УДК 681.391:543/545

ИСПОЛЬЗОВАНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ С ЦЕЛЬЮ ОБРАБОТКИ СИГНАЛОВ В ИНФОРМАЦИОННО-ИЗМЕРИТЕЛЬНЫХ СИСТЕМАХ ДЛЯ МУЛЬТИДЕТЕКТОРНОЙ ХРОМАТОГРАФИИ

Р.Т. Сайфуллин, С.С. Александров

Самарский государственный технический университет
Россия, 443100, г. Самара, ул. Молодогвардейская, 244

Рассматривается процесс формирования сигналов в мультиканальной хроматографической системе. Число каналов определяется либо числом используемых детекторов (газовая хроматография), либо числом длин волн, на которых выходные хроматографические сигналы регистрируются (жидкостная хроматография). Алгоритм обработки мультиканальных сигналов с использованием метода главных компонент состоит из следующих этапов: регистрация многоканальных хроматограмм на выходе хроматографа; формирование матрицы отсчетов; получение из матрицы отсчетов факторов; на основе расчета и анализа коэффициентов парной корреляции сравнение факторов и принятие решения.

Ключевые слова: *хроматограмма, мультиканальный сигнал, главные компоненты, коэффициент корреляции.*

Внедрение в исследовательскую практику многопараметрических информационно-измерительных систем (ИИС) на порядок увеличивает объем регистрируемой информации, при этом существенно усложняется анализ полученных данных. Особое место в структуре многопараметрических ИИС занимают приборы, позволяющие регистрировать данные большим количеством датчиков (детекторов) в течение длительного времени. Актуальным при этом является процесс автоматизации анализа и классификации огромного массива полученной информации. Одной из сфер применения подобного рода ИИС являются аналитические измерения.

Современные аналитические приборы могут производить огромное количество измерений. Однако из-за мультиколлинеарности доля полезной информации в таком массиве данных может быть относительно невелика. Для выделения значимой информации используются методы сжатия данных, основанные на представлении исходных данных через новые переменные существенно меньшей размерности, чем число исходных переменных. Сжатие данных позволяет представить полезную информацию в более компактном виде, удобном для визуализации и интерпретации. Одним из основных способов сжатия данных является метод главных компонент (МГК) [1]. Заметим, что следует различать главную компоненту (жен. род), определяемую в МГК, и химический компонент (муж. род), присутствующий в исследуемом образце. В первом случае это абстрактная величина, характеризуемая вектором нагрузок; во втором это реальное вещество, имеющее свой спектр. МГК является разновидностью мультикорреляционного

Раухат Талгатович Сайфуллин (д.т.н., проф.), профессор кафедры «Информационно-измерительная техника».

Сергей Сергеевич Александров, аспирант.

анализа и основан на обработке корреляционных матриц большой размерности.

При использовании МГК предполагается, что несколько измеряемых переменных сильно коррелируют друг с другом. Это означает, что либо они взаимно определяют друг друга, либо связь между ними обусловливается третьей величиной, которую непосредственно измерить нельзя [2]. Модель главных компонент в большей степени связана с этим предположением и дает возможность получить числовые значения этих третьих величин в виде набора линейно-независимых факторов (ЛНФ), которые описывают и воспроизводят исходную матрицу данных с необходимой точностью.

Алгоритм обработки мультиканальных сигналов с использованием МГК состоит из следующих этапов:

- регистрация многоканальных хроматограмм на выходе хроматографа;
- формирование на основе многоканальных хроматограмм матрицы отсчетов;
- получение факторов из матрицы отсчетов;
- сравнение факторов между собой на основе анализа и расчета коэффициентов парной корреляции.

