

Факторный анализ. Анализ главных компонент

Лекция 12

Что изучает факторный анализ

Факторный анализ — это статистический метод, применяемый для изучения взаимосвязей между переменными.

Предполагается, что наблюдаемые переменные, между которыми существует взаимосвязь, зависят от меньшего числа независимых переменных — **факторов**. При этом наблюдаемые переменные представляются в виде линейных комбинаций факторов и случайной ошибки.

Целью факторного анализа является выявление **факторов**, то есть скрытых независимых переменных.

Факторный анализ широко используется в психометрии, маркетинге, управлении продуктом, исследования операций и финансов.

Виды факторного анализа

Факторный анализ бывает:

1. **разведочным** (exploratory) — он осуществляется при исследовании структуры взаимосвязи между переменными (факторами) без предположения о числе факторов и их нагрузках;
2. **подтверждающим** (confirmatory) — он служит для проверки гипотез о числе факторов и их нагрузках.

В рамках курса рассматривается **только разведочный факторный анализ**.

Факторный анализ и анализ главных компонент

Существует два варианта факторного анализа:

1. метод (анализ) главных компонент (principal component analysis);
2. факторный анализ (factor analysis).

Они не тождественны математически. Появились в разное время и с разными целями, но сейчас по сути решают сходные задачи. Мы будем считать анализ главных компонент одним из вариантов разведочного факторного анализа.

Задачи

1. Сокращение числа переменных.
2. Измерение неизмеримого. Построение новых обобщенных показателей.
3. Наглядное представление многомерных наблюдений (проецирование данных).
4. Описание структуры взаимных связей между переменными, в частности, выявление групп взаимозависимых переменных.
5. Преодоление мультиколлинеарности переменных в регрессионном анализе.
6. **Заполнение пропущенных значений** (разреженные данные в рекомендательных системах).
7. И так далее...

Сокращение числа переменных

Исходные переменные (не все) заменяют на меньшее число новых искусственных переменных.

Новые переменные — факторы. Мы надеемся, что они содержат конденсированную информацию о нашем наборе данных.

Далее работают с факторами, а не с исходными переменными.

Работа портного

При массовом пошиве одежды используются характеристики человеческого тела: размер, полнота и рост. Все это — факторы, то есть искусственные переменные.

Найдены эмпирически, методом проб и ошибок.

Интерпретация их неоднозначна. Например, по одной из формул

$$\text{полнота} = (\text{длина окружности груди} - \text{длина окружности талии})/2$$

Пример Б. Шоу

Зависимости:

- ▶ Носит цилиндр — шире грудная клетка
- ▶ Абонемент на место в церкви — дольше живет
- ▶ Чаше моется — любит оперы Вагнера

Фактор: уровень дохода.

Техника сокращения числа переменных

Линейная зависимость между переменными равносильна тому, что коэффициент корреляции между ними равен единице.

Коррелирующие переменные, хорошо поддающиеся интерпретации

- ▶ Несколько переменных.
- ▶ В каждой паре коэффициент корреляции близок к 1.
- ▶ Это значит, что переменные линейно зависимы.
- ▶ Отбрасываем все переменные, кроме одной.

Нет оснований, чтобы оставить одну переменную

- ▶ Отбрасываем все коррелирующие переменные.
- ▶ Вместо них вводим новую переменную. Она аккумулирует в себе то общее, что измеряют исходные переменные.
- ▶ Такой переменной будет фактор или главная компонента.

Измерение неизмеримого

Как измерить. . .

- ▶ Как измерить любовь?
- ▶ Отношение пациента к своему доктору?
- ▶ Удовлетворенность сортом кофе?
- ▶ Как определить степень депрессии человека?
- ▶ Степень приверженности курению?
- ▶ Лояльность торговой марке?
- ▶ Вероятность разорения фирмы в течение следующего года?

Психология + измерения = психометрия

Исходные переменные отбирались так, чтобы косвенно измерять неизмеряемую величину.

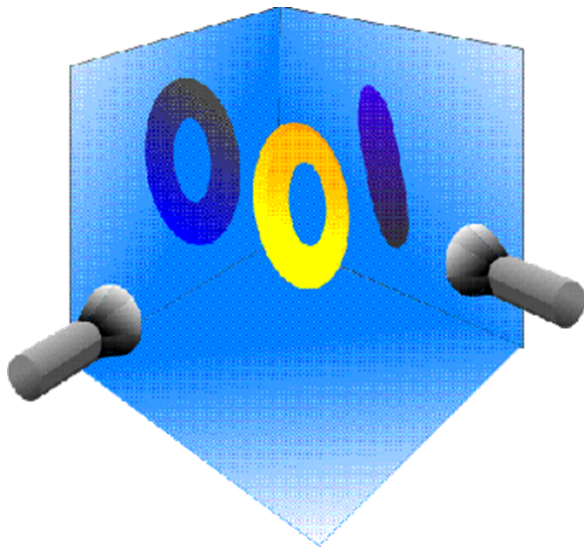
Из них комбинируются новые искусственные переменные — факторы.

