На правах рукописи

РОДИОНОВА ОКСАНА ЕВГЕНЬЕВНА

ИНТЕРВАЛЬНЫЙ МЕТОД ОБРАБОТКИ РЕЗУЛЬТАТОВ МНОГОКАНАЛЬНЫХ ЭКСПЕРИМЕНТОВ

01.04.01 – Приборы и методы экспериментальной физики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

доктора физико-математических наук

Москва 2008

Работа выполнена в Институте химической физики
им Н.Н. Семенова Российской Академии Наук

Официальные оппоненты: доктор физико-математических наук
Новиков Лев Васильевич

доктор физико-математических наук
профессор Спивак Семен Израилевич

доктор технических наук
профессор Русинов Леон Абрамович

Ведущая организация Учреждение Российской академии наук
Институт геохимии и аналитической
химии им. В.И. Вернадского РАН

Защита состоится « 20 » февраля 2009 г. в « 15 » часов на заседании диссертационного совета Д 002.034.01 при Институте аналитического приборостроения Российской Академии наук по адресу: 190103 С.-Петербург, Рижский пр. 26.

С диссертацией можно ознакомиться в научно-технической библиотеке Института аналитического приборостроения РАН по тому же адресу

Автореферат разослан « 13 » января 2009 г.

Ученый секретарь
диссертационного совета
кандидат физико-математических наук А.П. Щербаков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. По мере совершенствования приборной базы, экспериментальная физика начинает оперировать с большими массивами данных, которые содержат измерения сотен и тысяч объектов, при учете большого числа действующих факторов. Математическая обработка становится неотъемлемой составляющей сложного физического эксперимента. В некоторых случаях, именно методы извлечения полезной информации из экспериментальных результатов способствуют распространению той или иной техники эксперимента. Начиная с 70-х годов прошлого века, для анализа подобных данных используется хемометрический подход, суть которого состоит в двух принципах. Во-первых, это понижение размерности задачи с помощью проекционных методов, и, во-вторых, это использование формальных, линейных моделей для объяснения связей в данных. Ярким примером может служить инфракрасная спектроскопия в ближней области, которая до появления хемометрического подхода почти не применялась и считалась малоперспективной. Однако, использование такого подхода связано с двумя проблемами, которые не нашли еще своего окончательного решения. Во-первых, – это оценка неопределенности получаемых результатов, а во-вторых, – ограничение области применимости методов. На решение этих двух важных задач и направлено это исследование.

Главной идеей работы является интервальный подход, т.е. последовательное использование принципа ограниченности погрешностей. Этот принцип был впервые предложен Л. Канторовичем еще в 1962 г, однако до сих пор эта идея не получила должного признания и развития. В представленной работе главное внимание уделяется классической проблеме количественного анализа – решению задачи калибровки. При этом интервальный подход сочетается с проекционными методами, что дает в результате интервальный прогноз искомого физико-химического показателя. Этот метод назван простым интервальным оцениванием (ПИО). При его применении удается решить две задачи: установить область неопределенности прогноза и построить классификацию объектов, которая позволяет очертить область применения построенной ПИО модели.

Цель работы состоит:

– в разработке теоретических и прикладных аспектов интервального анализа результатов экспериментов. В том числе: построение интервальных моделей линейной калибровки, оценка индивидуальной неопределенности прогноза, создание системы классификации объектов, определение области применения построенных моделей;

– в написании алгоритмов обработки многоканальных сигналов и создании компьютерной системы анализа результатов эксперимента, позволяющей реализовать потенциальные возможности измерительных систем и приборов;

– в построении методологии совместного применения проекционных методов и ПИО при решении важных теоретических и практических задач интерпретации больших наборов данных многоканальных экспериментов.

Научная новизна работы определяется следующими результатами:

1) Предложен новый метод построения линейных калибровочных зависимостей – простое интервальное оценивание, сочетающий в себе проекционный подход с интервальным анализом погрешностей. Показано, что этим методом можно обрабатывать большие массивы мультиколлинеарных данных физических экспериментов, причем результаты прогноза представляются в интервальной форме.

2) Разработаны теоретические основы метода ПИО. Исследованы его основные свойства. Разработан новый подход к оцениванию индивидуальной неопределенности прогноза для каждого объекта. Показано, что этот метод можно считать свободным от вида распределения погрешности.

3) Разработан новый подход к классификации объектов, позволяющий естественным объектом очертить рамки, в которых может использоваться построенная модель. Это достигается с помощью определения статуса объекта различающего: надежные «внутренние объекты», существенные «граничные объекты», подозрительные «внешние объекты», выпадающие «абсолютно-внешние объекты» и разрушительные «выбросы».

Научная и практическая значимость работы состоит в том, что с помощью разработанного метода и с применением компьютерной программы SIC были решены несколько важных теоретических и практических задач интерпретации данных различных многоканальных экспериментов. В том числе:

- на примере обработки сигналов дифференциальной сканирующей калориметрии проведено сопоставление результатов, полученных методом ПИО, с традиционными регрессионными доверительными оценками;
- на примере анализа ИК-спектров в ближней области представлен новый метод выбора представительного поднабора исследуемых объектов;
- в области многомерного контроля процессов, представлен новый метод многомерной статистической оптимизации процессов, проиллюстрированный практическим примером;
- на примере распознавания фальсифицированных лекарств с помощью ИК-спектроскопии в ближней области построен новый подхода к решению задачи дискриминации – интервальный вариант метода ПЛС дискриминации;
- на примере определения следовых концентраций нефти в воде с помощью акустических измерений, проведено сопоставление предложенной в работе классификации объектов с известным методом замкнутых оболочек.

Достоверность результатов обеспечена высокой точностью используемых теоретических и численных методов, согласованностью аналитических и численных методов решения задач, сопоставлением теоретических и экспериментальных результатов, а также сравнением с результатами, полученными другими авторами.

Автором выносится на защиту:

интервальный подход, предназначенный для анализа больших массивов данных физических экспериментов, решения линейных задач калибровки и прогнозирования. Основу подхода составляет метод простого интервального оценивания

(ПИО), который позволяет:

- вычислять оценки неизвестных параметров модели в виде области в пространстве параметров;
- вычислять результаты прогноза отклика в интервальном виде;
- создать метод классификации статуса объектов и интерпретации прогнозных интервалов;
- построить процедуру распознавания исследуемых объектов;
- разработать метод формирования представительной выборки.

Компьютерная программа SIC для решения задач линейной калибровки.

Методология применения интервального подхода для решения задач интерпретации физико-химических данных, раскрытая на следующих примерах:

- анализ кинетических данных на примере оценки активности антиоксидантов;
- построение калибровки и классификации объектов в задаче определения следовых концентраций нефти в воде с помощью акустических измерений;
- распознавание фальсифицированных лекарств с помощью инфракрасной спектроскопии в ближней области;
- построения моделей пассивной и активной оптимизации при многомерном контроле процессов;
- формирование представительной выборки на примере определение влажности зерна с помощью инфракрасной спектроскопии в ближней области.

Апробация работы. Основные результаты диссертации докладывались и обсуждались на следующих конференциях и симпозиумах: Всероссийской конференции «Математические методы в химии» (Санкт-Петербург 2003), Gordon Research Conference (Williamston 2001), XVI Менделеевском съезде (Ст.-Петербург 1998), Scandinavian Symposiums on Chemometrics (Lahti 1997, Porsgrunn 1999, Copenhagen 2001, Lappeenranta 2007), II международной конференции «Экспериментальные методы в физике» (Барнаул 2001), CONFERENTIA CHEMOMETRICA (Budapest 1997; Tata 2002), Международной школе-конференции «Современные методы анализа многомерных данных» (Кострома 2002, Барнаул 2003, Пушкинские Горы 2004, Черноголовка 2005, Самара 2006, Казань 2008), Chemometrics in Analytical Chemistry (Лиссабон 2004, Монпелье 2008), Symposium on Computer Applications and Chemometrics in Analytical Chemistry (Балатон 2004), международной конференции "Идентификация систем и задачи управления" (Москва 2005), всероссийском совещании по интервальному анализу (Петергоф 2006), международном конгрессе по аналитическим наукам (Москва 2006).

Публикации. Основные результаты работы опубликованы в 33 статьях (рецензируемых журналах, книгах и сборниках) и 40 тезисах докладов на международных и всероссийских конференциях (Общее число публикации автора 73).

Структура и объем работы. Диссертационная работа изложена на 272 страницах, содержит 85 рисунков и 22 таблицы. Диссертация состоит из введения, двенадцати глав с описание теоретических и прикладных исследований, выводов,

приложения с описанием основных алгоритмов и списка литературы из 297 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Введение

Обосновывается актуальность работы, и формулируются основные цели исследования. Учитывая, что проекционные методы, называемые хемометрическими, малоизвестны в России, в первой части представлены основные принципы и методы хемометрики.

Описание основных понятий многомерной калибровки, изложение метода простого интервального оценивания и его применение к задачам анализа результатов различных физических экспериментов рассматриваются в трех частях диссертационной работы.

ЧАСТЬ I. МНОГОМЕРНЫЕ ДАННЫЕ И ФОРМАЛЬНЫЕ МОДЕЛИ

Главы 1-3

В этой части вводятся основные понятия и объекты, с которыми работает исследователь при математической обработке результатов физического эксперимента. Приводится краткое описание базовых методов, которые используются в работе совместно с новым интервальным подходом.

Данные. Результаты физических экспериментов, т.е. наборы данных – это основной предмет, рассматриваемый в работе. Простейший случай – это одномерные данные, например, значение оптической плотности на одной длине волны. Наиболее часто встречающиеся, это двухмодальные данные. Типичный пример – набор спектров, снятых для I объектов на J длинах волн, представленных матрицей $\mathbf{X}$ размерности ($I \times J$). Строка представляет объект (образец или наблюдение), а столбец – переменную (длину волны). В последнее время большое внимание уделяется и более сложным, т.н. многомодальным (n-way) наборам данных. Пример таких данных рассматривается в главе 9.

Результаты физических экспериментов могут объединяться в блоки. Простейший случай – это один блок $\mathbf{X}$. В регрессионном анализе используются данные, состоящие из двух и более блоков. Блок независимых переменных представляется матрицей $\mathbf{X}$ размерности ($I \times J$) (например, матрица спектров). Блок откликов представляется матрицей $\mathbf{Y}$ размерности ($I \times L$) (например, матрица концентраций).

Методы качественного анализа. В задачах качественного анализа участвует один блок данных. При анализе многоканальных данных используются методы их сжатия. 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 должно быть существенно меньше числа исходных переменных, и, во-вторых, потери от такого сжатия должны быть сопоставимы с шумом в данных. Эти методы можно интерпретировать как проекцию исходных данных на пространство меньшей размерности, образованное скрытыми (латентными) переменными.

Метод главных компонент (МГК, K. Pearson, 1901) состоит в декомпозиции исходной матрицы $\mathbf{X}(I\times J)$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E} = \sum_{k=1}^{K} \mathbf{t}_{\mathrm{k}}\mathbf{p}_{\mathrm{k}}^{\mathrm{t}} + \mathbf{E}\,. \qquad (3.1)$$

$\mathbf{T}$ $(I\times K)$ называется матрицей *счетов,* столбцы $\mathbf{t}_{\mathrm{k}}$ ортогональны, т.е. $\mathbf{T}^{\mathrm{t}}\mathbf{T}=\mathrm{diag}(\lambda_{\mathrm{k}})$ – диагональная матрица, причем λ_{k} являются собственными значениями матрицы $\mathbf{X}^{\mathrm{t}}\mathbf{X}$. $\mathbf{P}(J\times K)$ называется матрицей *нагрузок*, столбцы которой ортонормированны, т.е. $\mathbf{P}^{\mathrm{t}}\mathbf{P}=\mathbf{I}$. $\mathbf{E}(J\times K)$ – это матрица остатков. Величина K называется *числом главных компонент* (ГК). Выбор K, проводится с использованием критериев, показывающих точность достигнутой декомпозиции. Величины

$$\mu_{\mathrm{k}} = 100\sum_{i=1}^{I}\mathbf{t}_{\mathrm{ik}}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{\mathrm{ij}}^2\,, \qquad E_{\mathrm{k}} = 100\left(1 - \sum_{i=1}^{I}\sum_{j=1}^{J} e_{\mathrm{ij}}^2 \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} x_{\mathrm{ij}}^2\right), \qquad k = 1,\ldots,K \qquad (3.2)$$

называются нормированным *собственным значением* и *объясненной вариацией*.

Важнейшим преимуществом проекционных методов является возможность представление сложных данных физического эксперимента в более простом виде, допускающем простую графическую интерпретацию.

