

УДК 543.153

Хемометрика: достижения и перспективы

О.Е.Родионова, А.Л.Померанцев

Институт химической физики им. Н.Н.Семенова Российской академии наук
119991 Москва, ул. Косыгина, 4, факс (495)939–7483

Рассмотрены основные хемометрические методы и модели, используемые для решения задач качественного и количественного анализа, а также для аналитического контроля технологических процессов. Показаны достижения в области хемометрики за последние 20 лет. Обсуждены тенденции и перспективы ее развития.
Библиография — 228 ссылок.

Оглавление

I. Введение

1. История хемометрики и ее место в системе знаний

Со времени опубликования перевода на русский язык единственной (до недавнего времени) книги по хемометрике [1] прошло 20 лет и за это время многое изменилось. В настоящее время хемометрические методы используют в различных областях науки и техники. Данный обзор посвящен в основном аналитической химии, где можно выделить три направления применения хемометрики: качественный и количественный анализ, контроль химического анализа и планирование эксперимента.[2] Главное внимание уделено первому направлению, меньшее — второму, и практически ничего не сказано о третьем. Такая расстановка акцентов обусловлена тем, что именно в таком порядке возрастает степень информированности российских химиков о хемометрических методах. В отечественных научных журналах публикуется большое число работ, посвященных планированию эксперимента [3] и метрологии.[4]

Число публикаций по хемометрике в мире стремительно возрастает: 15 лет назад выходило ~100 статей в год, к настоящему времени — более 5000 в год. Поэтому при подготовке обзора мы прибегли к разумным ограничениям рассматриваемой области. Анализ химических данных — наиболее важное направление в хемометрике. В последнее время оно очень быстро и плодотворно развивается, при этом химикам-аналитикам предложены не только новые методы обработки данных, но и новые подходы к постановке экспериментов.

Хемометрика — это синтетическая дисциплина, находящаяся на стыке химии † и математики, и как это часто бывает с пограничными дисциплинами, до сих пор не имеет общепризнанного определения. Наиболее популярное определение принадлежит Д.Массарту,[6] который считал, что хемометрика — это химическая дисциплина, в которой применяют 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. Наверное, многие согласятся с таким определением. Однако область науки должна определяться не используемыми методами и инструментами, а целями и задачами, которые она преследует. Разумеется, проблема извлечения информации из исходных данных крайне важна как для практики, так и для развития теории, однако не менее важна и задача конструирования таких экспериментов, в которых можно получить данные, содержащие нужную информацию. Эти два разнозначных аспекта — извлечение информации из данных и получение данных с нужной информацией — нашли отражение в современном определении хемометрики, данном С.Волдом.[7] Хемометрика решает следующие задачи в области химии:

— как получить химически важную информацию из химических данных,

— как организовать и представить эту информацию,

— как получить данные, содержащие такую информацию.

† Хемометрика как самостоятельная поддисциплина в рамках аналитической химии появилась в 1974 г.[5] Ее основателями можно считать американца Б.Ковальски и шведа С.Волда — внука С.Аррениуса.

О.Е.Родионова. Кадидат физико-математических наук, старший научный сотрудник ИХФ РАН. Телефон: (495)939–7483, e-mail: oksana@chph.ras.ru
А.Л.Померанцев. Доктор физико-математических наук, ведущий научный сотрудник того же института. Телефон: (495)939–7483, e-mail: forecast@chph.ras.ru
Область научных интересов авторов: хемометрика, математическая статистика, химическая физика.

Дата поступления 23 августа 2005 г.

Бурное развитие хемометрики именно в конце 1970-х годов связано с появлением в это же время быстродействующей вычислительной техники, которая повсеместно стала доступна ученым и инженерам. Это позволило на практике воплотить многие сложные алгоритмы, особенно для анализа данных, полученных в многооткликовых и многофакторных экспериментах. Как следствие появилось более сложное оборудование, способное производить многократно большее число измерений. Однако оказалось, что большое количество данных еще не означает, что необходимой информации достаточно. Поэтому химики-аналитики стали активно применять хемометрические методы для извлечения такой информации и для подтверждения того, что сделанные выводы достоверны. В результате был достигнут первый несомненный успех. Оказалось, что очень часто традиционные аналитические методы, требующие больших затрат труда, времени, уникального оборудования, дорогих реактивов, могут быть заменены гораздо более быстрыми и дешевыми косвенными методами. Наиболее ярко эта тенденция проявилась при использовании ИК-спектроскопии, особенно в ближней области, прежде считавшейся малоинформативной из-за высокого и трудно устранимого шума, обусловленного интенсивным поглощением воды и эффектом рассеяния в спектрах отражения.[8] Поэтому первые работы по хемометрике были посвящены методам анализа спектроскопических данных,[9–11] построению градуировочных моделей (градуировок) с помощью метода главных компонент [12] и метода проекций на латентные структуры.[13]

Говоря об истории хемометрики, нельзя не упомянуть ученых, которые задолго до 1970-х годов заложили основы хемометрического подхода. Начать, очевидно, нужно с К.Гаусса, предложившего в 1795 г. метод наименьших квадратов. Первым хемометриком можно, по-видимому, считать и У.Госсета (известного под псевдонимом Стьюдент), работавшего аналитиком на пивоварне и уже в конце XIX в. начавшего применять методы анализа химических данных.[14] В начале XX в. появилась работа К.Пирсона,[15] в которой был предложен метод главных компонент, несколько позднее были опубликованы работы Р.Фишера — автора многочисленных статистических методов, таких как метод максимума правдоподобия и факторный анализ,[16] а также пионерских работ по планированию эксперимента.[17] Среди отечественных ученых следует отметить прежде всего В.Налимова, внесшего значительный вклад в теорию планирования химического эксперимента.[18]

Хемометрика зародилась и длительное время развивалась в рамках аналитической химии, и специалисты в этой области до сих пор остаются главными пользователями хемометрических методов. Однако со временем проявилась тенденция, расцененная некоторыми исследователями как выход хемометрики «из-под крыла» аналитической химии и превращение ее в самостоятельную дисциплину. Два обстоятельства дали основание для такого вывода. Во-первых, это усложнение математического аппарата, используемого в хемометрике. Десять лет назад химики-аналитики смогли усвоить и принять многомерный подход к анализу данных, т.е. такие методы, как проекция на латентные структуры [19] или разложение по сингулярным значениям.[20] Однако затем, в период повсеместного увлечения хемометриков новыми методами анализа данных (мультимодальным подходом,[21] вэйвлет-анализом,[22] методом опорных векторов [23] и т.п.), наметился некоторый разрыв между химиками и хемометриками: химики не понимали, что и зачем делают хемометрики, которые в свою очередь не понимали, почему их новые методы не востребованы в аналитической химии. Во-вторых, появились многочисленные приложения, в которых хемометрический подход стал успешно применяться в областях, далеких от аналитической химии, например при многомерном статистическом контроле процессов,[24] анализе изображений,[25] в биологии.[26] Такое непонимание проявилось и в том, что на последней конференции «Хемометрика в аналитической химии» (CAC-2004, Лиссабон)[27] многими участниками обсуждался вопрос — является ли хемометрика по-прежнему частью аналитической химии?

Из сказанного выше видно, что хемометрика тесно связана с математикой, особенно с математической статистикой. Большинство химиков-аналитиков понимают необходимость применения статистических методов в химическом анализе и используют их для вычисления средних величин, отклонений, пределов обнаружения, проверки гипотез и т.п. Они считают, что именно эти простые операции и составляют основу хемометрического подхода в аналитической химии. Однако лишь немногие исследователи понимают, что это не так, и могут использовать все разнообразие хемометрических методов для анализа химических данных.

Следует отметить, что для эффективного практического применения хемометрики совсем не обязательно знать, например, статистическую теорию метода главных компонент, достаточно понимать основы, базовые идеи этого подхода. А вот что действительно необходимо знать, так это методы подготовки данных, принципы отбора переменных, и самое важное — уметь правильно интерпретировать проекции данных (нагрузки и счета) в пространстве главных компонент. Этот навык, как показала многолетняя практика, можно приобрести и без глубоких математических познаний. Данный обзор написан в рамках концепции, в соответствии с которой основные принципы, методы и достижения хемометрики изложены с привлечением минимально необходимого числа математических формул, а геометрическая интерпретация превалирует над алгебраической.

Многие методы и алгоритмы, используемые в хемометрике, математики [28] справедливо считают плохо обоснованными. Специалисты в области хемометрики рассматривают 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,[29] они создают новые методы анализа и делают это так быстро, что математики, по словам американского статистика Д.Фридмана,‡ не успевают не только раскритиковать их за это, но и просто понять, что же происходит в хемометрике. Такой подход контрастирует с ситуацией, сложившейся в биометрике,[30] — области, которую, образно говоря, можно считать «старшей сестрой» хемометрики. Со времен Р.Фишера в биометрике традиционно применяют только хорошо апробированные классические методы математической статистики, такие как факторный анализ или линейный дискриминационный анализ. Вместе с тем специалисты в другой близкой области — психометрике [31] — активно разрабатывают новые подходы к анализу данных. Так, самый популярный в хемометрике метод проекции на латентные структуры был разработан Г.Волдом [32] именно для применения в этой области.§

Благодаря такому «агрессивному» подходу к анализу данных хемометрика нашла многочисленные применения в

‡ J.Friedman. *Boosting and Bagging*. Доступно на http://www.amstat.org/sections/spes/GRC2001.htm.

§ Интересно, что в начале 1970-х годов господствовало мнение, согласно которому проекционные методы малоприемлемы в естественных науках, но иногда могут быть полезны в общественных науках как методы отыскания эффективных комбинаций переменных (см.[33] том 2, стр. 48).

разных разделах химической науки (например, в физической химии для исследования кинетики,[34] в органической химии для предсказания активности соединений по их структуре (QSAR),[35] в химии полимеров,[36] в теоретической и квантовой химии [37]), а также в смежных и далеких от химии областях (например, в пивоварении,[38] астрономии,[39] при решении судебных споров о защите окружающей среды [40] и контроле качества производства полупроводников [41]) (см.¶).

Некоторые направления хемометрики развивались в СССР, а позднее в России. Так, еще в 1950-е годы в Харьковском государственном университете под руководством Н.П.Комаря проводили исследования по математическому описанию равновесий.[43] Позднее появились работы Л.А.Грибова [44] и М.Е.Эляшберга [45] по спектральным методам, Б.М.Марьянова по титриметрии,[46] Б.Г.Дерендяева и В.И.Вершинина по методам компьютерной идентификации органических соединений,[47] И.Г.Зенкевича по хроматографии.[48] Активное использование хемометрического подхода характерно [49] для научной школы академика Ю.А.Золотова.[2] Исследования в близкой к хемометрике области QSAR ведутся под руководством академика Н.С.Зефирова.[50] Метрологические аспекты и контроль качества химического анализа исследуются в работах В.И.Дворкина.[51] В Санкт-Петербургском государственном университете группа ученых под руководством Ю.Г.Власова работает над созданием сенсорных систем, известных под названием «электронный язык»,[52] а в Воронежской технологической академии разрабатывают аналогичные системы, известные как «электронный нос».[53] Во всех этих областях интенсивно используют хемометрические методы. В.Ф.Разумов и его коллеги из Института химической физики РАН (Черноголовка) применяют многомерные методы анализа данных при решении задач химической кинетики.[54, 55] За последние годы в России появились новые группы ученых, разрабатывающих и применяющих хемометрические методы: О.Е.Родионова,[56] А.Л.Померанцев,[57] А.Ю.Богомолов [58, 59] (в Москве); С.В.Кучерявский,[60] С.И.Жилин [61] (в Барнауле); С.В.Романенко [62] (в Томске); Е.В.Шабанова и И.Л.Васильев [63] (в Иркутске).

2. Информационное и программное обеспечение

Мы уже упоминали широко известную в России монографию [1], в которой отражено положение дел в хемометрике, сложившееся к середине 1980-х годов. В настоящее время наиболее полно хемометрические методы изложены в двухтомнике [64, 65], написанном группой авторов под руководством Д.Массарта. В этом издании наряду с подробным описанием основных хемометрических методов и приемов приведено много примеров их практического приложения. Кроме того, существует множество изданий, ориентированных на разный круг читателей. Так, студентам и специалистам в области аналитической химии, начинающим осваивать хемометрику, проще начать с монографии [42]; исследователям, занимающимся спектральным анализом, будут интересны книги [66, 67]. Много полезной информации можно найти в работе [68]. Нельзя не упомянуть монографию Е.Малиновского,[69] которую до сих пор многие химики-аналитики считают лучшим учебником в рассматриваемой области. Теоретические основы хемометрики изложены в работах [70, 71]. Недавно на русский язык переведен учебник [72], содержащий краткое описание методов хемометрики. Интересное введение в хемометрику написал Б.М.Марьянов.[73]

¶ Подробный анализ использования хемометрических методов в различных областях приведен в монографии Р.Бреретона,[42] к которой мы и отсылаем заинтересованного читателя.

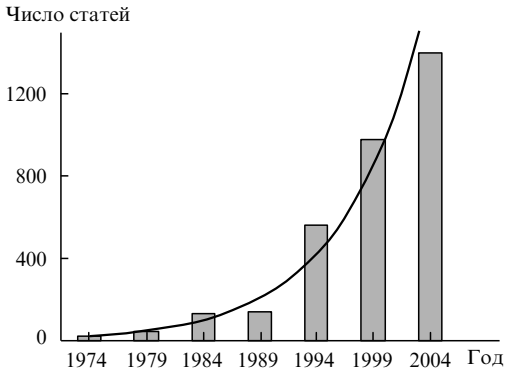

Рис. 1. Диаграмма, иллюстрирующая рост числа статей по хемометрике, на примере публикаций в периодических изданиях издательства Elsevier.

Небольшим тиражом (для участников трех научных школ по хемометрике) в России был издан сокращенный перевод учебника, написанного К.Эсбенсеном.[74]

Проблемам хемометрики посвящены два специализированных журнала — *Journal of Chemometrics* и *Chemometrics and Intelligent Laboratory Systems*. Статьи, в которых отражены результаты применения хемометрических методов для решения прикладных задач, регулярно печатают более 50 научных журналов, такие как *Analytical Chemistry*, *Analytica Chimica Acta*, *Analyst*, *Talanta*, *Trends in Analytical Chemistry*, *Journal of Chromatography*, *Computers and Chemical Engineering*, *Vibrational Spectroscopy* и т.д. Число публикаций, авторы которых используют хемометрические методы в качестве основного инструмента для анализа и обработки экспериментальных данных, растет с каждым годом (рис. 1).

Вопросы хемометрики специалисты рассматривают как на небольших региональных конференциях и семинарах, так и на регулярных международных конференциях. Наиболее авторитетными являются конференция «Хемометрика в аналитической химии» (Chemometrics in Analytical Chemistry)[27] и «Скандинавский симпозиум по хемометрике» (Scandinavian Symposium on Chemometrics).† В России, начиная с 2002 г., проходят ежегодные международные школы-симпозиумы «Современные методы анализа многомерных данных».[75–77] Теоретические и прикладные аспекты хемометрики широко представлены в виде интернет-ресурсов, в основном это англоязычные сайты,‡ однако есть и несколько российских ресурсов.§

В качестве программного обеспечения в хемометрике применяют специализированные пакеты программ,¶ позволяющие наглядно и быстро обрабатывать данные в интерактивном режиме. Широко используют и статистические

† *The 9th Scandinavian Symposium on Chemometrics (SSC9)*. Доступно на http://www.conference.is/ssc9

‡ *Home of Chemometry Consultancy*. Доступно на http://www.chemometry.com; *Chemometrics Literature Database*. Доступно на http://www.models.kvl.dk/ris/risweb.isa (1 мая 2005); *Chemometrics World*. Доступно на http://www.wiley.co.uk/wileychi/chemometrics/Home.html; *The Alchemist*. Доступно на http://www.chemweb.com/alchemist

§ *Российское хемометрическое общество*. Доступно на http://rcs.chph.ras.ru; *Хемометрика в России*. Доступно на http://www.chemometrics.ru

¶ *The Unscrambler*. Доступно на http://www.camo.no; *Eigenvector Research* Inc. Доступно на http://www.eigenvector.com; *Umetrics*. Доступно на http://www.umetrics.com

пакеты общего назначения.† Часто исследователи пишут процедуры сами, например в кодах MATLAB,‡ и публикуют их для свободного доступа в Интернете или в книгах (см.[78]).

