Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

«Национальный исследовательский университет

«Высшая школа экономики»

Факультет Бизнес-информатики

Отделение Прикладной математики и информатики

Кафедра Высшей математики на факультете экономики

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

на тему

**Снижение размерности многомерных показателей смешанной структуры**

Выполнил студент группы 471

Шалимова Юлия Андреевна

Научный руководитель:

доцент, кандидат наук,

Горяинова Елена Рудольфовна

Москва 2013

Оглавление

[Аннотация 3](#_Toc388370842)

[Abstract 4](#_Toc388370843)

[Ключевые слова 4](#_Toc388370844)

[Введение 4](#_Toc388370845)

[Глава 1 (теоретическая) 7](#_Toc388370846)

[1.1 Коротко о методе главных компонент. 8](#_Toc388370847)

[1.2 Факторный анализ. 11](#_Toc388370848)

[1.2.1 Модель факторного анализа 11](#_Toc388370849)

[1.2.2 Сравнение факторного анализа с методом главных компонент. 13](#_Toc388370850)

[1.2.3 Интерпретация матрицы нагрузок. 13](#_Toc388370851)

[1.2.4 Неоднозначность решения задачи факторного анализа. Выбор вращения факторов. 15](#_Toc388370852)

[1.2.5 Метод максимального правдоподобия 16](#_Toc388370853)

[1.2.6 Выбор числа общих факторов. 20](#_Toc388370854)

[1.3 Выводы по главе 23](#_Toc388370855)

[Глава 2 (техническая) 24](#_Toc388370856)

[2.1 Итерационный метод поиска оценок максимального правдоподобия для матрицы нагрузок и матрицы остаточных дисперсий 24](#_Toc388370858)

[2.2 Выводы по главе 26](#_Toc388370859)

[Глава 3 (экспериментальная) 26](#_Toc388370860)

[3.1 Адаптация метода максимального правдоподобия к нелинейным зависимостям. 26](#_Toc388370862)

[3.2 Генерация тестовых данных. 30](#_Toc388370863)

[3.3 Измерение эффективности метода 37](#_Toc388370864)

[3.4 Определение числа факторов. 39](#_Toc388370865)

[3.5 Результаты применения методов к искусственно смоделированным данным 40](#_Toc388370866)

[3.6 Демонстрация эффективности метода определения числа факторов, основанного на величине приращения объясняемого среднеквадратического отклонения 44](#_Toc388370867)

[3.7 Результаты и выводы по главе 48](#_Toc388370868)

[Глава 4 (исследовательская). 50](#_Toc388370869)

[4.1 Выбор реальных данных. 50](#_Toc388370871)

[4.2 Результаты применения методов к реальным данным. 53](#_Toc388370872)

[4.3 Выводы по главе. 63](#_Toc388370873)

[Заключение 64](#_Toc388370874)

[Приложение 66](#_Toc388370875)

[Список литературы 67](#_Toc388370876)

# Аннотация

С ростом научно-технического прогресса решение прикладных задач требует обработки и хранения все больших массивов данных, поэтому методы снижения размерности становятся все более востребованными. Однако для некоторых структур данных традиционные методы снижения размерности оказываются неэффективными. Данный дипломный проект нацелен на повышение эффективности одного из традиционных методов факторного анализа - метода максимального правдоподобия – для снижения размерности пространства признаков, связанных нелинейными типами зависимости.

В этой работе объясняются теоретические основы факторного анализа, и, в том числе, его традиционного метода максимального правдоподобия. Предлагается способ адаптации этого метода к работе с нелинейными типами зависимости, проверяется эффективность этой адаптации как на искусственно смоделированных, так и на реальных данных.

# Abstract

With a rapid technological change, solving practical tasks requires to process increasing amounts of data. That is why dimensionality reduction methods are assuming greater importance nowadays. However, these methods lose theoretical validity or even efficiency when applied to certain data structures, because they set significant limits on the investigated system. Namely, in these methods it is assumed that the initial features of the system are linearly interdependent.

The aim of this graduation project is to adapt one of the traditional methods of factor analysis - maximum-likelihood method – to the cases of nonlinearly dependent features.

This paper provides an overview of the theoretical basis of factor analysis and its maximum-likelihood method. The adaptations of this method to nonlinear relationships between features are proposed. The traditional and the modified methods are compared in terms of efficiency on both modeled and real data.

# Ключевые слова

Факторный анализ, метод главных компонент, метод максимального правдоподобия, нагрузочная матрица, варимакс, квартимакс, коэффициент ранговой корреляции Спирмена, коэффициент Крамера.

# Введение

При изучении объектов, каждый из который характеризуется большим количеством признаков, часто возникает необходимость описать эти объекты значительно меньшим числом признаков, сохранив при этом как можно больше важной информации об объектах. Целью этого может являться следующее:

* Образование наиболее информативных интегративных характеристик, таких как уровень жизни, качество продукции, успешность компании, состояние здоровья, уровень физической подготовки, размер одежды и т.д.
* Визуализация изучаемых объектов, с помощью проектирования значений их признаков на специально подобранную прямую, плоскость или трехмерное пространство.
* Сокращение объема хранимых данных.

Список можно продолжать.

Избыточность признаков может быть вызвана тем, что некоторые признаки отражают одну и ту же информацию, как, например, масса, измеренная в килограммах, и масса, измеренная в фунтах. Или же некоторые признаки могут отражать информацию, которая не имеет никакого отношения к теме исследования. В этих двух случаях число признаков может быть уменьшено за счет простого отбора. В английской литературе такая форма сокращения размерности носит название «feature selection».

Если избыточность обусловлена тем, что многие признаки находятся в сильной зависимости друг от друга, то обычно целесообразнее применить то, что называется «feature extraction», а именно - подвергнуть старые признаки таким преобразованиям, чтобы получить на выходе меньшее число новых признаков, потеряв при этом минимум информации.

В этой работе пойдет речь о факторном анализе – совокупности методов, осуществляющих уменьшение размерности согласно последнему из этих двух принципов. Большое преимущество факторного анализа состоит в том, что его модель интуитивно понятна исследователю, а методы – просты в реализации. Однако, факторный анализ накладывает весомое ограничение на изучаемую систему: он требует, чтобы признаки были линейно зависимы. Таким образом, методы факторного анализа становятся теоретически необоснованными или и вовсе неприменимыми для решения задач, в которых признаки связаны нелинейной зависимостью. Но, принимая во внимание вышеперечисленные достоинства факторного анализа, не хотелось бы совсем отказываться от его методов и в этом случае.

Таким образом, ***цель*** данного проекта - адаптировать один из методов факторного анализа - метод максимального правдоподобия – для работы с нелинейно зависимыми признаками. Поскольку в реальной жизни зачастую никакое предположение о типе зависимости между признаками не представляется обоснованным, то достижение этой цели должно означать повышение эффективности традиционного метода правдоподобия для решения реальных практических задач.

Для достижения этой цели перед автором этой работы стоят следующие основные задачи:

1. Изучение и программная реализация метода максимального правдоподобия.
2. Внесение в традиционный метод корректировок, необходимых для повышения эффективности его работы в случае нелинейно зависимых признаков.
3. Проверка корректности работы разработанных алгоритмов на искусственно смоделированных данных.
4. Сравнение эффективности адаптированных и традиционного методов на моделированных данных, связанных разными типами зависимости.
5. Применение методов к реальным данным с целью проверки и сравнения их эффективности для решения практических и научно-исследовательских задач.

Таким образом, ***объект*** данного исследования - это метод максимального правдоподобия и разработанные автором его модификации. ***Предметом*** исследования является эффективность этих методов для различных типов зависимостей между признаками (как сама по себе, так и в сравнении друг с другом). ***Метод*** исследования – это применение традиционного и модифицированных алгоритмов как к моделированным, так и к реальным данным с последующим анализом и сравнением результатов.

Данная работа имеет следующую структуру. В первых двух главах собран и проанализирован материал по теме данного исследования. В первой главе представлены теоретические основы факторного анализа и его традиционного метода максимального правдоподобия. Во второй главе рассказывается о компьютерной реализации этого метода. Уточним, что все результаты, представленные в первых двух главах, не являются плодами работы автора данного дипломного проекта. Последние две главы посвящены исследованию, проведенному непосредственно самим автором. В третьей главе описаны корректировки, внесенные автором в традиционный метод максимального правдоподобия, чтобы сделать его применимым к нелинейно зависимым признакам. В этой же главе продемонстрированы результаты проверки и сравнения эффективности традиционного и модифицированного методов на искусственно смоделированных данных. В последней – четвертой главе представлены и проанализированы результаты применения разработанных алгоритмов к реальным данным и сделаны выводы касательно эффективности традиционного и модифицированных методов для решения реальных практических задач.

# Глава 1 (теоретическая)

В этом разделе буду представлены теоретические основы факторного анализа, а так же его сравнение с родственным ему методом главных компонент. Автором было принято решение начать именно с метода главных компонент, поскольку его модель схожа с моделью факторного анализа, но принцип работы несколько проще. Таким образом, предварительное ознакомление с методом главных компонент способствует лучшему проникновению в суть факторного анализа.

## Коротко о методе главных компонент.

Метод главных компонент преобразует  старых признаков в  некоррелированных новых, а затем отбрасывает те, которые обладают наименьшей дисперсией. За счет этого и достигается снижение размерности.

Пусть центрированные случайные величины  - исходные признаки,  - их ковариационная матрица.  обычно неизвестна, но может быть оценена выборочной ковариационной матрицей.  - симметрична, следовательно, все ее собственные числа вещественны. Обозначим их  в порядке убывания. Предположим, что все собственные числа различны и положительны. Для большинства практических задач это предположение обычно верно. Более того, из симметричности  следует, что в  существует базис из собственных векторов этой матрицы. Обозначим его . Здесь  - собственный вектор единичной длины, соответствующий -му по величине собственному числу матрицы . В этом базисе  будет иметь вид , причем , где  - матрица, столбцами которой являются определенные выше .

***Определение***: Назовем линейную комбинацию  случайных величин  ***линейной комбинацией единичной нормы***, если .

Наша цель – найти  линейных комбинаций единичной нормы случайных величин , чтобы для каждого  -я линейная комбинация была некоррелирована с предыдущими  линейными комбинациями и обладала как можно большей дисперсией. Такие линейные комбинации называются главными компонентами системы случайных величин .

Рассмотрим случайный вектор  такой, что . Пусть  - его ковариационная матрица.  - центрированный случайный вектор, следовательно,  тоже центрирован. Тогда . Таким образом, мы получили, что  некоррелированы и .

Докажем, что  и есть искомые главные компоненты. Покажем, что  является линейной комбинацией случайных величин  с наибольшей дисперсией. Пусть  - произвольная линейная комбинация единичной нормы элементов . Из определения вектора  следует, что , где . При этом  также имеет единичную длину, поскольку при ортогональном преобразовании длина вектора остается прежней. Вычислим дисперсию . Поскольку ,



Вспомним что . Поэтому  максимальна при . А это и означает, что  является линейной комбинацией единичной нормы с наибольшей дисперсией.

