ГОСУДАРСТВЕННЫЙ КОМИТЕТ РОССИЙСКОЙ ФЕДЕРАЦИИ ПО ВЫСШЕМУ ОБРАЗОВАНИЮ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ В ЖИДКОСТНОЙ ХРОМАТОГРАФИИ

Дипломная работа студента 874 группы *К.В. Воронцова*
Научный руководитель чл.-корр. АЕН России, д.ф.-м.н. *К.В. Рудаков*

Москва — 1994

Содержание

§1. Введение 3
§2. Структура исходных данных и постановка задачи 6
§3. Алгоритмы вычисления k-разложений 9
§4. Оптимальные k-разложения 15
§5. Алгоритмы предварительной обработки данных 21
§6. Идентификация химических веществ 26
Список литературы 27

§1. Введение

В настоящей работе рассматривается один из подходов к распознаванию (идентификации) компонентов смеси химических веществ по данным хроматографического анализа. Предлагаемые алгоритмы ориентированы на работу с исходными данными системы экстренной токсикологии REMEDI (Bio-Rad Laboratories, США), предназначенной для оперативного анализа образцов плазмы крови.

В основе хроматографических методов лежит следующий простой принцип. Имеется трубка (*колонка*), заполненная поверхностно активным пористым веществом (*сорбентом*), способным на некоторое время захватывать (*адсорбировать*) молекулы других веществ. Сквозь колонку с постоянной скоростью пропускается однородный поток жидкости, называемой подвижной фазой или *элюентом*. В начальный момент времени в поток впрыскивается капля анализируемого раствора. Благодаря специальным свойствам сорбента скорости движения различных веществ вдоль колонки оказываются различными, и на выходе они появляются через разные промежутки времени (график зависимости суммарной концентрации от времени называют *хроматограммой*, а момент выхода отдельного вещества — его *временем удерживания*). Разделенные таким образом вещества подвергаются анализу и идентификации.

В REMEDI с этой целью используется сканирующий ультрафиолетовый (УФ) детектор, который через равные промежутки времени снимает спектры поглощения смеси, находящейся на выходе колонки. На основе этих данных компьютер производит идентификацию химических веществ. Алгоритм идентификации сопоставляет спектральные и хроматографические характеристики обнаруженных пиков концентрации со своей базой данных (более 300 лекарственных препаратов и их метаболитов). Для каждого хроматографического пика формируется список веществ-кандидатов, из которого методом комбинированного поиска с возвратами выбирается окончательное решение. Более полное описание данного алгоритма можно найти в работе [11].

Опыт эксплуатации REMEDI и анализ зарубежных публикаций позволили выявить следующие недостатки, присущие применяемому алгоритму.

1. Сравнивая анализируемый спектр со спектрами веществ из базы данных, алгоритм использует в основном «точечные» критерии, такие как совпадение точек максимумов и точек перегиба на спектрах, а также «скорректированных отношений» [11], вычисленных при некоторых длинах волн. Ясно, что такой подход заведомо ухудшает качество идентификации, поскольку игнорируется значительная часть информации, содержащейся в спектрах (каждый спектр состоит из 112 точек). Было замечено, что системе не всегда удается идентифицировать вещество, даже если оно присутствует в базе данных. Эти случаи поддаются следующей классификации.

Близкие времена удерживания у двух или большего числа веществ. В результате наложения пиков нескольких веществ первоначальная форма их спектров искажается и алгоритм с частотой около 10% начинает ошибаться при идентификации обнаруженных пиков или одного из них. Показателен случай, когда не идентифицировался фенобарбетал из-за того, что присутствовавшая в нем характерная технологическая примесь давала близкий пик.

Последовательность перекрывающихся пиков. При наложении большого числа пиков некоторые из них даже не обнаруживались, а частота правильной идентификации заметно снижалась. По этой причине затруднена идентификация многих снотворных и других веществ, элюирующих в начальной части хроматограммы [9].

Низкие концентрации вещества в образце. В этом случае ненадежная идентификация вполне оправдана, так как высота пика становится сравнимой с уровнем шума УФ-детектора. Как показали эксперименты [9], при отношении сигнала к шуму, равному 12, правильно идентифицируется не менее 95% изолированных пиков. Следует однако учитывать, что в алгоритме не делается попыток сгладить влияние шума путем усреднения спектров, что, как известно из статистики, должно привести к увеличению точности. Таким образом, имеется еще некоторый резерв для снижения порога идентифицируемости.

Вещества с нехарактерным спектром, лишенным точек максимума или точек перегиба. Поскольку эти точки используются при идентификации, их отсутствие приводит к увеличению вероятности неверной идентификации во всех выше перечисленных случаях.

2. Хотя в системе заложена возможность самообучения, в действительности занести в базу данных хроматографические и спектральные данные неизвестного вещества можно только при наличии отчетливого изолированного пика. На практике это условие выполняется довольно редко, что сильно снижает эффективность самообучения и требует проведения отдельных анализов исключительно с целью обучения системы.

3. Результаты предыдущих анализов сохраняются программой на жестком диске, однако если их число превысит максимально допустимое (около 130), все они будут уничтожены и заполнение диска начнется заново. Этот факт препятствует теоретически возможному «накоплению опыта» и связан с тем, что для каждого анализа целиком хранится файл исходных данных размером около 430 кбайт, который никак не преобразуется и не архивируется.

Перечисленные недостатки можно было бы устранить путем предварительной обработки данных, поступающих с УФ-детектора, состоящей в выделении истинных спектров и хроматограмм каждого вещества. Под *истинным спектром* будем понимать тот спектр, который давало бы данное вещество на отчетливом изолированном хроматографическом пике, не подверженном влиянию пиков с близкими временами удерживания. Аналогично определяется понятие *истинной*

хроматограммы. Поскольку истинный спектр является неизменной характеристикой вещества, он может быть использован для более надежной идентификации.

В настоящей работе рассматривается один из подходов к предварительной обработке данных, позволяющий устранить перечисленные недостатки. В общих чертах предлагаемый подход состоит в следующем [12,13].

Как станет видно из анализа структуры исходных данных, задача предобработки сводится к решению матричного уравнения $XY^T = Z$ относительно неизвестных матриц X и Y с неизвестным числом столбцов k, равным числу веществ в образце.

Простейший способ определить k — положить его наименьшим, при котором решения данного уравнения еще существуют. Однако на практике не имеет смысла говорить о точных решениях, так как элементы исходной матрицы Z заданы с погрешностями. В самом общем случае должна проявляться закономерность: чем больше погрешность исходных данных, тем меньше информации можно их них извлечь. Поскольку матрицы X и Y как раз и являются извлекаемой информацией, число k может уменьшиться при увеличении погрешности измерений. Отсюда, в частности, следует, что для его определения необходимо, как минимум, обладать априорной оценкой погрешности исходных данных.

Рассматриваемое матричное уравнение имеет бесконечно много решений, причем если известно одно из них, то любое другое может быть рассчитано по простым формулам. Таким образом, не лишена смысла задача вычисления произвольного решения указанного уравнения. Будет построен итерационный процесс, скорость сходимости которого тем выше, чем меньше погрешность исходных данных. В частности, если погрешность равна нулю, процесс сходится за один шаг из почти любого начального приближения.

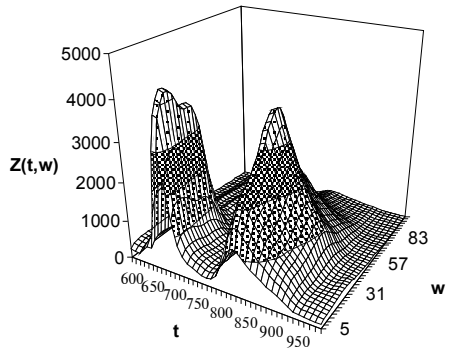
Задача решения матричного уравнения некорректно поставлена по Адамару (нет единственности). В то же время, для распознавания химических веществ необходимо иметь единственное описание истинного спектра и истинной хроматограммы каждого вещества. Для устранения неоднозначности (регуляризации) предлагается вводить дополнительные ограничения на вид хроматограмм, считая, что истинные хроматограммы описываются элементами достаточно узкого семейства функций (модели хроматографического пика). Будет показано, что если модель точна и удовлетворяет специальному условию нелинейности, то решение задачи регуляризации существует и единственно. При этом удается построить алгоритмы, позволяющие находить спектры и хроматограммы, достаточно близкие к истинным, опираясь на вычисленное ранее произвольное решение. Путем постепенного уточнения модели пика можно добиться приемлемой точности описания исходных данных.

