УДК 543

Хемометрический подход к исследованию больших массивов химических данных

О. Е. Родионова

ОКСАНА ЕВГЕНЬЕВНА РОДИОНОВА — кандидат физико-математических наук, старший научный сотрудник Института химической физики им. Н.Н. Семенова РАН. Область научных интересов: математическое моделирование, вычислительная математика, хемометрика, спектроскопия, математическая статистика.

119991 Москва, ул. Косыгина, 4, Институт химической физики им. Н.Н. Семенова РАН, E-mail rcs@chph.ras.ru

Введение

История становления хемометрики и ее место в системе знаний

Хемометрика — это синтетическая дисциплина, находящаяся на стыке химии и математики. Как самостоятельное научное направление внутри аналитической химии она возникла в 1974 в г.Сиэтле, США [1]. У ее истоков стояли американец Брюс Ковальски (B. Kowalski) и швед Сванте Волд (S. Wold) — внук великого ученого Сванте Аррениуса. Как это часто бывает, хемометрика до сих пор не имеет общепризнанного определения. Наиболее популярное принадлежит Д. Массарту (D. Massart) [2]: *хемометрика — это химическая дисциплина, применяющая математические, статистические и другие методы, основанные на формальной логике, для построения или отбора оптимальных методов измерения и планов эксперимента, а также для извлечения наиболее важной информации при анализе экспериментальных данных.*

Традиционно математические модели в химии строились так, чтобы в математической форме выразить те или иные законы химии и физики. Однако с совершенствованием и усложнением эксперимента появилась необходимость анализа очень больших массивов данных. В то же время всегда существовала необходимость моделирования, хотя бы в ограниченной области, таких процессов и зависимостей, которые не поддаются содержательному математическому описанию из-за сложности происходящих процессов или их неизученности.

Хемометрика использует по большей части «простые» эмпирические модели и с этой точки зрения является областью формального моделирования (поэтому та часть определения проф. Массарта, которая указывает на формальную логику, весьма существенна). Еще одной важной особенностью хемометрики, в отличие от «традиционного» подхода, является рассмотрение многофакторных данных в совокупности, без выделения предварительно одного или нескольких «существенных» факторов или переменных. В результате построенные многомерные эмпирические модели позволяют обнаружить новые связи или явления, так как учитывают скрытые объективно существующие связи внутри объекта.

Подобный подход отпугивает многих исследователей, считающих формальное моделирование поверхностным. Однако при этом не обращается внимание на то, что основной содержательный компонент исследования вовсе не исключается, а переносится на конструирование содержательного эксперимента и, самое главное, на интерпретацию и анализ полученных результатов.

С. Волд [3] так определил задачи, решаемые хемометрикой: *как получить химически важную информацию из химических данных, как организовать и представить эту информацию и как получить данные, содержащие такую информацию.*

То, что хемометрика возникла и начала бурно развиваться именно в начале 70-х годов прошлого столетия, явно связано с появлением в то же время быстродействующей вычислительной техники, которая стала повсеместно доступна ученым и инженерам. Это позволило воплотить многие сложные алгоритмы обработки данных, в особенности методы анализа результатов многооткликовых и многофакторных экспериментов. В свою очередь это побудило производителей измерительных приборов разрабатывать оборудование, способное производить многократно большее количество измерений. Для извлечения полезной информации из все возрастающего объема химических данных потребовалось применять хемометрические методы. В результате такого взаимодействия был достигнут первый несомненный успех хемометрики. Оказалось, что традиционные аналитические методы, требующие больших затрат труда, времени, уникального оборудования, дорогостоящих реактивов, могут быть заменены на косвенные хемометрические методы, гораздо более быстрые и дешевые. Наиболее ярко эта тенденция проявилась при использовании инфракрасной (ИК) спектроскопии, особенно в ближней области, прежде считавшейся малополезной из-за высокого и трудно устранимого шума, обусловленного интенсивным поглощением воды и эффектом рассеяния в спектрах отражения [4]. Поэтому первые работы по хемометрике были посвящены методам анализа спектроскопических данных [5—7], построению калибровочных моделей с помощью метода главных компонент [8] и метода проекций на латентные структуры [9].

Хемометрика зародилась и длительное время развивалась в рамках аналитической химии. Однако со временем обнаружилась тенденция, которую некоторые исследователи расценили как выход хемометрики

«из-под крыла» аналитической химии и превращение ее в самостоятельную дисциплину. Два обстоятельства дали повод к такому выводу. Первое — это усложнение математического аппарата, используемого в хемометрике [10—14]. Второе обстоятельство связано с появлением многочисленных приложений, в которых хемометрический подход с успехом применяется в областях, далеких от химии. Достаточно вспомнить о многомерном статистическом контроле процессов (MSPC) [15], об анализе изображений (MIA) [16], о биологических приложениях [17].

Хемометрика тесно связана с математикой, и их взаимоотношения заслуживают отдельного рассмотрения. Многие методы и алгоритмы, популярные в хемометрике, не вызывают восторга у математиков [18], справедливо считающих их плохо обоснованными с формальной точки зрения. Специалисты в области хемометрики всегда рассматривали свою деятельность как компромисс между возможностью и необходимостью, полагая, что главное — это практический результат, а не теоретическое обоснование невозможности его достижения. Сталкиваясь с практическими задачами интерпретации очень больших и сложно организованных массивов экспериментальных данных [19], хемометрики изобретают все новые и новые методы их анализа. Делают они это так быстро, что математики, по словам американского статистика Д. Фридмана (J. Friedman) [20], не успевают не только раскритиковать их за это, но и просто понять, что же происходит в хемометрике.

Благодаря такому «агрессивному» подходу к анализу данных, хемометрика нашла многочисленные приложения в самых разных смежных и далеких от химии областях. Она применяется в физической химии для исследования кинетики [21], в органической химии для предсказания активности соединений по их структуре (QSAR) [22], в химии полимеров [23], в теоретической и квантовой химии [24]. Диапазон использования хемометрики очень широк — от пивоварения [25] до астрономии [26]. Она применяется для решения судебных споров по вопросам защиты окружающей среды [27] и для контроля качества производства полупроводников [28]. Подробный анализ взаимодействия хемометрики с различными областями человеческой деятельности приведен в книге английского аналитика Р. Бреретона (R. Brereton) [29].

Некоторые направления хемометрики развивались и в СССР, и позднее в России. Так, например, еще в 1950-е годы в Харьковском университете под руководством Н.П. Комаря проводились исследования по математическому описанию химических равновесий [30]. Позднее появились работы Л.А. Грибова [31] и М.Е. Эляшберга по спектральным методам [32], Б.М. Марьянова по титриметрии [33], Б.Г. Дерендяева и В.И. Вершинина по методам компьютерной идентификации органических веществ [34], И.Г. Зенкевича по хроматографии [35]. Хемометрический подход активно используется в работах [36], выполняемых в научной школе Ю.А. Золотова [37]. Исследования в близкой к хемометрике области QSAR ведутся под руководством Н.С. Зефирова [38]. Метрологическим аспектам и вопросам контроля качества химического анализа посвящена монография В.И. Дворкина [39]. В С.-Петербургском университете группа ученых под руководством Ю.Г. Власова работает над созданием сенсорных систем, известных под названием «электронный язык» [40], а в Воронеже разрабатываются аналогичные методы, известные как «электронный нос» [41]. Во всех этих областях интенсивно используются хемометрические методы. Специалисты из Черноголовки применяют многомерные методы анализа данных при решении задач химической кинетики [42, 43]. За последние годы в России появились новые научные группы, разрабатывающие и применяющие хемометрические подходы: в Москве [44—47], в Барнауле [48, 49], в Томске [50]), в Иркутске [51]).

Информационное и программное обеспечение

Единственная широко известная в России книга по хемометрике [52] была опубликована на русском языке 19 лет назад. Она достаточно полно отражала состояние этой научной области в середине 1980-х годов. На сегодняшний день наиболее полным изложением хемометрических методов является двухтомник, написанный группой авторов под руководством Д. Массарта [53, 54]. Он включает описание основных хемометрических методов и приемов, большое количество практических приложений, обширный список литературы. Существует также множество книг и учебников, ориентированных на очень разный круг читателей. Для студентов и специалистов в области аналитической химии, проще начать с книги [29]; для исследователей, занимающихся в основном спектральным анализом, будут понятнее книги [55, 56]. Для освоения практики применения хемометрики полезна книга [57]. Нельзя также не упомянуть знаменитую книгу Е. Малиновского (E. Malinowski) [58], которую до сих пор многие аналитики считают лучшим учебником в этой области. Теоретические основы хемометрики изложены в работах [59, 60]. Недавно на русский язык был переведен учебник [61], содержащий краткое описание хемометрики. Небольшое, но очень полезное введение в хемометрику написано Б.М. Марьяновым [62]. Для участников трех конференций по хемометрике в России был издан сокращенный перевод популярного в мире учебника по хемометрике [63].

Проблемам хемометрики посвящены два специализированных журнала: Journal of Chemometrics и Chemometrics and Intelligent Laboratory Systems. Статьи, где описывается применение хемометрических методов в прикладных задачах, регулярно печатаются более чем в 50-ти научных журналах, в частности в Analytical Chemistry, Analytica Chimica Acta, Analyst, Talanta, Trends in Analytical Chemistry, Journal of Chromatography, Computers and Chemical Engineering, Vibrational Spectroscopy и т.д.

В мире проводятся как небольшие региональные конференции и семинары, так и регулярные международные конференции. Наиболее авторитетными являются конференция «Хемометрика в аналитической химии» (Chemometrics in Analytical Chemistry — CAC) [64] и «Скандинавский симпозиум по хемометрике» (Scandinavian Symposium on Chemometrics — SSC) [65]. В России, начиная с 2002 года, проходят ежегодные международные школы-симпозиумы «Современные методы анализа многомерных данных» [66—68]. Теоретические и прикладные аспекты хемометрики широко представлены в виде интернет-ресурсов. В боль-

шинстве своем это англоязычные страницы [69—72], есть и несколько российских сайтов [73, 74].

В качестве программного обеспечения в хемометрике используются специализированные пакеты программ [75—77], позволяющие наглядно и быстро обрабатывать данные в интерактивном режиме. Однако широко применяются и статистические пакеты общего назначения [78, 79]. Зачастую исследователи пишут процедуры сами, например в кодах MATLAB [80], и они публикуются для свободного применения, например в [60].

Термины и обозначения

В русском языке до сих пор не сложилась общепризнанная система хемометрических терминов. Некоторые понятия переводились ранее неверно или неточно. Например, фундаментальный хемометрический метод PLS первоначально расшифровывался как *Partial Least Squares* и на русский язык это переводилось как «частичные» или «частные наименьшие квадраты», что никак не соответствует сути метода. К счастью, в последнее время оригинальная трактовка аббревиатуры PLS изменилась на *Projection on Latent Structures*, что дословно переводится как «проекция на латентные структуры».

Термины *soft* и *hard*, часто используемые в хемометрике для характеристики методов моделирования, должны, по нашему мнению, переводиться как «формальный» и «содержательный», что точнее отражает суть этих понятий. Для понятия *N-way* мы используем термин «N-модальный». Может быть, это и не лучшее решение, но применение традиционного термина тензорного анализа «валентность» в химическом контексте мы сочли неудачным. Во многих случаях переводчики просто избегали давать русские названия ключевым хемометрическим понятиям, таким как *scores* и *loadings*, используя вместо них сложные эвфемизмы. Мы полагаем, что в хемометрике невозможно обойтись без понятий «счет» и «нагрузка» или их аналогов.

Хемометрика — это наука, использующая сокращения. В данном случае мы имеем в виду не понижение размерности данных, а то, что в хемометрике часто используются аббревиатуры: PCA, PLS, PCR, RMSEP и т. п. Несмотря на то, что у некоторых из них есть общепринятые русские аналоги, например PCA это МГК, PCR это РГК, в данном обзоре мы решили сохранить оригинальные английские аббревиатуры. Мы приводим список сокращений, использованных в статье, в котором курсивом выделены переводы, употребляемые впервые.

