27470

clidean norm of the error decreases. The results for normal density reconstruction are similar to the previous but the desired accuracy is achieved slower.

The accuracy of the described Monte Carlo method for reconstruction of densities is compared with other statistical methods as histograms and kernel density estimation. From Table 2 it is clear that Monte Carlo method is more precise than the histogram method. For normal density reconstruction Monte Carlo method gives accuracy of the same order with comparison of the used statistical methods.

Graphic presentation of the density function is shown in case of exponential distribution (left-hand plots) and normal distribution (right-hand plots) of Figure 1. The true density is represented by a solid line on Figure 1 and 2. The stars

Table 3: Characteristics of the sample for normal distribution.

Number of realizations	Min value in realization	Max value in realization	Probable error
100	-2.04589	3.01190	0.01855
1 000	-3.88507	3.01190	0.00422
10 000	-3.88507	3.44948	0.00123
100 000	-4.02841	3.98811	0.00034

Table 4: The Euclidean norm of the absolute error from normal density reconstruction with Monte Carlo method, histogram method and kernel density estimation. The number of the intervals on the mesh is equal to 50.

Number of realizations	100	1 000	10 000	100 000
Step of the mesh	0.10115	0.13793	0.14669	0.16033
Monte Carlo method	1.19988	0.16516	0.04508	0.01765
Histogram method	0.97351	0.23474	0.07544	0.07355
Kernel density estimation	0.14004	0.07344	0.03231	0.01403

represent results obtained using Monte Carlo method. It is obvious from figures that with increasing the number of realizations the reconstructed values fit better to the corresponding density curve for all considered methods. This clearly shows that high accuracy can only be obtained using sufficiently large samples.

5. Conclusion

In the present paper a Monte Carlo method and non-parametric methods for density reconstruction are considered. Their accuracy in solving the investigated problem is compared. The obtained results show that the considered algorithm using Monte Carlo method gives the same accuracy with some advantage with comparison of the other mentioned methods for fixed number of realizations of random variable and fixed step of mesh.

Acknowledgments

This work was supported by Center of Excellence BIS-21 Grant ICA1-2000-70016.

