

Министерство образования науки Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

А.Н.Баженов, Т.О.Яворук

Тензорные разложения и их применение в
флуориметрии.

Учебное пособие

Санкт-Петербург

2021

УДК 519.9 Р32

А в т о р ы:

А.Н.Баженов, Т.О.Яворук. Тензорные разложения и их применение в флуориметрии. – СПб., 2021. – 47 с.

Учебное пособие соответствует образовательному стандарту высшего образования Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого» по направлению подготовки бакалавров 01.03.02 и магистров 01.04.02 «Прикладная математика и информатика», по дисциплинам «Математическая статистика», «Стохастические модели и анализ данных» и «Обработка сигналов».

Проблематика использования тензорных разложений в последнее десятилетие вышла на качественно новый уровень. Это относится к охвату задач и круга конечных пользователей, поддержке программными средствами, повышению размерности задач, развитию новых математических методов. Всё это мотивирует изучение метода студентами и начинающими исследователями.

В издании приводятся краткие теоретические сведения и практические примеры применения тензорных преобразований в флуориметрии. Намечены пути дальнейших исследований в этой области.

Пособие адресовано всем, кто интересуется применением математики к решению практических задач, в первую очередь, в области анализа многомерных данных.

Материал апробирован в учебных курсах для студентов кафедры «Прикладная математика» Физико-Механического Института СПбПУ.

©Санкт-Петербургский политехнический университет Петра Великого, 2021

Ключевые слова: анализ многомерных данных, тензорные разложения, PARAFAC, Tucker3, ALS, флуориметрия.

СОДЕРЖАНИЕ

1	Введение	5
2	Постановка задачи обработки флуориметрических данных	6
3	Обозначения	10
4	Теория	11
4.1	Произведение Адамара (Шура)	11
4.2	Произведение Кронекера	11
4.3	Тензорное произведение	12
4.4	Произведение Хатри-Рао	13
4.5	k -ранг	13
4.6	Псевдообращение Мура-Пенроуза	13
4.7	PARAFAC	14
4.8	PARAFAC-ALS	16
4.9	Tucker3	18
4.10	Tucker2	19
4.11	Tucker1	19
4.12	PARAFAC2	20
4.13	Единственность разложения	20
4.14	Примеры представления тензоров	21
4.14.1	PARAFAC	21
4.14.2	PARAFAC-ALS	22
4.14.3	Tucker3	23
4.14.4	Tucker2	24
4.14.5	Tucker1	24
4.14.6	PARAFAC2	25
4.15	Теория — заключение	25
5	Результаты	26
5.1	Permafrost	26
5.2	Lake Kivu	30
5.3	Lake Baikal	34
5.4	Обработка данных	39

6	Использованные и разработанные методы и программы	42
7	Заключение	43
	Список литературы	45
	Предметный указатель	47

1 Введение

Учебное пособие посвящено применению тензорных разложений (ТР) к анализу данных. Исторически тензорные разложения сначала были применены к анализу психометрических данных во второй половине XX века. С тех пор эта ветвь анализа получила широкое применение и в других методах анализа данных.

Значительной областью применения ТР стала хемометрика (хемометрия). Подробное обсуждение моделей, методов и практики использования математических приёмов ТР, а также обширная библиография даны в обстоятельном обзоре О.Е.Родионовой и А.Л.Померанцева [1]. Именно в этой области, в силу массового применения методов и аппаратуры, дающих на выходе работы многомерные данные, имеет место большое количество публикаций.

Это, безусловно, положительный момент. При этом, ввиду того, что пользователями этого оборудования являются люди, для которых математика не является основным видом деятельности, математические аспекты методов в публикациях по хемометрике представлены недостаточно подробно. Как правило, приводятся ссылки на использованные методы, и обсуждаются результаты, а не способ их получения и обоснованность использованных методов.

Бурное развитие вычислительной техники и прикладного программного обеспечения также внесло свой вклад в способ изложения в оригинальных публикациях. К весьма краткому изложению математической стороны вопроса добавились ссылки на использование программных средств и библиотек. В результате, из публикаций выпало и изложение собственно вычислительных аспектов работы, которая возложена на общедоступные программные средства. В числе первых тензорные разложения были реализованы средствами Matlab [2] и R [3].

Тем временем, поле применения тензорных преобразований очень сильно расширилось и вышло далеко за рамки психометрии и хемометрии. Наверное, любое перечисление таких применений будет неполным или слишком уж обширным.

Начнём с применений, где данные изначально зависят от двух независимых переменных. Сюда попадают все двумерные объекты, такие как изображения, любые карты, микроскопические изображения, данные, полученные при сканировании двух параметров (линейных, угловых, длин волн, энергий), и т.п. Пространство параметров может быть и частотным. Символически это можно выразить как

$$(X, Y) \equiv 2D,$$

$$X \equiv \{Ex \oplus W \oplus L_x \oplus \dots\}, \quad Y \equiv \{Em \oplus H \oplus L_y \oplus \dots\}$$

Применительно к флуориметрии Ex, Em относятся к спектрам возбуждения и эмиссии образца. Для других применений возможны пары переменных, таких как $W(idth), H(eight); L_x, L_y$.

При добавлении к двумерности ещё одного параметра: нумерации образцов, времени, глубины резкости, увеличения, получаем 3-мерные наборы данных. В качестве конкретных примеров можно привести конфокальную микроскопию, гиперспектральные измерения: от

микроскопических до глобальных объектов, термо-рентгеновские, люминисцентные и другие картирования, различные виды томографии. В таком случае, получаем 3-мерные объекты,

$$(X, Y, Z) \equiv 3D,$$

$$Z \equiv \{Samples \oplus T \oplus H \oplus \dots\}$$

с которыми можно проводить исследования.

Далее следуют применения тензорных методов в машинном обучении, которое создало отдельную область вычислений, и вызвало к жизни многочисленные программные пакеты, такие как `PyTorch`, `TensorFlow`, `TensorLy`. Доступность этих пакетов существенно расширило число пользователей и приложений.

Наконец, более широкое использование суперкомпьютеров повлекло за собой активное повышение размерности решаемых задач. В качестве примера можно привести молекулярную динамику и задачи квантовой химии, где ставится задача решения реалистичного описания материи на низшем уровне организации материального мира. Это существенно многомерные задачи. В последние 15 лет появились методы разложения многомерных тензоров — тензорные поездки (TT, `Tensor Train`), поддерживаемые соответствующими библиотеками [4].

Подводя итог, можно констатировать, что проблематика использования тензорных преобразований в последнее десятилетие вышла на качественно новый уровень по охвату задач и круга конечных пользователей (см., например, электрофизиологические исследования [5]), поддержке программными средствами, повышению размерности задач, развитию новых математических методов. Всё это мотивирует изучение метода студентами и начинающими исследователями.

2 Постановка задачи обработки флуориметрических данных

В нашей публикации мы ставим задачу продемонстрировать использование ТР на примере конкретной проблемы, имеющей место в области глобальной климатологии [8]. Речь идёт о процессах в вечной мерзлоте, которые влияют на климат Земли. Для анализа проб используется метод флуориметрии.

Особенности флуориметрических данных. Суть метода флуориметрии состоит в следующем: растворенная проба сканируется источником света с перестраиваемой длиной волны. Молекулы, содержащиеся в растворе поглощают свет и, в свою очередь, испускают свет, который анализируется по длине волны.

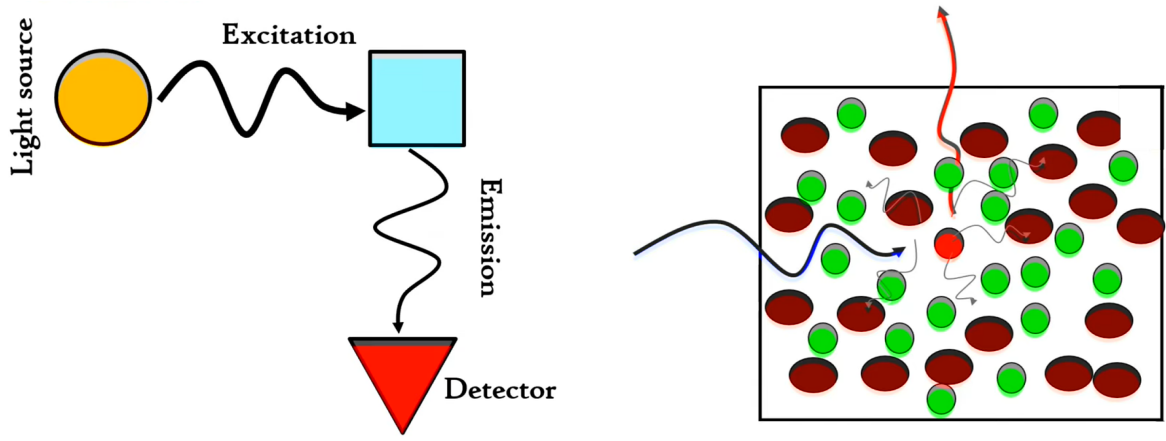
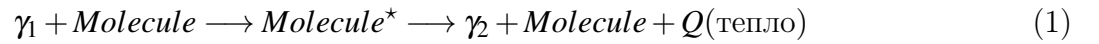


Рис. 1: Принцип флуориметрии. [20]

На левом рисунке представлена схема организации флуориметрических измерений. На правом рисунке показан процесс возбуждения фотоном (темная стрелка) молекулы (красный объект), с последующим излучением более длинноволнового фотона (красная стрелка). Тонкие линии представляют тепловое движение молекул без излучения.

Для единичного фотона процесс возбуждения молекулы образца и последующей эмиссии можно представить следующим образом:



где γ_1, γ_2 — поглощенный и испущенный фотон, *Molecule* — молекула вещества в исходном (основном) состоянии, *Molecule** — молекула вещества в возбуждённом состоянии, *Q* — тепловые потери на нагрев.

Поглощаются и испускаются не одиночные длины волн, а спектры, имеющие достаточно широкие границы. Таким образом, для каждой пробы *k* имеются двумерные данные

$$S_k(X, Y),$$

где *X, Y* — длины волн поглощения и эмиссии, *S_k* — интенсивность излучения.

Ввиду потерь на рассеяние энергии (нагрев) энергия фотонов эмиссии ниже, чем поглощённых: $E(\gamma_1) < E(\gamma_2)$. На языке длин волн этот факт описывается как $\lambda_2 > \lambda_1$ для каждого индивидуального процесса (1). В результате спектр эмиссии смещён в сторону более длинных волн. Символически это можно выразить как отношение порядка в пространстве длин волн

$$X \preceq Y \quad (Ex \preceq Em). \quad (2)$$

Совокупность данных проб конкретно исследуемого объекта

$$\{S_k(X, Y)\}, \quad k = 1, 2, \dots, n \quad (3)$$

образует выборку.

Пример спектров поглощения и эмиссии. Приведём примеры, иллюстрирующие соотношение (2). Наиболее важными являются спектры поглощения и излучения протеиногенных аминокислот тирозина, фенилаланина и триптофана. Эти спектры приведены на рисунке 2.

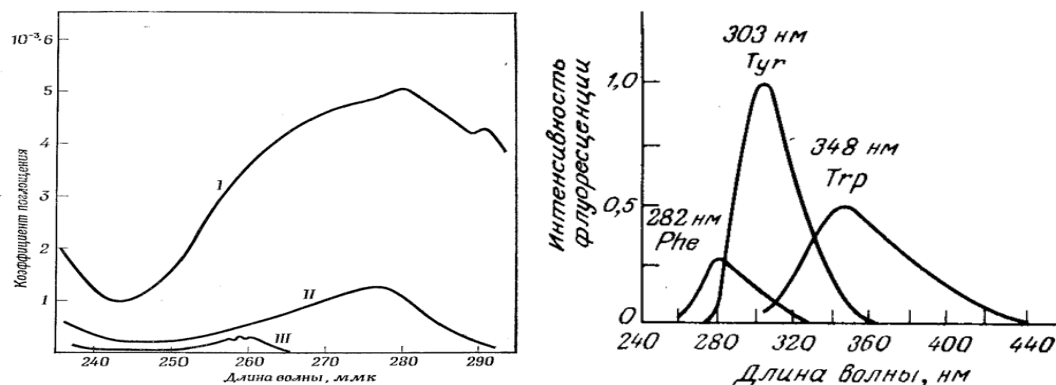


Рис. 2: Спектры поглощения и излучения протеиногенных аминокислот.

Видно, что максимумы спектров поглощения приходятся на длины волн 260-280 нм, а излучения — на 280-350 нм.

Сигнал от органического вещества. Содержательный анализ результатов может быть произведен на основе моделей конкретной предметной области. В случае флуориметрии образцов растворённых органических материалов (DOM — diluted organic matter), искомыми факторами являются концентрации базовых компонент органического материала — биогенные аминокислоты и различного рода вещества, характеризующие природное происхождение материала, такие как гуминовые кислоты. Примеры этих компонент в данных флуориметрии приведены в работе [6].

