НИУ ВШЭ, ОП «Политология», 2022
Курс «Введение в многомерный статистический анализ»

Домашнее задание 3
Рубанов Владислав БПТ 201

Теоретическая часть

Задача 1.

Дана ковариационная матрица для двумерного массива, состоящего из переменных X и Y:

$$\Sigma = \begin{pmatrix} 9 & 2 \\ 2 & 6 \end{pmatrix}$$

Реализуем **метод главных компонент** вручную по шагам, разобранным на лекции. Получим на выходе две вещи: собственные значения и матрицу поворота.

Начнем с поиска собственных значений матрицы Σ.

Решим следующее характеристическое уравнение:

$$det(\Sigma - \lambda I) = 0$$

$$det(\begin{pmatrix} 9 & 2 \\ 2 & 6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}) = 0$$
$$\begin{vmatrix} 9-\lambda & 2 \\ 2 & 6-\lambda \end{vmatrix} = 0$$

Посчитаем определитель и решим полученное квадратное уравнение.

$$(9-\lambda)\cdot(6-\lambda) - 2\cdot 2 = 0$$
$$54 - 9\lambda - 6\lambda + \lambda^2 - 4 = 0$$
$$\lambda^2 - 15\lambda + 50 = 0$$

Решим квадратное уравнение через дискриминант.

$$D = b^2 - 4ac = (-15)^2 - 4\cdot 1\cdot 50 = 25$$
$$\lambda_1 = \frac{-b+\sqrt{D}}{2a} = \frac{15+5}{2} = 10$$
$$\lambda_2 = \frac{-b-\sqrt{D}}{2a} = \frac{15-5}{2} = 5$$

Итак, мы нашли собственные значения ковариационной матрицы $\Sigma = \begin{pmatrix} 9 & 2 \\ 2 & 6 \end{pmatrix}$. Очевидно, что первым собственным значением будет 10 ($\lambda_1 = 10$), а вторым — 5 ($\lambda_2 = 5$), поскольку собственное значение соответствует дисперсии главной компоненты (ГК), а мы знаем, что дисперсия первой ГК всегда больше дисперсии второй ГК и т.д. Таким образом, матрица собственных значений S будет выглядеть следующим образом:

$$S = \begin{pmatrix} 10 & 0 \\ 0 & 5 \end{pmatrix}$$

Далее **найдем собственные векторы** ковариационной матрицы Σ, наложив ограничение на их длину, равное 1 для однозначности решения.

Итак, из определения собственных векторов мы знаем, что

$$\Sigma \cdot \vec{a} = \lambda \cdot \vec{a}$$

где $\vec{a}$ — собственный вектор, а λ — собственное значение матрицы Σ. Значит,

$$\begin{cases} \Sigma \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \lambda_1 \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} \\ a_{11}^2 + a_{21}^2 = 1 \end{cases}$$

И

$$\begin{cases} \Sigma \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = \lambda_1 \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} \\ a_{12}^2 + a_{22}^2 = 1 \end{cases}$$

Найдем первый собственный вектор $\vec{a_1} = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$

$$\begin{pmatrix} 9 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = 10 \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$$
$$\begin{pmatrix} 9a_{11} + 2a_{21} \\ 2a_{11} + 6a_{21} \end{pmatrix} = \begin{pmatrix} 10a_{11} \\ 10a_{21} \end{pmatrix}$$

Теперь составим систему линейных уравнений, добавив ограничение на длину собственного вектора.

$$\begin{cases} 9a_{11} + 2a_{21} = 10a_{11} \\ 2a_{11} + 6a_{21} = 10a_{21} \\ a_{11}^2 + a_{21}^2 = 1 \end{cases}$$

Итак, мы можем видеть, что первые два уравнения получились *линейно зависимыми*. Значит, мы можем выбрать из них любое и выразить через него третье уравнение. Возьмем первое: так будет удобнее, а также перенесем правую часть уравнения налево.

$$\begin{cases} -a_{11} + 2a_{21} = 0 \\ a_{11}^2 + a_{21}^2 = 1 \end{cases}$$

$$-a_{11} + 2a_{21} = 0$$
$$a_{11} = 2a_{21}$$

Подставим во второе уравнение:

$$\begin{aligned} 4a_{21}^2 + a_{21}^2 &= 1 \\ 5a_{21}^2 &= 1 \\ a_{21}^2 &= \frac{1}{5} \\ a_{21} &= \pm\frac{1}{\sqrt{5}} \\ a_{21} &= -\frac{1}{\sqrt{5}} \end{aligned}$$

Берем **отрицательный знак** так, чтобы у элементов a_{21} и a_{22} (далее) получился отрицательный знак, как было сказано делать для того, чтобы мы получили корректную матрицу поворота при транспонировании матрицы из собственных векторов. Следовательно,

$$\begin{aligned} a_{11} &= 2a_{21} \\ a_{11} &= -\frac{2}{\sqrt{5}} \end{aligned}$$

Итак, мы нашли первый собственный вектор матрицы Σ. $\vec{a_1} = \begin{pmatrix} -\frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{pmatrix}$

Подставим его в матрицу собственных векторов A:

$$A = \begin{pmatrix} -\frac{2}{\sqrt{5}} & a_{12} \\ -\frac{1}{\sqrt{5}} & a_{22} \end{pmatrix}$$

Теперь найдем **второй собственный вектор**. Подставим второе собственное значение в уравнение выше.

$$\begin{pmatrix} 9 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = 5 \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}$$
$$\begin{pmatrix} 9a_{12} + 2a_{22} \\ 2a_{12} + 6a_{22} \end{pmatrix} = \begin{pmatrix} 5a_{12} \\ 5a_{22} \end{pmatrix}$$

