ИЗВЕСТИЯ РАН. СЕРИЯ ФИЗИЧЕСКАЯ, 2022, том 86, № 11, с. 1593–1597

УДК 519.6

РЕГРЕССИЯ ЭКСПЕРИМЕНТАЛЬНЫХ КРИВЫХ МЕТОДОМ ОРТОГОНАЛИЗОВАННЫХ ПОЛИНОМОВ

© 2022 г. О. И. Топор[1, *], А. А. Белов[1, 2], Л. В. Бородачев[1]

[1]*Федеральное государственное бюджетное образовательное учреждение высшего образования "Московский государственный университет имени М.В. Ломоносова", физический факультет, Москва, Россия*

[2]*Федеральное государственное автономное образовательное учреждение высшего образования "Российский университет дружбы народов", Москва, Россия*

******E-mail: topor.oi15@physics.msu.ru*

Поступила в редакцию 30.06.2022 г.
После доработки 15.07.2022 г.
Принята к публикации 22.07.2022 г.

Впервые проведено количественное сравнение двух способов регрессии: на основе ортогональных и неортогональных полиномов. В качестве теста построена модельная задача, имитирующая эксперименты по измерению скоростей химических реакций. Показаны преимущества метода ортогональных полиномов.

DOI: 10.31857/S036767652211031X

ВВЕДЕНИЕ

Задача регрессии экспериментальных кривых является важной частью интерпретации натурного эксперимента. В общем случае задача ставится следующим образом [1]. Пусть имеется набор экспериментальных точек: аргументов x_i и значений функции $u_i \pm \delta_i$, где δ_i – абсолютные погрешности измерений. Требуется приблизить зависимость $u(x)$ некоторой априорно выбранной кривой $\varphi(x)$, содержащей свободные параметры $a_j, 1 \le j \le n$.

Чаще всего в качестве $\varphi(x)$ используют обобщенный многочлен

$$\varphi(\vec{a}, x) = \sum_{j=1}^{n} a_j \varphi_j(x). \qquad (1)$$

Параметры a_j выбирают так, чтобы среднеквадратичное уклонение кривой (1) от экспериментальных точек было минимальным

$$F(\vec{a}) \equiv \sum_{i=1}^{I} \frac{(u_i - \varphi(x_i))^2}{\delta_i^2} \to \min. \qquad (2)$$

Если экспериментальные погрешности велики, то задача является некорректной: один и тот же набор данных может достоверно описываться различными аппроксимантами $\varphi(x)$. При этом функция $F(\vec{a})$ является многоэкстремальной. Так, экспериментальные точки, расположенные далеко от кривой $\varphi(x)$, слабо влияют на $F(\vec{a})$ и на выбор параметров a_j. Если одна из точек расположена близко к кривой $\varphi(x)$, а остальные точки сравнительно далеко, то $F(\vec{a})$ имеет локальный минимум. Вблизи другой точки возникнет другой локальный минимум и т.д.

В этом случае из физического смысла выбирают специальный вид $\varphi(x)$ с небольшим числом параметров a_j либо вводят регуляризацию с помощью стабилизатора Тихонова [2, 3].

В ряде случаев за счет выбора специальных переменных экспериментальную зависимость удается привести к полиномиальной, в частности – линейной. Например, зависимость скоростей K химических реакций от температуры T описывается законом Аррениуса [4]

$$K(T) = A\exp(-E/T), \qquad (3)$$

где A, E – подгоночные параметры. В переменных $1/T - \lg K$ эта зависимость превращается в прямую.

Традиционно для регрессии таких данных используют аппроксимацию полиномами. В качестве базиса выбирают степени $\varphi_j(x) = x^j$, $j = 0,1,2,\ldots$ Этот базис является неортогональным: скалярные произведения

$$\langle \varphi_j, \varphi_k \rangle = \sum_{i=1}^{I} \frac{\varphi_j(x_i)\varphi_k(x_i)}{\delta_i^2}, \qquad (4)$$

отличны от нуля при $j \ne k$.

