Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и вычислительной физики
Кафедра «Прикладная математика»

КУРСОВАЯ РАБОТА ПО ТЕМЕ: «МЕТОД ГЛАВНЫХ КОМПОНЕНТ» ПО ДИСЦИПЛИНЕ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»

Выполнил студент
Войнова Алёна Игоревна
группы 3630102/80201

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

Санкт-Петербург
2021

Содержание

Список таблиц

Список иллюстраций

1 Постановка задачи

- Провести анализ данных, которые представленны датасетом [3].
- Применить метод главных компонент на представленных данных

2 Теория

Метод Главных Компонент (англ. Principal Components Analysis, PCA) — один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Изобретен К. Пирсоном (англ.Karl Pearson) в 1901 г.

Применяется во многих областях, таких как распознавание образов, компьютерное зрение, сжатие данных и т. п. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных или к сингулярному разложению матрицы данных.

Иногда метод главных компонент называют преобразованием Кархунена-Лоэва (англ. Karhunen-Loeve)[1] или преобразованием Хотеллинга (англ. Hotelling transform).

2.1 Формальное описание

Пусть имеется матрица переменных X размерностью (IJ), где I – число образцов (строк), а J – это число независимых переменных (столбцов), которых, как правило, много $(J >> 1)$. В методе главных компонент используются новые, формальные переменные $t_a(a = 1, \dots A)$, являющиеся линейной комбинацией исходных переменных $x_j(j = 1, \dots J)$

$$t_a = p_{a1}x_1 + \cdots + p_{aJ}x_J$$

С помощью этих новых переменных матрица X разлагается в произведение двух матриц T и P

$$X = TP^T + E = \sum_A a = 1(t_a P_a^t + E)$$

Матрица T называется матрицей счетов (scores). Ее размерность (IA). Матрица P называется матрицей нагрузок (loadings). Ее размерность (JA). E – это матрица остатков, размерностью (IJ)

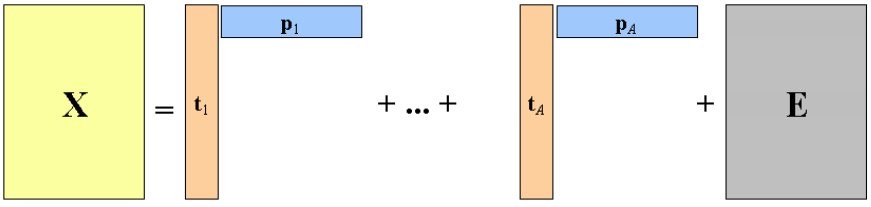

Рис. 1: Разложение по главным компонентам

Новые переменные t_a называются главными компонентами (Principal Components), поэтому и сам метод называется методом главных компонент (PCA). Число столбцов – t_a в матрице T, и p_a в матрице P, равно A, которое называется числом главных компонент (PC). Эта величина заведомо меньше числа переменных J и числа образцов I.

Важным свойством PCA является ортогональность (независимость) главных компонент. Поэтому матрица счетов T не перестраивается при увеличении числа компонент, а к ней просто прибавляется еще один столбец – соответствующий новому направлению. Тоже происходит и с матрицей нагрузок P.

2.2 Алгоритм

Чаще всего для построения PCA счетов и нагрузок, используется рекуррентный алгоритм, который на каждом шагу вычисляет одну компоненту. Сначала исходная матрица X преобразуется и превращается в матрицу E_0, $a = 0$. Далее применяют следующий алгоритм.

1. Выбрать начальный вектор t
2. $pt = t^t E_a / t^t t$
3. $p = p/(p^t p)$
4. $t = E_a p / p^t p$
5. Проверить сходимость, если нет, то идти на 2

После вычисления очередной (a-ой) компоненты, полагаем $t_a = t$ и $p_a = p$. Для получения следующей компоненты надо вычислить остатки $E_{a+1} = E_a \breve{} t p^t$ и применить к ним тот же алгоритм, заменив индекс a на $a + 1$.

