ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2000, том 10, № 1, с. 25–34

ОРИГИНАЛЬНЫЕ СТАТЬИ

УДК 519.233.5+519.252

© А. Л. Буляница, В. Е. Курочкин

ОЦЕНИВАНИЕ НЕОБХОДИМОГО ЧИСЛА ТОЧЕК НАБЛЮДЕНИЯ ПРИ ПОСТРОЕНИИ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Решается задача определения числа n точек наблюдения $x_1, ..., x_n$, позволяющего исключить критические точки и обеспечить приемлемую точность построения регрессионной модели. В основе лежат методы выявления критических точек (разбалансировки и риска) x_j, предложенные Хьюбером.

ВВЕДЕНИЕ

Одним из этапов обработки экспериментальных данных является построение регрессионной модели, выявляющей функциональную зависимость вектора наблюдений **Y** с координатами $y_1,..., y_n$ от вектора точек наблюдений **X** с координатами $x_1, ... , x_n$.

Многие экспериментальные зависимости $\mathbf{Y} = F(\mathbf{X})$ можно линеаризовать и использовать двухпараметрическую линейную модель вида

$$y_j = \theta_1 + \theta_2 x_j + u_j, \qquad (1)$$

где θ_1 — начальное смещение, θ_2 — параметр положения, u_j — случайная ошибка оценивания. Величины u_j одинаково распределены, имеют нулевое среднее и ограниченную дисперсию.

В этом случае метод наименьших квадратов (МНК) обеспечивает несмещенность оценок параметров модели (1) и наименьшую дисперсию ошибки оценивания $y_1, ..., y_n$. Вектор оцененных (подогнанных) значений $\mathbf{Y}^*$ получают по формуле $\mathbf{Y}^* = \mathbf{HY}$. При этом

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \qquad (2)$$

Матрица плана **X** определена точками наблюдений $x_1, ..., x_n$ и для модели (1) принимает вид

$$\begin{bmatrix} 1 \ldots 1 \\ x_1 \ldots x_n \end{bmatrix}^T. \qquad (3)$$

Индекс T обозначает транспонирование.

Особенности применения термина «линейный» к модели (1) поясняются Кендаллом и Стьюартом [1]. В отличие от решенной Кендаллом и Стьюартом [2] задачи оптимального выбора $x_1, ..., x_n$ с целью минимизации дисперсии МНК-оценок параметров модели (1), в нашем случае x_j либо задаются предварительно, либо определяются характером исследуемого процесса.

Хьюбер [3] называет **H** матрицей подгонки и связывает наличие критических точек наблюдения x_n — точек разбалансировки — с ее диагональными элементами $H_{jj} = h_j$.

Здесь необходимо отметить, что наличие погрешностей в величине наблюдений в таких точках в силу их положения может повлиять на параметры зависимости подгонки непредсказуемым образом, а следовательно, и на оцениваемый вектор **Y***.

Классификация точек наблюдения x_j, согласно [3], производится в соответствии с условиями:

$$H_{jj} > 0{,}5, \qquad (4)$$

$$0{,}2 < H_{jj} \le 0{,}5, \qquad (5)$$

$$H_{jj} \le 0{,}2. \qquad (6)$$

Условие (4) позволяет квалифицировать x_j как точку разбалансировки, и ее использование в построении модели (1) при наличии погрешности в наблюдении y_j приведет к практически неконтролируемому возрастанию ошибки регрессионного оценивания. При выполнении (5) x_j считается точкой риска, и ее использования следует избегать. Выполнение условия (6) говорит о надежности использования точек x_j для наблюдения величины y_j.

Величины h_j явно определяются выражением

$$h_j = \frac{S_2 - 2S_1x_j + S_0x_j^2}{S_0S_2 - S_1^2}, \qquad (7)$$

где $S_k = \sum_{i=1}^{n}(x_i)^k, \quad k = 0{,}1{,}2$.

Хьюбером [3] доказаны два важнейших свойства h_j:

1. $h_j \in [0;1]$, $j = 1, ..., n,$ поскольку **H** является проективной матрицей.

2. $\mathrm{tr}(\mathbf{H}) = p$, где p — число параметров регрессионной модели. В нашем случае (1) $p = 2$.

В Приложении доказаны остальные свойства h_j:
- — инвариантность относительно элементарных преобразований x_j: изменения масштаба (Δ) и линейного сдвига;
- — m-кратный повтор измерений в каждой точке x_j приводит к уменьшению всех h_j в m раз, т. е. $h_j^* = h_j/m$, $j = 1,..., n$;
- — $h_j \geq 1/n$, $j = 1,..., n$;
- — $h_j^{\max} \geq 2/n$;
- — удвоение измерения в одной из точек x_j приводит к заведомому выполнению (5) или (6) применительно к измерению в указанной точке;
- — добавление какого-либо нового измерения приводит к уменьшению всех h_j, за исключением, быть может, одного.

Расположение совокупности x_j будем называть стратегией измерения. Выбор стратегии связан с характером исследуемой зависимости $\mathbf{Y}=F(\mathbf{X})$, должен обеспечить ее линеаризацию и адекватность построения регрессионной модели (1). Кроме того, в понятии стратегия необходимо учитывать две разные возможности получения величин y_j в точках x_j:

а. По схеме принципиально неповторимых (уникальных) измерений — так называемая схема измерений без дублирования. Подобным свойством обладает кинетическая кривая (x_j — величины, определяемые моментами времени измерения t_j, y_j — величина отклика).

