

Application of Tensors in Multivariate Component Analysis of Categorized Features

A. A. Banin¹DOI: [10.18255/1818-1015-2025-4-330-359](https://doi.org/10.18255/1818-1015-2025-4-330-359)¹Cherepovets State University, Cherepovets, Russia

MSC2020: 62H25

Research article

Full text in Russian

Received August 1, 2025

Revised October 15, 2025

Accepted October 29, 2025

When modeling social processes and phenomena, it is often necessary to process data related to categorized features, identify cause-and-effect relationships between such data, and determine the most significant indicators. A study of existing approaches to analyzing dependencies between categorized variables revealed several problems when applying these methods to multidimensional categorized data (tensors). Therefore, this article proposes an approach to studying dependencies between such variables using multidimensional component analysis. This approach involves applying tensor unfolding matrices obtained for each of its axes (categorized features). The method allows for the construction of integral characteristics (components) based on the elements of the original tensor, the formation of component loading matrices, and the calculation of the tensor core, which has fewer gradations of categorized features (lower number of dimensions in the tensor axes) than the original tensor. The article proposes a method for ranking the gradations of categorized variables by the degree of cumulative influence of component loadings, based on the calculation of vector norms. The described approach to studying dependencies between multidimensional categorized variables is demonstrated using a three-dimensional tensor with the shape (4;10;10) and categorized features: nosology group, field of activity, and group of professionally significant qualities. The algorithm for analyzing multidimensional categorized data using multidimensional component analysis, discussed in this article, is intended to be incorporated as an analytical tool into the regional information and analytical portal "PERSPEKTIVA-PRO". This tool can be used to develop a digital support trajectory for people with disabilities and special needs, taking into account their personal and variable characteristics.

Keywords: component analysis; tensor; mode-n matricization; commutation matrix; vectorization; Kronecker product; singular value decomposition

INFORMATION ABOUT THE AUTHORS

Banin, Alexander A. (corresponding author)	ORCID iD: 0000-0003-4242-1471 . E-mail: aabanin1@chsu.ru PhD, Associate Professor
---	--

Funding: The work was carried out within the framework of scientific research included in the plans of scientific work of scientific organizations and educational institutions of higher education carrying out scientific research at the expense of the federal budget on the topic "Research and modeling of variable trajectories of psychological, pedagogical and clinical-psychological support for the professional development of persons with disabilities in the digital environment" (Ministry of Science and Higher Education of the Russian Federation, FEGN-2023-0005).

For citation: A. A. Banin, "Application of tensors in multivariate component analysis of categorized features", *Modeling and Analysis of Information Systems*, vol. 32, no. 4, pp. 330–359, 2025. DOI: [10.18255/1818-1015-2025-4-330-359](https://doi.org/10.18255/1818-1015-2025-4-330-359).

Применение тензоров в многомерном компонентном анализе категоризованных признаков

А. А. Банин¹DOI: [10.18255/1818-1015-2025-4-330-359](https://doi.org/10.18255/1818-1015-2025-4-330-359)¹Череповецкий государственный университет, Череповец, Россия

УДК 519.254

Научная статья

Полный текст на русском языке

Получена 1 августа 2025 г.

После доработки 15 октября 2025 г.

Принята к публикации 29 октября 2025 г.

При моделировании социальных процессов и явлений зачастую приходится обрабатывать данные, относящиеся к категоризованным признакам, выявлять причинно-следственные связи между такими данными, выделять наиболее существенные показатели. Исследование существующих подходов к анализу зависимостей между категоризованными переменными выявило ряд проблем при применении этих методов для многомерных категоризованных данных (тензоров). Поэтому в статье предлагается подход для изучения зависимостей между такими переменными с использованием многомерного компонентного анализа. Данный подход предполагает применение матриц развертки тензора, полученных для каждой его оси (категоризованного признака). Метод позволяет построить интегральные характеристики (компоненты) по элементам исходного тензора, сформировать матрицы компонентных нагрузок и рассчитать ядро тензора, имеющего меньшее число градаций категоризованных признаков (меньшее число измерений на осях тензора), чем исходный тензор. В статье предложен метод ранжирования градаций категоризованных переменных по степени совокупного влияния на них компонентных нагрузок, основанный на вычислении векторных норм. Изложенный подход к изучению зависимостей между многомерными категоризованными переменными продемонстрирован на примере трехмерного тензора с формой (4;10;10) и категоризованными признаками: группа нозологии, сфера деятельности, группа профессионально значимых качеств. Рассмотренный в статье алгоритм изучения многомерных категоризованных данных с применением многомерного компонентного анализа предполагается включить как аналитический инструмент информационно-аналитического регионального портала «ПЕРСПЕКТИВА-PRO», который может быть использован для разработки траекторий цифрового сопровождения лиц с инвалидностью и лиц с ОВЗ с учетом их личностных и вариативных характеристик.

Ключевые слова: компонентный анализ; тензор; матрица развертки; коммутационная матрица; векторизация; произведение Кронекера; сингулярное разложение

ИНФОРМАЦИЯ ОБ АВТОРАХ

Банин, Александр Анатольевич
(автор для корреспонденции)

ORCID iD: [0000-0003-4242-1471](https://orcid.org/0000-0003-4242-1471). E-mail: aabanin1@chsu.ru
Канд. физ.-мат. наук, доцент

Финансирование: Работа выполнена в рамках научных исследований, включаемых в планы научных работ научных организаций и образовательных организаций высшего образования, осуществляющих научные исследования за счёт средств федерального бюджета по теме «Исследование и моделирование вариативных траекторий психолого-педагогического и клиничко-психологического сопровождения профессионального становления лиц с инвалидностью в условиях цифровой среды» (Министерство науки и высшего образования Российской Федерации, FEGN-2023–0005).

Для цитирования: А. А. Банин, “Application of tensors in multivariate component analysis of categorized features”, *Modeling and Analysis of Information Systems*, vol. 32, no. 4, pp. 330–359, 2025. DOI: [10.18255/1818-1015-2025-4-330-359](https://doi.org/10.18255/1818-1015-2025-4-330-359).

Введение

При моделировании социальных процессов и явлений зачастую приходится обрабатывать данные, относящиеся к категоризованным признакам, выявлять причинно-следственные связи между такими данными, выделять наиболее существенные показатели, строить и оценивать интегральные характеристики таких процессов и явлений. В рамках решения задачи по созданию научно-практического инструментария профориентации, образования и содействия трудоустройству инвалидов и лиц с ОВЗ для регионального информационно-аналитического портала «ПЕРСПЕКТИВА-PRO» [1] в настоящей работе рассматривается математическая модель комплексного оценивания развития профессионально значимых качеств (ПЗК) инвалидов и лиц с ОВЗ (респондентов) с учетом их вариативных характеристик. К таким характеристикам относятся, например, регион проживания, социальный статус, возраст, состояние здоровья, уровень образования. Большинство таких вариативных характеристик являются категоризованными переменными, т. е. переменными, возможные значения которых задаются конечным числом состояний (категорий, градаций). В основе математической модели лежит компонентный анализ для многомерных массивов (тензоров), осями (измерениями) которых являются категоризованные переменные. Компоненты в модели строятся математически и представляют собой новые, искусственно созданные градации осей тензора, которые линейно комбинируют исходные категории измерений тензора, независимы и сохраняют ключевую информацию. При этом количество новых градаций осей тензора меньше их первоначального числа. Данная работа является продолжением исследования [2], в котором компонентный анализ применялся для матрицы (двумерного тензора), которая строилась отдельно для каждого респондента. Осями такого тензора были сфера деятельности (выделено 10 сфер деятельности) и группы ПЗК (выделено 10 групп), а элементами выступали оценки респондентом своих ПЗК (целочисленные значения от 1 до 5), рассчитанные в соответствии с методикой эталонных значений [1]. Данная модель позволяет выделить сферы профессиональной деятельности, наиболее успешные для профессиональной самореализации инвалидов и лиц с ОВЗ с учетом их ПЗК. В настоящей же статье рассматривается математическая модель ранжирования градаций категоризованных признаков на основе построенных компонент для многомерных категоризованных переменных. Модель может быть использована как для одного респондента так, и для группы респондентов. В частности, модель, представленная в работе [2], является частным случаем данной математической модели, реализованной для одного респондента с двумя категоризованными признаками. В статье предлагаемый метод исследования зависимостей между многомерными категоризованными признаками продемонстрирован на примере трехмерного тензора со следующими переменными (осями): группа нозологии, сфера деятельности, группа ПЗК.

1. Обзор связанных работ

Использование многомерного статистического инструментария для обработки и анализа категоризованных данных, описывающих социальные аспекты личной и общественной жизни, является довольно распространенным подходом. Рассмотрим примеры применения математико-статистических методов, составляющие основу данного инструментария, в прикладных задачах с категоризованными переменными социальной направленности.

Большинство инструментов обработки категоризованных данных можно разбить на две группы. К первой группе относятся методы построения различных регрессионных уравнений, в которых категоризованные переменные являются объясняющими (независимыми) переменными или объясняемыми (зависимыми) переменными. В первом случае включение таких переменных в уравнение регрессии происходит путем добавления фиктивных переменных, причем, если категоризованный признак имеет k градаций, то количество фиктивных переменных для этого признака должно

быть $k - 1$. В случае зависимых категоризованных переменных используются модели дискретного выбора, в частности логит-модель или пробит-модель.

Так в работе [3] при изучении влияния статуса индивида на его занятость и продолжительность рабочей недели с помощью пробит-модели рассчитывается индекс склонности — вероятность того, что индивид имеет статус инвалида (чтобы разделить понятия плохого здоровья и статуса инвалида, подтвержденного документально). В качестве ковариант (объясняющих переменных) наряду с количественными переменными в модель включены и категоризованные переменные: самооценка здоровья, пол, уровень образования, место проживания. А для оценки качества модели формируется контрольная группа из индивидов, не имеющих статус инвалида, с наиболее близкими значениями индекса склонности к значениям этого индекса у индивидов, имеющих статус инвалида. Также в данном исследовании построена пробит-регрессия для оценки влияния статуса инвалида на вероятность его занятости. При этом в спецификацию были включены все переменные, которые использовались в пробит-модели индекса склонности.

В статье [4] исследуется социальное самочувствие населения в зависимости от того, что интересует людей в данный момент и какие происходят изменения в их предпочтениях и интересах. Большинство переменных изначально являются категоризованными признаками. Поэтому в этой работе предлагается их заменить на соответствующие количественные переменные, представляющие собой индексы социального самочувствия (для зависимых переменных) и индексы поисковой активности — индексы поисковых запросов (для объясняющих переменных). Для сокращения числа объясняющих переменных в исследовании выполняется факторный анализ, главной целью которого является объединение поисковых запросов в категории, характеризующие тот или иной аспект социального самочувствия, на основе факторных нагрузок. При этом факторный анализ используется как метод обнаружения взаимосвязей между значениями переменных посредством исследования структуры ковариационных и корреляционных матриц. Для выделения категорий запросов, которые в большей степени коррелируют с индексами социального самочувствия, в статье применяется модель байесовского усреднения, направленная на отбор наиболее надежных (робастных) объясняющих переменных. На заключительном этапе моделирования результаты байесовской модели уточняются с помощью регрессионного анализа. В качестве зависимой переменной этой регрессии выступает какой-либо индекс социального самочувствия, а независимых переменных — отобранные с помощью байесовской модели усреднения категории поисковых запросов. Таким образом, чтобы выявить зависимость между категоризованными признаками, в данной работе предлагается заменить их количественными переменными (индексами), а затем, используя факторный анализ и модель байесовского усреднения, сформировать категории поисковых запросов и оценить регрессионную зависимость между индексами социального самочувствия и категориями поисковых запросов.

В исследовании [5] оценивается эффект воздействия одной категоризованной переменной (высшего образования) на другую категоризованную переменную (здоровье индивида) с помощью параметрического и непараметрического подходов, представленных соответственно многомерной рекурсивной пробит-моделью и мэтчингом. В статье рассматриваются категоризованные признаки с двумя состояниями (бинарные переменные), которые разбиты на три группы: переменные исхода (самооценка собственного здоровья в двух вариантах, наличие хронических заболеваний), переменные воздействия (наличие у индивида высшего образования), контрольные переменные (переменные, отражающие ответы респондентов на вопросы об их индивидуальных характеристиках, социально-экономическом статусе и поведении в отношении собственного здоровья). Параметрический подход, представленный в данной работе, включает оценивание рекурсивной системы бинарных уравнений, в которую входят отдельные уравнения для переменной исхода и переменной воздействия. Кроме того, в модели учитывается возможная эндогенность контрольных переменных,

входящих в уравнение исхода как объясняющие переменные, за счет добавления соответствующих уравнений в систему. Многомерные пробит-модели оцениваются с помощью метода максимального правдоподобия в предположении, что случайные ошибки, входящие в уравнения системы, имеют совместное нормальное распределение с нулевым вектором математических ожиданий и некоторой ковариационной матрицей. Непараметрический подход к оценке причинно-следственных связей между категоризованными признаками представлен в данной работе методом мэтчинга, который не требует предположений о функциональной форме модели и распределении случайных ошибок модели исхода. Суть метода заключается в разделении индивидов на две группы (группа воздействия, группа контроля) в зависимости от того какое значение принимает категоризованная переменная воздействия, и подбора для каждого объекта из группы воздействия наиболее похожие на него по контрольным переменным объекты из группы контроля. Затем оценка ненаблюдаемого потенциального исхода для данного объекта из группы воздействия строится на основе наблюдаемых исходов объектов из группы контроля, соответствующих этому объекту. Аналогичная процедура проводится для каждого объекта из контрольной группы. Подбор объектов, наиболее похожих на данный объект из противоположной ему группы, осуществляется с помощью вычисления следующих метрик: расстояния Махаланобиса, использующего ковариационную матрицу для контрольных переменных, и оцениваемой с помощью логистической регрессии меры склонности к воздействию — условной вероятности того, что объект подвергается воздействию. После процедуры мэтчинга строится логистическая регрессия для каждой переменной исхода на переменную воздействия и контрольные переменные. Полученные оценки параметров уравнения использовались для расчета эффекта воздействия — вероятностей того, что конкретный индивид попадает в соответствующую категорию переменной исхода. Для применения данного непараметрического метода необходимы наблюдаемые значения переменных исхода для каждого респондента, а также выполнение ряда допущений: 1) отбор объектов в контрольную группу и группу воздействия происходит исключительно по наблюдаемым характеристикам и не зависит от потенциальных исходов; 2) любой объект выборки с положительной вероятностью может попасть как в группу воздействия, так и в группу контроля; 3) потенциальные исходы одного объекта не зависят от назначения воздействия другим объектам.

