УДК 61:004.438:303.722.2
МРНТИ 50.53, 83.77

АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ И ФАКТОРНЫЙ АНАЛИЗ В ПРОГРАММНОЙ СРЕДЕ R

В.Л. ЕГОШИН[1], Н.В. САВВИНА[2], А.М. ГРЖИБОВСКИЙ[1-2]

[1]. Северный государственный медицинский университет, г. Архангельск, Россия
Северо-Восточный федеральный университет, г. Якутск, Россия

Егошин В.Л. – http://orcid.org/0000-0002-8407-3789
Саввина Н.В. – http://orcid.org/0000-0003-2441-6193
Гржибовский А.М. – https://orcid.org/0000-0002-5464-0498

For citing/
библиографиялық сілтеме/
библиографическая ссылка:

Egoshin VL, Savvina NV, Grjibovski AM. Principal components analysis and factor analysis in R. West Kazakhstan Medical Journal 2020; 62(1):6–14.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. R бағдарламалық ортадағы негізгі компоненттерді талдау және факторлық талдау. West Kazakhstan Medical Journal 2020; 62(1):6–14.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. Анализ главных компонент и факторный анализ в программной среде R. West Kazakhstan Medical Journal 2020; 62(1):6–14.

Principal components analysis and factor analysis in R

V.L.Egoshin[1], N.V.Savvina [2], A.M. Grjibovski [1-2]
[1]Northern State Medical University, Arkhangelsk, Russia
[2]North-Eastern Federal University, Yakutsk, Russia

In this paper we describe basic principles of using R package for principal components analysis and factor analysis. We present step-by-step guidelines and syntax for the analysis using practical example with real and freely available data to simplify educational process. In addition to the syntax we present R outputs and their interpretation.
Keywords: *R, factor analysis, principal components analysis, syntax, listing.*

R бағдарламалық ортадағы негізгі компоненттерді талдау және факторлық талдау

В.Л. Егошин[1], Н.В. Саввина[2], А.М. Гржибовский[1-2]
[1]Солтүстік мемлекеттік медицина университеті, Архангельск, Ресей
[2]Солтүстік-Шығыс федеральды университеті, Якутск, Ресей

Бұл жұмыста факторлық талдау мен негізгі компоненттерді талдауды жүзеге асыру үшін R бағдарламалық ортасын қолданудың негізгі принциптері ұсынылған. Білім алушылардың практикалық жұмысы үшін қолжетімді нақты мәліметтерді қолдана отырып практикалық мысал түрінде R қадамдық алгоритм мен синтаксис ұсынылған. Синтаксистен басқа, нәтижелері R шығаратындай, сондай-ақ олардың интерпретациясы түрінде ұсынылған.
Негізгі сөздер: *R, факторлық талдау, негізі компоненттерді талдау, синтаксис, листинг.*

Анализ главных компонент и факторный анализ в программной среде R

В.Л. Егошин[1], Н.В. Саввина[2], А.М. Гржибовский[1-2]
[1]Северный государственный медицинский университет, г. Архангельск, Россия
[2]Северо-Восточный федеральный университет, г. Якутск, Россия

В данной работе представлены основные принципы применения программной среды Rдля осуществления факторного анализа и анализа главных компонент. Представлен пошаговый алгоритм и синтаксис в Rдля в виде практического примера с использованием реальных данных, находящихся в свободном доступе для практической работы обучающихся. Помимо синтаксиса представлены результаты в том виде, как их выдает R, а также их интерпретация.
Ключевые слова: *R, факторный анализ, анализ главных компонент, синтаксис, листинг.*

Введение

Анализ главных компонент – метод многомерного статистического анализа, используемый для сокращения размерности пространства переменных с минимальной потерей полезной информации. Предложен К. Пирсоном в 1901 г. Фа́кторный анализ – метод, применяемый для изучения связей между переменными. Анализ главных компонент следует применять, если необходимо уменьшить количество коррелирующих переменных. Факторный анализ рекомендуется использовать при желании оценить теоретические модели латентных факторов, основанных на наблюда-

Гржибовский А.М.
e-mail: andrej.grjibovski@gmail.com

Received/
Келіп түсті/
Поступила:
28.02.2020

Accepted/
Басылымға қабылданды/
Принята к публикации:
13.03.2020

ISSN 2707-6180 (Print)

Published by West Kazakhstan Marat Ospanov Medical University

емых переменных. Методы используются при анализе многомерных данных в биоинформатике, психологии, социологии, эконометрике, анализ главных компонент особо значим в системах компьютерного зрения, распознавания лиц, сжатия данных. Выполнение этих методов анализа, в том числе с использованием программы R, довольно подробно изложена в различных источниках: Jolliffe (2002), Costelloand Osborne (2005), Bergmanetal. (2009), DiStefano, Zhu, and Mîndrilă (2009), Abdi and Williams (2010), Brian and Torsten (2011), Kuhn and Johnson (2013), Crawley (2013), Кабаков (2014), Gaskin and Happell (2014), Osborne (2015), Lopes (2017), Wright (2017), Nistrup (2019) и других.

