УДК 519.25

НЕПАРАМЕТРИЧЕСКИЙ МЕТОД ВОССТАНОВЛЕНИЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ ПО НАБЛЮДЕНИЯМ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

А.Д. Крыжановский
А.А. Пастушков[@]

МИРЭА - Российский технологический университет, Москва 119454, Россия
[@]Автор для переписки, e-mail: Pastushkov.A@mail.ru

В ходе исследования статистических характеристик поля, образованного локально неоднородными областями, возникает задача восстановления по результатам экспериментальных наблюдений функции плотности вероятности с несколькими вершинами. Применять в этом случае параметрические методы восстановления плотности вероятности, как правило, крайне затруднительно. Поэтому для восстановления плотности вероятности имеет смысл применять непараметрические методы. Обычно используемый для этих целей метод Розенблатта-Парзена обладает невысокой точностью и скоростью сходимости. Метод, предложенный в работе Ченцова Н.Н., обладает более высокой точностью и скоростью сходимости, однако для многовершинных распределений его скорость сходимости также невелика. Аналогичные выводы можно сделать относительно метода, предложенного в работе Вапника В.Н. Поэтому проблема разработки методики восстановления многовершинной плотности вероятности по результатам экспериментальных наблюдений становится весьма актуальной. В работе предложен непараметрический метод восстановления многовершинной плотности вероятности случайного процесса по наблюдениям случайной величины. Метод является регулярным в смысле регуляризации Тихонова и, как показывает анализ и решение тестовых задач, обладает достаточно высокой точностью и скоростью сходимости.

Ключевые слова: непараметрические методы, функция распределения, плотность вероятности, функции отсчетов, ряд Уиттекера, квазирешение.

A NONPARAMETRIC METHOD OF RECONSTRUCTING PROBABILITY DENSITY ACCORDING TO THE OBSERVATIONS OF A RANDOM VARIABLE

A.D. Kryzhanovsky,
A.A. Pastushkov[@]

MIREA – Russian Technological University, Moscow 119454, Russia
[@]Corresponding author e-mail: Pastushkov.A@mail.ru

When investigating the statistical characteristics of a field formed by locally inhomogeneous regions, the problem of reconstructing the probability density function with several vertices on the basis of the results of experimental observations arises. In this case, it is very difficult to apply parametric methods for reconstructing the probability density. Therefore, to restore the probability density, it makes sense to use non-parametric methods of recovery. The Rosenblatt-Parzen method usually used for these purposes has low accuracy and convergence rate. The method proposed in the work of Chentsov N.N. has higher accuracy and convergence rate. However, for multi-vertex distributions its convergence rate is also low. Similar conclusions can be drawn regarding the method proposed in the work of Vapnik V.N. Thus, the problem of developing a technique for reconstructing the multi-vertex probability density on the basis of the results of experimental observations becomes very urgent. The article suggests a nonparametric method of reconstructing probability density according to the observations of a random variable. The method is regular in the sense of Tikhonov regularization and, as the analysis and solution of test problems show, it has sufficiently high accuracy and convergence rate.

Keywords: nonparametric methods, distribution function, probability density, sampling function, Whittaker series, quasisolution.

Введение

При исследовании статистических характеристик поля излучения рассеянного и/или излученного поверхностью, образованной локально неоднородными областями, возникает задача восстановления функции плотности вероятности с несколькими вершинами. В этом случае для восстановления плотности вероятности целесообразно применять непараметрические методы восстановления. Однако известные приемы обладают малой скоростью сходимости [1]. Согласно литературным данным, скорость сходимости в зависимости от применяемого метода, имеет оценку $O\left(n^{-\frac{1}{3}}\right)$ [2–5] или $O\left(n^{-\frac{1}{2}+\frac{1}{4k}}\right)$ [6], где n – число независимых наблюдений случайной величины, а k связано с числом интервалов гистограммы m соотношением $k = ln(n)/(2ln(m))$.