3. Обозначения и термины

В обзоре использованы следующие обозначения. Скалярные переменные выделены курсивом, например *s*. Векторы (столбцы) обозначены прямыми полужирными строчными буквами, например **x**, а матрицы — прописными, например **W**; мультимодальные матрицы выделены курсивом, например ***G***. Элементы массивов обозначены той же, но строчной буквой. Например, w_{ij} означает элемент матрицы **W**, индекс *i* обозначает строку матрицы и изменяется от 1 до *I*; индекс *j* соответствует номеру столбца и меняется от 1 до *J*. Аналогичные обозначения применены и для других индексов, например $a = 1, \ldots, A$. Операция транспонирования обозначается верхним индексом t, например $\mathbf{X}^{\mathrm{t}}$.

В русскоязычной научной литературе до сих пор не сложилась общепризнанная система хемометрических терминов. Некоторые понятия ранее были переведены неверно или неточно. Например, фундаментальный в хемометрике метод PLS первоначально расшифровывался как *Partial Least Squares*. На русский язык это переводилось как «частичные (или частные) наименьшие квадраты», что не соответствовало сути метода. В последнее время трактовка аббревиатуры PLS изменилась на *Projection on Latent Structures*, что в дословном переводе означает «проекция на латентные структуры». Мы считаем, что именно так и следует называть этот метод. Термины soft и hard, часто используемые в хемометрике для характеристики методов моделирования, должны, по нашему мнению, переводиться словами «формальный» и «содержательный», которые точнее отражают их суть. При переводе понятия *N*-way мы использовали термин «*N*-модальный». Может быть, это и не лучшее решение, но применение традиционного термина тензорного анализа «валентность» в контексте аналитической химии мы сочли неудачным. Во многих случаях переводчики просто избегали давать русские названия ключевым понятиям хемометрики, например таким как scores и loadings, используя вместо них сложные эвфемизмы. Мы полагаем, что в хемометрике невозможно обойтись без таких понятий, как счета и нагрузки или их аналогов.

Специалисты в области хемометрики, как и в любой другой области знаний, часто используют аббревиатуры — сокращенные названия методов, алгоритмов, специальных терминов. Несмотря на то, что у некоторых из них есть общепринятые русские аналоги, в этом обзоре мы дали аббревиатуры от оригинальных английских названий. Ниже приведен список использованных сокращений (курсивом выделены переводы, употребляемые впервые).

ALS (Alternating Least–Squares) — *чередующиеся наименьшие квадраты;* 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; MIA (Multivariate Image Analysis) — многомерный анализ изображений; MSC (Multiplicative Signal Correction или Multiplicative Scatter Correction) — *множественная коррекция сигнала* или *мультипликативная коррекция рассеяния*; MSPC (Multivariate Statistical Process Control) — многомерный статистический контроль процессов; NAS (Net Analyte Signal) — *полезный аналитический сигнал*; NIPALS (Non-linear Iterative Projections by Alternating Least-Squares) — *нелинейное итерационное проецирование при помощи чередующихся наименьших квадратов*; OSC (Orthogonal Signal Correction) — ортогональная коррекция сигнала; 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 (Qualitative Structure–Activity Relationship) — количественная связь структура–активность; RMSEC (Root-Mean Square Error of Calibration) — среднеквадратичный остаток градуировки; 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 (SIMple Partial Least Squares regression) — *элементарные последовательные наименьшие квадраты*; SMCR (Self-Modeling Curve Resolution) — *метод автомодельного разрешения кривых*; SPC (Statistical Process Control) — статистический контроль процессов; SVD (Singular Value Decomposition) — разложение по сингулярным значениям; SVM (Support Vector Machine) — метод опорных векторов; WFA (Window Factor Analysis) — оконный факторный анализ.

II. Данные и модели, используемые в химическом анализе

1. Химические данные и информация

Экспериментальные данные — основной объект хемометрики. Следуя классификации, предложенной в работе [79], рассмотрим типичную структуру химических данных (рис. 2).

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

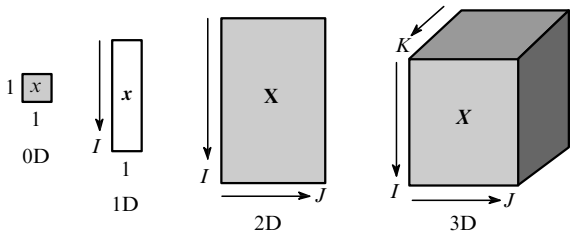

Рис. 2. Графическое представление данных разной модальности.

† *SPSS*. Доступно на http://www.spss.com; *STATISTICA*. Доступно на http://www.statsoftinc.com

‡ *MATLAB*. Доступно на http://www.mathworks.com

модальные данные, т.е. набор результатов нескольких измерений, относящихся к одному образцу. Примерами таких данных являются спектр или хроматограмма. С математической точки зрения их можно интерпретировать как 1D-вектор (столбец или строку), каждый элемент которого соответствует некоторой переменной (длине волны, времени удерживания). Число переменных определяет размерность данных.

Следующий распространенный тип химических данных — двухмодальные данные. Их представляют 2D-матрицей — таблицей из чисел, имеющей I строк и J столбцов. Типичным примером таких данных может служить набор спектров, снятых для I образцов на J длинах волн. Каждая строка в 2D-матрице представляет объект (в данном случае образец), а каждый столбец — переменную (длину волны). Отнесение данных к объектам (образцам) или переменным (каналам) имеет большое значение для их интерпретации. Однако не всегда такое разделение очевидно. Например, при анализе данных, полученных методом высокоэффективной жидкостной хроматографии с диодно-матричным детектированием, для 30 точек по времени на 28 длинах волн мы можем составить матрицу из 28 строк и 30 столбцов, а можем, наоборот, считать длины волн переменными, а времена удерживания — объектами. В большинстве случаев в эволюционных (развивающихся во времени) экспериментах объекты соответствуют временам, т.е. образец, изменяющийся во времени, рассматривают как серию образцов.[42]

По мере развития гибридных методов[2] большое внимание стали уделять трех- (и более) модальным данным.[80] Их можно представить 3D-матрицей, геометрический образ которой — параллелепипед, в котором каждое ребро соответствует своему типу переменной. Пример четырех- и даже восьмимодальных данных можно найти в работе[80].

Данные можно объединять в блоки. В простейшем случае — это один блок **X**. Такой случай чаще встречается в качественном анализе, например в задаче разделения спектров. В количественном анализе, основанном на регрессионных зависимостях, используют данные, объединенные в два (и более) блока. Блок предикторов (например, 2D-матрица спектров **X**) и блок откликов (например, 1D-вектор концентраций **y**) составляют набор стандартных данных, по которым строят градуировочную модель

$$\mathbf{y} = \mathbf{Xb}.$$

Встречаются и более сложные структуры данных, включающие три (и более) блока.[81] Для их анализа применяют специальные методы маршрутного моделирования.[82]

Может показаться, что такая систематизация данных — размерность, модальность, блочность — носит формальный характер и представляет интерес только для математиков, но не для химиков. Однако это не так. За последние годы кардинально изменились критерии того, какие данные можно считать большими. Если в начале 1970-х годов матрицу данных (например, спектров) считали большой, если в ней было 20 столбцов (переменных, например длин волн) и 100 строк (объектов, например образцов), то в настоящее время большой можно считать матрицу с 1 000 000 столбцов и 400 000 строк.[29] При обработке таких массивов их приходится делить на блоки и интерпретировать по очереди. Чтобы разделение не было формальным, необходимо участие опытного химика, понимающего суть дела. Понятие модальности тоже ввели 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 гиперспектральные измерения,[83] и использованием микрочипов[84] структура данных будет усложняться.

Основная задача хемометрики состоит в извлечении из массива данных нужной химической информации. Понятие информации — ключевое для хемометрики. Что считать информацией, зависит от решаемой задачи. Иногда достаточно знать, что некоторое вещество присутствует в системе, но часто необходимо получить и количественные значения. Данные могут содержать нужную информацию, они даже могут быть избыточными, а могут и не содержать информации. Но во всех случаях данные включают нежелательную составляющую — шум (например, погрешности), который скрывает нужную информацию.

Для иллюстрации рассмотрим следующий идеализированный эксперимент. Положим, что мы имеем систему, состоящую из смеси трех веществ A, B и C без примесей, и известны точные (без погрешностей) спектры $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ всех компонентов. Слово «спектры» употреблено в общем смысле: это могут быть любые многомерные данные, например хроматограммы, в которых λ — время удерживания. Требуется определить концентрации компонентов по спектру смеси $x(\lambda)$, который также можно получить без погрешностей. Если каждый спектр содержит значения для 30 длин волн (времен) λ, то для решения этой задачи можно составить 30 уравнений относительно трех неизвестных концентраций c_A, c_B и c_C:

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

Очевидно, что для получения нужной информации столько уравнений не требуется, можно оставить только три, соответствующие любым[§] трем длинам волн. Таким образом, исходные данные (30-мерный 1D-вектор) избыточны по отношению к искомой информации: используя три любые точки спектра, мы будем получать одни и те же значения концентраций.

Рассмотрим более реалистичный пример, допустив, что все спектры содержат некоторую случайную погрешность. Тогда концентрации, определенные по разным тройкам длин волн, будут отличаться. Эти оценки можно усреднить и получить более точные значения концентраций. Заметим, что этого же можно достичь с помощью повторных экспериментов. Однако такой путь неэффективен, поскольку требует больших затрат сил и времени. Гораздо проще уменьшить неопределенность количественного анализа за счет увеличения числа переменных (каналов, длин волн) в одном эксперименте. Этот вывод отражает первый важный принцип хемометрики — использование многомерного подхода при конструировании экспериментов и анализе их результатов.

Выше отмечалось, что данные всегда (или почти всегда) содержат шум различной природы. Это могут быть случайные погрешности, сопровождающие эксперимент: сдвиг базовых линий, погрешности в определении сигналов, неточности при подготовке и проведении эксперимента. Во многих случаях шум — это часть данных, не содержащая искомой информации. Так, если бы в рассмотренном выше примере нужно было определить концентрацию только двух веществ A и B, то вещество C было бы нежелательной примесью, а вклад от него — шумом. Что считать шумом, а что — информацией? Данный вопрос всегда решается с учетом поставленных целей и методов, используемых для ее достижения. В этом заключается второй принцип хемометрического подхода к анализу данных.

Шум и избыточность в данных проявляют себя через корреляционные связи между переменными. Возвращаясь к

§ Строго говоря, не любым: необходимое условие — система должна иметь единственное решение.

идеализированному примеру, можно заметить, что в матрице «чистых» спектров **S**, имеющей размерность 3×30 (3 строки (образца) на 30 столбцов (длин волн)), только три столбца будут линейно независимыми. Зафиксировав эту тройку, любой четвертый столбец можно представить в виде их линейной комбинации. Разумеется, то, что их ровно три, не случайность — ведь именно столько веществ присутствует в нашей системе. Это число называется рангом матрицы **S**, и оно играет важную роль в хемометрическом анализе. Рассматривая тот же пример в более реалистичном варианте (с учетом погрешностей) можно заметить появление дополнительных корреляций в данных. Это произойдет если, например, концентрация третьего вещества C будет существенно меньше погрешности (шума). Этих данных уже недостаточно для надежного определения всех трех концентраций, и эффективный ранг матрицы будет равен 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.¶

Понятие эффективного (химического) ранга и скрытых (латентных) переменных, число которых равно этому рангу, лежит в основе третьего важнейшего принципа хемометрики.[69] Проиллюстрируем его применение на следующем примере. Предположим, что имеется несколько (I) смесей веществ A, B и C, но их точные спектры $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ неизвестны. Можно получить спектры этих образцов в виде двухмодальных данных, и построить матрицу **X** размером $I \times 30$. Путем обычного математического анализа можно определить ранг данной матрицы. Это число дает информацию о том, сколько компонентов присутствует в системе или по крайней мере о том, сколько их можно различить.

Таким образом, в массиве химических данных почти всегда имеются внутренние скрытые связи между переменными, приводящие к множественным корреляциям — мультиколлинеарности. Такое свойство может проявляться как избыточность данных, что позволяет повысить качество оценок. Однако при неправильном методе обработки данных мультиколлинеарность может негативно сказаться на качестве анализа. Например, нельзя применять метод множественной линейной регрессии в условиях мультиколлинеарности.[74] Для регрессионного анализа таких данных необходимо использовать специальные методы, например метод ридж-регресии[86] или проекционные подходы.[71]

Существенным источником шума в данных может быть отбор образцов. Теория пробоотбора, значительный вклад в которую внес П.Жи,[87] приобрела большую популярность в последнее время.[88] Ее многочисленные приложения можно найти в публикации[89], полностью посвященной этой теме. Другая проблема, с которой может столкнуться химик-аналитик — это пропуски в данных,[90] обусловленные разными причинами: из-за отказа приборов, вследствие выхода за пределы обнаружения, нехватки образцов для исследования и т.п. Большинство хемометрических методов не допускают пропусков в данных, поэтому для заполнения пропусков используют специальные методы, среди которых наиболее распространен метод на основе итерационного алгоритма. Каждая итерация состоит из двух шагов. На первом шаге оценку параметров модели проводят так, как будто данные известны полностью. Для этого пропуски заполняют некоторыми априорно допустимыми значениями, например средними по окружающим элементам массива данных. На втором шаге с помощью полученной модели находят наиболее вероятные значения пропущенных данных и переходят к следующей итерации. Для заполнения пропусков также используют подход, основанный на методе максимума правдоподобия.[91] Детали таких алгоритмов в большой степени зависят от модели описания данных.

¶ Различие в понятиях причинности и корреляции забавно проиллюстрировано в книге[85], в которой приведен пример высокой положительной корреляции между числом жителей и числом аистов в городе Ольденбург (Германия) за период с 1930 по 1936 гг. Разумеется, эти две переменные связаны между собой корреляционными связями, возникающими из-за того, что в системе присутствует третья скрытая переменная, с которой они обе связаны причинными связями.

2. Общая методология анализа данных. Модели и методы

Хемометрические методы анализа данных можно разделить на две группы, соответствующие двум главным задачам: 1) исследование данных, например, с целью классификации и дискриминации; 2) предсказание новых значений, например с целью градуировки. Методы первой группы, как правило, оперируют с одним блоком данных, а второй — как минимум с двумя блоками (предикторов и откликов). В зависимости от поставленных целей, методы решения могут быть направлены на предсказание в рамках диапазона условий эксперимента (интерполяция) или за пределами этого диапазона (экстраполяция). Методы разделяют на формальные (soft), называемые также «черными», и содержательные (hard) — «белые». При использовании формальных моделей[92] данные описывают эмпирической зависимостью (как правило, линейной), справедливой в ограниченном диапазоне условий. В этом случае необязательно знать механизм исследуемого процесса, однако такой метод не позволяет решать задачи экстраполяции. Параметры формальных моделей лишены физического смысла и должны интерпретироваться с помощью соответствующих математических методов. Содержательное моделирование[93] базируется на физико-химических принципах и позволяет экстраполировать поведение системы в новых условиях. Параметры «белой» модели имеют физический смысл, и их значения могут помочь при интерпретации найденной зависимости. Однако такой метод может быть применен, только если модель известна *a priori*. Каждый из подходов имеет достоинства и недостатки,[36] и у каждого из них есть свои сторонники и противники. Исторически сложилось так, что в нашей стране интенсивно развивался содержательный метод, а в других странах — преимущественно формальный метод. За последнее время появилось много работ, авторы которых рассматривают так называемые «серые» модели,[94] объединяющие сильные стороны обоих методов. Проиллюстрируем разные подходы к моделированию примерами из аналитической химии.

