

ОБЩИЕ
ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.6

МОДЕЛИРОВАНИЕ СТРУКТУРЫ ДАННЫХ С ПОМОЩЬЮ
БЛОЧНОГО ТЕНЗОРНОГО РАЗЛОЖЕНИЯ:
РАЗЛОЖЕНИЕ ОБЪЕДИНЕННЫХ ТЕНЗОРОВ
И ВАРИАЦИОННОЕ БЛОЧНОЕ ТЕНЗОРНОЕ РАЗЛОЖЕНИЕ
КАК ПАРАМЕТРИЗОВАННАЯ МОДЕЛЬ СМЕСЕЙ¹⁾

© 2021 г. И. В. Оселедец^{1,2,*}, П. В. Харюк^{1,2,3,**}

¹ 121205 Москва, Большой бульвар, 30, стр. 1, Сколковский институт науки и технологий, Россия

² 119333 Москва, ул. Губкина, 8, Институт вычислительной математики им. Г.И. Марчука РАН, Россия

³ 119991 Москва, Ленинские горы, 1, стр. 52, МГУ им. М.В. Ломоносова,
Факультет вычислительной математики и кибернетики, Россия

*e-mail: ivan.oseledets@gmail.com

**e-mail: kharyuk.pavel@gmail.com

Поступила в редакцию 24.12.2020 г.

Переработанный вариант 24.12.2020 г.

Принята к публикации 14.01.2021 г.

Развивается идея использования тензорных разложений в качестве параметрической модели группового анализа данных. Представлены две модели на основе блочного тензорного разложения: детерминистическая и вероятностная, с использованием различных форматов слагаемых. Установлена связь между блочным тензорным разложением и смесями непрерывных латентных вероятностных моделей: на основе блочного тензорного разложения построена модель смеси распределений со структурированным представлением. Модели были протестированы в задаче кластеризации набора цветных изображений и данных электрической активности мозга. Результаты показывают, что предложенные подходы способны к выделению релевантной индивидуальной составляющей данных. Библ. 54. Фиг. 4. Табл. 5.

Ключевые слова: групповой анализ данных, блочное тензорное разложение, машинное обучение, анализ компонент, модель смеси распределений.

DOI: 10.31857/S004446692105015X

1. ВВЕДЕНИЕ

В процессе поиска зависимостей в данных важно учитывать их природу. Ряд собранных наборов данных может быть представлен в тензорном виде (как многомерные массивы), и учет их размерности представляется важным не только для эффективной обработки, но и для потенциальной интерпретации найденных параметров в конкретном прикладном исследовании. Более того, подобные данные часто являются избыточными и могут быть приближены меньшим числом параметров. Тензорные разложения предоставляют возможность построения подобных приближений. Различные тензорные форматы, такие как канонический (полилинейный, CPD) или формат Таккера (см. [1], [2]), нашли успешное применение в задачах анализа данных (см. [3], [4]). В данной работе в качестве основы для моделей мы использовали так называемое *блочное тензорное разложение* (BTD), которое является одним из способов обобщения классических тензорных форматов и было предложено в [5]–[7].

В условиях ряда исследований требуется извлекать общую и индивидуальную информацию из данных. Например, подобные исследования возникают в когнитивных, медицинских и иных задачах изучения мозга по группе из нескольких человек (см., например, [8]–[10]); другим примером может служить химическая экспертиза набора образцов (см. [11]–[13]). Естественно ожидать от подобных наборов данных наличие общих (групповых) и индивидуальных частей, и реконструкцию соответствующих частей из данных мы будем называть *групповым анализом данных*.

¹⁾ Работа выполнена при финансовой поддержке РФФИ (проект № 16-31-00494-мол_а).

Предложенные в данной работе модели структуры данных основаны на ВТД со слагаемыми различных типов, что позволяет моделировать групповые и индивидуальные части различными форматами.

Предложенные тензорные модели структуры данных основаны на концепте “связанного многомерного анализа компонент” (см. [14], [15]), реализованного в виде ВТД с дополнительными условиями; кроме того, реализована модель смеси вероятностных распределений, параметризованная через ВТД и названная *вариационным блочным тензорным разложением* (VBTD). В предыдущих работах по групповому анализу данных (см. [16], [17]) специальный случай первой (детерминистической) модели (а именно, условное ВТД со слагаемыми в (Lg, 1) формате) был рассмотрен как составляющий элемент обработки данных для задачи классификации. В настоящей работе модели были протестированы в условиях задачи кластеризации. Сведения об использованных в работе обозначениях приведены в п. 10.1.

2. СТРУКТУРИРОВАНИЕ НАБОРОВ ДАННЫХ ДЛЯ ГРУППОВОГО АНАЛИЗА

Одна из главных целей группового анализа данных может быть сформулирована как выделение общей информации из комбинированного набора данных. Вопрос в том, что может быть названо “общей информацией”. В данной работе мы придерживаемся распространенной в области обработки сигналов точки зрения на природу данных: предположения о том, что наблюдения $\{X_i\}_{i=1}^N$ генерируются в результате преобразования активности нескольких источников, представленных в векторном виде. Адаптация такой точки зрения к групповому анализу данных ведет к разделению источников на общие, S_{com} , и индивидуальные, $\{S_i\}_{i=1}^N$.

Следующий шаг состоит в предположении об отделимости (в некотором смысле) общих и индивидуальных источников. Принято считать, что зависимость наблюдений от всех источников аддитивна.

Предположение группового анализа данных. Наблюдения $\{X_i\}_{i=1}^N$ могут быть приближены аддитивной функциональной зависимостью от общих, S_{com} , и индивидуальных, $\{S_i\}_{i=1}^N$, источников

$$X_i \approx F_i(S_{\text{com}}, S_i) = f_i(S_{\text{com}}) + g_i(S_i), \quad i = \overline{1, N}. \quad (1)$$

В случае, если наблюдения X_i двумерны (являются матрицами), распространенным способом моделирования указанной зависимости является линейное смешивание:

$$f_i(S_{\text{com}}) = S_{\text{com}} B_{f_i}^T, \quad g_i(S_i) = S_i B_{g_i}^T, \quad i = \overline{1, N}, \quad (2)$$

где B_{f_i} , B_{g_i} – матрицы коэффициентов общих и индивидуальных источников соответственно для образца X_i . Модель факторизации данных ставит вопрос о том, как найти требуемые параметры разложения. В [18] предложен ответ для общей части в виде группового метода независимых компонент (group ICA). Zhou и др. (см. [19]) построили алгоритм вычисления обеих частей разложения через приближение матрицы общих источников, ортогональной всем индивидуальным источникам. В [20] общая часть оценивается через малоранговое разложение склеенных матричных данных $X = [X_1 \dots X_N] \approx S_{\text{com}} [B_{f_1} \dots B_{f_N}]^T$ и каждая индивидуальная часть вычисляется как малоранговое разложение данных после удаления групповой части: $X_i - S_{\text{com}} B_{f_i}^T \approx S_i B_{g_i}^T$.

В случае многомерных данных X_i размерами $n_1 \times \dots \times n_d$, как правило, используются предварительно матризованные наблюдения. Однако в этом случае рассматриваемая модель игнорирует размерность входных данных. Детерминистическая ВТД модель, предложенная в данной работе, обобщает подход для больших размерностей следующим образом. Рассмотрим матрицу развертки тензорных данных по выбранной моде источников $k \in \{1, 2, \dots, d\}$ (например, отвечающей пикселям или времени) и применим к ней сформулированные ранее предположения (1), (2):

$$\text{unfold}_k(X_i) = (X_i)_{(k)} \approx S_{\text{com}} B_{f_i}^T + S_i B_{g_i}^T, \quad i = \overline{1, N}. \quad (3)$$

Далее учтем тензорную структуру слагаемых в правой части (3). Также вместо вычисления набора связанных разложений объединим данные $\{X_i\}_{i=1}^N$ вдоль новой групповой оси, получив общий тензор данных X размерами $n_1 \times \dots \times n_d \times N$, и наложим условия на специальную структуру параметров, отвечающих групповой моде. Объединение данных требует выполнения условий однородности для $\{X_i\}_{i=1}^N$: все образцы должны иметь равную размерность с одинаковыми размерами мод.

3. ДЕТЕРМИНИСТИЧЕСКАЯ ГРУППОВАЯ МОДЕЛЬ НА ОСНОВЕ
ВТД РАЗЛОЖЕНИЯ ОБЪЕДИНЕННЫХ ДАННЫХ

После учета предположений из разд. 2 модельная структура входных данных примет следующий вид:

$$X \approx \sum_{i=1}^N (X_i(S_{\text{com}}, \theta_{\text{com},i}) + X_i(S_i, \theta_{\text{ind},i})) \circ \mathbf{e}_i, \tag{4}$$

где \mathbf{e}_i есть i -й столбец единичной матрицы I_N , \circ обозначает внешнее произведение, $\theta_{\text{com},i}$, $\theta_{\text{ind},i}$ – общие и индивидуальные параметры, соответствующие тому или иному тензорному формату и отвечающие общим, S_{com} , и индивидуальным, S_i , источникам. Сосредоточимся на двух тензорных форматах: формате Таккера и $(Lg, 1)$ (специальном каноническом с параметром P , отвечающим числу мод с полноразмерными фактор-матрицами), который был предложен в [21]. В случае $(Lg, 1)$ формата для $T_i = X_i(S_i, \theta_{\text{ind},i})$ соответствующая сумма может быть переписана в следующем виде:

$$\sum_{i=1}^N T_i \circ \mathbf{e}_i = \sum_{i=1}^N \left[[C_1^{[i]}, \dots, C_P^{[i]}, \mathbf{c}_{P+1}^{[i]} E^{[i]}, \dots, \mathbf{c}_d^{[i]} E^{[i]}] \right] \circ \mathbf{e}_i = [[C_1, \dots, C_d E, I_N E]], \tag{5}$$

где умножение фактор-матриц $\{C_j\}_{j=P+1}^d$ на $E = [E^{[1]}, \dots, E^{[N]}]$ дублирует необходимое количество раз их столбцы. Выражение справедливо для обеих частей разложения: индивидуальных и общих. В последнем случае дополнительно предполагаем, что $X_i(S_{\text{com}}, \theta_{\text{com},i}) = p_i T_G$, с дополнительными параметрами $\sum_{i=1}^N p_i = 1$, $p_i \geq 0$, которые позволяют регулировать присутствие общей части T_G . Из этого предположения следует, что групповые части модели могут быть представлены как $T_G \circ \mathbb{1}_{N \times 1}$, где $\mathbb{1}_{N \times 1}$ – вектор из единиц размерами $N \times 1$. Сводя все предположения воедино, а также задав число компонент (ранги) для индивидуальных и общих слагаемых, $L = [L_1 \dots L_N L_{N+1}]$, приходим к следующей модели:

$$X \approx [[C_1, \dots, C_d, C_{d+1}], \quad C_k = \begin{cases} [\overbrace{C_k^{[1]}, \dots, C_k^{[N+1]}}^{L_1}], & k = \overline{1, P}, \\ [\mathbf{c}_j^{[1]}, \dots, \mathbf{c}_j^{[N+1]}] E, & k = \overline{P+1, d}, \quad j = k - P, \\ [I_N \mathbf{p}] E, & k = d + 1, \end{cases} \tag{6}$$

где $E = [\mathbb{1}_{1 \times L_1} \otimes \mathbf{e}_1, \dots, \mathbb{1}_{1 \times L_{N+1}} \otimes \mathbf{e}_{N+1}]$, \otimes – произведение Кронекера, \mathbf{p} – вектор параметров, определенный ранее. Однако вариативность групповой части в данной модели может быть слишком ограниченной. Для моделирования более сложной зависимости от общей части можно использовать формат Таккера со схожими условиями на фактор-матрицу групповой оси. Таким образом, полностью $(Lg, 1)$ групповую модель (далее GLRO) дополним похожей моделью, в которой общая часть представлена в формате Таккера (далее GTLD):

$$X \approx [[G; A_1, \dots, A_d, A_{d+1}] + [C_1, \dots, C_d, C_{d+1}], \tag{7}$$

$$A_{d+1} = \text{diag}(\mathbf{p}), \quad C_k = \begin{cases} [\overbrace{C_k^{[1]}, \dots, C_k^{[N]}}^{L_1}], & k = \overline{1, P}, \\ [\mathbf{c}_j^{[1]}, \dots, \mathbf{c}_j^{[N]}] E, & k = \overline{P+1, d}, \quad j = k - P, \\ I_N, & k = d + 1, \end{cases}$$

где последняя мода отвечает групповой оси, как и в (6).

