Лекция 6

ФАКТОРНЫЙ, КОМПОНЕНТНЫЙ И ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Факторный анализ. Основные принципы

В отличие от кластерного анализа, методы **факторного анализа** применяются, когда неизвестные факторы ищут в форме количественных переменных.

Факторный анализ - это совокупность методов, которые на основе реально существующих связей объектов (признаков) позволяют выявить латентные (неявные) обобщающие характеристики организационной структуры.

Факторный анализ. Основные принципы

При этом предполагается, что наблюдаемые переменные являются линейной комбинацией факторов.

Под *фактором* понимается гипотетическая непосредственно не измеряемая, скрытая (латентная) переменная в той или иной мере связанная с исходными наблюдаемыми переменными.

К факторному анализу в широком смысле относятся:

- метод главных компонент,
- методы многомерного шкалирования, применяемые для формирования факторного пространства по информации о близости объектов,
- методы кластерного анализа, применяемые для описания неколичественных факторов.

Основные цели факторного анализа:

1) **сокращение числа переменных** (редукция данных);

2) **определение структуры взаимосвязей** между переменными (классификация переменных);

3) **косвенные оценки признаков**, неподдающихся непосредственному измерению;

4) **преобразование исходных переменных** к более удобному для интерпретации виду.

Факторный и компонентный анализ – для решения задачи снижения размерности

- **Факторный и компонентный анализ** в большинстве случаев проводятся совместно.
- **Компонентный анализ** является методом определения структурной зависимости между случайными переменными. В результате его использования получается сжатое описание малого объема, несущее почти всю информацию, содержащуюся в исходных данных.

Факторный и компонентный анализ – для решения задачи снижения размерности

- **Факторный анализ** является более общим методом преобразования исходных переменных по сравнению с компонентным анализом.
- ***Факторный анализ предназначен*** для выявления действия различных факторов и их комбинаций на величину результативного признака. При этом сокращается число переменных и определяется структура взаимосвязей между переменными.

Факторный анализ. Особенности

1) *факторный анализ*, в противоположность контролируемому эксперименту, опирается в основном на наблюдения над естественным варьированием переменных;

2) при использовании *факторного анализа* совокупность переменных, изучаемых с точки зрения связей между ними, не выбирается произвольно: сам метод позволит выявить основные факторы, оказывающие существенное влияние в данной области;

.

Факторный анализ. Особенности

3) *факторный анализ* не требует предварительных гипотез, наоборот, он сам может служить методом выдвижения гипотез, а также выступать критерием гипотез, опирающихся на данные, полученные другими методами;

4) *факторный анализ* не требует априорных предположений относительно того, какие переменные независимы, а какие зависимы, метод не преувеличивает причинно-следственные связи и решает вопрос об их мере в процессе дальнейших исследований.

Применение методов факторного анализа включает три этапа:

1) **выделение первоначальных факторов;**

2) **вращение выделенных факторов с целью облегчения их интерпретации в терминах исходных переменных** (в частности, для исключения отрицательных значений);

3) **содержательная интерпретация новых факторов в предметных терминах**, что является творческой задачей исследователя, выходящей за рамки предлагаемого формального метода.

Наиболее часто **факторный анализ** используется для выявления в наблюдаемых признаках $x_1, x_2, \ldots, x_k$ некоторых латентных (скрытых) переменных f_m, называемых *факторами.*

Гипотеза о наличии этих факторов основана на предположении о существовании чего-то общего в наблюдаемых признаках.

Гипотетические факторы обладают следующими свойствами:

1. Они образуют линейно независимый набор переменных, т.е. ни один из факторов (компонент) не выводится как линейная комбинация остальных.

2. Переменные, являющиеся гипотетическими факторами, можно разделить на два основных вида – общие и характерные факторы. Они отличаются структурой весов в линейном уравнении, которое выводит значение наблюдаемой переменной из гипотетических факторов.

Общий фактор имеет несколько переменных с ненулевым весом или факторной нагрузкой, соответствующей этому фактору.

При этом фактор называется *общим*, если хотя бы две его нагрузки значительно отличаются от нуля.

Гипотетические факторы обладают следующими свойствами:

Характерный фактор имеет только одну переменную с ненулевым весом (т.е. только одна переменная от него зависит).

---

3. Всегда предполагается, что общие факторы не коррелируют с характерным фактором, также характерные факторы не коррелированы между собой.

---

4. Обычно предполагается, что число общих факторов меньше, чем число наблюдаемых переменных, однако число характерных факторов принимают равным числу наблюдаемых переменных.

Факторный анализ. Основные принципы

Набор «новых» признаков объясняет большую часть общей изменчивости наблюдаемых данных, а поэтому передают большую часть информации, заключенной в первоначальных наблюдениях.

Особенностью такого преобразования признаков, осуществляемого при помощи процедуры, называемой «вращением факторов» и приводящей к определению «нагрузок» признаков на агрегированные факторы, является то, что оно осуществляется без существенной потери информации.

