САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Институт прикладной математики и механики

Высшая школа прикладной математики и вычислительной физики

Отчет
по курсовой работе
по дисциплине
«Математическая статистика»
на тему
«Применение метода главных компонент в анализе результатов флуориметрии»

Выполнил студент:
Колосков Александр
группа: 3630102/80301

Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2021 г.

Содержание

Список иллюстраций

1 Постановка задачи

1. Методом главных компонент обработать результаты флюориметрии, представленные в виде EEM (Excitation Emission Matrix).

2. Проанализировать вклад главных компонент в суммарную дисперсию. Выбрать первые k компонент, вклад которых в суммарную дисперсию $\approx 90\%$.

3. Визуализировать исходные данные и главные компоненты. На основании визуализации сделать выводы о физическом смысле построенных главных компонент.

2 Теория

2.1 Постановка задачи метода главных компонент

В компонентном анализе ищется такое линейное преобразование

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

где $\widehat{x} = (x_1, \dots, x_d),\ \widehat{f} = (f_1, \dots, f_d)$ - 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 $\mathbf{E}f_i = 0,\ \mathbf{D}f_i = 1,\ i = 1, \dots, d$; всегда для простоты предполагается, что $\mathbf{E}x_i = 0,\ i = 1, \dots, d$. В этом случае дисперсия выражается как

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

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

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

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

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

2. Построим ковариационную матрицу

$$C = \frac{1}{n-1} XX^T. \tag{4}$$

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.
Данные для анализа были предоставлены научным руководителем (архив ToBazhenov, папка VD_DOM_Permafrost, 17 образцов за исключением образцов с несовпадающими размерностями матриц) вместе со статьей [5].

Для применения метода главных компонент, трехсторонний массив данных [17 Samples X 76 Emission's lambda X 351 Excitation's lambda] данных должен быть развернут в двумерную матрицу [17 Samples X 26676 Parameters]. После анализа рассчитанные параметры главных компонент нужно снова свернуть в двухсторонний массив [76 Emission's lambda X 351 Excitation's lambda].

Для увеличения точности метода главных компонент(см. [6]) и выразительности графического представления результатов было принято решение сделать «срезки» и аппроксимацию медианным фильтром так называемых областей реллеевского рассеяния - областей высокой интенсивности флуорисценции, находящихся вне зоны полезных данных. Для графического представления было выбрано представление линиями уровня.