ANN (artificial neural network) — искусственная нейронная сеть;

DASCO (discriminant analysis with shrunk covariance matrices) — *дискриминантный анализ с сокращенной ковариационной матрицей*;

EFA (evolving factor analysis) — *эволюционный факторный анализ*;

GA (genetic algorithm) — генетический алгоритм;

IA (immune algorithm) — иммунный алгоритм;

INLR (implicit non-linear latent variable regression) — *неявная нелинейная регрессия на латентных переменных*;

ITTFA (iterative target transformation factor analysis) — *итерационный целевой факторный анализ*;

KNN (k-nearest neighbours) — классификация по K ближайшим соседям;

LOO (leave-one-out) — метод перекрестной проверки с исключением по одному образцу;

NIPALS (non-linear iterative projections by alternating least-squares) — *нелинейное итерационное проецирование при помощи чередующихся наименьших квадратов*;

PARAFAC (parallel factor analysis) — параллельный факторный анализ;

PAT (process analytical technology) — аналитический контроль процессов;

PC (principal component) — главная компонента;

PCA (principal component analysis) — метод главных компонент;

PCR (principal component regression) — регрессия на главные компоненты;

PLS (projection on latent structures) — проекция на латентные структуры;

PLS-DA (PLS discriminant analysis) — дискриминантный анализ с помощью регрессии на латентные структуры;

PMN (penalized minimum norm projection) — *проекции с помощью штрафных функций минимума нормы*;

QPLS (quadratic PLS) — квадратичный PLS;

QSAR (quantitative structure-activity relationship) — количественная связь «структура—активность»;

RMSEC (root-mean square error of calibration) — среднеквадратичный остаток калибровки;

RMSEP (root-mean square error of prediction) — среднеквадратичный остаток прогноза;

SIMCA (soft independent modeling of class analogy) — *формальное независимое моделирование аналогий классов*;

SIMPLISMA (Simple-to-use interactive self-modeling mixture analysis) — *простой интерактивный автомодельный анализ смесей*;

SIMPLS (Statistically Inspired Modification of PLS) — *статистически мотивированный* PLS;

SMCR (self-modeling curve resolution) — *метод автомодельного разрешения кривых*;

SVD (singular value decomposition) — разложение по сингулярным значениям;

SVM (support vector machine) — метод опорных векторов;

WFA (window factor analysis) — оконный факторный анализ

В статье используются следующие обозначения. Скалярные переменные выделяются курсивом, например s. Векторы (столбцы) обозначаются прямыми жирными строчными буквами, например $\mathbf{x}$, а матрицы — прописными буквами, например $\mathbf{X}$. Мультимодальные матрицы выделяются еще и курсивом, например $\boldsymbol{G}$. Элементы массивов обозначают той же, но строчной буквой, например x_{ij} — это элемент матрицы $\mathbf{X}$. Индекс i обозначает строку матрицы; он изменяется от 1 до I. Индекс j соответствует столбцу, он изменяется от 1 до J. Аналогичные обозначения применяются и для других индексов, например $a = 1,\ldots, A$. Операция транспонирования обозначается верхним индексом t, например $\mathbf{X}^t$.

Основы хемометрического подхода

Химические данные и информация

Химические данные — это основной объект, с которым работает хемометрика. Следуя классификации, предложенной в работе [81], рассмотрим «типичное» устройство данных (см. рис. 1).

Простейший случай — это одномерные данные (0D), т.е. просто одно число, например, значение оптической плотности, которое может быть получено измерением на монохроматическом фотометре.

Более сложный случай — это многомерные, *одномодальные* данные (1D), т.е. набор результатов нескольких измерений, относящихся к одному образцу. Примерами таких данных являются спектр, хроматограмма, набор дескрипторов. С математической точки зрения их можно интерпретировать как 1D-вектор (столбец или строка), каждый элемент которого соответствует некоторой переменной (длина волны в спектроскопии, время удерживания в хроматографии). Число переменных определяет размерность данных.

Следующий, наиболее распространенный тип химических данных, — это *двухмодальные* данные, которые представляются 2D-матрицей — таблицей из чисел, имеющей *I* строк и *J* столбцов. Типичный пример — набор спектров, снятых для *I* образцов на *J* длинах волн. Каждая строка в такой матрице представляет объект (в данном случае образец), а каждый столбец — переменную (длину волны). Отнесение данных к объектам (образцам) или к переменным (каналам) имеет большое значение для их интерпретации.

В последнее время большое внимание уделяется *трех- (и более) модальным данным* [82]. Их можно представить в виде параллелепипеда (3D-матрицы), в котором каждое ребро соответствует своему типу переменной. Пример четырех- и даже восьмимодальных данных можно найти в [82].

Данные могут объединяться в блоки. Простейший случай — это один блок **X**. Такой случай чаще встречается в качественном анализе, например, в задаче разделения спектров и концентраций. Количественный анализ, основанный на регрессионных зависимостях, использует данные, состоящие из двух и более блоков. Блок предикторов (например, 2D-матрица спектров **X**) и блок откликов (например, 1D- вектор концентраций **y**) составляют набор стандартных данных, по которым строится калибровочная модель $\mathbf{y} = \mathbf{Xb}$. Встречаются данные и более сложной структуры, объединяющей три и более блоков данных [83]. Для их анализа применяются специальные методы, называемые маршрутным моделированием (path modeling) [84].

Может показаться, что такая систематизация данных — размерность, модальность, блочность, носит несколько формальный характер и может представлять интерес только для математиков. Это не совсем так. В последние годы кардинально изменились представления о том, какие данные можно считать большими. Если в начале 1970-х годов большой считалась матрица данных (например спектров), состоящая из 20 столбцов (переменных, например длин волн) и 100 строк (объектов, например образцов), то сейчас, с развитием техники эксперимента, большой может считаться матрица с 1 000 000 столбцов и 400 000 строк [19]. При обработке таких массивов их приходится разделять на блоки и интерпретировать поочередно. Это разделение нельзя проводить формально, здесь обязательно требуется участие опытного химика, понимающего суть дела. Понятие модальности тоже придумали не математики. Это естественный ответ на потребность анализа данных гибридных и эволюционных экспериментов, число которых увеличивается по мере развития инструментальной базы. С внедрением новых аналитических методов, таких как гиперспектральный анализ [85] и микрочипы [86], сложность данных будет только нарастать.

Основная задача хемометрики состоит в извлечении из данных нужной химической информации. Понятие *информация* является ключевым в хемометрике, поэтому дадим подробные комментарии. Что является информацией — зависит от цели решаемой задачи. В некоторых случаях достаточно знать, что искомое вещество присутствует в системе, а в других — необходимо получить количественные значения. Экспериментальные данные могут содержать нужную информацию, они даже могут быть избыточными, но иногда информации в данных может не быть совсем. Все данные содержат шум, например погрешности, которые скрывают нужную информацию.

Для иллюстрации рассмотрим следующий идеализированный эксперимент. Имеется система, состоящая из трех веществ A, B и C без посторонних примесей. Предположим, что абсолютно точно известны

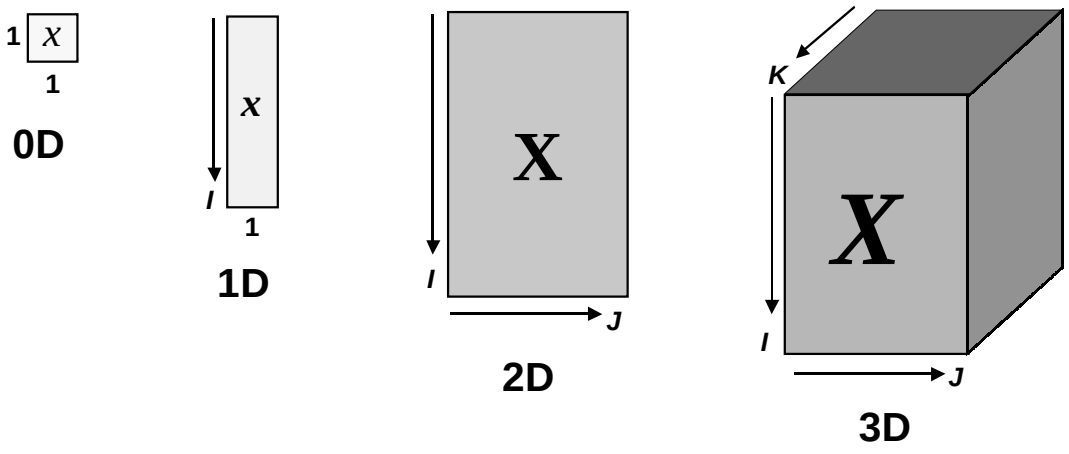

Рис. 1. Графическое представление данных разной модальности

спектры $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ всех компонентов. Заметим, что слово «спектры» употребляется в самом общем смысле. Это могут быть любые многомерные данные, например хроматограммы, и тогда λ — это время удерживания. Требуется определить концентрации веществ по спектру смеси $x(\lambda)$,которые также можно получить без погрешностей. Если каждый спектр содержит значения отклика для 30 длин волн (или времен удерживания) λ, то для решения этой задачи можно составить 30 уравнений относительно трех неизвестных концентраций c_A, c_B и c_C:

$$x(\lambda_1) = c_A s_A(\lambda_1) + c_B s_B(\lambda_1) + c_C s_C(\lambda)$$
$$....$$
$$x(\lambda_{30}) = c_A s_A(\lambda_{30}) + c_B s_B(\lambda_{30}) + c_C s_C(\lambda_{30})$$

Ясно, что столько уравнений решать не нужно, можно оставить только три из них, соответствующие любым* трем длинам волн, и по ним получить нужную информацию. Таким образом, оказывается, что исходные данные (30-мерный 1D-вектор) избыточны по отношению к искомой информации — используя любые три точки спектра, получим одни и те же значения концентраций.

Сделаем теперь пример более реалистичным, допуская, что все спектры содержат некоторую случайную погрешность. Тогда оценки концентраций, определяемые по разным «тройкам» длин волн, будут различаться. Эти оценки можно усреднить и получить концентрации с лучшей точностью. Заметим, что того же результата можно достичь и с помощью повторных экспериментов. Однако этот путь не эффективен, поскольку требует больших затрат сил и времени. Гораздо проще уменьшать неопределенность результатов количественного анализа за счет увеличения числа переменных (каналов, длин волн) в одном-единственном эксперименте. Этот вывод является первым важным принципом хемометрики — *использование многомерного подхода* при конструировании экспериментов и анализе их результатов.

Данные всегда (или почти всегда) содержат в себе нежелательную составляющую, называемую *шумом.* Природа шума может быть различной. Это могут быть случайные погрешности, сопровождающие эксперимент (сдвиг базовых линий, погрешности в определении сигналов, неточности в подготовке и проведении эксперимента). Но во многих случаях шум — это та часть данных, которая не содержит искомой информации. Так, например, если в рассмотренном выше примере требовалось определить концентрации только двух веществ A и B, то вклад от вещества C был бы шумом, нежелательной примесью. *Что считать шумом, а что информацией* — решается с учетом поставленных целей и методов, используемых для ее достижения. Это второй принцип хемометрического подхода к анализу данных.

Шум и избыточность в данных обязательно проявляются через *корреляционные связи* между переменными. Возвратимся к идеализированному примеру. Можно заметить, что в матрице «чистых» спектров, имеющей размерность 3×30, только три столбца данных линейно независимы. Зафиксировав эту тройку, любой четвертый столбец можно представить в виде их линейной комбинации. То, что линейно независимых столбцов ровно три, не является случайностью, ведь именно такое число веществ присутствует в нашей системе. Это число, называемое рангом матрицы, играет важную роль в хемометрическом анализе. В более реалистичном варианте рассматриваемой задачи (наличие погрешностей) можно заметить появление дополнительных корреляций в данных, например, если концентрация третьего вещества C будет существенно меньше шума. Тогда этих данных окажется уже недостаточно для надежного определения всех трех концентраций, и эффективный ранг матрицы станет равным двум. Таким образом, погрешности в данных могут привести к появлению не систематических, а случайных связей между переменными. Очевидно, что в первом случае имеют место причинные, а во втором — корреляционные связи. Знание эффективного (химического) ранга и соответственно скрытых, *латентных переменных*, число которых равно этому рангу, является третьим важнейшим принципом хемометрики [58].

Проиллюстрируем его применение примером, который будет некоторым усложнением уже рассмотренной ранее трехкомпонентной системы. Предположим, что имеется несколько (I) смесей веществ A, B и C, но индивидуальные спектры этих веществ $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ неизвестны. Проведя анализ, можно получить спектры этих образцов — двухмодальные данные, т.е. матрицу $\mathbf{X}$ размерности I×30. Подвергнув матрицу $\mathbf{X}$ обычному математическому анализу, можно определить ее ранг. Это число дает важную информацию о том, сколько компонентов присутствует в системе или, по крайне мере, сколько их можно различить.

Таким образом, в химических данных почти всегда присутствуют внутренние, скрытые связи между переменными, приводящие к множественным корреляциям — коллинеарностям. Такое свойство данных называется *мультиколлинеарностью.* Оно может проявляться как избыточность данных, что позволяет улучшить качество оценок. Вместе с тем при неправильном выборе метода обработки данных мультиколлинеарность может негативно сказаться на качестве анализа. Так, например, применение метода множественной линейной регрессии в условиях мультиколлинеарности совершенно неприемлемо [63]. Для регрессионного анализа таких данных следует привлекать специальные методы, например ридж-регресию [87] или проекционные подходы [60].

