УДК 519.254

И. В. Пономарев, В. В. Славский

РАВНОМЕРНО НЕЧЕТКАЯ МОДЕЛЬ ЛИНЕЙНОЙ РЕГРЕССИИ*

В работе изучается нечеткая и стандартная модель линейной регрессии в случае одного аргумента. Даются геометрическая интерпретация нечеткой модели и ее сравнение со стандартной моделью линейной регрессии. Исследуется вычислительная сложность нечеткой модели линейной регрессии. Указываются эффективные алгоритмы решения, имеющие сложность порядков $O(n \log n)$, $O(n^2)$, и их реализация в среде MatLab.

Ключевые слова: нечеткая модель линейной регрессии, сложность вычислений.

Введение

В классической постановке задача регрессии может быть сформулирована следующим образом: имеется некоторое изучаемое устройство F — «черный ящик», на вход которого подается значение x_i, а на выходе получается наблюдаемое значение y_i. Значение y_i зависит от x_i, от устройства ящика и от случайных или скрытых параметров. Простейшая математическая модель устройства F может быть сформулирована в виде равенства

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

где $f \in \Phi$ — однозначная числовая функция из некоторого семейства Φ, описывающего гипотетическое устройство F, ε_i — величины отклонения выходных значений реального устройства F от его модели f.

Задача состоит в выборе $f \in \Phi$, для которой ε_i — минимальны в некотором смысле. При статистическом подходе обычно предполагают, что ε_i — случайные независимые нормально распределенные величины с нулевым математическим ожиданием и некоторой дисперсией. Эти предположения основаны на вероятностном подходе и при построении математической модели не всегда оправданы. Хотя вероятностный подход к регрессии имеет многие важные приложения, но в некоторых ситуациях возникают проблемы:

- Количество наблюдений неадекватно, т. е. небольшой набор данных (x_i, y_i).
- Трудности с проверкой предположения о вероятностном законе распределения ε_i.
- Неясность в закономерности между входными и выходными переменными.

В данной работе исследуется одна из возможных моделей регрессии, основанная на теории нечетких множеств Л. Заде. Одной из первых работ в этом направлении стала

*Работа выполнена при финансовой поддержке РФФИ (проект 08-01-98001-р_сибирь_а), а также ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг. (гос. контракт № 02.740.11.0457).

ISSN 1818-7897. Вестник НГУ. Серия: Математика, механика, информатика. 2010. Т. 10, вып. 2. С. 118–134

статья [9], в которой были смягчены некоторые из предположений классического статистического подхода, связанные с неуверенностью в закономерности между входящими и выходящими данными. Основная модель Танака имеет вид линейной нечеткой функции:

$$Y = A_0 + A_1 x, \tag{1}$$

где x — известные переменные, A_i — нечеткие треугольные числа с центром c_i и шириной w_i $A_i = (c_i, w_i)$. Другие авторы также начинают рассмотрение с такой же зависимостью, из недавних работ можно указать [6; 10]. Принципиальные отличия начинают появляться на этапе формирования функционала качества модели. Обычно этот функционал и соответственно качество модели зависят от степени нечеткости коэффициентов. Нечеткие регрессионные модели в настоящее время активно исследуются и применяются [7].

В данной работе предложена более универсальная «равномерно нечеткая» модель зависимости, которая, с одной стороны, примыкает к классическому регрессионному анализу путем замены квадратичной нормы на Чебышескую равномерную норму, с другой — содержит частный случай модели Танака, когда A_1 — четкое число. Предложенный подход позволяет избежать увеличения рассеяния зависимой переменной при возрастании амплитуды независимой переменной [8]. Дается геометрическая интерпретация этой задачи, указываются эффективные алгоритмы решения и их реализация в среде MatLab.

Будем предполагать, что устройству F соответствует равенство

$$A = f(x),$$

где $f \in \Phi$ — нечеткая числовая функция из некоторого семейства Φ, описывающего гипотетическое устройство F, т. е. аргументу x сопоставляется нечеткое числовое значение $A = f(x)$.

Определение 1. Нечетким числом A называется множество $\{(y, \mu_A(y))\}$, где $y \in R$, $\mu_A(y) : R \to [0, 1]$, и функция принадлежности $\mu_A(y)$ которого является выпуклой и унимодальной.

Определение 2. Функция принадлежности $\mu_A(y)$ называется строго унимодальной на промежутке $[a, b] \subset R$, если она непрерывна на промежутке $[a, b]$, а также существует некоторая точка $c \in [a, b]$ такая, что:

- функция $\mu_A(y)$ строго монотонно возрастает на промежутке $[a, c]$;
- функция $\mu_A(y)$ строго монотонно убывает на промежутке $[c, b]$;
- функция $\mu_A(y)$ принимает свое максимальное значение в точке c. Точка c называется модой нечеткого числа.

Наблюдаемое значение $y_i \in R$, соответствующее $x_i \in R$, будем рассматривать как дефазификацию нечеткого числа A_i. $\mu_{A_i}(y_i)$ — степень достоверности этого наблюдаемого значения.

Определение 3. Величину, равную

$$\delta(f) = \min_{i=1,\ldots,n} \{\mu_{A_i}(y_i)\},$$

назовем степенью достоверности модели f.

Далее будем предполагать, что функция принадлежности $\mu_A(y)$ имеет конкретный вид

$$\mu_A(y) = \varphi\left(\frac{|f_0(x) - y|}{\sigma}\right),$$

где $\varphi : [0,\infty) \to [0,1]$ — фиксированная убывающая функция $\varphi(0) = 1$, $\sigma > 0$ — параметр, $f_0(x)$ – однозначная числовая функция, равная моде нечеткого числа $A = f(x)$. Функция f_0 принадлежит некоторому фиксированному семейству Φ_0 числовых функций. Очевидно, что

$$\delta(f) = \varphi\left(\frac{\max_{i=1,\dots,n}|f_0(x_i) - y_i|}{\sigma}\right).$$

Таким образом, задача нахождения наиболее достоверной модели свелась к нахождению

$$\max_{f_0\in\Phi_0}\delta(f) = \varphi\left(\frac{\alpha}{\sigma}\right),$$

где

$$\alpha = \min_{f_0\in\Phi_0}\max_{i=1,\dots,n}|f_0(x_i) - y_i|.$$

Константу σ определяем из условия нормировки достоверности модели, например $\varphi\left(\frac{\alpha}{\sigma}\right) = 0{,}95$. Для линейной регрессии семейство Φ_0 состоит из линейных функций вида $y = kx + b$. Соответствующая математическая задача сводится к нахождению минимума

$$\alpha_\infty = \min_{k,b}\max_{i=1,\dots,n}|kx_i + b - y_i|.$$

Знаком ∞ будем подчеркивать происхождение этой величины от чебышевской нормы в силу известного равенства

$$\max_i\{|y_i|\} = \lim_{p\to+\infty}\Big(\sum_i |y_i|^p\Big)^{\frac{1}{p}} = \lim_{p\to+\infty}\|y\|_p.$$

