

Санкт-Петербургский политехнический университет
Высшая школа прикладной математики и
вычислительной физики, ИПММ

Направление подготовки
01.03.02 «Прикладная математика и информатика»

Отчет по курсовой работе
по дисциплине «Математическая статистика»
на тему «Метод главных компонент»

Выполнил студент гр. 3630102/80201
Кирпиченко С. Р.
Руководитель
Баженов А. Н.

Санкт-Петербург
2021

Содержание

	Страница
1 Постановка задачи	5
2 Теория	5
2.1 Постановка задачи метода главных компонент	5
2.2 Алгоритм	5
2.3 Описание прикладной области	6
3 Реализация	6
4 Результаты	7
4.1 Примеры образцов	7
4.2 Пакет образцов 1	8
4.3 Пакет образцов 2	10
4.4 Пакет образцов 3	13
5 Обсуждение	15

Список иллюстраций

	Страница
1 Географическое расположение областей сбора данных	6
2 Несколько исходных образцов	7
3 Сглаженные исходные образцы	8
4 Гистограмма распределения дисперсий по компонентам, пакет 1	8
5 График первой полученной компоненты, пакет 1	9
6 График второй полученной компоненты, пакет 1	9
7 Разбиение главной компоненты 1 на области, пакет 1	9
8 Разбиение главной компоненты 2 на области, пакет 1	10
9 Гистограмма распределения дисперсий по компонентам, пакет 2	10
10 График первой полученной компоненты, пакет 2	11
11 График второй полученной компоненты, пакет 2	11
12 Разбиение главной компоненты 1 на области, пакет 2	12
13 Разбиение главной компоненты 2 на области, пакет 2	12
14 Гистограмма распределения дисперсий по компонентам, пакет 2	13
15 График первой полученной компоненты, пакет 3	13
16 График второй полученной компоненты, пакет 3	13
17 Разбиение главной компоненты 1 на области, пакет 3	14
18 Разбиение главной компоненты 2 на области, пакет 3	14

Список таблиц

	Страница
1 Таблица аминокислот из [4]	6
2 Сопоставление областей с данными, пакет 1	10
3 Сопоставление областей с данными, пакет 2	12
4 Сопоставление областей с данными, пакет 3	15

1 Постановка задачи

2 Теория

2.1 Постановка задачи метода главных компонент

В компонентном анализе ищется такое линейное преобразование

$$\hat{x} = L\hat{f}, \quad (1)$$

где $\hat{x} = (x_1, \dots, x_d)$, $\hat{f} = (f_1, \dots, f_d)$ - векторы-столбцы случайных величин и $L = \|l_{ij}\|$ - квадратная матрица размером $d \times d$, в которой случайные величины f_1, \dots, f_d некоррелированы и нормированы $E f_i = 0$, $D f_i = 1$, $i = 1, \dots, d$; всегда для простоты предполагается, что $E x_i = 0$, $i = 1, \dots, d$. В этом случае дисперсия выражается как

$$D x_i = l_{i1}^2 + \dots + l_{id}^2, \quad i = 1, \dots, d$$

Следовательно, суммарная дисперсия $\{x_i\}_{i=1}^d$ равна

$$\sum_{i=1}^d D x_i = \sum_{i=1}^d l_{i1}^2 + \dots + \sum_{i=1}^d l_{id}^2 \quad (2)$$

Отыскание представления (1) эквивалентно определению d таких нормированных линейных комбинаций y_1, \dots, y_d переменных x_1, \dots, x_d (т.е. сумма квадратов коэффициентов равна 1), что для каждого $k = 1, \dots, d$ y_k имеет наибольшую дисперсию среди всех нормированных линейных комбинаций при условии некоррелированности с предыдущими комбинациями y_1, \dots, y_{k-1} . Такие линейные комбинации y_1, \dots, y_d называются *главными компонентами* системы случайных величин x_1, \dots, x_d .

2.2 Алгоритм

Пусть дана d - мерная выборка (X_1, \dots, X_n) .

1. Составим матрицу

$$X = \begin{bmatrix} x_1^1 & \dots & x_n^1 \\ x_1^2 & \dots & x_n^2 \\ \dots & \dots & \dots \\ x_1^d & \dots & x_n^d \end{bmatrix} \quad (3)$$

2. Построим ковариационную матрицу

$$C = \frac{1}{n-1} X X^T. \quad (4)$$

3. C диагонализуемая, то есть представима в виде

$$C = P^T \Lambda P, \quad (5)$$

где P^T есть ортонормированная матрица, содержащая собственные векторы матрицы C , или *главные компоненты*, а Λ - диагональная матрица, содержащая соответствующие главным компонентам собственные числа матрицы C . Причем, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$, и λ_i есть вклад компоненты f_i в суммарную дисперсию $x_1 \dots x_d$, равную в силу (2) $\lambda_1 + \dots + \lambda_d = \text{tr}(\Lambda)$

4. Для проекции X на множество главных компонент с индексами i_1, \dots, i_k составим матрицу P , столбцами которой будут являться собственные вектора v_{i_1}, \dots, v_{i_k} . Тогда проекцией X на множество главных компонент с индексами i_1, \dots, i_k будет являться

$$Y = P X. \quad (6)$$

2.3 Описание прикладной области

На объект подается излучение частотой от видимого спектра до ультрафиолета. Излучение поглощается молекулами, которые, в свою очередь, излучали свет с некоторой частотой. Варьируя частоту подаваемого излучения, строятся графики ответного излучения (частота + интенсивность). Предлагается построить главные компоненты данных, представленных руководителем, выделить в них максимумы, проанализировать их расположение на соответствие некоторой аминокислоте с помощью табл. 1 и сравнить результаты для двух пакетов данных.

