МЕРЫ ВАРИАБЕЛЬНОСТИ ОЦЕНОК ПАРАМЕТРОВ В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ ПО ИНТЕРВАЛЬНЫМ ДАННЫМ

С.П. ШАРЫЙ,
ФИЦ ИВТ и НГУ,
Новосибирск, Россия
shary@ict.nsc.ru

Аннотация. *Представлены две возможныые конструкции количественной меры вариабельности оценки параметров в задаче восстановления линейной зависимости по интервальным данным. Они показывают степень изменчивости и неоднозначности оценки, и необходимость их введения вызвана тем, что в условиях интервальных данных ответ к задаче, как правило, неединствен. Обсуждаются содержательные мотивации и смысл новых величин с точки зрения других известных характеристик данных. Применение предложенных мер вариабельности иллюстрируется на конкретных примерах.*

Ключевые слова: *задача восстановления зависимостей, интервальная неопределённость данных, метод максимума совместности, мера вариабельности.*

Цель нашей работы — представить две возможные конструкции количественной меры вариабельности для оценки параметров функциональной зависимости в интервальном анализе данных. Это сравнительно молодая ветвь современной науки об обработке данных, которая не опирается на аппарат теории вероятностей, но широко использует методы интервального анализа (см., к примеру, [1, 2, 3]).

Термином «вариабельность» мы называем степень изменчивости и неоднозначности оценки, и необходимость её введения диктуется тем обстоятельством, что в задачах обработки интервальных данных ответ, как правило, неединствен. Обычно мы получаем целое множество различных оценок, одинаково пригодных в качестве ответов к задаче и согласующихся с её данными. То, насколько мало или обширно это множество, как раз таки характеризуется термином «вариабельность».

Мы рассматриваем задачу восстановления линейной зависимости вида

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \qquad (1)$$

в которой $x_1, x_2, \ldots, x_n$ – независимые переменные (называемые также *входными* или *предикторными* переменными), y – зависимая переменная (называемая также *выходной* или *критериальной* переменной), а $\beta_1, \beta_2, \ldots, \beta_n$ – некоторые коэффициенты. Эти неизвестные коэффициенты должны быть определены на основе ряда измерений значений $x_1, x_2, \ldots, x_n$ и y. Результаты измерений неточны, и мы предполагаем, что они являются некоторыми интервалами, которые дают двусторонние границы точных значений измеренных величин. Будем считать, что в результате i-го измерения получаются такие интервалы $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{in}, \boldsymbol{y}_i$, что истинное значение x_1 лежит в $\boldsymbol{x}_{i1}$, истинное значение x_2 лежит в $\boldsymbol{x}_{i2}$ и т. д. вплоть до y, истинное значение которого лежит в $\boldsymbol{y}_i$. Иными словами, мы рассматриваем интервальные данные

$$\begin{array}{c} \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n}, \boldsymbol{y}_1, \\ \boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \ldots, \boldsymbol{x}_{2n}, \boldsymbol{y}_2, \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ \boldsymbol{x}_{m1}, \boldsymbol{x}_{m2}, \ldots, \boldsymbol{x}_{mn}, \boldsymbol{y}_m, \end{array} \qquad (2)$$

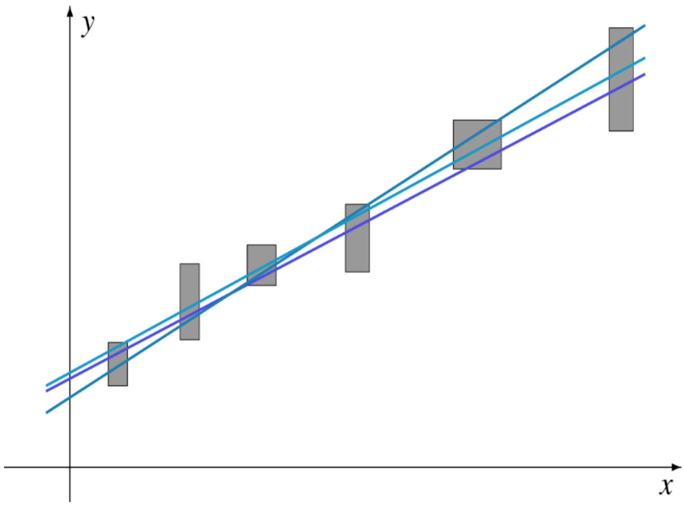
а также построенные из них матрицу $\boldsymbol{X} = (\boldsymbol{x}_{ij})$ и вектор $\boldsymbol{y} = (\boldsymbol{y}_i)$.

Нам необходимо найти или оценить коэффициенты β_j, $j = 1, 2, \ldots, n$, для которых линейная функция (1) «наилучшим образом» приближает данные (2) (см. рисунок).

