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

Список иллюстраций

Список таблиц

1 Постановка задачи

2 Теория

2.1 Постановка задачи метода главных компонент

В компонентном анализе ищется такое линейное преобразование

$$\widehat{x} = L\widehat{f}, \tag{1}$$

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

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

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

$$\sum_{i=1}^{d} \mathbf{D}x_i = \sum_{i=1}^{d} l_{i1}^2 + \cdots + \sum_{i=1}^{d} l_{id}^2 \tag{2}$$

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

2.2 Алгоритм

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

1. Составим матрицу
$$X = \begin{bmatrix} x_1^1 & \ldots & x_n^1 \\ x_1^2 & \ldots & x_n^2 \\ \ldots & \ldots & \ldots \\ x_1^d & \ldots & x_n^d \end{bmatrix} \tag{3}$$

2. Построим ковариационную матрицу
$$C = \frac{1}{n-1} X X^T. \tag{4}$$

3. C диагонализуемая, то есть представима в виде
$$C = P^T \Lambda P, \tag{5}$$
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, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d > 0$, и λ_i есть вклад компоненты f_i в суммарную дисперсию $x_1 ... x_d$, равную в силу (2) $\lambda_1 + ... + \lambda_d = tr(\Lambda)$

4. Для проекции X на множество главных компонент с индексами $i_1, ...i_k$ составим матрицу P, столбцами которой будут являться собственые вектора $v_{i_1}, ..., v_{i_k}$. Тогда проекцией X на множество главных компонент с индексами $i_1, ...i_k$ будет являться
$$Y = PX. \tag{6}$$

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» и 2 научные статьи: [4], [5].

4 Результаты

4.1 Примеры образцов

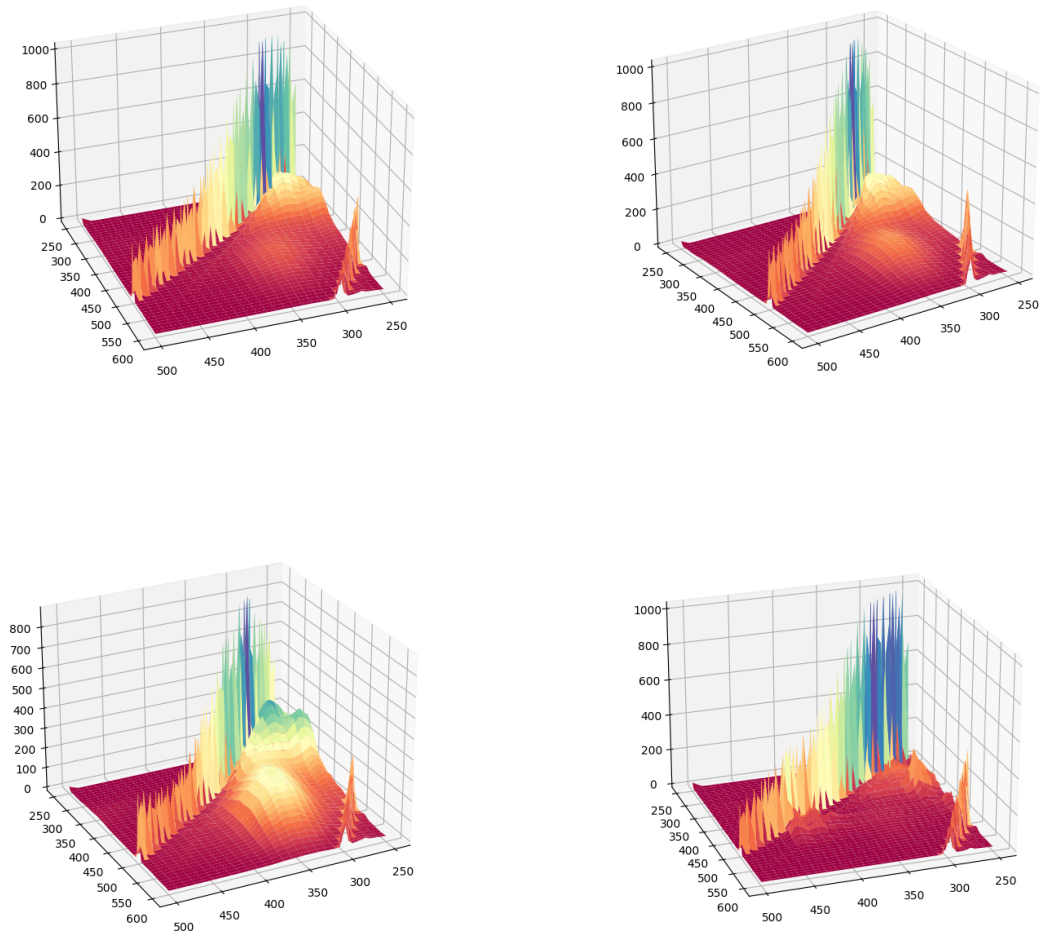

Рис. 1: Несколько исходных образцов

Исходные данные включают в себя 30 образцов, 4 из которых для примера изображены на рис. 1. 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 1 образцы преобразованы следующим образом:

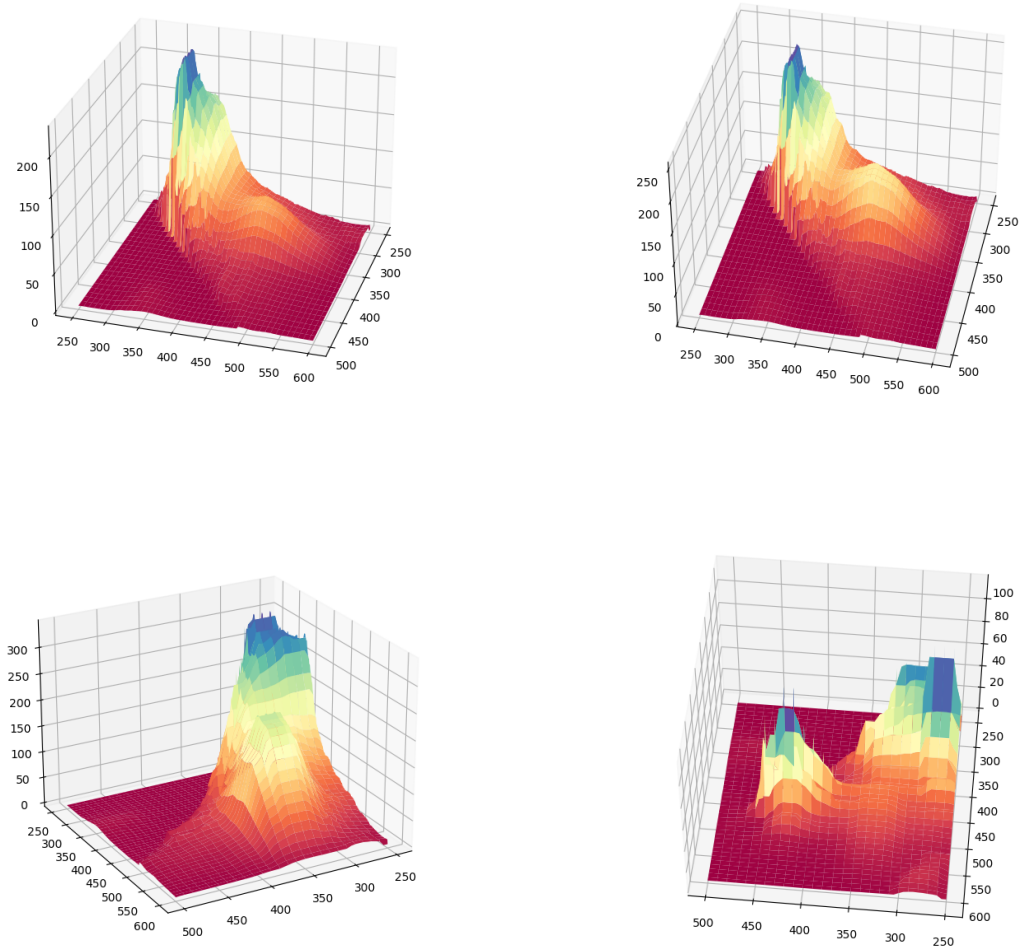

Рис. 2: Сглаженные исходные образцы

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

По итогам применения метода главных компонент были получены следующие результаты:

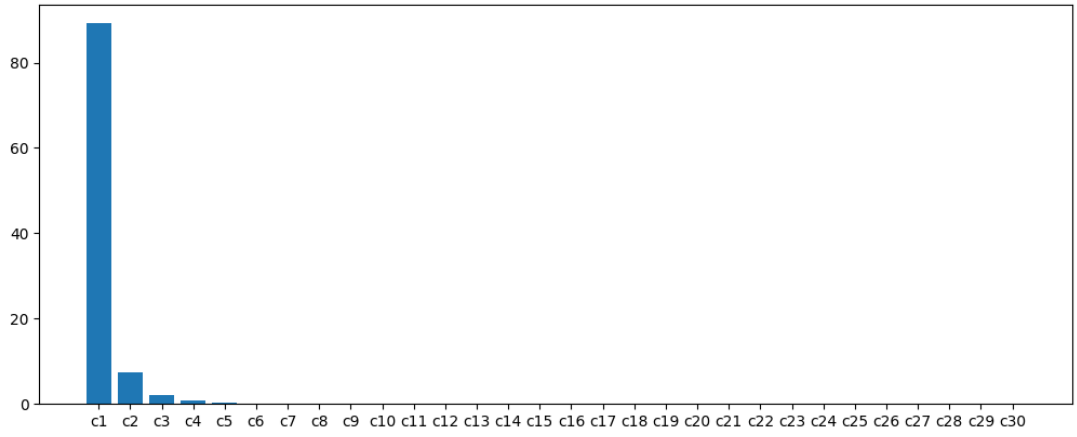

Рис. 3: Гистограмма распределения дисперсий по компонентам

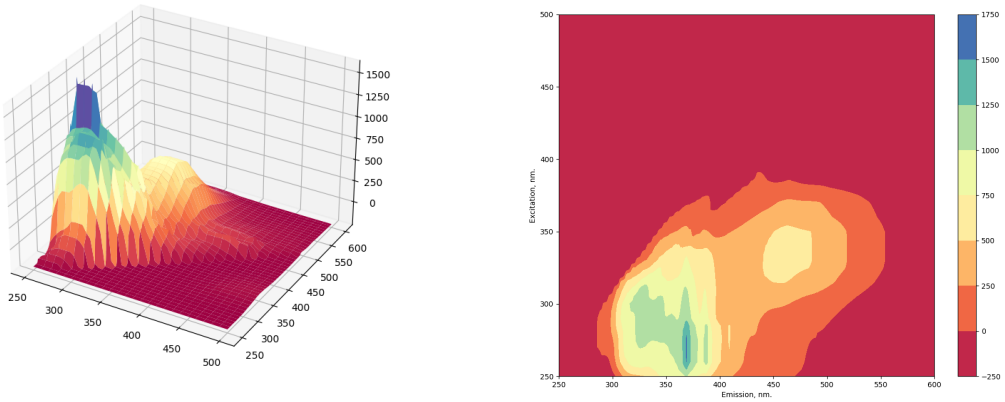

Рис. 4: График первой полученной компоненты

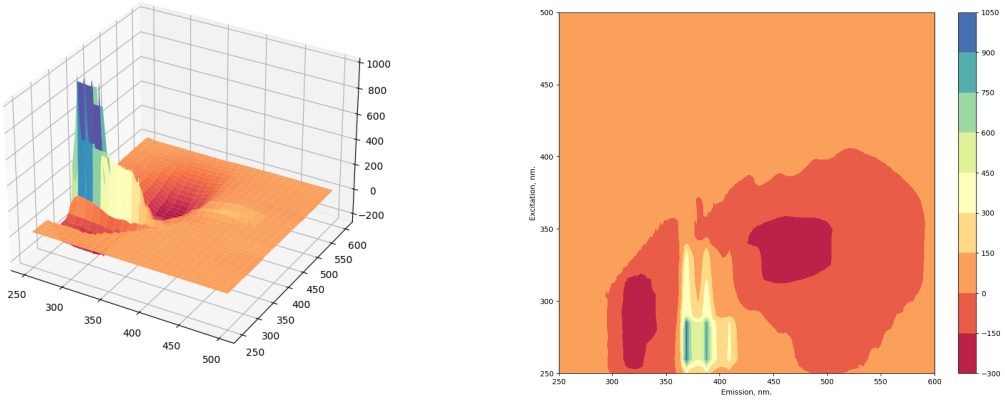

Рис. 5: График второй полученной компоненты

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

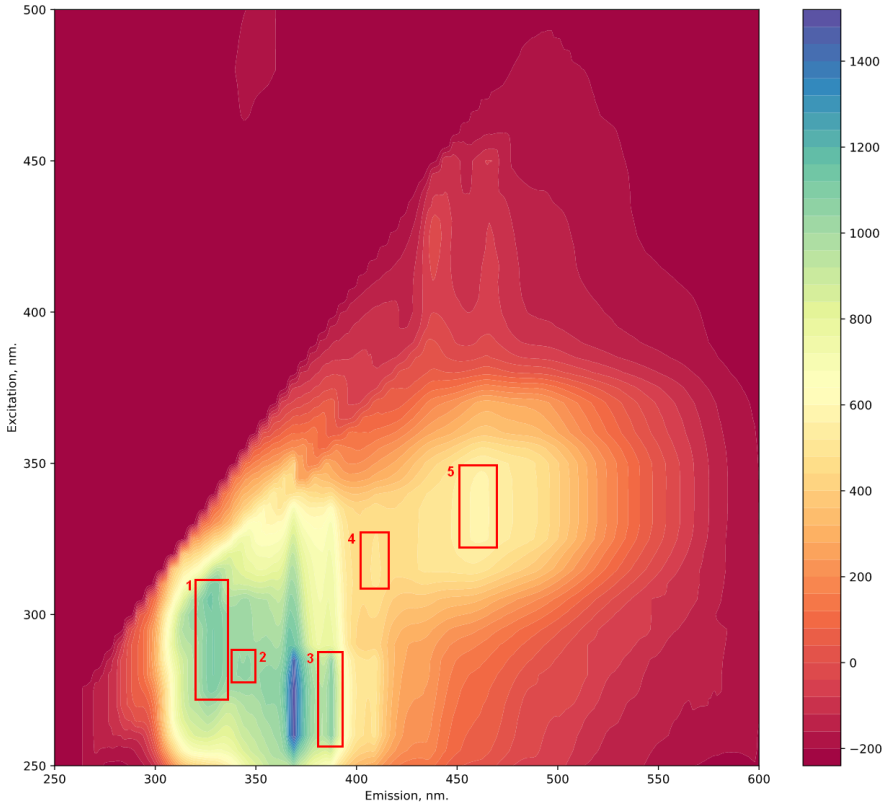

Рис. 6: Разбиение главной компоненты 1 на области

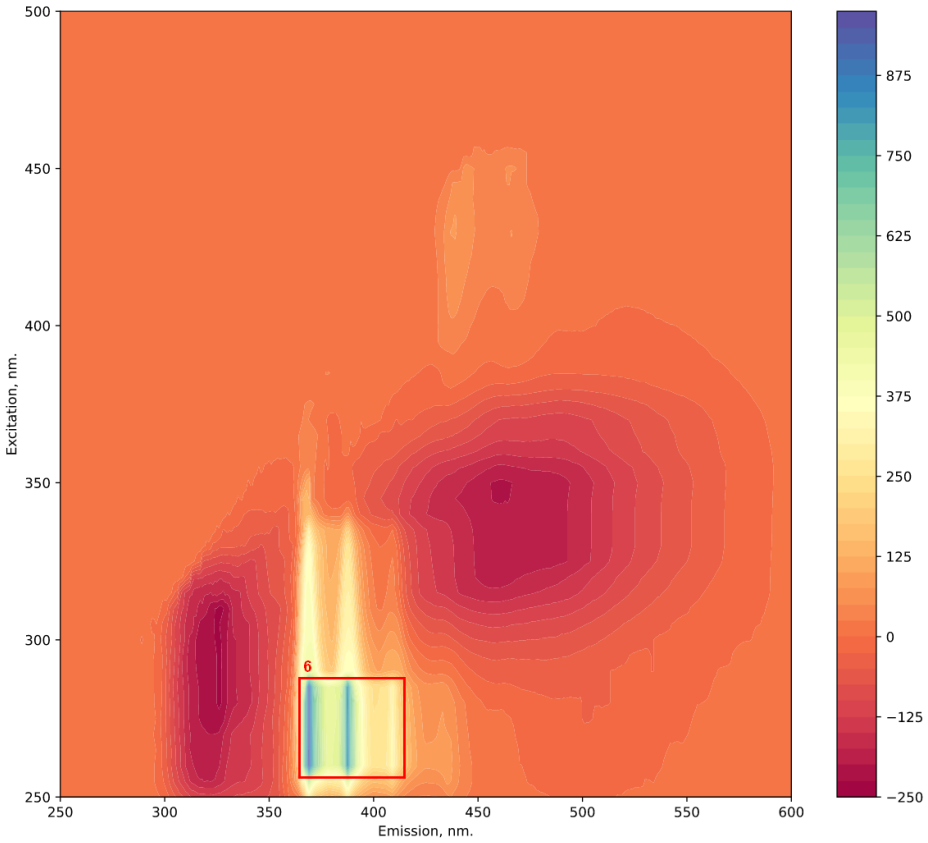

Рис. 7: Разбиение главной компоненты 2 на области

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 1: Сопоставление областей с данными

5 Обсуждение

1. Из гистограммы (рис. 3) видно, что первая компонента покрывает примерно 90% дисперсии исходных данных. Первые две рассмотренные компоненты покрывают почти все 100% дисперсии;
2. Максимумы на рис. 4 и рис. 5 обозначают области, в которых дисперсия данных среди представленных образцов максимальна. Некоторым из этих максимумов удалось сопоставить вещества из статьи [4] (табл. 1), однако область глобального максимума в первой компоненте не удалось соотнести ни с какой аминокислотой в [4];
3. Рельеф первой главной компоненты в целом похож на большую часть исходных данных (например, можно визуально сопоставить рис. 4 и первые три образца из рис. 2).

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.