Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчёт по курсовой работе по дисциплине«Математическая статистика»

Выполнил студент:
Василевский Елисей Александрович
группа: 3630102/70201

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

Санкт-Петербург 2020

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Есть набор 2D данных, следы жизни в геологических объектах.Смысл двумерности следующий. На объект подается излучение, просто свет, от ближнего ультрафиолетового до видимомого. Длина волны — первая переменная x_1. Когда свет с заданной x_1 попадает в объект, его поглощают молекулы и в свою очередь, излучают свет с длинами волны x_2 примерно в том же диапазоне. То, что они излучают записывается в виде графика $I(x_1 = const, x_2)$ это обычный график. Далее, x_1 варьируются, и формируется $I(x_1, x_2)$. Функция 2-х переменных.Пики на графике I можно идентифицировать с излучением протеиногенных аминокислот, т.е. это остатки органической жизни, хоть во льдах, хоть на метеоритах. Известна область для каждой аминокислоты в координатах (x_1, x_2).

Есть данные из Арктики и центральной Африки. Они объединены в группы. Для каждой пробы посчитать интегралы интенсивности искомых аминокислот. Дальше — выяснить, есть разница между группами или нет.

2 Теория

2.1 Корреляционный момент (ковариация) и коэффициент корреляции

Корреляционным моментом, иначе ковариацией, двух случайных величин X и Y называется математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий .

$$K = cov(X,Y) = M[(X - x)(Y - y)] \tag{1}$$

Коэффициентом корреляции ρ двух случайных величин X и Y называется отношение их корреляционного момента к произведению их средних квадратических отклонений:

$$\rho = \frac{K}{\sigma_x \sigma_y} \tag{2}$$

Коэффициент корреляции — это нормированная числовая характеристика, являющаяся мерой близости зависимости между случайными величинами к линейной

2.2 Выборочные коэффициенты корреляции

2.2.1 Выборочный коэффициент корреляции Пирсона

Пусть по выборке значений $\{x_i,y_i\}_1^n$ двумерной с.в. (X, Y) требуется оценить коэффициент корреляции $\rho = \frac{K}{\sqrt{DXDY}}$. Естественной оценкой для ρ служит его статистический аналог в виде выборочного коэффициента корреляции, предложенного К.Пирсоном, —

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2 \frac{1}{n}\sum(y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}, \tag{3}$$

где K, s_X, s_Y — выборочные ковариация и дисперсии с.в. X и Y .

2.3 Метод главных компонент

Способ снижения размерности данных, который преобразует большое число скоррелированных переменных, называемых главными компонентами.

Пусть есть матрица входных данных

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... \\ x_{N1} & x_{N2} & ... & x_{Nk} \end{bmatrix}$$

Выполняется предварительная стандартизация данных.

$$x'_{ij} = \frac{x_{ij} - \bar{X}_j}{\sigma(X_j)}, i = \overline{1,N}, j = \overline{1,k} \tag{4}$$

$$X' = \begin{bmatrix} x'_{11} & x'_{12} & ... & x'_{1k} \\ x'_{21} & x'_{22} & ... & x'_{2k} \\ ... & ... & ... & ... \\ x'_{N1} & x'_{N2} & ... & x'_{Nk} \end{bmatrix}$$

Вычисляется ковариационная матрица нормированных данных.

$$cov(X') = \begin{bmatrix} 1 & r'_{12} & ... & r'_{1k} \\ r'_{21} & 1 & ... & r'_{2k} \\ ... & ... & ... & ... \\ r'_{N1} & r'_{N2} & ... & 1 \end{bmatrix}$$

r_{ij}- коэффициент корреляции между i-ым и j-ым признаками

Замечание. Условие для использования метода главных компонент - коррелируемость признаков.

Строится линейная комбинация нормированных исходных признаков с определенными весовыми коэффициентами (главный компонент):

$$Z_j = a_{1j}X'_1 + a_{2j}X'_2 + ... + a_{kj}X'_k \tag{5}$$

где a_{ij} - весовые коэффициенты.

