Учреждение Российской академии наук
Институт химической физики им Н.Н. Семенова РАН

На правах рукописи

Родионова Оксана Евгеньевна

Интервальный метод обработки результатов многоканальных экспериментов

01.04.01 – Приборы и методы экспериментальной физики

Диссертация

на соискание ученой степени

доктора физико-математических наук

Москва 2008

Оглавление

Введение

Работа посвящена разработке нового подхода, объединяющего современные проекционные методы и метод простого интервального оценивания, применяемого при решении важных теоретических и практических задач интерпретации результатов многоканальных экспериментов. Показано, что подобный подход позволяет обрабатывать сложные наборы экспериментальных данных, пронизанных внутренними связями.

Описание экспериментальных данных, построение модели и предсказание новых значений – это одна из старейших, но вечно актуальных задач, которая активно применяется при исследовании различных физических и химических явлений. Традиционно, математические модели строились так, чтобы в математической форме выразить те или иные законы химии и физики. Однако, с совершенствованием и усложнением эксперимента, появилась необходимость анализа очень больших массивов данных. В то же время всегда существовала необходимость моделирования, хотя бы в ограниченной области, таких процессов и зависимостей которые не поддаются содержательному математическому описанию из-за сложности происходящих процессов или их неизученности. Это привело к потребности применения формальных методов моделирования и породило новую область, называемую хемометрикой. Она появилась осенью 1974 года, в городе Сиэтле, США [1]. У ее истоков стояли два человека: американец Брюс Ковальски (B. Kowalski) и швед Сванте Волд (S. Wold) – внук Сванте Аррениуса (S. Arrhenius). Как это часто бывает с подобными дисциплинами, хемометрика до сих пор не имеет общепризнанного определения. Наиболее популярное определение принадлежит Д. Массарту (D. Massart), [2] который считал, что *хемометрика – это 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 практики. 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 накопленных данных по-прежнему остается крайне важной, как с практической, так и с теоретической точки зрения, однако сейчас становится очевидным, что не менее важной

является и задача конструирования таких экспериментов, которые могут предоставить 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 [3]. Хемометрика решает следующие задачи: *как получить физически/химически важную информацию из экспериментальных данных, как организовать и представить эту информацию, и как получить данные, содержащую такую информацию*.

То, что формальные методы многомерного анализа больших массивов экспериментальных данных, или 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 многооткликовых и многофакторных экспериментов. В свою очередь, это 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 исследователю. Это подвигло их активно применять математические методы для извлечения такой информации и для подтверждения того, что сделанные при этом выводы достоверны. В результате такого взаимодействия был достигнут первый несомненный успех. 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 [4]. Первые работы по хемометрике были посвящены методам анализа спектроскопических данных [5-7], построению для них калибровочных моделей с помощью метода главных компонент [8] и метода проекций на латентные структуры [9].

Говоря об истории развития методов многомерного анализа данных, 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 [11] на пивоварне Гиннеса, где он работал аналитиком. В начале 20 века появилась работа К. Пирсона (K. Pearson) [10], 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 [12], а также пионерских работ [13] по планированию эксперимента. Среди советских ученых следует отметить, прежде всего, В. Налимова, внесшего значительный вклад в теорию планирования эксперимента [14].

Хемометрика зародилась, и длительное время развивалась внутри аналитической химии. Однако со временем обнаружилась тенденция, которую некоторые исследователи расценили, как выход хемометрики из-под крыла аналитической химии и превращение ее в самостоятельную дисциплину. Два обстоятельства дали повод к такому выводу. Во-первых, это усложнение математического аппарата, используемого при анализе многофакторных экспериментов. Десять лет назад экспериментаторы смогли усвоить и принять многомерный подход к анализу данных, т.е. такие методы как проекция на латентные структуры (ПЛС) [15] или разложение по сингулярным значениям (SVD) [16]. Однако потом, в период повального увлечения новыми методами анализа данных: мультимодальным подходом (n-way) [17], вэйвлет-анализом (wavelet) [18], методом опорных векторов (SVM) [19] и т.п., наметился некоторый разрыв между экспериментаторами и теоретиками. Второе обстоятельство, приведшее к отдалению хемометрики от аналитической химии, связано с появлением многочисленных приложений, в которых хемометрический подход с успехом применялся в областях, далеких от аналитической химии. Достаточно вспомнить о гиперспектральном анализе и анализе изображений (MIA) [20], многомерном статистическом контроле процессов (MSPC) [21], а также о многочисленных биофизических и биологических приложениях [22].

Методы многомерного анализа данных тесно связаны с математикой и, в особенности, с математической статистикой, откуда они черпает свои идеи. Большинство экспериментаторов понимают необходимость применения статистики в физическом и 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, и лишь немногие исследователи решаются пойти дальше и действительно использовать хемометрику для анализа своих данных. Большинство

экспериментаторов 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.

Взаимоотношения хемометрики и математики заслуживают отдельного рассмотрения. Многие методы и алгоритмы, популярные в хемометрике, не вызывают восторга у математиков [24], 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 [25], хемометрики изобретают все новые и новые методы их анализа. Делают они это так быстро, что математики, по словам американского статистика Д. Фридмана (J. Friedman), не успевают не только раскритиковать их за это, но и просто понять, что же происходит в этой области. Такой подход контрастирует с ситуацией, сложившейся в биометрике [26], 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 дискриминантный анализ. С другой стороны, специалисты, работающие в другой близкой дисциплине – психометрике [27], традиционно активно разрабатывали новые подходы к анализу данных. Так, самый популярный в хемометрике метод ПЛС, был изобретен Г. Волдом (H. Wold) [28] именно для применения в этой области. Забавно, 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 методы отыскания эффективных комбинаций переменных*» [29, 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 [30], в органической химии для предсказания активности соединений по их структуре (QSAR) [31], в химии полимеров [32], в теоретической и квантовой химии [33]. Хемометрика используется в самых разнообразных областях – от пивоварения [34], до астрономии [35]. Она применяется для решения судебных споров о защите окружающей среды [36] и для контроля качества производства полупроводников [37]. Подробный анализ взаимодействий хемометрики с различными областями человеческой деятельности приведен в книге английского аналитика Р. Бреретона (R. Brereton) [38].

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 [39]. Позднее появились работы Л. Грибова [40] и М. Эляшберга по спектральным методам [41], Б. Марьянова по титраметрии [42], Б. Дерендяева и В. Вершинина по методам компьютерной идентификации органических соединений [43], И. Зенкевича по хроматографии [44]. Исследования в близкой к хемометрике области QSAR ведутся под руководством Н. Зефирова [45]. Метрологические аспекты и контроль качества химического анализа исследуются в работах В. Дворкина [46] и Ю. Карпова [47]. В С.-Петербургском университете группа ученых под руководством Ю. Власова работает над созданием сенсорных систем, известных под названием «электронный язык» [48], а в Воронеже разрабатываются аналогичные методы, известные как «электронный нос» [49]. 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 [50, 51]. За последние годы в России появились новые группы ученых, разрабатывающих и применяющих хемометрические подходы: в Москве [52-55], в Барнауле [56, 57], в Томске [58], в Иркутске [59].

Информационное и программное обеспечение. Единственная широко-известная в России книга по хемометрике была переведена и опубликована 20 лет назад [60]. Она ярко отражала положение дел в этой области, сложившееся в середине 80-х годов. На сегодняшний день наиболее полным изложением хемометрических методов является двухтомник, написанный группой авторов под руководством Д. Массарта [61, 62]. Он включает подробное описание основных 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 [38]; исследователям, занимающимся, в основном, спектральным анализом, будут понятнее книги [63, 64]. Для практического применения очень полезна книга [65]. Также нельзя не упомянуть знаменитую книгу Е. Малиновского (E. Malinowski) [66], которую до сих пор многие экспериментаторы считают лучшим учебником в этой области. Теоретические основы хемометрики были изложены в работах [67, 68]. Недавно на русский язык был переведен учебник [69], содержащий краткое описание хемометрики в одной из своих глав. Небольшое, но очень полезное введение в хемометрику написал Б. Марьянов [70]. Маленьким тиражом (для участников четырех конференций по хемометрики в России) был издан сокращенный перевод самого популярного в мире учебника по многомерному анализу данных, написанного К. Эсбенсеном (K. Esbensen) [71].

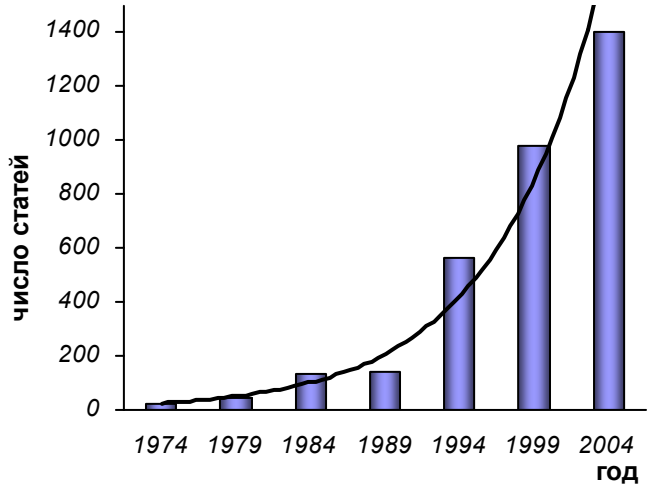


Рис 1. Число статей по хемометрике, опубликованных в журналах издательства Elsevier

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 Vibrational Spectroscopy, Analytica Chimica Acta, Computers and Chemical Engineering, Analyst, Talanta, 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.) [72] В качестве программного обеспечения применяются специализированные пакеты программ [73-75], позволяющие наглядно и быстро обрабатывать данные в интерактивном режиме. Однако широко применяются и статистические пакеты общего назначения [76, 77]. Часто исследователи пишут

процедуры сами, например, в кодах MATLAB [78], и они публикуются для свободного применения, например [68].

Широкое распространение и применение методов многомерного анализа данных в первую очередь обусловлено тем, что главной своей целью этот подход видит в решении конкретных, в основном физических и химических задач, а потом находит уже существующие или разрабатывает новые математические и статистические приемы и алгоритмы.

Эта работа выполнялась в тесном сотрудничестве со многими коллегами. Большой вклад в разработку алгоритмов и написание программ, так же в обсуждение результатов работы внес А.Л. Померанцев (ИХФ РАН), оригинальная физико-химическая моделей для ДСК эксперимента была предложена Е.В. Быстрицкой (ИХФ РАН). Большое влияние на проведение работ в области хемометрики оказал K.H. Esbensen (Aalborg University Esbjerg). Экспериментальные данные, используемые в работе, были получены А.А. Крючковым (НИИКП), L. P. Houmøller и K.H. Esbensen (Aalborg University Esbjerg).

Многомерные данные и формальные модели

В этом разделе в систематическом виде вводятся основные понятия и объекты, с которыми работает исследователь при анализе результатов физического эксперимента с помощью хемометрических методов. Материал этого раздела представлен в работах [72, 79].

Обозначения и термины. В работе используются следующие обозначения. Скалярные переменные выделяются курсивом, например *s*. Векторы (столбцы) обозначаются прямыми жирными строчными буквами, например **x**, а матрицы – заглавными, например **W**. Мультимодальные матрицы еще и выделяются курсивом, например ***G***. Элементы массивов обозначают той же, но строчной буквой, например w_{ij} – это элемент матрицы **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 k=1,..., K. Операция транспонирования обозначается верхним индексом t, например $\mathbf{X}^t$. В описании алгоритмов, верхний индекс, так же обозначает номер итерации, например $\mathbf{p}^m$, обозначает вектор **p**, вычисленный на *m*-ой итерации.

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* используется термин *N-модальный*. Может быть, это и не лучшее решение, но применение традиционного термина тензорного анализа «валентность» в физическом/химическом контексте, будет неудачным. Во многих случаях переводчики просто избегали давать русские названия ключевым хемометрическим понятиям, таким как *scores* и *loadings*, используя вместо них сложные

эвфемизмы. Однако в хемометрике невозможно обойтись без понятий *счета* и *нагрузки*, или их аналогов.

Хемометрика – это наука о сокращениях. В данном случае имеется в виду не понижение размерности данных, а то, что в хемометрике часто используются аббревиатуры: PCA, PLS, PCR, RMSEP и т. п. У некоторых из них есть общепринятые русские аналоги, которые и используются в дальнейшем в тексте. Например, PCA – это МГК (метод главных компонент), PCR – это РГК (регрессия на главные компоненты), PLS – ПЛС (проекция на латентные структуры). А так же имеется множество сокращений, не имеющих устоявшихся русских аналогов. В тексте эти аббревиатуры используются в английской нотации. Ниже они расшифрованы.

ALS (alternating least-squares) – чередующиеся наименьшие квадраты;

ANN (artificial neural network) – искусственная нейронная сеть;

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

EFA (evolving factor analysis) – эволюционный факторный анализ;

GA (genetic algorithm) – генетический алгоритм;

IA (immune algorithm) – иммунный алгоритм;

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

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

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

LOO (leave one out) – метод перекрестной проверки с исключением по одному объекту;

MIA (multivariate image analysis) – многомерный анализ изображений;

MSC (multiplicative signal correction или multiplicative scatter correction) – множественная коррекция сигнала или мультипликативная коррекция рассеяния;

MSPC (multivariate statistical process control) – многомерный статистический контроль процессов;

NAS (net analyte signal) – полезный аналитический сигнал;

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

OSC (orthogonal signal correction) – ортогональная коррекция сигнала;

PARAFAC (parallel factor analysis) – параллельный факторный анализ;

PAT (process analytical technology) – методы анализа процессов;

PC (principal component) – главная компонента (ГК);

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

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

QPLS (quadratic PLS) – квадратичный PLS;

QSAR (qualitative structure-activity relationship) – количественная связь структура-активность;

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

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

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

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

SIMPLS (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.

1. Модели, используемые при обработке экспериментальных результатов

1.1. Экспериментальные данные и информация

Экспериментальные данные – это основной объект, рассматриваемый в работе. . Следуя классификации, предложенной в работе [81], рассмотрим типичное устройство данных физических экспериментов (см.Рис. 1.1).

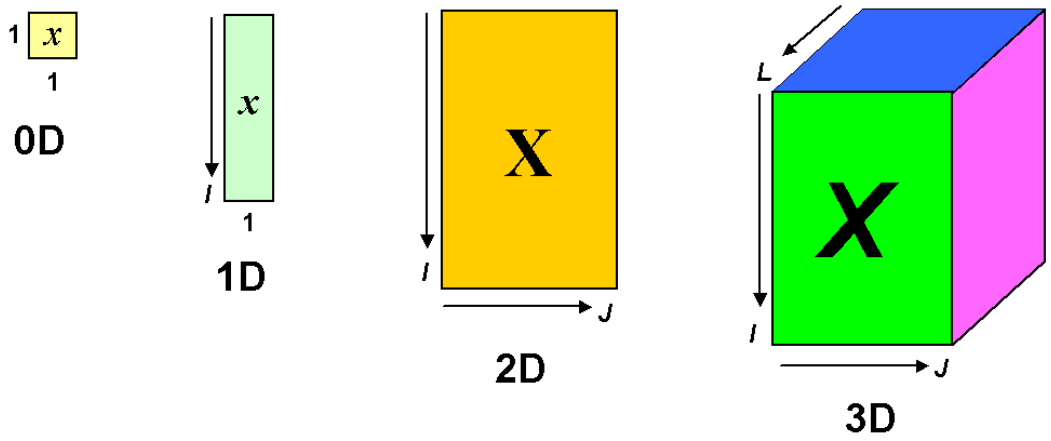


Рис. 1.1 Графическое представление данных разной модальности

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 объекту. 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 физических 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 объектов 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, можно составить матрицу из 28

строк и 30 столбцов, а можно, наоборот, считать длины волн переменными, а времена удержания – объектами. В большинстве случаев, в эволюционных (то есть, развивающихся во времени) физических экспериментах, объекты соответствуют временам, то есть образец, меняющийся во времени, рассматривается как серия объектов [38].

С интенсивным развитием гибридных методов [69] большое внимание уделяется *трех (и более) модальным данным* [82]. Их можно представить в виде параллелепипеда (3D матрицы), в котором каждое ребро соответствует своему типу переменной. В разделе 3.2 подробно рассматривается 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 [82].

Результаты физических экспериментов, или данные для анализа, могут объединяться в блоки. Простейший случай – это один блок **X**. Такой случай чаще встречается в качественном анализе, например, в задаче разделения спектров и концентраций. Количественный анализ, основанный на регрессионных зависимостях, использует данные, состоящие из двух и более блоков. Блок предикторов (например, 2D-матрица спектров **X**) и блок откликов (например, 1D- вектор концентраций **y**) составляют набор стандартных данных, по которым строится калибровочная модель **y=Xb**. Встречаются данные и более сложной структуры, объединяющей три и более блоков данных [84]. Для их анализа применяются специальные методы, называемые маршрутным моделированием (path modeling) [85]. Может показаться, что такая систематизация данных: размерность, модальность, блочность, носит несколько формальный характер и может представлять интерес только для математиков, но не для экспериментаторов. 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 физического эксперимента, большой может считаться матрица с 1 000 000 столбцов и 400 000 строками [25]. При обработке таких массивов, их приходится разделять на блоки и интерпретировать по очереди. Это

разделение нельзя проводить формально, тут обязательно требуется участие опытного физика , 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 методов анализа, таких как гиперспектральные данные [86] и микрочипы [87], сложность данных будет только нарастать.

Основная задача хемометрики состоит в извлечении из данных нужной физической/химической *информации*. Понятие информации является ключевым, поэтому на нем следует остановиться подробнее. 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), который скрывает 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

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

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 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 – *использование многомерного подхода* при конструировании физических 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 важнейший принцип.

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 многомерном 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 связи [88]. Понятие эффективного ранга и скрытых, *латентных переменных*, число которых равно этому рангу, является третьим важнейшим принципом [66]. 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 объектов – 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 физических 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 , мультколлинеарность может негативно сказаться на качестве анализа. Так, например, применение множественной линейной регрессии в условиях мультиколлинеарности совершенно неприемлемо [71]. Для регрессионного анализа таких данных надо применять специальные методы, например ридж-регресию [89], или проекционные подходы [68].

Существенным источником шума в данных может быть отбор образцов. Теория *пробоотбора*, значительный вклад в которую внес П. Жи (P. Gy) [90], приобрела большую популярность в последнее время [91]. Многочисленные приложения можно найти в специальном выпуске [92], полностью посвященном этой теме. Другая проблема, с которой может столкнуться исследователь – это *пропуски в данных* [93]. Это может случиться по разным причинам: отказы приборов, выход за пределы обнаружения, нехватка объектов 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, основанный на методе максимума правдоподобия [94]. Детали таких алгоритмов в большой степени зависят от того, какая модель используется для описания данных.

1.2. Модели и методы

Рассмотрев устройство данных физического эксперимента, перейдем к методам их анализа. Далее основные методы будут описаны более подробно, а этот раздел 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, калибровка. Методы первой группы оперируют, как правило, с одним блоком данных, а в калибровке 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 [95], результаты физического эксперимента 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 [96] 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 [32], 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 [97], объединяющие сильные

стороны обоих методов. Проиллюстрируем разные подходы к моделированию примерами.

Важными объектами математического моделирования 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 [98]. С другой стороны, в работе [99] замечено, что по своей форме титриметрические кривые напоминают обратные гиперболические и тригонометрические функции. Исходя из этого, предлагается использовать формальные, «черные» зависимости, составленные из функций arcsin, arccos и т.п. Компромиссный, «серый» подход предложен в работе [42], где заменой переменных содержательная модель преобразуется в кусочно-линейную. Затем для оценки параметров применяется метод *чередующихся наименьших квадратов* (ALS) [100], 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 [101] 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 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 представительным. Первый набор, называемый обучающим или калибровочным, используется для идентификации модели, т.е. для оценки ее параметров. Второй набор, называемый проверочным или тестовым 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 с помощью проверочного набора (методом *тест-валидации*). В некоторых случаях объем данных слишком мал для такой проверки. Тогда применяют другой метод – перекрестной проверки (*кросс-валидация*) [102]. В этом методе проверочные значения вычисляют с помощью следующей процедуры. Некоторую фиксированную долю (например, первые 10% объектов) 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 [71]. 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 данных физического эксперимента, 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 [103]. Неопределенность в

количественном анализе характеризуется либо числом – стандартным отклонением [104], либо интервалом – доверительным [105] или прогнозным [52]. В качественном анализе применяется метод проверки статистических гипотез [106], в котором неопределенность характеризуется через вероятность принятия неверного решения [107]. Методы оценки неопределенности при моделировании многомерных [108] и многомодальных [109] данных вызывают большой интерес исследователей. Для описания различных аспектов надежности метода применяются специальные характеристики: специфичность, селективность, предел обнаружения, отношение сигнал/шум [70]. Актуальным методом их определения является подход с использованием концепции [111] *полезного аналитического сигнала* (NAS). Многомерный вектор NAS определяется как та часть полного сигнала (спектра), которая используется для моделирования и прогноза [112]. Оставшаяся часть сигнала, включающая погрешности, вклады от посторонних компонентов, рассматривается как шум. Концепции NAS была применена к задаче определения предела обнаружения при анализе двух- [113] и трехмодальных [114] данных. Полученные результаты нашли многочисленные практические приложения, одно из которых рассмотрено в разделе 3.2.

Надежность метода сильно зависит от того, какие данные были использованы для построения и проверки соответствующей модели. Наличие выбросов [115] или малоинформативных данных снижает точность модели, и наоборот, присутствие представительных, влиятельных объектов в эксперименте [116] существенно улучшает качество модели. Оценка влиятельности данных может проводиться классическими регрессионными методами [117], а может выполняться с помощью нестатистических процедур [52]. При использовании построенной модели для определения интересующих нас показателей, мы сталкиваемся с похожими проблемами. Может оказаться, что метод не применим к некоторым объектам (выброс в прогнозе [118]) или дает очень неточный результат. Оценка неопределенности метода не в среднем [119], а для индивидуальных объектов – это сложная задача, над решением которой работают сейчас разные группы исследователей [120]. Именно их усилия определяют успешное решение таких практически важных задач как перенос калибровок с одного прибора на другой [121], отбор переменных [122], построение робастных [123] методов анализа данных.

1.3. Подготовка данных и обработка сигналов

Важным условием правильного моделирования и, соответственно, успешного физико-химического анализа, является *предварительная подготовка* данных, которая включает различные преобразования исходных, «сырых» экспериментальных значений. Простейшими преобразованиями является центрирование и нормирование [124].

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**=($m_1$**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. Оно понижает эффективный ранг модели на единицу и может улучшать точность описания. Это преобразование можно рассматривать как проецирование на нулевую главную компоненту [9], поэтому оно всегда применяется в методах ПЛС и МГК. 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** справа на матрицу **W**, т.е. $\widetilde{\mathbf{X}} = \mathbf{XW}$. Матрица **W** – это диагональная матрица размерности *J*×*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** слева на диагональную матрицу **W,** т.е. $\widetilde{\mathbf{X}} = \mathbf{WX}$. При этом размерность **W** равна *I*×*I*, а ее элементы w_{ii} – это обратные значения стандартных отклонений строк $\mathbf{x}_i^t$. Комбинация центрирования и нормирования по столбцам $\widetilde{x}_{ij} = \left(x_{ij} - m_j\right) \Big/ d_j$ 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 [68]. В тоже время, к этому преобразованию нужно относится с большой осторожностью, т.к. оно может сильно исказить результаты качественного анализа [38].

Помимо этих линейных преобразований используются и нелинейные трансформации результатов эксперимента. Так, в БИК спектроскопии часто применяется преобразование Кубелки-Мунка (Kubelka-Munck) [125]. Цель этого и других трансформаций, например, преобразования Бокса-Кокса (Box-Cox) [29] – линеаризация модели. Часто простые операции с данными – логарифмирование [52], извлечение корня [32] помогают существенно улучшить модель.

Результаты физического эксперимента 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) [38, 126]. 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 рассмотрим подробнее. Метод *множественной коррекции сигнала*, называемый также мультипликативной коррекцией рассеяния (MSC) [67] был первоначально разработан [127] для БИК спектроскопии и базировался на идеях работы [125]. Процедура MSC преобразования устроена очень просто. Сначала определяется «базовый спектр» $\mathbf{m}^{\mathrm{t}} = \left(\mathbf{x}_1^{\mathrm{t}} + \ldots + \mathbf{x}_I^{\mathrm{t}}\right)/I$ 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 $\tilde{\mathbf{x}}_i^{\mathrm{t}} = \left(\mathbf{x}_i^{\mathrm{t}} - a_i\right)/b_i$. Параметры множественной коррекции a_i и b_i могут определяться не по всем переменным, а только по некоторому (скользящему) окну [128].

Второй метод, а точнее группа методов, называемых *ортогональной коррекцией сигнала* (OSC) [129] 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 ПЛС компонент K, необходимых для моделирования данных. Существует много вариантов этого метода, первоначально предложенного в [130], и развитого в работах [131, 132]. 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 ПЛС. Подробное изложение метода и MATLAB код приведен в [132].

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 [133], и количественном анализе [134]. Для отбора переменных применяются различные методы: генетический алгоритм [135], оптимизация Парето (Pareto) [136], «складного ножа» (jack-knife) [122]. Особую важность отбор переменных приобретает в тех методах, где аналитический сигнал непрерывно зависит от канала, например в спектрометрии [137]. Здесь отбор переменных осуществляется целыми блоками, как в методе [132], или в работе [138]. Помимо отбора переменных используется и *отбор объектов*, т.е. строк $\mathbf{x}_i^t$ в матрице **X** (как и соответствующих им значений в матрице откликов **Y)**. Отбор объектов также позволяет улучшить качество модели, но особенно он важен для обнаружения выбросов [118], при переносе калибровок с одного прибора на другой [121, 139, 140]. Новый подход к классификации и отбору объектов изложен в работе [52].

Обработка сигналов. Обработка сигналов с помощью различных преобразований и фильтров играет важную роль в физическом анализе [141, 142]. 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 сделать 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 [143]. 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 свертки спектры дают частотно-временное представление с разным разрешением, зависящим от ширины диапазона. Вэйвлет анализ часто предшествует методам моделирования, что позволяет применять их к очень большим наборам данных без потери информации [144] .Этот метод часто применяется для сжатия и сглаживания одно и двухмодальных ИК и ЯМР спектров [145].

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 исследователей. 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 [146], еще встречаются.

1.4. Результат главы 1

В этой главе введена формальная классификация данных независимо от содержания физической информации, которая в них заложена. Данная классификация необходима для правильного подбора математических методов обработки экспериментальных данных.

Важным является разделение моделей для описания данных на формальные, «черные» и содержательные, «белые» в зависимости от того, что является источником для построения той или иной модели. Традиционное содержательное моделирование, основанное на знание тех или иных физических и химических законах, в настоящее время зачастую уступает место формальному, эмпирическому моделированию, из-за сложности эксперимента и большого объема получаемых физических данных. В результате, происходит с одной стороны чрезмерное усложнения содержательной модели, а с другой стороны для построения такой модели приходится прибегать к многочисленным допущениям и упрощениям, что приводит временами к моделированию некоторой идеализированной ситуации, далекой от условий реального эксперимента. Однако, при применении формальных моделей, работа по математической обработке экспериментальных данных ни в коем случае не превращается в формально-математические упражнения. Понимание физической сути исследования с необходимостью используется при конструировании эксперимента, а так же при интерпретации результатов моделирования.

В последнем разделе этой главы рассмотрены различные методы предварительной подготовки данных. Такие преобразования данных, в основном, зависят от применяемого экспериментального оборудования, и во многих случаях помогаю существенно улучшить качество последующего моделирования.

2. Методы качественного анализа: исследование, классификация и дискриминация

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, при многомерном подходе, 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 физического эксперимента, используя новые скрытые переменные. При этом должны выполняться два условия. Во-первых, число новых переменных (эффективный 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.

В этой главе подробно рассмотрен метод главных компонент (МГК), который лежит в основе целого класса проекционных методов, используемых для сжатия исходных данных. Так же рассмотрены методы классификации и дискриминации, применяемые для качественного анализа данных.

2.1. Метод главных компонент

Наиболее популярным способом сжатия данных является *метод главных компонент* (МГК) [15]. Он дает основу для других аналогичных проекционных методов, включая эволюционный факторный анализ (EFA) [147], оконный факторный анализ (WFA) [148], итерационный целевой факторный анализ (ITTFA) [149], а также многих методов классификации, например, *формального независимого моделирования аналогий классов* (SIMCA) [150]. 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** [71]

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E} = \sum_{k=1}^{K} \mathbf{t}_{\mathrm{k}} \mathbf{p}_{\mathrm{k}}^{\mathrm{t}} + \mathbf{E} \qquad (2.1)$$

В этом уравнении **T** называется матрицей *счетов* (scores) , **P** – матрицей *нагрузок* (loadings), а **E** – матрицей остатков (См.Рис. 2.1). Число столбцов – $\mathbf{t}_k$ в матрице **T** и $\mathbf{p}_k$ в матрице **P** – равно эффективному рангу матрицы **X**. Эта величина *K* называется *числом главных компонент* (ГК) и она, естественно, меньше, чем число столбцов в матрице **X**.

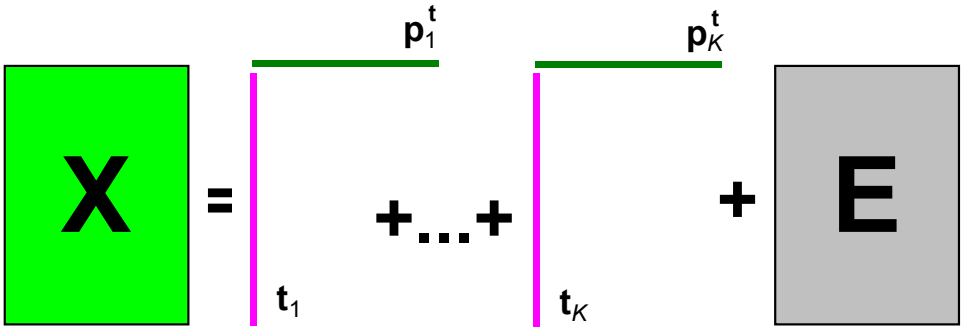


Рис. 2.1. Графическое представление метода главных компонент

Для иллюстрации МГК, мы опять вернемся к примеру, рассмотренному в конце раздела 1.1. Матрица спектров смесей **X** может быть представлена как произведение матрицы концентраций **C** и матрицы спектров чистых компонентов **S**

$$\mathbf{X}=\mathbf{C}\mathbf{S}^t+\mathbf{E} \qquad (2.2)$$

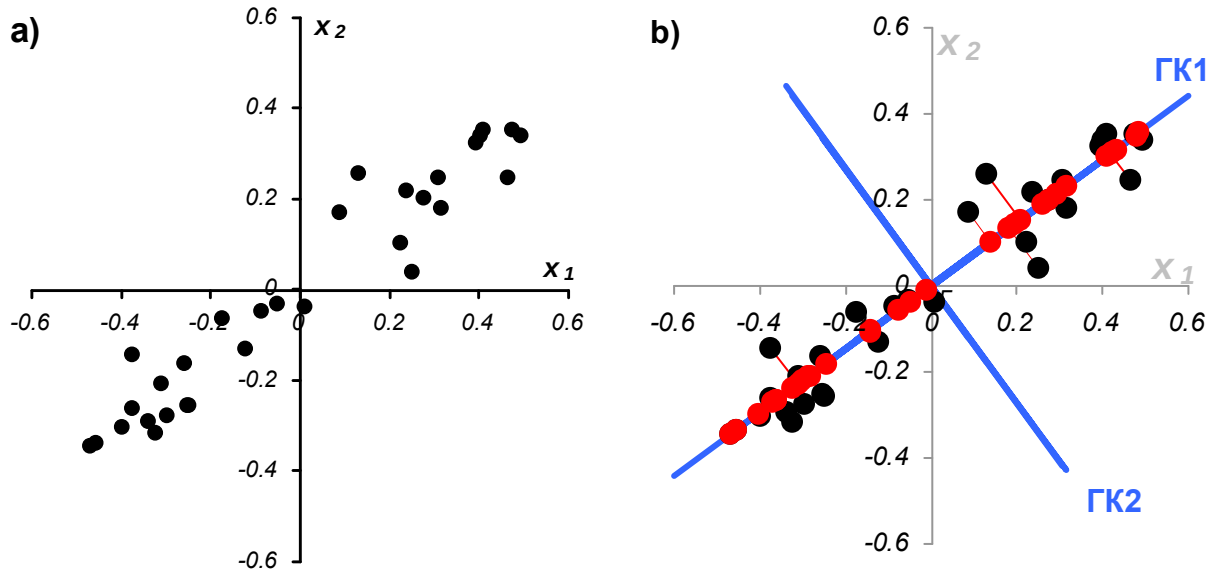
Число строк в матрице **X** равно числу объектов (*I*), и каждая ее строка соответствует спектру одного объекта, снятому для *J* длин волн. Число строк в матрице **C** также равно *I*, а вот число столбцов соответствует числу компонентов в смеси (*K*=3). Матрица чистых спектров присутствует в разложении (2.2) в транспонированном виде, т.к. количество ее строк равно числу длин волн (*J*), а число столбцов равно *K*. Как уже отмечалось выше, при анализе реальных экспериментальных данных, отягощенных погрешностями, представленными матрицей **E**, эффективный ранг *K* 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) [151]. В этой области можно выделить два направления. Первое использует метод автомодельного разрешения кривых [152] (SMCR) и оно ориентированно, прежде всего, на приложение к гибридной хроматографии [153]. Для анализа применяются методы формального моделирования (PCA, EFA), которые не

используют содержательное знание об исследуемой системе. В рамках этого подхода можно отметить метод SIMPLISMA [154], применяющий простой, но весьма эффективный подход, основанный на отборе переменных [155]. Второе направление, напротив, учитывает априорную информацию о процессах, применяя «серые» модели [156]. Это направление находит свое приложение при исследовании кинетики [30] и термодинамики [157] . Ключевым моментом в таких задачах является определение величины эффективного ранга системы – числа главных компонент K [158]. В идеале предсказанные спектры **S** и концентрации **C** близки к истинным значениям, хотя их никогда невозможно восстановить точно. Причина этого не только в погрешностях эксперимента, но и в том, что спектры могут частично перекрываться. Когда МГК применяется для разделения данных на физически осмысленные компоненты, как в уравнении (2.2), 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 физико-химических данных. В этом случае предполагается, что в исследуемой системе имеются скрытые, латентные переменные, влияющие на поведение системы. Число латентных переменных, K, существенно меньше числа исходных переменных J. При этом латентные переменные являются линейной комбинацией исходных переменных. В этом случае матрицы счетов **T** и нагрузок **P** уже нельзя интерпретировать как спектры и концентрации, а число главных компонент K – 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 МГК. На Рис. 2.2a 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 (ГК1) 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 (ГК2) соответствует наибольшему изменению в остатках, показанных на Рис. 2.2b отрезками, перпендикулярными оси $\mathbf{p}_1$.



Данные в исходных координатах Данные в координатах главных компонент

Рис. 2.2. Графическая иллюстрация метода главных компонент

Этот тривиальный пример показывает, что метод главных компонент осуществляется последовательно, шаг за шагом. На каждом шаге исследуются остатки $\mathbf{E}_k$, среди них выбирается направление наибольшего изменения, данные проецируются на эту ось, вычисляются новые остатки, и т. д. Этот алгоритм называется NIPALS [71]. Другой популярный алгоритм сжатия данных – разложение по сингулярным значениям (SVD)[159] – строит ту же декомпозицию (2.1) без итераций. Остановка итерационной процедуры, или, другими словами, выбор числа главных компонент K, проводится с использованием критериев, показывающих точность достигнутой декомпозиции. Пусть исходная матрица $\mathbf{X}$ имеет размер: I строк и J столбцов, и в разложении (2.1) участвуют K главных компонент. Величины

$$\mu_k = 100 \sum_{i=1}^{I} \mathbf{t}_{ik}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2, \qquad E_k = 100\left(1 - \sum_{i=1}^{I}\sum_{j=1}^{J} e_{ij}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2\right), \qquad k = 1,\ldots,K \tag{2.3}$$

называются нормированным *собственным значением* и *объясненной вариацией*. Их обычно изображают на графике в зависимости от числа k, тогда резкое изменение величин (2.3) указывает на нужное значение числа главных компонент. Для правильного выбора K, необходимо либо использовать проверочный набор, либо метод перекрестной проверки, так как это описано в разделе 1.2. Уравнения (2.1) не содержат в себе свободного члена, поэтому для декомпозиции данных их следует сначала отцентрировать и, иногда, нормировать (см. раздел 1.3).

Метод главных компонент можно интерпретировать как проецирование данных на подпространство меньшей размерности. При этом матрицу исходных, экспериментальных данных $\mathbf{X}(I\times J)$ можно трактовать как набор из I точек (строк в матрице **X,** изучаемых объектов) в пространстве размерности J (столбцов матрицы **X**), т.е. $\mathbf{x}_i \in R^J$. Метод главных компонент проецирует точки $\mathbf{x}_i$ на подпространство меньшей размерности K, переходя от исходную матрицу $\mathbf{X}(I\times J)$ к матрице счетов $\mathbf{T}(I\times K)$, где $\mathbf{t}_i \in R^K$ (обычно $K<<J$). Иными словами, для анализа данных, МГК проецирует точки из пространства R^J в пространство R^K . Возникающие при этом остатки **E** рассматриваются как шум, не содержащий значимой физической информации. В этом подпространстве можно ввести меру близости объектов, называемую *расстоянием Махаланобиса* (Mahalanobis) [160], с помощью которой удается решить многие проблемы качественного анализа. Другим мощным инструментом анализа данных в проекционном подпространстве является *прокрустово* (Procrustes) вращение [161].

Одним из важнейших преимуществ проекционных методов, в том числе и МГК, является возможность представить сложные данные физического эксперимента в более простом виде, так чтобы исследователь смог «увидеть» результаты экспериментов в простой графической интерпретации. В этом контексте МГК можно рассматривать как прием, при котором данные заключаются в эллипсоид, и вычисляются главные оси этого эллипсоида. В результате, первые две главные оси (две первые главные компоненты) образуют плоскость, при проекции на которую видны наибольшая вариация в данных. Каждая следующая ось эллипсоида будет меньше или равна предыдущей, и таким образом, в этом направлении вариации в данных будут все менее заметны. В общем случае, если структура данных такова, что в них имеются какие-то группы или кластеры, обычно это видно при исследовании проекций, построенных для первых нескольких компонент. Поэтому, при исследовании данных методом МГК, особое внимание уделяется графикам счетов и нагрузок. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый объект 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 [38] 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 объектов, 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: [71, 162].

Дополнительно, можно привести и алгебраическую интерпретацию МГК, которая полезна как для понимания метода, так и для его численной реализации. С алгебраической точки зрения, МГК можно рассматривать как решение задачи на собственные значения,

$$\mathbf{X}^t\mathbf{X}\mathbf{p}=\lambda\mathbf{p}$$

т.е., для исходной матрицы $\mathbf{X}(I\times J)$ найти K собственных векторов $\mathbf{p}_k$ (векторы нагрузок), соответствующих первым K наибольшим собственным значениям матрицы $\mathbf{X}^t\mathbf{X}$. При этом векторы счетов t_k вычисляется как

$$\mathbf{t}_k=\mathbf{X}\mathbf{p}_k ,$$

т.е. являются линейной комбинацией исходных значений матрицы $\mathbf{X}$, столбцы $\mathbf{t}_k$ ортогональны и

$$\mathbf{t}_k^t\mathbf{t}_k = \lambda_k ,$$

а столбцы матрицы $\mathbf{P}$ ортонормированны, т.е. $\mathbf{P}^t\mathbf{P}=\mathbf{I}$ – единичной матрице. Подробный алгоритм МГК приведен в Приложении (см. раздел 13.1).

Рассмотрим пример практического использования МГК в физико-химическом анализе. В работе [163] 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 диффузного рассеяния $R(\lambda)$ были сняты с помощью прибора Bomem MB160 с приставкой Powder Samplir, в диапазоне 3800–10000 $см^{-1}$ (1069 волновых чисел) без специальной подготовки образцов. Исходные данные были преобразованы как $-\log R$, центрированы и подготовлены процедурой MSC [71], рассмотренной в разделе 1.3. Они показаны на Рис. 2.3

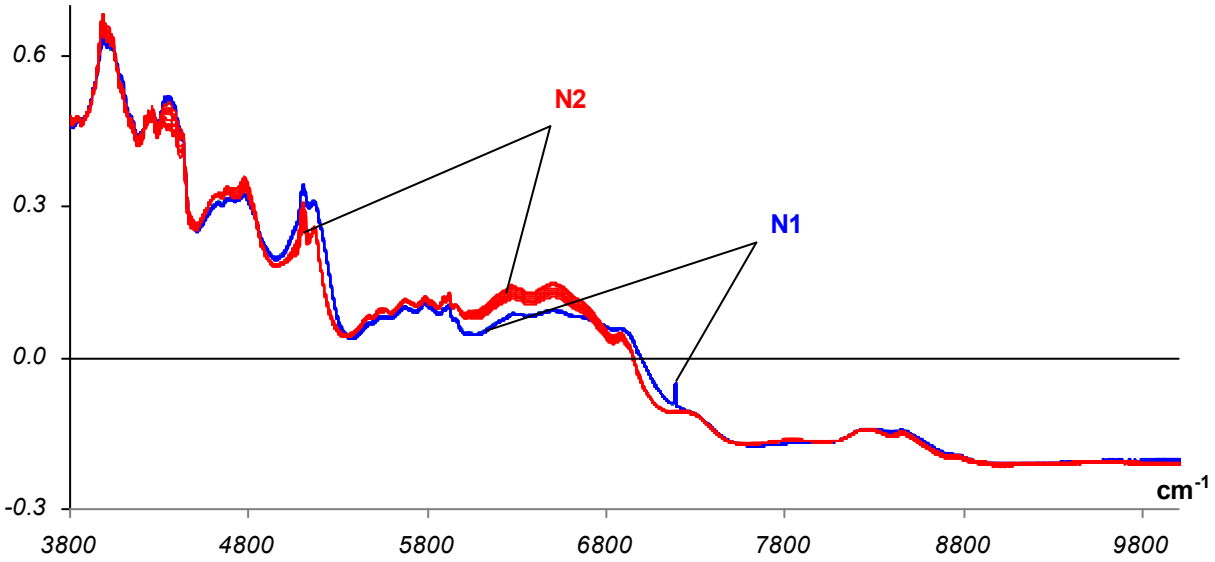
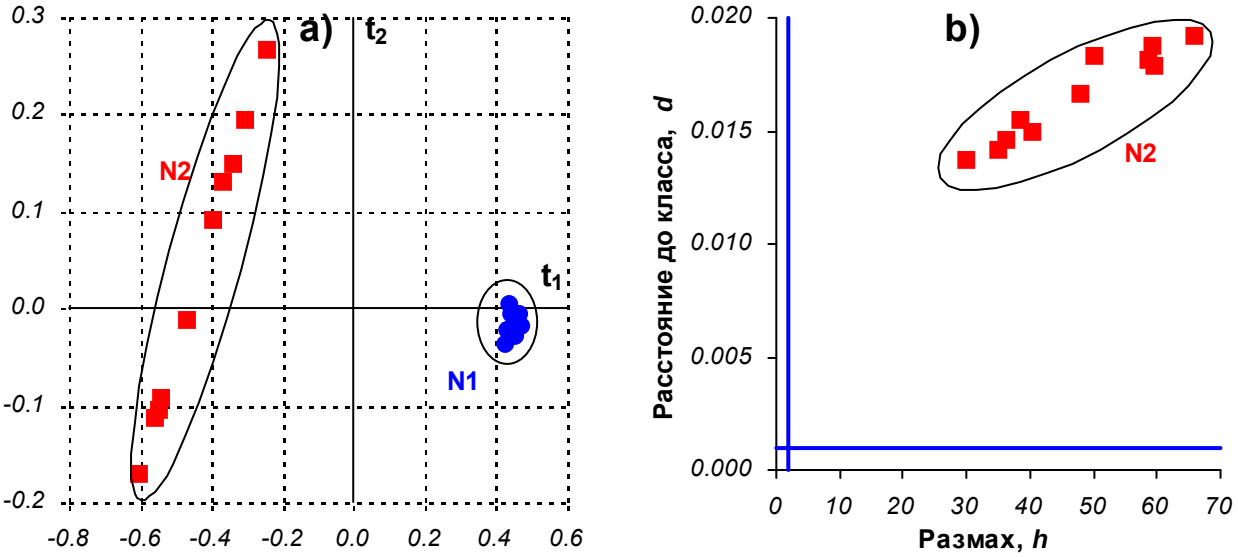


Рис. 2.3. Спектры, преобразованные процедурой MSC. N1 – истинные таблетки, N2 –фальсификат

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



Классификация на графике МГК счетов Классификация методом SIMCA

Рис. 2.4. Определение фальсифицированных лекарств. N1 – истинные таблетки, N2 –фальсификат

На Рис. 2.4а показан график МГК счетов ($\mathbf{t}_1$, $\mathbf{t}_2$) этих спектров. На нем четко видны 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%.

2.2. Классификация и дискриминация

Рассмотренный пример относится к задачам классификации. Это весьма широкий класс задач качественного анализа, в которых требуется установить принадлежность объекта 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 калибровочный набор объектов, про который имеется априорная информация о принадлежности к классам. Методы решения задач классификации без обучения основаны, главным образом, на МГК декомпозиции с последующим анализом [164] расстояний между классами, построением дендрограмм, использованием нечетких множеств [165] и т.п. В работе [166] применялось прокрустово вращение, а в работах [167-169] – расстояние Махаланобиса. Однако, в тех случаях, когда возможно проведение дискриминации, т.е. классификации с обучением, этим методам следует отдавать предпочтение.

Калибровочный набор объектов используется для построения модели классификации, т.е. набора правил, с помощью которых новый объект 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 объектов. В физико-химическом анализе классификация применяется к наборам мультиколлинеарных данных (спектры, хроматограммы), поэтому дискриминантная модель почти всегда многомерна и основана на соответствующих проекционных подходах – МГК, ПЛС. Можно отметить использование линейного дискриминантного анализа в БИК спектроскопии [170], а также канонического дискриминантного анализа [171]. Одним из самых популярных подходов является метод *формального независимого моделирования аналогий классов* (SIMCA [172]), разработанный С. Волдом [150].

В основе метода SIMCA лежит предположение о том, что все объекты в одном классе имеют сходные свойства, но и обладают индивидуальными особенностями. При построении дискриминантной модели необходимо учитывать только сходство, отбрасывая особенности как шум. Для этого каждый класс из калибровочного набора независимо моделируется МГК с разным числом главных компонент K. 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

$$d = \sqrt{\frac{1}{J-A}\sum_{j=1} e_{\mathrm{j}}^2} \tag{2.4}$$

Эта величина сравнивается со среднеквадратичным остатком внутри класса

$$d_0 = \sqrt{\frac{1}{(I-A-1)(J-A)}\sum_{ij} e_{\mathrm{ij}}^2} \tag{2.5}$$

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

$$h = \frac{1}{I} + \sum_{k=1}^{K} \frac{\tau_{\mathrm{k}}^2}{\mathbf{t}_{\mathrm{k}}^{\mathrm{t}}\mathbf{t}_{\mathrm{k}}} \tag{2.6}$$

Здесь τ_{k} – это проекция нового объекта (счет) на главную компоненту k, а $\mathbf{t}_{\mathrm{k}}$ – это вектор, содержащий счета всех обучающих объектов в классе.

На Рис. 2.4b показано применение метода SIMCA для дискриминации таблеток. В качестве класса использовались подлинные таблетки, а на графике показаны расстояния d и h от объектов 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 различимы.

Помимо метода SIMCA, для дискриминации данных физических экспериментов используется также метод DASCO [173], который похож на SIMCA, метод классификация по ближайшему соседу (KNN) [174], метод опорных векторов (SVM) [175-176] и многие другие. Очень мощным инструментом является метод дискриминантного анализа с помощью регрессии на латентные структуры – PLS-DA [177]. Идея этого подхода состоит в том, что правила дискриминации для K классов

задаются линейными регрессионными уравнениями вида **XB**=**D**, где **X** – это полная матрица всех исходных данных (*I*×*J*), **B** – это матрица неизвестных коэффициентов (*J*×*K*), а **D** – это специальная матрица (*I*×*K*), которая состоит из нулей и единиц. При построении матрицы **D** единицы ставят только в те строки (объекты), которые принадлежат классу, соответствующему номеру столбца. Регрессионная задача решается методом ПЛС (см. раздел 3.1), что позволяет в дальнейшем применять построенную регрессию для предсказания принадлежности новых объектов. Для этого строится прогноз отклика нового объекта, и результат сравнивается с нулем или единицей.

2.3. Трехмодальные методы

Метод главных компонент был разработан для данных, имеющих вид двухмодальной 2D-матрицы. Однако, в последнее время, исследователи все чаще имеют дело с трех- и более модальными данными, которые имеют более сложную структуру, например, параллелепипед (Рис. 2.5). Источником таких данных служат, например, гибридные [178, 179] и эволюционные методы [180]. 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 физико-химического анализа, приведено в книге [82]. Краткое введение в методы анализа трехмодальных данных представлено в статье [181]. Эти же алгоритмы используются для обработки данных, полученных в результате гиперспектральных измерений [86], а также для анализа изображений [17].

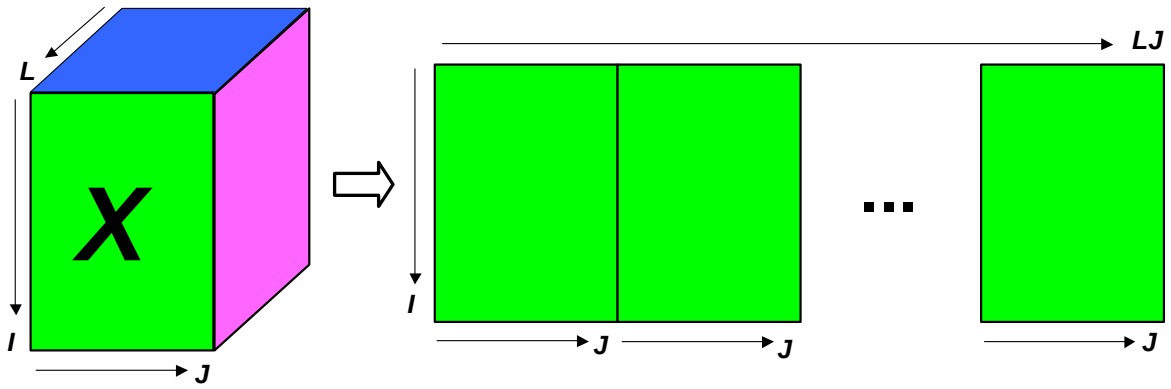


Рис. 2.5. Графическое представление метода развертывания в плоскую матрицу

Метод *развертывания* (unfolding) [182],– это простейший способ анализа трехмодальных данных, с помощью которого 3D-матрица $\boldsymbol{X}$ размерности $I \times J \times L$ разворачивается в обычную 2D-матрицу $^{u}\mathbf{X}$ размерности $I \times JL$ (см. Рис. 2.5). При этом I называется «основной» модой. Далее возможно обычное применение метода главных компонент (см. раздел 2.1). Такой подход часто оказывается эффективным, как, например в [32], 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 $I \times J \times L$ к 2D-матрице $I \times LJ$ уже не учитывается, что измерения x_{ilj} и x_{il+1j} являются соседними, что может быть существенно.

Алгоритм *Tucker3* [183] 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. 2.6. Каждый элемент исходной 3D-матрицы $\boldsymbol{X}$ можно записать в виде суммы (2.7),

$$x_{ijl} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip} b_{jq} c_{lr} g_{pqr} + e_{ijl} \qquad (2.7)$$

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.

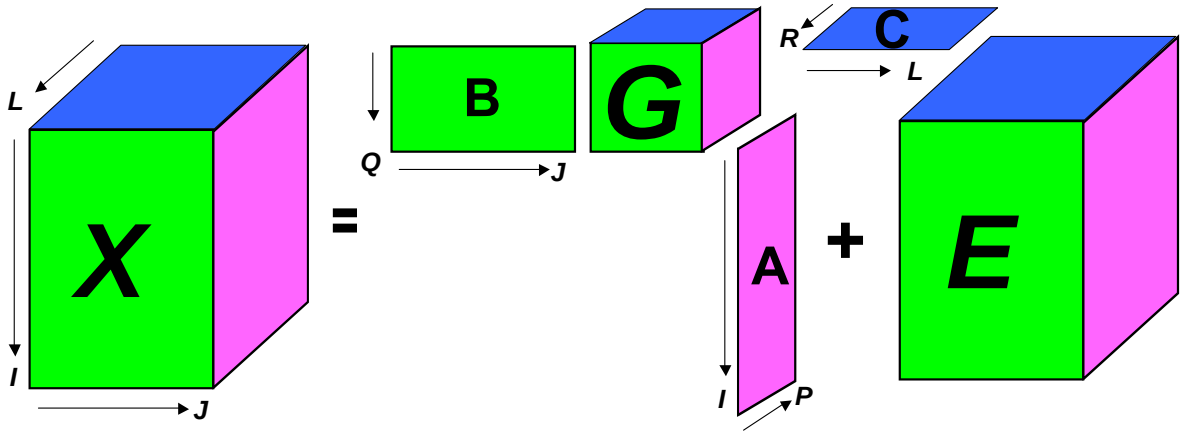


Рис. 2.6. Графическое представление модели Tucker3

Метод *PARAFAC* (parallel factor analysis) [181] 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_{ijl}

$$x_{ijl} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{lr} + e_{ijl} \qquad (2.8)$$

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. 2.7.

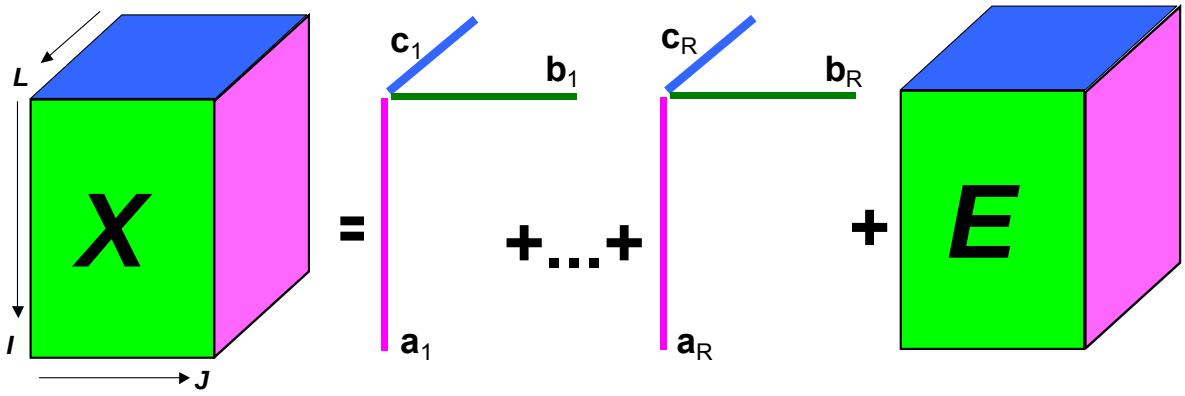


Рис. 2.7. Графическое представление модели PARAFAC с *R* компонентами

MATLAB код алгоритма PARAFAC можно найти в [181]. Так как матрицы нагрузок в разложении (2.8) 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 [183], а алгоритмы всех рассмотренных методов декомпозиции трехмодальных данных приведены в [184].

2.4. Результаты главы 2

Проекционные методы позволяют анализировать результаты многоканальных экспериментов, представляя очень большие наборы данных в компактной и наглядной форме. Они дают возможность выявлять существующие содержательные зависимости, как между переменными, так и между объектами. Такой подход влияет на выбор физического/химического метода исследования. Проекционные методы позволяют широко использовать низкоселективные, но часто более простые в технике эксперимента

методы или более быстрые. Селективность метода повышается за счет соответствующей математической обработки.

Проекционный подход так же позволяет эффективно решать задачи классификации и дискриминации; при этом каждый образец рассматривается в совокупности, со всеми своими свойствами, как единый объект. Скрытые зависимости между отдельными свойствами одного и того же объекта учитываются при математическом анализе всего массива экспериментальных данных.

3. Методы количественного анализа: калибровка

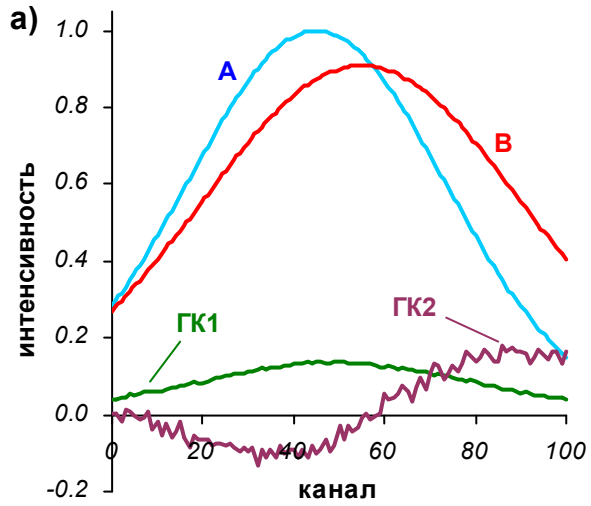
В задачах количественного анализа участвуют два блока данных. Первый блок **X** – это матрица аналитических сигналов (например, спектров, хроматограмм, и т.п.). Второй блок **Y** – это матрица соответствующих показателей (например, концентраций). Число строк (I) в этих матрицах равно количеству измеряемых объектов , число столбцов (J) в матрице **X** соответствует числу каналов (длин волн), на которых записывается сигнал, и, наконец, число столбцов (L) в матрице **Y** равно числу . откликов. Задача калибровки состоит в построении математической модели, связывающей блоки **X** и **Y**, с помощью которой можно в дальнейшем предсказывать значения показателей **y** по новой строке значений сигнала **x** [9].

По виду математических моделей, а, следовательно, и по методам отыскания неизвестных параметров различают линейную и нелинейную калибровку. В этой главе рассмотрены преимущества и недостатки обоих подходов. Так же представлены методы многомодальной калибровки. Применение таких методов, в первую очередь, связано с необходимостью обработки результатов сложных физико-химических экспериментов, например гибридных методов.

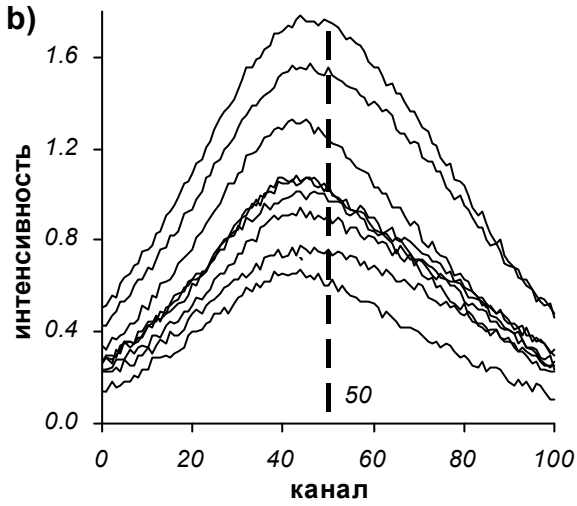
3.1. Линейная калибровка

Простейшая калибровочная модель – это *одномерная регрессия* (J=1, L=1), т.е. модель $y=a+bx$ [185], которая соответствует одному каналу аналитического сигнала. С помощью методов классического регрессионного анализа можно строить более сложную, множественную регрессию ($I>J$, L=1), в которой участвуют несколько каналов, т.е. **y=Xb** [29]. 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* калибровки [186]. 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) [69]. Второй случай называется обратной калибровкой (**Y=C**, **X=S**). Этот подход является сейчас господствующим, поскольку он удобнее с практической точки зрения, т.к. непосредственно предсказывает искомый показатель (концентрацию **C**) по

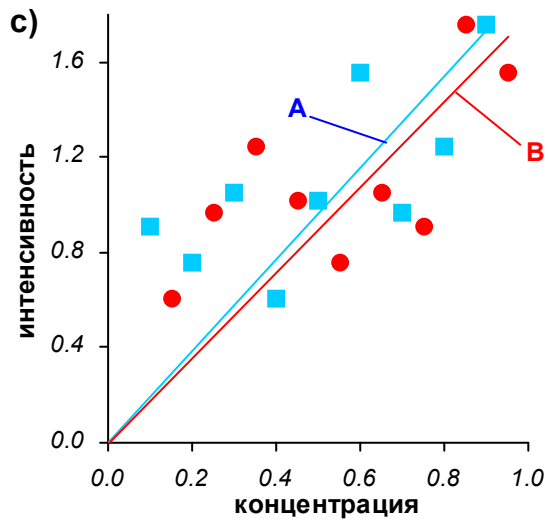
измеренному сигналу (спектру **S**). Кроме того, современные регрессионные методы (РГК, ПЛС) позволяют работать с данными, в которых погрешности присутствуют в обоих блоках.



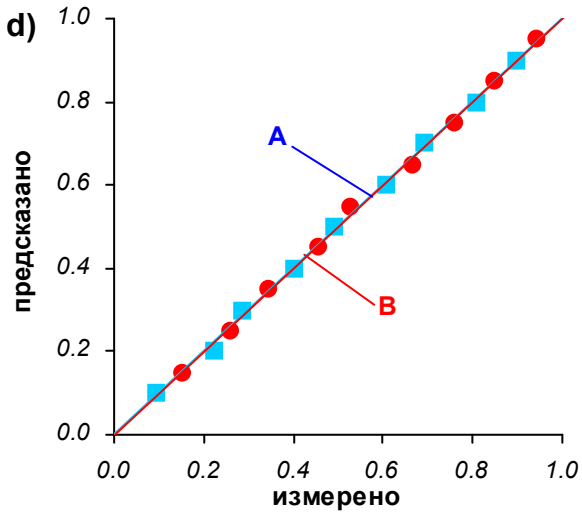
Спектры чистых (A, B) и главных компонент (ГК1, ГК2)



Модельные спектральные данные



Одномерная калибровка.



Калибровка методом PCR

Рис. 3.1. Модельный пример построения различных калибровок. ■ - A, ● - B

Для иллюстрации различных методов калибровки, мы вновь используем пример, введенный в разделе 1.1. Теперь наполним его конкретным содержанием, смоделировав данные **X** и **Y**. Рассмотрим смесь двух веществ A и B (L=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. 3.1а (кривые A и B).

Таб. 3.1 Концентрации веществ A и B в условных единицах

	Обучающий набор									Проверочный набор				
A	0.90	0.80	0.70	0.60	0.50	0.40	0.20	0.30	0.10	0.55	0.75	0.25	0.35	1.45
B	0.85	0.35	0.25	0.95	0.45	0.15	0.55	0.65	0.75	0.70	0.50	0.30	0.80	0.20

Спектры сильно перекрываются, так, что невозможно выделить какие-то «селективные» каналы для оценки концентраций. На соседнем рисунке Рис. 3.1b представлены девять модельных спектров (I=9) различных смесей A и B (см. Таб. 3.1), в которые внесена случайная погрешность со стандартным отклонением 0.05. Они будут использоваться как калибровочный набор.

Для построения одномерной калибровки мы взяли интенсивности $s(\lambda_{50})$ девяти сигналов для канала 50 и изобразили их на Рис. 3.1c в зависимости от концентраций c_A и c_B веществ A (квадраты) и B (круги). Соответствующие калибровочные зависимости $s=bc$ показаны прямыми A и B.

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

$$\mathrm{RMSEC} = \sqrt{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2 \Big/ F}\,, \qquad (3.1)$$

где y_i и $\hat{y}_i$ соответственно, измеренные и предсказанные значения отклика (концентрации) для образцов сравнения i=1,..., I. Величина F – это число степеней свободы [38], и она равна I–1 для одномерной регрессии без свободного члена. Ясно, что чем меньше RMSEC, тем точнее описываются обучающие данные. Кроме того, качество калибровки характеризуется еще и *коэффициентом корреляции* R^2 между величинами $\mathbf{y}$ и $\hat{\mathbf{y}}$ – чем он ближе к единице, тем лучше. Соответствующие значения даны в Таб. 3.2.

Таб. 3.2. Модельный пример. Точность калибровки

	Одномерная калибровка		Метод главных компонент		Проекция на латентные структуры (ПЛС2)		Проекция на латентные структуры (ПЛС1)	
	A	B	A	B	A	B	A	B
RMSEC	0.209	0.469	0.077	0.051	0.070	0.044	0.069	0.045
R^2	0.80	0.21	0.93	0.97	0.94	0.98	0.95	0.98

Эти результаты подтверждают, что из-за недостатка «приборной» селективности одномерная калибровка неудовлетворительна. Калибровка с помощью множественной регрессии будет рассмотрена ниже, а сейчас покажем, как работает многомерная модель, построенная с помощью *регрессии на главные компоненты* (РГК [89]).

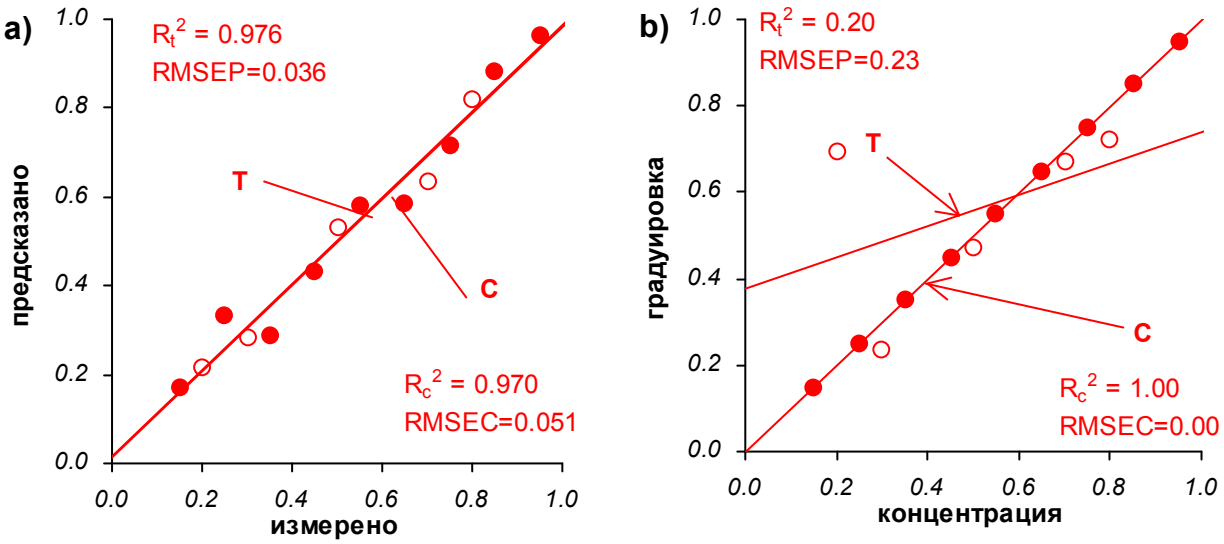
В методе РГК используется обратная калибровка, так, что **Y**=**C**, **X**=**S**. Применяя МГК, матрицу **X** можно разложить по формуле (2.1), причем в нашем примере, очевидно, K=2. Получившиеся векторы нагрузок $\mathbf{p}_1$ и $\mathbf{p}_2$ показаны на Рис. 3.1a (кривые ГК1 и ГК2). Сравнивая этот 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 калибровки методом РГК показаны на Рис. 3.1d, где изображены вычисленные с помощью модели значения концентраций $\hat{\mathbf{y}}$ в зависимости от соответствующих известных значений **y** (квадраты для A и круги для B), а также калибровочные линии, которые сливаются. Приведенные в Таб. 3.2 величины RMSEC и R^2, свидетельствуют о том, что метод РГК позволяет достичь высокой «математической» селективности и получить оценки концентраций веществ A и B с точностью, гораздо лучшей, чем в одноканальной калибровке. В методе РГК число степеней свободы в уравнении (3.1) равно, $F=I–K$.

Ранее уже отмечалось, что каждая формальная модель нуждается в полноценной проверке. В нашем примере такая проверка проводилась с помощью проверочного набора (тест-валидация), состоящего из пяти объектов (смесей A и B). На Рис. 3.2a представлены результаты проверки для вещества B. Здесь, в координатах «известно-предсказано», изображены девять объектов, участвовавших в калибровке (круги) и пять проверочных объектов (окружности). Там же приведены значения остатков калибровки (RMSEC) и предсказания (RMSEP), а также коэффициенты корреляции для калибровочного (R_c^2) и проверочного (R_t^2) наборов. Среднеквадратичный остаток предсказания (RMSEP) вычисляется аналогично RMSEC (формула (3.1)), но только для объектов из проверочного набора:

$$\mathrm{RMSEP} = \sqrt{\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 \Big/ M} \qquad (3.2)$$

При этом M равно числу таких объектов. Видно, что метод главных компонент выдерживает проверку – калибровочная (C) и проверочная (T) линии сливаются.

Рассмотрим, в этом контексте, метод калибровки с помощью *множественной регрессии*. Т.к. калибровочный набор состоит из девяти объектов, то для построения модели мы можем использовать не более восьми каналов (*I*>*J*), например: 1, 14, 27, и т.д.



Метод главных компонент Множественная регрессия

Рис. 3.2. Проверка калибровок в модельном примере.
Обучающий (●, C) и проверочный (○, T) наборы.

На Рис. 3.2b показаны результаты – калибровка (C) и проверка (T) – для метода множественной регрессии. Поскольку в этом случае число калибровочных объектов всего на единицу больше числа каналов, то калибровочная прямая в точности проходит через все точки обучающего набора (круги), поэтому RMSEC=0, и R_c^2=1. Однако проверка показывает, что построенная модель абсолютно неприемлема для предсказания. Это – типичный пример *переоценки* модели [68], когда точность описания калибровочных 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* [187]. 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.

Таб. 3.3. Модельный пример. Точность предсказания

	Метод главных компонент		Проекция на латентные структуры (ПЛС2)		Проекция на латентные структуры (ПЛС1)	
	A	B	A	B	A	B
RMSEP	0.054	0.036	0.049	0.032	0.048	0.032
R^2	0.99	0.98	0.99	0.98	0.99	0.98

В настоящее время самым популярным методом многомерной калибровки является метод *проекции на латентные структуры* (ПЛС), который уже упоминался выше. Он во многом похож на метод РГК, так как в его основе лежит проекционный подход. Однако РГК – это двухэтапный метод: на первом этапе применяется МГК, который анализирует структуру матрицы **X**, строит ортогональный базис в пространстве счетов ($K \leq J$), и, проецирует исходные данные на пространство меньшей размерности, тем самым, преодолевая проблему мультиколлинеарности в исходных данных; и только на втором этапе применяется множественная регрессия, т.е. вычисляются регрессионные коэффициенты. Что касается ПЛС метода, то он одноэтапный. Декомпозиция исходных данных и вычисление регрессионных коэффициентом производятся одновременно. Кроме того, если в исследуемых данных имеется несколько откликов ($L>1$), то при использовании РГК, строится отдельная регрессионная модель для каждого отклика, тогда как при использовании проекций на латентные структуры возможно построение одной общей регрессионной модели. Такой вариант ПЛС метода называется ПЛС2. Так же по-разному проводится декомпозиция исходных данных: для ПЛС - это одновременная декомпозиция матриц **X** и **Y** (Рис. 3.3)

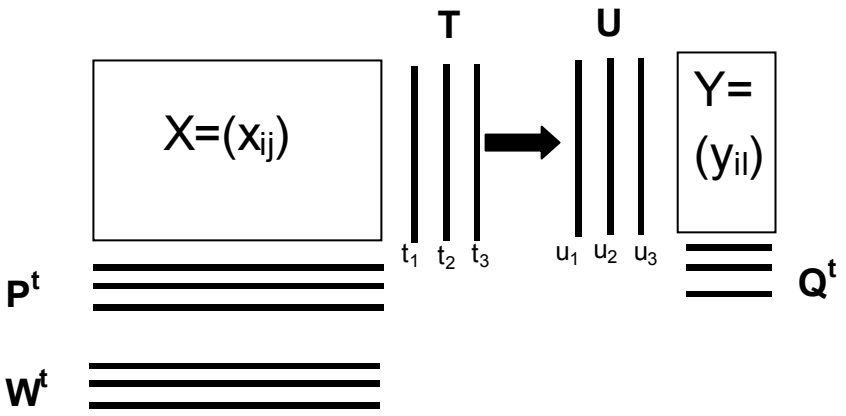


Рис. 3.3. Проекция на латентные структуры (ПЛС). Схема метода.

Проекции строятся согласованно – так, чтобы максимизировать корреляцию между соответствующими векторами X-счетов $\mathbf{t}_k$ и Y-счетов $\mathbf{u}_k$. Для этого, процедура ПЛС проецирования начинается с поиска такого вектора весов $\mathbf{w}_k$, чтобы вектор $\mathbf{t}_k=\mathbf{X}\mathbf{w}_k$ наилучшим образом описывал вектор откликов $\mathbf{y}_l$. Мерой близости между векторами $\mathbf{x}_j$ и $\mathbf{y}_l$ в ПЛС методе служит величина скалярного произведения $(\mathbf{x}_j^t,\mathbf{y}_l)$, поэтому $\mathbf{w}=(\mathbf{X}^t,\mathbf{y})$. Для устойчивости вычислений, вектор $\mathbf{w}$ шкалируется, так чтобы его длина равнялась единице. Определив вектор X-счетов $\mathbf{t}_k$, можно вычислить вектор X-нагрузок $\mathbf{p}_k$ для матрицы $\mathbf{X}$ и вектор Y-нагрузок $\mathbf{q}_k$ для матрицы $\mathbf{Y}$.

В результате применения ПЛС-метода исходную матрицу $\mathbf{X}$ можно представить в следующем виде

$$X_{ij} = \sum_{k=1}^{K} t_{ik} p_{kj} + e_{ij} \quad \text{или в матричном виде: } \mathbf{X}=\mathbf{T}\mathbf{P}^t+\mathbf{E}, \qquad (3.3)$$

При этом X-счета $\mathbf{t}_k$ ортогональны, а произведение $\mathbf{T}\mathbf{P}^t$ является хорошим приближением исходной матрицы $\mathbf{X}$. В отличие от МГК, векторы X-нагрузок $\mathbf{p}_k$ не ортогональны.
Одновременно происходит и декомпозиция матрицы откликов $\mathbf{Y}$

$$y_{il} = \sum_{k=1}^{K} u_{ik} q_{kl} + g_{il} \quad \text{или в матричном виде:} \quad \mathbf{Y}=\mathbf{U}\mathbf{Q}^t+\mathbf{G},$$

где $\mathbf{u}_k$ и $\mathbf{q}_k$– векторы счетов и нагрузок для матрицы откликов $\mathbf{Y}$. Ту же матрицу можно выразить через матрицу X-счетов $\mathbf{T}$, в виде

$$y_{il} = \sum_{k=1}^{K} t_{ik} q_{kl} + f_{il} \quad \text{или в матричном виде:} \quad \mathbf{Y}=\mathbf{T}\mathbf{Q}^t+\mathbf{F},$$

Y-остатки f_{il}, представляют разницу между известными откликами, и откликами, вычисленными в процессе моделирования. Именно они и составляют матрицу Y-остатков $\mathbf{F}$. В отличие от X-счетов, векторы $\mathbf{u}_k$ не ортогональны.

Используя полученное разложение, одновременно получаем оценки значений регрессионных коэффициентов для вычисления откликов в виде

$$\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{F}\,,\ \mathbf{A} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}\mathbf{Q}^t$$

С алгебраической точки зрения метод ПЛС, аналогично МГК (см. раздел 2.1), можно рассматривать как решение задачи на собственные значения. Первый вектор взвешенных нагрузок, $\mathbf{w}_1$, является собственным вектором, соответствующим максимальному собственному значению объединенной ковариационной матрицы $\mathbf{X}^t\mathbf{Y}\mathbf{Y}^t\mathbf{X}$. Последующие векторы $\mathbf{w}_k$, являются собственными векторами остаточной матрицы $\mathbf{Z}_k^t\mathbf{Y}\mathbf{Y}^t\mathbf{Z}_k$, где $\mathbf{Z}_k=\mathbf{Z}_{k-1}-\mathbf{T}_{k-1}\mathbf{P}^t_{k-1}$. В свою очередь, первый вектор X-счетов, $\mathbf{t}_1$,

является собственным вектором, соответствующим максимальному собственному значению о матрицы $\mathbf{XX}^t\mathbf{YY}^t$, а последующие векторы счетов $\mathbf{t}_k$, собственными векторами матриц $\mathbf{Z}_k\mathbf{Z}_k{}^t\mathbf{YY}^t$.

Графики векторов ПЛС счетов и нагрузок, так же как и в МГК, интенсивно используются для визуализации исследуемых данных. Кроме того, появляются важные графики зависимостей ($\mathbf{t}_k$,$\mathbf{u}_k$), которые показывают связь между X-переменными и Y-переменными в пространстве счетов (Рис. 3.4). На подобных графиках можно найти выбросы в X-переменных, в Y-переменных, а так же выбросы, относительно корреляционной связи X-Y. Рис. 3.4a демонстрирует сильную корреляцию между предикторами и откликами т.к. линия тренда проходит с наклоном, равным 1.0, и все объекты расположены недалеко от диагонали. Рис. 3.4b выявляет существенную нелинейную зависимость между предикторами и откликами. Такие данные желательно предварительно преобразовать (см. раздел 1.3) перед моделированием.

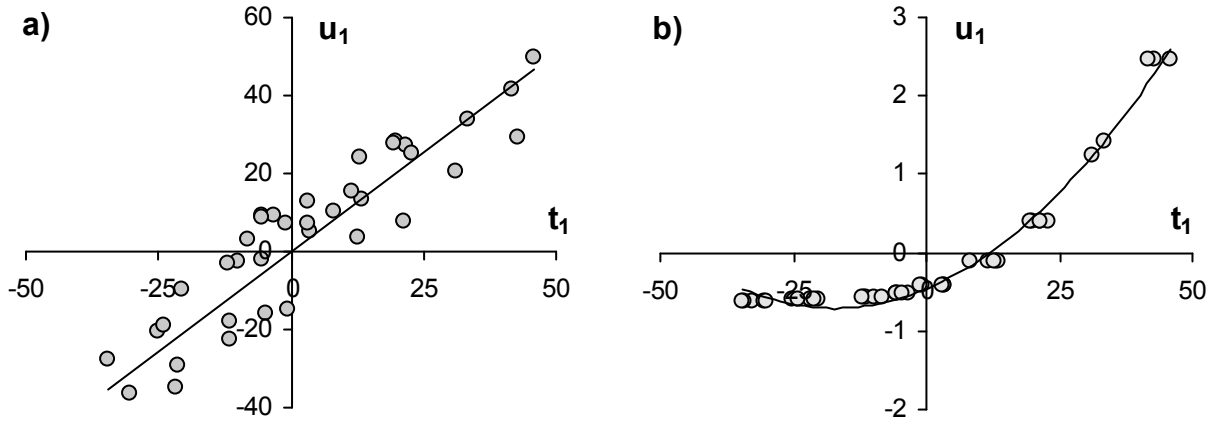
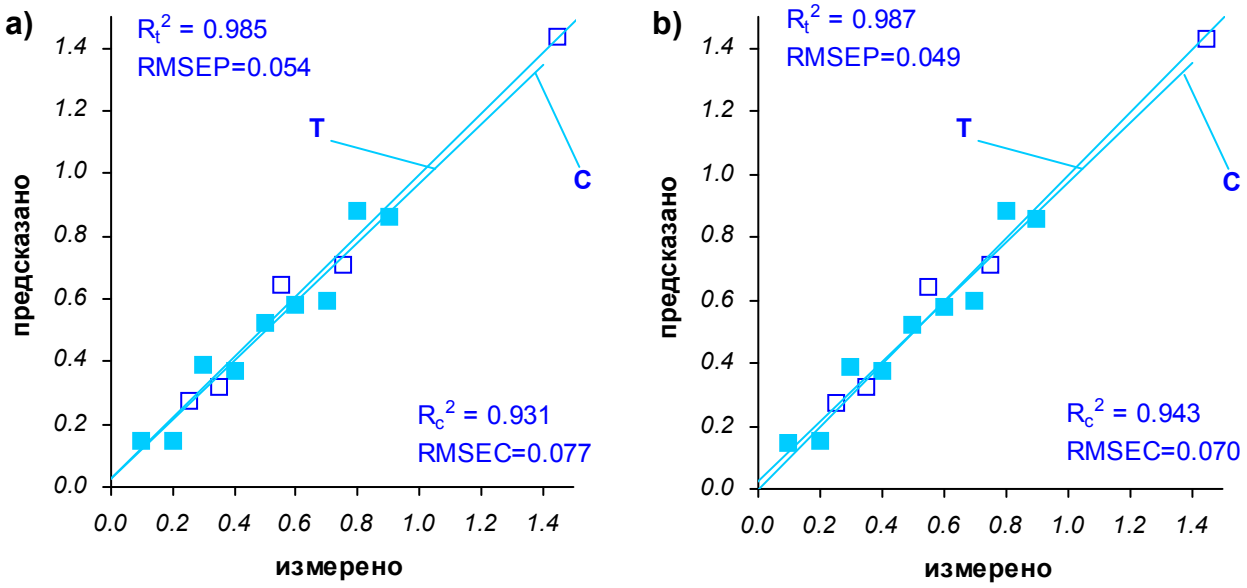


Рис. 3.4. Графики счетов ($\mathbf{t}_k$,$\mathbf{u}_k$) показывают зависимость между X-переменными и Y-переменными.

