УДК 519.8

Н.М Оскорбин, А.В. Максимов, С.И. Жилин

Построение и анализ эмпирических зависимостей методом центра неопределенности

Рассматривается задача построения и анализа эмпирической зависимости $y = f(\alpha, x)$ скалярной переменой y, $(y \in R)$ и факторами $x \in R^n$, проводимых с использованием данных (y_j, x_j), $j = 1, \dots, N$, полученных в N наблюдениях. Наряду с наблюдениями имеется доступная для использования дополнительная (априорная) информация о виде зависимости $y = f(\alpha, x)$ и ее свойствах, условиях наблюдения, характере изменения переменных и др. Задача построения зависимости сводится к выбору на основе всей имеющейся непротиворечивой информации вида функции f и определению оценок вектора параметров α.

Задача анализа зависимости включает в себя выявление и устранение противоречий в исходных данных, проверку гипотез о виде искомой зависимости (линейная, квадратичная и др.), ее свойствах и оценку степени работоспособности найденной функции, т.е. пригодности эмпирической зависимости на практике. Представляют интерес также количественные или качественные суждения о ценности совокупности имеющейся информации и отдельных ее составляющих, о скорости ее старения, способах обновления исходных данных и т.д.

Перечисленные задачи анализа эмпирических зависимостей в настоящее время еще далеки от точных математических постановок. Эти постановки могут быть получены с использованием концепции информационных систем [1–3]. В работе [3] приведены способы выявления противоречий в совокупности информации, что позволяет выделять ситуации при построении эмпирических моделей, в которых форма зависимости или уровни погрешностей измерения переменных выбраны неверно. Очевидно, что прежде чем проводить обработку, необходимо устранить противоречивость наблюдений и исходных предположений. В данной работе рассматривается фрагмент общей задачи построения и анализа зависимостей, который включает определение оценок параметров искомой зависимости в рамках фиксированной ее структуры и интервалов достоверности найденных значений этих параметров.

Традиционно задачи построения и анализа эмпирических зависимостей решаются методами теории вероятностей и математической статистики, в частности, методами регрессионного анализа. В этом случае предполагается, что вид функции искомой зависимости задан верно, а невязки $\varepsilon = y - f(\alpha, x)$ имеют нормальное распределение. Эти предположения во многих прикладных задачах не выполняются [1, с. 140]. Кроме того, недостаточно полно используется априорная информация. Устранение противоречий в исходной информации не проводится. Выявлено, что в регрессионном анализе оценки пригодности найденной эмпирической зависимости искажаются условиями проведения эксперимента [1, с. 147]. Эти и другие обстоятельства ограничивают область применимости вероятностных методов в задачах обработки результатов наблюдений.

В данной работе используется нестатистический подход к построению эмпирических зависимостей, который предложен Л.В. Канторовичем [2, с. 701], развит в работах [3–10] и использует возможности линейного программирования для записи условий обработки данных с учетом всех имеющихся информационных соотношений между значениями наблюдаемых переменных. Вся совокупность способов обоснований и приемов построения и анализа эмпирических зависимостей на основе предлагаемого подхода названа методом центра неопределенности (МЦН) [3, с. 64–69].

В этом методе оценки $\hat{\alpha}_i$, $(i = 0, 1, \dots, n)$ вектора неизвестных коэффициентов $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_n)$ искомой линейной многофакторной зависимости

$$f(\alpha, x) = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n, \; x \in X \qquad (1)$$

отыскиваются по таблице экспериментальных данных, полученных в N наблюдениях

$$(y_i, x_{1j}, \dots, x_{nj}), \; j = 1, \dots, N, \qquad (2)$$

в каждом из которых значение искомой функции $f(\alpha, x_j)$ при фиксированных значениях вектора факторов $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})$ удовлетворяет неравенству

$$y_j - \varepsilon_j^- \le f(\alpha, x_j) \le y_j + \varepsilon_j^+, \; j = 1, \dots, N. \qquad (3)$$

Здесь $(y_j - \varepsilon_j^-)$, $(y_j + \varepsilon_j^+)$ — соответственно оценки нижнего и верхнего значений выходной переменной y в точке x_j, а ε_j^-, ε_j^+ — соответственно нижняя и верхняя оценки невязок уравнения (1), найденные по условиям j-го наблюдения за входными и выходной переменными.

