УДК 004.94

С.В. Лучкова, Т.О. Перемитина

Повышение наглядности отображения результатов факторного анализа

Рассмотрен метод, отображающий структурную связь данных и позволяющий увеличить наглядность численных результатов факторного анализа. Описаны метод факторного анализа с вращением и классификацией, применяемый для анализа многомерных данных, а также метод построения дендрограммы, основанный на кластерном анализе, увеличивающий наглядность результатов. Приведены экспериментальные исследования анализа данных о валютных парах с применением описанных методов.

Ключевые слова: факторный анализ, классификация, дендрограмма, увеличение наглядности отображения результатов.

В научных исследованиях приходится иметь дело с объектами различной природы. Для описания свойств и поведения таких объектов требуются большие объемы разнородной информации. Выявление закономерностей, построение моделей и расшифровка результатов предполагают умение извлекать на научно-методической основе, требуемую информацию из наблюдений природных объектов, а также исследований фактических значений параметров, характеризующих эти объекты [1].

При исследовании многомерных объектов часто возникают вопросы о повышении наглядности рассматриваемых объектов и получаемых результатов исследования. Повышение наглядности связано как со сжатием исследуемых данных, так и с графическим представлением результатов. Для решения вопроса о сжатии данных можно воспользоваться методом факторного анализа, который позволяет это сделать. Факторный анализ является многомерным статистическим методом и в настоящее время широко используется в различных областях исследовательской деятельности – в экономике, статистике, нейрофизиологии, политологии, психологии и т.д. В исследовании сложных объектов факторный анализ позволяет выявить скрытые взаимосвязи объекта и их структурные особенности. Для решения вопроса о повышении наглядности результатов факторного анализа можно воспользоваться такими графическими методами: классификация, построение гистограмм и структурных связей с помощью дендрограммы.

Цель данной работы – описать факторный анализ с вращением, применяемый для анализа многомерных данных, метод построения дендрограммы, который отображает структурную связь исследуемых данных и повышает наглядность факторного анализа, а также показать применение описанных методов на примере анализа данных о валютных парах.

Алгоритм факторного анализа с вращением. Факторный анализ является одним из разделов многомерного статистического анализа. Он основан на нормальном распределении, то есть каждый из используемых признаков изучаемого объекта должен иметь нормальный закон распределения. Факторный анализ исследует внутреннюю структуру ковариационной и корреляционной матриц системы признаков изучаемого объекта [2].

Пусть в изучаемом объекте отобрано N записей. В каждой из них измерены значения K параметров и получены значения случайных многомерных нормально распределенных величин. Эти значения случайных многомерных величин обусловлены различными причинами, которые называются факторами. Предполагается, что число этих факторов всегда меньше, чем число K измеряемых параметров изучаемого объекта. Эти факторы являются скрытыми, их нельзя непосредственно измерить, и поэтому они представляются гипотетическими. Однако имеются методы их выявления, которые и составляют сущность факторного анализа [3, 4].

Алгоритм факторного анализа с вращением будет выглядеть так:

Вход: таблица наблюдений без пропусков (с восстановленными данными).

Шаг 0. Загружаем данные и выбираем анализируемые параметры.

Шаг 1. Нормируем данные.

Шаг 2. Рассчитываем матрицы ковариации и корреляции.

Шаг 3. Вычисляем собственные числа и собственные вектора (применяя разложение Холецкого, *LU*-разложение [5]).

Шаг 4. Рассчитываем факторы и вычисляем долю влияния каждого из факторов на значения параметров.

Шаг 5. Выявляем наиболее значимые факторы.

Шаг 6. Воссоздаем в факторном координатном пространстве изучаемый объект (отображаем на пространственном графике).

Шаг 7. Применяем ортогональное вращение методом «Варимакс» [6] для увеличения критерия качества каждого фактора.

Шаг 8. Классифицируем данные, применяя метод *K*-средней кластеризации [7] для разделения данных на классы. Если достигнуто условие завершения анализа, *Шаг 9, иначе Шаг 0.*

Шаг 9. Выводим результат.

