Построение многомерной градуировки методом простого интервального оценивания

Померанцев А.Л., Родионова О.Е.

Институт химической физики РАН, Москва, 119991, Косыгина 4, polycert@chph.ras.ru

Простое интервальное оценивание (ПИО) – это метод линейного моделирования и построения интервальных оценок прогноза в многомерной градуировке. 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 дает новые возможности для построения содержательной классификации влиятельности образцов. Метод основывается на единственном предположении об ограниченности ошибок и использует алгоритмы линейного программирования для анализа данных. Такой подход значительно отличается от традиционных регрессионных методов используемых в хемометрике и потому с трудом воспринимается аналитиками. В статье дается простое объяснение метода ПИО, иллюстрированное простейшими модельными и реальными примерами.

Ключевые слова: многомерная градуировка, интервальное оценивание, метод ПИО, линейное программирование, классификация образцов.

Введение

Описание экспериментальных данных, построение модели и предсказание новых значений, т.е. то, что кратко называется градуировкой или калибровкой, – это одна из старейших, но вечно актуальных задач, которая активно применяется в аналитической химии [1]. Суть задачи многомерной градуировки (ММГ) состоит в следующем. Пусть имеются экспериментальные данные, представленные двумя матрицами: **X** – это матрица аналитических сигналов (например, спектров) и **Y** – это матрица соответствующих химических показателей (например, концентраций). Число строк в этих матрицах равно количеству исследованных образцов, число столбцов в матрице **X** соответствует числу каналов (длин волн), на которых записывается сигнал, и, наконец, число столбцов в матрице **Y** равно числу химических показателей, т.е. откликов. На основе набора образцов сравнения {**X,Y**}, требуется построить математическую модель **Y**=**X*a***, с помощью которой можно предсказывать новые значения откликов ***y*** по заданной, новой строке значений аналитического сигнала ***x***. Оценивание такой модели – это сложная, некорректно поставленная математическая задача [6], однако применение именно многомерной модели дает значительный выигрыш в точности по сравнению с простой градуировкой по нескольким «характеристическим каналам» [2].

Со времен Гаусса (1794) для анализа экспериментальных данных использовался регрессионный подход, в основе которого лежит принцип минимизации отклонений

вычисленных модельных значений $\hat{y}$, от соответствующих экспериментальных величин y – метод наименьших квадратов [3]. Развитие этого подхода – метод главных компонент (1901) [4], метод максимума правдоподобия (1912) [5], ридж-регрессия (1963) [6], проекция на латентные структуры (1975) [7], и др. – позволили применять его для сложных, некорректно поставленных задач, например в спектроскопии, где число неизвестных параметров (длин волн) значительно больше, чем число исследованных образцов [8]. Однако все эти методы дают результат предсказания в виде точечной оценки, тогда как на практике часто нужна интервальная оценка, учитывающая неопределенность прогноза. Построение доверительных интервалов традиционными статистическими методами невозможно из-за сложности задачи [9], а использование имитационных методов затруднительно из-за большого времени расчетов [10].

В 1962 Канторович [11] предложил другой подход к анализу данных – заменить минимизацию суммы квадратов отклонений на систему неравенств, которая решается с помощью методов линейного программирования. В этом случае результат прогноза сразу имеет вид интервала, поэтому этот метод и был назван «*простым интервальным оцениванием*» (ПИО). В свое время эта идея не получила должного признания и развития, что было связано, по-видимому, с недостаточным быстродействием компьютеров. В 80-90-ые годы с помощью этого метода было выполнено несколько интересных прикладных работ [12-19], в том числе и в области аналитической химии [18]. Итоги этих исследований были подведены в монографии [20], где подробно рассматривается основная задача, решаемая авторами упомянутых выше работ. Это – задача интервальной оценки *параметров* моделей, погружение области возможных значений этих параметров в гиперкуб, параллелепипед, эллипсоид, и т.п.

Такая постановка задачи представляется нам неплодотворной и малоперспективной, что и было подтверждено практикой – за последние 10 лет новые работы в этом направлении не замечены. В тоже время мы считаем, что идея Канторовича может дать интересные результаты, если рассматривать ММГ как задачу построение интервального прогноза *отклика y*. В этом случае удается решить две равно важные практические задачи. Во-первых, установить область неопределенности [21] для прогноза искомого отклика (химического показателя), т.е. оценить *точность* построенной градуировки, индивидуально для каждого образца. Во-вторых, используя подход ПИО, можно построить *классификацию* образцов [22], т.е. установить индивидуальные особенности каждого образца, определяемые по его взаимоотношениям, как с моделью, так и с другими образцами сравнения. Общеизвестными примерами такой классификации

являются такие понятия как *выброс* (образец, резко выделяющийся из общей закономерности) или *экстремальный образец* (находящийся в периферийной области модели и оказывающий значительное влияние на ее построение). Не смотря на широкое употребление этих понятий в различных исследованиях [23-31], не существует их общепризнанных определений и методов обнаружения. Метод ПИО может восполнить этот пробел.

Теоретические аспекты метода ПИО были опубликованы в работах [22, 32], а результаты его практического применения нашли свое отражение в публикациях [21, 33, 34, 43]. Однако этот метод значительно отличается от традиционного, привычного регрессионного подхода, применяемого в задачах ММГ. Его «философия», математический аппарат, терминология непривычны для аналитиков. Исходя из этого, мы предлагаем вниманию читателей максимально простое объяснение метода ПИО, основанное на простейших одно- и двумерных примерах, с помощью которых, тем не менее, удается объяснить и продемонстрировать все основные идеи и результаты.

В первой части работы мы представляем доводы, обосновывающие основной постулат метода ПИО – ограниченность ошибок. Приводятся как теоретические, так и практические аргументы в защиту этого постулата. Вторая часть статьи посвящена подробному рассмотрению простейшего модельного примера, в котором все необходимые вычисления можно провести с помощью карандаша и бумаги. С его помощью вводятся и иллюстрируются основные понятия, используемые в ПИО методе, и показывается, к каким выводам приводит последовательное применение принципа ограниченности ошибок. В третьей части работы разбирается реальный пример – предсказание октанового числа бензина по ИК спектру [35, 36] – классическая задача ММГ. В Приложении дается формальное, математически строгое описание метода ПИО.

1. Почему ошибки ограничены

Основным предположением ПИО является ограниченность ошибки измерения. Такой подход к интерпретации экспериментальных данных нуждается в некотором обосновании. При анализе данных, стандартным допущением является принцип нормальности ошибок. Это либо явно, либо молчаливо предполагается. Однако допущение о нормальном распределении ошибки неоднократно подвергалось критике с самых разных позиций. Многочисленные исследования (см. например, [37, 38]) показывают, что обычно ошибка измерения скорее ограничена, чем нормальна. Характерно, что большинство аналитиков не связывают с принципом нормальности факт

неограниченности ошибок. На прямой вопрос о том, как часто исследователю приходилось обрабатывать данные, 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 аналитики, часто превышает 10^{+6} [39], так что в них уверенно можно было бы ожидать 20-30 «нормальных» значений, выходящих за этот порог. Примечательно мнение авторов работы [30], которые отмечают: «действительно, в реальных исследованиях химик часто может, в какой-то степени, отобрать образцы, что приводит к распределению, которое скорее равномерно, чем нормально».

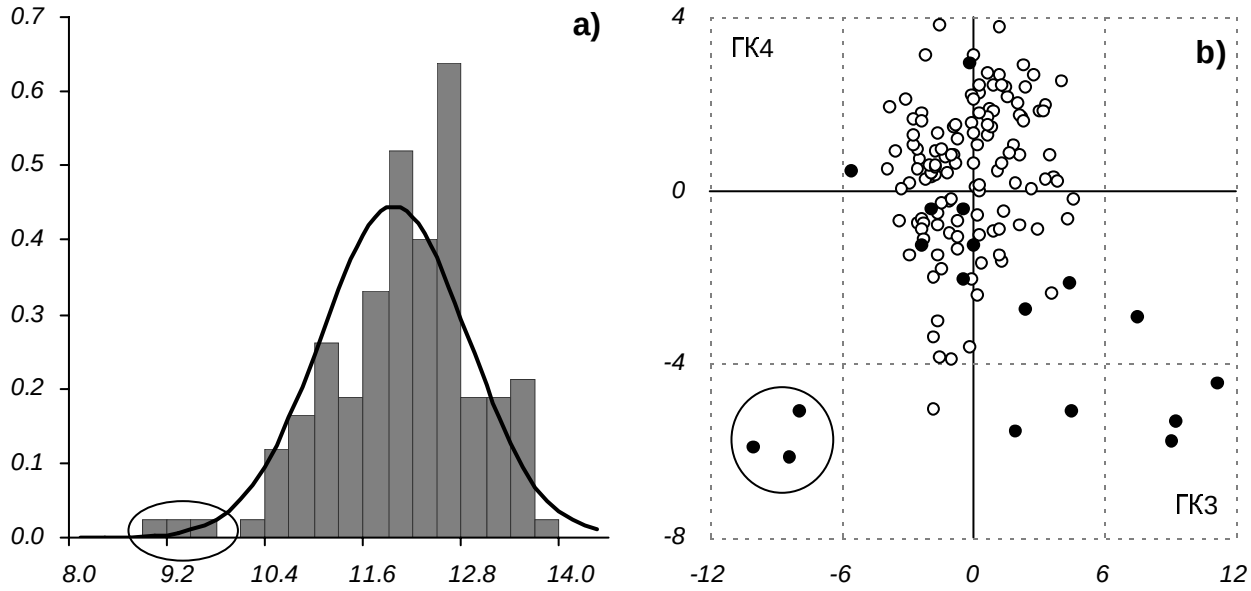


Рис. 1 ММГ содержания воды в зерне по БИК спектрам на исходном наборе из 141 образца. Гистограмма распределения содержания воды в зерне (a) и график ПЛС счетов ГК3/ГК4 (b). Закрашенные точки отмечают «подозрительные образцы».

Рассмотрим типичный пример, подтверждающий эту точку зрения. Классической задачей ММГ является определение характеристик качества зерна по БИК спектрам [40]. В рассматриваемом примере измерения проводились на приборе InfraLUM FT-10 в диапазоне 8000-14000 $см^{-1}$, а прогнозируемым аналитическим показателем является содержание воды 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 [41] второго порядка с окном в 3 точки; (4) нормализацию каждого спектра вдоль спектральных линий,

(5) центрирование и нормировка всех спектров по образцам. Данные ***y*** также усреднялись по трем повторным измерениям, а затем центрировались и нормировались. Модель многомерной градуировки строилась по 141 образцу в спектральной области 9000 – 11000 $см^{-1}$ методом проекций на латентные структуры ПЛС (см. раздел 3.2 Методы). Для построения ММГ достаточно четырех главных ПЛС компонент, которые объясняют, соответственно, 99% и 90% дисперсии **X** и ***y***.