Аналогичными рассуждениями доказывается, что  имеет наибольшую дисперсию среди всех возможных линейных комбинаций единичной нормы элементов , некоррелированных с . И так далее до .

Таким образом, мы доказали, что  является -ой главной компонентой системы случайных величин , и ее дисперсия равна -му по величине собственному числу ковариационной матрицы этих случайных величин.

Дадим полученным главным компонентам наглядную интерпретацию. Рассмотрим случайные величины  такие, что .  некоррелированы и нормированы (имеют единичную дисперсию). В матричном виде выражение  следующее: . Тогда  выражается через  следующим образом: . Таким образом, мы получили систему



Или, в матричном виде, , где -ый столбец  матрицы  имеет вид .

Такое выражение случайных величин  удобно тем, что их дисперсии наглядно выражаются через коэффициенты матрицы :

.

Тогда суммарная дисперсия случайных величин  вычисляется по формуле .

Величина  называется вкладом компоненты  в дисперсию , а  - вкладом компоненты  в суммарную дисперсию случайных величин .

Ранее было получено, что , где  - -ый столбец матрицы . Собственные вектора  мы выбрали единичной длины, следовательно, . Вспомним, что  упорядочены по убыванию. Следовательно,  имеет наибольший вклад в суммарную дисперсию,  обладает вторым по величине вкладом, и так далее. Вклад последней компоненты  в суммарную дисперсию является наименьшим.

Предположим, что вклад компоненты  настолько мал, что им можно пренебречь. Тогда систему  можно описать не , а  случайными величинами. В общем случае, если величина  (т.е. вклад последних  компонент) мала, то -мерный вектор  можно охарактеризовать всего лишь  параметрами .

Таким образом, в методе главных компонент задача снижения размерности сводится к нахождению собственных чисел и собственных векторов ковариационной матрицы исходных признаков. Поэтому, с развитием вычислительной техники метод главных компонент приобрел большое практическое значение.

## Факторный анализ.

### Модель факторного анализа

Факторный анализ представляет собой совокупность методов, имеющих разные принципы работы, но объединенных предположением о том, что изменчивость в значениях наблюдаемых признаков в основном обусловлена наличием небольшого числа (меньшего, чем количество признаков) скрытых причин, общих для всех признаков. Эти причины называются общими факторами. Оставшаяся доля изменчивости каждого признака объясняется присутствием «частного фактора», который влияет только на этот признак и ни на какой другой.

Факторный анализ, так же как и метод главных компонент, основан на исследовании ковариационной (или корреляционной) матрицы исходных признаков. Однако если метод главных компонент работает только с дисперсиями признаков и их линейных комбинаций, то факторный анализ интересуется ковариациями (корреляциями) признаков. Большой коэффициент корреляции между двумя признаками говорит о том, что среди общих факторов имеются такие, которые сильно влияют на эти два признака одновременно. Таким образом, работая с корреляционной матрицей исходных признаков, можно установить, как именно признаки зависят от общих факторов. И тогда размерность изучаемой системы может быть сокращена с числа признаков до числа общих факторов.

Модель факторного анализа выглядит следующим образом:



Здесь  - непосредственно наблюдаемые значения признаков,  - называются общими факторами, поскольку они порождают зависимость между признаками,  - называются частными факторами, т.к.  влияет только на значение -го признака.  - действительные числа. Отметим, что значения и общих, и частных факторов непосредственно не наблюдаемы.

Кроме того, принимаются следующие допущения:

* Общие факторы  - некоррелированные случайные величины
* 
* Частные факторы  независимы как между собой, так и с .
* 

Дисперсии  частных факторов называются остаточными дисперсиями.

В матричном виде наша система записывается как . Матрица  называется матрицей нагрузок. Ее элемент  называется нагрузкой -го фактора на -й признак.

Таким образом, в факторном анализе задача снижения размерности сводится к оцениванию матрицы нагрузок на основании наблюдений над случайными величинами .

### Сравнение факторного анализа с методом главных компонент.

Сходство факторного анализа с методом главных компонент состоит в том, что в обоих методах снижение размерности признакового пространства достигается за счет отыскания некоторых факторов (их должно быть меньше, чем исходных признаков), изменчивость которых в наибольшей степени объясняет изменчивость всего первоначального набора признаков.

Основное различие заключается в следующем. В методе главных компонент при исследовании  признаков ищется тоже  таких факторов, что объясняет всю изменчивость исходной системы. А затем, те факторы, которые объясняют наименьшую долю изменчивости, отбрасываются как несущественные, за счет чего и достигается снижение размерности. Причем эти  факторов ищутся в виде линейных комбинаций исходных признаков, что означает отсутствие каких-либо значительных предположений о структуре изучаемой системы. А в факторном анализе сокращение размерности производится исходя из *изначального предположения*, что изменчивость в значениях признаков является результатом влияния меньшего числа факторов-причин. Поэтому, в некотором смысле, в рамках модели факторного анализа сама возможность сокращения размерности более обоснована. Еще различие одно различие заключается в том, что факторный анализ, в отличие от метода главных компонент, учитывает также наличие специфических факторов, что делает его модель более приближенной к реальной жизни.

Тем не менее, несмотря на эти отличия, нельзя отрицать, что факторный анализ и метод главных компонент чрезвычайно похожи. Более того, авторы многих научных изданий относят метод главных компонент к методам факторного анализа.

### Интерпретация матрицы нагрузок.

Возвращаясь к непосредственно факторному анализу, прежде всего, необходимо поговорить о математической трактовке матрицы . Из предположений модели следует, что

.

Если еще предполагается, что случайные величины  нормированы, то  представляет собой корреляционную матрицу случайных векторов  и , т.е. .

Теперь перейдем к содержательной интерпретации нагрузочной матрицы. Предположим, что нам удалось каким-то образом оценить матрицу . Если -ая строка полученной матрицы содержит только маленькие элементы, то можно считать, что  некоррелирован с другими признаками, а поскольку мы допускаем только линейные зависимости, то и независим. Более того, из предположений модели следует, что . Если все нагрузки на -ый признак малы, то основная доля его дисперсии приходится на частный фактор. Таким образом, мы можем считать  шумовым признаком.

Если  - единственный большой элемент как в -ой строке, так и в -ом столбце, то  также следует считать независимым с другими признаками. Однако в этом случае некоторая доля его дисперсии объясняется влиянием на него -го общего фактора. Дальнейшее использование этого заключения зависит от специфики решаемой исследователем задачи.

Наконец, если существуют два признака  и  такие, что нагрузки каждого фактора для них примерно одинаковы, причем нагрузка хотя бы одного фактора велика (т.е. -ая и -ая строки матрицы  почти равны, и имеют хотя бы по одному большому элементу), то  и  зависимы между собой. Таким образом, эти два признака можно заменить их некоторой комбинацией, или же один из признаков можно просто напросто исключить из рассмотрения. В любом случае, появляется возможность уменьшить количество признаков на один.

### Неоднозначность решения задачи факторного анализа. Выбор вращения факторов.

В модели факторного анализа, как нагрузочная матрица, так и общие факторы могут быть определены только с точностью до ортогонального преобразования. В самом деле, рассмотрим ортогональную матрицу  размером . Обозначим случайный вектор  как . Из ортогональности  следует, что компоненты вектора  являются центрированными и нормированными и взаимно некоррелированными случайными величинами, равно как и общие факторы модели. Более того, равенство  можно переписать в виде . Таким образом, случайные величины  тоже являются общими факторами модели, но соответствуют другой нагрузочной матрице. Мы обозначили ее за .

Вспомним, что умножение на ортогональную матрицу можно рассматривать как поворот. Таким образом, общие факторы и нагрузки в модели факторного анализа определяются с точностью до вращения.

Такая неоднозначность сопряжена не только с неудобствами. Применение факторного анализа принято считать успешным, если его результаты легко интерпретируются. Таким образом, получив каким-либо из методов факторного анализа нагрузочную матрицу, мы можем с помощью вращения добиться ее наилучшей возможной интерпретации.

Существует множество различных методов вращения факторного пространства. В этой работе будут рассмотрены два наиболее часто применимые из них – варимакс и квартимакс.

Методом «варимакс» находится матрица с элементами  такими, что значение выражения  максимально. Столбцы полученной матрицы будут максимально сильно отличаться друг от друга, что означает разделение факторов за счет максимального уменьшения числа признаков, связанных с каждым фактором.

Квартимакс ищет такую матрицу , что максимизируется значение выражения . В этом случае строки полученной матрицы будут различаться максимально сильно, что означает, что с каждым признаком связано минимальное возможное число факторов.

### Метод максимального правдоподобия

Как уже было сказано ранее, факторный анализ представляет собой совокупность методов, имеющих общую модель, но разные принципы работы. Одним из этих методов является метод максимального правдоподобия. Он представляет собой частный случай классического метода максимального правдоподобия для оценивания неизвестных параметров распределения случайных величин.

Напомним принцип работы классического метода максимального правдоподобия. Пусть рассматривается случайный вектор . Нам известен «тип» распределения случайного вектора , но неизвестны некоторые параметры этого распределения. Метод максимального правдоподобия оценивает эти параметры путем максимизации некоторой функции от этих параметров, называемой функцией правдоподобия. В случаях дискретного и непрерывного распределений случайного вектора  определение функции правдоподобия разное.

Пусть  - выборка такая, что при любом   распределен так же, как и . Причем  независимы между собой. Пусть  - реализация этой выборки. Рассмотрим функцию .

Если  - дискретный случайный вектор, то функция правдоподобия  определятся как .

Если  имеет непрерывное распределение, то определение следующее:

, где  - функция плотности случайной величины .

Следует отметить, что в большинстве случаев в качестве  используется тождественная функция. В этом случае, определения функции правдоподобия для дискретного и непрерывного случаев превращаются в  и , соответственно. Несмотря на некоторую излишнюю конкретику, эти два определения являются более распространенными.

Идея метода кажется интуитивно понятной. Выражаясь простыми словами, он максимизирует вероятность увидеть те значения случайного вектора, которые мы и без того уже видим. Таким образом, кажется разумным считать оценки, которые доставляют максимум функции правдоподобия, близкими к реальным значениям параметров. Тем не менее, следует отметить, что оценки, полученные методом максимального правдоподобия, могут быть смещенными, но все же являются асимптотически несмещёнными, [состоятельными](http://ru.wikipedia.org/wiki/%D0%A1%D0%BE%D1%81%D1%82%D0%BE%D1%8F%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0), а при выполнении определённых условий регулярности - асимптотически эффективными и асимптотически нормальными.

Перейдем к методу максимального правдоподобия для факторного анализа. Начнем с того, что в нем предполагается, что общие факторы  имеют стандартное нормальное распределение , частные факторы  имеют нормальное распределение: .

Обозначим ковариационную матрицу случайных величин  как . Из предположений модели следует, что





В матричном виде эти равенства выглядят как , где  - матрица, диагональными элементами которой являются остаточные дисперсии, а всеми остальными – нули.

