ПРИМЕНЕНИЕ ЛИНЕЙНОГО ПОИСКА ДЛЯ РАСЧЕТА КАНОНИЧЕСКОГО ТЕНЗОРНОГО РАЗЛОЖЕНИЯ СПЕКТРОВ МОЛЕКУЛЯРНОЙ ФЛУОРЕСЦЕНЦИИ

И. Н. Крылов, И. В. Селиверстова, Т. А. Лабутин*

УДК 535.372

https://doi.org/10.47612/0514-7506-2023-90-1-90-96

Московский государственный университет имени М. В. Ломоносова, Москва, Россия; e-mail: timurla@laser.chem.msu.ru

(Поступила 26 октября 2022)

Выделение групп флуорофоров в растворенном органическом веществе с помощью канонического тензорного разложения PARAFAC трехмерных спектров флуоресценции возбуждение/испускание широко используется при изучении природных вод, однако его расчет, особенно на стадии валидации, требует очень больших временных затрат. Рассмотрены несколько стратегий ускорения канонического тензорного разложения для спектров молекулярной флуоресценции морских вод. Показано, что стратегии с оптимизацией большого количества параметров экстраполяции не позволяют достичь значительного ускорения из-за больших временных затрат на эту операцию. Предложено решение, когда оптимизация шага проводится для одной переменной один раз на несколько итераций алгоритма. Подобный подход позволяет достичь ускорения расчетов с использованием линейного поиска. Максимальное ускорение в 2.3 раза достигнуто при использовании стратегии линейного поиска, в которой шаг экстраполяции является степенной функцией номера итерации, хотя в этом случае на некоторых стадиях работы алгоритма наблюдается коллинеарность последовательных шагов.

Ключевые слова: *молекулярная флуоресценция, трехмерные спектры испускание/возбуждение, флуорофоры, каноническое тензорное разложение PARAFAC, линейный поиск.*

Recovery of fluorophore groups in dissolved organic matter using the PARAFAC canonical tensor decomposition of fluorescence excitation-emission matrix (EEM) is widely used in the study of natural waters. However, fitting the PARAFAC model, especially for its validation, is very time consuming. Several strategies for accelerating the PARAFAC fitting to the EEM of sea waters were considered. It was shown that strategies with optimization of a large set of hyperparameters do not result in significant acceleration due to high time costs for this operation. It was proposed to perform optimization for one variable once for several iterations of the algorithm. This approach made it possible to achieve acceleration of calculations using line search strategy. The maximum acceleration by 2.3 times was achieved for the line search strategy using the extrapolation step in a power function of the iteration number, although in this case, sequential steps are collinear at some stages of the algorithm.

Keywords: *molecular fluorescence, fluorescence excitation–emission matrix, canonical tensor decomposition PARAFAC, line search.*

Введение. В спектре флуоресценции растворенного органического вещества (РОВ) в природных водах нельзя выделить полосы индивидуальных соединений из-за их исключительно сложного состава, поэтому принято рассматривать группы флуорофоров РОВ — условные соединения, которым приписывают определенные оптические характеристики, связанные с происхождением и трансформацией органического вещества в водах [1]. Для повышения возможностей по характеризации от-

---

IMPLEMENTATION OF LINE SEARCH FOR PARAFAC ANALYSIS OF FLUORESCENCE EXCITATION-EMISSION MATRIX

I. N. Krylov, I. V. Seliverstova, T. A. Labutin* (*Lomonosov Moscow State University, Moscow, Russia; e-mail: timurla@laser.chem.msu.ru*)