Далее предполагается, что y — количественный показатель, а факторы $x_1, \ldots, x_n$ в отдельности могут быть количественными и (или) качественными (принимающими значение $1, 2, \ldots, m, 1 < m \leq n$). Они могут отражать комплексные, нелинейные и сочетанные эффекты влияния факторов моделируемого процесса. Обозначим через $A(N)$ множество значений вектора коэффициентов α, удовлетворяющих системе неравенств (3). Основным принципом обработки наблюдений в рассматриваемом методе является равнозначность всех элементов множества $A(N)$, т.е. разных значений коэффициентов α искомой зависимости (1). Термин для $A(N)$ — "множество неопределенности значений α" — подчеркивает этот принцип [3]. Отдельные точки множества $A(N)$, в частности, каким-либо образом задаваемый "центр неопределенности" [3], могут выступать в виде характеристик части или всех точек множества $A(N)$, удобных для анализа или практического использования искомой зависимости (1). Отметим, что указанный подход лежит в рамках естественнонаучной традиции обработки результатов наблюдений.

Информационная суть МЦН выясняется в сравнении его с методом наименьших квадратов (МНК) при поиске значений коэффициентов и регрессионным и корреляционным анализами при проверке гипотез (доверительных интервалов оценок коэффициентов α и значений переменной y в заданной точке x, наличии выбросов в исходных данных и др.). В МНК этапы оценки коэффициентов и этап анализа уравнения разделены и проводятся на разных математических моделях, а для проведения анализа эмпирических зависимостей необходимо достаточное число разнообразных наблюдений ($N > n$), выполнимость предположений о нормальности, независимости и др. [1, с. 140]. В случае невыполнения этих предположений МНК не содержит способов их контроля. Используемый в этом методе анализ остатков теоретически не обоснован и не гарантирует даже обнаружения выбросов.

В отличие от МНК в МЦН этапы построения модели и ее анализ не разделяются и выполняются с использованием системы неравенств (3) методами линейного программирования. При наличии противоречий в исходных данных решения задач оценки параметров и анализ не проводят, поскольку в этом случае система неравенств (3) становится несовместной ($A(N)$ — пустое множество), и предпринимаются усилия по устранению противоречий либо делаются выводы о несоответствии принятого вида функции информационным взаимосвязям моделируемого объекта. Это обстоятельство используется для проверки и других гипотез. Таким образом, **в вероятностном подходе используется информация о согласованности наблюдений (принцип максимального правдоподобия), а в МЦН — информация о их непротиворечивости.**

Сформулируем далее две задачи обработки данных МЦН: задачу оценивания значения функции $f(\alpha, x)$ при любом фиксированном $x \in X$ и задачу оценивания значения любого из коэффициентов $\alpha_i, i = 0, \ldots, n$. Первая задача сводится к нахождению двух функций $y^{\text{н}}(x)$ и $y^{\text{в}}(x)$ путем решения следующих двух задач линейного программирования:

$$\begin{aligned} y^{\text{н}}(x) &= \min_{\alpha \in A(N)} (\alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n); \\ y^{\text{в}}(x) &= \max_{\alpha \in A(N)} (\alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n). \end{aligned} \tag{4}$$

При решении второй задачи — задачи оценивания параметров зависимости (1) — находим $\alpha_i^{\text{н}}$ и $\alpha_i^{\text{в}}$ из условий

$$\left.\begin{aligned} \alpha_i^{\text{н}} &= \min_{\alpha \in A(N)} \alpha_i; \\ \alpha_i^{\text{в}} &= \max_{\alpha \in A(N)} \alpha_i, \end{aligned}\right. \quad i = 0, 1, \ldots, n. \tag{5}$$