Еще один подход применения регрессионного анализа для оценки зависимости между категоризованными признаками представлен в статье [6]. В данной работе исследуется связь удовлетворенности работой с удовлетворенностью ее отдельными аспектами (заработной платой, условиями труда, возможностью профессионального роста). Для этого рассматриваются четыре категоризованные переменные, значения которых формируются по опросам индивидов (по каждой переменной дается ответ респондентом по пятибалльной шкале от 1 (полностью удовлетворен) до 5 (совсем не удовлетворен)). Построение регрессионной зависимости удовлетворенности работой в целом от ее компонент опирается на предположении, что за каждым ответом стоит ненаблюдаемая непрерывная случайная величина, которая сводится к дискретной шкале интервальной группировкой. Так, например, уровень удовлетворенности работой в целом J_i для i -го индивида связан с «истинным» (ненаблюдаемым) уровнем удовлетворенности J_i^* следующим образом:

$$J_i = \begin{cases} 1, & J_i^* < \bar{J}_1, \\ 2, & \bar{J}_1 \leq J_i^* < \bar{J}_2, \\ 3, & \bar{J}_2 \leq J_i^* < \bar{J}_3, \\ 4, & \bar{J}_3 \leq J_i^* < \bar{J}_4, \\ 5, & \bar{J}_4 \leq J_i^*, \end{cases}$$

где $\bar{J}_1, \bar{J}_2, \bar{J}_3, \bar{J}_4$ — границы интервалов группировки, неизвестные заранее и подлежащие оцениванию. Аналогично выполняется группировка и для остальных категоризованных признаков. В предположении, что «истинные» величины категоризованных переменных имеют совместное нормальное распределение с нулевыми математическими ожиданиями, единичными дисперсиями и неизвестной корреляционной матрицей, методом максимального правдоподобия оцениваются границы интервалов группирования ненаблюдаемых уровней удовлетворенности и элементы корреляционной матрицы. Используя оценки элементов корреляционной матрицы, рассчитываются коэффициенты уравнения линейной регрессии — зависимости «истинной» удовлетворенности работой в целом от «истинных» удовлетворенностей ее отдельных аспектов. Данный метод был выбран в виду характера переменных модели: все они являются упорядоченными дискретными переменными с ограниченным набором значений.

Ко второй группе инструментов оценивания зависимостей между категоризованными признаками относятся методы, использующие таблицы сопряженности признаков [7], в которых результаты статистического обследования n объектов по категоризованным признакам $x^{(1)}, \dots, x^{(m)}$ ($m \geq 2$) представлены в виде частот $n_{i_1 \dots i_m}$ — числа объектов (из общего числа n обследованных), у которых значения признаков $x^{(1)}, \dots, x^{(m)}$ зафиксированы на уровне $i_1, \dots, i_m$ градаций соответственно.

В статье [8] проводится сравнение регрессии с категоризованными предикторами (признаками) — регрессии, в которой для таких признаков используются фиктивные переменные, и логлинейного анализа — метода изучения таблиц сопряженности признаков, в котором связь между двумя и более категоризованными переменными исследуется при помощи натурального логарифма теоретических частот. Использование логлинейного анализа для оценки связи между категоризованными признаками по мнению авторов статьи подходит по следующим причинам: 1) такой анализ логически допускает оперирование понятиями зависимой и независимой переменных; 2) данный метод работает как с количественными, так и с категоризованными шкалами, не требуя преобразовывать последние в фиктивные переменные; 3) позволяет не только находить многомерные связи и взаимодействия, но и определять детерминанты явлений.

Подход к изучению категоризованных переменных с применением логлинейного анализа представлен в работе [9], в которой оцениваются зависимости между факторами развития риска инфаркта миокарда, его осложнениями и локализацией очага поражения. В исследовании характер данных таков, что адекватный анализ имеющихся наблюдений может быть выполнен на основе таблиц сопряженности признаков с применением логлинейных моделей. Логлинейный анализ в данной статье осуществляется для двухвходовых таблиц сопряженности признаков, при этом теоретические частоты рассчитываются через наблюдаемые в зависимости от числа степеней свободы (насыщенная модель и ненасыщенные модели: модель независимости, модель с отсутствием влияния одного фактора, равновероятная модель). Гипотеза независимости признаков (эффект взаимодействия признаков равен нулю) проверяется с использованием критерия χ^2 , число степеней свободы которого равно числу клеток в таблице сопряженности за вычетом числа независимых параметров (общее число параметров минус число ограничений, накладываемых на параметры).

В статье [10] рассматриваются методы представления и совместной обработки многомерных данных объектов числовой и нечисловой природы в пространстве действительных чисел. Анализ таких данных выполняется в два этапа. На первом этапе все данные в матрице типа «объект — признак», измеряемые в количественных шкалах, преобразуются в номинальную шкалу и представляются в виде таблицы сопряженности признаков. Также на этом этапе данные таблицы используются для выявления статистически значимой связи между двумя признаками и построения логлинейных моделей. В случае многомерных таблиц сопряженности признаков из всего разнообразия моделей предлагается выбирать несколько относительно простых моделей. При этом выбор следует начинать с построения насыщенной модели (включает все эффекты), а при построении

ненасыщенной модели нужно учитывать только те эффекты, которые существенно отличаются от нуля. Также нужно учитывать, что при росте числа категоризованных признаков и их градаций растет и число ячеек в многомерной таблице сопряженности признаков, и как следствие в практических исследованиях число ячеек может превысить число наблюдений (во многих ячейках оказываются нулевые элементы). В таких случаях либо логлинейный анализ неприменим, либо таким ячейкам искусственно присваивается значение 0.5. На втором этапе используется метод анализа соответствий, когда таблица сопряженности признаков преобразуется в числовую матрицу, к которой применимы операции над действительными числами. Это позволяет решать задачи, связанные с представлением многомерных объектов в виде точек в координатном пространстве малой размерности (факторный анализ, метод главных компонент, компонентный анализ). В данном исследовании выполняется одномерный компонентный анализ, основанный на теореме Эккарта–Юнга, отдельно для строк и для столбцов преобразованной двумерной таблицы сопряженности признаков. Выбор размерности пространства осуществляется по доле вклада собственных значений в общую инерцию.

В работе [11] рассматривается применение логлинейного анализа для многомерных таблиц сопряженности признаков, в которых число категоризованных переменных больше двух. Обычно, в виде таких таблиц могут быть представлены результаты обработки социологических анкет. Основным условием применения данного метода является независимость и категориальная шкала переменных. Логлинейная модель, представленная в исследовании, — это множественная регрессия, в которой зависимая переменная — натуральный логарифм соответствующей теоретической частоты таблицы сопряженности признаков, а независимые переменные — категоризованные признаки и их совместные воздействия. Чтобы число параметров не превышало числа ячеек таблицы сопряженности признаков, в модели накладываются ограничения на эти параметры в виде некоторых балансовых уравнений. Параметры логлинейной модели оцениваются методом максимального правдоподобия, при этом предполагается, что наблюдаемые частоты имеют пуассоновское или полиномиальное распределение. Функция правдоподобия в этой модели показывает насколько «правдоподобно» получить наблюдаемые частоты при заданных параметрах, которые порождают ожидаемые частоты. В статье рассматривается логлинейная модель для трехмерной таблицы сопряженности признаков. Так если A, B, C есть категоризованные переменные, где A принимает I значений, B — J значений, C — K значений, то модель имеет вид:

$$\ln(n_{ijk}^*) = u_0 + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}, \quad (*)$$

где

- n_{ijk}^* — теоретические частоты, $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$;
- u_0 — общее среднее значение;
- u_i^A — эффект i -й категории переменной $A, i = 1, \dots, I$;
- u_j^B — эффект j -й категории переменной $B, j = 1, \dots, J$;
- u_k^C — эффект k -й категории переменной $C, k = 1, \dots, K$;
- u_{ij}^{AB} — эффект совместного воздействия i -й и j -й категорий двух переменных A и B соответственно;
- аналогично $u_{ik}^{AC}, u_{jk}^{BC}, u_{ijk}^{ABC}$ — эффекты совместного воздействия.

При этом должны выполняться следующие ограничения (балансовые уравнения):

$$\begin{aligned} \sum_i u_i^A &= \sum_j u_j^B = \sum_k u_k^C = 0, \\ \sum_i u_{ij}^{AB} &= \sum_j u_{ij}^{AB} = \sum_i u_{ik}^{AC} = \sum_k u_{ik}^{AC} = \sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = 0, \end{aligned}$$

$$\sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = 0.$$

При этом модель (*) называется насыщенной, если она включает все факторы и их возможные взаимодействия. Если же модель (*) содержит только часть статистически значимых факторов и их взаимодействий, то она называется ненасыщенной. Для определения значимости эффекта его исключают из модели и рассчитывают следующие статистики:

- 1) $\chi^2 = \sum_i \sum_j \sum_k \frac{n_{ijk} - n_{ijk}^*}{n_{ijk}^*}$ (хи-квадрат Пирсона);
- 2) $G^2 = 2 \sum_i \sum_j \sum_k n_{ijk} \ln \left(\frac{n_{ijk}}{n_{ijk}^*} \right)$ (хи-квадрат метода максимального правдоподобия);

где n_{ijk} — наблюдаемые частоты, n_{ijk}^* — теоретические частоты. Если статистики окажутся выше своего критического уровня при заданном уровне значимости, то эффект считается значимым и должен присутствовать в модели. В противном случае эффект может быть исключен из модели. Аналогичным образом данная модель может быть обобщена и на случай, когда таблица сопряженности признаков содержит более трех категоризованных переменных.

В статье [12] рассматривается еще один инструмент обработки категоризованных признаков — категориальный метод главных компонент, который использует преобразование категориальных данных в количественные латентные переменные, чтобы затем применить к ним стандартный РСА. Данный метод применяется для решения задачи снижения размерности пространств исходных данных, измеренных в любых шкалах. Он позволяет одновременно анализировать как интервальные, порядковые, номинальные и дихотомические переменные, так и качественные переменные. Исходными данными выступает матрица «объект — признак»:

$$X_{n \times p} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix},$$

где переменная x_{ij} принимает одно из L_j различных значений, соответствующее градации категоризованного признака x_j (L_j — число категорий признака), $i = 1, \dots, n$, $j = 1, \dots, p$. Результатом метода является таблица с нагрузками по каждому признаку, что позволяет объединить эти признаки в группы (или факторы). Алгоритм обработки данных в категориальном методе главных компонент включает два этапа: 1) «оцифровка» переменных; 2) редукция размерности данных. Математическая формализация метода предполагает нахождение матрицы интегральных характеристик (объектных счетов) Z размера $n \times r$ таким образом, чтобы функция $\sigma(Z, W)$ принимала минимальное значение

$$\sigma(Z, W) = \sum_{j=1}^p \text{tr} \left[(Z - G_j W_j)^T (Z - G_j W_j) \right] \rightarrow \min$$

при ограничениях центрированности и ортонормированности:

$$Z^T 1_n = 0_r, Z^T Z = n I_r,$$

где G_j — $n \times L_j$ — матрица индикаторов для исходной переменной x_j :

$$G_j(i, l_j) = \begin{cases} 1, & \text{если объект } i \text{ относится к категории } l_j, \\ 0, & \text{в противном случае,} \end{cases}$$

а W_j — $L_j \times r$ — матрица переменной x_j , содержащая координаты всех ее категорий в r -ом пространстве, 1_n — единичный вектор размерности $n \times 1$, 0_r — нулевой вектор размерности $r \times 1$, I_r — единичная

матрица размером $r \times r$. Оптимизация целевой функции при данных ограничениях осуществляется с помощью итерационного алгоритма PRINCALS, который минимизирует функцию потерь $\sigma(Z, W)$ через чередующуюся оптимизацию параметров. Алгоритм находит оптимальные коды (числовые замены для категорий переменных) для категорий и главные компоненты одновременно.