Часто объектами математического моделирования в аналитической химии служат титриметрические процессы, отличающиеся многообразием химических реакций и регистрируемых сигналов. Уравнения кривых титрования нередко весьма сложны и не могут быть записаны в явной форме относительно регистрируемого сигнала. Это затрудняет применение содержательных моделей для решения обратной задачи — оценки параметров по измеренным точкам кривой. Тем не менее, используя современные вычислительные системы, такую задачу все же можно решить в рамках «белого» моделирования.[95] В работе[96] замечено, что по форме титриметрические кривые напоминают графики обратных гиперболических и тригонометрических функций. Исходя из этого предлагается использовать формальные («черные») зависимости, составленные из тригонометрических функций arcsin, arccos и т.п. В соответствии с компромиссным («серым») подходом, предложенным в работе[46], с помощью замены переменных содержательная модель преобразуется в кусочно-линейную. Затем для оценки параметров применяют метод ALS,[97] суть которого состоит в после-

довательном приближении модели к данным: сначала линейными регрессионными методами оценивают линейные параметры при фиксированных значениях нелинейных, а затем нелинейные оценивают в процедуре наискорейшего спуска при найденных ранее фиксированных оценках линейных параметров. Процедуры чередуются до сходимости результатов.

Интерес к «черным» и «серым» методам моделирования обусловлен трудностями выбора и подтверждения правильности содержательной модели. Во многих случаях все сводится к простому перебору небольшого числа конкурирующих зависимостей, в результате которого обычно выбирают простейшую модель с минимальной невязкой. Однако это не доказывает правильности выбранного метода и может привести к грубым ошибкам. Часто исследователи используют модели, справедливо названные «розовыми»,† в основе которых лежат идеализированные зависимости, плохо соответствующие реальным артефактам, присутствующим в данных, — дрейфам базовых линий, аномальным погрешностям и т.п. Формальные многофакторные линейные модели и надлежащие методы их анализа гораздо лучше «приспособлены» к учету артефактов. Такие модели работают и в тех случаях, когда о содержательном физико-химическом подходе не может быть и речи. Основанием для использования линейных моделей служит тот факт, что любую, даже очень сложную, но непрерывную зависимость в достаточно малой области можно представить в виде линейной функции. Принципиальным в таком случае является вопрос о том, какую область можно считать допустимой, иначе говоря, вопрос о границах применимости построенной формальной модели. Ответ на него дают методы проверки (валидации) моделей.

При надлежащем построении модели исходный массив данных состоит из двух независимо полученных достаточно представительных наборов. Первый набор — обучающий — используют для идентификации модели, т.е. для оценки ее параметров. Второй набор — проверочный — служит только для проверки модели. Построенную модель применяют к данным из проверочного набора, и полученные результаты сравнивают с проверочными данными методом тест-валидации. По итогам сравнения принимают решение о правильности и точности моделирования. В некоторых случаях объем данных слишком мал для такой проверки. Тогда применяют другой метод — перекрестной проверки (кросс-валидации).[98] В соответствии с этим методом 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 применяют ее к исключенному набору. На следующем цикле исключенные данные возвращают и удаляют уже другую часть данных (следующие 10%) и опять строят модель, которую применяют к исключенным данным. Эту процедуру повторяют до тех пор, пока все данные не побывают в числе исключенных (в нашем случае — 10 циклов). Наиболее (но неоправданно) распространен вариант перекрестной проверки, в котором данные исключаются по одному (LOO). В регрессионном анализе используют также проверку методом коррекции размахом.[74] Следует отметить, что та или иная проверочная процедура должна применяться не только в количественном, но и в качественном анализе при решении задач дискриминации и классификации.

Любой результат, полученный при анализе и моделировании экспериментальных данных, заключает в себе неопределенность. Количественная оценка или качественное суждение могут измениться при повторном эксперименте в результате проявления разнообразных случайных и систематических погрешностей как изначально присутствующих в исходных данных, так и внесенных на стадии моделирования.[99] Неопределенность в количественном анализе характеризуется либо числом — стандартным отклонением,[100] — либо интервалом — доверительным[101] или прогнозным.[56] В качественном анализе применяют метод проверки статистических гипотез,[102] в котором неопределенность характеризуется вероятностью принятия неверного решения.[103] Методы оценки неопределенности при моделировании многомерных[104] и многомодальных[105] данных вызывают большой интерес хемометриков. Для описания различных аспектов надежности аналитического метода применяют специальные характеристики: специфичность, селективность, предел обнаружения, отношение сигнал : шум.[73] Актуальным методом их определения является подход с использованием концепции NAS.[106] Многомерный вектор NAS определяется как часть полного сигнала (спектра), которую используют для моделирования и прогноза.[107] Оставшуюся часть сигнала, включающую погрешности и вклады от посторонних компонентов, рассматривают как шум. Концепция NAS была применена к задаче определения предела обнаружения при анализе двух- (см.[108]) и трехмодальных[109] данных. Полученные результаты нашли многочисленные практические приложения, одно из которых рассмотрено ниже.

Надежность аналитического метода в значительной мере зависит от того, какие данные были использованы для построения и проверки соответствующей модели. Наличие выбросов[110] или малоинформативных данных снижает точность модели, и наоборот, присутствие представительных (влиятельных) образцов в эксперименте[111] существенно улучшает качество модели. Оценку влиятельности данных можно проводить классическими регрессионными методами,[112] а можно выполнять с помощью нестатистических процедур.[56] При использовании построенной модели для определения интересующих показателей сталкиваются с похожими проблемами. Может оказаться, что метод неприменим к некоторым образцам (выброс в прогнозе[113]) или дает неточный результат. Оценка неопределенности метода не в среднем,[114] а индивидуально для образцов — сложная задача, над решением которой работают разные группы исследователей.[115] Именно их усилия определяют успешное решение таких практически важных задач, как перенос градуировочных моделей с одного прибора на другой,[116] отбор переменных,[117] построение робастных методов анализа данных.[118]

III. Методы качественного анализа. Исследование, классификация и дискриминация

1. Метод главных компонент

Современные приборы могут легко проводить огромное число измерений в единицу времени. Например, если использовать *in situ* спектроскопический датчик для получения спектра на 300 длинах волн каждые 15 с, то за час работы он даст матрицу данных размерностью 240×300, т.е. 72 000 чисел. Однако вследствие мультиколлинеарности доля полезной информации в таком массиве может быть относительно невелика. Для выделения полезной информа-

† См. О.Н.Карпухин. Глобальные (стратегические) проблемы практического применения сложных математико-статистических методов (хемометрики). Доклад на четвертом международном симпозиуме «Современные методы анализа многомерных данных» (WSC-4). Черноголовка, 14–18 февраля 2005 г.. Доступно на http://www.chemometrics.ru/articles/karpukhin

ции в хемометрике используют методы сжатия данных (в отличие от традиционного подхода, когда из данных выделяют только результаты отдельных особо значимых измерений). Чтобы представить исходные данные в этих методах, используют новые скрытые переменные. При этом должны выполняться два условия. Во-первых, число новых переменных (химический ранг) должно быть существенно меньше числа исходных переменных, и, во-вторых, потери от такого сжатия данных должны быть сопоставимы с шумом в них. Сжатие данных позволяет представить полезную информацию в более компактном виде, удобном для визуализации и интерпретации.

Для сжатия данных чаще всего используют метод PCA,[19] который лежит в основе других аналогичных хемометрических методов, включая EFA,[119] WFA,[120] ITTFA,[121] а также многих методов классификации, например метода SIMCA.[122] Метод главных компонент заключается в декомпозиции исходной 2D-матрицы **X**, т.е. в представлении ее в виде произведения двух 2D-матриц **T** и **P** (см.[74]),

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

В этом уравнении **T** называется матрицей счетов (scores), **P** — матрицей нагрузок (loadings), а **E** — матрицей остатков (рис. 3). Число столбцов — $\mathbf{t}_a$ в матрице **T** и $\mathbf{p}_a$ в матрице **P** — равно эффективному (химическому) рангу матрицы **X**. Эту величину обозначают A и называют числом главных компонент, естественно, оно меньше числа столбцов в матрице **X**.

Для иллюстрации метода PCA вернемся к примеру, рассмотренному в разделе II.1. Матрица спектров смесей **X** может быть представлена как произведение матрицы концентраций **C** и матрицы спектров чистых компонентов **S**

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

Число строк в матрице **X** равно числу образцов (I), и каждая 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 точных спектров присутствует в разложении (2) в транспонированном виде, так как число ее строк равно числу длин волн (J), а число столбцов равно A. Как отмечено выше, при анализе реальных экспериментальных данных, отягощенных погрешностями, представленными матрицей **E**, эффективный ранг A может не совпадать с реальным числом компонентов в смеси. Чаще он бывает больше за счет влияния неконцентрационных факторов, например температуры.

Задача разделения экспериментальной матрицы **X** на «чистые» составляющие, соответствующие концентрациям **C** и спектрам **S** (понимаемым в обобщенном смысле), — предмет особой области в хемометрике, названной разделением кривых (curve resolution).[123] В этой области можно выделить два направления. В первом используют метод автомодельного разрешения кривых (SMCR)[124] и оно ориентировано в основном на приложение к гибридной хроматографии.[125] Для реализации автомодельного подхода применяют методы формального моделирования, например PCA, EFA, которые не используют содержательное знание об исследуемой системе. В рамках этого подхода можно выделить метод SIMPLISMA,[126] применяющий простой, но весьма эффективный подход, основанный на отборе переменных.[127] Во втором направлении, напротив, учитывают априорную информацию о процессах и применяют «серые» модели.[128] Это направление находит приложение при исследовании кинетики [34] и термодинамики.[129] Ключевым моментом в таких задачах является определение химического ранга системы — числа главных компонент A.[130] В идеальном случае предсказанные спектры **S** и концентрации **C** должны быть близки к истинным, хотя их невозможно восстановить точно. Причина этого не только в погрешностях эксперимента, но и в том, что спектры могут частично перекрываться. Когда PCA применяют для разделения данных на химически осмысленные компоненты, как в уравнении (2), его часто называют факторным анализом, в отличие от формального анализа главных компонент.[131]

Метод главных компонент эффективен не только в задачах разделения. Он применяется при анализе любых химических данных. В этом случае матрицы счетов **T** и нагрузок **P** уже нельзя интерпретировать как спектры и концентрации, а число главных компонент A — как число химических компонентов, присутствующих в исследуемой системе. Тем не менее, даже формальный анализ счетов и нагрузок оказывается очень полезным для понимания структуры данных. Дадим простейшую двумерную иллюстрацию метода PCA.

Данные, состоящие только из двух переменных x_1 и x_2, которые связаны сильной корреляцией, представлены на рис. 4,*a*. Те же данные в новых координатах представлены на рис. 4,*b*. Вектор нагрузок $\mathbf{p}_1$ первой главной компоненты (PC1) определяет направление новой оси, вдоль которой происходит наибольшее изменение данных. Проекции всех исходных точек на эту ось составляют вектор $\mathbf{t}_1$. Вторая

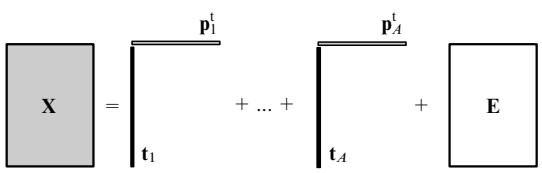

Рис. 3. Графическое представление метода главных компонент.

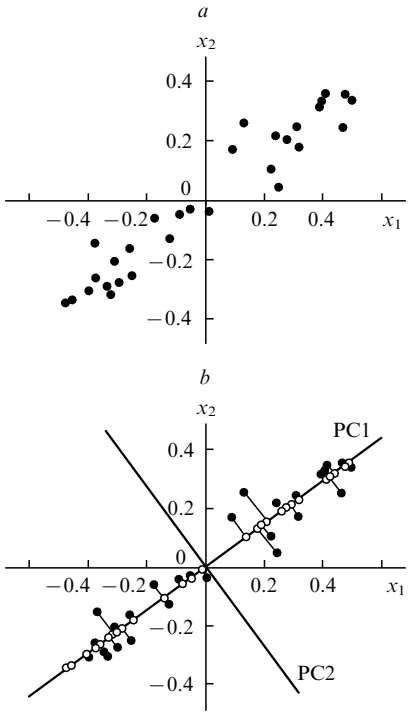

Рис. 4. Графическая иллюстрация метода главных компонент. *a* — данные в исходных координатах, *b* — данные в координатах главных компонент.

главная компонента $\mathbf{p}_2$ ортогональна первой, и ее направление (PC2) соответствует наибольшему изменению в остатках (показаны 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 выбирают направление наибольшего изменения, данные проецируют на эту ось, вычисляют новые остатки и т.д. (алгоритм NIPALS).[74] В соответствии с другим популярным алгоритмом сжатия данных — SVD (см.[132]) — строят ту же декомпозицию (1) без итераций. Выбор числа главных компонент A (другими словами, остановку итерационной процедуры) проводят с использованием критериев, показывающих точность достигнутой декомпозиции. Пусть исходная матрица $\mathbf{X}$ имеет I строк и J столбцов, и в разложении (1) участвуют A главных компонент. Величины

$$\mu_a = 100\sum_{i=1}^{I}\mathbf{t}_{ia}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2\,, \qquad (3)$$

$$E_a = 100\left(1 - \sum_{i=1}^{I}\sum_{j=1}^{J}\mathrm{e}_{ij}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2\right), \quad a = 1, \ldots, A$$

называют нормированным собственным значением и объясненной дисперсией соответственно. Их обычно изображают графиком зависимости от числа a. Резкое изменение этих величин указывает на нужное значение числа главных компонент. Для правильного выбора A необходимо использовать метод тест-валидации либо кросс-валидации.

Уравнения (1) не содержат свободного члена, поэтому для декомпозиции данных иногда их следует сначала отцентрировать (т.е. вычесть среднее по столбцам) и нормировать.