Геометрически данная задача сводится к нахождению полосы, заключенной между двумя параллельными прямыми минимальной «ширины вдоль оси OY» и содержащей множество точек

$$\Omega = \{(x_i, y_i) : i = 1,\dots,n\}.$$

Замечание 1. При вероятностном подходе к данной задаче находят минимум квадратичного отклонения

$$\alpha_2 = \min_{k,b}\Big(\sum_{i=1}^n\big(kx_i + b - y_i\big)^2\Big)^{\frac{1}{2}},$$

при этом среднее значение отклонений ε_i будет равно нулю:

$$\frac{1}{n}\sum_{i=1}^n(kx_i + b - y_i) = 0.$$

Данная задача имеет явное решение

$$k = \frac{n\left(\sum_{i=1}^n x_iy_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{D(x)},$$

$$b = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i y_i\right)\left(\sum_{i=1}^n x_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} = \overline{y} - k \cdot \overline{x},$$
$$\alpha_2 = \left(nD(y)(1 - r_{xy}^2)\right)^{\frac{1}{2}},$$

где $\overline{x}$ — среднее арифметическое, $D(x)$ — дисперсия, r_{xy} — коэффициент парной корреляции.

Сложность вычисления этих выражений (число операций умножения или деления) пропорциональна n. Заметим, что в данной постановке задачи она эквивалентна нахождению максимума

$$\max_{k,b} \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(kx_i + b - y_i)^2}{2\sigma^2}\right) = \max_{k,b} \prod_{i=1}^n \frac{1}{\sigma} p\left(\frac{|kx_i + b - y_i|}{\sigma}\right),$$

где $p(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$, что соответствует принципу максимального правдоподобия, а параметр σ находится из условия нормировки.

Замечание 2. Введем обозначение для минимума p-отклонения:

$$\alpha_p = \alpha_p(X, Y) = \min_{k,b}\Big(\sum_{i=1}^n |kx_i + b - y_i|^p\Big)^{\frac{1}{p}},$$

где $1 \le p \le \infty$, $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_n\} \in R^n$.

§ 1. Основные результаты

Теорема 1. *Пусть $n > 1$ и $x_i \ne x_j$ при $i \ne j$. Тогда существует единственная прямая $y = k_0 x + b_0$, на которой достигается минимум*

$$\alpha_\infty = \min_{k,b} \max_{i=1,\ldots,n} |kx_i + b - y_i|.$$

ДОКАЗАТЕЛЬСТВО. Функции $|kx_i + b - y_i|$ от аргументов k, b — выпуклы вниз, следовательно верхняя огибающая этих функций

$$\beta(k, b) = \max_{i=1,\ldots,n} |kx_i + b - y_i|$$

также выпукла вниз. Для произвольного набора чисел $\{c_i\}$ справедливо равенство

$$\max_{i=1,\ldots,n} |c_i - b| = \frac{\max_{i=1,\ldots,n} c_i - \min_{i=1,\ldots,n} c_i}{2} + \left|\frac{\max_{i=1,\ldots,n} c_i + \min_{i=1,\ldots,n} c_i}{2} - b\right|. \quad (2)$$

Используя (2) и полагая

$$g_1(k) = \frac{\max_{i=1,\ldots,n}(kx_i - y_i) - \min_{i=1,\ldots,n}(kx_i - y_i)}{2},$$
$$g_2(k) = \frac{\max_{i=1,\ldots,n}(kx_i - y_i) + \min_{i=1,\ldots,n}(kx_i - y_i)}{2},$$

получим равенство

$$\beta(k, b) = g_1(k) + |b + g_2(k)|.$$

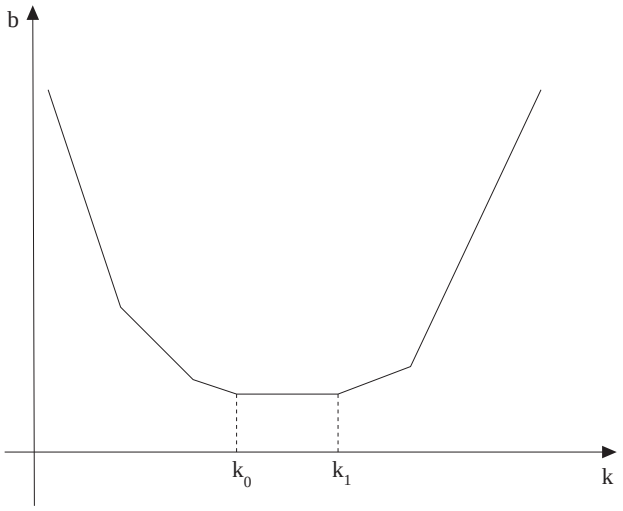

Рис. 1. Функция $g_1(x)$

Следовательно, полагая $b = -g_2(k)$, получим

$$\alpha_\infty = \min_{k,b} \beta(k,b) = \min_k g_1(k). \tag{3}$$

Докажем единственность прямой $y = k_0x + b_0$, на которой достигается минимум. Предположим противное, что таких прямых две: $y = k_0x + b_0$ и $y = k_1x + b_1$.

Если $k_0 = k_1$, то согласно формуле $\beta(k,b) = g_1(k) + |b + g_2(k)|$ получим, что $b_0 = b_1$. Следовательно, $k_0 \neq k_1$ и

$$g_1(k_0) = g_1(k_1).$$

Функции

$$\begin{aligned} \varphi^+(k) &= \frac{\max\limits_{i=1,\dots,n}(kx_i - y_i)}{2}, \\ \varphi^-(k) &= \frac{\min\limits_{i=1,\dots,n}(kx_i - y_i)}{2} \end{aligned}$$

от аргумента k кусочно-линейные и соответственно выпуклые вниз и вверх. Следовательно, функция $g_1(k)$ выпуклая вниз и равна константе на отрезке $[k_0, k_1]$ (рис 1). Поэтому сужения функций $\varphi^+(k)$ и $\varphi^-(k)$ на отрезок $[k_0, k_1]$ — линейные функции с одним и тем же угловым коэффициентом

$$\begin{aligned} \varphi^+(k) &= \frac{kx_{i_0} - y_{i_0}}{2}, \quad k \in [k_0, k_1], \\ \varphi^-(k) &= \frac{kx_{i_1} - y_{i_1}}{2}, \quad k \in [k_0, k_1], \end{aligned}$$

т. е. $x_{i_0} = x_{i_1}$, $y_{i_0} \neq y_{i_1}$ — получили противоречие с условием теоремы. Существование минимума следует из того, что кусочно-линейная функция $g_1(k) = \varphi^+(k) - \varphi^-(k)$ при $k \to +\infty$ имеет угловой коэффициент, равный

$$\frac{\max\limits_{i=1,\dots,n} x_i - \min\limits_{i=1,\dots,n} x_i}{2} > 0,$$

соответственно при $k \to -\infty$ угловой коэффициент равен

$$-\frac{\max\limits_{i=1,\ldots,n} x_i - \min\limits_{i=1,\ldots,n} x_i}{2} < 0.$$

Следовательно, выпуклая вниз функция $g_1(k)$ обязана достигать своего минимального значения (см. рис. 1).

