УДК 519.7

© 2011 г. **А.В. Лапко**, д-р техн. наук,
В.А. Лапко, д-р техн. наук
(Институт вычислительного моделирования СО РАН, Красноярск)
(Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева, Красноярск)

ИССЛЕДОВАНИЕ НЕПАРАМЕТРИЧЕСКОЙ РЕГРЕССИИ В УСЛОВИЯХ ЧАСТИЧНЫХ СВЕДЕНИЙ О ВИДЕ ВОССТАНАВЛИВАЕМОЙ ЗАВИСИМОСТИ

Исследуется модифицированная непараметрическая регрессия, которая обеспечивает учет априорных сведений о виде восстанавливаемых зависимостей. Устанавливаются свойства ее асимптотической несмещенности и осуществляется анализ результатов вычислительных экспериментов.
Ключевые слова: непараметрическая регрессия, стохастические зависимости, априорные сведения, асимптотические свойства.

Введение

Для наиболее полного учета априорной информации о виде восстанавливаемых зависимостей и экспериментальных данных о ее локальном поведении широко используются гибридные модели [1]. Традиционные гибридные модели сочетают в одном решающем правиле преимущество параметрических и непараметрических аппроксимаций. При этом единое решающее правило образуют параметрическая модель восстанавливаемой зависимости и корректирующая ее функция непараметрического типа, которые строятся в одном и том же пространстве переменных. Полученные результаты были развиты на условия наличия частичных априорных сведений о виде восстанавливаемых зависимостей в ограниченном пространстве признаков [2]. Основная проблема применения гибридных моделей состоит в выборе вида корректирующей функции, которая является трудно формализуемой. Для ее обхода предлагается использовать непараметрическую регрессию, синтез которой основан на обобщении априорной информации о виде восстанавливаемых зависимостей и экспериментальных данных об их локальном поведении.

Цель работы состоит в обосновании возможности учета априорных сведений о виде восстанавливаемых закономерностей при синтезе непараметрических моделей стохастических зависимостей, основанных на оценках плотности вероятности типа Розенблатта-Парзена [3].

Синтез модифицированной непараметрической регрессии

Пусть об однозначной зависимости

$$y = y\ (x), \quad \forall x \in R^k \qquad (1)$$

известно ее частичное описание

$$\bar{y}_1 = F\left(\bar{x}_1, a\right), \quad \forall \bar{x}_1 \in R^{k2}, \quad k2 < k$$

из класса линейных полиномов относительно некоторого ограниченного набора признаков $\bar{x}_1$ из $x = (x_1, \bar{x}_1)$. Здесь $x_1 = \left(x_{1v}, v = \overline{1, k1}\right)$ при $k = k1 + k2$. Имеется выборка $V = \left(x_1^i, \bar{x}_1^i, y^i, i = \overline{1,n}\right)$ экспериментальных данных, составленная из статистически независимых значений переменных x, y исследуемой зависимости (1). Параметры a полинома $F(\bar{x}_1, a)$ будем считать заданными.

Необходимо осуществить синтез модифицированной непараметрической модели $\bar{\bar{y}}(x)$ зависимости (1), совмещающей в одном решающем правиле всю имеющуюся априорную информацию.

На основании априорных сведений, организуя вычислительный эксперимент, сформируем промежуточную обучающую выборку

$$V^1 = \left(x_1^i, \bar{y}_1^i = F\left(\bar{x}_1^i, a\right), y^i, i = \overline{1,n}\right).$$

Известно, что оптимальным решающим правилом в смысле минимума среднеквадратического отклонения является условное математическое ожидание [4]

$$\bar{y} = j_1(x_1, \bar{y}_1) = \int_{-\infty}^{+\infty} y\ p\left(\frac{y}{x_1, \bar{y}_1}\right) dy.$$

В качестве приближения по эмпирическим данным V^1 кривой регрессии $\bar{y} = j_1(x_1, \bar{y}_1) = j(x)$ примем статистику

$$\bar{\bar{y}}(x) = \sum_{i=1}^{n} y^i b_i(x), \qquad (2)$$

где $b_i(x) = \dfrac{\prod\limits_{n=1}^{k1} \Phi\left(\dfrac{x_{1v} - x_{1v}^i}{c_v}\right) \Phi\left(\dfrac{\bar{y}_1 - \bar{y}_1^i}{c}\right)}{\sum\limits_{i=1}^{n} \prod\limits_{n=1}^{k1} \Phi\left(\dfrac{x_{1v} - x_{1v}^i}{c_v}\right) \Phi\left(\dfrac{\bar{y}_1 - \bar{y}_1^i}{c}\right)}$.

