МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Машинное обучение»

Тема: Понижение размерности пространства признаков

Студент гр. 8303 _____ Гришин К. И.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Ознакомиться с методами понижения размерности данных из библиотеки *Scikit Learn*.

Ход выполнения работы

Загрузка данных

1. Загрузить датасет по ссылке: https://www.kaggle.com/uciml/glass. Данные представлены в виде csv таблицы.
2. Создать Python скрипт. Загрузить датасет в датафрейм, и разделить данные на описательные признаки и признак отображающий класс. (отсортированные данные приведены на рис. 1).

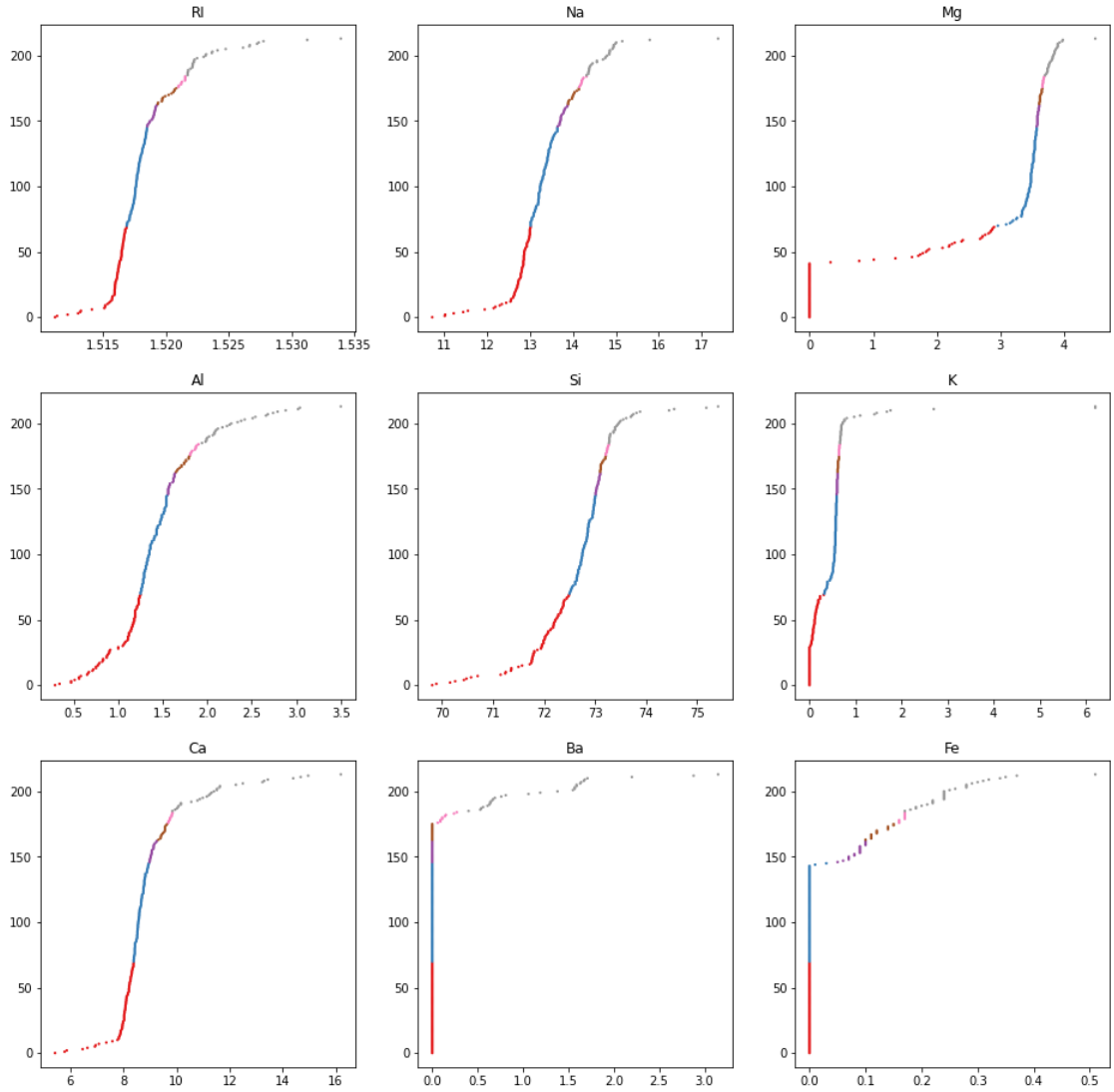

Рисунок 1. Отсортированные данные.

3. Провести нормировку данных к интервалу [0 1].
4. Построить диаграммы рассеяния для пар признаков (рис. 2). Самостоятельно определите соответствие цвета на диаграмме и класса в датасете (рис. 3)