Коэффициенты a_j зависят от случайных величин δ_i и поэтому сами являются случайными величинами. Для них хорошо известны классические оценки стандартного уклонения на основе распределения Стьюдента [5]. Сами коэффициенты округляют в пределах этих стандартных уклонений. Однако из-за неортогональности базиса случайные величины a_j оказываются коррелированными. Поэтому их нельзя округлять независимо: это может вносить существенную погрешность в аппроксимирующую кривую [1].

Чтобы преодолеть эту трудность, в [1, 6] был предложен метод аппроксимации с помощью полиномов, ортогонализованных на множестве экспериментальных точек в смысле скалярного произведения (2). В указанных работах были получены оценки доверительных интервалов для коэффициентов регрессии и для аппроксимирующей кривой.

В данной работе проведено количественное сравнение двух способов регрессии: на основе ортогонализованных и неортогональных полиномов. В качестве теста построена модельная задача, имитирующая эксперименты по измерению скоростей химических реакций. Показаны преимущества метода ортогонализованных полиномов. Насколько нам известно, такие процедуры тестирования ранее не применялись.

ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ ПО СКОРОСТЯМ ХИМИЧЕСКИХ РЕАКЦИЙ

В химических экспериментах непосредственно измеряют концентрации реагирующих веществ в зависимости от времени, и по этим профилям определяют скорости реакций K при фиксированной температуре T и давлении p (см., например, [4]).

Как правило, по каждой реакции опубликовано [7–10] много экспериментальных работ, причем диапазон условий в них частично перекрывается, частично различается. Из-за неизбежных погрешностей эксперимента результаты различных авторов отличаются друг от друга, причем нередко эти отличия значительны. Поэтому вся совокупность экспериментальных данных в координатах $1/T - \lg K$ выглядит как размытая прямая. Например, для реакции $C_2H_6 \to 2CH_3$ данные разных авторов для K при фиксированной температуре могут различаться до 20–30 раз (т.е. до 1.5 порядков) [11]. При этом полный диапазон изменения K составлял ~14 порядков.

Такие данные, бесспорно, представительны для тестирования методов регрессии. Однако, чтобы провести адекватное количественное сравнение, необходимо иметь не только массив исходных экспериментальных данных, но и точную кривую, которая в натурном эксперименте неизвестна.

МОДЕЛЬНЫЕ ДАННЫЕ

В данной работе в качестве "экспериментального" материала использовались модельные данные, которые генерировались по следующему правилу. Пусть

$$u = k_0 x + b_0 \tag{5}$$

есть точная "экспериментальная" прямая. Она определена во всем рассматриваемом диапазоне аргумента x.

Пусть имеется M "лабораторий". Для каждой лаборатории зададим диапазон аргумента $[x_{min}, x_{max}]$. В этом диапазоне зададим среднюю кривую данной лаборатории

$$u = kx + b. \tag{6}$$

Здесь k, b есть случайные величины со средними значениями соответственно k_0, b_0 и стандартными отклонениями соответственно δk, δb. Разница средней кривой (5) и точной (4) есть систематическая погрешность данной лаборатории.

Далее возьмем I точек x_i, равномерно распределенных на отрезке $[x_{min}, x_{max}]$. Для каждой точки вычислим u_i по формуле (5) и добавим к этому значению гауссову случайную величину с нулевым средним и стандартным уклонением δu. Последняя величина есть случайная погрешность отдельного измерения. Полная погрешность каждого измерения определялась как разность экспериментальной ординаты и точной прямой $\delta_i = u_i - u(x_i)$.

Таким образом, совокупность точек и их погрешностей $\{x_i, u_i, \delta_i\}$ для всех "лабораторий" являются исходными экспериментальными данными. Для каждой из них известны случайная, систематическая и полная погрешности. Уменьшение величин δk, δb, δu приводит к пропорциональному уменьшению указанных погрешностей.

Пример таких модельных данных приведен на рис. 1. Здесь $M = 4$, $I = 160$, $k_0 = -1$, $b_0 = 2$, $\delta k = \delta b = \delta u = 0.3$. Насколько нам известно, такие процедуры тестирования для задач регрессии экспериментальных данных ранее не применялись.

