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
бионформатике и генетике»

Выполнил студент:
Величко Арсений
группа: 3630102/80201

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

Санкт-Петербург
2021 г.

Содержание

Страница

1 Постановка задачи **3**

2 Теория **3**
2.1 Метод главных компонент 3
2.2 k-means 3
2.3 Обработка информации о геноме 3

3 Реализация **4**

4 Результаты **4**
4.1 Бактерия Streptomyces coelicolor 4
4.2 Бактерия Caulobacter vibrioides 6

5 Обсуждение **8**

Список иллюстраций

Страница

1 Проекция на пространство первых 2 главных компонент в случае пространств 1-, 2-, 3- и 4-буквенных слов 4
2 Проекция на пространство первой главной компоненты в случае пространств 1-, 2-, 3- и 4-буквенных слов 5
3 Кластеризованный рисунок для проекции в пространстве 3-буквенных слов 6
4 Проекция на пространство первых 2 главных компонент в случае пространств 1-, 2-, 3- и 4-буквенных слов 6
5 Проекция на пространство первой главной компоненты в случае пространств 1-, 2-, 3- и 4-буквенных слов 7
6 Кластеризованный рисунок для проекции в пространстве 3-буквенных слов 8

1 Постановка задачи

Исследовать применимость метода главных компонент (principal component analysis, PCA) в биоинформатике и генетике, в частности, при анализе геномных последовательностей. Визуально представить полученные результаты.

2 Теория

2.1 Метод главных компонент

Пусть дана p-мерная выборка $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$.

1. Составим ковариационную матрицу C:

$$C = \mathbf{E}\left[(\boldsymbol{X} - \mathbf{E}\left[\boldsymbol{X}\right])(\boldsymbol{X} - \mathbf{E}\left[\boldsymbol{X}\right])^T\right]$$

2. Так как $C = C^T$ и $\forall z \in \mathbb{R}^p \; z^T C z > 0$, у C существует спектральное разложение вида

$$C = V \Lambda V^T. \tag{1}$$

   Столбцы V представляют собой ортонормированные собственные вектора матрицы C.

3. Проекцией изначальной выборки на множество первых k главных компонент назовем величину

$$\widehat{X} = \boldsymbol{X} V_k, \tag{2}$$

   где V_k составлена из первых k столбцов матрицы V.

2.2 k-means

Пусть дана p-мерная выборка $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, которую нужно разбить на k ($k \leq n$) непересекающихся подгрупп (кластеров). В качестве начального приближения возьмем k различных точек из $\boldsymbol{X}$. Положим их за $\{c_l^{(0)}\}_{l=1}^k$ (центры). Теперь разобьем все точки из $\boldsymbol{X}$ на k кластеров $K_1^{(0)}, \ldots, K_k^{(0)}$. Номер m кластера $K_m^{(0)}$, в который следует отнести точку x из выборки $\boldsymbol{X}$, определяется следующим образом:

$$m = \arg\min_{1 \leq l \leq k} \rho(c_l^{(0)}, x), \tag{3}$$

где ρ - евклидова метрика на $\mathbb{R}^p$.
Рассмотрим общий (i-ый) шаг алгоритма:

1. Рассчитать $\{c_l^{(i)}\}_{l=1}^k$:

$$\{c_l^{(i)}\}_{l=1}^k = \left\{\mathbf{E}\left[K_l^{(i-1)}\right]\right\}_{l=1}^k. \tag{4}$$

2. Выйти, если центры кластеров не изменились. В ином случае рассчитать новые кластеры.

