

Проблема оценивания плотности вероятности по эмпирическим данным

Карандеев Д.А., Эйсымонт И.М.

(Институт Проблем Управления РАН; Москва; Всероссийский заочный финансово-экономический институт; Москва)

В настоящее время проблемы управления большими системами и принятия решений при большом количестве влияющих факторов часто может быть решена только с использованием достаточно точных расчетов на ЭВМ. В связи с этим возникает задача построения точных оценок плотности вероятности на основе выборок очень малого объема. Для выборок малого объема построенная в настоящей работе оценка дает более точное приближение относительно существующих. Кроме того, оценка является более устойчивой относительно выбора константы регуляризации.

1. Введение

Известно, что плотность распределения вероятностей $p(x)$ - функция, удовлетворяющая следующему уравнению:

$$\int \theta(z-x)p(x)dx = F(z), \quad (1)$$

где $F(z)$ - функция распределения, $\theta(z-x)$ - функция Хевисайда. Таким образом, функция $p(x)$ является решением интегрального уравнения Фредгольма первого рода. Задача решения такого уравнения является некорректно поставленной [5]. Для решения таких задач широко применяется метод регуляризации по Тихонову. Будем искать решение операторного уравнения:

$$Af = F, \quad (2)$$

где A - оператор, осуществляющий взаимно-однозначное отображение элементов $f(x)$ множества Φ_1 метрического пространства E_1 в элементы $F(x)$ множества Φ_2 .

метрического пространства E_2 . Рассмотрим последовательность случайных функций $F_n(x)$, $n = 1, 2, \dots$, такую, что:

$$\rho_{E_2}(F, F_n) \xrightarrow{P} 0, \text{ при } n \rightarrow \infty.$$

Суть метода регуляризации состоит в том, что по последовательности $F_n(x)$ строится последовательность функций $p_n(x)$, минимизирующих функционал:

$$R(f, F_n) = \beta_{E_2}^2(Af, F_n) + \alpha_n \Omega(f), \quad (3)$$

где $\Omega(f)$ - стабилизирующий функционал, а константы регуляризации $\alpha_n \rightarrow 0$ при $n \rightarrow \infty$.

Для стабилизирующего функционала $\Omega(f)$, удовлетворяющего следующим трем условиям:

- 1) точное решение $p(x)$ уравнения (2) принадлежит $D(\Omega)$ - области определения регуляризующего функционала $\Omega(f)$,
- 2) функционал $\Omega(f)$ принимает на $D(\Omega)$ только вещественные неотрицательные значения,
- 3) все множества $M_c = \{f : \Omega(f) \leq C\}$, $C \geq 0$ являются компактами в метрике $\rho_{E_2}(f_1, f_2)$,

доказаны следующие теоремы [1].

Теорема 1. Если для каждого n выбирается положительное α_n такое, что $\alpha_n \rightarrow 0$ при $n \rightarrow \infty$, то для любых положительных μ и v найдется такой номер $N=N(\mu, v)$, что при всех $n > N$ элементы p_n , минимизирующие функционал (3), удовлетворяют неравенству:

$$P\{\rho_{E_2}(p_n, p) > v\} \leq P\{\rho_{E_2}^2(F, F_n) > \mu \alpha_n\},$$

где $p(x)$ - точное решение операторного уравнения (2) с правой частью F .

Теорема 2. Пусть E_1 - гильбертово пространство, $\Omega(f) = \|f\|^2$ и выполнены остальные условия теоремы 1. Тогда для любого $\varepsilon > 0$ найдется такой номер $N=N(\varepsilon)$, что при $n > N(\varepsilon)$:

$$P\left(\|p_n - p\|^2 > \varepsilon\right) < 2P\left(\rho_{E_n}^2(F_n, F) > \frac{\varepsilon}{2}\alpha_n\right).$$

Эти две теоремы при правильном выборе константы регуляризации позволяют решить задачу (1) [5]. Проблема выбора константы регуляризации - это отдельная задача, которая рассматривается в [2]. Подобная задача рассматривалась для оценки плотности вероятностей из класса $L_2(-\pi, \pi)$, из класса функций, k -я производная которых интегрируема с квадратом на (a, b) и других классов. В случае, когда плотность $p(x) \in L_2(-\infty, \infty)$, была получена оценка Розенблатта-Парзена, которая строилась как решение уравнения (1) с эмпирической функцией распределения в правой части [1]. Настоящая работа посвящена построению оценки $p(x) \in L_2(-\infty, \infty)$, когда правая часть уравнения (1) - это некоторая непрерывная оценка функции распределения, построенная по выборке конечного объема.

2. Построение оценки

Пусть $X = \{x_1, x_2, \dots, x_n\}$ - выборка из неизвестного распределения. Рассмотрим следующую непрерывную функцию-полигон:

1) по точкам выборки построим новую сетку $Z = \{z_0, z_1, \dots, z_n\}$, где¹

$$z_0 = x_1, \quad z_i = \frac{x_{i+1} + x_i}{2}, \quad i = 1, \dots, n-1, \quad z_n = x_n;$$

2) значения функции в узлах сетки определим так:

$$\tilde{F}_n(z_n) = f(z_{n-1}) + \frac{1}{n}, \quad \tilde{F}_n(z_0) = 0;$$

3) по полученным точкам построим кусочно-линейную непрерывную функцию-полигон $F_n(x)$:

$$F_n(x) = \frac{1}{n} \sum_{j=0}^{n-1} \left(j + \frac{x - z_j}{(z_{j+1} - z_j)} \right) [\theta(x - z_j) - \theta(x - z_{j+1})] + \theta(x - z_n). \quad (4)$$

¹ Вообще говоря, z_0 и z_n могут быть выбраны иначе, например, $z_0 = 2x_1 - x_2$, а $z_n = 2x_n - x_{n-1}$, достаточно только выполнения условий: $z_0 < z_1$ и $z_{n-1} < z_n$.

Поскольку для рассматриваемого непрерывного полигона $F_n(x)$ и для эмпирической функции распределения $S_n(x)$, которая определяется формулой:

$$S_n(x) = \begin{cases} 0, & x < x_1, \\ \frac{k}{n}, & x_k \leq x < x_{k+1}, \quad k=1, 2, \dots, n-1, \\ 1, & x \geq x_n, \end{cases}$$

справедливо неравенство:

$$\sup_x |F(x) - F_n(x)| \leq \sup_x |F(x) - S_n(x)| + \frac{1}{n},$$

то справедлив следующий аналог теоремы Гливенко-Кантелли о сходимости эмпирической функции распределения к истинной функции распределения [4].