Название	Excitation	Emission
Humic acid-like	320-350	420-480
Humic acid-like	250-260	380-420
Mariane humic acids	310-320	380-420
Protein-like containing Tryptophan	270-280	300-320
Tryptophan and Protein-like Related to Biological	270-280	320-350
Fulvic acids	245-255	410-440

Таблица 1: Таблица аминокислот из [4]

Два первых пакета данных были собраны на озере Киву (Восточная Африка, одно из Великих Африканских озёр). Третий пакет - на Байкале.

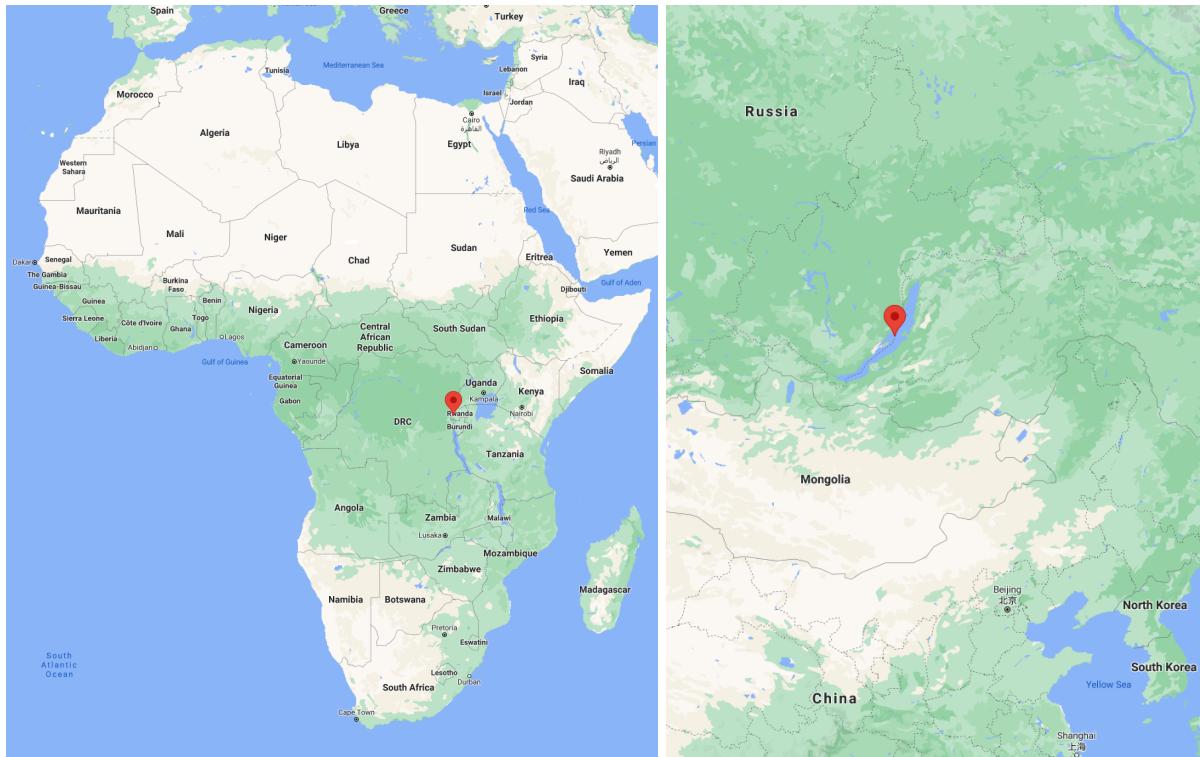


Рис. 1: Географическое расположение областей сбора данных

3 Реализация

Лабораторная работа выполнена на языке Python в среде PyCharm с использованием библиотек numpy, matplotlib.pyplot. Метод главных компонент был взят из модуля decomposition библиотеки sklearn.

Научным руководителем предоставлено 30 образцов исходных данных «Hexane_extr_Kivu_Lake», 145 образцов «Kivu_220», 15 образцов «Baikal220» и 2 научные статьи: [4], [5]. Первый пакет образцов частично входит во второй, но есть и

различные семплы.

