——— МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ ———— В ПРИБОРОСТРОЕНИИ

УДК 543.426; 543.9

© Л. В. Новиков, В. В. Куркина

МЕТОД ОЦЕНКИ ПАРАМЕТРОВ СПЕКТРАЛЬНЫХ ПИКОВ

Предлагается новый экономичный алгоритм оценки параметров сигналов в масс-спектрометрии, хроматографии и др. приложениях, представляющих собой последовательность пиков на фоне шумов. Развивается традиционный подход обнаружения вершины пиков по пересечению первой производной нулевой линии. С целью повышения надежности обнаружения начала, конца, вершины и седловины между пиками, предлагается сравнивать величины производных в трех точках скользящего окна данных.

Кл. сл.: обработка сигналов, аналитические спектры, оценка параметров, идентификация пиков

введение

Оценка параметров сигналов, содержащих информацию о качественном и количественном составе анализируемых веществ, было и остается важной задачей в обработке данных приборов многокомпонентного анализа в аналитической химии, биологии и медицине. При этом требования к качеству обнаружения и особенно оценке параметров сигналов в присутствии шума, базовой линии и в условиях неполного разделения пиков постоянно растут, т. к. они определяют стедостоверности решений, принимаемых пень по результатам анализа. В частности, в приборах генетического анализа при расшифровке последовательности нуклеотидов необходимо определить кроме традиционных параметров — время выхода и амплитуда пиков — также ширину, площадь и признак полного или частичного разделения пиков. Все эти параметры нужны для оценки показателей качества расшифровки последовательности нуклеотидов в цепочке по методике Phred [1].

Традиционные методы анализа данных аналитических приборов, как правило, включают предварительную обработку с целью минимизировать влияние на качество оценки информативных параметров сигнала случайного и детерминированного шумов, выбросов, базовой линии и эффекта неполного разделения пиков. В работе [2] рассматривается несколько программных пакетов предварительной обработки, применяемых, в частности, в исследованиях метаболомики на основе хроматографии и масс-спектрометрии.

Пакет XCMS [3], выполняет операции фильтрации, идентификации, определения площади и выравнивания времени удерживания пиков. В алгоритме обнаружения пиков используется метод согласованной фильтрации с применением в качестве фильтра второй производной гауссовой модели формы пика. Величина порога выбирается по уровню отношения сигнал/шум. Ширина пика определяется по точкам пересечения нуля отфильтрованных данных в зоне пика.

В пакете MetSign [4] обнаружение пиков производится по пересечению нулевой линии первой и второй производной хроматограммы. На основе модели в виде экспоненциально-модифицированной гауссианы выполняется уточнение параметров пиков методом подгонки кривых.

Программные пакеты MetAlign [5] и MZmine [6] выполняют коррекцию базовой линии, фильтрацию шума и артефактов в спектре масс, детектирование пиков, нормализацию, формирование таблиц пиков, визуализацию результатов.

В статье [7] авторы используют популярный фильтр Савицкого—Голея для вычисления производных первых трех порядков в скользящем окне данных. По первой производной определяют зоны пиков по шкале времени хроматограммы. Если величина отрицательной области второй производной превышает порог, пик считается обнаруженным. По числу пересечений нулевой линии третьей производной в зоне пика судят о наличии или отсутствии неразделенных соседних пиков. Значения времени удерживания, амплитуды и ширины пика определяются по данным, полученным из второй производной сигнала.

В работе [8] авторы используют вторую производную гауссовой кривой, взятую с обратным знаком ("мексиканская шляпа"), в качестве ядра оператора для вычисления сглаженной второй производной хроматографического сигнала. Обнаружение пика производится по превышению второй производной порога, равному 10-кратной интенсивности шума. Интенсивность пиков определяется в точках максимума по интенсивности нормированной второй производной храматограммы, время удерживания — по положению максимумов на временной шкале и, наконец, ширина пиков по разности между временами окончания и начала пиков на уровне, равном 3-кратной интенсивности шума.

