МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное   
учреждение высшего образования   
«Южный федеральный университет»

Институт математики, механики   
и компьютерных наук им. И. И. Воровича

Кафедра информатики и вычислительного эксперимента

Григорян Георгий Зоргевич

**Использование стресс-функции для определения размерности пространства характеристик в задачах DATA MINING**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА  
по направлению подготовки  
02.03.02 – Фундаментальная информатика и информационные технологии

**Научный руководитель** –   
доц., к. ф.-м. н. Нестеренко Виктор Александрович

Допущено к защите:  
заведующий кафедрой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Михалкович С.С.

**Оглавление**

Оглавление

[Введение 3](#_Toc104843007)

1. [Постановка задачи 3](#_Toc104843008)

[2. Принцип работы алгоритма 3](#_Toc104843009)

[3. Реализация алгоритма 4](#_Toc104843010)

[3.1 Подготовка данных 4](#_Toc104843011)

[3.2 Вычисление ковариационной матрицы 5](#_Toc104843012)

[3.3 Извлечение собственных векторов и собственных чисел 7](#_Toc104843013)

# Введение

На протяжении многих лет объемы данных, предоставленных людям, стремительно увеличиваются. Из-за этого перед человечеством встает проблема извлечения полезной информации из них. Поэтому на помощь к людям приходит такая технология как Data Mining. Data Mining — это интеллектуальный анализ данных. Данная технология пришла на замену прикладной статистике, следовательно отсюда проистекает изобилие методов и алгоритмов. Сам же термин “Data Mining” часто переводится как добыча данных, извлечение информации. Одной из важных задач в Data Mining является уменьшение размерности. Есть множество методов, позволяющих то сделать, но в своей работе я опишу и реализую Метод Главных Компонент либо же Анализ Основных Компонент (PCA, Principal Component Analysis).

# 1.Постановка задачи

Как говорилось ранее, в своей работе я выбрал и реализовал Метод Главных Компонент для редукции размерности пространства данных.

Основные этапы выполнения задания:

1. Выбор метода уменьшения размерности.
2. Изучить материалы и реализовать метод.
3. Обработка данных.

# 2. Принцип работы алгоритма

Это способ выявления закономерностей в данных и выражения данных таким образом, чтобы подчеркнуть их сходства и различия. Поскольку закономерности может быть трудно найти в данных большой размерности, где роскошь графического представления недоступна, PCA является мощным инструментом для анализа данных.

# 3. Реализация алгоритма

Вычисление главных компонент может быть сведено к вычислению [сингулярного разложения](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D0%B7%D0%BB%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5) матрицы данных или к вычислению [собственных векторов](https://ru.wikipedia.org/wiki/%D0%A1%D0%BE%D0%B1%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B2%D0%B5%D0%BA%D1%82%D0%BE%D1%80) и [собственных значений](https://ru.wikipedia.org/wiki/%D0%A1%D0%BE%D0%B1%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5_%D0%B7%D0%BD%D0%B0%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) [ковариационной матрицы](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%B2%D0%B0%D1%80%D0%B8%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BC%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0) исходных данных.  В моем случае алгоритм реализован через нахождение собственных значений и векторов ковариационной матрицы на языке C#.

## Подготовка данных

Для демонстрации работы алгоритма был сгенерирован data set состоящий из вещественных чисел ([Рис. 1](#_Подготовка_данных)). Размерность пространства данных равнялось 4 признакам, такая размерность взята для проверки редукции к разному количеству главных компонент.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 1.0 | 2.73446908 | 11.01792737 | 98.47615116 |
| 2.0 | 4.35122722 | 12.38619263 | 20.46282367 |
| 3.0 | 7.21132988 | 11.48804931 | 67.17093415 |
| 4.0 | 11.24872601 | 24.10099224 | 77.66304739 |
| 5.0 | 9.58103444 | 30.21117481 | 142.25100662 |
| 6.0 | 12.09865079 | 32.83617975 | 127.60618803 |
| 7.0 | 13.78706794 | 42.4237342 | 199.82062948 |
| 8.0 | 13.85301221 | 51.14366228 | 110.97707692 |
| 9.0 | 15.29003911 | 47.1998583 | 218.21847896 |
| 10.0 | 18.0998018 | 60.19540684 | 269.12150528 |

Рис. 1. Таблица сгенерированных данных.

Для корректной работы PCA необходимо центрировать данные, т.е. вычесть из каждого значения столбца среднее арифметическое этого столбца (Рис. 2). Таким образом среднее арифметическое центрированных данных будет равняться нулю.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| -4,50000000 | -8,09106677 | -21,28239040 | -34,70063301 |
| -3,50000000 | -6,47430863 | -19,91412514 | -112,71396050 |
| -2,50000000 | -3,61420597 | -20,81226846 | -66,00585002 |
| -1,50000000 | 0,42319016 | -8,19932553 | -55,51373678 |
| -0,50000000 | -1,24450141 | -2,08914296 | 9,07422245 |
| 0,50000000 | 1,27311494 | 0,53586198 | -5,57059614 |
| 1,50000000 | 2,96153209 | 10,12341643 | 66,64384531 |
| 2,50000000 | 3,02747636 | 18,84334451 | -22,19970725 |
| 3,50000000 | 4,46450326 | 14,89954053 | 85,04169479 |
| 4,50000000 | 7,27426595 | 27,89508907 | 135,94472111 |

Рис. 2. Таблица центрированных данных.

## Вычисление ковариационной матрицы

Для начала стоит разобраться что такое ковариация. Ковариация или корреляционный момент — мера зависимости одной случайной величины от другой. В нашем случае формула, по которой вычисляется ковариация значений двух измерений будет выглядеть так:

Рис.3. Формула ковариации двух векторов.

где , среднее арифметическое значений измерений X, Y.

Ковариационная матрица для n-мерного пространства признаков будет выглядеть так:

Рис.4 Ковариационная матрица

где одно из возможных измерений.

Рис.5 Пример ковариационной матрицы для данных с размерностью пространства.

В матрице коэффициентов ковариационной матрицы имеет значение знаки этих коэффициентов. Если знак – это:

* положительное число, то две переменные прямо пропорциональны, то есть второй увеличивается или уменьшается вместе с первым.
* отрицательное число, то переменные обратно пропорциональны, то есть второй увеличивается, когда первый уменьшается, и наоборот.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 9,16666666 | 14,34448508 | 52,34245948 | 193,94202309 |
|  | 14,34448508 | 23,84180375 | 81,40260965 | 300,65742902 |
|  | 52,34245947 | 81,40260965 | 312,47012667 | 1122,85505928 |
|  | 193,94202309 | 300,65742902 | 1122,85505928 | 5789,75302839 |

Рис. 5. Ковариационная матрица сгенерированных данных данных.

## 3.3 Извлечение собственных векторов и собственных чисел

Что бы определить Главные Компоненты необходимо экстраполировать собственные векторы (eigenvectors) и собственные числа (eigenvalues) из ковариационной матрицы. Главная Компонента ­­­­­­­– это новая переменная, комбинированная таким образом, что новые переменные не коррелированы между собой, и основная информация об исходных переменных помещается в первых компонентах.