Сравнительный анализ методов Розенблатта-Парзена и структурной минимизации риска для аппроксимации плотностей вероятностей случайных величин

С. В. Поршнев, А. С. Копосов, Е. И. Березовик

Уральский Федеральный Университет имени первого Президента России Б. Н. Ельцина
620002, Екатеринбург, ул. Мира, 19

e-mail: s.v.porshnev@urfu.ru, alexkopas@gmail.com, miss.berezovik@mail.ru

Аннотация. Проводится сравнительный анализ результатов применения методов аппроксимации Розенблатта-Парзена (АРП) и структурной минимизации риска (СМР) для аппроксимации плотностей вероятностей (ПВ) случайных величин с ограниченной областью рассеяния. Известны два подхода к решению этой задачи: параметрический и непараметрический. В соответствие с первым подходом на основе априорной информации выбирают вид функции распределения (ФР) случайной величины, зависящей от некоторого набора параметров, и меру близости между теоретической и экспериментальной ФР. В основе непараметрической статистики лежит подход, позволяющий получать адаптивные оценки эмпирических ФР в виде некоторых функционалов, независящих от вида выбираемой на основе априорной информации ФР. В АРП метод восстановления плотности распределения экспериментальной выборки основан на предположении о том, что ФР оценивается локально в каждой точке с помощью элементов обучающей выборки из некоторой окрестности данной точки. При этом общая ФР есть некоторая линейная комбинация известных ядерных функций. В методе СМР оценка ПР ищется в виде разложения по системе тригонометрических функций. Для сравнительного анализа были использованы случайные величины с одно-, двух и трехмодовыми ПВ. Для оценки качества аппроксимации анализируемых методов использовалось значение интегральной погрешности. Получены оценки точности аппроксимации и времени вычисления ПВ, каждым из выбранных методов. Для проведения анализа построены сводные таблицы точности аппроксимации и времени вычислении ПВ. Сделаны выводы о достоинствах и недостатках методов. Предложены рекомендации по использованию того или иного метода в зависимости от размера исходной выборки.

Ключевые слова: ограниченная область рассеяния, непараметрическая статистика, эмпирическая функция распределения, аппроксимация Розенблатта-Парзена, метод структурной минимизации риска.

1. Введение

Восстановление функции распределения по выборке случайных данных, полученных в результате проведения тех или иных экспериментов, является одной из основных задач прикладной математической статистики [1], которая имеет важное практическое значение, например, при решении задач прочностной надежности элементов и объектов нефтегазового оборудования [2]. Данная задача имеет следующую постановку: по экспериментальной выборке из генеральной совокупности значений $X_i, i=\overline{1,N}$ найти соответствующую функцию распределения (ФР) $F(y)=\Pr\{X \le y\},$ связанную с ПР $f(y)$ следующим соотношением:

$$F(y)=\int_{-\infty}^{y} f(\xi)d\xi, \tag{1}$$

соответственно,

$$f(y)=\frac{dF(y)}{dy}. \tag{2}$$

Известны два подхода к решению этой задачи: параметрический и непараметрический. В соответствие с первым подходом на основе априорной информации выбирают вид ФР случайной величины $X_i,$ зависящей от некоторого набора параметров, и меру близости между теоретической и экспериментальной ФР:

$$F_N(y)=\frac{1}{N}\sum_{i=1}^{N}\Theta(y-x_i), \tag{3}$$

где функция Хэвисайда

$$\Theta(y-x_i)=\begin{cases}1, \text{ при } y-x_i \ge 0,\\ 0, \text{ при } y-x_i < 0,\end{cases}$$

также, вообще говоря, зависящую от вида распределения [4]. Далее находят оценки значений параметров ФР, обеспечивающих максимальную близость теоретической ФР и эмпирической ФР. Существование решения обсуждаемой задачи обеспечивает центральная теорема математической статистики, согласно которой с ростом объема выборки N функция $F_N(y)$ с вероятностью, равной единице, равномерно приближается по ФР к $F(y)$:

$$\Pr\{\lim_{N\to\infty}\sup|F_N(y)-F(y)|=0\}=1.$$

В основе непараметрической статистики лежит подход, позволяющий получать адаптивные оценки эмпирических ФР в виде некоторых функционалов, независящих от вида выбираемой на основе априорной информации ФР [2]. Для этого разработан целый ряд известных методов [2, 6–9], в том числе: метод гистограмм, ме-

тод «гребенка», метод ближайших соседей, метод разложения по базисным функциям, аппроксимация Розенблатта-Парзена и ряд других. Работоспособность методов непараметрической статистики и целесообразность их применения при анализе экспериментальных данных подтверждается результатами, полученными различными исследователями, см., например, [3].

Напомним, следуя [2], что данный метод восстановления плотности распределения экспериментальной выборки основан на предположении о том, что ФР оценивается локально в каждой точке x_i с помощью элементов обучающей выборки из некоторой окрестности x_i. При этом общая функция вероятности $F(y)$ есть некоторая линейная комбинация известных функций:

$$F(y) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{y - x_i}{h}\right), \tag{4}$$

где $K(t), t = (y - x_i)/h$ — ядерная функция, удовлетворяющая следующим условиям:

а) $K(t)$ — монотонно неубывающая функция, область значений которой принадлежит интервалу [0,1];

б) $K(t) = 1 - K(t)$ — функция, симметричная относительно 0;

в) $h_N \to 0$ при $N \to \infty$.

