Лекция 7.

Факторный анализ

Термин факторный анализ впервые ввел Thurstone, 1931

- Факторный анализ в современной статистике - совокупность методов, которые на основе реально существующих связей признаков, объектов или явлений позволяют выявлять *латентные* обобщающие характеристики организованной структуры и механизма развития изучаемых явлений или процессов.
- Понятие латентности является ключевым и означает неявность характеристик, раскрываемых при помощи методов факторного анализа.

К.Иберла: "Основная цель факторного анализа состоит в выявлении гипотетических величин, или факторов, по большому числу экспериментальных данных. …факторный анализ является методом, упорядочивающим кажущуюся хаотичность изучаемого явления и генерирующим новые гипотезы"

Факторный анализ - это выявление и обоснование действия различных признаков и их комбинаций на исследуемый процесс путем снижения их размерности.

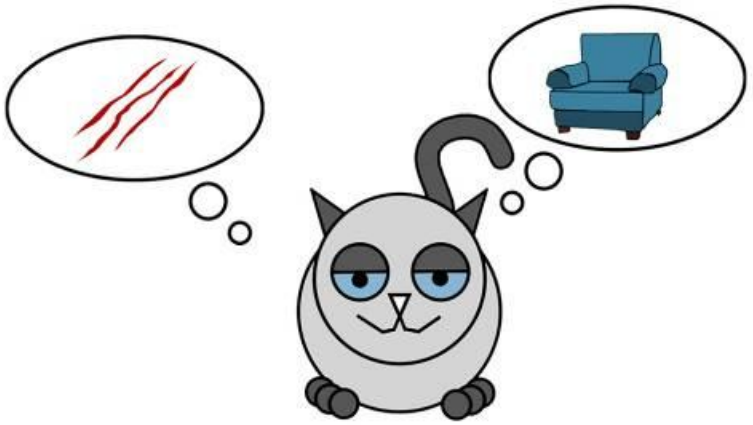
Такая задача решается путем «сжатия» исходной информации и выделения из нее наиболее «существенной» информации, т.е. описание объектов меньшим числом обобщенных признаков, называемых факторами.

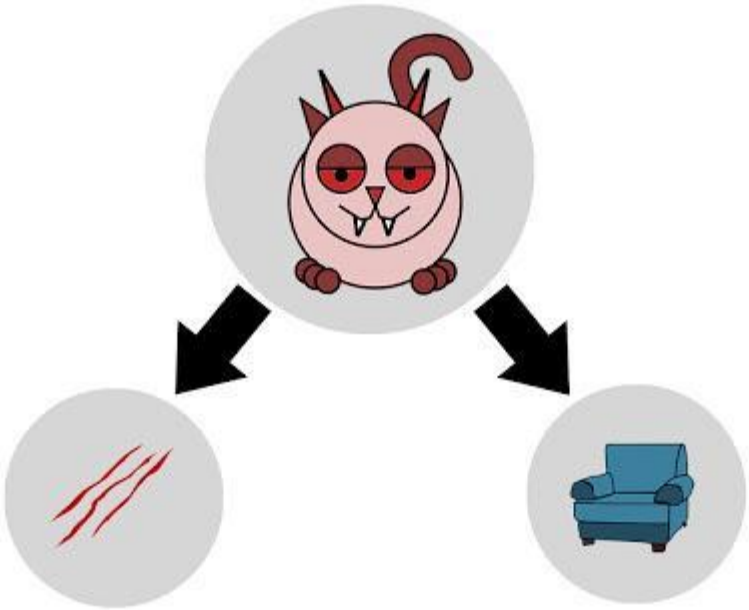
Модель факторного анализа связана с предположением, что связь между набором переменных обусловлена некой другой величиной, не поддающейся непосредственному измерению Измеряемые величины являются формой проявления фактора, объясняющего наблюдаемые связи.

Метод главных компонент и собственно факторный анализ

- *ФА предполагает разложение ф-ров на **общие** и **характерные**.*
- *В отличие от МГК не утверждается, что наблюдаемые признаки могут быть однозначно вычислены (без потери информации) по значениям общих факторов **f**.*