Существенным источником шума в данных может быть способ отбора образцов. Теория *пробоотбора*, значительный вклад в которую внес П. Жи (P. Gy) [88], приобрела большую популярность в последнее время [89]. Многочисленные приложения теории пробоотбора можно найти в специальном выпуске [90], полностью посвященном этой теме.

Другая проблема, с которой может столкнуться исследователь, — это *пропуски в данных* [91]. Это может случиться по разным причинам: отказы измерительного прибора, выход за пределы обнаружения, нехватка образцов и т.п. Большинство хемометрических методов не допускает пропусков в данных, поэтому для их заполнения используют специальные

* Строго говоря — не совсем любым. Необходимо, чтобы система имела единственное решение.

приемы, среди которых самым популярным является итерационный алгоритм. Каждая его итерация состоит из двух шагов. На первом шаге проводится оценка параметров модели так, как будто данные известны полностью. При этом пропуски заполняются некоторыми априорно допустимыми значениями, например средними по окружающим элементам массива данных. На втором шаге с помощью полученной модели находятся наиболее вероятные значения пропущенных данных и совершается следующая итерация. Для заполнения пропусков используется также подход, основанный на методе максимума правдоподобия [92]. Детали таких алгоритмов в большой степени зависят от того, какая модель используется для описания данных.

Модели, методы

Рассмотрев устройство данных, перейдем к методам их анализа. Далее основные хемометрические методы будут описаны более подробно, а этот раздел посвятим вопросам методологии.

Хемометрические методы можно разделить на две группы, соответствующие двум главным задачам: *исследованию* данных, например классификации и дискриминации, и *предсказанию* новых значений, например калибровка. Методы первой группы оперируют, как правило, с одним блоком данных, а для калибровки необходимы, как минимум, два блока — предикторов и откликов. В зависимости от поставленных целей, методы решения могут быть направлены на предсказание внутри диапазона условий эксперимента (*интерполяция*) или за его пределами (*экстраполяция*).

Существенным является разделение методов на формальные (soft), называемые также «черными», и содержательные (hard), или «белые». При использовании формальных моделей [93] данные описываются эмпирической зависимостью (как правило, линейной), справедливой в ограниченном диапазоне условий эксперимента. В этом случае не нужно знать, как устроен механизм исследуемого процесса, однако такой метод не позволяет решать задачи экстраполяции. Параметры формальных моделей лишены физического смысла и должны интерпретироваться соответствующими математическими методами. Содержательное моделирование [94] базируется на физико-химических принципах и позволяет экстраполировать поведение системы на новые условия. Параметры «содержательной» модели имеют физический смысл, и их значения могут помочь при интерпретации найденной зависимости. Однако этот метод может быть применен только в том случае, если модель известна априори. Каждый из подходов имеет свои сильные и слабые стороны [23] и у каждого есть свои сторонники и противники. За последнее время появилось много работ, в которых рассматриваются так называемые «серые» модели [95], объединяющие сильные стороны обоих методов.

Интерес к «черным» (формальным) и «серым» методам моделирования обусловлен большими трудностями выбора и подтверждения правильности содержательной модели. Во многих случаях все сводится к простому перебору внутри короткого набора конкурирующих зависимостей, в результате которого обычно выбирается наипростейшая модель, минимально отклоняющаяся от данных. Но такой подход не гарантирует правильности выбранного метода и может привести к грубым ошибкам. Часто исследователи используют модели, которые О.Н. Карпухин [96] справедливо назвал «розовыми», — это идеализированные зависимости, плохо соответствующие реальным артефактам, присутствующим в данных (дрейф базовых линий, погрешности, не подчиняющиеся нормальному распределению, и т.п.). Формальные многофакторные линейные модели и надлежащие методы их анализа гораздо лучше приспособлены к учету «неидеальности» данных. Они работают и в тех случаях, когда ни о какой содержательной физико-химической модели не может быть и речи. Обоснованием для использования линейных моделей служит тот факт, что любую, даже очень сложную, но непрерывную зависимость можно представить как линейную функцию параметров в достаточно малой области изменения этих параметров. Принципиальным моментом здесь является то, какую область можно считать допустимой, иначе говоря, насколько широко можно применять построенную формальную модель. Ответ на этот вопрос дают *методы проверки* (валидации) моделей.

При надлежащем построении модели исходный массив данных состоит из двух независимо полученных наборов, каждый из которых является достаточно представительным. Первый набор, называемый обучающим, используется для идентификации модели, т.е. для оценки ее параметров. Второй набор, называемый проверочным, служит для проверки модели. Построенная модель применяется к данным из проверочного набора, и полученные результаты сравниваются с проверочными данными. Таким образом принимается решение о правильности, точности моделирования методом *тест-валидации*. В некоторых случаях, когда объем данных слишком мал для такой проверки, применяют метод перекрестной проверки — *кросс-валидации* [97]. По этому методу проверочные значения вычисляют с помощью следующей процедуры. Некоторую фиксированную долю, например первые 10% образцов, исключают из исходного набора данных. Затем строят модель, используя только оставшиеся 90% данных, и эту модель применяют к данным исключенного набора. На следующем цикле исключенные данные возвращают, и удаляется уже другая порция данных (следующие 10%). Снова строится модель, которая применяется к исключенным данным. Такая процедура повторяется до тех пор, пока все данные не пройдут стадию исключения (в нашем случае — 10 циклов). Наиболее популярен (хотя и неоправданно) вариант перекрестной проверки, в котором данные исключаются по одному (LOO). В регрессионном анализе используется также проверка методом коррекции размахом, которая описана в [63].

Любой результат, полученный при анализе и моделировании экспериментальных данных, несет в себе неопределенность. Количественная оценка или качественное суждение могут измениться при повторном эксперименте вследствие воздействия различных случайных и систематических погрешностей как присутствующих в исходных данных, так и вносимых на стадии моделирования [98]. *Неопределенность* в количественном анализе характеризуется либо числом — стандартным отклонением [99], либо интервалом — доверительным [100] или прогнозным [44]. В качественном анализе применяется метод проверки стати-

стических гипотез [101], в котором неопределенность характеризуется через вероятность принятия неверного решения [102]. Методы оценки неопределенности при моделировании многомерных [103] и многомодальных [104] данных вызывают большой интерес у специалистов в области хемометрики.

Надежность аналитического метода сильно зависит от того, какие данные были использованы для построения и проверки соответствующей модели. Наличие выбросов [105] или малоинформативных данных снижает точность модели и, наоборот, присутствие влиятельных образцов в эксперименте [106] существенно улучшает качество модели. Оценка влиятельности данных может проводиться как классическими регрессионными методами [107], так и с помощью нестатистических процедур [44]. При использовании построенной модели для определения интересующих показателей обычно сталкиваются с похожими проблемами. Может оказаться, что метод не применим к некоторым образцам (выброс в прогнозных данных [108]) или дает очень неточный результат. Оценка неопределенности метода не в среднем [109], а для индивидуальных образцов является сложной задачей, над решением которой работают в настоящее время разные группы исследователей [см. например 110]. Именно их усилия определяют успешное решение таких практически важных задач, как перенос калибровок с одного прибора на другой [111], отбор переменных [112], построение робастных методов анализа данных [113].

Исследование данных, задачи классификации и дискриминации

Метод главных компонент

Современные приборы позволяют производить огромное количество измерений. Например, если использовать *in situ* спектроскопический датчик для записи спектра на 300 длинах волн через каждые 15 с, то за один час работы можно получить матрицу данных размерности 300×240, т.е. 72000 чисел. Однако из-за мультиколлинеарности доля полезной информации в таком массиве может быть относительно невелика. Для выделения полезной информации в хемометрике используются методы сжатия данных (в отличие от традиционного подхода, когда из данных выделяют отдельные, особо значимые результаты измерения). Идея этих методов состоит в том, чтобы представить исходные данные, используя новые скрытые переменные. При этом должны выполняться два условия. Во-первых, число новых переменных (химический ранг) должно быть существенно меньше, чем число исходных переменных, и, во-вторых, потери от такого сжатия данных должны быть сопоставимы с шумом в данных. Методы сжатия данных позволяют представить полезную информацию в более компактном виде, удобном для визуализации и интерпретации.

Наиболее популярным способом сжатия данных является *метод главных компонент* (PCA) [10]. Он дает основу для других аналогичных хемометрических методов, включая эволюционный факторный анализ (EFA) [114], оконный факторный анализ (WFA) [115], итерационный целевой факторный анализ (ITTFA) [116], а также для многих методов классификации, например, формального независимого моделирования аналогий классов (SIMCA) [117]. С математической точки зрения метод главных компонент — это декомпозиция исходной 2D-матрицы **X**, т.е. представление ее в виде произведения двух 2D-матриц **T** и **P** [63]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{t} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{t} + \mathbf{E} \qquad (1)$$

где **T** — матрица счетов (scores); **P** — матрица нагрузок (loadings); **E** — матрица остатков (см. рис. 2). Число столбцов $\mathbf{t}_a$ в матрице **T** и $\mathbf{p}_a$ в матрице **P** равно эффективному (химическому) рангу матрицы **X**. Величина A называется *числом главных компонент* (PC), и она, естественно, меньше числа столбцов в матрице **X**.

Для иллюстрации метода PCA вернемся к примеру смеси трех веществ, рассмотренному ранее. Матрица спектров смесей веществ **X** может быть представлена как произведение матрицы концентраций **C** и матрицы спектров чистых компонентов **S**:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{t} + \mathbf{E} \qquad (2)$$

Число строк в матрице **X** равно числу образцов (I), каждая ее строка соответствует спектру одного образца, снятому для J длин волн. Число строк в матрице **C** также равно I, а число столбцов соответствует числу компонентов в смеси (A = 3). Матрица чистых спектров **S** в разложении (2) представлена в транспонированном виде, т.е. количество ее строк равно числу длин волн (J), а число столбцов равно A.

Задача разделения экспериментальной матрицы **X** на «чистые» составляющие, соответствующие концентрациям **C** и спектрам **S** (понимаемым в обобщенном смысле) — предмет особой области в хемометрике, называемой *разделением кривых* (curve resolution) [118]. В этой области можно выделить два направления.

Первое использует метод автомодельного разрешения кривых (SMCR) [119] и оно ориентировано прежде всего на приложение к гибридным методам анализа (хроматография) [120]. Для реализации автомодельного подхода применяются методы формального моделирования (PCA, EFA), которые не требуют содержательного знания об исследуемой системе. В рамках этого направления можно отметить метод SIMPLISMA [121], применяющий простой, но весьма эффективный подход, основанный на отборе переменных [122].

Второе направление, напротив, учитывает априорную информацию о процессах и использует «серые»

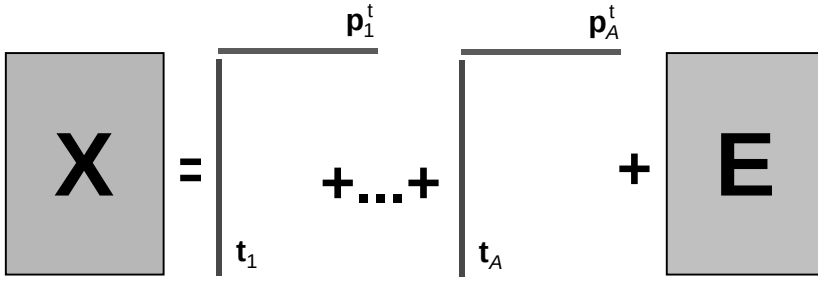

Рис. 2. Графическое представление метода главных компонент

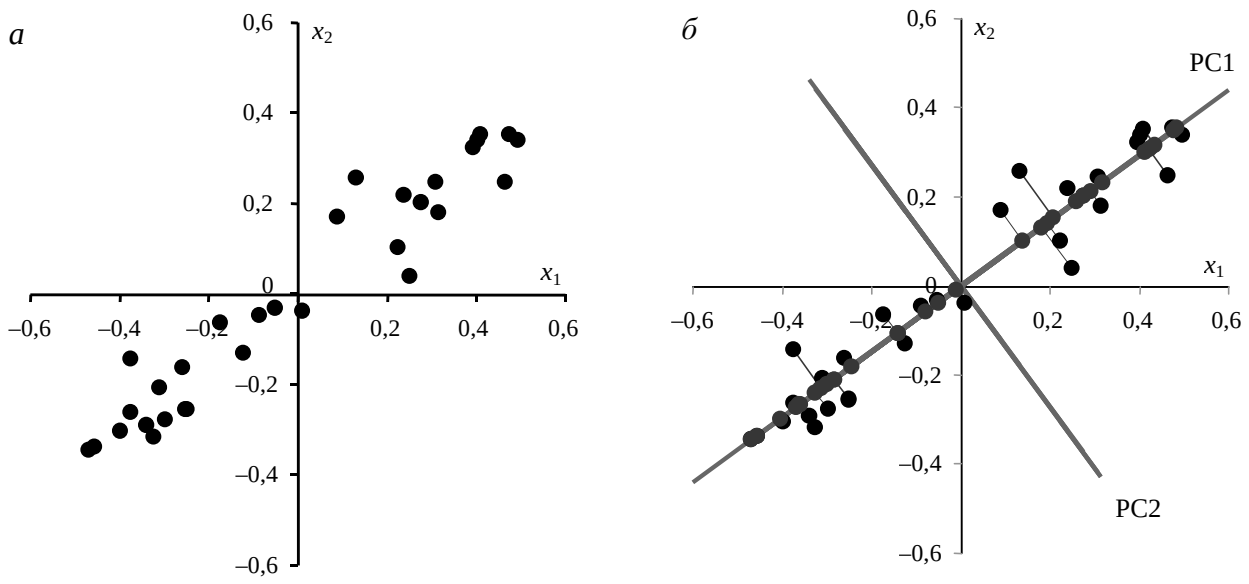

Рис. 3. Графическое изображение набора данных в методе главных компонент:

а — данные в исходных координатах переменных x_1 и x_2; *б* — данные в координатах главных компонент

модели [123]. Это направление находит свое приложение при исследовании кинетики [121] и термодинамики [124]. Ключевым моментом в таких задачах является определение величины химического ранга системы — числа главных компонент A [125]. В идеале предсказанные спектры **S** и концентрации **C** близки к истинным значениям, хотя их невозможно восстановить точно. Причина этого не только в погрешностях эксперимента, но и в том, что спектры могут частично перекрываться. Если PCA применяется для разделения данных на химически осмысленные компоненты, как в уравнении (2), то метод обычно называют факторным анализом (в отличие от формального анализа главных компонент).

Метод главных компонент эффективен не только в приложении к задачам разделения. Он пригоден для исследования любых химических данных [126, 127]. В этом случае матрицы счетов **T** и нагрузок **P** уже нельзя интерпретировать как спектры и концентрации, а число главных компонент A — как число химических компонент, присутствующих в исследуемой системе. Тем не менее даже формальный анализ матриц счетов и нагрузок оказывается весьма полезным для понимания устройства данных. Дадим простейшую двумерную иллюстрацию метода PCA.

На рис. 3*а* показаны данные, состоящие только из двух переменных x_1 и x_2, которые связаны сильной корреляцией. На рис. 3*б* те же данные представлены в новых координатах. Вектор нагрузок $\mathbf{p}_1$ первой главной компоненты (PC1) определяет направление новой оси, вдоль которой наблюдается наибольшее изменение данных. Проекции всех исходных точек на эту ось составляют вектор $\mathbf{t}_1$. Вторая главная компонента $\mathbf{p}_2$ ортогональна первой, и ее направление (PC2) соответствует наибольшему изменению в остатках, показанных на рис. 3*б* отрезками, перпендикулярными оси $\mathbf{p}_1$.

Этот тривиальный пример показывает, что метод главных компонент осуществляется последовательно, шаг за шагом. На каждом шаге исследуются остатки $\mathbf{E}_a$, среди них выбирается направление наибольшего изменения, данные проецируются на эту ось, вычисляются новые остатки и т.д. Этот алгоритм называется NIPALS [63]. Другой популярный алгоритм сжатия данных – разложение по сингулярным значениям (SVD)[128] – строит ту же декомпозицию (1) без итераций. Остановка итерационной процедуры или, другими словами, выбор числа главных компонент A проводится с использованием критериев, показывающих точность достигнутой декомпозиции. Пусть исходная матрица **X** имеет размерность: I строк и J столбцов, и в разложении (1) участвуют A главных компонент. Величины

$$\mu_a = 100 \sum_{i=1}^{I} t_{ia}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2 ,$$

$$E_a = 100\left(1 - \sum_{i=1}^{I}\sum_{j=1}^{J} e_{ij}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2\right), \; a = 1, \ldots, A \qquad (3)$$

называются нормированным *собственным значением* и *объясненной дисперсией.* Их обычно изображают на графике в зависимости от числа a, тогда резкое изменение величин (3) указывает на нужное значение числа главных компонент. Для правильного выбора A необходимо использовать метод тест-валидации либо кросс-валидации (как это описано выше). Уравнения (1) не содержат в себе свободного члена, поэтому для декомпозиции данных их следует сначала отцентрировать (т.е. вычесть среднее по столбцам) и иногда нормировать.

Метод главных компонент можно трактовать как проецирование данных на подпространство меньшей размерности. Возникающие при этом остатки **E** рассматриваются как шум, не содержащий значимой химической информации. В этом подпространстве

можно ввести меру близости образцов, называемую *расстоянием Махаланобиса* (Mahalanobis) [129], с помощью которой удается решать многие задачи качественного анализа. Другим мощным инструментом анализа данных в проекционном подпространстве является *прокрустово* (Procrustes) вращение [130].

При исследовании данных методом PCA особое внимание уделяется графикам счетов и нагрузок. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый образец изображается в координатах ($\mathbf{t}_i$, $\mathbf{t}_j$), чаще всего — ($\mathbf{t}_1$, $\mathbf{t}_2$). Близость двух точек означает их схожесть, т.е. положительную корреляцию. Точки, расположенные под прямым углом, являются некоррелированными, а расположенные диаметрально противоположно — имеют отрицательную корреляцию. Если график счетов используется для анализа взаимоотношений образцов, то график нагрузок применяется для исследования роли переменных. На графике нагрузок каждая переменная отображается точкой в координатах ($\mathbf{p}_i$, $\mathbf{p}_j$), например ($\mathbf{p}_1$, $\mathbf{p}_2$). Анализируя его аналогично графику счетов, можно понять, какие переменные связаны, а какие независимы. Совместное исследование парных графиков счетов и нагрузок также может дать много полезной информации о данных [63].

Классификация и дискриминация

Классификация — это весьма широкий круг задач качественного химического анализа, в которых требуется установить принадлежность образца к некоторому классу. Задачи классификации можно разделить на две большие группы. К первой группе относятся так называемые задачи *без обучения* (unsupervised). Они названы так потому, что в них не используется обучающий набор образцов и их можно рассматривать как разновидность исследовательского анализа. Задачи второй группы — классификация *с обучением* (supervised), называются также задачами *дискриминации*. В них применяется обучающий набор образцов, о которых имеется априорная информация относительно принадлежности к классам. Методы решения задач классификации без обучения основаны главным образом на декомпозиции матрицы данных по методу PCA с последующим анализом [131] расстояний между классами, построением дендрограмм, использованием нечетких множеств [132] и т.п. В работе [133] применяется метод прокрустово вращение, а в работах [134—136] — расстояние Махаланобиса. В тех случаях, когда возможно проведение дискриминации, т.е. классификации с обучением, этим методам следует отдавать предпочтение.

Обучающий набор образцов используется для построения модели классификации, т.е. набора правил, с помощью которых исследуемый образец может быть отнесен к тому или иному классу. После того, как модель (или модели) построена, ее необходимо проверить, используя методы тест- или кросс-валидации, и определить, насколько она точна. При успехе проверки — модель готова к практическому применению, т.е. к предсказанию принадлежности новых образцов.

В аналитической химии классификация применяется к наборам мультиколлинеарных данных (спектры, хроматограммы), поэтому дискриминационная модель почти всегда многомерна и основана на соответствующих проекционных подходах — PCA, PLS. Можно отметить использование линейного дискриминантного анализа в ближней ИК спектроскопии [137], а также канонического дискриминантного анализа [138]. Одним из самых популярных подходов является метод формального независимого моделирования аналогий классов (SIMCA [139]), разработанный С. Волдом [117].

В основе метода SIMCA лежит предположение о том, что все объекты одного класса имеют сходные свойства, но обладают индивидуальными особенностями. При построении дискриминационной модели необходимо учитывать только сходство, отбрасывая особенности как шум. Для этого каждый класс из обучающего набора образцов моделируется независимо методом PCA с разным числом главных компонент *A*. После этого вычисляются расстояния между классами, а также расстояния от каждого класса до нового объекта. В качестве таких метрик используются две величины. Одна величина — это расстояние *d* от объекта до класса, которое вычисляется как среднеквадратичное значение остатков **e**, возникающих при проецировании объекта на класс

$$d = \sqrt{\frac{1}{J - A}\sum_{j=1} e_j^2}$$

Эта величина сравнивается со среднеквадратичным остатком внутри класса

$$d_0 = \sqrt{\frac{1}{(I - A - 1)(J - A)}\sum_{ij} e_{ij}^2}$$

Другая величина определяет расстояние от объекта до центра класса, которое вычисляется как *размах* (квадрат расстояния Махаланобиса):

$$h = \frac{1}{I} + \sum_{a=1}^{A} \frac{\tau_a^2}{\mathbf{t}_a^t \mathbf{t}_a}$$

где τ_a — проекция нового образца (счет) на главную компоненту *a*; $\mathbf{t}_a$ — вектор, содержащий счета всех обучающих образцов в классе.

Помимо метода SIMCA для дискриминации химических данных используются также метод DASCO [140], схожий с SIMCA, метод классификации по ближайшему соседу (KNN) [141], метод опорных векторов (SVM) [142, 143] и многие другие. Мощным инструментом является метод дискриминантного анализа с помощью регрессии на латентные структуры — PLS-DA [144].

Трехмодальные методы

Метод главных компонент был разработан для обработки данных, имеющих вид двухмодальной 2D-матрицы. Однако в последнее время исследователи все чаще имеют дело с трех- и более модальными данными, которые обладают более сложной структурой, например в виде параллелепипеда (рис. 4). Такие данные получают, например в гибридных [145, 146] и эволюционных методах анализа [147]. Для сжатия таких массивов данных применяются специальные подходы (три наиболее часто используемые кратко рассмотрены в этом разделе). Полное и систематическое описание этих методов вместе с многочисленными примерами их применения в задачах химического

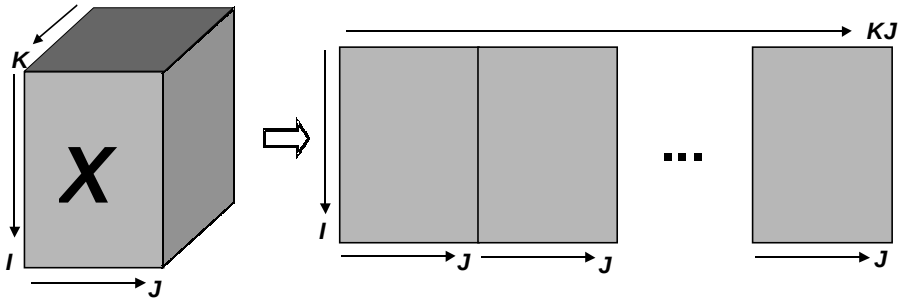

Рис. 4. Графическое представление метода развертывания в плоскую матрицу

анализа приведено в книге [82]. Краткое введение в методы анализа трехмодальных данных представлено в статье [148]. Эти же алгоритмы используются для обработки данных, полученных в результате гиперспектральных измерений [85], а также для анализа изображений [12].

Метод развертывания (unfolding) [149] — это простейший способ анализа трехмодальных данных, с помощью которого 3D-матрица ***X*** размерности $I \times J \times K$ разворачивается в обычную 2D-матрицу $^{u}\mathbf{X}$ размерности $I \times KJ$ (рис. 4). При этом I называется «основной» модой. Далее возможно обычное применение метода главных компонент. Такой подход часто оказывается эффективным, как, например в [23], и при решении 3D-QSAR задач [150, 151]. Однако он имеет ряд недостатков. Во-первых, в качестве основной моды можно выбирать любое из трех направлений, т.е. имеет место неоднозначность развертывания. Во-вторых, при таком подходе теряется связь между соседними точками, так как с переходом от 3D-матрицы $I \times J \times K$ к 2D-матрице $I \times KJ$ уже не учитывается, что измерения x_{ikj} и x_{ik+1j} являются соседними, что может быть существенно.

Алгоритм Tucker3 [152] позволяет обрабатывать трехмодальные данные, сохраняя их первоначальную структуру, а следовательно, и последовательность измерений, например порядок длин волн спектра или последовательность точек по времени на хроматограмме. Исходные 3D-данные ***X*** разлагаются на три обычные 2D-матрицы нагрузок (**A**, **B**, **C**) и трехмодальный керн-массив ***G*** (рис. 5). Каждый элемент исходной 3D-матрицы ***X*** можно записать в виде суммы:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk} \qquad (4)$$

где a, b и c — элементы матриц нагрузок, каждая из которых соответствует своей моде; g — элементы керн-массива ***G***. При этом число главных компонент по каждому направлению (P, Q, R) может быть различным.