Замечание 3. Пусть $\varphi^+(k)$ и $\varphi^-(k)$ — выпуклые соответственно вниз и вверх кусочно-линейные функции, имеющие конечное число звеньев, и такие, что любая касательная к графику одной из них разделяет графики этих функций. Тогда в плоскости $\{k, b\}$ определено конечное множество прямых $\{kx_i - y_i : i = 1, \ldots, n\}$, каждая из которых содержит одно звено $\varphi^+(k)$ или $\varphi^-(k)$ и таких, что

$$\begin{aligned} \varphi^+(k) &= \frac{\max\limits_{i=1,\ldots,n} (kx_i - y_i)}{2}, \\ \varphi^-(k) &= \frac{\min\limits_{i=1,\ldots,n} (kx_i - y_i)}{2}. \end{aligned}$$

Обозначим через $M(\varphi^+, \varphi^-) = \left\{(x, y) : \varphi^-(k) \le \frac{kx-y}{2} \le \varphi^+(k), \forall k \in R\right\}$ соответствующее множество точек на плоскости $\{x, y\}$.

Теорема 2. *Пусть $\varphi^+(k)$ и $\varphi^-(k)$ — выпуклые соответственно вниз и вверх кусочно-линейные функции такие, что любая касательная к графику одной из них разделяет графики этих функций. Тогда множество $M(\varphi^+, \varphi^-)$ выпуклое, точки*

$$\Omega_{(\varphi^+, \varphi^-)} = \{(x_i, y_i) : i = 1, \ldots, n\}$$

крайние точки множества $M(\varphi^+, \varphi^-)$.

ДОКАЗАТЕЛЬСТВО. Согласно теореме Каратеодори [2] выпуклая оболочка множества Ω совпадает с объединением всех треугольников с вершинами из Ω. Пусть точка $A_0 = (x_0, y_0)$ принадлежит выпуклой комбинации трех точек $A_i = (x_i, y_i)$, $A_j = (x_j, y_j)$, $A_s = (x_s, y_s)$

$$A_0 = \lambda_1 A_i + \lambda_2 A_j + \lambda_3 A_s,$$

где $\lambda_1 + \lambda_2 + \lambda_3 = 1$, $\lambda_1, \lambda_2, \lambda_3 \ge 0$. Имеем неравенства:

$$\begin{aligned} \varphi^-(k) &\le \frac{kx_i - y_i}{2} \le \varphi^+(k), \\ \varphi^-(k) &\le \frac{kx_j - y_j}{2} \le \varphi^+(k), \\ \varphi^-(k) &\le \frac{kx_s - y_s}{2} \le \varphi^+(k), \end{aligned}$$

где $k \in R$. Умножая эти неравенства на $\lambda_1, \lambda_2, \lambda_3$ и складывая, получим

$$\varphi^-(k) \le \frac{kx_0 - y_0}{2} \le \varphi^+(k), \forall k \in R.$$

Следовательно,

$$A_0 \in M\left(\varphi^+, \varphi^-\right).$$

Проверим, что точки множества $\Omega_{(\varphi^+,\varphi^-)} = \{(x_i, y_i) : i = 1, \ldots, n\}$ крайние для множества $M(\varphi^+, \varphi^-)$. Пусть точка $A_s = (x_s, y_s) \in \Omega_{(\varphi^+,\varphi^-)}$ соответствует прямой, содержащей звено ломаной $\varphi^+(k)$, т. е. на некотором отрезке $k \in [k_0, k_1]$ выполняется равенство

$$\varphi^+(k) = x_s k - y_s.$$

Предположим, что точка A_s — середина отрезка $[B, C] \subset M(\varphi^+, \varphi^-)$, где $B = (x^*, y^*)$, $C = (x^{**}, y^{**})$. Тогда при $k \in [k_0, k_1]$

$$\begin{aligned} \frac{1}{2}\left[(x^* k - y^*) + (x^{**} k - y^{**})\right] &= \varphi^+(k), \\ (x^* k - y^*) &\le \varphi^+(k), \\ (x^{**} k - y^{**}) &\le \varphi^+(k). \end{aligned}$$

Следовательно $A_s = B = C$, что и требовалось доказать.

Определение 4. Пусть $A_0 = (x_0, y_0)$ — точка на плоскости переменных $\{x, y\}$, $\alpha \ge 0$. Рассмотрим множество прямых $y = kx + b$, для которых $y_0 - \alpha \le kx_0 + b \le y_0 + \alpha$, и обозначим через

$$L_\alpha(A_0) = \{(k, b) : y_0 - \alpha \le kx_0 + b \le y_0 + \alpha\}$$

соответствующее множество коэффициентов прямых. Множество $L_\alpha(x_0, y_0)$ замкнутое и выпуклое — полоса (в плоскости переменных $\{k, b\}$), заключенная между двумя параллельными прямыми.

Теорема 3. *Пусть*

$$\alpha(i, j, t) = \inf\left\{\alpha : L_\alpha(x_i, y_i) \cap L_\alpha(x_j, y_j) \cap L_\alpha(x_t, y_t) \ne \varnothing\right\},$$

тогда справедливо равенство

$$\min_{k,b} \beta(k, b) = \max_{\{i,j,t\}} \left\{\alpha(i, j, t)\right\}.$$

ДОКАЗАТЕЛЬСТВО. Согласно теореме Хелли [4], если для любых трех множеств семейства Σ выпуклых, замкнутых, ограниченных подмножеств плоскости их пересечение непусто, то

$$\cap\{A : A \in \Sigma\} \ne \varnothing.$$

Пусть $\alpha_0 = \max\limits_{\{i,j,t\}} \{\alpha(i, j, t)\}$, тогда любые три множества $L_{\alpha_0}(x_i, y_i) \cap L_{\alpha_0}(x_j, y_j) \cap$ $\cap L_{\alpha_0}(x_t, y_t) \ne \varnothing$, следовательно, имеется общая точка (k, b) всех множеств $L_{\alpha_0}(x_i, y_i)$, т. е.

$$-\alpha_0 \le kx_i + b - y_i \le \alpha_0, \quad i = 1, \ldots, n.$$

Следовательно, все точки $A_i = (x_i, y_i)$ находятся внутри полосы, симметричной относительно прямой $kx + b$ ширины, равной $2\alpha_0$ (вдоль оси OY). Следовательно, $\alpha_0 \ge$ $\ge \min\limits_{k,b} \beta(k, b)$. Обратное неравенство следует непосредственно из определения $\min\limits_{k,b} \beta(k, b)$.