В силу того что как канонический и (Lr, 1) форматы, так и формат Таккера одинаковым образом разделяют оси тензора, имеет смысл указать общий случай ВТД со слагаемыми во всех трех форматах:

$$X \approx \underbrace{\sum_{m=1}^M \left[G^{[m]}; A_1^{[m]}, \dots, A_{d+1}^{[m]} \right]}_{\text{в формате Таккера}} + \underbrace{\left[B_1, \dots, B_{d+1} \right]}_{\text{канонический}} + \underbrace{\left[C_1, \dots, C_{d+1} \right]}_{(Lr,1)}. \quad (8)$$

Отметим, что программная реализация предложенных моделей позволяет вычислять параметры такого разложения более общего вида. В групповом анализе данных также можно моделировать индивидуальные части в формате Таккера, но это потребует увеличения числа гиперпараметров (рангов) для каждого отдельного слагаемого. В случае же (Lr, 1) формата потребуется указать лишь один вектор гиперпараметров L , который задает ранги всех полноразмерных фактор-матриц, отвечающих индивидуальным слагаемым.

Помимо гиперпараметров, упомянутых ранее (P , L , и/или ранги Таккера), предложенный подход требует выбора некоторого подмножества среди первых P осей, $\Omega \subseteq \{1, 2, \dots, P\}$, для которых вводится условие отделимости соответствующих общих и индивидуальных фактор-матриц (аналогично [19]) с тем, чтобы убрать пересечения между общей и индивидуальными активностями:

$$\begin{aligned} (C_\gamma^{[N+1]})^T C_\gamma^{[i]} &= 0, \quad \text{для модели (6),} \\ (A_\gamma)^T C_\gamma^{[i]} &= 0, \quad \text{для модели (7),} \quad \forall i = \overline{1, N}, \quad \forall \gamma \in \Omega. \end{aligned} \quad (9)$$

В данной работе мы остановились на выборе множества Ω , состоящего из одной моды.

4. ВАРИАЦИОННОЕ БЛОЧНОЕ ТЕНЗОРНОЕ РАЗЛОЖЕНИЕ КАК МОДЕЛЬ СМЕСИ РАСПРЕДЕЛЕНИЙ

Рассмотрим альтернативный подход к параметризации данных с помощью ВТД. Первым отличием является вероятностный характер модели: векторизация наблюдаемых данных предполагается реализацией случайного вектора \mathbf{x} , зависящего от латентного (скрытого) представления; процесс моделирования состоит в детализации функции плотности распределения $p(\mathbf{x})$, а результат обучения потенциально применим также и для генерации новых данных (генеративная модель). Другое важное отличие состоит в том, что индивидуальные параметры модели специфичны для каждой отдельной группы наблюдений (кластера) в предположении, что известно точное число кластеров K , а не для каждого отдельного представителя.

4.1. Параметризованные смеси распределений

В общих чертах, предложенная вероятностная модель на основе ВТД возникает в результате параметризации смеси непрерывных латентных моделей в предположении о структурированности матриц смешивания. Частным случаем подобных смесей является смесь вероятностных главных компонент (mixture of PPCAs, см. [22]). Базовая модель вероятностных главных компонент (PPCA) (см. [23]) представляет собой параметризацию наблюдений \mathbf{x} с помощью скрытого (латентного) представления, имеющего стандартное нормальное распределение, $\mathbf{z} \sim N(\mathbf{0}, I)$: $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$, $p(\mathbf{x}|\mathbf{z}) = N(\boldsymbol{\mu} + W\mathbf{z}, L^{-1})$, где $\boldsymbol{\mu}$ – постоянная составляющая среднего (сдвиг), W – матрица смешивания, $L^{-1} = \sigma^2 I$ – ковариационная матрица (в более общем случае факторного анализа L^{-1} не обязательно является скалярной). В модели смеси вероятностных непрерывных скрытых представлений предполагается, что $p(\mathbf{x})$ является мультимодальным распределением (см. [24]), порожденным вероятностной смесью нескольких распределений, каждая мода которого параметризована своим непрерывным скрытым представлением:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k p_k(\mathbf{x}), \quad p_k(\mathbf{x}) = \int p_k(\mathbf{x}|\mathbf{z}_k)p(\mathbf{z}_k)d\mathbf{z}_k, \quad 0 \leq \pi_k \leq 1, \quad \sum_{k=1}^K \pi_k = 1, \quad (10)$$

где $p_k(\mathbf{x}|\mathbf{z}_k) = N(\boldsymbol{\mu}_k + W_k \mathbf{z}_k, L_k^{-1})$, $p(\mathbf{z}_k) = N(\mathbf{0}, I)$. Если проводить аналогии с разд. 3, то в среднем наблюдения \mathbf{x} предполагаются сгенерированными в результате взвешенной суммы нескольких

процессов линейного смешивания. Коэффициенты $\{\pi_k\}_{k=1}^K$ выступают в роли априорных вероятностей принадлежности к тому или иному кластеру. Соответствующие апостериорные вероятности вычисляются по правилу Байеса и имеют вид

$$\gamma_k(\mathbf{x}) = p(k|\mathbf{x}) = \frac{\pi_k p_k(\mathbf{x})}{\sum_l \pi_l p_l(\mathbf{x})}. \quad (11)$$

Обратим также внимание, что априорные вероятности принадлежности к тому или иному кластеру в построенной модели были параметризованы через исходные наблюдения и параметры кластера $\pi_k \equiv \pi_k(\mathbf{x}, W_k)$ (подробнее см. следующий п. 4.2.).

Дальнейшая связь с ВТД устанавливается следующим образом: для каждого $k = \overline{1, K}$ положим $\mu_k = 0$, обозначим реализацию вектора \mathbf{z}_k как \mathbf{Z}_k и рассмотрим $W_k \mathbf{Z}_k$ как векторизацию некоторого тензора, структура которого зависит от выбора того или иного тензорного формата, например:

– канонический (полилинейный, СР):

$$\mathbf{Z}_k = \text{vec}(\text{diag}(\hat{\lambda}_k)) \in \mathbb{R}^{R_k \times 1}, \quad W_k = C_k^{(d)} \odot \dots \odot C_k^{(1)}; \quad (12)$$

– (Lr, 1) с $1 < P < d$, $\mathbf{L}_k = (L_{k;1}, \dots, L_{k;R_k})$, $M_k = \sum_i L_{k;i}$ (СР специального вида):

$$\mathbf{Z}_k \in \mathbb{R}^{M_k \times 1}, \quad W_k = \tilde{C}_k^{(d)} \odot \dots \odot \tilde{C}_k^{(1)}, \quad \tilde{C}_k^{(p)} = \begin{cases} C_k^{(p)} \in \mathbb{R}^{n_p \times M_k}, & p \leq P, \\ \underbrace{C_k^{(p)} E_k}_{\text{повтор столбцов } C_k^{(p)} \in \mathbb{R}^{n_p \times R_k}}, & p > P; \end{cases} \quad (13)$$

– Таккера с ядром в роли латентного представления (“Tucker-core”):

$$\mathbf{Z}_k = \text{vec}(\hat{G}_k) \in \mathbb{R}^{r_1 \dots r_d \times 1}, \quad W_k = A_k^{(d)} \otimes \dots \otimes A_k^{(1)}; \quad (14)$$

– Таккера с фактор-матрицей в роли латентного представления (“Tucker-factor”), $G \in \mathbb{R}^{r_0 \times r_1 \times \dots \times r_d}$:

$$\mathbf{Z}_k = \text{vec}(\hat{\mathbf{a}}_k^{(0)}) \in \mathbb{R}^{r_0 \times 1}, \quad W_k = \left((A_k^{(d)} \otimes \dots \otimes A_k^{(1)}) (G_k)_{(0)}^T \right) \otimes I; \quad (15)$$

– тензорного поезда (ТТ) [25] с ядрами $G_k^{(i)} \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$, $I_{N_{i\dots i}} = I_{n_{i+1} \dots n_d}$:

$$\mathbf{Z}_k \in \mathbb{R}^{r_d \times 1}, \quad W_k = (I_{N_{1\dots 1}} \otimes G_k^{(1)}) \dots (I_{N_{d-1\dots d-1}} \otimes G_k^{(d-1)}) G_k^{(d)}. \quad (16)$$

Таким образом, правдоподобие модели *вероятностного ВТД-смешивания* принимает следующий вид:

$$p(\mathbf{x}|\mathbf{z}_1, \dots, \mathbf{z}_K) = \sum_{k=1}^K \pi_k N(\mathbf{x}|W_k \mathbf{z}_k, L_k^{-1}), \quad W_k \mathbf{z}_k = \text{vec}(T_k(\mathbf{z}_k)), \quad (17)$$

ковариационные матрицы L_k^{-1} полагаются либо имеющими скалярный вид (изотропный шум), либо диагональными. Обратная связь модели с классическим ВТД разложением может быть прослежена через условное математическое ожидание \mathbf{x} при известных реализациях скрытых представлений $\{\mathbf{z}_k\}_{k=1}^K$, результат соответствует развертке тензора в ВТД формате с π_k в роли весов для слагаемых:

$$E[\mathbf{x}|\mathbf{z}_1, \dots, \mathbf{z}_K] = \sum_{k=1}^K \pi_k W_k \mathbf{z}_k = \text{vec}(\hat{T}). \quad (18)$$

Помимо введенного правдоподобия, требуется также определить распределение для скрытых представлений $\{\mathbf{z}_k\}_{k=1}^K$. В силу того что правдоподобие моделируется с помощью нормального распределения, в качестве априорных распределений $p(\mathbf{z}_k)$ также выберем нормальное (как сопряженное к правдоподобию, см. [24]):

$$p(\mathbf{z}_k) = N(\mathbf{m}_k, \Lambda_k^{-1}) = N(\mathbf{m}_k(\mathbf{x}, W_k), \Lambda_k^{-1}). \quad (19)$$

Отличием от случая РРСА является использование диагональных матриц Λ_k^{-1} и ненулевых средних \mathbf{m}_k . Как и в случае вариационных автоэнкодеров (см. [26]), для средних \mathbf{m}_k была введена зависимость от входных данных и параметров разложения, в данном случае $\mathbf{m}_k(\mathbf{x}, W_k) = (W_k^T W_k)^{-1} W_k^T \mathbf{x} = W_k^+ \mathbf{x}$. Такой выбор объясняется следующими соображениями: рассмотрим условное среднее $E[\mathbf{x}|k] = W_k \mathbf{m}_k$, и для конкретной реализации \mathbf{x} данное выражение есть ни что иное как проекция в пространство столбцов матрицы W_k , $W_k \mathbf{m}_k(\mathbf{x}, W_k) = P_{W_k} \mathbf{x}$, т.е. в итоге кластерное среднее определяется соответствующей проекцией наблюдения.

Параметры ковариационных матриц Λ_k^{-1} и L_k^{-1} полагаются реализациями из лог-нормальных распределений с нулевым средним и единичной дисперсией. Наконец, укажем итоговое выражение для модели вариационного ВТД разложения (VBTD) с нормальными априорным распределением и правдоподобием:

$$p(\mathbf{x}, \{\mathbf{z}_k, \Lambda_k, L_k^{-1}\}_{k=1}^K) = \sum_{k=1}^K p_k N(\ln L_k^{-1} | 0, I) N(\ln \Lambda_k^{-1} | 0, I), \quad (20)$$

$$p_k = \pi_k(\mathbf{x}, W_k) N(\mathbf{x} | W_k \mathbf{z}_k, L_k^{-1}) N(\mathbf{z}_k | \mathbf{m}_k(\mathbf{x}, W_k), \Lambda_k^{-1}), \quad W_k \mathbf{z}_k = \text{vec}(T_k(\mathbf{z}_k)).$$

Текущая реализация предполагает, что параметры $\{W_k\}_{k=1}^K$ являются числовыми величинами, однако возможно построение более общей версии разложения, если моделировать их как случайные величины.

Аналогично детерминистической модели, введем групповую часть разложения, \mathbf{z}_g , а также введем обозначение $\hat{\mathbf{z}}_k = [\mathbf{z}_k \mathbf{z}_g]^T$. Полагая независимость обеих частей скрытого представления $\mathbf{z}_g \perp \mathbf{z}_k$, получаем $p(\hat{\mathbf{z}}_k) = p(\mathbf{z}_k, \mathbf{z}_g) = p(\mathbf{z}_g) p(\mathbf{z}_k)$, и используя свойства нормального распределения, приходим к следующему виду правдоподобия:

$$p_k(\mathbf{x} | \hat{\mathbf{z}}_k) = N(\mathbf{x} | W_k \mathbf{z}_k + W_g \mathbf{z}_g, L_k^{-1} + L_g^{-1}), \quad k = \overline{1, K}, \quad (21)$$

где W_g – параметры разложения, отвечающие общему скрытому представлению \mathbf{z}_g , L_g^{-1} – ковариационная матрица для группового слагаемого.

