МИНОБРНАУКИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Хемометрика и фармацевтический анализ

Учебно-методическое пособие

А.И. Сливкин
П.М. Карлов

Воронеж, 2020

УДК 615(075.32)

Утверждено Научно-методическим советом фармацевтического факультета протокол №

Рецензент: доктор химических наук, профессор, заслуженный деятель науки РФ, зав. Кафедрой аналитической химии ВГУ В.Ф. Селеменев

В пособии приведены краткие теоретические основы многопараметрической калибровки в хемометрическом анализе. Рассмотрено практическое использование хемометрических методов, представленных на примере сравнительного многопараметрического и классического однофакторного способов калибровки в ИК-спектроскопии.

Предназначено для ординаторов, обучающихся по специальности 33.08.03 «фармацевтическая химия и фармакогнозия»

Для специальности - 33.08.03. «Фармацевтическая химия и фармакогнозия»

Введение

С появлением хемометрических методов в аналитической химии, фармацевтическом анализе произошли коренные изменения. Термин «хемометрический» метод далее будет относиться к любому методу многопараметрической калибровки. В отличие от классической калибровки по одному параметру системы (однопараметрической калибровки), которая основана на вариации одного определенного параметра (в случае спектральных данных, на вариации площади или высоты одного пика), в методе многопараметрической калибровки используется весь спектр вещества. Преимущества такого метода очевидны – привлекается больше спектральной информации, что дает возможность идентифицировать и соотносить с составом даже малейшие изменения в спектре образца.

На данный момент существует множество публикаций по данной теме, однако математическая терминология, широко используемая в них, малопонятна для провизора-аналитика. Тем не менее, новые статистические методы оценки являются необходимым инструментом в современном анализе и дают возможность изучать сложные системы, исследование которых было невозможно до относительно недавнего времени.

Одна из целей при создании данного пособия, - это в простых терминах описать многопараметрические методы калибровки, в связи с этим намеренно избегаются серьезные математические описания.

В первой главе представлены теоретические основы метода хемометрической оценки. Во второй главе даны объяснения функциональности метода на примере алгоритма регрессии парциальных наименьших квадратов (**P**artial **L**east **S**quares, PLS) и показываются преимущества многопараметрической калибровки по сравнению с классической однопараметрической. В третьей главе описан хемометрический анализ тестового образца и стандартные действия,

которые следует предпринимать для оценки результатов анализа. Таким образом, в первых трех главах рассматриваются лишь теоретические аспекты многопараметрической калибровки и не даются практические советы по оптимизации PLS – модели.

Практическое использование хемометрических методов и методы оптимизации важных параметров модели описаны в главах 4, 5 и 6. Эти главы предназначены непосредственно для практикующего провизора-аналитика. После изучения этих глав, даже непрофессионал в области хемометрического анализа сможет быстро построить калибровочную модель и обеспечить получение оптимальных результатов, что очень важно для фундаментализации фармацевтического анализа

Три вышеупомянутые главы являются основными в руководстве. Главная задача данного руководства – ознакомить и научить пользователя легко и правильно применять методику многопараметрической калибровки. Для понимания информации необязательно детальное знание теории и статистических методов (т.е. глав 2 и 3).

В главе 7 представлен словарь для быстрого восприятия значений всех новых важных статистических терминов, используемых в специальной литературе. Как уже было упомянуто выше, знание математической терминологии необязательно для того, чтобы успешно производить хемометрическую калибровку. В залючении (глава 8) представлены краткие выводы, перспективы метода, а также производится сравнительный анализ многопараметрического и классического однофакторного методов калибровки.

Данное учебное пособие направлено на по возможности краткое теоретическое обоснование предлагаемого метода и предоставление практических рекомендаций для рутинной лабораторной практики. Оно предназначено для того, чтобы научить пользователя создавать

оптимальные хемометрические модели для любых задач, не пользуясь при этом глубокими познаниями в теории многопараметрической калибровки.

Тем не менее, ошибочно было бы думать, что к представленной методике можно подходить легкомысленно. Качество анализа существенно зависит от того, насколько аккуратно и внимательно аналитик пользуется программным обеспечением для хемометрического анализа.

Метод многопараметрической калибровки в аналитической и фармацевтической химии

В данной и следующей главах приведены теоретические основы метода многопараметрической калибровки. Здесь нет серьезных математических описаний алгоритмов. Тем не менее, методика и способ ее подачи могут быть непривычными для некоторых аналитиков в области фармакопейного анализа и потребуют от них определенных усилий в понимании метода. Основная задача пособия дать именно практическое описание методики, однако некий экскурс в теорию поможет лучше разобраться в методе многопараметрической калибровки.

1. Теоретические аспекты

Для того чтобы описать метод многопараметрической калибровки, прежде всего, необходимо дать представление о практической процедуре развития метода. В общих чертах, любой количественный аналитический метод направлен на то, чтобы определить зависимость параметра системы Y от измеренного параметра X данной системы. Как правило, результат достигается в два этапа: первый – калибровка, второй – анализ (прогноз). В процессе построения калибровки ищется корреляция между измеренным количественным параметром X и свойством системы Y.

Корреляция описывается калибровочной моделью:[1]

$$Y = X * b \qquad (2\text{-}1)$$

где b^{1} - калибровочная функция, часто называемая коэффициентом регрессии, или b- коэффициентом:

$$b = (X^T * X)^{-1} * X^T * Y \qquad (2\text{-}2)$$

Значения параметров X и Y записаны в форме матриц. При необходимости представить данные спектроскопических измерений, интенсивности спектров записываются в строках X-матрицы. Каждый следующий образец, таким образом, будет соответствовать следующей строке в матрице. Соответствующие значения Y будут записаны в соответствующих строках Y-матрицы. T – операция транспонирования ассоциированных матриц.

После проведения калибровки осуществляется анализ. Применив калибровочную модель к измеренному параметру X, можно определить значение Y тестового образца. Описанный процесс представлен на схеме 2.1.

Шаг 1: Калибровка.

Матрица данных X		***Матрица данных Y***		***Калибровочная функция b***

Шаг 2: Анализ

Матрица		***Калибровочна***		***Матрица***

данных X		***я функция b***		***данных Y***

Рис. 2.1 Схема количественного определения веществ

В случае количественного анализа ближнего инфракрасного спектра (БИК), инфракрасного спектра (ИК) или спектра комбинационного рассеяния (КР), измеренными параметрами системы являются, как правило, спектры поглощения или испускания, а определить необходимо значение концентрации исследуемых веществ. Для этого можно воспользоваться либо методом однопараметрической калибровки (с использованием одной переменной), либо методом многопараметрической калибровки.

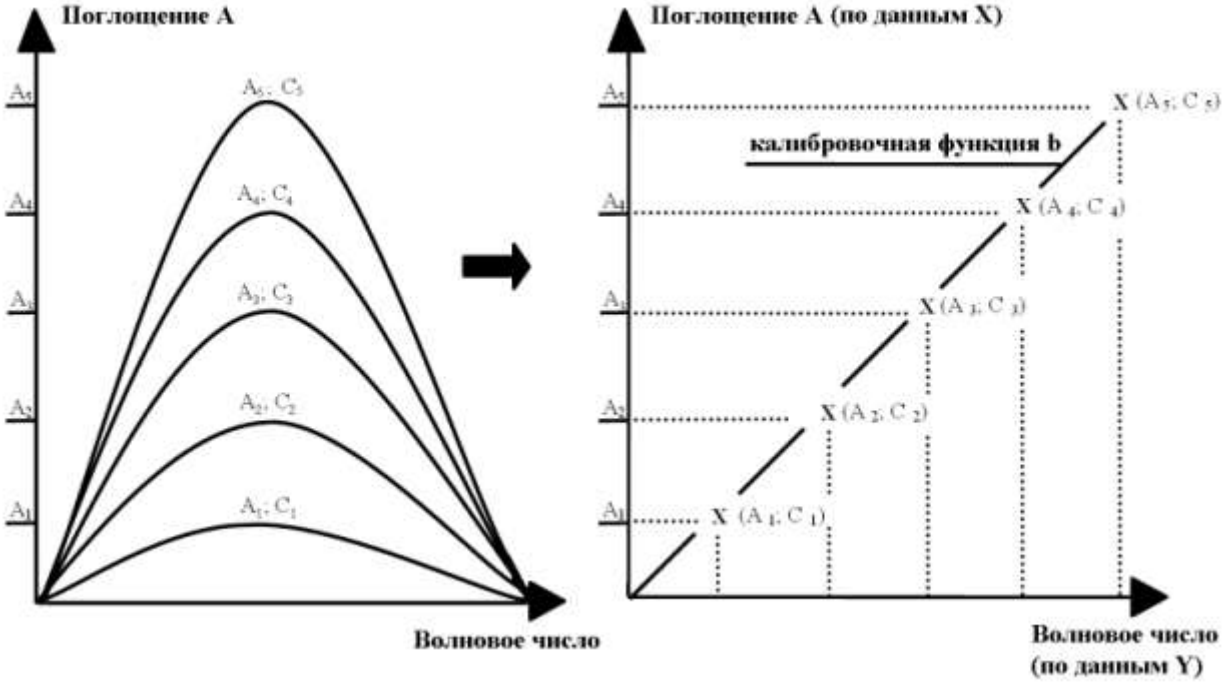


Рис. 2.2 Калибровка спектра поглощения

На рисунках 2.2 и 2.3 показан метод однопараметрической оценки по пику поглощения образца. Были исследованы пять образцов с концентрациями от C_1 до C_5. Эти значения соответствовали значениям величины Y. В результате измерений получены пять величин X, от A_1 до A_5. (см. рис. 2.2. слева).

В случае однопараметрической калибровки значения пиков поглощения расположены на графике напротив значений концентрации вещества (см. рис. 2.2 справа). Калибровочная функция, полученная из данных поглощения/концентрации, позволяет в дальнейшем вычислять значения концентрации на основании измеренных значений поглощения и наоборот.

Анализ нового тестового образца проводится путем спектроскопических измерений и определения величины поглощения A_P в максимуме. Действие калибровочной функции, полученной ранее, на эту величину дает искомые результаты (рис. 2.3). В случае однокомпонентной системы достаточно производить оценку спектров только на одной длине волны. Для двухкомпонентной системы процедура усложняется, так как необходимо произвести измерения для второй длины волны, также характерной для системы. Таким образом, для каждого дополнительного компонента необходимо добавить в калибровочную модель значение величины его поглощения на дополнительной, подходящей длине волны.

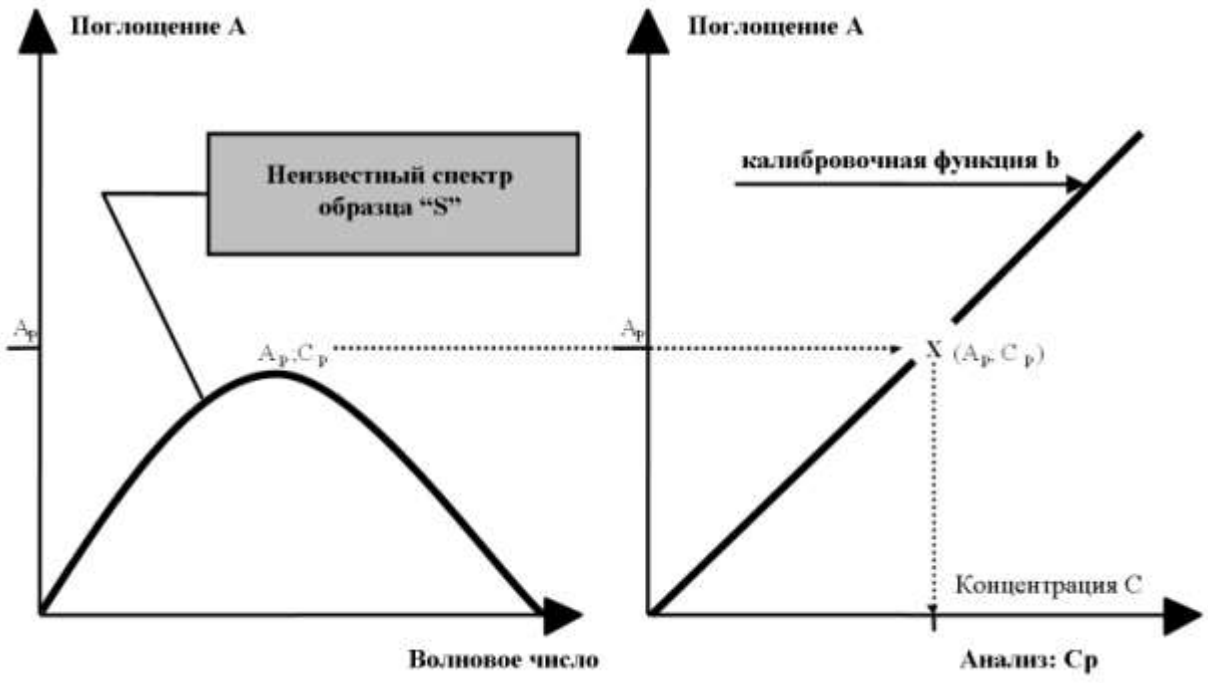


Рис. 2.3 Анализ спектра поглощения

В большинстве случаев метод однопараметрической калибровки не дает достаточно точную информацию вследствие следующих недостатков:

- Концентрация аналита коррелирует лишь с одной точкой спектра, а значит, при оценке образцов, не прошедших калибровку невозможно распознать ни выбросы, ни наличие неизвестных примесей. Другими словами, по высоте и площади пиков нельзя делать выводы о соответствии структуры измеренных спектров калибровочным данным.
- Статистические колебания сигналов, такие как шум детектора, непосредственно переплетаются с данными концентраций. Для того, чтобы быть уверенными в результатах приходится производить множественные измерения образца, а затем усреднять полученные данные.
- Для того, чтобы получать удовлетворительные результаты калибровки многокомпонентных систем, необходимо наличие достаточно четкого разграничения максимумов пиков. Однако во многих случаях, особенно когда речь идет об инфракрасной спектроскопии, этого просто невозможно сделать.
- При анализе многокомпонентных систем принимается линейная зависимость значения поглощения от длины волны для всех аналитов (закон Бугера–Ламберта–Бера). Значения поглощения располагаются напротив значений концентрации, затем применяется калибровочная функция b (см. рис. 2.2.), что зачастую не является истинным для реальных систем. Межмолекулярное взаимодействие или температурные эффекты могут искажать соответствующие линии. Более того, к некоторым методам, например к часто используемому в ИК – спектроскопии методу диффузионного отражения, закон Бугера–Ламберта–Бера неприменим.

Таким образом, однопараметрическая калибровка в случае многокомпонентных систем зачастую не дает верных результатов. В таких

случаях следует прибегать к методам многопараметрической калибровки, таким как множественная линейная регрессия (**M**ultiple **L**inear **R**egression, MLR), принципиальная компонентная регрессия (**P**rincipal **C**omponent **R**egression, PCR) и регрессия парциальных наименьших квадратов (**P**artial **L**east **S**quares, PLS).

2. Регрессия парциальных наименьших квадратов(PLS)

Со сравнительным анализом этих хемометрических методов можно ознакомиться в ранее опубликованных работах. Поскольку PLS-алгоритм применяется наиболее часто, именно он будет описан ниже. Ввиду достаточной громоздкости и сложности математических описаний, не будем приводить их в полном объеме. Подробную информацию можно найти в литературных источниках.

Для проведения PLS-регрессии для данной системы необходимо соотнести спектральную информацию с соответствующими значениями концентраций. Важно уметь распознавать изменения этих параметров и определять их зависимость друг от друга. Для этого следует произвести измерения большого количества образцов. Для математического описания полученных данных, оба параметра системы (X и Y) записывают в виде матричных элементов, после чего образуют их собственные векторы. (см. рис. 2.4). Эти векторы называются факторами первичных компонентов и могут быть использованы для прогноза концентраций вместо исходных спектров, так как содержат всю необходимую информацию об исследуемых системах. Преимущества такой замены очевидны: существенная для анализа информация из большого набора данных сжимается и выражается факторами, которые можно использовать для калибровки.

В случае PLS-регрессии собственные векторы сортируются по нисходящей. Первый фактор характеризует основные изменения

наблюдаемого спектра. Для калибровочной модели он наиболее важен. С возрастанием числа факторов привлекается все больше спектральной информации и характеризуются даже малейшие изменения в структуре спектра. Отсюда вытекает важное следствие: Низшие факторы характеризуют важные изменения спектральной структуры, когда как высшие факторы, в основном, отражают распределения спектрального шума.

Чрезвычайно важно выбрать оптимальное количество факторов для создания наиболее правильной PLS-модели. Когда факторов слишком мало, спектральная структура распознается недостаточно полно. Вследствие этого соответствующая регрессия неизбежно приводит к неудовлетворительным результатам анализа. Если же количество факторов слишком велико, неизбежны искажения результатов вследствие включения в анализ спектральных шумов.

При PLS-регрессии матрица спектральных данных X и матрица концентрационных данных Y сводятся к нескольким факторам. Для этого исходные матрицы представляются как суммы A произведений так называемых векторов оценки t_i на вектора загрузки p_i и q_i соответственно[3]:

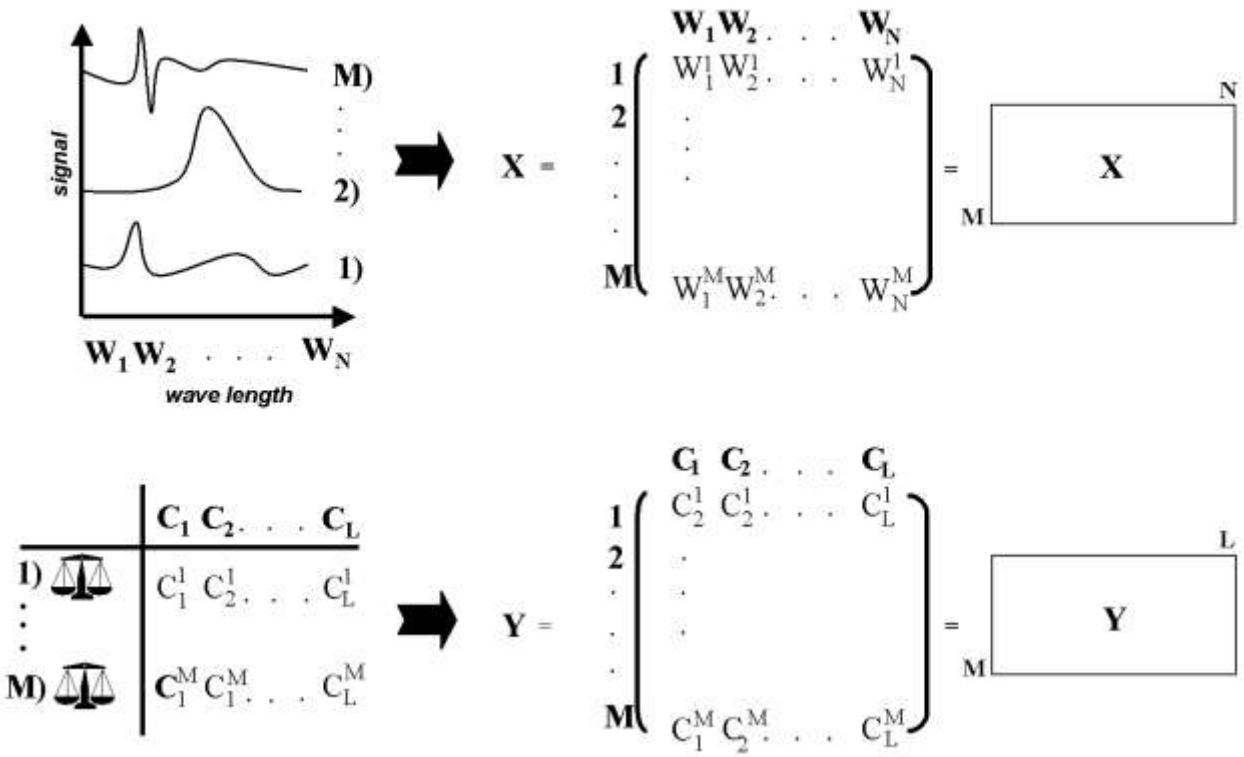


Рис.2.4 Кодировка спектральных и концентрационных данных в матричной форме. В этом примере было измерено M калибровочных образцов и N длин волн положения линий в полученных спектрах записаны в строки (M,N) – матрицы спектральных данных (матрица X). Аналогично, значения L компонент записаны в (M,L) – концентрационную матрицу.

