УДК. 004.42

АЛГОРИТМ ВЫБОРА ШИРИНЫ ПОЛОСЫ ПРОПУСКАНИЯ В МЕТОДЕ ЯДЕРНОЙ ОЦЕНКИ ПЛОТНОСТИ

Т. Д. Лыу, аспирант РГРТУ, Рязань, Россия;
orcid.org/0000-0003-3347-1469, e-mail: dattlhp96@gmail.com

Непараметрическая оценка плотности является важным инструментом статистического анализа, но традиционные методы, такие как ядерные оценки случайной величины с правилом нормального соответствия, страдают от чрезмерного сглаживания и недостаточной адаптивности. ***Целью данной работы*** *является модернизация метода выбора ширины полосы пропускания на основе метода подключаемых модулей, который полностью исключает использование правил нормального соответствия, негативно влияющих на производительность методов подключаемых модулей. Полученные результаты показывают более точную оценку плотности по сравнению с правилом нормального соответствия, что подтверждено снижением интегральной квадратичной ошибки до 90 %. Модернизированный метод обеспечивает повышенную точность оценки плотности для сложных распределений без численной оптимизации, сохраняя вычислительную эффективность.*

Ключевые слова: *ширина полосы пропускания, непараметрическая оценка плотности, ядерная оценка плотности.*

DOI: 10.21667/1995-4565-2025-93-41-50

Введение

Непараметрическая оценка плотности представляет собой значимый инструмент в статистическом анализе данных. Непараметрическая оценка может быть применена, в частности, для анализа мультимодальности, асимметрии или иных структурных особенностей распределения данных. Кроме того, данный метод используется в задачах классификации и дискриминантного анализа. Непараметрическая оценка плотности демонстрирует свою эффективность и в вычислительных методах Монте-Карло. В отличие от параметрического подхода, при котором модель задается через ограниченное число параметров с последующей их оценкой на основе принципа максимального правдоподобия, непараметрическая оценка плотности выступает альтернативой. Преимущество непараметрического подхода заключается в значительно большей гибкости при моделировании заданного набора данных, а также в отсутствии влияния смещения, связанного с выбором модели, что отличает его от классического подхода. На сегодняшний день наиболее распространенным методом непараметрической оценки плотности является ядерная оценка плотности (ЯОП) случайной величины [1].

Несмотря на обширный объем литературы, посвященной данной теме, остается множество спорных вопросов, касающихся реализации и практической эффективности ядерных оценок плотности. Во-первых, наиболее популярная техника выбора ширины полосы пропускания на основе метода подключаемых модулей [2] негативно подвержена влиянию так называемого правила нормального соответствия, которое по сути представляет собой построение предварительной нормальной модели данных, от которой зависит производительность метода выбора полосы пропускания. Хотя оценщики, основанные на методе подключаемых модулей, демонстрируют хорошие результаты при приблизительном соблюдении предположения о нормальности, на концептуальном уровне использование правила нормального соответствия подрывает изначальную мотивацию применения непараметрического метода.

Во-вторых, популярный ядерный оценщик плотности с гауссовым ядром не обладает локальной адаптивностью, что часто приводит к высокой чувствительности к выбросам, появ-

лению ложных пиков и общей неудовлетворительной производительности в плане смещения – тенденции к сглаживанию пиков и впадин плотности [3].

В-третьих, большинство ядерных оценщиков страдают от пограничного смещения, например, когда данные неотрицательны, – явление, обусловленное тем, что большинство ядер не учитывают специфические знания о домене данных.

Эти проблемы были в определенной степени смягчены благодаря внедрению более сложных ядер, чем простое гауссово ядро. Ядра более высокого порядка использовались как способ повышения локальной адаптивности и уменьшения смещения, однако они имеют недостатки, заключающиеся в том, что не обеспечивают корректных неотрицательных оценок плотности и требуют большого объема выборки для достижения хорошей производительности. Проблема недостаточной локальной адаптивности была решена введением адаптивных ядерных оценщиков. К ним относятся баллонные оценщики, оценщики ближайших соседей и ядерные оценщики с переменной шириной полосы пропускания, ни один из которых не дает подлинных плотностей, что делает их в некоторой степени неудовлетворительными [3]. Другие предложения, такие как адаптивные оценщики на основе точек выборки, являются вычислительно сложными (быстрое преобразование Фурье не может быть применено), а в некоторых случаях их интеграл не равен единице. Пограничные ядерные оценщики, специально разработанные для устранения пограничного смещения, либо не обладают адаптивностью вдали от границ, либо не приводят к подлинным плотностям [4]. Таким образом, литература изобилует частичными решениями, которые затрудняют создание единой всеобъемлющей структуры для разрешения этих проблем.