Классификация и дискриминация. Это широкий класс задач качественного анализа, в которых требуется установить принадлежность объекта к некоторому классу. Эффективным подходом является метод *формального независимого моделирования аналогий классов* –SIMCA, (S. Wold, 1976). В этом методе каждый класс из обучающего набора независимо моделируется с помощью МГК с разным числом главных компонент K. После этого вычисляются расстояния между классами, а также расстояния от каждого класса до нового объекта. Используются расстояние d от объекта до класса, которое сравнивается с величиной d_0,

$$d = \sqrt{\frac{1}{J-K}\sum e_{\mathrm{j}}^2}\,,\quad d_0 = \sqrt{\frac{1}{(I-K-1)(J-K)}\sum_{ij} e_{\mathrm{ij}}^2}\,, \qquad (3.3)$$

а также расстояние от объекта до центра класса, называемое *размахом*

$$h = \sum_{k=1}^{K} \frac{\tau_{\mathrm{k}}^2}{\mathbf{t}_{\mathrm{k}}^{\mathrm{t}}\mathbf{t}_{\mathrm{k}}} \qquad (3.4)$$

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

Методы количественного анализа: калибровка. В задачах количественного анализа участвуют два блока данных: $\mathbf{X}$ – это матрица предикторов, а $\mathbf{Y}$ – это матрица откликов. Задача многомерной калибровки (ММК) состоит в построении математической модели, связывающей блоки $\mathbf{X}$ и $\mathbf{Y}$, с помощью которой можно предсказывать значения показателей y по новой строке значений $\mathbf{x}$.

По виду математических моделей, а, следовательно, и по методам отыскания неизвестных параметров, различают линейную и нелинейную калибровку. В третьей главе подробно рассмотрены линейные методы, кратко представлена нелинейная калибровка, рассмотрены методы многомодальной калибровки.

Линейная калибровка. Для решения задачи многомерной калибровки

$$\mathbf{Y}=\mathbf{XA}+\mathbf{E} \qquad (3.5)$$

используются проекционные методы, которые решают проблему вырожденности матрицы $\mathbf{X}^t\mathbf{X}$ и дают устойчивые оценки для неизвестных коэффициентов **A**. При использовании *регрессии на главные компоненты* – РГК, калибровка осуществляется в два этапа: на первом этапе к матрице **X** применяется МГК (3.1); на втором этапе к матрице счетов **T** применяется множественная регрессия:

$$\hat{\mathbf{a}} = \mathbf{P}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{P}^t\mathbf{X}^t\mathbf{y}\,. \qquad (3.6)$$

В настоящее время наиболее популярен метод *проекций на латентные структуры* (ПЛС, H. Wold, 1973). В ПЛС методе, декомпозиция **X** и **Y** производится одновременно

$$\mathbf{X}=\mathbf{TP}^t+\mathbf{E} \text{ и } \mathbf{X}=\mathbf{TW}^t+\mathbf{E},\ \mathbf{Y}=\mathbf{UQ}^t+\mathbf{F}\,, \qquad (3.7)$$

Столбцы матрицы **W** образуют ортонормированную систему. Оценки регрессионных коэффициентов имеют вид

$$\hat{\mathbf{A}} = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}\mathbf{Q}^t\,. \qquad (3.8)$$

Точность калибровки и предсказания принято характеризовать величинами

$$RMSEC = \sqrt{\sum_{i=1}^{I}(y_i^c - \hat{y}_i^c)^2 \Big/ F} \quad , RMSEP = \sqrt{\sum_{i=1}^{M}(y_i^t - \hat{y}_i^t)^2 \Big/ M} \qquad (3.9)$$

где y_i и $\hat{y}_i$ соответственно, измеренные и предсказанные значения физического показателя (индекс *c* определяет калибровочные объекты, *t* – объекты из проверочного набора). *I* – это число объектов в калибровочном наборе, *M* – в проверочном, а *F* – это число степеней свободы.

Существенным недостатком таких методов является то, что все они дают результат предсказания в виде точечной оценки, тогда как на практике часто нужна интервальная оценка, учитывающая неопределенность прогноза. Для преодоления этого недостатка предлагается использовать метод простого интервального оценивания.

ЧАСТЬ II. МЕТОД ПРОСТОГО ИНТЕРВАЛЬНОГО ОЦЕНИВАНИЯ

4. Объяснение ПИО метода

Интервальный подход основывается на следующей идее – заменить минимизацию суммы квадратов отклонений (метод наименьших квадратов) на систему неравенств, которая решается с помощью линейного программирования. Ранее был выполнен ряд важных прикладных работ, в частности получены интересные результаты по анализу информационной ценности кинетических измерений (С. Спивак, 1984). Кроме того, проводились исследования, направлены на построение интервальной оценки параметров моделей (метод центра неопределенностей), что оказалось малоплодотворным.

Использование такого подхода может дать интересные результаты, если рассматривать многомерную калибровку (ММК) как задачу построение интервального прогноза отклика. В этом случае результат прогноза сразу имеет вид интервала, поэтому этот метод называется «простым интервальным оцениванием» (ПИО). ПИО метод значительно отличается от привычного регрессионного подхода. Поэтому, перед строгим изложением математических аспектов ПИО,

приведено его элементарное объяснение, основанное на простейших примерах.

4.1. Почему погрешности ограничены

Исходным предположением ПИО является ограниченность погрешностей измерений, что принципиально отличается от стандартного допущения о нормальности погрешностей. Характерно, что большинство исследователей не связывают с принципом нормальности факт неограниченности погрешностей. Практика показывает, что на стадии предварительной обработки исследователи удаляют величины, лежащие за порогом трех или четырех стандартных отклонений. В то же время, объем данных, с которым работают экспериментаторы, часто превышает 10^{+6}, так что в них уверенно можно было бы ожидать 20-30 «нормальных» значений, выходящих за 4σ.

Еще один довод в пользу ограниченности погрешностей появляется при применении проекционных подходов. Так как эти методы используют формальные линейные модели, которые приближают исследуемые зависимости лишь на ограниченном участке, то, при построении таких моделей, периферийные объекты, которые могут нарушить линейность, обычно удаляют.

4.2. Модельный пример

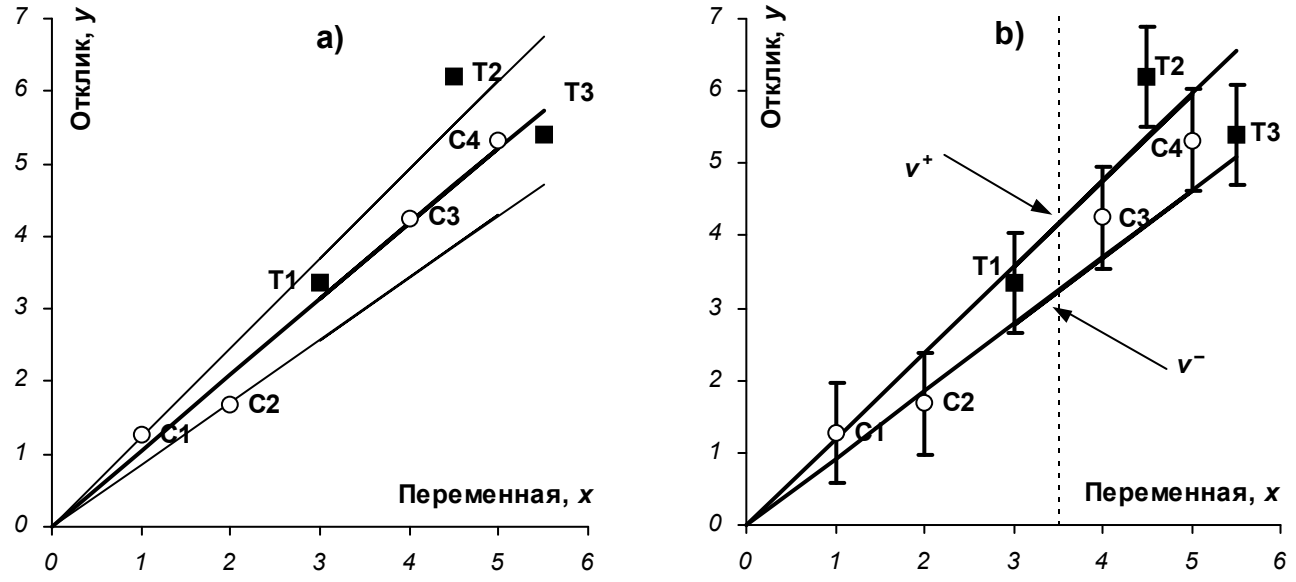
Рассматривается простейшая одномерная регрессия

$$y = xa + \varepsilon. \qquad (4.1)$$

Основным предположением метода ПИО является постулат об ограниченности погрешности измерения ε,. утверждающий, что никакая погрешность ε не может превосходить по абсолютной величине некоторую константу β,

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

Исследуем выводы, непосредственно вытекающие из этого постулата.



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

Метод ПИО: I – интервалы ошибок, ▬ границы интервалов предсказания

Рис. 4.1 Одномерный модельный пример: ○- калибровочные и ■- проверочные объекты

На Рис. 4.1 приведены модельные данные, построенные для регрессии (4.1)

при a=1. Погрешность измерения моделировалась с использованием равномерного распределения шириной 1.4, (β=0.7). Объекты C1-C4, используются как калибровочные, объекты T1-T3 – проверочные. Не смотря на примитивность примера, с его помощью можно объяснить все основные свойства метода ПИО.

МНК калибровка. Используя калибровочные данные (x_i,. y_i), i=1-4 и стандартную методику обработки, можно найти МНК оценку параметра a

$$\hat{a} = \frac{\bar{y}}{\bar{x}} = 1.004, \quad \text{где } \bar{x} = \frac{1}{4}\sum_1^4 x_i, \quad \bar{y} = \frac{1}{4}\sum_1^4 y_i, \tag{4.3}$$

и предсказать значения отклика y во всех точках x, как калибровочных, так и новых, оценить дисперсию погрешности ε и построить доверительный интервал:

$$\hat{y} = \hat{a}x; \quad s^2 = \frac{1}{3}\sum_1^4 \left(y_i - \hat{y}_i\right)^2 = 0.078 \; ; \; \hat{y}^{\pm} = \hat{y} \pm s\frac{x}{2\bar{x}}\mathrm{t}_3(P), \tag{4.4}$$

$\mathrm{t}_3(P)$ — это квантиль распределения Стьюдента с 3 степенями свободы для вероятности P. Границы доверительных интервалов приведены на Рис. 4.1a (P=0.95).

ПИО калибровка. Предположим, значение β известно. Из (4.1) и (4.2) следует, что для каждой пары (x_i,. y_i) из калибровочного набора выполняется условие

$$\left| y_i - ax_i \right| \le \beta, \quad \text{или в эквивалентной форме} \quad a_i^{\min} \le a \le a_i^{\max}, \tag{4.5}$$

где

$$a_i^{\min} = \frac{y_i - \beta}{x_i} \qquad a_i^{\max} = \frac{y_i + \beta}{x_i} \; .i=1,\ldots,4, \tag{4.6}$$

Неравенства (4.5) должны выполняться для всех калибровочных объектов. Так может быть только тогда, когда значений параметра a лежат в интервале

$$a^{\min} \le a \le a^{\max},$$
$$a^{\min} = \max_{1\le i\le 4} a_i^{\min}, \quad a^{\max} = \min_{1\le i\le 4} a_i^{\max}; \quad a^{\min} = 0.92, \; a^{\max} = 1.19 \tag{4.7}$$

Интервал (4.7) определяет *область допустимых значений* (ОДЗ) параметра a, т.е. такие значения, которые не противоречат экспериментальным данным. Когда параметр a меняется в интервале (4.7), то соответствующая величина отклика y=ax в произвольной точке x ограничена значениями:

$$v^- \le y \le v^+, \qquad v^- = a^{\min}x, \qquad v^+ = a^{\max}x \tag{4.8}$$

Таким образом построена интервальная оценка параметра a (4.7), которая является аналогом точечной МНК-оценки $\hat{a}$. Кроме того, найдены и прогнозные интервалы (4.8) для отклика y, справедливые, как для калибровочных, так и для любых других (новых) объектов (Рис. 4.1b).

Отметим очевидный факт, что построение калибровки методом ПИО в нашем примере «держится» только на двух объектах: C2 и C4. Они задают границы (4.7) возможных значений параметра a, поэтому мы вправе назвать эти объекты *граничными*. Прочие калибровочные объекты C1 и C3 несущественны; их можно удалить из калибровочного набора, и результат останется прежним. Это очень важное свойство метода ПИО, которое находит применение в задаче вы-

бора представительного набора объектов.

4.3. Сходимость интервальных оценок

На другом простом примере проведено сравнение интервальной ПИО-оценки с обычной оценкой метода максимума правдоподобия. Рассматривается выборка $\mathbf{x}=(x_1,..., x_n)$ из нормального распределения $N(\alpha,\sigma^2)$, усеченного на интервале $[\alpha-\beta, \alpha+\beta]$, $\beta=\kappa\sigma$. Требуется построить оценку среднего значения α при известных значениях β и κ, и исследовать ее сходимость, т.е. зависимость точности от объема выборки n.

Оценка α по методу максимума правдоподобия или моментов строится как среднее по выборке $a_{\mathrm{ML}}=\frac{1}{n}\sum_{i=1}^{n}x_i$, и ее точность можно охарактеризовать приближенным доверительным интервалом $\mathrm{Prob}\left(\left| a_{\mathrm{ML}}-\alpha \right| < \beta h_{\mathrm{ML}}\right)=P$, где

$$h_{\mathrm{ML}}(P)=\frac{x_{0.5(1-P)}}{\sqrt{n}}\varphi(\kappa) \qquad (4.9)$$

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

С другой стороны, интервальная оценка имеет вид $a_{\mathrm{SIC}}=\left[\min(x_i+\beta), \max(x_i-\beta)\right]$. Нормированную полуширину (ПИО размах) этого интервала, соответствующую доверительной вероятности P, можно записать в виде:

$$h_{\mathrm{SIC}}(P)=-\frac{\ln(1-P)}{n}2\psi(\kappa). \qquad (4.10)$$

Функции $\varphi(\kappa)$ и $\psi(\kappa)$ зависят от параметра κ, который определяет, как проводится отсечение (κ=0.2, 1, 2, 3, 4). При κ=0.2 распределение близко к равномерному, а при κ=4 практически неотличимо от не усеченного нормального распределения.

