ХЕМОМЕТРИКА В АНАЛИТИЧЕСКОЙ ХИМИИ

О. Е. Родионова, А. Л. Померанцев

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

Рассмотрены основные хемометрические методы, используемые в аналитической химии. Проанализированы итоги развития хемометрики за последние 20 лет, обсуждены тенденции и перспективы ее роста. Дан обзор методов и моделей, применяемых для решения задач качественного и количественного анализов, а также для аналитического контроля процессов.

Библиография – 245 ссылок

ОГЛАВЛЕНИЕ

1. ВВЕДЕНИЕ

1.1. История хемометрики и ее место в системе знаний

Эта статья является первым большим обзором хемометрики в российской научной периодике. Со времени опубликования перевода единственной, до недавнего времени, книги по хемометрике [1] прошло более 15 лет, и многое существенно изменилось. Сейчас хемометрические методы используются в различных областях науки и техники. Однако этот обзор посвящен, в основном, аналитической химии, где можно выделить три основные направления применения хемометрики: качественный и количественный анализ, контроль химического анализа, и планирование эксперимента [2]. Главное внимание мы уделим первому направлению, меньшее второму, и практически ничего не будет сказано о третьем. Такая расстановка акцентов диктуется тем, что именно в таком порядке возрастает степень информированности российских аналитиков о хемометрических методах. Традиционно сильная школа отечественных статистиков, занимающихся планированием эксперимента [3], обеспечила большое количество публикаций по этой теме. Неплохо обстоит дело и с теми разделами хемометрики, которые связаны с метрологией [4]. В то же время за прошедшие пятнадцать лет число публикаций по хемометрике выросло от 100 до 5000 в год. Это делает задачу обзора всей этой науки практически невыполнимой и понуждает нас к разумным ограничениям в выборе представляемой области. Анализ химических данных – это самая важная часть хемометрики. В последнее время это направление очень быстро и плодотворно развивается, предлагая аналитикам не только новые методы обработки данных, но и новые подходы к постановке экспериментов.

Хемометрика, как самостоятельная поддисциплина внутри аналитической химии, появилась осенью 1974 года, в городе Сиэтле, США [5]. У ее истоков стояли два человека: американец Брюс Ковальски (B. Kowalski) и швед Сванте Волд (S. Wold) – внук Сванте Аррениуса. Хемометрика – это синтетическая дисциплина, находящаяся на стыке химии и математики. Как это часто бывает с подобными дисциплинами, хемометрика до сих пор не имеет общепризнанного определения. Наиболее популярное определение принадлежит Д. Массарту (D. Massart), [6] который считает, что *хемометрика – это химическая дисциплина, применяющая математические, статистические и другие методы, основанные на формальной логике, для построения или отбора оптимальных методов измерения и планов эксперимента, а также для извлечения наиболее важной информации при анализе экспериментальных данных*. С таким определением согласятся, наверное, многие практи-

ки. Однако область науки должна определятся не через методы и инструменты, которые она использует, а через цели и задачи, которые она преследует. Разумеется, задача извлечения информации из исходных данных по-прежнему остается крайне важной, как с практической, так и с теоретической точки зрения, однако, сейчас все лучше понятно, что не менее важной является и задача конструирования таких экспериментов, которые могут доставить данные, в которых содержится нужная информация. Эти два разнозначных аспекта – извлечение информации из данных и получение данных с нужной информацией – нашли свое 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, и как получить данные, содержащую такую информацию*.

То, что хемометрика родилась и начала бурно развиваться именно в начале 70-х годов, явно связано с появлением в то же время быстродействующей вычислительной техники, которая стала повсеместно доступна ученым и инженерам. Это позволило практически воплотить многие сложные алгоритмы обработки данных, в особенности методы анализа многоотказовых и многофакторных экспериментов. Это, в свою очередь, побудило производителей приборов разрабатывать более сложное оборудование, способное производить многократно большее количество измерений. Однако вскоре оказалось, что большее количество данных еще не означает большее количество информации, необходимой аналитикам. Это подвигло их активно применять хемометрические методы для извлечения такой информации и для подтверждения того, что сделанные при этом выводы достоверны. В результате такого взаимодействия был достигнут первый несомненный успех хемометрики. Оказалось, что очень часто традиционные аналитические методы, требующие больших затрат труда, времени, уникального оборудования, дорогих реактивов, могут быть заменены на косвенные методы, которые гораздо быстрее и дешевле. Наиболее ярко эта тенденция проявилась при использовании инфракрасной (ИК) спектроскопии, особенно в ближней области (БИК), прежде считавшейся малополезной 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 для них градуировочных моделей с помощью метода главных компонент [12] и метода проекций на латентные структуры [13].

Говоря об истории хемометрики нельзя не отметить ученых, которые еще задолго до 70-х заложили основы хемометрического подхода. Начать, очевидно, нужно с К. Гаусса

(K. Gauss), который в 1795 году ввел метод наименьших квадратов. Первым практикующим хемометриком следует, по-видимому, считать У. Госсета (W. Gosset), известного под псевдонимом Стьюдент, который в конце 19 века применял методы анализа данных [15] на пивоварне Гиннесса, где он работал аналитиком. В начале 20 века появилась работа К. Пирсона (K. Pearson) [14], в которой был предложен метод главных компонент, несколько позднее работы Р. Фишера (R. Fisher) – автора многочисленных статистических методов, таких как метод максимума правдоподобия и факторного анализа [16], а также пионерских работ [17] по планированию эксперимента. Среди советских ученых следует отметить, прежде всего, В. Налимова, внесшего значительный вклад в теорию планирования химического эксперимента [18].

Хемометрика зародилась, и длительное время развивалась внутри аналитической химии, и аналитики до сих пор остаются главными пользователями хемометрических методов. Однако со временем обнаружилась тенденция, которую некоторые исследователи расценили, как выход хемометрики из-под крыла аналитической химии и превращение ее в самостоятельную дисциплину. Два обстоятельства дали повод к такому выводу. Во-первых, это усложнение математического аппарата, используемого в хемометрике. Десять лет назад аналитики смогли усвоить и принять многомерный подход к анализу данных, т.е. такие методы как проекция на латентные структуры (PLS) [19] или разложение по сингулярным значениям (SVD) [20]. Однако потом, в период повального увлечения хемометриков новыми методами анализа данных: мультимодальным подходом (n-way) [21], вэйвлет-анализом (wavelet) [22], методом опорных векторов (SVM) [23] и т.п., наметился некоторый разрыв между практикующими аналитиками и хемометриками. Аналитики не понимали, что и зачем делают хемометрики, а те, в свою очередь, не понимали, почему их новые методы не востребованы в аналитической практике. Второе обстоятельство, приведшее к некоторому охлаждению этих «родственных отношений», связано с появлением многочисленных приложений, в которых хемометрический подход с успехом применялся в областях, далеких от аналитической химии. Достаточно вспомнить о многомерном статистическом контроле процессов (MSPC) [24], об анализе изображений (MIA) [25], или о биологических приложениях [26]. Кризис доверия привел к тому, что на последней конференции «Хемометрика в аналитической химии» (CAC-2004, Лиссабон [27]) вопрос о том, является ли хемометрика по-прежнему частью аналитической химии, обсуждался многими участниками.

Хемометрика тесно связана с математикой и, в особенности, с математической статистикой, откуда она черпает свои идеи. Большинство аналитиков понимают необходи-

мость применения статистики в химическом анализе и используют ее для вычисления средних, отклонений, пределов обнаружения, проверки гипотез и т.п. Часто именно эти простые приемы и называют хемометрическим подходом в аналитической химии, и лишь немногие исследователи решаются пойти дальше и действительно использовать хемометрику для анализа своих данных. Большинство аналитиков-практиков не любят математику, и сложные уравнения пугают их. Однако для эффективного практического применения хемометрики совсем не обязательно знать статистическую теорию метода главных компонент, достаточно понимать основы, базовые идеи этого подхода. А вот что действительно необходимо знать – это методы подготовки данных, принципы отбора переменных, и, самое главное, надо уметь правильно интерпретировать проекции данных (нагрузки и счета) в пространстве главных компонент. Хотя этот навык, как показывает многолетняя практика обучения хемометрике «без уравнений», можно приобрести и без глубоких математических познаний. Это обзор написан в рамках той же концепции, когда все основные принципы, методы и достижения хемометрики излагаются с минимально необходимым числом математических формул, и когда геометрическая интерпретация превалирует над алгебраической.

Взаимоотношения хемометрики и математики заслуживают отдельного рассмотрения. Многие методы и алгоритмы, популярные в хемометрике, не вызывают восторга у математиков [28], которые справедливо считают их плохо обоснованными с формальной точки зрения. Хемометрики всегда рассматривали свою деятельность как компромисс между возможностью и необходимостью, полагая, что главное – это практический результат, а не теоретическое обоснование невозможности его достижения. Сталкиваясь с практическими задачами интерпретации очень больших и сложно организованных массивов экспериментальных данных [29], хемометрики изобретают все новые и новые методы их анализа. Делают они это так быстро, что математики, по словам американского статистика Д. Фридмана (J. Friedman) [30], не успевают не только раскритиковать их за это, но и просто понять, что же происходит в этой хемометрике. Такой подход контрастирует с ситуацией, сложившейся в биометрике [31], которую можно считать, в каком-то смысле, старшей сестрой хемометрики. Со времен Фишера биометрики традиционно применяют только хорошо апробированные, классические методы математической статистики, такие как факторный анализ, или линейный дискриминационный анализ. С другой стороны, специалисты, работающие в другой близкой дисциплине – психометрике [32], традиционно активно разрабатывали новые подходы к анализу данных. Так, самый популярный в хемометрике метод PLS был изобретен Г. Волдом (H. Wold) [33] именно для применения в этой области. За-

бавно, что в начале 70-х годов господствовало мнение, что проекционные методы: «*малоприемлемы в физических, технических и биологических науках. Они могут быть полезны иногда в общественных науках как метод отыскания эффективных комбинаций переменных*» [34, т.2. стр.48].

Благодаря такому «агрессивному» подходу к анализу данных, хемометрика нашла многочисленные применения в самых разных – смежных и далеких от химии областях. Она применяется в физической химии для исследования кинетики [35], в органической химии для предсказания активности соединений по их структуре (QSAR) [36], в химии полимеров [37], в теоретической и квантовой химии [38]. Хемометрика используется в самых разнообразных областях – от пивоварения [39], до астрономии [40]. Она применяется для решения судебных споров о защите окружающей среды [41] и для контроля качества производства полупроводников [42]. Подробный анализ взаимодействий хемометрики с различными областями человеческой деятельности приведен в книге английского аналитика Р. Бреретона (R. Brereton) [43], к которой мы и отсылаем заинтересованного читателя.

Некоторые направления хемометрики развивались и в СССР, и позднее в России. Так, например, еще в 50-е годы в Харьковском университете под руководством Н. Комаря проводились исследования по математическому описанию равновесий [44]. Позднее появились работы Л. Грибова [45] и М. Эляшберга по спектральным методам [46], Б. Марьянова по титраметрии [47], Б. Дерендяева и В. Вершинина по методам компьютерной идентификации органических соединений [48], И. Зенкевича по хроматографии [49]. Хемометрический подход активно используется в работах [50], выполняемых в научной школе Ю. Золотова [2]. Исследования в близкой к хемометрике области QSAR ведутся под руководством Н. Зефирова [51]. Метрологические аспекты и контроль качества химического анализа исследуются в работах В. Дворкина [52]. В С.-Петербургском университете группа ученых под руководством Ю. Власова работает над созданием сенсорных систем, известных под названием «электронный язык» [53], а в Воронеже разрабатываются аналогичные методы, известные как «электронный нос» [54]. Во всех этих областях интенсивно используются хемометрические методы. В. Разумов и его коллеги из Черноголовки применяют многомерные методы анализа данных при решении задач химической кинетики [55, 56]. За последние годы в России появились новые группы ученых, разрабатывающих и применяющих хемометрические подходы: в Москве (О. Родионова [57], А. Померанцев [58], А. Богомолов [59, 60]), в Барнауле (С. Кучерявский [61], С. Жилин [62]), в Томске (С. Романенко [63]), в Иркутске (Е. Шабанова и И. Васильев [64]).

1.2. Информационное и программное обеспечение

Мы уже упоминали единственную широко-известную в России книгу по хемометрике [1]. Она ярко отражала положение дел в хемометрике, сложившееся в середине 80-х годов. На сегодняшний день наиболее полным изложением хемометрических методов является двухтомник, написанный группой авторов под руководством Д. Массарта [65, 66]. Он включает подробное описание основных хемометрических методов и приемов, большое количество практических приложений, а так же обширный список литературы. Помимо этого, существует множество книг и учебников, ориентированных на очень разный круг читателей. Так, для студентов и специалистов в области аналитической химии, начинающих осваивать хемометрику, проще начать с книги [43]; исследователям, занимающимся, в основном, спектральным анализом, будут понятнее книги [67, 68]. Для практического применения очень полезна книга [69]. Также нельзя не упомянуть знаменитую книгу Е. Малиновского (E. Malinowski) [70], которую до сих пор многие аналитики считают лучшим учебником в этой области. Теоретические основы хемометрики были изложены в работах [71, 72]. Недавно на русский язык был переведен учебник [73], содержащий краткое описание хемометрики. Небольшое, но очень полезное введение в хемометрику написал Б. Марьянов [74]. Маленьким тиражом (для участников двух конференций по хемометрики в России) был издан сокращенный перевод самого популярного в мире учебника по хемометрике, написанного К. Эсбенсеном (K. Esbensen) [75].

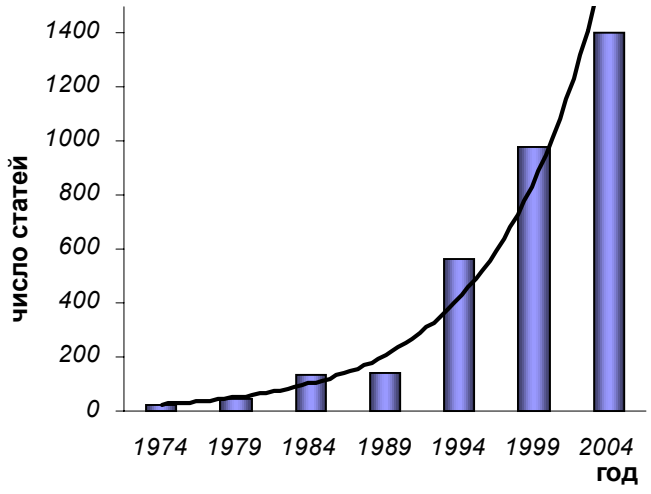


Рис. 1. Число статей по хемометрике, опубликованных в журналах издательства Elsevier

Проблемам хемометрики посвящены два специализированных журнала: *Journal of Chemometrics* и *Chemometrics and Intelligent Laboratory Systems*. Статьи, где хемометрические методы используются в прикладных задачах, регулярно печатаются более чем в 50-ти научных журналах, таких как Analytical Chemistry, Analytica Chimica Acta, Analyst, Talanta, Trends in Analytical Chemistry, Journal of Chromatography, Computers and Chemi-

cal Engineering, Vibrational Spectroscopy и т.д. Число статей, использующих хемометрические методы в качестве основного инструмента для анализа и обработки экспериментальных данных, возрастает с каждым годом (см. Рис. 1).

В мире проводятся как небольшие региональные конференции и семинары, так и регулярные международные конференции. Наиболее авторитетными являются конференция *Хемометрика в аналитической химии* (Chemometrics in Analytical Chemistry – CAC) [27] и *Скандинавский симпозиум по хемометрике* (Scandinavian Symposium on Chemometrics – SSC) [76]. В России, начиная с 2002 года, проходят ежегодные международные школы-симпозиумы [77-79] «Современные методы анализа многомерных данных». Теоретические и прикладные аспекты хемометрики широко представлены в виде интернет-ресурсов. По большей части это англоязычные страницы [80-83], однако есть и несколько российских ресурсов [84-85].

В качестве программного обеспечения в хемометрике применяются специализированные пакеты программ [86-88], позволяющие наглядно и быстро обрабатывать данные в интерактивном режиме. Однако широко применяются и статистические пакеты общего назначения [89, 90]. Часто исследователи пишут процедуры сами, например, в кодах MATLAB [91], и они публикуются для свободного применения, например [72].

1.3. Обозначения и термины

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

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

soft и *hard*, часто используемые в хемометрике для характеристики методов моделирования, должны, по нашему мнению, переводиться словами *формальный* и *содержательный*. Это точнее отражает суть этих понятий. 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 хемометрическим понятиям, таким как *scores* и *loadings*, используя вместо них сложные эвфемизмы. Мы полагаем, что в хемометрике невозможно обойтись без понятий *счета* и *нагрузки*, или их аналогов.