Теперь составим систему линейных уравнений, добавив ограничение на длину собственного вектора.

$$\begin{cases} 9a_{12} + 2a_{22} = 5a_{12} \\ 2a_{12} + 6a_{22} = 5a_{22} \\ a_{12}^2 + a_{22}^2 = 1 \end{cases}$$

Мы снова получили два линейно зависимых уравнения. Возьмем второе: так будет удобнее, а также перенесем правую часть уравнения налево.

$$\begin{cases} 2a_{12} + a_{22} = 0 \\ a_{12}^2 + a_{22}^2 = 1 \end{cases}$$

Выразим a_{22}:

$$\begin{aligned} 2a_{12} + a_{22} &= 0 \\ a_{22} &= -2a_{12} \end{aligned}$$

Подставим в первое уравнение:

$$\begin{aligned} a_{12}^2 + 4a_{12}^2 &= 1 \\ 5a_{12}^2 &= 1 \\ a_{12}^2 &= \frac{1}{5} \\ a_{12} &= \pm\frac{1}{\sqrt{5}} \\ a_{12} &= \frac{1}{\sqrt{5}} \end{aligned}$$

Берем **положительный знак**, чтобы элемент a_{22} получил отрицательный знак, как было упомянуто ранее. Следовательно,

$$\begin{aligned} a_{22} &= -2a_{12} \\ a_{22} &= -2 \cdot \frac{1}{\sqrt{5}} \\ a_{22} &= -\frac{2}{\sqrt{5}} \end{aligned}$$

Итак, мы нашли второй собственный вектор матрицы Σ. $\vec{a_2} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix}$

Подставим его в матрицу собственных векторов A:

$$A = \begin{pmatrix} -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}$$

Ура! Мы нашли оба собственных вектора ковариационной матрицы Σ.
Можем дополнительно убедиться в том, что два вектора ортогональны:

$$\vec{a_1} \cdot \vec{a_2} = \left(-\frac{2}{\sqrt{5}} \cdot \frac{1}{\sqrt{5}}\right) + \left(-\frac{1}{\sqrt{5}} \cdot -\frac{2}{\sqrt{5}}\right) = 0$$

Теперь проверим себя в R.

```
A <- matrix(c(9, 2, 2, 6), nrow = 2, byrow = TRUE)
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 10  5
##
## $vectors
##            [,1]       [,2]
## [1,] -0.8944272  0.4472136
## [2,] -0.4472136 -0.8944272
```

Во-первых, мы можем заметить, что собственные значения были найдены правильно. **Во-вторых**, казалось бы, можно увидеть что-то странное в получившейся матрице собственных векторов.

Но нет: это действительно данные значения — убедимся в этом:

```
1/sqrt(5)

## [1] 0.4472136

2/sqrt(5)

## [1] 0.8944272
```

Ура! Матрица из собственных векторов была найдена верно!

Теперь осталось лишь превратить ее в матрицу поворота для перехода к новому базису, транспонировав:

$$A^T = \begin{pmatrix} -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}^T = \begin{pmatrix} -\frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}$$

Матрица поворота найдена!

Задача 2.

Найдем значение первой главной компоненты для наблюдения A, если про него известно следующее: центрированно-нормированное значение X равно 5.5, а центрированно-нормированное значение Y равно 3.2.

Мы знаем, что

$$PC = XA^T$$

Значит, для нахождения первой ГК для наблюдения A нам потребуется первый вектор из матрицы поворота, соответствующий первой ГК:

$$PC_1 = \begin{pmatrix} 5.5 \\ 3.2 \end{pmatrix} \cdot \begin{pmatrix} -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix} = \left(5.5 \cdot -\frac{2}{\sqrt{5}}\right) + \left(3.2 \cdot \frac{1}{\sqrt{5}}\right) \approx -3.49$$

Итак, значение первой ГК для наблюдения A равно примерно $-3{,}49$.

Практическая часть

Задача 1.

Загрузим данные, которые содержатся в файле `protests.csv`, и **сохраним** их в датафрейм `pro`.

```
pro <- read.csv("protests.csv")
```

Выберем из датафрейма `pro` перечисленные в описании столбцы и **сохраним** их в отдельный датафрейм `d`. **Удалим** строки с пропущенными значениями.

```
library(tidyverse)
d <- pro %>% dplyr::select(country, year, duration, part2,
                           reaction, arrested, wounded)
d <- na.omit(d)
```