дельных флуорофоров при оптическом изучении природных вод широко применяется флуоресцентный метод с регистрацией трехмерных спектров испускания/возбуждения [2]. Простейшим алгоритмом обработки трехмодальных данных является их развертывание в обычную матрицу, однако имеется неоднозначность развертывания и, самое главное, теряется связь между соседними точками [3]. Поэтому в последнее время широкое развитие получил алгоритм канонического тензорного разложения, основанный на параллельном факторном анализе (PARAFAC), который позволяет обрабатывать трехмерные спектры флуоресценции с сохранением их первоначальной структуры [4] и получать данные об индивидуальных соединениях или классах близких по оптическим свойствам веществ [5]. Несомненные достоинства канонического тензорного разложения — единственность решения (хотя оно требует значительных расчетных ресурсов для больших наборов данных), соответствие физической модели спектров молекулярной флуоресценции [6], выделение спектров возбуждения и испускания отдельных флуорофоров, а также вкладов отдельных флуорофоров без использования дополнительной априорной информации. Сложным является выбор количества компонент в каноническом тензорном разложении [7]: хотя единственное решение существует для любого количества компонент, не у каждого из них есть физический смысл. Распространен анализ делением на половины (split-half analysis) [8], который заключается в сравнении разложений двух половин набора данных. Для анализа спектров флуоресценции нами предложен вариант статистического рассмотрения результатов многократного разложения на половины исходного набора — метод “случайного” деления на половины [9]. Количество поднаборов из двух половин из n образцов $C_n^{n/2}$ растет быстрее, чем $2^{n/2}$, и велико даже для небольших n. Требуемое время расчетов возрастает многократно, даже если выборку для статистической оценки качества модели сделать небольшой долей от этого числа [9]. Соответственно, необходимо разрабатывать алгоритмы для ускорения расчетов PARAFAC. В данной работе рассмотрено применение алгоритма акселерации — линейного поиска.

Теория. На примере трехмерного тензора $\boldsymbol{X}$, собранного из спектров возбуждения/испускания флуоресценции, задачу канонического тензорного разложения [10] можно представить как

$$\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} \sum_{i,j,k} \left(\sum_{r}^{R} A_{i,r} B_{j,r} C_{k,r} - X_{i,j,k} \right)^2 = \left\| \boldsymbol{A}\left(\boldsymbol{B} \odot \boldsymbol{C}\right) - \boldsymbol{X} \right\|^2,$$

где $\odot$ — произведение Хатри—Рао, или поколоночное произведение Кронекера.

Если первое измерение тензора соответствует длинам волн испускания, второе — длинам волн возбуждения флуоресценции, третье — образцам, то столбцы матрицы $\boldsymbol{A}$ содержат величины, пропорциональные спектрам испускания различных флуорофоров, матрицы $\boldsymbol{B}$ — спектрам их возбуждения, а матрицы $\boldsymbol{C}$ — их вкладам в спектр каждого образца. Это выполняется при условии, если параметр R (количество компонентов) правильно выбран для данного тензора $\boldsymbol{X}$ [5]. Назначить конкретные единицы измерения значениям матриц без наложения дополнительных ограничений на решение не позволяет неопределенность шкалы, поскольку из имеющегося решения всегда можно сделать другое решение с такой же нормой невязки путем умножения его на комбинацию констант:

$$\sum_r A_{i,r} B_{j,r} C_{k,r} = \sum_r \left(\alpha_r A_{i,r}\right)\left(\beta_r B_{j,r}\right)\left(\gamma_r C_{k,r}\right) = \sum_r \tilde{A}_{i,r} \tilde{B}_{j,r} \tilde{C}_{k,r}, \text{ если } \alpha_r \beta_r \gamma_r = 1 \forall r .$$

Модель также обладает неопределенностью перестановки: если столбцы каждой из трех матриц поменять местами одинаковым образом, получится еще одно решение с такой же нормой невязки. Универсального решения этой задачи в замкнутой форме не существует, поэтому для его поиска, как правило, применяют покоординатный спуск (alternating least squares), решая подзадачи для каждой матрицы $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ отдельно, фиксируя две другие матрицы. У этих подзадач существуют решение, выражаемое в явном виде через псевдообратные матрицы [11], и критерии значимости параметров и адекватности модели [12]. В целом, для решения задачи необходимо выполнять следующие шаги до достижения сходимости:

$$\boldsymbol{A}_i \leftarrow \boldsymbol{X}\left(\boldsymbol{C}_{i-1} \odot \boldsymbol{B}_{i-1}\right)^+,$$

$$\boldsymbol{B}_i \leftarrow \boldsymbol{X}\left(\boldsymbol{A}_{i-1} \odot \boldsymbol{C}_{i-1}\right)^+,$$

$$\boldsymbol{C}_i \leftarrow \boldsymbol{X}\left(\boldsymbol{B}_{i-1} \odot \boldsymbol{A}_{i-1}\right)^+,$$

где $\boldsymbol{A}^+$ — псевдообратная к матрице $\boldsymbol{A}$.