Теорема 3. Пусть $F_n(x)$ - функция, заданная формулой (4), а $F(x)$ - искомая функция распределения. Тогда

$$P\left\{\sup_x |F(x) - F_n(x)| \xrightarrow{n \rightarrow \infty} 0\right\} = 1.$$

Пусть искомая плотность распределения $p(x) \in L_2(-\infty, \infty)$. Будем искать $p(x)$ как решение уравнения (1) с функцией $F_n(x)$ в правой части. Согласно методу регуляризации, решение (1) может быть найдено путем минимизации в L_2 функционала (3), который в нашем случае имеет вид:

$$\begin{aligned} R_{n,\alpha} &= \left\| \int_{-\infty}^z f(x) dx - F_n(x) \right\|_{L_2}^2 + \alpha_n \|f(x)\|_{L_2}^2 = \\ &= \|Af(x) - F_n(x)\|_{L_2}^2 + \alpha_n \|f(x)\|_{L_2}^2, \end{aligned}$$

где оператор A согласно (1) задается формулой:

$$A(f(x)) = \int_{-\infty}^{+\infty} \theta(z-x) f(x) dx. \quad (5)$$

Лемма 4. Минимум функционала $R_{n,\alpha}$ достигается на решении уравнения:

$$A^* Af(x) - A^* F_n(x) + \alpha_n f(x) = 0, \quad (6)$$

где A - линейный оператор, а A^* - оператор, сопряженный к A .

Доказательство: Для отыскания минимума найдем производную Фреше от функционала

$$R(f) = \|Af(x) - F_n(x)\|^2 + \alpha_n \|f(x)\|^2$$

по функции $f(x)$.

$$\begin{aligned} R(f+h) - R(f) &= \|Af(x) + Ah(x) - F_n(x)\|^2 + \alpha_n \|f(x) + h(x)\|^2 - \\ &\quad - \|Af(x) - F_n(x)\|^2 - \alpha_n \|f(x)\|^2 = \\ &= \|Af(x) - F_n(x)\|^2 + \|Ah(x)\|^2 + 2(Af(x) - F_n(x), Ah(x)) + \alpha_n \|f(x)\|^2 + \\ &\quad + \alpha_n \|h(x)\|^2 + 2\alpha_n (f(x), h(x)) - \|Af(x) - F_n(x)\|^2 - \alpha_n \|f(x)\|^2 = \\ &= 2(Af(x) - F_n(x), Ah(x)) + 2\alpha_n (f(x), h(x)) + o(\|h(x)\|) = \\ &= 2(A^*(Af(x) - F_n(x)), h(x)) + 2\alpha_n (f(x), h(x)) + o(\|h(x)\|) = \\ &= 2[\{A^*(Af(x) - F_n(x)) + \alpha_n f(x)\}, h(x)] + o(\|h(x)\|). \end{aligned}$$

Выражение в фигурных скобках есть производная Фреше функционала $R_{n,\alpha}$ по $f(x)$, а следовательно, минимум достигается на решении уравнения:
 $A^*Af(x) - A^*F_n(x) + \alpha_n f(x) \equiv 0$. Лемма доказана.

Согласно (5) уравнение (6) примет вид:

$$\int_{-\infty}^{\infty} \theta(t-x) \left[\int_{-\infty}^{\infty} \theta(t-\tau) f(\tau) d\tau - F_n(t) \right] dt + \alpha_n f(x) = 0. \quad (7)$$

Теорема 5. Решением уравнения (7) является функция:

$$\begin{aligned} p(x) &= \frac{1}{2n} \left[\sum_{j=0}^{n-1} \operatorname{sign}(x - z_j) \lambda_j e^{-\frac{|x-z_j|}{\sqrt{\alpha_n}}} + \right. \\ &\quad + \frac{1}{z_n - z_{n-1}} e^{-\frac{|x-z_n|}{\sqrt{\alpha_n}}} \operatorname{sign}(x - z_n) - \frac{1}{z_1 - z_0} e^{-\frac{|x-z_0|}{\sqrt{\alpha_n}}} \operatorname{sign}(x - z_0) + \\ &\quad \left. + 2 \sum_{j=0}^{n-1} \frac{1}{(z_{j+1} - z_j)} (\theta(z_{j+1} - x) - \theta(z_j - x)) \right], \end{aligned} \quad (8)$$

$$\text{где } \lambda_j = \frac{1}{(z_j - z_{j+1})} = \frac{1}{(z_{j+1} - z_j)}.$$

Доказательство: Применим к уравнению (7) преобразование Фурье (в смысле обобщенных функций), учитывая, что преобразование Фурье свертки функций равно произведению преобразований Фурье этих функций, получаем уравнение:

$$\left(-\frac{1}{iu} + \pi\delta(u)\right) \left[\left(-\frac{1}{iu} + \pi\delta(u)\right) \hat{f}(u) - \phi(F_n(x)) \right] + \alpha_n \hat{f}(u) = 0, \quad (9)$$

где $\phi(F_n(x)) = \int_{-\infty}^{\infty} F_n(x) e^{-iux} du$ - преобразование Фурье функции $F_n(x)$, $\hat{f}(u)$ - преобразование Фурье функции $f(x)$. Преобразуем выражение (4) для $F_n(x)$ к следующему виду:

$$F_n(x) = \frac{1}{n} \sum_{j=0}^{n-1} \left(j - \frac{z_j}{(z_{j+1} - z_j)} \right) [\theta(x - z_j) - \theta(x - z_{j+1})] + \\ + \frac{1}{n} \sum_{j=0}^{n-1} \frac{x}{(z_{j+1} - z_j)} [\theta(x - z_j) - \theta(x - z_{j+1})] + \theta(x - z_n).$$

Пользуясь преобразованиями Фурье для функции Хевисайда и функции $x\theta(x-a)$ [3], получаем преобразование Фурье функции $F_n(x)$:

$$\phi(F_n(x)) = \frac{i}{nu} \sum_{j=0}^{n-1} (e^{-iuz_{j+1}} - e^{-iuz_j}) \left(j - \frac{i}{u(z_{j+1} - z_j)} \right) + \frac{i}{nu} \sum_{j=1}^{n-1} e^{-iuz_{j+1}} + \pi\delta(u) - \frac{i}{u} e^{-iua}.$$