На Рис. 1 представлены гистограмма распределения значений содержания воды в зерне (a) и график ПЛС счетов в координатах ГКЗ/ГК4 (b). Применяя к этим данным обычный статистический анализ, можно заметить, что они не противоречат гипотезе о нормальном распределении откликов. Даже три экстремальных образца, отмеченные на графиках, выглядят «допустимыми» – вероятности их появления равны: 0.03, 0.21, и 0.38. Тем не менее, действуя согласно обычной процедуре ММГ, мы удалили все образцы, отмеченные на Рис. 1b закрашенными точками, как выпадающие, и провели новую градуировку модели. Результаты обработки цензурированных данных (124 образца) показаны на Рис. 2.

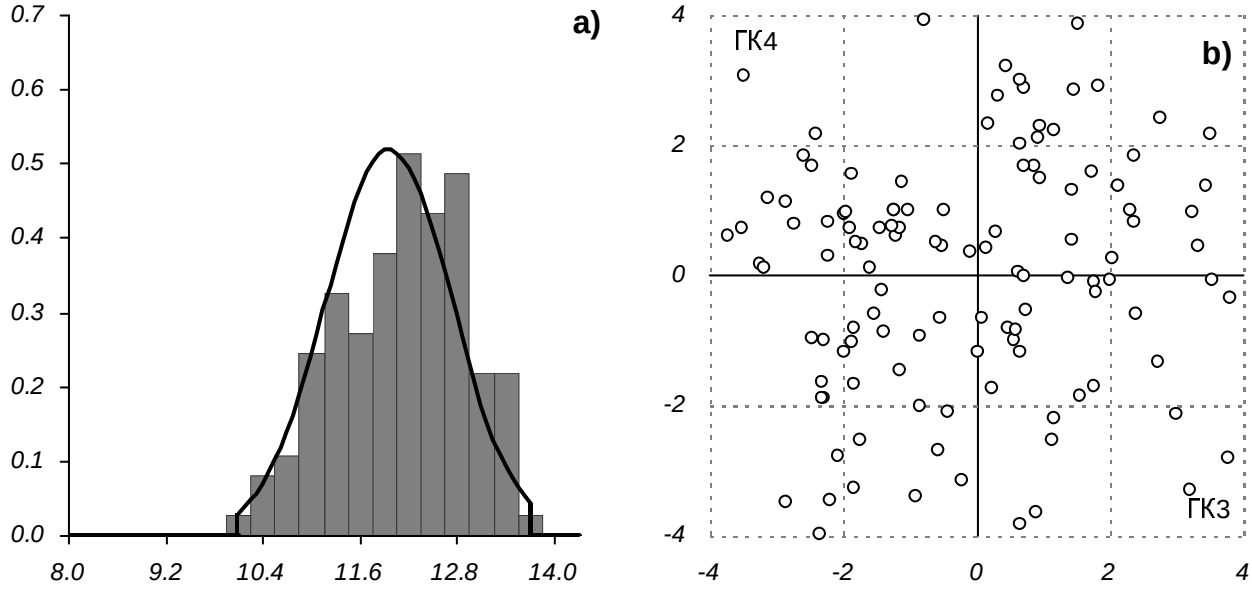


Рис. 2 ММГ содержания воды в зерне по БИК спектрам на цензурированном наборе из 124 образцов. Гистограмма распределения содержания воды в зерне (a) и график ПЛС счетов ГКЗ/ГК4 (b).

Теперь ПЛС модель объясняет, соответственно, 99% и 92% дисперсии X и y. На графике счетов (b) уже не видно никаких подозрительных образцов. 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 ошибки, которые подчиняются не нормальному, а, скорее, усеченному нормальному распределению. В следующем разделе мы увидим, какие выводы следуют из факта ограниченности ошибок, рассматриваемого далее как постулат.

2. Объяснение ПИО метода. Одномерный пример

2. 1. Модельный пример

Объясним, как работает метод ПИО, используя простейшую одномерную регрессию

$$y = xa + \varepsilon \qquad (1)$$

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

Основным предположением метода ПИО является постулат об ограниченности ошибки измерения ε., который можно сформулировать следующим образом. Никакая ошибка ε не может превосходить по абсолютной величине некоторую константу β, т.е. –

$$\mathrm{Prob}(|\varepsilon| > \beta) = 0 \qquad (2)$$

Исследуем простейшие выводы, немедленно вытекающие из этого постулата. В Табл. 1 (столбцы 1 и 2) и на Рис. 3 приведены модельные данные, построенные нами для регрессии (1) при a=1. Ошибка измерения в отклике y моделировалась с использованием равномерного распределения шириной 1.4, т.е., в этом примере, β=0.7

Табл. 1 Модельные данные и результаты их обработки

Образцы	x	y	$\hat{y}$	$\hat{y}^-$	$\hat{y}^+$	$a^{\min}$	$a^{\max}$	v^-	v^+	h	r	$\lvert r\rvert+h$
0	1	2	3	4	5	6	7	8	9	10	11	12
C1	1.0	1.28	1.04	0.86	1.23	0.58	1.98	0.92	1.19	0.19	0.31	0.51
C2	2.0	1.68	2.09	1.72	2.46	0.49	**1.19**	1.85	2.38	0.38	-0.62	1.00
C3	4.0	4.25	4.18	3.43	4.92	0.89	1.24	3.70	4.76	0.76	0.03	0.79
C4	5.0	5.32	5.22	4.29	6.15	**0.92**	1.20	4.62	5.95	0.95	0.05	1.00
T1	3.0	3.35	3.13	2.58	3.69	0.88	1.35	2.77	3.57	0.57	0.26	0.83
T2	4.5	6.19	4.70	3.86	5.53	1.22	1.53	4.16	5.36	0.86	2.05	2.91
T3	5.5	5.40	5.74	4.72	6.76	0.85	1.11	5.08	6.55	1.05	-0.60	1.64

В этом примере использован очень короткий набор данных (образцов сравнения), который разбит на две части. Первые четыре образца, обозначенные С1-С4, являются обучающим набором, используемым для построения модели. На Рис. 3 они изображены открытыми кружками. Последние три образца, обозначенные как Т1-Т3, являются проверочными образцами, для которых строится прогноз. Они показаны на Рис. 3 закрашенными квадратами. Не смотря на примитивность этого примера, с его помощью можно объяснить все основные свойства метода ПИО.

2.2 МНК градуировка

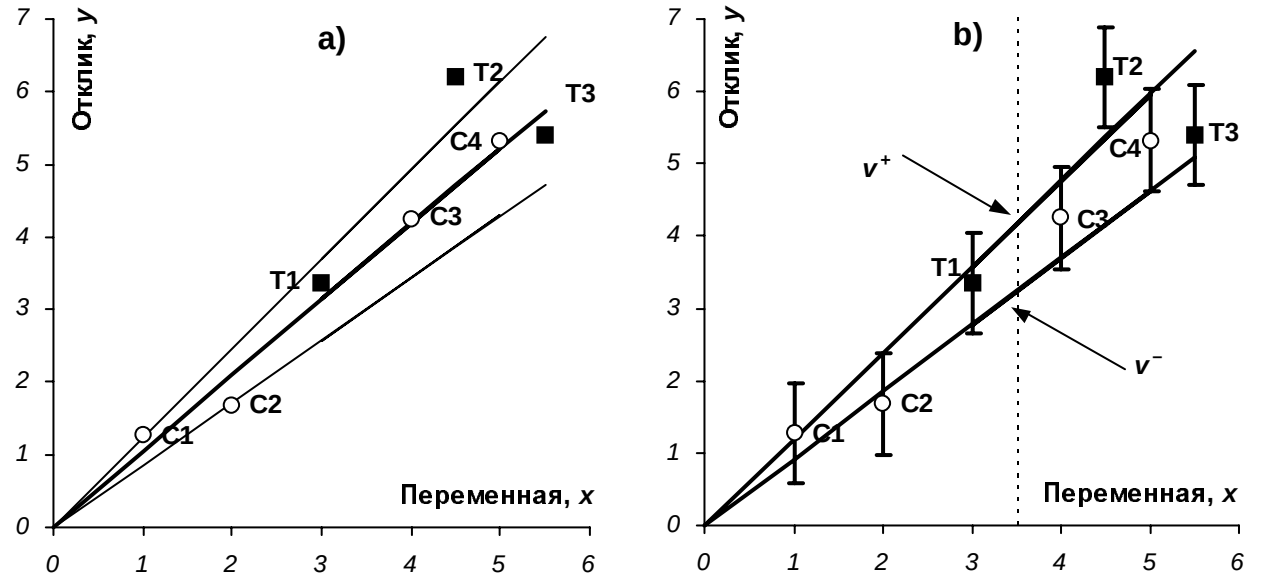
Начнем, однако, с традиционного метода наименьших квадратов (МНК) [5]. Используя обучающие данные (x_i,. y_i), i=1 – 4 (столбцы 1 и 2 в Табл. 1, образцы С1-С4), можно найти МНК оценку параметра a

$$\hat{a} = \frac{\bar{y}}{\bar{x}} = 1.003, \quad \text{где } \bar{x} = \frac{1}{4}\sum_{1}^{4} x_i, \quad \bar{y} = \frac{1}{4}\sum_{1}^{4} y_i, \tag{3}$$

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

$$\hat{y} = \hat{a}x, \tag{4}$$

(столбец 3 в Табл. 1 и Рис. 3а; жирная линия).



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

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

Рис. 3 Одномерный модельный пример. O- обучающие образцы, ■- проверочные образцы

Мы можем оценить дисперсию ошибки ε, применив хорошо известную формулу [3]

$$s^2 = \frac{1}{3}\sum_1^4 (y_i - \hat{y}_i)^2 = 0.078\,, \qquad (5)$$

а также построить доверительные интервалы для отклика

$$\hat{y}^{\pm} = \hat{y} \pm s\frac{x}{2\bar{x}}\mathrm{t}_3(P)\,. \qquad (6)$$

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, Табл.1 и на Рис. 3а, тонкие линии.

2. 3. ПИО градуировка

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 не известна. Позже мы рассмотрим, как можно разрешить эту проблему.

Из уравнения регрессии (1) и принципа ограниченности ошибки (2) следует, что для каждого образца (x_i,. y_i), i=1 – 4 из обучающего набора выполняется условие

$$\left| y_i - ax_i \right| \le \beta \qquad (7)$$

или, в эквивалентной форме –

$$a_i^{\min} \le a \le a_i^{\max}\,, \qquad (8)$$

где

$$a_i^{\min} = \frac{y_i - \beta}{x_i} \qquad a_i^{\max} = \frac{y_i + \beta}{x_i}\,. \qquad (9)$$

Значения (9) приведены в 6-ом и 7-ом столбцах Табл. 1. Неравенства (8) должны выполняться одновременно для всех обучающих образцов, т.е. для i=1, 2, 3, 4. Очевидно, что так может быть только для тех значений параметра a, которые лежат в интервале

$$a^{\min} \le a \le a^{\max}\,, \qquad (10)$$

где

$$a^{\min} = \max_{1\le i\le 4} a_i^{\min}\,, \qquad a^{\max} = \min_{1\le i\le 4} a_i^{\max}\,. \qquad (11)$$

Эти значения выделены жирным шрифтом в соответствующих столбцах Табл. 1.