Выход: таблицы с рассчитанными данными и графические отображения.

Алгоритм построение дендрограммы. Дендрограмма – это иерархическое дерево или граф без циклов, построенный по определенной матрице, отражающей меру близости исследуемых элементов. Дендрограмма позволяет отобразить графически взаимные связи между объектами из заданного множества. Для создания дендрограммы требуются матрица сходства (или различия), которая определяет уровень сходства между парами объектов, и метод построения, который определяет способ пересчёта матрицы сходства (различия) после объединения (или разделения) очередных двух объектов в кластер [8-10]. В работах по кластерному анализу описано несколько метрик:

1) одиночная связь (расстояние ближайшего соседа);
2) полная связь (расстояние наиболее удаленных соседей);
3) средняя связь (среднее расстояние между всеми парами объектов);
4) центроидный метод (расстояние между центрами тяжести объектов);
5) метод Уорда (мера – это прирост суммы квадратов расстояний до центров кластеров, получаемый в результате объединения объектов) [8].

Так как в нашем случае мы используем построение дендрограммы для улучшения наглядности результатов факторного анализа, то за матрицу сходства возьмем корреляционную матрицу и представим табличный результат ковариационной матрицы в виде дендрограммы.

Для построения дендрограммы воспользуемся центроидным методом.

Алгоритм построения дендрограммы будет выглядеть так:

Вход: корреляционная матрица

Шаг 1. Рассчитываем матрицу весов.

Шаг 2. Находим два элемента для объединения.

Шаг 3. Делаем перерасчет матрицы расстояний. Если достигнуто условие завершения, то *Шаг 4*, иначе *Шаг 2*.

Шаг 4. Выводим результат.

Выход: графическое отображение корреляционной матрицы в виде графа-дерева (дендрорамы).

Эксперимент. Для анализа была сформирована выборка данных о 10 валютных парах (табл. 1) за полугодовой период (08.2012–03.2013). Пара – это кросс-курс, который образуется отношением каждой валюты в паре к доллару США [11, 12].

Таблица 1

Исследуемые данные

№ п/п	Обозначение	Характеристика
1	AUDCHF	Австралийский доллар и швейцарский франк
2	AUDUSD	Австралийский доллар и доллар США
3	CHFJPY	Швейцарский франк и японская иена
4	EURJPY	Евро и японская иена
5	EURUSD	Евро и доллар США
6	GBPUSD	Фунты стерлингов и доллар США
7	NZDUSD	Новозеландский доллар и доллар США
8	USDCAD	Доллар США и канадский доллар
9	USDCHF	Доллар США и швейцарский франк
10	USDJPY	Доллар США и японская иена

Факторный анализ. Для выявления структуры данных рассмотрим значения корреляционной таблицы (табл. 2).

Таблица 2

Корреляционная матрица

Признаки	Коэффициенты корреляции				
	AUDCHF	AUDUSD	CHFJPY	EURJPY	EURUSD
AUDCHF	1	0,423	–0,540	–0,522	–0,745
AUDUSD	0,423	1	–0,048	–0,046	0,117
CHFJPY	–0,540	–0,048	1	0,997	0,797
EURJPY	–0,522	–0,046	0,997	1	0,806
EURUSD	–0,745	0,117	0,797	0,806	1
GBPUSD	0,014	0,451	–0,575	–0,585	–0,073
NZDUSD	–0,435	0,380	0,710	0,715	0,829
USDCAD	–0,150	–0,371	0,634	0,632	0,193
USDCHF	0,832	–0,136	–0,547	–0,528	–0,869
USDJPY	–0,414	–0,086	0,983	0,984	0,689