После подстановки полученного выражения в (9), уравнение принимает вид:

$$\left(-\frac{1}{iu} + \pi\delta(u)\right) \left[-\frac{1}{iu} \hat{f}(u) + \pi\delta(u) \int_{-\infty}^{\infty} f(x) e^{-iux} dx - \right. \\ \left. - \frac{i}{un} \sum_{j=0}^{n-1} (e^{-iuz_{j+1}} - e^{-iuz_j}) \left(j - \frac{i}{u(z_{j+1} - z_j)} \right) - \frac{i}{un} \sum_{j=1}^{n-1} e^{-iuz_{j+1}} - \pi\delta(u) + \frac{i}{u} e^{-iua} \right] + \alpha_n \hat{f}(u) = 0. \quad (10)$$

В силу свойств δ -функции Дирака ($\delta(u)f(u) = \delta(u)f(0)$) и плотности вероятности:

$$\pi\delta(u) \int_{-\infty}^{\infty} f(x) e^{-iux} dx = \pi\delta(u).$$

Члены в квадратных скобках, не содержащие величины $\hat{f}(u)$, могут быть преобразованы к виду:

$$-\frac{1}{iu^2} \sum_{j=1}^{n-1} \lambda_j e^{-iux_j} - \frac{1}{iu^2} \left(\frac{e^{-iux_n}}{(z_n - z_{n-1})} - \frac{e^{-iux_0}}{(z_1 - z_0)} \right).$$

После такого преобразования уравнение (10) примет вид:

$$\begin{aligned} & -\frac{1}{u^2} \hat{f}(u) + \pi\delta(u) \frac{1}{iu} \hat{f}(u) + \alpha_n \pi \hat{f}(u) - \\ & - \left(\frac{1}{iu} + \pi\delta(u) \right) \frac{1}{iu^2} \left[\sum_{j=1}^{n-1} \lambda_j e^{-iux_j} + \left(\frac{e^{-iux_n}}{(z_n - z_{n-1})} - \frac{e^{-iux_0}}{(z_1 - z_0)} \right) \right] = 0. \end{aligned}$$

Это уравнение эквивалентно следующему:

$$\begin{aligned} & \hat{f}(u) - \pi\delta(u) i \hat{f}(u) u + \alpha_n u^2 \hat{f}(u) + \\ & + \frac{1}{n} \left(\frac{1}{iu} + \pi\delta(u) \right) \left[\sum_{j=1}^{n-1} \lambda_j e^{-iux_j} + \left(\frac{e^{-iux_n}}{(z_n - z_{n-1})} - \frac{e^{-iux_0}}{(z_1 - z_0)} \right) \right] = 0. \end{aligned}$$

В силу свойств функции Дирака:

$$\begin{aligned} & \pi\delta(u) i \hat{f}(u) u = 0, \\ & \pi\delta(u) \left[\sum_{j=1}^{n-1} \lambda_j e^{-iux_j} + \left(\frac{e^{-iux_n}}{(z_n - z_{n-1})} - \frac{e^{-iux_0}}{(z_1 - z_0)} \right) \right] = 0, \end{aligned}$$

а оставшиеся члены дают следующее выражение для $\hat{f}(u)$:

$$\hat{f}(u) = -\frac{1}{iu(1+\alpha_n u^2)} \left[\sum_{j=1}^{n-1} e^{-iux_j} \lambda_j + \left(\frac{e^{-iux_n}}{(z_n - z_{n-1})} - \frac{e^{-iux_0}}{(z_1 - z_0)} \right) \right].$$

Применяя обратное преобразование Фурье к функции $\hat{f}(u)$ и учитывая, что:

$$\phi_{inv} \left(\frac{e^{-iux}}{u(1+\alpha_n u^2)} \right) = -\frac{i}{2} \left(-1 + 2\theta(a-x) + e^{\frac{1}{\sqrt{\alpha}}(\sigma-x)} \theta(x-a) - e^{-\frac{1}{\sqrt{\alpha}}(a-x)} \theta(a-x) \right),$$

получаем решение уравнения (6). Это и будет оценка (8). Теорема доказана.

Теоремы 1, 2, 3 дают сходимость построенной оценки в L_2 .

3. Сравнение двух оценок

На основе эксперимента сравним новую оценку (8) с известной оценкой Розенблатта-Парзена с экспоненциальным ядром:

$$p(x) = \sum_{j=1}^n \frac{1}{2n\sqrt{\alpha}} e^{-\frac{|x-x_j|}{\sqrt{\alpha}}} \quad (11)$$

На основе выборки объемом $n=40$, состоящей из смеси трех нормальных распределений $N(m=2, \sigma=2)$, $N(m=4, \sigma=0.2)$ и $N(m=1, \sigma=0.2)$, взятых с весами $1/4$, $1/4$ и $1/2$, построены две оценки: оценка (11) (см. рис. 1) и оценка (8) (см. рис. 2). Для сравнения на рисунках приведен график истинного распределения. Константа регуляризации в обоих случаях вычислялась как $\alpha_n = 1/n$.