По сравнению с градиентными и Ньютон-подобными подходами, где происходит оптимизация всех параметров сразу, покоординатный спуск требует меньше оперативной памяти [13]. В некото-

рых случаях применение регуляризации совместно с линейным поиском может улучшить результаты разложения [14]. Метод Левенберга—Маркуардта может приводить к локальному минимуму в отличие от покоординатного спуска с линейным поиском [15].

Во многих случаях сходимость покоординатного спуска затруднена из-за малого размера шага в одном направлении [16]. Этим можно воспользоваться, выполняя экстраполирующие шаги каждый раз после обычных шагов алгоритма:

$$\boldsymbol{A}_i \leftarrow \boldsymbol{X}\left(\boldsymbol{C}_{i-1} \odot \boldsymbol{B}_{i-1}\right)^{+},$$
$$\boldsymbol{B}_i \leftarrow \boldsymbol{X}\left(\boldsymbol{A}_{i-1} \odot \boldsymbol{C}_{i-1}\right)^{+},$$
$$\boldsymbol{C}_i \leftarrow \boldsymbol{X}\left(\boldsymbol{B}_{i-1} \odot \boldsymbol{A}_{i-1}\right)^{+},$$
$$\boldsymbol{A}_{i+1} \leftarrow \boldsymbol{A}_i + \alpha_A\left(\boldsymbol{A}_i - \boldsymbol{A}_{i-1}\right),$$
$$\boldsymbol{B}_{i+1} \leftarrow \boldsymbol{B}_i + \alpha_B\left(\boldsymbol{B}_i - \boldsymbol{B}_{i-1}\right),$$
$$\boldsymbol{C}_{i+1} \leftarrow \boldsymbol{C}_i + \alpha_C\left(\boldsymbol{C}_i - \boldsymbol{C}_{i-1}\right).$$

Если выбор коэффициентов экстраполяции α_A, α_B, α_C занимает меньше времени, чем расчет произведений Хатри—Рао и матриц, псевдообратных к ним, то использование экстраполяции позволяет снизить время вычисления разложения.

Особенно заметными оказываются проблемы, связанные с медленной сходимостью, когда один или несколько компонентов канонического тензорного разложения коллинеарны [15]. В этом случае до тех пор, пока шаг вычисления псевдообратных матриц может дать полезное направление, линейная экстраполяция дает улучшение невязки [17].

Набор данных и методы. Использован набор спектров флуоресценции морских вод, собранных во время 63 рейса (2015 г.) исследовательского судна “Академик Мстислав Келдыш” [9], состоящий из 30 образцов по 58 длинам волн возбуждения и 391 длине волны испускания. Программная реализация алгоритма линейного поиска выполнена на языке программирования R с использованием пакетов albatross [9] версии 0.3-5 и multiway [18] версии 1.0-6. Решение задачи канонического тензорного разложения останавливали при относительном изменении нормы невязки $<10^{-8}$. Перед выполнением расчетов зону, включающую сигнал рассеяния, из спектров удалили и интерполировали с использованием алгоритма LOESS [19].

Рассмотриваем три стратегии линейного поиска: полную, одномерную и степенную. Полная стратегия линейного поиска заключается в минимизации функции от переменных α_A, α_B, α_C, соответствующих шагам экстраполяции по измерениям исходного тензора:

$$\min_{\alpha_A,\alpha_B,\alpha_C} \sum_{i,j,k}\left(X_{i,j,k} - \sum_r\left(A_{i,r} + \alpha_A\Delta A_{i,r}\right)\left(B_{j,r} + \alpha_B\Delta B_{j,r}\right)\left(C_{k,r} + \alpha_C\Delta C_{k,r}\right)\right)^2.$$

Функцию минимизировали с помощью алгоритма L-BFGS-B [20] с начальными значениями 0.2 и ограничениями в [0; 1000] для всех трех переменных. Итерацию пропускали, если не удавалось снизить функцию потерь по сравнению с шагом без экстраполяции.

