Н.Н. Двоерядкина, Н.А. Чалкина

ФАКТОРНЫЙ АНАЛИЗ ПРИ ИССЛЕДОВАНИИ СТРУКТУРЫ ДАННЫХ

In article possibility of use of one of multidimensional methods of the analysis for studying of structure of the initial data on the basis of a correlation matrix is considered.

Собирая данные, исследователь руководствуется определенными гипотезами, информация относится к избранным предмету и теме исследования, но нередко представляет собой сырой материал, в котором нужно изучить структуру показателей, характеризующих объекты, а также выявить однородные группы объектов. Такая работа может быть осуществлена с помощью факторного анализа.

Идея метода факторного анализа состоит в сжатии матрицы признаков в матрицу с меньшим числом переменных, сохраняющую почти ту же самую информацию, что и исходная матрица, т.е. сконцентрировать исходную информацию, выражая большое число рассматриваемых признаков через меньшее число более емких внутренних характеристик явления, которые, однако, не поддаются непосредственному измерению. При этом предполагается, что наиболее емкие характеристики окажутся одновременно и наиболее существенными, определяющими.

Задачами факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения данных или как метод структурной классификации.

Материалом для факторного анализа служат корреляционные связи, а точнее, – коэффициенты корреляции Пирсона, которые вычисляются между переменными, включенными в обследование. Иными словами, факторному анализу подвергают корреляционные матрицы или матрицы интеркорреляций.

Главное понятие факторного анализа – фактор. Это искусственный статистический показатель, возникающий в результате специальных преобразований таблицы коэффициентов корреляции между изучаемыми признаками, или матрицы интеркорреляций.

Основные результаты факторного анализа выражаются в наборах факторных нагрузок и факторных весов.

Факторные нагрузки – это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором. Таблица факторных нагрузок содержит столько строк, сколько признаков, и столько столбцов, сколько факторов.

Для построения матрицы факторных нагрузок необходимо найти собственные числа и собственные векторы корреляционной матрицы. Нормированные координаты собственных векторов являются элементами матрицы факторных нагрузок.

Факторными весами называют количественные значения выделенных факторов для каждого из имеющихся объектов. Объекту с большим значением факторного веса присуща большая степень проявления свойств, определяемых данным фактором. Поэтому положительные

факторные веса соответствуют тем объектам, которые обладают степенью проявления свойств больше средней, а отрицательные – тем объектам, для которых степень проявления свойств меньше средней. Количество строк в таблице факторных весов совпадает с числом объектов, количество столбцов соответствует числу факторов.

Таким образом, данные о факторных нагрузках позволяют сформулировать выводы о наборе исходных признаков, отражающих тот или иной фактор, и об относительном весе отдельного признака в структуре каждого фактора. В свою очередь данные о факторных весах определяют ранжировку объектов по каждому фактору.

Набор методов факторного анализа в настоящее время достаточно велик, насчитывает десятки различных подходов и приемов обработки данных. В основе каждого метода факторного анализа лежит математическая модель, описывающая соотношения между исходными признаками и обобщенными факторами. Рассмотрим наиболее распространенные методы.

Центроидный метод. Основан на предположении, что каждый из исходных признаков X_i может быть представлен как функция небольшого числа общих факторов $F_1,F_2,\ldots,F_k$ и характерного фактора U_i:

$$X_i = \sum a_{ik} F_k + U_i \,. \qquad (1)$$

Факторы F построены так, чтобы наилучшим способом (с минимальной погрешностью) представить X. В этой модели «скрытые» переменные F_k называются общими факторами, а переменные U_i – специфическими («характерными», «уникальными») факторами.

При этом считается, что каждый общий фактор имеет существенное значение для анализа всех исходных признаков, т.е. фактор F_i – общий для всех X_i. В то же время изменения в специфическом факторе U_i воздействуют на значения только соответствующего признака X_i. Таким образом, специфический фактор U_i отражает ту специфику признака X_i, которая не может быть выражена через общие факторы.

Метод главных компонент. В основе модели для выражения исходных признаков через факторы здесь лежит предположение, что число общих факторов равно числу исходных признаков, а специфические факторы отсутствуют вообще:

$$X_i = \sum a_{ik} F_k \,. \qquad (2)$$

Уравнения определяют систему преобразования одних параметров в другие. Поскольку число факторов равно числу исходных параметров, задача искомого преобразования решается однозначно, т.е. факторные нагрузки определяются в этом методе однозначно.

Каждая из переменных F_i называется здесь i-й главной компонентой. Метод главных компонент состоит в построении факторов – главных компонент, каждый из которых представляет линейную комбинацию исходных признаков.

