МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное агентство по образованию

ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики, информатики и механики
Кафедра Математических методов исследования операций

Ноздрин Станислав Сергеевич

Отчет по лабораторной работе на тему:

Факторный анализ

3 курс, 7 группа

Преподаватель
И. А. Титова

Воронеж, 2020 г.

Содержание

1. Введение

Факторный анализ является важным инструментом для решения практических задач при разработке рекомендательных систем (и не только), который используется прежде всего для редукции данных и определения структуры взаимосвязи данных.

Факторный анализ переживал несколько волн развития и какое-то время был объявлен "плохой"математикой(как, например, и нейронные сети в первые этапы развития) в первую очередь за высокий уровень субъективности и интуиции при анализе результатов, а также "пренебрежением"к некоторым математическим понятиям(например, к вопросу о высоком или невысоком уровне скоррелированности полученных факторов и исходных признаков)

Задачи, которые решает факторный анализ:

1) Сокращение числа переменных

Факторный анализ позволяет сократить число переменных в n-ое количество раз и ,как мы надеемся, сохранить всю или важную информацию о данных, однако это невсегда возможно.

Этот инструмент позволяет объединить переменные, которые обладают общим свойством и в какой-то степени похожи(обладают высокой корреляцией), в один фактор, измеряющие то же, что и исходные переменные.

2)Измерение "неизмеримого"

Часто мы имеем дело с косвенными признаками, которые объясняют что-то одно, но с разных сторон. Факторный анализ позволяет объединить эти внешние признаки в один и измерить это свойство.

3)Выявление структуры зависимости данных

В данном контексте факторный анализ позволяет описать структуру данных прежде всего через призму корреляционной зависимости между признаками.

Например, если несколько переменных имеют сильную корреляцию, то они зависимы, а если слабую, то нет. Поэтому "зависимые"переменные стоит объединить в один фактор, а переменные, которые не имеют высоких корреляций с другими переменными, стоит оставить, т.к они объясняют что-то своё(содержат уникальную информацию).

Обращу внимание, что здесь и проявляется один из моментов, почему факторный анализ называют "плохой"математикой, т.к. здесь происходит пренебрежение математическими понятиями о корреляции, т.е о наивном предположении, например, что низкий уровень корреляции указывает на отстутствие зависимости между признаками, что совершенно не так(т.к можно отрицать лишь линейную зависимость).

4)Проецирование данных

Факторный анализ позволяет проецировать многомерные данные на двумерную/-трехмерную плоскость. Задача заключается в отыскании лучшей проекции, которая позволит описать данные наилучшим образом.

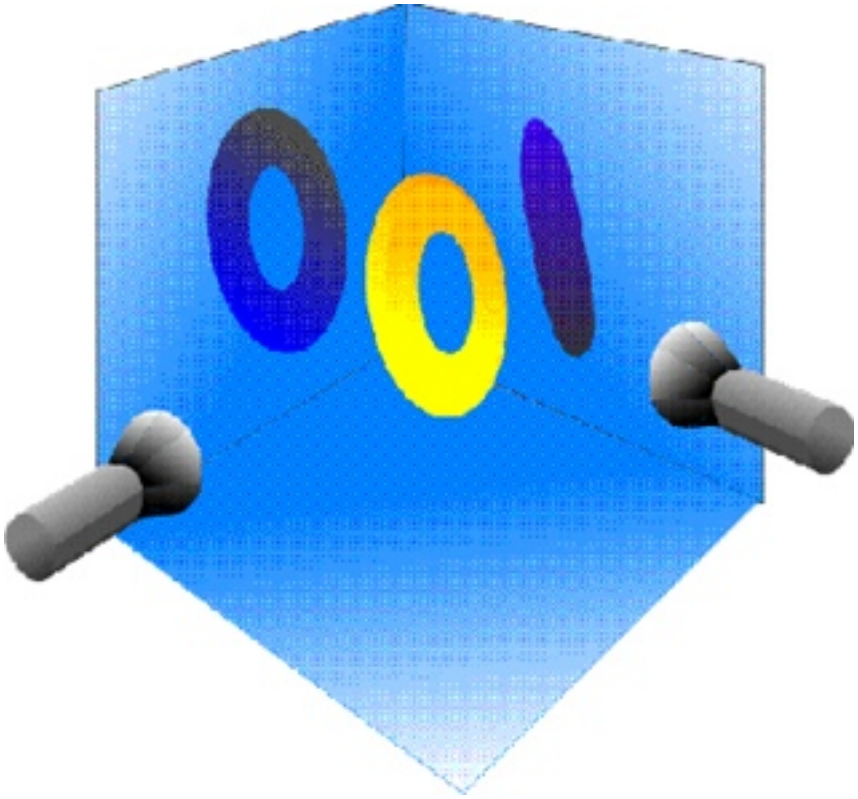
Однако все зависит от того, что мы хотим увидеть:

а)Сократить число переменных и сохранить "расстояния"между исходными примерами

b)Найти такую проекцию, которая лучше разделит несколько "облаков"точек (используется при задачах класификации/кластеризации)

Также это важный пункт для улучшения визуализации данных. Например, существует алгоритм t-SNE(как пакет в языке Python,например), который позволяет сократить размерность данных и визуализировать их.

Пример поиска наилучшей проекции



5)Преодоление мультиколленарности

Часто в случае многомерных задач, когда требуется построить предиктивную модель, мы сталкиваемся с проблемой большого количества скоррелированных данных, что влияет на качество модели. Когда таких признаков очень много, то "отбрасывание"лишних превращается в мучительный процесс. На помощь приходит факторный анализ, который позволяет найти такие факторы, которые будут ортогональны между собой в пространстве и будут "независимы".Правда, существует издержка, в виде потери ясности интерпритации полученных признаков и их влияния на модель.

2. Теоретическая часть

Когда заходит разговор о факторном анализе, существует небольшая путаница.

Дело в том, что факторным анализом называют как классический факторный анализ, так и метод главных компонент.

Метод главных компонент наиболее популярен сегодня, он фактически проводится, когда применяется SVD-разложение, поэтому начнем с него.

2.1. Анализ главных компонент

Итак, пусть у нас есть вектор признаков $X = (X_1, ..., X_k)$, где k - количество признаков.

Задача состоит в том, чтобы найти наиболее информативную проекцию, наибольшое облоко точек, т.е точки с наибольшим разбросом. Как измерить степень разброса? Будем это делать с помощью дисперсии.

Итак, будем искать компоненту(фактор) в виде линейной комбинации признаков с наибольшей дисперсией.

$Y_1 = a_{11}X_1 + a_{12}X_2 + .. + a_{1k}X_k$

$D[Y_1] \to max$

Так как дисперсия зависит от параметров вектора $\vec{a}$, то введем дополнительное условие: $a_i a_j = 1$, где $a_i = (a_{i1}, .., a_{ik})$

Итак, мы построили первый фактор, теперь нужно построить второй таким же образом:

$Y_2 = a_{21}X_1 + a_{22}X_2 + .. + a_{2k}X_k$, но этот фактор должен быть ортогонален в пространстве признаков первому, то есть мы должны потребовать независимость, т.е нескоррелированость этих двух компонент, то есть:

$D[Y_2] \to max$

$corr(Y_1, Y_2) = 0$

Тогда для k-ой компоненты получаем задачу:

$Y_k = a_{k1}X_1 + a_{k2}X_2 + .. + a_{kk}X_k$

$D[Y_k] \to max$

$corr(Y_k, Y_1) = 0$

$corr(Y_k, Y_2) = 0$

...

$corr(Y_k, Y_{k-1}) = 0$

Нахождение параметров **a** сводится к решению задачи:

$Ra = \lambda a$, где R - матрица ковариаций(корреляций) вектора $\mathbf{X}$, а λ - собственное значение

Можно работать с матрицей ковариацией, но тогда для сохранения интерпритации следует стандартизировать данные(т.е перейти к корреляциям)

Решив задачу из линейной алгебры $|R - \lambda I = 0|$, получим k собственных векторов, а также, что $D[Y] = \lambda$

То есть задача поиска компоненты с максимальной дисперсией сводится к поиску максимальному собственному значению

Также сделаем вывод, что $\sum \lambda_i = trace(R)$

Так как R - матрица корреляций, то получаем $\sum \lambda_i = k$

Расположив λ_i в порядке убывания, выберем то количество собственных чисел(факторов), которое нужно в конкретной задаче в зависимости от того, какое количество информации они "покроют".

