НЕПАРАМЕТРИЧЕСКИЕ ЯДЕРНЫЕ ОЦЕНКИ НАДАРАЯ – ВАТСОНА И ОБЛАСТЬ ИХ ЗАДАНИЯ

Е. Г. Красногир

Бегорусский государственный университет
Минск, Беларусь
E-mail: krasnahir@bsu.by

В статье рассматривается непараметрическая ядерная оценка Надарая – Ватсона функции регрессии и полученные на ее основе непараметрические оценки функций дрейфа и диффузии диффузионного случайного процесса. Установлено, что они являются работоспособными лишь на интервале, содержащем наблюдаемые значения случайного процесса. Вне данного интервала поведение непараметрических оценок не будет совпадать с поведением оцениваемых функций. Теоретические результаты подтверждаются вычислительными экспериментами.

Ключевые слова: непараметрическая ядерная оценка Надарая – Ватсона, стохастическое дифференциальное уравнение, диффузионный случайный процесс.

Современный этап развития вычислительной техники позволяет эффективно решать все большее количество сложных научных и производственных задач. Однако формализация математической модели с использованием соотношений с точностью до неизвестных параметров для решения данных задач зачастую невозможна, так как формальные связи между причиной и следствием неясны или неизвестны. Данная проблема особенно остро ощущается при изучении явлений стохастического характера. Поэтому все чаще исследователи обращаются к непараметрическим методам анализа, позволяющим по эмпирическим данным сформировать математическую модель явления. Во многих случаях такой подход оказывается единственно возможным. В данной работе в применении к случайным процессам диффузии исследуются непараметрические ядерные оценки, задаваемые всего двумя характеристиками – ядерной функцией (обычно это некоторая плотность вероятностей) и положительным коэффициентом, называемым параметром размытости.

Пусть $\{X_i, Y_i\}$, $i = \overline{1,n}$ – двумерная выборка объема n из двумерного случайного вектора (X,Y), а $m(x) = E[Y|X = x] \in \boldsymbol{R}$, $x \in \boldsymbol{R}$. Предположим, что X и Y имеют совместную плотность и маргинальные плотности. Непараметрической ядерной оценкой Надарая – Ватсона [1], [2] регрессии $m(x)$ называется функция

$$m_h(x) = \frac{\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}, \; x \in \boldsymbol{R}, \tag{1}$$

где $K_h(y) = K(y/h)$, $K(y)$, $y \in \boldsymbol{R}$ – ядро (ядерная функция), $h > 0$ – параметр размытости.

Используя (1), получим непараметрические оценки функций дрейфа и диффузии диффузионного случайного процесса.

Пусть $X_0, X_1, ..., X_n$ – наблюдения стационарного случайного процесса X_t соответственно в моменты времени t_0, $t_1 = t_0 + \Delta$, ..., $t_n = t_0 + n\Delta$, где $\Delta > 0$ – фиксированный временной интервал. X_t порождается стохастическим дифференциальным уравнением

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,$$

где W_t – стандартный одномерный винеровский процесс, $t \geq 0$.

Функция $\mu : \boldsymbol{R} \to \boldsymbol{R}$ называется функцией дрейфа или дрейфом, $\sigma : \boldsymbol{R} \to \boldsymbol{R}_+$ – диффузией (функцией диффузии) ($\boldsymbol{R}_+$ – множество неотрицательных действительных чисел).

Если в (1) взять $Y_i = (X_{i+1} - X_i)/\Delta$, то непараметрическая ядерная оценка функции дрейфа может быть вычислена по формуле [3]

$$\mu_h(x) = \frac{\sum_{i=0}^{n-1}(X_{i+1} - X_i)K_h(x - X_i)}{\Delta\sum_{i=0}^{n-1}K_h(x - X_i)}. \qquad (2)$$

Если в (1) взять $Y_i = (X_{i+1} - X_i)^2/\Delta$, то непараметрическая ядерная оценка функции диффузии может быть вычислена по формуле [3]

$$\sigma_h^2(x) = \frac{\sum_{i=0}^{n-1}(X_{i+1} - X_i)^2 K_h(x - X_i)}{\Delta\sum_{i=0}^{n-1}K_h(x - X_i)}. \qquad (3)$$

Рассмотрим область определения функций $m_h(x)$, $\mu_h(x)$ и $\sigma_h^2(x)$. Очевидно, что они определены для тех x, для которых знаменатель в (1) – (3) не равен нулю.