Может оказаться, что они измеряют исследуемую характеристику.

Методика семантического дифференциала

Наглядное представление многомерных данных

Проекции хорошие и плохие



Метод главных компонент

Рассмотрим случайный вектор $X_1, X_2, \dots, X_k$.

Задача 1. Найти $Y_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1k}X_k$ такую, что $D(Y_1)$ максимальна.

Дополнительное условие 1: $\vec{a}_1 \vec{a}_1^T = 1$, где $\vec{a}_1 = (a_{11}, a_{12}, \dots, a_{1k})$.

Зачем нам максимальная дисперсия?

Фактически мы вводим новую переменную Y_1 .

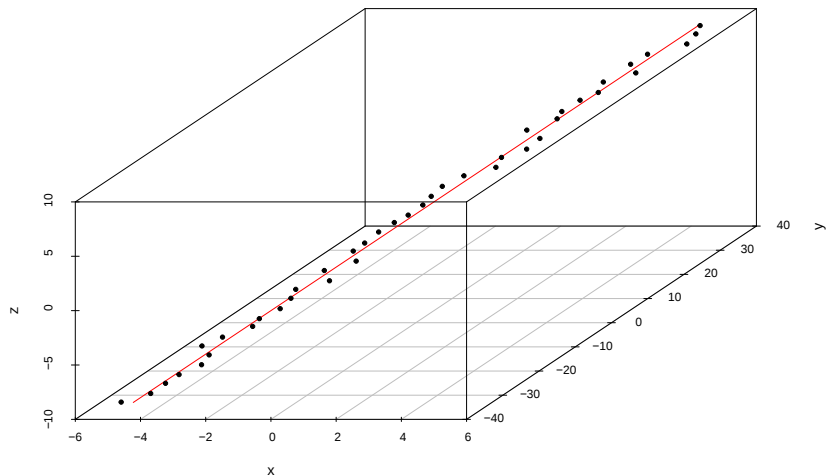
Информативность переменной будем измерять ее дисперсией.

Если дисперсия равна нулю, то никакие изменения X_i не приведут к изменениям новой переменной. Такая переменная просто ничего не выражает.

Таким образом, мы хотим построить наиболее информативную новую переменную.

Y_1 — первая главная компонента, наиболее информативный комплекс переменных.

Новая переменная вместо трех исходных



Продолжаем выделять главные компоненты

Задача 2. Найти $Y_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2k}X_k$ такую, что $D(Y_2)$ максимальна.

Дополнительное условие 2.1: $\vec{a}_2 \vec{a}_2^T = 1$, где $\vec{a}_2 = (a_{21}, a_{22}, \dots, a_{2k})$.

Дополнительное условие 2.2: $\text{corr}(Y_2, Y_1) = 0$

Независимости переменных Y_2 и Y_1 добиться не получится и мы заменяем ее более слабым требованием — отсутствием корреляции.

Продолжаем выделять главные компоненты

Задача 3. Найти $Y_3 = a_{31}X_1 + a_{32}X_2 + \dots + a_{3k}X_k$ такую, что $D(Y_3)$ максимальна.

Дополнительное условие 3.1: $\vec{a}_3 \vec{a}_3^T = 1$, где $\vec{a}_3 = (a_{31}, a_{32}, \dots, a_{3k})$.

Дополнительное условие 3.2: $\text{corr}(Y_3, Y_1) = 0$

Дополнительное условие 3.3: $\text{corr}(Y_3, Y_2) = 0$

Майские ассоциации



Идея: Аббакумов В.Л. Факторный анализ

Главная компонента — шампур, на который нанизаны точки.

Завершаем выделять главные компоненты

Задача k. Найти $Y_k = a_{k1}X_1 + a_{k2}X_2 + \dots + a_{kk}X_k$ такую, что $D(Y_k)$ максимальна.

Дополнительное условие k.1: $\vec{a}_k \vec{a}_k^T = 1$, где $\vec{a}_k = (a_{k1}, a_{k2}, \dots, a_{kk})$.

Дополнительное условие k.2: $\text{corr}(Y_k, Y_1) = 0$

Дополнительное условие k.3: $\text{corr}(Y_k, Y_2) = 0$

...

Дополнительное условие k.k: $\text{corr}(Y_k, Y_{k-1}) = 0$

Критерий для определения числа главных компонент, которые мы хотим оставить

Каждая переменная вносит в общий котел единицу информации (переменные нормализованы, их дисперсии равны 1). Общее число переменных равно k

$$\sum_{i=1}^k D(Y_i) = k.$$

Если мы оставляем главную компоненту с дисперсией 0.5, то эта компонента несет информации меньше, чем было у исходных переменных.