Посмотрим на первые наблюдения в получившемся наборе данных, на его структуру и описательные статистики:

```
head(d)
```

```
##       country year duration part2 reaction arrested wounded
## 1 netherlands   75        1    60       60        0       0
## 2 netherlands   75        1    50       60        0       0
## 3 netherlands   75        1   100       60        0       0
## 4 netherlands   75        1     1       60        0       0
## 5 netherlands   75        1     1       60        0       0
## 6 netherlands   75        1   850       60        0       0
```

```
str(d)
```

```
## 'data.frame': 8562 obs. of  7 variables:
##  $ country : chr  "netherlands" "netherlands" "netherlands" "netherlands" ...
##  $ year    : int  75 75 75 75 75 75 75 75 75 75 ...
##  $ duration: int  1 1 1 1 1 1 1 7 1 1 ...
##  $ part2   : num  60 50 100 1 1 850 850 500 500 5000 ...
##  $ reaction: int  60 60 60 60 60 60 60 60 16 60 ...
##  $ arrested: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ wounded : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "na.action")= 'omit' Named int [1:459] 62 88 193 398 484 565 612 623 627 661 ...
##   ..- attr(*, "names")= chr [1:459] "62" "88" "193" "398" ...
```

```
summary(d)
```

```
##    country               year          duration         part2
##  Length:8562        Min.   :75.00   Min.   :1.000   Min.   :     1
##  Class :character   1st Qu.:79.00   1st Qu.:1.000   1st Qu.:     5
##  Mode  :character   Median :82.00   Median :1.000   Median :   150
##                     Mean   :82.38   Mean   :1.387   Mean   :  4561
##                     3rd Qu.:86.00   3rd Qu.:1.000   3rd Qu.:  1000
##                     Max.   :89.00   Max.   :9.000   Max.   :999998
##     reaction        arrested          wounded
```

```
##  Min.   :10.00   Min.   :  0.000   Min.   : 0.0000
##  1st Qu.:60.00   1st Qu.:  0.000   1st Qu.: 0.0000
##  Median :60.00   Median :  0.000   Median : 0.0000
##  Mean   :56.05   Mean   :  1.927   Mean   : 0.8178
##  3rd Qu.:60.00   3rd Qu.:  0.000   3rd Qu.: 0.0000
##  Max.   :60.00   Max.   :900.000   Max.   :98.0000
```

Агрегируем данные по странам и годам. Проверим, что все получилось.

```
final <- d %>% group_by(country, year) %>%
  summarise(duration = sum(duration),
            part2 = sum(part2),
            reaction = sum(reaction),
            arrested = sum(arrested),
            wounded = sum(wounded))
final <- as.data.frame(final)

head(final)

##   country year duration  part2 reaction arrested wounded
## 1  france   75      360 629455    11919      136      14
## 2  france   76      260 360693    10845       63      31
## 3  france   77      287 715452    10834       70       7
## 4  france   78      272 246551     8395      104      31
## 5  france   79      187 539567     8944      112     122
## 6  france   80      443 460437    13108      123      84
```

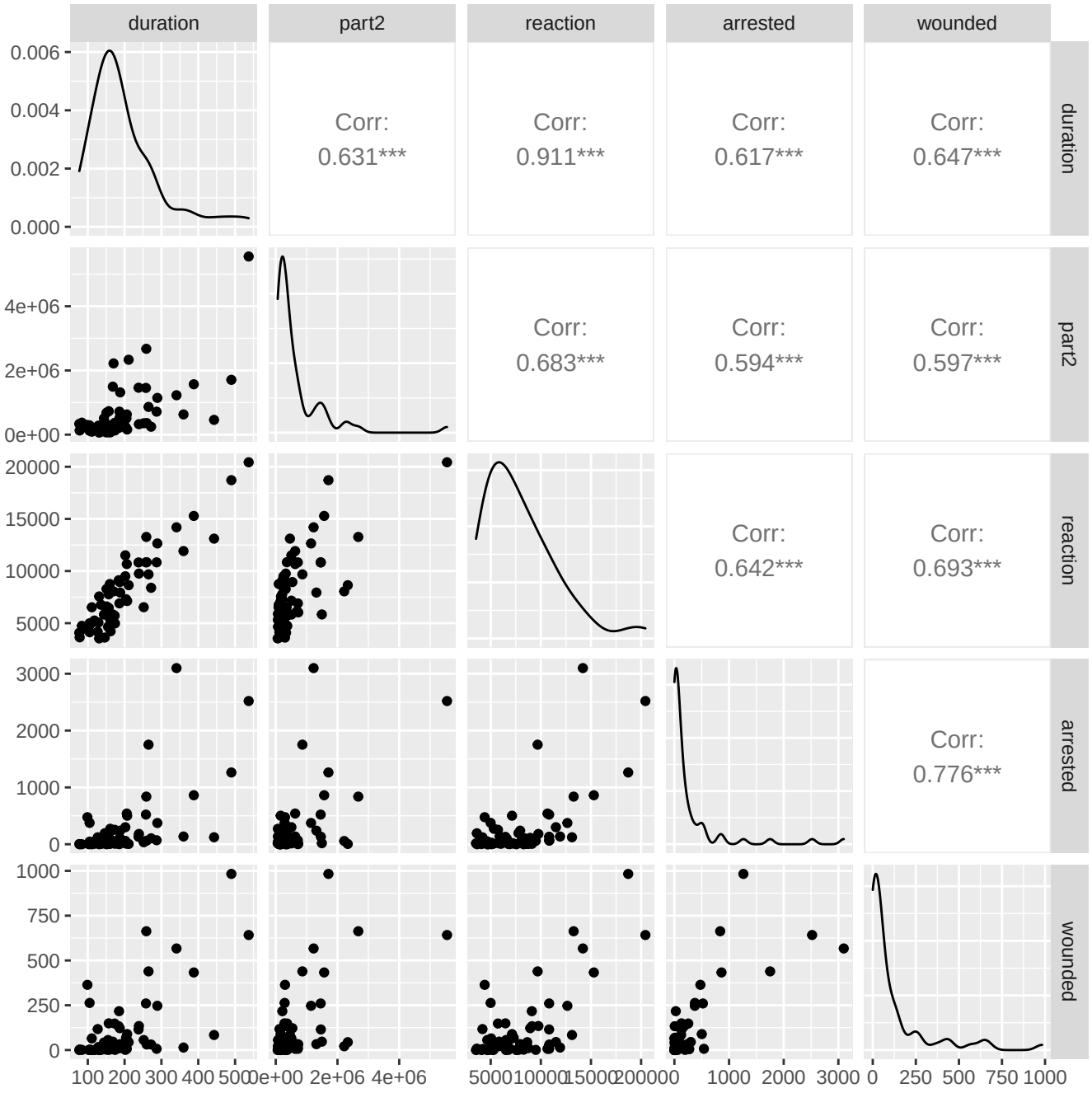
Сохраним в датафрейм `m` все переменные из `final`, *кроме года и названия страны*, и снова посмотрим на структуру данных и описательные статистики.

```
m <- final %>% dplyr::select(duration, part2, reaction,
                             arrested, wounded)

str(m)

## 'data.frame': 60 obs. of  5 variables:
##  $ duration: int  360 260 287 272 187 443 204 184 170 211 ...
##  $ part2   : num  629455 360693 715452 246551 539567 ...
##  $ reaction: int  11919 10845 10834 8395 8944 13108 7295 9114 8051 8642 ...
##  $ arrested: int  136 63 70 104 112 123 11 7 57 5 ...
##  $ wounded : int  14 31 7 31 122 84 69 133 21 44 ...

summary(m)

##     duration         part2            reaction        arrested
##  Min.   : 77.0   Min.   :  62572   Min.   : 3537   Min.   :   0.0
##  1st Qu.:142.0   1st Qu.: 158335   1st Qu.: 5206   1st Qu.:  10.5
##  Median :171.5   Median : 324446   Median : 7144   Median :  61.0
##  Mean   :197.9   Mean   : 650896   Mean   : 7999   Mean   : 275.1
##  3rd Qu.:238.2   3rd Qu.: 715708   3rd Qu.: 9701   3rd Qu.: 242.8
##  Max.   :537.0   Max.   :5549448   Max.   :20429   Max.   :3099.0
##     wounded
```

```
##  Min.   :  0.00
##  1st Qu.:  5.75
##  Median : 37.50
##  Mean   :116.70
##  3rd Qu.:124.75
##  Max.   :983.00
```