Теорема 4. *Справедлива формула*

$$\alpha(i,j,t) = 4 \cdot \frac{S_{ijt}}{l_{ij} + l_{jt} + l_{ti}},$$

где S_{ijt} *— площадь треугольника* $A_i(x_i, y_i)$, $A_j(x_j, y_j)$, $A_t(x_t, y_t)$, $l_{ij} = |x_i - x_j|$, $l_{jt} = |x_j - x_t|$, $l_{ti} = |x_t - x_i|$.

ДОКАЗАТЕЛЬСТВО. Равенство следует непосредственно из рис. 3.

Замечание 4. Задача нахождение всех $\alpha(i,j,t)$ имеет сложность порядка n^3, так как число сочетаний $C_n^3 = \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}$. Ниже будут указаны гораздо более эффективные алгоритмы.

§ 1.1. *Сравнение нечеткой и стандартной линейной регрессии*

Изучим сначала в общем виде, как связаны между собой минимальные p-отклонения $\alpha_p(X, Y)$.

Пусть V — конечномерное векторное пространство, $\|\cdot\|_1$, $\|\cdot\|_2$ — две нормы на V. Тогда существуют константы $0 < C_1 \le C_2$ такие, что

$$C_1 \le \frac{\|X\|_2}{\|X\|_1} \le C_2, \quad \forall X \in V.$$

Пусть $L \subset V$ — векторное подпространство, $V' = V/L$ — фактор-пространство. Тогда на V' — определены индуцированные нормы

$$\|z + L\|_1' = \min_{u \in L} \|z + u\|_1,$$
$$\|z + L\|_2' = \min_{u \in L} \|z + u\|_2,$$

где $z + L \in V/L$ — класс смежности элемента $z \in V$.

Лемма 1. *Существует константы* C_1', C_2' *такие, что* $C_1 \le C_1' \le C_2' \le C_2$ *и*

$$C_1' \le \frac{\|z + L\|_2'}{\|z + L\|_1'} \le C_2', \quad \forall \{z + L\} \in V/L,$$

константы C_1', C_2', *вообще говоря, зависят от выбора* L.

ДОКАЗАТЕЛЬСТВО. В силу конечномерности L, существует $u \in L$ такой, что

$$\|z + L\|_2' = \|z + u\|_2,$$
$$\|z + L\|_1' \le \|z + u\|_1 \le \frac{\|z + u\|_2}{C_1} = \frac{\|z + L\|_2'}{C_1}.$$

Следовательно,

$$C_1 \le \frac{\|z + L\|_2}{\|z + L\|_1}.$$

Следовательно,

$$C_1' = \inf_{z \in V} \frac{\|z + L\|_2}{\|z + L\|_1} \ge C_1.$$

Аналогично устанавливается существование $C_2' \ge C_2$.

Замечание 5. Пусть $V = R^n$ — n-мерное арифметическое пространство, введем обозначения для норм:

$$\|u\|_2 = \Big(\sum_{i=1}^{n} |u_i|^2\Big)^{\frac{1}{2}} \quad \text{— Евклидова норма,}$$

$$\|u\|_\infty = \max_i |u_i| \quad \text{— Чебышевская норма,}$$

$$\|u\|_1 = \sum_{i=1}^{n} |u_i| \quad \text{— Хемингова норма.}$$

Справедливы неравенства

$$\begin{cases} \|u\|_2 & \le \sqrt{n}\|u\|_\infty \\ \|u\|_\infty & \le \|u\|_2 \end{cases}, \quad \begin{cases} \|u\|_1 & \le n\|u\|_\infty \\ \|u\|_\infty & \le \|u\|_1 \end{cases}, \quad \begin{cases} \|u\|_1 & \le \sqrt{n}\|u\|_2 \\ \|u\|_2 & \le \|u\|_1 \end{cases}.$$

Следствие 1. *Справедливо неравенство*

$$1 \le C_1' \le \frac{\alpha_2(X,Y)}{\alpha_\infty(X,Y)} \le C_2' \le \sqrt{n}, \tag{4}$$

константы C_1', C_2', вообще говоря, зависят от L, т. е. от X.

ДОКАЗАТЕЛЬСТВО. Пусть $X = \{x_1, x_2, \ldots, x_n\}$, $E = \{1, \ldots, 1\}$, $Y = \{y_1, \ldots, y_n\} \in \in R^n$, обозначим через $L \subset V = R^n$ векторное подпространство вида

$$L = \{kX + bE : k, b \in R\},$$

тогда

$$\min_{k,b} \max_{i=1,\ldots,n} |kx_i + b - y_i| = \|Y + L\|_\infty',$$

$$\min_{k,b} \Big(\sum_{i=1,\ldots,n} |kx_i + b - y_i|^2\Big)^{\frac{1}{2}} = \|Y + L\|_2'.$$

Искомое неравенство следует из леммы 1.

Для вывода более точных оценок констант C_1', C_2' выведем одну формулу минимального квадратичного отклонения $\alpha_2(X,Y)$.

Обозначим через $X_s = \{x_1, \ldots, \widehat{x}_s, \ldots, x_n\}$, $Y_s = \{y_1, \ldots, \widehat{y}_s, \ldots, y_n\}$, где символ $\widehat{\ }$ означает, что соответствующий элемент удален из списка. Справедливы следующие утверждения.

Лемма 2.

$$M(X) = \frac{n-1}{n} M(X_s) + \frac{1}{n} x_s.$$

ДОКАЗАТЕЛЬСТВО.

$$M(X) = \frac{\sum\limits_{j=1}^{n} x_j}{n} = \frac{\sum\limits_{j=1}^{n-1} x_j}{n} + \frac{x_s}{n} = \frac{n-1}{n} M(X_s) + \frac{1}{n} x_s.$$

Лемма 3.

$$\operatorname{cov}(X,Y) = \frac{n-1}{n}\operatorname{cov}(X_s,Y_s) + \frac{1}{n-1}\left(x_s - M(X)\right)\left(y_s - M(Y)\right) = \\ = \frac{n-1}{n}\operatorname{cov}(X_s,Y_s) + \frac{n-1}{n^2}\left(x_s - M(X_s)\right)\left(y_s - M(Y_s)\right).$$

ДОКАЗАТЕЛЬСТВО.