Анализ представленных выше работ по применению многомерных статистических методов для обработки категоризованных переменных, описывающих социальные аспекты жизни, и исследованию взаимосвязей между ними показывает, что распространенными инструментами изучения являются: регрессионный анализ (пробит-модель, логит-модель, модель с фиктивными переменными), факторный анализ, мэтчинг, анализ таблиц сопряженности признаков (логлинейный анализ, метод анализа соответствий), категориальный метод главных компонент.

В отличие от рассмотренных методов, которые в основном применяются к двумерным массивам данных, в настоящей статье предлагается математическая модель многомерного компонентного анализа для обработки категоризованных данных и выявления сложных многомерных закономерностей (многомерных паттернов) между категоризованными признаками с использованием тензорного представления. При этом оси (моды) такого тензора соответствуют категоризованным признакам, а измерения по осям — градациям (категориям) этих признаков.

2. Метод многомерного компонентного анализа

2.1. Основные определения

Следуя работе [13], для описания многомерного компонентного анализа определим следующие основные понятия: тензор, мультииндекс, матрица развертки, векторизация тензора.

Пусть d — целое число, $d \geq 1$.

Определение 1. d -мерным тензором называется массив данных, который характеризуется следующими параметрами:

- 1) d — количество осей тензора (размерность тензора);
- 2) форма тензора $(n_1, \dots, n_d)$ — упорядоченный набор целых чисел, описывающий количество измерений на каждой оси тензора, $n_i \geq 1, i = 1, \dots, d$;
- 3) тип данных.

Тензор будем обозначать как $\underline{X}$, а его элементы $\underline{X}(i_1, \dots, i_d), i_k = 1, \dots, n_k, k = 1, \dots, d$.

Рассмотрим кортеж индексов $(i_1, \dots, i_d)$, где $i_k \in \{1, \dots, n_k\}, k = 1, \dots, d$.

Определение 2. Мультииндексом, который соответствует кортежу $(i_1, \dots, i_d)$, называется показатель, определяемый по правилу либо *лексикографического* порядка (первый индекс в кортеже меняется медленнее всех, а последний — быстрее всех):

$$\sigma = \overline{i_1 \dots i_d} = i_d + (i_{d-1} - 1)n_d + (i_{d-2} - 1)n_d n_{d-1} + \dots + (i_1 - 1)n_d n_{d-1} \dots n_2, \quad (1)$$

либо *обратного лексикографического* порядка (последний индекс в кортеже меняется медленнее всех, а первый — быстрее всех):

$$\sigma = \overline{i_1 \dots i_d} = i_1 + (i_2 - 1)n_1 + (i_3 - 1)n_1 n_2 + \dots + (i_d - 1)n_1 n_2 \dots n_{d-1}. \quad (1')$$

Определение 3. Пусть выбрано произвольное $k \in \{1, \dots, d\}$. Мультииндексом, соответствующим кортежу $(i_1, \dots, i_{k-1}, i_{k+1}, \dots, i_d)$ (исключается ось с номером k) называется мультииндекс $\sigma_k = \overline{i_1 \dots i_{k-1} i_{k+1}, \dots, i_d}$, определяемый равенством (1) или (1') для $(d-1)$ -мерного тензора с формой $(n_1, \dots, n_{k-1}, n_{k+1}, \dots, n_d)$.

Всюду далее, если специально не оговорено, используется мультииндекс, рассчитанный по правилу лексикографического порядка.

Рассмотрим d -мерный тензор $\underline{X}$ с формой $(n_1, \dots, n_d)$ и пусть $n = \prod_{i=1}^d n_i$.

Определение 4. Векторизацией $\underline{\text{vec}}(\underline{X})$ d -мерного тензора $\underline{X}$ с формой $(n_1, \dots, n_d)$ называется n -мерный вектор-столбец, элементы которого определяются по правилу

$$(\underline{\text{vec}}(\underline{X}))_{\sigma} = \underline{X}(i_1, \dots, i_d), \quad (2)$$

где мультииндекс σ рассчитывается в соответствии с формулой (1).

Замечание. Если матрицу рассматривать как двумерный тензор, а мультииндекс рассчитывать по правилу обратного лексикографического порядка, то операция векторизации такого тензора будет равносильна обычной операции векторизации матрицы, когда матрица преобразуется в вектор-столбец последовательным размещением ее столбцов один под другим.

Определение 5. Матрицей развертки $X_{(k)}$ d -мерного тензора $\underline{X}$ с формой $(n_1, \dots, n_d)$ соответствующей оси $k \in \{1, \dots, d\}$, называется матрица размера $n_k \times (n/n_k)$, элементы которой определяются по правилу

$$(X_{(k)})_{i_k, \sigma_k} = \underline{X}(i_1, \dots, i_k, \dots, i_d), \quad (3)$$

где $i_k = 1, \dots, n_k$, а мультииндекс σ_k рассчитывается в соответствии с определением 3.

Определение 6. Коммутационной матрицей $K_{m,n} = (K(\alpha, \beta))$, $\alpha = 1, \dots, n$, $\beta = 1, \dots, m$ называется блочная квадратная матрица порядка mn , блок которой $K(\alpha, \beta)$ есть матрица размера $m \times n$ с элементами $(K(\alpha, \beta))_{ij} = \begin{cases} 1, & i = \beta, j = \alpha; \\ 0, & \text{в остальных случаях} \end{cases}$, $i = 1, \dots, m$, $j = 1, \dots, n$.

Замечание. Отметим следующие свойства коммутационной матрицы [14]:

- 1) коммутационная матрица является матрицей перестановок;
- 2) $K_{m,n}^T = K_{m,n}^{-1} = K_{n,m}$, где T — операция транспонирования матрицы;
- 3) $K_{n,1} = K_{1,n} = I_n$, где I_n — единичная матрица размера n ;
- 4) если $\underline{X} = X$ — матрица размера $m \times n$, то операция векторизации тензора $\underline{\text{vec}}(\underline{X})$ и операция векторизации матрицы $\text{vec}(X)$ связаны следующими равенствами:

$$\underline{\text{vec}}(\underline{X}) = \underline{\text{vec}}(X) = \text{vec}(X^T) = K_{m,n} \text{vec}(X), \quad (4)$$

где $K_{m,n}$ — коммутационная матрица. При этом равенство (4) верно и для транспонированной матрицы X^T .

2.2. Многомерный компонентный анализ для категоризованных переменных

Как известно [14], задачей многомерного компонентного анализа является построение аппроксимации d -мерного тензора $\underline{X}$ с формой $(n_1, \dots, n_d)$, в которой каждый элемент данного тензора заменяется линейной комбинацией элементов также d -мерного тензора $\underline{Z}$, но с формой $(n'_1, \dots, n'_d)$, где $n'_i \leq n_i$ для всех $i = 1, \dots, d$. При этом тензор $\underline{Z}$ должен сохранять всю или почти всю информацию о различиях между исходными многомерными данными. В настоящей работе многомерный компонентный анализ применяется к тензорам, оси которых соответствуют категоризованным переменным, а градации этих переменных — измерениям по их осям. Элементами таких тензоров являются числовые значения, которые количественно описывают взаимосвязь или результат измерения, соответствующий данной комбинации категорий.

Вначале рассмотрим случай, когда двумерные тензоры $\underline{X} = X$ и $\underline{Z} = Z$ представляют собой матрицы соответственно размеров $n_1 \times n_2$ и $n'_1 \times n'_2$. Тогда задача аппроксимации заключается

в нахождении матрицы Z , а также матриц A и B соответственно размеров $n_1 \times n'_1$ и $n_2 \times n'_2$, так, чтобы выполнялось равенство

$$X = AZB^T + E, \quad (5)$$

где матрица E имеет размеры $n_1 \times n_2$, а матрицы A и B удовлетворяют условиям (являются полуортогональными)

$$A^T A = I_{n'_1}, \quad B^T B = I_{n'_2}, \quad n'_1, n'_2 \leq \text{rang}(X). \quad (6)$$

Используя оператор векторизации матриц, операцию произведения Кронекера и их свойства, представим равенство (5) в следующем виде

$$\text{vec}(X^T) = \text{vec}(BZ^T A^T + E^T) = \text{vec}(BZ^T A^T) + \text{vec}(E^T) = (A \otimes B)\text{vec}(Z^T) + \text{vec}(E^T),$$

где $A \otimes B$ — произведение Кронекера матриц A и B . Применяя равенство (4), запишем полученное равенство в виде

$$\underline{\text{vec}}(X) = (A \otimes B)\underline{\text{vec}}(Z) + \underline{\text{vec}}(E)$$

или

$$x = Cz + e, \quad (7)$$

где

$$x = \underline{\text{vec}}(X), z = \underline{\text{vec}}(Z), e = \underline{\text{vec}}(E), C = A \otimes B. \quad (8)$$

Поскольку элементами тензора являются числовые значения некоторой измеряемой величины (например, оценка, рейтинг, вероятность, количество), то определим целевую функцию отклонения исходного тензора $\underline{X}$ от его аппроксимации как сумму квадратов отклонений (как квадрат евклидовой нормы вектора e)

$$\phi(C, z) = (x - Cz)^T (x - Cz). \quad (9)$$

Следовательно, задача двумерного компонентного анализа заключается в нахождении матриц A и B , удовлетворяющих условиям (6), также вектора z , при которых функция (9) принимает наименьшее значение. Решение этой задачи основано на теореме Эккарта–Юнга и представлено в монографии [14].

Обобщая эту двумерную модель на тензоры произвольной размерности, получим задачу многомерного компонентного анализа, заключающуюся в нахождении матриц $A_1, \dots, A_d$, называемых матрицами компонентных нагрузок, и тензора $\underline{Z}$ таких, что функция (9) принимает наименьшее значение, где

$$x = \underline{\text{vec}}(\underline{X}), z = \underline{\text{vec}}(\underline{Z}), C = A_1 \otimes \dots \otimes A_d, \quad (10)$$

а прямоугольные матрицы A_i размера $n_i \times n'_i$ удовлетворяют условию полуортогональности

$$A_i^T A_i = I_{n'_i}, i = 1, \dots, d. \quad (11)$$

Тензор $\underline{Z}$ называется *ядром тензора* $\underline{X}$, поскольку каждый элемент тензора $\underline{X}$ есть линейная комбинация, коэффициентами которой являются элементы матрицы C . Измерениями осей тензора $\underline{Z}$ выступают новые, математически искусственно созданные градации переменных — *компоненты* (или *концепты*). Поэтому составляющие ядра тензора представляют собой интегральные характеристики взаимовлияния исходных переменных, которые показывает силу взаимодействия между компонентами разных осей тензора. Нулевые или малые составляющие ядра тензора указывают на отсутствие связи между компонентами.

Аналогичная задача аппроксимации многомерного тензора, называемая аппроксимацией Таккера, также рассматривается в работе [15].

2.3. Решение задачи многомерного компонентного анализа и матрицы развертки

Решение задачи многомерного компонентного анализа основано на сингулярном разложении вещественных прямоугольных матриц.

Теорема 1 (сингулярное разложение вещественных матриц [14]). Пусть A — вещественная матрица размера $t \times q$, имеющая ранг r . Тогда матрица A может быть представлена единственным образом в виде $A = V\Sigma W^T$, где V и W — вещественные полуортогональные матрицы размеров $t \times r$ и $q \times r$ соответственно ($V^T V = I_r$, $W^T W = I_r$), а Σ — вещественная диагональная матрица размера $r \times r$, по главной диагонали которой расположены в порядке убывания положительные квадратные корни из собственных значений матрицы либо AA^T (если $t \leq q$), либо $A^T A$ (если $t > q$), называемые сингулярными числами матрицы A , а столбцы матриц V и W называются соответственно левыми и правыми сингулярными векторами матрицы A . При этом, если $t \leq q$, то столбцы матрицы V , упорядоченные в соответствие с сингулярными числами, есть ортонормированные собственные векторы матрицы AA^T , а матрица W определяется равенством $W = A^T V \Sigma^{-1}$. Если $t > q$, то столбцы матрицы W , упорядоченные в соответствии с сингулярными числами, есть ортонормированные собственные векторы матрицы $A^T A$, а матрица V определяется равенством $V = A W \Sigma^{-1}$.

Замечание. Способ построения сингулярного разложения вещественной матрицы, описанный в теореме 1, позволяет однозначно определить это разложение.

Используя сингулярное разложение вещественных матриц, решение задачи многомерного компонентного анализа может быть найдено следующим образом.

Обозначим

$$n = \prod_{i=1}^d n_i; \quad n' = \prod_{i=1}^d n'_i; \quad \alpha_0 = 1, \alpha_k = n_1 \cdot n_2 \dots \cdot n_k, k = 1, \dots, d; \quad \beta_k = \frac{n}{\alpha_{d-k}}, k = 0, 1, \dots, d.$$

Теорема 2 ([14]). Минимум функции (9) равен $x^T x - z^T z$ и достигается, когда матрицы $A_1, \dots, A_d$ и вектор z удовлетворяют следующим условиям:

$$z = (A_1 \otimes \dots \otimes A_d)^T x, \quad (12)$$

где для всех $i = 1, \dots, d$ A_i — матрица размера $n_i \times n'_i$, состоящая из левых сингулярных векторов, соответствующих n'_i максимальным сингулярным числам матрицы $X_i A_{[i]}$. При этом матрица $A_{[i]}$ размера $(n/n_i) \times (n'/n'_i)$ задается равенством

$$A_{[i]} = A_1 \otimes \dots \otimes A_{i-1} \otimes A_{i+1} \otimes \dots \otimes A_d, \quad (13)$$

а матрица X_i размера $n_i \times (n/n_i)$ удовлетворяет условию

$$\text{vec}(X_i) = (I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}) x. \quad (14)$$

Покажем, что матрицы X_i , удовлетворяющие равенству (14), непосредственно связаны с матрицами развертки $X_{(i)}$ тензора $\underline{X}$.