Отлично. Получилось 60 наблюдений и 5 переменных, с которыми мы будем работать. *Посмотрим на распределение* наших величин:

```
library(GGally)
ggpairs(m)
```



Можно увидеть, что переменные очень сильно скоррелированы. **Значит, применение МГК оправдано.**

Задача 2.

Реализуем на переменных из датафрейма `m` метод главных компонент и выведем его результаты.

```
pca <- prcomp(m, center = TRUE, scale = TRUE)
pca

## Standard deviations (1, .., p=5):
## [1] 1.9294196 0.7419670 0.6484467 0.4713142 0.2901804
##
## Rotation (n x k) = (5 x 5):
##                 PC1        PC2        PC3         PC4         PC5
## duration -0.4603603  0.4646695 -0.3527759  0.12505015  0.65731445
## part2    -0.4183310  0.1693309  0.8884696 -0.04182172  0.07210334
## reaction -0.4748435  0.4033850 -0.2408012 -0.01540626 -0.74403065
## arrested -0.4345467 -0.5955446 -0.0561144  0.67204217 -0.04130654
## wounded  -0.4458255 -0.4878692 -0.1582305 -0.72851587  0.08631902
```

Запишем выражение для **второй** главной компоненты:

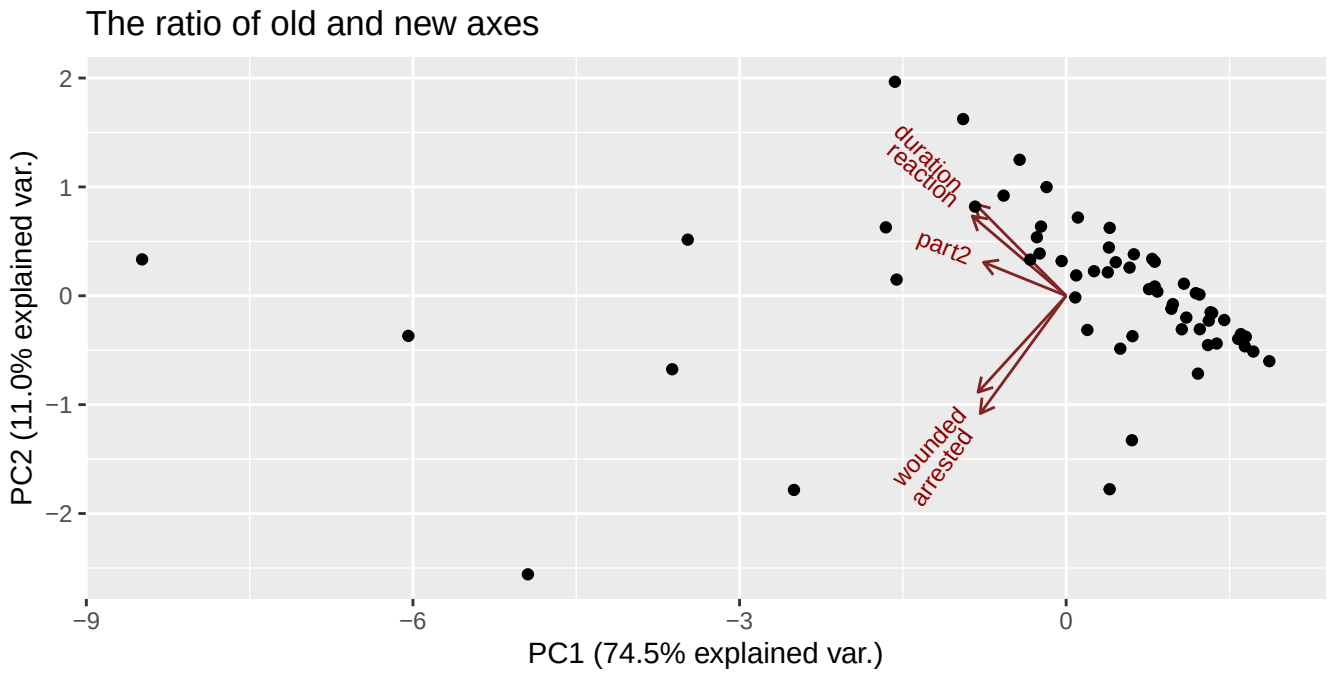
$$PC_2 = 0.4646695 \times duration + 0.1693309 \times part2 + 0.4033850 \times reaction + (-0.5955446) \times arrested + (-0.4878692) \times wounded$$

Задача 3.