Правило (Кайзер, Kaiser): оставляем главные компоненты с дисперсией > 1 .

Дополнение (Жолиффе, Jolliffe): лучше оставлять те, что с дисперсией > 0.8 — это безопаснее.

Пример: анализ поведения самцов колюшки

Фиксировались различные элементы поведения самца колюшки (stickleback), включая реакцию на помещенную в аквариум модель самца. Эксперимент проводился над 53 разными самцами. Регистрировались следующие элементы поведения.

- ▶ LUNGES — Количество ударов модели самца-соперника.
- ▶ BITES — Количество укусов модели самца-соперника.
- ▶ ZIGZAGS — Плавание зигзагом, которое является частью поведения, направленного на привлечение самок.
- ▶ NEST — Число действий, связанных с построением гнезда.
- ▶ SPINES — Число раз, когда топорщился колючий верхний плавник.
- ▶ DNEST — Суммарная длительность времени, проведенного за построением гнезда.
- ▶ BOUT — Количество элементов поведения, характеризующих готовность к схватке.

Источник: *PCA Example 3: Male stickleback behaviour.*

```

# Читаем данные
stick <- read.csv2("stickle.csv")

# Выполняем анализ главных компонент
# с предварительной стандартизацией данных
stick.pca = prcomp(stick, center = TRUE, scale. = TRUE)

# Смотрим результаты анализа
summary(stick.pca)

```

Importance of components%s:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.5127	1.2059	0.9895	0.9413	0.8679	0.63625	0.48412
Proportion of Variance	0.3269	0.2077	0.1399	0.1266	0.1076	0.05783	0.03348
Cumulative Proportion	0.3269	0.5346	0.6745	0.8011	0.9087	0.96652	1.00000

Результаты

Importance of components%:

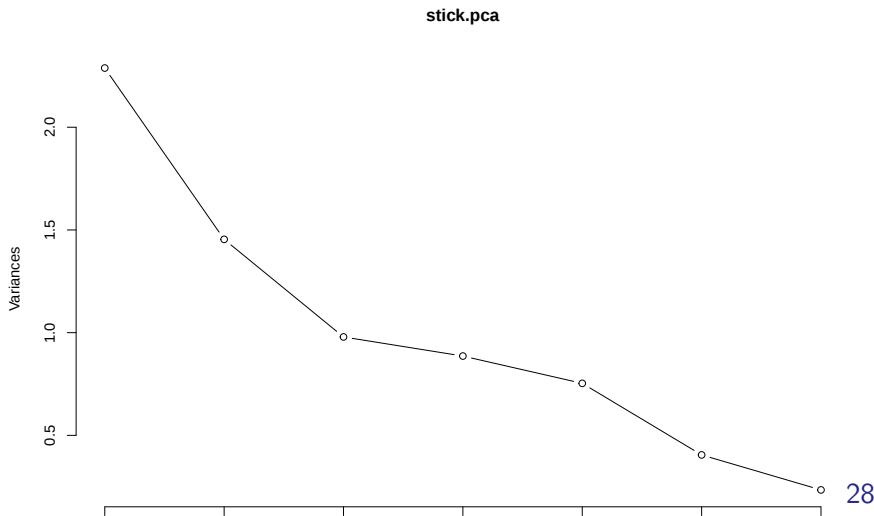
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.5127	1.2059	0.9895	0.9413	0.8679	0.63625	0.48412
Proportion of Variance	0.3269	0.2077	0.1399	0.1266	0.1076	0.05783	0.03348
Cumulative Proportion	0.3269	0.5346	0.6745	0.8011	0.9087	0.96652	1.00000

Cumulative Proportion (кумулятивная доля) позволяет оценить, какую долю дисперсии объясняют первые главные компоненты (ГК):

- ▶ две первые ГК (PC1 и PC2) объясняют 53.5% дисперсии;
- ▶ три первые ГК объясняют уже 67.4% от дисперсии;
- ▶ и т. д.

Сколько нужно ГК? График “каменистая осыпь”

```
# График scree plot помогает выбрать нужное число ГК  
# По оси X - число ГК, по Y - дисперсия  
plot(stick.pca, type = "l")
```



Сколько нужно ГК? Правило Жолиффе

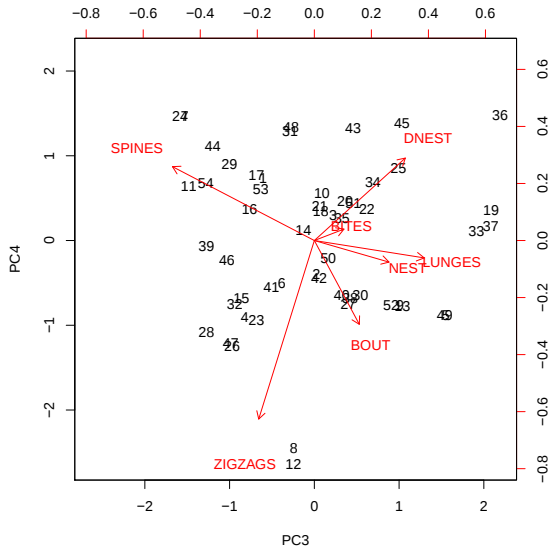
```
stick.pca$sdev^2 > 0.8
```

```
## [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
```