При анализе главных компонент выполняется следующее:

1. стандартизация данных (centerandscale)
2. расчет собственных векторов (Eugenvectors) и собственных значений (Eugenvalues) из ковариационной или корреляционной матрицы
3. сортировка собственных значений (Eugenvalues) в порядке убывания и выбор K наибольших собственных векторов (Eugenvectors)
4. создание проекционной матрицы из выбранных собственных векторов (Eugenvectors)
5. трансформация оригинального набора данных через для получения - размерного субпространства

Цель этих действий – преобразовать данные из начального состояния в субпространство с-размерностью, где обычно меньше начальной размерности . Создаваемые при этом компоненты являются действительно ортогональными линейными комбинациями для максимизации общей дисперсии. При выполнении факторного анализа создаются новые, латентные переменные способом, сходным с используемым при анализе главных компонент. Предполагается, что создаваемые латентные переменные отражают некоторую общую сущность исходных переменных, вычисляемую при оценке вариабельности.

Предварительным этапом для выполнения анализа главных компонент и факторного анализа может быть стандартизация данных. Она выполняется для того, чтобы сделать переменные сравнимыми, и необходима если данные 1) измерены в разных шкалах (например, м, кг, сек), 2) имеют значимо отличающиеся величины. Трансформация выполняется следующим образом , где – среднее арифметическое, – стандартное отклонение. В R может быть выполнена функцией scale

Анализ главных компонент и факторный анализ в R

В R можно выполнить анализ главных компонент и факторный анализ используя функции базового пакета. Также в R имеется много пакетов для анализа многомерных данных, включающих функции для выполнения анализа главных компонент и факторного анализа. Возможности функций этих пакетов более широкие чем у базовых. В работе использованы базовые пакеты R 3.6.1 (RCoreTeam2019), а также пакеты dplyr (Wickham, François, etal. 2019), ggplot2 (Wickham 2016), tidyr (Wickham and Henry 2019), входящие в пакет tidyverse (Wickham, Averick, et al. 2019), factoextra (Kassambara and Mundt 2019), psych (Revelle 2019), ggrepel (Slowikowski 2019), knitr (Xie 2015). Работа выполнена в IDERStudio. Для выполнения анализа главных компонент и факторного анализа в работе использованы функции базового пакета stats: prcomp, factanal.

Использованные данные

В качестве данныхдля демонстрации применения анализа главных компонент используются результаты соревнований по легкоатлетическому десятиборью на Олимпийскихиграх2016года,опубликованныенасайте Википедии https://en.wikipedia.org/wiki/Athletics_at_the_2016_Summer_Olympics_–_Men%27s_decathlon. Десятиборье – спортивная дисциплина, в которой мужчины в течение двух дней соревнуются в 10 видах лёгкой атлетики: беге на 100 м, прыжках в длину, толкании ядра, прыжках в высоту, беге на 400 м, барьерном беге на 110 м, метании диска, прыжках с шестом, метании копья и беге на 1500 м. Результаты в отдельных видах “переводятся” в очки с помощью таблиц. Победителем становится спортсмен. набравший наибольшее количество очков. Используются данные о спортсменах, занявших места с 1 по 23.

Словарь для набора данных

название переменной	объяснение переменной
rank	занятое место
athlete	имя, фамилия спортсмена
overall_points	набранное количество очков
run100m	результат в беге на 100 м, сек
longjump	результат в прыжках в длину, м
shotput	результат в толкании ядра, м
highjump	результат в прыжках в высоту, м
run400m	результат в беге на 400 м, сек
hurdles110m	результат в беге на 110 м с барьерами, сек
discus	результат в метании диска, м
polevault	результат в прыжках в высоту с шестом, м
javelin	результат в метании копья, м
run1500m	результат в беге на 1500 м, сек

Листинг 1 показывает данные о результатах выступлений лучших шести спортсменов, некоторые показатели описательной статистики. Корреляционная матрица выведена с использованием функции пакета psych (рис.1).

Листинг 1

```
# используемые пакеты

library(tidyverse)
```

```
library(ggrepel)
library(factoextra)

decathlon <-read_csv("../data/decathlon.csv")
# Top 6
decathlon %>%
select(1:8) %>%
head() %>%
knitr::kable(caption ="Таб.1 6 лучших")

decathlon %>%
select(1, 2, 9:13) %>%
head() %>%
knitr::kable(caption ="Таб.2 6 лучших продолжение")
# показатели описательной статистики
psych::describe(decathlon[, 4:13], fast =TRUE) %>%
knitr::kable(digits =3, caption =»Таб.3 Некоторые показатели описательной статистики»)
# корреляционная матрица

psych::pairs.panels(decathlon[, 4:13], gap =0)
```

Оценка возможностей использования анализа главных компонент и факторного анализа

Предварительно рекомендуется определить критерий сферичности Бартлетта (Bartlett) и критерий Кайзера-Мейера-Олкина (Kaiser, Meyer, Olkin). Тесты выполняются для корреляционных матриц исходных данных.Критерий сферичности Бартлетта оценивает изучаемые данные на возможность их сжатия со значимыми результатами: нулевая гипотеза предполагает, что переменные ортогональные, не коррелируют. p-value менее уровня значимости предполагает возможность выполнения PCA.