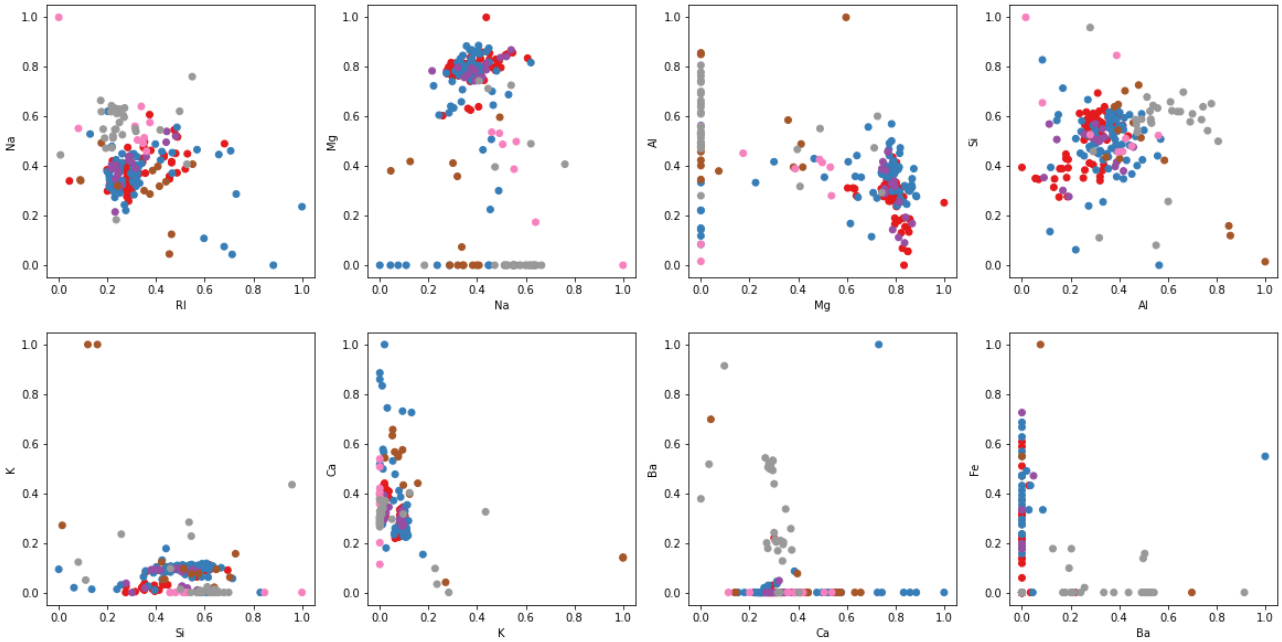

Рисунок 2. Диаграммы рассеяния пар признаков.

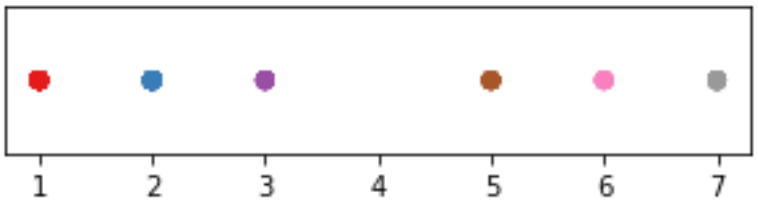

Рисунок 3. Соответствие номера класса и его цвета.

Метод главных компонент

1. Используя метод главных компонент (*PCA*). Проведите понижение размерности пространства до размерности 2.
2. Выведите значение объясненной дисперсии в процентах и собственные числа, соответствующие компонентам (табл. 1).

компонента	1	2
объясненная дисперсия в процентах	0.454296	0.179901
собственные числа	5.104931	3.212457

Таблица 1. Объясненная дисперсия и собственные числа.

3. Постройте диаграмму рассеяния после метода главных компонент (рис. 4).

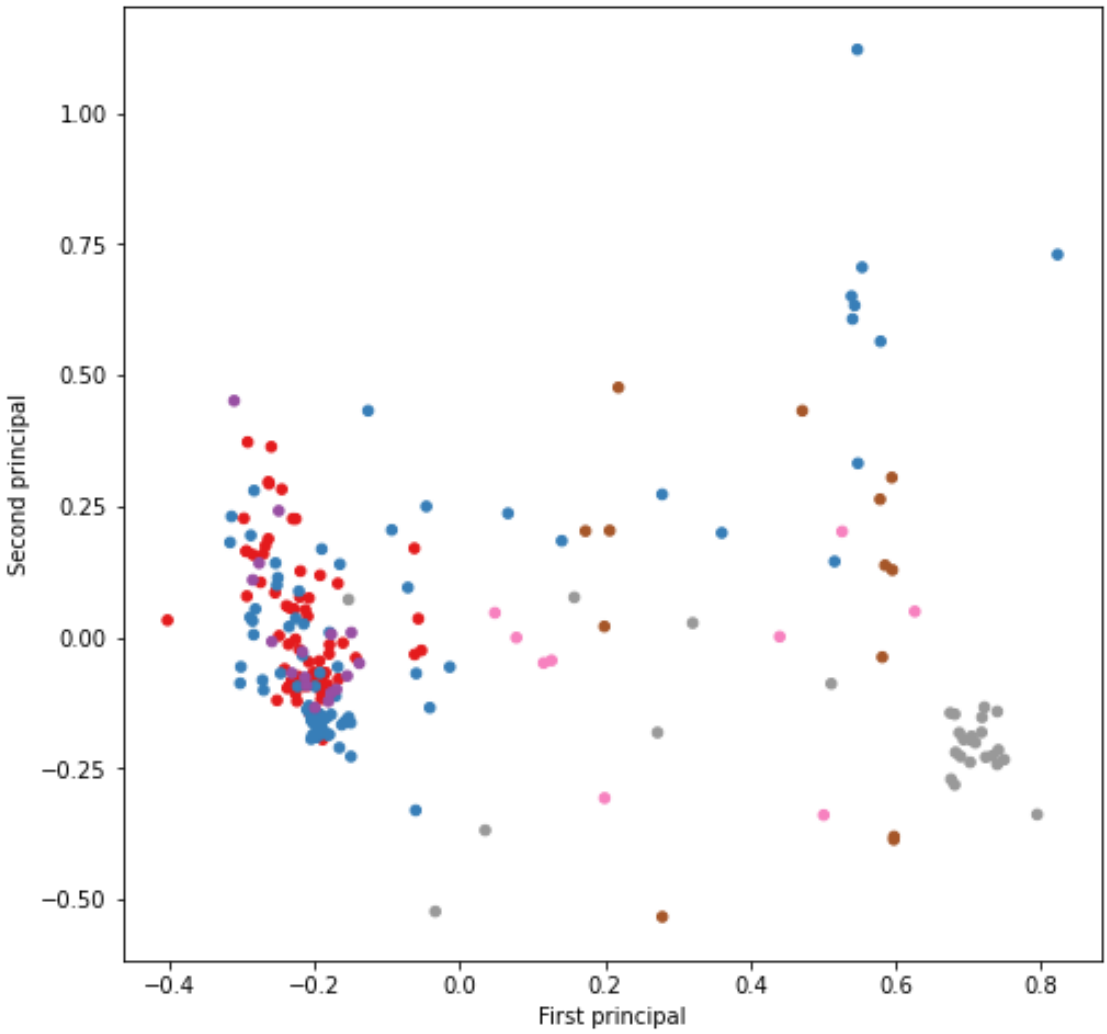

Рисунок 4. Диаграмма рассеяния первых двух главных компонент.