Здесь h — параметр «размытости», определяющий гладкость получаемой оценки.

Соответственно, ПР вычисляется по формуле

$$f(y) = \frac{1}{N \cdot h}\sum_{i=1}^{N} k\left(\frac{y - x_i}{h}\right), \tag{5}$$

где $k(y) = \frac{d}{dy}K(y)$.

На практике наиболее часто в качестве ядерных функций $k(y)$ используются функции, представленные в табл. 1 [3].

Оптимальные значения ядерной функции и параметра h находятся из условия достижения информационным функционалом

$$J = \int_{-\infty}^{\infty} \ln k(t) \cdot f(t) dt \tag{6}$$

максимального значения, которое, как очевидно, выполняется при $k(t) = f(t)$ [7, 8].

Результаты исследования особенностей аппроксимации Розенблатта-Парзена в задаче аппроксимации одномодальных распределений дискретных и непрерывных случайных величин с ограниченной областью рассеяния изложены в [10] и [11], соответственно.

Таблица 1. Ядерные функции, наиболее часто используемые на практике

№	*Ядро*	*Формула*
1	Нормальное	$k(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$
2	Лапласа	$k(t)=\frac{1}{2}e^{-\lvert t\rvert}$
3	Фишера	$k(t)=\frac{1}{2\pi}\left(\frac{\sin(t/2)}{t/2}\right),\lvert t/2\rvert\le\pi$
4	Коши	$k(t)=\frac{1}{\pi}\left(\frac{1}{1+t^2}\right)$
5	Логистическое	$k(t)=\frac{e^{-t}}{(1+e^{-t})^2}$
6	Епанечникова	$k(t)=\frac{3\left(1-t^2/5\right)}{4\sqrt{5}},\ \lvert t\rvert\le\sqrt{5}$
7	Равномерное	$k(t)=\frac{1}{2},\ \lvert t\rvert\le 1$
8	Треугольное	$k(t)=1-\lvert t\rvert,\ \lvert t\rvert\le 1$
9	Квадратичное	$k(t)=\frac{3}{4}\left(1-t^2\right),\ \lvert t\rvert\le 1$

Также известен альтернативный подход к непараметрической аппроксимации, описанный в [15, 16], в соответствие с которым неизвестная ПР $f(t)$ предполагается непрерывной и сосредоточенной на отрезке $t\in[0,1]$, а оценка ПР ищется в виде разложения по системе тригонометрических функций

$$\varphi_j(t)=\sqrt{\frac{4}{\pi}}\cos\left((2j-1)\frac{\pi}{2}t\right),\ j=1,2,\ldots,$$

следующим образом

$$f^{(N)}(t)=\sum_{j=1}^{N}\lambda_j\varphi_j(t), \qquad (7)$$

где λ_j — коэффициенты разложения. Здесь число тригонометрических функций N («сложность» оценки) и значения коэффициентов разложения λ_j находятся с помощью метода структурной минимизации риска [15, 16].

Однако сравнительного анализа данных методов аппроксимации ПР случайных последовательностей, а также соответствующих рекомендаций по выбору используемого в конкретной ситуации метода, многочисленных публикациях по не-

параметрической статистике обнаружить не удается. В этой связи исследование данных методов представляет практический интерес.

В статье обсуждаются результаты сравнительного анализа оценок ПР распределений случайных последовательностей, вычисленных с помощью аппроксимации Розенблатта-Парзена и метода структурной минимизации риска, с точки зрения затрат вычислительных ресурсов (время вычисления) и точности аппроксимации ПР

2. Методика исследования

В качестве объекта исследования были использованы случайные числа с ограниченной областью рассеяния. Их выбор обусловлен тем, что параметры большого числа реальных технических систем относятся к данному классу случайных распределений [4].

Напомним, что физическая модель случайной величины с ограниченной областью рассеяния (СВООР) была предложена А. Эйнштейном и Смолуховским [4, 5]. В соответствие с данной моделью СВООР порождают значения траектории броуновской частицы, совершающая одномерные случайные блуждания на отрезке $[a, b]$, от границ которого она испытывает абсолютно упругие отражения. Можно показать [14], что ПР данной случайной величины вычисляется формуле

$$f(x;\mu,\sigma,a,b) = A\left[\varphi(x;\mu,\sigma,a,b) + \sum_{g=0}^{\infty}\varphi^{\pm}_{2g+1}(x;\mu,\sigma,a,b) + \sum_{g=1}^{\infty}\varphi^{\pm}_{2g}(x;\mu,\sigma,a,b)\right], \qquad (8)$$

где A — нормировочный коэффициент, определяемый из условия

$$\int_a^b f(\xi;\mu,\sigma,a,b)d\xi = 1,$$

$$\varphi(x;\mu,\sigma,a,b) = \exp[-(x-\mu)^2/2\sigma^2],$$

$$\varphi^{\pm}_{2g+1}(x;\mu,\sigma,a,b) = \exp[-(x-x^{\pm}_{2g+1})^2/2\sigma^2],$$

$$\varphi^{\pm}_{2g}(x;\mu,\sigma,a,b) = \exp[-(x-x^{\pm}_{2g})^2/2\sigma^2].$$

Из (8) видно, что ПР представляет собой линейную комбинацию плотностей нормального закона, центры распределений которых находятся по следующим формулам

$$x^{\pm}_{2g} = \pm 4g(b-a)+\mu,\ x^{\pm}_{2g+1} = \pm(4g+2)(b-a)-\mu,$$

где $g = 0, 1, \ldots$.

