

ОЦЕНКА ПОГРЕШНОСТИ КОЛИЧЕСТВЕННОГО СПЕКТРОХИМИЧЕСКОГО АНАЛИЗА МНОГОКОМПОНЕНТНОЙ СМЕСИ В ПРИСУТСТВИИ СЛУЧАЙНО ИЗМЕНЯЮЩЕЙСЯ БАЗОВОЙ ЛИНИИ СПЕКТРА

И. М. Дубровкин, В. А. Ананьев

ERROR ESTIMATION IN THE QUANTITATIVE SPECTROCHEMICAL ANALYSIS OF MULTICOMPONENT MIXTURE WITH RANDOMLY VARYING SPECTRAL BASELINE

I. M. Dubrovkin, V. A. Ananiev

Работа выполнена в рамках НИР, финансируемой из средств Министерства образования и науки РФ в рамках государственного задания КемГУ.

Погрешности количественного анализа многокомпонентных смесей, измеренных по спектрам поглощения, возмущенных случайно изменяющейся базовой линией, могут быть значительно уменьшены путем введения ковариационной матрицы шумов при расчетах по методу наименьших квадратов. Эффективность предложенного метода продемонстрирована на примере анализа препарата Экседрин.

The errors of quantitative analysis of multicomponent mixtures using absorption spectra perturbed by randomly varying baseline could be significantly reduced by the introduction of the noise covariance matrix in least squares estimation. The effectiveness of the proposed method is demonstrated on the example of Excedrin drug analysis.

Ключевые слова: количественный спектрохимический анализ, прямая калибровка, погрешности, случайно изменяющаяся базовая линия.

Keywords: quantitative spectrochemical analysis, direct calibration, errors, randomly varying base line.

Количественный спектрохимический анализ многокомпонентной смеси известного качественного состава является одним из самых распространенных методов аналитической химии [1]. Например, количественный анализ лекарственных препаратов можно проводить по спектрам оптического поглощения растворов, полученных после их растворения. Так как спектр поглощения каждого компонента смеси в растворе заранее известен, то для анализа обычно используют метод прямой калибровки (метод **K**-матрицы) [1]. Столбцы матрицы **K** представляют собой значения коэффициентов молярного поглощения индивидуальных компонентов смеси для заданных длин волн. Предполагается, что эти столбцы линейно независимы. При сильном перекрытии спектров индивидуальных компонентов используются методы хемометрики [2].

Классический метод прямой калибровки основан на использовании идеальной модели спектра многокомпонентной смеси, предполагающей присутствие только гомоскедатичного (обладающего постоянной дисперсией) и некоррелированного шума. Это допущение не всегда оправдано из-за нестабильности условий регистрации спектра и специфики его цифровой обработки [3]. Кроме этого, отклонение от идеальной модели может быть обусловлено наличием неконтролируемых примесей в растворе. Визуально указанные эффекты могут проявляться в спектре поглощения в форме базовой линии (фона), для устранения которой используется предварительная обработка спектров: алгебраическая коррекция фона [1], производная спектрометрия [4], вейвлет-преобразование [5]. Как правило, полная компенсация базовой линии практически невозможна, поэтому в спектре смеси всегда будут оставаться ее "следы" со случайно изменяющимися параметрами.

Целью настоящей работы является оценка влияния случайно изменяющейся базовой линии на точностные характеристики количественного анализа по методу прямой калибровки.

Теория

В статье используются общепринятые термины линейной алгебры. Заглавные и строчные буквы, выделенные жирным шрифтом, обозначают матрицы и векторы-столбцы, соответственно, скаляры выделены курсивом. Надстрочные индексы **T** и **-1** используются для обозначения транспонированной и обратной матриц, соответственно.

В соответствии с законом Бугера-Ламберта-Бера спектр оптического поглощения смеси, состоящей из **y** компонентов, можно записать в виде

$$y = Kcd + \epsilon \tag{1}$$

где **K** – $m \times n$ матрица, столбцы которой составлены из коэффициентов молярного поглощения чистых компонентов смеси для заданных длин волн; **m** – число заданных длин волн; **c** – $n \times 1$ вектор концентраций компонентов; **d** – толщина поглощающего слоя (кюветы), которая для простоты изложения принята равной 1 см; **ε** – $m \times 1$ вектор нормального шума с нулевым средним и ковариационной матрицей $\sigma_y^2 \mathbf{V}$. Для идеальной модели **V** является единичной матрицей, а σ_y^2 равно дисперсии шумов. При наличии случайно изменяющейся базовой линии матрица **V** отлична от единичной (см. Приложение).