Проинтерпретируем две первые главные компоненты.

Для этого нам понадобится выдача из предыдущего задания (результат реализации МГК), а также специальный график из пакета `ggbiplot`, который наложит старые оси на новые.

```
library(ggbiplot)
ggbiplot(pca, scale=0) + ggtitle("The ratio of old and new axes")
```



Итак, можно заметить, что в **первую главную компоненту (ГК)** все переменные входят с примерно одинаковым весом. Несмотря на то, что у них у всех отрицательный знак, важно, что он одинаковый. Таким образом, мы можем утверждать, что первая ГК является чем-то вроде общего **индекса „качества“ протеста**, учитывающим все переменные (*длительность протеста, число участников, реакция на протест, число арестованных, число раненых*) с одинаковым весом и знаком. К тому же, из графика можно заметить, что только эта первая ГК **объясняет сразу 74.5% информации** (дисперсии), что уже очень много само по себе. Также на графике более интуитивно заметно, что стрелочки, которыми отмечены направления исходных координатных осей, сонаправлены относительно друг друга. К тому же, это логично, учитывая высокую степень скоррелированности исходных показателей.

Во **вторую же ГК** с **высоким положительным** коэффициентом входят такие переменные, как *длительность протеста и реакция на него*, **с маленьким положительным** — *число участников* и **с большим отрицательным** — *число арестованных и раненых протестующих.*
Особенно хорошо это заметно на **графике**: две стрелочки, отвечающие за число арестованных и раненых протестующих (`arrested` и `wounded`) сильно направлены вниз, если посмотреть со стороны PC_2, стрелочки, отвечающие за длительность и реакцию (`duration` и `reaction`) — направлены сильно

вверх, а число участников (`part2`) — направлено положительно, но слабее предыдущих двух. Таким образом, данную ГК можно назвать **„мерой миролюбивости и длительности“ (и немного качества) протеста**, как странно это бы не звучало. Наибольшее значение данного индекса получат такие наблюдения, которые были дольше, привлекли большую реакцию, задействовали относительно большое число людей, но не повлекли за собой столкновений (арестов и ранений). Может быть, это относится к забастовкам или массовым шествиям и другим конструктивным видам проявлений протестной активности.

То есть, если мы будем брать 1-ю ГК в качестве некоторого интегрального индекса, она покажет нам общее „качество“ самого протеста через выраженность основных показателей, представленных в наборе данных.

Если мы возьмем первые две ГК, вторая ГК также будет отражать степень мирности протеста и его „конструктивности“ (отсутствие жертв и при этом большую реакцию).

Задача 4.

Теперь, когда мы подумали над содержательной интерпретацией полученных ГК, можно воспользоваться формальными (и не очень) методами, которые помогут нам определить: **сколько главных компонент можно извлечь**.

Воспользуемся основными методами, представленными на лекции:

1. ***Содержательная интерпретация.***

   На предыдущем шаге было замечено, что **первые две ГК** являются хорошо объяснимыми с содержательной точки зрения.

   В третью ГК с очень высоким положительным весом входит только число участников, остальные показатели — с отрицательным и не таким высоким (а число арестованных почти с нулевым весом). То есть можно сказать, что третья ГК отвечает практически исключительно за численность участников. Несмотря на то, что это довольно полезная информация, можно сказать, что, если мы будем брать 3 ГК, начиная с нее начнется некоторое **дублирование информации** (численность и до этого хорошо учитывалась). Поэтому в целях снижения размерности я бы предпочел взять первые 2 ГК, исходя из содержательных аргументов. Содержательно они вполне удовлетворяют нашим целям, а также просты в визуализации и интерпретации.

2. ***Выбор по информативности.***

   Следующий метод: выбрать столько ГК, чтобы они объясняли не менее 75-80% дисперсии исходных данных.

   Посмотрим на выдачу функции `summary`:

```
summary(pca)

## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.9294 0.7420 0.6484 0.47131 0.29018
## Proportion of Variance 0.7445 0.1101 0.0841 0.04443 0.01684
## Cumulative Proportion  0.7445 0.8546 0.9387 0.98316 1.00000
```

   Итак, можно заметить, что первая ГК (как было замечено ранее) уже объясняет примерно 74% информации, а первые две ГК в сумме — больше 85%. Этого более, чем достаточно для минимизации потери информации. Поэтому, исходя из метода информативности, я также бы выбрал **две первые ГК**.

Однако следует заметить, что выбор одной первой ГК также не является плохим. Если наша цель — построить единый интегральный индекс для протестной активности — одна ГК хорошо бы с этим справилась.