В [2] описан метод, имеющий более высокую скорость сходимости, однако, как показывает тестирование данного метода, он не всегда имеет высокую точность.

Уравнения для определения плотности вероятности по результатам наблюдения случайной величины

В соответствии с определением, функция плотности вероятности случайного процесса $\omega(x)$ является абсолютно интегрируемой функцией, удовлетворяющей условиям $\omega(x) \geq 0, \int_{-\infty}^{\infty} \omega(x)dx = 1$.

Рассмотрим без ограничения общности случай, когда плотность вероятности $\omega(x)$ задана на конечном интервале $[0, \pi]$, равна нулю на границах интервала, непрерывна и имеет непрерывную производную на этом интервале. Отсюда следует, что она принадлежит множеству функций с ограниченной вариацией V_C, удовлетворяющих на отрезке $[0, \pi]$ условиям $\max(\omega(x)) \leq M$, $V_a^b \omega \leq K$, (M и K – постоянные, одни и те же для всех $\omega \in V_C$). Так как $\omega(x)$ интегрируема на $[0, \pi]$, то функция распределения $F(x) = \int_0^x \omega(t)dt$ является

абсолютно непрерывной [7], ограниченной и имеет непрерывную производную $\omega(x)$, то есть $F(x) \in V_C$. Поскольку функции $F(x)$ и $\omega(x)$ интегрируемы, то, следовательно, интегрируемы в квадрате на интервале [0, π], т.е. принадлежат одновременно множествам V_C и L_2 $[a,b]$ [8].

Предположим далее, что функция $F(x)$, а, следовательно, и функция $\omega(x)$ [9] являются функциями с финитным спектром $F(x)$, $\omega(x) \in$ B. Тогда функции $F(x)$ и $\omega(x)$ можно представить в виде разложений в ряд Уиттекера:

$$L_n(F,x) = \sum_{k=0}^{n} F\left(\frac{\pi k}{n}\right) \cdot \mathrm{sinc}\left(n\left(x - \frac{\pi k}{n}\right)\right) \tag{1}$$

$$L_n(\omega,x) = \sum_{k=0}^{n} \omega\left(\frac{\pi k}{n}\right) \cdot sinc\left(n\left(x - \frac{\pi k}{n}\right)\right)$$

где $\mathrm{sinc}\left(n\left(x - \frac{\pi k}{n}\right)\right) = \sin\left(n\left(x - \frac{\pi k}{n}\right)\right) \Big/ n\left(x - \frac{\pi k}{n}\right)$ – система функций отсчетов ортогональная и полная на множестве $\mathrm{B} \subset \mathrm{L}_2$ $[a,b]$.

Поскольку ранее было принято, что функции $F(x)$ и $\omega(x)$ имеет ограниченную вариацию, то на любом подынтервале интервала [0, π] её можно представить в виде абсолютно сходящихся последовательностей [10] $\lim_{n\to\infty} \|L_n(F,x) - F(x)\| \to 0$, $\lim_{n\to\infty} \|L_n(\omega,x) - \omega(x)\| \to 0$.

Представим $F\left(\frac{\pi k}{n}\right)$ в виде

$$F\left(\frac{\pi k}{n}\right) = \int_0^{\frac{\pi k}{n}} \omega(t)dt \cong \frac{\pi}{n} \sum_{i=0}^{k} \omega\left(\frac{\pi i}{n}\right) \tag{2}$$

Подставляя (2) в (1), получим:

$$F(x) = \frac{\pi}{n} \sum_{k=0}^{\infty} \sum_{i=0}^{k} \omega\left(\frac{\pi i}{n}\right) sinc\left(n\left(x - \frac{\pi k}{n}\right)\right) = \frac{\pi}{n} \sum_{i=0}^{\infty} \omega\left(\frac{\pi i}{n}\right) \sum_{k=i}^{\infty} sinc\left(n\left(x - \frac{\pi k}{n}\right)\right)$$

Полагая далее, что x принимает дискретные значения $x_j = \frac{\pi}{n} j, j = 0...\infty$, перепишем последнее выражение в виде операторного уравнения

$$\mathbf{F} = \mathbf{A}_{\infty,\infty}\boldsymbol{\omega},$$

где $\mathbf{A}_{\infty,\infty}$ – линейный оператор.