Если при построении непараметрической оценки используется ядро, принимающее ненулевые значения на конечном интервале (например, ядро Епанечникова [4]

$$K(x) = 3(1 - x^2)I(|x| \leq 1)/4$$

($I(\cdot)$ – индикаторная функция, равная единице при выполнении условий в скобках и нулю в противном случае)), то наибольшим интервалом, на котором знаменатель не равен нулю, будет $(X_{\min} - h,\ X_{\max} + h)$, где $X_{\min}$ и $X_{\max}$ – соответственно минимальное и максимальное значение среди всех наблюдений X_i, $i = \overline{1,n}$. Таким образом, непараметрическую ядерную оценку фактически можно построить только на вышерассмотренном интервале, который обычно является небольшим.

Рассмотрим теперь такие ядра, которые не равны нулю для любого $x \in \boldsymbol{R}$. В этом случае непараметрические оценки (1) – (3) будут определены при $x \in \boldsymbol{R}$. Примером вышеуказанных ядер является гауссовское ядро [5]

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}.$$

Поэтому возникает вопрос, корректно ли рассматривать непараметрические ядерные оценки вида (1) – (3) за пределами интервала $(X_{\min} - h,\ X_{\max} + h)$.

ОГРАНИЧЕННОСТЬ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ ФУНКЦИИ РЕГРЕССИИ ПРИ ИСПОЛЬЗОВАНИИ ГАУССОВСКОГО ЯДРА

Используя гауссовское ядро, рассмотрим поведение определенной на всей числовой прямой непараметрической оценки функции регрессии.

Найдем пределы $m_h(x)$ (1) при $x \to -\infty$ и $x \to +\infty$.

Пусть $x \to -\infty$. Перепишем (1) как

$$m_h(x) = \frac{K_h(x - X_{\min})\sum_{i=1}^{n} Y_i \dfrac{K_h(x - X_i)}{K_h(x - X_{\min})}}{K_h(x - X_{\min})\sum_{i=1}^{n} \dfrac{K_h(x - X_i)}{K_h(x - X_{\min})}} = \frac{\sum_{i=1}^{n} Y_i \exp\left\{-\dfrac{(x - X_i)^2 - (x - X_{\min})^2}{2h^2}\right\}}{\sum_{i=1}^{n} \exp\left\{-\dfrac{(x - X_i)^2 - (x - X_{\min})^2}{2h^2}\right\}} =$$

$$= \frac{\sum_{i=1}^{n} Y_i \exp\left\{\dfrac{(X_i - X_{\min})(2x - X_i - X_{\min})}{2h^2}\right\}}{\sum_{i=1}^{n} \exp\left\{\dfrac{(X_i - X_{\min})(2x - X_i - X_{\min})}{2h^2}\right\}}.$$

Так как $X_{\min}$ – минимальное значение среди всех X_i, $i = \overline{1,n}$, то $X_i - X_{\min} \geq 0$, $i = \overline{1,n}$, и при $x \to -\infty$

$$\exp\left\{\frac{(X_i - X_{\min})(2x - X_i - X_{\min})}{2h^2}\right\} \to I(X_{\min} = X_i).$$

Таким образом,

$$\lim_{x \to -\infty} m_h(x) = \frac{1}{t}\sum_{j \in J} Y_j, \qquad (4)$$

где J – множество индексов i таких, что $X_{\min} = X_i$, t – количество наблюдений, равных $X_{\min}$.