Метод PARAFAC (parallel factor analysis) [148] отличается от модели Tucker 3 тем, что каждая мода представляется одним и тем же числом главных компонент R (рис. 6). Разложение строится так, чтобы минимизировать сумму квадратов остатков e_{ijk}:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \qquad (5)$$

Основным достоинством этого метода является единственность разложения. Так, если исследуется смесь нескольких химических веществ, то при правильном выборе числа главных компонент матрицы нагрузок представляют чистые спектры исходных веществ.

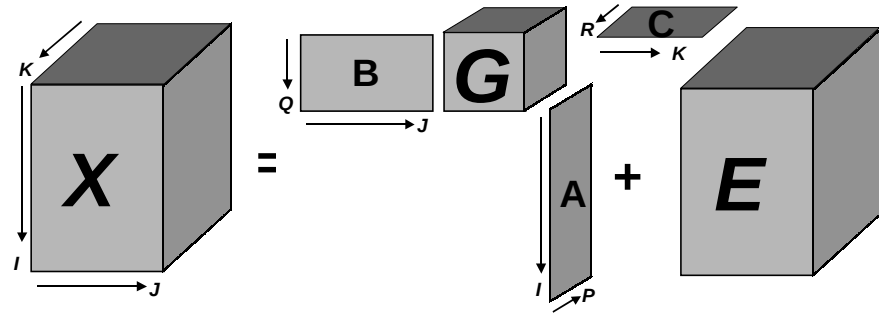

Рис. 5. Графическое представление модели Tucker3

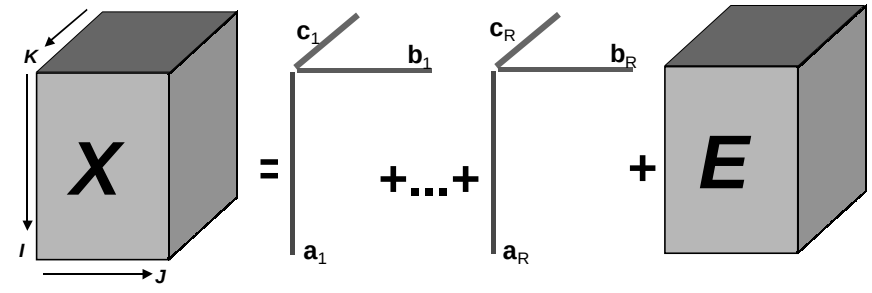

Рис. 6. Графическое представление модели PARAFAC с *R* компонентами

MATLAB-код алгоритма PARAFAC можно найти в [148]. Так как матрицы нагрузок в разложении (5) определяются с помощью итерационной процедуры, этот метод требует очень большого объема вычислений. В настоящее время ведутся работы, направленные на ускорение вычислительных процедур. Последние достижения и их критический анализ представлены в [152], алгоритмы всех рассмотренных методов декомпозиции трехмодальных данных приведены в [153].

Моделирование и предсказание

Линейная калибровка

В задачах количественного анализа участвуют два блока данных. Первый блок **X** — это матрица предикторов (спектры, хроматограммы, набор дескрипторов и т.п.); второй блок **Y** — это матрица откликов (концентрации, активности и т.д.). Число строк (I) в этих матрицах равно количеству

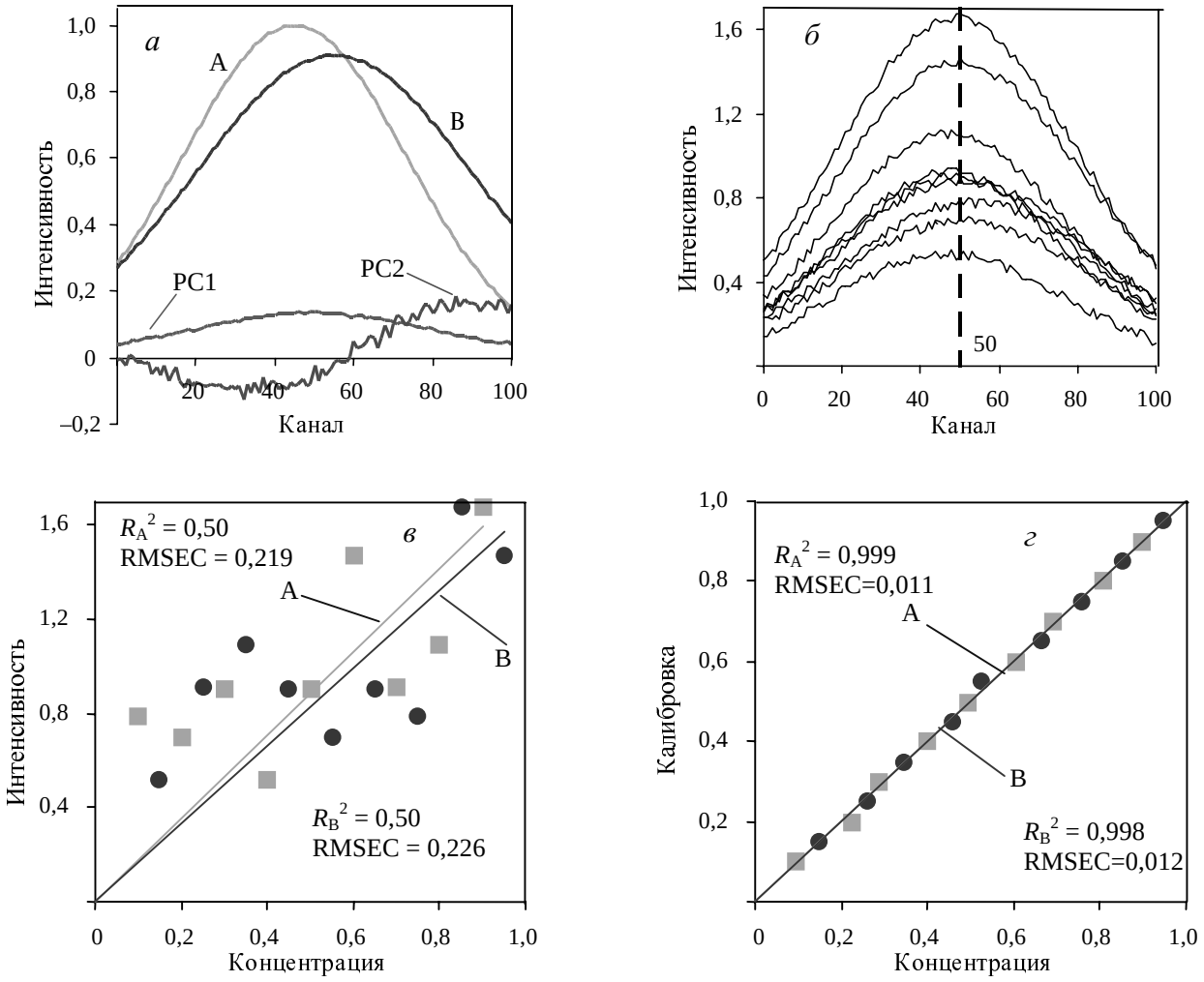

Рис. 7. Модельный пример построения калибровочной зависимости различными методами:

а — спектры чистых веществ (A, B) и главных компонент (PC1, PC2); *б* — модельные спектральные данные; *в* — одномерная калибровка; *г* — калибровка методом PCR. ■ — A, ● — B

образцов, число столбцов (J) в матрице **X** соответствует числу каналов (длин волн), на которых записывается сигнал, и наконец, число столбцов (K) в матрице **Y** равно числу химических показателей, т.е. откликов. Задача калибровки, или градуировки, как ее принято называть в более узком смысле в аналитической химии, состоит в построении математической модели, связывающей блоки **X** и **Y**, с помощью которой можно в дальнейшем предсказывать значения показателей **y** по новой строке значений **x** [9]. Простейшая калибровочная модель — это *одномерная регрессия* (J = 1, K = 1), т.е. модель $y = a + bx$ [154], которая соответствует одному каналу аналитического сигнала. С помощью методов классического регрессионного анализа можно строить более сложную, множественную регрессию ($I > J$, K = 1), в которой участвуют несколько каналов, т.е. **y** = **Xb** [155]. При использовании этих методов обычно предполагается, что значения факторов x_{ij} известны точно, а погрешности присутствуют только в блоке **y**. Современные регрессионные методы (PCR, PLS) позволяют работать с данными для случая, когда погрешности содержатся в обоих блоках.

Для иллюстрации различных методов калибровки снова используем ранее введенный пример смеси веществ. Теперь наполним его конкретным содержанием, смоделировав данные **X** и **Y**. Рассмотрим смесь двух веществ A и B (K = 2) и предположим, что имеется некоторый прибор, позволяющий измерять аналитический сигнал s (спектр) на 101 канале (J = 101). Соответствующие спектры «чистых» веществ ($c_A = c_B = 1$) приведены на рис. 7*а* (кривые A и B). Спектры сильно перекрываются, так что невозможно выделить какие-либо «селективные» каналы для оценки концентраций. На рис. 7*б* представлены девять модельных спектров (I = 9) различных смесей A и B, в которые внесена случайная погрешность со стандартным отклонением 0,05. Эти модельные спектры используем как обучающий набор.

Для построения одномерной калибровки мы взяли интенсивности $s(\lambda_{50})$ для канала 50 и изобразили их на рис. 7*в* в зависимости от концентраций c_A и c_B веществ A (квадраты) и B (круги); соответствующие калибровочные зависимости $s = bc$ — прямые A и B.

Точность калибровки принято характеризовать *среднеквадратичным остатком калибровки* (RMSEC), который вычисляется по формуле

$$\mathrm{RMSEC} = \sqrt{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2 \Big/ F} \qquad (6)$$

где y_i и $\hat{y}_i$ — соответственно известные и предсказанные значения отклика для образцов сравнения $i = 1, \ldots, I$; F — число степеней свободы [29], оно равно $I - 1$ для одномерной регрессии (без свободного члена).

Ясно, что чем меньше RMSEC, тем точнее описываются обучающие данные. Кроме того, качество калибровки характеризуется также *коэффициентом корреляции* R^2 между величинами **y** и $\hat{\mathbf{y}}$ — чем он ближе к единице, тем лучше точность калибровки (соответствующие значения даны на рис. 7*в*). Графики, приведенные на рис. 7*в*, показывают, что из-за недостатка «приборной» селективности одномерная калибровка неудовлетворительна. Калибровка с помощью множественной регрессии будет рассмотрена ниже, а здесь покажем, как работает многомерная модель, построенная с помощью *регрессии на главные компоненты* (PCR [87]).

В данном случае матрица предикторов состоит из спектральных данных, а матрица откликов содержит значения концентраций (т.е. **X = S,** а **Y = C)**. Применяя метод PCA, матрицу **X** можно разложить по формуле (1), причем в нашем примере $A = 2$. Получившиеся векторы нагрузок $\mathbf{p}_1$ и $\mathbf{p}_2$ показаны на рис. 7*а* (кривые PC1 и PC2). Сравнивая этот и соседний (б) графики видно, что первая главная компонента описывает гладкий тренд в данных, тогда как вторая компонента представляет «зашумленные» отклонения от этого тренда. Полученная матрица счетов **T** используется как блок независимых факторов (предикторов) в регрессии на блок откликов **Y**, т.е. **Y = TB**. Результаты калибровки методом PCR показаны на рис. 7*г*, где изображены предсказанные значения концентраций $\hat{\mathbf{y}}$ в зависимости от соответствующих известных значений **y** (квадраты для A и круги для B), а также регрессионные прямые, которые сливаются. Приведенные на этом же графике величины RMSEC и R^2 свидетельствуют о том, что метод PCR позволяет достичь высокой «математической» селективности и получить оценки концентраций веществ A и B с точностью, гораздо лучшей, чем в случае одноканальной калибровки. В методе PCR число степеней свободы в уравнении (6) равно $F = I - A$.

Ранее уже отмечалось, что каждая хемометрическая модель нуждается в полноценной проверке. В нашем примере такая проверка проводилась с помощью проверочного набора (тест-валидация), состоящего из пяти образцов (смесей A и B). На рис. 8*а* представлены результаты проверки модели для оценки концентрации вещества B. В координатах «известно—предсказано» («концентрация—калибровка») приведены данные для девяти образцов, которые были использованы в калибровке (черные круги), и пять образцов проверочного набора (окружности). Здесь же указаны значения остатков калибровки (RMSEC) и проверки (RMSEP), а также коэффициенты корреляции для обучающего (R_c^2) и проверочного (R_t^2) наборов. Среднеквадратичный остаток проверки RMSEP вычисляется аналогично RMSEC (формула 6), но только с использованием образцов из проверочного набора. При этом число степеней свободы F равно числу этих образцов. Из рис. 8*а* видно, что метод главных компонент выдерживает проверку — калибровочная (C) и проверочная (T) прямые сливаются. Рассмотрим в этом контексте метод калибровки с помощью *множественной регрессии*. Поскольку обучающий набор состоит из девяти образцов, то для построения модели можно использовать не более восьми каналов ($I > J$), например: первый, четырнадцатый, двадцать седьмой, и т.д.