Цель данной работы заключается в получении модернизированного алгоритма выбора ширины полосы пропускания на основе метода подключаемых модулей, который полностью исключает использование правил нормального соответствия, который негативно влияет на производительность метода подключаемых модулей. Модернизированный метод подключаемых модулей является подлинно «непараметрическим», поскольку не требует предварительной нормальной модели данных. Более того, наш подход к методу подключаемых модулей не включает числовую оптимизацию, а также незначительно медленнее вычисления по правилу нормального соответствия [4, 5].

Сначала мы описываем ядерный оценщик плотности с гауссовым ядром и объясняем, как его можно рассматривать как частный случай сглаживания с использованием диффузионного процесса. Затем ядерный оценщик плотности с гауссовым ядром используется для обоснования наиболее общей линейной диффузии, которая обладает набором ключевых сглаживающих свойств. Мы анализируем асимптотические свойства полученного оценщика и объясняем, как вычислить асимптотически оптимальную ширину полосы пропускания методом подключаемых модулей. Наконец, практические преимущества модели демонстрируются на примерах моделирования с использованием некоторых хорошо известных наборов данных [3, 6, 7].

Теоретическая часть

При заданных N независимых реализациях $X_N \equiv \{X_1,...,X_N\}$ из неизвестной непрерывной функции плотности вероятности (ФПВ) f на X, ядерный оценщик плотности с гауссовым ядром определяется следующим образом:

$$\widehat{f}(x;h) = \frac{1}{N}\sum_{i=1}^{N}\phi(x, X_i; h), \quad x \in \mathrm{R}, \tag{1}$$

где $\phi(x, X_i; h) = \frac{1}{\sqrt{2\pi h}} e^{-x(x-X_i)^2/(2h)}$ является гауссовой функцией плотности вероятности (ядром) с положением X_i и масштабом $\sqrt{h}$. Масштаб называют шириной полосы пропускания. Многочисленные исследования были сосредоточены на оптимальном выборе h в (1), поскольку эффективность $\widehat{f}$ как оценщика f критически зависит от его значения [2].

Хорошо изученным критерием, используемым для определения оптимального h, является средняя интегральная квадратичная ошибка (MISE):

$$\mathrm{MISE}\left\{\widehat{f}\right\}(h) = \mathrm{E}_f \int \left[\widehat{f}(x;h) - f(x)\right]^2 dx,$$

которая удобно разлагается на компоненты интегрального квадрата смещения и интегральной дисперсии:

$$\mathrm{MISE}\left\{\widehat{f}\right\}(h) = \int \left(\mathrm{E}_f\left[\widehat{f}(x;h)\right] - f(x)\right)^2 dx + \int \mathrm{Var}_f\left[\widehat{f}(x;h)\right] dx,$$

где $\left(\mathrm{E}_f\left[\widehat{f}(x;h)\right] - f(x)\right)^2$ – точечное смещение f, $\mathrm{Var}_f\left[\widehat{f}(x;h)\right]$ – точечная дисперсия f.

Отметим, что операторы математического ожидания и дисперсии применяются к случайной выборке X_N. Средняя интегральная квадратичная ошибка (MISE) зависит от ширины полосы пропускания $\sqrt{h}$ и f довольно сложным образом. Анализ упрощается, если рассматривать асимптотическое приближение к MISE, обозначаемое как AMISE, при условиях согласованности, когда $h = h_N$ зависит от размера выборки N таким образом, что $h_N \to 0$ и $N\sqrt{h_N} \to \infty$ при $N \to \infty$, а f является дважды непрерывно дифференцируемой функцией [2]. Асимптотически оптимальная ширина полосы пропускания представляет собой минимизатор AMISE.

Ключевое наблюдение относительно ядерного оценщика плотности с гауссовым ядром (1) заключается в том, что он является единственным решением диффузионного уравнения в частных производных (УЧП):

$$\frac{\partial}{\partial t}\widehat{f}(x;h) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\widehat{f}(x,h), \qquad x \in \mathrm{X}, h > 0, \tag{2}$$

где X ≡ R, с начальным условием $\widehat{f}(x;0) = \Delta(x)$, причем $\Delta(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - X_i)$ представляет собой эмпирическую плотность данных X_N.

Чтобы проиллюстрировать преимущество формулировки в виде уравнения в частных производных (УЧП) над более традиционной формулировкой (1), рассмотрим случай, когда область данных известна как X ≡ [0,1]. Аналитическое решение этого УЧП в данном случае имеет вид:

$$\widehat{f}(x;h) = \frac{1}{N}\sum_{i=1}^{N} K(x, X_i; h), \qquad x \in [0,1], \tag{3}$$

где ядро K задается как:

$$K(x, X_i; h) = \sum_{k=-\infty}^{\infty} \phi(x, 2k + X_i; h) + \phi(x, 2k - X_i; h), \qquad x \in [0,1], \tag{4}$$

Свойства ядерного оценщика плотности с гауссовым ядром:

Представлены технические детали доказательств свойств диффузионного оценщика. Мы используем обозначение $\|\cdot\|$ для указания евклидовой нормы на R.