Матрица  обычно неизвестна, но может быть оценена выборочной ковариационной матрицей. Пусть  - независимые наблюдения над случайным вектором . В общем случае, выборочная ковариационная матрица  вычисляется по формуле: , где . Но мы предполагаем, что  - центрированный случайный вектор, поэтому формула имеет вид: .

Пусть  - реализация выборки . В методе максимального правдоподобия для факторного анализа используется выборочная ковариационная матрица  в качестве функции  (см. определение функции правдоподобия). Тогда функцией правдоподобия является совместная функция плотности элементов матрицы , вычисленная для .

Воспользуемся следующим утверждением. Если случайный вектор  имеет нормальное распределение , то совместная функция плотности элементов его выборочной ковариационной матрицы  имеет вид :



Где  - множитель, зависящий только от  и ,  - элементы матрицы . Это распределение называется распределением Уишарта и обозначается . Поскольку оцениваемым параметром является матрица , мы можем опустить не зависящие от нее множители. Выражение примет вид . Для удобства прологарифмируем полученное выражение: . Поскольку логарифм является монотонно возрастающей функцией, то полученное логарифмированием выражение достигает максимума в той же точке, что и первоначальное. Вновь опуская члены, не зависящие от , получаем нашу функцию правдоподобия: . Поскольку во время преобразований мы поменяли знак, оценками максимального правдоподобия будут те значения  и , которые доставляют ***минимум*** функции . Дифференцируя ее по  и  и приравнивая полученные частные производные к нулю, мы получим необходимые уравнения на  и . Проведя необходимые алгебраические преобразования, мы получаем систему:



Здесь  обозначает матрицу, у которой на главной диагонали стоят диагональные элементы матрицы , а все остальные элементы – нули. Матрица , как уже говорилось, имеет вид . В разделе 1.2.4 доказывалось, что матрица  может быть определена только с точностью до умножения на произвольную ортогональную матрицу. Поэтому для решения этой системы нам требуется дополнительное условие на . Потребуем, чтобы матрица  была диагональной, причем ее диагональные элементы были расположены в порядке убывания. Таким образом, мы свели задачу поиска оценок максимального правдоподобия для  и  к решению этой системы с дополнительным условием. Искать ее решение предполагается итерационным методом, который будет представлен при описании технической составляющей данного дипломного проекта.

### Выбор числа общих факторов.

В предыдущих главах предполагалось, что число общих факторов  заранее известно исследователю. Тем не менее, на практике это обычно не так. Таким образом, одной из задач факторного анализа является определение этого числа.

На сегодняшний день существует множество как теоретически обоснованных, так чисто эмпирических методов решения этой задачи. К первым относится, например, критерий значимости, основанный на методе отношения правдоподобия. Согласно этому критерию, сначала принимается некое начальное значение . При отсутствии каких-либо априорных предположений относительно числа факторов за начальное значение берется . Затем необходимо проверить гипотезу : «число общих факторов меньше либо равно » против альтернативы : «число общих факторов строго больше ». Если  будет отвергнута в пользу альтернативы, то  следует увеличить на единицу и заново проверить гипотезу , но уже с новым значением . Алгоритм останавливается, когда основная гипотеза будет принята в первый раз. Текущее значение  и следует считать числом общих факторов.

Теперь о том, как именно необходимо проверять гипотезу  можно переформулировать более формально: «ковариационная матрица признаков  размером  может быть представлена в виде суммы диагональной матрицы с положительными элементами на диагонали и матрицы  ранга ». (Соответственно, : «не может»). Потребуем, чтобы выполнялось неравенство: . По сути, это требование того, чтобы  было мало по сравнению с .

Для проверки  против  воспользуемся методом отношения правдоподобия.

Пусть некоторая случайная величина  имеет распределение , причем вид функции нам известен, а о параметре  известно лишь то, что он принадлежит некому допустимому множеству: . Пусть проверяется гипотеза  против альтернативы . Согласно методу отношения правдоподобия, для проверки  против  на уровне значимости необходимо вычислить значение статистики , где - выборка, порождаемая случайной величиной , а  - функция правдоподобия этой выборки. Если это значение попадает в критическую область , тогда гипотеза  отвергается в пользу альтернативы . Значение параметра выбирается так, чтобы критерий имел заданный уровень значимости , т.е. вероятность попадания значения статистики в критическую область при условии верности гипотезы была меньше либо равна .

При выполнении некоторых условий, (которые выполняются в том числе тогда, когда наблюдения независимы и имеют нормальное распределение, т.е. в нашем случае), если объем выборки стремится к бесконечности, то, при верности гипотезы , статистика  сходится по распределению к случайной величине, распределенной согласно закону хи-квадрат , где . Точные формулировки этих условий приведены в: (Roger R. Davidson, William E. Lever, 1967).

Вернемся к проверке гипотезы о числе факторов. В нашем случае, в качестве  выступает случайный вектор признаков . Неизвестный параметр распределения этого случайного вектора – его ковариационная матрица . Тогда множество  - все симметричные положительно определенные матрицы размера .  - множество матриц, которые могут быть представлены виде суммы диагональной матрицы с положительными элементами на диагонали и матрицы  ранга . Тогда  достигается при значении параметра  равном , где  и  - оценки нагрузочной матрицы  и матрицы остаточных дисперсий , полученные методом максимального правдоподобия. Очевидно также, что  достигается при значении , равном выборочной ковариационной матрице . Подставляя эти значения в нашу функцию правдоподобия (формула (1) пункт 1.2.5), получаем: 



Здесь под  имеется ввиду след матрицы .

Логарифмируя это выражение и умножая на , получаем .

Асимптотическое распределение этой величины - , где . Вычислим значение степеней свободы . Размерность множества всех симметричных положительно определенных матриц размера равно . Число всех параметров в выражении равно . Более того, в пункте 2.2.5 на нагрузочную матрицу  было наложено дополнительное условие, обеспечивающее ее однозначность. А именно, было потребовано, чтобы матрица  была диагональной, причем ее диагональные элементы были расположены в порядке убывания. Это условие накладывает дополнительных ограничений на матрицы  и . Таким образом, . Тогда . Заметим, что условие , наложенное в начале этого пункта, обеспечивает положительность числа .

В качестве границы критической области выступает квантиль .

Таким образом, если значение статистики оказывается больше, чем квантиль , то гипотеза  отвергается в пользу альтернативы .

Среди более эмпирических методов определения числа факторов хотелось бы выделить критерий, основанный на величине доли объясняемой дисперсии. Суть этого критерия состоит в том, что факторы ранжируются по величине доли объясняемой суммарной дисперсии признаков. А затем, те факторы, которые не объясняют долю дисперсии , превышающую некоторый заранее заданный порог, исключаются из рассмотрения.

## Выводы по главе

Итак, в этой главе был дан краткий обзор теоретических основ факторного анализа и, в частности, его метода максимального правдоподобия. В том числе были рассмотрены такие важные задачи, как выбор метода вращения факторного пространства и определение числа общих факторов. Помимо прочего, было проведено сравнение факторного анализа с родственной ему техникой – методом главных компонент, что способствует лучшему пониманию принципа работы первого.

# Глава 2 (техническая)



## Итерационный метод поиска оценок максимального правдоподобия для матрицы нагрузок и матрицы остаточных дисперсий

Напомним, что мы свели задачу поиска оценок максимального правдоподобия для  и  к решению системы  с дополнительным условием на матрицу . Из системы выводится равенство . Оно понадобится нам немного позже.

Решение этой системы с условием можно найти следующим итерационным методом. Пусть имеются некоторые начальные приближения матриц  и  -  и . Обозначим -ые строки матриц  и  как  и , соответственно. На вход алгоритму подаются выборочная ковариационная матрица  и матрицы  и .

На первом шаге первой итерации мы вычисляем:

* вектор-строку 
* вектор-строку 
* число 

и принимаем  в качестве второго приближения для .

Если у нас более одного фактора, мы переходим ко второму шагу первой итерации. На нем мы вычисляем:

* вектор-строку 
* число 
* вектор-строку 
* число 

и объявляем  вторым приближением для .

Если имеется третий фактор, то делаем третий шаг первой итерации, т.е. вычисляем

* вектор-строку 
* два числа  and 
* вектор-строку 
* число 

и берем  за второе приближение для , и так далее. Если всего у нас  факторов, то нам необходимо сделать  таких шагов. В конце первой итерации мы получим второе приближение для матрицы . Обозначим его . Второе приближение матрицы , обозначаемое , мы получаем, подставляя элементы  в равенство .

Матрицы  и  принимаются за начальные для второй итерации алгоритма. На ней мы аналогичным способом получаем третьи приближения матриц  и , обозначаемые  and . И так далее.

Условия сходимости описанного алгоритма неизвестны, однако на практике сходимость есть, причем довольно быстрая.

В процессе работы над дипломным проектом этот алгоритм был реализован на Матлабе. Сразу оговоримся, что, для улучшения качества работы метода, вместо выборочной ковариационной матрицы признаков была использована выборочная корреляционная. В качестве начальных приближений  и  были приняты матрица размера , состоящая из случайных величин, равномерно распределенных на отрезке  и диагональная матрица размера , все диагональные элементы которой равны .

## Выводы по главе

Итак, в этой главе был представлен итерационный алгоритм, реализующий традиционный метод максимального правдоподобия.

Еще раз подчеркнем, что все результаты и выводы, представленные в этой и предыдущей главе, не являются результатами исследования автора данного дипломного проекта. До этого момента автор лишь анализировал и обобщал информацию, представленную в различных учебных пособиях по факторному анализу, и описывал результаты других исследований в данной области.

# Глава 3 (экспериментальная)

В этой и следующей главах мы переходим к описанию исследования, проведенного непосредственно самим автором. Если какие-то результаты, представленные в этой главе, позаимствованы из других источников, это будет явно указано.

В этой главе будут предложены способы адаптации традиционного метода максимального правдоподобия для работы с признаками, связанными монотонным нелинейным и немонотонным типами зависимости. Затем будет проверена и сравнена эффективность традиционного метода максимального правдоподобия и его модификаций на основании результатов серии применения этих методов к искусственно смоделированным данным, связанным различными типами зависимости.



## Адаптация метода максимального правдоподобия к нелинейным зависимостям.

Как показано в первой главе, в факторном анализе предполагается, что значения признаков ***линейно*** зависят от общих факторов, а, следовательно, и между собой. Однако, на практике это предположение редко является обоснованным.

Как уже было сказано, в факторном анализе в качестве меры связи между признаками используется ковариация или же, гораздо чаще, корреляция. Однако корреляция является мерой ***линейной*** зависимости между двумя случайными величинами, поэтому в случае нелинейной зависимости признаков она теряет в информативности. Более того, если между признаками имеется зависимость немонотонного типа, то вообще пропадает смысл использовать корреляцию или ковариацию, как меру связи между ними.