Существует негласное правило: стоит отбросить все факторы, собственные значения которых меньше 0.8, но оно необязательное.

На практике удобно построить график каменистой осыпи.

2.2. Факторный анализ

Итак, пусть у нас также есть вектор исходных признаков $X = (X_1, ..., X_k)$, а также вектор скрытых переменных $F = (F_1, ..., F_l)$

Тогда скажем, что признак - это есть линейная комбинация факторов, а также ещё **необъясненная часть - U**:

$X_i = a_{i1}F_1 + a_{i2}F_2 + .. + a_{il}F_l + U_i$

Тогда, чем больше $D[U_i]$, тем хуже факторы объясняют исходную переменную.

Сведем задачу к задаче линейной алгебры:

$X = AF + U$, где:

X - матрица исходных признаков

A - матрица

F - матрица факторов

U - необъясненная часть

Также сформируем некоторые предположения:

1) $E[X] = 0$, то есть работать с корреляциями

2) $E[F] = 0$ и $D[F] = 1$,

то есть надеемся, что **матрица корреляций F - единичная матрица**

3) $Corr(U_i, U_j) = 0$ and $Corr(U_i, F_k) = 0$

Итак, найдем матрицу коэффициентов **A** следующим образом:

Пусть матрице признаков **X** соответствует матрица корреляций **R**

А матрице $\tilde{X}$ соотсвутствет матрица корреляций $\tilde{R}$

Где $\tilde{X}$ - матрица объсненной части(без учета U) факторами исходных признаков

Тогда попытаемся найти такую матрицу коэффициентов, чтобы:

$R \approx \tilde{R}$, а конкретно сведем задачу к:

$\sum(r_{ij} - \tilde{r_{ij}}) \to min$

Но тогда мы столкнемся с проблемой неединственности решения. Мы можем получить наилучшее с помощью вращения $varimax$, который позволяет принимать значения коэффициентов a_{ij} более близкими к 0 или 1.

К сожалению, он не везде реализован(например, в R он есть, а в Python - нет).

3. Решение практической задачи

3.1. Знакомство с данными

Конечно, смысл в факторном анализе заключается в том, чтобы резко сократить число переменных(с 300 до 40, например). Но так как задача учебная, то будем работать с данными с небольшим количеством переменных.

Итак, данные представляю из себя опрос 35 участников по поводу их предочтений в выборе напитков(пьет или нет)

Каждый столбец - это вариант напитка, их можно выбирать несколько.

1. **COKE** - Кока-кола
2. **DCOKE** - Диетическая Кока-кола
3. **DPEPSI** - Диетический Пепси
4. **D7UP** - Диетический 7ап
5. **PEPSI** - Пепси
6. **SPRITE** - Спрайт
7. **TAB** - Обычная вода
8. **D7UP** - 7ап