Метод главных компонент

- Преобразование (вращение) факторов приводит к получению бесконечного множества решений, среди которых нужно выбрать те, которые облегчают интерпретацию вновь полученных факторов.
- Для того, чтобы выразить большое число откликов через малое число факторов, наиболее часто используется **метод главных компонент**.
- Это метод основан на ортогональном проектировании исходного многомерного пространства в пространство меньшей размерности, в котором точки-наблюдения имеют наибольший разброс.

Метод главных компонент

Главные компоненты получаются из исходных переменных путем целенаправленного *вращения*, т.е. как линейные комбинации исходных переменных.

Вращение производится таким образом, чтобы *главные компоненты* были ортогональны и имели максимальную дисперсию среди возможных линейных комбинаций исходных переменных X.

При этом переменные не коррелированы между собой и упорядочены по убыванию дисперсии (первая компонента имеет наибольшую дисперсию).

Общая дисперсия после преобразования остается без изменений.

Метод главных компонент

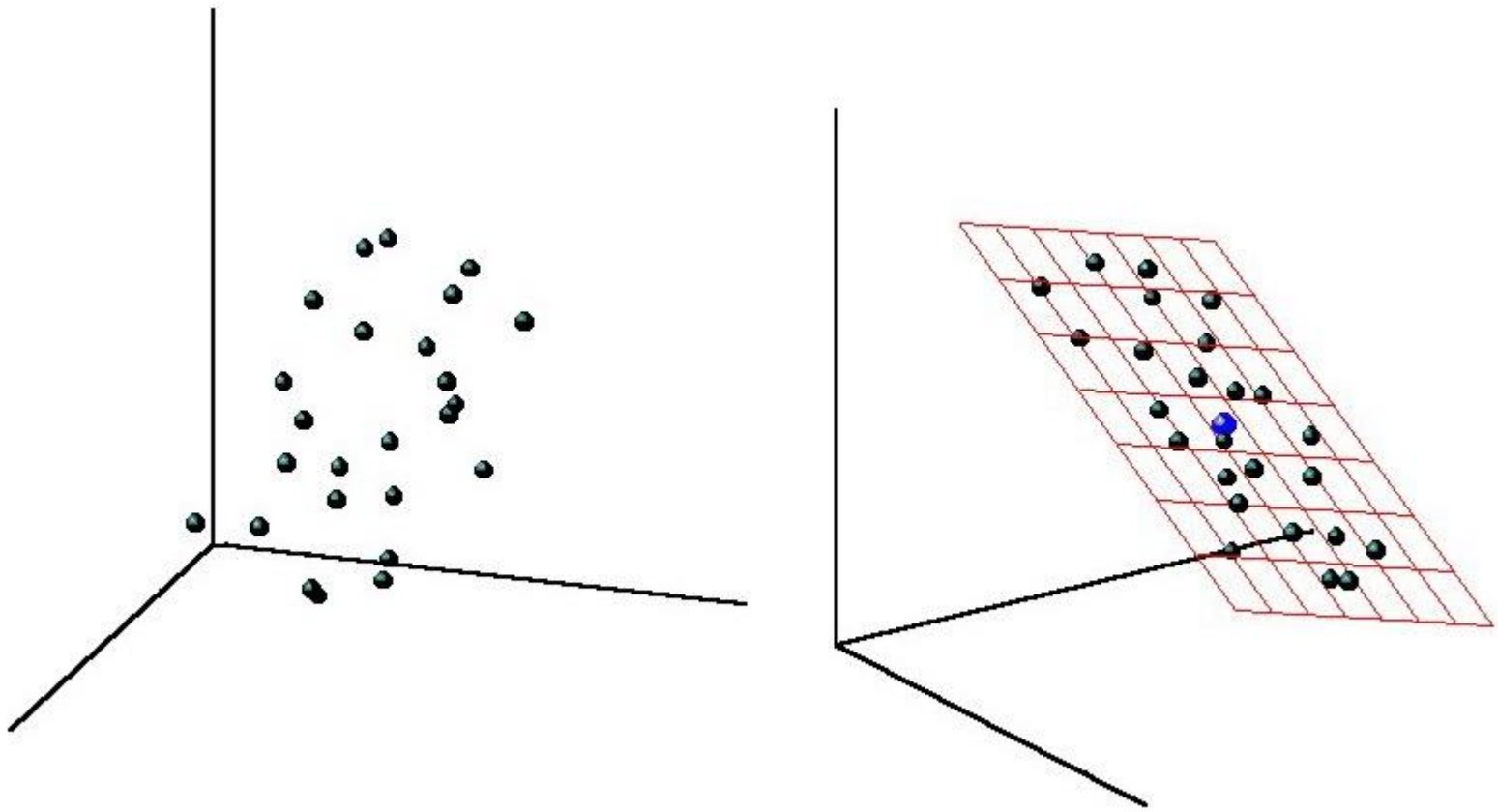

Рис. Графическое представление метода главных компонент

Метод главных компонент

Вспомогательные переменные можно спроектировать на подпространство факторов, чтобы сделать выводы об этих переменных, даже если они не участвовали непосредственно в вычислениях.

То есть, вспомогательные переменные используются только для интерпретации результатов.

