Многомерные модели анализа данных

Курс лекций
«Методы многомерного анализа в социологических исследованиях»
(лекция 1-2)

Преподаватель: Цихончик Надежда Васильевна, старший преподаватель кафедры философии и социологии СГНиМК САФУ

План лекции

1. Понятие о многомерных методах анализа данных
2. Регрессионный анализ
3. Факторный анализ
4. Дискриминантный анализ
5. Кластерный анализ
6. Многомерное шкалирование

Методы многомерного анализа (multivariate analysis methods)

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ [multidimensional, multivariate statistical analysis] — раздел математической статистики, объединяющий методы изучения статистических данных, которые являются значениями многомерных качественных или количественных признаков

Классификация многомерных методов

По назначению:

- Методы предсказания (экстраполяции): множественный регрессионный и дискриминантный анализ
- Методы классификации: варианты кластерного анализа (без обучения) и дискриминантный анализ
- Структурные методы: факторный анализ и многомерное шкалирование

Классификация многомерных методов

По исходным предположениям о структуре данных:

- Методы, исходящие из предположения о согласованной изменчивости признаков: факторный, множественный регрессионный, отчасти – дискриминантный анализ
- Методы, исходящие из предположения о том, что различия между объектами можно описать как расстояние между ними: кластерный анализ, многомерное шкалирование

Классификация многомерных методов

По виду исходных данных:

- Методы, использующие в качестве исходных данных только признаки, измеренные у группы объектов: множественный регрессионный, дискриминантный, факторный анализ
- Методы, исходными данными для которых могут быть попарные сходства (различия) между объектами: кластерный анализ и многомерное шкалирование

2 вопрос лекции. Регрессионный анализ

- Цель множественного регрессионного анализа (МРА) – изучение взаимосвязи одной переменной (зависимой, результирующей) от нескольких других переменных (зависимых, исходных)
- Наиболее часто этот метод применяется для предсказания результата (обучения, деятельности) по ряду предварительно измеренных характеристик

Основные задачи МРА

1. Определение того, в какой мере «зависимая» переменная связана с совокупностью «независимых переменных», какова статистическая значимость этой взаимосвязи. Показатель – коэффициент множественной корреляции (КМК) и его статистическая значимость по Ф-критерию Фишера.
2. Определение существенности вклада каждой «независимой» переменной в оценку «зависимой» переменной, отсев несущественных для предсказания «независимых» переменных. Показатель – регрессионные коэффициенты β, их статистическая значимость по критерию Стьюдента
3. Анализ точности предсказания и вероятных ошибок оценки «зависимой» переменной. Показатель – квадрат КМК, интерпретируемый как доля дисперсии «зависимой» переменной, объясняемая совокупностью «независимых» переменных. Вероятные ошибки предсказания анализируются по расхождению (разности) действительных значений «зависимой» переменной и оцененных при помощи модели МРА.
4. Оценка (предсказание) неизвестных значений «зависимой» переменной по известным значениям «независимых» переменных. Осуществляется по вычисленным параметрам множественной регрессии.

Исходные данные МРА

Исходной для МРА является матрица данных, включающая в себя НП и ЗП, измеренные для группы объектов (испытуемых).

Главное требование к исходным данным – отсутствие линейных взаимосвязей между переменными, когда одна переменная является линейной производной другой переменной; переменные должны быть измерены на метрической шкале (интервалов или отношений) и иметь нормальное

Регрессионный анализ

- основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной
- Уравнение регрессии выглядит следующим образом: Y=a+b*X
- При помощи этого уравнения переменная Y выражается через константу a и угол наклона прямой (или угловой коэффициент) b, умноженный на значение переменной X. Константу a также называют свободным членом, а угловой коэффициент - коэффициентом регрессии или B-коэффициентом

Регрессионный анализ

- Остаток - это отклонение отдельной точки (наблюдения) от линии регрессии (предсказанно значения)

Лекция «Основы анализа данных»
http://www.intuit.ru/department/database/datamining/8/4.html

Регрессионный анализ

Таблица 8.3а. Регрессионная статистика

Регрессионная статистика	
Множественный R	0,998364
R-квадрат	0,99673
Нормированный R-квадрат	0,996321
Стандартная ошибка	0,42405
Наблюдения	10