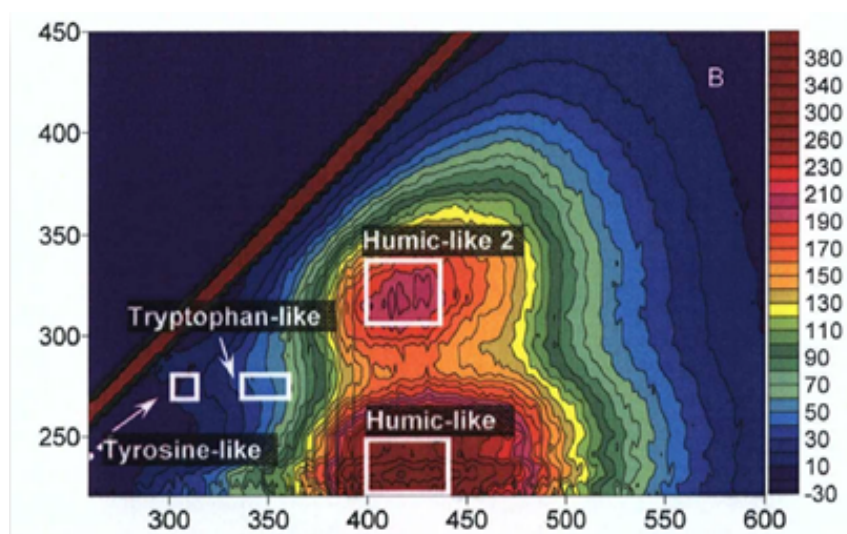


Рис. 3: Спектральные компоненты образцов растворённых органических материалов.

Таким образом, задача состоит в определении спектров эмиссии/возбуждения и их вклада в общий сигнал различных веществ в образцах.

Фон. Помимо полезной информации, в данных, полученных с помощью флуориметрии, присутствуют мешающие сигналы. Они имеют различное происхождение. Во-первых, это случайные изменения сигнала, которые присутствуют для всех длин волн. Этот вид фона существенен только при слабых сигналах аналита. Более существенны сигналы от рассеяния света в растворителе поскольку относительное содержание вещества растворителя существенно превышает концентрацию собственно пробы. Основные из этих процессов — Рэлеевское и Рамановское рассеяние. Они порождают в данных полосы сигналов, которые могут накладываться на полезные сигналы. Есть и другие явления, которые влияют на численные результаты, такие как поглощение пробой в случае высоких концентраций.

Математические аспекты задачи. Рассмотрим математические и образовательные аспекты нашей работы.

Тензорные разложения применяются для определения факторов n -мерных массивов данных. Конкретным ТР, используемым в нашей работе, является параллельный факторный анализ, PARAFAC. PARAFAC предназначен для определения факторов-нагрузок в трёхмерных массивах данных. В хемометрии полученные факторы имеют определённый смысл: спектры эмиссии, спектры возбуждения и относительной концентрации веществ в образцах [10]. Сами образцы представляют собой двумерные массивы. Последовательность таких образцов образует трёхмерный массив - тензор третьего порядка.

Преимущество PARAFAC модели заключается в единственности разложения т.е. отсутствии свободы вращения [11], [12] и способности обработки зашумлённых данных.

PARAFAC применяется и в других областях. В психометрии набор данных - ответы на вопросы, нескольких субъектов, опрошенных несколько раз. Так же PARAFAC можно использовать в анализе изображений — определение изменяющихся составляющих и т.д. Однако, более популярным его использованием является определение состава смеси веществ.

Целью работы является определение спектров компонент эмиссии и возбуждения и определение ранга разложения по данным, полученным с помощью флуориметра из разных географических регионов: Вечная мерзлота (Permafrost), Африка (Lake Kivu), Азия (Lake Baikal).

Задачи:

1. Привести собственные примеры представлений тензоров;
2. Описать алгоритм получения компонент;
3. Реализовать один из алгоритмов получения компонент;
4. Получить компоненты разложения предоставленных данных;
5. Проанализировать полученные результаты.

3 Обозначения

Будем использовать обозначения в основном как в книге [13].

a - скаляр

a - вектор

A - матрица

a_i - элементы-вектора матрицы или элементы-скаляры вектора

a_{ij} - элементы-скаляры матрицы

$\underline{A}$ - тензор

$\otimes$ - произведение Кронекера

$\triangle$ - тензорное произведение

$\circ$ - произведение Адамара

i, j, k, p, q, r - индексы

I, J, K, P, Q, R - максимальные значения индексов

4 Теория

Приведём характеристику применения тензорных разложений из обзора [9].

На рубеже тысячелетий использование форматов данных с тензорной структурой было признано основной концепцией для преодоления “проклятия размерности” в многомерном численном моделировании. Руководящим принципом тензорных методов является аппроксимация многомерных функций и операторов, основанных на выбранном заранее разделении переменных и выполнении вычислений в тензорных структурах низкой размерности.

Современные приложения тензорных методов включают сложные многомерные задачи, возникающие в науках о материалах, биологии, стохастическом моделировании, обработке сигналов, машинном обучении и интеллектуальном анализе данных, финансовой математике и т.д.

Обзор и исчерпывающая библиография по традиционным инструментам полилинейной аппроксимации в информатике, основанной на моделях Таккера и других, представлены в [12], [13].

Будем использовать обозначения в основном как в книге [13]. Перейдём к основным понятиям.

4.1 Произведение Адамара (Шура)

Произведение Адамара представляет собой поэлементное произведение всех соответствующих составляющих двух матриц A и B одинаковой размерности. Пример:

$$C = A \circ B = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \circ \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{31}b_{31} & a_{32}b_{32} & a_{33}b_{33} \end{pmatrix}$$

Произведение Адамара используется в алгоритмах сжатия с потерями (JPEG), и в различных программных средах, в том числе на графических процессорах.

4.2 Произведение Кронекера

Это бинарная операция над матрицами произвольного размера $A = A[m \times n]$ и $B = B[p \times q]$. Результатом является блочная матрица $C = C[mp \times nq]$, каждый блок которой - это вторая матрица умноженная на определённый элемент первой матрицы.

$$C = A \otimes B = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & a_{13}B \\ a_{21}B & a_{22}B & a_{23}B \\ a_{31}B & a_{32}B & a_{33}B \end{pmatrix}$$

4.3 Тензорное произведение

Введём тензорное произведение для векторов и матриц [13]. Тензорное произведение очень похоже на произведение Кронекера (и на тензорное произведение для графов) и обозначаются они обычно одним и тем же символом, но в контексте публикации они будут различаться и применяться только для трёх векторов - триады. Такая триада так же называется трёхмерным тензором единичного ранга (rank-one tensor). Используемое обозначение — $\triangle$.

$$B = a \triangle b = a \otimes b^T = ab^T = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \begin{pmatrix} b_1 & b_2 & b_3 \end{pmatrix} = \begin{pmatrix} a_1 b_1 & a_1 b_2 & a_1 b_3 \\ a_2 b_1 & a_2 b_2 & a_2 b_3 \\ a_3 b_1 & a_3 b_2 & a_3 b_3 \end{pmatrix}$$

Тензорное произведение матрицы и вектора - это тензор третьего порядка, представленный в виде матриц-срезов.

$$\underline{X} = A \triangle c = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \triangle \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

$$X_i = c_i A = \begin{pmatrix} c_i a_{11} & c_i a_{12} & c_i a_{13} \\ c_i a_{21} & c_i a_{22} & c_i a_{23} \\ c_i a_{31} & c_i a_{32} & c_i a_{33} \end{pmatrix}, \quad i = \overline{1, 3}$$

$$\underline{Z} = a \triangle D = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \triangle \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \triangle \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$$

$$Z_i = \begin{pmatrix} a_1 d_{1i} & a_1 d_{2i} & a_1 d_{3i} \\ a_2 d_{1i} & a_2 d_{2i} & a_2 d_{3i} \\ a_3 d_{1i} & a_3 d_{2i} & a_3 d_{3i} \end{pmatrix}, \quad i = \overline{1, 3}$$

$$\underline{Y} = B \triangle c$$

$$Y_i = \begin{pmatrix} a_1 b_1 c_i & a_1 b_2 c_i & a_1 b_3 c_i \\ a_2 b_1 c_i & a_2 b_2 c_i & a_2 b_3 c_i \\ a_3 b_1 c_i & a_3 b_2 c_i & a_3 b_3 c_i \end{pmatrix} = c_i ab^T$$

$$Y = \begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix}$$

Стоит заметить, что матрицу, полученную с помощью произведения Кронекера, можно интерпретировать как тензор [14].

Замечание (об определениях). Порядком тензора назовём его размерность (скаляр - нуле-

вой порядок, вектор - первый порядок, матрица - второй, ...). Понятие ранга используется для количества единичных тензоров на которые можно разложить исходный тензор. Конец замечания.

4.4 Произведение Хатри-Рао

Произведение Хатри-Рао - произведение Кронекера для всех соответствующих столбцов:

$$C = A \odot B = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \odot \begin{pmatrix} b_1 & b_2 & b_3 \end{pmatrix} = \begin{pmatrix} a_1 \otimes b_1 & a_2 \otimes b_2 & a_3 \otimes b_3 \end{pmatrix}$$

4.5 k -ранг

k -ранг матрицы - это максимальное количество столбцов, которые можно случайно выбрать, и они всегда окажутся линейно независимыми (ЛНЗ).

Пример [13]:

$$A = \begin{pmatrix} 1 & 1 & 2 & 5 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 4 & 2 \end{pmatrix}; \quad \text{rank}(A) = 3, \quad k_A = 1.$$

4.6 Псевдообращение Мура-Пенроуза

Псевдообращение Мура-Пенроуза является обобщением обратной матрицы.

Для $A = A_{m \times n}$ псевдообратная к A определяется как $A^+ = A_{n \times m}^+$ удовлетворяет следующим условиям:

1. $AA^+A = A$
 AA^+ отображает векторы-столбцы матрицы A на себя
2. $A^+AA^+ = A^+$
3. $(AA^+)^* = AA^+$, AA^+ - эрмитова ($C^T = \overline{C}$)
4. $(A^+A)^* = (A^+A)$, A^+A - тоже эрмитова.

Пример:

$$y = Xb + e$$

$$\min_b \|y - Xb\|^2$$

Решение: $b = X^+y$.

Применим SVD разложение $X = USV^T$. Тогда получим

$$X^+ = VS^+U^T,$$

$$\text{где } s_{ii}^+ = \begin{cases} 1/s_{ii}, & s_{ii} \neq 0, \\ 0, & s_{ii} = 0. \end{cases}$$

4.7 PARAFAC

Идея тензорного разложения заключается в представлении исходного тензора $\underline{X}$ в виде суммы триад. Минимальное количество триад, на которые можно разложить тензор, называется рангом тензора.

При определении такого разложения возникает несколько проблем: определение ранга тензора и получение векторов триад. Может оказаться, что ранг разложения известен заранее, и тогда первая проблема исчезает, если нет - то это предмет исследования. Некоторые выводы можно сделать если разделить данные на группы (обычно две), проделать одинаковые манипуляции и сравнить результаты. Иногда, когда важна точность в описании данных (т.е. в них нет шума или он не существен), некоторым критерием определения ранга могут служить аномалии в поведении нормы тензора ошибок. Для решения второй проблемы существует алгоритм ALS, есть и другие алгоритмы, но в данной работе рассматривается только этот. Данный алгоритм подразумевает описание тензора с помощью векторов с некоторой точностью. Для определения разности между полученными элементами и элементами исходных данных вводится тензор ошибок:

$$\underline{E} = \underline{X} - \hat{\underline{X}},$$

$\hat{\underline{X}}$ — тензор, полученный с помощью алгоритма.

Стоит заметить, что в реальных данных $\underline{E}$ имеет существенную норму.

Псевдо-рангом тензора называется ранг тензора $\hat{\underline{X}}$ или, что то же самое - количество полученных триад, описывающих данные.

В итоге исходный тензор можно представить в виде:

$$\underline{X} = \sum_{r=1}^R a_r \triangle b_r \triangle c_r + \underline{E},$$

где a_r, b_r, c_r - вектора, $\underline{E}$ - тензор ошибок, R - псевдо-ранг тензора.

Существуют и другие способы представления PARAFAC разложения.

Пусть A - матрица, составленная из векторов a_r , B - матрица, составленная из векторов b_r , C - матрица, составленная из векторов c_r , k - номер фронтального среза тензора, на диагонали матрицы D_k стоит k -ая строка матрицы C , остальные элементы - нули, $X = \begin{pmatrix} X_1 & X_2 & \dots & X_K \end{pmatrix}$, $E = \begin{pmatrix} E_1 & E_2 & \dots & E_K \end{pmatrix}$ - матрицы, составленные из последовательных фронтальных срезов.

$$1. X_k = AD_k B^T$$

Пример:

$$A = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \quad B = \begin{pmatrix} b_1 & b_2 \end{pmatrix} \quad C = \begin{pmatrix} c_1 & c_2 \end{pmatrix}$$

$$X_i = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} c_{1i} & 0 \\ 0 & c_{2i} \end{pmatrix} \begin{pmatrix} b_1^T \\ b_2^T \end{pmatrix} = c_{1i}a_1b_1^T + c_{2i}a_2b_2^T$$

2. $X = AD(I \otimes B^T)$

Пример:

Пусть количество фронтальных срезов $K = 3$.

$$\begin{aligned} X &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} c_{11} & 0 & c_{12} & 0 & c_{13} & 0 \\ 0 & c_{21} & 0 & c_{22} & 0 & c_{23} \end{pmatrix} \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} b_1^T \\ b_2^T \end{pmatrix} \right] = \\ &= \begin{pmatrix} a_1c_{11} & a_2c_{21} & a_1c_{12} & a_2c_{22} & a_1c_{13} & a_2c_{23} \end{pmatrix} \begin{pmatrix} b_1^T & 0 & 0 \\ b_2^T & 0 & 0 \\ 0 & b_1^T & 0 \\ 0 & b_2^T & 0 \\ 0 & 0 & b_1^T \\ 0 & 0 & b_2^T \end{pmatrix} = \\ &= \begin{pmatrix} c_{11}a_1b_1^T + c_{21}a_2b_2^T & c_{12}a_1b_1^T + c_{22}a_2b_2^T & c_{13}a_1b_1^T + c_{23}a_2b_2^T \end{pmatrix} \end{aligned}$$

3. $X = AH(C^T \otimes B^T)$,

$H = I_{R \times RR}$ - последовательные срезы тензора $\underline{I}$, на супердиагонали которого стоят единицы.