Спектральные данные:

$$X = t_1p_1^T + t_2p_2^T + t_3p_3^T + \ldots + t_Rp_R^T + F \qquad \textit{(2-3)}$$

Концентрационные данные:

$$Y = t_1q_1^T + t_2q_2^T + t_3q_3^T + \ldots + t_Rq_R^T + G \qquad \textit{(2-4)}$$

Схематическое представление уравнения (2-3)[3]:

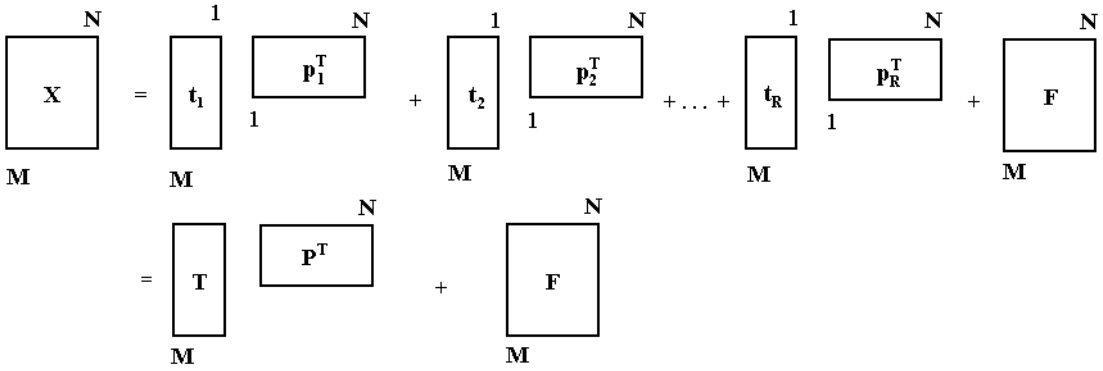


Рис. 2.5 Схематическая диаграмма для факторизации матрицы спектральных данных X

Ранг R отражает количество факторов, а T означает транспонирование соответствующих векторов загрузки. F и G – остаточные матрицы спектральных и концентрационных данных соответственно, которые отвечают за изменения структуры данных, не учитывающиеся в процессе факторизации.

Как правило, количество измеренных величин поглощения сильно превышает число компонентов. Система, таким образом, является «переопределенной» и можно выявить зависимость не только от какой-то одной спектральной характеристики (как максимум пика в случае однопараметрической калибровки), а от всей спектральной структуры. Полученная в процессе калибровки по всему спектру информация гораздо шире, чем полученная методом однопараметрической калибровки. Более того, становится возможным определять выбросы и принимать решение относительно того, приводит ли наличие неизвестных компонент, не связанных с набором данных (примесей), к спектральным изменениям. В отличие от однопараметрической калибровки, можно рассматривать спектральные характеристики, взятые со сторон пика. Таким образом, даже

сильно перекрывающиеся полосы могут быть подвергнуты анализу, а также могут быть исследованы области, содержащие много шума.

Важность PLS-регрессии в аналитической химии возрастает из-за одновременной взаимозависимой факторизации параметров X и Y. При оценке спектра поглощения можно предположить, что изменения спектральных данных происходят из-за изменения концентраций аналита. Это означает, что изменение спектральных данных должно приводить к соответствующим изменениям спектра. Следовательно, векторы оценки концентрационных и спектральных данных должны быть одинаковы. Однако, в случае реальных систем при построении матриц математическими методами, ошибки пробоподготовки и методов сравнения при определении значений концентраций, вносят погрешность измерений и спектральный шум, что приводит к различию векторов оценки. В PLS-методе предполагается идентичность векторов оценки для двух наборов данных при заданном числе факторов. Они выбираются так, чтобы иметь наименьшее возможное отклонение от исходных значений. Это является компромиссом между удобством факториального описания образцов и повышением корреляции между набором данных.

Алгоритм PLS1 рассматривает значения концентраций только одного аналита. Все другие данные интерпретируются как распределение, т.е. концентрационная Y-матрица является вектором. В алгоритме PLS2 для построения модели используются концентрации всех компонент системы. Для определения нового образца производится одновременный анализ всех веществ, прошедших калибровку. Так, в отличие от калибровки PLS1, все данные концентрационной матрицы должны коррелировать с данными спектральной матрицы, вследствие чего предсказание PLS2 менее информативно, чем предсказание PLS1. Таким образом, предпочтительным механизмом проведения калибровки является PLS1. Этот механизм успешно

применяется для всех откалиброванных компонент при проведении анализа многокомпонентной системы.

3. Оценка хемометрических моделей и анализ тестовых образцов

В предыдущей главе описан метод создания калибровочной модели на основе PLS-регрессии спектроскопических данных и соответствующих значений концентраций. Эта модель содержит ограниченное число векторов оценки и загрузки, посредством которых записываются соответственно спектральные и концентрационные данные, что позволяет рассчитать калибровочную функцию b и на ее основании проводить анализ тестовых образцов.

Таким образом, измеряются спектры тех образцов, которые необходимо изучить. Путем комбинации спектральных данных с калибровочной функцией b, из уравнения (2-1) можно получить значения концентраций аналитов:

$$Y_{Analysis} = X_{Analysis} \cdot b \qquad \textit{(3-1)}$$

где $X_{Analysis}$ – спектральные данные анализируемого образца,

$Y_{Analysis}$ – значения концентраций анализируемого образца.

Расчет калибровочной функциии b непосредственно позволяет определить концентрации аналита из данных соответствующего спектра. Степень корреляции спектральных данных и данных концентраций напрямую влияет на качество анализа. При хорошей корреляции анализ получится довольно точным, а при плохой результаты никогда не будут удовлетворительными. Следовательно, необходимо найти калибровочную функцию b с хорошей корреляцией, что в свою очередь даст возможность получить наилучшие результаты анализа.

Для этого необходимо проверить, а значит, оценить полученную модель. Такая оценка подразумевает прогнозирование при помощи построенной хемометрической модели значения концентраций определенного числа образцов с известной концентрацией аналита. Сравнение прогнозируемых и истинных значений показывает точность модели. Для оценки используют большое число различных параметров модели и набор параметров, дающий наименьшую погрешность прогноза, определяет лучший метод. Проверка разных хемометрических методов позволяет определять выбросы, наиболее удобные частотные диапазоны и дает возможность определять оптимальное число факторов. Возможны два типа проверки: внутренняя проверка (проверка калибровочных образцов) и внешняя (проверка тестовых образцов).

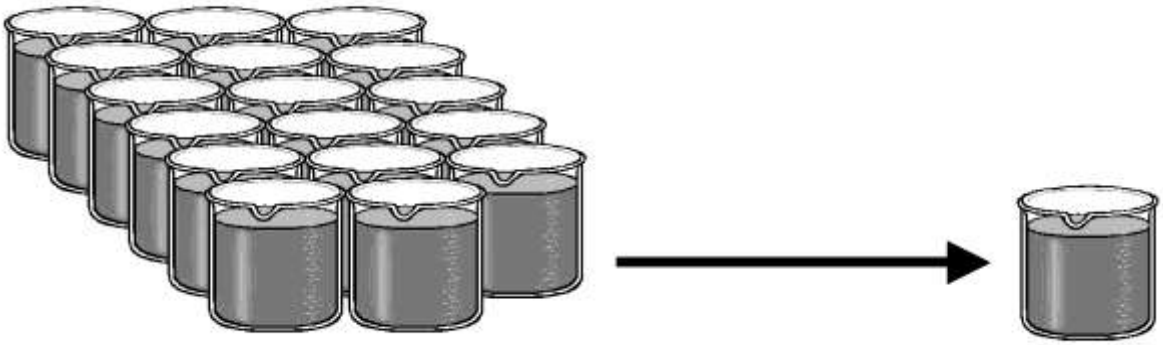
При внутренней проверке из калибровочного набора извлекаются отдельные образцы (на усмотрение пользователя). При помощи оставшихся образцов создается хемометрическая модель и используется для анализа извлеченных образцов. Сравнив результаты с реальными значениями концентраций, легко определить, насколько точно модель прогнозирует образцы. Пользователь заранее извлекает образцы, чтобы быть уверенным в том, что результаты проверки независимы. Независимость набора данных – особенно важный момент. Только в этом случае реально получить действительно точные результаты анализа.

Для оценки полного набора данных извлеченные образцы помещаются обратно, а затем анализируется новый набор тестовых образцов. Такой процесс извлечения образцов, их анализа и возвращения обратно продолжается последовательно до тех пор, пока все образцы не будут проанализированы. Сравнение значений, полученных при анализе, и исходных данных позволяет рассчитать прогнозируемую погрешность полного набора данных, а именно RMSECV (среднеквадратичную

погрешность внутренней проверки). Это мера, определяющая среднюю точность прогноза хемометрической модели. Чем меньше эта погрешность, тем выше качество модели.

При внутренней проверке важно изъять из набора совсем немного образцов, так модель, основанная на оставшемся наборе данных должна быть очень близкой к исходной модели. Если набор данных содержит менее 50 образцов, настоятельно рекомендуется брать для внутренней проверки не более одного образца.

Весь процесс схематически изображен на следующем рисунке:



калибровочный образец: разработка методики

тестовый образец: проверка методики

Рис. 3.1 Схема внутренней проверки

Последовательность внутренней проверки:

1. Взять образец из калибровочного набора.
2. Создать модель на основе оставшихся образцов.
3. Проанализировать извлеченный тестовый образец, рассчитать погрешность анализа для этого образца: Y_1^{meas}-Y_1^{pred}.
4. Вернуть образец на место и взять следующий. Рассчитать новую модель и сделать прогноз для нового образца: Y_2^{meas}-Y_2^{pred}.

5. Повторять действие 4 до тех пор, пока все M-образцы калибровочного набора не будут единожды проанализированы; рассчитать среднюю погрешность прогноза RMSECV.

$$RMSECV = \sqrt{\frac{1}{M} \cdot \sum_{i=1}^{M} (Y_i^{meas} - Y_i^{pred})^2} = \sqrt{\frac{1}{M} \cdot PRESS} \qquad \textbf{(3-2)}$$

Второй метод проверки качества модели – внешняя проверка. В отличие от внутренней, здесь для создания модели берутся все образцы калибровочного набора. Модель остается постоянной для дальнейшей проверки, т. е. анализируемые образцы не извлекаются из калибровочного набора данных. Для того чтобы оценить погрешность прогноза, последующие образцы измеряются и их добавляют в так называемый «тестовый набор». При проверке анализируются образцы только из тестового набора, то есть при внешней проверке набор данных подразделяется на набор калибровочных образцов и набор образцов для анализа. В отличие от внутренней проверки не происходит обмена между двумя наборами образцов. Сравнение результатов анализа с исходными данными концентраций позволяет рассчитать RMSEP (среднеквадратичную погрешность прогноза). Эта величина также является количественным показателем прогнозируемой точности модели. Качественные модели характеризуется низким значением RMSEP. Результаты внешней и внутренней проверки должны быть сравнимыми. В противном случае, полагается, что проанализировано слишком мало образцов для создания подходящей модели. На рис. 3.2 представлена схема внешней проверки:

различные наборы образцов

калибровочный образец: разработка методики

тестовый образец: проверка методики

Рис. 3.2 Схема внешней проверки

Последовательность внешней проверки:

1. Создать модель на основании всех калибровочных спектров.
2. Оценить модель, используя отдельный набор тестовых образцов с известными значениями концентраций, рассчитать среднеквадратичную погрешность прогноза RMSEP из соответствующих погрешностей анализа:

$$\text{RMSEP} = \sqrt{\frac{1}{M} \cdot \sum_{i=1}^{M} \left(Y_i^{meas} - Y_i^{pred}\right)^2} \qquad \textit{(3-3)}$$

Вывод: внешняя проверка отличается от внутренней тем, что при внешней проверке образцы не извлекаются из общего калибровочного набора. Модель постоянна при любом анализе. При внутренней проверке, напротив, оценка модели происходит на основании только калибровочного набора данных, и даже при ограниченном количестве образцов можно создать достаточно большое число наборов данных для создания и проверки модели.

4. Практические рекомендации по разработке калибровочной модели

Теория создания PLS-модели рассматривалась в предыдущих главах. Данная глава посвящена практическим аспектам: от сбора спектральных данных до анализа тестовых образцов. Обычно, процесс построения калибровки разбивают на шесть этапов:

Шаг 1: Ввод спектральных и концентрационных данных

Прежде чем приступить к расчету модели, необходимо задать спектры и ввести соответствующие значения концентраций для отдельных компонентов, а также определить наборы спектров для калибровки и проверки. Если набор данных достаточно велик, рекомендуется задавать одинаковое количество спектров и для калибровки, и для проверки. В том случае, если доступен лишь незначительный объем данных, необходимо создать полный калибровочный набор. Тогда не следует определять тестовые спектры, а последующая оценка модели будет проводиться исключительно путем внутренней проверки.

Шаг 2: Предварительная обработка данных

На этом этапе происходит выбор метода предварительной обработки спектральных данных. Часто необходимо исключать так называемые плывущие базовые линии. На практике для оптимизации PLS-модели зачастую приходится вычитать прямую линию, нормализовать вектор или брать первую производную спектра.

Шаг 3: Выбор оптимального диапазона частоты

Выбор оптимального диапазона частоты является определяющим для создания качественной PLS-модели. При построении модели нужно выбирать такой диапазон частот спектра, в котором можно четко наблюдать зависимость изменений спектральных данных и данных концентрации.

Степень зависимости легко оценить, используя коэффициент смешанной корреляции R^2 (см. шаг 4).

Шаг 4: Проверка и оптимизация метода

Правильность выбора методов предварительной обработки данных и диапазона частоты для поставленной задачи оценивается в результате проверки. Здесь рассчитываются такие важные параметры, как коэффициент смешанной корреляции R^2 и средние погрешности прогнозирования RMSECV или RMSEP. Кроме того, проводится автоматическое распознавание выбросов (см. Главу 5). Результаты отражаются в отчете. Для оптимизации метода коэффициент смешанной корреляции R^2 и соответствующая средняя погрешность сводятся в таблице для всех видимых комбинаций предварительной обработки данных и окон частоты (задача специалиста по применению найти значимые комбинации, здесь априори нельзя дать никакие общие рекомендации). В таблице 4 показан способ сведения результатов проверки.

Таблица 4.1 Форма для сравнения качества оценки результатов

№	Предварительная обработка данных	Частотный диапазон [$см^{-1}$]	Оптимальный порядок	Коэффициент смешанной корреляции R^2	Средняя погрешность прогнозирования	Приме-чания

Для калибровки берутся такие установочные параметры, при которых значение R^2 велико, а соответствующая средняя погрешность прогнозирования низка. Более того, целесообразно во многих случаях

выбирать такие установочные параметры, которые дают сравнительно высокие результаты проверки при меньшем количестве факторов. В процессе проверки легко распознаются потенциальные выбросы. Их выдают слишком высокие значения F или FProb. Если в результате независимой проверки образцов подтверждается, что такие значения были получены в результате ошибки измерения, их следует исключить из набора данных.

Шаг 5: Калибровка

Только после того, как из калибровочного набора данных исключены все выбросы и выбраны наиболее оптимальные параметры системы, можно приступать к созданию окончательной версии модели. В процессе калибровки рассчитываются векторы оценки и загрузки и определяется коэффициент b (см. Главу 2). После сортировки значений их можно применять для анализа новых образцов.

Шаг 6: Анализ

На этой последней стадии оптимизированная хемометрическая модель используется для анализа новых образцов. Одновременно проверяется достоверность анализа с помощью характеристических параметров. Одна из опций – это расчет так называемого «расстояния Махаланобиса». Спектральные структуры окончательного калибровочного набора сравниваются со структурой спектра аналита. Если в спектре окажутся «неподходящие» структуры или значения компонента аналита выпадают из калибровочного диапазона, имеет место превышение расстояния Махаланобиса (см. Главу 7).

Еще один способ, часто применяемый для определения выбросов – расчет спектрального остатка. Рассчитывается разница между измеренным спектром и теоретически ожидаемым в результате факторного анализа калибровочного спектра. Чем меньше эта разница (т.е. чем меньше остаток),

тем достовернее результаты анализа (см. Главу 7). Величина спектрального остатка и расстояния Махаланобиса являются количественными показателями качества анализа. Есть и другие статистические параметры, с помощью которых определяются выбросы, но здесь они не приводятся.

Таким образом, анализ дает важную информацию по двум направлениям: он показывает анализируемые значения образца и помогает определить выбросы. Т.е. в случае возникновения ошибки, ошибочное измерение отразится в некорректных результатах анализа, что пользователь сможет сразу определить.

5. Критерии отбора при создании PLS-калибровки; Оптимизация метода

В принципе, существует бесконечное множество возможностей для создания хемометрической модели. Однако, как правило, можно получить удовлетворительные результаты анализа путем оптимизации только нескольких параметров системы.

В связи с этим особую важность приобретают выбор спектрального диапазона калибровки, выбор образцов и способов первичной обработки данных. В этой главе даются наиболее важные критерии отбора, которые позволят химикам-аналитикам получать удовлетворительные результаты анализа, оптимизируя лишь некоторые параметры системы.

5.1. Отбор диапазона концентраций и значений компонент

Калибровочный диапазон концентраций должен быть немного шире, чем предполагаемый диапазон концентраций в образце. Тогда модель будет более «устойчива» в значениях тех компонент, которые лежат на границе калибровочного диапазона. Это важно, например, в тех случаях, когда необходимо уверенно определять партии дефектных изделий, появляющиеся во время производственного процесса. Значения

концентраций в таких образцах, как правило, имеют существенные отклонения, и поэтому калибровочный диапазон концентраций должен быть шире, чтобы давать надежные результаты. На рис. 5.1 для сравнения даны реперные значения и значения, полученные при анализе.

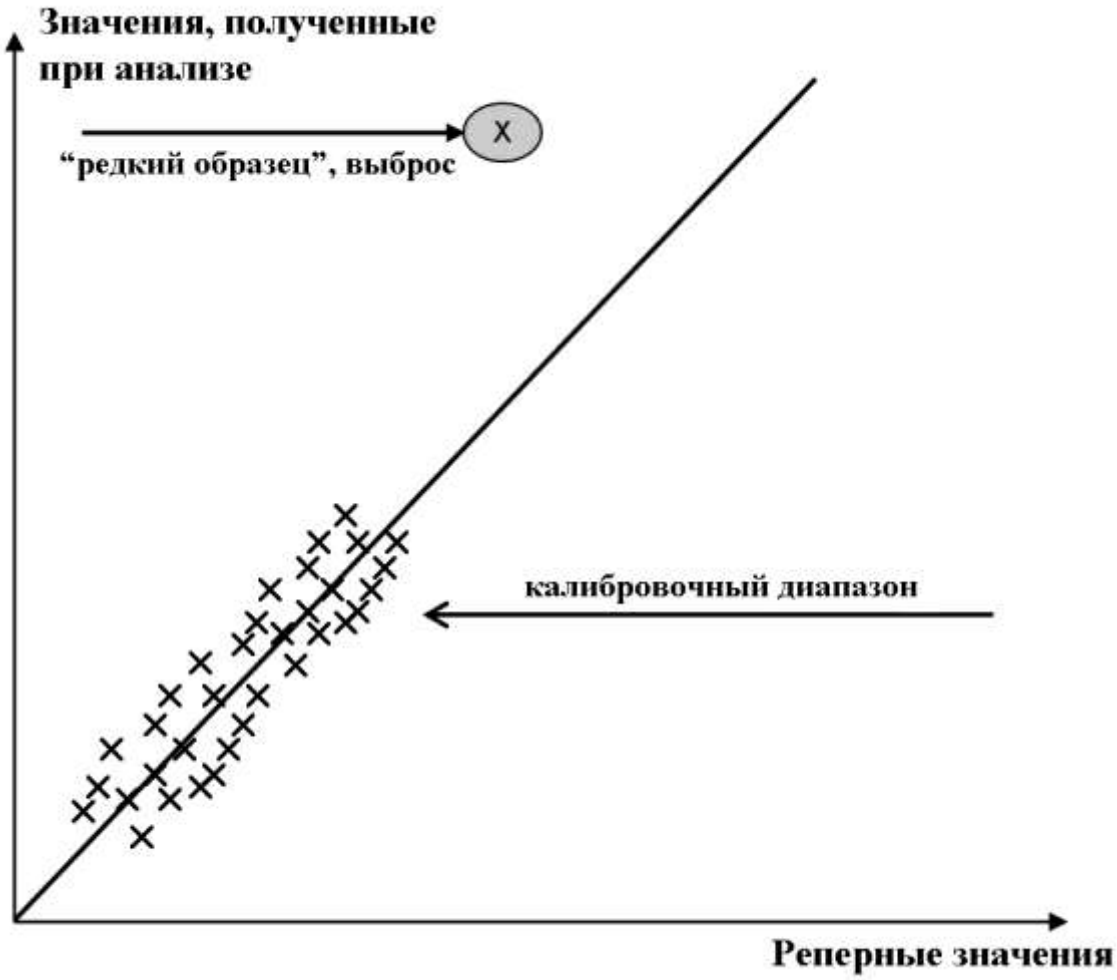


Рис.5.1 Выбор диапазона концентраций: Образцы с концентрацией «типичного» диапазона хорошо лежат на корреляционной зависимости. Аналит с не типичной концентрацией определяется как выброс.