В статистике (2) ядерные функции $\Phi(u)$ удовлетворяют условиям H:

$\Phi(u) = \Phi(-u), \ 0 \le \Phi(u) < \infty$;

$\int \Phi(u) du = 1, \ \int u^2 \Phi(u) du = 1; \ \int u^m \Phi(u) du < \infty, \ 0 \le m < \infty$;

$c = c(n)$; $c_v = c_v(n)$; $v = \overline{1, k1}$ – коэффициенты размытости ядерных функций, значения которых убывают с ростом объема n обучающей выборки. Здесь и далее бесконечные пределы интегрирования опускаются.

Оптимизация модифицированной непараметрической регрессии (2) по коэффициентам размытости ядерных функций c, c_v, $v = \overline{1, k1}$ осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки

среднеквадратической ошибки аппроксимации искомой зависимости.

При оценивании зависимости в ситуации $x=(x_1,\bar{x}_1)$ сначала вычисляется $\bar{y}_1=F(\bar{x}_1,a)$, а затем по данным $(x_1,\bar{y}_1)$ в соответствии со статистикой (2) определяется значение $\bar{\bar{y}}(x)$.

Асимптотические свойства модифицированной непараметрической регрессии

Для упрощения изложения результатов аналитических исследований будем считать, что в частичном наборе признаков $x_{1\nu}$, $\nu=\overline{1,k1}$ их количество $k1 = 1$. В качестве ядерной функции примем функцию вида

$$\Phi(u)=\begin{cases}\frac{1}{2} & \forall \quad |u|<1,\\ 0 & \forall \quad |u|\geq 1.\end{cases}$$

В этом случае непараметрическая регрессия (2) запишется:

$$\bar{\bar{y}}(x)=\frac{\sum_{i=1}^{n} y^i\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\bar{y}_1-\bar{y}_1^i}{c}\right)}{\sum_{i=1}^{n}\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\bar{y}_1-\bar{y}_1^i}{c}\right)}. \qquad (3)$$

Тогда справедливо следующее утверждение.

Теорема. Пусть 1) частичные сведения $\bar{y}_1=F(\bar{x}_1,a)$ о виде восстанавливаемой зависимости (1) принадлежат к классу линейных полиномов; 2) функция $j(x)$ и плотность вероятности $p(x)$ ограничены вместе со своими производными до второго порядка включительно; 3) ядерные функции $\Phi(u)$ являются положительными, симметричными и нормированными; 4) последовательности коэффициентов размытости $c_1(n)$, $c(n)$ ядерных функций таковы, что при $n\to\infty$ их значения стремятся к нулю. Тогда непараметрическая регрессия (3) обладает свойством асимптотической несмещенности относительно оптимального решающего правила $j(x)$.

Доказательство. Представим модель (3) в виде

$$\bar{\bar{y}}(x)=\frac{(nc_1c)^{-1}\sum_{i=1}^{n} y^i\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\bar{y}_1-\bar{y}_1^i}{c}\right)}{(nc_1c)^{-1}\sum_{i=1}^{n}\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\frac{\bar{y}_1-\bar{y}_1^i}{c}\right)}=\frac{\bar{z}_1(x)}{\bar{z}_2(x)}. \qquad (4)$$

Проведем преобразования

$$M\frac{\bar{z}_1(x)}{\bar{z}_2(x)}=M\left[\frac{\bar{z}_1(x)}{M\,\bar{z}_2(x)}+\frac{\bar{z}_1(x)}{\bar{z}_2(x)\,M\,\bar{z}_2(x)}\left(M\,\bar{z}_2(x)-\bar{z}_2(x)\right)\right], \qquad (5)$$

где M – знак математического ожидания.