Пример:

Количество фронтальных срезов $K = 3$.

$$\begin{aligned} X &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right) \left[\begin{pmatrix} c_1^T \\ c_2^T \end{pmatrix} \otimes \begin{pmatrix} b_1^T \\ b_2^T \end{pmatrix} \right] = \\ &= \begin{pmatrix} a_1 & 0 & 0 & a_2 \end{pmatrix} \begin{pmatrix} c_{11}b_1^T & c_{12}b_1^T & c_{13}b_1^T \\ c_{11}b_2^T & c_{12}b_2^T & c_{13}b_2^T \\ c_{21}b_1^T & c_{22}b_1^T & c_{23}b_1^T \\ c_{21}b_2^T & c_{22}b_2^T & c_{23}b_2^T \end{pmatrix} = \\ &= \begin{pmatrix} c_{11}a_1b_1^T + c_{21}a_2b_2^T & c_{12}a_1b_1^T + c_{22}a_2b_2^T & c_{13}a_1b_1^T + c_{23}a_2b_2^T \end{pmatrix} \end{aligned}$$

4. $X = A(C \odot B)^T$

Пример:

Количество фронтальных срезов $K = 3$.

$$\begin{aligned} X &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \left[\begin{pmatrix} c_1 & c_2 \end{pmatrix} \odot \begin{pmatrix} b_1 & b_2 \end{pmatrix} \right]^T = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} c_{11}b_1 & c_{21}b_2 \\ c_{12}b_1 & c_{22}b_2 \\ c_{13}b_1 & c_{23}b_2 \end{pmatrix}^T = \\ &= \begin{pmatrix} c_{11}a_1b_1^T + c_{21}a_2b_2^T & c_{12}a_1b_1^T + c_{22}a_2b_2^T & c_{13}a_1b_1^T + c_{23}a_2b_2^T \end{pmatrix} \end{aligned}$$

4.8 PARAFAC-ALS

Алгоритм ALS заключается в решении задачи оптимизации: минимизации тензора ошибок.

Для двулинейной модели:

$$\min_{A,B} \|X - AB^T\|^2$$

1. init A , call it $\hat{A}$
2. calc $\hat{B} = X(\hat{A}^+)^T$
3. calc $\hat{A} = X(\hat{B}^+)^T$
4. until convergence. Then $B = \hat{B}$, $A = \hat{A}$

Для трилинейной модели:

$$\min_{A,B,C} \|X - \hat{X}\|^2,$$

$$\hat{X} = \hat{X}(A, B, C)$$

1. Init param
2. $\hat{A} = \arg \min_S \|X - \hat{X}(S, B, C)\|^2$
3. $\hat{B} = \arg \min_T \|X - \hat{X}(\hat{A}, T, C)\|^2$
4. $\hat{C} = \arg \min_U \|X - \hat{X}(\hat{A}, \hat{B}, U)\|^2$
5. go to 2 until convergence

Замечание. В некоторых случаях могут быть известны некоторые свойства компонент, которые можно (и нужно) учитывать при их определении.

Важным свойством в получении спектров эмиссии, возбуждения и концентраций является неотрицательность компонент. В таком случае в задачу оптимизации нужно добавить условие неотрицательности на составляющие матриц. Конец замечания.

PARAFAC:

$$\min_{A,B,C} \|X - A(C \odot B)^T\|^2$$

Обозначим $Z = C \odot B$, в итоге перешли к двулинейной модели, отсюда $A = X(Z^T)^+ = XZ(Z^T Z)^{-1}$.
 B и C - аналогично.

1. Init B and C
2. $Z = C \odot B$
 $A = X_{I \times JK} Z (Z^T Z)^{-1}$
3. $Z = C \odot A$
 $B = X_{J \times IK} Z (Z^T Z)^{-1}$
4. $Z = B \odot A$
 $C = X_{K \times JI} Z (Z^T Z)^{-1}$
5. go to step 2 until convergence

Но это не очень удобное представление для программирования. Матрицы, составленные из фронтальных срезов, могут иметь большие размеры. Поэтому заметим:

$$XZ = \begin{pmatrix} X_1 & \dots & X_K \end{pmatrix} \begin{pmatrix} D_1 B^T \\ \dots \\ D_K B^T \end{pmatrix} = X_1 B D_1 + \dots + X_K B D_K$$

$$Z^T Z = (C \odot B)^T (C \odot B) = (C^T C) \odot (B^T B).$$

Перепишем алгоритм:

1. Init B and C
2. $A = (\sum_{k=1}^K X_k B D_k) ((C^T C) \odot (B^T B))^{-1}$
3. $B = (\sum_{k=1}^K X_k A D_k) ((C^T C) \odot (A^T A))^{-1}$
4. $d_k = ((B^T B) \odot (A^T A))^{-1} ((A^T X_k B) \odot I) \mathbf{1}, k = 1, \dots, K$
5. go to step 2 while $\|E\| > \varepsilon$ or $\|\hat{X}_i - \hat{X}_{i-1}\| > \varepsilon$

$\|E\|$ - Евклидова норма, d_k - k -ая строка матрицы C , $I = I_{R \times R}$ - единичная матрица,
 $\mathbf{1} = \mathbf{1}_{R \times 1}$ - вектор, состоящий из единиц.

Начальное приближение можно выбрать случайное или из соображений сингулярных векторов. Существуют две проблемы, возникающие в связи с этим: попадание в локальный минимум и медленная скорость сходимости. Первую можно попробовать решить, выбрав разные начальные приближения. Появление второй может быть связано с сильно коррелированными исходными данными или их отсутствием.

4.9 Tucker3

Существует семейство тензорных разложений Tucker [13]. PARAFAC является частным случаем тензорного разложения Tucker3.

Триадная форма записи Tucker3:

$$\underline{X} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} a_p \Delta b_q \Delta c_r$$

Матричная форма записи Tucker3:

$$X = AG(C \otimes B)^T,$$

где $X = X_{I \times J \times K} = (X_1 \ X_2 \ \dots \ X_K)$, $A = A_{I \times P}$, $B = B_{Q \times J}$, $C = C_{R \times K}$,

$$\underline{G} = G_{P \times Q \times R} = (G_1 \ G_2 \ \dots \ G_R)$$

$\underline{G}$ — ядро тензорного разложения (core tensor).

Заметим, что $\underline{G}$ может не быть единичным, диагональным или кубическим. В PARAFAC

$$\underline{G} = \underline{I}$$

— тензор, на супер диагонали которого стоят единицы.

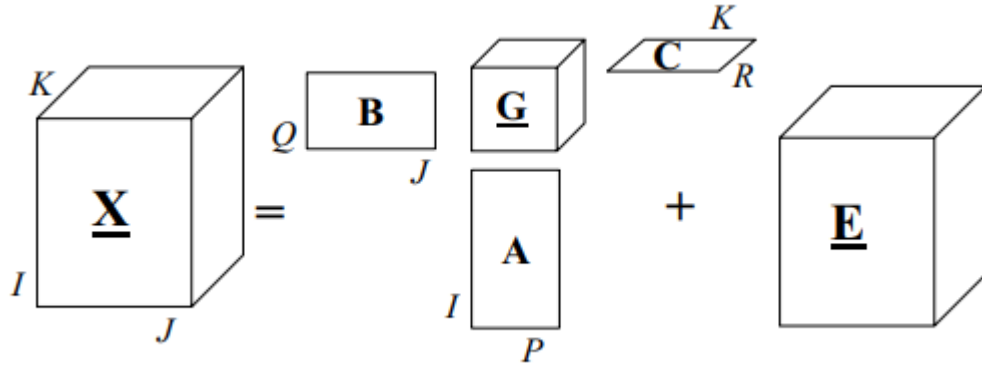


Рис. 4: Tucker3. [13]

Без дополнительных условий не обладает свойством единственности разложения (обладает свободой вращения):

$$X = AG(C \otimes B)^T = APP^{-1}G(C \otimes B)^T = \tilde{A}\tilde{G}(C \otimes B)^T$$

4.10 Tucker2

Триадная форма записи:

$$\underline{X} = \sum_{p=1}^P \sum_{q=1}^Q a_p \triangle b_q \triangle g_{pq}$$

g_{pq} - вектор (tube) тензора $\underline{G}$

Одна из матричных форм записи:

$$X_{I \times JK} = A G_{P \times QR} (I \otimes B)^T$$

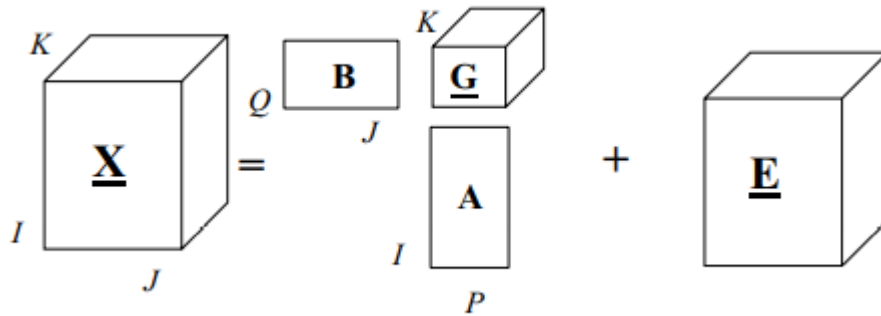


Рис. 5: Tucker2. [13]

4.11 Tucker1

Триадная форма записи:

$$\underline{X} = \sum_{p=1}^P a_p \triangle g_p$$

a_p - вектор, g_p - матрица (фронтальный срез тензора $\underline{G}$)

Одна из матричных форм записи:

$$X_{I \times JK} = A G_{P \times QR}$$

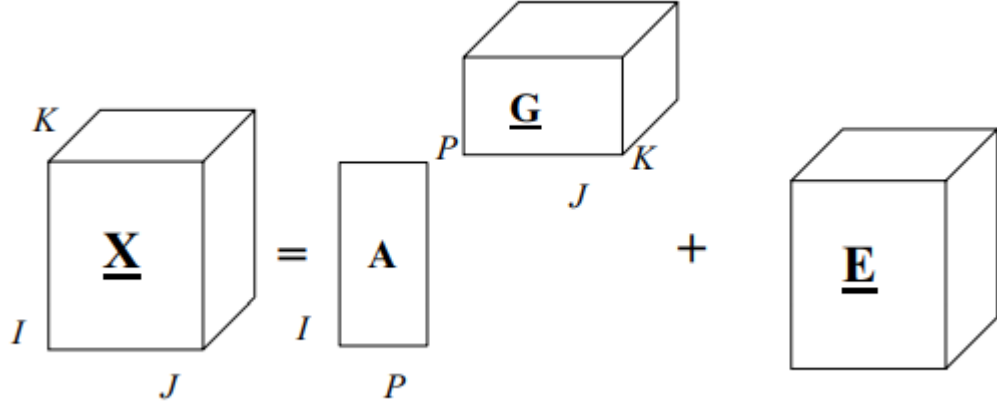


Рис. 6: Tucker1

4.12 PARAFAC2

Опишем представление исходного тензора. Допустим, существует K матриц X_k (образцы, экземпляры), каждая из них имеет J столбцов (какие-то переменные) и I_k строк (наблюдения). Причём у разных матриц количество строк может отличаться. Для таких данных применим PARAFAC2.

Матричное представление:

$$X_k = U_k S_k V^T + E_k,$$

где $U_k = (U_k)_{I_k \times R}$, $S_k = (S_k)_{R \times R}$ - диагональная, $V = V_{J \times R}$

В [18] описан алгоритм ALS для получения разложения, [19] приведён алгоритм разложения и условия единственности разложения. Без дополнительных условий [12] не обладает свойством единственности разложения.