Критерий Кайзера-Мейера-Олкина оценивает, насколько изучаемые данные подходят для факторного анализа, является мерой выборочной достаточности общей дисперсии. Значения от 0,5 до 0,7 считаются посредственными, от 0,7 до 0,8 - хорошими, от 0,8 и выше очень хорошими.

Выполнение тестов - в листинге 2.

Таб.1. 6 лучших

rank	athlete	overall_points	run100m	longjump	shotput	highjump	run400m
1	Ashton Eaton	8893	10.46	7.94	14.73	2.01	46.07
2	Kévin Mayer	8834	10.81	7.60	15.76	2.04	48.28
3	Damian Warner	8666	10.30	7.67	13.66	2.04	47.35
4	Kai Kazmirek	8580	10.78	7.69	14.20	2.10	46.75
5	Larbi Bourrada	8521	10.75	7.52	13.78	2.10	47.98
6	Leonel Suárez	8460	11.21	7.14	14.27	2.07	48.15

Таб.2. 6 лучших продолжение

rank	athlete	hurdles110m	discus	polevault	javelin	run1500m
1	Ashton Eaton	13.80	45.49	5.2	59.77	263.33
2	Kévin Mayer	14.02	46.78	5.4	65.04	265.49
3	Damian Warner	13.58	44.93	4.7	63.19	264.90
4	Kai Kazmirek	14.62	43.25	5.0	64.60	271.25
5	Larbi Bourrada	14.15	42.39	4.6	66.49	254.60
6	Leonel Suárez	14.48	47.07	4.9	72.32	268.32

Таб.3. Некоторые показатели описательной статистики

	vars	n	mean	sd	min	max	range	se
run100m	1	23	10.903	0.248	10.30	11.32	1.02	0.052
longjump	2	23	7.314	0.316	6.73	7.94	1.21	0.066
shotput	3	23	14.145	1.049	11.49	15.76	4.27	0.219
highjump	4	23	2.022	0.093	1.77	2.19	0.42	0.019
run400m	5	23	48.866	1.178	46.07	50.81	4.74	0.246
hurdles110m	6	23	14.617	0.633	13.58	16.51	2.93	0.132
discus	7	23	44.217	4.139	34.91	53.24	18.33	0.863
polevault	8	23	4.848	0.304	4.40	5.40	1.00	0.063
javelin	9	23	61.311	6.421	46.42	72.32	25.90	1.339
run1500m	10	23	274.128	10.043	254.60	293.07	38.47	2.094

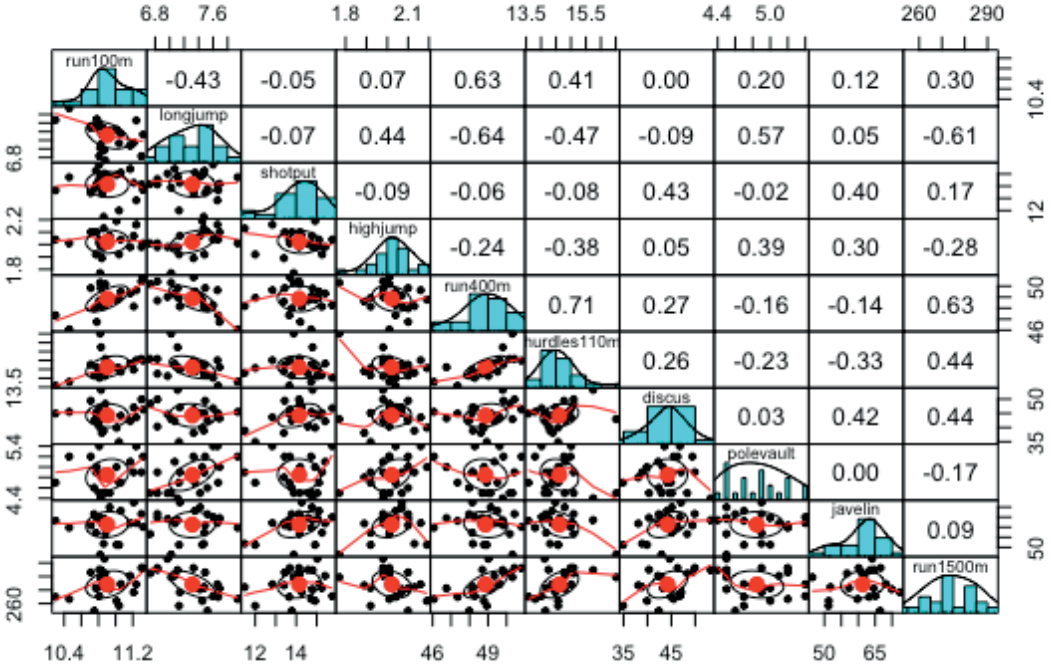

Рис.1. Корреляционная матрица исходных данных

Листинг 2

```
# критерий сферичности Бартлетта
psych::cortest.bartlett(cor(decathlon[4:13]))
$chisq
[1] 539.4296

$p.value
[1] 6.132506e-86

$df
[1] 45
# КритерийКайзера-Мейера-Олкина
psych::KMO(cor(decathlon[4:13]))
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = cor(decathlon[4:13]))
Overall MSA = 0.5
MSA for each item =
 run100m longjump shotput highjump run400m
 0.31 0.57 0.66 0.64 0.78
hurdles110m discus polevault javelin run1500m
 0.55 0.35 0.27 0.29 0.74
```

