

ОБРАБОТКА ДАННЫХ: КАК УЛУЧШИТЬ ГРАДУИРОВКУ

Ю.Каламбет, к.ф.-м.н., С.Мальцев, к.ф.-м.н., ЗАО "Амперсанд" kalambet@ampersand.ru
Ю.Козьмин, Институт биоорганической химии РАН

Обобщен опыт работы по разработке программного обеспечения для хроматографии и поддержки пользователей. Обсуждаются пути повышения точности результатов химического анализа вообще и хроматографического в частности. Для простоты изложения обсуждение ограничивается примерами на основе линейных и прямо пропорциональных градуировок. Особое внимание уделено выбору модели градуировки и ее проверке.

Есть разные способы повышения точности химического анализа, мы будем говорить о тех, которые связаны с обработкой данных. Речь пойдет об измерениях, в которых применяется градуировочная характеристика (далее – ГХ) – функциональная зависимость аналитического сигнала от содержания аналита, выраженная в виде формулы [1]. Ограничимся случаем, когда определение параметров формулы производится методом наименьших квадратов [2]. Рассмотрим, от каких факторов зависит доверительный интервал решения задачи химического анализа – предсказания содержания аналита по величине аналитического сигнала. Обратим особое внимание на начальный участок ГХ, по которому вычисляются пределы обнаружения и определения.

Предел обнаружения основан на статистической проверке гипотезы о том, что содержание аналита равно нулю. Предел определения соответствует содержанию, для которого аналитический сигнал в три-пять раз больше стандартного отклонения, характерного для данной интенсивности сигнала. Посмотрим, какими средствами можно понизить величину доверительного интервала без ущерба доказательности анализа.

Доверительный интервал (ДИ) формально существует только для определенной измеренной или предсказанной величины. Обобщим

это понятие на ГХ, приписав ДИ каждой точке ГХ (заметьте, что ДИ откладывается вверх и вниз от линии градуировки). Сначала обратим внимание на то, что есть несколько разных по смыслу доверительных интервалов и мы должны их различать (рис.1).

Доверительная область (ДО) градуировочной характеристики (та область, куда с заданным уровнем вероятности попадает ГХ) зависит от того, насколько точно оценены коэффициенты функциональной зависимости, и для линей-

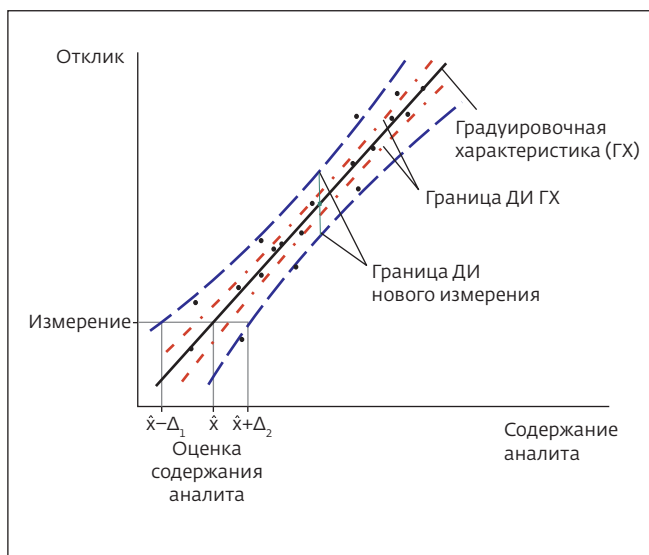


Рис.1. Градуировка и ее доверительные интервалы

ной регрессии описывается формулой (5)*. ДИ нового измерения ограничивает ту область, в которую с заданным уровнем вероятности попадает очередное измерение, по точности не отличающееся от градуировочных. Соответственно, границы доверительного интервала расширяются (формула (6)).

Доверительный интервал предсказания (восстановление содержания аналита по отклику детектора) получается путем пересечения ДО градуировочного графика горизонтальной линией; он не симметричен относительно предсказанного содержания аналита и в большинстве случаев может быть вычислен только с помощью компьютера. Формулы для ДИ предсказания, основанные на делении ДИ нового измерения на коэффициент пропорциональности ГХ (см., например, [5]), приближительные и графически получаются проведением через точки концов ДИ линий, параллельных градуировке (рис.2).

УСРЕДНЯТЬ ИЛИ НЕ УСРЕДНЯТЬ?