§2. Структура исходных данных и постановка задачи

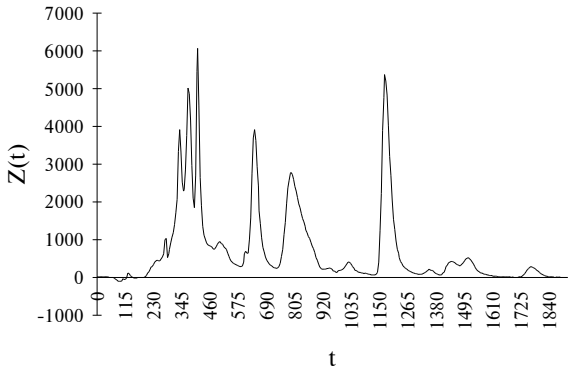
Рассмотрим структуру исходных данных (рис. 1). В каждый из моментов времени $t_1,\ldots,t_n$, $n=1920$, в компьютер поступает набор m чисел — сигналов, зарегистрированных УФ-детектором на длинах волн $w_1,\ldots,w_m$, $m=112$. Величина сигнала z прямо пропорциональна взвешенной сумме концентраций C всех веществ, находящихся в данный момент на выходе колонки. Весовые коэффициенты F определяются способностью веществ поглощать излучение соответствующих длин волн и, очевидно, не зависят от времени. В то же время концентрации веществ в подвижной фазе не могут зависеть от длины волны. Таким образом, совокупность исходных данных можно рассматривать как таблично заданную функцию двух переменных

$$z(t,w)=\sum_{s=1}^{k} C_s(t)\,F_s(w),$$

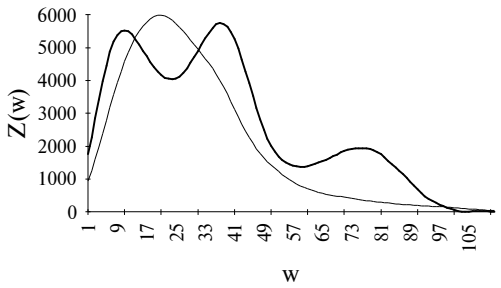
где k — число химических веществ в образце, $C_s(t)$ — истинная хроматограмма s-ого вещества, $F_s(w)$ — истинный спектр s-ого вещества.



(а)



(б)



(в)

Рис. 1. Структура исходных данных:
а - фрагмент исходной поверхности;
б - характерный вид хроматограмм (фиксирована длина волны);
в - характерный вид спектров (фиксированы 2 момента времени).

Задача восстановления истинных спектров и хроматограмм может рассматриваться как частный случай более общей задачи нахождения структурного описания элемента множества, заданного путем обобщенного индуктивного определения.

Говорят, что множество M задано путем *обобщенного индуктивного определения*, если указано подмножество M_0 непроизводных элементов, набор операций, с помощью которых можно порождать новые элементы множества M из уже имеющихся, и указано, что множество M состоит из

тех и только тех элементов, которые могут быть получены таким образом.

Элементы множества M разумно представлять в виде *структурных описаний* (формул), составленных из описаний отдельных непроизводных элементов и порядка выполнения операций над ними. Если при этом удается достаточно просто описывать сами непроизводные элементы, то такой подход позволяет представлять элементы множества M в компактном виде и, кроме того, иметь информацию об их «внутреннем строении». Как правило, наличие такой информации позволяет эффективно вычислять оценки близости при решении задач распознавания.

В нашем случае множеством непроизводных элементов являются функции с разделенными переменными. Набор операций включает в себя операции сложения и умножения на число. Допустимая функция (элемент порожденного множества) задется с погрешностями в узлах прямоугольной сетки. Требуется найти структурное описание этой функции. Искомое описание должно быть единственным и устойчивым по отношению к погрешностям исходных данных, в частности, малые погрешности не должны приводить к изменению числа непроизводных элементов в описании.

Будем называть *непроизводной поверхностью* всякую функцию двух переменных вида $f(x)g(y)$, где $f(x)$ и $g(y)$ — произвольные функции, определенные на множествах $\Omega_x \subseteq \mathbb{R}$ и $\Omega_y \subseteq \mathbb{R}$ соответственно, а *допустимой поверхностью* — всякую функцию двух переменных, представимую в виде суммы конечного числа непроизводных поверхностей.

Сформулируем задачу предварительной обработки данных следующим образом. Пусть допустимая поверхность

$$z(x,y) = \sum_{s=1}^{k} f_s(x) g_s(y) \tag{2.1}$$

задана с погрешностями в узлах прямоугольной сетки (ξ_i, η_j), $\xi_i \in \Omega_x$, $\eta_j \in \Omega_y$, $i = 1,\ldots,n$, $j = 1,\ldots,m$ и пусть для определенности $n \le m$. Число непроизводных поверхностей k неизвестно. Требуется найти структурное описание допустимой поверхности, в данном случае определить k и восстановить значения функций $f_s(x)$, $g_s(y)$ в узлах сетки.

Обозначим через $M_{n,m}$ пространство действительных матриц размера $n \times m$ и введем матрицы

$$X = \left[f_s(\xi_i)\right] \in M_{n,k} ,$$
$$Y = \left[g_s(\eta_j)\right] \in M_{m,k} ,$$
$$Z = \left[z(\xi_i, \eta_j)\right] \in M_{n,m}.$$

Записав соотношение (2.1) в узлах сетки, получим матричное уравнение

$$XY^T = Z \tag{2.2}$$

относительно неизвестных матриц X и Y с неизвестным числом столбцов k. Для определения k потребуем, чтобы искомое разложение вида (2.1) состояло из минимального числа слагаемых.

Определение. Матрица $Z \in M_{n,m}$ называется *k-разложимой*, если k — наименьшее число, при котором существуют матрицы $X \in M_{n,k}$ и $Y \in M_{m,k}$, удовлетворяющие уравнению $XY^T = Z$. Упорядоченная пара матриц (X, Y) называется *k-разложением Z*.

Множество всех k-разложений Z будем обозначать через M_Z^k.

Лемма 2.1. *Матрица $Z \in M_{n,m}$ является k-разложимой тогда и только тогда, когда ее ранг равен k. Всякое k-разложение состоит из матриц полного ранга.*

Доказательство. Достаточность. Пусть матрица Z имеет ранг k. Тогда она не может быть q-разложимой при $q<k$, иначе ее ранг был бы меньше k. С другой стороны, всякая матрица ранга k представима в виде скелетного разложения [3] — произведения $n \times k$- и $k \times m$-матриц, причем в качестве столбцов первой из них можно взять k линейно независимых столбцов Z. Таким образом, матрица Z является k-разложимой.

Необходимость. Пусть (X, Y) — произвольное k-разложение Z. Матрицы X, Y и Z, очевидно, не могут иметь ранг, больший k. Они также не могут иметь ранг, меньший k, так как иначе, по доказанному, матрица Z оказалась бы q-разложимой при $q<k$. Лемма доказана.

Пусть (X, Y) — произвольное k-разложение Z. Тогда согласно лемме матрицы X^TX и Y^TY невырождены и (2.2) равносильно каждому из равенств

$$Y = Z^TX(X^TX)^{-1}, \quad X = ZY(Y^TY)^{-1}. \tag{2.3}$$

Лемма 2.2. *Пусть (X, Y) — какое-либо k-разложение матрицы Z. Тогда множество всех ее k-разложений имеет вид*

$$M_Z^k = \{\ (XA, YA^{-1T}) :\ A \in M_{k,k},\ \mathrm{rg}A = k\}.$$

Доказательство. Очевидно, всякая пара матриц указанного вида является k-разложением Z. Верно и обратное: для любого k-разложения (U, V) найдется невырожденная $k \times k$-матрица A такая, что $U = XA$, $V = YA^{-1T}$. Действительно, используя (2.3), легко проверить истинность этих равенств, если положить $A = Y^TV(V^TV)^{-1}$. Причем A невырождена в силу оценок ранга сверху и снизу: $k = \mathrm{rg}U = \mathrm{rg}XA \le \mathrm{rg}A \le k$. Таким образом, множество указанного вида и множество всех k-разложений матрицы Z совпадают. Лемма доказана.