Одномерная стратегия линейного поиска заключается в минимизации функции от одной переменной, соответствующей шагу экстраполяции, общему для всех трех измерений исходного тензора:

$$\min_{\alpha} \sum_{i,j,k}\left(X_{i,j,k} - \sum_r\left(A_{i,r} + \alpha\Delta A_{i,r}\right)\left(B_{j,r} + \alpha\Delta B_{j,r}\right)\left(C_{k,r} + \alpha\Delta C_{k,r}\right)\right)^2.$$

Функцию минимизировали с помощью безградиентного метода одномерной оптимизации [21] с ограничениями в [0; 1000].

Степенная стратегия линейного поиска [22] заключается в том, чтобы делать шаг экстраполяции, равный степенной функции от номера итерации n:

$$\alpha = n^{1/p}.$$

По умолчанию $p = 3$. Каждый раз, когда экстраполяция не приводит к снижению невязки, шаг не используют, а перед следующей попыткой параметр p увеличивают на 1.

Расчет модели стандартным алгоритмом, а также с выбранными стратегиями линейного поиска повторяли 512 раз со случайными равномерно распределенными начальными значениями параметров для получения статистической оценки эффективности алгоритмов ускорения расчета модели.

Результаты и их обсуждение. Использование полной стратегии показало значительное снижение количества итераций, требуемых для решения задачи канонического тензорного разложения, однако суммарное время на ее решение выросло из-за ресурсоемкой операции по определению оптимальной длины шага: вычисление произведений Кронекера и псевдообратных матриц в стандартном алгоритме происходит быстрее. По умолчанию алгоритм L-BFGS-B останавливается, если относительное изменение функции потерь $<10^7\varepsilon$, где ε — машинная точность. Только снижение критерия остановки до $10^{12}\varepsilon$ и использование оптимизации один раз на 10 стандартных итераций алгоритма обеспечивало достижение паритета по времени расчета модели.

Аналогично полной стратегии оптимизировано использование одномерного линейного поиска: оценку шага проводили только каждые четыре итерации, а точность решения α ослаблена с $\pm(|\alpha|\sqrt{\varepsilon}+\sqrt[4]{\varepsilon}/3)$ до $\pm(|\alpha|\sqrt{\varepsilon}+\sqrt[8]{\varepsilon}/3)$. Это позволило в два раза превзойти по времени стандартный вариант алгоритма без линейного поиска, не снижая точности конечного решения.

Степенная стратегия линейного поиска сама по себе снижает общее время вычислений, а также позволяет достигать существенного ускорения в 2.3 раза при достижении сходимости с необходимой точностью, если выполнять ее каждые две итерации. На рис. 1 показаны времена, затрачиваемые на решение задачи канонического тензорного разложения с использованием рассмотренных стратегий линейного поиска.

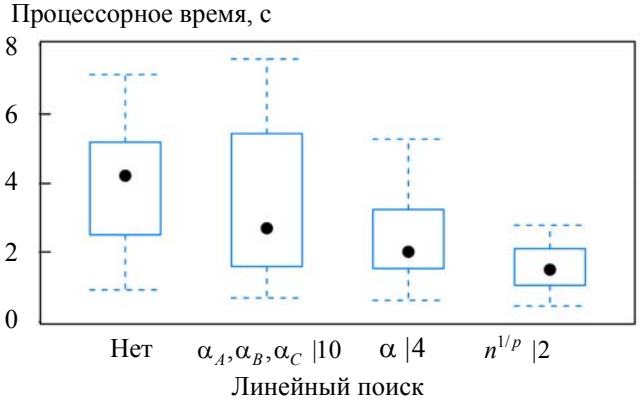

Рис. 1. Время решения задачи канонического тензорного разложения: “Нет” — без линейного поиска; $\alpha_A,\alpha_B,\alpha_C|10$ — полная стратегия раз в 10 итераций; $\alpha|4$ — одномерная стратегия раз в 4 итерации; $n^{1/p}|2$ — степенной шаг раз в две итерации; точка — медиана, прямоугольник — область между 25 и 75 перцентилями, штриховая линия — размах