Обозначим словом "уровень" (жаргон характерен для программ обработки хроматографических данных) результат усреднения аналитического отклика по всем измерениям (точкам) с одинаковым содержанием аналита. При этом необходимо, чтобы число точек для каждого уровня было одинаковым; обозначим его m .

Градуировку можно построить двумя способами: используя все измеренные точки (рис.3а) или используя уровни (рис.3б). Коэффициенты регрессии, построенной по точкам, будут в точности равны коэффициентам регрессии, построенной по уровням с одинаковым числом точек. Наличие большого числа степеней свободы при градуировке по точкам совсем не означает, что окажется возможным построить аппроксимирующий полином высокой степени. Ранг матрицы, используемой при расчете профиля u , из формулы (1), не превышает числа уровней, и степень полинома не может превышать число уровней минус один, как и в случае градуировки по уровням.

Теперь посмотрим, чем отличаются доверительные интервалы для этих случаев. Обратим внимание на первые два сомножителя формулы доверительного интервала (формула (1)). Нетрудно понять, что их произведение пред-

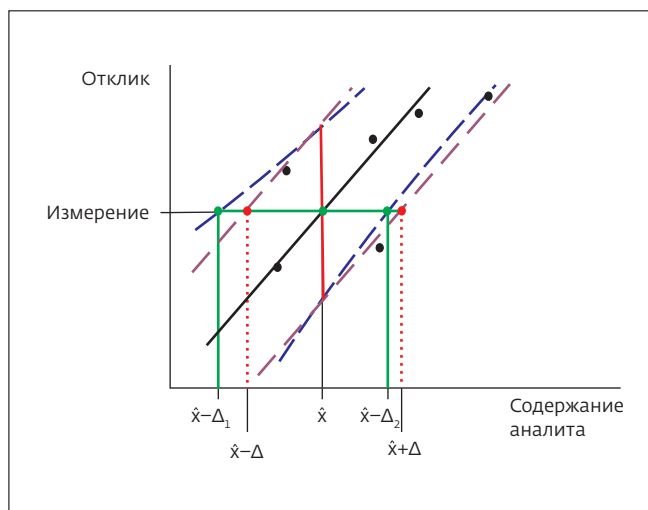


Рис.2. $\hat{x}$ – предсказанное содержание аналита в анализе (предсказание); $\hat{x}-\Delta$, $\hat{x}+\Delta$ – приближенная оценка ДИ; $\hat{x}-\Delta_1$, $\hat{x}+\Delta_2$ – правильная оценка ДИ предсказания

ставляет собой оценку точности любого единичного измерения из тех, по которым строится градуировка. Эта оценка используется при расчете всех вариантов ДИ.

Если мы используем точки, то получим достаточно надежную оценку дисперсии единичного измерения: величина коэффициента Стьюдента t стабилизируется по мере роста числа степеней свободы.

Если мы используем уровни (усредненные точки), то число степеней свободы такой зависимости резко падает, и коэффициент Стьюдента растет. Второй сомножитель – стандартное отклонение точек от прямой – падает. Растет также и третий сомножитель формулы (1) – уровень профиля вероятности $\sqrt{u}$. Доверительный интервал градуировки, построенной по уровням, может быть как больше, так и меньше интервала, полученного по всем точкам, но будет заведомо менее надежным из-за меньшей статистики. При большом числе уровней доверительная область градуировки по уровням совпадет с ДО градуировки по точкам.

ДИ нового измерения, построенный по уровням, позволяет построить ДИ предсказания только для среднего из m измерений, тогда как по градуировке с использованием точек оценку предсказания можно получить как для однократного измерения, так и для любого числа повторений.