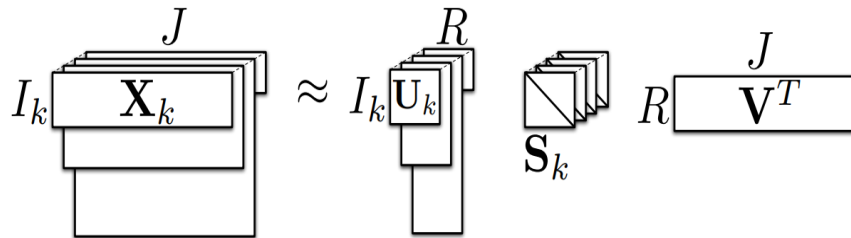


Рис. 7: PARAFAC2 [18]

4.13 Единственность разложения

Единственность подразумевается без учёта масштабирования или перестановок компонент матриц, т.е. элементы матриц, отвечающие одним компонентам сохраняют пропорции между элементами, и порядок записи компонент в матрицах неважен (как сумма слагаемых).

Достаточное условие для того чтобы тензорное разложение PARAFAC было единственным:

$$k_A + k_B + k_C \geq 2R + 2,$$

k_A, k_B, k_C - k -ранги соответствующих матриц, R - количество компонент PARAFAC подробнее см. [11] и [15].

4.14 Примеры представления тензоров

Тензор, представленный в виде фронтальных срезов:

$$X_1 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

4.14.1 PARAFAC

Опишем его как сумму двух триад:

$$\underline{X} = a_1 \triangle b_1 \triangle c_1 + a_2 \triangle b_2 \triangle c_2$$

где a_1, b_1 и a_2, b_2 - вектора, описывающие компоненты-матрицы, c_1, c_2 - вектора концентрации:

$$a_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad b_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad c_1 = \begin{pmatrix} 1 \\ 0 \\ 1/2 \end{pmatrix}$$

$$a_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad b_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad c_2 = \begin{pmatrix} 0 \\ 1 \\ 1/2 \end{pmatrix}$$

Матричные формы записи:

$$\begin{pmatrix} X_1 & X_2 & X_3 \end{pmatrix} = A I_{2 \times 2 \times 2} (C \otimes B)^T = A (C \odot B)^T,$$

где $A = \begin{pmatrix} a_1 & a_2 \end{pmatrix}$, $B = \begin{pmatrix} b_1 & b_2 \end{pmatrix}$, $C = \begin{pmatrix} c_1 & c_2 \end{pmatrix}$,
 $I = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right)$ - развёртка тензора, на главной диагонали которого стоят единицы.

4.14.2 PARAFAC-ALS

Начальное приближение:

$$B = \begin{pmatrix} 0.5 & 1 \\ 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0.5 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}$$

Зависимость точности от количества итераций:

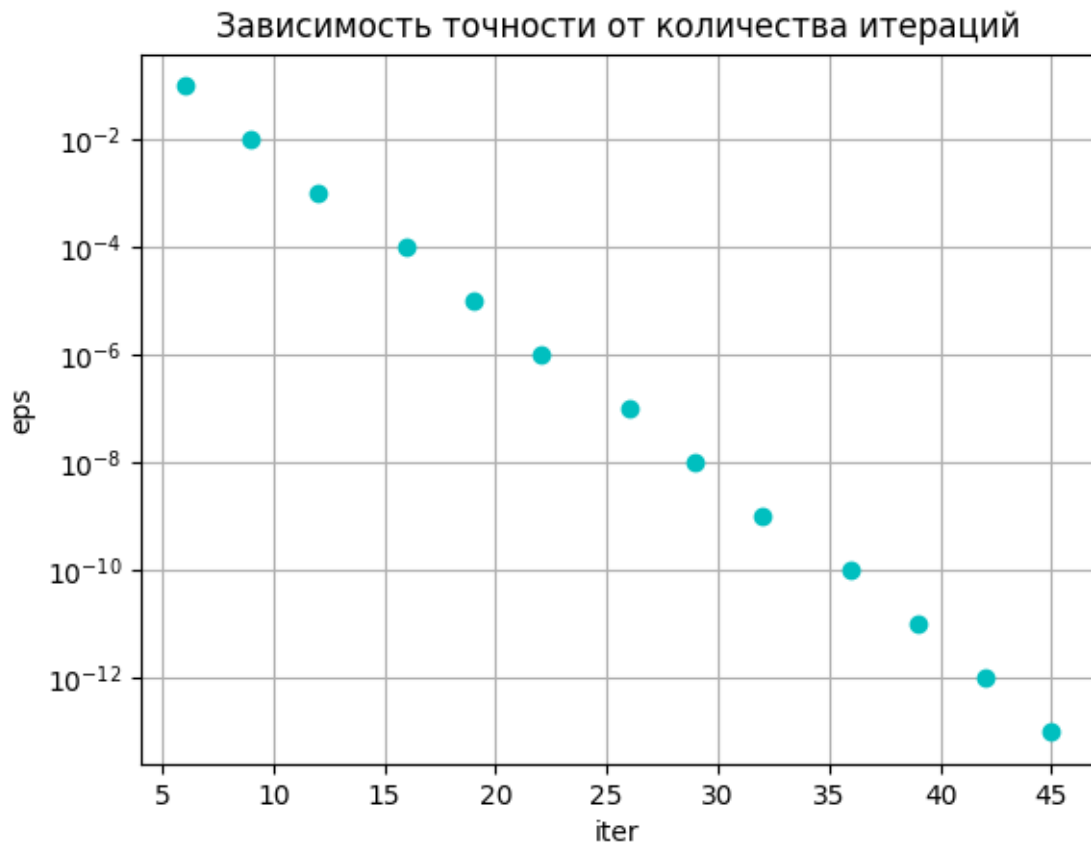


Рис. 8: Зависимость точности от количества итераций

Результат:

$$A = \begin{pmatrix} 5.88e-01 & -8.39e-15 \\ 5.88e-01 & 3.63e-01 \\ 1.36e-14 & 3.63e-01 \end{pmatrix} \quad B = \begin{pmatrix} 9.60e-01 & -7.51e-14 \\ 9.60e-01 & 1.83e+00 \\ 3.93e-14 & 1.83e+00 \end{pmatrix}$$

$$C = \begin{pmatrix} 1.76e+00 & -5.24e-14 \\ 6.16e-14 & 1.49e+00 \\ 8.84e-01 & 7.49e-01 \end{pmatrix}$$

Полученный тензор:

$$X_1 = \begin{pmatrix} 1. & 1. & 4.1e-14 \\ 1. & 1. & 6.1e-15 \\ 2.3e-14 & -1.2e-14 & -3.5e-14 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 3.5e-14 & 1.2e-14 & -2.3e-14 \\ -6.1e-15 & 1. & 1. \\ -4.1e-14 & 1. & 1. \end{pmatrix}$$

$$X_3 = \begin{pmatrix} 0.50 & 0.5 & 8.9e-15 \\ 0.50 & 1.0 & 0.50 \\ -8.9e-15 & 0.5 & 0.50 \end{pmatrix}$$

4.14.3 Tucker3

Предыдущий пример в одном из представлений Tucker3:

$$(X_1 \ X_2 \ X_3) = A G_{2 \times 2 \times 2} (C \otimes B)^T = A (C \odot B)^T,$$

где $A = \begin{pmatrix} a_1 & a_2 \end{pmatrix}$, $B = \begin{pmatrix} b_1 & b_2 \end{pmatrix}$, $C = \begin{pmatrix} c_1 & c_2 \end{pmatrix}$, G - core-array

$$G = \left(\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array} \right)$$

Пример с не кубическим $\underline{G}$:

$$X = (X_1 \ X_2 \ X_3) = \left(\begin{array}{ccc|ccc|ccc} 0 & 2 & 0 & 1 & 2 & 1 & 1 & 3 & 1 \\ 0 & 0 & 2 & 1 & 1 & 2 & 1 & 1 & 3 \end{array} \right)$$

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad C = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad G = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array} \right)$$

Триадная форма записи:

$$\underline{X} = g_{111}a_1\triangle b_1\triangle c_1 + g_{231}a_2\triangle b_3\triangle c_1 + g_{112}a_1\triangle b_1\triangle c_2 + g_{322}a_3\triangle b_2\triangle c_2 + g_{232}a_2\triangle b_3\triangle c_2$$

PARAFAC разложение:

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \end{pmatrix}$$

$$B = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \end{pmatrix}$$

$$C^T = \begin{pmatrix} 2 & 2 & 3 \\ -2 & -1 & -2 \\ -2 & -1 & -2 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} c_1^T \\ c_2^T \\ c_3^T \\ c_4^T \end{pmatrix}$$

4.14.4 Tucker2

Первый пример в одном из представлении Tucker2:

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$G = \left(\begin{array}{cc|cc|cc} 1 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0.5 \end{array} \right)$$

4.14.5 Tucker1

Первый пример в одном из представлений Tucker1:

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$G = \left(\begin{array}{ccc|ccc|ccc} 1 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0.5 & 0.5 \end{array} \right)$$

4.14.6 PARAFAC2

Исходный тензор:

$$X_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad X_2 = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix} \quad X_3 = \begin{pmatrix} 2 & 2 & 0 \end{pmatrix}$$

Одно из PARAFAC2 представлений:

$$U_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad U_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad U_3 = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$S = \left(\begin{array}{cc|cc|cc} 1 & 0 & 0.5 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0.5 & 0 & 2 \end{array} \right)$$

$$V = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

4.15 Теория — заключение

Тензорные разложения дают большую гибкость в представлении данных. Выбор конкретного вида разложения существенно зависит от характера данных и способа их содержательной интерпретации.

5 Результаты

Перейдём к результатам исследования конкретных данных. Наиболее значимые с точки зрения специалистов являются образцы вечной мерзлоты [8] — раздел 5.1. Другие типы образцов, из озёр разных регионов планеты, разделы 5.2 и 5.3, представляют интерес с точки зрения математических аспектов обработки данных. Например, в них может быть другое соотношение сигнала и фона, количество пиков полезного сигнала и т.п.

5.1 Permafrost

Данные:

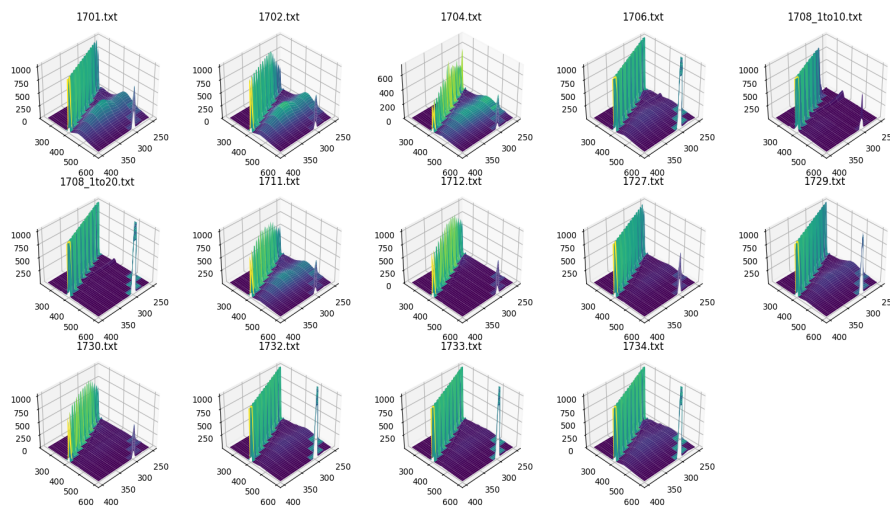


Рис. 9: Исходные данные

На данных (см. Рис. 9) можно заметить пики Рэля с коэффициентами наклона $k = 1$; $k = 2$ и Рамана с коэффициентами наклона $k = 1.12$; $k = 2.2$ (см. страницу 9). Они мешают определению спектров эмиссии и возбуждения. Заменяем значения этих прямых на средние значения стоящих рядом точек. Ширина этих прямых - характеристика прибора.

Преобразованные данные, вид сверху:

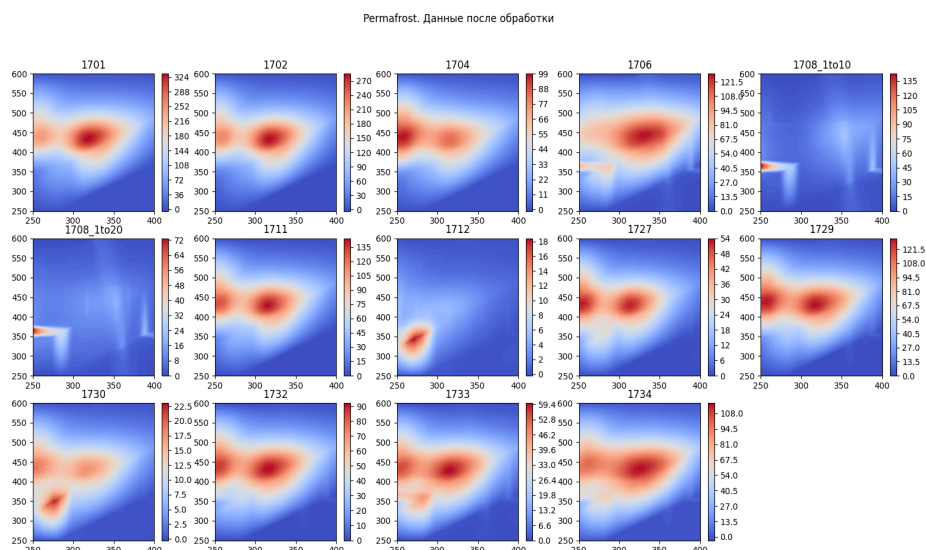


Рис. 10: Преобразованные исходные данные

Применим PARAFAC разложение на все данные с разными рангами разложения:



Рис. 11: Permafrost. Все данные. Разные ранги

Заметим, что для ранга равного шести (см. Рис. 11) пики некоторых спектров совпадают, и концентрации ведут себя одинаково, что может говорить о слишком большом ранге.