ПЛС метод применяется в двух видах, ПЛС1 – если модель строится только для одного отклика, и ПЛС2 – если одновременно регрессионная модель строится сразу для нескольких откликов. Детальное описание ПЛС алгоритмов приведено в Приложении, раздел 13.3.



ПЛС2 ПЛС1

Рис. 3.5 Применение ПЛС метода к модельному примеру. Анализ вещества А

Применим ПЛС метод к тому же модельному примеру смеси двух веществ, который рассматривался в начале этого раздела (Рис. 3.1a,b). При этом воспользуемся как ПЛС2, так и ПЛС1 алгоритмами. Для наглядности, результаты моделирования и предсказания для вещества А приведены на Рис. 3.5. Основные характеристики «качества» ПЛС1 и ПЛС2 моделей приведены в Таб. 3.2 и Таб. 3.3 . Из результатов расчетов видно, что, например, для вещества А, в среднем, т.е. по показателю RMSEC, точность калибровки по сравнению с РГК улучшилась на 9% при ПЛС2 моделировании, и на 13% при ПЛС1 моделировании. Что касается точности прогнозирования, то она улучшилась на 8% и 11% соответственно. Аналогичные показатели получились и для вещества В.

Так как в ПЛС методе декомпозиция матрицы **X** происходит с учетом корреляции с откликами Y, этот метод часто позволяет получить либо более точную, либо более простую модель, с меньшим числом ГК. Поэтому ПЛС регрессия лучше описывает сложные связи, используя при этом меньшее число главных компонент. Детальное описание метода ПЛС можно найти в [67, 71]. Этот подход послужил основой для очень многих методов калибровки, используемых в многомерном анализе данных, таких как SIMPLS [188], PMN [189], робастный PLS [190], Ridge PLS [191], и многих других подходов [192].

Существенным недостатком проекционных регрессионных методов (РГК, ПЛС и пр.) является то, что 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, как это делается в классическом регрессионном анализе [89] невозможно из-за сложности задачи. Так как оценки параметров в регрессионных моделях, основанных на проекционных методах, не являются независимыми (при $K<J$), то, с точки зрения классического статистического подхода, их доверительные интервалы неограниченны. К тому же, оценки параметров являются смещенными [119]. Использование имитационных методов [102] затруднительно из-за большого времени расчетов [105].

В [83] была предложена формула Мартенса (Martens) для оценки неопределенности прогноза для индивидуального объекта l, т.е. для оценки дисперсии ошибки предсказания

$$Var(\hat{y}_l - y_l) = R(\mathbf{y}_{val})\left(1 - \frac{K+1}{I}\right)\left(h_l + \frac{R(\mathbf{x}_l)}{R(\mathbf{X}_{val})} + \frac{1}{I}\right) \qquad (3.4)$$

где $R(\boldsymbol{y}_{val})$ – это среднеквадратичный остаток в откликах для объектов из проверочного набора, h_l – это размах (2.6), т.е. расстояние от предсказываемого объекта до модели, $R(\boldsymbol{x}_l)$ – это среднеквадратичный остаток для значений аналитического сигнала объекта l, $R(\mathbf{X}_{val})$ – значение среднеквадратичных остатков для аналитических сигналов (матрицы $\mathbf{X}$), для объектов из проверочного набора, I- количество объетов в калибровочном наборе, K число главных компонент. Формулу (3.4) ни в коем случае нельзя использовать для вычисления прогнозных интервалов при предсказании значения отклика y. Однако (3.4) можно рассматривать как дополнительную диагностику: если значения $Var(\hat{y}_l - y_l)$ велико, то надежность прогноза для l-ого объекта является низкой. Достоинством формулы Мартенса является то, что все величины в ее правой части могут быть явно вычислены и не требуют априорного знания величин погрешностей измерения.

Другой, статистический подход к оценке неопределенности в прогнозе представлен в работах [104, 108, 109] и нашел свое отражение в недавно опубликованном отчете IUPAC [110]. Учитывая все неопределенности при измерениях, моделировании и прогнозировании, в общем случае оценка дисперсии ошибки предсказания для нового объекта l может быть записана в следующем виде

$$Var(\hat{y}_l - y_l) = h_l \bullet (\sigma^2(\varepsilon) + \sigma^2(\Delta\mathbf{y}) + \|\boldsymbol{\beta}\|^2 \sigma^2(\Delta\mathbf{X})) + \sigma^2(\varepsilon_l) + \|\boldsymbol{\beta}\|^2 \sigma^2(\Delta\mathbf{x}) \qquad (3.5)$$

где h_l – это размах (2.4) для объекта l, $\sigma^2(\varepsilon)$ – дисперсия погрешностей моделирования, $\sigma^2(\Delta\mathbf{X})$ и $\sigma^2(\Delta\mathbf{y})$– дисперсии погрешностей измерения предиктора и отклика для объектов из обучающего набора, $\boldsymbol{\beta}$ - это гипотетический «истинный» вектор регрессионных параметров. Попытка включить в формулу все неопределенности приводит к тому, что (3.5) непосредственно вычислить невозможно. Слагаемые в правой части либо должны быть известны априори, например, погрешности измерения, либо оценены дополнительно, например $\|\boldsymbol{\beta}\|$.

Л. Канторович [193] предложил кардинально 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. Позднее этот метод был назван «простым интервальным оцениванием» (ПИО) [194]. Метод ПИО, его теоретические аспекты, а так же многочисленные приложения, представлены в данной работе.

3.2. Многомодальная регрессия

Методы многомерной калибровки естественно обобщаются на случай, в котором $\boldsymbol{X}$ и $\boldsymbol{Y}$ блоки являются N-модальными матрицами [82]. Эта регрессия может быть построена различными способами. Используя методы, описанные в разделе 2.3 (PARAFAC, Tucker3), блок предикторов $\boldsymbol{X}$ раскладывается в произведение 2D-матриц нагрузок, с помощью которых проводится оценка параметров. Эти методы можно рассматривать как обобщение метода РГК для многомодальных данных. Обобщением метода ПЛС является Tri-PLS декомпозиция 3D-матрицы $\boldsymbol{X}$, которую можно представить в «матрицированном» виде [196]

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

Здесь $^{u}\mathbf{X}$ – это 2D матрица (размерности $I \times LJ$), получаемая при развертке 3-D матрицы $\boldsymbol{X}$ (размерности $I \times L \times J$), как это показано на Рис. 2.5. $\mathbf{T}$ – это 2D-матрица счетов (размерности $I \times K$), а $^{u}\mathbf{P}$ – это 2D матрица весов (размерности $K \times LJ$), 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 ПЛС, счета $\mathbf{T}$ выбираются таким

образом, чтобы максимизировать корреляцию между векторами $\mathbf{t}_a$ и $\mathbf{u}_a$. Сама регрессионная задача **U**=**TB** решается традиционным способом.

Математический аппарат, используемый в многомодальной калибровке, довольно сложен. Однако в настоящее время, существуют программные продукты [197], позволяющие исследователям легко справляться с математическими трудностями. В литературе имеются многочисленные примеры использования мультимодальной калибровки в физико-химическом анализе: кинетическая спектрофотомерия для определения пестицидов [198], разрешение налагающихся пиков в ВЭЖХ-ДДМ [199], определение следовых концентраций металлов [200].

Проиллюстрировать использование многомодальной регрессии на примере [179]. В этой работе 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 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 объектов 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 7×210×37, в режиме детектирования по отдельным ионам размерность этой матрицы равна 7×8×22. Блок откликов – это 1D вектор **y**, состоящий из 7 концентраций.

Для построения калибровок 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 [114].

3.3. Нелинейная калибровка

В некоторых случаях, например в рассмотренных выше задачах титрования, построить линейную калибровку невозможно. Кроме того, линейный подход требует большого количества данных, которые не всегда доступны. В этом случае используются два альтернативных подхода: множественная нелинейная регрессия или многомерная нелинейная калибровка. В этом разделе мы рассмотрим оба эти подхода.

Нелинейный регрессионный анализ [201] 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, задачи [98]. Так, этот подход применялся при анализе активности антиоксидантов [32], для решения обратной кинетической задачи [30, 97], в уже упомянутом титровании [202, 203]. В работе [53] 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 [80]. Для учета нелинейных эффектов предлагаются разнообразные усовершенствования [84] обычного метода ПЛС: INLR [204], GIFI-PLS [205], QPLS [206]. Помимо нелинейного ПЛС, при формальном моделировании, активно применяется метод *искусственных нейронных сетей* (ANN) [207, 208], 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 физиков и химиков, которые начали применять их для классификации [209], дискриминации [49] и калибровки [210, 211]. Затем, однако, наметилось некоторое охлаждение интереса, и использование ANN заметно снизилось. Причина заключена все в той же проблеме переоценки моделей, о которой шла речь выше. 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 при формальном моделировании [212, 213]. Метод GA, и его разновидность – иммунный алгоритм (IA), полезны в тех случаях, когда задача физико-

химического анализа не поддается формализации в терминах обычных целевых функций, например при разрешении многокомпонентных перекрывающихся хроматограмм [214]. Пример практического применения различных нелинейных подходов в хемилюминесцентном анализе приведен в работе [215].

3.4. Результаты главы 3.

Рассмотрены основные подходы, применяющиеся при количественном анализе. С математической точки зрения, модели количественного анализа – это регрессионные задачи. При использовании формального подхода, модели могут быть как линейные, так и нелинейные относительно неизвестных параметров. При использовании содержательного подхода к моделированию, модели по большей части являются нелинейными.

Применение формальных моделей на основе проекционных методов доказало свою высокую эффективность, а также позволило решить широкий класс задач, которые до этого, при использовании содержательного подхода, решить было невозможно. Однако одним из узких мест при применении таких методов как РГК и ПЛС, является оценка неопределенности в прогнозе не в среднем, а индивидуально для каждого нового объекта, и связанная с этим оценка надежности результатов прогноза. В настоящее время, этот вопрос активно разрабатывается. Ему посвящены многочисленные исследования по теории применения проекционных методов для решения задач калибровки.

Метод простого интервального оценивания

Простое интервальное оценивание (ПИО) – это метод линейного моделирования и построения интервальных оценок прогноза в многомерной калибровке. ПИО дает результат в удобном интервальном виде, учитывающем все имеющиеся неопределенности: ошибки измерения предикторов и откликов, погрешности билинейного моделирования, и т.п. Кроме того, метод ПИО предоставляет новые возможности для построения содержательной классификации влиятельности объектов.

ПИО метод основывается на идеи Л. Канторовича [193] высказанной в 1962 году, а именно – при анализе 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 называется «*простым интервальным оцениванием*» (ПИО). В свое время эта идея не получила должного признания и развития, что было связано, по-видимому, с недостаточным быстродействием компьютеров. В 80-х-90-х гг., используя эту идею, был выполнен ряд важных прикладных работ [216-219], в частности получены интересные результаты по анализу информационной ценности кинетических измерений [220, 221], а так же работы в области аналитической химии [222]. Кроме того, проводились исследования, направлены на построение интервальной оценки параметров моделей (метод центра неопределенностей), что оказалось малоплодотворным. Итоги этих исследований были подведены в монографии [195], где подробно рассматривается основная задача, решаемая авторами. Это – задача интервальной оценки *параметров* моделей, погружение области возможных значений этих параметров в гиперкуб, параллелепипед, эллипсоид, и т.п.

Такая постановка задачи представляется неплодотворной и малоперспективной, что и было подтверждено практикой – за последние 10 лет новые работы в этом направлении не замечены. В тоже время, идея Канторовича может дать интересные результаты, если рассматривать многомерную калибровку (ММК) как задачу построение интервального прогноза *отклика y* [194]. В этом случае удается решить две равно важные практические задачи. Во-первых, установить область неопределенности [223] для прогноза искомого отклика, т.е. оценить *точность* построенной калибровки, индивидуально для каждого объекта. Во-вторых, используя подход ПИО, можно построить систему *классификации* объектов [52], т.е. установить индивидуальные особенности каждого объекта,

определяемые по его взаимоотношениям, как с моделью, так и с другими объектами. Общеизвестными примерами такой классификации являются такие понятия как *выброс* (объект, резко выделяющийся из общей закономерности) или *экстремальный объект* (находящийся в периферийной области модели и оказывающий значительное влияние на ее построение). Не смотря на широкое употребление этих понятий в различных исследованиях [116, 118, 122, 132, 224-228], не существует их общепризнанных определений и методов обнаружения. Метод ПИО может восполнить этот пробел.

Однако ПИО метод значительно отличается от традиционного, привычного регрессионного подхода, применяемого в задачах многомерной калибровки. Его «философия», математический аппарат, терминология непривычны для экспериментаторов. Исходя из этого, перед строгим изложением математических аспектов ПИО, предлагается максимально простое объяснение, основанное на простейших одномерном примере, с помощью которого, тем не менее, удается продемонстрировать все основные идеи и результаты. Основные материалы четвертой главы опубликованы в [229].

4. Объяснение ПИО метода

4.1. Почему погрешности ограничены

Основным предположением ПИО является ограниченность погрешности измерения. Такой подход к интерпретации данных физических экспериментов нуждается в некотором обосновании. При анализе данных, стандартным допущением является принцип нормальности погрешностей. Это либо явно, либо молчаливо предполагается. Хорошо известно, что постулат о нормальном распределении погрешности неоднократно подвергался критике с самых разных позиций. Многочисленные исследования (см. например, [230, 231]) показывают, что обычно погрешность измерения скорее ограничена, чем нормальна. Характерно, что большинство экспериментаторов не связывают с принципом нормальности факт неограниченности погрешностей. На прямой вопрос о том, как часто исследователю приходилось обрабатывать результаты экспериментов, в которых присутствовали значения, лежащие за порогом четырех стандартных отклонений (4σ), как правило, следовал ответ, что если такие величины и встречались, то они безжалостно удалялись еще на стадии предварительной обработки. В то же время объем данных, с которым работают сейчас ученые, часто превышает 10^{+6} [80], так что в них уверенно можно было бы ожидать 20-30 «нормальных» значений, выходящих за этот порог. Примечательно мнение авторов работы [116], которые отмечают: «действительно, в реальных исследованиях исследователь часто может, в какой-то степени, отобрать образцы, что приводит к распределению, которое скорее равномерно, чем нормально».

Рассмотрим типичный пример, подтверждающий эту точку зрения. Классической задачей ММК является определение характеристик качества зерна по БИК спектрам [140]. В рассматриваемом примере измерения проводились на приборе InfraLUM FT-10 в диапазоне 8000-14000 $см^{-1}$, а прогнозируемым показателем является процентное содержание влаги в зерне. Спектральные данные **X** подготавливались по процедуре включающей: (1) усреднение спектров по трем повторным измерениям; (2) логарифмирование; (3) сглаживание спектров по алгоритму Савицкого-Голэя [126] второго порядка с окном в 3 точки; (4) нормализацию каждого спектра вдоль спектральных линий, (5) центрирование и нормировка всех спектров по объектам. Данные ***y*** также усреднялись по трем повторным измерениям, а затем центрировались и нормировались. Модель многомерной калибровки строилась по 141 образцу в

спектральной области 9000 – 11000 см$^{-1}$ методом проекций на латентные структуры ПЛС (см. раздел 3.1). Для построения ММК достаточно четырех главных ПЛС компонент, которые объясняют, соответственно, 99% и 90% вариации **X** и **y**.

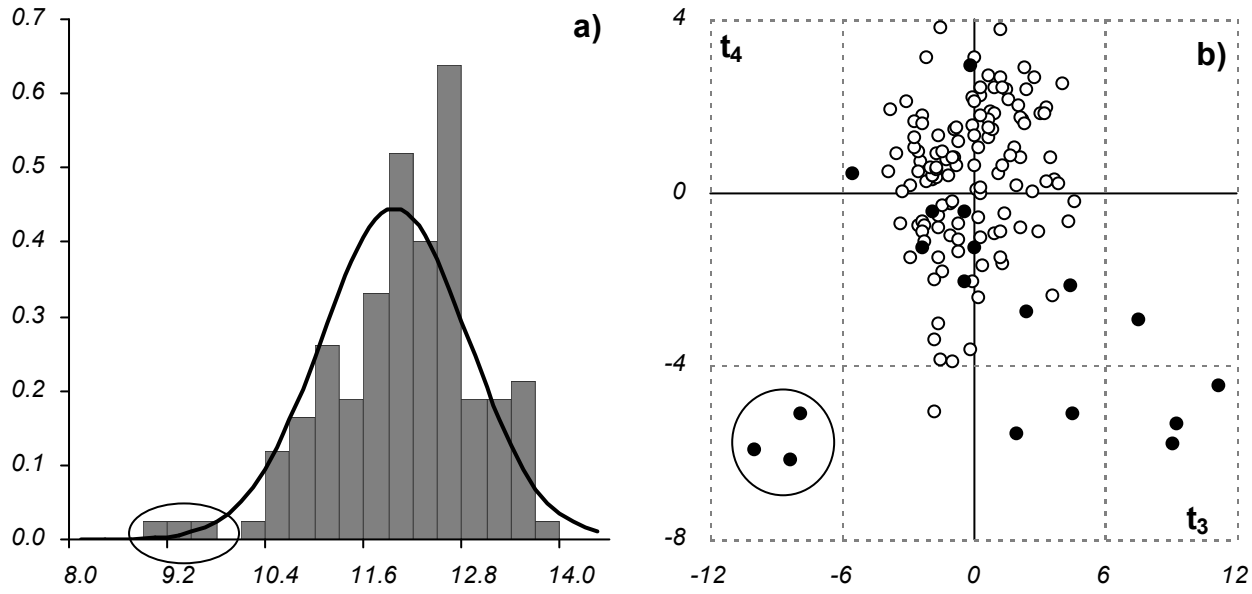


Рис. 4.1 ММК содержания влаги в зерне по БИК спектрам на исходном наборе из 141 образца. Гистограмма распределения содержания влаги в зерне (a) и график ПЛС счетов $\mathbf{t_3}$-$\mathbf{t_4}$ (b). Закрашенные точки отмечают «подозрительные объекты».

На Рис. 4.1 представлены гистограмма распределения значений содержания влаги в зерне (a) и график ПЛС счетов в координатах $\mathbf{t_3}$-$\mathbf{t_4}$ (b). Применяя к этим данным обычный статистический анализ, можно заметить, что они не противоречат гипотезе о нормальном распределении откликов. Даже три экстремальных объекта, отмеченные на графиках, выглядят «допустимыми» – вероятности их появления равны: 0.03, 0.21, и 0.38. Тем не менее, действуя согласно обычной процедуре ММК, мы удалили все объекты, отмеченные на Рис. 4.1b закрашенными точками, как выпадающие, и провели новую калибровку модели. Результаты обработки цензурированных данных (124 объекта) показаны на Рис. 4.2.

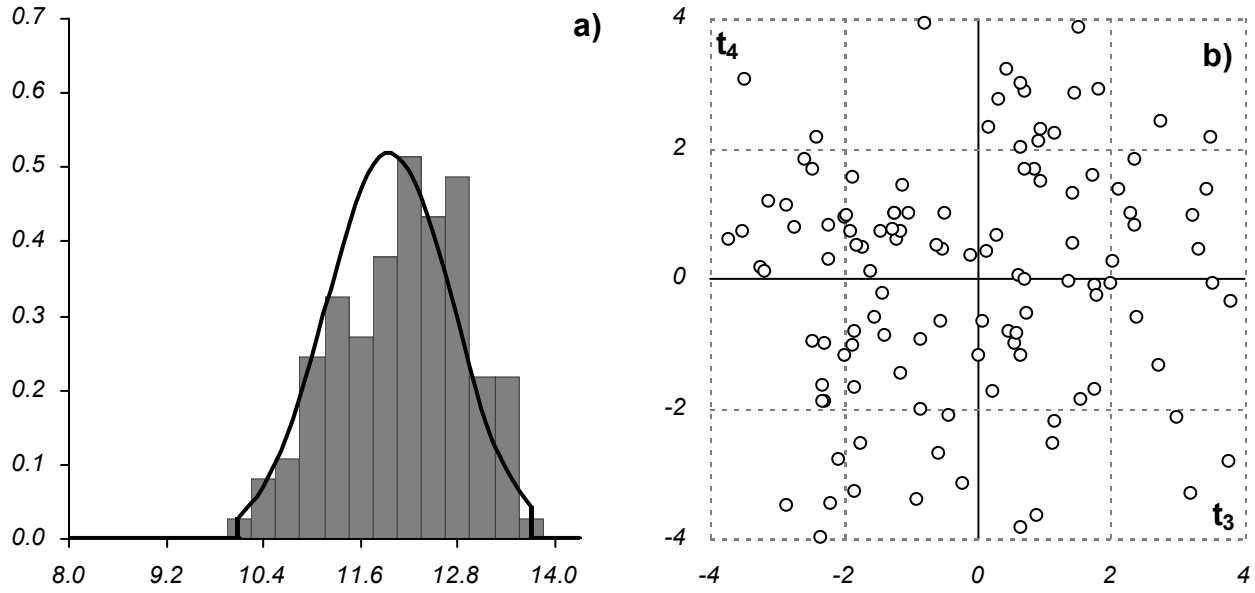


Рис. 4.2 ММК содержания влаги в зерне по БИК спектрам на цензурированном наборе из 124 объектов. Гистограмма распределения содержания влаги в зерне (a) и график ПЛС счетов $\mathbf{t_3}$-$\mathbf{t_4}$ (b).

Теперь ПЛС модель объясняет, соответственно, 99% и 92% вариации **X** и **y**. На графике счетов (b) уже не видно никаких подозрительных объектов. Однако при этом распределение отклика соответствует уже *усеченному* нормальному распределению, обрезанному на расстоянии $\pm 2.5\sigma$ от центра.

Рассмотренный пример показывает, что, следуя традиционным методам анализа и обработки данных, мы получаем ограниченные погрешности, которые подчиняются не нормальному, а, скорее, усеченному нормальному распределению. В следующем разделе мы увидим, какие выводы следуют из факта ограниченности погрешностей, рассматриваемого далее как постулат.

4.2. Модельный пример

Объясним, как работает метод ПИО, используя простейшую одномерную регрессию

$$y = \mathrm{x}a + \varepsilon \qquad (4.1)$$

Более строгое математическое изложение дано в разделе 5, из которого будут использоваться лишь несколько простейших формул.

Основным предположением метода ПИО является постулат об ограниченности погрешности измерения ε., который можно сформулировать следующим образом.

Никакая погрешность ε не может превосходить по абсолютной величине некоторую константу β, т.е. –

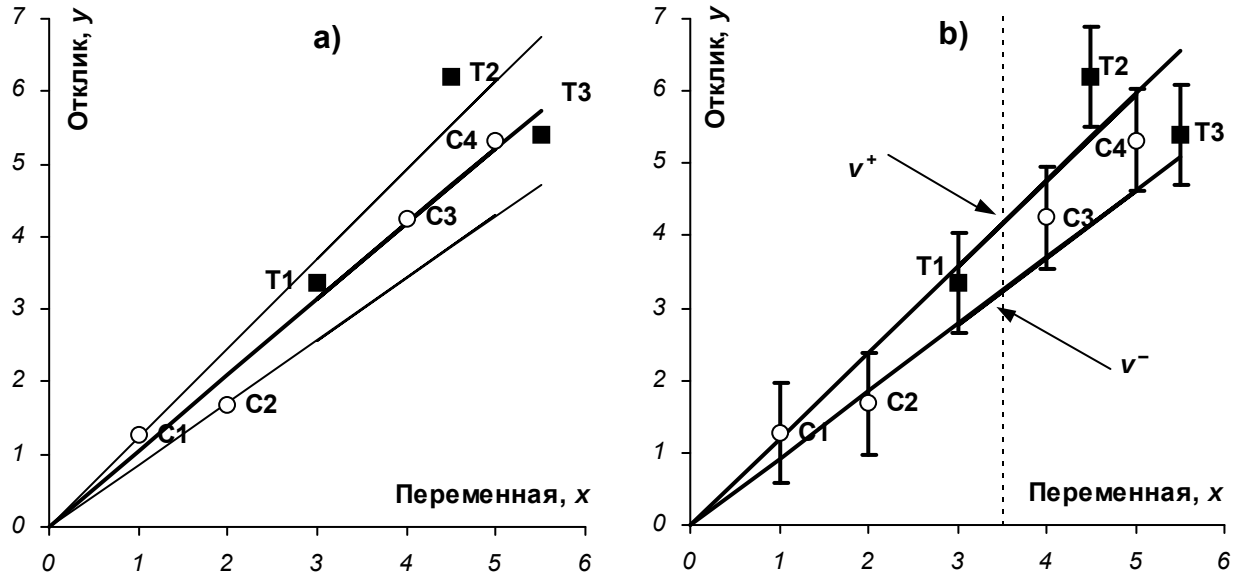
$$\text{Prob}(|\varepsilon| > \beta) = 0 \qquad (4.2)$$

Исследуем простейшие выводы, немедленно вытекающие из этого постулата. В Таб. 4.1 (столбцы 1 и 2) и на Рис. 4.3 приведены модельные данные, построенные нами для регрессии (4.1) при a=1. Погрешность измерения в отклике y моделировалась с использованием равномерного распределения шириной 1.4, т.е., в этом примере, β=0.7

Таб. 4.1 Модельные данные и результаты их обработки

Объекты	x	y	$\hat{y}$	$\hat{y}^-$	$\hat{y}^+$	$a^{\min}$	$a^{\max}$	v^-	v^+
0	1	2	3	4	5	6	7	8	9
C1	1.0	1.28	1.04	0.86	1.23	0.58	1.98	0.92	1.19
C2	2.0	1.68	2.09	1.72	2.46	0.49	**1.19**	1.85	2.38
C3	4.0	4.25	4.18	3.43	4.92	0.89	1.24	3.70	4.76
C4	5.0	5.32	5.22	4.29	6.15	**0.92**	1.20	4.62	5.95
T1	3.0	3.35	3.13	2.58	3.69	0.88	1.35	2.77	3.57
T2	4.5	6.19	4.70	3.86	5.53	1.22	1.53	4.16	5.36
T3	5.5	5.40	5.74	4.72	6.76	0.85	1.11	5.08	6.55

В этом примере использован очень короткий набор данных, который разбит на две части. Первые четыре объекта, обозначенные C1-C4, являются калибровочным набором, используемым для построения модели. На Рис. 4.3 они изображены открытыми кружками. Последние три объекта, обозначенные как T1-T3, являются проверочными объектами, для которых строится прогноз. Они показаны на Рис. 4.3 закрашенными квадратами. Не смотря на примитивность этого примера, с его помощью можно объяснить все основные свойства метода ПИО.



Метод наименьших квадратов. ▬ МНК прогноз, ─ границы доверительных интервалов

Метод ПИО: I – интервалы погрешностей, ▬ границы предсказанных интервалов

Рис. 4.3 Одномерный модельный пример. ○- калибровочные объекты, ■- проверочные объекты

МНК калибровка. Начнем, с традиционного метода наименьших квадратов (МНК) [89]. Используя калибровочные данные (x_i,. y_i), i=1 – 4 (столбцы 1 и 2 в Таб. 4.1, объекты C1-C4) и стандартную методику обработки, можно найти МНК оценку параметра a

$$\hat{a} = \frac{\bar{y}}{\bar{x}} = 1.003, \quad \text{где } \bar{x} = \frac{1}{4}\sum_1^4 x_i, \quad \bar{y} = \frac{1}{4}\sum_1^4 y_i, \tag{4.3}$$

и предсказать значения отклика y во всех точках x, как калибровочных, так и новых

$$\hat{y} = \hat{a}x, \tag{4.4}$$

(столбец 3 в Таб. 4.1 и Рис. 4.3a; жирная линия).

Мы также можем оценить дисперсию погрешности ε, применив хорошо известную формулу [201]

$$s^2 = \frac{1}{3}\sum_1^4 (y_i - \hat{y}_i)^2 = 0.078, \tag{4.5}$$

а также построить доверительные интервалы для отклика

$$\hat{y}^{\pm} = \hat{y} \pm s\frac{x}{2\bar{x}}\mathrm{t}_3(P). \tag{4.6}$$

Здесь $\mathrm{t}_3(P)$ — это квантиль распределения Стьюдента с тремя степенями свободы для вероятности P. Значения доверительных пределов для P=0.95 приведены в столбцах 4 и 5, Таб. 4.1 и на Рис. 4.3a, тонкие линии.

ПИО калибровка. Рассмотрим теперь, как эти же данные интерпретируются методом ПИО. Будем предполагать, что мы знаем, что β =0.7. В большинстве практических приложений дело обстоит сложнее и величина β заранее неизвестна. Позже мы рассмотрим, как можно разрешить эту проблему.

Из уравнения регрессии (4.1) и принципа ограниченности погрешности (4.2) следует, что для каждого объекта (x_i,. y_i), i=1 – 4 из калибровочного набора выполняется условие

$$\left| y_i - a x_i \right| \le \beta \tag{4.7}$$

или в эквивалентной форме –

$$a_i^{\min} \le a \le a_i^{\max}, \tag{4.8}$$

где

$$a_i^{\min} = \frac{y_i - \beta}{x_i} \qquad a_i^{\max} = \frac{y_i + \beta}{x_i}. \tag{4.9}$$

Значения (4.9) приведены в 6-ом и 7-ом столбцах Таб. 4.1. Неравенства (4.8) должны выполняться одновременно для всех калибровочных объектов, т.е. для i=1, 2, 3, 4. Очевидно, что так может быть только для тех значений параметра a, которые лежат в интервале

$$a^{\min} \le a \le a^{\max}, \tag{4.10}$$

где

$$a^{\min} = \max_{1 \le i \le 4} a_i^{\min}, \qquad a^{\max} = \min_{1 \le i \le 4} a_i^{\max}. \tag{4.11}$$

Эти значения выделены жирным шрифтом в соответствующих столбцах Таб. 4.1.

Интервал (4.10) определяет *область допустимых значений* (ОДЗ) параметра a, т.е. таких значений, которые не противоречат экспериментальным данным. Очевидно, что когда параметр a меняется в интервале (4.10), то соответствующая величина отклика y=ax в произвольной точке x ограничена значениями.

$$v^- \le y \le v^+, \tag{4.12}$$

где

$$v^- = a^{\min} x, \qquad v^+ = a^{\max} x. \tag{4.13}$$

Таким образом, построена интервальная оценка параметра a (4.10), которая является аналогом точечной оценки $\hat{a}$, получаемой с помощью МНК. Кроме того,

найдены и прогнозные интервалы (4.13) для отклика y, справедливые, как для калибровочных, так и для любых других (новых) объектов.

Рассмотрим графическую интерпретацию метода ПИО. На Рис. 4.3b приведены те же данные, что и на Рис. 4.3a, но теперь каждая точка сопровождается интервалом погрешностей (вертикальные отрезки) полушириной β =0.7. При построении ПИО-оценок нужно рассмотреть все возможные прямые, проходящие через начало координат, такие, чтобы каждая из них «зацепила» все интервалы погрешностей для всех четырех калибровочных объектов. Из графика видно, что нижним пределом является линия, проходящая через нижнюю точку интервала погрешностей объекта C4. Верхним пределом является прямая, проходящая через верхнюю точку интервала ошибок объекта C2. Все линии, заключенные между этими границами, очевидно, будут удовлетворять поставленным условиям (4.7) и, наоборот, любая прямая, лежащая вне этого угла, будет противоречить этим условиям. Границы показаны на Рис. 4.3b двумя жирными линиями v^+ и v^-.

Отметим очевидный факт, что построение калибровки методом ПИО в нашем примере «держится» только на двух объектах: C2 и C4. Именно они задают границы (4.10) возможных значений параметра a, так, что мы вправе назвать эти объекты *граничными*. Прочие калибровочные объекты C1 и C3 несущественны; их можно удалить из обучающего набора, и результат останется прежним. Это очень важное свойство метода ПИО, которое находит применение в задаче выбора представительного набора объектов [52]. Итак, было показано, что все объекты из калибровочного набора в методе ПИО разделяются на две группы: наиболее важные, граничные объекты, на которых держится модель, и несущественные, внутренние объекты, которые можно удалить из калибровочного набора и модель при этом не изменится.

Статус объектов. Рассмотрим, что может произойти с ПИО моделью, точнее с ее ОДЗ, при добавлении в калибровочный набор нового объекта. Очевидно, что ОДЗ может только уменьшиться. Например, при добавлении объекта T3, верхняя граница (линия v^+) пройдет уже ниже прежней, так чтобы «зацепить» верхнюю границу интервала погрешностей для T3; при этом $a^{\max}$ станет равной 1.11, вместо 1.16 (Таб. 4.1). Это свойство метода ПИО, называемое *состоятельностью* (см. утверждение(5.3)), важно в теоретическом плане. Оно показывает, что при увеличении числа объектов в калибровочном наборе, неопределенность ПИО оценок уменьшается. При этом если

максимальная погрешность выбрана правильно, или она, по крайней мере, не меньше β, то истинное значение параметра a всегда будет находиться внутри области допустимых значений (4.10). Это свойство, называемое *несмещенностью* (см. (5.5) в главе 5), также важно для понимания и обоснования метода ПИО. Однако не всякий новый объект, включенный в калибровочный набор, приводит к уточнению модели. Например, объект T1 не изменит ПИО модель. Это видно из Рис. 4.3b, где прогнозный интервал полностью лежит внутри соответствующего интервала погрешности T1, а также из 6-го и 7-го столбцов Таб. 4.1. Другой случай – это объект T2. Его интервал погрешности не пересекается с прогнозным интервалом, поэтому при добавлении T2 в калибровочный набор произойдет разрушение модели, поскольку система неравенств (4.7) станет несовместной. Это также хорошо видно из Таб. 4.1 – минимум по столбцу 7 (1.11) станет меньше, чем максимум по столбцу 6 (1.22). Таким образом, можно классифицировать новые объекты по трем группам по отношению к тому, как они повлияют на модель, если их добавить в калибровочный набор, т.е. определить статус (влиятельность) каждого объекта. Во-первых, можно выделить класс *внутренних* объектов, которые не меняют модель, и класс *внешних* объектов, которые меняют модель. Кроме того, из внешних объектов можно выделить еще группу *выбросов*, т.е. объекты, которые нельзя добавлять в калибровочный набор (при данном значении β), так как они разрушают модель.

4.3. Сходимость интервальных оценок

В математической статистике по большей части используются гладкие оценки, для которых имеет место фундаментальное неравенство Крамера-Рао [232]. Пусть $f_\alpha(x)$ – плотность распределения случайной величины x, зависящая от неизвестного параметра α., непрерывно дифференцируема по α, и пусть a – непрерывная оценка α со смещением $b(\alpha)$=E(a).Тогда для дисперсии оценки a выполняется неравенство (Крамера-Рао)

$$\mathrm{D}(a) \geq \frac{\left[1+b'(\alpha)\right]^2}{nI(\alpha)} \tag{4.14}$$

Здесь n – это объем выборки, а функция $I(\alpha)$ называется функцией информации Фишера.

Неравенство (4.14) устанавливает два фундаментальных свойства гладких оценок. Во-первых, из него следует, что неопределенность таких оценок не может быть меньше определенной в (4.14) величины, и только в лучшем случае (для т.н. эффективных

оценок) достигать этого предела. Во-вторых, видно, что с ростом объема выборки n, точность этих оценок изменяется как n^{-1}.

ПИО-оценки не являются гладкими, поэтому для них не выполняется неравенство Крамера-Рао. Это открывает теоретическую возможность построения “сверх эффективных” оценок, которые сходятся быстрее, чем n^{-1}. Проиллюстрируем эту идею на простом примере.

Рассмотрим следующий случай. Пусть имеется выборка $\mathbf{x}=(x_1,\ldots, x_n)$ из нормального распределения $N(\alpha,\sigma^2)$, усеченного на интервале $[\alpha-\beta, \alpha+\beta]$, $\beta=\kappa\sigma$. Соответствующая функция плотность распределения имеет вид

$$f(x\,|\,\alpha,\beta,\kappa)=\begin{cases}0 & , \quad x<\alpha-\beta \\ A(\beta,\kappa)\exp\left(-\frac{1}{2}\left(\frac{\kappa(x-\alpha)}{\beta}\right)^2\right) & , \quad \alpha-\beta\le x\le\alpha+\beta \\ 0 & , \quad x>\alpha+\beta\end{cases} \tag{4.15}$$

где

$$A(\beta,\kappa)=\frac{\kappa}{\beta}\left[\sqrt{2\pi}\,\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right]^{-1},$$

а кумулятивная функция

$$F(x)=\frac{1}{2}\left[1+\frac{\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\frac{x-\alpha}{\beta}\right)}{\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)}\right]. \tag{4.16}$$

Параметр $\kappa=\beta/\sigma$ определяет, как проводится отсечение. На Рис. 4.4 показаны несколько функций плотностей таких распределений в координатах $(x-\alpha)/\beta$ для различных значений $\kappa=\beta/\sigma$. Видно, что случай $\kappa=0.2$ очень близок к равномерному распределению, а случай $\kappa=4$ практически неотличим от полного, неусеченного нормального распределения.

Требуется построить оценку математического ожидания α при известном значении β и κ и исследовать ее сходимость, т.е. зависимость точности оценки от объема выборки n.

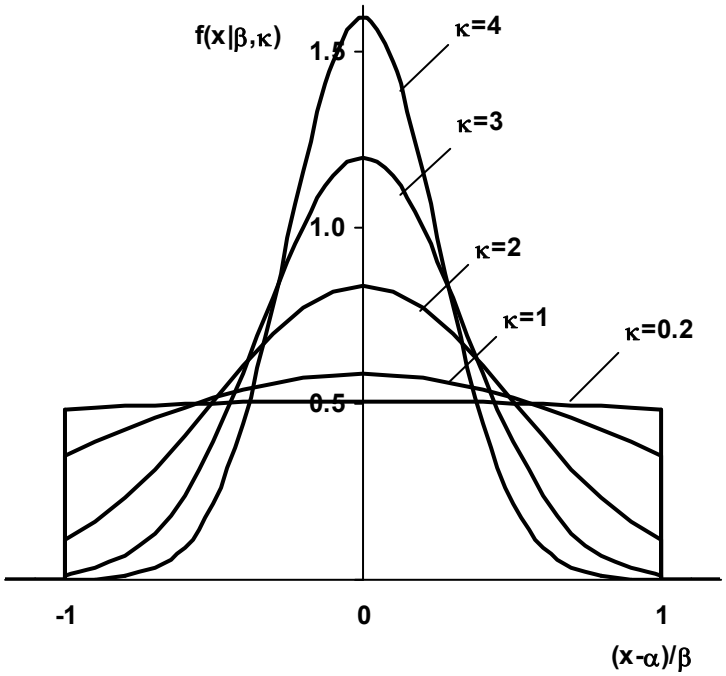


Рис. 4.4 Функции плотностей усеченного нормального распределения (4.15) в зависимости от значений κ

Традиционная оценка.

Обычная оценка α методом максимума правдоподобия (ММП) или моментов строится как среднее по выборке

$$a_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.17}$$

Оценка a_{ML} является несмещенной, $\mathrm{E}(a_{\mathrm{ML}}) = \alpha$ и эффективной. При достаточно больших n (n>10) она является нормальной со средним α и дисперсией

$$\mathrm{V}^2(a_{\mathrm{ML}}) = \frac{\beta^2}{n}\varphi^2(\kappa); \qquad \varphi^2(\kappa) = \frac{1}{\kappa^2}\left[1 - \frac{\sqrt{2}\kappa\exp\left(-\frac{\kappa^2}{2}\right)}{\sqrt{\pi}\,\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)}\right]. \tag{4.18}$$

Точность оценки (4.17) можно охарактеризовать приближенным доверительным интервалом

$$\mathrm{Prob}\left(\left| a_{\mathrm{ML}} - \alpha \right| < \beta\, h_{\mathrm{ML}}\right) = P, \tag{4.19}$$

где

$$h_{\mathrm{ML}}(P) = \frac{x_{0.5(1-P)}}{\sqrt{n}}\varphi(\kappa) \tag{4.20}$$

– это нормированная полуширина доверительного интервала (ML размах), а x_γ – это квантиль нормального распределения.

Интервальная оценка

Для той же выборки, оценка методом ПИО строится как интервал

$$a_{\mathrm{SIC}} = \left[\max(x_i - \beta), \min(x_i + \beta)\right] \tag{4.21}$$

Смысл этого определения ясен из Рис. 4.5. Ширину интервала (4.21) можно выразить как

$$l_{\mathrm{SIC}} = \min(x_i + \beta) - \max(x_i - \beta) = \left(\beta - \alpha + \min(x_i)\right) + \left(\alpha + \beta - \max(x_i)\right) \tag{4.22}$$

Учитывая, что случайные величины $\alpha + \beta - \max(x_i)$ и $\beta - \alpha + \min(x_i)$ одинаково распределены и (при больших значениях *n)* независимы, получаем

$$l_{\mathrm{SIC}} = 2\left(\alpha + \beta - \max(x_i)\right) \tag{4.23}$$

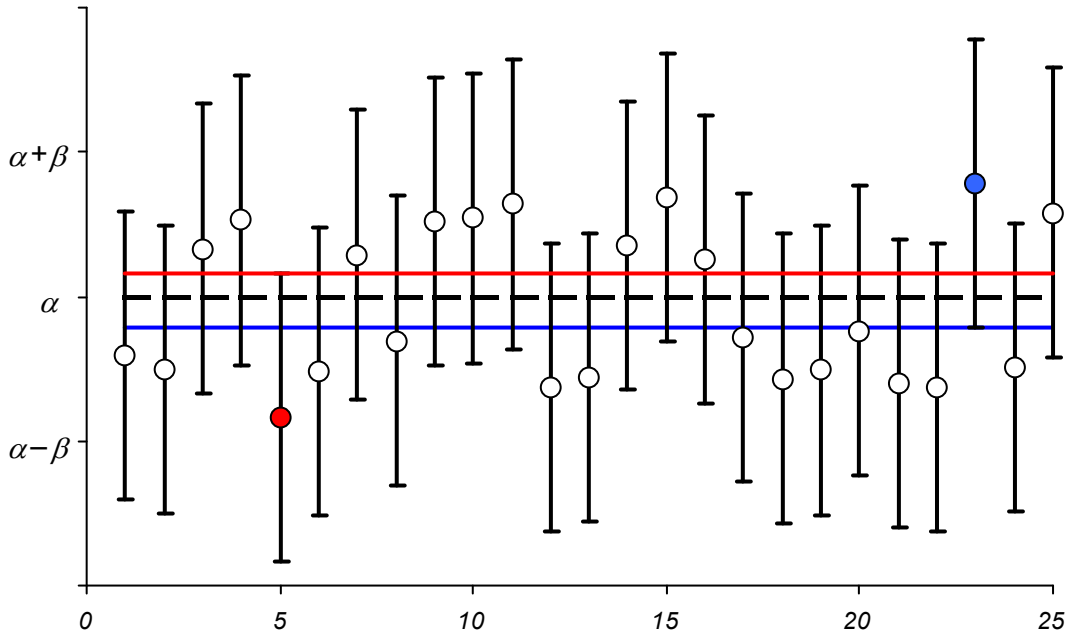


Рис. 4.5 Построение интервальной ПИО оценки

Величина l_{SIC} является случайной. Для определения ее распределения рассмотрим случайную величину

$$x_{\max} = \max(x_i) \tag{4.24}$$

Ее распределение имеет вид [233]

$$G(z) = \mathrm{Prob}\left(x_{\max} < z\right) = \left[F(z)\right]^n, \tag{4.25}$$

где функция F определена (4.16). Поэтому,

$$H(y) = \mathrm{Prob}\left(l_{\mathrm{SIC}} < y\right) = \mathrm{Prob}\left(2(\alpha+\beta-x_{\max}) < y\right) = \mathrm{Prob}\left(x_{\max} > \alpha+\beta-0.5y\right) = \\ = 1 - G\left(\alpha+\beta-0.5y\right) \tag{4.26}$$

Используя известную формулу математического анализа

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \mathrm{e}^x \tag{4.27}$$

получаем

$$H(y) \approx 1 - \exp\left(-\frac{ny}{4\beta\psi(\kappa)}\right), \tag{4.28}$$

где ψ(κ) – это сложная функция, которую, однако, можно аппроксимировать выражением

$$\psi(\kappa) \approx \exp\left(\frac{1}{3}\kappa^2\right)$$

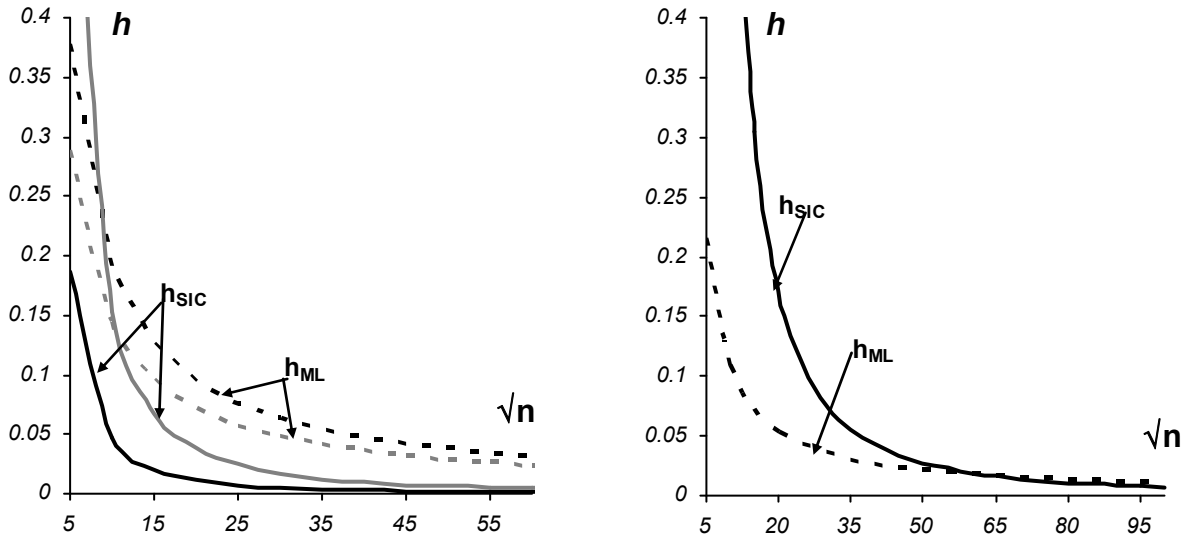
Полагая $H(y)=P$, находим нормированную полуширину (ПИО размах) для ПИО интервала, соответствующую доверительной вероятности P.

$$h_{\mathrm{SIC}}(P) = -\frac{\ln(1-P)}{n} 2\psi(\kappa) \tag{4.29}$$

Сравнивая формулы (4.20) и (4.29) можно сделать следующие выводы.

1. Видна общая тенденция, с ростом объема выборки ширина интервала неопределенности для оценки, полученный с помощью интервального подхода уменьшается в $\sqrt{n}$ раз быстрее, чем для ММП оценки.

2. При небольших объемах выборки и различных значениях κ выводы не столь однозначны. Рис. 4.6 позволяет сравнить величины (4.20) и (4.29) в зависимости от объема выборки и уровня отсечения, κ., при заданной вероятности P=0.9. Для $\kappa\leq 1$, и $n>10$, h_{SIC} меньше h_{ML}. Для $\kappa>1$, при малых объемах выборки, интервал неопределенности для ММП оценки оказывается меньше, чем для ПИО оценки, однако с увеличением объема выборки это соотношение меняется. Так, при заданной вероятности P=0.9, и κ=2, h_{SIC} становится меньше h_{ML} только тогда, когда объем выборки больше 100; а для κ=3 это происходи при n>3500 (Рис. 4.6b)



a) Черные кривые (сплошная и пунктир) при κ=0.2; серые кривые (сплошная и пунктир) при κ=2

b) κ=3

Рис. 4.6 Полуширина интервала неопределенности в зависимости от объема выборки, при заданной вероятности P=0.9

Таким образом, на этом примере показано, что для усеченных распределений ПИО-оценка лучше эффективной оценки ММП и действительно является "сверх эффективной". В случае же нормального, не усеченного, закона распределения оценка ММП лучше. Чем ближе усеченный закон распределения к неусеченному (большие значения κ), тем больше должен быть объем выборки, для того, чтобы ПИО-оценка могла конкурировать с ММП оценкой.

При рассмотрении более сложной, линейной регрессионной модели, величины в (4.20) и (4.28) будут зависеть не от n, а от матрицы плана эксперимента **X**. Доказательство аналогичного утверждения в общем случае наталкивается на существенные математические сложности, и поэтому не может быть предъявлено. С другой стороны, исследование этого вопроса выходит за рамки настоящей работы и может составить, по мнению автора, предмет более глубокого изучения специалистами в области математической статистики.

4.4. Результат главы 4

В этой главе показано, что основной постулат ПИО метода об ограниченности погрешностей соответствует сложившейся практике обработки данных физических экспериментов. Главное (и единственное) предположение об ограниченности

погрешности, является не недостатком, а преимуществом метода, так как, с практической точки зрения, оно выглядит более обоснованным, чем традиционное допущение о нормальности, а, следовательно, и неограниченности погрешностей.

Метод ПИО не использует никаких исходных предположений о виде погрешности, кроме ее ограниченности. Тем самым его можно считать методом, свободным от вида распределения.

Простейший линейный пример демонстрирует все основные свойства ПИО метода и позволяет сравнить результаты ПИО калибровки и прогноза с применением традиционного МНК.

Оценки, построенные на основе экстремальных статистик, такие как ПИО интервалы, являются, по-видимому, более эффективными, чем традиционные гладкие оценки при условии использования функций распределения, определенных на ограниченных носителях.

5. Описание метода ПИО

ПИО метод – это метод построения *линейных* моделей калибровки и прогнозирования для широкого круга задач анализа многоканальных экспериментальных данных. Эта глава представляет систематическое описание метода и его свойств в общем виде. Основные результаты этой главы представлены в [194]

5.1. Область допустимых значений

Рассмотрим модель линейной многомерной калибровки

$$\mathbf{y} = \mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon}\,, \qquad (5.1)$$

где $\mathbf{y}$ – это I-мерный вектор откликов; $\mathbf{a}$ – это J-мерный вектор параметров; $\mathbf{X}$ – это $(I\times J)$-мерная матрица предикторов (независимых переменных), $\boldsymbol{\varepsilon}$ – это вектор ошибок. Будем считать, что погрешности ε ограничены, т.е., что существует такая величина $\beta>0$, называемая максимальной погрешностью, что

$$\mathrm{Prob}\{|\varepsilon| > \beta\} = 0, \text{ и что для любых } 0<b<\beta \;\; \mathrm{Prob}\{|\varepsilon| > b\} > 0 \qquad (5.2)$$

где $\mathrm{Prob}\{\bullet\}$ обозначает вероятность события. Симметричность и гомоскедастичность погрешности ε, а также предположение об отсутствии погрешностей в матрице $\mathbf{X}$, не являются принципиальными и могут быть в дальнейшем отброшены. Для начала, будем полагать, что величина β известна.

Назовем пару $(\mathbf{x}_i, y_i)$, $i=1, \ldots, I$, калибровочным образцом. Здесь вектор $\mathbf{x}_i^t$ – это i-ая строка матрицы $\mathbf{X}$, соответствующая отклику y_i в уравнении (5.1). Согласно условию (5.2), для каждого $i=1,\ldots,I$ справедливы неравенства

$$y_i^- \le \mathbf{x}_i^t\mathbf{a} \le y_i^+, \qquad y_i^- = y_i - \beta, \qquad y_i^+ = y_i + \beta \qquad (5.3)$$

Естественно, что истинное значение вектора параметров, обозначаемое далее $\boldsymbol{\alpha}$, неизвестно. Однако можно рассмотреть все векторы $\boldsymbol{a}$, которые удовлетворяют этим неравенствам. Значения $\boldsymbol{a}$, которые удовлетворяют условию (5.3) для данного объекта i, образуют полосу $S(\mathbf{x}_i, y_i)$ в пространстве параметров R^J. Положение и ширина этой полосы определяются значениями $(\mathbf{x}_i, y_i)$. Рассмотрим все объекты из калибровочного набора и соответствующие им полосы. Очевидно, что вектор параметров $\mathbf{a}$ удовлетворяет всем неравенствам (5.3) одновременно тогда и только тогда, когда он принадлежит всем полосам.

Определение 5.1

Область допустимых значений (ОДЗ) A для параметров **a** системы (5.1) – это множество в пространстве параметров, образованное пересечением всех полос

$$A = \bigcap_{i=1}^{I} S(\mathbf{x}_i, y_i) \tag{5.4}$$

Область A – это замкнутый выпуклый многогранник [234, 235], образованный границами пересекающихся полос. A – это случайное множество, поскольку оно построено с использованием случайных величин **y**.

5.2. Свойства ОДЗ

Для любой модели заданной уравнением (5.1), ОДЗ A обладает следующими свойствами.

Утверждение 5.1

Область A является *несмещенной* оценкой параметра $\boldsymbol{\alpha}$.

В теории доверительных областей [236] это означает, что вероятность покрытия областью ложного значения не больше, чем вероятность покрытия истинного значения. Непосредственно из определения ОДЗ следует, что истинное значение $\boldsymbol{\alpha}$ всегда принадлежит A:

$$\mathrm{Prob}\{\boldsymbol{\alpha} \in A\} = 1 \tag{5.5}$$

В частности это означает, что если область A состоит только из одного элемента, т.е. $A=\{\mathbf{a}\}$, то этот элемент равен точному значению параметра $\boldsymbol{\alpha}$. Интересно отметить, что в ПИО анализе такая ситуация возможна даже в случае, когда количество объектов в модели конечно ($I<\infty$), в отличие от традиционного статистического подхода.

Утверждение 5. 2.

Область A *ограничена* тогда и только тогда [234, 235], когда матрица **X** имеет полный ранг

$$\mathrm{rank}\ \mathbf{X} = J. \tag{5.6}$$

Для доказательства этого утверждения сначала докажем следующую лемму.

Лемма

Существует такая область $A_1 \subset R^J$, что

$$A \subset A_1 \tag{5.7}$$

и

$$A_1 = \left\{\mathbf{a} \in R^J : (\mathbf{Xa} - \mathbf{z})^t \mathbf{W} (\mathbf{Xa} - \mathbf{z}) < r^2\right\}, \tag{5.8}$$

где $\mathbf{W} = diag(w_1,...,w_p)$ любая диагональная матрица с положительными коэффициентами,

$$r^2 = \mathbf{e}^t \mathbf{W} \mathbf{e}$$

и

$$\mathbf{z} = \frac{1}{2}\left(\mathbf{y}^+ + \mathbf{y}^-\right) \qquad \mathbf{e} = \frac{1}{2}\left(\mathbf{y}^+ - \mathbf{y}^-\right).$$

Лемма может быть легко доказана, если принять во внимание следующее очевидное утверждение: для любых векторов **c** и **b** с неотрицательными компонентами, из условия $\mathbf{c} < \mathbf{b}$ следует $\mathbf{c}^t\mathbf{c} < \mathbf{b}^t\mathbf{b}$. (Неравенства понимаются по-компонентно)

Следствия из этой леммы помогут нам доказать Утверждение 5.2. Обозначим через $\lambda_1 > \lambda_2 > ... > \lambda_J$ собственные значения матрицы $\mathbf{X}^t\mathbf{X}$.

Сначала докажем, что если матрица **X** имеет полный ранг, то область *A* ограничена.

Доказательство Утверждения 5.2.

Рассматривается ОДЗ *A,* которая определена системой неравенств

$$\mathbf{y}^- < \mathbf{X}\mathbf{a} < \mathbf{y}^+ \qquad (5.9)$$

Если матрица **X** имеет полный ранг, т.е. выполнено условие (5.6), тогда у матрицы $\mathbf{X}^t\mathbf{X}$ имеется *J* положительных собственных значения $\lambda_1 > \lambda_2 > ... > \lambda_J > 0$. Для построения множества A_1 (5.8), в качестве диагональных элементов матрицы **W** можно взять обратные величины собственных значений матрицы $\mathbf{X}^t\mathbf{X}$, т.е. $w_i = \lambda_i^{-1}$, тогда область A_1 будет представлять собой эллипсоид, покрывающий ОДЗ, и определяться формулой

$$A_1 = \left\{\mathbf{a} \in R^J : (\mathbf{X}\mathbf{a} - \mathbf{z})^t \operatorname{diag}(\lambda_1^{-1}, \lambda_2^{-1},..., \lambda_J^{-1})(\mathbf{X}\mathbf{a} - \mathbf{z}) < r^2\right\}, \qquad (5.10)$$

Следовательно, область $A \subseteq A_1$ ограничена.

Теперь докажем, обратное утверждение, что если область *A* ограничена, то матрица **X** полного ранга.

Будем доказывать от противного. Пусть

$$\text{rank } \mathbf{X} < J \qquad (5.11)$$

но область *A* ограничена.

Т.к. $A \neq \varnothing$, то существует хотя бы одна точка $\mathbf{a}_0 \in A$, такая, что

$$\mathbf{X}\mathbf{a}_0 = \mathbf{y}_0 \qquad (5.12)$$

и

$$\mathbf{y}^- < \mathbf{y}_0 < \mathbf{y}^+. \qquad (5.13)$$

Рассмотрим систему уравнений

$$\mathbf{Xa} = \mathbf{y}_0 \qquad (5.14)$$

относительно вектора **a**. Так как матрица **X** имеет неполный ранг, то система (5.14) задает линейное многообразие L (гиперплоскость) значений **a**, для которых, в силу (5.13), выполняется условие (5.9). Поэтому $L \subseteq A$. Однако L неограничено, как всякое линейное многообразие, следовательно, неограниченно множество A. QED

Из Утверждения 5.2 следует, что если система (5.1) мультиколлинеарна, т.е. rank **X** < J, то до использования ПИО метода необходимо применить какую-либо процедуру регуляризации. Например, можно использовать стандартный хемометрический подход [67, 235] и спроецировать исходные данные (5.1) на подпространство меньшей размерности

$$\mathbf{y} = \mathbf{TP}^t\mathbf{a} + \mathbf{f} = \mathbf{Tq} + \mathbf{f}, \qquad (5.15)$$

где матрица счетов **T** имеет полный ранг $K<J$ и затем применить метод ПИО к этой системе.

Утверждение 5.3

Область A является *состоятельной* оценкой параметра $\boldsymbol{\alpha}$, т.е., по построению,

$$\mathrm{Prob}\{A \cap \boldsymbol{\alpha}\} = 1 \quad \text{при} \quad I \to \infty \qquad (5.16)$$

при тех же «слабых» условиях [237]:

$$\lambda_j \to \infty \quad \text{при} \quad I \to \infty,$$

что и в МНК. Это свойство означает, что при увеличении количества калибровочных объектов, область A стягивается к истинному значению $\boldsymbol{\alpha}$.

Утверждение 5.4

Область A образована не всеми объектами из калибровочного набора, а только некоторыми, называемыми *граничными.* Поэтому, из калибровочного набора можно исключить все объекты, кроме граничных, и ОДЗ при этом не изменится.

Доказательство этого утверждения очевидно.

5.3. Предсказание отклика

Рассмотрим задачу предсказания отклика y для некоторого нового вектора **x** по модели (5.1). Если параметр **a** меняется внутри ОДЗ A, то, очевидно, что предсказываемое значение $\mathbf{y}=\mathbf{x}^t\mathbf{a}$ принадлежит интервалу

$$V = [v^-, v^+] \qquad (5.17)$$

где

$$v^- = \min_{\mathbf{a}\in A}(\boldsymbol{x}^t\mathbf{a}), \qquad v^+ = \max_{\mathbf{a}\in A}(\boldsymbol{x}^t\mathbf{a}). \tag{5.18}$$

Интервал V является результатом прогноза методом ПИО. Для его вычисления не нужно строить область A в явном виде, т.к. решения задач (5.18) могут быть найдены с помощью стандартных методов линейного программирования [238, 239], которые используются для нахождения оптимального значения (минимума или максимума) линейной функции на выпуклом замкнутом множестве – многограннике A. Известно, что оптимальное значение достигается в одной из вершин этого многогранника. Стандартная процедура симплекс-метода, целенаправленно (двигаясь либо в сторону увеличения, либо в сторону уменьшения) исследует значения функции в вершинах многогранника, до поиска оптимума. Свойства симплекс метода, например условия сходимости за конечное число шагов хорошо изучены [238], а сама вычислительная процедура разработана в деталях [239] и является частью многих стандартных математических пакетов, например [76, 78]. Она подробно рассмотрена в главе 7.

5.4. Оценка β

Для применения метода ПИО необходимо знать величину максимальной погрешности β. Обычно она неизвестна и, вместо β, используется некоторая оценка b. Понятно, что в этом случае ОДЗ, A, зависит от b и что $A(b)$ монотонно расширяется с увеличением b –

$$b_1 > b_2 \quad \Rightarrow \quad A(b_1) \supset A(b_2). \tag{5.19}$$

В случае, когда имеется последовательность оценок $b_1>b_2>\ldots \geq\beta$, сходящаяся к β, то Утверждения 5.1-5.4 будут выполняться и для $A(b_n)$. Кроме того, очевидно, что

$$A(0) = \varnothing, \quad A(\infty) \neq \varnothing \tag{5.20}$$

Из (5.19)-(5.20) следует, что существует *минимальное* значение b, при котором $A(b) \neq \varnothing$. Это значение может быть принято в качестве оценки величины β

$$b_{\min} = \min\{b, \quad A(b) \neq \varnothing\} \tag{5.21}$$

Оценка (5.21) является состоятельной, но смещенной, т.к. $b_{\min}\leq\beta$ для любого количества объектов I в калибровочном наборе. Она задает нижний предел всех возможных значений β. Это, несомненно, полезная характеристика калибровочного набора и модели, но помимо $b_{\min}$ необходимо оценить и верхнюю границу максимальной погрешности.

Очевидно, что любая разумная оценка b должна зависеть от двух обстоятельств:

1. Число объектов в калибровочном наборе. Чем больше объектов, тем ближе величина b к β.

2. Тяжесть крыльев функции распределения ошибок. Чем крылья легче, тем хуже эта оценка.

Применяя традиционный статистический подход [233], можно построить такую оценку b, что Prob$\{b>\beta\}>P$ и, при этом, оценка b максимально близка к β.

Рассмотрим $\hat{\mathbf{y}}$ – некоторую точечную (регрессионную) оценку вектора $\mathbf{y}$, остатки $\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}$, и величины

$$r_i=\frac{e_i}{\sqrt{1-h_i}},\ \text{где}\ h_i=\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \tag{5.22}$$

Положим

$$b=b_{\text{reg}}=\max\left(|r_1|,...,|r_I|\right),\ \text{и}\ w=\frac{1}{b}\sqrt{\frac{1}{I}\sum_{i=1}^{I}e_i^2}\ . \tag{5.23}$$

Исходное предположение состоит в том, что

$$\frac{b}{\beta}\sim \mathrm{N}\left(M(w,I),D^2(I)\right), \tag{5.24}$$

где N(0,1) – это стандартное нормальное распределение. Тогда величина

$$b_{\text{SIC}}=\frac{b}{M-x_PD}, \tag{5.25}$$

где x_P – это квантиль нормального распределения, представляет верхний доверительный предел для β. Действительно

$$\Pr\left\{b_{\text{SIC}}\ge\beta\right\}=\Pr\left\{\frac{b}{\beta}\ge M-x_PD\right\}=\Phi(x_P)=P\ , \tag{5.26}$$

Для подтверждения предположения (5.24), а также для построения функций M и D, было проведено статистическое моделирование, которое характеризуется следующими свойствами.

1. Рассматривались шесть калибровочных наборов с различным количеством объектов, I=10, 20, 50, 100, 250.
2. Для каждого такого набора погрешность моделировалась распределенной на интервале [-1, 1], т.е при β=1. Заметим, что формула (5.25) является инвариантной относительно масштабирования, т.е. при замене β на aβ, т.к. при этом b

заменяется на ab, а величина w не меняется. Поэтому при моделировании достаточно использовать β=1.

3. Погрешность моделировалась с помощью семи нормальных распределений, усеченных на [-1, 1], $N(0, \kappa^{-2})$, где κ =0.2, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0. (См.раздел 4.3 и Рис. 4.4). Когда параметр κ=0.2, получается распределение, близкое к равномерному, а при κ=3 – почти нормальное распределение. Таким образом, рассматривался достаточно широкий круг распределений, которые, в основном, и встречаются в экспериментальной практике.
4. Для каждого I и κ, 500 раз моделировали набор объектов, т.е. для каждой пары параметров было создано 500 различных калибровочных наборов. Для каждого такого набора вычислялись оценки b и w (5.23).
5. Так как на практике, параметр κ неизвестен, результаты моделирования для всех значений параметра κ при фиксированном I объединялись. Т.е. для каждого значения I было вычислено 3500 пар (b, w).

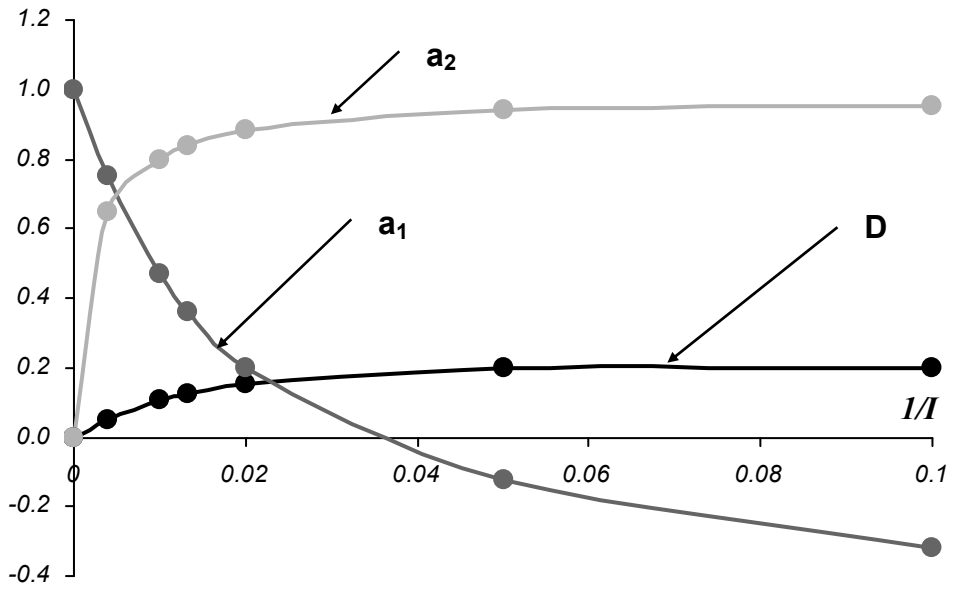


Рис. 5.1 Приближение статистик вычисленных по выборкам эмпирическими функциями (5.28) и (5.29)

Анализ данных показал, что условное математическое ожидание M может быть приближено формулой

$$\mathrm{E}(b \mid w) = M(I, w) = a_1(I)w + a_2(I), \qquad (5.27)$$

где

$$a_1(I)=\frac{1-172.812I^{-1.5287}}{1+172.812I^{-1.1626}} \quad a_2(I)=0.9572\left[1-\exp\left(-17.8528/\sqrt{I}\right)\right], \tag{5.28}$$

а условная дисперсия D – формулой

$$D(I)=0.2011\left[1-\exp\left(-73.4113/I\right)\right]. \tag{5.29}$$

При построении этих функции, принимались во внимание следующие очевидные свойства

$$a_1(1)\approx -1, \quad a_1(\infty)=1, \quad a_2(\infty)=0, \quad D(\infty)=0$$

На Рис. 5.1 показано, как эти зависимости приближают экспериментальные значения, вычисленные по выборкам, полученным при имитационном моделировании.

Правильность предположения (5.25) подтверждается графиком (Рис. 5.2), на котором представлены выборочные вероятности того, что условные стандартизованные значения больше, чем соответствующий квантиль нормального распределения x_P

$$P_c(P)=\Pr\left\{\frac{b-M(I,w)}{D(I)}>-x_P\right\} \tag{5.30}$$

Если предположение (5.25) верно, то P_c=P

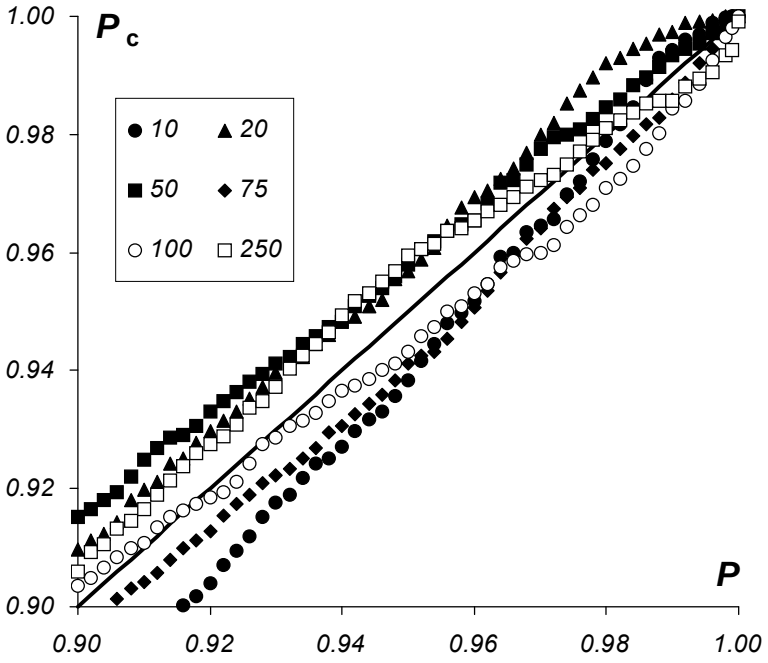


Рис. 5.2 Выборочная вероятность P_c (точки) относительно теоретической вероятности P вычисленной для нормального распределения с параметрами (5.27)

Из Рис. 5.2 видно, что на интервале $0.90 < P < 1$ эмпирические значения достаточно хорошо приближают теоретические вероятности.

Рис. 5.3 показывает зависимость $C(I, w)$

$$C(I,w) = \left(M - x_P D\right)^{-1}, \tag{5.31}$$

относительно величины w для различных объемов выборки (I=10, 20,..., 250) и значения квантиля нормального распределения $x_{P,}$ вычисленного для вероятности P=0.95. Значения, посчитанные при моделировании (точки), хорошо согласуются с вычисленными (сплошные линии) с помощью функций (5.27)-(5.29). Рост функции (5.31) для I=10 и I=20 объясняется отрицательной корреляцией между параметрами b и w.

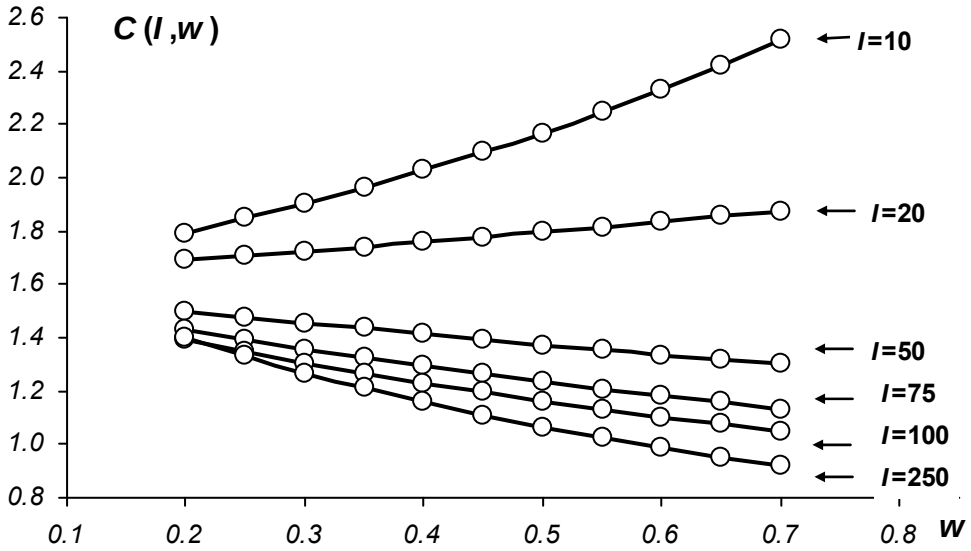


Рис. 5.3. Корректирующая функция C(I,w) для различного объема выборки (I).
Точки – значения, полученные при моделировании, сплошные линии – вычисленные значения.

Таким образом, мы получили однозначный алгоритм для вычисления b_{SIC} для исследуемого калибровочного набора из I объектов:

1. вычисление значений b и w (5.23) с помощью метода наименьших квадратов (либо любого проекционного метода, например МГК или ПЛС, если этого требует конкретная задача)
2. вычисление корректирующей функции $C(I,w)$, определенной уравнением (5.31), относительно параметров, вычисленных по формулам (5.28) и (5.29)
3. вычисление $b_{SIC} = b\, C(I,w)$

Именно значение b_{SIC} в качестве оценки β в дальнейшем используется в ПИО методе для определения прогнозных интервалов и для классификации объектов. При построении оценки b_{SIC} учитывались два важных свойства, во-первых, это оценка сверху, т.е. $b_{SIC} \geq \beta$, но с другой стороны b_{SIC} строилась так, чтобы быть максимально близкой по значению к β.

В принципе, возможно построение и других оценок для величины β. Для грубой оценки, можно представить такое практическое правило, которое называется правилом ‘1-2-3-4 сигма’. Если предположить, что среднеквадратичный остаток моделирования RMSEC $\cong 1\sigma$, тогда $b_{\min} \cong 2\sigma$, $b_{reg} \cong 3\sigma$ и $b_{SIC} \cong 4\sigma$. Конечно, это правило отражает только тенденцию, при этом, как уже говорилось, оценка β зависит от количества объектов в калибровочном наборе, однако многочисленные практические примеры подтверждают справедливость этого правила. В основе этого правила лежат следующие соображения. Для любого распределения погрешности ε, максимальная погрешность, т.е. размах, не может быть меньше 2σ. В крайнем случае, когда ошибка распределена равномерно, β=1.7σ [233]. Далее, для обычного объема выборки (I<1000), маловероятно ожидать выбросы, которые располагаются за границами 3σ. Граница в 4σ дает нам уверенность, что новые объекты не будут выпадать за эту границу. Таким образом, можно утверждать, что при всех практически возможных функциях распределения, значения максимальной погрешности будут располагаться между тремя и четырьмя σ.

В Таб. 5.1 приведены оценки β, вычисленные для реальных практических примеров рассмотренных в главах 8-12.

Таб. 5.1. Иллюстрация правила ‘1-2-3-4σ’

Данные	I	RMSEC/σ	$b_{\min}$/σ	b_{reg}/σ	b_{SIC}/σ
Нефть	40	1.0	2.3	3.0	4.1
ДСК	11	1.0	1.8	2.0	3.7
Пшеница	139	1.0	2.8	3.0	3.7

Точность ПИО моделирования.

Значения $b_{\min}$ и b_{SIC} полностью характеризуют точность ПИО моделирования, т.е.

$$b_{\min} \le \beta \le b_{SIC} \tag{5.32}$$

При этом о точности моделирования можно сказать следующее:

1. Любое априорное значение β допустимо только в том случае, если оно больше или равно $b_{\min}$.
2. Моделирование с помощью ПИО методов с параметром b_{SIC} гарантирует, что для объектов из калибровочного набора, ‘истинное’ значение отклика расположено внутри соответствующего прогнозного интервала.

3. Даже в наихудшем случае, полуширина прогнозного интервала для объектов из калибровочного набора меньше или равна b_{SIC}.
4. Обе оценки β: b_{min} (определенная в (5.21) и b_{SIC} (определенная в (5.25)) - являются состоятельными. Это означает, что для любого значения β из интервала (5.32)) выполняются свойства 2, 4 из раздела 5, а свойства 1 , 3 выполняются асимптотически.

Вычисление вероятности покрытия.

Уравнения (5.25) и (5.26) показывают, как вычислить оценку b_{SIC}, которая обеспечивает верхний 95-% доверительный интервал для неизвестного значения максимальной погрешности β

$$\text{Prob}\{b_{SIC} > \beta\} = 0.95\,.$$

Из этого можно было бы предположить, что та же вероятность 0.95 является вероятностью покрытия истинного значения прогнозным интервалом. Однако это не так.

Предположим, что истинное значение отклика y не попадает в интервал $V(b_{SIC})$ (5.17), построенный для найденного значения b_{SIC}. Это событие возможно тогда, когда одновременно реализуются два независимых события.

4. Оценка b_{SIC} меньше β.
5. Истинное значение y находится на краю интервала $V(\beta)$, там, где интервал $V(\beta)$ не пересекается с интервалом $V(b_{SIC})$.

Для того чтобы вычислить вероятности этих событий, введем случайную величину

$$z=b_{SIC}/\beta,$$

которая изменяется на интервале [0,1] при $0 \le b_{SIC} \le \beta$. Из 5.4 известно, что для заданного значения w, оценка b (5.23) может считаться нормально распределенной $N(M, D^2)$, с параметрами заданными в (5.28) и (5.29) . Эти параметры зависят от оценки среднеквадратичного отклонения w и объема выборки, т.е. количества объектов в калибровочном наборе, I. Следовательно, оценку $b_{SIC}=bC(I, w)$, тоже можно рассматривать как нормально распределенную. Поэтому вероятность первого события равна

$$P_1(z) = \text{Pr}\{b_{SIC} < \beta\} = \text{Pr}\{z < 1\} \approx \Phi\left(\frac{z - m_z}{\sigma_z}\right) \qquad (5.33)$$

где Ф – это кумулятивная функция нормального распределения, $\sigma_z = DC$, $m_z = MC$. Следовательно, можно легко посчитать значение $P_1(z)$ для любых заданных величин w и I.

Для того, чтобы определить вероятность второго события $P_2(z)$, необходимо иметь в виду, что эта вероятность имеет два граничных значения: $P_2(0)=1$ и $P_2(1)=0$. Фактически, эта величина показывает, с какой вероятностью истинное значение лежит внутри прогнозного интервала. Рассмотрим вероятность попадания истинного значения в малый интервал Δ. С точки зрения классической теории доверительных интервалов, эта вероятность больше, когда Δ расположен в середине интервала, и меньше, если Δ расположен на краю интервала. Поэтому, можно предположить, что вероятность $P_2(z)$ зависит от формы функции распределения погрешностей. Используя параметр формы распределения κ, введенный в 5.4, можно записать приближенную формулу для $P_2(z)$

$$P_2(z) = 2\frac{\Phi(\Xi) - \Phi(\Xi z)}{2\Phi(\Xi) - 1} \tag{5.34}$$

Очевидно, что существует хотя и очень сложная, но однозначная зависимость между параметром формы распределения κ и величиной w^2.

