



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:	Израелян Ева Арамовна
Группа:	РК6-54Б
Тип задания:	лабораторная работа
Тема:	Спектральное и сингулярное разложение

Студент

подпись, дата

Израелян Е.А.

Фамилия, И.О.

Преподаватель

подпись, дата

Фамилия, И.О.

Москва, 2021

Содержание

Спектральное и сингулярное разложения		3
1	Задание	3
2	Цель выполнения лабораторной работы	4
3	Разработка функции $\text{rsa}(A)$	4
4	Вычисление главных компонент	4
5	Заключение	6

Спектральное и сингулярное разложения

1 Задание

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).

Задача 25(спектральное и сингулярное разложения)

Требуется (базовая часть):

1. Написать функцию $\text{pca}(A)$, 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.
 - 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 ($M = \text{malignant}$, $B = \text{benign}$), и оставшиеся 30 элемент соответствуют характеристикам опухоли.
3. Найти главные компоненты указанного набора данных, используя функцию $\text{pca}(A)$.
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.

2 Цель выполнения лабораторной работы

Цель выполнения лабораторной работы – познакомиться с методом главных компонент, основой которого является сингулярное разложение, применить его на практике для работы с набором данных о пациентах с опухолью.

3 Разработка функции pca(A)

Для начала требовалось произвести масштабирование признаков/показаний, а точнее было необходимо пересчитать все значения x исходной матрицы по формуле $(x - \bar{x})/\sigma_x$, где \bar{x} – это выборочное среднее по данному признаку, а σ_x – выборочное стандартное отклонение. Для этого использовать функция StandardScaler() из sklearn.preprocessing.

Затем на основе полученной матрицы A необходимо сформировать ковариационную матрицу K вида:

$$K = \nu A^T A,$$

где коэффициент $\nu = \frac{1}{m-1}$.

При помощи функции numpy.linalg.eig были определены собственные числа и вектора матрицы K . Затем собственные числа были отсортированы в убывающую последовательность, после чего были отсортированы соответствующие им собственные вектора.

В Листинге 1 представлен код функции pca(A), возвращающей список главных компонент и список соответствующих стандартных отклонений.

Листинг 1. Код функции pca(A)

```
1 def pca(A):
2     scaler = StandardScaler()
3     A = scaler.fit_transform(A)
4     v = 1 / (A.shape[0] - 1)
5     K = A.T @ A * v
6     eig_values, eig_vectors = np.linalg.eig(K)
7     eig_values_sorted = eig_values.argsort()[::-1]
8     pc = eig_vectors[:, eig_values_sorted]
9     eig_values = eig_values[eig_values_sorted]
10    std = np.sqrt(eig_values)
11    return pc, std
```

4 Вычисление главных компонент

В наборе данных Breast Cancer Wisconsin Dataset первый столбец содержит в себе ID пациентов, а второй - диагноз (М - опухоль злокачественная, В - доброкачественная).

Остальные 30 соответствуют характеристикам опухоли. Для того чтобы считать эти данные, была использована функция `pandas.read_csv`. Отбросив первые два столбца, мы получим набор данных `features`, с которым будем дальше работать.

Передав в качестве аргумента матрицу `features` в ранее разработанную нами функцию, реализующую метод главных компонент, получим главные компоненты указанного набора данных и соответствующие им стандартные отклонения. На рис. 1 продемонстрированы полученные стандартные отклонения, соответствующие номерам главных компонент.