Оценка вектора **c** по обобщенному методу наименьших квадратов [4] равна:

$$\hat{c} = (\mathbf{K}^T \mathbf{V}^{-1} \mathbf{K})^{-1} \mathbf{K}^T \mathbf{V}^{-1} y, \tag{2}$$

с дисперсионной матрицей

$$\mathbf{D}_{\hat{c}} = \sigma_y^2 (\mathbf{K}^T \mathbf{V}^{-1} \mathbf{K})^{-1}. \tag{3}$$

Дисперсия вектора \hat{c} :

$$\sigma_{\hat{c}}^2 = \text{tr}(\mathbf{D}_e) + 2 \sum_{\substack{j \leq n \\ j > i \geq 1}} D_{ij} D_{ji}, \quad (4)$$

где tr – след матрицы; D_{kl} – элемент матрицы \mathbf{D} . Наличие второго члена в выражении (4) показывает, что концентрации отдельных компонентов смеси коррелированы.

Численной оценкой дисперсии вектора концентраций является его среднеквадратическая погрешность

$$MSE(\alpha) = \text{mean} \left(\sum_{k=1}^n (\hat{c}_k - c_k^{mix})^2 + 2 \sum_{\substack{j \leq n \\ j > i \geq 1}} (\hat{c}_i - c_i^{mix})(\hat{c}_j - c_j^{mix}) \right), \quad (5)$$

где mean обозначает усреднение повторных численных оценок; c_k^{mix} – точная концентрация k -го компонента в смеси.

Для подобных по форме и сильно перекрывающихся спектров чистых компонентов расчет обратной матрицы в выражении (2) будет очень чувствителен к погрешностям измерения спектров. Для стабилизации обращения матриц используется метод регуляризации по Тихонову [5], согласно которому

$$\hat{c} = (\mathbf{K}^T \mathbf{V}^{-1} \mathbf{K} + \alpha \mathbf{I})^{-1} \mathbf{K}^T \mathbf{V}^{-1} \mathbf{y}, \quad (6)$$

где α – параметр регуляризации; \mathbf{I} – единичная матрица. Мы будем подбирать значение параметра регуляризации таким образом, чтобы минимизировать относительную среднеквадратическую погрешность оценки \hat{c} в уравнении (6):

$$RMSE(\alpha)_{\min} = \min_{\alpha} MSE(\alpha) / \sum_{k=1}^n (c_k^{mix})^2. \quad (7)$$

Оценка погрешностей анализа таблеток Экседрин

Спектры Экседрина и его ингредиентов (рис. 1) в цифровой форме были получены путем компьютерного сканирования опубликованных спектров [6].

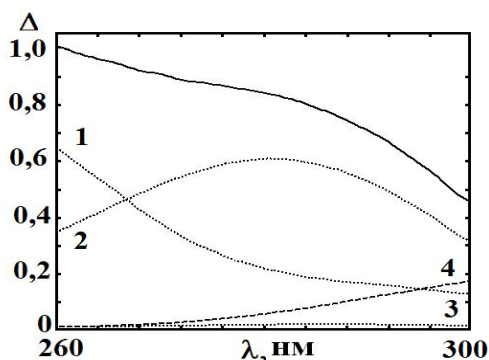


Рис. 1. Спектры препарата Экседрин (—) и его ингредиентов (...) (парацетамол (1), кофеин (2) и аспирин (3), в соотношении 0.4464:0.1071:0.4464, соответственно). Суммарная концентрация равна 20 мг/л в смеси вода-метанол-ледяная уксусная кислота (соотношение указанных компонентов 69:28:3). Спектр салициловой кислоты (4) с концентрацией 4 мг/л в той же смеси. Воспроизводится с разрешения "http://www.dissolutiontech.com"

Было взято 100 равноотстоящих отсчетов в области 260 – 300 нм. К спектру раствора таблеток Экседрин (рис. 1) добавлялся полиномиальный фон (см. уравнение (A1) Приложения) со случайными параметрами. Численные оценки корня относительной среднеквадратической погрешности ($RRMSE$) производились путем усреднения 10000 статистически независимых расчетов. Для вычислений и построения графиков была написана программа на языке MATLAB.

Результаты численного эксперимента представлены на рис. 2.

Анализ полученных данных позволяет сделать следующие выводы:

1. Использование матрицы \mathbf{V} , рассчитанной в строгом соответствии с дисперсиями коэффициентов полиномиального фона (см. Приложение) свидетельствуют о хорошем совпадении теоретических и численных оценок (рис. 2а).

2. Пренебрежение случайным фоном при больших стандартных отклонениях его наклона (σ_{a_1}) значительно увеличивает погрешность определения вектора концентраций (рис. 2б).

