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

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Машинное обучение»

Тема: Понижение размерности пространства признаков

Студенка гр. 8303 _____ Самойлова А.С.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

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

Ход выполнения работы

Загрузка данных

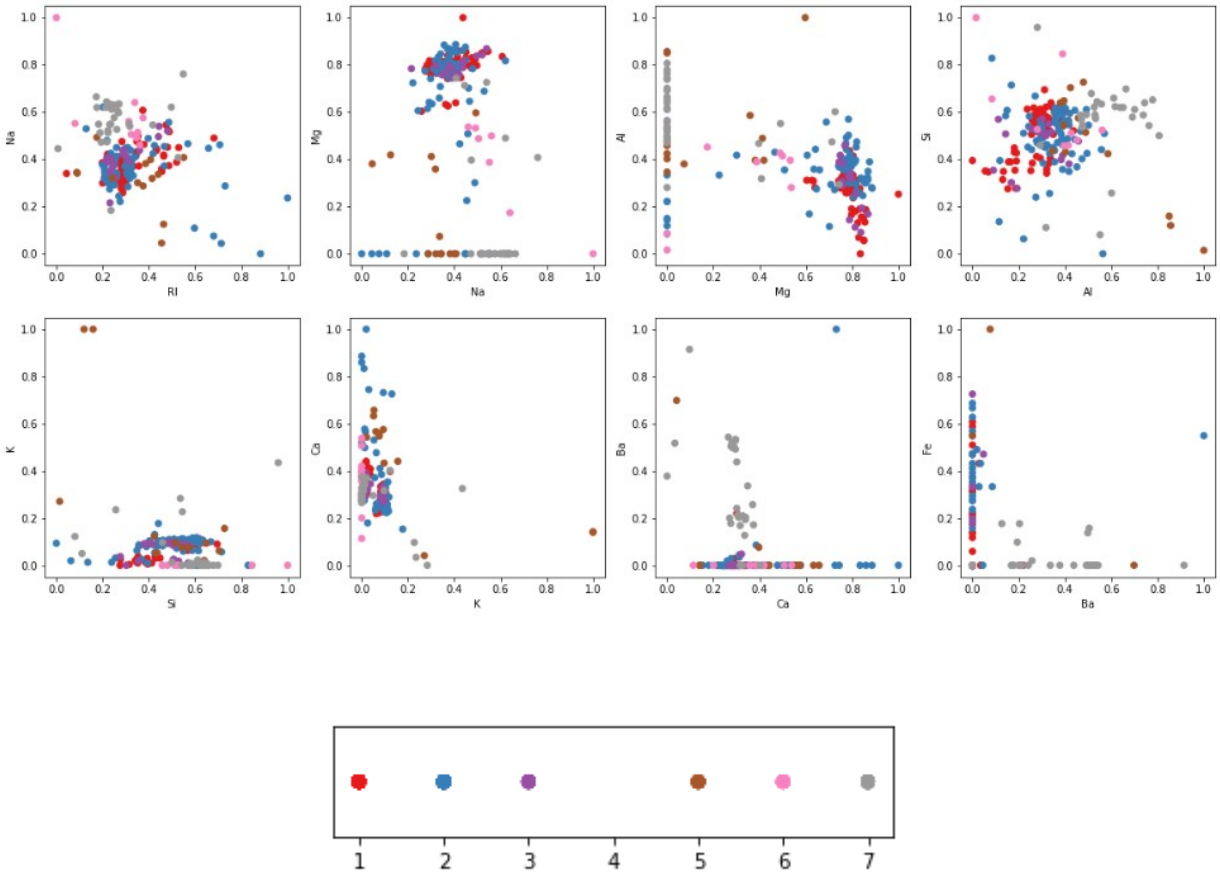
1. Загрузить датасет (https://www.kaggle.com/uciml/glass)
2. Создать *Python* скрипт. Загружзить датасет в датафрейм, и разделить данные на описательные признаки и признак отображающий класс:

```
           RI     Na    Mg    Al     Si     K    Ca    Ba   Fe  Type
0     1.52101  13.64  4.49  1.10  71.78  0.06  8.75  0.00  0.0     1
1     1.51761  13.89  3.60  1.36  72.73  0.48  7.83  0.00  0.0     1
2     1.51618  13.53  3.55  1.54  72.99  0.39  7.78  0.00  0.0     1
3     1.51766  13.21  3.69  1.29  72.61  0.57  8.22  0.00  0.0     1
4     1.51742  13.27  3.62  1.24  73.08  0.55  8.07  0.00  0.0     1
..        ...    ...   ...   ...    ...   ...   ...   ...  ...   ...
209   1.51623  14.14  0.00  2.88  72.61  0.08  9.18  1.06  0.0     7
210   1.51685  14.92  0.00  1.99  73.06  0.00  8.40  1.59  0.0     7
211   1.52065  14.36  0.00  2.02  73.42  0.00  8.44  1.64  0.0     7
212   1.51651  14.38  0.00  1.94  73.61  0.00  8.48  1.57  0.0     7
213   1.51711  14.23  0.00  2.08  73.36  0.00  8.62  1.67  0.0     7

[214 rows x 10 columns]
```