Таблица 8.3б. Коэффициенты регрессии

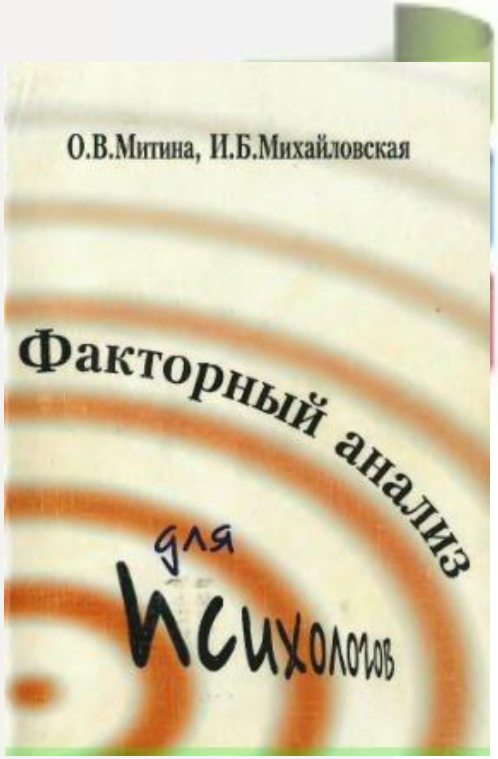
	Коэффициенты	Стандартная ошибка	t-статистика
Y-пересечение	2,694545455	0,33176878	8,121757129
Переменная X 1	2,305454545	0,04668634	49,38177965
* Приведен усеченный вариант расчетов			

Таблица 8.3в. Остатки

Наблюдение	Предсказанное Y	Остатки	Стандартные остатки
1	9,610909091	-0,610909091	-1,528044662
2	7,305454545	-0,305454545	-0,764022331
3	11,91636364	0,083636364	0,209196591
4	14,22181818	0,778181818	1,946437843
5	16,52727273	0,472727273	1,182415512
6	18,83272727	0,167272727	0,418393181
7	21,13818182	-0,138181818	-0,34562915
8	23,44363636	-0,043636364	-0,109146047
9	25,74909091	-0,149090909	-0,372915662
10	28,05454545	-0,254545455	-0,636685276

Таблица 8.4. Результаты прогнозирования переменной Y

x	Y(прогнозируемое)
11	28,05455
12	30,36
13	32,66545
14	34,97091
15	37,27636
16	39,58182



3 вопрос лекции. Факторный анализ

Факторный анализ

- многомерный статистический метод, применяемый для изучения взаимосвязей между значениями переменных
- (Factor analysis) Метод, используемый для определения скрытых психологических переменных личности или скрытых переменных в вопросах тестов, которые выявляются при обработке корреляционной матрицы.
- Главными целями факторного анализа являются: (1) *сокращение* числа переменных (редукция данных) и (2) *определение структуры* взаимосвязей между переменными, т.е. *классификация переменных*

Виды факторного анализа

два основных типа факторного анализа:

- *эксплораторный* (разведочный) - используется на ранних этапах исследования как инструмент для объединения в группы первичных переменных и для порождения гипотез относительно структуры латентных факторов
- *конфирматорный* (подтверждающий гипотезу) - используется на более поздних стадиях работы для подтверждения уже выстроенной гипотезы о латентной структуре

Факторный анализ. Немного истории

- Точный момент возникновения метода факторного анализа определить достаточно трудно.
- Если отсчитывать его историю от изобретения Ф. Гальтоном коэффициента корреляции, то это середина 1880-х гг. Работая с антропометрическими данными, Пирсон в 1901 г. выдвинул идею «главных осей»,
- рождение факторного анализа как метода исследования связывают с публикацией в 1904 г. статьи Спирмэна «Объективное определение и измерение общего интеллекта». На основе статистического анализа тестов Спирмэн выдвинул двухфакторную теорию интеллекта

Факторный анализ. Немного истории

- В нашей стране обсуждение основ факторного анализа началось еще в 1930-х гг.
- Новый этап развития этого метода в СССР начался в 1950-х гг. в антропологии
- Небылицын (1960) - называя факторный анализ скорее искусством, предоставляющим немалый простор для субъективных интерпретаций и выводов, автор все же предлагает психологам познакомиться с теорией, основными предпосылками, логикой и техникой этого метода
- свое окончательное название на русском языке метод факторного анализа получил благодаря работе Теплова
- имена коллег, наиболее часто использующих факторный анализ сегодня, - «отцы-основатели» психосемантического направления — В. Ф. Петренко и А. Г. Шмелев

Факторный анализ

- Переменные, входящие в одно подмножество и коррелирующие между собой, но в значительной степени независимые от переменных из других подмножеств, образуют **факторы**
- Цель факторного анализа — идентифицировать явно не наблюдаемые факторы с помощью множества наблюдаемых переменных.
- В основе парадигмы использования факторного анализа лежит предположение о том, что выделяемые факторы отражают глубинные процессы (латентные, не наблюдаемые, не измеряемые), являющиеся причиной корреляций первичных (наблюдаемых, измеряемых) переменных. Другими словами, факторы (глубинные параметры) детерминируют (определяют) первичные наблюдаемые переменные и могут быть использованы для объяснения комплексных явлений. Наблюдаемые корреляции между первичными переменными возникают из-за того, что их детерминируют одни и те же факторы.