Есть случаи, когда использование уровней необходимо, – например, когда дисперсия

* См. врезку

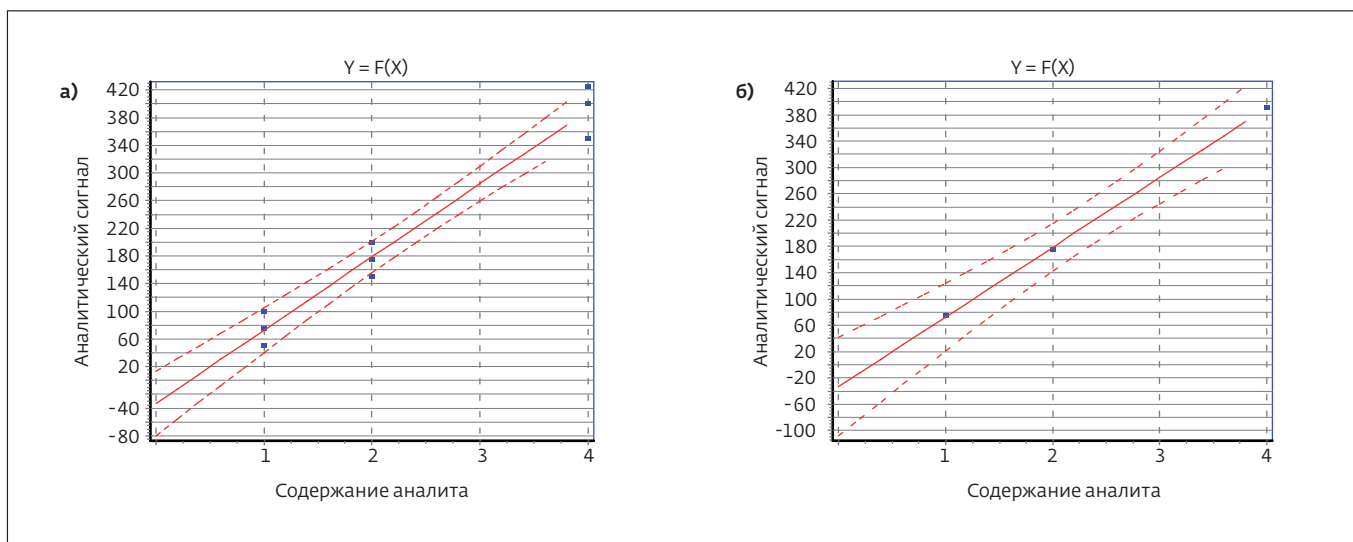


Рис.3. ДО градуировки, построенная а) по точкам; б) по уровням

повторных измерений существенно меньше дисперсии регрессии (определение слова "существенно" в этом контексте зависит от требуемого уровня доверительной вероятности; для уровня 0,95 при нескольких независимых источниках ошибок принято пренебрегать случайными составляющими с дисперсией менее 1/10 от дисперсии основного источника погрешности; по стандартному отклонению это чуть больше трех). Представим ситуацию, когда повторные измерения делаются с помощью дозирования пробы автосамплером из одной пробирки. Одновременно предположим, что случайная составляющая ошибки дозирования

много меньше систематической (рис.4а). Тогда воспроизводимость для любого уровня будет очень хороша, но не будет отражать действительной ошибки измерения (рис.4б). В этом случае повторные измерения уровня просто не имеют смысла, и для построения регрессии можно использовать как среднее, так и любое одно измерение этого уровня.

ВЫБОР МОДЕЛИ

При построении градуировочной зависимости доверительные интервалы чаще всего считаются по формуле (5), которая выполняется для линейной градуировки, не проходя-