2.3 Обработка информации о геноме

Информация, необходимая живой клетке для функционирования, зашифрована в одной длинной молекуле ДНК. Эта информация может быть представлена в виде непрерывного текста, содержащего всего 4 буквы - A, C, G, T (отражающие одно из 4 азотистых оснований, встречающихся в ДНК). Отдельные участки геномной последовательности, имеющие особенную структуру (начинающиеся и заканчивающиеся с определенных трехбуквенных комбинаций), называются генами.
Ниже привёден фрагмент геномной последовательности:

GAATTCTTAACGTCCTGAGACACGACAGCGACCTC
TGACCGGACTCGTTCCGCGTCTTTGGACAATCGGG
ATTCAGACTTCGGGGGATGCGGCGCAGGCTTGGGG
ATGATAGGCGAGCAATGCGACCGTTGATCACAGCG

Назовем *словом* любую непрерывную подпоследовательность букв в тексте. Так как текст не содержит разделителей, то разбиение на слова данного текста не единственно. Будем поочередно разбивать текст на 1-, 2-, 3-, 4-буквенные слова. Количество возможных уникальных слов каждой длины равно соответственно $4^1, 4^2, 4^3, 4^4$.
Пусть имеется текст, состоящий из m букв, который необходимо проанализировать на содержание t-буквенных слов ($1 \leq t \leq 4$). Разобьем текст на k фрагментов $X_1, \ldots, X_k$ длиной $l \leq \frac{m}{k}$. В каждом фрагменте посчитаем количество каждого из 4^t t-буквенных слов. Размерность каждого X_i таким образом равна 4^t. К полученной выборке применим метод главных компонент.

3 Реализация

Лабораторная работа выполнена в среде MATLAB версии R2019b с использованием следующих пакетов:

- Statistics and Machine Learning Toolbox.

4 Результаты

В опытах текст делится на фрагменты длиной $l \leq 300$.

4.1 Бактерия Streptomyces coelicolor

Результаты для генома бактерии Streptomyces coelicolor:

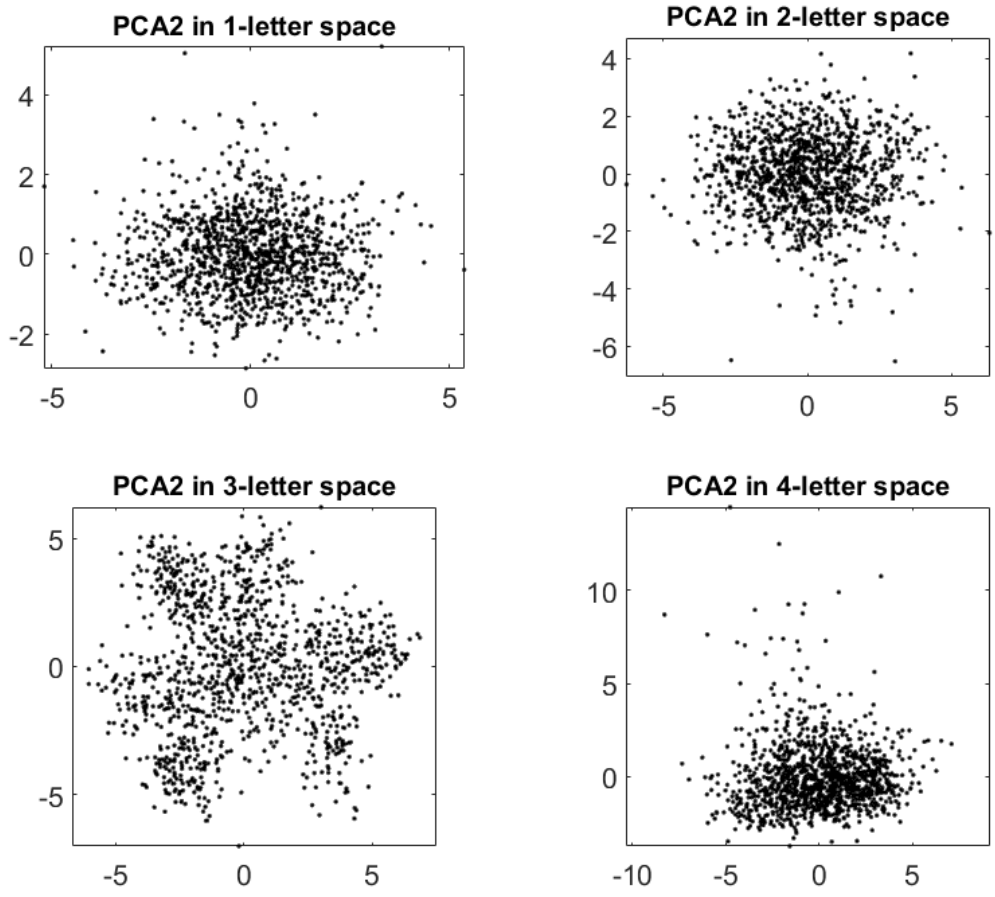

Рис. 1: Проекция на пространство первых 2 главных компонент в случае пространств 1-, 2-, 3- и 4-буквенных слов