Рассмотрим поведение ошибки при разных рангах и точности:

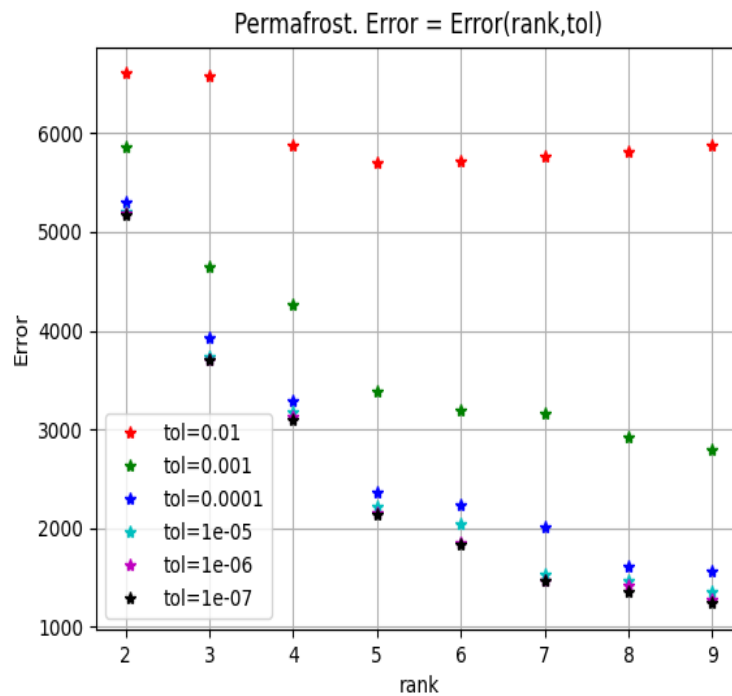


Рис. 12: Permafrost. Поведение ошибки

На рисунке 12 можно заметить anomальное поведение ошибки при меньшей точности, что может говорить о возможном ранге разложения. Изобразим компоненты при рангах, равных пяти и четырём. В левой части - компоненты, полученные после обработки только пиков Рэлея, в правой отсутствуют пики Рэлея и Рамана.

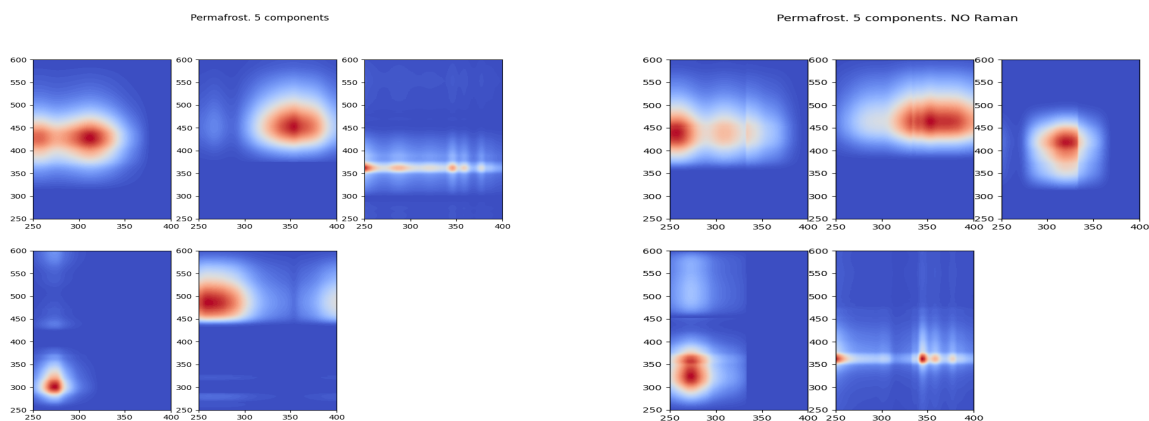


Рис. 13: Permafrost. Компоненты. Rank = 5

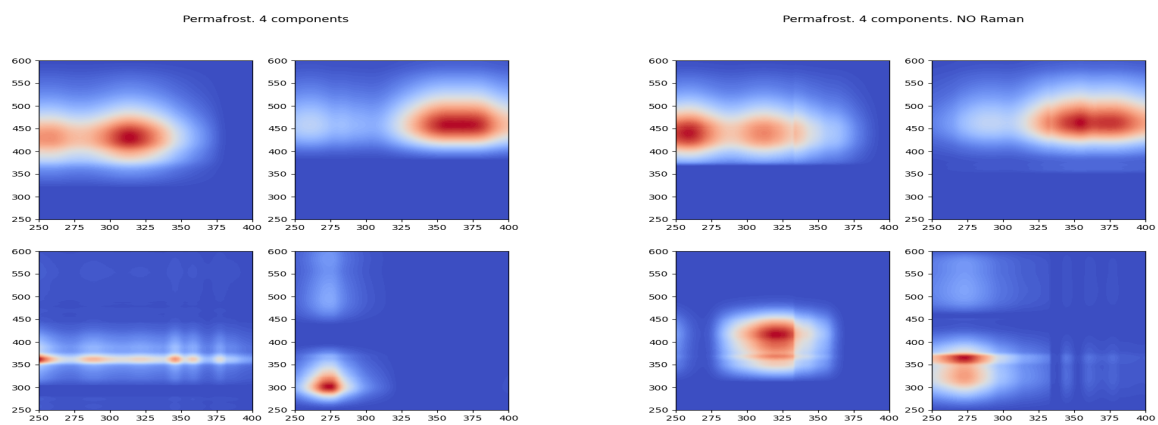


Рис. 14: Permafrost. Компоненты. Rank = 4

5.2 Lake Kivu

Исходные данные после обработки - отсечение пиков Рэля и Рамана:

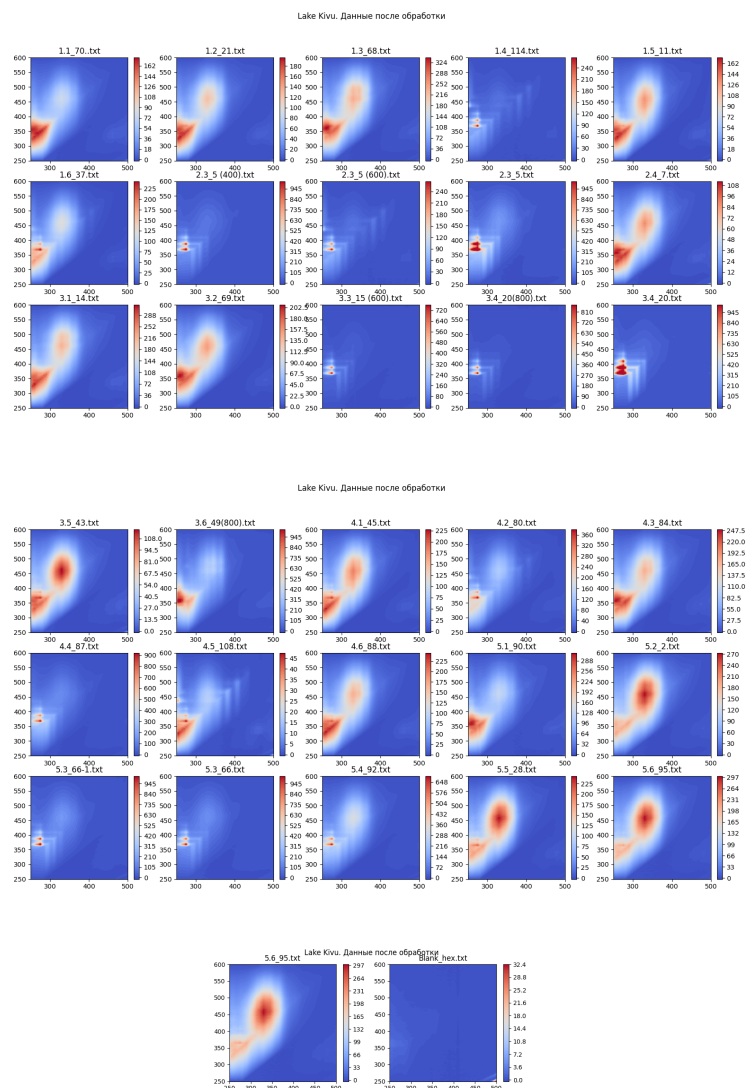


Рис. 15: Kivu. Исходные данные без пиков Рэля и Рамана

Последний файл не рассматривается далее, т.к. он содержит только артефакты - пики Рэля, Рамана и другие.

Применим PARAFAC для разделения компонент.

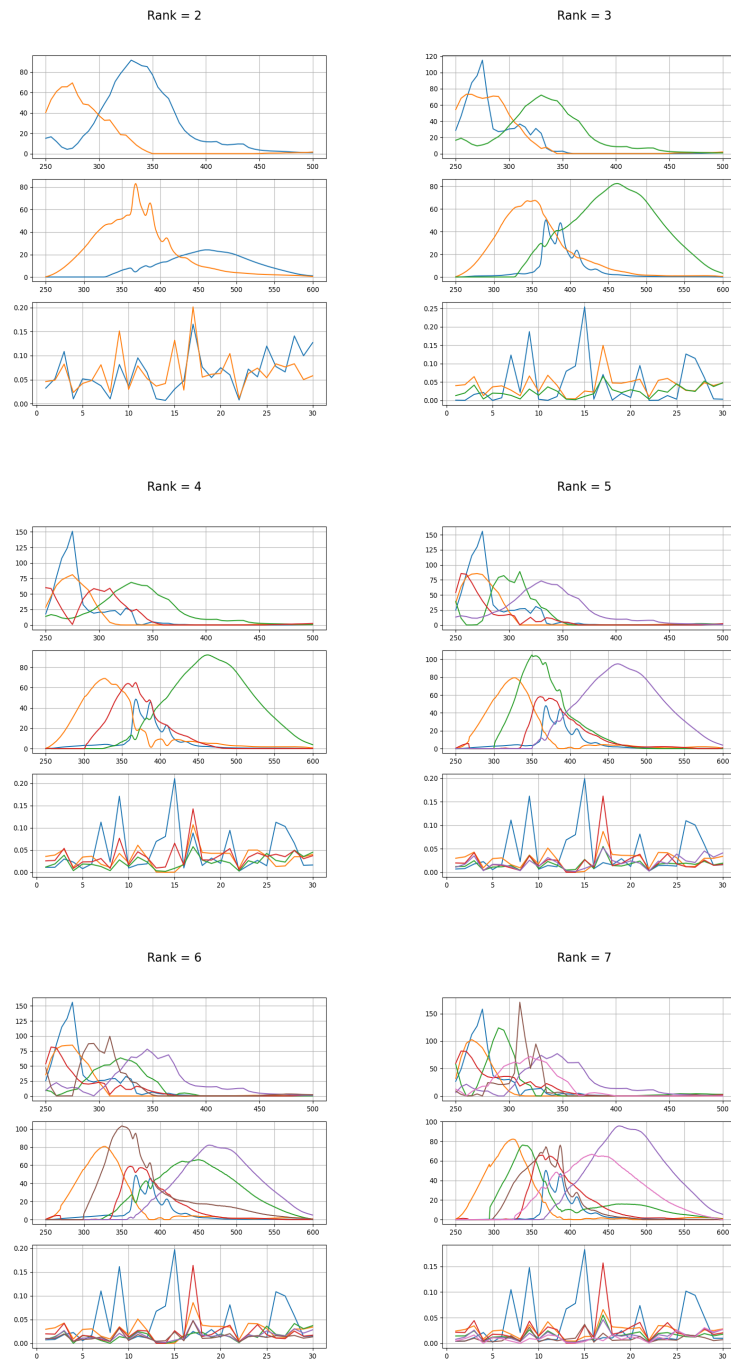


Рис. 16: Kivu. Все данные. Разные ранги

Последим за поведением ошибки:

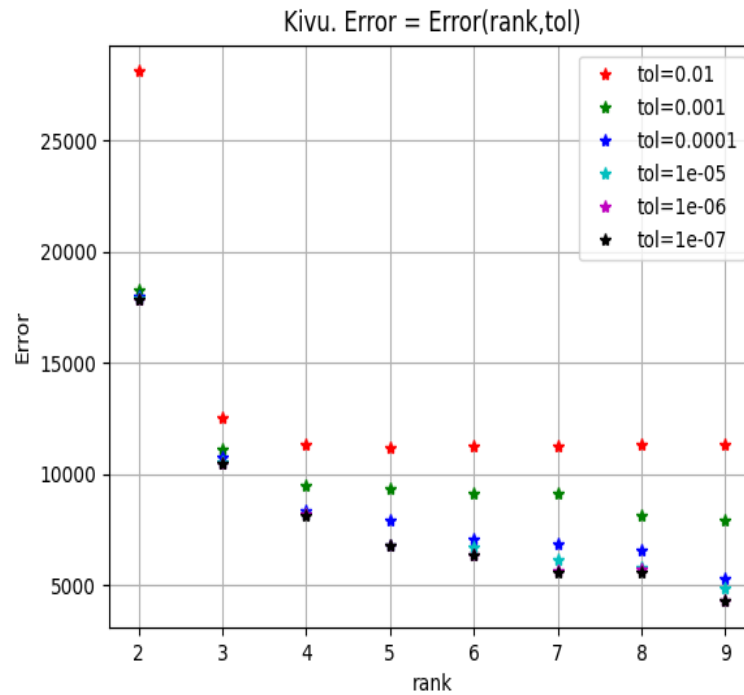


Рис. 17: Kivu. Поведение ошибки

Аномальное поведение наблюдается для рангов равных пяти, шести и семи. Также заметим, что при этих же рангах некоторые факторы-компоненты сильно похожи, что может говорить о слишком большом ранге. Изобразим компоненты для 4-7 рангов.