Условия накладываемые на главные компоненты

1. Дисперсии разброса проекций объектов на главные компоненты должны быть убывающими:

$$\sigma^2(Z_1) > \sigma^2(Z_2) > ... > \sigma^2(Z_k)$$

2. Главные компоненты линейно независимы:

$$cov(Z) = \begin{bmatrix} \sigma^2(Z_1) & 0 & ... & 0 \\ 0 & \sigma^2(Z_2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma^2(Z_k) \end{bmatrix}$$

3. Суммы выборочных дисперсий по исходным признакам(нормированные) и главным компонентам равны:

$$\sum_{j=1}^{k} \sigma^2(Z_j) = \sum_{j=1}^{k} \sigma^2(X'_j)$$

Для нахождения весовых коэффициентов нужно найти собственные значения корреляционной матрицы наблюдаемых показателей:

$$det(cov(X') - \lambda E) = 0 \tag{6}$$

и найти соответствующие собственные вектора

$$a_j = (a_{1j},a_{2j},...,a_{kj})^T, |a_j| = 1 \tag{7}$$

которые по сути и являются весовыми коэффициентами, т.е можно сокращенно записать

$$Z = X'A \tag{8}$$

где

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1k} \\ a_{21} & a_{22} & ... & a_{2k} \\ ... & ... & ... & ... \\ a_{k1} & a_{k2} & ... & a_{kk} \end{bmatrix}$$

3 Реализация

Курсовая работа выполнена с помощью встроенных средств языка программирования R в среде разработки RStudio.

4 Результаты

4.1 Выборочные коэффициенты корреляции для главных компонент полученных из исходных данных

№	образец из Арктики	образец из Африки	Коэффициент корреляции (r)
1	1701	1.3_68	0.823
2	1701	2.4_7	0.847
3	1701	3.1_14	0.783
4	1701	3.2_69	0.794
5	1701	4.1_45	0.835
6	1701	4.3_84	0.831
7	1702	1.5_11	0.794
8	1702	3.1_14	0.814
9	1702	3.5_43	0.766
10	1702	4.3_84	0.797
11	1702	4.6_88	0.755
12	1704	3.1_14	0.847
13	1704	4.1_45	0.755
14	1704	4.3_84	0.800
15	1711	1.2_21	0.808
16	1711	1.6_37	0.759
17	1711	2.4_7	0.833
18	1711	3.1_14	0.812
19	1711	4.6_88	0.831
20	1728	1.2_21	0.849
21	1728	4.3_84	0.812
22	1732	1.5_11	0.788
23	1732	2.4_7	0.820
24	1732	3.1_14	0.831
25	1732	4.3_84	0.777
26	1732	4.6_88	0.824

Таблица 1: Образцы у которых коэффициент корреляции $r > 0.75$

№	образец из Арктики	образец из Африки	Коэффициент корреляции (r)
1	1701	3.5_43	0.882
2	1701	5.2_2	0.856
3	1701	5.5_28	0.866
4	1701	5.6_95	0.862
5	1702	2.4_7	0.870
6	1702	4.1_45	0.852
7	1704	1.2_21	0.912
8	1704	1.5_11	0.896
9	1704	2.4_7	0.878
10	1704	4.6_88	0.875
11	1706	5.4_92	0.882
12	1711	1.5_11	0.854
13	1711	4.1_45	0.854
14	1712	4.4_87	0.913
15	1712	5.3_66	0.863
16	1727	4.2_80	0.855
17	1728	1.5_11	0.903
18	1728	2.4_7	0.917
19	1728	3.1_14	0.879
20	1728	4.1_45	0.886
21	1728	4.6_88	0.855
22	1729	1.2_21	0.901
23	1729	1.5_11	0.932
24	1729	2.4_7	0.915
25	1729	3.1_14	0.942
26	1729	4.1_45	0.882
27	1729	4.3_84	0.863
28	1729	4.6_88	0.913
29	1730	2.3_5	0.938
30	1730	3.4_20	0.903
31	1732	4.1_45	0.909