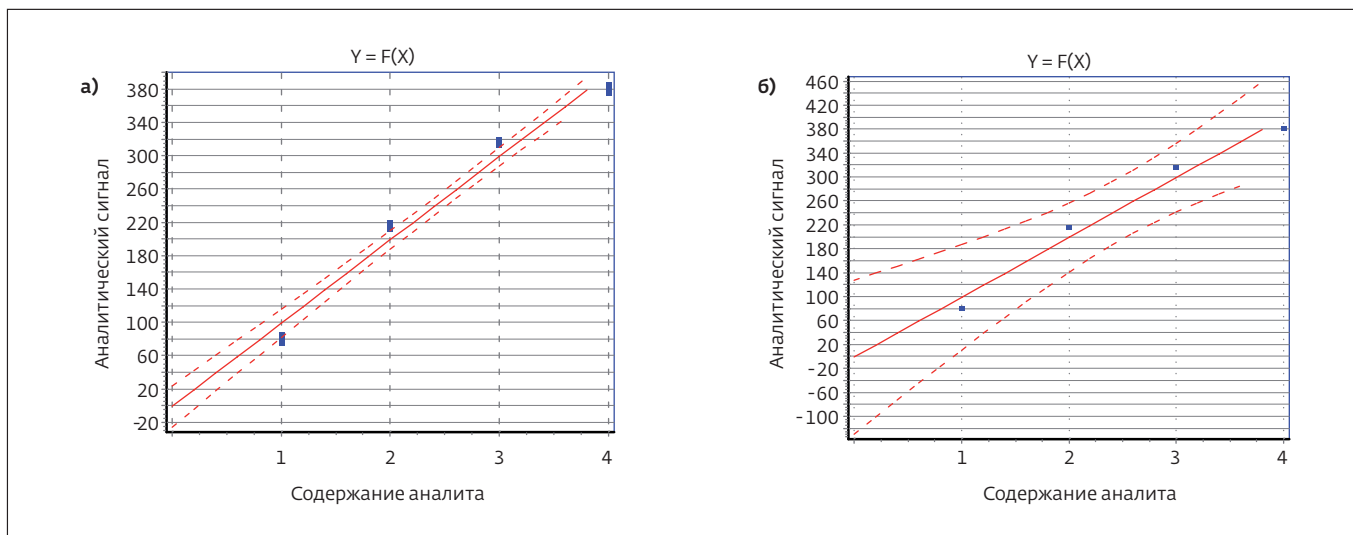


Рис.4. ДИ линейной градуировки, построенный а) по точкам с четырьмя повторными измерениями с малой дисперсией, б) по уровням. Правильный результат оценки ДИ регрессии на рисунке (б)

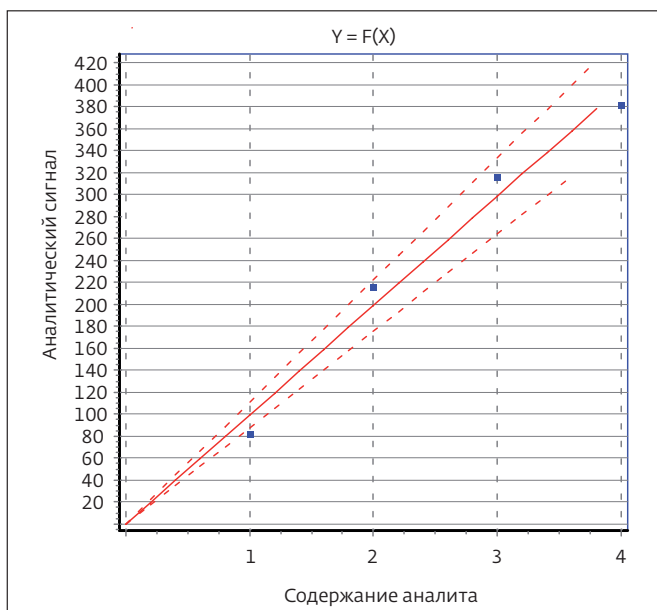


Рис.5. ДИ прямо пропорциональной градуировки.
Границы ДИ градуировки в предположении ее линейности приведены на рис.4б

щей через начало координат. В то же время, если есть априорные сведения, позволяющие ожидать прямо пропорциональной зависимости аналитического сигнала от содержания

аналита, можно использовать модели градуировки без свободного члена. У такой градуировки на один коэффициент меньше, и, соответственно, у решения больше степеней свободы. Формулы доверительных интервалов такой градуировки (7) и (8) также изменяются (рис.5). К сожалению, в нормативной литературе эти формулы не используются [5], к прямо пропорциональной градуировке там обычно применяют формулы, характерные для линейной градуировки.

Вблизи начала координат ДИ прямо пропорциональной ГХ становится существенно меньше ДИ нового измерения, и во многих случаях ДИ градуировки в этой области можно пренебречь. Безусловно, кроме теоретических предположений нужны статистические свидетельства того, что прямо пропорциональная градуировка допустима. Таким свидетельством является, к примеру, попадание точки начала координат в ДО линейной ГХ.

Следует иметь в виду, что близость свободного члена уравнения линейной регрессии к нулю не означает, что градуировка является прямо пропорциональной. Предположение о прямой пропорциональности должно быть под-



25 лет

МультиХром

ЗАО «Амперсенд»

версия 3

- Управление более чем 200 хроматографами и отдельными устройствами различных производителей: Agilent, Gilson, Knauer, Metrohm, Люмэкс, Аквилон, Хроматек, ПортЛаб, Химавтоматика и др.
- Оригинальные методы предельного подавления шумов, адаптивной интерполяции пиков и разделения перекрывающихся пиков по форме.
- Спектральный анализ и многоканальная хроматография: идентификация и расчет количества компонента по спектру пика, разделение перекрывающихся пиков с помощью факторного анализа.

