Липецкий государственный технический университет

Кафедра прикладной математики

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ

Лекция 6

6. Методы снижения размерности моделей и обнаружение в моделях скрытой структуры

Составитель - Сысоев А.С., к.т.н., доц.

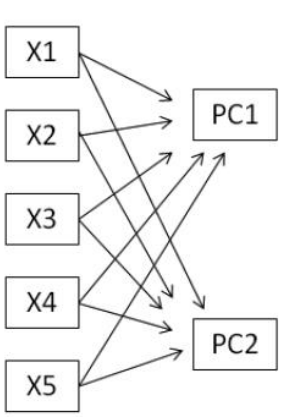
Липецк – 2021

Outline

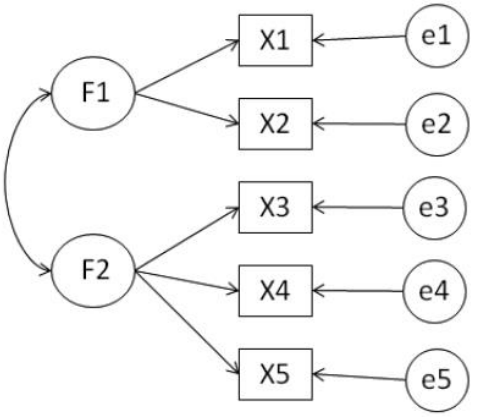
6. Методы снижения размерности и обнаружения скрытой структуры

Анализ главных компонент (principal components analysis, PCA) – это способ снижения размерности данных, который преобразует большое число скоррелированных переменных в гораздо меньший набор нескоррелированных переменных, называемых *главными компонентами*.

Факторный анализ (exploratory factor analysis, EFA) - методы, которые обнаруживают скрытую структуру в имеющемся наборе переменных. Этот анализ позволяет найти меньший набор лежащих в основе, или латентных структурных компонентов, которые могут объяснить взаимосвязи между наблюдаемыми, или явными, переменными.



(a) Модель анализа главных компонент



(b) Модель факторного анализа

6.1. Анализ главных компонент и факторный анализ в R

Пакет `psych`

Функции	Описание
`principal()`	Анализ главных компонент с возможностью поворота осей*
`fa()`	Факторный анализ методом главных осей, минимальных остатков, взвешенных наименьших квадратов или наибольшего правдоподобия
`fa.parallel()`	График собственных значений (scree plot) с параллельным анализом
`factor.plot()`	Графическое изображение результатов факторного анализа или анализа главных компонент
`fa.diagram()`	Графическое изображения матриц нагрузок факторного анализа или анализа главных компонент
`scree()`	График собственных значений для факторного анализа и анализа главных компонент

ШАГИ

1. *Подготовить данные*. Результаты и PCA, и EFA выводятся из корреляций между наблюдаемыми переменными. Можно использовать в качестве аргументов функций principal() и fa() либо исходную таблицу данных, либо корреляционную матрицу.

2. *Выбрать факторную модель*. Нужно решить, что лучше подходит для исследовательских задач – PCA (снижение размерности данных) или EFA (обнаружение скрытой структуры).

3. *Решить, сколько компонент/факторов выделять.*

4. *Выделить компоненты/факторы.*

5. *Повернуть компоненты/факторы.*

6. *Интерпретировать результаты.*

7. *Вычислить значения компонент или факторов.*

6.2. Главные компоненты

Главные компоненты – это линейные комбинации наблюдаемых переменных.

Первая главная компонента $PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k$ - это взвешенная комбинация k наблюдаемых переменных, которая учитывает наибольшую дисперсию (долю изменчивости) исходного набора переменных.

Вторая главная компонента – это линейная комбинация, которая учитывает наибольшую дисперсию исходного набора переменных при условии, что она ортогональна первой главной компоненте (то есть не коррелирует с ней).