Показано, что, в рассматриваемой задаче, ПИО-оценка эффективней оценки ММП, начиная с некоторого объема выборки n_0, которая зависит от параметра κ. Чем ближе усеченный закон распределения к нормальному (большие значения κ), тем больше должен быть объем выборки.

Результат главы 4. Показано, что 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 распределения погрешности, кроме ее ограниченности. Тем самым его можно считать методом, свободным от вида распределения.

5. Описание метода ПИО

Эта глава представляет систематическое описание метода ПИО, вводятся основные определения, приводятся доказательства в общем виде.

5.1. Область допустимых значений

Рассмотрим модель линейной многомерной калибровки

$$\mathbf{y} = \mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon}\,, \tag{5.1}$$

где $\mathbf{y}$ ($I\times1$)– это вектор откликов; $\mathbf{a}$ ($J\times1$)– это вектор параметров; $\mathbf{X}$ ($I\times J$)– это матрица предикторов, $\boldsymbol{\varepsilon}$ ($I\times1$) – это вектор погрешностей. Ограниченность погрешности $\boldsymbol{\varepsilon}$ означает, что существует такая величина $\beta>0$, что

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

Для любого объекта ($\mathbf{x}_\mathrm{i}$, y_i) из калибровочного набора, i=1, …, I можно записать

$$y_\mathrm{i}^{-} \le \mathbf{x}_\mathrm{i}^{\mathrm{t}}\mathbf{a} \le y_\mathrm{i}^{+}, \qquad y_\mathrm{i}^{-} = y_\mathrm{i} - \beta, \qquad y_\mathrm{i}^{+} = y_\mathrm{i} + \beta\,. \tag{5.3}$$

Значения $\mathbf{a}$, удовлетворяющие (5.3) для данного объекта i, образуют полосу $S(\mathbf{x}_\mathrm{i}, y_\mathrm{i})$ в пространстве параметров R^J. 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 $\mathbf{a}$ системы (5.1) – это множество в пространстве параметров:

$$A=\bigcap_{i=1}^{I} S(\mathbf{x}_\mathrm{i}, y_\mathrm{i}) \quad \text{или} \quad A=\left\{\mathbf{a}\in \mathrm{R}^J : \mathbf{y}^{-} < \mathbf{X}\mathbf{a} < \mathbf{y}^{+}\right\} \tag{5.4}$$

A **-** это замкнутый выпуклый многогранник. При этом A является случайным множеством, поскольку оно построено с использованием случайных величин y.

5.2. Свойства ОДЗ

Показано, что для любой модели, заданной уравнением (5.1), ОДЗ A обладает следующими свойствами.

Свойство 1. Область A является *несмещенной* оценкой параметра $\boldsymbol{\alpha}$.

Непосредственно из определения ОДЗ следует, что истинное значение $\boldsymbol{\alpha}$ всегда принадлежит A: $\mathrm{Prob}\{\boldsymbol{\alpha}\in A\}=1$.

Свойство 2. Область A ограничена тогда и только тогда, когда матрица $\mathbf{X}$ имеет полный ранг, т.е. rank $\mathbf{X}$= J .

Это означает, что если система (5.1) мультиколлинеарна, т.е. rank $\mathbf{X}$ < J, то до использования ПИО метода, необходимо применить какую-либо процедуру регуляризации. Например, спроецировать исходные данные (5.1) на подпространство меньшей размерности

$$\mathbf{y} = \mathbf{T}\mathbf{P}^{\mathrm{t}}\mathbf{a} + \mathbf{f} = \mathbf{T}\mathbf{q} + \mathbf{f}, \tag{5.5}$$

где матрица $\mathbf{T}$ имеет полный ранг $K<J$, а затем применить метод ПИО к (5.5).

Свойство 3. Область A является состоятельной оценкой параметра $\boldsymbol{\alpha}$,

$$\mathrm{Prob}\{A\cap\boldsymbol{\alpha}\}=1 \quad \text{при} \quad I\to\infty \tag{5.6}$$

при тех же «слабых» условиях, что и в МНК, т.е. $\lambda_j\to\infty$ при $I\to\infty$.

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

Свойство 4 Область A образована не всеми объектами из калибровочного набо-

ра, а только некоторыми, называемыми граничными.

Это означает, что из калибровочного набора можно исключить все объекты, кроме граничных, и ОДЗ при этом не изменится.

5.3. Предсказание отклика

Используя ОДЗ A, построенную для модели (5.1) или (5.5), можно предсказать значение отклика y для любого вектора $\mathbf{x}$. Если параметр $\mathbf{a}$ меняется внутри ОДЗ A, то значение $y=\mathbf{x}^t\mathbf{a}$ принадлежит интервалу

$$V=[v^-, v^+] \quad \text{где} \quad v^- = \min_{\mathbf{a}\in A}(\boldsymbol{x}^t\mathbf{a}), \qquad v^+ = \max_{\mathbf{a}\in A}(\boldsymbol{x}^t\mathbf{a}) \tag{5.7}$$

Интервал V является результатом прогноза методом ПИО. Для его вычисления не требуется строить область A в явном виде, т.к. значения v^- и v^+ могут быть найдены с помощью стандартных методов линейного программирования.

Кроме того, имеется еще интервал калибровки U, который характеризует меру неопределенности в модели

$$U=[\,y-\beta\,,\,y+\beta\,]\,. \tag{5.8}$$

Величина прогнозного интервал V индивидуальна для каждого объекта, а величина интервала калибровки U – общая для всех объектов. Взаимное расположение этих интервалов (Рис. 6.2а) характеризует "качество" прогноза.

5.4. Оценка β

Как правило, величина β неизвестна и, вместо нее, используется некоторая оценка b. Согласно определению (5.4), ОДЗ A зависит от b, и $A(b)$ монотонно расширяется с увеличением b –

$$b_1 > b_2 \quad \Rightarrow \quad A(b_1) \supset A(b_2), \quad A(0)=\varnothing, \quad A(\infty)\neq\varnothing \tag{5.9}$$

Из (5.9) следует, что существует минимальное значение b, при котором $A(b)\neq\varnothing$. Это значение может быть принято в качестве оценки величины β

$$b_{\min} = \min\{b, \quad A(b)\neq\varnothing\}\,. \tag{5.10}$$

Предложенная оценка (5.10) является состоятельной, но смещенной, т.к. $b_{\min}\leq\beta$. Она задает нижний предел всех возможных значений β. Поэтому необходимо оценить и верхнюю границу максимальной погрешности.

Очевидно, что любая разумная оценка b должна зависеть от двух показателей: (1) числа объектов в калибровочном наборе; чем больше объектов, тем ближе величина b к β; (2) тяжести крыльев функции распределения погрешностей; чем крылья легче, тем хуже эта оценка. Применяя традиционный статистический подход к регрессионным остаткам $\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}$, можно построить такую оценку b, что $\mathrm{Prob}\{b>\beta\}>P$ и, при этом, оценка b максимально близка к β. Имитационное моделирование, проведенное для различного числа объектов с использованием различных ограниченных распределений ошибки, показывает, что оценка

$$b_{\mathrm{SIC}}=b_{\mathrm{reg}}\,C(I,\,s^2,\,P),\ \text{где}\ b_{reg}=\max\left(\left|\frac{e_1}{\sqrt{1-h_1}}\right|,\ldots,\left|\frac{e_I}{\sqrt{1-h_I}}\right|\right),\ h_i=\mathbf{x}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_i \tag{5.11}$$

может быть принята за искомый верхний предел β с вероятностью P. Здесь

$C(I, s^2, P)$ – эмпирическая функция, которая зависит от числа объектов в калибровочном наборе I и от оценки дисперсии остатков s^2.

Значение b_{SIC} в качестве оценки β в дальнейшем используется в ПИО методе для определения прогнозных интервалов и для классификации объектов.

Точность ПИО моделирования. Значения b_{min} и b_{SIC}

$$b_{min} \le \beta \le b_{SIC} \qquad (5.12)$$

полностью характеризуют точность ПИО моделирования, т.е.

- Любое априорное значение β допустимо только в том случае, если оно больше или равно b_{min}.
- Моделирование с помощью ПИО методов с параметром b_{SIC} гарантирует, что для объектов из калибровочного набора, ‘истинное’ значение отклика расположено внутри соответствующего прогнозного интервала.
- Даже в ‘наихудшем’ случае, полуширина прогнозного интервала для объектов из калибровочного набора меньше или равна b_{SIC}.
- Обе оценки β: b_{min} (5.10) и b_{SIC} (5.11) – являются состоятельными. Это означает, что для любого значения β из интервала (5.12) выполняются свойства 2 и 4 из раздела 5.2, а свойства 1 и 3 выполняется асимптотически.

Результат главы 5. В этой главе приведены основные понятия и доказаны основные свойства ПИО метода.

1. Дано определение и доказаны основные свойства области допустимых значений параметров A. Показано, то 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 ПИО-интервалы, устанавливают индивидуальную неопределенность прогноза отклика для каждого нового объекта.

6. Классификация статуса объектов

6.1. Характеристики статуса объектов

Для характеристики качества прогноза и формализации понятий «похожих» и «непохожих» объектов в рамках метода ПИО вводятся следующие определения.

Пусть имеется ПИО модель, построенная с помощью набора калибровочных объектов $(\mathbf{x}_i, y_i)$, $i=1, \ldots, I$, которая характеризуется своей ОДЗ A, (5.4). Рассмотрим новый объект, т.е. пару $(\mathbf{x}, y)$, с которым связана своя полоса $S(\mathbf{x}, y)$, определенная неравенствами $y - \beta \le \mathbf{x}^t\mathbf{a} \le y + \beta$. Тогда взаимное положение полосы $S(\mathbf{x}, y)$ и области A характеризует статус объекта (см. Рис. 6.1).

Определение 6.1 Объект $(\mathbf{x}, y)$ называется *внутренним*, если он не изменяет ОДЗ, т.е. $A \cap S(\mathbf{x}, y) = A$, иначе, $\left|\mathbf{x}^t\mathbf{a} - y\right| \le \beta$ для $\forall\ a \in A$.

Любой объект из калибровочного набора, по построению, является внутренним (Рис. 6.1 a,b).

Определение 6.2 Объект ($\mathbf{x}_i$, y_i) из калибровочного набора называется *граничным*, если существует такой параметр $\mathbf{a}\in A$, что $\left|\mathbf{x}_i^t\mathbf{a}-y_i\right|=\beta$.

Граничные объекты формируют ОДЗ, и, поэтому, являются наиболее важными среди объектов калибровочного набора (Рис. 6.1 а).

Определение 6.3 Объект ($\mathbf{x}$, y) называется *внешним*, если он уменьшает ОДЗ, т.е. $A \cap S(\mathbf{x}, y) \neq A$, иначе, $\exists\ \mathbf{a}\in A$ что $\left|\mathbf{x}^t\mathbf{a}-y\right|>\beta$.

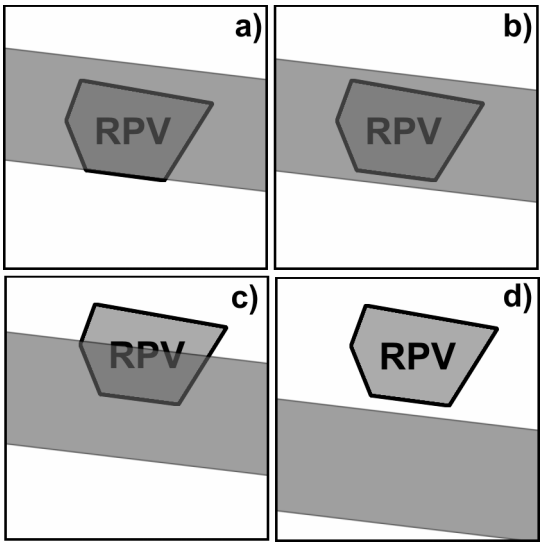


Рис. 6.1 Возможные положения полосы нового объекта по отношению к данной ОДЗ в пространстве параметров

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 ($\mathbf{x}$, y) называется – *выбросом*, если он уничтожает ОДЗ, т.е. $A \cap S(\mathbf{x}, y) =\varnothing$, иначе, $\left|\mathbf{x}^t\mathbf{a}-y\right|>\beta$ для $\forall\ \mathbf{a}\in A$. (Рис. 6.1d)

Определение 6.5 Объект ($\mathbf{x}$, y) называется *абсолютно внешним*, если для любого значения y $A \cap S(\mathbf{x}, y) \neq A$.