Пусть в результате эксперимента сформирована матрица данных *X*. Это могут быть, например, многоканальные хроматограммы, регистрируемые на выходах многоволнового детектора на основе диодной матрицы (жидкостная хроматография) [2], либо сигналы на выходах детекторов разного принципа действия, соединенных последовательно или параллельно (газовая хроматография) [3] (см. рисунок). Число каналов определяется либо числом используемых детекторов, либо числом длин волн, на которых регистрируются выходные хроматографические сигналы.

ИИС для мультидетекторной хроматографии с параллельно подключенными детекторами представлена на рис. *а*, с последовательно подключенными детекторами – на рис. *б*.

Автоматический пробоотборник подает в поток газа-носителя определенное количество анализируемой смеси. В хроматографической колонке осуществляется разделение смеси на отдельные составляющие компоненты, попадающие в детектор. Детектор регистрирует присутствие веществ, отличающихся по физическим или физико-химическим свойствам от газа-носителя, и преобразует возникающие изменения в электрический сигнал.

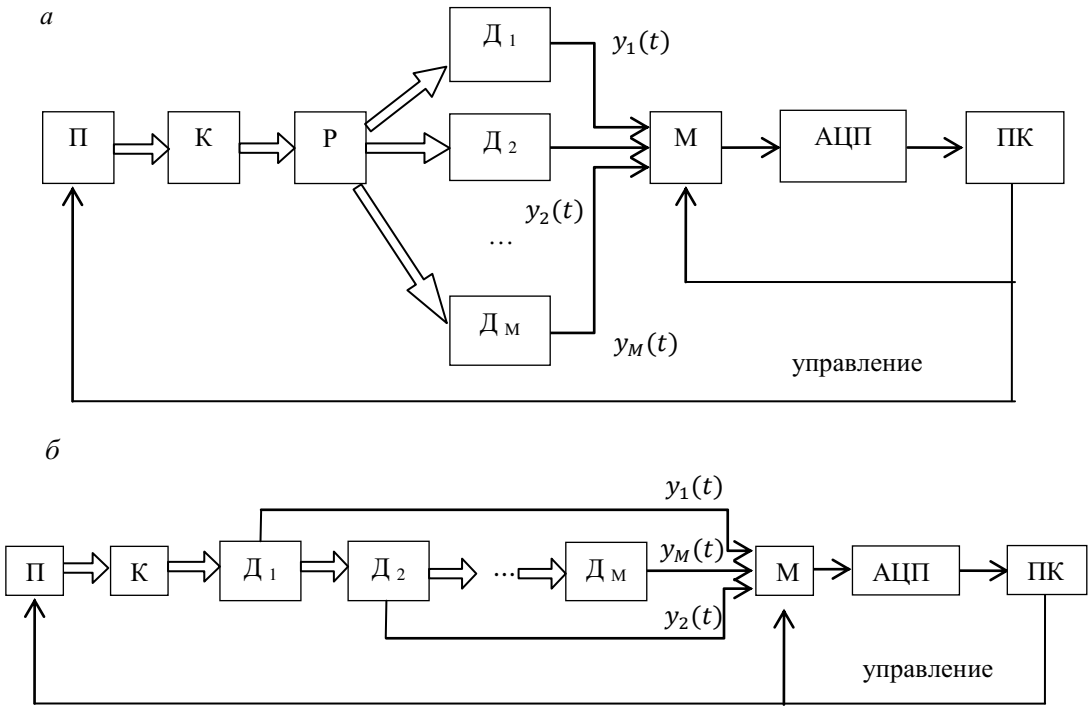
Используют следующие типы детекторов: ПИД – пламенно-ионизационный детектор, ДТП – детектор по теплопроводности (катарометр), ЭЗД – электронно-захватный детектор, ПФД – пламенно-фотометрический детектор, ТИД – термоионный детектор, ФИД – фотоионизационный детектор. Детекторы могут объединяться в аналитическом модуле в различных комбинациях.