ОРТОГОНАЛИЗОВАННЫЕ ПОЛИНОМЫ

Опишем построение полиномов, ортогонализованных на множестве экспериментальных точек [1, 6]. Определим средние величины

$$\overline{x^q} = \frac{1}{\Delta}\sum_{j=1}^{J} \delta_j^{-2} x_j^q, \quad \Delta = \sum_{j=1}^{J} \delta_j^2. \tag{7}$$

Пусть $\varphi_j(x)$ есть многочлен степени j. Он имеет j нулей, которые обозначим через $c_{j,k}$, где $1 \le k \le j$. Представим этот многолчен в виде

$$\varphi_n(x) = \prod_{m=1}^{n}(x - c_{n,m}). \tag{8}$$

Очевидно, $\varphi_0(x) = 1$. Многочлен $\varphi_1(x) = x - c_{1,1}$ выберем так, чтобы он был ортогонален $\varphi_0(x)$ в смысле скалярного произведения (4): $\langle\varphi_0, \varphi_1\rangle = 0$. Отсюда нетрудно найти $c_{1,1} = \overline{x}$. Многочлен $\varphi_2(x) = (x - c_{2,1})(x - c_{2,2})$ определим так, чтобы он был ортогонален $\varphi_0(x)$ и $\varphi_1(x)$. Это приводит к квадратному уравнению относительно корней $c_{2,1}$, $c_{2,2}$

$$c^2 + Pc + Q = 0,$$
$$P = -\frac{\overline{x^3} - \overline{x^2}\,\overline{x}}{\overline{x^2} - \overline{x}^2}, \quad Q = \frac{\overline{x^3}\,\overline{x} - \overline{x^2}^2}{\overline{x^2} - \overline{x}^2}. \tag{9}$$

Аналогично строятся кубический полином $\varphi_3(x)$ и многочлены более высоких степеней.

Подставляя разложение по таким полиномам в (2) и проводя минимизацию, найдем коэффициенты a_j

$$a_j = \frac{\langle u, \varphi_j\rangle}{\langle\varphi_j, \varphi_j\rangle}. \tag{10}$$

Чтобы найти доверительные интервалы коэффициентов, проварьируем в (12) экспериментальные данные u_i в пределах их стандартных уклонений δ_i. Величина $\delta(a_j)_i = (\partial a_j / \partial u_i)\delta_i$ есть возмущение коэффициента a_j, вызванное погрешностью i-го измерения. Просуммируем квадраты таких возмущений по всем значениям i и извлечем квадратный корень. Это даст стандартное уклонение коэффициента a_j. Оно равно

$$\delta a_j = \frac{1}{\langle\varphi_j, \varphi_j\rangle^{1/2}}. \tag{11}$$

Таким образом, данный подход позволяет вычислить не только коэффициенты разложения, но и их доверительные интервалы. Поскольку функции φ_j ортогональны, коэффициенты разложения некоррелированы. Поэтому каждый коэффициент можно округлять независимо от других в пределах его доверительного интервала. Коэффициенты a_j можно считать недостоверными, если $|a_j| < \delta a_j$. Недостоверные коэффициенты необходимо отбрасывать и ряд (1) соответственно обрывать.

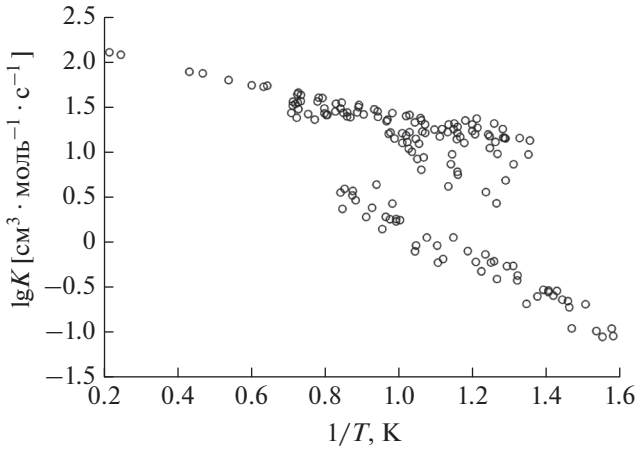

Рис. 1. Пример реализации модельных данных.