Структура (алгоритм) анализа

1. *Подготовка исходной матрицы данных*
2. *Вычисление матрицы взаимосвязей признаков*
3. *Факторизация* (при этом необходимо указать количество факторов, выделяемых в ходе факторного решения, и метод вычисления).
4. *Вращение* — преобразование факторов, облегчающее их интерпретацию
5. *Подсчет факторных значений* по каждому фактору для каждого наблюдения
6. *Интерпретация данных*

1. Подготовка исходных данных

- Практически во всех процедурах любой программы факторного анализа в качестве исходных данных используются матрицы. *Матрица* — это прямоугольная (в частном случае квадратная) таблица чисел, в которой, как правило, горизонтальные линии (строки, ряды) соответствуют наблюдениям (объектам), а вертикальные линии (столбцы) — переменным

Факторный анализ

Обязательные условия факторного анализа

- Все признаки должны быть количественными.
- Число признаков должно быть в два раза больше числа переменных.
- Выборка должна быть однородна.
- Исходные переменные должны быть распределены симметрично.
- Факторный анализ осуществляется по коррелирующим переменным

2. Вычисление матрицы взаимосвязей признаков

- Процедура факторного анализа начинается с вычисления *матрицы взаимосвязей* переменных между собой (это квадратная матрица, размер которой равен количеству переменных).
- Наиболее распространенная мера взаимосвязи (используемая в факторном анализе в 95% случаев) — это корреляционная связь

3. Факторизация

- Проблемы:

1. критериев, которые позволяли бы проверить правильность найденного решения, не существует
2. после выделения факторов возникает бесконечное множество вариантов вращения, базирующихся на тех же исходных переменных, но дающих разные решения
3. факторный анализ довольно часто применяют с целью спасти плохо продуманное исследование

3. Факторизация

1. гипотеза относительно того, какие факторы могли бы описывать предметную область. Статистически очень важно, чтобы экспериментальное исследование было достаточно широким и можно было бы выделить не менее пяти-шести гипотетических факторов
2. выбор переменных для наблюдения - *маркерные переменные* - маркерные переменные должны быть в высокой степени взаимосвязаны с одним и только одним фактором и иметь по нему высокие нагрузки вне зависимости от того, с помощью какого алгоритма выделялись и вращались факторы

Матрицы, наиболее часто используемые в факторном анализе

Обозначение	Название	Размер	Описание
R	Матрица взаимосвязей	$p \times p$	Взаимосвязи между переменными
D	Матрица нестандартизированных данных	$N \times p$	Первичные данные — нестандартизированные значения наблюдений по первичным переменным
Z	Матрица стандартизированных данных	$N \times p$	Стандартизованные значения наблюдений по первичным переменным
F	Матрица значений факторов	$N \times f$	Стандартизированные значения наблюдений по факторам
A	Матрица факторных нагрузок Матрица факторного отображения	$p \times f$	Коэффициенты регрессии для общих факторов при условии, что наблюдаемые переменные являются линейной комбинацией факторов. В случае ортогонального вращения — взаимосвязи между переменными и факторами
B	Матрица коэффициентов значений факторов	$p \times f$	Коэффициенты регрессии для вычисления значений факторов с помощью значений переменных
S	Структурная матрица	$p \times f$	Взаимосвязи между переменными и факторами

Продолжение таблицы 1

S	Структурная матрица	$p \times f$	Взаимосвязи между переменными и факторами
Φ	Матрица корреляций факторов	$f \times f$	Корреляции между факторами
L	Матрица собственных значений (диагональная)	$f \times f$	Собственные значения (характеристические, латентные корни); каждому фактору соответствует одно собственное число
V	Матрица собственных векторов	$p \times f$	Собственные (характеристические) вектора; каждому собственному числу соответствует один собственный вектор

Примечание. При указании размера дается количество рядов × количество столбцов: p — количество переменных, N — количество наблюдений, f — количество ров или компонент. Если матрица взаимосвязей R не вырождена и имеет ранг равный p, то тогда фактически выделяется p собственных чисел и собственных ров, а не f. Однако интерес представляют только f из них. Поэтому оставшиеся p—f не показываются.

матрицам S и Φ применяется только косоугольное вращение, к остальным — ортогональное и косоугольное.