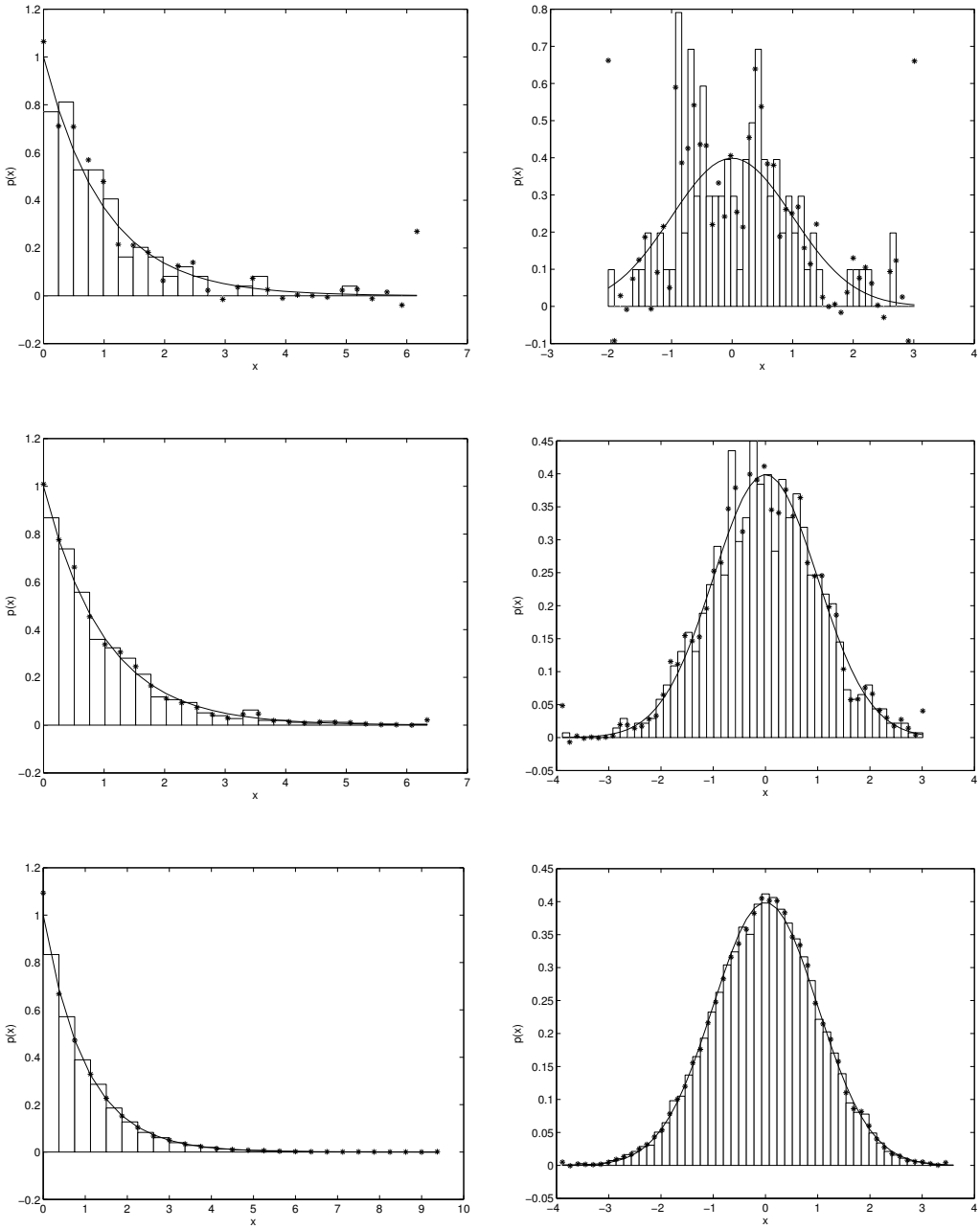

Figure 1: Exponential/normal density estimation with Monte Carlo method and histogram method. Sample sizes $N = 100, 1\,000, 10\,000$. Number of intervals is: 25 (exponential distribution) and 50 (normal distribution).

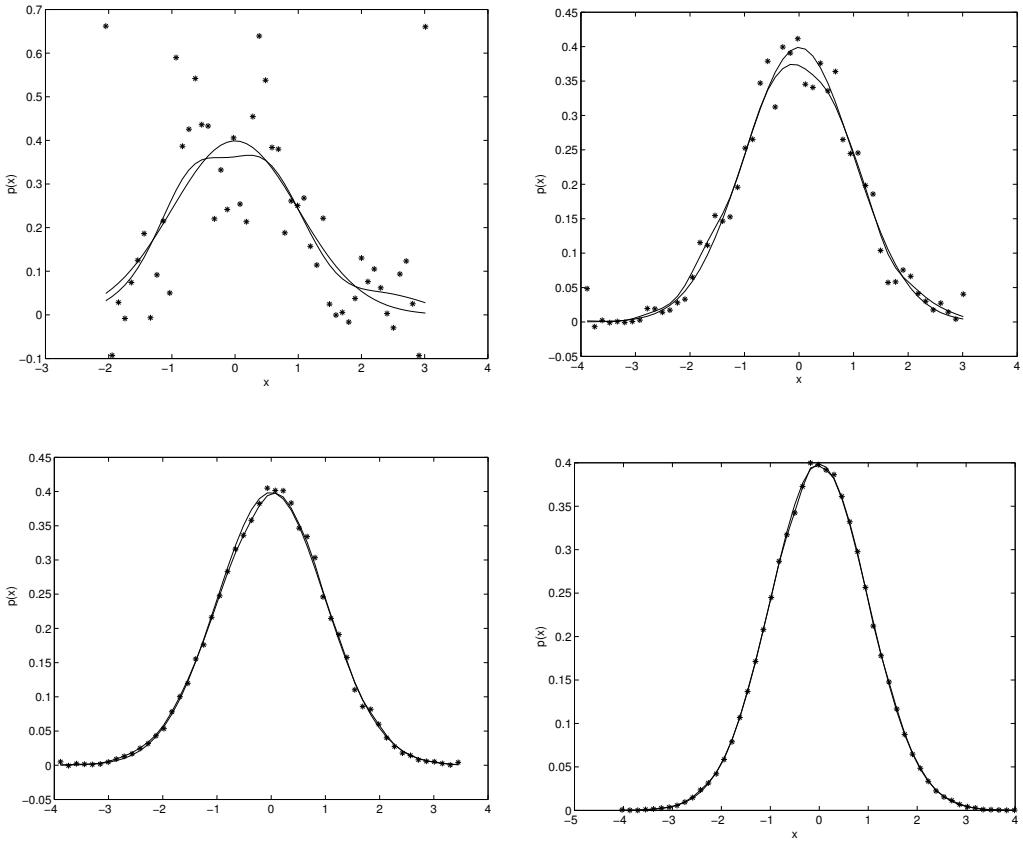

Figure 2: Normal density estimation with Monte Carlo method and kernel density estimation. Sample sizes $N = 100, 1\,000, 10\,000, 100\,000$. Number of intervals is 50.

REFERENCES

[1] I. Dimov, A. Karaivanova. Overconvergent Monte Carlo Methods for Density-function Modelling Using B-Splines, Advances in Numerical Methods and Appl., World Scientific, pp. 85–93, 1994.

[2] J. M. Hammersley, D. C. Handscomb. Monte Carlo Methods, Jonh Wiley & Sons, inc., New York, London, Methuen, 1964.

[3] W. L. Martinez, A. R. Martinez. Computational Statistics Handbook with MATLAB, Chapman & Hall/CRC, 2002.

[4] B. W. Silverman. Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[5] I. M. Sobol. Monte Carlo Numerical Methods, Nauka, Moscow, 1973, (in Russian).

Sofiya Ivanovska
Central Laboratory for Parallel Processing
Bulgarian Academy of Sciences
Acad. G. Bontchev St., Bl.25A
1113 Sofia, Bulgaria
e-mail: sofia@copern.bas.bg