УДК 004.622

РАЗРАБОТКА ГИБРИДНОГО МЕТОДА ФАКТОРНОГО АНАЛИЗА ДЛЯ НЕПОЛНОГО НАБОРА МЕДИЦИНСКИХ ДАННЫХ

О. А. Попова, преподаватель кафедры ТГМУ, Тюмень, Россия;
orcid.org/ 0009-0006-3530-5703, e-mail: popovaOA@tyumsmu.ru

В работе представлен сравнительный анализ эффективности работы известных методов факторного анализа: PLS, FastICA, BFA и MLFA, а также вновь разработанных гибридных методов PLS-NN и PLS-RF. Основной задачей исследования было выявление методов, обеспечивающих наилучшую точность объяснения дисперсии в целевой переменной и наилучшее соответствие модели данным. Результаты показали, что метод PLS объясняет значительную долю дисперсии в целевой переменной, демонстрируя хорошие показатели подгонки модели, однако наблюдались признаки избыточной сложности модели. Метод FastICA продемонстрировал высокую объясняющую способность, но выявлены потенциальные проблемы переобучения. Методы BFA и MLFA показали неудовлетворительные результаты, характеризующиеся отрицательной прогностической производительностью и неудовлетворительными значениями показателей подгонки модели. По результатам исследования был выбран метод PLS для дальнейшего усовершенствования и подгонки. С целью повышения его эффективности была применена гибридизация, что позволило значительно улучшить качество модели и ее соответствие данным. Анализ результатов работы гибридных методов факторного анализа (PLS-NN и PLS-RF) показал, что оба метода обладают высокой способностью объяснять вариацию в исходных данных. Однако метод PLS-NN превзошел метод PLS-RF по ряду показателей, таких как коэффициент детерминации, информационные критерии AIC и BIC, а также показатели RMSEA и SRMR, что свидетельствует о лучшей подгонке модели к данным и меньшем уровне ошибки аппроксимации. Таким образом, исследование подтверждает, что метод PLS-NN является предпочтительным для использования в условиях рассматриваемого набора данных благодаря своей точности, объясняющей способности и качеству подгонки модели.

Ключевые слова: *медицинские данные, методы факторного анализа, PLS, FastICA, BFA, MLFA, гибридные методы факторного анализа, Random Forest, Neural Networks, коэффициент детерминации, информационные критерии, предобработка входных данных.*

DOI: 10.21667/1995-4565-2025-91-87-103

Введение

В современном мире экспертные системы играют все более значимую роль в автоматизации рутинных задач, связанных с обработкой больших данных, прогнозированием результатов и обучением моделей искусственного интеллекта. Проблема проектирования экспертных систем приобрела особую актуальность в связи с указом президента Российской Федерации от 10.10.2019 г. № 490, который ставит целью скорейшее развитие и внедрение искусственного интеллекта на всей территории РФ: «Повышение качества услуг в сфере здравоохранении (включая профилактические обследования, диагностику, основанную на анализе изображений, прогнозирование возникновения и развития заболеваний, подбор оптимальных дозировок лекарственных препаратов, сокращение угроз пандемий, автоматизацию и точность хирургических вмешательств)» [**Ошибка! Источник ссылки не найден.**]. Таким образом, проектирование экспертных систем является приоритетным направлением и продолжает интенсивно развиваться благодаря появлению новых алгоритмов проектирования, совершенствованию методов машинного обучения и синергетическим связям со смежными областями искусственного интеллекта, такими как компьютерное зрение, обработка естественного языка и робототехника [2]. В процессе построения экспертных систем, в частности систем поддержки принятия решений (СППР), предобработка и анализ входных данных игра-

ют решающую роль. Так как предобработка позволяет очистить данные от ошибок, пропусков и шума, она тем самым повышает точность предсказаний СППР. Предобработка данных помимо «стандартной очистки» позволяет объединить разрозненные данные из различных источников, таких как лабораторные анализы, анкетные данные и результаты инструментальных исследований, и выявить скрытые взаимосвязи между переменными, что может не быть очевидными при первоначальном анализе. Это особенно важно в медицине, где взаимосвязи между различными показателями могут быть сложными и неочевидными, а от качества и точности интерпретации результатов зависят здоровье и жизнь пациентах [3].

Теоретические исследования

Одним из наиболее эффективных методов предобработки входных медицинских данных, в условиях разрозненности, мультиколлинеарности и многообразия признаков является факторный анализ. Факторный анализ позволяет выделить основные компоненты, которые определяют структуру данных, выявить скрытые взаимосвязи между признаками, уменьшить размерность данных, устранить шум и ошибки. При интерпретации медицинских данных факторный анализ может помочь в выявлении скрытых закономерностей и взаимосвязей между различными клиническими показателями, что может привести к более глубокому пониманию механизмов развития заболеваний и улучшению качества медицинской помощи. В результате число переменных, участвующих в исследовании, может быть существенно ограничено.

Основные идеи факторного анализа были заложены английским психологом и антропологом Ф. Гальтоном [4]. Факторный анализ в большинстве источников описывается как раздел многомерного статистического анализа, объединяющий методы оценки размерности множества наблюдаемых переменных посредством исследования структуры корреляционных матриц. При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов [5, 6]. Основная цель факторного анализа – выявить набор основных факторов, которые могут объяснить большинство вариаций в данных. Факторный анализ является важнейшим инструментом в работе по предобработке данных для разработки СППР (систем поддержки принятия решений). Определение важных факторов в наборе данных позволяет выявить ключевые переменные, которые оказывают наибольшее влияние на результаты, и упростить структуру данных, что облегчает их анализ и интерпретацию [7]. Для анализа мультиколлинеарных, разнородных комплексных медицинских данных (КМД, комплексный набор данных по пациенту, состоящий из физиологических, анатомических и биохимических показателей здоровья) более предпочтителен стохастический вид факторного анализа, так как он учитывает случайную природу данных и может выделить основные факторы, влияющие на вариабельность данных. В отличие от стохастического факторного анализа детерминированный факторный анализ может столкнуться с проблемами вычислительной неустойчивости и невозможности определить уникальное решение. Также при работе с разными типами данных **с**тохастический факторный анализ может быть более гибким и адаптируемым к разнородным данным, поскольку он может использовать различные распределения и модели для различных типов данных. Стохастический факторный анализ также эффективен в условиях, когда данные содержат ошибки или шум, и может выделить основные факторы, влияющие на изменения в данных. Это позволяет оценить их вклад в общую вариабельность данных и получить более точное представление о структуре данных.

В целом стохастический факторный анализ является мощным инструментом для анализа комплексных медицинских данных, который может помочь выявить основные факторы, влияющие на вариабельность данных, и оценить их вклад в общую дисперсию данных [8-10]. Существует достаточно много работ, посвященных факторному анализу, его применению к разным видам данных, а также разработке новых и улучшенных версий этого метода. Так,

авторы работы [11] предлагают использовать глубокое обучение (глубокий факторный анализ, DFA). Модель представлена с использованием иерархической конструкции сверточного факторного анализа с разреженными факторными нагрузками и оценками. Вычисление параметров модели, зависящих от слоя, реализовано в байесовской настройке с использованием сэмплера Гиббса и вариационного байесовского (VB) анализа, которые явно используют сверточную природу расширения. В работе [12] авторы предлагают вариационный байесовский алгоритм, который аппроксимирует апостериорное распределение за долю вычислительных затрат MCMC. Авторы работ [13,14] анализируют два различных мультипликативных алгоритма для NMF (неотрицательной матрицы факторизации). Они отличаются лишь незначительно мультипликативным коэффициентом, используемым в правилах обновления. Один алгоритм может быть показан для минимизации обычной ошибки наименьших квадратов, в то время как другой минимизирует обобщенную дивергенцию Кульбака – Лейблера. Монотонность сходимости обоих алгоритмов может быть доказана с помощью вспомогательной функции, аналогичной той, которая используется для доказательства сходимости алгоритма ExpectationMaximization. Алгоритмы также могут быть интерпретированы как градиентный спуск с диагональным масштабированием, где коэффициент масштабирования оптимален выбран для обеспечения конвергенции.

