|  |  |
| --- | --- |
| **ё**Gerb-BMSTU_01 | **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 |  | Сергеева Диана Константиновна |
| Группа |  | РК6-56Б |
| Тип задания |  | лабораторная работа |
| Тема лабораторной работы |  | Спектральное и сингулярное разложения |

Студент **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_Сергеева Д.К.\_\_**

*подпись, дата фамилия, и.о.*

Преподаватель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_Соколов А.П. \_\_**

*подпись, дата фамилия, и.о.*

*Москва, 2021 г.*

Оглавление

[Задание на лабораторную работу 3](#_Toc90828647)

[Цель выполнения лабораторной работы 4](#_Toc90828648)

[Выполненные задачи 4](#_Toc90828649)

[1. Разработана функция для метода главных компонент для понижения размерности данных 5](#_Toc90828650)

[2. Построен график зависимости стандартного отклонения, соответствующего номерам главных компонент 5](#_Toc90828651)

[3. Построены проекции на главные компоненты 6](#_Toc90828652)

[Заключение 7](#_Toc90828653)

[Список использованных источников 7](#_Toc90828654)

# Задание на лабораторную работу

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.

Требуется (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/854843>.

–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, классифицирующий 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>).

1. Найти главные компоненты указанного набора данных, используя функцию pca(A).
2. Вывести на экран стандартные отклонения, соответствующие номерам главных компонент.
3. 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.

# Цель выполнения лабораторной работы

**Цель выполнения лабораторной работы** – изучение метода главных компонент для понижения размерности данных. Изучение стандартного отклонения.

# Выполненные задачи

1. Разработана функция для метода главных компонент для понижения размерности данных.
2. Построен график зависимости стандартного отклонения, соответствующего номерам главных компонент.
3. Построены проекции на главные компоненты

# Разработана функция для метода главных компонент для понижения размерности данных

Реализована функция *psa(A)*, на вход которой подаётся матрица данных о пациентах .

Поиск главных компонент и стандартных отклонений осуществляется на основе матрицы центрированных данных. Нахождение такой матрицы реализовано с помощью функции *get\_normalized\_data\_matrix(A)*. Из курса лекций известно, что такая матрица вычисляется следующим образом:

*,*

где – матрица единиц; – матрица, у которой все элементы равны 0, а на главной диагонали стоят 1; – изначальная матрица; – количество столбцов в матрице .

Для нахождения главных компонент и стандартных отклонений, воспользуемся функцией *numpy.linalg.eig* из библиотеки *numpy.* Функция возвращает собственные числа и собственные вектора матрицы.

По теореме о главных компонентах из курса лекций: главными компонентами матрицы центрированных данных являются её сингулярные вектора. Стандартное отклонение -ой главной компоненты при этом равняется:

, (1)

где – -ое сингулярное число (, - количество столбцов в матрице).

То есть собственные вектора и есть искомые главные компоненты. Стандартные отклонения рассчитываются по формуле (1), причем сингулярное число вычисляется как корень из соответствующего собственного числа, вычисленного с помощью функции *numpy.linalg.eig*.

