УДК 735.29

ИССЛЕДОВАНИЕ МНОГОМЕРНЫХ ДАННЫХ МЕТОДАМИ ФАКТОРНОГО АНАЛИЗА

Кучеров М.Н.
Научный руководитель: доцент, к.ф.-м.н. Баранова И.В.
Сибирский Федеральный Университет
Институт Математики и фундаментальной

RESEARCHING OF MULTIDIMENSIONAL DATA BY FACTOR ANALYSIS METHODS

Kucherov M.N.
Scientific Supervisor: Assoc. Prof., PhD. Baranova I.V.
Siberian Federal University, Institute of Mathematics and Computer Science,
Russia, Krasnoyarsk, Svobodny str., 79, 660041
E-mail: matrix-elf@yandex.ru

This paper considers the concepts and application of factor analysis. In work are described the basic techniques of classical factor analysis: principal component and centroid methods. Also work contains the solution of the practical task of research multifactor statistics using foregoing methods.

В данной статье рассматривается понятие факторного анализа, его основные проблемы и методы. Также в работе решается практическая задача применения ряда методов факторного анализа для изучения многомерных статистических данных.

Факторный анализ − многомерный статистический метод, применяемый для изучения взаимосвязей между значениями переменных. Делается предположение о том, что несколько измеряемых переменных коррелируют между собой. Эта ситуация возникает в ситуации, когда они взаимно определяют друг друга или связь между этими переменными обуславливается какой-то третьей величиной, которую непосредственно измерить нельзя. Возникает задача, можно ли по данным переменным выделить величину, так называемый фактор, который объяснил бы наблюдаемые связи. Под термином ***фактор*** подразумевается математическая величина, получаемая на основе наблюдений.

Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических связей корреляций между наблюдаемыми переменными. В основе методики факторного анализа лежит анализ коэффициентов корреляции наблюдаемых переменных.

Схема решения и основные проблемы факторного анализа

Целью факторного анализа является получение из исходной матрицы данных Y размерности (n,m) матрицу P размерности (r,n), где r < n.

Любой метод факторного анализа начинается с $Y = (y_{ij})$ − матрицы исходных данных. Здесь $i = 1,\ldots,m$ − переменные, а $j = 1,\ldots,n$ − индивидуумы (объекты). По ней вычисляется корреляционная матрица $R = (r_{ik})$, где $i,k = 1,\ldots,m$ − коэффициенты корреляции соответствующих переменных. По главной диагонали корреляционной матрицы проставляются оценки общностей (сумм квадратов нагрузок факторов) и

получают редуцированную корреляционную матрицу $R_h = \left(r_{ik}^h\right),\ i,k = 1,\ldots,m$ − остаточные коэффициенты корреляции. Это составляет проблему общности, которая состоит в установлении оценок общностей. Из матрицы R_h с помощью определенных способов извлекают факторы, получая в результате матрицу факторного отображения $A = \left(A_{ij}\right)$, элементами которой являются факторные нагрузки. Здесь $i = 1,\ldots,r$ − факторы, $j = 1,\ldots,m$ − переменные. Столбцы матрицы *A* ортогональны и занимают произвольную позицию в отношении переменных, определяемую методом выделения факторов. Возможно большое число матриц *A*, которые будут одинаково хорошо воспроизводить R_h по равенству $R_h = A \cdot A'$. Из них должна быть выбрана одна, что составляет проблему вращения. Решение проблемы вращения одним из нескольких способов приводит к матрице $V = \left(v_{ij}\right)$ − факторной матрице после поворота, $i = 1,\ldots,r$ − факторы, $j = 1,\ldots,m$ − переменные. Последняя проблема факторного анализа касается оценки значений факторов для каждого индивидуума путем нахождения $P = \left(p_{ij}\right)$ − матрицы значений факторов, $i = 1,\ldots,r$ − факторы, $j = 1,\ldots,m$ − индивидуумы (объекты).