4 Результаты

4.1 Примеры образцов

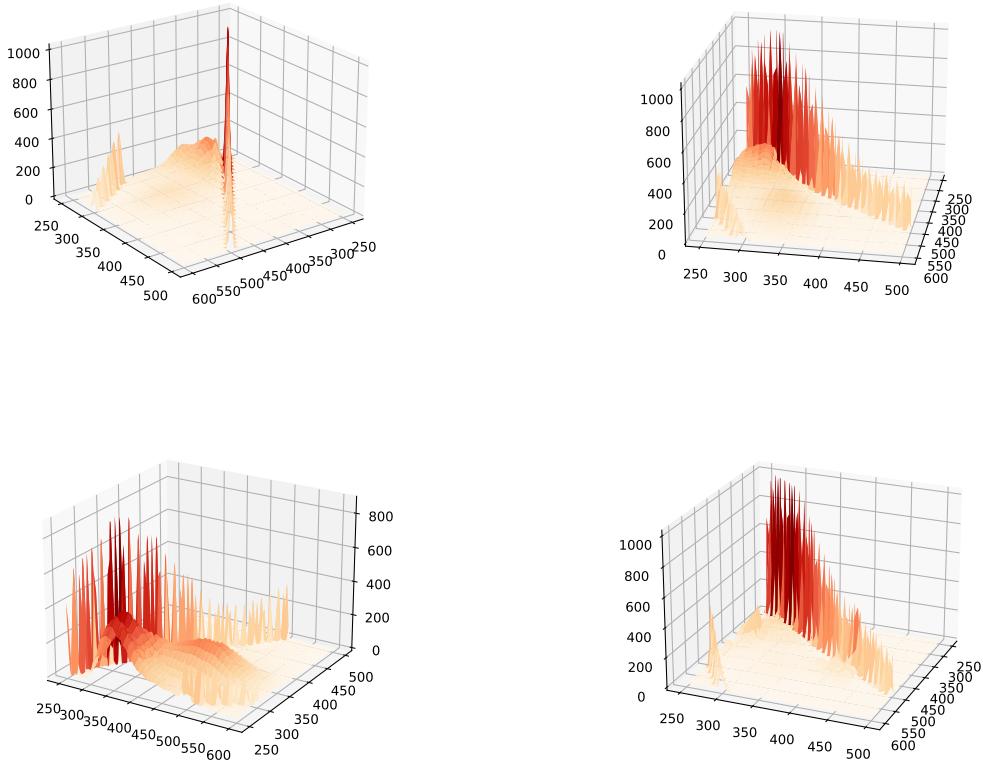


Рис. 2: Несколько исходных образцов

4 образца исходных данных для примера изображены на рис. 2. Как видим, на всех графиках присутствует шум - две полосы резких пиков. Чтобы убрать данные выбросы из исходных данных, был применен медианный фильтр с размером окна (10, 4). В результате этой операции приведенные на 2 образцы преобразованы следующим образом:

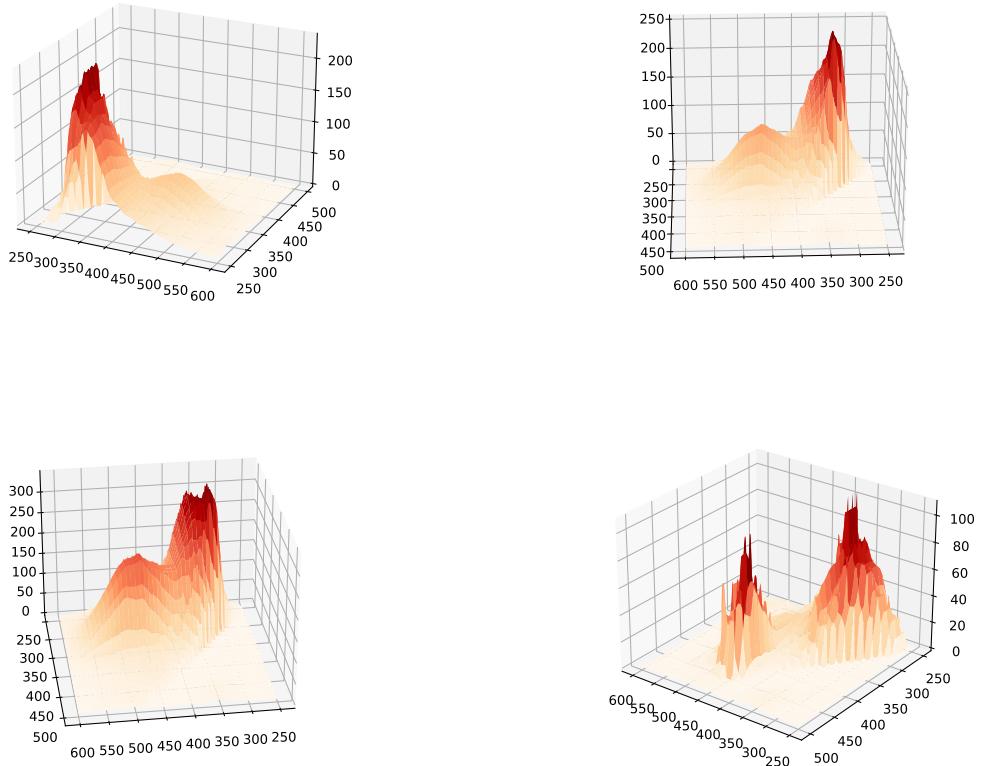


Рис. 3: Сглаженные исходные образцы

Как видим, характер и количество экстремумов (исключая описанный шум) было сохранено.

4.2 Пакет образцов 1

По итогам применения метода главных компонент были получены следующие результаты:

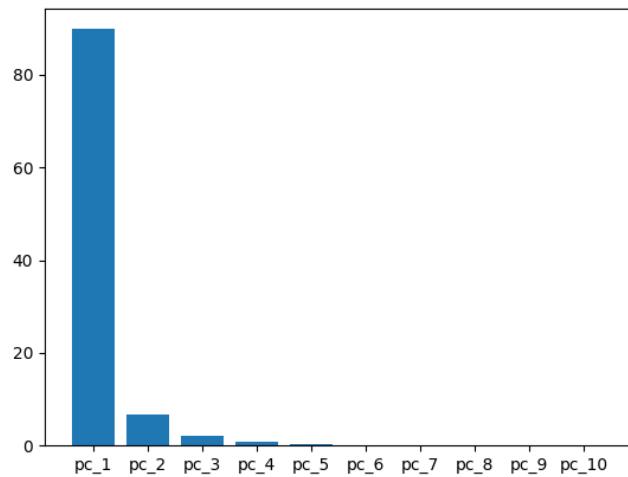


Рис. 4: Гистограмма распределения дисперсий по компонентам, пакет 1

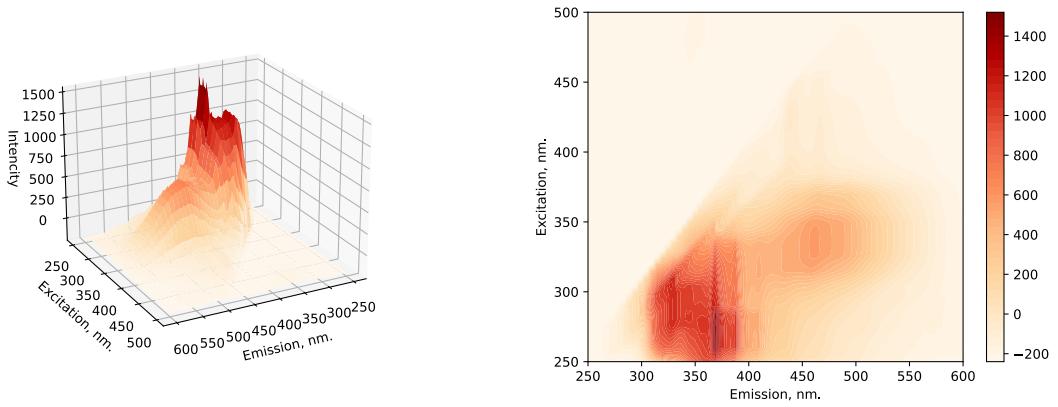


Рис. 5: График первой полученной компоненты, пакет 1

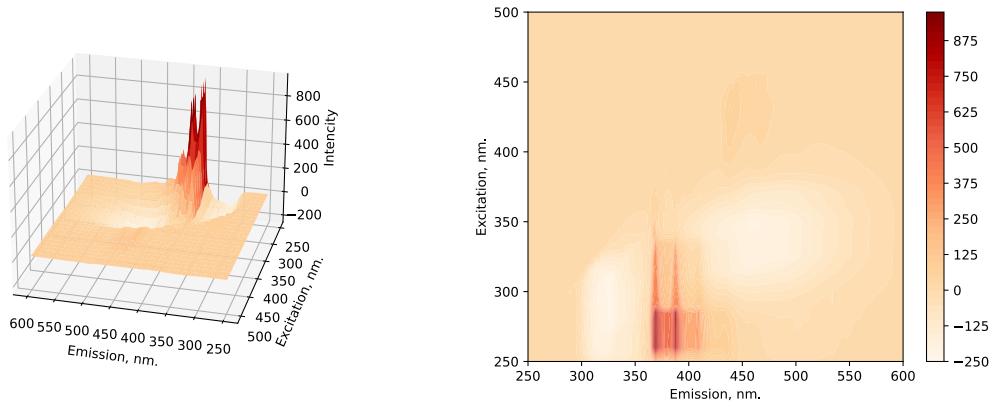


Рис. 6: График второй полученной компоненты, пакет 1

Отметим области максимальной дисперсии на рассматриваемых главных компонентах для сопоставления их с областями из рис. 7, [4].

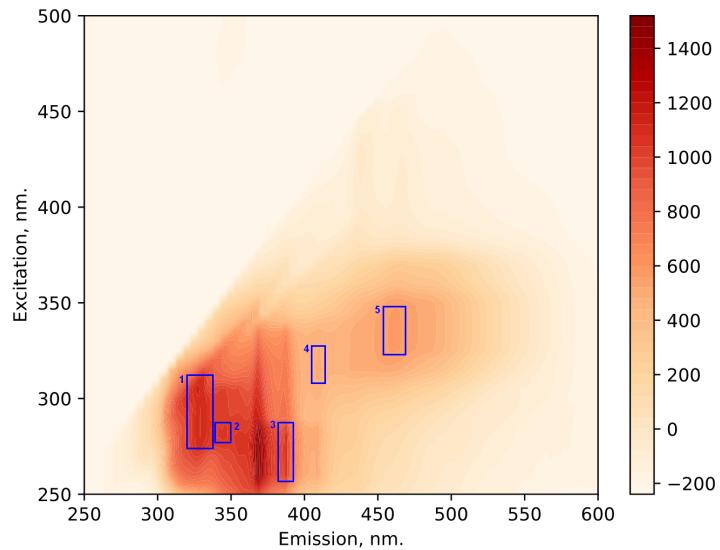


Рис. 7: Разбиение главной компоненты 1 на области, пакет 1

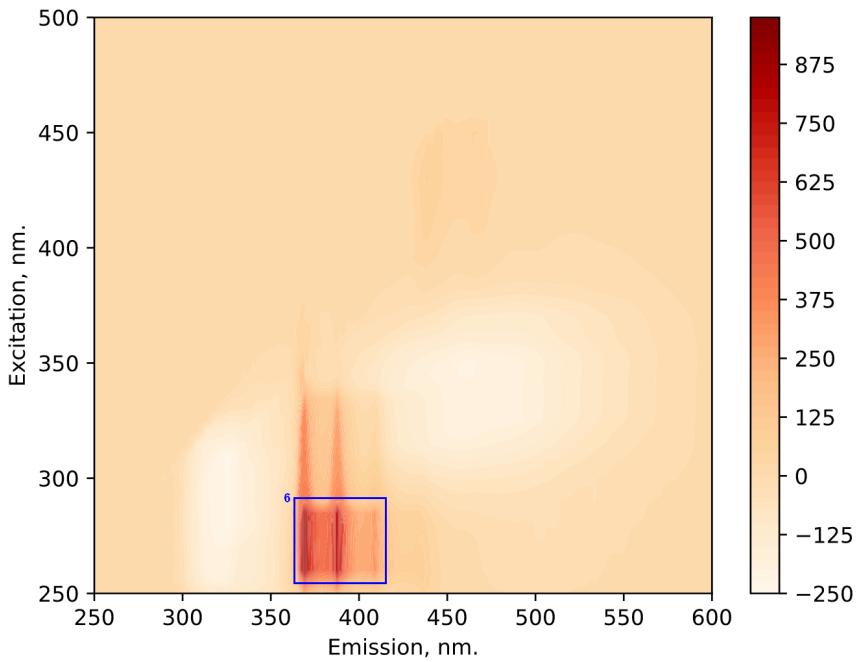


Рис. 8: Разбиение главной компоненты 2 на области, пакет 1

Номер области	Сопоставленное из [4] вещество
1	Protein-like containing Tryptophan
2	Tryptophan and Protein-like Related to Biological
3	Humic acid-like
4	Marine humic acids
5	Humic acid-like
6	Humic acid-like