Хемометрика – это наука о сокращениях. В данном случае мы имеем в виду не понижение размерности данных, а то, что в хемометрике часто используются аббревиатуры: PCA, PLS, PCR, RMSEP и т. п. Несмотря на то, что у некоторых из них есть общепринятые русские аналоги, например PCA это МГК, PCR это РГК, в этом обзоре мы решили сохранить оригинальные английские аббревиатуры. Разумеется, все эти решения не бесспорны и мы были бы рады обсудить терминологические проблемы с заинтересованными коллегами. Мы приводим список сокращений, использованных в статье, в котором 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) – *дискриминационный 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 ana-

lytical 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) – дискриминационный 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

2. ДАННЫЕ И МОДЕЛИ, ИСПОЛЬЗУЕМЫЕ В ХИМИЧЕСКОМ АНАЛИЗЕ

2.1. Химические данные и информация

Экспериментальные данные – это основной объект, с которым работает хемометрика. Следуя классификации, предложенной в работе [93], рассмотрим типичное устройство химических данных. (см. Рис. 2)

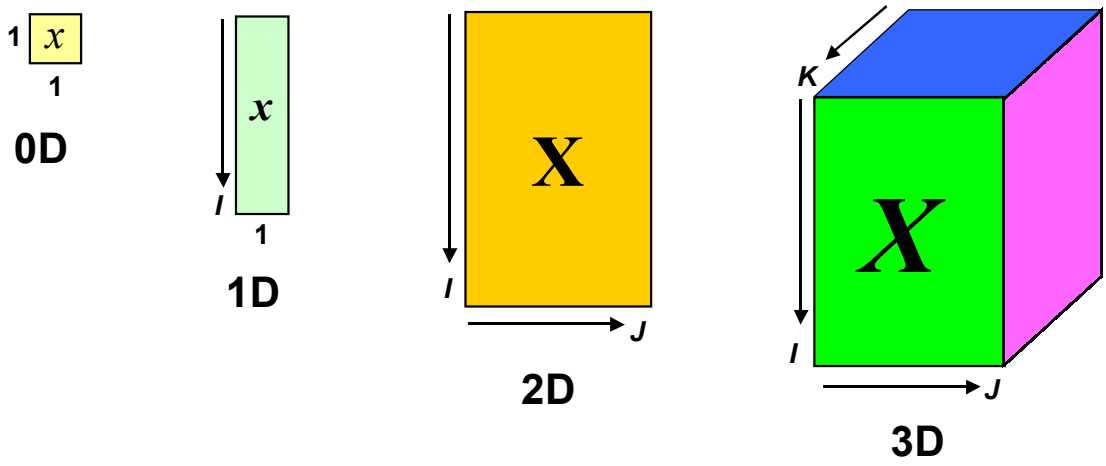


Рис. 2. Графическое представление данных разной модальности

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 – это многомерные, *одномодальные* данные, т.е. набор из 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, наиболее распространенный тип химических данных, это *двухмодальные* данные, которые представляются 2D-матрицей – таблицей из чисел, имеющей I строк и J столбцов. Типичный пример это набор спектров, снятых для I образцов на J длинах волн. Каждая строка в такой матрице представляет объект (в данном случае, образец), а каждый столбец – переменную (длину волны). Отнесение данных к объектам (образцам) или переменным (каналам) имеет большое значение для их интерпретации. Хотя не всегда такое разделение очевидно. Например, при анализе данных, полученных методом ВЭЖХ-ДДМ 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

удержания – объектами. В большинстве случаев, в эволюционных (то есть, развивающихся во времени) экспериментах, объекты соответствуют временам, то есть образец, меняющийся во времени, рассматривается как серия образцов [43].

С интенсивным развитием гибридных методов [2] большое внимание уделяется *трех (и более) модальным данным* [94]. Их можно представить в виде параллелепипеда (3D матрицы), в котором каждое ребро соответствует своему типу переменной. В разделе 4.2 настоящего обзора мы подробно рассматриваем типичный пример, в котором сочетание газовой хроматографии и масс-спектрометрии (ГХ-МС) применяется для количественного определения кленбутерола в биологических матрицах. В этом случае имеется одна мода, соответствующая стандартным образцам, и две моды, представляющие переменные: время удерживания и отношение массы к заряду. Пример четырех и даже восьми модальных данных можно найти в [94].

Данные могут объединяться в блоки. Простейший случай – это один блок **X**. Такой случай чаще встречается в качественном анализе, например, в задаче разделения спектров и концентраций. Количественный анализ, основанный на регрессионных зависимостях, использует данные, состоящие из двух и более блоков. Блок предикторов (например, 2D-матрица спектров **X**) и блок откликов (например, 1D- вектор концентраций **y**) составляют набор стандартных данных, по которым строится градуировочная модель **y**=**Xb**. Встречаются данные и более сложной структуры, объединяющей три и более блоков данных [95]. Для их анализа применяются специальные методы, называемые маршрутным моделированием (path modeling) [96]. Может показаться, что такая систематизация данных: размерность, модальность, блочность, носит несколько формальный характер и может представлять интерес только для математиков, но не для аналитиков. Дело обстоит не совсем так. За последние годы кардинально изменилась оценка того, какие данные можно считать большими. Если в начале 70-х матрица данных (например, спектров) считалась большой, когда в ней имелось 20 столбцов (переменных, например, длин волн) и 100 строк (объектов, например, образцов), то сейчас, с развитием техники эксперимента, большой может считаться матрица с 1 000 000 столбцов и 400 000 строками [29]. При обработке таких массивов, их приходится разделять на блоки и интерпретировать по очереди. Это разделение нельзя проводить формально, тут обязательно требуется участие опытного химика, понимающего суть дела. Модальность тоже придумали не математики. Это естественный ответ на потребность анализа данных гибридных и эволюционных экспериментов, число которых все увеличивается по мере развития инструментальной базы. С внедрением новых

аналитических методов, таких как гиперспектральные данные [97] и микрочипы [98], сложность данных будет только нарастать.

Основная задача хемометрики состоит в извлечении из данных нужной химической *информации*. Понятие информации является ключевым для хемометрики, поэтому мы дадим подробные комментарии к этой теме. Что является информацией, зависит от сути решаемой задачи. В некоторых случаях достаточно знать, что некоторое вещество присутствует в системе, но в других уже необходимо получить и количественные значения. Данные могут содержать нужную информацию, они даже могут быть избыточными, но в некоторых случаях информации в данных может не быть совсем. Во всех случаях данные содержат шум (например, погрешности), которые скрывают нужную информацию. Для иллюстрации рассмотрим следующий идеализированный эксперимент. Пусть есть система, состоящая из смеси трех веществ A, B и C без посторонних примесей. Предположим, что абсолютно точно известны спектры $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ всех компонентов. Заметим, что слово «спектры» употребляется в самом общем смысле. Это могут быть любые многомерные данные, например, хроматограммы, в которых λ – это время удержания. Требуется определить концентрации 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

$$x(\lambda_1)=c_A s_A(\lambda_1)+c_B s_C(\lambda_1)+c_C s_C(\lambda_1)$$

....

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

Совершенно ясно, что столько уравнений не нужно – можно оставить только три из них, соответствующие любым[*)] трем длинам волн и получить нужную информацию. Таким образом, видно, что исходные данные (30 мерный 1D-вектор) избыточны по отношению к искомой информации – используя любые три точки спектра, мы будем получать одни и те же значения концентраций. Сделаем теперь пример более реалистичным допуская, что все спектры содержат некоторую случайную погрешность. Тогда оценки концентраций, определяемые по разным тройкам длин волн, будут отличаться. Эти оценки можно усреднить и получить концентрации с лучшей точностью. Заметим, что того же можно было бы достичь и с помощью повторных экспериментов. Однако этот путь не эффективен, поскольку

[*)] Строго говоря – не совсем любым. Необходимо, чтобы система имела единственное решение.

требует больших затрат сил и времени. Гораздо проще уменьшать неопределенность количественного анализа за счет увеличения числа переменных (каналов, длин волн) в одном единственном эксперименте. Этот вывод является первым, важным принципом хемометрики – *использование многомерного подхода* при конструировании экспериментов и анализе их результатов.

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

Шум и избыточность в данных обязательно проявляют себя через *корреляционные связи* между переменными. Возвращаясь к идеализированному примеру, можно заметить, что в матрице «чистых» спектров, имеющей размерность 3×30, 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, это не случайность – ведь именно столько веществ присутствует в нашей системе. Это число называется рангом матрицы, и оно играет важную роль в хемометрическом анализе. Рассматривая тот же пример в более реалистичном варианте, с присутствием погрешностей, можно заметить появление дополнительных корреляций в данных. Это произойдет, например, если концентрация третьего вещества C будет существенно меньше шума. Тогда эти данные будут уже недостаточными 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 связями. Различие в понятиях причинности и корреляции забавно проиллюстрировано в книге [99], где приведен пример высокой положительной корреляции между числом жителей и количеством аистов в городе Ольденбург (Германия) за период 1930-36 годов. Разумеется, эти две переменные связаны между собой корреляционными связями, которые возникают из-за того, что в системе присутствует третья, скрытая переменная, с которой

они обе связаны причинными связями. Понятие эффективного (химического) ранга и скрытых, *латентных переменных*, число которых равно этому рангу, является третьим важнейшим принципом хемометрики [70]. Проиллюстрируем его применение примером, который будет некоторым усложнением уже рассмотренной ранее трехкомпонентной системы. Предположим, что имеется несколько (I) смесей веществ A, B и C, но их чистые спектры $s_A(\lambda)$, $s_B(\lambda)$, $s_C(\lambda)$ теперь не известны. Проведя анализ, можно получить спектры этих образцов – двухмодальные данные, т.е. матрицу **X** размером $I\times30$. Подвергнув матрицу **X** обычному математическому анализу, можно определить ее ранг. Это число дает важную информацию о том, сколько компонентов присутствует в системе или, по крайне мере, сколько их можно различить.

Таким образом, в химических данных почти всегда присутствуют внутренние, скрытые связи между переменными, приводящие к множественным корреляциям – коллинеарностям. Такое свойство данных называется *мультиколлинеарностью*. Оно может проявляться как избыточность данных, что позволяет улучшить качество оценок. С другой стороны, при неправильном методе обработки данных, мультколлинеарность может негативно сказаться на качестве анализа. Так, например, применение множественной линейной регрессии в условиях мультиколлинеарности совершенно неприемлемо [75]. Для регрессионного анализа таких данных надо применять специальные методы, например ридж-регресию [100], или проекционные подходы [72].

Существенным источником шума в данных может быть отбор образцов. Теория *пробоотбора*, значительный вклад в которую внес П. Жи (P. Gy) [101], приобрела большую популярность последнее время [102]. Многочисленные приложения можно найти в специальном выпуске [103], полностью посвященном этой теме. Другая проблема, с которой может столкнуться аналитик – это *пропуски в данных* [104]. Это может случиться по разным причинам: отказы приборов, выход за пределы обнаружения, нехватка образцов для исследования, и т.п. Большинство хемометрических методов не допускают пропусков в данных, поэтому для их заполнения используют специальные приемы, среди которых самым популярным является итерационный алгоритм. Каждая его итерация состоит из двух шагов. На первом шаге проводится оценка параметров модели так, как будто данные известны полностью. Для этого пропуски заполняются некоторыми априорно допустимыми значениями, например средними по окружающим элементам. На втором шаге, с помощью полученной модели, находятся наиболее вероятные значения пропущенных данных, и совершается следующая итерация. Для заполнения пропусков используется также подход, ос-

нованный на методе максимума правдоподобия [105]. Детали таких алгоритмов в большой степени зависят от того, какая модель используется для описания данных.

2.2. Модели и методы

Рассмотрев устройство данных, перейдем к методам их анализа. Далее основные хемометрические методы будут описаны более подробно, а этот раздел будет посвящен общей методологии. Хемометрические методы можно разделить на две группы, соответствующие двум главным задачам: *исследование* данных, например, классификация и дискриминация, и *предсказание* новых значений, например, градуировка. Методы первой группы оперируют, как правило, с одним блоком данных, а в градуировке необходимы, как минимум, два блока – предикторов и откликов. В зависимости от поставленных целей, методы решения могут быть направлены на предсказание внутри диапазона условий эксперимента (*интерполяция*) или за его пределами (*экстраполяция*). Существенным является разделение методов на *формальные* (soft), называемые также «черными», и *содержательные* (hard), или «белые». При использовании формальных моделей [106], данные описываются эмпирической зависимостью (как правило, линейной), справедливой в ограниченном диапазоне условий. В этом случае не нужно знать, как устроен механизм исследуемого процесса, однако такой метод не позволяет решать задачи экстраполяции. Параметры формальных моделей лишены физического смысла и должны интерпретироваться соответствующими математическими методами. Содержательное моделирование [107] 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 тогда, когда модель известна априори. Каждый из подходов имеет свои сильные и слабые стороны [37], и у каждого из них есть свои сторонники и противники. Исторически сложилось так, что в России интенсивно развивался содержательный подход, тогда как на западе отдавали преимущество формальным методам. За последнее время появилось много работ, в которых рассматриваются так называемые «серые» модели [108], объединяющие сильные стороны обоих методов. Проиллюстрируем разные подходы к моделированию примерами из аналитической химии.

Важными объектами математического моделирования в аналитической химии являются 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, т.е. для оценивания параметров по измеренным точкам кривой. Тем не менее, такую задачу можно все же решить в рамках «белого» моделирования, используя современные вычислительные системы [109]. С другой стороны, в работе [110] замечено, что по своей форме титриметрические кривые напоминают обратные гиперболические и тригонометрические функции. Исходя из этого, предлагается использовать формальные, «черные» зависимости, составленные из функций arcsinh, arccos и т.п. Компромиссный, «серый» подход предложен в работе [47], где заменой переменных содержательная модель преобразуется в кусочно-линейную. Затем для оценки параметров применяется метод *чередующихся наименьших квадратов* (ALS) [111], суть которого состоит в последовательном приближении модели к данным – сначала линейными регрессионными методами определяются оценки линейных параметров, при фиксированных значениях нелинейных, а затем нелинейные оцениваются в процедуре наискорейшего спуска, при найденных ранее фиксированных оценках линейных параметров. Процедура чередуется до сходимости.

Интерес к «черным» и «серым» методам моделирования обусловлен большими трудностями выбора и подтверждения правильности содержательной модели. Во многих случаях все сводится к простому перебору внутри короткого набора конкурирующих зависимостей, в результате которого обычно выбирается наипростейшая модель с минимальной невязкой. Однако это не доказывает правильность выбранного метода и может приводить к грубым ошибкам. Часто исследователи используют модели, которые О. Карпухин [112] справедливо назвал «розовыми» – это идеализированные зависимости, плохо соответствующие реальным артефактам, присутствующим в данных: дрейфам базовых линий, ненормальным погрешностям, и т.п. Формальные, многофакторные линейные модели и надлежащие методы их анализа гораздо лучше приспособлены к учету таких «неидеальностей». Они работают и в тех случаях, когда ни о какой содержательной, физико-химической модели не может быть и речи. Обоснованием для использования линейных моделей служит тот факт, что любую, даже очень сложную, но непрерывную зависимость можно представить как линейную функцию параметров в достаточно малой области. Принципиальным моментом здесь является то, какую область можно считать допустимой, иначе говоря, насколько широко можно применять построенную формальную модель. Ответ на этот вопрос дают *методы проверки* (валидации) моделей.

При надлежащем построении модели исходный массив данных состоит из двух независимо полученных наборов, каждый из которых является достаточно представитель-

ным. Первый набор, называемый обучающим, используется для идентификации модели, т.е. для оценки ее параметров. Второй набор, называемый проверочным, служит только для проверки модели. Построенная модель применяется к данным из проверочного набора, и полученные результаты сравниваются с проверочными данными. Таким образом принимается решение о правильности, точности моделирования методом *тест-валидации*. В некоторых случаях объем данных слишком мал для такой проверки. Тогда применяют другой метод – перекрестной проверки (*кросс-валидация*) [113]. В этом методе 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 возвращаются, и удаляется уже другая порция данных (следующие 10%), и опять строится модель, которая применяется к исключенным данным. Эта процедура повторяется до тех пор, пока все данные не побывают в числе исключенных (в нашем случае – 10 циклов). Наиболее (но неоправданно) популярен вариант перекрестной проверки, в котором данные исключаются по одному (LOO). В регрессионном анализе используется также проверка методом коррекции размахом, которая описана в [75]. Следует отметить, что та или иная проверочная процедура должна применяется не только в количественном, но и в качественном анализе при решении задач дискриминации и классификации.

Любой результат, полученный при анализе и моделировании экспериментальных данных, несет в себе *неопределенность*. Количественная оценка или качественное суждение могут измениться при повторном эксперименте в результате действия разнообразных случайных и систематических погрешностей, как присутствующих в исходных данных, так и вносимых на стадии моделирования [114]. Неопределенность в количественном анализе характеризуется либо числом – стандартным отклонением [115], либо интервалом – доверительным [116] или прогнозным [57]. В качественном анализе применяется метод проверки статистических гипотез [117], в котором неопределенность характеризуется через вероятность принятия неверного решения [118]. Методы оценки неопределенности при моделировании многомерных [119] и многомодальных [120] данных вызывают большой интерес хемометриков. Для описания различных аспектов надежности аналитического метода применяются специальные характеристики: специфичность, селективность, предел обнаружения, отношение сигнал/шум [74]. Актуальным методом их определения является подход с использованием концепции [121] *полезного аналитического сигнала* (NAS). Многомерный вектор NAS определяется как та часть полного сигнала (спектра), которая используется

для моделирования и прогноза [122]. Оставшаяся часть сигнала, включающая погрешности, вклады от посторонних компонентов, рассматривается как шум. Концепции NAS была применена к задаче определения предела обнаружения при анализе двух- [123] и трехмодальных [124] данных. Полученные результаты нашли многочисленные практические приложения, одно из которых рассмотрено в разделе 4.2.

Надежность аналитического метода сильно зависит от того, какие данные были использованы для построения и проверки соответствующей модели. Наличие выбросов [125] или малоинформативных данных снижает точность модели, и наоборот, присутствие представительных, влиятельных образцов в эксперименте [126] существенно улучшает качество модели. Оценка влиятельности данных может проводиться классическими регрессионными методами [127], а может выполняться с помощью нестатистических процедур [57]. При использовании построенной модели для определения интересующих нас показателей, мы сталкиваемся с похожими проблемами. Может оказаться, что метод не применим к некоторым образцам (выброс в прогнозе [128]) или дает очень неточный результат. Оценка неопределенности метода не в среднем [129], а для индивидуальных образцов – это сложная задача, над решением которой работают сейчас разные группы исследователей [130]. Именно их усилия определяют успешное решение таких практически важных задач как перенос градуировок с одного прибора на другой [131], отбор переменных [132], построение робастных [133] методов анализа данных.