Метод главных компонент

При решении задачи измерения распределенных параметров y_n индивидуумов (или объектов), *n* точек сосредоточены в nмерном пространстве с n осями в облаке вокруг общего центра тяжести. Это облако точек наблюдений в общем случае имеет овальную форму и называется эллипсоидом. В геометрическом плане метод главных факторов состоит в том, что:

1. определяют самую длинную ось эллипсоида. Она является первой главной осью, которая должна пройти через центр тяжести.
2. Устанавливают подпространство (в данном случае плоскость), которое перпендикулярно к первой главной оси и которое проходит через центр тяжести. В этом подпространстве находится следующая по величине ось скопления точек и т. д., пока не будут определены последовательно все главные оси.

Центроидный метод

Синонимом названия «центроидный метод» является «метод центра тяжести». Это название объясняет принцип метода. Положение первой координатной оси должно быть определено так, чтобы она проходила через центр тяжести скопления точек. Факторное отображение можно рассматривать как размещение *m* точек-переменных в *r*-мерном пространстве, причем отдельные точки или векторы представляют переменные. Чтобы получить однозначное положение системы координат, условливаются, что первая ось должна проходить через центр тяжести скопления точек-переменных. Вторая ось выбирается перпендикулярно к первой, третья − перпендикулярно ко второй и т.д.

При применении метода главных факторов и центроидного метода возникает вопрос: когда должен быть закончен процесс выделения факторов? Общепризнанного метода определения числа факторов, подлежащих выделению, не существует. Наиболее известным является *критерий Кайзера* и *Дикмана*: факторы, вклады которых (сумма квадратов факторных нагрузок) в полную дисперсию меньше единицы, имеют долю дисперсии, меньшую единичной дисперсии переменных. Такие факторы не должны выделяться.

Применение методов факторного анализа для исследования многофакторных статистических данных

В работе была решена практическая задача, заключающая в исследовании медицинской статистики с помощью изученных методов факторного анализа. Статистика состояла из двух частей: в первой части были приведены показатели для пациентов с удачным исходом оперативного вмешательства, во второй − с неудачным. Состояние здоровья каждого пациента оценивалось 114 показателями, такими как: группа крови, резус, рост, вес, место проживания, возраст, курение, наличие заболеваний органов системы дыхания, кровеносной системы, перенесенных инфекций, хронических заболеваний, характеристики проведенного лечения и госпитализации (недостаточная квалификация персонала, неправильный диагноз, ошибки лечения, отсутствие профилактики) и т.д.

Для решения задачи статистика была разделена на 2 группы (в соответствии с исходом). Для каждой группы была составлена корреляционная матрица. Для корреляционной матрицы найдены характеристические корни и векторы. Затем, с помощью умножения каждого собственного вектора на корень из соответствующего собственного числа, были получены векторы нагрузок. Из этих векторов была составлена матрица, в которой каждая строчка представляет собой фактор, ячейками которого являются нагрузки.

В первый фактор для группы пациентов с удачным исходом оперативного вмешательства были включены 8 переменных, среди которых наибольшую нагрузку имели показатели, связанные с наличием заболеваний кровеносной системы, перенесенных инфекций и хронических болезней. Для второй группы пациентов наибольшие нагрузки в первом факторе имеют показатели проведенного лечения и госпитализации (неправильный диагноз, несвоевременное обращение, отсутствие необходимого оборудования), а уже затем − наличие хронических болезней и перенесенных инфекций.

В результате исследования можно сделать вывод, что для улучшения исходов операций, специалистам в области здравоохранения следует обратить наибольшее внимание на выявленные показатели проведенного лечения и госпитализации и принять меры по повышению квалификации медицинского персонала и закупкам необходимого оборудования.

СПИСОК ЛИТЕРАТУРЫ

1. Иберла. К. Факторный анализ. – М.: Наука, 1972. – 318 с.
2. Лоули. Дж. Факторный анализ как статистический метод. – М.: Наука, 1967. – 136 с.
3. Осипов. Г. Методы измерения в социологии. – М.: Наука, 2003. – 120 с.
4. Хартман. Г. Современный факторный анализ. – М.: Статистика, 1972. – 484 с.
5. Guttman L. Multiplle rectilinear prediction and the resolution into components // Psychometrika. – 1940. – №5. – 75-99 p.
6. Thurstone L.L. Multiple factor analysis. – Chicago, 1961. – 250 p.