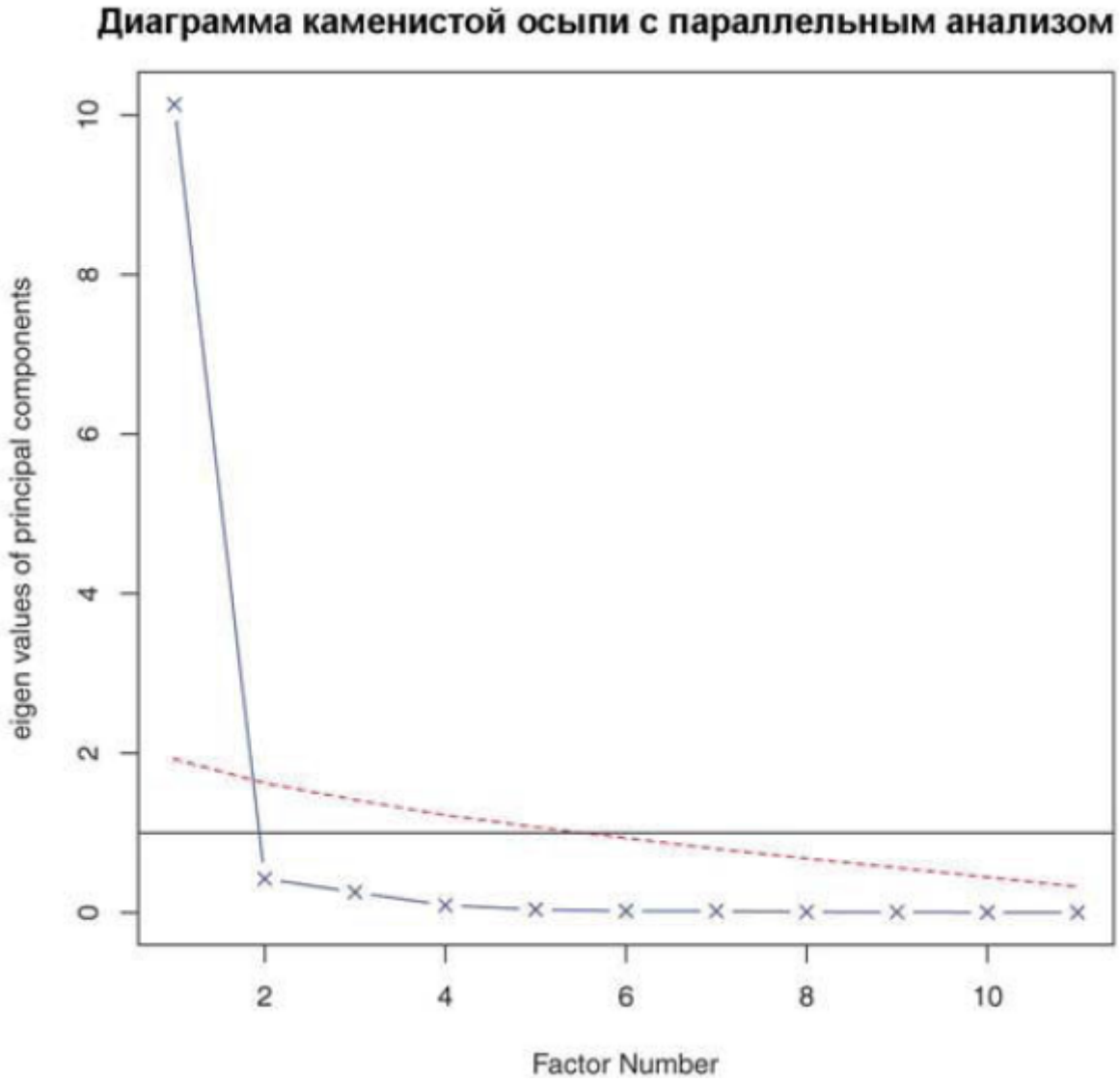
ВЫБОР НЕОБХОДИМОГО ЧИСЛА КОМПОНЕНТ

Существует несколько критериев для определения числа компонент в PCA:

- имеющийся опыт и теоретические соображения;
- объяснение заданной доли дисперсии исходных переменных (например, 80%);
- изучение собственных значений матрицы корреляций между всеми переменными.