Метод главных компонент можно трактовать как проецирование данных на подпространство меньшей размерности. Возникающие при этом остатки $\mathbf{E}$ рассматривают как шум, не содержащий значимой химической информации. В этом подпространстве можно ввести меру близости образцов, называемую расстоянием Махаланобиса (Mahalanobis),[133] с помощью которой удается решить многие проблемы качественного анализа. Другим мощным методом анализа данных в проекционном подпространстве является метод прокрустова вращения.[134]

При исследовании данных методом PCA особое внимание уделяют графикам счетов и нагрузок. Они несут информацию о структуре данных. 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. Применяя этот подход в задачах хроматографического анализа,[42] можно, например, установить, что линейные участки на графике счетов соответствуют областям чистых компонентов на хроматограмме, искривленные участки представляют области наложения пиков, а число таких участков соответствует числу различных компонентов в системе. Если график счетов используют для анализа взаимоотношений образцов, то график нагрузок применяют для исследования роли переменных. На графике нагрузок каждую переменную отображают 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 позволяет получать полезную информацию из данных.[74]

Приведем пример практического использования PCA в химическом анализе. В работе [135] рассмотрена возможность

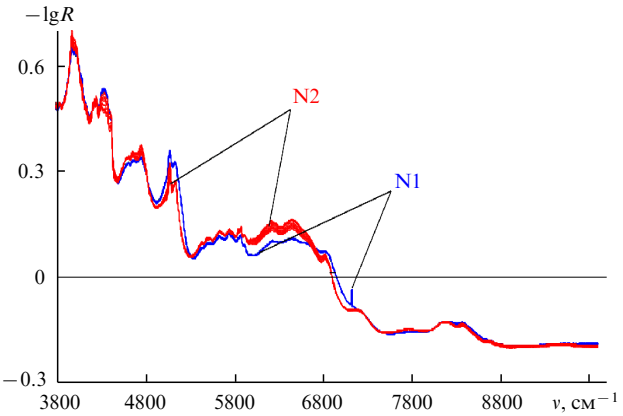

Рис. 5. Спектры, преобразованные процедурой MSC.[135]
Здесь и на рис. 6 приняты обозначения: N1 — истинные таблетки, N 2 — фальсифицированные таблетки.

применения ИК-спектроскопии в ближней области для обнаружения фальсифицированных лекарств. Были исследованы образцы истинных (N1, 10 штук) и поддельных (N2, 10 штук) таблеток популярного спазмолитического средства. Двадцать спектров диффузного рассеяния $R(\lambda)$ были сняты с помощью прибора «Bomem MB160» с приставкой «Powder Samplir» в диапазоне 3800 – 10 000 см$^{-1}$ (1069 длин волн) без специальной подготовки образцов. Исходные данные были преобразованы к виду $-\lg R$, центрированы и подготовлены процедурой MSC (рис. 5).[74] Отрицательные значения сигнала обусловлены тем, что для фона и спектров образцов использовали различные регулировки усиления.

На графике PCA счетов ($\mathbf{t}_1$, $\mathbf{t}_2$) этих спектров (рис. 6,*а*) четко видны две группы точек, соответствующих истинным и фальсифицированным таблеткам. Разброс точек в группе N2 (контрафакт) существенно больше, чем в группе N1 (оригиналы). Это можно объяснить лучшим контролем качества при легальном производстве. В этом примере достаточно использовать только две главные компоненты, для которых $\mu_1 = 94\%$, $\mu_2 = 4.9\%$, $E_2 = 99\%$.

2. Классификация и дискриминация

Рассмотренный ниже пример относится к задачам классификации. Это весьма широкий класс задач качественного химического анализа, в которых требуется установить принадлежность образца к некоторому классу. Задачи классификации можно разделить на две группы. К первой относятся так называемые задачи без обучения (unsupervised), в которых не используют обучающий набор, и их можно рассматривать как разновидность исследовательского анализа. Именно этот подход применен в рассмотренном примере с фальшивыми таблетками. Задачи второй группы — классификация с обучением (supervised), которые также называют задачами дискриминации. Для их решения применяют обучающий набор образцов, о которых имеется априорная информация о принадлежности к классам. Методы решения задач классификации без обучения основаны главным образом на PCA-декомпозиции с последующим анализом расстояний между классами,[136] построением дендрограмм, использованием нечетких множеств [137] и т.п. В работе [138] применялось прокрустово вращение, а в работах [139–141] — расстояние Махаланобиса. Однако если возможно проведение дискриминации, то этим методам следует отдавать предпочтение.

Обучающий набор образцов используют для построения модели классификации, т.е. свода правил, с помощью кото-

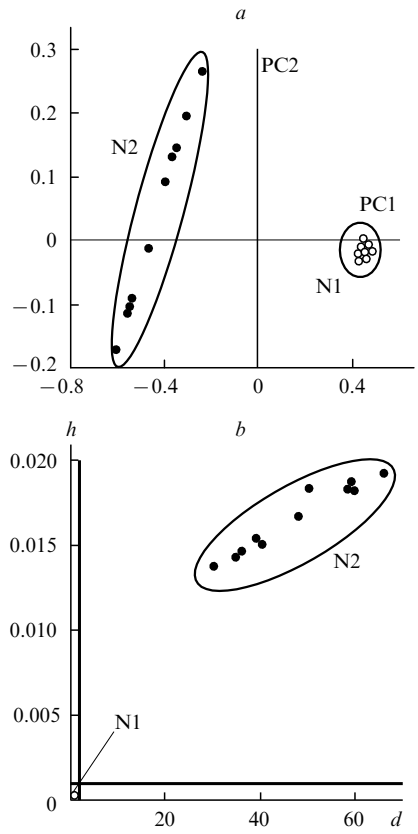

Рис. 6. Определение фальсифицированных лекарств с использованием PCA-счетов (*a*) и метода SIMCA (*b*).

рых новый образец может быть отнесен к тому или иному классу. После того как модель (или модели) построена, ее необходимо проверить с помощью методов тест- или кроссвалидации и определить, насколько она точна. Если проверка успешная, модель готова к практическому применению. В аналитической химии обычно классифицируют наборы мультиколлинеарных данных (спектры, хроматограммы), поэтому дискриминантная модель почти всегда многомерна и основана на соответствующих проекционных подходах — PCA, PLS. Можно отметить использование линейного дискриминантного анализа в ИК-спектроскопии в ближней области,[142] а также канонического дискриминантного анализа.[143] Одним из наиболее распространенных является метод SIMCA,[144] разработанный С.Волдом.[122]

В основе метода SIMCA лежит предположение о том, что все объекты в одном классе имеют как сходные свойства, так и индивидуальные особенности. При построении дискриминационной модели необходимо учитывать только сходство, отбросив особенности как шум. Для этого каждый класс из обучающего набора независимо моделируют, используя метод PCA с разным числом главных компонент A. После этого вычисляют расстояния между классами, а также расстояния от каждого класса до нового объекта. В качестве таких метрик используют две величины. Расстояние от объекта до класса (d) находят как среднеквадратичное значение остатков **e**, возникающих при проецировании объекта на класс

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

Эту величину сравнивают со среднеквадратичным остатком внутри класса

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

Вторую величину — расстояние от объекта до центра класса (h) — вычисляют как размах (квадрат расстояния Махаланобиса)

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

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

Применение метода SIMCA для дискриминации таблеток иллюстрирует рис. 6,*b*. В качестве класса использовали подлинные таблетки, а на графике показаны расстояния d и h от образцов подделок до этого класса. Вертикальная и горизонтальная линии определяют правила, по которым новый объект может быть отнесен к классу настоящих таблеток. Видно, что все образцы фальшивых таблеток находятся далеко от класса подлинных, поэтому легко могут быть дискриминированы. В приведенном масштабе точки, соответствующие образцам истинных таблеток, сливаются в одну, которая практически совпадает с началом координат.

Для дискриминации химических данных помимо метода SIMCA используют похожий на него метод DASCO,[145] а также методы KNN,[146] SVM [147, 148] и многие другие. Мощным инструментом является метод PLS–DA.[149] Его идея состоит в том, что дискриминационные правила для K классов задают линейными регрессионными уравнениями вида

$$\mathbf{XB} = \mathbf{D},$$

где **X** — полная матрица всех исходных данных ($I \times J$), **B** — матрица неизвестных коэффициентов ($J \times K$), а **D** — специальная матрица ($I \times K$), состоящая из нулей и единиц. При построении матрицы **D** единицы ставят только в те строки (образцы), которые принадлежат классу, соответствующему номеру столбца. Регрессионная задача решается методом PLS (см. ниже), что позволяет в дальнейшем применять построенную регрессию для предсказания принадлежности новых образцов. Для этого прогнозируют отклик нового образца, и результат сравнивают с нулем или единицей.

3. Трехмодальные методы

Метод главных компонент разработан для анализа данных, которые можно представить в виде двухмодальной 2D-матрицы. Однако в последнее время химики-аналитики все чаще имеют дело с трех- (и более) модальными данными более сложной структуры. Такие данные получают, используя, например, гибридные [150, 151] и эволюционные методы.[152] Для сжатия массивов данных применяют специальные подходы, три из которых (чаще всего используемых) кратко рассмотрены в этом подразделе. Наиболее полное и систематическое описание этих методов вместе с многочисленными примерами их применения при решении задач химического анализа приведено в монографии [80]. Краткий обзор методов и алгоритмов, используемых для анализа трехмодальных данных, представлен в статье [153]. Эти же алгоритмы применяют для обработки данных, полученных в результате гиперспектральных измерений,[83] а также для анализа изображений.[21]

Метод развертывания (unfolding)[154] — простейший способ анализа трехмодальных данных, с помощью которого 3D-матрица $\boldsymbol{X}$ размерности $I \times J \times K$ разворачивается в

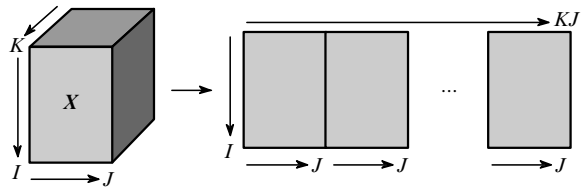

Рис. 7. Графическое представление метода развертывания в плоскую матрицу.

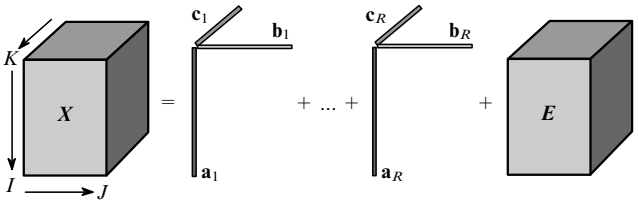

Рис. 9. Графическое представление модели PARAFAC с R компонентами.

обычную 2D-матрицу $^{u}\mathbf{X}$ размерности $I \times JK$ (рис. 7). При этом I называют «основной» модой. После развертывания можно применять метод главных компонент (см. раздел III.1). Такой подход часто оказывается эффективным (см., например,[36]), хотя он имеет ряд недостатков; во-первых, в качестве основной моды можно выбирать любое из трех направлений, т.е. имеется неоднозначность развертывания; во-вторых, теряется связь между соседними точками, так как при переходе от 3D- к 2D-матрице уже не учитывается, что измерения x_{ikj} и x_{ik+lj} являются соседними, а это может быть существенно.

Алгоритм Tucker3 (см.[155]) позволяет обрабатывать трехмодальные данные, сохраняя их первоначальную структуру, а значит, и последовательность измерений, например порядок длин волн в спектре, либо последовательность точек по времени в хроматограмме. Исходные 3D-данные $\boldsymbol{X}$ представляют в виде трех обычных 2D-матриц нагрузок (**A**, **B**, **C**) и трехмодального керн-массива $\boldsymbol{G}$. Схема такого разложения данных приведена на рис. 8. Каждый элемент исходной 3D-матрицы $\boldsymbol{X}$ можно записать в виде суммы

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

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

Метод PARAFAC (см.[153]) отличается от модели Tucker3 тем, что каждая мода представляется одним и тем же числом главных компонент R. Разложение строят так, чтобы минимизировать сумму квадратов остатков e_{ijk}

$$x_{ijk} = \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr} + \mathrm{e}_{ijk}\,. \qquad (5)$$

Основным достоинством этого метода является единственность разложения. Так, если исследовали смесь нескольких химических веществ, то при правильном выборе числа главных компонент матрицы нагрузок представляют точные спектры исходных веществ. Графическая схема метода PARAFAC представлена на рис. 9.

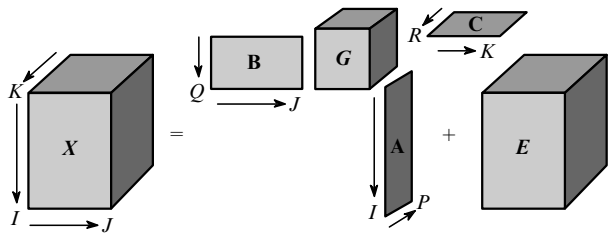

Рис. 8. Графическое представление модели Tucker3.

Коды MATLAB алгоритма PARAFAC можно найти в работе [150]. Так как матрицы нагрузок в разложении (5) определяют с помощью итерационной процедуры, этот метод требует очень большого объема вычислений. В настоящее время ведутся работы по ускорению вычислительных процедур. Критический анализ последних достижений в этой области проведен авторами работы [155]. Алгоритмы всех рассмотренных методов декомпозиции трехмодальных данных приведены в исследовании [156].

IV. Методы количественного анализа. Градуировочные модели

1. Линейная градуировка

Для решения задач количественного анализа [2] используют два блока данных. Первый блок **X** — матрица аналитических сигналов (например, спектров, хроматограмм и т.п.); второй блок **Y** — матрица соответствующих химических показателей (например, концентраций). Число строк (I) в этих матрицах равно числу образцов сравнения, число столбцов (J) в матрице **X** соответствует числу каналов (длин волн), на которых записывается сигнал, число столбцов (K) в матрице **Y** равно числу химических показателей, т.е. откликов. Цель градуировки — построение математической модели, связывающей блоки **X** и **Y**, с помощью которой можно в дальнейшем предсказывать значения показателей **y** по новой строке значений аналитического сигнала **x**.[13]

Простейшая градуировочная модель — это одномерная регрессия ($J = 1$, $K = 1$)[157]

$$y = a + bx,$$

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,[33]

$$\mathbf{y} = \mathbf{Xb}.$$

При использовании этих моделей обычно предполагают, что значения факторов x_{ij} известны точно, а погрешности присутствуют только в блоке **y**. В связи с этим различают два подхода к построению модели: первый называют прямой градуировкой, второй — обратной градуировкой.[158] При первом подходе в качестве независимых факторов используют химические показатели (**X** = **C**), а в качестве откликов — спектральные измерения (**Y** = **S**). Ранее считали, что прямая модель лучше соответствует предположению о безошибочности блока **X** и кроме того, она согласуется с законом Бугера – Ламберта – Бера.[72] При втором подходе **Y** = **C**, **X** = **S**. В настоящее время данный подход превалирует в хемометрике, поскольку он удобнее с практической точки зрения, так как непосредственно предсказывает нужный аналитический показатель (например, концентрацию **C**) по измеренному сигналу (спектру **S**). Кроме того, современные

регрессионные методы (PCR, PLS) позволяют работать с данными, в которых погрешности присутствуют в обоих блоках.

Для иллюстрации различных методов градуировки вернемся к примеру, рассмотренному в разделе II.1. Теперь мы наполним его конкретным содержанием, смоделировав данные **X** и **Y**. Положим, что имеется смесь двух веществ A и B ($K = 2$) и прибор, позволяющий измерять аналитический сигнал s (спектр) на 101 канале ($J = 101$). Соответствующие спектры «чистых» веществ ($c_A = c_B = 1$) представлены на рис. 10,*a* (кривые *A* и *B*). Спектры сильно перекрываются, поэтому невозможно выделить «селективные» каналы для оценки концентраций. На рис. 10,*b* представлены девять модельных спектров ($I = 9$) различных смесей веществ A и B, в которые внесена случайная погрешность со стандартным отклонением 0.05. Они будут использованы как обучающий набор.

Для построения одномерной градуировки мы взяли интенсивности $s(\lambda_{50})$ девяти сигналов для канала 50 и изобразили их на рис. 10,*c* в зависимости от концентраций c_A и c_B веществ A (точки *1*) и B (точки *2*). Соответствующие градуировочные зависимости $s = bc$ показаны прямыми.