Индивидуальные значения компонентов должны быть равномерно распределены по всему диапазону концентраций. Не рекомендуется

включать данные образцов в калибровочный набор данных, если значения концентраций этих образцов заметно выходят за рамки обычного диапазона (см. рис. 5.1). Если требуется расширить диапазон концентраций существующей модели, необходимо создать более широкий набор данных.

5.2. Подбор репрезентативных образцов для калибровки

Еще один важная проблема – проблема внешних воздействий, возникающих в процессе измерений, которых невозможно избежать во время пробоподготовки. Примером может послужить спектроскопическое in-line измерение (измерение внутри технологического процесса с помощью отвода проб) химических процессов. Зачастую нет возможности провести подготовку или термостатирование образца. Иногда, в принципе технически осуществимая подготовка образца не производится из-за своей дороговизны.

Хемометрические методы – важнейший инструмент, который помогает решить эту проблему. Перекрывание или деформацию сигналов аналита, вызванных загрязнением или колебанием температуры, можно учитывать при факторизации спектров. Поскольку все релевантные параметры системы сохраняются в независимых блоках информации (факторах), можно легко избавиться от помех. Для этого при калибровке все потенциальные помехи моделируются и сохраняются как независимые факторы. При сравнении этих значений с реперными значениями аналита, алгоритм мгновенно определит информацию, которая исходит не из самого аналита. Таким образом, соответствующие структуры не будут использоваться для прогноза новых тестовых образцов. Другими словами, PLS-алгоритм помогает различать аналитически релевантные и аналитически бесполезные структуры спектра. При калибровке помехи определяются и при дальнейшем анализе не учитываются.

Задачей разработчика метода является измерение образцов в реальных условиях. Следовательно, необходимо брать во внимание *все* потенциальные помехи, вмешивающиеся в систему, для того, чтобы «научить» алгоритм их распознавать и отбрасывать. Все колебания, которые могут иметь место в реальной ситуации, должны быть включены в калибровочный набор. Только в этом случае можно быть уверенными, что образец репрезентативен и подходит для разработки метода. Поэтому нецелесообразно проводить калибровку при идеальных условиях. Измерение чистейших, хорошо термостатированных образцов, действительно, дает очень мало ошибок при анализе, однако созданная на основании таких измерений модель не может считаться надежной для практической работы в реальных условиях.

На следующем примере продемонстрируем влияние температуры на образец:

Благодаря способности создавать водородные связи, спектр воды очень чувствителен к изменениям температуры. Когда температура повышается, молекулы H_2O начинают двигаться быстрее, что ведет к разрыву водородных связей. Существующая плотность электронов возвращается на ОН-связь, что приводит к увеличению «силовой константы». Молекула колеблется быстрее, и соответствующая полоса поглощения в спектре БИК смещается в сторону более высоких частот (волновым числам). Чем выше температура воды, тем к более высоким частотам движется полоса. Более того, поскольку вода расширяется с повышением температуры, результирующее уменьшение плотности приводит к понижению высоты сигнала. Таким образом, при повышении температуры понижается сигнальная интенсивность спектра, и одновременно максимум пика смещается к более высоким длинам волн (см. рис. 5.2).

Даже для больших температурных изменений, с различной заселенностью единичных колебательно-вращательных уровней, можно определить расширение и искажение полос поглощения. Для простоты, графическое описание этих деформаций здесь не приводится.

Калибровка по термостатированным водным растворам не даст верных результатов, если в дальнейшем анализ будет производиться при другой температуре.

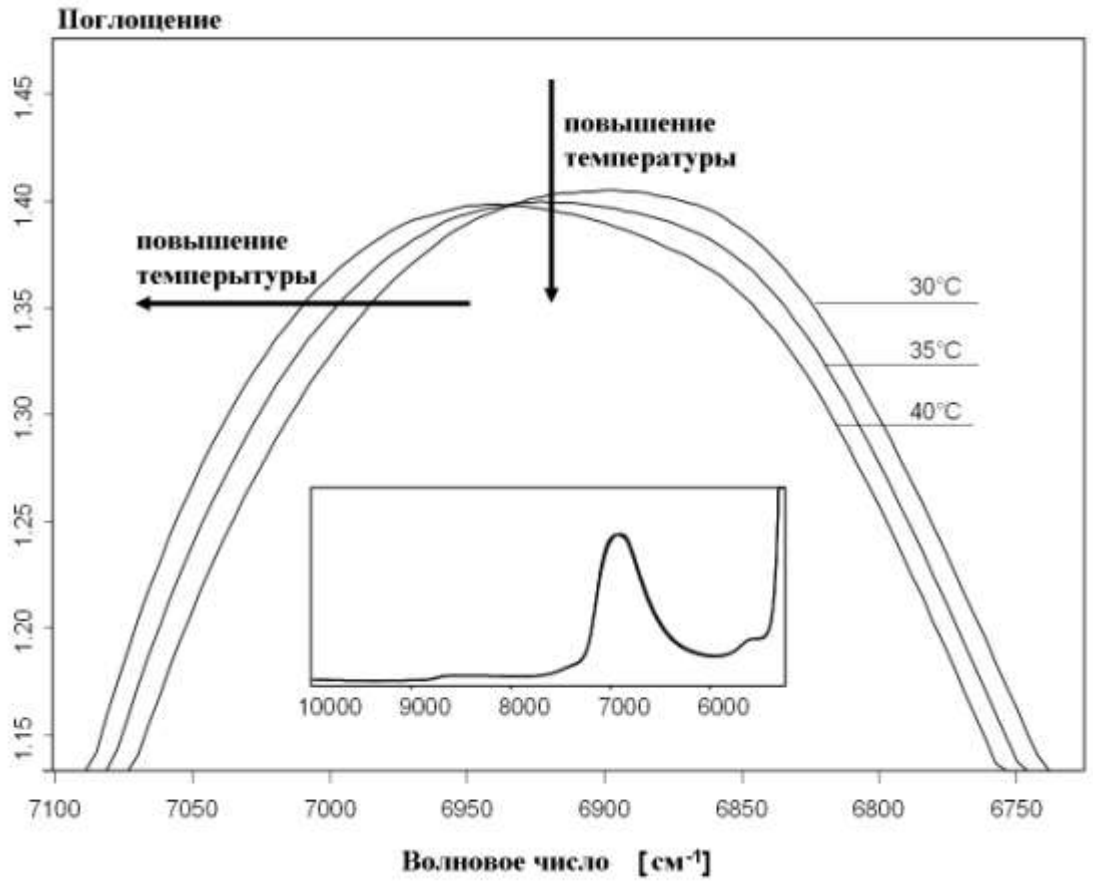


Рис. 5.2 Влияние температуры на первый обертон воды (оптическая длина пути 1 мм; в качестве образца сравнения – воздух)

Если калибровка проводилась в определенных условиях, необходимо во время анализа поддерживать те же условия. Необходимо учитывать, что

хемометрические методы должны учитывать изменения состава и качества продукта, даже в других условиях проведения измерений. В основном, метод работает только в тех условиях, при которых он разрабатывался и распознаются только те помехи, которые учитывались при калибровке. Вследствие этого утверждение модели является последним шагом.

Построение и утверждение метода оказывается сравнительно легким, если нет необходимости производить корреляцию между реперными значениями и соответствующими значениями ошибок измерения или температурных изменений. Чтобы скомпенсировать нежелательные ошибки в многопараметрической калибровке, корректируют рассчитанные при анализе значения при помощи внутренней поправочной функции. Алгоритм PLS1 ищет зависимость между компонентом и соответствующей спектральной структурой, и другая информация не учитывается. Причина возникновения погрешности – загрязнение образца, влияние другого вещества или спектрального шума – для анализа не важна (см. Глава 2).

Следовательно, для калибровки достаточно брать данные, которые были получены в достоверных условиях. Реперные значения мешающих компонент определять не требуется. Как было показано в вышеописанном примере (рис. 5.2), для калибровки водных растворов важно проводить измерения образцов в определенном репрезентативном диапазоне температур. Если же изменение качества продукта должно отражаться в конечной модели, следует добавить в существующую модель дополнительный информативный набор образцов.

5.3. Влияние на прибор окружающей среды

Репрезентативные калибровочные образцы – основа стабильной хемометрической модели. Очевидно, что внешние воздействия влияют не только на образец, но и на сам прибор, с помощью которого проводятся

измерения. В случае оптической спектрометрии речь идет как о влиянии температуры окружающей среды, так и об атмосферных воздействиях, например диффузии CO_2 и паров воды, попадающих внутрь спектрометра и/или испаряющихся из него. Обратите внимание на следующий пример:

Диффузия CO_2 и H_2O ведет к сильным помехам в случае спектра среднего ИК-диапазона, в случае же ближней ИК-спектроскопии, чтобы заметно повлиять на результаты измерений, достаточно одной только воды, растворенной в окружающей среде и проникающей в оборудование.

На рисунке 5.3 отражено влияние влаги, попадающей в спектрометр ближнего ИК-диапазона. Верхний спектр измерен сразу после реперных замеров. В нем не отражено никакого влияния дополнительного поглощения воды. Затем прибор поместили в точку с высокой влажностью воздуха. Нижний спектр снят примерно через неделю, причем не проводилось никаких новых реперных измерений. Легко распознать очень четкие структуры в области примерно 7500 - 6700 $см^{-1}$ и 5800 - 5000 $см^{-1}$, наложенные на изначальный спектр. Помехи в обоих случаях настолько сильны, что оригинальные структуры уже невозможно различить.

Спектральные области с такими помехами не могут быть включены в оценку. Следует разрабатывать методы, не учитывающие такие спектральные диапазоны. Если такой возможности нет, следует либо в кратчайшие сроки измерить новый реперный спектр, либо продуть прибор сухим воздухом.

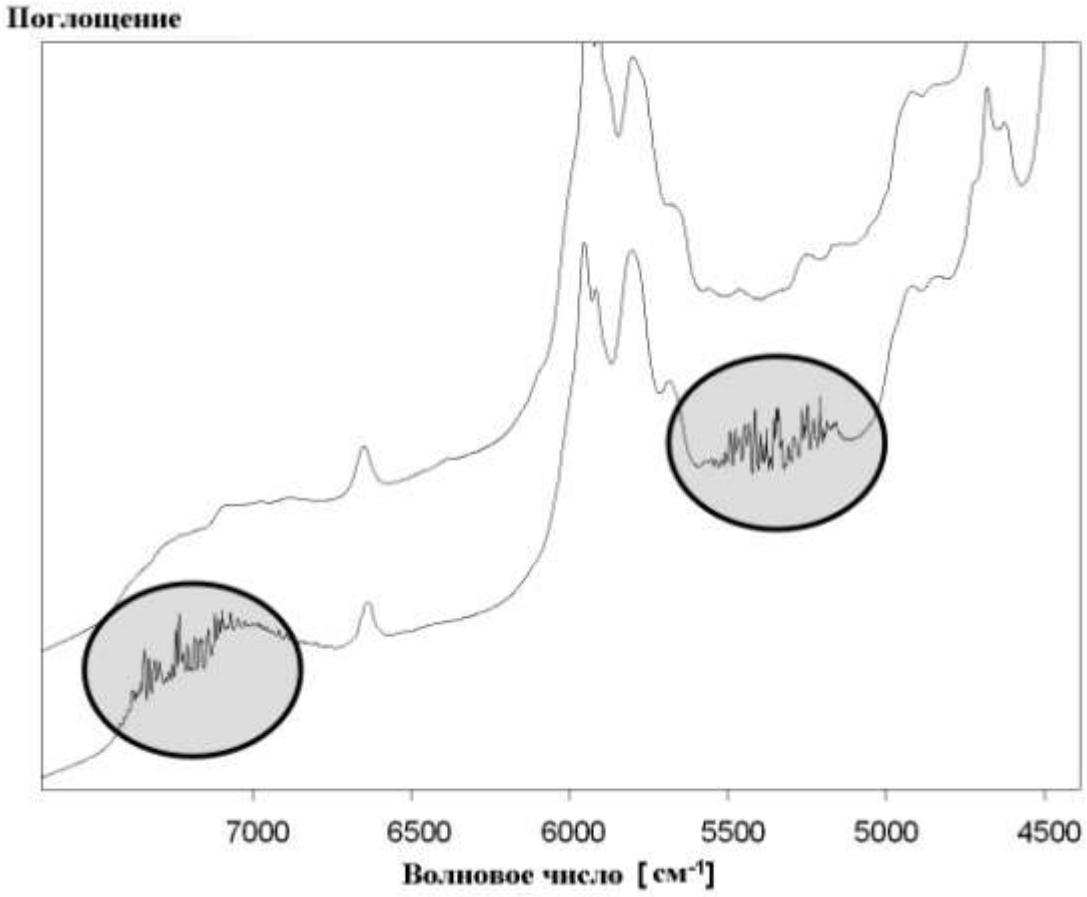


Рис. 5.3 Спектры поглощения до (верхний) и после (нижний) проникновения влажного воздуха в спектрометр ближнего ИК-диапазона (MATRIX-F, Bruker Optik GmbH). Спектры сдвинуты по оси поглощения для сравнения

5.4. Проблема коллинеарности

В принципе, существует два метода создания калибровочных наборов образцов. Во-первых, можно измерять реальные образцы и с помощью независимых аналитических методов определять реперные значения. Кроме того, можно синтезировать образцы в лаборатории. Важно не только равномерно распределить значения концентраций по всему диапазону. В случае многокомпонентных смесей необходимо, чтобы для значений концентраций изменения не были коллинеарны, то есть концентрации

соответствующих компонентов не должны увеличиваться или уменьшаться равным образом в разных образцах (см. рис. 5.4).

При коллинеарном наборе данных PLS-алгоритм не сможет четко приписать конкретные спектральные полосы соответствующим значениям компонент. Метод, разработанный таким образом, совершенно бесполезен для анализа неколлинеарных наборов данных.

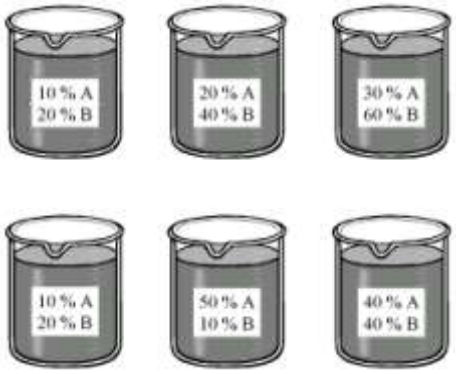


Рис. 5.4 ***Верхний ряд*****: Коллинеарный набор данных с тремя образцами, содержащими компоненты “А” и “В”. Значения единичных компонент меняется во всех образцах одинаково, т.е. набор данных коллинеарен.** ***Нижний ряд*****: Набор данных не коллинеарен, т.е. изменение состава не систематическое.**

Применение коллинеарных калибровочных наборов допустимо только при уверенности, что значения компонентов образцов, которые нужно анализировать, коллинеарны. Например, в случае любых двухкомпонентных систем. Здесь при увеличении концентрации одного компонента концентрация другого соответственно (т.е. коллинеарно) уменьшается, потому что оба значения всегда составляют 100%. Однако следует добавить, что при всей многообразности сложных многокомпонентных систем и они также закономерно коллинеарны. Так, для изделий из пластика степень

полимеризации обычно возрастает в точном соответствии с уменьшением содержания мономера. Таким образом, степень полимеризации может быть очень точно измерена с использованием спектроскопии ближнего ИК-диапазона. Определяется концентрация базовых материалов (а не собственно полимеризация), но благодаря прямой (коллинеарной) связи обоих параметров можно легко провести анализ. Данная методика универсальна. Сегодня при помощи спектроскопии ближнего ИК-диапазона можно легко определять многие физические параметры, исследовать вещества неактивные в ИК или определять концентрации аналитов, находящиеся ниже пределов обнаружения. Это возможно, когда их значения коллинеарны со значениями впоследствии определяемых компонент.

Метод недействителен, если значения компонент могут изменяться независимо друг от друга. Тогда необходимо следить, чтобы значения статистически распределялись по всем образцам. Следовательно, не рекомендуется создавать стандарты для калибровки методом разведения (разбавления), так как кроме аналита разбавляются и другие компоненты. В этом случае алгоритм не дифференцирует конкретные значения - и при проверке результат анализа окажется совершенно бесполезным.

В связи с этим не рекомендуется измерять набор калибровочных стандартов по нарастающей или нисходящей последовательности значений концентраций. Системные изменения в образцах, происходящие в процессе измерения, такие как, например, повышение температуры, могут имитировать корреляцию в свойствах системы. Однако измерение образцов повторно на более поздней стадии анализа и сравнение результатов со спектрами, полученными ранее, позволяет обеспечить исключение систематических изменений из калибровки.

Важно иметь в виду проблему коллинеарности с самого начала, еще на стадии создания калибровочных наборов данных. Позднее будет сложно распознать, не изменяется ли коллинеарно какой-то компонент системы. Следующий пример иллюстрирует вышесказанное.

Таблица 5.3 Примеры коллинеарных наборов данных.

No.	**X**	**X .3,9**	**X .0,4**	**X / 120**	**(X.0,3) + 15**	**100 – 2X**	**50 - (X.1,7)**
1	1	3,9	0,4	0,00833333	15,3	98	48,3
2	2	7,8	0,8	0,01666667	15,6	96	46,6
3	5	19,5	2	0,04166667	16,5	90	41,5
4	7	27,3	2,8	0,05833333	17,1	86	38,1
5	8	31,2	3,2	0,06666667	17,4	84	36,4
6	9	35,1	3,6	0,075	17,7	82	34,7
...	...	...	...	...	...	...	...

Числа в каждом столбце получены простыми преобразованиями из чисел в столбце «X». Следовательно, они коллинеарны. Это касается не только первого столбца. Числа каждого столбца коллинеарны по отношению к числам последующего столбца. Относительные изменения каждого значения компонента одинаковы для всех образцов. И тем не менее, не всегда можно распознать коллинеарность с первого взгляда. Значит необходимо с самого начала уделять внимание независимому распределению индивидуальных значений компонентов.

5.5. Измерения фона

Целесообразно время от времени проводить новые измерения фона. Тогда компенсируются внешние воздействия на измерительное

оборудование (т.е. спектрометр, оптоволокно, кюветы и т.д.), и достигается высокая точность результатов анализа.

Выбирая реперные вещества, следует учитывать, что они должны быть максимально инертными и не показывать полос поглощения в измеряемом спектральном диапазоне. В ближней ИК-спектроскопии целесообразно производить измерения фона при следующих условиях:

- В газовой фазе:

Средний ИК	Вакуумирование или продувка сухим воздухом или азотом
Ближний ИК	Естественные условия, реже вакуумирование или продувка сухим воздухом или азотом

- В жидкой фазе:

Средний ИК	Естественные условия, иногда чистые растворители (только при постоянной температуре (см. ниже)
Ближний ИК	Естественные условия

- В твердой фазе:

Средний ИК	Сухой бромид калия или йодид цезия
Ближний ИК	Тефлон или шероховатые металлические поверхности, рассеивающие свет (главным образом, золото)

В спектроскопии комбинационного рассеяния, как правило, измерения фона не проводятся.

Влияние на измерения атмосферных газов (особенно CO_2 и испарений H_2O) настолько велико в среднем ИК-диапазоне, что приходится продувать спектрометры азотом.

В отличие от среднего, ближний ИК-диапазон не столь чувствителен, и заметное влияние на него оказывают лишь рассеянные в атмосфере пары воды. Их влияние обычно компенсируют регулярными реперными измерениями (например, раз в час). Следовательно, в ближнем ИК-диапазоне измерения можно проводить уже без дальнейших мер предосторожности. Однако, далеко не во всех случаях можно так поступать. Оборудование в течение нескольких месяцев обычно работает при одних данных измерения фона. Изменения влажности окружающего воздуха влияют на результаты анализов в спектральных диапазонах 7500 - 6700 $см^{-1}$ и 5800 - 5000 $см^{-1}$ (см. выше). Если важно произвести оценку в этих диапазонах, следует продуть приборы азотом или очищенным от масел воздухом.