Продолжение табл. 2

Признаки	Коэффициенты корреляции				
	GBPUSD	NZDUSD	USDCAD	USDCHF	USDJPY
AUDCHF	0,014	–0,435	–0,150	0,832	–0,414
AUDUSD	0,451	0,380	–0,371	–0,136	–0,086
CHFJPY	–0,575	0,710	0,634	–0,547	0,983
EURJPY	–0,585	0,715	0,632	–0,528	0,984
EURUSD	–0,073	0,829	0,193	–0,869	0,689
GBPUSD	1	–0,068	–0,877	–0,272	–0,696
NZDUSD	–0,068	1	0,154	–0,694	0,631
USDCAD	–0,877	0,154	1	0,092	0,723
USDCHF	–0,272	–0,694	0,092	1	–0,384
USDJPY	–0,696	0,631	0,723	–0,384	1

Довольно логично, что валюта, входящая в одну пару, будет хорошо коррелировать с другой парой, в которую она также входит. Однако это не единственная зависимость, которую можно увидеть с помощью корреляционной матрицы.

Так, пара CHFJPY (швейцарский франк – японская иена) коррелирует с парой EURJPY (евро – японская иена) и USDJPY (доллар США – японская иена). Объяснение такой корреляции довольно простое: как швейцарский франк, так и японская иена – это валюты двух небольших государств. Причем франк тесно связан с евро и поэтому зачастую динамика пары CHFJPY копируется движением EURJPY. Однако можно заметить, что данная пара также коррелирует с парой USDCAD, EURUSD и NZDUSD и обладает обратной корреляцией с AUDCHF, USDCHF, GBPUSD. Это объясняется тем, что динамика курса подвержена влиянию множества сторонних факторов, которые не всегда связаны с тем, что происходит в Швейцарии и Японии [13].

Динамика же пары EURJPY определяется тем, что иена является валютой-убежищем, то есть инвесторы всего мира начинают ее скупать сразу же, как экономика кажется мрачной или туманной, и иена дорожает относительно доллара, который также является валютой-убежищем. На динамику влияют и усилия монетарных властей Японии – как только иена укрепляется сильней, чем власти могут допустить, Центробанк сразу готов прибегнуть к валютным интервенциям. А отсюда и объяснение корреляций данной пары с EURUSD, NZDUSD и USDCAD [13].

Интересными также являются связи для самой распространенной пары EURUSD: положительная связь наблюдается с NZDUSD, EURJPY, CHFJPY и USDJPY, а отрицательные с USDCHF и AUDCHF. Это довольно спокойная пара, динамика изменений практически всегда плавная, так как манипулирование с ее курсом затруднительно даже для крупных инвесторов. Однако для пары характерна неоднократная смена тенденции в течение дня, зачастую из-за новостей, в равной мере приходящих как из США, так и из Европы. А отрицательная связь с парами AUDCHF и USDCHF объясняется зависимостью данных пар от динамики доллара США, и к тому же важной особенностью пары USDCHF считается зеркальная корреляция с EURUSD [13].

Анализ корреляционной матрицы признаков позволяет выявить структуру взаимосвязей, которую графически удобно представлять в виде иерархической дендрограммы (рис. 1), которая подчеркивает выделенные корреляционные взаимосвязи.

Согласно графику (рис. 2) для анализа достаточно использовать 3–4 фактора, которые полностью объяснят структуру и зависимости в исследуемых данных.

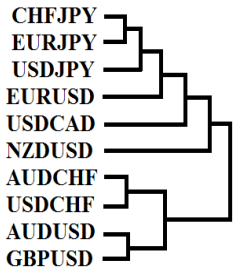

Рис. 1. Дендрограмма (структура взаимосвязи валютных пар)

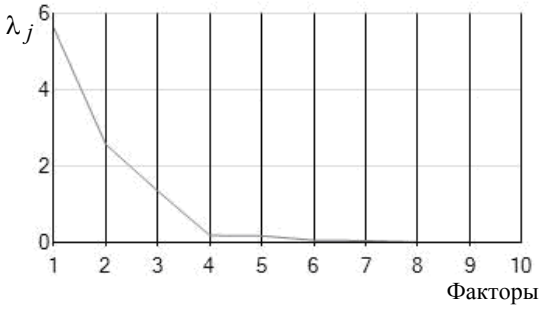

Рис. 2. Зависимость собственных чисел и факторов

Рассмотрим признаковую структуру выбранных 4 факторов (табл. 3).