4 Результаты

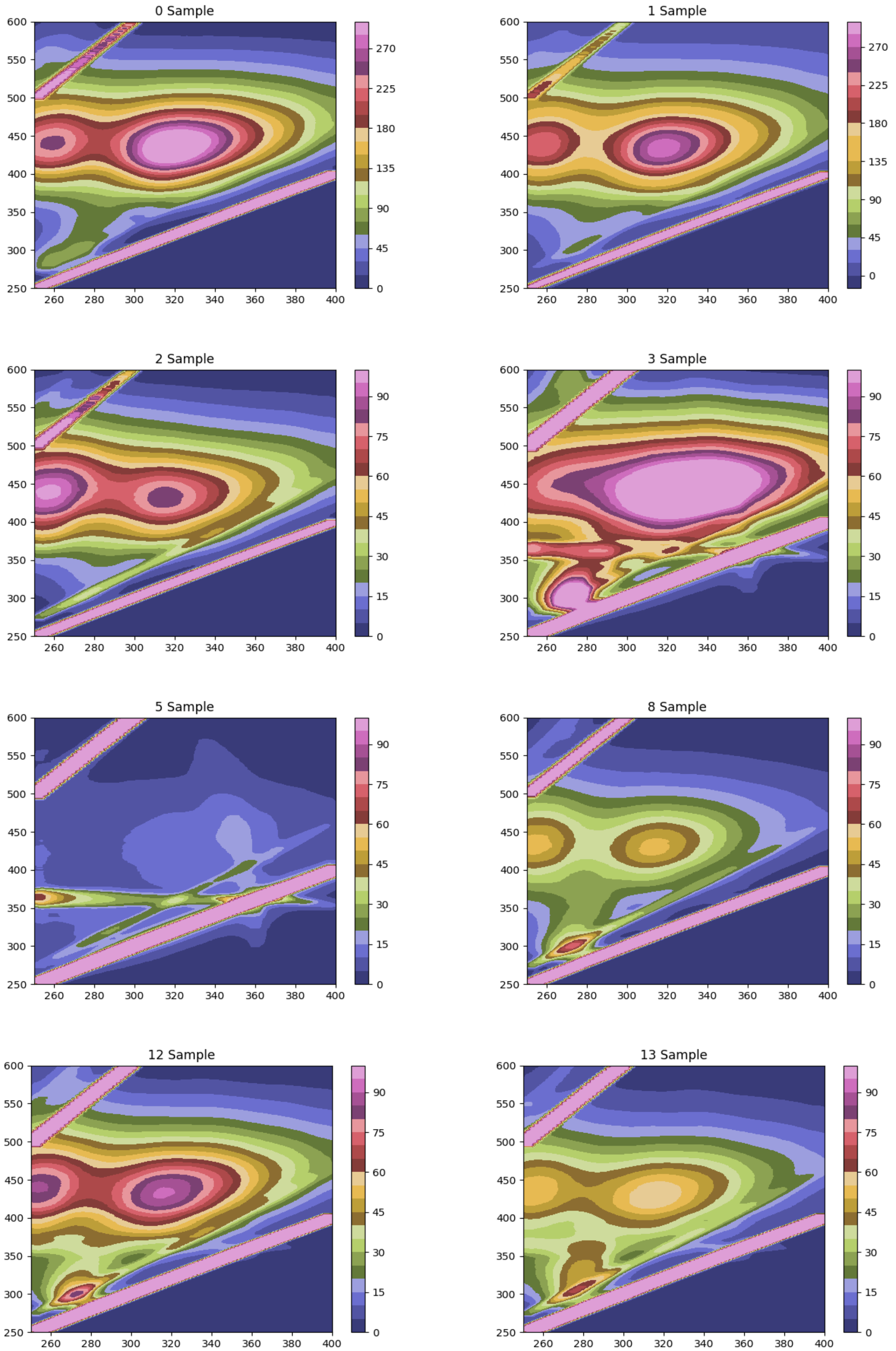

Рис. 1: Примеры образцов флуориметрии

4.1 Главные компоненты

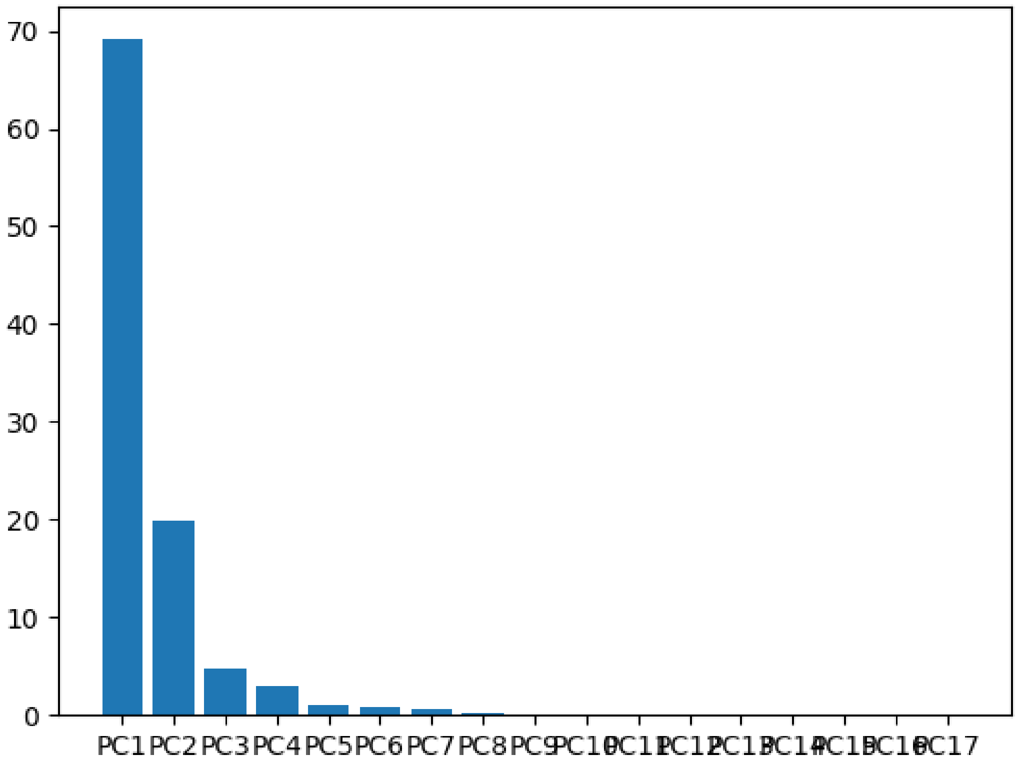

Рис. 2: Вклад компонент в суммарную дисперсию (в процентах)