4. Проанализируйте и обоснуйте полученные результаты
    - Данные класса 1 сконцентрированы одним облаком в отрицательной части первой компоненты и по середине второй.
    - Данные класса 2 сконцентрированы одним облаком, а также имеют высокодисперсный разброс в сторону увеличения компонент.
    - Данные класса 3 сконцентрированы одним облаком также как и класс 1, однако имеют при этом меньшую дисперсию.
    - Данные класса 5 не сконцентрированы одним облаком, однако представляют собой две полосы параллельные второй компоненте в плоскости первых двух главных компонент.
    - Данные класса 6 также неконцентрированны, как и класс 5, однако при этом имеют большую дисперсию в плоскости первых двух главных компонент
    - Данные класса семь по большей части сконцентрированы в конце положительной части первой главной компоненты и в отрицательной части второй, однако имеют разброс в сторону уменьшения первой компоненты.

5. Изменяя количество компонент, определите количество, при котором компоненты объясняют не менее 85% дисперсии данных

1	2	3	4	5
45.43%	63.42%	76.07%	85.87%	92.73%

Таблица 2. Объясненная дисперсия для указанного количества компонент.

Исходя из таблицы 3 можно сделать вывод, что первые 4 компоненты объясняют больше 85% дисперсии данных.

6. Используя метод *inverse_transform* восстановите данные, сравните с исходными

Данные были восстановлены из первых двух компонент. Результат восстановления представлен в виде таблицы с дисперсиями (табл. 3) и в виде диаграмм рассеяния (рис. 5)

σ^2	**Ri**	**Na**	**Mg**	**Al**	**Si**	**K**	**Ca**	**Ba**	**Fe**
исх.	0.0178	0.0151	0.1032	0.0242	0.0191	0.0110	0.0175	0.0249	0.0365
восст.	0.0129	0.0034	0.1018	0.0134	0.0026	0.0008	0.0147	0.0100	0.0112

Таблица 3. Значения дисперсии исходных и восстановленных данных.

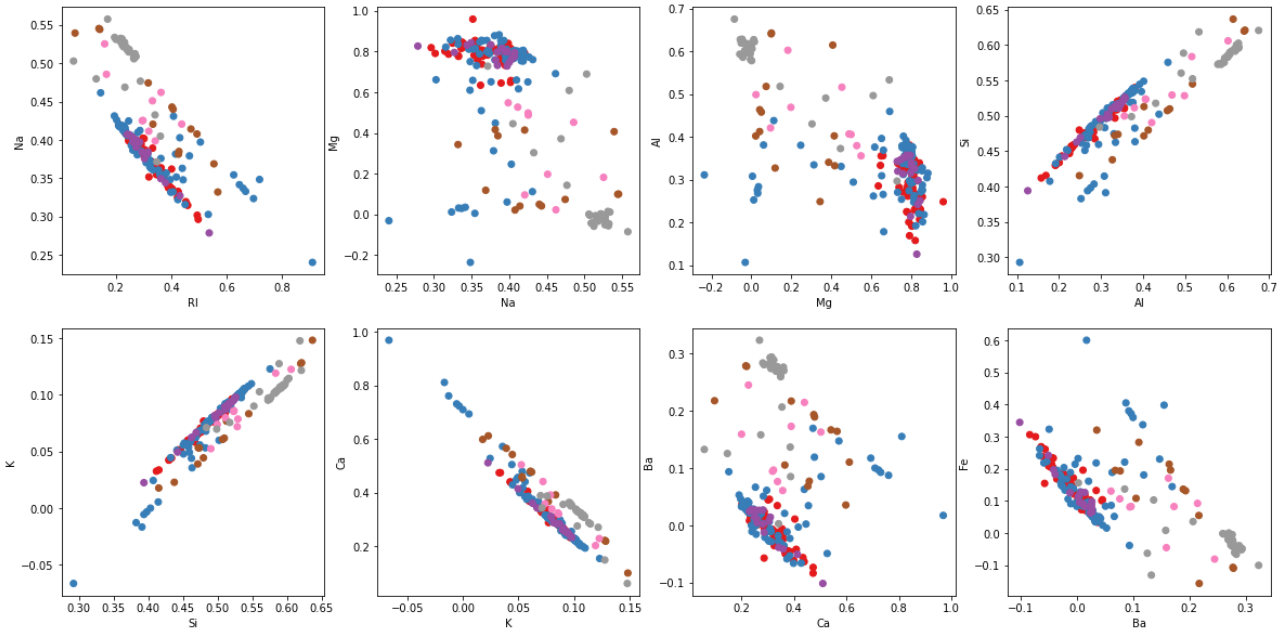

Рисунок 5. Диаграммы рассеяния после восстановления первых двух компонент.

Исходя из данных полученных в таблице 3 можно сделать вывод, что восстановление данных по первым двум компонентам для данного набора данных не имеет большого смысла. Относительно не сильно дисперсия поменялась у элементов Ri, Mg и Ca, у остальных элементов дисперсия уменьшилась в разы, что привело к сильному повреждению данных, что видно на диаграммах с рисунка 5.

Проведем такой же анализ для первых четырех компонент. Значения дисперсий для восстановленных данных указаны в таблице 4, диаграммы рассеяния – на рисунке 6.

σ^2	**Ri**	**Na**	**Mg**	**Al**	**Si**	**K**	**Ca**	**Ba**	**Fe**
исх.	0.0178	0.0151	0.1032	0.0242	0.0191	0.0110	0.0175	0.0249	0.0365
восст.	0.0169	0.0047	0.1028	0.0185	0.0168	0.0019	0.0166	0.0175	0.0356

Таблица 4. Значения дисперсии исходных и восстановленных данных.