Вывод: будем использовать первые 4 ГК.

biplot() помогает увидеть корреляции (нагрузки)

```
biplot(stick.pca, choices = 3:4, scale = 0)
```



Нагрузки ГК

```
stick.pca$rotation
```

##		PC1	PC2	PC3	PC4	PC5
##	LUNGES	0.4701279	-0.3104661	0.4816326	-0.07619368	-0.18130460
##	BITES	0.4589796	-0.5073716	0.1299271	0.05001379	-0.08378925
##	ZIGZAGS	-0.2481177	-0.3789323	-0.2438164	-0.78255836	-0.02177523
##	NEST	-0.4353556	-0.4002646	0.3273391	-0.09465088	0.11356803
##	SPINES	0.1639491	-0.4652081	-0.6218028	0.32376702	0.41154609
##	DNEST	-0.4217978	-0.2242316	0.3987999	0.36138125	0.39063733
##	BOUT	0.3353027	0.2751091	0.1971586	-0.36730814	0.79042294
##		PC6	PC7			
##	LUNGES	-0.05536321	-0.639436182			
##	BITES	-0.08045943	0.706424046			
##	ZIGZAGS	-0.34605259	-0.052700865			
##	NEST	0.72100117	0.037466458			
##	SPINES	0.11564146	-0.287219657			
##	DNEST	-0.57406155	-0.004980871			
##	BOUT	0.08914314	0.073385142			

Факторный анализ

Как факторы объясняют переменные?

Рассмотрим случайный вектор

$$\mathbf{X}^T = (X_1, X_2, \dots, X_k).$$

Предполагается, что существуют скрытые переменные — **факторы**

$$\mathbf{F}^T = (F_1, F_2, \dots, F_p),$$

от которых зависят измеримые переменные $\mathbf{X}$

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{ik}F_p + U_i, \quad i = 1, \dots, k$$

Причем $p < k$.

Матричная запись

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{ik}F_p + U_i$$

$\Downarrow$

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \mathbf{U}$$

где $A = (a_{ij})$, $i = 1, \dots, k$, $j = 1, \dots, p$, $\mathbf{U}^T = (U_1, U_2, \dots, U_k)$

Факторы $\mathbf{F}$ называют **общими** (common factors).

Элементы матрицы $\mathbf{A}$ называются **факторными нагрузками** (factor loadings).

Элементы вектора $\mathbf{U}$ называются **характерными факторами** (specific variates).

Дополнительные предположения

1.

$$E(\mathbf{X}) = 0$$

Не теряем общности, так как будем рассматривать корреляции.

2.

$$E(\mathbf{F}) = 0, \quad D(\mathbf{F}) = I$$

Хотим, чтобы факторы объясняли разное, то есть были независимы. На деле придется ограничиться некоррелированностью факторов.

3. Остатки нельзя объяснить факторами, т.е. элементы вектора $\mathbf{U}$, не коррелируют друг с другом и с факторами.

Структура дисперсии переменных

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{ik}F_p + U_i, \quad i = 1, \dots, k.$$

Считаем дисперсию переменных

$$D(X_i) = \sum_{j=1}^p a_{ij}^2 + D(U_i).$$

Общность (communality) $h_i^2 = \sum_{j=1}^p a_{ij}^2$ — часть дисперсии переменной, объясненной общими факторами.

Уникальность (specific variance или unique variance) — часть дисперсии переменной, объясненной характерными факторами.

Общность позволяет измерить качество объяснения дисперсии факторами. Например, при $h_i^2 = 0.9$ можно утверждать, что факторами переменная X_i объяснена на 90%.

Отступление: ковариационная и корреляционная матрицы

- ▶ $\Sigma = (\Sigma_{ij})$ — ковариационная матрица;
- ▶ $\mathbf{R} = (r_{ij})$ — корреляционная матрица.

$$\Sigma_{ij} = \text{cov}(X_i, X_j) = \sum (X_i - \bar{X}_i)(X_j - \bar{X}_j),$$

$$r_{ij} = \frac{\text{cov}(X_i, X_j)}{\sqrt{\sum (X_i - \bar{X}_i)^2} \sqrt{\sum (X_j - \bar{X}_j)^2}}.$$

Корреляционная матрица

Пусть

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{ik}F_p + U_i, \quad i = 1, \dots, k.$$

Поскольку

$$\text{cov}(X_i, X_j) = \sum_{m=1}^k a_{im}a_{mj},$$

то имеет место соотношение

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \text{diag}(D(\mathbf{U}))$$

$\mathbf{R}$ — матрица корреляций случайного вектора $\mathbf{X}$.