На рис. 8*б* показаны результаты калибровки (C) и проверки (T), полученные методом множественной

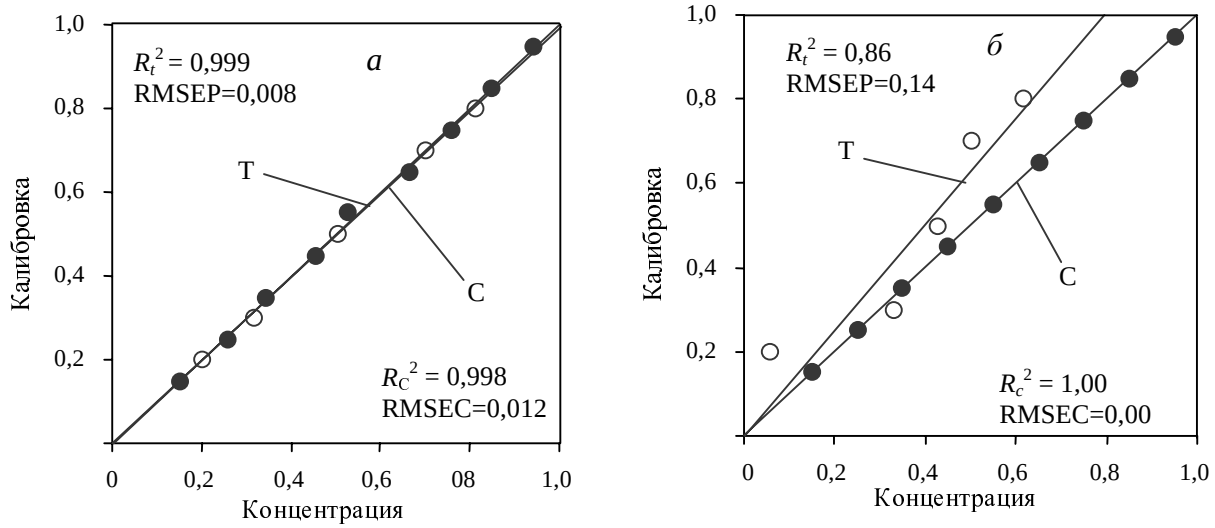

Рис. 8. Проверка калибровок в модельном примере:

а — метод главных компонент; *б* — множественная регрессия. Обучающий (●, C) и проверочный (○, T) наборы

регрессии. Поскольку в этом случае число образцов всего на единицу больше числа каналов (8), то калибровочная прямая C точно проходит через все точки обучающего набора (черные круги), поэтому RMSEC = 0, и $R_c^2 = 1$. Однако проверка показывает неудовлетворительное качество калибровки: точность ее на порядок хуже точности, достигаемой по методу PCR, а проверочная прямая T значительно расходится от калибровочной C. Это — типичный пример *переоценки* модели [60], когда точность описания обучающих данных значительно лучше, чем качество прогнозирования.

Проблема сбалансированности описания данных рассматривается во многих работах А. Хоскюлдссона (A. Höskuldsson), который в 1988 г. ввел новую концепцию моделирования — так называемый *H-принцип* [156]. Согласно этому принципу точность моделирования (RMSEC) и точность прогнозирования (RMSEP) связаны между собой. Улучшение RMSEC неминуемо влечет ухудшение RMSEP, поэтому их нужно рассматривать совместно. Именно по этой причине множественная линейная регрессия, в которой всегда участвует явно избыточное число параметров, неизбежно приводит к неустойчивым моделям, непригодным для практического применения.

В настоящее время самым популярным методом многомерной калибровки в хемометрике является метод *проекции на латентные структуры* (PLS). Он во многом схож с методом PCR, но с тем существенным отличием, что в методе PLS проводится одновременная декомпозиция матриц **X** и **Y**:

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{UQ}^t + \mathbf{F} \qquad (7)$$

Проекции строятся согласованно так, чтобы максимизировать корреляцию между соответствующими векторами X-счетов $\mathbf{t}_a$ и Y-счетов $\mathbf{u}_a$. Поэтому PLS регрессия гораздо лучше описывает сложные связи, используя при этом меньшее число главных компонент. Детальное описание метода PLS приведено в [63]. Этот подход послужил основой для очень многих методов калибровки, используемых в хемометрике, таких как SIMPLS [157], PMN [158], робастный PLS [159], Ridge PLS [160] и многих других подходов [161].

Все представленные методы дают результат предсказания в виде точечной оценки, тогда как на практике часто нужна *интервальная оценка*, учитывающая неопределенность результатов прогноза. Построение доверительных интервалов традиционными статистическими методами невозможно из-за сложности задачи [109], а использование имитационных методов [97] затруднительно из-за большого времени расчетов [100]. В 1962 г. Л.В. Канторович [162] предложил другой подход к анализу данных — заменить минимизацию суммы квадратов отклонений на систему неравенств, которая решается с помощью методов линейного программирования. В этом случае результат прогноза сразу имеет вид интервала, поэтому этот метод и был назван «простым интервальным оцениванием» (ПИО) [23, 44]. С помощью этого метода было выполнено несколько работ в области аналитической химии [163].

Многомодальная регрессия

Методы многомерной калибровки естественно обобщаются на случай, когда ***X*** и ***Y*** блоки являются N-модальными матрицами [82]. Многомодальная регрессия может быть построена различными способами. С помощью методов, описанных выше (PARAFAC, Tucker3), блок предикторов ***X*** раскладывается в произведение 2D-матриц нагрузок, с помощью которых проводится оценка параметров. Эти методы можно рассматривать как обобщение метода PCR для многомодальных данных. Обобщением метода PLS является Tri-PLS декомпозиция 3D-матрицы ***X***, которую можно представить в «матрицированном» виде [164]:

$$^u\mathbf{X} \approx \mathbf{T} \cdot {}^u\mathbf{P}$$

где $^u\mathbf{X}$ — 2D-матрица (размерность $I \times KJ$), получаемая при развертке 3D-матрицы ***X*** (размерность $I \times K \times J$); **T** — 2D-матрица счетов (размерность $I \times A$); $^u\mathbf{P}$ — 2D-матрица весов (размерность $A \times KJ$), которая, в свою очередь, является разверткой для 3D-матрицы ***P***, представляемой как тензорное произведение двух 2D-матриц

$$\boldsymbol{P} = {}^J\mathbf{P} \otimes {}^K\mathbf{P}$$

Декомпозиция блока ***Y*** проводится аналогично:

$$^u\mathbf{Y} \approx \mathbf{U} \cdot {}^u\mathbf{Q}$$

Здесь так же, как и в обычном методе PLS, матрица счетов **T** выбирается таким способом, чтобы максимизировать корреляцию между векторами $\mathbf{t}_a$ и $\mathbf{u}_a$. Сама регрессионная задача $\mathbf{U} = \mathbf{TB}$ решается традиционным способом.

Математический аппарат, используемый при многомодальной калибровке, довольно сложен. Однако в настоящее время существуют программные продукты [165], позволяющие исследователям легко справляться с математическими трудностями. В литературе имеются многочисленные примеры использования мультимодальной калибровки в химии: применение кинетической спектрофотометрии для определения пестицидов [166], разрешение налагающихся пиков в ВЭЖХ с диодно-матричным детектором [167], определение следовых концентраций металлов [168], применение газовой хроматографии/масс-спектрометрии для определения следовых концентраций кленбутерола в биологических образцах [169], использование QSAR для предсказания антибактериальной активности соединений [170].

В работе [170] антибактериальная активность вещества (матрица ***Y***) представлена как функция уменьшения роста бактерий на 50% ($\log 1/IC_{50}$) в зависимости от продолжительности действия (2—10 часов) и уровня pH (5,6—8,0), т.е. блок данных ***Y*** есть 3D-матрица, а матрица дескрипторов **X** — это обычная 2D-матрица. Для исследования и моделирования стандартный метод PLS не подходит, поэтому перед применением PLS используется метод развертывания данных, чтобы проследить, как биологическая активность зависит от времени и уровня pH.

Нелинейная калибровка

В некоторых случаях, например задачах титрования, в сложных QSAR задачах, построить линейную

калибровочную зависимость невозможно. Кроме того, линейный подход требует большого количества данных, которые не всегда доступны. В этом случае используются два альтернативных подхода: множественная нелинейная регрессия или многомерная нелинейная калибровка.

Нелинейный регрессионный анализ [171] можно с успехом применять для решения задач количественного анализа в том случае, если число переменных невелико. Кроме того, необходимо располагать содержательной моделью, связывающей блоки предикторов **X** и блок откликов **Y**. По-видимому, круг таких задач не очень широк — в него входят, почти исключительно, кинетические и титриметрические задачи [172]. Так, этот подход применялся при анализе активности антиоксидантов [23], для решения обратной кинетической задачи [21, 95], для обработки результатов титрования [173, 174]. В работе [45] содержится подробный анализ проблем, с которыми сталкивается исследователь, применяющий нелинейный регрессионный анализ.

Альтернативой классической регрессии является формальный подход, который не требует знания содержательной модели, но предполагает наличие большого числа данных [175]. Для учета нелинейных эффектов предлагаются разнообразные усовершенствования [83] обычного метода PLS: INLR [176], GIFI-PLS [177], QPLS [178].

Помимо нелинейного PLS, в хемометрике активно применяется мето*д искусственных нейронных сетей* (ANN) [179, 180], имитирующий распространение сигналов в коре головного мозга. Этот метод с успехом используется для интерполяции функций и классификации. В последние 10 лет нейронные сети привлекли к себе большое внимание химиков, которые начали применять их для классификации [181], дискриминации [41] и калибровки [182, 183]. Затем, однако, наметилось некоторое охлаждение интереса, и использование ANN в хемометрике заметно снизилось. Причина заключена все в той же проблеме переоценки моделей, о которой шла речь выше.

Другим интересным методом нелинейного моделирования, имитирующим биологические процессы, является *генетический алгоритм* (GA), с успехом применяемый в хемометрике [184, 185]. Метод GA и его разновидность — иммунный алгоритм (IA) полезны в тех случаях, когда задача химического анализа не поддается формализации в терминах обычных целевых функций, например при разрешении многокомпонентных перекрывающихся хроматограмм [186]. Пример практического применения различных нелинейных подходов в хемилюминесцентном анализе приведен в работе [187].

Заключение

В статье рассмотрены основные достижения хемометрики за последние 20 лет. В то же время за пределами обзора остались очень многие актуальные направления и приложения как близкие к химии, так и далекие от нее, например методы *аналитического контроля процессов* (PAT). Это практическая область [188—190], в которую хемометрика органично привнесла свои подходы и методы [191—195] и где она бурно развивается.

Несмотря на всевозможные и разнообразные приложения, хемометрика несомненно является химической дисциплиной. Ее широкое распространение и применение в первую очередь обусловлено тем, что главной своей целью хемометрика видит решение конкретных, в основном химических задач, а потом находит уже существующие или разрабатывает новые математические и статистические приемы и алгоритмы. Развитие таких родственных областей, как биометрика и технометрика, которые зародились гораздо раньше, пошло по иному пути. Основное внимание в этих областях уделялось совершенствованию статистического аппарата, в результате теперь их можно скорее отнести к математико-статистическим дисциплинам, весьма мало востребованным биологами и инженерами.

Несмотря на все имеющиеся объективные и субъективные трудности, мы с оптимизмом оцениваем перспективы развития хемометрики в отечественной науке. Наблюдается растущий интерес к этой дисциплине как со стороны химиков-аналитиков, так и со стороны других специалистов — физиков и математиков. Сравнивая современное положение дел с ситуацией, имевшей место еще пять—семь лет назад, нельзя не отметить значительный рост публикаций российских ученых в отечественных и международных журналах, посвященных хемометрике.

Однако, если российские ученые хотят быть действительно активными участниками этого процесса, нужно предпринять срочные меры по развитию хемометрики. По нашему мнению, необходимо значительно поднять уровень преподавания хемометрики в университетах. Для этого следует подготовить (написать или перевести) учебник по хемометрике, разработать несколько типовых программ обучения для химиков-аналитиков, технологов, инженеров и т.п. Считаем, что можно ставить вопрос о введении соответствующих специальностей в магистерских специализациях, а также в кандидатских и докторских советах. И конечно же надо незамедлительно решить вопрос о подписке на ведущие журналы в этой области: Journal of Chemometrics и Chemometrics and Intelligent Laboratory Systems, которые сейчас нельзя найти ни в одной российской библиотеке.