В работе показано, что при добавлении в калибровочный набор дополнительного I+1 –го объекта, в зависимости от его статуса, происходят следующие изменения с ОДЗ A. Если объект является внутренним, то 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. $A_{I+1}=\varnothing$. Классификация объектов проявляется не только во взаимном расположении полос и ОДЗ в пространстве параметров, но и во взаимном положении калибровочного, U (5.8) и прогнозного, V (5.7) интервалов. Это подтверждается следующими утверждениями, доказанными в работе.

Утверждение 6.1 Для всех *калибровочных* объектов выполняется условие

$$V_i \cap U_i = V_i, \quad i=1,...,I.$$

Утверждение 6.2 Объект является *внутренним* тогда и только тогда, когда

$$V_i \cap U_i = V_i .$$

Утверждение 6.3 Калибровочный объект ($V_i \subseteq U_i$) является *граничным* тогда и только тогда, когда

$$\max(V_i)=\max(U_i) \text{ либо } \min(V_i)=\min(U_i).$$

Утверждение 6.4 Объект является *выбросом* тогда и только тогда, когда

$$V \cap U =\varnothing.$$

Утверждение 6.5 Объект является *абсолютно-внешним* тогда и только тогда, когда для любого значения y

$$V \cap U \neq V.$$

6.2. Диаграмма статуса объектов (ДСО)

Для того чтобы процедуру классификации объектов сделать максимально простой и наглядной, в работе введены следующие величины.

Определение 6.6. *ПИО-остатком* называется величина –

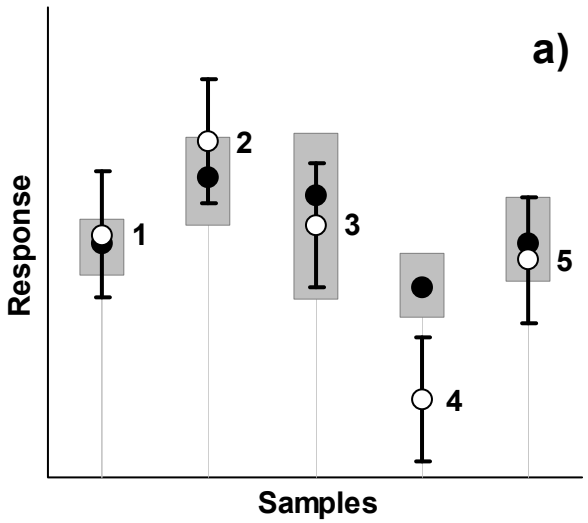
$$r(\mathbf{x}, y) = \frac{1}{\beta}\left(y - \frac{v^{+}(\mathbf{x}) + v^{-}(\mathbf{x})}{2} \right), \tag{6.1}$$

которая характеризует β-нормализованное *смещение*.

Определение 6.7. *ПИО-размахом* называется величина –

$$h(\mathbf{x}) = \frac{1}{\beta}\left(\frac{v^{+}(\mathbf{x}) - v^{-}(\mathbf{x})}{2} \right), \tag{6.2}$$

которая характеризует β-нормализованную *воспроизводимость*. Целесообразность этих определений раскрывается следующими утверждениями.



Интервал калибровки (черный), интервал предсказания (серый), (○) – опорное значение, (●) – предсказанное значение

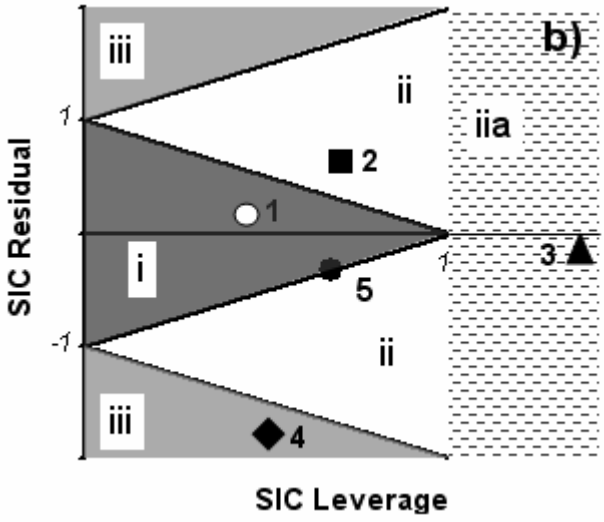


Диаграмма статуса объектов. i – внутренние, ii-внешние, iia- абсолютно внешние, iii- выбросы

Рис. 6.2. Результаты ПИО прогноза

Утверждение 6.6 Все калибровочные объекты удовлетворяют неравенству $\left| r(\mathbf{x}, y) \right| \le 1 - h(\mathbf{x})$.

Утверждение 6.7 Объект ($\mathbf{x}$, y) является *внутренним* тогда и только тогда, когда $\left| r(\mathbf{x}, y) \right| \le 1 - h(\mathbf{x})$.

Утверждение 6.8 Калибровочный объект ($\mathbf{x}_i$, y_i) является *граничным*, тогда и только тогда, когда $|r(\mathbf{x}_i, y_i)| = 1 - h(\mathbf{x}_i)$.

Утверждение 6.9 Объект ($\mathbf{x}$, y) является *выбросом* тогда и только тогда, когда $\left| r(\mathbf{x}, y) \right| > 1 + h(\mathbf{x})$

Утверждение 6.10 Объект ($\mathbf{x}$, y) является *абсолютно-внешним* тогда и только тогда, когда $h(\mathbf{x}) > 1$.

Используя Определения 6.6-6.7 и Утверждения 6.6-6.10, можно построить диаграмму статуса объектов (ДСО), прототип которой показан на Рис. 6.2b.

6.3. Классификация новых объектов

Когда модель ММК применяется к новым объектам, соответствующие значения *y* неизвестны. Поэтому нельзя вычислить ПИО-остаток, *r* (6.1), но всегда можно определить величину ПИО-размаха, *h* (6.2). Если для нового объекта $h>1$ (область *iia* на Рис. 6.2b), то этот объект является абсолютно-внешним. Для любого калибровочного набора можно сконструировать область в пространстве предикторов (счетов), за пределами которой располагаются абсолютно внешние объекты. Следующее утверждение определяет эту область.

Утверждение 6.11. Пусть D – это область в пространстве предикторов, образованная всеми возможными линейными комбинациями взвешенных векторов предикторов (или счетов) $\mathbf{x}_i$ из калибровочного набора, такими что

$$\mathbf{x} = \sum_{i=1}^{I} \frac{\lambda_i}{h(\mathbf{x}_i)} \mathbf{x}_i, \quad \sum_{i=1}^{I} |\lambda_i| \le 1 \ . \tag{6.3}$$

Тогда все абсолютно внешние объекты будут расположены вне этой области.

Результаты главы 6. Показано, что для решения задач многомерной калибровки, ПИО подход позволяет ввести новый метод классификации объектов. Он базируется на определениях 6.1-6.5 и утверждениях 6.1-6.5. При этом нет необходимости в явном виде строить ОДЗ в пространстве параметров.

Для ПИО классификации достаточно построить диаграмму статуса объектов, которая, независимо от исходной размерности задачи, представляется в двумерном пространстве. Позиция каждого объекта на ДСО определяет, подобен ли изучаемый объект объектам из калибровочного набора, и тем самым, задает разумные границы применимости построенной калибровки.

7. Программная реализация ПИО метода

Разработанные аспекты ПИО метода были реализованы в программе SIC (Simple Interval Calculations), которая работает под управлением системы Excel, входящей в стандартный пакет Microsoft Office. Приведено описание структуры программы, которая состоит из целого набора процедур: (1) предварительной подготовки данных; (2) проекционных регрессионных методов (МГК, РГК, ПЛС 1, ПЛС 2); (3) процедуры приведения исходной задачи к стандартной форме линейной оптимизационной модели; (4) стандартной процедуры Симплекс-метода для решения линейной оптимизационной задачи; (5) вычисления результатов, построения ДСО.

Вся входная информации представляется в виде таблиц рабочих листов Excel. С помощью программы SIC можно получить следующую информацию:

- результаты интервального прогноза отклика $[v^-, v^+]$;
- точечную регрессионную оценку откликов (РГК, ПЛС);
- оценки параметра β: $b_{\min}$ и b_{SIC};
- ПИО-остаток и ПИО-размах;
- диаграмму статуса объектов.

Информация выводится как в числовом, так и в графическом виде. Устройство программы SIC соответствует современным требованиям. Все действия выполняются либо с помощью диалогового окна, либо осуществляются с помо-

щью VBA процедур. Программа SIC – это инструмент, созданный для интервального и регрессионного анализа результатов сложных многофакторных физических экспериментов. Программа систематически используется в работе.

ЧАСТЬ III. ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ МЕТОДА ПРОСТОГО ИНТЕРВАЛЬНОГО ОЦЕНИВАНИЯ

Эта часть работы посвящена методология применения интервального подхода для интерпретации результатов различных многоканальных экспериментов.

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

В этой главе излагаются общие принципы, применяемые при совместном использовании проекционных регрессионных методов и метода ПИО. Проекционный подход базируется на концепции «скрытых (латентных) переменных», на которых строится проекционное подпространство. Возможность наглядного представления сложных многофакторных данных физического эксперимента в проекционном пространстве позволяет исследователю лучше понять и объяснить изучаемые явления. При этом возникает необходимость (1) охарактеризовать свойства каждого отдельного объекта относительно всей группы объектов и построенной модели; (2) очертить область действия модели, а, следовательно, и надежность прогноза. Метод ПИО представляет систему классификации объектов, а так же набор однозначных правил для определения статуса (роли) каждого объекта.

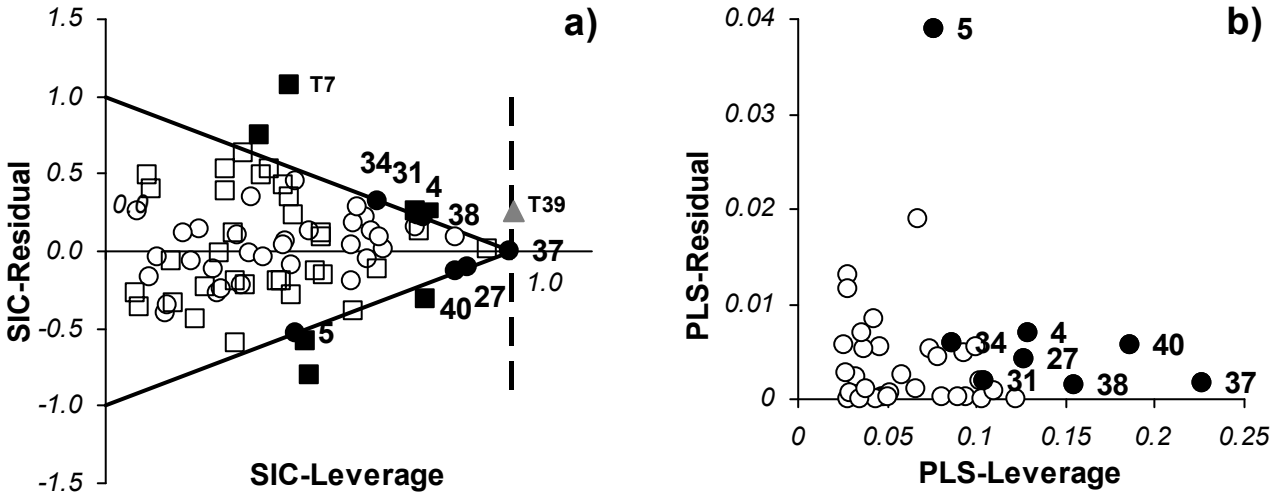
8.1. Эксперимент. Измерение следовых концентраций нефти в воде с помощью акустических измерений

Результаты ПИО классификации демонстрируются 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.

Матрица предикторов **X** состоит из акустических спектров (преобразованные с помощью быстрого преобразования Фурье) на 1024 частотах; вектор откликов **y** – это известные стандартные концентрации нефти (0, 2.5, 5, 10, 20, 50, 100, 300 ppm.). С помощью ПЛС метода построена модель, основанная на двух ГК, при этом *RMSEC*=0.12, и *RMSEP*=0.24.

8.2. Исследование калибровочного набора

Сравнение графиков на Рис. 8.1 a) и b) показывает, какую новую информацию предоставляет ПИО метод, по сравнению с обычным ПЛС методом. Сравнивая ДСО (Рис. 8.1 a) с графиком влиятельности Рис. 8.1 b видно, что все наиболее влиятельные объекты (NN 37, 38 и 40), а так же объект, имеющий максимальное значения остатка моделирования (N5), являются граничными по ПИО классификации.



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

Рис. 8.1 Определение следовых концентраций нефти в воде.
Калибровочный набор, объекты: ○- внутренние, ● - граничные.
Проверочный набор, объекты: □ - внутренние, ■ - внешние, ▲ - абсолютно внешние

ПИО классификация позволяет однозначно определять все наиболее влиятельные объекты среди калибровочного набора (Утв.6.8). Концепция граничных объектов имеет смысл не только внутри самого метода ПИО, она объективно характеризует изучаемую структуру данных физического эксперимента.

8.3. Исследование проверочного набора

Важным аспектом ПИО классификации является определение статуса объектов проверочного набора. В проверочном наборе обнаружено 32 внутренних и 8 внешних объектов (Рис. 8.1 а). Внешними объекты могут быть по двум причинам: (1) большая ошибка в измерениях откликов; (2) погрешность моделирования. Прогноз на такие объекты, например T7, является ненадежным. Объект T39 является абсолютно внешним. Такие объекты по структуре данных в предикторах отличается от калибровочных объектов. Величина прогнозных интервалов для них всегда больше, чем β.

