УДК 519.62

О ВЛИЯНИИ ЯДЕРНОЙ ФУНКЦИИ НА КАЧЕСТВО НЕПАРАМЕТРИЧЕСКОГО ОЦЕНИВАНИЯ ПЛОТНОСТИ ПО ВЫБОРКАМ ЗАВИСИМЫХ ОТСЧЕТОВ

Е.Г. Красногир

Белорусский государственный университет, Беларусь

E-mail krasnahir@bsu.by

Рассматривается влияние сходимости аппроксимации интегральной среднеквадратичной ошибки на непараметрическую оценку плотности для различных ядер в применении к одному из простейших стационарных диффузионных случайных процессов.

Ключевые слова: *Ядерное непараметрическое оценивание плотности, выбор параметра размытости, радиус сходимости степенного ряда, стационарный случайный процесс.*

В настоящее время существует большой класс задач, при решении которых нет возможности воспользоваться тем или иным известным семейством распределений. Поэтому все большее применение находят непараметрические модели, преимущество которых состоит в том, что они не требуют какой-либо априорной информации об оцениваемой функции. Непараметрическое оценивание успешно используется в таких областях, как распознавание образов, информационная проходка баз данных, машинное зрение, машинное обучение [1]. Однако отказ от априорной информации всегда приводит к ухудшению качества. Поэтому хотя непараметрическое ядерное оценивание сегодня является стандартным способом анализа различных данных, до сих пор не разрешен вопрос, как достичь качественного оценивания и какой из параметров взять в качестве оптимального.

Существует большое количество работ, где для построения непараметрических оценок используются выборочные данные, которые являются независимыми в совокупности (см. например [1], [2], [3], [4]). В данной статье мы рассмотрим аппроксимацию меры качества оценки (интегральной среднеквадратической ошибки оценивания MISE (Mean Integrated Squared Error)) степенным рядом в случае зависимых данных.

При использовании любой аппроксимации всегда возникает вопрос ее близости к аппроксимируемой функции. Так как мы будем иметь дело со степенным рядом, то естественно возникает вопрос его сходимости. Нашей задачей является нахождение радиуса сходимости для различных ядер и определение его влияния на оптимальный параметр размытости.

Пусть X является стационарным случайным процессом с неизвестной маргинальной плотностью $f(x)$ и пусть $X_1, \ldots, X_n$ – отсчеты этого процесса соответственно в моменты времени $t+\Delta, t+2\Delta, \ldots, t+(n-1)\Delta, t+n\Delta$ ($\Delta>0$). В качестве оценки функции $f(x)$ рассмотрим статистику [2]

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

где h – параметр размытости, который должен удовлетворять условиям состоятельности [2]:

$$h \to 0,\ nh \to \infty \text{ при } n \to \infty, \tag{2}$$

а $K(x)$ – ядерная функция, имеющая свойства симметричной плотности распределения вероятностей [3], то есть

$$\int_{-\infty}^{+\infty} K(x)dx = 1,\ K(x) \ge 0,\ K(x) = K(-x). \tag{3}$$

Функцию $f_h(x)$, определяемую выражением (1), будем называть непараметрической оценкой маргинальной плотности $f(x)$. Так как (1) состоит из двух компонентов, то качество оценивания определяется выбором ядра $K(x)$ и параметра размытости h. Для нахождения h необходимо выбрать некий критерий или расстояние между истинной плотностью и её непараметрической оценкой, а затем минимизировать данное расстояние по h. Одним из наиболее простых критериев является интегральная среднеквадратичная ошибка оценивания $MISE(h)$ [2]:

$$MISE(h) \equiv \int_{-\infty}^{+\infty} E[(f_h(x) - f(x))^2]dx. \tag{4}$$

Так как найти оптимальный параметр размытости в явной форме из выражения (4) не удается, его определяют приближенно с определенной точностью. Поскольку по условию состоятельности (2) при больших объемах выборки параметр h должен быть достаточно малым, обычно представляют (4) в виде степенного ряда Тейлора по параметру h:

$$MISE(h) = \sum_{k=0}^{\infty} a_k h^k \tag{5}$$

и ограничиваются несколькими первыми членами разложения. Тогда задача поиска оптимального параметра размытости принимает вид

$$\sum_{k=1}^{m} a_k h^k \to \min_h, \tag{6}$$

где m – некоторое (обычно небольшое) натуральное число.