- Связь признаков

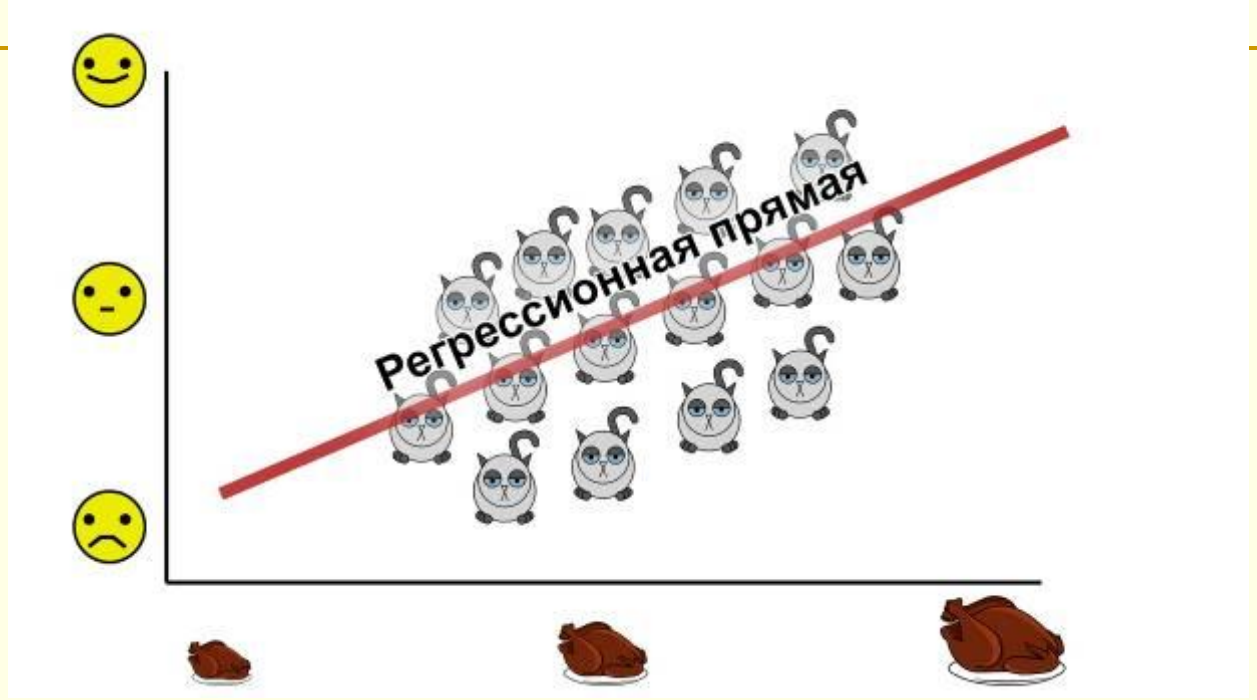


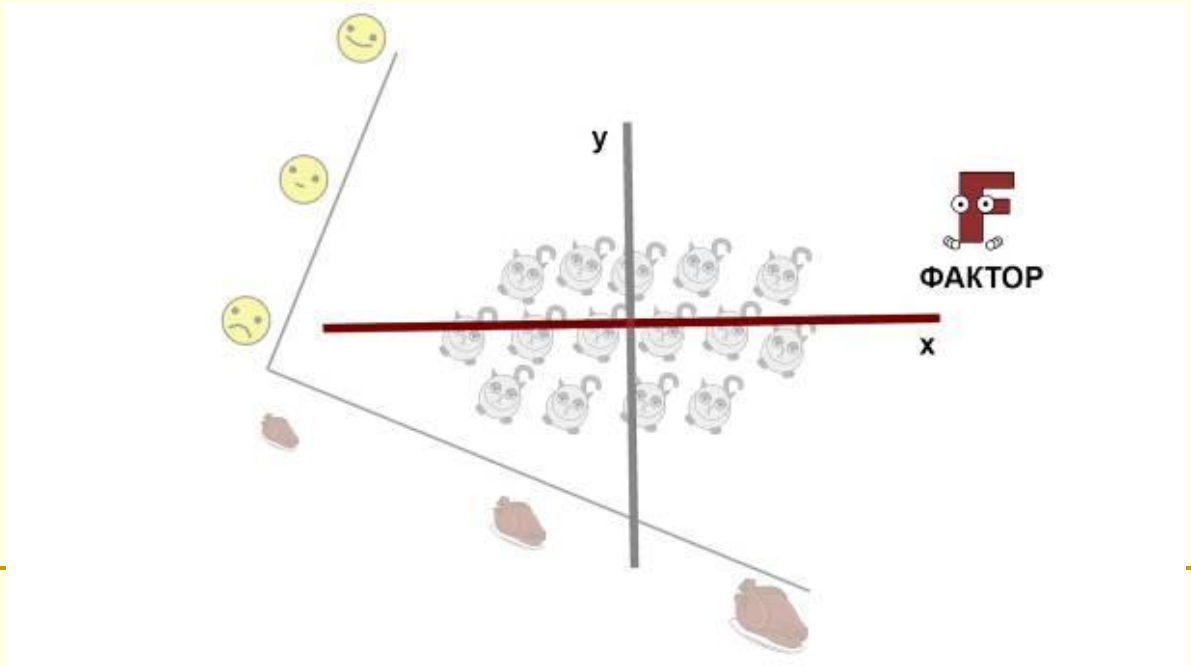