Пусть теперь $x \to +\infty$. Преобразуем (1):

$$m_h(x) = \frac{K_h(x - X_{\max})\sum_{i=1}^{n} Y_i \dfrac{K_h(x - X_i)}{K_h(x - X_{\max})}}{K_h(x - X_{\max})\sum_{i=1}^{n} \dfrac{K_h(x - X_i)}{K_h(x - X_{\max})}} = \frac{\sum_{i=1}^{n} Y_i \exp\left\{-\dfrac{(x - X_i)^2 - (x - X_{\max})^2}{2h^2}\right\}}{\sum_{i=1}^{n} \exp\left\{-\dfrac{(x - X_i)^2 - (x - X_{\max})^2}{2h^2}\right\}} =$$

$$= \frac{\sum_{i=1}^{n} Y_i \exp\left\{-\dfrac{(X_{\max} - X_i)(2x - X_i - X_{\max})}{2h^2}\right\}}{\sum_{i=1}^{n} \exp\left\{\dfrac{(X_{\max} - X_i)(2x - X_i - X_{\max})}{2h^2}\right\}}.$$

Так как $X_{\max}$ – максимальное значение среди всех X_i, $i = \overline{1,n}$, то $X_{\max} - X_i \geq 0$, $i = \overline{1,n}$, и при $x \to +\infty$

$$\exp\left\{-\frac{(X_{\max}-X_i)(2x-X_i-X_{\max})}{2h^2}\right\}\to I(X_{\max}=X_i).$$

Таким образом,

$$\lim_{x\to\infty} m_h(x)=\frac{1}{s}\sum_{j\in G} Y_j\ , \tag{5}$$

где G – множество индексов i таких, что $X_{\max}=X_i$, s – количество наблюдений, равных $X_{\max}$.

Для непараметрической оценки функции дрейфа $\mu_h(x)$ вида (2) $Y_i=(X_{i+1}-X_i)/\Delta$. Поэтому если $i\in J$, то $Y_i\ge 0$, а если $i\in G$, то $Y_i\le 0$. Таким образом, $\lim\limits_{x\to-\infty}\mu_h(x)>0$, а $\lim\limits_{x\to\infty}\mu_h(x)<0$. Для оценки диффузии вида (3) выполняется $\lim\limits_{x\to\pm\infty}\sigma_h^2(x)>0$.

Итак, непараметрические ядерные оценки функций регрессии, дрейфа и диффузии при $x\to\pm\infty$ являются ограниченными, тогда как функции $\mu(x)$ и $\sigma(x)$, вообще говоря, могут быть неограниченными.

ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ: ПОВЕДЕНИЕ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК ДРЕЙФА И ДИФФУЗИИ В ЗАВИСИМОСТИ ОТ ОБЛАСТИ ЗАДАНИЯ (ГАУССОВСКОЕ ЯДРО)

Пусть имеются наблюдения (рисунок 1) процесса X_t ($n=7500$, $\Delta=1/250$), удовлетворяющего стохастическому дифференциальному уравнению

$$dX_t=k(\theta-X_t)dt+\sigma\sqrt{X_t}\,dW_t, \tag{6}$$

где $k=0{,}21459$, $\theta=0{,}08571$, $\sigma^2=0{,}00613$. Данная модель рассматривается в [6] при обсуждении вопроса нелинейности непараметрической ядерной оценки функции дрейфа. $X_{\min}=0{,}021012$, $X_{\max}=0{,}11921$.