Теорема 3. Пусть $X_{(i)}$ — матрица развертки тензора $\underline{X}$, соответствующая оси i , а X_i — матрица, удовлетворяющая условию (14). Тогда верны следующие равенства

$$x = \underline{\text{vec}}(\underline{X}) = (I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}) \text{vec}(X_{(i)}), \quad i = 1, \dots, d, \quad (15)$$

$$x = \underline{\text{vec}}(\underline{X}) = \text{vec}(X_{(d)}), \quad (16)$$

$$\text{vec}(X_i) = (I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}^2) \text{vec}(X_{(i)}), \quad i = 1, \dots, d. \quad (17)$$

Доказательство. 1. Докажем справедливость равенства

$$\text{vec}(X_{(i)}) = (I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}})x, \quad i = 1, \dots, d. \quad (18)$$

Заметим, что матрица $I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}}$ является блочно-диагональной, у которой на главной диагонали расположены матрицы $K_{n_i, \beta_{d-i}}$. Занумеруем эти блоки индексом $q = 1, \dots, \alpha_{i-1}$. Так как матрица $K_{n_i, \beta_{d-i}}$ является матрицей перестановок, то и матрица $I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}}$ также будет матрицей перестановок, в которой 1 расположены в следующем порядке номеров столбцов (строки следуют в обычном порядке от первой до последней):

$$1 + (s-1)\beta_{d-i}, \quad s = 1, \dots, n_i; \quad 2 + (s-1)\beta_{d-i}, \quad s = 1, \dots, n_i; \dots; \quad \beta_{d-i} + (s-1)\beta_{d-i}, \quad s = 1, \dots, n_i,$$

где строки матрицы $I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}}$ соответствуют блоку $q = 1$.

Аналогично для произвольного блока q ($q = 1, \dots, \alpha_{i-1}$) матрицы $I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}}$ порядок будет следующий:

$$1 + (q-1)\beta_{d-i}n_i + (s-1)\beta_{d-i}, \quad s = 1, \dots, n_i; \dots; \quad \beta_{d-i} + (q-1)\beta_{d-i}n_i + (s-1)\beta_{d-i}, \quad s = 1, \dots, n_i,$$

Следовательно, элементы вектора $x = \text{vec}(X)$ при умножении слева на матрицу перестановок $(I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}})$ будут расположены в соответствующем векторе $(I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}})x$ в следующей последовательности (для блока $q = 1$):

$$\begin{aligned} & \left(\underline{X}(1, \dots, 1, \dots, 1), \underline{X}(1, \dots, 2, \dots, 1), \dots, \underline{X}(1, \dots, n_i, \dots, 1) \right)^T, \\ & \left(\underline{X}(1, \dots, 1, \dots, 2), \underline{X}(1, \dots, 2, \dots, 2), \dots, \underline{X}(1, \dots, n_i, \dots, 2) \right)^T, \\ & \dots, \\ & \left(\underline{X}(1, \dots, 1, n_{i+1}, \dots, n_d), \underline{X}(1, \dots, 2, n_{i+1}, \dots, n_d), \dots, \underline{X}(1, \dots, n_i, n_{i+1}, \dots, n_d) \right)^T, \end{aligned}$$

что является соответственно $1, 2, \dots, (n_d \cdot \dots \cdot n_{i+1})$ столбцами матрицы развертки $X_{(i)}$.

Аналогично, выписывая элементы вектора $(I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}})x$ для остальных блоков $q = 2, \dots, \alpha_{i-1}$ матрицы $I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}}$, получим другие столбцы матрицы $X_{(i)}$. Таким образом, верно равенство (18).

2. Так как матрица $I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}}$ невырожденная, поскольку является матрицей перестановок, то используя равенство (18) и свойства коммутационной матрицы, произведения Кронекера, векторизации матриц, получим

$$x = (I_{\alpha_{i-1}} \otimes K_{n_i, \beta_{d-i}})^{-1} \text{vec}(X_{(i)}) = (I_{\alpha_{i-1}}^{-1} \otimes K_{n_i, \beta_{d-i}}^{-1}) \text{vec}(X_{(i)}) = (I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}) \text{vec}(X_{(i)}),$$

т. е. равенство (15).

3. Подставив в равенство (15) $i = d$, получим

$$x = (I_{\alpha_{d-1}} \otimes K_{\beta_0, n_d}) \text{vec}(X_{(d)}) = (I_{n/n_d} \otimes K_{1, n_d}) \text{vec}(X_{(d)}) = (I_{n/n_d} \otimes I_{n_d}) \text{vec}(X_{(d)}) = I_n \text{vec}(X_{(d)}) = \text{vec}(X_{(d)}),$$

т. е. равенство (16).

4. Подставив в равенство (14) соотношение (15), получим

$$\begin{aligned} \text{vec}(X_i) &= (I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i})(I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}) \text{vec}(X_{(i)}) = ((I_{\alpha_{i-1}} \cdot I_{\alpha_{i-1}}) \otimes (K_{\beta_{d-i}, n_i} \cdot K_{\beta_{d-i}, n_i})) \text{vec}(X_{(i)}) = \\ &= (I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}^2) \text{vec}(X_{(i)}), \end{aligned}$$

т. е. равенство (17). □

Замечание. Так как матрица $I_{\alpha_{i-1}} \otimes K_{\beta_{d-i}, n_i}^2$ является матрицей перестановок, а матрицы X_i и $X_{(i)}$ имеют одинаковый размер $n_i \times (n/n_i)$, то X_i получается из $X_{(i)}$ перестановкой ее элементов. Поэтому удобно формировать и хранить исходные данные не в виде тензора $\underline{X}$, а в виде его матриц развертки $X_{(i)}$, и по формуле (17) с использованием операции, обратной векторизации матриц (функция `reshape` [14]), находить матрицы X_i . Таким образом, матрицы развертки $X_{(i)}$ используются в решении задачи многомерного компонентного анализа.

2.4. Итерационный алгоритм нахождения матриц компонентных нагрузок

Как видно из теоремы 2 для каждого $i = 1, \dots, d$ матрица A_i состоит из левых сингулярных векторов матрицы $X_i A_{[i]}$, где матрица $A_{[i]}$ зависит от матриц $A_1, \dots, A_{i-1}, A_{i+1}, \dots, A_d$, согласно равенству (13). Поэтому получить точное выражение для матриц A_i не представляется возможным. В связи чем в настоящей статье рассматривается итерационный алгоритм нахождения данных матриц, основанный на подходе, описанном в работе [14].

Пусть тензор $\underline{X}$ представлен в виде его матриц развертки $X_{(i)}, i = 1, \dots, d$.

В соответствии с теоремами 2 и 3 данный итерационный алгоритм включает следующие шаги:

1. Зададим априорно $n'_i \leq \text{rang}(X_{(i)}), i = 1, \dots, d$, (выбор значения n'_i может быть обусловлен, например, задачей снижения размерности или содержательным смыслом задачи).
2. Определим матрицы $X_i = \text{reshape}(\text{vec}(X_{(i)}), n_i, n/n_i)$, где $\text{vec}(X_{(i)})$ рассчитывается по формуле (17), $i = 1, \dots, d$.
3. Зададим произвольным образом начальные значения $A_1^{(0)}, A_2^{(0)}, \dots, A_d^{(0)}$ матриц $A_1, A_2, \dots, A_d$ например, $A_i^{(0)} = [e_1, \dots, e_{n'_i}]$, где для всех $s = 1, \dots, n'_i$ e_s — это n_s -мерный вектор-столбец, элементы которого равны

$$e_{sj} = \begin{cases} 1, & s = j; \\ 0, & s \neq j; \end{cases}, j = 1, \dots, n_s.$$

4. На каждой итерации $k = 0, 1, 2, \dots$ последовательно для каждого $i = 1, \dots, d$ найдем по формуле (13) матрицу $A_{[i]}^{(k)} = A_1^{(k+1)} \otimes \dots \otimes A_{i-1}^{(k+1)} \otimes A_{i+1}^{(k)} \otimes \dots \otimes A_d^{(k)}$, а также матрицу $Q_i^{(k)} = X_i A_{[i]}^{(k)}$, для которой вычислим ее сингулярное разложение $Q_i^{(k)} = V_i^{(k)} \Sigma_i^{(k)} [W_i^{(k)}]^T$ (теорема 1); и далее определим матрицу $A_i^{(k)}$, состоящую из первых n'_i столбцов матрицы $V_i^{(k)}$ (соответствуют n'_i максимальным сингулярным числам матрицы $Q_i^{(k)}$).
5. Итерационный процесс продолжаем до тех пор, пока не будет выполняться условие остановки: для заданной точности $\varepsilon > 0$

$$MSE(A_i^{(k)}, A_i^{(k-1)}) = \frac{1}{n_i \cdot n'_i} \|A_i^{(k)} - A_i^{(k-1)}\|_{l_2}^2 < \varepsilon, i = 1, \dots, d. \quad (19)$$

Обозначим через $k^* (k^* \geq 1)$ номер итерации, для которой выполнено условие остановки (19). Определим матрицу $C = A_1 \otimes \dots \otimes A_d$, где $A_i = A_i^{(k^*)}, i = 1, \dots, d$. По формулам (12) и (16) вычислим вектор $z = C^T \text{vec}(X_{(d)})$ и найдем ядро тензора $\underline{X}$ — тензор $\underline{Z} = \text{reshape}(z, (n'_1, \dots, n'_d))$.

2.5. Интерпретация результатов многомерного компонентного анализа

Для произвольного i -го категоризованного признака (i -го измерения тензора) рассмотрим его матрицу компонентных нагрузок $A_i, i \in \{1, \dots, d\}$. И пусть $s_i \in \{1, \dots, n_i\}$ и $k_i \in \{1, \dots, n'_i\}$ есть некоторые произвольные категория и i -ориентированная компонента данного признака (моды) соответственно. Так как коэффициенты матрицы A_i по модулю не превосходят 1 (следует из условия (11)), то ее элемент $A_i(s_i, k_i)$ показывает вклад (корреляцию, вес) категории s_i в k_i -ю компоненту i -го категоризованного признака.

Для определения содержательного смысла компоненты k_i для моды i выбираем градации этой категоризованной переменной с наибольшими по абсолютной величине корреляциями (с наибольшими весами) этих градаций с данной компонентой. На практике часто используют эмпирические правила определения силы связи, например, $|A_i(s_i, k_i)| > 0,7$ — очень сильная связь (отличная), $|A_i(s_i, k_i)| > 0,5$ — существенная связь (хорошая), $|A_i(s_i, k_i)| > 0,2$ — умеренная связь (приемлемая, но требует осторожности в интерпретации), $|A_i(s_i, k_i)| \leq 0,2$ — слабая связь (данная компонента плохо представляет эту категорию). Затем интерпретируем компоненту как некоторый кластер отобранных по корреляции градаций i -го категоризованного признака (общее понятие или латентный фактор, которые описывают эти градации).

Важно заметить следующее, касающееся интерпретации компонент. Во-первых, когда несколько градаций имеют равные или почти равные значения нагрузок (корреляций) на один и тот же фактор (компоненту), то это означает, что данные градации являются взаимозаменяемыми индикаторами одной и той же компоненты. Их вклад в формирование фактора равноценен. Во-вторых, алгоритм вычисления сингулярных векторов, которые и являются компонентными нагрузками, может с одинаковой вероятностью находить вектор или противоположный ему вектор. Оба решения являются математически корректными и содержат одну и ту же информацию, поэтому в случае необходимости (для более содержательной интерпретации компонент) матрицу компонентных нагрузок можно умножить на (-1) . При этом, согласно формуле (12) элементы ядра тензора изменят знак, если число матриц компонентных нагрузок, умноженных на (-1) будет нечетно, и останутся без изменений — если четно.

Ключевая особенность интерпретации результатов многомерного компонентного анализа состоит в том, что концепты разных категориальных признаков взаимодействуют друг с другом через ядро тензора, что в свою очередь позволяет интерпретировать сложные многомерные комплексные закономерности в данных. При этом ядро тензора показывает силу взаимодействия между концептами разных категориальных признаков: чем больше по абсолютной величине элемент ядра, тем сильнее взаимодействие между концептами, комбинации которых соответствует это значение.

Одна и та же градация i -го категоризованного признака может иметь значимые нагрузки на несколько компонент этой переменной. Интегральным показателем того, насколько хорошо категория s_i представлена во всех компонентах своего i -го признака, может служить некоторая векторная норма, рассчитанная для вектора нагрузок этой категории на компоненты данной категориальной переменной. В настоящей статье в качестве такой векторной нормы предлагается использовать норму Гельдера l_p , рассчитанную для нескольких показателей: $p = 1; 2; \infty; 4$.