Непосредственно из лемм 2.1 и 2.2 следует

Теорема 2.1. *Если ранг матрицы Z равен k, то она k-разложима и*

$$M_Z^k = \{\ (XA, YA^{-1T}) :\ A \in M_{k,k},\ \mathrm{rg}A = k\},$$

где матрица X составлена из k линейно независимых столбцов Z, матрица Y определяется по формуле $Y = Z^T X (X^T X)^{-1}$.

Таким образом, при данном определении непроизводной поверхности для всякой допустимой поверхности можно указать бесконечно много различных структурных описаний. Для обеспечения единственности описания необходимо сужать множество непроизводных поверхностей путем введения дополнительных ограничений на функции $f(x)$ или $g(y)$. Эта задача будет рассмотрена в §4.

§3. Алгоритмы вычисления *k*-разложений

Согласно лемме 2.2, множество всех *k*-разложений однозначно определяется каким-либо одним его элементом. Поэтому практический интерес представляет задача вычисления произвольного *k*-разложения данной матрицы. Заметим, что использование с этой целью теоремы 2.1 нецелесообразно, когда элементы матрицы Z заданы с погрешностями, поскольку ее ранг может не совпадать с числом непроизводных поверхностей *k*. Возможный выход состоит в приближенном решении матричного уравнения (2.2) и определении k из условия достаточно точного (в пределах погрешностей Z) выполнения равенства (2.2).

Введем в $M_{n,m}$ евклидову норму $\|A\| = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 \right)^{1/2}$, где $A = \left[a_{ij} \right]_{n \times m}$.

Определение. Матрица Z называется *k, δ-разложимой*, если k— наименьшее число, при котором существует *k*-разложимая матрица W такая, что $\|Z - W\| \le \delta$. Пара матриц (X, Y) называется *k, δ-разложением Z*, если она является *k*-разложением W.

Рассмотрим симметрическую неотрицательно определенную матрицу ZZ^T. Все ее собственные числа сосредоточены на отрезке действительной оси $[0, \|Z\|^2]$, а соответствующие им собственные векторы могут быть выбраны так, чтобы они составляли ортонормированную систему векторов в R^n. Докажем следующую теорему, примыкающую к известным результатам [8] о сингулярном разложении матриц.

Теорема 3.1. *Пусть заданы матрица* $Z \in M_{n,m}$ *и действительное число* $\delta \ge 0$. *Пусть* $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n$ *— все собственные числа матрицы* ZZ^T. *Если k — наименьшее число, при котором выполняется условие*

$$\lambda_1 + \ldots + \lambda_k \ge \|Z\|^2 - \delta^2, \tag{3.1}$$

то матрица Z является k, δ-разложимой, причем множество наилучших в смысле евклидовой нормы k, δ-разложений Z имеет вид

$$\{ (XA, Z^T X A^{-1T}) : \ A \in M_{k,k}, \ \mathrm{rg} A = k \},$$

где $n \times k$-матрица X составлена из ортонормированных собственных векторов ZZ^T, отвечающих k наибольшим собственным числам.

Доказательство.

Определим функционал $J_q : M_{n,q} \times M_{m,q} \to \mathbb{R}$ по формуле

$$J_q(U,V) = \left\| Z - UV^T \right\|^2,$$

полагая при этом $\operatorname{rg} U = \operatorname{rg} V = q$, и найдем пару матриц (U,V), доставляющую J_q глобальный минимум. Дифференцируя J_q по U и V, получим систему нормальных уравнений для данной задачи наименьших квадратов (необходимые условия минимума):

$$\begin{cases} U^T(Z - UV^T) = 0, \\ (Z - UV^T)V = 0 \ . \end{cases} \tag{3.2}$$

Покажем, что матрица $R = V^T V U^T U$ диагонализируема. Так как матрица $U^T U$ симметрическая и положительно определенная, найдется невырожденная матрица $S \in M_{q,q}$ такая, что $S^{-1T} U^T U S^{-1} = E$. Далее, так как матрица $SV^T V S^T$ симметрическая и невырожденная, найдется ортогональная матрица $T \in M_{q,q}$ такая, что $T(SV^T V S^T)T^{-1} = \Lambda$, где $\Lambda = \operatorname{diag}(d_1, ..., d_q)$ — диагональная матрица. Положим $A=TS$. Тогда, как легко проверить, $ARA^{-1} = \Lambda$, то есть R преобразованием подобия с невырожденной преобразующей матрицей A приводится к диагональному виду.

Произведем в системе (3.2) невырожденную замену переменных

$$U = XA, \quad V = YA^{-1T}. \tag{3.3}$$

Тогда столбцы новых матриц X и Y будут ортогональны:

$$X^T X = E, \quad Y^T Y = \Lambda,$$

а система (3.2) перейдет в эквивалентную

$$\begin{cases} Y = Z^T X, \\ ZZ^T X = X\Lambda. \end{cases}$$

Второе из полученных уравнений означает, что $d_1, \ldots, d_q$ являются собственными числами, а столбцы матрицы X — соответствующими им собственными векторами матрицы ZZ^T.

Для получения не только необходимого, но и достаточного условия минимума вычислим значение функционала J_q^* при U и V, удовлетворяющих (3.2). Поскольку $J_q(U,V) = J_q(X,Y)$, имеем:

$$J_q^* = \operatorname{tr}(Z - XY^T)^T (Z - XY^T) = \operatorname{tr} ZZ^T - \operatorname{tr}\Lambda = \left\| Z \right\|^2 - (d_1 + ... + d_q).$$

Глобальный минимум функционала $J_q(U,V)$ будет достигаться тогда и только тогда, когда $d_1, \ldots, d_q$ будут q наибольшими собственными числами матрицы ZZ^T. Причем, если $q=k$, то в силу условия (3.1)

$$\left\| Z - UV^T \right\|^2 = J_k^* = \left\| Z \right\|^2 - \lambda_1 - \ldots - \lambda_k \le \delta^2.$$

Если же $q<k$, то для любых $U \in M_{n,q}$ и $V \in M_{m,q}$ справедливо неравенство

$$\left\| Z - UV^T \right\|^2 \geq J_q^* = \left\| Z \right\|^2 - \lambda_1 - \ldots - \lambda_q > \delta^2 .$$

Таким образом, матрица Z оказывается k,δ-разложимой. При этом те и только те k,δ-разложения (U,V) минимизируют $J_k(U,V)$, которые преобразованием (3.3) связаны с парой $(X, Z^T X)$, в которой матрица X составлена из ортонормированных собственных векторов ZZ^T, отвечающих k наибольшим собственным числам.

Теорема доказана.

Теорема 3.1 позволяет найти k,δ-разложение Z, если задано δ и известен набор собственных чисел и векторов матрицы ZZ^T. Однако при увеличении размера матрицы Z затраты машинного времени на нахождение спектра существенно возрастают. Поэтому целесообразно рассматривать другие алгоритмы, не требующие явного решения задачи на собственные значения.

Исследуем возможность вычисления k,δ-разложений с помощью итерационного процесса, основанного на соотношениях (2.3).

Введем следующие обозначения. Пусть $X \in M_{n,k}$ — матрица ранга k. Линейную оболочку столбцов матрицы X обозначим через $\mathscr{L}(X)$. Проекционная матрица

$$P_X = E - X(X^T X)^{-1} X^T ,$$

если рассматривать ее как линейный оператор из $\mathbb{R}^n$ в $\mathbb{R}^n$, ставит в соответствие вектору его проекцию на ортогональное дополнение подпространства $\mathscr{L}(X)$. Далее будут использоваться следующие свойства проекционных матриц.

1. Матрица P_X симметрична ($P_X^T = P_X$), идемпотентна ($P_X P_X = P_X$) и вырождена ($P_X X = 0$, ранг P_X равен n-k).

2. Матрица P_X инвариантна относительно выбора базиса в подпространстве $\mathscr{L}(X)$: $P_{XA} = P_X$ для любой невырожденной $A \in M_{k,k}$.