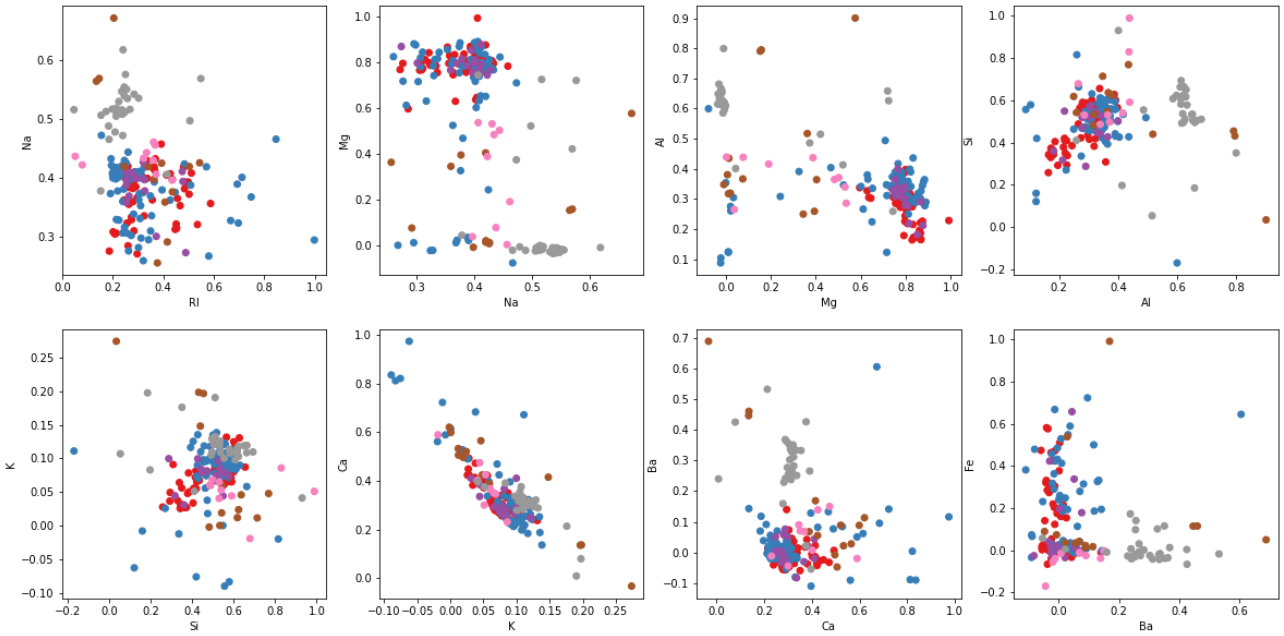

Рисунок 6. Диаграммы рассеяния после восстановления первых четырех компонент.

В случае с четырьмя компонентами восстановление данных прошло более успешно, теперь дисперсия близка к исходной уже для пяти элементов Ri, Mg, Si, Ca и Fe.

Однако несмотря на это, диаграммы рассеяния с рисунка 5 демонстрируют все еще достаточно сильное повреждение данных, хотя уже можно заметить похожие с исходными данными рисунки.

Данное повреждение данных наблюдается из потери 15% дисперсии.

7. Исследуйте метод главных компонент при различных параметрах *svd_solver*

   Параметр *svd_solver* можно установить в «*full*», «*arpack*» и «*randomized*». Различные методы *SVD* дают одинаковый результат для данного набора данных. Диаграммы рассеяния первых двух компонент для каждого метода *SVD* представлены на рисунке 7.

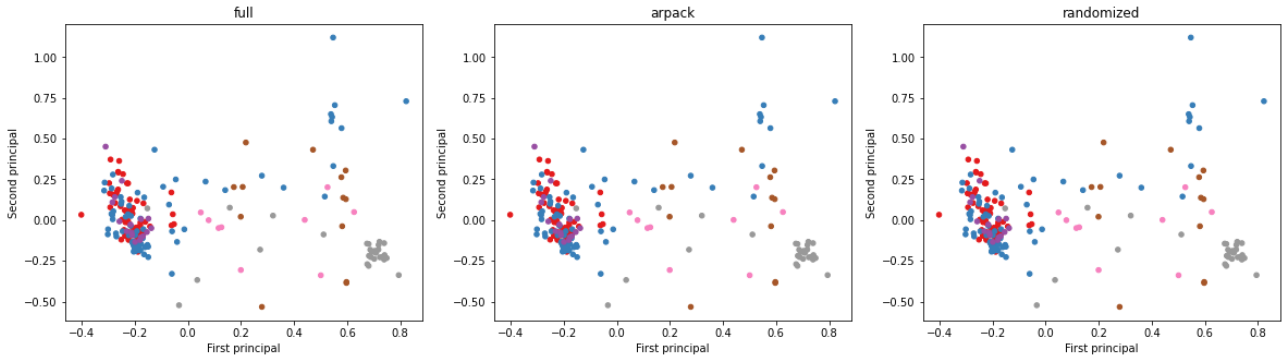

Рисунок 7. Диаграммы рассеяния для различных методов *SVD*.

Объясненные дисперсии в процентах и собственные числа для каждого метода *SVD* не отличаются и указаны в таблице 5.

компонента	1	2	3	4
объясненная дисперсия	0.454296	0.179901	0.126495	0.097978
собственные числа	5.104931	3.212457	2.693745	2.370751

Таблица 5. Объясненная дисперсия в процентах и собственные числа для каждого метода *SVD*.

Модификация метода главных компонент

1. По аналогии с *PCA* исследуйте *KernelPCA* для различных параметров *kernel* и различных параметрах для ядра.

   Диаграмма рассеяния первых двух компонент для *kernel=linear* приведена на рисунке 8. Значения собственных чисел ядреной матрицы и объясненная дисперсия приведены в таблице 6.

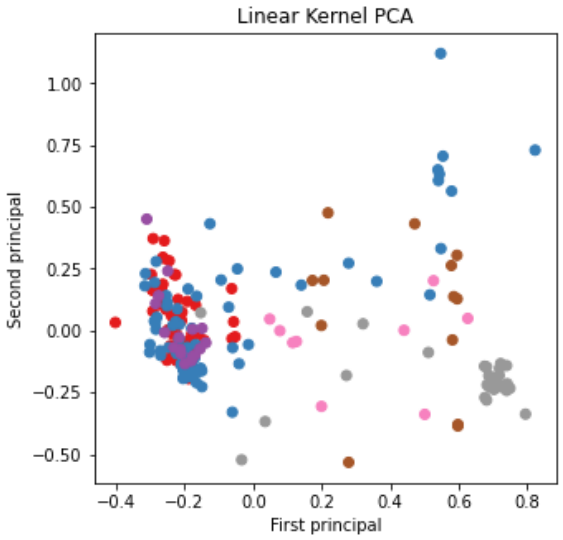

Рисунок 8. Диаграмма рассеяния первых двух компонент, *kernel=linear*.

	1	2	3	4	explained
linear	26.060318	10.319879	7.256264	5.620459	0.85867

Таблица 6. Собственные числа и объясненная дисперсия, *kernel=linear*.