Отметим, что, используя (8), можно создавать двух- и трех-модальные распределения СВООР:

$$f_{2\mathrm{mod}}(x) = f_1(x,\mu_1,\sigma_1,a_1,b_1)\cdot\alpha + f_2(x,\mu_2,\sigma_2,a_2,b_2)\cdot(1-\alpha), \qquad (9)$$

$$f_{3\mathrm{mod}}(x) = f_1(x,\mu_1,\sigma_1,a_1,b_1)\cdot\alpha_1 + f_2(x,\mu_2,\sigma_2,a_2,b_2)\cdot\alpha_2 + \\ + f_3(x,\mu_3,\sigma_3,a_3,b_3)\cdot(1-\alpha_1-\alpha_2). \quad (10)$$

Для сравнительного анализа были использованы СВООР, сгенерированные в соответствии с (8)–(10). Параметры ПР выбирались аналогичные использованным ранее при исследовании сравнения точности оценивания параметров одно- и двухмодальных ПР с помощью генетических алгоритмов и аппроксимации Розенблатта-Парзена [12–14].

Параметры распределений, в соответствие с которыми генерировались случайные величины, представлены в табл. 2–4.

Таблица 2. Параметры одномодальных распределений

Номер распределения	μ	σ	*a*	*b*
1	50	10	0	100
2	50	20	0	100
3	50	30	0	100
4	30	20	0	100

Таблица 3. Параметры двумодальных распределений

Номер распределения	μ_1	σ_1	a_1	b_1	μ_2	σ_2	a_2	b_2	α
5	30	15	0	100	70	5	0	100	0,5
6	30	10	0	100	70	15	0	100	0,7
7	20	15	0	100	80	10	0	100	0,5
8	30	10	0	100	70	10	0	100	0,4

Таблица 4. Параметры трехмодального распределения

Номер распределения	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3	$a_1 = a_2 = a_3$	$b_1 = b_2 = b_3$	α_1	α_2
9	20	5	50	5	70	5	0	100	0,3	0,3
10	20	7	55	5	70	5	0	100	0,3	0,3
11	20	5	50	5	70	5	0	100	0,2	0,3
12	20	10	50	7	70	5	0	100	0,3	0,3

В проведенных экспериментах были использованы 12 наборов параметров (*NumTypes* = 12) по 4 набора для каждого типа распределений. Для каждого набора параметров генерировались выборки следующих размеров: 30, 50, 100, 200, 300, 500. (*NumSelectionCount* = 6). Для каждого набора параметров и размера выборки вычислялось количество реализаций выборки (*NumExp* = 10).

Для оценки качества аппроксимации анализируемых методов использовалось значение интегральной погрешности, вычисляемое относительно теоретической функции распределения случайной последовательности по следующей формуле:

$$\Delta_{integr} = \frac{\sum_i (F_{teor}(x_i) - F_{pract}(x_i))^2}{\sum_i F_{teor}(x_i)^2}, \qquad (11)$$

Для каждой реализации вычислялась аппроксимация плотности вероятности по методу Розенблатта-Парзена и по методу структурной минимизации риска, а также соответствующие интегральные погрешности Δ_{integr} и время вычисления t. Затем интегральная погрешность и время вычисления усреднялись по ансамблю реализаций.

3. Анализ результатов

Примеры результатов оценивания ПР одномодальных, двумодальных и трехмодальных выборочных СП с ограниченной областью рассеяния для каждого из описанных выше наборов параметров представлены на рис. 1–3.

Усредненные по ансамблям реализаций значения времени, затраченного для вычислений аппроксимаций ПР СП, представленных на рис. 1–3, приведены в табл. 5–6. Усредненное по набору распределений время вычисления в секундах в зависимости от числа элементов выборки представлена на рис. 4.

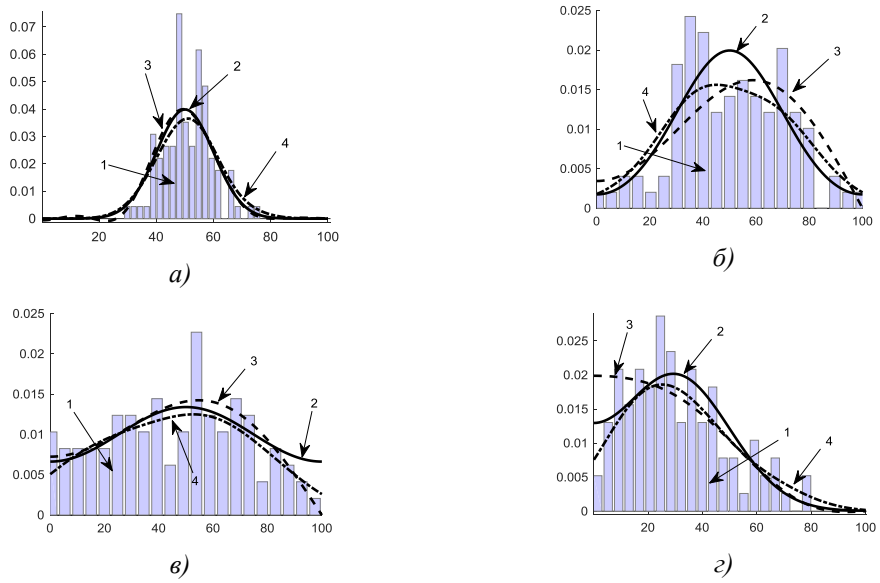

Рисунок 1. Результаты аппроксимации одномодальной ПР СП с ограниченной областью рассеяния, размер выборки $N = 500$: а) СП № 1; б) СП № 2; в) СП № 3, г) СП № 4; 1 — гистограмма выборки; 2 — теоретическая ПР, 3 — аппроксимация ПР методом структурной минимизации риска; 4 — аппроксимация Розенблатта-Парзена