На практике вместо матрицы матрицы корреляций генеральной совокупности $\mathbf{R}$ мы используем выборочную матрицу корреляций.

Дальше мы работаем, убеждая себя что выборочная матрица корреляций похожа на “истинную” и чем больше у нас наблюдений, тем больше эти матрицы будут похожи.

Неоднозначность решения

Решение

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \text{diag}(D(\mathbf{U}))$$

— матрица $\mathbf{A}$ — определено с точностью до ортогонального преобразования (вращения) $\mathbf{Q}$

$$\mathbf{A} = \mathbf{A}\mathbf{Q}$$

Эти вращения $\mathbf{Q}$ допускают различные интерпретации.

Среди всех решений нужно выбрать самое “симпатичное”. Но какое? В литературе существует масса рецептов.

Слишком много субъективизма при выборе. Из-за этого факторный анализ иногда называют “плохой статистикой”.

Когда применять факторный анализ, а когда анализ главных компонент?

Обсуждение на Cross Validated: *Differences on exploratory factor analysis, confirmatory factor analysis and principal component analysis*

Вывод сообщества:

1. Выполняйте анализ главных компонент, если хотите просто сократить количество наблюдаемых скоррелированных переменных, заменив их меньшим набором независимых составных переменных.
2. Выполняйте факторный анализ, если хотите построить теоретическую модель, в которой наблюдаемые переменные объясняются влиянием скрытых факторов.

Факторный анализ поведения самцов колюшки

##	LUNGES	BITES	ZIGZAGS	NEST
##	Min. : 41.0	Min. : 3.00	Min. : 0.000	Min. : 0.000
##	1st Qu.: 99.0	1st Qu.: 34.25	1st Qu.: 2.250	1st Qu.: 0.000
##	Median :127.0	Median : 60.00	Median : 5.000	Median : 0.000
##	Mean :131.2	Mean : 69.06	Mean : 7.278	Mean : 1.426
##	3rd Qu.:156.8	3rd Qu.: 84.25	3rd Qu.:10.750	3rd Qu.: 1.000
##	Max. :277.0	Max. :175.00	Max. :26.000	Max. :15.000
##	SPINES	DNEST	BOUT	
##	Min. : 5.00	Min. : 0.00	Min. : 6.00	
##	1st Qu.:14.00	1st Qu.: 0.00	1st Qu.: 23.50	
##	Median :16.00	Median : 0.00	Median : 45.00	
##	Mean :18.43	Mean : 34.26	Mean : 89.31	
##	3rd Qu.:24.75	3rd Qu.: 73.00	3rd Qu.:125.00	
##	Max. :46.00	Max. :199.00	Max. :306.00	

Выводы из первичного осмотра

- ▶ Больше половины самцов вообще не строит гнездо.
- ▶ Средние значения и дисперсии у разных переменных существенно различаются.
- ▶ Среди минимальных значений три нуля, дважды нули составляют больше половины всех наблюдений.

Структура корреляций между переменными

```
round(cor(x), 2)
```

##	LUNGES	BITES	ZIGZAGS	NEST	SPINES	DNEST	BOUT
## LUNGES	1.00	0.69	-0.14	-0.16	0.06	-0.23	0.23
## BITES	0.69	1.00	-0.04	-0.15	0.37	-0.22	0.12
## ZIGZAGS	-0.14	-0.04	1.00	0.35	0.07	0.09	-0.16
## NEST	-0.16	-0.15	0.35	1.00	-0.05	0.51	-0.31
## SPINES	0.06	0.37	0.07	-0.05	1.00	-0.05	-0.04
## DNEST	-0.23	-0.22	0.09	0.51	-0.05	1.00	-0.24
## BOUT	0.23	0.12	-0.16	-0.31	-0.04	-0.24	1.00

Выводы: нет больших значений корреляции, значит нет существенного дублирования информации.

Факторный анализ нацелен на удаление дублирующих переменных, а здесь этого нет.

Попытка выполнить факторный анализ с 5 факторами

```
stick.fa = factanal(x, factors=5, method = "mls",  
                    scores = "regression")
```

```
Error in factanal(x, factors = 5, method = "mls",  
                  scores = "regression") :  
  5 факторов слишком много для 7 переменных
```

Факторный анализ с 2 факторами

```
stick.fa2 = factanal(x, factors=2, scores = "regression")
```

Посмотрим на результаты

Call:

```
factanal(x = x, factors = 2, scores = "regression")
```

Uniquenesses:

LUNGES	BITES	ZIGZAGS	NEST	SPINES	DNEST	BOUT
0.517	0.005	0.861	0.165	0.860	0.673	0.875

Loadings:

	Factor1	Factor2
LUNGES	0.676	-0.164
BITES	0.992	-0.108
ZIGZAGS		0.373
NEST		0.912
SPINES	0.374	
DNEST	-0.160	0.549
BOUT		-0.344

	Factor1	Factor2
SS loadings	1.614	1.430
Proportion Var	0.231	0.204
Cumulative Var	0.231	0.435

Test of the hypothesis that 2 factors are sufficient.