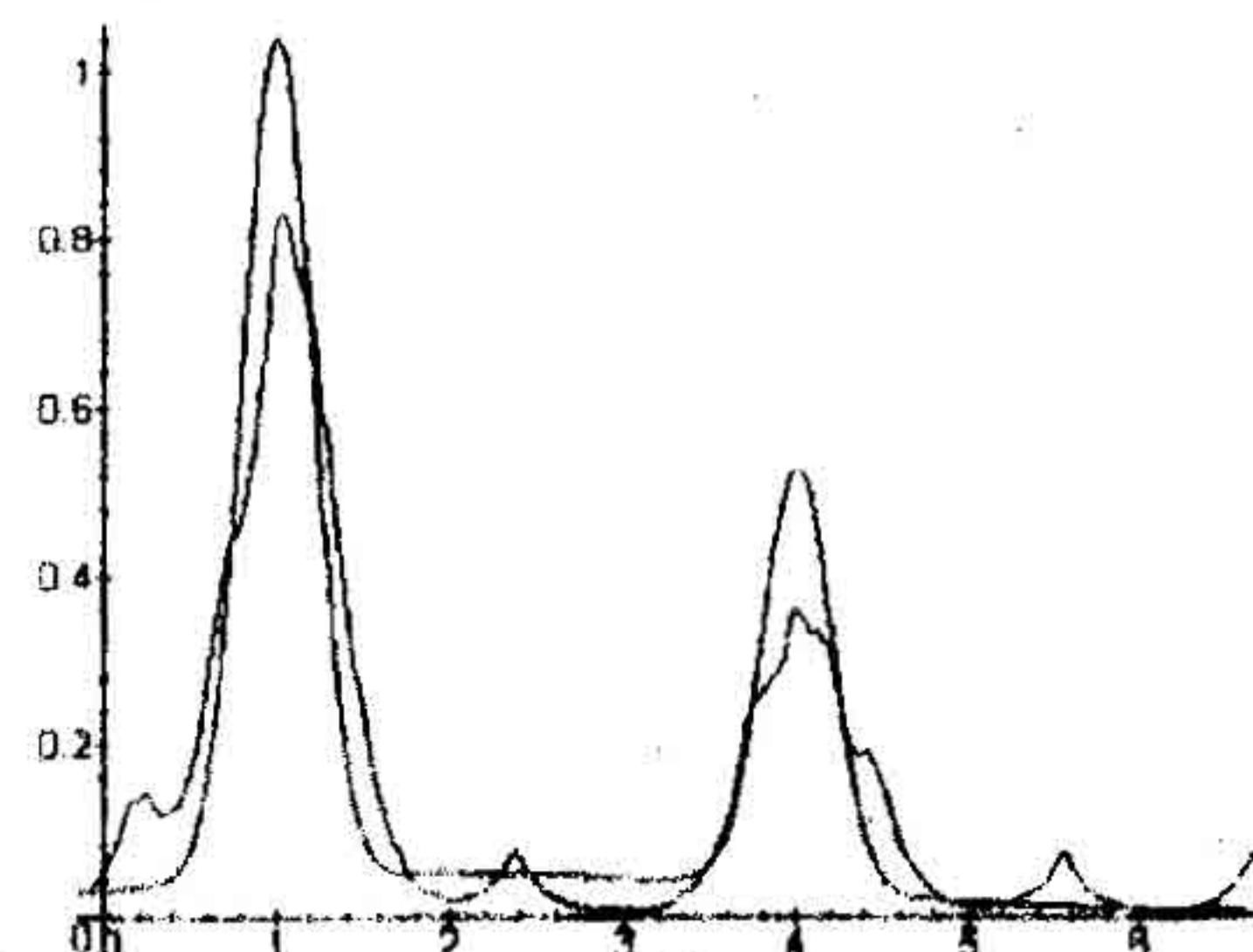


Рис. 1

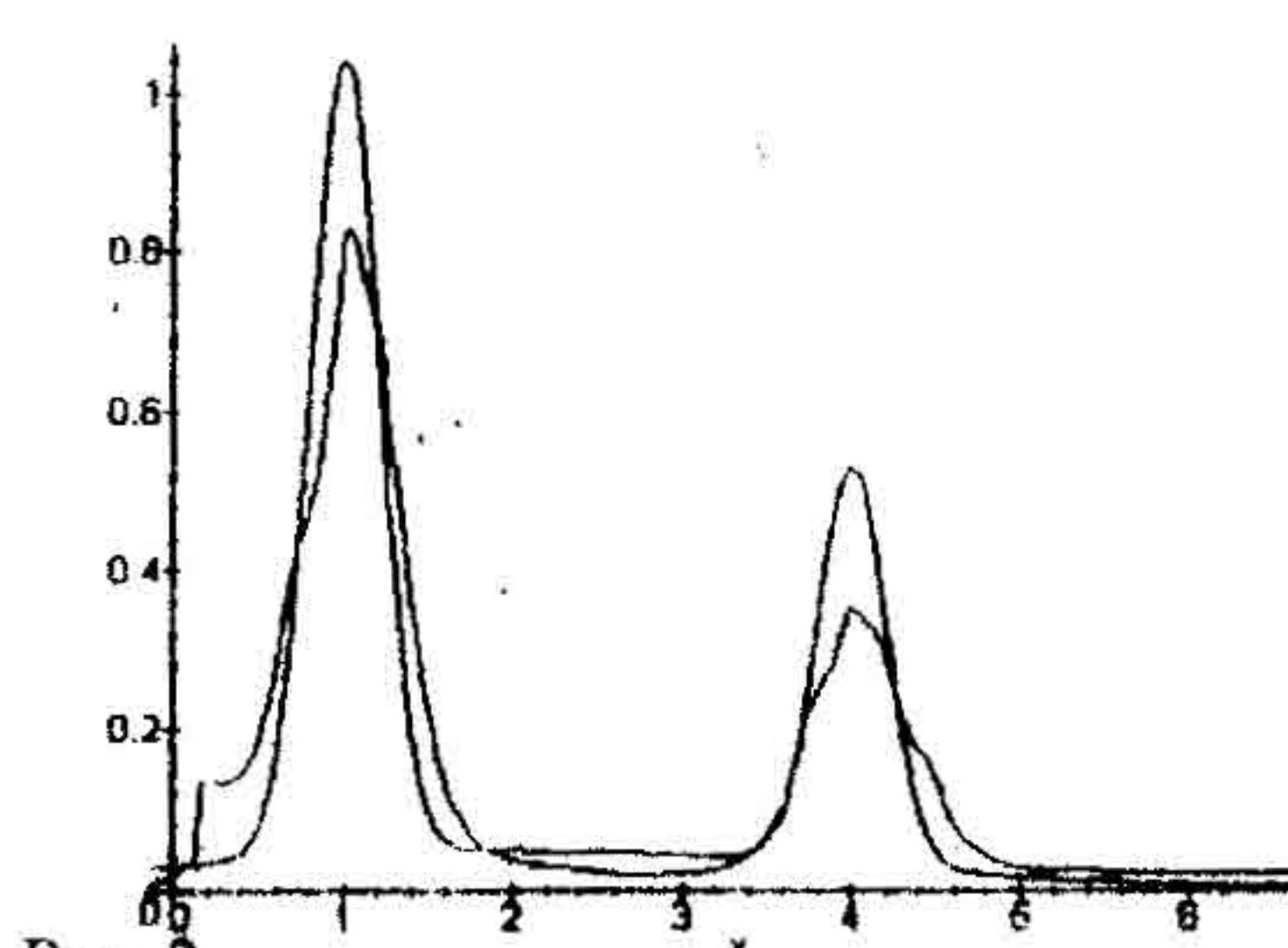


Рис. 2

Оценки для той же выборки с $\alpha_n = 0.01$ приведены на рис. 3 - оценка (11), рис. 4 - новая оценка. Из приведенных рисунков видно, что оценка (8) имеет гораздо меньше локальных максимумов. Кроме того, на рис. 1 и 3 в точке 2 оценка (11) практически обращается в ноль, что не соответствует действительности, т.к. одно из рассматриваемых распределений нормальное $N(m=2, \sigma=2)$. В то же время оценка (8) в точке 2 дает значение существенно отличное от нуля (рис. 2, 4).