4.2. Априорное решающее правило

Рассмотрим подробнее априорные вероятности принадлежности к кластерам, $\{\pi_k\}_{k=1}^K$. В общем виде они могут иметь сложную структуру, учитывающую дополнительные зависимости, и одной из наиболее важных является зависимость от входных данных \mathbf{x} , $\pi_k \equiv \pi_k(\mathbf{x}, \boldsymbol{\theta}_{\pi_k})$. От подобной функциональной зависимости требуется выполнение условия нормировки, т.е. $\pi_k(\mathbf{x}, \boldsymbol{\theta}_{\pi_k}) \geq 0$, $\sum_{k=1}^K \pi_k(\mathbf{x}, \boldsymbol{\theta}_{\pi_k}) = 1$. Распространенный способ учета такого условия состоит в использовании многопеременной логистической функции (softmax) на результате отображения некоторой (дифференцируемой) функции $a_k(\mathbf{x}, \boldsymbol{\theta}_{\pi_k})$ (см. [24]):

$$\pi_k \equiv \pi_k(\mathbf{x}, \boldsymbol{\theta}) = \text{softmax}_k(a(\mathbf{x}, \boldsymbol{\theta}_{\pi})) = \frac{\exp(a_k(\mathbf{x}, \boldsymbol{\theta}_{\pi_k}))}{\sum_i \exp(a_i(\mathbf{x}, \boldsymbol{\theta}_{\pi_i}))}. \quad (22)$$

Смеси распределений с параметризованными априорными вероятностями известны как модели коллектива экспертов (см. [24]). В случае, если для нормализации используется функция $\text{softmax}(a(\mathbf{x}, \boldsymbol{\theta}_{\pi}))$, отображение $a(\mathbf{x}, \boldsymbol{\theta}_{\pi})$ должно быть задано таким образом, чтобы значения элементов вектора-результата были прямо пропорциональны близости наблюдения \mathbf{x} к соответствующим кластерам. Выбор параметризации для предложенной модели на основе ВТД был мотивирован результатами работы [19] и нашим опытом из [27], в которой аналогичное отображение использовалось для построения классификаторов на основе разложений Таккера, которые, в частности, позволили получить более устойчивые результаты при изменении оборудования сбора данных и метода химической экстракции в сравнении с рядом иных рассмотренных классификаторов.

Параметризация основана на главном угле (см. [28]) между пространствами источников (столбцы фактор-матрицы для соответствующей оси) и соответствующей матризацией наблюдения. Пусть имеется d -мерное тензорное наблюдение X_i , матрица развертки параметров $A_k^{(m)}$ при выбранной моде источников $m \in \{1, 2, \dots, d\}$, $k = \overline{1, K}$, и их сингулярные разложения, $(X_i)_{(m)} = U_i \Sigma_i V_i^T$ и $A_k^{(m)} = U_k \Sigma_k V_k^T$, тогда отображение $a_k(\mathbf{x}, \boldsymbol{\theta}_{\pi_k})$ может быть вычислено следующим образом:

$$a_k(X_i, A_k^{(m)}) = \tilde{\mathbf{u}}^T \Sigma_k^{-1} V_k^T (A_k^{(m)})^T (X_i)_{(m)} V_i \Sigma_i^{-1} \tilde{\mathbf{v}} = \tilde{\mathbf{u}}^T U_k^T U_i \tilde{\mathbf{v}} = \sigma_{\max}(U_k^T U_i), \quad (23)$$

где $\tilde{\mathbf{u}}, \tilde{\mathbf{v}}$ – левый и правый сингулярные векторы для матрицы $U_k^T U_i$, отвечающие ее максимальному сингулярному значению $\sigma_{\max}(\cdot)$.

5. ОБУЧЕНИЕ МОДЕЛЕЙ

5.1. Нелинейная задача наименьших квадратов

Для предложенных детерминистических моделей GLRO и GTLD вычисление параметров производится через условную оптимизацию нелинейных наименьших квадратов. Для заданного набора d -мерных тензоров $\{X_i\}_{i=1}^N$ одинаковых размеров, объединенных вдоль групповой оси в $(d + 1)$ -мерный тензор X , задача поиска параметров $\boldsymbol{\theta}$ принимает следующий вид:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \|X - T(\boldsymbol{\theta})\|_F^2 &= \min_{\boldsymbol{\theta}_g, \{\boldsymbol{\theta}_i\}_{i=1}^N} \left\| X - \left(T(\boldsymbol{\theta}_g) + \sum_{i=1}^N T(\boldsymbol{\theta}_i) \right) \right\|_F^2 \\ \text{s.t. } h(\boldsymbol{\theta}_g, \{\boldsymbol{\theta}_i\}_{i=1}^N) &= 0, \end{aligned} \quad (24)$$

где через $\boldsymbol{\theta}_g, \{\boldsymbol{\theta}_i\}_{i=1}^N$ обозначены параметры общих и индивидуальных частей соответственно (согласно (6) или (7)), и в выражении $h(\boldsymbol{\theta}_g, \{\boldsymbol{\theta}_i\}_{i=1}^N) = 0$ собраны условия на фактор-матрицы групповой моды и условия отделимости (9).

Среди существующих методов оптимизации параметров в задаче нелинейных наименьших квадратов был использован ряд методов первого и второго порядка (выражения для градиентов и матрицы Гессе приведены в п. 10.2., 10.3.), а также метод попеременных наименьших квадратов (ALS). Для вычисления параметров различных тензорных разложений существуют готовые Матлаб пакеты (например, Tensorlab, см. [29]), однако для проведения исследований из области машинного обучения более распространено использование языка Python, на котором и был реализован вычислительный код для настоящей работы.

В приложении моделей к реальным данным становится критичным учитывать вычислительные затраты по времени, поэтому мы ограничили число итераций до 10. Предварительные эксперименты на наборе данных ETN80 показали, что наиболее быстрым из рассмотренных методов является проекционный ALS, при этом в случае использования методов второго порядка не наблюдалось значимого прироста качества приближения. По этим причинам для последующих экспериментов был использован проекционный метод ALS.

5.2. Стохастический вариационный вывод (SVI)

Вероятностные модели способны ассимилировать новые наблюдения и делать предсказания с помощью статистического вывода. Для моделей со скрытым представлением \mathbf{z} в общем случае для этого требуется вычисление апостериорного распределения $p(\mathbf{z}|\mathbf{x}) = p(\mathbf{x}, \mathbf{z})/p(\mathbf{x})$. Часто проблему представляет вычисление $p(\mathbf{x})$, и для статистического вывода используются различные приближения. Вариационный вывод представляет собой метод преобразования задачи вывода к задаче оптимизации. Основная идея заключается в замене точного апостериорного распределения на некоторого представителя из определенного семейства распределений, который как можно точнее приближает истинное апостериорное (подробнее о методе стохастического вариационного вывода см., например, [30]).

Непосредственная оптимизация параметров распределения производилась через максимизацию нижней грани правдоподобия (ELBO) (см. [31]). Для заданных моделью правдоподобия

$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{in}})$ и априорного распределения $p(\mathbf{z}|\boldsymbol{\theta}_z)$, $p_{\theta}(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{in}})p(\mathbf{z}|\boldsymbol{\theta}_z)$, а также для вариационного распределения $q(\mathbf{z}|\phi) = q_{\phi}(\mathbf{z})$ вариационная нижняя грань возникает в уравнении $\ln p_{\theta}(\mathbf{x}) = \text{ELBO} + \text{KL}(q_{\phi}(\mathbf{z})\|p_{\theta}(\mathbf{z}|\mathbf{x}))$ и определяется как

$$\text{ELBO} \equiv E_{q_{\phi}(\mathbf{z})}[\ln p_{\theta}(\mathbf{x}, \mathbf{z}) - \ln q_{\phi}(\mathbf{z})], \quad (25)$$

$\text{KL}(q_{\phi}(\mathbf{z})\|p_{\theta}(\mathbf{z}|\mathbf{x})) = \int q_{\phi}(\mathbf{z}) \ln \frac{q_{\phi}(\mathbf{z})}{p_{\theta}(\mathbf{z}|\mathbf{x})} d\mathbf{z}$ – дивергенция Кульбака–Лейблера, статистическая мера различий между двумя распределениями. Важно, что метод ELBO поддерживает обучение по подвыборкам и оценки Монте-Карло для стохастического градиента. Для реализации предложенной модели и оптимизации ее параметров методом стохастического вариационного вывода был использован пакет Pyro (см. [32]), в котором вычисления базируются на построении динамических вычислительных графов с помощью pytorch (см. [33]).

Для модели VBTD (20) было использовано следующее вариационное распределение:

$$p(\{\mathbf{z}_k, \Lambda_k, \sigma_k^2\}_{k=1}^K) = \prod_{k=1}^K \gamma_k^{z_{\pi_k}} \cdot \prod_{k=1}^K N(\mathbf{z}_k | \hat{\mathbf{m}}_k, \hat{\Lambda}_k^{-1}), \quad (26)$$

$$\gamma_k = \frac{p_k}{\sum_l p_l}, \quad z_{\pi_k} \in \{0, 1\}, \quad \sum_{k=1}^K z_{\pi_k} = 1, \quad p(z_{\pi_k} = 1) = \pi_k, \quad p(\mathbf{z}_{\pi}) = \prod_{k=1}^K \pi_k^{z_{\pi_k}},$$

где $N(\mathbf{z}_k | \hat{\mathbf{m}}_k, \hat{\Lambda}_k^{-1})$ – истинные апостериорные распределения для каждой отдельной модели непрерывного смешивания на основе нормального распределения.

6. КРАТКИЙ ОБЗОР ИНЫХ МОДЕЛЕЙ

В [34] представлена практическая методология использования специальных матричных разложений для группового анализа данных. Calhoun и др. строят групповой анализ данных функциональной магнитно-резонансной томографии (фМРТ) на основе метода независимых компонент (ICA). В деталях обсуждаются проблемы выбора числа компонент, моделирования зависимости данных от общих компонент, а также статистической валидации результатов. Метод ориентирован на выделение только общей активности, включает в себя двухшаговое выделение главных компонент (на уровне отдельных субъектов и на групповом уровне, для объединения найденных индивидуальных главных компонент) с последующим выделением независимых компонент. Вкратце модель выглядит следующим образом:

$$X_i \underset{\substack{\text{(PCA, инд.)}}}{\approx} Y_i V_i^T, \quad Z \underset{\substack{\text{(PCA, общ.)}}}{\approx} [Y_1, \dots, Y_N] W, \quad Z \underset{\substack{\text{(ICA)}}}{\approx} S A^T, \quad X_i \underset{\substack{\text{(ICA)}}}{\approx} S A^T W_i^T V_i^T = S(V_i W_i A)^T = S M_i^T, \quad (27)$$

где столбцы матрицы S – источники, M_i – матрица смешивания для наблюдения X_i , $i = \overline{1, N}$.

Несколькими годами позже появился обзор [18] прикладных исследований, использующих групповой метод независимых компонент (GICA), в частности, для анализа фМРТ данных, вызванных потенциалов, данных точечного нуклеотидного полиморфизма. В частности, в обзоре перечислены исследования, в которых метод независимых компонент адаптирован к задаче группового анализа данных; подходы сгруппированы в пять категорий: (1) построение отдельных факторизаций для каждого отдельного субъекта с последующим поиском взаимосвязей между ними; (2)–(3) объединение данных по одной из осей (пространственный/временной ICA); (4) усреднение данных по групповой оси; (5) тензорные разложения. Кроме того, авторы демонстрируют возможность использования ICA для мультимодального (т.е. включающего разнородные данные) анализа через модели общего (joint) и параллельного (parallel) ICA.

Другая известная модель, тензорный вероятностный метод независимых компонент (TRICA), основана на совмещении канонического разложения и вероятностного ICA (см. [10]). В данном подходе нет строгого разделения параметров на общие и индивидуальные, а матрица смешивания имеет структуру произведения Хатри–Рао между фактор-матрицами групповой оси и оси времени.