Ввиду ограниченности значений $\overline{=}_{y(x)=\frac{\bar{z}_1(x)}{\bar{z}_2(x)}}$ свойства статистики (3) зависят от асимптотического поведения $M(\bar{z}_1(x))$, $M(\bar{z}_2(x))$. Вычислим

$$M(\bar{z}_2(x))=(nc_1c)^{-1}\sum_{i=1}^{n}\int\ldots\int\Phi\left(\frac{x_1-x_1^i}{c_1}\right)\Phi\left(\sum_{n=1}^{k2}a_\nu\frac{\bar{x}_{1\nu}-\bar{x}_{1\nu}^i}{c}\right)\times$$

$$\times\, p\left(x_1^i,\bar{x}_{1\nu}^i,\nu=\overline{1,k2}\right)dx_1^i d\bar{x}_{1n}^{\mathrm{i}}\ldots d\bar{x}_{1k2}^i.$$

Так как $x_1^i,\bar{x}_{1\nu}^i,\nu=\overline{1,k2}$ являются значениями одних и тех же случайных величин $t,t_\nu,\nu=\overline{1,k2}$ с плотностью вероятности $p\left(t,t_\nu,\nu=\overline{1,k2}\right)$, то

$$M(\bar{z}_2(x))=(c_1c)^{-1}\int\ldots\int\Phi\left(\frac{x_1-t}{c_1}\right)\Phi\left(\sum_{\nu=1}^{k2}\frac{a_\nu}{c}(\bar{x}_{1\nu}-t_\nu)\right)\times$$

$$\times\, p\left(t,t_\nu,\nu=\overline{1,k2}\right)dtdt_1\ldots dt_{k2}.$$

Проведем замену переменных $u=\frac{(x_1-t)}{c_1}$, $u_\nu=\frac{a_\nu(\bar{x}_{1\nu}-t_\nu)}{c}$. После несложных преобразований получим

$$M(\bar{z}_2(x))=\frac{c^{k2-1}}{\prod_{\nu=1}^{k2}a_\nu}\int\ldots\int\Phi(u)\,\Phi\left(\sum_{\nu=1}^{k2}u_\nu\right)\times$$

$$\times\, p\left(x_1-c_1u,\bar{x}_{1\nu}-\frac{c}{a_\nu}u_\nu,\nu=\overline{1,k2}\right)du\,du_1\ldots\,du_{k2}. \qquad (6)$$

Разложим функцию

$$p\left(x_1-c_1u,\bar{x}_{1\nu}-\frac{c}{a_\nu}u_\nu,\nu=\overline{1,k2}\right)$$

в ряд Тейлора в точке $x=\left(x_1,x_{1\nu},\nu=\overline{1,k2}\right)$ и преобразуем (6) с учетом свойств:

$$\frac{1}{2^{k2-1}}\int_{-1}^{1}\ldots\int_{-1}^{1}\Phi\left(\sum_{\nu=1}^{k2}u_\nu\right)du_1\ldots du_{k2}=1,$$

$$\frac{1}{2^{k2-1}}\int_{-1}^{1}\ldots\int_{-1}^{1}u_t\Phi\left(\sum_{\nu=1}^{k2}u_\nu\right)du_1\ldots du_{k2}=0,\ t=\overline{1,k2},$$

$$\frac{1}{2^{k2-1}}\int_{-1}^{1}\ldots\int_{-1}^{1}u_t^2\ \Phi\left(\sum_{\nu=1}^{k2}u_\nu\right)du_1\ldots du_{k2}=b^{\,2}.$$

В результате при $n\to\infty$ имеем:

$$M(\bar{z}_2(x))\sim\frac{2^{k2-1}c^{k2-1}}{\prod_{\nu=1}^{k2}a_\nu}\times$$

$$\times\left[p(x)+\frac{c_1^2}{2}p_{x_1}^{(2)}(x)+\frac{c^2b^{\,2}}{2^{k2}}\sum_{\nu=1}^{k2}\frac{1}{a_\nu^{\,2}}p_\nu^{(2)}(x)+0\left(c^4\right)\right], \qquad (7)$$

где $p_{x_1}^{(2)}(x)$, $p_v^{(2)}(x)$ – вторые производные плотности вероятности $p\left(x_1, \bar{x}_{1v}, n = \overline{1,k2}\right)$ по переменным x_1, $\bar{x}_{1v}, v = \overline{1,k2}$ соответственно.