3. Через проекционную матрицу выражается решение задачи наименьших квадратов [3,8]:

$$\min_X \left\| AX - B \right\|^2 = \left\| P_A B \right\|^2 .$$

Последний факт позволяет определить синус угла $\varphi \in [0, \pi/2]$ между подпространствами $\mathscr{L}(X)$ и $\mathscr{L}(X_0)$, $X_0 \in M_{n,k}$ как максимальное расстояние от единичного вектора одного из подпространств до другого подпространства:

$$\sin\varphi = \max_{a \in M_{k,1}} \frac{\left\| P_X X_0 a \right\|}{\left\| X_0 a \right\|} . \qquad (3.4)$$

Теорема 3.2. *Пусть (X,Y) — наилучшее в смысле евклидовой нормы k,δ-разложение матрицы $Z \in M_{n,m}$ ранга n, начальное приближение*

$X_0 \in M_{n,k}$ *удовлетворяет условию* $\operatorname{rg} Y_0^T Y = k$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ — *все собственные числа матрицы* ZZ^T.

Тогда для итерационного процесса

$$Y_r = Z^T X_{r-1}, \quad X_r = ZY_r (Y_r^T Y_r)^{-1}, \; r = 1,2,\ldots \tag{3.5}$$

справедлива оценка

$$\left\| Z - X_r Y_r^T \right\| \leq \delta + \delta\sqrt{k}\, \frac{\mu^{r-1}}{\cos\varphi}, \tag{3.6}$$

где $\mu = \delta^2 \left\| \Lambda^{-1} \right\|$, $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_k)$, φ — *угол между* $\mathscr{L}(X)$ *и* $\mathscr{L}(X_0)$.

Доказательство. Поскольку Z и X_0 — матрицы полного ранга, то при любом $r > 0$ матрицы X_r и Y_r имеют ранг k, следовательно, вычисление обратной матрицы в (3.5) всегда возможно. Введем обозначения:

$$W = ZZ^T,$$
$$S_r = X(X_r^T X)^{-1} X_r^T, \; r = 0,1,\ldots$$

Используя соотношения (3.5), легко проверить равенства:

$$Z - X_r Y_r^T = (E - S_{r-1})(Z - XY^T) P_{Y_r},$$
$$S_r = X(X_0^T W^r X)^{-1} X_0^T W^r.$$

Поскольку (X,Y) — наилучшее в смысле евклидовой нормы k,δ-разложение Z, то, согласно теореме 3.1, можно без ограничения общности предполагать, что матрица X составлена из ортонормированных собственных векторов матрицы W. Тогда справедливы равенства $X^T X = E$ и $WX = X\Lambda$, следовательно,

$$W^r X = X\Lambda^r,$$
$$P_X W^r = (P_X W P_X)^r.$$

Используя представление единичной матрицы $E = P_X + XX^T$, найдем:

$$Z - X_r Y_r^T = (P_X - X\Lambda^{1-r}(X_0^T X)^{-1} X_0^T (P_X W P_X)^{r-1})(Z - XY^T) P_{Y_r}.$$

Учитывая, что при умножении произвольной матрицы на проекционную ее норма может только уменьшиться, получим оценку

$$\left\| Z - X_r Y_r^T \right\| \leq \delta + \left\| \Lambda^{1-r} \right\| \left\| (X_0^T X)^{-1} X_0^T \right\| \delta^{2r-1} \leq \delta + \delta\mu^{r-1} \left\| S_0 \right\|.$$

Оценим теперь $\|S_0\|$. Так как матрица S_0 не изменится, если X_0 умножить справа на произвольную невырожденную $k \times k$-матрицу, будем считать, что столбцы X_0 ортонормированы. Тогда

$$\left\| S_0 \right\|^2 = tr(X_0^T XX^T X_0)^{-1} \leq k / \lambda_0,$$

где λ_0 — минимальное собственное число матрицы $G = X_0^T XX^T X_0$. Поскольку матрица G симметрическая и положительно определенная, λ_0 может быть определено по формуле:

$$\lambda_0 = \min_{a \in M_{k,1}} \frac{a^T G a}{a^T a}.$$

Делая в G подстановку $XX^T = E - P_X$, с помощью (3.4) находим:

$$\lambda_0 = 1 - \max_a \frac{\|P_X X_0 a\|^2}{\|X_0 a\|^2} = \cos^2 \varphi.$$

Заметим, что $\cos\varphi \neq 0$ в силу невырожденности матрицы $X_0^T X$. Таким образом, справедлива оценка $\|S_0\| \le \sqrt{k} / \cos\varphi$, откуда немедленно следует требуемое неравенство.

Теорема доказана.

Как показывает (3.6), при достаточно малой среднеквадратичной погрешности начальных данных процесс (3.5) позволяет достичь приемлемой точности уже на первом шаге. В частности, если исходная матрица k-разложима ($\delta = 0$), равенство $Z = X_1 Y_1^T$ выполняется точно для любого X_0, удовлетворяющего условию $\mathrm{rg} X_0^T X = k$.

Формулы (3.5) можно видоизменить так, чтобы исключить вычисление обратных матриц:

(3.7) $$Y_r = Z^T X_{r-1} B_r, \quad X_r = Z Y_r, \; r = 1, 2, \ldots,$$

где невырожденная $k \times k$-матрица B_r подбирается так, чтобы столбцы матрицы Y_r были ортонормированы. При этом, как легко убедиться, произведение $X_r Y_r^T$ не изменится.

Пусть, например, умножение справа на B_r означает применение ортогонализации Грама-Шмидта [3, 8] к столбцам матрицы $Z^T X_{r-1}$. Тогда становится возможным изменить порядок вычислений, чередуя итерации отдельных столбцов матриц X_r и Y_r с добавлением к ним новых столбцов, и определять число k из условия достижения заданной точности. Опишем этот алгоритм подробнее.

Заметим, что при достаточно малых δ в силу теоремы 2.1 начальным приближением для (3.7) может послужить матрица X_0, составленная из линейно независимых столбцов Z. Возьмем, например, в качестве первого столбца X_0 столбец Z с максимальной нормой. Поскольку матрицы B_r – верхние треугольные, зная лишь первый столбец X_0, уже можно вычислить первые столбцы Y_1, X_1, Y_2, X_2, и т. д. Проведя N итераций, получим первый столбец x_N^1 искомой матрицы X_N. В качестве второго столбца начального приближения X_0 возьмем тот столбец Z, расстояние от которого до $\mathscr{L}(x_N^1)$ максимально. Аналогичным образом проведя несколько итераций, получим вектор x_N^2. Третий столбец X_0 также выберем из матрицы Z, но уже из условия максимума расстояния до $\mathscr{L}(x_N^1, x_N^2)$. И так далее, пока не окажется, что найдена матрица $X_N = \left[x_N^1 \ldots x_N^q\right]$, для которой $\left\|P_{X_N} Z\right\|$ меньше заданного

числа *d*. Запишем данный алгоритм в удобном для машинной реализации виде.

Алгоритм 3.1.

Вход: $Z \in M_{n,m}$; $d \ge 0$; число итераций $N > 0$.

Выход: $q > 0$; $X_N \in M_{n,q}$, $Y_N \in M_{m,q}$: $\left\| X_N Y_N^T - Z \right\| \le d$, $X_N^T X_N = E$.

Обозначения: $n \times m$-матрицы $Z_t = \left[z_t^1 \dots z_t^m \right]$, $t = 1, \dots, q$;

$n \times q$-матрицы $X_r = \left[x_r^1 \dots x_r^q \right]$, $r = 1, \dots, N$;

$m \times q$-матрицы $Y_r = \left[y_r^1 \dots y_r^q \right]$, $r = 1, \dots, N$;

функция $\operatorname{norm}(x) = x / \|x\|$.

1. $t := 1$; $Z_1 := Z$;
2. $j_t := \arg \max\limits_{1 \le j \le m} \left\| z_t^j \right\|$;
3. для $r = 1, \dots, N$ повторить:

$$\tilde{x}_r^t := \begin{cases} z_1^{j_t}, & r = 1, \\ Z y_{r-1}^t, & r > 1; \end{cases}$$

$$x_r^t := \operatorname{norm}\left(\tilde{x}_r^t - \sum_{s=1}^{t-1} x_r^s (x_r^s)^T \tilde{x}_r^t \right); \quad y_r^t := Z^T x_r^t;$$

4. вычислить $Z_{t+1} := Z_t - x_N^t (y_N^t)^T$;
5. если $\left\| Z_{t+1} \right\| > d$, то $t := t + 1$ и повторить 2-5;
6. иначе положить $q := t$.