Пусть вектор $a_i(s_i)$ — это строка s_i матрицы A_i . Тогда векторные нормы будут следующими:

$$1) \text{ манхэттенское расстояние } l_1 : \|a_i(s_i)\|_{l_1} = \sum_{k_i=1}^{n'_i} |A_i(s_i, k_i)|, \quad (20)$$

$$2) \text{ евклидова норма } l_2 : \|a_i(s_i)\|_{l_2} = \left(\sum_{k_i=1}^{n'_i} (A_i(s_i, k_i))^2 \right)^{1/2}, \quad (21)$$

$$3) l_\infty\text{-норма: } \|a_i(s_i)\|_{l_\infty} = \max_{k_i=1, \dots, n'_i} |A_i(s_i, k_i)|, \quad (22)$$

$$4) \text{ норма Гельдера } l_4 : \|a_i(s_i)\|_{l_4} = \left(\sum_{k_i=1}^{n'_i} (A_i(s_i, k_i))^4 \right)^{1/4}. \quad (23)$$

Использование *евклидовой нормы* l_2 обусловлено оценкой *коммунальности* градации категоризованного признака: доли вариации, которую компоненты моды объясняют в данных по конкретной градации этой же моды. Коммунальность рассчитывается как сумма квадратов компонентных нагрузок в строке s_i матрицы A_i (квадрат евклидовой нормы).

Целью вычисления *манхэттенского расстояния* l_1 является оценка суммарного вклада градации во все компоненты, когда компонентные нагрузки учитываются линейно и без уменьшения их абсолютных значений.

Норма l_∞ используется для того, чтобы оценить наибольший вклад, который вносит градация категориального признака в компоненты данной моды.

Поскольку при $|x| < 1$ (из условия (11) следует, что элементы матрицы компонентных нагрузок по абсолютной величине не более 1) функция x^4 убывает значительно быстрее, чем x^2 , то интегральный показатель совокупного влияния градации на все компоненты моды, представленный как *норма Гельдера* l_4 будет сильнее штрафоваться, чем при использовании евклидовой нормы l_2 , за компонентные нагрузки с абсолютными значениями, близкими к нулевым. Поэтому основная цель вычисления нормы l_4 — это более существенно учесть в интегральном показателе, в отличие от остальных норм, те компоненты моды, которые имеют слабую связь с градацией соответствующего категориального признака.

Так как для каждой градации категоризованной переменной рассчитываются четыре векторные нормы, каждая из которой учитывает определенную особенность при оценке влияния этой градации на компоненты данной моды, то для упорядочивания градаций этого категоризованного признака по их корреляции с компонентами предлагается следующий метод ранжирования, основанный на значениях вычисленных норм. При этом предполагаем, что все интегральные показатели (векторные нормы), оценивающие влияние градаций на компоненты моды, равнозначны.

Для каждой нормы по убыванию ее величины проранжируем градации i -го категориального признака следующим образом: градации с наибольшей нормой присвоим ранг 1, следующей градации по убыванию нормы — ранг 2 и т. д. В результате каждая градация $s_i = 1, \dots, n'_i$ будет иметь по четыре ранга. Найдем сумму полученных рангов и проранжируем на основании этого значения градации i -го категориального признака по степени его влияния на все компоненты данной моды: наименьшей сумме рангов соответствует наибольшее влияние.

2.6. Сравнительная оценка с альтернативными методами анализа многомерных категоризованных переменных

Ключевые особенности данных, которые в настоящей статье моделируются тензорами с категоризованными осями, заключаются в следующем: 1) ни одна из категоризованных переменных не является «явно» зависимой от остальных категоризованных признаков, но при этом предполагается, что существуют сложные многомерные паттерны в данных; 2) по некоторым категоризованным переменным к одному объекту может относиться не одна категория, а несколько или все категории; 3) совокупность категоризованных осей может описывать качественные характеристики различных областей, относящихся к людям, объектам, событиям; 4) не требуется предположения о распределении ни самих данных, ни ошибок модели.

Так методы оценки регрессионных зависимостей с категоризованными переменными: регрессионные уравнения с фиктивными переменными, пробит-регрессия, регрессия с заменой категоризованных признаков на количественные переменные в виде специальных индексов, многомерная рекурсивная пробит-модель, мэтчинг с логистической регрессией и дискретная регрессионная модель множественного выбора, — не могут быть использованы для оценки зависимостей между категоризованными данными, имеющими описанную выше первую особенность. Кроме того, некоторые из указанных методов требуют предположения о распределении ошибок модели.

Наличие в данных второй особенности не позволяет «напрямую» применить к совокупности категоризованных переменных логлинейный анализ, метод анализа соответствий (один из методов многомерного шкалирования для категоризованных переменных) и категориальный метод главных компонент. Чтобы исключить эту особенность, но сохранить всю информацию, необходимо вместо тех категоризованных переменных, по которым к одному объекту относятся несколько категорий, определить новые категоризованные признаки в виде комбинаций градаций исходных переменных. При этом количество градаций и сами градации новых категоризованных переменных будут определяться возможными значениями элементов тензора, соответствующими комбинации категорий исходных признаков. В результате число категоризованных переменных значительно возрастает, что приводит к росту вычислительной сложности алгоритмов логлинейного анализа, анализа соответствий и категориального метода главных компонент. А также компоненты будут формироваться не для исходных категоризованных переменных, а для вновь определенных переменных. Кроме того, теряется содержательная интерпретация результатов. Так, например, качественное ограничение итерационного алгоритма PRINCALS, реализующего категориальный метод главных компонент: 20–30 переменных для содержательного анализа [16]. Для логлинейного анализа такое качественное ограничение составляет 10 переменных, и, кроме того, в данном методе требуется, что наблюдаемые частоты имеют пуассоновское или полиномиальное распределение. Также логлинейный анализ позволяет только обнаруживать и исследовать внутренние взаимосвязи между категоризованными переменными, но не выделять в этих переменных латентные классы (компоненты). Для метода анализа соответствий качественное ограничение составляет 10–20 переменных.

Кроме анализа соответствий к методам многомерного шкалирования для категоризованных переменных относится нелинейное многомерное шкалирование на основе нейросетей. Этот подход использует нейронную сеть (автоэнкодер), а входными данными являются матрицы индикаторов для категоризованных переменных (как в категориальном методе главных компонент) или эмбединг-кодированные категоризованные признаки (преобразованные в числовые векторные представления категоризованные переменные). Данный метод может улавливать сложные нелинейные зависимости, но обучение нейронной сети требует большого объема данных. Реализация метода требует использования фреймворков глубокого обучения, таких как TensorFlow/Keras или PyTorch.

Если есть основания полагать, что в категоризованных данных присутствует мультилинейная структура и критически важна оценка неопределенности (например, важна вероятность принадлежности градации соответствующей компоненте), то для анализа категоризованных признаков можно использовать байесовское СР-разложение для этих переменных. Этот метод моделирует категоризованные наблюдения через латентные непрерывные факторы, для которых априорно задаются вероятностные распределения, и функцию связи (softmax). В результате получим не только компоненты для каждой моды, но и полную оценку неопределенности для всех параметров модели, что критически важно для принятия решений в условиях неполных данных. Самое основное ограничение этого метода — это вычислительная сложность (количество параметров растет экспоненциально с увеличением размерности тензора, требует вычисления градиента правдоподобия по всем параметрам, что для больших тензоров и моделей с softmax может быть очень медленным).

В отличие от рассмотренных методов, компонентный анализ может применяться к многомерным массивам данных, в том числе и имеющим ключевые особенности, описанные выше, без построения новых категоризованных переменных (как в категориальном методе главных компонент, логлинейном анализе или анализе соответствий) и без оценки различных регрессионных уравнений как дискретного выбора со значительным числом фиктивных переменных (как в логит-моделях, пробит-моделях) так и множественной регрессии с заменой категориальных переменных на специальные индексы. Также при использовании компонентного анализа не требуется

каких-то допущений относительно вероятностных распределений (как в байесовском СР-разложении), не требуется больших объемов данных (как в методах, использующих нейронные сети), и не теряется с увеличением числа категоризованных переменных содержательная интерпретация результатов.

3. Пример многомерного компонентного анализа категоризованных переменных

3.1. Описание данных и определение тензора

Рассмотрим применение представленного выше алгоритма многомерного компонентного анализа для тензора, модами которого являются категоризованные признаки.

В рамках разработанного цифрового сервиса по сопровождению профориентации, образования и трудоустройства лиц с ОВЗ и инвалидностью (информационно-аналитического регионального портала «ПЕРСПЕКТИВА-PRO») предлагается система профессиональной ориентации данной категории лиц, которая включает заполнение опросника. Опросник позволяет выявлять уровень развития профессионально значимых личностных качеств (ПЗК), которые дают возможность лицам с ОВЗ и инвалидностью эффективно реализовывать свой личностный потенциал [1]. В опросник экспертами были включены 50 ПЗК, разбитые на 10 групп по 5 ПЗК в каждой группе (см. таблицы 1–3). Т. е. были выделены те значимые качества, которые, по мнению экспертов, имеют наибольшее влияние на успешное профессиональное самоопределение индивидов. Уровень выраженности этих качеств в соответствии с методикой [1] оценивается респондентами по 5 балльной шкале (от 1 до 5). При этом участнику опроса предлагается из пяти ПЗК, входящих в одну группу, выбрать то, которое по субъективной оценке присуще ему больше всего (этому выбранному качеству присваивается 5 баллов). Из оставшихся четырех он выбирает следующее присущее ему качество и т. д. (соответственно присваиваются баллы: 4 балла — второму, 3 балла — третьему и т. д.). Такую самооценку собственных профессионально значимых качеств респондент проводит по всем 10 группам ПЗК.

Так же на экспертном уровне для профессионального самоопределения индивидов были выделены 10 сфер человеческой деятельности (см. таблицы 1–3) и для каждой из этих 10 сфер определены по 10 наиболее часто встречающихся и, соответственно, профессионально значимых качеств (по одному качеству из каждой группы ПЗК), называемые эталонными. При этом каждое значимое качество является эталонным для двух сфер деятельности. Например, по ЭКН такими эталонными ПЗК являются: аккуратность (ПЗК1), желание быть в социуме (ПЗК2), умение искать и анализировать информацию (ПЗК3), математические способности (ПЗК4), умение выбирать нужную информацию из большого объема (ПЗК5), точно воспроизвести материал по памяти спустя долгое время (ПЗК6), предпочтение решать практические задачи (ПЗК7), умение адаптироваться к новым условиям (ПЗК8), умение находить решение (ПЗК9), умеренный темп деятельности (ПЗК 10). Полный перечень эталонных ПЗК для каждой сферы деятельности представлен в монографии [1].

Результаты опроса респондентов (самооценки ПЗК), распределенные на соответствующие сферы деятельности согласно перечню эталонных ПЗК, для использования в многомерном компонентном анализе предлагается формировать в виде d -мерного тензора $\underline{X}$ ($d \geq 3$), имеющего форму $(n_1, \dots, n_{d-2}, n_{d-1}, n_d)$, где оси (моды) $1, \dots, d-2$ относятся к категоризованным переменным, описывающим вариативные характеристики инвалидов и лиц с ОВЗ: регион проживания, социальный статус, возрастная группа, группа нозологии, уровень образования и т. д., а оси $d-1$ и d — к категоризованным переменным: сфера деятельности и группы ПЗК.

Обозначим балл эталонного качества из группы профессионально значимых качеств $i_d \in \{\text{ПЗК1}, \dots, \text{ПЗК10}\}$, в которую входит это качество, распределенный на сферу деятельности $i_{d-1} \in \{\text{ЭКН}, \dots, \text{УПР}\}$ и относящийся к участнику опроса, имеющего вариативные характеристики $(i_1, \dots, i_{d-2})$, как $q_{(i_1, \dots, i_{d-2})}(i_{d-1}, i_d)$. Заметим, что один и тот же набор значений вариативных характеристик $(i_1, \dots, i_{d-2})$ может быть у нескольких респондентов. Такую ситуацию можно было бы

Table 1. List of categorized variables for nosological groups**Таблица 1.** Список категоризованных переменных для групп нозологии

Сокращение	Наименование	Количество респондентов
ДВГ	Нарушения опорно-двигательного аппарата	127
МНТ	Ментальные нарушения	310
СНС	Сенсорные нарушения	337
ОБЩ	Общие заболевания	126

Table 2. List of categorized variables for fields of activity**Таблица 2.** Список категоризованных переменных для сфер деятельности

Сокращение	Наименование
ЭКН	Экономическая сфера
ЭКЛ	Экологическая сфера
НАУ	Научная сфера
ХУД	Художественная сфера
ИНФ	IT-сфера
ТЕХ	Производственно-технологическая сфера
МЕД	Медицинская и спортивно-оздоровительная сфера
СОЦ	Сфера социального обслуживания
ПЕД	Педагогическая сфера
УПР	Управленческая сфера

учесть, добавив в тензор дополнительную моду (ось) «объекты». Но поскольку для каждого объекта (участника опроса) соответствует единственный набор значений категориальных признаков, то такой тензор будет содержать большое количество нулевых элементов. Эти нулевые значения оценок ПЗК положим тем грациям вариативных характеристик, которые не относятся к данному объекту. Следовательно, чтобы не формировать разреженный тензор, и чтобы учесть самооценки ПЗК всех участников опроса, в статье предлагается в качестве элементов тензора рассматривать средние оценки значимых качеств, рассчитанные для респондентов с одинаковыми значениями вариативных характеристик.