3. Зависимость относительной среднеквадратической погрешности оценки вектора концентрации от стандартного отклонения постоянного члена полинома (σ_{a_0}) мала по сравнению с её зависимостью от стандартных отклонений коэффициентов первой (σ_{a_1}) и второй степени (σ_{a_2}). Таким образом, погрешность анализа весьма чувствительна к случайным изменениям наклона некомпенсированной базовой линии.

4. Нами обнаружено что, при оптимальном выборе матрицы \mathbf{V} регуляризация решения уравнения (1) малоэффективна. Однако при больших стандартных отклонениях наклона базовой линии в случае $\mathbf{V} = \mathbf{I}$, благодаря регуляризации погрешность анализа значительно снижается (рис. 2в).

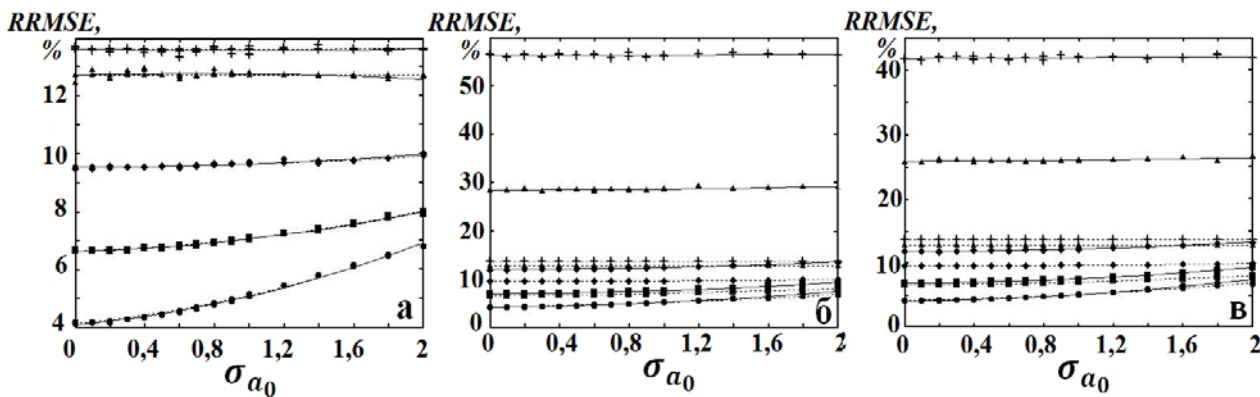


Рис. 2. Теоретические (...) и численные (—) оценки погрешности вектора концентраций компонентов препарата Экседрин. (а) Точная ковариационная матрица V (б). $V = I$ (в). Метод регуляризации при $V = I$. $\sigma_y = 0,005$, $\sigma_{a_1} = 0$ (●), 1×10^{-4} (■), 2×10^{-4} (◆), 5×10^{-4} (▲), 10×10^{-4} (+) $\sigma_{a_2} = \sigma_{a_1} / 10$

В качестве практического примера оценим погрешность анализа таблеток Экседрин, в спектре раствора которых присутствует фон, обусловленный наличием салициловой кислоты, являющейся продуктом гидролиза аспирина. Будучи очень слабым, этот фон в значительной степени искажается шумами спектрометра и поэтому меняется при многократной регистрации спектра. Для моделирования этого процесса спектр раствора салициловой кислоты (рис. 1), возмущенный нормальным шумом $\sigma_y = 0,005$, был аппроксимирован полиномом второй степени. При 10^5 повторениях статистически независимых численных экспериментов были получены следующие значения дисперсий полиномиальных коэффициентов: $\sigma_{a_0} = 0,61$; $\sigma_{a_1} = 0,028$; $\sigma_{a_2} = 0,00027$, на основании которых была рассчитана матрица V . Полученные результаты: RRMSE 11.7 % и 20.9 % для оптимальной матрицы V и $V = I$, соответственно, свидетельствуют о высокой эффективности учета переменного фона при количественном анализе многокомпонентной смеси.

Заключение

Методика, разработанная в данной статье, позволяет в каждом конкретном случае оценить погрешность аналитической процедуры при наличии гетероскедастичных коррелированных шумов, обусловленных случайно изменяющейся базовой линией спектра. Сравнение погрешности эксперимента с погрешностью, полученной путем компьютерного моделирования с использованием оптимальной ковариационной матрицы шумов может быть обоснованием целесообразности дальнейшего улучшения используемого для анализа смеси спектрофотометрического метода, например, изменением pH раствора.

