

$$\lg K_{S_1} = \lg \frac{B W}{S a_{\infty} \gamma R T} - \frac{B n F}{2,3 R T} (\varphi_{n,a} - \varphi^0) - q_1 \lg C_{S_1} ;$$

$$K_{S_2} = 1 \cdot 10^{-10} \text{ см/сек} ; K_{S_1} = 7 \cdot 10^{-8} \text{ см/сек} .$$

Рассчитанные значения констант указывают на необратимость процессов разряда-ионизации мышьяка в исследуемых условиях.

#### Литература

1. А.Г.Стромберг, Л.Н.Попова. Электрохимия, 4, 39, 1968.
2. Н.А.Колпакова. Диссертация, ТПИ, Томск, 1968.
3. H. Matsuda, *Jayabe, Z. Elektrochem*, 63, II64, 1959.
4. X.З.Брайнина. Инверсионная вольтамперометрия твердых фаз. М., "Химия", 1972.
5. Э.А.Захарова. Изв. ТПИ, I28, 53, 1964.
6. Э.К.Спирин. Диссертация, ТПИ, Томск, 1965.
7. Pourbaix M, *Atlas d'equilibres electrochimiques Paris*, 1963.

### СГЛАЖИВАНИЕ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ НЕПАРАМЕТРИЧЕСКОГО ПОДХОДА

А.И.Рубан

В задачах исследования химико-технологических объектов неотъемлемой частью является сглаживание результатов эксперимента, т.е. построение регрессионных зависимостей между переменными. Для этих целей обычно используют метод наименьших квадратов (МНК). Однако он имеет ряд существенных ограничений, сужающих область его применения.

В данной работе предложен общий подход к сглаживанию результатов эксперимента. Он более универсален, чем МНК, и приводит к сравнительно простым вычислительным алгоритмам.

#### 1. Постановка задачи.

Считаем, что имеется система случайных переменных  $Y, X_1, \dots, X_k$ , связанных между собой неизвестной статической зависимостью.

Необходимо на основе экспериментальных данных

$$\begin{aligned} & Y_1, x_{11}, \dots, x_{1k}, \\ & Y_2, x_{21}, \dots, x_{2k}, \\ & \dots \dots \dots \dots \dots \dots \\ & Y_n, x_{n1}, \dots, x_{nk}, \end{aligned} \tag{I.I}$$

где  $l$  - номер измерения, определить оценку  $\hat{M}\{Y|x_1, \dots, x_k\}$  регрессионной поверхности  $M\{Y/x_1, \dots, x_k\}$  - средней зависимости между

Известно / 1 /, что условное математическое ожидание  $M\{Y|x_1, \dots, x_k\}$  удовлетворяет критерию наименьших средне-квадратичных отклонений

$$M\{[Y - M\{Y|x_1, \dots, x_k\}]^2\} = \min_{\eta} M\{[Y - \eta(x_1, \dots, x_k)]^2\} \quad (1.2)$$

Поэтому на практике обычно используют следующий подход.

Задается вид оценки средней зависимости  $Y$  от  $X_1, \dots, X_k$

$$\hat{M}\{Y|x_1, \dots, x_k\} = \hat{\eta}(x_1, \dots, x_k, \alpha) \quad (1.3)$$

с точностью до вектора параметров  $\alpha$ . При этом используется априорная информация о  $M\{Y|x_1, \dots, x_k\}$ . Если таковая отсутствует, то в качестве (1.3) берется отрезок ряда Тейлора. В случае нелинейной зависимости приходится либо учитывать значительное число членов, либо разбивать область  $(X_1, \dots, X_k)$  на части, в каждой из которых допускается линейная аппроксимация. В обоих случаях размерность вектора  $\alpha$  велика и намного превышает  $K$  - размерность области "входных" переменных.

После выбора зависимости (1.3) параметры  $\alpha$  определяются из критерия (1.2), в котором вместо оператора математического ожидания  $M$  берется его оценка

$$J(\alpha) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\eta}(x_{i1}, \dots, x_{ik}, \alpha))^2 = \min_{\alpha} \quad (1.4)$$

Это простейший вид критерия наименьших квадратов.

Метод наименьших квадратов имеет свои особенности, которые существенно ограничивают область его применения.

1. При анализе результатов в МНК считается, что  $(X_1, \dots, X_k)$  измеряются точно. Хотя в настоящее время существует решение задачи при наличии малых аддитивных ошибок измерения переменных  $(X_1, \dots, X_k)$ , но результаты достаточно сложны и на практике не используются.

2. Все помехи приведены к аддитивным ошибкам измерения  $Y$ .

3. Эти ошибки не зависят от "входов"  $(X_1, \dots, X_k)$

4. Известно параметрическое задание поверхности регрессии (1.3).

5. Наилучшие свойства у оценок  $\alpha$  получаются, когда они

линейно входят в оценку регрессионной поверхности (1.3) и когда аддитивные ошибки измерения  $y$  имеют нормальный закон распределения.