б. По схеме измерений с дублированием. Например, калибровочная характеристика (x_j — концентрация j-й пробы, y_j — информативный параметр). В этом случае неоднократный повтор измерения в любой точке x_j теоретически возможен.

Вторым элементом стратегии измерений является задание зависимости $x_j = f(j)$, $j = 1,..., n$. Такое задание обусловлено необходимостью линеаризации зависимости $\mathbf{Y} = F(\mathbf{X})$ и может в значительной степени определяться методикой измерения.

Наиболее распространенной стратегией измерения представляется равномерная по x_j вида В1 — $f(j) = j\Delta$. Эта стратегия оптимальна для исследования процессов, близких к линейным, в том числе при анализе градуировочных зависимостей.

В рамках последующих стратегий В2–В5 осуществлено преобразование координат x_j, позволяющее линеаризовать зависимость $\mathbf{Y} = F(\mathbf{X})$ и свести ее к виду (1).

С помощью стратегии В2 — $x_j = j^2\Delta$, где $x_j = t_j$ — возможна линеаризация экстракционного процесса [4, 5], описываемого уравнением (8)

$$y_j = \alpha(C_0)\sqrt{t_j}\,. \qquad (8)$$

Здесь параметр положения α является функцией исходной концентрации C_0 экстрагируемого вещества: $\alpha = C_0\sqrt{\pi D}$, D — коэффициент диффузии.

Равномерный по времени съем данных в этих же условиях приведет к стратегии измерения В3: $x_j = \sqrt{j}\,\Delta$. Сама же стратегия В3 позволит линеаризовать квадратичную зависимость типа $E_j = mV_j^2/2$ при необходимости оценивания массы m.

Стратегией измерения В4 — $x_j = \log(j)\Delta$ — линеаризуется кинетическая кривая, характерная для процессов радиоактивного распада.

Кривую изменения оптической плотности при фотометрических измерениях можно линеаризовать с помощью стратегии измерения В5: $x_j = \alpha^{j-1}\Delta$ при $\alpha \neq 1$.

Целью работы является обоснование необходимых условий измерения на основе расчета минимально необходимого числа точек наблюдения n, что позволит исключить из числа x_j критические точки (разбалансировки или риска) при любой из заданных стратегий измерения В1–В5.

1. РАСЧЕТ НЕОБХОДИМОГО ЧИСЛА ТОЧЕК НАБЛЮДЕНИЯ В СЛУЧАЕ БЕЗ ДУБЛИРОВАНИЯ

1.1. Исследование равномерной стратегии измерения В1

В этом случае выражения (7) примут вид

$$S_1 = n(n+1)/2\,,\quad S_2 = n(n+1)(2n+1)/6\,,$$

$$h_j = \frac{2}{n(n-1)}\left[2n+1-6j+\frac{6j^2}{n+1}\right]. \qquad (9)$$

Согласно (9), h_j симметричны относительно середины диагонали

$$j^* = (n+1)/2 \qquad (10)$$

и имеют минимум по j в одной или двух (в зависимости от четности n) ближайших к j^* точках. Максимальными элементами, таким образом, являются h_1 и h_n. Указанное свойство, безусловно, коррелирует с замечанием Кендалла и Стьюарта [2] о том, что «отрезок прямой наиболее эффективно определяется своими крайними точками».

Исключение точек разбалансировки (цель 1) в x_1 и x_n связано с (4) в форме $h_1 = h_n < 0{,}5$ и сводится к решению квадратного неравенства

$$\frac{4n-2}{n(n+1)} < 0{,}5\,,$$

что в целых числах дает $n \geq 7$.

Отсутствие точек риска (цель 2) связано с (5) и сводится к $n \geq 19$.

Асимптотические оценки h_j в условиях $n \to \infty$ принимают вид

$$h_j^{\min} \to 1/n, \quad h_1 = h_n \to 4/n .$$

1.2. Исследование стратегий измерения B2–B5

Аналогичным образом проводится исследование других стратегий B2–B5. Результаты расчетов сведены в табл. 1.

1.3. Обсуждение результатов

На основе представленных выше рассуждений и данных табл. 1 можно утверждать, что

1. Исходя из инвариантности элементов подгоночной матрицы **H**, очевидно, что изменение интервала дискретности или задержка начала наблюдений при сохранении их числа *n не способны* исключить критические точки. Единственным способом их исключения при сохранении выбранной стратегии измерения B1–B5 является увеличение числа наблюдений до величины, указанной в табл. 1.

2. Выбор стратегии измерений B4, способствующей наилучшей линеаризации логарифмических процессов, при определенных условиях (больших основаниях α) неизбежно сохраняет точки разбалансировки или точки риска.

3. В тех случаях, когда форма процесса допускает выбор различных стратегий измерения, этот выбор может явиться средством борьбы с критическими точками. Например, кривая, построенная по 25 точкам, имеет потенциальные точки риска (см. табл. 1) в том случае, если стратегией измерения были B2, B3 или B5. Однако стратегия B1 исключает такую возможность.

4. Потенциальными точками разбалансировки являются наиболее удаленные от средней точки наблюдения x_j. Для равномерной стратегии B1 таких точек будет 2, а для неравномерных (например, B2–B5) — одна (см. табл. 1).