Таким образом, элементы тензора $\underline{X}$ определим по следующему правилу

$$\underline{X}(i_1, \dots, i_{d-2}, i_{d-1}, i_d) = \frac{1}{N(i_1, \dots, i_{d-2})} \sum_{(i_1, \dots, i_{d-2})} q_{(i_1, \dots, i_{d-2})}(i_{d-1}, i_d), \quad i_k = 1, \dots, n_k, \quad k = 1, \dots, d, \quad (24)$$

где $N(i_1, \dots, i_{d-2})$ — число респондентов, имеющих одинаковые вариативные характеристики, соответствующие вектору значений категоризованных признаков $(i_1, \dots, i_{d-2})$.

В настоящем примере наряду с переменными: сфера деятельности и группы ПЗК рассматривается еще одна категоризованная переменная — группа нозологии (см. таблицу 1).

В таблицах 1–3 представлены градации категоризованных переменных, их обозначения, которые используются в последующих таблицах, количество респондентов, принимавших участие в опросе.

3.2. Исходные данные

На основе данных опроса инвалидов и лиц с ОВЗ по формуле (24) были рассчитаны элементы трехмерного тензора $\underline{X}$ с формой $(4; 10; 10)$, где 1-я ось (1-я мода) — группы нозологии, 2-я ось (2-я мода) — сферы деятельности, 3-я ось (3-я мода) — группы ПЗК. Тензор был сформирован в виде отдельных срезов по 1-й моде (по группам нозологии). Результаты этих расчетов представлены в таблицах 4–7. В таблицах содержатся средние оценки ПЗК, являющихся эталонными для данных

Table 3. List of categorized variables for PZK groups**Таблица 3.** Список категоризованных переменных для групп ПЗК

Сокращение	Наименование
ПЗК1	Личностные качества: аккуратность, наблюдательность, ответственность, самоконтроль, целеустремленность
ПЗК2	Личностные характеристики: увлечённость, желание быть в социуме, оптимизм, нескритичность к себе, инициативность
ПЗК3	Умения: прикладные умения (ручной труд), умение работать в жёстких временных рамках, умение выразительно и грамотно говорить и писать, умение планировать свою деятельность, умение искать и анализировать информацию
ПЗК4	Способности: математические, конструктивно-технические, творческие, организаторские, физические
ПЗК5	Внимание: работать с несколькими объектами и документами одновременно, быстро переключаться на другой вид работы, длительно сосредотачиваться на одном задании, выбирать нужную информацию из большого объема, не отвлекаться на посторонние помехи
ПЗК6	Память: вы можете быстро запомнить большой объем информации, вы можете точно воспроизвести материал по памяти спустя долгое время, вам легче запомнить то, что вы увидели, вы лучше запоминаете информацию на слух, у вас хорошая двигательная память
ПЗК7	Мышление: вы умеете выделять главное; вы можете предлагать нестандартные решения; вы умеете устанавливать связь между причиной и следствием; вы предпочитаете решать задачи в уме; вы предпочитаете решать практические задачи
ПЗК8	Адаптивность: эмоциональная сдержанность, стрессоустойчивость, умение адаптироваться к новым условиям, умение выдерживать значительные физические нагрузки, умение выдерживать значительные эмоциональные нагрузки
ПЗК9	Взаимодействие с другими людьми: умение ставить задачи, умение организовывать других, умение находить решение, умение учитывать мнение других, умение работать в команде
ПЗК10	Темп деятельности: монотонная деятельность в течение длительного периода времени, монотонная деятельность в течение короткого периода времени, деятельность в быстром темпе в течение длительного времени, деятельность в быстром темпе в течение короткого периода времени, умеренный темп деятельности

сфер деятельности. Так, например, в таблице 4 по строке ЭКН (экономическая сфера) указаны средние самооценки лиц, имеющих нарушения опорно-двигательного аппарата, тех качеств, которые по экспертному мнению являются профессионально значимыми для этой сферы деятельности (максимальная оценка — 5 баллов, минимальная оценка — 1 балл): *аккуратность* — 2.543 балла, *желание быть в социуме* — 2.858 балла, *умение искать и анализировать информацию* — 3.606 балла, *математические способности* — 2.614 балла, *умение выбирать нужную информацию из большого объема* — 3.465 балла, *точно воспроизвести материал по памяти спустя долгое время* — 2.646 балла, *предпочтение решать практические задачи* — 3.118 балла, *умение адаптироваться к новым условиям* — 3.252 балла, *умение находить решение* — 3.197 балла, *умеренный темп деятельности* — 3.850 балла.

Table 4. Average ratings of PZK (musculoskeletal disorders)

	ПЗК1	ПЗК2	ПЗК3	ПЗК4	ПЗК5	ПЗК6	ПЗК7	ПЗК8	ПЗК9	ПЗК10
ЭКН	2.543	2.858	3.606	2.614	3.465	2.646	3.118	3.252	3.197	3.850
ЭКЛ	3.016	3.488	3.606	2.480	2.819	3.906	3.315	2.394	2.764	2.661
НАУ	3.047	3.488	2.756	3.024	2.701	2.646	2.748	3.016	3.197	2.331
ХУД	3.047	3.402	2.756	3.803	2.969	2.929	2.748	3.000	3.685	2.661
ИНФ	2.803	2.449	2.740	3.079	3.047	2.646	3.118	3.339	3.685	3.850
ТЕХ	3.591	2.803	2.740	3.079	2.969	2.874	3.339	3.252	2.906	2.331
МЕД	2.543	2.803	2.567	2.480	3.047	3.906	3.315	3.339	2.906	3.276
СОЦ	2.803	3.402	2.567	3.803	2.819	2.874	3.339	2.394	2.764	3.276
ПЕД	3.016	2.858	3.331	3.024	3.465	2.646	2.480	3.000	2.449	2.882
УПР	3.591	2.449	3.331	2.614	2.701	2.929	2.480	3.016	2.449	2.882

Таблица 4. Средние оценки ПЗК (нарушения опорно-двигательного аппарата)**Table 5.** Average ratings of PZK (mental disorders)

	ПЗК1	ПЗК2	ПЗК3	ПЗК4	ПЗК5	ПЗК6	ПЗК7	ПЗК8	ПЗК9	ПЗК10
ЭКН	3.142	2.858	3.113	2.339	3.187	2.652	3.435	3.174	2.990	3.648
ЭКЛ	3.168	3.526	3.113	3.055	3.003	3.884	3.129	2.990	2.729	3.090
НАУ	2.819	3.526	2.448	2.713	2.452	2.419	2.703	2.635	2.990	2.342
ХУД	2.819	3.548	2.448	3.765	2.887	2.965	2.703	2.990	3.603	3.090
ИНФ	2.626	2.177	3.684	3.129	3.471	2.419	3.435	3.210	3.603	3.648
ТЕХ	3.245	2.890	3.684	3.129	2.887	3.081	3.200	3.174	3.277	2.342
МЕД	3.142	2.890	2.535	3.055	3.471	3.884	3.129	3.210	3.277	3.061
СОЦ	2.626	3.548	2.535	3.765	3.003	3.081	3.200	2.990	2.729	3.061
ПЕД	3.168	2.858	3.219	2.713	3.187	2.652	2.532	2.990	2.400	2.858
УПР	3.245	2.177	3.219	2.339	2.452	2.965	2.532	2.635	2.400	2.858

Таблица 5. Средние оценки ПЗК (ментальные нарушения)

3.3. Расчет матриц компонентных нагрузок и интерпретация результатов

На основе этих срезов в соответствие с формулой (3) для каждой оси тензора были построены матрицы развертки $X_{(1)}, X_{(2)}, X_{(3)}$ соответственно размеров $4 \times 100, 10 \times 40, 10 \times 40$. Далее по формуле (17) найдены векторы $\text{vec}(X_1), \text{vec}(X_2), \text{vec}(X_3)$ и с использованием функции `reshape` определены матрицы X_1, X_2, X_3 (второй шаг итерационного алгоритма нахождения матриц компонентных нагрузок). Затем, применяя итерационный алгоритм, для каждой категоризованной переменной были вычислены матрицы компонентных нагрузок $A_1 = A_1^{(k*)}, A_2 = A_2^{(k*)}, A_3 = A_3^{(k*)}$. Расчеты выполнены для случая, когда $n'_1 = n'_2 = n'_3 = 3$, т. е. для каждой моды в работе предложено выделить по 3 компоненты. Точность в условии остановки (19) задавалась равной 0.1. При этом в момент остановки $MSE(A_1^{(k*)}, A_1^{(k*-1)}) = 0.0201, MSE(A_2^{(k*)}, A_2^{(k*-1)}) = 0.0042, MSE(A_3^{(k*)}, A_3^{(k*-1)}) = 0.0805$.

Для удобства интерпретации полученных результатов все матрицы компонентных нагрузок были умножены на -1 . Результаты вычислений представлены в таблицах 8–10. Так же в данных таблицах представлены значения векторных норм (20)–(23) и выполнены в соответствие с рассмотренным ранее методом ранжирования градаций по каждому категоризованному признаку по степени совокупного влияния данных градаций на компоненты этого признака.

При интерпретации компонент для категоризованного признака — группы нозологии будем исходить из следующего: 1) компонента касается не столько конкретного диагноза, сколько взаимодействия разных видов нарушений в контексте ограничений по развитию ПЗК; 2) компонентные нагрузки показывают степень ограничения развития ПЗК: положительные значения — увеличение

Table 6. Average ratings of PZK (sensory impairments)

	ПЗК1	ПЗК2	ПЗК3	ПЗК4	ПЗК5	ПЗК6	ПЗК7	ПЗК8	ПЗК9	ПЗК10
ЭЖН	3.021	2.914	3.234	2.552	3.110	2.694	3.344	3.187	2.988	3.582
ЭКЛ	3.050	3.531	3.234	3.009	3.104	3.866	3.042	2.804	2.748	2.991
НАУ	2.807	3.531	2.404	2.777	2.493	2.460	2.786	2.712	2.988	2.558
ХУД	2.807	3.490	2.404	3.653	2.887	2.861	2.786	3.074	3.501	2.991
ИНФ	2.671	2.291	3.507	3.009	3.407	2.460	3.344	3.223	3.501	3.582
ТЕХ	3.451	2.774	3.507	3.009	2.887	3.119	3.252	3.187	3.258	2.558
МЕД	3.021	2.774	2.614	3.009	3.407	3.866	3.042	3.223	3.258	2.988
СОЦ	2.671	3.490	2.614	3.653	3.104	3.119	3.252	2.804	2.748	2.988
ПЕД	3.050	2.914	3.240	2.777	3.110	2.694	2.576	3.074	2.504	2.881
УПР	3.451	2.291	3.240	2.552	2.493	2.861	2.576	2.712	2.504	2.881

Таблица 6. Средние оценки ПЗК (сенсорные нарушения)**Table 7.** Average ratings of PZK (general medical conditions)

	ПЗК1	ПЗК2	ПЗК3	ПЗК4	ПЗК5	ПЗК6	ПЗК7	ПЗК8	ПЗК9	ПЗК10
ЭЖН	3.032	2.960	3.333	2.548	3.397	2.421	3.429	3.381	3.008	3.778
ЭКЛ	3.262	3.714	3.333	2.714	2.921	4.056	3.183	2.706	2.770	2.857
НАУ	2.754	3.714	2.627	2.825	2.444	2.333	2.651	2.635	3.008	2.365
ХУД	2.754	3.310	2.627	3.889	2.746	2.937	2.651	3.000	3.548	2.857
ИНФ	2.571	2.310	3.333	3.024	3.492	2.333	3.429	3.278	3.548	3.778
ТЕХ	3.381	2.706	3.333	3.024	2.746	3.254	3.405	3.381	3.341	2.365
МЕД	3.032	2.706	2.444	2.714	3.492	4.056	3.183	3.278	3.341	2.905
СОЦ	2.571	3.310	2.444	3.889	2.921	3.254	3.405	2.706	2.770	2.905
ПЕД	3.262	2.960	3.262	2.825	3.397	2.421	2.333	3.000	2.333	3.095
УПР	3.381	2.310	3.262	2.548	2.444	2.937	2.333	2.635	2.333	3.095

Таблица 7. Средние оценки ПЗК (общие заболевания)

ограничений по развитию ПЗК, отрицательные значения — уменьшение ограничений по развитию ПЗК, связанных именно со взаимодействием разных видов нарушений.

Так как все градации первой моды тензора (группы нозологии) имеют равные или почти равные значения нагрузок на компоненту №1 (по всем градациям умеренная связь с компонентой, но близка к существенной связи), то это означает, что данные категории являются взаимозаменяемыми индикаторами одной и той же латентной черты индивида. Их вклад в формирование этой компоненты абсолютно идентичен, и, заболевания, входящие в каждую из групп нозологии, в одинаковой степени увеличивают ограничения по формированию и развитию значимых качеств лиц с ОВЗ, необходимых для их профессиональной самореализации. Таким образом, компонента №1 — это класс лиц с ОВЗ и инвалидностью, которые имеют сходный уровень физических способностей и когнитивных функций несмотря на различные нарушения здоровья (*инвалиды и лица с ОВЗ с сопоставимыми возможностями*).