Таблица 3

Факторные нагрузки, собственные значения и веса признаков

Признаки / факторы	F_1	F_2	F_3	F_4
AUDCHF	**0,670**	0,311	**0,668**	0,045
AUDUSD	0,089	–0,546	**0,813**	–0,168
CHFJPY	**–0,978**	0,108	0,117	–0,022
EURJPY	**–0,976**	0,116	0,132	0,012
EURUSD	**–0,878**	–0,428	–0,047	0,011
GBPUSD	0,494	**–0,844**	–0,078	–0,112
NZDUSD	**–0,768**	–0,427	0,319	0,245
USDCAD	–0,585	**0,731**	0,053	–0,248
USDCHF	**0,662**	**0,683**	0,251	0,132
USDJPY	**–0,934**	0,272	0,182	0,004
Собственные значения λ_j	5,623	2,572	1,347	0,183
Вес факторов, %	56,230	25,720	13,470	1,830

Примечание. Коэффициенты являются значимыми ($\beta = 0,05$) при их абсолютном значении не менее 0,624.

Анализ признаковой структуры фактора F_1 показывает, что нагрузка этого фактора значимо определяется парами AUDCHF (0,670) и USDCHF (0,662), а также имеет значимую отрицательную связь с CHFJPY (–0,978), EURJPY (–0,976), USDJPY (–0,934), EURUSD (–0,878) и NZDUSD (–0,768).

Структура же фактора F_2 определяется нагрузкой пар USDCAD(0,731) и USDCHF (0,683), но обладает отрицательной связью с парой GBPUSD (–0,844), а структура фактора F_3 определяется положительной нагрузкой пар AUDUSD (0,813) и AUDCHF (0,668).

Пространственная структура данных (рис. 3) на основе 1-го и 4-го фактора с применением k-средней кластеризации выделяет 3 класса валютных пар. Согласно разделению на классы мы получили, что в первый класс вошли 5 валютных пар: AUDCHF, AUDUSD, NZDUSD, USDCAD, USDCHF, во второй класс – 2 пары: EURUSD, GBPUSD и в третий класс – 3 пары: CHFJPY, EURJPY и USDJPY.

С помощью процедуры вращения выявим наиболее интерпретируемые факторы. После вращения 1-го и 4-го фактора на 40° (табл. 4) структура данных (рис. 4) выстраивается в более простую структуру и показывает, что для характеристики данных достаточно 3 факторов, а 4-й фактор избыточен.

Проведенное исследование показывает, что факторный анализ позволяет выявить взаимосвязи исследуемых валютных пар и провести анализ признаковой и факторной структуры. А за счет комбинации факторного анализа с дендрограммой и классификацией, мы можем разделить исследуемые данные на группы, выявить в них главные валютные пары, рассмотреть их особенности и характерные черты.