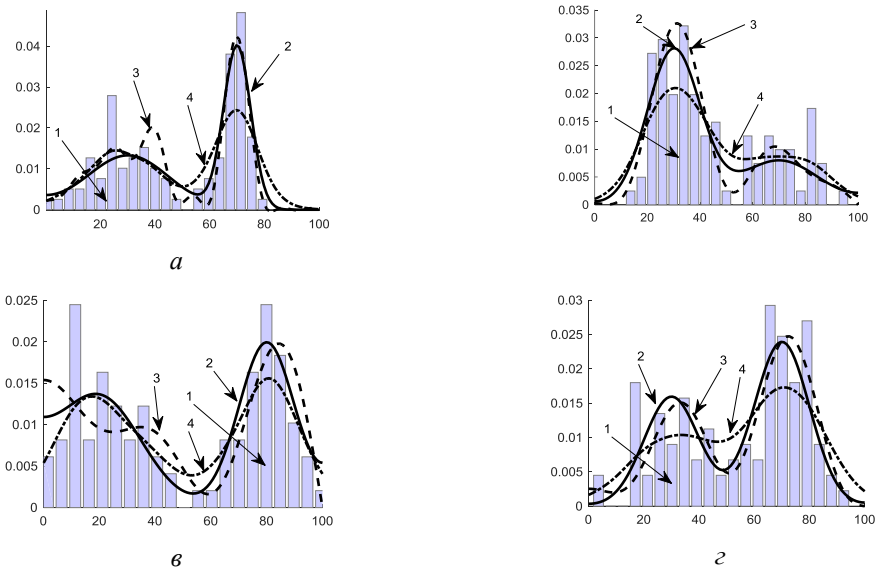

Рисунок 2. Результаты аппроксимации двумодальной ПР СП с ограниченной областью рассеяния, размера выборки $N = 500$: а) СП № 5; б) СП № 6; в) СП № 7; г) СП № 8; 1 — гистограмма выборки; 2 — теоретическая ПР, 3 — аппроксимация ПР методом структурной минимизации риска; 4 — аппроксимация Розенблатта-Парзена

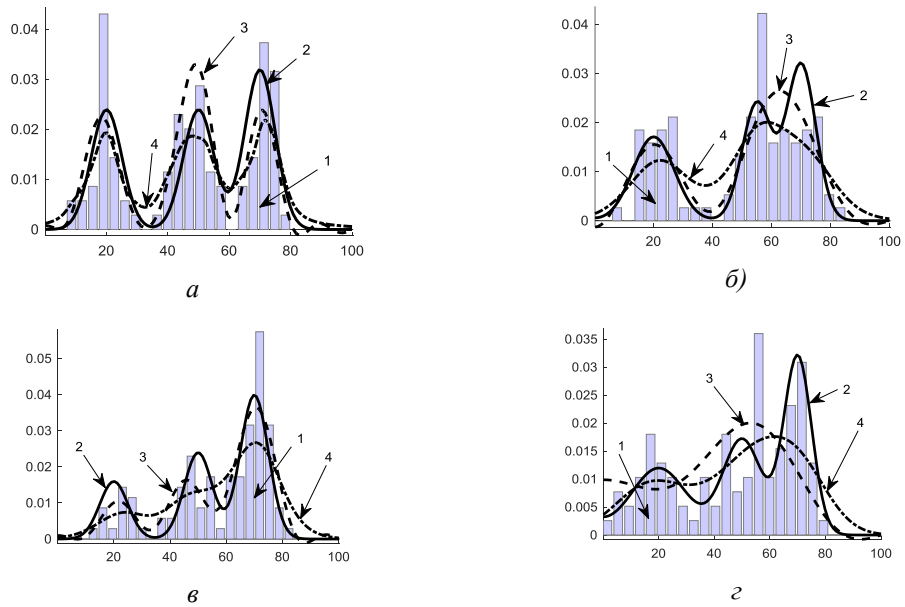

Рисунок 3. Результаты аппроксимации трехмодальной ПР СП с ограниченной областью рассеяния, размера выборки $N = 500$: а) СП № 9; б) СП № 10; в) СП № 11; г) СП № 12; 1 — гистограмма выборки; 2 — теоретическая ПР, 3 — аппроксимация ПР методом структурной минимизации риска; 4 — аппроксимация Розенблатта-Парзена

Таблица 5. Значения времени, затраченного для вычислений аппроксимаций ПР СП с помощью аппроксимации Розенблатта-Парзена, в секундах

