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	Морозов Александр Юрьевич
Группа	РК6-54Б
Тип задания	Лабораторная работа № 4
Тема лабораторной работы	Спектральное и сингулярное разложения

Студент _____ **Морозов А.Ю.**
подпись, дата *фамилия, и.о.*

Преподаватель _____ **Соколов А. П.**
подпись, дата *фамилия, и.о.*

Москва, 2021 г.

Оглавление

Задание на лабораторную работу 3
Цель выполнения лабораторной работы 5
Выполненные задачи 5
1. Базовая часть 5
2. Продвинутая часть 8
Заключение 8
Список использованных источников 9

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. 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).

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, соответствующие 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.

Продвинутая часть

1. Построить лапласианы (матрицы Кирхгофа) L для трех графов:
   - полный граф $G1$, имеющий 10 узлов;
   - граф $G2$, изображенный на рисунке 1;
   - граф $G3$, матрица смежности которого хранится в файле https://archrk6.bmstu.ru/index.php/f/854844,

   где лапласианом графа называется матрица $L = D - A$, где A – матрица смежности и D – матрица, на главной диагонали которой расположены степени вершин графа, а остальные элементы равны нулю.
2. Доказать, что лапласиан неориентированного невзвешенного графа с n вершинами является положительно полуопределенной матрицей, имеющей n неотрицательных собственных чисел, одно из которых равно нулю.
3. Найти спектр каждого из указанных графов, т.е. найти собственные числа и вектора их лапласианов. Какие особенности спектра каждого из графов вы можете выделить? Какова их связь с количеством кластеров?
4. Найти количество кластеров в графе $G3$, используя второй собственный вектор лапласиана. Для демонстрации кластеров выведите на графике исходную матрицу смежности и ее отсортированную версию.

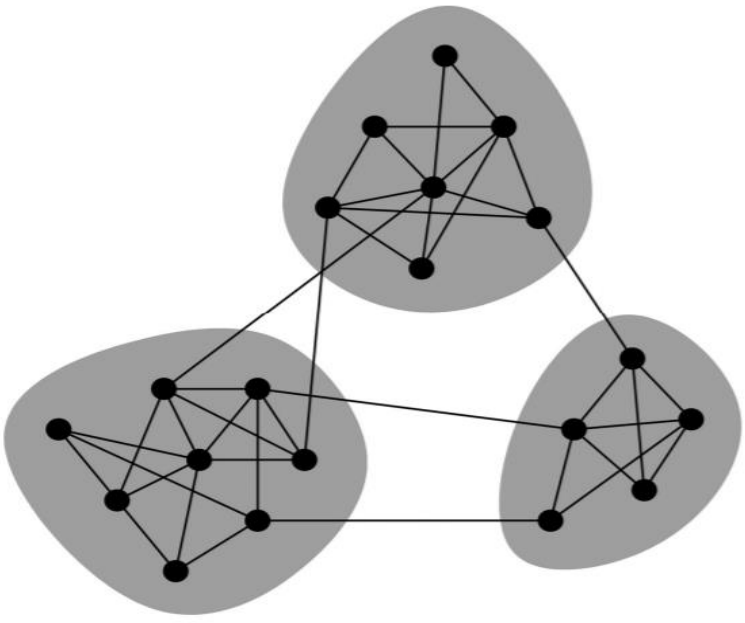

Рис. 1: Граф, содержащий три кластера

Цель выполнения лабораторной работы

Цель выполнения лабораторной работы – изучить и применить на практике один из наиболее удачных методов понижения размерности – метод главных компонент. Написать программу для проведения этого исследования и проинтерпретировать полученные в результате работы программы графики функций и прочие визуализации.

Выполненные задачи

Базовая часть

1. Разработка функции, принимающей на вход прямоугольную матрицу данных A и возвращающей список главных компонент и список соответствующих стандартных отклонений.
2. Импорт и обработка данного датасета.
3. Вывод на экран стандартных отклонений, соответствующих номерам главных компонент.
4. Вывод на экран двух первых главных компонент и всех точек. Доказательство того, что проекции на первые две главные компоненты достаточно для разделения опухолей по типам.

Продвинутая часть

Не выполнял

1. Базовая часть

1.1.

Матрица данных $X \in \mathbb{R}^{m \times n}$ имеет вид:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{12} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

где строки матрицы – измерения данных, а столбцы – показатели данных.

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

$$A = \left(E - \frac{1}{m} e e^T\right) X$$

где E – единичная матрица, а e – единичный вектор (ee^T – соответственно матрица единиц).

Функция, решающая эту задачу, представлена в листинге 1.

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

Листинг 1: Функция, вычисляющая матрицу центрированных данных

На вход функции $pca(A)$ передаётся матрица центрированных данных. Прежде всего необходимо найти матрицу Грама по формуле:

$$K = A^T A$$

Далее при помощи функции $linalg.eig()$ были найдены собственные числа и вектора матрицы Грама.

После получения собственных чисел можно вычислить сингулярные числа по формуле:

$$\sigma_i = \sqrt{\lambda_i},$$

где λ – ненулевые собственные числа матрицы Грама.