Каждая компонента связана с собственным значением корреляционной матрицы. Первой главной компоненте (principal component, PC) соответствует наибольшее собственное значение, второй компоненте – второе по величине собственное значение и т. д. Согласно *критерию Кайзера-Харриса* (Kaiser-Harris), следует использовать компоненты, у которых собственные значения превышают единицу.

6.2. Главные компоненты



6.2. Главные компоненты

ВЫДЕЛЕНИЕ ГЛАВНЫХ КОМПОНЕНТ

`principal(r, nfactors=, rotate=, scores=),`

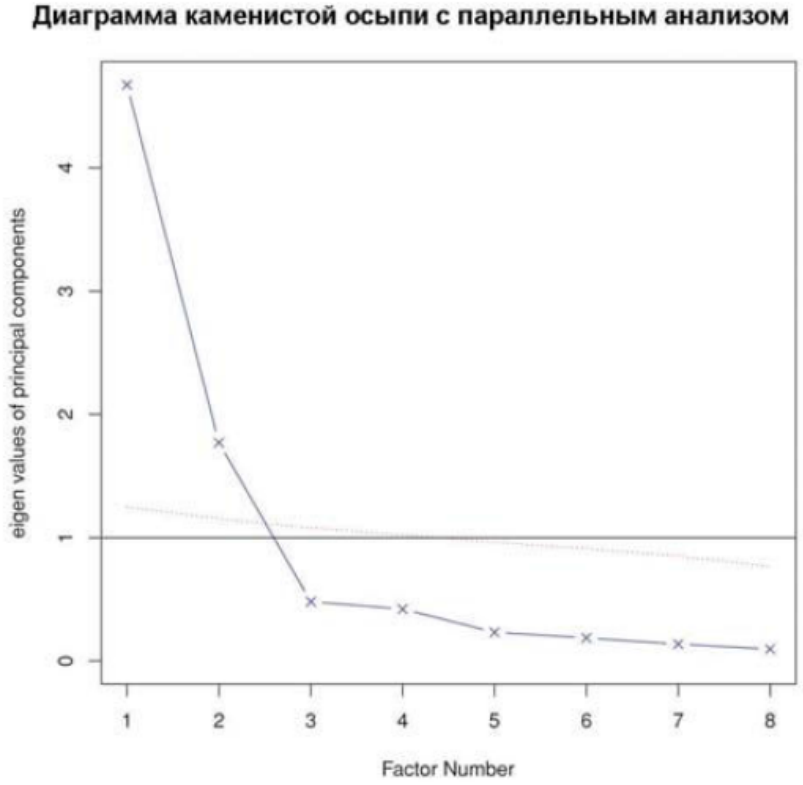
где `r` – корреляционная матрица, или исходная таблица данных; `nfactors` – определяет число главных компонент, которые нужно выделить (по умолчанию одна); `rotate` – указывает, какой тип вращения нужно применить (по умолчанию варимакс); `scores` – определяет, нужно ли рассчитывать значения главных компонент (по умолчанию – нет).

```
> pc <- principal(USJudgeRatings[,-1], nfactors=1)
> pc
Principal Components Analysis
Call: principal(r = USJudgeRatings[, -1], nfactors=1)
Standardized loadings based upon correlation matrix
      PC1   h2    u2
INTG 0.92 0.84 0.157
DMNR 0.91 0.83 0.166
DILG 0.97 0.94 0.061
CFMG 0.96 0.93 0.072
DECI 0.96 0.92 0.076
PREP 0.98 0.97 0.030
FAMI 0.98 0.95 0.047
ORAL 1.00 0.99 0.009
WRIT 0.99 0.98 0.020
PHYS 0.89 0.80 0.201
RTEN 0.99 0.97 0.028
                 PC1
SS loadings    10.13
Proportion Var  0.92
```

