ISSN 0868–5886 *НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2003, том 13, № 2, с. 96–99*

ОРИГИНАЛЬНЫЕ СТАТЬИ

УДК 543.545: 519.25

© И. А. Леонтьев

ОБРАБОТКА ДАННЫХ В ЗАДАЧАХ ЭЛЕКТРОФОРЕЗА

Электрофоретические задачи требуют обработки данных. В работе рассматриваются методы сглаживания и поиска экстремумов для задач капиллярного электрофореза. Эти методы могут применяться и для многих других задач. В работе также рассмотрены методы для просчета некоторых параметров гауссовых пиков. Это может помочь в количественном анализе.

ВВЕДЕНИЕ

Информационные сигналы, полученные с детекторов прибора, как правило, подлежат первичной и вторичной обработкам. К первичной обработке относят методы оценивания информационных параметров, к вторичной — математическую обработку и анализ. Одной из распространенных задач является задача поиска экстремальных точек. Точкой локального экстремума называют такую точку x^*, для которой существует число $\delta > 0$ такое, что $F(x^*) < F(y)$ для всех $y \in N(x^*, \delta)$, $y \neq x^*$ [1] (для минимума) или $F(x^*) > F(y)$ (для максимума).

Однако в реальных физических экспериментах, где всегда имеются шумы, такое определение вряд ли подходит, т. к. большинство точек являются экстремумом (максимумом или минимумом). Именно поэтому необходимо применять процедуры сглаживания и фильтрации при первичной обработке. Эти методы могут применяться для широкого круга задач.

В работе рассматриваются методы поиска экстремумов для задач капиллярного электрофореза. Экстремумами в таких задачах являются пики, а анализ состоит в обнаружении и расчете их параметров, в том числе временны́х. Пиком в таких задачах можно назвать набор последовательных точек, значения которых значительно превышают значения точек вне этого набора. Анализ данных с помощью программного обеспечения, поставляемого с приборами капиллярного электрофореза, часто базируется на уже существующем и доказавшем свою пригодность алгоритме для высокопроизводительной жидкостной хроматографии [2].

ОПИСАНИЕ АЛГОРИТМА ПОИСКА ЭКСТРЕМУМОВ

В этом предлагаемом алгоритме определение пиков основано на изменениях первой производной сигнала. Пик считается обнаруженным, когда первая производная превысит некое пороговое значение, которое нужно установить в соответствии с заданными критериями (рис. 1). Действительное время начала пика определяется точкой, где первая производная равна нулю. Конец пика определяется аналогично. Достоверность этого алгоритма зависит от правильного выбора порогового значения производной и реальной ширины самого пика [2].

Как правило, в электрофорезе пик имеет Гауссову форму, т. е. $f(x) = a_1 \cdot e^{-\frac{(x-x_0)^2}{2\sigma^2}}$. Максимум первой производной достигается в точке

$$x = x_0 - \sigma. \qquad (1)$$

Максимальное значение производной в этой точке будет $\dot{f}_{\max} = a_1 \cdot \sqrt{e} / \sigma$. Следовательно, пороговое значение первой производной выбирается таким, чтобы оно было меньше, чем $a_1 \cdot \sqrt{e} / \sigma$ для всех пиков.

Как показывает практика, в капиллярном электрофорезе часто возникают ситуации, когда появляются небольшие отрицательные шумовые выбросы сигнала в окрестности точек, где первая производная превышает пороговое значение (рис. 2). Эту точку легко принять за конец пика, т. к. первая производная меняет знак и, следовательно, проходит через нуль. В такой ситуации центр пика будет определен заведомо неправильно. Для устранения этого недостатка необходимо применять предварительное сглаживание данных.

Для сглаживания данных существует множество методов. Здесь будет рассмотрен довольно простой метод Савитского—Голая. Метод применяется для серии данных $f_i = f(t_i)$, где $t_i = t_0 + i \cdot h$ и i — целое число. Каждое значение данных f_i заменяется линейной комбинацией g_i самого значения и значений нескольких соседей, т. е.