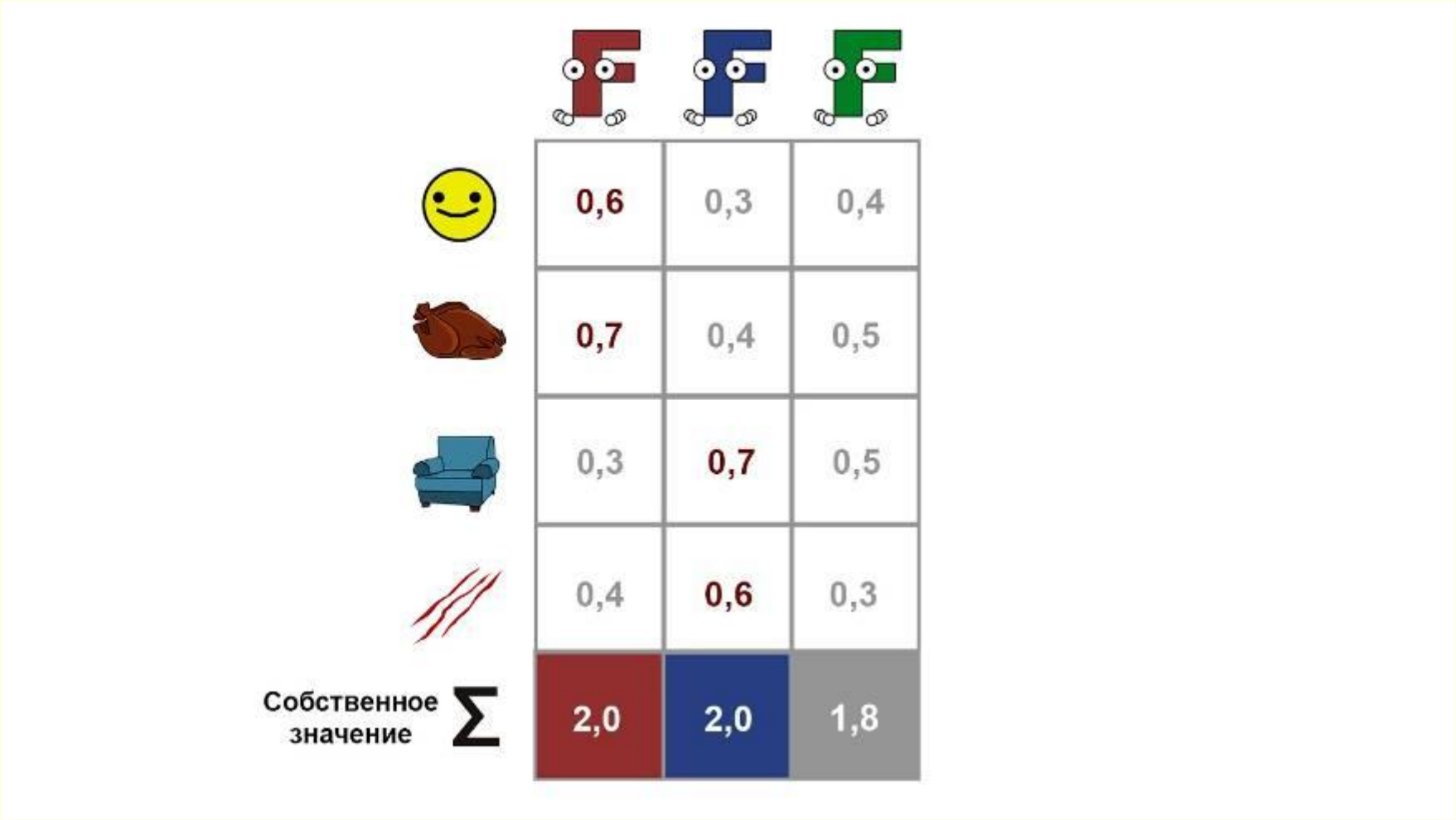
- Латентная переменная?