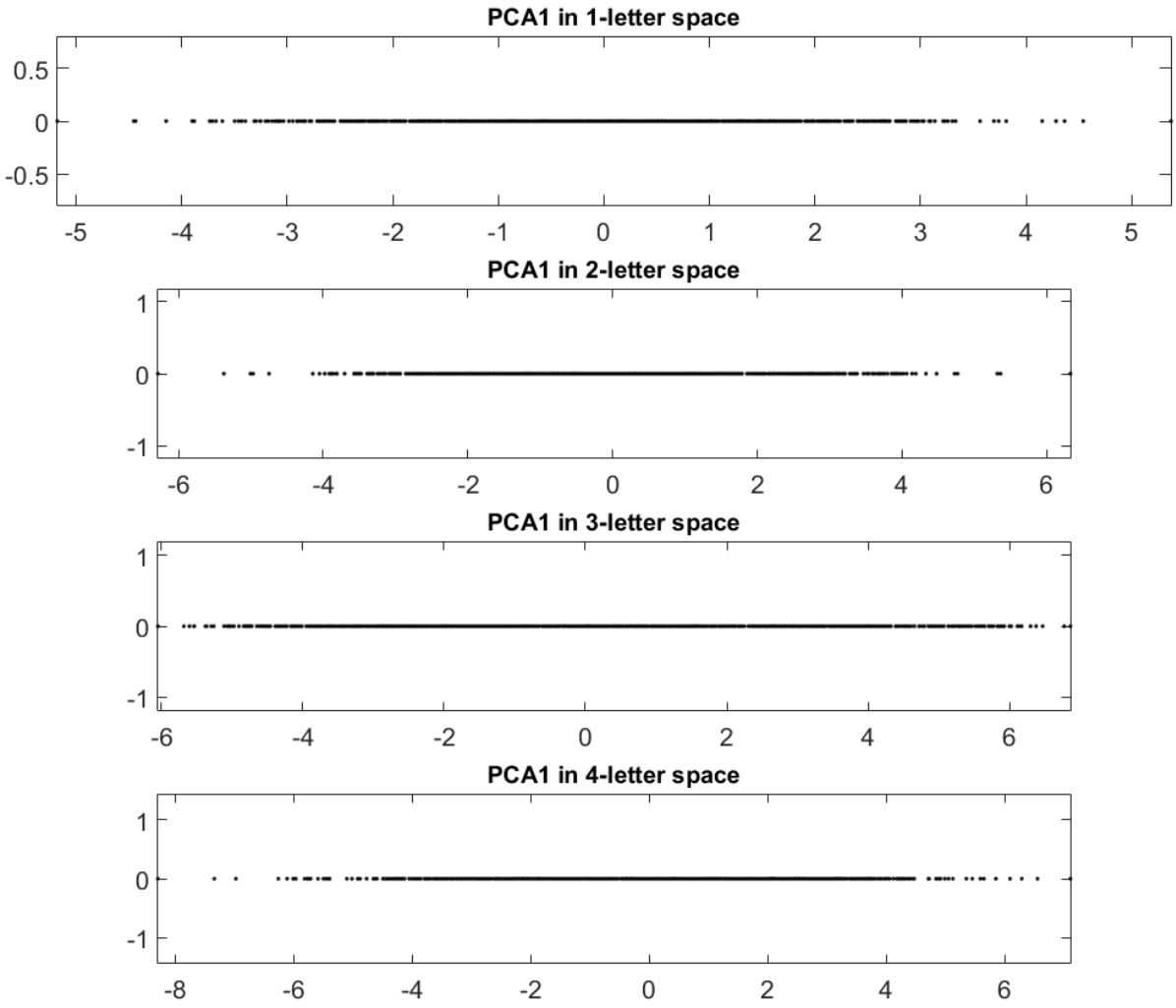

Рис. 2: Проекция на пространство первой главной компоненты в случае пространств 1-, 2-, 3- и 4-буквенных слов