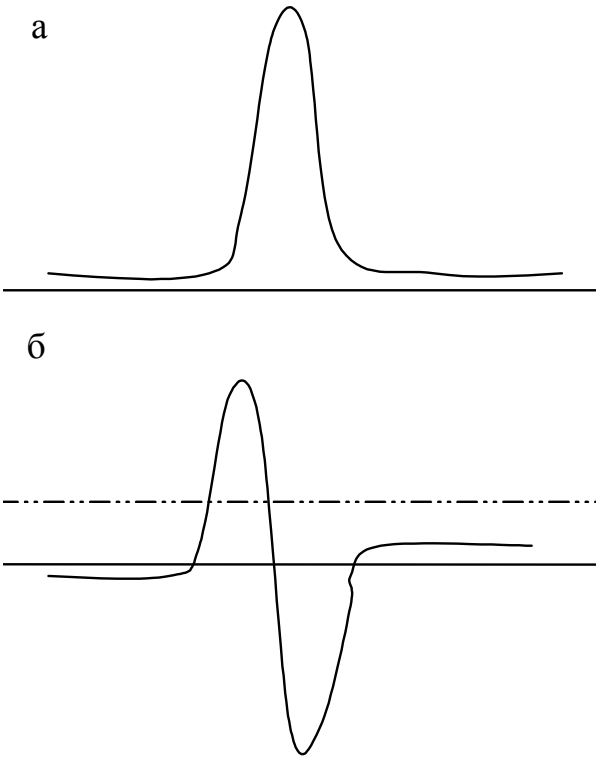

Рис. 1. Выбор порогового значения.
а — сигнал-пик; б — производная от сигнала; штрих пунктирной линией обозначен порог обнаружения пика

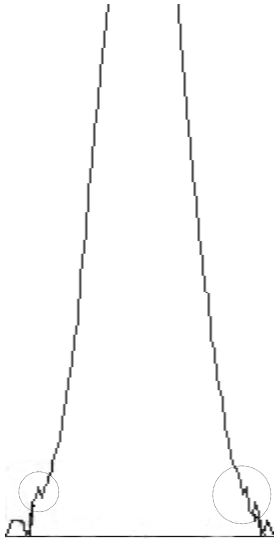

Рис. 2. Шумовые выбросы на фронтах сигнала

$$g_i = \sum_{n=-n_l}^{n_r} c_n f_{i+n}\,, \qquad (2)$$

где n_l — число точек слева от точки i, а n_r — справа. В простейшем случае — это усреднение значения, если $c_n = 1/(n_l + n_r + 1)$. В случае, когда подвергаемая сглаживанию информационная зависимость является линейной, метод не вносит искажения. Что касается величин локальных максимумов, то метод всегда уменьшает максимальное значение функции. Наиболее подвержены искажению узкие пики, которые менее всего походят на линейную функцию в пределах соседних точек. Можно показать, что такие экстремумы после применения сглаживания становятся не только меньше, но еще и незначительно увеличиваются по ширине.

Если пренебречь увеличением ширины пика, то критерием для выбора числа соседей n в данном методе сглаживания будет $g^* \geq k \cdot f(x_0)$, где g^*— значение для точки максимума, вычисленное из (2), k — величина допустимого уменьшения экстремума и $f(x_0)$ — реальное значение экстремума.

Главным недостатком алгоритма поиска экстремумов высокопроизводительной жидкостной хроматографии является его смещенность. В алгоритме находится не точка экстремума, а точка, где первая производная превышает пороговое значение. Смещение будет всегда больше, чем σ, т. к. максимум первой производной достигается в точке $x_0 - \sigma$ согласно (1). Таким образом, метод нуждается в процедуре уточнения центра экстремума.

Если форма пика известна, то наиболее эффективным методом нахождения его центра и прочих параметров является метод наименьших квадратов. Метод обладает свойством оптимальности, состоящим в том, что он дает несмещенные оценки, имеющие минимальную дисперсию [3].

ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ ПИКА

Если $f(t) = x_1 \cdot e^{x_2 (t - x_3)^2}$ и имеется набор измерений f_i, то задача сводится к поиску

$$\min_{x_1, x_2, x_3} F(x_1, x_2, x_3) = \sum_{i=1}^{m} (f_i - x_1 \cdot e^{x_2 (t_i - x_3)^2})^2 . \qquad (3)$$

Эта задача носит название нелинейной задачи наименьших квадратов [4]. Задача оптимизации состоит в минимизации $F(x)$.

Нам необходимо найти такую точку x^*, что $F(x^*) \leq F(x)$ для всех допустимых точек x, которые близки к x^*. Такая точка называется локальным минимумом.

Алгоритмы для минимизации функции n переменных разрабатываются уже свыше 140 лет.

Наиболее распространенными являются метод наискорейшего спуска и многомерный метод Ньютона.

Метод наискорейшего спуска был предложен Коши в 1845 г. Суть метода состоит в следующем. Обозначим вектор $(x_1,\ldots, x_n)^T$ через $\mathbf{x}$ и предположим, что функция $F(\mathbf{x})$ имеет непрерывные частные производные нескольких порядков. Для фиксированного $\mathbf{x}$ и меняющегося α совокупность векторов $(\mathbf{x}, -\alpha\nabla F)$ представляет собой луч, исходящий из точки $\mathbf{x}$. Известно, что $-\nabla F(\mathbf{x})$ — это направление "с холма" для функции $F(\mathbf{x})$ [4], т. е. для достаточно малого положительного значения α значения функции будут убывать: $F(\mathbf{x} - \alpha\nabla F) < F(\mathbf{x})$ [4]. После этого ищется значение α $(0 < \alpha < \infty)$, минимизирующее F в направлении $-\nabla F(\mathbf{x})$. Это уже одномерная минимизация. Найдя этот минимум, начинают поиск вдоль полупрямой наискорейшего спуска, исходящей из новой точки $\mathbf{x}$ [4].

Достоинство метода наискорейшего спуска заключается в том, что он всегда сходится, если функция $F(\mathbf{x})$ имеет непрерывные производные. Но в некоторых случаях он сходится довольно медленно (иногда требуется более ста итераций).

Этого можно избежать, применяя n-мерный аналог метода Ньютона. Функция F записывается в виде

$$F(\mathbf{x}) \approx F(\mathbf{x}) + \mathbf{p}^T\nabla F(\mathbf{x}) + \\ + \frac{1}{2}\mathbf{p}^T\nabla^2 F(\mathbf{x})\mathbf{p} = F(\mathbf{x}) + Q(\mathbf{p}). \quad (4)$$

Чтобы получить шаг p, минимизируется квадратичная функция $Q(\mathbf{p})$, при помощи построения ее градиента по p

$$\nabla_p Q(\mathbf{p}) = \\ = \nabla_p(\mathbf{p}^T\nabla F(\mathbf{x}) + \frac{1}{2}\mathbf{p}^T\nabla^2 F(\mathbf{x})\mathbf{p}) = \\ = F(\mathbf{x}) + \nabla^2 F(\mathbf{x})\mathbf{p}. \quad (5)$$

Приравнивая его к нулю, получаем

$$\nabla^2 F(\mathbf{x})\mathbf{p} = -\nabla F(\mathbf{x}). \quad (6)$$

Это система n линейных уравнений относительно n неизвестных $\mathbf{p} = (p_1, p_2,\ldots, p_n)^T$.

Итак, $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p} = \mathbf{x}_k - \nabla^2 F(\mathbf{x}_k)^{-1}\nabla F(\mathbf{x}_k)$.

Метод Ньютона в n-мерном случае обладает тем же свойством быстрой сходимости, что и в одномерном случае, а именно он сходится квадратично в окрестности решения:

$$\left\|\mathbf{x}_{k+1} - \mathbf{x}^*\right\|_n \le \beta\left\|\mathbf{x}_k - \mathbf{x}^*\right\|_n^2, \quad (7)$$

где β — некоторая неотрицательная константа, зависящая от $F(\mathbf{x})$.