3. МЕТОДЫ КАЧЕСТВЕННОГО АНАЛИЗА: ИССЛЕДОВАНИЕ, КЛАССИФИКАЦИЯ И ДИСКРИМИНАЦИЯ

3.1. Метод главных компонент

Современные приборы могут легко производить огромное количество измерений. 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 300×240, т.е. 72000 чисел. Однако, из-за мультиколлинеарности, доля полезной информации в таком массиве может быть относительно невелика. Для выделения полезной информации в хемометрике используются методы сжатия данных (в отличие от традиционного подхода, когда из данных выделялись только отдельные особо значимые измерения). Идея этих методов состоит в том, чтобы представить исходные данные, используя новые скрытые переменные. При этом должны выполнятся два условия. Во-первых, число новых переменных (химический ранг) должно быть существенно меньше, чем число исходных переменных, и, во-вторых, потери от такого сжатия данных должны быть сопоставимы с шумом в данных. Сжатие данных позволяет представить полезную информацию в более компактном виде, удобном для визуализации и интерпретации.

Наиболее популярным способом сжатия данных является *метод главных компонент* (PCA) [19]. Он дает основу для других аналогичных хемометрических методов, включая эволюционный факторный анализ (EFA) [134], оконный факторный анализ (WFA) [135], итерационный целевой факторный анализ (ITTFA) [136], а также многих методов классификации, например, *формального независимого моделирования аналогий классов* (SIMCA) [137]. С математической точки зрения метод главных компонент – это декомпозиция исходной 2D-матрицы **X**, т.е. представление ее в виде произведения двух 2D-матриц **T** и **P** [75]

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

В этом уравнении **T** называется матрицей *счетов* (scores) , **P** – матрицей *нагрузок* (loadings), а **E** – матрицей остатков (См. Рис. 3). Число столбцов – $\mathbf{t}_{\mathrm{a}}$ в матрице **T** и $\mathbf{p}_{\mathrm{a}}$ в матрице **P** – равно эффективному (химическому) рангу матрицы **X**. Эта величина *A* называется *числом главных компонент* (PC) и она, естественно, меньше, чем число столбцов в матрице **X**.

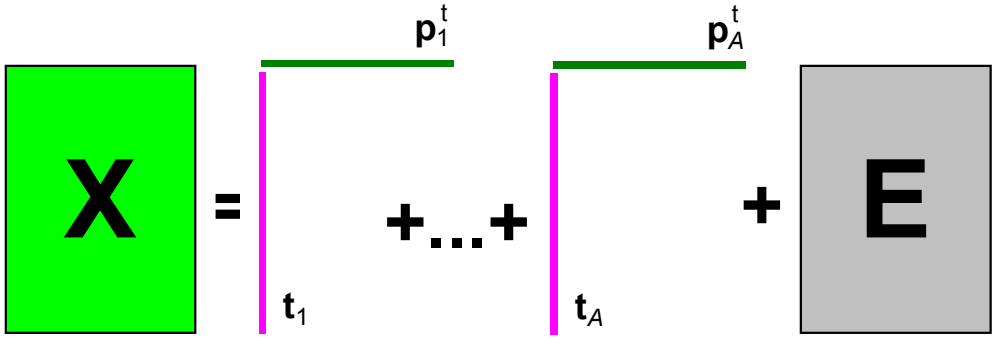


Рис. 3. Графическое представление метода главных компонент

Для иллюстрации метода PCA, мы опять вернемся к примеру, рассмотренному в конце раздела 2.1. Матрица спектров смесей **X** может быть представлена как произведение матрицы концентраций **C** и матрицы спектров чистых компонентов **S**

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

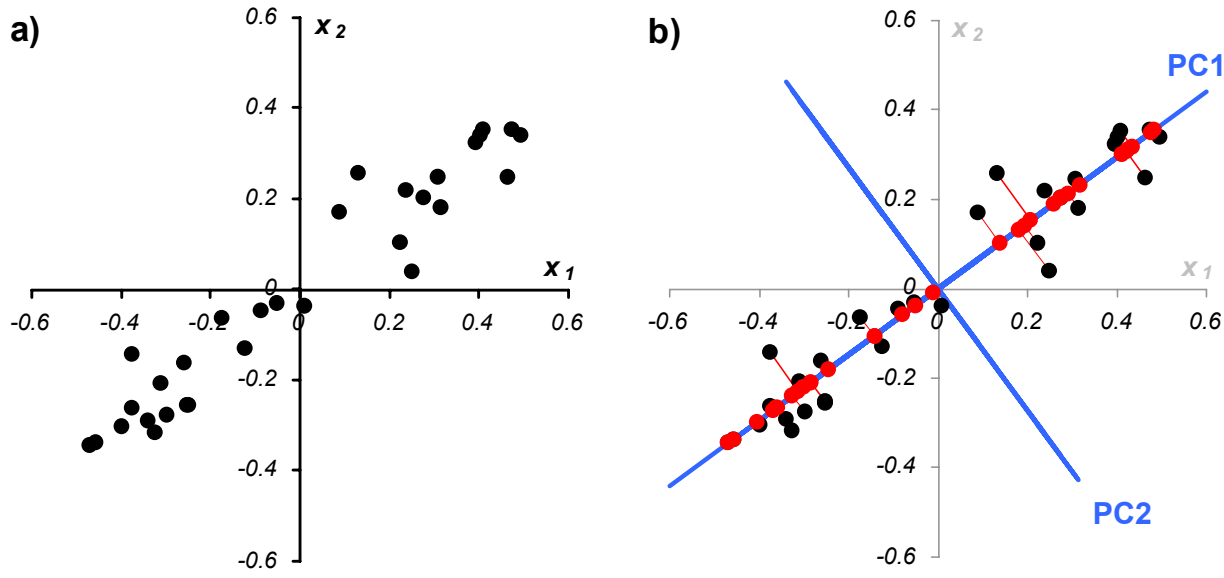
Число строк в матрице **X** равно числу образцов (*I*), и каждая ее строка соответствует спектру одного образца, снятому для *J* длин волн. Число строк в матрице **C** также равно *I*, а вот число столбцов соответствует числу компонентов в смеси (*A*=3). Матрица чистых спектров присутствует в разложении (2) в транспонированном виде, т.к. количество ее строк равно числу длин волн (*J*), а число столбцов равно *A*. Как уже отмечалось 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 неконцентрационных факторов, например, температуры.

Задача разделения экспериментальной матрицы **X** на «чистые» составляющие, соответствующие концентрациям **C** и спектрам **S** (понимаемым в обобщенном смысле), составляет предмет особой области в хемометрике, называемой *разделением кривых* (curve resolution) [138]. В этой области можно выделить два направления. Первое использует метод автомодельного разрешения кривых [139] (SMCR) и оно ориентированно, прежде всего, на приложение к гибридной хроматографии [140]. Для анализа применяются методы формального моделирования (PCA, EFA), которые не используют содержательное знание об исследуемой системе. В рамках этого подхода можно отметить метод SIMPLISMA [141], применяющий простой, но весьма эффективный подход, основанный на отборе переменных [142]. Второе направление, напротив, учитывает априорную информацию о процессах, применяя «серые» модели [143]. Это направление находит свое приложение при исследовании кинетики [35] и термодинамики [144] . Ключевым моментом в таких задачах является определение величины химического ранга системы – числа главных компонент *A* [145]. В идеале

предсказанные спектры **S** и концентрации **C** близки к истинным значениям, хотя их никогда не возможно восстановить точно. Причина этого не только в погрешностях эксперимента, но и в том, что спектры могут частично перекрываться. Когда PCA применяется для разделения данных на химически осмысленные компоненты, как в уравнении (2), он часто называется *факторным анализом*, в отличие от формального анализа главных компонент.

Метод главных компонент эффективен не только в задачах разделения. Он применяется при исследовательском 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 устройства данных. Дадим простейшую двумерную иллюстрацию метода PCA.

На Рис. 4a показаны данные, состоящие только из двух переменных x_1 и x_2, которые связаны сильной корреляцией. На соседнем рисунке те же данные представлены в новых координатах. 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 главная компонента $\mathbf{p}_2$ ортогональна первой, и ее направление (PC2) соответствует наибольшему изменению в остатках, показанных на Рис. 4b отрезками, перпендикулярными оси $\mathbf{p}_1$.



Данные в исходных координатах Данные в координатах главных компонент

Рис. 4. Графическая иллюстрация метода главных компонент

Этот тривиальный пример показывает, что метод главных компонент осуществляется последовательно, шаг за шагом. На каждом шаге исследуются остатки $\mathbf{E}_a$, среди них выбирается направление наибольшего изменения, данные проецируются на эту ось, вычисляются новые остатки, и т. д. Этот алгоритм называется NIPALS [75]. Другой популярный алгоритм сжатия данных – разложение по сингулярным значениям (SVD)[147] – строит ту же декомпозицию (1) без итераций. Остановка итерационной процедуры, или, другими словами, выбор числа главных компонент A, проводится с использованием критериев, показывающих точность достигнутой декомпозиции. Пусть исходная матрица $\mathbf{X}$ имеет размер: I строк и J столбцов, и в разложении (1) участвуют A главных компонент. Величины

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

называются нормированным *собственным значением* и *объясненной дисперсией*. Их обычно изображают на графике в зависимости от числа a, тогда резкое изменение величин (3) указывает на нужное значение числа главных компонент. Для правильного выбора A, необходимо использовать метод тест-валидации, либо кросс-валидации, так как это описано в разделе 2.2. Уравнения (1) не содержат в себе свободного члена, поэтому для декомпозиции данных их следует сначала отцентрировать и, иногда, нормировать. Подробнее о методах подготовки данных будет рассказано в разделе 5.1.

Метод главных компонент можно трактовать как проецирование данных на подпространство меньшей размерности. Возникающие при этом остатки $\mathbf{E}$ рассматриваются как шум, не содержащий значимой химической информации. В этом подпространстве можно ввести меру близости образцов, называемую *расстоянием Махаланобиса* (Mahalanobis) [148], с помощью которой удается решить многие проблемы качественного анализа. Другим мощным инструментом анализа данных в проекционном подпространстве является *прокрустово* (Procrustes) вращение [149].

При исследовании данных методом PCA, особое внимание уделяется графикам счетов и нагрузок. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый образец изображается в координатах ($\mathbf{t}_i$, $\mathbf{t}_j$), чаще всего – ($\mathbf{t}_1$, $\mathbf{t}_2$). Близость двух точек означает их схожесть, т.е. положительную корреляцию. Точки, расположенные под прямым углом, являются некоррелированными, а расположенные диаметрально противоположно – имеют отрицательную корреляцию. Применяя этот подход в задачах хроматографического анализа [43] 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 сложном кластере. Если график счетов используется для анализа взаимоотношений образцов, то график нагрузок применяется для исследования роли переменных. На графике нагрузок каждая переменная отображается точкой в координатах ($\mathbf{p}_i$, $\mathbf{p}_j$), например ($\mathbf{p}_1$, $\mathbf{p}_2$). Анализируя его аналогично графику счетов, можно понять, какие переменные связаны, а какие независимы. Совместное исследование парных графиков счетов и нагрузок, также может дать много полезной информации о данных: [75].

Рассмотрим пример практического использования PCA в химическом анализе. В работе [150] проверяется возможность применения БИК спектроскопии для обнаружения фальсифицированных лекарств. Исследовались образцы истинных (N1, 10 штук) и поддельных таблеток (N2, 10 штук) популярного спазмолитического средства. Двадцать спектров диффузионного рассеяния $R(\lambda)$ были сняты с помощью прибора Bomem MB160 с приставкой Powder Samplir, в диапазоне 3800–10000 см$^{-1}$ (1069 длин волн) без специальной подготовки образцов. Исходные данные были преобразованы как -log$R$, центрированы и подготовлены процедурой MSC [75], рассмотренной в разделе 5.1. Они показаны на Рис. 5

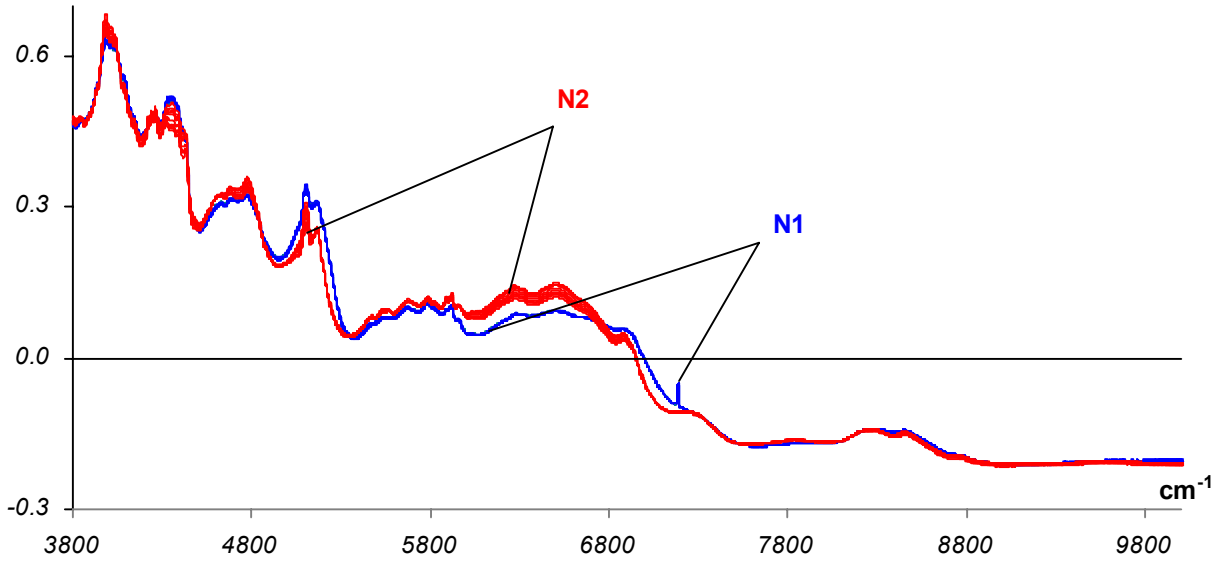
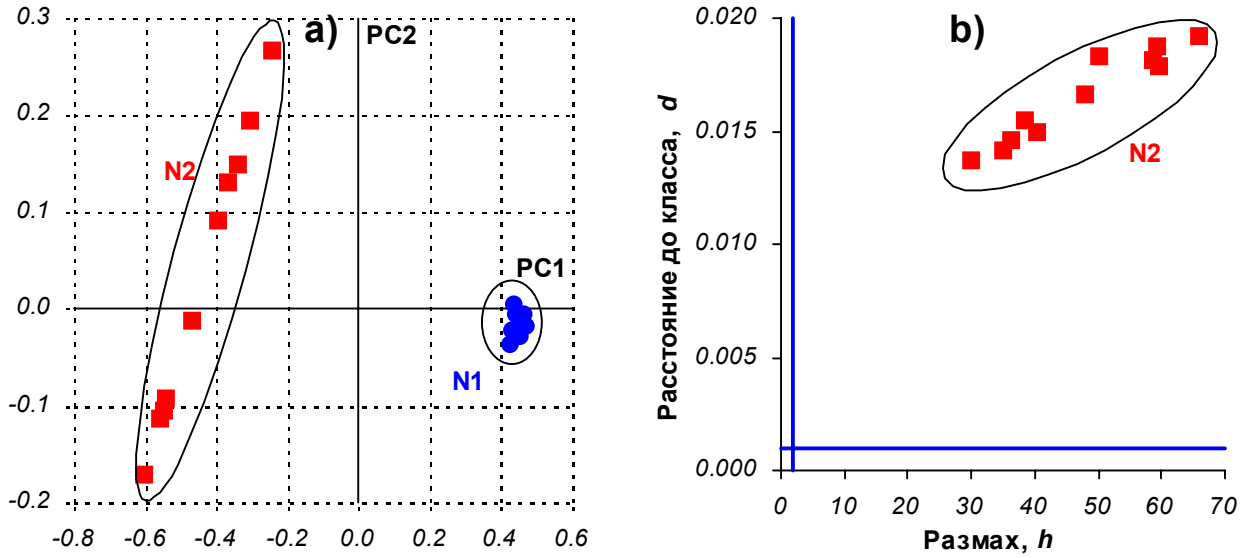


Рис. 5. Спектры , преобразованные процедурой MSC. N1 – истинные таблетки, N2 –фальсификат

Отрицательные значения сигнала объясняются тем, что для фона и спектров образцов были использованы различные регулировки усиления.

На Рис. 6а показан график PCA счетов ($\mathbf{t}_1$, $\mathbf{t}_2$) этих спектров. На нем четко видно две группы точек соответствующих истинным и фальсифицированным таблеткам. Разброс точек в группе N2 (контрафакт) существенно больше, чем в группе N1 (оригиналы). Это

может объясняться лучшим контролем при легальном производстве. В этом примере достаточно использовать только две главные компоненты, для которых μ_1=94% μ_2=4.9%, E_2= 99%.



Классификация на графике PCA счетов

Классификация методом SIMCA

Рис. 6. Определения фальсифицированных лекарств. N1 – истинные таблетки, N2 –фальсификат

3.2. Классификация и дискриминация

Рассмотренный 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 большие группы. К первой относятся так называемые задачи *без обучения* (unsupervised). Они названы так, потому, что в них не используется обучающий набор и их можно рассматривать как разновидность исследовательского анализа. Именно этот подход применялся в рассмотренном примере с фальшивыми таблетками. Задачи второй группы – классификация *с обучением* (supervised), называются также задачами *дискриминации*. В них применяется обучающий набор образцов, про которых имеется априорная информация о принадлежности к классам. Методы решения задач классификации без обучения основаны, главным образом, на PCA декомпозиции с последующим анализом [151] расстояний между классами, построением дендрограмм, использованием нечетких множеств [152] и т.п. В работе [153] применялось прокрустово вращение, а в работах [154-156] – расстояние Махалано-

биса. Однако, в тех случаях, когда возможно проведение дискриминации, т.е. классификации с обучением, этим методам следует отдавать предпочтение.