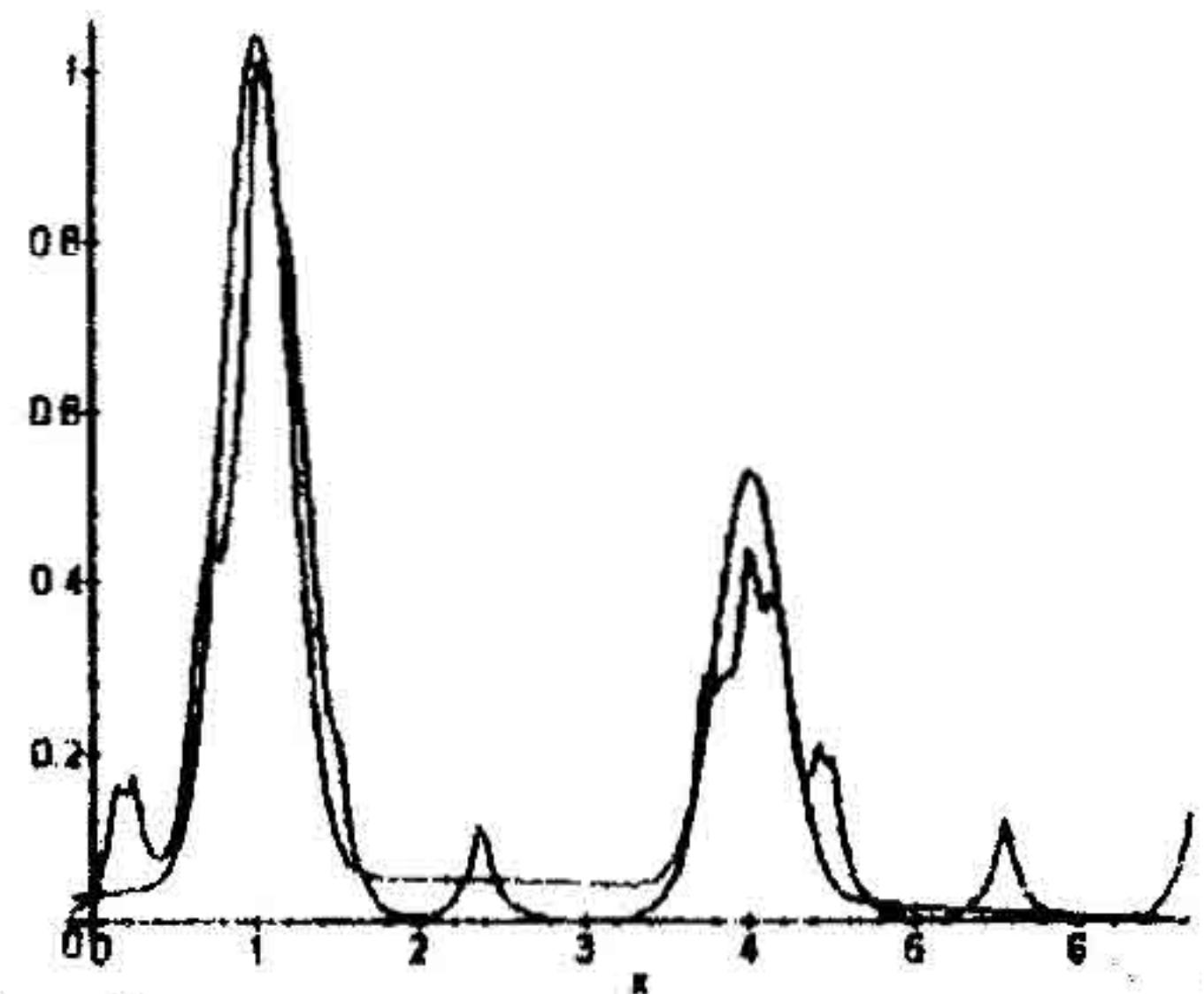


Рис. 3

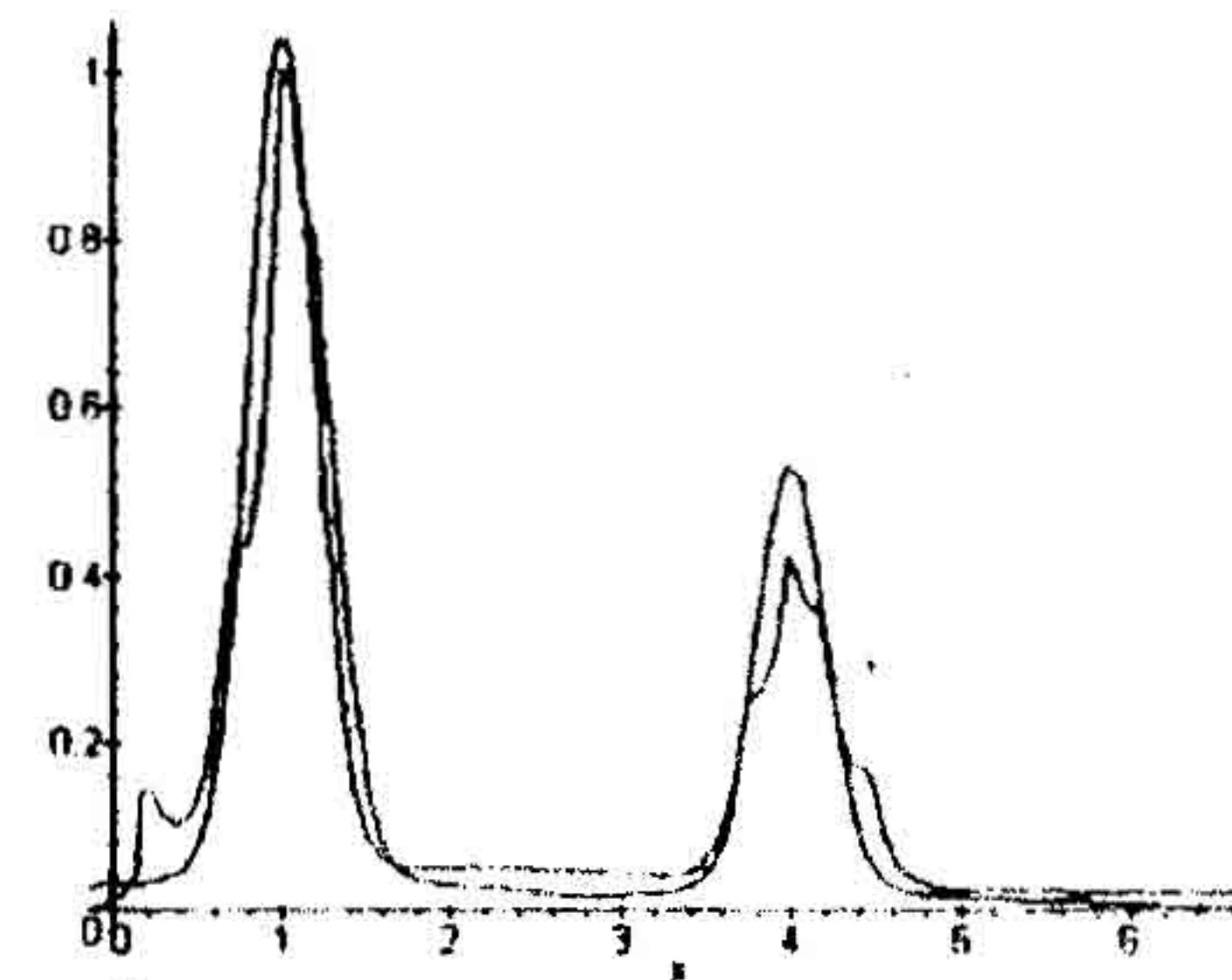


Рис. 4

На рис. 5 приведена оценка (11), построенная на основе выборки объема $n=20$ из распределения Коши ($m = 1, \sigma = 1$). Для той же выборки оценка (8) приведена на рис. 6. Константа регуляризации в обоих случаях $\alpha_n = 0.01$.

Оценка, приведенная на рис. 5, похожа на смесь нескольких нормальных распределений из-за нескольких резких всплесков, в то время, как на рис. 6 видно, что это выборка из распределения с одним максимумом. Оценки для той же выборки с $\alpha_n = 1/n$ приведены на рис. 7 - оценка (11), новая оценка (8) - рис. 8.