В трудах [15] авторы предлагают использовать новый оценщик, основанный на спектральных свойствах матрицы корреляции выборок Спирмена в многомерной настройке, где размерность и размер выборки пропорционально стремятся к бесконечности. Авторы работы [16] используют новый подход к факторному анализу, с помощью унифицированной байесовской структуры. Их подход использует промежуточные вращения факторов на протяжении всего процесса обучения, что значительно повышает эффективность априорных значений, вызывающих разреженность. Эти автоматические вращения к разреженности встроены в алгоритм PXL-EM, байесовский вариант расширенного по параметрам EM для обнаружения апостериорного режима. Итеируя между мягким порогом малых факторных нагрузок и преобразованиями факторного базиса, получают резкое ускорение, устойчивость к плохим инициализациям и лучше ориентированные разреженные решения. Авторы работы [17] предлагают канонический корреляционный анализ на основе декомпозиции для многомерных наборов данных (D-CCA), это новый метод разложения, который определяет общие и отличительные матрицы из $\ell 2$ пространства случайных величин, а не традиционно используемое евклидово пространство, с тщательным построением ортогональных отношений между отличительными матрицами. Показано, что предлагаемые оценщики общих и отличительных матриц являются последовательными и имеют лучшую производительность, чем некоторые современные методы как в смоделированных данных, так и в реальном анализе. В работе [18] авторы предлагают использовать факторный анализ как процедуру вменения, которая использует факторы, оцененные из высокого блока, вместе с повторно повернутыми нагрузками, оцененными из широкого блока, для вменения пропущенных значений в панели данных. Предполагается, что сильная факторная структура сохраняется для полной панели данных и ее подблоков и общий компонент может быть последовательно оценен с четырьмя различными скоростями сходимости без необходимости регуляризации или итерации. Автор работы [19] предлагает алгоритм проведения факторного анализа на базе автокорреляционной нейронной сети. Нейронная сеть данного типа обладает способностью автокорреляции входного и выходного сигнала. В этой нейронной сети для осуществления обратного распространения ошибки от максимума корреляции входного и выходного сигналов добавляется дополнительный слой нейронов с весовыми коэффициентами равными значениям входного сигнала. Максимизация корреляции входного и выходного сигналов приводит к реализации факторного анализа и вычисления главных компонент в случае меньшего числа нейронов на выходном слое, чем на входном слое нейронной сети. Этот же автор в работе [20] предлагает новый подход к проведению факторного анализа на базе метода кластеризации данных *k*-средних и последующего факторного вращения. Факторный анализ выделяет из множества

исходных показателей *k* главных компонент или факторов, с наибольшей точностью аппроксимирующих разброс и распределение исходных данных. Такие главные компоненты формируют факторную структуру исходных данных. В качестве направлений и положений главных компонент могут быть использованы различные характеристические точки исходной структуры данных. В данной работе предлагается использовать центры кластеров исходных данных. В результате метод *k*-средних позволяет найти факторную структуру в исходном многомерном пространстве данных из положений *k* центров выделенных кластеров. Последующее факторное вращение по оригинальному критерию интерпретируемости позволяет найти простую факторную структуру.

Постановка задачи

Комплексные клинические медицинские данные (ККМД) часто представляют собой разнотипный массив данных с высокой размерностью, в котором особое значение имеет корреляция между переменными, что делает анализ данных трудоемким и затрудняет интерпретацию результатов. Таким образом, в условиях разрозненности, мультиколлинеарности и многообразия признаков медицинских данных возникает *проблема* в необходимости адаптации существующих методов или в разработке новых методов предобработки входных данных. Но несмотря на огромное многообразие работ, посвященных факторному анализу данных, так и не найдены методы, адаптированные к специфике комплексных медицинских данных небольшого объема. Так, при анализе работы методов предобработки данных каждый метод специфично работал только с определенным качественным составом медицинских данных и не обеспечивал универсального решения для комплексных задач, возникающих в практике участкового врача. Например, методы, основанные на линейной регрессии, могут не учитывать нелинейные зависимости между переменными, а алгоритмы кластеризации могут оказаться неэффективными в условиях высокой размерности, большого числа выбросов и разнотипного состава данных.

Таким образом, для успешного подбора методов предобработки для ККМД необходимо учитывать специфику набора медицинских данных, таких как: неравномерное распределение, разрозненность, разнородность, разнотипность и мультиколлинеарность данных, наличие скрытых факторов и корреляции между признаками. Поэтому следует разрабатывать гибридные модели, которые соединяют традиционные статистические методы и современные алгоритмы машинного обучения. Такие модели могут быть более устойчивыми и точными, предоставляя дополнительные инструменты для интерпретации комплексных данных. Таким образом, решение *проблемы* адаптации методов факторного анализа медицинских данных ставит *цель* работы: разработку гибридного метода факторного анализа для небольшого набора комплексных клинических медицинских данных с целью реализации в модуле СППР. Интеграция таких методов в СППР может значительно улучшить качество и своевременность оказания медицинской помощи, повысить уровень диагностики и индивидуализировать подход к каждому пациенту.

Задача работы состоит в изучении, анализе, апробировании существующих методов факторного анализа, сравнении и выборе наиболее эффективного метода, дальнейшей его адаптации путем разработки нового гибридного метода факторного анализа под имеющийся качественный состав небольшого набора ККМД. Исходными данными является реальный датасет (реестр) с результатами медицинских исследований 104 пациентов с подтвержденным диагнозом неалкогольной жировой болезни печени (НАЖБП) на конкретной стадии заболевания. Реестр содержит информацию об анатомических, физиологических, биохимических и аппаратных методах исследования, проведенных в ходе диагностики печени. Реестр медицинских данных представлен в виде файла Excel (.xlsx) и содержит 21 параметр: «Sex», «Age», «Diagnosis», «F», «S», «Height», «Weight», «BMI», «Waist circumference», «Gamma glutamyl transferase (GGT)», «Triglycerides», «High density lipoprotein cholesterol», «Low-density lipoprotein cholesterol», «Alanine aminotransferase (ALT)», «Alkaline phosphatase», «To-

tal protein», «Total cholestero», «Aspartate aminotransferase (AST)», «Albumen», «Total bilirubin», «Glucose». Данный датасет был проверен с помощью методов оценки пригодности выборки для проведения факторного анализа KMO (Kaiser-Meyer-Olkin) и теста Бартлетта (для проверки гипотезы корреляции между переменными, чтобы оправдать использование факторного анализа). По результатам проверки были получены следующие значения: KMO = 0,619, что означает, что данные приемлемы для факторного анализа и находятся на нижней границе; тест Барлетта p-value <0,05 указывает на то, что переменные значимо коррелируют и данные подходят для факторного анализа. Значение показателя суммарной объясненной дисперсии составил – 58,43 % что является приемлемым для оценки репрезентативности и выявления основных факторов. Однако выборка проходит по нижней границе условно принятых рекомендаций 5-10 наблюдений на признак (104 наблюдения на 21 признак), и такой объем, хотя и позволяет провести факторный анализ и получить предварительные результаты, но все же накладывает определенные ограничения на исследование. Несмотря на то, что данная выборка может быть достаточной для проведения предварительного анализа или получения предварительных выводов, для более надежных и всесторонних выводов требуется больший объем данных. Более крупная выборка с большим количеством наблюдений и признаков поможет лучше представить популяцию и минимизировать ошибки и искажения в результатах. Таким образом, хотя текущая выборка является репрезентативной в определенных пределах, но для повышения надежности и валидности результатов рекомендуется проводить дальнейшие исследования с более обширными данными. На данном этапе выводы рассматриваются как предварительные и рекомендуется интерпретировать их с учетом ограниченного объема данных.

В данном исследовании проводится комплексный анализ методов факторного анализа, применимых к небольшому набору комплексных клинических медицинских данных, с целью определения наиболее эффективного подхода и разработки гибридного метода, адаптированного к специфическим особенностям имеющегося набора данных. Для анализа методов поиска скрытых факторов были выбраны методы стохастического и мультивариантного факторного анализа: метод многоуровневого факторного анализа *(MLFA),* метод независимых компонент *(ICA),* метод байесовского факторного анализа *(BFA)*, метод частичных наименьших квадратов *(Partial Least Squares, PLS).*

Методология

Этапы методологии исследования включают в себя несколько важных шагов: описание и анализ входных данных, предобработка входных комплексных клинических медицинских данных алгоритмами машинного обучения (очистка, нормализация и стандартизация данных), изучение и анализ существующих методов факторного анализа, разработка гибридного метода факторного анализа, который объединяет сильные стороны различных подходов для создания более эффективного метода. Этап разработки включает тестирование исследуемых методов факторного анализа на имеющемся наборе медицинских данных, анализ их эффективности с последующей гибридизацией эффективного метода для создания гибридной модели.