Первая главная компонента F_1 определяет такое направление в пространстве исходных признаков, по которому совокупность объектов (точек) имеет наибольший разброс (дисперсию). Вторая главная компонента F_2 строится с расчетом, чтобы ее направление было ортогонально направлению F_1 и она объясняла как можно большую часть остаточной дисперсии – и т.д., вплоть до m-й главной компоненты F_m. Так как выделение главных компонент происходит в убывающем порядке с точки зрения доли объясняемой ими дисперсии, то признаки, входящие в первую главную компоненту, оказывают максимальное влияние на дифференциацию изучаемых объектов.

Для определения количества факторов существует несколько критериев:

1. *Критерий Кайзера* предлагает отобрать только факторы, с собственными значениями, бóльшими 1.

2. *Критерий каменистой осыпи* является графическим методом, впервые предложенным Кэттелем (Cattell). Необходимо изобразить собственные значения корреляционной матрицы, расположенные в убывающем порядке в виде графика (по оси абсцисс порядковый номер числа, по оси ординат – его значение) (рис. 1).

Кэттель предложил найти такое место на графике, где убывание собственных значений корреляционной матрицы слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь» («осыпь» – геологический термин, обозначающий обломки горных пород, скапливающиеся в нижней части скалистого склона.

В соответствии с этим критерием можно оставить в этом примере два или три фактора.

Первый критерий (*критерий Кайзера*) сохраняет иногда слишком много факторов, в то время как второй (*критерий каменистой осыпи*) – слишком мало; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. Обычно исследуется несколько решений с бóльшим или меньшим числом факторов, и затем выбирается одно, наиболее интерпретируемое.

Результаты факторного анализа будут успешными, если удается дать содержательную интерпретацию выявленных факторов, исходя из смысла показателей, которые их характеризуют. Данная стадия работы весьма ответственна: она требует от исследователя четкого представления о содержательном смысле показателей, привлеченных для анализа и на основе которых выделены факторы.

Поэтому при предварительном тщательном отборе показателей для факторного анализа следует руководствоваться их содержательным смыслом, а не стремлением к включению в анализ как можно большего их числа.

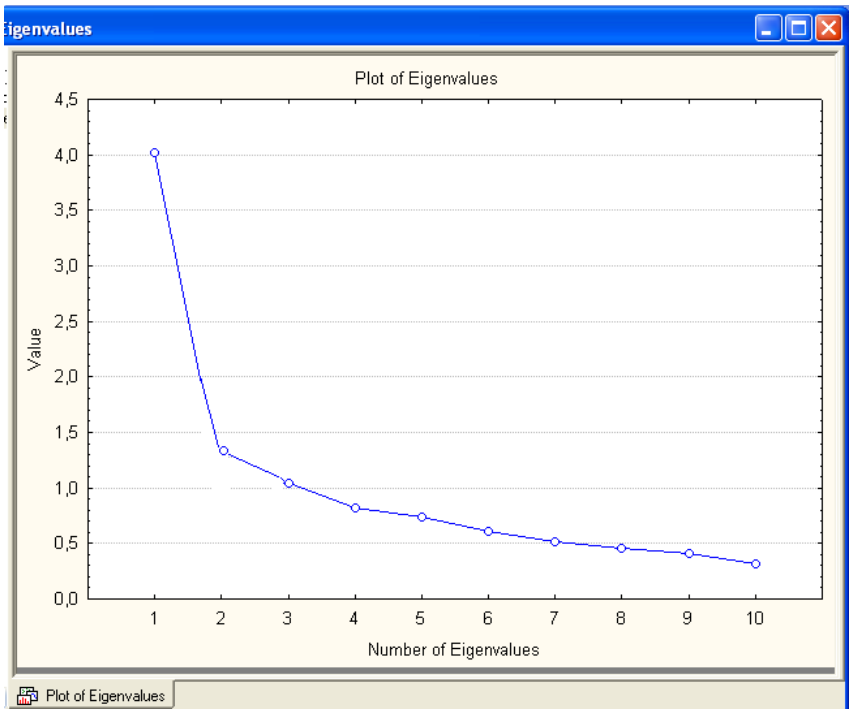

Рис. 1. Критерий каменистой осыпи.

Алгоритмы факторного анализа отличаются трудоемкостью, их полное выполнение возможно при условии использования технических средств.

В программе *Statistica* факторный анализ проводится с помощью модуля *Multivariate Exploratory Techniques*. Переход к процедуре факторного анализа осуществляется посредством пункта *Factor Analysis*. В открывшемся окне необходимо указать тип формата данных (*Raw Data* или *Correlation Matrix*), переменные для анализа и нажать *OK*. Далее выбрать конкретный метод факторизации корреляционной матрицы – *Extraction method* (например, *Principal components* – метод главных компонент), максимальное число факторов (*maximum of factors*), на первом этапе обычно равное числу переменных, минимальное собственное число (*minimum eigenvalue)*, равное 0, и нажать *OK* (рис. 2).