6.2. Главные компоненты

Промеры тела у девочек

```
library(psych)
fa.parallel(Harman23.cor$cov, n.obs=302, fa="pc", ntrials=100,
            show.legend=FALSE, main="Диаграмма каменистой осыпи
➥с параллельным анализом")
```



6.2. Главные компоненты

```
> library(psych)
> PC <- principal(Harman23.cor$cov, nfactors=2, rotate="none")
> PC
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 2, rotate = "none")
Standardized loadings based upon correlation matrix
                 PC1   PC2   h2    u2
height          0.86 -0.37 0.88 0.123
arm.span        0.84 -0.44 0.90 0.097
forearm         0.81 -0.46 0.87 0.128
lower.leg       0.84 -0.40 0.86 0.139
weight          0.76  0.52 0.85 0.150
bitro.diameter  0.67  0.53 0.74 0.261
chest.girth     0.62  0.58 0.72 0.283
chest.width     0.67  0.42 0.62 0.375
                 PC1  PC2
SS loadings     4.67 1.77
Proportion Var  0.58 0.22
Cumulative Var  0.58 0.81
```

ВРАЩЕНИЕ ГЛАВНЫХ КОМПОНЕНТ

Вращение – это набор математических приемов трансформации матрицы нагрузок компонент в другую, более легко интерпретируемую.

Способы вращения различаются по тому, остаются ли получившиеся компоненты нескоррелированными (*ортогональное вращение*) или же допускается их корреляция (*наклонное вращение*).

Наиболее распространенный тип ортогонального вращения – это ***варимакс*** (varimax), при котором делается попытка очистить столбцы матрицы нагрузок так, чтобы каждая компонента была определена ограниченным набором переменных (то есть в каждом столбце будет лишь несколько больших нагрузок и много очень малых).

6.2. Главные компоненты

```
> rc <- principal(Harman23.cor$cov, nfactors=2, rotate="varimax")
> rc
Principal Components Analysis
Call: principal(r = Harman23.cor$cov, nfactors = 2, rotate = "varimax")
Standardized loadings based upon correlation matrix
                 RC1  RC2   h2    u2
height          0.90 0.25 0.88 0.123
arm.span        0.93 0.19 0.90 0.097
forearm         0.92 0.16 0.87 0.128
lower.leg       0.90 0.22 0.86 0.139
weight          0.26 0.88 0.85 0.150
bitro.diameter 0.19 0.84 0.74 0.261
chest.girth     0.11 0.84 0.72 0.283
chest.width     0.26 0.75 0.62 0.375
                 RC1  RC2
SS loadings     3.52 2.92
Proportion Var 0.44 0.37
Cumulative Var 0.44 0.81
```

6.2. Главные компоненты

ВЫЧИСЛЕНИЕ ЗНАЧЕНИЙ ГЛАВНЫХ КОМПОНЕНТ

```
> library(psych)
> rc <- principal(Harman23.cor$cov, nfactors=2, rotate="varimax")
> round(unclass(rc$weights), 2)
                 RC1   RC2
height          0.28 -0.05
arm.span        0.30 -0.08
forearm         0.30 -0.09
lower.leg       0.28 -0.06
weight         -0.06  0.33
bitro.diameter -0.08  0.32
chest.girth    -0.10  0.34
chest.width    -0.04  0.27
```

```
PC1 = 0.28*height + 0.30*arm.span + 0.30*forearm + 0.29*lower.leg -
      0.06*weight - 0.08*bitro.diameter - 0.10*chest.girth -
      0.04*chest.width
```

```
PC2 = -0.05*height - 0.08*arm.span - 0.09*forearm - 0.06*lower.leg +
       0.33*weight + 0.32*bitro.diameter + 0.34*chest.girth +
       0.27*chest.width
```

6.3. Разведочный факторный анализ

$$X_i = a_1F_1 + a_2F_2 + \ldots + a_pF_p + U_i,$$

где X_i – это i-ая наблюдаемая переменная (i = 1…k), F_j – это общие факторы (j = 1…p), и $p < k$. U_i – это уникальная составляющая переменной X_i (не объясненная общими факторами). Параметр a_1 можно интерпретировать как степень вклада каждого фактора в наблюдаемую переменную.