Категориями, определяющими компоненту №2, являются сенсорные нарушения (очень сильная связь с компонентой), которые ограничивают поток информации извне (нарушения зрения, слуха) и общие заболевания (существенная связь с компонентой) — соматические (телесные) заболевания, не связанные напрямую с органами чувств (заболевания внутренних органов). Две другие группы нозологии: нарушения опорно-двигательного аппарата (ОДА) и ментальные нарушения имеют умеренную и слабую связь с этой компонентой соответственно. При этом сенсорные нарушения являются доминирующими и увеличивающими степень ограничения развития ПЗК. С другой стороны, общие заболевания снижают ограничения по развитию значимых качеств инвалидов и лиц

Table 8. Component loading matrix (nosological groups)**Таблица 8.** Матрица компонентных нагрузок (группы нозологии)

Компонента	№1	№2	№3	I_1	Ранг I_1	I_2	Ранг I_2	I_∞	Ранг I_∞	I_4	Ранг I_4	Сумма рангов	Итого
Сингулярное число	59.977	1.583	0.819										
ДВГ	0.501	-0.233	0.830	1.564	2	0.997	1	0.830	1	0.858	1	5	1
МНТ	0.501	0.119	-0.196	0.816	4	0.551	4	0.501	4	0.504	4	16	4
СНС	0.499	0.737	-0.133	1.369	3	0.900	3	0.737	2	0.773	2	10	3
ОБЩ	0.499	-0.623	-0.505	1.626	1	0.944	2	0.623	3	0.726	3	9	2

с ОВЗ, входящих в данный кластер. Такая ситуация означает, что именно сенсорные нарушения являются первопричиной основных ограничений ПЗК в жизни человека, а общие заболевания являются фактором риска развития дефицита ПЗК. Следовательно, лечение общего заболевания является основанием для компенсации сенсорного нарушения, а значит, ведет к снижению ограничений по развитию физических способностей и когнитивных функций. Т. е. наличие общих заболеваний через компенсационную зависимость приводит у таких респондентов к снижению ограничений по развитию ПЗК. Таким образом, компонента № 2 — это класс инвалидов и лиц с ОВЗ, имеющих *соматогенные сенсорные нарушения*.

Аналогичным образом анализ нагрузок для компоненты № 3, позволяет выделить класс инвалидов и лиц с ОВЗ, у которых доминирующим (очень сильная связь с компонентой) является нарушение ОДА (трудности передвижения, поражения верхних конечностей), которые увеличивают степень ограничения развития ПЗК, а наличие общего заболевания (существенная связь с компонентой) через компенсационные механизмы приводит к снижению этих ограничений. Объяснить этот компенсационный эффект общего заболевания при нарушениях ОДА можно запуском адаптационных механизмов, когда системное заболевание заставляет человека выстраивать комплексную систему управления здоровьем, которая, как побочный положительный эффект, помогает более эффективно предотвращать и минимизировать вторичные ограничения развития, связанные с ОДА, и, как следствие, способствовать развитию значимых качеств и инвалидов и лиц с ОВЗ. Таким образом, компонента № 3 — это класс инвалидов и лиц с ОВЗ, имеющих *соматогенные нарушения ОДА*.

Ранжирование групп нозологии, позволило упорядочить их по совокупному влиянию на выделенные компоненты. Так нарушения ОДА в большей степени по сравнению с другими заболеваниями оказывают влияние на развитие у инвалидов и лиц с ОВЗ их физических способностей и когнитивных функций, что, в свою очередь, отражается на формировании ПЗК. Второе место по влиянию на развитие ПЗК оказывают общие заболевания, на третьем — сенсорные нарушения, на четвертом — ментальные нарушения (трудности мышления, трудности социальной адаптации, трудности соблюдения темпа работы).

Кроме того, из значений евклидовой нормы, оценивающей коммунальность градаций, можно отметить, что 99.7 % вариации оценок ПЗК, относящихся к инвалидам и лицам с ОВЗ, имеющими нарушения ОДА, объясняется выделенными компонентами (группы инвалидов и лиц с ОВЗ). Аналогично для инвалидов и лиц с ОВЗ, имеющих общие заболевания — 94.4 %, сенсорные нарушения — 90 %, ментальные нарушения — 55.1 %.

Как видно из таблицы 9, все категории второй моды тензора (сферы деятельности) имеют умеренную положительную связь с компонентой № 1. Тем не менее для интерпретации этой компоненты выделим три сферы деятельности, имеющие наибольшие компонентные нагрузки. Это эко-

Table 9. Component loading matrix (fields of activity)**Таблица 9.** Матрица компонентных нагрузок (сферы деятельности)

Ком- поне- нта	№1	№2	№3	l_1	Ранг l_1	l_2	Ранг l_2	l_∞	Ранг l_∞	l_4	Ранг l_4	Сум- ма ран- гов	И т о г
Син- гуля- рное число	60.061	3.372	2.313										
ЭКН	0.325	-0.327	0.257	0.910	5	0.528	5	0.327	7	0.406	6	23	5.5
ЭКЛ	0.329	0.522	-0.223	1.075	2	0.657	3	0.522	3	0.546	3	11	3
НАУ	0.292	0.240	0.431	0.963	4	0.573	4	0.431	4	0.461	4	16	4
ХУД	0.323	-0.039	0.080	0.442	10	0.335	10	0.323	9	0.323	9	38	10
ИНФ	0.327	-0.713	-0.087	1.127	1	0.789	1	0.713	1	0.720	1	4	1
ТЕХ	0.323	0.111	-0.042	0.476	9	0.344	9	0.323	8	0.324	8	34	8
МЕД	0.327	-0.069	-0.604	1.000	3	0.690	2	0.604	2	0.616	2	9	2
СОЦ	0.319	0.148	-0.339	0.806	6	0.488	7	0.339	6	0.394	7	26	7
ПЕД	0.304	0.077	0.381	0.762	7	0.493	6	0.381	5	0.415	5	23	5.5
УПР	0.290	0.091	0.241	0.622	8	0.388	8	0.290	10	0.320	10	36	9

логическая сфера, IT-сфера, медицинская и спортивно-оздоровительная сфера. Тогда компонента № 1 представляет собой такую сферу деятельности, в которой для профессиональной самореализации инвалидам и лицам с ОВЗ необходимы ПЗК, являющиеся ключевыми для выделенных выше сфер деятельности. Такой компонентой № 1 может быть *персонализированная экологическая медицина и wellness-аналитика*. Эта сфера деятельности предполагает создание персональных цифровых рекомендаций по укреплению здоровья, физической активности и профилактике заболеваний, на основе данных о состоянии окружающей среды и индивидуальных показателях здоровья человека (например, мобильного приложения «Персональный экологический врач» для горожан, спортсменов, людей с хроническими заболеваниями или корпоративных wellness-платформ для компаний).

Для компоненты № 2 доминирующей категорией является IT-сфера (очень сильная связь с компонентой), а также на эту компоненту определяющее ее смысл оказывает влияние и экологическая сфера. Причем это должна быть такая сфера деятельности, в которой для успешной профессиональной самореализации инвалидам и лицам с ОВЗ нужны профессиональные качества, значимые для самореализации в экологической сфере, но значимые качества для самореализации в IT-сфере должны быть как можно меньше выражены. Такой компонентой № 2 может быть *практическая аналоговая эко-активность*. Это направление деятельности предполагает осознанный отказ от цифровых технологий там, где это экологически и социально оправдано, и замена их аналоговыми или природными решениями для снижения совокупного экологического ущерба и укрепления реального благополучия (например, консультирование и создание решений для жизни с минимальным углеродным следом, основанных на отказе от энергоемких цифровых привычек или организация программ в заповедных зонах, где полный или частичный отказ от цифровых устройств является условием для глубокого экологического просвещения и волонтерства).

Аналогичный анализ нагрузок для компоненты № 3 содержательно определяет эту компоненту как сферу деятельности, качествами которой для успешной профессиональной самореализации инвалидов и лиц с ОВЗ должны оставаться ПЗК, необходимые для научной сферы (умеренная положительная связь с компонентой). Но при этом значимые качества для медицинской и спортивно-оздоровительной сферы должны быть слабо выражены (существенная отрицательная связь с компонентой), что не позволяет респонденту успешно самореализоваться в этой сфере. Поэтому тако-

Table 10. Component loading matrix (PZK Group)**Таблица 10.** Матрица компонентных нагрузок (группы ПЗК)

Ком- поне- нта	№1	№2	№3	I_1	Ранг I_1	I_2	Ранг I_2	I_∞	Ранг I_∞	I_4	Ранг I_4	Сум- ма ран- гов	И т о г
Син- гуля- рное число	60.005	2.282	2.028										
ПЗК1	0.316	-0.296	0.105	0.716	7	0.445	7	0.316	9	0.365	7	30	8
ПЗК2	0.317	-0.183	-0.309	0.808	5	0.478	5	0.317	6	0.377	5	21	5
ПЗК3	0.316	-0.583	-0.230	1.129	2	0.702	2	0.583	2	0.599	2	8	2
ПЗК4	0.316	-0.067	0.530	0.913	4	0.621	3	0.530	3	0.546	3	13	3
ПЗК5	0.316	0.008	0.016	0.340	10	0.317	10	0.316	8	0.316	10	38	10
ПЗК6	0.317	0.500	-0.601	1.418	1	0.843	1	0.601	1	0.671	1	4	1
ПЗК7	0.317	0.267	-0.036	0.620	8	0.416	8	0.317	5	0.351	8	29	7
ПЗК8	0.316	0.144	0.295	0.756	6	0.456	6	0.316	7	0.366	6	25	6
ПЗК9	0.317	0.395	0.321	1.032	3	0.599	4	0.395	4	0.460	4	15	4
ПЗК10	0.316	-0.189	-0.091	0.595	9	0.379	9	0.316	10	0.326	9	37	9

му индивиду можно рекомендовать научную сферу, исследования которой относятся к медицине или спортивно-оздоровительной деятельности. Такой компонентой № 3 может быть биомедицина — это фундаментальная дисциплина, которая применяет принципы естественных наук (биологии, химии, физики) к клинической медицине.

Согласно проведенному ранжированию сфер деятельности можно отметить, что наиболее значимыми с точки зрения их совокупного влияния на выделенные компоненты являются ИТ-сфера (1 место), медицинская и спортивно-оздоровительная сфера (2 место), экологическая сфера (3 место), научная сфера (4 место). Поэтому можно считать, что ПЗК этих сфер деятельности, являются наиболее востребованными для успешной профессиональной самореализации инвалидов и лиц с ОВЗ.

Те же сферы деятельности и в том же порядке располагаются по величине их коммуналности: 78.9 % вариации оценок ПЗК, которые являются ключевыми для ИТ-сферы, объясняется выделенными компонентами, для медицинская и спортивно-оздоровительная сфера — 69 %, для экологической сферы — 65.7 %, для научной сферы — 57.3 %.

Из таблицы 10 видно, что все градации третьей моды тензора (группы ПЗК) имеют равные или почти равные значения нагрузок на компоненту № 1 (по всем градациям умеренная связь с компонентой). Это означает, что данные категории являются взаимозаменяемыми индикаторами одной и той же латентной черты индивида, т. е. развитие любого качества из данных групп ПЗК положительно коррелирует с развитием этой латентной характеристики. Такой компонентой № 1 является высший уровень управления познавательной деятельностью — исполнительные функции. Исполнительные функции — это набор сложных когнитивных процессов, которые отвечают за планирование, принятие решений, контроль импульсов, переключение между задачами и регуляцию поведения.

Для компоненты № 2 наибольшую корреляцию с этой компонентой демонстрируют качества из группы ПЗК3 — умения (существенная отрицательная связь с компонентой) и качества из группы ПЗК6 — память (умеренная, но близка к существенной, связь с компонентой). Следовательно, интерпретировать компоненту № 2 как *ригидность мышления (или низкая когнитивная гибкость)*. Ригидный ум склонен к созданию жестких схем, шаблонов и повторяющихся процедур. В такой

Table 11. Tensor components

Таблица 11. Компоненты тензора

Мода тензора	Компонента № 1	Компонента № 2	Компонента № 3	Категория с наибольшим совокупным влиянием	Категория с наибольшим значением комму-нальности
Группы нозологии	Инвалиды и лица с ОВЗ с сопоставимыми возможностями	Соматогенные сенсорные нарушения	Соматогенные нарушения ОДА	Нарушения ОДА	Нарушения ОДА
Сферы деятельности	Персонализированная экологическая медицина и wellness-аналитика	Практическая аналоговая эко-активность	Биомедицина	IT-сфера	IT-сфера
Группы ПЗК	Исполнительные функции	Ригидность мышления	Высокая когнитивная гибкость	Память	Память

стабильной и предсказуемой «ментальной среде» процесс запоминания может быть очень эффективным. Но при этом ригидность не позволяет адаптировать навыки под новую задачу, среду или инструмент, т. е. ригидность серьезно мешает формированию новых, особенно сложных, умений.

Для компоненты № 3 наибольшую корреляцию с этой компонентой имеют качества из группы ПЗК4 — способности (существенная положительная связь с компонентой) и качества из группы ПЗК6 — память (существенная отрицательная связь с компонентой). В этом случае компоненту № 3 можно определить как *высокая когнитивная гибкость*. Чрезмерная гибкость может нарушать процессы консолидации и извлечения воспоминаний, что мешает развитию памяти. Но при этом именно гибкость лежит в основе формирования и применения сложных, адаптивных способностей.