Пусть $h = h_N$ таково, что $\lim_{N\to\infty} h_N = 0$ и $\lim_{N\to\infty} N\sqrt{h_N} = \infty$. Предположим, что f'' является непрерывной функцией с интегрируемым квадратом. Интегральный квадрат смещения и интегральная дисперсия ядерного оценщика плотности с гауссовым ядром (1) имеют асимптотическое поведение следующим образом:

$$\left\| E_f\left[\widehat{f}(\cdot;h)\right] - f \right\|^2 = \frac{1}{4}h^2 \left\| f'' \right\|^2 + o(h^2), \qquad N \to \infty, \tag{5}$$

$$\int \mathrm{Var}_f\left[\widehat{f}(x;h)\right] dx = \frac{1}{2N\sqrt{\pi h}} + o\left(\left(N\sqrt{h}\right)^{-1}\right), \qquad N \to \infty. \tag{6}$$

Соответственно асимптотическое приближение первого порядка для MISE, обозначаемое как AMISE, задается следующим образом:

$$\mathrm{AMISE}\left\{\widehat{f}\right\}(h)=\frac{1}{4}h^2\left\|f''\right\|^2+\frac{1}{2N\sqrt{\pi h}}. \tag{7}$$

Асимптотически оптимальное значение h является минимизатором AMISE.

$$h^*=\left(\frac{1}{2N\sqrt{\pi}\left\|f''\right\|^2}\right)^{2/5}, \tag{8}$$

что дает минимальное значение:

$$\mathrm{AMISE}\left\{\widehat{f}\right\}(h^*)=N^{-4/5}\frac{5\left\|f''\right\|^{2/5}}{4^{7/5}\pi^{2/5}}. \tag{9}$$

Алгоритм выбора ширины полосы пропускания

Мы поясним, как оценить ширину полосы пропускания $\sqrt{h^*}$ в (8) для ядерного оценщика плотности с гауссовым ядром (1). Здесь мы представляем новую процедуру выбора ширины полосы пропускания методом подключаемых модулей, основанную на идеях из [2], для достижения хорошей практической производительности. Отличительной особенностью предложенного метода является то, что он не использует правила нормального соответствия и, таким образом, полностью основан на данных.

Из (8) очевидно, что для вычисления оптимального значения h^* для ядерного оценщика плотности с гауссовым ядром (1) необходимо оценить функционал $\left\|f''\right\|^2$.

Таким образом, мы рассматриваем задачу оценки $\left\|f^{(j)}\right\|^2$ для произвольного целого числа $j \geq 1$. Равенство $\left\|f^{(j)}\right\|^2=(-1)^j\mathrm{E}_f[f^{(2j)}(X)]$ предполагает два возможных оценщика методом подключаемых модулей. Первый из них заключается в следующем:

$$(-1)^j\widehat{\mathrm{E}_f f^{(2j)}}:=\frac{(-1)^j}{N}\sum_{k=1}^{N}\widehat{f}^{(2j)}(X_k;h_j)]=\frac{(-1)^j}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\phi^{(2j)}(X_k;X_m;h_j), \tag{10}$$

где $\widehat{f}$ – ядерный оценщик плотности с гауссовым ядром (1). Второй оценщик определяется следующим образом:

$$\begin{aligned}\left\|\widehat{f^{(j)}}\right\|^2:=\left\|\widehat{f}^{(j)}(\cdot;h)\right\|^2&=\frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\int_R\phi^{(j)}(x,X_k;h_j)\phi^{(j)}(x,X_m;h_j)dx\\&=\frac{(-1)^j}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\phi^{(2j)}(X_k;X_m;2h_j),\end{aligned} \tag{11}$$

где $\frac{(-1)^j}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\phi^{(2j)}(X_k;X_m;2h_j)$ представляет собой упрощение, легко вытекающее из того факта, что гауссово ядро ϕ удовлетворяет следующим уравнению Чепмена-Колмогорова [10]:

$$\int_{\mathrm{X}}K(x_1,x_0;h_1)K(x_2,x_1;h_2)dx_1=K(x_2,x_0;h_1+h_2).$$

Для заданной ширины полосы пропускания оба оценщика $(-1)^j\widehat{\mathrm{E}_f f^{(2j)}}$ и $\left\|\widehat{f^{(j)}}\right\|^2$ направлены на оценку одной и той же величины, а именно $\left\|f^{(j)}\right\|^2$. Мы выбираем h_j таким образом, чтобы оба оценщика (10) и (11) были асимптотически эквивалентны в смысле среднеквадратичной ошибки. Другими словами, мы выбираем $h_j=h_j^*$ так, чтобы оба значения

$(-1)^j \widehat{\mathrm{E}_f f^{(2j)}}$ и $\left\| \widehat{f^{(j)}} \right\|^2$ имели одинаковую асимптотическую среднеквадратичную ошибку. Это приводит к следующему Предложение.