Тем не менее, в случаях линейной зависимости признаков модель факторного анализа достаточно хорошо описывает действительность, проста для понимания и является эффективной для многих реальных ситуаций. Поэтому не хотелось бы совсем отказываться от методов факторного анализа, если признаки нелинейно зависимы. Естественным образом возникает вопрос: а нельзя ли внести небольшие изменения в методы факторного анализа, чтобы основная идея осталась прежней, но методы стали эффективными и для случаев нелинейной зависимости тоже.

В качестве этих изменений предлагается следующее. Попробуем заменить коэффициент корреляции на меру связи, которая будет более информативной в случае нелинейной зависимости.

Предположим, заранее известно, что зависимость между признаками монотонного типа. В этом случае в качестве меры связи предлагается использовать коэффициент ранговой корреляции Спирмена.

Пусть  - независимые наблюдения над случайным вектором признаков . Тогда  являются значениями -го признака для этих  наблюдений. Рассмотрим два признака  и . Пусть  - ранг элемента  в выборке . Другими словами, если мы упорядочим  по возрастанию,  окажется под номером . Аналогично,  - ранг элемента  в выборке . Пусть  и  - соответствующие средние арифметические рангов, т.е. , . Тогда коэффициент ранговой корреляции двух признаков  и  вычисляется по формуле:



Матрицу, состоящую из коэффициентов ранговой корреляции, мы обозначаем как .

По факту, коэффициент ранговой корреляции Спирмена показывает, насколько хорошо зависимость между двумя случайными величинами может быть описана монотонной функцией. Поэтому он не подходит, если даже предположение о монотонности зависимости между признаками не представляется обоснованным. В этом случае предлагается заменить коэффициенты корреляции на коэффициенты Крамера.

Чтобы вычислить коэффициент Крамера для двух признаков  and , в первую очередь необходимо построить таблицу сопряженности этих признаков. В нашем случае таблица сопряженности строится следующим образом. Упорядочим каждую из выборок  и  в порядке возрастания и разобьем полученные последовательности на заданное число интервалов равной длины. Тогда на пресечении -ой строки и -го столбца таблицы сопряженности должно стоять число наблюдений, у которых значения -го и -го признаков попали соответственно в -ый и -ый интервалы. Элемент, стоящий на пересечении последней строки и -го столбца должен равняться числу наблюдений, чьи значения -го признака попали в -ый интервал, т.е. просто сумме элементов -го столбца. Для элемента на пересечении -ой строки и последнего столбца все симметрично. На пересечении последней строки и последнего столбца находится общее число наблюдений. Таким образом, таблица сопряженности признаков  and  имеет вид:

Таблица 3.1. Таблица сопряженности

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Интервал |  | Интервал |  | Интервал |  |
| Интервал |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Интервал |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Интервал |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Отметим, что для простоты мы разбили значения обоих признаков на равное число интервалов. (У нас оно обозначено как ). Однако, в общем случае число интервалов может быть разным.

Следующим шагом к получению коэффициента Крамера для признаков  и  является вычисление статистики хи-квадрат для этих признаков. Она вычисляется на основании таблицы сопряженности по следующей формуле:



Тогда формула для вычисления коэффициента Крамера следующая:



Следует отметить, что эта формула верна только если значения признаков были разбиты на равное число интервалов. В общем случае в знаменателе дроби вместо  должно стоять минимальное из этих чисел.

Матрицу, состоящую из коэффициентов Крамера, мы обозначаем как .

Попробуем подать на вход методу максимального правдоподобия вместо выборочной корреляционной матрицы  матрицу  или , в зависимости от того, что нам известно о типе связи между признаками.

Необходимо признать, что, заменяя выборочную корреляционную матрицу  на  или , мы жертвуем теоретической обоснованностью. Кроме того, возникают некоторые неясности. Например, цель измененного метода такая же как и у первоначального – оценить нагрузочную матрицу . Но давайте вспомним, что именно мы называем матрицей . В прошлых главах мы предполагали, что каждый признак может быть представлен как линейная комбинация общих факторов плюс частный фактор. Матрицей  мы называли матрицу из коэффициентов этих линейных комбинаций. Однако сейчас мы допускаем, что признаки могут нелинейно зависеть друг от друга, а, следовательно, и от общих факторов. Значит, их нельзя представить в виде линейных комбинаций факторов. Так что же теперь мы подразумеваем под ?

Тем не менее, несмотря на то, что наша замена одного коэффициента другим носит чисто эмпирический характер, в случае нелинейной зависимости признаков модифицированные методы оказываются более эффективными, чем традиционный. Это будет продемонстрировано в следующих разделах.

## Генерация тестовых данных.

Для проверки корректности работы методов и сравнения их эффективности для каждого типа зависимости необходимо иметь достаточное количество тестовых данных, связанных зависимостями линейного, монотонного и немонотонного типа. Под зависимостью, например, монотонного типа понимается нефункциональная зависимость, которая хорошо приближается монотонной функцией. (Аналогично для линейного и немонотонного типов). Причем, эти функции должны быть нам заранее известны. Более того, эти данные должны иметь такую структуру, чтобы было заранее понятно, каким должен оказаться результат, если метод корректно работает на этих данных.

Было решено моделировать тестовые данные следующей структуры. Генерируется выборка размера , элементами которой являются вектора размерности . -ая компонента -го наблюдения означает реализацию-го признака для -го объекта. Причем генерация выборки происходит так, чтобы признаки образовывали группы. В каждой группе зависимость между признаками должна хорошо приближаться заранее заданной функцией: линейной, нелинейной, но монотонной или немонотонной. Признаки из разных групп должны быть независимы. В терминах факторного анализа такую структуру следует понимать как: на все признаки каждой группы влияет единственный один и тот же общий фактор, и больше этот фактор не влияет ни на какие другие признаки. Таким образом, у нас каждой группе соответствует один общий фактор. Других общих факторов нет.

Был выбран именно такой принцип генерации, потому что по таким данным хорошо понятно, каким должен получиться результат применения метода в случае его правильной работы. А именно, строки полученной нагрузочной матрицы тоже должны образовывать группы. Строки внутри одной группы должны быть почти равны. (Т.е. их поэлементные разности должны быть очень маленькими). Причем, в каждой строке должно быть по одному большому элементу – в том столбце, который соответствует фактору, влияющему на ту группу, к которой относится признак, соответствующий этой строке. Остальные элементы этой строки должны быть близки к нулю.

Таким образом, если метод работает, например, только для монотонного типа зависимости, то строки его нагрузочной матрицы образуют столько групп, в скольких тип зависимости между признаками носит монотонный характер.

Далее будет подробно описан способ генерации данных вышеуказанной структуры, которые в дальнейшем используются в данной работе для демонстрации полученных автором результатов.

Для демонстрации результатов были сгенерированы четыре группы признаков, по три признака в каждой. В первых двух группах признаки связаны зависимостью линейного типа, в третьей – монотонного нелинейного, в четвертой – немонотонного. В первой группе линейный тип зависимости обусловливается высокими коэффициентами корреляции между признаками. Зависимости в трех оставшихся группах – это «зашумленные» функциональные зависимости. Другими словами, эти зависимости приближаются с высокой точностью некоторыми (конкретными) линейной, монотонной нелинейной и немонотонной функциями, которые далее обозначаются ,  и  соответственно. Иллюстрации этих четырех зависимостей представлены ниже.

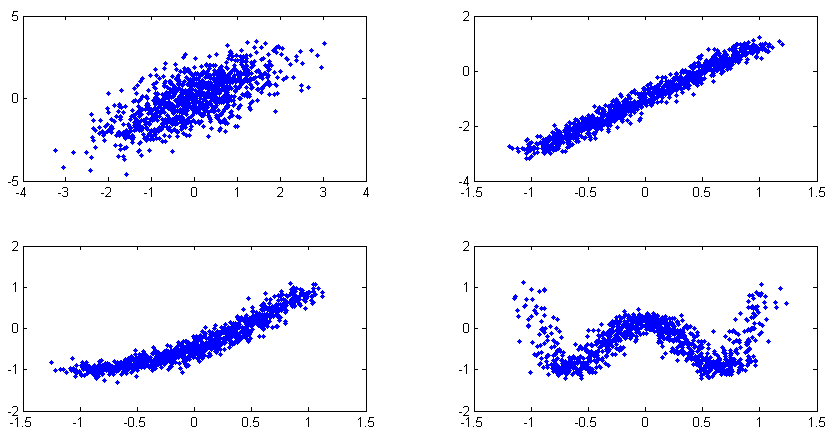


Рисунок 3.1. Зависимости, связывающие признаки в группах демонстрационных данных

Чтобы избежать путаницы, далее будем говорить, что в первой группе признаки связаны зависимостью «почти линейного» типа, а во второй – просто линейного. В программе сгенерированная выборка хранится в виде матрицы  размера , где  - количество признаков,  - количество наблюдений. В нашем случае число признаков  равно двенадцати, число наблюдений  для наилучшей наглядности было выбрано равным десяти тысячам, но следует подчеркнуть, что и на выборках гораздо меньшего объема методы работают так же, как и на больших. Элемент матрицы соответствует -ой компоненте -го наблюдения. Строки этой матрицы также образовывают группы. Строки внутри каждой группы получаются одна из другой применением соответствующей функции. Или же случайные величины, порождающие эти строки, имеют заданный коэффициент корреляции.

Первая группа признаков была сгенерирована следующим образом. Первый признак  - это стандартная нормальная случайная величина. Второй признак – это нормальная случайная величина, имеющая коэффициент корреляции с , равный 0.7. Аналогично, - нормальная случайная величина, имеющая с  коэффициент корреляции 0.7.

Необходимо более подробно пояснить, как происходила генерация реализаций случайных величин, имеющих заранее заданный коэффициент корреляции. Сначала дадим нашей задаче математическую трактовку. Пусть имеются независимые случайные величины  и . Рассмотрим случайную величину . Необходимо подобрать такое значение параметра , чтобы коэффициент корреляции случайных величин  и  был равен заранее заданной константе . Причем , поскольку коэффициент корреляции может лежать только в этих пределах. Заметим, что дисперсия случайной величины  равна:



Сначала найдем ковариацию случайных величин  и .



Тогда их корреляция равна:



Выражаем из этой формулы .

Предположим, что  неотрицательно. Тогда:







Аналогично, при отрицательном  получаем:



Итого:



Теперь перейдем к алгоритмической реализации этого метода.

Для генерации данных, имеющих заданный коэффициент корреляции, была создана функция «generateCorr». Реализующий эту функцию фрагмент написанного в Матлабе программного кода приведен в Приложении. Функция «generateCorr» работает следующим образом. Она принимает на вход три целых числа: r, n и corr, где corr . r означает количество признаков в нашей генерируемой группе, n – число наблюдений, corr – заданный коэффициент корреляции. Затем создается матрица размера r×n. Первая строка этой матрицы заполняется стандартными нормальными случайными величинами. Вводится переменная . В начале ей присваивается значение дисперсии каждого элемента первой строки, т.е. единица. Затем в цикле для каждой строки матрицы, со второй по последнюю, выполняется следующая последовательность шагов:

* вычисляется текущее значение переменной α по формуле . По сути, это и есть формула (3), только теперь под случайной величиной мы понимаем стандартную нормальную случайную величину. Значит, в числителе равно единице. А текущее значение переменной на данной итерации цикла равно дисперсии элементов предыдущей строки.
* происходит заполнение текущей строки следующим образом: i-ый элемент текущей строки получается умножением i-го элемента предыдущей строки на текущее значение параметра α и прибавлением к полученному произведению стандартной нормальной случайной величины, (под которой и подразумевается из математической трактовки).
* переменной  присваивается новое значение, равное , т.е. теперь в переменной хранится значение дисперсии элементов текущей строки (см. формулу для (2)).