Пусть j принимает конечное число значений, $j = 1..n$, тогда выражение для $\mathbf{F}$ можно записать в следующем виде

$$F_j = \frac{\pi}{n} \sum_{i=0}^{n} \omega_i \sum_{k=i}^{n} sinc\left(\pi\left(j-k\right)\right) = A_{n,n}\omega, x_j = \frac{\pi j}{n}, j = 0..n$$

где $\boldsymbol{A}_{n,n}$ – матрица с общим элементом вида $a_{i,j} = \sum_{k=i}^{n} \boldsymbol{sinc}\left(\pi\left(j-k\right)\right) = \begin{cases} 1, j \geq i \\ 0, j < i \end{cases}$, т.е. матрица $\boldsymbol{A}_{n,n}$ – верхняя треугольная, составленная из единиц. Обратная матрица от матрицы $\boldsymbol{A}_{n,n}$, $\boldsymbol{A}_{n,n}^{-1}$

является двухдиагональной матрицей, по главной диагонали которой стоят единицы. Вторая диагональ, расположенная над главной, состоит из отрицательных единиц.

Норма матрицы $\mathbf{A}_{n,n}$ ограничена при любых конечных n, и последовательность матриц $\mathbf{A}_{n,n}$ сходится по норме к оператору $\mathbf{A}_{\infty,\infty}$, действительно, как следует из [10]:

$$\lim_{n\to\infty}\left\|\mathbf{A}_{\infty,\infty}\mathbf{x}-\mathbf{A}_{n,n}\mathbf{x}\right\|=\lim_{n\to\infty}\left\|F(x)-\mathbf{A}_{n,n}\mathbf{x}\right\|\to 0,$$

$$\lim_{\substack{n\to\infty\\ m\to\infty}}\left\|\mathbf{A}_{m,m}\mathbf{x}-\mathbf{A}_{n,n}\mathbf{x}\right\|=\lim_{\substack{n\to\infty\\ m\to\infty}}\left\|\left(F(x)-\mathbf{A}_{m,m}\mathbf{x}\right)-\left(F(x)-\mathbf{A}_{n,n}\mathbf{x}\right)\right\|\le\lim_{n\to\infty}\left\|F(x)-\mathbf{A}_{m,m}\mathbf{x}\right\|+$$
$$+\lim_{m\to\infty}\left\|F(x)-\mathbf{A}_{n,n}\mathbf{x}\right\|\to 0.$$

Следовательно, оператор $\mathbf{A}_{\infty,\infty}$ является ограниченным оператором [11].

Линейный ограниченный оператор $\mathbf{A}_{\infty,\infty}$ взаимно однозначно отображает $\frac{\pi}{n}\boldsymbol{\omega}$, принадлежащее пространству Банаха l_2 в $\frac{\pi}{n}\mathbf{F}$, принадлежащее тому же пространству, следовательно, оператор $\mathbf{A}_{\infty,\infty}$ имеет ограниченный обратный оператор [7].

Рассмотрим операторное уравнение с приближенно заданной левой частью $\mathbf{F}_\varepsilon=\mathbf{A}_{\infty,\infty}\boldsymbol{\omega}$, где $F_\varepsilon(x)=F(x)+\varepsilon(x)$ – эмпирическая функция распределения;

ε – погрешность, относительно которой будем полагать, что она является отрезком центрированной случайной функции с финитным спектром, непрерывной и имеющей ограниченную вариацию.

В силу линейности пространства $B\subset L_2\,[a,b]$ функция $F_\varepsilon(x)$ так же, как и функция $F(x)$, будет принадлежать пространству $B\subset L_2\,[a,b]$.