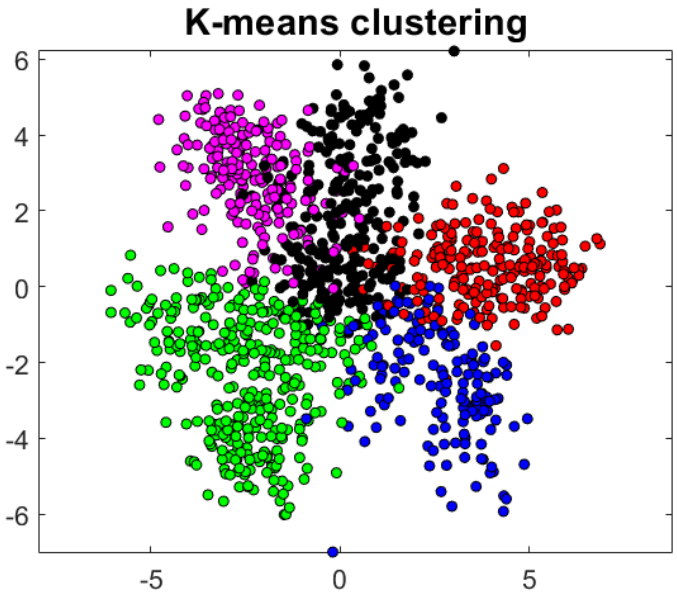

Рис. 3: Кластеризованный рисунок для проекции в пространстве 3-буквенных слов

4.2 Бактерия Caulobacter vibrioides

Результаты для генома бактерии Caulobacter vibrioides:

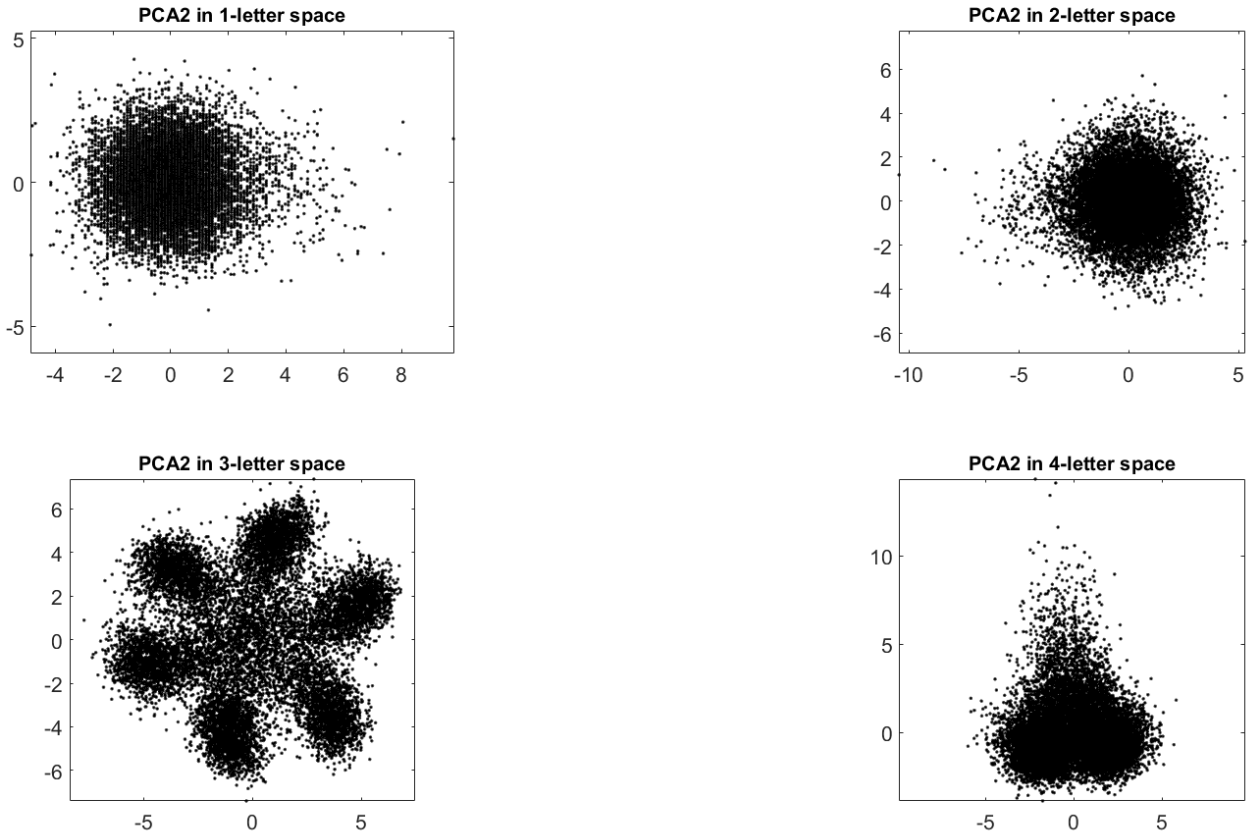

Рис. 4: Проекция на пространство первых 2 главных компонент в случае пространств 1-, 2-, 3- и 4-буквенных слов