1.4. Влияние малых вариаций точек наблюдения x_j на величины элементов h_j

Эффект влияния малых вариаций можно оценить с помощью матрицы влияния **D**, элементы которой определяются как $d_{ij} = \left.\dfrac{\partial h_i}{\partial \varepsilon_j}\right|_{\varepsilon_j \to 0}$, где ε_j — малые вариации x_j ($i, j = 1,..., n$).

В явном виде элементы матрицы **D** при выбранной стратегии измерения B1 представимы в форме:

$$d_{ij} = \frac{2(j-i)\Delta - n(2j-n-1)(S_2 - 2S_1 i + ni^2)}{\Delta^2} + \frac{2(-S_1 + nj)}{\Delta}\delta_{ij}, \tag{11}$$

где δ_{ij} — символ Кронекера, $\Delta = n^2(n^2-1)/12$, S_1 и S_2 определяются (9).

Вследствие симметрии h_j (10) относительно j^* при стратегии B1, матрица **D** имеет регулярную псевдокососимметричную структуру. (При повороте элемента матрицы вокруг ее условного центра — элемента $d_{j^*j^*}$ на 180 градусов получается тот же элемент с противоположным знаком).

Таблица 1. Характеристики стратегий измерения B1–B5

Параметр	Стратегия				
	B1	**B2**	**B3**	**B4**	**B5**
№ max элемента	1 (n)	1	n	n	1
№ min элемента	$(n+1)/2$	$4n/9$	$n/\sqrt{3}$	$1+\log\left(\dfrac{\alpha^n-1}{n}\right)/\log(\alpha)$	$n(2\pi n)^{\frac{1}{2n}}/e$
Цель 1, n_{min}	7	9	9	$6\left(\sqrt{2}-\alpha\right)^{-\frac{1}{2}}$	14
Цель 2, n_{min}	19	30	30	$7\left(\sqrt{5}/2-\alpha\right)^{-\frac{2}{3}}/2$	72
h_1	$4/n$	$9/n$	$9/(4n)$	$1/n$	$\log^2(n)/n$
$h_j^{\min}$	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$
h_n	$4/n$	$3/n$	$6/n$	$(\alpha^2-1)/\alpha^2$	$2/n$

Для частного случая $n = 4$ ($h_1 = h_4 = 0{,}7$; $h_2 = h_3 = 0{,}3$), а матрица влияния **D** примет вид

$$D = \begin{bmatrix} -0{,}18 & +0{,}24 & +0{,}06 & -0{,}12 \\ +0{,}08 & -0{,}14 & +0{,}04 & +0{,}02 \\ -0{,}02 & -0{,}04 & +0{,}14 & -0{,}08 \\ +0{,}12 & -0{,}06 & -0{,}24 & +0{,}18 \end{bmatrix}. \qquad (12)$$

Так как для линейной регрессионной модели (1), независимо от стратегии измерений и вариации x_j, tr(**H**) = 2, то сумма по столбцам элементов матрицы **D** есть нуль. Нулевая сумма строк матрицы **D** является следствием инвариантности элементов **H** относительно постоянного смещения.

Воздействие малых вариаций x_j — ε_j на элементы матрицы **H** является суперпозицией простейших воздействий (при только одном $\varepsilon_j \neq 0$) и определяется как $\delta h_i \approx \sum_{j=1}^{n} d_{ij}\varepsilon_j$.

Например, пусть n = 4, а x_j примерно удовлетворяют стратегии B1: $x_1 = 1{,}05$; $x_2 = 2{,}00$; $x_3 = 2{,}90$; $x_4 = 4{,}00$. Можно полагать, что $\varepsilon_1 = -0{,}05$; $\varepsilon_2 = 0{,}00$; $\varepsilon_3 = -0{,}10$; $\varepsilon_4 = 0{,}00$. Оценки элементов матрицы **H**, полученные на основе (11) и (12) (приближенные) и точные значения элементов матрицы **H**, соответствующие указанной стратегии, представлены в табл. 2. Невозмущенной оценкой следует считать элементы матрицы **H**, соответствующие «идеальной» стратегии измерения B1.

Аппарат матриц влияния позволяет оценивать элементы матрицы **H**, в том числе соответствующие критическим точкам наблюдения, не явным расчетом, а через некоторую близкую «идеальную», возможно, нереализуемую стратегию измерения. В частности, возможен учет влияния отклонения от выбранной стратегии вследствие субъективного фактора (задержка времени съема данных, отклонение концентрации от заданного значения при калибровке и т.п.).

1.5. Выводы

1. В данном разделе рассмотрены условия организации экспериментов в рамках схемы без дублирования измерений, позволяющие избегать критических точек наблюдения. Отсутствие таких точек позволяет исключить неконтролируемое возрастание ошибки регрессионного оценивания и позволяет объяснить указанную ошибку исключительно статистическим выбросом наблюдаемого значения y_j.

2. Аппарат матриц влияния позволяет рассчитывать элементы подгоночной матрицы **H** на основе какой-либо близкой стратегии измерения (например, B1–B5), а также определить, какие ошибки оператора (вариации x_j) способны оказать наибольшее влияние на появление (исключение) точек разбалансировки или риска.

2. РАСЧЕТ НЕОБХОДИМОГО ЧИСЛА ТОЧЕК НАБЛЮДЕНИЯ В СЛУЧАЕ ИЗМЕРЕНИЙ С ДУБЛИРОВАНИЕМ

Можно утверждать, что при условии принципиальной возможности повтора измерений в любой точке x_j неравенства (5) или (6) могут выполняться при одной из измерительных схем: 2+1+2, 1+2+2 или 2+2+1. То есть при общем числе измерений 5, выполненных в трех различных точках x_j.