Далее происходит нормировка и аналого-цифровое преобразование полученного сигнала. В мультидетекторном хроматографе выходные сигналы детекторов поступают на вход АЦП через мультиплексор, который осуществляет циклический поочередный опрос всех каналов хроматографа. Для каждого из детекторов ПК формирует зависимости сигнала от времени (хроматограммы).

Сигнал каждого из детекторов с номерами $m \in \{1,\ 2,\ \ldots\ ,\ M\}$ (M – общее число детекторов) представляется в виде совокупности дискретных отсчетов, взятых в моменты времени с номерами $n = 1,\ 2,\ \ldots\ ,\ N$, где N – общее число отсчетов. Тогда в векторной форме сигнал может быть представлен как

$$X = (X_1,\ X_2,\ \ldots\ ,\ X_n,\ \ldots\ ,\ X_N) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2n} & \cdots & x_{2N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{M1} & x_{M2} & \cdots & x_{Mn} & \cdots & x_{MN} \end{pmatrix},$$

где $X_n = (x_{1n},\ \ldots\ ,\ x_{mn},\ \ldots\ ,\ x_{Mn})^T$ – значение сигнала в момент времени с номером $n \in \{1,\ 2,\ \ldots\ ,\ N\}$; x_{mn} – значение m-й компоненты сигнала в указанный момент времени, $m \in \{1,\ 2,\ \ldots\ ,\ M\}$; T – знак транспонирования.



ИИС для мультидетекторной хроматографии:
а – параллельное подключение детекторов; *б* – последовательное подключение детекторов;
П – автоматический пробоотборник, К – колонка, Р – распределитель (делитель потока), $Д_1$–$Д_М$ – детекторы, М – мультиплексор, АЦП – аналого-цифровой преобразователь, ПК – персональный компьютер

Таким образом, зарегистрированные хроматограммы могут быть представлены в виде матрицы отсчетов *X*. Например, при средней длительности хроматограммы 17,05 мин, интервале дискретизации $\Delta = 1$ с, использовании 8 детекторов (M=8) формируется матрица X размера (8×1024). В терминах МГК для матрицы исходных данных X размерностью $M \times N$ M – число образцов (число объектов наблюдения), N – число переменных.

Исходная матрица данных X подвергается предварительной обработке, которая включает операции центрирования и нормирования. Тогда для матрицы Z центрированных и нормированных значений переменных элементы матрицы Z вычисляют по формуле

$$z_{mn} = \frac{x_{mn} - \bar{x}_n}{s_n},$$

где x_{mn} – m-е значения n-й компоненты вектора X_n, $m = 1, 2, \ldots, M$; $n = 1, 2, \ldots, N$;

$\bar{x}_n$ – оценка математического ожидания компонент вектора X_n:

$$\bar{x}_n = \frac{\sum_m x_{mn}}{M},$$

s_n – оценка среднеквадратического отклонения компонент вектора X_n:

$$s_n = \sqrt{\frac{\sum_m (x_{mn} - \bar{x}_n)^2}{M - 1}}.$$

МГК заключается в нахождении для исходных данных такого их ортогонального преобразования в новую систему координат, для которого выполняются следующие условия:

– выборочная дисперсия данных максимальна вдоль первой координаты;

– выборочная дисперсия данных вдоль k-й координаты максимальна при условии ортогональности первым (k-1) координатам.

Следовательно, направления базисных векторов будут выбраны так, что коэффициент ковариации между проекциями исходного набора данных на различные координатные оси будет равен нулю.

В векторной форме проекция многоканального хроматографического сигнала на главные компоненты может быть представлена в виде

$$Y = AZ,$$

где Z – исходный многоканальный сигнал (центрированный и нормированный) размерности ($M \times N$); A – матрица преобразования размерности ($K \times M$) (матрица нагрузок); Y – некоррелированный многомерный временной ряд (размерности ($K \times N$)), представляющий собой проекцию исходного сигнала на K главных компонентов.

Процедура построения матрицы A включает в себя следующие шаги.