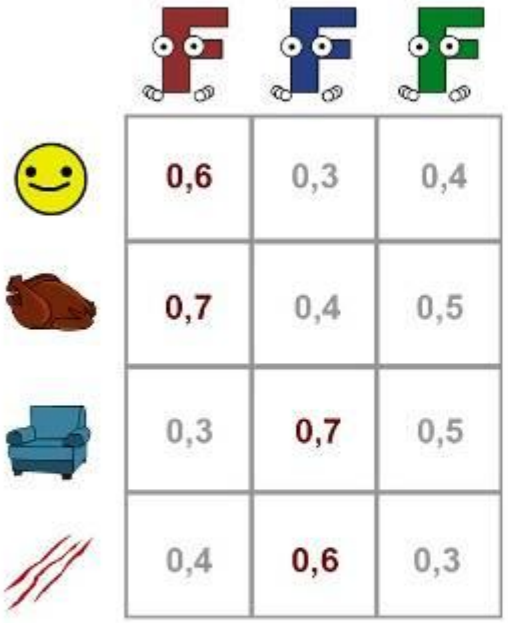
Царапучесть?

				
	1	0,9	0,2	0,3
	0,9	1	0,1	0,2
	0,2	0,1	1	0,8
	0,3	0,2	0,8	1


Регрессионная прямая

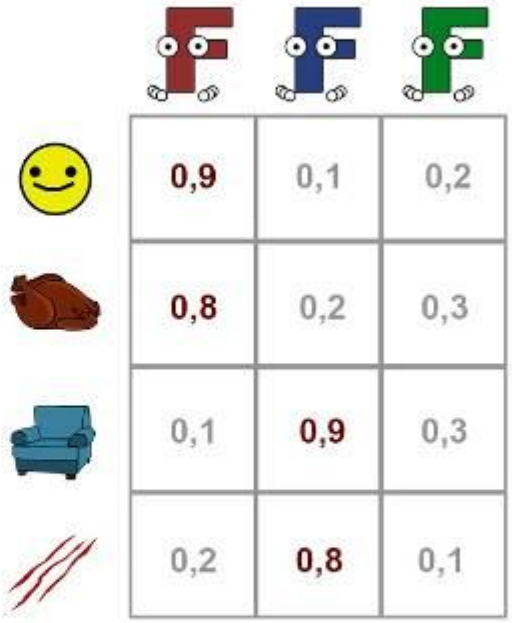

y
ФАКТОР
x


0,6
0,3
0,4
0,7
0,4
0,5
0,3
0,7
0,5
0,4
0,6
0,3
Собственное значение
Σ
2,0
2,0
1,8



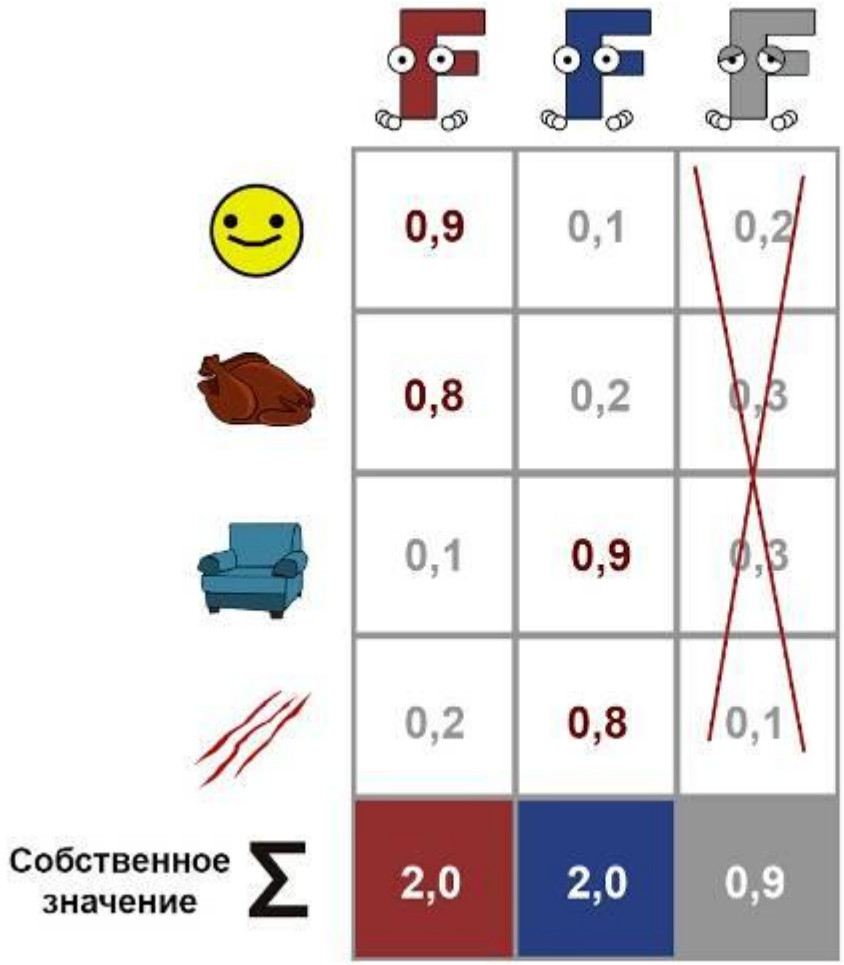
	0,6	0,3	0,4
	0,7	0,4	0,5
	0,3	**0,7**	0,5
	0,4	**0,6**	0,3