Точность градуировки принято характеризовать величиной RMSEC

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

где y_i и $\hat{y}_i$ — соответственно известные и предсказанные значения химического показателя (концентрации) для образцов сравнения $i = 1, \ldots, I$; F — число степеней свободы:[42] $F = I - 1$ для одномерной регрессии без свободного члена. Очевидно, чем меньше RMSEC, тем точнее описываются обучающие данные. Кроме того, качество градуировки характеризуется и коэффициентом корреляции R^2 между величинами **y** и $\hat{\mathbf{y}}$: чем он ближе к единице, тем выше качество градуировки. Соответствующие значения приведены в подрисуночной подписи к рис. 10,*c*. Графические зависимости подтверждают, что из-за недостатка «приборной» селективности одномерная градуировка неудовлетворительна. Градуировка с помощью множественной регрессии будет рассмотрена ниже.

Покажем, как работает многомерная модель, построенная с помощью метода PCR.[86] В методе PCR используют обратную градуировку, так что **Y** = **C**, **X** = **S**. Применяя метод PCA, матрицу **X** можно разложить по формуле (1), причем для нашего примера $A = 2$. Получившиеся векторы нагрузок $\mathbf{p}_1$ и $\mathbf{p}_2$ показаны на рис. 10,*a* (кривые PC1 и PC2). При сравнении графиков, приведенных на рис. 10,*a*,*b*, видно, что первая главная компонента описывает гладкий тренд в данных, тогда как вторая компонента представляет зашумленные отклонения от этого тренда. Полученную матрицу счетов **T** используют как блок независимых факторов (предикторов) в регрессии на блок откликов **Y**, т.е. **Y** = **Tb**. Результаты градуировки методом PCR представлены на рис. 10,*d*, на котором изображены предсказанные значения концентраций $\hat{\mathbf{y}}$ в зависимости от соответствующих известных значений **y** (точки *1* для вещества A и точки *2* для вещества B), а также линии регрессий, которые сливаются. Значения RMSEC и R^2, приведенные для этого графика, свидетельствуют о том, что метод PCR позволяет достичь высокой «математической» селективности и получить оценки концентраций веществ A и B с гораздо большей точностью, чем при одноканальной градуировке. В методе PCR число степеней свободы в уравнении (6) составляет

$$F = I - A.$$

Выше было сказано о том, что каждая хемометрическая модель нуждается в полноценной проверке. В нашем примере такую проверку проводили с помощью проверочного набора (тест-валидация), состоящего из пяти образцов (смесей A и B). Результаты проверки для вещества B представлены на рис. 11,*a*. Здесь в координатах «измерено – предсказано», приведены данные для девяти образцов, участвовавших в градуировке, и пяти проверочных образцов. Также приведены значения среднеквадратичных остатков RMSEC и RMSEP и коэффициенты корреляции для обучающего (R_c^2) и проверочного (R_t^2) наборов. Значения RMSEP вычисляют аналогично RMSEC (см. формулу (6)), но только для образцов из проверочного набора. При этом число степеней сво-

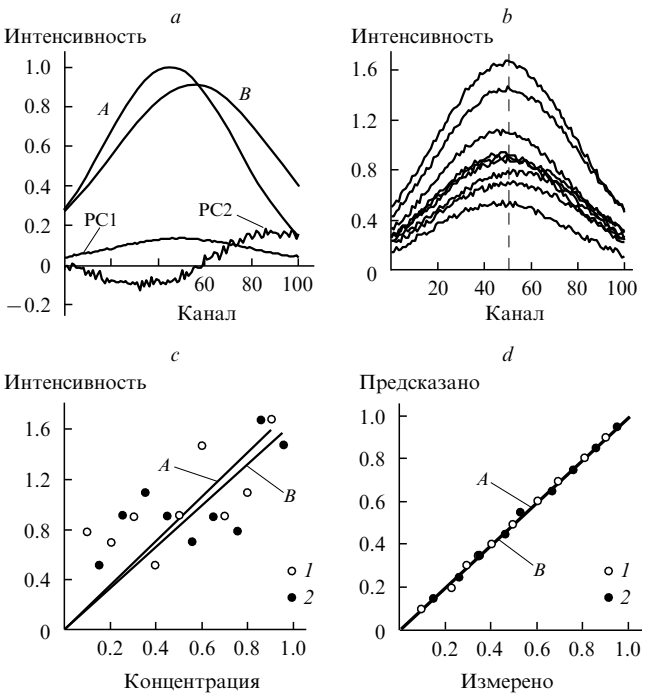

Рис. 10. Примеры построения различных градуировок.
a — спектры чистых (кривые *A*, *B*) и главных компонент (кривые PC1, PC2); *b* — модельные спектральные данные; *c* — одномерная градуировка, $R_A^2 = 0.50$, RMSEC = 0.26, $R_B^2 = 0.46$, RMSEC = 0.219; *d* — градуировка методом PCR, $R_A^2 = 0.999$, RMSEC = 0.011, $R_B^2 = 0.998$, RMSEC = 0.012.

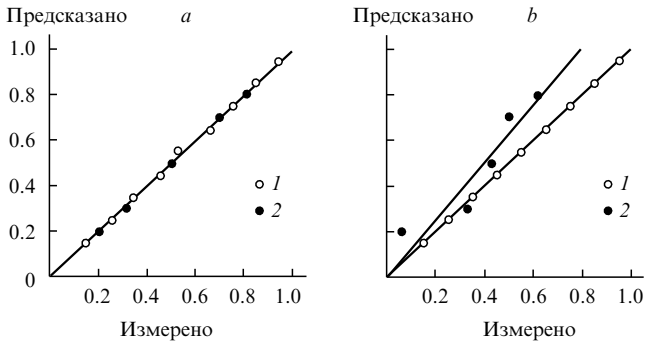

Рис. 11. Проверка градуировок в модельном примере с использованием метода главных компонент (*a*) и множественной регрессии (*b*).
a: $R_t^2 = 0.999$, RMSEC = 0.008, $R_c^2 = 0.998$, RMSEC = 0.12;
b: $R_t^2 = 0.86$, RMSEC = 0.14, $R_c^2 = 1.0$, RMSEC = 0. *1* — обучающий набор, *2* — проверочный набор.

боды F равно числу таких образцов. Видно, что метод главных компонент выдерживает проверку: градуировочная и проверочная линии сливаются.

Рассмотрим в этом контексте метод градуировки с помощью множественной регрессии. Обучающий набор состоит из девяти образцов, поэтому для построения модели мы можем использовать не более восьми каналов ($I > J$), например первый, четырнадцатый, двадцать седьмой и т.д. Результаты градуировки и проверки для метода множественной регрессии представлены на рис. 11,*b*. Поскольку число образцов всего на единицу больше числа каналов, то градуировочная прямая точно проходит через все точки обучающего набора (точки *1*), поэтому RMSEC = 0, и $R_c^2 = 1$. Однако проверка показала неудовлетворительное качество такой градуировки: точность на порядок ниже точности в методе PCR, а проверочная прямая не совпадает с градуировочной. Это — типичный пример переоценки модели:[71] точность описания обучающих данных значительно выше точности прогнозирования.

Проблема сбалансированности описания данных рассмотрена во многих работах А.Хоскюлдссона, который ввел новую концепцию моделирования — так называемый H-принцип.[159] Согласно этому принципу, точность моделирования, оцениваемая параметром RMSEC, и точность прогнозирования, оцениваемая параметром RMSEP, связаны между собой. Улучшение RMSEC влечет ухудшение RMSEP, поэтому их нужно рассматривать совместно. Именно по этой причине множественная линейная регрессия, в которой всегда участвует явно избыточное число параметров, приводит к неустойчивым моделям, непригодным для практического применения.

В настоящее время наиболее распространенным методом многомерной градуировки в хемометрике является метод PLS. Он во многом похож на метод PCR, с тем существенным отличием, что в PLS проводится одновременная декомпозиция матриц **X** и **Y**

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E}, \qquad (7)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{t}} + \mathbf{F}.$$

Проекции строят согласованно — так, чтобы максимизировать корреляцию между соответствующими векторами X-счетов $\mathbf{t}_a$ и Y-счетов $\mathbf{u}_a$. Поэтому регрессия в методе PLS гораздо лучше описывает сложные связи, при этом используется меньшее число главных компонент. Детально метод PLS рассмотрен в книге [74]. Этот подход послужил основой для очень многих методов градуировки, используемых в хемометрике, таких как SIMPLS,[160] PMN,[161] робастный PLS,[162] ридж-PLS [163] и других.[164]

Однако все эти методы дают предсказания в виде точечной оценки, тогда как на практике часто нужна интервальная оценка, учитывающая неопределенность прогноза. Построение доверительных интервалов традиционными статистическими методами невозможно из-за сложности задачи,[114] а использование имитационных методов [98] затруднительно из-за длительности расчетов.[101] Л.Канторович [165] предложил заменить минимизацию суммы квадратов отклонений на систему неравенств, которая решается с помощью методов линейного программирования. В таком случае результат прогноза сразу имеет вид интервала, поэтому данный метод был назван «простым интервальным оцениванием».[36, 56] С его помощью выполнено несколько работ в области аналитической химии.[166]

2. Многомодальная регрессия

Методы многомерной градуировки естественно обобщаются на случай, в котором блоки $\boldsymbol{X}$ и $\boldsymbol{Y}$ представлены N-модальными матрицами.[80] При этом регрессию можно построить различными способами. Используя методы PARAFAC и Tucker3, блок предикторов $\boldsymbol{X}$ представляют в виде произведения 2D-матриц нагрузок, с помощью которых оценивают параметры. Эти методы можно рассматривать как обобщение метода PCR для многомодальных данных. Обобщением метода PLS является Tri – PLS-декомпозиция 3D-матрицы $\boldsymbol{X}$, которую можно представить в виде [167]

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

Здесь ${}^{\mathrm{u}}\mathbf{X}$ — 2D-матрица (размерности $I \times KJ$), полученная при развертке 3D-матрицы $\boldsymbol{X}$ (размерности $I \times K \times J$) (см. рис. 7); **T** — 2D-матрица счетов (размерности $I \times A$); ${}^{\mathrm{u}}\mathbf{P}$ — 2D-матрица весов (размерности $A \times KJ$), которая в свою очередь является разверткой для 3D-матрицы $\boldsymbol{P}$, представляемой как тензорное произведение двух 2D-матриц

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

Декомпозицию блока $\boldsymbol{Y}$ проводят аналогично

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

Здесь, как и в обычном методе PLS, счета **T** выбирают так, 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 градуировке, довольно сложен. Однако в настоящее время существуют программные продукты,‡ позволяющие химикам легко справляться с математическими трудностями. В литературе можно найти многочисленные примеры использования мультимодальной градуировки в химическом анализе. Так, этот метод используют в спектрофотометрии для определения пестицидов,[168] в высокоэффективной жидкостной хроматографии с диодно-матричным детектированием для разрешения налагающихся пиков,[169] для определения следовых концентраций металлов [170] и др.

В статье [171] рассмотрено применение газовой хроматомасс-спектрометрии для определения следовых концентраций кленбутерола в биологических образцах. В последнее время этот метод широко применяют для анализа следовых количеств органических веществ. Однако сложность биологических объектов, а также низкий уровень содержания исследуемого вещества приводят к тому, что оценка предела обнаружения существенно зависит от способа математической обработки экспериментальных данных. В рассматриваемой работе были приготовлены 7 стандартных образцов с известными концентрациями кленбутерола. Масс-спектрометрическое детектирование осуществляли как в режиме полного сканирования (210 ионов), так и в режиме детектирования по восьми отдельным ионам. Полученные данные имеют трехмодальную структуру: первая мода — образцы, вторая — масс-спектры, третья — хроматограммы. В режиме полного сканирования получена 3D-матрица предикторов $\boldsymbol{X}$ размерности $7 \times 210 \times 37$, в режиме детектирования по отдельным ионам размерность этой матрицы равна $7 \times 8 \times 22$. Блок откликов представляет собой 1D-вектор **y**, включающий семь значений концентраций.

Для построения градуировок использовали различные трехмодальные алгоритмы: PARAFAC, PARAFAC2, Tucker3, а также Tri – PLS. Лучшим оказался метод Tri – PLS, так как дал наименьший предел обнаружения. Сопоставление результатов, полученных с применением этого метода и стандартной одномерной методики, показало значительное снижение предела обнаружения: в режиме полного сканиро-

‡ См., R.Bro, C.A.Andersson. *The N-Way Toolbox for MATLAB*. Version 2.02 (2003). Доступно на http://www.models.kvl.dk/source

вания с 283 до 20.91 мг·кг$^{-1}$, при сканировании по отдельным ионам с 73.95 до 26.32 мг·кг$^{-1}$. Для вычисления предела обнаружения использовали концепцию NAS.[124]

3. Нелинейная градуировка

Иногда, например в рассмотренных выше задачах титрования, построить линейную градуировку невозможно. Кроме того, линейный подход требует большого количества данных, которые не всегда доступны. В этом случае возможны два альтернативных подхода: в соответствии с первым используют множественную нелинейную регрессию, в соответствии со вторым — многомерную нелинейную градуировку. Ниже рассмотрены оба подхода.

Нелинейный регрессионный анализ[172] можно успешно применять для решения задач количественного анализа, если число переменных невелико. Кроме того, необходимо располагать содержательной моделью, связывающей блоки **X** и **Y**. По-видимому, круг таких задач не очень широк — в основном это кинетические (в том числе титриметрические) задачи.[95] Данный подход применяли, например, при анализе активности антиоксидантов,[36] для решения обратной кинетической задачи,[34,94] в уже упомянутом титровании.[173,174] В работе[57] проведен подробный анализ проблем, с которыми сталкивается исследователь, использующий этот подход.

Альтернативой классической регрессии является формальный подход, который не требует знания содержательной модели, но предполагает наличие большого числа данных.[78] Для учета нелинейных эффектов предложены такие методы, как INLR,[175] GIFI–PLS,[176] QPLS,[177] включающие разнообразные усовершенствования обычного метода PLS.[81] Помимо нелинейного PLS в хемометрике активно применяют метод ANN,[178,179] имитирующий распространение сигналов в коре головного мозга. Этот метод успешно используют для интерполяции функций. Примерно 10 лет назад метод нейронных сетей привлек к себе внимание химиков, которые начали применять его для классификации,[180] дискриминации[53] и градуировки.[181,182] Затем, однако, наметилось некоторое снижение интереса и использовать метод ANN в хемометрике стали заметно реже. Причина заключена все в той же упомянутой выше проблеме переоценки моделей. При использовании нейронных сетей очень трудно правильно оценить степень сложности модели, что приводит к неустойчивому и ненадежному прогнозу. Другим интересным методом нелинейного моделирования, имитирующим биологические процессы, является метод GA.[183,184] Этот метод и его разновидность — метод IA — полезны в тех случаях, когда задача химического анализа не поддается формализации в терминах обычных целевых функций, например при разрешении многокомпонентных перекрывающихся хроматограмм.[185] Примеры практического применения различных нелинейных подходов в хемилюминесцентном анализе рассмотрены в работе[186].

V. Подготовка данных и обработка сигналов

1. Подготовка данных

Важным условием правильного моделирования и, соответственно, успешного химического анализа является предварительная подготовка данных, которая включает различные преобразования исходных («сырых») экспериментальных значений. Простейшими преобразованиями являются центрирование и нормирование.[187] Центрирование — вычитание из исходной матрицы **X** некоторой матрицы **M**

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{M}.$$

Обычно центрирование проводят по столбцам: для каждого вектора $\mathbf{x}_j$ вычисляют средние значения

$$m_j = \frac{x_{1j} + \ldots + x_{Ij}}{I},$$

тогда

$$\mathbf{M} = (m_1\mathbf{l}, \ldots, m_J\mathbf{l}),$$

где **l** — вектор из единиц размерности I. Иногда центрирование проводят и по строкам: вычисляют средние значения по строкам, которые вычитают из соответствующих строк $\mathbf{x}_i^{\mathrm{t}}$. В случае мультимодальных данных центрирование можно проводить по каждой моде отдельно. Центрирование необходимо, если модель однородна, т.е. не содержит свободного члена — как в уравнениях (1) и (7). После такой операции химический ранг модели понижается на единицу и может повыситься точность описания. Центрирование можно рассматривать как проецирование на нулевую главную компоненту,[13] поэтому его всегда применяют в методах PCA и PLS. Однако центрирование не следует применять, если в данных есть пропуски.

