

ЧИСЛЕННЫЕ ОПЕРАЦИИ НАД ЯДЕРНЫМИ ОЦЕНКАМИ В ЗАДАЧЕ ВОССТАНОВЛЕНИЯ ФУНКЦИИ ПЛОТНОСТИ ВЕРОЯТНОСТИ

Б. С. Добронез, О. А. Попова

*Сибирский федеральный университет, Институт космических и информационных технологий,
660074, Красноярск*

УДК 519.24

В статье рассматриваются новые подходы к повышению точности ядерных оценок и численные операции над плотностями вероятности. Для повышения точности оценок функции плотности вероятности применяется экстраполяция Ричардсона. Используя правило Рунге, рассмотрен алгоритм оценок вторых производных функции плотности вероятности. Показано, что использование экстраполяции Ричардсона, позволяет построить сплайны, аппроксимирующие функции плотности вероятности. Рассмотрена численная реализация арифметических операций над функциями плотности вероятности.

Ключевые слова: Численный вероятностный анализ, повышение точности, ядерные оценки, сплайны, арифметические операции, сглаживание.

Введение

Задача восстановления неизвестной функции плотности вероятности и оценка ее производных по эмпирическим данным с использованием численных методов является одним из наиболее важных вопросов в области прикладных исследований [2, 3, 9]. Важно знать оценку для математического ожидания нормы погрешности построенной эмпирической функции плотности вероятности [10].

Оценка производных функция плотности вероятности имеет важное значение для приложений. Подробный обзор существующих методов оценки производных и библиография представлены в [7]. Использование оценок вторых производных дает возможность получить реалистичные оценки математических ожиданий оценки погрешности восстановления функция плотности вероятности в норме l_2 . Знание этих оценок позволяет рассчитать оптимальный параметр сглаживания h [7].

Одно из первых правил для практической оценки погрешности было предложено К. Рунге в начале XX века. Это правило широко использовалось сначала в области квадратурных вычислений, затем в разностных методах и методе конечных элементов. Это правило основано на декомпозиции приближенного решения u^h как суммы [1]

$$u^h = u + h^k v + O(h^{k+m}), \quad (1)$$

где u есть искомое точное решение, v — неизвестная функция и h — малый параметр дискретизации, который чаще всего рассматривается, как шаг разностной сетки. Целое значение k характеризует порядок точности приближенного решения, и $m > 0$ характеризует порядок точности приближенного решения членом погрешности $h^k v$. Поскольку u и v не зависят от h , для параметра $h/2$ справедливо разложение:

$$u^{h/2} = u + \left(\frac{h}{2}\right)^k v + O(h^{k+m}). \quad (2)$$

Вычтем его из (1) избавляясь от u :

$$u^h - u^{h/2} = v \left(\frac{h}{2}\right)^k (2^k - 1) + O(h^{k+m}).$$

Отсюда можно определить главный член погрешности:

$$u^{h/2} - u \approx \frac{u^h - u^{h/2}}{2^k - 1}. \quad (3)$$

Поскольку в формуле (3) отброшен остаточный член порядка $O(h^{k+m})$, то полученное выражение не приводит к гарантированной оценке, но при достаточно малых h дает представление о величине погрешности численного решения.

Экстраполяция Ричардсона является общим методом для получения результатов высокой точности по формулам низкого порядка [1].

Объединим две аппроксимации таким способом, чтобы исключить ошибку порядка h^k . Умножая (2) на 2^k и вычитая из (1), получаем

$$u = \frac{2^k}{2^k - 1} u^{h/2} - \frac{1}{2^k - 1} u^h + O(h^{k+m}).$$

В статье рассматривается подход к повышению точности восстановления функции плотности вероятности на основе эмпирических данных. Подход основан на использовании экстраполяции Ричардсона, для этого используются линейные комбинации ядерных оценок с разными параметрами сглаживания h . Применение правила Рунге позволяет строить оценки второй производной от функции плотности вероятности. В отличие от известных методов, этот подход не требует дифференцирования ядерных оценок или вычисления конечных разностей от эмпирической функции плотности вероятности.