До вращения

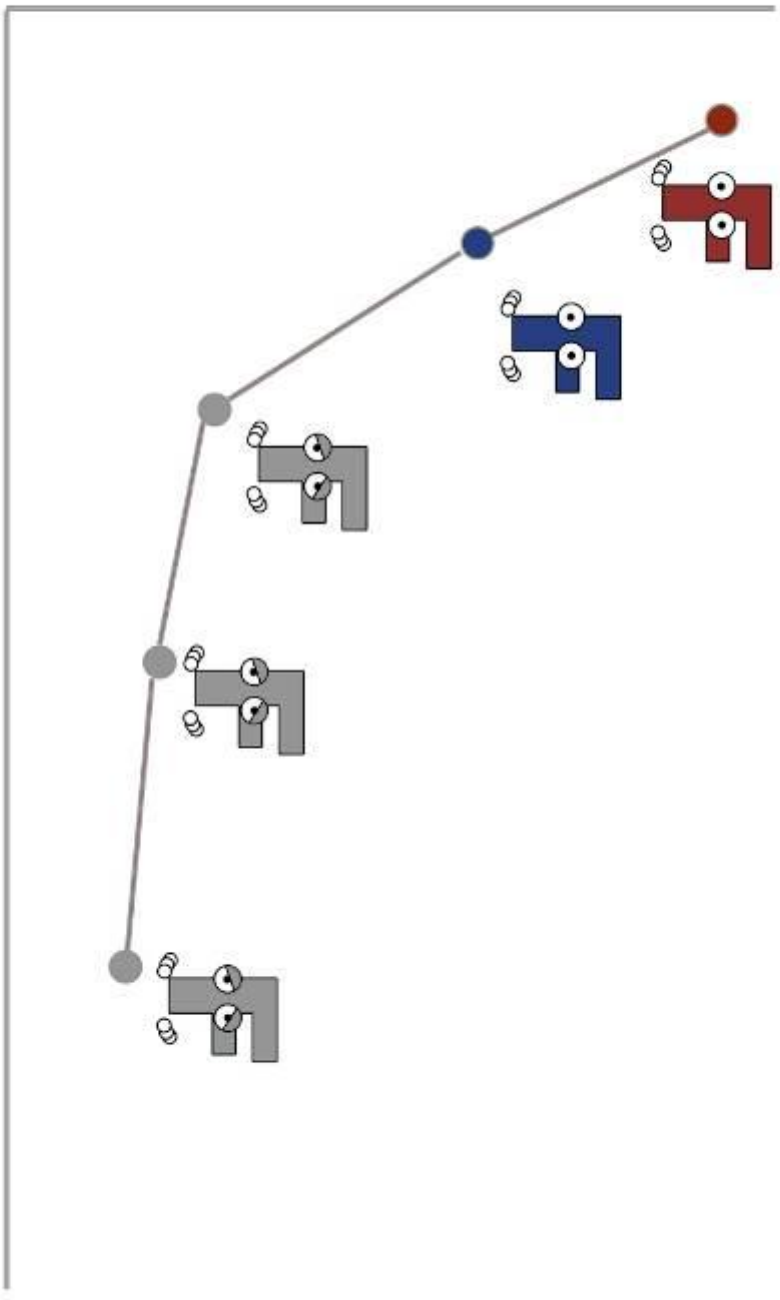


	0,9	0,1	0,2
	0,8	0,2	0,3
	0,1	**0,9**	0,3
	0,2	**0,8**	0,1

После вращения


0,9
0,1
0,2
0,8
0,2
0,3
0,1
0,9
0,3
0,2
0,8
0,1
Собственное значение
Σ
2,0
2,0
0,9

Собственное значение



- В результате измерения мы имеем дело с набором элементарных признаков *Xi*, измеренных по нескольким шкалам. Это – *явные переменные.* Если признаки изменяются согласованно, то можно предположить существование определенных общих причин этой изменчивости, т.е. существование некоторых скрытых (латентных) факторов. Задача анализа – найти эти факторы.
- Так как факторы представляют собой объединение определенных переменных, из этого следует, что эти переменные связаны друг с другом, т.е. обладают корреляцией/ковариацией, причем большей между собой, чем с другими переменными, входящими в другой фактор.
- Методы отыскания факторов и основываются на использовании коэффициентов корреляции (ковариации) между переменными. Факторный анализ дает ***нетривиальное*** решение, т.е. решение нельзя предвидеть, не применяя специальную технику извлечения факторов.