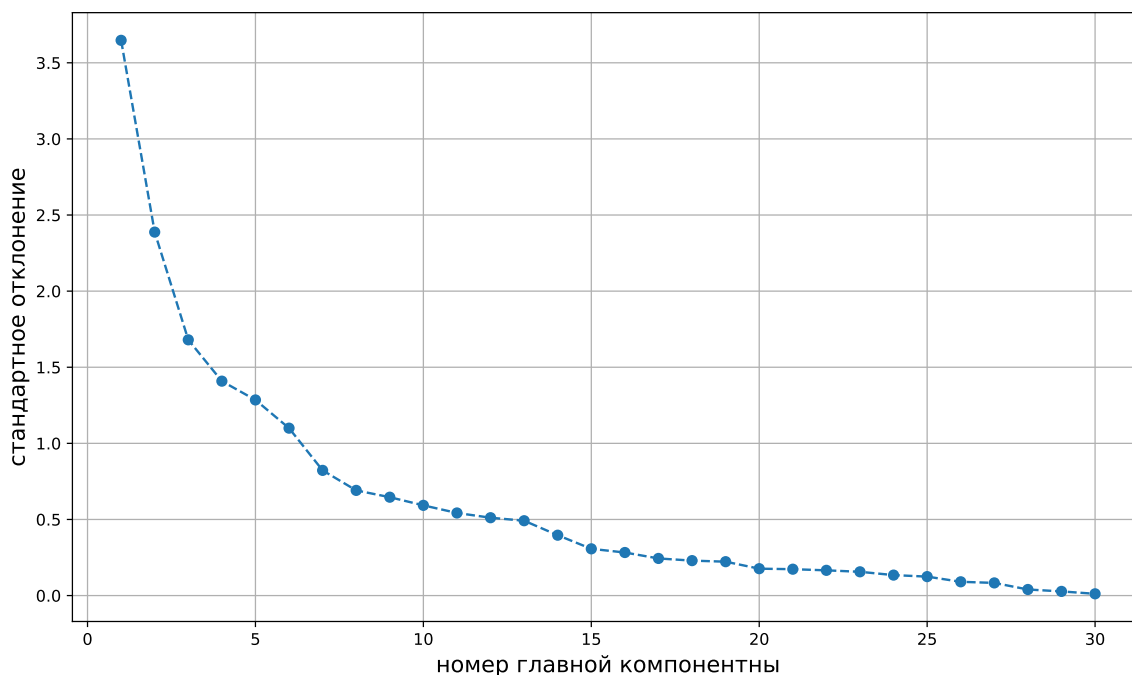


Рис. 1. Стандартные отклонения, соответствующие номерам главных компонент

Из графика видно, что наибольшие стандартные отклонения у первых двух компонент. Это говорит о том, что первые две характеристики представляют наибольший интерес для нас, так как содержат в себе больше информации.

Убедимся в том, что проекций на первые две главные компоненты достаточно для сепарации типов опухолей. Для того чтобы спроецировать данные на главные компоненты, матрицу исходных данных нужно было умножить на главные компоненты, которые были рассчитаны ранее.

На рисунке 2 представлены результаты проекции матрицы признаков на две первые главные компоненты. Зелёным цветом обозначены пациенты с доброкачественной опухолью (В), а жёлтым — с злокачественной (М). Из графика можно сделать вывод, что произведённой проекции было достаточно для сепарации типов опухолей, так как множества слабо пересекаются.

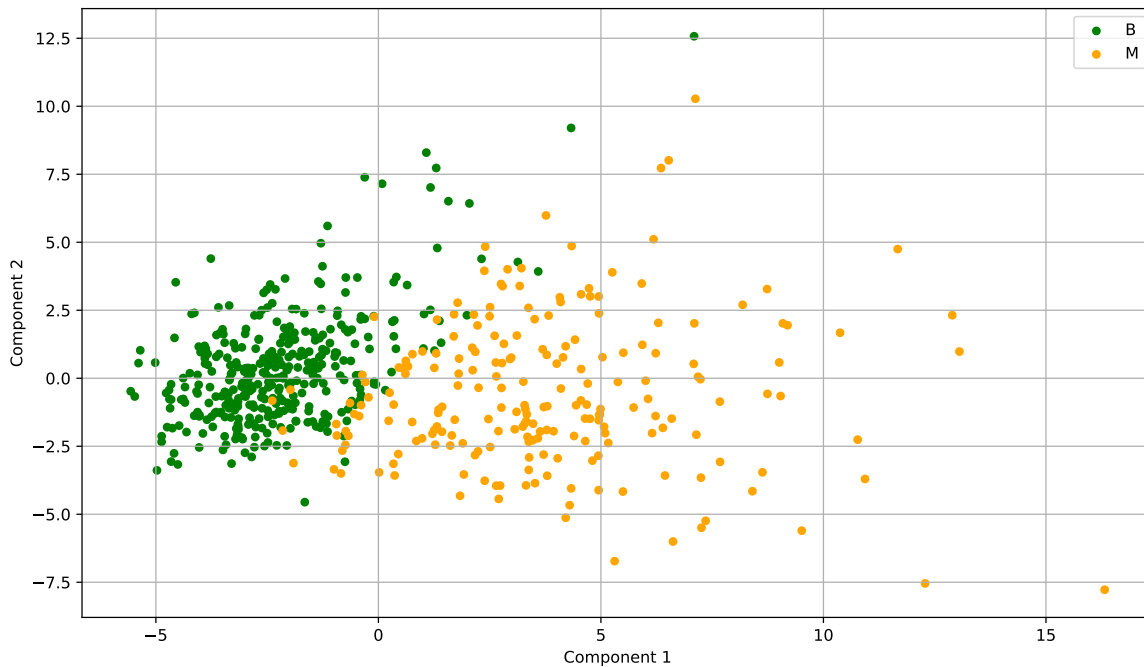


Рис. 2. Проекция матрицы на две первые главные компоненты

5 Заключение

1. Была разработана функция $\text{pca}(A)$ для вычисления главных компонент набора данных Breast Cancer Wisconsin Dataset, хранящего данные о пациентах с опухолью.
2. На основе полученных результатов был получен график стандартных отклонений, соответствующих номерам главных компонент. Первые две главные компоненты имели наибольшее стандартное отклонение, потому использовались далее для проекции.
3. Была произведена проекция на две первые главные компоненты, которая показала, что проекций на первые две 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

Израелян Е.А.. Отчет о выполнении лабораторной работы по дисциплине «Вычислительная математика». [Электронный ресурс] — Москва: 2021. — 7 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 РК6-54Б, Израелян Е.А.

2021, осенний семестр