ЛИТЕРАТУРА

1. *Geladi P., Esbensen K.* Chemom. Intell. Lab. Syst., 1990, v. **7**, p. 197.
2. *Massart D.L.* Chemometrics: a textbook. Elsevier, N. Y., 1988.
3. *Wold S.* Chemom. Intell. Lab. Syst., 1995, v. 30, p. 109.
4. *Blanco M., Villarroya I.* Trends Anal. Chem., 2002, v. 21, p. 240.
5. *Osborne B.G., Fearn T.* Near Infrared Spectroscopy in Food Analysis, Longman Scientific and Technical. Harlow, Essex, England, 1986.
6. *Blanco M. , Coello J., Iturriaga H., Maspoch S.,. Rovira E.* J. Pharm. Biomed. Anal., 1997,v. 16, p. 255.
7. *Espinosa A., Lambert D., Valleur M.* Hydrocarbon Process, 1995, v. 74, p. 86.
8. *Næs T., Irgens C., Martens H.* Appl. Stat., 1986, v. 35, p. 195.
9. *Martens H., Næs T.* Trends Anal. Chem., 1984, v. 3, p. 204.
10. *Wold S., Esbensen K., Geladi P.* Chemom. Intell. Lab. Syst., 1987, v. 2, p. 37.

11. *Shrager R.I.* Ibid., 1986, v. 1, p. 59.
12. *Geladi P., Grahn H.* Multivariate Image Analysis. Wiley, Chichester , 1996.
13. *Walczak B., Massart D.L.* Trends Anal. Chem., 1997, v. 16, p. 451.
14. *Belousov A.I., Verzakov S.A., von Frese J.* J. Chemom., 2002, v. 16, p. 482.
15. *Nomikos P., MacGregor J.F.* Am. Inst. Chem. Engin. J., 1994, v. 40, p. 1361.
16. *Geladi P., Esbensen K.* J. Chemom., 1991, v. 5, p. 97.
17. *Schaeferling M., Schiller S., Paul H., Kruschina M., Pavlickova P., Meerkamp M., Giammasi C., Kambhampati D.* Electrophoresis, 2002, v. 23, p. 3097.
18. *Frank I.E., Friedman J.H.* Technometrics, 1993, v. 35, p. 109.
19. *Wold S., Berglund A., Kettaneh N.* J. Chemom., 2002, v. 16, p. 377.
20. *Friedman J.* Lect. at the Gordon Conf. on Statist. and Chem. Engineering, Williamstown, MA, 2001. Доступно на http://www.amstat.org/sections/spes/GRC2001.htm [1 мая 2005].
21. *Родионова О.Е., Померанцев А.Л.* Кинетика и катализ, 2004, т. 45, с. 485.
22. *Koh H.-L., Yau W.-P., Ong P.-S., Hegde A.* Drug Discov. Today, 2003, v. 8, p. 889.
23. *Pomerantsev A.L., Rodionova O.Ye.*Chemom. Intell. Lab. Syst., 2005,v. 79, p.73.
24. *Грибов Л.А.* Математические методы и ЭВМ в аналитической химии, М., 1989.
25. *Siebert K.J.* J. Am. Soc. Brew. Chem., 2001, v. 59, p. 147.
26. *Varmuza K., Werther W., Krueger F.R., Kissel J., Schmid E.R.* Int. J. Mass Spectrom., 1999, v. 189, p. 79.
27. *Johnson W., Ehrlich R.* Environ. Forensics, 2002, v. 3, p. 59.
28. *Wise B.M., Gallagher N.B., Martin E.B.* J. Chemom., 2001, v. 15, p. 285.
29. *Brereton R.G.* Chemometrics: Data analysis for the laboratory and chemical plant. Chichester: Wiley, UK, 2003.
30. *Комарь Н.П.* Основы качественного химического анализа. Харьков, 1955.
31. *Грибов Л.А., Баранов В.И., Эляшберг М.Е.* Безэталонный молекулярный спектральный анализ. Теоретические основы. М.: Едиториал УРСС, 2002.
32. *Эляшберг М.* Успехи химии, 1999, т. 68, с. 579.
33. *Марьянов Б., Зарубин А., Шумар С.* Ж. аналит. химии, 2003, т. 58, с. 1126.
34. *Вершинин В.И., Дерендяев Б.Г., Лебедев К.С.* Методы компьютерной идентификации органических соединений. М.: Академкнига, 2002.
35. *Zenkevich I.G., Krõnicz B.* Chemom. Intell. Lab. Syst., 2003, v. 67, p. 51.
36. *Pletnev I.V., Zernov V.V.* Anal. Chim. Acta, 2002, v. 455, p. 131.
37. *Золотов Ю.А.* Аналитическая химия: проблемы и достижения. М.: Наука, 1992.
38. *Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С.* Успехи химии, 2003, т. 72, с. 706.
39. *Дворкин В.И.* Метрология и обеспечение качества количественного химического анализа. М.: Химия, 2001.
40. *Власов Ю.Г., Легин А.В., Рудницкая А.М.* Успехи химии, 2005, в печати.
41. *Калач А.В., Коренман Я.И., Нифталиев С.И.* Искусственные нейронные сети — вчера, сегодня, завтра. Воронеж: Воронеж. гос. технол. акад., 2002.
42. *Казаков С.П., Рябенко А.А., Разумов В.Ф.* Оптика и спектроскопия, 1999, т. 86, с. 537.
43. *Разумов В.Ф., Алфимов М.В.* Ж. науч. и прикл. физики, 2003, т. 46, с. 28.
44. *Rodionova O.Ye., Esbensen K.H., Pomerantsev A.L.* J. Chemom., 2004, v. 18, p. 402.
45. *Bystritskaya E.V., Pomerantsev A.L., Rodionova O.Ye.* Ibid., 2000, v. 14, p. 667.
46. *Bogomolov A., McBrien M.*Anal. Chim. Acta, 2003, v. 490, p. 41.
47. *Bogomolov A., McBrien M.*US Patent, US-2004-0126892-A1, 2004.
48. *Kucheryavski S., Polyakov V., Govorov A.* In: Progress in Chemometrics Research. Ed A.L. Pomerantsev. N. Y.: NovaScience Publishers, 2005, p. 3—11.
49. *Оскорбин Н.М., Максимов А.В., Жилин С.И.* Изв. АлтГУ, 1998, № 1, с. 35.
50. *Romanenko S.V. Stromberg A.G., Selivanova E.V., Romanenko E.S.* Chemom. Intell. Lab. Syst., 2004, v. 73, p. 7.
51. *Васильева И.Е, Кузнецов А.М, Васильев И.Л, Шабанова Е.В.* Ж. аналит. химии, 1997, т. 52, с. 1238.
52. *Шараф М.А., Иллмен Д.Л., Ковальски Б.Р.* Хемометрика. Пер. с англ. М.: Мир, 1987. [M. Sharaf, D. Illman, B. Kowalski. Chemometrics. N. Y.: Wiley, 1986].
53. *Massart D.L., Vandeginste B.G., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J.* Handbook of Chemometrics and Qualimetrics. Part A. Amsterdam: Elsevier, 1997.
54. *Vandeginste B.G., Massart D.L., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J.* Handbook of Chemometrics and Qualimetrics. Part B. Amsterdam: Elsevier, 1998.
55. *Næs T., Isaksson T., Fearn T., Davies T.* Multivariate Calibration and Classification, Christerer, UK, 2002.
56. *Kramer R.* Chemometric Techniques for Quantitative Analysis. Marcel-Dekker, 1998.
57. *Beebe K.R., Pell R.J., Seasholtz M.B.* Chemometrics: a Practical Guide. N. Y.: Willey, 1998.
58. *Malinowski E.R.* Factor Analysis in Chemistry. N. Y.: Wiley, 2nd edn, 1991.
59. *Martens H., Næs T.* Multivariate calibration. N. Y.: Wiley, 1989.
60. *Höskuldsson A.* Prediction Methods in Science and Technology. V. 1, Thor Publishing, Copenhagen, Denmark, 1996.
61. Аналитическая химия. Проблемы и подходы (в 2-х т.). Под. ред. Р. Кельнера, Ж.-М. Мерме, М. Отто, Г.М. Видмера. Пер. с англ. М.: Мир АСТ, 2004 [Analytical Chemistry. The Approved Text to FECS Curriculum Analytical Chemistry, Wiley-VCH, Weinheim].
62. *Марьянов Б.М.* Избранные главы хемометрики. Томск: Изд-во Томского ун-та, 2004.
63. *Эсбенсен К.* Анализ многомерных данных. Сокр. пер. с англ. Под ред. О. Родионовой. Изд-во ИПХФ РАН, 2005 [K.H. Esbensen. Multivariate Data Analysis — In Practice 4-th Ed., CAMO, 2000].
64. *Ferreira M.M.C.* J. Chemom., 2004, v. 18, p. 385.
65. The 9th Scandinavian Symposium on Chemometrics (SSC9) Доступно на http://www.conference.is/ssc9/ [1 мая 2005].
66. *Esbensen K., Rodionova O., Pomerantsev A., Startsev O., Kucheryavskiy S.* J. Chemom., 2003, v. 17, p. 422.
67. *Rodionova O.Ye.* Chemom. Intell. Lab. Syst., 2003, v. 67, p. 194.
68. *Kucheryavski S., Marks C., Varmuza K.* Ibid., 2005, v. 78, p. 138.
69. Home of Chemometry Consultancy. Доступно на http://www.chemometry.com/ [1 мая 2005].
70. Chemometrics literature database. Доступно на http://www.models.kvl.dk/ris/risweb.isa [1 мая 2005].
71. Chemometrics World. Доступно на http://www.wiley.co.uk/-wileychi/chemometrics/Home.html [1 мая 2005].
72. The Alchemist. Доступно на http://www.chemweb.com/-alchemist/ [1 мая 2005].

