**Министерство науки и высшего образования Российской Федерации**
**Федеральное государственное бюджетное образовательное учреждение высшего образования**
**«Московский государственный технический университет имени Н.Э. Баумана**
**(национальный исследовательский университет)»**
**(МГТУ им. Н.Э. Баумана)**

ФАКУЛЬТЕТ *Робототехники и комплексной автоматизации*

КАФЕДРА *Системы автоматизированного проектирования (РК-6)*

# ОТЧЕТ О ВЫПОЛНЕНИИ ЛАБОРАТОРНОЙ РАБОТЫ

по дисциплине: «Вычислительная математика»

| | |
|---|---|
| Студент | Журавлев Николай |
| Группа | РК6-52Б |
| Тип задания | Лабораторная работа №4 |
| Тема лабораторной работы | Спектральное и сингулярное разложения |

Студент _____ **Журавлев Н. В.**
*подпись, дата* *фамилия, и.о.*

Преподаватель _____ **Першин А. Ю**
*подпись, дата* *фамилия, и.о.*

Москва, 2021 г.

# Оглавление

## Введение

Спектральное разложение (разложение на собственные числа и вектора) и сингулярное разложение, то есть обобщение первого на прямоугольные матрицы, играют настолько важную роль в прикладной линейной алгебре, что тяжело придумать область, где одновременно используются матрицы и не используются указанные разложения в том или ином контексте. В базовой части лабораторной работы мы рассмотрим метод главных компонент (англ. Principal Component Analysis, PCA), без преувеличения самый популярный метод для понижения размерности данных, основой которого является сингулярное разложение. В продвинутой части мы рассмотрим куда менее очевидное применение разложений, а именно одну из классических задач спектральной теории графов – задачу разделения графа на сильно связанные компоненты (кластеризация на графе).

## Цель выполнения лабораторной работы

Мы рассмотрим метод главных компонент (англ. Principal Component Analysis, PCA), самый популярный метод для понижения размерности данных, основой которого является сингулярное разложение.

## Задачи на лабораторную работу

### Базовая часть

1. Написать функцию *pca(A)*, принимающую на вход прямоугольную матрицу данных A и возвращающую список главных компонент и список соответствующих стандартных отклонений.
2. Скачать набор данных Breast Cancer Wisconsin Dataset: https://archrk6.bmstu.ru/index.php/f/85484391.

   – Указанный датасет хранит данные 569 пациентов с опухолью, которых обследовали на предмет наличия рака молочной железы. В каждом обследовании опухоль была проклассифицирована экспертами как доброкачественная (benign, 357 пациентов) или злокачественная (malignant, 212 пациентов) на основе детального исследования снимков и анализов. Дополнительно на основе снимков был автоматически выявлен и задокументирован ряд характеристик опухолей: радиус,

площадь, фрактальная размерность и так далее (всего 30 характеристик). Постановку диагноза можно автоматизировать, если удастся создать алгоритм, классифицирущий опухоли исключительно на основе этих автоматически получаемых характеристик. Указанный файл является таблицей, где отдельная строка соответствует отдельному пациенту. Первый элемент в строке обозначает ID пациента, второй элемент – диагноз (M = malignant, B = benign), и оставшиеся 30 элемент соответствуют характеристикам опухоли (их детальное описание находится в файле https://archrk6.bmstu.ru/index.php/f/854842).

3. Найти главные компоненты указанного набора данных, используя функцию *pca(A)*.

4. Вывести на экран стандартные отклонения, соответсвующие номерам главных компонент.

5. Продемонстрировать, что проекций на первые две главные компоненты достаточно для того, чтобы произвести сепарацию типов опухолей (доброкачественная и злокачественная) для подавляющего их большинства. Для этого необходимо вывести на экран проекции каждой из точек на экран, используя *scatter plot*.

## Выполненные задачи

## Базовая часть

### 1. Разработка функции pca(A)

Была разработана функция pca(A) для принятия на вход прямоугольной матрицы данных A и возвращения списка главных компонентов и списка соответствующих стандартных отклонений.