Диаграмма рассеяния первых двух компонент для *kernel=poly*, при значениях по умолчанию, приведена на рисунке 9. Диаграммы рассеяния при различных параметрах ядра приведены на рисунке 10. Значения собственных чисел ядреной матрицы и объясненная дисперсия приведены в таблице 7.

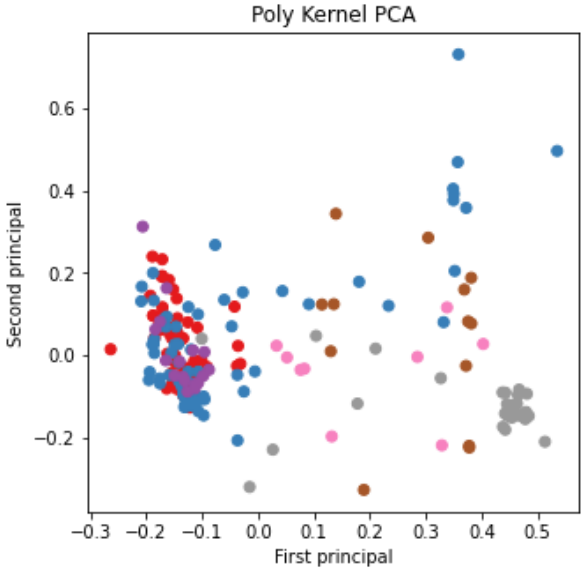

Рисунок 9. Диаграмма рассеяния первых двух компонент, *kernel=poly*.

	1	2	3	4	explained
default	10.918196	4.319377	3.118851	2.367917	0.849032
gamma = 1	344.004959	139.713954	108.149960	77.611329	0.803165
gamma = 0.5	93.580987	37.372218	28.330788	20.680771	0.821348
degree = 1	2.895591	1.146653	0.806252	0.624495	0.858670
degree = 4	16.329092	6.467189	4.727374	3.555738	0.842860
coef0 = 0	0.131363	0.058304	0.045710	0.033413	0.762103
coef0 = 10	889.810298	352.277001	248.452475	192.015368	0.857763

Таблица 7. Собственные числа и объясненная дисперсия для различных параметров ядра, *kernel=poly*.

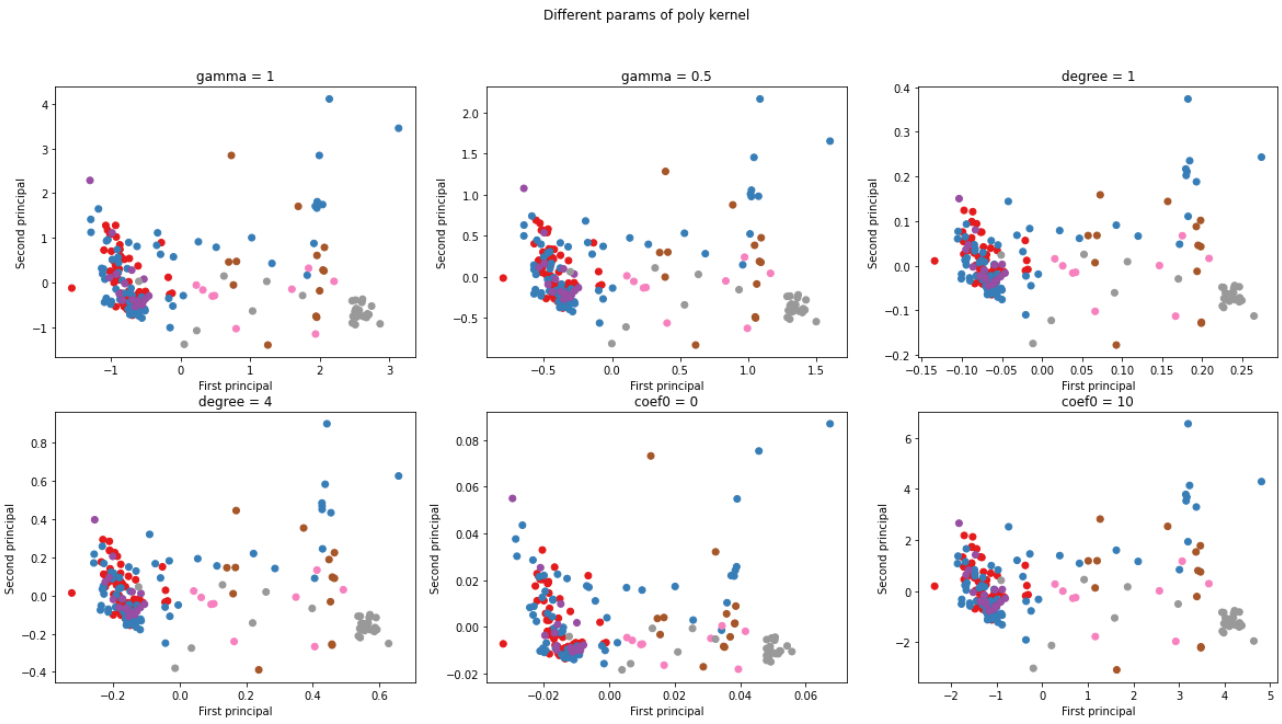

Рисунок 10. Диаграмма рассеяния первых двух компонент для различных параметров ядра, *kernel=poly*.

Диаграмма рассеяния первых двух компонент для *kernel=rbf*, при значениях по умолчанию, приведена на рисунке 11. Диаграммы рассеяния при различных параметрах ядра приведены на рисунке 12. Значения собственных чисел ядреной матрицы и объясненная дисперсия приведены в таблице 8.

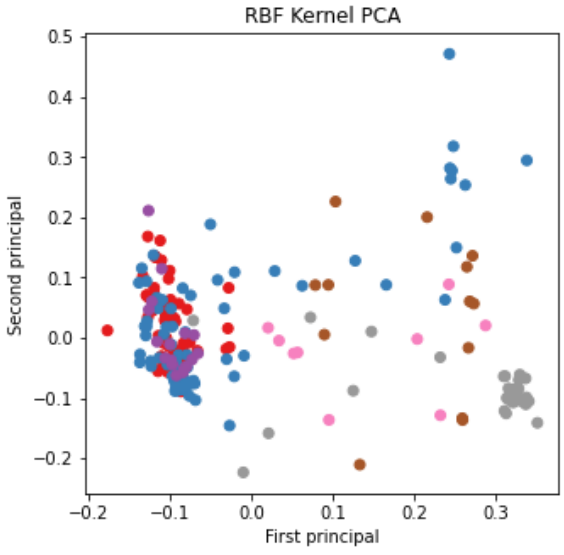

Рисунок 11. Диаграмма рассеяния первых двух компонент, *kernel=rbf*.