Многомерная оптимизация даже при значительном пропуске дорогостоящих итераций может замедлить решение задачи, хотя в случае большого числа запусков можно говорить о небольшом уменьшении общего времени расчетов; одномерная оптимизация в большинстве случаев обеспечивает значимое преимущество по сравнению со стандартным алгоритмом без линейного поиска, степенная стратегия консистентно выигрывает по времени расчетов у исходного варианта без ускорения.

На рис. 2 показаны совместные распределения процессорного времени, затраченного на вычисление разложений, и количеств итераций. Видно, что использование методов оптимизации заметно поднимает стоимость одной итерации по сравнению с методом без линейного поиска, что приходится компенсировать пропуском итераций. Пропуская слишком много итераций или ослабляя критерии остановки оптимизатора, получаем размеры шагов экстраполяции недостаточного качества, которые не приведут к решению быстрее, чем неускоренный метод. В отличие от стратегий, использующих оптимизацию, степенной метод практически не увеличивает вычислительную стоимость итераций, что позволяет применять его чаще, даже если предложенные шаги не всегда оптимальны.

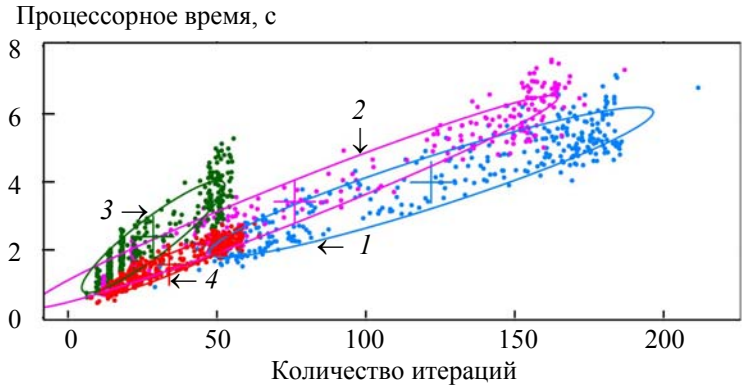

Рис. 2. Распределения процессорного времени и количества итераций при решении задачи канонического тензорного разложения с использованием различных стратегий линейного поиска: *1* — “Нет” (•), *2* — $\alpha_A,\alpha_B,\alpha_C|10$ (•), *3* — $\alpha|4$ (•), *4* — $n^{1/p}|2$ (•)

На рис. 3 на примере индивидуальных траекторий показаны косинусы углов между последовательными шагами с использованием различных стратегий линейного поиска: все параметры канонического разложения представлены в виде одного вектора **x**, после чего вычислен косинус:

$$\cos\left(\mathbf{x}_{i+1}-\mathbf{x}_i,\,\mathbf{x}_i-\mathbf{x}_{i-1}\right)=\frac{\left(\mathbf{x}_{i+1}-\mathbf{x}_i\right)^T\left(\mathbf{x}_i-\mathbf{x}_{i-1}\right)}{\sqrt{\left\|\mathbf{x}_{i+1}-\mathbf{x}_i\right\|\left\|\mathbf{x}_i-\mathbf{x}_{i-1}\right\|}}.$$

Метод без линейного поиска практически сразу начинает предлагать шаги в одном и том же направлении, тогда как одномерная стратегия и степенной метод позволяют полностью использовать текущее направление и начать следующий шаг в другом направлении. (Метод без ускорения требует гораздо больше шагов, чем другие методы, поэтому график не продолжается дальше 30 итераций.) Ближе к концу траектории степенная стратегия перестает улучшать невязку и приводит к нескольким коллинеарным шагам.