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

Калибровочная модель используется для предсказания откликов новых объектов. Если новый объект плохо согласуется с моделью, результат предсказания будет плохим (большая неопределенность), или даже неверным (предсказанное значение и прогнозный интервал далеки от истинного значения). В работе предлагается новый метод определения выпадающих объектов, основанный на методе ПИО. Этот метод сравнивается с известным методом выпуклых оболочек (Fernandez, 2002).

Для обнаружения выпадающих объектов, предлагается построить в пространстве предикторов область, которая определяет абсолютно внешние объекты (Утв. 6.11). Для каждого объекта $\mathbf{x}_i$ из калибровочного набора вычисляются координаты точек $\mathbf{x}_i^b$, образующих границу области по формуле

$\mathbf{x}_i^b = \mathbf{x}_i 2\beta\left(v^+(\mathbf{x}_i) - v^-(\mathbf{x}_i)\right)$. Существенным отличием является то, что метод выпуклых оболочек учитывает только значения предикторов, в то время как метод ПИО принимает во внимание еще и результаты моделирования отклика.

Результаты главы 8. На примере анализа результатов многоканальных акустических измерений, показано, что при объединении метода ПИО с известными методами билинейного моделирования (РГК, ПЛС) появляется новый инструмент для анализа сложных многофакторных данных. Визуализация многоканальных экспериментальных данных помогает проследить имеющиеся физические зависимости, оценить качество проведенного эксперимента.

Основой для такой визуализации служит классификация статуса объектов, основанная на следующих правилах.

1. Калибровочные объекты делятся на два класса: граничные, наиболее важные объекты, и внутренние объекты, являющиеся избыточными. (Утв. 6.6-6.7).
2. Проверочные объекты можно разделить (Утв. 6.7-6.10) на два основных класса: внутренние (типичные) и внешние объекты. Среди внешних объектов дополнительно выделяются абсолютно внешние объекты и выбросы.
3. Для новых объектов, имеется правило (Утв. 6.10), выделяющее абсолютно внешние объекты. Это является существенным достижением ПИО метода, так как гарантирует, что, применяя модель для новых измерений, мы не выйдем за область действия модели.

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

Традиционно для задач анализа кинетических данных применяется содержательное физико-химическое моделирование, базирующееся на основных кинетических принципах. Оно позволяет получать оценки параметров с высокой точностью, но применимо только тогда, когда модель процесса известна априори. Альтернативой является формальный подход, в котором кинетическая модель явно не используется. При этом экспериментальные данные описываются линейной многофакторной моделью, справедливой в ограниченном диапазоне условий. Использование одного и того же набора данных, позволяет сравнить оба подхода и сделать выводы о том, в каком случае какой подход предпочтительнее.

9.1. Эксперимент. Оценка активности антиоксидантов

Антиоксиданты (АО) – это специальные добавки, которые замедляют термоокислительное старение полимеров. Основной характеристикой эффективности АО является *период индукции*, измеряемый в процессе длительного термостарения. Альтернативой является подход, использующий метод дифференциальной сканирующей калориметрии (ДСК), с последующей математической обработкой полученных данных. В эксперименте исследовались 25 образцов АО. Были изготовлены пленки *полипропилена* (ПП) с АО в концентрациях 0.05% , 0.07%, и 0.1%. 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). Были получены данные, в которых матрица **X** состоит из температур начала окисления (ТНО), определенных в ДСК эксперименте. Они образуют трех модальный блок. Данные **Y** – это значения периодов индукции (ИП), полученные с помощью длительного термического старения.

9.2. Формальное моделирование

Экспериментальные данные обрабатывались с помощью метода ПЛС – для калибровки, и метода ПИО – для построения прогнозных интервалов. Исходные **X** данные раскладывались в плоскую матрицу (25 × 15). Для каждой концентрации строилась отдельная модель. Ввиду гетероскедастичности, из значения **Y** извлекался квадратный корень. Результаты приведены в Таб. 9.1, а на Рис. 9.1

Таб. 9.1 Прогноз 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 $(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2)$	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 $(\mathbf{w}_i)$	1.038	1.151	1.397	0.934	1.204	1.476
6.	Корреляция $(\mathbf{w}_1, \mathbf{w}_2)$	0.202	0.007	0.028	0.202	0.007	0.028
7.	Корреляция $(\mathbf{y}, \mathbf{w}_i)$	0.815	0.846	0.836	–0.184	–0.161	–0.113

9.3. Содержательное моделирование

В этом подходе для каждого АО строится своя кинетическая модель - всего 25 нелинейных регрессионных (НЛР) моделей. Для каждой модели матрица **X** – это 5 скоростей нагрева × 3 концентрации АО, векторы **y** – это 3 значения ИП. Процедура калибровки состоит из двух шагов. На первом шаге строится модель, описывающая расход антиоксиданта в ходе ДСК – это калибровка **X** данных. На втором шаге, строится модель для описания расхода АО в ходе длительного термического старения – калибровка **Y**

$$y = \left[\frac{E_c}{RT_e} + \ln(A_0) - c \right] \exp\left(\frac{E_a}{RT_e} - a \right) . \qquad (9.1)$$

В этой функции участвуют те же кинетические параметры a, E_a, c и E_c, что и в первой модели. Их оценки находятся на первом шаге, а на втором применяется процедура переноса ошибок для того, чтобы оценить неопределенность в прогнозе. В качестве инструмента для вычислений использовалась программа Fitter. Сводные результаты НЛР прогноза приведены в Таб. 9.1 и на Рис. 9.1

9.4. Сравнение методов

Из Таб. 9.1 и Рис. 9.1, можно сделать следующие выводы. Оба метода имеют близкую точность (ряд 1 в Таб. 9.1) и смещение (ряд 2). Неопределенность прогноза становится больше, когда начальная концентрация АО увеличивается. В целом, ПЛС/ПИО метод дает лучшие результаты для малых начальных кон-

центраций АО, а НЛР лучше для больших концентраций. Однако точечные оценки ($\hat{\mathbf{y}}_m$, m=1 – это НЛР оценка, m=2 – это ПЛС/ПИО оценка) в среднем близки (см. ряд 3).

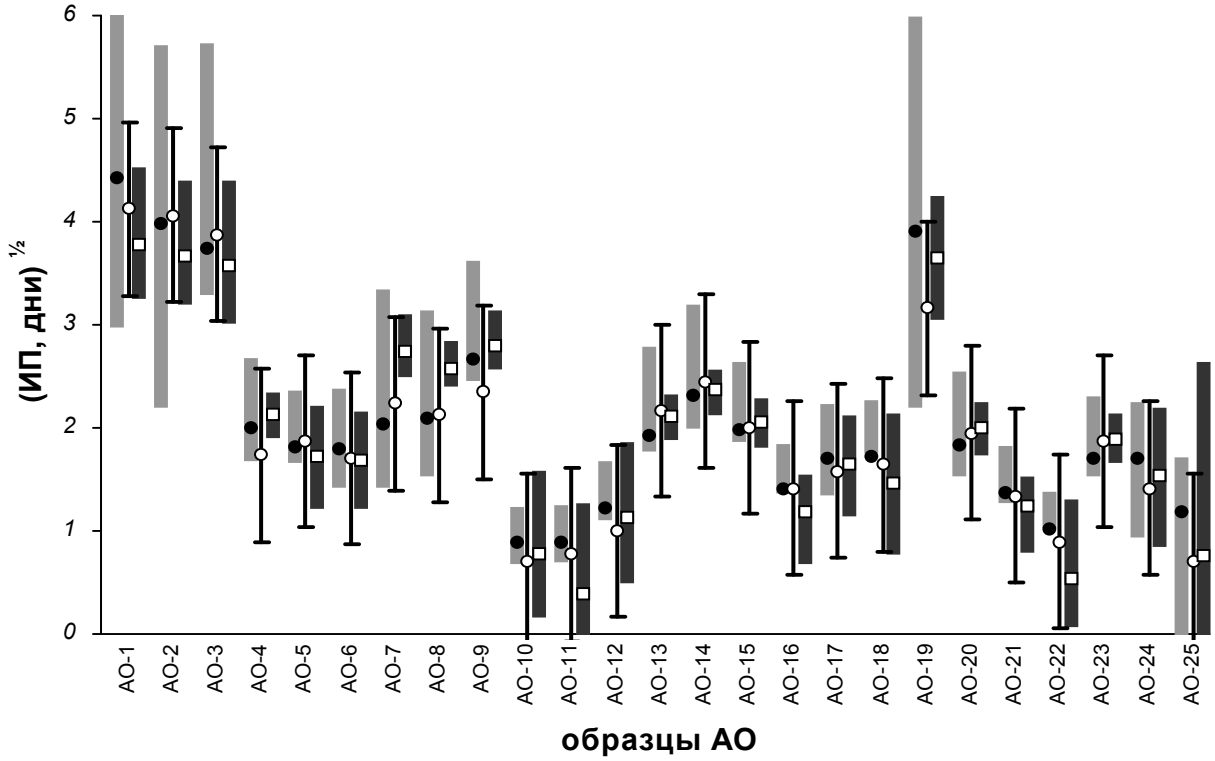


Рис. 9.1. Результаты прогноза ИП для различных образцов АО с начальной концентрацией 0.05. Точки (●) и серые прямоугольники представляют содержательное (НЛР) предсказание. Квадраты (□) и черные прямоугольники изображают формальное (ПЛС/ПИО) моделирование. Точки (○) соответствуют измеренным значениям с вертикальными отрезками, которые показывают погрешность измерения (калибровки) β. Из всех величин извлечен квадратный корень

Оба метода хорошо моделируют значения **X**, но содержательный метод (НЛР) делает это немного лучше (ряд 4). Видно (ряд 5), что ширина прогнозных интервалов растет с начальной концентрацией АО. Это следует из формулы для содержательной модели (9.1), представляющую зависимость ИП от начальной концентрации АО. В тоже время, в формальной ПЛС/ПИО 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. Интервальные оценки близки в среднем (ряд 5), хотя доверительные интервалы ($\mathbf{w}_1$) могут сильно отличаться от прогнозных интервалов ($\mathbf{w}_2$) для отдельных объектов (см. ряд 6 и Рис. 9.1). Последний ряд Таб. 9.1 показывает, что ширина $\mathbf{w}_1$ растет с увеличением значения периода индукции для всех начальных концентраций АО, тогда как ширина $\mathbf{w}_2$ не зависит от **y**. Это свидетельствует о том, что преобразование откликов, действительно дало ожидаемый эффект в ПЛС/ПИО моделировании, но не смогло исправить результаты НЛР моделирования.

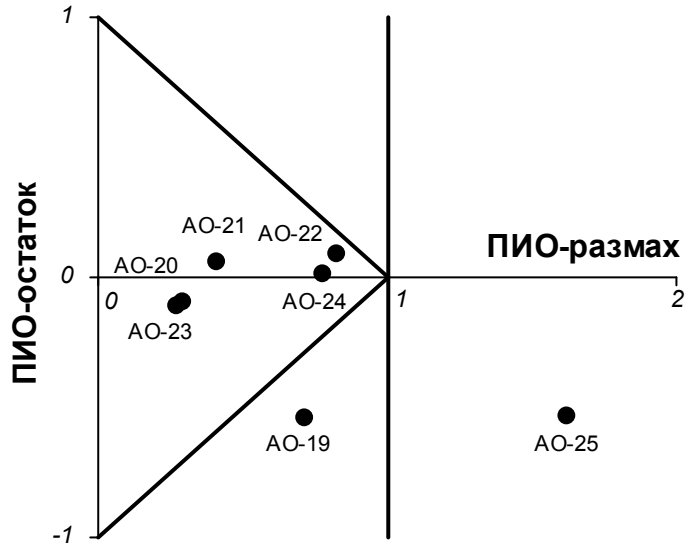


Рис. 9.2 ДСО объектов из проверочного набора с А0=0.07

Для любого метода важно ограничить область его применимости. Содержательное моделирование (НЛР) может использоваться для предсказания ИП для различных концентраций АО и при разных температурах экспозиции. Однако мы не можем указать границы допустимой экстраполяции. При формальном моделировании экстраполяция исключается, а область применимости модели дает ДСО (Рис. 9.2). Все объекты, расположенные внутри треугольника (АО-20, 21, ..., 24) – внутренние, прогноз на них надежен. Объекты АО-19 и АО-25 внешние. Они не противоречат модели, но прогноз на них менее надежен. Тому могут быть две причины: большой размах (АО-25) и смещение (АО-19).

Результаты главы 9. Продемонстрированы два подхода к решению одной и той же задачи – проверки активности АО. Предлагается заменить длительное термостарение на быстрые измерения с помощью ДСК. Отмечено принципиальное различие между областями применимости математических моделей и различная тактика планирования эксперимента. Показано, что в случае, когда целью исследования является предсказание поведения некоторой полимерной системы, содержательный подход предпочтителен. В случае, когда исследователь желает сравнить активность различных АО, формальная модель лучше отвечает такой постановке.