Интервал (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 (10), то соответствующая величина отклика $y=ax$ в произвольной точке x ограничена значениями

$$v^- \le y \le v^+, \tag{12}$$

где

$$v^- = a^{\min} x, \qquad v^+ = a^{\max} x. \tag{13}$$

Эти величины приведены в столбцах 8 и 9 Табл. 1.

Таким образом, построена интервальная оценка параметра a (10), которая является аналогом точечной оценки $\hat{a}$, получаемой с помощью МНК. Кроме того, найдены и прогнозные интервалы (13) для отклика y, справедливые, как для обучающих, так и для любых других (новых) образцов.

Рассмотрим графическую интерпретацию метода ПИО. На Рис. 3b приведены те же данные, что и на Рис. 3a, но теперь каждая точка сопровождается интервалом ошибок (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 ошибок для всех четырех обучающих образцов. Из графика видно, что нижним пределом является линия, проходящая через нижнюю точку интервала ошибок образца C4. Верхним пределом является прямая, проходящая через верхнюю точку интервала ошибок образца C2. Все линии, заключенные между этими границами, очевидно, будут удовлетворять поставленным условиям (7) и, наоборот, любая прямая, лежащая вне этого угла, будет противоречить этим условиям. Границы показаны на Рис. 3b двумя жирными линиями v^+ и v^-.

Отметим очевидный факт, что построение градуировки методом ПИО в нашем примере «держится» только на двух образцах: C2 и C4. Именно они задают границы (10) возможных значений параметра a, так, что мы вправе назвать эти образцы *граничными*. Прочие обучающие образцы C1 и C3 несущественны; мы можем удалить их из обучающего набора, и результат останется прежним. Это очень важное свойство метода ПИО, которое находит применение в задаче выбора представительного набора образцов [22].

Итак, мы показали, что все образцы из обучающего набора в методе ПИО разделяются на две группы: наиболее важные, граничные образцы, на которых держится модель, и несущественные, внутренние образцы, которые можно удалить из обучающего набора и модель при этом не изменится.

2. 4. Статус образцов

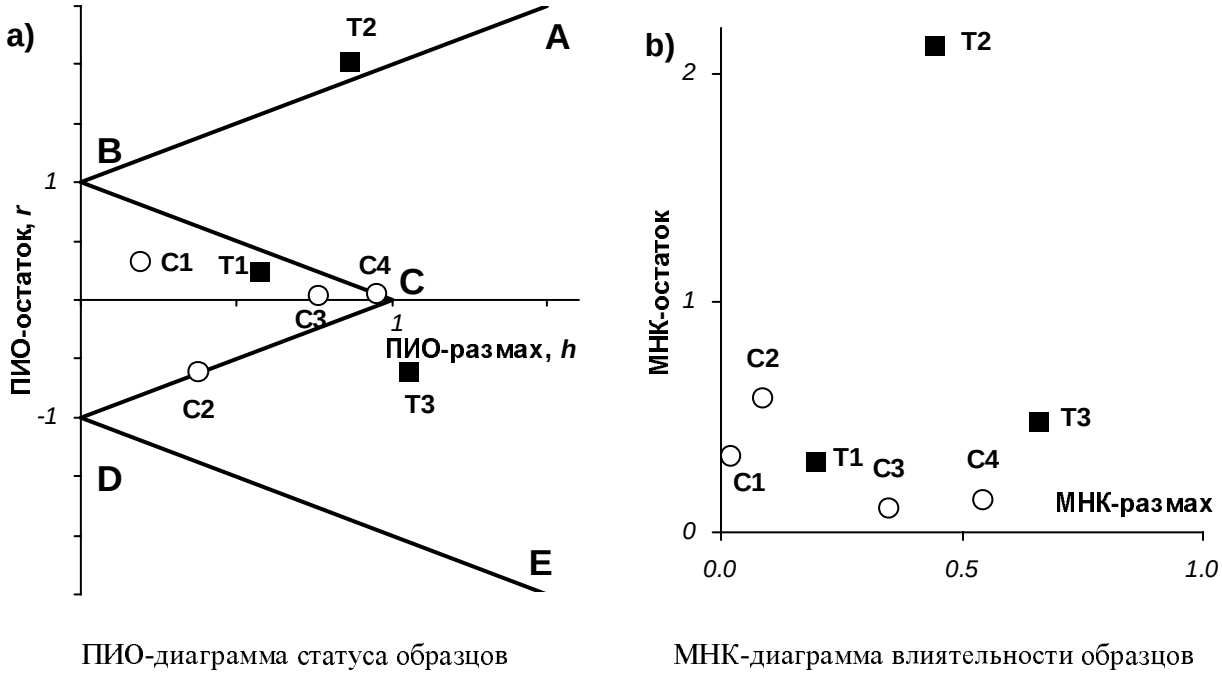
Поразмышляем теперь о том, что может произойти с ПИО-моделью, точнее с ее ОДЗ, при добавлении в обучающий набор нового образца. Очевидно, что ОДЗ может только уменьшится, но не увеличится. Например, при добавлении образца Т3, верхний предел (линия v^+) пройдет уже ниже прежнего, так чтобы «зацепить» верхнюю границу интервала ошибок для Т3; при этом a^{max} станет равной 1.11, вместо 1.16 (См. Табл. 1). Это свойство метода ПИО, называемое *состоятельностью* (см. (А8) в Приложении), важно в теоретическом плане. Оно показывает, что при увеличении числа образцов в обучающем наборе, неопределенность ПИО-оценок уменьшается. При этом, если величина предела ошибки выбрана правильно, или она, по крайней мере, не меньше, чем β, то истинное значение параметра a всегда будет находиться внутри области допустимых значений (10). Это свойство, называемое *несмещенностью* (см. (А5) в Приложении), также важно для понимания и обоснования метода ПИО.

Однако не всякий новый образец, включенный в обучающий набор, приводит к уточнению модели. Например, образец Т1 не изменит ПИО-модель. Это видно из Рис. 3b, где прогнозный интервал полностью лежит внутри соответствующего интервала ошибок Т1, а также из 6-го и 7-го столбцов Табл. 1. Другой случай – это образец Т2. Его интервал ошибок не пересекается с прогнозным интервалом, поэтому при добавлении Т2 в обучающий набор произойдет разрушение модели, т.к. система неравенств (7) станет несовместной. Это также хорошо видно из Табл. 1 – минимум по столбцу 7 (1.11) станет меньше, чем максимум по столбцу 6 (1.22). Таким образом, можно классифицировать новые образцы по трем группам по отношению к тому, как они повлияют на модель, если их добавить в обучающий набор, т.е. определить статус (влиятельность) каждого образца. Во-первых, можно выделить класс *внутренних* образцов, которые не меняют модель, и класс *внешних* образцов, таких, которые изменяют модель. Кроме того, из внешних образцов можно выделить еще группу *выбросов*, т.е. таких образцов, которые нельзя добавлять в обучающий набор (при данном значении β), т.к. они разрушают модель.

2.5. ПИО-классификация статуса образцов

Исследование статуса образцов на графике X-Y неудобно, а в многомерном случае просто невозможно. Для того, что провести такой анализ в общем случае, вводятся две переменные, отражающие свойства образца: ПИО-остаток r, и ПИО-размах h (уравнения (А11) и (А12) в Приложение). Получив прогнозные интервалы (12) легко вычислить и эти

величины. Для нашего примера значения r и h приведены в столбцах 10 и 11 Табл. 1 и изображены на *диаграмме статуса образцов* (ДСО) на Рис. 4а в координатах (h, r).



ПИО-диаграмма статуса образцов МНК-диаграмма влиятельности образцов

Рис. 4. Определение статуса образцов в одномерном модельном примере.
○- обучающие образцы, ■- проверочные образцы

На этой диаграмме обучающие образцы показаны открытыми кругами, а проверочные образцы – закрашенными квадратами, так же, как и на Рис. 3. Жирная ломанная ABCDE ограничивает области различных статусов образцов. Форма этой ломаной определяется двумя фундаментальными неравенствами ((A13) и (A15) в Приложении), связывающими величины h и r. Из диаграммы видно, что все образцы из обучающего набора лежат внутри треугольника BCD (их статус – внутренние, для них $|r|+h \leq 1$, см. столбец 12, Табл.1), причем образцы C2 и C4 расположены на его границах (их статус – граничные, для них $|r|+h = 1$). Проверочный образец T1 также попал в треугольник внутренних образцов, т.к. $|r|+h = 0.83<1$. Образец T2 лежит ниже прямой DE, что свидетельствует о том, что он является выбросом. Для него $|r|-h = 2.91 > 1$ (см. неравенство (A15) в Приложении). Образец T3 – это внешний образец, причем из диаграммы видно, что ни при каком значении своего остатка r, он не попадет в область внутренних образцов. 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 градуировки. Такие образцы называются *абсолютно-внешними*.

Итак, показано, что метод ПИО позволяет ввести новую классификацию всех объектов ММГ – как из обучающего набора, так и проверочных и новых образцов. Эта классификация называется *классификацией статуса образцов* (КСО). Она базируется на определениях (A11) и (A12), а также на утверждениях (A13)-(A16), приведенных в Приложении. Практическое применение КСО сводится к вычислению значений r и h, построению соответствующей ДСО и исследованию расположения образцов на этом графике.