Следуя приведенной технологии вычислений, найдем асимптотическое выражение для

$$M\left(\bar{\bar{z}}_1(x)\right) = (nc_1c)^{-1}\sum_{i=1}^{n}\int\ldots\int y^i \Phi\left(\frac{x_1 - x_1^i}{c_1}\right)\Phi\left(\sum_{v=1}^{k2} a_v \frac{\bar{x}_{1v} - \bar{x}_{1v}^i}{c}\right)\times$$

$$\times p\left(y^i, x_1^i, \bar{\bar{x}}_{1v}^i, v = \overline{1,k2}\right) dy^i dx_1^i \, d\bar{\bar{x}}_{11}^i \ldots d\bar{\bar{x}}_{1k2} =$$

$$= \frac{c^{k2-1}}{\prod_{v=1}^{k2} a_v}\int\ldots\int j\left(x_1 - c_1 u, \bar{\bar{x}}_{1v} - \frac{c}{a_v}u_v, v = \overline{1,k2}\right)\times$$

$$\times \Phi(u)\Phi\left(\sum_{v=1}^{k2} u_v\right) p\left(x_1 - c_1 u, \bar{\bar{x}}_{1v} - \frac{c}{a_v}u_v, v = \overline{1,k2}\right) du\, du_1 \ldots du_{k2} \sim$$

$$\sim \frac{2^{k2-1} c^{k2-1}}{\prod_{v=1}^{k2} a_v}\left[j(x)p(x) + \frac{c_1^2}{2}\left(j(x)p(x)\right)_{x_1}^{(2)} + \right.$$

$$\left. + \frac{c^2 b^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{a_v^2}\left(j(x)p(x)\right)_{x_{1v}}^{(2)} + 0\left(c_1^2 c^2, c_1^4 c_v^4, v = \overline{1,k2}\right)\right]. \qquad (8)$$

В выражении (8) $\left(j(x)p(x)\right)_{x_1}^{(2)}$, $\left(j(x)p(x)\right)_{x_{1v}}^{(2)}$ – вторые производные произведения двух функций по переменным x_1, $\bar{x}_{1v}, v = \overline{1,k2}$ соответственно.

Подставим выражения (7) и (8) в (5), получим:

$$M\left(\bar{\bar{y}}(x)\right) \sim M\frac{\bar{\bar{z}}_1(x)}{\bar{\bar{z}}_2(x)} \sim$$

$$\sim \frac{j(x)p(x) + \frac{c_1^2}{2}\left(j(x)p(x)\right)_{x_1}^{(2)} + \frac{c^2 b^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{a_v^2}\left(j(x)p(x)\right)_{x_{1v}}^{(2)}}{p(x) + \frac{c_1^2}{2}p_{x_1}^{(2)}(x) + \frac{c^2 b^2}{2^{k2}}\sum_{v=1}^{k2}\frac{1}{a_v^2}p_v^{(2)}(x)}. \qquad (9)$$

Из анализа выражения (9) следует, что при $c_1 = c_1(n) \to 0$, $c = c(n) \to 0$ с ростом $n \to \infty$ изучаемая статистика (3) обладает свойством асимптотической несмещенности.

Замечание. При $k2 = 2$ полученные результаты могут быть использованы при исследовании свойств статистических моделей, основанных на методе группового учета аргументов [5]. Идея метода заключается в построении последовательности моделей

$$\bar{y}_j = \vec{j}_j(x_j, \bar{y}_{j-1}), i = \overline{1,m}. \qquad (10)$$

Ранее не используемая в моделях $\bar{y}_t$, $t=\overline{1,j-1}$ компонента x_j вектора аргументов x обеспечивает в наборе с $\bar{y}_{j-1}$ минимальное расхождение значений $\bar{y}_j$ с экспериментальными данными. На каждом этапе процедуры (10) искомая зависимость оценивается в пространстве двух переменных $(x_j, \bar{y}_{j-1})$.