3. ***Метод Кайзера.***

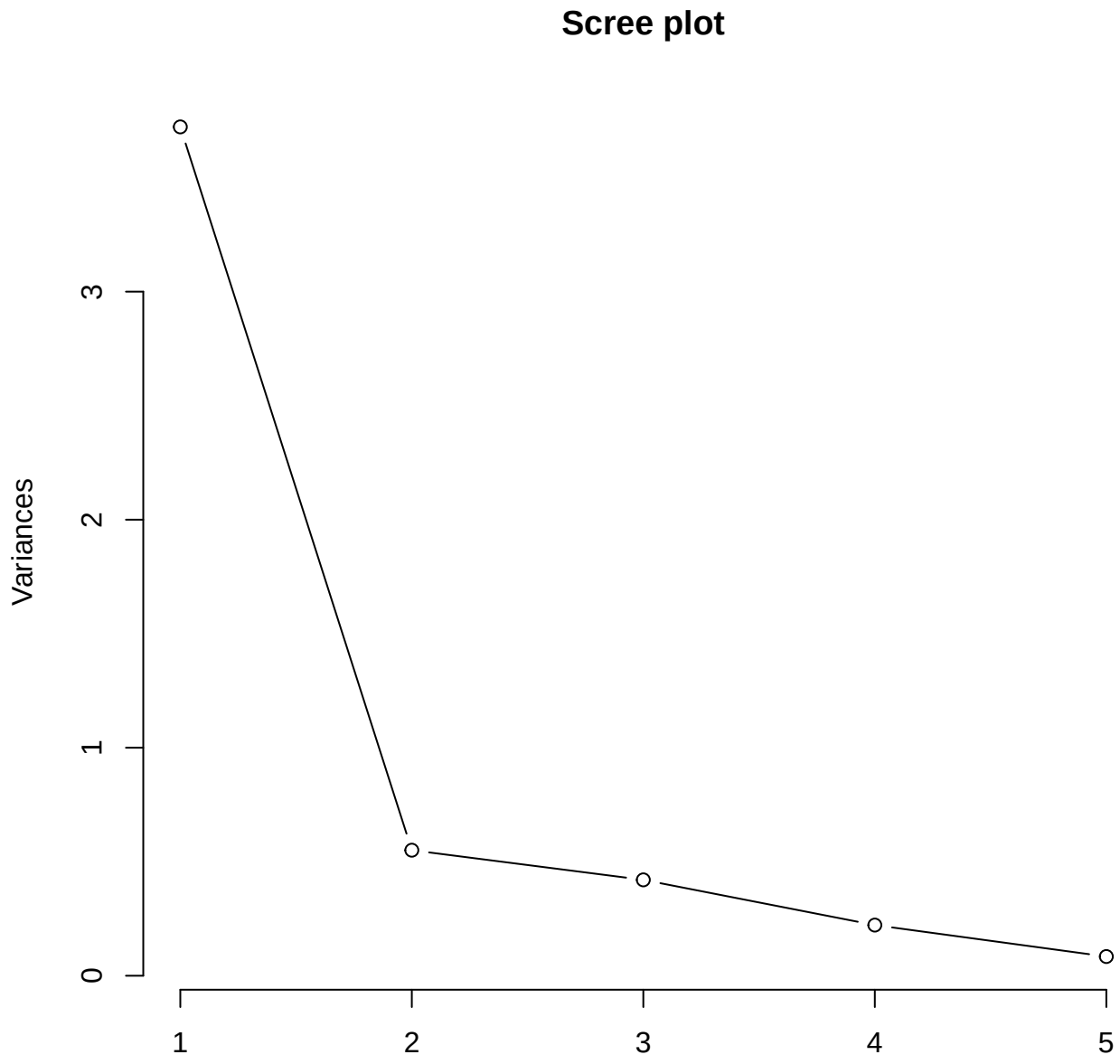
   Согласно методу кайзера, нам нужно выбрать столько ГК, у скольких собственные значения больше 1. Как известно, собственные значения ковариационной матрицы в контексте МГК — это дисперсии соответствующих ГК — поэтому нам лишь необходимо возвести стандартные отклонения из предыдущего пункта в квадрат.

   Таким образом, согласно этому методу, следует выбрать **только первую ГК**: ее дисперсия равна 3.723.

4. ***Метод Кеттела (визуальный).***

   Далее будет представлен график „scree plot“ или „каменистой осыпи“:

```
plot(pca, type = "l", main = "Scree plot")
```



Слева как раз-таки видны дисперсии (или собственные значения в чистом виде).

График показывает, что на 2-й компоненте есть излом: дисперсия начинает убывать не так сильно, т.е. добавление новых ГК не сильно увеличивает процент изменчивости исходных данных.

Таким образом, судя по графику, нам следует взять **две первые ГК**.

Итак, настало время взвесить все реализованные методы.

Обычно выбор числа ГК, которые следует извлечь, зависит от поставленной задачи. Поскольку в качестве задачи не стояло построение единого интегрального индекса, состоящего из одной ГК (несмотря на то, что, как выяснилось, это вышло бы неплохо), согласно большинству методов (кроме метода Кайзера) нам ***следует выбрать две первые ГК для дальнейшей работы.***

Хотелось бы отметить, что это вполне **соответствует как моим ожиданиями, так и содержательным соображениям**, поскольку было видно, что первые две ГК в полной мере объясняют большую часть изменчивости данных, а также легко интерпретируются. Выбор большего числа ГК не принес бы большого прироста в объяснении вариации, а также усложнил бы интерпретацию.

Задача 4. (Дополнительная)

Мне также показалось очень интересным учесть дополнительный фактор в виде страны, в которой происходили протесты, поскольку этот фактор можно попробовать довольно интересно визуализировать.

Я решил проделать это в начале на агрегированных данных, а затем — на первичных (с большим числом наблюдений). Включать в датасет год не представляется полезным, поскольку он может смешать результаты МГК.

