ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ КАТЕГОРИАЛЬНОГО МЕТОДА ГЛАВНЫХ КОМПОНЕНТ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

И. В. Долгих
(Г. Новосибирск, ФГБОУ ВО Новосибирский государственный педагогический университет)
iradolgikh@bk.ru

POSSIBILITIES OF APPLICATION OF CATPCA IN EDUCATIONAL RESEARCH

I. V. Dolgikh
(Novosibirsk, Novosibirsk State Pedagogical University)

Abstract. In this article we give an overview of the possibilities of the categorical method of principal components as a correct alternative to the widely used classical method of principal components. We view CatPCA as a method of Educational Data Mining and see the further development of this method in pedagogical research.

Keywords: Categorical Principal Components Analysis, Educational Data Mining, factor analysis.

Интерес педагогов и психологов к использованию в исследованиях математических методов из года в год продолжает расти [1]. Связано это, в том числе, с необходимостью обработки и анализом информации, которая представлена в базах данных образовательных учреждений и архивах. А. Х. Кутлалиев считает, что «сверхзадача любого анализа – извлечь максимум информации из имеющихся данных» [2]. Особенностью работы с данными в гуманитарных сферах, является то, что большинство зависимостей в психологии и педагогике имеют характер не функциональной, а статистической связи. В статистической связи между двумя элементами есть элементы случайности, поэтому она проявляется как тенденция [3].

Среди задач психолого-педагогических исследований часто встречаются такие, как описание структуры изучаемых показателей, выявление однородных групп респондентов и описание их типов поведения, сравнение и т.д. Перечисленные задачи могут быть решены при помощи методов частотных распределений и таблиц смежности [4]. Вместе с тем увеличивается количество более сложных и комплексных методов: факторный и кластерный анализ, дисперсионный анализ, многомерное шкалирование, анализ соответствий, деревья классификаций и т. д. Часто встречаются ситуации, когда для целей исследований не достаточно анализа первичных статистических характеристик результатов опросов. Если установлено, что между переменными существуют взаимосвязи, которые могли бы определяться латентными факторами, то рекомендуется использовать факторный анализ [5].

Одним из наиболее распространенных методов выявления факторов является метод главных компонент (Principal Components Analysis, PCA), который широко используется как инструмент разведывательного анализа, изучения структуры данных на предварительных этапах исследования [5]. Однако, как и любой метод, он имеет свои методологические ограничения. Во-первых, метод может быть использован только с числами, в математическом смысле этого слова. Во-вторых, в PCA предполагается, что переменные связаны между собой линейно. Линейность предполагает, что может быть рассчитан коэффициент корреляции Пирсона, который «ловит» линейную связь между переменными. В-третьих, модель предполагает, что изучаемые характеристики имеют нормальное распределение, но в действительности, исследователь скорей всего не знает характер распределения изучаемых переменных.

Применительно к психолого-педагогическим исследованиям (речь идет об образовательных данных) указанные ограничения существенны. Большинство данных измерено на шкалах более низкого уровня, чем количественные. Под измерением будем понимать алгоритмическую операцию, которая данному наблюдаемому состоянию объекта, процесса, явления ставит в соответствие определенное обозначение: число, символ и др. [6]. Предположение о линейности также накладывает ограничения на то, к каким данным целесообразно применять метод главных компонент. В случае привлечения метода главных компонент ис-

следователь принимает предпосылку метода о том, что изучаемые переменные линейно связаны между собой, но часто ему неизвестен характер взаимосвязи между переменными. Таким образом, мы наблюдаем задачу рассматривать и анализировать структуру переменных и несоответствие наиболее часто используемого метода разведывательного факторного анализа методом главных компонент данной задаче в силу методологических ограничений последнего. Данная проблема может быть решена посредством использования метода, который бы корректно работал с данными, измеренными по шкалам порядкового и номинального типов.

Категориальный метод главных компонент появился как результат слияния классического метода главных компонент и оптимального шкалирования [4] . Серьезным отличием классического и категориального метода главных компонент является то, что «входными» данными первого является корреляционная или ковариационная матрица, а «входными» данными второго – данные, то есть матрица n×m, где n – респонденты, m – переменные. Этот факт является преимуществом, поскольку позволяет исследователю одновременно включить в анализ переменные, измеренные на разных уровнях. На «выходе» исследователь имеет решение с указанным количеством главных компонент, факторные нагрузки и факторные значения [4].

В литературе и статистических программах категориальный метод главных компонент известен под разными наименованиями, каждое из которых появилось как реакция на уже имеющиеся названия этого алгоритма анализа данных. Так, например, категориальный метод главных компонент доступен в SPSS с 1999 года в блоке «Категориальные данные» (Categories). Метод назывался PRINCALS (Principal Components Altering Least Squares) в версиях до 12.0, позднее был переименован в CATPCA (Categorical Principal Components Analysis).