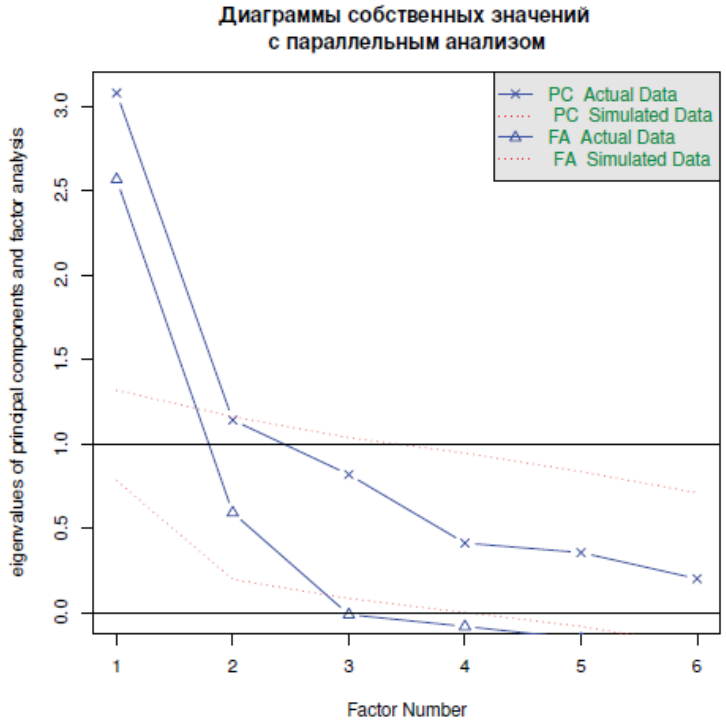
Пример: невербальная оценка общего умственного развития (general), тест на завершение фигур (picture), тест блочных конструкций (blocks), тест с лабиринтом (maze), тест на понимание прочитанного (reading) и тест на словарный запас (vocab).

```
> options(digits=2)
> covariances <- ability.cov$cov
> correlations <- cov2cor(covariances)
> correlations
        general picture blocks maze reading vocab
general    1.00    0.47   0.55 0.34    0.58  0.51
picture    0.47    1.00   0.57 0.19    0.26  0.24
blocks     0.55    0.57   1.00 0.45    0.35  0.36
maze       0.34    0.19   0.45 1.00    0.18  0.22
reading    0.58    0.26   0.35 0.18    1.00  0.79
vocab      0.51    0.24   0.36 0.22    0.79  1.00
```

6.3. Разведочный факторный анализ

ОПРЕДЕЛЕНИЕ ЧИСЛА ИЗВЛЕКАЕМЫХ ФАКТОРОВ

```
> library(psych)
> covariances <- ability.cov$cov
> correlations <- cov2cor(covariances)
> fa.parallel(correlations, n.obs=112, fa="both", n.iter=100,
              main=" Диаграммы собственных значений с параллельным
➥анализом")
```



6.3. Разведочный факторный анализ

ВЫДЕЛЕНИЕ ОБЩИХ ФАКТОРОВ

```
fa(r, nfactors=, n.obs=, rotate=, scores=, fm=),
```