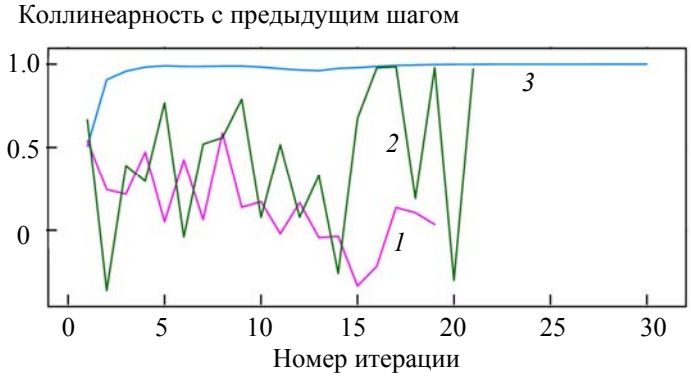

Рис. 3. Косинус угла между соседними шагами как функция от номера итерации при использовании различных алгоритмов линейного поиска: $\alpha_A,\alpha_B,\alpha_C$ (*1*), $n^{1/p}|2$ (*2*) и “Нет” (*3*)

На рис. 4 представлены результаты анализа спектров флуоресценции. Для выбора количества компонент набор данных перемешивали, разделяли на половины (64 раза) и сравнивали результаты разложения двух половин по величине среднего косинуса угла между соответствующими друг другу столбцами матриц ***A*** и ***B***. Хотя многие пары четырехкомпонентных разложений после сравнения оказываются ниже порога в 0.95 [23], медиана для среднего косинуса угла между соответствующими компонентами выше данного порога, что не достигается для пятикомпонентных моделей. Это позво-

ляет считать разложение на четыре компоненты правильным решением для данного набора данных. Три компоненты коррелируют друг с другом и с индексом гумификации образца [24]. Компонента, связанная с флуоресценцией триптофана, ожидаемо демонстрирует отрицательную корреляцию с индексом гумификации, поскольку связана с автохтонным органическим веществом [25].

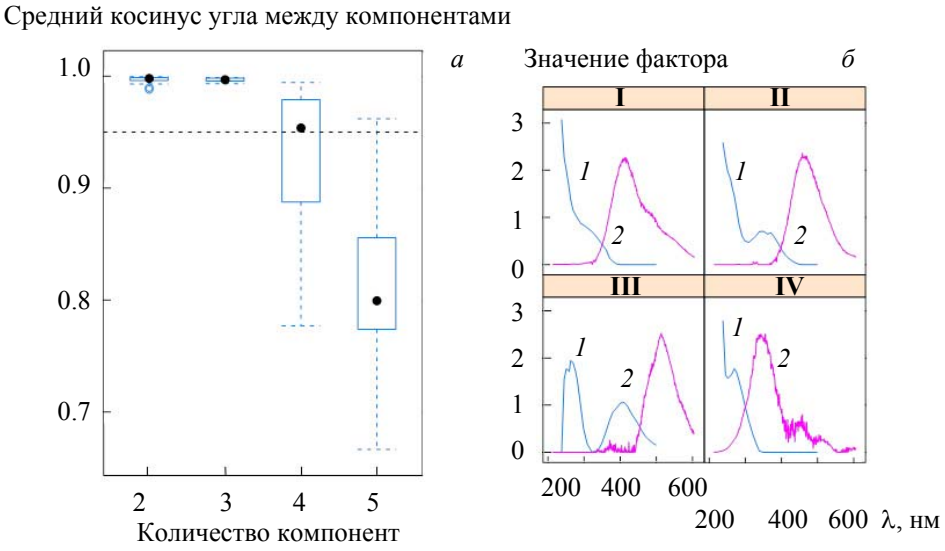

Рис. 4. Валидация канонического тензорного разложения методом случайного деления на половины (*а*) и спектры групп I—IV флуорофоров на основании PARAFAC-разложения (*б*): возбуждение (*1*) и испускание (*2*)

Заключение. Полная, одномерная и степенная стратегии линейного поиска применены к задаче канонического тензорного разложения набора спектров молекулярной флуоресценции (испускания/возбуждения) морских вод с установленным ранее числом флуорофоров, равным четырем. Ускорение расчетов обеспечивают одномерная и степенная стратегии при ослаблении критерия сходимости и выборе оптимального шага не на каждой итерации. Оптимальным представляется использование степенной стратегии с начальной степенью экстраполяции 1/3, что позволяет добиться максимального ускорения работы алгоритма для канонического тензорного разложения PARAFAC, проводить разложение больших наборов данных (несколько сотен спектров) и увеличить количество разбиений при проведении статистического рассмотрения качества полученной модели для повышения надежности выбора числа компонент (флуорофоров).