10. Применение метода ПИО к задачам классификации на примере распознавания фальсифицированных лекарств с помощью ИК-спектроскопии в ближней области

Многомерный подход эффективно используется в задачах классификации. В этой главе проводится сравнение известных методов: МГК и SIMCA и нового подхода, объединяющего известный метод ПЛС дискриминации (M. Sjöström, 1986) с методом ПИО. Сравнение проводится на примере распознавания фальшивых лекарственных средств с помощью БИК-спектроскопии.

10.1. Комбинированный метод: ПЛС дискриминация и метод ПИО

Для разделения объектов на Q различных классов используется калибровочный набор, включающий объекты из всех Q классов. В качестве предикторов используется $\mathbf{X}$ ($I\times J$) – матрица признаков (например, БИК-спектры). В качестве откликов $\mathbf{Y}$ вводится матрица искусственных переменных, т.н. матрица принадлежности классу. Число столбцов в $\mathbf{Y}$ равно числу классов Q. Для всех объектов из класса q (q=1,…, Q), y_q равно 1, а для остальных –1. Затем строится ПЛС2 модель, и для нового объекта вычисляется прогноз, по которому определяют при-

надлежность объекта к классу. Предлагается дополнить метод ПЛС дискриминации методом ПИО. Для калибровочного набора ПИО метод позволяет очертить границы классов, а для новых объектов – оценить их близость к классу.

10.2. Эксперимент 1. Исследование таблеток - БИК спектры диффузного рассеяния

Исследовались образцы пищеварительного фермента в виде таблеток, всего 75 объектов: 11 серий подлинных (G1 – G11) и 4 серии фальсифицированных таблеток (F1 – F4), по 5 таблеток в серии. В качестве матрицы **X** использовались БИК спектры диффузного рассеяния $R(\lambda)$ на участке 4000–7500 см$^{-1}$. (1750 волновых чисел), преобразованные как –log$R$. МГК и SIMCA не дают надежного разделения на классы. Для того чтобы повысить надежность, основное внимание при моделировании должно уделяться различию между классами, а не индивидуальным особенностям объектов внутри класса.

Для этого использовался метод ПЛС дискриминации совместно с ПИО методом. Калибровочный набор формировался как из подлинных (G1-G3, по 5 таблеток), так и из фальшивых (F1, F2 по 4 таблетки) образцов. Построенная модель, надежно различает фальшивые и настоящие таблетки, а так же дает дополнительную информацию о различных сериях образцов. Результаты предсказания на проверочном наборе представлены на Рис. 10.1. Образцы серий G надежно распознаются как подлинные, а серий F – как фальшивые.

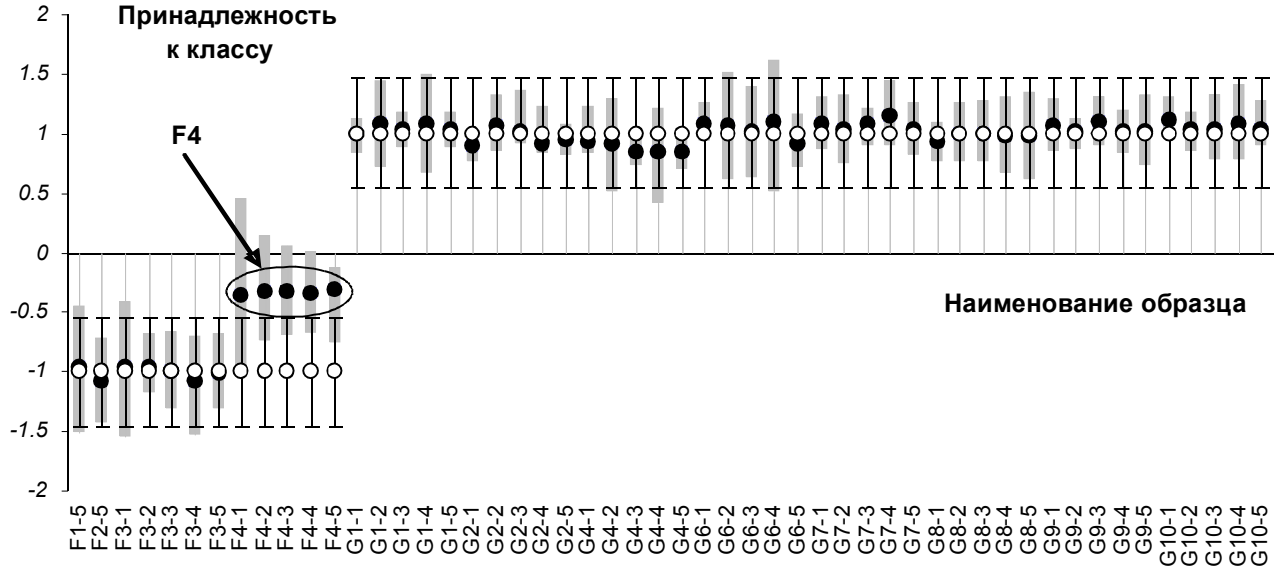


Рис. 10.1 БИК-спектроскопия, определение фальсифицированных лекарств (таблетки), проверочный набор. ПЛС модель с 2 ГК, b_{min}=0.23 и b_{SIC}=0.46. Черный интервал –ПИО калибровка, серый – ПИО предсказанный , (○) – опорные значения, (●) –ПЛС предсказанные значения

Для образцов серии F4 можно заключить следующее: (1) их нельзя отнести к классу подлинных образцов; (2) они существенно отличаются от серий F1-F3. Для фальсифицированных образцов характерен больший разброс внутри серии, а также между сериями, по сравнению с подлинными образцами.

10.3. Эксперимент 2. Исследование ампул - БИК спектры пропускания

Исследовался глюкокортикостеройдный препарат в ампулах с 4% водным раствором активного вещества; две серии подлинных образцов, G1 и G2, и одна

серия поддельных, F1, по 15 ампул в каждой серии. Измерялись БИК спектры пропускания, ампулы не вскрывались. Использовались две спектральные области: 5500 – 6400 $см^{-1}$ и 7200 – 9000 $см^{-1}$, всего 702 волновых числа. Так же, как и в предыдущем примере, результаты применения МГК и SIMCA не дают надежного разделения на классы, часть образцов подлинных лекарств из проверочного набора классифицируется как не входящие в этот класс.

ПИО моделирование показывает, что величина интервала калибровки достаточно велика, т.е. построенная модель несет в себе существенную неопределенность. Сходными являются и результаты распознавания для проверочного набора.

Результаты главы 10. Математическая обработка результатов БИК-спектроскопии позволяет разработать быструю и не требующую специальной пробоподготовки процедуру распознавания фальшивых лекарств.

Проведено сравнение различных методов классификации. Показано, что точечных оценок, получаемых методом ПЛС дискриминации, не достаточно для надежного разделения классов и распознавания новых объектов, т.к. понятие "близости" к классу должно иметь численное выражение. Дополнение ПЛС дискриминации методом ПИО дает следующие преимущества: (1) интервал калибровки позволяет очертить точную границу каждого класса; (2) интервал предсказания позволяет численно охарактеризовать близость объекта к тому или иному классу; (3) ПИО классификация статуса объектов позволяет охарактеризовать однородность объектов внутри класса, а так же выявить группы объектов с особыми свойствами, отличающими их от объектов предопределенных классов.

11. Методы анализа процессов

Современный многомерный контроль процессов 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. Для осуществления многомерного статистического контроля процессов – МСКП (MacGregor, 1995), собирается информация об изучаемом процессе: инструментальные показатели **X** и выходные переменные, **Y**. На основе набора (**X**, **Y**) строится линейная модель калибровки, с помощью которой проверяется, находится ли процесс внутри допустимых границ. Эта глава посвящена расширению метода МСКП. Предлагается подход, определяющий действия по оптимизации процесса в режиме *in-line*, названный многомерной статистической оптимизацией процессов (МСОП). Для его реализации используется сочетание ПЛС регрессии и метода ПИО.

11.1. Описание исследуемого процесса

Теоретические разработки иллюстрируются примером многостадийного технологического процесса, представленного 25 инструментальными переменными **X** (J=25), и одной выходной переменной **y**, характеризующей "качество" результата. Данные (**X**, **y**), состоят из I=154 объектов (наблюдений). Весь процесс разделен на 7 стадий (L=7), каждую из которых можно описать входными,

текущими и выходными переменными. Набор данных (**X**, **y**) разделен (по столбцам) на L блоков, соответствующих стадиям процесса: $\mathbf{X}=(\mathbf{X}_{\mathrm{I}}, \mathbf{X}_{\mathrm{II}},\ldots, \mathbf{X}_{\mathrm{L}})$. Последний блок L+1 состоит из переменной Y=**y**. Данные преобразованы так, что каждая переменная, включая y, изменяется в интервале (–1, +1). Значения вне этого интервала считаются недопустимыми. Предполагается, что наивысшее качество характеризуется y=+1, а наихудшее соответствует y=–1. Данные также разделены на калибровочный набор (102 объекта), и проверочный набор (52 объекта).

11.2. Контроль процесса

Исследуется *пассивная оптимизация* новым методом *расширяющегося МСКП*, опирающимся на мульти-блоковую регрессию (A. Höskuldsson, 2001).

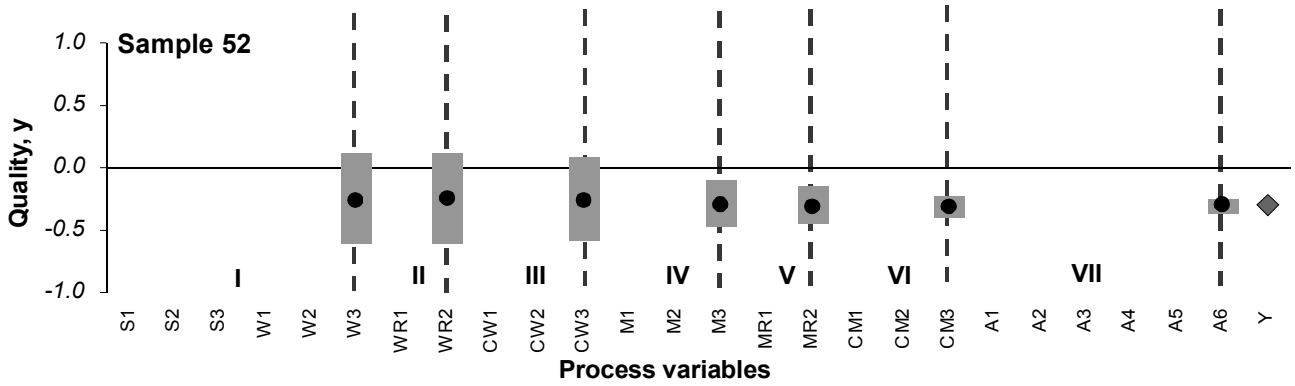


Рис. 11.1 Предсказание показателя «качество» на каждой стадии процесса для объекта из проверочного набора, ПИО интервал (серый прямоугольник), ПЛС прогноз (черные кружки). Ромб в правой части – измеренное значение y.

Используя весь набор данных, можно построить полную ПЛС модель

$$\mathrm{XY}: \mathbf{X} \Rightarrow \mathbf{y}\,, \qquad (11.1)$$

в которой задействовано K главных компонент. На тех же данных можно построить серию из L–1 ПЛС регрессионных моделей

$$\mathrm{XY_I}: \mathbf{X}_{(\mathrm{I})} \Rightarrow \boldsymbol{y}, \qquad \mathrm{XY_{II}}: \mathbf{X}_{(\mathrm{II})} \Rightarrow \boldsymbol{y}, \qquad \ldots, \qquad \mathrm{XY_{L\text{-}1}}: \mathbf{X}_{(\mathrm{L\text{-}1})} \Rightarrow \boldsymbol{y}. \qquad (11.2)$$

Здесь каждая модель обозначается оператором $\mathrm{XY_M}$, который представляет регрессию X-блока, $\mathbf{X}_{(\mathrm{M})}$, на Y-блок, **y**. Все модели (11.2) используют одно и то же число ГК, которое выбирается при анализе полной модели (11.1).

Целью моделирования является предсказание выходной переменной **y** на каждой (M-ой) стадии процесса. Для оценки неопределенности применяется метод ПИО. Результат *расширяющегося МСКП* приведен на Рис. 11.1.

11.3. Оптимизация процесса

Рассматривается задача выбора значений инструментальных переменных по ходу процесса: определение переменных $\mathbf{X}_{(\mathrm{M})}$, которые являются входными для следующей (M-ой) стадии процесса. При этом необходимо придерживаться двух основных принципов. Новые величины переменных, во-первых, должны повышать значение y; и, во-вторых, значения этих переменных должны находиться внутри допустимых контролируемых границ. Предлагаемый подход базируется на концепции статуса объектов метода ПИО.

В общем случае рассматриваются два блока инструментальных переменных

– **X** и **Z**, и соответствующий им вектор **y**. Целью является предсказания величины *y* для набора переменных (**x**, **z**), в котором значения **x** известны, а **z** неизвестны. Требуется найти такие значения **z**, которые будут оптимизировать *y*, при условии, что $\mathbf{z} \in L_z$ – области допустимых значений **z**.