После того, как построено пространство из главных компонент, новые образцы X_{new} могут быть на него спроецированы, иными словами – определены матрицы их счетов T_{new}. В методе PCA это делается очень просто $T_{new} = X_{new} P$

3 Реализация

Курсовая работа выполнена с помощью средств языка программирования **Python** в среде разработки **Jupyter**. Исходный код лабораторной работы приведён в приложении.

4 Результаты

4.1 Анализ данных

Датасет состоит из морфологических данных и данных о стабильных изотопах 13 видов печеночных птиц Cinclodes, а также метаданные музейных коллекций для

каждого образца[2]. Датасет включает в себя 439 экземпляров птиц, каждый из которых описан 80 признаками.

Данные содержает следующую информацию: локализация образца (страна, регион, местность), вес, размерные данные (перьев, крыльев, головы, туловища), музеи, в которых представлен экземпляр и изотопные характеристики.

Для построения следующих графиков были выбраны все образцы определенного вида и взяты средние значения их характеристик.

4.2 Исследование характеристик в зависимости от вида птиц

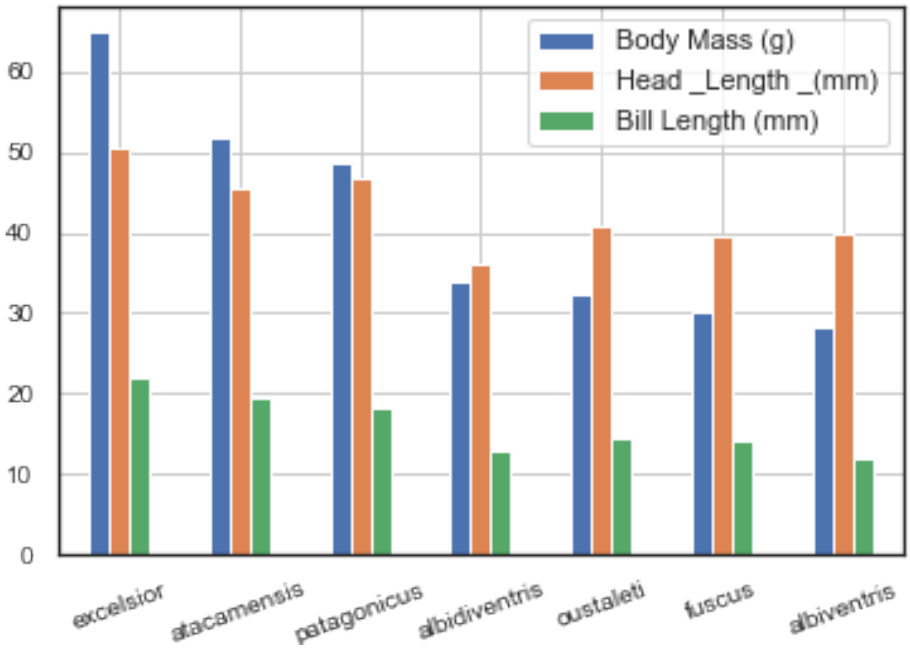

Рис. 2: Зависимость размеров птиц от вида

Вид **excelsior** имеет самые внушительные размеры, в то время как вид **albiventris** проигрывает в этих характеристиках.