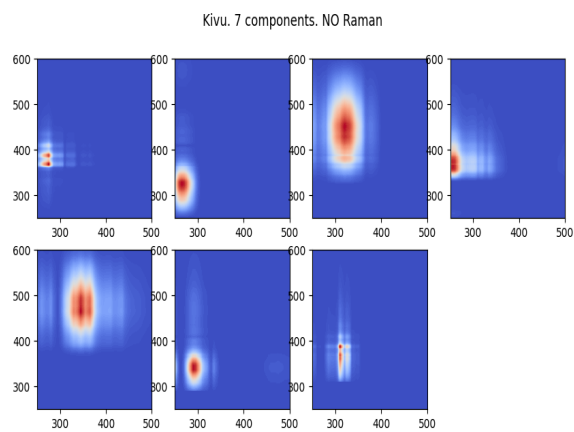
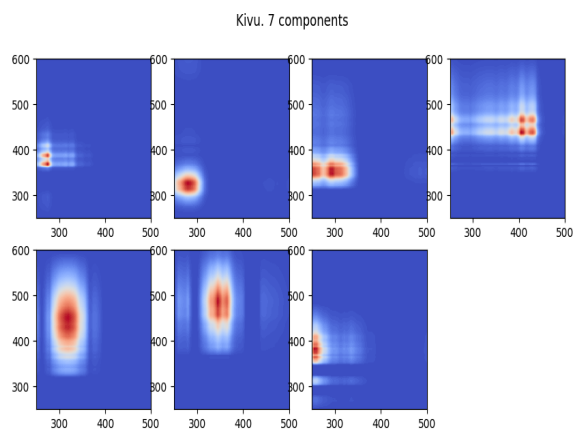
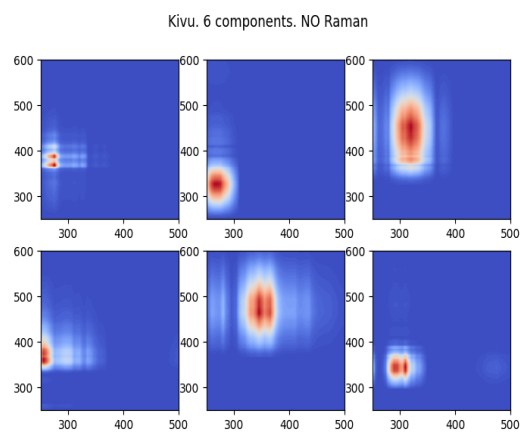
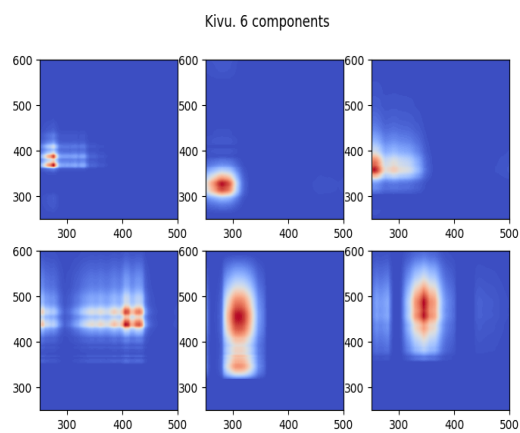
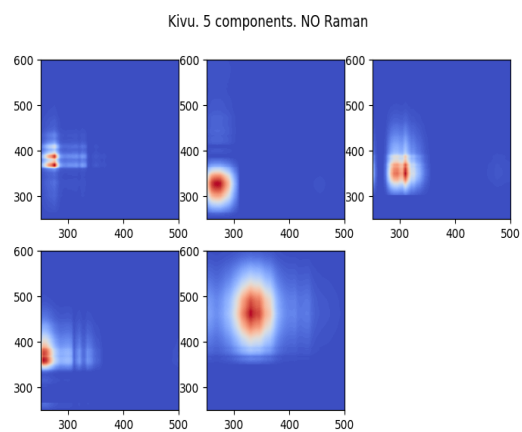
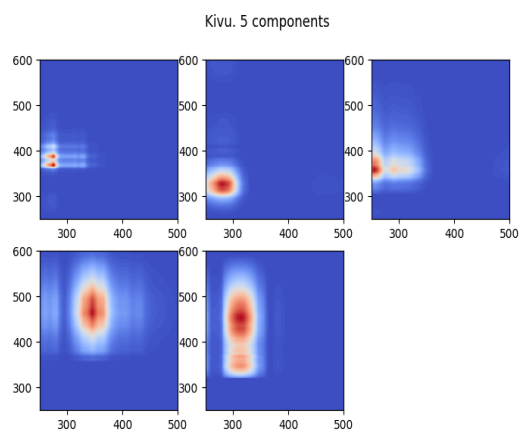
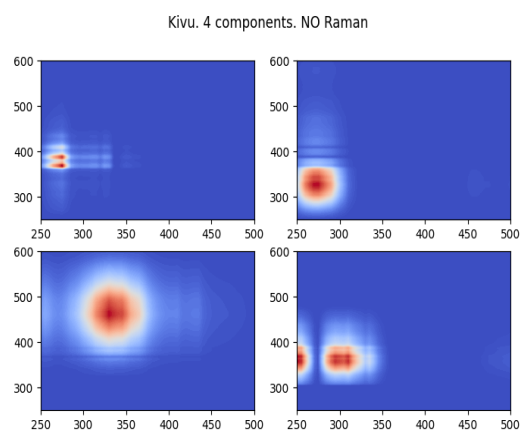
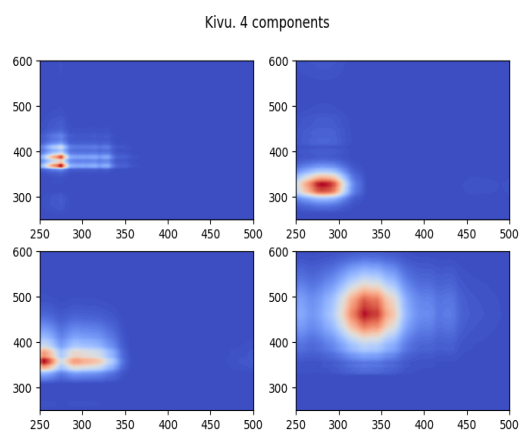


Рис. 18: Kivu. Компоненты

На Рис. 18 можно заметить как на определение компонент повлияло отсечение пиков Рамана. Плохо определялись компоненты, находящиеся вблизи этой прямой. Возможно, иногда не стоит подвергать обработке некоторые артефакты.

5.3 Lake Baikal

Исходные данные:

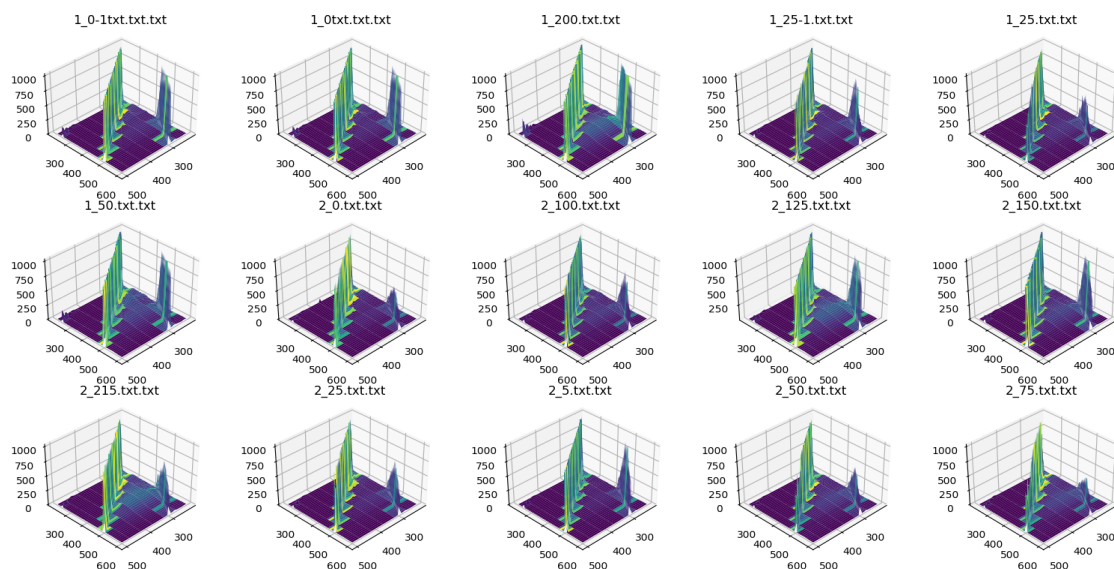


Рис. 19: Baikal. Исходные данные

Данные, после того как убрали артефакты:

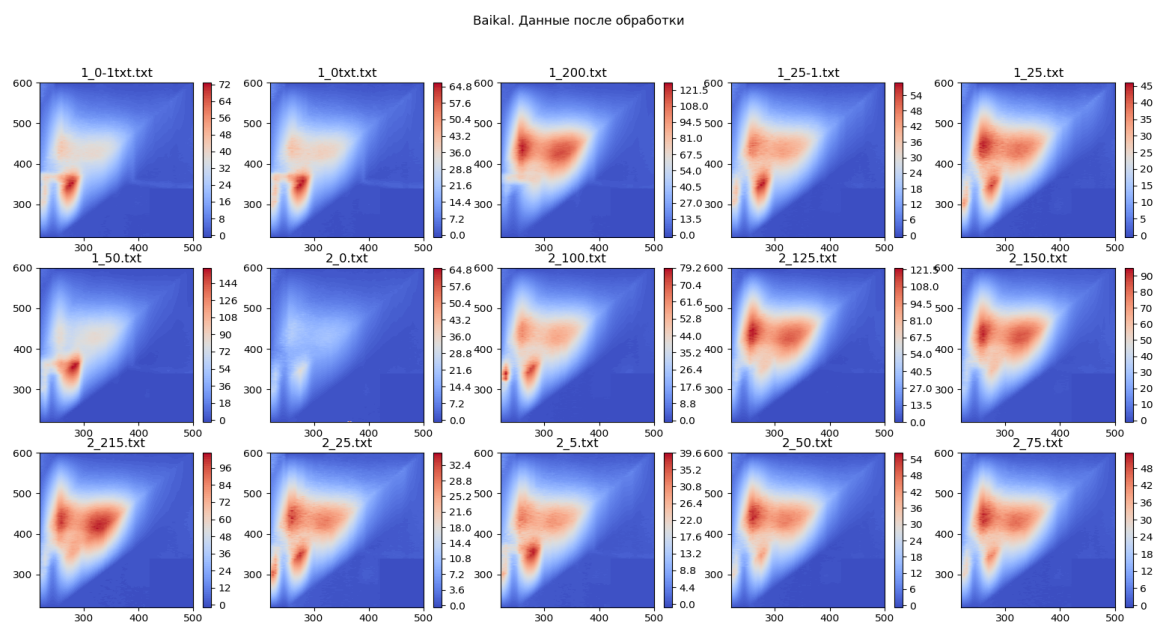


Рис. 20: Baikal. Исходные данные без артефактов

Заменим, что данные похожи между собой, что может плохо сказаться на определении компонент.

Применим PARAFAC:



Рис. 21: Байкал. Все данные. Разные ранги

Проследим за изменением ошибки:

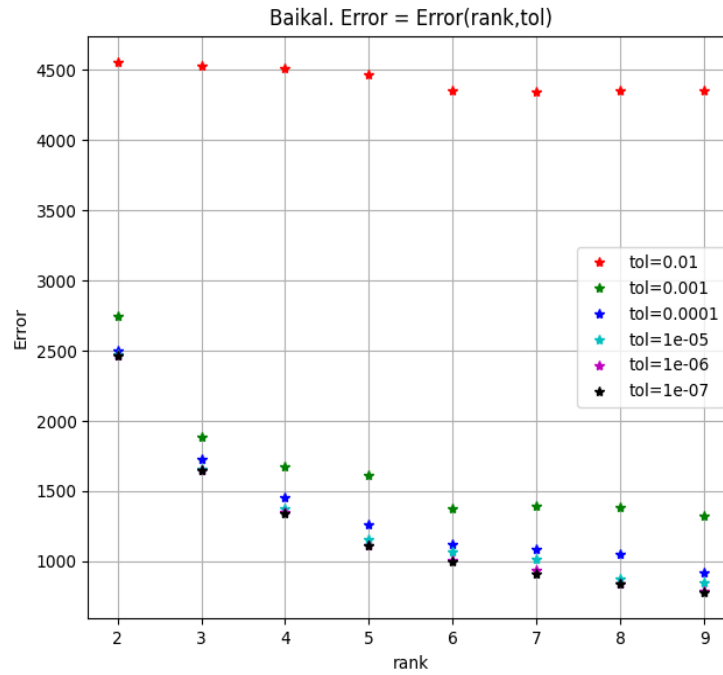


Рис. 22: Baikal. Поведение ошибки

Аномалии наблюдаются для рангов, равных шести и семи. Но полученные кривые имеют несколько пиков, что можно объяснить не самым лучшим отсечением пиков Рамана и Рэлея. Изобразим компоненты для рангов с 4 по 7. На рисунках ниже в левой части, компоненты получены только после отсечения пиков Рэлея, в правой после отсечения пиков Рэлея и Рамана.