*Листинг 1. Реализация функции pca(A)*

```
def pca(A):
    A = np.array(A)
    K = A.T @ A

    lambd, q = np.linalg.eig(K)

    sig = np.sqrt(lambd)
```

```
    q_sort = q[:, np.flip(np.argsort(sig))]
    return q_sort, np.flip(np.sort(np.std(A, axis=0)))
```

**2. Был скачан и занесён в программу набор данных Breast Cancer Wisconsin Dataset.**

Указанный датасет хранит данные 569 пациентов с опухолью, которых обследовали на предмет наличия рака молочной железы. В каждом обследовании опухоль была проклассифицирована экспертами как доброкачественная (benign, 357 пациентов) или злокачественная (malignant, 212 пациентов) на основе детального исследования снимков и анализов. Дополнительно на основе снимков был автоматически выявлен и задокументирован ряд характеристик опухолей: радиус, площадь, фрактальная размерность и так далее (всего 30 характеристик). Постановку диагноза можно автоматизировать, если удастся создать алгоритм, классифицирущий опухоли исключительно на основе этих автоматически получаемых характеристик. Указанный файл является таблицей, где отдельная строка соответствует отдельному пациенту. Первый элемент в строке обозначает ID пациента, второй элемент – диагноз (M = malignant, B = benign), и оставшиеся 30 элемент соответствуют характеристикам опухоли

3. **Найти главные компоненты указанного набора данных.**

Из исходной матрицы можно получить матрицу центрированных данных A, по формуле:

$$A = \left(E - \frac{\bar{e}\bar{e}^T}{m}\right)$$

*На рис. 1 представлены данные 1 и 2 компоненты данной матрицы.*

*Листинг 2. Реализация центрирализации матрицы A*

```
def normalized_data_matrix(X):
    m = X.shape[0]
    return (np.eye(m) - 1 / m * np.ones((m, m))) @ X
```

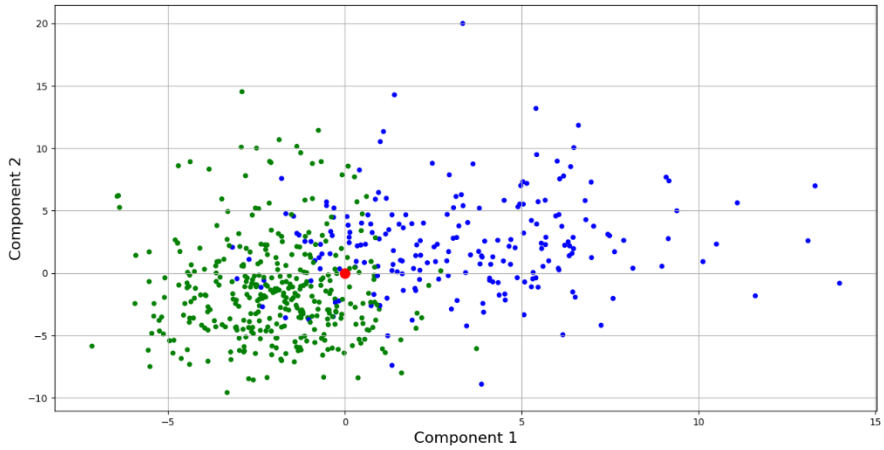


*Рисунок 1. Матрица центрированных данных A, где зеленый – malignant, синий – benign, а красная точка – центр координат.*

4. **Вывести на экран стандартные отклонения.**

Посчитаем стандартное отклонение, для каждого компонента, затем составим график зависимости стандартного отклонения от номеров главных компонент.