Номер распределения	*Тип распределения*	*Число элементов СП. N*					
		30	*50*	*100*	*200*	*300*	*500*
1	одномодальное	0.3096	0.1541	0.5226	1.9442	4.2366	11.6722
2		0.0719	0.1480	0.5000	2.0391	4.7897	12.1768
3		0.0760	0.1554	0.5237	2.0065	4.4858	12.5198
4		0.0790	0.1570	0.5457	2.0518	4.4958	12.8216
$\bar{t}$		0.1342	0.1536	0.5230	2.0104	4.5020	12.2976
σ_t		0.1170	0.0040	0.0186	0.0481	0.2262	0.4932
5	двухмодальное	0.0737	0.1703	0.6775	2.7666	4.7466	15.1806
6		0.1060	0.2725	0.7612	2.5337	7.0093	17.2169
7		0.0874	0.2399	0.7832	3.3426	5.4030	15.0317
8		0.0714	0.1561	0.5271	2.0253	4.5812	16.4738
$\bar{t}$		0.0846	0.2097	0.6873	2.6670	5.4350	15.9757
σ_t		0.0159	0.0556	0.1161	0.5465	1.1079	1.0507
9	трехмодальное	0.0786	0.1520	0.5852	2.0329	4.3215	12.7809
10		0.0715	0.1538	0.5331	2.0054	4.7785	15.0850
11		0.0836	0.1835	1.1009	2.8176	4.6185	12.4203
12		0.0697	0.1535	0.5225	1.9883	4.6274	12.5336
$\bar{t}$		0.0758	0.1607	0.6854	2.2110	4.5865	13.2050
σ_t		0.0064	0.0152	0.2784	0.4048	0.1913	1.2624
$\bar{t}$	Все распределения	0.0982	0.1747	**0.6319**	**2.2962**	**4.8412**	**13.8261**
σ_t		0.0673	0.0399	**0.1771**	**0.4573**	**0.7432**	**1.8634**

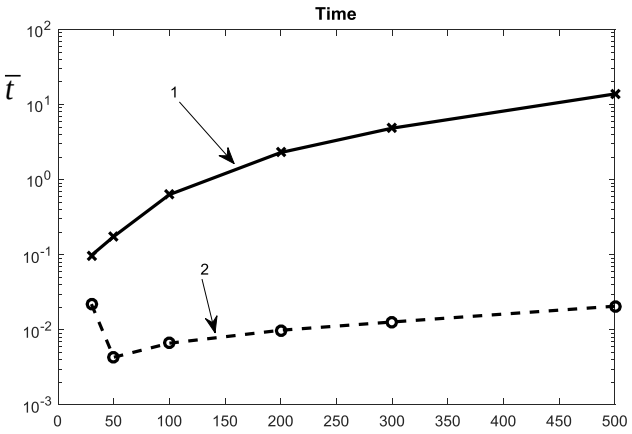

Рисунок 4. Усредненное по набору распределений время вычисления в секундах $\bar{t}$ в зависимости от числа элементов выборки N. 1 — метод Розенблатта-Парзена; 2 — метод структурной минимизации риска.

Усредненные по ансамблям реализаций интегральные погрешности представлены в табл. 7–8; усредненная по набору распределений интегральная погрешность в зависимости от числа элементов выборки показана на рис. 5.

Таблица 6. Значения времени, затраченного для вычислений аппроксимаций ПР СП с помощью в метода структурной минимизации риска, в секундах

Номер распределения	*Тип распределения*	*Число элементов СП. N*					
		30	50	100	200	300	500
1	одномодальное	0.2277	0.0056	0.0062	0.0107	0.0128	0.0195
2		0.0039	0.0041	0.0076	0.0094	0.0137	0.0187
3		0.0034	0.0043	0.0062	0.0086	0.0116	0.0181
4		0.0033	0.0043	0.0059	0.0103	0.0119	0.0179
$\bar{t}$		0.0596	0.0046	0.0065	0.0097	0.0125	0.0186
σ_t		0.1121	0.0007	0.0008	0.0009	0.0009	0.0007
5	двухмодальное	0.0028	0.0045	0.0064	0.0098	0.0108	0.0200
6		0.0041	0.0048	0.0074	0.0100	0.0155	0.0197
7		0.0038	0.0055	0.0090	0.0096	0.0124	0.0178
8		0.0026	0.0032	0.0055	0.0079	0.0153	0.0221
$\bar{t}$		0.0033	0.0045	0.0071	0.0093	0.0135	0.0199
σ_t		0.0008	0.0010	0.0015	0.0010	0.0023	0.0018
9	трехмодальное	0.0031	0.0036	0.0060	0.0101	0.0121	0.0173
10		0.0027	0.0038	0.0059	0.0087	0.0113	0.0420
11		0.0037	0.0043	0.0077	0.0140	0.0133	0.0161
12		0.0031	0.0037	0.0061	0.0095	0.0121	0.0187
$\bar{t}$		0.0032	0.0038	0.0064	0.0106	0.0122	0.0235
σ_t		0.0004	0.0003	0.0008	0.0023	0.0008	0.0124
$\bar{t}$	Все распределения	0.0220	0.0043	0.0066	0.0099	0.0127	0.0207
σ_t		0.0648	0.0007	0.0010	0.0015	0.0015	0.0069