Полученные результаты допускают выполнение анализа главных компонент (p-value< 0,001) и невысоко оценивают возможности выполнения факторного анализа (MSA = 0.5).

Анализ главных компонент с использованием функции prcomp

Функция prcomp базового пакета stats может быть выполнена в виде

prcomp(x, center = TRUE, scale. = TRUE, …)

Аргументы: - изучаемая матрица изучаемых исходных данных, - логическое значение, указывающее на центрирование переменных, - логическое значение, указывающее на стандартизацию данных в исходных переменных

Объект prcomp из матрицы результатов выступлений, используя централизацию и стандартизацию данных.

Функция summary позволяет вывести важные для PCA результаты, отражающие такие параметры главных компонент, как

- Standard deviation – стандартное отклонение для вектора с центрированными и стандартизированными данными
- Proportion of Variance – доля дисперсии данных учитываемых в компоненте
- Cumulative Proportion – суммируемая доля дисперсии

Proportion of Variance и Cumulative Proportion рассчитываются из стандартного отклонения, и могут быть представлены графически.

Кроме того, созданный объект PCA (в нашем случае d_pca) имеет следующие значения:

- center, scale, sdev – это, соответственно, центральное значение, стандартизованное значение и стандартные отклонения каждой главной компоненты
- rotation – отражает связь между начальными переменными и главными компонентами, это своеобразные “веса” – значения, которыми старые переменные отдельных видов спорта входят в новую синтетическую переменноу
- x – значение каждой записи исходных данных представленное в главных компонентах

Создание объекта класса prcomp и получение данных о нём представлено в листинге 3

Листинг 3

```
# создание матрицы данных
deca<-as.matrix(decathlon[, 4:13])
rownames(deca) <-decathlon[, 1] %>%pull()

# создание объекта prcomp
d_pca<-prcomp(deca, center =TRUE, scale. =TRUE)
summary(d_pca)
Importance of components:
 PC1 PC2 PC3 PC4 PC5
Standard deviation 1.8716 1.4056 1.2427 0.96326 0.83646
Proportion of Variance 0.3503 0.1976 0.1544 0.09279 0.06997
Cumulative Proportion 0.3503 0.5479 0.7023 0.79511 0.86507
 PC6 PC7 PC8 PC9
Standard deviation 0.71391 0.6316 0.45152 0.40717
Proportion of Variance 0.05097 0.0399 0.02039 0.01658
Cumulative Proportion 0.91604 0.9559 0.97632 0.99290
 PC10
Standard deviation 0.2664
Proportion of Variance 0.0071
Cumulative Proportion 1.0000
# значения первых 3 столбцоы
d_pca$rotation[, 1:3]
PC1 PC2 PC3
run100m 0.28459328 0.063154786 -0.54900687
longjump -0.45158520 0.052142898 -0.15390500
shotput 0.04174919 0.487206307 0.30257333
highjump -0.26393997 0.240899018 -0.42606725
run400m 0.46931787 -0.009296962 -0.25809345
hurdles110m 0.41930402 -0.136306126 -0.12150115
discus 0.17349269 0.530025814 0.05742241
polevault -0.21372887 0.143497303 -0.56117213
javelin -0.05812655 0.580527220 0.04460781
run1500m 0.41041029 0.201727002 0.01564220
# значение первых шести строк первых трех главных компонентов
d_pca$x[, 1:3] %>%head()
PC1 PC2 PC3
1 -3.6219003 0.4030445 1.00750879
2 -1.7983677 1.7064321 -0.27769086
3 -2.8217561 -0.0314924 1.74943432
4 -2.0326847 0.4616188 -0.06429364
5 -2.1148261 -0.1399968 0.49722701
6 -0.1608041 1.5356724 -0.58361723
```

Выбор количества компонент

Правило **Kaiser-Guttman** предполагает отбирать компоненты с собственным значением (дисперсией) более 1 (единицы).Критерий «каменистой осыпи» графический метод, предложен Cattell в 1966 г., на графике необходимо определить место, где убывание собственных значений слева направо максимально замедляется.