3. Факторизация

3. Матрица взаимосвязей должна быть *факторизуемой,* т.е. корреляции в ней должны быть больше 0.3
4. Переменная с низким квадратом множественной корреляции с другими переменными и слабой взаимосвязью со всеми значимыми факторами представляет собой *постороннюю переменную*. Ее лучше исключить из модели.

4. Вращение

- *Поворот* факторов — это процесс поиска наиболее легко интерпретируемого решения для данного количества факторов
- Вращение обычно применяется после выделения факторов для максимизации высоких корреляций и минимизации низких
- Существуют два основных класса поворотов: *ортогональный* и *косоугольный*
- Существуют многочисленные методы вращения, но чаще всего используется поворот ***варимакс,*** представляющий собой процедуру максимизации дисперсий.

4. Варимакс-вращение

- Этот поворот максимизирует дисперсии факторных нагрузок, делая высокие нагрузки выше, а низкие ниже для каждого из факторов.
- У матрицы после поворота низкие факторные нагрузки ниже, а высокие выше, чем у матрицы до поворота. Подчеркнутая разница нагрузок облегчает интерпретацию фактора, позволяет однозначно выбрать сильно взаимосвязанные с ним переменные
- *Матрица преобразования* — это матрица синусов и косинусов угла Ψ, на который выполняется поворот. (Отсюда и название преобразования — *поворот*, потому что с геометрической точки зрения происходит поворот осей вокруг начала координат факторного пространства)

5. Подсчет факторных значений

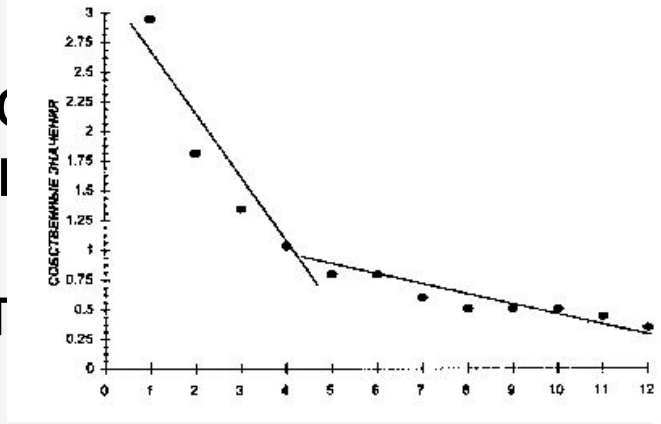
1. Общность переменной – доля дисперсии фактора. Например, первый фактор объясняет 50% дисперсии переменных. Второй фактор объясняет 48% дисперсии переменных и (в силу ортогональности вращения) два фактора в сумме объясняют 98% дисперсии переменных.
2. Доля дисперсии решения, объясняемая фактором, — **доля *ковариации***

Адекватная факторная модель

- Нахождение наиболее адекватной факторной модели связано с определением количества факторов
- ***несколько часто употребляемых критериев:***
- различные правила, формулируемые в терминах собственных чисел;
- критерий следа (отсеивания);
- критерии значимости, связанные с методами максимального правдоподобия и наименьших квадратов;
- критерий, основанный на величине долей дисперсий факторов;
- критерий факторных нагрузок;
- критерий интерпретируемости и инвариантности.

Адекватная факторная модель: методы отбора

1. оценка собственных чисел и введение критерия значимости фактора при наличии собственного числа > 1
2. анализ «следа» - на графике виден отчетливый излом между крутым наклоном первых факторо и постепенным убыванием остальных. Этот постепенный «сход на нет от найденной точки получил название «*scree*» (след)



Адекватная факторная модель

- вычислительная процедура факторного анализа представляет собой многоступенчатый процесс, допустимо принимать решение о количестве остающихся факторов на различных этапах расчета — либо в процессе выделения факторов, либо после этого. Однако лишь на последних этапах получают важную информацию о количестве факторов, которые следует оставить.
- Основная стратегия при этом состоит в том, чтобы вначале выделить на один фактор больше, а затем либо отбросить его, либо оставить на основании дальнейших результатов анализа и дополнительных критериев

Простота структуры

- Простая структура имеет не слишком сильно взаимосвязанные между собой факторы.
- Несколько переменных сильно взаимосвязаны с каждым фактором и только один фактор сильно взаимосвязан с каждой переменной.
- Другими словами, столбцы матрицы факторных нагрузок *A*, определяющие факторы по отношению к переменным, имеют несколько высоких и много низких значений, тогда как строки матрицы *A*, определяющие переменные по отношению к факторам, имеют по одному высокому значению.
- **Строки с более чем одной высокой факторной нагрузкой** соответствуют переменным, считающимся сложными в связи с тем, что они отражают влияние более чем одного фактора.
- Обычно следует избегать сложных переменных, поскольку они затрудняют интерпретацию факторов.