Обучающий набор образцов используется для построения модели классификации, т.е. набора правил, с помощью которых новый образец может быть отнесен к тому или другому классу. После того, как модель (или модели) построена, ее необходимо проверить, используя методы тест- или кросс-валидации, и определить насколько она точна. При успехе проверки, модель готова к практическому применению, т.е. к предсказанию принадлежности новых образцов. В аналитической химии классификация применяется к наборам мультиколлинеарных данных (спектры, хроматограммы), поэтому дискриминационная модель почти всегда многомерна и основана на соответствующих проекционных подходах – PCA, PLS. Можно отметить использование линейного дискриминационного анализа в БИК спектроскопии [157], а также канонического дискриминационного анализа [158]. Одним из самых популярных подходов является метод *формального независимого моделирования аналогий классов* (SIMCA [159]), разработанный С. Волдом [137].

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

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

Эта величина сравнивается со среднеквадратичным остатком внутри класса

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

Вторая величина определяет расстояние от объекта до центра класса, и она вычисляется как *размах* (квадрат расстояния Махаланобиса).

$$h = \frac{1}{I} + \sum_{a=1}^{A} \frac{\tau_{\mathrm{a}}^2}{\mathbf{t}_{\mathrm{a}}^{\mathrm{t}}\mathbf{t}_{\mathrm{a}}}$$

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

На Рис. 6b показано применение метода SIMCA для дискриминации таблеток. В качестве класса использовались 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 таблеток и, поэтому, легко могут быть дискриминированы.

Помимо метода SIMCA, для дискриминации химических данных используется также метод DASCO [160], который похож на SIMCA, метод классификация по ближайшему соседу (KNN) [161], метод опорных векторов (SVM) [162-163] и многие другие. Очень мощным инструментом является метод дискриминационного анализа с помощью регрессии на латентные структуры – PLS-DA [164]. Идея этого подхода состоит в том, что дискриминационные правила для K классов задаются линейными регрессионными уравнениями вида $\mathbf{XB}=\mathbf{D}$, где $\mathbf{X}$ – это полная матрица всех исходных данных ($I\times J$), $\mathbf{B}$ – это матрица неизвестных коэффициентов ($J\times K$), а $\mathbf{D}$ – это специальная матрица ($I\times K$), которая состоит из нулей и единиц. При построении матрицы $\mathbf{D}$ единицы ставят только в те строки (образцы), которые принадлежат классу, соответствующему номеру столбца. Регрессионная задача решается методом PLS (см. раздел 4.1), что позволяет в дальнейшем применять построенную регрессию для предсказания принадлежности новых образцов. Для этого строится прогноз отклика нового образца, и результат сравнивается с нулем или единицей.

3.3. Трехмодальные методы

Метод главных компонент был разработан для данных, имеющих вид двухмодальной 2D-матрицы. Однако, в последнее время, аналитики все чаще имеют дело с трех- и более модальными данными, которые имеют более сложную структуру, например, параллелепипед (Рис. 7). Источником таких данных служат, например, гибридные [166, 167] и эволюционные методы [168]. Для сжатия таких массивов применяются специальные подходы; три наиболее часто используемых будут кратко рассмотрены в этом подразделе. Самое полное и систематическое описание этих методов, вместе с многочисленными примерами их применения в задачах химического анализа, приведено в книге [94]. Краткое введение в методы анализа трехмодальных данных представлено в статье [165]. Эти же алгоритмы ис-

пользуются для обработки данных, полученных в результате гиперспектральных измерений [97], а также для анализа изображений [21].

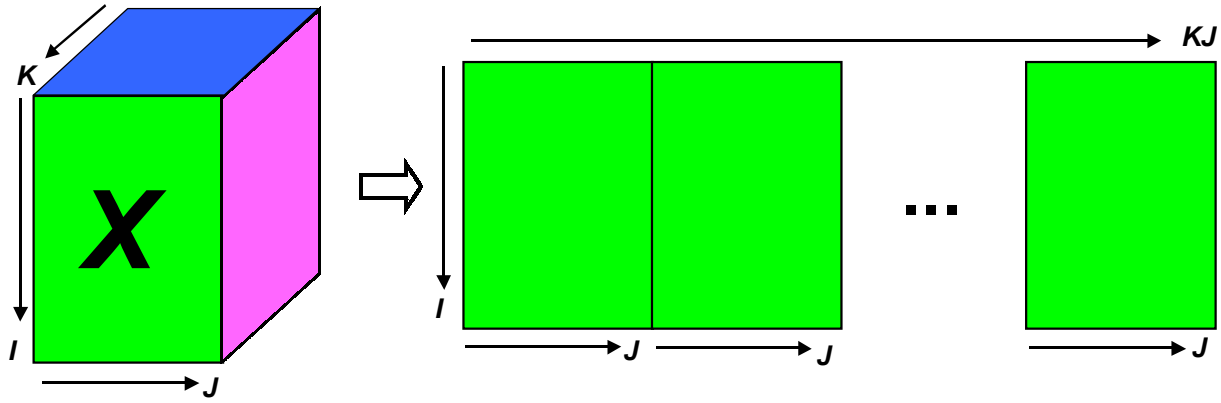


Рис. 7. Графическое представление метода развертывания в плоскую матрицу

Метод *развертывания* (unfolding) [169],– это простейший способ анализа трехмодальных данных, с помощью которого 3D-матрица $\boldsymbol{X}$ размерности $I \times J \times K$ разворачивается в обычную 2D-матрицу $^{u}\mathbf{X}$ размерности $I \times JK$ (см. Рис. 7). При этом I называется «основной» модой. Далее возможно обычное применение метода главных компонент (см. раздел 3.1). Такой подход часто оказывается эффективным, как, например в [37], хотя он и имеет ряд недостатков. Во-первых, в качестве основной моды, можно выбирать любое из трех направлений, т.е. имеется неоднозначность развертывания. Во-вторых, при таком подходе теряется связь между соседними точками, т.к. при переходе от 3D-матрицы $I \times J \times K$ к 2D-матрице $I \times KJ$ уже не учитывается, что измерения x_{ikj} и x_{ik+1j} являются соседними, что может быть существенно.

Алгоритм *Tucker3* [170] позволяет обрабатывать трехмодальные данные, сохраняя их первоначальную структуру, а, следовательно, и последовательность измерений, например порядок длин волн спектра, либо последовательность точек по времени в хроматограмме. Исходные 3D-данные $\boldsymbol{X}$ разлагаются на три обычные 2D-матрицы нагрузок (**A**, **B**, **C**) и трехмодальный керн-массив $\boldsymbol{G}$. Схема этого разложения изображена на Рис. 8. Каждый элемент исходной 3D-матрицы $\boldsymbol{X}$ можно записать в виде суммы (4),

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

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

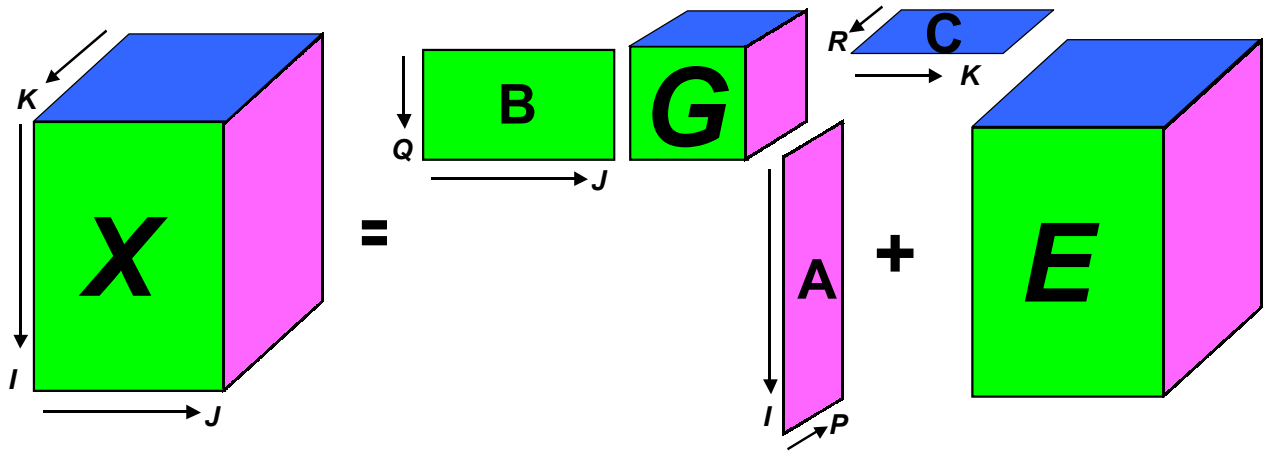


Рис. 8. Графическое представление модели Tucker3

Метод *PARAFAC* (parallel factor analysis) [165] отличается от модели Tucker3 тем, что каждая мода представляется одним и тем же числом главных компонент R. Разложение строится так, чтобы минимизировать сумму квадратов остатков e_{ijk}

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

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

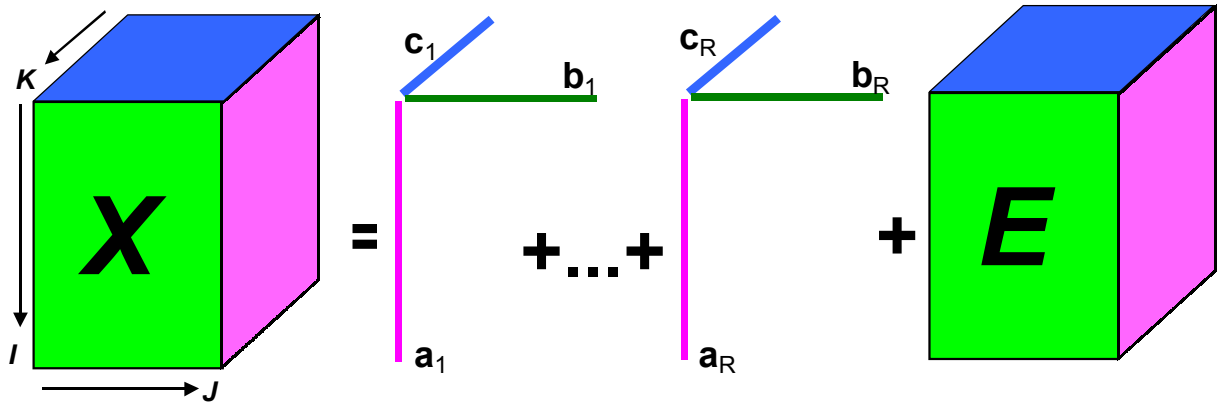


Рис. 9. Графическое представление модели PARAFAC с R компонентами

MATLAB код алгоритма PARAFAC можно найти в [165]. Так как матрицы нагрузок в разложении (5) определяются с помощью итерационной процедуры, этот метод требует очень большого объема вычислений. В настоящее время ведутся работы по ускорению вычислительных процедур. Последние достижения и их критический анализ представлены в [170], а алгоритмы всех рассмотренных методов декомпозиции трехмодальных данных приведены в [171].

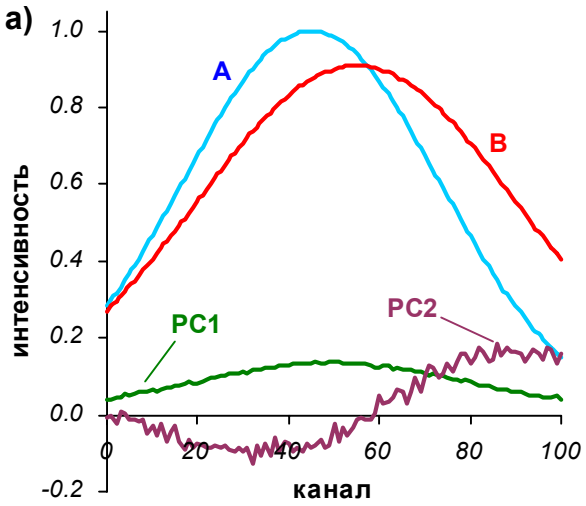
4. МЕТОДЫ КОЛИЧЕСТВЕННОГО АНАЛИЗА: ГРАДУИРОВКА

4.1. Линейная градуировка

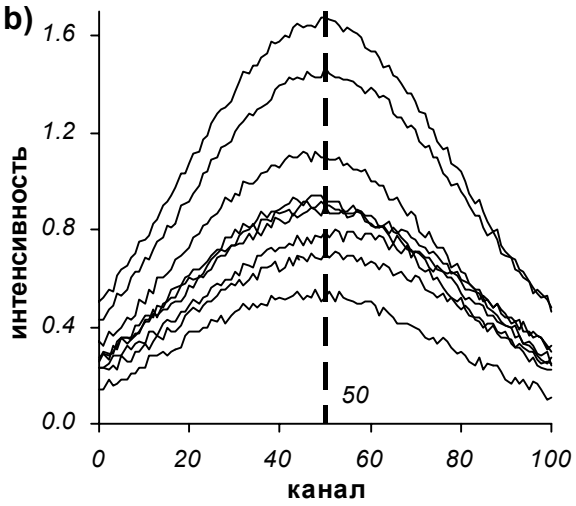
В задачах количественного анализа [2] участвуют два блока данных. Первый блок **X** – это матрица аналитических сигналов (например, спектров, хроматограмм, и т.п.). Второй блок **Y** – это матрица соответствующих химических показателей (например, концентраций). Число строк (I) в этих матрицах равно количеству образцов сравнения, число столбцов (J) в матрице **X** соответствует числу каналов (длин волн), на которых записывается сигнал, и, наконец, число столбцов (K) в матрице **Y** равно числу химических показателей, т.е. откликов. Задача градуировки состоит в построении 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), т.е. модель $y=a+bx$ [172], которая соответствует одному каналу аналитического сигнала. С помощью методов классического регрессионного анализа можно строить более сложную, множественную регрессию ($I>J$, K=1), в которой участвуют несколько каналов, т.е. **y**=**Xb** [34]. При использовании этих методов, обычно предполагается, что значения факторов x_{ij} известны точно, а погрешности присутствуют только в блоке **y**. В связи с этим, различают два подхода к построению модели: *прямая и обратная* градуировки [173]. В первом случае, в качестве независимых факторов используют химические показатели (**X**=**C**), а в качестве откликов – спектральные измерения (**Y**=**S**). Ранее считалось, что прямая модель лучше соответствует предположению о безошибочности блока **X**, а, кроме того, она еще и согласуется с законом Бугера-Ламберта-Бера (Bouguer-Lambert-Beer) [73]. Второй случай называется обратной градуировкой (**Y**=**C**, **X**=**S**). Этот подход является сейчас господствующим в хемометрике, поскольку он удобнее с практической точки зрения, т.к. непосредственно предсказывает нужный аналитический показатель (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, мы вновь используем пример, введенный в разделе 2.1. Теперь мы наполним его конкретным содержанием, смоделировав данные **X** и **Y**. Рассмотрим смесь двух веществ A и B (K=2) и предположим, что имеется некоторый прибор, позволяющий измерять аналитический сигнал s (спектр) на 101

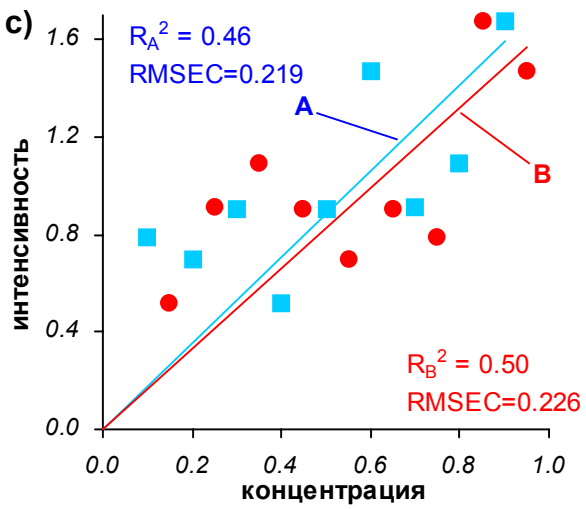
канале (J=101). Соответствующие спектры «чистых» веществ (c_A=c_B=1) показаны на Рис. 10a (кривые A и B). Спектры сильно перекрываются, так, что невозможно выделить какие-то «селективные» каналы для оценки концентраций. На соседнем рисунке Рис. 10b представлены девять модельных спектров (I=9) различных смесей A и B, в которые внесена случайная погрешность со стандартным отклонением 0.05. Они будут использоваться как обучающий набор.



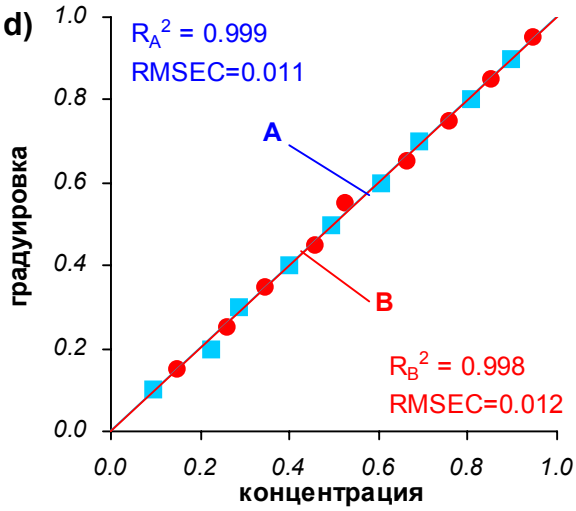
Спектры чистых (A, B) и главных компонент (PC1, PC2)



Модельные спектральные данные



Одномерная градуировка.