	1	2	3	4	explained
default	5.351453	2.018054	1.495738	1.110905	0.849938
gamma = 1	28.217500	9.985555	8.079237	5.530154	0.816641
gamma = 0.5	18.713311	6.434521	5.347800	3.647029	0.826683

Таблица 8. Собственные числа и объясненная дисперсия для различных параметров ядра, *kernel=rbf*.

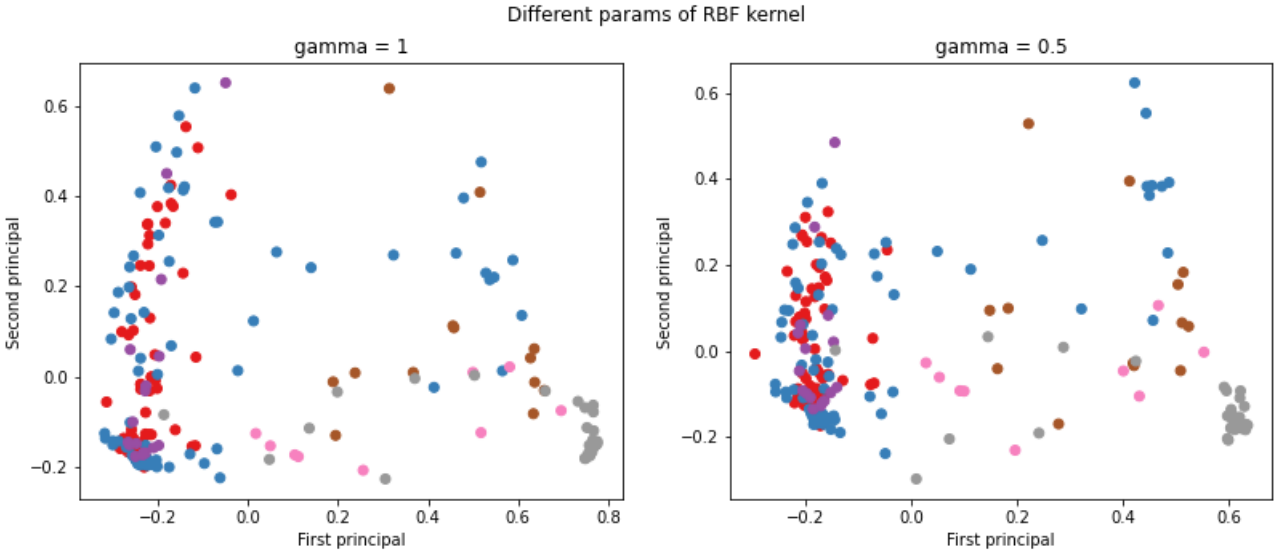

Рисунок 12. Диаграмма рассеяния первых двух компонент для различных параметров ядра, *kernel=rbf*.

Диаграмма рассеяния первых двух компонент для *kernel=sigmoid*, при значениях по умолчанию, приведена на рисунке 13. Диаграммы рассеяния при различных параметрах ядра приведены на рисунке 14. Значения собственных чисел ядреной матрицы и объясненная дисперсия приведены в таблице 9.

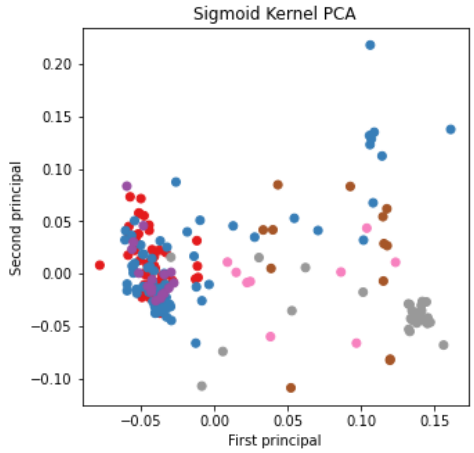

Рисунок 13. Диаграмма рассеяния первых двух компонент, *kernel=sigmoid*.

	1	2	3	4	explained
default	1.006181	0.39983750	0.2740985	0.2161195	0.861865
gamma = 1	1.652027	0.7699351	0.4662115	0.3311741	0.872191
gamma = 0.5	2.202103	0.9114499	0.5714274	0.4644157	0.870135
coef0 = 0	2.852693	1.129250	0.7910028	0.6142272	0.859284
coef0 = 10	1.8762e-08	7.4674e-09	5.0884e-09	4.0282e-09	0.862512

Таблица 9. Собственные числа и объясненная дисперсия для различных параметров ядра, *kernel=sigmoid*.