- Поддержка капиллярного электрофореза с дополнительными методами обработки данных.
- Специализированный модуль для гель-проникающей хроматографии.
- Широкие возможности форматирования – создание документов заданного образца; выдача отчетов в форматах PDF, RTF, HTML; статистические отчеты для группы хроматограмм.
- Поддержка требований «Надлежащей лабораторной практики» (GLP, GMP) и электронного документооборота.

support@ampersand.ru

www.multichrom.ru

+7(499)196-52-90

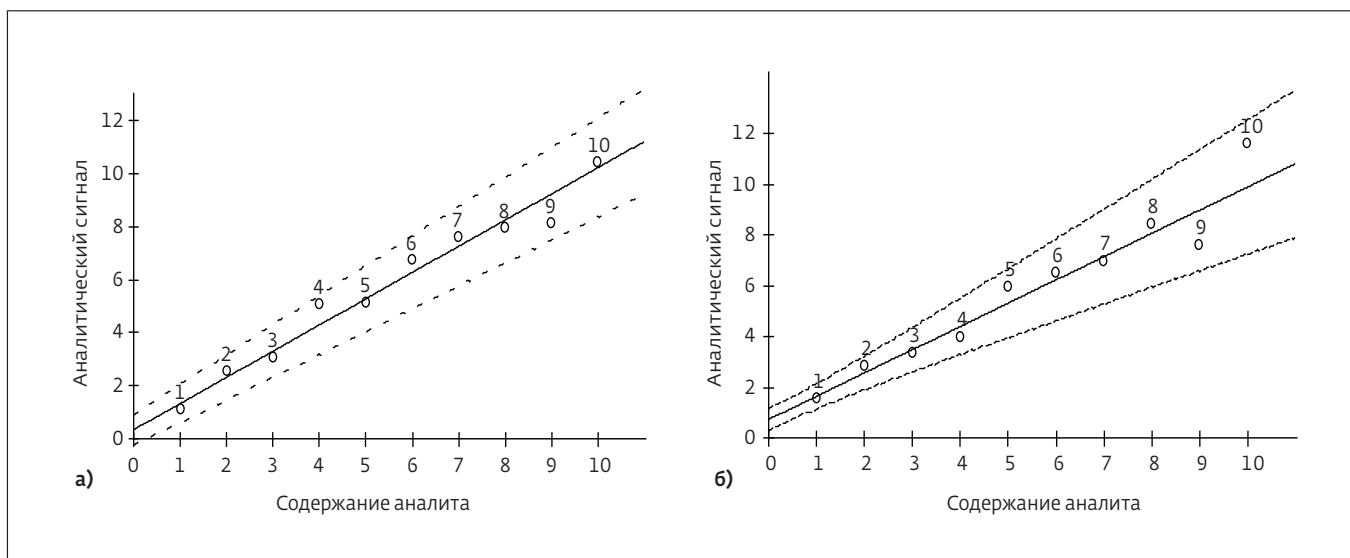


Рис.6. Распределение ДИ нового измерения для случаев, когда ошибка пропорциональна корню из сигнала (а) или сигналу (б)

креплено моделью физического процесса измерения аналитического сигнала.

Если градуировка ведется по точкам и число степеней свободы достаточно велико, решение о значимости коэффициентов градуировки можно принимать на основе очень простого критерия: минимальности доверительного интервала градуировки для всего интервала исследуемых содержаний аналита. Если степень полинома слишком мала, остаточная сумма квадратов возрастает, и, несмотря на увеличение числа степеней свободы, ДИ растет. При избыточной степени аппроксимирующего полинома дисперсия почти не меняется, тогда как число степеней свободы падает и доверительный интервал тоже растет. Минимум доверительного интервала как раз и покажет правильную степень для конкретного случая. При поиске минимума следует иметь в виду значимость отличий ДИ: даже если численно ДИ третьей степени меньше ДИ второй, но разница не превышает ДИ оценки ДИ (здесь формулы не приводим), то гипотезу о том, что ДИ третьей степени меньше ДИ второй, придется отклонить. Такой же алгоритм годится для упрощения формулы (например, принятия решения о значимости свободного члена).

Обратите внимание, что оптимальная аппроксимация может оказаться разной для разных интервалов содержаний аналита. При этом нельзя допускать, чтобы выбор пара-

метров зависимости основывался только на соображении минимума ДИ, модель процесса измерения аналитического сигнала имеет первостепенное значение.

ВЗВЕШЕННАЯ РЕГРЕССИЯ

Метод наименьших квадратов хорошо работает в том случае, когда ожидаемые ошибки всех измерений одинаковы (ошибка равна квадратному корню дисперсии сигнала для данного содержания аналита). Однако на практике это не так. Ошибка измерения часто растет с ростом сигнала. Закон изменения ошибки зависит от конкретной ситуации. В одной из предыдущих статей [7] мы приводили решения для популярных случаев, когда ожидаемая ошибка пропорциональна аналитическому сигналу (постоянная относительная ошибка) и квадратному корню сигнала (ситуация, характерная для детектора радиоактивности или масс-спектрометрического детектора).