Этапы факторного анализа

А. Формирование цели. Разведочный (эксплораторный) и конфирматорный анализ.

Б. Выбор совокупностей признаков и объектов. Один из самых ответственных этапов, в значительной степени влияющий на результаты. Следует тщательно проанализировать и обосновать выбор совокупности признаков, шкал измерения и представительность множества объектов.

В. Получение исходной факторной структуры. Центроидный метод – алгоритмический подход, главные компоненты – аппроксимационный метод оценки параметров модели, метод максимального правдоподобия – теоретико-вероятностная парадигма.

Г. Вращение факторной структуры.

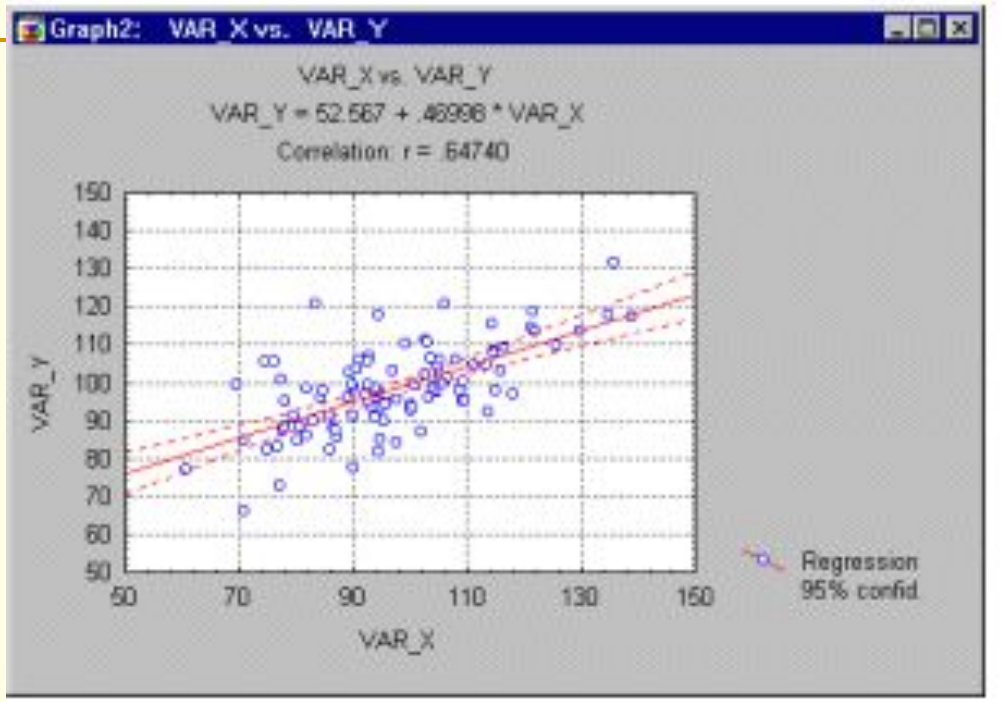
Д. Выявление факторов второго порядка.

Е. Интерпретация и использование решений.

Главные цели факторного анализа:

(1) сокращение числа переменных (редукция данных)
(2) определение структуры взаимосвязей между переменными, т. е. классификация переменных.

Поэтому факторный анализ используется либо как метод сокращения данных, либо как метод классификации



- Если определить новую переменную на основе линии регрессии, то такая переменная будет включить в себя наиболее существенные черты обеих переменных. Фактически, сокращается число переменных. Новый фактор (переменная) - линейная комбинация двух исходных переменных.
- Если пример с двумя переменными распространить на большее число переменных, то вычисления становятся сложнее, однако основной принцип представления двух или более зависимых переменных одним фактором остается в силе.

В этом – суть идеи анализа главных компонент.