Основное преимущество данного подхода заключено в возможности определения оптимальных (в смысле критерия наименьших квадратов) оценок при малых объемах выборок.

Если условия I-5 не выполняются, то для построения оценки регрессионной кривой  $-\hat{\mu}\{y/x_1, \dots, x_k\}$  - можно применить описанный ниже подход, основанный на использовании оценок совместных плотностей вероятности, нашедших широкое распространение в непараметрической статистике. Для одномерного случая (один "вход", один "выход") некоторые результаты содержатся в работах / 2-4 /.

2. Метод решения.

Кривая регрессии однозначно определяется совместной плотностью вероятности  $f(y, x_1, \dots, x_k) = f(y, \bar{x})$  :

$$\rho(\bar{x}) = \mu\{y|\bar{x}\} = \int_{-\infty}^{\infty} y f(y, \bar{x}) dy / \int_{-\infty}^{\infty} f(y, \bar{x}) dy. \quad (2.1)$$

Заменяем плотность  $f(y, \bar{x})$  ее оценкой  $\hat{f}_n(y, \bar{x})$  и получаем оценку для искомой кривой регрессии / 4 /.

$$\hat{\rho}_n(\bar{x}) = \int_{-\infty}^{\infty} y \hat{f}_n(y, \bar{x}) dy / \int_{-\infty}^{\infty} \hat{f}_n(y, \bar{x}) dy. \quad (2.2)$$

Здесь индекс  $n$  означает, что оценки  $\hat{f}_n(y, \bar{x})$  и  $\hat{\rho}_n(\bar{x})$  строятся на основе выборки (1.1) объема  $n$ .

Оценки  $\hat{f}_n(y, \bar{x})$  можно выбирать в различной форме. Однако при недостатке априорной информации желательно иметь оценки с устойчивыми к изменениям этой информации свойствами. Остановимся на использовании парзеновских оценок

$$\hat{f}(y, \bar{x}) = n^{-\frac{4}{k+5}} \sum_{i=1}^n \prod_{j=1}^k \alpha_j K_j \left[ \alpha_j n^{\frac{1}{k+5}} (x_i - x_{ij}) \right] \alpha_0 K_0 \left[ \alpha_0 n^{\frac{1}{k+5}} (y - y_i) \right] \quad (2.3)$$

Здесь коэффициенты размытости и ядра  $K_0, K_1, \dots, K_k$  выбраны оптимальными / 5 /:

$$K_j[\xi] = K[\xi] = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} \xi^2\right), & |\xi| \leq \sqrt{5} \\ 0, & |\xi| \geq \sqrt{5} \end{cases} \quad (2.4)$$

$$j = 0, 1, \dots, k$$

Постоянные положительные коэффициенты  $\{\alpha_j\}$  определяются видом неизвестной плотности  $f(y, \bar{x})$ . Далее неопределенность относительно этих коэффициентов мы сохраним, чтобы с помощью их выбора решить наилучшим образом основную задачу о построении сглаживающей линии регрессии. Заметим, что при любых  $0 < \alpha_j < \infty$ ,  $j = 1, \dots, K$  оценка  $\hat{f}_n(y, \bar{x})$  сходится при  $n \rightarrow \infty$  к истинной плотности  $f(y, \bar{x})$ , если выборочные значения (1.1) независимы при различных  $n$ .

Подставляем (2.3) в (2.2) и получаем оценку поверхности регрессии

$$\hat{p}_n(\bar{x}, \alpha) = \frac{\sum_{i=1}^n y_i \prod_{j=1}^K K_j [\alpha_j n^{\frac{1}{K+5}} (x_i - x_{ij})]}{\sum_{i=1}^n \prod_{j=1}^K K_j [\alpha_j n^{\frac{1}{K+5}} (x_i - x_{ij})]} \quad (2.5)$$

Затем параметры  $\alpha = (\alpha_1, \dots, \alpha_K)$  выбираем из требования наилучшего сглаживания (1.4).

Преимущества описанного метода построения нелинейных регрессионных зависимостей.

1. Отпадает необходимость проводить параметризацию регрессионной кривой. Процедура параметризации довольно сложная, требует наличия значительной априорной информации, и поэтому в многомерном случае обычно исследователи ограничиваются линейной или квадратичной аппроксимацией.

2. При оптимальном выборе структуры коэффициентов размерности число подстраиваемых параметров  $\alpha_1, \dots, \alpha_K$  имеет размерность пространства "входных" переменных -  $K$ , т.е. задача подстройки  $\alpha$  из критерия (1.4) по размерности совпадает с задачей построения линейного регрессионного уравнения. Однако нелинейная зависимость от  $\alpha$  несколько усложняет процедуру их поиска. Ниже будет приведен непараметрический алгоритм их подстройки, основанный на использовании уравнения типа (2.5). Следует заметить, что с ростом  $n$  чувствительность функции качества  $J(\alpha)$  по параметрам  $\alpha$  падает и при любом  $0 < \alpha < \infty$  условия сходимости оценки  $\hat{p}_n(\bar{x}, \alpha)$  к  $p(\bar{x})$  выполняются. Т.е., если мы располагаем большим объемом выборки  $n$ , то выбор  $\alpha$  становится не критичен, и при любом их наборе мы получим хорошее качество сглаживания экспериментальных значений.