Но, как и в одномерном случае, у метода Ньютона есть недостатки. Например, он может не сходиться. Но если матрица $\nabla^2 F^{-1}$ положительно определена, т. е. удовлетворяет условию $\mathbf{z}^T\nabla^2 F^{-1}\mathbf{z}$ для всех $\mathbf{z} \ne 0$, то в этом случае направление метода Ньютона гарантированно будет направлением спуска [4].

Недостатком метода Ньютона является необходимость вычислять матрицу вторых производных. Другой недостаток заключается в том, что вычисление шага $\mathbf{p}$ требует решения системы n линейных уравнений. Метод наискорейшего спуска не страдает этими недостатками, но он приводит к большим потерям в скорости сходимости [4].

Для того чтобы получить альтернативные эффективные и практичные методы, можно аппроксимировать гессиан в ходе минимизации функции. Эти методы основаны на аппроксимации гессиана секущими и являются обобщением метода секущих. Если $B_k \approx \nabla^2 F_k = F(\mathbf{x})$, то шаг на $k^{\text{-й}}$ итерации будет определяться из системы

$$B_k\cdot\mathbf{p} = -\nabla F_k. \quad (8)$$

Таким образом, мы получаем вариант метода Ньютона, в котором используется приближенный гессиан. Новая аппроксимация гессиана выбирается так, чтобы

$$\mathbf{B}_{k+1}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \nabla F_{k+1} - \nabla F_k. \quad (9)$$

В одномерном случае это однозначно определяет $\mathbf{B}_{k+1}$. При более высоких размерностях, для того чтобы определить $\mathbf{B}_{k+1}$, необходимы дополнительные условия.

Поскольку нам необходимо на каждой итерации вычислять обратную матрицу $\mathbf{B}^{-1}$, можно еще упростить метод Ньютона, используя алгоритм Шермана—Моррисона для приближенного вычисления обратной матрицы. Предположим, что для матрицы $\mathbf{B}_k$ вычислена обратная $\mathbf{B}_k^{-1}$. Пусть на следующей итерации матрица

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \mathbf{u}\cdot\mathbf{v}^T, \quad (10)$$

где $\mathbf{u}$ и $\mathbf{v}$ — векторы размерности n. Тогда матрицу $\mathbf{B}_{k+1}^{-1}$ можно вычислять по формуле

$$\mathbf{B}_{k+1}^{-1} = \mathbf{B}_k^{-1} + \boldsymbol{\alpha}(\mathbf{B}_k^{-1}\mathbf{u})(\mathbf{v}^T\mathbf{B}_k^{-1}), \quad (11)$$

где $\boldsymbol{\alpha} = 1/(1 - \mathbf{v}^T\mathbf{B}_k^{-1}\mathbf{u})$. Это будет стоить $O(n^2)$ арифметических операций по сравнению с $O(n^3)$ операций при стандартном вычислении новой обратной матрицы.

ЗАКЛЮЧЕНИЕ

В статье описан метод для обработки информационных данных в капиллярном электрофорезе. Метод позволяет обнаружить и определить параметры пиков. Для определения параметров пиков использовался аналог метода Ньютона, в котором не требуется ни вычисления вторых производных, ни решения систем уравнений.

СПИСОК ЛИТЕРАТУРЫ

1. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация. М.: Мир, 1985. 509 с.
2. *Wanders B. J.* Data analysis in capillary electrophoresis // Handbook of Capillary Electrophoresis. CRC Press Inc., 1997. P. 449–450.
3. *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. 110 с.
4. *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. М.: Мир, 2001. 500 с.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 16.04.2003.

DATA ANALYSIS IN CAPILLARY ELECTROPHORESIS APPLICATIONS

I. A. Leontyev

Institute for Analytical Instrumentation RAS, Saint-Petersburg

The electrophoretic problems involve data processing. Methods of smoothing and search of extrema for problems of capillary electrophoresis are described. These methods can be applied to many other problems. In this work a technique for calculation of some characteristics of Gaussian peaks is also considered. This may help in quantitative analysis.