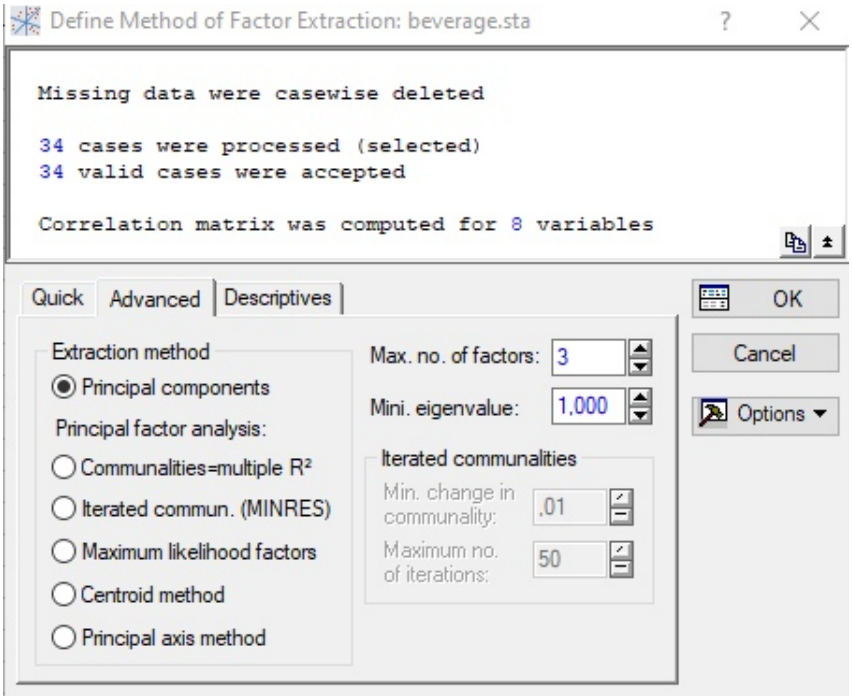
Таблица данных

	1 numb.obs	2 COKE	3 D_COKE	4 D_PEPSI	5 D_7UP	6 PEPSI	7 SPRITE	8 TAB	9 SEVENUP
1	1	1	0	0	0	1	1	0	1
2	2	1	0	0	0	1	0	0	0
3	3	1	0	0	0	1	0	0	0
4	4	0	1	0	1	0	0	1	0
5	5	1	0	0	0	1	0	0	0
6	6	1	0	0	0	1	1	0	0
7	7	0	1	1	1	0	0	1	0
8	8	1	1	0	0	1	1	0	1
9	9	1	1	0	0	0	1	1	1
10	10	1	0	0	0	1	0	0	1
11	11	1	0	0	0	1	1	0	0
12	12	0	1	0	0	0	0	1	0
13	13	0	0	1	1	0	1	0	1
14	14	1	0	0	0	0	1	0	0
15	15	0	1	1	0	0	0	1	0
16	16	0	0	0	0	1	1	0	0
17	17	0	1	0	0	0	1	0	0
18	18	1	1	0	0	1	0	0	0
19	19	1	0	0	0	0	0	0	1
20	20	1	1	1	0	1	0	0	0
21	21	1	0	0	0	1	0	0	0
22	22	1	0	0	0	1	0	0	0
23	23	0	1	0	1	0	0	1	0
24	24	1	1	0	0	1	0	0	0
25	25	0	1	1	1	0	0	0	0
26	26	0	1	0	1	0	0	1	0
27	27	0	1	0	0	0	0	1	0
28	28	1	0	0	0	0	1	0	1

3.2. Решение

В качестве метода выберем аналих главных компонент(Principal Components), а в качестве максимального количества факторов - 3.

Настройка факторного анализа



Посмотрим на корреляции в данных. Отметим, что сильно скоррелированных переменных нет.

Настройка факторного анализа

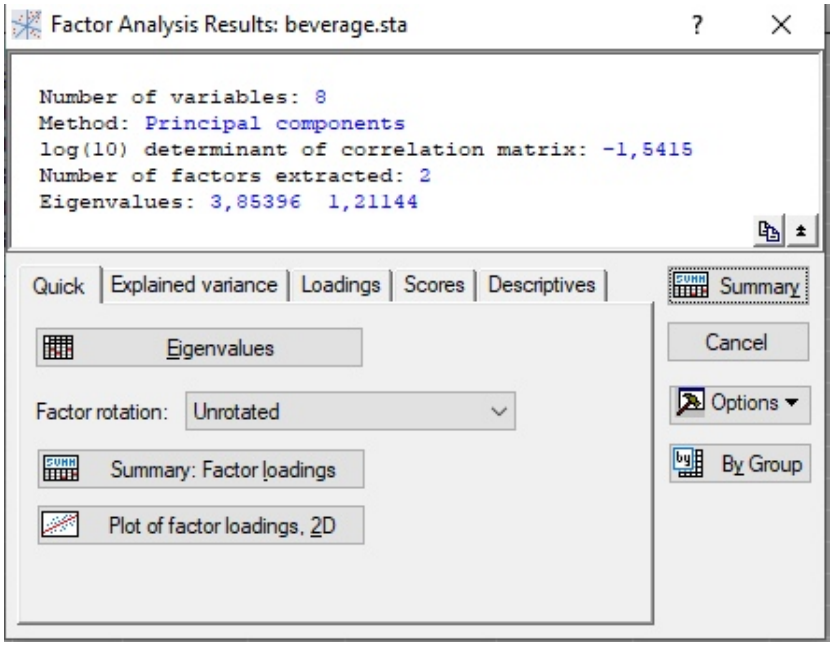
Correlations (beverage.sta)
Casewise deletion of MD
N=34