Позднее данные методы были рассмотрены с более общей точки зрения Guo, Pagnoni (см. [35]). В работе предложен модифицированный EM-алгоритм для максимизации правдоподобия с учетом трех вариантов структуры матрицы смешивания:

$$M_{\text{TPICA}} = M^{\text{m}_{\text{subj}}} \otimes M^{\text{m}_{\text{time}}}, \quad M_{\text{GICA}} = [M_1 \dots M_N], \quad M_{\text{gTPICA}} = \begin{bmatrix} M_1^{\text{m}_{\text{subj}}} \otimes M_1^{\text{m}_{\text{time}}} \\ \vdots \\ M_N^{\text{m}_{\text{subj}}} \otimes M_N^{\text{m}_{\text{time}}} \end{bmatrix}. \quad (28)$$

Дополнительно авторами рассмотрена методология выбора оптимальной структуры матрицы смешивания на основе критерия отношения правдоподобия, валидация которой была произведена на фМРТ данных.

Работа [36] посвящена построению группового анализа данных электрической активности мозга (ЭЭГ) на основе неотрицательных матричных разложений с учетом как групповой, так и индивидуальных составляющих: разделение индивидуальных фактор-матриц $\{S_i^{(I)}\}_{i=1}^N$, как и увеличение степени схожести между групповыми фактор-матрицами $\{S_i^{(C)}\}_{i=1}^N$, достигалось с помощью введения соответствующих слагаемых в оптимизируемый функционал:

$$\min_{S_i^{(C)}, S_i^{(I)}, M_i} \sum_{i=1}^N \left[\|X_i - [S_i^{(C)}, S_i^{(I)}] M_i^T\|_F^2 + \gamma \| [S_i^{(C)}, S_i^{(I)}] \|_F^2 + \sum_{j=1}^{i-1} \left(\alpha \|S_i^{(C)} - S_j^{(C)}\|_F^2 - \beta \|S_i^{(I)} - S_j^{(I)}\|_F^2 \right) \right], \quad \text{s.t.} \quad S_i^{(C)}, S_i^{(I)}, M_i \geq 0. \quad (29)$$

В [37] представлен вариант модели группового анализа данных на основе неотрицательного разреженного канонического разложения и концепта “связанных тензорных разложений”, в которой с помощью иерархического ALS алгоритма (вариант блочно-координатного спуска, каждая итерация которого состоит в последовательном обновлении строк/столбцов фактор-матриц, см. [38], [39]) оцениваются статистически независимые индивидуальные компоненты. Оптимизационная задача для вычисления разложения имеет вид (условия неотрицательности и разреженности опущены):

$$\min \sum_{i=1}^N \|X_i - [\Lambda; U_1^{(i)}, \dots, U_d^{(i)}]\|_F^2 \quad (30)$$

s.t. $U_k^{(i)}|_{i=1}^N = U_k, \quad \|U_k[:, r]\|_F = 1, \quad r = \overline{1, R}, \quad k = \overline{1, K}, \quad K < d, \quad \Lambda = \text{Diag}(\lambda_1, \dots, \lambda_R).$

В [40] была предложена модель на основе разложения Таккера с двумя фактор-матрицами для случая 3D данных. Общие компоненты моделируются как первые C столбцов левой сингулярной матрицы соответствующей фактор-матрицы, а индивидуальные компоненты вычисляются как первые I_k столбцов левой сингулярной матрицы для проекции той же фактор-матрицы на пространство, ортогональное пространству общих компонент.

В [19] был предложен метод выделения общей и индивидуальной активностей на основе матричной модели (COBE). Предполагаемая структура данных схожа с рассматриваемой в [36], но модель выделяет общие источники иным образом:

$$\min_{A, A_i} \sum_{i=1}^N \|X_i - \hat{A} \hat{B}_i^T - A_i B_i^T\|_F^2 \quad (31)$$

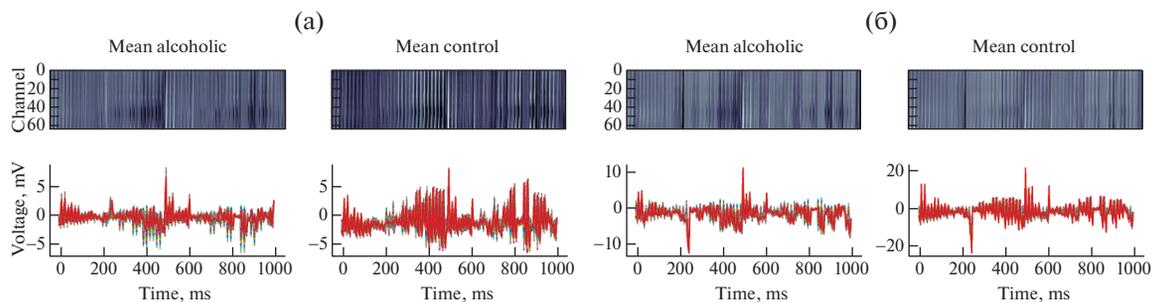
s.t. $\hat{A}^T \hat{A} = I_C, \quad A_i^T A_i = I_{n_i - C}, \quad \hat{A}^T A_i = 0, \quad i = \overline{1, N},$

где \hat{A} – общие источники (C столбцов), A_i – индивидуальные источники.

Из существующих расширений ВТД выделим вариант разложения со слагаемыми в формате Таккера, предложенный в [41] для случая совмещенных фактор-матриц. Предложенные вычислительные алгоритмы опираются на расширенный QZ алгоритм и метод Якоби. Кроме того, важно отметить, что идея использования ВТД разложения с разным типом слагаемых не является новой, например, в [42] были использованы одновременно слагаемые в формате Таккера и каноническом формате для построения приближений классических потенциалов.



Фиг. 1. ETh80, изображения разных классов (а) и выборочные ракурсы 10 объектов класса “чашка” (б). Набор данных доступен по адресу: <http://datasets.d2.mpi-inf.mpg.de/eth80/eth80-contours.tgz>



Фиг. 2. Усредненные данные ЭЭГ для первого (а) и третьего (б) экспериментов; на изображениях сверху сигналы изображены в 2D виде, где более светлый цвет соответствует большим значениям магнитуды; разные цвета на изображениях снизу отвечают разным каналам. Набор данных доступен по ссылке: <https://archive.ics.uci.edu/ml/datasets/eeg+database>

7. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

7.1. Программная реализация

Все модели и вычислительные эксперименты были реализованы на языке программирования Python с использованием дистрибутива Anaconda (см. [43]), который включает в себя различные пакеты для научных вычислений. В данном исследовании были использованы следующие пакеты: numpy [44], scipy [45], pandas [46], scikit-learn [47], matplotlib [48], seaborn [49], pyro [32]. Весь код опубликован в следующих репозиториях: <https://github.com/kharyuk/gbtd>; <https://github.com/kharyuk/vbtd>. Эксперименты были систематизированы с помощью Jupyter Notebooks (см. [50]).

7.2. Наборы данных

ETh-80. ETh80 (см. [51]) состоит из изображений различных объектов, снятых с разных ракурсов. Всего доступно восемь классов: “яблоко”, “машина”, “корова”, “чашка”, “собака”, “лошадь”, “груша”, “томат”; каждый класс включает 10 объектов (наблюдений) (фиг. 1). Полученный в результате набор данных имеет размеры $(N_{\text{samples}}, N_{\text{pixels}}, N_{\text{angles}}, N_{\text{colors}})$, где $N_{\text{angles}} = 41$. Разрешение изображений было понижено до 32×32 , откуда $N_{\text{pixels}} = 1024$.

SMNI EEG (ERP). Набор данных SMNI EEG состоит из измерений электрической активности мозга по $N_{\text{electrode}} = 64$ электродам для двух групп субъектов: страдающих алкогольной зависимостью и контрольной группы. Согласно описанию, данные были записаны с частотой 256 Гц и нарезаны на фрагменты длиной в 1 с, содержащие момент предъявления стимула. Данные были собраны для $N_{\text{condition}} = 3$ условий: предъявление одиночного стимула (изображения), двух совпадающих и двух различных стимулов. Данные были дополнительно усреднены по повторностям, помеченные как испорченные образцы были исключены, и итоговый набор данных был представлен как тензор размерами $(N_{\text{subject}}, N_{\text{time}}, N_{\text{electrode}}, N_{\text{condition}})$, где $N_{\text{subject}} = 119$, из них 76 субъектов из первой группы, 43 из контрольной (фиг. 2).

Таблица 1. Сравнительные результаты моделей с подобранными гиперпараметрами для набора данных ETN80

		ARI	AMI	FMI
НАС	Raw ¹	.568	.749	.638
	COBE ¹	.643 ± .023	.774 ± .016	.691 ± .019
	COBE ²	.573 ± .013	.780 ± .005	.663 ± .009
	GICA ¹	.722 ± .000	.794 ± .000	.753 ± .000
	GICA ²	.577 ± .000	.802 ± .000	.661 ± .000
	GLRO ¹	.472 ± .093	.665 ± .060	.559 ± .074
	GLRO ²	.471 ± .098	.693 ± .076	.580 ± .067
	GTLD ¹	.497 ± .071	.668 ± .047	.573 ± .054
	GTLD ²	.493 ± .072	.699 ± .043	.585 ± .047
Kmeans	Raw	.533 ± .068	.690 ± .047	.597 ± .055
	COBE ³	.571 ± .049	.729 ± .030	.635 ± .037
	GICA ³	.559 ± .057	.712 ± .042	.624 ± .046
	GLRO ²	.461 ± .077	.645 ± .062	.543 ± .062
	GLRO ³	.445 ± .065	.645 ± .047	.535 ± .050
	GTLD ³	.473 ± .065	.652 ± .050	.550 ± .052
GMM	Raw	.475 ± .072	.639 ± .064	.549 ± .060
	COBE ³	.333 ± .262	.438 ± .341	.376 ± .293
	GICA ⁴	.384 ± .229	.513 ± .301	.439 ± .258
	GLRO ²	.473 ± .079	.651 ± .065	.553 ± .063
	GLRO ³	.466 ± .072	.662 ± .054	.554 ± .054
	GTLD ³	.476 ± .073	.656 ± .056	.555 ± .058

Примечание. Формат ячеек таблицы: “ $x \pm y$ ”, где x – среднее значение, y – стандартное отклонение. Для НАС за $R_{average}$ и $R_{complete}$ обозначены пересчет расстояний через усреднение и взятие максимума, за ρ_{corr} , ρ_{cos} , $\rho_{Canb.}$ – корреляция, косинусное расстояние и расстояние Канберра; Raw¹: ρ_{corr} , $R_{average}$; COBE¹: $r_c = 2$, ρ_{cos}/ρ_{corr} , $R_{complete}$; COBE²: $r_c = 2$, ρ_{cos}/ρ_{corr} , $R_{average}$; COBE³: $r_c = 4$; GICA¹: $r_c = 2$, $r_i = \overline{3,5}$, ρ_{cos}/ρ_{corr} , $R_{complete}$; GICA²: $r_c = 2$, $r_i = \overline{4,5}$, $\rho_{Canb.}$, $R_{average}$; GICA³: $r_c = 4$, $r_i = 3$; GICA⁴: $r_c = 2$, $r_i = 2$; GLRO¹: $r_c = 4$, $r_i = 4$, ρ_{corr} , $R_{complete}$; GLRO²: $r_c = 4$, $r_i = 3$, для НАС – $\rho_{Canb.}$, $R_{average}$; GLRO³: $r_c = 5$, $r_i = 2$; GTLD¹: $r_c = 5$, $r_i = 3$, ρ_{corr} , $R_{complete}$; GTLD²: $r_c = 4$, $r_i = 4$, $\rho_{Canb.}$, $R_{average}$; GTLD³: $r_c = 5$, $r_i = 4$.

7.3. Оценка качества методов кластеризации

Кластеризация представляет собой одну из стандартных задач машинного обучения, в которой требуется сгруппировать различные образцы по заданному критерию близости. Более формально, указывается некоторый функционал близости $\rho(X_i, X_j)$, и для входных данных $\{X_i\}_{i=1}^N$ необходимо ввести разбиение на непересекающиеся подмножества (кластеры) таким образом, что объекты внутри одного кластера были более близки друг к другу, чем к образцам из прочих кластеров, в смысле $\rho(X_i, X_j)$.