Можно заметить, что треугольная форма области внутренних образцов на ДСО (Рис. 4a) напоминает обычную диаграмму влиятельности [7], показанную на Рис. 4b. На этой диаграмме все образцы модельного примера изображены в координатах: МНК-размах

$$h_{\text{МНК}} = \boldsymbol{x}^{\text{t}}\left(\mathbf{X}^{\text{t}}\mathbf{X}\right)^{-1}\boldsymbol{x} = x_{\text{i}}^2 \Big/ \sum_{i=1}^{4} x_{\text{i}}^2$$

против нормированного МНК-остатка.

$$r_{\text{МНК}} = \frac{1}{\beta}(y - \hat{y})$$

Сходство между диаграммами статуса и влиятельности вытекает из хорошо известного статистического соотношения [8], которое связывает *погрешность* (среднеквадратичную ошибку градуировки, RMSEC), *воспроизводимость* (стандартную ошибку, SEC) и *смещение* (систематическую погрешность, BIAS)

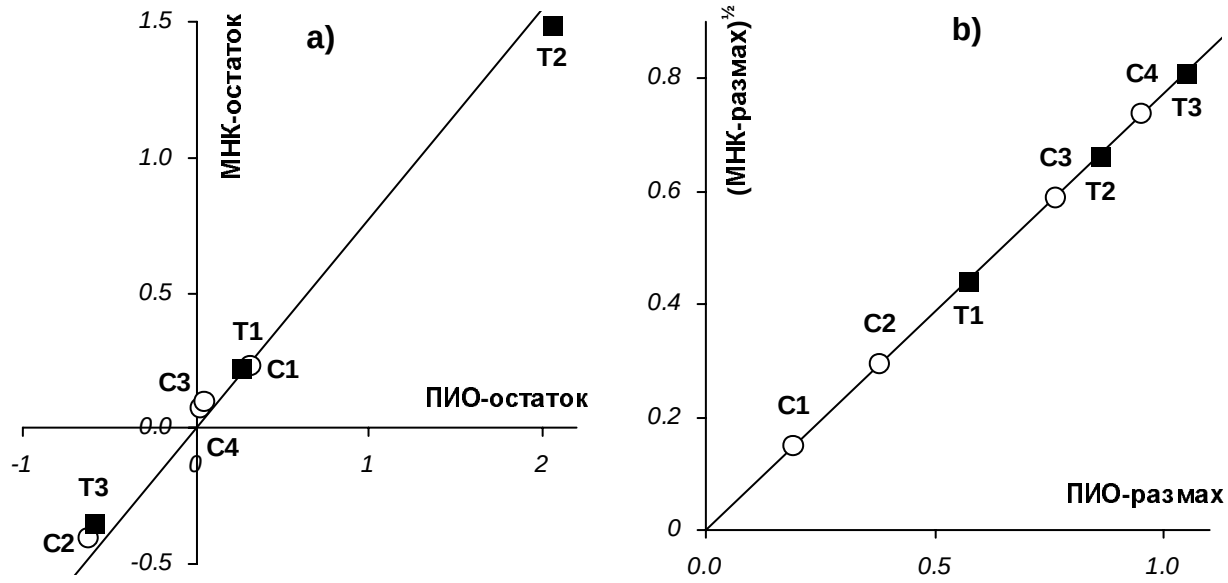
$$\text{RMSEC}^2 \approx \text{SEC}^2 + \text{BIAS}^2. \qquad (14)$$

В методе ПИО, в котором величина β является погрешностью, ПИО-размах h характеризует (нормализованную) воспроизводимость, а ПИО-остаток r отвечает за (нормализованное) смещение, уравнение (14) может быть представлено в виде

$$\beta^2 = \beta^2 h^2(\mathbf{x}) + \beta^2 r^2(\mathbf{x},y), \qquad (15)$$

который, действительно, совпадает с уравнением (A13) в Приложении. С другой стороны, мы должны признать существенное отличие между уравнениями (14) и (15). Оно состоит в том, что последнее уравнение справедливо для каждого образца, тогда как уравнение (14) имеет смысл только для всей совокупности образцов в данных, т.е. в среднем.

В заключение этого раздела приведем два графика (Рис. 5 a и b), свидетельствующие о тесной связи между характеристиками образцов в МНК и ПИО методах.



Сравнение МНК и ПИО остатков

Сравнение МНК и ПИО размахов

Рис. 5. Характеристики образцов в одномерном модельном примере.
○- обучающие образцы, ■- проверочные образцы

Из Рис. 5a видна сильная корреляция (R^2=0.999) между МНК и ПИО остатками. Связь между размахами более сложная (Рис. 5b), 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 (12), тогда как в методе наименьших квадратов размах пропорционален дисперсии прогноза [42], 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 рассмотрена в работах [21, 43].

2.6. Как оценить β

Построение оценки b параметра β – максимальной погрешности – это довольно сложная статистическая процедура, которая кратко описана в Приложении и более подробно в работе [32]. Для понимания сути дела важно лишь то, что, в зависимости от характера распределения ошибки, построенная оценка b оказывается, как правило, в пределах 2σ-4σ, где σ – это дисперсия распределения. Очевидно, что для любого усеченного распределения величина β не может быть меньше 2σ. Крайний случай – это

равномерное распределение, в котором $\beta = 1.71\sigma$. Затем, для обычного числа образцов в эксперименте (скажем, менее 1000) мы не можем ожидать появление экстремальных значений за пределом 3σ. Наконец, предел 4σ дает нам гарантию, что и новые образцы никогда не перейдут эту границу. В рассмотренном примере $\beta = 2.5\sigma$, что несколько больше чем должно быть при равномерном распределении, которое мы использовали при моделировании данных. Такой результат вполне объясним, так как оценка параметра β вычислялась по очень маленькой обучающей выборке, всего четыре образца.

Естественно поставить вопрос о том, насколько такие вариации в оценке погрешности β могут повлиять на результаты КСО. Поскольку в методе ПИО характеристики образцов: ПИО-остаток r и ПИО-размах h, определены как относительные величины, деленные на β (см. (A11) и (A12) в Приложении), завышенная оценка погрешности β не влияет на результаты КСО. Совсем по-другому обстоит дело с прогнозными интервалами в методе ПИО. Они увеличиваются с ростом оценки b, и, при $b=\beta$, эти интервалы, по определению, покрывают истинное значение с вероятностью 1. С другой стороны, в работе [32] было показано, что прогнозные интервалы, построенные для оценки b_{SIC} ((A21), при P=0.90), вместо истинного значения β, имеют вероятность покрытия не меньшую, чем 0.9999. Этот результат подтверждает, что не только предложенная классификация статуса образцов, но и весь метод ПИО, в целом, может использоваться на практике.

Чтобы проиллюстрировать это утверждение, сравним МНК и ПИО оценки в нашем примере. Сопоставляя графики a и b на Рис. 3, можно увидеть, что 95% доверительный коридор для МНК прогноза шире, чем прогнозный коридор метода ПИО. Если рассчитать вероятность попадания в прогнозный ПИО интервал с помощью формулы (6), то она окажется равной 0.91. Заметим, что мы взяли достаточно большое значение $\beta = 2.5\sigma$, которое соответствует нормальной вероятности Prob[-2.5, +2.5]=0.99.

Таким образом, рассмотренный простой модельный пример свидетельствует о двух важных обстоятельствах. Во-первых, использование неограниченного (нормального) распределения для построения доверительных оценок приводит к излишне широким интервалам. Во-вторых, даже при небольшом количестве данных, метод ПИО дает интервалы разумной ширины, которые соответствуют действительности. Чтобы подтвердить последнее утверждение, мы 100000 раз повторили моделирование нашего примера. И ни разу истинное значение $y=x$ не вышло за пределы прогнозных ПИО интервалов.

3. Реальный пример. Многомерная модель

3.1. Данные

Мы использовали хорошо известный дидактический пример предсказания октанового числа бензина [35, 36] для демонстрации того, что метод ПИО может с успехом применяться к реальным задачам ММГ, в том числе и в условиях мультиколлинеарности. В этом примере матрица **X** состоит из БИК спектров поглощения, полученных для 226 длин волн в области 1100-1550 нм. Они показаны на Рис. 6. Компоненты вектора ***y*** являются результатами соответствующих лабораторных измерений октановых чисел [47]. Данные разделены на два набора, в каждом из которых октановые числа меняются в диапазоне 87-93. Первый, обучающий набор состоит из 24 образцов промышленных бензинов (№№1-24), и он используется для построения ММГ. Во второй набор включены 13 образцов, которые служат для проверки прогнозных свойств модели. Важно, что этот проверочный набор содержит четыре образца (№№10-13), которые представляют этилированные бензины, отсутствующие в обучающем наборе. Мы будем называть проверочные наборы без этих образцов (№№1-9), и с ними (№№1-13), соответственно, коротким и длинным проверочным набором.

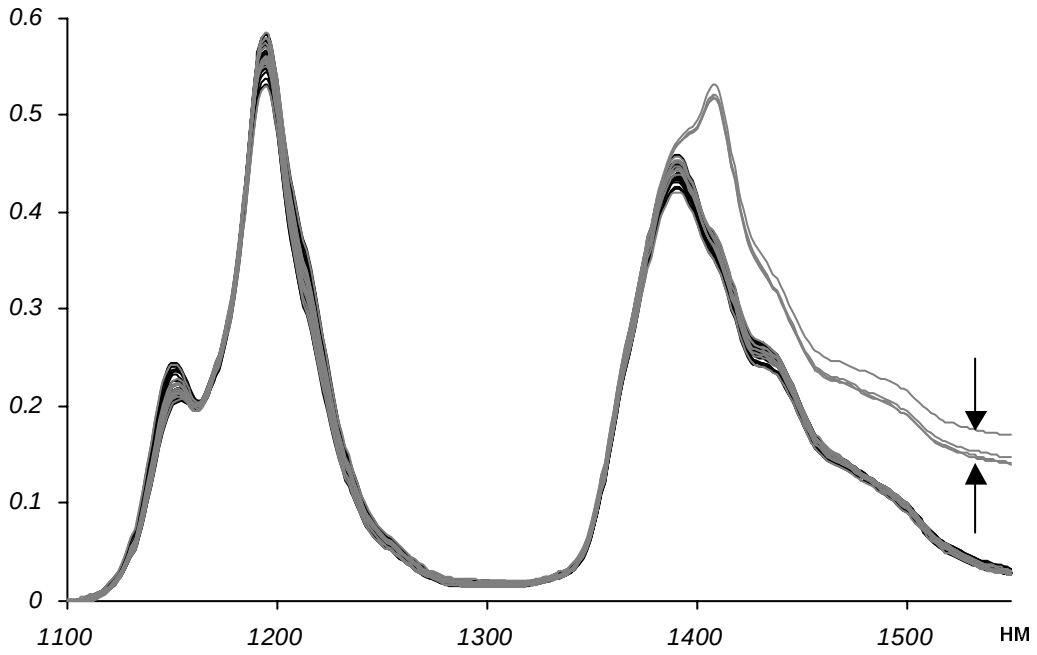


Рис. 6. Спектры поглощения образцов бензинов. Этилированные образцы показаны стрелками.