Исходя из доказанного ранее свойства инвариантности, будем полагать следующие точки измерения: $x_{1,2} = 0$; $x_{3,4} = 1$ и $x_5 = z$.

В этом случае значения h_j в явном виде равны соответственно

$$h_{1,2}(x = 0) = \frac{2 + z^2}{2(2z^2 - 2z + 3)},$$
$$h_{3,4}(x = 1) = \frac{z^2 - 2z + 3}{2(2z^2 - 2z + 3)}, \qquad (13)$$
$$h_5(x = z) = \frac{2z^2 - 2z + 1}{2z^2 - 2z + 3}.$$

Проверка выполнения неравенства $h_j \le 0{,}5$ применительно ко всем следовым элементам подгоночной матрицы **H** позволяет сформулировать ограничения, накладываемые на измерение $x_5 = z$.

При этом если $z<0$, то схема измерения 1+2+2; если $0 < z < 1$, то 2+1+2 и при $z>1$ — 2+2+1. При $z=0$ или $z=1$ реализуется схема 3+2 или 2+3 соответст-

Таблица 2. Расчет элементов подгоночной матрицы на основе матрицы влияния

Элемент	Вид оценки		
	невозмущенная	точная	приближенная
h_1	0,7000	0,6839	0,6850
h_2	0,3000	0,2999	0,3000
h_3	0,3000	0,2857	0,2850
h_4	0,7000	0,7304	0,7300

венно. Результаты исследования указанных схем измерения сведены в табл. 3.

Таким образом, измерительная схема вида 2+1+2 обеспечивает исключение точек разбалансировки независимо от соотношения между точками измерения x_j. Однако указанная схема принципиально нереализуема в случае построения кинетической кривой или при иной схеме без дублирования.

3. ИССЛЕДОВАНИЕ РОЛИ ТОЧЕК РАЗБАЛАНСИРОВКИ ПРИ РЕШЕНИИ ЗАДАЧИ ОЦЕНИВАНИЯ ПАРАМЕТРОВ РЕГРЕССИОННОЙ МОДЕЛИ (1)

3.1. Влияние точек разбалансировки на точность оценивания параметров линейной регрессионной модели (1)

Как правило, калибровочные характеристики $y_j = \varphi(x_j)$ являются кусочно-детерминированными линейными трендами первого порядка. Таким образом, для построения адекватной линейной регрессионной модели (1) можно выбрать любую стратегию $x_j = f(j)$, наиболее удобной из которых является стратегия S1: $x_j = j$. В рамках указанной стратегии значения концентраций x_j равномерно изменяются в исследуемом диапазоне концентраций x_1–x_n. Методически более обоснованной представляется другая стратегия измерения S2: $x_j = 1/j$. Такой выбор стратегии обоснован тем, что калибровочные растворы получены разбавлением наиболее концентрированного раствора концентрации x_1 в j раз, $j = 1,..., n$. (Последовательность точек измерения x_j в этом случае упорядочена по убыванию).

В работе [6] показан пример калибровочной характеристики хемосенсора для определения концентрации ионов железа (Fe^{3+}). Характеристика представляет собой совокупность двух линейных трендов первого порядка. Первый линейный участок соответствует диапазону концентраций Fe^{3+} от $2{,}5\cdot10^{-5}$ до $1\cdot10^{-4}$ моль/л. Линейная модель калибровочной характеристики имеет вид

$$A\cdot10^2 = -0{,}3000+888{,}0\cdot C_{Fe}. \qquad (14)$$

Здесь A — отклик хемосенсора, C — концентрация ионов Fe^{3+}.

В соответствии с выбранными стратегиями измерения S1 и S2 калибровочная характеристика должна строиться по следующим точкам (при $n = 4$):

$$\begin{aligned}
&\text{S1:}\quad x_1 = 2{,}5\cdot10^{-5},\ y_1 = 1{,}92\cdot10^{-2},\\
&\qquad x_2 = 5\cdot10^{-5},\ y_2 = 4{,}14\cdot10^{-2},\\
&\qquad x_3 = 7{,}5\cdot10^{-5},\ y_3 = 6{,}36\cdot10^{-2},\\
&\qquad x_4 = 1\cdot10^{-4},\ y_4 = 8{,}58\cdot10^{-2};\\
&\text{S2:}\quad x_1 = 2{,}5\cdot10^{-5},\ y_1 = 1{,}92\cdot10^{-2},\\
&\qquad x_2 = 3{,}3\cdot10^{-5},\ y_2 = 2{,}66\cdot10^{-2},\\
&\qquad x_3 = 5\cdot10^{-5},\ y_3 = 4{,}14\cdot10^{-2},\\
&\qquad x_4 = 1\cdot10^{-4},\ y_4 = 8{,}58\cdot10^{-2}.
\end{aligned} \qquad (15)$$

Согласно ранее полученным результатам, подобные стратегии измерения не гарантируют отсутствия точек разбалансировки, так как условие $n \geq 5$ не выполняется. Элементы матрицы **H** (7) для указанных стратегий будут соответственно

S1: $h_1 = 0{,}70,\ h_2 = 0{,}30,\ h_3 = 0{,}30,\ h_4 = 0{,}70;$

S2: $h_1 = 0{,}47,\ h_2 = 0{,}35,\ h_3 = 0{,}25,\ h_4 = 0{,}93.$

Задав отклонение результата измерения y_j от (15) на величину $\delta = +0{,}1\cdot10^{-2}$, получим регрессионное уравнение, построенное по точкам (x_j, y_j), отклоненное от (14). Величина отклонения зависит от h_j, соответствующего точке ошибочного измерения x_j. Полученные регрессионные уравнения представлены в табл. 4.