Таблица 2: Сопоставление областей с данными, пакет 1

4.3 Пакет образцов 2

По итогам применения метода главных компонент были получены следующие результаты:

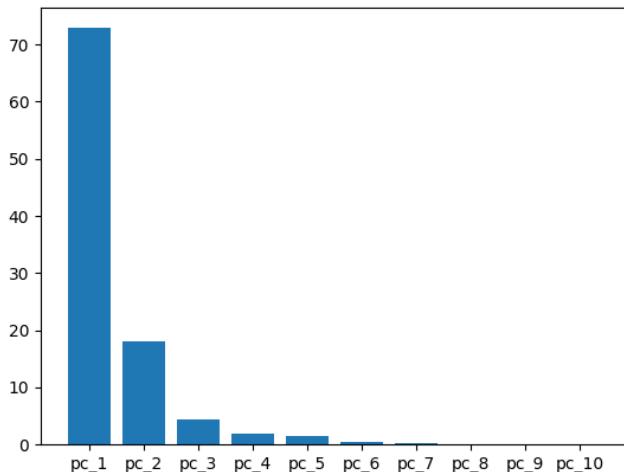


Рис. 9: Гистограмма распределения дисперсий по компонентам, пакет 2

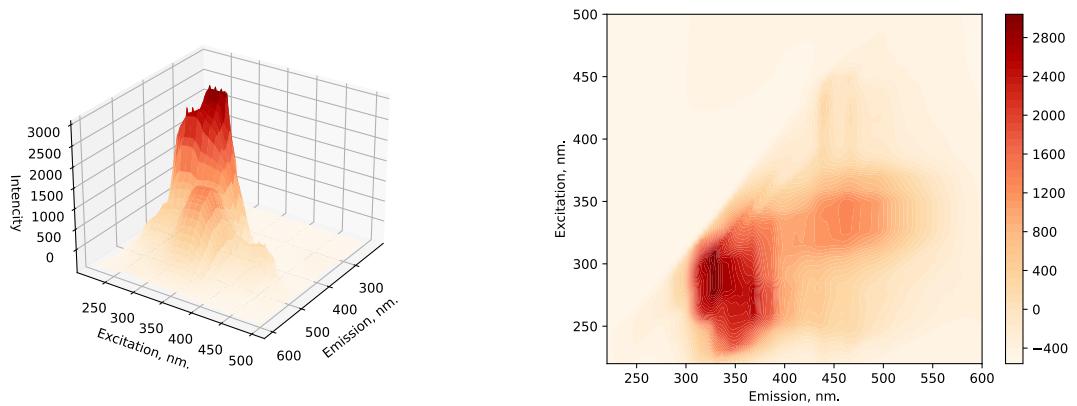


Рис. 10: График первой полученной компоненты, пакет 2

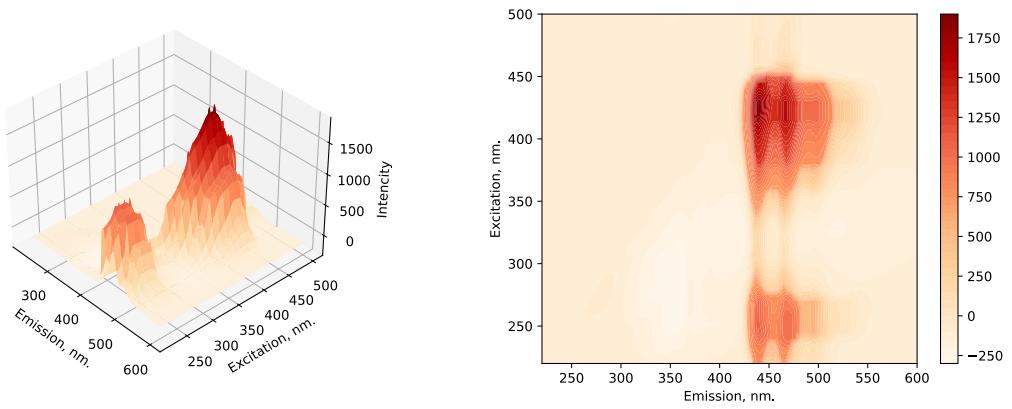


Рис. 11: График второй полученной компоненты, пакет 2

Отметим области максимальной дисперсии на рассматриваемых главных компонентах для сопоставления их с областями из рис. 7, [4].