3.2. Методы

При построении ММГ основная математическая трудность состоит в обращении матрицы $\mathbf{X}^t\mathbf{X}$, которая в нашем случае имеет размерность 226×226. Если бы эта матрица была невырождена (имела бы полный ранг [5]), то для градуировки методом наименьших квадратов можно было бы обратить эту матрицу и найти оценки неизвестных параметров модели $\hat{\boldsymbol{a}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{y}$. Однако, в нашем примере, как и в большинстве практических задач, эта матрица вырождена. Поэтому, в соответствии со свойством (А6) в Приложении, применение метода ПИО невозможно. Для преодоления этой трудности используются различные методы регуляризации данных, например, метод главных компонент, ридж-регрессия, и т.д. Мы использовали ПЛС – метод *проекций на латентные структуры* [35]. Суть метода состоит в одновременной декомпозиции матрицы $\mathbf{X}$ и вектора $\boldsymbol{y}$ в виде

$$\mathbf{X}=\mathbf{T}\mathbf{P}^t+\mathbf{E} \qquad \boldsymbol{y}=\mathbf{T}\boldsymbol{q}+\boldsymbol{f}, \qquad (16)$$

где $\mathbf{T}$ называется матрицей *счетов*, $\mathbf{P}$ и $\boldsymbol{q}$, соответственно, – матрицей и вектором *нагрузок*, а $\mathbf{E}$ и $\boldsymbol{f}$, соответственно, – матрицей и вектором *остатков*. При построении этой декомпозиции учитывают три обстоятельства. Во-первых, столбцы $\boldsymbol{t}_i$ в матрице $\mathbf{T}$ являются линейными комбинациями столбцов $\boldsymbol{x}$ матрицы $\mathbf{X}$, т.е. $\boldsymbol{t}_i = \mathbf{X}\boldsymbol{w}_i$. Во-вторых, коэффициенты $\boldsymbol{w}$ выбираются так, чтобы максимизировать корреляцию между откликом $\boldsymbol{y}$ и вектором $\boldsymbol{t}_i$. В-третьих, число столбцов в матрице счетов $\mathbf{T}$ и матрице нагрузок $\mathbf{P}$ должно быть равно эффективному рангу матрицы $\mathbf{X}$. Эта величина – k, называется числом *главных компонент* (ГК) и она, естественно, меньше, чем число столбцов в матрице $\mathbf{X}$. Важным преимуществом метода ПЛС является возможность визуальной интерпретации данных на графиках счетов, что мы уже использовали в разделе 1 (см. Рис. 1 и Рис. 2). При прогнозировании нового (проверочного) образца $\boldsymbol{x}$, он проецируется в вектор счетов $\boldsymbol{t}$, для которого затем применяется регрессионная модель (16).

Традиционно задача ММГ задается в однородном виде $\boldsymbol{y}=\mathbf{X}\boldsymbol{a}$, так что $\boldsymbol{y}=\mathbf{0}$, при $\boldsymbol{x}=\mathbf{0}$. Для того, чтобы согласовать исходные данные ($\mathbf{X}_{\text{raw}}$, $\boldsymbol{y}_{\text{raw}}$) с этой моделью, они *центрируются*

$$\boldsymbol{y} = \boldsymbol{y}_{\text{raw}} - m_0\mathbf{1}, \qquad \mathbf{X} = \mathbf{X}_{\text{raw}} - \left(m_1\mathbf{1}, m_2\mathbf{1}, \ldots, m_p\mathbf{1}\right)$$

Здесь m_0 – это среднее значение вектора откликов $\boldsymbol{y}$, а m_i – это средние значения, вычисленные для всех столбцов матрицы $\mathbf{X}_{\text{raw}}$. Кроме того, часто бывает необходимо провести и *нормирование* данных. Это делается для того, чтобы усреднить вклады от различных переменных. Если данные не нормировать, то результат может зависеть от некоторых переменных, которые могут иметь большую дисперсию, но малую

регрессионную значимость. Нормирование данных ($\mathbf{X}_{raw}$, $\boldsymbol{y}_{raw}$) – это умножение на диагональные матрицы $\mathbf{X} = \mathbf{X}_{raw}\,\mathbf{S}_X$ и $\boldsymbol{y} = \boldsymbol{y}_{raw}\,\mathbf{S}_y$. Диагональные элементы матриц $\mathbf{S}$ обычно выбирают равными обратным стандартным отклонениям s_{ii}, вычисленными для соответствующих столбцов $\mathbf{X}_{raw}$ и $\boldsymbol{y}_{raw}$, т.е.. $\mathbf{S}_{ii}=(s_{ii})^{-1}$.

Ознакомится с методом ПЛС можно по многочисленным, например [7, 8], но труднодоступным в России монографиям. За последнее время появилось, наконец, верное, но краткое изложение этого метода в учебнике [1]. Подробное изложение на русском языке ПЛС, также как и других методов многомерного анализа данных, содержится в книге [35].

Используя метод ПЛС в нашем примере, можно спроецировать исходную задачу ММГ на двумерное подпространство, где новая задача является уже невырожденной.

$$\boldsymbol{y} = m_0\mathbf{1} + \mathbf{T}\boldsymbol{a} + \boldsymbol{\varepsilon}.$$

Здесь m_0 – это среднее значение $\boldsymbol{y}$, а $\mathbf{T}$ – это матрица счетов размером $n\times 2$. При таком количестве главных компонент (k=2) объясняется 97% дисперсии $\mathbf{X}$ и 98% дисперсии $\boldsymbol{y}$.

3.3. Градуировка

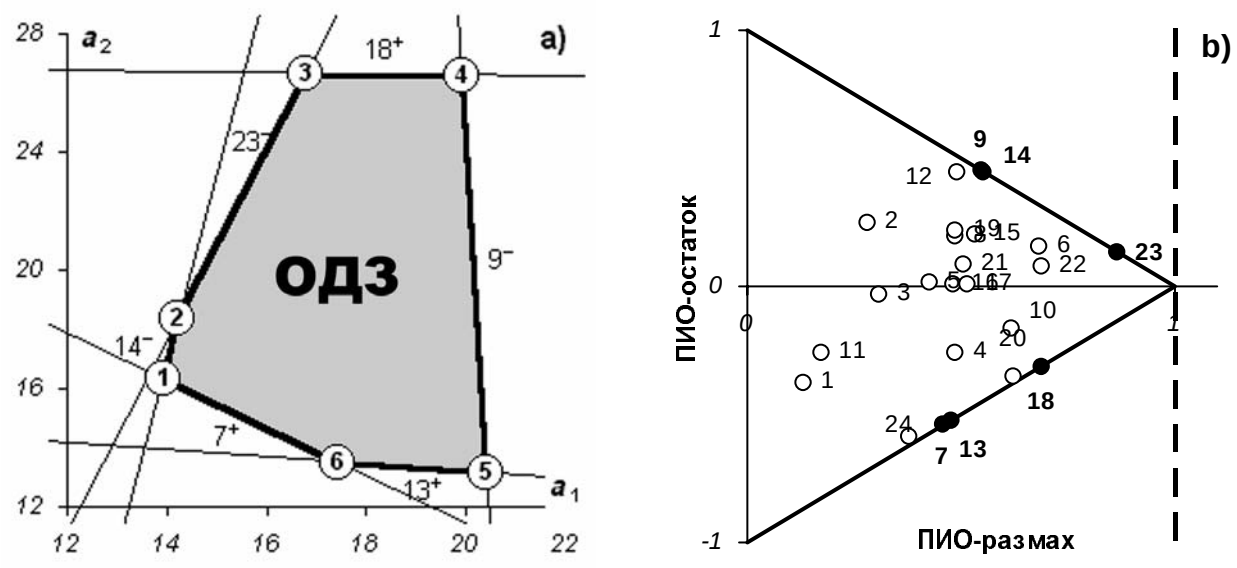
Для использования метода ПИО нужно определить значение максимальной погрешности β, так, как это описано в Приложении. При этом необходимо принять в расчет то, что проекционные методы обязательно увеличивают общую погрешность за счет ошибок моделирования. Это происходят из-за того, что билинейные (ПЛС, РГК) модели являются только аппроксимациями сложных систем. Поэтому максимальная погрешность β всегда больше, чем отдельно взятая ошибка измерения отклика.

Применяя формулу (А19) из Приложения, получаем b_{min}=0.484. Это означает, что для рассматриваемых данных нельзя применять метод ПИО с b < 0.484, т.к. ОДЗ будет пустой. Величина b_{min} дает нижнюю границу оценки максимальной погрешности β, но нам необходима и соответствующая оценка сверху. Используя уравнение (А21) из Приложения при P=0.90 , мы получаем оценку b_{SIC}=0.880. Это значение b используется далее во всех вычислениях в качестве оценки максимальной погрешности β. Оценив среднеквадратичную ошибку градуировки (RMSEC)

$$s_{cal} = \sqrt{\frac{1}{n_{cal}}(\boldsymbol{y}_{cal} - \hat{\boldsymbol{y}}_{cal})^2} = 0.268,$$

можно сравнить точность моделирования методами ПЛС и ПИО: b_{min}/s_{cal}=1.81 и b_{SIC}/s_{cal} =3.28.

Для применения метода ПИО нет необходимости явно конструировать ОДЗ, тем более что для размерности большей, чем 2 – это очень сложная задача [20]. Однако, в рассматриваемом примере p=k=2, поэтому для большей иллюстративности, а также для того, чтобы объяснить технику метода ПИО, мы показывает ОДЗ на Рис. 7a.



Область допустимых значений:
— граничные линии, ○- вершины

Диаграмма статуса обучающих образцов
○- внутренние, ● - граничные образцы

Рис. 7. Прогноз октанового числа. ПЛС модель с 2 главными компонентами. Обучающий набор

Также как и в одномерном примере, ОДЗ образована не всеми 24 образцами из обучающего набора, а только граничными образцами. В нашем случае имеется шесть таких граничных объектов (№№. 7, 9, 13, 14, 18, 23). Они обозначены закрашенными точками на диаграмме статуса обучающих образцов, показанной на Рис. 7b. Все граничные образцы лежат на границе треугольника, т.е. для них $|r|+h$ =1. Именно эти объекты образуют ОДЗ, как это показано на Рис. 7a, где каждая линия соответствует либо уравнению $\boldsymbol{t}_i^{\mathrm{t}}\boldsymbol{a} = y_i - m_0 + b$ (обозначены знаком “+”), либо уравнению $\boldsymbol{t}_i^{\mathrm{t}}\boldsymbol{a} = y_i - m_0 - b$ (обозначены знаком “–“). Числа около линий соответствуют номерам граничных образцов, а ОДЗ обведена жирной ломаной линией.

Итак, результатами ПИО градуировки являются: оценка b максимальной погрешности β, и набор граничных образцов, образующих ОДЗ.

3.4. Прогноз

Рассчитаем теперь методом ПИО прогнозные интервалы. При этом для каждого проверочного образца нужно вычислить величины v^- и v^+ (уравнение (A10) в

Приложении), которые определяют границы индивидуальных прогнозных интервалов. Эта оптимизационная задача решается с помощью метода *линейного программирования*, который позволяет получить решение в общем случае, без явного представления ОДЗ.

Задача линейного программирования [44] заключается в максимизации/ минимизации линейной функции при наличии линейных ограничений. В общем случае эту задачу можно представить в так называемом каноническом виде

$$\min_{\boldsymbol{a}} \left\{ \boldsymbol{c}^{\mathrm{t}}\boldsymbol{a}, \text{ при условиях } \mathbf{T}\boldsymbol{a} = \boldsymbol{d} \text{ и } \boldsymbol{a} \geq 0 \right\}$$

где $\boldsymbol{a} \in \mathrm{R}^{\mathrm{p}}$ – это вектор неизвестных параметров, $\boldsymbol{c} \in \mathrm{R}^{\mathrm{p}}$ – коэффициенты целевой функции. Матрицу $\mathbf{T} \in \mathrm{R}^{\mathrm{n} \times \mathrm{p}}$ называют матрицей условий, а вектор $\boldsymbol{d} \in \mathrm{R}^{\mathrm{n}}$ – вектором ограничений. Любая система линейных неравенств может быть приведена к каноническому виду с помощью введения дополнительных переменных. Избыточные переменные добавляют к системе, чтобы исключить ограничения типа «меньше, чем», а остаточные переменные добавляют, чтобы исключить условия «больше чем». Кроме того, любая задача максимизации может быть преобразована в задачу минимизации изменением знаков коэффициентов целевой функции [46].