В работах [9, 10] используется функция "мексиканская шляпа" в качестве материнского вейвлета для непрерывного вейвлет-преобразования (НВП) данных. Выполняя НВП при различных величинах масштаба материнского вейвлета, получают двумерный спектр на плоскости время/масштаб, содержащий на каждом масштабе (шкале) пики, порожденные исходными данными. Максимумы пиков, положения которых близки, формируют хребты, объединяющие эти максимумы по всем шкалам. Пик считается обнаруженным, если сумма его максимальных значений по шкалам (по хребту) превышает заданный порог. Положение пика определяется по отсчету при минимальном значении масштаба. Благодаря применению НВП в широком диапазоне масштабов метод позволяет разделить близкие пики без заметной седловины между ними. Точные значения параметров пиков определяют методом подгонки кривых. Применение "мексиканской шляпы" и вообще второй производной данных позволяет, кроме того, удалить базовую линию, линейную в пределах окна дифференцирующей функции.

Алгоритм детектирования пиков в тандеме (газовый хроматограф—спектрометр ионной мобильности) (GC-IMS) [11] основан на вычислении первой и второй частных производных двумерных данных. Величины интенсивности, гауссовой кривизны и матрицы Гессе в каждой точке двумерной поверхности сравниваются с соответствующими порогами, по результатам чего принимается решение о наличии пиков.

Таким образом, существующие методы оценки параметров пиков в той или иной степени используют дифференцирование исходных (сырых) данных. В настоящей работе предлагается алгоритм, в котором с целью повышения надежности обнаружения особенностей сигналов, имеющих вид пика на фоне шумов, используется метод вычисления производных не в одной, а в трех точках скользящего окна данных для принятия решения о наличии пика и оценки его параметров.

ТЕОРИЯ

Данные аналитических приборов y(t) многокомпонентного анализа веществ можно представить в виде аддитивной смеси K пиков $\{x_k(t), k = 1, ..., K\}$, например, гауссовой формы



Рис. 1. Модель аналитического сигнала, состоящий из ста пиков гауссовой формы при отношении сигнал/шум 2.5.

Ширина пиков линейно возрастает с увеличением і

и шума n(t) (предполагается, что базовая линия удалена):

$$y(t) = \sum_{k=1}^{K} x_k(t) + n(t),$$

где $x_k(t) = A_k \exp\left\{-(t-t_k)^2/\mu_k^2\right\}$; t — непрерывное или дискретное время $t = i \, \delta t$ (i = 1, 2, ..., N), N — число отсчетов данных, δt — интервал дискретизации; A_k, t_k, μ_k — интенсивность, положение и среднеквадратическая ширина k-го пика. Пример такого аналитического спектра, состоящий из ста пиков гауссовой формы с добавлением белого шума, приведен на рис. 1.

Массив дискретных данных $y(i\delta t) = y(i)$ образует *N*-мерный вектор

$$\mathbf{y} = (y(1), y(2), ..., y(N)).$$

Сформируем текущую выборку из M+1 отсчетов с центром в точке *i* исходных данных:

$$\mathbf{y}_{w} = (y_{w}(1), y_{w}(2), ..., y_{w}(M/2), ..., y_{w}(M+1)) = = (y(i - M/2), ..., y(i), ..., y(i + M/2)),$$

где M выбирается всегда четным из условия $0.5 \div 0.7$ от ширины пика на половине высоты в отсчетах. При этом во избежание скачков на границах массива **y**, полагаем, что если $i - M/2 \le 0$, то $y_w(i - M/2) = y(1)$, и если i + M/2 > N, то $y_w(i + M/2) = y(N)$.



Рис. 2. Аппроксимация данных полиномом второго порядка в скользящем окне из 15 точек (M = 14) на сетке (j = -7, ..., j = 0, ..., j = 7). t_l, t_0, t_r — точки аппроксимирующей параболы, в которых вычисляются первые производные

Аппроксимируем отсчеты в окне y_w полиномом второй степени (параболой) на сетке j = (-M/2,...,0,...,M/2):

$$\tilde{y}(j)=p_1j^2+p_2j+p_3,$$

где *p*₁, *p*₂, *p*₃ — коэффициенты полинома.