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - M(X)\right)\left(y_i - M(Y)\right)}{n} = \frac{\sum_{i\neq s}\left(x_i - M(X)\right)\left(y_i - M(Y)\right)}{n} + \\ + \frac{\left(x_s - M(X)\right)\left(y_s - M(Y)\right)}{n} = \frac{\sum_{i\neq s}\left(x_i - \frac{n-1}{n}M(X_s) - \frac{x_s}{n}\right)\left(y_i - \frac{n-1}{n}M(Y_s) - \frac{y_s}{n}\right)}{n} + \\ + \frac{\left(x_s - M(X)\right)\left(y_s - M(Y)\right)}{n} = \\ = \frac{\sum_{i\neq s}\left(\left(x_i - M(X_s)\right) + \left(\frac{1}{n}M(X_s) - \frac{x_s}{n}\right)\right)\left(\left(y_i - M(Y_s)\right) + \left(\frac{1}{n}M(Y_s) - \frac{y_s}{n}\right)\right)}{n} + \\ + \frac{\left(x_s - M(X)\right)\left(y_s - M(Y)\right)}{n} = \frac{\sum_{i\neq s}\left(x_i - M(X_s)\right)\left(y_i - M(Y_s)\right)}{n} + \\ + \frac{\sum_{i\neq s}\left(x_i - M(X_s)\right)\left(M(Y_s) - y_s\right)}{n^2} + \frac{\sum_{i\neq s}\left(y_i - M(Y_s)\right)\left(M(X_s) - x_s\right)}{n^2} + \\ + \frac{\sum_{i\neq s}\left(M(X_s) - x_s\right)\left(M(Y_s) - y_s\right)}{n^3} + \frac{\left(x_s - M(X)\right)\left(y_s - M(Y)\right)}{n} = \\ = \frac{n-1}{n}\operatorname{cov}(X_s,Y_s) + \frac{n-1}{n^3}\left(x_s - M(X_s)\right)\left(y_s - M(Y_s)\right) + \frac{1}{n}\left(x_s - M(X)\right)\left(y_s - M(Y)\right) = \\ = \frac{n-1}{n}\operatorname{cov}(X_s,Y_s) + \left(\frac{1}{n(n-1)} + \frac{1}{n}\right)\left(x_s - M(X)\right)\left(y_s - M(Y)\right) = \\ = \frac{n-1}{n}\operatorname{cov}(X_s,Y_s) + \left(\frac{n-1}{n^3} + \frac{(n-1)^2}{n^3}\right)\left(x_s - M(X_s)\right)\left(y_s - M(Y_s)\right).$$

Лемма 4.

$$D(X) = \frac{n-1}{n}D(X_s) + \frac{1}{n-1}\left(x_s - M(X)\right)^2 = \frac{n-1}{n}D(X_s) + \frac{n-1}{n^2}\left(x_s - M(X_s)\right)^2.$$

ДОКАЗАТЕЛЬСТВО.

$$D(X) = \operatorname{cov}(X,X) = \frac{n-1}{n}\operatorname{cov}(X_s,X_s) + \frac{1}{n-1}\left(x_s - M(X)\right)\left(x_s - M(X)\right) = \\ = \frac{n-1}{n}D(X_s) + \frac{1}{n-1}\left(x_s - M(X)\right)^2.$$

$$D(X) = \operatorname{cov}(X,X) = \frac{n-1}{n}\operatorname{cov}(X_s,X_s) + \frac{n-1}{n^2}\left(x_s - M(X_s)\right)\left(x_s - M(X_s)\right) = \\ = \frac{n-1}{n}D(X_s) + \frac{n-1}{n^2}\left(x_s - M(X_s)\right)^2.$$

Теорема 5. *Справедлива формула для минимального квадратичного отклонения* $\alpha_2(X,Y)$*:*

$$\alpha_2^2(X,Y) = \frac{n^3\left(D(x)D(y) - \operatorname{cov}^2(x,y)\right)}{n^2 D(x)} = 4\cdot\frac{\frac{1}{3!}\sum_{i,j,k}S_{ijk}^2}{n^2D(x)} = 4\cdot\frac{\frac{1}{3!}\sum_{i,j,k}S_{ijk}^2}{\frac{1}{2!}\sum_{i,j}l_{ij}^2},$$

где S_{ijk} *— площадь треугольника* $A_i(x_i,y_i)$, $A_j(x_j,y_j)$, $A_k(x_k,y_k)$, $l_{ij} = |x_i - x_j|$.

ДОКАЗАТЕЛЬСТВО. Справедливость первого равенства непосредственно следует из формулы $\alpha_2^2 = n^3 D(y)\left(1 - r_{xy}^2\right)$ [1]. Покажем, что $\frac{1}{2!}\sum_{i,j} l_{ij}^2 = n^2 D(x)$.

$$\frac{1}{2!}\sum_{i,j} l_{ij}^2 = \frac{1}{2!}\sum_{i,j}\left((x_i - \bar{x}) - (x_j - \bar{x})\right)^2 =$$
$$= \frac{1}{2!}\sum_{i,j}(x_i - \bar{x})^2 - \sum_{i,j}(x_i - \bar{x})(x_j - \bar{x}) + \frac{1}{2!}\sum_{i,j}(x_j - \bar{x})^2 =$$
$$= \frac{n^2}{2}D(x) + \frac{n^2}{2}D(x) = n^2 D(x).$$

Осталось доказать, что $\frac{4}{3!}\sum_{i,j,k} S_{ijk}^2 = n^3\left(D(x)D(y) - \operatorname{cov}^2(x,y)\right)$. Доказательство проведем индукцией по числу точек.

База индукции. Рассмотрим три точки $A_1(x_1, y_1)$, $A_2(x_2, y_2)$, $A_3(x_3, y_3)$. Составим уравнение парной регрессии $y = kx + b$ и найдем соответствующие прогнозные значения $A_1'(x_1, y_1')$, $A_2'(x_2, y_2')$, $A_3'(x_3, y_3')$, где $y_i' = kx_i + b$ (рис. 2).

