



Министерство науки и высшего образования Российской Федерации
федеральное государственное бюджетное образовательное учреждение
высшего профессионального 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 Вадимович
Группа:	РК6-52Б
Тип задания:	Лабораторная работа
Тема:	Спектральное и сингулярное разложение

Студент

подпись, дата

Журавлев Н.В.

Фамилия, И.О.

Преподаватель

подпись, дата

Першин А.Ю.

Фамилия, И.О.

Москва, 2021

Содержание

Спектральное и сингулярное разложения	3
1 Введение	3
2 Цель выполнения лабораторной работы	3
3 Задачи на лабораторную работу	3
4 Выполненные задачи	4
5 Разработка функции $\text{rsa}(A)$	4
6 Занесение в программу набор данных Breast Cancer Wisconsin Dataset .	5
7 Нахождение главных компонент указанного набора данных.	5
8 Вывод на экран стандартные отклонения.	6
9 Демонстрация, того что проекция на первые две главные компоненты достаточно для сепарации типов опухолей.	7
10 Заключение	8

Спектральное и сингулярное разложения

1 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).

2 Цель выполнения лабораторной работы

Цель выполнения лабораторной работы – рассмотреть метод главных компонент (англ. Principal Component Analysis, PCA), самый популярный метод для понижения размерности данных, основой которого является сингулярное разложение.

3 Задачи на лабораторную работу

Базовая часть:

1. Написать функцию $pca(A)$, принимающую на вход прямоугольную матрицу данных A и возвращающая 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, 4 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>).

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`.

4 Выполненные задачи

1. Написана функция `pca(A)`, принимающая на вход прямоугольную матрицу данных `A` и возвращающую список главных компонент и список соответствующих стандартных отклонений.
2. Скачан необходимый набор данных.
3. Найдены главные компоненты указанного набора данных, используя функцию `pca(A)`.
4. Выведены на экран стандартные отклонения, соответствующие номерам главных компонент.
5. Продемонстрировано, что проекций на первые две главные компоненты достаточно для того, чтобы произвести сепарацию типов опухолей (доброкачественная и злокачественная) для подавляющего их большинства.

5 Разработка функции `pca(A)`

Была разработана функция `pca(A)`, принимающая на вход прямоугольную матрицу данных `A`, и возвращающая список главных компонент и список соответствующих стандартных отклонений (листинг 1).

Листинг 1. Реализация функции `pca(A)`

```
1 def pca(A):
2     A = np.array(A)
3     K = A.T @ A
4
5     lambd, q = np.linalg.eig(K)
6
7     sig = np.sqrt(lambd)
```

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

6 Занесение в программу набор данных Breast Cancer Wisconsin Dataset

Набор данных хранится в файле с названием wdbc.data. После прочтения данных из файла, функция делает 2 массива: в первом диагнозы, а во втором отбросим первые 2 столбца (id пациента и диагноз). Реализация - листинг (2).

Листинг 2. Реализация функции composite_trapezoid

```
1 def get_some_data():
2     data = pd.read_csv('wdbc.data', delimiter=',', header=None)
3     diagnosis = np.array(data[:, 1:2].T[0])
4     X = np.array(data[:, 2:])
5     result = X.astype(np.float32, copy=True)
6     return result, diagnosis
```

7 Нахождение главных компонент указанного набора данных.

Из исходной матрицы делается матрица центрированных данных A (рисунок 1), по формуле:

$$A = (E - \frac{1}{m}ee^T)X,$$

где E единичная матрица, ee^T - матрица единиц. Вывод матрицы и подсчёт главных компонент через функцию pca представлен на листинге (3).

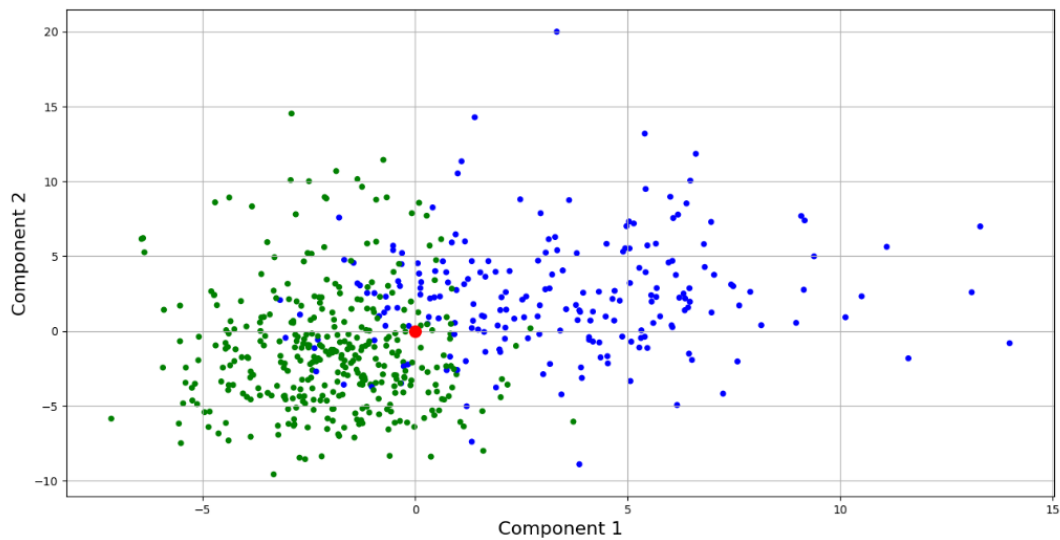


Рис. 1. Матрица центрированных данных, где зеленый – malignant, синий – benign, а красная точка – центр координат.