В листинге 4 показан вывод собственных значений с использованием функции пакета factoextraget_eig, их можно также получить из результатов оценки объекта d_pca, функция screeplot из базового пакета покажет график “каменистая осыпь” (рис. 2). Графики на рис. 3,4,5 выполнены с использованием функций пакета ggplot2

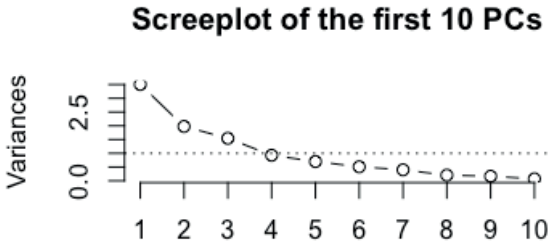

Рис.2. График ‘каменистая осыпь’ базовыми средствами

Листинг 4

```
# Extract eigenvalues/variances
get_eig(d_pca) %>%
knitr::kable(digits =3, caption =”Таб 4. Собственные значения / дисперсия”)
# собственные значения - вектор дисперсий для главных компонент
(pca.var<-d_pca$sdev**2)
 [1] 3.50304654 1.97579363 1.54436020 0.92786192
0.69965748
 [6] 0.50966106 0.39896652 0.20386854 0.16579040
0.07099371
# дисперсии главных компонент в процентах
(pca.var2 <-round(pca.var/sum(pca.var) *100, 1))
```

Таб 4. Собственные значения / дисперсия

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.503	35.030	35.030
Dim.2	1.976	19.758	54.788
Dim.3	1.544	15.444	70.232
Dim.4	0.928	9.279	79.511
Dim.5	0.700	6.997	86.507
Dim.6	0.510	5.097	91.604
Dim.7	0.399	3.990	95.593
Dim.8	0.204	2.039	97.632
Dim.9	0.166	1.658	99.290
Dim.10	0.071	0.710	100.000

```
[1] 35.0 19.8 15.4 9.3 7.0 5.1 4.0 2.0 1.7 0.7
# Visualize eigenvalues/variances
screeplot(d_pca, type =”l”, npcs =10, main =”Screeplot of the first 10 PCs”)
abline(h =1, lty =3)
tibble(
x =1:10,
y = pca.var
) %>%
ggplot(aes(x, y)) +
geom_point() +
geom_line() +
geom_hline(yintercept =1, lty =2) +
scale_x_continuous(breaks =1:10, labels =paste0(“PC”, as.character(1:10))) +
labs(x =””, y =”Eugenvalue”)
```

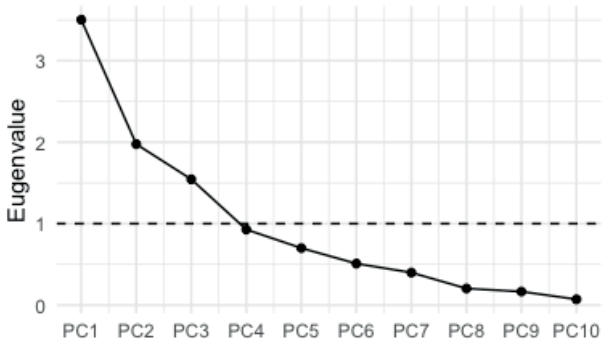

Рис.3. График ‘каменистая осыпь’ и правило Kaiser-Guttman

```
tibble(
x =1:10,
y = pca.var /sum(pca.var)
) %>%
ggplot(aes(x, y)) +
geom_col(fill =”grey70”) +
expand_limits(y =c(0, .4)) +
geom_text(aes(label =paste0(as.character(round(y *100, 1)), “%”)), vjust =-.2, size =3) +
scale_x_continuous(breaks =1:10, labels =paste0(“PC”, as.character(1:10))) +
scale_y_continuous(labels = scales::percent_format()) +
labs(x =””, y =”Proportion of Variance”)
```

```
tibble(
x =1:10,
y = pca.var /sum(pca.var),
cumy =cumsum(y)
) %>%
ggplot(aes(x, cumy)) +
geom_col(fill =”grey70”) +
geom_text(aes(label =paste0(as.character(round(cumy *100, 1)), “%”)), vjust =-.2, size =3) +
expand_limits(y =0) +
scale_x_continuous(breaks =1:10, labels =paste0(“PC”, as.character(1:10))) +
scale_y_continuous(labels = scales::percent_format()) +
labs(x =””, y =”Cumulative Proportion”)
```

Представление данных с использованием главных компонент

При выполнении анализа главных компонент можно построить ряд графиков с использованием первой и второй главных компонент, объясняющих почти 55 % вариабельности. (листинг 5). На графиках в соответствии со значениями первой и второй главных компонент показаны на рис.6 участники соревнований, на рис.7 - исходные переменные.При графической оценке главных компонент популярно использование PCAbiplot - двухмерной диаграммы, преставляющей связь между рядами и столбцами исходного набора данных (рис.8).В точечной диаграмме (рис.9) эллипсы выделяют области групповых переменных, точки большего диаметра - центры области эллипса. В качестве групповых признаков использована созданная в соответствии с занятым на соревнованиях местом бинарная переменная.