Следовательно, мы получили операторное уравнение, связывающее эмпирическую функцию распределения $\mathbf{F}_\varepsilon$ с плотностью вероятности случайного процесса $\boldsymbol{\omega}$:

$$\mathbf{F}_\varepsilon=\mathbf{A}_{\infty,\infty}\boldsymbol{\omega} \tag{3}$$

Алгоритм восстановления плотности вероятности по результатам наблюдения случайной величины

Поскольку пространство l_2, которому принадлежат $\frac{\pi}{n}F_\varepsilon\left(\frac{\pi i}{n}\right),\frac{\pi}{n}\omega\left(\frac{\pi i}{n}\right)$, строго выпукло и компактно, и операторы $\mathbf{A}_{\infty,\infty}$, $\mathbf{A}^{-1}{}_{\infty,\infty}$, $\mathbf{A}_{n,n}$ и $\mathbf{A}^{-1}{}_{n,n}$ ограничены, то в соответствии с [12] для уравнения (3) можно построить квазирешение. Это решение единственно и является регулярным в смысле регуляризации Тихонова.

Решение уравнения будем искать методом конечномерной аппроксимации [13], заменив исходное уравнение уравнением

$$\mathbf{F}_\varepsilon=\mathbf{A}_{n,m}\boldsymbol{\omega}_{m,1}=\mathbf{A}_{n,n}\mathbf{S}_{n,m}\mathbf{a}_{m,1}=\mathbf{U}_{n,m}\mathbf{a}_{m,1}$$

где $\mathbf{A}_{n,n}$ – матрица с общим элементом $a_{i,j}=\sum_{k=j}^{n}sinc\left(\pi(i-k)\right),\ \ i,j=0...n$;

$\mathbf{S}_{n,m}$ – матрица с общим элементом $s_{i,j}=\sqrt{\frac{4}{\pi}}\cos\left((2j-1)\frac{\pi}{2}\frac{i}{n}\right),\ \ i=0...n,\ j=0...m.$,

откуда получим:

$$\boldsymbol{a} = \left(\boldsymbol{U}^T \boldsymbol{U}\right)^{-1} \boldsymbol{U}^T \widetilde{\boldsymbol{F}}.$$

Соответственно, для плотности вероятности найдем выражение, которое представлено ниже:

$$\omega(x) = \sum_{j=0}^{m} a_j \sqrt{\frac{4}{\pi}} \cos\left((2j-1)\frac{\pi}{2}x\right).$$

В связи с тем, что решение, найденное методом конечномерной аппроксимации, в общем случае единственным не является [13], то дополнительно потребуем, чтобы m выбиралось из условия m равно минимальному значению, при котором выполняется равенство

$$\mathbf{p}_{n,1}^T \mathbf{S}_{n,m} \mathbf{a}_{n,1} \frac{\pi}{n} = 1,$$

где **p** – единичный вектор.

Результаты восстановления плотности вероятности по предложенной в работе методике и методике работы [2] приведены ниже на рис.1 и 2. Пунктиром на рисунках обозначено точное значение функции плотности вероятности.