Таблица 3. Исследование схем измерения (с дублированием)

Схема			Условие Хьюбера	Критическая точка	Дополнительные требования
x_1	x_2	x_3			
2	2	—	= 0,5	x_1, x_2	—
2	3	—	= 0,5	x_1	—
3	2	—	= 0,5	x_2	—
2	1	2	< 0,5	x_1	$x_2 > (x_1+x_3)/2$
2	1	2	< 0,5	x_3	$x_2 < (x_1+x_3)/2$
1	2	2	< 0,5	x_1	$(x_2 - x_1) < (\sqrt{3}-1)(x_3 - x_2)/2$
2	2	1	< 0,5	x_3	$(x_3 - x_2) < (\sqrt{3}-1)(x_2 - x_1)/2$

Роль точки разбалансировки как фактора искажения коэффициентов линейной регрессионной модели (1) особенно ярко проявляется применительно к изменению α_1 — тангенса угла наклона. Подобный эффект особенно опасен при экстраполяции калибровочной зависимости на более широкий диапазон концентраций.

В работе [7] анализируется отклик хемосенсорного анализатора концентрации ионов меди (Cu^{2+}). Показано, что на начальном этапе измерений, например 1–16 с, отклик удовлетворяет уравнению (8).

Возможно ввести такие единицы измерения информативного параметра, чтобы при условии отсутствия помех измерения отклик хемосенсора удовлетворял уравнению $y_i = 1{,}000\sqrt{t_i}$.

Для различного выбора стратегии измерения проиллюстрируем динамику изменения оценок параметров тренда в случае внесения ошибки измерения информативного параметра равной δ=0,2 (см. табл. 5).

Сохраняется тенденция наиболее сильного влияния на изменения коэффициентов регрессионной модели (1) измерений с большими значениями h_j. Для равномерной стратегии измерения B1 — это x_1 и x_4, для стратегии B3 — x_1.

3.2. Методы борьбы с точками разбалансировки

Как было показано ранее, одним из способов борьбы с точками разбалансировки является увеличение числа измерений n. Этот способ приемлем для всех случаев, в том числе и для однократных измерений (схема без дублирования). Другой способ — точное или приближенное дублирование в критических точках измерения, то есть для x_j, в которых h_j принимает наибольшие значения. Очевидно, что в случае принципиально неповторимых измерений возможно только *приближенное* дублирование.

Как показано в Приложении, критической точкой измерения, для которой соответствующий элемент матрицы разбалансировки h_j принимает наибольшее значение, может быть либо $x_{\min}$, либо $x_{\max}$. Если функция *f(j)*, описывающая стратегию измерения, выпукла вверх, то есть $d^2 f / dj^2 < 0$, то критической точкой является $x_{\min}$. В противном случае — $x_{\max}$. Для стратегии B1 $f(j) = j$ и $f'' = 0$. В этом случае обе точки $x_{\min}$ и $x_{\max}$ являются критическими.

В нашем случае, если критическими являются $x_j = 1$ (или $x_j = 4$), следует провести дополнитель-

Таблица 4. Связь ошибки оценивания параметров модели (14) с величиной элементов h_j

Стратегия	Искаженное измерение	h_j	Уравнение
S1	x_1	0,70	$A\cdot 10^2 = -0{,}2000 + 876{,}0\cdot C_{Fe}$
	x_2	0,30	$A\cdot 10^2 = -0{,}2500 + 884{,}0\cdot C_{Fe}$
	x_3	0,30	$A\cdot 10^2 = -0{,}3000 + 892{,}0\cdot C_{Fe}$
	x_4	0,70	$A\cdot 10^2 = -0{,}3500 + 900{,}0\cdot C_{Fe}$
S2	x_1	0,47	$A\cdot 10^2 = -0{,}2318 + 879{,}8\cdot C_{Fe}$
	x_2	0,35	$A\cdot 10^2 = -0{,}2445 + 882{,}3\cdot C_{Fe}$
	x_3	0,25	$A\cdot 10^2 = -0{,}2702 + 887{,}2\cdot C_{Fe}$
	x_4	0,93	$A\cdot 10^2 = -0{,}3471 + 902{,}0\cdot C_{Fe}$

Таблица 5. Влияние выбора стратегии измерения на точность оценивания параметров регрессионной модели

Стратегия	t_j	x_j	h_j	Уравнение
B1	1	1,00	0,70	$0{,}2000 + 0{,}9400\cdot x_j$
	4	2,00	0,30	$0{,}1000 + 0{,}9800\cdot x_j$
	9	3,00	0,30	$0{,}0000 + 1{,}0200\cdot x_j$
	16	4,00	0,70	$-0{,}1000 + 1{,}0600\cdot x_j$
B3	1	1,00	0,82	$0{,}2313 + 0{,}9327\cdot x_j$
	6	2,45	0,26	$0{,}0760 + 0{,}9904\cdot x_j$
	11	3,32	0,33	$-0{,}0172 + 1{,}025\cdot x_j$
	16	4,00	0,59	$-0{,}0901 + 1{,}052\cdot x_j$

ное измерение, например в точках $x_j = 1{,}05$ или $x_j = 3{,}95$ соответственно. Если же реальный временно́й интервал между исходными измерениями мал, то схема приближенного дублирования критических измерений из-за ограниченного быстродействия технически неосуществима.