Анализ результатов вычислительных экспериментов

На основании данных вычислительных экспериментов сравнивалась эффективность статистики (2) и традиционной непараметрической регрессии

$$\tilde{y}(x)=\frac{\sum_{i=1}^{n} y^i \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}{\sum_{i=1}^{n} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}. \qquad (11)$$

В качестве искомой зависимости (1) принимался полином второй степени

$$y(\mathrm{x}) = x_1^2 + 2x_2^2 + x_1x_2 + x_3 + 0{,}5x_4 + 2x_5, \qquad (12)$$

каждый аргумент которого принимает значения из интервала $x_v \in [0;1]$, $v=\overline{1,5}$ с равномерным законом распределения. Частичные сведения о восстанавливаемой зависимости в соответствии с условиями теоремы определяются линейным полиномом

$$\bar{y}_1 = x_3 + 0{,}5x_4 + 2x_5.$$

При формировании обучающей выборки $V=\left(x_v^i, v=\overline{1,5}, y^i, i=\overline{1,n}\right)$ на значения функции (12) накладывалась аддитивная помеха

$$y^i = \psi\left(x^i\right)\left(1+2\left(\varepsilon^i - 0{,}5\right)r\right), \qquad (13)$$

где e – случайная величина с равномерным законом в диапазоне $[0;1]$; r – параметр, определяющий уровень шума.

При синтезе непараметрических моделей (2), (11) использовалась ядерная функция Епанечникова, а их оптимизация по коэффициентам размытости осуществлялась в режиме «скользящего экзамена» из условия минимума среднеквадратического критерия. При этом полагалось, что значения коэффициентов размытости $c_v = c$, $v=\overline{1,5}$ для непараметрической регрессии и $\bar{c}_v = \bar{c}$, $v=1,2$ – для модели (2), так как интервалы изменения аргументов восстанавливаемой зависимости априори одинаковые. В качестве критерия эффективности моделей (2), (11) принимались среднеквадратические отклонения W_2, W_{11} их значений от функции (12), которые оценивались по контрольной выборке V_k объема $n_k = 10000$. При этом ситуации из выборки V_k, в которых исследуемые непараметрические модели не идентифицируют значения функции (12), не участвуют в формировании критериев их эффективности. Доля таких ситуаций не превышает значений 0,06 от объема контрольной выборки.

Вычислительные эксперименты при фиксированных условиях исследования осуществлялись 60 раз. По полученным результатам определялись зависимо-

сти статистических оценок среднеквадратических отклонений W_2, W_{11} соответственно непараметрических моделей (2), (11) от объема n обучающей выборки и параметра r уровня помех в выражении (13), накладываемых на значения восстанавливаемой функции (12) (рис. 1).