3. Провести нормировку данных к интервалу [0, 1]

4. Построить диаграммы рассеяния для пар признаков. Определить соответствие цвета на диаграмме и класса в датасете.

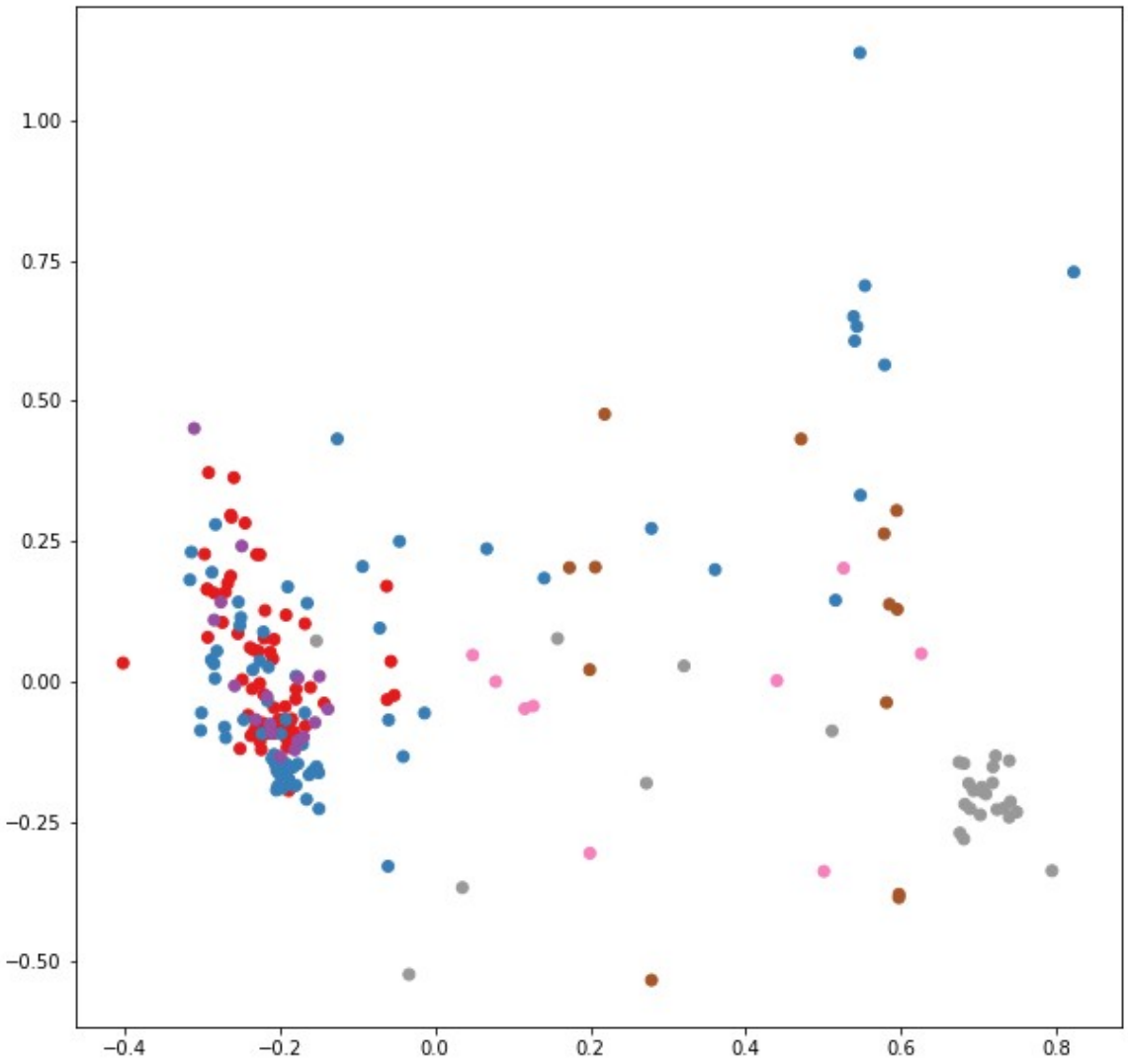


Метод главных компонент

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

Компонента	1	2
Объясненная дисперсия в %	0.45429569	0.17990097
Собственные числа	5.1049308	3.21245688

3. Построить диаграмму рассеяния после метода главных компонент



4. Проанализировать и обосновать полученные результаты
   - Класс 1: данные сконцентрированы в отрицательной части первой компоненты и в центральной части второй компоненты.
   - Класс 2: большая часть данных находится в отрицательных частях первой и второй компанент, с разбросом в сторону увеличения компонент
