МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных систем

Факультет автоматики и вычислительной техники

Кафедра систем автоматизации управления

**Метод главных компонент: составление фондового индекса**

Отчет по лабораторной работе №2  
по дисциплине

«Большие данные»

Выполнил студент гр. ИТб-4302-02-00 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ /Вершинин П.А./

(Подпись)

Руководитель ст. преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ /Родионов К.В./

(Подпись)

Работа защищена с оценкой «\_\_\_\_\_\_\_\_\_\_\_» «\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2023 г.

Цель: научиться использовать метод главных компонент (PCA) для создания улучшенного индекса Доу-Джонса и анализа данных о стоимости акций 30 крупнейших компаний США.

В ходе работы необходимо выполнить следующее задание:

* загрузить данные;
* обучить модели РСА;
* применить РСА к данным и извлечь первую компоненту;
* загрузить информацию об индексе Доу-Джонса и вычислить корреляцию;
* определить компании с наибольшим весом в первой компоненте.

Для начала загружаем данные o стоимости акций 30 компаний на закрытии торгов за каждый день в файле «close\_prices.csv». Используем библиотеку «pandas» для этой цели и извлекаем только столбцы с ценами акций, исключая столбец с датой. Чтобы загрузить данные необходимо выполнить следующий код:

|  |
| --- |
| import pandas as pd  def get\_data(data):  del data['date']  return data  data\_close\_prices =pd.read\_csv('close\_prices.csv')  data\_close\_prices =get\_data(data\_close\_prices) |

Вызываем функцию для удаления столбца «date» из загруженных ранее данных.

Метод главных компонент понижает размерность данных оптимальным с точки зрения сохранения дисперсии способом. Количество новых признаков является основным параметром метода главных компонент. Одним из методов является выбор компонент, объясняющих не менее определенной доли дисперсии.

Обучаем PCA, указывая количество компонент (в данном случае 10) и вычисляем, скольких компонент хватит, чтобы объяснить 90% дисперсии данных. Для этого необходимо выполнить следующий код:

|  |
| --- |
| from sklearn.decomposition import PCA  import numpy as np  model = PCA(n\_components=10)  model.fit(data\_close\_prices)  print(f"Для объяснения 90% дисперсии достаточно {np.argmax(np.cumsum(model.explained\_variance\_ratio\_) >= 0.9)+1} компонент") |

Для объяснения 90% дисперсии достаточно 4 компонент.

Применяем обученное преобразование PCA к исходным данным и извлекаем значения первой компоненты, выполнив следующий код:

|  |
| --- |
| transformed = model.transform(data\_close\_prices)  first\_component = transformed[:, 0] |

На рисунке 1 показана часть первой компоненты.

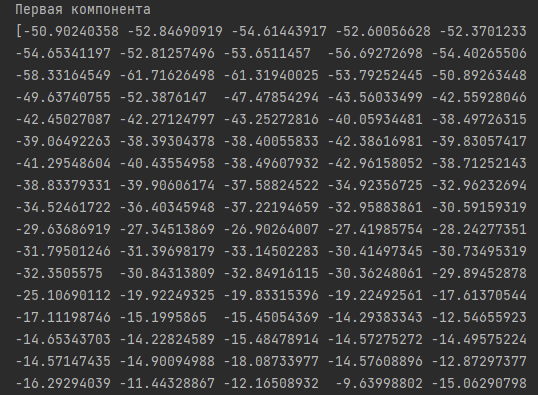


Рисунок 1 – Часть первой компоненты

Далее загружаем информацию об индексе Доу-Джонса из файла «djia\_index.csv» и вычисляем корреляцию Пирсона между первой компонентой и индексом Доу-Джонса. Для этого необходимо выполнить следующий код:

|  |
| --- |
| data\_djia\_index = pd.read\_csv('djia\_index.csv')  djia\_index = data\_djia\_index['^DJI']  correalence = np.corrcoef(first\_component, djia\_index)[0,1] |

Корреляция Пирсона между первой компонентой и индексом Доу-Джонса равна 0.91.

На последнем этапе определяем компанию с наибольшим весом в первой компоненте, что позволяет понять, какая компания оказывает наибольшее влияние на эту компоненту. Для этого необходимо выполнить следующий код:

|  |
| --- |
| weights = model.components\_[0]  max\_weight\_index = np.argmax(weights)  company\_with\_max\_weight = data\_close\_prices.columns[1:][max\_weight\_index]  max\_weight = weights[max\_weight\_index]  print(f"Компания с наибольшим весом в первой компоненте: {company\_with\_max\_weight}")  print(f"Вес компании в первой компоненте: {max\_weight:.2f}") |

На рисунке 2 представлен результат выполнения кода.



Рисунок 2 – Результат анализа

Аббревиатура компании с наибольшим весом в первой компоненте – «V». Вес данной компании в первой компоненте равен 0,58.

В ходе выполнения лабораторной работы была продемонстрирована возможность применения метода главных компонент для анализа фондового рынка. Данный метод позволяет выявлять скрытые взаимосвязи между акциями различных компаний и определять наиболее влиятельные компании.