Если для данных известны истинные метки, качество кластеризации можно измерить следующим образом. Рассмотрим два разбиения, $U = \{U_i\}$, $V = \{V_j\}$, где U отвечает истинным меткам, и введем на них четыре стандартные величины: число истинно-положительных (TP), истинно-отрицательных (TN), ложноотрицательных (FN), ложноположительных (FP) пар:

$$\begin{aligned}
 TP &= |\{(a_k, a_l) | a_k, a_l \in U_i, a_k, a_l \in V_j\}|, & TN &= |\{(a_k, a_l) | a_k \in U_i, a_l \in U_{i_2}, a_k \in V_{j_1}, a_l \in V_{j_2}\}|, \\
 FP &= |\{(a_k, a_l) | a_k \in U_i, a_l \in U_{i_2}, a_k, a_l \in V_j\}|, & FN &= |\{(a_k, a_l) | a_k, a_l \in U_i, a_k \in V_{j_1}, a_l \in V_{j_2}\}|.
 \end{aligned}
 \tag{32}$$

Также определим следующие вспомогательные величины:

$$P_{ij} = |U_i \cap V_j|/N, \quad P_i = |U_i|/N, \quad P_j = |V_j|/N, \quad H(W) = E[-\log P(W)], \tag{33}$$

где W – некоторое разбиение, $H(\cdot)$ – энтропия.

Таблица 2. Сравнительные результаты моделей с подобранными гиперпараметрами для набора данных SMNI EEG

		ARI	AMI	FMI
НАС	Raw ¹	.161	.104	.601
	COBE ¹	.092 ± .060	.057 ± .044	.577 ± .029
	COBE ²	.077 ± .053	.057 ± .032	.591 ± .047
	GICA ¹	.253 ± .000	.198 ± .000	.715 ± .000
	GLRO ¹	.048 ± .075	.036 ± .044	.621 ± .059
	GTLD ¹	.060 ± .078	.046 ± .037	.597 ± .052
	GTLD ²	.009 ± .047	.058 ± .060	.568 ± .038
Kmeans	Raw	.196 ± .015	.123 ± .009	.624 ± .009
	COBE ²	.168 ± .059	.106 ± .037	.608 ± .033
	GICA ²	.236 ± .004	.153 ± .005	.646 ± .006
	GLRO ²	.039 ± .054	.027 ± .037	.676 ± .080
	GTLD ²	.126 ± .094	.083 ± .063	.674 ± .051
GMM	Raw	.156 ± .049	.103 ± .032	.597 ± .027
	COBE ²	.151 ± .060	.101 ± .038	.597 ± .031
	GICA ²	.196 ± .092	.127 ± .059	.630 ± .147
	GLRO ³	.037 ± .059	.018 ± .044	.709 ± .037
	GLRO ⁴	.033 ± .021	.021 ± .021	.720 ± .006
	GTLD ²	.052 ± .056	.038 ± .049	.712 ± .030
	GTLD ³	.058 ± .060	.034 ± .041	.704 ± .032

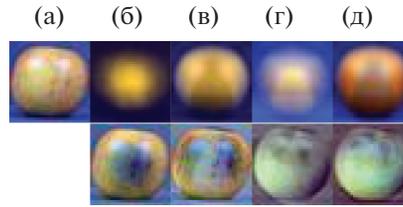
Примечание. Формат ячеек таблицы: “ $x \pm y$ ”, где x – среднее значение, y – стандартное отклонение. Для НАС за R_{complete} обозначен пересчет расстояний через взятие максимума, за ρ_{corr} , ρ_{cos} , ρ_{RBF} , ρ_{l_2} , ρ_{Canb} . – корреляция, косинусное, RBF, l_2 расстояния и расстояние Канберра; Raw¹: для НАС – $\rho_{l_2}/\rho_{\text{cos}}/\rho_{\text{RBF}}/\rho_{\text{corr}}$, R_{complete} ; COBE¹: $r_c = 3$, $\rho_{l_2}/\rho_{\text{cos}}/\rho_{\text{RBF}}/\rho_{\text{corr}}$, R_{complete} ; COBE²: $r_c = 4$, для НАС – $\rho_{l_2}/\rho_{\text{cos}}/\rho_{\text{RBF}}/\rho_{\text{corr}}$, R_{complete} ; GICA¹: $r_c = 2$, $r_i = 1$, ρ_{l_2} , R_{complete} ; GICA²: $r_c = 1$, $r_i = 2$; GLRO¹: $r_c = 4$, $r_i = 5$, ρ_{Canb} , R_{complete} ; GLRO²: $r_c = 3$, $r_i = 5$; GLRO³: $r_c = 3$, $r_i = 4$; GLRO⁴: $r_c = 1$, $r_i = 3$; GTLD¹: $r_c = 1$, $r_i = 3$, ρ_{corr} , R_{complete} ; GTLD²: $r_c = 4$, $r_i = 3$, для НАС – ρ_{Canb} , R_{complete} ; GTLD³: $r_c = 2$, $r_i = 1$.

Используя величины (32), (33), определим три стандартных показателя качества, индекс Фолькс–Мэллоуз (FMI), как геометрическое среднее между точностью и полнотой, скорректированный индекс Рэнда (ARI), который схож с долей верных ответов, и скорректированная вза-

имная информация (AMI), $C_N^2 = \binom{N}{2}$:

$$\begin{aligned}
 \text{RI} &= (TP + TN)/C_N^2, & \text{MI} &= \sum_{i=1}^I \sum_{j=1}^J P_{ij} \log \frac{P_{ij}}{P_i P_j}, \\
 \text{ARI} &= \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}, & \text{AMI} &= \frac{\text{MI} - E[\text{MI}]}{\max(H(U), H(V)) - E[\text{MI}]}, \\
 \text{FMI} &= \sqrt{\frac{TP}{TP + FP} \frac{TP}{TP + FN}}.
 \end{aligned} \tag{34}$$

В силу того что модели GLRO и GTLD были построены как расширение модели, предложенной в [19] (будем ссылаться на нее по названию алгоритма COBE), модель COBE была также использована для сравнения. Другой альтернативой для сравнения качества была выбрана модель группового метода независимых компонент GICA (см. [18]). Оригинальная версия модели была разработана для выделения только общей активности; в наших экспериментах была воспроизведена процедура выделения общих независимых компонент, дополненная вычитанием восстановленных общих частей из исходных данных (т.е. контрастированием). Выделение индивиду-



Фиг. 3. Пример изображения из набора EТН80 и соответствующие ему индивидуальная и групповая части: (а) оригинал; (б) COBE, $r_c = 2$; (в) GICA, $r_c = 2, r_i = 4$; (г) GLRO, $r_c = 4, r_i = 3$; (д) GTLD, $r_c = 5, r_i = 3$. (б)–(д) – Верхний ряд изображений отвечает групповым частям, нижний – индивидуальным.

альных частей в случае COBE было проведено способом, описанным в оригинальной статье, для GLRO и GTLD были использованы соответствующие слагаемые полного разложения. Для моделей были использованы разные сочетания общих рангов, $r_c = \overline{1,5}$, и рангов индивидуальных, $r_i = \overline{1,5}$ (кроме COBE, где используются только r_c). Для GICA r_c соответствовал одновременно числу вторичных главных компонент и числу независимых компонент. В случае GTLD все ранги Таккера полагались равными r_c .

Контрастированные и необработанные данные были кластеризованы с помощью следующих алгоритмов: иерархическая агломеративная кластеризация HAC (см. [52]), K-средних (см. [52]) и кластеризации на основе модели смеси гауссовских распределений GMM (см. [52]) (с диагональными ковариационными матрицами). В табл. 1, 2 представлены результаты, полученные для EТН80 и SMNI EEG соответственно. Результаты для всех методов контрастирования были усреднены по 10 запускам; для алгоритмов K-средних/GMM результаты дополнительно усреднены по 20 разным инициализациям для каждого запуска.

На фиг. 3 приведены примеры общих и индивидуальных частей, полученных разными способами для одного из изображений из данных EТН80.

Кластеризация с помощью предложенной VBTD модели может быть выполнена через использование как априорного правила $\pi_k(x, \theta_{\pi_k})$, так и через вычисление апостериорных вероятностей γ_k . На наборах данных были протестированы различные конфигурации VBTD моделей с фиксированными рангами (равными 3). Индивидуальные слагаемые моделировались в (Lr, 1) формате. Результаты для лучших пяти конфигураций на апостериорном правиле приведены в табл. 3 в смысле AMI и в табл. 4 в смысле ARI. Все эксперименты проведены с ограничением на число итераций $N_{it} = 100$ при фиксированном случайном состоянии.

Таблица 3. VBTD, топ-5 конфигураций в смысле апостериорного AMI

Данные	Формат слагаемых		Шум, L_k^{-1}	Априорное правило			Апостериорное правило		
	(инд.)	(груп.)		AMI	ARI	FMI	AMI	ARI	FMI
EТН80	(Lr, 1)	–	изотр.	.646 (.619)	.335 (.320)	.512 (.498)	.683 (.640)	.412 (.404)	.560 (.538)
	(Lr, 1)	ТТ	диаг.	.585 (.529)	.383 (.272)	.522 (.486)	.650 (.340)	.392 (.116)	.544 (.378)
	(Lr, 1)	Таккер (ядро)	изотр.	.709 (.458)	.508 (.257)	.616 (.455)	.646 (.420)	.454 (.175)	.567 (.405)
	(Lr, 1)	ТТ	изотр.	.427 (.382)	.169 (.139)	.412 (.389)	.623 (.496)	.395 (.298)	.527 (.429)
	(Lr, 1)	–	диаг.	.541 (.507)	.255 (.235)	.484 (.467)	.600 (.541)	.336 (.255)	.531 (.484)
SMNI	Таккер (ядро)	–	изотр.	.032 (–.005)	.081 (.001)	.641 (.528)	.166 (.166)	.215 (.215)	.708 (.708)
	Таккер (фактор)	–	изотр.	.047 (.013)	.094 (.039)	.705 (.697)	.139 (.000)	.060 (.000)	.731 (.731)
	Таккер (ядро)	–	диаг.	.026 (–.010)	.053 (.010)	.712 (.710)	.128 (.000)	.186 (.000)	.731 (.731)
	ТТ (Lr, 1)	–	диаг.	.017 (.000)	.046 (.026)	.644 (.644)	.104 (.091)	.161 (.147)	.601 (.596)
	(Lr, 1)	–	изотр.	.044 (.001)	.098 (.020)	.669 (.560)	.090 (.006)	.162 (.004)	.660 (.585)

Примечание. Формат ячеек таблицы: “x(y)”, где x – максимальное значение индекса, y – после 100 итераций.

Таблица 4. VBTD, топ-5 конфигураций в смысле апостериорного ARI

Данные	Формат слагаемых		Шум, L_k^{-1}	Априорное правило			Апостериорное правило		
	(инд.)	(груп.)		AMI	ARI	FMI	AMI	ARI	FMI
ETH80	(Lr, 1)	Таккер (ядро)	изотр.	.709 (.458)	.508 (.257)	.616 (.455)	.646 (.420)	.454 (.175)	.567 (.405)
	TT	—	диаг.	.310 (.238)	.188 (.091)	.394 (.344)	.555 (.463)	.433 (.271)	.501 (.410)
	(Lr, 1)	—	изотр.	.646 (.619)	.335 (.320)	.512 (.498)	.683 (.640)	.412 (.404)	.560 (.538)
	(Lr, 1)	TT	изотр.	.427 (.382)	.169 (.139)	.412 (.389)	.623 (.496)	.395 (.298)	.527 (.429)
SMNI	(Lr, 1)	TT	диаг.	.585 (.529)	.383 (.272)	.522 (.486)	.650 (.340)	.392 (.116)	.544 (.378)
	Таккер (ядро)	—	изотр.	.032 (−.005)	.081 (.001)	.641 (.528)	.166 (.166)	.215 (.215)	.708 (.708)
	Таккер (ядро)	—	диаг.	.026 (−.010)	.053 (.010)	.712 (.710)	.128 (.000)	.186 (.000)	.731 (.731)
	(Lr, 1)	—	изотр.	.044 (.001)	.098 (.020)	.669 (.560)	.090 (.006)	.162 (.004)	.660 (.585)
	TT	—	диаг.	.017 (.000)	.046 (.026)	.644 (.644)	.104 (.091)	.161 (.147)	.601 (.596)
	(Lr, 1)	(Lr, 1)	диаг.	.035 (.000)	.074 (−.021)	.720 (.689)	.071 (.057)	.136 (.117)	.731 (.640)

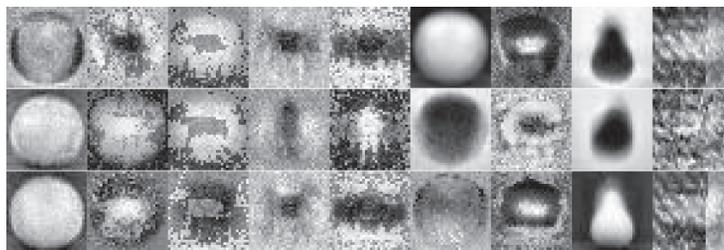
Примечание. Формат ячеек таблицы: “ $x(y)$ ”, где x – максимальное значение индекса, y – после 100 итераций.