Влияние эффекта увеличения числа измерений (до $n = 5$) на точность оценивания параметров регрессионной модели при сохранении измерительных стратегий B1 и B3 иллюстрируется данными табл. 6.

Уменьшение величин h_j приводит к уменьшению погрешности оценивания информативного параметра α. При $n = 4$ наибольшее отклонение составляло ±6,0 % для стратегии B1 и –6,7 % для B3; при $n = 5$ — ±5,3 % и –6,2 %, соответственно.

Рекомендованный способ точного и приближенного дублирования, в частности, означает и добавление одной точки к общему числу измерений.

Эффект от точного и приближенного дублирования в точке x_1 при стратегии измерения B1 сводится к следующим результатам:

а. *Точное дублирование*

Измерение в точках –1, 1, 2, 3, 4.

Элементы матрицы разбалансировки –0,412; 0,412; 0,206; 0,294; 0,676.

Помеха измерения в дублированной критической точке x_1 приводит к уравнению линейной регрессии $0{,}1176+0{,}9647\cdot x_j$. Погрешность оценивания равна –3,5 %. Помеха в недублированной критической точке x_4 позволяет получить регрессионное уравнение вида $-0{,}0784+1{,}0540\cdot x_j$. Погрешность оценивания +5,4 %.

б. *Приближенное дублирование*

Измерение в точках –1; 1,05; 2; 3; 4.

В случае появления помехи на одном из дублированных измерений регрессионные уравнения примут вид соответственно $0{,}1200+0{,}9638\cdot x_j$ и $0{,}1167+0{,}9653\cdot x_j$.

Последствия от помехи в недублированном измерении x_4 будут практически такие же, как и в случае точного дублирования.

Для равномерной стратегии измерения B1 эффект дублирования измерения в одной из двух критических точках наиболее слаб, так как в обеих критических точках значения h_j одинаково. В этом случае более сильным будет эффект от дублирования измерения в обеих критических точках — x_1 и x_4, при сохранении равномерности стратегии и общего числа измерений $n = 5$.

Такую схему измерений называли 2+1+2, и, как было показано ранее, ее использование гарантирует выполнение условия $h_j < 0{,}5$.

Такая схема предполагает измерения в точках 1; 1; 2,5; 4; 4. В случае помехи измерения в одной из двух критических точек $x_{1,2}$ или $x_{4,5}$ уравнения линейной регрессии примут вид соответственно $0{,}1233+0{,}9667\cdot x_j$ и $-0{,}0440+1{,}033\cdot x_j$. То есть погрешность определения информативного параметра составляет менее 3,4 %.

В случае однократных измерений выполнение условий Хьюбера требует значительно большего числа измерений. Кроме того, при некоторых выбранных стратегиях оно не всегда выполнимо. Необходимое увеличение числа точек измерения при использовании различных стратегий иллюстрируется данными табл. 7.

Таким образом, техническая возможность повтора измерений в экстремальных точках x_j позволяет существенно сократить необходимое число измерений — до 5 измерений в 3 различных точках. Техническая или принципиальная невозможность повтора измерений требует существенно большего числа измерений, особенно при стратегиях, характеризующихся существенной кривизной зависимости *f(j)*.

Таблица 6. Влияние числа и расположения точек измерения на оценку параметров регрессионной модели (1)

Стратегия	t_j	x_j	h_j	Уравнение
B1	1,00	1,00	0,60	$0{,}1733+0{,}9467\cdot x_j$
	3,06	1,75	0,30	$0{,}1067+0{,}9733\cdot x_j$
	6,25	2,50	0,20	$0{,}0400+1{,}0000\cdot x_j$
	10,6	3,25	0,30	$-0{,}0267+1{,}0270\cdot x_j$
	16,0	4,00	0,60	$-0{,}0933+1{,}0530\cdot x_j$
B3	1,00	1,00	0,73	$0{,}2090+0{,}9379\cdot x_j$
	4,75	2,18	0,25	$0{,}0931+0{,}9805\cdot x_j$
	8,50	2,92	0,20	$0{,}0204+1{,}0070\cdot x_j$
	12,3	3,50	0,31	$-0{,}0366+1{,}0280\cdot x_j$
	16,0	4,00	0,50	$-0{,}0858+1{,}0460\cdot x_j$

Таблица 7. Необходимое число измерений для исключения точек разбалансировки

Стратегия *f(j)*		Способ измерения		
		Однократное	**Повтор в критической точке**	**Повторы в экстремальных точках**
j		7	6	5
j^2		9	7	5
$\sqrt{j}$		9	6	5
log(j)		14	6	5
α^{j-1}	α=1,1	11	7	5
	α=1,2	13	7	5
	α=1,4	50	6	5
	α=2,0	—	6	5

Наличие точек разбалансировки — измерений x_j с большими значениями h_j — приводит к резкому повышению ошибки оценивания параметров регрессионной модели (1), в особенности тангенса угла наклона зависимости — α_1. Подобный эффект крайне нежелателен, поскольку для большинства кинетических процессов именно α_1 является информативным параметром, определяемым концентрацией целевой компоненты. Применительно к калибровочным характеристикам подобные ошибки оценивания параметров модели приведут к ошибкам оценивания концентрации, особенно в случае экстраполяции заложенного регрессионного уравнения на более широкий диапазон концентраций.