Для корректной работы алгоритмов машинного обучения тщательная предобработка входных данных является обязательным этапом. Модели машинного обучения демонстрируют свою эффективность при работе с «чистыми» данными, которые характеризуются нормализованными значениями, что позволяет избежать искажения результатов из-за различий в масштабах. Кроме того, категориальные данные должны быть переведены в числовой формат для точной интерпретации алгоритмами. Отсутствие шума также является важным фактором, позволяющим сосредоточиться на значимых сигналах. Данные должны быть полными, без пропусков значений, что обеспечивает полную картину анализируемой информации. Наконец, небольшая размерность данных упрощает обработку и анализ, что делает их более пригодными для использования в моделях машинного обучения. Таким образом, тщательная

предобработка данных является ключевым фактором в достижении высоких результатов в задачах машинного обучения.

Комплексные клинические медицинские данные (ККМД) часто имеют высокую размерность, разнотипность и содержат ошибки, вызванные человеческим фактором. Следовательно, на этапе предподготовки ККМД необходимо провести анализ размерности датафрейма и типа данных. Затем данные должны быть преобразованы в числовой формат посредством перекодирования категориальных переменных, при этом необходимо изменить аномальные значения, удалить дубликаты и оставить только строки с информативными параметрами значений в массиве данных.

Такой подход позволяет получить «чистые» данные, готовые для использования в моделях машинного обучения, что в конечном итоге приводит к более точным и высоким результатам. Использование программных средств для анализа и преобразования данных позволяет автоматизировать процесс подготовки данных и повысить его эффективность.

Для обработки и анализа данных были использованы: табличный процессор *MS EXEL* и язык программирования Python в облачной среде программирования *GoogleColab*. *MS EXEL* позволил быстро и эффективно подготовить данные для дальнейшего анализа, удалив дубликаты и аномальные значения, а также проведя дискретизацию данных.

Python был выбран благодаря своей простоте и широкому спектру библиотек для машинного обучения и интеллектуального анализа данных: *TensorFlow, PyTorch, Scikit-learn, Matplotlib*, *Scipy, Pandas*. Облачная среда *GoogleColab* предоставила возможность использования *GPU* или *TPU* для выполнения вычислительно интенсивных задач, что ускорило процесс анализа данных.

При создании и анализе структуры данных была использована библиотека с открытым исходным кодом *Pandas*. Эта мощная библиотека предоставляет инструменты для группировки и визуализации данных, а также для формирования сводных таблиц и выборки информации на основе заданных критериев.

Для преобразования категориальных данных в числовые значения был применен метод векторизации *OrdinalEncoder* из библиотеки *Scikit-learn. Scikit-learn* – это библиотека на языке программирования Python, с открытым исходным кодом, применяется для создания моделей в машинном обучении (Machine Learning). Она предоставляет широкий спектр алгоритмов для задач классификации, регрессии, кластеризации и других задач машинного обучения. Библиотека также включает в себя инструменты для предварительной обработки данных, выбора признаков, оценки модели и визуализации результатов, а также легко интегрируется с другими библиотеками, широко используемыми в машинном обучении и *Data Science*, включая *Matplotlib*, Plotly, *NumPy, Pandas DataFrame* и *SciPy*. Метод *OrdinalEncoder* выполняет порядковое кодирование категориальных переменных, что позволяет преобразовывать их в числовые значения с учетом определённого порядка, данное преобразование необходимо в моделях машинного обучения, которые требуют числовых входных данных.

Для обеспечения корректной работы алгоритмов машинного обучения необходимо нормализовать атрибуты ККМД, которые имеют широкие диапазоны значений. Для этого был использован метод *MinMaxScaler* из библиотеки *Scikit-learn*. Этот метод линейно масштабирует данные до фиксированного диапазона, обычно от 0 до 1, что позволяет привести различные масштабы признаков к единому виду. Нормализация данных помогает предотвратить доминирование признаков с большими диапазонами значений над признаками с меньшими диапазонами, что может привести к более точным результатам моделирования.

Для эффективного использования модели машинного обучения в рекомендательной системе данные были стандартизированы с помощью метода *StandardScaler* из библиотеки *Scikit-learn*. Этот метод удаляет среднее значение и масштабирует данные до единичной дисперсии, что представлено формулой (1):

$$z = (x - u) / s, \quad (1)$$

где x – исходное значение; u – среднее значение обучающих выборок или ноль, если параметр with_mean=False; s – стандартное отклонение обучающих выборок или единица, если параметр with_std=False.

Центрирование и масштабирование происходят независимо для каждого признака путем вычисления соответствующей статистики по выборкам в обучающем наборе. Затем среднее значение и стандартное отклонение сохраняются для использования в последующих данных. Стандартизация данных позволяет улучшить производительность модели машинного обучения и повысить точность прогнозирования.

Метод многоуровневого факторного анализа *(MLFA, Multi-Level Factor Analysis*) – это статистический метод, позволяющий проанализировать сложные данные с многоуровневой структурой [21-24]. Этот метод расширяет возможности традиционного факторного анализа, поскольку он учитывает не только корреляцию между переменными, но также их иерархическую организацию данных. Метод *MLFA* основывается на факторной модели, где каждая переменная представлена как линейная комбинация нескольких факторов. В многоуровневом факторном анализе *(MLFA)* факторы могут быть структурированы иерархически на нескольких уровнях, что дает возможность учитывать сложные взаимосвязи между переменными. Цель метода факторного анализа – найти матрицу факторных нагрузок Λ размером $p \times k$, формула (2), тогда матрица данных X будет иметь вид согласно формуле (3).

$$\Lambda = (\Lambda 1, \Lambda 2 ... \Lambda_k), \tag{2}$$

где $\Lambda 1$ – матрица факторных нагрузок для уровня 1 ($k1$ факторов); $\Lambda 2$ – матрица факторных нагрузок для уровня 2 ($k2$ факторов); Λk – матрица факторных нагрузок для уровня k; p – количество переменных; k – количество факторов.

$$X = \Lambda F + \varepsilon, \tag{3}$$

где F – матрица факторов размером $n \times k$; n – количество наблюдений; p – количество переменных; X – матрица данных, ε – матрица ошибок размером $n \times p$.

Метод независимых компонент *(ICA)* – это статистический метод, применяемый для выделения независимых компонент из смешанных сигналов. Он основывается на предположении о том, что исходные сигналы не зависят друг от друга и не коррелируют, то есть не имеют линейной зависимости. Метод *ICA* был впервые предложен французским математиком и информатиком Пьером Комоном (Pierre Comon) в 1994 году. Однако термин «независимый компонентный анализ» был введен в 1995 году Херкулом (Aapo Hyvärinen) и Ойя (Erkki Oja) [25-27]. Основные принципы *ICA* основываются на независимости (исходные сигналы независимы и не коррелируют друг с другом), нелинейности (смешанные сигналы представляют собой нелинейную комбинацию исходных сигналов), не стационарности (смешанные сигналы могут меняться во времени).

В начале работы метода *ICA* данные подготавливаются путем центрирования и нормализации. Центрирование данных, представлено формулой (4), выполняется путем вычитания среднего значения из каждого столбца матрицы X.

$$X_c = X - E[X], \tag{4}$$

где $E[X]$ – матрица средних значений размером $1 \times m$; l – количество наблюдений (строк) в матрице; m – количество признаков (столбцов) в матрице.

Затем проводится нормализация данных путем деления каждого столбца матрицы X_c на его стандартное отклонение, представлено формулой (5).

$$X_n = \frac{X_c}{\sigma}, \tag{5}$$

где σ – матрица стандартных отклонений размером $1 \times m$.