В работе ученых Ф. Лагона и Ф. Падовано приводятся аргументы в пользу применения именно этого метода анализа данных [4]. Одним из таких аргументов является минимизация потерь информации. Также преимуществом CatPCA в [4] называют, корректную работу с порядковыми данными. Нелинейный метод главных компонент не делает предположений относительно характера распределения переменных, а также согласно выбранному уровню шкалирования анализирует расстояния между соседними значениями шкалы. Имеется возможность осуществить нелинейное преобразование данных, соответствующее конкретному факторному решению. Также отметим, что нелинейный метод главных компонент используется в направлении Data Mining [7].

Проведение разведывательного анализа категориальным методом главных компонент будет реализовано на базе исследования «Повышение удовлетворенности потребителей качеством образовательных услуг в бакалавриате и магистратуре». Таким образом, в наших дальнейших исследованиях CatPCA будет рассматриваться как метод Educational Data Mining.

ЛИТЕРАТУРА

1. Кузьмин Р. И., Макарова Л. Н. Специфика корреляционно-регрессионного моделирования в рамках психолого-педагогических исследований // Вестник Тамбовского университета. Серия: Естественные и технические науки. 2012. Т. 17. № 3. С.1059-1067.

2. Кутлалиев А. Х. Методы статистического анализа в прикладных социальных исследованиях. 1955 – 2005 гг. – М.: ГФК Русь, 2007. С. 10

3. Коростелкин Б. Г. Применение методов корреляционного и факторного анализа в психолого-педагогических исследованиях // Вестник Челябинского университета. Серия 5. Педагогика, Психология, 2001. № 1. -С. 46-55.

4. Lagona F.,·Padovano F. A nonlinear principal component analysis of the relationship between budget rules and fiscal performance in the European Union// Public Choice, 2007, №130, pp. 401-436//

5. Долгих И. В. Совершенствование методов оценивания сущности брендов с использованием факторного анализа // Материалы 54 международной научной студенческой конференции (МНСК-2016), 16–20 апр. 2016 г. – Новосибирск: Изд-во НГТУ, 2016. – С. 20-22 - ISBN 978-5-4437-0498-2

6. Эконометрика: учеб.-метод. комплекс для дистанц. Обучения / сост.: А. Л. Осипов, В. Н. Храпов. – Новосибирск: Изд-во СибАГС, 2002. – 172 с.

7. Joshua B. Tenenbaum, Vin de Silva, John C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science Vol 290, 22 December 2000, 2319-2323.

ПРОБЛЕМА КАЧЕСТВА ИСХОДНЫХ ДАННЫХ В МАТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ В РФ

С. В. Романчуков
(г. Томск, Томский политехнический университет)
inoy@vtomske.ru

PROBLEM OF THE INITIAL DATA QUALITY IN THE MATHEMATICAL MODELING OF SOCIAL AND ECONOMIC PROCESSES IN THE RUSSIAN FEDERATION

Sergey Romanchukov
(Tomsk, Tomsk Polytechnic University)

Abstract. This article is devoted to the problems of the initial data quality in the model development of socio-economic processes in Russia. It addresses to such weaknesses in modern Russian sociology and comparativistics as insufficient degree of automation and research software support, low degree of standardization of research procedures and incompatibility of obtained results from different regions. The main task of this article is to outline the boundaries of the problem and re-energize the process of its discussion.

Keywords: social studies, data analysis, region, models, indexes.

Введение. В текущем состоянии российской и мировой экономики и социальной сферы особую важность приобретают исследования, позволяющие оценить динамику развития регионов, построить описательные и прогностические модели, позволяющие повысить эффективность управления за счёт повышения точности прогнозирования происходящих социально-экономических процессов и возможных последствий запланированных мероприятий. Значимость изысканий в данной сфере подтверждается так же количеством осуществляемых на территории РФ проектов по региональной компаративистике, однако, зачастую подобные проекты оказываются ограничены рамками сугубо социологии и математической статистики, в то время как более активное применение математических методов, информационных и сетевых технологий, методик вычислительного эксперимента ограничено в силу целого ряда проблем. Результаты применения таких методов крайне зависимы от качества, объёма, достоверности, доступности, полноты и степени формализации исходных данных.

Проблема недостаточной информатизации социальных исследований. Разумеется, данная проблема актуальна не первый год, и по этой теме опубликовано достаточно большое количество материалов, однако даже их авторы зачастую отмечают тот факт, что разработанные в рамках информационных технологий компьютерные приёмы решения социологических задач, остающиеся неизвестными большинству социологов и психологов, "указанный автор (прим. Давыдов А.А.) долго и тщетно пытался привлечь внимание к ним социологов"[1] и потому зачастую не внедряются в реальную каждодневную практику исследовательских групп. Кроме того акценты в большинстве подобных публикаций смещены в сторону задач анализа и обработки социологических данных (как например работы Толстовой Ю. Н., Давыдова А.А. и других), которые является важными и значимыми, но отнюдь не