Важным требованием при стандартизации спектров и методов является использование в качестве репера инертных материалов, а также материалов с малым количеством полос поглощения или не имеющих полос поглощения в заданном диапазоне. Как уже говорилось, в (Б)ИК-спектроскопии высота, ширина, форма и расположение полос поглощения зависят от внешней температуры. Если реперные измерения и измерения образцов не проводить при абсолютно одинаковых температурах, то появятся артефакты реперных материалов, при условии, что материалы имеют полосы поглощения в наблюдаемом спектральном диапазоне. Причем появление артефактов следует ожидать на тех же частотах, где лежат полосы поглощения материалов. На рис. 5.4 приведены спектры БИК воды:

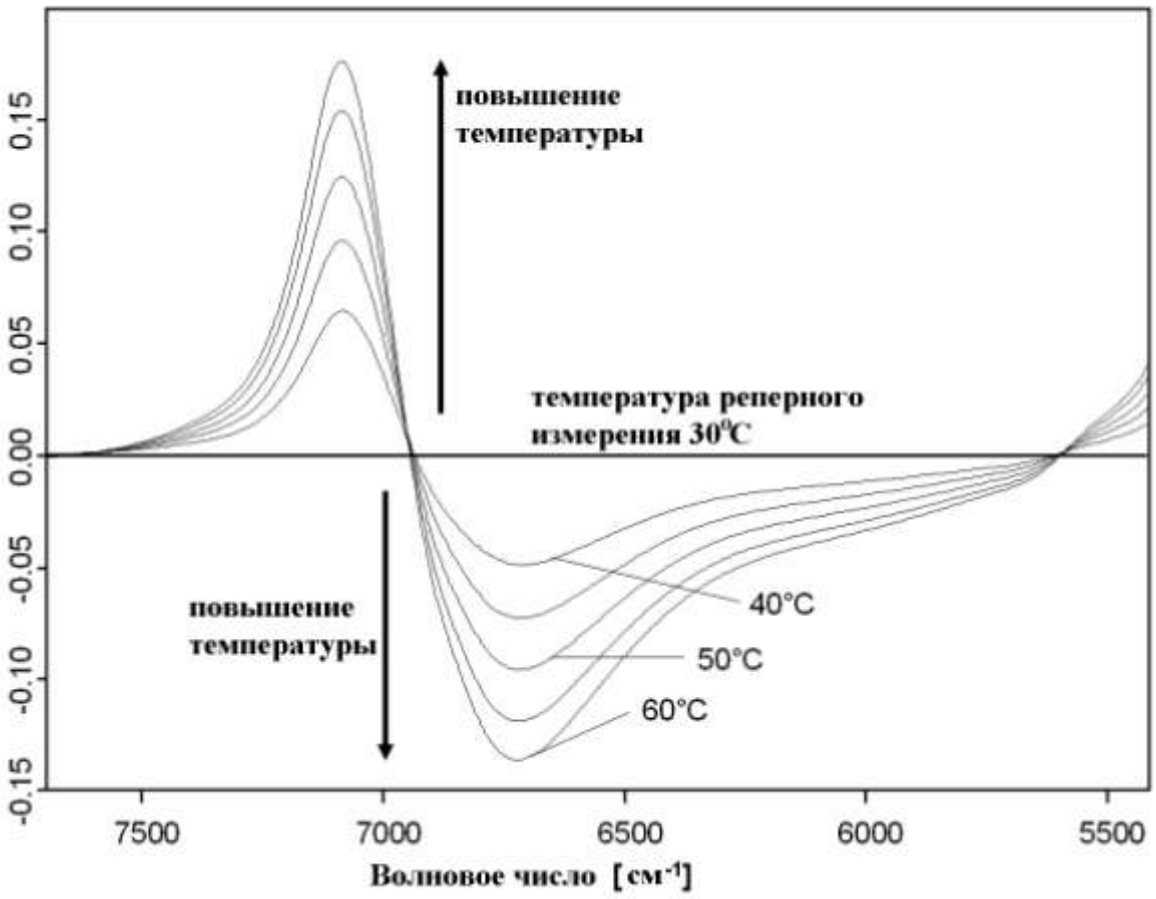


Рис. 5.4 Спектры сдвига первого обертона воды с ростом температуры. Для измерения и фона, и образца использовали чистую воду. Фон регистрировали при 30°C, образцы измеряли в диапазоне температур 40 - 60 °C.

В приведенном эксперименте были измерены те же образцы, что и в рис. 5.2. Однако, в рис. 5.2 реперные точки регистрировали по воздуху, и никаких артефактов не было. В данном эксперименте в качестве репера использовали воду, что позволило наблюдать структуры, наложенные на исходный спектр, которые в свою очередь вносят ошибку в хемометрическую модель. При измерении по воздуху наблюдается лишь небольшое отклонение (см. рис. 5.2), которое совсем немного влияет на результаты анализа. В колебательной спектроскопии нецелесообразно брать

матрицу образца для измерений фона. Уравнение, наиболее широко используемое в химическом анализе:

[спектр вещества] – [спектр матрицы] = [чистый спектр аналита],
или
[матрица + аналит] – [матрица] = [аналит]

Это уравнение неприменимо в (Б)ИК-спектроскопии и спектроскопии комбинационного рассеяния (исключением может быть только абсолютно одинаковые температуры при измерении фона и образца). Предпочтительней проводить измерения фона по воздуху (или азоту) или по реперному материалу, который не дает полос поглощения в соответствующем спектральном диапазоне.

5.6. Важность параметров измерения и реперных значений

В анализе (Б)ИК-Фурье спектров и спектров комбинационного рассеяния, число спектральных данных обычно существенно превышает число компонент. Каждый компонент ассоциируется не с определенной точкой спектра (как при однопараметрической калибровке), а со всем спектром. Такая система намного более информативна, а PLS-алгоритм позволяет до определенной степени скомпенсировать статистические погрешности в калибровочных данных. Однако очевидно, что при неполном наборе данных невозможно создать строгую модель. Значит от качества спектрометра (его разрешения, устойчивости, коэффициента сигнал/шум и особенно точности измерений) зависит качество создаваемой модели.

В данном контексте особенную важность приобретает качество реперных данных и аккуратность пробоподготовки образцов. Опыт показывает, что в большинстве случаев некачественные хемометрические

модели появляются именно от невнимательного отношения к технике либо халатности при подготовке образцов. Ясно, что если допускать колебания температур при измерениях или исследовать негомогенные или нечистые материалы, это приведет к серьезным ошибкам. Ошибки спектрометра или неточности самой PLS-модели обычно не ведут к фатальным последствиям, так что необходимо иметь в виду, что в первую очередь именно внимательный подход к технике процесса и надежные реперные методы определяют качество анализа.

Если качество реперных образцов невозможно или очень сложно улучшить, целесообразно сделать несколько повторных замеров того же образца и усреднить результаты. Тогда усредняются и статистические погрешности, а значит, возможные выбросы оказывают существенно меньшее влияние на конечный результат анализа. Те же методы действительны и для спектрометрического анализа, то есть в исключительных случаях допустимо проведение повторных измерений и усреднение результатов.

5.7. Выбор спектральных данных

В принципе, нет никаких ограничений при выборе спектральных данных для построения PLS-калибровки. Можно выбирать данные из интерферограмм, из однолучевых спектров, спектров пропускания, спектров поглощения и т.д. В большинстве случаев целесообразно использовать спектры поглощения, поскольку действует закон Бугера-Ламберта-Бэра. Спектры поглощения лучше отражают линейные отношения между значениями поглощения и данными концентрации, а при факторизации PLS-алгоритма линейные отношения предпочтительны.

5.8. Выбор методов внутренней и внешней оценки

Хемометрическая модель строится на основания калибровочных данных, и затем ее качество тестируется с помощью тестовых спектров. Существует два метода такой проверки: внутренняя (на основе калибровочных образцов) и внешняя (на основе тестовых образцов).

Внешняя проверка: Берут два независимых набора образцов, один для калибровки системы, другой для проверки соответствующей модели. Оба набора должны содержать одинаковое число образцов, и в каждом наборе весь диапазон концентраций должен быть распределен равномерно по системе.

Внутренняя проверка: Если доступных образцов недостаточно (менее 50), приходится обходиться без отдельного тестового набора. Тогда надежно проверить качество модели можно только при помощи внутренней проверки.

На практике внутреннюю проверку проводят для первичной оценки метода. Только если она гарантирует, что был измерен информативный набор спектров, есть смысл проводить внешнюю проверку. А внешняя проверка предпочтительна, когда доступно большое количество образцов, так как такая проверка значительно экономит время расчетов по сравнению с внутренней проверкой.

Большинство программ предлагают оба варианта проверки. Это помогает аналитику-практику оценивать устойчивость метода. Проводится проверка следующим образом:

Первый шаг (как описано выше) – выбор лучших параметров модели с помощью внутренней проверки. Затем записывают результаты R^2 и RMSECV. Второй шаг – набор данных разбивают на две равные части. Один «пакет» - калибровочный набор данных, другой – тестовый. Затем, строят калибровочную модель на ранее определенных параметрах, проводят внешнюю проверку и записывают значения R^2 и RMSECV. Третий шаг –

тестовый и калибровочный наборы данных меняют местами и проверяют с теми же параметрами модели (вторая проверка тестового набора). Результаты R^2 и RMSEP сравнивают с прежде полученными в результате внешней проверки значениями, так же, как при внутренней проверке. Таблица 5.3 показывает погрешность анализа.

Таблица 5.3 Проверка стабильности калибровки

	Значение ошибки (внутренняя проверка)	**Значение ошибки (тестовый набор 1)**	**Значение ошибки (тестовый набор 2)**
Стабильная калибровка	0,10	0,11	0,10
Нестабильная калибровка	0,10	0,25	0,33

Если набор данных достаточно информативен и количественно выдержан (а это обязательное условие для создания устойчивой модели), все значения окажутся в пределах небольшой и допустимой погрешности. Однако, если при внутренней проверке погрешность существенно больше, чем при внешней, это может означать, что было измерено недостаточное количество образцов. Следовательно, для внутренней проверки, в противовес внешней, в калибровочный набор требуется включать приблизительно в два раза больше образцов. При недостаточном количестве образцов результаты проверки серьезно искажаются.

И наоборот, если набор данных достаточно велик, при делении пополам общего количества калибровочных спектров, образующиеся

наборы будут иметь схожие ошибки. В этом случае можно полагать, что калибровка окажется устойчивой. Более того, в основном рекомендуется проводить как метод оптимизации с первым тестовым набором образцов и проверку модели на устойчивость, так и вторую проверку тестового набора. Чем больше различных комбинаций протестировать, тем легче определить приводит ли деление набора данных к потере точности метода и следует ли измерять дополнительные образцы.

Независимо от типа проверки, важно не включать одни и те же образцы в калибровочный и тестовый набор данных. А также при многократных измерениях одних и тех же образцов недопустимо разделять отдельные спектры между двумя наборами данных (см. главу 6: «PLS-регрессия – безграничная точность анализа»).

5.9. Выбор спектральных диапазонов

PLS-регрессия – так называемый метод полного спектра, то есть чем больше спектральных характеристик, тем больше спектральной информации о компоненте и тем лучше соответствующая модель. Однако необходимо иметь в виду, что включения спектральных шумов или полос поглощения вмешивающихся компонентов могут ухудшить качество модели. В этих случаях PLS-алгоритм вычисляет спектральные структуры, не относящиеся к исследуемому компоненту, а результаты анализа существенно ухудшаются.

На рис. 5.6 изображен БИК-спектр воды (измерения произведены методом пропускания, оптическая длина пути 2 мм). Между 5200 $см^{-1}$ и 4000 $см^{-1}$ – полное поглощение, ниже 4000 $см^{-1}$ – серьезные включения спектрального шума. Оба диапазона не могут быть использованы для калибровки.

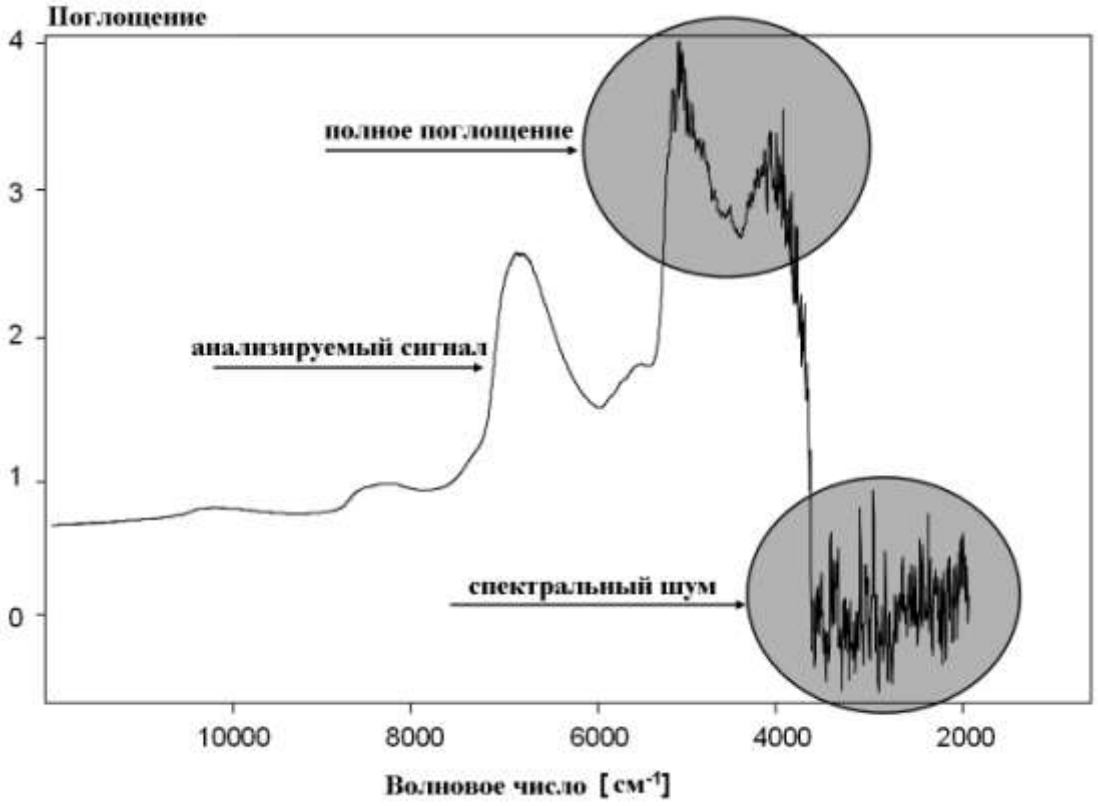


Рис. 5.6 Спектр воды в кювете (оптическая длина пути 2 мм)

При поиске оптимума рекомендуется измерять все образцы по всему спектральному диапазону, а затем искать спектральные особенности. Обычно оптимальными являются полосы поглощения со значениями, лежащими между 0.7 и 1.0. Когда измерения проводят на современных Фурье-спектрометрах, для калибровки, как правило, можно брать значения поглощения до 2.5. Однако важно, чтобы детектор работал линейно по всему диапазону (т.е. охлаждаемые детекторы InAs или InGaAs). Сигналы большей величины не рассматривают, так как они дают менее надежные результаты при измерениях.

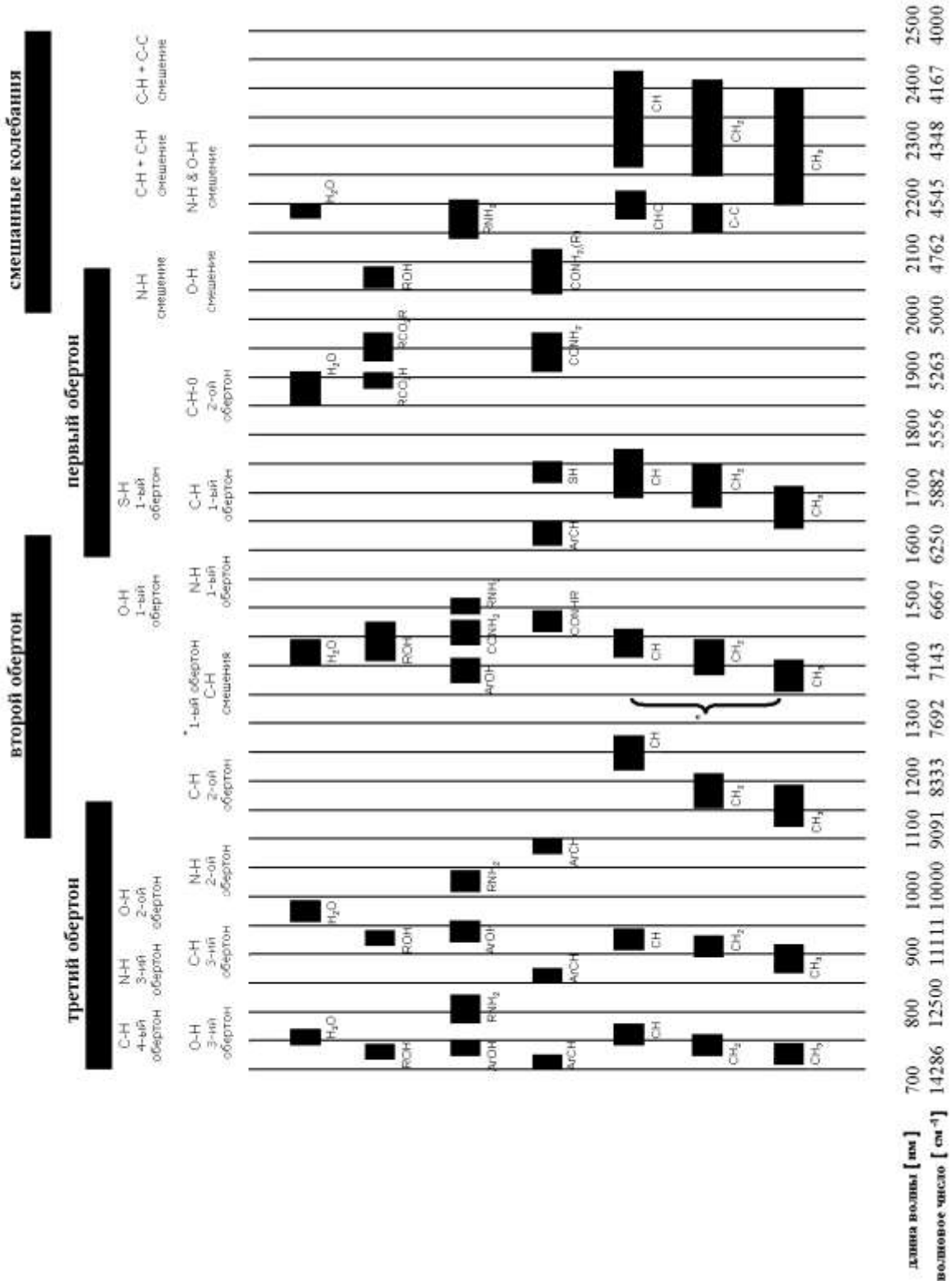
В любом случае, рекомендуется убирать последовательно целые группы сигналов. Так, в ИК, БИК-спектроскопии и спектроскопии комбинационного рассеяния у многих веществ (за небольшим исключением) наблюдаются сигналы в большом спектральном диапазоне и

зачастую не требуется искать выделенную структуру. В таблице 5.3 и на рис. 5.7 представлен краткий обзор частотных диапазонов, обычно рассматриваемых в БИК-спектроскопии (более подробная информация дана в приложении).

Таблица 5.3 Поглощение сигналов в ближнем ИК-диапазоне

Группа	**Частотный диапазон [см-1]**	**Наименование**
Алифатические углеводороды	6.300 - 5.500	1. Обертон CH-stretching
	9.100 - 7.800	2. Обертон CH-stretching
	5.000 - 4.100	смешение
	7.700 - 6.900	смешение
Ароматические углеводороды	ca. 6.000	1. Обертон CH-stretching
	ca. 9.000	2. Обертон CH-stretching
	4.700 - 4.000	смешение
	7.300 - 6.900	смешение
Карбоксильная кислота	ca. 6.900 ca. 5.250 4.900 - 4.600	1. Обертон CH-stretching 2.Обертон CO-stretching смешение
Амины	7.000 - 6.500 5.200 - 4.500	1. Обертон NH-stretching смешение
Вода (очень сильное поглощение)	7.500 - 6.400 5.400 - 4.900	1. Обертон OH-stretching смешение

Рис. 5.7 Полосы поглощения различных функциональных групп в ближнем ИК-диапазоне



5.10. Выбор первичной обработки данных

Кроме выбора правильного частотного диапазона, первичная обработка данных – второй по важности вопрос при создании модели. Цель первичной обработки данных - смоделировать спектры таким образом, чтобы PLS-алгоритм мог установить хорошую корреляцию между спектральными данными и данными концентраций.

Отсутствие первичной обработки данных: первичная обработка данных не проводится.

Вычитание постоянного сдвига: спектры сдвигаются линейно таким образом, чтобы значение y стало равным нулю.