После подготовки данных определяется количество независимых компонент s с помощью критерия информационного критерия (*InfoMax*) или подобного алгоритма оптимизации. Затем инициализируется матрица весов W размером $m \times$s, которая может быть выполнена случайным образом. Для обновления матрицы весов W (максимизирует функцию контраст-

ности L), используется алгоритм градиентного спуска. Функция контрастности L используется для оценки качества разделения смешанных сигналов на их независимые компоненты, представлено формулой (6), обновление весов в алгоритме градиентного спуска представлено формулой (7).

$$L = I(Y) = H(Y) - \sum\left[H_{y_i}\right], \tag{6}$$

где $I(Y)$ – информационный критерий; $H(Y)$ – энтропия матрицы Y; $H(y_i)$ – энтропия i-й компоненты матрицы Y.

$$W_{new} = W_{old} - \mu \cdot \frac{\partial L}{\partial W}, \tag{7}$$

где μ – скорость обучения; L – функция контрастности; $\partial L/\partial W$ – градиент функции контрастности по матрице весов W.

Этот процесс повторяется до тех пор, пока не будет достигнуто заданное значение функции контрастности или не будет превышено максимальное количество итераций.

Таким образом, модель ICA можно представить, как линейную комбинацию независимых компонент S, смешанных с помощью матрицы A, с добавлением шума N, представлено формулой (8).

$$X = AS + N, \tag{8}$$

где A – матрица смешивания размером $m \times k$; S – матрица независимых компонент размером $n \times k$; N – матрица шума размером $n \times m$.

Цель метода ICA – найти матрицу W, которая представляет собой обратную матрицу A, чтобы восстановить независимые компоненты S, представлено формулой (9).

$$W = A\hat{}(-1), \tag{9}$$

где W – матрица размером $k \times m$, которая представляет собой обратную матрицу A.

Восстановление независимых компонент S выполняется с помощью матрицы весов W путем умножения наблюдаемых данных X на матрицу W. Это означает, что матрица W может быть использована для разделения смешанных сигналов на их независимые компоненты, приведено формулой (10).

$$S = WX, \tag{10}$$

где S – матрица независимых компонент размером $n \times k$. В работе был использован метод *FastICA*, он использует более быстрый и эффективно сходящийся алгоритм для вычисления независимых компонент, чем стандартный метод *ICA*. *FastICA* предоставляет опцию регуляризации, позволяющую управлять количеством детализации в извлеченных компонентах.

Метод стохастического байесовского факторного анализа (*Bayesian Factor Analysis* (*BFA*)) [28, 29] представляет собой метод статистического анализа, который использует байесовский вывод, что позволяет включать априорные знания о параметрах модели. Модель предполагает существование скрытых переменных (факторов), которые объясняют корреляции между наблюдаемыми переменными. Метод позволяет интегрировать различные источники данных, а также учитывать структурные особенности модели. Оценка параметров модели в методе *BFA* часто осуществляется с использованием таких методов, как метод Монте-Карло по образцу (цепь Маркова Монте-Карло, MCMC), позволяющий производить выборки из апостериорного распределения параметров, учитывая как наблюдаемые данные, так и априорные распределения. По формуле следует учитывать, что ошибки независимы и нормально распределены со средним значением 0 и ковариационной матрицей Ψ. Однако предполагается, что ковариационная матрица является полностью положительно-определенной матрицей, которая в среднем априори диагональна для представления традиционных представлений о модели, содержащей «общие» и «специфические» факторы. Также предполагается, что оценки факторов не являются фиксированными [представлено формулой (11)], а являются случайными нормально распределенными переменными со средним значением 0 и ковариацией $R = I_m$ (где I_m – единичная матрица) при том, что оценки факторов и ошибки независимы. Предположение о том, что оценки факторов не являются фиксированными, обу-

словлено байесовским подходом, где при анализе все параметры модели (включая факторы) рассматриваются как случайные величины, а не как фиксированные значения. Это позволяет учесть априорные знания и неопределенность в данных. Оценки факторов генерируются с использованием методов, таких как MCMC (Марковские цепи Монте-Карло), которые позволяют получать выборки из апостериорного распределения параметров. Предположение о случайности факторов делает модель более гибкой и позволяет учитывать изменчивость данных. Это важно, когда данные имеют сложную структуру или содержат шум. В формуле $fi \sim N(0, I_m)$ указано, что вектор оценок факторов *fi* моделируется как случайная величина с нормальным распределением, где I_m – единичная матрица. Это стандартное предположение, которое упрощает вычисления и обеспечивает идентифицируемость модели.

$$\left(x_i \mid \mu, \lambda, f_i, m\right) = \mu + \lambda + f_i + \varepsilon_i, \tag{11}$$

где μ – вектор среднего значения генеральной совокупности; Λ – матрица факторной нагрузки; f_i – вектор оценок факторов; m – количество факторов; ε_i – вектор ошибок (размером $p \times 1$), который предполагается нормально распределённым с $\varepsilon_i \sim N(0, \Psi)$; Ψ – ковариационная матрица ошибок; N – нормальное распределение; p – число переменных.

Также оценки факторов и ошибки предполагаются независимыми. Данное утверждение важно для корректной интерпретации модели и разделения влияния факторов и случайных ошибок. Факторы (скрытые переменные, которые объясняют корреляции между наблюдаемыми переменными) и ошибки (случайные отклонения, которые не объясняются факторами) не влияют друг на друга. Математически это выражается в том, что ковариация между вектором оценок факторов *fi* и вектором ошибок ε_i равна нулю: $Cov(f_i, \varepsilon_i) = 0$. В статистике предполагается, что ошибки – это случайные отклонения, которые возникают из-за различных факторов, таких как шум в данных, неточности измерений или пропуски. Если ошибки зависят от величин факторов, это может привести к систематическим смещениям в оценках, что нарушит предпосылки модели. Также независимость факторов и ошибок позволяет использовать стандартные методы оценки, такие как метод максимального правдоподобия (*ML*) или байесовские методы, без риска того, что ошибки будут влиять на оценки факторов. Эти методы учитывают структуру скрытых переменных и обеспечивают более надежные и устойчивые результаты. Независимость факторов и ошибок способствует большей устойчивости выводов, что позволяет делать более обоснованные рекомендации на основе полученных данных. Важно проводить диагностику модели, чтобы проверить, соблюдаются ли предположения о независимости. Это может быть сделано с помощью различных тестов и графиков, таких как анализ остатков (например, график остатков против предсказанных значений) и тесты на автокорреляцию (например, тест Дарбина – Уотсона).

Метод частичных наименьших квадратов *(Partial Least Squares, PLS)* [30, 31] представляет собой мультивариантный метод факторного анализа, предназначенный для изучения взаимоотношений между двумя наборами переменных – *X* и *Y*. Метод *PLS* сочетает свойства метода главных компонент и множественной регрессии. Сначала он выделяет набор скрытых факторов, которые объясняют, как можно больше ковариации между независимыми и зависимыми переменными. Затем на этапе регрессии предсказываются значения зависимых переменных с использованием декомпозиции независимых переменных. Механизм работы метода *PLS* реализуется итеративно, путем нахождения последовательных пар факторов в наборе *X* и наборе *Y*, которые объясняют наибольшую часть ковариации между двумя наборами переменных. В каждом шаге итерации *PLS* находит фактор в наборе *X*, который объясняет наибольшую часть вариации в наборе *Y*, а затем находит фактор в наборе *Y*, который объясняет наибольшую часть вариации в наборе *X*. Этот процесс повторяется до тех пор, пока не будет найдено заданное число факторов или до тех пор, пока не будет объяснена большая часть ковариации между двумя наборами переменных. Таким образом, суть метода *PLS* заключается в одновременном разложении матриц данных *X* и *Y* на линейные комбинации их столбцов, а затем в использовании этих линейных комбинаций для построения регрессионной модели.

Общая базовая модель многомерного метода *PLS* представлена формулами (12) и (13):

$$X = TP^T + E, \quad (12)$$

$$Y = UQ^T + F, \quad (13)$$

где X – это $n \times m$ матрица предикторов; Y – это $n \times p$ матрица ответов; T, U – это $n \times l$ матрицы, которые являются, соответственно проекциями X (оценка X, факторная матрица) и проекциями Y (оценки Y); P, Q – это матрицы $m \times l$ и $p \times l$ (загрузка матриц); E, F – это члены ошибок, которые предполагаются независимыми и одинаково распределенными случайными нормальными величинами.