Каноническую задачу линейного программирования можно решить симплекс методом, который представляет собой последовательный перебор угловых точек, при котором значение целевой функции $\boldsymbol{c}^{\mathrm{t}}\boldsymbol{a}$ убывает от итерации к итерации. В результате находится решение – точка $\boldsymbol{a}$, которая одновременно является допустимой (удовлетворяющей всем ограничениям) и оптимальной (дающей минимальное значение). Симплекс метод [44-46] является хорошо известным алгоритмом, который включен во многие пакеты программного обеспечения, например, [47]. Он не требует явного конструирования многогранника, а вычисляет допустимые вершины алгебраически, используя соответствующие системы линейных уравнений.

В нашем примере, ОДЗ, образованная линейными ограничениями – это многоугольник, который имеет шесть вершин, пронумерованных для наглядности (см. Рис. 7а). Для того чтобы проиллюстрировать суть симплекс-метода, найдем прогнозный интервал для первого образца из проверочного набора. Мы будем обозначать его индексом test. Используя обычный ПЛС алгоритм, можно найти проекцию этого 226-мерного вектора $\boldsymbol{x}_{\text{test}}$ на плоскость главных компонент, т.е. вычислить вектор счетов $\boldsymbol{t}_{\text{test}} =$ (– 0.0689; 0.0343). Для определения границ прогнозного интервала $[v^{-}, v^{+}]$, необходимо решить две задачи линейного программирования

$$v^- = \min_{\boldsymbol{a}} \boldsymbol{t}_{\text{test}}^{\text{t}}\boldsymbol{a} \ , \qquad v^+ = \max_{\boldsymbol{a}} \boldsymbol{t}_{\text{test}}^{\text{t}}\boldsymbol{a}$$

где вектор параметров $\boldsymbol{a}$ удовлетворяет ограничениям

$$y_i - m_0 - \beta \le \boldsymbol{t}_i^{\text{t}}\boldsymbol{a} \le y_i - m_0 + \beta, \quad i = 1,2,..,24$$

т.е. $\boldsymbol{a}$ лежит внутри ОДЗ, показанной на Рис. 7b. Решив эти задачи, можно получить интервальный прогноз для искомого отклика $y_{\text{test}} = m_0 + \boldsymbol{t}_{\text{test}}^{\text{t}}\boldsymbol{a}$,

$$v^- + m_0 \le y_{\text{test}} \le v^+ + m_0 ,$$

Table 2 Построение прогнозного интервала методом ПИО

Вершина	a_1	a_2	$\boldsymbol{t}_{\text{test}}^{\text{t}}\boldsymbol{a}$	y_{test}
1	13.91	16.36	-0.398	88.85
2	14.22	18.36	-0.351	88.90
3	**16.79**	**26.66**	**-0.244**	**89.01**
4	19.91	26.61	-0.461	88.79
5	**20.41**	**13.16**	**-0.956**	**88.30**
6	17.43	13.51	-0.739	88.52

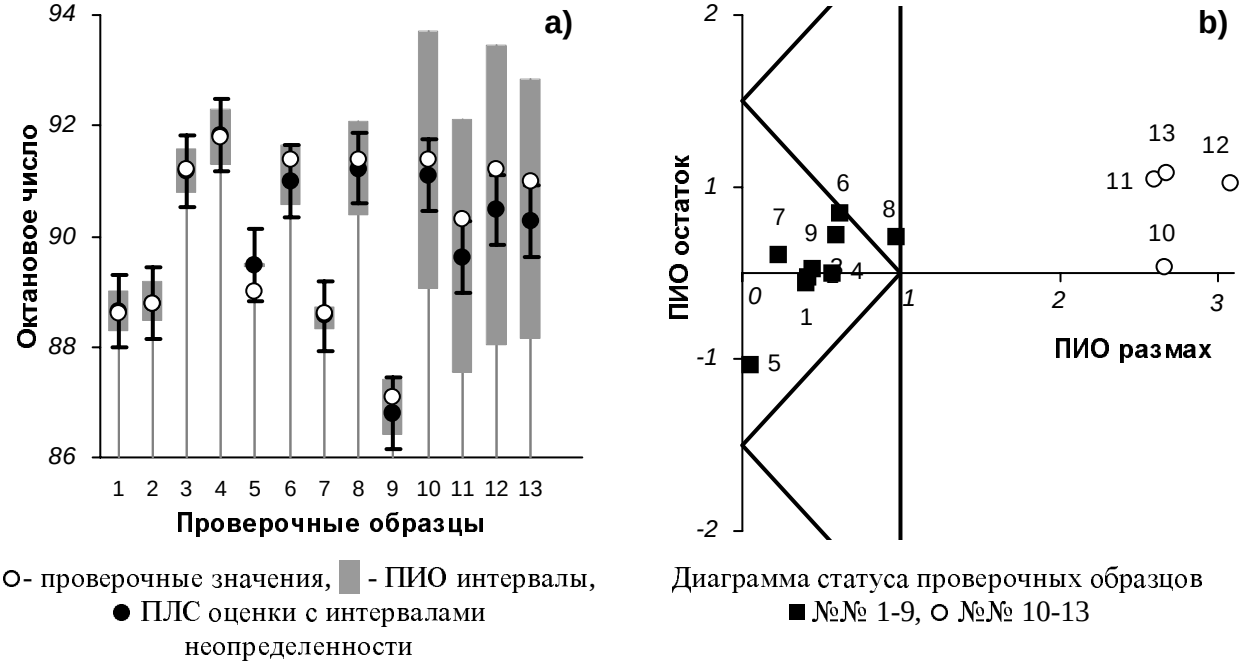
В Табл. 2 представлены координаты всех шести вершин ОДЗ (a_1 и a_2) и соответствующие значения откликов. Из этой таблицы можно определить прогнозный интервал $88.30<y_{\text{test}}<89.01$, соответствующий вершинам 5 (минимум) и 3 (максимум). На самом деле, в сложной задаче нет необходимости действовать так неэффективно и проверять каждую вершину в отдельности – достаточно применить стандартный симплекс-алгоритм. Первая допустимая вершина, определяемая симплекс алгоритмом – это вершина 1. Для нахождения минимума (v^-), алгоритм проходит следующий путь 1→ 6 → 5, а для определения максимума (v^+) – путь 1 → 2 → 3 (см. Рис. 7a).

3.5. Результаты

Сравним прогнозные интервалы для проверочных образцов, вычисленные методом ПИО, с оценками $\hat{\boldsymbol{y}}_{\text{test}}$, полученными методом ПЛС. Для этого мы будем использовать традиционную оценку точности ПЛС-прогноза – среднеквадратичную ошибку предсказания (RMSEP) –

$$s_{\text{test}} = \sqrt{\frac{1}{n_{\text{test}}}\left(\boldsymbol{y}_{\text{test}} - \hat{\boldsymbol{y}}_{\text{test}}\right)^2} \ .$$

Величина s_{test}, определенная методом перекрестной самопроверки с исключением по одному образцу (LOO) [7], равна 0.322. Если новые образцы принадлежат к тому же типу (качественно и количественно), что и обучающие образцы, то можно ожидать для них примерно такую же ошибку предсказания. В рассматриваемом примере дело обстоит именно так для короткого проверочного набора. Соответствующие интервалы неопределенности $[\hat{y}_{test} \pm 2s_{test}]$ показаны на Рис. 8a черными отрезками, а ПИО интервалы – серыми прямоугольниками. Разумеется, применять такой подход к выпадающим проверочным образцам (№№ 10-13) было бы неверно.



○- проверочные значения, ▮ - ПИО интервалы, ● ПЛС оценки с интервалами неопределенности

Диаграмма статуса проверочных образцов ■ №№ 1-9, ○ №№ 10-13

Рис. 8. Прогноз октанового числа. ПЛС модель с 2 главными компонентами. Проверочные данные.

ПИО интервалы для последних четырех выпадающих образцов очень велики. В обычном проекционном подходе они трактуются как выбросы [35]. Эти же образцы легко могут быть идентифицированы и на диаграмме статуса образцов (Рис. 8b). При использовании ДСО, образцы №№ 10-13 определяются как абсолютно-внешние, т.е. как такие образцы, которые совершенно не похожи на образцы из обучающего набора. Исследуя график на Рис. 8a, можно заметить, что проверочные значения (открытые точки), так же как и значения ПЛС-прогноза (закрашенные точки) лежат внутри ПИО-интервалов (серые прямоугольники), и что интервалы неопределенности, определенные в методе ПЛС (черные интервалы), согласуются с ПИО-интервалами для короткого проверочного набора образцов (№№1-9). В тоже время, размер ПИО-интервала

индивидуален для каждого нового образца и поэтому он более информативен по сравнению со средним значением неопределенности, вычисленным в методе ПЛС. Что касается выпадающих образцов (№№ 10-13), то ПИО-интервалы немедленно сигнализируют об их аномальности. Можно также отметить, что для «нормальных» образцов ПИО-интервалы меньше, чем соответствующие ПЛС-интервалы неопределенности.

Этот пример показывает, как метод ПИО отвечает на вопросы, которые всегда являются важными для аналитика.

1. Оценка максимальной погрешности β определяет точность градуировки и задает границу воспроизводимости для всех образцов, которые подобны образцам из обучающего набора.

2. Прогнозные интервалы, полученные методом ПИО, устанавливают индивидуальную неопределенность прогноза отклика для каждого нового образца.

3. Позиция каждого образца на диаграмме статуса позволяет определить, подобен ли этот объект образцам из обучающего набора, и тем самым, задает разумные границы применимости построенной градуировки.

Применительно к рассмотренному примеру градуировки октанового числа по БИК-спектрам, было бы естественно установить, что в повседневной практике эта методика могла бы применяться только к тем новым образцам, которые являются *внутренними* по отношению к построенной ММГ. Только в этом случае можно гарантировать, что точность прогноза будет не хуже, чем точность градуировки, которая примерно равна точности традиционного лабораторного метода определения октанового числа. С другой стороны, если бы эта методика предназначалась для проведения научно-исследовательских работ, то было бы достаточно предупредить, что она не может применяться для *абсолютно-внешних* образцов, поскольку они сильно отличаются в структуре своих предикторов от стандартных обучающих образцов.

4. Выводы

Мы полагаем, что представленный метод простого интервального оценивания может быть полезен для практического использования в задачах анализа многомерных данных. Метод ПИО имеет некоторые преимущества по сравнению с традиционным регрессионным подходом.