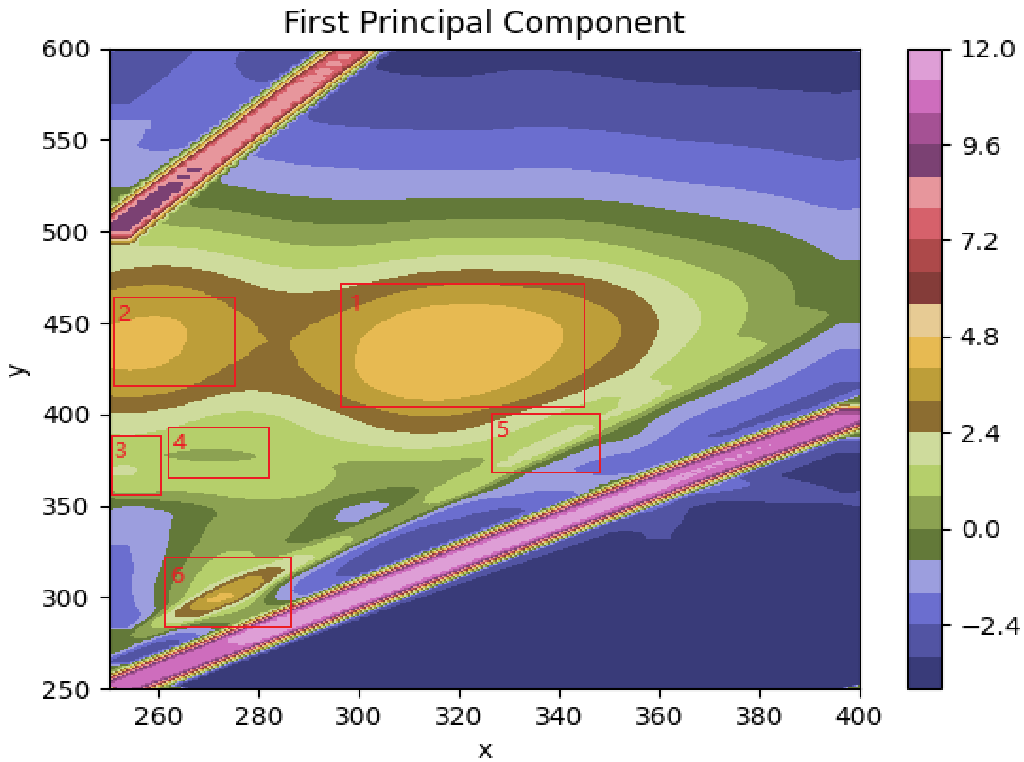

Рис. 3: Первая главная компонента

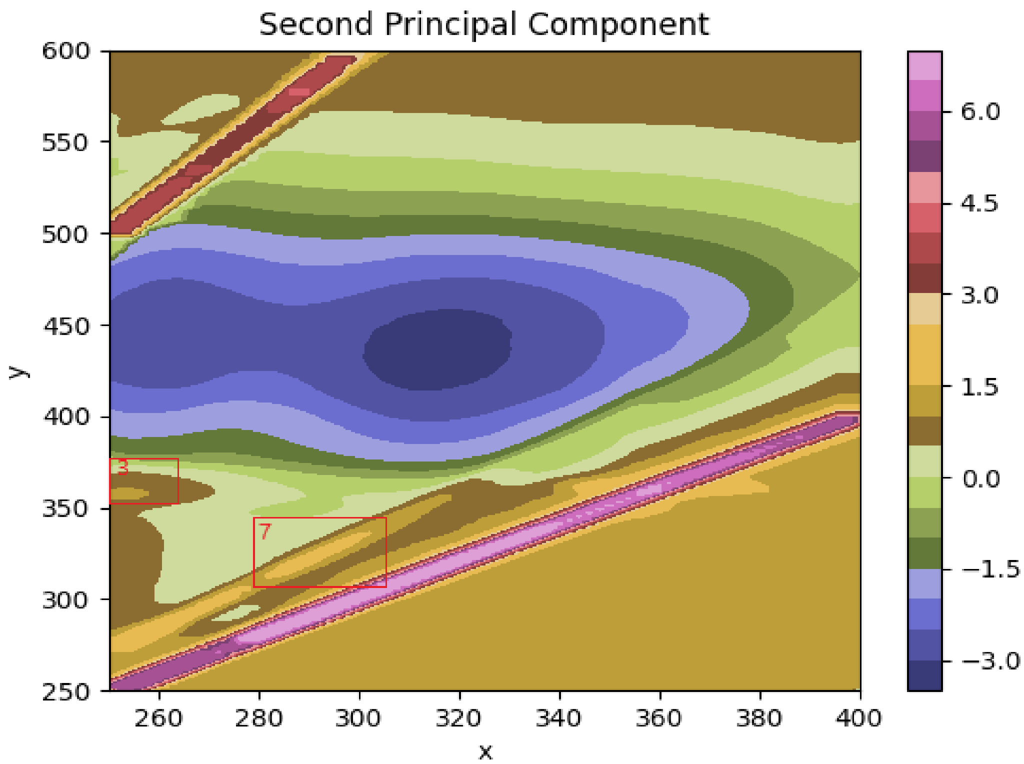

Рис. 4: Вторая главная компонента

5 Обсуждение

1. Из графика (2) видно, что для покрытия $\approx 90\%$ дисперсии исходных данных требуется первые две главные компоненты.

2. На рисунках (3), (4) линиями уровня представлены первая и вторая главные компоненты. Кроме того, эмпирически выявлены области локальной выпуклости главных компонент как функции переменных Emission's lambda(x), Excitation's lambda(y). Эти области можно интерпретировать как области наибольшей дисперсии интенсивности флуоресценции между образцами.
   Следует отметить, что данные области зачастую совпадают с областями локальных максимумов конкретных образцов. Для примера: область 1 совпадает с областью локального максимума образцов 0, 1, 3, 8, 12, 13; при этом для образца 5 эта область является областью сравнительно малой интенсивности, что объясняет большую дисперсию между образцами в области 1, отраженную в главной компоненте 1.

Репозиторий

`https://github.com/KoloskovAleksandr/MathStatLabs2021`

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.: Книжный дом «ВИБРО-КОМ», 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 с.

[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.