Разложения X и Y производятся таким образом, чтобы максимизировать ковариацию между T и U.

Экспериментальные исследования

Биомедицинские данные являются высоко коррелированными и часто не обладают нормальным распределением, что связано с множеством факторов, таких как сложность биологических систем, ограниченность выборки и артефакты измерений. В ходе предобработки данных в исследуемом наборе были удалены дубликаты и аномальные значения, проведена дискретизация данных, порядковое кодирование категориальных переменных, нормализация и стандартизация. Для проверки степени мультиколлинеарности данных был использован фактор инфляции дисперсии (VIF), который показывает уровень нестабильности оценок коэффициентов регрессии при условии мультиколлинеарности (высококоллинеарные показатели были удалены). Так как $1 < \text{VIF} < 5$ обычно считается приемлемым уровнем мультиколлинеарности, то из набора были удалены показатели со значением выше 5 (2 признака). Диапазон $1 < \text{VIF} < 5$ считается приемлемым уровнем мультиколлинеарности, что соответствует общепринятым эмпирическим правилам в статистике и анализе данных [32]. Для подбора оптимального количества компонент в исследованном наборе данных был использован метод снижения размерности PCA, далее с помощью метода Кайзера были вычислены собственные значения (eigenvalues), которые отражают долю вариации, объясняемую каждым фактором. Оценка точности и эффективности методов факторного анализа была проанализирована с помощью метрик: объясненной дисперсии (explained variance), критерия сферичности Бартлетта (Chi-square), коэффициента детерминации (R^2), информационных критериев BIC и AIC, критериев подгонки модели: RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), TLI (Tucker-Lewis Index). Результаты работы каждого метода факторного анализа были проанализированы с помощью 10 критериев. Результаты работы методов факторного анализа приведены в таблице 1.

Метод PLS объясняет меньше половины дисперсии в целевой переменной (0,46), значение хи-квадрат для модели PLS (87,36) ниже, чем для нулевой модели (162,38), что указывает на лучшую подгонку модели к данным по сравнению с нулевой моделью. Коэффициент детерминации R^2 (0,46) близок к объясненной дисперсии, что указывает на то, что модель PLS объясняет около 46 % дисперсии в целевой переменной. Значения AIC и BIC относительно высоки (15,36 и 32,54 соответственно), что может указывать на избыточную сложность модели. Значения RMSEA, SRMR, CFI и TLI близки к 0, что свидетельствует о хороших показателях подгонки модели.

Таблица 1 – Результаты работы методов факторного анализа
Table 1 – Calculation Results

Метод	Explained variance	Chi-square	Chi-square for null model	R^2	AIC	BIC	RMSEA	SRMR	CFI	TLI
PLS	0,46	87.36	162.38	0,46	15,36	32,54	0,109	1	0,46	0,97
FastICA	6,99	145.98	162.38	0,70	-89,38	-72,2	0,15	1,3	0,10	0,07
BFA	8,40	191.02	162.38	0,63	-71,91	-54,73	0,17	1,5	-0,17	-0,56
MLFA	8,06	190.17	162.38	0,62	-70,17	-52,99	0,17	1,49	-0,17	-0,55

Метод FastICA показал некорректный результат дисперсии в целевой переменной (6,99), возможно, что данный критерий не подходят для оценки метода FastICA. Значение хи-квадрат для модели FastICA (145,98) ниже, чем для нулевой модели, что указывает на хорошую подгонку модели к данным по сравнению с нулевой моделью. Коэффициент детерминации R^2 (0,70) показал высокий результат, что указывает на то, что модель объясняет значительную часть дисперсии в данных. Отрицательные значения AIC и BIC (–89,38 и –72,2 соответственно) свидетельствуют, что модель слишком сильно приспосабливается к данным (логарифм правдоподобия велик относительно числа параметров модели). Значения RMSEA, CFI и TLI близки к 0, что свидетельствует о хороших показателях подгонки модели, несмотря на возможное переобучение. Однако критерий SRMR показал некорректный результата (1,3), что указывает на непригодность этой модели для данных, а также наличие мультиколлинеарности, когда если есть высокая корреляция между независимыми переменными, что может привести к высоким значениям SRMR.

При реализации метода BFA также был получен некорректный результат дисперсии в целевой переменной (8,40), что возможно связано с несоответствием критерия либо настройки параметров метода. Значение хи-квадрат для модели BFA (191,02) выше, чем для нулевой модели, что указывает на худшую подгонку модели к данным по сравнению с нулевой моделью и другими методами. Отрицательные значения AIC и BIC (–71,91 и –54,73 соответственно) свидетельствуют, что модель слишком сильно приспосабливается к данным, в данном случае лучше провести дополнительные тесты для исключения переобучения модели. Коэффициент детерминации R^2 (–71,91) отрицательный, что неверно и указывает на плохую прогнозную производительность модели BFA. Значения RMSEA, SRMR, CFI и TLI относительно высоки и отрицательны, что свидетельствует о плохих показателях подгонки модели.

Метод MLFA показал результат дисперсии в целевой переменной (8,06), близкую к BFA (8,40), что возможно связано с несоответствием критерия либо настройки параметров метода. Значение хи-квадрат для модели MLFA (190,17) также выше, чем для нулевой модели, что указывает на худшую подгонку модели к данным по сравнению с нулевой моделью и другими методами. Отрицательные значения AIC и BIC (–70,17 и –52,99 соответственно) свидетельствуют, что модель слишком сильно приспосабливается к данным, в данном случае лучше провести дополнительные тесты для исключения переобучения модели. Коэффициент детерминации R^2 (–70,17) отрицательный, что указывает на плохую прогнозную производительность модели MLFA. Значения RMSEA, SRMR, CFI и TLI относительно высоки и отрицательны, что свидетельствует о плохих показателях подгонки модели.

Нормальное распределение результатов работы методов факторного анализа представлено на рисунке 1. Высокие пики плотности распределения указывают на значения факторов, которые чаще встречаются в данных. Широкие кривые плотности указывают на большую вариабельность в значениях факторов, в то время как узкие кривые могут указывать на более однородные данные. При визуальном анализе результатов работы методов факторного анализа видно, что метод BFA и метод MLFA показывают более четкие и узкие распределения, это означает, что они лучше объясняют структуру данных с точки зрения интерпретации. Однако нужно учитывать сложность и мультиколлинеарность структуры медицинских данных при интерпретации и анализе результатов. Данную структурную особенность медицинских данных хорошо демонстрируют методы PLS и FastICA.

При анализе оценок результатов работы методов факторного анализа метод PLS показал лучшие результаты для небольшого набора комплексных медицинских данных. Он демонстрирует приемлемое объяснение дисперсии в целевой переменной (46 %), показывает лучшую подгонку к данным по сравнению с нулевой моделью, имеет близкие к идеальным показателям для RMSEA, SRMR, CFI и TLI, а также не страдает от серьезных недостатков, обнаруженных в других методах (BFA, MLFA), таких как переобучение, мультиколлинеарность и плохая прогнозная производительность. Метод FastICA также заслуживает внимания благодаря высокому R^2 и хорошим показателям подгонки (за исключением SRMR). Однако

некорректный результат дисперсии в целевой переменной и потенциальная мультиколлинеарность снижают уверенность в этом методе. Визуализация оценок результатов работы методов факторного анализа приведена на рисунке 2.

Для улучшения работы метода PLS было принято использовать более сложные, гибридные модели, которые могут лучше объяснить сложные зависимости в данных. Таким образом, были разработаны ансамблевые сочетания методов PLS и метода случайного леса (PLS-RF), PLS и метода нейронных сетей (PLS-NN). Для реализации гибридного метода, основанного на нейронных сетях, была использована нейронная сеть – многослойный персептрон. Результаты работы гибридных методов представлены в таблице 2.