Числовой пример подтвердил теоретические обоснования и показал хорошее качество представленного подхода.

1 Ядерные оценки функции плотности вероятности

Для оценки функция плотности вероятности часто используются непараметрические методы. Отметим, что вплоть до середины 50-х годов в качестве единственного подхода для построения непараметрической оценки функция плотности вероятности использовалась гистограмма. Первые важные результаты в области применения ядерных оценок для функция плотности вероятности были получены в работах М. Розенблатта, Э. Парзена и Н. Ченцова.

В общем виде ядерная оценка может быть записана в виде

$$\hat{f}^h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - \xi_i}{h}\right) = \frac{1}{Nh} \sum_{i=1}^N K_h(x - \xi_i),$$

где $K_h(t) = K(t/h)/h$.

Обозначим

$$K_h(x, \xi_i) = K\left(\frac{x - \xi_i}{h}\right).$$

где ξ случайная величина с функцией плотности вероятности $f(x)$.

Тогда математическое ожидание

$$M[\hat{f}^h(x)] = M[K_h(x, \xi)]$$

и

$$\sigma_N = D[\hat{f}^h(x)] = \frac{1}{N} D[K_h(x, \xi)].$$

Значение математического ожидания можно записать, как

$$M[K_h(x, \xi)] = \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x - t}{h}\right) f(t) dt = \int_{-\infty}^{\infty} K(\eta) f(x - h\eta) d\eta.$$

Заметим, что

$$f(x - h\eta) = f(x) - hf'(x)\eta + \frac{h^2}{2} f''(x)\eta^2 + \frac{h^3}{6} f^{(3)}(x)\eta^3 + O(h^4).$$

$$\mathbb{M}[K_h(x, \xi)] = f(x) \int_{-\infty}^{\infty} K(\eta) d\eta - hf'(x) \int_{-\infty}^{\infty} \eta K(\eta) d\eta + \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} K(\eta) \eta^2 d\eta + \frac{h^3}{6} f^{(3)}(x) \int_{-\infty}^{\infty} K(\eta) \eta^3 d\eta + O(h^4).$$

Пусть ядро K удовлетворяет требованиям

$$\int_{-\infty}^{\infty} K(\eta) d\eta = 1, \quad \int_{-\infty}^{\infty} \eta K(\eta) d\eta = 0 \quad \text{и} \quad \int_{-\infty}^{\infty} \eta^3 K(\eta) d\eta = 0.$$

Обозначим

$$\int_{-\infty}^{\infty} \eta^2 K(\eta) d\eta = \sigma_K^2.$$

Тогда

$$\mathbb{M}[\hat{f}^h(x)] = \mathbb{M}[K_h(x, \xi)] = f(x) + \sigma^2 h^2 f''(x)/2 + O(h^4)$$

и

$$\mathbb{M}[\hat{f}^h(x) - f(x)] = \sigma^2 h^2 f''(x)/2 + O(h^4).$$

Определим $f^h(x)$

$$f^h(x) = \mathbb{M}[\hat{f}^h(x)] = f(x) + \sigma^2 h^2 f''(x)/2 + O(h^4). \quad (4)$$

и $f^{2h}(x)$

$$f^{2h}(x) = \mathbb{M}[\hat{f}^{2h}(x)] = f(x) + 4\sigma^2 h^2 f''(x)/2 + O(h^4). \quad (5)$$

Далее оценим

$$\begin{aligned} \mathbb{D}[K_h(x, \xi)] &= \mathbb{M}[(\frac{1}{h} K(\frac{x-\xi}{h}))^2] - (\mathbb{M}[\frac{1}{h} K(\frac{x-\xi}{h})])^2. \\ \mathbb{M}[(\frac{1}{h} K(\frac{x-\xi}{h}))^2] &= \frac{1}{h} \int_{-\infty}^{\infty} K^2(\eta) f(x - h\eta) d\eta = \frac{f(x)}{h} \int_{-\infty}^{\infty} K^2(\eta) d\eta + \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} K^2(\eta) \eta^2 d\eta + O(h^4). \end{aligned}$$