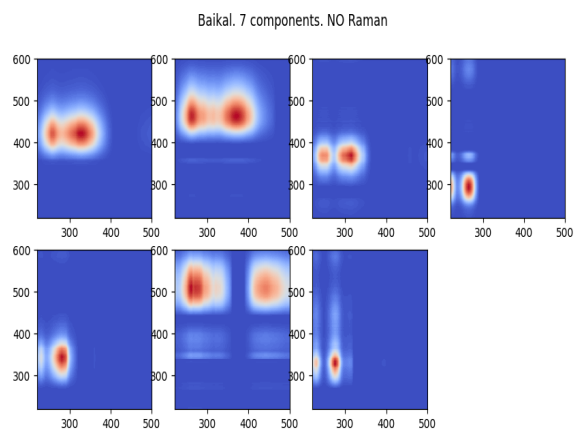
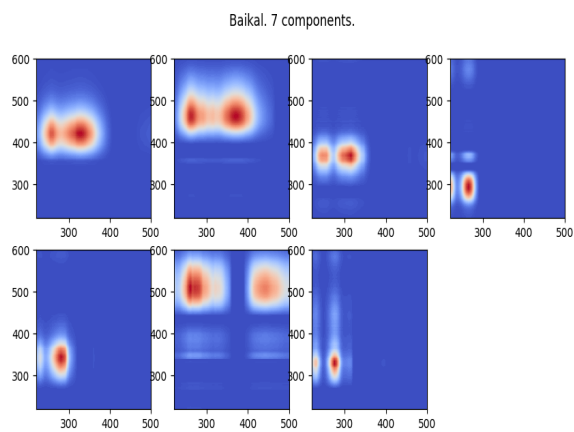
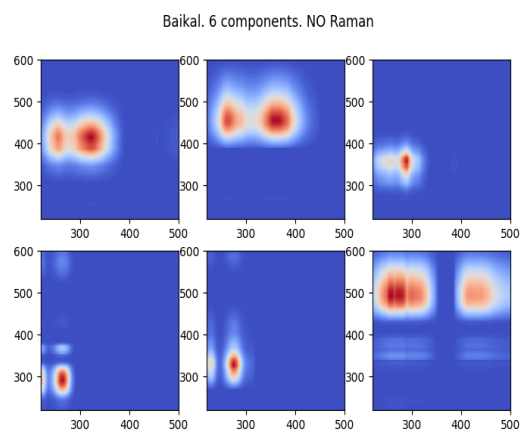
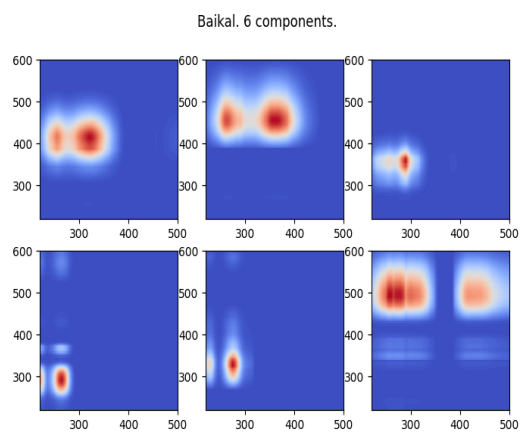
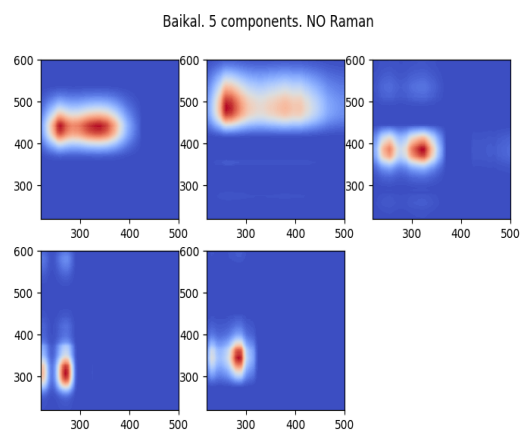
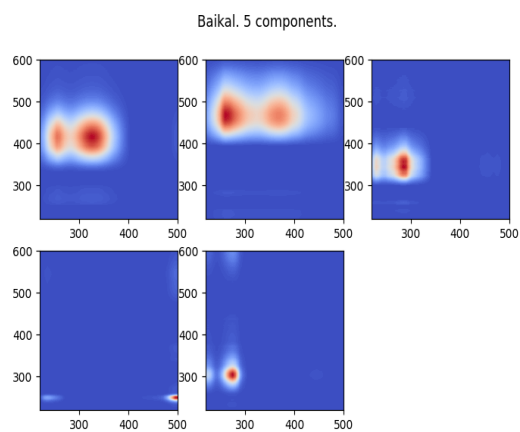
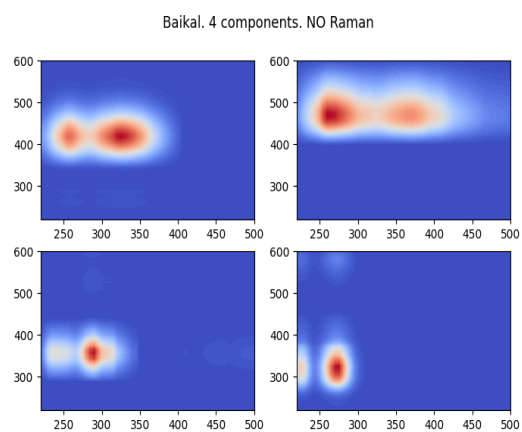
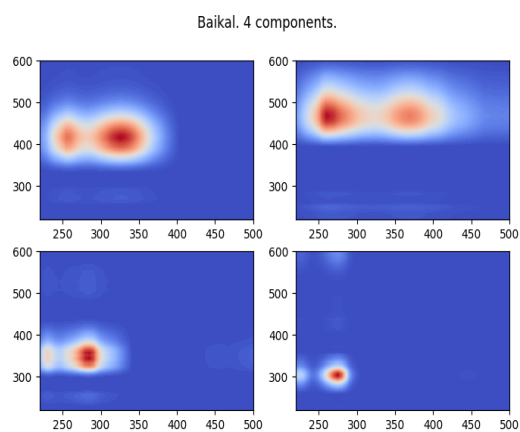


Рис. 23: Baikal. Компоненты

5.4 Обработка данных

Поскольку в флуориметрии используют данные, полученных при реальных измерениях, от образцов разной природы, концентрации, с возможным присутствием загрязнений и т.п., бывает необходима их предварительная обработка.

В книге [13] этому вопросу посвящена отдельная глава, в которой рассмотрены различные варианты предобработки.

Конспективно можно выделить следующие виды предобработки:

- устранение или учет сдвигов данных
- центрирование
- шкалирование (растяжение по осям)
- учёт пропусков данных
- устранение систематического фона

Изменение компонент в зависимости от предварительной обработки исходных данных (см. Permafrost): Без обработки:

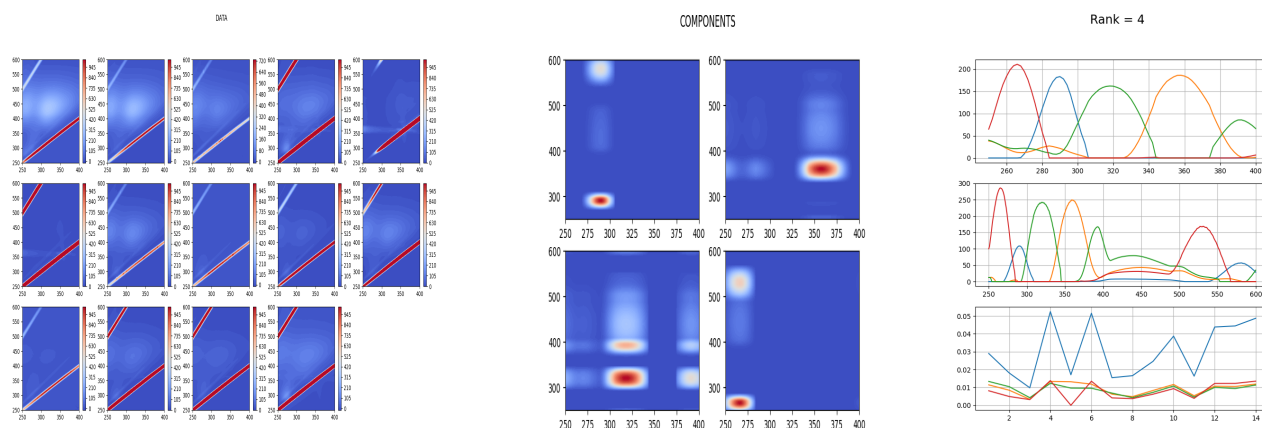


Рис. 24: Компоненты, полученные без предварительной обработки данных

После отсечения пиков Рэлея:

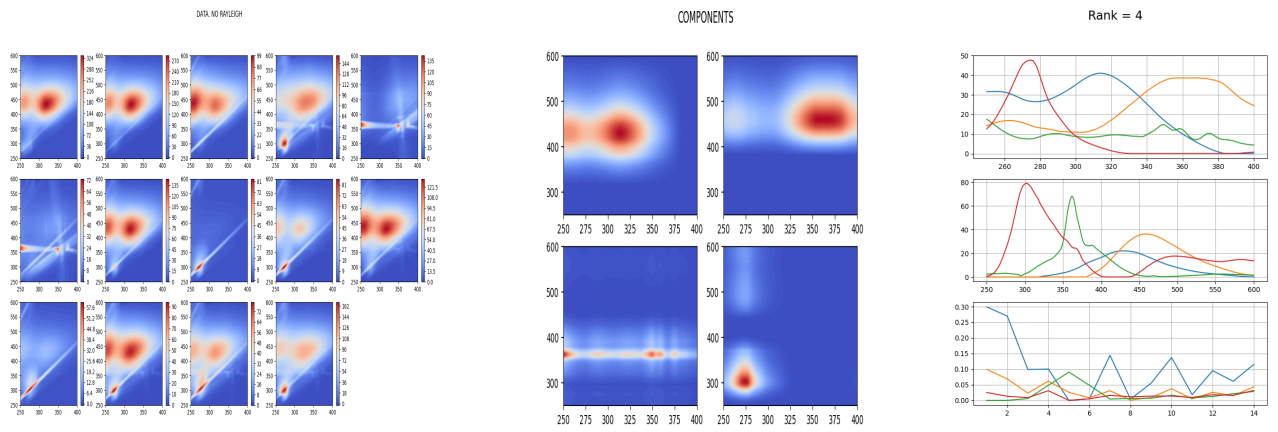


Рис. 25: Компоненты, полученные после отсечения пиков Рэля

Отсечение пиков Рэля и Рамана:

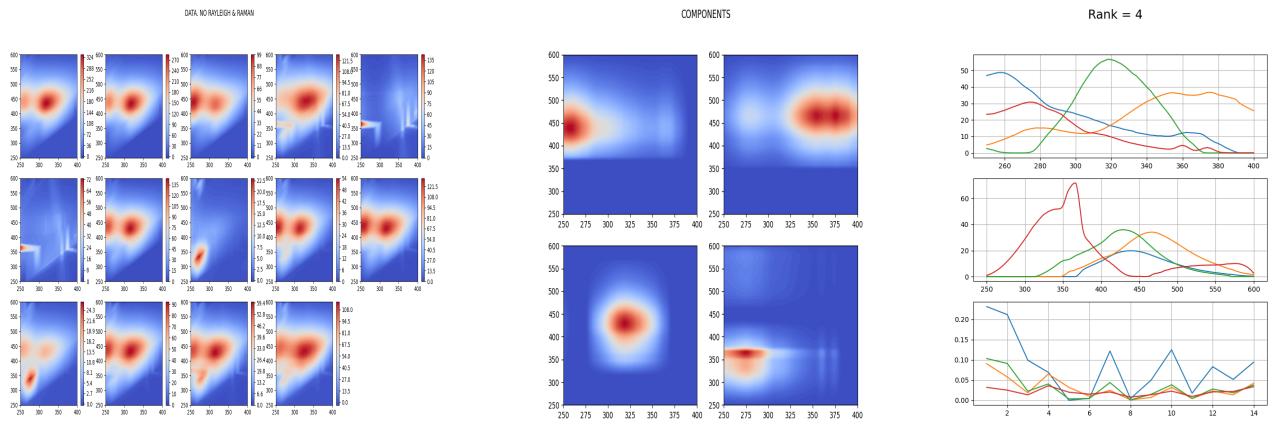


Рис. 26: Компоненты, полученные после отсечения пиков Рэля и Рамана

Замена нулями:

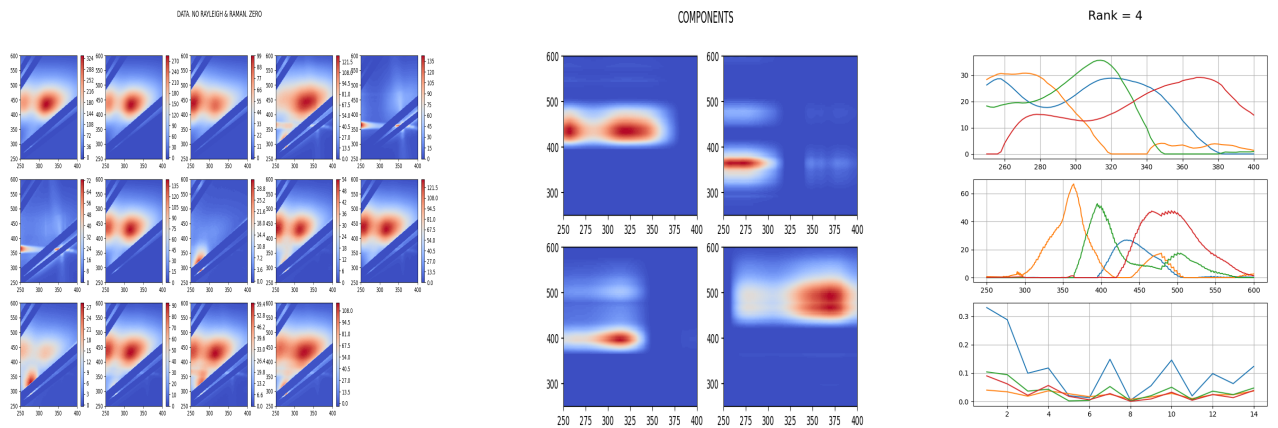


Рис. 27: Компоненты, полученные после замены пиков Рэля и Рамана нулями

Интерпретация результатов. Содержательную оценку результатов можно провести на основе представлений о характере базовых процессов флуориметрии, а также представлений о самих образцах. Характер образцов и ожидаемые спектры поглощения-эмиссии лежат за рамками математического исследования и должны проводиться специалистами в предметной области.

Что касается математических оценок, то наиболее валидным фактом является смещение компонент спектра эмиссии в длинноволновую область, что представлено в виде отношения порядка между спектрами поглощения - эмиссии (2). Придать количественный характер этому факту можно различными способами.

Физики, в частности, используют функции корреляции

$$R_{XY}(\tau) = \sum_{\lambda} X(\lambda)Y(\lambda + \tau), \quad (4)$$

где λ — длины волн фотонов, τ — сдвиг между длинами волн в спектрах поглощения - эмиссии. Функция корреляции имеет максимум при наиболее характерном сдвиге спектра эмиссии относительно спектра поглощения. Вычислив

$$\Delta\lambda = \arg \max_{\tau} R_{XY}(\tau)$$

для всех компонент тензорного разложения, можно проводить содержательный анализ на основе представлений о базовых реакциях (1).

С математической точки зрения, может быть содержательной установление какой-либо меры расстояния между спектрами поглощения - эмиссии **dist** и дальнейшей работы с этой мерой. В силу положительности спектров, возможно использование, например, аппарата нечётких множеств или иных подходов, описывающих математические объекты с двусторонними ограничениями, для которых установлены отношения упорядочивания. В качестве одного из кандидатов можно указать интервальный анализ с использованием полной интервальной арифметики Каухера.

Также возможно описание спектров поглощения и излучения как твинов, интервалов с интервальными границами. В книге [21] обсуждаются различные операции с этими объектами.

6 Используемые и разработанные методы и программы

Вычисления проводились с помощью языка программирования Python 3.8 в среде разработки PyCharm. Написан алгоритм PARAFAC для искусственных примеров и другие вспомогательные функции для обработки исходных данных.

GitHub: <https://github.com/Tatiana655/Practice2021>

В работе с реальными данными использовались алгоритмы тензорных преобразований из библиотеки tensorly.

7 Заключение

Тензорные преобразования представляют интерес как с теоретической, так и с практической точки зрения, и являются некоторым инструментом для решения многомерных задач.

Тензорное разложение PARAFAC заключается в приближённом описании трёхмерного тензора суммой триад. В хемометрии каждый вектор триады имеет свою семантику. К сожалению, на реальных данных определение ранга разложения вызывает затруднения, а также интерпретацию полученным результатам может дать только специалист другого направления.

В работе приведены примеры представлений тензоров с помощью моделей PARAFAC и Tucker3, а также приведён пример получения компонент с использованием алгоритма PARAFAC-ALS. На практике удалось пронаблюдать одно из преимуществ рассматриваемого тензорного разложения - адекватное определение компонент на зашумлённых данных.

Дальнейшие исследования. В настоящее время ведутся интенсивные поиски улучшения скорости работы вычислительного ядра PARAFAC и его модификаций, таких как PARAFAC2, CANDELINC, DEDICOM, PARATUCK, PARATUCK2 и иерархических видов метода ALS. В значительном числе публикаций рассматриваются различные модификации вычислений ALS.

Типичной является публикация [16], в которой рассмотрен набор модификаций ALS (всего 4) и проведён ряд вычислительных экспериментов с тензорами размерности $30 \times 30 \times 30$ и $100 \times 100 \times 100$ и рангом разложения 20 и 25 соответственно. Эксперименты проводились для случайных матриц A, B, C и для плохообусловленных задач с группой малых собственных значений. Изучены основные закономерности скорости сходимости для 4 алгоритмов. В целом они демонстрируют существенный выигрыш по отношению к традиционным схемам вычислений на случайных матрицах. В случае плохой обусловленности анализ сложнее. Авторы приводят обсуждение результатов и предлагают дальнейшие пути исследования.

Современный обзор методов ускорения приведён в только что вышедшей работе [17]. В ней рассмотрены в общей сложности 10 методов для PARAFAC2, базового и иерархического. Показано, что алгоритм PARAFAC2-ALS5, использующий ускоренные градиенты Нестерова, достигает максимальной скорости сходимости при сохранении низкой доли локальных минимумов решений. Показано, что этот новый метод в среднем в 13 раз быстрее по сравнению с алгоритмом PARAFAC2-ALS без ускорения, в то время как обычно используемый алгоритм экстраполяции строки в пакете N-way toolbox в 3 раза для одного и того же моделируемого набора данных.

Перечисленные примеры свидетельствуют, что тензорные разложения заслуживают активного изучения на образовательном и исследовательском уровне. Введение раздела о тензорных преобразованиях в курсы вычислительной математики и в практику даст обучающимся и исследователям современные эффективные инструменты вычислений. Необходимы также исследования новых вычислительных схем, альтернативных ALS, для использования тензорных разложений для статистики данных с интревальной неопределённостью. Это по-

требует разработки вопроса применения вычислительных методов негладкой оптимизации в PARAFAC и подобные методы.

Применительно к задачам флуориметрии, можно предложить более детальное исследование рангов разложения и применения альтернативных вычислительных схем для конкретного типа данных, подобно тому, как это сделано в работе [17].

Благодарности. Авторы выражают благодарность сотрудникам ВНИИОкеангеология им. И.С.Грамберга П.Б.Семёнову, А.А. Письменюк, С.А.Мальшеву за постановку задачи, предоставленные данные и полезные обсуждения.

Также выражаем признательность в.н.с. ФТИ им.А.Ф.Иоффе д.ф.-м.н. С.А.Руколайне за полезные консультации в области тензорных разложений.

Список литературы

- [1] О. Е. Родионова, А. Л. Померанцев, “Хеометрика: достижения и перспективы”, Усп. хим., 75:4 (2006), 302–321; Russian Chem. Reviews, 75:4 (2006), 271–287
- [2] R. Bro . The N-way Toolbox. he N-way Toolbox version 1.8.0.0 (5.2 MB) by Tools for fitting multi-way (tensor) models such as PARAFAC
- [3] EEM package.
<https://cran.r-project.org/web/packages/EEM/vignettes/vignette.html>
- [4] И.Оселедец. TT-Toolbox. <https://github.com/oseledets/TT-Toolbox>
- [5] S.K. Schmitz, P.P. Hasselbach, B. Ebisch, A. Klein, G. Pipa and R.A.W. Galuske. Application of Parallel Factor Analysis (PARAFAC) to electrophysiological data. //Front. Neuroinform., 30 January 2015 | <https://doi.org/10.3389/fninf.2014.00084>
- [6] W. Chen, P. Westerhoff, Paul Westerhoff, J.A. Leenheer, K. Booksh. Fluorescence Excitation-Emission Matrix Regional Integration to Quantify Spectra for Dissolved Organic Matter January 2004 Environmental Science and Technology 37(24):5701-10 DOI: 10.1021/es034354c
- [7] R.Rivas, G.Leon, J.Leal, J.Ramirez, F.Romero. Characterization of dissolved organic matter in an agricultural wastewater irrigated soil, in semi arid Mexico. Rev. Int. Contam. Ambient vol.33 no.4 Ciudad de México nov. 2017 <https://doi.org/10.20937/rica.2017.33.04.03>
- [8] Semenov, P.B.; Pismeniuk, A.A.; Malyshev, S.A.; Leibman, M.O.; Streletskaya, I.D.; Shatrova, E.V.; Kizyakov, A.I.; Vanshtein, B.G. Methane and Dissolved Organic Matter in the Ground Ice Samples from Central Yamal: Implications to Biogeochemical Cycling and Greenhouse Gas Emission. Geosciences 2020, 10, 450. <https://doi.org/10.3390/geosciences10110450>
- [9] B.N. Khoromskij. Tensors-structured numerical methods in scientific computing: Survey on recent advances. // Chemometrics and Intelligent Laboratory Systems 110 (2012) 1–19
- [10] Sanchez, E., Kowalski, B. R. (1990). Tensorial resolution: A direct trilinear decomposition. Journal of Chemometrics, 4(1), 29–45.
- [11] Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear Algebra and Its Applications, 18(2), 95–138.
- [12] Kolda, T. G., Bader, B. W. (2009). Tensor Decompositions and Applications. SIAM Review, 51(3), 455–500.
- [13] Smilde A., Bro R., and Geladi P. (2004), Multi-Way Analysis: Applications in the Chemical Sciences, Wiley, West Sussex, England. 369.

- [14] Burdick, D. S. (1995). An introduction to tensor products with applications to multiway data analysis. *Chemometrics and Intelligent Laboratory Systems*, 28(2), 229–237.
- [15] Sidiropoulos, N. D., Bro, R. (2000). On the uniqueness of multilinear decomposition of N-way arrays. *Journal of Chemometrics*, 14(3), 229–239.
- [16] Д.М.Меркулов, Н.К.Тупица. Об ускоренных методах поиска канонического тензорного разложения. *Труды МФТИ. Том 12, № 4 (48) (2020)*
- [17] H. Yu, D. Augustijn, R. Bro, Accelerating PARAFAC2 algorithms for nonnegative complex tensor decomposition, *Chemometrics and Intelligent Laboratory Systems*,
[https:// doi.org/10.1016/j.chemolab.2021.104312](https://doi.org/10.1016/j.chemolab.2021.104312).
- [18] <https://doi.org/10.1145/3097983.3098014>
- [19] Kiers, H. A. L., ten Berge, J. M. F., Bro, R. (1999). PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics*, 13(3-4), 275–294.
- [20] https://www.youtube.com/watch?v=-SdZi8LKLog&t=1338s&ab_channel=QualityAndTechnology
- [21] Sainz M.A., Armengol J., Calm R., Herrero P., Jorba L.J., Vehi J. *Modal Interval Analysis: New Tools for Numerical Information*. – Cham, Switzerland: Springer, 2014. – (Lecture Notes in Mathematics; vol. 2091).

Предметный указатель

- k -ранг, 13
- Matlab, 5
- PyTorch, TensorFlow, TensorLy, 6
- R, 5
- core tensor, 18
- PARAFAC, 14
- PARAFAC2, 20, 43
- Tensor Train, тензорные поезда, 6
- Tucker1, 19
- Tucker2. [13], 19
- Tucker3, 18
- Произведение Хатри-Рао, 13
- Псевдообращение Мура-Пенроуза, 13
- алгоритм ALS, 16
- двулинейная модель, 16
- иерархический ALS, 43
- матрица-срез, 12
- матричная форма записи Tucker3, 18
- методы негладкой оптимизации, 44
- пики Рамана, 9
- пики Рэлея, 9
- произведение Адамара (Шура), 11
- произведение Кронекера, 11
- псевдоранг тензора, 14
- ранг тензора, 14
- растворённое ограниченное вещество, 8
- супердиагональ, 15
- тензорное произведение, 12
- тензорные разложения, 5
- триада, 12
- триадная форма записи Tucker3, 18
- трилинейная модель, 16
- ускоренные градиенты Нестерова, 43
- флуориметрия, 8
- хеометрика, 5