Второе простейшее преобразование данных — нормирование. Это преобразование, в отличие от центрирования, не меняет структуры данных, а просто изменяет вес их различных частей при обработке. Нормирование можно проводить по каждой моде. Нормирование по столбцам — это умножение исходной матрицы **X** слева на матрицу **W**

$$\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X},$$

где **W** — диагональная матрица размерности $J \times J$. Обычно диагональные элементы w_{jj} равны обратным значениям стандартного отклонения

$$d_j = \sqrt{\sum_{i=1}^{I} (x_{ij} - m_j)^2 / I}$$

по столбцу $\mathbf{x}_j$. Нормирование по строкам (которое также называют нормализацией) — это умножение матрицы **X** справа на диагональную матрицу **W**

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}.$$

При этом размерность **W** равна $I \times I$, а ее элементы w_{ii} — обратные значения стандартных отклонений строк $\mathbf{x}_i^{\mathrm{t}}$. Комбинацию центрирования и нормирования по столбцам

$$\tilde{x}_{ij} = \frac{x_{ij} - m_j}{d_j}$$

называют автошкалированием.

Нормирование данных часто применяют с целью уравнять вклад в модель от различных переменных (например, в гибридном методе жидкостная хроматография–масс-спектрометрия), учесть гетероскедастические погрешности, или чтобы обрабатывать совместно разные блоки данных. Нормирование также можно рассматривать как метод, позволяющий стабилизировать вычислительные алгоритмы.[71] В то же время к этому преобразованию нужно относиться с большой осторожностью, так как оно может исказить результаты качественного анализа.[42]

Помимо линейных используют и нелинейные преобразования результатов эксперимента. Так, при анализе данных ИК-спектроскопии в ближней области часто применяют преобразование Кубелки–Мунка.[188] Цель этого и других трансформаций, например преобразования Бокса–Кокса,[33] — линеаризация модели. Часто простые операции с данными, такие как логарифмирование[56] или извлечение корня,[36] помогают существенно улучшить модель.

Исходные данные почти всегда содержат погрешности, как случайные, так и систематические. Чтобы уменьшить влияние случайного шума, применяют различные способы сглаживания данных, например скользящее среднее, метод Савицкого–Голея.[42,189] Гораздо сложнее нивелировать влияние систематических погрешностей — удалить систематический сдвиг в данных. Если этот сдвиг постоянен, то его убирают, используя центрирование. В случае линейных или квадратичных зависимостей от переменной (например, длины волны) помогает численное дифференцирование. В случае более сложных зависимостей используют специальные методы, два из которых мы рассмотрим.

Метод множественной коррекции сигнала, называемый также мультипликативной коррекцией рассеяния (MSC)[70] был первоначально разработан [190] для ИК-спектроскопии в ближней области и базировался на идеях, высказанных в работе [188]. Процедура MSC-преобразования проста. Сначала определяют «базовый спектр» как среднее по всем строкам матрицы **X**.

$$\mathbf{m}^{\mathrm{t}} = \frac{\mathbf{x}_1^{\mathrm{t}} + \ldots + \mathbf{x}_I^{\mathrm{t}}}{I}$$

Затем для каждой строки $\mathbf{x}_i^{\mathrm{t}}$ строят регрессию

$$\mathbf{x}_i^{\mathrm{t}} = a_i + b_i\mathbf{m}^{\mathrm{t}} + \mathbf{e}_i^{\mathrm{t}}$$

и определяют коэффициенты a_i и b_i. Преобразованные данные получают из уравнения

$$\tilde{\mathbf{x}}_i^{\mathrm{t}} = \frac{\mathbf{x}_i^{\mathrm{t}} - a_i}{b_i}.$$

Параметры множественной коррекции a_i и b_i можно определять не по всем переменным, а только по некоторому (подвижному) окну.[191]

Второй метод (точнее группа методов) OSC (см.[192]) отличается тем, что для преобразования матрицы предикторов **X** используют блок откликов **Y**. Этот метод применяют при подготовке данных для решения задач количественного анализа. Идея OSC состоит в том, чтобы удалить из блока **X** все систематические зависимости, не связанные с моделируемым откликом, т.е. часть матрицы **X**, которая ортогональна матрице **Y**. При этом должен увеличиться коэффициент корреляции R^2 и уменьшиться число PLS-компонент A, необходимых для моделирования данных. Существует много вариантов этого метода, первоначально предложенного в работе [193] и развитого в статьях [194,195]. Процедура в методе OSC, как и в методе PLS, осуществляется последовательно по шагам. На каждом шаге из матрицы **X** удаляют часть, связанную с одной OSC-компонентой. Для определения части матрицы

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2,$$

ортогональной **Y**, т.е.

$$\mathbf{Z} = \mathbf{Y}^{\mathrm{t}}\mathbf{X}_2 = 0,$$

применяют алгоритм, аналогичный PLS. Подробно метод OSC и коды MATLAB рассмотрены в работе [195].

Альтернативой методам коррекции сигнала MSC и OSC является подход, в котором качество модели улучшается за счет отбора переменных. Полезность отбора, т.е. исключение из исходного массива данных **X** некоторых столбцов $\mathbf{x}_j$, подтверждена как теоретическими исследованиями, так и результатами экспериментов. Такой подход используют и в качественном,[196] и в количественном анализе.[197] Для отбора переменных применяют различные методы: генетический алгоритм,[198] оптимизацию Парето,[199] «складного ножа».[117] Особо важен отбор переменных в тех случаях, когда аналитический сигнал непрерывно зависит от канала, например при анализе спектрометрических данных.[200] Здесь переменные отбирают блоками, как в методах, рассмотренных в статьях [195,201]. Помимо отбора переменных используется и отбор образцов, т.е. строк $\mathbf{x}_i^{\mathrm{t}}$ в матрице **X** (как и соответствующих им значений в матрице откликов **Y**). Отбор образцов также позволяет улучшить качество модели, но особенно он важен для обнаружения выбросов,[113] при переносе градуировок с одного прибора на другой.[116,202,203] Новый подход к классификации и отбору образцов изложен в работе [56].

2. Обработка сигналов

Обработка аналитических сигналов с помощью различных преобразований и фильтров играет важную роль в химическом анализе.[204,205] Так, преобразование Фурье по сути произвело революцию в ЯМР-, ИК- и рентгеновской спектроскопии за последние 20 лет. Исходные данные регистрируются уже не в форме привычных спектров, а в виде временных рядов, в которых вся спектроскопическая информация перемешана, и для восстановления спектров необходимо математическое преобразование. Одной из основных причин применения фурье-спектрометрии является увеличение отношения сигнал/шум, при этом появляется возможность провести эксперимент примерно в 100 раз быстрее, чем при использовании обычного спектрометра. Например, это позволило сделать спектроскопию ЯМР на ядрах ^{13}C обычным аналитическим методом, несмотря на низкую чувствительность ядер ^{13}C к внешнему магнитному полю. Используя методы импульсной спектроскопии ЯМР, можно накапливать сигналы при большом числе импульсов и суммировать их. Одновременно с фурье-спектроскопией появилось множество методов, улучшающих качество полученных данных. Такие методы включают так называемую фурье-деконволюцию (развертку, разделение сигналов), различные манипуляции с исходными данными во временном домене и последующее применение преобразования Фурье.

Другим современным методом обработки сигналов является вэйвлет-анализ.[206] С его помощью можно кодировать, сжимать и моделировать большие массивы данных, содержащих тысячи переменных. Такой анализ является естественным продолжением и развитием методов Фурье. Недостаток разложения Фурье заключается в том, что его базисные функции непрерывно зависят от времени, поэтому они не пригодны для представления данных, зависящих от времени. В вэйвлет-анализе используют базисные функции с ограниченным диапазоном изменения аргумента, которые удовлетворяют специальным требованиям шкалирования диапазона. Эти функции сдвигаются вдоль оси сигнала и получаемые в результате сверки спектры дают частотно-временное представление с разным разрешением, зависящим от ширины диапазона. Если вэйвлет-анализ предшествует методам PCA или PLS, то эти методы можно применять к анализу очень больших массивов данных без потери информации.[207] Вэйвлет-анализ часто используют для сжатия и сглаживания одно- и двухмодальных ИК-спектров и спектров ЯМР.[208]

Иногда требуются методы, позволяющие быстро сглаживать сигналы в режиме реального времени. Одним из таких методов является фильтр Калмана. С его помощью можно моделировать, например, изменения кинетики в ходе процесса. Общая идея фильтра Калмана состоит в том, чтобы уточнять модель по мере развития процесса. Как только новые данные становятся доступны, модель дополняется и совершенствуется. С появлением быстродействующих и мощных компьютеров нужда в фильтре Калмана практически отпала, хотя отдельные работы, в которых его применяют, еще встречаются (см., например,[209]).

VI. Заключение

1. Аналитический контроль процессов

Мы рассмотрели основные достижения хемометрики за последние 15–20 лет и их применение в аналитической химии. Вне рамок обзора остались очень многие актуальные направления и приложения, как близкие, так и далекие от аналитической химии. Одно направление заслуживает особого рассмотрения, поскольку в нем наиболее ярко проявились тенденции и перспективы развития хемометрики и аналитической химии. Речь идет о методах аналитического контроля процессов.

В 1930-х годах американский статистик У.Шухарт предложил использовать статистические методы для контроля технологических процессов.[210] Его идея очень проста: если собрать и статистически обработать данные о нормально функционирующем производстве за длительное время, то для контролируемого технологического показателя x можно установить пределы $[x_{\min}, x_{\max}]$, внутри которых процесс развивается нормально. Выход показателя x за эти пределы сигнализирует о каких-то нарушениях, требующих немедленного вмешательства. Такой метод был назван статистическим контролем процессов. Его стали успешно применять на практике (карты Шухарта). Однако со временем, по мере усложнения технологий, оказалось, что нельзя контролировать каждый показатель x_i отдельно, независимо от других показателей. Это часто приводило к ошибочным решениям — ложным тревогам, браку и т.п. Дело в том, что измеряемые технологические показатели $x_1, x_2, \ldots$, как правило, связаны (коррелированы) между собой и их необходимо рассматривать совместно. Ситуация во многом напоминает анализ многоканальных химических данных (например, спектров). С учетом такой аналогии Дж.МакГрегор[211] разработал новый подход к решению этой задачи — MSPC.[212] Он предложил использовать метод PCA для анализа многомерных данных и строить контрольные пределы в пространстве счетов с помощью расстояния Махаланобиса. Идея оказалась чрезвычайно плодотворной и нашла многих сторонников. Усилиями специалистов в области хемометрики были разработаны специальные методы для многомерного контроля периодических (например, биохимических) процессов, основанные на трехмодальной градуировке,[24] для анализа сложных процессов были созданы иерархические,[213] блочные[214] и маршрутные методы.[82] Появились работы, авторы которых предлагали не только контролировать, но и оптимизировать процессы.[215] Эти теоретические разработки начали применять на практике, прежде всего в пищевой[216] и фармацевтической[217] отраслях промышленности. Были созданы системы контроля качества производства полимеров,[218] цветных металлов,[219] полупроводников.[41]

Таким образом, в хемометрике возникло новое направление[220] для решения задач, далеких от задач от аналитической химии. Однако оказалось, что традиционно используемые в разных отраслях промышленности датчики и сенсоры не могут дать информации, необходимой для контроля сложных (прежде всего фармацевтических) процессов. Возникла острая нужда в методах мониторинга химических реакций в режиме реального времени и даже *in situ*.[221] Для этой цели подошли традиционные аналитические методы, прежде всего УФ- (см.[94]) и ИК- (см.[222]) спектрометрические. Для контроля химических реакций и процессов оказались востребованы хемометрические методы разрешения кривых и оценки кинетических констант по спектроскопическим[223] и хроматографическим[224] данным. Сочетание MSPC с аналитическими методами мониторинга, а также с методами контроля качества химического анализа[225] дало толчок развитию нового направления в аналитической химии — аналитическому контролю процессов.[226] Методы PAT включают планирование, анализ и контроль критических переменных, характеризующих состояние производственных материалов и процессов в режиме реального времени (т.е. в процессе производства).

Главная проблема, с которой сталкивается современная промышленность, — обеспечение высокого качества конечного продукта. С учетом усиливающейся глобальной конкуренции и быстро меняющихся запросов рынка становится понятной необходимость введения на производстве эффективного контроля процессов в режиме реального времени, и хемометрика играет определяющую роль в решении этой задачи. Решение Федеральной комиссии США по контролю за лекарствами (FDA),[227] вышедшее в сентябре 2004 г., законодательно подтвердило эту роль.

В рамках хемометрики созданы замечательные методы и алгоритмы, однако они очень медленно и неохотно признавались регулирующими органами во всех развитых странах. Объяснить это можно тем, что идеи многомерного подхода труднее для восприятия и визуализации, чем традиционных одномерных методов, которые не всегда могут дать полную картину происходящего. Например, гораздо проще заявить, что качество продукта определяется высотой некоторого пика для определенной длины волны, чем объяснить, что это качество связано с тем, попадает ли проекция всего спектра в определенную область в пространстве PLS-счетов, найденную с помощью расстояния Махаланобиса и т.п.

За всю историю развития хемометрики принят лишь один нормативный документ в этой области.[228] С принятием документа о PAT хемометрика неизбежно станет легитимным инструментом для всех компаний, желающих следовать руководящим принципам FDA. По нашему мнению, принятие такого нормативного документа ознаменует кардинальный поворот в технологии, новую парадигму производства, кредо которой — сделать качество неотъемлемым атрибутом продукта. Принципиальное отличие этой парадигмы от существующей — отказ от принципа стандартизации и унификации в пользу гибкого оперативного управления будущим качеством продукта на всех стадиях производства, начиная от анализа сырья. Простейший пример: по качеству продукция предприятий общественного питания, работающих в рамках парадигмы «стандартизации», безусловно, хуже домашней пищи, производимой в условиях гибкого оперативного контроля. Однако внедрение системы PAT позволит и массовым производителям обеспечить столь же высокое качество при сохранении больших объемов выпуска продукции.

2. Перспективы развития

Какова же роль аналитической химии в становлении этой новой технологической парадигмы? Как и чем могут химики ответить на этот вопрос? Нам представляется, что в развитии аналитики будут преобладать следующие тенденции. Во-первых, объекты анализа станут более сложными и комплексными. Технологические потребности будут ставить перед исследователями не частные вопросы — сколько вещества X в пробе, — а общие — получится ли продукт нужного качества из используемого сырья или правильно ли развивается химическая реакция в данной колонке? Во-вторых, методы анализа будут изменены таким образом, чтобы обеспечить получение необходимых данных не в лаборатории (*at line*), а непосредственно на производстве, в режиме реального времени (*in line*). В-третьих, резко увеличится объем многомодальных и многомерных данных. Усилится роль гибридных и композиционных методов анализа. В-четвертых, искомая химическая информация будет «очень глубоко

спрятана» в этих данных, и более того, она будет все менее формализована, что потребует применения тонких методов ее извлечения. В-пятых, изменится организация аналитического эксперимента — вместо исследования одной пробы в одном опыте необходимо будет использовать системный подход, в соответствии с которым множество разных проб автоматически испытывают одновременно различными методами в разных условиях. Такой массовый компьютеризованный эксперимент (который можно наблюдать в настоящее время, например, в технологии микрочипов) станет рутинной аналитической практикой. В-шестых, акцент в аналитическом исследовании будет перенесен на биологические объекты и биохимические процессы, а также на исследование технологических процессов в целом.