Однако при этом возникает вопрос сходимости полученного степенного ряда (5). Если оптимальный параметр, полученный при минимизации функции (4), находится вне интервала сходимости ряда, то при любом m решение задачи (6) не будет одновременно оптимальным параметром для (4).

Обозначим совместную плотность отсчётов X_k и X_{k+j}, $k = \overline{1, n-j}$, $j = \overline{1, n-k}$, стационарного случайного процесса X через $f_j(t, s)$. Тогда из (4) можно получить

$$MISE(h) = \int_{-\infty}^{+\infty} f^2(x)dx - 2\sum_{m=0}^{\infty}\left(\frac{(-h)^m \mu_m(K)}{m!}\int_{-\infty}^{\infty} f(x) f^{(m)}(x)dx\right) +$$

$$+\frac{\nu_2(K)}{hn} + \frac{2}{n^2}\sum_{m=0}^{\infty}\frac{(-h)^m}{m!}\left(\sum_{k=0}^{m} C_m^k \mu_k(K)\mu_{m-k}(K)\left[\sum_{j=1}^{n-1}(n-j)\int_{-\infty}^{\infty} {}_j f_{k,m-k}^{(m)}(x,x)dx\right]\right),$$

где

$$\mu_r(K) = \int_{-\infty}^{\infty} u^r K(u)du,\ \nu_2(K) = \int_{-\infty}^{\infty} K^2(u)du,\ {}_j f_{k,m-k}^{(m)}(x,x) = \left.\frac{\partial^m f_j}{\partial t^k \partial s^{m-k}}\right|_{t=x, s=x}.$$

Так как ядро является симметричной функцией, то коэффициенты $\mu_r(K) = 0$ для нечётных значений r. Используя это свойство, получаем следующее утверждение:

Утверждение 1. *Пусть соответственно маргинальная и совместная функция, $f(x)$ и $f_j(t, s)$, наблюдений процесса X необходимое число раз дифференцируемы в окрестностях точек x и (x, x), причём $f^{(k)}(\pm\infty) = 0$ для всех k. Тогда для h, принадлежащих интервалу сходимости ряда Тейлора и ядер $K(x)$, удовлетворяющих условию (3), функция $MISE(h)$ преобразуется к виду*

$$MISE(h) = \int_{-\infty}^{+\infty} f^2(x)dx - 2\sum_{m=0}^{\infty}\left(\frac{h^{2m}\mu_{2m}(K)}{(2m)!}\int_{-\infty}^{\infty} f(x) f^{(2m)}(x)dx\right) + \frac{\nu_2(K)}{hn} +$$

$$+\frac{2}{n^2}\sum_{m=0}^{\infty}\frac{h^{2m}}{(2m)!}\left(\sum_{k=0}^{m} C_{2m}^{2k}\mu_{2k}(K)\mu_{2m-2k}(K)\left[\sum_{j=1}^{n-1}(n-j)\int_{-\infty}^{\infty} {}_j f_{2k,2m-2k}^{(2m)}(x,x)dx\right]\right). \tag{7}$$

Опустив не зависящие от h слагаемые, получим функцию, которую нужно минимизировать для поиска оптимального параметра. Обозначим её $Q(h)$.

Как мы видим, в (7) присутствует два степенных ряда. Для проверки их сходимости можно воспользоваться утверждением 2 [5]:

Утверждение 2. *Если существует* $\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| =:: l$, *то радиус сходимости R степенного ряда* $\sum_{k=0}^{\infty} a_k x^k$ *определяется как* $R = \frac{1}{l}$.

В связи с тем, что члены рядов в (7) являются сложными функциональными выражениями, найти радиусы сходимости в явном виде не представляется возможным. Поэтому мы исследуем вопрос сходимости для некоторых ядер в предположении, что процесс X является одним из

простейших диффузионных случайных процессов – процессом, который описывается следующим стохастическим дифференциальным уравнением [6]:

$$dX = k(\theta - X)dt + \sigma\, dW(t), \tag{8}$$

где $W(t)$ – стандартный винеровский процесс, а k, θ, σ – постоянные параметры. Иногда говорят, что (8) задает модель Васичека. Ввиду линейности уравнения (8) относительно X, случайный процесс, порождаемый им, является нормальным марковским с маргинальной плотностью вероятностей

$$f(x) = \frac{1}{\sqrt{2\pi D}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{\sqrt{D}} \right)^2 \right],$$