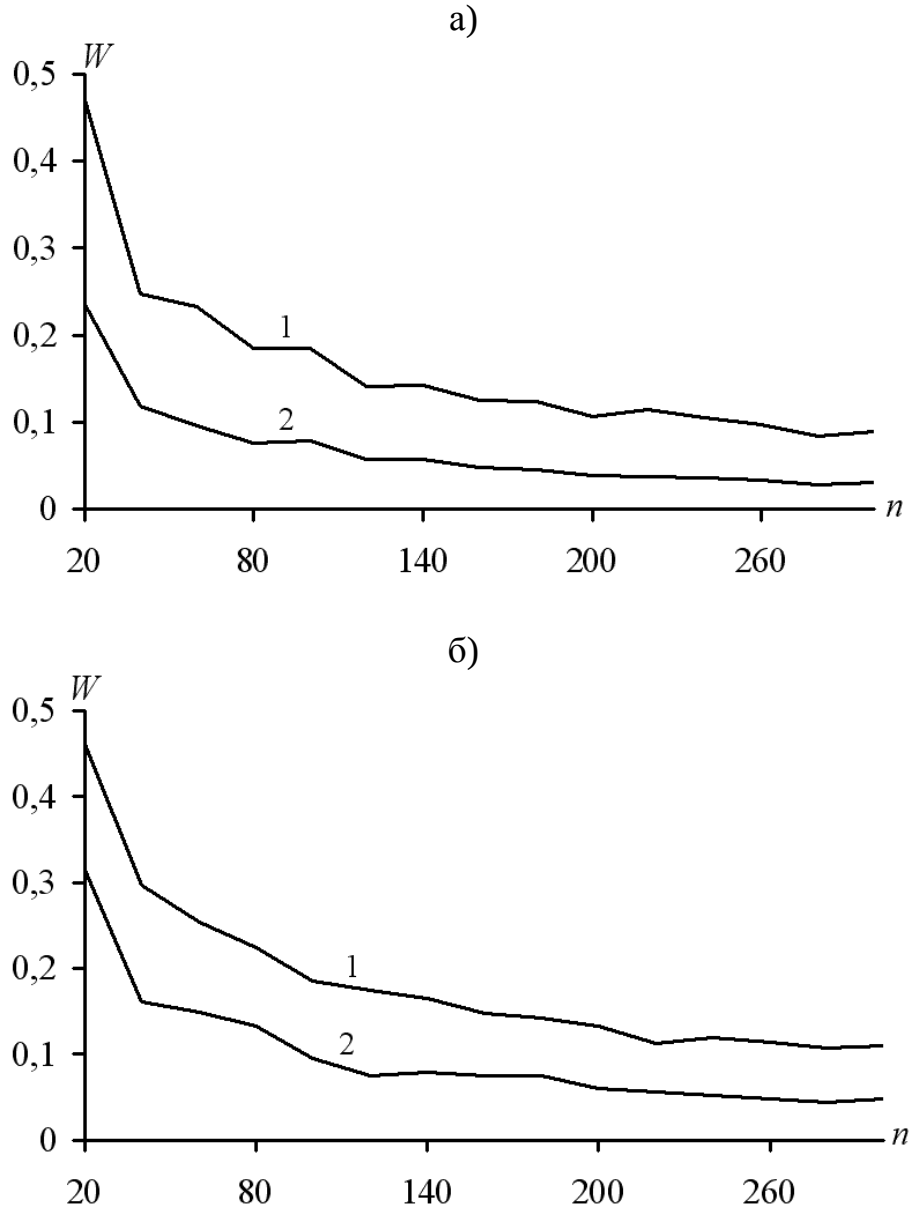

Рис. 1. Зависимости статистических оценок среднеквадратических отклонений традиционной непараметрической регрессии (11) (кривая 1) и ее модификации (2) (кривая 2) от объема n обучающей выборки при значении параметра r=0,05 (а); r=0,2 (б).

На всем диапазоне изменения n модифицированная непараметрическая регрессия (2) имеет более высокие аппроксимационные свойства по сравнению с традиционной непараметрической регрессией (11). Данная закономерность сохраняется с ростом уровня помех. Значения критериев эффективности W_2 и W_{11} достоверно отличаются при различных объемах обучающих выборок. Причем дисперсия среднеквадратического отклонения W_{11} непараметрической регрессии (11) имеет большее значение, чем для модифицированной регрессии (2) (рис. 2).

Эффективность модифицированной непараметрической модели (2) объяс-

няется возможностью снижения ее размерности за счет использования априорных сведений о наличии линейной взаимосвязи между переменными исследуемой зависимости. Данное заключение согласуется с результатами исследования гибридных моделей стохастических зависимостей [1].