Аналогично выводится выражение доверительного коридора полученной аппроксимации

$$\delta\varphi(x) = \pm\left(\sum_{j=1}^{n}\delta a_j^2\varphi_j^2(x)\right)^{1/2}, \tag{12}$$

причем суммируются только члены с достоверными коэффициентами.

РЕЗУЛЬТАТЫ

Мы провели серию расчетов для описанной выше модельной задачи с различными значениями параметров δk, δb, δu. Для простоты в каждом расчете эти величины были одинаковыми $\delta k = \delta b = \delta u \equiv \sigma$. В каждом расчете мы строили аппроксимацию двумя способами: а) по неортогональному базису $\varphi_0 = 1$, $\varphi_1 = x$ и б) по ортогонализованному базису $\varphi_0 = 1$, $\varphi_0 = x - \overline{x}$. Эти аппроксимации определялись для всего рассматриваемого диапазона аргумента x.

Для каждого способа вычислялись оценки доверительных интервалов для коэффициентов δa_0, δa_1 и аппроксимирующей кривой $\delta\varphi(x)$. Эти оценки можно трактовать как апостериорные оценки погрешности. Вычислялось также отличие k_0 и b_0 от соответствующих коэффициентов аппроксимации и точной кривой $u(x)$ от аппроксиманты $\varphi(x)$. Эти разности можно рассматривать как фактическую погрешность относительно точного ответа.

На рис. 2 показано отношение апостериорных оценок к фактической точности для обоих коэффициентов и среднеквадратичной нормы апостериорной оценки к среднеквадратичной норме фактической точности для аппроксимирующей кривой. Здесь выбраны $I = 160$, $M = 4$. По горизонтали отложен параметр σ. Чтобы охватить широкий диапазон значений σ, график построен в двойном логарифмическом масштабе.

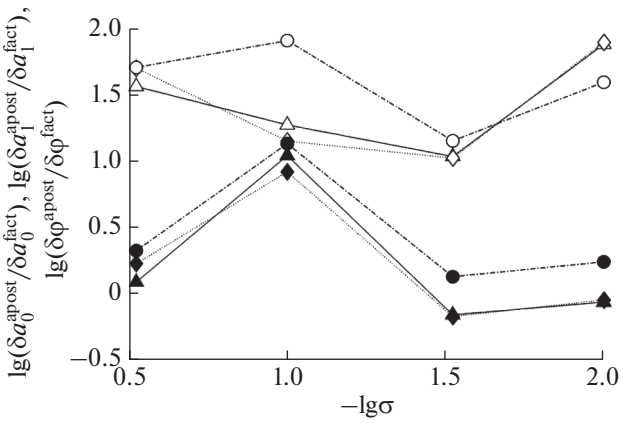

Рис. 2. Отношение оценок доверительных интервалов к фактическим погрешностям. ▲ – a_0, ◆ – a_1, ○ – $\|\delta\varphi\|$.

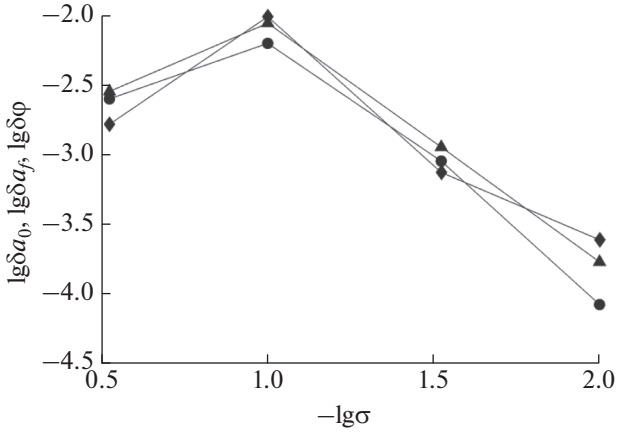

Рис. 3. Сходимость коэффициентов аппроксимации и аппроксимирующей кривой при уменьшении экспериментальных погрешностей. Обозначения соответствуют рис. 2.