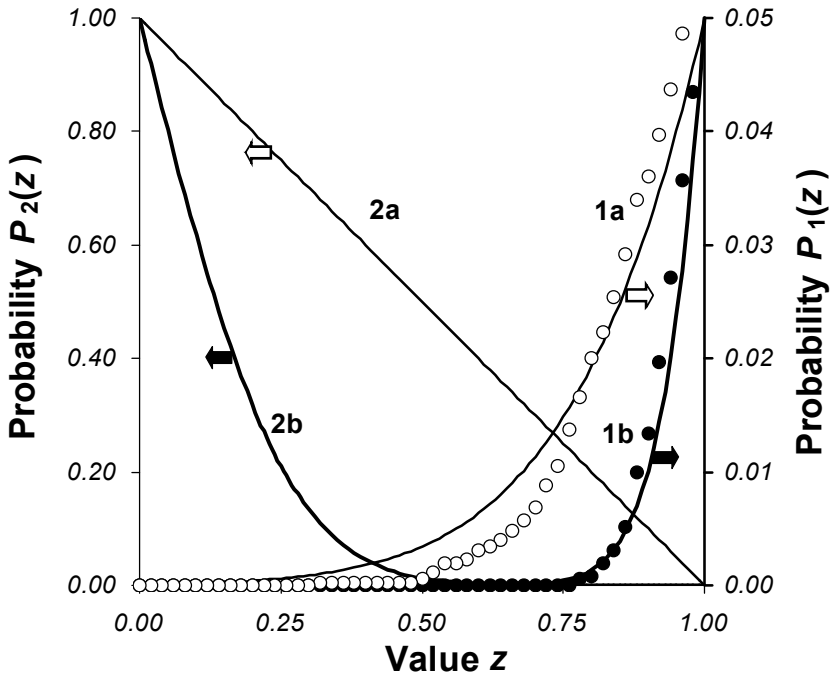


Рис. 5.4 Вероятности $P_1(z)$ и $P_2(z)$ для w=0.6, I=10 (a)и для I=100, w=0.2, (b) и соответствующие выборочные вероятности P_1 для I=10 (белые круги ○) и для I=100 (черные круги ●)

На Рис. 5.4 показаны два крайних случая распределения погрешности: равномерное распределение (т.е. κ=0.2, w=0.6) для n=10; и нормальное распределение, усеченное на 5σ для I=100. Вероятности $P_1(z)$ (правая ось Y) представлена кривой 1a (первый случай) и кривой 1b (второй случай). Соответствующие вероятности $P_2(z)$ (левая ось Y) представлены кривыми 2a (первый случай) и 2b (второй случай). На Рис. 5.4 так же представлены выборочные вероятности, рассчитанные на данных, получившихся в процессе моделирования для I=10 (белые круги) и для I=100 (черные круги). Эти результаты показывают, что выражение для вероятности $P_1(z)$ (5.33) дает хорошее приближение к результатам, полученным в статистическом моделировании.

Теперь необходимо перемножить вероятности (5.33) и (5.34), и проинтегрировать по z на интервале [0,1]. В результате мы получаем, что в худшем случае, когда w=0.6, I=10 вероятность перекрытия прогнозным интервалом точного значения равна 0.9983, а в лучшем случае, когда w=0.2, I=100, значение этой вероятности больше чем 0.9999999. Это показывает, что с практической точки зрения, т.е. полагая, что 0.9999 ≡ 1, мы можем считать, что неравенство (5.32) задает точные границы для неизвестного параметра β.

5.5. Результат главы 5.

В этой главе приведены основные понятия и свойства ПИО метода. Показано, как метод отвечает на вопросы, которые всегда являются важными для исследователя при анализе данных.

1. Ведено понятия и приведены основные свойства области допустимых значений параметров A. Показано, что область A является множественным аналогом точечной оценки неизвестных параметров в регрессионном анализе.
2. Показано, что оценка максимальной погрешности β определяет точность калибровки и задает границу воспроизводимости для всех объектов, которые подобны объектам из калибровочного набора.
3. Показано, что прогнозные интервалы, полученные методом ПИО, устанавливают индивидуальную неопределенность прогноза отклика для каждого нового объекта.

В основе ПИО метода лежит единственное предположение об ограниченности погрешностей. Именно это предположение приводит к оценке параметров модели в виде целой области A. В свою очередь ОДЗ A порождает интервальную оценку откликов y.

6. Классификация статуса объектов

Для анализа «качества» простроенной модели, и возможности использовать ее для прогноза, применяют различные методы проверки [68] – с помощью тестового набора, перекрестной проверки и т. п. Учитывая, что линейная калибровка (5.1) – это задача интерполяции, естественно полагать, что результат прогноза будет «хорошим» в том случае, когда проверочные, или новые объекты, будут «похожи» на объекты из обучающего набора. В этой главе будет рассмотрен вопрос классификации статуса объектов, т.е. взаимоотношений новых объектов моделирования со старыми объектами, использованными на стадии построения модели. Основной материал этой главы представлен в [52].

6.1. Характеристики статуса объектов

Для характеристики качества прогноза и формализации понятия «похожих» и «непохожих» объектов в рамках ПИО, можно ввести следующие понятия.

Пусть имеется ПИО модель, построенная с помощью набора калибровочных объектов $(\mathbf{x}_i, y_i)$, $i=1, \ldots, I$. Эта модель характеризуется своей областью допустимых значений (ОДЗ) A, которая определенна как пересечение полос (Определение 5.1, уравнение (5.4)). Рассмотрим некоторый (новый) объект, т.е. пару $(\mathbf{x}, y)$, с которым связана своя полоса $S(\mathbf{x}, y)$, определенная неравенствами

$$y - \beta \le \mathbf{x}^t \mathbf{a} \le y + \beta, \qquad (6.1)$$

Тогда взаимное положение полосы $S(\mathbf{x}, y)$ и области A характеризует статус объекта (Рис. 6.1). Заметим, что в качестве нового объекта $(\mathbf{x}, y)$ мы можем использовать как действительно новый объект, не входящий в калибровочный набор, так и один из объектов из этого набора. В последнем случае мы сможем оценить индивидуальное влияние каждого обучающего объекта на результаты калибровки.

Определение 6.1

Объект $(\mathbf{x}, y)$ называется – *внутренним*, если он не изменяет ОДЗ, т.е.

$$A \cap S(\boldsymbol{x}, y) = A$$

Иными словами $\left|\mathbf{x}^t\mathbf{a} - y\right| \le \beta$ для любого $\mathbf{a} \in A$ (Рис. 6.1 a,b).

Любой объект из калибровочного набора, по построению, является внутренним.

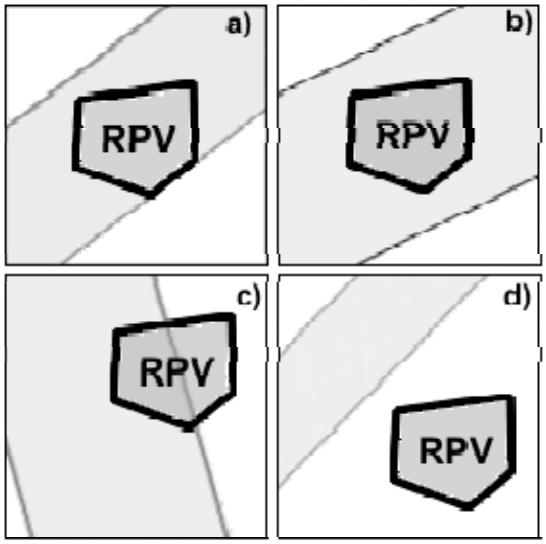


Рис. 6.1 Возможные положения полосы нового объекта по отношению к данной ОДЗ в пространстве параметров

Определение 6.2

Объект (**x**, y) из калибровочного набора называется граничным, если существует такой параметр $\mathbf{a} \in A$, что $\left|\mathbf{x}^t \mathbf{a} - y\right| = \beta$ (Рис. 6.1a).

Граничные объекты формируют ОДЗ, и, поэтому, являются наиболее важными среди объектов калибровочного набора. Очевидно, что если в калибровочном наборе оставить только эти объекты, то ОДЗ не изменится. Таким образом, все остальные образы из калибровочного набора можно считать в некотором смысле «избыточными» или «несущественными».

Определение 6.3

Объект (**x**, y) называется– *внешним*, если он уменьшает ОДЗ, т.е.

$$A \cap S(\boldsymbol{x}, y) \neq A$$

Иными словами, $\left|x^t \mathbf{a} - y\right| > \beta$ для некоторых значений параметра $\mathbf{a} \in A$. (Рис. 6.1 c,d).

Согласно определениям 6.1 и 6.3 все объекты делятся на внутренние и внешние. Однако, среди внешних объектов можно провести более детальное различие.

Определение 6.4

Объект (**x**, y) называется – *выбросом*, если он уничтожает ОДЗ, т.е.

$$A \cap S(\boldsymbol{x}, y) = \varnothing$$

Иными словами, $\left|\mathbf{x}^t \mathbf{a} - y\right| > \beta$ для любого $\mathbf{a} \in A$ (Рис. 6.1d).

Определение 6.5

Объект ($\mathbf{x}$, y) называется – *абсолютно внешним*, если для любого значения y

$$A \cap S(\mathbf{x}, y) \neq A$$

Существенным в этом определении является то, что определить, является ли объект абсолютно внешним или нет, можно в любом случае, даже тогда, когда соответствующее значение отклика y не известно.

Для того, чтобы понять смысл такой классификации, рассмотрим, какие изменения могут произойти с моделью, т.е. с ОДЗ A, при добавлении нового объекта в калибровочный набор. Если объект является внутренним, то при его добавлении, ОДЗ не изменится, т.е. $A_{I+1}=A_I$. Если объект является внешним, но не выбросом, то ОДЗ уменьшится, т.е. $A_{I+1}\subset A_I$, а добавленный объект станет граничным. Если объект является выбросом, то ОДЗ исчезает, т.е. $A_{I+1}=\varnothing$. (Здесь A_I обозначает ОДЗ, которая была построена с помощью калибровочного набора, состоящего из I объектов.)

Как следует из пояснений к определениям 6.1-6.5, классификация объектов проявляется не только во взаимном расположении полос и ОДЗ в пространстве параметров, но и во взаимном положении калибровочного и прогнозного интервалов. Напомним, что в результате применения ПИО метода мы получаем два интервала. Первый – это *калибровочный интервал* U, который характеризует меру неопределенности в калибровке (черный интервал на Рис. 6.2а),

$$U=[\, y-\beta\,,\, y+\beta\,] \qquad (6.2)$$

Второй – это *прогнозный,* или ПИО интервал V (5.17), который характеризует меру неопределенности в прогнозе (серый интервал на Рис. 6.2а); величина прогнозного интервала индивидуальна для каждого объекта. При этом не делается различия между объектами из калибровочного или тестового набора. Это дает нам возможность определить статус объекта в терминах V и U интервалов.

Утверждение 6.1

Для всех *калибровочных* объектов выполняется условие

$$V_i \cap U_i = V_i, \qquad i=1,...,I \qquad (6.3)$$

Утверждение 6.2

Объект является *внутренним* тогда и только тогда, когда

$$V \cap U = V \qquad (6.4)$$

Утверждение 6.3

Калибровочный объект ($V_i \subseteq U_i$) является *граничным* тогда и только тогда, когда

$$\max(V_i)=\max(U_i) \quad \text{либо} \quad \min(V_i)=\min(U_i) \tag{6.5}$$

Утверждение 6.4

Объект является *выбросом* тогда и только тогда, когда

$$V \cap U = \varnothing \tag{6.6}$$

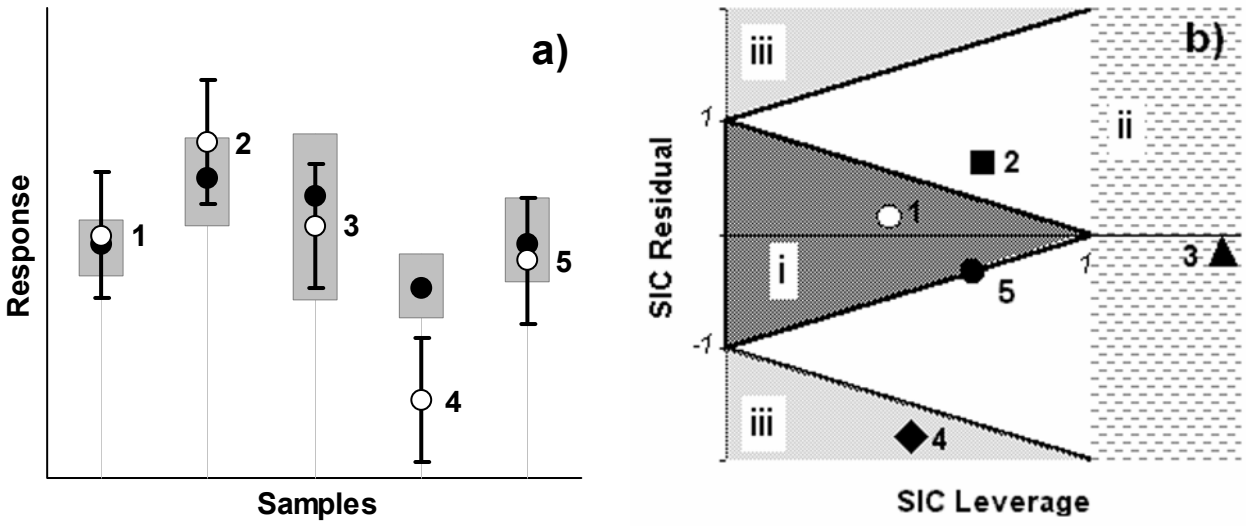
Утверждение 6.5

Объект является *абсолютно-внешним* тогда и только тогда, когда для любого значения y

$$V \cap U \neq V \tag{6.7}$$

Доказательство этих утверждений очевидно следует из определений 5.1 и 6.1-6.5.

Исследуя Рис. 6.2а, так же как раньше Рис. 6.1, можно заметить, что для разных объектов взаимное положение интервалов калибровки и предсказания различно. В некоторых случаях прогнозный интервал полностью лежит внутри интервала калибровки (объект 1, Рис. 6.2а). Это хороший случай, так как погрешность предсказания такого объекта меньше, чем погрешность моделирования, т.е. мы имеем дело с *внутренними* объектами.



Калибровочный интервал (черный), интервал предсказания (серый), (○) – референтное значение, (●) – предсказанное значение

Диаграмма статуса объектов. i – внутренние, ii- внешние, iia- абсолютно внешние, iii- выбросы

Рис. 6.2. Результаты ПИО прогноза.

Противоположная картина наблюдается для объекта 3, когда погрешность предсказания больше погрешности калибровки. Объект 3 не может быть калибровочным образцом; это однозначно проверочный объект. Согласно Утверждению 6.5, это *абсолютно внешний* объект. У объекта 2, который является *внешним*, прогнозный

интервал смещен относительно интервала калибровки. Это также объект из проверочного набора, причем, в данном случае, возможна ошибка в значениях отклика *y*. Объект 4 демонстрирует наихудший возможный случай, когда прогнозный и калибровочный интервалы не имеют ни одной общей точки; согласно Утверждению 6.4 — это *выброс*. Такая ситуация возможна тогда, когда объект по своей структуре, значениям **x**, сильно отличается от калибровочных объектов и не может быть предсказан данной моделью. Последний 5-ый объект представляет особый интерес, так как у него одна из границ прогнозного интервала совпадает с границей интервала калибровки. Если этот объект из калибровочного набора, то это *граничный* объект (Утверждение 6.2), т.е. один из тех объектов, которые образуют ОДЗ.

Таким образом, график Рис. 6.2а одновременно представляет результаты ПИО прогноза для каждого объекта (серый интервал), показывает взаимное положение между прогнозным интервалом и измеренным значением отклика *y*, а так же между калибровочным и прогнозным интервалами. Все это вместе и характеризует «качество» прогноза, и представляет интерпретацию статуса объектов уже в пространстве откликов.

6.2. Диаграмма статуса объектов (ДСО)

Для того, чтоб процедуру классификации объектов сделать максимально простой и наглядной, вводятся следующие величины.

Определение 6.6

Величина –

$$r(\mathbf{x}, y) = \frac{1}{\beta}\left(y - \frac{v^{+}(\mathbf{x}) + v^{-}(\mathbf{x})}{2} \right) \tag{6.8}$$

называется *ПИО-остатком.*

Величина r представляет разницу между центром прогнозного интервала $[v^{+}, v^{-}]$ и значением y (нормированную на β), поэтому r характеризует *смещение*.

Определение 6.7

Величина –

$$h(\boldsymbol{x}) = \frac{1}{\beta}\left(\frac{v^{+}(\boldsymbol{x}) - v^{-}(\boldsymbol{x})}{2} \right). \tag{6.9}$$

называется *ПИО-размахом.*

Величина h вычисляется как полуширина прогнозного интервала, деленная на максимальную погрешность, поэтому h характеризует β-нормализованную *воспроизводимость*.

Утверждение 6.6

Все калибровочные объекты удовлетворяют неравенству

$$\left| r(\mathbf{x}, y) \right| \le 1 - h(\mathbf{x}), \tag{6.10}$$

Доказательство. Из уравнений (5.3) следует, что для любого объекта из калибровочного набора,

$$| (\mathbf{x}_i, \mathbf{a}) - y_i | \le \beta,$$

Это неравенство справедливо для любого значения параметра **a** из ОДЗ В том числе это справедливо и для тех значений **a**, которые являются решением оптимизационной задачи (5.18). Таким образом, для любого объекта ($\mathbf{x}_i$, y_i) из калибровочного набора справедливо

$$\left| v_i^- - y_i \right| \le \beta \quad \text{и} \quad \left| v_i^+ - y_i \right| \le \beta$$

или

$$v_i^- \ge y_i - \beta \text{ и } v_i^+ \le y_i + \beta \tag{6.11}$$

Следовательно, прогнозный интервал V_i , построенный для калибровочного объекта ($\mathbf{x}_i$, y_i), находится внутри интервала калибровки, т.е. $V_i \subseteq U_i$. Вычитая из обеих частей неравенств (6.11) выражение

$$\frac{v_i^+ + v_i^-}{2}$$

и применяя Определения 6.6-6.7, введенные для h и r , получаем

$$r \le 1 - h \text{ и } r \ge h - 1$$

QED

Используя величины r и h, классификацию можно легко провести без явного конструирования ОДЗ [194].

Утверждение 6.7

Объект ($\boldsymbol{x}$, y) является *внутренним* тогда и только тогда, когда

$$\left| r(\boldsymbol{x}, y) \right| \le 1 - h(\boldsymbol{x}) \tag{6.12}$$

Согласно Утверждению 6.2 внутренним является объект, для которого прогнозный интервал лежит внутри интервала калибровки. Доказательство Утверждения 6.7 аналогично доказательству Утверждения 6.6.

Утверждение 6.8.

Калибровочный объект ($\boldsymbol{x}_i$, y_i) является *граничным*, тогда и только тогда, когда

$$|r(\boldsymbol{x}_i,y_i)| = 1 - h(\boldsymbol{x}_i) \qquad (6.13)$$

Доказательство. Понятие граничного объекта, вводится только для объектов из калибровочного набора, для которого $V_i \subseteq U_i$. Используя Утверждение 6.3, надо доказать, что совпадают либо правые, либо левые границы этих интервалов (либо эти интервалы просто равны, т.е. совпадают и правые и левые границы). Рассмотрим случай правой границы, для левой доказательства аналогичны. Используя определения (6.8), (6.9) для правой границы можно записать,

$$r(\boldsymbol{x}_i,y_i) = 1 - h(\boldsymbol{x}_i) \quad \Leftrightarrow \quad \frac{1}{\beta}\left(y_i - \frac{v_i^+ + v_i^-}{2}\right) = 1 - \frac{1}{\beta}\left(\frac{v_i^+ - v_i^-}{2}\right)$$

откуда после простых арифметических преобразований получаем

$$y_i - \frac{v_i^+ + v_i^-}{2} = \beta - \frac{v_i^+ - v_i^-}{2} \quad \Leftrightarrow \quad y_i + \beta = v_i^+$$

Что и требовалось доказать.

Утверждение 6.9

Объект ($\boldsymbol{x}$, y) является *выбросом* тогда и только тогда, когда

$$\left| r(\boldsymbol{x}, y) \right| > 1 + h(\boldsymbol{x}) \,. \qquad (6.14)$$

Доказательство. По Утверждению 6.4, объект является выбросом, если калибровочный и прогнозный интервалы не пересекаются, т.е., либо

$$\frac{1}{\beta}\left(y_i - \frac{v_i^+ + v_i^-}{2}\right) > 1 + \frac{1}{\beta}\left(\frac{v_i^+ - v_i^-}{2}\right)$$

либо

$$\frac{1}{\beta}\left(y - \frac{v_i^+ + v_i^-}{2}\right) < -1 - \frac{1}{\beta}\left(\frac{v_i^+ - v_i^-}{2}\right)$$

После простых алгебраически преобразований, получаем

$$v^- \geq y + \beta \;\text{ или }\; v^+ \leq y - \beta$$

Что и означает, что интервалы $U_i \cap V_i = \varnothing$.

Утверждение 6.10

Объект ($\boldsymbol{x}$, y) является *абсолютно-внешним* тогда и только тогда, когда

$$h(\boldsymbol{x}) > 1 \qquad (6.15)$$

Доказательство. Абсолютно-внешний, это такой объект, у которого ширина прогнозного интервала больше ширины интервала калибровки, т.е.

$$v_i^+(\mathbf{x}) - v_i^-(\mathbf{x}) > 2\beta \;\Leftrightarrow\; \frac{v_{\mathrm{i}}^+(\mathbf{x}) - v_{\mathrm{i}}^-(\mathbf{x})}{2\beta} > 1$$

Учитывая определение ПИО-размаха (6.9), получаем (6.15).

Используя Определения (6.8)-(6.9) и Утверждения 6.6-6.10, уравнения (6.12)-(6.15), можно построить диаграмму статуса объектов (ДСО), прототип которой показан на Рис. 6.2b. При любой размерности исходных данных (**X, *y***) и для любого числа параметров, ДСО является двумерной диаграммой, и это делает ее мощным инструментом в ММК. Утверждения 6.6-6.10 и формулы (6.12)-(6.15) делят плоскость «ПИО-остаток (*r*)» – «ПИО-размах (*h*)» на три области, каждая из которых соответствует одной из трех категорий объектов: внутренние (область i на Рис. 6.2b), внешние (вне области i) и выбросы (область iii).

Можно заметить, что треугольная форма области внутренних объектов на ДСО (Рис. 6.2b) напоминает обычную диаграмму влиятельности, когда объекты изображаются в координатах МНК-остатки

$$r_{\text{МНК}} = |y - \hat{y}|$$

относительно МНК-размаха [63]

$$h_{\text{МНК}} = \mathbf{x}^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\mathbf{x} = x_{\mathrm{i}}^2 \Big/ \sum_{i=1}^{I} x_{\mathrm{i}}^2$$

Пример такой диаграммы влиятельности приведен на Рис. 8.5b для модели определения следовых концентраций нефти в воде. Сходство между диаграммами статуса и влиятельности вытекает из хорошо известного статистического соотношения [67],

$$RMSEC^2 \approx SEC^2 + BIAS^2, \qquad (6.16)$$

которое связывает погрешность (среднеквадратичный остаток калибровки, *RMSEC* (3.1), воспроизводимость (стандартную погрешность, *SEC*) и смещение (систематическую погрешность, *BIAS*)

При этом, квадрат стандартной погрешности является оценкой дисперсии остатков моделирования и вычисляется как

$$SEC^2 = (I - K - 1)^{-1} \sum_{i=1}^{I} \left(\hat{y}_{\mathrm{i}} - y_{\mathrm{i}} - \overline{(\hat{\mathbf{y}} - \mathbf{y})}\right)^2, \qquad (6.17)$$

Смещение, которое характеризует систематическую составляющую погрешности, вычисляется по формуле

$$BIAS = \frac{\sum_{i=1}^{I}(\hat{y}_i - y_i)}{I} = \overline{(\hat{\mathbf{y}} - \mathbf{y})}.$$

Точнее, для того чтобы формула (6.16) была справедливой, необходимо в (6.17) делить не на (*I*-*K*-1), а на *I*.

В методе ПИО, в котором величина β является погрешностью, ПИО-размах *h* характеризует (нормализованную) воспроизводимость, а ПИО-остаток *r* отвечает за (нормализованное) смещение, уравнение (6.16) может быть представлено в виде

$$\beta^2 = \beta^2 h^2(\boldsymbol{x}) + \beta^2 r^2(\boldsymbol{x}, y), \qquad (6.18)$$

С другой стороны, необходимо признать существенное отличие между уравнениями (6.16) и (6.18). Оно состоит в том, что последнее уравнение справедливо для каждого объекта, тогда как уравнение (6.16) имеет смысл только для всей совокупности объектов, т.е. в среднем.

Необходимо отметить, что концепция влиятельности объектов широко используется в стандартном регрессионном подходе. Мера влиятельности объекта может быть использована для определения выбросов при вычислении регрессионных коэффициентов. Объекты, обладающие высокой мерой влиятельности, должны рассматриваться и анализироваться с особой тщательностью, так как, в случае удаления такого объекта из обучающей выборки, регрессионная модель существенно меняется. В обычной множественной регрессии, в качестве меры влиятельности объекта часто используют расстояние Кука (Cook) [240], хотя возможны и другие подходы [241]. Расстояние Кука D_i для объекта *i*, входящего в регрессионную модель (5.1) вычисляется по формуле

$$D_i = \frac{(\mathbf{a}_{(-i)} - \mathbf{a})^t \mathbf{X}^t \mathbf{X} (\mathbf{a}_{(-i)} - \mathbf{a})}{s^2 (J+1)}, \qquad (6.19)$$

где **a** и $\mathbf{a}_{(-i)}$– это оценки регрессионных коэффициентов, вычисленные для всех *I* объектов, и для случая, когда объект *i* исключен из обучающей выборки, *J* – количество столбцов в матрице **X**. Уравнение (6.19) может быть переписано в эквивалентном виде

$$D_i = \frac{1}{(J+1)} \frac{h_i}{1 - h_i} r_i^2 . \qquad (6.20)$$

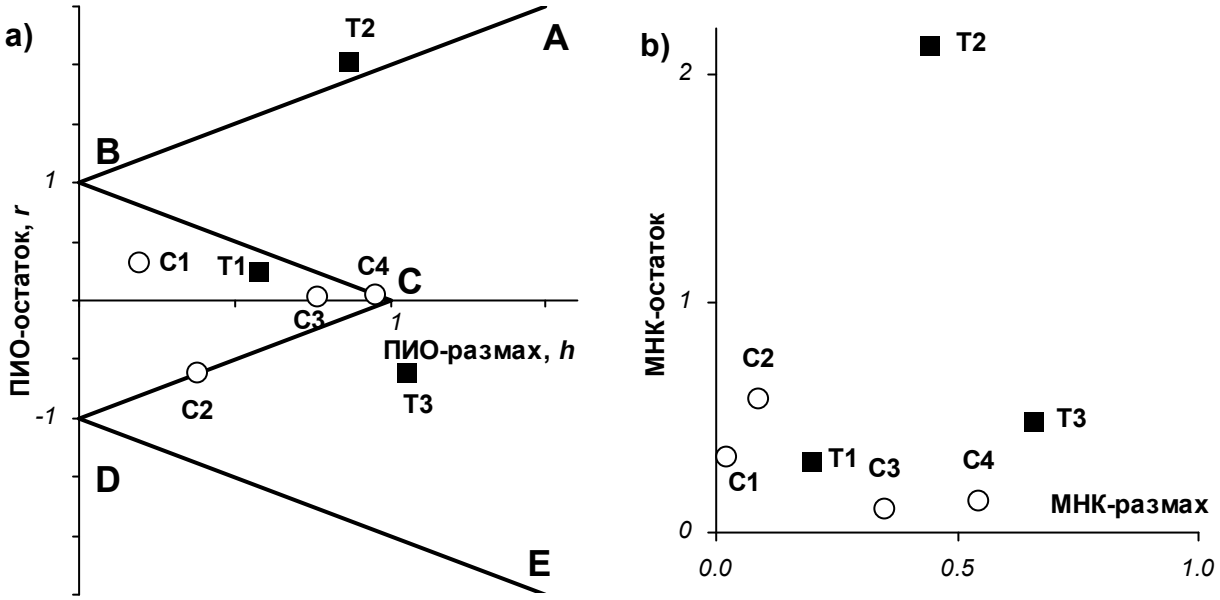
где h_i – это величина размаха, вычисленного для объекта *i* (2.6). Откуда видно, что D_i является функцией от «стьюдентизированного» остатка (т.е. остатка деленного на свою стандартную ошибку) и размаха.

Такая же мера влиятельности используется и для проекционных регрессионных методов, с заменой **X** на **T** и **a** на **q**. Однако в этом случае, переход от формулы (6.19) к формуле (6.20) не так очевиден.

Таким образом, треугольная форма области внутренних объектов на ДСО и графиков влиятельности отражает тот факт, что для того чтобы объект был влиятельным для данной модели (т.е. в большей степени, чем другие объекты, влиял на определение величин регрессионных коэффициентов), такому образцу не достаточно иметь только большей размах или только больший остаток, чем у остальных образов. Наиболее влиятельными будут объекты, у которых велики оба эти показателя.

6.3. Классификация объектов. Одномерный модельный пример

Возвращаясь к простейшему примеру из Раздела 4.2, посмотрим, как для него выглядит ДСО.



ПИО-диаграмма статуса объектов МНК-диаграмма влиятельности объектов

Рис. 6.3. Определение статуса объектов в одномерном модельном примере.
○- обучающие объекты, ■- проверочные объекты

Получив прогнозные интервалы (4.12), легко вычислить ПИО-остаток r, и ПИО-размах h (уравнения (6.8) и (6.9)).

Для нашего примера значения r и h приведены в столбцах 10 и 11 Таб. 6.1 и изображены на *диаграмме статуса объектов* (ДСО) на Рис. 6.3а в координатах (h, r). На этой диаграмме калибровочные объекты показаны незакрашенными кругами, а

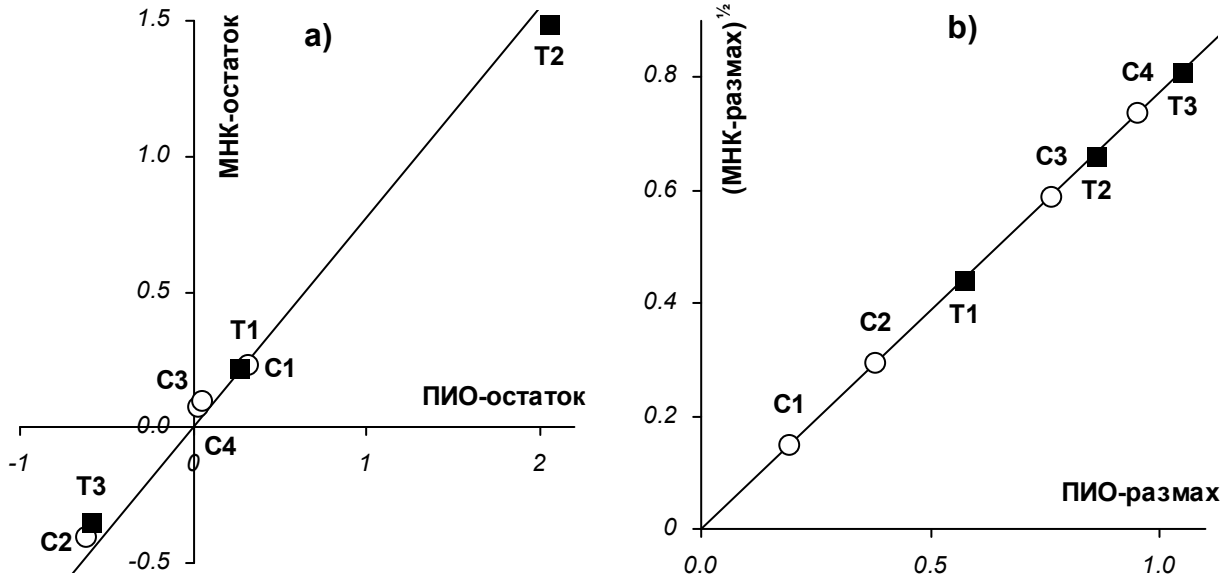
проверочные объекты – закрашенными квадратами, так же, как и на Рис. 4.3. Жирная ломанная ABCDE ограничивает области различных статусов объектов. Форма этой ломаной определяется двумя фундаментальными неравенствами ((6.12) и (6.14)), связывающими величины h и r.

Таб. 6.1 Модельные данные и результаты их обработки (Продолжение Таб. 4.1)

Объекты	v^-	v^+	h	r	$\|r\|+h$
0	8	9	10	11	12
C1	0.92	1.19	0.19	0.31	0.51
C2	1.85	2.38	0.38	–0.62	1.00
C3	3.70	4.76	0.76	0.03	0.79
C4	4.62	5.95	0.95	0.05	1.00
T1	2.77	3.57	0.57	0.26	0.83
T2	4.16	5.36	0.86	2.05	2.91
T3	5.08	6.55	1.05	–0.60	1.64

Из диаграммы видно, что все объекты из калибровочного набора лежат внутри треугольника BCD (их статус – внутренние, для них $|r|+h \le 1$, см. столбец 12, Таб. 6.1), причем объекты C2 и C4 расположены на его границах (их статус – граничные, для них $|r|+h = 1$). Проверочный объект T1 также попал в треугольник внутренних объектов, т.к. $|r|+h = 0.83<1$. Объект T2 лежит ниже прямой DE, что свидетельствует о том, что он является выбросом. Для него $|r|-h = 2.91 > 1$ (см. неравенство (6.14)). Объект T3 – это внешний объект, причем из диаграммы видно, что ни при каком значении своего остатка r, он не попадет в область внутренних объектов. Для него $h = 1.05 > 1$. Это свидетельствует о том, что значение его переменной x содержит в себе новую, существенную информацию, которая отсутствует в модели калибровки, т.е. это *абсолютно-внешний объект* (6.15).

Следующие два графика (Рис. 6.4a и b), так же свидетельствуют о тесной связи между характеристиками объектов в МНК и ПИО методах.



Сравнение МНК и ПИО остатков

Сравнение МНК и ПИО размахов

Рис. 6.4 Характеристики объектов в одномерном модельном примере.
○- обучающие объекты, ■- проверочные объекты

Из Рис. 6.4a видна сильная корреляция (R^2=0.999) между МНК и ПИО остатками. Связь между размахами более сложная (Рис. 6.4b), но и здесь наблюдается корреляция (R^2=1.000) между корнем квадратным из МНК размаха и ПИО размахом. Такое соотношение с очевидностью вытекает из определений этих величин. Размах в методе ПИО пропорционален ширине прогнозного интервала (4.12), тогда как в методе наименьших квадратов размах пропорционален дисперсии прогноза [68], которая определяет ширину доверительного интервала, пропорциональную корню квадратному из дисперсии. Разумеется, в более сложных задачах связь между МНК и ПИО характеристиками будет не такой простой, но основная тенденция сохраняется. Проблема схожести и различия МНК и ПИО методов, сравнение прогнозных ПИО интервалов и доверительных МНК интервалов, подробно рассмотрено в Разделе 9.5 и в работах [32, 223].

6.4. Классификация новых объектов

Обычно, когда модель ММК применяется к новым объектам, соответствующие значения y неизвестны. Поэтому нельзя вычислить ПИО-остаток, r (6.8), но всегда можно определить ПИО-размах, h (6.9). Легко видеть, что если размах нового объекта больше,

чем единица (h>1, область ii на Рис. 6.2b), то этот объект никогда не может быть отнесен к типу внутренних, ни при каком значении y, т.е. он является абсолютно-внешним (6.15).

Таким образом, даже тогда, когда информации о референтных значениях **y** отсутствует, тем не менее, можно сказать, что для абсолютно-внешних объектов ширина прогнозного интервал будет больше ширины калибровочного интервала. Используя определение ПИО-размаха (6.9), можно записать

$$h(\lambda \boldsymbol{x})=|\lambda| h(\boldsymbol{x}), \tag{6.21}$$

Отсюда видно, что для любого объекта из калибровочного набора $\boldsymbol{x}_i$, вектор $\lambda\boldsymbol{x}_i$ будет абсолютно внешним образцом тогда и только тогда, когда $|\lambda|>1/h(\boldsymbol{x}_i)$. Таким образом, для любого калибровочного набора можно сконструировать область в пространстве предикторов (счетов), за пределами которой располагаются абсолютно внешние объекты. Приведем утверждение, определяющее такую область.

Утверждение 6.11.

Пусть D – это область в пространстве предикторов, образованная всеми возможными линейными комбинациями взвешенных векторов предикторов (или счетов) $\boldsymbol{x}_i$ из калибровочного набора, такими что

$$\mathbf{x}=\sum_{i=1}^{I}\frac{\lambda_i}{h(\mathbf{x}_i)}\mathbf{x}_i, \quad \sum_{i=1}^{I}|\lambda_i|\le 1 \tag{6.22}$$

Тогда все абсолютно внешние объекты будут расположены вне этой области.

Доказательство следует из утверждения 6.10 и формулы (6.21).

Формулы (6.22) и (6.9) определяют алгоритм для вычисления координат точек, формирующих область абсолютно внешних образов. Для каждого объекта $\mathbf{x}_i$ из калибровочного набора в пространстве предикторов (счетов) вычисляются координаты точки $\mathbf{x}_i^b$ образующие границу области по следующей формуле

$$\mathbf{x}_i^b = \mathbf{x}_i 2\beta\left(v^+(\mathbf{x}_i)-v^-(\mathbf{x}_i)\right) \tag{6.23}$$

Для двумерной задачи, такую область легко построить в явном виде (Рис. 6.5). При этом каждая точка (объект), соединяется прямой линией с центром. Координаты граничных точек вычисляются по формуле (6.23) и они располагаются симметрично от центра на этой прямой линии. Соединив все построенные таким образом точки, получаем многоугольник, за пределами которого расположена область абсолютно внешних образов.

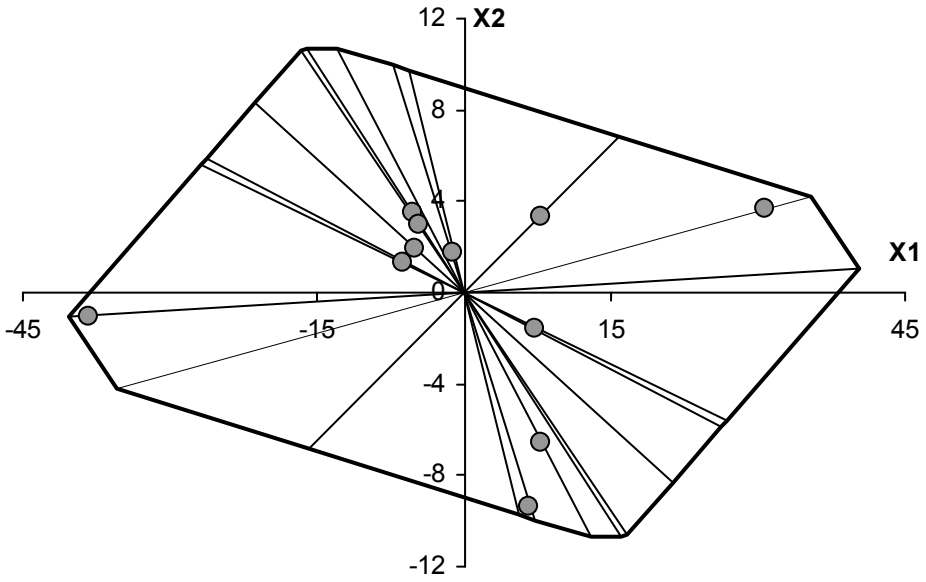


Рис. 6.5 Двумерный случай. ● - объекты из калибровочного набора, черная ломаная линия образует границу абсолютно вешних объектов

Здесь мы вновь видим определяющую роль граничных объектов. Именно они задают вершины многоугольника, образующего область абсолютно внешних объектов. Другие объекты из калибровочного набора задают точки, лежащие на прямых, соединяющих эти вершины. Подробно техника построения разных областей в пространстве предикторов (или счетов) рассмотрена в главе 8.

6.5. Результаты главы 6

В этой главе показано, что в общем случае, для решения задач ММК, ПИО подход позволяет ввести новый метод классификации объектов. Он базируется на определениях 6.1-6.5 и утверждениях 6.1-6.5. Для определения статуса объекта нет необходимости в явном виде строить ОДЗ в пространстве параметров. Такое построение является очень сложной задачей, особенно, если пространство параметров имеет размерность больше чем два или три, а количество объектов в калибровочном наборе, а следовательно и количество полос, формирующих эту область велико. Для ПИО классификации достаточно построить диаграмму статуса объектов, которая независимо от исходной размерности задачи, представляется в двумерном пространстве, т.е. на плоскости. Позиция каждого объекта на диаграмме статуса позволяет определить, подобен ли изучаемый объект объектам из калибровочного набора, и тем самым, задает разумные

границы применимости построенной калибровки, что крайне важно для задач формального моделирования

7. Программная реализация ПИО метода

Важно не только разработать теоретические аспекты метода, но и воплотить его в виде удобного инструмента, с помощью которого исследователь мог бы легко анализировать результаты сложных многофакторных физических экспериментов. Для этого нужно детально разработать и объединить все необходимые вычислительные алгоритмы и на их базе создать программу, которая будет служить таким инструментом. В этой главе рассматриваются алгоритмы и программа, реализующая ПИО метод.

Одним из достоинств ПИО метода является то, что в его основе лежат хорошо известные численные алгоритмы. Это проекционные методы (МГК и ПЛС) для регуляризации исходной задачи, а так же, симплекс-метод, для решения линейных оптимизационных задач. Проекционные методы подробно рассмотрены в разделах 2.1 и 3.1, а алгоритмы приведены в приложении (разделы 13.1 и 13.3). Задача линейной оптимизации и ее использование в методе ПИО рассмотрены в разделе 7.1, а симплекс-алгоритм описан в приложении (раздел 13.4). Оригинальным является алгоритм приведения исходной задачи (5.1) к задаче линейного программирования в каноническом виде (раздел 7.2).

7.1. Задача линейного программирования. Основные понятия.

Задачи оптимизации нечасто встречаются при обработке результатов физических экспериментов, в отличие от экономических задач. Однако, как уже отмечалось, и ПЛС метод, являющийся сейчас одним из базовых инструментов хемометрики, был в свое время позаимствован из эконометрики [3]. Для объяснения алгоритма ПИО метода необходимо напомнить некоторые базовые понятия и определения из области исследования операций [239].

В самом общем виде *задача оптимизации* ставится следующем образом: даны множество $A \subset \mathrm{R}^J$ и числовая функция $f(\mathbf{a})$, определенная на множестве A; требуется найти точку максимума или минимума функции f на A. Принято записывать задачу на максимум в виде

$$f(\mathbf{a}) \to \max \ , \ \mathbf{a} \in A \ , \qquad (7.1)$$

при этом f называют *целевой функцией*, A*- допустимым множеством*, любой элемент $\mathbf{a} \in A$ *– допустимой точкой* задачи (7.1).

Решением задачи (7.1), или точкой максимума функции f на множестве A, называется точка $\mathbf{a}^* \in A$, такая что

$$f(\mathbf{a}^*) \geq f(\mathbf{a}) \text{ при всех } \mathbf{a} \in A \qquad (7.2)$$

Эта запись эквивалентна следующей

$$f(\mathbf{a}^*) = \max_{\mathbf{a} \in A} f(\mathbf{a})$$

Величина

$$f^* = \sup_{\mathbf{a} \in A} f(\mathbf{a}) \qquad (7.3)$$

т.е. точная верхняя грань f на A, называется *значением задачи* (7.1). Отметим, что f^* существует всегда.

По аналогии с (7.1) можно записать задачу минимизации функции f на множестве A в виде

$$f(\mathbf{a}) \to \min \ , \ \mathbf{a} \in A \qquad (7.4)$$

Заменяя в (7.2), (7.3) знак «≥» на «≤» и символ «sup» на «inf», получаем определения для понятий решения и значения задачи (7.4).

Очевидно, что задача (7.4) эквивалентна, в смысле совпадения множества решений, задаче

$$-f(\mathbf{a}) \to \max \ , \ \mathbf{a} \in A$$

Поэтому в теоретических рассмотрениях достаточно ограничиться рассмотрением лишь одного из этих двух типов оптимизационных задач. В дальнейшем будем рассматривать преимущественно задачу на максимум (7.1).

Задача называется задачей *условной оптимизации*, если A – собственное подмножество R^J. Простейший класс условных оптимизационных задач образуют задачи *линейного программирования* (ЛП). Так называется задача вида (7.1) или (7.4), где целевая функция f линейна, а допустимое множество A задается системой из конечного числа линейных неравенств и уравнений

$$f(\mathbf{a}) = \sum_{j=1}^{J} c_j a_j \to \max \ \ \mathbf{a} \in A$$

$$A = \{\mathbf{a} \in R^J \ \Big| \sum_{j=1}^{J} x_{ij} a_j \leq b_i \ , \ i = 1,...,k \ \ \sum_{j=1}^{J} x_{ij} a_j = b_i \ , i = k+1,...,I\}$$

где c_j, x_{ij}, b_i – заданные числа. С формально-математической точки зрения, ПИО модель как раз и является задачей линейного программирования.

Геометрическая интерпретация.

Задачи ЛП с двумя переменными можно дать наглядную геометрическую интерпретацию, позволяющую увидеть ряд важных общих свойств. Рассмотрим задачу вида

$$c_1 a_1 + c_2 a_2 \rightarrow \max$$
$$A = \left\{ \mathbf{a} : \sum_{j=1}^{2} x_{ij} a_j \le b_i ,\ i = 1,...,I \right\} \qquad (7.5)$$

При J=2 множество A можно рассматривать как пересечение полуплоскостей Рис. 7.1 (в общем случае – это пересечение полупространств).

Рассмотрим целевую функцию

$$c_1 a_1 + c_2 a_2 = \gamma \qquad (7.6)$$

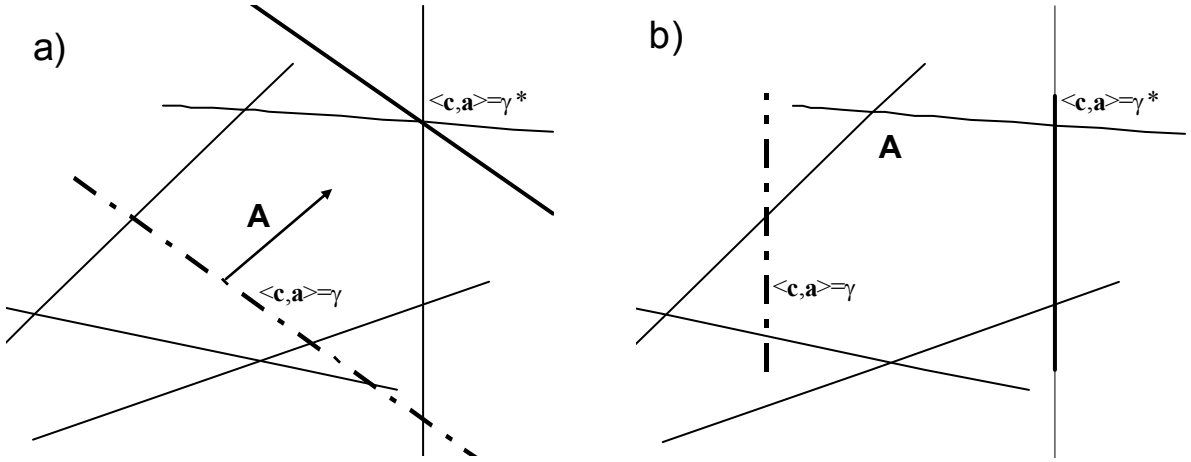
Это уравнение, при различных значениях γ описывает семейство параллельных прямых (в общем случае, при J>2, это семейство параллельных гиперплоскостей). Вектор **c** направлен в сторону возрастания целевой функции. Рассмотрим некоторую точку $\mathbf{a}_0 \in A$. Ей соответствует значение целевой функции $c_1 a_{01} + c_2 a_{02} = \gamma_0$. Теперь будем перемещать прямую (7.6) в направлении **c**, т.е. в направлении увеличения γ. Задачу (7.5) можно переформулировать следующим образом: найти максимальное значение γ^* среди всех γ, при котором прямая (7.6) имеет непустое пересечение с допустимым множеством A. Пересечение прямой $c_1 a_1 + c_2 a_2 = \gamma^*$ с областью A и есть решений задачи (7.5), а γ^* - ее значение. Так же, при J=2, легко изобразить случаи, когда множество A неограниченно, но решение существует; когда решение существует, но не единственно; когда A неограниченно и решение неограниченно.

В общем случае, можно указать следующие свойства задач ЛП:

1. Допустимое множество задач ЛП является многогранным. (Многогранные множества в R^J как раз и определяются как множество решений систем линейных неравенств). Это множество имеет конечное число вершин (угловых точек).

2. Если решение существует и единственно, то им является одна из угловых точек допустимого множества (Рис. 7.1a). Если же решений много (Рис. 7.1b), то среди них обязательно имеется угловая точка.

3. Отсутствие допустимых решений происходит тогда, когда ограничения одновременно выполняться не могут, т.е. введенные ограничения противоречивы и область допустимых решений A=∅.



Решение существует и единственно Решение существует, но не единственно

Рис. 7.1. Пример решение задачи линейного программирования (7.5)

Таким образом, чтобы найти решение задачи ЛП, если оно существует, достаточно перебрать все угловые точки (вершины) допустимого множества и выбрать ту, где целевая функция принимает максимальное значение.

Численный метод решения задачи ЛП

Для использования общего метода решения задач ЛП любая исходная модель предварительно должна быть представлена в некоторой форме, которую называют *канонической формой линейной оптимизационной модели* [238]. При этом

- все ограничения записываются в виде равенств с неотрицательной правой частью;
- значения всех переменных модели неотрицательны;
- целевая функция подлежит максимизации;

То есть, каноническая задача ЛП имеет вид

$$\begin{aligned} &\sum_{j=1}^{J} c_j a_j \to \max \ , \\ &\sum_{j=1}^{J} x_{ij} a_j = b_i \ , i = 1,...,I \\ &a_j \geq 0 \ , j = 1,...,J \end{aligned} \tag{7.7}$$

или в более компактной векторной форме

$$\begin{aligned} &\langle \mathbf{c}, \mathbf{a} \rangle \to \max, \\ &\mathbf{X}\mathbf{a} = \mathbf{b}, \\ &\mathbf{a} \geq 0 \end{aligned} \tag{7.8}$$

При этом **c** называется вектором коэффициентов целевой функции, **X** – матрицей ограничений, **b** – вектор правых частей ограничений.

Любая задача ЛП может быть приведена к каноническому виду, для этого используют следующие стандартные приемы.

Для ограничений

1. Исходные ограничения, записанные в виде неравенств типа ≤ (≥), можно представить в виде равенств, прибавляя остаточную переменную к левой части ограничения (вычитая избыточную переменную из левой части).

2. Правую часть равенства можно сделать неотрицательной, умножая обе части на –1.

Для переменных.

1. Любую переменную, не имеющую ограничения в знаке, можно представить как разность двух неотрицательных переменных $y_i = y_i' - y_i''$ где $y_i', y_i'' \geq 0$. Такую подстановку следует использовать во всех ограничениях, которые содержат исходную переменную y_i, а также в выражении для целевой функции.

2. При необходимости в левую часть каждого уравнения добавляют искусственные переменные. Это необходимо тогда, когда уравнение не содержит «очевидного» начального базисного решения.

Для целевой функции

Целевая функция линейной оптимизационной модели, представленной в канонической форме, подлежит максимизации. Максимизация некоторой функции эквивалентна минимизации той же функции, взятой с противоположным знаком. Эквивалентность означает, что при одной и той же совокупности ограничений оптимальные значения переменных будут одинаковыми.

Основным численным методом решения задач ЛП является *симплекс-метод*. В его вычислительной схеме реализуется упорядоченный процесс, при котором, начиная с некоторой исходной допустимой угловой точки, осуществляется последовательный переход от одной допустимой угловой точки к другой до тех пор, пока не будет найдена точка соответствующая оптимальному решению. Решение, соответствующее исходной точке называют начальным решением. Выбор каждой последующей угловой точки определяется следующими двумя правилами.

1. Каждая последующая угловая точка должна быть смежной с предыдущей, т.е. этот переход осуществляется по границам (ребрам) пространства решений. Перед новым

переходом каждая из полученных точек проверяется на оптимальность. Переход производится в направлении наибольшего изменения целевой функции.

2. Обратный переход к предшествующей угловой точке не может производиться.

Алгебраическое представление угловой точки

Для определения угловых точек алгебраическим путем, приравнивают нулю такое количество переменных, которое равно разности между количеством неизвестных и числом уравнений. В этом состоит сущность свойства *однозначности* экстремальных точек. Так любая точка внутренней области пространства решений вообще не имеет ни одной нулевой переменной, а любая не угловая точка, лежащая на границе, всегда имеет лишь одну ненулевую переменную.

Будем считать, что линейная модель канонической формы содержит m уравнений и n ($m \le n$) неизвестных (правые части ограничений – неотрицательные). Тогда все допустимые угловые точки определяются как все однозначные неотрицательные решения системы m уравнений, в которых n-m переменных равны нулю.

Однозначные решения такой системы уравнений называются *базисными решениями*. Если базисное решение удовлетворяет требованию неотрицательности правых частей, оно называется *допустимым базисным решением*. Переменные, имеющие нулевое значение, называются *небазисными переменными,* остальные – *базисными переменными*.

Максимальное число итерации при использовании симплекс-метода равно максимальному числу базисных решений задачи ЛП, представленной в каноническом виде, и не превышает

$$C_m^n = n!/\left[(n-m)!\,m!\right]$$

Смежные угловые точки отличаются только одной переменной в каждой группе (нулевых и ненулевых переменных). Поэтому, при реализации симплекс-метода, последующую угловую точку можно определить путем замены одной из текущих небазисных переменных (нулевых) переменных текущей базисной переменной. Поэтому вводятся понятия *включаемой переменной* – это небазисная в данный момент переменная, которая будет включена в базис на следующей итерации (при переходе к смежной экстремальной точке); *исключаемая переменная* – это та базисная переменная. которая на следующей итерации подлежит исключению из множества базисных переменных. Подробно алгоритм симплекс-метода описан в приложении (раздел 13.4)

7.2. ПИО метод как задача линейного программирования

Формирование системы ограничений, допустимой области *A* в каноническом виде.

Вернемся к модели линейной калибровки (5.1), состоящей из I объектов, и J неизвестных параметров. Без ограничения общности можно считать, что матрица предикторов **X** имеет полный ранг. Если это не так, то сначала задача (5.1) регуляризируется, например, с помощью одного из проекционных методов, и вместо задачи (5.1) рассматривается задача (5.15) .Следовательно, можно считать, что rank **X**=J. Так же будем полагать, что максимальная погрешность β известна. Тогда каждое двустороннее неравенство типа (5.3), порождаемое образцом из обучающего набора, приводит к двум неравенствам в системе ограничений задачи ЛП, т.е. вместо (5.3), для i=1,..,I записываем

$$\begin{aligned} \mathbf{x}_i^t\mathbf{a} &\geq y_i - \beta \\ \mathbf{x}_i^t\mathbf{a} &\leq y_i + \beta \end{aligned} \qquad (7.9)$$

и для определения допустимой области A получаем систему из $2I$ неравенств, которые надо привести к каноническому виду (7.7) или (7.8) используя правила, описанные в разделе 7.1.

Отметим два важных обстоятельства. Во-первых, если правая часть неравенства отрицательна, то неравенство умножается на –1. Во-вторых, неравенства, с помощью введения дополнительных переменных, преобразуются в равенства. К левой части каждого неравенства вида «≤» прибавляется остаточная переменная $sr_i \geq 0$. Из левой части каждого неравенства вида «≥» вычитается избыточная переменная $ss_i \geq 0$, а так же прибавляется искусственная переменная $r_i \geq 0$, для того чтобы найти начальное приближение, используемое на Этапе 1 симплекс-метода (раздел 13.4). В итоге, допустимая область A**,** это множество таких значений вектора параметров **a**, которые удовлетворяют системе уравнений:

$$\begin{aligned} &\sum_{j=1}^{J} x_{ij}a_j + sr_{i1} = b_{i1}\,,\ sr_{i1} \geq 0 \quad ,\quad i1 = 1,...,I1 \\ &\sum_{j=1}^{J} x_{ij}a_j - ss_{i2} + r_{i2} = b_{i2}\,,\ ss_{i2} \geq 0\,,\ r_{i2} \geq 0\ ,\ i2 = 1,...,I2 \\ &i = 1,...,I \quad ,\quad 2I = I1 + I2 \end{aligned} \qquad (7.10)$$

где b_{i1} и b_{i2} – это правые части неравенств (7.9); x_{ij} – элементы матрицы **X**; a_j неизвестные параметры; индекс *i1*, отвечает за неравенства вида «≤», а *i2* за неравенства вида «≥». Общее количество неизвестных параметров в системе (7.10) $L= 3I+I2$.

Единственно чего не хватает системе (7.10) для того, чтобы представлять задачу ЛП в каноническом виде, это условия на неотрицательность переменных, что все $a_j \geq 0$. С содержательной точки зрения, любые значения a_j могут быть как положительные, так и отрицательные. Если воспользоваться стандартным приемом, и ввести вместо каждой переменной a_j две новых переменных $a_j = a'_j - a''_j$ где $a'_j, a''_j \geq 0$, то мы значительно увеличим размерность исходной задачи. С другой стороны, так как область *A* ограничена, то ее можно заключить в эллипсоид вида (5.10).

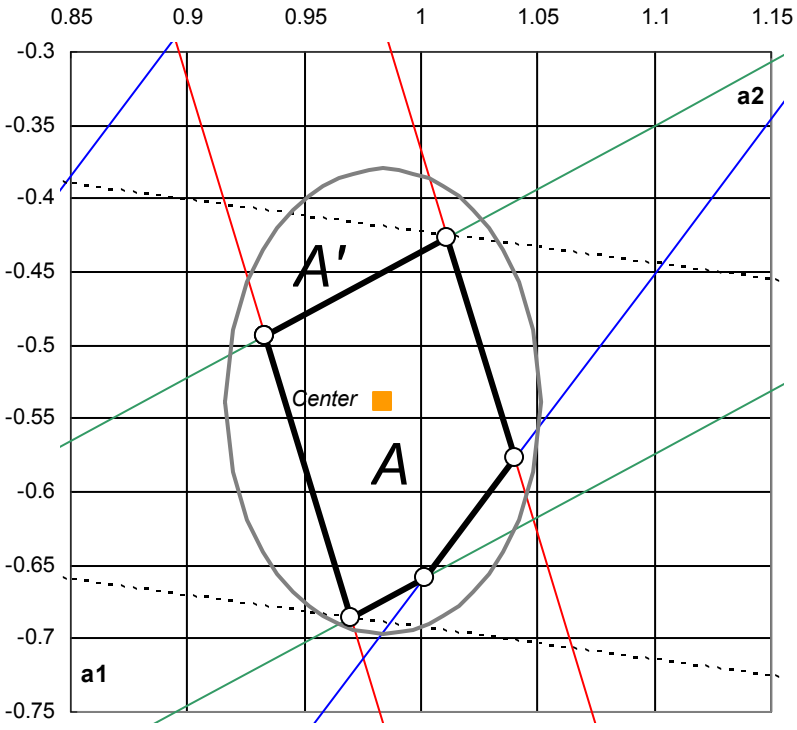


Рис. 7.2 Двумерный случай. Область *A*- ОДЗ ; ограничивающий эллипс *A'*.

Двумерный случай изображен на Рис. 7.2. Вычислив положение центра и величину главных осей эллипсоида, его можно сдвинуть так, чтобы весь эллипсоид, а с ним и ОДЗ располагались в положительной части пространства. Это гарантирует условие $a_j \geq 0$. После того как задача ЛП решена и найдены экстремумы, для получения правильных результатов, эллипсоид, а с ним и все параметры сдвигаются в обратную сторону. Формулы для вычисления центра и главных осей эллипсоида приведены в приложении

(раздел 13.5). После такого преобразования, задача ЛП принимает канонический вид и готова для вычислений.

7.3. Основные свойства, возможности, требования и ограничения программы SIC

ПИО метод можно непосредственно использовать для анализа экспериментальных данных только в том случае, когда матрица **X** невырожденная. В современных задачах многомерного анализа многоканальных данных такая ситуация крайне редкая. Поэтому, ПИО метод применяется не к исходной задачи (5.1), а к задаче (5.15), которая получается после регуляризации матрицы **X** с помощью одного из проекционных методов. Это либо МГК, либо ПЛС, по выбору пользователя. Вся программа ПИО-метода состоит из следующих основных алгоритмов:

1. Предварительная подготовка данных
2. Проекционные регрессионные методы (МГК, РГК, ПЛС 1, ПЛС 2)
3. Процедуры приведения исходной задачи к стандартной форме линейной оптимизационной модели.
4. Стандартная процедура Симплекс-метода для решения линейной оптимизационной задачи.
5. Вычисление результатов, построения ДСО.

Для программной реализации метода ПИО была разработана программа, которая называется SIC – Simple Interval Calculations. SIC – это программа, которая работает под управлением системы Excel, входящей в стандартный пакет Microsoft Office. Используя этот инструмент, можно решать задачи моделирования, предсказания, и определения статуса объектов применительно к линейным моделям. Вся входная и выходная информации представляется как таблицы и графики в Excel.

С помощью SIC можно получить следующую информацию:

- результаты интервального прогноза[v^{-} , v^{+}];
- точечную регрессионную оценку для значений откликов $\hat{y}$ с помощью выбранного регрессионного метода (РГК, ПЛС);
- оценки параметра β: $b_{\min}$ и b_{SIC}
- ПИО-остаток и ПИО-размах

Информация выводится как в числовом, так и в графическом виде.

Дополнительно, используя программу SIC можно воспользоваться ее функциями, и вывести на лист Excel основные результаты проецирования: матрицы счетов и матрицы нагрузок.

Программа SIC получает всю исходную информацию из открытой рабочей книги Excel. Эта информация должна быть помещена в таблицах непосредственно на рабочий лист (данные **X** и **Y**). Пользователь может организовать рабочее пространство так, как ему удобно – использовать только один лист для размещения всей информации, или несколько листов, или даже листы в разных рабочих книгах. Полученные результаты также выводятся как таблицы на листах рабочей книги и в виде графиков (см. пример на Рис. 7.4).

Программа не имеет ограничений на размер входной информации – количество предикторов, и откликов в данных. Размер данных ограничен только объемом памяти, который поддерживает операционная система, установленная на компьютере.

В состав программы входят два файла Sic.dll и Sic.xla. Для инсталляции программы, необходимо разместить файлы так, как указано в Таб. 7.1.

Таб. 7.1 Размещение файлов программы SIC

Файл	**Excel 7.0/Win 95**	**Excel 8.0 (9.0) / Win 98 (NT, ME)**
Sic.dll	*Root**OS*	*Root**OS*
Sic.xla	*Root**MO*\Office\Library	*Root**OS*\Profiles*Username*\Application Data\Microsoft\Addins

Символ *Root* означает диск , на котором находится соответствующая директория, например, *Root*= ***C***: ; символ *OS* означает имя директории , в которую установлена операционная система , например *OS*=***Windows*** ; символ *MO* означает имя директории , где находится система Microsoft Office, например *MO*= ***Program Files**Microsoft Office*** ; символ *Username* означает имя директории , содержащей индивидуальные настройки , например *Username*= ***All Users***.

После загрузки Excel необходимо выполнить подключение надстройки SIC.xla с помощью команды **Сервис / Надстройки (Tools/Add-Ins)**.

7.4. Входная информация для программы SIC

Для того чтобы активировать программу SIC, нужно использовать главное меню системы Excel, выбрав **Сервис (Tools)/SIC.** После этого появится Диалог программы SIC (Рис. 7.3), с помощью которого вводится вся необходимая информация и начинает

работать программа. Диалог разбит на семь отдельных областей, рассмотрим каждую из них подробно.

Область **Regression**

Как отмечалось в разделе 5.1, перед тем как начать ПИО вычисления, задачу необходимо регуляризировать, применяя МГК или ПЛС. Выбор метода осуществляется в поле **Method**. Программа позволяет выбрать МГК либо ПЛС1, либо ПЛС2. Количество главных или ПЛС компонент выбирается в поле **PCs**. Можно точно указать количество ГК, для которых должна быть произведена декомпозиция. Но можно в этом поле указать интервал, например «*2-4*», тогда данные будут обрабатываться три раза, для 2-х, 3-х и 4-х компонент, и в итоге получатся три результата. Можно так же указать «*All*», и тогда программа будет считать столько раз, сколько возможно вычислить ГК. Для каждой компоненты будет выведен свой результат. Программа вычисляет результаты не только ПИО прогноза, но и регрессионные результаты в соответствии с тем типом регрессии, который был выбран в поле **Method.**

Области данных: **Training Set**, **Test set**.

Для того, чтобы объяснить программе, какие именно данные составляют обучающий набор, адрес этих данных надо передать программе. Массивы **X** и**Y** могут располагаться на разных рабочих листах Excel и даже в разных рабочих книгах. Отдельно вводится область **X** данных, отдельно область **Y**. Данные могут быть записаны на рабочих листах Excel в транспонированном виде, об этом необходимо сообщить программе, отметив соответствующие поле **t,** расположенное непосредственно за полями **X** и **Y**. Аналогичным образом вводятся данные о проверочном массиве в области **Test Set.** Можно указывать как адрес области данных (Рис. 7.3, область **Test Set**), так и области, котором даны имена с помощью стандартных средств Excel (Рис. 7.3, область **Training Set**). В качестве **Test Set** возможно указать обучающий набор. В этом случае «прогноз» будет сделан на обучающие объекты, что позволяет автоматически вычислить статус всех объектов обучающего набора.

Область проверочных данных так же используется и для прогноза новых объектов. В этом случае поле **Y** остается незаполненным.

Область **Validation.**

Здесь указывается, какой именно способ проверки используется для модели. Возможны следующие режимы:

1. *Test Set* – проверка проводится с помощью тестового массива, и в область данных **Test Set** необходимо ввести ссылки на соответствующие данные, $\mathbf{X}_{test}$, $\mathbf{Y}_{test}$;
2. *FullCross* (*1 Out*) - полная перекрестная проверка, с выкидыванием на каждом шаге по одному образцу;
3. *Systematic1122* –перекрестная проверка с систематическим выкидыванием *k*-объектов. Размер сегмента определяется пользователем с помощью поля **Segments**;
4. *Systematic1212* –перекрестная проверка с систематическим выкидыванием каждого *k*-ого объекта. Параметр *k* определяется в поле **Segments**;

При использовании одного из режимов 2-4, область **Test Set** становится неактивной, так как при этих режимах проверочный массив не нужен.

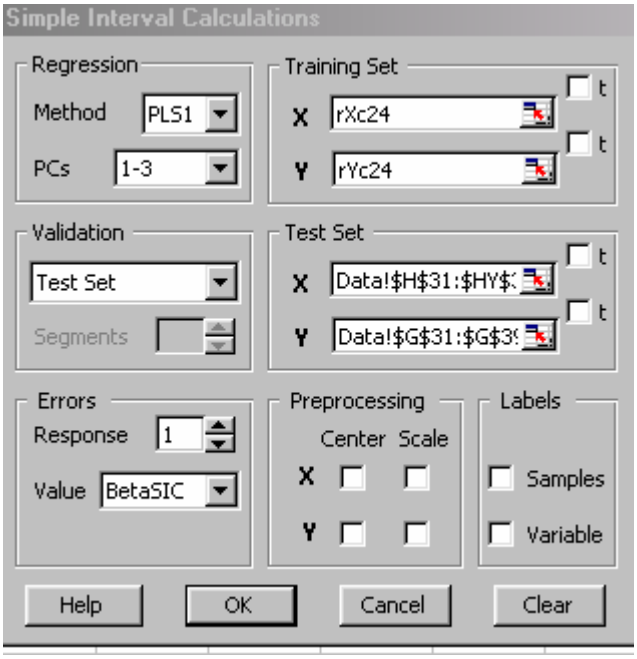


Рис. 7.3 Диалоговое окно программы SIC.

Область **Error**

Используется для того, чтобы объяснить программе, с каким значением β будут проводиться ПИО вычисления. В поле **Value** можно:

1. указать конкретное значение β, если оно известно пользователю;
2. выбрать «*Minimum*», тогда программа вычислит b_{min} согласно формуле (5.21), и все последующие вычисления будет производить для $\beta = b_{min}$;

3. выбрать «*BetaSIC*», тогда программа сначала вычисляет значение b_{SIC} по формуле (5.25), а затем все последующие вычисления производятся для $\beta = b_{SIC}$.

Над полем **Value** расположено поле **Response**. Если модель одноотликовая, то в этом поле стоит «*1*» по умолчанию. Если откликов несколько, то для каждого из них можно выбрать свое значение β, т.е. установить требуемое значение в поле **Value**.

Область **Preprocessing**..

Перед обработкой, данные можно центрировать и/или шкалировать, отдельно **X**, отдельно **Y** значения. Для этого отмечаются соответствующие контрольные поля в области **Preprocessing**.

Область **Labels**

Если поля **Samples** и/или **Variables** отмечены, это значит, что в областях данных содержатся имена объектов и/или переменных, которые не надо учитывать при загрузке данных в память.

После нажатия кнопки OK программа начинает счет, если все необходимые поля правильно заполнены, либо выдает сообщение об ошибке, с комментарием, в чем именно ошибка состоит.

Все действия, выполняемые пользователем с помощью диалогового окна, могут быть проделаны в автоматическом режиме с помощью *функций SIC*. Для этого нужно написать программу (макрос) на встроенном в Excel языке VBA [242]. Кроме того, такие программы можно создавать и автоматически. Детали того, как это делается, приведены в разделе 7.6.

7.5. Результаты работы программы SIC

После окончания работы программа SIC создает новый рабочий лист в активной книге Excel и туда выводит как числовую информацию, в виде таблиц Excel, так и графическую.

В числовом виде выводятся следующие результаты. Для всех объектов указанных в области **Test Set,** если используется режим проверки **Validation** = *Test Set* **,** либо для всех объектов из обучающего набора (область **Training Set**), если выбран любой другой режим **Validation**:

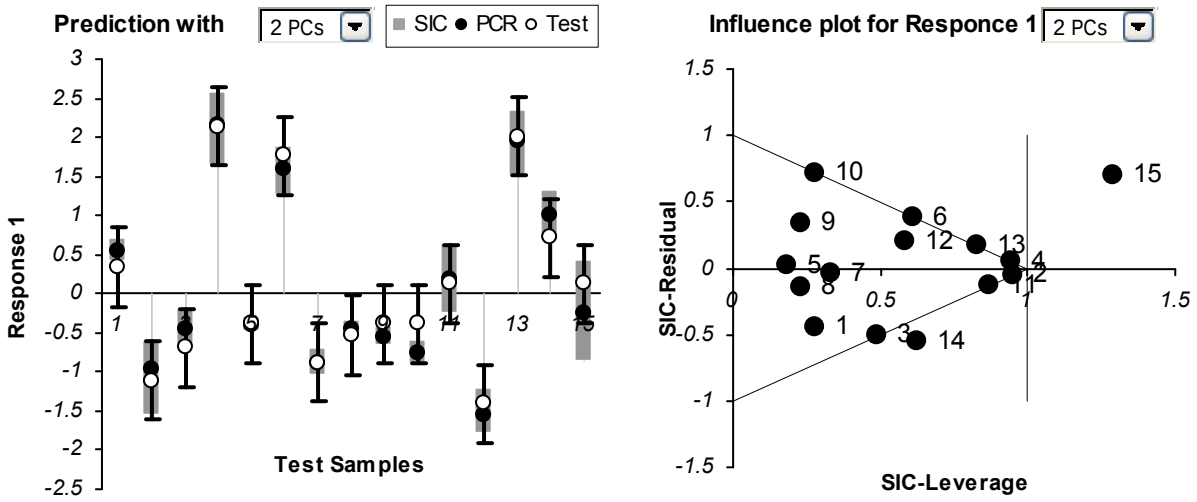
- y_i^+ - правая граница ПИО интервала
- $\hat{y}_i$ - оценка с помощью регрессии,

- $\bar{y}_i$ - правая граница ПИО интервала
- y_i – стандартное проверочное значение, ссылка на которое вводилось в поле **Test Set/Y**, или **Training Set/Y**
- ПИО остаток r_i (6.8)
- ПИО размах h_i (6.9).
- Значения для y_i и r_i не выводятся, если вычисления производятся для новых объектов, для которых значения откликов неизвестны.

Так же выводится следующая общая информация:

- количество главных компонент, с котором проводился счет;
- значение β, используемое при данных расчетах;
- RMSEP;

Кроме того, выводится дополнительная информация относительно того, где расположены исходные таблицы данных, и какие дополнительные установки были произведены с помощью основного диалога программы SIC. Эта информация может понадобиться пользователю, если он захочет повторить счет в том же режиме.



a) График прогнозных значений

b) Диаграмма статуса объектов

Рис. 7.4 Графические результаты работы программы SIC. Пример

Результат работы так же представлен двумя графиками. На графике прогнозных значений (Рис. 7.4а) наглядно показана информация из пункта 1 числовой информации. Здесь, белые точки обозначают измеренные значения откликов, черные точки – это оценки полученные регрессией, серые интервалы – это прогнозные ПИО интервалы SIC,

черные – калибровочные ПИО интервалы. На Рис. 7.4 b) изображены те же объекты, что и на соседнем графике, но уже на диаграмме статуса объектов.

7.6. Автоматизация работы с программой SIC

Если в процессе исследований пользователю необходимо провести большой численный эксперимент, при котором надо будет много раз обращаться программе SIC, то это удобнее сделать в автоматическом режиме с помощью VBA [242] процедур и используя функции SIC. Важно помнить о том, что прежде чем использовать эти функции в программах, необходимо установить ссылку на надстройку SIC.xla в Microsoft Visual Basic. Для этого в активном модуле VBA надо выбрать **References** в меню **Tools**, а затем отметить файл SIC.xla в **Available References**. Для всех функций SIC действует следующее правило - если возвращаемое значение равно нулю , то произошла ошибка . Для вывода этой ошибки используется функция `SICError`.

`SICError`

Выводит информацию о произошедшей ошибке .

Синтаксис

`SICError`

Аргументы

`Нет`.

Для того чтобы передать программе SIC входную информацию, произвести обработку данных, и вывести информацию в память и/или на рабочие листы Excel, используются специально разработанные функции программы SIC.

`SICMake`

Передает входную информацию в программу SIC и начинает процесс обработки данных.

Синтаксис

SICMake([sXcalibration,] [sYcalibration,] [sXtest,] [sYtest,] [kMethod,] [sPCs,] [kValidation,][kSegments,] [dErrorValues][kXcalibrationT,] [kYcalibrationT,] [kXtestT,] [kYtestT,] [kXcente,] [kXweight,] [kYcenter,] [kYweight,]

Аргументы

`sXcalibration`

Ссылка на область предикторов обучающего массива. В диалоговом окне соответствует полю **X** области **Training Set**.

sYcalibration

Ссылка на область откликов обучающего массива. В диалоговом окне соответствует полю **Y** области **Training Set**.

sXtest

Ссылка на область предикторов проверочного массива. В диалоговом окне соответствует полю **X** области **Test Set**. Если выбран режим проверки отличный от *Test Set*, этот аргумент не используется.

sYtest

Ссылка на область откликов проверочного массива. В диалоговом окне соответствует полю **Y** области **Test Set**. Если выбран режим проверки отличный от *Test Set*, этот аргумент не используется. Если производится прогноз на новые объекты, и значения откликов неизвестны, то этот аргумент так же не используется.

kMethod

Указывает на метод декомпозиции. При **kMethod=0** используется МГК, **kMethod=1** используется ПЛС, **kMethod=2** используется ПЛС2,

sPCs

Указывает на количество главных компонент, которое должно быть использовано при декомпозиции. При **kMethod=0** используется МГК; **kMethod=1** используется ПЛС-1; **kMethod=2** используется ПЛС-2,

kValidation

Указывает на режим проверки. При **kValidation=0** проверка с помощью тестового массива; **kValidation=1** проверка с помощью полной кросс-валидации; **kValidation=2** проверка с помощью сегментированной кросс-валидации, сегменты выводятся в порядке «*Systematic1122*»; **kValidation=3** проверка с помощью сегментированной кросс-валидации, сегменты выводятся в порядке «*Systematic1212*».

kSegments

При **kValidation=2** ***или*** **3,** вводится количество сегментов для сегментированной кросс-валидации. В других режимах этот параметр вводить не надо.

dErrorValues

Указывает на величину β которая используется при ПИО вычислениях. При **dErrorValues =-1** , сначала определяется минимальное значение β , а потом производятся ПИО вычисления. При **dErrorValues =-2** , вычисляется $\beta = b_{SIC}$, а потом производятся ПИО вычисления. Если **dErrorValues = «положительное число»**, то ПИО вычисления производятся с указанным значением β.

kXcalibrationT

Если **kXcalibrationT=True** массив предикторов X обучающего набора перед вычислением надо транспонировать. При **kXcalibrationT=False**, транспонирование не требуется. Необязателен, по умолчанию **kXcalibrationT=False.**

kYcalibrationT

Если **kYcalibrationT=True** массив откликов Y обучающего набора перед вычислением надо транспонировать. При **kYcalibrationT=False**, транспонирование не требуется. Необязателен, по умолчанию **kYcalibrationT=False.**

kXtestT

Если **kXtestT=True** массив предикторов X проверочного набора перед вычислением надо транспонировать. При **kXtestT=False**, транспонирование не требуется. Необязателен, по умолчанию **kXtestT=False.**

kYtestT

Если **kYtestT=True** массив откликов Y проверочного набора перед вычислением надо транспонировать. При **kYtestT=False**, транспонирование не требуется. Необязателен, по умолчанию **kYtestT=False.**

SICResultsBack

Возвращает результаты, полученные с помощью программы SIC.

SICResultsBack([Maximum,] [Minimum,] [Regression,][Test,] [Errors,] [PCs,] [RMSEP,] [Leverage,] [Residuals,] [Boundary])

Некоторые аргументы являются массивами . При этом предполагается установка Option Base 1 в модуле VBA. Необходимо заранее зарезервировать достаточное число элементов этих массивах

Аргументы.

Maximum

Этот аргумент получает массив правых границ ПИО интервала, т.е. он состоит из y_i^+.

Minimum

Этот аргумент получает массив левых границ ПИО интервала, т.е. он состоит из y_i^-.

Regression

Этот аргумент получает массив значений откликов, полученный с помощью регрессионного метода, т.е. он состоит из $\hat{y}_i$.

Test

Этот аргумент получает массив проверочных значений откликов.

Errors

Этому аргументу присваивается значение параметра β, которое использовалось при расчетах. Если расчеты проводились для нескольких откликов и/или нескольких значений ГК, то этот аргумент является массивом.

PCs

Этому аргументу присваивается количество ГК, которое использовалось при расчетах. Если расчеты проводились для нескольких откликов и/или нескольких значений ГК, то этот аргумент является массивом.

RMSEP

Этому аргументу присваивается значение среднеквадратичного остатка при прогнозе (см.(3.2)) по результатам регрессионного оценивания. Если расчеты проводились для нескольких откликов и/или нескольких значений ГК, то этот аргумент является массивом.

Leverage

Этот аргумент получает массив значений ПИО размаха интервала (см.((6.9))

Residuals

Этот аргумент получает массив значений ПИО остатков интервала (см. (6.9))

Boundary

Этот аргумент получает массив значений,

$$g_i = 1 - h(\boldsymbol{x}_i) - |r(\boldsymbol{x}_i, y_i)|$$

используемый для определения граничных объектов. Если объект принадлежит к обучающему набору и $|g_i| < 1.e\text{-}8$, то объект можно считать граничным.

SICDimensionsBack

Возвращает текущие размерности зарегистрированных данных

SICDimensionsBack([PCs,] [Variables,] [Samples,] [Responses,] [Tests])

Аргументы

PCs

Этот аргумент возвращает количество главных компонент для которых проводились вычисления.

Variables

Этот аргумент возвращает количество переменных в матрице предикторов **X**.

Samples

Этот аргумент возвращает количество объектов в обучающем наборе.

Responses

Этот аргумент возвращает количество откликов (столбцов) в матрице **Y**.

Tests

Этот аргумент возвращает количество объектов в проверочном наборе.