Для искомых величин в сформулированных двух задачах имеем:

$$\begin{gathered} y^{\text{н}}(x) \leq f(\alpha, x) \leq y^{\text{в}}(x), \; x \in X; \\ \alpha_i^{\text{н}} \leq \alpha_i \leq \alpha_i^{\text{в}}, \; i = 0, 1, \ldots, n. \end{gathered} \tag{6}$$

Неравенства (6) являются простым следствием системы неравенств (3). Если при решении задачи (4) (или (5)), выяснится, что система неравенств (3) несовместна ($A(N)$ — пустое множество), то исходные предпосылки и (или) результаты наблюдений в совокупности противоречивы. Анализ противоречивых данных — особый раздел рассматриваемого метода.

Системе неравенств (3) могут удовлетворять также бесконечно удаленные точки. В этом случае исходной информации для построения искомой зависимости недостаточно и необходимо либо продолжить экспериментальное изучение объекта моделирования, либо привлечь дополнительные сведения априорного характера об объекте моделирования. В частном случае задачи поиска интервальной оценки коэффициента

α в зависимости $y(x) = 1 + \alpha x$ из выражений (5) имеем при $x_j > 0$, $j = 1, \ldots, N$ следующие конечные формулы [4, с. 143]:

$$\alpha^{\text{н}} = \max_{1 \le j \le N} \left[(y_j - \varepsilon_j^- - 1)/x_j \right]; \qquad \alpha^{\text{в}} = \min_{1 \le j \le N} \left[(y_j + \varepsilon_j^+ - 1)/x_j \right]. \tag{7}$$

Рассмотрим далее две особенности обработки данных в МЦН. Первая касается способов учета априорных сведений об объекте моделирования. Пусть, например, при построении зависимости (1) известно, что коэффициент $\alpha_1 <= \alpha_2$. Тогда указанное ограничение на выбор допустимых значений коэффициентов α непосредственно включается в систему неравенств (3) и множество неопределенности $A(N)$ может сократиться. Степень такого сокращения при включении любой строки с номером j связана с уровнем информационной нагрузки, соответствующей j -й "порции информации". К априорной информации, кроме вышеприведенных неравенств, могут быть отнесены требования ограниченности информационной нагрузки всех или части наблюдений.

Вторая особенность относится к способу представления множества неопределенности $A(N)$, точное описание которого является сложной задачей. Предложено использовать [3] простые аппроксимации этого множества, в частности, в виде гиперпрямоугольника

$$\hat{A}(N) = \left\{ \alpha \mid \alpha_i^{\text{н}} \le \alpha_i \le \alpha_i^{\text{в}},\ i = 0, 1, \ldots, n \right\}, \tag{8}$$

где $\alpha_i^{\text{н}}, \alpha_i^{\text{в}}, i = 0, 1, \ldots, n$ находим решением 2(n+1) задач линейного программирования (5). Существенно отметить, что аппроксимация $\hat{A}(N)$ включает исходное множество $A(N)$, и тем самым, при замене множества $A(N)$ на множество $\hat{A}(N)$ не вносится ложной информации. Наглядной характеристикой множества $\hat{A}(N)$ является его представление с использованием $2(n+1)$ чисел $\hat{\alpha}_i,\ \Delta\alpha_i,\ i = 0, 1, \ldots, n$ [3], где

$$\hat{\alpha}_i = \tfrac{1}{2}(\alpha_i^{\text{н}} + \alpha_i^{\text{в}}); \qquad \Delta\alpha_i = \tfrac{1}{2}(\alpha_i^{\text{в}} - \alpha_i^{\text{н}}). \tag{9}$$

Здесь точку $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_n)$ можно рассматривать в качестве точечных оценок искомых параметров зависимости (1) при симметричных в (3) невязках, т.е. при $\varepsilon_j^- = \varepsilon_j^+, j = 1, \ldots, N$.