где `r` – это корреляционная матрица или таблица исходных данных; `nfactors` определяет число факторов, которое нужно выделить (1 по умолчанию); `n.obs` – число наблюдений (если анализируется корреляционная матрица); `rotate` определяет тип вращения факторов (по умолчанию облимин, `oblimin`); `scores` указывает, нужно ли вычислять значения факторов (по умолчанию – нет); `fm` задает метод факторного анализа (по умолчанию минрез, `minres`).

```
> fa <- fa(correlations, nfactors=2, rotate="none", fm="pa")
> fa
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 2, rotate = "none", fm = "pa")
Standardized loadings based upon correlation matrix
          PA1   PA2   h2   u2
general 0.75  0.07 0.57 0.43
picture 0.52  0.32 0.38 0.62
blocks  0.75  0.52 0.83 0.17
maze    0.39  0.22 0.20 0.80
reading 0.81 -0.51 0.91 0.09
vocab   0.73 -0.39 0.69 0.31
                PA1  PA2
SS loadings    2.75 0.83
Proportion Var 0.46 0.14
Cumulative Var 0.46 0.60
```

6.3. Разведочный факторный анализ

ВРАЩЕНИЕ ФАКТОРОВ

Ортогональное

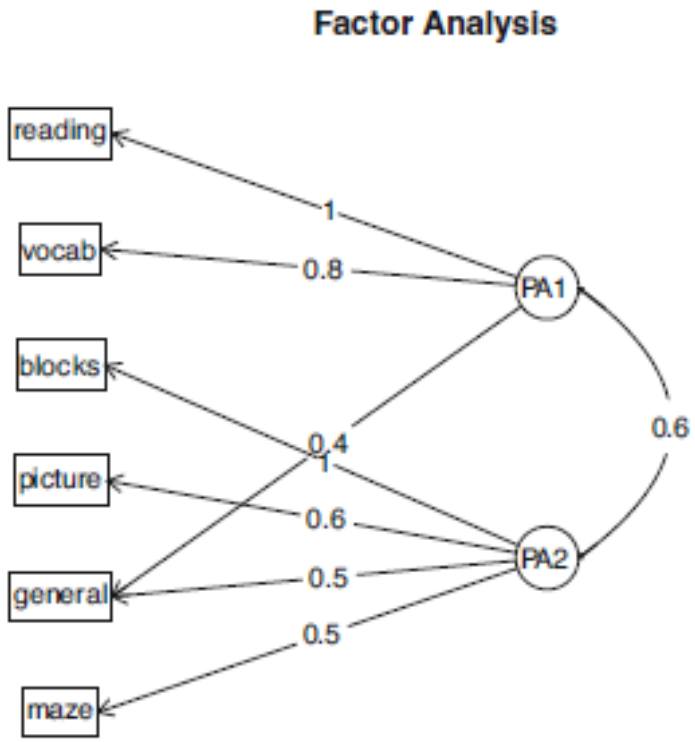
```
> fa.varimax <- fa(correlations, nfactors=2, rotate="varimax", fm="pa")
> fa.varimax
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 2, rotate = "varimax", fm = "pa")
Standardized loadings based upon correlation matrix
          PA1  PA2   h2   u2
general 0.49 0.57 0.57 0.43
picture 0.16 0.59 0.38 0.62
blocks  0.18 0.89 0.83 0.17
maze    0.13 0.43 0.20 0.80
reading 0.93 0.20 0.91 0.09
vocab   0.80 0.23 0.69 0.31
                PA1  PA2
SS loadings    1.83 1.75

Proportion Var 0.30 0.29
Cumulative Var 0.30 0.60
```

Наклонное

```
> fa.promax <- fa(correlations, nfactors=2, rotate="promax", fm="pa")
> fa.promax
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 2, rotate = "promax", fm = "pa")
Standardized loadings based upon correlation matrix
           PA1   PA2   h2   u2
general  0.36  0.49 0.57 0.43
picture -0.04  0.64 0.38 0.62
blocks  -0.12  0.98 0.83 0.17
maze    -0.01  0.45 0.20 0.80
reading  1.01 -0.11 0.91 0.09
vocab    0.84 -0.02 0.69 0.31
                PA1  PA2
SS loadings    1.82 1.76
Proportion Var 0.30 0.29
Cumulative Var 0.30 0.60
 With factor correlations of
     PA1  PA2
PA1 1.00 0.57
PA2 0.57 1.00
```