где $D = \dfrac{\sigma^2}{2k}$. Совместная плотность

$$f_j(t,x) = \frac{1}{2\pi D\sqrt{1-\rho^{2j}}} \exp\left[-\frac{1}{2}\left(\frac{t-\theta-(x-\theta)\rho^j}{\sqrt{D(1-\rho^{2j})}} \right)^2 - \frac{1}{2}\left(\frac{x-\theta}{\sqrt{D}} \right)^2 \right],$$

где $\rho = e^{-k\Delta}$ – коэффициент корреляции между соседними отсчетами.

Итак, для рассматриваемой модели случайного процесса из (7) мы получим

$$Q(h) = -2\sum_{m=1}^{\infty}\left(\frac{(-1)^m h^{2m} \mu_{2m}(K)(2m-1)!!}{2^{m+1}\sqrt{D^{2m+1}}(2m)!\sqrt{\pi}} \right) + \frac{\nu_2(K)}{hn} +$$

$$+ \frac{2}{n^2}\sum_{m=1}^{\infty}\left(\frac{(-1)^m h^{2m}(2m-1)!!}{2^{m+1}\sqrt{D^{2m+1}}\sqrt{\pi}(2m)!} \left[\sum_{j=1}^{n-1} \frac{(n-j)}{\sqrt{(1-\rho^j)^{2m+1}}} \right] \left(\sum_{k=0}^{m} C_{2m}^{2k} \mu_{2k}(K)\mu_{2m-2k}(K) \right) \right). \tag{9}$$

Исследуем зависимость радиуса сходимости данного ряда от выбора ядерной функции. Найдем радиусы сходимости первого и второго ряда в (9). Для первого ряда

$$\left| \frac{a_{m+1}}{a_m} \right|_1 = \frac{\mu_{2m+2}(K)(2m+1)!!\, 2^{m+1}\sqrt{D^{2m+1}}(2m)!\sqrt{\pi}}{2^{m+2}\sqrt{D^{2m+3}}(2m+2)!\sqrt{\pi}\mu_{2m}(K)(2m-1)!!} = \frac{\mu_{2m+2}(K)}{2D(2m+2)\mu_{2m}(K)}, \tag{10}$$

для второго

$$\left| \frac{a_{m+1}}{a_m} \right|_2 = \frac{(2m+1)!!\, 2^{m+1}\sqrt{D^{2m+1}}(2m)!\sqrt{\pi}}{2^{m+2}\sqrt{D^{2m+3}}(2m+2)!\sqrt{\pi}(2m-1)!!} \times$$

$$\times \left[\sum_{j=1}^{n-1} \frac{(n-j)}{\sqrt{(1-\rho^j)^{2m+3}}} \right] \left[\sum_{j=1}^{n-1} \frac{(n-j)}{\sqrt{(1-\rho^j)^{2m+1}}} \right]^{-1} \frac{S_{2m+2}}{S_{2m}} =$$

$$= \frac{1}{4D(m+1)(1-\rho)} \left[\sum_{j=1}^{n-1} \frac{n-j}{\sqrt{\left(\sum_{i=0}^{j-1} \rho^i \right)^{2m+3}}} \right] \left[\sum_{j=1}^{n-1} \frac{n-j}{\sqrt{\left(\sum_{i=0}^{j-1} \rho^i \right)^{2m+1}}} \right]^{-1} \frac{S_{2m+2}}{S_{2m}} \sim$$

$$\sim \frac{1}{4D(m+1)(1-\rho)} \frac{S_{2m+2}}{S_{2m}}, \tag{11}$$

где

$$S_{2m} = \sum_{k=0}^{m} C_{2m}^{2k} \mu_{2k}(K)\mu_{2m-2k}(K). \tag{12}$$

В таблице 1 представлены компоненты формулы (10), необходимые для вычисления величины $\left| \dfrac{a_{m+1}}{a_m} \right|_1$ и ее предела при $m \to \infty$ для различных ядер.