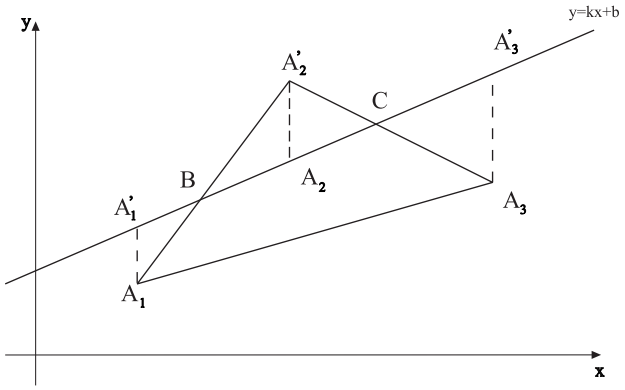

Рис. 2. База индукции

$\Delta A_1 A_1' B \sim \Delta A_2 A_2' B$, следовательно,

$$S_{A_2A_2'B} = \left(\frac{A_2A_2'}{A_1A_1'}\right)^2 S_{A_1A_1'B}, \quad S_{A_1A_1'B} = \frac{1}{2}\left(\frac{A_1A_1'}{A_1A_1' + A_2A_2'}\right)(x_2 - x_1)\cdot A_1A_1'.$$

$\Delta A_3 A_3' C \sim \Delta A_2 A_2' C$, следовательно,

$$S_{A_2A_2'C} = \left(\frac{A_2A_2'}{A_3A_3'}\right)^2 S_{A_3A_3'C}, \quad S_{A_3A_3'C} = \frac{1}{2}\left(\frac{A_3A_3'}{A_2A_2' + A_3A_3'}\right)(x_3 - x_2)\cdot A_3A_3'.$$

$A_1A_1'A_3'A_3$ — трапеция, и, значит, ее площадь выражается равенством

$$S_{A_1A_1'A_3'A_3} = \frac{1}{2}\left(A_1A_1' + A_3A_3'\right)(x_3 - x_1).$$

Выразим площадь треугольника $A_1A_2A_3$:

$$S_{A_1A_2A_3} = S_{A_1BCA_3} + S_{A_2A_2'B} + S_{A_2A_2'C} = S_{A_1BCA_3} + \left(\frac{A_2A_2'}{A_1A_1'}\right)^2 S_{A_1A_1'B} +$$
$$+ \left(\frac{A_2A_2'}{A_3A_3'}\right)^2 S_{A_3A_3'C} = S_{A_1A_1'A_3'A_3} + \left(\frac{A_2A_2'}{A_1A_1'} - 1\right)^2 S_{A_1A_1'B} + \left(\frac{A_2A_2'}{A_3A_3'} - 1\right)^2 S_{A_3A_3'C}.$$

Подставив в последнее равенство полученные выражения площадей соответствующих фигур, получаем, что

$$2S_{A_1A_2A_3} = \left(A_1A_1' + A_3A_3'\right)(x_3 - x_1) + \left(A_2A_2' - A_1A_1'\right)(x_2 - x_1) + \\ + \left(A_2A_2' - A_3A_3'\right)(x_3 - x_2) = A_1A_1' \cdot (x_3 - x_2) + A_2A_2' \cdot (x_3 - x_1) + A_3A_3' \cdot (x_1 - x_2) = \\ = (y_1' - y_1) \cdot (x_3 - x_2) + (y_2 - y_2') \cdot (x_3 - x_1) + (y_3' - y_3) \cdot (x_1 - x_2).$$

Возведя равенство в квадрат и выполнив некоторые тождественные преобразования, имеем

$$4S_{A_1A_2A_3}^2 = \sum_i (y_i - y_i')^2 \frac{1}{2} \sum_{i,j} (x_i - x_j)^2 - \big[((y_1' - y_1)(x_3 - x_1) - (y_2 - y_2)'(x_3 - x_2))^2 + \\ + ((y_1' - y_1)(x_2 - x_1) - (y_3' - y_3)(x_3 - x_2))^2 + ((y_1' - y_1)(x_3 - x_1) - (y_2 - y_2')(x_3 - x_2))^2\big].$$

Покажем, что второе слагаемое равно нулю. Это означает, что каждое из входящих в него выражений равно нулю. Можно показать, что вектор остатков $(y_1 - y_1', y_2 - y_2', y_3 - y_3')$ перпендикулярен вектору независимой переменной (x_1, x_2, x_3), т. е. $x_1(y_1 - y_1') + x_2(y_2 - \\ - y_2') + x_3(y_3 - y_3') = 0$. Проведя тождественные преобразования, получим

$$x_1(y_1 - y_1') + x_2(y_2 - y_2') + x_3(y_3 - y_3') + \\ + x_3(y_1 - y_1') - x_3(y_1 - y_1') + x_2(y_2 - y_2') - x_2(y_2 - y_2') = 0, \\ (y_1 - y_1')(x_1 - x_3) - (y_2 - y_2')(x_3 - x_2) + x_3(y_1 + y_2 + y_3 - y_1' - y_2' - y_3') = 0, \\ (y_1' - y_1)(x_3 - x_1) - (y_2 - y_2')(x_3 - x_2) = 0.$$

Аналогично показывается равенство нулю двух оставшихся выражений. Следовательно,

$$\sum_i (y_i - y_i')^2 = \alpha_2 = 4 \frac{\frac{1}{3!} S_{i,j,k}}{\frac{1}{2} \sum_{i,j} (x_i - x_j)^2}.$$

Индукционный переход. Предположим, что равенство верно для $n - 1$ точки. Докажем его верность для n точек. Пусть $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$ — соответственно X-координаты и Y-координаты множества $\Omega = \{A_i\,(x_i, y_i) : i = 1, \ldots, n\}$ точек, обозначим через $\Omega_s = \{A_1, \ldots, \widehat{A_s}, \ldots, A_n\}$, где $\widehat{\ }$ означает, что соответствующий элемент удален из списка. Представим множества Ω в виде объединения n множеств Ω_s, содержащих по $n - 1$ элементу. Для каждого из этих множеств выполнено предположение индукции, следовательно,

$$\frac{4}{3!} \sum_{i,j,k \in \Omega} S_{ijk}^2 = \frac{4}{3!(n-3)} \sum_s \sum_{i,j,k \in \Omega_s} S_{ijk}^2 = \frac{(n-1)^3}{n-3} \sum_s \left(D(X_s)D(Y_s) - \operatorname{cov}^2(X_s, Y_s)\right).$$

Воспользовавшись леммами 3 и 4, имеем

$$D(X_s)D(Y_s) = \frac{n^2}{(n-1)^2} D(X)D(Y) - \frac{n^2}{(n-1)^3} D(X)\,(y_s - M(Y))^2 - \\ - \frac{n^2}{(n-1)^3}\,(x_s - M(X))^2\,D(Y) + \frac{n^2}{(n-1)^4}\,(x_s - M(X))^2\,(y_s - M(Y))^2,$$

$$\operatorname{cov}^2(X_s,Y_s)=\frac{n^2}{(n-1)^2}\operatorname{cov}^2(X,Y)+\frac{n^2}{(n-1)^4}\left(x_s-M(X)\right)^2\left(y_s-M(Y)\right)^2-$$
$$-\frac{2n^2}{(n-1)^3}\operatorname{cov}(X,Y)\left(x_s-M(X)\right)\left(y_s-M(Y)\right).$$

Подставим полученные выражение в исходную формулу и, проведя суммирование, получим

$$\frac{(n-1)^3}{n-3}\sum_s\left(D(X_s)D(Y_s)-\operatorname{cov}^2(X_s,Y_s)\right)=$$
$$=\frac{(n-1)^3}{n-3}\left(\frac{n^3}{(n-1)^2}-\frac{2n^3}{(n-1)^3}\right)\cdot\left(D(X)D(Y)-\operatorname{cov}^2(X,Y)\right)=$$
$$=n^3\left(D(X)D(Y)-\operatorname{cov}^2(X,Y)\right).$$

Используя полученное равенство, можно уточнить неравенство (4).