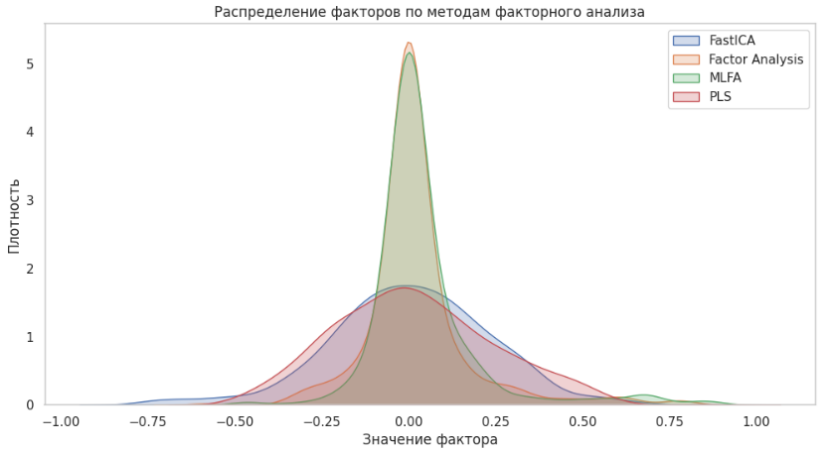

Рисунок 1 – Распределение факторов по методам факторного анализа
Figure 1 – Distribution of factors by factor analysis methods

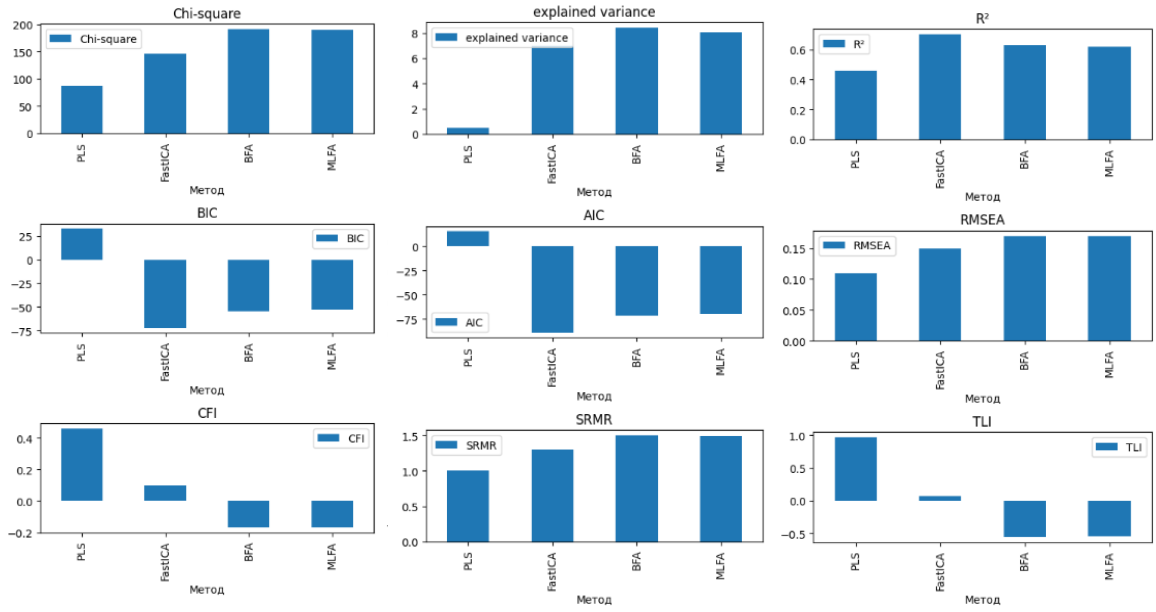

Рисунок 2 – Оценка результатов работы методов факторного анализа
Figure 2 – Evaluation of factor analysis methods results

Таблица 2 – Результаты работы гибридных методов факторного анализа
Table 2 – Results of hybrid methods of factor analysis

Метод	Explained variance	Chi-square	Chi-square for null model	R^2	AIC	BIC	RMSEA	SRMR	CFI	TLI
PLS-RF	0,98	2,87	162,38	0,98	-273,19	-256,01	0,02	0,2	0,98	2,06
PLS-NN	0,99	1,37	162,38	0,99	-342,03	-324,85	0,01	0,1	0,99	2,08

Анализ результатов работы гибридных методов факторного анализа показал, что оба метода имеют высокие значения объясненной дисперсии (98 % и 99 %, соответственно), что свидетельствует о хорошей способности моделей объяснять вариацию в исходных данных. Однако метод PLS-NN справился лучше. Значения показателя Chi-square в обеих моделях указывают на хорошее соответствие между наблюдаемыми и предсказанными значениями. При этом у метода PLS-NN этот показатель заметно ниже (1,37 против 2,87), что говорит о лучшей подгонке модели к данным. Коэффициент детерминации (R^2) также подтверждает превосходство метода PLS-NN, так как он достигает значения 0,99, тогда как у PLS-RF этот показатель равен 0,98. Информационные критерии AIC и BIC для обеих моделей имеют отрицательные значения, что указывает на их адекватность. Однако у метода PLS-NN эти показатели существенно ниже (–342,03 и –324,85 против –273,19 и –256,01), что делает эту модель предпочтительной. Значение RMSEA для модели PLS-NN составило 0,01, что находится в пределах идеального диапазона (< 0,05) и свидетельствует о хорошем соответствии модели данным. Для модели PLS-RF значение RMSEA составило 0,02, что также попадает в допустимый диапазон (< 0,05), но ниже, чем у модели PLS-NN. Значение показателя SRMR для модели PLS-NN составило – 0,10, что находится в пределах хорошего соответствия (< 0,10). В то же время значение SRMR для модели PLS-RF составило 0.20, что указывает на приемлемое, но менее оптимальное соответствие модели данным. Значения индексов CFI и TLI для обеих моделей находятся в диапазоне хороших значений (> 0,90), хотя у метода PLS-NN они несколько выше (CFI = 0,99, TLI = 2,08) по сравнению с методом PLS-RF (CFI = 0,98, TLI = 2,06), что свидетельствуют о лучшем качестве модели и лучшем соответствии данным по сравнению с методом PLS-RF. Визуализация оценок результатов работы гибридных методов факторного анализа приведена на рисунке 3. Нормальное распределение результатов работы гибридных методов факторного анализа представлено на рисунке 4. Метод PLS-RF показал более высокие пики плотности распределения факторов, что указывает на более простой подход к интерпретации состава данных. Метод PLS-NN показал широкие кривые распределения значений, что соответствует большей вариабельности в значениях факторов и составу исследуемых данных. Таким образом, при анализе оценок результатов работы гибридных методов, можно заключить, что метод PLS-NN справился лучшие по сравнению с методом PLS-RF. Метод PLS-NN обеспечивает более точную подгонку модели к данным, лучшую объясняющую способность, лучшие показатели оценки качества модели, включая в себя низкую ошибку аппроксимации.

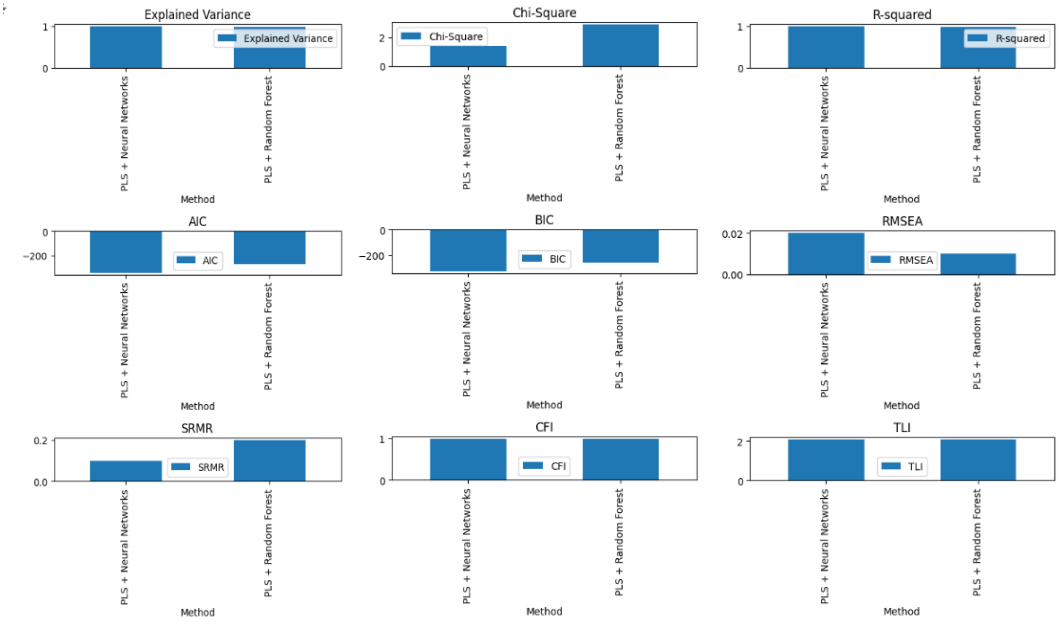

Рисунок 3 – Оценка результатов работы гибридных методов факторного анализа
Figure 3 – Evaluation of the results of factor analysis hybrid methods