Отличие Факторного анализа от Метода главных компонент

- Результатом ФА является модель, в явном виде описывающая зависимость наблюдаемых переменных от скрытых факторов (МГК это описательный анализ данных);
- ФА предусматривает ошибку моделирования (**специфический фактор**) для каждой из наблюдаемых переменных. МГК пытается объяснить **всю** изменчивость, включая шум, зависимостью от главных компонент;
- В **МГК** главные компоненты - ***линейные комбинации наблюдаемых переменных.*** В **ФА** наблюдаемые переменные - ***линейные комбинации общих и специфических факторов***;
- Получаемые в результате ФА факторы могут быть использованы для интерпретации наблюдаемых данных;
- Главные компоненты некоррелированны (ортогональны), факторы - не обязательно;
- МГК можно рассматривать как частный случай ФА, когда все специфические факторы приняты равными нулю, а общие факторы ортогональны.

Главные методы факторного решения:

1) метод максимального правдоподобия;
2) метод наименьших квадратов – метод главных осей;
3) центроидный метод;
4) альфа-факторный анализ (Кайзер, 1965),
5) анализ образов (Гуттман, 1953; Харрис, 1962),
6) метод главных факторов;
7) анализ главных компонент (Хотеллинг, 1933).

Вращение факторов

Процесс поиска оптимальной факторной структуры.

Л. Терстоун считал, что цель исследования заключается в поиске “простой структуры” или попытке объяснить большое число переменных меньшим числом факторов. При поиске простой структуры следует иметь в виду следующее:

- целесообразно стремиться к получению для каждой переменной максимального числа больших факторных нагрузок по одним факторам;
- и одновременно наибольшего количества минимальных факторных нагрузок по другим факторам.

В предельном случае самая простая структура получается тогда, когда все переменные располагаются на соответствующих факторных осях, т.е. имеют ненулевые факторные нагрузки только по одному фактору, а по остальным – нулевые.

Два класса методов вращения

- методы ортогонального вращения, когда при повороте осей координат, угол между факторами остается прямой (и, следовательно, - факторы не связаны между собой
- методы косоугольного (облического) вращения, когда первоначальное ограничение о некоррелированности факторов снимается.

Методы ортогонального вращения:

варимакс, квартимакс, эквимакс.

- Варимакс – наиболее часто используемый на практике метод, цель - минимизировать количество переменных, имеющих высокие нагрузки на данный фактор.
- Квартимакс - в определенном смысле противоположен варимаксу, т.к. минимизирует количество факторов, необходимых для объяснения данной переменной. Поэтому он усиливает интерпретабельность переменных. Квартимакс – вращение приводит к выделению одного из общих факторов с достаточно большими нагрузками на большинство переменных.
- Эквимакс и биквартимакс - два схожих метода, представляющих собой комбинацию варимакса и квартимакса.
- Специальные исследования (Л. Кайзер, 1958) свидетельствуют в пользу преимущественного использования варимакса при прочих равных условиях.

Методы косоугольного вращения:

- позволяют упростить факторное решение за счет введения предположения о коррелированности факторов о возможности существования факторов более высокого порядка, объясняющих наблюдаемую корреляцию.
- Основное преимущество косоугольного вращения - в возможности проверки ортогональности получаемых факторов: если в результате вращения получаются действительно ортогональные факторы, то можно быть уверенным в том, что ортогональность / независимость им действительно свойственна, а не является следствием использования метода ортогонального вращения.
- В статистических пакетах - метод облимин.

Форма представления результатов факторного анализа

Основные результаты факторного анализа выражаются в наборах **факторных нагрузок** и **факторных весов**. Можно оценить действительные значения факторов для отдельных наблюдений - **факторные оценки** (эти значения используются, чтобы провести дальнейший анализ факторов).

- Фактор называется **генеральным** (general), если все его нагрузки значительно отличаются от нуля.
- **Общий** фактор (common) – когда хотя бы две нагрузки значительно отличаются от нуля.
- **Характерный** (unique) фактор – представляет только одну переменную.
- Число высоких нагрузок переменной на общие факторы называется ее **сложностью** (complexity).
- **Пространство общих факторов** – пространство наименьшей размерности, в котором можно представить все переменные в виде векторов. При геометрической интерпретации факторами являются координатные оси, на которые натянуто пространство общих факторов. Эти оси-факторы **нормированы**, т.к. их длина приведена к единице (поскольку дисперсия фактора должна быть равна 1).
- **Полное факторное пространство** натянуто на все факторы – как общие, так и характерные.