5. Подсчет факторных значений

Переменная	Factor Loadings (Варимаксимальный н) (база данных.sta) Извлечение: Основные компоненты (Marked loadings are > ,700000) **Фактор 1**	Фактор 2
актив	0,870935	0,108304
общ	0,546735	0,439535
впеч	0,895009	0,108347
власть	0,719035	0,037098
ДЗД	0,803067	-0,228268
и.кнф	-0,556287	0,540316
творч	0,768984	0,184016
п.пом	0,117832	0,935611
манип	0,103660	0,764653
позн	0,927405	0,159064
Expl.Var	4,805917	2,081309
Prp.Totl	0,480592	0,208131

Факторные нагрузки - это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки.

6. Интерпретация факторов

- Чтобы интерпретировать фактор, исследователь пытается найти глубинное измерение, объединяющее группу переменных, имеющих по нему высокие нагрузки
- Процедура наименования фактора (присвоения ему названия или какого-то ярлыка) — процесс, требующий одновременно и творчества и научной обоснованности.

3 вопрос лекции. Факторный анализ

- http://www.statsoft.ru/home/textbook/modules/stfacan.html Электронный учебник Statsoft
- http://www.learnspss.ru/hndbook/glava19/cont4.htm Пример факторного анализа из области психологии
- http://psychlib.ru/mgppu/mit/MIT-001-.HTM О. В. Митина, И. Б. Михайловская. ФАКТОРНЫЙ АНАЛИЗ ДЛЯ ПСИХОЛОГОВ. Учебное пособие. М., 2001.

4 вопрос лекции. Дискриминантный анализ

Дискриминантный анализ

- метод многомерной статистики, предназначенных для 1) описания различий между классами и 2) классификации объектов, не входивших в первоначальную выборку обучающую

СВЯЗЬ С РЕГРЕССИОННЫМ И ДИСПЕРСИОННЫМ

Таблица 18.1. Сходства и отличия между дисперсионным, регрессионным и дискриминантным анализом

	Дисперсионный анализ	Регрессионный анализ	Дискриминантный анализ
Сходства			
Число зависимых переменных	Одна	Одна	Одна
Число независимых переменных	Несколько	Несколько	Несколько
Отличия			
Природа зависимой переменной	Метрическая	Метрическая	Категориальная
Природа независимой переменной	Категориальная	Метрическая	Метрическая

Требования к данным

- В модели должно быть не менее двух классов
- в каждом классе - не менее двух объектов из обучающей выборки,
- число дискриминантных переменных не должно превосходить объем обучающей выборки за вычетом двух объектов
- Дискриминантные переменные должны быть количественными и линейно независимыми (не должны коррелировать друг с другом)

СТАТИСТИКИ, СВЯЗАННЫЕ С ДИСКРИМИНАНТНЫМ АНАЛИЗОМ

- Каноническая корреляция
- Центроид
- Классификационная матрица
- Коэффициенты дискриминантной функции
- Дискриминантные показатели
- F-статистика и ее значимость
- Средние группы и групповые стандартные отклонения
- Объединенная межгрупповая корреляционная матрица
- Нормированные коэффициенты дискриминантных функций
- Структурные коэффициенты корреляции
- Общая корреляционная матрица
- Коэффициент л Уилкса

5 вопрос лекции. Кластерный анализ

Кластерный анализ

Кластер	Муж	30-50 лет	>50 лет	Рук.	Мед	Льготы	з/п	стаж	Образов.
1	80%	90%	5%	70%	10%	12%	95%	30%	30%
2	40%	35%	45%	13%	60%	70%	60%	40%	20%
3	50%	70%	10%	5%	30%	20%	70%	20%	50%

- **Кластерный анализ предназначен для разбиения совокупности объектов на однородные группы (кластеры или классы). По сути это задача многомерной классификации данных**

Задача кластерного анализа

заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов G на m (m – целое) кластеров (подмножеств) Q1, Q2, …, Qm, так,
чтобы каждый объект Gj принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и
тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными

Задачи кластерного анализа

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Проблемы кластерного анализа

- элементы (в нашем случае банки) характеризуются большим количеством факторов, которые имеют разные единицы измерения и разные абсолютные величины, буквально не сопоставимые друг с другом и несущие разный объем информации;
- первоначально неизвестно число кластеров, на которое необходимо разбить исходную совокупность элементов, и визуальные наблюдения в многомерном случае просто не приводят к успеху;
- какие метрики использовать в качестве меры расстояния (меры близости) между элементами;
- какую целевую функцию или метод использовать для объединения элементов в кластеры.