Последним шагом будет вычисление стандартных отклонений по формуле:

$$s = \sqrt{\nu}\sigma,$$

Где $\nu = \frac{1}{m-1}$

Готовая функция $pca(A)$ представлена в листинге 2.

```python
def pca(A):
    K = A.T @ A
    val, vect = np.linalg.eig(K)
    sing_val = np.sqrt(val)
    return vect, np.sqrt(1. / (A.shape[0] - 1)) * sing_val
```

Листинг 2: Функция, реализующая метод главных компонент.

1.2.

Датасет Breast Cancer Wisconsin Dataset хранит данные 569 пациентов с опухолью, которых обследовали на предмет наличия рака молочной железы. В каждом обследовании опухоль была проклассифицирована экспертами как доброкачественная (benign, 357пациентов) 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). Указанный датасет 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 элементов соответствуют характеристикам опухоли.

Для чтения данных из файла была использована функция $read\_scv()$ из библиотеки $pandas$.

Первые 2 столбца были отброшены. Они соответствуют ID пациента и его диагнозу. Соответственно, прямоугольная матрица X будет состоять из 569 строк и 30 столбцов.

1.3

На рисунке 2 представлена зависимость выборочного стандартных отклонений от номеров главных компонент.

Несложно заметить, что у первых двух главных компонент значения стандартных отклонений значительно больше, чем у всех остальных. Следовательно, можно сделать вывод о том, что проекции точек на первые две главные компоненты достаточно для разделения опухолей по типам.

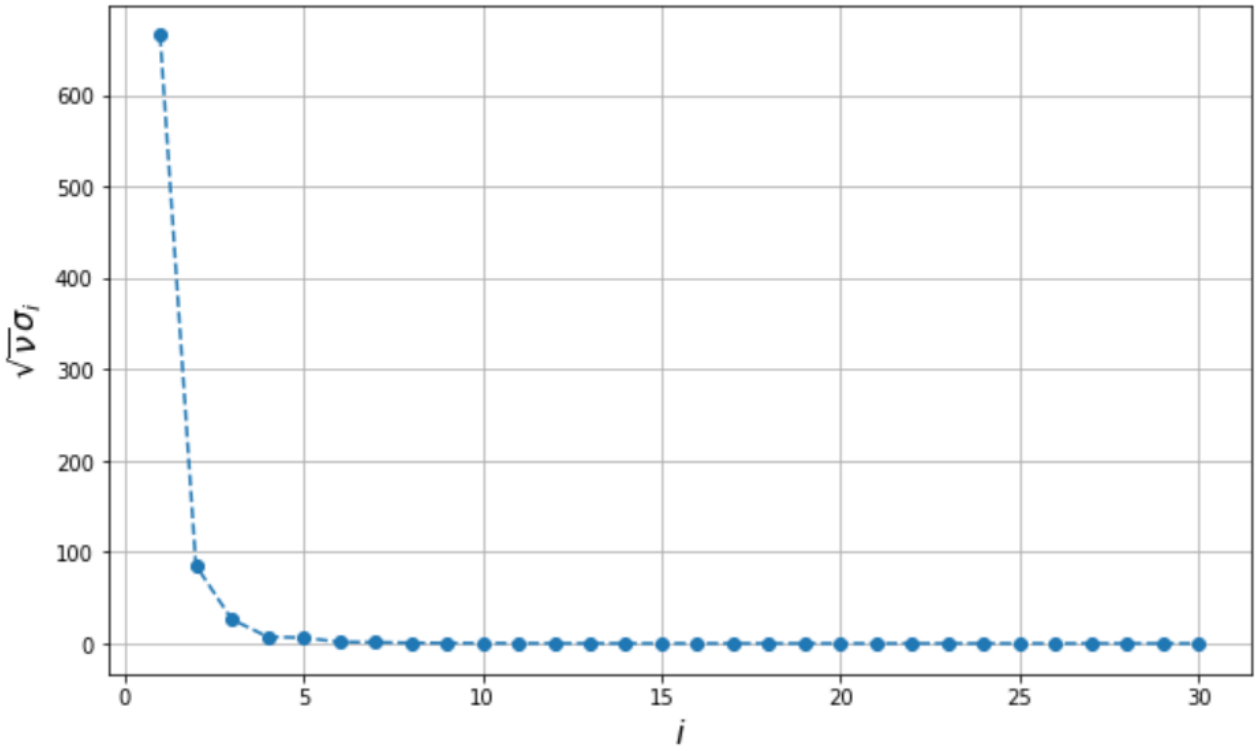

Рис. 2: Зависимость выборочного стандартных отклонений от номеров главных компонент

1.4

Проверим справедливость утверждения о достаточности проекции точек на первые 2 главные компоненты для разделения опухолей по типам. Для этого достаточно посмотреть на рисунок 3. На нём красным цветом обозначены злокачественные опухоли ($malignant$), зелёным – доброкачественные ($benign$), синим цветом показана первая главная компонента,

оранжевым – вторая. Данные главные компоненты формируют ортонормированный базис. Невооружённым взглядом видно, что разделение опухолей по типам проведено с достаточной точностью.