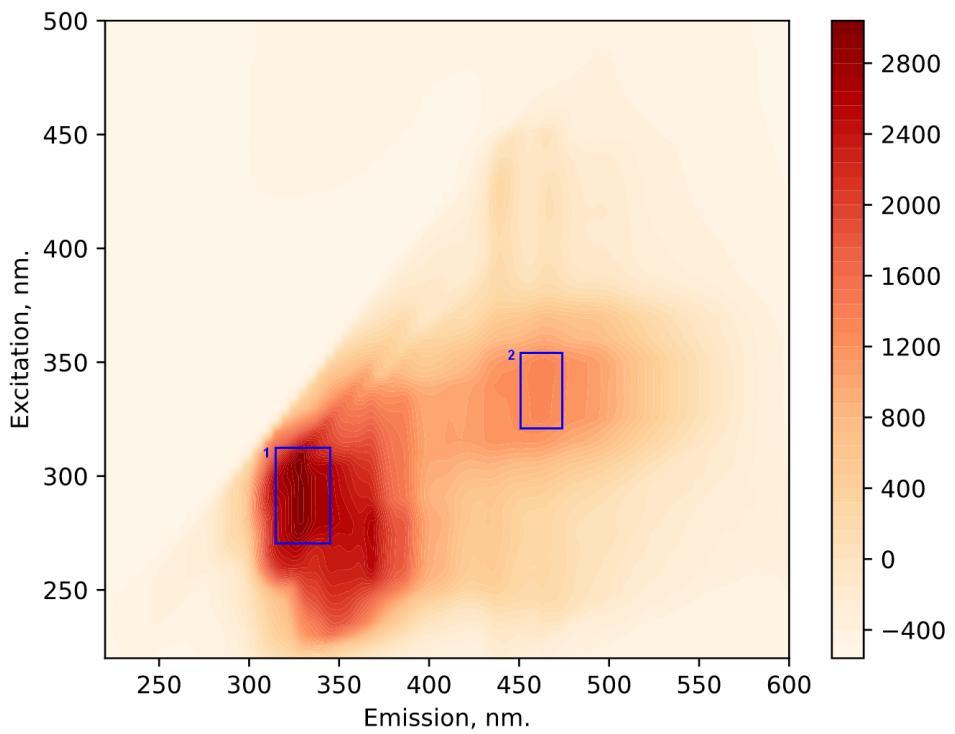


Рис. 12: Разбиение главной компоненты 1 на области, пакет 2

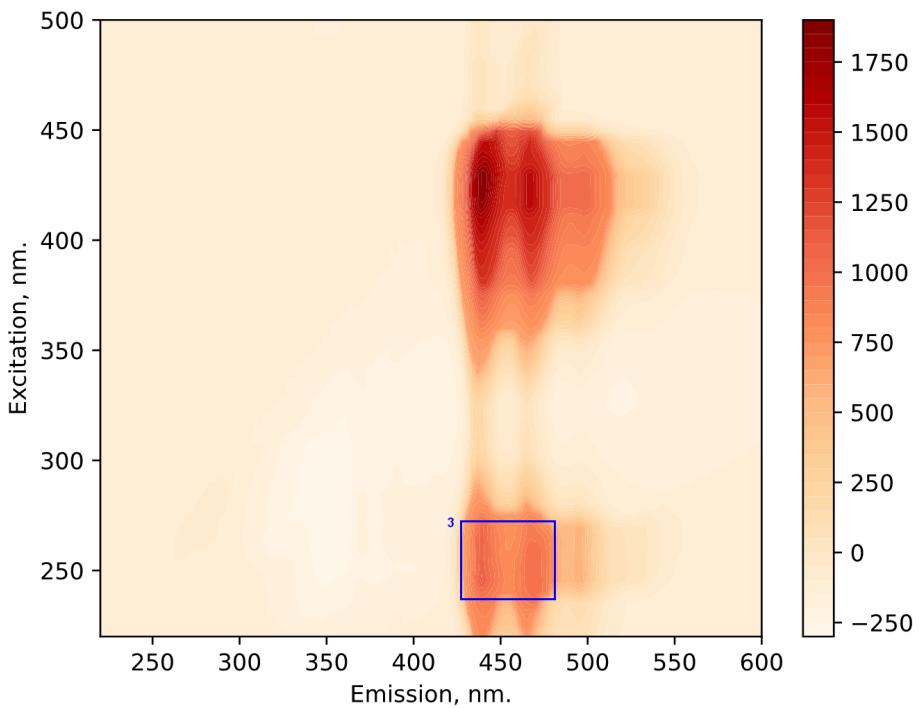


Рис. 13: Разбиение главной компоненты 2 на области, пакет 2

Номер области	Сопоставленное из [4] вещество
1	Protein-like containing Tryptophan
2	Humic acid-like
3	Humic acid-like

Таблица 3: Сопоставление областей с данными, пакет 2

4.4 Пакет образцов 3

По итогам применения метода главных компонент были получены следующие результаты:

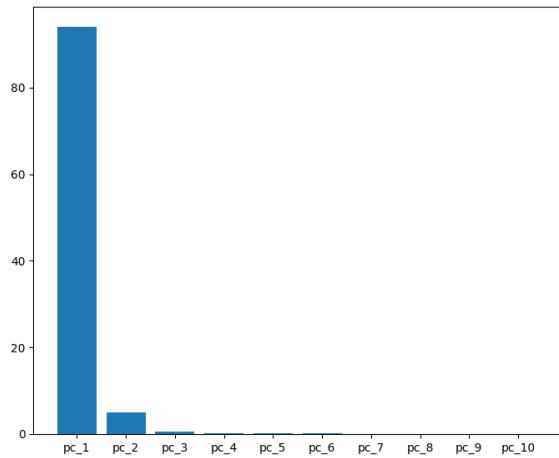


Рис. 14: Гистограмма распределения дисперсий по компонентам, пакет 2

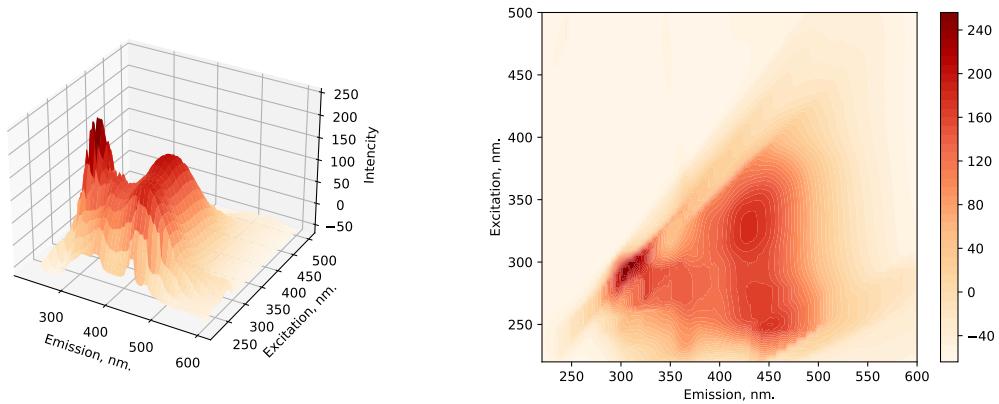


Рис. 15: График первой полученной компоненты, пакет 3

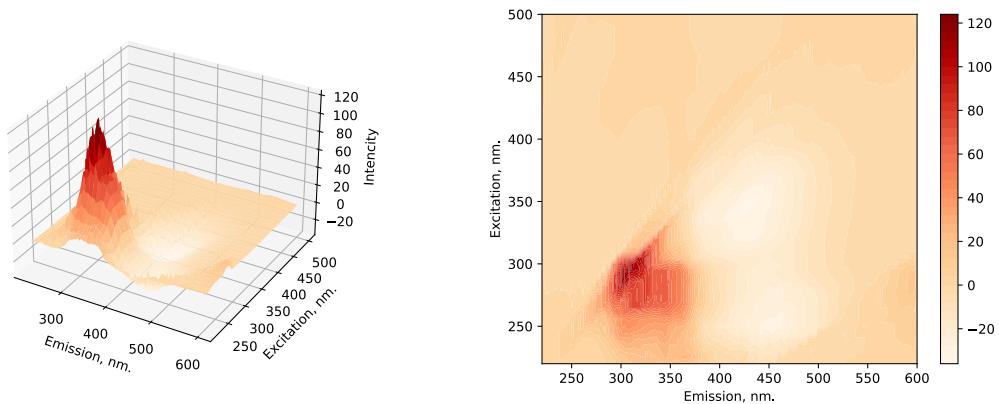


Рис. 16: График второй полученной компоненты, пакет 3

Отметим области максимальной дисперсии на рассматриваемых главных компонентах для сопоставления их с областями из рис. 7, [4].