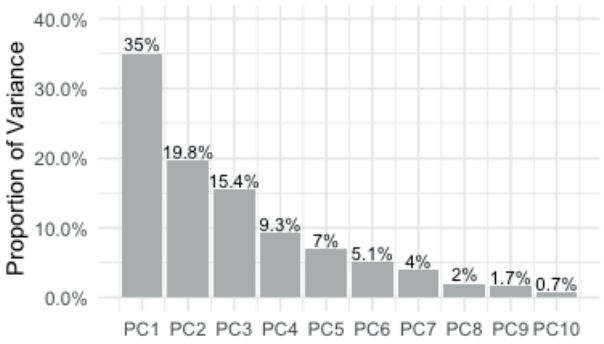

Рис.4. Доля объяснённой дисперсии для каждой компоненты

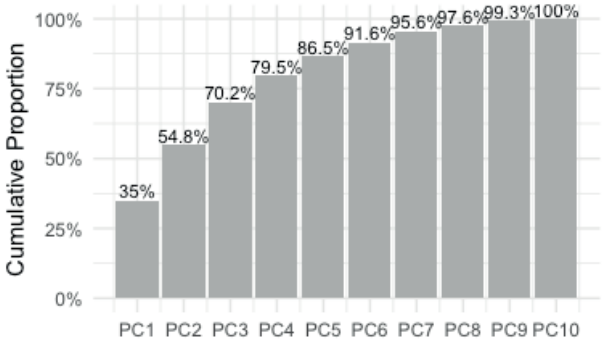

Рис.5. Суммированная доля объяснённой дисперсии

Листинг 5

```
tibble(
PC1 = d_pca$x[, 1],
PC2 = d_pca$x[, 2],
score = decathlon$overall_points,
rank = decathlon$rank
) %>%
ggplot(aes(PC1, PC2, label = rank, color = score)) +
geom_point(show.legend =FALSE) +
geom_text_repel(show.legend =FALSE, size =3) +
scale_color_gradient(low =”grey90”, high =”black”) +
labs(
```

```
title =”PC1 vs. PC2”,
x =paste0(“PC1 (“, as.character(pca.var2[1]), “ %)”),
y =paste0(“PC2 (“, as.character(pca.var2[2]), “ %)”)
)
```

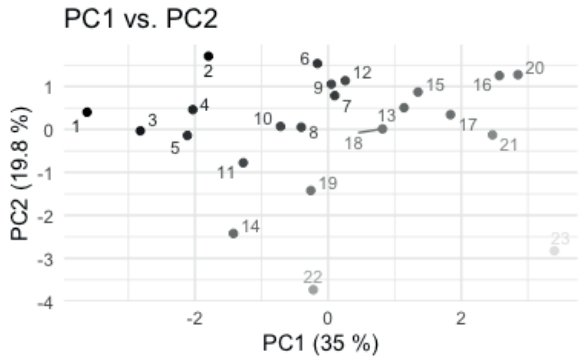

Рис.6. Первая и вторая главные компоненты

```
# Graphofvariables: defaultplot
# fviz_pca_var(d_pca)

tibble(
PC1 =d_pca$rotation[, 1],
PC2 =d_pca$rotation[, 2],
vid =rownames(d_pca$rotation)
) %>%
ggplot(aes(PC1, PC2, label =vid)) +
geom_point() +
geom_text_repel() +
geom_vline(xintercept =0, lty =3) +
geom_hline(yintercept =0, lty =3) +
labs(
title =””,
x =paste0(“PC1 (“, as.character(pca.var2[1]), “ %)”),
y =paste0(“PC2 (“, as.character(pca.var2[2]), “ %)”)
)
```

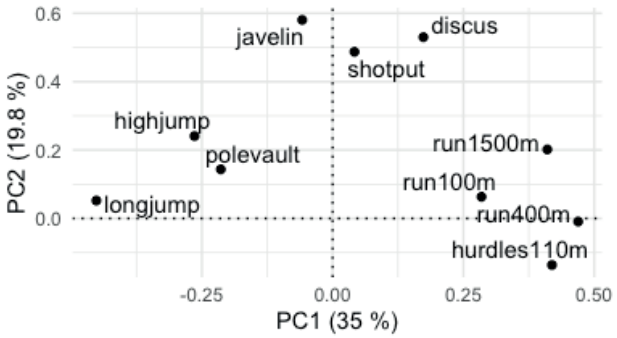

Рис.7. График переменных

```
fviz_pca_biplot(d_pca)
```

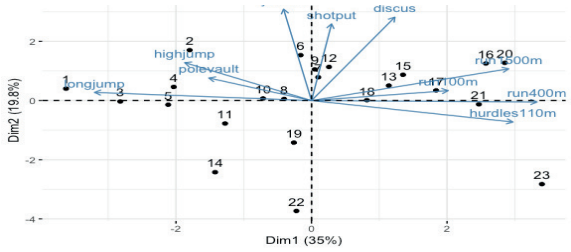

Рис.8. PCA biplot

```
# создание групповой переменной
decathlonl <-read_csv(“../data/decathlon.csv”) %>%
mutate(r_group =factor(ifelse(rank <=6, “Top 6”, “Others”)))

fviz_pca_ind(d_pca,
geom.ind =c(“point”, “text”),
pointshape =21,
pointsize =2,
fill.ind =factor(decathlonl$r_group),
col.ind =”black”,
palette =c(“grey60”, “black”),
addEllipses =TRUE,
col.var =”black”,
repel =TRUE,
legend.title =”Top 6”
) +
labs(title =”Индивидуальные значения - PCA”)
```

Факторный анализ с использованием функции factanal

Функция factanal базового пакета stats может быть выполнена в виде

factanal(x, factors, scores, rotation, …)

Аргументы: - изучаемая матрица, - количество создаваемых факторов, - вычисление значений для исходных данных в новых латентных переменных, - вращение, используемый при проведении факторного анализа.