Применение: убирается постоянный сдвиг, который может появится, к примеру, в результате различия в назначенном усилении детектора.

Вычитание прямой линии: В каждом частотном диапазоне при помощи метода PLS прямая линия подгоняется под спектр, а затем вычитается из соответствующего спектра.

Применение: Убирается линейный наклон сдвига (см. рис. 5.8).

Нормализация вектора: Прежде всего, спектры центрируют. Затем вычисляют сумму всех квадратов всех значений Y и делят соответствующие спектры на квадратный корень суммы. Так называемый нормальный вектор конечного спектра всегда равен 1.

Применение: В принципе спектр содержит два вида информации: о высоте полос и об их структуре. В результате нормализации информация о высоте перестает браться в расчет, а остается лишь информация о структуре. Нормализацию применяют, например, для того, чтобы убрать влияние различия в оптическом пути для случаев измерений методом пропускания. Оптический путь изменяет высоту сигнала, но не его структуру. Также при измерениях диффузного отражения можно минимизировать влияние различной плотности материалов.

Нормализация минимума-максимума (для спектров поглощения)**:** Спектры линейно сдвигаются так, что минимум значений Y становится равен нулю. Затем спектры расширяют по оси так, что максимум значений Y равняется двум единицам поглощения (рис. 5.8).

Применение: схожее с тем, которое дает нормализация вектора.

Коррекция множественного разброса значений: Прежде всего, рассчитывают средний спектр по данным всех спектров калибровочного набора. Затем каждый спектр X (i) изменяют в соответствии с:

$$X(i) = u + v * X(i) \qquad (5\text{-}1)$$

Коэффициенты u и v выбирают так, чтобы разность между измененным вектором X (i) и средним спектром была минимальной.

Применение: Часто применяют при измерениях методом диффузного отражения.

Первая производная: расчет первой производной спектра: вычисляется первая производная спектра (рис. 5.8).

Применение: Когда первая производная рассчитана, заметные сигналы становятся более четко выраженными относительно плоских полос. Метод применяется для того, чтобы выделить заметные, но все же слабо выраженные по сравнению с огромными широкополосными структурами черты. Также важно, что метод применяется для расчета широких полос, что часто происходит в методе спектроскопии ближнего ИК-диапазона. После расчета такие структуры получают более четкую форму, что

значительно облегчает дальнейшие расчеты. Метод также позволяет устранить наклон базовой линии.

Когда расчет первой производной используется в качестве метода первичной обработки данных, следует иметь в виду, что также увеличиваются спектральные шумы. Они накладываются на спектр и могут исказить сигналы аналита. Рекомендуется увеличить число накоплений спектров. Допустимо небольшое сглаживание шумов.

Вторая производная: расчет второй производной соответствующего спектра.

Применение: В отличие от расчета первой производной, при данном методе можно рассчитать даже чрезвычайно плоские спектры. При этом влияние спектральных шумов обычно настолько велико, что рассчитывать спектры можно только в очень ограниченном диапазоне.

На рис. 5.8 показаны различные методы первичной обработки данных и их влияние на вид спектра ближнего ИК-диапазона (измерения человеческой руки оптоволоконным датчиком). На оригинальном спектре виден небольшой сдвиг базовой линии и ее небольшой дрейф. Дрейф можно убрать вычитанием прямой линии (на рисунке – пунктирная линия, а сдвиг – нормализацией минимума-максимума (на рисунке – точечная линия). Первая производная изначального спектра (на рисунке – линия «точка-тире») для наглядности расширена и сдвинута к более высоким значениям поглощения. Наблюдается относительное усиление четких структур по сравнению с изначальным спектром.

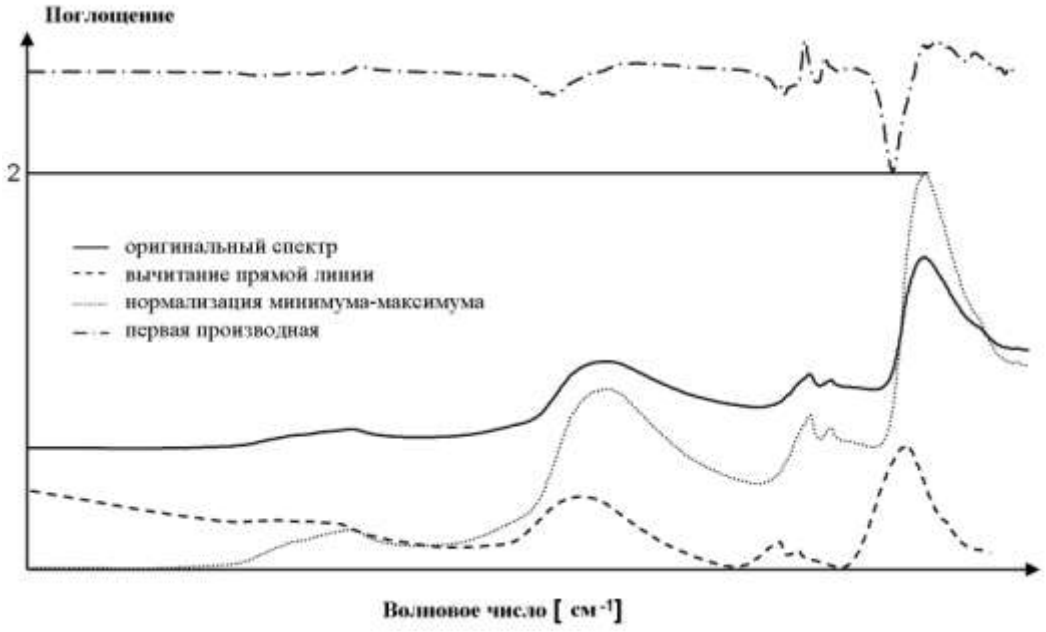


Рис. 5.8 Спектр ближнего ИК-диапазона человеческой руки; измерения методом диффузного отражения

Выбор оптимального метода сильно зависит от изучаемой системы. Опыт показывает, что в подавляющем большинстве случаев вычитание прямой линии, нормализация спектра или первая производная дают лучшие результаты. В небольшом числе случаев, лучшие результаты дает комбинация двух методов. Зачастую ряд обработок дает одинаково хорошие результаты (см. Главу 6), так что возможно применение различных методов первичной обработки данных.

5.11. Выбор подходящего числа факторов

При PLS-калибровке данные концентраций и спектральные данные сначала записываются в форме матриц, а затем сводятся только к нескольким факторам. Число факторов хемометрической модели называют «рангом». Выбор количества рангов – важнейший момент, от которого зависит качество анализа. Если выбрать недостаточное число факторов, будет трудно объяснить изменения, происходящие в концентрационных и

спектральных данных. Зависимость между двумя наборами данных столь мала, что на ее основании невозможно сделать выводы. При выборе слишком большого числа факторов в расчет берутся даже малейшие изменения в наборе данных, такие, например, как спектральные шумы, и в калибровке участвует масса несущественной для анализа информации, что далеко не лучшим образом влияет на качество. И в том, и другом случае, такие модели существенно искажают результаты. Таким образом, важно подобрать оптимальное число факторов, которое будет гарантировать минимизацию ошибок в анализе.

Существует множество индикаторов, которые позволяют определить оптимальное число факторов для конкретной модели: средняя погрешность прогноза (RMSECV для внутренней проверки и/или RMSEP – для внешней) при правильном выборе числа факторов будет минимальной, а R^2 примет максимальное значение. То есть, найти оптимальное число факторов нетрудно. Сначала следует рассчитать R^2 и среднюю погрешность, а также построить график зависимости R^2 и средней погрешности от ранга. Ранг можно считать оптимальным, когда упомянутые характеристики показывают оптимальные значения и/или существенно не изменяются по мере увеличения числа факторов. Если несколько рангов дают удовлетворительные результаты, рекомендуется выбирать модель с наименьшим количеством факторов (см. Главу 6).

Внимание: *Проверку метода можно проводить только на независимых тестовых спектрах, то есть спектры ни в коем случае не должны быть включены в калибровочный набор данных. Это может появиться при измерении, когда одного образца могут быть определены как выбросы. Для проверки тестового набора должны быть измерены новые образцы.*

5.12. Выбор удобных калибровочных образцов; распознавание выбросов

Если калибровочный набор содержит выбросы, их легко распознать после проверки модели. Индикаторами наличия выбросов являются:

Большое значение спектрального остатка;

Большое значение F;

Величина FProb = 0,99 и выше.

Относительно высокая погрешность анализа, которая приводит к высоким значениям «расхождения» (величины, показывающие расхождение между реперными значениями и значениями, полученными в результате анализа).

В большинстве случаев полезно графически сравнить реперные значения и значения, полученные в результате анализа. У выбросов значения сильно отличаются от соответствующих реперных. После того, как выбросы извлечены из калибровочного набора, разрабатывают новую модель, которая при проверке уже не содержит выбросов. Может получиться, что даже после извлечения выбросов, модель не является оптимальной. Так бывает, когда выбросов слишком много, поэтому они существенно влияют на качество модели. Тогда, нужно оптимизировать ее шаг за шагом, последовательно извлекая выбросы.

Внимание: *Всегда необходимо разобраться, не возникли ли выбросы в результате некорректно совершенных измерений. Исключение типичных для эксперимента, но неподходящих образцов недопустимо, поскольку модель должна быть устойчивой для правильной оценки разнообразных образцов в реальных условиях.*

Часто в качестве выбросов фиксируют образцы, лежащие вне калибровочного диапазона. Однако причиной такого результата являются уже не ошибки в измерениях, а недостаточная устойчивость метода. Тогда

при дальнейших измерениях задачей разработчика будет равномерно распределить значения компонентов по всему диапазону, а не убирать те немногие значения выбросов, которые появились в данной модели.

Не следует слепо следовать результатам проверки и убирать все образцы, помеченные как выбросы. Обычно хемометрическое программное обеспечение считает значение достоверным только на основе статистических параметров. Очевидно, что рассматривая эти значения без знания характерных параметров веществ, невозможно делать выводы, является ли обозначенный спектр выбросом. Следовательно, только сам аналитик, произведя независимое исследование спектра, может в каждом конкретном случае исключать значения.

5.13. оценка результатов проверки

Одной из важнейших целей разработчика является проверка пригодности разработанного набора данных для решения поставленной задачи. Самый простой способ – многократное измерение образцов и последовательное нанесение на график полученных реальных значений против прогнозируемых. В идеальном случае реперные значения и значения результатов анализа для всех измерений окажутся в пределах предполагаемой погрешности.

Однако это не относится к любой калибровочной модели. Иногда значения RMSECV или RMSEP не устраивают исследователя. Тогда его задачей является найти разумное объяснение данному факту и принять меры для улучшения результата. Если же это невозможно, важно быстро оценить, что выбранный аналитический метод непригоден для анализа образцов, и разработка такого метода должна быть прекращена. Ниже приведены два примера описания этого.

На рис. 5.9 сравниваются реальные и прогнозируемые значения для двух калибровок, построенных с недостаточной точностью. Каждый образец был измерен 3 раза. На рис. показано, что значения далеко отнесены от биссектрисы. Однако, с левой стороны видно, что измерение образцов каждой серии осуществлено с высокой точностью, поскольку находятся примерно на одном расстоянии от прямой линии. Другими словами: Измерение не является верным, но произведено точно. Это позволяет сделать вывод о том, что аналитический метод будет удобен для решения задач последующего измерения. Возможно, причины неточности в чем-то другом.

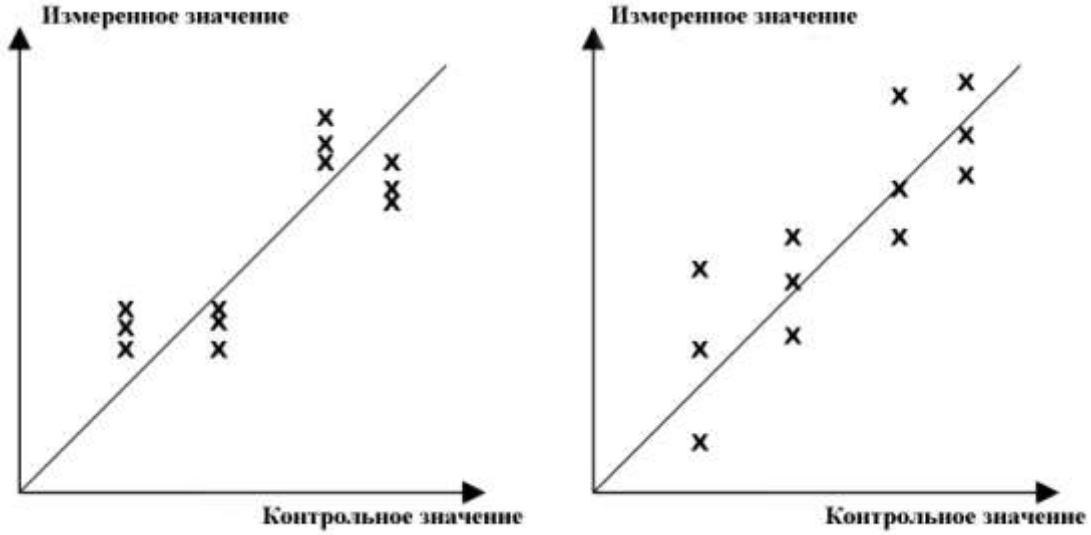


Рис. 5.9 Представление двух калибровок с недостаточной точностью

Например, при неточных реперных значениях или некачественной подготовке образцов могут появиться статистические отклонения значений анализа от реальных значений. При проведении измерений в разных температурных режимах или с различными начальными условиями могут возникать систематические ошибки. На рис. 5.9 показано, что в таком

случае точки могут скапливаться выше или ниже прямой линии. Более того, к подобному результату может привести неточность хемометрического метода. Например, для описания сложных многокомпонентных систем требуется измерить относительно большое число репрезентативных калибровочных стандартов. В противном случае алгоритм не может правильно связать конкретные структуры спектра аналита с соответствующими компонентами. Данные анализа тестовых образцов будут точными, но неверными, и модель окажется неприменимой для дальнейшей работы со спектрами.

Во многих случаях довольно просто найти причины плохого анализа, если отдельно значения измерены с высокой точностью. Гораздо сложнее, если наблюдается существенный разброс конкретных точек (правый график на рис. 5.9). Здесь значения не только неверны, но и невоспроизводимы. В таких случаях причина либо в неправильной подготовке образцов, либо в начальных условиях. Например, если материал неоднороден или для измерений брали слишком малую часть материала, точных значений не получится, даже при проведении точных локальных измерений. При оценке результатов это необходимо иметь в виду. Если невозможно принять адекватные меры для улучшения результатов, часто имеет смысл добавить образцы в калибровочную модель. Если и после этого не достигается удовлетворительного соответствия прогнозируемым значениям, приходится смириться с тем, что модель оказалась неприменимой для данной задачи.

5.14. Выполнение и утверждение методов

Важнейшими задачами разработчика являются обработка репрезентативных калибровочных данных и учет возможных ошибок в измерениях или в работе оборудования. И поскольку приходится не только

изучать образцы и уметь проводить измерения, но и разбираться в технике измерений и оценке метода, от аналитика требуются глубокие знания.

В каждой компании должен быть специалист по калибровке. Причем в каждой отрасли промышленности должен быть квалифицированный специалист по калибровке. Помимо собственно калибровки, он должен оценивать предполагаемую точность в контексте выполнимости анализа. Также, он должен предлагать «инструкцию» последовательного осуществления метода.

Рекомендуется периодически проверять достоверность хемометрического метода. Можно проверять метод с помощью независимого тестового набора, например, раз в несколько месяцев или после определенного количества замеров. В некоторых отраслях промышленности, например, в фармацевтической промышленности, такие проверки обязательны (препарат допускают к продаже только при наличии документа, удостоверяющего, что он прошел очередную проверку).

Но даже если такой регламентации нет, необходимо «наблюдать» за тем, как работает метод. Метод, изначально работающий хорошо может со временем становится хуже. Этому факту может быть несколько объяснений: Во-первых, причиной этого может быть ухудшение состояния самого прибора (кювет, оптоволоконных датчиков и самого оптоволокна). Кроме того, на качество анализа может влиять изменение положения прибора, и разное качество сырья. Регулярно следить за состоянием методики особенно важно, когда дело касается калибровки природных материалов или нефтехимического сырья.

Для многих применений важно как можно раньше определить медленное ухудшение качества анализа. Современные задачи аналитики и технологии не позволяют поводить улучшение метода, если он перестает подходить для данной задачи. Установка метода должна быть

продолжительным процессом и являться «фоном» повседневного рутинного анализа.

Если становится ясно, что анализ тестовых образцов систематически дает сбои, сначала необходимо сравнить новые спектры со снятыми ранее калибровочными спектрами. В большинстве случаях причину появления выбросов удается определить довольно быстро. Выявив причину, необходимо расширить набор образцов новыми подходящими образцами. Здесь следует заметить, что бессмысленно продолжительное время добавлять новые спектры, не разобравшись предварительно в причине возникновения ошибок. Существует опасность того, что в этих образцах не окажется необходимой для стабильного метода информации. Очень большие наборы данных далеко не всегда репрезентативны.

Следующая проблема, с которой можно столкнуться – невозможность воссоздать в лаборатории те же условия, что и при разработке метода. Кроме того, часто оказывается невозможным выбрать образец и проанализировать его с помощью реперного метода, другими словами, собрать репрезентативный калибровочный набор. Поэтому полезно сначала подобрать образцы, а затем делать выводы относительно соответствующих значений, основываясь на рутинном анализе. Если и это невозможно, тогда метод сильно ограничивается и может работать только как индикатор процесса. Но иногда должно использоваться построения метода в лаборатории, что дает неточные результаты, но с другой стороны является зеркалом относительного продвижения реакции по верному пути.

Суммируя все выше сказанное, задача квалифицированного аналитика – определить удовлетворяет ли метод требованиям конкретного процесса.

6. Практический пример

В предыдущей главе описаны все существенные параметры системы, а также даны советы по созданию оптимального PLS-алгоритма. В данной главе на практическом примере будут продемонстрированы наиболее эффективные средства создания калибровочной модели. Кроме того, пользователю будет показано действие ошибочного метода проверки, приводящее к получению неправильной калибровочной модели. Также будут указаны критерии, которые необходимы для распознавания таких моделей.

6.1. Разработка и проверка метода

Для PLS-калибровки важно правильно выбрать диапазоны частот, количество факторов и грамотно провести предварительную обработку данных. Количество факторов зависит от того, сколько параметров в системе. Это количество невозможно рассчитать с помощью теоретических методов. Поэтому их вычисляют эмпирически следующим образом:

Измеряют большое число тестовых образцов. В зависимости от поставленной задачи набор калибровочных данных колеблется от 20 до 100 образцов. Чем сложнее состав образца, тем больше спектров требуется для калибровки. Так, двухкомпонентные системы калибруются на основании всего нескольких спектров. При исследовании сложных систем, таких как органические смеси, или при изучении физических параметров нефтехимических продуктов, например, требуются гораздо более серьезные усилия.

При разработке модели калибровочные спектры и их соответствующие реперные значения даются в программном обеспечении PLS. После определения нужных частотных диапазонов и предварительной

обработки данных проводят калибровку. Затем проверяют качество калибровки (см. Главу 4). На усмотрение пользователя проводят либо внутреннюю, либо внешнюю проверку. Часто предпочтительна внутренняя проверка, так как для калибровки и последующей проверки берут все спектры. Не происходит никакой потери данных при определении внешнего набора данных (см. Главу 3).

Качество калибровки легко оценить, рассчитав коэффициент корреляции R^2 и величину погрешности (RMSECV или RMSEP). Примером может служить анализ смеси метанола CH_3OH, этанола C_2H_5OH и пропанола C_3H_7OH. Измерения проводились на спектрометре фирмы «BRUKER» MATRIX-F в спектральном диапазоне от 10,000 $см^{-1}$ до 4, 000 $см^{-1}$ (оптическая длина пути кювет: 2 мм, спектральное разрешение 8 $см^{-1}$,подсоединение с помощью оптоволокна длиной 50 м). В общей сложности измерили 30 смесей, в диапазоне концентраций от 0 до 100%. На рис. 6.1 приведена выборка соответствующих спектров.

Уже на примере трехкомпонентной смеси налицо сильное перекрывание сигналов аналита. Наблюдаются следующие четыре основных группы сигналов: комбинационные колебания группы СОН при 4800 $см^{-1}$, первые обертона групп CH_2 и CH_3 (6000 $см^{-1}$ - 5500 $см^{-1}$), группы СОН (7300 $см^{-1}$ - 6000 $см^{-1}$) и вторые обертона групп CH_2 и CH_3 (8800 $см^{-1}$ - 7800 $см^{-1}$). Выше 9000 $см^{-1}$ сигнала не наблюдается, ниже 4400 $см^{-1}$ наблюдается область большого спектрального шума, что можно объяснить сильными потерями света в оптоволокне.