Variable	COKE	D_COKE	D_PEPSI	D_7UP	PEPSI	SPRITE	TAB	SEVENUP
COKE	1,00	-0,60	-0,52	-0,61	0,67	0,20	-0,70	0,37
D_COKE	-0,60	1,00	0,42	0,36	-0,47	-0,31	0,69	-0,33
D_PEPSI	-0,52	0,42	1,00	0,40	-0,38	-0,24	0,36	-0,18
D_7UP	-0,61	0,36	0,40	1,00	-0,48	-0,20	0,43	-0,14
PEPSI	0,67	-0,47	-0,38	-0,48	1,00	-0,02	-0,65	0,10
SPRITE	0,20	-0,31	-0,24	-0,20	-0,02	1,00	-0,34	0,30
TAB	-0,70	0,69	0,36	0,43	-0,65	-0,34	1,00	-0,27
SEVENUP	0,37	-0,33	-0,18	-0,14	0,10	0,30	-0,27	1,00

Нажимаем **OK** и получаем результат факторного анализа

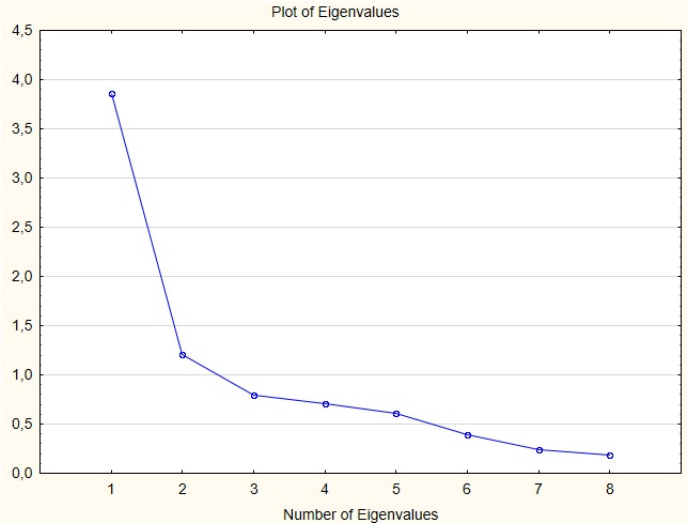
Итак, STATISTICA посчитала оптимальным количеством в виде двух факторов, с соответсвующими значениями λ.

Результат



Посмотрим на график каменистой осыпи с количеством факторов и значениями λ.

График каменистой осыпи для λ



Итак, значение $\lambda_1 = 0.385$, то есть первый фактор объясняет около 40 процентов

информации. $\lambda_2 = 0.12$, а $\lambda_3 = 0.08$, что в принципе указывает на то, что здесь может быть и 3 фактора.

3.3. Интерпретация

И наконец, посмотрим на корреляции факторов с исходными переменными

Таблица корреляций

Variable	Factor Loadings (Varimax normalized) (beverage.sta) Extraction: Principal components (Marked loadings are >,700000) Factor 1	**Factor 2**
COKE	-0,867695	-0,224393
D_COKE	0,667202	0,433561
D_PEPSI	0,599080	0,215425
D_7UP	0,709610	0,060913
PEPSI	-0,868956	0,147996
SPRITE	-0,069478	-0,823256
TAB	0,777387	0,328512
SEVENUP	-0,155058	-0,721528
Expl.Var	3,448781	1,616622
Prp.Totl	0,431098	0,202078

Отмечу, что в настройках был выбран **varimax normalized**, для того, чтобы найти наилучшие коэффициенты со стандартизированными данными.

Итак, **1 фактор** имеет высокую корреляцию с диетическими напитками, а также с употреблением обычной воды и высокую отрицательную корреляцию с кока-колой. Т.е первый фактор отвечает за **любителей здоровых напитков**

2 фактор имеет высокую отрицательную корреляцию с такими напитками как **SEVENUP** и **SPRITE** и низкие со всеми остальными. Т.е второй фактор вмещает себя **нелюбителей SEVENUP и SPRITE, но любителей всего и разного**