Таким образом, в полученной матрице коэффициент корреляции между случайными величинами, порождающими ее первую и вторую строки будет равен corr, коэффициент корреляции между случайными величинами, порождающими ее вторую и третью строки будет равен corr, и так далее.

Принцип генерации оставшихся трех групп следующий. Пусть имеют усеченное стандартное нормальное распределение. А  - нормальное распределение . Тогда значения признаков  -  вычисляются по следующим формулам:



Технически эта задача была решена следующим образом. Была создана функция «generateData» (см. Приложение).Она принимает на вход массив функций (теперь уже имеются ввиду математические функции) и два целых числа – m и n. m означает количество признаков в каждой группе (группы одинакового размера), n – количество наблюдений. Тогда количество всех признаков (далее оно обозначается r) равно произведению размера массива функций и m. Затем создается матрица из нулей размера r×n. Заполнение этой матрицы происходит следующим образом. Первая строка матрицы заполняется случайными величинами (правильнее сказать, их реализациями), имеющими усеченное стандартное нормальное распределение. Вторая строка матрицы получается применением первой функции из массива к первой строке, третья - применением первой функции ко второй строке и так далее до m-ой строки. На данный момент мы задали значения всех признаков первой группы. (m+1)-я строка матрицы заполняется опять усеченными стандартными нормальными случайными величинами и все действия повторяются для следующих m строк. Теперь мы задали значения признаков второй группы. И так далее, пока не заполним все строки матрицы. После этого добавляем к каждому элементу матрицы нормально распределенную случайную величину с нулевым математическим ожиданием и дисперсией, равной 0.01, т.е. добавляем к нашим данным шум.

Отметим, что с помощью одного вызова функции «generateData» можно смоделировать сразу несколько групп взаимно зависимых признаков, в то время как одним вызовом функции «generateCorr» моделируется только одна такая группа.

Хотелось бы пояснить, почему используется ***усеченное*** стандартное нормальное распределение вместо обычного. Предположим, что мы все же используем обычное стандартное нормальное распределение при заполнении первых строк каждой группы. Пусть во входном массиве функций имеется многочлен достаточно высокой степени (например, четвертой). Тогда, если при заполнении первой строки той группы, строки внутри которой будут связаны с помощью этого многочлена, значение хотя бы одного элемента этой строки попадет в выброс нормального распределения, то в последующих строках это число будет возведено в очень высокую степень. Тогда некоторые элементы полученной на выходе матрицы будут сильно отличаться от всех остальных. В этом случае, число обусловленности матрицы оказывается очень велико. Плохо обусловленную матрицу тяжело или даже невозможно численно обратить, а метод максимального правдоподобия (и его модификации тоже) широко используют операцию обращения, что приводит к накоплению ошибки или тому, что метод не сходится (т.е. работает бесконечно).

Для генерации случайных величин, имеющих усеченное стандартное нормальное распределение, автором была создана функция «randn\_cut» (см. Приложение). Эта функция работает следующим образом. На вход принимаются два целых числа a и b. Затем создается матрица размера a×b, заполненная стандартными нормальными случайными величинами. (Это делается с помощью встроенной в Матлаб функции «randn(a,b)»). Затем те элементы матрицы, значения которых оказались больше единицы или меньше минус единицы (т.е. попали в выбросы нормального распределения), заменяются на новые стандартные нормальные случайные величины. Затем такое же замещение происходит с новой матрицей. И так до тех пор, пока все элементы не будут лежать в интервале от минус единицы до единицы. (С вероятностью 1 это условие в какой-то момент окажется выполненным).

## Измерение эффективности метода

Чтобы иметь возможность формально сравнивать эффективность методов для разных типов зависимости, нам необходим некоторый коэффициент эффективности.

Предположим, что у нас имеется некий гипотетический метод, идеально работающий на наших данных. Этот метод возвращает некоторую «эталонную» нагрузочную матрицу. Далее будем говорить, ***что из двух «реальных» методов более эффективным является тот, который выдает нагрузочную матрицу, столбцы которой ближе (в смысле среднеквадратической ошибки) к столбцам эталонной матрицы***.

Для наших демонстрационных данных эталонная нагрузочная матрица имеет вид:



Причем, столбцы этой матрицы могут быть расположены не в этом порядке, а произвольно. Однако, без потери общности можно зафиксировать этот порядок. Первый столбец соответствует первой группе признаков, второй – второй группе и т.д. Таким образом, получив одним из наших трех методов нагрузочную матрицу, сравниваем ее столбцы с эталонными столбцами. Для этого сначала необходимо понять, какому из эталонных столбцов соответствует каждый реальный столбец. Для этого для каждого эталонного столбца, начиная с первого, ищем среди реальных столбцов такой, для которого значение среднеквадратической ошибки минимально. (Здесь - количество признаков – значит, для наших демонстрационных данных ; - -ый элемент данного эталонного столбца;  - -ый элемент текущего реального столбца). (На следующей итерации этот столбец мы уже не рассматриваем). Этим столбцом окажется тот, который соответствует той же группе, что и эталонный. Полученное минимальное значение  и считаем коэффициентом эффективности метода для этой группы.

## Определение числа факторов.

Напомним, что одной из задач данного дипломного проекта является применение трех рассматриваемых методов к реальным данным. Поскольку при работе с реальными данными число общих факторов неизвестно, автору данной работы было необходимо реализовать какой-либо из методов определения этого числа. Сначала автором был реализован метод, основанный на критерии значимости (см. раздел 1.2.6). Однако, во время экспериментов с моделированными данными автору не удалось добиться удовлетворительной работы этого метода на нелинейно зависимых признаках. Поскольку реальные данные вполне могут быть связаны нелинейным типом зависимости, возникла необходимость реализации более эффективного для этого случая метода.

Было решено определять число факторов следующим образом. Сначала проводим метод максимального правдоподобия, задав ему число факторов, равное числу признаков . Затем для каждого столбца полученной нагрузочной матрицы  вычисляем значение , где  - -ый элемент -го столбца. В пункте 1.2.3 было показано, что . Отсюда сумма квадратов элементов -го столбца нагрузочной матрицы означает, сколько суммарной дисперсии признаков мы объясняем, добавляя к имеющимся  факторам -ый. Соответственно, тогда  - это сколько суммарного среднеквадратического отклонения признаков мы объясняем, добавляя к имеющимся  факторам -ый. Поскольку метод максимального правдоподобия (а значит и его модификации тоже) упорядочивают столбцы нагрузочной матрицы в порядке убывания «значительности» соответствующих этим столбцам факторов, то и  будут упорядочены в порядке убывания.

Теперь вспомним, что в методе максимального правдоподобия мы используем выборочную корреляционную, а не ковариационную матрицу, что равносильно тому, что мы предполагаем нормированность признаков.

Таким образом, если для некоторого столбца значение  оказалось значительно больше единицы, это означает, что фактор, соответствующий этому столбцу, влияет как минимум на два признака. Значит, этот фактор – значительный, и мы его оставляем. Если  получилось близко к единице, то скорее всего, этот фактор влияет только на один признак, причем другие факторы на этот признак не влияют. Оставлять или отбрасывать такие факторы – зависит от цели исследования. Если же для какого-то фактора  получилось значительно меньше единицы, то можно сделать вывод, что этот фактор объясняет только шум, и отбросить его.

Итак, мы определили, сколько факторов мы оставляем. Опять запускаем метод, но уже заранее задаем ему это число факторов. Полученный результат считаем окончательным.

Следует подчеркнуть, что этот метод не является разработкой автора данного дипломного проекта, поскольку он использует ту же идею, что и критерий, основанный на величине доли объясняемой дисперсии (см. раздел 1.2.6). Отличие состоит лишь в том, что факторы ранжируются не по величине доли объясняемой дисперсии, а по величине приращения объясняемого среднеквадратического отклонения.

Эксперименты с искусственно смоделированными данными подтвердили эффективность этого метода. Это будет продемонстрировано в следующем разделе.

## Результаты применения методов к искусственно смоделированным данным

Итак, до начала экспериментов предполагалось, что традиционный метод максимального правдоподобия работает только для линейных зависимостей и тех, которые хорошо аппроксимируются линейными. Метод, использующий в качестве меры связи признаков коэффициент ранговой корреляции Спирмена (далее – «второй метод»), подходит для любых зависимостей монотонного типа. А от метода, который использует коэффициенты Крамера (далее – «третий метод»), ожидалась корректная работа на всех рассматриваемых типах зависимости.

Для проверки этих предположений, автором данной работы была проведена серия применений всех трех методов к тестовым данным. В каждом эксперименте данные имели структуру, которая была описана в разделе 3.1, менялись только функции, объединяющие признаки в группы.

Результаты экспериментов оказались следующими. Традиционный метод максимального правдоподобия действительно хорошо работает для линейных зависимостей и тех, которые хорошо приближаются линейными. В том числе он оказывается эффективным и для большинства зависимостей монотонного типа, поскольку они тоже могут быть приближены линейными, пусть и менее точно. Для зависимостей немонотонного типа он, как и ожидалось, не работает.

Второй метод хорошо работает для любых зависимостей монотонного типа. Причем с точки зрения эффективности он не отличается от традиционного как для линейного, так и для монотонного. На зависимостях немонотонного типа он тоже не работает.

Третий метод, как и предполагалось, работает для всех рассматриваемых типов зависимости, однако на всех, кроме немонотонного, уступает двум другим методам в эффективности.

Продемонстрируем эти результаты на данных с четырьмя группами признаков (ранее мы везде называли эти данные демонстрационными).

Поскольку сейчас нашей целью является продемонстрировать и сравнить эффективность методов, мы будем заранее задавать им истинное число факторов – четыре. Демонстрация эффективности нашего «эмпирического» метода определения числа факторов будет представлена в следующем разделе.

Результаты применения к демонстрационным данным традиционного метода максимального правдоподобия оказались следующими. Нагрузочная матрица:



Как видим, строки нагрузочной матрицы образуют только три группы – соответствующие группам признаков с нелинейным монотонным, линейным и «почти линейным» типами зависимости (сверху вниз.)

Коэффициенты эффективности для групп в порядке с первой по четвертую:



Применение к этой матрице методов вращения «варимакс» и «квартимакс» не внесло в нее никаких существенных изменений. Практически не изменились также и коэффициенты эффективности.