Предложение 1. Оценки $(-1)^j \widehat{\mathrm{E}_f f^{(2j)}}$ и $\left\| \widehat{f^{(j)}} \right\|^2$ имеют одинаковую асимптотическую среднеквадратическую ошибку, когда:

$$h_j^* = \left(\frac{1+1/2^{j+1/2}}{3} \frac{1\times 3\times 5\times \ldots \times (2j-1)}{N\sqrt{\pi/2}\left\| f^{(j+1)} \right\|^2} \right)^{2/(3+2j)} . \quad (12)$$

Предложение 1 было доказано в [1].

Так, например:

$$h_2^* = \left(\frac{8+\sqrt{2}}{24} \frac{3}{N\sqrt{\pi/2}\left\| f^{(3)} \right\|^2} \right)^{2/7} . \quad (13)$$

является нашем выбором полосы пропускания для оценки $\left\| f'' \right\|^2$. Мы оцениваем каждый h_j^* по:

$$\hat{h}_j^* = \left(\frac{1+1/2^{j+1/2}}{3} \frac{1\times 3\times 5\times \ldots \times (2j-1)}{N\sqrt{\pi/2}\left\| \widehat{f^{(j+1)}} \right\|^2} \right)^{2/(3+2j)} . \quad (14)$$

Вычисление $\left\| \widehat{f^{(j+1)}} \right\|^2$ требует оценки самого h_{j+1}^*, что, в свою очередь, требует оценки h_{j+2}^* и так далее, как видно из формул (11) и (14). Мы сталкиваемся с задачей оценки бесконечной последовательности $\{h_{j+k}^*, k \geq 1\}$. Однако очевидно, что при заданном h_{l+1}^* для некоторого $l > 0$ мы можем рекурсивно оценить все значения, а затем оценить само h^* на основе (8). Это мотивирует использование селектора ширины полосы пропускания методом подключаемых модулей [1-2], который определяется следующим образом:

1. Для заданного целого числа $l > 0$ оцените h_{l+1}^* с помощью (12) и $\left\| f^{(l+2)} \right\|^2$, вычисленных при предположении, что f является нормальной плотностью с математическим ожиданием и дисперсией, оцененными по данным. Обозначим эту оценку как $\hat{h}_{l+1}^*$.

2. Используйте $\hat{h}_{l+1}^*$ для оценки $\left\| f^{(l+1)} \right\|^2$ с помощью оценщика методом подключаемых модулей (11) и h_l^* с помощью (14). Затем используйте h_l^* для оценки h_{l-1}^* и так далее, пока не получим оценку $\hat{h}_2^*$.

3. Используйте оценку $\hat{h}_2^*$ для вычисления $\hat{h}^*$ из (8).

Таким образом, *l*-стадийный прямой селектор ширины полосы пропускания методом подключаемых модулей включает оценку l функционалов $\{\left\| f^{(j)} \right\|, 2 \leq j \leq l+1\}$ с использованием оценщика методом подключаемых модулей (11). Процедуру можно описать более абстрактно следующим образом. Обозначим функциональную зависимость $\hat{h}_j^*$ от $\hat{h}_{j+1}^*$ в формуле (14) как:

$$\hat{h}_j^* = \gamma_j(\hat{h}_{j+1}^*).$$

Тогда очевидно, что $\hat{h}_j^* = \gamma_j(\gamma_{j+1}(\hat{h}_{j+2}^*)) = \gamma_j(\gamma_{j+1}(\gamma_{j+2}(\hat{h}_{j+3}^*))) = \ldots$. Для упрощения записи мы определяем композицию следующим образом:

$$\gamma^{[k]}(h) = \gamma_1(...\gamma_{k-1}(\gamma_k(h))...), \quad k \geq 1.$$

Анализ формул (14) и (8) показывает, что оценка h^* удовлетворяет следующим условиям:

$$\hat{h}^* = \xi\hat{h}_1^* = \xi\gamma^{[1]}\left(\hat{h}_2^*\right) = \xi\gamma^{[2]}\left(\hat{h}_3^*\right) = \cdots = \xi\gamma^{[l]}\left(\hat{h}_{l+1}^*\right),$$

$$\xi = \left(\frac{6\sqrt{2}-3}{7}\right)^{2/5} \approx 0,9.$$

Тогда для заданного целого числа $l > 0$, l-стадийный прямой селектор ширины полосы пропускания методом подключаемых модулей заключается в вычислении:

$$\hat{h}^* = \xi\gamma^{[l]}(h_{l+1}^*),$$

где h_{l+1}^* оценивается с помощью (12), предполагая, что f в $\left\|f^{(l+2)}\right\|^2$ является нормальной плотностью с математическим ожиданием и дисперсией, оцененными по данным. Наиболее слабым местом этой процедуры является предположение, что истинная f представляет собой гауссову плотность для вычисления $\left\|f^{(l+2)}\right\|^2$. Это предположение может привести к произвольно плохим оценкам h^*, например, когда истинная f значительно отличается от гауссовой. Вместо этого мы предлагаем найти решение нелинейного уравнения:

$$h = \xi\gamma^{[l]}(h), \tag{15}$$

Для некоторого l, используя либо итерацию фиксированной точки, либо метод Ньютона с начальным приближением $h = 0$. Версия с итерацией фиксированной точки формализована в следующем алгоритме.

АЛГОРИТМ 1 (Усовершенствованный метод Шизера-Джонса). Для заданного $l > 2$ выполните следующие шаги:

1. Шаг 1: Инициализируйте с $u_0 = \lambda$, где λ – машинная точность, и $n = 0$;
2. Шаг 2: Задайте $u_{n+1} = \xi\gamma^{[l]}(u_n)$;
3. Шаг 3: Если $|u_{n+1} - u_n| < \lambda$, остановитесь и установите $\hat{h}^* = u_{n+1}$; в противном случае установите $n := n+1$ и повторите с шага 2;
4. Шаг 4: Предоставьте ядерный оценщик плотности с гауссовым ядром (1), вычисленный при $\hat{h}^*$, как окончательную оценку f, и $\hat{h}_2^* = \gamma^{[l-1]}(u_{n+1})$ как ширину полосы пропускания для оптимальной оценки $\left\|f^{''}\right\|^2$.

На рисунке 1 показан блок-схема Алгоритма 1 для выбора оптимальной ширины полосы пропускания.

Численный опыт свидетельствует о следующем. Во-первых, алгоритм фиксированной точки не терпит неудачу в поиске корня уравнения $h = \xi\gamma^{[l]}(h)$. Во-вторых, корень, по-видимому, является уникальным. В-третьих, решения уравнений $h = \xi\gamma^{[5]}(h)$ и $h = \xi\gamma^{[l+5]}(h)$ для любого $l > 0$ не различаются в практически значимом смысле. Другими словами, увеличение стадий правила выбора ширины полосы пропускания сверх $l = 5$ не приносит дополнительных преимуществ. Мы рекомендуем установить $l = 5$. Наконец, численная процедура вычисления $\gamma^{[5]}(h)$ выполняется быстро при реализации с использованием дискретного косинусного преобразования [8].

Метод подключаемых модулей, описанный в Алгоритме 1, демонстрирует превосходящую практическую производительность по сравнению с существующими реализациями метода подключаемых модулей, включая конкретное правило решения уравнения Шизера и Джонса [1-2]. Поскольку мы заимствуем многие плодотворные идеи, описанные в [2], мы называем наш новый алгоритм усовершенствованным методом Шизера-Джонса (УШД).