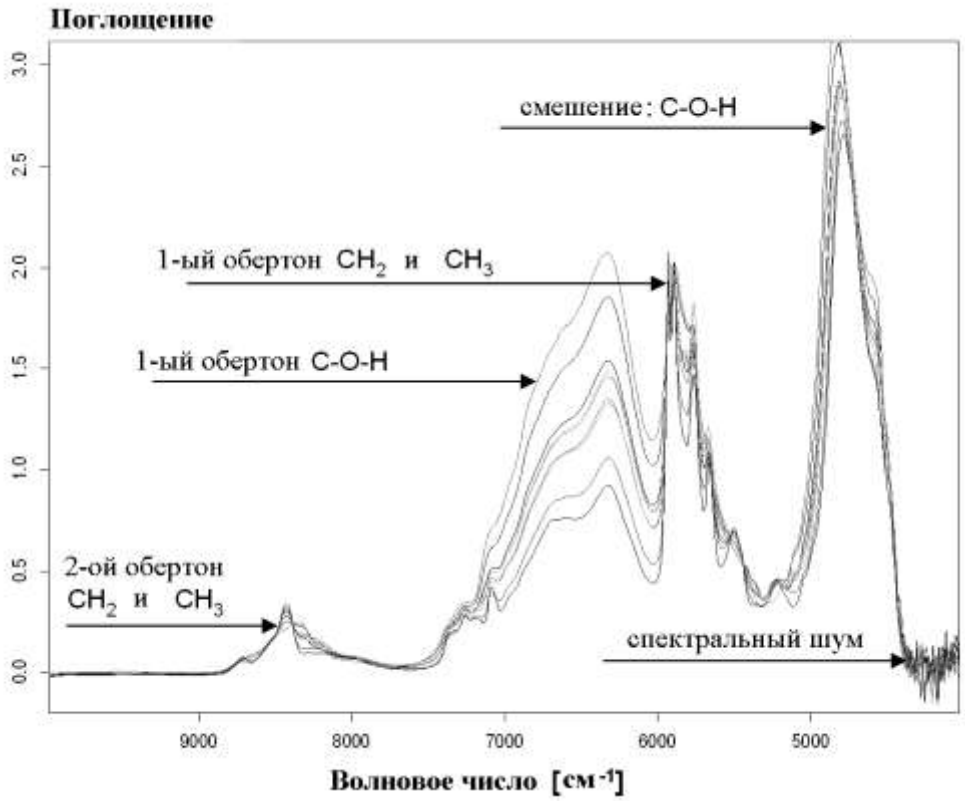


Рис. 6.1 БИК спектры смесей метанола, этанола и пропанола (кювета соединена с оптоволокном длиной 50 м, длина оптического пути 2 мм)

Затем проводят PLS-калибровку системы. Рекомендуется проводить первую калибровку по всему диапазону частот. Однако следует выбирать места с низкими шумами. Области на границах окон прозрачности или те, где полосы поглощения больше 2,5, обычно характеризуются очень низкой интенсивностью света. Следовательно, в результате сигнал будет содержать очень много шума и не может быть использован для исследований. Значит, сигналы ниже 4400 $см^{-1}$ и в районе 4800 $см^{-1}$ не рассматриваются. Далее принимаем во внимание, что выше 9000 $см^{-1}$ не наблюдается существенного поглощения, первым окном для калибровки будет диапазон между 9000 $см^{-1}$ и 5200 $см^{-1}$.

Если проводить проверку с увеличением числа факторов, обычно сначала результаты анализа улучшаются. Чем выше выбранный фактор и чем больше привлекается для анализа спектральной информации, тем лучше результаты. Однако улучшение результатов наблюдается до определенного момента. Начиная с такого «критического» числа факторов в модель добавляется все больше спектрального шума и качество результатов изменяется.

Данный процесс представлен на рис. 6.2. Средние погрешности прогноза для PLS-регрессии концентрации метанола в 30 смесях CH_3OH, C_2H_5OH и C_3H_7OH представлены в зависимости от количества факторов. Оценку метода проводили путем внутренней проверки. Сначала с увеличением числа факторов улучшаются результаты анализа. Начиная с семи факторов, результаты ухудшаются, т.к. модель «перенасыщается». Следовательно, для того, чтобы получить оптимальные результаты анализа, лучше всего остановиться на шести факторах. Средняя погрешность прогноза 0,07%. Это вполне реальное значение анализа на спектрометре ближнего ИК-диапазона для чистых жидких многокомпонентных смесей.

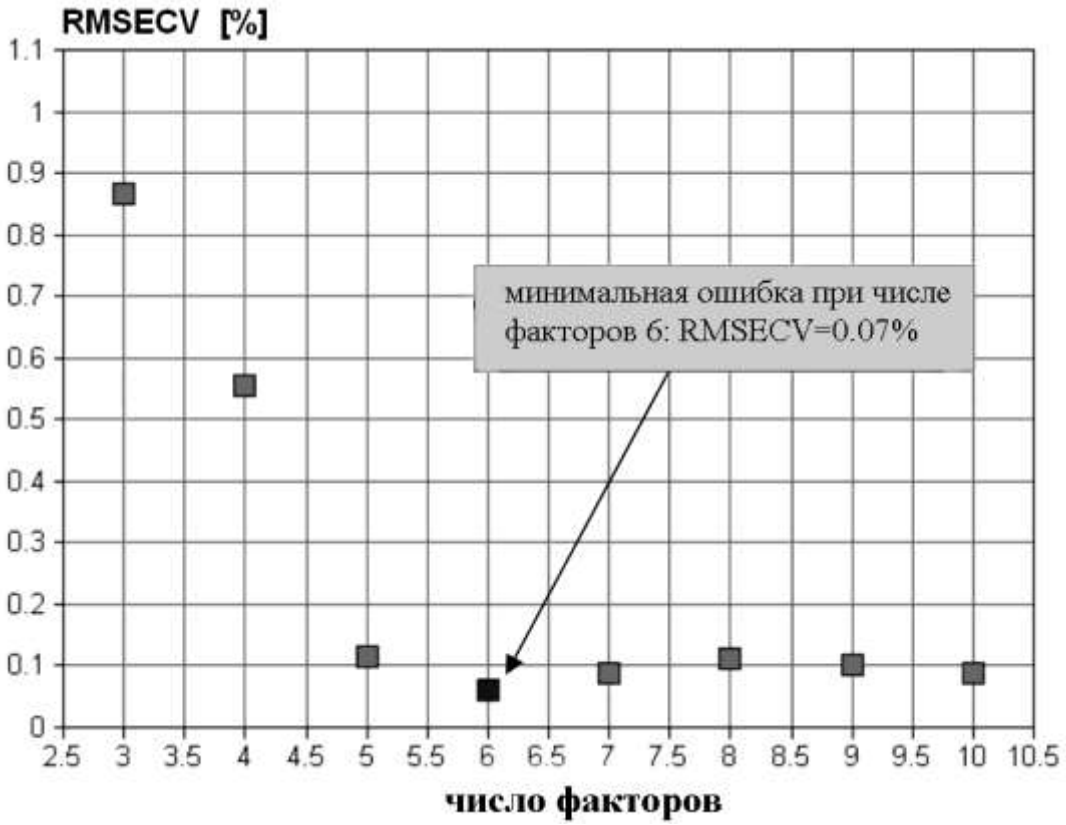


Рис. 6.2 Ошибка прогноза метанола в зависимости от числа факторов PLS-регрессии. Анализируемая смесь метанол/этанол/пропанол.

На рис. 6.3 показаны результаты сравнения значений анализа и соответствующих реперных данных для шестифакторной модели. В целом, эти значения соответствуют друг другу. В этом случае использовали *независимые* тестовые спектры, т.е. исследуемый спектр не содержался в калибровочном тестовом наборе. Следовательно, примерно таких же хороших результатов можно ожидать в будущем при анализе спиртовых смесей.

Таким образом, легко найти оптимальное число факторов для определенной заданной калибровочной модели. Следовательно, остается только подобрать наиболее подходящий метод для решения данной задачи, а именно способ предварительной подготовки данных и спектральный

диапазон. Так как общий ответ на эти вопросы дать невозможно, действовать в каждом конкретном случае приходится только методом проб и ошибок.

Для этого, значения систематически изменяют и рассчитывают индивидуально для увеличивающегося количества факторов. Самые большие значения коэффициента R^2 и/или минимальные ошибки прогноза и характеризуют лучшую модель. Следовательно, все значимые варианты в диапазонах частот и методы предварительной обработки данных тестируются последовательно до тех пор, пока не будет найдена оптимальная модель. Для того, чтобы найти подходящие диапазоны частот, зачастую достаточно сгруппировать подходящие для анализа спектральные диапазоны (см. «Выбор спектрального диапазона» в Главе 5). Как правило, нет необходимости находить отдельные точки спектра.

Глубокие познания в математике не нужны для того, чтобы подобрать методы предварительной обработки данных или выбрать диапазоны частот. При измерении жидкостей хорошим значением R^2 считается цифра более 99%, а для твердых веществ - более 90%. Если эти значения заметно отличаются от приведенных, модель не будет качественной.

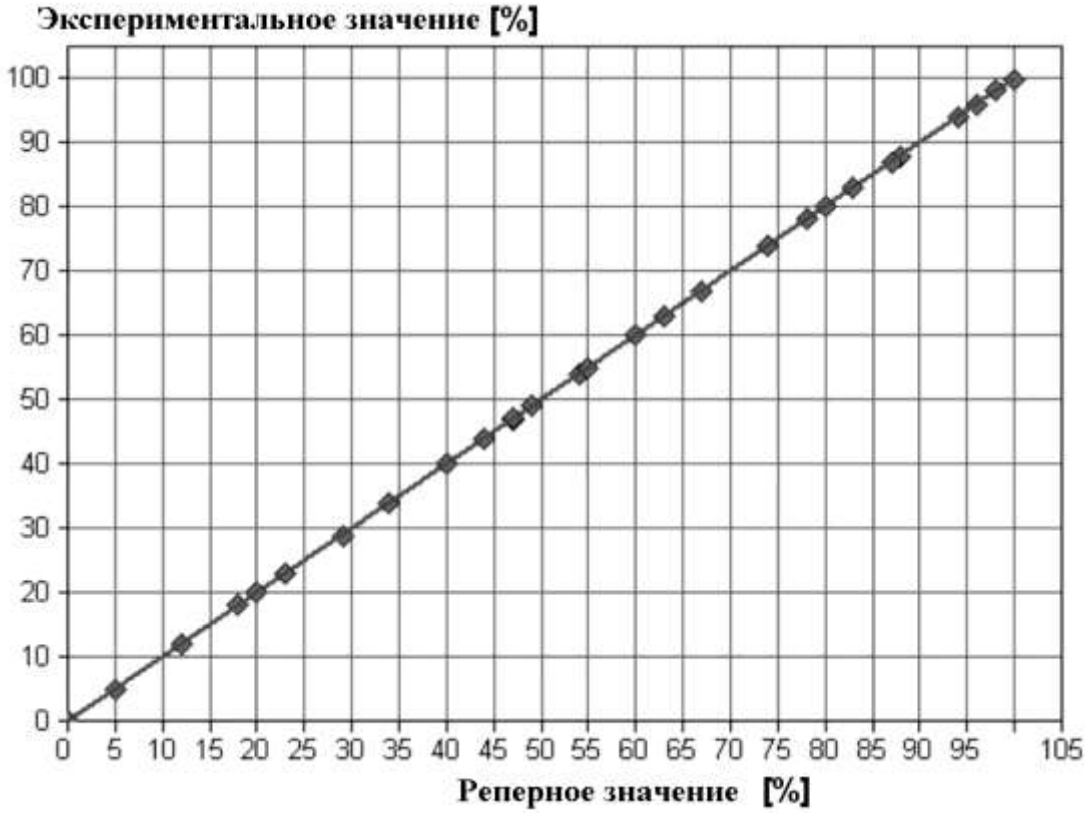


Рис. 6.3 Сравнение реперных и экспериментальных значений для PLS-регрессии определения концентрации метанола в смеси метанол/этанол/пропанол; RMSECVmin = 0,07% (частотный диапазон: 9000 см$^{-1}$ – 5200 см$^{-1}$, число факторов: 6; подготовка данных перед анализом методом «вычитания прямой линии»)

Для того чтобы сравнить модели, рекомендуется свести наиболее важные параметры в таблице, как приведено ниже. Для наглядности здесь показаны результаты лишь пяти проверок, в реальности же один за одним приводятся данные 30 и более проверок.

Таблица 6.1 Метод оптимизации анализа методом ближней инфракрасной спектроскопии для определения концентрации метанола в смеси метанол/этанол/пропанол.

№	Предвари-тельная обработка данных	Частотный диапазон [$см^{-1}$]	Оптимальное число факторов	Коэффициент смешанной корреляции R^2[%]	Значение погрешности прогноза	Примечания
1	Нет	9,000-5,200	9	99.8	0.16%	Общий спектр
2	Вычитание прямой линии	9,000-5,200	6	>99.9	0.07%	Общий спектр
3	Нормализация вектора	9,000-5,200	8	99.6	0.42%	Общий спектр
4	Вычитание прямой линии	7,000-5,200	8	>99.9	0.07%	1 обертоны
5	Вычитание прямой линии	6,000-5,200	7	>99.9	0.07%	нет ОН
...	...	...	...	...	...	...

В первых трех строках приведены калибровки по всему спектральному диапазону 9000 $см^{-1}$ – 5200 $см^{-1}$ для разных способов предварительной подготовки данных. Как видно из таблицы, ВПЛ (вычитание прямой линии) лучше всего подходит в качестве метода предварительной обработки данных (Метод № 2 в таблице 6.1). Коэффициент R^2 больше, погрешности прогноза меньше, чем при обработке Методами №1 и №3. Если взглянуть дальше на диапазоны частот (методы № 4 и №5), становится очевидным, что сделать модель еще лучше не представляется возможным.

Если пренебречь вторым обертоном колебаний групп CH_2- и CH_3- между 8800 $см^{-1}$ и 7800 $см^{-1}$, это не повлияет на результаты анализа. Также

не повлияет на результат пренебрежение сильным поглощением ОН-группы приблизительно при 6900 $см^{-1}$. Средняя погрешность нескольких моделей равна 0,07%. Следовательно, на первый взгляд все три модели в равной степени подходят для определения концентрации метанола. Тем не менее, рекомендуется брать модель с наименьшим количеством факторов. Чем меньше число факторов, с которыми работает метод, тем стабильнее модель. Т.е., в данном примере нужно проводить калибровку в спектральном диапазоне между 9900 $см^{-1}$ и 5200 $см^{-1}$ и использовать ВПЛ для предварительной подготовки данных (Метод №2).

На основе этих данных можно составить метод анализа тестовых образцов. Наиболее важные результаты всегда следует записывать. На рис. 6.4 представлен образец отчета по описанному примеру.

Validation Report

Method developer:	Oliver Hardy
Last change of method:	05.10.2002
Instrument:	MATRIX-F, serial no. 101, Bruker Optik GmbH
Software:	OPUS QUANT, version 4.2 (13.03.2003)
Method file:	Alcohol.q2
Standards (total):	30
Calibration spectra:	30
Data block:	Absorbance spectra
Components (total):	3
Frequency ranges:	1
Number of selected data points:	1143
Data preprocessing:	Subtraction of a straight line
Frequency range:	9,000 – 5,200 cm^{-1}
Spectral resolution:	8 cm^{-1}
Name of measurement experiment:	NIR_Alcohols.xpm

Summarization for Methanol:

Conzentration range:	0 - 100%
Type of validation:	Cross Validation
No. of samples leaving out:	1
Optimum rank:	6
Coefficient of determination:	99,99%
Mean error of prediction RMSECV:	0,07%

Rank	R^2	RMSECV	Recom. Rank
1	82.56	12,00	
2	99.68	2,45	
3	99.92	0,88	
4	99.95	0,56	
5	99.98	0,12	
6	99.99	0,07	*
7	99.99	0,09	
8	99.98	0,11	
9	99.98	0,10	
10	99.99	0,09	

Results of cross validation for rank 6:

file name	reference value	analysis	difference	possible outliers
05ALK1.1	0	-0.307	0.307	
05ALK2.1	100	99.560	0.440	*
05ALK3.1	20.001	20.153	-0.152	
05ALK4.1	33.364	33.262	0.102	
05ALK5.1	49.666	49.750	-0.084	
05ALK6.1	24.942	24.912	0.030	
...				...
05ALK30.1	85.501	85.490	0.011	

Place, Date:

_____ _____

Signature (Method Developer) Signature (Release)

Рис. 6.4 Отчет о проверке

Может показаться странным, что в приведенном примере несколько разных аналитических моделей дают одинаковые результаты. Равнозначность нескольких хемометрических моделей объясняется, исходя из факторизации спектров. В известной степени индивидуальные факторы представляют «единицы информации», которые в свою очередь отражают определенные свойства (и/или совокупность свойств) образца. Таким свойством системы является, например, концентрация аналита. При успешной факторизации PLS-алгоритм распознает существенные для анализа факторы и связывает их с соответствующими свойствами системы (т.е. концентрацией аналита). Как правило, так получается при большом количестве спектральных диапазонов, так как при этом большинство веществ дают доступные для анализа сигналы в более чем одном частотном диапазоне. Поскольку каждый из этих диапазонов состоит из множества точек (т.е. содержит много аналитической информации), статистически систему можно определить для всех этих спектральных диапазонов. Следовательно, в большинстве случаев доступен выбор калибровочных моделей сравнимого качества, которые дают одинаково удовлетворительные результаты анализа.

Следующий важный момент вытекает из факторизации спектров. В случае однопараметрической калибровки при анализе многокомпонентных смесей требуется отделять отдельные сигналы аналита. Каждый компонент приписан определенной длине волны в определенной области[2]. При многопараметрической калибровке такой необходимости нет. Она позволяет оценить несколько компонентов на основании одних и тех же спектральных структур и методов предварительной обработки данных. Поскольку при факторизации спектры разбиваются на независимые единицы информации, нет необходимости разделять их вручную. Существенное преимущество по

сравнению с классической однопараметрической калибровкой видно в случае, когда сигналы сильно перекрываются.

6.2. Анализ и определение выбросов

Для того чтобы проанализировать новые тестовые образцы, необходимо провести спектральные измерения и проанализировать на основании предварительно разработанных и оптимизированных методов. После этого рассчитываются расстояние Махаланобиса или спектральный остаток. Эти значения нужны для определения выбросов.

Распознавать выбросы очень важно для оценки результатов. Наличие выбросов вполне возможно, если анализируемое вещество имеет загрязнения или было некорректно измерено. Выбросы легко распознаются по возрастанию расстояния Махаланобиса или по спектральному остатку. Чем хуже зависимость между тестируемым спектром и концентрациями, тем выше соответствующие значения. Если расстояние Махаланобиса больше, чем его соответствующая пороговая величина, образец квалифицируется как выброс.

Расстояние Махаланобиса и спектральный остаток являются количественными характеристиками, показывающими качество анализа. Если значения оказываются ниже пороговой величины, данные анализа можно считать надежными. Следовательно, образец, не подходящий для измерений отслеживается программным обеспечением и исследователь получает информацию об этом.

Результаты анализа записывают в отчете, примером которого может служить рис. 6.5.

Analysis Report

Operator:	Stan Laurel
Date:	10.09.2003
Instrument:	MATRIX-F, serial no. 101, Bruker Optik GmbH
Software:	OPUS QUANT, version 4.2 (13.3.2003), Bruker Optik GmbH
Method file:	Alcohol.q2
Method developer:	Oliver Hardy
Last change of method:	05.10.2002
Product group:	Alcohol
Measurement method:	NIR
Name of measurement experiment:	NIR_Alkocols.xpm

No.	File name	Path:	Component	Analysis	Mahalanobis-Distance	Limit for Mahal.-Distance	Outliers
1	Gin.1	D:\Alcohol	Methanol	-0.026379%	0.434	0.201	*
2	Gin.2	D:\ Alcohol	Methanol	70.331%	0.031	0.201	
3	Gin.3	D:\ Alcohol	Methanol	100.01%	0.530	0.201	*
4	Gin.4	D:\ Alcohol	Methanol	24.919%	0.167	0.201	
5	Gin.5	D:\ Alcohol	Methanol	50.007%	0.197	0.201	
6	Rum.1	D:\ Alcohol	Methanol	66.687%	0.091	0.201	
7	Rum.2	D:\ Alcohol	Methanol	0.01794%	0.148	0.201	
8	Rum.3	D:\ Alcohol	Methanol	75.112%	0.173	0.201	
9	Rum.4	D:\ Alcohol	Methanol	25.375%	0.188	0.201	
10	Rum.5	D:\ Alcohol	Methanol	33.403%	0.080	0.201	
11	Whiskey.1	D:\ Alcohol	Methanol	43.964%	0.089	0.201	
12	Whiskey.2	D:\ Alcohol	Methanol	13.755%	0.161	0.201	
13	Whiskey.3	D:\ Alcohol	Methanol	36.508%	0.016	0.201	
14	Whiskey.4	D:\ Alcohol	Methanol	26.603%	0.011	0.201	
15	Whiskey.5	D:\ Alcohol	Methanol	41.486%	0.058	0.201	

Place, Date:

_____ Signature (Operator)

_____ Signature (Release)

Рис. 6.5 Отчет об анализе

6.3. PLS-регрессия: обеспечение безграничной точности

Оптимизация хемометрических моделей и анализ тестового образца описаны в первых двух разделах данной главы. В данной главе авторы сосредоточатся на обсуждении возможных источников погрешностей. Важнейшими объектами здесь являются выбор тестовых спектров для оценки модели и проверки метода.