Выберем в интервале
$$\left[-\frac{M}{2}, \frac{M}{2}\right]$$
 три точки t_l, t_0, t_r $(l - \text{left}, r - \text{right})$ такие, что $t_l = j = -\left[\frac{M}{4}\right], t_0 = j = 0, t_r = j = \left[\frac{M}{4}\right],$ где $\left[\frac{M}{4}\right]$ -

целое число, большее или равное M/4. На рис. 2 показан фрагмент спектра с окном данных из пятнадцати отсчетов (M = 14) и аппроксимирующая парабола с точками t_l , t_0 и t_r .

Величина первой производной параболы в точке t_l будет равна

$$d_l = 2p_1 t_l + p_2 \tag{1}$$

и в точке t_r :

$$d_r = 2p_1 t_r + p_2 \,. \tag{2}$$

Величина второй производной параболы равна *p*₁.

Величина параболы в точке j = 0 равна усредненному значению в скользящей выборке исходных данных в точке і

$$y(i) = \tilde{y}(0) = p_3$$
. (3)

ОПИСАНИЕ АЛГОРИТМА

Используя формулы (1), (2), (3) рассмотрим алгоритм обнаружения пиков и их параметров. В начале обработки исходных данных по этим формулам в программе устанавливаются признаки обнаружения начала и вершины пика:

$$PSS = \begin{cases} 0, & \text{вершина пика не обнаружена,} \\ 1, & \text{вершина пика обнаружена.} \end{cases}$$

Условие начала пика имеет вид

$$d_l > 0; \ d_r > th_{slope}; \ p_3 > th_{min} \ \text{M} \ PS = 0,$$
 (4)

где th_{\min} — порог нижнего уровня, выбираемого по стандартному отклонению шума σ , $th_{\min} = (1 \div 2)\sigma$, th_{slope} — порог по наклону фронта пика, для гауссовой формы пика выбирается из интервала $th_{\text{slope}} = 0.003 \div 0.007$.

После обнаружения начала пика устанавливаются признаки PS = 1 и PSS = 0.



Рис. 3. Фрагмент спектра с аппроксимирующей параболой (утолщенная линия) в начале пика. th_{\min} — порог нижнего уровня; ts_k — начало пика



Рис. 4. Фрагмент спектра с аппроксимирующей параболой (утолщенная линия) на вершине пика.

 th_{\max} — порог максимального уровня; $t_{\max k}$ — положение вершины k-го пика

Начало *k*-го пика определяется по положению точки t_i на шкале отсчетов $i: ts_k = \arg(t_i)_i$ при условии PS = 1 (см. рис. 3).

Условие вершины пика имеет вид

 $PS = 1; d_1 > 0; d_r < 0; p_1 < 0$ и $max(\tilde{y}) > th_{max}$, (5) где th_{max} — порог максимального уровня, выбираемый из условия

$$th_{\max} = (4 \div 6)\sigma$$
.

Оценка амплитуды \hat{A}_k и положение вершины t_{\max_k} *k* -го пика определяются по точке максимума i_{\max} параболы:

$$\hat{A}_{k} = \max\left(\tilde{y}\right), \ t_{\max_{k}} = \arg\left(\max\left(\tilde{y}\right)\right)_{i} \tag{6}$$

по шкале отсчетов і (см. рис. 4).

$$\operatorname{abs}\left(t_{\max_{k}}-i_{0}\right)<3.$$

После обнаружения вершины пика устанавливаются признаки PS = 0 и PSS = 1.