The chi square statistic is 8.93 on 8 degrees of freedom.

The p-value is 0.349

1. Уникальности (Uniquenesses)

Большая уникальность означает, что не объясненная факторами часть данной переменной слишком велика.

Uniquenesses:

LUNGES	BITES	ZIGZAGS	NEST	SPINES	DNEST	BOUT
0.517	0.005	0.861	0.165	0.860	0.673	0.875

ZIGZAGS, SPINES, DNEST и BOUT объяснены плохо, а BITES и NEST — хорошо.

Переменная может оказаться “сама по себе” и не дублироваться никакими другими переменными.

Рекомендация: если переменная плохо объяснена факторами, такую переменную лучше не вводить в факторный анализ. Она характеризует что-то свое, уникальное.

Итак, в строке Uniquenesses мы ищем большие дисперсии — необъясненные переменные. Устраняем их из анализа.

2. Нагрузки (Loadings)

Loadings — это матрица A , она же матрица корреляций. Наша цель: понять, что измеряет фактор, дать факторам интерпретацию.

Большая по модулю корреляция переменной и фактора в Loadings означает, что фактор объясняет большую часть той переменной с которой у него большая корреляция.

Что измеряет Factor1? Factor2?

R оставляет пустые места на месте близких к нулю корреляций. Делается по умолчанию, для наглядности. С этой же целью результат округлен до трех знаков после запятой.

```
##
## Loadings:
##           Factor1 Factor2
## LUNGES    0.676  -0.164
## BITES     0.992  -0.108
## ZIGZAGS           0.373
## NEST           0.912
## SPINES    0.374
## DNEST     -0.160   0.549
## BOUT              -0.344
##
##           Factor1 Factor2
## SS loadings      1.614   1.430
## Proportion Var   0.231   0.204
## Cumulative Var   0.231   0.435
```

Интерпретация факторов

- ▶ Factor1 объясняет агрессивное поведение рыбок (LUNGES, BITES, SPINES).
- ▶ Factor2 имеет значительные положительные корреляции с действиями, направленными на привлечение самок и гнездование и отрицательные корреляции с агрессивными действиями. Вывод: Factor2 измеряет “домовитость”.

Значения факторов: stick.fa2\$scores

[1,]	-1.07897741	-0.54533088	[28,]	0.12209717	-0.35401111
[2,]	-0.31089890	-0.52665030	[29,]	-0.23769866	-0.48525593
[3,]	-0.92864935	-0.24781929	[30,]	0.23573850	-0.60725038
[4,]	1.58038917	-0.25673511	[31,]	0.44076580	0.68280954
[5,]	0.16828696	3.90608312	[32,]	-0.67526086	1.26561999
[6,]	0.20506010	-0.13222161	[33,]	-0.23498106	-0.11664566
[7,]	2.41000144	-0.33990326	[34,]	-1.02175866	-0.09401053
[8,]	-0.09451796	0.02829634	[35,]	0.71070606	-0.46294377
[9,]	0.99916681	-0.57061718	[36,]	-0.24792737	0.73718584
[10,]	-0.77820621	0.35119307	[37,]	1.90972845	-0.11937353
[11,]	-0.45668318	-0.14449738	[38,]	-0.03288298	-0.50967439
[12,]	-0.04842546	0.02615393	[39,]	-0.04083386	-0.38522547
[13,]	0.99932821	-0.57271005	[40,]	0.03618076	-0.50503145
[14,]	-1.06724282	-0.51537565	[41,]	0.25264276	-0.11000196
[15,]	-0.65004922	1.29001096	[42,]	-0.38115852	-0.53194470
[16,]	-0.84168380	-0.48577959	[43,]	-1.47954210	0.28327741
[17,]	-1.26351961	-0.55659996	[44,]	-0.28477514	-0.49045267
[18,]	-0.92858720	-0.25600495	[45,]	-0.80417043	0.94418658
[19,]	2.02562646	-0.11154315	[46,]	-0.08901428	-0.39068073
[20,]	1.44230040	-0.47130027	[47,]	-1.40732651	-0.61057866
[21,]	-0.84701241	0.34692992	[48,]	0.64960697	0.70571731
[22,]	-0.40268422	-0.50755445	[49,]	0.07595694	3.90612192
[23,]	1.58124121	-0.26553933	[50,]	0.18050147	1.04498153
[24,]	2.27086931	-0.34683843	[51,]	0.84614532	-0.47699773
[25,]	1.46299625	-0.06137600	[52,]	-0.33973851	-0.67018364
[26,]	-1.54676783	-0.61804087	[53,]	-1.59181069	-0.55347328
[27,]	0.05161044	-0.61034743	[54,]	-0.54416178	0.09795328