Градуировка методом PCR

Рис. 10. Модельный пример построения различных градуировок. ■ - A, ● - B

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

Точность градуировки принято характеризовать величиной *среднеквадратичного остатка градуировки* (RMSEC), который вычисляется по формуле

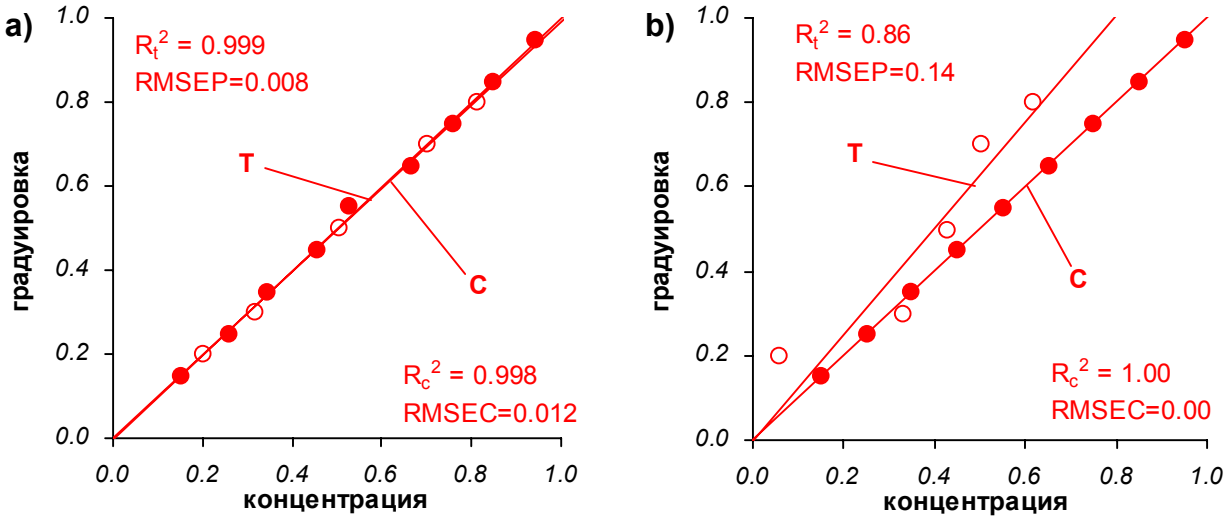
$$\mathrm{RMSEC} = \sqrt{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2 \Big/ F} \qquad (6)$$

где y_i и $\hat{y}_i$ соответственно, известные и предсказанные значения химического показателя (концентрации) для образцов сравнения i=1,..., I. Величина F – это число степеней свободы [43], и она равна I–1 для одномерной регрессии без свободного члена. Ясно, что чем меньше RMSEC, тем точнее описываются обучающие данные. Кроме того, качество градуировки характеризуется еще и *коэффициентом корреляции* R^2 между величинами **y** и $\hat{\mathbf{y}}$ – чем он ближе к единице, тем лучше. Соответствующие значения даны на Рис. 10c. Этот график подтверждает, что из-за недостатка «приборной» селективности одномерная градуировка неудовлетворительна. Градуировка с помощью множественной регрессии будет рассмотрена ниже, а сейчас покажем, как работает многомерная модель, построенная с помощью *регрессии на главные компоненты* (PCR [100]).

В методе PCR используется обратная градуировка, так, что **Y**=**C**, **X**=**S**. Применяя метод PCA, матрицу **X** можно разложить по формуле (1), причем в нашем примере, очевидно, A=2. Получившиеся вектора нагрузок $\mathbf{p}_1$ и $\mathbf{p}_2$ показаны на Рис. 10a (кривые PC1 и PC2). Сравнивая этот и соседний (b) графики, можно увидеть, что первая главная компонента описывает гладкий тренд в данных, тогда как вторая компонента представляет зашумленные отклонения от этого тренда. Полученная матрица счетов **T** используется как блок независимых факторов (предикторов) в регрессии на блок откликов **Y**, т.е. **Y**=**Tb**. Результаты градуировки методом PCR показаны на Рис. 10d, где изображены предсказанные значения концентраций $\hat{\mathbf{y}}$ в зависимости от соответствующих известных значений **y** (квадраты для A и круги для B), а также градуировочные линии, которые сливаются. Приведенные на этом же графике величины RMSEC и R^2, свидетельствуют о том, что метод PCR позволяет достичь высокой «математической» селективности и получить оценки концентраций веществ A и B с точностью, гораздо лучшей, чем в одноканальной градуировке. В методе PCR число степеней свободы в уравнении (6) равно, F=I–A.

Ранее уже отмечалось, что каждая хемометрическая модель нуждается в полноценной проверке. В нашем примере такая проверка проводилась с помощью проверочного набора (тест-валидация), состоящего из пяти образцов (смесей A и B). На Рис. 11a представлены результаты проверки для вещества B. Здесь, в координатах «известно-предсказано», изображены девять образцов, участвовавших в градуировке (круги) и пять

проверочных образцов (окружности). Там же приведены значения остатков градуировки (RMSEC) и проверки (RMSEP), а также коэффициенты корреляции для обучающего (R_c^2) и проверочного (R_t^2) наборов. Среднеквадратичный остаток проверки (RMSEP) вычисляется аналогично RMSEC (формула (6)), но только для образцов из проверочного набора. При этом F равно числу таких образцов. Видно, что метод главных компонент выдерживает проверку – градуировочная (C) и проверочная (T) линии сливаются. Рассмотрим, в этом контексте, метод градуировки с помощью *множественной регрессии*. Т.к. обучающий набор состоит из девяти образцов, то для построения этой модели мы можем использовать не более восьми каналов ($I>J$), например: 1, 14, 27, и т.д.



Метод главных компонент Множественная регрессия

Рис. 11. Проверка градуировок в модельном примере.
Обучающий (●, C) и проверочный (○, T) наборы.

На Рис. 11b показаны результаты – градуировка (C) и проверка (T) – для метода множественной регрессии. Поскольку в этом случае число образцов всего на единицу больше числа каналов, то градуировочная прямая в точности проходит через все точки обучающего набора (круги), поэтому RMSEC=0, и R_c^2=1. Однако проверка показывает неудовлетворительное качество такой градуировки – точность на порядок хуже точности в методе PCR, а проверочная прямая (T) значительно отличается от градуировочной (C). Это – типичный пример *переоценки* модели [72], когда точность описания обучающих данных значительно лучше, чем качество прогнозирования. Проблема сбалансированности описания данных рассматривается во многих работах А. Хоскюлдссона (A. Höskuldsson), который в 1988 году ввел новую концепцию моделирования – так называемый *H-принцип* [174]. Согласно этому принципу точность моделирования (RMSEC) и точность прогнозирования (RMSEP) связаны между собой. Улучшение RMSEC неминуемо влечет ухудшение

RMSEP, поэтому их нужно рассматривать совместно. Именно по этой причине множественная линейная регрессия, в которой всегда участвует явно избыточное число параметров, неизбежно приводит к неустойчивым моделям, непригодным для практического применения.

В настоящее время самым популярным методом многомерной градуировки в хемометрике является метод *проекции на латентные структуры* (PLS), который уже упоминался выше. Он во многом похож на метод PCR, с тем существенным отличием, что в PLS проводится одновременная декомпозиция матриц **X** и **Y**

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{t}+\mathbf{E} \qquad \mathbf{Y}=\mathbf{U}\mathbf{Q}^{t}+\mathbf{F}. \qquad (7)$$

Проекции строятся согласованно – так, чтобы максимизировать корреляцию между соответствующими векторами X-счетов $\mathbf{t}_a$ и Y-счетов $\mathbf{u}_a$. Поэтому PLS регрессия гораздо лучше описывает сложные связи, используя при этом меньшее число главных компонент. Детальное описание метода PLS приведено в [75]. Этот подход послужил основой для очень многих методов градуировки, используемых в хемометрике, таких как SIMPLS [175], PMN [176], робастный PLS [177], Ridge PLS [178], и многих других подходов [179].

Однако все эти методы дают результат 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 [129], а использование имитационных методов [113] затруднительно из-за большого времени расчетов [116]. В 1962 г. Л. Канторович [180] предложил другой подход к анализу данных – заменить минимизацию суммы квадратов отклонений на систему неравенств, которая решается с помощью методов линейного программирования. В этом случае результат прогноза сразу имеет вид интервала, поэтому этот метод и был назван «простым интервальным оцениванием» (ПИО) [37, 57]. С помощью этого метода было выполнено несколько работ в области аналитической химии [181].

4.2. Многомодальная регрессия

Методы многомерной градуировки естественно обобщаются на случай, в котором ***X*** и ***Y*** блоки являются N-модальными матрицами [94]. Эта регрессия может быть построена различными способами. Используя методы, описанные в разделе 3.3 (PARAFAC, Tucker3), блок предикторов ***X*** раскладывается в произведение 2D-матриц нагрузок, с помощью которых проводится оценка параметров. Эти методы можно рассматривать как обобщение

метода PCR для многомодальных данных. Обобщением метода PLS является Tri-PLS декомпозиция 3D-матрицы $\boldsymbol{X}$, которую можно представить в «матрицированном» виде [182]

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

Здесь $^{u}\mathbf{X}$ – это 2D матрица (размерности $I \times KJ$), получаемая при развертке 3-D матрицы $\boldsymbol{X}$ (размерности $I \times K \times J$), как это показано на Рис. 7. $\mathbf{T}$ – это 2D-матрица счетов (размерности $I \times A$), а $^{u}\mathbf{P}$ – это 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 $\boldsymbol{P} = {}^{J}\mathbf{P} \otimes {}^{K}\mathbf{P}$. Декомпозиция блока $\boldsymbol{Y}$ проводится аналогично $^{u}\mathbf{Y} \approx \mathbf{U} \cdot {}^{u}\mathbf{Q}$. Здесь так же, как и в обычном методе PLS, счета $\mathbf{T}$ выбираются таким способом, чтобы максимизировать корреляцию между векторами $\mathbf{t}_a$ и $\mathbf{u}_a$. Сама регрессионная задача $\mathbf{U}=\mathbf{TB}$ решается традиционным способом.

Математический аппарат, используемый в многомодальной градуировке, довольно сложен. Однако в настоящее время, существуют программные продукты [183], позволяющие аналитикам легко справляться с математическими трудностями. В литературе имеются многочисленные примеры использования мультимодальной градуировки в химическом анализе: кинетическая спектрофотометрия для определения пестицидов [184], разрешение налагающихся пиков в ВЭЖХ-ДДМ [185], определение следовых концентраций металлов [186], и, наконец, работа [187], которую мы рассмотрим более подробно.

В этой статье исследуется применение газовой хромато-масс-спектрометрии (ГХ-МС) для определения следовых концентраций кленбутерола в биологических образцах. В последнее время метод ГХ-МС является самым популярным среди всех гибридных методов. Он широко применяется в аналитических лабораториях, занимающихся следовым анализом органических веществ. Однако сложность биологических объектов, совместно с низким уровнем содержания исследуемого вещества, делает оценку предела обнаружения существенно зависимой от способа математической обработки экспериментальных данных. В рассматриваемой работе были приготовлены 7 стандартных образцов с известными концентрациями кленбутерола. Масс-спектрометрическое детектирование осуществлялось как в режиме полного сканирования (210 ионов), так и в режиме детектирования по отдельным, 8 ионам. Полученные данные имеют трехмодальную структуру, первая мода – образцы, вторая – масс-спектры, третья мода – хроматограммы. В режиме полного сканирования получается 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 – это 1D вектор $\mathbf{y}$, состоящий из 7 концентраций.

Для построения градуировок использовались различные трехмодальные алгоритмы: PARAFAC, PARAFAC2, Tucker3, а также Tri-PLS. Сравнение этих методов показало, что Tri-PLS является наилучшим, т.к. он дает наименьший предел обнаружения. Сопоставление этого метода со стандартной одномерной методикой показало значительное снижение предела обнаружения: в режиме полного сканирования с 283 мг $кг^{-1}$ до 20.91 мг $кг^{-1}$, при сканировании по отдельным ионам с 73.95 мг $кг^{-1}$ до 26.32 мг $кг^{-1}$. Для вычисления предела обнаружения использовалась концепция NAS [124].

4.3. Нелинейная градуировка

В некоторых случаях, 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 используются два альтернативных подхода: множественная нелинейная регрессия или многомерная нелинейная градуировка. В этом разделе мы рассмотрим оба эти подхода.

Нелинейный регрессионный анализ [188] может с успехом применяться для решения задач количественного анализа в случае, когда число переменных невелико. Кроме того, для его применения необходимо располагать содержательной моделью связывающей блоки **X** и **Y**. По-видимому, круг таких задач не очень широк – в него входят, почти исключительно, кинетические, в том числе титраметрические, проблемы [109]. Так, этот подход применялся при анализе активности антиоксидантов [37], для решения обратной кинетической задачи [35, 108], в уже упомянутом титровании [189, 190]. В работе [58] содержится подробный анализ проблем, с которыми сталкивается исследователь, применяющий этот подход.

Альтернативой классической регрессии является формальный подход, который не требует знания содержательной модели, но предполагает наличие большого числа данных [92]. Для учета нелинейных эффектов предлагаются разнообразные усовершенствования [95] обычного метода PLS: INLR [191], GIFI-PLS [192], QPLS [193]. Помимо нелинейного PLS, в хемометрике активно применяется метод *искусственных нейронных сетей* (ANN) [194, 195], имитирующий распространение сигналов в коре головного мозга. Этот метод с успехом используется для интерполяции функций и классификации. Последние 10 лет нейронные сети привлекли к себе большое внимание химиков, которые начали применять их для классификации [196], дискриминации [54] и градуировки [197, 198]. Затем, однако, наметилось некоторое охлаждение интереса, и использование ANN в хемометрике заметно снизилось. Причина заключена все в той же проблеме переоценки моделей, о которой шла речь вы-

ше. При использовании нейронных сетей очень трудно установить правильную степень сложности модели, что приводит к неустойчивому и ненадежному прогнозу. Другим интересным методом нелинейного моделирования, имитирующим биологические процессы, является *генетический алгоритм* (GA), с успехом применяемый в хемометрике [199, 200]. Метод GA, и его разновидность – иммунный алгоритм (IA), полезны в тех случаях, когда задача химического анализа не поддается формализации в терминах обычных целевых функций, например при разрешении многокомпонентных перекрывающихся хроматограмм [201]. Пример практического применения различных нелинейных подходов в хемилюминесцентном анализе приведен в работе [202].

5. ПОДГОТОВКА ДАННЫХ И ОБРАБОТКА СИГНАЛОВ

5.1. 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 является центрирование и нормирование [203]. *Центрирование* – это вычитание из исходной матрицы **X** матрицы **M**, т.е. $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{M}$. Обычно центрирование проводится по столбцам: для каждого вектора $\mathbf{x}_j$ вычисляются средние значения $m_j = \left(x_{1j} + \ldots + x_{Ij}\right) \Big/ I$, тогда **M**=($m_j$**1**,..., m_J**1**), где **1** – это вектор из единиц размерности *I*. Иногда центрирование проводится и по строкам. Тогда вычисляют средние значения по строкам, которое вычитается из соответствующей строки $\mathbf{x}_i^t$. В мультимодальных данных центрирование может проводиться по каждой моде отдельно. Центрирование необходимо в тех случаях, когда модель однородна, т.е. не содержит свободного члена – как в уравнениях (1) и (7). Оно понижает химический ранг модели на единицу и может улучшать точность описания. Это преобразование можно рассматривать как проецирование на нулевую главную компоненту [13], поэтому оно всегда применяется в методах PCA и PLS. Однако центрирование не применимо в том случае, когда в данных имеются пропуски.

Второе простейшее преобразование данных – это *нормирование*. Это преобразование, в отличие от центрирования, не меняет структуру данных, а просто изменяет вес различных частей данных при обработке. Нормирование также может проводиться по каждой моде. Нормирование по столбцам – это умножение исходной матрицы **X** слева на матрицу **W**, т.е. $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$. Матрица **W** – это диагональная матрица размерности $J \times J$. Обычно диагональные элементы w_{jj} равны обратным значениям стандартного отклонения $d_j = \sqrt{\sum_{i=1}^{I} \left(x_{ij} - m_j\right)^2 \Big/ I}$ по столбцу $\mathbf{x}_j$. Нормирование по строкам (называемое также нормализацией) – это умножение матрицы **X** справа на диагональную матрицу **W,** т.е. $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$. При этом размерность **W** равна $I \times I$, а ее элементы w_{ii} – это обратные значения стандартных отклонений строк $\mathbf{x}_i^t$. Комбинация центрирования и нормирования по столбцам $\tilde{x}_{ij} = \left(x_{ij} - m_j\right) \Big/ d_j$ называется *автошкалированием*. Нормирование данных часто применяют для того, чтобы уравнять вклад в модель от различных переменных (например, в

гибридном методе ЖХ-МС), учесть гетероскедастические погрешности, или для того, чтобы обрабатывать совместно разные блоки данных. Нормирование также можно рассматривать как метод, позволяющий стабилизировать вычислительные алгоритмы [72]. В тоже время, к этому преобразованию нужно относится с большой осторожностью, т.к. оно может сильно исказить результаты качественного анализа [43].

Помимо этих линейных преобразований используются и нелинейные трансформации результатов эксперимента. Так, в БИК спектроскопии часто применяется преобразование Кубелки-Мунка (Kubelka-Munck) [204]. Цель этого и других трансформаций, например, преобразования Бокса-Кокса (Box-Cox) [34] – линеаризация модели. Как уже отмечалось в разделе 4.3, часто простые операции с данными – логарифмирование [57], извлечение корня [37] помогают существенно улучшить модель.

Исходные данные почти всегда содержат погрешности, как случайные, так и систематические. Для того чтобы уменьшить влияние случайного шума, применяют различные методы сглаживания данных: скользящее среднее, Савицкого-Голея (Savitzky-Golay) [43, 205]. Удаление систематического сдвига в данных, так называемой «базовой линии», представляет более сложную задачу. В случае, когда этот сдвиг постоянен, то он убирается центрированием. Для линейных или квадратичных зависимостей от переменной (длины волны) помогает численное дифференцирование. Для более сложных зависимостей используются специальные метолы, два из которых мы рассмотрим. Метод *множественной коррекции сигнала*, называемый также мультипликативной коррекцией рассеяния (MSC) [71] был первоначально разработан [206] для БИК спектроскопии и базировался на идеях работы [204]. Процедура MSC преобразования устроена очень просто. Сначала определяется «базовый спектр» $\mathbf{m}^{\mathrm{t}} = \left(\mathbf{x}_1^{\mathrm{t}} + \ldots + \mathbf{x}_I^{\mathrm{t}}\right) \Big/ I$ как среднее по всем строкам матрицы **X**. Затем для каждой строки $\mathbf{x}_i^{\mathrm{t}}$ строится регрессия $\mathbf{x}_i^{\mathrm{t}} = a_i + b_i \mathbf{m}^{\mathrm{t}} + \mathbf{e}_i^{\mathrm{t}}$ на этот спектр, и определяются коэффициенты a_i и b_i. Преобразованные данные получаются из уравнения $\tilde{\mathbf{x}}_i^{\mathrm{t}} = \left(\mathbf{x}_i^{\mathrm{t}} - a_i\right) \Big/ b_i$. Параметры множественной коррекции a_i и b_i могут определяться не по всем переменным, а только по некоторому (подвижному) окну [207].