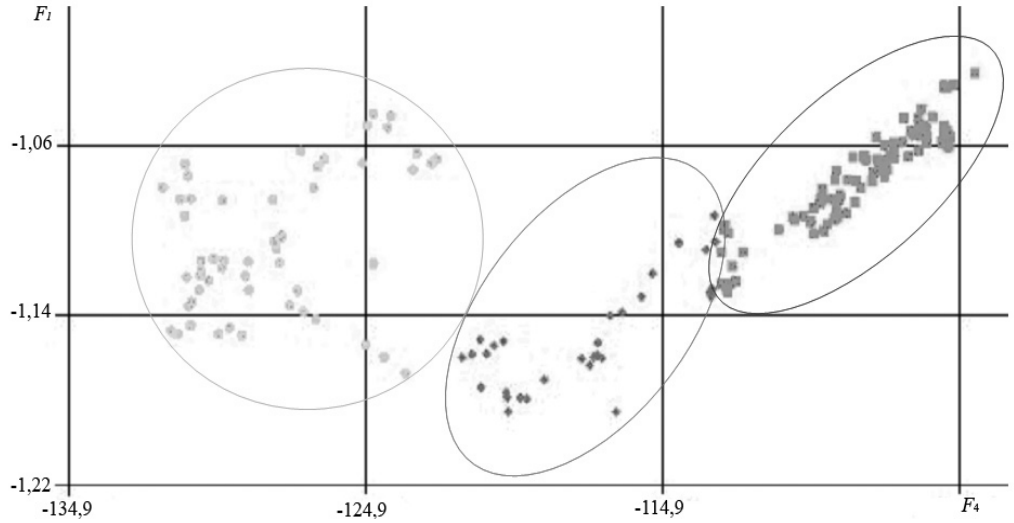

Рис. 3. Пространственная структура данных с разделением на классы методом *k*-средней кластеризации

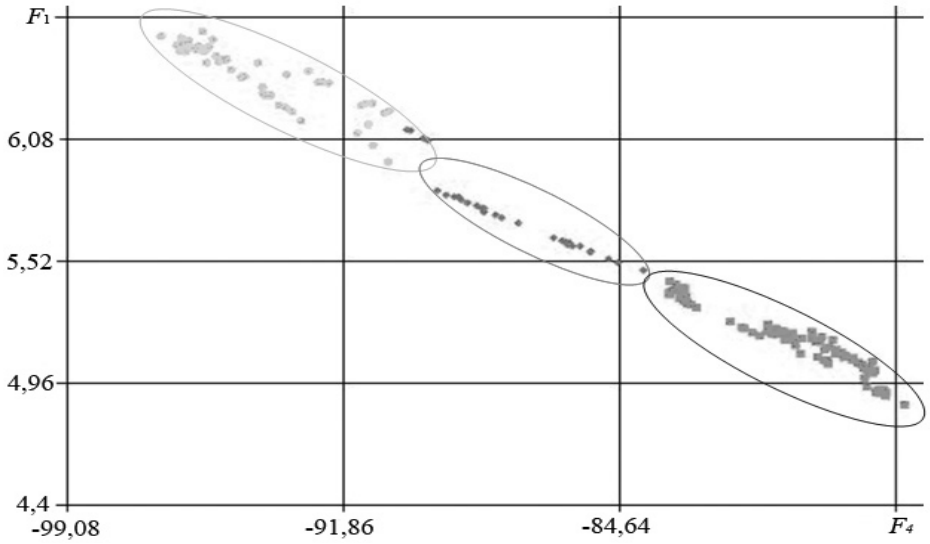

Рис. 4. Пространственная структура данных после вращения

Таблица 4

Факторные нагрузки, собственные значения и веса признаков после вращения

Признаки / факторы	F_1	F_2	F_3	F_4
AUDCHF	1,078	0,164	0,430	0,027
AUDUSD	0,084	–0,544	0,819	–0,171
CHFJPY	–0,924	0,108	0,118	–0,022
EURJPY	–0,300	0,288	0,534	0,039
EURUSD	–0,830	–0,427	–0,047	0,011
GBPUSD	0,467	–0,840	–0,078	–0,114
NZDUSD	–0,726	–0,425	0,321	0,250
USDCAD	–0,553	0,728	0,053	–0,253
USDCHF	0,626	0,680	0,253	0,135
USDJPY	–0,883	0,271	0,183	0,004
Собственные значения λ_j	5,026	2,551	1,366	0,190
Вес факторов, %	50,260	25,510	13,660	1,900
Критерий «Варимакс» до вращения	0,246	0,191	0,117	0,018
Критерий «Варимакс» после вращения	**0,250**	**0,190**	**0,118**	**0,019**

Заключение. В работе рассмотрен метод построения дендрограммы, отображающий структурную связь данных и позволяющий увеличить наглядность численных результатов факторного анализа. Описан метод факторного анализа с вращением и классификацией и *k*-средней классификаци-

ей для проведения анализа, включающего в себя выявления взаимосвязей различных признаков объектов, главных действующих факторов, анализа их признаковой структуры и анализа факторной структуры. Более того, метод позволяет воссоздать в факторном координатном пространстве облик изучаемого объекта и указать его характерные признаки и отличительные особенности.