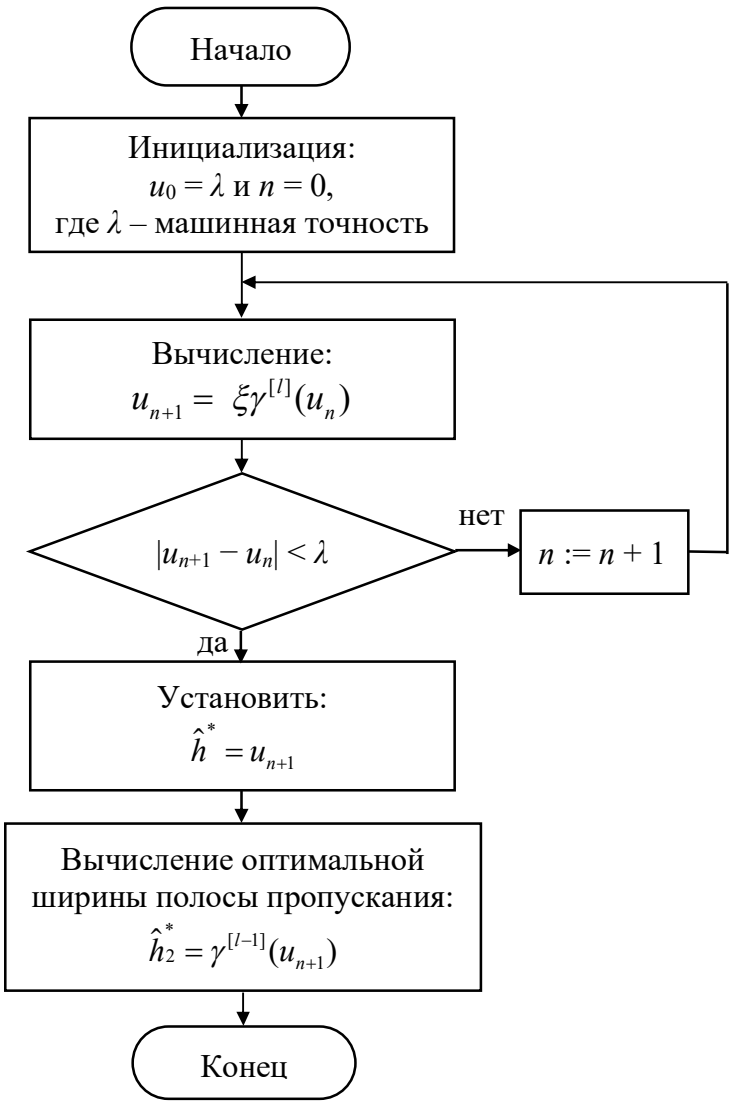

Рисунок 1 – Блок-схема Алгоритма 1 для выбора ширины полосы пропускания
Figure 1 – Flow chart of Algorithm 1 for bandwidth selection

Результаты моделирования

Рассмотрим, например, случай, когда данные выборки представляют собой смесь двух нормальных распределений $N\left(-2,\left(\frac{1}{2}\right)^2\right)$ и $N\left(2,1^2\right)$, что создаёт бимодальное распределение с двумя чётко выраженными пиками при $x = -2$ и $x = 2$.

Такое распределение идеально подходит для проверки производительности методов оценки плотности, таких как новый алгоритм усовершенствованный Шизера-Джонс метод и правило нормального соответствия. Результат моделирования Алгоритма 1 на языке Matlab [9] представлен на рисунке 2.

На рисунке 2 показан сравнение оценок плотности с использованием улучшенного правила Шитера-Джонса (синяя пунктирная линия из точек) и правила нормального соответствия (красная пунктирная линия) на выборке данных с бимодальным распределением $N\left(-2,\left(\frac{1}{2}\right)^2\right)$ и $N\left(2,1^2\right)$. Чёрная сплошная линия представляет истинную плотность, а гистограмма отображает распределение выборки.

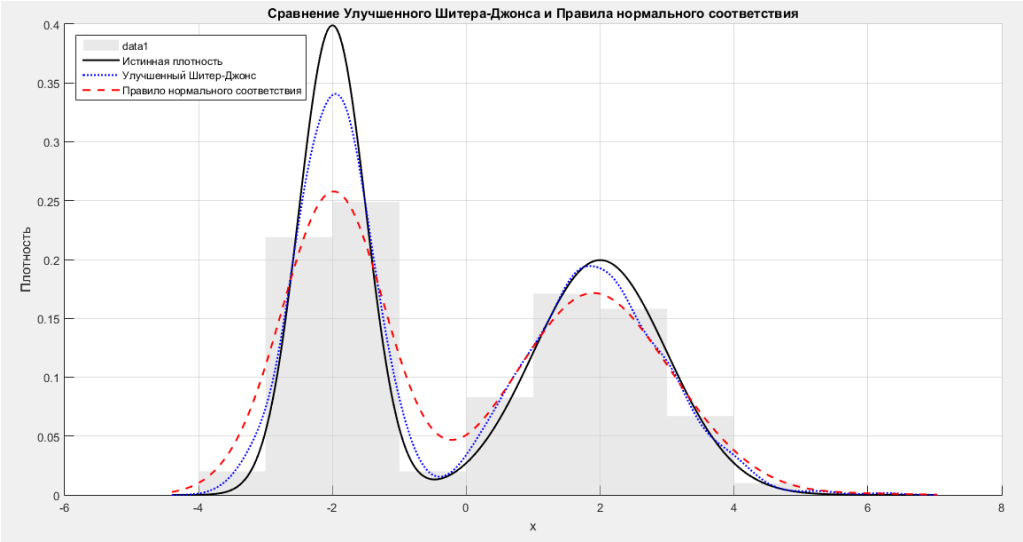

Рисунок 2 – Сравнение оценок плотности с использованием усовершенствованного правила Шизер-Джонса в Алгоритме 1 и исходного правила, использующего правило нормального соответствия
Figure 2 – Comparison of density estimates using the improved Sheather-Jones rule in Algorithm 1 and the original rule using normal reference rule

Результаты моделирования показывают, что улучшенное правило выбора ширины полосы Шитера-Джонса демонстрирует более точную оценку истинной плотности по сравнению с правилом нормального соответствия. На гистограмме данных видно, что выборка имеет бимодальное распределение с двумя выраженными пиками. Улучшенный метод Шитера-Джонса лучше улавливает эти пики, в то время как правило нормального соответствия склонно к чрезмерному сглаживанию, что приводит к потере деталей в структуре плотности.