В отличие от детерминистических моделей, в которых каждый отдельный образец раскладывается на общие и индивидуальные составляющие, VBTD модель выделяет индивидуальные части для отдельного кластера (своеобразные кластерные “тренды”). Эта особенность позволяет визуализировать выделенные источники для каждого отдельного кластера. На фиг. 4 изображены примеры для набора данных ETH80.

8. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Если ориентироваться на индекс FMI, все алгоритмы показывают достаточно оптимистичные результаты. Однако индекс не является устойчивым к предсказаниям с константными значениями, что делает его особенно ненадежным в случае несбалансированных наборов данных с малым числом кластеров (SMNI EEG). В то же время показатели AMI и ARI отражают качество кластеризации более надежным образом. Следует отметить особенность набора данных SMNI EEG: в целом, все использованные алгоритмы показали невысокое качество кластеризации; предполагаем, что результаты можно улучшить с помощью дополнительной предварительной обработки данных, которая выделит более тонкие различия между группами субъектов, но этот вопрос оставлен за рамками работы.

GLRO и GTLD модели показали несколько худшие результаты в сравнении с другими рассмотренными методами, что может быть объяснено особенной структурой контрастированных данных. В случае GICA или COBE предположение о малоранговости применялось только к об-



Фиг. 4. Примеры источников, построенных с помощью VBTD. Конфигурация модели: (Lr, 1) формат индивидуальных слагаемых (ранг 3, $P = 2$), TT формат для общей части ($r = (1, 3, 3, 3)$), диагональная ковариационная матрица шума. Столбцы отвечают кластерам (последний из них – общей части), строки – отдельным восстановленным источникам (без упорядочивания).

щей части разложения, которая затем вычиталась из данных. В моделях GLRO и GTLD малоранговое представление строится также и для индивидуальных частей. Предварительно настроенные VBTD модели показали качество кластеризации, сравнимое с GICA и COBE. Вместе с тем в текущей реализации модели мы столкнулись с нестабильностями процесса обучения, которые в ряде случаев проявляют себя заметным зазором между лучшими значениями показателей и значениями на последней итерации. Хотя рассмотренные матричные модели имеют высокие показатели качества, данные модели опираются на избыточное представление данных. В отличие от них, предложенная VBTD модель не требует вычисления параметров для каждого отдельного образца, что делает ее предпочтительнее с точки зрения затрат на хранение параметров. Кроме того, в модели может быть учтено масштабирование по числу размерностей (например, с помощью формата квантизированного тензорного произведения QTT, см. [53]). Используемый метод для вычисления параметров, стохастический вариационный вывод через оптимизацию вариационной нижней грани правдоподобия, поддерживает обучение по подвыборкам, что делает модель масштабируемой в смысле числа данных.

Важно отметить, что хотя в некоторых случаях удалось получить высокое качество кластеризации на исходных данных без какой-либо предварительной обработки, подходы такого плана не дают возможности для последующей интерпретации внутренней структуры данных.

9. ВЫВОДЫ И ДАЛЬНЕЙШАЯ РАБОТА

Для связанного многомерного анализа компонент были предложены новые модели на основе блочного тензорного разложения объединенных вдоль новой оси данных с дополнительными условиями на параметры (GLRO, GTLD), а также вероятностная модель вариационного блочного тензорного разложения со слагаемыми в различных форматах. Вычисление параметров детерминистических моделей производилось через решение задачи нелинейных наименьших квадратов; для вычисления параметров вероятностной модели был использован метод стохастического вариационного вывода.

Как и матричные модели GICA и COBE, предложенные модели на основе BTD показали способность к приближению групповой и индивидуальной активности, что было проверено методом кластеризации контрастированных (т.е. сохранивших только индивидуальную часть) данных. Заметим, что результаты кластеризации зависят не только от метода контрастирования, но и от метода кластеризации. Вместе с тем, если общие и индивидуальные компоненты разложения потребуется использовать в смысле, отличном от признакового представления для кластеризации, то для выбора модели потребуется использование иного способа тестирования.

Одним из недостатков предложенных моделей GLRO и GTLD является большая вычислительная сложность, которая может быть компенсирована возможностью выделения факторов, специфичных для разных мод, в случае малого числа образцов. Для выделения зависимостей из больших наборов данных более перспективным выбором выглядит VBTD модель.

Дальнейшая работа над VBTD моделью включает рассмотрение распределений компонент, отличных от нормального, решение проблемы выбора рангов или увеличение степени разреженности через вероятностное моделирование, переход к полностью вероятностной версии разложения, в которой все параметры модели рассматриваются как случайные величины. Перспективным направлением является усовершенствование модели через искусственную тензоризацию данных.

ПРИЛОЖЕНИЕ

1. Используемые в работе обозначения

В настоящей статье использовались следующие обозначения.

Определение. Многомерный массив с вещественными элементами $X \in \mathbb{R}^{n_1 \times \dots \times n_d}$ называется (вещественным) *тензором*. Элемент тензора X , отвечающий мульти-индексу (i_1, \dots, i_d) , обозначается как $X[i_1, \dots, i_d]$. Его k -я ось называется также модой размера n_k . Любой тензор может быть записан как вектор-столбец с помощью операции векторизации. Соответствие между индексами оригинала и векторизованного результата задается правилом

$$f(i_1, \dots, i_d) = 1 + \sum_{k=1}^d \left[(i_k - 1) \prod_{l=1}^{k-1} n_l \right], \quad \text{vec}(X)[f(i_1, \dots, i_d)] = X[i_1, \dots, i_d], \quad i_k = \overline{1, n_k}, \quad k = \overline{1, d}.$$

Таблица 5. Используемые в работе обозначения

$\overline{1}$	Массив из единиц	$X_{(k)}$	Матризация тензора X по оси k
$i = \overline{1, I}$	$i \in \{1, \dots, I\}$	A^+	Псевдообращение Мура–Пенроуза
$(\dots)_{(k)}$	Операция без k -го составляющего	$X \times_k Y$	Перемножение тензоров по k -й оси
$\text{diag}(A)$	Главная диагональ матрицы	$\text{diag}(\mathbf{v})$	Диагональная матрица, \mathbf{v} – диагональ
$\text{off diag}(A)$	Матрица A с нулями на диагонали	$\text{Diag}(\mathbf{p})$	Диагональный тензор, \mathbf{p} – диагональ
$\widehat{\text{vec}}(X)$	Векторизация тензора X (строка)	$\text{vec}(X)$	Векторизация тензора X (столбец)
$\langle X, Y \rangle$	Внутреннее произведение	$X \circ Y$	Внешнее произведение двух тензоров
$\llbracket C_1, \dots, C_d \rrbracket$	Канонический формат	$A \odot B$	Матричное произведение Хатри–Рао
$\llbracket G; A_1, \dots, A_d \rrbracket$	Формат Таккера	$A \otimes B$	Матричное кронекерово произведение

Для преобразования тензора в вектор-строку использовано схожее обозначение: $\widehat{\text{vec}}(\cdot)$. Для тензоров размерности d также вводится операция матризации, для чего в случае $d > 2$ объединяют его оси. Операцию можно вводить различными способами, здесь рассмотрим подход с сохранением в неизменном виде одной из осей с номером k . Будем называть такую операцию разверткой по индексу k : $\text{unfold}_k(X^{n_1 \times \dots \times n_k \times \dots \times n_d}) = X_{(k)} = \hat{X}^{n_k \times \prod_{l \neq k} n_l}$. Кроме того, если у тензора есть мода с некоторым номером l , размер n_l которой может быть разложен на множители, $n_l = \prod_{p=1}^P n_{l,p}$, к такому тензору можно применить операцию *тензоризации* (искусственного повышения размерности). Использовано обозначение $\text{Tens}(\cdot)$. Операция требует принять соглашение о пересчете индексов, в данной статье применялся стандартный порядок “по столбцам” (так называемый фортран-порядок).

Пусть имеется два тензора $X \in \mathbb{R}^{n_1^{(x)} \times \dots \times n_k \times \dots \times n_d^{(x)}}$, $Y \in \mathbb{R}^{n_1^{(y)} \times \dots \times n_k \times \dots \times n_d^{(y)}}$ размерностями d_1 и d_2 , и пусть для них имеется общая мода с номером k размером n_k . В этом случае определим операцию *умножения по k -й моде*, $X \times_k Y = Z$, между X и Y по следующему поэлементному правилу:

$$Z[\dots, i_{k-1}, i_{k+1}, \dots, j_{k-1}, j_{k+1}, \dots] = \sum_{l=1}^{n_k} X[\dots, i_{k-1}, l, i_{k+1}, \dots] Y[\dots, j_{k-1}, l, j_{k+1}, \dots],$$

где $i_p = \overline{1, \dots, n_p^{(x)}}$, $j_q = \overline{1, \dots, n_q^{(y)}}$, $p = \overline{1, d_1}$, $q = \overline{1, d_2}$. В литературе данную операцию также называют свертыванием (см. [2]).

Фробениусова норма для тензора X определяется выражением вида $\|X\|_F^2 = \langle \text{vec}(X), \text{vec}(X) \rangle$, где $\langle \cdot, \cdot \rangle$ – скалярное (внутреннее) произведение между двумя векторами. Операцию можно обобщить на случай тензоров без дополнительной их векторизации: результат эквивалентен последовательному свертыванию аргументов-тензоров одинакового размера по всем их осям. Если же имеется некоторое подмножество осей, по которым перемножение выполнять не следует, то будем обозначать такую ситуацию следующим образом, используя скобки внутреннего произведения: $\langle X, Y \rangle_{\{\alpha \in \Omega\}} = X \times_{k \in \{\dots, d\} \setminus \Omega} Y$. Иными словами, подстрочные скобки для операции означают, что операция выполняется для всех объектов (по всем индексам), кроме указанных внутри этих скобок. В дополнение к внутреннему, внешнее произведение $X \circ Y = Z$ между двумя тензорами X, Y определяется как $Z[i_1, \dots, i_{d_1}, j_1, \dots, j_{d_2}] = X[i_1, \dots, i_{d_1}] \cdot Y[j_1, \dots, j_{d_2}]$.

Отметим еще два важных произведения. Кронекерово произведение двух матриц A и B размерами (n_1, n_2) и (m_1, m_2) определяется как $(A \otimes B)[m_1(i-1) + k, m_2(j-1) + l] = A[i, j]B[k, l]$. Если обе матрицы A, B имеют одинаковое количество столбцов (т.е. выполнено $n_2 = m_2$), то для них определено произведение Хатри–Рао, которое удобно использовать в контексте развертки тензоров, представленных в каноническом формате: $(A \odot B)[m_1(i-1) + k, j] = A[i, j]B[k, j]$.

Для удобства ряд обозначений, использованных в работе, приведен в табл. 5. Более подробный разбор базовых операций и тензорных разложений можно найти в [1].