SICSettingsBack

Эта функция возвращает всю входную информацию, которая была введена функцией **SICMake**.

SICSettingsBack([Settings])

Аргументы

Settings

Этот аргумент является массивом, в элементы которого записана вся входная информация, полученная с помощью функции **SICMake**

SICOutput

Это функция выводит результат расчетов с помощью программы SIC. Функция создает новый лист активной рабочей книги Excel и туда выводит результат.

SICOutput([kPlots])

Аргументы

kPlots

С помощью этого аргумента пользователь объясняет программе, нужно ли выводить графическую информацию. Если **kPlots:=True**, то графики выводятся, в противном случае выводится только цифровая информация.

7.7. Функции рабочего листа программы SIC

Дополнительными возможностями программы SIC является набор специально определенных макрофункции для вычисления матриц счетов и нагрузок при декомпозиции. Этими функциями можно пользоваться как обычными функциями рабочего листа Excel. Для этого надо активизировать Мастер функций и выбрать категорию **Определенные пользователем**. При этом в списке имен функций будут представлены описанные ниже функции.

ScoresPCA

Возвращает в выделенную область рабочей книги матрицу счетов, посчитанную по МГК

Синтаксис

```
ScoresPCA (rMatrixX, [nPCs])
```

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу счетов **T**.

LoadingsPCA

Возвращает в выделенную область рабочей книги матрицу нагрузок, посчитанную по МГК

Синтаксис

```
LoadingsPCA (rMatrixX, [nPCs])
```

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу счетов **P**.

ScoresPLS

Возвращает в выделенную область рабочей книги матрицу счетов предикторов посчитанную по методу ПЛС 1

Синтаксис

ScoresPLS (rMatrixX, rMatrixY, [nPCs])

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположен вектор откликов **y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу счетов **T**.

UScoresPLS

Возвращает в выделенную область рабочей книги матрицу счетов откликов, посчитанную по методу ПЛС 1

Синтаксис

UScoresPLS (rMatrixX, rMatrixY, [nPCs])

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположен вектор откликов **y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу счетов **U**.

LoadingsPLS

Возвращает в выделенную область рабочей книги матрицу нагрузок предикторов посчитанную по методу ПЛС 1

Синтаксис

LoadingsPLS (rMatrixX, rMatrixY, [nPCs])

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположен вектор откликов **y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу нагрузок **P**.

WLoadingsPLS

Возвращает в выделенную область рабочей книги матрицу взвешенных нагрузок предикторов, посчитанную по методу ПЛС 1

Синтаксис

```
WLoadingsPLS (rMatrixX, rMatrixY, [nPCs])
```

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположен вектор откликов **y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу нагрузок **W**.

QLoadingsPLS

Возвращает в выделенную область рабочей книги вектор нагрузок откликов посчитанную по методу ПЛС 1

Синтаксис

```
QLoadingsPLS (rMatrixX, rMatrixY, [nPCs])
```

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположен вектор откликов **y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в вектор нагрузок **q**.

ScoresPLS2

Возвращает в выделенную область рабочей книги матрицу счетов предикторов посчитанную по методу ПЛС 2

Синтаксис

ScoresPLS2 (rMatrixX, rMatrixY, [nPCs])

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположена матрица откликов **Y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу счетов **T**.

UScoresPLS2

Возвращает в выделенную область рабочей книги матрицу счетов откликов посчитанную по методу ПЛС 2

Синтаксис

UScoresPLS2 (rMatrixX, rMatrixY, [nPCs])

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположена матрица откликов **Y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу счетов **U**.

LoadingsPLS2

Возвращает в выделенную область рабочей книги матрицу нагрузок предикторов посчитанную по методу ПЛС 2

Синтаксис

LoadingsPLS2 (rMatrixX, rMatrixY, [nPCs])

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположена матрица откликов **Y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу нагрузок **P**.

WLoadingsPLS2

Возвращает в выделенную область рабочей книги матрицу взвешенных нагрузок предикторов посчитанную по методу ПЛС 2

Синтаксис

```
WLoadingsPLS2 (rMatrixX, rMatrixY, [nPCs])
```

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположена матрица откликов **Y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу нагрузок **W**.

QLoadingsPLS2

Возвращает в выделенную область рабочей книги вектор нагрузок откликов посчитанную по методу ПЛС 1

Синтаксис

```
QLoadingsPLS2 (rMatrixX, rMatrixY, [nPCs])
```

Аргументы

rMatrixX

Область на рабочем листе, где расположена матрица предикторов **X**

rMatrixY

Область на рабочем листе, где расположена матрица откликов **Y**

nPCs

Необязателен. Указывает максимальное количество ГК , которые должны быть включены в матрицу нагрузок **Q**.

Для всех этих функций действуют следующие правила:

- если область для вывода результата меньше, чем соответствующая матрица результата, то выводится только начальная часть матрицы, которая помещается в эту область;

- если отмечена более широкая область, то в излишние клетки выводится сообщение об ошибке *#N/A*;
- если параметр **`nPCs`** не задан, то выводится максимально возможное для отмеченной области количество столбцов матрицы счетов/ нагрузок;
- если параметр **`nPCs`** задан, но **`nPCs`** > rank**X ,** то декомпозиция делается для максимально возможного количества ГК, а оставшиеся столбцы заполняются сообщением об ошибке *#N/A*.

7.8. Результаты главы 7

В этой главе приведено описание структуры программы SIC, которая состоит из целого набора процедур. Основные алгоритмы, используемые для реализации ПИО, метода разработаны и хорошо обоснованы с вычислительной точки зрения. Представление регрессионной задачи в виде задачи линейного программирования непривычно для большинства экспериментаторов, поэтому в этой главе приведены основные сведения из области линейной оптимизации и дано детальное описание переход от исходной задачи к задаче ЛП. Так же дано подробное описание программы SIC - надстройки для программы Excel. С ее помощью можно проводить интервальную и регрессионную обработку данных и выполнять ПИО классификацию статуса объектов, использовать встроенные функции, для вычисления основных матриц в проекционных методах. Устройство системы SIC соответствует современным требованиям к подобным программам. Программа имеет простой и привычный дружественный интерфейс. Вся входная информация задается на листах рабочей книги Excel. Вся выходная информация, таблицы и графики, так же выводятся в стандартном виде на рабочие листы Excel. Для того чтобы передать данные в программу и начать обработку используется диалоговое окно.

Все действия, которые выполняются с помощью Диалога SIC, могут быть осуществлены и с помощью VBA процедур. Для этого применяются функции SIC. Используя их можно написать подпрограммы для ввода и обработки данных, получить результаты моделирования. Приведено подробное описание всех функций.

В программа SIC так же имеются дополнительные функции, осуществляющие декомпозицию исходной задачи одним из выбранных методов :МГК, ПЛС-1, ПЛС-2. Для работы с этими функциями не надо открывать диалог программы SIC, их используют как обычные функции рабочего листа Excel.

Программа SIC – это мощный, современный инструмент, созданный для интервального и регрессионного анализа результатов сложных многофакторных физических экспериментов. Программа систематически используется в этой работе для решения всех теоретических и прикладных задач. Не смотря на свою новизну, программа уже применяется не только в Институте химической физики РАН (Москва), но и в лаборатории ACSAB Alborg (Esbjerg, Дания). Программа представлялась на международных и российских конференциях.

Теоретические и практические аспекты применения метода простого интервального оценивания

Потребность в моделировании физико-химических процессов и зависимостей, которые не поддаются содержательному математическому описанию из-за их сложности или не изученности, вызвала необходимость применения формальных математических (хемометрических) методов. Использование формальных методов имеет свои особенности. Во-первых, по большей части – это «простые» эмпирические модели, справедливые в весьма ограниченной области. Поэтому явное описание области действия модели, либо проверка вновь исследуемых объектов на принадлежность этой области, является неотъемлемой частью формального моделирования. Во-вторых, в отличие от «традиционного подхода», при формальном моделировании рассматриваются многофакторные данные в совокупности, не выделяя предварительно один или несколько «существенных» факторов или переменных. В результате построенные многомерные эмпирические модели позволяют обнаружить новые связи или явления, так как учитывают скрытые, объективно существующие связи внутри системы.

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

В этом разделе рассматриваются различные методологические аспекты применения хемометрических методов с использованием практических примеров. Так как метод ПИО является новым среди формальных методов, в этом разделе сравниваются результаты и выводы, получаемые с помощью метода ПИО, с уже «традиционными» для формального моделирования методами РГК и ПЛС.

Кроме того, сравниваются подход и область применимости содержательного и формального моделирования к анализу одного и того же эксперимента.

Также рассматривается методология применения метода ПИО для решения трех конкретных задач. Это – выбор представительного выборки из обучающего набора, задача расширяющегося моделирования при решении задачи аналитического контроля процесса, и применение метода ПИО для решения задач дискриминации.

8. Применение проекционных методов совместно с методом ПИО на примере анализа многоканальных акустических измерений. Наглядное представление многофакторных данных

Формальные методы моделирования широко используют проекционный подход, который основывается на концепции «скрытых (латентных) переменных», или базисных векторов, на которых строится проекционное подпространство. Одним из основных достоинств применения проекционного подхода к построению моделей является возможность представления в явном виде скрытых закономерностей и структур, присущих данным физических экспериментов. При этом важную роль играет понятие «изменение значений». Основным предположением, которое используют при нахождении таких «скрытых закономерностей», является то, что направления, по которым происходят наибольшие изменения в данных, так или иначе связаны с этими закономерностями. Применение проекционных методов приводит к существенному понижению размерности задачи, и, как следствие, к возможности изобразить изучаемые объекты наглядно в 2-х и 3-х мерных пространствах. Для этого используются графики счетов, которые показывают взаимное расположение объектов (см. раздел 3.1). У исследователей, которые впервые сталкиваются с графиками счетов, сразу возникает вопрос: «какие единицы отложены по осям координат?». Вопрос этот естественный, так как исходные данные являются результатами некоторых физических или химических измерений. Однако на графиках счетов объекты представлены в относительных, безразмерных единицах, так как изучаются не сами значения, а их изменения и положения относительно друг друга. Графики счетов применяются для идентификации групп, обнаружения тенденций, нахождения общих свойств и т.д. Правила перехода из пространства исходных переменных в проекционное пространство описываются векторами нагрузок. Как и в случае счетов, векторы нагрузок можно изобразить графически – один вектор относительно другого. График нагрузок показывает, какой вклад вносит каждая исходная переменная в каждую скрытую переменную (компоненту). Если рассматривать совместно графики счетов и нагрузок, то можно выявить влияние исходных переменных на скрытые структуры данных.

Возможность наглядного представления сложных многофакторных данных физического эксперимента в проекционном пространстве позволяет исследователю лучше понять и объяснить изучаемые явления.

При детально анализе графиков счетов возникает необходимость охарактеризовать свойства каждого отдельного объекта относительно всей группы объектов и построенной модели. Например, понять, какой из образов являет наиболее важным, охарактеризовать роль объекта среди всего набора данных количественно, классифицировать различные типы выбросов и т.д. В стандартных проекционных методах такая классификация делается полуэмпирическим путем и, во многом, зависит от интуиции исследователя и его понимания физического существа изучаемого процесса.

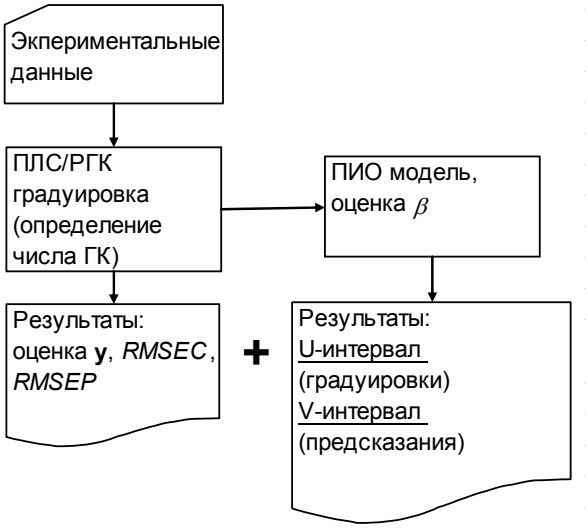


Рис. 8.1 Общая схема применения проекционных методов совместно с методом ПИОЧ

Еще один вопрос, который неизбежно возникает при формальном моделировании, это область действия модели, а, следовательно, и надежность прогноза. Существует множество эмпирических приемов, как это можно контролировать, но общепризнанного, точного подхода не существует.

Метод ПИО заполняет эти пробелы и представляет законченную систему классификации объектов, и набор однозначных правил для определения статуса (роли) каждого объекта в исследуемом наборе данных, в совокупности с математической моделью (глава 6). Схема совместного использования проекционных регрессионных методов и метода ПИО изображена на Рис. 8.1.Основной материал этой главы опубликован в [52]

8.1. Эксперимент. Измерение следовых концентраций нефти в воде с помощью акустических измерений

Результаты ПИО классификации демонстрируется на примере применения акустических измерений с последующей математической обработкой экспериментальных данных для количественного определения следовых концентраций нефти в промышленных сточных водах в режиме реального времени. Эксперимент разрабатывался и проводился группой проф. К. Есбенсена (Esbensen) в лаборатории прикладной хемометрики в Telemark Institute of Technology (Норвегия) [243], и далее в

университете Ålborg, Esbierg (Дания). Позднее данные, полученные в этом эксперименте, анализировались с помощью ПИО метода [52].

Для проведения эксперимента была построена лабораторная установка, позволяющей имитировать поведение потока жидкости в трубопроводе. Для создания турбулентности в потоке использовались измерительные диафрагмы. Применение акустического метода исследования в данном случае возможно потому, что при наличии нефти в воде, физические свойства потока изменяются. Даже очень небольшие примеси нефти уменьшают поверхностное натяжение жидкости, что и отражается в изменениях акустического сигнала. Вибрационные сенсоры измеряли сигнал, который далее усиливался, и, с помощью быстрого преобразования Фурье, превращался из сигнала, зависящего от времени, в частотный, который в дальнейшем и назывался акустическим спектром. Пример двух акустических спектров с различной концентрацией нефти (2.5 ppm и 300 ppm) приведен на Рис. 8.2.

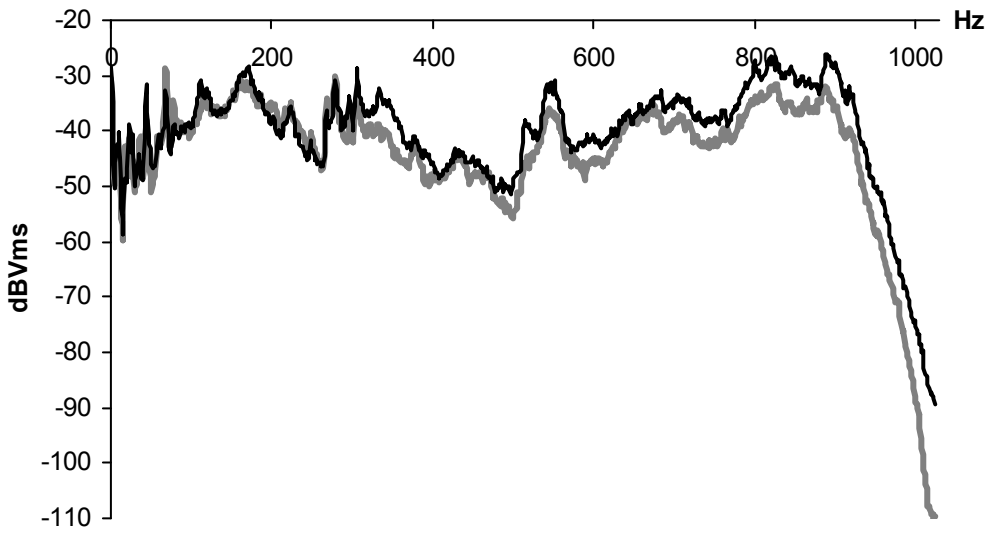


Рис. 8.2 Пример акустических спектров. – объект с концентрацией 2.5 ppm
– объект с концентрацией 300 ppm

Для построения калибровочной модели, связывающей значения акустического сигнала с концентрацией нефти в воде, было приготовлено 40 образцов с различной концентрацией нефти: 0, 2.5, 5, 10, 20, 50, 100, 300 ppm. По пять образцов каждой концентрации. Такой же набор из 40 образцов был приготовлен для проверки методики.

В качестве матрицы предикторов **X** использовалась матрица акустических спектров на 1024 частотах; вектор откликов **y** – это известные стандартные концентрации нефти.

8.2. Исследование калибровочного набора

Для построения калибровки использование обычной линейной регрессионной модели

$$\mathbf{y}=\mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon}$$

невозможно из-за вырожденности матрицы **X**. Поэтому, для регуляризации задачи, применялся метод ПЛС1 (раздел 3.1)

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{t}+\mathbf{E} \qquad\qquad \mathbf{y}=\mathbf{T}\mathbf{q}+\mathbf{f}.$$

Однако применение метода ПЛС к исходным данным показало существенно нелинейную связь между предикторами **X** и откликами **y**. Эта нелинейность отчетливо видна на графике X-счетов $\mathbf{t}_k$ относительно Y-счетов, $\mathbf{u}_k$ для первой ПЛС компоненты (Рис. 8.3).

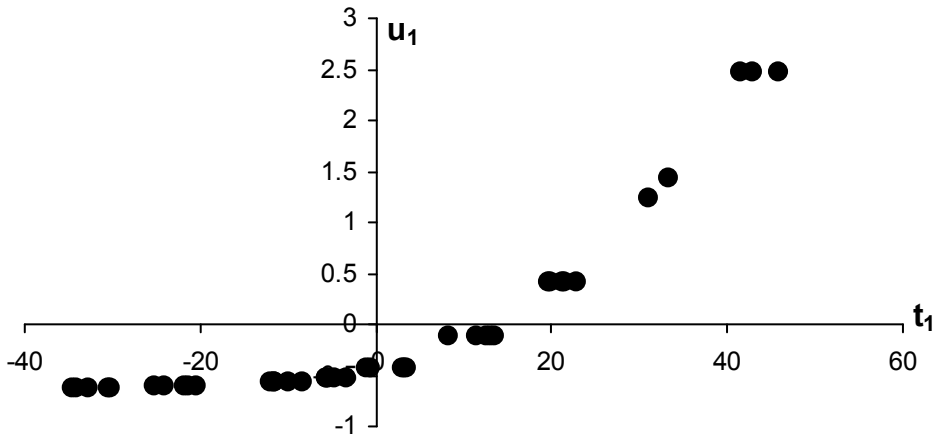


Рис. 8.3 Определение следовых концентраций нефти в воде. ПЛС модель с использованием исходных (не преобразованных) данных.

Для ПЛС1, при наличие только одного отклика, **u**-векторы представляют из себя просто усеченные векторы **y,** т.е.

$$\mathbf{u}_1=\mathbf{y}_1, \qquad \mathbf{u}_к=\mathbf{u}_{k-1}-(\mathbf{y}^{T}\mathbf{t}_k)/(\mathbf{t}_k{}^{T}\mathbf{t}_k)\,\mathbf{t}_k, \quad k=2,3,..,K \tag{8.1}$$

Напомним, что в основе метода ПЛС лежит предположение о линейной связи между **X** и **y**. Существенное отклонение от линейности может привести к большим ошибкам, как в калибровке, так и при предсказании новых объектов. Поэтому очень важно обнаружить такую нелинейную связь, особенно когда она проявляется на первой ГК, и постараться скорректировать модель. При использовании метода ПЛС, производится одновременная декомпозиция предикторов и откликов, причем главные направления выбираются так, чтобы максимально описать корреляцию между **y** и **X** (см. раздел 3.1 и [67]). При этом счета $\mathbf{t}_k$ – это новые координаты точек из матрицы **X**, а счета $\mathbf{u}_k$ выделяют ту часть структуры данных **y**, которая лучше всего описывается матрицей **X**

вдоль рассматриваемой главной компоненты *k.* Наиболее важными, как и МГК, являются графики для первых главных компонент.

Для того чтобы стало возможным построить линейную модель на всем диапазоне изменения *y*, отклики были преобразованы по формуле

$$y=\lg(1+y_{raw}) \; . \qquad (8.2)$$

Это позволило построить ПЛС модель с 2 главными компонентами, которая описывает 60% изменений в матрице **X** (E_x=60%), но при этом 99.9% изменений в **y** (E_y=99.9%), где

$$E_x = 100\left(1-\sum_{i=1}^{I}\sum_{j=1}^{J} e_{ij}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2\right), \qquad E_y = 100\left(1-\sum_{i=1}^{I} f_i^2 \Big/ \sum_{i=1}^{I} y_i^2\right)$$

Среднеквадратичный остаток моделирования (3.1) RMSEC=0.051, а среднеквадратичный остаток предсказания, вычисленный на проверочном наборе (3.2) RMSEP=0.092. Используя обратное к (8.2) преобразование можно выразить погрешности моделирования и прогноза в исходных единицах измерений, тогда RMSEC=0.12, и RMSEP=0.24. Это можно считать хорошей точностью, учитывая диапазон изменения откликов.

Черные и белые кружки на Рис. 8.4 обозначают объекты калибровочного набора. Видно, что преобразование (8.2) позволило избавиться от нелинейности.

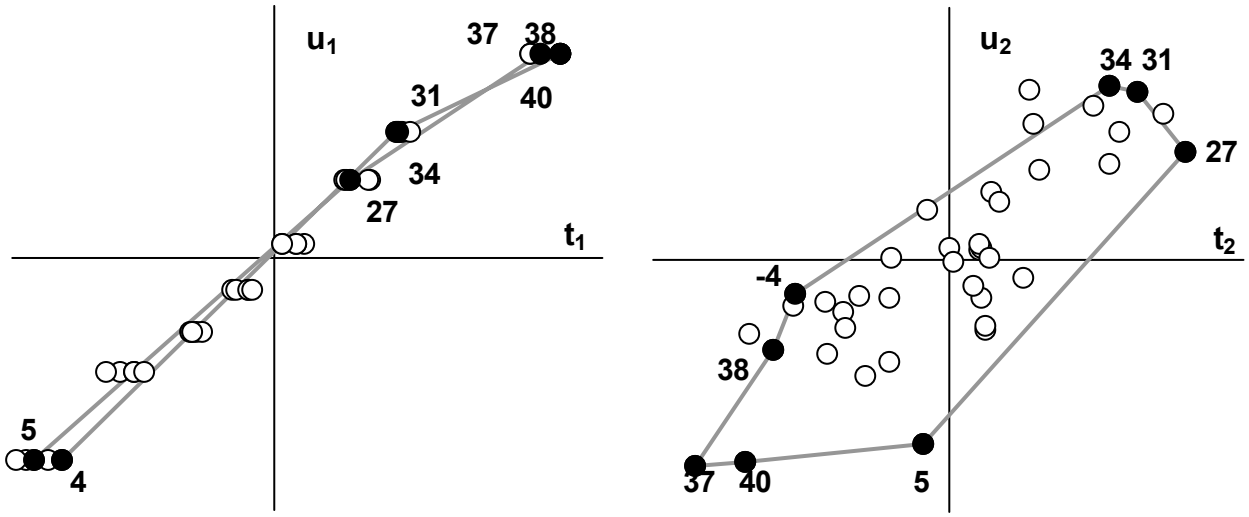
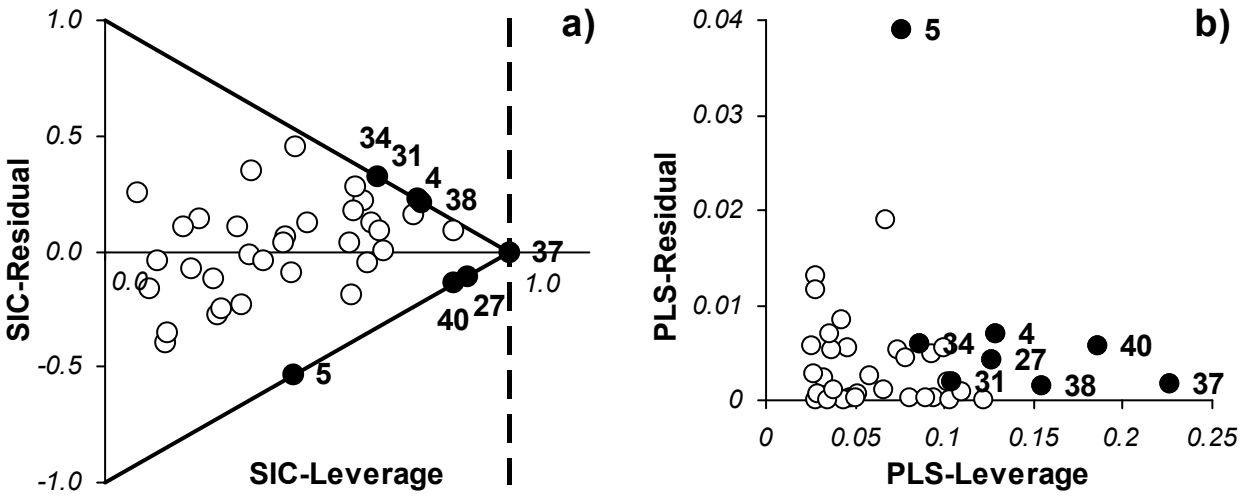


Рис. 8.4 Определение следовых концентраций нефти в воде. ПЛС модель с 2 ГК. График счетов **t-u**. Калибровочный набор: ○- внутренние объекты, ● - граничные объекты, серая линия соединяет граничные объекты

Используя ПЛС преобразования, строим ПИО модель и соответствующую ДСО (Рис. 8.5 а); при этом оценка b_{SIC}=0.227. ПИО метод обнаружил восемь граничных объектов (6.13), их номера в калибровочном наборе обозначены на графиках.



ПИО диаграмма статуса объектов. График влиятельности объектов по **y**

Рис. 8.5 Определение следовых концентраций нефти в воде.
Калибровочный набор: O- внутренние объекты, ● - граничные объекты

Объекты из калибровочного набора на Рис. 8.4 обозначены в соответствии с ПИО классификацией: черные кружки – это граничные объекты, белые кружки – внутренние объекты. Сравнивая Рис. 8.4 и Рис. 8.5а видно, что объекты, называемые граничными, действительно максимально удалены от центра облака точек. Серый многоугольник отделяет область модели, ограниченную ПИО граничными объектами. На графиках так же видно, что некоторые внутренние объекты выходят за границы многоугольников. Это можно легко объяснить, так как графики счетов $\mathbf{t}_k$-$\mathbf{u}_k$ это всего лишь проекции всей модели на определенные плоскости.

Интересным является вопрос, а какую новую информацию предоставляет ПИО метод, по сравнению с обычным ПЛС методом. Сравнивая ДСО (Рис. 8.5а) с графиком влиятельности Рис. 8.5b видно, что все наиболее влиятельные (раздел 6.1) объекты (NN 37, 38 и 40), а так же объект, имеющий максимальное значения остатка моделирования (N5), в то же время являются граничными по ПИО классификации. Графики, представленные на Рис. 8.5 показывают, что ПИО классификация позволяет находить все наиболее влиятельные объекты среди калибровочного набора. Для определения таких объектов метод ПИО дает однозначное и простое правило (Утверждение 6.8). Таким образом, можно сделать вывод, что концепция граничных объектов имеет смысл и

полезна не только внутри самого метода ПИО, а так же объективно характеризует исследуемую структуру данных физического эксперимента.

8.3. Исследование проверочного набора

Важным аспектом ПИО классификации является возможность определения статуса объектов проверочного массива. Оценка величины прогнозного интервала индивидуально для каждого объекта, и как следствие, разделение проверочных объектов на различные классы является преимуществом метода ПИО, по сравнению с традиционным регрессионным подходом.

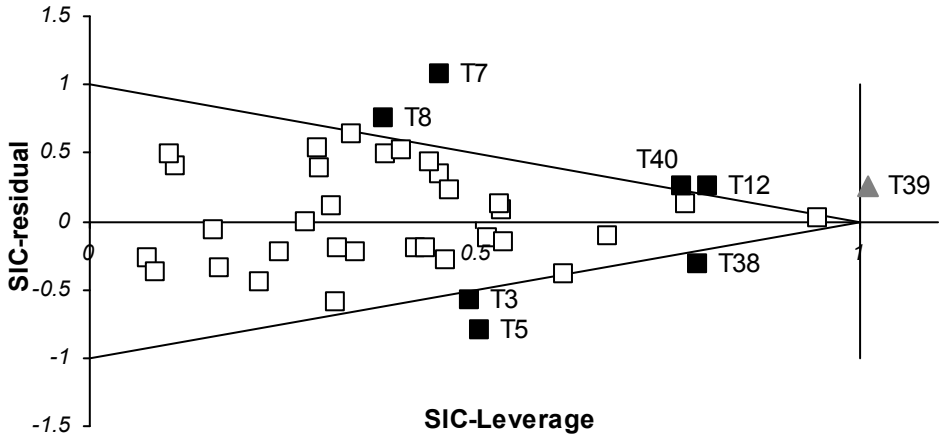


Рис. 8.6 Определение следовых концентраций нефти в воде. Диаграмма статуса объектов, Проверочный набор: □ - внутренние объекты, ■ - внешние объекты, ▲ - абсолютно внешние объекты.

Согласно ПИО классификации (Утверждения 6.7,6.9, 6.10), в проверочном наборе обнаружено 32 внутренних объекта и восемь внешних. Объекты NN 3, 5, 7, 8, 12, 38, 40 внешние, но при этом они расположены достаточно близко к области модели. Внешними эти объекты могут быть по двум причинам: во-первых, они могут содержать большую погрешность в значениях откликов; во-вторых, связь между откликами и предикторами (X-y) несколько иная, чем для калибровочных объектов, т.е. сказывается погрешность моделирования. После того, как такие объекты выявлены с помощью ДСО, их последующий содержательный анализ производится исследователем. Кроме того, объект N 39 (серый треугольник на Рис. 8.6) является абсолютно внешним. Такой объект по структуре данных в предикторах отличается от калибровочных объектов. Важно, что ни при каких значениях отклика y такой объект не станет внутренним. Величина прогнозных интервалов для таких объектов всегда больше чем β.

Рассмотрим проекции проверочных объектов на пространство, определенное ПЛС-моделью, и их положения относительно области, очерченной граничными объектами. На графиках (Рис. 8.7) внутренние объекты обозначены белыми квадратами, внешние – черными, а абсолютно внешний объект изображен серым треугольником, серая ломаная линия ограничивает область модели. Эта область построена на основе граничных объектов и перенесена с Рис. 8.4. Хотя на Рис. 8.7 видно периферийное положение внешних объектов, в отличие от ДСО четкого разграничения между внутренними и внешними объектами получить не удается.

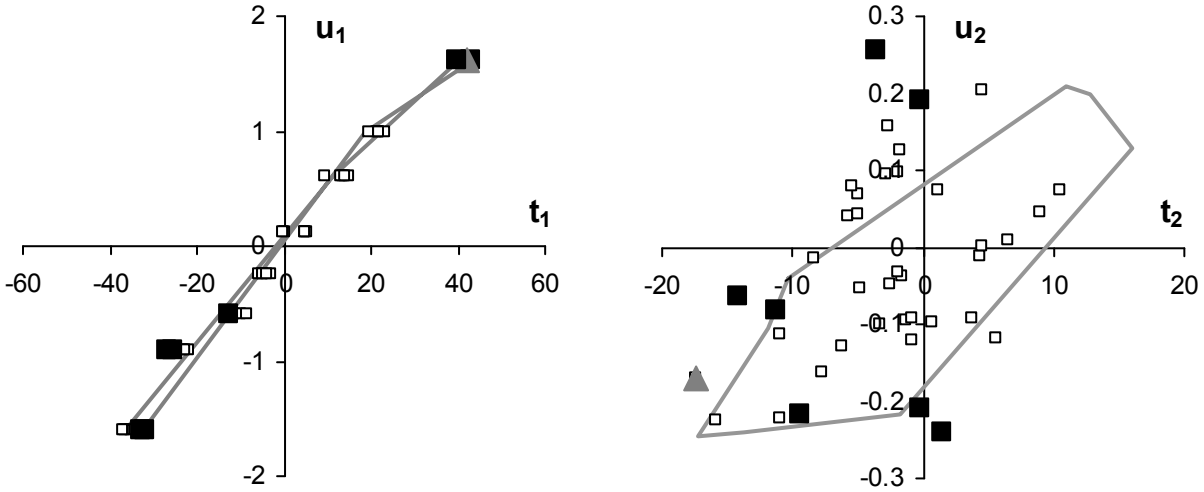


Рис. 8.7 Определение следовых концентраций нефти в воде. ПЛС модель с 2 ГК. График счетов **t-u.** Проверочный набор: □ - внутренние объекты, ■ - внешние объекты, ▲ - абсолютно внешние объекты, серая линия соединяет граничные объекты обучающего набора

Принадлежность объекта к тому или иному классу определяет качество прогноза, т.е. величину и положение прогнозного интервала, вычисленного ПИО методом. При ПЛС моделировании неопределенность в прогнозе оценивают двумя способами. Чаще всего это *оценка в среднем*, с помощью величины RMSEP (3.2), что никак не отражает индивидуальных особенностей конкретного объекта. Второй широко известный способ, это использование весьма приближенной формулы для оценки отклонений в прогнозе для *l*-ого объекта (3.4), предложенной Мартенсом [83] и реализованной в программе UNSCRAMBLER [83]. Обращает на себя внимание то, что неопределенность в прогнозе нового объекта, вычисляемая по формуле (3.4), зависит напрямую от проверочного набора и лишь косвенно от самой модели. Хотя значение отклика для *l*-ого объекта предсказывается непосредственно по модели. Это означает, что если взять другой проверочный набор, то неопределенность в прогнозе для того же объекта может быть

другой, хотя само предсказанное значение не изменится. Сравним результаты ПИО прогноза с интервалами неопределенности, вычисленными по формуле (3.4). На Рис. 8.8, для наглядности, изображены лишь 12 объектов из проверочного набора, белыми кружками обозначены известные проверочные значения откликов, черными кружками – ПЛС прогноз. Черные отрезки вокруг них, это удвоенное стандартное отклонение ($\pm 2s(\hat{y}_l)$) вычиненные по формуле (3.4), серые интервалы – это результат предсказания метода ПИО.

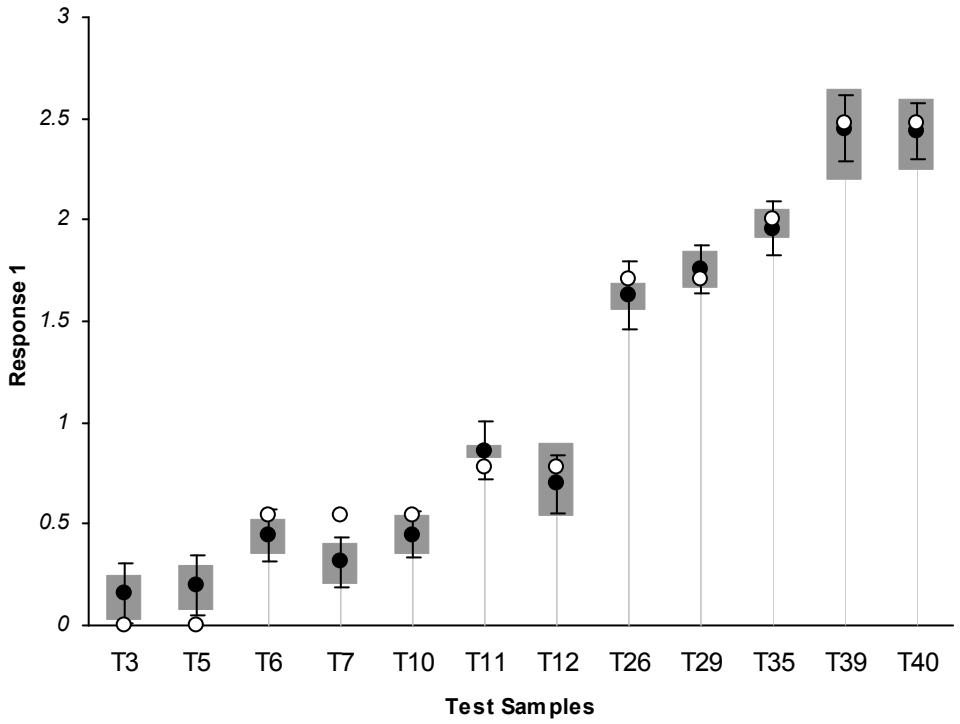


Рис. 8.8 Определение следовых концентраций нефти в воде. Проверочный набор., (●) –ПЛС предсказанное значение $\pm 2s(\hat{y}_l)$ серый интервал – ПИО прогноз, (о) – опорное значение

В среднем, оба построенных интервала близки, однако видно, что для объектов, которые являются внутренними для ПИО модели, неопределенность в прогнозе, вычисленная по формуле (3.4) завышена, например, для образцов NN 11, 26, 35, а для абсолютно внешнего объекта N 39, наоборот, занижена.

Таким образом, ПИО метод не только позволяет выявить граничные объекты в калибровочном наборе, но и представляет подробную информацию для индивидуальной классификации объектов проверочного набора.

8.4. Исследование выбросов

После того как калибровочная модель построена, она обычно используется для предсказания откликов для новых объектов. Если предсказываемый объект плохо согласуется с моделью, результат предсказания будет плохим, т.е. будет большая неопределенность в прогнозе, или хуже того, неверным, когда предсказанное значение и прогнозный интервал будут лежать далеко от измеренного значения. Обычно, такие объекты называются выбросами в прогнозе, однако общепринятой теории трактовки таких выбросов пока не существует, хотя такие попытки и делаются [118-120]. С точки зрения метода ПИО, величина интервала предсказания для новых объектов, которые сильно отличаются по структуре данных в предикторах, всегда будет больше, чем максимальная погрешность калибровки, β. Такие объекты называются абсолютно внешними, для их определения есть строгое правило (Утверждение 6.10). При этом применение формулы (6.15) не требует знания измеренного значения отклика, что и требуется для определения выбросов для новых объектов.

Сравним подход, предлагаемый методом ПИО с методом выпуклых оболочек [118].

Метод выпуклых оболочек

Этот метод проверяет, лежит ли новый объект внутри границ выпуклой оболочки, построенной на основе периферийных объектов в пространстве предикторов, или нет. Метод применим для любой размерности пространства предикторов/счетов, однако в двух- и трехмерном случае он особенно нагляден. В рассматриваемом нами примере, ПЛС модель построена с использованием 2 ГК, следовательно, пространство счетов в данном случае представляет плоскость $\mathbf{t}_1$-$\mathbf{t}_2$, и все построения можно наглядно продемонстрировать на графиках. Построение выпуклой оболочки осуществляется следующим образом:

1. В пространстве счетов (предикторов), среди объектов калибровочного набора определяется объект, A1, максимально удаленный от центра облака точек. В нашем случае, это объект N 37 (см. Рис. 8.9).
2. Первая грань выпуклой оболочки соединяет точку A1 с такой точкой A2, что все оставшиеся объекты и центр облака распложены по одну сторону от этой грани. В нашем случае, это прямая между объектами N37 и N11.

3. Следующая грань строится по тому же правилу, начиная с точки A2. Построение проводится до тех пор, пока граница не получится замкнутой. Построенный таким образом многогранник называется первичной оболочкой.

После того как первичная оболочка построена, она может быть использована для определения выбросов. Для этого вектор предикторов **x** нового объекта проецируется на пространство модели, т.е. в на пространство счетов. Если проекция объекта лежит внутри первичной оболочки, то объект хороший (Рис. 8.9, объект r1). Если же объект находится за границами оболочки (Рис. 8.9, объект r2), то такой объект трактуется как выброс.

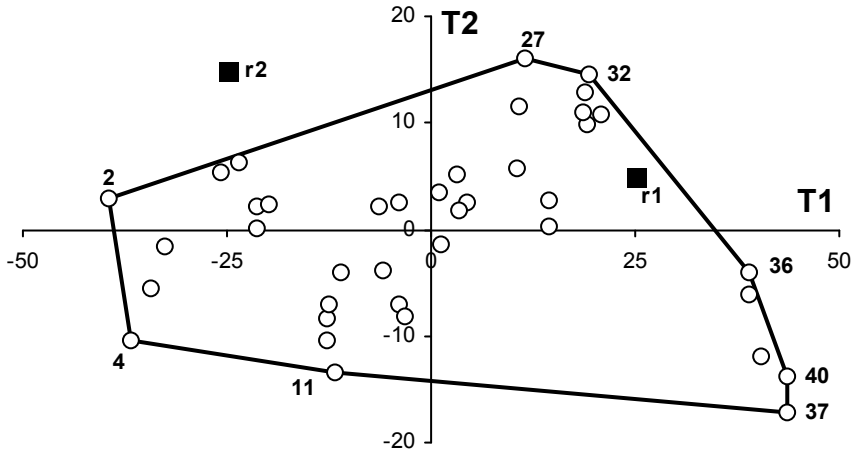


Рис. 8.9 Определение следовых концентраций нефти в воде. ПЛС модель с 2 ГЛ. ○ объекты из калибровочного набора, —. выпуклая оболочка, ■- проверочные объекты.

Для того чтобы этот метод сделать более чувствительным, в работе [118] предлагается дополнительно учесть погрешность моделирования, и построить вокруг первичной оболочки вторичную границу.

Для наглядности, опишем способ построения вторичной оболочки в двумерном пространстве.

1. Возьмем одну из вершин первичной оболочки, A. Вычислим расстояние от вершины A, до двух смежных вершин A_1 и A_2. Найдем минимальное из двух расстояний, например, $d(A_1)$.
2. Расстояние $d(A_1)$ откладывается на отрезке [A, A_2], начиная от вершины A. Тем самым, на отрезке [A,A_2]. получается новая точка A_{new}, и строится треугольник, $A_1A_{new}A$.
3. Вычисляется середина стороны A_1A_{new}, - точка $H_{1/2}$.

4. Прямая линия, соединяющая точки $H_{1/2}$ и A, продляется за область первичной оболочки на величину R_i и получается вершина вторичной оболочки.
5. Такая процедура проводится со всеми вершинами первичной оболочки. Новые вершины вторичной оболочки соединяются, образуя расширенную границу для определения выбросов. Вторичная оболочка изображена на Рис. 8.10, сплошная черная линия.

   Для вычисления величины R_i используется формула

$$R_i = \frac{R(x_i)}{R(\mathbf{X}_{cal})} \qquad (8.3)$$

где $R(x_I)$ – среднеквадратичный остаток в предикторах для i-ого объекта, $R(\mathbf{X}_{cal})$-среднеквадратичный остаток в предикторах для всех обучающих объектов, кроме i-го объекта. Т.е. для вычисления всех R_i необходимо провести процедуру подобную перекрестной проверки, с выкидыванием поочередно объектов, образующих вершины первичной оболочки.

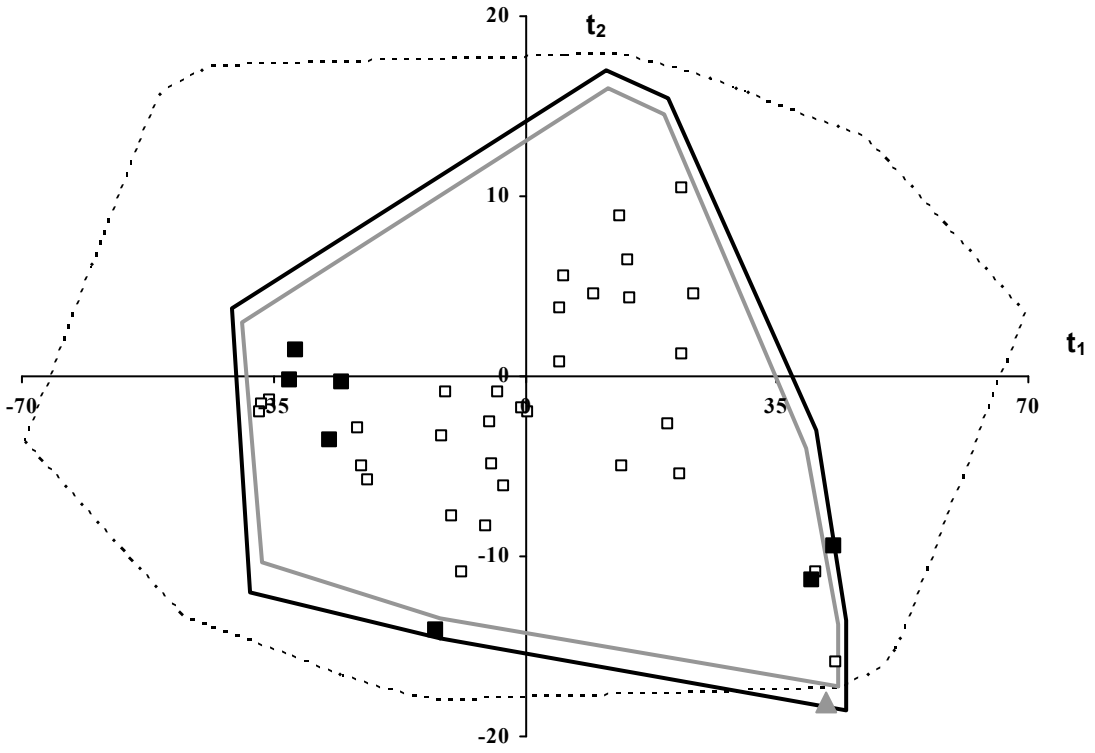


Рис. 8.10 Определение следовых концентраций нефти в воде. ПЛС модель с 2 ГК. Проверочные объекты: □ - внутренние объекты, ■ - внешние объекты, ▲ - абсолютно внешний объект (T39), — первичная оболочка, — вторичная оболочка, ---- область абсолютно внешних объектов.

Из Рис. 8.10 видно, что, хотя оболочки строились по объектам из калибровочного набора, все объекты проверочного набора (кроме трех объектов) лежат внутри первичной

оболочки. Что касается вторичной оболочки, то в ней лежат все проверочные объекты, даже объект Т39 (серый треугольник в нижней части графика), который классифицировался методом ПИО как абсолютно внешний, и прогноз на который не надежен.

Область абсолютно внешних объектов

ПИО метод также предлагает способ построения в пространстве предикторов области, которая определяет выбросы, т.е., в терминологии ПИО, абсолютно внешние объекты. Для построения такой области используется утверждение 6.11. Там же, в разделе 6.4, приведен алгоритм ее построения. Для построения границы области абсолютно внешних объектов используются значения проекции калибровочных объектов на область предикторов, а также результаты ПИО моделирования (формула (6.23)). Результат построения приведен на Рис. 8.10, пунктирная линия. В результате получается, что область внутри границы абсолютно внешних объектов больше, чем внутренние области, ограниченные первичной и вторичной оболочками. В то же время, видно, что абсолютно внешний объект (Т39), классифицируется ПИО методом как выброс, в отличие от метода оболочек. Существенным отличием в способе построения этих границ является то, что метод оболочек учитывает только значения предикторов калибровочных образов, в то время как метод ПИО принимает во внимание и результаты моделирования отклика.

8.5. Результаты главы 8

В этой главе, на примере анализа результатов многоканальных акустических измерений, показано, как можно объединить новый метод ПИО с хорошо известными в хемометрике методами билинейного моделирования (РГК, ПЛС). В результате появляется новый мощный инструмент для визуализации и детального анализа сложных многофакторных данных. Этот инструмент дает возможность определять статус каждого объекта, как из калибровочного, так и из проверочного наборов. При использовании современных сложных методов исследования, исследователь в результате проведения эксперимента получает большие наборы экспериментальных данных, которые сложно как осознать, так и проанализировать. Поэтому, достаточно часто, часть данных вообще не принимается во внимание, либо полученные данные обрабатываются крайне формально. Возможность визуализации больших массивов многоканальных экспериментальных данных возвращает исследователя к сути изучаемых явлений,

помогая проследить имеющиеся физические зависимости, оценить качество проведенного эксперимента.

Основой для такой визуализации служит классификация статуса объектов, которая непосредственно вытекает из метода ПИО.

Метод ПИО предлагает однозначные правила для классификации.

1. Все калибровочные объекты можно разделить на два класса: граничные, наиболее важные объекты, которые формируют модель и внутренние объекты, которые являются избыточными для формирования модели. (Утверждения 6.6-6.8).

2. Все проверочные объекты можно детально классифицировать (Утверждения 6.7, 6.9, 6.10). Их можно разделить на два основных класса: внутренние, эти объекты наиболее типичны и похожи на объекты из калибровочного набора, и внешние объекты. Среди внешних объектов можно ввести дополнительное разделение: абсолютно внешние объекты, эти объекты сильно отличаются от объектов калибровочного набора по предикторам, **X** матрица; выбросы – это объекты, противоречащие данной модели.

3. Для новых объектов, у которых значения откликов *y* неизвестно, имеется строгое правило (Утверждения 6.10), выделяющее абсолютно внешние объекты, которые плохо совместимы с построенной моделью. Это является существенным достижением ПИО метода, так как гарантирует, что при использовании модели для вновь полученных объектов мы не выйдем за область действия модели, т.е. будем находиться в условиях интерполяции, что чрезвычайно важно при формальном моделировании.

9. Сравнение содержательного и формального подхода к интерпретации кинетических данных на примере анализа данных ДСК эксперимента и длительного термостарения

Анализ кинетических данных – это необходимый этап в исследовании механизма процессов и предсказания практически важных свойств. Оценка кинетических параметров по экспериментальным данным, т.е. решение обратной задачи химической кинетики, как отдельное направление, сформировалось в 70х – 80х годах 20-го века. В разработке этого направления принимали участие как математики [244, 245], так и кинетики [246-249]. Традиционно, для решения подобных задач, применяется *содержательное* физико-химическое моделирование [96, 249-251], которое базируется на основных кинетических принципах и позволяет получать оценки параметров с высокой точностью. Однако такой метод может быть применен только тогда, когда модель процесса известна априори. Альтернативой является формальный подход, в котором кинетическая модель явно не используется. При таком подходе [71, 95, 100, 252] данные физического эксперимента описываются линейной многофакторной моделью, справедливой в ограниченном диапазоне условий

Оба подхода имеют свои сильные и слабые стороны. Исторически сложилось так, что в России интенсивно развивался содержательный подход, тогда как на западе отдавали преимущество формальным методам [253].

В этой главе проводится сопоставление этих методов, рассматриваются методологические вопросы, актуальные для анализа кинетических данных. Здесь применен как содержательный подход, основанный на нелинейном регрессионном анализе, так и формальный подход, включающий проекции на латентные структуры, в сочетании с методом простого интервального оценивания. Использование одного и того же массива данных, позволяет сравнить оба подхода и сделать выводы о том, в каком случае, какой подход предпочтительнее.

Сопоставление содержательного и формального подходов к анализу кинетических данных проводится на примере оценки эффективности антиоксидантов (АО) в полиолефинах [254]. Необходимо сразу оговориться, что речь не идет о предсказании возможной активности *будущего соединения* хорошо известными методами QSAR [31] по известным значениям молекулярных дескрипторов. В рассматриваемом примере экспериментально проверяется качество уже *существующих соединений*, о которых

имеются веские основания предполагать, что они могут быть использованы в качестве АО. Для этого используются кинетические данные, полученные методом дифференциальной сканирующей калориметрии (ДСК). Такой подход позволяет значительно ускорить процесс тестирования АО, отказавшись от традиционного длительного и дорогостоящего метода, который предполагает выдержку образцов в печах в течение 1-3 месяцев. Кинетические данные представляют значения температур начала окисления, полученные методом ДСК для различных скоростей нагрева. Они образуют **X** матрицу. В качестве откликов, используется матрица **Y**, состоящая из значений периодов индукции окисления, полученных традиционным путем. Мы использовали две калибровочные модели Y=F(X): формальную линейную зависимость, и содержательную нелинейную модель, в которой учитываются современные представлений о механизме окисления полиолефинов [255]. Для построения первой применяется множественная линейная регрессия [68], в которой прогнозные интервалы получаются методом ПИО [52]. Во второй, содержательной модели используется нелинейный регрессионный анализ [53] и строятся традиционные доверительные интервалы с помощью метода последовательного байесовского оценивания [256]. Таким образом, для одного и того же массива экспериментальных данных, были применены методы формального и содержательного моделирования. Это позволило сравнить полученные результаты, и понять, в каком случае, какой подход предпочтительнее. Основной материал этой главы опубликован в работах [223, 257]

9.1. Оценка активности антиоксидантов

Исследование эффективности антиоксидантов в полиолефинах – это длительный и дорогой процесс. Стандартная процедура тестирования АО предполагает длительную выдержку образцов при постоянной температуре. Однако можно применять и альтернативный подход, который позволяет предсказывать активность АО быстрее. Идея использовать для этого метод ДСК была предложена и исследована ранее [53, 254, 258]. В этих работах применялся как содержательный [258], так и формальный [254] подходы для построения модели. Однако использованные данные были небольшого размера. В этой главе и работе [223] исследуется довольно представительный набор образцов АО, который позволил решить две важные задачи. Первая – это практический вопрос о применимости предлагаемого альтернативного метода. Вторая задача состоит в

сравнении содержательного и формального подходов к моделированию на одном и том же наборе экспериментальных данных.

Антиоксиданты – это специальные добавки, которые замедляют термоокислительное старение полимеров. Они защищают материал от окисления в ходе изготовления материала и при эксплуатации изделий. Взаимодействуя со свободными радикалами, АО обрывает цепи, но при этом и сам расходуется. АО полностью подавляет окисление до тех пор, пока его концентрация не упадет ниже некоторого критического значения. Поэтому основной характеристикой эффективности АО является *период индукции* (ИП), время, в течение которого концентрация АО достаточна велика. Чем больше период индукции (при той же температуре) – тем более эффективен АО. По сложившейся практике период индукции определяется с помощью *длительного термического старения* образцов при температуре 120^{o}C-140^{o}C. Время выдержки зависит от качества АО и может составлять от 1 дня для плохого АО, до 100 дней для хорошего АО. В течение этого времени экспериментатор обследует образцы в поисках явных признаков деградации – растрескивания, пожелтения, и т. п. Разумеется, такой метод является ненадежным и требует больших затрат времени и средств.

Альтернативой этому методу является подход, использующий ДСК измерения, с последующей математической обработкой полученных данных. Дифференциальная сканирующая калориметрия – это метод, в котором сигналом является величина теплового потока от образца, нагреваемого с постоянной скоростью. Если в образце происходит химическая реакция с ненулевым тепловым эффектом, то величина ДСК сигнала пропорциональна скорости реакции. Применительно к рассматриваемой задаче, это означает, что пока концентрация АО в образце еще достаточна для подавления цепного окисления, то ДСК сигнал является константой, однако, начиная с некоторой температуры, он начинает резко расти. Эта температура называется *температурой начала окисления* (ТНО). Проводя эксперименты при разной *скорости нагрева v*, можно получить несколько значений ТНО.

9.2. Эксперимент

В эксперименте использовались 25 образцов АО обозначенные как АО-1, … , АО-25. Некоторые из них (например, АО-1, АО-2 и АО-3) являются стандартными добавками, используемыми в промышленности. Другие образцы – это новые вещества, о которых имеются обоснованные предположения об их эффективности. Все образцы были

добавлены в порошок *полипропилена* (ПП) в концентрациях 0.05% (500 ppm), 0.07%, и 0.1%. После смешивания, композиция была подвергнута экструзии при температуре 250^{o}C и превращена в пленки толщиной 0.25 мм. Эти пленки и были исходным материалом для проведения ДСК измерений и долговременного старения. Старение проводилось в термошкафах при температуре 140^{o}C. В результате этого эксперимента мы получили значения периодов индукции (ИП), выраженные в днях.

ДСК измерения проводились в температурном диапазоне от 150^{o}C до 350^{o}C, где наблюдается экзотермический максимум, связанный с окислением полимера. При этом использовались пять различных скоростей нагрева 2, 5, 10, 15, 20 (град/мин). Значения температур начала окисления (ТНО) вычислялись с использование программы Fitter [259], способом, объясненным в [53].

В результате всех экспериментов мы получили данные, которые схематически изображены на Рис. 9.1. Данные ***X*** являются значениями температур начала окисления (ТНО) полученными в ДСК эксперименте. Они образуют трех модальный (3-way) [82] блок: 25 образцов АО × 3 концентрации АО × 5 скоростей нагрева. Данные **Y** – это значения периодов индукции (ИП), полученные с помощью длительного термического старения. Они образуют двух модальный блок (матрицу): 25 образцов АО × 3 концентрации АО.

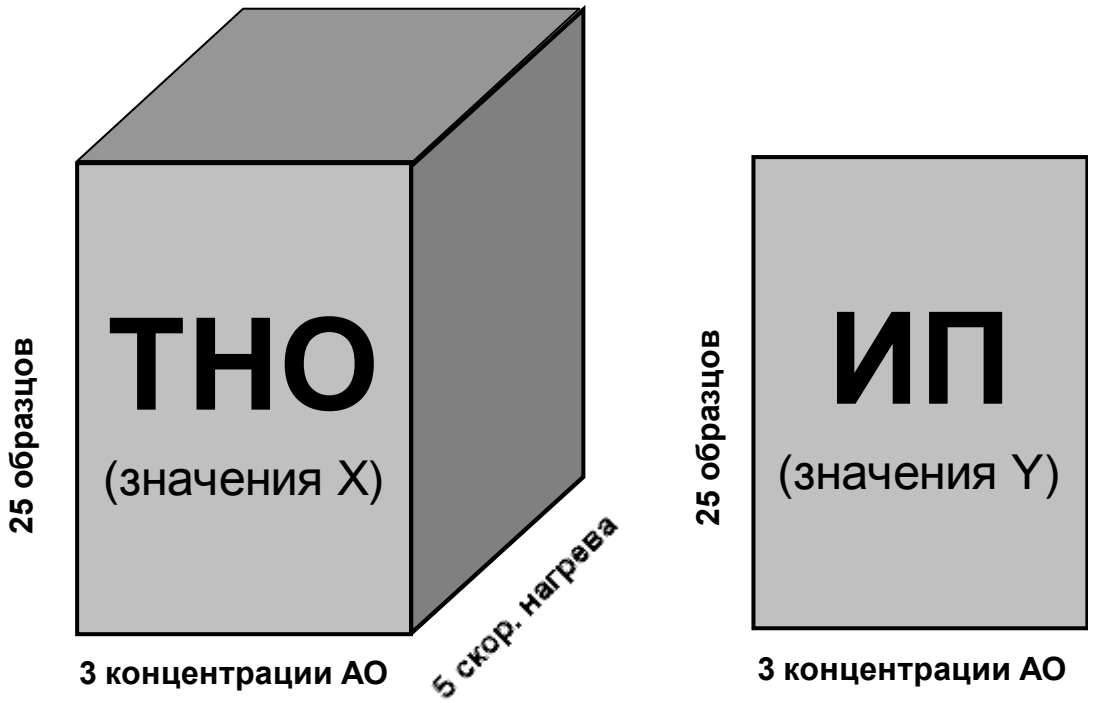


Рис. 9.1 Устройство экспериментальных данных

9.3. Формальное моделирование

Полученный массив данных обрабатывался с помощью формального подхода, который объединяет метод *проекций на латентные структуры* (ПЛС) [67, 68] для калибровки и метод *простого интервального оценивания* (ПИО) [52, 194] для построения прогнозных интервалов.

Исходные 3-х модальные ***X*** данные раскладывались в плоскую матрицу (25 × 15), так, как показано на Рис. 9.2. Важным является то обстоятельство, что погрешность в значениях периода индукции не постоянна, а растет с величиной ТНО. Это может быть объяснено двумя обстоятельствами. Первое связано с вышеупомянутым методом визуальной инспекции, который обладает худшей точностью для более стабильных (т.е. долгоживущих) АО. Вторая причина заключается в процедуре приготовления образцов, когда малое количество АО смешивается с большим количеством ПП. Это приводит к некоторой неизбежной неоднородности распределения АО в ПП. Для того чтобы скомпенсировать непостоянство погрешности измерения ИП, мы извлекли квадратный корень из всех значений периодов индукции. Кроме того, при ПЛС моделировании, мы центрировали **X** и **Y** данные.

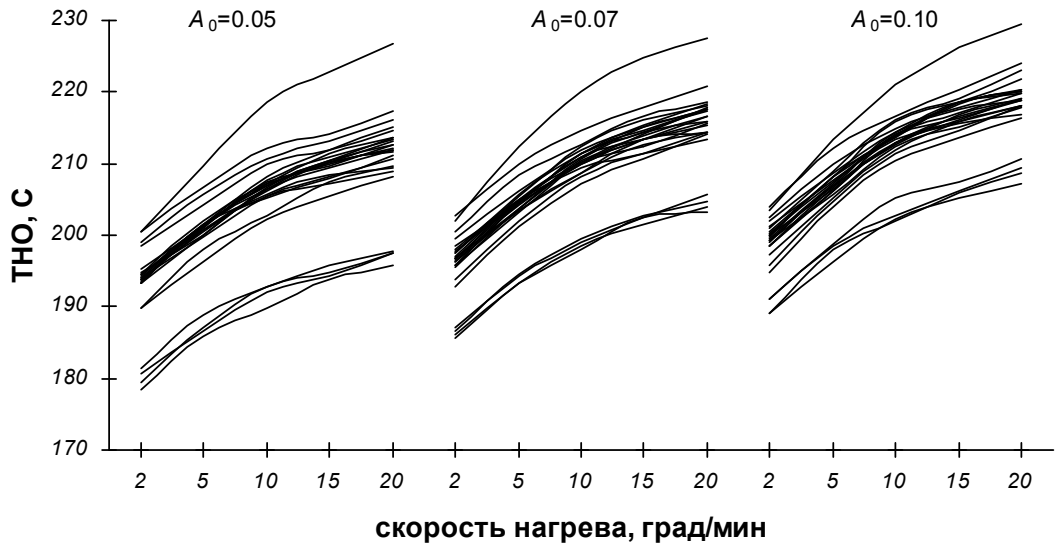


Рис. 9.2 Устройство **X** данных для ПЛС моделирования

При построении линейной модели калибровки

$$\mathbf{y}=\mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon} \qquad (9.1)$$

основная математическая трудность, как и при обработке результатов эксперимента в главе 8, состоит в обращении матрицы $\mathbf{X}^t\mathbf{X}$ (см. раздел 3.1), которая в нашем случае

имеет размерность 15 × 15. Если бы эта матрица была не вырождена (имела бы полный ранг), то такую калибровку можно было бы выполнить методом наименьших квадратов и найти оценки неизвестных параметров модели $\hat{\mathbf{a}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$. Однако, в нашем примере, как и в большинстве практических задач, эта матрица вырождена. В данном случае для регуляризации задачи используется ПЛС метод (раздел. 13.3)

Применяя метод ПЛС, мы разделили полный набор данных на две части. *Калибровочный* набор включал восемнадцать образцов (АО-1 – АО-18); эти данные использовались для построения модели. Второй, *проверочный* набор данных, состоял из семи образцов (АО-19 – АО-25). Они использовались для проверки модели. Для каждого значения A_0 исходной концентрации АО была построена своя собственная ПЛС модель

$$\mathbf{X}=\mathbf{T}\mathbf{P}^t+\mathbf{E} \qquad \mathbf{y}=\mathbf{T}\mathbf{q}+\mathbf{f}, \tag{9.2}$$

с двумя главными компонентами. Основные характеристики этих моделей приведены в Таб. 9.1

Таб. 9.1 Главные характеристики ПЛС моделей, построенных для различных начальных значений концентраций АО

$\mathbf{A_0}$	$\mathbf{X_{expl}}$	$\mathbf{Y_{expl}}$	**RMSEC**	$\mathbf{R^2_{cal}}$	$\mathbf{R^2_{test}}$	**β**
0.05	99%	92%	0.287	0.96	0.99	0.84
0.07	99%	88%	0.342	0.93	0.99	1.02
0.10	99%	84%	0.395	0.91	0.97	1.20

В этой таблице для каждой начальной концентрации АО (A_0) представлены следующие величины, характеризующие качество построенной модели. Объясненные вариации в **X** и в ***y*** вычисляются как усредненные нормированные невязки в модели (9.2) по калибровочному набору ($\mathbf{X}_{cal}$, $\mathbf{y}_{cal}$):

$$X_{expl} = 1 - \frac{\left\|\mathbf{X}_{cal} - \mathbf{T}\mathbf{P}^t\right\|^2}{\left\|\mathbf{X}_{cal}\right\|^2} \qquad Y_{expl} = 1 - \frac{\left\|\mathbf{y}_{cal} - \mathbf{T}\mathbf{q}\right\|^2}{\left\|\mathbf{y}_{cal}\right\|^2}$$

Они характеризуют то, насколько точно ПЛС метод приблизил (объяснил) исходные данные. Среднеквадратичный остаток калибровки (3.1) вычисляется по данным калибровочного набора, его можно записать как

$$\mathrm{RMSEC} = \frac{\left\|\mathbf{y}_{cal} - \mathbf{T}\mathbf{q}\right\|}{\sqrt{I_{cal} - K}}, \tag{9.3}$$

Он характеризует точность моделирование отклика. В приведенных выше формулах I_{cal} – это число данных в калибровочном наборе, а K – это число главных компонент. Норма матрицы (вектора) – это корень квадратный из суммы квадратов ее (его) компонент. Коэффициенты корреляции измеренных значений с предсказанными значениями для калибровочного (R^2_{cal}) и проверочного (R^2_{test}) наборов, вычисляются обычным образом.

Используя ПЛС метод, мы получаем среднюю, точечную оценку периода индукции. Для того чтобы найти интервальную прогнозную оценку, применяется метод ПИО. Величина β характеризует ошибку калибровки и может быть оценена традиционными статистическими методами (см. Раздел 5.4). Значения β, найденные для исследуемых данных, представлены в последнем столбце Таб. 9.1.

На Рис. 9.3 показаны результаты ПЛС/ПИО прогноза исследуемых данных для исходной концентрации АО $A_0 = 0.05$. Здесь предсказанные значения периода индукции (ИП) изображены в зависимости от соответствующих измеренных значений, причем из обеих величин извлечен квадратный корень. Каждый объект (открытые точки представляют калибровочные объекты, а закрытые – проверочные) показан вместе с двумя интервалами погрешностей. Горизонтальные отрезки представляют калибровочные интервалы для измеренных величинах ИП, и они равны удвоенному значению максимальной погрешности β, для всех объектов (6.2). Вертикальные отрезки соответствуют прогнозным интервалам (5.17), и они различны для разных объектов. Видно, что для всех калибровочных объектов прогнозный интервал всегда меньше или равен калибровочному интервалу. Это является фундаментальным свойством, непосредственно следующим из теории ПИО метода (см. утверждения 6.1 и 6.2).

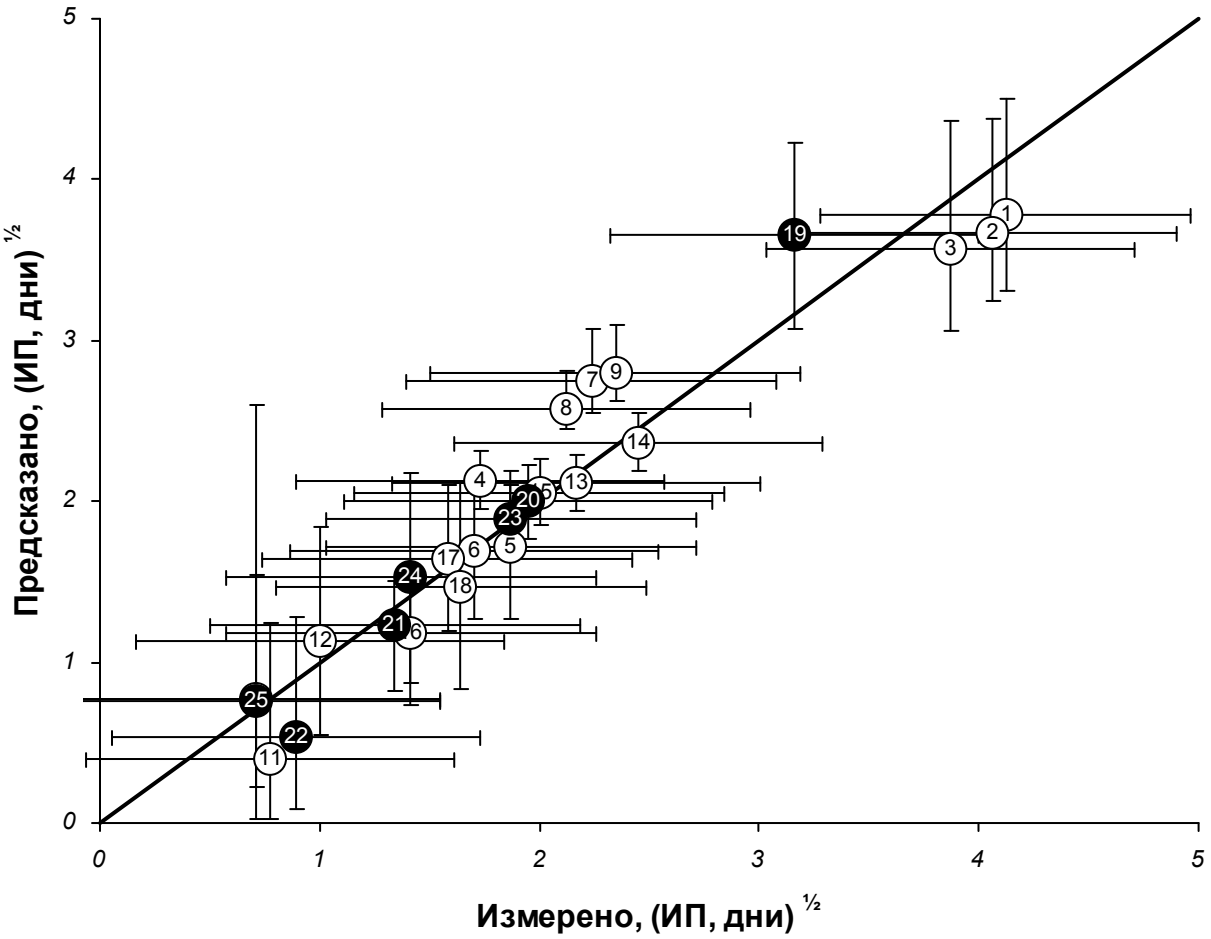


Рис. 9.3 Прогноз периода индукции (ИП) формальным методом для АО с начальной концентрацией 0.05. Корень квадратный из предсказанных и измеренных значении. Открытые точки (○) – калибровочные, а закрытые точки (●) –проверочные объекты. Горизонтальные отрезки показывают β (ошибку моделирования), а вертикальные отрезки – ПИО (прогнозные) интервалы

9.4. Содержательное моделирование

Содержательные модели для прогноза периода индукции окисления строятся для каждого типа АО индивидуально. Для этого мы выделяем в полном массиве данных, показанном на Рис. 9.1, 25 горизонтальных «срезов» и получаем, соответственно, 25 наборов данных: матрицы $\mathbf{X}_i$ размерностью 5 × 3, и векторы $\mathbf{y}_i$ длиной 3.

Процедура калибровки этих данных состоит из двух шагов. На первом шаге строится модель, которая описывает расход антиоксиданта в ходе ДСК эксперимента – это калибровка **X** данных. Регрессионная модель является неявной функцией, связывающей температуру начала окисления (ТНО) T, начальную концентрацию АО A_0, и скорость нагрева v. Эта функция нелинейно зависит от неизвестных кинетических параметров, которые оцениваются методом *нелинейного регрессионного анализа* (НЛР). На втором шаге моделирования мы строим модель для описания расхода АО в ходе

длительного термического старения – калибровка **Y** данных. Эта регрессионная модель явно выражает период индукции как функцию температуры экспозиции T_e и начальной концентрации АО A_0. В этой функции участвуют те же кинетические параметры, что и в первой модели. Их оценки мы находим на первом шаге, а на втором мы применяем специальную процедуру переноса ошибок для того, чтобы оценить неопределенность в прогнозе периода индукции.

Рассмотрим первый шаг моделирования. В ходе старения материала происходит расход антиоксиданта. Период индукции определяется моментом времени, когда АО израсходуется до некоторого критического значения A_c, которое зависит от температуры по закону Аррениуса [255]

$$A_c = k_c \exp\left(-\frac{E_c}{RT}\right) \tag{9.4}$$

Расход АО в ходе окисления идет термически, по закону

$$\begin{aligned} &\frac{dA}{dt} = -kA \\ &A(0) = A_0 \end{aligned} \tag{9.5}$$

где A – текущая концентрация АО, A_0 – начальная концентрация АО, k – константа скорости, зависящая от температуры по закону Аррениуса

$$k = k_a \exp\left(-\frac{E_a}{RT}\right)$$

В ДСК эксперименте образцы разогреваются со скоростью v от начальной температуры T_0=293К

$$T(t) = T_0 + vt\,.$$

В этом случае решение уравнения (9.5) можно представить в виде

$$A(t) = A_0 \exp\left(-k_0 Z(t)\right) \tag{9.6}$$

где функция $Z(t)$ имеет вид

$$Z(t) = \int_0^t \exp\left(-\frac{E_a}{R(T_0 + vs)}\right) ds\,.$$

Этот интеграл может быть представлен с использованием стандартной интегральной показательной функции $\mathrm{E}_n(z)$

$$\mathrm{E}_n(z) = \int_1^\infty t^{-n} e^{-zt}\, dt, \quad n = 0,1,2,...; \quad z > 0\,.$$

в виде, зависящем не от времени t, а от температуры T

$$Z(T)=\frac{1}{v}\left[T\,\mathrm{E}_2\left(\frac{E_\mathrm{a}}{RT}\right)-T_0\,\mathrm{E}_2\left(\frac{E_\mathrm{a}}{RT_0}\right)\right]. \qquad (9.7)$$

Подставив (9.7) в (9.6) и приравняв текущую концентрацию АО критической концентрации A_c из уравнения (9.4), получаем уравнение

$$A_0\exp\left(-k_0 Z(T)\right)=k_\mathrm{c}\exp\left(-\frac{E_\mathrm{c}}{RT}\right), \qquad (9.8)$$

которое неявно определяет зависимость температуры начала окисления T от неизвестных параметров и известных условий эксперимента. Однако практически использовать это уравнение очень неудобно. Приведем его в форму, которая более подходит для оценивания параметров. После логарифмирования и некоторых упрощений получаем

$$\exp(a)\,\mathrm{E}_2\left(\frac{E_a}{RT}\right)T+v\left(c-a-\frac{E_\mathrm{c}}{RT}\right)=0, \qquad (9.9)$$

где $a=\ln(k_\mathrm{a})$ и $c=\ln(k_\mathrm{c})$.

В этом уравнении температура T рассматривается как отклик, а скорость нагрева v и начальная концентрация A_0 – как предикторы (факторы эксперимента). Величины a, E_a, c и E_c – это неизвестные параметры, которые оцениваются по экспериментальным данным. При поиске оценок мы должны разрешать уравнение (9.9) относительно переменной T

$$T=T(v,A_0;\ a,E_\mathrm{a},c,E_\mathrm{c}) \qquad (9.10)$$

многократно, для каждого заданного набора значений предикторов v и A_0, и искать такие значения (оценки) параметров a, E_a, c, и E_c, которые минимизируют сумму квадратов отклонений

$$\min_{a,E_\mathrm{a},c,E_\mathrm{c}}\sum_{ij}\left[Y_{ij}-T(v_j,A_{0i};a,E_\mathrm{a},c,E_\mathrm{c})\right]^2. \qquad (9.11)$$

Здесь Y_{ij} – это экспериментальные значения температуры начала окисления (ТНО), индекс i соответствует трем начальным концентрациям АО, а индекс j нумерует 5 скоростей нагрева v.

При минимизации суммы (9.11) мы сталкиваемся со значительными трудностями. Во-первых, мы не можем представить функцию (9.10) явно, и даже не можем неявно выразить ее через элементарные функции. Во-вторых, очевидно, что уравнение (9.9) задает функцию, которая нелинейно зависит от неизвестных параметров. И, наконец, минимизация суммы (9.11) является *жесткой* [246 – 248] задачей. Все эти

обстоятельства превращают подгонку модели в довольно сложную вычислительную задачу. Для ее решения использовалось специальное программное обеспечение, Fitter [53, 259], которое обладает возможностями для решения подобных задач. В этой программе реализован стабильный градиентный метод минимизации, который базируется на технике матричной экспоненты [250–251, 260]. Этот метод показал высокую эффективность при решении жестких задач, превосходящую популярный метод Левенберга-Марквардта [261, 262]. Кроме того, в программе Fitter, регрессионное уравнение можно задавать в естественной алгебраической форме, которая допускает и неявные выражения. Последнее, и очень важное свойство программы, состоит в том, что в ней вычисление производных проводится автоматически через символьное дифференцирование, что позволяет поддерживать высокую точность вычислений.

На Рис. 9.4 представлен пример калибровки первой модели для АО-1. Показаны экспериментальные данные (точки) и соответствующие калибровочные кривые.

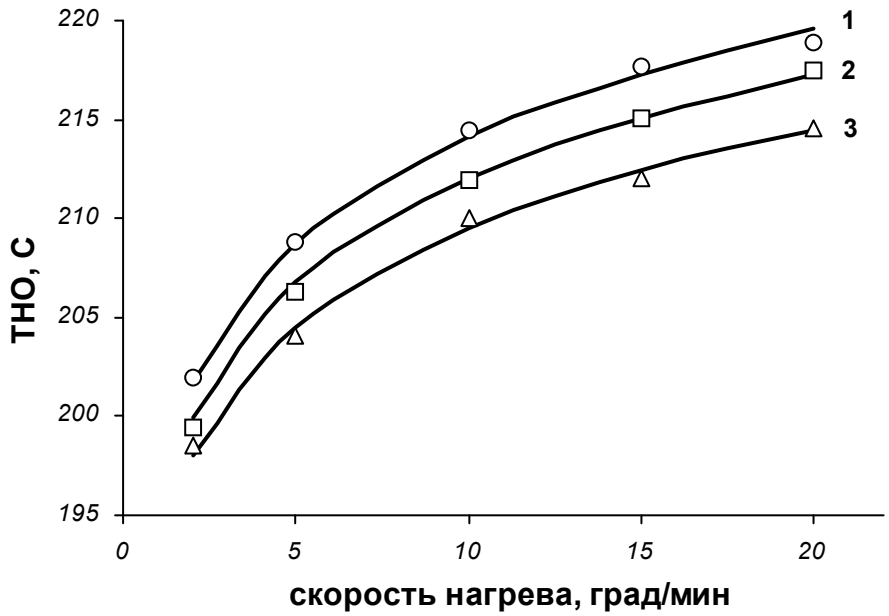


Рис. 9.4. Значения температур начала окисления (ТНО, точки и кривые) в зависимости от скорости нагрева для начальных концентраций АО: 0.1 (**1**, ○), 0.07 (**2**, □), 0.05 (**3**, △). ПП+ АО-1

Рассмотрим теперь второй шаг моделирования, на котором строится модель для описания расхода АО в ходе длительного термического старения. Решение уравнения (9.5) при постоянных условиях T=Const, имеет вид

$$A(t) = A_0 \exp(-kt).$$

Для того чтобы найти период индукции t_{ind}, мы должны приравнять текущую концентрацию $A(t)$ критической концентрации A_{c}, определенной в уравнении (9.5). Таким образом,

$$t_{\text{ind}} = \left[\frac{E_{\text{c}}}{RT_{\text{e}}} + \ln(A_0) - c \right] \exp\left(\frac{E_{\text{a}}}{RT_{\text{e}}} - a \right), \qquad (9.12)$$

где T_{e} – это температура экспозиции (старения). В нашем случае это 140°C, т.е. $T_{\text{e}} = 413\text{K}$.

В уравнении (9.12) мы должны использовать оценки параметров $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{E}_a, \hat{c}, \hat{E}_c)$, найденные на первом шаге калибровки. Для того чтобы получить не только точечную оценку периода индукции, но и определить ее доверительный интервал, необходимо принять во внимание неопределенность в этих оценках. К сожалению, мы не можем применить к уравнению (9.12) обычный способ переноса ошибок [201]

$$\text{var}(f) = \frac{\partial f(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}^{\text{T}} \text{cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) \frac{\partial f(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}},$$

из-за нелинейности модели. Точный доверительный интервал может быть получен с помощью статистического моделирования, точнее – методом свободного моделирования [263]. В этом методе доверительный интервал для функции $f(\boldsymbol{\theta})$ вычисляется как процентиль выборки $\{f(\boldsymbol{\theta}_1), f(\boldsymbol{\theta}_2), \ldots\}$, в которой значения $f(\boldsymbol{\theta}_{\text{i}})$ получаются в результате моделирования параметров $\boldsymbol{\theta}_{\text{i}}$ в соответствие с нормальным распределением $N\left(\hat{\boldsymbol{\theta}}, \text{cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\right)$.