Данные для кластерного анализа

- Кластерный анализ можно применять к интервальным данным, частотам, бинарными данным. Важно, чтобы переменные изменялись в
сравнимых шкалах
- Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений к некоторой величине,
отражающей определенные свойства данного показателя

Кластер

- Кластер – это совокупность однородных элементов, идентичных объектов, образующих группу единиц
- Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.
- Центр кластера - это среднее геометрическое место точек в пространстве переменных.
- Радиус кластера - максимальное расстояние точек от центра кластера.

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

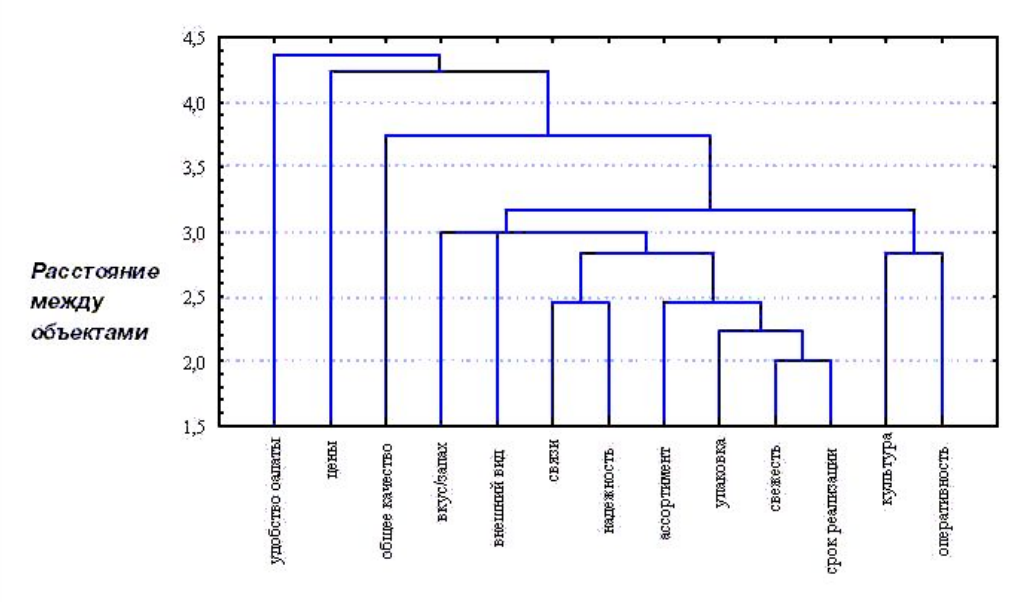
- иерархические;
- неиерархические.

В качестве основных методов анализа пакет **STATISTICA** предлагает **Joining** (**tree clustering**) – группу иерархических методов (7 видов), которые используются в том случае, если число кластеров заранее неизвестно, и **K-Means Clustering** (метод *K*-средних), в котором пользователь заранее определяет количество кластеров.

Методы кластерного анализа: иерархические

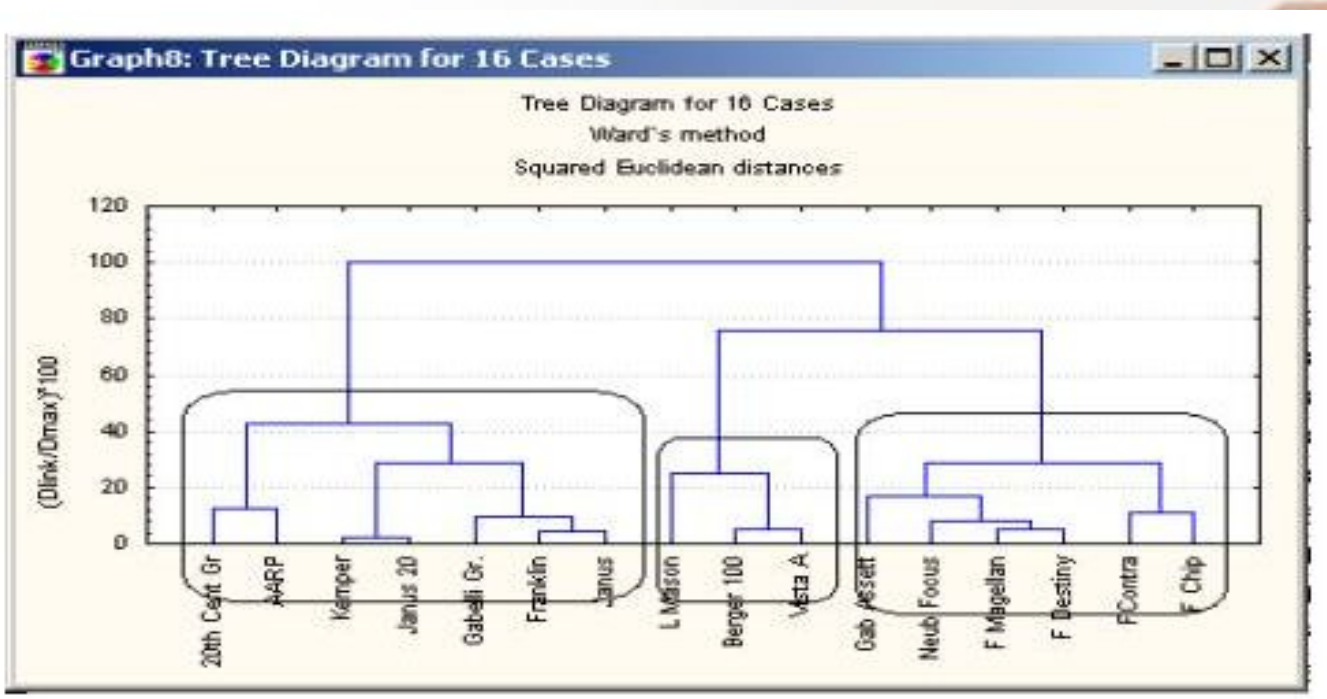
- Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие
- используются при небольших объемах наборов данных
- Преимуществом является их наглядность
- связаны с построением дендрограмм