Для проверки PLS-модели совершенно необходим репрезентативный набор тестовых образцов. Они должны покрывать весь диапазон концентраций для калибровки и отражать всевозможные вариации образца. Кроме того, следует принимать во внимание все потенциальные изменения

в окружающей обстановке, такие как, например, флуктуации температур или проникновение влаги в аппаратуру. Только в этом случае можно получить надежные данные относительно ожидаемых погрешностей.

Исходя из вышесказанного, непозволительно выбирать тестовые образцы, уже включенные в калибровочный набор[10]. Если для создания набора данных каждый образец измеряли несколько раз, *все* результаты измерений этого образца включают либо в тестовый набор, либо в калибровочный. В случае внутренней проверки, все спектры образца «принадлежат тестируемому спектру».

Выбор идентичных данных для тестового и калибровочного наборов – задача первостепенной важности. В этом можно легко убедиться, взглянув на уравнения (2-1) и (2-3). При расчете коэффициента регрессии b спектральные данные и данные концентраций вносят непосредственно в уравнение. Если во время проверки функция b коррелирует со спектральными данными, исходя из которых вычисляли функцию (т.е. тестовый и калибровочный наборы идентичны), этот результат совпадет с рассчитанными ранее данными концентрации (см. уравнение (2-3)). Такая «реконструкция» введенных изначально реперных значений будет тем точнее, чем больше факторов выбрано для калибровки.

Таким образом, при достаточно большом числе факторов можно добиться хорошего соответствия между тестовым образцом и соответствующим спектром в калибровочном наборе, т.е. получится то же значение результатов анализа, которое было заложено в модель в процессе калибровки. При этом спектральный шум не снижает качество результатов, поскольку амплитуды шума тестовых и калибровочных тестов одинаковы. Это называется оценкой с так называемыми «зависимыми образцами», потому что образцы уже «распознаны» моделью.

Очевидно, что результаты, полученные при проверке «независимых образцов», совершенно бесполезны[10]. Продемонстрируем это на следующем примере. Проанализированы 30 спектров описанной ранее смеси, содержащей метанол, этанол и пропанол. На этот раз выбираются не точные концентрации, а случайные значения от 0 до 100 %, совершенно несвязанные с реальными значениями компонентов.

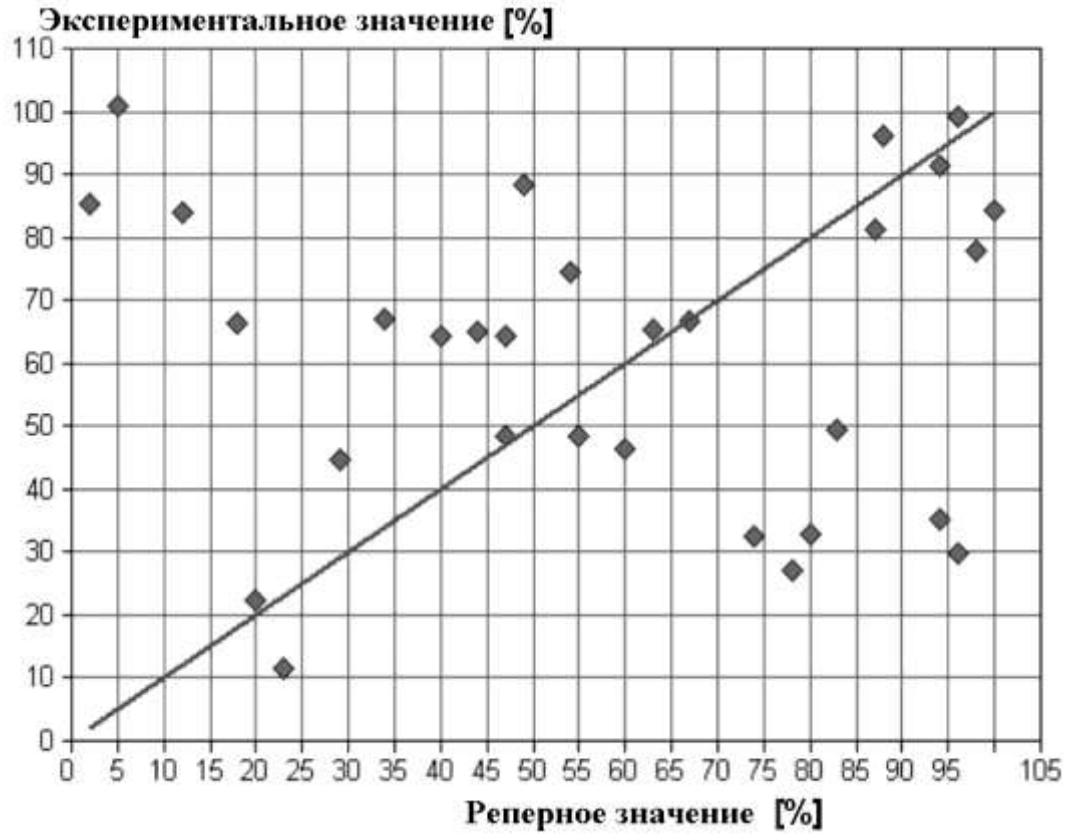


Рис. 6.6 Проверка PLS-регрессии для определения метанола в смеси метанол/этанол/пропанол с независимыми образцами для модели с 13 факторами. Концентрация реперных значений в диапазоне 0 – 100 %. Проверка показывает, что анализ невозможен.

Аналитически корректная проверка данной модели показывает, что она однозначно не подходит для прогноза указанных значений концентраций. Это показано на рис. 6.6. Реперные и аналитические значения никак не связаны, например, PLS-анализ метанола концентрацией 5% определяется моделью как 102% метанола. Другой пример: образец с

содержанием 96% определяется как 29%. Как и ожидалось, анализ невозможен.

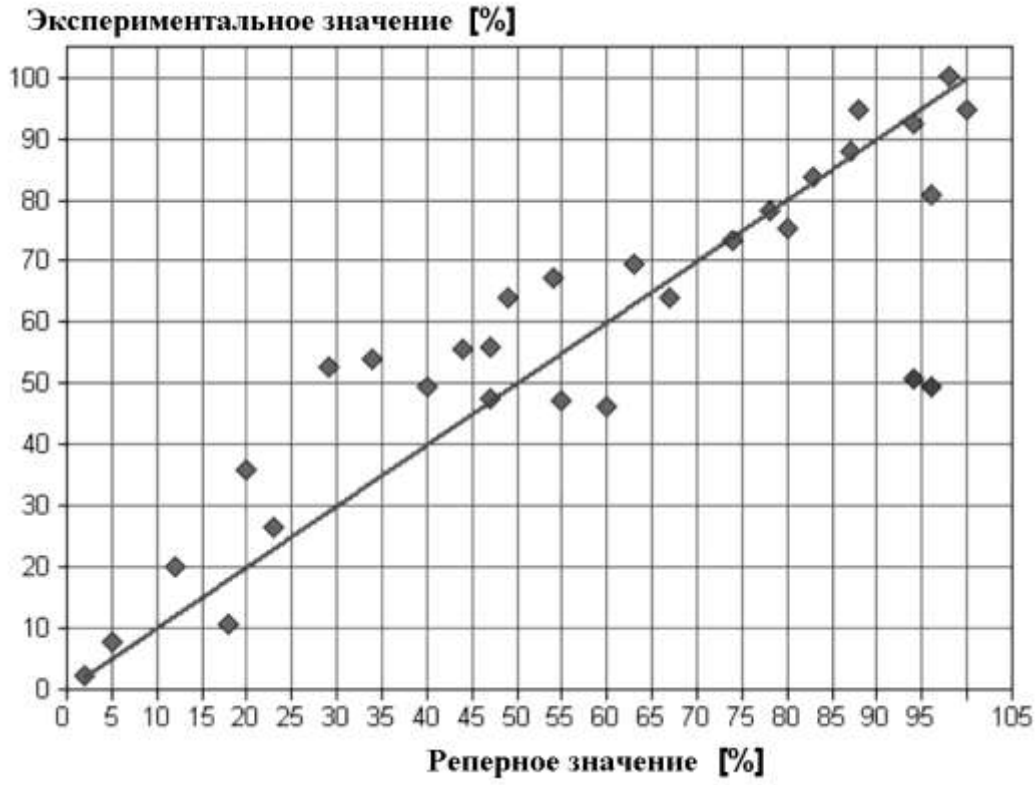


Рис. 6.7 Проверка PLS-регрессии для определения концентрации метанола в смеси метанол/ этанол/пропанол с зависимыми образцами для модели с 7 факторами. Реперные значения концентраций в диапазоне 0 – 100 % (RMSEE = 17,5 %).

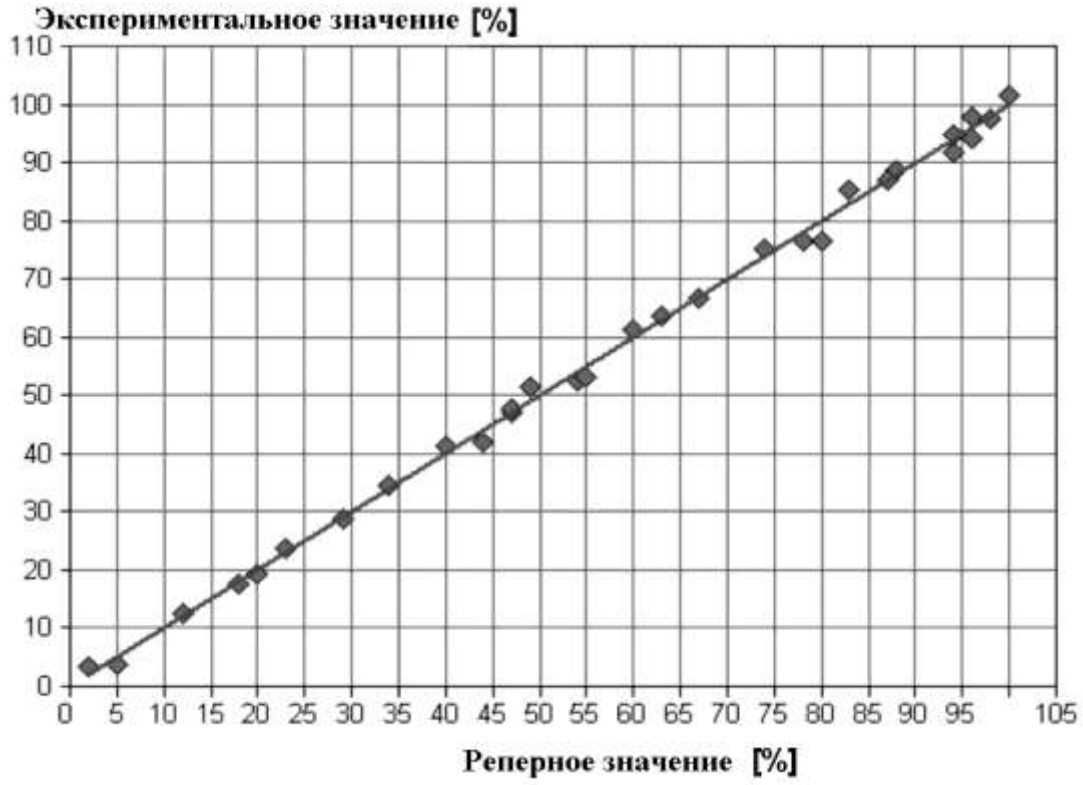


Рис. 6.8 Проверка, аналогичная рис. 6.7 для модели с 13 факторами (RMSEE = 0,42 %).

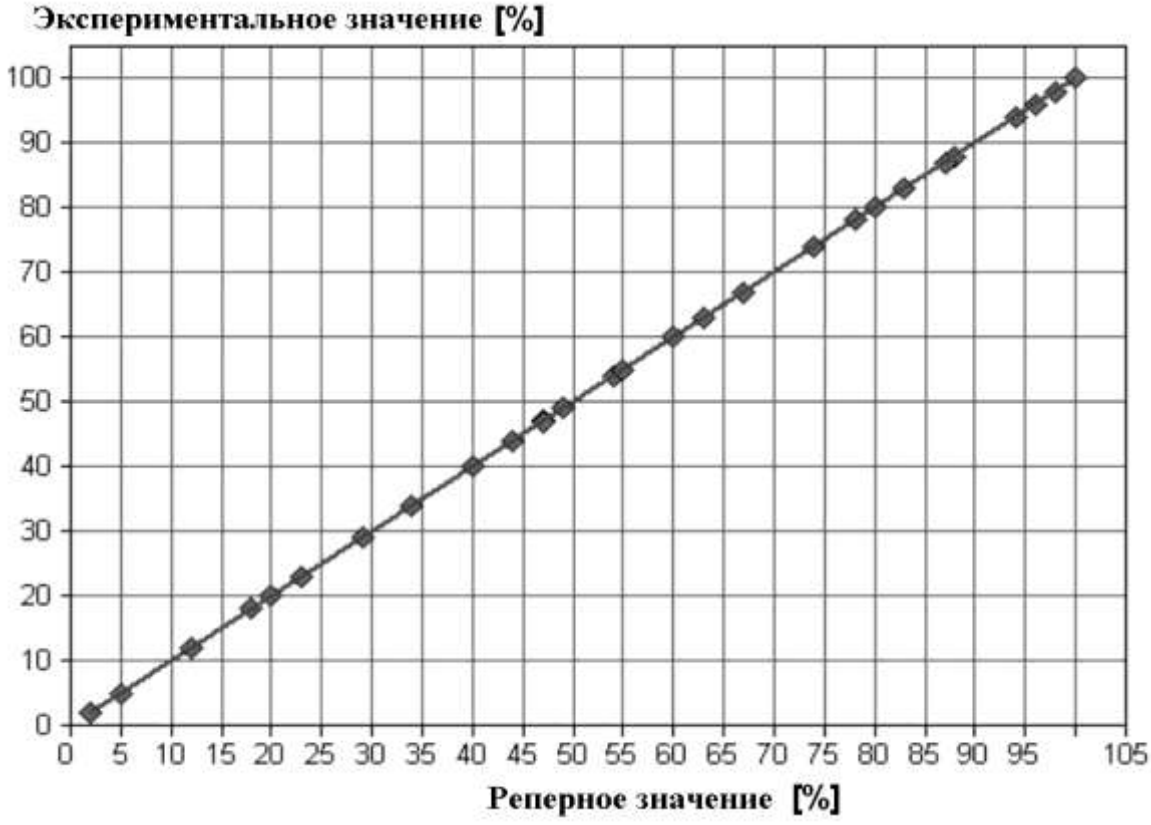


Рис. 6.9 Проверка, аналогичная рис. 6.7 для модели с 16 факторами (RMSEE = 0,04 %).

Ситуация радикально меняется, если в разработке метода используются те же образцы, что и в проверке, т.е. если проверочный набор данных собран из зависимых образцов. Даже для 7-факторной модели наблюдается четкая зависимость между «настоящими» значениями и прогнозируемыми анализом значениями (см. рис. 6.7).

Точность анализа может быть повышена, если выбрать 13 или 16 факторов (рис. 6.8 и 6.9). Соответствующие средние погрешности будут 17,5% для 7-факторной модели, 0,42% - для 13-факторной и 0,04% - для 16-факторной. Следовательно, 16-факторная калибровка показывает лучшие результаты, чем аналитически корректная (см. рис. 6.3).

Из примеров видно, что даже при относительно небольшом числе факторов, значения могут быть воспроизводимыми. То есть, можно получать весьма хорошие результаты на неприемлемых тестовых спектрах. Однако такой метод не выдерживает проверки с реальными образцами.

Результаты проверки легко оценить на практике. С одной стороны, можно проверить характеристические параметры, например, среднюю погрешность анализа. Как уже упоминалось, ошибка прогноза *должна* проходить через оптимальные значения при возрастании числа факторов. Если, значения постепенно улучшаются при возрастании факторов, для проверки были выбраны зависимые тестовые спектры. Это показано на рис. 6.10. С ростом числа факторов, ошибка анализа падает до 0% при числе факторов, равном 6. И напротив, при правильной проверке модели минимальная погрешность, равная 0,07% наблюдается при 6 факторах, и далее не уменьшается (рис. 6.2).

С другой стороны, достоверность модели можно проверить простым измерением образцов. Для этого измеряют и анализируют небольшое количество образцов. Полученная погрешность должна оказаться в тех же пределах, что и определенная ранее погрешность прогноза (RMSECV или RMSEP). Если оказывается, что средняя погрешность проверочного набора данных заметно меньше погрешности измеренных образцов, то для проверки выбраны неадекватные образцы.

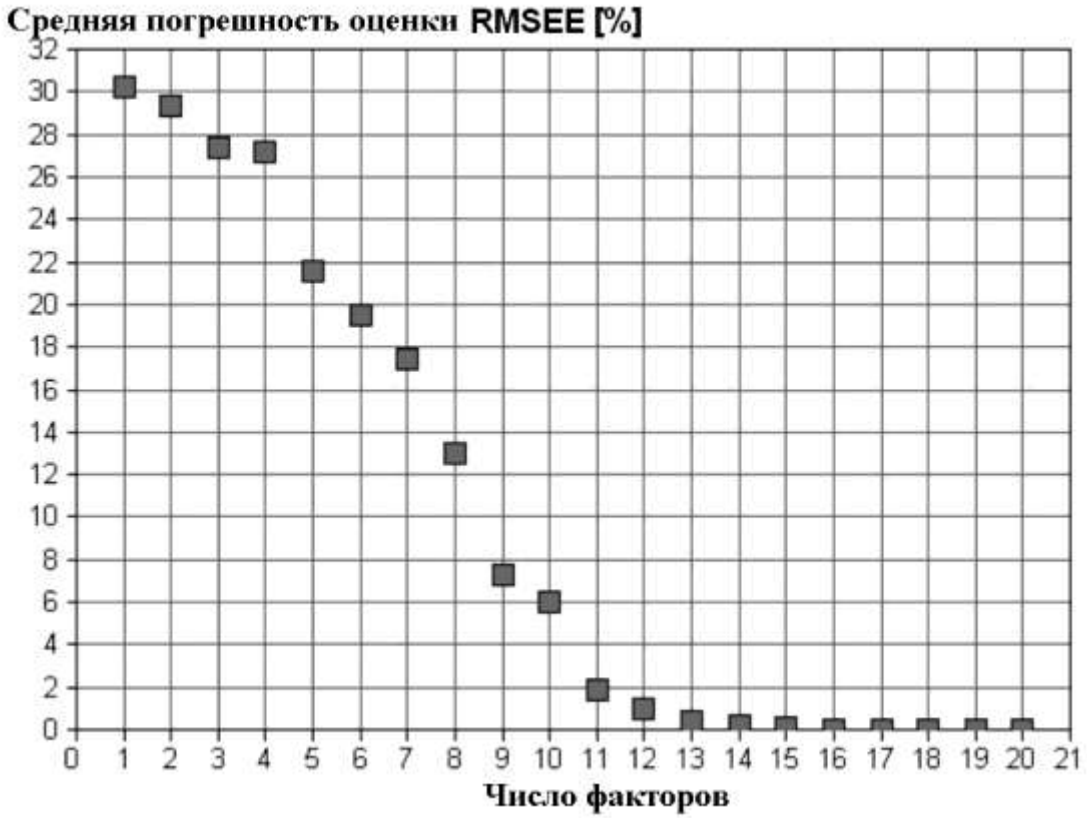


Рис. 6.10 Средняя ошибка определения в зависимости от числа факторов для смеси метанол/этанол/пропанол. Реперные значения концентраций в диапазоне 0 – 100 %. Проверка осуществляется с использованием зависимых тестовых образцов. Пример недопустимой проверки можно распознать по ошибке анализа, которая постепенно уменьшается с увеличением числа факторов.