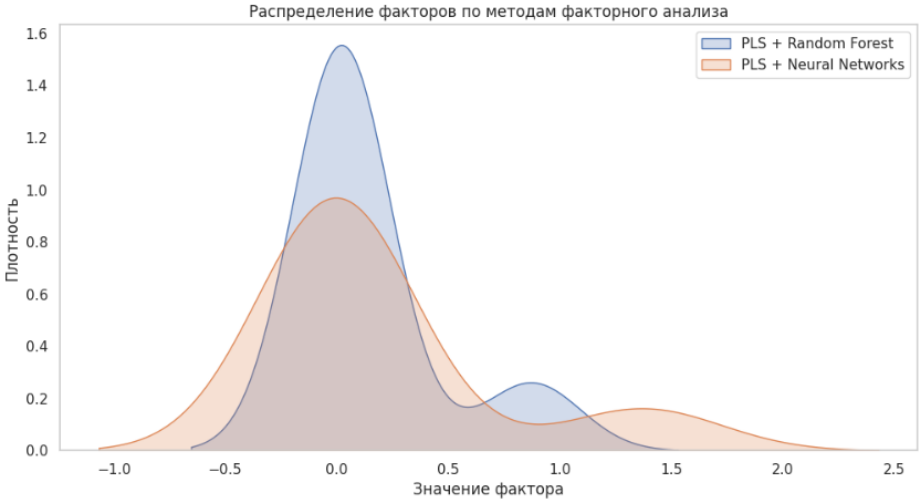

Рисунок 4 – Распределение факторов гибридных методов факторного анализа
Figure 4 – Factor Distribution in Hybrid Factor Analysis Methods

Заключение

В статье рассмотрены и проанализированы популярные методы факторного анализа (PLS, FastICA, BFA, MLFA) для состава комплексных клинических медицинских данных. В результате исследования выбран метод факторного анализа, максимально соответствующий архитектуре данных, – метод PLS. В статье предложены и проанализированы новые гибридные способы реализации метода факторного анализа PLS, это метод PLS-NN и метод PLS-RF. По результатам сравнения гибридных методов метод PLS-NN показал лучшие результаты по ряду показателей оценки качества модели и более точную подгонку модели к данным. Таким образом, метод гибридного факторного анализа PLS-NN будет рекомендован как инструмент для предварительной обработки ККМД при построении рекомендательной системы.

Библиографический список

1. Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации». URL: http://publication.pravo.gov.ru/document/view/0001201910110003. (дата обращения:01.09.2024).

2. **Кузенкова Н.Н.** Система поддержки принятия врачебных решений – цифровой инструмент врача поликлиники // Московская медицина. 2022. № 1(47). С. 54-55.

3. **Дубошинский Р.И., Колосов В.С., Немков А.Г. и др**. Анализ субъективных факторов, влияющих на освоение врачами функциональных возможностей медицинских информационных систем // Менеджер здравоохранения. 2024. № 8. С. 83-90.

4. **Galton F.** Co-relations and their measurement, chiefly from anthropometric data // Proceedings of the Royal Society of London. 1888. No.45. Pp.135-145.

5. Факторный, дискриминантный и кластерный анализ // сборник работ под ред. Енюкова И.С. М.: Финансы и статистика. 1989. 215 с.

6. **Harman H.H.** Modern factor analysis // University of Chicago Press. 1976.

7. **Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнёва Е.А., Смирнов В.И.** Технологии комплексного интеллектуального анализа клинических данных // Вестник РАМН. 2016. № 2. С. 160-171.

8. **Lawley D.N. Maxwell A.E.** Factor analysis as a statistical method (2-е изд.). Butterworths. 1971.

9. **Joreskog K.G**. Some contributions to maximum likelihood factor analysis // Psychometrika. 1966. No. 32(4). Pp. 443-482.

10. **Bentler P.M.** Multivariate analysis with latent variables: Causal modeling // Annual Review of Psychology. 1980. No.31. Pp. 419-456.

11. **Liu J., Zhang Y., Wang Z.** Deep Learning with Hierarchical Convolutional Factor Analysis // Transactions on Neural Networks and Learning Systems. 2013. No.35. Pp. 1887-1901.

12. **Hansen B., Avalos-Pacheco A., Russo M., De Vito R**. A Variational Bayes Approach to Factor Analysis // Springer Proceedings in Mathematics & Statistics. 2023. Vol. 435.

13. **Lee D.D, Seung S.H.** Learning the parts of objects by nonnegative matrix factorization // Nature. 1999. No.401. Pp. 788-791.

14. **Lee D.D, Seung S.H.** Algorithms for nonnegative matrix factorization // Adv Neural Inform Process Syst. 2001. No13. Pp. 556-562.

15. **Qiu J., Li Z., Yao J.** Robust Estimation for Number of Factors in High Dimensional Factor Modeling via Spearman Correlation Matrix // Journal of the American Statistical Association. 2024. Pp.1-13.

16. **Rockova V., George E.I**. Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity // Journal of the American Statistical Association. 2016. No. 111(516). Pp.1608-1622.

17. **Shu H., Wang X., Zhu H.** D-CCA: A Decomposition-Based Canonical Correlation Analysis for High-Dimensional Datasets // Journal of the American Statistical Association. 2019. No. 115(529). Pp. 292-306.

18. **Bai J., Ng S.** Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data // Journal of the American Statistical Association. 2021. No. 116 (536). Pp.1746-1763.

19. **Шовин В.А.** Автокорреляционная нейронная сеть факторного анализа // МСиМ. 2018. № 3 (47). С. 61-67.

20. **Шовин В.А.** Факторный анализ на базе метода К-средних // МСиМ. 2018. № 4 (48). С. 78-84.

21. **Pearson K.** On lines and planes of closest fit to systems of points in space // Philosophical Magazine. 1901. No. 2(11). Pp. 559-572.

22. **Cattell R.B.** The multiple abstract variance analysis equations and solutions: for nature-nurture research on continuous variables // Psychometrika. 1960. No. 25(2). Pp. 163-183.

23. **Cattell R.B.** The theory of fluid and crystallized intelligence: A critical experiment // Journal of Experimental Psychology. 1963. No. 66(3). Pp. 299-306.

24. **Cattell R.B.** Factor analysis: An introduction and manual for the psychologist and social scientist. Harper & Row. 1965.

25. **Cattell R.B.** The scree test for the number of factors // Multivariate Behavioral Research. 1966. No. 1(2). Pp. 245-276.

26. **Comon P.** Independent component analysis, a new concept? // Signal Processing. 1994. No. 36(3). Pp. 287-314.

27. **Hyvarinen A., Oja E.** A fast fixed-point algorithm for independent component analysis // Neural Computation. 1997. No. 9(7). Pp. 1483-1492.

28. **Bayes Thomas and Price Richard.** An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton M.A. and F.R.S. // Philosophical Transactions of the Royal Society of London. 1763. No. 53. Pp. 370-418.

29. **Афифи А., Эйзен С.** Статистический анализ: Подход с использованием ЭВМ. М.: Мир, 1982. 488 с.

30. **Legendre A.M.** On Least Squares. From D E Smith, A Source Book in Mathematics, McGraw-Hill 1929 and Dover. 1959. Vol. II. Pp. 576-579.

31. **Линник Ю.В.** Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. 2-е изд. М. 1962.

32. **Belsley D.A., Kuh E., Welsch R.E.** Regression Diagnostics; Identifying Influence Data and Source of Collinearity. John Wiley & Sons. New York. 1980. Pp. 15-292.