Пример доверительного прогноза, построенного с помощью этой техники, для образца АО-18 показан на Рис. 9.5 Доверительная вероятность равна P=0.90. Доверительные интервалы для других образцов представлены на Рис. 9.6.

Важно, что при содержательном (НЛР) моделировании мы не использовали измеренные в ходе длительного старения значения периодов индукции. Поэтому мы можем использовать их при проверке нашей модели, вычислив среднеквадратичный остаток прогноза (RMSEP (3.2)),

$$RMSEP = \frac{\left\| \mathbf{y}_{\text{test}} - \mathbf{Tq} \right\|}{\sqrt{n_{\text{test}} - r}},$$

который показана в Таб. 9.2.

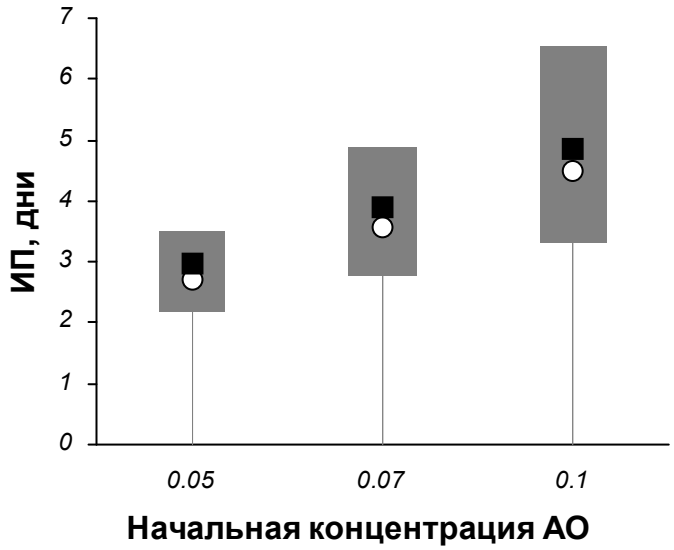


Рис. 9.5. Прогноз периода индукции (ИП) для различных начальных концентраций АО. Доверительные интервалы для вероятности P=0.90 показаны серыми прямоугольниками. Черные квадраты (■) представляют точечные оценки ИП, а открытые точки (○) соответствуют измеренным значениям. ПП+АО-18

В тоже время мы не можем оценить такую традиционную меру точности, как среднеквадратичный остаток калибровки RMSEC (9.3), поскольку на первой стадии моделирования мы моделировали не период индукции (ИП), а температуру начала окисления (ТНО) как функцию v и A_0. Однако можно найти остаточную сумму квадратов для каждого АО, и по ним вычислить среднее отклонение в оценках ТНО. Эти величины показаны в ряду 4 Таб. 9.2.

Таб. 9.2. Статистические характеристики прогноза содержательным (НЛР) и формальным (ПЛС/ПИО) методами

Начальная концентрация АО	НЛР (i=1, CI)			ПЛС/ПИО (i=2, PI)		
	0.05	0.07	0.10	0.05	0.07	0.10
1. RMSEP	0.242	0.246	0.272	0.239	0.251	0.336
2. Смещение	0.087	0.058	0.040	0.011	0.004	0.002
3. Корреляция $(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2)$	0.953	0.934	0.916	0.953	0.934	0.916
4. Среднее $(\mathbf{X} - \hat{\mathbf{X}})^2$		0.224		0.286	0.286	0.286
5. Среднее ($\boldsymbol{w}_i$)	1.038	1.151	1.397	0.934	1.204	1.476
6. Корреляция ($\boldsymbol{w}_1$, $\boldsymbol{w}_2$)	0.202	0.007	0.028	0.202	0.007	0.028
7. Корреляция ($\boldsymbol{y}$, $\boldsymbol{w}_i$)	0.815	0.846	0.836	–0.184	–0.161	–0.113

9.5. Сравнение методов

Займемся сравнением результатов содержательного (НЛР) и формального (ПЛС/ПИО) моделирования, которые были получены на одно и том же наборе экспериментальных данных. Эти модели используют данные различным образом. У нас имеются 25 содержательных моделей, построенных для каждого АО в отдельности, и 3 формальные модели, построенные для каждой начальной концентрации АО.

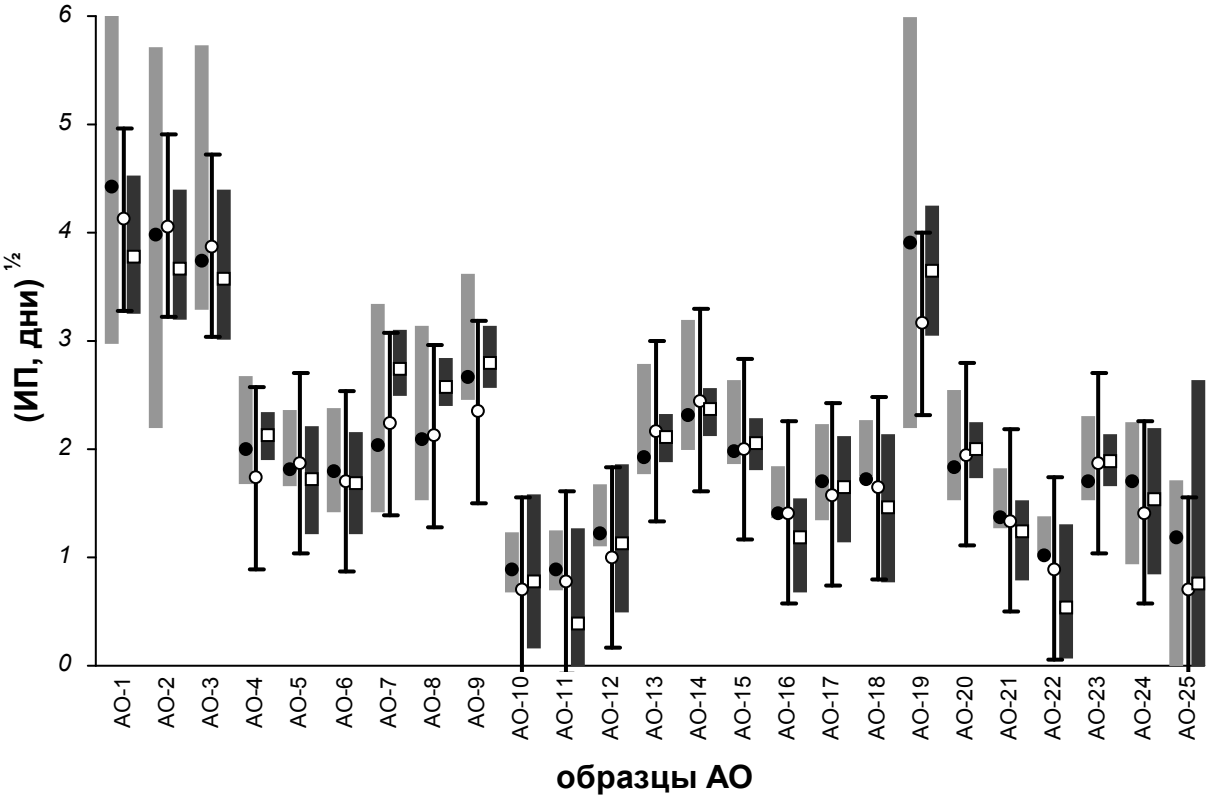


Рис. 9.6. Результаты прогноза периода индукции для различных образцов АО с начальной концентрацией 0.05. Черные точки (●) и серые прямоугольники представляют содержательное (НЛР) предсказание. Открытые квадраты (□) и черные прямоугольники изображают формальное (ПЛС/ПИО) моделирование. Открытые точки (○) соответствуют измеренным значениям с вертикальными отрезками, которые показывают погрешность измерения (калибровки) β. Из всех величин извлечен квадратный корень.

На Рис. 9.6 собраны все результаты, полученные каждым методом для начальной концентрации АО равной 0.05. На этом графике представлены результаты содержательного (НЛР) моделирования: – черные точки и серые прямоугольники доверительных интервалов для вероятности 0.90. Результаты формального метода (ПЛС/ПИО) показаны открытыми квадратами и черными прямоугольниками прогнозных интервалов. Референтные значения периодов индукции, измеренные методом длительного термического старения, представлены открытыми точками с интервалами калибровки, которые равны β – максимальной погрешности. Для того чтобы сделать

сравнение более наглядным, мы извлекли квадратный корень из всех значений периодов индукции (ИП). Такие же графики могут быть построены для других наборов данных, соответствующих начальной концентрации АО 0.07 и 0.10. Они выглядят аналогично.

В Таб. 9.2 представлены некоторые общие статистические характеристики содержательного (НЛР) и формального (ПЛС/ПИО) методов. Здесь использованы следующие обозначения для измеренных и предсказанных X (ТНО) и Y (ИП) значений. $\mathbf{X}$ и $\hat{\mathbf{X}}$ – это соответственно измеренные и предсказанные значения температур начала окисления (ТНО); $\boldsymbol{y} = \sqrt{\text{ИП}}$ – это корень квадратный из измеренного значения периода индукции (ИП); $\hat{\boldsymbol{y}}_i$ – это соответствующие векторы, полученные после извлечения квадратного корня из точечных оценок периода индукции (ИП), вычисленных содержательным (НЛР, i=1) и формальным (ПЛС/ПИО, i=2) методами; $\boldsymbol{w}_i$ – равны ширине доверительных (НЛР, i=1) и прогнозных (ПЛС/ПИО, i=2) интервалов, причем из всех интервалов также извлечен квадратный корень. Для сокращения мы будем использовать символы CI и PI для обозначения, соответственно, доверительных и прогнозных интервалов. Первые – были получены методом содержательного моделирования, НЛР, а вторые – формальным методом ПЛС/ПИО.

Из Таб. 9.2 и Рис. 9.6, можно сделать следующие выводы. Оба метода имеют близкую точность (ряд 1 в Таб. 9.2) и смещение (ряд 2). Точность прогноза становится хуже, когда начальная концентрация АО увеличивается. В целом, ПЛС/ПИО метод дает лучшие результаты для малых начальных концентраций АО, тогда как НЛР лучше для больших концентраций. Однако, точечные оценки в среднем близки (см. ряд 3). Оба метода хорошо моделируют значения $\mathbf{X}$ (т.е. величины температур начала окисления ТНО), но содержательный метод (НЛР) делает это немного лучше (ряд 4).

Интервальные оценки также очень близки в среднем (ряд 5), хотя доверительные интервалы (CI) могут сильно отличаться от прогнозных интервалов (PI) для отдельных образцов (см. ряд 6 и Рис. 9.6). Последний ряд Таб. 9.2 показывает, что ширина CI растет с увеличением значения периода индукции для всех начальных концентраций АО, тогда как ширина PI не зависит от y. Это свидетельствует о том, что примененное нами преобразование откликов $y = \sqrt{\text{ИП}}$, действительно дало ожидаемый эффект в ПЛС/ПИО моделировании, но не смогло исправить результаты НЛР моделирования.

Из Таб. 9.2 (ряд 5) видно, что ширина предсказанных интервалов растет с начальной концентрацией АО. Это видно и из содержательной модели (9.12), которая

представляет зависимость значения ИП t_{ind} от A_0. С другой стороны, в формальной ПЛС модели это никак нельзя было предвидеть. По-видимому, этот факт является фундаментальным свойством исследуемой полимерной системы, а именно, чем больше добавлено АО в образец, тем хуже мы можем предсказать его период индукции. Важно, что и содержательный, и формальный методы в этом смысле дают сходные результаты. При этом результаты обоих методов могут сильно отличаться для отдельных образцов. Из Рис. 9.6 мы видим, что для некоторых образцов (например, АО-5, 6, 10, 11) доверительные интервалы (CI) меньше, чем прогнозные интервалы (PI), тогда как для других образцов (например, АО-1, 2, 3) имеет место обратное. Было бы интересно понять, почему индивидуальные доверительные и прогнозные интервалы так отличаются.

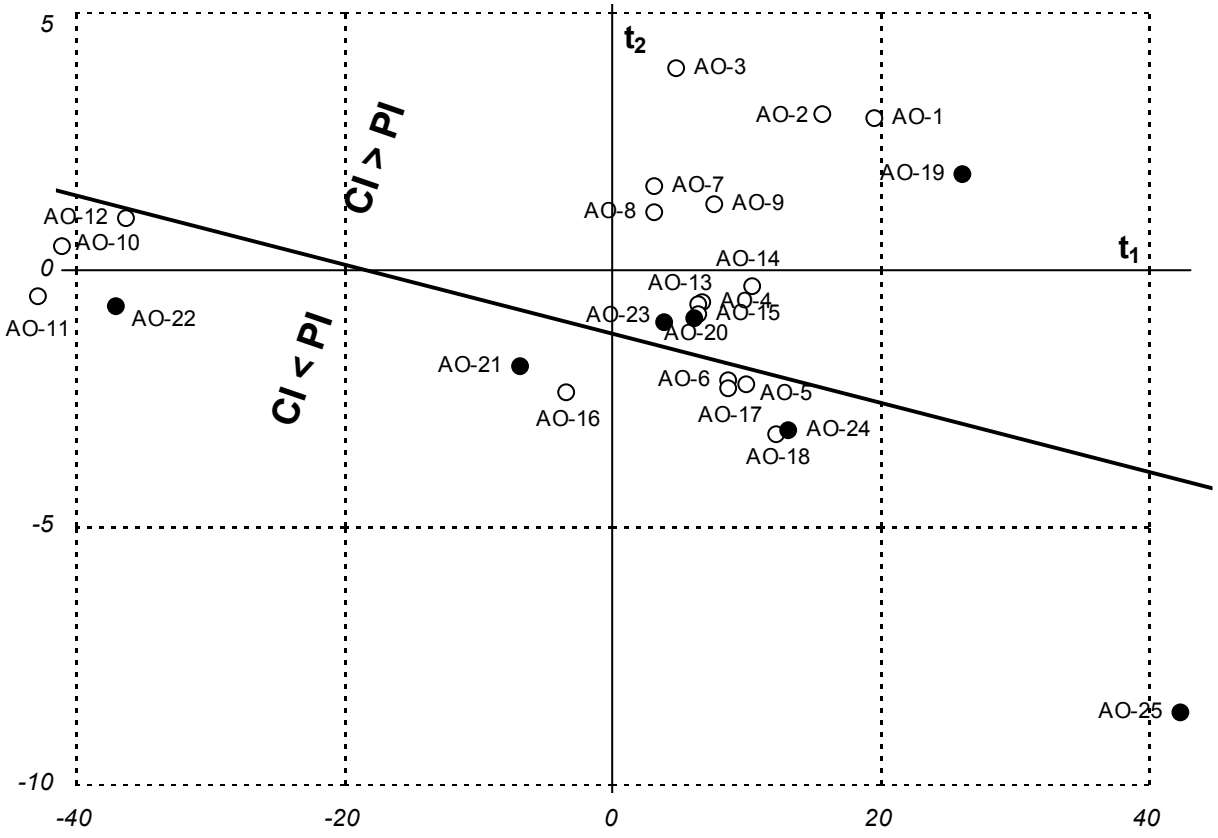


Рис. 9.7 График ПЛС счетов для образцов с A_0= 0.05. Открытые точки (○) представляют калибровочные, а закрашенные (●) – проверочные объекты. Линия разделяет объекты, для которых доверительные интервалы (CIs) меньше (больше), чем прогнозные интервалы (PIs).

На Рис. 9.7 показан график ПЛС счетов (т.е. проекций X данных на пространство двух главных ПЛС компонент) для образцов с начальной концентрацией АО 0.05. Восемнадцать калибровочных объектов (АО-1, … , АО-18) представлены открытыми точками, а шесть проверочных объектов (АО-19, … , АО-25) отмечены закрытыми

точками. Видно, что образец АО-25 является очевидным выбросом (или, точнее, внешним образцом в ПИО теории классификации объектов). Предсказанное значение для АО-25 мало отличается от измеренного значения, однако этот объект находится очень далеко от центра ПЛС модели. Поэтому неопределенность в прогнозе АО-25 высока, и он имеет наибольший ПИО интервал (см. Рис. 9.6), являясь при этом единственным проверочным образцом, для которого погрешность предсказания больше, чем погрешность моделирования. Из Рис. 9.7 мы также видим, что все образцы, для которых доверительные интервалы (CI с помощью НЛР) меньше, чем прогнозные интервалы (PI с помощью ПИО), расположены в левой нижней части графика, ниже прямой линии. Это прямая разделяет образцы на две группы: для которых CI > PI, и для которых CI < PI.

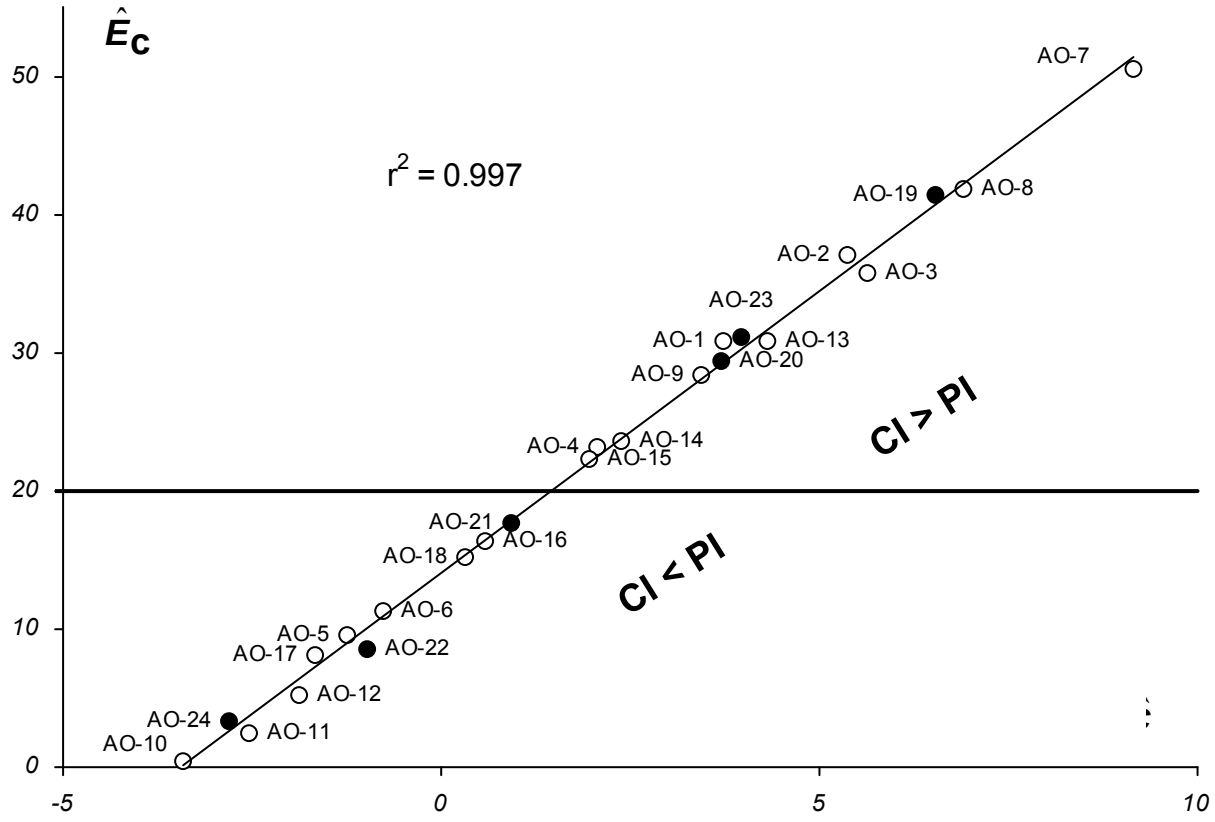


Рис. 9.8. Корреляция между оценками предэкспоненты *c* и соответствующей энергией активации E_c. Открытые точки (○) представляют калибровочные объекты, а закрашенные (●) – проверочные. Тонкая линия показывает корреляционный тренд. Толстая линия разделяет образы, для которых доверительные интервалы (CIs) меньше (больше) чем прогнозные интервалы (PIs).

Такой результат представляется нам удивительным. Мы ожидали обнаружить некоторый порядок, структуру, в объектах, которая была бы связана с шириной интервалов, т.е. с точностью предсказания. Однако мы обнаруживаем другую структуру, которая разделяет образцы по тому, какой метод – содержательный или формальный,

является для них лучшим. В таком случае было бы интересно изучить другой график, связанный уже с содержательным, НЛР методом, и посмотреть, есть ли там похожие структуры. На Рис. 9.8 представлен такой график, на котором показана корреляционная зависимость между оценками параметров E_c и c (см уравнение (9.9)). Важно отметить, что оценки параметров в обоих парах (a, E_a) и (c, E_c) сильно коррелированны. В уравнении Аррениуса часто случается [264], что корреляция между оценками предэкспоненциального множителя и энергии активации близка к 1. В нашем случае: $\mathrm{cor}(\hat{a}, \hat{E}_a) = 0.995, \mathrm{cor}(\hat{c}, \hat{E}_c) = 0.997$. Рис. 9.8 дает возможность посмотреть на проблему с другой стороны. На нем ясно видно, что все образцы, для которых содержательный (НЛР) подход лучше (CI < PI) расположены в нижней части графика, под толстой линией. Переход соответствует критической концентрации АО равной 0.016 при T=140^{o}C.

График счетов на Рис. 9.7 представляет только одно подмножества объектов, а именно образцы с начальной концентрацией АО равной 0.05. Другие графики: для A_0=0.07, и A_0=0.10 имеют сходную структуру и здесь не приводятся. Однако оценки параметров (c, E_c) были получены с помощью содержательного моделирования, в котором все образцы с различными начальными концентрациями АО обрабатывались совместно методом НЛР, поэтому Рис. 9.8 представляет весь набор данных. Это обстоятельство позволяет сделать следующий вывод, который кажется важным для выбора метода калибровки. Такое свойство как «доверительный интервал (CI), построенный методом нелинейной регрессии, больше (меньше) чем прогнозный интервал (PI), построенный методом простого интервального оценивания» не зависит от величины начальной концентрации АО в образце, а зависит от величины критической концентрации АО, которая, в свою очередь, является свойством самого АО. Таким образом, для выбранного АО метод НЛР всегда будет лучше (хуже) метода ПЛС/ПИО вне зависимости от начальной концентрации этого АО. Для АО с низким значением критической концентрации A_c точнее будет содержательный подход (НЛР), а для АО с высоким значением A_c формальный (ПЛС/ПИО).

Другой важный аспект, который должен быть рассмотрен при сравнении методов моделирования, состоит в ограничениях на область применимости каждого метода. Содержательное моделирование (НЛР) имеет здесь очевидное преимущество, т.к. оно может использоваться для предсказания периода индукции для различных концентраций АО и при разных температурах экспозиции. На Рис. 9.9 показан прогноз периода

индукции АО-18 при начальной концентрации 0.04 для температур экспозиции в интервале 80°C <*T* < 200°C. Разумеется, эти условия не были исследованы в эксперименте, а представленные результаты были получены экстраполяцией уравнения (9.12) на эти условия. Конечно, формальный метод моделирования не может дать подобного прогноза. Между прочим, температура *T*=200°C имеет большое значение, т.к. при ней происходит переработка полипропилена. С другой стороны применение содержательной нелинейной модели к условиям, лежащим далеко от области эксперимента, довольно рискованно. В частности, эта модель не может быть использована для экстраполяции на комнатную температуру. Поэтому нельзя прогнозировать срок службы полипропилена, используя метод ДСК. Иными словами, мы не можем указать точные границы применимости содержательного подхода.

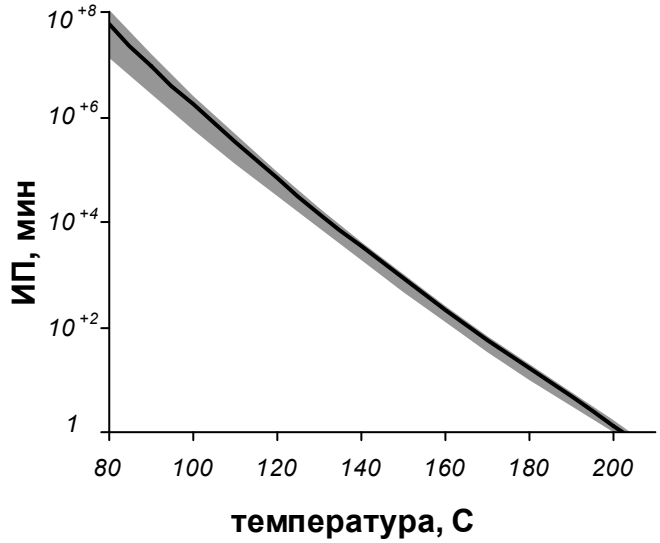


Рис. 9.9 Прогноз периода индукции (ИП) в зависимости от температуры экспозиции для АО-18 с начальной концентрацией 0.04. Линия представляет среднее значение, а серый коридор показывает 0.95 доверительный интервал.

Ситуация с формальным ПЛС/ПИО методом совершенно другая. Здесь можно точно указать соответствующие границы применимости. На Рис. 9.10 показан график классификации ПИО-статусов проверочных объектов, который используется для установления этих ограничений (подробное объяснение теории ПИО классификации дано в разделе 6). Каждый объект из проверочного набора (АО-18, ..., АО-25) изображается на графике статусов в координатах. *ПИО-остаток–ПИО-размах.* Положение проверочного (или нового) объекта на этом графике определяет возможность построения для него прогноза. Все образцы, расположенные внутри треугольника (АО-

20, 21, ..., 24) – внутренние. Они полностью согласуются с моделью и прогноз на них наиболее надежен. Обратный случай означает, что такие объекты лежат вне модели, они называются внешними (образцы АО-19 и АО-25). Внешние объекты не противоречат модели, но прогноз на них менее надежен. Тому могут быть две причины: большой размах (АО-25) и смещение (АО-19). Таким образом, используя технику ПИО-статусов, можно легко классифицировать новый объект и тем самым ограничить область применимости формального ПЛС/ПИО метода.

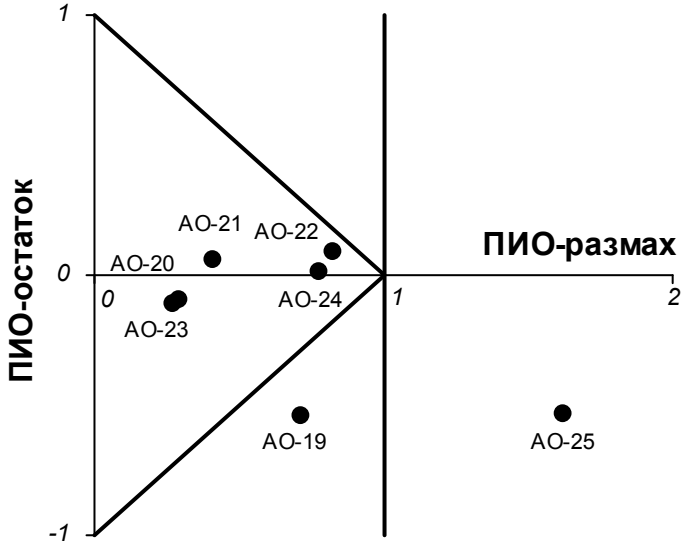


Рис. 9.10. График классификации ПИО-статусов объектов из проверочного набора с A_0=0.07.

Обсуждая условия применимости содержательного и формального методов, необходимо отметить принципиально разную природу областей применимости. В содержательном подходе таковой является область в пространстве факторов (T и A_0), в которую модель может быть экстраполирована. При этом мы имеем дело с одним и тем же АО, который был предварительно исследован методом ДСК. При формальном подходе таковой является область новых антиоксидантов, к которым может быть применена построенная ПЛС/ПИО модель. При этом условия эксперимента, т.е. начальная концентрация АО и скорость нагрева в ДСК, должны быть теми же самыми, что и в исходном калибровочном наборе. С этим соображением связано и еще одно обстоятельство. Допустим, необходимо улучшить модель и достичь большей точности в прогнозе. Для этого надо применять различные планы уточняющего эксперимента. В формальной модели надо добавить новые образцы АО, отличающиеся от исходного калибровочного набора. В содержательной модели, для каждого АО строится своя

собственная модель, поэтому для ее улучшения надо провести новые эксперименты с другими начальными концентрациями того же АО. Это повысит точность прогноза именно этого АО и не будет иметь никакого влияния на качество предсказания для других АО.

9.6. Результаты главы 9

В этой главе продемонстрированы два подхода к решению одной и той же практической задачи – проверки активности АО. Практическим результатом является возможность замены длительного и дорогостоящего процесса термического старения быстрым ДСК методом, с последующей обработкой полученных данных содержательным (НЛР) или формальным (ПЛС/ПИО) методом. Оба подхода к моделированию дают удовлетворительную точность прогноза периода индукции, которая не хуже, чем ошибка в традиционном методе.

Каждый метод моделирования имеет свои преимущества и недостатки. Содержательный подход позволяет получать результаты прогноза, экстраполированные на условия (температура и концентрация), лежащие за областью эксперимента. Однако сложно ограничить область такой экстраполяции, поэтому иногда получаемый результат не надежен. С другой стороны, формальный метод имеет строгую область применимости, очерченную с помощью техники ПИО статуса. В тоже время, ПЛС/ПИО подход не может быть применен для предсказания периода индукции в условиях, отличающихся от условий калибровочного эксперимента. Оба метода имеют близкую точность предсказания и обнаруживают схожее поведение по отношению к условиям прогноза, т.е., чем больше начальная концентрация АО, тем хуже точность предсказания. Однако содержательный подход дает лучшие результаты для АО с низким уровнем критической концентрации, т.е. для АО с малым периодом индукции, тогда как формальный метод лучше в противном случае. Это является фундаментальным, внутренним свойством АО, которое сохраняется при изменении начальной концентрации АО в образце.

Таким образом, в случае, когда целью исследований является предсказание поведения некоторой данной полимерной системы, содержательный подход предпочтителен. В случае, когда исследователь желает сравнить активность различных АО, формальная модель лучше отвечает такой постановке.

Формальный подход не требует глубоких знаний о природе исследуемого явления. Математические модели, используемые в этом методе, линейны относительно неизвестных параметров и поэтому просты для применения и интерпретации. Однако для построения таких моделей часто требуется *угадать* подходящее *преобразование* факторов или отклика, приводящее к линеаризации системы. В рассмотренном примере прогноза активности АО мы удачно извлекли квадратный корень из всех экспериментальных значений периода индукции и тем самым достигли нужного нам результата. Поэтому применение формального подхода требует от исследователя некоторого опыта решения подобных задач. В тоже время, будучи правильно примененным, этот подход дает надежные результаты в самых различных областях. Он хорошо работает при решении задач классификации или *интерполяции*, т.е. прогнозирования на (примерно) те же условиях, в которых проходил эксперимент. Для задач экстраполяции формальный подход применять нельзя.

Содержательный подход устанавливает связь между факторами и откликом эксперимента, учитывая современные представления о физических и химических процессах, проходящих в системе. При этом получаются довольно сложные математические модели, так что для их обработки требуется применение специальных вычислительных методов. Однако если такие модели адекватно описывают исследуемый процесс, то они позволяют делать *экстраполяцию* – прогноз за пределы области факторов, исследованной в эксперименте.

10. Применение метода ПИО к задачам классификации на примере распознавания фальшивых лекарств с помощью ИК-спектроскопии в ближней области

Многомерный подход эффективно используется в задачах качественного анализа. Как уже говорилось в разделе 2.2, это широкий класс задач, в которых требуется установить принадлежность объекта к некоторому классу либо разделить объекты на специфические группы и кластеры.

Методы классификации без обучения, такие как МГК (см. разделы 2.1- 2.2) хорошо справляются с мультиколлинеарными данными, используя проекционный подход и отделяя содержательную информацию от шума. При построении проекционной модели направления главных компонент выбираются так, чтобы описать наибольшие изменения в исследуемых данных, при этом не используется априорная информация о принадлежности объекта тому или иному классу. Однако вовсе не обязательно направления главных компонент совпадают с направлениями, по которым происходит наилучшее разделение между классами. Один из способов введения информации о принадлежности классу в явном виде является использования подхода основанного на методе ПЛС. Этот метод был предложен С. Волдом (S.Wold) и получил название ПЛС дискриминация [80]. Добавление к ПЛС дискриминации информации, получаемой с помощью метода ПИО, позволяет повысить надежность распознавания.

В этой главе представлена методика совместное использование ПИО и ПЛС дискриминации. Такой комбинированный подход применяется к решению задач распознавания фальшивых лекарственных средств с помощью ИК спектроскопии в ближней области (БИК-спектроскопии). . Основные результаты этой главы опубликованы в [163].

10.1. Распознавание фальсифицированных лекарств с помощью инфракрасной спектроскопии в ближней области

Проблема фальсифицированных лекарственных средств во всем мире становится все более серьезной. Сведения о поддельных лекарствах стали достоянием ВОЗ в 1982 году и касались в основном случаев обнаружения фальсификатов в развивающихся странах. В 1992 году на совместной конференции ВОЗ и Международной федерации ассоциаций фармацевтических производителей (International Federation of Pharmaceutical

Manufacturers Associations) было дано определение термина «фальсифицированное лекарственное средство» [265]: фальсифицированное лекарственное средство – это лекарственное средство, предумышленно маркированное таким образом, чтобы ввести в заблуждение в отношении состава и/или изготовителя».

В 2003 г. на конференции FIP в Сиднее дается немного обновленное определение [266]: «фальсификация в отношении лекарственных средств означает преднамеренное и неправильное обозначение в отношении подлинности, состава и/или источника готового лекарственного средства или ингредиента для приготовления лекарственного средства». На конференции было также указано, что фальсификация может касаться как оригинальных, так и воспроизведенных лекарств, а также традиционных лекарственных средств.

На сегодняшний день выделяют следующие разновидности фальсификатов [266].

1. Не содержащие действующие вещества, в том числе и указанные на этикетке – «пустышки» (около 43%).

2. Содержащие действующие вещества, не указанные на этикетке (около 7%). Обычно в таких «препаратах» дорогое действующее вещество заменяется менее активным и более дешевым аналогом. Возможна также продажа «высокоактивных» препаратов, например растительного происхождения, в которые, на самом деле, включены действительно активные лекарственные вещества (например, стероиды или эфедрин и его производные).

3. Содержащие действующие вещества, указанные на этикетке, но изготовленные другим производителем (до 50%). Они включают в себя до 21% подделок с низким содержанием указанного на этикетке действующего вещества, до 5% имеющих ненадлежащую упаковку и до 24% имеющих низкое качество изготовления самого препарата.

По данным ВОЗ [266], на развивающиеся страны приходится около 70% оборота всех поддельных лекарств, на развитые – около 30%. По терапевтическим группам: противомикробные средства 28%, гормональные средства 22% (включая стероиды 10%), противогистаминные средства 17%, сосудорасширяющие средства 7%, средства, применяемые при сексуальных расстройствах 5%, противосудорожные средства 2%, другие лекарственные средства 19%.

По некоторым данным, более 12% лекарств, обращающихся в России, являются подделками. Во многих случаях эти фальшивки представляют большую опасность. Так,

известен случай [265], когда использование поддельного препарата, содержавшего диэтиленгликоль, повлекло смерть более 500 человек, в большинстве детей.

В последнее время все чаще встречаются поддельные препараты, имеющие высокое «качество» изготовления, что затрудняет их обнаружение. Тесты на растворимость или простые цветные реакции позволяют выявить лишь грубые подделки. Сложные химические методы, такие как высокоэффективная газовая хроматография, более эффективны, но они требуют большого времени [267, 268]. Кроме того, когда подделка содержит нужное количество активного ингредиента, но отличается присутствием других, часто очень опасных добавок, необходимы дополнительные исследования.

В качестве эффективного метода для быстрого обнаружения поддельных лекарств предлагается использовать метод БИК спектроскопии в сочетании с последующей математической обработкой результатов измерений. В спектральном диапазоне 1000-2500 nm (4500 - 9000 cm^{-1}) свет может проникать на несколько миллиметров вглубь образца. Это обстоятельство делает возможным использование т.н. диффузного отражения, в котором свет, проходящий через образец, рассеивается внутри него, отражается и, затем, детектируется. Таким образом, можно получать информацию не только о химическом составе, но и о физической структуре образца. К преимуществам такого подхода следует отнести: простоту проведения типового анализа и отсутствие пробоподготовки. Однако результаты экспериментов содержат требуемую информацию в скрытом виде. Для ее извлечения необходимо применять специальные математические методы. Следовательно, для проведения типовых анализов необходима предварительная работа по построению математической модели для каждого конкретного типа лекарства. Предлагается два способа построения таких моделей, первый это метод SIMCA, описанный в разделе 2.2 а так же более информативный метод ПЛС дискриминации в сочетании с ПИО моделированием.

10.2. Комбинированный метод: ПЛС дискриминация и метод ПИО

Предположим, что надо разделить объекты на Q различных классов. Для построения модели используется калибровочный набор, который составляется из объектов всех Q классов. Матрица предикторов **X** – это матрица наблюдений, состоящая из I объектов и J переменных. В качестве матрицы откликов **Y** вводится матрица искусственных переменных или матрица принадлежности классу. Количество столбцов в **Y** равно количеству классов Q. Для всех объектов из класса q (q=1,2…Q), y_q равно 1, а

для объектов, не принадлежащих *q*-ому классу, значения откликов равны -1. Пример такой модели для трех классов в схематическом виде представлен на Рис. 10.1. Далее, используя набор данных (**X**,**Y**), строится ПЛС2 модель. Для каждого нового объекта, вычисляется прогнозное значение $\hat{\mathbf{y}}^t$, по которому и определяют принадлежность объекта тому или иному классу. Такой подход относится к методам классификации с обучением, или методам дискриминации.

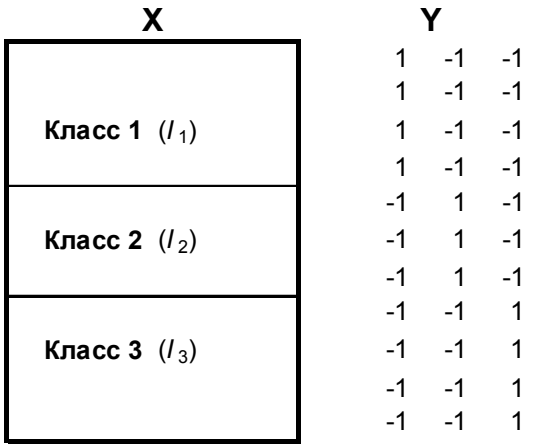


Рис. 10.1 Схема метода ПЛС дискриминации.

Очевидно, что значения отклика вычисляются с некоторой погрешностью. Для оценки погрешности предлагается в дополнении к методу ПЛС дискриминации использовать ПИО метод.

Для объектов из калибровочного набора необходимо оценить погрешность моделирования, для того чтобы понять, не перекрываются ли классы представленные в калибровочном наборе. В ПИО методе для этих целей используется интервал калибровки (6.2) который представляет оценку максимальной погрешности.

Для распознавания новых объектов необходимо оценить близость значения предсказанного значения $\hat{\mathbf{y}}^t$ к тому, или иному классу. В ПИО методе для этого служит интервал предсказания (5.17), характеризующий неопределенность в прогнозе индивидуально для каждого объекта.

10.3. Эксперимент 1. Исследование таблеток. БИК спектры диффузного рассеяния

Исследовались образцы пищеварительного ферментного средства Мезим (панкреатин). Таблетки плохо растворимы в воде, не растворимы в спирте, содержат

главным образом трипсин и амилазу, стандартизуются биологическим путем. В нашем распоряжении имелся достаточно представительный набор образцов, состоящий из одиннадцати различных серий подлинных таблеток (обозначенных как G1 – G11) и четырех серий фальсифицированных таблеток (обозначенных F1 – F4). Каждая серия представлена 5 таблетками из одной упаковки. Всего 75 образцов.

Спектры диффузного рассеяния $R(\lambda)$ были сняты с помощью прибора Spectrum One NTS в диапазоне 4000–1000 см$^{-1}$ с разрешением 8 см$^{-1}$. Измерения проводились без специальной подготовки образцов. Для анализа данных был выбран наиболее информативный участок 4000–7500 см$^{-1}$. (1750 длин волн). Исходные данные преобразовывались как $-\log R$. Далее данные центрировались и предварительно преобразовывались с помощью процедурой MSC [127], рассмотренной в разделе 1.3. Они показаны на Рис. 10.2

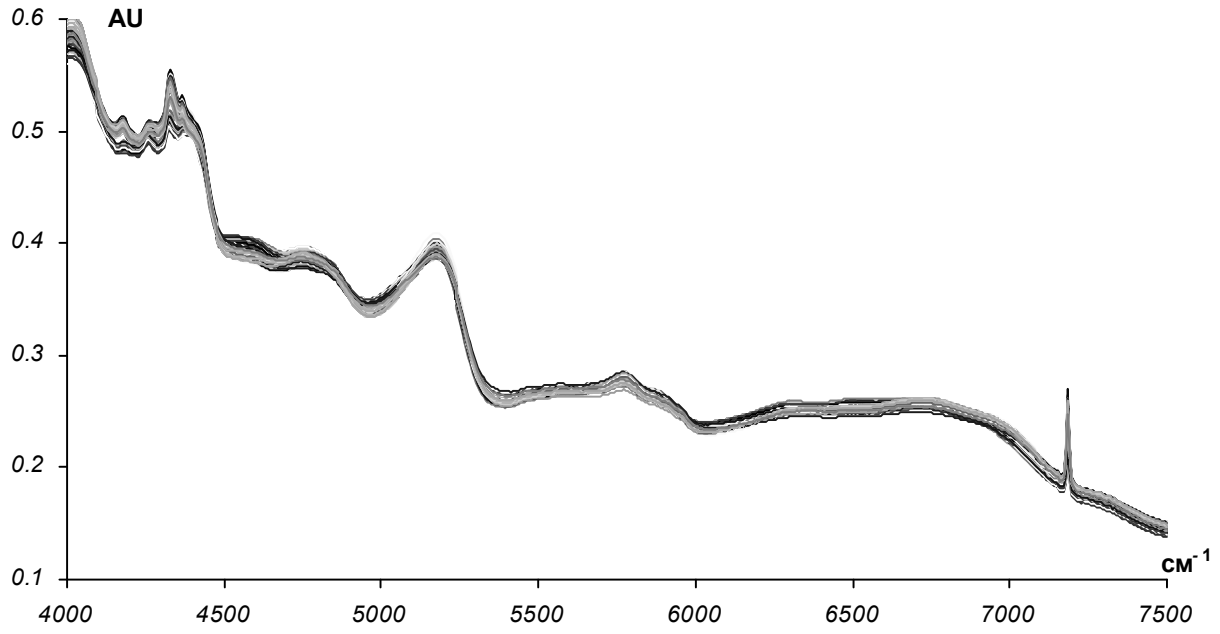


Рис. 10.2 Определение фальсифицированных лекарств (таблетки). Экспериментальные данные после MSC.

Таким образом, исходный набор данных, т.е. матрица **X** имеет размерность (75×1750). Целью исследования является построение математической модели, которая может надежно разделить настоящие и поддельные образцы. При этом необходимо проверить модель как на ошибки первого рода, когда поддельные образцы классифицируются как подлинные; так и ошибки второго рода, когда подлинные таблетки будут отнесены к фальсификатам. Исследование осложняется тем, что

исходным сырьем служит природный материал, ферментный препарат из поджелудочной железы убойного скота, который сам по себе имеет естественную вариабельность.

10.4. Математическая обработка результатов эксперимента

Сначала проводилась классификация без обучения, так как это описано в разделе 2.2. Для этого калибровочного набор формируется только из подлинных образцов 3-х различных серий. Эти 15 образцов используются для построения модели по МГК. Все остальные объекты собираются в проверочный набор (см. Рис. 10.3). Объекты из проверочного набора проецируются на подпространство главных компонент, образованное моделью.

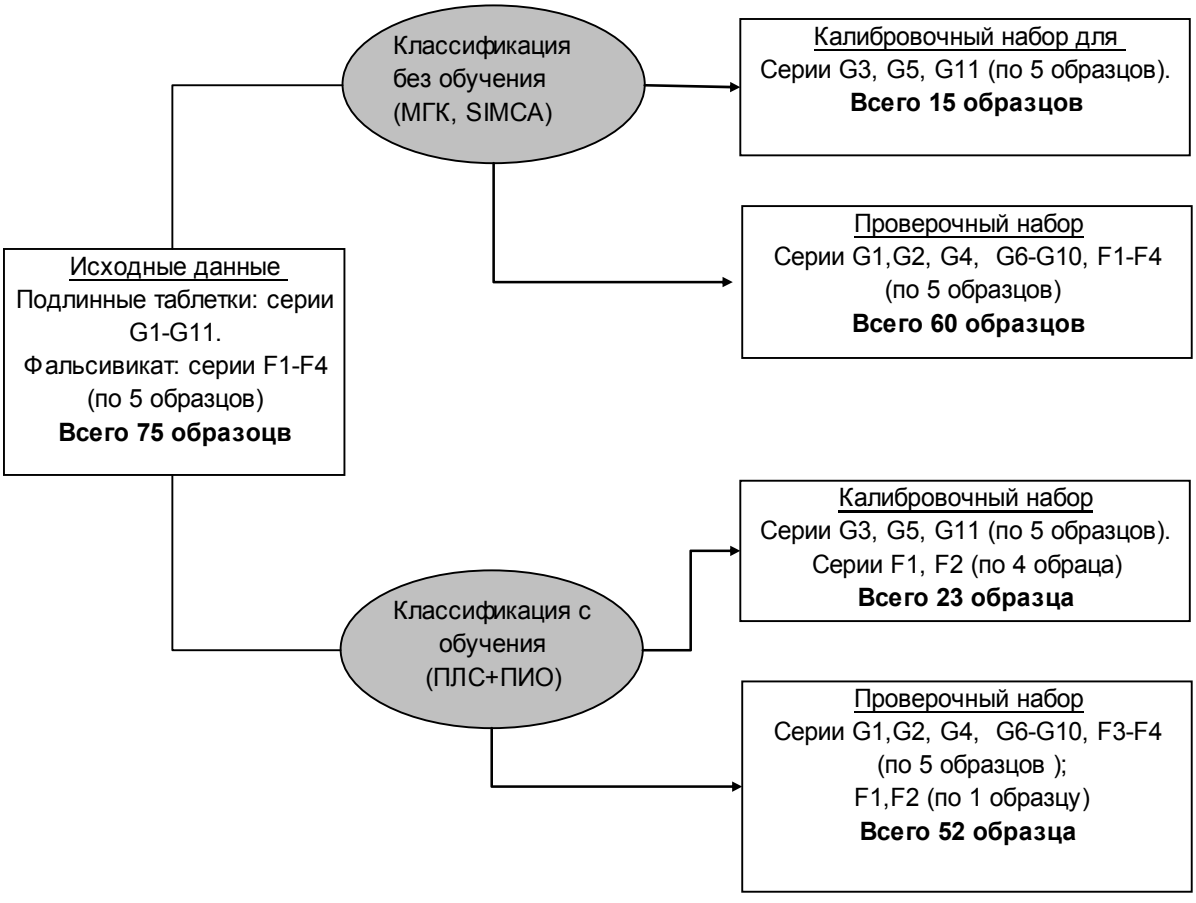


Рис. 10.3 Схема разбиения на калибровочный и проверочныq наборы

На Рис. 10.4 представлены объекты из калибровочного набора (черные кружки) на графиках счетов ($\mathbf{t}_1$-$\mathbf{t}_2$) и ($\mathbf{t}_1$-$\mathbf{t}_3$). Кроме того, на каждом из графиков представлена 95%-ая доверительная область Хотеллинга, которая на плоскости имеет вид эллипса. Главные оси эллипса вычисляются по формуле

$$[\hat{Var}(\mathbf{t}_k) F_{2,I-2,\alpha} 2(I^2 - K)/(I(I-2)]^{1/2}, \qquad (10.1)$$

где $\hat{Var}(\mathbf{t}_k)$ - это оценка дисперсии $\mathbf{t}_k$, I – число образцов в калибровочном наборе, K число главных компонент в модели , F - квантиль F-распределения с 2, I-2 степенями свободы и уровнем значимости α (в нашем случае α=0.05), k- номер главной компоненты, оси по которой строится эллипс.

После проецирования проверочных объектов на подпространство ГК, видно, что практически все образцы подлинных лекарств (незакрашенные круги на Рис. 10.4) расположены внутри области модели. Однако в эту область попадают и часть поддельных лекарств (закрашенные квадраты Рис. 10.4). Усложнение модели, т.е. увеличение количества главных компонент, не дает улучшения результата распознавания. Таким образом, можно заключить, что применение только МГК не позволяет надежно разделить исследуемые образцы.

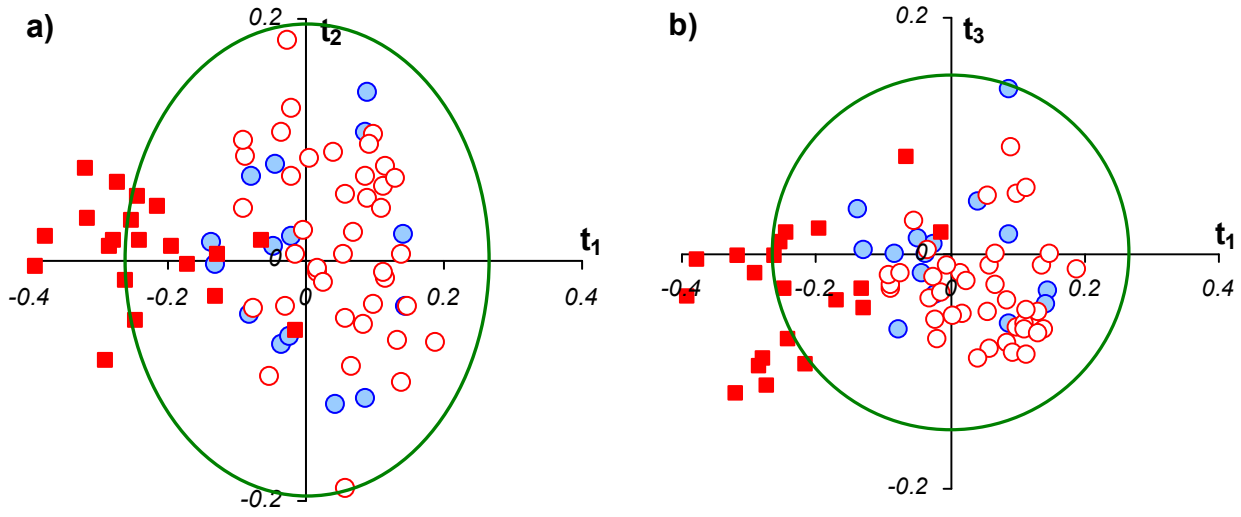


График счетов ($\mathbf{t}_1$-$\mathbf{t}_2$) График счетов ($\mathbf{t}_1$-$\mathbf{t}_3$)

Рис. 10.4 Определение фальсифицированных лекарств (таблетки). МГК. ●-калибровочный набор, ○- подлинные образцы, ■-фальсифицированные образцы

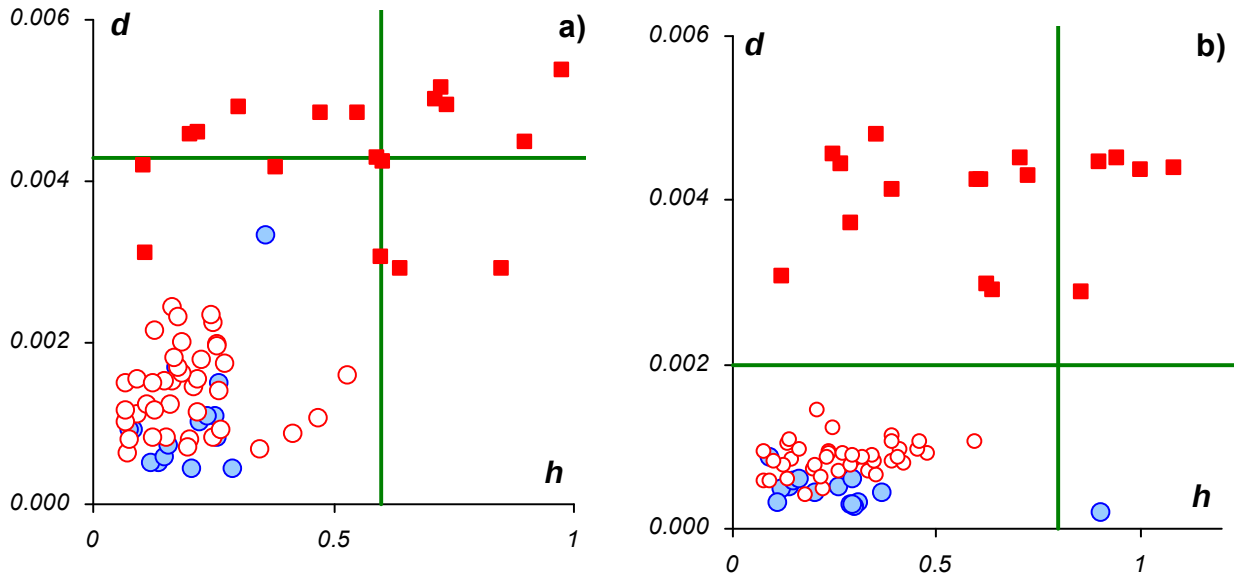
Далее, воспользуемся построенной МГК моделью и применим метод SIMCA, как это описано в разделе 2.2. В данном случае нам необходимо построить модель только для одного класса – класса подлинных объектов. Использовалось два варианта моделей с 2 и 3 ГК (Рис. 10.5). Объекты из калибровочного набора, так же как и на предыдущем графике изображены закрашенными кружками на плоскости (d –h). Где d – это среднеквадратичный остаток моделирования, т.е. расстояние от объекта до

подпространства главных компонент, а h это МГК-размах, т.е. расстояние от объекта до центра модели, внутри подпространства ГК. Для определения области модели строятся критические уровни. По оси d, критический уровень вычисляется с помощью среднеквадратичного остатка внутри класса, величины d_0 (2.5). По оси h критический уровень вычисляется как удвоенное значение среднего размаха для объектов внутри модели.

$$h_{crit} = 2\frac{K+1}{I}, \qquad (10.2)$$

где I – это количество образцов в калибровочном наборе, а K – число главных компонент в модели.

Далее, для всех объектов из проверочного набора, вычисляются значения d (2.4) и h (2.6). Как видно из Рис. 10.5, модель, построенная с использование двух ГК, правильно классифицирует образцы подлинных лекарств (незакрашенные кружки), но при этом опять относит часть фальсифицированных лекарств к классу подлинных. Для надежного разделения необходимо использовать модель, построенную на трех ГК (Рис. 10.5 b).



МГК модель с 2 главными компонентами МГК модель с 3 главными компонентами

Рис. 10.5 Определение фальсифицированных лекарств (таблетки). Модель SIMCA. ●- калибровочный набор, ○- подлинные образцы, ■-фальсифицированные образцы

Сложность построения таких моделей заключается в том, что при моделировании класса подлинных образов необходимо включать в модель только общие особенности класса и, по возможности, отбрасывать индивидуальные особенности каждого объекта.

Это достигается тем, что модель не должна быть "слишком подробной", т.е. содержать много ГК. В рассматриваемом случае, модель, построенная на трех ГК, уже рассматривает один объект из калибровочного набора как экстремальный. На Рис. 10.4b, этот объект виден как закрашенный кружок лежащий хоть и близко, но за границей доверительного эллипса. На Рис. 10.5b, этот же объект располагается за критической границей по оси *h*.

Для того чтобы основное внимание при моделирование удалялось различию между классами, предлагается использовать метод ПЛС дискриминации, описанный в разделе 10.2.

Калибровочный набор формируется теперь как из подлинных, так и из фальсифицированных образцов. В проверочный набор также вошли объекты из обоих классов (см. Рис. 10.3). При этом, в проверочный набор включены, как образцы поддельных лекарств из серий, участвующих в построении модели (по одной таблетки из каждой такой серии), так и образцы из новых серий поддельных лекарств.

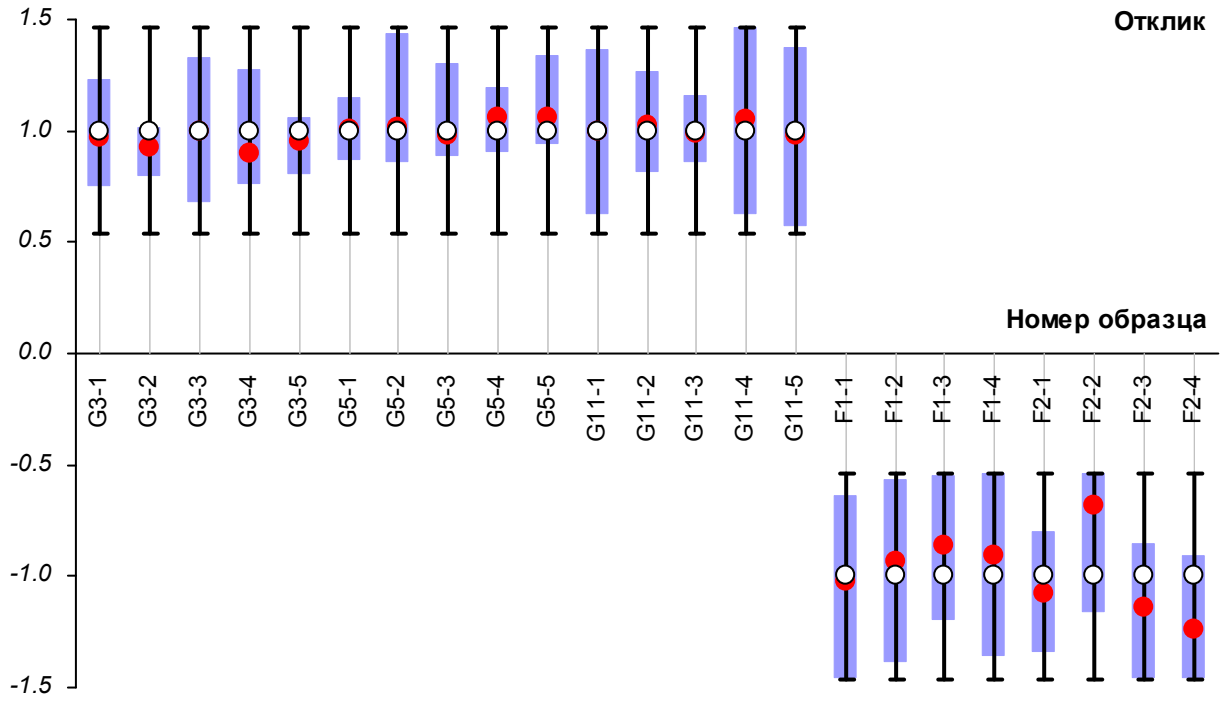


Рис. 10.6 Определение фальсифицированных лекарств (таблетки). ПЛС модель с двумя ГК. Калибровочный набор.

Если необходимо различить всего два класса, то при применении метода ПЛС дискриминации вводится всего один искусственный вектор **y**. При этом, для объектов,

принадлежащих одному классу, в нашем случае – классу подлинных лекарств, значения y_i=1, а для объектов принадлежащих другому классу, в нашем случае фальсифицированным лекарствам, y_i= –1. ПЛС модель, построенная с помощью двух ГК, описывает 79% вариации матрицы **X**, но при этом 99 % вариации **y**. Таким образом, введя искусственную переменную принадлежности классу, при построении модели основное внимание было перенесено на различие между классами, а не на вариацию внутри одного класса. Результаты ПЛС моделирования представлены на Рис. 10.6. Закрашенные кружки – это значения **y**, вычисленные с помощью ПЛС модели, незакрашенные кружки – это опорные значения. Видно, что модель хорошо разделяет классы объектов. При этом вычисленные и опорные значения близки, особенно для класса подлинных образцов (образцы G3-1 – G11-5). Что касается класса фальсифицированных образцов (образцы F1-1 – F2-4), здесь расхождения между вычисленными и присвоенными значениями уже больше.

Для того чтобы оценить неопределенность в прогнозе и сделать дискриминацию более надежной, применим к построенной ПЛС модели метод ПИО. Для ПИО модели вычислены значения b_{min}=0.23 и b_{SIC}=0.46. При этом обнаружено шесть граничных объектов, пять из которых принадлежат классу фальшивых лекарств, и только один классу подлинных. Результаты ПИО моделирования представлены на Рис. 10.6. Черные отрезки – это максимальная погрешность, равная ± $b_{SIC,}$ а прямоугольники – это индивидуальные ПИО интервалы. ПИО метод так же подтверждает надежное разделение на классы среди объектов калибровочного набора, даже с учетом индивидуальных погрешностей.

Далее применим ПЛС и ПИО модели к проверочному набору (Рис. 10.7). Результаты ПЛС прогноза показывают, что и проверочные объекты можно разделить. Все образцы подлинных лекарств (G1-1 – G10-5), представляющие восемь различных серий распознаются правильно. Образцы фальшивых лекарств F1-5 и F2-5 взяты из серий F1 и F2. Эти серии вошли в калибровочный набор (см. Рис. 10.3). Образцы F3-1 – F3-5 принадлежат фальшивой серии F3. Относительно всех этих объектов можно заключить, что они похожи на фальсифицированные образцы калибровочного набора, и их принадлежность фальсифицированному набору очевидна. Что касается образцов F4-1 – F4-5 (фальсифицированная серия F4), то их принадлежность фальсифицированному набору не так очевидна.

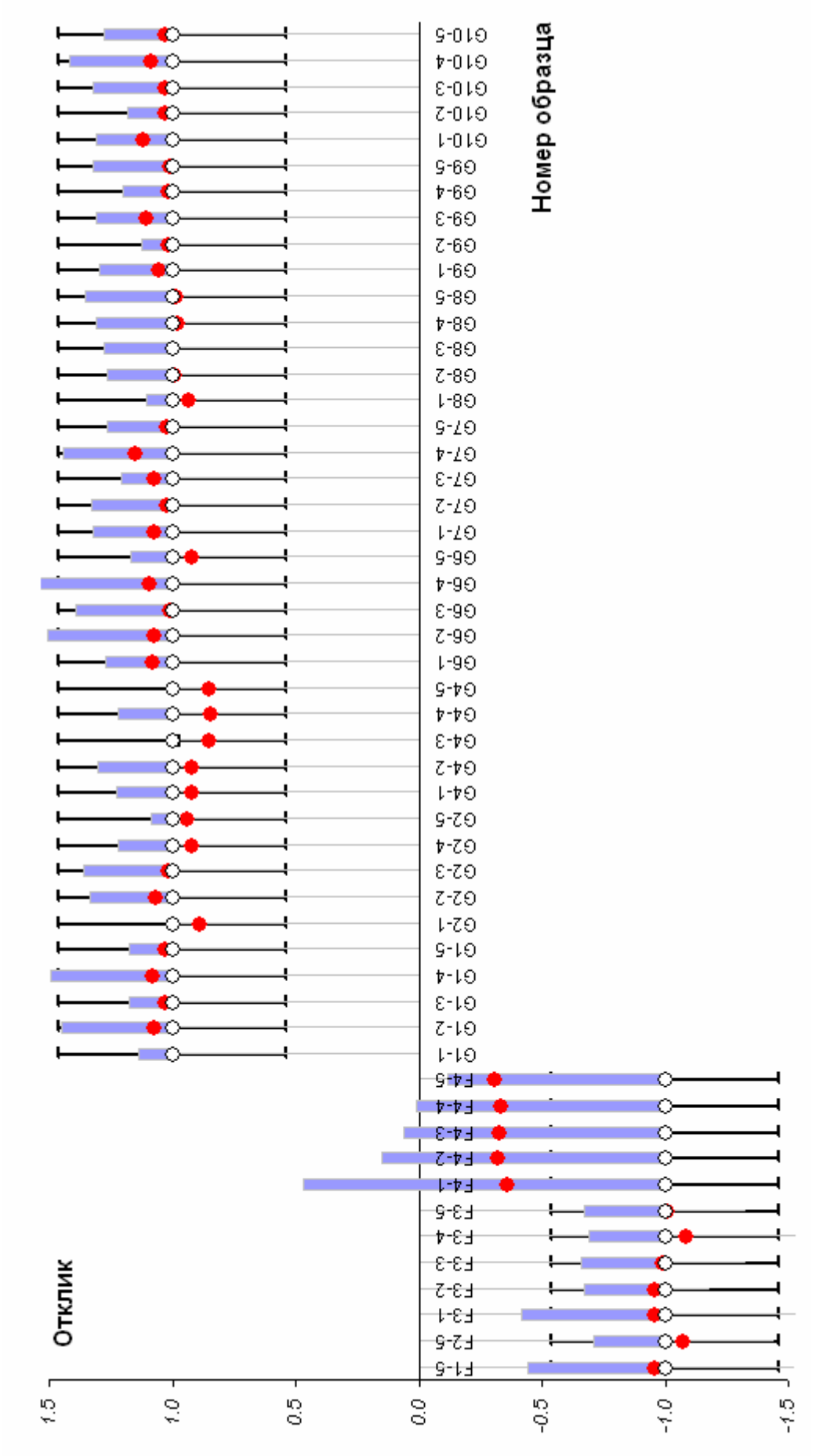


Рис. 10.7 Определение фальсифицированных лекарств (таблетки). ПЛС модель с 2 ГК, ПИО моделирование.
Проверочный набор. а) –Предсказание;

Для более детального анализа воспользуемся результатами ПИО моделирования. Из Рис. 10.7 видно, что хотя прогнозные значения образцов F4-1 – F4-5 лежат далеко от значений, определенных для каждого из классов, но их индивидуальные прогнозные интервалы (прямоугольники) не пересекаются с интервалами калибровки класса подлинных объектов (черные отрезки). Следовательно, образцы F4-1 – F4-5 нельзя отнести к классу подлинных объектов. Они, конечно, расположены ближе к классу фальшивых объектов, но их прогнозные интервалы лишь незначительно пересекаются с интервалами калибровки класса фальшивых объектов. Экстремальное положение этих объектов видно и на ДСО (Рис. 10.8). Таким образом, относительно образцов из серии F4 можно заключить следующее: (1) эти объекты очевидно нельзя отнести к классу подлинных объектов; (2) образцы серии F4 существенно отличаются от серий F1-F3. Можно предположить, что фальсификат F4 был сделан иным производителем, чем серии F1-F3, либо серия F4 была произведена из другого сырья.

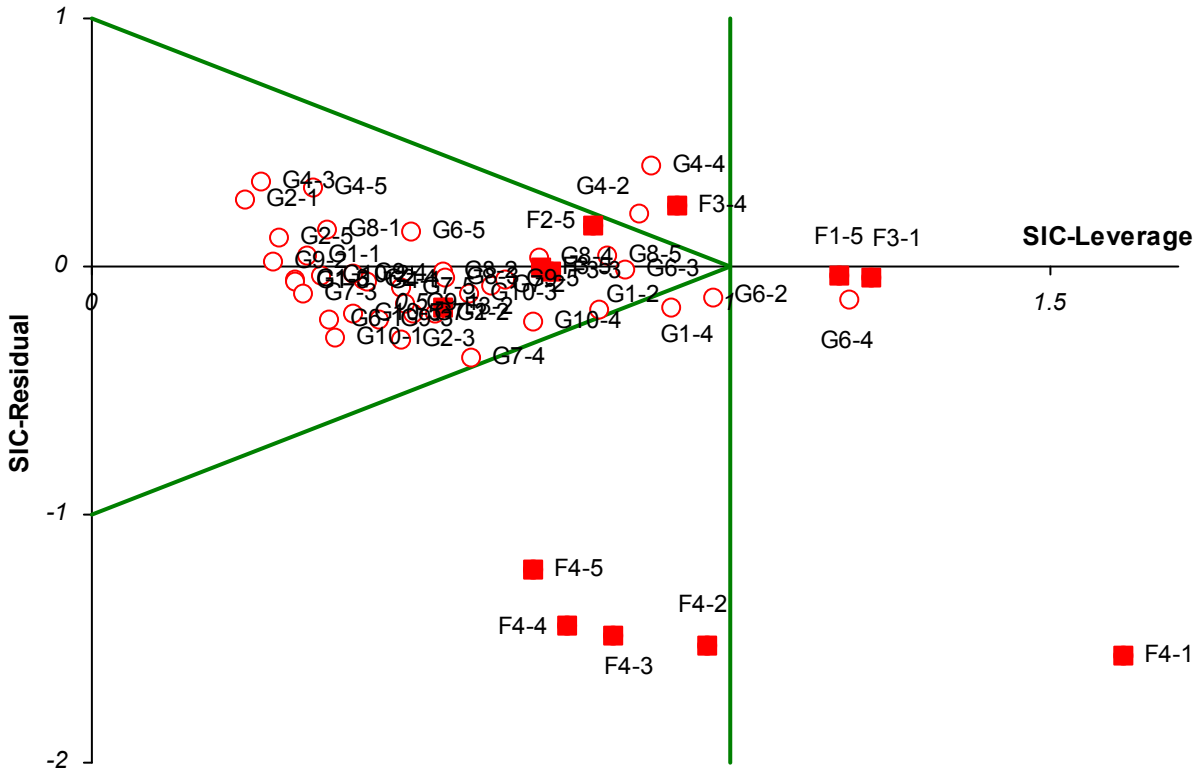


Рис. 10.8 Определение фальсифицированных лекарств (таблетки). ПЛС модель с 2 ГК
Диаграмма статуса объектов : ○- подлинные образцы, ■-фальсифицированные образцы

Продолжая исследовать ДСО (Рис. 10.8), можно заметить, что среди подлинных образцов есть один экстремальный объект (G6-4), который классифицируется как

абсолютно внешний (Опр.6.5), при этом у этого объекта практически нулевое значение ПИО остатка (6.8). Этот результат, скорее всего, следует отнести к погрешностям измерения. Кроме этого еще четыре подлинных образца из проверочного набора (G1-4, G4-2, G4-4 и G6-2) классифицируются как внешние (Опр.6.3), хотя они и расположены очень близко к области модели. По всей видимости, это отражает естественную вариацию внутри разных серий подлинных образцов. То же самое можно сказать и про образцы F1-5 и F3-1, относящиеся к классу фальшивок. Они являются абсолютно внешними по отношению к построенной модели. Но, так как их ПИО остатки близки к нулю, они надежно классифицируются как фальшивые. Как показывает опыт, для фальсифицированных образцов вообще характерен больший разброс между образцами внутри серии, а так же между сериями по сравнению с подлинными образцами. По всей видимости, это можно объяснить низкой технологической дисциплиной производителей фальсификата.

10.5. Эксперимент 2. Исследование ампул - БИК спектры пропускания

Исследовался глюкокортикостеройдный препарат Дексаметазон в растворимой форме, дексамитазон 21-фосфата натриевой соли. Исследуемые образцы – это запаянные ампулы темного стекла с 4% водным раствором активного вещества. Объем ампулы 1 мл. Исследование препаратов для инъекций затруднено тем, что желательно проводить эксперимент, не вскрывая самих ампулы. С другой стороны, так как практически все препараты для инъекций являются водными растворами, а вода характеризуется сильными полосами поглощения в БИК области, то выделить сигнал, получаемый от действующего вещества на фоне воды затруднительно.

В нашем распоряжении имелось две серии подлинных образцов, G1 и G2, и одна серия поддельных, серия F1. Каждая серия состояла из 15 ампул, т.е. всего 45 образцов.

Измерения проводились на приборе Bomem MB160 FT NIR в диапазоне 5500см$^{-1}$ – 10000 см$^{-1}$, с разрешением 4 см$^{-1}$, при этом ампулы не вскрывались. Спектры пропускания $T(\lambda)$ преобразовывались как $-\log T$ (см. Рис. 10.9).

Для исследования использовались две информативные спектральные области: 5500 – 6400 см$^{-1}$ и 7200 – 9000 см$^{-1}$. Всего 702 длин волн. Таким образом, матрица исходных экспериментальных данных $\mathbf{X}$ имеет размерность (45× 702). Данные центрировались и предварительно преобразовывались с помощью процедуры MSC (раздел 1.3).

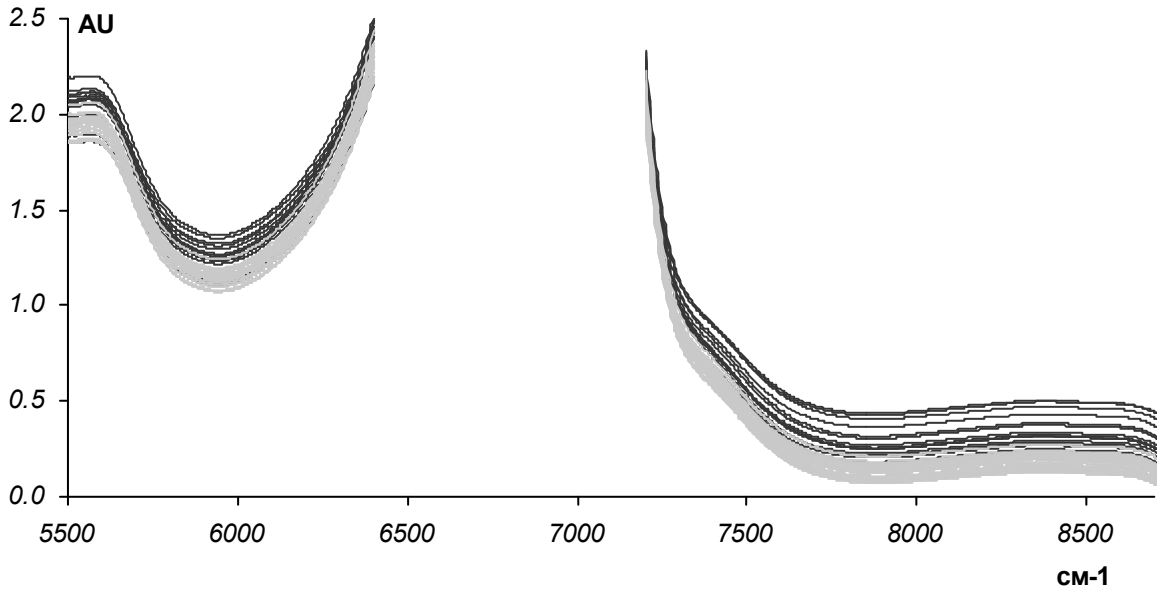


Рис. 10.9 Определение фальсифицированных лекарств (ампулы). Исходные спектры.

10.6. Математическая обработка результатов эксперимента.

Сначала, как и в разделе 10.4, проводится классификация без обучения. Для этого формируется калибровочный набор, состоящий из 12 ампул серии G1. Остальные 33 образца: 3 ампулы серии G1 и по 15 ампул серий G2 и F2, отнесены к проверочному набору.

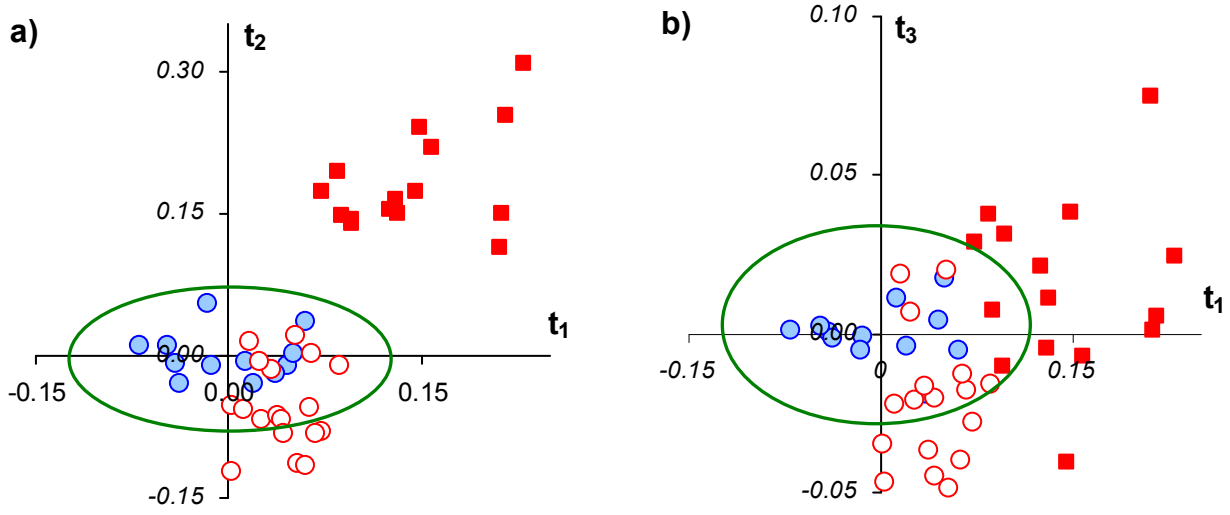
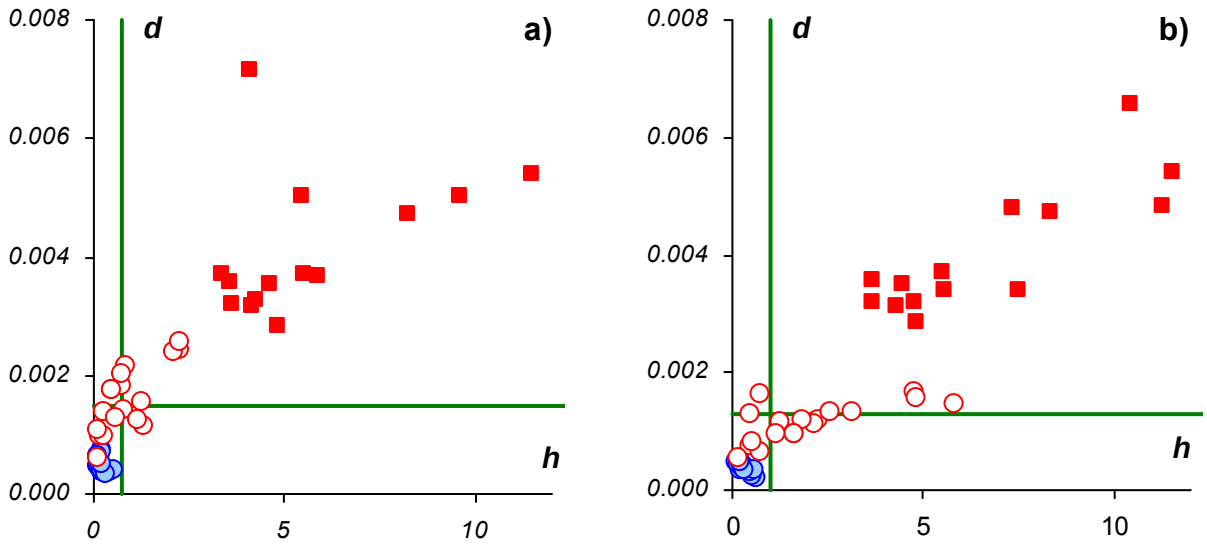


График счетов (t_1-t_2) График счетов (t_1-t_3)

Рис. 10.10 Определение фальсифицированных лекарств (ампулы). МГК . ●- калибровочный набор, ○- подлинные образцы, ■-фальсифицированные образцы

На Рис. 10.10 представлены результаты применения МГК. На графиках счетов видно, что разброс между образцами поддельных лекарств (закрашенные квадраты) больше, чем между настоящими образцами; видна и тенденция к кластеризации групп образцов из G и F серий. Примерно половина образцов подлинных лекарств (не закрашенные кружки на Рис. 10.10) лежат внутри доверительной области (10.1). При этом, в эту же область попадают и часть поддельных лекарств. Увеличение числа главных компонент не дает улучшения результата дискриминации Рис. 10.10b. Таким образом можно заключить, что применение только МГК не позволяет разделить исследуемые образцы.

К построенной МГК модели с двумя и тремя ГК применим метод SIMCA и представим результаты на плоскости d –h: среднеквадратичный остаток относительно размаха (см. Рис. 10.11). Условные обозначения объектов на Рис. 10.11 те же, что и на предыдущем рисунке. Как видно, метод SIMCA позволяет надежно отделить фальшивые образцы от настоящих образцов калибровочного набора, но при этом возникает ошибка второго рода: большая часть подлинных объектов распознается как фальшивые. Таким образом, в данном случае, мы опять не получили надежного способа распознавания.