2. Структура матрицы Гессе для ВТД разложения с тремя типами слагаемых

Пусть требуется приблизить данный тензор $T \in \mathbb{R}^{n_1 \times \dots \times n_d}$ в формате (8). Обозначим через \mathbf{x} вектор-столбец, составленный из векторизации всех параметров модели:

$$\mathbf{x} = \left[\dots \widehat{\text{vec}}(C_k^{[s]}) \dots \widehat{\text{vec}}(A_k^{[m]}) \dots \widehat{\text{vec}}(G^{[m]}) \dots \right]^T, \quad (35)$$

где $C_k^{[s]}$ есть k -я фактор-матрица для s -го слагаемого в каноническом формате, $A_k^{[m]}$ есть k -я фактор-матрица для m -го слагаемого в формате Таккера, $G^{[m]}$ – ядро Таккера для m -го слагаемого в формате Таккера, и пусть $F = F(\mathbf{x})$ – восстановленный тензор с помощью текущего приближения параметров:

$$F(\mathbf{x}) = \sum_{m=1}^M \left[G^{[m]}; A_1^{[m]}, \dots, A_d^{[m]} \right] + [C_1, \dots, C_d]$$

(как в (8)). Для дальнейших целей введем дополнительное обозначение $Z(\mathbf{x}) = F(\mathbf{x}) - T$. Соответствующая оптимизационная задача может быть сформулирована в виде задачи нелинейных наименьших квадратов, градиент и матрица Гессе которой имеют специальный вид:

$$\begin{aligned} \min_{\mathbf{x}} f(\mathbf{x}) &= \frac{1}{2} \| F(\mathbf{x}) - T \|_F^2, \quad H[f](\mathbf{x}) = J^T(\mathbf{x})J(\mathbf{x}) + Q(\mathbf{x}), \\ \nabla f(\mathbf{x}) &= J^T(\mathbf{x})\text{vec}(F(\mathbf{x}) - T), \quad Q(\mathbf{x}) = \sum_i \text{vec}(Z(\mathbf{x}))[i] \cdot \nabla^2(\text{vec}(Z(\mathbf{x}))[i]). \end{aligned} \quad (36)$$

Здесь $J(\mathbf{x})$ – матрица Якоби для $f(\mathbf{x})$:

$$J(\mathbf{x}) = \left[(V_{\{k\}}^{[s]} \otimes I_{n_k}) \dots (G_{(k)}^{[m]} V_{\{k\}}^{[m]T} \otimes I_{n_k}) \dots V^{[m]} \dots \right], \quad (37)$$

$k = \overline{1, d}$, $m = \overline{1, M}$, $s = \overline{1, S}$ и $V_{\{k\}}^{[s]} = C_d^{[s]} \odot \dots \odot C_{k+1}^{[s]} \odot C_{k-1}^{[s]} \odot \dots \odot C_1^{[s]}$, $V_{\{k\}}^{[m]} = A_d^{[m]} \otimes \dots \otimes A_{k+1}^{[m]} \otimes A_{k-1}^{[m]} \otimes \dots \otimes A_1^{[m]}$, $V^{[m]} = A_d^{[m]} \otimes \dots \otimes A_1^{[m]}$. Главная часть матрицы Гессе имеет следующую структуру:

$$J^T J = \begin{bmatrix} (\text{Gr}^{\text{CC}})_{k_1, k_2; s_1, s_2; \cdot, \cdot} & (\text{Gr}^{\text{CA}})_{k_1, k_2; s_1, \cdot; \cdot, m_2} & (\text{Gr}^{\text{CG}})_{k_1, \cdot; s_1, \cdot; \cdot, m_2} \\ (\text{Gr}^{\text{AC}})_{k_1, k_2; \cdot, s_2; m_1, \cdot} & (\text{Gr}^{\text{AA}})_{k_1, k_2; \cdot, \cdot; m_1, m_2} & (\text{Gr}^{\text{AG}})_{k_1, \cdot; \cdot, \cdot; m_1, m_2} \\ (\text{Gr}^{\text{GC}})_{k_1, k_2; \cdot, s_2; m_1, \cdot} & (\text{Gr}^{\text{GA}})_{\cdot, k_2; \cdot, \cdot; m_1, m_2} & (\text{Gr}^{\text{GG}})_{\cdot, \cdot; \cdot, \cdot; m_1, m_2} \end{bmatrix}, \quad (38)$$

где $k_1, k_2 = \overline{1, d}$, $m_1, m_2 = \overline{1, M}$, $s_1, s_2 = \overline{1, S}$, и соответствующие блоки имеют вид

$$\begin{aligned} \text{Gr}_{k_1, k_2; s_1, s_2; \cdot, \cdot}^{\text{CC}} &= (V_{\{k_1\}}^{[s_1]T} \otimes I_{n_{k_1}}) P_{1, k_2}^{k_1, 1} (V_{\{k_2\}}^{[s_2]} \otimes I_{n_{k_2}}), \\ \text{Gr}_{k_1, k_2; s_1, \cdot; \cdot, m_2}^{\text{CA}} &= (V_{\{k_1\}}^{[s_1]T} \otimes I_{n_{k_1}}) P_{1, k_2}^{k_1, 1} (V_{\{k_2\}}^{[m_2]} G_{(k_2)}^{[m_2]T} \otimes I_{n_{k_2}}), \\ \text{Gr}_{k_1, k_2; s_1, \cdot; \cdot, m_2}^{\text{CG}} &= (V_{\{k_1\}}^{[s_1]T} \otimes I_{n_{k_1}}) P_1^{k_1} (V^{[m_2]}), \\ \text{Gr}_{k_1, k_2; \cdot, s_2; m_1, \cdot}^{\text{AC}} &= (G_{(k_1)}^{[m_1]} V_{\{k_1\}}^{[m_1]T} \otimes I_{n_{k_1}}) P_{1, k_2}^{k_1, 1} (V_{\{k_2\}}^{[m_2]} G_{(k_2)}^{[m_2]T} \otimes I_{n_{k_2}}), \\ \text{Gr}_{k_1, \cdot; \cdot, \cdot; m_1, m_2}^{\text{AA}} &= (G_{(k_1)}^{[m_1]} V_{\{k_1\}}^{[m_1]T} \otimes I_{n_{k_1}}) P_1^{k_1} (V^{[m_2]}), \\ \text{Gr}_{\cdot, k_2; \cdot, \cdot; m_1, m_2}^{\text{GA}} &= (V^{[m_1]T} \otimes I_{n_{k_1}}) P_1^{k_1} (V^{[m_2]}), \\ \text{Gr}_{\cdot, \cdot; \cdot, \cdot; m_1, m_2}^{\text{GG}} &= (V^{[m_1]T}) (V^{[m_2]}), \end{aligned} \quad (39)$$

здесь P – коммутационные матрицы (“матрицы перестановки осей”), верхние и нижние индексы обозначают исходные и новые номера мод соответственно. Например, векторизация транспонирования матрицы A может быть записана в виде $P_{2,1}^{1,2} \text{vec}(A) = \text{vec}(A^T)$. В тензорном случае все прочие индексы при этом упорядочиваются согласно их естественной нумерации.

На практике слагаемое $Q(\mathbf{x})$ часто исключается по нескольким причинам: требует дополнительных вычислительных ресурсов, конфликтует с требованием оптимизационного метода на симметричность и положительную определенность приближения гессиана, а также по причине того, что в окрестности локального минимума его влиянием можно пренебречь. Вместе с тем

приводим выражения для вычисления данной квадратичной части гессиана, используя тензорное представление:

$$\text{Tens}\left(\frac{\partial^2 \mathbf{z}}{\partial \text{vec}(C_k^{[s_1]}) \partial \text{vec}(C_l^{[s_2]})} \cdot \mathbf{z}\right) = \begin{cases} \left\langle \left[\left[C_1^{[s_1]}, \dots, C_d^{[s_1]} \right] \right]_{C_k^{[s_1]}, C_l^{[s_1]} = I_{L_{s_1}}}, Z \right\rangle_{\{k,l\}}, & k \neq l, \quad s_1 = s_2, \\ 0, & \text{иначе,} \end{cases} \quad (40)$$

$$\text{Tens}\left(\frac{\partial^2 \mathbf{z}}{\partial \text{vec}(A_k^{[m_1]}) \partial \text{vec}(A_l^{[m_2]})} \cdot \mathbf{z}\right) = \begin{cases} \left\langle \left[\left[G^{[m_1]}, A_1^{[m_1]}, \dots, A_d^{[m_1]} \right] \right]_{A_k^{[m_1]} = I_{I_k^{[m_1]}}, A_l^{[m_1]} = I_{I_l^{[m_1]}}, Z \right\rangle_{\{k,l\}}, & k \neq l, \quad m_1 = m_2, \\ 0, & \text{иначе,} \end{cases} \quad (41)$$

$$\text{Tens}\left(\frac{\partial^2 \mathbf{z}}{\partial \text{vec}(A_k^{[m_1]}) \partial \text{vec}(G^{[m_2]})} \cdot \mathbf{z}\right) = \begin{cases} \left[\left[Z; A_1^{[m_1]}, \dots, A_d^{[m_1]} \right] \right]_{A_k^{[m_1]} = I_{I_k^{[m_1]}}, & m_1 = m_2, \\ 0, & m_1 \neq m_2, \end{cases} \quad (42)$$

$m_1, m_2 = \overline{1, M}$, $k, l = \overline{1, d}$, $\mathbf{z} = \text{vec}(Z(\mathbf{x}))$, $Z = Z(\mathbf{x})$, и все остальные вторые производные равны нулю. Важно отметить, что использование явного представления матрицы Гессе требует больших затрат по памяти. Для оптимизации расходования памяти имеет смысл использовать не полное представление, а использовать результат умножения матрицы на вектор, опираясь на структуру параметров, в том или ином итеративном методе, что и было сделано в используемой реализации.

3. Встраивание групповых условий в оптимизационную задачу

В данном приложении описано, как встроить сформулированные условия в оптимизационную задачу. Предложенная детерминистическая модель для группового анализа данных включает в себя условия на фактор-матрицу групповой оси, а также условие отделимости вида (9). Суммируем их для обоих вариантов модели, полной (Lr, 1), GLRO (6), и модели GTLD (7):

$$\begin{aligned} C_{d+1} &= [I_N \mathbf{p}], & A_{d+1} &= \text{diag}(\mathbf{p}), \\ p_1 + \dots + p_N &= p_{\text{cum}}, \quad p_i \geq p_{\min}, & p_1 + \dots + p_N &= p_{\text{cum}}, \quad p_i \geq p_{\min}, \\ (C_\gamma^{[N+1]})^T C_\gamma^{[i]} &= 0, \quad \gamma \in \Omega, \quad i = \overline{1, N}, & A_\gamma^T C_\gamma^{[i]} &= 0, \quad \gamma \in \Omega, \quad i = \overline{1, N}, \end{aligned} \quad (43)$$

где p_{\min} и p_{cum} – фиксированные константы, Ω – набор мод для условия отделимости, N – общее число образцов. Существуют разные способы учета этих условий в задаче (36), обратимся к двум методам: проекционному и методу множителей Лагранжа. В проекционном методе каждая итерация состоит из шага обновления решения безусловной задачи оптимизации с последующим его проектированием на область допустимости, заданной через условия. Первая часть проектора, отвечающая условию отделимости, является проектированием столбцов индивидуальных фактор-матриц на пространство, ортогональное столбцам общей фактор-матрицы: $P_\gamma^\perp(X) = (I - Y(Y^T Y)^{-1} Y^T) X$. Оставшаяся часть проектора является проекцией \mathbf{p} на внутренность l_1 шара с вырезом в окрестности нуля и положительными значениями координат (см. [54]), $\{\mathbf{y} \in \mathbb{R}^N \mid y_1 + \dots + y_N = p_{\text{cum}}, y_i \geq p_{\min} > 0\}$.

Второй метод состоит в использовании множителей Лагранжа, сводящих условную задачу оптимизации к безусловной. Запишем соответствующее слагаемое целевого функционала $G(\mathbf{x})$:

$$\begin{aligned} G(\mathbf{x}) &= \sum_{\gamma \in \Omega} \sum_{i=1}^N \langle \text{vec}(U_\gamma^T C_\gamma^{[i]}), \hat{\boldsymbol{\mu}}_{\gamma,i} \rangle - \kappa \cdot (p_{\text{cum}} - \langle \mathbf{p}, \mathbb{1} \rangle) - \langle \zeta, \mathbf{p} - p_{\min} \cdot \mathbb{1} \rangle, \\ U_\gamma &= \begin{cases} C_\gamma^{[N+1]}, & \text{для модели (6),} \\ A_\gamma, & \text{для модели (7),} \end{cases} \quad \begin{cases} C_{d+1} = [I_N \mathbf{p}], & \text{для модели (6)} \\ C_{d+1} = I_N, A_{d+1} = \text{diag}(\mathbf{p}), & \text{для модели (7).} \end{cases} \end{aligned} \quad (44)$$

Число дополнительных параметров можно уменьшить, ослабив условия на ортогональность групповых и индивидуальных фактор-матриц:

$$G(\mathbf{x}) = \sum_{\gamma \in \Omega} \frac{\mu_\gamma}{2} \sum_{i=1}^N \|U_\gamma^T C_\gamma^{[i]}\|_F^2 - \kappa \cdot (p_{\text{cum}} - \langle \mathbf{p}, \mathbb{1} \rangle) - \langle \zeta, \mathbf{p} - p_{\min} \cdot \mathbb{1} \rangle, \quad (45)$$