Таким образом, традиционный метод лучше всего обнаружил зависимости линейного и нелинейного монотонного типов, чуть хуже - «почти линейного» и оставил без внимания зависимость немонотонного типа.

Результат применения второго метода получился следующим. Нагрузочная матрица:



Коэффициенты эффективности для групп в порядке с первой по четвертую:



Методы вращения опять не внесли никаких существенных изменений.

Таким образом, второй метод, так же как и традиционный справился только с первыми тремя типами зависимости.

Как видим, коэффициенты эффективности практически в точности совпадают с коэффициентами традиционного метода, что подтверждает, что для искусственно смоделированных данных эти методы не отличаются в смысле эффективности.

Нагрузочная матрица, полученная третьим методом, оказалась следующей:



Коэффициенты эффективности:



Таким образом, из всех трех методов, только этот метод способен улавливать зависимости немонотонного типа. Однако для всех остальных типов он оказывается менее эффективным, чем два других метода. Особенно сильно он уступает им в случае «почти линейного» типа зависимости (в предыдущих двух случаях коэффициенты эффективности для этого типа были равны  и , а здесь мы видим только ). С помощью методов вращения удалось преобразовать эту матрицу в чуть более похожую на эталонную. Соответственно, и коэффициенты эффективности увеличились в среднем на одну сотую. Но первые три коэффициента все равно оказались меньше, чем полученные первыми двумя методами.

## Демонстрация эффективности метода определения числа факторов, основанного на величине приращения объясняемого среднеквадратического отклонения

Как было показано в предыдущем разделе, первые два метода не выделяют признаки четвертой группы в один фактор. Поэтому, если наша эмпирическая техника определения числа факторов работает правильно, то для первых двух методов первые три коэффициента  должны получиться значительно больше единицы, следующие три – близки к единице, а остальные шесть – значительно меньше единицы. Третий метод выделяет каждую из четырех групп в свой отдельный фактор. Поэтому для него, если наш способ определения числа факторов работает правильно, первые четыре коэффициента  должны получиться значительно больше единицы, а остальные восемь – значительно меньше единицы.

Применяем к демонстрационным данным традиционный метод максимального правдоподобия с максимальным числом факторов, равным двенадцати. Ниже проиллюстрированы значения приращений объясняемого среднеквадратического отклонения (т.е. значения коэффициентов )для двенадцати факторов:

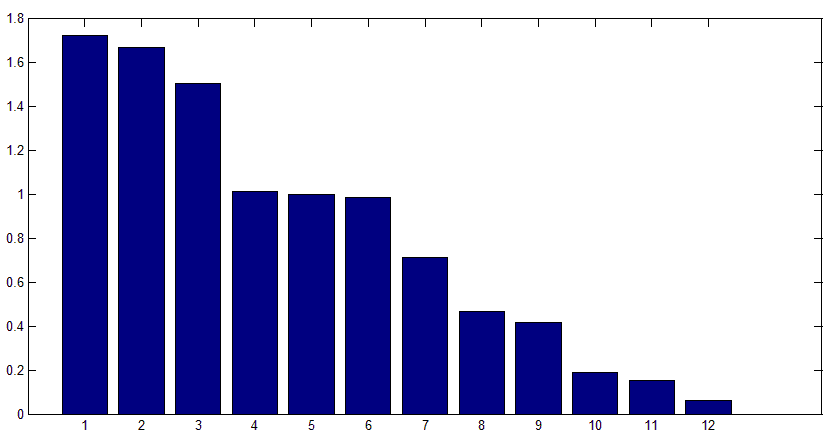


Рисунок 3.2. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для традиционного метода

Как видим, наши ожидания оправдались. Попробуем запустить традиционный метод для шести факторов. Полученная нагрузочная матрица:



Таким образом, первые три группы признаков традиционный метод выделяет каждую в свой фактор – этим факторам соответствуют первые три высоких коэффициента , а последние три признака он выделяет каждый в отдельный фактор – им соответствуют коэффициенты , близкие к единице. Такой вид нагрузочной матрицы свидетельствует о том, что в этом эксперименте метод определения числа факторов сработал правильно.

Теперь применим к нашим данным второй метод, предварительно задав ему максимальное число факторов (двенадцать). Полученные значения коэффициентов  для этих двенадцати факторов:

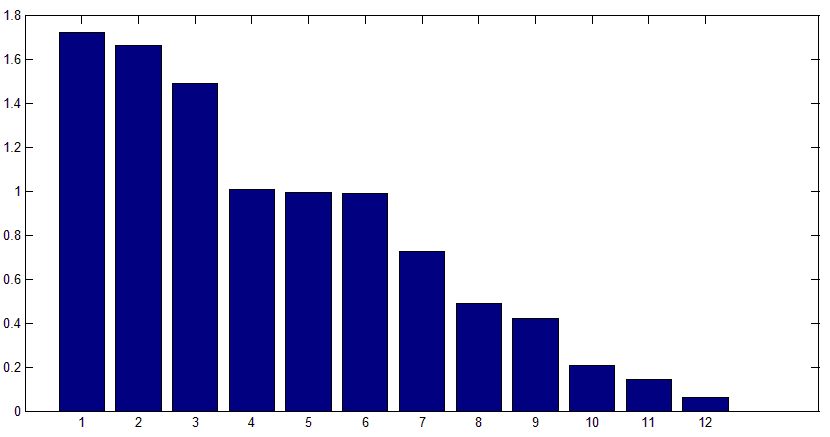


Рисунок 3.3. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для второго метода

Наши ожидания опять оправдались.

Запускаем этот же метод на шести факторах. Нагрузочная матрица:



Интерпретация результата такая же, как и для традиционного метода – в этом эксперименте метод определения числа опять факторов сработал правильно.

Применяем к данным третий метод, который был признан эффективным для всех четырех типов зависимости. Значения  для двенадцати факторов:

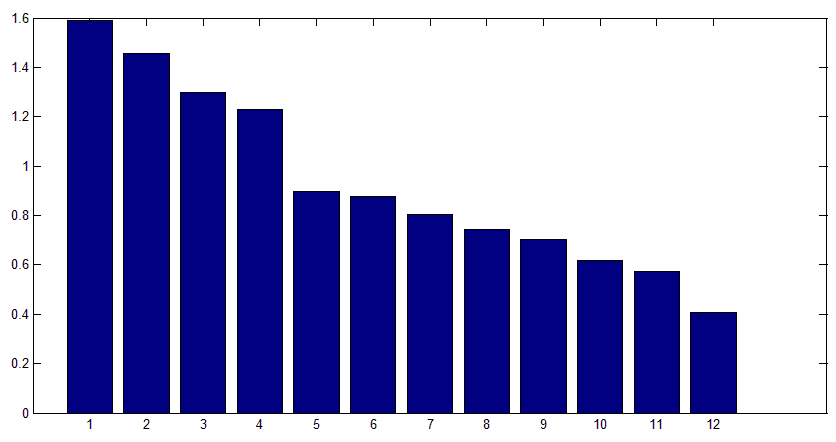


Рисунок 3.4. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для третьего метода

Как видим, больше единицы оказались значения для четырех факторов. Наши ожидания оправдались в третий раз, значит, можно заключить, что наш метод определения числа факторов работает правильно (по крайней мере, на этих данных).

На других искусственно смоделированных данных (той же структуры, но отличных от демонстрационных) этот эмпирический метод определения числа факторов демонстрировал столь же эффективную работу.

## Результаты и выводы по главе

Итак, в этой главе были предложены способы адаптации традиционного метода максимального правдоподобия для работы с признаками, связанными монотонным нелинейным и немонотонным типами зависимости. Эти способы заключаются в том, что вместо выборочной корреляционной матрицы исходных признаков на вход алгоритму подается матрица, состоящая из коэффициентов ранговой корреляции Спирмена, если предполагается что зависимость между признаками носит монотонный характер, или матрица, состоящая из коэффициентов Крамера, если никакое предположения о типе зависимости между признаками не представляется возможным. Также в этой главе были описаны способы вычисления двух вышеуказанных коэффициентов и дана их смысловая интерпретация.

Другая часть этой главы была посвящена экспериментам с искусственно смоделированными данными, проведенным с целью проверки и сравнения эффективности трех рассматриваемых методов для разных типов зависимости между признаками. В этой главе были подробно описаны походы к генерации тестовых данных, введено понятие эффективности метода и предложен способ измерения этой эффективности. Кроме того был подробно описан метод определения числа общих факторов, основанный на величине приращения объясняемого среднеквадратического отклонения и продемонстрирована эффективность работы этого метода.

Результаты проверки и сравнения эффективности методов оказались следующими. Первый метод - традиционный метод максимального правдоподобия – как и предполагалось автором, демонстрирует эффективную работу только в тех случаях, когда зависимость между исходными признаками может быть хорошо приближена линейной функцией. Второй метод – модификация традиционного метода максимального правдоподобия, использующая в качестве меры связи признаков коэффициенты ранговой корреляции Спирмена – как и ожидалось, оказалась эффективной только для монотонного типа зависимости. Третий метод – модификация, использующая коэффициенты Крамера – продемонстрировал удовлетворительную работу как на монотонных, так и на немонотонных типах зависимости, что тоже оправдало ожидания автора.

С точки зрения эффективности, в экспериментах с искусственно смоделированными данными первый и второй метод практически не отличаются. Третий метод, хоть и оказался единственным подходящим для немонотонного типа зависимости, для монотонного типа (а значит, и в том числе для линейного) значительно уступает двум другим методам в эффективности.

Все вышеизложенные результаты были продемонстрированы этой главе на конкретных примерах.

# Глава 4 (исследовательская).

В этой главе будут продемонстрированы результаты проверки и сравнения эффективности работы всех трех методов для ***реальных*** данных и сделаны выводы касательно их применимости для решения практических задач.



## Выбор реальных данных.

Поскольку нашей целью является все же проверка применимости разработанных методов в реальной жизни, то нам требуются такие данные, чтобы, исходя из логики и здравого смысла, было понятно, каким приблизительно должен получиться результат применения метода, если он действительно работает.

В качестве таких данных было решено выбрать еженедельные средние потребительские цены на некоторые продукты питания. В данном случае признаками являются товары, а наблюдениями – моменты времени. Напомним, что в модели факторного анализа предполагается, что наблюдения независимы и одинаково распределены. Но цены на товары растут с течением времени, поэтому, чтобы предположения модели выполнялись, под реализацией -го признака для -го наблюдения (далее обозначается как ) мы будем понимать не саму цену товара  в момент времени (далее обозначается ), а величину относительного прироста цены, т.е. .

Чтобы по этим данным было понятно, каким должен получиться результат работы метода в случае, если этот метод действительно работает, товары следует выбирать так, чтобы они, опять же, образовывали группы. В одной группе должны находиться товары, изменения цен которых взаимозависимы. Например, очевидно, что изменение цены на сметану находится в прямой зависимости от изменения цены на молоко.