Таблица 1

Ядро [2]	$K(u)$	$\mu_{2m}(K)$	$\left\|\frac{a_{m+1}}{a_m}\right\|_1$	$\lim\limits_{m\to\infty}\left\|\frac{a_{m+1}}{a_m}\right\|_1$
Епанечникова	$3(1-u^2)I(\|u\|\le 1)/4$	$\frac{3}{(2m+3)(2m+1)}$	$\frac{2m+1}{4D(m+1)(2m+5)}$	0
Квадратичное	$15(1-u^2)^2I(\|u\|\le 1)/16$	$\frac{15(2m-1)!!}{(2m+5)!!}$	$\frac{2m+1}{4D(m+1)(2m+7)}$	0
Треугольное	$(1-\|u\|)I(\|u\|\le 1)$	$\frac{1}{(m+1)(2m+1)}$	$\frac{2m+1}{4D(m+2)(2m+3)}$	0
Гауссовское	$(2\pi)^{-1/2}\exp(-u^2/2)$	$(2m-1)!!$	$\frac{2m+1}{4D(m+1)}$	$\frac{1}{2D}$
Равномерное	$I(\|u\|\le 1)/2$	$\frac{1}{2m+1}$	$\frac{2m+1}{4D(m+1)(2m+3)}$	0
Кубическое	$35(1-u^2)^3I(\|u\|\le 1)/32$	$\frac{105(2m-1)!!}{(2m+7)!!}$	$\frac{2m+1}{4D(m+1)(2m+9)}$	0

В таблице значение $I(|u|\le 1)$ равно либо 1 при выполнении условия в скобках либо 0 в противном случае.

Далее, в таблице 2 представлена величина S_{2m}, вычисленная по формуле (12), а также отношение $\frac{S_{2m+2}(K)}{S_{2m}(K)}$ для различных ядер.

Таблица 2

Ядро	$S_{2m}(K)$	$\frac{S_{2m+2}(K)}{S_{2m}(K)}$
Епанечникова	$\frac{9(m+1)(2m+5)2^{2m+4}(2m)!}{\Gamma(2m+7)}$	$\frac{4(m+2)(2m+1)}{(m+4)(2m+5)}$
Квадратичное	$\frac{225(m+1)(m+2)2^{2m+5}(2m)!}{(m+4)(m+5)\Gamma(2m+7)}$	$\frac{4(2m+1)(m+3)}{(2m+7)(m+6)}$
Треугольное	$\frac{8(2^{2m+2}-1)(2m)!}{\Gamma(2m+5)}$	$\frac{(2^{2m+4}-1)(2m+1)(m+1)}{(2^{2m+2}-1)(2m+5)(m+3)}$
Гауссовское	$\frac{(2m)!}{m!}$	$4m+2$
Равномерное	$\frac{2^{2m+1}(2m)!}{\Gamma(2m+3)}$	$\frac{4(2m+1)(m+1)}{(2m+3)(m+2)}$
Кубическое	$\frac{11025(m+1)(m+2)(m+3)2^{2m+7}(2m)!}{(m+5)(m+6)(m+7)\Gamma(2m+9)}$	$\frac{4(2m+1)(m+4)}{(2m+9)(m+8)}$

Из таблицы 2 и (11) следует, что $\lim\limits_{m\to\infty}\left|\frac{a_{m+1}}{a_m}\right|_2$ равен $\frac{1}{D(1-\rho)}$ для гауссовского ядра и нулю для оставшихся ядер.

Таким образом, из таблицы 1 и 2 и утверждения 2 получаем, что ряд (9) в случае гауссовского ядра сходится только для $0 \le h < \sqrt{D(1-\rho)} \equiv R$.

Для других вышерассмотренных ядер мы имеем сходимость для любых h.

При использовании гауссовского ядра параметр h – решение (6) – будет близок к $h_{MISE} = \arg\min MISE(h)$ только если

$$h_{MISE} \le \sqrt{D(1-\rho)}\ . \qquad (13)$$

В противном случае использование многочлена Тейлора любой степени будет приводить к значению $h < h_{MISE}$.

Все остальные ядра могут использоваться без всяких ограничений.