Дендрограмма



Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Определение количества кластеров



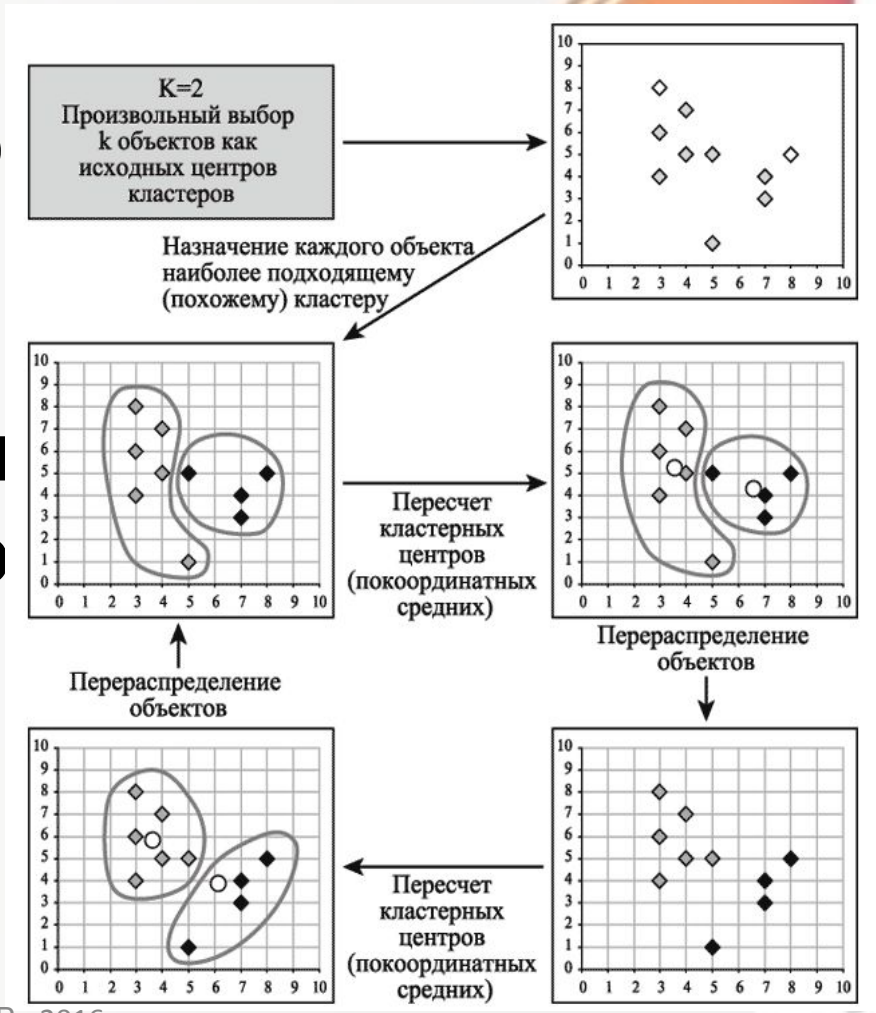
- способ сводится к определению скачкообразного увеличения некоторого коэффициента, который характеризует переход от сильно связанного к слабо связанному состоянию объектов

Методы кластерного анализа: неиерархические

- основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности
- В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки

K-Means Clustering (метод *K*-средних)

- для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеро



Сравнительный анализ иерархических и неиерархических методов кластеризации

- Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации. Ценой, которую приходится платить за эти достоинства метода, является слово "априори"

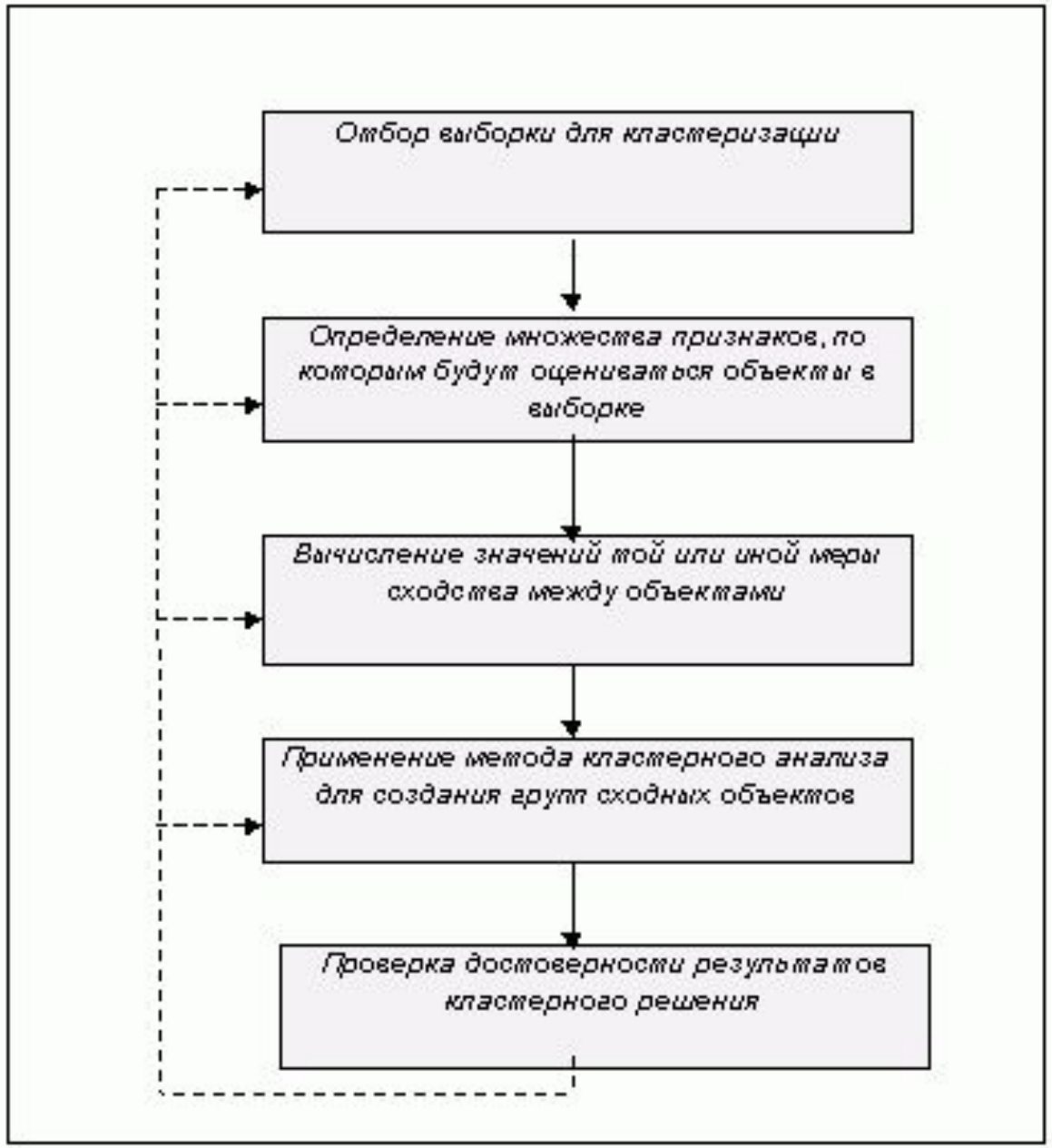


Рис. 1 Схема организации проведения кластерного анализа.

6 вопрос лекции. Многомерное шкалирование

Многомерное шкалирование

- семейство моделей и связанных с ними методов для представления данных о сходствах или различиях стимульных объектов либо др. элементов на основе заданной пространственной модели
- один из методов исследования структуры и снижения размерности пространства переменных
- Задача многомерного шкалирования в самом общем виде состоит в том, чтобы выявить структуру исследуемого множества стимулов

Спасибо за внимание!