Критерии значимости факторов:

- Критерии, основанные на собственных числах (чаще всего – вес больше 1). Кайзер отдает предпочтение этому критерию.
- Критерий, основанный на величине доли воспроизводимой дисперсии (например, 1 или 5 или 10%).
- Критерий отсеивания – Кэттелл (1965) предлагает отсекать те факторы, которые при графическом изображении собственных чисел дают практически горизонтальную линию. Кайзер (1970) отмечает, что это также субъективный критерий.
- Критерий интерпретируемости и инвариантности. Можно применить к одним и тем же данным комбинацию независимых критериев и принимать те результаты, которые подходят ко всем критериям. "Окончательное решение должно базироваться на его приемлемости с точки зрения научных представлений в данной области. Этот подход является"обходным маневром", но, к сожалению, а может быть и к счастью, мы вынуждены принять его, если хотим, чтобы нашими результатами могли воспользоваться другие исследователи"

Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.

Краткие ответы на часто возникающие вопросы

- Какой способ измерений необходим в факторном анализе? – Требуется, чтобы переменные измерялись хотя бы на уровне шкалы интервалов. Требование обусловлено тем, что входной информацией для факторного анализа являются элементы ковариационной матрицы.
- Значит ли, что исследователь должен всегда избегать использования факторного анализа, когда метризуемость пространства переменных не вполне ясна? – Нет, не обязательно. Многие переменные, как меры отношений и мнений, переменные при обработке результатов тестирования, не имеют точно определенной метрической основы. Однако предполагается, что порядковым переменным можно давать числовые значения, не нарушая их внутренних свойств. Во многом это определяется тем, что коэффициенты корреляции обладают свойством робастности по отношению к порядковым искажениям в измеряемых данных.

- Возможно ли использование тау-статистики Кендалла или гамма-статистики Гудмана и Крускала вместо обычных корреляций? – Нет, т.к. нет факторных моделей с порядковыми статистиками. Допускается эвристическое использование таких моделей без статистической интерпретации результатов.
- Можно ли использовать факторный анализ для дихотомических переменных? – Дихотомические переменные нельзя представить в рамках факторной модели. Поэтому никакие соображения, кроме чисто эвристических, не могут обосновать применение факторного анализа к дихотомическим переменным.
- В каких случаях возможно применение факторного анализа к данным, содержащим дихотомические переменные или переменные с конечным множеством значений? – В общем случае, чем шире множество значений, тем точнее результаты. В случае дихотомических переменных использование коэффициента корреляции Пирсона может быть оправдано, если решается задача нахождения кластеров переменных и если корреляции между исходными переменными невелики – не более 0,6 – 0,7.

Критерии значимости и устойчивость факторных решений

- В каких случаях используется метод максимального правдоподобия и связанные с ним критерии значимости, и каков минимальный объем выборки? – Чем больше объем, тем точнее хи-квадрат-аппроксимация. Лоули и Максвелл считают, что этот критерий применим, когда в выборке на 51 наблюдение больше, чем число переменных.
- Сколько переменных должно приходиться на один гипотетический фактор? – Тэрстоун считает, что не менее трех переменных. В целом исследователи сходятся, что переменных должно быть вдвое больше, чем факторов.

Другие статистические вопросы

- Что означает знак факторных нагрузок? – Сам по себе не имеет внутреннего содержания и не несет информации о зависимости между переменной и фактором. Однако надо сопоставлять между собой знаки для различных переменных при одном факторе.
- Что означают собственные значения, связанные с факторами, полученные после вращения? – В первоначальном факторном решении величина собственного значения несет информацию об относительной важности каждого фактора. Для факторного решения после вращения это свойство не сохраняется, т.к. в результате вращения определяются совсем другие факторы, поэтому неважно, какую долю дисперсии воспроизводит каждый из них.
- Можно ли включать в анализ переменные, некоторые из которых являются причинными для других?– В общем случае, переменные не должны быть причинными для других. Предполагается, что все переменные есть функции скрытых факторов. Однако при достаточном опыте можно применять факторный анализ и к причинным системам с более сложной структурой.

Применение факторного анализа в психологии

1. При конструировании тестов. Вопросы, которые имеют высокие нагрузки по каждому фактору, направлены на измерение одного итого же лежащего в их основе психологического конструкта. Вопросы с низкими нагрузками на факторы просто удаляются.
2. При проверке психометрических свойств опросников, особенно когда они используются в новых культурах или популяциях. Если в одной культуре, для которой опросник разработали, выделяется две шкалы, то для другой культуры тоже должно быть обнаружено два фактора. Если же такая факторная структура не обнаружена, то это значит, что опросник не работает в новой ситуации и его не следует использовать.
3. Для измерения значений понятий – так называемый семантический дифференциал, предложенный Е.Осгудом.