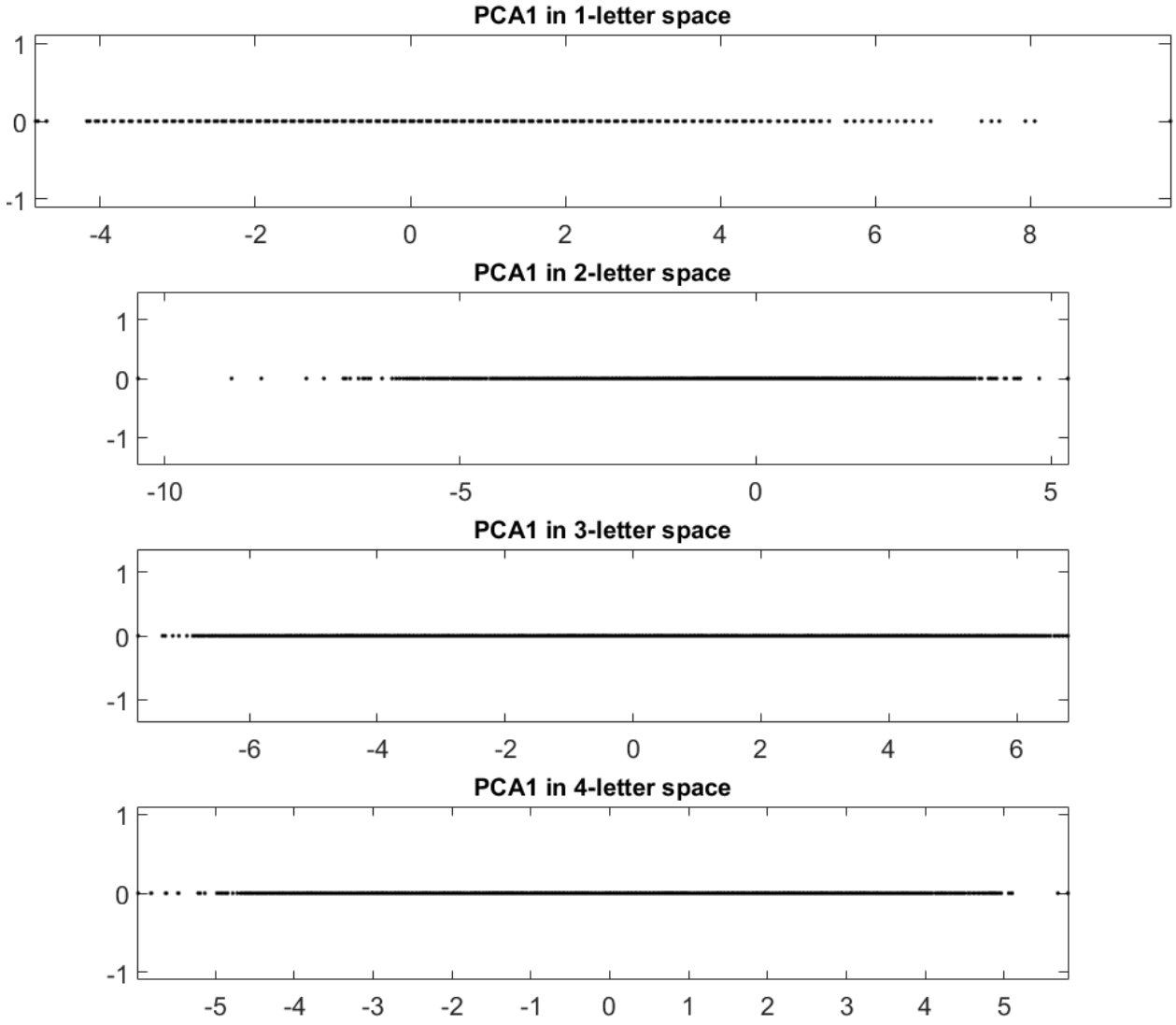

Рис. 5: Проекция на пространство первой главной компоненты в случае пространств 1-, 2-, 3- и 4-буквенных слов

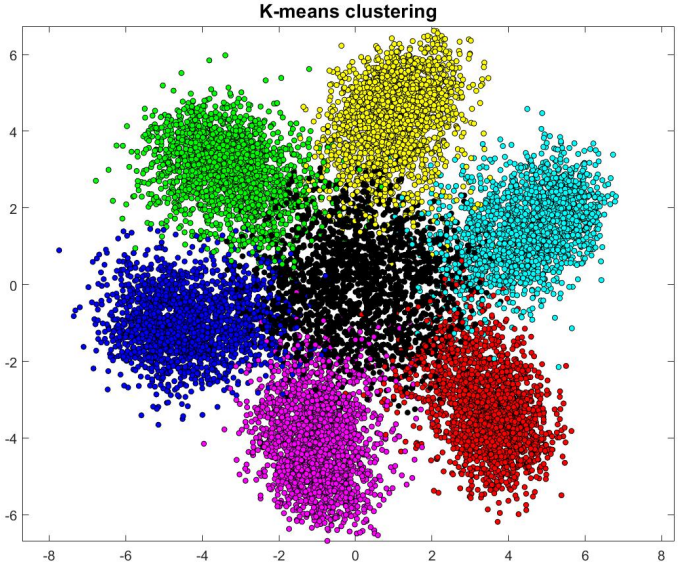

Рис. 6: Кластеризованный рисунок для проекции в пространстве 3-буквенных слов

5 Обсуждение

Основной вывод, который можно сделать из полученных изображений состоит в том, что геномный текст содержит информацию, которая кодируется неперекрывающимися тройками, потому что график, соответствующий тройкам, как видно, сильно структурирован, в отличие от изображений синглетов, дуплетов и четверок. Подробнее о результатах кластеризации, получаемых в такого типа исследованиях, можно почитать в [3].

Примечание

С кодом работы можно ознакомиться по ссылке: `https://github.com/ArsenyVelichko/MathStat`.

Список литературы

[1] Genbank FTP-site: `https://www.ncbi.nlm.nih.gov/genbank/ftp/`.

[2] Jackson, J.: A User's Guide to Principal Components (Wiley Series in Probability and Statistics). Wiley-Interscience (2003).

[3] Gorban, A.N., Zinovyev, A.Yu., and Popova, T.G.: Four basic symmetry types in the universal 7-cluster structure of 143 complete bacterial genomic sequences. In Silico Biology 5 (2005). URL: `https://arxiv.org/abs/q-bio/0410033`.