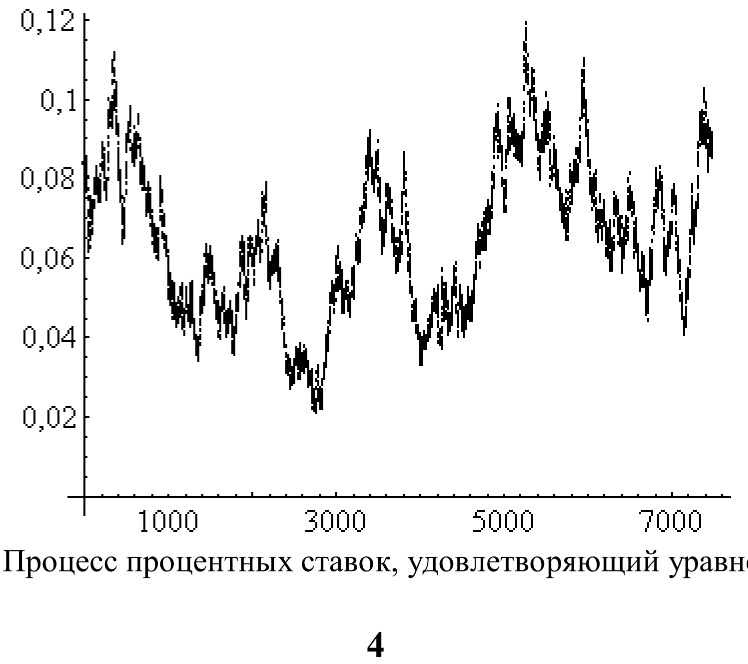

Рис.1. – Процесс процентных ставок, удовлетворяющий уравнению (6)

На рисунке 2 представлен график непараметрической ядерной оценки $\mu_h(x)$ (2) при $h = 0{,}0253$ (здесь и далее h находится с помощью минимизации функции кроссвалидации [5]

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n-1} K_h(X_i - X_j)Y_j}{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n-1} K_h(X_i - X_j)} \right)^2 w(X_i),$$

где $w(x)$ – весовая функция) для $x \in [-13; 4]$, ядро гауссовское. Очевидно, что данная оценка, в отличие от $\mu(x)$, является нелинейной функцией и на краях интервала достигает своих предельных значений, которые согласно (4) и (5) равны соответственно – 0,780614 и 0,252141. На рисунке 3 представлен фрагмент рисунка 2 (нижний график) для $x \in [0; 0{,}2]$ (этот промежуток рассматривается в [6]) в сравнении с истинной линейной функцией (верхний график).

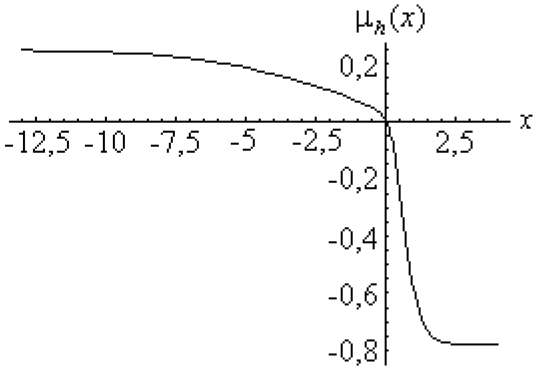

Рис.2. – Непараметрическая оценка функции дрейфа при использовании гауссовского ядра

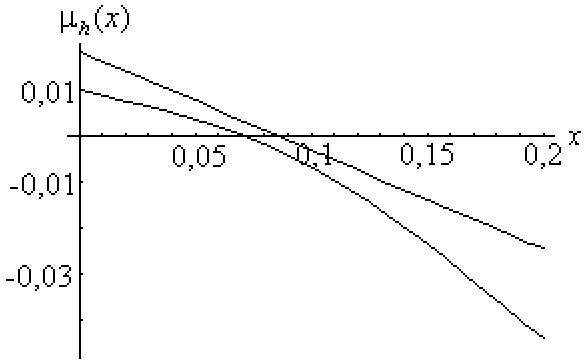

Рис.3. Фрагмент рисунка 1 (нижняя линия) в сравнении с истинной функцией дрейфа

Разность между $\mu_h(x)$ и $\mu(x)$ является наименьшей именно на интервале $x \in X = (X_{\min} - h,\ X_{\max} + h)$. Поэтому нельзя делать вывод о линейности или нелинейности непараметрической оценки дрейфа, основываясь на интервале большем, чем X.