В обоих случаях доверительный интервал в начале градуировки заметно сужается (рис.6). Если желательно уменьшить пределы обнаружения и определения, то за этот эффект стоит бороться. В самом деле, оценка дисперсии отдельного измерения, сделанная в предположении, что дисперсии всех точек одинаковы, даст завышенную оценку ДИ в начале градуировки и заниженную – в конце.

Описанные выше варианты взвешивания, при которых ошибка вблизи начала координат

Метод наименьших квадратов и расчет доверительных интервалов

Аппроксимированные значения методом наименьших квадратов [2] вычисляются по формуле

$$\hat{Y} = X \cdot \beta,$$

где

$\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}^T$ – вектор аппроксимированных значений;

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ 1 & x_2 & x_2^2 & \dots & x_2^p \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^p \end{pmatrix} - \text{матрица степеней по оси } x \text{ (временной);}$$

$\beta = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$ – вектор коэффициентов;

$Y = \{y_1, y_2, \dots, y_n\}^T$ – вектор откликов детектора, n – число точек, использованных при построении полиномиальной аппроксимации (ширина окна фильтра);

p – число параметров регрессии (столбцов в матрице X).

Доверительный интервал (ДИ) градуировки Δ_Y вычисляется как

$$\Delta_Y = t_{n-p}^{(1/2)\alpha} \cdot S \cdot \sqrt{u_*}, \quad (1)$$

$S^2 = \text{RSS}/(n - p)$ – оценка дисперсии σ^2 ошибки измерения;

$\text{RSS} = (Y - X \cdot \beta)^T \cdot (Y - X \cdot \beta)$ – остаточная сумма квадратов (Residual Sum of Squares);

$u_* = x_*^T (X^T \cdot X)^{-1} x_*$;

$x_*^T = \{1, x_*, \dots, x_*^p\}$;

x_* – положение, в котором производится аппроксимационная оценка;

$t_{m-\delta}^{\delta}$ – коэффициент Стьюдента для доверительной вероятности $(1-\delta)$ и m степеней свободы.

ДИ измерения

$$\Delta_Y = t_{n-p}^{(1/2)\alpha} \cdot S \cdot \sqrt{1+u_*}. \quad (2)$$

При априорно известном ДИ измерения C_Y формула ДИ градуировки принимает вид

$$\Delta_Y = C_Y \cdot \sqrt{u_*}, \quad (3)$$

и ДИ измерения

$$\Delta_Y = C_Y \cdot \sqrt{1+u_*}. \quad (4)$$

Линейная градуировка

В матрице X остается два первых столбца; вектор $x_*^T = \{1, x_*\}$.

ДИ регрессии равен

$$\Delta_Y = t_{n-p}^{(1/2)\alpha} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \quad (5)$$

где $\bar{x} = \frac{\sum x_i}{n}$;

ДИ нового измерения

$$\Delta_Y = t_{n-p}^{(1/2)\alpha} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \quad (6)$$

Прямо пропорциональная градуировка

В матрице X остается только второй столбец; вектор $x_*^T = \{x_*\}$ становится скаляром. ДИ регрессии равен

$$\Delta_Y = t_{n-1}^{(1/2)\alpha} \cdot S \cdot \frac{|x_*|}{\sqrt{\sum x_i^2}}, \quad (7)$$

ДИ нового измерения

$$\Delta_Y = t_{n-1}^{(1/2)\alpha} \cdot S \cdot \sqrt{1 + \frac{x_*^2}{\sum x_i^2}}. \quad (8)$$

обращается в ноль, малоприспособны для оценки пределов обнаружения и определения – ошибка не может быть ниже определенной, характерной для прибора и методики, величины. Более адекватная модель зависимости ожидаемой дисперсии от отклика может быть описана прямой с положительным свободным членом или аналогичной степенной зависимостью.