МГК модель с 2 главными компонентами

МГК модель с 3 главными компонентами

Рис. 10.11 Определение фальсифицированных лекарств (таблетки). Модель SIMCA. ●- калибровочный набор,
○- подлинные образцы, ■-фальсифицированные образцы

Перейдем к комбинированному методу ПЛС дискриминации совместно с методом ПИО. Для этого добавим в калибровочный набор, помимо имеющихся там 12 образцов

серии G1, еще 8 фальшивых образцов серии F1. Таким образом, калибровочный набор теперь состоит из 20 объектов. Для проверочного набора осталось 25 объектов: 7 образцов серии F1, 3 образца серии G1 и 15 образцов серии G2. Введем, как описано в разделе 10.4, вектор искусственных переменных **y**, отражающий принадлежность одному из двух классов. Для объектов, принадлежащих классу подлинных лекарств, значения y_i=1, а для объектов из класса фальсифицированных лекарств, y_i= –1. ПЛС модель, построенная с помощью двух ГК, описывает 94% вариации матрицы **X**, и 99 %вариации **y**. Результаты ПЛС моделирования представлены на Рис. 10.12: закрашенные кружки – это вычисленные значения откликов, а незакрашенные кружки – это опорные значения. Если судить только по точечным оценкам, то объекты калибровочного набора хорошо разделяются на два класса.

Однако, дополнив результаты ПЛС дискриминации ПИО моделированием, можно заметить, что величина интервала калибровки достаточно велика, т.е. построенная модель несет в себе существенную неопределенность. Для ПИО модели были вычислены значения b_{min}=0.48 и b_{SIC}=0.75, при этом область допустимых значений образована шестью граничными объектами.

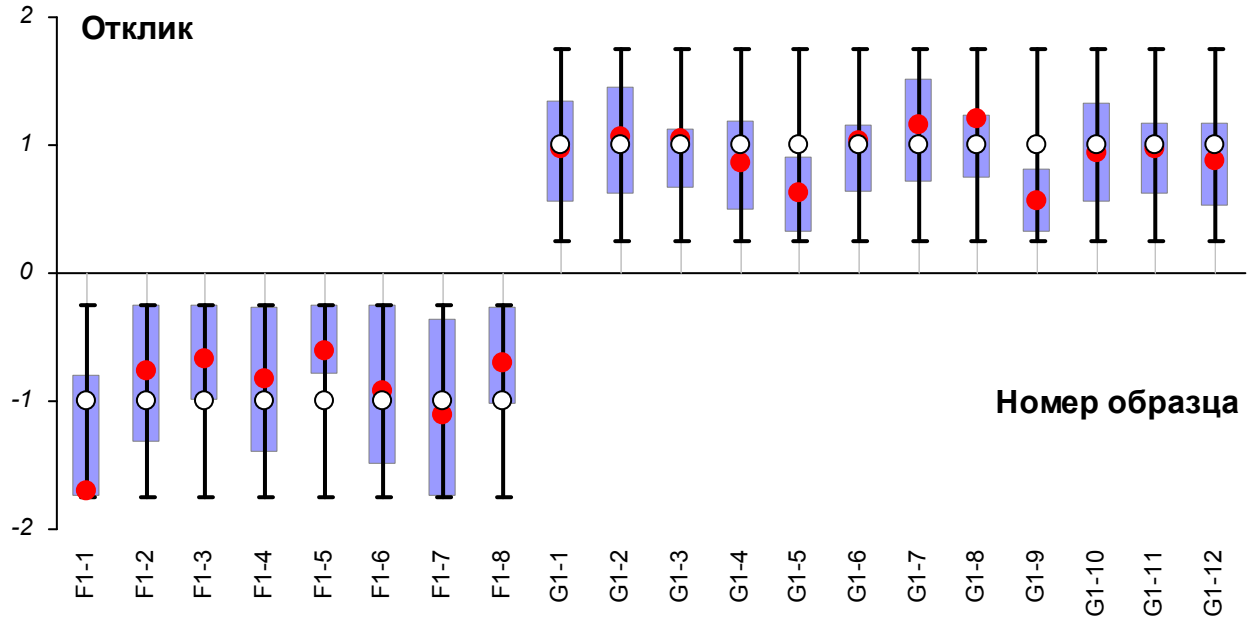


Рис. 10.12 Определение фальсифицированных лекарств (ампулы). ПЛС модель с 2 ГК. Калибровочный набор

ПИО метод показывает, что, несмотря на достаточно хорошие точечные оценки, ширина интервала калибровки (Рис. 10.12, черные отрезки) весьма велика, хотя классы и

не перекрываются, т.е. объекты калибровочного набора надежно разделяются на два класса.

Рассмотрим, как работает построенная модель на проверочном наборе (Рис. 10.13.) Если судить только по точечным оценкам, т.е. по результатам ПЛС дискриминации, то прогнозные значения для подлинных объектов лежат ближе к +1, а для фальшивых к –1. Однако ПИО метод дает более детальный анализ результатов. Образцы G1-13 – G1-15 из проверочного набора надежно предсказываются (величины их индивидуальных интервалов предсказания малы и они полностью лежат внутри интервала калибровки). Это и понятно, так как эти объекты из той же серии G1, которая использовалась в калибровочном наборе. Для большей части образцов серии G2 хотя и можно сказать, что они далеки от фальшивых образцов, входящих в калибровочный набор, но их прогноз не очень надежен, так как 11 из 15 образцов являются внешними (Опр. 6.3), по ПИО классификации статуса объектов. Фальсифицированные образцы распознаются надежно. По большей части, это, видимо, определяется тем, что в калибровочный набор были включены фальшивые образцы. При этом два образца из серии F1 , F1-9 и F1-15 являются внешними согласно ПИО классификации. Это говорит о существенном разбросе внутри серии фальшивых объектов.

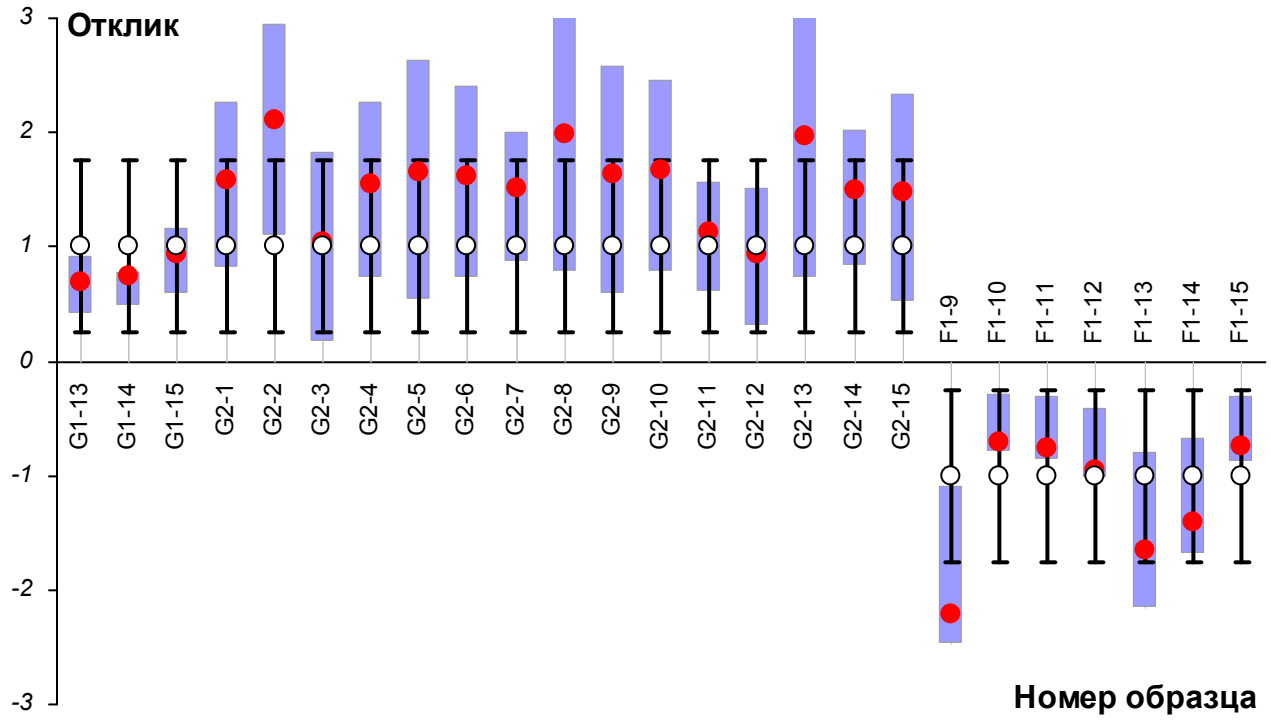


Рис. 10.13 Определение фальсифицированных лекарств (ампулы). ПЛС модель с 2 ГК. Проверочный набор

Таким образом, применение совместно метода ПЛС дискриминации и ПИО метода позволили, в данном случае, надежно распознать фальшивые ампулы, что не удалось сделать МГК и методом SIMCA. Однако следует отметить, что, в данном случае, экспериментальный набор был достаточно скудным. Ненадежность предсказания подлинных образцов говорит о том, что в калибровочный набор следует включить несколько различных серий, для того чтобы учесть естественный разброс между сериями подлинных лекарств. Для надежной проверки распознавания, желательно исследовать и несколько серий фальсифицированных образцов. Однако полученные результаты показывают перспективность предложенного метода.

10.7. Результаты главы 10

Математическая обработка результатов БИК-спектроскопии позволяет разработать быструю и не требующей требующую специальной пробоподготовки процедуру распознавания фальшивых лекарств.

Проведено сравнение различных методов классификации, как без обучения, так и с обучением. Показано, что точечные оценки, получаемые методом ПЛС дискриминации, не достаточны для надежного разделения классов и последующего распознавания новых объектов, так как понятие "близости" к классу должно иметь численное выражение.

Дополнение ПЛС дискриминации методом ПИО дает следующие преимущества.

1. Интервал калибровки (значение величины b_{SIC}), получаемый в результате ПИО моделирования, позволяет очертить точную границу каждого класса.

2. Индивидуальный ПИО интервал предсказания позволяет численно охарактеризовать близость объекта к тому или иному классу.

3. ПИО классификация статуса объектов позволяет охарактеризовать однородность образцов внутри одного класса, а так же выявить объекты или группы образцов с особыми свойствами, отличающими их от объектов предопределенных классов.

К практическим результатам данной главы следует отнести эффективность предложенного подхода к распознаванию фальшивых лекарств, сочетающего методы БИК спектроскопии с последующей математической обработкой экспериментальных данных, комбинированным методом ПЛС дискриминации совместно с методом ПИО. Полученные результаты показывают, что для надежного распознавания фальшивых лекарств необходим большой экспериментальный материал и критический отбор объектов, как в калибровочный, так и в проверочный наборы.

11. Методы анализа процессов

Это направление заслуживает особого внимания, поскольку в нем наиболее ярко проявились тенденции и перспективы развития общего подхода объединяющего физико-химические эксперименты, проводимые в режиме реального времени, с математическими методами многомерного анализа данных. Речь идет о многомерных методах *анализая процессов* (PAT).

В 30-х годах прошлого века американский статистик У. Шухарт (W. Shewhart) предложил использовать статистические методы для контроля хода технологических процессов [269]. 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* (СКП, в англоязычной литературе часто используется сокращение 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 показатели x_1, x_2,..., как правило, связаны между собой, т.е. коррелированны и должны рассматриваться совместно. Ситуация во многом напоминает анализ многоканальных экспериментов (например, спектров) эволюционирующих во времени. Эта аналогия позволила Дж. МакГрегору (J. MacGregor) [270] разработать новый подход к этой задаче, названный *многомерным статистическим контролем процессов* (МСКП) [271]. Он предложил использовать метод главных компонент (МГК) для анализа многомерных исторических данных и строить контрольные пределы в пространстве счетов с помощью расстояния Махаланобиса. Идея оказалась чрезвычайно плодотворной и нашла много последователей. Были разработаны специальные методы для многомерного контроля периодических (например, биохимических) процессов, основанные на трехмодальной калибровке [20], для анализа сложных, многоблочных процессов были созданы иерархические [272], блочные [273], и маршрутные методы [85, 274]. Появились работы, в которых предлагалось не только контролировать, но и оптимизировать ход процессов

[275]. Эти теоретические разработки начали применять на практике, прежде всего в фармацевтической [276] и пищевой [277] промышленности. Были созданы системы контроля качества производства полимеров [278], цветных металлов [279], полупроводников [37].

Таким образом, в хемометрике возникло новое направление [280], стремительно отдаляющееся от 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* [281]. Для этой цели прекрасно подошли традиционные физические методы, прежде всего, спектрометрия: УФ [97], и НПВО-ИК [282]. В том же контексте контроля химических реакций и процессов оказались востребованы хемометрические методы разрешения кривых и оценки кинетических констант по спектроскопическим [283] и хроматографическим [284] данным. Использование МСКП в сочетании с аналитическими методами мониторинга, а также методами контроля качества [285] породило новое направление – методы анализа процессов (МАП). В англоязычной литературе сейчас очень широко используется соответствующее английское сокращение – PAT (*process analytical technology*) [286]. МАП – 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 многих производящих компаний. Многомерный подход играет важную роль в решении этой задачи. Решение Федеральной комиссии по контролю за лекарствами (FDA) [287], 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 пространстве ПЛС-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 [288]. Теперь же, с принятием документа о МАП, хемометрика неизбежно станет легитимным инструментом для всех компаний, желающих следовать руководящим принципам FDA.

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 с анализа входного сырья.

Для осуществления МСКП собирается информация об изучаемом процессе, т.е. накапливаются значения инструментальных показателей для различных реализаций процесса. Так формируется матрица предикторов **X**. Каждая строка этой матрицы содержит наблюдения об одной реализации процесса, далее называемой образцом. Накапливаются так же данные, отражающие конечный результат, т.е. выходные переменные, **Y**. Набор данных (**X,Y**) называется историческим набором. Используя исторический набор, строится *линейная* модель калибровки, которая объясняет, как переменные **y** зависит от X-переменных и проверяет, находится ли процесс внутри допустимых границ. Исследуя построенную модель, можно предложить план действий по корректировке процесса. Однако такая оптимизация будет оптимизацией *post factum.* Наиболее ценной является оптимизация *in situ*, позволяющая осуществлять корректировку по ходу самого процесса, для того чтобы улучшить будущие выходные показатели.

Данный раздел посвящен расширению МСКП метода. Предлагается подход, который определяет действия по оптимизации процесса в режиме in-line. Такой поход предлагается назвать многомерной статистической оптимизацией процессов (МСОП) [275, 289]. Для реализации МСОП предлагается использовать комбинацию двух методов.

Первый, – это ПЛС регрессия [67, 68] с построением различных проекционных моделей (раздел 3.1), а второй – это метод ПИО (глава 5). Для демонстрации того, как работает предложенный подход, используется модельный примером многостадийного технологического процесса.

Основные результаты этой главы опубликованы в [275, 289]

11.1. Описание многостадийного процесса

Рассматривается многостадийный технологический процесс, который представлен двадцатью пятью инструментальными переменными x, характеризующими сам процесс, и одной выходной переменной y, назовем этот показатель "качеством" (например, чистота производимого продукта). Весь процесс разделен на семь стадий, обозначенных римскими цифрами. Каждую стадию можно описать входными, текущими и выходными, «будущими» переменными. Все переменные, которые измерены на предыдущих стадиях, считаются входными переменными (предикторами), и их значения фиксированы. Текущие переменные являются контролируемыми переменными, их можно изменять на текущей стадии. Все переменные, характеризующие последующие стадии на данный момент рассматриваются как отклики, которые, в принципе, можно предсказывать. По мере продвижения вдоль процесса роли переменных меняются.

Первая стадия (I) представлена шестью входными переменными(W1, W2, W3 и S1, S2, S3), которые отражают качество исходных продуктов S и W. На второй стадии (II), компонент W подвергается очистке. Процесс очистки характеризуется переменными WR1 и WR2. Переменные CW1, CW2, и CW3 (Стадия III) характеризуют свойства выходного продукта CW. Следующая стадия (IV) – это смешение компонентов S и CW. Результат смешения M характеризуют переменные M1, M2, и M3. Далее, смесь M подвергается дополнительной очистке (Стадия V). Процесс очистки характеризуют переменные MR1 и MR2, а качество выходного продукта CM описывается переменными CM1, CM2, и CM3 (Стадия VI). На последней стадии (VII) происходит окончательная доработка, которая делается с помощью малых добавок A1, …, A6. Выходная переменная (P=y) характеризует "качество" конечного продукта.

В конце каждой стадии можно проанализировать промежуточные результаты и скорректировать переменные, характеризующие текущую стадию. Анализ и регулировка параметров процесса должны производиться с учетом требований, накладываемых на текущие переменные, и с целью улучшения показателя выходной переменной, т.е.

переменной *y*. В примере предполагается, что все инструментальные переменные являются контролируемыми.

Естественно, что предлагать теоретически оправданные, но практически нереализуемые корректирующие действия бессмысленно. Поэтому, МСКП подход базируется на сборе и анализе накопленных исторических данных. Опыт предыдущих, удачных реализаций используется для анализа текущей ситуации.

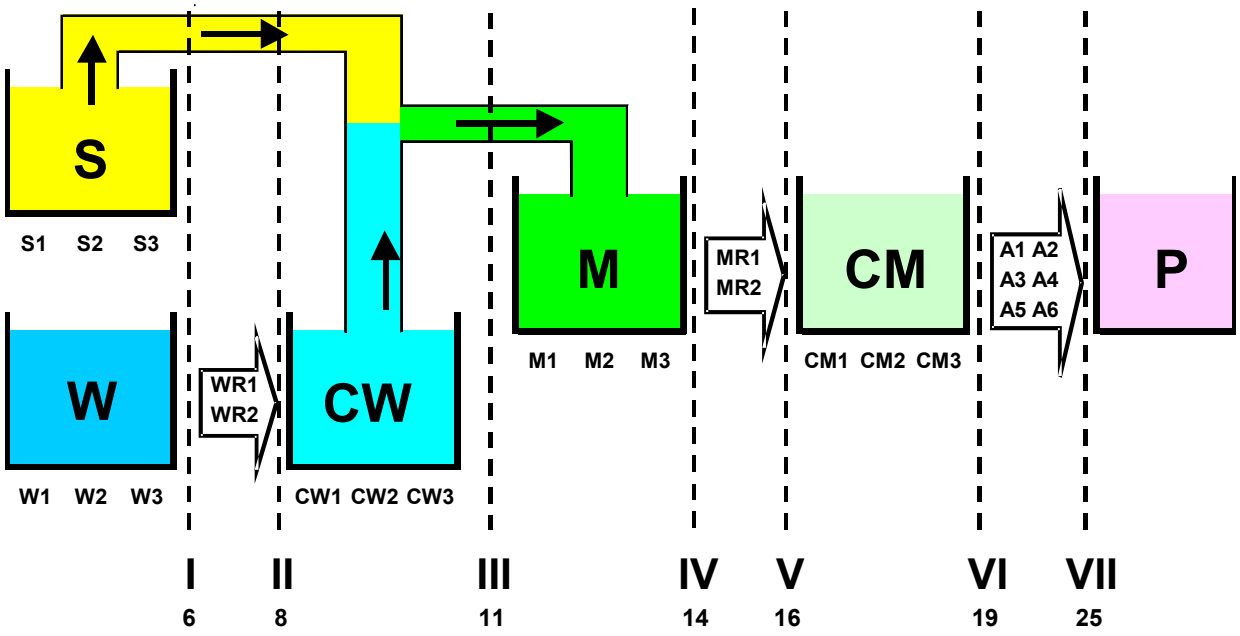


Рис. 11.1Схема многостадийного процесса.

11.2. Контроль процесса. Теория

Для начала рассмотрим аналитический контроль процесса без попыток вмешательства в него. Такой подход можно назвать *пассивной оптимизацией*. Для реализации этого подхода предлагается использовать метод *расширяющегося МСКП*, который основан на концепции мульти-блоковой регрессии [274, 290].

Предположим, что существует набор исторических данных (**X, y**), состоящий из *I* наблюдений, представляющих успешные реализации процесса. Каждое наблюдение (объект) представлено вектор-строкой (x_1, x_2, …, x_J, y) и соответствует полному циклу. Весь набор данных (**X**, **y**) разделен вертикально (по столбцам) на L блоков, соответствующих стадиям процесса:

$$\mathbf{X}=(\mathbf{X}_I, \mathbf{X}_{II},\ldots, \mathbf{X}_L).$$

Последний блок L+1 состоит из одной переменной, характеризующей выходную переменную **Y**=**y**.

В рассматриваемом примере (см. Рис. 11.1), блок $\mathbf{X}_{IV}$ – это матрица размерности (I×3), которая состоит из переменных M1, M2, и M3. Обозначение $\mathbf{X}_{(M)}$ используется для представления матрицы **X**, состоящей из переменных всех предыдущих M стадий. Для процесса, показанного на Рис. 11.1, блок $\mathbf{X}_{(III)}=(\mathbf{X}_{I}, \mathbf{X}_{II}, \mathbf{X}_{III})$, состоит из матрицы размерностью (I ×11). Применяя эти обозначения для M-той стадии процесса, видно, что матрица $\mathbf{X}_{(M-1)}$ является матрицей входных значений (предикторов), а матрица $\mathbf{X}_{M}$ обозначает текущие, контролируемые переменные. На первой стадии I, входных переменных нет, а для последней стадии L, $\mathbf{X}_{(L)}=\mathbf{X}$, т.е. все инструментальные переменные выступают в роли предикторов. Так как данные представляют собой измерения с различных приборов, перед обработкой их необходимо привести к унифицированному виду. Все данные центрированы (см. раздел 1.3), и шкалированы таким образом, что каждая переменная, включая выходную переменную y, изменяется в интервале (–1, +1). Все значения переменных вне этого интервала считаются недопустимыми. Так же предполагается, что наивысшее "качество" характеризует значение y=+1, а наихудшее, но допустимое, соответствует y=–1.

Основываясь на всем наборе данных, можно построить полную ПЛС1 модель

$$XY: \mathbf{X} \Rightarrow \mathbf{y}\ , \qquad (11.1)$$

используя K главных компонент. Оператор XY обозначает операцию, при которой по данным блока **X** (предикторы), строится значение **y** (отклик).

Также предполагается, что переменные, входящие в блок **X,** дополнительно модифицированы. Модификация заключается в умножении части переменных (столбцов) матрицы **X** на « –1». Это делается для того, чтобы все оценки регрессионных параметров **a** в уравнении регрессии

$$\mathbf{y} = \mathbf{Xa}+\boldsymbol{\varepsilon}\ , \qquad (11.2)$$

были положительными. Такая модификация приводит к тому, что любое увеличение значения переменной в **X** ведет к увеличению значения y**.** Это облегчает процедуру оптимизации, описанную в разделе 11.4. Для того, чтобы не вводить новых обозначений, символы **X** и **y** будут также использоваться и для модифицированного набора данных.

Используя исторические данные, можно построить серию из L-1 ПЛС1 регрессионной модели

$$XY_I: \mathbf{X}_{(I)} \Rightarrow \boldsymbol{y}, \qquad XY_{II}: \mathbf{X}_{(II)} \Rightarrow \boldsymbol{y}, \qquad \ldots, \qquad XY_{L-1}: \mathbf{X}_{(L-1)} \Rightarrow \boldsymbol{y}. \tag{11.3}$$

Здесь каждая модель обозначается оператором XY_M, который представляет регрессию X-блока, $\mathbf{X}_{(M)}$, на Y-блок, **y.** Каждая XY модель использует одно и тоже число ПЛС компонент *K*, которое выбирается при анализе полной модели (11.1).

Основной целью моделирования (11.3) является предсказания выходной переменной на каждой (M-ой) стадии процесса. Предсказанные значения ***y*** могут далее сравниваться с желаемым результатом. Если расхождение между прогнозируемым и стандартным значениями *y* велико, то это является сигналом о нежелательном течении процесса. В таком случае, можно скорректировать параметры следующей стадии (M+1), для того чтобы вернуть процесс к нормальному функционированию. Для проверки своего решения можно задать несколько вариантов параметров настройки стадии M+1 и, с помощью модели XY_{M+1}:$\mathbf{X}_{(M+1)} \Rightarrow \mathbf{y}$, оценить будущий результат. Таким образом, система (11.3) работает как консультант, который своими советами помогает в принятии решений. Однако такой «консультант» не может вычислить точное значение выходной переменной – при определении *y* всегда сохраняется некоторая неопределенность. Для оценки этой неопределенности применяется метод ПИО. Беря за основу уже построенные ПЛС модели с фиксированным количеством ПЛС компонент *K*, мы строим соответствующие ПИО модели и определяем максимальную погрешность β, как это описано в разделе 5.4.

Описанная выше конструкция и является расширяющимся МСКП.

11.3. Контроль процесса. Пример применения

В этом разделе демонстрируется применение теории расширяющегося МСКП к многостадийному процессу, представленному в разделе 11.1. В нашем распоряжении имеется набор исторических данных, состоящих из 154 реализаций (объектов). Данные центрированы и шкалированы, как описано в разделе 11.2.

Для построения полной регрессионной ПЛС модели (11.1) использовалось 7 главных компонент. Это число выбрано с помощью анализа результатов 10% сегментированной перекрестной проверки, которая осуществлялась следующим образом. Первые 10% объектов (в данном случае это 15 объектов), исключаются из моделирования. Параметры модели оцениваются на 90% данных, в нашем случае это 139 образцов. Далее эти параметры используются для предсказания тех 10% значений *y*, которые были исключены из моделирования. Такая операция повторяется 10 раз, и

каждый раз исключается новый набор объектов. Основные показатели качества калибровки и проверки приведены на Рис. 11.2. Здесь, в зависимости от числа ГК k, представлены следующие характеристики. Величины RMSEC (3.1) и RMSEP (3.2) характеризуют точность калибровки и прогноза в среднем, величины E_p и E_r,

$$E_p = 1 - \frac{\sum \mathrm{E}_{ij}^2}{\sum \mathrm{X}_{ij}^2} \qquad E_r = 1 - \frac{\sum f_i^2}{\sum y_i^2}, \tag{11.4}$$

показывают, какую долю X-данных и y-данных удалось описать при использовании k ПЛС ГК. В Таб. 11.1 эти же величины представлены в процентах, т.е. Cum(X)= 100%·E_p и Cum(Y)=100%·E_r.

На каждом шаге ПЛС декомпозиции вычисляется вектор X-счетов $\mathbf{t}_k$ и соответствующий ему вектор Y-счетов $\mathbf{u}_k$ (8.1). Корреляция между этими векторами

$$r_k = \mathrm{cor}\,(\mathbf{t}_k, \mathbf{u}_k) \tag{11.5}$$

служит важным индикатором, показывающим завершенность процесса ПЛС декомпозиции.

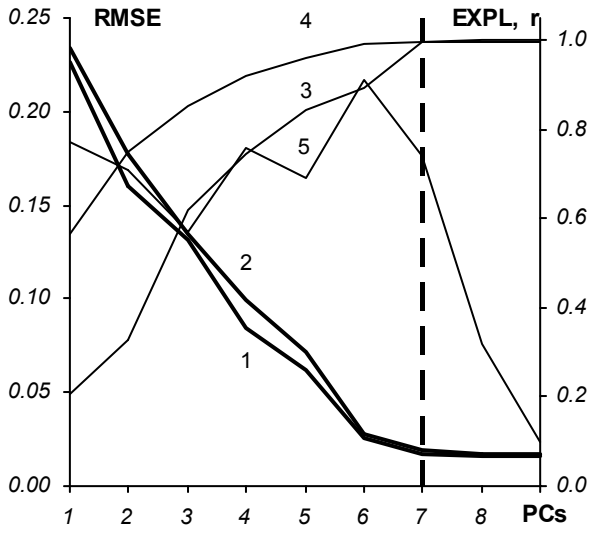


Рис. 11.2 Построение полной ПЛС модели.
Основные показатели относительно номера ГК. Левая ось Y: 1- RMSEC, 2-RMSEP. Правая осьY 3- E_p, 4- E_r, 5 – r=cor(t, u)

На основании анализа этих показателей число главных компонент следует выбрать равным шести или семи. Обычно рекомендуется выбирать минимально возможное количество ГК, для того чтобы избежать переопределенности, а, следовательно, и неустойчивости модели. На шести ГК коэффициент корреляции между векторами ($\mathbf{t}_k$, $\mathbf{u}_k$)

r_6=0.916 достигает максимума, при этом модель, объясняет 95% вариации в **X** матрице и 99% в **y**. К тому же значения RMSEC (3.1) и RMSEP (3.2) сравниваются на уровне 0.026. Однако более детальное изучения парных графиков счетов ($\mathbf{t}_k$, $\mathbf{u}_k$) Рис. 11.3 и Таб. 11.1 позволило принять другое решение.

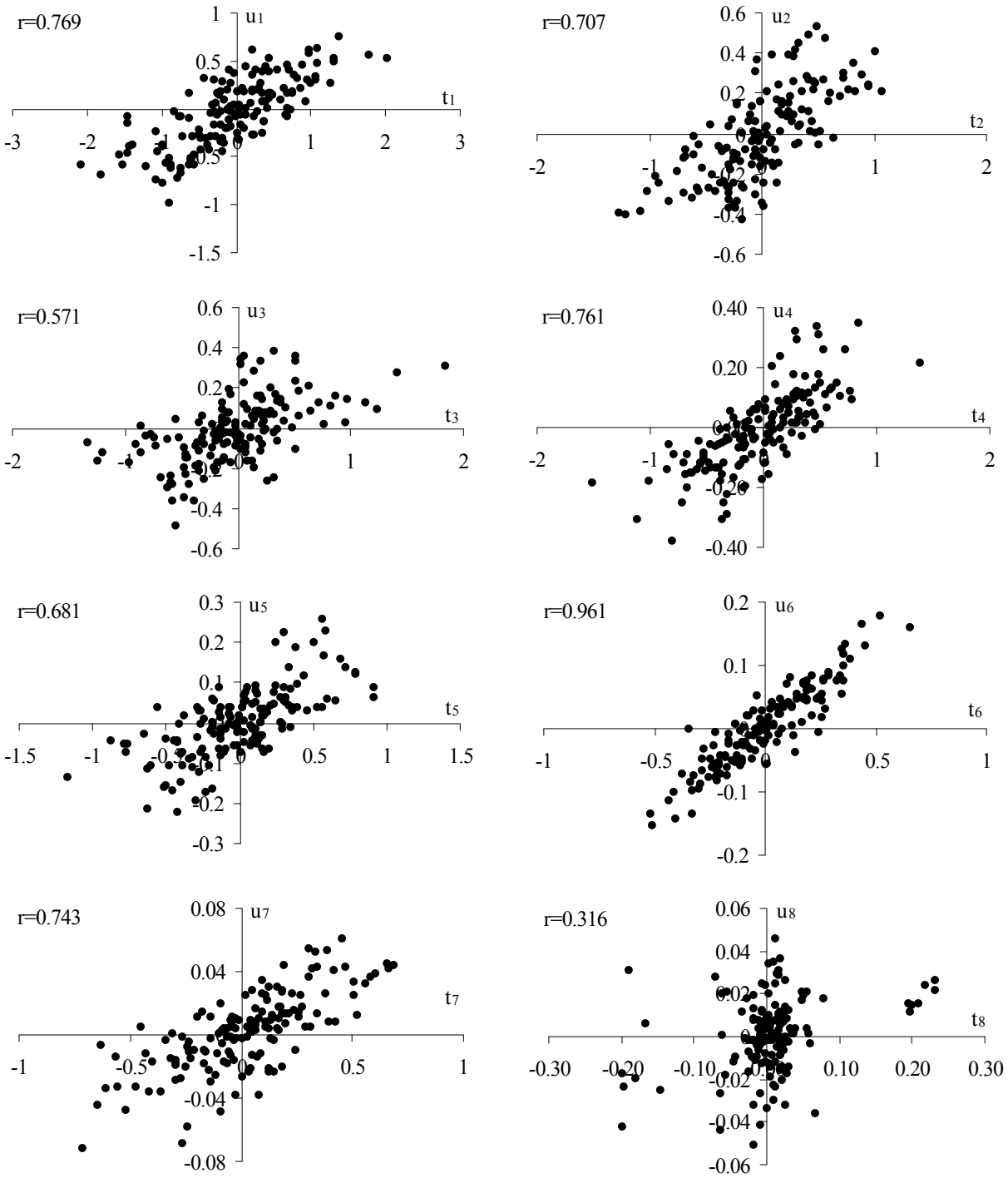


Рис. 11.3 . Графики X-счетов, t_k относительно Y-счетов, u_k, k=1,...,8

Рассмотрим подробнее графики счетов $\mathbf{t}_k$ относительно Y-счетов, $\mathbf{u}_k$. Для ПЛС1, при наличие только одного отклика, **u**-векторы представляют из себя просто усеченные векторы **y** (см.(8.1)**.** Для k=1,...,8, графики $\mathbf{u}_k$ относительно $\mathbf{t}_k$ показаны на Рис. 11.3

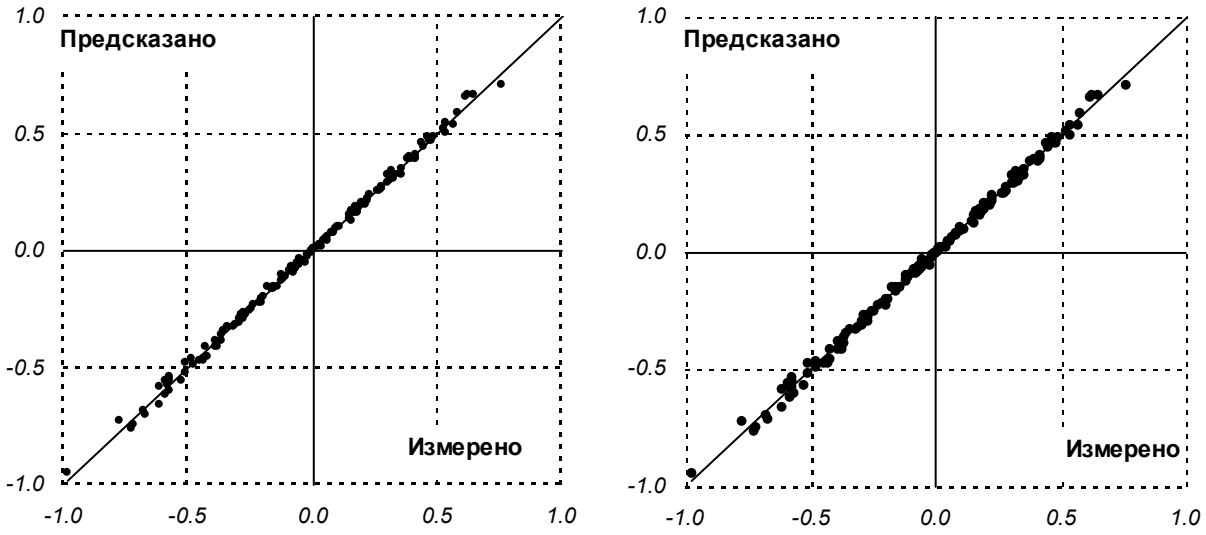
Эти графики подтверждают правильность выбора не шести, а семи компонент. Коэффициент корреляции между $\mathbf{t}_7$ и $\mathbf{u}_7$ достаточно велик и равен 0.713. Можно заметить, что изменения значения отклика колеблются между –0.08 и 0.08. Если эти значения меньше, чем точность измерения этого показателя, то необходимо выбирать шесть ГК. В Таб. 11.1 показано, какая часть изменений в **X** и **y** описывается на каждом шагу проецирования, т.е. при каждом увеличении размерности k.

Таб. 11.1 Общая доля X, Cum(X), и общая доля Y, Cum(Y), в процентах, которая описывается на каждом шаге проецирования

Размерность	1	2	3	4	5	6	7	8	9	10
Cum(**X**)	31.15	46.91	72.78	81.76	92.78	95.69	99.75	99.98	99.99	100.00
Cum(**Y**)	57.86	78.61	85.33	95.75	97.37	99.57	99.79	99.79	99.79	99.79

Из Таб. 11.1 видно, что при семи ГК, описывается 99.75% от **X** и 99.79% от **y**. Седьмая главная компонента описывает 4.1% от **X** и 0.2% от **y.** Это не очень большой вклад, но с вычислительной точки зрения существенное улучшение результата по сравнению с шестью компонентами. Сама по себе шестая главная компонента так же дает незначительный вклад в описание блоков X и Y. Однако, как видно из Рис. 11.3, коэффициент корреляции между счетами $\mathbf{t}_6$ и $\mathbf{u}_6$ очень высок, и равен 0.916. В этом проявляется один из недостатков метода ПЛС, когда существенный коэффициент корреляции описывает относительно незначительные части данных X и Y. Это является косвенным свидетельством того, что можно улучшить модель, если предварительно, перед проецированием, произвести отбор переменных и выбрать ту часть данных матрицы X, которая влияет на значения Y. Способам отбора переменных посвящена обширная литература, например [63, 65]. Так как в нашем случае точность описания достаточна, предварительного отбора переменных не делалось.

Для подтверждения правильности выбора семи ГК, полезно так же взглянуть на графики предсказанных значений относительно измеренных для модели из семи ГК (Рис. 11.4).



a) Значения, вычисленные при моделировании b) Значения, вычисленные по перекрестной проверке

Рис. 11.4 Предсказанных значений относительно измеренных для полной ПЛС модели из 7 ГК

На Рис. 11.4a представлена зависимость значений отклика, вычисленных относительно измеренных. Видно, что точки расположены близко к прямой, проходящей под углом 45°. Объясненная дисперсия Cum(**Y**)= 99.790%, а RMSEC=0.0166.

Значения y, вычисленные с помощью перекрестной проверки, представляют типичный случай, когда новые значения отклика оцениваются с помощью построенной модели (Рис. 11.4b). Коэффициент корреляции между измеренными и предсказанными значениями r=0.9989, а $Q^2=(r)^2$ =0.9977. Среднеквадратичный остаток прогноза RMSEP=0.017 показывает, что при использовании модели с семью главными компонентами, значения отклика в 95% случаях могут быть предсказаны с точностью не хуже чем ±2×0.017=±0.034, а в 99.99% случаев с точностью не хуже чем ±4×0.017=±0.068. В дальнейшем, уровень неопределенности, оцененный с помощью ПЛС моделирования, будет сравниваться с максимальной погрешностью β, вычисленной с помощью метода ПИО.

Правильность решения относительно выбора семи ГК была подтверждена при дополнительных исследованиях (см. раздел 12.8).

Теперь перейдем к построению расширяющихся ПЛС моделей (11.3). Для рассматриваемого примера строится серия из семи моделей, при этом в каждой из этих моделей, число главных компонент равно семи (кроме первой модели, где K=6), как было

определено для полной модели (11.1). Для каждой такой модели блок данных **X** состоит из всех переменных, которые измерялись на предыдущих стадиях, т.е.

$\mathbf{X}_{(I)}=\mathbf{X}_{I}=$(S1, S2, S3, W1, W2, W3),

$\mathbf{X}_{(II)}=(\mathbf{X}_{I}, \mathbf{X}_{II})=$(S1, S2, S3, W1, W2, W3, WR1, WR2), ... , и т.д.

Для того чтобы проверить каждую из построенных моделей и оценить точность ее предсказания, все данные были разделены на две части: калибровочный набор (102 объекта), и проверочный, или тестовый набор (52 объекта). Калибровочный набор используется для построения моделей, а тестовый набор используется для оценки неопределенности прогноза. Основные характеристики построенных калибровок представлены в Таб. 11.2.

Таб. 11.2 Основные характеристики расширяющегося ПЛС моделирования (11.3)

	Стадии:	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
Калибровка	$\mathbf{E_p}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\mathbf{E_r}$	0.78	0.85	0.84	0.90	0.99	0.99	1.00
	RMSEC	0.157	0.126	0.132	0.102	0.038	0.023	0.016
Проверка	$\mathbf{E_p}$	1.00	1.00	1.00	1.00	0.99	0.99	0.99
	$\mathbf{E_r}$	0.85	0.90	0.91	0.94	0.98	0.99	1.00
	RMSEP	0.148	0.122	0.123	0.095	0.045	0.024	0.018

На основании построенных ПЛС моделей строится серия соответствующих ПИО моделей. В Таб. 11.3 представлены основные характеристики, относящиеся к ПИО моделированию.

Таб. 11.3 Основные характеристики расширяющегося ПИО моделирования (11.3)

Стадии:	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
$\boldsymbol{b}$	0.80	0.70	0.60	0.50	0.40	0.20	0.10
$\boldsymbol{b}_{\mathbf{min}}$	0.51	0.39	0.40	0.23	0.08	0.06	0.04
$\boldsymbol{b}/\boldsymbol{s}_{\mathbf{c}}$	5.41	3.59	3.01	3.26	7.04	5.62	4.13
$\boldsymbol{b}_{\mathbf{min}}/\boldsymbol{s}_{\mathbf{c}}$	3.25	1.99	2.05	1.49	1.48	1.60	1.65
$\boldsymbol{w}$	0.68	0.65	0.56	0.49	0.43	0.21	0.09
$\boldsymbol{h}$	0.80	0.93	0.94	0.98	1.07	1.04	0.96

В Таб. 11.3 представлена следующая информация. Первая строка, обозначенная b – это верхняя граница β, вычисленная по формуле (5.25). Следующая строка, $b_{\min}$ – это оценка β, вычисленная по формуле (5.21). Как видно из таблицы, оба значения убывают с расширением модели. Это согласуется с предположением, что погрешность моделирования падает по мере увеличения объема данных. Следующие две строки представляют отношения b/s_c и $b_{\min}/s_c$, где s_c=RMSEC. Видно, что первое отношение близко к величине 4.5, а второе примерно 2. Этот результат согласуется с практическим правилом ‘1-2-3-4 сигма’, предложенным в разделе 5.4. Строка w содержит средние значения ширины ПИО интервала (4.13), вычисленные на проверочном наборе. Видно, что по мере расширения матрицы предикторов, неопределенность прогноза уменьшается. В строке h представлены средние значения ПИО размаха (6.9), вычисленные на проверочном наборе. Видно, что значения ПИО размаха мало меняются по мере расширения модели.

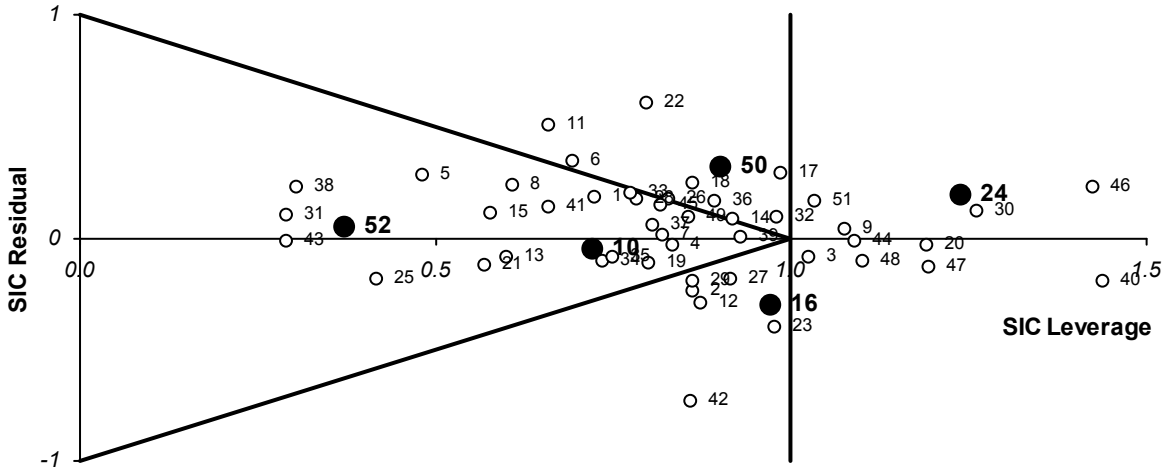


Рис. 11.5 Диаграмма статуса объектов. Полная ПЛС модель, после VII этапа.

На Рис. 11.5 представлена ДСО, построенная для проверочного набора с помощью полной модели (11.1). Полезно сравнить этот график с архетипом диаграммы статуса объектов (Рис. 6.2b). Видно, что среди 52 проверочных объектов, имеется 24 внутренних объекта, которые лежат внутри треугольной области модели (например, объекты №10 и №52). Оставшиеся 28 образов являются внешними, например, объекты №24, №46, №50. Среди внешних объектов можно выделить 11 абсолютно внешних, таких как объект №24. На ДСО не выявлено ни одного выброса. Для всех объектов значения ПИО размаха меньше чем 1.5, даже для объектов №40 и №46, занимающих самое правое положение на ДСО.

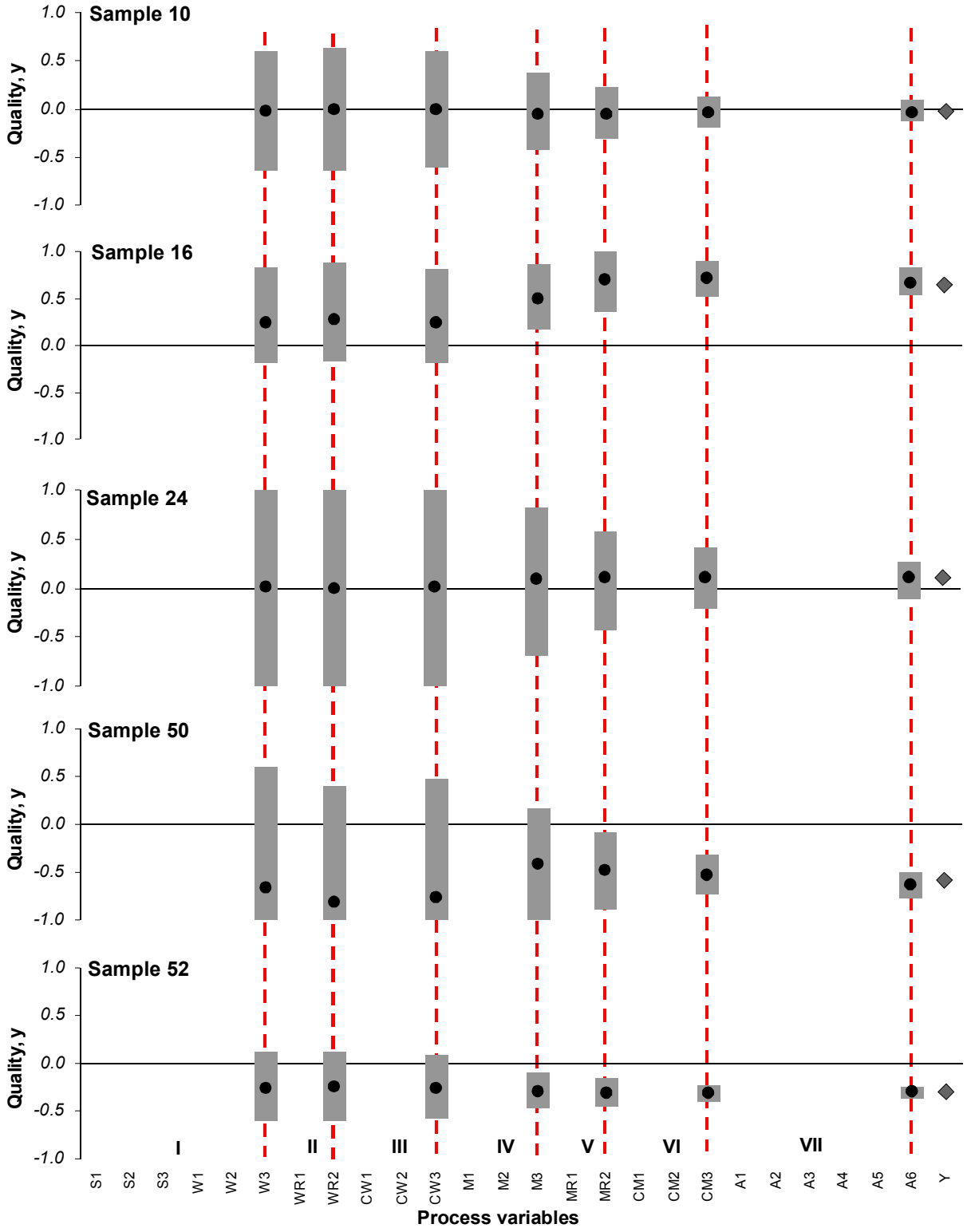


Рис. 11.6 Предсказание показателя «качество» на каждой стадии процесса для отобранных объектов. ПИО интервал (серый прямоугольник), ПЛС прогноз (черные кружки). Ромб в правой части графиков – проверочные значения *y*.

Таким образом, ДСО дает представление о качестве проверочного набора. Например, если бы все тестовые объекты были бы внутренними, то оценка точности прогноза на таком проверочном наборе была бы излишне оптимистичной.

Для того чтобы показать, как работает расширяющееся моделирование, были отобраны 5 наиболее характерных образов, представляющих все возможные случаи. Это объекты №№ 10, 16, 24, 50 и 52, которые отмечены крупными черными кружочками на Рис. 11.5. Результаты расширяющегося ПЛС моделирования и соответствующих ПИО моделей представлены на Рис. 11.6. Каждый график показывает результаты предсказания показателя *y*, который оценивается после завершения соответствующей стадии процесса. При этом, точечная ПЛС оценка обозначена черным кружком, а интервальная ПИО оценка – серым прямоугольником. Фактическое измеренное значение величины *y* обозначено ромбом.

Из Рис. 11.6 видно, что по мере продвижения вдоль процесса, ширина ПИО интервала уменьшается. При этом проверочное значение *y* всегда располагается внутри этого интервала. Так же видно, что ширина ПИО интервалов, а следовательно и степень неопределенности при прогнозе, меньше для внутренних объектов (№10 и №52) и наибольшая для абсолютно внешних объектов (№24).

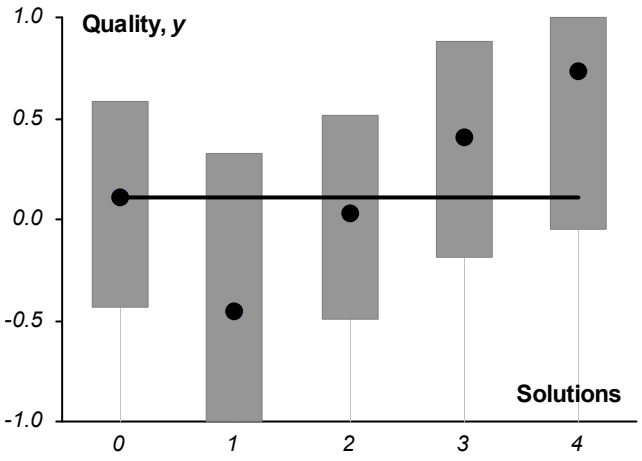


Рис. 11.7 Предсказания переменной «качество» при различных решениях на стадии V

Рассмотрим, как проводится пассивная оптимизация на примере объекта №24 в точке сразу после стадии IV. На следующем V этапе два параметра, переменные MR1 и MR2, могут быть подправлены таким образом, чтобы улучшить выходное значение, т.е. показатель *y*. Четыре возможные решения представлены на Рис. 11.7

Нулевое решение дано для сравнения и представляет измеренные показатели, реализованные в исторических данных. Решение 1 – это использование средних значений величин MR1 и MR2. Так как данные центрированы, то эти значения равны нулю. Решение 2 соответствует прогнозу по ПЛС2 модели $\mathbf{X}_{(IV)} \Rightarrow \mathbf{X}_{V,}$ которое будет подробно рассмотрено в разделе 11.4. Решения 3 и 4 – это некоторые экспертные значения, которые могут быть установлены опытным путем.

На Рис. 11.7 показаны предсказанные значения величин *y*: в виде точечной ПЛС оценки (черные кружки) и прогнозного интервала вычисленного по ПИО методу (серые прямоугольники). Горизонтальная линия представляет реально полученное значение *y* для объекта №24 и используется как контрольный уровень. Видно, что решения 1 и 2 не улучшают выходной показатель, в то время как решения 3 и 4 ведут к увеличению значения *y*. В то же время, изменения, прелагаемые решением 4, могут быть слишком резкими и практически нереализуемыми. Вопрос о том, как правильно выбрать корректирующие значения инструментальных переменных по ходу процесса рассматривается в следующем разделе.

11.4. Оптимизация процесса. Теория

В этом разделе объясняется, как осуществить *активную оптимизацию*, т.е. как выбрать оптимальные действия по корректировке переменных, которые могут производиться в конце каждой стадии процесса. Фактически это означает правильный выбор значений контролируемых переменных, которые являются входными для следующей стадии процесса. При выборе корректирующих действий необходимо придерживаться двух основных правил. Подправленные величины входных переменных, во-первых, должны увеличивать значения *y*; во-вторых, значения этих переменных должны находиться внутри допустимых контролируемых границ. Предлагаемый подход базируется на концепции статуса объектов метода ПИО.

Рассмотрим случай, когда изучаемый процесс состоит из трех блоков. Переменные в блоках **X** и **Z** – это известные данные предыдущих реализаций процесса, переменная **y** – это соответствующие этим реализациям процесса значения выходного показателя, и они также известны. Основной целью является предсказания величины *y* для новых значений инструментальных переменных (**x**, **z**). Однако, не все значения этих переменных известны на текущий момент. Значения переменных **z** (см. Рис. 11.8) неизвестны. Значения **x**, которые ассоциируются с блоком X, известны, и их можно

использовать как для предсказания значений **z**, так и значения *y*. Требуется найти такие «оптимальные» значения **z**, которые могут максимизировать будущее значение *y*. Таким образом, нас интересует не предсказания значений **z** само по себе, а поиск таких значений **z**, которые улучшают показатель *y*.

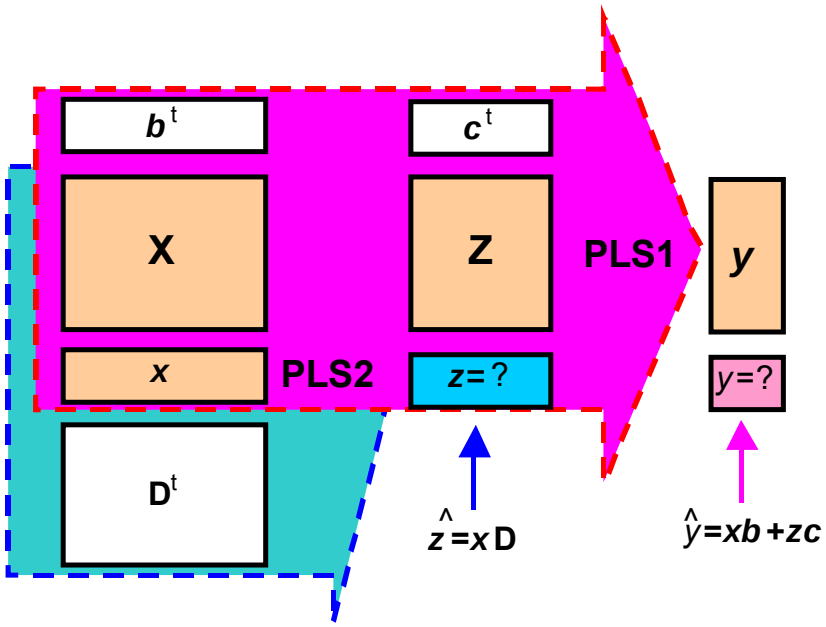


Рис. 11.8 Трехблочная схема процесса

Для поиска подходящего значения **z** можно воспользоваться двумя ПЛС моделями. Первая из них, это полная ПЛС1 модель, которая предсказывает значения откликов **y**, используя объединенный блок предикторов **X** и **Z**, т.е.

$$\text{XY}: (\mathbf{X}, \mathbf{Z}) \Rightarrow \mathbf{y} \qquad (11.6)$$

Обозначим через **b** и **c** регрессионные коэффициенты, полученные при построении модели XY на базе исторических данных (**X**, **Z**, ***y***), т.е.

$$\hat{\boldsymbol{y}} = \text{XY}(\mathbf{X}, \mathbf{Z}) = \mathbf{X}\boldsymbol{b} + \mathbf{Z}\boldsymbol{c}$$

Важно отметить, что вектор **a** исходной модели (11.2) как раз и состоит из векторов **b** и **c**. Задачу оптимизации в данном случае можно поставить следующим образом

$$f(\mathbf{z}) = (\mathbf{x}^{\mathrm{t}}\mathbf{b} + \mathbf{z}^{\mathrm{t}}\mathbf{c}) \rightarrow \mathbf{max} \quad \text{при условии, что } \mathbf{z} \in L_z, \qquad (11.7)$$

где L_z – это область допустимых значений **z**. Область L_z будет определена позднее, а сейчас отметим только самые общие соображения по выбору этой области.

Основная сложность в линейной оптимизации заключается не в поиске самого решение, а в ограничении той области L_z , на которой этот оптимум достигается. Дело в том, что поиск решений линейной задачи оптимизации на неограниченной области

приводит к несодержательным результатам. В первую очередь ограничения на область L_z накладываются физическими свойствами измеряемых величин, а так же условиями процесса. Без потери общности, эти ограничения можно записать в виде

$$\left| z_i \right| < 1, \text{ for } i = 1,2,... \qquad \text{и} \qquad \left| \mathbf{x}^t\mathbf{b} + \mathbf{z}^t\mathbf{c} \right| < 1$$

Учитывая дополнительное шкалирование данных (раздел 11.2), модель XY обладает тем свойством, что увеличение значения любой переменной z_i вектора $\mathbf{z}$ приводит к увеличению значения выходной переменной y. Таким образом, максимальное значение всегда достигается на верхней границе контролируемой инструментальной переменной. Следовательно, все, что нужно сделать для решения задачи оптимизации, это определить разумные границы изменения переменных процесса. Грубые границы можно получить, используя шкалирование данных. Однако границы для всех переменных в виде (–1, +1) слишком широкие, и оптимизация с учетом таких ограничений приводит к нереалистичному значению y, которое становится больше +1, т.е. выходит за предел своего допустимого уровня. Это происходит из-за сильной корреляции между инструментальными переменными, поэтому установить все значения этих переменных в состояние (+1, +1, …, +1) невозможно. Область L_z должна включать в себя только такие значения $\mathbf{z}$, которые не противоречат предшествующему опыту. Это означает, что возможные значения $\mathbf{z}$ должны согласовываться с ПЛС моделью (11.6), которая описывает исследуемый процесс. Здесь необходимо отметить, что такая постановка задачи тесно связана с другой, хорошо известной задачей многомерного анализа, а именно, с задачей «определения выбросов при прогнозе» [67, 118], которая рассматривалась в Главе 8. Предсказывая новый объект, необходимо проверять, подходит ли для этого объекта построенная модель калибровки, либо предсказываемый объект сильно отличается от объектов калибровочного набора, что может привести к сомнительным результатам прогноза. Так как при прогнозировании новых объектов значения отклика y неизвестно, предсказание выбросов может основываться только на значениях X переменных. Для оценки качества нового объекта, используются две меры близости объекта и модели. Первый критерий, это величина размаха, которая показывает, как далеко новый объект находится от центра модели, при этом расстояние рассматривается только на гиперплоскости главных компонент, образованной моделью. А второй критерий, – это трансверсальное расстояние от объекта до гиперплоскости модели. Обе меры являются случайными величинами, поэтому для их оценки предлагается применить статистические критерии.

Дадим формальное определение области L_z, используя метод ПИО. Пусть **h** – это вектор, состоящий из значений ПИО размаха, а вектор **d**

$$d_i = \sqrt{\mathbf{e}_i^t \mathbf{e}_i}, \quad \text{где } \mathbf{e}_i^t = \mathbf{x}_i^t - \mathbf{t}_i \mathbf{P}^t \tag{11.8}$$

это корень квадратный из X остатков, который определяет расстояние от объекта до гиперплоскости ПЛС модели. Векторы **h** и **d** можно получить на этапе проверки полной модели XY (11.6) с использованием проверочного набора ($\mathbf{X}_{test}$, $\mathbf{Z}_{test}$), либо с помощью перекрестной проверки. Эти значения нельзя подсчитать на этапе построения модели, так как для всех объектов из калибровочного набора значения ПИО размаха всегда меньше или равно 1 (6.12). Рассматривая векторы **h** и **d** как репрезентативную выборку, можно найти их критические значения, например, можно установить по 4 критических значения для каждого критерия:

$$\begin{aligned} l_0&=m_h, \quad l_1=m_h+s_h, \quad l_2=m_h+2s_h, \quad l_3=m_h+3s_h, \\ r_0&=m_d, \quad r_1=m_d+s_d, \quad r_2=m_d+2s_d, \quad r_3=m_d+3s_d, \end{aligned} \tag{11.9}$$

где m_h и m_d средние значения, а s_h и s_d стандартные отклонения, подсчитанные для векторов **h** и **d** соответственно. Эти уровни устанавливают критическое положение для нового объекта относительно полной ПЛС модели.

Используя границы (11.9), которые вычислены на основании исторических данных, можно определить, является ли новый вектор **z** допустимым значением с точки зрения оптимизации или нет. С этой целью, для объединенного вектора (**x**, **z**), вычисляются значения h_z и d_z и сравниваются с установленными критическими уровнями l_c и r_c. Если $h_z < l_c$, и $r_z < r_c$, тогда $\mathbf{z} \in L_z$. Используя разные критические уровни, введенные в (11.9), можно выбирать различные стратегии оптимизации: осторожную, решительную и т.д.

Теперь рассмотрим вторую модель, показанную на схеме Рис. 11.8. Это ПЛС2 регрессионная модель, которая предсказывает значения блока **Z** по блоку предикторов **X**, т.е.

$$\text{XZ: } \mathbf{X} \Rightarrow \mathbf{Z} \tag{11.10}$$

Используя исторические данные (**X**, **Z**), можно построить модель калибровки $\hat{\mathbf{Z}} = \text{XX}(\mathbf{X}) = \mathbf{XD}$, где **D** это матрица параметров модели. Применяя эту модель к новому вектору **x**, можно предсказать значения вектора **z**. Этот вектор, $\hat{\mathbf{z}} = \text{XX}(\mathbf{x})$, не является решением оптимизационной задачи (11.7); однако, $\hat{\mathbf{z}}$ очевидно является допустимым решением и, следовательно, принадлежит области L_z. В процессе оптимизации каждый

компонент вектора $\hat{\mathbf{z}}$ можно изменять до тех пор, пока новый вектор $\mathbf{z}^+$ будет оставаться в пределах области L_z. Это действие можно представить с помощью оператора G

$$\mathrm{G}(\hat{\mathbf{z}}) = \mathbf{z}^+, \qquad (11.11)$$

который определяет стратегию оптимизации. Очевидно, что благодаря дополнительному шкалированию X-переменных, вектор-строка G(**z**) должна быть, покомпонентно, больше или равна **z**. Единичный оператор

$$\mathrm{G}(\hat{\mathbf{z}}) = \hat{\mathbf{z}}. \qquad (11.12)$$

дает тривиальное решение, при котором никакого увеличения значений не происходит. Другие варианты оператора G рассматриваются в следующем разделе.

11.5. Оптимизация процесса. Пример применения

Покажем, как работает активная оптимизация применительно к процессу, описанному в разделе 11.1. Используя подход, предложенный в предыдущем разделе, корректирующие действия можно проводить двумя способами. Первый способ – это увеличение каждого компонента вектора **z**, до тех пор, пока не будет нарушено условие $\mathbf{z} \in L_z$. Однако это не лучший способ, так как компоненты вектора **z** связаны между собой. Второй путь более систематичный. Предлагается выбрать какой-либо общий метод оптимизации **z**, и применять его последовательно, на всех стадиях технологического процесса. При этом необходимо проверять, что оптимальные значения $\mathbf{z}^+$ согласуются с критерием (11.9). Второй способ, как более интересный для изучения и был применен.

Серия расширяющихся ПЛС1 моделей (11.3) уже была представлена в разделе 11.2

$$\mathrm{XY_I}{:}\mathbf{X}_{(\mathrm{I})} \Rightarrow \boldsymbol{y},\ \mathrm{XY_{II}}{:}\mathbf{X}_{(\mathrm{II})} \Rightarrow \boldsymbol{y},\ \ldots,\ \mathrm{XY_{L\text{-}I}}{:}\mathbf{X}_{(\mathrm{L\text{-}1})} \Rightarrow \boldsymbol{y}$$

Эти модели соответствуют моделям (11.6) в трехблочной схеме процесса. Кроме того, можно построить серию ПЛС2 моделей аналогичных (11.10), т.е.

$$\mathrm{XX_I}{:}\mathbf{X}_{(\mathrm{I})} \Rightarrow \mathbf{X}_{\mathrm{II}},\ \mathrm{XX_{II}}{:}\mathbf{X}_{(\mathrm{II})} \Rightarrow \mathbf{X}_{\mathrm{III}},\ \ldots.,\ \mathrm{XX_{VI}}{:}\mathbf{X}_{(\mathrm{VI})} \Rightarrow \mathbf{X}_{\mathrm{VII}} \qquad (11.13)$$

С помощью этих моделей можно предсказывать значения будущих переменных блока **X** (переменные **Z**), зная предыдущие значения X-переменных (**X**). Эти модели строятся на основе калибровочного набора, представленного в разделе 11.3. Каждая модель, обозначенная оператором $\mathrm{XX_M}$, является ПЛС2-регрессией блока $\mathbf{X}_{(\mathrm{M})}$ на блок $\mathbf{X}_{\mathrm{M+1}}$. Модели получаются разной степени сложности, т.е. у них разное количество ГК,

различные значения ПИО параметра β, и т.д. Основные характеристики этих моделей приведены в Таб. 11.4.

Таб. 11.4 Основные характеристики ПЛС2 моделей (11.13)

Модель	ГК	E_p	Переменные	E_r	b	s_c	b/s_c
$XX_I : \mathbf{X}_{(I)} \Rightarrow \mathbf{X}_{II}$	6	1.00	WR1	1.00	0.03	0.010	2.63
			WR2	1.00	0.05	0.012	3.79
$XX_{II} : \mathbf{X}_{(II)} \Rightarrow \mathbf{X}_{III}$	6	1.00	CW1	1.00	0.03	0.008	3.70
			CW2	1.00	0.05	0.011	4.16
			CW3	1.00	0.02	0.005	4.30
$XX_{III} : \mathbf{X}_{(III)} \Rightarrow \mathbf{X}_{IV}$	6	1.00	M1	0.99	0.12	0.027	4.40
			M2	1.00	0.06	0.015	4.11
			M3	0.98	0.14	0.033	4.11
$XX_{IV} : \mathbf{X}_{(IV)} \Rightarrow \mathbf{X}_{V}$	6	1.00	MR1	0.99	0.08	0.029	2.90
			MR2	0.99	0.06	0.020	3.03
$XX_V : \mathbf{X}_{(V)} \Rightarrow \mathbf{X}_{VI}$	7	1.00	CM1	1.00	0.06	0.014	4.14
			CM2	1.00	0.03	0.006	4.28
			CM3	1.00	0.01	0.002	4.13
$XX_{VI} : \mathbf{X}_{(VI)} \Rightarrow \mathbf{X}_{VII}$	8	1.00	A1	1.00	0.05	0.012	3.84
			A2	1.00	0.05	0.011	4.48
			A3	1.00	0.05	0.009	4.96
			A4	1.00	0.07	0.016	4.24
			A5	1.00	0.08	0.017	4.42
			A6	1.00	0.06	0.012	4.90

Например, модель XX_{III}: $\mathbf{X}_{(III)} \rightarrow \mathbf{X}_{IV}$ предсказывает четвертый блок переменных $\mathbf{X}_{IV}$, состоящий из переменных M1, M2 и M3 (матрица размерности 102×3), с помощью блока $\mathbf{X}_{(III)}$, состоящего из 11 переменных, представленных матрицей (102×11).

Соответствующая ПЛС2 модель использует шесть ГК. Блок предикторов моделируется с точностью, близкой к 100%. Отклики M1, M2 и M3 моделируются с точностью 99%, 100%, и 98 %, а оценки параметров β равны b=0.12, 0.06, и 0.14, соответственно. В сравнении со значениями RMSEC = s_c = 0.027, 0.015, 0.033, получаются следующие граничные отношения для b/s_c = 4.40, 4.11, и 4.11.

Рассматриваемый подход выражает одну из основных концепций многомерного анализа данных – какие переменные считать предикторами, а какие откликами, часто зависит от решаемой задачи [71]. В нашем случае, роли переменных меняются по ходу процесса. Что касается рассматриваемого примера, то модели (11.13) дают наиболее реалистичные оценки значений будущих переменных, которые согласуются с историческими данными. Эти модели могут применяться к любым новым реализациям процесса, для того чтобы предсказать переменные на следующем этапе производства, в блоке **Z**.

Рассмотрим различные возможные решения оптимизационной задачи (11.7). Для этого 52 объекта из проверочного набора будут рассматриваться как новые объекты, которые требуют оптимизации. Так как фактические значения y известны, то этот набор будет выступать в качестве контрольного набора. Наша цель – сравнить между собой различные методики, а так же сопоставить результаты применения оптимизационных методик с контрольными результатами.

Рассмотрим распределение контрольных объектов по их значениям переменной y. Разделим интервал [–1.0, +1.0] на десять равных корзин и разложим 52 объекта по этим корзинам, в зависимости от их y-значений; далее отшкалируем частоты на количество объектов и получим гистограмму распределения объектов контрольного набора. Из Рис. 11.9а видно, что объекты распределены симметрично со средним значением M= –0.10, и стандартным отклонением S=0.38.

Применительно к данному примеру общий алгоритм оптимизации выглядит следующим образом. Рассмотрим процесс после стадии I, когда следующий блок X переменных, $\mathbf{X}_{II}$, должен быть скорректирован. ПЛС2 модель XX_I: $\mathbf{X}_I \Rightarrow \mathbf{X}_{II}$ вычисляет оценки $\hat{\mathbf{X}}_{II}$ X-переменных, которые могут использоваться для управления процессом. Для того, чтобы получить оптимальное решение, применим некоторый увеличивающий оператор G, $G(\hat{\mathbf{X}}_{II}) = \mathbf{X}_{II}^{+}$ (11.11). После этого, значения $\mathbf{X}_{II}^{+}$ объединяются в один блок

со значениями $\mathbf{X}_{\mathrm{I}}$, и блок $\mathbf{X}^{+}_{(\mathrm{II})} = (\mathbf{X}_{\mathrm{I}}, \mathbf{X}^{+}_{\mathrm{II}})$ используется в качестве входных, предикторных переменных для моделирования следующей стадии, III.

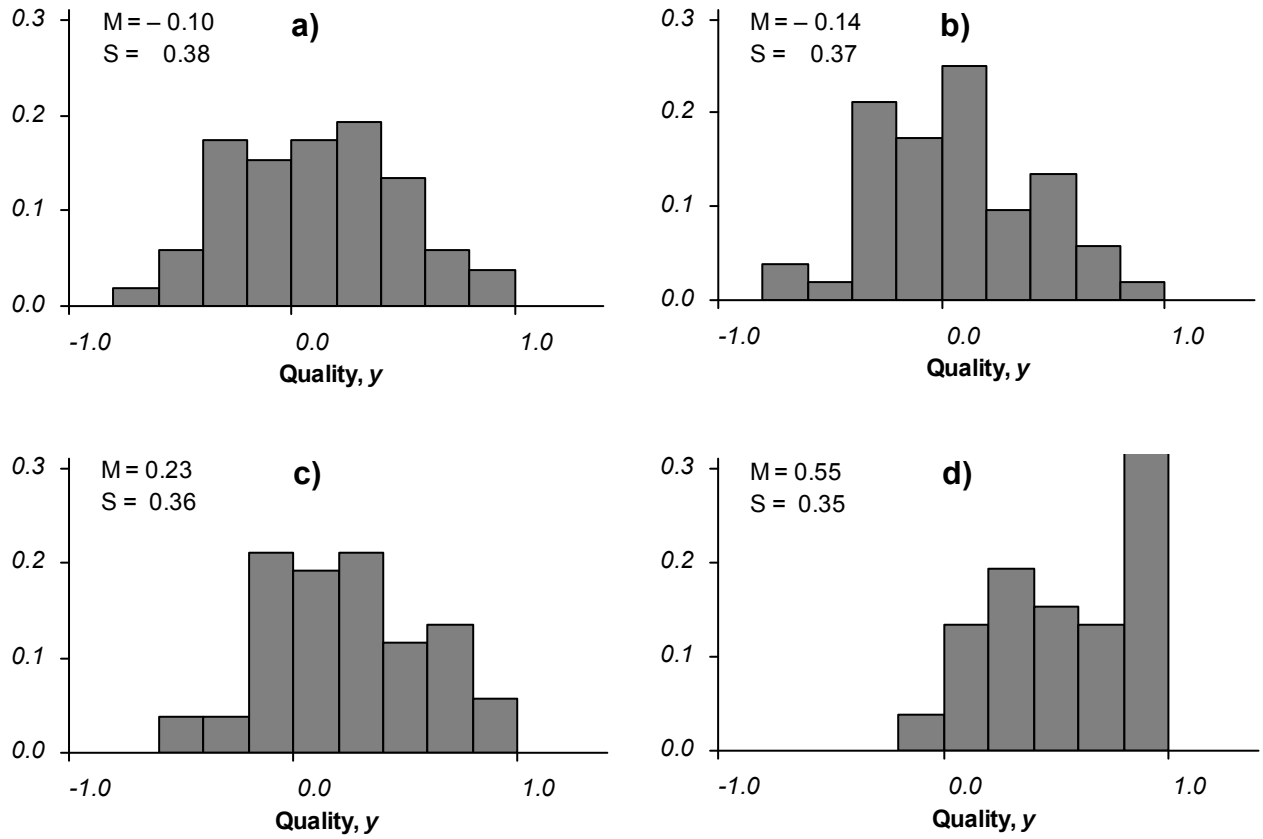


Рис. 11.9 Распределение объектов по переменной *y*

a) Контрольный набор (до оптимизации), b) Оптимизация, тип «внутренний»,

c) Оптимизация, тип «внешний» d) Оптимизация, тип «выбросы»

Затем блок $\mathbf{X}^{+}_{(\mathrm{II})}$ обрабатывается подобным же образом, т.е.

$$\mathrm{XX}_{\mathrm{II}}(\mathbf{X}^{+}_{(\mathrm{II})}) = \hat{\mathbf{X}}_{\mathrm{III}}$$
$$\mathrm{G}(\hat{\mathbf{X}}_{\mathrm{III}}) = \mathbf{X}^{+}_{\mathrm{III}}$$
$$(\mathbf{X}^{+}_{(\mathrm{II})}, \mathbf{X}^{+}_{\mathrm{III}}) = \mathbf{X}^{+}_{(\mathrm{III})}$$

Повторяя эту процедуру на каждой стадии, можно получить набор из X-переменных $\mathbf{X}^{+} = (\mathbf{X}_{\mathrm{I}}, \mathbf{X}^{+}_{\mathrm{II}}, \ldots \mathbf{X}^{+}_{\mathrm{VII}})$, который может быть использован на каждой стадии процесса для предсказания значений переменной качества *y*. Для этого используется расширяющееся моделирование, и строятся ПЛС1 модели вида (11.3)

$$\boldsymbol{y}^{+}_{\mathrm{M}} = \mathrm{XY}\left(\mathbf{X}^{+}_{(\mathrm{M})}\right), \quad \mathrm{M} = \mathrm{II}, \mathrm{III}, \ldots \mathrm{VII},$$

Они дают нам точечные оценки. Соответствующие ПИО модели дают и интервальные оценки, так как это описано в разделе 11.3. Такой алгоритм был реализован для трех различных увеличивающих операторов G (11.11), которые представляют разные стратегии оптимизации.

Первая стратегия использует единичный оператор G. В этом случае, в качестве оптимального решения используются оценки, вычисленные с помощью ПЛС2 моделей (11.13). Результаты такой оптимизации показаны на Рис. 11.9b. Видно, что гистограмма распределения показателя *y*, предсказанная с помощью $\mathbf{X}^+$ блока, т.е. $\mathbf{y}^+ = \mathrm{XY}(\mathbf{X}^+)$, очень похожа на гистограмму контрольных значений **y**.

Вторая стратегия предполагает более решительные действия по увеличению X-переменных. Она заключается в прибавление некоторой величины к каждой контролируемой переменной. Пусть g_1, g_2, … – это набор некоторых положительных чисел. Тогда воздействие оператора G на каждую переменную z_i из **z**, определим как

$$\mathrm{G}(\hat{z}_\mathrm{i}) = \hat{z}_\mathrm{i} + g_\mathrm{i} \ .$$

Для выбора величин g_1, g_2, …предлагается использовать концепцию ПИО метода, и в частности, использовать величину МП β, т.е. выбирать $g_\mathrm{i} = b_\mathrm{i}$, где b_i это оценка величины β для каждой Z-переменной.

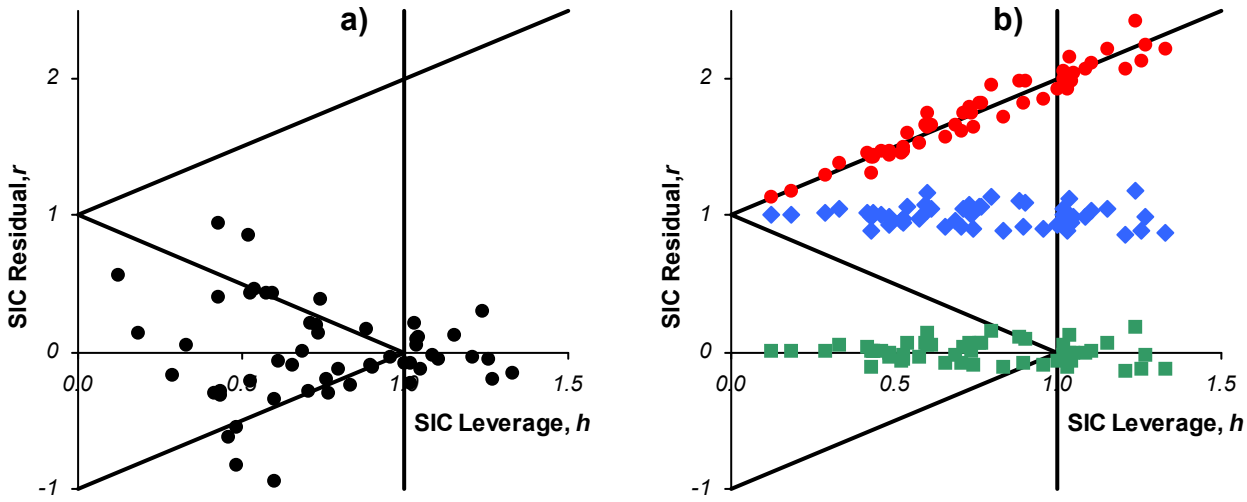
Для рассматриваемого примера, эти значения указаны в Таб. 11.4, колонка *b*. Например, модель $\mathrm{XX_I}$: $\mathbf{X}_{(\mathrm{I})} \Rightarrow \mathbf{X}_\mathrm{I}$ предсказывает второй блок $\mathbf{X}_\mathrm{II}$ (переменные WR1, WR2), который составляет матрицу (102×2), с помощью блока $\mathbf{X}_{(\mathrm{I})}$, состоящего из 6 первых переменных, представленных матрицей (102×6). Для этого используется ПЛС2 модель с шестью ГК, а соответствующие оценки β равны *b*=0.03 и 0.05. Применяя вторую стратегию, значения $\hat{x}_i$ заменяются на значения $x_i^+ = \hat{x}_i + b_i$. Такая оптимизация применяется к каждому образцу из контрольного набора и приводит к ощутимому увеличению показателя *y*, что, например, видно по увеличению среднего значения *M* =0.23–(–0.10)=0.33 относительно контрольного набора.