Таблица 2: Образцы у которых коэффициент корреляции $r > 0.85$

№	образец из Арктики	образец из Африки	Коэффициент корреляции (r)
1	1712	2.3_5	0.974
2	1712	3.4_20	0.980
3	1730	4.4_87	0.975

Таблица 3: Образцы у которых коэффициент корреляции $r > 0.95$

4.2 Изображения данных с максимальными выборочными коэффициентами корреляции

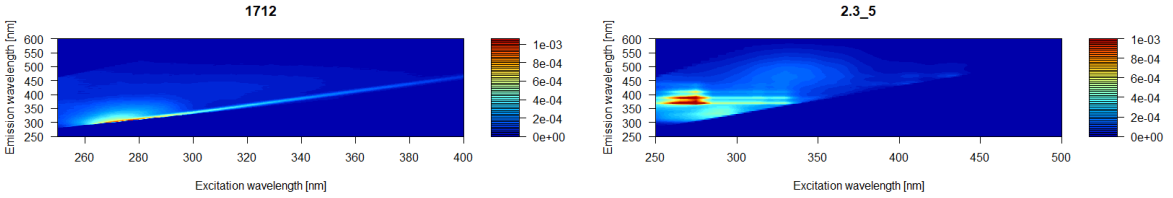

Рис. 1: Образцы 1712 и 2.3_5 , $r = 0.974$

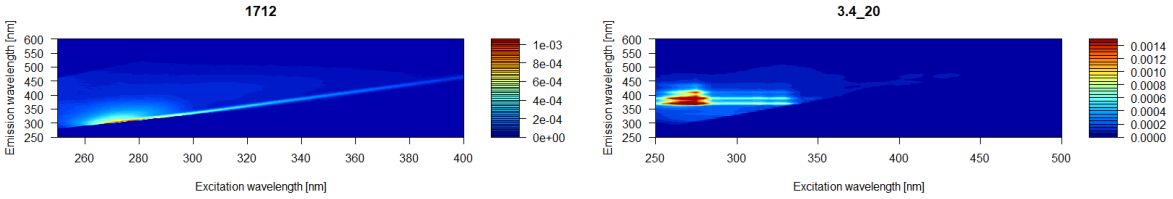

Рис. 2: Образцы 1712 и 3.4_20 , $r = 0.980$

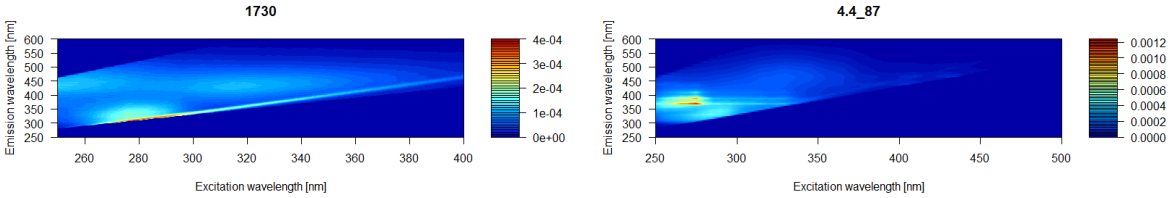

Рис. 3: Образцы 1730 и 4.4_87 , $r = 0.975$

4.3 Выборочные коэффициенты корреляции для главных компонент после выделения области наибольшей интенсивности

В качестве области наибольшей интенсивности выбрана область с границами [250 300; 250 450].

№	образец из Африки	образец из Арктики	Коэффициент корреляции (r)
1	1.4_114	1712	0.847
2	1.4_114	1727	0.767
3	2.3_5	1733	0.849
4	3.4_20	1706	0.761
5	3.5_43	1702	0.846
6	3.5_43	1704	0.782
7	3.5_43	1711	0.827
8	3.5_43	1728	0.839
9	3.5_43	1729	0.811
10	4.1_45	1701	0.826
11	4.3_84	1711	0.830
12	4.3_84	1729	0.780
13	4.5_108	1702	0.802
14	4.5_108	1711	0.805
15	4.5_108	1728	0.841
16	4.6_88	1701	0.841
17	5.1_90	1701	0.791
18	5.2_2	1732	0.830
19	5.3_66	1727	0.778
20	5.4_92	1734	0.777
21	5.5_28	1701	0.794
22	5.5_28	1702	0.824
23	5.5_28	1704	0.807
24	5.5_28	1711	0.845
25	5.5_28	1728	0.824
26	5.5_28	1729	0.818
27	5.6_95	1702	0.778
28	5.6_95	1704	0.758
29	5.6_95	1711	0.801
30	5.6_95	1728	0.785
31	5.6_95	1729	0.769

