*Белорусский государственный университет*

*Факультет радиофизики и компьютерных технологий*

**Лабораторная работа №3**

*«Метод главных компонент»*

Работу выполнил

Редько Александр

3 курс, 5КБ

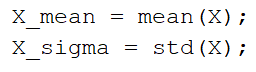
2023

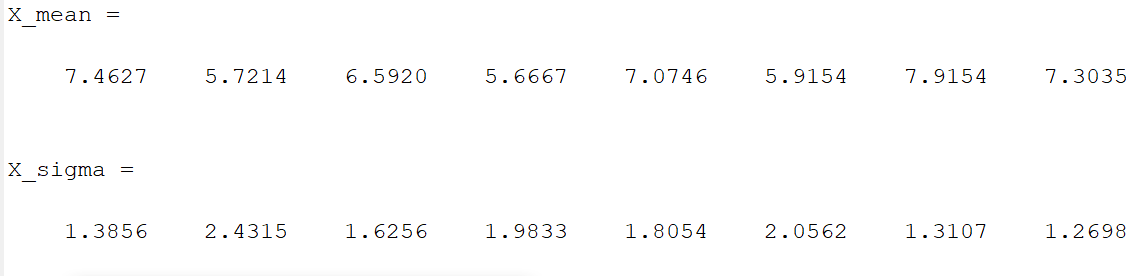
**Цель работы:** практическое освоение метода главных компонент для

решения задач снижения размерности и визуализации многомерных данных.

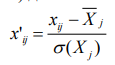
**Ход работы:** в качестве исходных данных для выполнения данной лабораторной работы был использован текстовый файл «data12.txt» из дополнений к лабораторному практикуму. В данном файле содержалась матрица 201 на 8.

Сначала необходимо провести нормировку данных для устранения неоднородности, поэтому были вычислены мат. ожидания и среднеквадратические отклонения по столбцам.

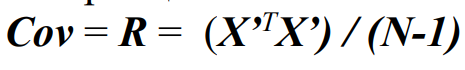




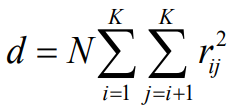
И произвести вычисления по следующей формуле:



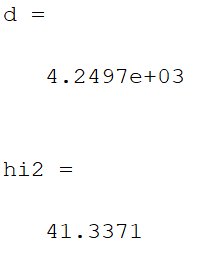
На следующем шаге необходимо построить ковариационную матрицу. Ввиду произведенной нормализации данных матрица ковариаций будет корреляционной матрицей исходных данных R порядка K × K.:



Прежде чем начинать анализ главных компонент целесообразно проверить, значимо ли отличается от единичной матрицы корреляционная матрица исходных нормированных данных. В предположении, что исходные данные подчиняются многомерному нормальному распределению, можно воспользоваться статистикой

,

где rij – наддиагональные элементы корреляционной матрицы R. Статистика d подчиняется χ2 -распределению с K(K-1)/2 степенями свободы. Если корреляционная матрица исходных данных не отличается от единичной матрицы, т. е. d ≤ χ2, вычисленное при заданном уровне доверительной вероятности и заданном числе степеней свободы, то применение метода главных компонент нецелесообразно.

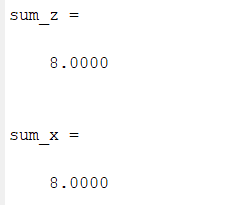


Оказалось, что d >hi2 (по формуле chi2inv) на 2 порядка.

С помощью eig необходимо найти собственные значения ковариационной матрицы и собственные векторы, что впоследствии позволило рассчитать проекции объектов на главные компоненты.

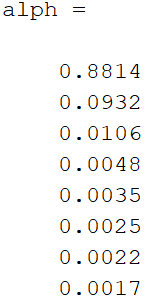
Чтобы удостовериться в правильности данных вычислений, необходимо посчитать сумму дисперсий. При выполнении следующего условия, вычисления прошли без ошибок: сумма выборочных дисперсий исходных признаков равна сумме выборочных дисперсий проекций объектов на главные компоненты



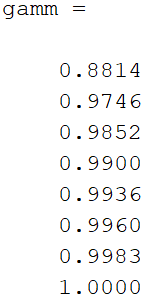


Кроме этого, рассчитываются:

а) Относительная доля разброса, приходящаяся на j-ую главную компоненту:



б) Относительная доля разброса, приходящаяся на i первых компонент:



И на основе первых двух компонент строится график рассеивания:



**Выводы:**

Таким образом, метод главных компонент позволяет описать большой набор признаков K небольшим числом главных компонент M, M << K, при этом различия между объектами зависят от доли изменчивости, связанной с данной главной компонентой. На практике мы научились это делать с помощью пакета Matlab на примере предложенных данных

**Контрольные вопросы:**

1. Для каких задач обработки экспериментальных данных используется метод главных компонент?

В задачах, связанных с наглядным представлением данных (визуализация данных), снижением размерности данных без существенной потери информативности данных или (сжатием данных), стремлением к лаконизму исследуемых данных (упрощение данных). Например, распознавание образов, компьютерное зрение, сжатие данных.

1. Чему равно математическое ожидание и дисперсия стандартизованной переменной?

М[] = нулю, а дисперсия единице.

1. Какова дисперсия i-й главной компоненты?

Соответствующему ей собственному значению корреляционной матрицы.

1. Какова относительная доля разброса, приходящаяся на j-ую главную компоненту?



1. Какова относительная доля разброса, приходящаяся на i первых главных компонент?



1. Как проверить значимость корреляционной матрицы исходных данных?

Вычислить статистику d. Статистика d подчиняется χ2 -распределению с K(K-1)/2 степенями свободы. Если корреляционная матрица исходных данных не отличается от единичной матрицы, т. е. d ≤ χ2, вычисленное при заданном уровне доверительной вероятности и заданном числе степеней свободы, то применение метода главных компонент нецелесообразно.

1. Какова интерпретация первых двух главных компонент?

Они содержат более 95% основной информации.

1. Для каких целей используются функции MATLAB: chi2inv, eig, flipud, fliplr, scatter?

chi2inv(p,nu) возвращает обратную кумулятивную функцию распределения хи-квадрат со степенями свободы nu, оцененную для значений вероятности в p.

e = eig(A) возвращает вектор-столбец, содержащий собственные значения квадратной матрицы A.

B = flipud(A) возвращает A с перевернутыми строками в направлении вверх-вниз (то есть вокруг горизонтальной оси). Если A – вектор-столбец, то возвращает вектор той же длины с обратным порядком его элементов. Если вектор-строка, то возвращает без изменений. Для многомерных массивов оперирует плоскостями, образованными первым и вторым измерениями.

fliplr – аналогично, только лево-право.

scatter(x,y) создает точечную диаграмму с круглыми маркерами в местах, указанных векторами x и y. Чтобы построить один набор координат, задайте x и y как векторы равной длины. Чтобы построить несколько наборов координат на одном и том же наборе осей, задайте по крайней мере один из x или y в качестве матрицы.