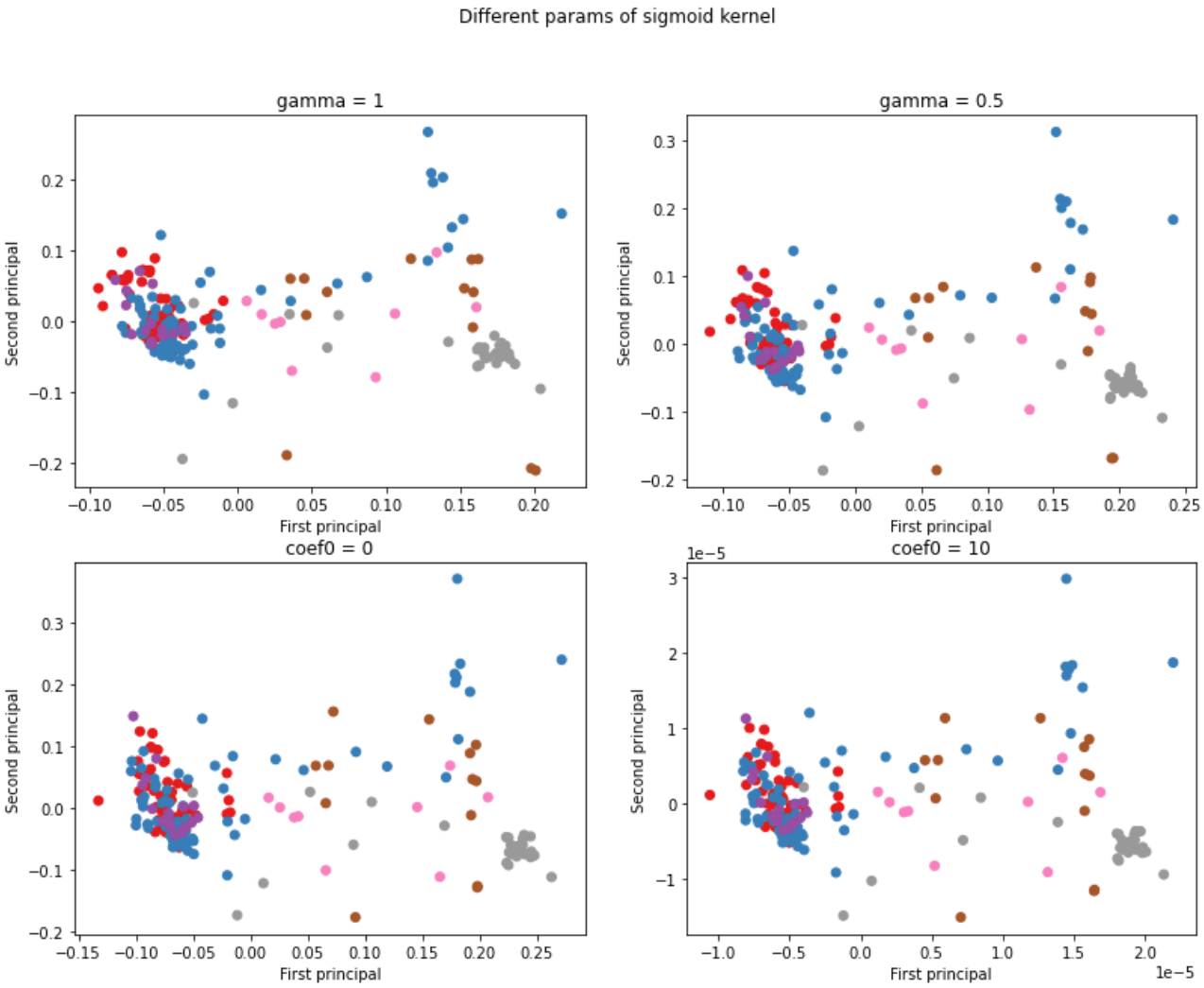

Рисунок 14. Диаграмма рассеяния первых двух компонент для различных параметров ядра, *kernel=sigmoid*.

Диаграмма рассеяния первых двух компонент для *kernel=cosine* приведена на рисунке 15. Значения собственных чисел ядреной матрицы и объясненная дисперсия приведены в таблице 10.

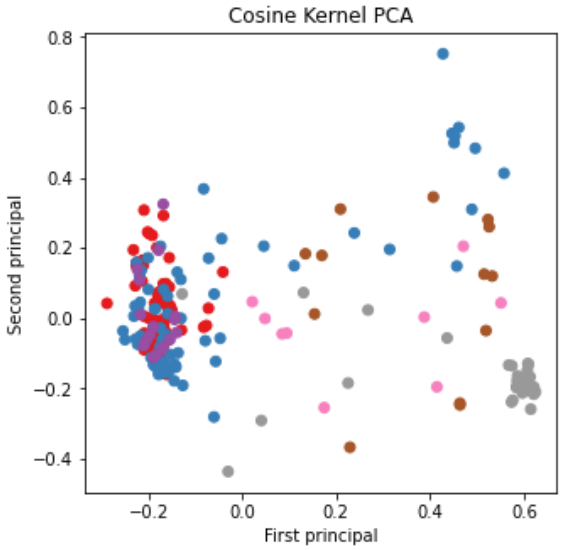

Рисунок 15. Диаграмма рассеяния первых двух компонент, *kernel=cosine*.

	1	2	3	4	explained
cosine	18.31403	6.475385	4.695999	3.578125	0.859944

Таблица 10. Собственные числа и объясненная дисперсия, *kernel=cosine*.

В ходе применения *KernelPCA* с различными ядерными функциями выяснено, что собственные числа ядерной матрицы для каждой функции отличаются как в масштабе, так и в соотношении друг с другом, однако объясненная дисперсия во всех случаях схожа. Худшую объясненную дисперсию четырех компонент для данного набора данных показало полиномиальное ядро с параметром *coef0=0*, лучшую – сигмовидное ядро с *gamma=1*.

Ядерные функции для каждого параметра (linear, poly, rbf, sigmoid, cosine) представлены в таблице 11.

kernel of *KernelPCA*	Ядерная функция
linear	$k(x, y) = x^T y$
poly	$k(x, y) = (\gamma x^T y + c_0)^d$
rbf	$k(x, y) = \exp(-\gamma \|x - y\|^2)$
sigmoid	$k(x, y) = \tanh(\gamma x^T y + c_0)$
cosine	$k(x, y) = \frac{xy^T}{\|x\|\|y\|}$
precomputed	Задается ядерной матрицей.

Таблица 11. Ядерные функции *KernelPCA*.

В уравнениях таблицы 11 у ядерных функций используются параметры, которые можно сообщить в *KernelPCA*:

- γ – *gamma*
- c_0 – *coef0*
- d – *degree*

2. Определите, при каких параметрах *KernelPCA* работает также как *PCA*.
   PCA работает аналогично *KernelPCA* с линейным ядром (*kernel=linear*).

3. Аналогично исследуйте *SparsePCA*
   SparsePCA позволяет проводить анализ разреженных компонент, таким образом чтобы данные можно было восстановить наиболее оптимально. Для решения задачи lasso, *SparsePCA* использует один из двух методов: *Lars* (least angle regression) – наименьшая угловая регрессия; *CD* (coordinate descent) – координатный спуск.
   Lasso – метод регрессионного анализа, который позволяет наиболее оптимально (для *SparsePCA*) сократить данные путем отбора.
   Диаграммы рассеянности первых двух компонент с использованием метода Lars и различными параметрами alpha представлены на рисунке 16, с использованием метода *CD* – на рисунке 17.