Все эти тенденции уже сейчас прослеживаются в аналитической химии. Роль химика-аналитика изменится: он неизбежно станет более аналитиком, чем химиком. Задачи, которые будет решать такой исследователь, сводятся к двум главным. Первая — как придумать эксперимент, чтобы получить данные, из которых, в принципе, можно извлечь нужную информацию. При этом искомой информацией может быть не количественный (концентрация) или качественный (да/нет) результат, а прогноз конечного состояния исследуемой системы. Вторая — как извлечь нужную информацию из данных, интерпретировать ее в категориях полезности и качества. Для решения таких задач исследователь должен использовать опыт и инструментарий хемометрики. Таким образом, хемометрика как неотъемлемая часть аналитической химии в значительной мере определяет направления ее развития.§

Важно подчеркнуть, что существенное расширение сферы применения аналитических методов должно базироваться на тесном сотрудничестве специалистов в области хемометрики не только с химиками, но с другими учеными, и в первую очередь с математиками, физиками, биологами.

Авторы благодарят О.Н.Карпухина (ИХФ РАН, Москва) и А.Ю.Богомолова (EMBL, Гамбург) за ценные советы при подготовке данной статьи, а также К.Эсбенсена (Университет Ольбург, Дания) за предпринятые им усилия в деле популяризации хемометрики в России.

Литература

1. М.А.Шараф, Д.Л.Иллмэн, Б.Р.Ковальски. *Хемометрика*. Мир, Москва, 1987
2. Ю.А.Золотов. *Аналитическая химия: проблемы и достижения*. Наука, Москва, 1992
3. Ю.В.Грановский. *Вестн. МГУ. Сер. 2. Химия*, **38**, 211 (1997)
4. Ю.А.Карпов, Т.М.Полховская. *Стандартизация и метрология в металлургическом производстве*. Изд-во МИСИС, Москва, 1989
5. P.Geladi, K.Esbensen. *Chemom. Intell. Lab. Syst.*, **7**, 197 (1990)
6. D.L.Massart. *Chemometrics: a Textbook*. Elsevier, New York, 1988
7. S.Wold. *Chemom. Intell. Lab. Syst.*, **30**, 109 (1995)
8. M.Blanco, I.Villarroya. *Trends Anal. Chem.*, **21**, 240 (2002)
9. B.G.Osborne, T.Fearn. *Near Infrared Spectroscopy in Food Analysis*. Longman Scientific and Technical, Harlow, Essex, 1986
10. M.Blanco, J.Coello, H.Iturriaga, S.Maspoch, E.Rovira. *J. Pharm. Biomed. Anal.*, **16**, 255 (1997)
11. A.Espinosa, D.Lambert, M.Valleur. *Hydrocarbon Process.*, **74**, 86 (1995)
12. T.Næs, C.Irgens, H.Martens. *Appl. Stat.*, **35**, 195 (1986)
13. H.Martens, T.Næs. *Trends Anal. Chem.*, **3**, 204 (1984)
14. W.S.Gosset («Student»). *Biometrika*, **6**, 1 (1908)
15. K.Pearson. *Philippine Mag.*, **2** (6), 559 (1901)
16. R.A.Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925
17. R.A.Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935
18. В.Налимов. *Применение математической статистики при анализе вещества*. Физматлит, Москва, 1960
19. S.Wold, K.Esbensen, P.Geladi. *Chemom. Intell. Lab. Syst.*, **2**, 37 (1987)
20. G.Golub, C.Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, 1996
21. P.Geladi, H.Grahn. *Multivariate Image Analysis*. Wiley, Chichester 1996
22. B.Walczak, D.L.Massart. *Trends Anal. Chem.*, **16**, 451 (1997)
23. A.I.Belousov, S.A.Verzakov, J.von Frese. *J. Chemom.*, **16**, 482 (2002)
24. P.Nomikos, J.F.MacGregor. *AIChE J.*, **40**, 1361 (1994)
25. P.Geladi, K.Esbensen. *J. Chemom.*, **5**, 97 (1991)
26. M.Schaeferling, S.Schiller, H.Paul, M.Kruschina, P.Pavlickova, M.Meerkamp, C.Giammasi, D.Kambhampati. *Electrophoresis*, **23**, 3097 (2002)
27. M.M.C.Ferreira. *J. Chemom.*, **18**, 385 (2004)
28. I.E.Frank, J.H.Friedman. *Technometrics*, **35**, 109 (1993)
29. S.Wold, A.Berglund, N.Kettaneh. *J. Chemom.*, **16**, 377 (2002)
30. G.Molenberghs. *Biometrics*, **61**, 1 (2005)
31. А.Г.Шмелев. *Вопросы психологии*, (5), 34 (1982)
32. H.Wold. In *Perspectives in Probability and Statistics*. Applied Probability Trust; University of Sheffield, Sheffield, 1975. P. 117
33. Н.Дрейпер, Г.Смит. *Прикладной регрессионный анализ. Т. 1, 2*. Финансы и статистика, Москва, 1987
34. О.Е.Родионова, А.Л.Померанцев. *Кинетика и катализ*, **45**, 485 (2004)
35. H.-L.Koh, W.-P.Yau, P.-S.Ong, A.Hegde. *Drug Discov. Today*, **8**, 889 (2003)
36. A.L.Pomerantsev, O.Ye.Rodionova. *Chemom. Intell. Lab. Syst.*, **79**, 73 (2005)
37. Л.Грибов. *Математические методы и ЭВМ в аналитической химии*. Наука, Москва, 1989
38. K.J.Siebert. *J. Am. Soc. Brew. Chem.*, **59**, 147 (2001)
39. K.Varmuza, W.Werther, F.R.Krueger, J.Kissel, E.R.Schmid. *Int. J. Mass Spectrom.*, **189**, 79 (1999)
40. G.W.Johnson, R.Ehrlich. *Environ. Forensics*, **3**, 59 (2002)
41. B.M.Wise, N.B.Gallagher, E.B.Martin. *J. Chemom.*, **15**, 285 (2001)
42. R.G.Brereton. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Wiley, Chichester, 2003
43. Н.П.Комарь. *Основы качественного химического анализа*. Харьков, 1955
44. Л.А.Грибов, В.И.Баранов, М.Е.Эляшберг. *Безэталонный молекулярный спектральный анализ. Теоретические основы*. Едиториал УРСС, Москва, 2002
45. М.Е.Эляшберг. *Успехи химии*, **68**, 579 (1999)
46. Б.М.Марьянов, А.Г.Зарубин, С.В.Шумар. *Журн. аналит. химии*, **58**, 1126 (2003)
47. В.И.Вершинин, Б.Г.Дерендяев, К.С.Лебедев. *Методы компьютерной идентификация органических соединений*. Наука, Москва, 2002
48. I.G.Zenkevich, B.Kránicz. *Chemom. Intell. Lab. Syst.*, **67**, 51 (2003)
49. I.V.Pletnev, V.V.Zernov. *Anal. Chim. Acta*, **455**, 131 (2002)
50. Н.М.Гальберштам, И.И.Баскин, В.А.Палюлин, Н.С.Зефиров. *Успехи химии*, **72**, 706 (2003)
51. В.И.Дворкин. *Метрология и обеспечение качества количественного химического анализа*. Химия, Москва, 2001
52. Ю.Г.Власов, А.В.Легин, А.М.Рудницкая. *Успехи химии*, **75**, 141 (2006)
53. А.В.Калач, Я.И.Коренман, С.И.Нифталиев. *Искусственные нейронные сети — вчера, сегодня, завтра*. Гос. технол. акад., Воронеж, 2002

§ Проецируя сказанное на практическую подготовку специалистов по хемометрике, хотелось бы обратить внимание на целесообразность создания специальных вузовских программ по этой дисциплине. Очевидно, оправдана и постановка вопроса о соответствующих специальностях в магистерских специализациях, а также о создании кандидатских и докторских советов.

54. С.П.Казаков, А.А.Рябенко, В.Ф.Разумов. *Оптика и спектроскопия*, **86**, 537 (1999)
55. В.Ф.Разумов, М.В.Алфимов. *Журн. науч. и прикл. фото- и кинематографии*, **46**, 28 (2003)
56. O.Ye.Rodionova, K.H.Esbensen, A.L.Pomerantsev. *J. Chemom.*, **18**, 402 (2004)
57. E.V.Bystritskaya, A.L.Pomerantsev, O.Ye.Rodionova. *J. Chemom.*, **14**, 667 (2000)
58. A.Bogomolov, M.McBrien. *Anal. Chim. Acta*, **490**, 41 (2003)
59. Пат. 2004-0126892-A1 США (2004)
60. S.Kucheryavski, V.Polyakov, A.Govorov. In *Progress in Chemometrics Research*. (Ed. A.L.Pomerantsev). NovaScience Publishers, New York, 2005. P. 3
61. Н.М.Оскорбин, А.В.Максимов, С.И.Жилин. *Изв. АлтГУ*, (1), 35 (1998)
62. S.V.Romanenko A.G.Stromberg, E.V.Selivanova, E.S.Romanenko. *Chemom. Intell. Lab. Syst.*, **73**, 7 (2004)
63. И.Е.Васильева, А.М.Кузнецов, И.Л.Васильев, Е.В.Шабанова. *Журн. аналит. химии*, **52**, 1238 (1997)
64. D.L.Massart, B.G.Vandeginste, L.M.C.Buydens, S.De Jong, P.J.Lewi, J.Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics. Part A*. Elsevier, Amsterdam, 1997
65. B.G.Vandeginste, D.L.Massart, L.M.C.Buydens, S.De Jong, P.J.Lewi, J.Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics. Part B*. Elsevier, Amsterdam, 1998
66. T.Naes, T.Isaksson, T.Fearn, T.Davies. *Multivariate Calibration and Classification*. Wiley, Christerer, 2002
67. R.Kramer. *Chemometric Techniques for Quantitative Analysis*. Marcel Dekker, New York, 1998
68. K.R.Beebe, R.J.Pell, M.B.Seasholtz. *Chemometrics: a Practical Guide*. Willey, New York, 1998
69. E.R.Malinowski. *Factor Analysis in Chemistry*. (2nd Ed.). Wiley, New York, 1991
70. H.Martens, T.Næs. *Multivariate Calibration*. Wiley, New York, 1989
71. A.Höskuldsson. *Prediction Methods in Science and Technology. Vol. 1*. Thor Publishing, Copenhagen, 1996
72. *Аналитическая химия. Проблемы и подходы. Т. 1, 2*. (Под ред. Р.Кельнера, Ж.-М.Мерме, М.Отто, Г.М.Видмера). Мир АСТ, Москва, 2004
73. Б.М.Марьянов. *Избранные главы хемометрики*. Изд-во ТГУ, Томск, 2004
74. К.Эсбенсен. *Анализ многомерных данных*. ИПХФ РАН, Черноголовка, 2005
75. K.Esbensen, O.Rodionova, A.Pomerantsev, O.Startsev, S.Kucheryavskiy. *J. Chemom.* **17**, 422 (2003)
76. O.Ye.Rodionova. *Chemom. Intell. Lab. Syst.*, **67**, 194 (2003)
77. S.Kucheryavski, C.Marks, K.Varmuza. *Chemom. Intell. Lab. Syst.* **78**, 138, (2005)
78. L.Eriksson, E.Johansson, N.Kettaneh-Wold, S.Wold. *Multi- and Megavariate Data Analysis*. Umetrics, Umeå, 2001
79. E.Sanchez, B.R.Kowalski. *J. Chemom.*, **2** 247 (1988)
80. A.Smilde, R.Bro, P.Geladi. *Multi-way Analysis with Applications in the Chemical Sciences*. Wiley, Chichester, 2004
81. S.Wold, J.Trygg, A.Berglund, H.Antti. *Chemom. Intell. Lab. Syst.*, **58**, 131 (2001)
82. A.Höskuldsson. *J. Chemom.*, **58**, 287 (2001)
83. P.Geladi, J.Burger, T.Lestanderet. *Chemom. Intell. Lab. Syst.*, **72**, 209 (2004)
84. G.H.W.Sanders, A.Manz. *Trends Anal. Chem.*, **19**, 364 (2000)
85. G.E.P.Box, W.G.Hunter, J.S.Hunter. *Statistics for Experimenters*, Wiley, New York, 1978
86. Е.З.Демиденко. *Линейная и нелинейная регрессии*. Финансы и статистика, Москва, 1981
87. P.Jy. *Sampling for Analytical Purposes*. Wiley, Chichester, 1989
88. W.Kleingeld, J.Ferreira, S.Coward. *J. Chemom.*, **18**, 121 (2004)
89. Special Issue.Tutorials on sampling. *Chemom. Intell. Lab. Syst.*, **74**, 1 (2004)
90. B.Walczak, D.L.Massart. *Chemom. Intell. Lab. Syst.*, **58**, 15 (2001)
91. P.R.C.Nelson, P.A.Taylor, J.F.MacGregor. *Chemom. Intell. Lab. Syst.*, **35**, 45 (1996)
92. H.Haario, V.-M.Taavitsainen. *Chemom. Intell. Lab. Syst.*, **44**, 77 (1998)
93. Э.Ф.Брин, А.Л.Померанцев. *Хим. физика*, **5**, 1674 (1986)
94. S.P.Gurden, J.A.Westerhuis, S.Bijlsma, A.K.Smilde. *J. Chemom.*, **15**, 101 (2001)
95. А.Л.Померанцев. Дис. д-ра физ.-мат. наук. ИХФ РАН, Москва, 2003
96. D.A.Morales. *J. Chemom.*, **16**, 247 (2002)
97. A.de Juan, M.Maeder, M.Martinez, R.Tauler. *Chemom. Intell. Lab. Syst.* **54**, 123 (2000)
98. Б.Эфрон. В кн. *Нетрадиционные методы многомерного статистического анализа*. Финансы и статистика, Москва, 1988. С. 19
99. *EURACHEM/CITAC Guide, Quantifying Uncertainty in Analytical Measurement*. (2nd Ed.). EURACHEM, Lisbon, Portugal, 2000
100. K.Faber, B.R.Kowalski. *Chemom. Intell. Lab. Syst.*, **34**, 283 (1996)
101. A.L.Pomerantsev. *Chemom. Intell. Lab. Syst.*, **49**, 41 (1999)
102. A.Pulido, I.Ruisánchez, R.Boqué, F.X.Rius. *Trends Anal. Chem.*, **22**, 647 (2003)
103. V.I.Vershinin. *Accredit. Qual. Assur.*, **9**, 415 (2004)
104. N.M.Faber. *Chemom. Intell. Lab. Syst.*, **64**, 169 (2002)
105. N.M.Faber, R.Bro. *Chemom. Intell. Lab. Syst.*, **61**, 133 (2002)
106. A.Lorber. *Anal. Chem.*, **58**, 1167 (1986)
107. J.Ferré, N.M.Faber. *Chemom. Intell. Lab. Syst.*, **69**, 123 (2003)
108. R.Boqué, N.M.Faber, F.Xavier Rius. *Anal. Chim. Acta*, **423**, 41 (2000)
109. R.Boqué; J.Ferré, N.M.Faber, F.Xavier Rius. *Anal. Chim. Acta*, **451**, 313 (2002)
110. I.Berget, T.Næs. *J. Chemom.*, **18**, 103 (2004)
111. D.Jouan-Rimbaud, D.L.Massart, C.A.Saby, C.Puel. *Anal. Chim. Acta*, **350**, 149 (1997)
112. M.Meloun, J.Militký, M.Hill, R.G.Brereton. *Analyst*, **127**, 433 (2002)
113. J.A.Fernandez Pierna, F.Wahl, O.E.de Noord, D.L.Massart. *Chemom. Intell. Lab. Syst.*, **63**, 27 (2002)
114. K.Faber. *Chemom. Intell. Lab. Syst.*, **52**, 123 (2000)
115. N.M.Faber, X.-H.Song, P.K.Hopke. *Trends Anal. Chem.*, **22**, 330 (2003)
116. E.Bouveresse, D.L.Massart. *Vib. Spectrosc.*, **11**, 3 (1996)
117. F.Westad, H.Martens. *J. Near Infrared Spectrosc.*, **8**, 117 (2000)
118. M.Hubert, S.Verboven. *J. Chemom.*, **17**, 438 (2003)
119. H.R.Keller, D.L.Massart. *Chemom. Intell. Lab. Syst.*, **12**, 209 (1992)
120. E.R.Malinowski. *J. Chemom.*, **6**, 29 (1992)
121. P.J.Gemperline. *Anal. Chem.*, **58**, 2656 (1986)
122. S.Wold. *Pattern Recogn.*, **8**, 127 (1976)
123. J.-H.Jiang, Y.Liang, Y.Ozaki. *Chemom. Intell. Lab. Syst.*, **71**, 1 (2004)
124. F.C.Sanchez, B.van de Borgaert, S.C.Rutan, D.L.Massart. *Chemom. Intell. Lab. Syst.*, **34**, 139 (1996)
125. H.Shen, B.Grande, O.M.Kvalheim, I.Eide. *Anal. Chim. Acta*, **446**, 313 (2001)
126. W.Windig, J.Guilment. *Anal. Chem.*, **63**, 1425 (1991)
127. A.Bogomolov, M.Hachey. In *Progress in Chemometrics Research*. (Ed. A.L.Pomerantsev). Nova Science Publishers, New York, 2005. P. 119
128. J.Diewok, A.de Juan, M.Marcel, R.Tauler, B.Lendl. *Anal. Chem.*, **76**, 641 (2003)
129. А.Ю.Богомолов, Т.Н.Ростовщикова, В.В.Смирнов. *Журн. физ. химии*, **69**, 1197 (1995)
130. H.A.Seipel, J.H.Kalivas. *J. Chemom.*, **18**, 306 (2004)
131. S.R.Crouch, A.Scheeline, E.S.Kirkor. *Anal. Chem.*, **72**, 53 (2000)
132. R.I.Shrager. *Chemom. Intell. Lab. Syst.*, **1**, 59 (1986)
133. R.De Maesschalck, D.Jouan-Rimbaud, D.L.Massart. *Chemom. Intell. Lab. Syst.*, **50**, 1 (2000)
134. J.M.Andrade, M.P.Gomez-Carracedo, W.Krzanowski, M.Kubista. *Chemom. Intell. Lab. Syst.*, **72**, 123 (2004)
135. O.Ye.Rodionova, L.P.Houmøller, A.L.Pomerantsev, P.Geladi, J.Burger, V.L.Dorofeyev, A.P.Arzamastsev. *Anal. Chim. Acta*, **549**, 151 (2005)
136. L.X.Sun, K.Danzer. *J. Chemom.*, **10**, 325 (1996)
137. A.J.Myles, S.D.Brown. *J. Chemom.*, **17**, 531 (2003)