Теорема 6. *Справедливо неравенство, связывающее* α_2 *и* α_∞*:*

$$\frac{\alpha_2}{\alpha_\infty}\le\sqrt{\frac{3n}{8}}.$$

ДОКАЗАТЕЛЬСТВО. Из рис. 3 легко следует неравенство для произвольных i, j, k:

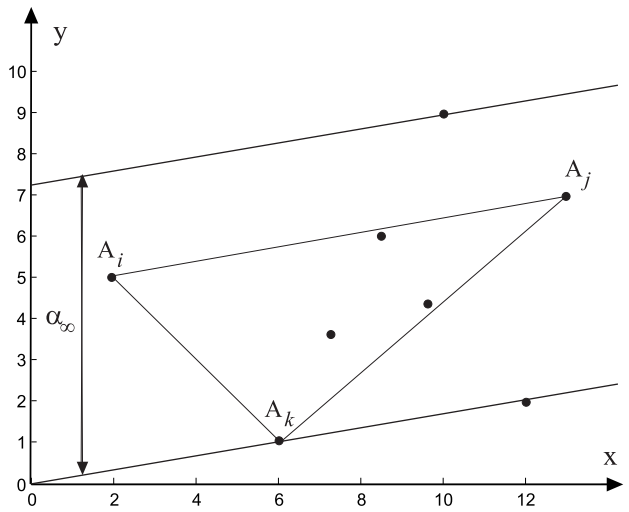

Рис. 3. Оценка для площади треугольника

$$S_{ijk}\le\frac{1}{2}\cdot\alpha_\infty\cdot\frac{1}{2}\big(|x_i-x_j|+|x_i-x_k|+|x_j-x_k|\big),$$

$$S_{ijk}^2\le\frac{1}{16}\alpha_\infty^2\cdot\big(|x_i-x_j|+|x_i-x_k|+|x_j-x_k|\big)^2.$$

Согласно неравенству Коши–Буняковского $\left(\sum z_i\right)^2\le n\sum z_i^2$, имеем

$$S_{ijk}^2\le\frac{3}{16}\alpha_\infty^2\cdot\left(|x_i-x_j|^2+|x_i-x_k|^2+|x_j-x_k|^2\right).$$

Суммируя, получим

$$\sum_{i,j,k} S_{ijk}^2 \le \frac{9}{16}\alpha_\infty^2 \cdot n \cdot \sum_{i,j} |x_i - x_j|^2,$$

$$4 \cdot \frac{1}{3!} \frac{\sum_{i,j,k} S_{ijk}^2}{\sum_{i,j} |x_i - x_j|^2} \le \frac{3n}{8}\alpha_\infty^2.$$

Откуда следует искомое неравенство

$$\frac{\alpha_2}{\alpha_\infty} \le \sqrt{\frac{3n}{8}}.$$

§ 1.2. *Экспериментальное моделирование нечеткой и стандартной линейной регрессии*

В общем случае установить точные оценки, связывающие α_2 и α_∞, затруднительно, но, используя систему MatLab, можно легко получить качественные свойства этих оценок.

Воспользуемся методом Монте-Карло для нахождения экспериментальных приближенных оценок отношения $C = \frac{\alpha_2}{\alpha_\infty}$ в зависимости от числа точек (дальнейшие вычисления выполнены в системе MatLab). На рис. 4 представлена зависимость отношения C от числа точек n.

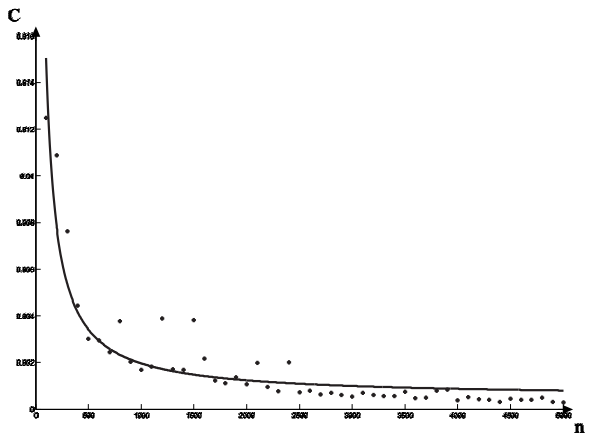

Рис. 4. Зависимость отношения α_2/α_∞ от числа точек

Аппроксимирующая функция имеет вид

$$C = 0{,}0005085 + \frac{1{,}455}{n}\,.$$

Полученная функция имеет статистически значимые коэффициенты и на 86,6 % объясняет вариацию отношения C.

§ 1.3. *Алгоритмы нахождения* $\min\limits_{k,b} \beta(k,b)$

§ 1.3.1. *Симплекс метод*

Задачу (3) нахождения $\min\limits_{k,b} \beta(k,b) = \min\limits_{k} g_1(k)$ можно свести к следующей задаче линейного программирования:

$$\begin{cases} \min\limits_{u,v,k} (u - v), \\ u \geq kx_i - y_i, \quad i = 1, \ldots, n, \\ v \leq kx_j - y_j, \quad j = 1, \ldots, n, \end{cases}$$

где u — верхняя огибающая, v — нижняя огибающая. Данную задачу можно решить, применяя симплекс метод (рис. 5).

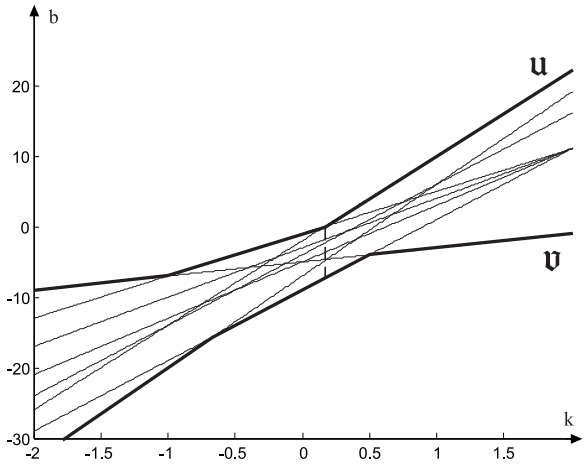

Рис. 5. Минимальное значение разности двух огибающих

Замечание 6. Сложность данного алгоритма порядка n^2.