Метод главных компонент

- Аналогично наблюдения можно разделить на *вспомогательные* и *активные* наблюдения для анализа.
- Только основные наблюдения будут участвовать в вычислениях главных компонент.
- Вспомогательные наблюдения позже проектируются на векторное подпространство, образованное факторами, которые были вычислены на основе переменных анализа и основных наблюдений.
- Выводы на основе вычисленных факторов применимы и к вспомогательным наблюдениям.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Дискриминантный анализ

Дискриминантный анализ является разделом многомерного статистического анализа, который включает в себя методы классификации многомерных наблюдений по принципу максимального сходства при наличии обучающих признаков.

В кластерном анализе рассматриваются методы многомерной классификации без обучения.

В дискриминантном анализе новые кластеры не образуются, а формулируется правило, по которому объекты подмножества подлежащего классификации относятся к одному из уже существующих (обучающих) подмножеств (классов), на основе сравнения величины дискриминантной функции классифицируемого объекта, рассчитанной по дискриминантным переменным, с некоторой константой дискриминации.

Дискриминантный анализ

Дискриминантный анализ – это общий термин, относящийся к нескольким тесно связанным статистическим процедурам.

Эти процедуры можно разделить на методы ***интерпретации межгрупповых различий*** – ***дискриминации*** и методы ***классификации наблюдений*** по группам.

Дискриминантный анализ

Задачи дискриминантного анализа

Задачи первого типа – задачи дискриминации (пример – в медицинской практике).

Второй тип задачи относится к ситуации, когда признаки принадлежности объекта к той или иной группе потеряны, и их нужно восстановить.

Задачи третьего типа связаны с предсказанием будущих событий на основании имеющихся данных.

Дискриминация

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть ***дискриминантными переменными***), которая бы оптимально разделила рассматриваемые группы.

Дискриминация

Линейная функция

$$d_{km} = \beta_0 + \beta_1 x_{1km} + \ldots + \beta_p x_{pkm}, \quad m = 1, \ldots, n, \; k = 1, \ldots, g$$

называется ***канонической дискриминантной функцией*** с неизвестными коэффициентами β_i

d_{km} – значение дискриминантной функции для ***m***-го объекта в группе ***k***

x_{ikm} – значение дискриминантной переменной X_i для ***m***-го объекта в группе ***k***.

С геометрической точки зрения дискриминантные функции определяют гиперповерхности в ***p***-мерном пространстве.

Дискриминация

Коэффициенты βi первой канонической дискриминантной функции d выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга.

Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой.

Аналогично определяются и другие функции.

Отсюда следует, что ***любая каноническая дискриминантная функция имеет нулевую внутригрупповую корреляцию с*** $d_1, d_2, \ldots, d_{g-1}$

Дискриминация

- Если число групп равно g, то число канонических дискриминантных функций будет на единицу меньше числа групп.
- Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций.
- Тогда графическое изображение объектов будет представлено в одно–, двух– и трехмерных пространствах.

Коэффициенты канонической дискриминантной функции

Для получения коэффициентов β_i канонической дискриминантной функции нужен статистический критерий различения групп.

Классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп.

Большая внутригрупповая вариация нежелательна, так как в этом случае любое заданное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним.

Коэффициенты канонической дискриминантной функции

Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции ***d***, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой

$$\lambda = \mathbf{B}(d)/\mathbf{W}(d) \qquad (2)$$

где **B** - межгрупповая и **W** внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

В некоторых работах вместо **W** используют матрицу рассеяния **T** объединенных данных.

Коэффициенты канонической дискриминантной функции

Рассмотрим ***максимизацию отношения*** (2) ***для произвольного числа классов***.

Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в ***k***-й группе;

n – общее число наблюдений по всем группам;

x_{ikm} – величина переменной ***i*** для ***m***-го наблюдения в ***k***-й группе;

$\bar{x}_{ik}$ – средняя величина переменной ***i*** в ***k***-й группе;

$\bar{x}_i$ – среднее значение переменной ***i*** по всем группам;

$T(u, v)$ – общая сумма перекрестных произведений для переменных ***u*** и ***v***;

$W(u, v)$ – внутригрупповая сумма перекрестных произведений для переменных ***u*** и ***v***.

$t_{ij} – T(x_i, x_j)$; $w_{ij} – W(x_i, x_j)$.

В модели дискриминации должны соблюдаться следующие условия:

1) число групп: $g \geq 2$;
2) число объектов в каждой группе: $n_i \geq 2$;
3) число дискриминантных переменных: $0 < p < (n - 2)$;
4) дискриминантные переменные измеряются в интервальной шкале;
5) дискриминантные переменные линейно независимы;
6) ковариационные матрицы групп примерно равны;
7) дискриминантные переменные в каждой группе подчиняются многомерному нормальному закону распределения.

Коэффициенты канонической дискриминантной функции

Рассмотрим задачу максимизации отношения (2) когда имеются g групп.

Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп.

Для этого вычислим матрицу рассеяния **T**, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних $\bar{x}_i$, $i = 1, \ldots, p$ по каждой переменной.

Элементы матрицы **T** определяются выражением

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n} (x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j)$$