Ранжирование групп ПЗК по степени их совокупного влияния на описанные выше компоненты позволяет выделить следующие наиболее значимые для профессиональной самореализации инвалидов и лиц с ОВЗ группы качеств: память (1 место), умения (2 место), способности (3 место), взаимодействие с другими людьми (4 место), личностные характеристики (5 место).

Те же группы ПЗК и в том же порядке располагаются по величине их коммунальности: 84.3 % вариации оценок ПЗК, характеризующих память индивида, объясняется выделенными компонентами, для умений — 70.2 %, для способностей — 62.1 %, для качеств, характеризующих взаимодействие с другими людьми, — 59.9 %, для личностных характеристик — 47.8 %.

Сводные результаты интерпретации матриц компонентных нагрузок представлены в таблице 11.

По формулам (10) и (12) было найдено ядро тензора (тензор $3 \times 3 \times 3$). При этом строки — это 1 мода, столбцы — 2 мода, блоки — 3 мода.

$$\underline{\underline{Z}} = \begin{pmatrix} 59.948 & -0.809 & 2.105 \\ -0.039 & 0.101 & 0.517 \\ -0.080 & -0.264 & 0.788 \end{pmatrix} \begin{pmatrix} 0.88 & -0.139 & 0.696 \\ 0.217 & 0.130 & 0.308 \\ -0.059 & -1.046 & -0.269 \end{pmatrix} \begin{pmatrix} -0.161 & -0.098 & -0.275 \\ -0.025 & 0.017 & -0.227 \\ -0.120 & -0.211 & 0.485 \end{pmatrix}.$$

Проинтерпретируем три самых больших по абсолютной величине элемента ядра тензора, поскольку они несут основную смысловую нагрузку. Зависимость компонент, образующих комби-

нацию (*Инвалиды и лица с ОВЗ с сопоставимыми возможностями; Персонализированная экологическая медицина и wellness-аналитика; Исполнительные функции*), соответствующую элементу 59.948, можно определить как *сферу персонализированной экологической медицины, создающей уникальные, адаптированные рабочие места, которые одновременно являются и профессиональной деятельностью, и средой для развития исполнительных функций у инвалидов и лиц с ОВЗ, имеющих сопоставимые возможности*. В свою очередь, эти сотрудники, обладая уникальной личной и компенсаторной экспертизой, повышают качество и инклюзивность продуктов и сервисов в этой сфере, делая ее более успешной и создавая дополнительные рабочие места. В результате такую сферу деятельности можно рассматривать как самоподдерживающуюся профессиональную экосистему, где ограничения превращаются в конкурентные преимущества, а работа — в инструмент развития и интеграции.

Взаимодействие компонент, входящих в комбинацию (*Инвалиды и лица с ОВЗ с сопоставимыми возможностями; Биомедицина; Исполнительные функции*), соответствующую элементу ядра тензора 2.105, можно представить в виде *модели экспертно-ориентированной биомедицины*, где личный опыт и уникальные компетенции, сформированные инвалидностью, становятся драйвером инноваций в биомедицине. Т.е. инвалиды и лица с ОВЗ, чьи уникально развитые исполнительные функции делают их исключительно ценными сотрудниками в сфере биомедицины и обеспечивают эту сферу «взглядом изнутри», тем самым ускоряя перевод фундаментальных исследований в практические решения, которые работают в реальной жизни.

Взаимодействие компонент, входящих в комбинацию (*Соматогенные нарушения ОДА; Практическая аналоговая эко-активность; Ригидность мышления*), соответствующую элементу ядра тензора –1.046, можно описать в виде *модели целенаправленного упрощения*, когда отрицательная зависимость трансформируется в профессиональную нишу, где ограничения по здоровью становятся конкурентными преимуществами. В сфере практической аналоговой эко-активности соматогенные нарушения ОДА и ригидность мышления перестают быть медицинскими ограничениями и становятся критически важными профессиональными квалификациями. Работа такого сотрудника характеризуется минимальным воздействием на природу, высочайшей методической точностью и безупречной долгосрочной последовательностью. Данная сфера деятельности (определению медленной, низкотехнологичная, циклическая) создает профессиональную нишу, где эти ограничения являются конкурентными преимуществами.

Остальные элементы (их значения близки к нулю или практически равны нулю) показывают либо слабую связь, либо несущественную связь между компонентами различных мод тензора, т.е. вклад в модель соответствующих комбинаций компонент тензора незначителен, либо соответствующие комбинации компонент не взаимодействуют.

Данные таблиц 8–10 показывают, что для всех матриц компонентных нагрузок сингулярное число, соответствующее первой компоненте, значительно больше двух других. Причем нагрузки на эту компоненту равные или почти равные для всех категорий. Это можно объяснить тем, что исходные данные (таблицы 4–7), рассчитанные как средние оценки ПЗК инвалидов и лиц с ОВЗ с различными нарушениями здоровья (с различной нозологией), мало отличаются по своим значениям. Небольшое различие между данными подтверждает метрика, основанная на RSMD, вычисленная для средних ПЗК по каждой паре таблиц 4–7 (по каждой паре видов нозологии). Результаты вычислений представлены в таблице 12.

Такие оценки ПЗК, мало зависящие от нозологии, связаны с тем, что участники опроса при самооценке выраженности у них ПЗК скорее всего не учитывают свои ограничения по здоровью. Поэтому для получения более содержательных результатов необходимо иметь оценки ПЗК инвалидов и лиц с ОВЗ, учитывающие в том числе и нозологию.

Предложенный многомерный компонентный анализ ПЗК дает возможность находить закономерности в «гиперпространстве» данных, где какая-то закономерность (паттерн) может быть

Table 12. RSMD for mean PZK between nosology groups

Группы нозологии	ДВГ	МНТ	СНС	ОБЩ
ДВГ	—	0.284	0.246	0.232
МНТ	0.284	—	0.094	0.154
СНС	0.246	0.094	—	0.146
ОБЩ	0.232	0.154	0.146	—

Таблица 12. RSMD для средних ПЗК между группами нозологии

скрыта в совместной вариации вдоль всех осей (категоризованных переменных) одновременно. Такая информация позволит разрабатывать траектории цифрового сопровождения лиц с инвалидностью и лиц с ОБЗ с учетом их вариативных характеристик, что является одной из основных задач информационно-аналитического регионального портала «ПЕРСПЕКТИВА-PRO».

Для реализации предложенного метода анализа категоризованных признаков для трехмерного тензора был разработан код на языке Python 3. В программе были использованы следующие библиотеки: Pandas — для обработки табличных данных, NumPy — для проведения матричных вычислений и Stats из SciPy — для ранжирования данных.

Заключение

В настоящей статье были изучены существующие подходы к анализу зависимостей между категоризованными переменными и выявлены проблемы при применении этих методов для многомерных категоризованных данных (тензоров). Поэтому в работе для исследования зависимостей между категоризованными переменными представлен подход с использованием многомерного компонентного анализа. Для реализации алгоритма многомерного компонентного анализа предлагается использовать матрицы развертки тензора, полученные для каждой его оси (категоризованного признака). Поэтому удобно формировать и хранить исходные данные не в виде тензора, а в виде матриц развертки.

Данный метод позволяет построить интегральные характеристики (компоненты) по элементам исходного тензора, сформировать матрицы компонентных нагрузок и рассчитать ядро тензора, имеющего меньшее число градаций категоризованных признаков (осей тензора). Тогда элементы исходного тензора могут быть аппроксимированы наилучшим образом с помощью линейных комбинаций компонент ядра, а коэффициентами этих линейных комбинаций будут элементы матрицы, представляющей собой произведение Кронекера матриц компонентных нагрузок. В статье предложен метод ранжирования градаций категоризованных переменных по степени совокупного влияния на них компонентных нагрузок, основанный на вычислении четырех векторных норм: манхэттенского расстояния, евклидовой нормы, l_∞ -нормы, нормы Гельдера с параметром $p = 4$. Для нахождения матриц компонентных нагрузок описан итерационный алгоритм их вычисления. Указан критерий остановки данного итерационного алгоритма.

Предложенный метод применения многомерного компонентного анализа для исследования зависимостей между категоризованными переменными продемонстрирован на примере трехмерного тензора с формой (4;10;10) и категоризованными признаками: группы нозологии, сферы деятельности, группы профессионально значимых качеств. Для данного примера итерационным методом, описанным в работе, вычислены матрицы компонентных нагрузок, когда выделяются 3 компоненты по каждой категоризованной переменной. Даны интерпретации компонент для каждой моды тензора, а также выделены многомерные зависимости между компонентами категоризованных признаков, которые соответствуют значимым элементам ядра тензора. Выполнено ранжирование градаций для каждой моды по совокупному влиянию градаций на компоненты мод и выделены наиболее влиятельные градации по каждому категоризованному признаку.

Предложенный в статье метод изучения зависимостей между многомерными категоризованными переменными предполагается включить как аналитический инструмент информационно-аналитического регионального портала «ПЕРСПЕКТИВА-PRO». Результаты исследований взаимосвязей между категоризованными переменными с применением данного инструментария будут использованы для разработки траекторий цифрового сопровождения лиц с инвалидностью и лиц с ОВЗ с учетом их личностных и вариативных характеристик, таких как регион проживания, социальный статус, возраст, состояние здоровья, уровень образования, сфера деятельности, профессионально значимые качества и т. п.

Благодарности

Автор благодарит анонимных рецензентов, ознакомившихся со статьей и сделавших ценные замечания, позволившие улучшить ее качество.

References

- [1] O. A. Denisova, *Tsifrovye resursy professionalizatsii molodykh invalidov*. Cherepovets State University, 2021, in Russian.
- [2] A. A. Banin and O. Y. Lyaginova, "Tensor data model for the development of trajectories for accompanying physically challenged people and people with disabilities due to their variable characteristics", *Cherepovets State University Bulletin*, no. 1, pp. 7–25, 2024, in Russian. doi: [10.23859/1994-0637-2024-1-118-17](https://doi.org/10.23859/1994-0637-2024-1-118-17).
- [3] A. Demianova and A. Lukiyanova, "The impact of disability status on labor supply in Russia", *Applied Econometrics*, vol. 44, pp. 50–74, 2016.
- [4] D. Fantazzini, M. Shackleina, and N. Yuras, "Big Data for computing social well-being indices of the Russian population", *Applied Econometrics*, vol. 50, pp. 43–66, 2018.
- [5] E. Kossova and M. Kosorukova, "Estimation of the treatment effect of higher education on health: Comparison of the multivariate recursive probit model and matching", *Applied Econometrics*, vol. 69, pp. 65–90, 2023.
- [6] A. Dubnovitskaya and K. Furmanov, "Job satisfaction in Russia: Wages, working conditions and promotion opportunities", *Applied Econometrics*, vol. 72, pp. 121–139, 2023.
- [7] S. Ayvazyan and V. Mkhitarian, *Prikladnaya statistika i osnovy ekonometriki. Uchebnik dlya vuzov*. YuNITI, 1998, in Russian.
- [8] P. Popova and A. Rotmistrov, "Regression with categorical predictors: Criticizing dummy-variable usage and log-linear analysis as an alternative approach", *Sociological Journal*, vol. 22, no. 3, pp. 8–31, 2016, in Russian. doi: [10.19181/socjour.2016.22.3.4583](https://doi.org/10.19181/socjour.2016.22.3.4583).
- [9] V. Bure, O. Glivinskaya, and A. Sotnikov, "Log-linear database analysis by cardiac infarction data among young and middle age patients", *Vestnik of Saint Petersburg University*, vol. 10, no. 1, pp. 35–41, 2010, in Russian.
- [10] O. Dolinina and R. Karimov, "Methods of presentation and joint processing of multidimensional data for objects of numeric and non-numeric origin", *Vestnik of Saratov State Technical University*, no. 1, pp. 100–109, 2006, in Russian.
- [11] E. Fomina, "The possibility of log-linear analysis for survey results processing", *PNRPU Sociology and Economics Bulletin*, no. 3, pp. 197–211, 2018, in Russian. doi: [10.15593/2224-9354/2018.3.16?!303.621.3-047.44](https://doi.org/10.15593/2224-9354/2018.3.16?!303.621.3-047.44).
- [12] E. Fomina, "Methods of analysis of categorical variables in sociological research", *Proceedings of Higher Educational Institutions. Sociology. Economy. Politics*, vol. 12, pp. 39–51, 2019, in Russian.

- [13] A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, and D. P. Mandic, “Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions”, *Foundations and Trends® in Machine Learning*, vol. 9, no. 4–5, pp. 249–429, 2016. DOI: [10.1561 / 22000000059](https://doi.org/10.1561/22000000059).
- [14] Y. Magnus and K. Neydekker, *Matrix differential calculation with applications in statistics and econometrics*, 3rd edition. Wiley, 2019.
- [15] G. Golub and V. Loan, *Matrix Computations*, 4th edition. JHU press, 2013.
- [16] M. Linting, J. J. Meulman, P. J. F. Groenen, and A. J. van der Koojj, “Nonlinear principal components analysis: Introduction and application”, *Psychological Methods*, vol. 12, no. 3, pp. 336–358, 2007. DOI: [10.1037/1082-989X.12.3.336](https://doi.org/10.1037/1082-989X.12.3.336).