Таблица 4: Образцы у которых коэффициент корреляции $r > 0.75$

№	образец из Африки	образец из Арктики	Коэффициент корреляции (r)
1	1.1_70	1701	0.900
2	1.1_70	1704	0.882
3	1.1_70	1711	0.937
4	1.2_21	1701	0.906
5	1.2_21	1704	0.932
6	1.4_114	1730	0.874
7	1.5_11	1701	0.899
8	1.5_11	1704	0.939
9	1.6_37	1702	0.856
10	1.6_37	1704	0.931
11	1.6_37	1711	0.874
12	1.6_37	1728	0.883
13	1.6_37	1729	0.924
14	2.4_7	1701	0.916
15	2.4_7	1702	0.938
16	2.4_7	1704	0.854
17	2.4_7	1711	0.896
18	2.4_7	1728	0.933
19	2.4_7	1729	0.915
20	3.1_14	1701	0.912
21	3.1_14	1704	0.909
22	3.1_14	1711	0.946
23	3.4_20	1730	0.935
24	3.4_20	1733	0.856
25	3.5_43	1701	0.877
26	4.1_45	1702	0.919
27	4.1_45	1711	0.927
28	4.1_45	1728	0.943
29	4.3_84	1701	0.904
30	4.3_84	1702	0.880
31	4.3_84	1728	0.856
32	4.4_87	1730	0.890
33	4.5_108	1704	0.912
34	4.5_108	1729	0.876
35	4.6_88	1702	0.939
36	4.6_88	1704	0.927
37	4.6_88	1711	0.943
38	5.3_66	1730	0.909
39	5.4_92	1712	0.904
40	5.5_28	1732	0.890
41	5.6_95	1732	0.896

Таблица 5: Образцы у которых коэффициент корреляции $r > 0.85$

№	образец из Африки	образец из Арктики	Коэффициент корреляции (r)
1	1.1_70	1702	0.963
2	1.1_70	1728	0.975
3	1.1_70	1729	0.951
4	1.2_21	1702	0.970
5	1.2_21	1711	0.955
6	1.2_21	1728	0.984
7	1.2_21	1729	0.987
8	1.5_11	1702	0.967
9	1.5_11	1711	0.956
10	1.5_11	1728	0.980
11	1.5_11	1729	0.991
12	2.3_5	1712	0.963
13	2.3_5	1730	0.967
14	3.1_14	1702	0.961
15	3.1_14	1728	0.970
16	3.1_14	1729	0.973
17	3.4_20	1712	0.953
18	4.1_45	1704	0.960
19	4.1_45	1729	0.979
20	4.4_87	1712	0.976
21	4.6_88	1728	0.967
22	4.6_88	1729	0.971
23	5.3_66	1712	0.989

Таблица 6: Образцы у которых коэффициент корреляции $r > 0.95$

5 Обсуждение

- Проанализировав результат можно сказать, что большая часть данных из одной группы имеет зависимость с данными из другой группы, т.к. коэффициент корреляции достаточно большой и стремится к 1. В связи с этим можно предположить о схожести образцов из Арктики и Африки.
- Проанализировав результат выделения исходных данных , можно сделать вывод о том ,что при обрезки области большей интенсивности количество коррелирующих образцов увеличивается.

6 Литература

- Максимов Ю. Д. Математическая статистика //СПб.: СПбГПУ. – 2004.
- Лекция по методу главных компонент // ИБ БГУ - 2020.
- Курс лекций по эконометрике //НИУ ВШЭ - 2016.

7 Приложения

- Репозиторий с исходным кодом: https://github.com/re1nex/mathstat