с теми же U_{γ} , что и в (44). Обозначим вектор производных по вспомогательным переменным для $G(\mathbf{x})$ как $g(\mathbf{x})$. Градиент для $g(\mathbf{x})$ в случае модели (7) имеет следующий вид:

$$\begin{aligned} \nabla g(\mathbf{x}) &= [\nabla g_{\mu_{\gamma_1}}(\mathbf{x}) \dots \nabla g_{\mu_{\gamma_{|\Omega|}}}(\mathbf{x}) \nabla g_{\kappa}(\mathbf{x}) \nabla g_{\zeta_1}(\mathbf{x}) \dots \nabla g_{\zeta_N}(\mathbf{x})], \\ \nabla g_{\mu_{\gamma_j}}(\mathbf{x})^T &= [\dots 0 \widehat{\text{vec}}(A_{\gamma_j} A_{\gamma_j}^T C_{\gamma_j}) 0 \dots 0 \widehat{\text{vec}}(C_{\gamma_j} C_{\gamma_j}^T A_{\gamma_j}) 0 \dots], \\ \nabla g_{\kappa}(\mathbf{x})^T &= [\dots 0 \mathbb{1}_{1 \times N} 0 \dots], \quad \nabla g_{\zeta_i}(\mathbf{x})^T = \begin{bmatrix} \dots & 0 & \underbrace{-1}_{i\text{-я позиция}} & 0 & \dots \end{bmatrix}. \end{aligned} \quad (46)$$

Наконец, каждая итерация требует решения системы с расширенной матрицей Гессе, $\tau = [\mu_{\gamma_1}, \dots, \mu_{\gamma_{|\Omega|}}, \kappa, \zeta_1, \dots, \zeta_N]^T$:

$$\begin{bmatrix} H[f](\mathbf{x}) & \nabla g(\mathbf{x}) \\ \nabla g(\mathbf{x})^T & 0 \end{bmatrix} \begin{bmatrix} \Delta \mathbf{x} \\ \Delta \tau \end{bmatrix} = \begin{bmatrix} -\nabla f(\mathbf{x}) \\ -g(\mathbf{x}) \end{bmatrix}. \quad (47)$$

СПИСОК ЛИТЕРАТУРЫ

1. Kolda T.G., Bader B.W. Tensor decompositions and applications // SIAM Rev. 2009. V. 51 Iss. 3. P. 455–500.
2. Cichocki A., Lee N., Oseledets I., Phan A.-H., Zhao Q., Mandic D.P. Tensor networks for dimensionality reduction and large-scale optimization: Part 1, low-rank tensor decompositions // Found. Trends Mach. Learn. 2016. V. 9. Iss. 4–5. P. 249–429.
3. Sidiropoulos N.D., De Lathauwer L., Fu X., Huang K., Papalexakis E.E., Faloutsos C. Tensor decomposition for signal processing and machine learning // IEEE Trans. Signal Proc. 2017. V. 65. Iss. 13. P. 3551–3582.
4. Phan A.-H., Cichocki A. Tensor decompositions for feature extraction and classification of high dimensional datasets // IEICE Nonlin. Theory and its App. 2010. V. 1. Iss. 1. P. 37–68.
5. De Lathauwer L. Decompositions of a higher-order tensor in block terms – part I: Lemmas for partitioned matrices // SIAM J. Matrix Anal. Appl. 2008. V. 30. Iss. 3. P. 1022–1032.
6. De Lathauwer L. Decompositions of a higher-order tensor in block terms – part II: Definitions and uniqueness // SIAM J. Matrix Anal. Appl. 2008. V. 30. Iss. 3. P. 1033–1066.
7. De Lathauwer L., Nion D. Decompositions of a higher-order tensor in block terms – part III: Alternating least squares algorithms // SIAM J. Matrix Anal. Appl. 2008. V. 30. Iss. 3. P. 1067–1083.
8. Prasad G., Jahanshad N., Aganj I., Lenglet C., Sapiro G., Toga A.W., Thompson P.M. Atlas-based fiber clustering for multi-subject analysis of high angular resolution diffusion imaging tractography // Proc. IEEE Int. Symp. Biomed. Imaging. 2011. P. 276–280.
9. Calhoun V.D., Adali T. Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery // IEEE Rev. Biomed. Eng. 2012. V. 5. P. 60–73.
10. Beckmann C.F., Smith S.M. Tensorial extensions of independent component analysis for multisubject fMRI analysis. // Neuroimage. 2005. V. 25. Iss. 1. P. 294–311.
11. Nazarenko D.V., Kharyuk P.V., Oseledets I.V., Rodin I.A., Shpigun O.A. Machine learning for LC–MS medicinal plants identification // Chemom. Intell. Lab. Syst. 2016. V. 156. P. 174–180.
12. Xia P., Bai Z., Liang T., Yang D., Liang Z., Yan X., Liu Y. High-performance liquid chromatography based chemical fingerprint analysis and chemometric approaches for the identification and distinction of three endangered panax plants in Southeast Asia // J. Sep. Sci. 2016. V. 39. Iss. 20. P. 3880–3888.
13. Smilde A., Bro R., Geladi P. Multi-way analysis: applications in the chemical sciences // UK: John Wiley & Sons, 2005.
14. Cichocki A., Mandic D., De Lathauwer L., Zhou G., Zhao Q., Caiafa C., Phan A.H. Tensor decompositions for signal processing applications: From two-way to multiway component analysis // IEEE Signal Process. Mag. 2015. V. 32. Iss. 2. P. 145–163.
15. Zhou G., Zhao Q., Zhang Y., Adali T., Xie Sh., Cichocki A. Linked component analysis from matrices to high-order tensors: Applications to biomedical data // Proc. IEEE. 2016. V. 104. Iss. 2. P. 310–331.
16. Харюк П.В. Групповой анализ данных на основе блочного канонического разложения // Сб. тезисов 59-й научной конференции МФТИ. М.: МФТИ, 2016.
17. Харюк П.В. Классификация сигналов с помощью блочного тензорного разложения в задаче группового анализа данных // Сб. тезисов XXIV Международной научной конференции Ломоносов-2017. М.: ООО “МАКС Пресс”, 2017. С. 152–153.
18. Calhoun V.D., Liu J., Adali T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. Neuroimage, 2009. V. 45. Iss. 1. P. S163–S172.
19. Zhou G., Cichocki A., Zhang Y., Mandic D.P. Group component analysis for multiblock data: Common and individual feature extraction. IEEE Trans. Neural Netw. Learn. Syst. 2016. V. 27. Iss. 11. P. 2426–2439.
20. Lock E.F., Hoadley K.A., Marron J.S., Nobel A.B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types // Ann. Appl. Stat. 2013. V. 7. Iss. 1. P. 523.
21. Sorber L., Van Barel M., De Lathauwer L. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(Lr, Lr, 1) terms, and a new generalization // SIAM J. Optim. 2013. V. 23. Iss. 2. P. 695–720.

22. *Tipping M.E., Bishop C.M.* Mixtures of probabilistic principal component analyzers // *Neural Comput.* 1999. V. 11. Iss. 2. P. 443–482.
23. *Tipping M.E., Bishop C.M.* Probabilistic principal component analysis // *J. R. Stat. Soc. Series B Stat. Methodol.* 1999. V. 61. Iss. 3. P. 611–622.
24. *Bishop C.M.* Pattern recognition and machine learning // NY: Springer, 2006.
25. *Oseledets I.V.* Tensor-train decomposition // *SIAM J. Sci. Comput.* 2011. V. 33. Iss. 5. P. 2295–2317.
26. *Kingma D.P., Welling M.* Auto-encoding variational bayes // arXiv:1312.6114, 2013.
27. *Kharyuk P., Nazarenko D., Oseledets I., Rodin I., Shpigun O., Tsitsilin A., Lavrentyev M.* Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task // *Sci. Rep.* 2018. V. 8. Iss. 1. N. 17053.
28. *Björck A., Golub G.H.* Numerical methods for computing angles between linear subspaces // *Math. of Computat.* 1973. V. 27. Iss. 123. P. 579–594.
29. *Vervliet N., Debals O., Sorber L., Van Barel M., De Lathauwer L.* Tensorlab user guide // 2016.
30. *Blei D., Ranganath R., Mohamed S.* Variational inference: Foundations and modern methods // NIPS Tutorial, 2016.
31. *Wingate D., Weber T.* Automated variational inference in probabilistic programming arXiv:1301.1299, 2013.
32. *Bingham E., Chen J.P., Jankowiak M., Obermeyer F., Pradhan N., Karaletsos T., Singh R., Szerlip P., Horsfall P., Goodman N.D.* Pyro: Deep universal probabilistic programming // *J. Mach. Learn. Res.* 2019. V. 20. Iss. 1. P. 973–978.
33. *Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L., Lerer A.* Automatic differentiation in PyTorch // NIPS Autodiff Workshop, 2017.
34. *Calhoun V.D., Adali T., Pearlson G.D., Pekar J.J.* A method for making group inferences from functional MRI data using independent component analysis // *Hum. Brain Mapp.* 2001. V. 14. Iss. 3. P. 140–151.
35. *Guo Y., Pagnoni G.* A unified framework for group independent component analysis for multi-subject fMRI data // *NeuroImage.* 2008. V. 42. Iss. 3. P. 1078–1093.
36. *Lee H., Choi S.* Group nonnegative matrix factorization for EEG classification // *Artif. Int. Stat.* 2009. P. 320–327.
37. *Yokota T., Cichocki A., Yamashita Y.* Linked PARAFAC/CP tensor decomposition and its fast implementation for multi-block tensor analysis // *Springer Int. Conf. Neural Inform. Proc.* 2012. P. 84–91.
38. *Cichocki A., Zdunek R., Amari S.-I.* Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization // *Springer Int. Conf. Indep. Comp. Anal. Sig. Sep.* 2007. P. 169–176.
39. *Gillis N., Glineur F.* Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization // *Neural Comput.* 2012. V. 24. Iss. 4. P. 1085–1105.
40. *Yokota T., Cichocki A.* Linked Tucker2 decomposition for flexible multi-block data analysis // *Springer Int. Conf. Neural Inform. Proc.* 2014. P. 111–118.
41. *Gong X.F., Lin Q.-H., Debals O., Vervliet N., De Lathauwer L.* Coupled rank-(Lm, Ln, 1) block term decomposition by coupled block simultaneous generalized Schur decomposition // *IEEE ICASSP.* 2016. P. 2554–2558.
42. *Khoromskij B., Khoromskaia V.* Low rank tucker-type tensor approximation to classical potentials // *Open Math.* 2007. V. 5. Iss. 3. P. 523–550.
43. *Continuum Analytics.* Anaconda software distribution // [Электронный ресурс] vers. 2-2.4.0, 11.2015. Дата обращения: 1.11.2019.
44. *Oliphant T.E.* A guide to NumPy // USA: Trelgol Publ. USA. V. 1. 2006.
45. *Jones E., Oliphant T., Peterson P.* SciPy: Open source scientific tools for Python // [Электронный ресурс] 2001. Дата обращения: 1.11.2019.
46. *McKinney W.* Data structures for statistical computing in python // *Proceed. of the 9th Python in Sci. Conf.* V. 2010. 445, P. 51–56.
47. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.* Scikit-learn: machine learning in Python // *J. of Mach. Learn. Res.* 2011. V. 12. P. 2825–2830.
48. *Hunter J.D.* Matplotlib: A 2D graphics environment // *Comput. Sci. Eng.* 2007. V. 9. Iss. 3. P. 90–95.
49. *Waskom M., Botvinnik O., O’Kane D., Hobson P., Lukauskas S., Gempferline D.C., Augspurger T., Halchenko Y., Cole J.B., Warmenhoven J., de Ruiter J., Pye C., Hoyer S., Vanderplas J., Villalba S., Kunter G., Quintero E., Bachant P., Martin M., Meyer K., Miles A., Ram Y., Yarkoni T., Williams M.L., Evans C., Fitzgerald C., Brian, Fonnesbeck C., Lee A., Qalieh A.* Seaborn: statistical data visualization // [Электронный ресурс], v.0.8.1, 09.2017. Дата обращения: 1.11.2019.
50. *Kluyver T., Ragan-Kelley B., Pérez F., Granger B., Bussonnier M., Frederic J., Kelley K., Hamrick J., Grout J., Corlay S., Ivanov P., Avila D., Abdalla S., Willing C.* Jupyter notebooks – a publishing format for reproducible computational workflows // *ELPUB 2016.* IOS Press, P. 87–90.
51. *Leibe B., Schiele B.* Analyzing appearance and contour based methods for object categorization // *IEEE CVPR.* 2003. V. 2.
52. *Hastie T., Tibshirani R., Friedman J., Franklin J.* The elements of statistical learning: data mining, inference and prediction // *Springer Series in Statistics*, 2009.
53. *Khoromskij B.N., Oseledets I.V.* QTT approximation of elliptic solution operators in higher dimensions // *Russ. J. Numer. Anal. Math. Model.* 2011. V. 26. Iss. 3. P. 303–322.
54. *Gupta M.D., Kumar S., Xiao J.* L1 projections with box constraints. // arXiv:1010.0141, 2010.