Следуя ПИО концепции классификации статуса объектов, естественно попробовать и наиболее решительный способ увеличения X-переменных. Это *третий вариант стратегии оптимизации*. В этом случае оператор G выбирается следующим образом

$$\mathrm{G}(\hat{z}_\mathrm{i}) = \hat{z}_\mathrm{i} + b_\mathrm{i}(1 + h_\mathrm{i}) \ ,$$

где z_i – это компоненты вектор строки **z**, h_i – это значения ПИО размаха, а b_i – это оценки МП вычисленные для каждой переменной, относящейся к блоку **Z**. Гистограмма для *y*-переменной, предсказанной с помощью такой оптимизации, представлена на Рис. 11.9d. Виден существенны выигрыш по сравнению с контрольным набором, так как среднее значение стало равным 0.55– (–0.10)= 0.65.

Классификация статуса объектов помогает выбирать различные стратегии оптимизации, а соответствующие ДСО иллюстрируют идею, заложенную в каждую стратегию. Невозможно показать все ДСО для всех этапов моделирования и для всех переменных, так как их будет очень много. Приведем только самые характерные. На Рис. 11.10 представлены ДСО для предсказания значений переменной x_7, (WR1) в ПЛС2 модели XX_I: $\mathbf{X}_I \Rightarrow \mathbf{X}_{II}$. Для вычисления ПИО остатков известные значения контрольных объектов используются в качестве стандартных значений откликов. Видно, что имеется большой разброс между ПИО остатками (Рис. 11.10a), а значения ПИО размаха не превышают 1.5. Всего в контрольном наборе выявлено 23 внутренних и 29 внешних объектов. Это типичная ситуация распределения объектов в проверочном наборе (ср. Рис. 11.5).



Контрольный набор (без оптимизации)

Типы оптимизации:
внутренняя (■), внешняя (◆); выбросы (●)

Рис. 11.10 ДСО для модели XX_I, переменная WR1

На Рис. 11.10b показано как изменяется положение 52 контрольных образов, после применение каждой из трех стратегии оптимизации. ДСО помогает понять каков статус каждого улучшенного объекта: является ли положение объекта странным или обычным

относительно значений исторических данных. В рассматриваемом случае, ПИО остатки вычисляются относительно переменных $\boldsymbol{x}_7^+$, (после оптимизации). Эти значения принимаются за стандартные величины (6.8). Здесь мы опять возвращаемся к проблеме интерпретации выбросов, которые в нашем случае рассматриваются как особые новые объекты, стандартные значения которых равны оптимальным значениям. Применяя теорию классификации статуса объектов, можно понять, насколько значения после оптимизации отличаются от средних, исторически обусловленных значений переменных.

ДСО (квадраты, Рис. 11.10b) показывает, что для *первой стратегии оптимизации* основная часть объектов после оптимизации остается внутренней (35 из 52-х). Поэтому, простейшую оптимизацию с помощью единичного оператора G (11.11) можно назвать оптимизацией «внутреннего» типа. Такая оптимизация не приводит к увеличению значения *y*, так как просто повторяет исторический опыт со всеми его достижениями и промахами.

При применении *второй стратегии* (ромбы на Рис. 11.10b), видно, что все объекты становятся внешними (ср. с Рис. 6.2b). Согласно разделу 6.1, такое положение объектов не противоречит модели. Подобное расположение объектов можно наблюдать и для других X-переменных на соответствующих ДСО. Поэтому, рассматриваемый подход можно назвать оптимизацией по «внешнему» типу.

Идея, лежащая в основе *третьей стратегией оптимизации* становится понятной, если сравнить положение объектов после оптимизации (кружки на Рис. 11.10b) с архетипом для ДСО (Рис. 6.2b). Видно, что все объекты расположились на границе выбросов, поэтому такая оптимизация называется оптимизацией «выбросов».

В результате применения различных стратегий оптимизации к рассматриваемому примеру, получается три набора $\mathbf{X}^+$ переменных, которые, в той или иной степени, улучшают выходное значение в сравнении с контрольными показателями. Эти наборы данных (в соответствии со стратегиями) можно называть: «внутренними», «внешними» и «выбросами». Необходимо отметить, что все эти определения статуса объектов относятся только к моделям вида XX, определенных в (11.13), и не относятся к полной модели процесса XY, задаваемой уравнениями (11.1). Было бы огромным разочарованием, если бы наборы X-переменных после оптимизации не согласовывались бы с общей моделью процесса XY, построенной на основе исторических данных. Это означало бы, что ограничение $\mathbf{z} \in L_z$, заданное условием (11.7) нарушается, т.е. значения переменные $\mathbf{X}^+$ противоречат общей модели процесса. Поэтому необходимо для всех

трех наборов $\mathbf{X}^+$ проверить выполнения условий (11.9), т.е. установить, что полученные оптимальные решения находятся недалеко от центра полной ПЛС модели, а так же, что трансверсальные расстояния до гиперплоскости образованной ГК полной модели также невелики. В Таб. 11.5 представлены распределения ПИО размаха, вычисленные для контрольного набора **X,** и для трех наборов улучшенных X-переменных, $\mathbf{X}^+$. Так же указаны значения для среднего (m_h) и для стандартного отклонения (s_h) для **h**, по каждой из изучаемых стратегий.

Таб. 11.5 Распределение ПИО размаха, **h** для различных стратегий оптимизации. Значения l_i вычисленные по формулам (11.9). В двух последних колонках представлены значения среднего и среднеквадратичного отклонения для **h**.

Стратегия оптимизации	$h \le l_0$	$l_0 < h \le l_1$	$l_1 < h \le l_2$	$l_2 < h \le l_3$	$h > l_3$	m_h	s_h
Контрольные значения	26	19	5	2	0	0.835	0.261
Внутренняя	36	12	3	1	0	0.723	0.277
Внешняя	28	17	5	1	1	0.809	0.305
Выбросы	26	11	10	5	0	0.854	0.385

Такие же результаты, но для значений среднеквадратичных остатков X, **d** (11.8), представлены в Таб. 11.6.

Таб. 11.6. Распределение среднеквадратичных X -остатков, **d** для различных стратегий оптимизации. Значения r_i вычислены по формулам (11.9). В двух последних колонках представлены значения среднего и среднеквадратичного отклонения для **d**.

Optimization strategy	$d \le r_0$	$r_0 < d \le r_1$	$r_1 < d \le r_2$	$r_2 < d \le r_3$	$d > r_3$	m_d	s_d
Control	26	17	9	0	0	0.052	0.031
Insiders	35	10	3	4	0	0.047	0.030
Outsiders	19	25	5	3	0	0.068	0.024
Outliers	0	9	29	10	4	0.105	0.027

Изучая Таб. 11.5 и Таб. 11.6 можно заключить, что оптимальные значения наборов $\mathbf{X}^+$ согласуются с ключевым условием $\mathbf{X}^+ \in L_z$. Средние значения для каждого из трех типов оптимизации так же подтверждают это. Однако, из Таб. 11.6 видно, что

оптимизация типа «выбросы» более «радикальная», чем «внешняя» оптимизация. Поэтому оптимизацию типа «выбросы» нужно применять с большей осторожностью.

Идея классификации статуса объектов дает нам инструмент, с помощью которого можно понять, какая стратегия оптимизации хорошая, какая плохая, и какой увеличивающий оператор G нужно применять в том, или ином случае. Для этого используются соответствующие ДСО, которые строятся на основе ПЛС2 моделей вида XX_M (11.13). Позицию подправленной x_i^+ переменной на ДСО можно трактовать следующим образом:

1. Никакого увеличения y не удается достичь, если переменная располагается во внутренней области модели (область i, Рис. 6.2b)
2. Если мы хотим достичь существенного увеличения значения переменной y, тогда оптимальные значения переменных x_i^+ должны располагаться во внешней области, между внутренней областью, и областью выбросов (область ii, Рис. 6.2b)
3. Значения x_i^+, расположенные в области выбросов (область iii, Рис. 6.2b) могут противоречить накопленным историческим данным о процессе, поэтому подобные значения можно использовать только с большой осторожностью.
4. Обязательной является проверка оптимальных значений $x_i^+ \in L_z$, т.е. их непротиворечивость модели процесса.

Интересно сравнить результаты расширяющегося ПЛС/ПИО моделирования для отобранных объектов, представленных на Рис. 11.6 с аналогичными графиками, полученными при оптимизации, $\mathbf{X}^+$. На Рис. 11.11 представлены результаты применения оптимизации по типу «выбросы». Обозначения на этом графики такие же, как и на Рис. 11.6, с одной лишь разницей, что добавлены белые кружки, которые показывают контрольные значения X–переменных. Видно, что представленная оптимизация улучшает выходной показатель процесса. Так же можно отметить, что значение y, предсказанное на промежуточной стадии II, ведут к ухудшению этого показателя. Такой прогноз получается с помощью промежуточной модели XY_{II}:$\mathbf{X}_{(II)} \Rightarrow \boldsymbol{y}$, которая содержит отрицательные коэффициенты регрессии, соответствующие переменным WR1 и WR2, в то время как у полной модели XY:$\mathbf{X} \Rightarrow \boldsymbol{y}$ все коэффициенты положительны.

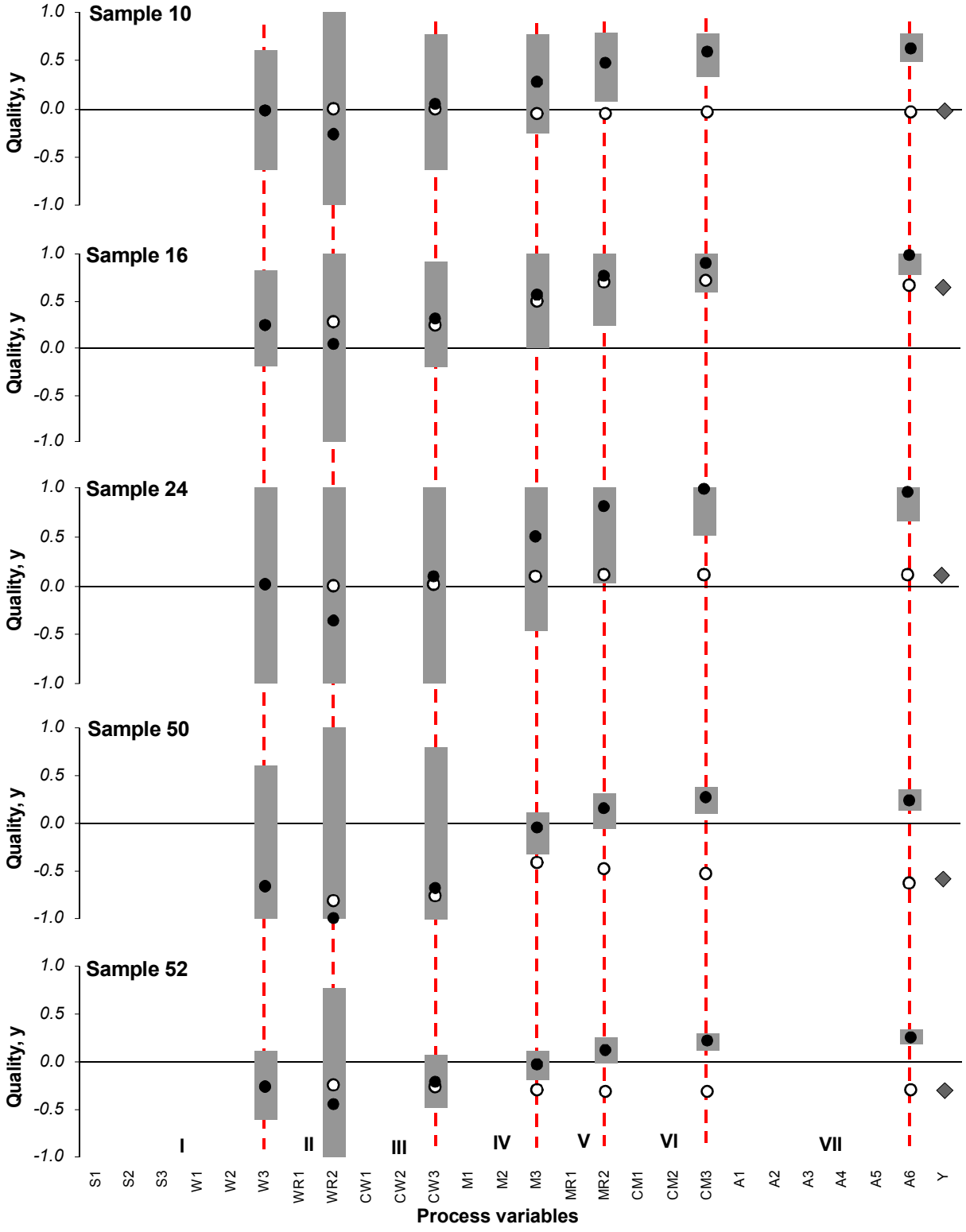


Рис. 11.11 Оптимизация по типу «выбросы». Предсказание выходного показателя на каждой стадии процесса. ПИО интервал (серый прямоугольник), ПЛС прогноз (счерные кружки). Ромб в правой части графиков – контрольные значения *y*, белые кружки – это контрольные значения X-переменных.

В связи с этим можно говорить о двух вариантах оптимизации: оптимизация относительно глобальной модели (11.1) и относительно промежуточной модели (11.3). При этом возможны два различных стиля оптимизации: глобальная и локальная. В первом случае значения всех переменных увеличиваются на каждой стадии, в то время как во втором случае, значения переменных, связанных с отрицательными коэффициентами в промежуточной модели (11.3), наоборот, уменьшаются. Глобальный стиль оптимизации, с использованием полной ПЛС модели (11.1), основывается на дополнительном шкалирование X переменных, описанном в разделе 11.2. Альтернативный, локальный подход опирается на промежуточные модели (11.3), где некоторые из коэффициентов могут быть отрицательными. Локальный подход при оптимизации также был опробован на примере изучаемого процесса, однако окончательный результат очень мало отличался от результата глобальной оптимизации. В данный момент этот эффект объяснить трудно, однако подобный результат, по-видимому, свидетельствует об устойчивости предложенной процедуры оптимизации.

В заключении необходимо отметить, что оптимизацию можно проводить и с иными целями, чем максимизация. Например, целью может быть найти такие значения переменных процесса, которые удерживают значения выходной переменной около среднего, нулевого значения.

11.6. Результаты главы 11

Первым результатом этой главы является разработка нового метода контроля процесса, названного расширяющимся многомерным статистическим контролем, т.е. расширяющимся МСКП. Метод основан на построении серии регрессионных ПЛС моделей, совместно с ПИО моделированием. ПЛС модели позволяют на каждой стадии процесса предсказывать точечные оценки выходного параметра, а ПИО метод добавляет к этой оценки интервал предсказания. Предлагается использовать целый набор ПЛС/ПИО моделей (11.3) для осуществления контроля над процессом. Представленный подход помогает предсказать результат планируемых действий по улучшению выходного показателя y на промежуточных стадиях процесса и осуществить пассивную оптимизацию. Такой подход может быть особенно важным в случае, когда длительность контролируемого процесса велика.

Вторым результатом этой главы является разработка подхода к активной оптимизации процесса, который основывается на блочном ПЛС и ПИО методе.

Многочисленные исследования указывают, что для улучшения выходного показателя *y* необходимы корректирующие действия которые, с одной стороны, остаются в рамках исследуемого процесса, а, с другой стороны, требуют вывода значений контролируемых переменных на границы возможных значений. Предложенный подход предполагает действия по регулированию контролируемых переменных на промежуточных стадиях, и предлагает набор стратегий таких действий. Активная оптимизация основывается на теории классификации статуса объектов, которая, в свою очередь, является неотъемлемой частью ПИО метода. Такой подход близок к традиционному многомерному статистическому контролю (МСКП), так как он так же опирается на накопленные данные о процессе, которые используются для моделирования. Поэтому представленный подход предлагается назвать многомерной статистической оптимизацией процессов (МСОП).

Существенным является способ формирования исторического набора данных. Эти данные служат основой для построения регрессионных моделей, формирования ограничений на корректирующие действия, и, в конечно счете, влияют на окончательный результат оптимизации. В рассмотренном примере отбор данных происходил следующем образом. Все случаи нетипичных или случайных реализации процесса не включались в набор, т.е., при формировании исторического набора производились обычные действия по исключению выбросов, однако, этот аспект требует дополнительного изучения.

12. Формирование представительной выборки объектов применительно к различным наборам многоканальных экспериментов

На практике, при решении задачи переноса калибровок с одного прибора на другой, при работе с очень большими наборами данных физических экспериментов, и в других случая возникает потребность выбрать из общего набора накопленных данных (**X**,**Y**), представительный более короткий набор. Необходимо, чтобы такая выборка отвечала двум основным требованиям: во-первых, она должна представлять всю вариабельность данных общего набора; во-вторых, число объектов в такой выборке должно быть существенно меньше, чем в полном наборе. Разработаны различные способы формирования такой выборки [116, 209, 132, 291] работающие более или менее эффективно в различных ситуациях. В дополнение к этому арсеналу методов, предлагается применить метод ПИО, в частности использовать классификацию статуса объектов. В этой главе подробно рассматривается, как применить ПИО подход для формирования представительной выборки [52], а также проводится сравнение полученных результатов с двумя наиболее известными и эффективными методами отбора представительных объектов: с методом Кеннарда-Стоуна (Kennard-Stone) [292] и D-оптимальным планированием [293].

Сам по себе термин «представительная» или «репрезентативная» выборка является неоднозначным и может трактоваться по-разному, в зависимости от поставленной задачи. В нашем случае, целью является выбор из имеющегося калибровочного набора данных наиболее влиятельных объектов, использование этих объектов в качестве нового калибровочного набора и на их основе построение калибровки, которая по своей предсказательной способности не уступала бы модели, построенной на полном калибровочного наборе. Такую представительную выборку можно назвать влиятельным набором.

При такой постановке задачи с необходимостью возникает вопрос оценки предсказательной способности модели. В отличие от традиционных моделей множественной регрессии, когда неопределенность в прогнозе характеризуется доверительными интервалами, в проекционных регрессионных моделях общепринятого критерия не существует [119]. Наиболее популярным является использование в качестве критерия среднеквадратичного остатка предсказания (RMSEP (6.2)), вычисленного на независимом проверочном наборе [63, 67]. Однако эту оценку необходимо применять с

осторожностью, так как величина RMSEP сильно зависит от «качества» этого проверочного набора. Например, если все объекты проверочного набора располагаются в близи центра калибровочной выборки, то RMSEP будет маленьким, и оценка предсказательной способности модели будет излишне оптимистичной.

Существенным недостатком критериев, основанных на оценке RMSEP, является то, что этот показатель оценивает модель только *в среднем*, и не является характеристикой неопределенности в прогнозе для каждого отдельного объекта. Поэтому RMSEP – это *необходимая, но недостаточная* характеристика, что вызывает потребность в более тщательном сравнении моделей. Для этого используются как «традиционный» для многомерной калибровки график влиятельности, который уже использовался в разделе 8.2, так и его аналог в ПИО методе – диаграмма статуса объектов (ДСО), представленная на Рис. 6.2b.

Основные результаты этой главы опубликованы в [52, 294]

12.1. Теория

Рассматривается регрессионная задача

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon}, \qquad (12.1)$$

где **X –** это ($I \times J$) матрица предикторов; **y** – это вектор откликов размерности I; **a** – это J-мерный вектор неизвестных параметров; **ε** – неизвестный вектор ошибок. Ранг матрицы **X** меньше J. Имея достаточно большой калибровочный набор (**X,y**) можно оценить неизвестные параметры **a** и в дальнейшем использовать модель (12.1) для предсказания отклика y для новых значений вектора **x**.

Нашей цель – является выделить из полного калибровочного набора (**X,y**) такой короткий влиятельный набор объектов (представительную выборку), который так же может быть использован для оценки параметров **a.** При этом точность предсказания y не должна заметно ухудшиться.

Метод граничных объектов

В соответствии с методом ПИО, все калибровочные объекты являются внутренними (Определение 6.1) относительно модели. Кроме того, среди этих объектов выделяются такие, которые играют специальную роль в калибровке. Такие объекты называются граничными (Определение 6.2), и они оказывают существенное влияние, так как ОДЗ A (5.4) в ПИО модели формируется не всеми объектами калибровочного набора,

а только граничными. Отсюда следует, что если все объекты, кроме граничных, убрать из калибровочного набора, то ОДЗ не изменится. А это означает, что граничные объекты являются наиболее влиятельными, по крайней мере, с точки зрения ПИО моделировании. Именно они и формируют представительную выборку по методу граничных объектов.

Метод Кеннарда-Стоуна

Метод Кеннарда-Стоуна [292] позволяет выбрать из исходного набора объектов, представительный поднабор, причем выбор осуществляется «равномерно» по всей области, где распределены объекты. При этом рассматриваются только данные в матрице **X**, значения **y** не учитываются. Это классический и широко используемый метод [209, 291, 295] выбирает объекты последовательно, один за другим. Первые два объекта выбираются так, чтобы между ними было максимальное расстояние из всех имеющихся объектов. Следующий объект, включаемый в выборку, это такой объект, который находится на максимальном расстоянии от первых двух отобранных объектов, и т.д. Предположим, что уже отобрано p объектов ($p<I$), тогда (p+1)-ый объект выбирается по критерию

$$\max_{p<r\le I}\left(\min\left(d_{1r},d_{2r},...d_{pr}\right)\right), \qquad (12.2)$$

где d_{lr} , l=1,...p, это квадрат Евклидова расстояния от объекта r из калибровочного набора, который еще не отобран, до всех p уже отобранных объектов. К достоинствам алгоритма надо отнести, во-первых, простоту реализации, и, во-вторых, то, что алгоритм может применяться к любой матрице предикторов, независимо от того, имеет она полный ранг или нет.

D- оптимальный план

Этот метод часто используется на практике[209, 296, 297]. D-оптимальный план выбирает объекты так, чтобы максимизировать определитель информационной матрицы $|\mathbf{X}^t\mathbf{X}|$ линейной регрессионной модели. Это итерационная процедура, которая начинается с некоторого произвольного начального плана, т.е. начального набора объектов. Количество объектов задается исследователем априори. На каждой итерации, каждый объект, содержащийся в плане, сравнивается с образцом из списка кандидатов. Если при таком сравнении выясняется, что объект из списка кандидатов оптимизирует текущий план, то он заменяет сравниваемый объект. Процедура заканчивается тогда, когда после

сравнения всех объектов из списка кандидатов, не достигается улучшения в смысле оптимизационного критерия [295]. Определитель информационной матрицы достигает максимального значения тогда, когда объекты, включенные в план, имеют максимальную вариацию в исследуемом наборе. Таким образом, по сравнению с процедурой Кеннарда-Стоуна, D-оптимальный план выбирает объекты наиболее удаленные от среднего значения. Когда число переменных, характеризующих каждый объект, велико и превышает число объектов, информационная матрица становится вырожденной и процедуру D-оптимального планирования нельзя применить непосредственно, а только после регуляризации задачи.

Обозначения и план исследования

Представленные выше методы формирования представительной выборки применялись к трем различным наборам экспериментальных данных. Первый экспериментальный набор, это определение влажности в зерне (описание эксперимента см. в разделе 12.2). Для того, чтобы продемонстрировать различные аспекты выбора влиятельного набора и свойства получившихся моделей, этот пример рассматривается детально. Вторая задача – это определение следовых концентраций нефти в воде (см. раздел 8.1). Этот пример показывает возможности различных методов отбора в случае, когда приходится иметь дело с достаточно небольшим набором данных и простой моделью, построенной всего на двух ГК. Третья задача – это моделирование многостадийного технологического процесса (см. раздел.11.1). Этот пример позволяет показать влияние объема представительной выборки на точность предсказания.

При рассмотрении каждого из этих примеров используются различные наборы и поднаборы объектов. Для того чтобы избежать путаницы и сделать изложение более понятным, будут использованы следующие общие обозначения. G (general) набор – это исходный набор объектов соответствующего примера. G набор делится на C (calibration) набор – калибровочный набор и T (test) набор – проверочный набор. Таким образом, C набор + T набор=G набор. Количество объектов в этих наборах обозначается как I_G, I_C и I_T, соответственно. ПЛС-модель, построенная с помощью калибровочного C набора проверяется на T наборе, такая модель обозначается Модель_C. Число главных компонент в такой модели фиксируется и сохраняется для всех последующих моделей, построенных на меньшем количестве объектов. К C набору по очереди применяется каждый из трех выше описанных способов формирования представительной выборки. В

результате образуются наборы: B (boundary) набор, K (Kennard_Stone) набор и D (D-optimal) набор (соответственно: граничный, Кеннарда-Стоуна, D –оптимальный) в зависимости от применяемого метода. При этом объекты анализируются не в исходном X–пространстве переменных, а в пространстве счетов, образованном главными компонентами Модели_C. Например, в методе Кеннарда-Стоуна расстояния ищутся не между векторами $\mathbf{x}_i$, а между их проекциями, т.е. векторами $\mathbf{t}_i$. Обозначения I_B, I_K и I_D используются для числа объектов в каждом из этих наборов.

Оставшиеся после выбора влиятельного набора объекты в C наборе называются «избыточными» (redundant). Они составляют наборы обозначенные: RB, либо RK, либо RD наборы, в зависимости от того, какой алгоритм выбора был применен. Очевидно, что каждый представительный набор, например B набор, совместно с соответствующим ему избыточным набором, в данном случае RB набором, составляют калибровочный C набор, т.е. B набор + RB набор= C набор. Калибровки, построенные на основе B, K, или D наборов, и проверенные с помощью T набора, называются Модель_B, Модель_K, и Модель_D соответственно. Каждая из этих моделей использует одно и тоже число ГК, определенное Моделью_C, однако каждая модель формирует свое проекционное ПЛС-подпространство. Значения среднеквадратичных остатков калибровки (RMSEC (3.1)), вычисленные для каждой модели, обозначаются RMSEC_C, RMSEC_B, RMSEC_K и RMSEC_D. Они вычисляются с использованием соответствующих наборов данных, например,

$$\text{RMSEC\_C} = \sqrt{\frac{1}{I_C - K}\sum_{i=1}^{I_C}\left(y_i - \hat{y}_i\right)^2}$$

где y_i – это измеренные значения отклика, $\hat{y}_i$ - значения, вычисленные с помощью калибровки на C наборе, K – число ГК. Аналогично, среднеквадратичные остатки предсказания (RMSEP (3.2)) для каждой из моделей, обозначаются RMSEP_C, RMSEP_B, RMSEP_K и RMSEP_D. Все эти значения вычисляются с помощью одного и того же T набора, но с использованием соответствующей модели, например,

$$\text{RMSEP\_C} = \sqrt{\frac{1}{I_T}\sum_{i=1}^{I_T}\left(y_i - \hat{y}_i^C\right)^2}$$

где y_i – измеренные значения из T набора, а $\hat{y}_i^C$ – соответствующие значения отклика, предсказанные с помощью Модели_C. Схематически все исследуемые наборы данных и модели изображены на Рис. 12.1.

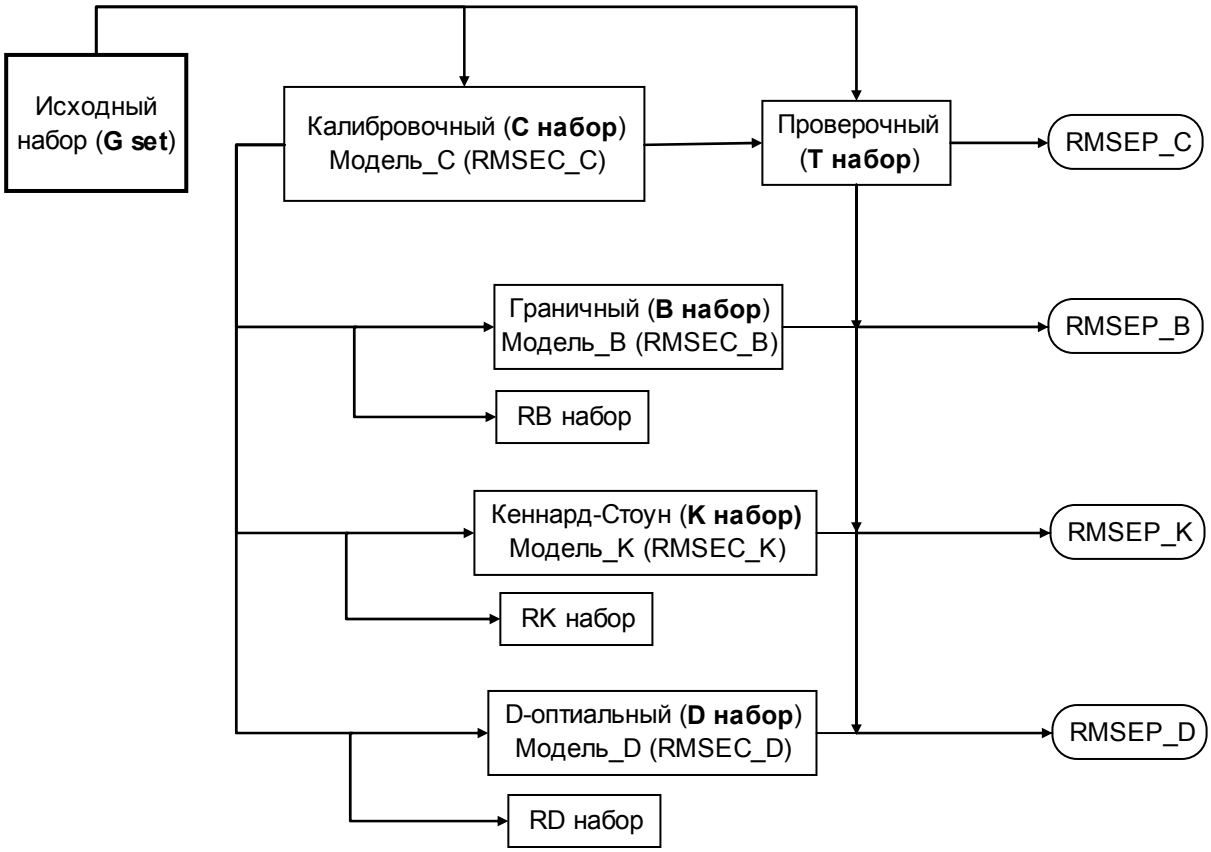


Рис. 12.1 Исследуемые наборы и соответствующие им модели.

Для каждого из рассматриваемых примеров строятся также несколько ПИО моделей, на основе соответствующих ПЛС моделей: Модель_C Модель_B и т.д. Однако для каждой из этих ПИО моделей используется одна и та же оценка величины β равная b_{SIC}. Эта оценка вычисляется для Модель_C и не меняется для других моделей, так же как и число главных компонент для ПЛС моделей.

Для того чтобы проверить, что выбранный короткий набор объектов Q (Q может быть B, K или D) действительно является представительным, выполняется следующая процедура:

1. Строится ПЛС модель, Модель_Q, на основе Q набора, с фиксированным число ГК, и соответствующая ПИО модель, с фиксированным значение b_{SIC}.
2. Модель_Q проверяется с помощью проверочного T набора.
3. Модель_Q используется для предсказания объектов из избыточного набора RQ. (12.3)
4. Результаты калибровки и предсказания сравниваются с результатами, полученными для Модели_C.

Для сравнения различных наборов и моделей используются несколько показателей: ПИО остатки, вычисляемые по формуле (6.8) и ПЛС остатки для Y-переменных $r_{\text{PLS}} = y - \hat{y}$; ПИО размах, вычисляемый по формуле (6.9) и ПЛС размах, вычисляемый по стандартной формуле (2.6)). В каждом конкретном случае будет понятно, для какой модели (C, B, K или D) вычислялась та или иная характеристика.

На всех рисунках этой главы (Рис. 12.3 - Рис. 12.7), будут использованы одни и те же обозначения для объектов. Все объекты из калибровочного набора, независимо от их последующей роли в различных моделях, обозначаются кружками. В соответствии с ПИО классификацией, закрашенные кружочки обозначаю граничные объекты, а незакрашенные – внутренние калибровочные объекты. Все объекты из проверочного набора обозначаются квадратами, а объекты из избыточного набора – треугольниками. Заштрихованные значки, выделяют некоторые особо важные объекты, которые анализируются в дальнейшем в тексте.

12.2. Эксперимент 1. Определение влажности зерна с помощью инфракрасной спектроскопии в ближней области.

Целью эксперимента является построение калибровки для определения процентного содержания влаги в пшенице. Для этого было собрано 142 объекта пшеницы. Влажность измерялась в лаборатории с помощью стандартного метода высушивания навесок. Результаты измерений использовались в качестве опорных значений и составили вектор откликов **y** размерности (142×1). Далее, для каждого из объектов три раза снимался БИК спектр в режиме пропускания в диапазоне 908–1120 nm. В качестве X-данных использовались результаты БИК-спектроскопии. При этом результаты трех измерений для каждого объекта усреднялись. Таим образом, исходная матрица X состояла из 142 объектов (строк), и значений интенсивности, измеренных на

118 длинах волн, являющихся переменными. Результаты измерений, % пропускания, пересчитывались в единицы поглощения [69].

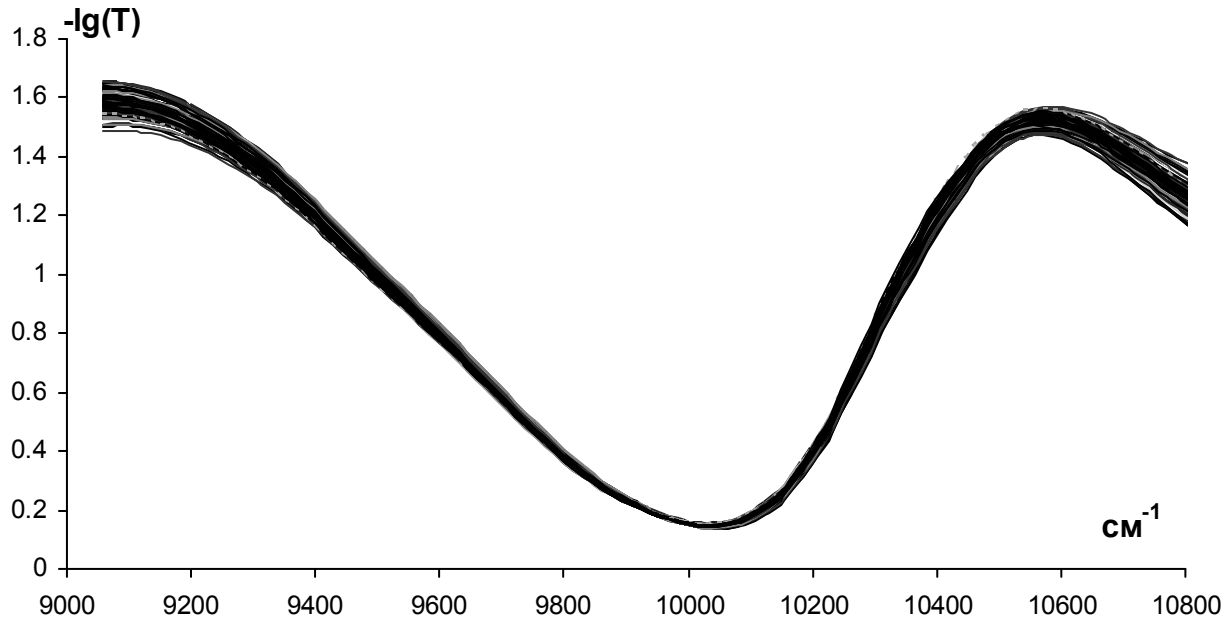


Рис. 12.2 Определение влажности зерна. Спектры поглощения.

Для предварительной обработки применялось центрирование и шкалирование на единичную дисперсию. Спектры исследуемых объектов после таких преобразований представлены на Рис. 12.2. Предварительный анализ выявил три объекта, которые можно считать выбросами (Рис. 4.1). После удаления выбросов (Рис. 4.2) набор данных составил 139 образцов, т.е. для калибровки использовалась матрица предикторов **X** размерности (139×118) и вектор откликов **y** размерности (139×1). Эти данные образуют исходный G набор. На его основе была построена ПЛС модель с четырьмя главными компонентами. Основные характеристики модели оценивались с помощью 10% перекрестной проверки (так же как в разделе 11.3). Результаты ПЛС моделирования в зависимости от числа ГК представлены в Таб. 12.1

В первой колонке Таб. 12.1 указано число главных компонент, используемых на каждом шагу проецирования. Во второй и третьей - какая часть изменений в **X** и **y** описывается при этом числе ГК. В последней колонке представлен коэффициент корреляции между вектором X-счетов $\mathbf{t}_k$ и соответствующим ему вектором Y-счетов $\mathbf{u}_k$ (11.5). Видна существенная корреляция между векторами $\mathbf{t}_k$ и $\mathbf{u}_k$ для первых четырех ГК.

Таб. 12.1. Определение влажности зерна. ПЛС модель для исходного набора (G набора) 139 образцов

ГК	Cum(X)	Cum(Y)	RMSEC	RMSEP (CV)	r_{tu}
1	31.84	51.42	0.622	0.638	0.72
2	86.86	61.93	0.551	0.565	0.47
3	95.58	81.51	0.384	0.400	0.72
4	**99.50**	**89.65**	**0.288**	**0.299**	**0.66**
5	99.70	89.81	0.285	0.301	0.13
6	99.77	90.11	0.281	0.300	0.17

В дальнейшем, для разделения G набора на калибровочный и проверочный, а так же для выбора коротких поднаборов, используются проекции значений **X** на ПЛС подпространство образованное четырьмя главными компонентами. Известно [116], что результат сравнения наборов данных зависит не только от данных самих по себе, но и от рабочего подпространства переменных, на котором это сравнение осуществляется, что, в свою очередь определяется моделью, например, числом ГК.

Используя ПЛС модель, построенную на G наборе, строим соответствующую ПИО модель, и оцениваем МП β, определяя две оценки b_{min}=1.03, b_{SIC}=1.5. Результаты этого моделирования: число ГК и оценки β, используются в дальнейшем при построении моделей на основе различных выборок из G набора, а также для анализа результатов.

12.3. Анализ данных на основе калибровочного и проверочного наборов - Модель_ C

Используя прием случайного выбора, разделяем исходный набор на калибровочный – C набор, состоящий из 99 объектов, и проверочный – T набор, состоящий из 40 объектов. Как говорилось в начале этой главы, качество проверочного набора относительно калибровочного набора и самой модели, существенно для правильной оценки предсказательных свойств модели. Желательно, чтобы калибровочный и проверочный наборы были очень похожи, тогда можно будет положиться на результаты такой проверки модели. Для сравнения двух наборов используем классические статистические методы. Во-первых, это обобщенный тест Бартлетта (Bartlett), который позволяет сравнить ковариационные матриц двух наборов. При этом вычисляются две ковариационные матрицы $\mathbf{S}_1$ и $\mathbf{S}_2$, по формулам

$$\mathbf{S}_1 = \frac{1}{I_1 - 1}\mathbf{X}_{1c}^{t}\mathbf{X}_{1c}$$

$$\mathbf{S}_2 = \frac{1}{I_2 - 1}\mathbf{X}_{2c}^t \mathbf{X}_{2c}$$

В данном случае матрицы $\mathbf{X}_1$ и $\mathbf{X}_2$ представляют два сравниваемых набора содержащих I_1 и I_2 объектов соответственно. Каждый объект является вектором-строкой, размерности $(1\times J)$. Матрицы $\mathbf{X}_{1c}$ и $\mathbf{X}_{2c}$ – центрированные по столбцам исходные матрицы.

Кроме того, необходимо вычислить еще и объединенную ковариационную матрицу **S,**

$$\mathbf{S} = \frac{(I_1 - 1)\mathbf{S}_1 + (I_2 - 1)\mathbf{S}_2}{I_1 - 1 + I_2 - 1} \tag{12.4}$$

Тогда величина C

$$C = 1/\nu \left[(I_1 + I_2 - 2)\ln|\mathbf{S}| - (I_1 - 1)\ln|\mathbf{S}_1| - (I_2 - 1)\ln|\mathbf{S}_2| \right]$$

$$\nu = 1 + \frac{2J^2 + 3J - 1}{6(J+1)}\left(\frac{1}{I_1 - 1} + \frac{1}{I_2 - 1} + \frac{1}{I_1 + I_2 - 2} \right)$$

распределена по χ^2 с $J(J+1)/2$ степенями свободы. Если статистика C меньше критического значения C_{crit}, то можно заключить, что ковариационные матрицы двух наборов статистически эквивалентны. Таким образом, тест Бартлетта позволяет сравнить ориентацию точек в пространстве, а так же дисперсию точек вокруг их средних значений.

Однако одного теста Бартлетта недостаточно. Дополнительно необходимо проверить, что скопление точек одного и другого наборов расположены примерно в одном и том же месте пространства. Для этого используется T^2 тест Хотеллинга (Hotelling) который проверяет, что расположение центров двух выборок в пространстве совпадает со статистической точки зрения. Расстояние между центрами двух наборов в пространстве можно вычислить с помощью формулы Махаланобиса

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{12.5}$$

где $\bar{\mathbf{x}}_1$ и $\bar{\mathbf{x}}_2$ векторы средних значений, а **S** – объединенная ковариационная матрица (12.4). Оценка (12.5) справедлива только в том случае, когда у двух выборок равны ковариационные матрицы. В нашем случае, это проверяется предыдущим тестом.

Тогда величина

$$F = \frac{I_1 I_2 (I_1 + I_2 - J - 1)}{J(I_1 + I_2)(I_1 + I_2 - 2)} D^2$$

подчиняется F-распределению с J и I_1+I_2-J-1 степенями свободы. Если вычисленное значение F меньше критического $F_{crit,}$, то расстояние между центрами двух выборок можно принять равным нулю. Подробное описание применения этих тестов для сравнения двух выборок описано в [116].

В нашем случае статистическое сравнение калибровочного и проверочного наборов дало следующие результаты: C=12.3 (C_{crit}=18.4) для теста Бартлетта и F=0.99 (F_{crit}=2.44) для T^2 теста Хотеллинга. Таким образом, так как калибровочный и проверочный наборы статистически одинаковы, можно полагаться на результаты проверки модели и оценки ее предсказательной способности с помощью проверочного набора.

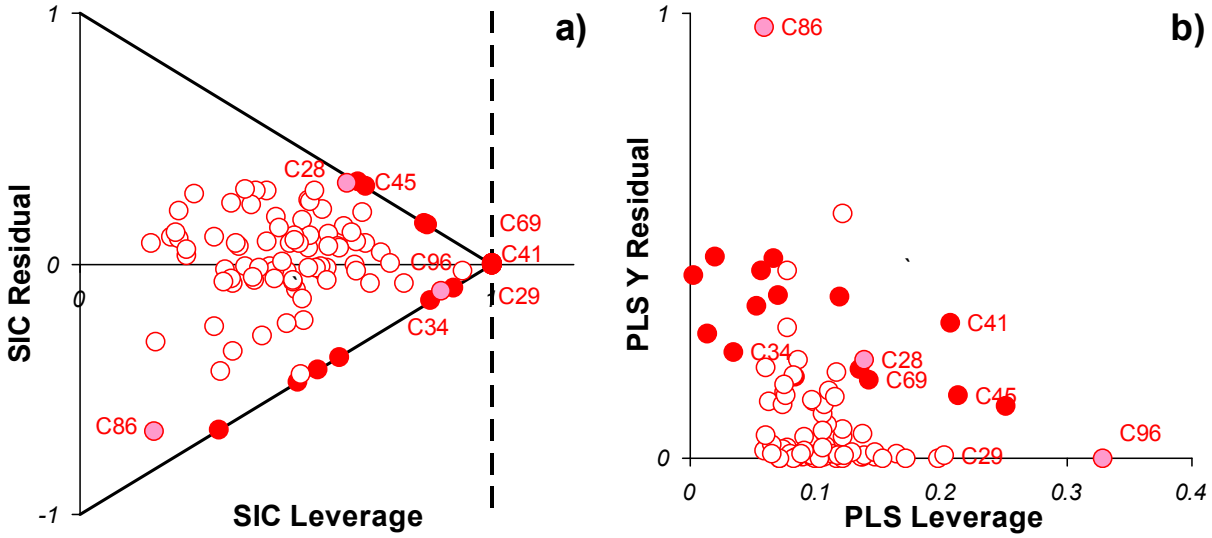
Для ПЛС модели, построенной на основе C набора, требуется четыре ГК, также как и для модели построенной на G наборе. Сводные данные результатов ПЛС регрессии приведены в Таб. 12.2, и они аналогичны результатам Таб. 12.1.

Таб. 12.2. Определение влажности зерна. ПЛС Модель_C

ГК	Cum(X)	Cum(Y)	RMSEC	RMSEP (CV)	r_{tu}
1	30.15	51.80	0.639	0.504	0.72
2	86.76	62.83	0.561	0.496	0.48
3	95.44	82.98	0.379	0.354	0.74
4	**99.45**	**90.74**	**0.280**	**0.313**	**0.63**
5	99.70	90.94	0.277	0.311	0.13
6	99.74	91.49	0.268	0.313	0.20

В данном случае, величина RMSEP вычислялась с помощью независимого проверочного T набора, который не использовался для построения калибровочной зависимости. Используя внешний проверочный набор, мы как бы имитируем результаты предсказания новых объектов, которые будут использовать в будущем, и которые не входят в калибровочный набор. Проверка с помощью независимого проверочного набора считается более надежной [71], чем перекрестная проверка, при условии представительности проверочного набора. В соответствии с обозначениями, введенными в разделе 12.1, полученная ПЛС модель называется Моделью_C.

На основе Модели_C, строим ПИО модель и соответствующую ей ДСО (Рис. 12.3а).



ПИО диаграмма статуса объектов

График влиятельности

Рис. 12.3. Определение влажности зерна. Модель_C на основе 4-х главных компонент. Калибровочный набор: ○-внутренние объекты, ● -граничные объекты

Следуя методу классификации статуса объектов (6.12), девятнадцать образцов из калибровочного набора, которые расположены на границе треугольника внутренних объектов, являются граничными. Далее, все объекты из калибровочного C набора обозначаются C1, C2,..., C99. Все объекты, принадлежащие проверочному T набору, обозначаются как T1, T2,..., T40.

Сравнивая ДСО (Рис. 12.3a) и традиционный график влиятельности [67] (Рис. 12.3b), видно, что все объекты, которые определяются как наиболее влиятельные на графике Рис. 12.3b, в то же время являются граничными с точки зрения ПИО классификации. Например, граничные объекты C29, C41, и C69 трактуются как влиятельные объекты обоими графиками. Если объекты очень похожи, то только один из них входит в набор граничных объектов. Так, влиятельный объект C28 классифицируется как внутренний, но при этом он расположен очень близко к границе треугольника и к образцу C45. Подобная же ситуация наблюдается и с объектами C96 (внутренний) и C34 (граничный). На графике влиятельности (Рис. 12.3b) эти объекты так же расположены очень близко друг к другу. Таким образом, теория ПИО классификации объектов позволяет выявить наиболее важные объекты в калибровочном наборе (Опр. 6.1-6.2).

Сравнение ДСО и графика влиятельности (Рис. 12.3) показывает, что концепция граничных объектов имеет смысл не только в пределах ПИО метода, но она

характеризует структуру исследуемых данных и может быть полезной для формирования выборки представительных объектов.

12.4. Граничная выборка, Модель_B

Согласно ПИО классификации, граничные объекты являются наиболее важными объектами, так как они формируют ОДЗ *A*. Остальные объекты калибровочного набора можно рассматривать как «избыточные». Необходимо понимать, что эта «избыточность» является условной и ее надо трактовать только в определенном смысле, т.е. эти объекты не оказывают влияния на размер области *A*.

Разделим C набор, на набор граничных объектов (B набор, I_B=19) и «избыточный» набор (RB набор, I_{RB} =80). RB набор формируется из тех объектов калибровочного C набора, которые не были включены в граничный B набор. В результате такого разделения, предполагаем, что, во-первых, B набор можно использовать вместо всего калибровочного C набора для построения калибровки, и, во-вторых, объекты из избыточного RB набора надежно предсказываются с помощью модели построенной на основе B набора.

Таб. 12.3. Определение влажности зерна. Сводная таблица. Предсказательные свойства ПЛС моделей с 4 ГК.

Модель	Обучающий набор	RMSEC	Проверочный набор	RMSEP
Модель _C	C набор	0.280	T набор	0.312
Модель_B	B набор	0.426	T набор	0.330
			RB набор	0.246
Модель_K	K набор	0.212	T набор	0.339
			RK набор	0.311
Модель_D	D набор	0.266	T набор	0.333
			RD набор	0.305

Для того чтобы показать, что B набор действительно является представительной выборкой, воспользуемся процедурой (12.3). Из Таб. 12.3 (ряды 1 и 2) видно, что точность калибровки (RMSEC_B) для Модели_B значительно хуже чем для Модели_C. Это легко объяснимо, так как для построения калибровочной зависимости в B набор были отобраны самые удаленные от центра объекты, а все «средние», типичные объекты были исключены. С другой стороны, если посмотреть на точность предсказания

(RMSEP) то она ухудшилась незначительно. Но это сравнение моделей "в среднем". Однако интересно также проанализировать неопределенность в прогнозе отдельно для каждого объекта из проверочного Т набора.

ПИО классификация объектов из проверочного Т набора одинакова, как для ПИО Модели_С, так и для ПИО Модели_В. Диаграмма статуса объектов (ДСО) приведена на Рис. 12.4. (ДСО, построенная на основе Модели_С выглядит также, поэтому здесь не приводится). Величина ПИО размаха для каждого из объектов Т набора, вычисленная с помощью Модели_С, совпадает с соответствующей величиной, вычисленной с помощью Модели_В. Это означает, что ширина интервала предсказания для каждого из проверочных объектов вычисленная по обеим моделям совпадает.

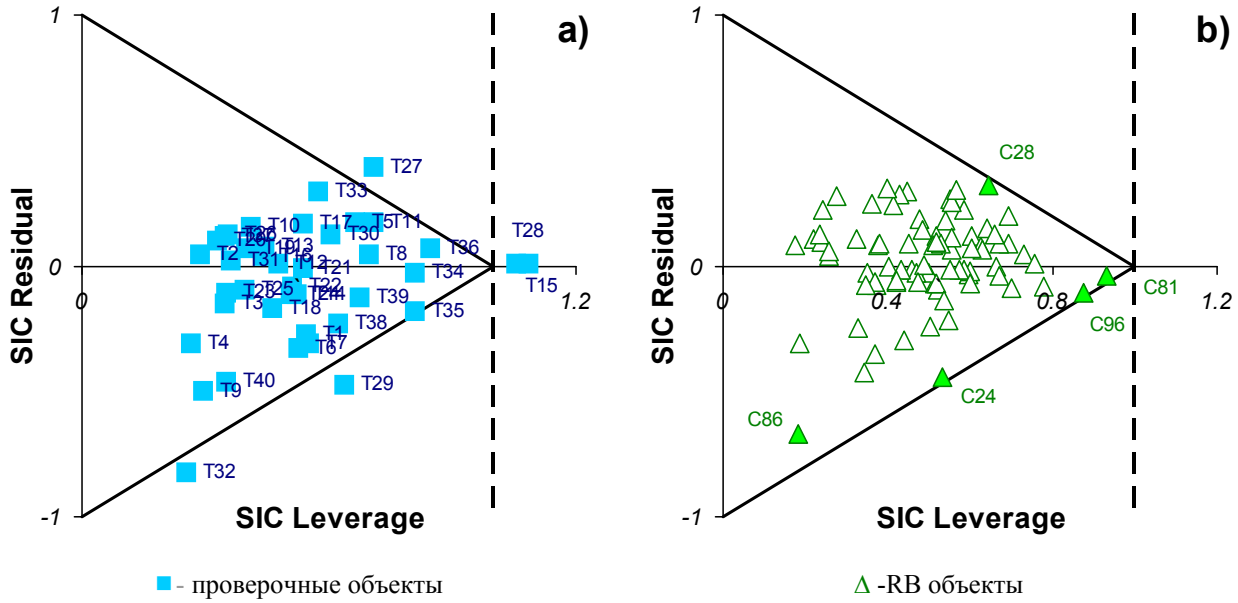
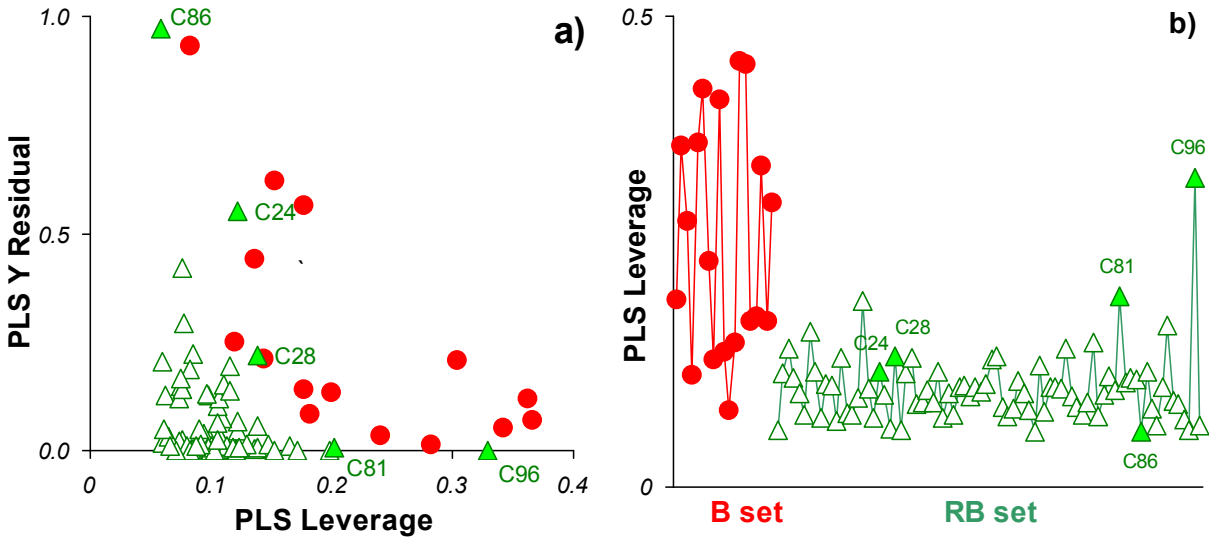


Рис. 12.4 Определение влажности зерна. ДСО для ПИО Модели_В. Предсказание проверочного Т и избыточного RB наборов

Несколько внешних объектов Т15, Т27, Т28, Т29, Т32 на ДСО показывают, что проверочный набор несколько отличается от В набора. В общем случае, наличие в проверочном наборе объектов, которые несильно, но отличаются от объектов калибровочного набора, важно для оценки устойчивости модели и более надежной оценки ее предсказательной способности.

Теперь рассмотрим как Модель_В предсказывает внутренние объекты, т.е. объекты, собранные в RB наборе. Если предположения относительно граничных образов справедливы, тогда RB набор должны составлять самые надежные «средние» объекты.

Сначала рассмотрим результат ПЛС предсказания. На Рис. 12.5a представлена зависимость ПЛС Y остатков, относительно ПЛС размаха. Видно, что граничные объекты (закрашенные кружки) являются наиболее влиятельными объектами. Одновременно, можно отметить, что значения ПЛС размаха для граничных объектов мало отличаются друг от друга. Мы не наблюдаем объектов с очень большой или очень маленькой величиной размаха (Рис. 12.5b). Это означает, что все объекты B набора играют в построении модели примерно одинаковую роль.



ПЛС Y остатки относительно ПЛС Y размаха

ПЛС Y размах для B и RB наборов

Рис. 12.5. Определение влажности зерна. ПЛС Модель_B с 4 ГК. ● -B набор, Δ- RB набор

Что касается тестовых объектов (из набора RB), то все они имеют достаточно маленький размах, а остатки колеблются от малых, до средних значений. Только у объекта C86 достаточно большой ПЛС остаток (Рис. 12.5a), но его ПЛС размах самый маленький, по сравнению со всем другими объектами.

Таким образом, можно заключить, что все объекты из RB набора расположены близко к центру модели и прогноз на них очень надежен. Это так же подтверждается результатами Таб. 12.3, строка 3. Также и ДСО (Рис. 12.4b) классифицирует все объекты из RB набора как внутренние.

Интересно сравнить расположение некоторых объектов на графике влиятельности (Рис. 12.5a), и на ДСО (Рис. 12.4b). У объекта C86 большой ПЛС остаток (Рис. 12.5a), и этот же можно сказать про ПИО остаток этого объекта (Рис. 12.4b). При этом его ПИО

размах мал, именно поэтому он и остается внутренним образцом. Одновременно, объект C86 имеет очень маленький ПЛС размах.

У объекта C96 наибольший ПЛС размах из всех объектов RB набора (Рис. 12.5b). Одновременно у этого объекта и наибольший ПИО размах (Рис. 12.4b), объект расположен вблизи границы внутренних объектов, и даже недалеко от границы абсолютно внешних объектов.

Таким образом, можно констатировать, что мы достигли поставленных целей:

(1) модель, построенная на основе граничных объектов, предсказывает проверочные объекты лишь немного хуже, чем модель, построенная на полном обучающем наборе;

(2) высокая точность предсказания объектов из RB набора подтверждает их «избыточность» для построения калибровки;

(3) граничный набор действительно является представительной выборкой, при этом его размер существенно меньше, чем калибровочный набор, а именно, 19 из 99 объектов.

Необходимо отметить, что значение RMSEP, вычисленное на RB наборе (Таб. 12.3) достаточно мало. Это легко объяснить, используя концепцию граничных объектов. После отбора наиболее важных (граничных) объектов из калибровочного набора, объекты которые остаются в этом наборе, являются наиболее типичными, «средними» объектами. Именно поэтому оценка RMSEP для них такая хорошая. Это соображение необходимо принимать во внимание при разделении исходного набора данных на калибровочный и проверочный наборы. Если проверочный набор будет состоять только из «хороших» объектов, величина RMSEP будет маленькой, и, следовательно, предсказательная способность модели будет сильно завышена.

На практике бывают ситуации, когда исследователь специально собирает калибровочный набор, как более широкий, включая в него некоторые экстремальные объекты, а проверяет калибровку на более узком проверочном наборе, считая, что в дальнейшем, на практике, разброс объектов будет примерно такой же, как в проверочном наборе. Это может придать большую устойчивость построенной калибровки. Однако, в общем случае, особенно для сравнения различных методов построения моделей, такой прием неприемлем. Например, в работе [291] исследователи предложили новый последовательный проекционный алгоритм (SPA) для отбора представительных объектов, проверяя свой метод с помощью независимого проверочного набора, вычисляя только значения RMSEP. При этом, исходный набор делился на калибровочный и

проверочный таким образом, что значения отклика в этих массивах были распределены равномерно во всем интервале изменения y. В тоже время, все объекты, которые при моделировании обнаруживали большие значения X-остатков или y-остатков, были включены в калибровочный набор. Понятно, что значение RMSEP вычисленное на проверочных наборах, сформированных таким образом, может быть излишне оптимистично, или, даже, вводить в заблуждение.

12.5. Сравнение репрезентативности различных выборок

В этом разделе мы постараемся сравнить различные стратегии отбора представительных выборок, а именно, сравнить новый метод граничных объектов с алгоритмами Кеннарда-Стоуна и D_оптимальным планированием, которые кратко изложены в разделе 12.1. Так как в двух последних методах размер представительного набора самими процедурами не определяется, мы будем отбирать в каждую из таких выборок по 19 объектов, столько же, как и в граничном наборе, т.е. $I_K = I_D = I_B = 19$. Из калибровочного C набора, анализируя матрицу X-счетов **T**, полученную с помощью Модели_G, сформируем K набор (по методу Кеннарда-Стоуна), а оставшиеся от калибровочного набора объекты соберем в «избыточный» RK набор. Далее, применим к K набору процедуру (12.3). В результате получаем ПЛС Модель_K с 4 ГК и соответствующую ей ПИО модель, а также и ДСО.

Та же процедура повторялась и для набора, сформированного с помощью D-оптимального алгоритма. Используя этот алгоритм, 19 объектов из C набора были отобраны в D набор, оставшиеся 80 «избыточных» объектов составили RD набор. Согласно процедуре (12.3) были построены ПЛС Модель_D, и, соответствующая ей ПИО модель, а также и ДСО. Еще раз необходимо подчеркнуть, что для проверки всех этих моделей использовался один и тот же проверочный T набор.

Сравнивая значения RMSEP, вычисленные для различных моделей (Таб. 12.3, строки 1, 2, 4, 6) видно, что, в среднем, точность предсказания этих моделей разная. Что касается ПИО классификации статуса объектов, то ДСО, построенные по Модели_B и Модели_D для проверочного набора совпадают. Для Модели_K на ДСО появляется дополнительный абсолютно внешний объект. Несмотря на это, ПИО размах для Модели_K иногда совпадает, но по большей части больше, чем ПИО размах для Модели_B, вычисленный для объектов проверочного набора. Так как величина ПИО размаха определяет ширину интервала предсказания, это означает, что предсказание с

помощью Модели_K будет менее точным чем с помощью Модели_B. Что касается Модели_D, то ПИО размах, вычисленный индивидуально для объектов T набора, совпадает с ПИО размахом Модели_B. Аналогичная ситуация проявляется и при сравнении значений ПЛС размаха (Рис. 12.6).

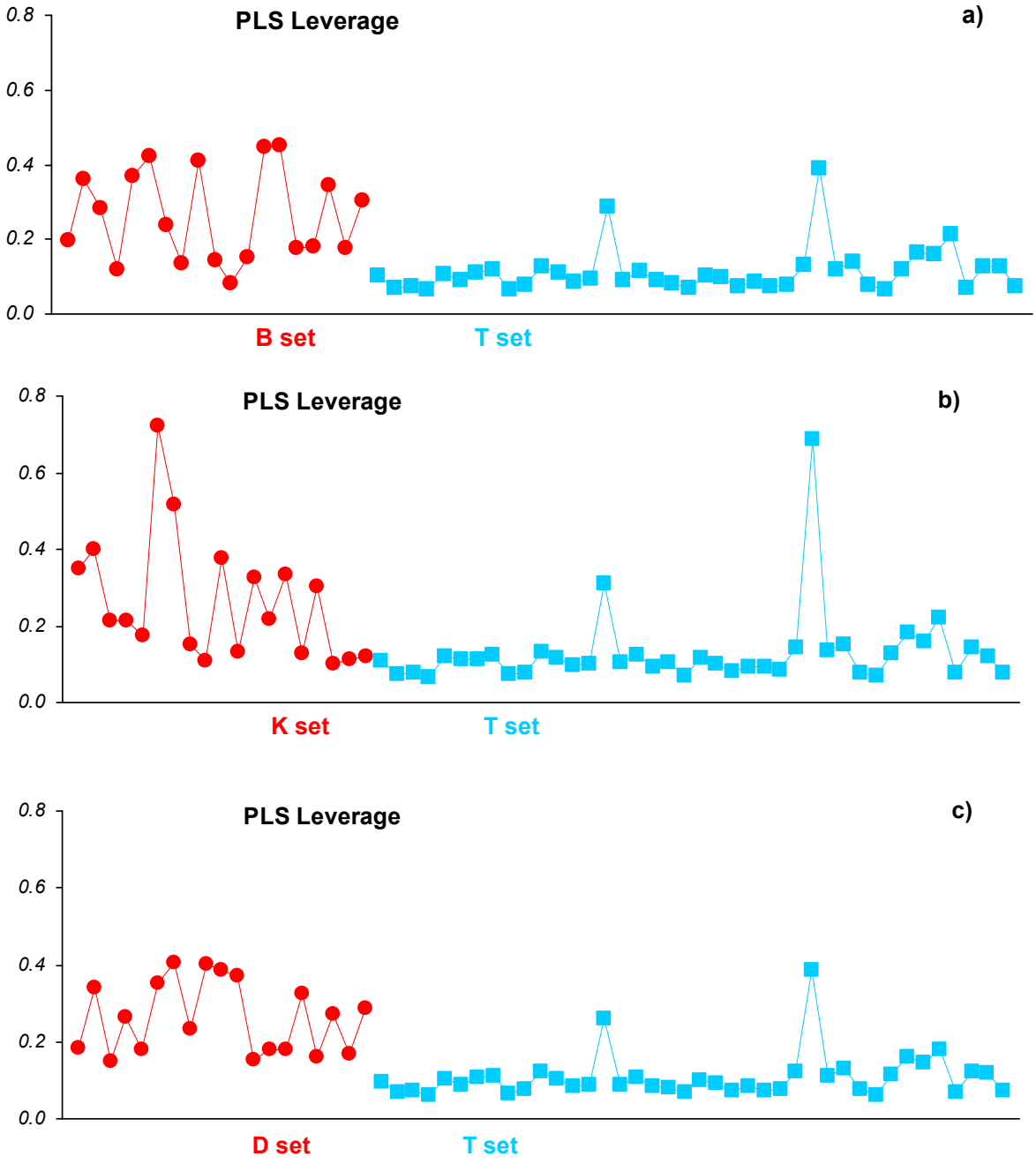
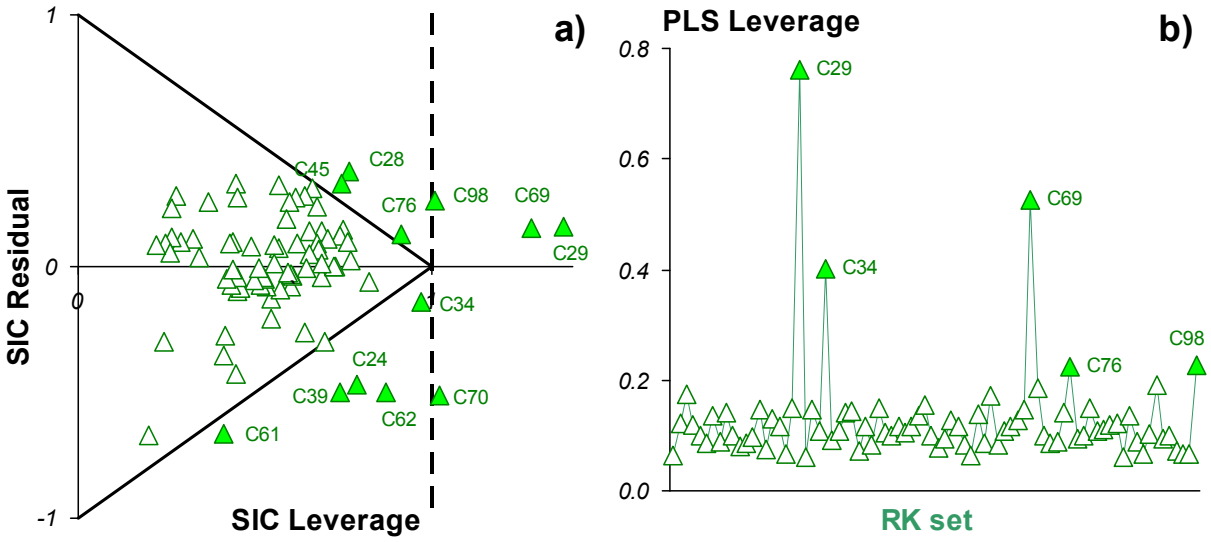


Рис. 12.6. Определение влажности зерна. Значения ПЛС размаха для различных моделей.
■ - Проверочный T набор, ● Калибровочные выборки (B, K и D наборы)

Анализируя полученные графики, можно отметить что наибольшие значения ПЛС размаха для калибровочных выборок в Модели_B (Рис. 12.6a) и Модели_D (Рис. 12.6c)

существенно меньше, чем в Модели_K (Рис. 12.6b). Распределение величин ПЛС размаха для Модели_B и Модели_D очень похожи, как для калибровочных выборок, так и для T набора. Из этого можно заключить, что наборы D и B имеют сходные свойства, а калибровки, построенные на их основе, обладают сравнимой предсказательной способностью. В этом отношении, модель, построенная на основе K набора, имеет худшие свойства.

Это можно объяснить, если вспомнить стратегию отбора объектов в различные короткие наборы. Стратегия граничных объектов и D-оптимальный алгоритмы стараются выбирать наиболее влиятельные, удаленные от центра модели объекты. В то время как метод Кеннарда-Стоуна отбирает объекты равномерно по всему набору.



ДСО для RK набора

ПЛС размах для объектов из RK набора

Рис. 12.7. Определение влажности зерна. Модель_K 4 ГК. Δ- RK объекты, ▲-выделенные RK объекты

Интересно проследить, как Модель_K предсказывает свои «избыточные» объекты, т.е. объекты RK набора. На ДСО (Рис. 12.7a) можно увидеть много объектов, которые классифицируются как внешние. С точки зрения ПИО подхода, K набор нельзя считать представительным, так как не все «избыточные» объекты оказываются внутренним. Среди RK набора выявлено 12 внешних объектов, из них 4 абсолютно внешние. Если обратиться к ПЛС диагностике, то значение RMSEP для RK набора близко к T набору, также в RK наборе имеются объекты с большими значениями ПЛС размаха, а объект C29 можно даже считать выбросом (Рис. 12.7b).

Что касается Модели_D, то она лучше, чем Модель_K предсказывает свои «избыточные» объекты. Согласно ПИО классификации, в RD наборе обнаружено только 7 внешних объектов, причем все они расположены очень близко к границе модели.

12.6. Различные калибровочные наборы

Для того чтобы подтвердить, что подход, основанный на граничных объектах, работает, и эффективный отбор представительной выборки произошел не благодаря удачному формированию калибровочного и проверочного наборов, проведем статистическое моделирование и 10 раз повторим следующую процедуру.

1. Набор исходных данных (G набор, I_G= 139) случайным образом делится на калибровочный (C набор, I_C= 99) и проверочный (T набор, I_T= 40).

2. Для каждой такой пары C и T наборов строятся ПЛС модель с 4 ГК, и соответствующая ей ПИО модель с b_{SIC}=1.5 (Модель_C)

3. Для каждого C-набора вычисляются свои B-, K-, и D-наборы и к ним применяется процедура (12.3).

Таб. 12.4. Определение влажности зерна. ПЛС модели с 4 ГК. Выбор граничных B наборов и оценка основных свойств для 10 калибровочных/проверочных наборов

#	I_B	Модель_C		Модель _B		Модель _K		Модель _D	
		RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP
1	18	0.258	0.359	0.328	0.372	0.209	0.362	0.155	0.362
2	19	0.309	0.227	0.456	0.249	0.304	0.281	0.289	0.267
3	19	0.280	0.312	0.426	0.330	0.212	0.339	0.266	0.335
4	21	0.292	0.281	0.471	0.305	0.253	0.304	0.295	0.325
5	24	0.289	0.287	0.449	0.278	0.305	0.293	0.245	0.311
6	21	0.292	0.281	0.471	0.305	0.253	0.304	0.295	0.325
7	18	0.290	0.292	0.469	0.278	0.264	0.283	0.258	0.289
8	21	0.284	0.304	0.423	0.317	0.202	0.328	0.244	0.319
9	22	0.277	0.315	0.477	0.329	0.274	0.334	0.224	0.348
10	21	0.295	0.276	0.453	0.318	0.206	0.315	0.234	0.342
Среднее		0.287	**0.293**	0.442	**0.308**	0.248	**0.314**	0.251	**0.322**

Естественно, что число граничных объектов и значения среднеквадратичных остатков (RMSE) получились немного разными для различных разбиений на C и T наборы. Но в целом результаты такого моделирования (Таб. 12.4) подтверждают эффективность ПИО подхода.

Последняя строка в Таб. 12.4 представляет средние значения (после 10 реализаций) для четырех видов моделей. Приведенный результат подтверждает, что граничные объекты, собранные из наиболее влиятельных объектов, действительно формируют представительную выборку. Эти влиятельные объекты существенны не только для ПИО подхода, они так же играют важную роль и в традиционном билинейном моделировании.

12.7. Эксперимент 2. Определение следовых концентраций нефти в воде

Вернемся к примеру, подробно рассмотренному в главе 8, и проверим, как работает процедура формирования представительной выборки с помощью граничных объектов. Напомним, что в качестве предикторов при построении калибровки используются Фурье преобразованные акустические спектры (1024 переменные), а в качестве откликов – следовые концентрации нефти в воде. Исходный набор состоит из 80 объектов (G набор), из них 40 объектов в калибровочном C наборе, и 40 – в проверочном T наборе. После предварительной подготовки исходных данных получаем ПЛС модель с 2 ГК (раздел 8.2). В C наборе обнаружено 8 граничных объектов, при этом оценки МП β следующие: b_{min}=0.154, и b_{SIC}=0.29. Так же, как в предыдущем эксперименте сформируем три набора B, K и D, и построим соответствующие им модели. Названия моделей имеют тот же смысл, что и модели представленные на Рис. 12.1. Предсказательная способность различных моделей представлена в Таб. 12.5.

Таб. 12.5**.** Следовые концентрации нефти в воде. ПЛС модели с 2 ГК. В таблице представлены значения RMSEP подсчитанные для проверочных наборов, указанных в каждом столбце.

Модель_C	Модель _B		Модель _K		Модель _D	
T набор	T набор	RB набор	T набор	RK набор	T набор	RD набор
0.092	0.100	0.070	0.122	0.111	0.106	0.08

Модель_B демонстрирует хорошие предсказательные свойства на T наборе, а также надежно предсказывает «избыточные» объекты из RB набора (колонка 3 в Таб. 12.5). Остальные модели также демонстрируют удовлетворительные предсказательные свойства, в среднем. ДСО, порождаемые соответствующими ПИО моделями, проводят

более детальный анализ и индивидуально классифицируют каждый из предсказанных объектов. Сводные данные по ПИО классификации представлены в Таб. 12.6. Величины указанные в клетках – это количество объектов, которые классифицируется как внешние или абсолютно внешние в Т наборе, а также в различных «избыточных» (RB, RK и RD) наборах.

Таб. 12.6. Следовые концентрации нефти в воде. ПИО модели. Результаты классификации для различных наборов Т

Калибровочный набор	Проверочный набор (I_T=40)		Избыточный набор (I_R=32)	
	Внешние	Абс. внешние	Внешние	Абс. внешние
С набор (Модель_С)	8	1	-	-
В набор (Модель_В)	10	0	3	0
К набор (Модель_К)	17	1	9	0
D набор (Модель_D)	10	1	5	0

Из Таб. 12.6 видно, что статус объектов, определенных с помощью Модели_С, несколько отличается от результатов, полученных с помощью Модели_В. Это небольшое рассогласование несущественно. Его можно объяснить тем, что граничный В набор состоит только из восьми объектов. Несмотря на это, рассматриваемый пример показывает эффективность применения метода граничных объектов для формирования представительной выборки и в случае небольшого исходного набора, и достаточно простой, всего 2 ГК, модели.

12.8. Эксперимент 3. Аналитический контроль процесса

Вернемся к примеру аналитического контроля процесса, подробно проанализированного в разделе 11.1 Матрица **X** состоит из 25 переменных. Вектор откликов **y** характеризует выходной показатель. Обучающий С набор состоит из 102 наблюдений (объектов), проверочный Т набор состоит из 52 объектов. Данные центрированы и шкалированы, так, как это описано в разделе 11.2. Как и в предыдущих примерах, Т набор используется для внешней проверки всех моделей. ПЛС Модель_С

использует 7 ГК, а в соответствующей ПИО модели оценки β следующие: b_{min}=0.040, и b_{SIC}=0.058. Согласно ПИО классификации, в С наборе обнаружено сорок шесть граничных объектов. После этого сформированы три коротких набора: B, K и D, по 46 объектов в каждом, и построены три ПЛС модели, в соответствии с блок-схемой, представленной на Рис. 12.1. Предсказательные свойства для этих моделей приведены в Таб. 12.7. Так же, как и в предыдущих примерах, Модель_B имеет предсказательную способность не хуже, чем Модель_C. У Модели_K и Модели_D предсказательная способность хуже.

Таб. 12.7. Аналитический контроль процесса. ПЛС модели с 7 главными компонентами. Величины в клетках представляю значения RMSEP, вычисленные для проверочных наборов, указанных в колонках

Модель_C	Модель _B		Модель _K		Модель _D	
T набор	T набор	RB набор	T набор	RK набор	T набор	RD набор
0.018	0.018	0.010	0.020	0.018	0.027	0.028

Интересно исследовать, как сильно перекрываются три сформированные выборки, т.е. сколько у них общих объектов. Графически эти три выборки представлены на Рис. 12.8

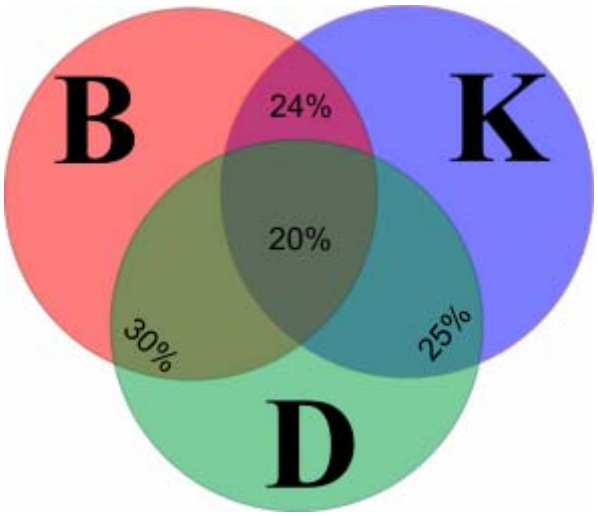


Рис. 12.8. Аналитический контроль процесса. Выборки по 46 объектов каждая.
B -B набор, K – K набор D – D набор

Проценты в пересекающихся областях показывают часть общих объектов в этих выборках. Видно, что только 20% объектов из калибровочного набора попали во все три

выборки. Это означает, что каждый метод формирует свой собственный представительный набор.

Зависимость точности предсказания от объема представительной выборки

В некоторых случаях, представительная выборка, составляющая 45% калибровочного набора, может показаться слишком большой. Поэтому важно исследовать, как влияет объем выборки на предсказательные свойства модели. Согласно ПИО методу, минимальное число граничных объектов определяются при $\beta = b_{min}$. В рассматриваемом примере аналитического контроля процессов минимальный граничный набор состоит из 8 объектов($I_B \geq 8$). Последовательно увеличивая b с $b = b_{min}$ до $b = b_{SIC}$, мы получаем расширяющийся B набор. Параллельно, для сравнения, применим метод Кеннарда-Стоуна и D оптимальное планирование и построим K наборы и D наборы, с таким же числом объектов. Для каждого из этих наборов строится ПЛС модель с 7 ГК, вычисляются значения RMSEC, а также значения RMSEP на одном и том же проверочном наборе T.

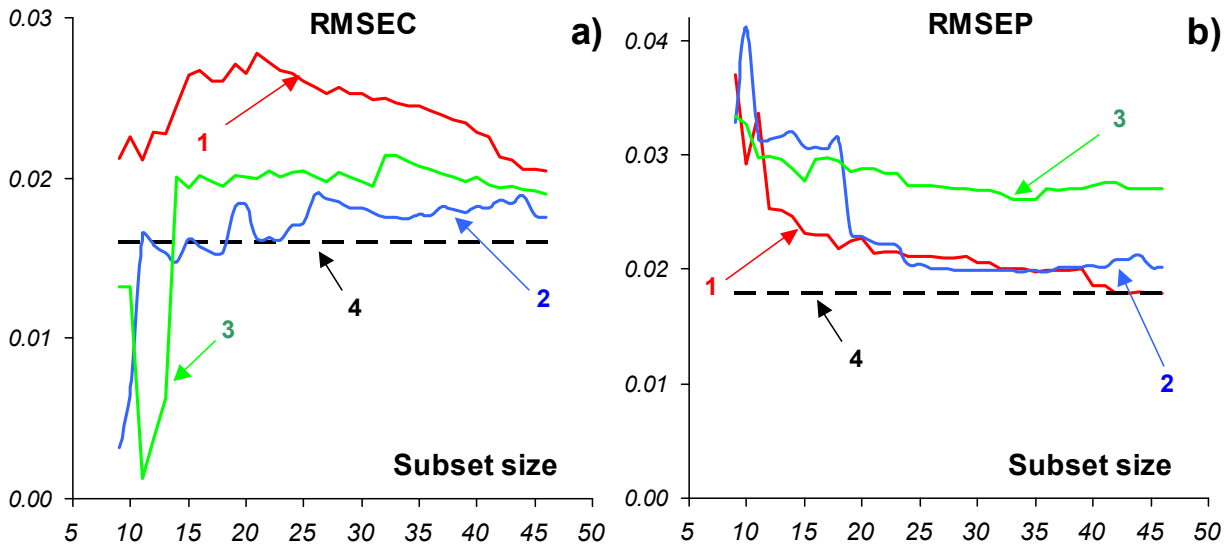


Рис. 12.9. Аналитический контроль процесса. ПЛС модели с 7 ГК. Зависимость RMSE от объема выборки1 – Модель_B, 2 – Модель _K, 3 – Модель _D, 4 – Модель_C

При таком подходе, можно рассматривать RMSEC и RMSEP как функции, зависящие от объема выборки (Рис. 12.9), которые вычисляются для трех моделей (Модель_B, Модель_K, Модель_D). Из Рис. 12.9b, кривая 1, видно, что для B наборов функция RMSEP(I_B) достаточно плавно убывает. Она быстрее, чем для наборов K и D стремится к предельному значению, равному RMSEP(I_C). При этом остаток калибровки

(Рис. 12.9а, кривая 1) RMSEC(I_B) все время остается наибольшим по сравнению с аналогичными значениями, вычисленными для наборов K и D.