Оценку ширины пика на половине высоты находим из условия равенства участка гауссовой модели в окрестности вершины пика и аппроксимирующей ее параболы (полагаем переменную *t* непрерывной):

$$\hat{A}_k \exp\left\{-\left(\frac{t-t_{\max_k}}{\mu_k}\right)^2\right\} = p_1 t^2 + p_2 t + p_3 = \tilde{y}(t).$$

В точке t_l параболы имеем

$$\hat{A}_k \exp\left\{-\left(\frac{t_l - t_{\max_k}}{\mu_k}\right)^2\right\} = \tilde{y}(t_l).$$

Прологарифмировав обе части равенства и выполнив преобразования, получим

$$\mu_k = \left| t_l - t_{\max_k} \right| \ln \left(\frac{\hat{A}_k}{\tilde{y}(t_l)} \right).$$

Ширина пика на половине высоты будет равна

$$\mu_{1/2_k} = \sqrt{2 \ln 2} \,\mu_k \,. \tag{7}$$

Площадь *k* -го пика гауссовой формы определяется по формуле

$$S_k = \sqrt{\pi} \, \hat{A}_k \, \mu_k \tag{8}$$

Условия конца пика имеют следующий вид: если

$$PSS = 1; \ d_l < 0; \ d_r < 0; \ p_3 < th_{\min}, \qquad (9)$$

то обнаружен одиночный (хорошо разделенный) пик.

При этом конец *k*-го пика определяется по положению точки t_k на шкале отсчетов *i*: $tf_k = \arg(t_k)_i$; устанавливается признак: пики разделены $R_k = 1$ (при условии $R_{k-1} \neq 0$) и признак PSS = 0.

Если условие (9) не выполняется, но выполняется условие

 $PSS = 1; d_l < 0; d_r > th_{slope}; p_3 > th_{min}$, (10) то обнаружена седловина между соседними пиками.

При этом конец *k*-го пика tf_k определяется по положению точки минимума параболы в седловине между пиками $t_{sad} = \arg(\min(\tilde{y}))_i$ по шкале отсчетов $i: tf_k = \arg(t_{sad})_i$ (см. рис. 5).



Рис. 5. Фрагмент спектра с аппроксимирующей параболой (утолщенная линия) на седловине пика.

 $t_{\rm sad}$ — положение седловины между пиками; tf_k — конец пика

Если при этом выполняется условие, например,

$$\tilde{y}(t_{\mathrm{sad}}) >= \frac{2}{3}\tilde{A}_k$$
,

т. е. если глубина седловины не превышает одной трети интенсивности пика, то можно считать, что соседние пики не разделены и устанавливается признак $R_k = 0$. В противном случае устанавливается признак $R_k = 1$.

Таким образом, в результате выполнения алгоритма детектирования последовательности пиков по приведенному выше алгоритму имеем оценки следующих параметров каждого пика:

- начало ts_k ,
- положение максимума $t_{\max k}$,
- интенсивность \hat{A}_k ,
- -площадь \hat{S}_{k} ,
- -конец tf_k ,
- признак разделения R_k .

моделирование

Проверка эффективности алгоритма производилась по модельному спектру, состоящему из 100 пиков гауссовой формы с добавлением белого шума (рис. 1) путем сравнения результатов обработки с алгоритмом, приведенным в работе [12]. Относительная погрешность оценки амплитуды, ширины и площади пиков производилась по десяти реализациям шума путем вычисления СКО параметров каждого пика и деления его на истинное значение параметра. Погрешность положения пика определялась по его отклонению от истинного значения. Максимальные значения погрешностей для обоих методов обработки сведены в таблицу. Кроме того, вычислялась вероятность правильного обнаружения как отношение числа обнаруженных пиков по всем десяти реализациям, положение которых отличается от заданных в модели не более, чем на четыре отсчета, к общему числу модельных пиков. Вероятность ложного обнаружения вычислялась как отношение числа обнаруженных пиков по всем десяти реализациям, не заданных в модели, к общему числу модельных пиков. Сравнение столбцов 1 и 2 таблицы показывает явное преимущество предлагаемого алгоритма как по уровню погрешности оценки параметров пиков, так и по величине вероятности ложного обнаружения.