UDC 004.622

DEVELOPMENT OF HYBRID FACTOR ANALYSIS METHOD FOR SMALL MEDICAL DATA SET

O. A. Popova, techer, TGMU, Tjumen, Rossia;
orcid.org/0009-0006-3530-5703, e-mail: popovaOA@tyumsmu.ru

The paper presents a comparative analysis of the effectiveness of well-known methods of factor analysis: PLS, FastICA, BFA and MLFA, as well as newly developed hybrid methods PLS-NN and PLS-RF. The main aim of the study was to identify methods that provide the best accuracy in explaining the variance in target variable and the best fit of the model to data. The results showed that PLS method explained a significant

proportion of the variance in target variable, demonstrating good model fit, but there were signs of excessive model complexity. FastICA method demonstrated high explanatory power, but potential overfitting problems were identified. BFA and MLFA methods showed unsatisfactory results, characterized by negative predictive performance and unsatisfactory values of model fit indicators. Based on the results of the study, PLS method was chosen for further improvement and adjustment. In order to increase its efficiency, hybridization was used, which significantly improved the quality of the model and its fit to the data. The analysis of the results of hybrid factor analysis methods (PLS-NN and PLS-RF) showed that both methods have high ability to explain variation in source data. However, PLS-NN method outperformed PLS-RF method in a number of indicators, such as the coefficient of determination, information criteria AIC and BIC, as well as RMSEA and SRMR indicators, which indicates a better fit of the model to data and a lower level of approximation error. In summary, the study confirms that PLS-NN is the method of choice to be used in considered dataset due to its accuracy, explanatory power and model fit quality. The problem of finding the characteristics of data channel of aircraft opto-electronic trajectory measurements is studied. ***The aim is to*** *find main channel quality indicators in a stationary mode with the transmission of priority and non-priority data opto-electronic means of trajectory measurements at different values of input stream intensity, window length and probability of frame distortion during transmission. Input stream from each measuring station is simple with a given priority level. Frame transmission through the channel is based on window control. Channel quality indicators using queuing system M/G/1 are found. The degree of frame error probability influence in case of transmission protocol window length while transmitting via communication channel on mediate values of frame number in a system, the time of frame waiting in a queue, determined for priority and non-priority frames is evaluated.*

Keywords: *medical data, factor analysis methods, PLS, FastICA, BFA, MLFA, hybrid methods of factor analysis, Random Forest, Neural Networks, coefficient of determination, information criteria, preprocessing of input data.*

DOI: 10.21667/1995-4565-2025-91-87-103

Referenes

1. Ukaz ot 10 oktjabrja 2019. № 490 «*O razvitii iskusstvennogo intellekta v Rossijskoj Federacii*». URL: http://publication.pravo.gov.ru/document/view/0001201910110003. (Date of request: 01.09.2024). (in Russian).

2. **Kuzenkova N.N.** Sistema podderzhki prinjatija vrachebnyh reshenij – cifrovoj instrument vracha polikliniki // Moskovskaja medicina. 2022, no. 1(47), pp. 54-55. (in Russian).

3. **Duboshinskij R.I., Kolosov V.S., Nemkov A.G. i dr.** Analiz sub#ektivnyh faktorov, vlijajushhih na osvoenie vrachami funkcional'nyh vozmozhnostej medicinskih informacionnyh system. *Menedzher zdravoohranenija*. 2024, no. 8, pp. 83-90. (in Russian).

4. **Galton F.** Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*. 1888, no.45, pp.135-145.

5. *Faktornyj, diskriminantnyj i klasternyj analiz* (Factorial, discriminant and cluster analysis). sbornik rabot pod red. Enjukova I.S. Moscow: Finansy i statistika. 1989, pp. 215. (in Russian).

6. **Harman, H. H.** Modern factor analysis. *University of Chicago Press*. 1976.

7. **Baranov A.A., Namazova-Baranova L.S., Smirnov I.V., Devjatkin D.A., Shelmanov A.O., Vishnjova E.A., Smirnov V.I.** Tehnologii kompleksnogo intellektual'nogo analiza klini-cheskih dannyh. *Vestnik RAMN*. 2016, no.2, pp. 160-171. (in Russian).

8. **Lawley D.N., Maxwell A.E.** *Factor analysis as a statistical method* (2 izd.). Butterworths. 1971.

9. **Joreskog K.G**. Some contributions to maximum likelihood factor analysis. *Psychometrika*. 1966, no. 32(4), pp. 443-482.

10. **Bentler P.M.** Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*. 1980, no.31, pp. 419-456.

11. **Liu J., Zhang Y., Wang Z.** Deep Learning with Hierarchical Convolutional Factor Analysis. *Transactions on Neural Networks and Learning Systems*. 2013, no.35, pp. 1887-1901.

12. **Hansen B., Avalos-Pacheco A., Russo M., De Vito R**. A Variational Bayes Approach to Factor Analysis. *Springer Proceedings in Mathematics & Statistics*. 2023, vol 435.

13. **Lee D.D., Seung S.H.** Learning the parts of objects by nonnegative matrix factorization. *Nature*. 1999, no.401, pp. 788-791.

14. **Lee D.D, Seung S.H.** Algorithms for nonnegative matrix factorization. *Adv Neural Inform Process Syst*. 2001, no. 13, pp. 556-562.

15. **Qiu, J., Li Z., Yao J.** Robust Estimation for Number of Factors in High Dimensional Factor Modeling via Spearman Correlation Matrix. *Journal of the American Statistical Association*. 2024, pp. 1-13.

16. **Rockova V., George E.I.** Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity. *Journal of the American Statistical Association*. 2016, no. 111 (516), pp. 1608-1622.

17. **Shu H., Wang X., Zhu H.** D-CCA: A Decomposition-Based Canonical Correlation Analysis for High-Dimensional Datasets. *Journal of the American Statistical Association*. 2019, no. 115(529), pp. 292-306.

18. **Bai J., Ng S.** Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data. *Journal of the American Statistical Association*. 2021, no. 116 (536), pp. 1746-1763.

19. **Shovin V.A.** Avtokorreljacionnaja nejronnaja set' faktornogo analiza. *MSiM*. 2018, no. 3 (47), pp. 61-67. (in Russian).

20. **Shovin V.A.** Faktornyj analiz na baze metoda K-srednih. *MSiM*. 2018, no. 4 (48), pp. 78-84. (in Russian).

21. **Pearson K.** On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. 1901, no. 2 (11), pp. 559-572.

22. **Cattell R.B.** The multiple abstract variance analysis equations and solutions: for nature-nurture research on continuous variables. *Psychometrika*. 1960, no. 25 (2), pp. 163-183.

23. **Cattell R.B.** The theory of fluid and crystallized intelligence: A critical experiment. *Journal of Experimental Psychology*. 1963, no. 66 (3), pp. 299-306.

24. **Cattell R.B.** *Factor analysis: An introduction and manual for the psychologist and social scientist*. Harper & Row. 1965.

25. **Cattell R.B.** The scree test for the number of factors. *Multivariate Behavioral Research*. 1966, no. 1(2), pp. 245-276.

26. **Comon P.** Independent component analysis, a new concept? *Signal Processing*. 1994, no. 36 (3), pp. 287-314.

27. **Hyvarinen A., Oja E.** A fast fixed-point algorithm for independent component analysis. *Neural Computation*. 1997, no. 9 (7), pp. 1483-1492.

28. **Bayes, Thomas, and Price, Richard** An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton M.A. and F.R.S. *Philosophical Transactions of the Royal Society of London*. 1763, no. 53, pp. 370-418.

29. **Afifi A., Jejzen S.** *Statisticheskij analiz: Podhod s ispol'zovaniem JeVM* (Statistical Analysis: A Computer-Based Approach). Moscow: Mi. 1982, 488 p. (in Russian).

30. **Legendre A.M.** *On Least Squares. From D E Smith, A Source Book in Mathematics, McGraw-Hill 1929 and Dover.* 1959, vol. II, pp. 576-579.

31. **Linnik Ju.V.** *Metod naimen'shih kvadratov i osnovy matematiko-statisticheskoj teorii obrabotki nabljudenij* (The Method of Least Squares and the Fundamentals of the Mathematical and Statistical Theory of Observation Processing). 2- izd. Moscow. 1962. (in Russian).

32. **Belsley D.A., Kuh E., Welsch R.E.** Regression Diagnostics; Identifying Influence Data and Source of Collinearity. *John Wiley & Sons*. New York. 1980, pp. 15-292.