В итоге получим

$$\sigma^2(x) = \mathbb{D}[\hat{f}(x)] = \frac{f(x)}{Nh} \|K\|_2^2 + \frac{f(x)^2}{N} + O(h/N).$$

Без ограничения общности, можно записать, что [7]

$$\mathbb{M}[(\hat{f}^h(x) - f(x))^2] = \sigma^2(x) + f''(x)^2 h^4/4 + O(h^6)$$

и

$$\sigma_I^2 = \int_{-\infty}^{\infty} \sigma^2(x) dx = \frac{\|K\|_2^2}{Nh} - \frac{\|f\|_2^2}{N} + O(h/N).$$

$$\mathbb{M}[\|\hat{f}^h - f\|_2^2] = \sigma_I^2 + \frac{\|f''\|_2^2 h^4}{4} + O(h^6) \quad (6)$$

2 Экстраполяция Ричардсона

Применим экстраполяцию Ричардсона к $f^h(x)$ и $f^{2h}(x)$. Далее умножим (4) на 1/4 и вычтем результат из (5). Исключая $\sigma^2 h^2 f''(x)/2$ из (4) и (5), мы получим

$$f(x) = \frac{4}{3} f^h(x) - \frac{1}{3} f^{2h}(x) + O(h^4).$$

Заметим, что мы построили приближение к функции $f(x)$

$$f_{cor}^h(x) = \frac{4}{3} \hat{f}^h(x) - \frac{1}{3} \hat{f}^{2h}(x). \quad (7)$$

с точностью $O(h^4)$.

С другой стороны, применяя правило Рунге, мы можем получить оценку

$$f''(x) = 2(f^h(x) - f^{2h}(x))/(3\sigma h^2) + O(h^2)$$

или

$$\|\hat{f}''\| = \frac{2}{3\sigma h^2} \|\hat{f}^h - \hat{f}^{2h}\| \quad (8)$$

3 Аппроксимация сплайнами

Как вытекает из (7), точность построенных оценок определяется из соотношения

$$M[f(x) - \hat{f}_{cor}(x)] = O(h^4),$$

что значительно точнее стандартных ядерных оценок. Важно отметить, что прямые вычисления $\hat{f}_{cor}(x)$ в произвольной точке x требуют значительных вычислительных затрат.

Рассмотрим вопрос построения сплайна s , аппроксимирующего $\hat{f}_{cor}(x)$, так чтобы выполнялась оценка

$$\|f - s\| \leq Ch^4.$$

Для этих целей построим в области $[a, b]$ носителя функции плотности вероятности $f(x)$ сетки

$$\omega_z = \{z_i = a + ih_z, i = 0, \dots, N_z\}, \quad \omega_x = \{x_i = a + ih_x, i = 0, \dots, N_x\}.$$

На сетке ω_z вычислим значения $f_i = \hat{f}_{cor}(z_i)$. Сплайн s будем строить на сетке ω_x . Краевые условия выберем следующим образом $s(a) = 0, s'(a) = 0, s(b) = 0, s'(b) = 0$

$$\sum_{i=1}^{N_z} (s(z_i) - f_i)^2 \rightarrow \min. \quad (9)$$

В случае кубических сплайнов, как классических так, и эрмитовых, задача (9) сводится к решению пяти-диагональной системы линейных алгебраических уравнений.

Для кубических сплайнов справедлива следующая оценка

$$\|f^\nu - s^\nu\| \leq Kh_x^{4-\nu} \|f^{(4)}\|, \quad (10)$$

где K — константа, не зависящая от h_x .