Приложение

Расчет ковариационной матрицы шума

Предположим, что в спектре, регистрируемом r раз, присутствует базовая линия, описываемая полиномом второй степени. Тогда отсчет i спектра t :

$$y_{it} = \delta_{it} + a_0 + a_1 i + a_2 i^2, \tag{A1}$$

где y_{it} – отсчет неискаженного спектра, обладающего нормальным шумом δ_{it} с нулевым средним и дисперсией σ_y^2 ; a_{qt} – q -й коэффициент полинома, остающийся постоянным в пределах данного спектра, однако, меняющийся случайным образом от спектра к спектру. Оценка элемента ковариационной матрицы, согласно данным работы [4]:

$$COV(y_k, y_l) = \frac{1}{r-1} \sum_{t=1}^r (y_{kt} - \bar{y}_k)(y_{lt} - \bar{y}_l), \tag{A2}$$

где черта означает среднее значение отсчета спектра по r повторениям. Из уравнения (A1) следует, что:

$$\bar{y}_i = \bar{y}_0 + \bar{a}_1 i + \bar{a}_2 i^2, \tag{A3}$$

Подставляя (A1) и (A3) в (A2), получим:

$$COV(y_k, y_l) = \sigma_y^2 \xi_{kl} + \frac{1}{r-1} \sum_{t=1}^r (\delta_{a_0 t} + \delta_{a_1 t} k + \delta_{a_2 t} k^2) (\delta_{a_0 t} + \delta_{a_1 t} l + \delta_{a_2 t} l^2), \tag{A4}$$

где ξ_{kl} – символ Кронекера ($\xi_{kl} = 1$ для $k = l$, иначе $\xi_{kl} = 0$), $\delta_{a_{qt}} = a_{qt} - \bar{a}_q$. Полагая, что $\delta_{a_{qt}}$ является нормально распределенной случайной величиной с нулевым средним и ковариационной матрицей $\sigma_{a_{qt}} \mathbf{I}$ (\mathbf{I} – единичная матрица) и, пренебрегая малым вкладом суммы перекрестных членов при достаточно большом r , уравнение (A4) приводится к виду:

$$COV(y_k, y_l) = \sigma_y^2 \xi_{kl} + \sigma_{a_0}^2 + \sigma_{a_1}^2 kl + \sigma_{a_2}^2 k^2 l^2$$

Окончательно, элементы матрицы V можно определить следующим образом:

$$V(k, l) = \xi_{kl} + \sigma_y^{-2} (\sigma_{a_0}^2 + \sigma_{a_1}^2 kl + \sigma_{a_2}^2 k^2 l^2). \tag{A5}$$

Литература

1. Бернштейн, И. Я. Спектрофотометрический анализ в органической химии/ И. Я. Бернштейн, Ю. А. Каминский. – Л.: Химия. – 1986. – 200 с.
2. Родионова, О. Е. Хемометрика: достижения и перспективы / О. Е. Родионова, А. Л. Померанцев // Успехи химии. – 2006. – Т. 75. – № 4. – С. 302 – 321.
3. Fernandez-Ahumada, E. A new formulation to estimate the variance of model prediction. Application to near infrared spectroscopy calibration / E. Fernandez-Ahumada, J. M. Roger, B. Palagos // Anal. Chim. Acta. – 2012. – V. 721. – P. 28 – 34.
4. Дубровкин, И. М. Химический анализ с использованием математического дифференцирования / И. М. Дубровкин // ЖАХ. – 1988. – Т. XLIV. – № 6. – С. 985 – 979.
5. Derczynski, U. Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm / U. Derczynski, K. Jetter, K. Molt [et al.] // Chemom. Intell. Lab. Syst. – 1999. – V. 47. – P. 179 – 187.
6. Себер, Дж. Линейный регрессионный анализ / Дж. Себер. – М.: Мир. – 1980. – 456 с.
7. Тихонов, А. Н. Методы решения некорректных задач / А. Н. Тихонов, В. Я. Арсенин. – М.: Наука. – 1986. – 288 с.
8. Simonian, M. H. Full spectrum quantitation: an advance UV. Visible spectroscopic method for multi-component dissolution testing / M. H. Simonian. – Режим доступа: <http://www.dissolutiontech.com>.

Информация об авторах:

Дубровкин Иосиф Михайлович – доктор физико-математических наук, заведующий отделом точных наук. Академический колледж Западная Галилея, Израиль, 972-4-6292910, dubrovkin@013.net.il.

Iosif M. Dubrovkin – Doctor of Physics and Mathematics, Head of Exact Sciences Department, Western Galilee College, Israel.

Ананьев Владимир Алексеевич – доктор химических наук, профессор кафедры аналитической химии КемГУ, eprlab@kemsu.ru.

Vladimir A. Ananiev – Doctor of Chemistry, Professor at the Department of Analytical Chemistry, Kemerovo State University.