На рисунке 4 изображена непараметрическая оценка (3) функции диффузии для $x \in [-0{,}5; 0{,}5]$ ($h = 0{,}0063$, ядро гауссовское). Видно, что $\sigma_h^2(x)$ на рассматриваемом промежутке также не является линейной и достигает своих предельных значений 0,000254 и 0,002437 (согласно (4) и (5)). Причем эти значения достигаются быстрее (то есть при x, которые находятся ближе к $X_{\min}$ и $X_{\max}$), чем в случае оценивания $\mu(x)$. На рисунке 5 представлен фрагмент рисунка 4 для $x \in [0; 0{,}2]$, а также истинная функция диффузии $\sigma^2(x)$ (прямая линия).

Очевидно, что разность между $\sigma_h^2(x)$ и $\sigma^2(x)$ является наименьшей на интервале X, вне которого оценка меняет свое поведение.

Таким образом, непараметрическую ядерную оценку функции регрессии независимо от ядра целесообразно рассматривать только на интервале $x \in (X_{\min} - h,\ X_{\max} + h)$. В противном случае можно получить оценку, которая не будет обладать свойствами, присущими оцениваемой функции.

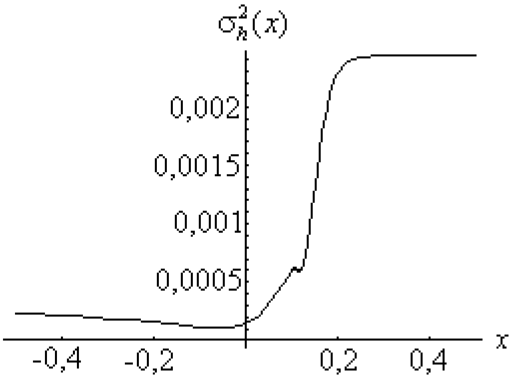

Рис.4. Непараметрическая оценка функции диффузии при использовании гауссовского ядра

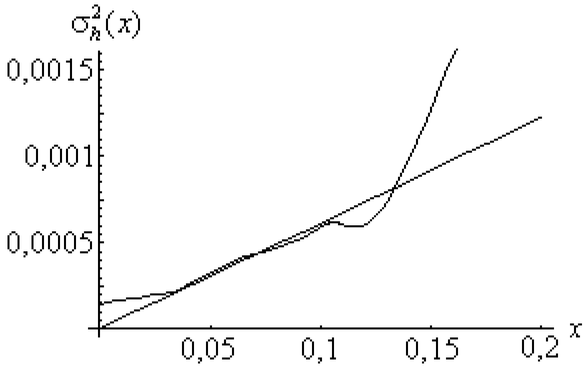

Рис.5. Фрагмент рисунка 3 в сравнении с истинной функцией диффузии (прямая линия)

ЛИТЕРАТУРА

1. Надарая, Э. А. Об оценке регрессии / Э. А. Надарая // Теория вероятностей и ее применение. 1964. Т. 9. № 1. С. 157–159.

2. Watson, G. S. Smooth regression analysis / G. S. Watson // Sankhya. Ser. A. 1964. V. 26. P. 359–372.

3. Stanton, R. A nonparametric model of term structure dynamics and the market price of interest rate risk / R. Stanton // J. of Finance. 1997. V. 52. № 5. P. 1973–2002.

4. Епанечников, В. А. Непараметрическая оценка многомерной плотности вероятности / В. А. Епанечников // Теория вероятностей и ее применение. 1969. Т. 14. № 1. С. 156–161.

5. Hardle, W. Applied nonparametric methods / W. Hardle, O. Linton // Handbook of Econometrics. Amsterdam, 1994. V. 4. P. 2295–2339.

6. Chapman, D. A. Is the Short Rate Drift Actually Nonlinear? / D. A. Chapman, N. D. Pearson // J. of Finance. 2000. V. 55. № 1. P. 355–388.