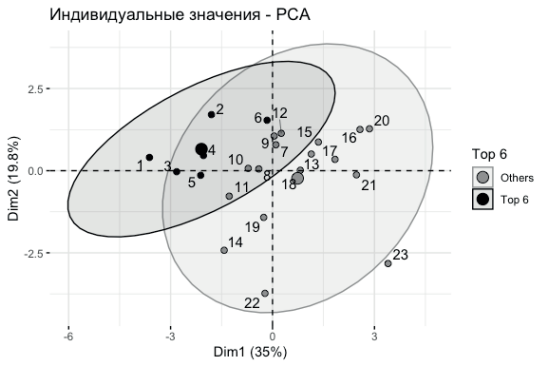

Рис.9. PCA

При выполнении функции оценивается тест на достаточность использования заданного количества факторов. Нулевая гипотеза предполагает достаточность избранного количеств факторов. Количество используемых факторов можно указать исходя из данных, полученных при проведении анализа главных компонент.

Для оценки при факторном анализе используются методы: Regression, Bartlett, Anderson-Rubin. В функции factanal доступные варианты для аргумента : regression, Bartlett, none.

Вращение – это математические методы трансформации матрицы для лучшей интерпретации. Выделяют ортогональное (orthogonal) и наклонное (oblique) вращение. При ортогональном вращении создаваемые переменные остаются нескорректированными, при наклонном - допускается их корреляция. Ортогональные методы включают varimax (используется наиболее ча-

сто), quartimax, equamax. Наклонные - directoblimin, quartimin, promax. В функции factanal доступные варианты для аргумента - none, varimax, promax.

Результат выполнения функции может быть выведен в виде

print(fa.object, digits = , cutoff = , sort = TRUE)

где - объект, созданный функцией factanal, - количество знаков после запятой, - отсекающее значение для показателей матрицы нагрузки (выводятся значения выше этой величины)

Показатели вывода включают:

- - уникальные значения вариабельности, не учитываемые в матрице нагрузки
- - показатели матрицы нагрузки

Пример факторного анализа в листинге 6.

Листинг 6

```
fa_d <-factanal(scale(decathlon[, 4:13]), factors =3, scores
="Bartlett", rotation ="varimax")
print(fa_d, digits =3, cutoff =.4, sort =TRUE)

Call:
factanal(x = scale(decathlon[, 4:13]), factors = 3, scores =
"Bartlett", rotation = "varimax")

Uniquenesses:
 run100m longjump shotput highjump run400m
 0.429 0.296 0.834 0.725 0.108
hurdles110m discus polevault javelin run1500m
 0.391 0.699 0.005 0.005 0.514

Loadings:
 Factor1 Factor2 Factor3
run100m 0.713
longjump -0.601 0.585
run400m 0.924
hurdles110m 0.687
run1500m 0.655
polevault 0.995
javelin 0.992
shotput 0.407
highjump 0.411
discus 0.447
Factor1 Factor2 Factor3
SS loadings 2.771 1.683 1.539
Proportion Var 0.277 0.168 0.154
Cumulative Var 0.277 0.445 0.599

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 24.6 on 18 degrees of freedom.
The p-value is 0.136
# первые 6 строк значений исходных данных в новых
переменных
fa_d$scores%>%head()
Factor1 Factor2 Factor3
[1,] -2.1498465 1.1891840 -0.4813694
[2,] -0.2448214 1.8504796 0.4446825
[3,] -1.8070200 -0.4274891 0.1857621
[4,] -1.3854539 0.5632827 0.3720406
[5,] -1.2122999 -0.7321417 0.7655130
[6,] -0.1264314 0.2757519 1.6986615
```

По результатам факторного анализа можно говорить о том, что первый фактор можно рассматривать как отражающий беговые возможности спортсменов, второй - прыжковые и третий - метательные. На рис.10 (листинг 7) эти показатели представлены в графическом виде: по оси абсцисс - значения для "бегового" фактора, по оси ординат - значения "прыжкового" фактора, размер круга зависит от показателей "метательного" фактора.