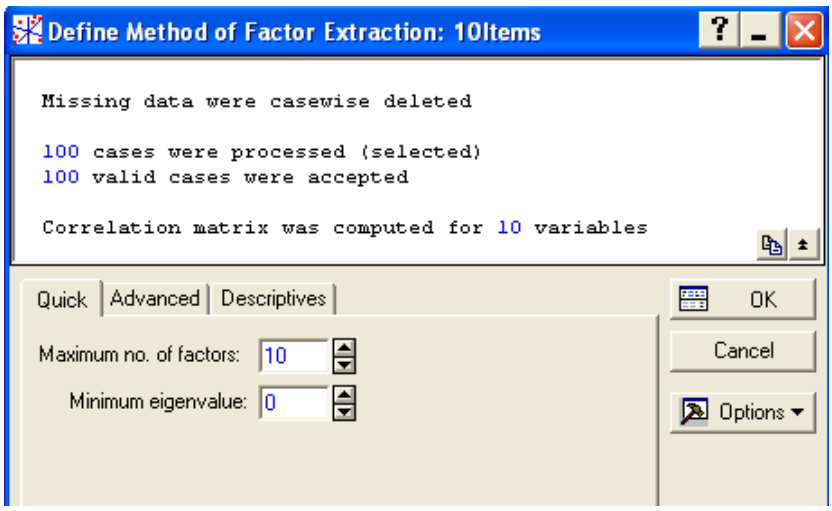

Рис. 2. Диалоговое окно факторного анализа.

После этого появится окно результатов факторного анализа, позволяющее просмотреть собственные значения корреляционной матрицы *(Eigenvalues)*, графическое изображение собственных чисел *(Scree plot)* матрицу факторных нагрузок (*Summary: Factor loading)* и факторных весов (*Factor scores).*

Рассмотрим на примере, как методами факторного анализа сконцентрировать исходную информацию, содержащуюся в трех переменных *X*, *Y*, *Z* в одной латентной характеристике *f*.

Для нахождения латентного фактора необходимо знать коэффициенты корреляции Пирсона для исходных данных (таблица).

Матрица коэффициентов корреляции

Переменные	Коэффициенты корреляции		
	X	*Y*	*Z*
X	1	0,06	0,15
Y	0,06	1	0,30
Z	0,15	0,30	1

Для построения матрицы факторных нагрузок необходимо найти собственные числа корреляционной матрицы *R*, решив уравнение:

$$\left|R-\lambda E\right|=0\,. \qquad (3)$$

Для полученной корреляционной матрицы *R*, представленной в таблице, собственные числа $\lambda_1=1{,}362$, $\lambda_2=0{,}953$, $\lambda_3=0{,}685$.

Согласно критерию Кайзера значимыми являются только факторы с собственными значениями, бóльшими 1. Нормированные координаты собственного вектора $v_1 = (-0{,}385; -0{,}626; -0{,}678)$, соответствующие собственному числу $\lambda_1 = 1{,}362$, находятся путем решения системы уравнений:

$$(R - \lambda_1 \cdot E) \cdot v = 0 \tag{4}$$

и последующей нормировке по формуле

$$v_{i\,норм} = \frac{v_i}{\sqrt{\sum_i v_i^2}}. \tag{5}$$

Элементы матрицы факторных нагрузок

$$A = v_1^T \cdot \sqrt{\lambda_1} = \begin{pmatrix} -0.45 \\ -0.731 \\ -0.792 \end{pmatrix} \tag{6}$$

являются коэффициентами корреляции между исходными переменными *X*, *Y*, *Z* и латентным фактором *f*. Их абсолютные значения показывают наличие достаточно значимой линейной связи между исходными переменными и найденным фактором, – следовательно, этот фактор можно рассматривать как переменную, связывающую между собой исходные переменные *X*, *Y*, *Z*.

Благодаря факторному анализу исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации и обеспечивает снижение размерности признакового пространства. Этот метод позволяет, учитывая эффект существенной многомерности данных, лаконичнее и проще объяснять многомерные структуры и характер взаимосвязей между ними. Сжатие информации получается за счет того, что число используемых факторов – новых единиц измерения – значительно меньше, чем было исходных признаков.

1. Дубров А.М. Многомерные статистические методы: Учеб. / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 2005. – 352 с.

2. Сошникова Л.А. Многомерный статистический анализ в экономике: Учеб. пос. для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер. – М.:ЮНИТИ-ДАНА, 2003. – 598 с.