Данный алгоритм не гарантирует, что найденная пара матриц (X_N, Y_N) будет *q,d*-разложением *Z*, так как в общем случае условие минимальности числа столбцов может не выполняться. Тем не менее, справедлива следующая

Теорема 3.3. *Пусть выполнены все условия теоремы 3.2. Пусть, кроме того,* $\delta = \left\| Z - XY^T \right\|$ *и* $\mu < 1$. *Тогда для любого* ε, $0 < \varepsilon^2 < \lambda_k$ *существует натуральное число* N_0 *такое, что при* $d^2 = \delta^2 + \varepsilon^2$ *и* $N \ge N_0$ *вычисленная алгоритмом 3.1 пара матриц будет k,d-разложением Z.*

Доказательство. Легко проверить, что условие окончания итераций (шаг 5) может быть записано в виде $\left\| Z - X_N Y_N^T \right\| \le d$.

Покажем, что число столбцов q матриц X_N и Y_N не может быть меньше *k*. Допустим противное. Тогда матрица *Z* была бы *p,d*-разложимой, $p \le q < k$, и выполнялось бы (см. теорему 3.1) неравенство

$$\lambda_1 + \dots + \lambda_p \ge \|Z\|^2 - d^2.$$

С другой стороны, поскольку $\delta = \left\| Z - XY^T \right\|$,

$$\lambda_1 + \dots + \lambda_p + \dots + \lambda_k = \|Z\|^2 - \delta^2.$$

Вычитая первое соотношение из второго, получим:

$$\lambda_{p+1}+\ldots+\lambda_k \le d^2-\delta^2=\varepsilon^2,$$

что противоречит условию $\varepsilon^2<\lambda_k$, следовательно, $q \ge k$.

Воспользовавшись оценкой (3.6), легко проверить, что если

$$N \ge 1+\left(\ln\mu\right)^{-1}\ln\left(\frac{\cos\varphi}{\sqrt{k}}\left(\sqrt{1+\varepsilon/\delta^2}-1\right)\right),$$

то условие окончания итераций будет выполняться при $q=k$, значит, пара матриц (X_N, Y_N) будет k,d-разложением Z.

Теорема доказана.

Итак, при подходящем выборе параметров алгоритма N и d удается определить одно из k,d-разложений Z. Отметим, что при $k << \min(n,m)$ этот алгоритм представляет собой эффективный метод сжатия исходных данных, так как позволяет заменить матрицу Z парой матриц существенно меньших размеров.

§4. Оптимальные *k*-разложения

Для получения единственного структурного описания заданной допустимой поверхности необходимо решать задачу регуляризации, накладывая дополнительные ограничения на множество непроизводных поверхностей. Привлекаемая для этого априорная информация должна удовлетворять двум требованиям. Ее использование должно быть обоснованным в рамках решаемой задачи и ее должно быть достаточно, чтобы отбросить бесконечное множество решений уравнения (2.2).

В данном случае выбор дополнительной информации был продиктован тем, что форма хроматографических пиков, определяемая протекающими в колонке физическими процессами, до известной степени одинакова (см. рис. 1б). Ниже будет показано (см. §5), что истинные хроматограммы можно довольно точно аппроксимировать элементами параметрического семейства функций.

Далее, в отличие от §2, *непроизводной поверхностью* будем называть всякую функцию двух переменных вида $f(x)g(y)$, где $f(x)$ и $g(y)$ определены на $\Omega_x \subseteq \mathbb{R}$ и $\Omega_y \subseteq \mathbb{R}$ соответственно, и $f(x)$ принадлежит заданному семейству функций $\mathscr{F}$.

При фиксированном наборе n чисел $\xi_i \in \Omega_x$, $i=1,\ldots,n$ семейство $\mathscr{F}$ индуцирует множество вектор-столбцов

$$\mathscr{F}_n=\left\{[f(\xi_1)\ldots f(\xi_n)]^T : f \in \mathscr{F}\right\}$$

и множество матриц размера $n \times k$ со столбцами из F_n

$$\mathscr{F}_{n,k}=\left\{[u_1\ldots u_k] : u_s \in \mathscr{F}_n,\ s=1,\ldots,k\right\}.$$

Пусть (X,Y) — известное k-разложение матрицы Z, и пусть для определенности столбцы X ортонормированы, $X^TX=E$. Задача состоит

в том, чтобы выяснить, каким условиям должно удовлетворять семейство $\mathscr{F}$, чтобы существовала единственная (с точностью до перестановки столбцов) матрица $U \in \mathscr{F}_{n,k}$ такая, что $\mathscr{L}(X) = \mathscr{L}(U)$, а также найти способ вычисления этой матрицы.

Определение. Семейство функций $\mathscr{F}$ называется *k-нелинейным* по набору точек $(\xi_1,\ldots,\xi_n)$, если для любых попарно различных вектор-столбцов $u_0, u_1,\ldots,u_k$ из множества $\mathscr{F}_n$ и любых действительных $a_1,\ldots,a_k$ выполняется условие $a_1u_1+\ldots+a_ku_k \neq u_0$.

Семейства функций, обладающие указанным свойством, существуют. Например, параметрическое семейство

$$\mathscr{F}^{\alpha} = \left\{ e^{-(x-\alpha)^2} : \alpha \in \mathrm{R} \right\}$$

k-нелинейно по любому набору из n различных точек для всех $k = 1,\ldots,n-1$. Для доказательства составим $n \times (k+1)$-матрицу из произвольных попарно различных вектор-столбцов $u_0, u_1,\ldots,u_k$, принадлежащих множеству $\mathscr{F}_n^{\alpha}$. В силу тождества $e^{-(x-\alpha)^2} = e^{-x^2}e^{2x\alpha}e^{-\alpha^2}$ первые ее (k+1) строк образуют квадратную матрицу, полученную умножением строк и столбцов обобщенной матрицы Вандермонда [3] на положительные числа. Опрелелитель последней отличен от нуля. Следовательно вектор-столбцы $u_0, u_1,\ldots,u_k$ линейно независимы, а указанное семейство, в силу их произвольности, k-нелинейно. Заметим, что данное семейство функций может рассматриваться как предельно упрощенная модель хроматографического пика.

В то же время очевидно, что всякое линейное семейство функций (такое что для любых $f_1(x), f_2(x) \in \mathscr{F}$ и $\alpha_1, \alpha_2 \in \mathbb{R}$ выполняется $\alpha_1 f_1(x) + \alpha_2 f_2(x) \in \mathscr{F}$), например, множество полиномов заданной степени, не может быть k-нелинейным.

Определение. Семейство функций $\mathscr{F}$ называется *k-точным* по набору точек $(\xi_1,\ldots,\xi_n)$ для данной матрицы $X \in M_{n,k}$ ранга k, если существует невырожденная $k \times k$-матрица A такая, что $XA \in \mathscr{F}_{n,k}$.

Теорема 4.1. *Если $\mathscr{F}$ — семейство функций, k-нелинейное и k-точное по набору точек $(\xi_1,\ldots,\xi_n)$ для данной матрицы X, то существует и единственна (с точностью до перестановки столбцов) матрица $U_* \in \mathscr{F}_{n,k}$ такая, что $\mathscr{L}(U_*) = \mathscr{L}(X)$.*

Доказательство. В силу k-точности семейства $\mathscr{F}$ в подпространстве $\mathscr{L}(X)$ найдутся k линейно независимых вектор-столбцов $u_1^*,\ldots,u_k^*$, принадлежащих $\mathscr{F}_n$. В силу k-нелинейности других элементов $\mathscr{F}_n$ в $\mathscr{L}(X)$ быть не может. Следовательно те и только те матрицы $U \in F_{n,k}$ удовлетворяют условию $\mathscr{L}(X) = \mathscr{L}(U)$, которые составлены из вектор-столбцов $u_1^*,\ldots,u_k^*$, взятых в произвольном порядке. Теорема доказана.

Из доказательства ясно, что k-нелинейность обеспечивает единственность, а k-точность — существование регуляризованного решения.

Заметим, что на практике гарантировать точное совпадение подпространств $\mathscr{L}(X) = \mathscr{L}(U)$ невозможно по многим причинам: из-за погрешностей исходных данных, погрешностей вычисления матрицы X и, самое существенное, неточности описания истинных хроматограмм элементами семейства $\mathscr{F}$. Ослабим поэтому требование k-точности: будем искать матрицу $U_* \in \mathscr{F}_{n,k}$, доставляющую минимум функционалу

$$\bar{J}(U) = \left\| P_U X \right\|^2,$$

который обращается в нуль тогда и только тогда, когда $\mathscr{L}(X) = \mathscr{L}(U)$. Таким образом, для k-точных семейств его минимизация должна приводить к искомому решению.