Процедура выбора стратегии измерения, включающая в себя соотношение между измерениями *f(j)* и общее число измерений n, является процедурой выбора либо моментов измерения (в случае регистрации отклика хемосенсора), либо калибровочных концентраций.

Рекомендации, предложенные в данной работе, — приближенное дублирование точек измерения и увеличение общего числа измерений — при возможности их практической реализации позволят значительно увеличить точность оценивания концентрации как на стадии анализа кинетической кривой отклика прибора, так и при использовании калибровочной зависимости в аналитических задачах.

Работа выполнена при частичной поддержке ФЦП «Интеграция» в рамках проекта № А0141 «Оптика и научное приборостроение», раздел «Поддержка УНЦ “Приборы и средства автоматизации для научных исследований”» на базе СПбГУАП и ИАнП РАН.

ПРИЛОЖЕНИЕ

Т1. Неравенство $0 \le h_j \le 1$ было доказано Хьюбером [3] на основании свойств подгоночной матрицы **H** (2).

Т2. Доказательство tr(**H**) = 2.

Осуществляется на основе явного выражения (7) простым суммированием h_j по j от 1 до n.

Т3. Инвариантность к элементарным преобразованиям — изменению масштаба и сдвигу.

Для доказательства рассмотрим вспомогательные стратегии измерений Z_j и T_j, связанных с эталонной X_j.

а. Инвариантность к изменению масштаба.

Пусть стратегия Z_j связана с X_j зависимостью $Z_j = X_j\Delta$. При этом

$$S_1^* = \sum_{j=1}^{n} Z_j = \sum_{j=1}^{n} (X_j\Delta) = \Delta\sum_{j=1}^{n} X_j = \Delta S_1,$$

$$S_2^* = \sum_{j=1}^{n} Z_j^2 = \sum_{j=1}^{n} (X_j\Delta)^2 = \Delta^2\sum_{j=1}^{n} X_j^2 = \Delta^2 S_2.$$

Таким образом,

$$nS_2^* - (S_1^*)^2 = \Delta^2\left[nS_2 - S_1^2\right], \qquad (\text{П1})$$

$$S_2^* - 2S_1^* Z_j + nZ_j^2 = \Delta^2\left[S_2 - 2S_1 X_j + nX_j^2\right]. \qquad (\text{П2})$$

Сокращение множителя Δ^2 в (П1) и (П2) приведет к совпадению их частного с выражением (7). Тем самым, учитывая произвольность Δ, h_j при

стратегии Z_j полностью совпадает с h_j при эталонной стратегии X_j.

б. Инвариантность к сдвигу.

Пусть стратегия T_j связана с эталонной как $T_j = X_j + \Delta$.

При этом

$$S_1^* = \sum_{j=1}^{n} T_j = \sum_{j=1}^{n} (X_j + \Delta) = \Delta n + \sum_{j=1}^{n} X_j = \Delta n + S_1,$$

$$S_2^* = \sum_{j=1}^{n} T_j^2 = \sum_{j=1}^{n} (X_j + \Delta)^2 =$$

$$= \Delta^2 n + 2\Delta \sum_{j=1}^{n} X_j + \sum_{j=1}^{n} X_j^2 = \Delta^2 n + 2\Delta S_1 + S_2.$$

Таким образом,

$$nS_2^* - (S_1^*)^2 = nS_2 - S_1^2, \qquad (\text{П3})$$

$$S_2^* - 2S_1^* Z_j + nZ_j^2 = S_2 - 2S_1 X_j + nX_j^2. \qquad (\text{П4})$$

Частное выражений (П3) и (П4) также совпадает с (7). Тем самым, учитывая произвольность Δ, h_j при выбранной стратегии T_j полностью совпадает с h_j при стратегии X_j.

Т4. m-кратный повтор измерений в каждой точке x_j приводит к уменьшению всех h_j в m раз: $h_j^* = h_j/m$, $j = 1,...,n$.

В случае m-кратного дублирования S_k увеличиваются в m раз ($k = 0,1,2$). При этом Δ увеличится в m^2 раз.

Следствие 1. Двукратный повтор измерений гарантирует, по крайней мере нежесткое (в форме равенства), невыполнение неравенства (4).

Следствие 2. Для построения регрессионной модели (1) в случае кинетической кривой данный метод исключения точек разбалансировки неприменим из-за принципиальной неповторимости каждого измерения.

Т5. $h_j \ge 1/n$, $j = 1,..., n$.

Очевидно, $S_0 = n$.

Пусть $x^* = S_1 / S_0$ и $\delta_j = x_j - x^*$.

Тогда

$$S_0 S_2 - S_1^2 = S_0 \sum_{j=1}^{n} \delta_j^2,$$

$$S_2 - 2S_1 x_j + S_0 x_j^2 = S_0 \delta_j^2 + (S_0 S_2 - S_1^2)/S_0.$$

В результате

$$h_j = \frac{1}{S_0} + \frac{\delta_j^2}{\sum_{j=1}^{n} \delta_j^2} \ge \frac{1}{S_0}. \qquad (\text{П5})$$

Т6. $h_j^{\max} \ge 2/n$.

Очевидно, что произвольный

$$\delta_i^2 \le \max \delta_i^2 \Rightarrow \sum_{i=1}^{n} \delta_i^2 \le n \max \delta_i^2.$$

$$h_i^{\max} = \frac{1}{S_0} + \frac{\max \delta_i^2}{\sum_{i=1}^{n} \delta_i^2} \ge \frac{1}{S_0} + \frac{1}{S_0} = \frac{2}{S_0}.$$

Следствие 1. Для заведомого выполнения (5) или (6) в форме ($h_j < 0{,}5$) требуется $n \ge 5$.