Рассмотрим пример. Пусть $k = 0{,}85837,\ \theta = 0{,}089102,\ \sigma^2 = 0{,}0021854$. Данные значения приводятся в [7] как оценки параметров модели Васичека (8), построенной на основе анализа реальных данных. Пусть $n = 1438$. Мы будем рассматривать гауссовское и равномерное ядро. Легко показать, что для модели Васичека выражение (4) равно

$$MISE(h) = \frac{1}{2\sqrt{\pi D}} + \frac{1}{2hn\sqrt{\pi}} - \frac{1}{\sqrt{\pi}\sqrt{D + 0{,}5h^2}} + \frac{1}{n^2\sqrt{\pi}} \sum_{j=1}^{n-1} \frac{n-j}{\sqrt{h^2 + D(1-\rho^j)}} \qquad (14)$$

для гауссовского ядра и

$$MISE(h) = \frac{1}{2\sqrt{\pi D}} + \frac{1}{2hn} - \frac{1}{h} Erf\left(\frac{h}{2\sqrt{D}}\right) + \frac{1}{n^2 h} \sum_{j=1}^{n-1} (n-j) \left[\frac{\exp\{-y^2(h)\} - 1}{y(h)\sqrt{\pi}} + \frac{Erf(y(h))}{h} \right], \qquad (15)$$

где $Erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$, $y(h) = \frac{h}{\sqrt{D(1-\rho^j)}}$, для равномерного.

Нашей целью является сравнение h_{MISE}, найденного минимизацией (14), и параметра $h_{(6)}$, полученного при решении задачи (6).

Возьмем гауссовское ядро и найдем h_{MISE} для (a) $\Delta = 1/6$ и (b) $\Delta = 1/12$ (см. таблицу 3). Из таблицы 3 следует, что условие (13) выполняется только в случае (a). Также для сравнения вычислим h_{MISE} для равномерного ядра с использованием (15) ($\Delta = 1/6$, $\Delta = 1/12$ и $\Delta = 1/24$).

Далее, найдем $h_{(6)}$ при $m = 2$ для вышерассмотренных ядер и Δ. Как видно из таблицы 3, все $h_{(6)}$ меньше h_{MISE}. Начнем увеличивать m, чтобы получить равенство этих параметров с точностью до седьмого знака после запятой. Через m^* обозначим такое значение m, при котором достигается вышеуказанная точность. Результаты представлены в таблице 3.

Таблица 3

Ядро	Гауссовское		Равномерное		
Δ	1/6	1/12	1/6	1/12	1/24
h_{MISE}	0,0107385	0,0128802	0,0185073	0,0220581	0,0269194
Радиус сходимости	0,0130265	0,00937432	∞	∞	∞
$h_{(6)}$ для m=2	0,0097677	0,00930027	0,017404	0,0172272	0,0141467
m^*	30	-	11	21	50

Мы видим, что для рассмотренных ядер и Δ (за исключением случая b для гауссовского ядра) значения $h_{(6)}$ и h_{MISE} становятся близкими для относительно малых m^*. Заметим также, что с уменьшением Δ величина m^* возрастает.

Рассмотрим теперь более подробно случай $\Delta = 1/12$ для гауссовского ядра. Как уже было сказано выше, в данной ситуации не выполняется условие (13). На рисунке 1 показана зависимость $h_{(6)}$ от m. Как мы видим, при увеличении m величина $h_{(6)}$ возрастает гораздо медленнее, чем в остальных рассмотренных в таблице 3 случаях. При этом оптимальный параметр не превосходит значения радиуса сходимости даже для больших m (при $m = 5000$ $h_{(6)} = 0{,}009371$, при $m = 10000$ $h_{(6)} = 0{,}009372$).

Таким образом, мы убедились, что радиус сходимости оказывает серьезное влияние на оптимальный параметр размытости, полученный минимизацией выражения (6). Рассмотрим теперь случай $\rho = 0$, что соответствует независимым выборочным значениям, и исследуем возможность применения функции $Q(h)$ (9) и выражения (6) для нахождения оптимального значения h для гауссовского ядра. Радиус сходимости в данной ситуации равен $\sqrt{D}$. Если оптимальное значение h будет находиться внутри интервала сходимости ряда (9), то в точке $h = \sqrt{D}$ функция $MISE(h)$ будет возрастать, соответственно ее производная в данной точке будет положительной. Найдем знак производной в точке $h = \sqrt{D}$:

$$MISE'(\sqrt{D}) = -\frac{1}{2Dn\sqrt{\pi}} + \frac{1}{2D\sqrt{\pi}\sqrt{1{,}5^3}} - \frac{(n-1)}{2Dn\sqrt{\pi}\sqrt{8}} = \frac{1}{2D\sqrt{\pi}}\left(-\frac{1}{n} + \frac{1}{\sqrt{1{,}5^3}} - \frac{(n-1)}{n\sqrt{8}}\right).$$

Отсюда следует, что при $n > \frac{1{,}5^{1{,}5}(2\sqrt{2}-1)}{2\sqrt{2}-1{,}5^{1{,}5}} \approx 3{,}4$ производная больше нуля, и минимальное значение параметра размытости находится в пределах интервала сходимости. Поэтому для $\rho = 0$ оптимальные значения h для функций $Q(h)$ и $MISE(h)$ будут совпадать.