Ортонормируем столбцы матрицы $U \equiv \left[u_1 \dots u_k\right]$, применив к ним процесс Грама-Шмидта, и обозначим полученную матрицу $\tilde{U} \equiv \left[\tilde{u}_1 \dots \tilde{u}_k\right]$. Определим последовательность проекционных матриц

$$P_0 = E_n;$$
$$P_s = P_{\tilde{u}_1} \dots P_{\tilde{u}_{s-1}} \equiv E - \tilde{u}_1 \tilde{u}_1^T - \dots - \tilde{u}_{s-1} \tilde{u}_{s-1}^T;\ s = 2, \dots, k.$$

Тогда

$$\tilde{u}_s = \frac{P_s u_s}{\left\| P_s u_s \right\|},\ s = 1, \dots, k$$

Вследствие элементарных свойств евклидовой матричной нормы

$$\bar{J}(U) = \left\| P_X \tilde{U} \right\|^2 = \sum_{s=1}^{k} \frac{\left\| P_X P_s u_s \right\|^2}{\left\| P_s u_s \right\|^2}.$$

Заметим, что при наличии k-точности $P_X P_s = P_X$, поэтому функционал $\bar{J}(U)$ можно заменить оценкой сверху

$$J(U) = \sum_{s=1}^{k} \frac{\left\| P_X u_s \right\|^2}{\left\| P_s u_s \right\|^2}, \tag{4.1}$$

не нарушив при этом требования $J(U) = 0 \Leftrightarrow \mathscr{L}(X) = \mathscr{L}(U)$. Введем обозначения для слагаемых в (4.1)

$$J_s(u) \equiv J_s(u_1, \dots, u_{s-1}, u) \equiv \frac{\left\| P_X u \right\|^2}{\left\| P_s u \right\|^2},\ s = 1, \dots, k \tag{4.2}$$

и рассмотрим последовательность задач минимизации

$$u_s^* = \arg \min_{u \in F_n} J_s(u_1^*, \dots, u_{s-1}^*, u),\ s = 1, \dots, k. \tag{4.3}$$

Если семейство $\mathscr{F}$ удовлетворяет условиям теоремы 4.1, то последовательная минимизация (4.3) приводит к искомому минимуму

функционала $J(U)$. Действительно, в силу k-точности существуют различные векторы $u_1^*,\dots,u_k^*$, доставляющие нулевые значения функционалам $J_1,\dots,J_k$ соответственно. В силу k-нелинейности эти векторы линейно независимы и составляют единственную (с точностью до перестановки столбцов) матрицу U_*, обращающую $J(U)$ в нуль.

Таким образом, минимизацию $J(U)$ по множеству $\mathscr{F}_{n,k}$ удается свести к существенно более простой задаче последовательной минимизации функционалов $J_1,\dots,J_k$ по множеству $\mathscr{F}_n$.

В общем случае, когда минимум $J(U)$ не равен нулю (нет k-точности), можно говорить лишь о его приближеной минимизации с помощью (4.3). Ясно однако, что чем лучше семейство $\mathscr{F}$ описывает функции $f(x)$, тем ближе к нулю минимум функционала $J(U)$, и тем ближе вычисленная с помощью (4.3) матрица U_* к матрице, минимизирующей $J(U)$. Независимо от того, выполняется условие k-точности или нет, можно ввести понятие оптимального k-разложения.

Определение. Назовем *оптимальным k-разложением* пару матриц $(X_0,Y_0)\in M_Z^k$, доставляющую минимум функционалу

$$\Phi(\tilde X,\tilde Y)=\left\|\tilde X-U_*\right\|,\ (\tilde X,\tilde Y)\in M_Z^k,$$

где U_* — матрица, вычисленная согласно (4.3).

Минимизируя Ф по множеству M_Z^k и используя лемму 2.2, легко получить формулы для расчета оптимального k-разложения:

$$X_0=XX^TU_*\ ,\ \ Y_0=Y(U_*^TX)^{-1}.$$

Рассмотрим частные случаи, конкретизируя вид семейства $\mathscr{F}$.

1. В задаче идентификации химических веществ в качестве регуляризирующего семейства функций можно вводить конечный набор истинных хроматограмм (или истинных спектров) веществ, включенных в обучающую выборку. В таком случае минимизация каждого из функционалов J_s сводится к линейному перебору элементов конечного множества функций. Причем, решив задачу предобработки, мы автоматически получаем решение задачи распознавания.

Такой подход, однако, имеет целый ряд недостатков. Он исключает возможность выделения спектра неизвестного вещества и приводит к неоправданному увеличению объема обучающей выборки, требуя при ее формировании слишком большого числа экспериментов с целью получения «чистых» изолированных пиков, от чего хотелось бы отказаться.

2. Пусть теперь $\mathscr{F}=\left\{f(x,\alpha):\alpha\in\mathbb{R}^p\right\}$ — параметрическое семейство функций, f дифференцируема по параметрам.

Опишем метод линеаризации для решения каждой из задач (4.3). Пусть векторы $u_1^*,\dots,u_{s-1}^*$ уже найдены и требуется определить вектор $u_s^*\in\mathscr{F}_n$, доставляющий минимум функционалу (4.2). Пусть α^0 —

начальное приближение вектора параметров. Предположим, что r-ое приближение α^r при некотором $r \geq 0$ уже известно. Обозначим

$$u^r = \left[f(\xi_i, \alpha^r) \right]_{n\times 1};$$

$$G = \left[\frac{\partial f(\xi_i, \alpha^r)}{\partial \alpha_q^r} \right]_{n\times p};$$

$$d = \alpha^{r+1} - \alpha^r.$$

Тогда в линейном приближении $u^{r+1} \approx u^r + Gd$. В качестве $r+1$-ого приближения a^{r+1} возьмем решение задачи минимизации дробно-квадратичного функционала

$$J_s(u^r + Gd) = \frac{\left\| P_X(u^r + Gd) \right\|^2}{\left\| P_s(u^r + Gd) \right\|^2}. \tag{4.4}$$

Приравнивая нулю производную по d, получим систему уравнений

$$\begin{cases} d = -\left[G^T P_X G \right]^{-1} \left(G^T P_X u^r - \mu G^T P_s (u^r - Gd) \right); \\ \mu = \dfrac{\left\| P_X(u^r - Gd) \right\|^2}{\left\| P_s(u^r - Gd) \right\|^2} \end{cases} \tag{4.5}$$

Для решения этой системы воспользуемся методом простой итерации. Положим на 0-ом шаге $\mu = 0$. Найдем из первого уравнения d и, подставив его во второе, определим μ. Будем повторять указанную процедуру, пока значения d и μ не перестанут существенно изменяться.

Описанный метод решения задач (4.3) хорошо зарекомендовал себя на практике, однако вопрос о теоретическом обосновании его сходимости остается открытым. Неясно также, всегда ли он позволяет избежать на s-ом шаге сходимости к одному из векторов $u_1^*, \ldots, u_{s-1}^*$, найденных на предыдущих шагах. Ниже приведем лишь некоторые свойства функционалов (4.2), предполагая, что выполняются условия теоремы 4.1.

1. Функционал $J_s(u)$ принимает значения из отрезка $[0,1]$, обращаясь в нуль ровно в $k+1-s$ точках $u_{s,}^*, \ldots, u_k^*$, непрерывен всюду за исключением точек $u_t^*, t < s$, в которых он не определен. Причем, в общем случае не существует предела $J_s(u)$ при $u \to u_t^*$.

2. Пусть $G_* = \left[\dfrac{\partial f(\xi_i, \alpha^*)}{\partial \alpha_q^*} \right]_{n\times p}$ — матрица производных в точке $u_t^* = \left[f(\xi_i, \alpha^*) \right]_{n\times 1}$, $t < s$. Тогда, линеаризуя функционал $J_s(u)$ в окрестности u_t^*, получим: $J_s(u) \to \dfrac{\left\| P_X G_* d \right\|^2}{\left\| P_s G_* d \right\|^2}$ при $d = \alpha - \alpha^* \to 0$. Легко

видеть, что если $P_X G_* = G_*$, то существует предел $\lim\limits_{u \to u_t^*} J_s(u) = 1$, то есть в точке u_t^* имеется выколотый максимум.