Абсолютные	(для і	положе	ния	пиков)	ИО	тносительные	среднеква;	дратич	еские о	шибки	(CKO)
оценки парам	етров	пиков	при	различ	ных	отношениях	сигнал/шум	1 для	метода,	изложе	енного
в [12], — 1 и п	предла	гаемого	э мет	года — 2	2						

Параметр	Отношение сигнал/шум									
	2.	.5	5		10		20			
	Метод расчета									
	1	2	1	2	1	2	1	2		
Положение пика (в отсчетах)	3.98	3	1.68	2	0.75	1	0.38	0		
Амплитуда	0.08	0.04	0.038	0.018	0.21	0.01	0.01	0.004		
Ширина	0.90	0.34	0.46	0.23	0.19	0.13	0.08	0.047		
Площадь	1.01	0.27	0.43	0.19	0.18	0.11	0.08	0.030		
Вероятность пра- вильного обнаруже- ния	0.980	0.999	1	1	1	1	1	1		
Вероятность ложно- го обнаружения	0.016	0.002	0.003	0.001	0.002	0	0.002	0		

ЗАКЛЮЧЕНИЕ

Алгоритм детектирования предложенным методом позволяет повысить достоверность обнаружения пиков благодаря низкому уровню ошибки оценок и вероятности ложного обнаружения при малых отношениях сигнал/шум. Это позволяет значительно понизить порог чувствительности приборов и расширить диапазон измеряемых концентраций.

СПИСОК ЛИТЕРАТУРЫ

- Ewing B., Green Ph. Base-calling of automated sequencer traces using Phred. II. Error probabilities // Genome Res. 1998. Vol. 8, no. 3. P. 186–194. Doi: 10.1101/gr.8.3.186.
- Cook D.W., Rutan S.C. Chemometrics for the analysis of chromatographic data in metabolomics investigations // J. Chemometrics. 2014. Vol. 28, no. 9. P. 681–687.
- Smith C.A., Want E.J., O'Maille G., Abagyan R., Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification // Anal. Chem. 2006. Vol. 78, no. 3. P.779–787. Doi: 10.1021/ac051437y.
- Wei X., Shi X., Kim S., Zhang L., Patrick J.S., Binkley J., McClain C., Zhang X. Data preprocessing method for liquid chromatography-mass spectrometry based metabolomics // Anal. Chem. 2012. Vol. 84, no. 18. P. 7963– 7971. Doi: 10.1021/ac3016856.
- Lommen A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing // Anal. Chem. 2009. Vol. 81, no. 8. P. 3079–3086. Doi: 10.1021/ac900036d.
- Katajamaa M., Oresic M. Processing methods for differential analysis of LC/MS profile data // BMC Bioinformatics. 2005. Vol. 6. P. 179–190. Doi: 10.1186/1471-2105-6-179.
- Vivó-Truyols G., Torres-Lapasió J.R., van Nederkassel A.M., Heyden Y.V., Massart D.L. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals. Part I: Peak detection // Journal of Chromatography A. 2005. Vol. 1096, no. 1-2. P. 133– 145. Doi: 10.1016/j.chroma.2005.03.092.
- Fredriksson M.J., Petersson P., Axelsson B.-O., Bylund D. An automatic peak finding method for LC-MS data using Gaussian second derivative filtering // J. Sep. Sci. 2009.

Vol. 32, no. 22. P. 3906–3918. Doi: 10.1002/jssc.200900395.

- Gregoire J.M., Dale D., van Dover B. A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data // Review of Scientific Instruments. 2011. Vol. 82, no. 1. 015105. Doi: 10.1063/1.3505103.
- Du P., Kibbe W.A., Lin S.M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching // Bioinformatics. 2006. Vol. 22, no. 17. P. 2059–2065. Doi: 10.1093/bioinformatics/btl355.
- Slodzinski R., Hildebrand L., Vautz W. Peak detection algorithm based on second derivative properties for two dimensional ion mobility spectrometry signals // Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives / Madjid Fathi (ed.). Springer-Verlag, Berlin, Heidelberg, 2013. P. 341–354.
- 12. *O'Haver T*. Interactive Signal Processing Tools. Peak Finding and Measurement. URL: http://terpconnect.umd.edu/~toh/spectrum/SignalProcessi ngTools.html.