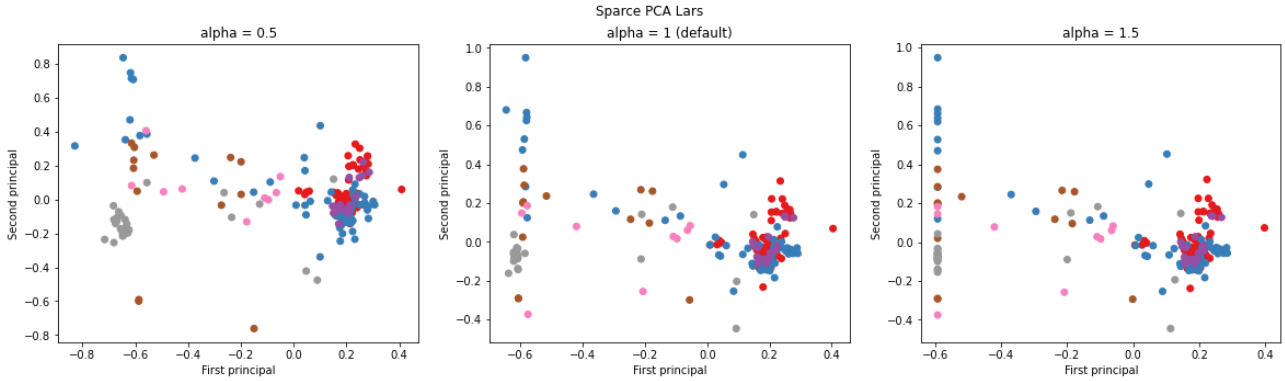

Рисунок 16. Диаграмма рассеянности первых двух компонент *SparsePCA* для различных *alpha, method=lars*.

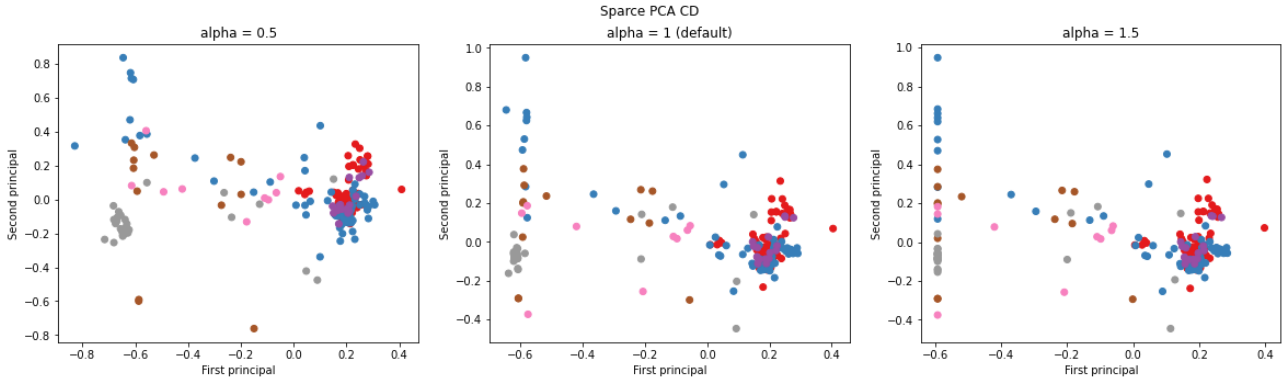

Рисунок 17. Диаграмма рассеянности первых двух компонент *SparsePCA* для различных *alpha, method=cd*.

4. Проанализируйте и обоснуйте полученные результаты

   Из рисунков 16 и 17 видно, что для данного набора данных выбор метода (*Lars* или *CD*) не имеет значения. Значение alpha определяет как сильно будут разрежены компоненты.

   При *alpha=0* разреживание не производится и *SparsePCA* аналогичен *PCA*.

Факторный анализ

1. Проведите понижении размерности используя факторный анализ *FactorAnalysis*.

   Факторный анализ – процесс выявления взаимосвязей между наблюдениями и поиска скрытых зависимостей. Сильно коррелирующие признаки объединяются и пространство признаков сокращается.

   Был проведен факторный анализ данных по двум компонентам. Диаграмма рассеяния для двух компонент приведена на рисунке 18.

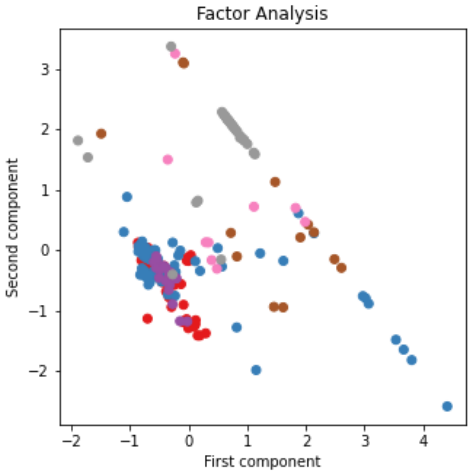

Рисунок 18. Диаграмма рассеяния для двух компонент факторного анализа.

2. Сравните полученные результаты с *PCA*

   Данные, полученные после метода главных компонент, тяжело поддаются анализу, в то время как факторный анализ позволяет выделить четкую корреляцию данных.

3. Объясните в чем разница между методом главных компонент и факторным анализом.

   - Метод главных компонент позволяет выделить признаки, вдоль которых лучше всего объясняется дисперсия
   - Факторный анализ объясняет корреляцию данных, а метод главных компонент – дисперсию.
   - Метод главных компонент представляет собой математический инструмент, позволяющий ориентировать данные лучшим образом, в

то время как факторный анализ представляет позволяет как-то интерпретировать результат.

- Метод главных компонент позволяет найти ортогональные компоненты, факторный анализ этого не гарантирует.

Вывод

В ходе лабораторной работы были изучены методы понижения размерности данных из библиотеки *Scikit Learn*.

Изучен метод PCA, а также его модификации KernelPCA и SparsePCA. Изучено влияние параметров встроенных функций ядра для KernelPCA. Изучено влияние различных методов решения SVD.

Также для понижения размерности использовался факторный анализ (FA). Выделены сходства и различия этих методов.