Во-первых, он не использует никаких исходных предположений о виде ошибки, кроме ее ограниченности. Тем самым его можно считать методом, свободным от вида распределения.

Во-вторых, он дает результат в удобном интервальном виде, учитывающем неопределенность прогноза отклика.

В-третьих, он позволяет естественно очертить рамки, в которых может использоваться построенная модель. Это достигается с помощью классификации статуса образцов, различающей: надежные «внутренние образцы», существенные «граничные образцы», подозрительные «внешние образцы», выпадающие «абсолютно-внешние образцы» и разрушительные «выбросы».

Главное (и единственное) предположение об ограниченности ошибки, по нашему мнению, является не недостатком, а преимуществом метода, т.к., с практической точки зрения, оно выглядит более обоснованным, чем традиционное допущение о нормальности ошибок.

Вычислительная процедура метода основана на известных алгоритмах линейного программирования и легко может быть реализована. Применение метода ПИО к реальным задачам дало результаты, которые хорошо согласуются с практическим опытом.

5. Программное обеспечение

Для ПЛС моделирования применялась программа The Unscrambler [53]. Простое интервальное оценивание (ПИО) производилось с помощью программного обеспечения, выполненного как надстройка (Add-In) для пакета Excel. В нем использовался алгоритм NIPALS [35] для билинейного моделирования, стандартный SIMPLEX алгоритм [44] для оптимизации, а также весь необходимый набор специальных процедур для предварительной обработки данных, преобразований, и т.п. Эта программа в настоящее время находится в стадии бета-тестирования.

6. Использованные сокращения

(Б)ИК	(Ближний) инфракрасный (спектр)
ДСО	Диаграмма статуса образцов
КСО	Классификация статуса образцов
МНК	Метод наименьших квадратов

ММГ	Многомерная градуировка
ОДЗ	Область допустимых значений
ПЛС	Проекция на латентные структуры (метод)
ПИО	Простое интервальное оценивание (метод)
РГК	Регрессия на главные компоненты
RMSEC	Среднеквадратичная ошибка градуировки
RMSEP	Среднеквадратичная ошибка предсказания

A. Приложение. Строгое описание метода ПИО

A.1 Область допустимых значений

Рассмотрим модель линейной многомерой градуировки

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{a}+\boldsymbol{\varepsilon}, \qquad \text{(A1)}$$

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

$$\text{Prob}\{|\varepsilon| > \beta\} = 0, \text{ и что для любых } 0<b<\beta \ \text{Prob}\{|\varepsilon| > b\} > 0 \qquad \text{(A2)}$$

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

Назовем пару $(\boldsymbol{x}_i, y_i)$, $i=1, \ldots, n$, обучающим образцом. Здесь вектор $\boldsymbol{x}_i^t$ – это i-ая строка матрицы $\mathbf{X}$, соответствующая отклику y_i в уравнении (A1). Согласно условию (A2), для каждого $i=1,\ldots,n$ справедливы неравенства

$$y_i^- \le \boldsymbol{x}_i^t\boldsymbol{a} \le y_i^+, \qquad y_i^- = y_i - \beta, \qquad y_i^+ = y_i + \beta \qquad \text{(A3)}$$

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 (A3) для данного образца i, образуют полосу $S(\boldsymbol{x}_i, y_i)$ в пространстве параметров R^p. Положение и ширина этой полосы определяются значениями $(\boldsymbol{x}_i, y_i)$. Рассмотрим все образцы из обучающего набора и соответствующие им полосы. Очевидно, что вектор параметров $\boldsymbol{a}$ удовлетворяет всем неравенствам (A3) одновременно тогда и только тогда, когда он принадлежит всем полосам.

Область *допустимых значений* (ОДЗ) A для параметров $\boldsymbol{a}$ системы (A1) – это множество в пространстве параметров, образованное пересечением всех полос

$$A = \bigcap_{i=1}^{n} S(\boldsymbol{x}_i, y_i) \qquad \text{(A4)}$$

Область A – это замкнутый выпуклый многогранник [49, 50], образованный границами пересекающихся полос. A – это случайное множество, поскольку оно построено с использованием случайных величин $\boldsymbol{y}$.

A.2 Свойства ОДЗ

ОДЗ A обладает следующими свойствами для любой модели заданной уравнением (A1).

Область A является *несмещенной* оценкой параметра $\boldsymbol{\alpha}$. Непосредственно из определения ОДЗ следует, что истинное значение $\boldsymbol{\alpha}$ всегда принадлежит A:

$$\mathrm{Prob}\{\boldsymbol{\alpha} \in A\} = 1 \qquad \text{(A5)}$$

Область A *ограничена* тогда и только тогда [49, 50], когда матрица $\mathbf{X}$ имеет полный ранг

$$\mathrm{rank}\mathbf{X} = p. \qquad \text{(A6)}$$

Из этого следует, что если система (A1) мультиколлинеарна, то до использования ПИО метода необходимо применить какую-либо процедуру регуляризации. Например, можно использовать стандартный подход [7, 32] и спроецировать исходные данные на подпространство меньшей размерности

$$\boldsymbol{y} = \mathbf{T}\mathbf{P}^{\mathrm{t}}\boldsymbol{a} + \boldsymbol{f} = \mathbf{T}\boldsymbol{q} + \boldsymbol{f}, \qquad \text{(A7)}$$

где матрица счетов $\mathbf{T}$ имеет полный ранг $k<p$ и затем применить метод ПИО к этой системе

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

$$\mathrm{Prob}\{A \cap \boldsymbol{\alpha}\} = 1 \quad \text{при} \quad n \to \infty \qquad \text{(A8)}$$

при тех же «слабых» условиях [51]:

$$\lambda_{\mathrm{p}} \to \infty \quad \text{при} \quad n \to \infty,$$

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

Область A образована не всеми обучающими образцами, а только некоторыми, называемыми *граничными*. Поэтому, из обучающего набора можно исключить все образцы, кроме граничных, и ОДЗ при этом не изменится.

A.3 Предсказание отклика

Рассмотрим задачу предсказания отклика y для некоторого нового вектора $\boldsymbol{x}$ по модели (A1). Если параметр $\boldsymbol{a}$ меняется внутри ОДЗ A, то, очевидно, что предсказываемое значение $\boldsymbol{y}=\boldsymbol{x}^{\mathrm{t}}\boldsymbol{a}$ принадлежит интервалу

$$V = [v^-, v^+] \tag{A9}$$

где

$$v^- = \min_{\boldsymbol{a} \in A} (\boldsymbol{x}^{\mathrm{t}} \boldsymbol{a}), \qquad v^+ = \max_{\boldsymbol{a} \in A} (\boldsymbol{x}^{\mathrm{t}} \boldsymbol{a}). \tag{A10}$$

Интервал V является результатом прогноза методом ПИО. Для его вычисления не нужно строить область A в явном виде, т.к. решения задач (A10) могут быть найдены с помощью стандартных методов линейного программирования [44, 45].

A.4. Классификация статуса образцов

Для численной характеристики качества прогноза методом ПИО вводятся следующие величины.

Величина –

$$r(\boldsymbol{x}, y) = \frac{1}{\beta} \left(y - \frac{v^+(\boldsymbol{x}) + v^-(\boldsymbol{x})}{2} \right) \tag{A11}$$

называется *ПИО-остатком*. Величина r представляет разницу между центром прогнозного интервала и значением y (нормированную на β), поэтому r характеризует *смещение*.

Величина –

$$h(\boldsymbol{x}) = \frac{1}{\beta} \left(\frac{v^+(\boldsymbol{x}) - v^-(\boldsymbol{x})}{2} \right). \tag{A12}$$

называется *ПИО-размахом*. Величина h вычисляется как полуширина прогнозного интервала, деленная на максимальную ошибку, поэтому h характеризует β-нормализованную *воспроизводимость*.

Очевидно, что при добавлении некоторого нового образца ($\boldsymbol{x}$, y) в обучающий набор, с ОДЗ A может произойти одно из следующих событий: 1) ОДЗ не меняется, т.е. $A_{n+1}=A_n$; 2) ОДЗ уменьшается, т.е. $A_{n+1} \subset A_n$; 3) ОДЗ исчезает, т.е. $A_{n+1}=\varnothing$. Здесь A_n обозначает ОДЗ, которая была построена с помощью обучающего набора, состоящего из n образцов. Первый случай соответствует образцу, который называется *внутренним*. Такие образцы полностью согласуются с моделью, поэтому им можно полностью доверять при прогнозе. Второй случай означает, что образец располагается вне имеющейся модели, поэтому он может быть назван *внешним* образцом. Внешние образцы не противоречат модели и, будучи добавлены в обучающий набор, улучшают точность моделирования. Однако пока такие образцы не включены в обучающий набор, прогноз на них ненадежен.

Это может быть обусловлено двумя причинами. Во-первых, ширина прогнозного интервала, т.е. ПИО-размах, может быть больше, чем ошибка градуировки, или имеется систематическая ошибка, что характеризуется ПИО-остатком. Наконец, третье событие, происходит тогда, когда новый образец полностью противоречит построенной модели. Такие образцы, очевидно, являются *выбросами* во всех смыслах этого слова и их нельзя использовать для предсказания.

Было показано [22, 32], что такая классификация легко может быть проведена без явного конструирования ОДЗ. Вместо этого, она проводится на базе следующих утверждений, связывающих величины r и h.

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

$$\left| r(\boldsymbol{x}, y) \right| \le 1 - h(\boldsymbol{x}) \qquad \text{(A13)}$$

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

$$|r(\boldsymbol{x}_i, y_i)| = 1 - h(\boldsymbol{x}_i) \qquad \text{(A14)}$$

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

$$\left| r(\boldsymbol{x}, y) \right| > 1 + h(\boldsymbol{x}) . \qquad \text{(A15)}$$

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

$$h(\boldsymbol{x}) > 1 \qquad \text{(A16)}$$

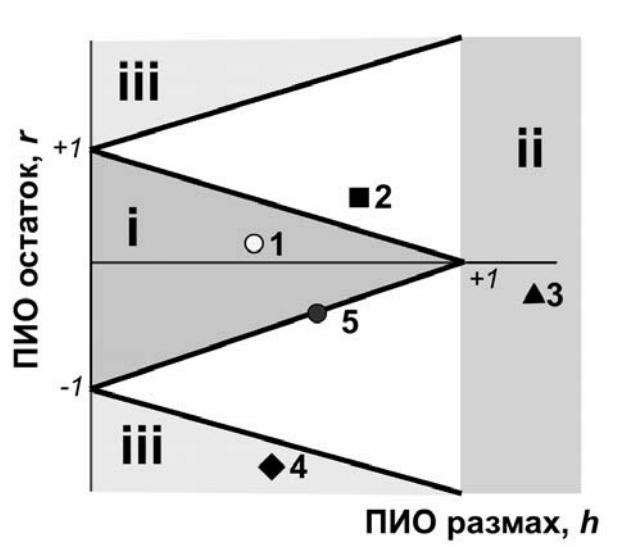


Рис. 9. Диаграмма статуса образцов.
i – внутренние образцы (○), ii- абсолютно-внешние образцы (▲), iii- выбросы (◆)

Используя определения (A11)-(A12) и утверждения (A13)-(A16), можно построить диаграмму статуса образцов (ДСО), прототип которой показан на Рис. 9. При любой размерности исходных данных ($\mathbf{X}$, $\boldsymbol{y}$) и для любого числа параметров, ДСО является двумерной диаграммой, и это делает ее очень мощным инструментом в ММГ. Утверждения (A13)-(A16) делят плоскость «ПИО-остаток (r)» – «ПИО-размах» на три области, каждая из которых соответствует одной из трех категорий образцов: внутренние (область i на Рис. 9), внешние (вне области i) и выбросы (область iii).