Институт аналитического приборостроения РАН, г. Санкт-Петербург (Новиков Л.В.)

Санкт-Петербургский государственный технологический институт (Технический университет) (Куркина В.В.)

Контакты: *Новиков Лев Васильевич*, novik38@mail.ru

Материал поступил в редакцию: 20.06.2017

THE METHOD FOR ESTIMATION OF SPECTRAL PEAK PARAMETERS

L. V. Novikov¹, V. V. Kurkina²

¹Institute for Analytical Instrumentation of RAS, Saint-Petersburg, Russia ²St. Petersburg State Technological Institute (Technical University), Russia

A new economical algorithm for estimating the parameters of signals in mass spectrometry, chromatography, and other applications, representing a sequence of peaks against a noise background, is proposed. A traditional approach for the peak top detection by the intersection the zero line by the first derivative is developing. In order to increase the reliability of detection of the beginning, end, vertex and saddle between peaks, it is proposed to compare the values of the derivatives at three points of the sliding data window.

Keywords: signal processing, analytical spectra, estimation of parameters, identification of peaks

REFERENCES

- Ewing B., Green Ph. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.*, 1998, vol. 8, no. 3, pp. 186–194. Doi: 10.1101/gr.8.3.186.
- Cook D.W., Rutan S.C. Chemometrics for the analysis of chromatographic data in metabolomics investigations. J. Chemometrics, 2014, vol. 28, no. 9, pp. 681–687.
- Smith C.A., Want E.J., O'Maille G., Abagyan R., Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, 2006, vol. 78, no. 3, pp.779–787. Doi: 10.1021/ac051437y.
- Wei X., Shi X., Kim S., Zhang L., Patrick J.S., Binkley J., McClain C., Zhang X. Data preprocessing method for liquid chromatography-mass spectrometry based metabolomics. *Anal. Chem.*, 2012, vol. 84, no. 18, pp. 7963– 7971. Doi: 10.1021/ac3016856.
- Lommen A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.*, 2009, vol. 81, no. 8, pp. 3079–3086. Doi: 10.1021/ac900036d.
- Katajamaa M., Oresic M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, 2005, vol. 6. pp. 179–190. Doi: 10.1186/1471-2105-6-179.
- 7. Vivó-Truyols G., Torres-Lapasió J.R., van Nederkassel A.M., Heyden Y.V., Massart D.L. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals. Part I: Peak detection. *Journal*

Contacts: *Novikov Lev Vasil'evich*, novik38@mail.ru

of Chromatography A, 2005. vol. 1096, no. 1-2, pp. 133–145. Doi: 10.1016/j.chroma.2005.03.092.

- Fredriksson M.J., Petersson P., Axelsson B.-O., Bylund D. An automatic peak finding method for LC-MS data using Gaussian second derivative filtering. *J. Sep. Sci.*, 2009, vol. 32, no. 22, pp. 3906–3918. Doi: 10.1002/jssc.200900395.
- Gregoire J.M., Dale D., van Dover B. A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data. *Review of Scientific Instruments*, 2011, vol. 82, no. 1, 015105. Doi: 10.1063/1.3505103.
- Du P., Kibbe W.A., Lin S.M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 2006, vol. 22, no. 17, pp. 2059–2065. Doi: 10.1093/bioinformatics/btl355.
- Slodzinski R., Hildebrand L., Vautz W. Peak detection algorithm based on second derivative properties for two dimensional ion mobility spectrometry signals. *Integration* of *Practice-Oriented Knowledge Technology: Trends and Prospectives*, Madjid Fathi (ed.), Springer-Verlag, Berlin, Heidelberg, 2013, pp. 341–354.
- 12. O'Haver T. Interactive Signal Processing Tools. Peak Finding and Measurement. URL: http://terpconnect.umd.edu/~toh/spectrum/SignalProcessi ngTools.html.

Article received in edition: 20.06.2017