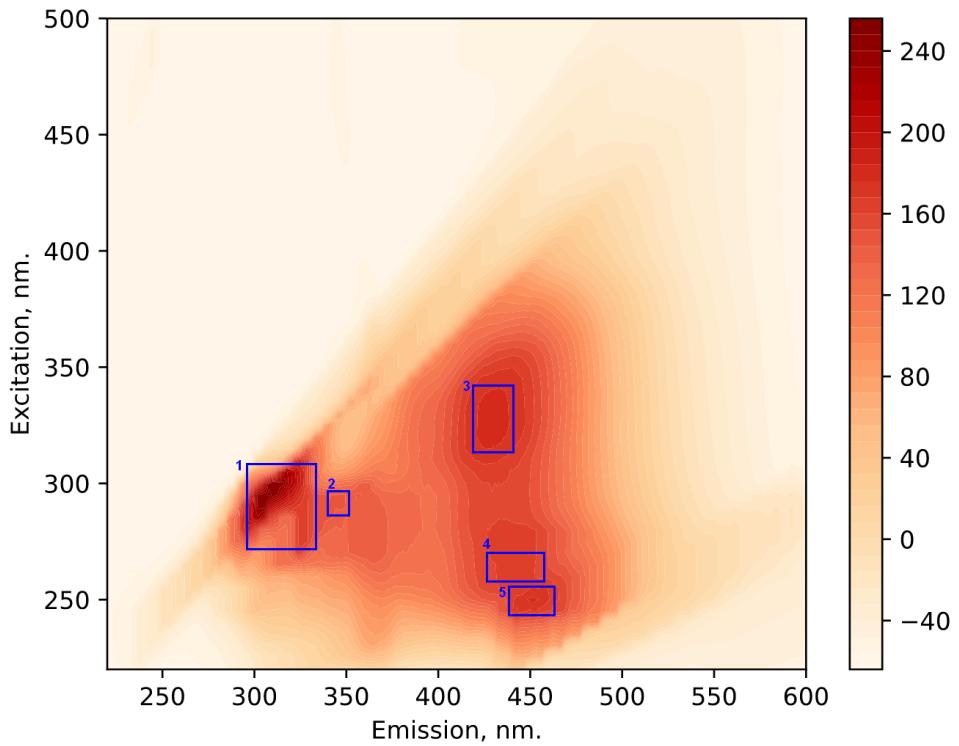


Рис. 17: Разбиение главной компоненты 1 на области, пакет 3

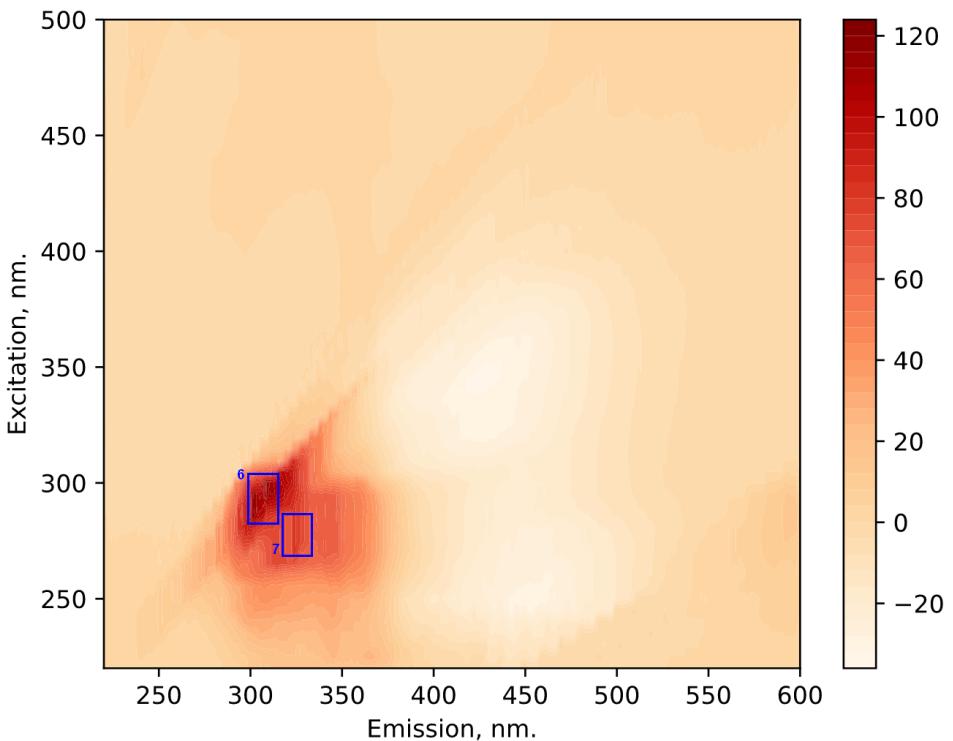


Рис. 18: Разбиение главной компоненты 2 на области, пакет 3

Номер области	Сопоставленное из [4] вещество
1	Protein-like containing Tryptophan
2	Tryptophan and Protein-like Related to Biological
3	Humic acid-like
4	Humic acid-like
5	Fulvic acids
6	Protein-like containing Tryptophan
7	Tryptophan and Protein-like Related to Biological

Таблица 4: Сопоставление областей с данными, пакет 3

5 Обсуждение

1. Из гистограмм (рис. 4, рис. 9 и рис. 14) видно, что первая компонента покрывает значительное количество дисперсии исходных данных. Первые три рассмотренные компоненты покрывают почти все 100% дисперсии во всех случаях;
2. На пакетах с малым числом образцов первая компонента покрывает большую часть дисперсии: 90% против 70% на пакете с большим числом образцов;
3. Максимумы на рис. 5, 6, 10, 11, 15, 16 обозначают области, в которых дисперсия данных среди представленных образцов максимальна. Некоторым из этих максимумов удалось сопоставить вещества из статьи [4] (табл. 2, 3), однако область максимума в первой компоненте в окрестности точки (360, 260), присутствующую во всех пакетах данных, не удалось соотнести ни с какой аминокислотой в [4];
4. Рельеф первой главной компоненты в целом похож на большую часть исходных данных (например, можно визуально сопоставить рис. 5 и первые три образца из рис. 3);
5. Во втором пакете данных области максимума получились более обширными и слаженными, вероятно, из-за возросшего количества образцов. Вследствие этого не удалось во втором случае выделить столько же веществ, сколько и в первом, - для них нет локальных максимумов на рис. 10;
6. Первые главные компоненты во всех пакетах имеют схожий рельеф, вторые же значительно различаются;
7. Во всех случаях первая главная компонента не содержит ярко выраженных оврагов, в отличие от второй. Это связано с требованием ортогональности главных компонент;
8. В обоих пакетах, относящихся к Киву, вторая главная компонента содержит максимум в окрестности точки (460, 410) (особенно выраженный во втором пакете), который вовсе не вписывается в рассматриваемую область рис. 7 в [4];
9. В пакете, относящемся к Байкалу, удалось выделить вещество, не присутствующее в пакетах, относящихся к Киву, - Fulvic acid. В свою очередь, в первом африканском пакете удалось выделить уникальную для него кислоту Marine humic acid. В целом же, спектр аминокислот схож и покрывает почти все вещества в регионах 4 и 5, рис. 7 в [4].

Исходный код

С исходным кодом программы и отчета можно ознакомиться в репозитории <https://github.com/Stasychbr/MatStat>.

Список литературы

- [1] Максимов Ю.Д. Математика. Теория и практика по математической статистике. Конспект-справочник по теории вероятностей : учеб. пособие / Ю.Д. Максимов; под ред. В.И. Антонова. – СПб. : Изд-во Политехн. ун-та, 2009. – 395 с. (Математика в политехническом университете).
- [2] Ивченко Г.И., Медведев Ю.И. Математическая статистика: Учебник. — М.: Издательство ЛКИ, 2014. — 352 с.
- [3] Айвазян, Бухштабер, Енюков, Мешалкин. Прикладная Статистика. Классификация и снижение размерности. - М.: Финансы и статистика, 1989. - 607 с.
- [4] Chen W., Westerhoff P., Leenheer J.A., Booksh K. Fluorescence Excitation-Emission Matrix Regional Integration to Quantify Spectra for Dissolved Organic Matter // Environ. Sci. Technol. 2003, 37, p. 5701-5710
- [5] Semenov P.B., et al. Methane and Dissolved Organic Matter in the Ground Ice Samples from Central Yamal: Implications to Biogeochemical Cycling and Greenhouse Gas Emission. // Geosciences. 2020: 450 c.
- [6] Dramichanin T., Ackovich L.L., Zekovich I., Dramichanin M. D. Detection of Adulterated Honey by Fluorescence Excitation-Emission Matrices // Hindawi Journal of Spectroscopy Volume 2018, Article ID 8395212, 6 p.