4.3 Исследование характеристик в зависимости от страны обитания птиц

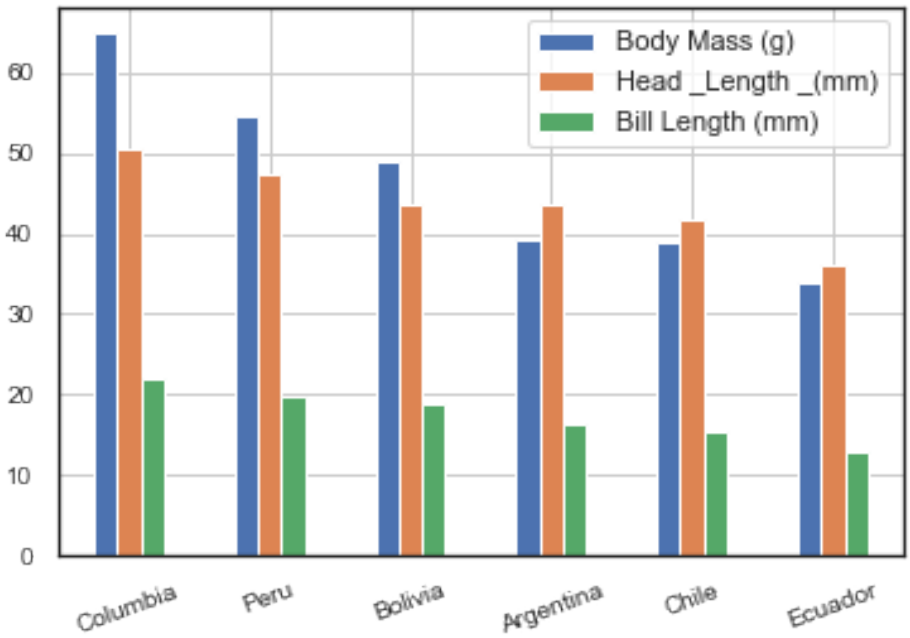

Рис. 3: Зависимость размеров птиц от страны

В **Колумбии** можно увидеть самых крупных птиц, а птицы - жители **Эквадора** достаточно компактны по своим размерам.

4.4 Корелляция данных

Рис. 4: Корреляция выборочных данных

4.5 PCA - метод главных компонент

При подготовке данных признаки, не имеющие численного представления (строки) были заменены на численное представление, а некоторые удалены. Так же были убраны строки с неизвестным неизвестными параметрами (NaN). В итоге остались 65 признаков и 34 элемента датасета.
Дальше данные стандартизировались, для корректной работы метода главных компонент.

4.5.1 Визуализация основных компонент

Применён PCA на подготовленный набор данных.

Пример показывает как количество компонент влияет на разделимость данных. График PC3 и PC4 явно не может разделить каждый класс, тогда как PC1 и PC2 показывает четкое разделение между каждым видом.