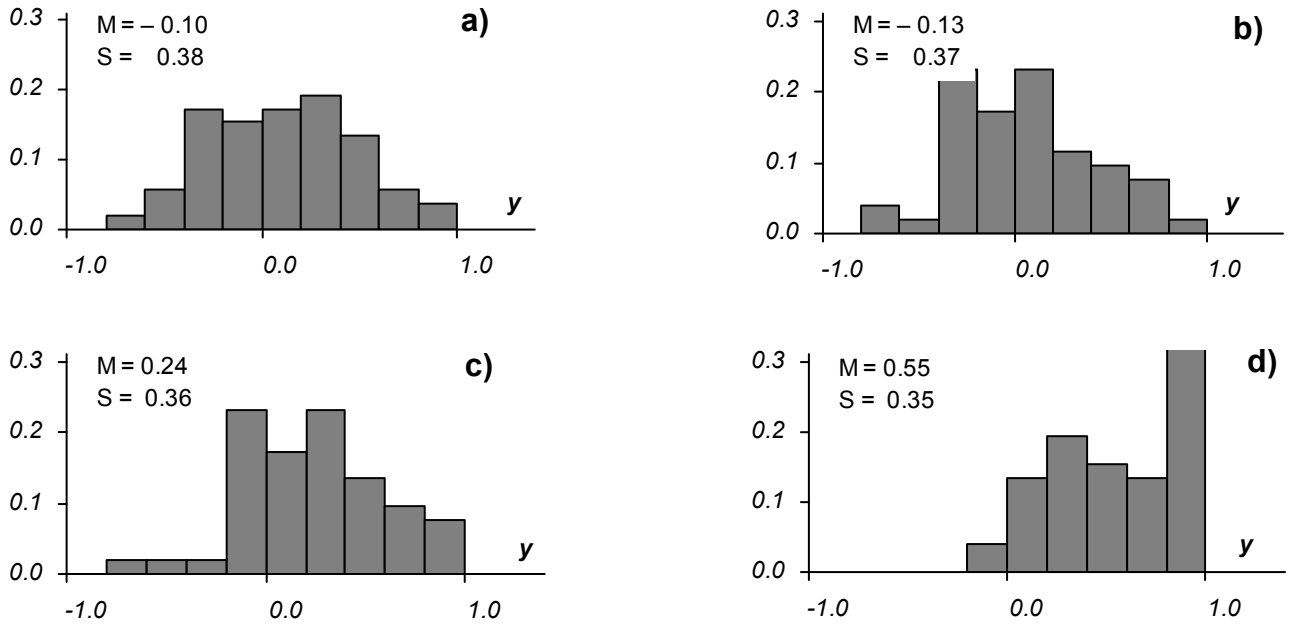


Рис. 11.2 Распределение объектов по переменной *y*.
a) Контрольный набор (до оптимизации), b) Оптимизация, тип «внутренний» - $G1(\hat{\mathbf{z}})$,
c) Оптимизация, тип «внешний» - $G2(\hat{\mathbf{z}})$, d) Оптимизация, тип «выбросы» - $G3(\hat{\mathbf{z}})$

Вычисление оптимальных значений $\mathbf{z}^+$ происходит в два этапа. На первом этапе, с использованием исторических данных (**X**, **Z**), строится ПЛС2 калибровка, и оцениваются значения $\hat{\mathbf{z}} = \mathrm{XX}(\mathbf{x})$. По построению, $\hat{\mathbf{z}}$ является допустимым решением (принадлежит области L_z). Вторым этапом является изменение каждого компонента вектора $\hat{\mathbf{z}}$ до тех пор, пока новый вектор $\mathbf{z}^+$ остается в пределах области L_z. Это действие можно представить с помощью оператора G: $G(\hat{\mathbf{z}}) = \mathbf{z}^+$, который определяет стратегию оптимизации. ПИО классификации статуса объектов дает инструмент, с помощью которого можно выбрать различные стратегии оптимизации. На Рис. 11.2 представлены результаты применения трех различных стратегий к набору из 52 объектов процесса.

Результаты главы 11. Разработан новый метод многомерного контроля процессов, расширяющийся МСКП, основанный на построении серии ПЛС моделей, совместно с ПИО моделированием. ПЛС модели на каждой стадии процесса предсказывают точечные оценки выходного параметра, а ПИО метод добавляет к этой оценке интервал предсказания. Такой подход помогает осуществить пассивную оптимизацию.

Разработан новый метод многомерной оптимизации процессов, который основывается на блочном ПЛС и ПИО методе. Показано, что для улучшения выходного показателя *y* необходимы такие корректирующие действия, которые, с одной стороны, остаются в рамках изучаемого процесса, а, с другой, требуют вывода значений контролируемых переменных на границы возможных значе-

ний. Предложенный подход включает регулирование контролируемых инструментальных переменных на промежуточных стадиях, и предлагает набор стратегий.

12. Формирование представительной выборки объектов применительно к различным наборам многоканальных экспериментов

При решении задачи переноса калибровок с одного прибора на другой, при работе с большими наборами данных физических экспериментов, и в других случая, возникает потребность выбрать из общего набора (**X**,**Y**), короткий, но представительный набор. Этот набор должен отвечать двум требованиям: во-первых, он должен представлять вариабельность полного набора данных; во-вторых, число объектов в новой выборке должно быть существенно меньше, чем в исходной.

12.1. Теория

В этой главе рассматривается применение ПИО для формирования представительной выборки – метод граничных объектов. Проводится сравнение результатов с двумя наиболее известными методами отбора: методом Кеннарда-Стоуна и D-оптимальным планированием.

Метод граничных объектов. В соответствии с методом ПИО, все калибровочные объекты являются внутренними (Опр. 6.1). Среди них выделяются граничные объекты (Опр. 6.2), которые формируют ОДЗ *A*. Поэтому набор граничных объектов составляет представительную выборку.

Метод Кеннарда-Стоуна осуществляет выбор объектов «равномерно» по всей области. В нем рассматривается только матрица **X**, значения **y** не учитываются. К достоинствам алгоритма надо отнести: (1) простоту реализации; (2) алгоритм может применяться к любой матрице **X**, независимо от ее ранга.

D-оптимальный план выбирает объекты так, чтобы максимизировать определитель информационной матрицы регрессии. Если число переменных в **X** превышает число объектов, D-оптимальную процедуру можно применять только после регуляризации задачи.

Обозначения и схема исследования. Эффективность приведенных методов исследуется на трех различных наборах многоканальных экспериментальных данных. Для этого используется несколько показателей: *RMSEC*, *RMSEP*, ПИО остатки (6.1) и ПЛС остатки для Y-переменных $r_{\text{PLS}} = y - \hat{y}$; ПИО размах (6.2) и ПЛС размах (3.4). Применяется одна и та же схема построения моделей (12.1)

1. Строится ПЛС модель, Модель_Q, на основе Q набора, с фиксированным числом ГК, и соответствующая ПИО модель, с фиксированным значением b_{SIC}. Здесь Q – это B, или K, или D.
2. Модель_Q проверяется с помощью проверочного T набора.
3. Модель_Q используется для предсказания объектов из избыточного набора RQ.
4. Результаты калибровки и предсказания сравниваются с результатами, полученными для Модели_C.

(12.1)

и способы выбора и проверки поднаборов (Рис. 12.1).

12.2. Эксперименты

Для сравнения различных методов анализируются следующие экспериментальные данные. Первый – это БИК спектры, используемые для определения содержания влаги в зернах пшеницы для 139 объектов. Спектры пропускания – $\lg T(\lambda)$ – были получены в диапазоне 908–1120 nm (118 длин волн). Второй пример – это определение следовых концентраций нефти в воде (раздел 8.1), третий пример – данные по многомерному контролю процесса (раздел 11.1).

12.3. Сравнение репрезентативности различных выборок

Для того чтобы сравнить подход, основанный на граничных объектах с уже известными методами, подробно изучался первый пример. Для этого, 10 раз повторялась следующая процедура.

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 строится 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 (12.1).

Результаты моделирования (Таб. 12.1) подтверждают эффективность ПИО подхода. Применение метода граничных объектов к двум другим примерам приводит к аналогичным результатам.

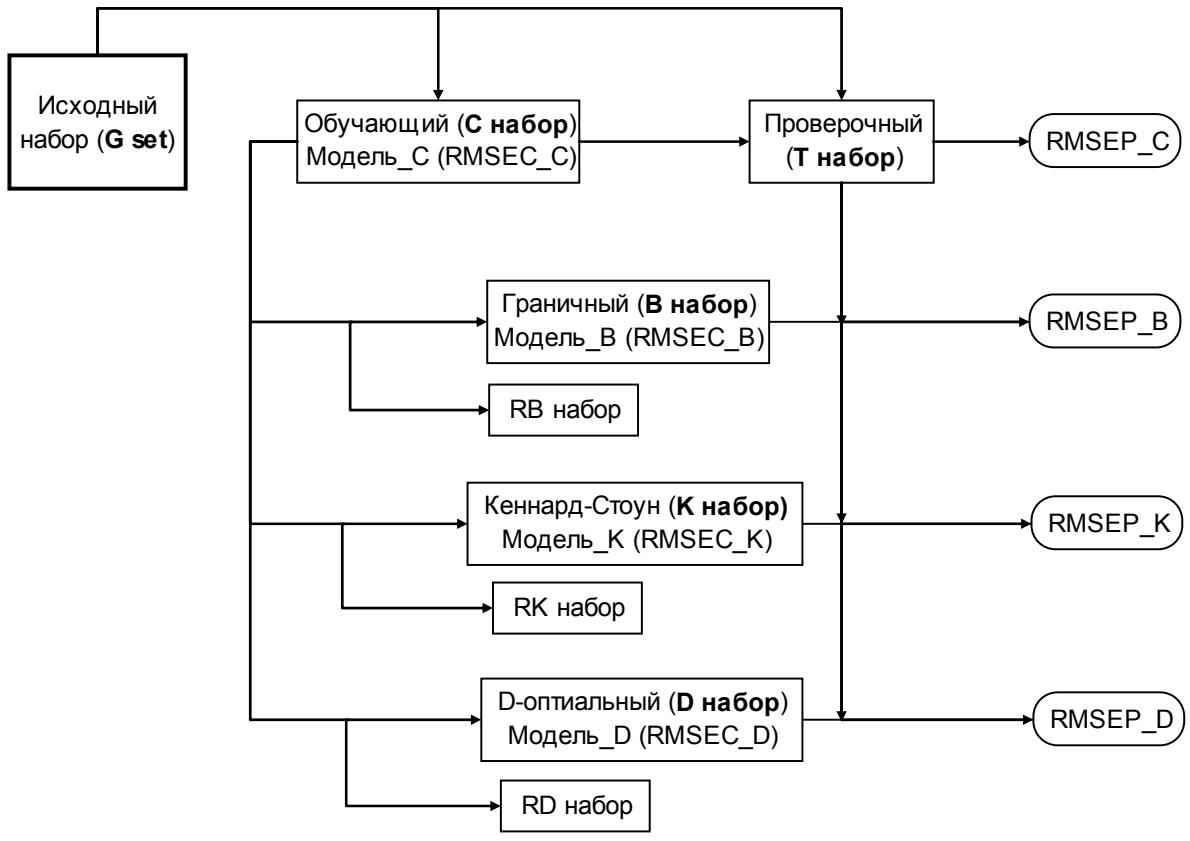


Рис. 12.1 Исследуемые наборы и соответствующие им модели

Показано, что предположение о граничных объектах, как наиболее влия-

тельных, подтверждается еще и тем фактом, что предсказание "избыточных" объектов, составленный из всех внутренних объектов калибровочного набора, осуществляется с наименьшей ошибкой предсказания, а все "избыточные" объекты классифицируются как внутренние.

Таб. 12.1 БИК-спектроскопия, определение влажности зерна ПЛС модели с 4 ГК. Средние значения по 10 калибровочным/проверочным наборам

#	I_B	Модель_C		Модель _B		Модель _K		Модель _D	
		RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP
Среднее		0.287	**0.293**	0.442	**0.308**	0.248	**0.314**	0.251	**0.322**

12.4. Зависимость точности предсказания от объема выборки

В некоторых случаях объем представительной выборки, определенной по методу граничных объектов, может показаться излишне большим. Так, например, для третьего набора данных, представляющих результаты многомерного контроля процессов, он составил 45% (46 из 102 объектов). Поэтому важно исследовать, как влияет объем выборки на предсказательные свойства модели. Согласно методу ПИО, минимальное число граничных объектов определяется при $\beta = b_{min}$. В рассматриваемом примере этот набор состоит из 8 объектов($I_B \geq 8$). Последовательно увеличивая b с $b = b_{min}$ до $b = b_{SIC}$, получаем расширяющийся B набор.

Параллельно, для сравнения, применяя метод Кеннарда-Стоуна и D оптимальное планирование, выбираются K - 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.