Видно, что для обоих способов аппроксимации апостериорная оценка является завышенной. При этом апостериорная оценка для метода ортогонализованных полиномов в 10 и более раз точнее, чем классические оценки для неортогональных полиномов. Это показывает, что регрессия по ортогонализованным полиномам является более надежной, чем по неортогональным.

На рис. 3 показаны апостериорные оценки погрешности δa_0, δa_1, $\delta\varphi$ в зависимости от σ для ортогонализованных полиномов. В качестве примера выбраны $I = 160$, $M = 4$. Видно, что при уменьшении σ указанные погрешности уменьшаются пропорционально σ. Таким образом, имеет место сходимость аппроксиманты и ее коэффициентов к точной кривой и точным значениям коэффициентов соответственно. Проводились расчеты и с другими значениями I и M, для них результаты оказались аналогичными.

ЗАКЛЮЧЕНИЕ

В вычислительной математике алгоритмы тестируются на задачах с известным точным решением. Погрешность расчета непосредственно находят как разность численного и точного решений. Далее исследуют поведение погрешности при уменьшении шага разностной сетки, при внесении возмущений в параметры задачи и т.д.

В данной работе построена аналогичная процедура количественного исследования для методов регрессии экспериментальных данных. Впервые проведено количественное сравнение аппроксимации ортогонализованными и неортогональными полиномами на модельных данных, имитирующих эксперименты по измерению скоростей химических реакций.

Показано, что оценки доверительных интервалов в методе ортогонализованных полиномов существенно (в 10 и более раз) точнее таковых для неортогональных полиномов.

Показано, что при уменьшении погрешности экспериментальных точек коэффициенты и аппроксиманта в методе ортогонализованных полиномов сходятся к точному ответу.

Работа выполнена при поддержке Совета по грантам Президента РФ (проект № МК-3630.2021.1.1).

СПИСОК ЛИТЕРАТУРЫ

1. *Днестровская Е.Ю., Калиткин Н.Н.* Регрессия экспериментальных кривых. Препринты ИПМ им. М.В. Келдыша. 1987. № 181.
2. *Белов А.А., Калиткин Н.Н.* // ДАН. 2016. Т. 470. № 3. С. 266.
3. *Белов А.А., Калиткин Н.Н.* // ЖВМиМФ. 2017. Т. 57. № 11. С. 7.
4. *Кондратьев В.Н., Никитин Е.Е.* Кинетика и механизм газоофазных химических реакций. М.: Наука, 1974
5. *Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф.* Справочник по теории вероятностей и математической статистике. М.: Наука, 1985.
6. *Белов А.А., Калиткин Н.Н.* // ЖВМиМФ. 2020. Т. 60. № 7. С. 105.
7. http://kinetics.nist.gov/kinetics.
8. *Burkholder J.B., Sander S.P., Abbatt J.P.D. et al.* Chemical kinetics and photochemical data for use in atmospheric studies. Evaluation No. 18. JPL Publication 15-10. Pasadena: Jet Propulsion Laboratory, 2015.
9. *Smith G.P., Golden D.M., Frenklach M. et al.* // GRI-Mech 3.0. Berkeley University of California, Gas Research Institute, 2002.
10. *Baulch D.L. et al.* // J. Phys. Chem. Ref. Data. 2005. V. 34. No. 3. P. 757.

11. *Топор О.И., Белов А.А., Федоров И.А.* // Изв. РАН. Сер. физ. 2021. Т. 85. № 2. С. 261; *Topor O.I., Belov A.A., Fedorov I.A.* // Bull. Russ. Acad. Sci. Phys. 2021. V. 85. No. 2. P. 196.

Regression of experimental data via the orthogonalized polynomial method

O. I. Topor[a], *, A. A. Belov[a, b], L. V. Borodachev[a]

[a] *Lomonosov Moscow State University, Moscow, 119991 Russia*

[b] *Peoples' Friendship University of Russia (RUDN University), Moscow, 117198 Russia*

**e-mail: topor.oi15@physics.msu.ru*

The problem of experimental data regression is of large practical importance. In the present work, we perform quantitative comparison of two regression methods: the orthogonalized polynomials and non-orthogonal ones. As a test, we construct a model problem imitating experimental data on chemical reactions. The advantages of the orthogonalized polynomial method are shown.