Если нашей целью является выбор такого короткого набора объектов, который мог бы использоваться вместо полного калибровочного набора, и при этом предсказательная способность модели не ухудшилась, то необходимо не менее 42 объектов B набора. Этот пример не только подтверждает влиятельность граничных объектов, он так же показывает, что объем выборки, предлагаемый методом ПИО очень близок к оптимальному.

12.9. Результаты главы 12

В этой главе представлен новый метод формирования представительной выборки на примере анализа результатов различных физических экспериментов. Метод граничных объектов основывается на ПИО подходе (теория классификации статуса объектов), объединенным с регрессионными проекционными методами (РГК, ПЛС).

Показано, что стратегия выбора граничных объектов является объективной и не зависит от того, кто ее применяет. При этом не требуется никакой дополнительной или априорной информации, которую невозможно было бы получить из исследуемого набора данных. Очень важно, что граничные объекты, которые формируют представительную выборку, являются влиятельными объектами не только с точки зрения ПИО моделирования, но они так же являются влиятельными объектами и при регрессионном моделировании.

Термин «избыточные» объекты, который применялся к внутренним объектам калибровочного набора, не следует трактовать непосредственно. Чем больше объектов в калибровочном наборе, тем точнее можно определить число главных компонент в ПЛС или РГК моделях, а так же получить более точную оценку b_{SIC}.

Наборы данных представлены тремя различными физическими экспериментами, рассмотренные в этой главе были порождены различными практическими задачами, они отличаются друг от друга, как по внутреннему устройству данных, так и по сложности построенных для них ПЛС моделей. Это позволило продемонстрировать работу метода в различных условиях.

Анализируя три стратегии формирования представительной выборки, можно сделать следующие выводы. Метод Кеннарда-Стоуна эффективен в том случае, когда исходный набор данных требуется разделить на равноценные наборы, например

обучающий и проверочный. Этот метод выбирает образы равномерно, поэтому он менее эффективен при отборе наиболее влиятельных объектов. Методы D оптимального планирования и граничных объектов работаю более эффективно при отборе влиятельных объектов. При этом у метода граничных объектов имеются несколько преимуществ. Во-первых, он однозначно определяет необходимое число объектов в представительной выборке для исследуемого набора данных и построенной модели. Во-вторых, при отборе объектов учитывается информация, как о значениях X переменных, так и Y переменных.

Необходимо заметить, что D оптимальное планирование часто используется для формирования калибровочного набора. При этом исследователь должен понимать, что после того, как он сформирует калибровочный набор из наиболее влиятельных объектов, оставшиеся для проверочного набора объекты будут описывать более узкий диапазон изменений, т.е. будут более «средними». При этом проверка модели с помощью такого проверочного набора может привести к обманчивым результатам.

В заключение хочется отметить, что выбор процедуры формирования представительной выборки существенно зависит от того, для чего эта выборка будет использована. Очень важно понимать сильные и слабые стороны каждого алгоритма, и те идеи, которые заложены в этих процедурах.

Заключение

Необходимость широкого применения и дальнейшего развития многомерных методов анализа данных востребована и тесно связана с тенденциями развития техники физико-химических экспериментов. Во-первых, *объекты анализа* становятся более сложными и комплексными. Во-вторых, *методы анализа* меняются таким образом, чтобы обеспечить получение необходимых данных в режиме реального времени (in line). В-третьих, резко увеличивается *объем данных*, которые повсеместно становятся многомерными и многомодальными. Увеличивается роль гибридных и композиционных методов анализа. В-четвертых, искомая физическая информация очень глубоко спрятана в этих данных, и все *менее формализована.*. В-пятых, прослеживается тенденция в изменении организации физического эксперимента – вместо исследования одной пробы в одном опыте, используется *системный подход*, в котором много разных проб автоматически испытываются одновременно разными методами, в разных условиях (например, технология микрочипов). В-шестых, акцент в исследовании все чаще переноситься на биологические объекты и биофизические процессы.

Все эти тенденции, ставят перед исследователем две главные задачи. Первая – как придумать, организовать, спланировать эксперимент с тем, чтобы получить данные, из которых, в принципе, можно получить нужную физическую информацию. Вторая – как извлечь и интерпретировать эту информацию. Для решения этих задач исследователь должен, в значительной мере, использовать опыт и инструментарий многомерного анализа. Однако широкое использование формальных методов затруднено методологическими проблемами. Исследователь, привыкший использовать в своей практике, пусть и весьма приближенные, но содержательные физические модели считает хемометрический подход слишком формальным и поверхностным. С другой стороны, в арсенале хемометрических методов до сих пор существуют плохо разработанные аспекты. Например, существенным недостатком проекционных регрессионных методов (РГК, ПЛС и пр.) является то, что все эти методы дают результат предсказания в виде точечной оценки, тогда как на практике часто нужна *интервальная оценка*, учитывающая неопределенность прогноза.

В работе рассмотрены теоретические, алгоритмические и методологические аспекты многомерных методов обработки больших массивов данных физических экспериментов. Объединение проекционных регрессионных методов с методом простого

интервального оценивания порождает мощный инструмент для решения задач многомерной калибровки. Такой подход позволяет обрабатывать наборы многоканальных данных, пронизанных внутренними связями, разделять полезную информацию и шум, представлять результат прогноза в интервальной форме, учитывающей неопределенность в прогнозе индивидуально для каждого объекта/измерения. Обобщая полученные результаты, можно сформулировать следующие **выводы**.

В теоретической части работы были получены следующие результаты.

1. Представлен и теоретически обоснован метод простого интервального оценивания (ПИО), предназначенный для решения линейных задач калибровки и прогнозирования больших массивов данных физических экспериментов. Доказаны основные свойства метода: ограниченность, состоятельность, несмещенность.

2. Обосновано предположение об ограниченности погрешностей, лежащее в основе метода ПИО. Показано, что это допущение является не недостатком, а преимуществом метода, так как, с практической точки зрения, оно более обоснованно, чем традиционное допущение о нормальности, а, следовательно, и неограниченности погрешностей.

3. Показано, что метод ПИО вычисляет оценки неизвестных параметров модели в виде области в пространстве параметров, что, в свою очередь, позволяет представить результаты прогноза отклика в интервальном виде, учитывающим все погрешности (измерения, моделирования и пр.). Это является существенным преимуществом в сравнении с традиционным регрессионным анализом, где результат прогноза – это точечная оценка

4. Приведены аргументы в пользу того, что ПИО-оценки, построенные на основе экстремальных статистик, являются более эффективными, чем традиционные гладкие оценки. Это открывает новое направление исследований в области прикладной статистики – построение суперэффективных оценок.

5. На основе метода ПИО разработан новый подход к классификации статуса объектов и интерпретации прогнозных интервалов. Введены новые понятия: ПИО-остаток и ПИО-размах, диаграмма статуса объектов (ДСО). Введены понятия внутренних, внешних, граничных объектов. Дано определение выбросов и абсолютно внешних объектов.

6. Разработаны новые методы статистического контроля процессов. Метод, названный расширяющимся многомерным статистическим контролем, основан на построении серии ПЛС моделей, совместно с ПИО моделированием. Он позволяет вычислять как точечные, так и интервальные оценки выходного параметра на промежуточных стадиях процесса. Предложен также метод активной оптимизации, разработаны различные стратегий оптимизации.

К практическим результатам работы следует отнести

7. Разработку общего алгоритм, объединяющего проекционные методы (РГК, ПЛС1, ПЛС2), метод линейного программирования (симплекс-метод), оригинальный алгоритм приведения задачи к каноническому виду, алгоритм определения статуса объектов и построения ДСО.

8. Создание компьютерной программы SIC – надстройки для программы Excel. С ее помощью можно проводить многоканальных сигналов, оценивать точность построенной модели, проводить классификацию объектов в зависимости от их влиятельности на модель, для объектов из обучающего набора, и оценить близость новых объектов к модели. Эффективность программы проверена с помощью имитационного моделирования, примеров описанных в литературе, а так же на большом числе реальных экспериментальных данных.

Созданная методология расширяет область применения многомерных методов для построения моделей калибровки по результатам многоканальных физических экспериментов. Новый подход объединяет проекционные методы и метод простого интервального оценивания.

9. На примере определения следовых концентраций нефти в воде показано, что разработанная классификации объектов имеет практическое значение не только в рамках метода ПИО, но и в рамках классических регрессионных моделей, а диаграмма статуса объектов является простым и удобным инструментом для визуализации и детального анализа результатов физических экспериментов.

10. Предложен новый метод выбора представительных (влиятельных) объектов из экспериментального набора данных, названный методом граничных образов, который может применяться как для переноса калибровочных моделей с одного прибора на другой, так и для эффективного уменьшения объема калибровочного набора. На примере обработки результатов БИК-спектроскопии зерна показана эффективность формирования представительной выборки.

11. На примере предсказания активности антиоксидантов проведено сравнение формального (ПИО) и содержательного (нелинейная регрессия) моделирования. Показано, что содержательный подход позволяет проводить экстраполяцию, однако при этом нельзя ограничить область экстраполяции. Формальный метод имеет строгую область применимости, очерченную с помощью техники ПИО статуса. Он дает надежные результаты при решении задач классификации или *интерполяции*. Показана практическая значимость построенных моделей, позволяющих заменить длительное термостарение быстрым экспериментом с помощью дифференциальной сканирующей калориметрии.

12. Показано, что дополнение стандартного метода ПЛС дискриминации методом ПИО повышает информативность при решении задач качественного анализа и распознавания. Предложена методика экспресс-распознавания фальсифицированных лекарств на основе БИК-спектроскопии.

Приложение

13. Алгоритмы

13.1. Метод главных компонент (NIPALS алгоритм)

Алгоритм предназначен для представления исходной матрицы **X**, размерностью *I* строк и *J* столбцов в виде:

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{t} + \mathbf{E}$$

или в векторном виде

$$x_{ij} = \sum_{k=1}^{K} t_{ik}\, p_{jk} + e_{ij(K)}$$

Индекс *K* обозначает количество главных компонент, включенных в разложение. Матрицы **T** (*I*×*K*) и **P** (*J*×*K*) называются, соответственно, матрицей счетов и нагрузок.

Алгоритм последовательно вычисляет все главные компоненты, начиная с той, которая соответствует максимальному собственному значению матрицы $\mathbf{X}^{T}\mathbf{X}$, и далее строго в порядке убывания собственных значений. Каждая ГК получается итерационно, повторяя регрессию **X** на счета **t** для уточнения нагрузок **p**, и регрессию **X** на найденные нагрузки **p** для уточнения **t**.

Предварительно переменные **X** шкалируют для выравнивания уровня шума. После этого переменные центрируют, например, вычитая среднее по каждой переменной, вычисленное на калибровочных объектах, и таким образом формируют матрицу $\mathbf{X}_0$

В начале процедуры, $k = 0$.

Шаг 1.

Задается текущее значение индекса $k=k+1$, номера искомой главной компоненты.

Шаг 2.

В качестве начального приближения, выбирается такой столбец $\mathbf{t}_k^0$, матрицы $\mathbf{X}_{a-1}$, который имеет наибольшую сумму квадратов элементов.

Шаг 3.

Уточнение приближения вектора нагрузок $\mathbf{p}_k^m$ для данной РС путем проекции матрицы $\mathbf{X}_{k-1}$ на $\mathbf{t}_k^m$, т.е.

$$\mathbf{p}_k^{mt} = (\mathbf{t}_k^{mt}\mathbf{t}_k^m)^{-1}\mathbf{t}_k^m\mathbf{X}_{k-1}$$

Шаг 4.

Нормирование вектора $\mathbf{p}_k^m$ для того, чтобы избежать вычислительных погрешностей:

$$\mathbf{p}_k^m = \mathbf{p}_k^m (\mathbf{p}_k^{m\,t} \mathbf{p}_k^m)^{-0.5}$$

Шаг 5.

Уточнение приближения счетов $\mathbf{t}_k^m$ для данной ГК путем проекции матрицы $\mathbf{X}_{a\text{-}1}$ на $\mathbf{p}_k^m$:

$$\mathbf{t}_k = \mathbf{X}_{k-1} \mathbf{p}_k^m (\mathbf{p}_k^{m\,t} \mathbf{p}_k^m)^{-1}$$

Шаг 6.

Вычисление приближения собственного значения τ_k^m ;

$$\tau_k^m = \mathbf{t}_k^{m\,t} \mathbf{t}_k^m$$

Шаг 7.

Проверка сходимости: если

$$\left| \tau_k^m - \tau_k^{m-1} \right| < \varepsilon \left| \tau_k^m + \tau_k^{m-1} \right|$$

где ε некоторая заданная константа, например ε=0.0001, то для данной РС итерации сошлись на m-ом шаге, т.е.

$$\mathbf{t}_k = \mathbf{t}_k^m \qquad \mathbf{p}_k = \mathbf{p}_k^m$$

Если нет, вернуться к шагу 3 и продолжить итерационные вычисления.

Шаг 8.

Вычитание вклада текущей главной компоненты.

$$\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k^t .$$

Если $k<K$, то перейти к шагу 1 для определения следующей РС. Ели же $k \geq K$, то процедура завершена.

13.2. Регрессия на главные компоненты - РГК

Алгоритм РГК работает в два этапа: вначале с помощью МГК производится декомпозиция матрицы **X**, а уже затем в пространстве главных компонент строят обычную множественную регрессию **Y** на **T**.

Регрессия на главные компоненты для *M* различных **Y**-переменных на *J* **X**-переменных эквивалентна *M* отдельным регрессиям на главные компоненты – по одной для каждой **Y**-переменной. В каждой такой регрессии используются одни и те же

переменные **X**. Таким образом, достаточно рассмотреть только случай однооткликовой задачи, когда **y –** это вектор.

РГК получается построением регрессии **y** на векторы счетов **t**, которые в свою очередь получаются из МГК-разложения матрицы **X**. Коэффициенты регрессии **a** для каждого **y** можно записать в виде

$$\mathbf{a}=\mathbf{Pq} \qquad \text{(A 1)}$$

где *X*-нагрузки $\mathbf{P}=\{\mathbf{p}_{jk}, j=1,2,...,J; k=1,2,...,K\}$ представляют собой МГК-нагрузки для рассматриваемых *K* главных компонент, а *Y*-нагрузки $\mathbf{q}=(q_1,...,q_K)^t$ находятся обычным образом, с помощью метода наименьших квадратов при построении регрессии **y** на, **T** что выражается моделью **y** = **Tq**+**f**. Так как столбцы матрицы **T** взаимно ортогональны по построению, это эквивалентно

$$\mathbf{q}=(\mathrm{diag}(1/\tau_k))\mathbf{T}^t\mathbf{y} \qquad \text{(A 2)}$$

Подставляя выражение (A 2) в (A 1) и заменяя **T** на **XP**, получаем РГК-оценку коэффициентов **a** в виде

$$\mathbf{a}=\mathbf{P}(\mathrm{diag}(1/\tau_k))\mathbf{P}^t\mathbf{X}^t\mathbf{y} \qquad \text{(A 3)}$$

которая часто используется в качестве определения РГК. Если количество ГК равно *J*, РГК дает те же коэффициенты **a**, что и МНК. Но так как переменные **X** часто коррелируют между собой и зашумлены, то оптимальное значение *K* существенно меньше чем *J*. При вычислении МНК-оценок используется деление на собственные значения τ_k которые в случае плохой обусловленности матрицы **X** близки к нулю, что делает определение коэффициентов **a** для МНК неустойчивым. Напротив, в РГК определение коэффициентов **a** устойчиво за счет отбрасывания таких ненадежных собственных значений.

13.3. Проекция на латентные структуры (ПЛС)

В общей форме ПЛС-модель записывается как

$$\mathbf{X}=\mathbf{T}\,\mathbf{P}^T+\mathbf{E} \text{ и } \mathbf{Y}=\mathbf{TQ}^T+\mathbf{F}$$

В этом разделе приводится NIPALS алгоритм, который последовательно компонента за компонентой вычисляет ПЛС-проекции матриц **X** и **Y.** В случае, когда есть несколько переменных в отклике, алгоритм является итерационным. Если модель одноткиковая, то вычисления производятся непосредственно, не прибегая к итерациям. Алгоритм состоит

из двух частей: (1) – это построение модели (калибровка); (2) – предсказание отклика на новых/тестовых объектов.

Однооткликовая модель (ПЛС1)

Калибровка.

Шаг 1.

Центрирование переменных **X** и **y**

$$\mathbf{X}_0 = \mathbf{X} - \mathbf{I}\bar{\mathbf{x}}^{\mathrm{t}} \text{ и } y_0 = \mathbf{y} - \mathbf{I}\bar{y}$$

где **I** – это единичная матрица соответствующего размера, $\bar{\mathbf{x}}$ -вектор средних значений матрицы **X**, вычисленных по столбцам. Выбирается максимальное значение главных компонент K так, чтобы оно было больше, чем число явлений, наличие которых предполагается в **X**. На начальном шаге полагаем $k = 1$:

Шаг 2.

Вектор взвешенных нагрузок $\mathbf{w}_{\mathrm{k}}$ вычисляется по формуле

$$\mathbf{w}_{\mathrm{k}} = c\mathbf{X}_{\mathrm{k-1}}^{\mathrm{t}}\mathbf{y}_{\mathrm{k-1}}$$

как МНК решение для локальной модели $\mathbf{X}_{\mathrm{k-1}} = \mathbf{y}_{\mathrm{k-1}}\mathbf{w}_{\mathrm{k}}^{\mathrm{t}} + \mathbf{E}$. При этом c –это нормировочный фактор, который делает длину $\mathbf{w}_{\mathrm{a}}$ равной единице, т.е.

$$c = (\mathbf{y}_{\mathrm{k-1}}^{\mathrm{t}}\mathbf{X}_{\mathrm{k-1}}\mathbf{X}_{\mathrm{k-1}}^{\mathrm{t}}\mathbf{y}_{\mathrm{k-1}})^{-0.5}$$

Шаг 3.

Определяются счета $\mathbf{t}_{\mathrm{k}}$ с помощью локальной модели

$$\mathbf{X}_{\mathrm{k\text{-}1}} = \mathbf{t}_{\mathrm{k}}\mathbf{w}_{\mathrm{k}}^{\mathrm{t}} + \mathbf{E}$$

МНК-решение имеет вид

$$\mathbf{t}_{\mathrm{k}} = \mathbf{X}_{\mathrm{k-1}}\mathbf{w}_{\mathrm{k}} \qquad \text{(A 4)}$$

учитывая, что $\mathbf{w}_{\mathrm{k}}^{\mathrm{t}}\mathbf{w}_{\mathrm{k}} = \mathbf{1}$

Шаг 4.

Вычисляются «спектральные» нагрузки $\mathbf{p}_{\mathrm{a}}$ с помощью локальной модели

$$\mathbf{X}_{\mathrm{k-1}} = \mathbf{t}_{\mathrm{k}}\mathbf{p}_{\mathrm{k}}^{\mathrm{t}} + \mathbf{E}$$

которая имеет МНК решение:

$$\mathbf{p}_{\mathrm{k}} = \mathbf{X}_{\mathrm{k-1}}^{\mathrm{t}}\mathbf{t}_{\mathrm{k}} / \mathbf{t}_{\mathrm{k}}^{\mathrm{t}}\mathbf{t}_{\mathrm{k}}$$

Шаг 5.

Определяются «химические» нагрузки q_{k} с помощью локальной модели

$$\mathbf{y}_{k-1} = \mathbf{t}_k q_k + \mathbf{f}$$

которая имеет МНК решение:

$$q_k = \mathbf{t}_k^t \mathbf{y}_{k-1} / \mathbf{t}_k^t \mathbf{t}_k$$

Шаг 6.

Вычитая вклад от найденной ГК, получаем новые остатки **X** и **y**,:

$$\mathbf{E}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k^t \quad \text{и} \quad \mathbf{f}_k = \mathbf{y}_{k-1} - \mathbf{t}_k q_k$$

Для этих остатков вычисляются различные статистические показатели, получаемые суммированием e_{ij} по всем объектам *i* и переменным *j*, а также суммированием f_i по всем *i* объектам.

Шаг 7.

Если *k*=*K*, переходим к Шагу 8. Иначе, заменяем старые величины $\mathbf{X}_{k-1}$ и $\mathbf{y}_{k-1}$ новыми значениями $\mathbf{E}_k$, $\mathbf{f}_k$ и увеличиваем номер ГК на единицу:

$$\mathbf{X}_k = \mathbf{E}_k \ , \ \mathbf{y}_k = \mathbf{f}_k \ k=k+1,$$

и возвращаемся на Шаг 2.

Шаг 8.

Для *K* ГК, вычисляем a_0 и **a**, которые будут использоваться в уравнении прогноза:

$$\hat{\mathbf{y}} = \mathbf{I}a_0 + \mathbf{X}\mathbf{a}\,,\ \mathbf{a} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}\mathbf{q}\,,\ a_0 = \bar{y} - \bar{\mathbf{x}}^t\mathbf{a} \qquad \text{(A 5)}$$

Прогноз

Полное предсказание. Для каждого нового объекта *i* = 1, 2, … выполняются Шаг 9 –Шаг 13. Если используется формула краткого предсказания, то выполняется Шаг 14.

Шаг 9.

Вычисляется:

$$\mathbf{x}_{i,0}^t = \mathbf{x}_i^t - \bar{\mathbf{x}}\mathbf{I}\,,$$

где $\bar{\mathbf{x}}$ - центр объектов калибровочного набора , *a*=1.

Шаг 10.

Находим $\mathbf{t}_{i,k}$ в соответствии с формулой (A 4) , т.е.

$$\mathbf{t}_{i,k} = \mathbf{x}_{i,k-1}^t \mathbf{w}_k\,.$$

Шаг 11.

Вычисляем новые остатки

$$\mathbf{x}_{i,k} = \mathbf{x}_{i,k-1} - \mathbf{t}_{ik}\mathbf{p}_k^t$$

Шаг 12.

Если $k < K$, то увеличиваем k на единицу и возвращаемся на Шаг 9. Если $k = K$, переходим к Шагу 13.

Шаг 13.

Предсказание y_i с помощью формулы

$$\hat{y}_i = \bar{y} + \sum_{k=1}^{K} \mathbf{t}_{ik} q_k$$

Вычисление статистик для остатков в $\mathbf{x}_{iK}$ и $\mathbf{t}_i$.

Шаг 14.

Краткое предсказание. Вместо шагов 9-13 можно найти $\hat{y}_i$, используя a_0 и $\mathbf{a}$ вычисленные на Шаге 8, (А 5) т.е.

$$\hat{y}_i = a_0 + \mathbf{x}_i^t \mathbf{a}$$

Многооткликовая модель (ПЛС2)

ПЛС2, это параллельный ПЛС-метод для нескольких переменных **Y.** Построение калибровочной модели похоже на ПЛС1, однако в данном случае алгоритм уже итерационный.

Калибровка.

Шаг 1.

Центрирование переменные $\mathbf{X}$ и $\mathbf{y}^t$

$$\mathbf{X}_0 = \mathbf{X} - \mathbf{I}\bar{\mathbf{x}}^t \text{ и } \mathbf{Y}_0 = \mathbf{Y} - \mathbf{I}\bar{\mathbf{y}}^t$$

где I – это единичная матрица соответствующего размера, $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ -векторы средних значений соответствующих матриц $\mathbf{X}$ и $\mathbf{Y}$ вычисленных по столбцам. Выбирается максимальное значение главных компонент K. На начальном шаге полагаем $k = 1$. В качестве начального приближения, выбирается такой столбец $\mathbf{u}_k^0$, матрицы $\mathbf{Y}_{k-1}$, который имеет наибольшую сумму квадратов элементов, полагаем $\mathbf{t}_k^0 = \mathbf{0}$.

Шаг 2.

Вектор взвешенных нагрузок для ГК k на m-ой итерации $\mathbf{w}_k^m$ вычисляется по формуле

$$\mathbf{w}_k^m = c\mathbf{X}_{k-1}^t \mathbf{u}_k^{m-1}$$

как МНК решение для локальной модели $\mathbf{X}_{k-1} = \mathbf{u}_k^{m\text{-}1}\mathbf{w}_k^{m\,t} + \mathbf{E}$. При этом c –это нормировочный фактор, который делает длину $\mathbf{w}_k^m$ равной единице, т.е.

$$c = (\mathbf{u}_k^{m\text{-}1\,t}\mathbf{X}_{k-1}\mathbf{X}_{k-1}^t\mathbf{u}_k^{m\text{-}1})^{-0.5}$$

Шаг 3.

Определяются счета $\mathbf{t}_k^m$ с помощью локальной модели

$$\mathbf{X}_{k\text{-}1} = \mathbf{t}_k^m\mathbf{w}_k^{m\,t} + \mathbf{E}$$

МНК-решение имеет вид

$$\mathbf{t}_k^m = \mathbf{X}_{k-1}\mathbf{w}_k^m \qquad \text{(A 6)}$$

учитывая, что $\mathbf{w}_k^{m\,t}\mathbf{w}_k^m = \mathbf{1}$

Шаг 4.

Вычисляются «спектральные» нагрузки $\mathbf{p}_k^m$ с помощью локальной модели

$$\mathbf{X}_{k-1} = \mathbf{t}_k^m\mathbf{p}_k^{m\,t} + \mathbf{E}$$

которая имеет МНК решение:

$$\mathbf{p}_k^m = \mathbf{X}_{k-1}^t\mathbf{t}_k^m \big/ \mathbf{t}_k^{m\,t}\mathbf{t}_k^m$$

Шаг 5.

Определяются «химические» нагрузки $\mathbf{q}_k^m$ с помощью локальной модели

$$\mathbf{Y}_{k-1} = \mathbf{t}_k^m\mathbf{q}_k^{m\,t} + \mathbf{F}$$

которая имеет МНК решение:

$$\mathbf{q}_k^m = \mathbf{t}_k^m\mathbf{Y}_{k-1}^t \big/ \mathbf{t}_k^{m\,t}\mathbf{t}_k^m$$

Шаг 6.

Проверяется сходимость вектору счетов $\mathbf{t}_k^m$, т.е. если для всех I объектов из калибровочного набора (i=1,2,...,I) выполнено условие

$$\left|t_{ki}^m - t_{ki}^{m-1}\right| \le \boldsymbol{\varepsilon}\left|t_{ki}^m + t_{ki}^{m-1}\right|$$

при достаточно малом значении ε, то итерации сошлись, и

$$\mathbf{t}_k = \mathbf{t}_k^m \;,\; \mathbf{u}_k = \mathbf{u}_k^m \;,\; \mathbf{q}_k = \mathbf{q}_k^m \;,\; \mathbf{w}_k = \mathbf{w}_k^m$$

Переходим на шаг 8.

Шаг 7.

Если сходимость не достигнута, то продолжаем итерации, вычисляем новое приближение $\mathbf{u}_k^m$, используя локальную модель

$$\mathbf{Y}_{k-1} = \mathbf{u}_k^m \mathbf{q}_k^{m^t} + \mathbf{F}$$

которая дает решение

$$\mathbf{u}_k^m = \mathbf{Y}_{k-1}\mathbf{q}_k^m \left(\mathbf{q}_k^{m^t}\mathbf{q}_k^m\right)^{-1}.$$

Полагаем m=m+1 и переходим к Шагу 2.

Шаг 8.

Вычитая вклад от найденной ГК, получаем новые остатки **X** и **Y**,:

$$\mathbf{E}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k^t \quad \text{и} \quad \mathbf{F}_k = \mathbf{Y}_{k-1} - \mathbf{t}_k \mathbf{q}_k$$

Для этих остатков вычисляются различные статистические показатели, получаемые суммированием e_{ij} по всем объектам i и переменным j, а также суммированием f_{il} по всем i объектам и откликам l.

Шаг 9.

Если k=K, переходим к Шагу 10. Иначе, заменяем старые величины $\mathbf{X}_{k-1}$ и $\mathbf{Y}_{k-1}$ новыми значениями $\mathbf{E}_k$, $\mathbf{F}_k$ и увеличиваем номер ГК на единицу:

$$\mathbf{X}_k = \mathbf{E}_k \ , \ \mathbf{Y}_k = \mathbf{F}_k \ , \ k=k+1,$$

Вновь выбираем для начала итерационного процесса $\mathbf{u}_k^0$, как столбец матрицы $\mathbf{Y}_{k-1}$, который имеет наибольшую сумму квадратов элементов, полагаем $\mathbf{t}_k^0 = \mathbf{0}$ и возвращаемся на Шаг 2.

Шаг 10.

Для K ГК, вычисляем $\mathbf{a}_0$ и **A**, которые будут использоваться в уравнении прогноза:

$$\hat{\mathbf{Y}} = \mathbf{a}_0^t + \mathbf{XA}\,,\ \mathbf{A} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}\mathbf{Q}^t\,,\ \mathbf{a}_0 = \bar{\mathbf{y}}^t - \bar{\mathbf{x}}^t\mathbf{A} \qquad \text{(A 7)}$$

Заметим, что **P** и **Q** не нормированы, **T** и **W** ортогональны, а **W** еще и нормирована.

Прогноз.

Прогнозное значение откликов в ПЛС-2 модели вычисляется так же как и для ПЛС-1. Полное предсказание. Для каждого нового объекта i = 1, 2, … выполняются шаги Шаг 9 –Шаг 13 алгоритма ПЛС-1. Единственное отличие, это использование вместо вектора **q** матрицу **Q**. Если используется формула краткого предсказания, то выполняется Шаг 14, который для ПЛС-2 имеет вид

$$\hat{\mathbf{y}}_i^t = \mathbf{a}_0^t + \mathbf{x}_i^t\mathbf{A}$$

13.4. Симплекс-метод.

Представлен двухэтапный симплекс-метод с введением искусственного базиса.

Этап 1. (Используется для нахождения начального базисного решения).

Вводятся искусственные переменные, необходимые для получения стартовой точки. Записывается новая целевая функция, предусматривающая минимизацию суммы искусственных переменных при исходных ограничениях, видоизмененных за счет введения искусственных переменных. Если минимальное значение новой целевой функции равно нулю (т.е. все искусственные переменные имеют в оптимуме нулевые значения), исходная задача имеет допустимое решение. Переходим к этапу 2. В противном случае, задача не имеет допустимых решений, и процесс вычислений заканчивается.

Этап 2. (Решение исходной задачи ЛП)

Оптимальное базисное решение, полученное на этапе 1, используется в качестве начального решения исходной задачи.

Алгоритм.

Шаг 0. Используя линейную модель канонической формы (7.7), определяют *начальное допустимое базисное решение* путем приравнивания к нулю n-m (небазисных) переменных.

Шаг 1. Из числа текущих небазисных (равных нулю) переменных выбирается *включаемая в новый базис переменная*, уравнение которой обеспечивает улучшение значения целевой функции. Если такой переменной нет, вычисления прекращаются так как текущее базисное решение оптимально. Иначе, переход на Шаг 2.

Шаг 2. Из числа переменных текущего базиса выбирается исключаемая переменная, которая должна принять нулевое значение (стать небазисной) при введении в состав базиса новой переменной.

Шаг 3. Находится новое базисное решение, соответствующее новым составам небазисных и базисных переменных. Осуществляется переход на шаг 1.

Замечание.

Шаг 0 алгоритма симплекс-метода, т.е. отыскание исходной точки по сложности сравним с решением самой задачи ЛП. В данной работе используется метод искусственного базиса и в результате получаем двухэтапный симплекс-метод.

13.5. Вычисление положения эллипсоида.

Целью является получение явной формулы для описания эллипсоида, который содержит в себе ОДЗ. Вычислив положение центра и главных осей эллипсоида, мы можем сдвинуть его так, чтобы он полностью располагался в положительной части пространства параметров R^K, а вместе с эллипсоидом и все значения $\mathbf{a} \in A$ будут положительными.

Рассматриваем линейную модель, после регуляризации матрицы предикторов, т.е после применение одного из проекционных методов с K главными компонентами

$$\mathbf{y} = \mathbf{Ta}+\boldsymbol{\varepsilon}\,, \qquad \text{причем} \qquad \mathbf{T}^t\mathbf{T}=\mathrm{diag}(\lambda_1,\lambda_2,...\lambda_K)=\boldsymbol{\Lambda}$$

Согласно методу ПИО можно записать, что

$$\mathbf{y}^- \le \mathbf{Ta} \le \mathbf{y}^+. \qquad \text{(A 8)}$$

Введем обозначения

$$\mathbf{z}=0.5(\mathbf{y}^- + \mathbf{y}^+) \quad , \quad \mathbf{e}=0.5(\mathbf{y}^+ - \mathbf{y}^-),$$

Тогда (A 8) можно записать в виде

$$\left|\mathbf{Ta}-\mathbf{z}\right| \le \mathbf{e} \qquad \Rightarrow \quad (\mathbf{Ta}-\mathbf{z})^t(\mathbf{Ta}-\mathbf{z}) \le \mathbf{e}^t\mathbf{e}\,.$$

раскрывая скобки и учитывая свойства матрицы **T**, получим

$$\mathbf{a}^t\boldsymbol{\Lambda}\mathbf{a} - 2\mathbf{a}^t\mathbf{X}^t\mathbf{z} + \mathbf{z}^t\mathbf{z} \le \mathbf{e}^t\mathbf{e}\,.$$

Обозначим

$$\mathbf{a}_0 = -\boldsymbol{\Lambda}^{-1}\mathbf{X}^t\mathbf{z} \quad \text{и} \quad \mathbf{u}=\mathbf{a}-\mathbf{a}_0\,, \qquad \text{(A 9)}$$

получим формулу эллипсоида в главных осях с центром $\mathbf{a}_0$

$$\mathbf{u}^t\boldsymbol{\Lambda}\mathbf{u} \le D \quad , \quad D=\mathbf{e}^t\mathbf{e}+\mathbf{z}^t\mathbf{T}\mathbf{a}_0-\mathbf{z}^t\mathbf{z} \qquad \text{(A 10)}$$

учитывая, что $\lambda_1 \ge \lambda_2 \ge ... \lambda_K > 0$, можно записать, что

$$\lambda_1 u_1^2 + \lambda_2 u_2^2 + ... + \lambda_K u_K^2 \le D\,.$$

тогда, для i=1,2,..K , величины главных осей эллипсоида вычисляются как $r_i = \sqrt{D/\lambda_i}$, и величина сдвига определяется формулой

$$\text{Shift}= - \min(s_1, s_2,...,s_K), \text{ где } s_i = -r_i + \mathbf{a}_{0i}$$

Если величина Shift<0, следовательно, все точки, принадлежащие эллипсоиду, имеют положительные координаты и сдвиг не требуется.

Литература

1. Geladi P., Esbensen K. Chemometrics, a growing and maturing discipline (Editorial). *Chemom. Intell. Lab. Syst*., **7**, 197 (1990)
2. Massart D.L. *Chemometrics: a textbook*. Elsevier, NY, 1988
3. Wold S. Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst*., **30**, 109 (1995)
4. Blanco M., Villarroya I. NIR spectroscopy: a rapid-response analytical tool. *Trends Anal. Chem*., **21**, 240 (2002)
5. Osborne B.G., Fearn T. *Near Infrared Spectroscopy in Food Analysis*. Longman Scientific and Technical, Harlow, Essex, England, 1986
6. Blanco M., Coello J., Iturriaga H., Maspoch S., Rovira E. Determination of water in ferrous lactate by near infrared reflectance spectroscopy with a fibre-optic probe. *J. Pharm. Biomed. Anal.,* **16**, 255 (1997)
7. Espinosa A., Lambert D., Valleur M. Use NIR technology to optimize plant operations. *Hydrocarbon Process,* **74**, 86 (1995)
8. Næs T., Irgens C., Martens H. Comparison of linear statistical methods for calibration of NIR instruments. *Appl. Stat.,* **35,** 195 (1986)
9. Martens H., Næs T. Multivariate calibration. I. Concepts and distinctions. *Trends Anal. Chem.,* **3**, 204 (1984)
10. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosopher Mag.,* **2** (6), 559 (1901)
11. Gosset W.S.(”Student”). The probable error of a mean. *Biometrika*, **6** , 1 (1908)
12. Fisher R.A. *Statistical methods for research workers.* Oliver and Boyd, Edinburgh,1925
13. Fisher R.A. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935
14. Налимов В. *Применение математической статистики при анализе вещества.* М, 1960
15. Wold S., Esbensen K., Geladi P. Principal component analysis. *Chemom. Intell. Lab. Syst.*, **2**, 37 (1987)
16. Shrager R.I. Chemical transitions measured by spectra and resolved using singular value decomposition. *Chemom. Intell. Lab. Syst.*, **1**, 59 (1986)
17. Geladi P., Grahn H. *Multivariate Image Analysis*. Wiley, Chichester, 1996

18. Walczak B., Massart D.L. Wavelets — something for analytical chemistry? *Trends Anal. Chem.,* **16** , 451, (1997)
19. Belousov A.I., Verzakov S.A., von Frese J. Application aspects of support vector machines. *J. Chemom.,* **16**, 482 (2002)
20. Gelad P., Esbensen K. Regression on multivariate images: Principal component regression for modeling, prediction and visual diagnostic tools. *J. Chemom.*, **5**, 97 (1991)
21. Nomikos P., MacGregor J.F. Monitoring batch processes using multiway principal component analysis. *American Inst. Chem. Engin. J.*, **40**, 1361 (1994)
22. SchaeferlingM., Schiller S. Pau H.l, Kruschina M., Pavlickova P., Meerkamp M., Giammasi C., Kambhampati D. Application of self-assembly techniques in the design of biocompatible protein microarray surfaces. *Electrophoresis*, **23**, 3097 (2002)
23. Ferreira M.M.C. 9th International Conference on Chemometrics in Analytical Chemistry (CAC-2004), Lisbon, Portugal. *J. Chemom*., **18**, 385 (2004)
24. Frank I.E., Friedman J.H. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109 (1993)
25. Wold S., Berglund A., Kettaneh N. New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P -with examples from pharmaceutical research and process modeling. *J. Chemom*., **16**, 377 (2002)
26. Molenberghs G. Biometry, Biometrics, Biostatistics, Bioinformatics,.., Bio-X. *Biometrics*, **61**, 1 (2005)
27. Шмелев А.Г. Традиционная психометрика и экспериментальная психосемантика: объектная и субъектная парадигмы анализа данных. *Вопросы Психологии*, **5**, 34 (1982)
28. Wold H. Perspectives in Probability and Statistics (papers in honour of M. S. Bartlett on the occasion of his 65 th birthday), Applied Probability Trust. В кн. *Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach.* University of Sheffield, p. 117, 1975
29. Дрейпер Н., Смит Г.. *Прикладной регрессионный анализ*, (в 2-х т.) Москва, Финансы и статистика, 1987 [N.R. Draper, H. Smith, Applied regression analysis, Wiley, N.Y.]
30. Родионова О.Е., Померанцев А.Л. Об одном методе решения обратной кинетической задачи по спектральным данным при неизвестных спектрах компонент. *Кинетика и катализ*, **45**, 485 (2004)

31. Koh H.-L., Yau W.-P., Ong P.-S., Hegde A. Current trends in modern pharmaceutical analysis for drug discovery. *Drug Discov. Today*, **8**, 889 (2003)
32. Pomerantsev A.L., Rodionova O.Ye. Hard and soft methods for prediction of antioxidants' activity based on the DSC measurements. *Chemom. Intell. Lab. Syst.*, **79**, 73 (2005)
33. Грибов Л.А. *Математические методы и ЭВМ в аналитической химии*, М. 1989
34. Siebert K.J. Chemometrics in Brewing - A Review. *J. Am. Soc. Brew. Chem.*, **59**, 147 (2001)
35. Varmuza K., Werther W., Krueger F.R., Kissel J., Schmid E.R. 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)
36. Johnson G.W., Ehrlich R. State of the art report on multivariate chemometric methods in environmental forensics. *Environ. Forensics*, **3**, 59 (2002)
37. WiseB.M., Gallagher N.B., Martin E.B. Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. *J. Chemom*., **15**, 285 (2001)
38. Brereton R.G.. *Chemometrics: Data analysis for the laboratory and chemical plant.* Wiley, Chichester, UK. 2003
39. Комарь Н.П. *Основы качественного химического анализа.* Харьков, 1955
40. Грибов Л.А., Баранов В.И., Эляшберг М.Е. *Безэталонный молекулярный спектральный анализ. Теоретические основы.* М. Едиториал УРСС, 2002
41. Эляшберг М. Экспертные системы для установления структуры органических молекул спектральными методами. *Успехи химии*, **68**, 579 (1999)
42. Марьянов Б., Зарубин А., Шумар С. Статистический анализ данных дифференцированного потенциометрического осадительного титрования бинарной смеси трех гетеровалентных ионов с помощью линейных характеристик. *Журн. аналит. химии*, **58**, 1126 (2003)
43. Вершинин В.И., Дерендяев Б.Г., Лебедев К.С. *Методы компьютерной идентификация органических соединений.* М. Академкнига, 2002
44. 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)

45. Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С. Нейронные сети как метод поиска зависимостей структура - свойство органических соединений. *Успехи химии*, **72**, 706 (2003)
46. Дворкин В.И.. *Метрология и обеспечение качества количественного химического анализа*. М. Химия, 2001
47. Карпов Ю.А., Полховская Т.М.. *Стандартизация и метрология в металлургическом производстве*. М. МИСИС, 1989
48. Власов Ю.Г., Легин А.В., Рудницкая А.М. Мультисенсорные системы типа электронный язык- новые возможности создания и применения химических сенсоров. *Успехи химии*, **75**, 141 (2006)
49. Калач А.В., Коренман Я. И., Нифталиев С.И. *Искусственные нейронные сети – вчера, сегодня, завтра.* Воронеж: Воронеж. гос. технол. акад., 2002
50. Казаков С.П., Рябенко А.А., Разумов В.Ф. 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)
51. Разумов В.Ф., Алфимов М.В. Фотохимия диарилэтиленов. *ЖНиПФ*, **46**, 28 (2003)
52. RodionovaO.Ye., Esbensen K.H., Pomerantsev A.L. Application of SIC (Simple Interval Calculation) for object status classification and outlier detection - comparison with PLS/PCR. *J. Chemom.,* **18**, 402 (2004)
53. Bystritskaya E.V., Pomerantsev A.L., Rodionova O.Ye. Non-linear regression analysis: new approach to traditional implementations. *J. Chemom.*, **14,** 667 (2000)
54. BogomolovA., McBrien M. Mutual peak matching in a series of HPLC/DAD mixture analyses. *Anal. Chim. Acta*, **490**, 41 (2003)
55. Bogomolov A., McBrien M. *Methods for Characterizing a Mixture of Chemical Compounds*, US Patent, US-2004-0126892-A1 (2004)
56. Kucheryavski S., Polyakov V., Govorov A. 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, pp. 3– 11, 2005
57. Оскорбин Н.М., Максимов А.В., Жилин С.И. Построение и анализ эмпирических зависимостей методом центра неопределенности. *Изв. АлтГУ*, **1**, 35 (1998)

58. Romanenko S.V., Stromberg, A.G., Selivanova E.V., Romanenko E.S. Resolution of the overlapping peaks in the case of linear sweep anodic stripping voltammetry via curve fitting. *Chemom. Intell. Lab. Syst.*, **73**, 7 (2004)
59. Васильева И.Е., Кузнецов А.М, Васильев И.Л, Шабанова Е.В. Калибровка методик атомно-эмиссионного анализа с компьютерной обработкой спектров. *Журн. аналит. химии,* **52**, 1238 (1997)
60. ШарафМ.А., Иллмэн Д.Л., Ковальски Б.Р. *Хемометрика*. Пер. с англ. М. Мир, 1987 [M. Sharaf, D. Illman, B. Kowalski. Chemometrics. NY: Wiley. 1986]
61. Massart D.L., Vandeginste B.G., Buydens L.M.C., De Jong, S. Lewi P.J., Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics Part A*. Elsevier, Amsterdam, 1997
62. VandeginsteB.G., Massart D.L., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics Part B*. Elsevier, Amsterdam, 1998
63. Næs T., Isaksson T., Fearn T., Davies T. *Multivariate Calibration and Classification*. Christerer, UK, 2002
64. Kramer R. *Chemometric Techniques for Quantitative Analysis*. Marcel-Dekker, 1998
65. Beebe K.R., Pell R.J., Seasholtz M.B. *Chemometrics: a Practical Guide*. Willey, N.Y., 1998
66. MalinowskiE.R. *Factor Analysis in Chemistry*. Wiley, N.Y., 2nd edn, 1991
67. Martens H., Næs T. *Multivariate calibration*. Wiley, New York. 1989
68. Höskuldsson A. *Prediction Methods in Science and Technology*. vol. 1, Thor Publishing, Copenhagen, Denmark, 1996
69. *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]
70. Марьянов Б.М. *Избранные главы хемометрики*. Томск: Из-во Том. ун-та, 2004
71. Эсбенсен К. *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]
72. Родионова О.Е., Померанцев А.Л. Хемометрика: достижения и перспективы. *Успехи химии*, **75** (4) 302-317(2006)
73. *The Unscramber*. Доступно на http://www.camo.no/ [3 мая 2007]

74. *Eigenvector Research, Inc*. Доступно на http://www.eigenvector.com/ [3 мая 2007]
75. *Umetrics*. Доступно на http://www.umetrics.com/ [3 мая 2007]
76. *SPSS*. Доступно на http://www.spss.com/ [3 мая 2007]
77. *STATISTICA*. Доступно на http://www.statsoftinc.com/ [3 мая 2007]
78. *MATLAB*. Доступно на http://www.mathworks.com/ [3 мая 2007]
79. Родионова О.Е. Хемометрический подход к исследованию больших массивов химических данных. *Рос. хим. ж. (Рос. хим. об-ва им . Д.И. Менделеева)*, **50**, 128 (2006)
80. ErikssonL., Johansson E., Kettaneh-Wold N., Wold S. *Multi- and Megavariate Data Analysis*. Umetrics, Umeå, 2001
81. Sanchez E., Kowalski B.R. Tensorial calibration: I. First-order calibration. *J. Chemom.*, **2** 247 (1988)
82. Smilde A., Bro R., Geladi P. *Multi-way Analysis with Applications in the Chemical Sciences*. John Wiley & Sons, Chichester, 2004
83. Hoy M., Steen K., Martens H. Review of partial least squares regression prediction error in Unscrambler. *Chemometrics Intell. Lab. Syst.,* **44,** 123 (1998)
84. Wold S., Trygg J., Berglund A., Antti H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.,* **58** , 131 (2001)
85. Höskuldsson A. Causal and path modelling. Hyperspectral imaging: calibration problems and solutions. *Chemom. Intell. Lab. Syst*., **58**, 287 (2001)
86. Geladi P., Burger J., Lestanderet T. Hyperspectral imaging: calibration problems and solutions. *Chemom. Intell. Lab. Syst*., **72**, 209 (2004)
87. Sander G.H.W., Manz A. Chip-based microsystems for genomic and proteomic analysis. *Trends Anal. Chem*., **19**, 364 (2000)
88. Box G.E.P., Hunter W.G., Hunter J.S.. *Statistics for Experimenters*. John Wiley & Sons Inc., NY, 1978
89. Демиденко Е.З. *Линейная и нелинейная регрессии*. Финансы и статистика, М, 1981
90. Jy P. *Sampling for Analytical Purposes*. John Wiley & Sons, Chichester, 1989
91. Kleingeld W., Ferreira J., Coward S. First World Conference on Sampling and Blending (WCSB1). *J. Chemom*., 18, 121 (2004)
92. 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 (2004)

93. Walczak B., Massart D.L. Tutorial. Dealing with missing data. *Chemom. Intell. Lab. Syst*., **58**, 15 (2001)
94. Nelson P.R.C., Taylor P.A., MacGregor J.F. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemom. Intell. Lab. Syst*., **35**, 45 (1996)
95. Haario H., Taavitsainen V.-M. Combining soft and hard modelling in chemical kinetic models. *Chemom. Intell. Lab. Syst*., **44**, 77 (1998)
96. Брин Э.Ф., Померанцев А.Л.. Классификация обратных задач кинетики гомогенных химических реакций. *Хим. физика*, **5**, 1674 (1986)
97. Gurden S.P., Westerhuis J.A., Bijlsma S., Smilde A.K. Modelling of spectroscopic batch process data using grey models to incorporate external information. *J. Chemom*., **15**, 101 (2001)
98. Померанцев А.Л. *Методы нелинейного регрессионного анализа для моделирования кинетики химических и физических процессов*. Дис. д-ра физ.-мат. наук, ИХФ РАН, Москва, 2003
99. Morales D. A. Mathematical modeling of titration curves. *J. Chemom*., **16**, 247 (2002)
100. de Juan A., Maeder M., Martınez M., Tauler R. Combining hard- and soft-modelling to solve kinetic problems. *Chemom. Intell. Lab. Syst*., **54**, 123 (2000)
101. Карпухин О.Н. 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/ [3 мая 2007]
102. Эфрон Б. *Нетрадиционные методы многомерного статистического анализа*. Москва, Финансы и Статистика, 1988 [B. Efron, *Ann. Stat.* , **7**, 1 (1979)].
103. *EURACHEM/CITAC Guide, Quantifying Uncertainty in Analytical Measurement*, 2nd ed., EURACHEM, Lisbon, Portugal, 2000
104. Faber K, Kowalski B. R. Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler. *Chemom. Intell. Lab. Syst.*, **34**, 283 (1996)
105. Pomerantsev A.L. Confidence Intervals for Non-linear Regression Extrapolation. *Chemom. Intell. Lab. Syst.*, **49**, 41 (1999)
106. Pulido A., Ruisánchez I., Boqué R., Rius F.X. Uncertainty of results in routine qualitative analysis. *Trends Anal. Chem*., **22**, 647 (2003)

107. Vershinin V.I. A priori method of evaluating uncertainties in qualitative chromatographic analysis:(probabilistic approach). *Accreditation and Quality Assurance*, **9**, 415 (2004)
108. Faber N.M. Uncertainty estimation for multivariate regression coefficients. *Chemom. Intell. Lab. Syst.*, **64**, 169 (2002)
109. Faber N.M.,. Bro R. Standard error of prediction for multiway PLS 1. Background and a simulation study. *Chemom. Intell. Lab. Syst.*, **61**, 133 (2002)
110. Olivieri A. C., Faber N.M., Ferre J. , Boque R., Kalivas J. H., Mark H. Ucertainy estimation and figures of merit for multivariate calibration. *Pure Appl.Chem.*,**78**, 633 (2006)
111. Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.,* **58**, 1167 (1986)
112. Ferré J., Faber N.M. Net analyte signal calculation for multivariate calibration. *Chemom. Intell. Lab. Syst.*, **69**, 123 (2003)
113. Boqué R., Faber N.M., Xavier Rius F. Detection limits in classical multivariate calibration models. *Anal. Chim. Acta,* **423**, 41 (2000)
114. Boqué R., Faber N.M., Xavier Rius F. Limit of detection estimator for second-order bilinear calibration. *Anal. Chim. Acta*, **451**, 313 (2002)
115. Berget I., Næs T. Using unclassified observations for improving classifiers. *J. Chemom.*, **18**, 103 (2004)
116. Jouan-Rimbaud D., Massart D.L., Saby C.A., Puel C. Characterization of the representativity of selected sets in multivariate calibration and pattern recognition. *Anal. Chim. Acta*, **350**, 149 (1997)
117. Meloun M., Militký J., Hill M., Brereton R.G. Crucial problems in regression modelling and their solutions. *Analyst*, **127**, 433 (2002)
118. Fernandez Pierna J.A., Wahl F., de Noord O.E., Massart D.L. Methods of outlier detection in prediction. *Chemom. Intell. Lab. Syst*., **63**, 27 (2002)
119. K. Faber. Comparison of two recently proposed expressions for partial least squares regression prediction error. *Chemom. Intell. Lab. Syst.*, **52**, 123 (2000)
120. Faber N.M., Song X.-H., Hopke P.K. Sample-specific standard error of prediction for partial least squares regression. *Trends Anal. Chem.*, **22**, 330 (2003)
121. Bouveresse E., Massart D.L. Standardization of near-infrared spectrometric instruments: A review. *Vibrat. Spectrosc.*, **11**, 3 (1996)

122. Westad F., Martens H. Variable selection in NIR based on significance testing in Partial Least Squares Regression (PLSR). *J. Near Infrared Spectros.*, **8**, 117 (2000)
123. Hubert M., Verboven S. A robust PCR method for high-dimensional regressors. *J. Chemom.*, **17**, 438 (2003)
124. Bro R., Smilde A.K. Centering and scaling in component analysis. *J. Chemom.*, **17**, 16 (2003)
125. Kubelka P., Munck F. Ein Beitrag zur Optik der Farbanstriche. *Zeits. F. techn. Physik,* **12**, 593 (1931)
126. Savitzky A., Golay M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, **36**, 1627 (1964)
127. Geladi P., MacDougall D., Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.*, **3**, 491 (1985)
128. Isaksson T., Kowalski B. Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products. *Appl. Spectrosc.*, **47**, 702 (1993)
129. Trygg J., Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.*, **17**, 53 (2003)
130. Wold S., Antti H., Lindgren F., Öhman J. Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.*, **44**, 175 (1998)
131. Fearn T. On orthogonal signal correction. *Chemom. Intell. Lab. Syst.*, **50**, 47 (2000)
132. Höskuldsson A. Variable and subset selection in PLS regression. *Chemom. Intell. Lab. Syst.*, **55**, 23 (2001)
133. Guo Q., Wu W., Massart D.L., Boucon C., de Jong S. Feature selection in principal component analysis of analytical data. *Chemom. Intell. Lab. Syst.*, **61**, 123 (2002)
134. Forina M., Lanteri S., Oliveros M.C. Selection of useful predictors in multivariate calibration. *Anal. Bioanal. Chem.,* **380**, 397 (2004)
135. Leardi R, Boggia R., Terrile M. Genetic algorithms as a strategy for feature selection. *J. Chemom.,* **6**, 267 (1992)
136. Kalivas J.H. Pareto calibration with built-in wavelength selection. *Anal. Chim. Acta,* **505**, 9 (2004)
137. Benoudjit N., Cools E., Meurens M., Verleysen M. Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear models. *Chemom. Intell. Lab. Syst*., **70**, 47 (2004)

138. Indahl U., Næs T. A variable selection strategy for supervised classification with continuous spectroscopic data. *J. Chemom.*, **18**, 53 (2004)
139. Feudale R.N., Woody N.A., Tan H., A.J. Myles, S.D. Brown, J. Ferré. Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst*., **64**, 181 (2002)
140. SulimaE.L., Zubkov V.A., Rusinov L.A. 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, pp. 196–203, 2005
141. 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)
142. Hansen P.W. Pre-processing method minimizing the need for reference analyses. *J. Chemom.,* **15**, 123 (2001)
143. Чуи К. *Введение в вэйвлеты*. М. Мир. 2001 [C.K. Chui. An Introduction to wavelets, Academic Press, 1992]
144. Trygg J., Wold S. PLS regression on wavelet compressed NIR spectra. *Chemom. Intell. Lab. Syst.*, **42**, 209 (1998)
145. Reinikainen S.-P. Wavelets in Compressing Spectral Data В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, pp. 21-36, 2005
146. Pan Y., Yoo C.K., Lee J.H., Lee I.-B. Process monitoring for continuous process with periodic characteristics. *J. Chemom.,* **18**, 69 (2004)
147. Keller H.R., Massart D.L. Evolving factor analysis. *Chemom. Intell. Lab. Syst.*, **12**, 209 (1992)
148. Malinowsk E.R. Window Factor Analysis: theoretical derivation and application to ow-injection analysis data. *J. Chemom.*, **6**, 29 (1992)
149. Gemperline P.J. Target transformation factor analysis with linear inequality constraints applied to spectroscopic-chromatographic data. *Anal. Chem.*, **58**, 2656 (1986)
150. Wold S. Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, **8**, 127 (1976)
151. Jiang J.-H., Liang, Y. Ozaki Y. Principles and methodologies in self-modeling curve resolution. *Chemom. Intell. Lab. Syst.*, **71**, 1 (2004)

152. Sanchez F.C., van de Borgaert B., Rutan S.C., Massart D.L. Multivariate peak purity approaches. *Chemom.Intell. Lab. Syst.*, **34**, 139 (1996)

153. Shen H., Grande B., Kvalheim O.M., Eide I. Automated curve resolution applied to data from multi-detection instruments. *Anal. Chim. Acta,* **446**, 311 (2001)

154. WindigW., Guilment J. Interactive self-modeling mixture analysis. *Anal. Chem.*, **63**, 1425 (1991)

155. Bogomolov A., Hachey M., Williams A. Software for interactive curve resolution using SIMPLISMA. В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, pp. 119-135, 2005

156. Diewok J., de Juan A., Marcel M., Tauler R., Lendl B. Application of a Combination of Hard and Soft Modeling for Equilibrium Systems to the Quantitative Analysis of pH-Modulated Mixture Samples. *Anal. Chem.*, **76**, 641 (2003)

157. БогомоловА.Ю., Ростовщикова Т.Н., Смирнов В.В. Комплексообразование хлорида железа(III) с хлористым водородом и водой в хлорорганическом растворителе. *Ж. Физ. Хим*., **69,** 1197 (1995)

158. Seipel H.A., Kalivas J.H. Effective rank for multivariate calibration methods. *J. Chemom.*, **18**, 306 (2004)

159. Shrager R.I. Chemical transitions measured by spectra and resolved using singular value decomposition. *Chemom. Intell. Lab. Syst.*, **1**, 59 (1986)

160. De Maesschalck R., Jouan-Rimbaud D., Massart D.L. Tutorial. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.*, **50**, 1 (2000)

161. Andrade J.M., Gomez-Carracedo M. P., Krzanowski W., Kubista M. Procrustes rotation in analytical chemistry, a tutorial. *Chemom. Intell. Lab. Syst.*, **72**, 123 (2004)

162. MikhailovE.V., Tupicina O.V., D.E. Bykov, Chertes K.L., Rodionova O.Ye., Pomerantsev A.L. Ecological assessment of landfills with multivariate analysis — A feasibility study. *Chemom. Intell. Lab.Syst*., **88** (1), 3-10 (2007)

163. RodionovaO.Ye., Houmøller L.P., Pomerantsev A.L., Gelad P., Burger J., Dorofeyev V.L., Arzamastsev A.P. NIR Spectrometry for Counterfeit Drug Detection. A Feasibility Study. *Anal. Chim. Acta*, **549**, 151 (2005)

164. SunL.X., Danzer K. Fuzzy cluster analysis by simulated annealing. *J. Chemom.*, **10**, 325 (1996)

165. Myle A.J., Brown S.D. Induction of decision trees using fuzzy partitions. *J. Chemom.*, **17**, 531 (2003)

166. González-ArjonaD., López-Pérez G., González A.G. Performing procrustes discriminant analysis with HOLMES. *Talanta,* **49**, 189 (1999)
167. Mark H. Use of Mahalanobis distances to evaluate sample preparation methods for near-infrared reflectance analysis. *Anal. Chem*., **59**, 790 (1987)
168. Gemperline P.J., Boyer N.R. Classification of near-infrared spectra using wavelength distances: comparison to the Mahalanobis distance and residual variance methods. *Anal. Chem.*, **67**, 160 (1995)
169. Mark H.L., Tunnell D. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Anal. Chem.*, **57**, 1449 (1985)
170. Indahl U., Sing N.S., Kirkhuus B., Næs T. Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise. *Chemom.Intell. Lab. Syst.*, **49**, 19 (1999)
171. Downey G., Boussion J., Beauchene D. Authentication of whole and ground coffee beans by near infrared reflectance spectroscopy. *J.Near Infrared Spectrosc.*, **2**, 85 (1994)
172. Flåten G.R., Grung B., Kvalheim O.M. A method for validation of reference sets in SIMCA modelling. *Chemom. Intell. Lab. Syst.*, **72**, 101 (2004)
173. Næs T., Indahl U. A unified description of classical classification methods for multicollinear data. *J. Chemom.*, **12**, 205 (1998)
174. McElhinney J., Downey G., Fearn T. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *J.Near Infrared Spectrosc.*, **7**, 145 (1999)
175. Zomer S., Brereton R., Carter J.F., Eckers C. 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)
176. Zernov V.V., Balakin K.V., Ivaschenko A.A., Savchuk N.P., Pletnev I.V. 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)
177. Sarker M., Rayens W. Partial least squares for discrimination. *J. Chemom*., **17**, 166 (2003)
178. Herrero A., Zamponi S., Marassi R.,. Conti P, Ortiz M.C., Sarabia L.A. Determination of the capability of detection of a hyphenated method: application to spectroelectrochemistry. *Chemom. Intell. Lab. Syst.*, **61**, 63 (2002)

179. Garcia I., Sarabia L., Ortiz M. C., Aldama J. M. 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)
180. Bijlsma S., Smilde A.K. Estimating reaction rate constants from two-step reaction: a comparison between two-way and three-way methods. *J. Chemom.*, **14**, 541 (2000)
181. Bro R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.*, **38**, 149 (1997)
182. Kiers H. Some procedures for displaying results from three-way methods. *J. Chemom.*, **14**, 151 (2000)
183. Faber N.M., Bro R., Hopke P.K. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemom. Intell. Lab. Syst.*, **65**, 119 (2003)
184. Andersson C.A. , Bro R. The N-way toolbox for MATLAB. *Chemom. Intell. Lab. Syst.*, **52** , 1 (2000)
185. del Rio F.J., Riu J., Rius F.X. Prediction intervals in linear regression taking into account errors on both axes. *J. Chemom.*, **15**, 773 (2001)
186. Brereton R.G. Introduction to multivariate calibration in analytical chemistry. *Analyst*, **125**, 2125 (2000)
187. Höskuldsson A. PLS Regression Methods. *J. Chemom.*, **2**, 211 (1988)
188. de Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.*, **18**, 251 (1993)
189. Li B., Morris A.J.,. Martin E.B. Generalized partial least squares regression based on the penalized minimum norm projection. *Chemom. Intell. Lab. Syst.*, **72**, 21 (2004)
190. Hubert M.,Vanden Branden K. Robust methods for partial least squares regression. *J. Chemom.,* **17**, 537 (2003)
191. Vigneau E., Devaux M., Qannari M., Robert P. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *J. Chemom.,* **11**, 239 (1997)
192. Geladi P. Some recent trends in the calibration literature. *Chemom. Intell. Lab. Syst*., **60**, 211 (2002)
193. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений. *Сиб. мат. журн.*, **3** , 701 (1962)

194. Rodionova O. Ye., Pomerantsev A.L. Principles of Simple Interval Calculations, В кн: *Progress in Chemometrics Research,* Pomerantsev AL (ed.). Nova Science Publishers: New York, pp43-64, 2005
195. Белов В.М., Суханов В.А., Унгер Ф.Г. *Теоретические и прикладные аспекты метода центра неопределенности*. Новосибирск, Наука, 1995
196. Bro R. Multi-way calibration. Multi-linear PLS. *J. Chemom.*, **10**, 47 (1996)
197. Bro R., Andersson C.A.. *The N-way Toolbox for MATLAB, Version 2.02, 2003*. Доступно на http://www.models.kvl.dk/source [3 мая 2007]
198. Ni Y., Huang C., Kokot S. Application of multivariate calibration and artificial neural networks to simultaneous kinetic-spectrophotometric determination of carbamate pesticides. *Chemom. Intell. Lab. Syst.,* **71**, 177 (2004)
199. Chen Z.P., Morris J., Martin E., Yu R.-Q., Liang Y.-Z., Gong F. Recursive evolving spectral projection for revealing the concentration windows of overlapping peaks in two-way chromatographic experiments. *Chemom. Intell. Lab. Syst*., **72**, 9 (2004)
200. Fernández F. M., Tudino M. B., Troccoli O. E. 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)
201. Бард Й. *Нелинейное оценивание параметров.* М.: Статистика, 1979. [Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974]
202. Barry D.M., Meites L. 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)
203. Марьянов Б. В кн. *Химики ТГУ на пороге третьего тысячелетия*. Томск, Изд -во ТГУ, сс. 48-58, 1998
204. Berglund A., Wold S. INLR, implicit non-linear latent variable regression. *J. Chemom.*, **11**, 141 (1997)
205. Berglund A., Kettaneh N.L.U., Wold S., Bendwell N.,Cameron D.R. The GIFI approach to non-linear PLS modelling. *J. Chemom.*, **15**, 321 (2001)
206. Wold S. Nonlinear partial least squares modelling II. Spline inner relation. *Chemom. Intell. Lab. Syst.*, **14**, 71 (1992)
207. Zupan J., Gasteiger J. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta*, **248**, 1 (1991)

208. Zupan J., Gasteiger J. *Neural Network for Chemists, An Introduction*. VCH, Weinheim, 1993.
209. Wu W., Walczak B., Massart D.L., Heuerding E., Erni F.E., Last I.R., Prebble K.A. Artificial neural networks in classificafion of NIR spectral data: Design of the training set. *Chemom. Intell. Lab. Syst.*, **33**, 35 (1996)
210. Smits J.R.M., Melssen W.J., Buydens L.M.C., Kateman G. Using artificial neural networks for solving chemical problems. Part I. Multi-layer feed-forward networks. *Chemom. Intell. Lab. Syst*., **22,** 165 (1994)
211. Melssen W.J., Smits J.R.M., Buydens L.M.C., Kateman G. 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)
212. Hibbert D.B. Genetic algorithms in chemistry. *Chemom. Intell. Lab. Syst.*, **19**, 277, (1993)
213. Leardi R. Genetic algorithms in chemometrics and chemistry: a review. *J. Chemom*., **15**, 559 (2001)
214. Shao X., Chen Z., Lin X. Resolution of multicomponent overlapping chromatogram using an immune algorithm and genetic algorithm. *Chemom. Intell. Lab. Syst.*, **50**, 91 (2000)
215. Tortajada-Genaro L.A., Campíns-Falcó P., Verdú-Andrés J., Bosch-Reig F. Multivariate versus univariate calibration for nonlinear chemiluminescence data. Application to chromium determination by luminol-hydrogen peroxide reaction. *Anal. Chim. Acta*, **450**, 155 (2001)
216. Вощинин А.П., Бочков А.Ф., Сотиров Г.Р. Методы анализа данных при интервальной нестатистической ошибке. *Завод. лаб*., **56**, 76 (1990)
217. Анисимов В.М., Померанцев А.Л., Новорадовский А.Г., Карпухин О.Н. Определение чувствительности объектов к полихроматическому световому воздействию. *Журн. прикл. спектрос*, **46,** 117 (1987)
218. Спивак С.И., Тимошенко В.И., Слинько М.Г. Применение метода выравнивания по П.Л. Чебышеву при построении кинетических модели сложной химической реакции. *Докл. АН СССР,* **192**, 580 (1970)
219. Слинько М.Г., Спивак С.И., Тимошенко В.И. О критериях определения параметров кинетических моделей. *Кинетика и катализ*, **13**, 1570 (1972)
220. Ахунов И.Р., Ахмадишин З. Ш., Спивак С.И. Математическая интерпретация кинетического эксперимента сложных реакций сопряженного окисления. *Хим. физика*, **12**, 1660 (1982)

221. Бахитова Р.Х. ,Спивак С.И. Нечеткие интервальные оценки в кинетике химических реакций. *Химия и хим. технол.*, **42**, 92 (1999)
222. Хлебников А.И. О методе центра неопределенности. *Ж. Аналит. химии,* **51,** 347 (1996)
223. Померанцев А.Л., Родионова О.Е. О двух подходах к анализу кинетических данных на примере предсказания активности антиоксидантов. *Кинетика и катализ,* **47**, 553 (2006)
224. Cook R.D. Detection of Influential Observations in Linear Regression. *Technometrics*; **19,** 15 (1977)
225. Cook R.D. Influential Observations in Linear Regression *J. Am. Statis. Ass.*, **74,** 169 (1979)
226. Andrews D.F., Pregibon D. Finding the outliers that matter. *J. Royal Statis. Soc.B*, **40,** 84 (1978)
227. Draper N.R., John J. A Influential Observations and Outliers in Regression. *Technometrics*, **23,** 21 (1981)
228. Næs T. The design of calibration in near infra-red reflectance analysis by clustering. *J. Chemometrics*, **1,** 121 (1987)
229. Померанцев А.Л., Родионова О.Е. Построение многомерной калибровки методом простого интервального оценивания. *Ж. Аналит. химии*, **61**, 1032 (2006)
230. Clancey V.J. Statistical methods in chemical analyses. *Nature*, **159,** 339 (1947)
231. Rajkó R. Treatment of model error in calibration by robust and fuzzy procedures. *Anal. Letters*, **27,** 215 (1994)
232. Боровков А.А. *Математическая статистика (Оценки параметров. Проверка гипотез).* М.: Наука, 1984
233. Гумбель Э. *Статистика экстремальных значений*. М. Мир, 1965 [Gumbel E Statistics of extremes*,* Columbia University Press: N.Y., 1962.]
234. Gass S. *Linear Programming* (4-th ed.). McGow-Hill: New York, 1975
235. Kuhn H.W., Tucker A.W. Linear Inequalities and Related Systems. *Ann. Math. Studies,* **38**, Princeton University Press: Princeton, N.J., 1956
236. Lehmann E.L. *Testing Statistical Hypotheses*. Wiley, New York, 1960
237. Eicker F. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Stat.*, **34**, 447 (1963)

238. Данциг Дж. *Линейное программирование, его применение и обобщение.* М. Прогресс, 1966 [Dantzing G.B. linear Programming and Extensions, Princeton University Press, Princeton, New Jersey, 1963]
239. Таха Х. *Введение в исследование операций*. М. Мир, 1985, т.1. [Taha H., Operations Research. An Introduction, (3-d ed), vol.1, MacMillan Publishing Co., N. Y., 1982]
240. Cook R.D. Detection of influential observations in linear regression. *Technometrics*, **21**, 15 (1977)
241. Belsley D.A., Kuh E., Welsch R.E. *Regression diagnostics. Identifying influential data and sources of collinearity.* John Wiley & Sons Ltd. New York, 1980
242. Walkenbach J. *Excel 2000 Power Programming with VBA*, 2001
243. Esbensen K.H., Halstensen M, Lied T.T., Saudland A, Svalestuen J., .de Silva S, Hope B. Acoustic chemometrics – from noise to information. *Chemom. Intell.Lab.Syst*., **44,** 61 (1998)
244. Лаврентьев М. М., Краева А.Г., Бухгейм А.В. *Обратная задача химической кинетики*. Новосибирск: ВЦ СО АН СССР, 1980
245. Спивак С.И, Горский В.Г. Неединственность решения задачи восстановления кинетических констант. *Докл. АН СССР*, **275**, 412 (1981)
246. *Применение вычислительной математики в химической и физической кинетике*. Под ред. Л.С. Полака, М.: Наука, 1969
247. Павлов Б.В., Родионова О.Е. Математическое моделирование сложных самоускоряющихся реакций. *Теор. основы хим. технологии*, **28**, 251 (1994)
248. Павлов Б.В., Родионова О.Е. Проблемы математического моделирования в неравновесной теории химических процессов. *Хим. физика*., **17**, 27 (1998)
249. Павлов Б.В., Брин Э.Ф. Обратная задача химической кинетики. *Хим. физика*, **3**(3), 393 (1984)
250. Павлов Б.В., Родионова О.Е. Численное решение систем линейных обыкновенных дифференциальных уравнений с постоянными коэффициентами. *Ж. вычил. матем. и матем. физ*., **34**, 622 (1994)
251. Павлов Б.В., Родионова О.Е. Методика усреднения при дискретизации кинетического интегро-дифференциального уравнения. *Ж. вычил. матем. и матем. физ*., **36**, 143 (1996)
252. Bijlsma S., Louwerse D.J., Windig W., Smilde A.K .Rapid estimation of rate constants using on-line SW-NIR and trilinear models. *Anal.Chim.Acta,* **376**, 339 (1998)

253. Pomerantsev A.L., Rodionova O.Ye. Chemometrics in Russia. *Chemom. Intell. Lab. Syst*, **48**, 121 (1999)
254. Pomerantsev A.L., Rodionova O.Ye. Prediction of antioxidants activity using DSC measurements. A feasibility study. В сб.: Zaikov et al (Eds) *Aging of polymers, polymer blends and polymer composites*, **1**, NovaScience Publishers, NY, pp.19-29, 2002, .
255. Shlyapnikov Yu.A In: *Development in Polymer Stabilization*, Applied Science Publishers, London, **5**, 1 (1981)
256. Pomerantsev A.L. Successive Bayesian estimation of reaction rate constants from spectral data *Chemometrics Intell.Lab.Syst*., **66** , 127 (2003)
257. Померанцев А.Л., Родионова О.Е. Содержательный и формальный подход к анализу кинетических данных. В сб. *Химическая и биологическая кинетика. Новые горизонты*. М. Химия, 2005 (ISBN: 5-98109-035-9), **1**, 124-172
258. Bystritskaya E.V., Pomerantsev A.L., Rodionova O.Ye. *Conferentia chemometrica (CC-97),* Budapest, Pos. 54, (1997)
259. *Fitter Solutions* [On line], http://polycert.chph.ras.ru/solution.htm [3 мая 2007]
260. Павлов Б.В., Повзнер А.Я. Об одном методе численного интегрирования систем обыкновенных дифференциальных уравнений. *Ж. Вычисл. Мат. Мат. Физ.*, **13**, 1056 (1973)
261. Marquardt D.W. An algorithm for least-squares estimation of non-linear parameters. *SIAM J.,* **11**, 431 (1963)
262. Levenberg K. A Method for the Solution of Certain Problems in Last Squares. *Quart. Appl. Math*., **2**, 164 (1944)
263. Pomerantsev A.L. Confidence intervals for nonlinear regression extrapolation. *Chemom. Intell.Lab.Syst.,* **49**, 41 (1999)
264. Родионова О.Е.,Померанцев А.Л. Оценивание параметров в уравнении Аррениуса. *Кинетика и катализ,* **46**, 329 (2005)
265. Counterfeit drugs. *Guidelines for the development of measures to combat counterfeit drugs*, WHO, Geneva (1999)
266. Counterfeit drugs: threat to Public Health. *55 World Health Assembly*, Geneva (2002)
267. Arzamastsev A.P., Dorofeyev V.L., Kochin V.Yu, et al. Using Thin-Layer Chromatography for the Fast Identification of the Fluoroquinolone Drugs. *World Congress of Pharmacy and Pharmaceutical Sciences 2002: 62nd International Congress of FIP*, **39** (2002)

268. Arzamastsev A.P., Dorofeyev V.L., Konovalov A.A., KochinV.Yu., Titov I.V. Determining Adulterated Drugs by Modern Analytical Techniques. *Pharmaceutical Chemistry Journal*, **38**, 166 (2004)
269. Shewhart W.A. *Economic Control of Quality of Manufactured Product*, Van Nostrand, New York, 1931
270. MacGregor J., Kourti Th. Statistical process Control of Multivariate Processes. *Control Engineering Practice*, **3**, 403 (1995)
271. Померанцев А.Л., Родионова О.Е. Многомерный статистический контроль процессов. *Методы менеджмента качества*, **6**, 15 (2002)
272. Kourti Th., MacGregor J. Process analysis, monitoring and diagnosis, using multivariate projection methods. Tutorial. *Chemom. Intell. Lab. Syst.*, **28**, 3, (1995)
273. Westerhuis J.A., Kourti Th., MacGregor J. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.*, **12**, 301 (1998)
274. Höskuldsson A., Rodionova O.Ye., PomerantsevA.L. Path Modelling and Process Control, *Chemom. Intell. Lab.Syst*., **88**,84 (2007)
275. Pomerantsev A.L., Rodionova O.Ye. Multivariate statistical process control and optimisation. В кн: *Progress in Chemometrics Research* (Ed: A.L. Pomerantsev) NovaScience Publishers, New York, pp. 209-227, 2005
276. Gabrielsson J., Lindberg N-O., Lundstedt T. Multivariate methods in pharmaceutical applications. *J. Chemom.*, **16**, 141 (2002)
277. Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemom. Intell. Lab. Syst.*, **46**, 133 (1999)
278. Yoo C.K.,Lee J.-M., Vanrolleghem P.A., Lee I.-B. On-line monitoring of batch processes using multiway independent component analysis. *Chemom. Intell. Lab. Syst.*, **71**, 151 (2004)
279. Baroni M., Benedetti P., Fraternale S. Scialpi, F., Vix P., Clementi S. The CARSO procedure in process optimization. *J. Chemom.*, **17**, 9 (2003)
280. Martens H., Martens M. *Multivariate Analysis of Quality: An Introduction,* John Wiley & Sons Ltd., Chichester, 2001
281. Dyson R.M., Hazenkamp M., Kaufmann K.,. Maeder M, Studer M., Zilian A. Modern tools for reaction monitoring: hard and soft modelling of non-ideal, on-line acquired spectra. *J. Chemom.*, **14**, 737 (2000)

282. Pöllänen K., Häkkinen A., Reinikainen S.-P., Louhi-Kultanen M., Nyström L. ATR-FTIR in monitoring of crystallization processes: comparison of indirect and direct OSC methods. *Chemom. Intell. Lab. Syst.,* **76**, 25 (2005)
283. Thurston T.J., Brereton R.G., Foord D.J., Escott R.E.A. Principal components plots for exploratory investigation of reactions using ultraviolet-visible spectroscopy: application to the formation of benzophenone phenylhydrazone. *Talanta*, **63**, 757 (2004)
284. Bezemer E., Rutan S.C. Multivariate curve resolution with non-linear fitting of kinetic profiles. *Chemom. Intell. Lab. Syst.*, **59**, 19 (2001)
285. Workman Jr., Creasy K.E., Doherty S., Bond L., Koch M., Ullman A., Veltkamp D.J. Process analytical chemistry. *Anal. Chem.*, **73**, 2705 (2001)
286. Gurden S.P., Martin E.B., Morris A.J. The introduction of process chemometrics into an industrial pilot plant laboratory. *Chemom. Intell. Lab. Syst.*, **44**, 319 (1998)
287. *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
288. *ASTM Standard E1655*. Standard Practices for Infrared Multivariate Quantitative Analysis, 1997
289. Pomerantsev A.L., Rodionova O.Ye., Höskuldsson A. Process control and optimization with simple interval calculation method. *Chemom. Intell. Lab.Syst*., **81** (2), 165 (2006)
290. Wagen L.E., Kowalski B. A multiblock partial least squares algorithm for investigation complex chemical systems. *J. Chemometrics*, **3,** 3 (1998)
291. Dantas Filho H.A., Harrop Galvao R.K., Ugulino Araujo M.C., da Silva C., Bezerra Saldaha T.C., Jose G.E., Pasquini C., Raimundo I.M., Rodrigues Rohwedder J.J. A strategy for selecting calibration samples for multivariate modelling. *Chemometrics Intell. Lab. Syst.*, **72,** 83 (2004)
292. Kennard R.W., Stone L.A. Computer Aided Design of Experiment. *Technometrics*, **11,** 137 (1969)
293. Федоров В.В. *Теория оптимального эксперимента.* Наука, Москва, 1971
294. Rodionova O.Ye., Pomerantsev A.L. Application of simple interval calculation method for representative subset selection. Тез.докл. на международной конф. *ICAS 2006,* Москва 2006

295. Rajer-Kanduc K.,Zupan J., Majcen N. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemometrics Intell. Lab. Syst,* **65,** 221 (2003)
296. Andersson P. M., Sjostro M., Wold S., Lundstedt T. Strategies for subset selection of parts of an in-house chemical library. *J. Chemometrics*, **15,** 353 (2001)
297. Cruciani G., Baroni M., Carosati E., Clementi M., Valigi R., Clementi S. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J. Chemometrics*, **18,** 146 (2004)