Реальную зависимость ожидаемой дисперсии от отклика необходимо выявлять экспериментально – здесь могут очень пригодиться уровни. Оценивая дисперсию измерений по набору точек, соответствующих одному содержанию аналита, следует иметь в виду, что для этой оценки тоже существует свой доверительный интервал, и для адекватной оценки

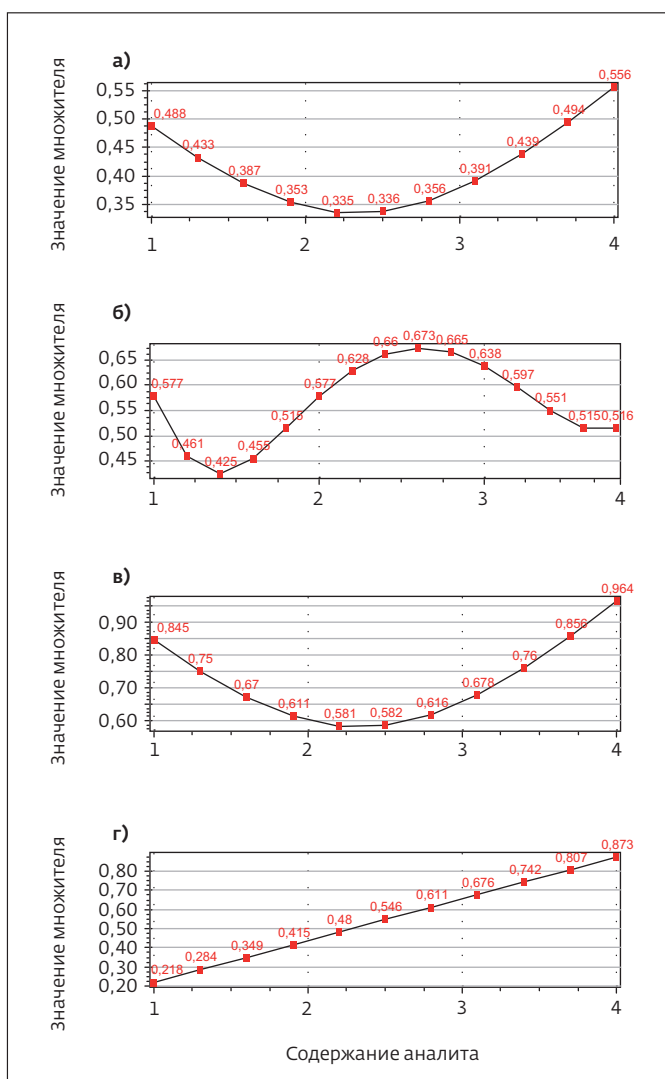


Рис.7. Профили $\sqrt{D}$ для линейной (а) и квадратичной (б) градуировки по точкам с рис.3а и линейной (в) и прямо пропорциональной (г) градуировки по уровням с рис.3б

дисперсии потребуется не менее девяти точек. Однако такую зависимость требуется строить всего один раз для типа анализа или даже для прибора, и усилия вполне могут окупиться доказательным понижением пределов обнаружения и определения. Профиль дисперсии может строиться в ходе разработки методики анализа, на этапе валидации аналитической системы и т.д. В некоторых случаях зависимость дисперсии от содержания аналита может быть посчитана теоретически на основании конструктивной схемы прибора и принципов измерения аналитического сигнала.

Априорно известная дисперсия измерения позволит строить рабочие градуировки на осно-

вании очень небольшого числа точек, поскольку ДИ отдельного измерения оценен заранее по достаточно большому числу точек и за счет этого адекватность оценки параметров градуировки возрастает.

ВЫБРОСЫ И ПРОВЕРКА (ВАЛИДАЦИЯ) ГРАДУИРОВКИ

Обычная процедура идентификации выбросов состоит в построении модели без анализируемой точки, оценке величины аналитического сигнала для содержания аналита в этой точке и сравнения прогнозированного значения с измеренным. В этом сравнении также участвуют доверительные интервалы, и в случае, если точка выходит за их пределы, она считается выбросом (ошибочным значением) и в построении модели в дальнейшем не участвует. Крайне желательно разбираться в причинах появления выбросов и по возможности устранять причины, а не выбросы.

Можно использовать процедуру фильтрации выбросов, основанную опять же на принципах минимизации доверительного интервала: выбросы – это точки, удаление которых приводит к значимому снижению доверительного интервала для определенного диапазона содержаний аналита. Для применения такой процедуры необходима априорная информация о дисперсии (распределении дисперсий) аналитического сигнала – выбросы не следует искать в случае, если дисперсия уже находится в пределах, характерных для данной методики.

Проверка адекватности рабочей градуировки может проводиться по критерию хи-квадрат, как в методе адаптивной фильтрации шума [6], причем этот тест следует применять как для точек, так и для уровней. Проблема проверки градуировки достаточно подробно освещена в нормативной литературе [3–5], так что мы не будем останавливаться на ней подробно.