1. Для исходного многоканального хроматографического сигнала Z производится расчет ковариационной матрицы $C = \{c_{ij}\}_{M \times M}$. Элементы ковариационной матрицы определяются как

$$c_{ij} = \mathrm{cov}\ (Z_i, Z_j) = \frac{1}{N-1} Z_i Z_j^T,\ i, j = \{1, 2, \ldots, M\},$$

где Z_i и Z_j – строки матрицы Z.

2. Осуществляется поиск собственных значений λ_i и собственных векторов F_i ковариационной матрицы C.

3. Матрица преобразования A формируется из первых K собственных векторов F_i, расположенных в порядке убывания соответствующих собственных значений λ_i ковариационной матрицы C:

$$A = (F_1, , F_2, \ldots, F_K)^T,$$

где $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \geq 0,\ 0 < K \leq M$.

МГК работает как фильтр: сигнал содержится в основном в проекции на первые главные компоненты, а в остальных компонентах пропорции шума намного выше. Оценку числа главных компонентов будем производить по правилу «сломанной трости» [4].

Набор нормированных собственных чисел $\frac{\lambda_i}{trC}$, $i = 1,\ 2,\ \ldots\ ,\ M$ (trC – след матрицы C) сравнивается с распределением длин обломков трости единичной длины, сломанной в (M-1)-й случайно выбранной точке (точки разлома выбираются независимо и равнораспределены по длине трости). Пусть L_i ($i = 1,\ 2,\ \ldots\ ,\ M$) – длины полученных кусков трости, занумерованные в порядке убывания длины: $L_1 \geq L_2 \geq \ldots \geq L_M$.

Математическое ожидание $\overline{L_i}$

$$l_i = \overline{L_i} = \frac{1}{M}\sum_{j=i}^{M}\frac{1}{j}.$$

По правилу сломанной трости K-й собственный вектор (в порядке убывания собственных чисел λ_i) сохраняется в списке главных компонент, если

$$\frac{\lambda_1}{trC} > l_1\ \ \&\ \ \frac{\lambda_2}{trC} > l_2\ \ \&\ \ \ldots\ \ \&\ \ \frac{\lambda_K}{trC} > l_K.$$

Сравнение многокомпонентных хроматограмм, соответствующих анализируемым образцам, осуществляется путем сравнения наборов расчетных ЛНФ с помощью матрицы коэффициентов корреляции ЛНФ исследуемых образцов и контрольного образца.

Основным показателем сходства или различия факторов при их сравнении может быть выбран коэффициент корреляции. Пусть сравниваются факторы F_1 и F_2. Коэффициент корреляции $R_{F_1F_2}$ показывает, являются ли сравниваемые величины линейно зависимыми, т. е. справедливость выполнения уравнения

$$F_1(i) = a + bF_2(i), \qquad (1)$$

где $F_1(i)$ и $F_2(i)$ – сравниваемые факторы, a и b – некоторые коэффициенты.

Чем меньше коэффициент корреляции, тем менее похожи сравниваемые объекты; чем больше – тем более похожи.

Согласно Джаффе [5] корреляция считается удовлетворительной, если $0{,}94 < R_{F_1F_2} < 0{,}97$; хорошей, если $0{,}97 < R_{F_1F_2} < 0{,}99$, и отличной при $R_{F_1F_2} > 0{,}99$. При $R_{F_1F_2} > 0{,}995$ уравнение (1) можно считать аналитической зависимостью.

Исходные матрицы мультидетекторных хроматограмм обычно описываются тремя-четырьмя ЛНФ. Максимальный вклад четвертого, пятого факторов, как правило, находится на уровне ошибок проведения хроматографического эксперимента. При идентификации образцов целесообразно ограничиться числовыми значениями первых трех факторов. Таким образом, критерием идентичности образца являются коэффициенты парной корреляции, соответствующие данному

образцу факторов $F_1 \div F_3$. Полностью идентичными образцами можно считать образцы, для которых коэффициенты парной корреляции факторов F_1 не ниже 0,99; F_2 не ниже 0,98; F_3 не ниже 0,96.