Для этого я создал дополнительный датасет, в который включил переменную, отвечающую за название страны, а затем перекодировал ее в целочисленный формат:

- 1 — Франция
- 2 — Германия
- 3 — Нидерланды
- 4 — Швейцария

```
final2 <- final %>% dplyr::select(country, duration, part2,
                            reaction, arrested, wounded)
final2$country <- ifelse(final2$country == 'france', 1,
                         ifelse(final2$country == 'germany', 2,
                                ifelse(final2$country == 'netherlands', 3, 4)))
```

И вновь реализовал МГК:

```
pca2 <- prcomp(final2, center = TRUE, scale = TRUE)
pca2

## Standard deviations (1, .., p=6):
## [1] 1.9651196 0.9776999 0.6969700 0.6413323 0.4669842 0.2593437
##
## Rotation (n x k) = (6 x 6):
##                 PC1         PC2         PC3         PC4         PC5         PC6
## country   0.2157406 -0.88737413 -0.33076306 -0.16774499  0.07820266  0.14954893
```

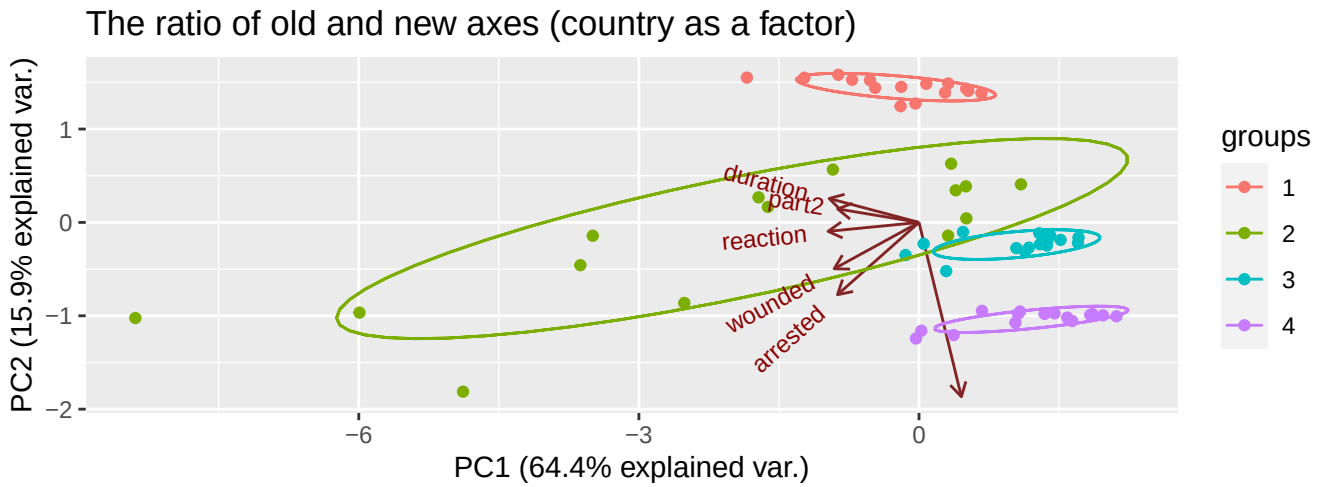
```
## duration -0.4574332  0.12174726 -0.48747642  0.25753182 -0.14864522  0.67073183
## part2    -0.4138855  0.06870162  0.06698305 -0.89563995  0.06773234  0.11284289
## reaction -0.4627353 -0.04548556 -0.52349568  0.08876387  0.08226924 -0.70364205
## arrested -0.4149594 -0.36791494  0.44350694  0.13275593 -0.68468562 -0.09659376
## wounded  -0.4319657 -0.23578487  0.42157997  0.27904282  0.70116948  0.10284901
```

Из-за включения новой переменной (которую R видит как количественный показатель) значения в матрице поворота поменялись, следовательно, это также отразится и на графике. Однако первая ГК все еще отвечает за общее „качество“ протеста: все коэффициенты, кроме страны, снова отрицательные. Но вот вторая ГК теперь в большей степени отвечает за разделение по странам (и немного по числу арестов и числу раненных — эти переменные имеют такой же знак).

Кроме того, вспоминая, что во вторую ГК, кроме страны, с таким же знаком идут число арестованных и раненых, можно заметить, что это также могло помочь такому явному разделению эллипсов: во Франции действительно были „спокойные“ годы, когда в протестах за год участвовало большое число людей, но число арестованных и раненых было очень маленьким (меньше 10 при более 700 тыс. участников), в отличие от Германии, где почти всегда было не менее 100 арестованных и раненых за год. При этом в Нидерландах и Швейцарии за год в принципе редко набиралось больше 10 арестованных и раненых, но и общая численность участников там была небольшой. Поэтому по PC_2 они находятся даже ниже Германии (высокая численность там „пересилила“ мирность выступлений в этих двух странах).

Теперь можно построить **график**, приняв переменную, ответственную за название страны, в качестве факторной и окрасив наблюдения соответствующими цветами:

```
ggbiplot(pca2, scale=0, groups = as.factor(final2$country), ellipse = T) +
  ggtitle("The ratio of old and new axes (country as a factor)")
```