Таблица 7. Интегральная погрешность Δ_{integr} в методе Розенблатта-Парзена

Номер распределения	*Тип распределения*	*Число элементов СП. N*					
		30	*50*	*100*	*200*	*300*	*500*
1	одномодальное	0.1002	0.0226	0.0322	0.0278	0.0252	0.0207
2		0.0987	0.0347	0.0209	0.0237	0.0142	0.0120
3		0.0518	0.0230	0.0198	0.0137	0.0143	0.0105
4		0.0672	0.0616	0.0249	0.0186	0.0145	0.0128
$\bar{\Delta}_{integr}$		0.0795	0.0355	0.0245	0.0209	0.0170	0.0140
$\sigma_{\Delta_{integr}}$		0.0239	0.0183	0.0056	0.0061	0.0055	0.0046
5	двухмодальное	0.3144	0.2001	0.1494	0.1068	0.0863	0.0766
6		0.1205	0.1482	0.0690	0.0470	0.0420	0.0361
7		0.2318	0.1371	0.0723	0.0501	0.0408	0.0331
8		0.1942	0.1441	0.0681	0.0613	0.0441	0.0350
$\bar{\Delta}_{integr}$		0.2152	0.1574	0.0897	0.0663	0.0533	0.0452
$\sigma_{\Delta_{integr}}$		0.0807	0.0289	0.0398	0.0277	0.0220	0.0209
9	трехмодальное	0.2575	0.2305	0.1426	0.1027	0.0805	0.0661
10		0.2121	0.1347	0.1150	0.0688	0.0630	0.0492
11		0.2497	0.1300	0.1706	0.0953	0.0850	0.0593
12		0.2517	0.1494	0.1407	0.0964	0.0771	0.0724
$\bar{\Delta}_{integr}$		0.2427	0.1612	0.1422	0.0908	0.0764	0.0617
$\sigma_{\Delta_{integr}}$		0.0207	0.0470	0.0227	0.0150	0.0095	0.0100
$\bar{\Delta}_{integr}$	Все распределения	0.1791	0.1180	0.0854	0.0594	0.0489	0.0403
$\sigma_{\Delta_{integr}}$		0.0872	0.0681	0.0558	0.0346	0.0286	0.0241

Таблица 8. Интегральная погрешность Δ_{integr} в методе структурной минимизации риска

Номер распределения	*Тип распределения*	*Число элементов СП. N*					
		30	*50*	*100*	*200*	*300*	*500*
1	одномодальное	0.0578	0.0797	0.0437	0.0561	0.0754	0.1855
2		0.0930	0.0734	0.0722	0.0540	0.0658	0.0896
3		0.1154	0.1120	0.0642	0.0314	0.0333	0.0359
4		0.1171	0.0724	0.0521	0.0629	0.0422	0.0642
$\overline{\Delta}_{integr}$		0.0958	0.0844	0.0580	0.0511	0.0542	0.0938
$\sigma_{\Delta_{integr}}$		0.0276	0.0187	0.0126	0.0137	0.0197	0.0649
5	двухмодальное	0.3600	0.2686	0.0868	0.2180	0.3120	0.4756
6		0.1738	0.1137	0.0512	0.0825	0.1538	0.3138
7		0.1850	0.1075	0.0997	0.0416	0.0482	0.0712
8		0.2137	0.1726	0.0521	0.0833	0.1684	0.1834
$\overline{\Delta}_{integr}$		0.2331	0.1656	0.0725	0.1064	0.1706	0.2610
$\sigma_{\Delta_{integr}}$		0.0862	0.0747	0.0246	0.0769	0.1084	0.1740
9	трехмодальное	0.2709	0.1666	0.0984	0.1068	0.1114	0.1405
10		0.3491	0.1187	0.1414	0.1287	0.1220	0.1631
11		0.3345	0.1228	0.0778	0.1598	0.1462	0.1712
12		0.4475	0.2752	0.1748	0.2158	0.2802	0.2800
$\overline{\Delta}_{integr}$		0.3505	0.1708	0.1231	0.1528	0.1650	0.1887
$\sigma_{\Delta_{integr}}$		0.0731	0.0729	0.0435	0.0473	0.0782	0.0622
$\overline{\Delta}_{integr}$	Все распределения	0.2265	0.1403	0.0845	0.1034	0.1299	0.1812
$\sigma_{\Delta_{integr}}$		0.1245	0.0691	0.0397	0.0645	0.0901	0.1248

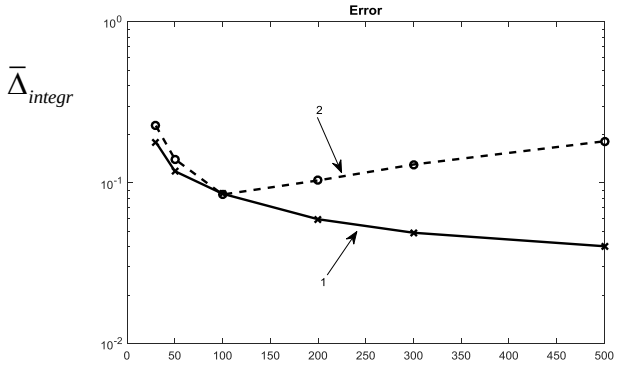

Рисунок 5. Усредненная по набору распределений интегральная погрешность $\overline{\Delta}_{integr}$ в зависимости от числа элементов выборки N. 1 — метод Розенблатта-Парзена; 2 — метод структурной минимизации риска