Задачу можно упростить, если вычислить в узлах сетки ω_x значения \hat{f}_i . Для этих целей будем использовать значения $\hat{f}_{cor}(z_i)$ и процедуры сглаживания. Например, для классических кубических сплайнов можно использовать метод скользящего среднего, метод взвешенной локальной регрессии, фильтр Савицкого–Голея. Следует стремиться, чтобы выполнялись оценки

$$|\hat{f}_i - f(x_i)| = O(h_x^4).$$

В этом случае построение сплайна сводится к решению трех диагональной системы линейных алгебраических уравнений и будет выполнена оценка (10).

В итоге кубический сплайн на отрезках $[x_{j-1}, x_j]$, $j = 1, \dots, N$ имеет два представления [8]:

$$\begin{aligned} s(x) = & M_{j-1}(x_j - x)^3/(6h_j) + M_j(x - x_{j-1})^3/(6h_j) + \\ & + (f_{j-1} - M_{j-1}h_j^2/6)(x_j - x)/h_j + (f_j - M_jh_j^2/6)(x - x_{j-1})/h_j, \end{aligned} \quad (11)$$

или

$$\begin{aligned} s(x) = & m_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 - m_j(x - x_{j-1})^2(x_j - x)/h_j^2 + \\ & + f_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 + f_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3, \end{aligned} \quad (12)$$

где $M_j = s''(x_j)$, $m_j = s'(x_j)$, $f_j = f(x_j)$.

Для эрмитовых кубических сплайнов в узлах сетки ω_x необходимо вычислить \hat{f}_i и значения производных \hat{f}'_i . Будем использовать фильтр Савицкого–Голея с кубическими полиномами [6]. В этом случае для построения эрмитовых кубических сплайнов достаточно локальных вычислений. На каждом интервале $[x_{j-1}, x_j]$, $j = 1, \dots, n$ эти сплайны представимы в виде

$$s(x) = f_{j-1}v((x - x_{j-1})/h_j) + f'_{j-1}w((x - x_{j-1})/h_j) + f_jv((x - x_j)/h_j) + f'_jw((x - x_j)/h_j),$$

где $v(x) = (|x| - 1)^2(2|x| + 1)$; $w(x) = x(|x| - 1)^2$.

4 Арифметические операции над плотностями вероятности

Реализация арифметических операций над двумя случайными величинами \mathbf{x} , \mathbf{y} основана на работе с их совместной функцией плотности вероятности $p(x, y)$.

Известны аналитические формулы для определения функции плотности вероятности результатов арифметических действий над случайными величинами. Например, для нахождения плотности вероятности $p_{x_1+x_2}$ суммы двух случайных величин $\mathbf{x}_1 + \mathbf{x}_2$ используется соотношение [11]

$$p_{x_1+x_2}(x) = \int_{-\infty}^{\infty} p(x-v, v)dv = \int_{-\infty}^{\infty} p(v, x-v)dv. \quad (13)$$

Плотность вероятности p_{x_1/x_2} частного двух случайных величин $\mathbf{x}_1/\mathbf{x}_2$ определяется выражением

$$p_{x_1/x_2}(x) = \int_0^{\infty} vp(xv, v)dv - \int_{-\infty}^0 vp(v, xv)dv. \quad (14)$$

Плотность вероятности $p_{x_1x_2}$ произведения двух случайных величин $\mathbf{x}_1\mathbf{x}_2$ представляется соотношением

$$p_{x_1x_2}(x) = \int_0^{\infty} (1/v)p(x/v, v)dv - \int_{-\infty}^0 (1/v)p(v, x/v)dv. \quad (15)$$

В случае, когда случайные величины \mathbf{x} , \mathbf{y} являются независимыми и имеют плотности вероятности вероятности, представленные кубическими сплайнами, s_x , s_y . Совместную плотность вероятности можно представить в виде произведения $p(x, y) = s_x s_y$. Поскольку кубический сплайн на каждом отрезке сетки представляет кубический полином, то $p(x, y)$ в случае вычисления интегралов (13)–(15) будет кусочно-полиномиальной функцией шестой степени. Наиболее удобными в этом случае будут квадратуры Гаусса с четырьмя внутренними узлами, которые точны на полиномах седьмой степени.