3. Получаемая оценка регрессионной кривой всегда явля-

ется асимптотически несмещенной и состоятельной (даже при наличии ошибок измерения "входов").

Остановимся теперь на алгоритме вычисления  $K$  параметров  $\alpha_1, \dots, \alpha_k$  из критерия (1.4). Эти параметры являются случайными величинами, ибо объем выборки конечен, а сами параметры  $\alpha$  зависят от выборочных значений (1.1) - случайных величин.  $J(\alpha)$  тоже случайная величина. Тогда, используя общую схему построения адаптивных алгоритмов / 6 /, получим итеративный алгоритм подстройки  $\alpha$

$$\alpha_j[e] = \frac{\sum_{i=1}^{e-1} \alpha_j[i] K \left[ \beta(e-1)^{1/6} (J^* - J(\alpha[i])) \right]}{\sum_{i=1}^{e-1} K \left[ \beta(e-1)^{1/6} (J^* - J(\alpha[i])) \right]} + \Delta \alpha_j[e] \quad (2.6)$$

$e = 2, 3, 4 \dots$

Здесь:  $e$  - номер итерации;  $K(\xi)$  - ядро (2.4);  $\beta$  - положительная константа, один из способов выбора которой приведен ниже;  $J(\alpha[i])$  - функция качества  $J$ , рассчитанная при  $\alpha_1 = \alpha_1[i]$ ,  $\alpha_2 = \alpha_2[i], \dots, \alpha_k = \alpha_k[i]$ ;

$$J^* = \min \{ J(\alpha[1]), \dots, J(\alpha[e-1]) \} \quad (2.7)$$

В алгоритме (2.6) на первых операциях величину  $J^*$  можно полагать равной нулю.

Опишем начальный этап расчета по (2.6). Задаем произвольно  $\alpha[1]$  и вычисляем  $J(\alpha[1]) = J_1$ . Выбираем  $\beta$  из условий, чтобы знаменатель (2.6) при  $e = 2$  не был равен нулю:

$$0 < \beta < \frac{\sqrt{5}}{J_1} \quad (2.8)$$

Затем на следующих итерациях выбранное  $\beta$  оставляем постоянным. Полагаем  $J^* = 0$ , вычисляем  $\alpha[2]$  и т.д.

3. Пример. Имеется ряд измеренных значений некоторой изменяющейся во времени переменной

$$y(t_1), y(t_2), \dots, y(t_n). \quad (3.1)$$

Зависимость от времени неизвестна и возникает необходимость отфильтровать ошибки измерения, т.е. получить сглаженную функцию времени  $\hat{y}(t)$ . Решение получаем с помощью алгоритма (2.5)

$$\hat{y}(t) = \frac{\sum_{i=1}^n y(t_i) K \left[ \alpha n^{1/6} (t - t_i) \right]}{\sum_{i=1}^n K \left[ \alpha n^{1/6} (t - t_i) \right]}, \quad (3.2)$$

где  $\alpha$  - скалярный параметр, расчет которого можно осуществить по формуле (2.6).

4. Заключение. Описанный в работе общий подход к построению алгоритмов сглаживания результатов эксперимента позволяет строить вычислительные схемы, просто реализуемые на ЭВМ. Метод универсален, поэтому автор рекомендует использовать его при исследовании сложных, плохо изученных процессов в объектах химии и химической технологии.

В настоящее время разрабатывается стандартная процедура алгоритма сглаживания. Запуск ее позволит инженерам-химикам-технологам в короткий срок обрабатывать на ЭВМ большие массивы экспериментальных данных с целью расчета математической взаимосвязи между параметрами.

#### Литература

1. В.С.Пугачев. Теория случайных функций. ГИ ФМЛ, М., 1962.
2. Э.А.Надарая. Замечания о непараметрических оценках плотности вероятности и кривой регрессии. Теория вероятностей и ее применение, т. 15, вып. I, 1970.
3. В.П.Живоглядов, А.В.Медведев. Непараметрические алгоритмы идентификации, распознавания образцов и дуального управления. В сб. "Исследование и оптимизация стохастических распределенных систем", Изд. "ИЛИМ", Фрунзе, 1971.
4. Н.Н.Апрушева, В.Д.Конаков. Использование непараметрических оценок в регрессионном анализе. Заводская лаборатория, 24, № 5, 1973.
5. А.П.Серых. Об использовании непараметрических оценок плотности в задачах распознавания образов. Труды СФТИ, вып.63, Математическая статистика и ее приложение, 1973.
6. В.П.Живоглядов, А.В.Медведев. Непараметрические стохастические алгоритмы управления и поиска экстремума. Автоматика и вычислительная техника, № 2, 1973.