Точечную оценку параметра α можно отыскивать также по точечной зависимости $f_t(\alpha, x)$, если она линейна по параметрам α, используя решение задач (4):

$$f_t(\alpha, x) = \frac{1}{2}(y^{\text{н}}(x) + y^{\text{в}}(x)).$$

Для несимметричных невязок понятие точечных оценок необходимо специально определить. Рассмотрим систему неравенств

$$y_j - k\varepsilon_j^- \le f(\alpha, x_j) \le y_j + k\varepsilon_j^+, \qquad k > 0,\ j = 1, \ldots, N, \tag{10}$$

которая при $k = 1$ совпадает с (3). Пусть k_0 – минимальное значение параметра k, при котором (10) имеет решение, а $A(N, k_0)$ — соответствующее множество неопределенности. Если $A(N, k_0)$ содержит только единственную точку α, то эту точку примем в качестве точечной оценки коэффициентов зависимости (1).

В иных случаях предлагается использовать метод получения точечных оценок при симметричных невязках, используя в задачах (4), (5) множества $A(N, k_0)$. В заключение отметим, что вопросы использования МЦН и его исследования рассмотрены в работах [5–11]. Сравнение МЦН с вероятностными методами проведено в [5,9]. Вопросы теоретического обоснования метода и его связи с интервальной математикой, методами выравнивания данных по П.Л. Чебышеву рассматривались в [10], применения для обработки выборок большого объема в [11]. Аспекты практического использования затронуты в работах [4–8, 10, 12].

Рассмотренный метод реализован в пакете программ MCN, который разработан и используется с 1987 г. в Алтайском государственном университете [11].

Литература

1. Люблинский Р.Н., Оскорбин Н.М. Методы декомпозиции при оптимальном управлении непрерывными производствами. Томск, 1979.

2. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журнал. 1962 Т. 3. N 5.

3. Оскорбин Н.М. Некоторые задачи обработки информации в управляемых системах // Синтез и проектирование многоуровневых иерархических систем. Барнаул, 1983.

4. Алгазин А.И., Шойхет Я.Н., Киселев В.И.,

Оскорбин Н.М. Оценка риска радиационного заражения населения Алтайского края // Семипалатинский полигон – Алтай. 1997. N 8.

5. Гузеев В.В., Суханов В.А., Белов В.М. Математическое обеспечение кинетического метода в аналитической химии // Математические методы и ЭВМ в аналитической химии: Программа конференции. М., 1986.

6. Белов В.М., Суханов В.А., Унгер Ф.Г. Метод центра неопределенности при расчете линейных градуировочных графиков и метрологических характеристик результатов химического анализа: Препринт / АН СССР. Сибирское отделение. Институт химии нефти. N 59. Томск, 1989.

7. Белов В.М., Суханов В.А., Гузеев В.В., Унгер Ф.Г. Оценивание параметров линейных физико-химических зависимостей прямоугольником метода центра неопределенности // Известия вузов. Физика. 1991. N 8.

8. Вощинин А.П., Сотиров Г.Р. Оптимизация в условиях неопределенности. Москва; София, 1989.

9. Вощинин А.П., Бочков А.Ф., Сотиров Г.Р. Метод анализа данных при интервальной нестатистической ошибке // Заводская лаборатория. Т. 56. 1990. N 7.

10. Белов В.М., Суханов В.А., Унгер Ф.Г. Теоретические и прикладные аспекты метода центра неопределенности. Новосибирск, 1995.

11. Ерохин Г.Н., Камышников А.И., Оскорбин Н.М. Обработка больших баз данных методами линейного программирования // Управление, математическое моделирование и оптимизация на базе ПЭВМ: Межвуз. сб. науч. раб. Барнаул, 1993.

12. Жилин С.И. Решение задачи трансформации векторных и растровых изображений с использованием методов линейного программирования // Проблемы предотвращения деградации земель Западной Сибири и осуществление государственного контроля за их использованием и охраной. М., 1997.