Из рис. 4–5 и табл. 5–8 видно:

– интегральная погрешность в методе Розенблатта-Парзена в среднем меньше, чем погрешность в методе структурной минимизации риска;

- разница в точности увеличивается с увеличением размера выборки анализируемой случайной величины;
- при малых размерах выборки (30 элементов) разница в погрешности является несущественной;
- метод структурной минимизации риска на 1–2 порядка быстрее производит вычисления. Данный результат объясняется тем, что в методе структурной минимизации риска задача поиска оптимальных параметров сводится к задаче решения системы линейных уравнений.
- независимо от вида распределений среднее время вычисления в методе Розенблатта-Парзена и в методе структурной минимизации риска увеличивается с увеличением размера выборки;
- независимо от вида распределений СКО времени вычисления в методе Розенблатта-Парзена увеличивается с увеличением размера выборки;
- модальность распределения в обоих методах не влияет на время вычисления;
- независимо от вида распределения интегральная погрешность в методе Розенблатта-Парзена уменьшается с увеличением размера выборки;
- интегральная погрешность в методе Розенблатта-Парзена и в методе структурной минимизации риска увеличивается с увеличением модальности распределения. Подобный результат можно объяснить увеличением «сложности» модели при возрастании количества составляющих распределений.

4. Выводы

Сравнительный анализ аппроксимации Розенблатта-Парзена и метода структурной минимизации риска позволяет сделать следующие выводы:

1. Реализован метод аппроксимации плотности вероятности случайной величины на основе метода структурной минимизации риска.

2. Проведены вычислительные эксперименты, подтверждающие эффективность метода структурной минимизации риска для различных видов распределений.

3. Получены оценки точности аппроксимации метода Розенблатта-Парзена и метода структурной минимизации риска в виде интегрального показателя, характеризующего в целом качество оценки плотности вероятности случайной последовательности.

4. Получены оценки времени вычисления метода Розенблатта-Парзена и метода структурной минимизации риска.

5. Получены оценки времени и точности вычисления методов в разрезе видов распределений: одномодальных, двумодальных, трехмодальных.

6. Для малых размеров выборки целесообразно использовать метод структурной минимизации риска, т. к. погрешность вычисления оказывается примерно одного порядка с методом Розенблатта-Парзена, а время вычислений, соответственно, на порядок меньше.

7. Для больших размеров выборки для получения более высокой точности аппроксимации целесообразно использовать метод Розенблатта-Парзена. Однако с увеличением размера выборки существенно увеличивается время вычислений.

Литература

[1] *Крамер Г.* Математические методы статистики. — М. : Мир, 1975.

[2] *Симахин В. А.* Робастные непараметрические оценки: адаптивные оценки взвешенного максимального правдоподобия в условиях статистической априорной неопределенности. — Saarbrucken, Germany : LAP LAMBERT Academic Publishing Gmb H&Co. KG, 2011.

[3] *Сызранцев В. Н., Невелев Я. П., Голофаст С. Л.* Расчет прочностной надежности изделий на основе методов непараметрической статистики. — Новосибирск : Наука, 2008.

[4] *Поршнев С. В., Овечкина Е. В., Каплан В. Е.* Теория и алгоритмы аппроксимации эмпирических зависимостей и распределений. — Екатеринбург : УрО РАН, 2006.

[5] *Эйнштейн А., Смолуховский М.* Брауновское движение : сб. статей. — Л. : ОНТИ — Гл. ред. общетех. лит., 1936.

[6] *Тарасенко Ф. П.* Непараметрическая статистика. — Томск : Изд-во Томского ун-та, 1976.

[7] *Уилкс С.* Математическая статистика. — М. : Наука, 1967.

[8] *Холлендер М.* Непараметрические методы статистики. — М. : Финансы и статистика, 1983.

[9] *Боровков А. А.* Математическая статистика. — М. : Наука, 1984.

[10] *Поршнев С. В., Копосов А. С.* Использование аппроксимации Розенблатта-Парзена для восстановления функции распределения дискретной случайной величины // *В мире научных открытий*. 2013. № 10 (46). С. 235–260.

[11] *Поршнев С. В., Копосов А. С.* Использование аппроксимации Розенблатта-Парзена для восстановления функции распределения непрерывной случайной величины с ограниченным одномодальным законом распределения // *Научный журнал КубГАУ*. 2013. № 08 (092). (http://ej.kubagro.ru/2013/08/pdf/76.pdf)

[12] *Поршнев С. В., Копосов А. С.* Методика оценивания параметров случайной величины со смешанным двумодальным законом распределения на основе совместного использования аппроксимации Розенблатта-Парзена, метода мнимых источников и

генетических алгоритмов // *Фундаментальные исследования*. 2014. № 8. Ч. 3. С. 583–589.

[13] *Поршнев С.В., Копосов А.С.* Аналитическое исследование особенностей случайных блужданий броуновской частицы в ограниченной области рассеяния // *Фундаментальные исследования*. 2013. № 4. Ч. 1. С. 57–64.

[14] *Поршнев С. В., Копосов А. С.* О выборе математических моделей распределений ограниченных случайных последовательностей // *Научный журнал КубГАУ*. 2012. № 10 (84). http://ej.kubagro.ru/2012/10/pdf/53.pdf

[15] *Вапник В. Н. и др.* Алгоритмы и программы восстановления зависимостей / под ред. В. Н. Вапника — М. : Наука, Гл. ред. физ.-мат. лит-ры, 1984.