Обычно, когда модель ММГ применяется к новым образцам, соответствующие значения y неизвестны. Поэтому нельзя вычислить ПИО-остаток, r (А11), но всегда можно определить ПИО-размах, h (А12). Легко видеть, что если размах нового образца больше, чем единица (h>1, область ii), то этот образец никогда не может быть отнесен к типу внутренних, ни при каком значении y. Подобные образцы образуют специальный класс объектов, которые называются *абсолютно-внешним* (утверждение (А16)).

А.5. 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 –

$$b_1 > b_2 \quad \Rightarrow \quad A(b_1) \supset A(b_2)\,. \qquad \text{(А17)}$$

Можно показать, что в случае, когда имеется последовательность оценок $b_1>b_2>\ldots \geq\beta$, сходящаяся к β, то свойства (А5)-(А8) будут выполняться и для $A(b_n)$. Кроме того, очевидно, что

$$A(0) = \varnothing, \quad A(\infty) \neq \varnothing \qquad \text{(А18)}$$

Из (А17)-(А18) следует, что существует *минимальное* значение b, при котором $A(b) \neq \varnothing$. Это значение может быть принято в качестве оценки величины β

$$b_{\min} = \min\{b, \quad A(b) \neq \varnothing\} \qquad \text{(А19)}$$

Оценка (А19) является состоятельной, но смещенной, т.к. $b_{\min}\leq\beta$. Она задает нижний предел всех возможных значений β. Это, несомненно, полезная характеристика обучающего набора и модели, но помимо $b_{\min}$ необходимо оценить и верхний предел максимальной погрешности.

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

1. Число образцов в обучающем наборе. Чем больше образцов, тем ближе b к β.

2. Тяжесть хвостов распределения ошибок. Чем легче хвосты, тем хуже эта оценка.

Применяя традиционный статистический подход [52], можно построить такую оценку b, что Prob{$b>\beta$}>P и, при этом, оценка b максимально близка к β. Рассмотрим $\hat{\boldsymbol{y}}$ – некоторую точечную (регрессионную) оценку вектора $\boldsymbol{y}$, остатки $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, и статистику

$$b_{\text{reg}}=\max(\,|e_1|,\ldots,|e_n|\,). \qquad \text{(А20)}$$

Статистическое моделирование, проведенное для различного числа образцов в обучающем наборе с использованием различных ограниченных распределений ошибки, показывает, что оценка

$$b_{SIC}=b_{reg}\, C(n, s^2, P) \qquad (A21)$$

является искомым верхним пределом β с вероятностью P. Эмпирическая функция C [32] зависит от n – числа образцов в обучающем наборе, и от s^2 – дисперсии остатков, которая характеризует тяжесть хвостов распределения ошибки.

Литература

1. 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, 2004г (Analytical Chemistry. The Approved Text to FECS Curriculum Analytical Chemistry, Wiley-VCH, Weinheim)
2. Марьянов Б.М. // Избранные главы хемометрики, Томск: Из-во Том. ун-та, 2004
3. Бард И.// Нелинейное оценивание параметров. М.: Статистика, 1979. (Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974)
4. Pirson K. // *Phil.Mag*., 1901, **2** (6), 559-572
5. Демиденко Е. З. //*Линейная и нелинейная регрессии*, М , Финансы и статистика, 1981
6. Тихонов А.Н.// *Докл. АН СССР*, 1963, т. **4**, 1035
7. Martens H., Naes T. //*Multivariate Calibration*, Wiley: New York, 1998.
8. Næs T., T. Isaksson T., Fearn T., Davies T. // *Multivariate Calibration and Classification*, NIR Publications, 2002
9. Faber K. //Comparison of two recently proposed expressions for partial least squares regression prediction error, *Chemom. Intell. Lab. Syst.* 2000; **52:** 123-134.
10. Pomerantsev A.L. //*Chemom. Intell.Lab.Syst*., 1999; **49**: 41
11. Канторович Л.В. // *Сиб. мат. журн*., 1962, **3** (5):701-709
12. Вощинин А.П., Бочков А.Ф., Сотиров Г.Р.// *Завод. лаб*., 1990, 56(7):76-95
13. Анисимов В.М., Померанцев А.Л., Новорадовский А.Г., Карпухин О.Н. // *Журн. прикл. спектрос*, 1987, **46**: 117-122
14. Ахунов И.Р., Ахмадишин З. Ш., Спивак С.И. // Химическая физика, 1982, 12: 1660-1665
15. Бахитова Р.Х, Спивак С.И. // Химия и химическая технология, 1999, 42(3), 92-96
16. Слинько М.Г. , Спивак С.И., Тимошенко В.И. // Кинетика и катализ 1972, 13 (6), 1570-1577

17. Спивак С.И., Тимошенко В.И, Слинько М.Г. // ДАН, 1970, 192 (3), 580-582

18. Белов В.М., Карбаинов Ю.А., Суханов В.А. и др.// Журнал аналитической химии, 1994, 49, (4), 370

19. Хлебников А.И. // Журнал аналитической химии, 1996, 51(3) 347-348

20. Белов В.М., Суханов В.А., Унгер Ф.Г.//*Теоретические и прикладные аспекты метода центра неопределенности*. Новосибирск: Наука, 1995

21. Померанцев А.Л., Родионова О.Е. // О двух подходах к анализу кинетических данных на примере предсказания активности антиоксидантов, представлено в *Кинетика и Катализ*, 2004

22. Rodionova O. Ye., Esbensen K. H., and Pomerantsev A.L. // Application of SIC (Simple Interval Calculation) for object status classification and outlier detection – comparison with PLS/PCR, *J. Chemometrics*, 2004, **18**:402-413

23. Westad F, Martens H. //Variable selection in NIR based on significance testing in Partial Least Squares Regression (PLSR). *J. Near Infrared Spectroscopy* 2000; **8**: 117– 124.

24. Cook R.D. //Detection of influential observations in linear regression. *Technometrics*; 1977, **19:** 15-18.

25. Cook R.D. //Influential observations in linear regression. *JASA*; 1979, **74**: 169-174.

26. Andrews D.F., Pregibon D. // Finding the outliers that matter. *J. Royal Statis. Soc.*; 1978, **B-40:** 84-93

27. Draper N.R., John J.A.// Influential observations and outlier in regression. *Technometrics*, 1981; **23:** 21-26.

28. Naes T. // The design of calibration in near infra-red reflectance analysis by clustering. *J. Chemometrics*; 1989, **1:** 121-134

29. Hoskuldsson A. //Variable and subset selection in PLS. *Chemometrics Intell. Lab. Syst.* 2001; **55**: 23-38.

30. 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* 1997; **350**: 149-161.

31. Fernandez Pierna J.A., Wahl F., de Noord O.E., Massart D.L. //Methods of outlier detection in prediction, *Chemom. Intell. Lab. Syst*. 2002; **63:** 27-39.

32. Rodionova O. Ye., Pomerantsev A.L. // Principles of Simple Interval Calculations, In: *Progress in Chemometrics Research* (ISBN: 1-59454-257-0), Pomerantsev AL (ed.). Nova Science Publishers: New York, 2005, 43-64

33. Pomerantsev A.L., Rodionova O. Ye.// Multivariate statistical process control and optimisation, *Ibid*, 209-227

34. 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*. Zaikov *et al* (Eds), vol 2., NovaScience Publishers: NY, 2002, 19-29.

35. Esbensen K.H.// *Multivariate Data Analysis – In Practice 4-th Ed.*, CAMO, 2000. [Анализ многомерных данных, сокр. пер. с англ. под ред. О.Родионовой, Из-во ИПХФ РАН, 2005]

36. А. Ланг // Измерение важнейших параметров бензина с помощью анализатора в ближней ИК-области спектра, *Нефтегазовые технологии*, 1994, N 9-10.

37. Clancey V.J. // Statistical Methods in Chemical Analyses, *Nature.* 1947; **159:** 339-340.

38. Rajkó R. //Treatment of model error in calibration by robust and fuzzy procedures; *Anal. Letters*, 1994; **27**: 215-228.

39. Eriksson L., Johansson E., Kettaneh-Wold N., Wold S.// *Multi- and Megavariate Data Analysis*, Umetrics, Umeå, 2001.

40. Sulima EL, Zubkov VA, Rusinov LA. // Specific features of practical implementation of calibration model transfer from a master instrument to slave NIR analyzers for analysis of main characteristics of wheat. In *Progress in Chemometrics Research* (ISBN: 1-59454-257-0), Pomerantsev AL (ed.). Nova Science Publishers: New York, 2005, 196–203.

41. Savitzky A., Golay M.J.E. //*Anal. Chem.*,1964, **36**, 1627

42. Hoskuldsson A.// *Prediction Methods in Science and Technology*, vol.1, Thor Publishing, Copenhagen, Denmark, 1996.

43. Pomerantsev A.L., Rodionova O.Ye. // Hard and soft methods for prediction of antioxidants' activity based on the DSC measurements, принята к печати в *Chemom. Intell. Lab. Syst*. 2005

44. Данциг Дж. // *Линейное программирование, его применение и обобщение*, М. Прогресс, 1966, 520С. [Dantzing G.B. linear Programming and Extensions, Princeton University Press, Princeton, New Jersey, 1963]

45. 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]

46. Карманов В.Г.// *Математическое программирование*, М. Физматлит, Изд.5, 2001, 263С.

47. Linear Programming Packages [On Line] http://www.ici.ro/camo/hlp.htm (1 мая 2005)

48. ГОСТ 8226-82 "Топливо моторное. Исследовательский метод определения октанового числа".
49. Gass S. // *Linear Programming* (4-th ed.) McGow-Hill: New York, 1975.
50. Kuhn H.W., Tucker A.W. // Linear Inequalities and Related Systems. *Ann. Math. Studies* **38**, Princeton University Press: Princeton, N.J., 1956.
51. Eicker F. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Stat*. 1963; **34:** 447-456.
52. Gumbel E. // *Statistics of extremes,* Columbia University Press: N.Y., 1962.
53. The Unscramber [Online] http://www.camo.no (1 января 2005)