Итак, в качестве признаков были выбраны следующие товары:

1. говядина
2. сосиски и сардельки
3. колбаса полукопченая и варено-копченая
4. колбаса вареная I сорта
5. говядина и свинина тушеная консервированная
6. масло сливочное
7. сметана
8. творог жирный
9. сыры сычужные твердые и мягкие
10. мука пшеничная
11. хлеб и булочные изделия из пшеничной муки

Еженедельные средние потребительские цены на эти продукты за период времени с января 2008 г. по апрель 2014 г. были взяты с сайта Федеральной службы государственной статистики: <http://www.gks.ru>.

Очевидно, что первые пять продуктов образуют одну «мясную» группу. Продукты 6. – 9. образуют «молочную» группу, а последние два продукта – это «мучная группа».

Таким образом, мы предполагаем, что, если метод на этих данных работает правильно, то он выделит каждую из групп в свой отдельный фактор. Однако, в отличие от эксперимента с искусственно смоделированными данными, мы не можем утверждать, что если на этих данных некий гипотетический метод работает идеально, то полученная им нагрузочная матрица должна иметь вид:



Потому что:

1. Реальные данные значительно «зашумлены»
2. Изменения цен на продукты из разных групп, скорее всего, зависимы, пусть и значительно слабее, чем внутри групп.
3. Зависимости внутри групп носят более сложный характер, чем в случае искусственно смоделированных данных. В качестве доказательства этого утверждения, ниже продемонстрирован график зависимости изменения цены на сосиски и сардельки от изменения цены на говядину.

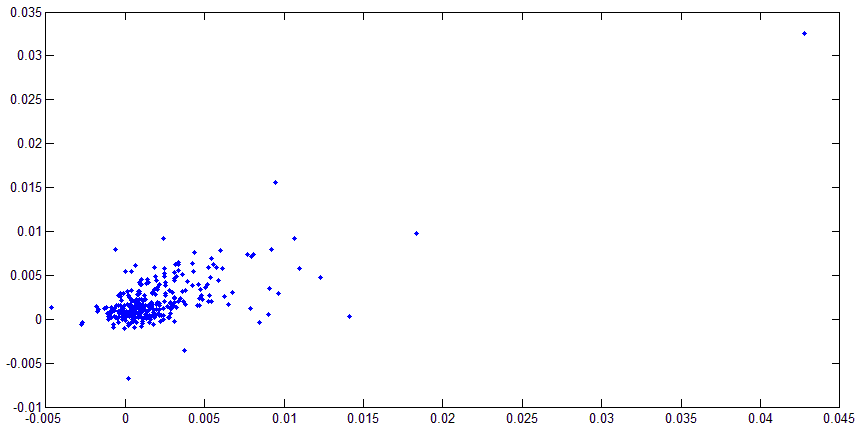


Рисунок 4.1. Зависимость изменения цены на сосиски и сардельки от изменения цены на говядину

Однако, нами выбраны такие данные, что мы заранее знаем, что матрица, полученная этим идеальным методом должна быть достаточно близка к матрице . Поэтому, при работе с этими реальными данными мы по-прежнему можем использовать коэффициенты эффективности, определенные в предыдущей главе, с той лишь разницей, что *небольшие* расхождения в значениях этих коэффициентов для двух методов не дают нам права утверждать, что один из них работает лучше, чем другой. (Потому что из того, что одна из матриц находится чуть ближе к матрице , не следует, что она ближе и к идеальной матрице). Но, если эти расхождения значительны, мы можем это утверждать.

Более того, мы не рассчитываем (по тем же самым причинам 1. – 3.), что наши методы будут работать на этих данных так же хорошо, как на искусственно смоделированных. Поэтому, мы не ожидаем, что как при эксперименте с искусственно смоделированными данными нам удастся сразу (т.е. без использования методов вращения) получить нагрузочную матрицу, которая будет близка к матрице . Нам достаточно того, чтобы строки первоначальной матрицы сгруппировались «по похожести». То есть первые пять строк полученной нагрузочной матрицы должны быть очень похожи друг на друга (т.е. их поэлементные разности должны быть малы), и значительно отличаться от всех остальных строк. Строки с шестой по девятую, аналогично, должны быть похожи друг на друга и ни на какие другие. И то же самое для последних двух строк. Однако с помощью методов вращения мы надеемся свести первоначально полученную матрицу к более близкой к эталонной.

## Результаты применения методов к реальным данным.

Первым протестируем на выбранных реальных данных традиционный метод. Сначала необходимо определить число факторов. Для этого вновь воспользуемся методом, основанным на величине приращения объясняемого среднеквадратического отклонения. Запускаем традиционный метод для максимального числа факторов, равного одиннадцати. Получаем следующие значения коэффициентов :

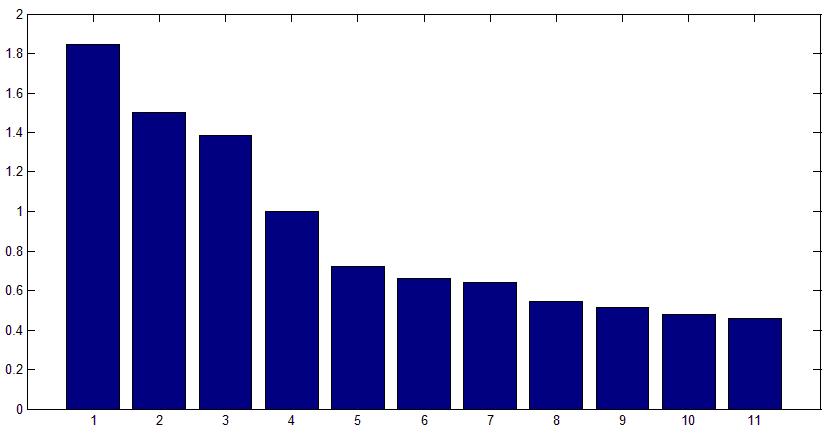


Рисунок 4.2. Приращения объясняемого среднеквадратического отклонения для факторов с первого по одиннадцатый для традиционного метода

 четвертого фактора получилось близко к единице. Возникает спорный вопрос – оставить или отбросить четвертый фактор. Учитывая, что мы работаем в предположении, что групп у нас три, все же отбросим.

Запускаем традиционный метод на трех факторах. Первоначальная нагрузочная матрица:



Разбиение на группы, несомненно, просматривается, но далеко не идеальное. Например, четвертая строка отличается первым элементом от остальных в своей группе, да и все строки второй группы лишь отдаленно похожи. Коэффициенты эффективности тоже оказались невелики:



Применим к полученной матрице метод вращения «варимакс»:



Коэффициенты эффективности:



Как видим, варимакс нас не подвел. Коэффициенты эффективности значительно превышают предыдущие. Полученную матрицу уже можно легко проинтерпретировать и эта интерпретация в основном совпадает с нашими ожиданиями. Каждая группа продуктов действительно выделяется в свой отдельный фактор. Это подтверждает, что изменение цен на продукты каждой группы имеет свою собственную тенденцию.

Тем не менее, можно заметить, что четвертая строка, соответствующая вареной колбасе, несколько отличается от других строк своей группы. То же самое можно сказать и седьмой строке, соответствующей сметане. Однако, вспомним что традиционный метод максимального правдоподобия предназначен только для работы с линейно зависимыми признаками. Поскольку мы применяем его к реальным данным, зависимость между которыми вполне может носить нелинейный характер, метод может выдавать несколько ошибочные результаты.

Результаты применения второго метода вращения – квартимакса практически точно совпадают с результатами варимакса. Вот они:



И коэффициенты эффективности:



Теперь перейдем ко второму методу, который использует в качестве меры связи коэффициент ранговой корреляции Спирмена. Выясняем число общих факторов. Полученные значения коэффициентов для одиннадцати оных:

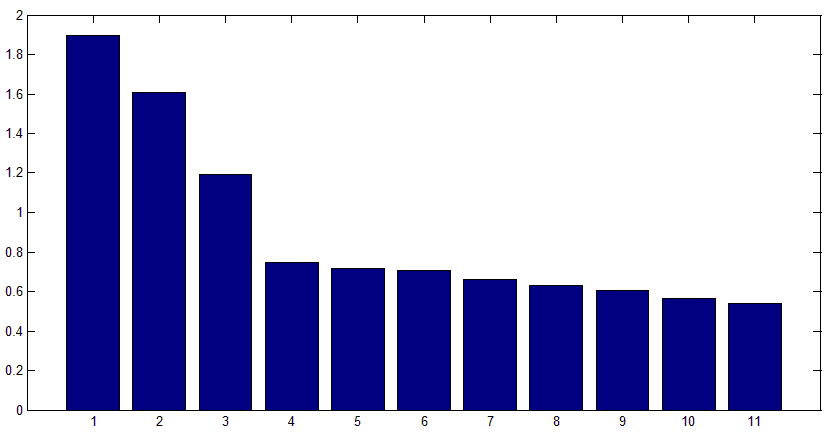


Рисунок 4.3. Приращения объясняемого среднеквадратического отклонения для факторов с первого по одиннадцатый для второго метода

Как и ожидалось, метод утверждает наличие трех факторов. Запускаем метод для них. Получаем первоначальную нагрузочную матрицу:



Коэффициенты эффективности:



Не вызывает сомнений, что этот метод действительно, как и ожидалось, объединяет в группы первые пять продуктов и последние два. Также очевидно, что он объединяет в группу продукты 6. – 8. Что касается 9-го продукта - сыра, то соответствующая ему строка все же несколько отличается от других строк «молочной» группы. Это может объясняться, например, тем, что, в отличие от сливочного масла, сметаны и творога, сыр необходимо достаточно долгое время выдерживать при изготовлении. Грубо говоря, если из одного и того же молока приготовить масло, сметану, творог и сыр, то сыр попадет на прилавок магазина значительно позже, чем три другие продукта. Таким образом, зависимость изменения цены на сыр от изменения цены на молоко вполне может несколько отличаться от зависимостей изменений цен на масло, сметану и творог от изменения цены на молоко. Однако при определении числа факторов, значение коэффициента  для четвертого фактора оказалось значительно меньше единицы. Из этого можно сделать вывод, что метод все же объединяет сыр в одну группу с другими молочными продуктами.

Применим к полученной матрице варимакс:



Коэффициенты эффективности:



Применим квартимакс:



Коэффициенты эффективности:



Как и в случае с традиционным методом, и варимакс, и квартимакс значительно улучшили интерпретируемость полученной нагрузочной матрицы. Как и ожидалось, метод утверждает, что каждая группа продуктов выделяется в свой особый фактор. В отличии от результатов традиционного метода, четвертая и седьмая строки, соответствующие вареной колбасе и сметане, не отличаются от других строк своих групп. Значительно прояснилась и ситуация с сыром: метод определенно выделяет его в тот же фактор, что и остальные молочные продукты, но утверждает, что его цена меняется несколько иначе, чем цена на масло, сметану и творог.