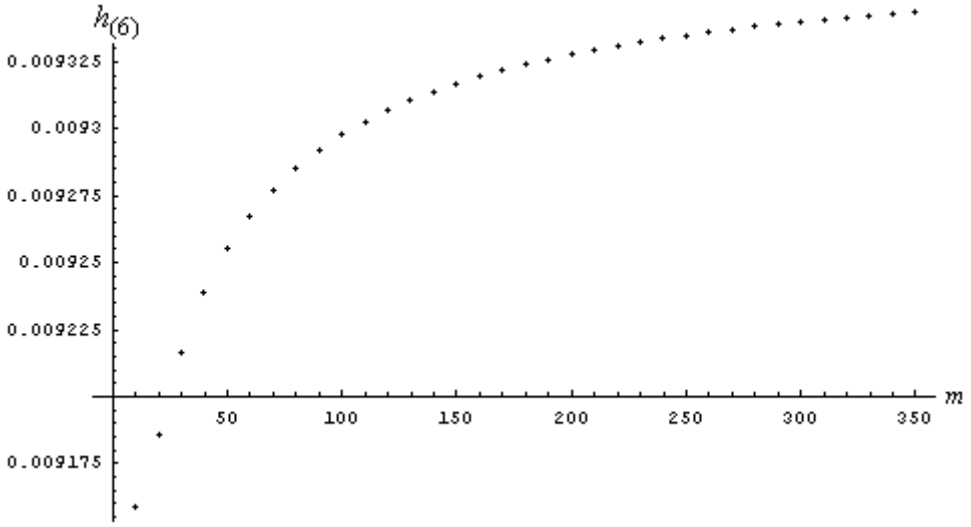

Рис. 1.

Изложенные выше результаты показывают, как влияет величина радиуса сходимости на параметр размытости при применении аппроксимации функции *MISE*(*h*) рядом Тейлора в случае зависимых наблюдений. Мы установили, что для модели Васичека наибольшие трудности возникают при использовании гауссовского ядра, так как если оптимальный параметр h_{MISE} функции *MISE*(*h*) находится вне интервала сходимости, то даже в случае большого количества членов аппроксимации не удается получить параметр размытости, близкий к h_{MISE}.

Также мы увидели, что с уменьшением величины Δ увеличивается количество членов аппроксимации, необходимых для получения *h*, близких к h_{MISE}. Для гауссовского ядра уменьшение Δ приводит к уменьшению радиуса сходимости. В случае независимых наблюдений ($\Delta = \infty$) радиус сходимости не будет влиять на оптимальный параметр. Таким образом, влияние зависимости может существенно осложнять поиск оптимального параметра размытости.

Литература:

1. *Raykar V.C., Duraiswami R.* Very fast optimal bandwidth selection for univariate kernel density estimation. Technical Report. – Univ. of Maryland, 2006.
2. *Turlach B.A*. Bandwidth selection in kernel density estimation: a review. Technical report. – Univ. Catholique de Louvain, 1993.
3. *Katkovnik V., Shmulevich I.* Kernel density estimation with adaptive varying window size // Pattern Recognition Letters. – 2002. – Vol. 23. – P. 1641–1648.
4. *Comaniciu D., Ramesh V., Meer P.* The variable bandwidth mean shift and data-driven scale selection // Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). – 2001. – P. 438-445.
5. *Воднев В.Т., Наумович А.Ф., Наумович Н.Ф.* Основные математические формулы. – Мн.: Вышэйшая школа, 1980. – 354 с.
6. *Vasicek O.* An Equilibrium Characterization of the Term Structure // Journal of Financial Economics. – 1977. – Vol. 5. – P. 177-188.
7. *Pritsker M.* Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models // Working paper. – Board of Governors of the USA Federal Reserve System, 1997.