   - Класс 3: данные сконцентрированы в отрицательной части первой компоненты и в центральной части второй компоненты (аналогично первому классу, но с меньшим разбросом)

- Класс 5: данные представляют два столбца параллельные второй компоненьте
- Класс 6: данные редставляют два столбца параллельные второй компоненьте (аналогично 5 классу, однако имеют большую дисперсию по первой компоненьте)
- Класс 7: большая часть данных находится в положительной части первой компоненьты и в отрицательной части второй, с разбросом в сторону уменьшения первой и увеличения второй компонент.

5. Изменяя количество компонент, определить 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%

Из представленной таблицы видно, что первые 4 компоненты объясняют 85,87% дисперсии данных.

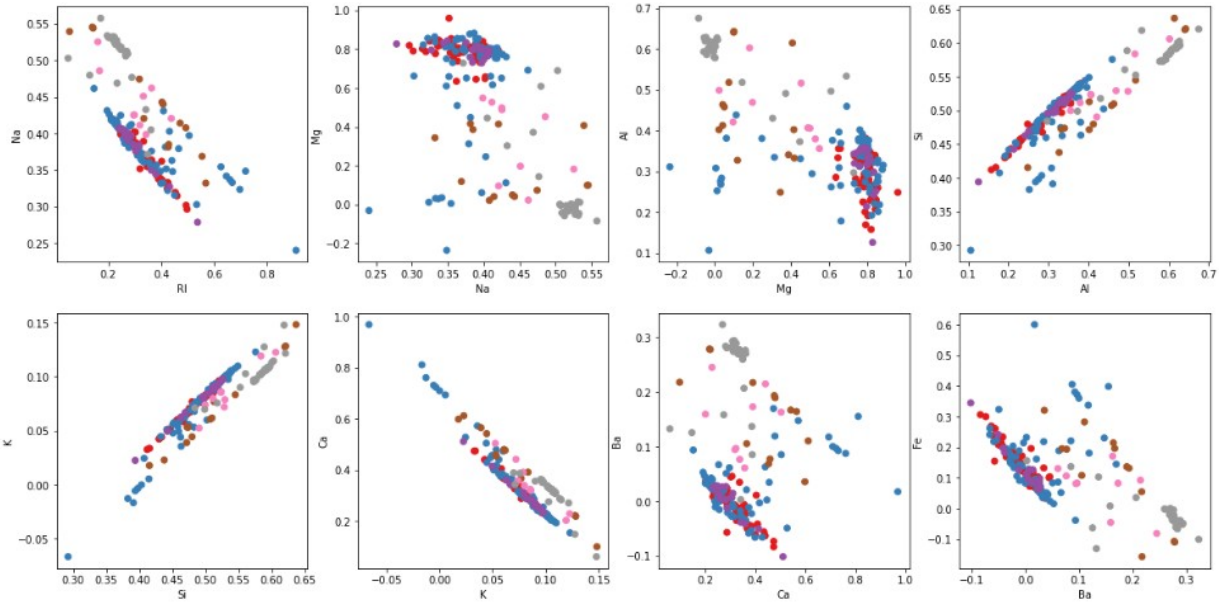
6. Используя метод inverse_transform восстановить данные, сравнить с исходными