Литература

1. Лучкова С.В. Применение программного комплекса «Нечеткая система на основе эволюционной стратегии» для задачи импутирования / С.В. Лучкова, Т.О. Перемитина // Информационные технологии. – 2013. – № 2. – С. 47–50.

2. Лоули Д. Факторный анализ как статистический метод / Д. Лоули, А. Максвелл. – М.: Мир, 1967. – 144 с.

3. Белонин М.Д. Факторный анализ в геологии / М.Д. Белонин, В.А. Голубева, Г.Т. Скублов. – М.: Недра, 1982 . – 269 с.

4. Ким Дж. Факторный, дискриминантный и кластерный анализ / Дж. Ким, Ч.У. Мюллер. – М.: Финансы и статистика, 1989. – 215 с.

5. Trefethen Lloyd N. Numerical linear algebra / Lloyd N. Trefethen, David Bau. – Philadelphia, USA: Society for Industrial and Applied Mathematics, 1997. – 263 p.

6. Харман Г. Современный факторный анализ. – М.: Статистика, 1972. – 483 с.

7. Vance F. Clustering and the continuous k-Means Algorithm // Los Alamos Science. – 1994. – № 22. – P. 138–144.

8. Гусев В.А. Алгоритм построения иерархической дендрограммы кластер-анализом в геолого-геохимических приложениях / В.А. Гусев, И.К. Карпов, А.И. Киселев // Изв. АН СССР. Сер. геологическая. – 1974. – № 8. – С. 61–67.

9. Мартюшева П.В. Кластерный анализ как инструмент менеджмента качества для обработки социологических опросов на промышленном предприятии / П.В. Мартюшева, О.В. Стукач // Доклады ТУСУР. – 2007. – № 1 (15). – С. 71–76.

10. Малинин П.В. Иерархический подход в задаче идентификации личности по голосу с помощью проекционных методов классификации многомерных данных / П.В. Малинин, В.В. Поляков // Доклады ТУСУР. – 2010. – № 1 (21), ч. 1. – С. 128–130.

11. Кросс-курсы и валютные пары [Электронный ресурс]. – Режим доступа: http://www.delta-stock.com/russia/resources/seminar.asp#crr, свободный (дата обращения: 18.06.2013).

12. Форекс-символы. Валютные символы с расшифровкой значения и краткой характеристикой [Электронный ресурс]. – Режим доступа: http://www.mt5.com/ru/forex_symbols, свободный (дата обращения: 18.06.2013).

13. Торговые инструменты. Валютные символы [Электронный ресурс]. – Режим доступа: http://www.forexcent.com/rus/help/trading_symbols.html, свободный (дата обращения: 18.06.2013).

Лучкова Софья Викторовна
Ассистент каф. автоматизации обработки информации (АОИ) ТУСУРа,
аспирантка Института химии нефти СО РАН
Тел.: 8-923-406-89-27
Эл. почта: sonetta27@gmail.com

Перемитина Татьяна Олеговна
Канд. техн. наук, доцент каф. АОИ ТУСУРа
Тел.: 8-903-954-69-25
Эл. почта: peremitinat@mail.ru

Luchkova S.V., Peremitina T.O.

Increasing the visibility of the display of factor analysis results

Thy paper considers a method of showing the structural relationship of data and increasing the visibility of the numerical results of factor analysis. We described the factor analysis method with varimax and classification, applied to the analysis of multidimensional data and a dendrogram method, increasing the visibility. The experimental research results are presented.

Keywords: factor analysis, classification, dendrogramma, increasing the visibility of display of the results.