В качестве примера, рассмотрим построение сплайна, аппроксимирующего $p_{x_1+x_2}$. Для этих целей в области носителя $p_{x_1+x_2}$ построим сетку $\omega = \{x_0, x_1, \dots, x_n\}$ и вычислим значения $f_i = p_{x_1+x_2}(x_i)$. Используя значения f_i и краевые условия $s'(x_0) = 0$, $s'(x_n) = 0$ на сетке ω построим кубический сплайн s . В этом случае справедлива оценка

$$\|p_{x_1+x_2}^{(\nu)} - s^{(\nu)}\| \leq Kh^{4-\nu} \|p_{x_1+x_2}^{(4)}\|, \quad \nu = 1, 2, 3.$$

Далее вычислим

$$\text{norm} = \int s(x)dx,$$

если $\text{norm} \neq 1$, то $s(x) := s(x)/\text{norm}$.

В случае, когда случайные величины \mathbf{x} , \mathbf{y} являются зависимыми, совместную функцию плотности вероятности необходимо вычислять отдельной процедурой.

5 Численные примеры

В качестве примера, мы рассмотрим задачу восстановления функции плотности вероятности случайной величины ξ суммы четырех равномерно распределенных на $[0, 1]$ случайных величин с использованием метода ядерных оценок.

Заметим, что функция плотности вероятности суммы n равномерно распределенных случайных величин может быть записана в виде

$$p_n(x) = \frac{1}{(n-1)!} (x^{n-1} - C_n^1(x-1)^{n-1} + C_n^2(x-2)^{n-1} - \dots) \quad (16)$$

где C_n^k коэффициенты биномиального разложения. При каждом фиксированном значении аргумента x суммы в скобках приведены только для тех членов, для которых величина $(x-k)$, $k = 1, 2, \dots$ неотрицательна.

Так, для $n = 4$ имеем:

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

На рис. 1 представлен численный пример оценки функции плотности вероятности случайной величины ξ , размерность выборки $N = 10^4$. Сплошная линия представляет точную функцию плотности вероятности $p(x)$, часть (а): “о” соответствует эмпирической функции плотности вероятности \hat{f}^h , $h = 0.1$, часть (б): “о” представляет f_{cor}^h восстановленную эмпирическую функцию плотности вероятности с использованием экстраполяции Ричардсона, часть (с): “о” — результат сглаживания f_{cor}^h .

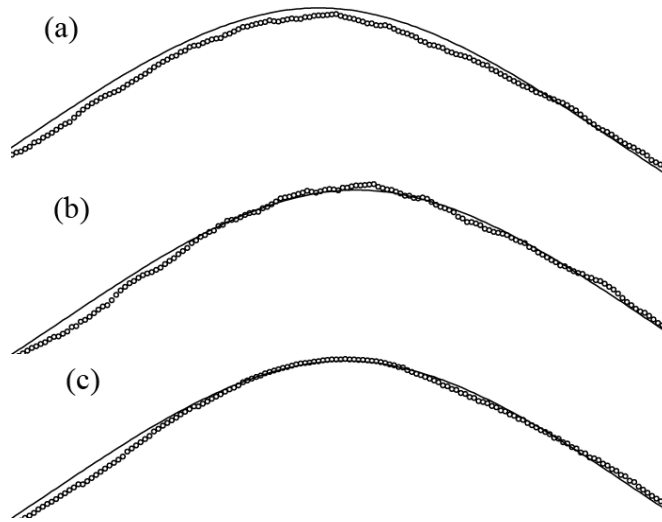


Рис. 1: (а): о — \hat{f}^h , (б): о — f_{cor}^h , (с): о — сглаживание f_{cor}^h

Таблица 1: Аппроксимация $\|f''\|$

| $N = 10^6$ | | $N = 10^4$ | |
|------------|-----------------|------------|-----------------|
| h | $\ \hat{f}''\ $ | h | $\ \hat{f}''\ $ |
| 0.05 | 1.6432 | 0.2 | 1.625162 |
| 0.1 | 1.6368 | 0.3 | 1.429925 |
| 0.2 | 1.52975 | 0.4 | 1.284578 |

Таблица 1 представляет оценку $\|f''\|$ через (8), точное значение $\|f''\| = 1.6431$.