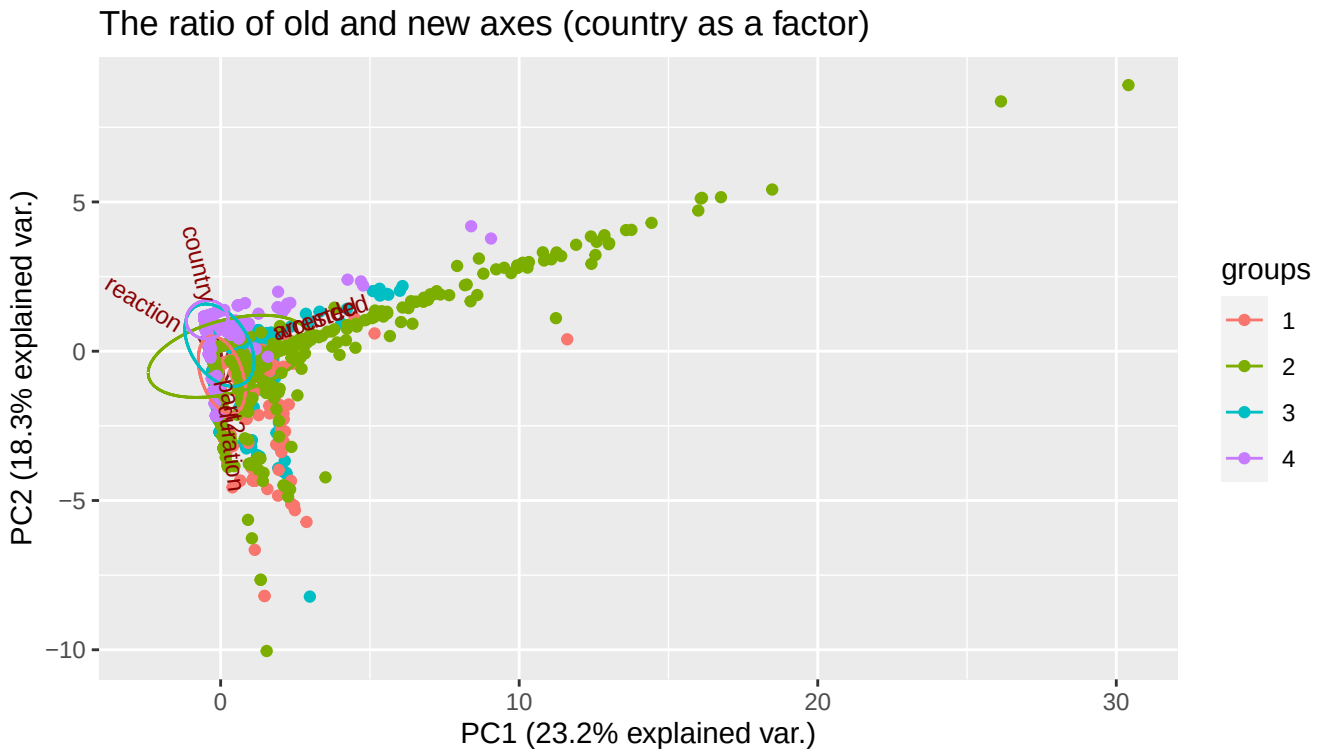


С одной стороны, можно заметить явное разделение групп наблюдений по странам по PC_2. Однако также можно заметить, что в Германии наблюдались более крупные протесты, в отличие от Нидерландов и Швейцарии. Франция в этом отношении находится между ними (по PC_1). Хотя нужно отметить, что в плане „качества“ (масштаба) протестов, Германии свойственнен большой разброс.

Попробуем сделать то же самое для первичных данных, т.е. без агрегирования по странам и годам.

```
d2 <- d %>% dplyr::select(country, duration, part2,
                          reaction, arrested, wounded)
d2$country <- ifelse(d2$country == 'france', 1,
                     ifelse(d2$country == 'germany', 2,
                            ifelse(d2$country == 'netherlands', 3, 4)))

pca3 <- prcomp(d2, center = TRUE, scale = TRUE)
pca3

## Standard deviations (1, .., p=6):
```

```
## [1] 1.1792099 1.0466346 1.0031641 0.9669215 0.9348030 0.8359952
##
## Rotation (n x k) = (6 x 6):
##                  PC1        PC2          PC3        PC4       PC5          PC6
## country  -0.14096960  0.5897486 -0.174408092  0.7600592 0.1459767 -0.053914890
## duration  0.07643380 -0.6379722 -0.461969406  0.2967362 0.5306548 -0.063926981
## part2     0.06059742 -0.2864196  0.853151000  0.3975000 0.1683742  0.008306428
## reaction -0.43344474  0.2695950  0.159327962 -0.3972143 0.7457146 -0.013756802
## arrested  0.62428422  0.2123973 -0.008538626 -0.0536940 0.2721887  0.698651534
## wounded   0.62690372  0.2129734  0.053302423 -0.1248566 0.1963910 -0.710376178

summary(pca3)

## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6
## Standard deviation     1.1792 1.0466 1.0032 0.9669 0.9348 0.8360
## Proportion of Variance 0.2318 0.1826 0.1677 0.1558 0.1456 0.1165
## Cumulative Proportion  0.2318 0.4143 0.5820 0.7379 0.8835 1.0000

ggbiplot(pca3, scale=0, groups = as.factor(d2$country), ellipse = T) +
  ggtitle("The ratio of old and new axes (country as a factor)")
```



Можно заметить, что получилось очень плохо.
Во-первых, ГК довольно сильно поменялись и скорее смешались между собой: интерпретация ухудшилась. Также объяснительная сила первых ГК также просела и смешалась с остальными: теперь отсчека в 80% объясненной информации достигается лишь при добавлении 5-й ГК, а собственные значения больше 1 сразу у 3-х первых ГК.
Во-вторых, график также получился менее удачным и интерпретируемым. Как минимум, можно заметить, что первые 2 ГК (ведь мы строим график на двумерной плоскости) объясняют только 43% информации. Явно вновь выделяется только Германия, **но в целом второй эксперимент следует признать неудачным, в отличие от первого.**

Надеюсь, мое дополнительное задание не заняло у Вас много времени!