ПЛАНИРОВАНИЕ ГРАДУИРОВКИ

Минимальные значения пределов обнаружения и определения совсем не означают, что доверительный интервал в любом диапазоне минимален (хотя такое и случается). Если исследователь заранее знает диапазон значений аналитического сигнала (интервал измерений), в котором планируется проводить химический анализ, можно заранее выбрать такой набор

градуировочных точек, который будет наилучшим образом соответствовать решаемой задаче.

Как мы уже обращали внимание, формула (3) доверительного интервала имеет два сомножителя – ДИ отдельного измерения и профиль доверительной вероятности. Если формула уже выбрана, то первый сомножитель зависит исключительно от числа измерений, и остается обратить внимание на второй. Второй сомножитель – профиль доверительной вероятности $\sqrt{u}$ – совсем не зависит от измеренных значений сигнала. Этот профиль формируется на основе значений аппроксимирующих функций при тех x_i , по которым будет строиться градуировка (рис.7). Минимум $\sqrt{u}$ для линейной градуировки находится в "центре масс", т.е. в точке, соответствующей среднему значению градуировочных точек по x , и значение в минимуме зависит исключительно от числа точек градуировки. Можно заранее подобрать такое минимальное число точек и набор положений градуировочных точек по оси содержания аналита, чтобы точность анализа была удовлетворительной. Как всегда, набор точек является компромиссом между потребностями точности анализа и валидации градуировки.

УМЕНЬШЕНИЕ ДИСПЕРСИИ ИЗМЕРЕНИЙ

Дисперсия измерений – самый важный фактор, уменьшающий ДИ. Основные факторы, влияющие на дисперсию, относятся к используемым приборам, устройствам дозирования, приготовлению градуировочных растворов при построении градуировки, повторным анализам и пробоподготовке при проведении анализа. Здесь мы не будем их касаться.

В хроматографии не следует забывать о математических способах увеличения точности анализа при малых концентрациях: оптимальном подавлении шумов [6] и аппроксимации формы пиков функцией, характерной для хроматографии, к примеру экспоненциально модифицированной гауссианой [8].

ЛИТЕРАТУРА

- ГОСТ Р 52361-2005. Контроль объекта аналитический.
- Себер Дж. Линейный регрессионный анализ. – Москва, МИР. Пер. с английского. Под ред. М.Б.Малютова, 1980.
- ГОСТ Р ИСО 11095-2007. Линейная калибровка с использованием образцов сравнения / ISO 11095-1996. Linear calibration using reference materials.
- ГОСТ Р 50.2.028-2003 ГСИ. Алгоритмы построения градуировочных характеристик средств измерений состава веществ и материалов и оценивание их погрешностей (неопределенностей). Оценивание погрешности (неопределенности) линейных градуировочных характеристик при использовании метода наименьших квадратов.
- Danzer K., Currie L.A. Guidelines for calibration in analytical chemistry, Part 1. Fundamentals and single component calibration. Pure & Applied Chemistry, 1998, 70, p.993-1014.
- Каламбет Ю.А., Мальцев С.А., Козьмин Ю.П. Фильтрация шумов: окончательное решение проблемы. – Аналитика, 2011, №1 (1), с.50–55.
- Каламбет Ю.А., Мальцев С.А. Доверительные интервалы градуировки при взвешенном МНК. – Аналитика, 2013, №4 (11), с.42–47.
- Kalambet Y., Kozmin Y., Mikhailova K., Nagaev I., Tikhonov P. Reconstruction of chromatographic peaks using the exponentially modified Gaussian function. – J. Chemometrics, 2011, 25: 352–356. doi: 10.1002/cem.1343.

ОАО «Волжский научно-исследовательский институт целлюлозно-бумажной промышленности»

Разработка и производство фильтровальных материалов

- Бумага крепированная упаковочная небеленая и беленая 70-120 г/м²
- Бумага фильтровальная лабораторная марок Ф, ФС, ФБ, ФМ
- Бумага и картон для средств индивидуальной защиты
- Бумага фильтровальная для очистки воздуха, масла, топлива
- Бумага электроизоляционная для кабелей
- Картон фильтровальный технический



**Отдел сбыта, тел.: (83631)4-97-11, 6-26-56
Факс: (83631)6-97-91, 4-99-65**

Россия, 425000, Республика Марий Эл, г.Волжск, ул.Щорса, д.4
e-mail: sale@marbum.ru

www.marbum.ru