3. Если семейство $\mathscr{F}$ однопараметрическое, то существует $\lim\limits_{u \to u_t^*} J_s(u)$.

§5. Алгоритмы предварительной обработки данных

Согласно постановке задачи исходные данные представлены в виде матрицы чисел размера 1920×112, которая рассматривается как трехмерная поверхность $z(x, y)$, заданная в узлах прямоугольной сетки. Задачу предобработки предлагается решать в четыре этапа:

1) подготовительный этап: выделение несущей поверхности и оценка амплитуды шума;

2) разбиение поверхности на фрагменты, содержащие относительно обособленные группы перекрывающихся пиков и отсечение «ровных» участков;

3) вычисление k, δ-разложения каждого из фрагментов;

4) вычисление оптимальных k-разложений при заданном регуляризирующем семействе функций специального вида.

В результате должны получить сжатое описание исходной поверхности, составленное из описаний отдельных пиков.

Будем придерживаться следующих соглашений и обозначений: время t и длина волны w измерены в таких единицах, что их целочисленные значения совпадают с номерами строк и столбцов соответственно исходной матрицы начальных данных Z; переменные x и t тождественны; переменные y и w тождественны.

Подготовительный этап предобработки. Прежде всего необходимо устранить систематическую погрешность, накладываемую на спектры веществ однородным потоком подвижной фазы (элюента). Кроме того, применение алгоритмов k, δ-разложения требует предварительной оценки среднеквадратичного шума. Предлагается довольно простой алгоритм, позволяющий решить сразу обе эти задачи.

Алгоритм основан на предположении, что любая хроматограмма имеет хотя бы один достаточно продолжительный «ровный» участок, не содержащий пиков. Как показывает опыт, на реальных хроматограммах это требование выполняется всегда.

Выделив такие участки и усреднив по ним значение $z(t, w)$ при каждой фиксированной длине волны w, получим спектр элюента $g_э(w)$. Теперь, чтобы снять с исходной поверхности систематическое возмущение, достаточно вычесть из нее цилиндрическую несущую поверхность с образующей $g_э(w)$:

$$z_{невозм}(t, w) = z(t, w) - g_э(w).$$

Среднеквадратичный шум δ легко оценивается при усреднении участков постоянства.

Выделение участков постоянства на хроматограмме $z(t, w)$ при фиксированном w можно производить согласно простому пороговому решающему правилу: если на интервале длиной не менее $T_{\max}$ функция отклоняется от своего среднего значения более чем на $d_{\max}$ менее $N_{\max}$ раз, то интервал считать участком постоянства. Если при фиксированном w не обнаружено ни одного участка постоянства,

следует увеличить порог $d_{\max}$ и повторить поиск участков постоянства. Опытным путем были найдены подходящие значения параметров: $T_{\max} = 40$, $d_{\max} = 10$, $N_{\max} = 1$. Алгоритм продемонстрировал неплохую устойчивость: при варьировании этих параметров в достаточно широких пределах удавалось во всех случаях (около 20 файлов исходных данных, взятых наугад) надежно выделять несущую поверхность.

Выделение массивов. Наличие продолжительных участков постоянства на хроматограммах, а также большая размерность матрицы исходных данных лишают смысла попытку разложения этой матрицы целиком. Более рационально разбить исходную поверхность на фрагменты, содержащие относительно обособленные группы пиков (назовем эти фрагменты *массивами*), а участки постоянства вовсе исключить из рассмотрения. Расположение *i*-ого массива будем характеризовать отрезком оси времени $\left[T_1^i, T_2^i\right]$. По оси длин волн массивы будут охватывать весь допустимый интервал.

Задача выделения массивов при кажущемся сходстве с задачей выделения несущей поверхности приводит к построению более «тонкого» алгоритма, поскольку приходится учитывать многие, часто противоречивые требования. Так, например, оказалось, что участки постоянства могут несколько отличаться по высоте (разность высот доходит иногда до 10δ), в то же время, хотелось бы идентифицировать отдельные пики высотой порядка 4δ и выше. Случайные выбросы, носящие характер шума, не должны быть выделены как массивы. Слишком длинные массивы целесообразно некоторым образом разбить на подмассивы, чтобы избежать решения задачи разложения слишком большой размерности. В то же время массивы должны быть по возможности более обособленными, и так далее.

Предлагается следующий алгоритм, позволяющий найти приемлемый компромисс между подобного рода требованиями и добиться надежного выделения массивов.

1. Используется кусочно-линейная аппроксимация нескольких (от 3 до 10) хроматограмм, взятых при заранее фиксированных значениях *w*. Аппроксимация строится так, чтобы погрешность на каждом участке не превышала δ.

2. Отбрасываются участки постоянства с коэффициентом наклона, по модулю не превышающим $a_{\max}$ и средним значением, не превышающим $Z_{\min}$. Остальные участки объявляются массивами.

3. Отбрасываются массивы продолжительностью менее $T_{\min}$ (шумовые выбросы), а также заканчивающиеся при $t < T_0$ (неинформативный начальный участок хроматограммы).

4. Массивы продолжительностью более $T_{\max}$ разбиваются на подмассивы по точкам локальных минимумов (смена знака коэффициента наклона с «−» на «+»). Разбиение осуществляется по той точке минимума, которая оптимизирует априори заданную штрафную функцию. Вид штрафной функции подбирается с тем расчетом, чтобы разбиения, ближайшие к середине массива имели больший приоритет (дробление не должно быть слишком мелким). Указанное правило

разбиения применяется к подмассивам до тех пор, пока продолжительность каждого из них не станет меньшей $T_{\max}$.

Как показали вычислительные эксперименты (около 20 исходных файлов), данный алгоритм позволяет добиться полного соответствия с «интуитивно наилучшим» разбиением при следующих значениях параметров: $a_{\max} = 0.5$, $Z_{\min} = 25$, $T_{\min} = 25$, $T_{\max} = 300$, $T_0 = 200$.

Разложение массивов. Полученные на предыдущем шаге фрагменты исходной поверхности (массивы) заданы матрицами, к которым далее применяется алгоритм 3.1 вычисления k,d-разложения. Как показали вычислительные эксперименты, целесообразно выбирать число итераций $N = 2$ и $d = (1.2 \div 1.8)\delta$, где δ — величина шума. Смысл последней рекомендации становится ясным из теоремы 3.3.

Уравнение хроматографического пика. Один из способов построения регуляризирующего семейства функций $\mathcal{F}$ состоит в рассмотрении физической модели протекающих в колонке процессов и получении на ее основе параметрических уравнений пика.

В простейшей классической модели [4] концентрация C адсорбата в подвижной фазе описывается гладкой функцией времени t и расстояния x вдоль направления движения потока. Функция $C(t,x)$ удовлетворяет дифференциальному уравнению

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - v(C)\frac{\partial C}{\partial x}, \tag{5.1}$$

описывающему два основных физических процесса: диффузию частиц адсорбата в элюенте (D — коэффициент диффузии) и перенос частиц адсорбата с переменной скоростью $v(C)$, зависящей от концентрации. При некоторых предположениях [4] эта зависимость описывается формулой

$$v(C) = \frac{1}{\Gamma(C)\omega_1 + \omega_2},$$

где ω_1, ω_2 — константы, не зависящие от концентрации, $\Gamma(C)$ — коэффициент Генри. Существует несколько подходов к оценке коэффициента Генри, отличающихся совокупностью предположений относительно однородности адсорбционных мест на поверхности адсорбента, взаимодействия адсорбированных молекул и толщине слоя адсорбированных молекул:

Таблица 5.1

Автор	Формул	однород-ность	взаимо-действие	толщина слоя
Лангмюр	$\Gamma(C) = \frac{b_1}{1 + b_2 C}$	да	нет	моно
Темкин	$\Gamma(C) = \frac{b_1}{C}\ln\frac{1 + b_2 C}{1 + b_3 C}$	нет	нет	моно

Фрейндлих	$\Gamma(C) = b_1 C^{b_2}$	нет	нет	моно
Фрумкин	$\Gamma(C) = \dfrac{b_1 e^{b_4 \Gamma(C)}}{1 + b_2 e^{b_4 \Gamma(C)} C}$	да	да	моно
БЭТ	$\Gamma(C) = \dfrac{b_1}{(1 - b_2 C)(1 + b_3 C)}$	да	нет	поли
	$\Gamma(C) = \dfrac{b_1}{1 + b_2 C} + \dfrac{b_3}{1 + b_4 C}$	нет	нет	моно

Здесь b_1, b_2, b_3, b_4 — константы.