[16] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М. : Наука, 1979.

Авторы:

Сергей Владимирович Поршнев — доктор технических наук, директор Учебно-научного центра «Информационная безопасность», Уральский Федеральный Университет имени первого Президента России Б. Н. Ельцина

Александр Сергеевич Копосов — кандидат технических наук, главный инженер по разработке Управления ИТ «Дом КлиК» ПАО Сбербанк России

Екатерина Игоревна Березовик — студент департамента информационных технологий и информатики, Уральский Федеральный Университет имени первого Президента России Б. Н. Ельцина

Comparative analysis of Rosenblatt-Parzen method and structural risk minimization method for approximation of the probability density functions of random variables

S. V. Porshnev, A. S. Koposov, E. I. Berezovik
Ural Federal University named after the first President of Russia B. N. Yeltsin
19, Mira St., Yekaterinburg, Russia 620002

e-mail: s.v.porshnev@urfu.ru, alexkopas@gmail.com, miss.berezovik@mail.ru

Abstract. In this article are considered the results of comparative analysis of Rosenblatt-Parzen approximation (ARP) and structural risk minimization (SRM) for approximation of probability density of random variables with a bounded scattering region problem. Two approaches to this problem are known: parametric and non-parametric. In accordance to the first approach based on a priori information choose the type of random variable distribution function (DF), which depends on set of parameters, and measure of proximity between theoretical and experimental distribution functions. Non-parametric statistics is based on the approach that allows getting adaptive assessments of empirical DF as some functionalities which do not depend on chosen type of DF based on a priori information. The method density distribution recovery of experimental sample in ARP is based on an assumption that DF is assessed locally in each point using elements of training set from some area of this point. And in this general DF is some linear combination of known nuclear functions. The assessment of density distri-

bution (DD) in SRM method is counted as a type of decomposition using system of trigonometrical functions. The random variables with one- two- and three-modules probability density were used for comparative analysis. The value of integrated error was used for assessment of approximation quality of analysed methods. The assessments of approximation accuracy and calculation time of DD were found via both methods. Summary tables of approximation accuracy and calculation time of DD were created for analysis. It was formulated conclusions about benefits and disadvantages of each method. It was suggested some recommendations for using this or those methods depending on size of source sample.
Keywords: distribution function, probability density, bounded scattering region problem, nonparametric statistics, empirical distribution function, Rosenblatt-Parzen approximation, fuzziness parameter, nuclear function, structural risk minimization method.

References

[1] *Kramer G.* (1975) *Matematicheskiye metody statistiki.* Moscow, Mir. [In Rus]
[2] *Simakhin V. A.* (2011) Robastnyye neparametricheskiye otsenki: adaptivnyye otsenki vzveshennogo maksimal'nogo pravdopodobiya v usloviyakh statisticheskoy apriornoy neopredelennosti. Saarbrucken, Germany, LAP LAMBERT [In Rus]
[3] *Syzrantsev V. N., Nevelev Ya. P., Golofast S. L.* Raschet prochnostnoy nadezhnosti izdeliy na osnove metodov neparametricheskoy statistiki. Novosibirsk, Nauka, 2008. [In Rus]
[4] *Porshnev S. V., Ovechkina Ye. V., Kaplan V. Ye.* (2006) Teoriya i algoritmy approksimatsii empiricheskikh zavisimostey i raspredeleniy. Yekaterinburg, UrO RAN. [In Rus]
[5] *Eynshteyn A., Smolukhovskiy M.* (1936) Braunovskoye dvizheniye : sb. statey. Leningrad, ONTI — Glavnaya redaktsiya obshchetekhnicheskoy literatury. [In Rus]
[6] *Tarasenko F. P.* (1976) Neparametricheskaya statistika. Tomsk : Izd-vo Tomskogo gos. un-ta. [In Rus]
[7] *Uilks S.* (1967) Matematicheskaya statistika. Moscow, Nauka. [In Rus]
[8] *Khollender M.* (1983) Neparametricheskiye metody statistiki. Moscow, Finansy i statistika. [In Rus]
[9] *Borovkov A. A.* (1984) Matematicheskaya statistika. Moscow, Nauka. [In Rus]
[10] *Porshnev S. V., Koposov A. S.* (2013) *V mire nauchnykh otkrytiy*, **10**(46):235–260. [In Rus]
[11] *Porshnev S. V., Koposov A. S.* (2013) *Nauchnyy zhurnal KubGAU*, 08:76 [In Rus]
[12] *Porshnev S. V., Koposov A. S.* (2014) *Fundamental'nyye issledovaniya*, (8):583–589. [In Rus]
[13] *Porshnev S. V., Koposov A. S.* (2013) *Fundamental'nyye issledovaniya*, (4):57–64. [In Rus]
[14] *Porshnev S. V., Koposov A. S.* (2012) *Nauchnyy zhurnal KubGAU*, 10:53 [In Rus]
[15] *Vapnik V. N. i dr.* (1984) Algoritmy i programmy vosstanovleniya zavisimostey. Moscow. [In Rus]
[16] *Vapnik V. N.* (1979) Vosstanovleniye zavisimostey po empiricheskim dannym. Moscow. [In Rus]