Работа с результатами интерпретации

```
# Кто самый драчливый?  
F1 <- stick.fa2$scores[,1]  
which.max(F1)
```

```
## [1] 7
```

```
max(F1)
```

```
## [1] 2.410001
```

```
# Кто самый домовитый?  
F2 <- stick.fa2$scores[,2]  
which.max(F2)
```

```
## [1] 49
```

```
max(F2)
```

```
## [1] 3.906122
```

Этапы работы с нагрузками

1. Смотрим большие корреляции.
2. Смотрим, что общего у тех переменных, с которыми у фактора большая корреляция.
3. Делаем выводы.

Что можно сделать еще

Попробовать выполнить анализ с тремя факторами.

Еще об интерпретации факторов

Из-за попыток дать интерпретацию там, где она неочевидна пошла плохая репутация факторного анализа.

Иногда нужны просто факторы, без интерпретации — для сокращения размерности.

Пример

Нужно построить регрессионную модель, а данные коллинеарны. Строим факторы. Они независимы друг от друга и никакой коллинеарности между ними быть не может. Затем строим регрессионную модель, но уже по факторам.

Конечно, лучше бы разобраться в сути переменных, но какой-то результат все же лучше, чем никакого. . .

Разберем список stick.fa2 подробнее

```
# Вопрос: итеративный алгоритм вычисления  
# факторных нагрузок (loadings) сошелся или нет?  
stick.fa2$converged
```

```
## [1] TRUE
```

```
# Вопрос: чему равны нагрузки (Loadings)?  
# Вспомним, что они - корреляции факторов и переменных  
# и по ним будем интерпретировать факторы  
stick.fa2$loadings
```

```
##  
## Loadings:  
##          Factor1 Factor2  
## LUNGES    0.676  -0.164  
## BITES     0.992  -0.108  
## ZIGZAGS           0.373  
## NEST           0.912  
## SPINES    0.374  
## DNEST     -0.160   0.549  
## BOUT              -0.344  
##  
##          Factor1 Factor2  
## SS loadings    1.614   1.430  
## Proportion Var  0.231   0.204  
## Cumulative Var  0.231   0.435
```

```
# Вопрос: чему равны дисперсии не объясненных  
# компонент переменных?  
# Та часть переменной, которая не объясняется  
# общими факторами (common factors), входит в  
# характерный фактор (unique factor).  
# uniquenesses - дисперсии этих факторов  
stick.fa2$uniquenesses
```

LUNGES	BITES	ZIGZAGS	NEST	SPINES	DNEST	BOUT
0.5165716	0.0050000	0.8607698	0.1648021	0.8602388	0.6732179	0.8747123

Вывод

- ▶ Переменные BITES и NEST хорошо объясняются факторами,
- ▶ Переменные SPINES и BOUT объяснены слабо.

```
# Вопрос: как выглядит матрица корреляций исходных переменных
# По существу, с помощью факторов приближаем именно ее!
stick.fa2$correlation
```

```
##          LUNGES          BITES          ZIGZAGS          NEST          SPINES
## LUNGES    1.00000000    0.68807674   -0.13933637   -0.16413256    0.05556711
## BITES     0.68807674    1.00000000   -0.04276105   -0.14888398    0.37338178
## ZIGZAGS   -0.13933637   -0.04276105    1.00000000    0.35189426    0.06781281
## NEST      -0.16413256   -0.14888398    0.35189426    1.00000000   -0.05253774
## SPINES     0.05556711    0.37338178    0.06781281   -0.05253774    1.00000000
## DNEST     -0.22856752   -0.21756303    0.09131506    0.51402827   -0.05109914
## BOUT       0.23331420    0.11731577   -0.16069256   -0.30587584   -0.04148252
##          DNEST          BOUT
## LUNGES   -0.22856752    0.23331420
## BITES    -0.21756303    0.11731577
## ZIGZAGS   0.09131506   -0.16069256
## NEST       0.51402827   -0.30587584
## SPINES    -0.05109914   -0.04148252
## DNEST      1.00000000   -0.24217747
## BOUT      -0.24217747    1.00000000
```