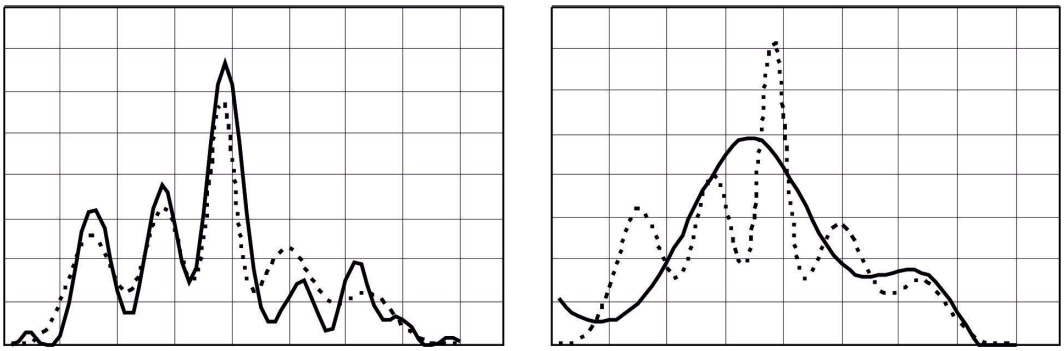

Рис. 1. Графики плотности вероятности, восстановленные по результатам измерения случайной величины, согласно предложенной методике (слева) и методике [2] (справа): объем выборки n = 64; относительная среднеквадратическая погрешность для найденного нами решения – 0.29, для решения, найденного по методике [2] – 0.45.

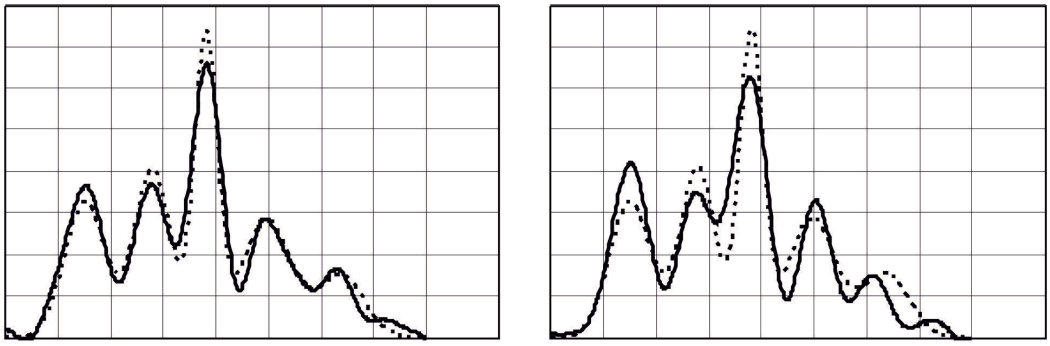

Рис. 2. Графики плотности вероятности, восстановленные по результатам измерения случайной величины, согласно предложенной методике (слева) и методике [2] (справа): объем выборки n = 512; относительная среднеквадратическая погрешность для предложенного нами решения – 0.11 (слева), для решения, найденного по методике [2] – 0.22.

Следует отметить, что алгоритм непараметрического восстановления плотности вероятности, предложенный нами, практически совпадает с алгоритмом, полученным ранее [14], однако предложенный в данной работе алгоритм обладает тем существенным преимуществом, что он не требует определения коэффициента регуляризации.

Полученные результаты можно обобщить на многомерный случай.

Пусть $f(x,y)$ является функцией с ограниченной плоской вариацией Тонелли, тогда при каждом фиксированном x_0 (y_0) функция $f(x, y_0)$, $f(x_0,y)$ будет иметь конечную вариацию $V(x, y_0)$ ($V(x_0, y)$). Таким образом, задачу восстановления плотности вероятности в двухмерном (многомерном[1]) случае можно свести к последовательному решению одномерных задач по сечениям y_i или x_i.

Анализ разработанного алгоритма

При решении тестовых задач последовательность случайных чисел с заданным законом распределения задавали следующим образом. Генерировали последовательность чисел $x_i = F(i/N)$, $i = 1...N$, здесь $F(x)$ – функция распределения, а N должно удовлетворять условию $N>>n$, где n – объем выборки, по которой находится эмпирическая функция распределения. Тестирование методики проводили для функций плотности вероятности с числом вершин от 2 до 7.

Установлено, что для всех видов функции плотности вероятности разработанная методика дает по величине относительной среднеквадратической погрешности выигрыш, в среднем, в два раза по сравнению с методикой работы [2].

