Санкт-Петербургский Политехнический Университет им. Петра Великого

Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №2 по дисциплине “Теория принятия экономических решений”

**Сокращение размерности выборки на основе метода PCA**

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Павлова Людмила Владимировна

Санкт-Петербург

2020 г.

Оглавление

[1 Постановка задачи 3](#_Toc57050017)

[2 Численные эксперименты 3](#_Toc57050018)

[2.1 Отладка метода с использованием модельных данных 3](#_Toc57050019)

[2.1.1 Исходные данные 3](#_Toc57050020)

[2.1.2 PCA-данные (2 главные компоненты) 4](#_Toc57050021)

[2.1.3 PCA-данные (3 главные компоненты) 5](#_Toc57050022)

[2.2 Отладка метода с использованием данных из репозитория “iris” 6](#_Toc57050023)

[2.2.1 Исходные данные 6](#_Toc57050024)

[2.2.2 PCA-данные 6](#_Toc57050025)

[2.3 Отладка метода с использованием данных из репозитория “german” 7](#_Toc57050026)

[2.3.1 PCA-данные 8](#_Toc57050027)

[3 Реализация 10](#_Toc57050028)

[4 Выводы 10](#_Toc57050029)

[5 Литература 10](#_Toc57050030)

[6 Приложения 10](#_Toc57050031)

# 1 Постановка задачи

## Описание метода PCA

Метод главных компонент (Principal Component Analysis, PCA) – один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.

PCA направлен на поиск линейно некоррелированных ортогональных осей, которые также известны как главные компоненты в -мерном пространстве, для проецирования точек данных на эти компоненты. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы или к сингулярному разложению матрицы исходных данных.

Первая PC определяет наибольшую вариацию в данных, каждая последующая объясняет все более меньшую долю вариации.

Критерий информативности:

где – собственные числа ковариационной матрицы (дисперсии) исходных признаков, расположенные в порядке невозрастания.

## 1.2 Задача

Отладить метод главных компонент на модельных коррелированных данных с их визуализацией, как исходных, так и в терминах главных компонент, и вычислить выборочные матрицы ковариаций. Так же применить PCA к данным из репозитория – “german” и “iris”.

На всех наборах данных производим центрирование () для более точного разложения.

## 1.3 Расчётные формулы

* Выборочные средние:
* Выборочная матрица ковариаций:

# **2 Численные эксперименты**

## 2.1 Отладка метода с использованием модельных данных

Пусть имеется выборка объёма , размерности , из нормального распределения .

- вектор средних, - матрица ковариаций признаков. Пусть .

### 2.1.1 Исходные данные

* Выборочная матрица ковариаций:
* Первые 5 строк таблицы:

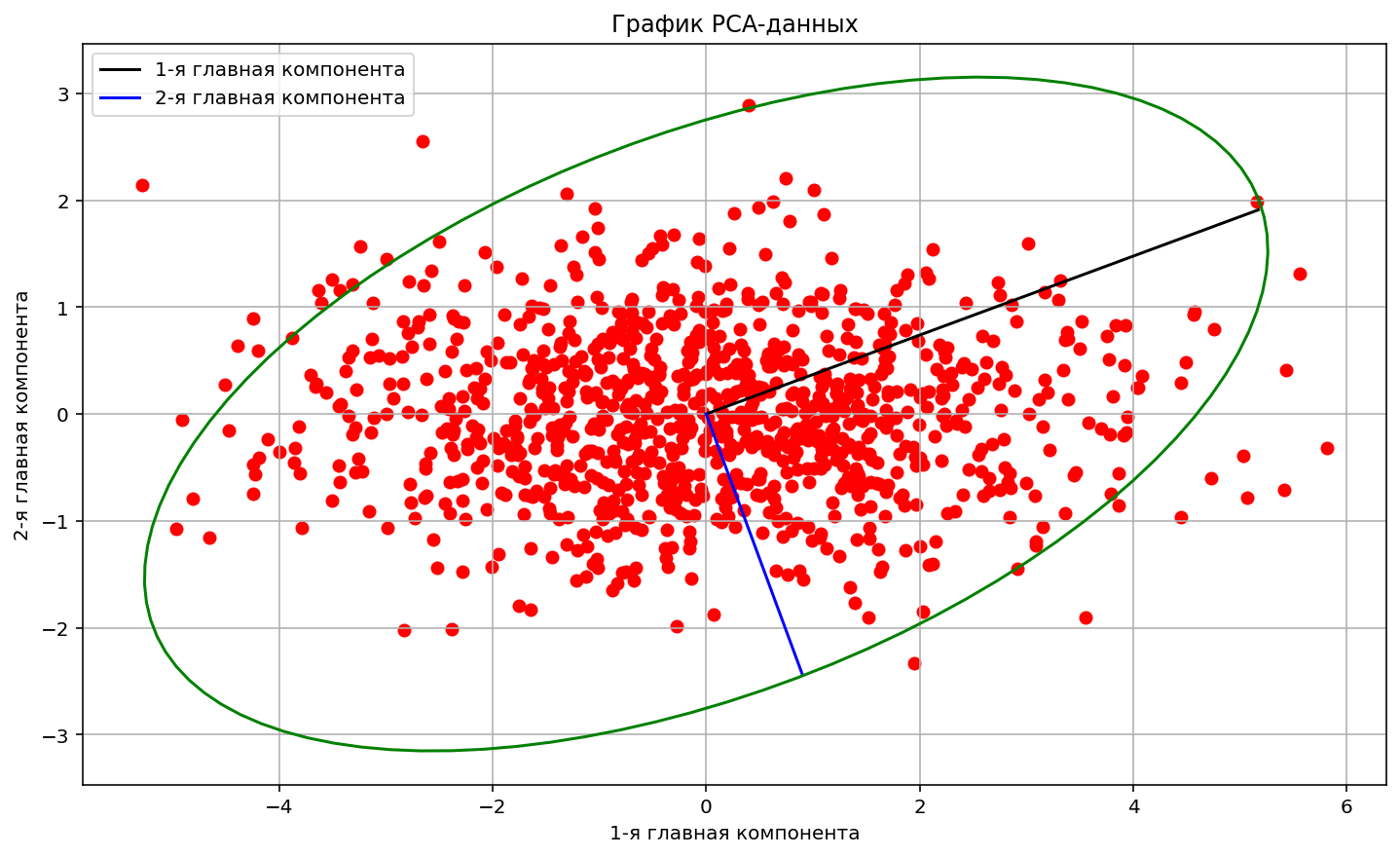
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Строка \ Признак** | **1-ый признак** | **2-ой признак** | **3-ий признак** | **4-ый признак** |
| **0** | -0.614131 | 0.200560 | 1.015251 | 1.829942 |
| **1** | 0.000016 | 1.242109 | 2.484203 | 3.726296 |
| **2** | 0.978808 | 1.941024 | 2.903241 | 3.865458 |
| **3** | -2.084482 | -0.220456 | 1.643571 | 3.507598 |
| **4** | 1.283508 | 2.644657 | 4.005806 | 5.366955 |

### 2.1.2 PCA-данные (2 главные компоненты)

* Выборочная матрица ковариаций:
* Первые 5 строк таблицы:

|  |  |  |
| --- | --- | --- |
| **Строка \ компонента** | **1-я главная компонента** | **2-я главная компонента** |
| **0** | -0.936433 | -0.054295 |
| **1** | 0.411214 | -0.228703 |
| **2** | 1.122467 | 0.311252 |
| **3** | -1.064819 | -1.400659 |
| **4** | 2.070431 | 0.065672 |

* График:



* Анализ полученных результатов:

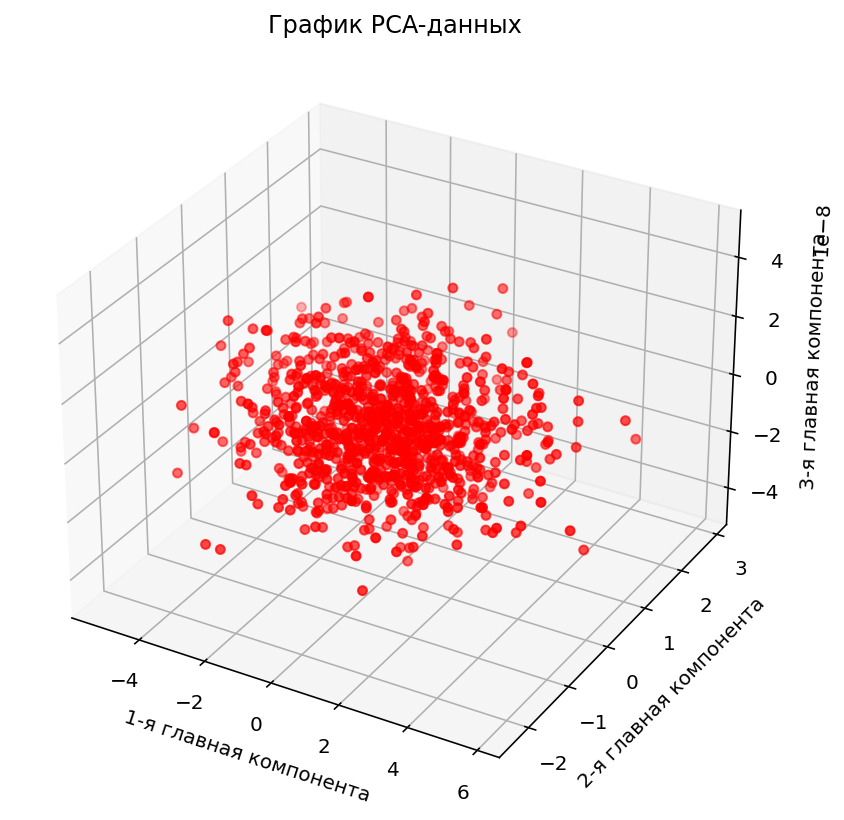
Как видим, отрезок большой полуоси эллипса объясняет основную часть дисперсии признаков, и в случае задачи сокращения размерности, спроецировав данные на этот отрезок мы потеряем наименьшее количество информации.

### 2.1.3 PCA-данные (3 главные компоненты)

* Выборочная матрица ковариаций для 3-х главных компонент:
* Выборочная матрица ковариаций для 4-х главных компонент:
* Первые 5 строк таблицы:

|  |  |  |  |
| --- | --- | --- | --- |
| **Строка \ компонента** | **1-я главная компонента** | **2-я главная компонента** | **3-я главная компонента** |
| **0** | -0.936433 | -0.054295 | 2.768957e-09 |
| **1** | 0.411214 | -0.228703 | -1.322754e-08 |
| **2** | 1.122467 | 0.311252 | 1.775848e-09 |
| **3** | -1.064819 | -1.400659 | 1.667514e-08 |
| **4** | 2.070431 | 0.065672 | 5.650787e-10 |

* График:



* Анализ полученных результатов:

Судя по PCA-графику и матрице выборочных корреляций, на разделимость данных метод PCA не сильно повлиял.

## 2.2 Отладка метода с использованием данных из репозитория “iris”

### 2.2.1 Исходные данные

* Выборочная матрица ковариаций:
* Первые 5 строк таблицы:

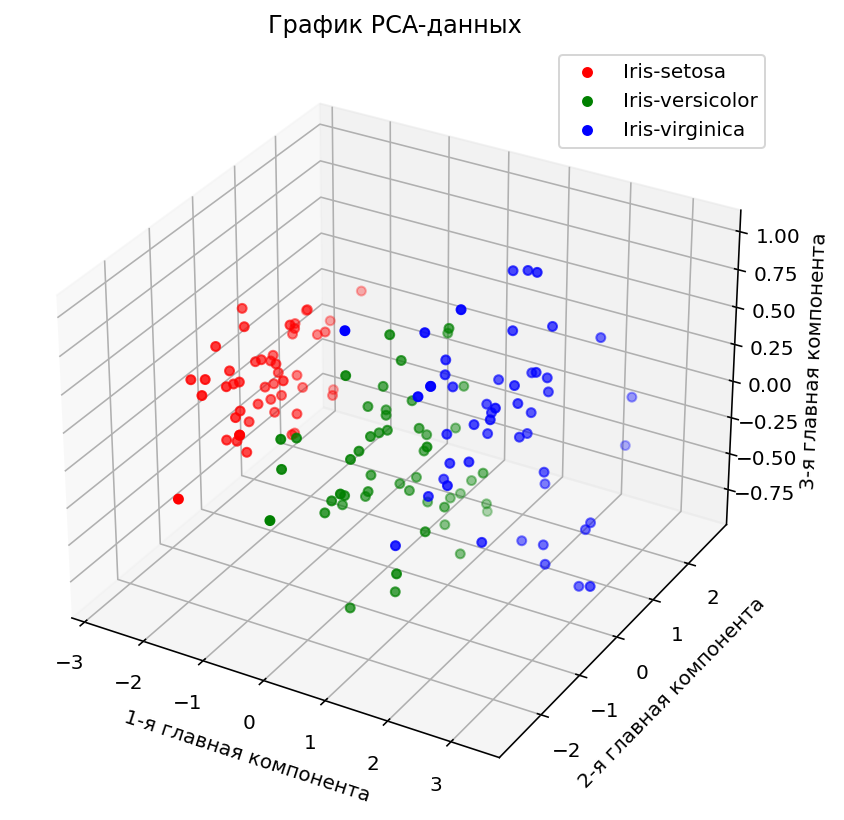
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Строка \ признак** | **Длина чашелистика** | **Ширина чашелистика** | **Длина лепестка** | **Ширина лепестка** | **Вид** |
| **0** | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |

### 2.2.2 PCA-данные

* Выборочная матрица ковариаций:
* Первые 5 строк таблицы:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Строка \ компонента** | **1-я главная компонента** | **2-я главная компонента** | **3-я главная компонента** | **Вид** |
| **0** | -2.264542 | 0.505704 | -0.121943 | Iris-setosa |
| **1** | -2.086426 | -0.655405 | -0.227251 | Iris-setosa |
| **2** | -2.367950 | -0.318477 | 0.051480 | Iris-setosa |
| **3** | -2.304197 | -0.575368 | 0.098860 | Iris-setosa |
| **4** | -2.388777 | 0.674767 | 0.021428 | Iris-setosa |

* График:



* Анализ полученных результатов:

Опять же судя по PCA-графику и матрице выборочных корреляций, на разделимость данных метод PCA не сильно повлиял. Данные как были слабо связаны, так и остались.

## 2.3 Отладка метода с использованием данных из репозитория “german”

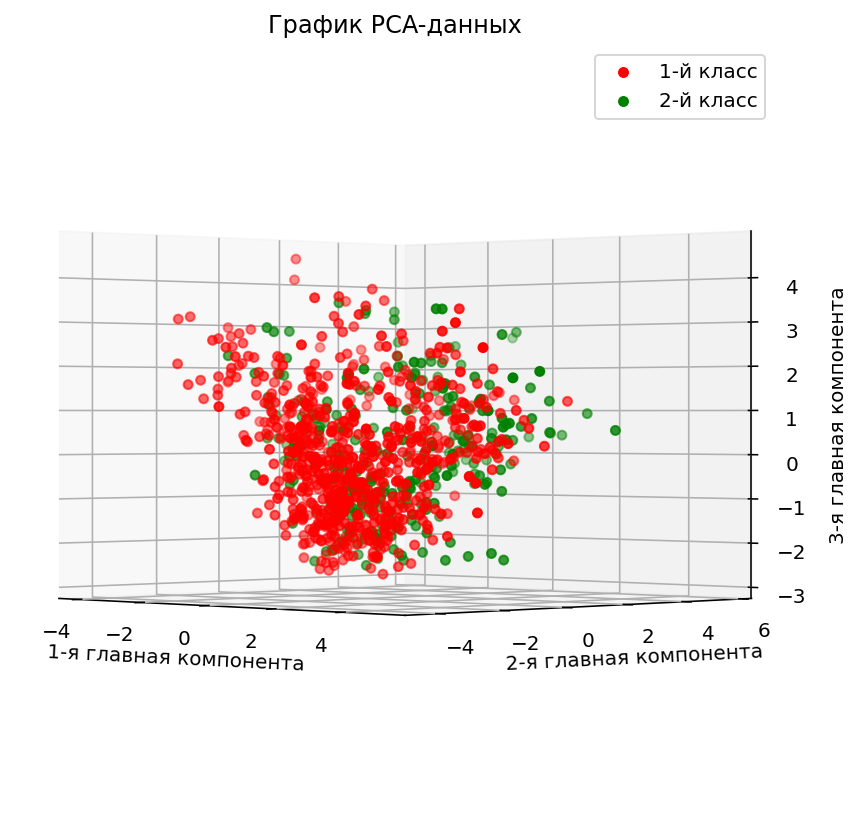
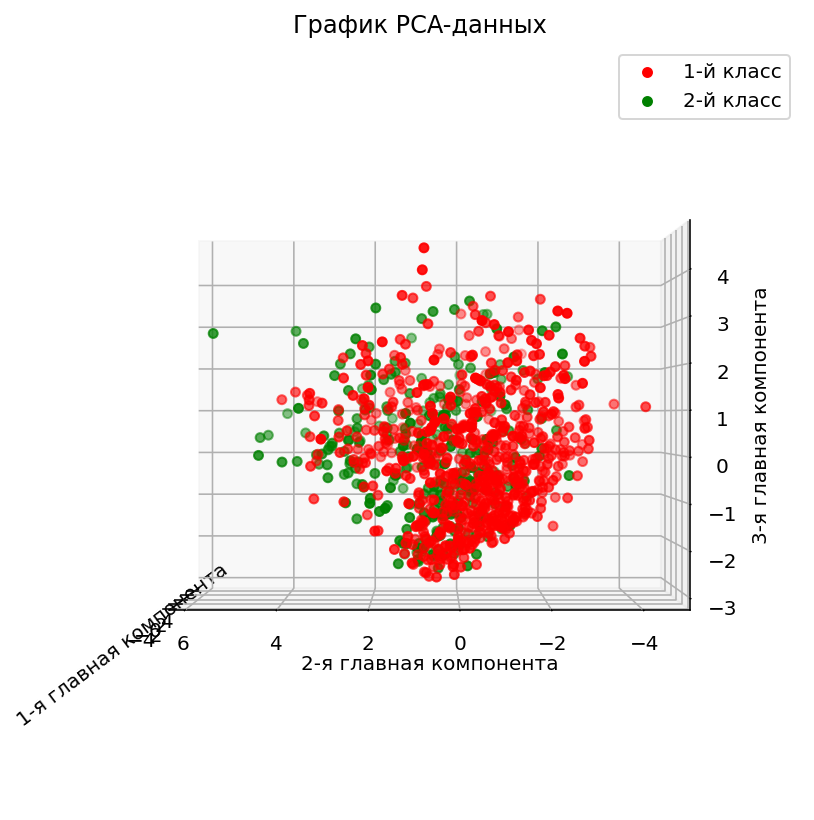
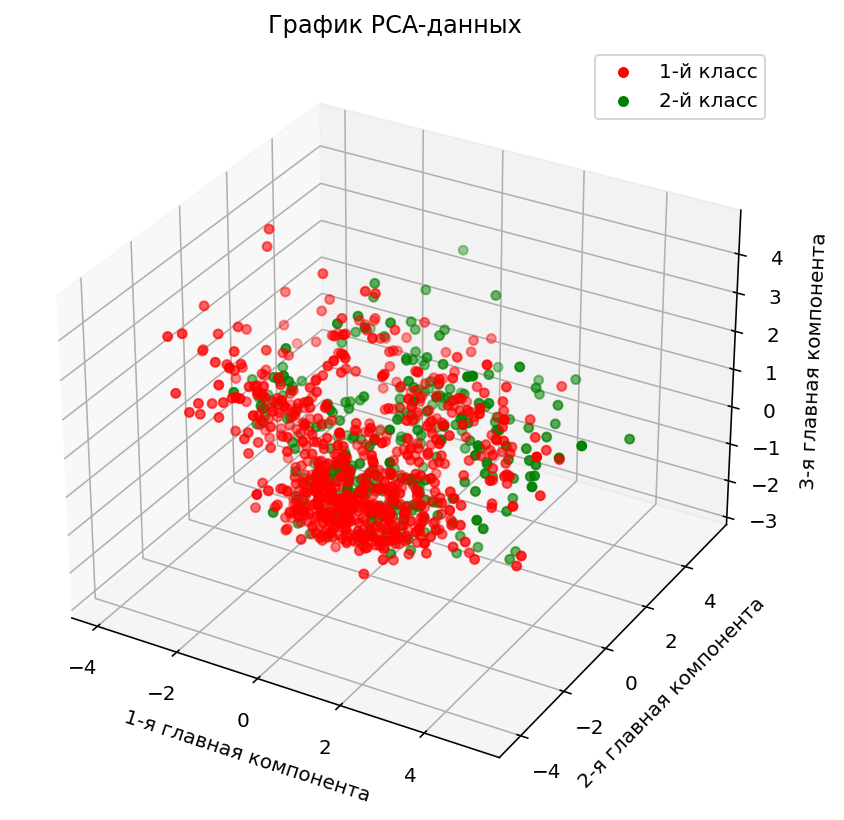
Рассмотрим данные из репозитория (файл *german.data-numeric*). Данные представляют 1000 объектов (1000 резюме различных людей), относящихся к одному из двух классов. Каждый объект характеризует набор из 24 численных признаков, являющихся характеристиками материального, финансового, семейного положения и его трудоустройства.