§ 1.3.2. *Метод Грэхема*

Исходя из геометрической интерпретации данной задачи, нам необходимо определить полосу, содержащую данное множество, имеющую минимальную вертикальную («вдоль оси OY») ширину. Для этого нам нужно найти выпуклую оболочку множества. Это можно сделать, используя метод обхода Грэхема [3]. Суть его состоит в следующем.

(1) Находится внутренняя точка множества.
(2) Используя найденную точку как начало полярной системы координат, упорядочим точки множества в соответствии с полярным углом.
(3) Просмотр начинается с точки p_1 с наименьшей абсциссой, заведомо являющейся вершиной выпуклой оболочки $i = 1$.
(4) Если тройка последовательных точек p_i, p_{i+1}, p_{i+2} образует внутренний угол, меньший π, то $i = i + 1$ и переходим к просмотру следующей тройки точек; в противном случае, точка p_{i+1} удаляется и $i = i - 1$.
(5) Просмотр заканчивается тогда, когда, проверив все точки, мы снова вернулись в точку p_1 (рис. 6).

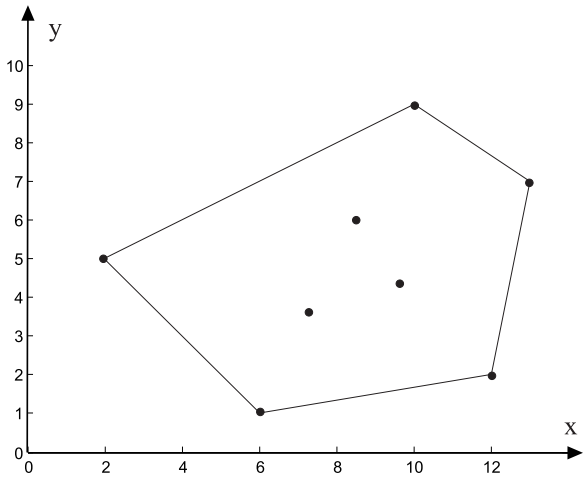

Рис. 6. Выпуклая оболочка множества

В результате будет получена упорядоченная последовательность вершин выпуклой оболочки $A_1, A_2, \ldots, A_s$. Затем ищутся всевозможные полосы, содержащие полученное выпуклое множество.

(1) Шаг $i = 1$.

(2) Выбирается одна из сторон полученной выпуклой оболочки A_iA_{i+1}, и определяется прямая, содержащая эту сторону $l_i:\ y = kx + b$.

(3) Для оставшихся вершин оболочки A_j находятся прямые $l_j:\ y = k(x - x_j) + y_j$ и проверяется условие, что полоса, ограниченная прямыми l_i и l_j, содержит все точки выпуклой оболочки.

(4) Операции (2)–(3) повторяются до тех пор, пока не будут рассмотрены все стороны выпуклой оболочки.

(5) Из найденных полос выбирается одна, имеющая наименьшую вертикальную ширину (рис. 7).

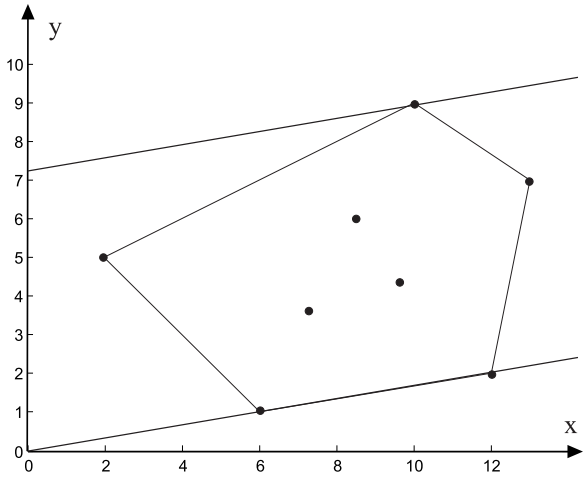

Рис. 7. Полоса минимальной ширины «вдоль оси OY»

Замечание 7. Рис. 5–7 получены по одному и тому же набору данных.

Замечание 8. Сложность данного алгоритма порядка $n \log n$ [3].

§ 1.3.3. *Алгоритм Quickhull в системе MatLab*

Данный алгоритм осуществляет алгоритм «быстрая оболочка» для нахождения выпуклой оболочки [5]. Этот алгоритм объединяет в себе двумерный Quickhull алгоритм с n-мерным алгоритмом «под-над» [3]. Главными преимуществами Quickhull является малая чувствительность к большому объему данных и погрешностям вычислений.

Список литературы

1. *Гмурман В. Е.* Теория вероятностей и математическая статистика. 4-е изд., доп.: Учеб. пособие для вузов. М.: Высш. шк., 1972.

2. *Лейхтвейс К.* Выпуклые множества. М.: Наука, 1985.

3. *Препарата Ф., Шеймос М.* Вычислительная геометрия: введение. М.: Мир, 1989.

4. *Хадвигер Г., Дебруннер Г.* Комбинаторная геометрия плоскости. М.: Наука, 1965.

5. *Barber C. B., Dobkin D. P., Huhdanpaa H. T.* The Quickhull Algorithm for Convex Hulls // ACM Transactions on Mathematical Software. 1996. Vol. 22. No. 4 (Dec. 1996). P. 469–483.

6. *Dug Hun Hong, Changha Hwang.* Support Vector Fuzzy Regression Machines // Fuzzy Sets Syst. 2003. Vol. 138. No. 2. P. 271–281.

7. *Kyung-Bin Song, Young-Sik Baek, Dug Hun Hong, Gilsoo Jang.* Short-Term Load Forecasting for the Holidays Using Fuzzy Linear Regression Method // IEEE Transactions on Power Systems. 2005. Vol. 20. No. 1. P. 96–101.

8. *Lu Jingli, Wang Ruili.* An Enhanced Fuzzy Linear Regression Model with More Flexible Spreads // Fuzzy Sets Syst. 2009. Vol. 160. No. 17. P. 2505–2523.

9. *Tanaka H., Uejima S., Asai K.* Linear Regression Analysis with Fuzzy Model // IEEE Transactions on Systems, Man and Cybernetics. 1982. Vol. 12. No. 6. P. 903–907.

10. *Shapiro A. F.* Fuzzy Regression and the Term Structure of Interest Rates Revisited // Proc. of the 14$^{\text{th}}$ Int. AFIR Colloquium. 2004. Vol. 1. P. 29–45.

Материал поступил в редколлегию 16.10.2008

Адреса авторов

ПОНОМАРЕВ Игорь Викторович
Алтайская государственная педагогическая академия
ул. Молодежная, 55, Барнаул, 656031, Россия
e-mail: igorpon@mail.ru

СЛАВСКИЙ Виктор Владимирович
Алтайская государственная педагогическая академия
ул. Молодежная, 55, Барнаул, 656031, Россия
e-mail: slavsky@uriit.ru