Оценим скорость сходимости предложенного метода. Если n выбраны таким образом, что для погрешностей правой части уравнения (3) и оператора $\delta\mathbf{A}_n$ выполняется неравенство $\delta\widetilde{\mathbf{F}} >> \delta\mathbf{A}_n$, то для оценки погрешности восстановленной плотности вероятности $\omega(x)$ получим соотношение $\delta\boldsymbol{\omega} \sim \delta\widetilde{\boldsymbol{F}}$ [13]. Взяв в качестве оценки погрешности $\delta\widetilde{\boldsymbol{F}}$ величину модуля отклонения между теоретической и эмпирической функциями распределения $\delta\widetilde{\boldsymbol{F}} \sim n^{-1/2}$ [10], получим оценку для скорости сходимости метода $O(n^{-1/2})$, что лучше, чем оценка скорости сходимости метода, предложенного в [1].

На основании проведенных исследований предложена методика решения задач восстановления многовершинных функций плотности вероятности и показана ее более эффективность (по скорости сходимости), чем методика работы [1], так как при большом числе вершин число интервалов группировки m гистограммы должно быть достаточно большим.

Литература:

1. Вапник В.Н., Стефанюк А.Р. Непараметрические методы восстановления плотности вероятности // Автоматика и телемеханика. 1978. № 8. С. 38–52.

2. Вапник В.Н., Глазкова Т.Г., Кощеев В.А. Алгоритмы и программы восстановления зависимостей / под ред. В.Н. Вапника. М.: Наука, 1984. 815 с.

3. Кропотов Ю.А. Методы оценивания моделей плотности вероятностей акустических сигналов в телекоммуникациях аудиообмена // Системы управления, связи и безопасности. 2017. № 1. С. 26–39.

[1]При размерности n > 2 функции F(x), ω(x) должна быть функцией с существенно ограниченной вариацией [14].

4. Куликов В.Б. Восстановление полимодальных плотностей вероятности по экспериментальным данным в структурах со стохастическими свойствами // Вестник Нижегородского университета им. Н.И. Лобачевского. Математическое моделирование. Оптимальное управление. 2014. № 1 (1). С. 248–256.

5. Лапко А.В., Лапко В.А. Анализ эффективности методов дискретизации интервала измерений случайной величины при оценивании плотности вероятности // Информатика и системы управления. 2015. № 3(45). С. 84–88

6. Ченцов Н.Н. Оценка неизвестной плотности распределения по наблюдениям // Докл. АН СССР. 1962. Т. 147. № 1. С. 45–48.

7. Колмогоров А.Н., Фомин С.И. Элементы теории функций и функционального анализа. М: Наука, 1976. 542 с.

8. Ильин В.А., Садовничий В.А., Сендов Бл.Х. Математический анализ. Начальный курс. М.: Изд-во МГУ, 1985. 662 с.

9. Хургин Я.И., Яковлев В.П. Финитные функции в физике и технике. М.: Наука, 1971. 408 с.

10. Трынин А.Ю. Необходимые и достаточные условия равномерной на отрезке синк-аппроксимации функций ограниченной вариации // Изв. Сарат. ун-та. Нов. серия. Сер. Математика. Механика. Информатика. 2016. Т. 16. Вып. 3. С. 288–298.

11. Смирнов В.И. Курс высшей математики: в 5-ти т. Т. 5. М.: Наука, 1974. 600 с.

12. Иванов В.К., Васин В.В., Танана В.П. Теория линейных некорректных задач и ее приложения. М.: Наука, 1978. 206 с.

13. Танана В.П. Методы решения операторных уравнений. М.: Наука, 1981. 156 с.

14. Пастушков А.А. Непараметрический метод восстановления плотности вероятности по наблюдениям случайной величины // Научный вестник МИРЭА. 2009. № 1(6). С. 57–61.

15. Вольпе А.И., Худяев С.И. Анализ в классах разрывных функций и уравнения математической физики. М.: Наука, 1975. 395 с.

References:

1. Vapnik V.N., Stefanyuk A.R. Nonparametric methods for recovering the probability density // Avtomatika i telemehanika (Automation and Remote Control). 1978. № 8. P. 38–52. (in Russ.).