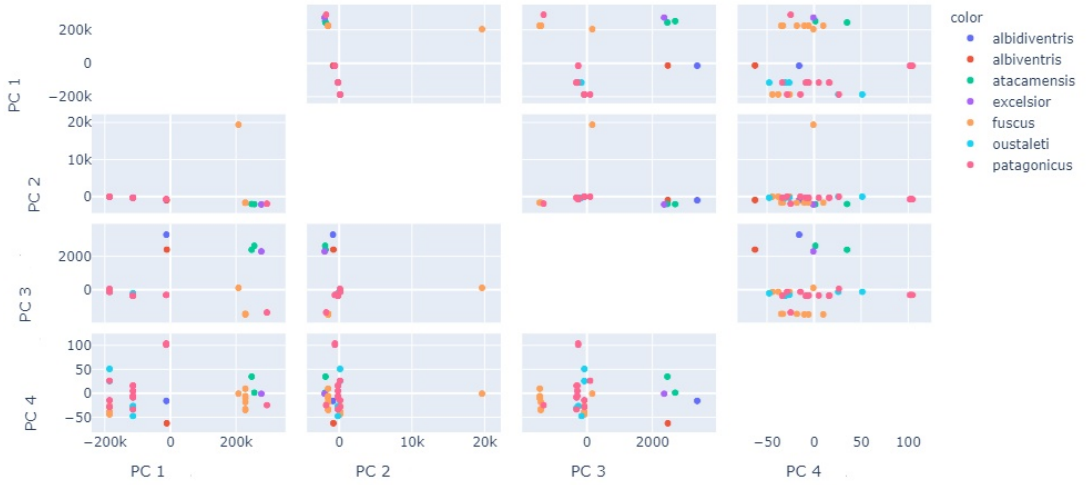

Рис. 5: Корреляция выборочных данных

4.5.2 Зависимость доли общей информации от количества компонент

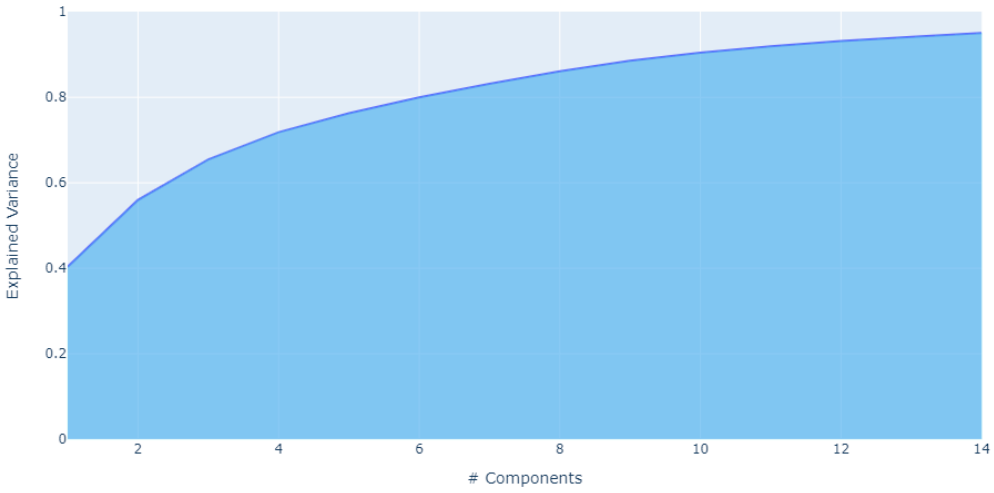

Рис. 6: Корреляция выборочных данных

1	2	3	4	5	6	7
0.416487	0.158065	0.096908	0.061427	0.040055	0.036198	0.032920
8	9	10	11	12	13	14
0.026988	0.023770	0.01928	0.014615	0.009893	0.009483	0.007696

Таблица 1: Доля информация в 14 компонентах

5 Обсуждение

В результате применения метода главных при попытке сохранить 95% информации получается 14 векторов (1), что в 4 раза меньше, чем количество признаков в исходных данных при потере информации лишь в 5%.
При этом, если рассматривать 23 компоненты, сохранятся все 99% информации.

- Преимущества PCA:

    1. Метод позволяет облегчить работу с данными, уменьшив число факторов, требующих внимания
    2. Помогает в построении более устойчивых моделей, выполняемых быстрее, чем было бы возможно для исходных входных полей.

- Недостатки PCA:

    1. Возможность непреднамеренного пренебрежения важными параметрами
    2. Использует ортогональную систему координат, что не всегда приводит к лучшим результатам

6 Приложения

URL: Выполненная лабораторная работа на GitHub
https://github.com/pikabol88/Math-Statistics/tree/main/coursework

Список литературы

[1] Gorban A. N., Kegl B., Wunsch D., Zinovyev A. Y. (Eds.), Principal Manifolds for Data Visualisation and Dimension Reduction, Series: Lecture Notes in Computational Science and Engineering 58, Springer, Berlin — Heidelberg — New York, 2007, XXIV, 340 p. 82 illus. ISBN 978-3-540-73749-0.

[2] Data from: Isotopic niches support the resource breadth hypothesis, http://en.wikipedia.org/wiki/Business_logic_layer, December 13, 2017.

[3] Data from: Isotopic niches support the resource breadth hypothesis, http://en.wikipedia.org/wiki/Business_logic_layer, December 13, 2017.

[4] PCA Visualization in Python https://plotly.com/python/pca-visualization/#what-about-dash