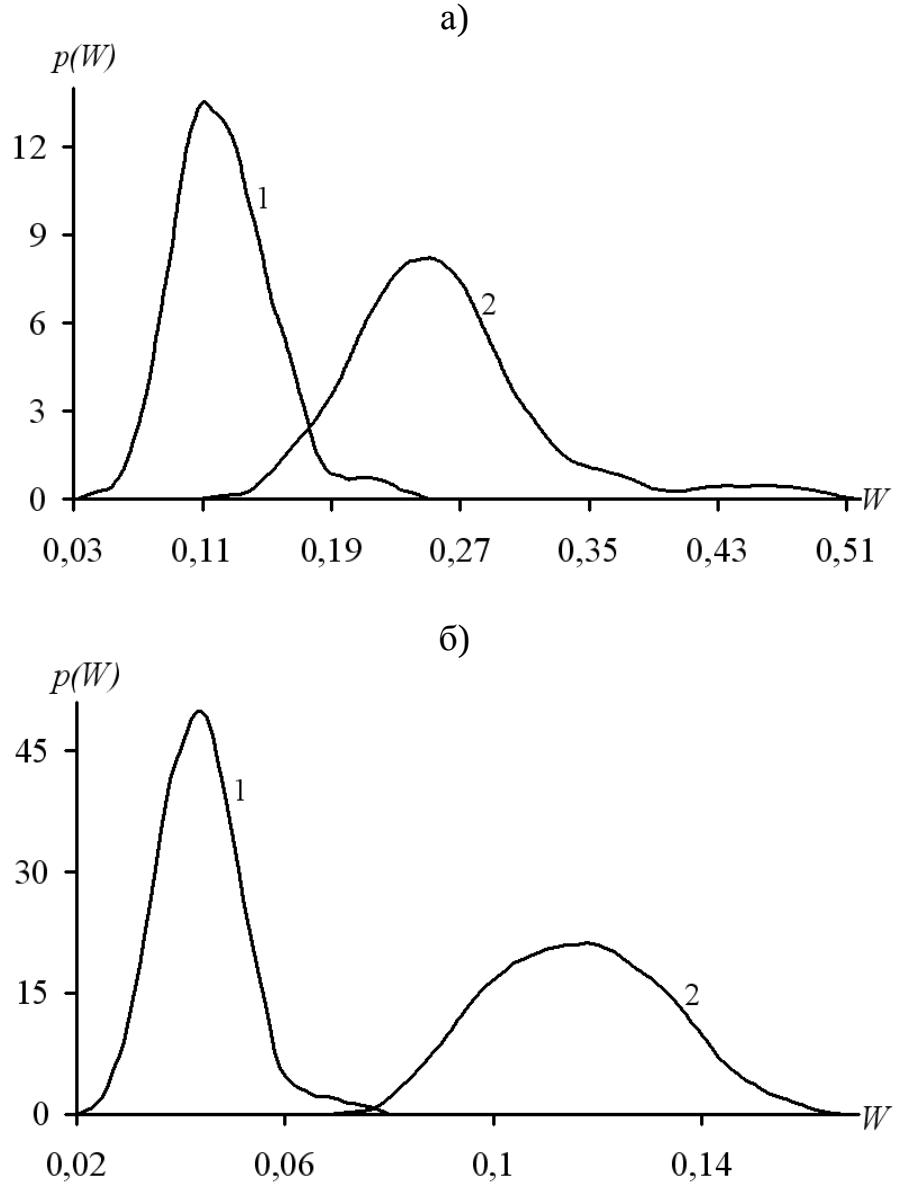

Рис. 2. Оценки плотностей вероятностей $p(W_2)$, $p(W_{11})$ среднеквадратических отклонений модифицированной непараметрической модели $\bar{\bar{y}}$ (кривая 1) и непараметрической регрессии $\tilde{y}$ (кривая 2). Условия эксперимента: объем обучающей выборки $n = 50$ (а), $n = 200$ (б); уровень шума в процедуре (13) $r = 0{,}1$.

Заключение

Традиционная непараметрическая регрессия, основанная на оценке плотности вероятности типа Розенблатта – Парзена, обобщена при построении статистических моделей в условиях наличия частичных сведений о виде восстанавливаемых зависимостей. Предлагаемая модифицированная непараметрическая регрессия обладает свойством асимптотической несмещенности. Это позволяет аналитически обосновать возможность частичного сжатия пространства признаков на

основе линейных преобразований, без существенной потери полезной информации.

Перспективное направление дальнейших исследований состоит в развитии предлагаемого подхода при анализе свойств статистических моделей, основанных на методе группового учета аргументов.

ЛИТЕРАТУРА

1. *Лапко А.В.,. Лапко В.А.* Гибридные модели стохастических зависимостей // Автометрия. – 2002. – №5. – С.38-48.
2. *Лапко В.А.* Синтез и анализ гибридных моделей стохастических зависимостей в условиях наличия их частного описания // Автометрия. – 2004. – №1. – С.51-59.
3. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. – 1962. – Vol.33. – P. 1065-1076.
4. *Надарая Э.А.* Непараметрические оценки кривой регрессии // Труды ВЦ АН ГССР. – 1965. – Вып.5. – С. 56 – 68.
5. *Ивахненко А.Г.* Непараметрический комбинированный алгоритм МГУА на операторах поиска аналогов // Автоматика. – 1990. – №5. – С. 14-27.

Статья представлена к публикации членом редколлегии Е.Л.Ереминым.

E-mail:
Лапко Александр Васильевич – lapko@icm.krasn.ru;
Лапко Василий Александрович – lapko@icm.krasn.ru.