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 20-33-90280).

[1] **R. M. Cory, D. M. McKnight.** Environ. Sci. Technol., **39**, N 21 (2005) 8142—8149
[2] **C. A. Stedmon, S. Markager, R. Bro.** Marine Chem., **82**, N 3-4 (2003) 239—254
[3] **О. Е. Родионова, А. Л. Померанцев.** Успехи химии, **75**, № 4 (2006) 302—321
[4] **R. Bro.** Chemometrics Intell. Lab. Systems, **38**, N 2 (1997) 149—171
[5] **C. M. Andersen, R. Bro.** J. Chemometrics, **17**, N 4 (2003) 200—215
[6] **C. A. Stedmon, R. Bro.** Limnology and Oceanography: Methods, **6**, N 11 (2008) 572—579
[7] **C. J. Hillar, L.-H. Lim.** J. ACM, **60**, N 6 (2013) 45(1—39)
[8] **W. S. DeSarbo.** An Application of PARAFAC to a Small Sample Problem, Demonstrating Preprocessing, Orthogonality Constraints, and Split-Half Diagnostic Techniques (Appendix), Rochester, New York, Social Science Research Network (1984)
[9] **I. N. Krylov, A. N. Drozdova, T. A. Labutin.** Chemometrics Intell. Lab. Systems, **207** (2020) 104176
[10] **F. L. Hitchcock.** J. Mathem. Phys., **6**, N 1-4 (1927) 164—189
[11] **В. С. Муха.** Изв. НАН Беларуси. Сер. физ.-мат. наук, **50**, № 2 (2016) 71—81
[12] **В. С. Муха.** Изв. НАН Беларуси. Сер. физ.-мат. наук, **50**, № 4 (2016) 53—60

[13] **R. Bro.** Multi-way Analysis in the Food Industry, The Netherlands, University of Amsterdam (1998)
[14] **C. Paulick, M. N. Wright, R. Verleger, K. Keller.** Chemometrics Intell. Lab. Systems, **137** (2014) 97—109
[15] **P. Comon, X. Luciani**, **A. L. F. de Almeida.** J. Chemometrics, **23**, N 7-8 (2009) 393—405
[16] **R. A. Harshman.** UCLA Working Papers in Phonetics, **16** (1970) 1—84
[17] **M. Rajih**, **P. Comon, R. A. Harshman.** SIAM J. Matrix Analysis and Applications, **30**, N 3 (2008) 1128—1147
[18] **N. E. Helwig.** Multiway: Component Models for Multi-Way Data, https://CRAN.R-project.org/package=multiway (дата обращения 24.06.2022)
[19] **W. S. Cleveland, S. J. Devlin.** J. Am. Statist. Ass., **83**, N 403 (1988) 596—610
[20] **R. H. Byrd, P. Lu, J. Nocedal, C. Zhu.** SIAM J. Sci. Comp., **16**, N 5 (1995) 1190—1208
[21] **R. P. Brent.** Algorithms for Minimization without Derivatives, Mineola, New York, Dover Publications (2002)
[22] **C. A. Andersson, R. Bro.** Chemometrics Intell. Lab. Systems, **52**, N 1 (2000) 1—4
[23] **K. R. Murphy, C. A. Stedmon, D. Graeber**, **R. Bro.** Anal. Methods, **5**, N 23 (2013) 6557
[24] **A. Zsolnay**, **E. Baigar**, **M. Jimenez, B. Steinweg, F. Saccomandi.** Chemosphere, **38**, N 1 (1999) 45—50
[25] **A. N. Drozdova, I. N. Krylov, A. A. Nedospasov, E. G. Arashkevich, T. A. Labutin.** Front. Marine Sci., **9** (2022)