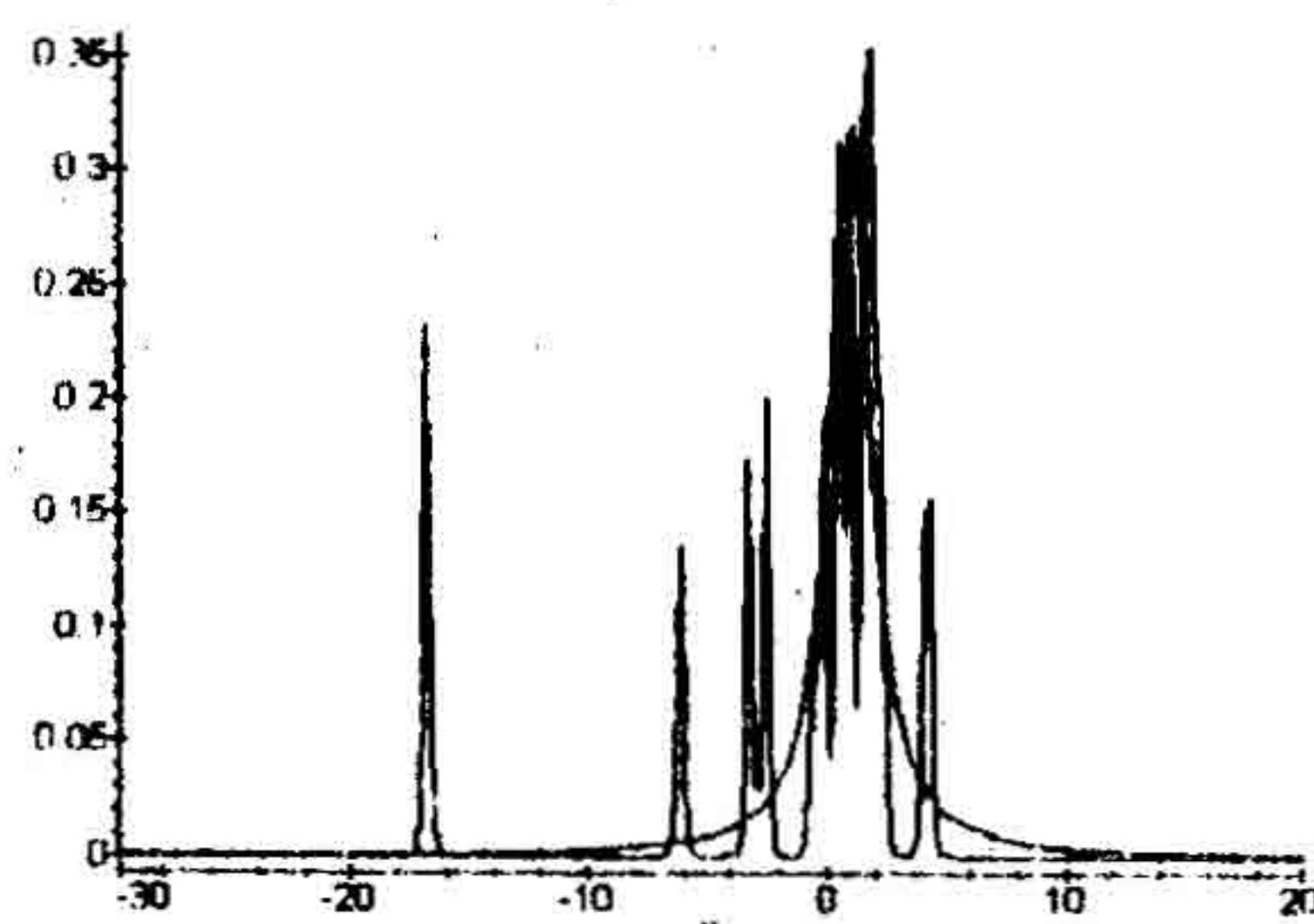


Рис. 5

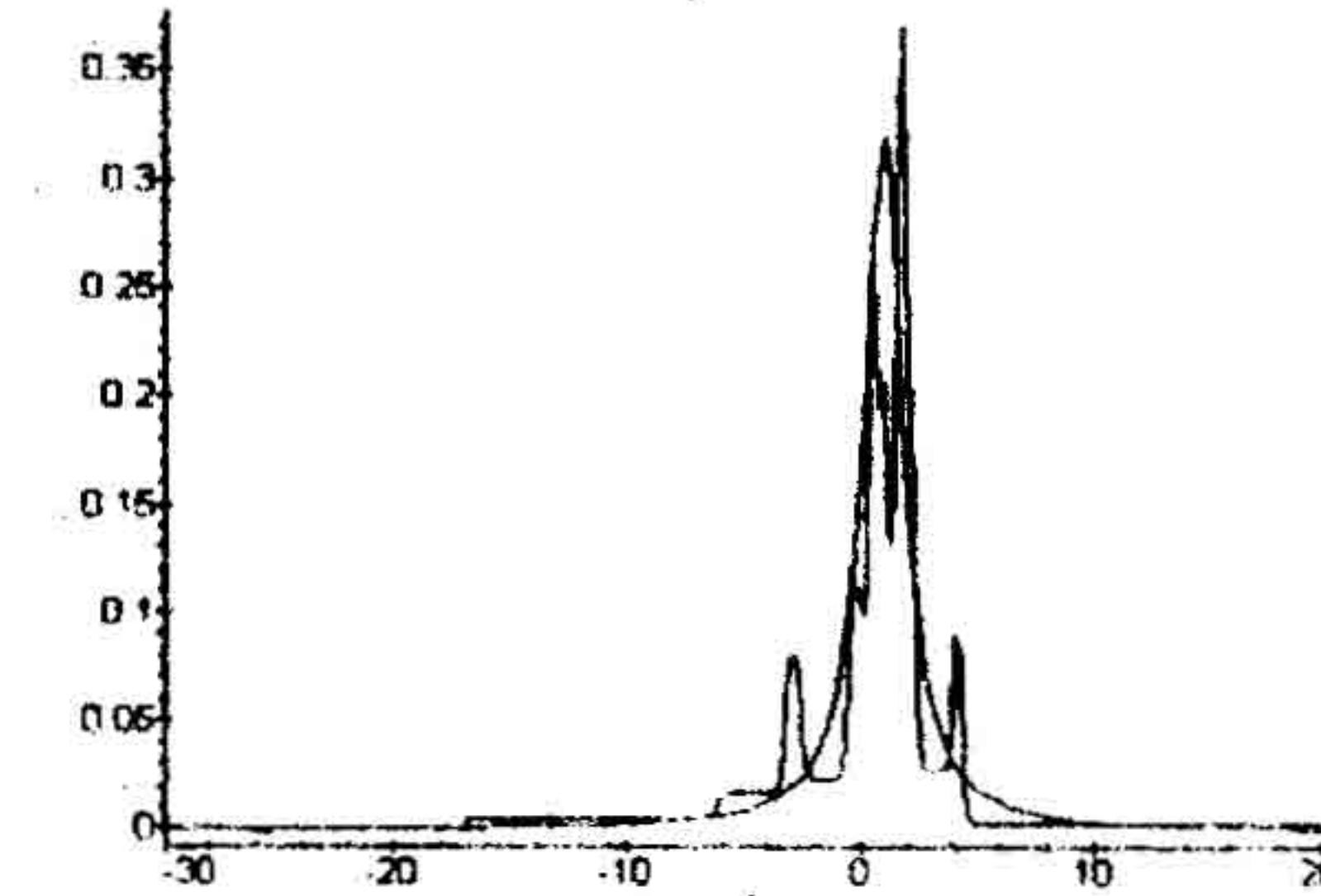


Рис. 6

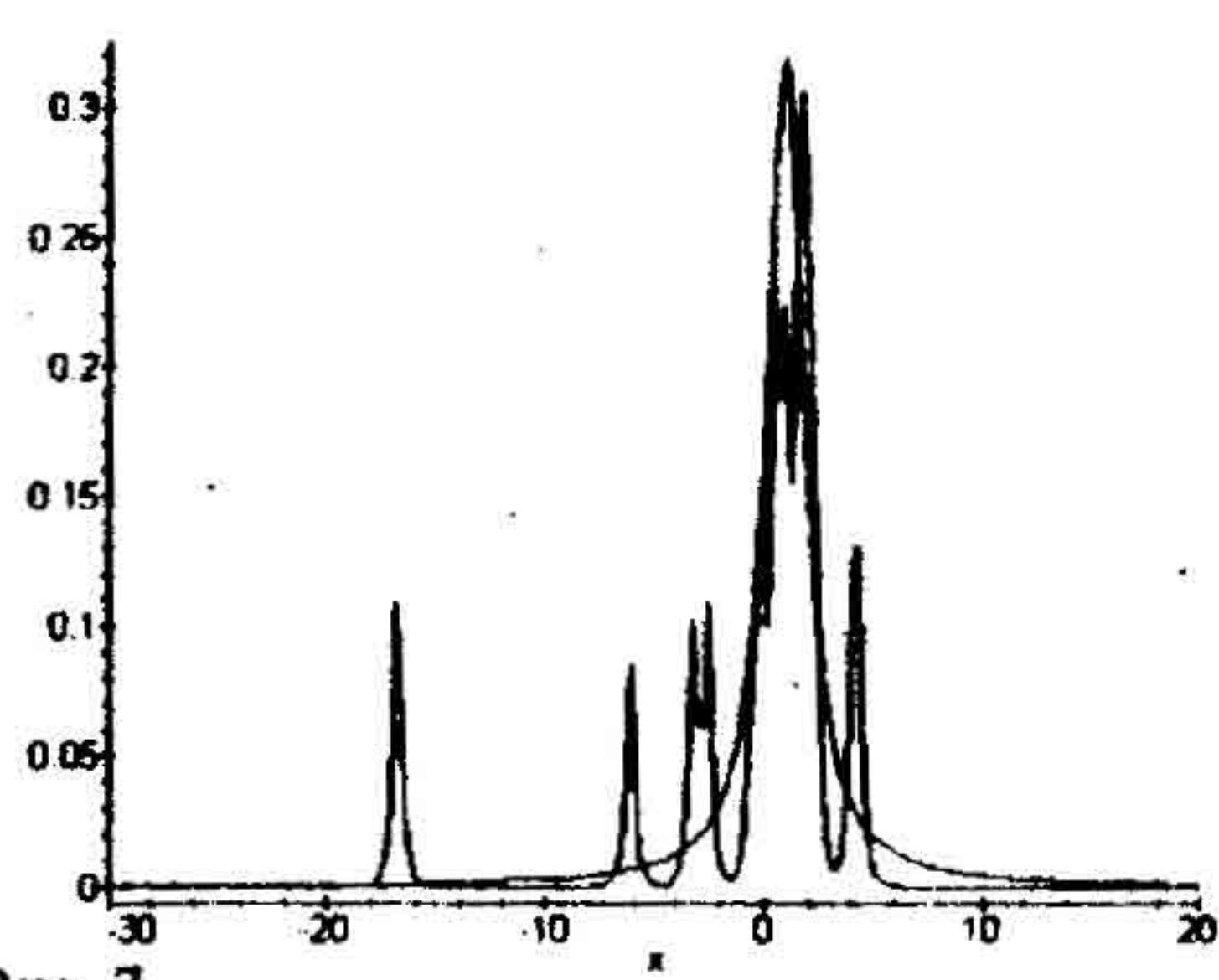
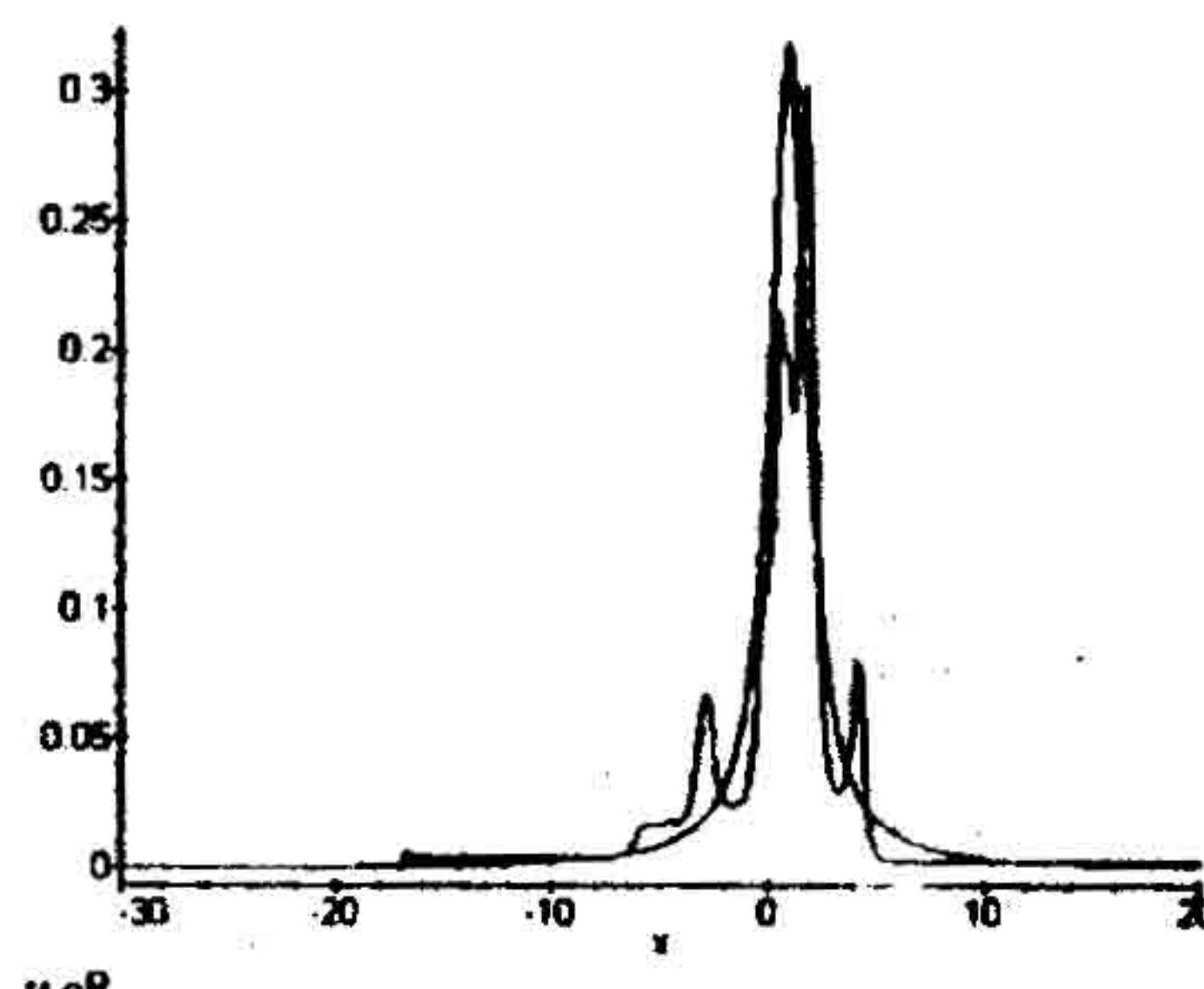


Рис. 7



и.с.8

Сравнительный анализ рассматриваемых оценок показал, что оценка (8), построенная на основе непрерывного полигона (4) для выборок из любых распределений, в некотором смысле лучше, чем оценка (11). В частности, оценка (8) дает лучшее приближение в случае, когда в восстанавливаемой плотности вероятности имеются "узкие" пики или "тяжелые хвосты". Экспериментальное сравнение проводилось на выборках из нормального распределения, гамма-распределения и распределения Коши. Кроме этого, эксперимент показал, что неправильно выбранный параметр регуляризации α_n гораздо меньше искажает оценку (8), чем оценку (11). Возможные способы подбора параметра регуляризации для рассматриваемых методов приведены в [2].

Литература:

- [1] Айду Ф.А., Вапник В.Н. Оценивание плотности вероятностей на основе метода стохастической регуляризации //Автоматика и Телемеханика. N4. 1988. С. 84-97.
- [2] Карапеев Д.А., Стефанюк А.Р. Выбор параметров частройки алгоритма при восстановлении функции плотности вероятности по эмтическим данным //Автоматика и Телемеханика. 1996. N10. С. 95-111.
- [3] Крейн С.Г. Функциональный анализ. М.: Наука, 1972.
- [4] Смирнов Н.В. Приближение законов распределения случайных величин по эмтическим данным //Успехи матем. наук. X (1944). С. 179-206.
- [5] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука. 1986.