Следствие 2. Максимальные значения могут принимать только h_1 или h_n — элементы матрицы разбалансировки, соответствующие измерениям в экстремальных точках. (Предполагается упорядоченность x_j по их возрастанию).

Т7. Удвоение измерения в одной из точек x_j приводит к выполнению условия $h_j \le 0{,}5$ в указанной точке.

Основные составляющие (7) после повторного измерения в точке x_j изменяются как

$$S_0 = S_0 + 1,\ S_1 = S_1 + x_j \text{ и } S_2 = S_2 + x_j^2.$$

Обозначив за Δ выражение $S_0 S_2 - S_1^2$, получим для новых элементов подгоночной матрицы h_i^* формулу

$$h_i^* = \frac{h_i + (x_i - x_j)^2 / \Delta}{1 + h_i}. \qquad (\text{П6})$$

Очевидно, что в случае $i = j$, то есть в точке удвоения измерения, $x_i = x_j$ и $h_j^* = h_j/(1+h_j)$.

Таким образом, $h_j^* < 0{,}5$ во всех случаях, за исключением $h_j = 1$.

Т8. Добавление измерения в новой точке z приводит к уменьшению h_j во всех точках измерения, за исключением, быть может, одной x_k при условии $x_k = x_j^{\circ}$, определяемой как

$$x_j^0 = x^* - \frac{S_0 S_2 - S_1^2}{S_0^2 (z - x^*)}.$$

По аналогии с (П6) добавление точки измерения $x_{n+1} = z$ приводит к матрице разбалансировки с элементами, определяемыми выражением

$$h_i^* = \frac{h_i + (x_i - z)^2 / \Delta}{1 + (S_2 - 2S_1 z + S_0 z^2)/\Delta}.$$

Неравенство $h_i^* \le h_i$ выполнится, если соблюдается

$$(x_i - z)^2 \le h_i (S_2 - 2S_1 z + S_0 z^2).$$

Представив h_i и $(x_i–z)$ в форме (П5), получим

$$h_i = 1/S_0 + S_0(x_i - x^*)^2/\Delta,$$
$$(S_2 - 2S_1 z + S_0 z^2)/\Delta = 1/S_0 + S_0(z - x^*)^2/\Delta.$$

Тогда требование $h_i^* \le h_i$ равносильно

$$\frac{(x_i - z)^2}{\Delta} \le \left(\frac{1}{S_0} + \frac{S_0(x_i - x^*)^2}{\Delta}\right)\left(\frac{1}{S_0} + \frac{S_0(z - x^*)^2}{\Delta}\right).$$

С учетом того, что

$$(x_i - z)^2 = \left[(x_i - x^*) - (z - x^*)\right]^2 =$$
$$(x_i - x^*)^2 + (z - x^*)^2 - 2(x_i - x^*)(z - x^*),$$

данное условие преобразуется к тривиальному утверждению

$$\left[\frac{1}{S_0} - \frac{S_0(x^* - x_i)(z - x^*)}{\Delta}\right]^2 \ge 0.$$

Таким образом, неравенство $h_i^* \le h_i$ выполняется для всех $i = 1,..., n$. При этом неравенство будет строгим для всех измерений, за исключением, быть может, x_k, при условии

$$S_0(x^* - x_k)(z - x^*)/\Delta = 1/S_0. \qquad (П7)$$

Это выражение совпадает с формулой для x_i^0. То есть, при дополнительном измерении в точке z останется неизменным элемент матрицы разбалансировки h_k, соответствующий измерению, проведенному в точке x_k, определяемой (П7). Все остальные h_i уменьшатся.

СПИСОК ЛИТЕРАТУРЫ

1. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. 899 с.
2. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. М.: Наука, 1976. 736 с.
3. *Хьюбер П.* Робастность в статистике. М.: Мир, 1984. 304 с.
4. *Тарасов В.В., Ягодин Г.А., Пичугин А.А.* Кинетика экстракции неорганических веществ // ВИНИТИ. Итоги науки и техники. Серия «Неорганическая химия». 1984. Т. 11. 187 с.
5. *Мусил Я., Новакова О., Кунц К.* Современная биохимия в схемах. М.: Мир, 1984. 216 с.
6. *Бурылов Д.А., Евстрапов А.А., Макарова Е.Д. и др.* Малогабаритный хемосенсорный анализатор // Журнал аналитической химии. 1997. Т. 52, № 5. С. 552–556.
7. *Kurochkin V.E., Makarova E.D.* Reflectance Spectrophotometry of Plasticized Membranes for the Design of fast Optical Chemosensors // Analytical Communication. 1996. V. 33, March. P. 115–116.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 21.02.99.

ESTIMATION OF THE NECESSARY NUMBER OF THE OBSERVATION POINTS UNDER CONSTRUCTION OF LINEAR REGRESSION MODELS

A. L. Bulianitsa, V. E. Kurochkin

Institute for Analytical Instrumentation RAS, Saint-Petersburg

The problem is being solved of determination of the number n of the observation points x_1, ..., x_n, which enables critical points excluding and provides the acceptable accuracy of regression model construction. As a basis, the methods of revealing critical points x_j (the unbalance and risk methods) proposed by Huber are used.