2. Vapnik V.N., Glazkova T.G., Kosheev V.A. Algorithms and programs for recovering dependencies / Ed. V.N. Vapnik. Moscow: Nauka Publ., 1984. 815 p. (in Russ.).

3. Kropotov Yu.A. Methods of estimation models of acoustic signals probability density in telecommunications audio-exchange systems // Sistemy upravleniya, svyazi i bezopasnosti (Systems of Control, Communications and Security). 2017. V. 1. P. 26–39. (in Russ.).

4. Kulikov V.B. Reconstruction of multimodal probability densities from experimental data in structures with stochastic properties // Vestnik Nizhegorodskogo universiteta imeni N.I. Lobachevskogo. Matematicheskoye modelirovaniye. Optimal'noye upravleniye (Vestnik of Lobachevsky University of Nizhniy Novgorod. Mathematic Modeling. Optimal Control). 2014. V. 1 (1). P. 248–256. (in Russ.).

5. Lapko A.V., Lapko V.A. The analysis of efficiency of decomposition methods of

measurements interval of the random variable at a probability density estimation // Informatika i sistemy upravleniya (Information Science and Control Systems). 2015. V. 3 (45). P. 84–88. (in Russ.).

6. Chentsov N.N. Estimation of the unknown distribution density by observations // Doklady Akademii Nauk SSSR. 1962. V. 147. № 1. P. 45–48. (in Russ.).

7. Kolmogorov A.N., Fomin S.I. Elements of the theory of functions and functional analysis. M.: Nauka Publ., 1976. 542 p.

8. Il'yin V.A., Sadovnichii V.A., Sendov Bl.X. Mathematical analysis. The initial course. Moscow: Publishing House of Moscow State University, 1985. 662 p.

9. Khurgin Ya.I., Yakovlev V.P. Finite functions in physics and engineering. Moscow: Nauka Publ., 1971. 408 p.

10. Trynin A.Yu. Necessary and sufficient conditions for the uniform on a segment sink-approximations functions of bounded variation // Izvestiya Saratovskogo Universiteta. Novaya Seriya. Matematika. Mekhanika. Informatika (Izvestiya of Saratov University. New Series. Series: Mathematics. Mechanics. Informatics). 2016. V. 16. № 3. P. 288–298. (in Russ.).

11. Smirnov V.I. Course of Higher Mathematics: in fives vols. V. 5. Moscow: Nauka Publ., 1974. 600 p. (in Russ.).

12. Ivanov V.K., Vasin V.V., Tanana V.P. Theory of linear ill-posed problems and its applications. Moscow: Nauka Publ., 1978. 206 p. (in Russ.).

13. Tanana V.P. Methods for solving operator equations. Moscow: Nauka Publ., 1981. 156 p. (in Russ.).

14. Pastushkov A.A. Nonparametric method for recovering the probability density from observations of a random variable // Nauchnyi vestnik MIREA (Scientific Bulletin of MIREA). 2009. № 1(6). P. 50–61. (in Russ.).

15. Volper A.I., Khudyaev S.I. Analysis in classes of discontinuous functions and equations of mathematical physics. Moscow: Nauka Publ., 1975 395 p. (in Russ.).

Об авторах:

Крыжановский Александр Дмитриевич, магистрант, кафедра информационных систем, Института кибернетики ФГБОУ ВО «МИРЭА - Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского д. 78).

Пастушков Александр Анатольевич, кандидат технических наук, доцент, доцент кафедры информационных систем Института кибернетики ФГБОУ ВО «МИРЭА - Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского д. 78).

About the authors:

Alexander D. Kryzhanovskiy, AD, Master of Arts, Department of Information Systems, Institute of Cybernetics, MIREA – Russian Technological University (119454, Russia, Moscow, 78 Vernadsky Ave.).

Alexander A. Pastushkov, Ph. D., Associate Professor, Associate of the Department of Information Systems, Institute of Cybernetics, MIREA – Russian Technological University (119454, Russia, Moscow, 78 Vernadsky Ave.).