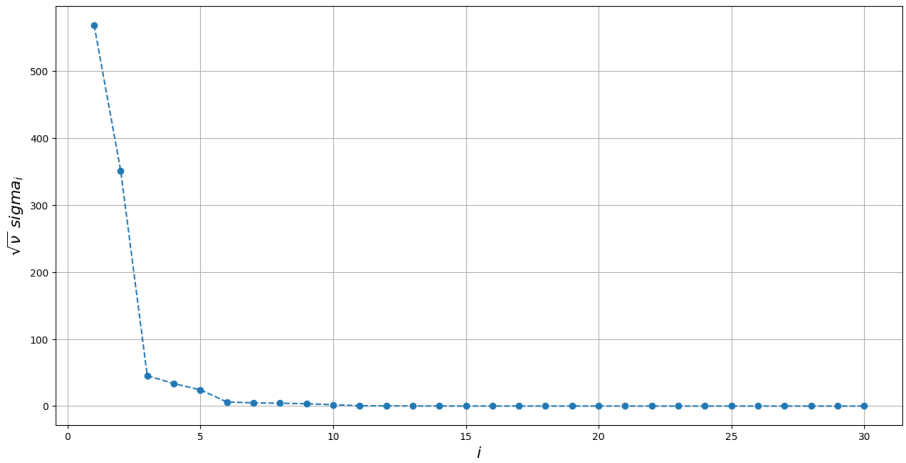


*Рисунок 2. Стандартные отклонения, соответствующие номерам главных компонент*

5. **Было продемонстрировано, что проекций на первые две главные компоненты достаточно для сепарации типов опухолей**

По рисунку 2 видно, что для первой и второй компоненты соответствуют наибольшие выборочные стандартные отклонения, а остальные главные компоненты можно отбросить, т.к. они не оказывают должного влияния на результат.

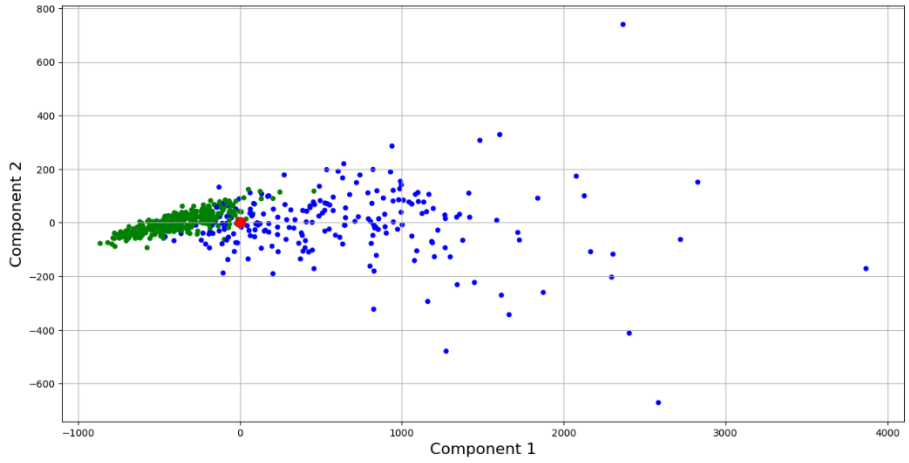


*Рисунок 3. Сепарация опухолей, где зеленый – malignant, синий – benign, а красная точка – центр координат*

*Листинг 3. Проекция матрицы A на 2 первые главные компоненты*

```
    A_K = normalized_X @ Q_T[:, :2]
    plot_data(A_K, diagnosis)
```

## Заключение

Была реализована функцию pca, с помощью которой были найдены главные компоненты и стандартное отклонение. Продемонстрировали, что проекции на две главные компоненты достаточно для того, чтобы произвести сепарацию типов опухолей.

## Список использованных источников


1. Першин А.Ю. Лекции по курсу «*Вычислительная математика*». Москва, 2018-2021. URL: https://archrk6.bmstu.ru/index.php/f/810046. (облачный сервис кафедры РК6).
2. Соколов, А.П., Першин, А.Ю «*Инструкция по выполнению лабораторных работ (общая)».* Москва: Соколов, А.П., Першин, А.Ю., 2018-2021. URL: *https://arch.rk6.bmstu.ru. (облачный сервис кафедры РК6).*