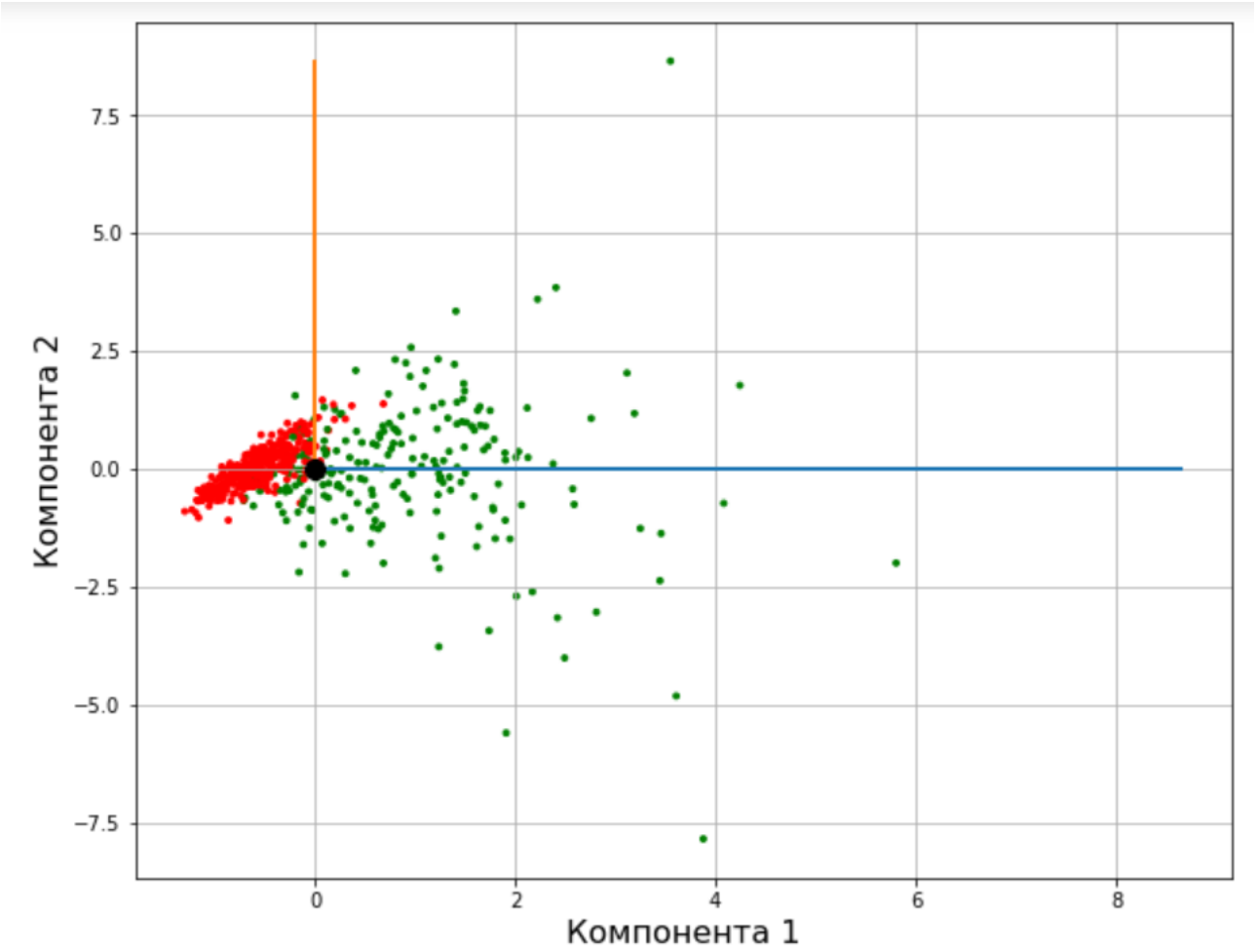

Рис. 3: Проекция точек на первые две главные компоненты.

2. Продвинутая часть

Не выполнял

Заключение

В ходе выполнения лабораторной работы:

1. Был рассмотрен и реализован на языке python один из самых ходовых методов понижения размерности – метод главных компонент
2. В качестве датасета для проведения исследования при помощи метода главных компонент были использованы данные о пациентах с опухолями.

Соответственно, было проведено разделение на доброкачественные и злокачественные опухоли.

3. В ходе работы выяснилось, что для качественного разделения достаточно двух первых главных компонент. Следовательно, размерность была понижена с 30 измерений до 2. Это показывает высокую эффективность метода.
4. Для того, чтобы сделать выводы, представленные в пунктах выше, были представлены некоторые визуализации данных.

Список использованных источников

1. Першин А.Ю. Лекции по курсу «Вычислительная математика». Москва, 2018-2021.
2. Першин А.Ю., Соколов А.П. Вычислительная математика, лабораторные работы (учебное пособие), МГТУ им. Баумана, Москва, 2018-2021.