Второй метод, а точнее группа методов, называемых *ортогональной коррекцией сигнала* (OSC) [208] отличаются тем, что для преобразования матрицы предикторов **X** используется второй блок – откликов **Y**. Эти методы применяются для подготовки данных в задачах количественного анализа. Идея OSC состоит в том, чтобы удалить из блока **X** все систематические зависимости, которые не связаны с моделируемым откликом, т.е. ту часть **X**,

которая ортогональна **Y**. При этом должен увеличится коэффициент корреляции R^2 и уменьшится число PLS компонент A, необходимых для моделирования данных. Существует много вариантов этого метода, первоначально предложенного в [209], и развитого в работах [210, 211]. Также как и в методе PLS, процедура OSC осуществляется последовательно, по шагам. На каждом шаге из матрицы **X** удаляется часть, связанная с одной OSC-компонентой. Для определения части матрицы $\mathbf{X}=\mathbf{X}_1+\mathbf{X}_2$ ортогональной **Y**, т.е. $\mathbf{Z}=\mathbf{Y}^t\mathbf{X}_2=\mathbf{0}$, применяется алгоритм, аналогичный PLS. Подробное изложение метода и MATLAB код приведен в [211].

Альтернативой методам коррекции сигнала MSC и OSC является подход, в котором качество модели улучшается *отбором переменных*. Полезность отбора, т.е. исключение из исходного массива данных **X** некоторых столбцов $\mathbf{x}_j$, подтверждается как теоретическими, так и практическими исследованиями. Такой подход используется и в качественном [212], и количественном анализе [213]. Для отбора переменных применяются различные методы: генетический алгоритм [214], оптимизация Парето (Pareto) [215], «складного ножа» (jack-knife) [132]. Особую важность отбор переменных приобретает в тех методах, где аналитический сигнал непрерывно зависит от канала, например в спектрометрии [216]. Здесь отбор переменных осуществляется целыми блоками, как в методе [211], или в работе [217]. Помимо отбора переменных используется и *отбор образцов*, т.е. строк $\mathbf{x}_i^t$ 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 [128], при переносе градуировок с одного прибора на другой [131, 218, 219]. Новый подход к классификации и отбору образцов изложен в работе [57].

5.2. Обработка сигналов

Обработка аналитических сигналов с помощью различных преобразований и фильтров играет важную роль в химическом анализе [220, 221]. Так *преобразование Фурье*, по сути, произвело революцию в ЯМР, ИК и рентгеновской спектроскопии за последние 20 лет. Теперь исходные данные уже не регистрируются в виде привычных спектров, а записываются 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 сде-

лать спектроскопию ЯМР на ядрах ^{13}C обычным аналитическим методом, несмотря на нечувствительность ядер ^{13}C. Импульсная спектроскопия ЯМР позволила накапливать сигналы для большого количества импульсов и суммировать их. Одновременно с Фурье-спектроскопией возникло большое количество методов улучшающих качество полученных данных, часто называемых Фурье деконволюцией (разверткой, разделением сигналов), которые включают в себя различные манипуляции с исходными данными во временном домене, и только потом применение преобразования Фурье.

Другим мощным современным методом обработки сигналов является *вэйвлет* (wavelet) анализ [222]. С его помощью можно кодировать, сжимать и моделировать большие массивы данных, которые содержат тысячи переменных. Вэйвлет анализ является естественным продолжением и развитием методов Фурье. Недостатком разложения Фурье является то, что его базисные функции непрерывно зависят от времени и поэтому они не пригодны для представления данных, зависящих от времени. В вэйвлет анализе применяются 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. Вэйвлет анализ часто предшествует методам PCA или PLS, что позволяет применять их к очень большим данным без потери информации [223] .Этот метод часто применяется для сжатия и сглаживания одно и двухмодальных ИК и ЯМР спектров [224].

Часто нужны методы, позволяющие сглаживать сигналы быстро, в реальном времени. Одним из таких методов является *фильтр Калмана* (Kalman), который еще совсем недавно, в конце 1980-х и в начале 1990-х, привлекал внимание многих хемометриков. С его помощью можно, например, смоделировать ход химической кинетики, не дожидаясь окончания процесса. В таких приложениях, как контроль процессов, часто нужно увидеть сглаженную кривую по ходу дела, в реальном времени. Общая идея фильтра Калмана состоит в том, чтобы уточнять модель по ходу развития процесса. Как только новые данные становятся доступны, модель дополняется и улучшается. С появлением быстрых и мощных компьютеров, нужда в фильтре Калмана практически исчезла, хотя отдельные работы, где он полезен [225], еще встречаются.

6. ЗАКЛЮЧЕНИЕ

6.1. Аналитический контроль процессов

Мы рассмотрели основные достижения хемометрики в аналитической химии за последние 15 – 20 лет. В то же время, за пределами обзора остались очень многие актуальные направления и приложения, как близкие к аналитической химии, так и далекие от нее. Одно направление заслуживает особого рассмотрения, поскольку в нем наиболее ярко проявились тенденции и перспективы развития как хемометрики в частности, так и аналитики, в целом. Речь идет о методах *аналитического контроля процессов* (PAT).

В 30-х годах прошлого века американский статистик У. Шухарт (W. Shewhart) предложил использовать статистические методы для контроля хода технологических процессов [226]. Его идея была очень проста – если собрать и статистически обработать исторические данные о нормально функционирующем производстве, 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, которые требуют немедленного вмешательства. Такой метод был назван *статистическим контролем процессов* (SPC) и начал с успехом применятся 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,..., как правило, связаны между собой, т.е. коррелированы и должны рассматриваться совместно. Ситуация во многом напоминает анализ многоканальных химических данных (например, спектров). Эта аналогия позволила Дж. МакГрегору (J. MacGregor) [227] разработать новый подход к этой задаче, названный *многомерным статистическим контролем процессов* (MSPC) [228]. Он предложил использовать метод главных компонент (PCA) для анализа многомерных исторических данных и строить контрольные пределы в пространстве счетов с помощью расстояния Махаланобиса. Идея оказалась чрезвычайно плодотворной и нашла много последователей. Усилиями хемометриков были разработаны специальные методы для многомерного контроля периодических (например, биохимических) процессов, основанные на трехмодальной градуировке [24], для анализа сложных, многоблочных процессов были созданы иерархические [229], блочные [230], и маршрутные методы [96]. Появились работы, в которых предлагалось не только контролировать, но и оптимизировать ход процессов [231]. Эти теоретические разработки начали применять на

ские разработки начали применять на практике, прежде всего в пищевой [232] и фармацевтической [233] промышленности. Были созданы системы контроля качества производства полимеров [234], цветных металлов [235], полупроводников [42].

Таким образом, в хемометрике возникло новое направление [236], стремительно отдаляющееся от аналитической химии. Однако время все расставило по своим местам. 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 реальном времени (*in-line*) и даже *in-situ* [237]. Для этой цели прекрасно подошли традиционные аналитические методы, прежде всего, спектрометрия: УФ [108], и НПВО-ИК [238]. В том же контексте контроля химических реакций и процессов оказались востребованы хемометрические методы разрешения кривых и оценки кинетических констант по спектроскопическим [239] и хроматографическим [240] данным. Использование MSPC в сочетании с аналитическими методами мониторинга, а также методами контроля качества химического анализа [241] породило новое направление в аналитической химии – аналитический контроль процессов (PAT) [242]. PAT – это система планирования, анализа и контроля критических переменных, характеризующих состояние производственных материалов и процессов в реальном времени (т.е. по ходу производства), с целью подтверждения качества производимого продукта.

Главная проблема, с которой сталкивается современная промышленность – это обеспечение постоянно высокого качества конечного продукта в типовом производственном процессе. Принимая во внимание увеличивающуюся глобальную конкуренцию и быстро меняющиеся потребности рынка, эффективный контроль процессов в реальном времени становится насущной потребностью всех производящих компаний. Аналитическая химия, в том числе и хемометрика, играет определяющую роль в решении этой задачи. Решение Федеральной комиссии по контролю за лекарствами (FDA) [243], вышедшее в сентябре 2004, законодательно подтвердило эту роль. Это очень важное событие. Хотя за 30 лет своего существования хемометрика создала много замечательных методов и алгоритмов, они очень медленно и неохотно признавались регулирующими органами во всем мире. Идеи многомерного подхода трудны для восприятия и визуализации в сравнении с традиционными одномерными методами, которые, однако, не всегда могут дать полную картину происходящего. Так, 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 некоторую область в про-

странстве PLS-счетов, определяемую с помощью расстояния Махаланобиса и т.д., и т.п. За все время существования хемометрики удалось принять лишь один, действительно хемометрический нормативный документ [244]. Теперь же, с принятием документа о PAT, хемометрика неизбежно станет легитимным инструментом для всех компаний, желающих следовать руководящим принципам FDA.

По нашему мнению, которое разделяют и многие коллеги, выход этого нормативного документа знаменует кардинальный поворот в технологии, новую *парадигму производства*, в которой главным лозунгом становится: «сделать качество неотъемлемым свойством продукта». Принципиальное отличие этой парадигмы от существующей ныне – это отказ от принципа стандартизации и унификации в пользу гибкого, оперативного управления будущим качеством на всех стадиях производства, начиная от анализа входного сырья. Приведем простейший пример, иллюстрирующий эту мысль. Предприятия общественного питания, работающие в рамках парадигмы «стандартизации», безусловно, проигрывают в качестве домашней еде, которая производится в условиях гибкого, оперативного контроля. Однако, внедрение системы PAT, позволит и массовым производителям обеспечить столь же высокое качество, при сохранении больших объемов выпуска продукции.

6.2. Перспективы развития

Какова же роль аналитической химии в этом процессе? Как и чем могут аналитики ответить на этот вызов времени? Что нужно изменить в российском «аналитическом процессе», для того, чтобы успешно вписаться в этот исторический поворот? Нам представляется, что развитие аналитики во всем мире будут определять следующие тенденции. Во-первых, *объекты анализа* станут более сложными и комплексными. Технологические потребности будут ставить перед аналитиками не частные вопросы – сколько вещества X в пробе, а общие вопросы – получится ли продукт нужного качества из этого сырья, или правильно ли развивается химическая реакция в этой колонке. Во-вторых, *методы анализа* будут меняться таким образом, чтобы обеспечить получение необходимых данных не в лаборатории (at line), а непосредственно на производстве, в реальном времени (in line). В-третьих, резко увеличится *объем данных*, которые повсеместно станут многомодальными и многомерными. Увеличится 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 самых изощренных методов ее извлечения. В-пятых, изменится организация аналитического экспе-

римента – вместо исследования одной пробы в одном опыте, будет использоваться *системный подход*, в котором много разных проб автоматически испытываются одновременно разными методами, в разных условиях. Такой массовый компьютеризованный эксперимент, пример которого мы уже видим в технологии микрочипов, станет рутинной аналитической практикой. В-шестых, акцент в аналитическом исследовании будет переноситься на *биологические объекты* и биохимические процессы, а также на исследование технологических процессов в целом.

Все эти тенденции, которые уже сейчас прослеживаются в аналитической химии, изменят и роль химика аналитика. Он неизбежно станет более *аналитиком, чем химиком*. Две главные задачи будут стоять перед этим исследователем. Первая – как придумать, организовать, спланировать эксперимент с тем, чтобы получить данные, из которых, в принципе, можно получить нужную информацию. При этом искомой информацией может быть не количественный (концентрация) или качественный (да/нет) результат, а прогноз финального состояния исследуемой системы в будущем, после прохождения ее через много стадий химических и физических превращений. Вторая – как извлечь эту информацию из данных, интерпретировать ее в категориях полезности и качества. Для решения этих задач исследователь должен, в значительной мере, использовать опыт и инструментарий хемометрики. Все это свидетельствует о том, что хемометрика, как неотъемлемая часть аналитической химии, в значительной мере определяет направления ее развития.

В 1825 году Огюст Конт (A. Comte), французский философ, основоположник позитивизма писал [245]: – "*Каждая попытка применить математические методы для исследования химических проблем должна рассматриваться как абсолютно абсурдная и противоречащая самому духу химии. Если математический анализ, когда-либо займет сколько-нибудь значительное место в химии – извращение, которое по счастью почти невероятно – это повлечет за собой повсеместно быстрое вырождение этой науки*". Это пророчество, к счастью, оказалось неверным, и мы становимся свидетелями роста применения математических методов в аналитической химии и одновременного распространения этой науки далеко за пределы ее обычного ареала. Если российские ученые хотят быть не только свидетелями, но и активными участниками этого процесса, нам необходимо предпринять срочные меры по развитию хемометрики. По нашему мнению, необходимо значительно усилить уровень преподавания хемометрики в университетах. Для этого нужно подготовить (написать или перевести) учебник по хемометрике, разработать несколько типовых программ обучения для химиков-аналитиков, технологов, инженеров и т.п. Считаем, что можно ставить вопрос о соответствующих специальностях в магистерских спе-

циализациях, а также в кандидатских и докторских Советах. И уж, конечно, надо незамедлительно решить вопрос о подписке на ведущие журналы в этой области: Journal of Chemometrics и Chemometrics and Intelligent Laboratory Systems, которые сейчас нельзя найти ни в одной российской библиотеке.

Несмотря на все имеющиеся объективные и субъективные трудности, мы с оптимизмом оцениваем перспективы развития хемометрики в России. Наблюдается растущий интерес к этой науке как со стороны химиков-аналитиков, так и среди других специалистов – физиков и математиков. Сравнивая современное положение дел с ситуацией, имевшей место еще пять-семь лет назад, нельзя не отметить значительный рост публикаций российских ученых в отечественных и международных журналах, посвященных хемометрике. Аналитическая химия является разработчиком и носителем определенной идеологии и методологии, обеспечивает профессионализм в анализе. Последние 20 лет доказали, что включение хемометрики, как неотъемлемой части, в методологию химического анализа позволяет значительно расширить арсенал аналитических методов, сделать их более эффективными и быстрыми. Хемометрика позволяет значительно расширить сферу применения аналитических методов, для этого необходимо тесное сотрудничество аналитиков с другими учеными: математиками, физиками, биологами.

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

7. ЛИТЕРАТУРА

1. М.А. Шараф, Д.Л. Иллмэн, Б.Р. Ковальски. *Хемометрика*, Пер. с англ. М. Мир: 1987 [M. Sharaf, D. Illman, B. Kowalski. Chemometrics. NY: Wiley. 1986]
2. Ю.А. Золотов. *Аналитическая химия: проблемы и достижения*, М. Наука 1992
3. Ю.В. ГрановскийУспехи и проблемы. *Вест. МГУ Сер 2 Химия*, **38**, 211 (1997)
4. Ю.А. Карпов, Т.М. Полховская. *Стандартизация и метрология в металлургическом производстве*: М. МИСИС 1989
5. P. Geladi, K. Esbensen. Chemometrics, a growing and maturing discipline (Editorial). *Chemom. Intell. Lab. Syst*., **7**, 197 (1990)
6. D.L. Massart. *Chemometrics: a textbook*, Elsevier, NY, 1988
7. S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst*., **30**, 109 (1995).
8. M. Blanco, I. Villarroya. NIR spectroscopy: a rapid-response analytical tool, *Trends Anal. Chem*., **21**, 240 (2002)
9. B.G. Osborne, T. Fearn. *Near Infrared Spectroscopy in Food Analysis*, Longman Scientific and Technical, Harlow, Essex, England, 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. MartensComparison of linear statistical methods for calibration of NIR instruments. *Appl. Stat.,* **35,** 195 (1986)
13. H. Martens, T. Næs. *Multivariate calibration. I. Concepts and distinctions. Trends Anal. Chem.,* **3**, 204 (1984)
14. K. Pearson. *On lines and planes of closest fit to systems of points in space, Philippine Mag.,* **2** (6), 559 (1901)
15. W.S. Gosset (”Student”). *The probable error of a mean. Biometrika*, **6** , 1 (1908)
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* М, 1960
19. S. Wold, K. Esbensen, P. Geladi. *Principal component analysis. Chemom. Intell. Lab. Syst.*, **2**, 37 (1987).

20. R.I. Shrager. Chemical transitions measured by spectra and resolved using singular value decomposition. *Chemom. Intell. Lab. Syst.*, **1**, 59 (1986)
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. Applicational aspects of support vector machines. *J. Chemom.,* **16**, 482 (2002)
24. P. Nomikos, J.F. MacGregor. *Monitoring batch processes using multiway principal component analysis. American Inst. Chem. Engin. 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. Application of self-assembly techniques in the design of biocompatible protein microarray surfaces, *Electrophoresis*, **23**, 3097 (2002)
27. M.M.C. Ferreira. 9th International Conference on Chemometrics in Analytical Chemistry (CAC-2004), Lisbon, Portugal, 20–23 September 2004 , *J. Chemom*., **18**, 385 (2004)
28. I.E. Frank, J.H. Friedman. *A statistical view of some chemometrics regression tools (with discussion). Technometrics*, **35**, 109 (1993)
29. S. Wold, A. Berglund, N. Kettaneh. *New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production) -with examples from pharmaceutical research and process modeling J. Chemom*., **16**, 377 (2002)
30. J. Friedman.Boosting and bagging, Lect. at the Gordon Conf. on Statist. and Chem. Engineering, Williamstown, MA, 2001. Доступно на http://www.amstat.org/sections/spes/GRC2001.htm [1 мая 2005].
31. G. Molenberghs. Biometry, Biometrics, Biostatistics, Bioinformatics,.., Bio-X *Biometrics*, **61**, 1 (2005)
32. А.Г. Шмелев. *Традиционная психометрика и экспериментальная психосемантика: объектная и субъектная парадигмы анализа данных, Вопросы Психологии*, №5, 34 (1982)
33. H. Wold. В кн.Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Perspectives in Probability and Statistics (papers in honour of M. S. Bartlett on the occasion of his 65 th birthday)*, Applied Probability Trust, University of Sheffield, 1975, p. 117.
34. Н. Дрейпер, Г. Смит. *Прикладной регрессионный анализ*, (в 2-х т.) Москва, Финансы и статистика, 1987 [N.R. Draper, H. Smith, Applied regression analysis, Wiley, N.Y.]