138. D.González-Arjona, G.López-Pérez, A.G.González. *Talanta*, **49**, 189 (1999)
139. H.Mark. *Anal. Chem.*, **59**, 790 (1987)
140. P.J.Gemperline, N.R.Boyer. *Anal. Chem.*, **67**, 160 (1995)
141. H.L.Mark, D.Tunnell. *Anal. Chem.*, **57**, 1449 (1985)
142. U.Indahl, N.S.Sing, B.Kirkhuus, T.Naes. *Chemom. Intell. Lab. Syst.*, **49**, 19 (1999)
143. G.Downey, J.Boussion, D.Beauchene. *J. Near Infrared Spectrosc.*, **2**, 85 (1994)
144. G.R.Flaten, B.Grung, O.M.Kvalheim. *Chemom. Intell. Lab. Syst.*, **72**, 101 (2004)
145. T.Næs, U.Indahl. *J. Chemom.*, **12**, 205 (1998)
146. J.McElhinney, G.Downey, T.Fearn. *J. Near Infrared Spectrosc.*, **7**, 145 (1999)
147. S.Zomer, R.Brereton, J.F.Carter, C.Eckers. *Analyst*, **129**, 175 (2004)
148. V.V.Zernov, K.V.Balakin, A.A.Ivaschenko, N.P.Savchuk, I.V.Pletnev. *J. Chem. Inf. Comput. Sci.*, **43**, 2048 (2003)
149. M.Sarker, W.Rayens. *J. Chemom.*, **17**, 166 (2003)
150. A.Herrero, S.Zamponi, R.Marassi, P.Conti, M.C.Ortiz, L.A.Sarabia. *Chemom. Intell. Lab. Syst.*, **61**, 63 (2002)
151. R.Manne, B.-V.Grande. *Chemom. Intell. Lab. Syst.*, **50**, 35 (2000)
152. S.Bijlsma, A.K.Smilde. *J. Chemom.*, **14**, 541 (2000)
153. R.Bro. *Chemom. Intell. Lab. Syst.*, **38**, 149 (1997)
154. H.Kiers. *J. Chemom.*, **14**, 151 (2000)
155. N.M.Faber, R.Bro, P.K.Hopke. *Chemom. Intell. Lab. Syst.*, **65**, 119 (2003)
156. C.A.Andersson, R.Bro. *Chemom. Intell. Lab. Syst.*, **52**, 1 (2000)
157. F.J.del Rio, J.Riu, F.X.Rius. *J. Chemom.*, **15**, 773 (2001)
158. R.G.Brereton. *Analyst*, **125**, 2125 (2000)
159. A.Höskuldsson. *J. Chemom.*, **2**, 211 (1988)
160. S.de Jong. *Chemom. Intell. Lab. Syst.*, **18**, 251 (1993)
161. B.Li, A.J.Morris, E.B.Martin. *Chemom. Intell. Lab. Syst.*, **72**, 21 (2004)
162. M.Hubert, K.Vanden Branden. *J. Chemom.*, **17**, 537 (2003)
163. E.Vigneau, M.Devaux, M.Qannari, P.Robert. *J. Chemom.*, **11**, 239 (1997)
164. P.Geladi. *Chemom. Intell. Lab. Syst.*, **60**, 211 (2002)
165. Л.В.Канторович. *Сиб. мат. журн.*, **3**, 701 (1962)
166. В.М.Белов, В.А.Суханов, Ф.Г.Унгер. *Теоретические и прикладные аспекты метода центра неопределенности*. Наука, Новосибирск, 1995
167. R.Bro. *J. Chemom.*, **10**, 47 (1996)
168. Y.Ni, C.Huang, S.Kokot. *Chemom. Intell. Lab. Syst.*, **71**, 177 (2004)
169. Z.P.Chen, J.Morris, E.Martin, R.-Q.Yu, Y.-Z.Liang, F.Gong. *Chemom. Intell. Lab. Syst.*, **72**, 9 (2004)
170. F.M.Fernández, M.B.Tudino, O.E.Troccoli. *Anal. Chim. Acta*, **433**, 119 (2001)
171. I.Garcia, L.Sarabia, M.C.Ortiz, J.M.Aldama. *Anal. Chim. Acta*, **515**, 55 (2004)
172. Й.Бард. *Нелинейное оценивание параметров*. Статистика, Москва, 1979
173. D.M.Barry, L.Meites. *Anal. Chim. Acta*, **68**, 435 (1974)
174. Б.Марьянов. В кн. *Химики ТГУ на пороге третьего тысячелетия*. Изд-во ТГУ, Томск, 1998. С. 48
175. A.Berglund, S.Wold. *J. Chemom.*, **11**, 141 (1997)
176. A.Berglund, N.L.U.Kettaneh, S.Wold, N.Bendwell, D.R.Cameron. *J. Chemom.*, **15**, 321 (2001)
177. S.Wold. *Chemom. Intell. Lab. Syst.*, **14**, 71 (1992)
178. J.Zupan, J.Gasteiger. *Anal. Chim. Acta*, **248**, 1 (1991)
179. J.Zupan, J.Gasteiger. *Neural Network for Chemists. An Introduction*. VCH, Weinheim, 1993
180. W.Wu, B.Walczak, D.L.Massart, E.Heuerding, F.E.Erni, I.R.Last, K.A.Prebble. *Chemom. Intell. Lab. Syst.*, **33**, 35 (1996)
181. J.R.M.Smits, W.J.Melssen, L.M.C.Buydens, G.Kateman. *Chemom. Intell. Lab. Syst.*, **22**, 165 (1994)
182. W.J.Melssen, J.R.M.Smits, L.M.C.Buydens, G.Kateman. *Chemom. Intell. Lab. Syst.*, **23**, 267 (1994)
183. D.B.Hibbert. *Chemom. Intell. Lab. Syst.*, **19**, 277 (1993)
184. R.Leardi. *J. Chemom.*, **15**, 559 (2001)
185. X.Shao, Z.Chen, X.Lin. *Chemom. Intell. Lab. Syst.*, **50**, 91 (2000)
186. L.A.Tortajada-Genaro, P.Campíns-Falcó, J.Verdú-Andrés, F.Bosch-Reig. *Anal. Chim. Acta*, **450**, 155 (2001)
187. R.Bro, A.K.Smilde. *J. Chemom.*, **17**, 16 (2003)
188. P.Kubelka, F.Munck. *Z. Tech. Phys.*, **12**, 593 (1931)
189. A.Savitzky, M.J.E.Golay. *Anal. Chem.*, **36**, 1627 (1964)
190. P.Geladi, D.MacDougall, H.Martens. *Appl. Spectrosc.*, **3**, 491 (1985)
191. T.Isaksson, B.Kowalski. *Appl. Spectrosc.*, **47**, 702 (1993)
192. J.Trygg, S.Wold. *J. Chemom.*, **17**, 53 (2003)
193. S.Wold, H.Antti, F.Lindgren, J.Öhman. *Chemom. Intell. Lab. Syst.*, **44**, 175 (1998)
194. T.Fearn. *Chemom. Intell. Lab. Syst.*, **50**, 47 (2000)
195. A.Höskuldsson. *Chemom. Intell. Lab. Syst.*, **55**, 23 (2001)
196. Q.Guo, W.Wu, D.L.Massart, C.Boucon, S.de Jong. *Chemom. Intell. Lab. Syst.*, **61**, 123 (2002)
197. M.Forina, S.Lanteri, M.C.Cerrato Oliveros, C.Pizarro Millan. *Anal. Bioanal. Chem.*, **380**, 397 (2004)
198. R.Leardi, R.Boggia, M.Terrile. *J. Chemom.*, **6**, 267 (1992)
199. J.H.Kalivas. *Anal. Chim. Acta*, **505**, 9 (2004)
200. N.Benoudjit, E.Cools, M.Meurens, M.Verleysen. *Chemom. Intell. Lab. Syst.*, **70**, 47 (2004)
201. U.Indahl, T.Naes. *J. Chemom.*, **18**, 53 (2004)
202. R.N.Feudale, N.A.Woody, H.Tan, A.J.Myles, S.D.Brown, J.Ferré. *Chemom. Intell. Lab. Syst.*, **64**, 181 (2002)
203. E.L.Sulima, V.A.Zubkov, L.A.Rusinov. In *Progress in Chemometrics Research*. (Ed. A.L.Pomerantsev). Nova Science Publishers, New York, 2005. P. 196
204. J-H Jiang, Y.Ozaki, M.Kleimann, H.W.Siesler. *Chemom. Intell. Lab. Syst.*, **70**, 83 (2004)
205. P.W.Hansen. *J. Chemom.*, **15**, 123 (2001)
206. К.Чуи. *Введение в вэйвлеты*. Мир, Москва, 2001
207. J.Trygg, S.Wold. *Chemom. Intell. Lab. Syst.*, **42**, 209 (1998)
208. S.-P.Reinikainen. In *Progress in Chemometrics Research*. (Ed. A.L.Pomerantsev). Nova Science Publishers, New York, 2005. P. 21
209. Y.Pan, C.K.Yoo, J.H.Lee, I.-B.Lee. *J. Chemom.*, **18**, 69 (2004)
210. W.A.Shewhart. *Economic Control of Quality of Manufactured Product*. Van Nostrand, New York, 1931
211. J.MacGregor, Th.Kourti. *Control Engineering Practice*, **3**, 403 (1995)
212. А.Л.Померанцев, О.Е.Родионова. *Методы менеджмента качества*, **6**, 15 (2002)
213. Th.Kourti, J.MacGregor. *Chemom. Intell. Lab. Syst.*, **28**, 3, (1995)
214. J.A.Westerhuis, Th.Kourti, J.F.Macgregor. *J. Chemom.*, **12**, 301 (1998)
215. A.L.Pomerantsev, O.Ye.Rodionova. In *Progress in Chemometrics Research*. (Ed. A.L. Pomerantsev). Nova Science Publishers, New York, 2005. P. 209
216. R.Bro. *Chemom. Intell. Lab. Syst.*, **46**, 133 (1999)
217. J.Gabrielsson, N-O.Lindberg, T.Lundstedt. *J. Chemom.*, **16**, 141 (2002)
218. C.K.Yoo, J.-M.Lee, P.A.Vanrolleghem, I.-B.Lee. *Chemom. Intell. Lab. Syst.*, **71**, 151 (2004)
219. M.Baroni, P.Benedetti, S.Fraternale, F.Scialpi, P.Vix, S.Clementi. *J. Chemom.*, **17**, 9 (2003)
220. H.Martens, M.Martens. *Multivariate Analysis of Quality: an Introduction*. Wiley, Chichester, 2001
221. R.M.Dyson, M.Hazenkamp, K.Kaufmann, M.Maeder, M.Studer, A.Zilian. *J. Chemom.*, **14**, 737 (2000)
222. K.Pöllänen, A.Häkkinen, S.-P.Reinikainen, M.Louhi-Kultanen, L.Nyström. *Chemom. Intell. Lab. Syst.*, **76**, 25 (2005)
223. T.J.Thurston, R.G.Brereton, D.J.Foord, R.E.A.Escott. *Talanta*, **63**, 757 (2004)
224. E.Bezemer, S.C.Rutan. *Chemom. Intell. Lab. Syst.*, **59**, 19 (2001)
225. J.Workman Jr., K.E.Creasy, S.Doherty, L.Bond, M.Koch, A.Ullman, D.J.Veltkamp. *Anal. Chem.*, **73**, 2705 (2001)
226. S.P.Gurden, E.B.Martin, A.J.Morris. *Chemom. Intell. Lab. Syst.*, **44**, 319 (1998)

227. *Guidance for Industry PAT — a Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Veterinary Medicine (CVM), Office of Regulatory Affairs (ORA), September 2004, Pharmaceutical CGMPs

228. *ASTM Standard E1655*. Standard Practices for Infrared Multivariate Quantitative Analysis, 1997

CHEMOMETRICS: ACHIEVEMENTS AND PROSPECTS

O.Ye.Rodionova, A.L.Pomerantsev
N.N.Semenov Institute of Chemical Physics, Russian Academy of Sciences
4, Ul. Kosygina, 119991 Moskow, Russian Federation, Fax +7(495)939–7483

The key chemometric methods and models used to solve the problems of qualitative and quantitative analysis and for analytical control in chemical industry are considered. The achievements in the field of chemometrics made in the last 20 years are surveyed. The trends and prospects for its development are discussed.
Bibliography — 228 references.

Received 23rd August 2005