Листинг 3. Вывод матрицы центрированных данных A и подсчёт главных компонент

```

1 normalized_X = normalized_data_matrix(X)
2 plot_data(normalized_X, diagnosis)
3 Q_T, std = pca(normalized_X)

```

8 Вывод на экран стандартные отклонения.

Вычисление стандартного отклонения, для каждого компонента происходит по формуле (1) (представлено на рисунке (2))

$$\sqrt{\nu}\sigma_j, \quad (1)$$

где σ_j - j -ое сингулярное число

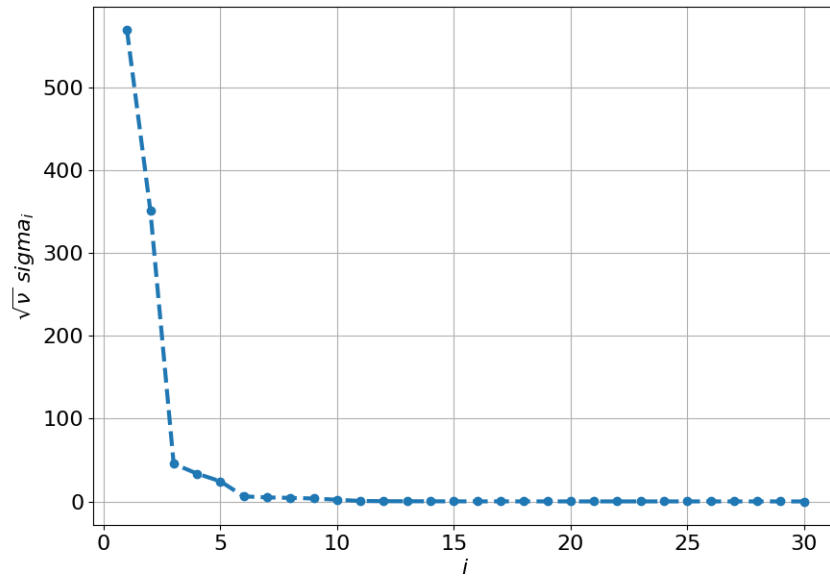


Рис. 2. Стандартные отклонения, соответствующие номерам главных компонент.

9 Демонстрация, того что проекция 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. Произведена проекция матрицы A на 2 первые главные компоненты (листинг 4). Итоговый график для сепарации типов опухолей изображён на рисунке (3).

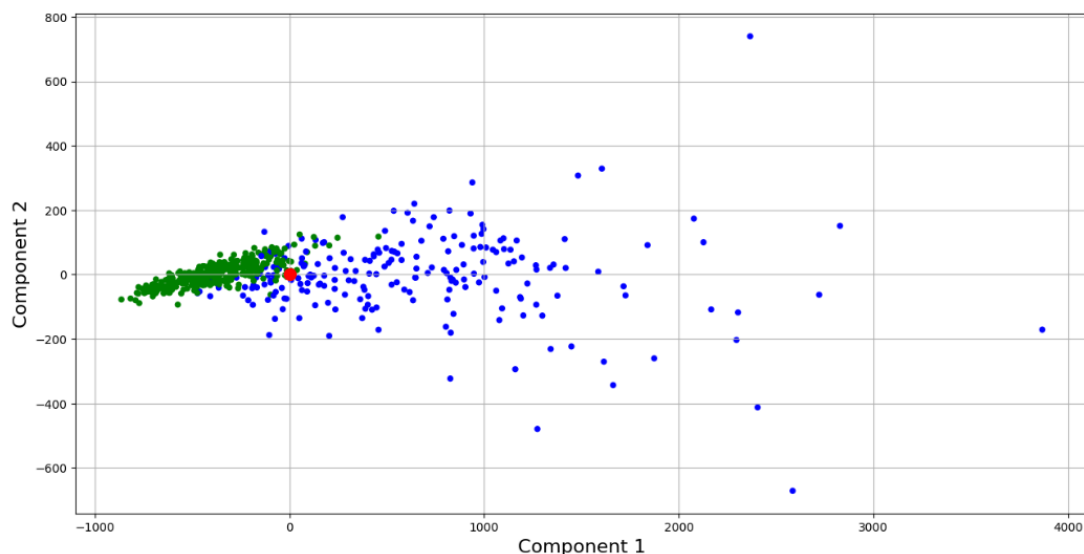


Рис. 3. Сепарация опухолей, где зеленый – malignant, синий – benign, а красная точка – центр координат.

Листинг 4. Проекция матрицы A на 2 первые главные компоненты

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

10 Заключение

В процессе выполнения лабораторной работы:



1. Применён метод PCA для нахождения главных компонент для набора данных Breast Cancer Wisconsin Datas.
2. Продемонстрировано, 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. С. 140.

Выходные данные

Журавлев Н.В.. Отчет о выполнении лабораторной работы по дисциплине «Вычислительная математика». [Электронный ресурс] — Москва: 2021. — 9 с. URL: <https://sa2systems.ru:88> (система контроля версий кафедры РК6)

Постановка:  ассистент кафедры РК-6, PhD А.Ю. Першин
Решение и вёрстка:  студент группы РК6-52Б, Журавлев Н.В.

2021, осенний семестр