Наблюдение: много больших корреляций, говорящих о взаимосвязи переменных.
Можно надеяться на успех применения факторного анализа.

```
# Значение критерия оптимальности в конце  
# процесса минимизации, число подсчетов функции  
# (критерия) и ее градиента  
stick.fa2$criteria
```

```
##          objective counts.function counts.gradient  
##          0.1840413          16.0000000          16.0000000
```

```
# Вопрос: сколько факторов было заказано?  
stick.fa2$factores
```

```
## [1] 2
```

```
# Вопрос: сколько степеней свободы?  
stick.fa2$dof
```

```
## [1] 8
```

```
# Вопрос: каким методом оцениваются факторные нагрузки?  
stick.fa2$method
```

```
## [1] "mle"
```

Заметим, что на сегодня других методов в пакете R нет.

```
# Матрица вращения факторов  
stick.fa2$rotmat
```

```
##           [,1]      [,2]  
## [1,] 0.9934798 -0.1140087  
## [2,] 0.1140087  0.9934798
```

```
# Вопрос: какие значения принимают факторы в  
# каждом конкретном случае  
stick.fa2$scores
```

```
# Значения статистики теста значимости и  
# его P-значение, если их можно вычислить.  
stick.fa2$STATISTIC
```

```
## objective  
## 8.926005
```

```
stick.fa2$PVAL
```

```
## objective  
## 0.3485766
```

В нашем случае со значимостью плохо. С другой стороны и применение теста к распределениям, далеким от нормального, сомнительно.

```
# Вопрос: сколько было наблюдений?  
stick.fa2$n.obs
```

```
## [1] 54
```

```
# Вопрос: какая команда выполнялась?  
stick.fa2$call
```

```
## factanal(x = x, factors = 2, scores = "regression")
```

Рождаемость и социоэкономические
показатели в Швейцарии в 1888 году

Описание данных swiss

Рассматриваются стандартизированные индикаторы рождаемости и показатели социального и экономического развития в 47 франко-говорящих провинциях Швейцарии примерно в 1888 году.

Данные содержат 47 наблюдений и 6 переменных. Все переменные измерены в процентах, то есть лежат в интервале [0, 100].

- ▶ Fertility — *Ig*, показатель рождаемости, описан как “обычный стандартизированный показатель рождаемости”.
- ▶ Agriculture — процент мужчин в провинции, работающих в сельском хозяйстве.
- ▶ Examination — процент призывников провинции, получивших высшие оценки на экзамене при поступлении в армию.
- ▶ Education — процент призывников провинции, чье образование превышает уровень начальной школы (primary school).
- ▶ Catholic — процент католиков (остальные протестанты).
- ▶ Infant_Mortality — детская смертность, процент проживших меньше одного года среди родившихся.

```
# Подключаем набор данных
data(swiss)
# Обычно с матрицы корреляций начинают анализ
cor(swiss)
```

```
##          Fertility Agriculture Examination  Education  Catholic
## Fertility      1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture    0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination   -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education     -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic       0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality 0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##
## Infant.Mortality
## Fertility      0.41655603
## Agriculture    -0.06085861
## Examination    -0.11402160
## Education      -0.09932185
## Catholic       0.17549591
## Infant.Mortality 1.00000000
```

Вывод: почти все корреляции большие, переменные взаимозависимы, можно надеяться на успех применения факторного анализа.

Факторный анализ с 2 факторами

```
swiss.fa2 <- factanal(swiss, factors=2, scores = "regression")
swiss.fa2$uniquenesses
```

##	Fertility	Agriculture	Examination	Education
##	0.41966825	0.49169873	0.27033067	0.00500000
##	Catholic	Infant.Mortality		
##	0.06071104	0.96046873		

Выводы:

- ▶ переменные Education и Catholic хорошо объясняются факторами;
- ▶ переменная Infant.Mortality практически не объяснена.

Можно попробовать исключить Infant.Mortality из числа факторов.

Что собой представляют факторы?

```
swiss.fa2$loadings
```

```
##  
## Loadings:  
##           Factor1 Factor2  
## Fertility      -0.652   0.393  
## Agriculture    -0.631   0.333  
## Examination     0.685  -0.510  
## Education       0.997  
## Catholic        -0.124   0.961  
## Infant.Mortality      0.175  
##  
##           Factor1 Factor2  
## SS loadings      2.311   1.481  
## Proportion Var    0.385   0.247  
## Cumulative Var    0.385   0.632
```

Интерпретация факторов

Второй фактор измеряет **доминирующую религию**. Его корреляция с переменной Catholic близка к единице.

Первый фактор измеряет, **насколько район продвинулся на пути от сельскохозяйственного общества к индустриальному**.

Действительно, фактор имеет отрицательную корреляцию с переменными Fertility и Agriculture. Высокие значения этих переменных характерны для сельскохозяйственного общества.

В то же время, фактор имеет положительную корреляцию с переменными Examination и Education. Высокие значения этих переменных характерны для индустриального общества

Т.е. присутствует разделение на протестанские индустриальные и католические сельскохозяйственные провинции.

Для каждого кантона можно получить число, характеризующее степень продвинутости данного кантона от сельского хозяйства к промышленному производству.