Таблица 2: Уточнение \hat{f} по формуле (7), $N = 10^6$

| h | $\ \hat{f}^h - f\ $ | $\ f_{cor}^h - f\ $ |
|------|---------------------|---------------------|
| 0.3 | 0.01039296 | 0.00198244 |
| 0.35 | 0.01302775 | 0.00172713 |
| 0.4 | 0.01654605 | 0.00215959 |

Таблица 2 показывает сравнение математического ожидания ошибки до и после уточнения.

Заключение

Рассмотрены подходы повышения точности аппроксимации функции плотности вероятности по эмпирическим данным. Эти подходы основаны на правиле Рунге и экстраполяции Ричардсона. Используя правило Рунге построены оценки вторых производных функции плотности вероятности. Это позволяет повысить точность оценок функции плотности вероятности на два порядка h . Знание оценок вторых производных позволяет выбрать оптимальные параметры h для построения ядерных оценок. Для оценки достоверности построенной эмпирической функции плотности вероятности предложена процедура построения апостериорных оценок.

Заметим, что использование ядерных оценок для восстановления функции плотности вероятности в некоторой точке x , требует вычисления значительного числа (сравнимого с размерностью выборки) функций ядер. В тех случаях, когда необходимо частое использование эмпирической функции плотности вероятности, предложен подход основанный на сплайн аппроксимации. В случае использования полиномиальных сплайнов вычисления эмпирической функции плотности вероятности сводятся лишь к вычислению соответствующего полинома.

Дальнейшее развитие данного подхода, предполагается в направлении построения эффективных процедур сглаживания и использования бутстреп технологий.

Список литературы

- [1] Марчук Г.И., Шайдулов В.В. Повышение точности решений разностных схем. М.: Наука, 1979 г. 319 с.
- [2] Dobronets B.S., Popova O.A. Numerical probabilistic analysis under aleatory and epistemic uncertainty // Reliable Computing. 2014. V. 19, P. 274–289.
- [3] Dobronets B., Popova O.: Numerical Probabilistic Approach for Optimization Problems. Scientific Computing, Computer Arithmetic, and Validated Numerics. Lecture Notes in Computer Science 9553. Springer International Publishing, 2016, 43–53.
- [4] Dobronets B.S., Popova O.A. Improving the accuracy of the probability density function estimation // Journal of Siberian Federal University — Mathematics and Physics, V. 10 (1), (2017), 16–21.
- [5] Dobronets B.S., Popova O.A. The numerical probabilistic approach to the processing and presentation of remote monitoring data// Journal of Siberian Federal University — Engineering and Technologies. 2016. V. 9(7), P. 960–971.
- [6] Savitzky A., Golay M. J. Smoothing and differentiation of data by simplified least squares procedures // Anal. Chem. 1964. Vol. 36. P. 1627–1639
- [7] Scott R.W.: Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, New York (2015)
- [8] Альберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. М.: Мир, 1972. 318 с.
- [9] Добронетц Б.С., Попова О.А. Численный вероятностный анализ неопределенных данных. Красноярск: Сибирский федеральный университет, Институт космический и информационных технологий. 2014.
- [10] Попова О.А. Информационный подход к апостериорным оценкам погрешности численного моделирования // Информатизация и связь. 2016. № 2. С. 40–43.
- [11] Гнеденко Б. В. Курс теории вероятностей. М.: Наука, 1988. 448 с.

*Борис Станиславович Добронетц — д.ф.-м.н., профессор, институт космических и информационных технологий,
Сибирский федеральный университет;
e-mail: BDobronets@yandex.ru;*

*Ольга Аркадьевна Попова — к.т.н., доцент, институт космических и информационных технологий,
Сибирский федеральный университет;
e-mail: OlgaArc@yandex.ru.*

Дата поступления — 29 мая 2017 г.