Коэффициенты эффективности для этого метода оказались несколько выше, чем для традиционного. (Мы сравниваем только те, которые получаются после применения методов вращения). Однако это превышение значительно только для второго коэффициента, соответствующего молочной группе. Поэтому можно утверждать лишь то, что второй метод оказывается более эффективным, чем традиционный, только для второй группы продуктов. Для двух других групп их эффективность одинакова.

Применяем третий метод. Полученные значения коэффициента  для одиннадцати факторов следующие:

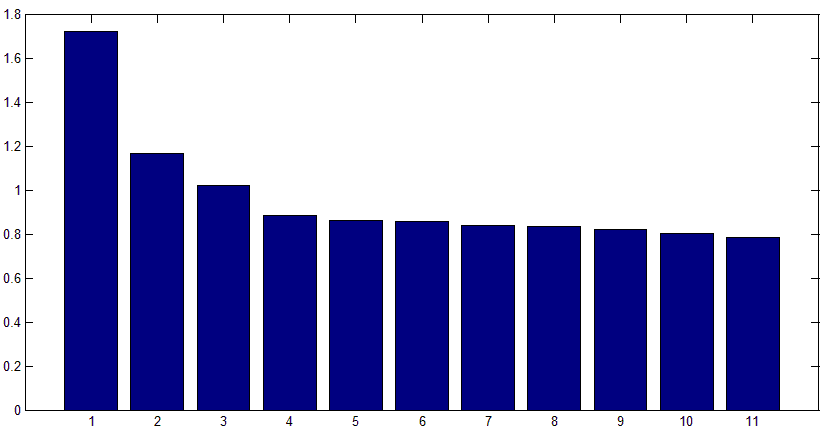


Рисунок 4.4. Приращения объясняемого среднеквадратического отклонения для факторов с первого по одиннадцатый для третьего метода

Опять же спорный вопрос – сколько факторов оставить три или два. Учитывая наши априорные соображения, оставляем три. Получаем следующую нагрузочную матрицу:



И коэффициенты эффективности:



Разбиение строк на группы «похожести» оказалось несколько более четким, чем в случае предыдущих двух методов, однако коэффициенты эффективности – несколько ниже. Посмотрим, повысят ли их методы вращения.

Применяем варимакс:



Коэффициенты эффективности:



Применяем квартимакс:



Коэффициенты эффективности:



Как видим, девятая строка, соответствующая сыру, опять несколько «выбивается» из своей группы. Поскольку то же самое наблюдалось в результатах второго метода, скорее всего, это не ошибка – цена на сыр действительно меняется немного не так, как цены на другие молочные продукты. Строки, соответствующие вареной колбасе и сметане, как и в предыдущем случае, не выбиваются из своих групп. На основании этого можно утверждать, что изменения цен на эти продукты имеют ту же тенденцию, что и изменения цен на другие продукты мясной и молочной групп соответственно, а результаты, полученные традиционным методом, объясняются тем, что метод не предназначен для нелинейных типов зависимости.

Что касается коэффициентов эффективности, они все равно оказываются ниже, чем коэффициенты предыдущих двух методов, полученные после вращения.

Таким образом, возможны два варианта:

1. Наше предположение касательно структуры данных верно, и тогда третий метод на этих данных менее эффективен, чем другие два метода.
2. В структуре данных имеются некоторые неучтенные нами зависимости. Тогда низкие коэффициенты эффективности, (которые на самом-то деле показывают лишь то, насколько результаты метода соответствуют нашим предположениям), наоборот, говорят о том, что третий метод способен замечать такие типы зависимостей, которые не видят предыдущие два метода (что, между прочим, подтвердилось для искусственно смоделированных данных). В поддержку этого предположения стоит заметить, что в нагрузочной матрице, полученной третьим методом, поэлементные разности строк, принадлежащих одной и той же группе, получаются меньше, чем в случае двух других методов. Т.е. третий метод лучше разбивает продукты в группы «похожести». Таким образом, если бы под эффективностью метода мы изначально понимали не расстояние нагрузочной матрицы до матрицы , а расстояние между строками нагрузочной матрицы, соответствующими одной группе, то третий метод был бы признан более эффективным, чем традиционный метод.

## Выводы по главе.

Итак, если под степенью эффективности метода понимать то, насколько столбцы полученной им нагрузочной матрицы близки (в смысле среднеквадратической ошибки) к столбцам нагрузочной матрицы, которая была бы получена гипотетическим идеальным для наших данных методом, то результаты экспериментов с реальными данными оказались следующими. Наиболее эффективным (для конкретных этих данных) оказался метод, использующий коэффициенты ранговой корреляции Спирмена. От него совсем немного отстает традиционный метод максимального правдоподобия. Наименее эффективным оказался метод, использующий коэффициенты Крамера.

Однако, если проанализировать полученные этими методами нагрузочные матрицы на основании одного только здравого смысла, не пытаясь формализовать понятие эффективности, то с точки зрения автора данного дипломного проекта, метод, использующий коэффициенты Крамера все же превосходит традиционный метод максимального правдоподобия, поскольку полученная им матрица обладает лучшей интерпретируемостью. (Если конкретнее, то в отличии от традиционного метода, метод, использующий коэффициенты Крамера, не утверждает, что тенденции изменений цен на вареную колбасу и сметану несколько отличаются от тенденций изменений цен на другие мясные и молочные продукты соответственно.)

Итак, можно утверждать, что с помощью каждого из трех рассмотренных методов нам удалось снизить размерность реальных данных с одиннадцати до трех. Это означает что, все три метода применимы для работы над реальными практическими задачами. Однако, поскольку нагрузочные матрицы этих методов все же несколько различаются, то для достижения наилучших результатов, автор предлагает при решении реальных задач использовать все три метода, анализировать полученные ими результаты, помня о том, какой из методов для какого типа зависимости предназначен. И уже на основании этого анализа принимать окончательное решение.

# Заключение

Итак, в рамках дипломного проекта автором был разработан программный модуль, реализующий метод максимального правдоподобия в модели факторного анализа. Этот модуль по выборочной корреляционной матрице признаков вычисляет значения нагрузочной матрицы и матрицы остаточных дисперсий, которые максимизируют правдоподобие данной выборочной корреляционной матрицы. Кроме того, были реализованы метод определения числа общих факторов, основанный на критерии значимости и эмпирический метод определения числа общих факторов, основанный на величине приращения объясняемого среднеквадратического отклонения, который в дальнейшем был использован при работе с реальными данными.

Кроме того, был предложен способ адаптации традиционного метода максимального правдоподобия для работы с признаками, связанными нелинейными типами зависимости. Этот способ заключается в замещении выборочного коэффициента корреляции признаков на другие меры связи – коэффициент ранговой корреляции Спирмена и коэффициент Крамера, которые являются более информативными для нелинейно зависимых случайных величин.

Для сравнения качества работы традиционного метода правдоподобия и двух его модификаций, автором было формализовано понятие эффективности этих методов. Исходя из этого понятия, эффективность методов была сравнена как на искусственно смоделированных признаках, связанных различными типами зависимости, так и на реальных данных.

Результаты этого сравнения для искусственно смоделированных данных оказались следующими. Традиционный метод максимального правдоподобия хорошо работает в случаях, когда признаки связаны линейными зависимостями и теми, которые хорошо приближаются линейными. В том числе он оказывается эффективным и для большинства зависимостей монотонного типа, поскольку они тоже могут быть приближены линейными функциями, пусть и менее точно. Для зависимостей немонотонного типа этот метод не работает.

Второй метод хорошо работает для любых зависимостей монотонного типа. Причем, с точки зрения степени эффективности, он не отличается от традиционного метода. На зависимостях немонотонного типа он тоже не работает.

Третий метод удовлетворительно работает как на монотонных, так и на немонотонных типах зависимости, однако на монотонных уступает двум другим методам в эффективности.

Учитывая, что традиционный метод максимального правдоподобия оказался совершенно неэффективным для немонотонного типа зависимости, а метод, использующий коэффициенты Крамера, справился с этим типом зависимости вполне успешно, можно сделать вывод, что автору данного дипломного проекта удалось адаптировать традиционный метод максимального правдоподобия для работы с нелинейно зависимыми признаками.

Результаты экспериментов с реальными данными оказались следующими. Каждым из трех методов удалось добиться снижения размерности исследуемых данных. Следовательно, все три метода применимы для решения реальных практических задач. Наиболее эффективным признан метод, использующий коэффициенты ранговой корреляции Спирмена, следующим по эффективности – традиционный метод максимального правдоподобия, и наименее эффективным – метод, использующий коэффициенты Крамера.

Однако, если отказаться от формального понятия эффективности, и сравнивать качество работы методов на реальных данных, исходя из здравого смысла, то, возможно, метод использующий коэффициенты Крамера тоже превосходит традиционный метод максимального правдоподобия.

В любом случае, метод, использующий коэффициенты ранговой корреляции Спирмена, оказывается эффективнее традиционного метода. Таким образом, можно сделать вывод, что автору данного данной работы удалось с помощью модификаций повысить эффективность традиционного метода максимального правдоподобия для решения реальных практических задач, что означает выполнение основной задачи дипломного проекта.

# Приложение

Фрагмент программного кода, написанного в Матлабе, реализующий функцию «generateCorr»:

function X=generateCorr(r,n,corr)

X=zeros(r,n);

X(1,:)=randn(1,n);

DX=1;

for j=2:r

alpha=sqrt((corr^2/(1-corr^2))/DX);

X(j,:)=X(j-1,:)\*alpha+randn(1,n);

DX=alpha^2\*DX+1;

End

Фрагмент программного кода, реализующий функцию «generateData»:

function X=generateData(F,m,n)

r=size(F,1)\*m;

X=zeros(r,n);

for i=0:size(F,1)-1

f=F{i+1};

X(i\*m+1,:)=randn\_cut(1,n);

for j=2:m

X(i\*m+j,:)=f(X(i\*m+j-1,:));

end

end

X=X+randn(r,n)\*0.1;

Фрагмент программного кода, реализующий функцию «randn\_cut»:

function X=randn\_cut(a,b)

X=randn(a,b);

E=(X>1)|(X<-1);

while nnz(E)~=0

X(E)=randn(size(X(E)));

E=(X>1)|(X<-1);

end

# Список литературы

**Roger R. Davidson, William E. Lever** The limiting distribution of the likelihood ratio statistic under a class of local alternatives [Книга]. - [б.м.] : Florida State University, 1967.

**Горяинова Е. Р., Панков А. Р., Платонов Е. Н.** Прикладные методы анализа статистических данных [Книга]. - Москва : Высшая Школа Экономики, 2012.

**Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка, М. С. Олдендерфер, Р. К. Блэшфилд** Факторный, дискриминантный и кластерный анализ [Книга]. - Москва : Финансы и статистика, 1989.

**Дронов С.В.** Многомерный статистический анализ [Книга]. - Барнаул : Алтайский государственный университет, 2003.

**Ивченко Г. И., Медведев Ю. И.** Введение в математическую статистику [Книга]. - Москва : ЛКИ, 2010.

**Калинина В. Н., Соловьев В. И.** Введение в многомерный статистический анализ [Книга]. - Москва : ГУУ, 2003.