Эксперименты с правилами нормального соответствия

Результат, показанный на рисунке 2, не является единичным случаем, когда правила нормального соответствия демонстрируют низкую эффективность. Мы провели всестороннее моделирование, чтобы сравнить усовершенствованный метод Шизера – Джонса (УШД) (Алгоритм 1) с исходным алгоритмом Шизера – Джонса (ШД) [1-2].

Таблица 1 представляет средние результаты по 10 независимым испытаниям для ряда различных тестовых случаев. Второй столбец отображает целевую плотность $f(x)$, третий столбец показывает размер выборки, использованный в экспериментах. Последний столбец содержит критерий для сравнения:

$$R = \frac{\left\| \hat{f}(\cdot;\hat{h}^{*}) - f \right\|^{2}}{\left\| \hat{f}(\cdot;h_{ШД}) - f \right\|^{2}},$$

где R – отношение интегральной квадратичной ошибки нового оценщика УШД к интегральной квадратичной ошибке исходного оценщика ШД, $h_{ШД}$ – ширина полосы пропускания, вычисленная с использованием исходного метода Шитера-Джонса [1-2].

Результаты, представленные в Таблице 1, подтверждают превосходство усовершенствованного метода Шитера – Джонса (УШД) над исходным методом Шитера – Джонса (ШД) почти во всех случаях. Наиболее значительное улучшение наблюдается в случае 2 при N=10^6, где значение R = 0,1. Это означает, что интегральная квадратичная ошибка УШД составляет всего 10 % от ошибки ШД, что эквивалентно улучшению на 90% по сравнению с исходным методом. Такой результат подчеркивает высокую эффективность УШД для сложных распределений и подтверждает преимущество отказа от правила нормального соответствия, позволяющего лучше адаптироваться к реальной структуре данных.

Таблица 1 – Результаты по 10 независимым испытаниям для ряда различных тестовых случаев [3]

Table 1 – Results from 10 separate trials across various test scenarios [3]

Случай	f(x)	*N*	*R*
1	$\frac{1}{2}N(0,1)+\sum_{k=0}^{4}\frac{1}{10}N\left(\frac{k}{2}-1,\left(\frac{1}{10}\right)^2\right)$	10^3 10^4	0,72 0,94
2	$\frac{49}{100}N\left(-1,\left(\frac{2}{3}\right)^2\right)+\frac{49}{100}N\left(1,\left(\frac{2}{3}\right)^2\right)+\frac{1}{350}\sum_{k=0}^{6}N\left(k-\frac{2}{3},\left(\frac{1}{100}\right)^2\right)$	10^5 10^6	0,35 0,1
3	$\frac{1}{10}N(0,1)+\frac{9}{10}N\left(0,\left(\frac{1}{10}\right)^2\right)$	10^3 10^5	1,01 1,00
4	$\frac{1}{2}N\left(0,\left(\frac{1}{10}\right)^2\right)+\frac{1}{2}N(5,1)$	10^2 10^3	0,31 0,7
5	$\sum_{k=0}^{7}\frac{1}{8}N\left(3\left(\left(\frac{2}{3}\right)^k-1\right),\left(\frac{2}{3}\right)^{2k}\right)$	10^3 10^4	0,69 0,84
6	$\frac{1}{2}N\left(-12,\frac{1}{4}\right)+\frac{1}{2}N\left(12,\frac{1}{4}\right)$	10^2 10^3	0,33 0,64
7	$\frac{3}{4}N(0,1)+\frac{1}{4}N\left(\frac{3}{2},\left(\frac{1}{3}\right)^2\right)$	10^3 10^4	1,02 1,00
8	$\frac{2}{7}\sum_{k=0}^{2}N\left(\frac{12k-15}{7},\left(\frac{2}{7}\right)^2\right)+\frac{1}{21}\sum_{k=8}^{10}N\left(\frac{2k}{7},\left(\frac{1}{21}\right)^2\right)$	10^3 10^4	0,45 0,27
9	$\frac{2}{3}N(0,1)+\frac{1}{3}N\left(0,\left(\frac{1}{10}\right)^2\right)$	10^2 10^3	0,78 0,93

Примечание: $N(\mu,\sigma^2)$ обозначает гауссову плотность с математическим ожиданием μ и дисперсией σ^2.

Заключение

В данной работе предложен алгоритм выбора ширины полосы пропускания на основе метода подключаемых модулей. Полученные результаты показывают, что улучшенное правило выбора ширины полосы Шитера – Джонса, реализованное в Алгоритме 1, обеспечивает более высокую точность оценки плотности и снижает интегральные квадратичные ошибки до 90 % по сравнению с правилом нормального соответствия.