На сегодняшний день разработано немало методов решения поставленной задачи, но мы будем работать с методом максимума совместности [4], который является, по-видимому, наиболее универсальным и хорошо разработанным. В этом методе оценка $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)$ для параметров линейной зависимости (1) берётся как аргумент максимума специального «распознающего функционала», некоторой функции $\mathrm{Tol} : \mathbb{R}^n \to \mathbb{R}$, с помощью которой определяется количественная мера совместности интервальной системы уравнений, построенной по данным задачи:

$$\mathrm{Tol}\,(\beta, \boldsymbol{X}, \boldsymbol{y}) \;=\; \min_{1\le i\le m} \left\{ \mathrm{rad}\,\boldsymbol{y}_i - \left| \mathrm{mid}\,\boldsymbol{y}_i - \sum\nolimits_{j=1}^{n} \boldsymbol{x}_{ij}\beta_j \right| \right\},$$

где $\mathrm{rad}\,\boldsymbol{y}_i$ и $\mathrm{rad}\,\boldsymbol{y}_i$ — радиус и середина интервала $\boldsymbol{y}_i$.



В качестве величины, характеризующей чувствительность и вариабельность оценки вектора параметров $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)$ линейной зависимости (1), которая получена методом максимума согласования по данным (3), мы предлагаем величины

$$\mathrm{IVE}(\boldsymbol{X}, \boldsymbol{y}) = \sqrt{n}\;\max_{\mathbb{R}^n}\,\mathrm{Tol}\cdot\left(\min_{X\in\boldsymbol{X}}\,\mathrm{cond}_2\,X\right)\cdot\frac{\left\|\arg\max_{\mathbb{R}^n}\mathrm{Tol}\right\|}{\|\hat{y}\|}$$

и

$$\mathrm{IDE}(\boldsymbol{X}, \boldsymbol{y}) = \sqrt{n}\;\max_{\mathbb{R}^n}\,\mathrm{Tol}\cdot\mathrm{cond}_2\mathcal{X}\cdot\frac{\left\|\arg\max_{\mathbb{R}^n}\mathrm{Tol}\right\|}{\|\hat{y}\|}.$$

В выписанных формулах переменные и величины имеют следующий смысл:

n – размерность вектора параметров восстанавливаемой функции;

$\|\cdot\|$ – евклидова норма векторов (2-норма), определяемая для векторов $y \in \mathbb{R}^n$ как

$$\| y \| = \left(\sum_{i=1}^{n} y_i^2\right)^{1/2};$$

$\operatorname{cond}_2 X$ — спектральное число обусловленности матрицы X, определяемое как

$$\operatorname{cond}_2 X = \sigma_{\max}(X) \big/ \sigma_{\min}(X),$$

т.е. как отношение максимального и минимального сингулярных чисел матрицы X; $\operatorname{cond}_2 X$ является распространением на прямоугольный случай известного в вычислительной линейной алгебре понятия числа обусловленности квадратной матрицы;

$\mathcal{X}$ – специальная точечная матрица, называемая *матрицей комбинаций концов* для интервальной матрицы $\boldsymbol{X}$, имеющая размеры $N \times n$, где $N \leq m \cdot 2^n$, и составленная из концов интервальных элементов вдоль каждой строки матрицы $\boldsymbol{X}$;

$\hat{\mathbf{y}}$ – это «наиболее представительная» точка интервального вектора данных $\boldsymbol{y}$, определяемая как

$$\hat{\mathbf{y}} = \frac{1}{2}\left(|\operatorname{mid} \boldsymbol{y} + \operatorname{rad} \boldsymbol{y}| + |\operatorname{mid} \boldsymbol{y} - \operatorname{rad} \boldsymbol{y}|\right),$$

где операции «mid» и «rad» применяются покомпонентно. Мы показываем, что в случае неотрицательного $\max \operatorname{Tol}$ величины IVE и IDE дают, фактически, оценки размеров множества решений задачи.

Литература

1. «Интервальный анализ и его приложения» — тематический веб-сайт http://www.nsc.ru/interval/
2. Moore R.E., Kearfott R.B., Cloud M.J. Introduction to interval analysis. – SIAM: Philadelphia, 2009.
3. Шарый С.П. Конечномерный интервальный анализ. – ИВТ СО РАН: Новосибирск, 2003-2021. – Электронная книга, см. http://www.nsc.ru/interval/?page=Library/InteBooks
4. Shary S.P Weak and strong compatibility in data fitting problems under interval uncertainty // Advances in Data Science and Adaptive Analysis. – 2020. – Vol. 12, No. 1. – 2050002.
5. Shary S.P. A variability measure for estimates of parameters in interval data fitting. 20 p. Deposited in arXiv as document arXiv:2003.05126