Листинг 7

```
tibble(
running = fa_d$scores[, 1],
jumping = fa_d$scores[, 2],
throwing = fa_d$scores[, 3],
rank =1:23
) %>%
ggplot(aes(running, jumping, label = rank)) +
geom_point(aes(size = throwing), shape =21) +
geom_text_repel(vjust =-1) +
scale_size(name ="Throw", breaks =c(-2, -1, 0, 1, 2), range =c(1,
6))
```

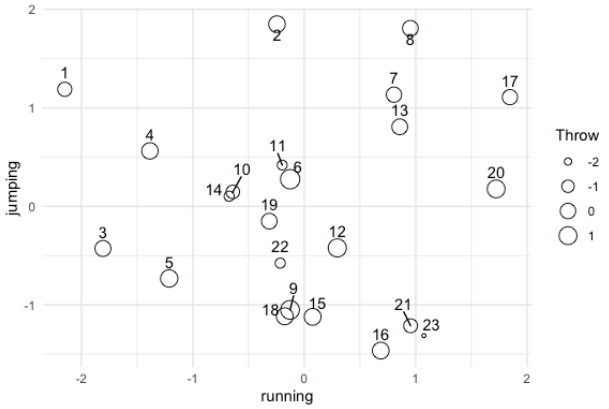

Рис.10. Результаты факторного анализа

Работа с R

Программная среда R является свободно распространяемым кросс-платформенным программным средством, использующимся для статистических вычислений и визуализации данных. Дистрибутивы R доступны на сайтах The Comprehensive R Archive Network, https://cran.r-project.org, Microsoft R Application Network, https://mran.microsoft.com/download. Удобным IDE (integrated development environment, интегрированная среда разработчика) для программы R является программа R Studio, свободно распространяемый дистрибутив может быть загружен на сайте R Studio IDE, https://www.rstudio.com/products/rstudio/. В наших более ранних публикациях (Егошин В.Л. 2018 и другие) мы уже касались вопросов применения программной среды R в биомедицинских исследованиях. Использованный в работе файл с набором данных и скрипт с кодом доступны на сайте https://github.com/valegoshin/Paper_Scripts.

Список литературы / References:

1. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(4): 433–59.
2. Bergman P, Grjibovski AM, Hagströmer M, Sallis JF, Sjöström M. The association between health enhancing physical activity and neighbourhood environment among Swedish adults - A population-based cross-sectional study. International Journal of Behavioral Nutrition and Physical Activity. 2009;6:8.
3. Everitt B, Hothorn T. An Introduction to Applied Multivariate Analysis with R. New York: Springer, 2011.
4. Costello AB, Osborne JW. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research and Evaluation. 2005;10(7).
5. Crawley MJ. The R Book. Chichester, West Sussex, United Kingdom: Wiley, 2013.
6. DiStefano C, Zhu M, Mîndrilă D. Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research and Evaluation. 2009;14(20).
7. Gaskin CJ, Happell B. On Exploratory Factor Analysis: A Review of Recent Evidence, an Assessment of Current Practice, and Recommendations for Future Use. Int J Nurs Stud. 2014:51(3):511–21.
8. Jolliffe IT. Principal component analysis. New York: Springer, 2002.
9. Kassambara A, Mundt F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses, 2019. https://CRAN.R-project.org/package=factoextra.
10. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer, 2013.
11. Lopes M. Dimensionality Reduction — Does Pca Really Improve Classification Outcome?https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-classification-outcome-6e9ba21f0a32.
12. Nistrup P. Principal Component Analysis (Pca) 101, Using R. 2019. https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff.
13. Osborne JW. What is rotating in exploratory factor analysis? Practical Assessment, Research and Evaluation. 2015;20(2):1–7.
14. R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
15. Revelle, W. Psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois: Northwestern University, 2019. https://CRAN.R-project.org/package=psych.
16. Slowikowski K. 2019. Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot 2'. https://CRAN.R-project.org/package=ggrepel.
17. Wickham H. Ggplot 2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. https://ggplot2.tidyverse.org.
18. Wickham H, Averick M, Bryan J, Chang W, D'Agostino L, Grolemund G et al. "Welcome to the tidyverse." Journal of Open Source Software. 2019;4(43):1686. https://doi.org/10.21105/joss.01686.
19. Wickham H, Romain F, Henry L, Müller K. 2019. Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.
20. Wickham H, Henry L. 2019. Tidyr: Tidy Messy Data. https://CRAN.R-project.org/package=tidyr.
21. Wright A. The current state and future of factor analysis in personality disorder research. Personality Disorders: Theory, Research, and Treatment. 2017;8(1):14–25.
22. Xie Y. Dynamic Documents with R and Knitr. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC, 2015. https://yihui.name/knitr/.
23. Егошин ВЛ, Саввина НВ, Иванов СВ, Гржибовский АМ. Основы работы в программной среде R при анализе биомедицинских данных. Экология человека. 2018;7:55–64.
    Egoshin VL, Savvina NV, Ivanov SV, Grjibovski AM. Basic Principles of Biomedical Data Analysis in R. Ekologiya cheloveka [Human Ecology]. 2018;7:55–64. [In Russian]
24. Кабаков РИ. R в действии: Анализ и визуализация данных в программе R/ пер. С англ. Полины а. Волковой. Москва: ДМК Пресс, 2014.
    Kabacoff RI. R in action: data analysis and graphics in R. Moscow: DMK Press, 2014. [In Russian]