Таким образом, предложенный метод значительно улучшает точность непараметрической оценки плотности, особенно для сложных распределений, и открывает перспективы для дальнейших исследований в области адаптивных оценок.

Библиографический список

1. **Wand M.P., Jones M.C.** Kernel Smoothing. London: Chapman and Hall. 1995. 226 p.

2. **Sheather S.J., Jones M.C.** A reliable data-based bandwidth selection method for kernel density estimation // Journal of the Royal Statistical Society, Series B. 1991. Vol. 53. Pp. 683-690.

3. **Marron J. S., Wand M. P.** Exact mean integrated error // Annals of Statistics. 1992. Vol. 20. № 2. Pp. 712-736.

4. **Васильев В.И.** Асимптотические свойства ядерных оценок плотности вероятности // Журнал вычислительной математики и математической физики. 2008. Т. 48. № 4. С. 567-575.

5. **Сильверман Б.В.** Оценка плотности для статистики и анализа данных. М.: Мир. 1989. 270 с.

6. **Орлов А.И.** Ядерные оценки плотности в непараметрической статистике // Заводская лаборатория. Диагностика материалов. 2010. Т. 76. № 5. С. 54-60.

7. **Лапко А.В., Лапко В.А.** Оптимизация ширины полосы пропускания в ядерной оценке плотности вероятности // Автоматика и телемеханика. 2015. № 9. С. 145-156.

8. **Horova I., Kolacek J., Zelinka J.** Kernel Smoothing in MATLAB: Theory and Practice of Kernel Smoothing. 2012. 244 p.

9. **Кулаичев А.П.** Ядерные оценки плотности и их реализация в MATLAB // Программные продукты и системы. 2012. № 3. С. 112-118.

10. **Гнеденко Б.В.** Курс теории вероятностей. М.: Наука. 1988. 448 с.

UDC. 004.42

BANDWIDTH SELECTION ALGORITHM BASED ON KERNEL DENSITY ESTIMATION METHOD

T. D. Luu, post-graduate student, RSREU, Ryazan, Russia;
orcid.org/0000-0003-3347-1469, e-mail: dattlhp96@gmail.com

Nonparametric density estimation serves as a key instrument in statistical analysis, however, traditional methods like kernel estimates relying on normal reference rule are hindered by excessive smoothing and limited adaptability. ***The aim of this work*** *is to modernize bandwidth selection method based on plug-in approach which completely eliminates the use of normal reference rules that negatively affect the performance of plug-in methods. The obtained results demonstrate more accurate density estimation compared to normal reference rule as evidenced by a reduction in integrated squared error of up to 90 %. A modernized method provides improved density estimation accuracy for complex distributions without numerical optimization while maintaining computational efficiency.*

Keywords*: bandwidth, nonparametric density estimation, kernel density estimation.*

DOI: 10.21667/1995-4565-2025-93-41-50

References

1. **Wand M.P., Jones M.C.** *Kernel Smoothing*. London: Chapman and Hall, 1995, 226 p.

2. **Sheather S.J., Jones M.C.** A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B.* 1991, vol. 53, pp. 683-690.

3. **Marron J. S., Wand M. P.** Exact mean integrated error. *Annals of Statistics.* 1992, vol. 20, no. 2, pp. 712-736.

4. **Vasiliev V.I.** Asimptoticheskie svoystva yadernykh otsenok plotnosti veroyatnosti. *Zhurnal vychislitel'noy matematiki i matematicheskoy fiziki.* 2008, vol. 48, no. 4, pp. 567-575 (in Russian).

5. **Silverman B.W.** *Otsenka plotnosti dlya statistiki i analiza dannykh* (Density estimation for statistics and data analysis). Moscow: Mir. 1989. 270 p (in Russian).

6. **Orlov A.I.** Yadernye otsenki plotnosti v neparametricheskoy statistike. *Zavodskaya laboratoriya*. *Diagnostika materialov.* 2010, vol. 76, no. 5, pp. 54-60 (in Russian).

7. **Lapko A.V., Lapko V.A.** Optimizatsiya shiriny polosy propuskaniya v yadernoy otsenke plotnosti veroyatnosti. *Avtomatika i telemekhanika.* 2015, no. 9, pp. 145-156 (in Russian).

8. **Horova I., Kolacek J., Zelinka J.** *Kernel Smoothing in MATLAB: Theory and Practice of Kernel Smoothing.* 2012, 244 p.

9. **Kulaichev A.P.** Yadernye otsenki plotnosti i ikh realizatsiya v MATLAB. *Programmnye produkty i sistemy.* 2012, no. 3, pp. 112-118 (in Russian).

10. **Gnedeko B.V.** *Kurs teorii veroyatnostey* (Course in probability theory). Moscow: Nauka. 1988. 448 p. (in Russian).