7. Основные термины

В хемометрической литературе при описании различных методов используется множество статистических параметров. Для проведения и оптимизации количественного анализа, как правило, необязательно знать конкретные формулы. Однако в данной главе приводятся основные термины, часто употребляемые в специальной литературе.

***b-coefficient*:** см. калибровочная функция.

Смещение: ордината линии регрессии

Корректировка смещения и наклона: корректировка смещения приводит к тому, что значения линии регрессии сводятся к первоначальным посредством вычитания смещения (т.е. ординаты) в соответствующих точках. Корректировка наклона приводит значение наклона к единице.

В спектроскопии корректировки смещения и наклона применяют при переносе калибровки с одного прибора на другой. Метод разрабатывается на центральном приборе, а затем используется на других, в дальнейшем их будем называть рабочими приборами. Специфичные системные изменения рабочего прибора рассчитывают заново и приводят к ожидаемым для центрального прибора значениям при помощи соответствующей корректировки. Затем определяют стандарты для рабочего прибора, на которые следует опираться при измерениях. Расчеты на рабочем спектрометре проводят так же, как на центральном. Вкратце их можно описать так:

Спектры, полученные при измерениях на обоих приборах для единичных образцов, отображаются в виде матрицы спектральных данных. ^{M}X – матрица спектральных данных центрального прибора, а ^{S}X - матрица спектральных данных рабочего прибора. На основании известных значений концентраций калибровочных образцов, рассчитывают коэффициент регрессии ^{M}b, который связывает спектральные данные и данные

концентраций для измерений на центральном приборе (ср. с уравнением 2-1)

$$^{M}Y = {}^{M}X * {}^{M}b \qquad (7\text{-}1)$$

Для измерений на рабочем приборе нет никакой необходимости делать дополнительную калибровку. Для того, чтобы рассчитать матрицу данных концентраций для рабочего прибора достаточно применить коэффициент ^{M}b, уже полученный при работе на центральном приборе:

$$^{s}Y = {}^{s}X * {}^{M}b \qquad (7\text{-}2)$$

Если спектры центрального и рабочего прибора идентичны ($^{M}X=^{S}X$), результаты анализа одинаковы ($^{M}Y=^{S}Y$), то есть можно проводить измерения на любом приборе. В спектроскопии комбинационного рассеяния такие измерения дают результаты с высокой точностью.

Если спектры одинаковых образцов, измеренные на центральном и рабочем приборах отличаются, можно попытаться нивелировать при помощи корректировки смещения и наклона. Для прогнозируемых значений рабочих приборов, строятся соответствующие значения рабочих приборов, определенные дополнительным измерением образцов. Если результаты обоих измерений совпадают, регрессия единичных значений Y будет выглядеть как биссектриса. Если спектры центрального и рабочего прибора отличаются, биссектриса будет искажена. Следовательно, необходимо скорректировать расчеты так, чтобы придать регрессии вид биссектрисы. Если простой коррекции, ее называют коррекцией смещения:

$$^{S}Y_{corrected} = (^{S}X * {}^{M}b) - \textit{смещение} \qquad (7\text{-}3)$$

Если же необходимо скорректировать и наклон линии регрессии, коррекцию называют коррекцией смещения/наклона:

$$^{S}Y_{corrected} = ((^{S}X * {}^{M}b)) - \textit{смещение}) * \textit{наклон}^{-1} \qquad (7\text{-}4)$$

Наклон и смещение – способы коррекции, которые позволяют непосредственно рассчитать концентрацию аналита.

Калибровочная функция: В процессе калибровки измеряют некоторое число образцов с известными значениями концентраций. Калибровочная функция b дает возможность скоррелировать свойство системы Y (т.е. концентрацию аналита) с эмпирически полученным значением X (т.е. спектром):

$$b = (X^{T} * X)^{-1} * X^{T} * Y \qquad (7\text{-}5)$$

X и Y записывают в виде матриц. В специальной литературе функцию b часто называют b-коэффициентом или коэффициентом регрессии. Эта функция позволяет рассчитать концентрации аналита непосредственно из спектральных данных (метод описан в главах 2 и 3):

$$Y = X * b \qquad (7\text{-}6)$$

Опытные специалисты часто используют коэффициент b для того, чтобы найти подходящие для разработки метода области спектра. В эти области, содержащие важную информацию об аналите, основной вклад вносит коэффициент регрессии.

Коэффициент определения: см. R^2.

Коррелограмма: Коррелограма отображает степень корреляции между спектральными данными и данными концентраций для данного числа факторов. Значения от -1 до +1 охватывают области с высокой корреляцией. В области, близкой к нулю, корреляция низкая, и эти области нельзя использовать для калибровки:

$$\text{Коррелограмма} = \frac{\sum_{i=1}^{M} (Y_i - Y_m) \cdot (A_i - A_m)}{\sqrt{\sum_{i=1}^{M} (Y_i - Y_m)^2} \cdot \sqrt{\sum_{i=1}^{M} (A_i - A_m)^2}} \quad (7\text{-}7)$$

Разность: Разница между реальной концентрацией образца i и прогнозируемой:

$$\text{Разность}_i = Yi^{\text{измеренное}} - Yi^{\text{прогнозируемое}} \quad (7\text{-}8)$$

Ошибка анализа: см. Ошибка прогноза.

Ошибка прогноза: Качество калибровки определяется ее точностью при анализе тестового образца. В процессе разработки метода точность прогноза проверяется, рассчитываются средние погрешности анализа для всех образцов, RMSECV (см. ниже) для внутренней проверки и RMSEP (см. ниже) – для внешней.

Объяснимая дисперсия: см. R^2.

Фактор: Матрица данных концентраций и матрица спектральных данных разбиваются на пары векторов факторизации (оценки и загрузки) с помощью алгоритма PLS (см. уравнения 2-3 и 2-4). Каждая пара этих векторов называется факторами.

Значения F и FProb: Значения F используют для распознавания выбросов калибровочного набора. Обычно их можно вывести из значений

концентраций и значений спектров измеренного образца. Различают два вида значений F. Одни рассчитывается непосредственно из спектрального остатка, другие – определяют по разнице реальных и прогнозируемых значений (предсказанных с помощью хемометрической модели). Чем больше значение F, тем вероятнее, что речь идет о выбросе.

Расчет значения F для определения спектрильных (7-9) и концентрационных (7-10) выбросов:

$$FValuei = \frac{(M-1)\cdot(SpecRes_i)^2}{\sum_{j\neq i}(SpecRes_j)^2} \qquad (7\text{-}9)$$

$$FValuei = \frac{(M-1)\cdot(Differ_i)^2}{\sum_{j\neq i}(Differ_j)^2} \qquad (7\text{-}10)$$

F-значение концентрационного выброса рассчитывают на основе разности измеренных значений и прогнозируемых значений образца i. M – число образцов в калибровочном наборе данных.

Величина и распределение F-значений единичного компонента зависит от применения. Для того чтобы понять, может ли F-значение оказаться выбросом, это значение следует сравнить со значениями F других образцов из калибровочного набора. Таким образом можно получить так называемое значение FProb. Оно показывает возможное наличие выбросов в распределении всех значений F:

$$FProbi = \frac{\int_0^{FValue} f(FValue)\,d(FValue)}{\int_0^{\infty} f(FValue)\,d(FValue)} \qquad (7\text{-}11)$$

Если значение F равно нулю, значение FProb также равно нулю (т.е. вероятность того, что мы имеем дело с выбросом равна 0%). Бесконечно большое значение F ведет к значению FProb, равному 1 (то есть к вероятности 100%)

Рычаг: Это мера влияния образца на модель PLS. Выражаясь математическим языком, это расстояние Махаланобиса единичных обрацов (см. расстояние Махаланобиса). Значение рычага для выброса будет существенно выше, чем для других образцов.

Расстояние Махаланобиса: В процессе факторизации полученные спектры разбиваются на разные факторы и спектральные остатки (см. уравнение 2-3), причем это относится как к калибровочным образцам, так и к анализируемым. Если для анализа тестового образца используют алгоритм PLS, необходимо проверить, можно ли на основании этой модель исследовать спектры. Знание расстояния Махаланобиса дает возможность проверить, как спектры аналита «подходят» к спектрам калибровочного набора данных.

Расстояние Махаланобиса характеризуется разницей между измеренными спектрами аналитов и средним значением всех спектров калибровочного набора данных, используемых для построения матрицы спектральных данных для заданного числа образцов. Чем больше разница, тем больше значение расстояния Махаланобиса. Объяснений этому может быть несколько. Внешние воздействия, такие как загрязнение образцов или температурные сдвиги ведут к искажению симметрии сдвигов, что в свою очередь приводит к увеличению расстояния Махаланобиса. Также эта величина растет, если в анализ попадают образцы, лежащие вне диапазона концентраций. Расстояние Махаланобиса – количественная мера, показывающая надежность анализа, так как указывает на выбросы и на

образцы с реперными значениями, выходящими за рамки нужных значений концентраций.

При PLS-регрессии для всех спектров расчитывается расстояния Махаланобиса. Из этих результатов определяется наиболее подходящее значение, для которого можно достоверно рассчитать спектр с заданной калибровочной функции. Значения, лежащие за определенным порогом, потенциально могут оказаться выбросами.

Если спектральные данные факторизуются согласно уравнению (2-3), расстояние Махаланобиса h_i определяется следующим образом:

$$h_i = t_i^T (X^T * X)^{-1} * t_i \qquad (7\text{-}12)$$

где расчет производится для R факторов. Если единичные факториальные векторы t_i не были рассчитаны для неизвестного образца, но из калибровочного спектра, они также называются рычагом. Значения рычага калибровочных образцов показывает степень их влияния на PLS-модель.

MDL (Минимальная описываемая длина): Эмпирически полученное число для определения оптимального числа факторов:

$$\text{MLD} = M \ln \frac{SSE}{M} + R \ln M \qquad (7\text{-}13)$$

PRESS (Прогнозируемая остаточная суммарная ошибка квадратов): Сумма всех квадратичных ошибок прогноза. Величина помогает оптимизировать число векторов PLS:

$$\text{PRESS} = \sum_{i=1}^{M} (Y_i^{meas} - Y_i^{pred})^2 \qquad (7\text{-}14)$$

PWS (Весомый спектр): Счетчик уравнения коррелограммы. Большие значения PWS указывают не только спектральные области с низкой корреляцией наборов данных (как коррелограмма), но и с достаточно высокой интенсивностью полос. При разработке калибровочной модели необходимо найти области с большой коррелограммой и с большими значениями PWS:

$$PWS = \sum_{i=1}^{M} (Y_i - Y_m) \cdot (E_i - E_m) \qquad (7\text{-}15)$$

R^2: Коэффициент определения показывает процент отклонения в значениях компонент, полученного при прогнозе. Часто его называют еще *объяснимой дисперсией*. Чем выше коэффициент, тем лучше корреляция между спектральными данными и данными концентраций. При низком R^2, как правило, и результаты анализа плохие. Причинами плохой корреляции могут быть: выбор неподходящих параметров модели (при этом коэффициент определения намного меньше 90%), неточные реперные данные или наличие выбросов в калибровочном наборе.

$$R^2 = \left[1 - \frac{SSE}{\sum_{i=1}^{M} (Y_i - Y_m)^2} \right] \cdot 100 \qquad (7\text{-}16)$$

Коэффициент регрессии: см. Калибровочная функция.

Ранг: Количество факторов для PLS-калибровки.

Остаток: Когда невозможно объяснить расхождения данных концентрации или спектральных данных с помощью факторизации, то, что факторизация не может описать, называют остатком. Соответствующие значения – значениями остатка (см. уравнения 2-3 и 2-4). Остаток – это разница между реальными данные и данными, полученными при факторизации, причем это в равной степени относится как к спектральным данным, так и к концентрационным данным.

Наибольший интерес в хемометрии представляет остаточное значение спектральной остаточной матрицы F (см. уравнение 2-4):

$$F = X - (t_1 p_1^T + t_2 p_2^T + t_3 p_3^T + \ldots + t_R p_R^T) \qquad (7\text{-}17)$$

Чем больше эта величина, тем меньше информации о спектральной структуре можно объяснить факторизацией.

Так же остаточная матрица концентрационных данных G описывает ту часть калибровочного набора данных, который не объясняет факторизация.

$$Resi = Y_i - Y_i^{\text{прогнозируемое}} \qquad (7\text{-}18)$$

Знать значения остатка важно не только при оценке калибровочных моделей. По остаткам спектра аналита можно распознавать выбросы, вычислив разницу между измеренным спектром x_i и спектром s_i, который «ожидался» при факторизации:

$$SpecResi = \sqrt{\sum_{j=1}^{M} (x_{i,j} - s_{i,j})^2} \qquad (7\text{-}19)$$

Сумму необходимо посчитать для всех частотных значений (индекс j). Здесь x_i –измеренные спектры, а s_i – связанные спектры:

$$s_i = t_{1,i} * p_{1,i}^T + t_{2,i} * p_{2,i}^T + t_{3,i} * p_{3,i}^T + ... + t_{R,i} * p_{R,i}^T \quad (7\text{-}20)$$

Чем больше остаток (т.е. чем больше разница между реальным спектром и спектром, ожидаемым после факторизации), тем с большей вероятностью образец окажется выбросом. Таким образом определяют примеси, которые накладываются на спектр и приводят к увеличению спектрального остатка.

Помимо расчета расстояния Махаланобиса, это значение наиболее часто используется для определения выбросов.

RMSECV (Среднеквадратичная погрешность внутренней проверки): Это количественная мера для определения точности, с которой делалась проверка. То же, что RMSEP для внешней проверки.

$$\text{REMSECV} = \sqrt{\frac{1}{M} \cdot \sum_{i=1}^{M} (Y_i^{meas} - Y_i^{pred})^2} = \sqrt{\frac{1}{M} \cdot PRESS} \quad (7\text{-}21)$$

RMSEE (Среднеквадратичная погрешность оценки): Величина помогает рассчитать погрешность при анализе калибровочных образцов. *Внимание*: RMSEE нельзя использовать для проверки модели, так как не проводится анализ независимых тестовых спектров, см. главу 6-C:

$$\sqrt{\frac{SSE}{M - R - 1}}$$

RMSEE = (7-22)

RMSELC (Среднеквадратичная погрешность коррекции рычага): Эту величину рассчитывают во время калибровки и с ее помощью оценивают ожидаемую погрешность анализа для тестового образца.

$$RMSELC = \sqrt{\frac{1}{M} \cdot \sum_{i=1}^{M} \left(\frac{Res_i}{1 - Lever_i} \right)^2} \qquad (7\text{-}23)$$

***RMSEP (Среднеквадратичная погрешность прогноза)*:** Величина, определяющая точность анализа тестового набора образцов. См. RMSECV для внутренней проверки.

$$RMSEP = \sqrt{\frac{1}{M} \cdot \sum_{i=1}^{M} (Y_i^{meas} - Y_i^{pred})^2} \qquad (7\text{-}24)$$

SECV (Стандартная погрешность внутренней проверки): см. RMSECV

SEE (Стандартная погрешность оценки): см. RMSEP

***SEP (Стандартная погрешность прогноза)*:** см. RMSEP

Спектральный остаток: см. Остаток

***Коррекция наклона*:** см. коррекция смещения.

SSE (Суммарная квадратичная погрешность): Сумма квадрата остатков. Чем большей полученное SEE, тем хуже модель объясняет отклонения наборов данных:

$$\sum_{i=1}^{M} (Res_i)^2$$

SSE = (7-25)

Заключение

И так, при разработке метода многопараметрической калибровки используется совершенно другой подход, чем при создании однопараметрической калибровки. В случае последней делается попытка найти сигнал, который можно количественно оценить, и чтобы при этом его не перекрывали сигналы других спектральных структур. При многопараметрической калибровке спектральные участки намеренно комбинируют. При этом перекрывание сигналов не имеет никакого значения. При помощи тестирования всех интересующих –провизора-аналитика комбинаций и проведения последовательной проверки можно найти комбинацию, подходящую для разработки метода. Как правило, не требуется досконально изучать информацию, касающуюся спектров отдельных субстанций, участвующих в анализе.

Очевидна необходимость автоматизировать весь процесс разработки и оптимизации метода. Поскольку положение всех функциональных групп в БИК-спектроскопии и спектроскопии комбинационного рассеяния хорошо известны, можно с помощью соответствующего программного обеспечения автоматизировать тестирование всех возможных параметров.

На рис. 8.1 можно увидеть результат такой автоматизации. Приведены данные для примера, описанного в главе 6.

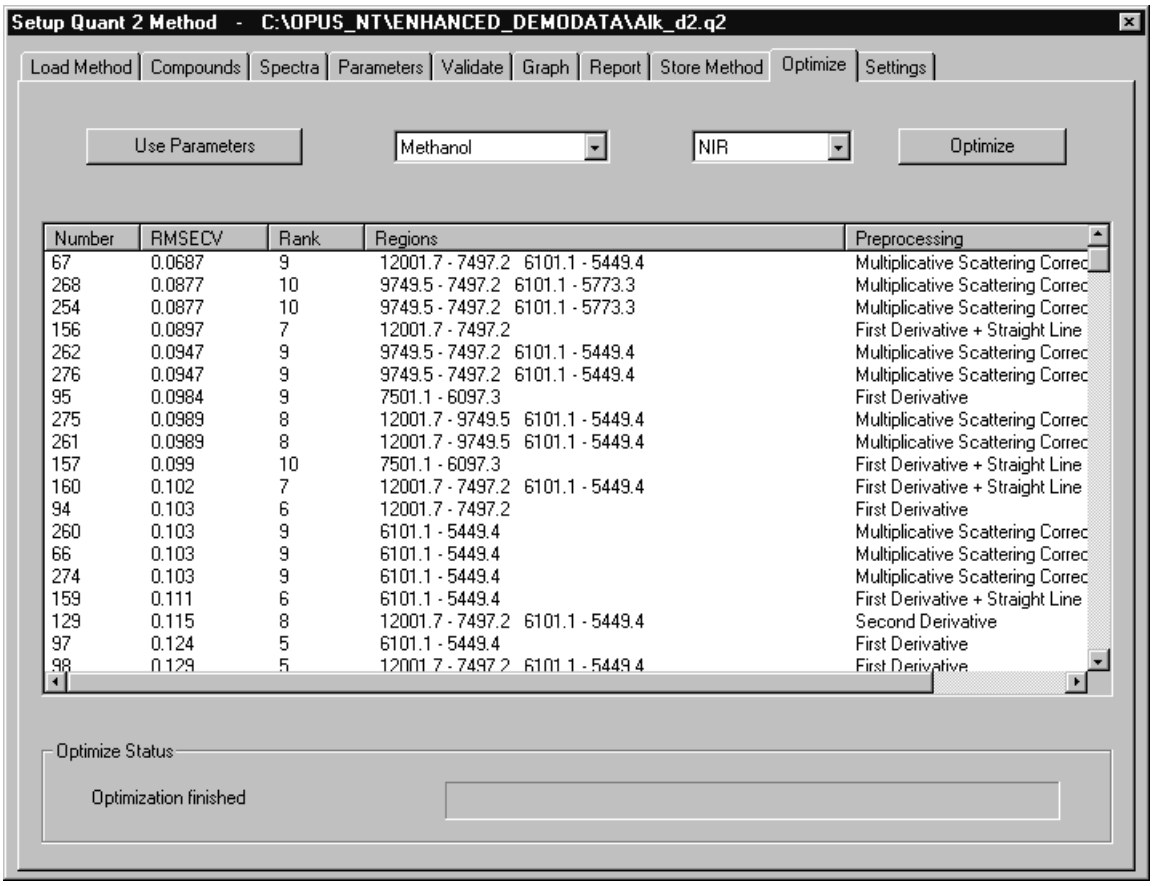


Рис. 8.1 Результат автоматической оптимизации, выполненного при помощи программного обеспечения OPUS/QUANT (Bruker Optik GmbH). На рисунке изображен список калибровочных моделей с наименьшими ошибками анализа.

Получен тот же результат, как и при неавтоматизированном методе. Отображено множество различных моделей, каждая из которых показывает сравнительно неплохой результат анализа. Опыт показывает, что таким данным можно в целом доверять. Разработчики не должны ограничивать себя лишь неавтоматизированной оптимизацией калибровочной модели. Теперь это можно доверить соответствующему программному обеспечению.

Сегодня навыки провизора-аналитика проявляются не столько в том, чтобы оптимизировать метод, сколько для того, чтобы правильно выбрать и

подготовить репрезентативные образцы, учесть возможные влияния окружающей среды на образцы и прибор, точно провести измерения и верно оценить полученные результаты, то есть провести те действия, которые компьютер выполнить не может.