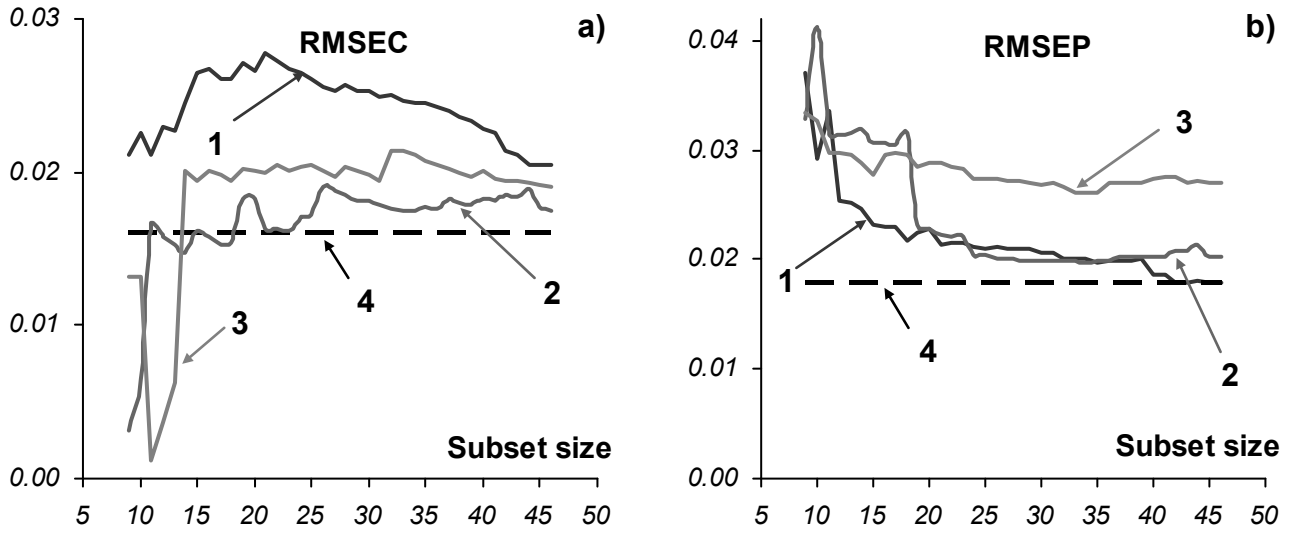


Рис. 12.2. Многомерный контроль процесса. ПЛС модели с 7 ГК. Зависимость RMSE от объема выборки 1 – Модель_B, 2 – Модель _K, 3 – Модель _D, 4 – Модель_C

Величины *RMSEC* и *RMSEP* можно рассматривать как функции, зависящие от объема выборки (Рис. 12.2), которые вычисляются для трех моделей (Мо-

дель_B, Модель_K, Модель_D). Из Рис. 12.2b, кривая 1, видно, что для B наборов функция $RMSEP(I_B)$ убывает быстрее, чем для наборов K и D, и стремится к предельному значению, $RMSEP(I_C)$. При этом отклонение в калибровке (Рис. 12.2a, кривая 1) $RMSEC(I_B)$ остается наибольшим по сравнению с аналогичными значениями, вычисленными для наборов K и D, т.е. B-набор аккумулирует в себе наиболее влиятельные калибровочные объекты.

Из Рис. 12.2 b также видно, что для формирования представительного короткого набора необходимо не менее 42 объектов. Это подтверждает, что объем выборки, предлагаемый методом ПИО близок к оптимальному.

Результаты главы 12. В этой главе подробно рассмотрен новый метод формирования представительной выборки. Метод граничных объектов основывается на методе ПИО (теория классификации статуса объектов), объединенным с проекционными методами (РГК, ПЛС). Показано, что стратегия выбора граничных объектов является объективной, т.е. не требует никакой дополнительной информации. Метод граничных объектов имеет следующие преимущества. Во-первых, он однозначно определяет необходимое число объектов в представительной выборке. Во-вторых, при отборе объектов, учитывается информация, как о значениях приборных X- переменных, так и Y- переменных.

Три исследованных набора многоканальных данных были порождены различными практическими задачами, они отличаются друг от друга, как по внутреннему устройству, так и по сложности построенных ПЛС моделей. Это доказывает эффективность метода для анализа различных физических экспериментов.

Основные теоретические и прикладные результаты работы

В работе рассмотрены теоретические, алгоритмические и методологические аспекты метода простого интервального оценивания (ПИО) в применении к обработке больших массивов данных многоканальных экспериментов. Обобщая полученные результаты, можно сформулировать следующие выводы:

1. Объединение проекционных регрессионных методов с методом простого интервального оценивания порождает мощный инструмент для решения задач многомерной калибровки. Такой подход позволяет обрабатывать большие наборы данных физических экспериментов, пронизанных внутренними связями, разделять полезную информацию и шум, представлять результат прогноза в интервальной форме, учитывающей неопределенность в прогнозе индивидуально для каждого объекта/измерения.

2. Предположение об ограниченности погрешностей, лежащее в основе метода ПИО, является не недостатком, а преимуществом метода, так как, с практической точки зрения, оно более обоснованно, чем традиционное допущение о нормальности, и, следовательно, неограниченности погрешностей.

3. Приведены аргументы в пользу того, что ПИО-оценки, построенные на основе экстремальных статистик, являются более эффективными, чем традиционные гладкие оценки.

4. На основе метода ПИО разработан новый подход к классификации статусов

объектов и интерпретации прогнозных интервалов. Введены новые понятия: ПИО-остаток и ПИО-размах, диаграмма статуса объектов (ДСО). Даны определения понятиям внутренние, внешние, граничные объекты. Дано определение выбросов и абсолютно внешних объектов. Показано, что разработанная классификация имеет практическое значение и в рамках классических регрессионных моделей. Диаграмма статусов объектов является удобным инструментом для визуального анализа сложных сигналов. Эффективность предложенного подхода продемонстрирована на ряде примеров, в том числе, на примере многоканальных акустических измерений для определения следовых концентраций нефти в воде.

5. Разработаны новые методы статистического контроля процессов. Метод, названный расширяющимся многомерном 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 метод активной оптимизации, разработаны различные стратегии оптимизации.

6. Предложен новый метод выбора представительных (влиятельных) объектов из экспериментального набора данных, названный методом граничных объеектов, 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.

7. Проведено сравнение формального (ПИО) и содержательного (нелинейная регрессия) моделирования. Показано, что содержательный подход позволяет проводить экстраполяцию, однако не ограничивает ее область. Формальный метод имеет строгую область применимости, определенную с помощью диаграммы статуса объектов. Показана практическая значимость построенных моделей, позволяющих заменить длительное термостарение быстрым экспериментом с помощью дифференциальной сканирующей калориметрии.

8. Показано, что дополнение стандартного метода ПЛС дискриминации методом ПИО повышает информативность при решении задач классификации. Предложена методика экспресс-распознавания фальсифицированных лекарств на основе БИК-спектроскопии.

9. Компьютерная программа SIC позволяет на практике применить предложенную методику, объединяющую проекционные регрессионные методы и ПИО моделирование. С ее помощью можно проводить обработку наборов многоканальных сигналов, оценивать точность калибровки, проводить классификацию объектов.

Основное содержание диссертации опубликовано в работах:

1. Павлов Б.В., Родионова О.Е. Математическое моделирование сложных самоускоряющихся реакций. *Теор. основы хим. технологии*, **28**, 251-258 (1994).
2. Павлов Б.В., Родионова О.Е. Численное решение систем линейных обыкно-

венных дифференциальных уравнений с постоянными коэффициентами. *Ж. вычисл матем. и матем. физ.*, **34**, 622-627 (1994).

3. Павлов Б.В., Родионова О.Е. Методика усреднения при дискретизации кинетического интегро-дифференциального уравнения. *Ж. вычисл. матем. и матем. физ.*, **36**, 143-161 (1996).
4. Павлов Б.В., Родионова О.Е. Проблемы математического моделирования в неравновесной теории химических процессов. *Хим. физ.*, **17**, 27-40 (1998).
5. Bystritskaya E.V., Pomerantsev A.L., Rodionova O.Ye. Prediction of the aging of polymer materials. *Chemom. Intell. Lab. Syst.* , **47**, 175-179 (1999).
6. Bystritskaya E.V., Pomerantsev A.L., Rodionova O.Ye. Evolutionary Design of Experiment for Accelerated Aging Tests. *Polymer Testing*, **19**, 221-229 (1999).
7. Pomerantsev A.L., Rodionova O.Ye. Chemometrics in Russia. *Chemom. Intell. Lab. Syst*, **48**, 121-129 (1999).
8. Bystritskaya E.V., Pomerantsev A.L., Rodionova O.Ye. Nonlinear Regression Analysis: New Approach to Traditional Implementations. *J. Chemometrics*, 14, 667-692 (2000).
9. Зобов В.Е., Лундин А.А., Родионова О.Е. К теории формы спектров ядерного магнитного резонанса в гетероядерных спиновых системах. *Хим.физ.* **19** (2), 39-43, (2000).
10. Зобов В.Е., Лундин А.А., Родионова О.Е. К теории форм спектров ЯМР в спиновых системах с двумя сортами яде. *Хим. физ.*, **19** (12), 26-40 (2000).
11. Зобов В.Е., Лундин А.А., Родионова О.Е. О форме спектров поглощения ЯМР и кросс релаксации в гетероядерных спиновых системах. *ЖЭТФ*, **120**, 619-636 (2001).
12. Померанцев А.Л., Кротов А.С., Родионова О.Е. *Компьютерная система FITTER для регрессионного анализа экспериментальных данных*, Учебное пособие, Барнаул, Из -во АГУ, 2001.
13. Померанцев А.Л., Родионова О.Е. Надстройка FITTER (FITTER). *Свидетельство об официальной регистрации № 2002611562 от 11.09.02*.
14. Pomerantsev A.L., Rodionova O.Ye. Prediction of Antioxidants Activity Using DSC Measurements. A Feasibility Study. In *Aging of polymers, polymer blends and polymer composites*, Eds: .E. Zaikov, A.L. Buchachenko and V.B. Ivanov, **2**, 19-29, Nova science Publishers, NY, 2002 (ISBN 1-59033-256-3).
15. Родионова О.Е., Померанцев А.Л. Об одном методе решения обратной кинетической задачи по спектральным данным при неизвестных спектрах компонент. *Кинетика и катализ*, **45**, 485-497 (2004).
16. Rodionova O.Ye., Pomerantsev A.L. Prediction of Rubber Stability by Accelerated Aging Test Modeling. In *Leading Edge Research on Polymers and Composites*, Eds: Monakov et al, cc. 105-124, Nova science Publishers, NY 2004, (ISBN:1-59033-975-4).
17. Rodionova O. 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. Chemometrics*, **18**, 402-413 (2004).

18. Rodionova O.Ye., Pomerantsev A.L. Principles of Simple Interval Calculations. In: *Progress In Chemometrics Research*, Ed.: A.L. Pomerantsev, 43-64, NovaScience Publishers, NY, 2005, (ISBN: 1-59454-257-0.)
19. Pomerantsev A.L., Rodionova O.Ye. Multivariate Statistical Process Control and Optimization. *Там же*, 209-227.
20. Semenchenko A.S., Semenchenko S.M., Rodionova O.Ye., Pomerantsev A.L. Explanatory data analysis of fish culture in Siberian lakes. *Там же*, 313-321.
21. Rodionova O.Ye., Pomerantsev A.L. Prediction of Rubber Stability by Accelerated Aging Test Modeling. *J Appl Polym Sci*, **95**, 1275-1284 (2005).
22. Померанцев А.Л., Родионова О.Е. Содержательный и формальный подход к анализу кинетических данных. В сб. *Химическая и биологическая кинетика. Новые горизонты*. М. Химия, **1**, 124-172, 2005 (ISBN: 5-98109-035-9).
23. Родионова О.Е., Померанцев А.Л. Оценивание параметров в уравнении Аррениуса. *Кинетика и катализ*, **46**, 329–332 (2005).
24. Pomerantsev A.L., Rodionova O.Ye. Hard and soft approaches to analysis of kinetic data. In: *Chemical and Biochemical kinetics. New horizons*, Eds. E.B. Burlakova, A.E. Shilov, S.D. Varfolomeev, G.E. Zaikov, Brill Academic Publishers, Leiden-Boston, **1**, 80-107, 2005.
25. Rodionova O.Ye., Houmøller L.P., Pomerantsev A.L., Geladi P., Burger J., Dorofeyev V.L, Arzamastsev A.P. NIR spectrometry for counterfeit drug detection. *Anal. Chim. Acta*, **549**, 151-158 (2005).
26. 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**, 73-83 (2005).
27. Pomerantsev A.L., Rodionova O.Ye., Höskuldsson A. Process Control and Optimization with Simple Interval Calculation Method. *Chemom. Intell. Lab. Syst.*, **81**, 165-179 (2006).
28. Родионова О.Е. Хемометрический подход к исследованию больших массивов химических данных. *Рос. хим. ж.* (*Ж. Рос. хим. об–ва им. Д.И. Менделеева*), **50**, 128-144 (2006).
29. Померанцев А.Л., Родионова О.Е. О двух подходах к анализу кинетических данных на примере предсказания активности антиоксидантов. *Кинетика и катализ*, **47**, 553–565 (2006).
30. Померанцев А.Л., Родионова О.Е. Построение многомерной градуировки методом простого интервального оценивания. *Ж. аналит. химии*, **61**, 1032-1047 (2006).
31. Родионова О.Е., Померанцев А.Л. Хемометрика: достижения и перспективы. *Успехи химии*, **75,** 302-317 (2006).
32. 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)
33. Höskuldsson A., Rodionova O. Ye., Pomerantsev A.L. Path modeling and process control. *Chemom. Intell. Lab.Syst.,* **88**, 84-99 (2007).