К первому классу относятся объектов, ко второму - .

### 2.3.1 PCA-данные

* Выборочная матрица ковариаций:
* Первые 5 строк таблицы:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Строка \ компонента** | **1-я главная компонента** | **2-я главная компонента** | **3-я главная компонента** | **Класс** |
| **0** | 0.013234 | -2.879658 | 0.778201 | 1 |
| **1** | 0.262257 | 0.976117 | -2.366935 | 2 |
| **2** | -2.186073 | -1.928887 | 1.823518 | 1 |
| **3** | 1.494757 | 1.342950 | 0.994195 | 1 |
| **4** | 1.234749 | -0.307253 | 1.973626 | 2 |

* График:
* Анализ полученных результатов:

Составим таблицу, отображающую % объясняемой информации при определённом количестве главных компонент, варьирующемся от 1 до 24:

|  |  |
| --- | --- |
| **Главных компонент** | **% объясняемой информации** |
| **1** | 10.492873 |
| **2** | 19.332331 |
| **3** | 27.055640 |
| **4** | 34.145547 |
| **5** | 40.945621 |
| **6** | 46.450270 |
| **7** | 51.522670 |
| **8** | 56.350965 |
| **9** | 61.030663 |
| **10** | 65.242266 |
| **11** | 69.185770 |
| **12** | 72.894039 |
| **13** | 76.479698 |
| **14** | 79.930596 |
| **15** | 83.193489 |
| **16** | 86.124819 |
| **17** | 88.818291 |
| **18** | 91.349070 |
| **19** | 93.701382 |
| **20** | 95.778902 |
| **21** | 97.200233 |
| **22** | 98.506423 |
| **23** | 99.338577 |
| **24** | 100.000000 |

Вместе 3 главные компоненты объясняют 27.05% информации:

1-я объясняет 10.492825260851063 % дисперсии

2-я объясняет 8.834699735550705 % дисперсии

3-я объясняет 7.723395116876455 % дисперсии

Мы сократили размерность с 24 до 3 признаков. Наблюдались как большие, так и малые значения в выборочной матрице корреляции исходной выборки. Следовательно, сложно сделать какие-то однозначные выводы. Наблюдается некоторая зависимость, но данные достаточно плохо разделимы.

Как видно из таблицы, чем больше главных компонент, тем медленнее приближается % информативности к значению 100. Если взять за порог информативности 95 %, то достаточно оставить 20 главных компонент.

# 3 Реализация

Была использована среда *IPython Notebook* (язык *Python 3.8.2*): модули *numpy* для работы с массивами и математических расчётов, *pandas* - для хранения данных в таблицах, *matplotlib.pyplot* – для построения графиков; *PCA* из модуля *sklearn.decomposition* для стандартизации матрицы наблюдений и *display* из модуля *IPython.display* для отображения pandas’ских таблиц в браузере.

# 4 Выводы

PCA – это проверенный и эффективный метод снижения размерности данных, особенно в случае когда эти признаки в данных имеют заметную линейную коррелированность. Чем сильнее линейная корреляция, тем меньше потеря информации.

# 5 Литература

[Основы работы с *numpy* (отдельная глава курса)](https://stepik.org/course/401)

Лекции по ТПЭР, “Метод PCA”, Павлова Л. В., 2020 г.

# 6 Приложения

[Код лабораторной](https://github.com/MeShootIn/economics-decision-theory/blob/main/lab_2/main.ipynb)