Данные, восстановленные по 2 компонентам:

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
data var	0.017772	0.015079	0.103201	0.024191	0.019130	0.011030	0.017494	0.024916	0.036503
recovered data var	0.012907	0.003405	0.101767	0.013433	0.002585	0.000811	0.014657	0.010019	0.011216

Из таблицы видно, что дисперсия не сильно поменялась только у Ri, Mg, Ca, а у остальных элементов она сильно уменьшилась.

Данные были восстановлены не корректно, что видно на диаграммах:

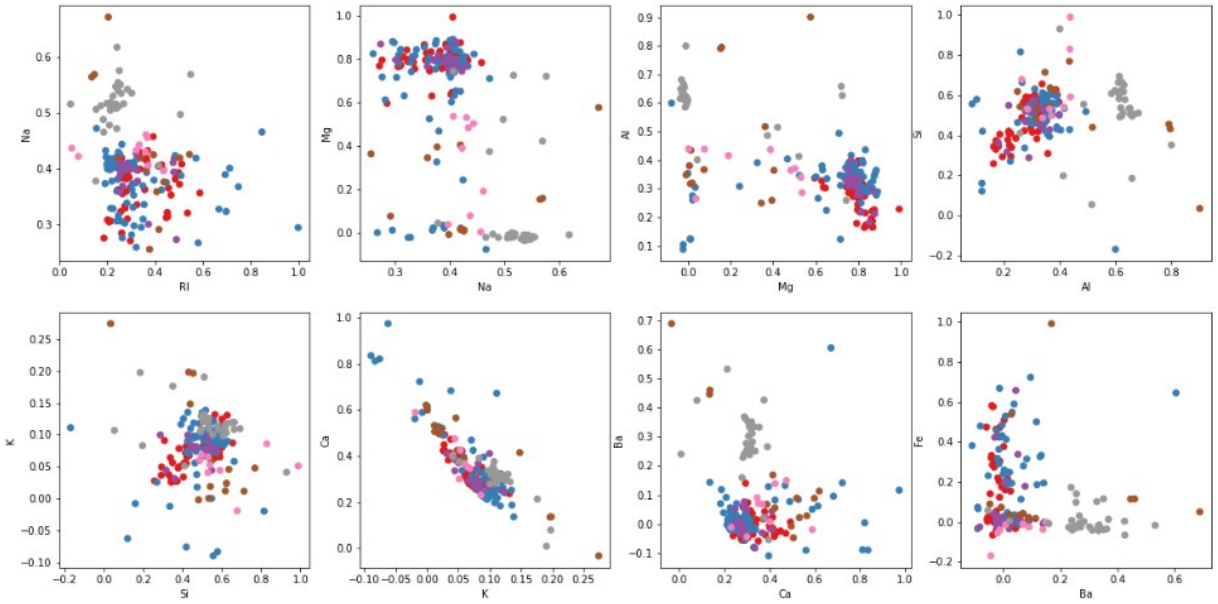


Данные, восстановленные по 4 компонентам:

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
data var	0.017772	0.015079	0.103201	0.024191	0.019130	0.011030	0.017494	0.024916	0.036503
recovered data var	0.016902	0.004732	0.102760	0.018471	0.016811	0.001918	0.016578	0.017499	0.035582

Из таблицы видно, что дисперсия близка к исходной уже у большего числа элементов: Ri, Mg, Si, Ca и Fe.