Решение уравнения (5.1) при начальном условии $C(0, x) = C_0 \delta(x)$, где $\delta(x)$ — δ-функция, и фиксированном $x = L_0$, где L_0 — длина колонки, может быть представлено в виде [2]:

$$C(t) = \frac{k_1}{\sqrt{t}} \exp\left\{ -\frac{1}{k_2 t} \left(k_3 - t v(C) \right)^2 \right\}, \tag{5.2}$$

где k_1, k_2, k_3 — коэффициенты, не зависящие от концентрации и времени.

Формула (5.2) не позволяет выразить $C(t)$ в явном аналитическом виде. Для вычисления $C(t)$ используются итерационные численные методы поиска корня функции на отрезке [5]. Для формирования вектора значений функции вида (5.2) в ряде близко отстоящих точек более эффективным оказалось численно решать задачу Коши (дифференциальное уравнение для $C(t)$ можно получить, продифференцировав (5.2) по t).

При фиксированном k_1 (например, равном 1) полученное параметрическое семейство функций становится k-нелинейным. Условие k-точности в общем случае не выполняется, поскольку ни одна из формул, приведенных в таблице 5.1 не позволяет идеально точно описывать реальные хроматографические пики. Тем не менее, как показали вычислительные эксперименты по аппроксимации изолированных пиков, в семействе (5.2) всегда находились функции, приближающие пики со среднеквадратичной погрешностью, составляющей не более 2 - 6% от высоты пика. Таким образом, полученное семейство функций позволяет находить оптимальные k-разложения и, соответственно, набор спектров и хроматограмм, достаточно близких к истинным.

Вычисление оптимального k-разложения массива $\left[T_1^i, T_2^i \right]$ проводится по схеме (4.2-5). Основную сложность при этом представляют выбор хороших начальных приближений для метода линеаризации (4.4-5) и аппроксимация пиков, по которым прошла граница раздела массивов.

Для поиска начальных приближений отбирается некоторое количество q хроматограмм при заранее фиксированных значениях длины волны (соответствующую $n \times q$-матрицу обозначим через X_0).

Пусть аппроксимируется первый пик, $s = 1$. На отобранных хроматограммах найдем наибольший локальный максимум из числа тех, которые удается надежно аппроксимировать вогнутой параболой. По коэффициентам параболы рассчитаем начальные приближения для параметров k_2 и k_3 семейства функций (5.2). Остальные параметры, входящие в функцию $v(C)$, зададим так, чтобы полученная кривая имела наиболее характерную для хроматографических пиков несимметричную форму.

Пусть, отправляясь из найденного начального приближения, удалось определить вектор u_1^*, доставляющий минимум функционалу $J_1(u)$. Возникает следующая задача: хотелось бы вычесть найденный пик из хроматограмм X_0 с тем, чтобы в дальнейшем определять начальные приближения для других пиков. Однако, как легко показать, точно вычесть пик можно лишь тогда, когда известны все остальные.

Воспользуемся следующим эвристическим приемом. После каждого s-ого шага будем вычитать из хроматограмм X_0 все найденные пики:

$$X_s = X_0 - \left[u_1^* \dots u_s^*\right] A, \quad A \in M_{s,q},\ s = 1, \dots, k,$$

определяя матрицу A из условия минимума нормы $\|X_s\|^2$. Решение этой задачи легко выразить через ортонормированные векторы $\tilde{u}_1^*, \dots, \tilde{u}_s^*$:

$$X_s = X_0 - \tilde{u}_1^* (\tilde{u}_1^*)^T X_0 - \dots - \tilde{u}_s^* (\tilde{u}_s^*)^T X_0.$$

При расчетах будем пользоваться более простой рекуррентной формулой

$$X_s = X_{s-1} - \tilde{u}_s^* (\tilde{u}_s^*)^T X_{s-1}.$$

Как показали вычислительные эксперименты, указанный способ корректировки начальной матрицы хроматограмм X_0 позволяет получать приемлемые начальные приближения для метода линеаризации (4.4-5).

Для аппроксимации пиков, по которым прошла граница раздела, используется двухпроходная схема перебора массивов.

На первом проходе в каждом i-ом массиве с границами $\left[T_1^i, T_2^i\right]$ аппроксимируются лишь те пики, у которых на этом отрезке удается выделить локальный максимум. Для каждого пика вычисляется носитель — отрезок времени, на котором значение функции превышает уровень шума.

На втором проходе для каждого массива число носителей, пересекающихся с отрезком $\left[T_1^i, T_2^i\right]$, сравнивается с числом k, найденным алгоритмом 3.1. Если первое оказывается меньшим, аппроксимация пиков на отрезке продолжается согласно прежнему алгоритму, но с учетом всех пиков, носители которых перекрывают отрезок $\left[T_1^i, T_2^i\right]$.

§6. Идентификация химических веществ

По завершении предварительной обработки исходных данных идентификация (распознавание) веществ становится относительно простой задачей. Каждое вещество характеризуется присущими только ему истинным спектром и временем удерживания. Остальные параметры хроматограммы в значительной степени определяются характеристиками колонки и поэтому не используются для идентификации.

Время удерживания данного вещества может быть легко рассчитано по коэффициентам уравнения пика. Для сравнения спектров достаточно вычислить их среднеквадратическое или максимальное отклонение. Возможно также построение других метрик, в том числе параметрических.

Далее строится параметрическое семейство метрик, учитывающее при сравнении пиков как спектры, так и времена удерживания. Для выбора оптимальной метрики, обеспечивающей корректное распознавание, должны быть использованы методы алгебраического подхода к обучению распознаванию образов [6-7].

В целях тестирования описанных алгоритмов была разработана специализированная программа, позволяющая:

— считывать произвольные участки допустимой поверхности из файла исходных данных или файла-описания и выводить их на экран в виде двумерных и трехмерных графиков;
— выделять несущую поверхность;
— разбивать исходную поверхность на массивы;
— аппроксимировать изолированные пики элементами нескольких параметрических семейств функций;
— вычислять произвольные и оптимальные разложения массивов и составлять из них описание всей поверхности;
— осуществлять преобразование файла исходных данных в более компактный (5 - 15 кбайт) файл-описание оптимального разложения.

Программа разработана для IBM PC AT под MS DOS 5.0 на языке программирования Turbo Pascal 6.0.

Список литературы

[1] Васильев Ф.П. Численные методы решения экстремальных задач. М.: Наука, 1988. 549 с.

[2] Владимиров В.С. Уравнения математической физики. М.: Наука, 1971. 509 с.

[3] Гантмахер Ф.Р. Теория матриц. М.: Наука, 1988.

[4] Гольберт К.А., Вигдергауз М.С. Курс газовой хроматографии. М.: Химия, 1974. 407 с.

[5] Демидович Б.П., Марон И.А. Основы вычислительной математики. М.: Физматгиз, 1963. 659 с.

[6] Журавлев Ю.И. Об алгебрическом подходе к решению задач распознавания или классификации. // Проблемы кибернетики. 1979. Вып.33. С.5-68.

[7] Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. I-III. // Кибернетика. 1977. 4. С.14-21. 1977. 6. С.21-27. 1978. 2. С.35-43.

[8] Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. М.: Наука, 1986.

[9] Adams A.K., Essien H., Binder S.R. // Ann. Biol. Clin. 1991. V. 49. P. 291-297.

[10] Binder S.R., Regalia M., Biaggi-McEachern M., Mazhar M. // J. Chromatogr. 1989. V. 473. P. 325.

[11] Fell A.F., Clark B.J., Scott H.P. // J. Chromatogr. 1984. V. 316. P. 423.

[12] Воронцов К.В. Предварительная обработка данных для решения специального класса задач распознавания // ЖВМиМФ (статья находится в печати).

[13] Воронцов К.В., Кобзева Н.В. Численные методы обработки данных для высокоэффективной жидкостной хроматографии // В сб. «Информационные проблемы клинической токсикологии», 1993.