Пусть в результате эксперимента получены факторы F_1^1, F_2^1, F_3^1 для первого образца и факторы F_1^2, F_2^2, F_3^2 для второго образца. Пусть коэффициенты корреляции факторов, полученные для этих двух образцов, имеют значения, представленные в таблице:

Значения коэффициентов парной корреляции

Фактор	F_1^1	F_2^1	F_3^1	F_1^2	F_2^2	F_3^2
F_1^1	1,000	0,000	-0,220	0,996	-0,015	-0,125
F_2^1	0,000	1,000	-0,305	-0,012	0,984	-0,184
F_3^1	-0,220	-0,305	1,000	-0,270	-0,471	0,960
F_1^2	0,996	-0,012	-0,270	1,000	-0,015	-0,180
F_2^2	-0,015	0,984	-0,471	-0,015	1,000	-0,358
F_3^2	-0,125	-0,184	0,960	-0,180	-0,358	1,000

Следовательно, коэффициенты парной корреляции соответствующих факторов таковы: коэффициент корреляции между факторами F_1^1 и F_1^2 равен 0,996; между факторами F_2^1 и F_2^2 равен 0,984; между факторами F_3^1 и F_3^2 равен 0,960.

Значения этих коэффициентов могут быть использованы для выявления идентичности образцов. Критерии сходства могут быть установлены по Джаффе: похожими считаются образцы, для которых коэффициенты парной корреляции соответствующих факторов выше 0,94.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Померанцев А.Л.* Метод главных компонент // Российское хемометрическое общество. – 2008 [Электронный ресурс]. – Режим доступа: http://rcs.chph.ras.ru/
2. *Гаврилина В.А., Сычев С.Н.* Исходные гипотезы для распознавания многокомпонентных физико-химических систем комбинацией «высокоэффективная жидкостная хроматография – метод главных компонент» // Сорбционные хроматографические процессы. – 2012. – Т. 12. – Вып. 5. – С. 798-805.
3. *Сайфуллин Р.Т., Александров С.С.* Определение качественного и количественного состава компонентов сложных смесей с использованием мультидетекторного хроматографа // Вестник Самарского государственного технического университета. Сер. Технические науки. – 201. – № 4 (40). – С. 77-83.
4. *Cangelosi R., Goriely A.* Component retention in principal component analysis with application to DNA microarray data // Biology Direct. – 2007. – 2:2 [Электронный ресурс]. – Режим доступа: http://biology-direct.com/content/2/1/2
5. *Jaffe H.H.* Reexaminatio of the Hammett equation // Chem. Rev., 1953. – V. 53. – № 2. – p. 191-254.

Статья поступила в редакцию 15 января 2015 г.

USING THE PRINCIPAL COMPONENT ANALYSIS FOR SIGNAL PROCESSING IN THE MULTIDETECTOR IMS CHROMATOGRAPHY

R.T. Saifullin, S.S. Aleksandrov

Samara State Technical University
244, Molodogvardeyskaya st., Samara, 443100, Russia

The process of signal formation in a multichannel chromatographic system is treated. The number of channels is determined either by the number of detectors (gas chromatography) or by the number of wavelengths of spectrometric detector signal (liquid chromatography). The algorithm of multi-channel signals processing using the principal-component method consists of the following stages: registrating chromatographic detector output (multi-channel chromatograms)t; forming a matrix of samples; obtaining the parameters from the matrix samples on the basis of the pair-correlation coefficients analysis and calculation, comparing the parameters, decision- making.

Keywords: *chromatogram, multichannel signal, the main components, correlation coefficient.*

Rauhat T. Saifullin (Dr. Sci. (Techn.)), Professor.
Sergey S. Aleksandrov, Postgraduate Student.