6.3. Разведочный факторный анализ

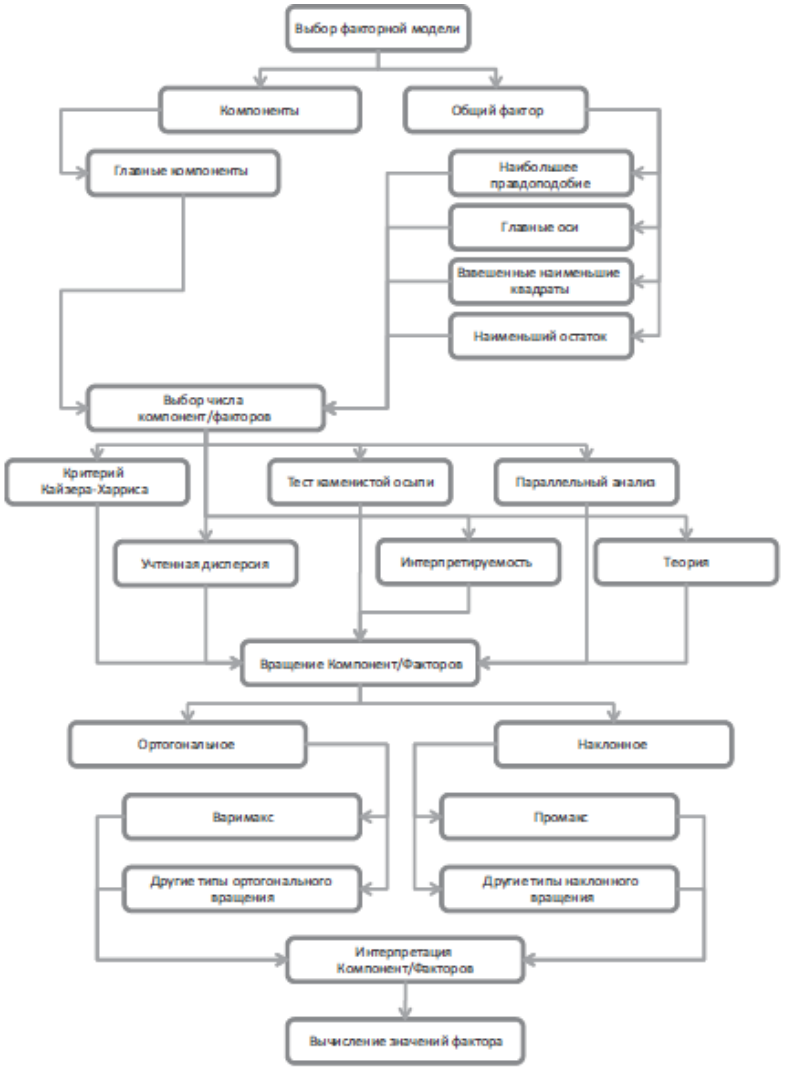
```
fa.diagram(fa.promax, simple=FALSE)
```



6.3. Разведочный факторный анализ

ЗНАЧЕНИЯ ФАКТОРОВ

```
> fa.promax$weights
          [,1]  [,2]
general 0.080 0.210
picture 0.021 0.090
blocks  0.044 0.695
maze    0.027 0.035
reading 0.739 0.044
vocab   0.176 0.039
```



Библиографический список

Кабаков Р. К. (2014) R в действии. Анализ и визуализация данных на языке R Издательство: ДМК Пресс, 580 с.

Шипунов А. Б., Балдин Е. М., Волкова П. А., Коробейников А. И., Назарова С. А., Петров С. В., Суфиянов В. Г. (2012) Наглядная статистика. Используем R! - М.: ДМК Пресс, 298 с.