35. О.Е. Родионова, А.Л. Померанцев. *Об одном методе решения обратной кинетической задачи по спектральным данным при неизвестных спектрах компонент, Кинетика и катализ*, **45**, 485 (2004)
36. H.-L. Koh, W.-P. Yau, P.-S. Ong, A. Hegde. *Current trends in modern pharmaceutical analysis for drug discovery. Drug Discov. Today*, **8**, 889 (2003)
37. A.L. Pomerantsev, O.Ye. Rodionova. Hard and soft methods for prediction of antioxidants' activity based on the DSC measurements, *Chemom. Intell. Lab. Syst.*, в печати (2005)
38. Л. Грибов. *Математические методы и ЭВМ в аналитической химии*, М. 1989
39. K.J. Siebert. *Chemometrics in Brewing - A Review. J. Am. Soc. Brew. Chem.*, **59**, 147 (2001).
40. K. Varmuza, W. Werther, F.R. Krueger, J. Kissel, E.R. Schmid. *Organic substances in cometary grains: Comparison of secondary ion mass spectral data and Californium-252 plasma desorption data from reference compounds Int. J. Mass Spectrom.*, **189**, 79 (1999).
41. G.W. Johnson, R. Ehrlich. *State of the art report on multivariate chemometric methods in environmental forensics. Environ. Forensics*, **3**, 59 (2002).
42. B.M. Wise, N.B. Gallagher, E.B. Martin. *Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. J. Chemom.*, **15**, 285 (2001)
43. R.G. Brereton. *Chemometrics: Data analysis for the laboratory and chemical plant.* Wiley, Chichester, UK. 2003
44. Н.П. Комарь *Основы качественного химического анализа*. Харьков, 1955
45. Л.А. Грибов, В.И. Баранов, М.Е. Эляшберг. *Безэталонный молекулярный спектральный анализ. Теоретические основы.* М. Едиториал УРСС, 2002.
46. М. Эляшберг. Экспертные системы для установления структуры органических молекул спектральными методами *Успехи химии*, **68**, 579 (1999)
47. Б. Марьянов, А. Зарубин, С. Шумар. Статистический анализ данных дифференцированного потенциометрического осадительного титрования бинарной смеси трех гетеровалентных ионов с помощью линейных характеристик *Журн. аналит. химии*, **58**, 1126 (2003)
48. В.И. Вершинин, Б.Г. Дерендяев, К.С. Лебедев. *Методы компьютерной идентификация органических соединений* М. Академкнига, 2002
49. Zenkevich I.G.; Kránicz B. Choice of nonlinear regression functions for various physicochemical constants within series of homologues *Chemom. Intell. Lab. Syst.,* **67,** 51 (2003)
50. I.V. Pletnev, V.V. Zernov. *Classification of metal ions according to their complexing properties: a data driven approach Anal. Chim. Acta*, **455**, 131 (2002)

51. Н.М. Гальберштам, И.И. Баскин, В.А. Палюлин, Н.С. Зефиров. Нейронные сети как метод поиска зависимостей структура - свойство органических соединений *Успехи химии*, **72**, 706 (2003).
52. В.И. Дворкин. *Метрология и обеспечение качества количественного химического анализа*, М. Химия, 2001.
53. Ю.Г. Власов, А.В. Легин, А.М. Рудницкая. *Мультисенсорные системы типаэлектронный язык- новые возможности создания и применения химических сенсоров , Успехи химии*, 2005, в печати
54. А.В. Калач, Я. И. Коренман, С.И. Нифталиев. *Искусственные нейронные сети – вчера, сегодня, завтра.* Воронеж: Воронеж. гос. технол. акад. 2002.
55. С.П. Казаков, А.А. Рябенко, В.Ф. Разумов. Спектрофотометрическое исследование фотоинициированного превращения 4а,4в-дигидродибензфенантрена в транс-1,2-ди(2-нафтил)этилен. Доказательство одноквантового механизма методом сингулярного разложения, *Оптика и спектроскопия*, **86**, 537 (1999)
56. В.Ф. Разумов, М.В. Алфимов. Фотохимия диарилэтиленов, *ЖНиПФ*, **46**, 28 (2003)
57. O.Ye. Rodionova, K.H. Esbensen, A.L. Pomerantsev. Application of SIC (Simple Interval Calculation) for object status classification and outlier detection - comparison with PLS/PCR, *J. Chemom.,* **18**, 402 (2004)
58. E.V. Bystritskaya, A.L. Pomerantsev, O.Ye. Rodionova. *Non-linear regression analysis: new approach to traditional implementations J. Chemom.*, **14,** 667 (2000)
59. A. Bogomolov, M. McBrien. Mutual peak matching in a series of HPLC/DAD mixture analyses, *Anal. Chim. Acta*, **490**, 41 (2003)
60. A. Bogomolov, M. McBrien. Methods for Characterizing a Mixture of Chemical Compounds, US Patent, US-2004-0126892-A1, (2004)
61. S. Kucheryavski, V. Polyakov, A. Govorov. Analysis of simulated fracture surfaces using AMT and fractal geometry methods, В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, 2005, pp. 3– 11.
62. Н.М. Оскорбин, А.В. Максимов, С.И. Жилин. Построение и анализ эмпирических зависимостей методом центра неопределенности, *Изв. АлтГУ*, №1, 35 (1998)
63. S.V. Romanenko A.G. Stromberg, E.V. Selivanova, E.S. Romanenko. . *Resolution of the overlapping peaks in the case of linear sweep anodic stripping voltammetry via curve fitting Chemom. Intell. Lab. Syst.*, **73**, 7 (2004)

64. И.Е Васильева, А.М Кузнецов, И.Л Васильев, Е.В. Шабанова. Градуировка методик атомно-эмиссионного анализа с компьютерной обработкой спектров *Журн. аналит. химии.* **52**, 1238 (1997)
65. 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
66. 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
67. T. Næs, T. Isaksson, T. Fearn, T. Davies. *Multivariate Calibration and Classification*, Christerer, UK, 2002
68. R. Kramer. *Chemometric Techniques for Quantitative Analysis*, Marcel-Dekker, 1998
69. K.R. Beebe, R.J. Pell, M.B. Seasholtz. *Chemometrics: a Practical Guide*, Willey, N.Y., 1998
70. E.R. Malinowski. *Factor Analysis in Chemistry*, Wiley, N.Y., 2nd edn, 1991
71. H. Martens, T. Næs. *Multivariate calibration*. Wiley, New York. 1989
72. A. Höskuldsson. *Prediction Methods in Science and Technology*, vol. 1, Thor Publishing, Copenhagen, Denmark, 1996
73. *Аналитическая химия. Проблемы и подходы* (в 2-х т.), под. ред. Кельнер Р., Мерме Ж.-М., Отто М., Видмер Г.М., пер. с анг., М., Мир АСТ, 2004 [Analytical Chemistry. The Approved Text to FECS Curriculum Analytical Chemistry, Wiley-VCH, Weinheim]
74. Б.М. Марьянов. *Избранные главы хемометрики*, Томск: Из-во Том. ун-та, 2004
75. К. Эсбенсен. *Анализ многомерных данных*, сокр. пер. с англ. под ред. О.Родионовой, Из-во ИПХФ РАН, 2005 [K.H. Esbensen. *Multivariate Data Analysis – In Practice 4-th Ed*., CAMO, 2000]
76. *The 9th Scandinavian Symposium on Chemometrics (SSC9)* Доступно на http://www.conference.is/ssc9/ [1 мая 2005]
77. K. Esbensen, O. Rodionova, A. Pomerantsev, O. Startsev, S. Kucheryavskiy. *Meeting Report Second Russian Winter School on Chemometrics (WSC-2), Feb. 28–March 06, Barnaul and Belokuriha, Altai, Russia J. Chemom*. **17**, 1, 2003
78. O. Ye. Rodionova. Second Winter School on Chemometrics, *Chemom. Intell. Lab. Syst.,* **67,** 194 (2003)
79. S. Kucheryavski, C. Marks, K. Varmuza. *Meeting report: Fourth winter symposium on chemometrics—WSC4 ''Multivariate Data Analysis'', 15–18 February, 2005, Institute of Problems of Chemical Physics, Chernogolovka, Chemom. Intell. Lab. Syst.* **78**, 138, (2005)

80. *Home of Chemometry Consultancy*. Доступно на http://www.chemometry.com/ [1 мая 2005]
81. *Chemometrics literature database*. Доступно на http://www.models.kvl.dk/ris/risweb.isa [1 мая 2005]
82. *Chemometrics World*. Доступно на http://www.wiley.co.uk/wileychi/chemometrics/Home.html [1 мая 2005]
83. *The Alchemist*. Доступно на http://www.chemweb.com/alchemist/ [1 мая 2005]
84. *Российское Хемометрическое Общество*. Доступно на http://rcs.chph.ras.ru/ [1 мая 2005]
85. *Хемометрика в России*. Доступно на http://www.chemometrics.ru/ [1 мая 2005]
86. *The Unscramber*. Доступно на http://www.camo.no/ [1 мая 2005]
87. *Eigenvector Research, Inc*. Доступно на http://www.eigenvector.com/ [1 мая 2005]
88. *Umetrics*. Доступно на http://www.umetrics.com/ [1 мая 2005]
89. *SPSS*. Доступно на http://www.spss.com/ [1 мая 2005]
90. *STATISTICA*. Доступно на http://www.statsoftinc.com/ [1 мая 2005]
91. *MATLAB*. Доступно на http://www.mathworks.com/ [1 мая 2005]
92. L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold. *Multi- and Megavariate Data Analysis*, Umetrics, Umeå, 2001
93. E. Sanchez, B.R. Kowalski. *J. Chemom.*, **2** 247 (1988)
94. A. Smilde, R. Bro, P. Geladi. *Multi-way Analysis with Applications in the Chemical Sciences*, John Wiley & Sons, Chichester, 2004.
95. S. Wold, J. Trygg, A. Berglund, H. Antti. *Some recent developments in PLS modeling Chemom. Intell. Lab. Syst.,* **58** , 131 (2001)
96. A. Höskuldsson. *Causal and path modelling. J. Chemom*., **58**, 287 (2001).
97. P. Geladi, J. Burger, T. Lestanderet. *Hyperspectral imaging: calibration problems and solutions, Chemom. Intell. Lab. Syst*., **72**, 209 (2004)
98. G.H.W. Sanders, A. Manz. *Chip-based microsystems for genomic and proteomic analysis Trends Anal. Chem*., **19**, 364 (2000)
99. G.E.P. Box, W.G. Hunter, J.S. Hunter. *Statistics for Experimenters*, John Wiley & Sons Inc., NY, 1978
100. Е.З. Демиденко. *Линейная и нелинейная регрессии*, Финансы и статистика, М, 1981
101. P. Jy. *Sampling for Analytical Purposes*. John Wiley & Sons, Chichester, 1989
102. W. Kleingeld, J. Ferreira, S. Coward. *First World Conference on Sampling and Blending (WCSB1), J. Chemom*., 18, 121 (2004)

103. Special Issue. : 50 years of Pierre Gy's Theory of Sampling Proceedings: First World Conference on Sampling and Blending (WCSB1) Tutorials on sampling. *: Theory and Practice Chemom. Intell. Lab. Syst*., **74**, 1-236 (2004)
104. B. Walczak, D.L. Massart. *Tutorial. Dealing with missing data. Chemom. Intell. Lab. Syst*., **58**, 15 (2001)
105. P.R.C. Nelson, P.A. Taylor, J.F. MacGregor. Missing data methods in PCA and PLS: Score calculations with incomplete observations *Chemom. Intell. Lab. Syst*., 35, 45 (1996)
106. H. Haario, V.-M. Taavitsainen. *Chemom. Intell. Lab. Syst*., **44**, 77 (1998)
107. Э.Ф. Брин, А.Л. Померанцев. *Хим. физика*, **5**, 1674 (1986)
108. S.P. Gurden, J.A. Westerhuis, S. Bijlsma, A.K. Smilde. *Modelling of spectroscopic batch process data using grey models to incorporate external information. J. Chemom*., **15**, 101 (2001)
109. А.Л. Померанцев. Методы нелинейного регрессионного анализа для моделирования кинетики химических и физических процессов. Дис. д-ра физ.-мат. наук, ИХФ РАН, Москва, 2003.
110. D. A. Morales. *Mathematical modeling of titration curves, J. Chemom*., **16**, 247 (2002)
111. A. de Juan, M. Maeder, M. Martınez, R.Tauler. *Combining hard- and soft-modelling to solve kinetic problems, Chemom. Intell. Lab. Syst*. **54**, 123 (2000)
112. О.Н. Карпухин. Глобальные (стратегические) проблемы практического применения сложных математико-статистических методов (хемометрики), Док. на 4-ом междун. симп. "Современные методы анализа многомерных данных" (WSC-4). Черноголовка, 14-18 февраля, 2005. Доступно на http://www.chemometrics.ru/articles/karpukhin/ [1 мая 2005]
113. Б. Эфрон. *Нетрадиционные методы многомерного статистического анализа*, Москва, Финансы и Статистика, 1988 [B. Efron, *Ann. Stat.* , **7**, 1 (1979)].
114. *EURACHEM/CITAC Guide, Quantifying Uncertainty in Analytical Measurement*, 2nd ed., EURACHEM, Lisbon, Portugal, 2000.
115. K Faber, B. R. Kowalski. *Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler. Chemom. Intell. Lab. Syst.*, **34**, 283 (1996)
116. A.L. Pomerantsev. Confidence Intervals for Non-linear Regression Extrapolation, *Chemom. Intell. Lab. Syst.*, **49**, 41 (1999)
117. A. Pulido, I. Ruisánchez, R. Boqué, F.X. Rius. *Uncertainty of results in routine qualitative analysis, Trends Anal. Chem*., **22**, 647 (2003)

118. V.I. Vershinin. *A priori method of evaluating uncertainties in qualitative chromatographic analysis:(probabilistic approach). Accreditation and Quality Assurance*, **9**, 415 (2004)
119. N.M. Faber. *Uncertainty estimation for multivariate regression coefficients. Chemom. Intell. Lab. Syst.*, **64**, 169 (2002)
120. N.M. Faber, R. Bro. *Standard error of prediction for multiway PLS 1. Background and a simulation study, Chemom. Intell. Lab. Syst.*, **61**, 133 (2002)
121. A. Lorber. *Anal. Chem.* **58**, 1167 (1986)
122. J. Ferré, N.M. Faber. *Net analyte signal calculation for multivariate calibration, Chemom. Intell. Lab. Syst.*, **69**, 123 (2003)
123. R. Boqué, N.M. Faber, F. Xavier Rius. *Detection limits in classical multivariate calibration models, Anal. Chim. Acta,* **423**, 41 (2000)
124. R. Boqué; J. Ferré, N.M. Faber, F. Xavier Rius. *Limit of detection estimator for second-order bilinear calibration, Anal. Chim. Acta*, **451**, 313 (2002)
125. I. Berget, T. Næs. Using unclassified observations for improving classifiers *J. Chemom.*, **18**, 103 (2004)
126. D. Jouan-Rimbaud, D.L. Massart, C.A. Saby, C. Puel. *Characterization of the representativity of selected sets in multivariate calibration and pattern recognition. Anal. Chim. Acta*, **350**, 149 (1997)
127. M. Meloun, J. Militký, M. Hill, R.G. Brereton. *Crucial problems in regression modelling and their solutions,Analyst*, **127**, 433 (2002)
128. J.A. Fernandez Pierna, F. Wahl, O.E. de Noord, D.L. Massart. *Methods of outlier detection in prediction Chemom. Intell. Lab. Syst.*, **63**, 27 (2002)
129. K. Faber. *Comparison of two recently proposed expressions for partial least squares regression prediction error, Chemom. Intell. Lab. Syst.*, **52**, 123 (2000)
130. N.M. Faber, X.-H. Song, P.K. Hopke. *Sample-specific standard error of prediction for partial least squares regression, Trends Anal. Chem.*, **22**, 330 (2003)
131. E. Bouveresse, D.L. Massart. *Standardization of near-infrared spectrometric instruments: A review. Vibrat. Spectrosc.*, **11**, 3 (1996)
132. F. Westad, H. Martens. *Variable selection in NIR based on significance testing in Partial Least Squares Regression (PLSR). J. Near Infrared Spectros.*, **8**, 117 (2000)
133. M. Hubert, S. Verboven. *A robust PCR method for high-dimensional regressors. J. Chemom.*, **17**, 438 (2003)
134. H.R. Keller, D.L. Massart. Evolving factor analysis. *Chemom. Intell. Lab. Syst.*, **12**, 209 (1992)