73. Российское хемометрическое общество. Доступно на http://rcs.chph.ras.ru/ [1 мая 2005].
74. Хемометрика в России. Доступно на http://www.chemometrics.ru/ [1 мая 2005].
75. The Unscramber. Доступно на http://www.camo.no/ [1 мая 2005].
76. Eigenvector Research, Inc. Доступно на http://www.eigenvector.com/ [1 мая 2005].
77. Umetrics. Доступно на http://www.umetrics.com/ [1 мая 2005].
78. SPSS. Доступно на http://www.spss.com/ [1 мая 2005].
79. STATISTICA. Доступно на http://www.statsoftinc.com/ [1 мая 2005].
80. MATLAB. Доступно на http://www.mathworks.com/ [1 мая 2005].
81. *Sanchez E., Kowalski B.R.* J. Chemom., 1988, v. 2 , p. 247.
82. *Smilde A., Bro R., Geladi P.* Multi-way Analysis with Applications in the Chemical Sciences. Chichester: John Wiley & Sons, 2004.
83. *Wold S., Trygg J., Berglund A., Antti H.* Chemom. Intell. Lab. Syst., 2001, v. 58, p. 131.
84. *Höskuldsson A.* J. Chemom., 2001, v. 58, p. 287.
85. *Geladi P., Burger J., Lestanderet T.* Chemom. Intell. Lab. Syst., 2004, v. 72, p. 209.
86. *Sanders G.H.W., Manz A.* Trends Anal. Chem., 2000, v. 19, p. 364.
87. *Демиденко Е.З.* Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981.
88. *Jy P.* Sampling for Analytical Purposes. Chichester: John Wiley & Sons, 1989.
89. *Kleingeld W., Ferreira J., Coward S.* J. Chemom., 2004, v. 18, p. 121.
90. Special Issue. Tutorials on sampling. Chemom. Intell. Lab. Syst., 2004, v. 74, p. 1—236.
91. *Walczak B., Massart D.L.* Ibid., 2001, v. 58, p. 15.
92. *Nelson P.R.C., Taylor P.A., MacGregor J.F.* Ibid., 1996, v. 35, p. 45.
93. *Haario H., Taavitsainen V.-M.* Ibid., 1998, v. 44, p. 77.
94. *Брин Э.Ф., Померанцев А.Л.* Хим. физика, 1986, т. 5, с. 1674.
95. *Gurden S.P., Westerhuis J.A., Bijlsma S., Smilde A.K.* J. Chemom., 2001, v. 15, p. 101.
96. *Карпухин О.Н.* Глобальные (стратегические) проблемы практического применения сложных математико-статистических методов (хемометрики), Докл. на 4-ом междун. симп. «Современные методы анализа многомерных данных» (WSC-4). Черноголовка, 14—18 февраля, 2005. Доступно на http://www.chemometrics.ru/-articles/karpukhin/ [1 мая 2005].
97. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988 [B. Efron, Ann. Stat., **7**, 1 (1979)].
98. EURACHEM/CITAC Guide, Quantifying Uncertainty in Analytical Measurement. 2nd ed., EURACHEM, Lisbon, Portugal, 2000.
99. *Faber K., Kowalski B.R.* Chemom. Intell. Lab. Syst., 1996, v. 34, p. 283.
100. *Pomerantsev A.L.* Ibid., 1999, v. 49, p. 41.
101. *Pulido A., Ruisánchez I., Boqué R., Rius F.X.* Trends Anal. Chem., 2003, v. 22, p. 647.
102. *Vershinin V.I.* Accreditation and Quality Assurance, 2004, v. 9, p. 415.
103. *Faber N.M.* Chemom. Intell. Lab. Syst., 2002, v. 64, p. 169.
104. *Faber N.M., Bro R.* Ibid., 2002, v. 61, p. 133.
105. *Berget I., Næs T.* J. Chemom., 2004, v. 18, p. 103.
106. *Jouan-Rimbaud D., Massart D.L., Saby C.A., Puel C.* Anal. Chim. Acta, 1997, v. 350, p. 149.
107. *Meloun M., Militký J., Hill M., Brereton R.G.* Analyst, 2002, v. 127, p. 433.
108. *Fernandez Pierna J.A., Wahl F., de Noord O.E., Massart D.L.* Chemom. Intell. Lab. Syst., 2002, v. 63, p. 27.
109. *Faber K.* Ibid., 2000, v. 52, p. 123.
110. *Faber N.M., Song X.-H., Hopke P.K.* Trends Anal. Chem., 2003, v. 22, p. 330.
111. *Bouveresse E., Massart D.L.* Vibrat. Spectrosc., 1996, v. 11, p. 3.
112. *Westad F., Martens H.* J. Near Infrared Spectros., 2000, v. 8, p. 117.
113. *Hubert M., Verboven S.* J. Chemom., 2003, v. 17, p. 438.
114. *Keller H.R., Massart D.L.* Chemom. Intell. Lab. Syst., 1992, v. 12, p. 209.
115. *Malinowski E.R.* J. Chemom., 1992, v. 6, p. 29.
116. *Gemperline P.J.* Anal. Chem., 1986, v. 58, p. 2656.
117. *Wold S.* Pattern Recognition, 1976, v. 8, p. 127.
118. *Jiang J.-H., Liang Y., Ozaki Y.* Chemom. Intell. Lab. Syst., 2004, v. 71, p. 1.
119. *Sanchez F.C., van de Borgaert B., Rutan S.C., Massart D.L.* Ibid., 1996, v. 34, p. 139.
120. *Shen H., Grande B., Kvalheim O.M., Eide I.* Anal. Chim. Acta, 2001, v. 446, p. 313.
121. *Windig W., Guilment J.* Anal. Chem., 1991, v. 63, p. 1425.
122. *Bogomolov A., Hachey M.* In: Progress in Chemometrics Research. Ed. A.L. Pomerantsev. N.Y.: NovaScience Publishers, 2005, p. 119—135.
123. *Diewok J., de Juan A., Marcel M., Tauler R., Lendl B.* Anal. Chem., 2003, v. 76, p. 641.
124. *Богомолов А. Ю., Ростовщикова Т.Н., Смирнов В.В.* Ж. физ. химии, 1995, т. 69, с. 1197.
125. *Seipel H.A., Kalivas J.H.* J. Chemom., 2004, v. 18, p. 306.
126. *Rodionova O.Ye., Houmuller L.P., Pomerantsev A.L., Geladi P., Burger J., Dorofeyev V.L., Arzamastsev A.P.* Anal. Chim. Acta, 2005, v. 549, p. 151.
127. *Guo Q., Wu W., Massart D.L., Boucon C., de Jong S.* Chemom. Intell. Lab. Syst., 2002, v. 61, p. 123.
128. *Shrager R.I.* Ibid., 1986, v. 1, p. 59.
129. *De Maesschalck R., Jouan-Rimbaud D., Massart D.L.* Ibid., 2000, v. 50, p. 1.
130. *Andrade J.M., Gomez-Carracedo M. P., Krzanowski W., Kubista M.* Ibid., 2004, v. 72, p. 123.
131. *Sun L.X., Danzer K.* J. Chemom., 1996, v. 10, p. 325.
132. *Myles A.J., Brown S.D.* Ibid., 2003, v. 17, p. 531.
133. *González-Arjona D., López-Pérez G., González A.G.* Talanta, 1999, v. 49, p. 189.
134. *Mark H.* Anal. Chem., 1987, v. 59, p. 790.
135. *Gemperline P.J., Boyer N.R.* Ibid., 1995, v. 67, p. 160.
136. *Mark H.L., Tunnell D.* Ibid., 1985, v. 57, p. 1449.
137. *Indahl U., Sing N.S., Kirkhuus B., Næs T.* Chemom. Intell. Lab. Syst., 1999, v. 49, p. 19.
138. *Downey G., Boussion J., Beauchene D.* J. Near Infrared Spectrosc., 1994, v. 2, p. 85.
139. *Flåten G.R., Grung B., Kvalheim O.M.* Chemom. Intell. Lab. Syst., 2004, v. 72, p. 101.
140. *Næs T., Indahl U.* J. Chemom., 1998, v. 12, p. 205.
141. *McElhinney J., Downey G., Fearn T.* J. Near Infrared Spectrosc., 1999, v. 7, p. 145.
142. *Zomer S., Brereton R., Carter J.F., Eckers C.* Analyst, 2004, v. 129, p. 175.
143. *Zernov V.V., Balakin K.V., Ivaschenko A.A., Savchuk N.P., Pletnev I.V.* J. Chem. Inf. Comput. Sci., 2003, v. 43, p. 2048.

144. *Sarker M., Rayens W.* J. Chemom., 2003, v. 17, p. 166.
145. *Herrero A., Zamponi S., Marassi R., Conti P., Ortiz M.C., Sarabia L.A.* Chemom. Intell. Lab. Syst., 2002, v. 61, p. 63.
146. *Garcia I., Sarabia L., Ortiz M.C., Aldama J.M.* Anal. Chim. Acta, 2004, v. 515, p. 55.
147. *Bijlsma S., Smilde A.K.* J. Chemom., 2000, v. 14, p. 541.
148. *Bro R.* Chemom. Intell. Lab. Syst., 1997, v. 38, p. 149.
149. *Kiers H.*J. Chemom., 2000, v. 14, p. 151.
150. *Hou T., Xu X.* Chemom. Intell. Lab. Syst., 2001, v. 56, p. 123.
151. *Hasegawa K. , Morikami K., Shiratori Y., Ohtsuka T., Aoki Yu., Shimma N.* Ibid., 2003, v. 69, p. 51.
152. *Faber N.M., Bro R., Hopke P.K.* Ibid., 2003, v. 65, p. 119.
153. *Andersson C.A., Bro R.* Ibid., 2000, v. 52, p. 1.
154. *del Rio F.J., Riu J., Rius F.X.* J. Chemom., 2001, v. 15, p. 773.
155. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. В 2-х т. М.: Финансы и статистика, 1987 [N.R. Draper, H. Smith, Applied regression analysis. N. Y.: Wiley].
156. *Höskuldsson A.* J. Chemom., 1988, v. 2, p. 211.
157. *de Jong S.* Chemom. Intell. Lab. Syst., 1993, v. 18, p. 251.
158. *Li B., Morris A.J., Martin E.B.* Ibid., 2004, v. 72, p. 21.
159. *Hubert M., Vanden Branden K.* J. Chemom., 2003, v. 17, p. 537.
160. *Vigneau E., Devaux M., Qannari M., Robert P.* Ibid., 1997, v. 11, p. 239.
161. *Geladi P.* Chemom. Intell. Lab. Syst., 2002, v. 60, p. 211.
162. *Канторович Л.В.* Сиб. матем. ж., 1962, т. 3, с. 701.
163. *Белов В.М., Суханов В.А., Унгер Ф.Г.* Теоретические и прикладные аспекты метода центра неопределенности. Новосибирск: Наука, 1995.
164. *Bro R.* J. Chemom., 1996, v. 10, p. 47.
165. *Bro R., Andersson C.A.* The N-way Toolbox for MATLAB, Version 2.02, 2003. Доступно на http://www.models.kvl.dk/-source [1 мая 2005].
166. *Ni Y., Huang C., Kokot S.* Chemom. Intell. Lab. Syst., 2004, v. 71, p. 177.
167. *Chen Z.P., Morris J., Martin E., Yu, R.-Q. Liang Y.-Z., Gong F.* Ibid., 2004, v. 72, p. 9.
168. *Fernández F.M., Tudino M.B., Troccoli O.E.* Anal. Chim. Acta, 2001, v. 433, p. 119.
169. *Garcia I., Sarabia L., Ortiz M.C., Aldama J.M.* Ibid., 2004, v. 515, p. 55.
170. *Eriksson L., Gottfries J., Johansson E., Wold S.* Chemom. Intell. Lab. Syst., 2004, v. 73, p. 73.
171. *Бард Й.* Нелинейное оценивание параметров. М.: Статистика, 1979. [*Y. Bard* Nonlinear Parameter Estimation. N. Y.: Academic Press, 1974].
172. *Померанцев А.Л.* Дисс. ... д-ра физ.-мат. наук. ИХФ РАН, Москва, 2003.
173. *Barry D.M., Meites L.* Anal. Chim. Acta, 1974, v. 68, p. 435.
174. *Марьянов Б.* В кн.: Химики ТГУ на пороге третьего тысячелетия. Томск: Изд-во ТГУ, 1998, с. 48—58.
175. *Eriksson L., Johansson E., Kettaneh-Wold N., Wold S.* Multi- and Megavariate Data Analysis, Umetrics, Umeå, 2001.
176. *Berglund A., Wold S.* J. Chemom., 1997, v. 11, p. 141.
177. *Berglund A., Kettaneh N.L.U., Wold S., Bendwell N., Cameron D.R.* Ibid., 2001, v. 15, p. 321.
178. *Wold S.* Chemom. Intell. Lab. Syst., 1992, v. 14, p. 71.
179. *Zupan J., Gasteiger J.* Anal. Chim. Acta, 1991, v. 248, p. 1.
180. *Zupan J., Gasteiger J.* Neural Network for Chemists, An Introduction. VCH, Weinheim, 1993.
181. *Wu W., Walczak B., Massart D.L., Heuerding E., Erni F.E., Last I.R., Prebble K.A.* Chemom. Intell. Lab. Syst., 1996, v. 33, p. 35.
182. *Smits J.R.M., Melssen W.J., Buydens L.M.C., Kateman G.* Ibid., 1994, v. 22, p. 165.
183. *Melssen W.J., Smits J.R.M., Buydens L.M.C., Kateman G.* Ibid., 1994, v. 23, p. 267.
184. *Hibbert D.B.* Ibid., 1993, v. 19, p. 277.
185. *Leardi R.* J. Chemom., 2001, v. 15, p. 559.
186. *Shao X., Chen Z., Lin X.* Chemom. Intell. Lab. Syst., 2000, v. 50, p. 91.
187. *Tortajada-Genaro L.A., Campíns-Falcó P., Verdú-Andrés J., Bosch-Reig F.* Anal. Chim. Acta, 2001, v. 450, p. 155.
188. *Bro R.* Chemom. Intell. Lab. Syst., 1999, v. 46, p. 133.
189. *Gabrielsson J., Lindberg N-O., Lundstedt T.* J. Chemom., 2002, v. 16, p. 141.
190. *Yoo C.K., Lee J.-M., Vanrolleghem P.A., Lee I.-B.*Chemom. Intell. Lab. Syst., 2004, v. 71, p. 151.
191. *MacGregor J., Kourti Th.* Control Engineering Practice, 1995, v. 3, p. 403.
192. *Померанцев А.Л., Родионова О.Е.* Методы менеджмента качества, 2002, № 6, с. 15.
193. *Pomerantsev A.L., Rodionova O.Ye.* In: Progress in Chemometrics Research. Ed. A.L. Pomerantsev. N. Y.: NovaScience Publishers, 2005, p. 209—227.
194. *Baroni M., Benedetti P., Fraternale S., Scialpi F., Vix P., Clementi S.* J. Chemom., 2003, v. 17, p. 9.
195. *Martens H., Martens M.* Multivariate Analysis of Quality: An Introduction. Chichester: John Wiley & Sons Ltd., 2001.