# Построен график зависимости стандартного отклонения, соответствующего номерам главных компонент

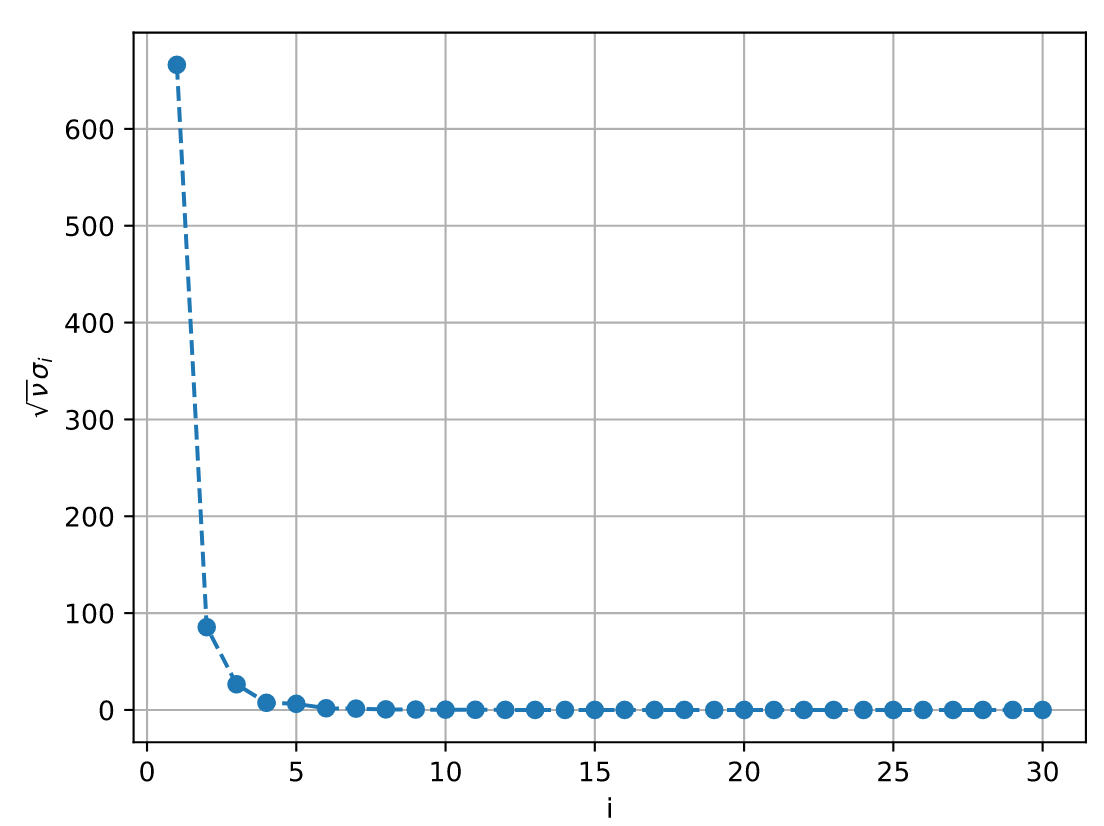


Рис. 1 – Зависимость стандартного отклонения от номера главной компоненты

По рис. 1 видно, что самой первой главной компоненте соответствует наибольшее стандартное отклонение. С увеличением номера главной компоненты разброс значений стандартных отклонений уменьшается.

# Построены проекции на главные компоненты

Для построения требуется произвести масштабирование показаний. То есть все значения необходимо пересчитать по формуле: , где – выборочное среднее значение по данному признаку, – выборочное стандартное отклонение.

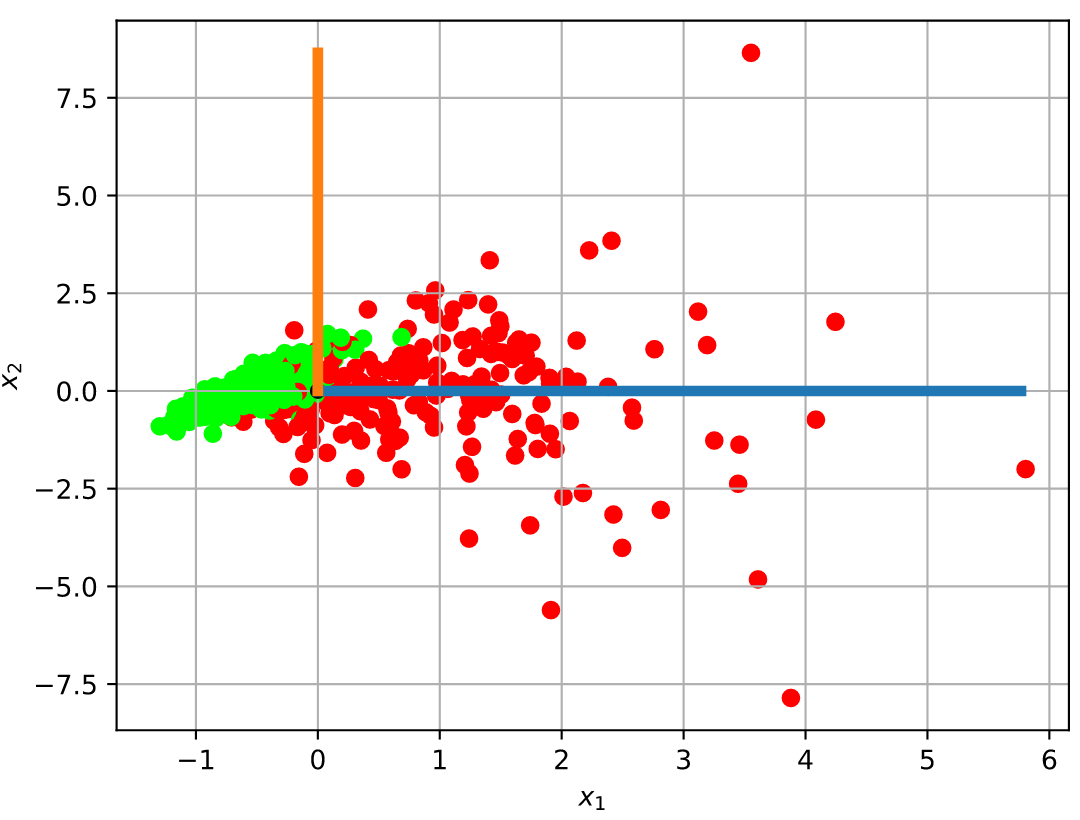


Рис. 2 – Проекция каждой точки на первые 2 главные компоненты

На рис. 2 зелёные точки соответствую доброкачественным опухолям, красные – злокачественным опухолям.

Вдоль двух координат достаточно большой разброс как вдоль первой главной компоненты, так и вдоль 2 главной компоненты.

# Заключение

По результатам лабораторной работы мы научились применять метод главных компонент для понижения размерности данных. Научились строить график зависимости стандартного отклонения от номера главной компоненты и график проекции точек на главные компоненты.

# Список использованных источников

1. Першин А.Ю. Лекции по курсу «Вычислительная математика. Москва, 2018-2021, С. 140.