135. E.R. Malinowski. *Window Factor Analysis: theoretical derivation and application to ow-injection analysis data. J. Chemom.*, **6**, 29 (1992)
136. P.J. Gemperline. Target transformation factor analysis with linear inequality constraints applied to spectroscopic-chromatographic data. *Anal. Chem.*, **58**, 2656 (1986).
137. S. Wold. *Pattern recognition by means of disjoint principal components models. Pattern Recognition*, **8**, 127 (1976)
138. J.-H. Jiang, Y. Liang, Y. Ozaki. Principles and methodologies in self-modeling curve resolution. *Chemom. Intell. Lab. Syst.*, **71**, 1 (2004)
139. F.C. Sanchez, B. van de Borgaert, S.C. Rutan, D.L. Massart. *Chemom.Intell. Lab. Syst.*, **34**, 139 (1996)
140. H. Shen, B. Grande, O.M. Kvalheim, I. Eide. *Anal. Chim. Acta,* **446**, 313 (2001)
141. W. Windig, J. Guilment. Interactive self-modeling mixture analysis. *Anal. Chem.*, **63**, 1425 (1991)
142. A. Bogomolov, M. Hachey. В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, 2005, pp. 119-135.
143. J. Diewok, A. de Juan, M. Marcel, R. Tauler, B. Lendl. *Anal. Chem.*, **76**, 641 (2003)
144. А. Ю. Богомолов, Т.Н. Ростовщикова, В.В. Смирнов, *Ж. Физ. Хим*. **69,** 1197 (1995)
145. H.A. Seipel, J.H. Kalivas. *Effective rank for multivariate calibration methods, J. Chemom.*, **18**, 306 (2004)
146. S.R. Crouch, A. Scheeline, E.S. Kirkor. *Kinetic determinations and some kinetic aspects of analytical chemistry. Anal. Chem.*, **72**, 53 (2000)
147. R.I. Shrager. *Chemical transitions measured by spectra and resolved using singular value decomposition. Chemom. Intell. Lab. Syst.*, **1**, 59 (1986)
148. R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart. *Tutorial. The Mahalanobis distance, Chemom. Intell. Lab. Syst.*, **50**, 1 (2000)
149. J.M. Andrade, M. P.Gomez-Carracedo, W. Krzanowski, M. Kubista. Procrustes rotation in analytical chemistry, a tutorial, *Chemom. Intell. Lab. Syst.*, **72**, 123 (2004)
150. O.Ye. Rodionova, L.P. Houmøller, A.L. Pomerantsev, P. Geladi, J. Burger, V.L. Dorofeyev, A.P. Arzamastsev. *NIR Spectrometry for Counterfeit Drug Detection. A Feasibility Study, Anal. Chim. Acta*, в печати (2005)
151. L.X. Sun, K. Danzer. *J. Chemom.*, **10**, 325 (1996)
152. A.J. Myles, S.D. Brown. *Induction of decision trees using fuzzy partitions, J. Chemom.*, **17**, 531 (2003)
153. D. González-Arjona, G. López-Pérez, A.G. González. *Talanta,* **49**, 189 (1999)

154. H. Mark. *Use of Mahalanobis distances to evaluate sample preparation methods for near-infrared reflectance analysis, Anal. Chem*., **59**, 790 (1987)
155. P.J. Gemperline, N.R. Boyer. *Classification of near-infrared spectra using wavelength distances: comparison to the Mahalanobis distance and residual variance methods, Anal. Chem.*, **67**, 160 (1995)
156. H.L. Mark, D. Tunnell. *Qualitative near-infrared reflectance analysis using Mahalanobis distances, Anal. Chem.*, **57**, 1449 (1985)
157. U. Indahl, N.S. Sing, B. Kirkhuus, T. Næs. *Chemom.Intell. Lab. Syst.*, **49**, 19 (1999)
158. G. Downey, J. Boussion, D. Beauchene. *J.Near Infrared Spectrosc.*, **2**, 85 (1994)
159. G.R. Flåten, B. Grung, O.M. Kvalheim. *A method for validation of reference sets in SIMCA modelling, Chemom. Intell. Lab. Syst.*, **72**, 101 (2004)
160. T. Næs, U. Indahl, *J. Chemom.*, **12**, 205 (1998)
161. J. McElhinney, G. Downey, T. Fearn. *J.Near Infrared Spectrosc.*, **7**, 145 (1999)
162. S. Zomer, R. Brereton, J.F. Carter, C. Eckers. *Support vector machines for the discrimination of analytical chemical data: application to the determination of tablet production by pyrolysis-gas chromatography-mass spectrometry Analyst*, **129**, 175 (2004)
163. V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, I.V. Pletnev. *Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions J. Chem. Inf. Comput. Sci*., **43**, 2048 (2003)
164. M. Sarker, W. Rayens. *Partial least squares for discrimination. J. Chemom*., **17**, 166 (2003)
165. R. Bro. *PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst.*, **38**, 149 (1997)
166. A. Herrero, S. Zamponi, R. Marassi, P. Conti, M.C. Ortiz, L.A. Sarabia. *Determination of the capability of detection of a hyphenated method: application to spectroelectrochemistry Chemom. Intell. Lab. Syst.*, **61**, 63 (2002)
167. I. Garcia, L. Sarabia, M. C. Ortiz, J. M. Aldama. *Anal. Chim. Acta*, **515**, 55 (2004)
168. S. Bijlsma, A.K. Smilde. *Estimating reaction rate constants from two-step reaction: a comparison between two-way and three-way methods. J. Chemom.*, **14**, 541 (2000)
169. H. Kiers. Some procedures for displaying results from three-way methods, *J. Chemom.*, **14**, 151 (2000)
170. N.M. Faber, R. Bro, P.K. Hopke. *Recent developments in CANDECOMP/PARAFAC algorithms: a critical review, Chemom. Intell. Lab. Syst.*, **65**, 119 (2003)
171. C.A. Andersson, R. Bro. *The N-way toolbox for MATLAB, Chemom. Intell. Lab. Syst.*, **52** , 1 (2000)

172. F.J. del Rio, J. Riu, F.X. Rius. *Prediction intervals in linear regression taking into account errors on both axes. J. Chemom.*, **15**, 773 (2001)
173. R.G. Brereton.Introduction to multivariate calibration in analytical chemistry. *Analyst*, **125**, 2125 (2000)
174. A. Höskuldsson. PLS Regression Methods *J. Chemom.*, **2**, 211 (1988)
175. S. de Jong. *SIMPLS: an alternative approach to partial least squares regression Chemom. Intell. Lab. Syst.*, **18**, 251 (1993)
176. B. Li, A.J. Morris, E.B. Martin. *Generalized partial least squares regression based on the penalized minimum norm projection Chemom. Intell. Lab. Syst.*, **72**, 21 (2004)
177. M. Hubert, K.Vanden Branden. *Robust methods for partial least squares regression J. Chemom.,* **17**, 537 (2003)
178. E. Vigneau, M. Devaux, M. Qannari, P. Robert. *Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration, J. Chemom.,* **11**, 239 (1997)
179. P. Geladi. *Some recent trends in the calibration literature Chemom. Intell. Lab. Syst*., **60**, 211 (2002)
180. Л.В. Канторович. *Сиб. мат. журн.*, **3** , 701 (1962)
181. В.М. Белов, В.А. Суханов, Ф.Г. Унгер. *Теоретические и прикладные аспекты метода центра неопределенности*. Новосибирск: Наука, 1995
182. R. Bro. Multi-way calibration. Multi-linear PLS, *J. Chemom.*, **10**, 47 (1996)
183. R. Bro, C.A. Andersson. *The N-way Toolbox for MATLAB, Version 2.02, 2003*. Доступно на http://www.models.kvl.dk/source [1 мая 2005]
184. Y. Ni, C. Huang, S. Kokot. *Application of multivariate calibration and artificial neural networks to simultaneous kinetic-spectrophotometric determination of carbamate pesticides Chemom. Intell. Lab. Syst.,* **71**, 177 (2004)
185. Z.P. Chen, J. Morris, E. Martin, R.-Q. Yu, Y.-Z. Liang, F. Gong. *Recursive evolving spectral projection for revealing the concentration windows of overlapping peaks in two-way chromatographic experiments Chemom. Intell. Lab. Syst*., **72**, 9 (2004)
186. F. M. Fernández, M. B. Tudino, O. E. Troccoli. *Multicomponent kinetic determination of Cu, Zn, Co, Ni and Fe at trace levels by first and second order multivariate calibration, Anal. Chim. Acta*, **433**, 119 (2001)
187. I. Garcia, L. Sarabia , M.C. Ortiz, J.M. Aldama. *Three-way models and detection capability of a gas chromatography-mass spectrometry method for the determination of clenbuterol in*

several biological matrices: the 2002/657/EC European Decision Anal. Chim. Acta, **515**, 55 (2004)

188. Й. Бард. *Нелинейное оценивание параметров*. М.: Статистика, 1979. [Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974]
189. D.M. Barry, L. Meites. *Titrimetric applications of multiparametric curve-fitting. Part 1 Potentiometric titrations of weak bases with strong acids at extreme dilutions. Anal. Chim. Acta*, **68**, 435 (1974)
190. Б. Марьянов. В кн. *Химики ТГУ на пороге третьего тысячелетия*. Томск, Изд -во ТГУ, 1998, сс. 48-58
191. A. Berglund, S. Wold. *INLR, implicit non-linear latent variable regression J. Chemom.*, **11**, 141 (1997)
192. A. Berglund, N.L.U. Kettaneh, S. Wold, N. Bendwell, D.R.Cameron. *The GIFI approach to non-linear PLS modelling. J. Chemom.*, **15**, 321 (2001)
193. S. Wold. *Chemom. Intell. Lab. Syst.*, **14**, 71 (1992)
194. J. Zupan, J. Gasteiger. *Anal. Chim. Acta*, **248**, 1 (1991)
195. J. Zupan, J. Gasteiger. *Neural Network for Chemists, An Introduction*. VCH, Weinheim, 1993.
196. W. Wu, B. Walczak, D.L. Massart, E. Heuerding, F.E. Erni, I.R. Last, K.A. Prebble. Artificial neural networks in classificafion of NIR spectral data: Design of the training set, *Chemom. Intell. Lab. Syst.*, 33, 35 (1996)
197. J.R.M. Smits, W.J. Melssen, L.M.C. Buydens, G. Kateman. *Using artificial neural networks for solving chemical problems. Part I. Multi-layer feed-forward networks. Chemom. Intell. Lab. Syst*., **22** 165 (1994)
198. W.J. Melssen, J.R.M. Smits, L.M.C. Buydens, G. Kateman. *Using artificial neural networks for solving chemical problems. Part II. Kohonen self-organising feature maps and Hopfield networks. Chemom. Intell. Lab. Syst.*, **23**, 267 (1994)
199. D.B. Hibbert.Genetic algorithms in chemistry. *Chemom. Intell. Lab. Syst.*, **19**, 277, (1993)
200. R. Leardi. *Genetic algorithms in chemometrics and chemistry: a review. J. Chemom*., **15**, 559 (2001)
201. X. Shao, Z. Chen, X. Lin. *Resolution of multicomponent overlapping chromatogram using an immune algorithm and genetic algorithm, Chemom. Intell. Lab. Syst.*, **50**, 91 (2000)
202. L.A. Tortajada-Genaro, P. Campíns-Falcó, J. Verdú-Andrés, F. Bosch-Reig. *Multivariate versus univariate calibration for nonlinear chemiluminescence data Application to chro-*

mium determination by luminol-hydrogen peroxide reaction Anal. Chim. Acta, **450**, 155 (2001)

203. R. Bro, A.K. Smilde. *Centering and scaling in component analysis, J. Chemom.*, **17**, 16 (2003)
204. P. Kubelka, F. Munck. *Z. Tech. Phys.,* **12**, 593 (1931)
205. A. Savitzky, M.J.E. Golay. *Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem.*, **36**, 1627 (1964)
206. P. Geladi, D. MacDougall, H. Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat, *Appl. Spectrosc.*, **3**, 491 (1985)
207. T. Isaksson, B. Kowalski. *Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products, Appl. Spectrosc.*, **47**, 702 (1993)
208. J. Trygg, S. Wold. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter, *J. Chemom.*, **17**, 53 (2003)
209. S. Wold, H. Antti, F. Lindgren, J. Öhman. *Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst.*, **44**, 175 (1998)
210. T. Fearn. On orthogonal signal correction, *Chemom. Intell. Lab. Syst.*, **50**, 47 (2000)
211. A. Höskuldsson. *Variable and subset selection in PLS regression. Chemom. Intell. Lab. Syst.*, **55**, 23 (2001)
212. Q. Guo, W. Wu, D.L. Massart, C. Boucon, S. de Jong. *Feature selection in principal component analysis of analytical data, Chemom. Intell. Lab. Syst.*, **61**, 123 (2002)
213. M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizarro Millan. Selection of useful predictors in multivariate calibration *Anal. Bioanal. Chem.,* **380**, 397 (2004)
214. R. Leardi, R. Boggia, M. Terrile. *Genetic algorithms as a strategy for feature selection, J. Chemom.,* **6**, 267 (1992)
215. J.H. Kalivas. *Pareto calibration with built-in wavelength selection, Anal. Chim. Acta,* **505**, 9 (2004)
216. N. Benoudjit, E. Cools, M. Meurens, M. Verleysen. *Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear models Chemom. Intell. Lab. Syst*., **70**, 47 (2004)
217. U. Indahl, T. Næs. *A variable selection strategy for supervised classification with continuous spectroscopic data, J. Chemom.*, **18**, 53 (2004)
218. R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré. *Transfer of multivariate calibration models: a review Chemom. Intell. Lab. Syst*., **64**, 181 (2002)

219. E.L. Sulima, V.A. Zubkov, L.A. Rusinov. Specific features of practical implementation of calibration model transfer from a master instrument to slave NIR analyzers for analysis of main characteristics of wheat. В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, 2005, pp. 196–203.
220. J-H Jiang, Y. Ozaki, M. Kleimann, H.W. Siesler. *Resolution of two-way data from on-line Fourier-transform Raman spectroscopic monitoring of the anionic dispersion polymerization of styrene and 1,3-butadiene by parallel vector analysis (PVA) and window factor analysis (WFA) Chemom. Intell. Lab. Syst.*, **70**, 83 (2004)
221. P.W. Hansen. Pre-processing method minimizing the need for reference analyses, *J. Chemom.,* **15**, 123 (2001)
222. К. Чуи. *Введение в вэйвлеты*, М. Мир. 2001 [C.K. Chui. An Introduction to wavelets, Academic Press, 1992]
223. J. Trygg , S. Wold. *Chemom. Intell. Lab. Syst.*, **42**, 209 (1998)
224. S.-P. Reinikainen. Wavelets in Compressing Spectral Data В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, 2005, pp. 21-36.
225. Y. Pan, C.K. Yoo, J.H. Lee, I.-B. Lee. Process monitoring for continuous process with periodic characteristics, *J. Chemom.,* **18**, 69 (2004)
226. W.A Shewhart. *Economic Control of Quality of Manufactured Product*, Van Nostrand, New York, 1931
227. J. MacGregor, Th. Kourti. Statistical process Control of Multivariate Processes. *Control Engineering Practice*, **3**, 403 (1995)
228. А.Л. Померанцев, О.Е. Родионова. Многомерный статистический контроль процессов *Методы менеджмента качества*, **6**, 15 (2002)
229. Th. Kourti, J. MacGregor. *Process analysis, monitoring and diagnosis, using multivariate projection methods. Tutorial. Chemom. Intell. Lab. Syst.*, **28**, 3, (1995)
230. J.A. Westerhuis, Th. Kourti, J.F. Macgregor. *Analysis of multiblock and hierarchical PCA and PLS models, J. Chemom.*, **12**, 301 (1998)
231. A.L. Pomerantsev, O.Ye. Rodionova. В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, 2005Multivariate statistical process control and optimisation,, pp. 209-227
232. R. Bro.Exploratory study of sugar production using fluorescence spectroscopy and multiway analysis, *Chemom. Intell. Lab. Syst.*, **46**, 133 (1999)
233. J. Gabrielsson, N-O. Lindberg, T. Lundstedt. *Multivariate methods in pharmaceutical applications, J. Chemom.*, **16**, 141 (2002)

234. C.K. Yoo, J.-M. Lee, P.A. Vanrolleghem, I.-B. Lee. *On-line monitoring of batch processes using multiway independent component analysis , Chemom. Intell. Lab. Syst.*, **71**, 151 (2004)

235. M. Baroni, P. Benedetti, S. Fraternale, F. Scialpi, P. Vix, S. Clementi. *The CARSO procedure in process optimization, J. Chemom.*, **17**, 9 (2003)

236. H. Martens, M. Martens. *Multivariate Analysis of Quality: An Introduction.* John Wiley & Sons Ltd.: Chichester, 2001.

237. R.M. Dyson, M. Hazenkamp, K. Kaufmann, M. Maeder, M. Studer, A. Zilian. *Modern tools for reaction monitoring: hard and soft modelling of non-ideal, on-line acquired spectra. J. Chemom.*, **14**, 737 (2000)

238. K. Pöllänen, A. Häkkinen, S.-P. Reinikainen, M. Louhi-Kultanen, L. Nyström. *ATR-FTIR in monitoring of crystallization processes: comparison of indirect and direct OSC methods, Chemom. Intell. Lab. Syst.,* **76**, 25 (2005)

239. T.J. Thurston, R.G. Brereton, D.J. Foord, R.E.A. Escott. *Principal components plots for exploratory investigation of reactions using ultraviolet-visible spectroscopy: application to the formation of benzophenone phenylhydrazone. Talanta*, **63**, 757 (2004)

240. E. Bezemer, S.C. Rutan. *Multivariate curve resolution with non-linear fitting of kinetic profiles. Chemom. Intell. Lab. Syst.*, **59**, 19 (2001)

241. Jr. Workman, J., K.E. Creasy, S. Doherty, L. Bond, M. Koch, A. Ullman, D.J.Veltkamp. Process analytical chemistry. *Anal. Chem.*, **73**, 2705 (2001)

242. S.P. Gurden, E.B. Martin, A.J. Morris. *The introduction of process chemometrics into an industrial pilot plant laboratory. Chemom. Intell. Lab. Syst.*, **44**, 319 (1998)

243. *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

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

245. A. Comte. *Cours de philosophie positive*, 1830, Paris