Однако на диаграммах видно, что данные всё равно сильно искажены:

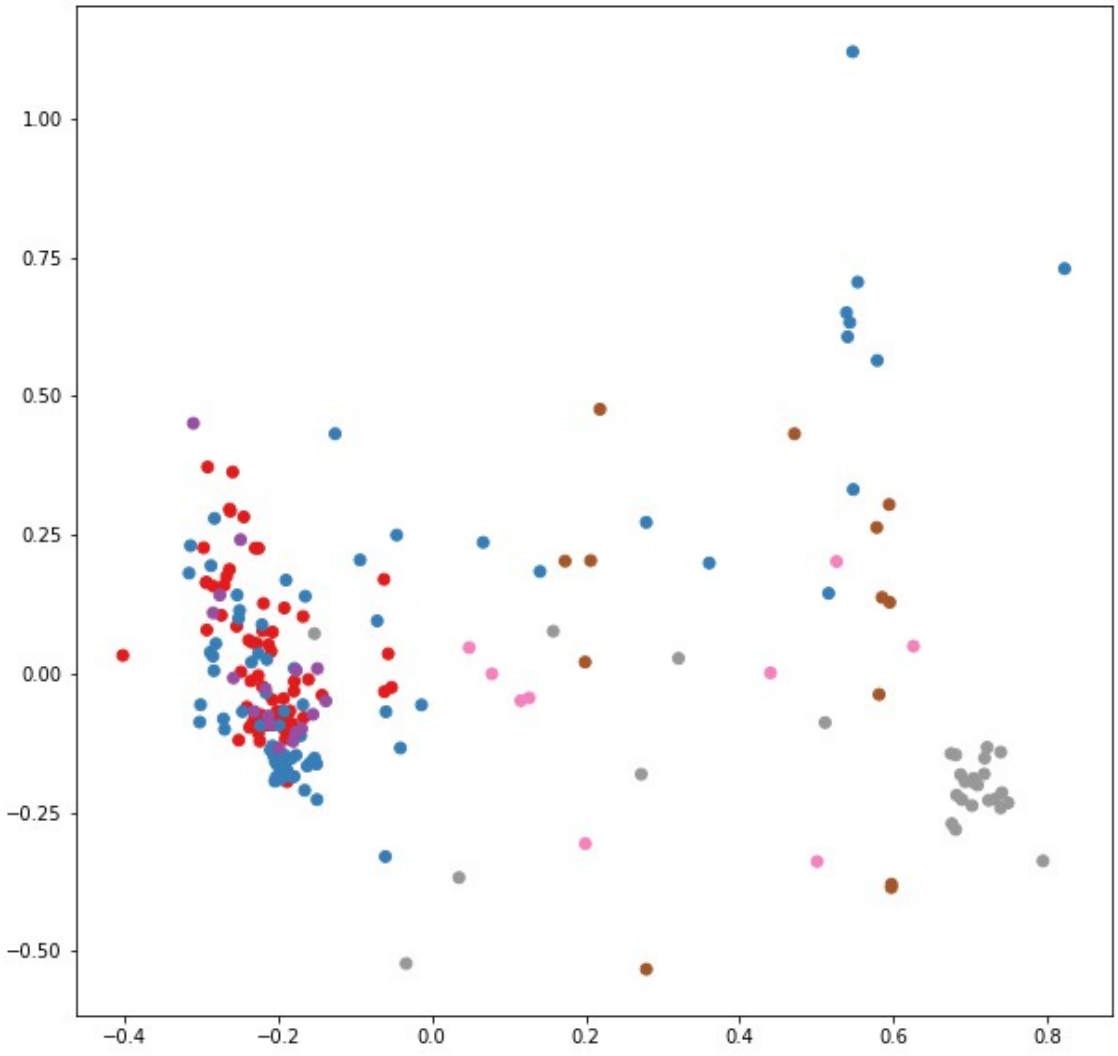


7. Исследовать метод главных компонент при различных параметрах *svd_solver*.

   Параметр *svd_solver* может быть установлен в значения:

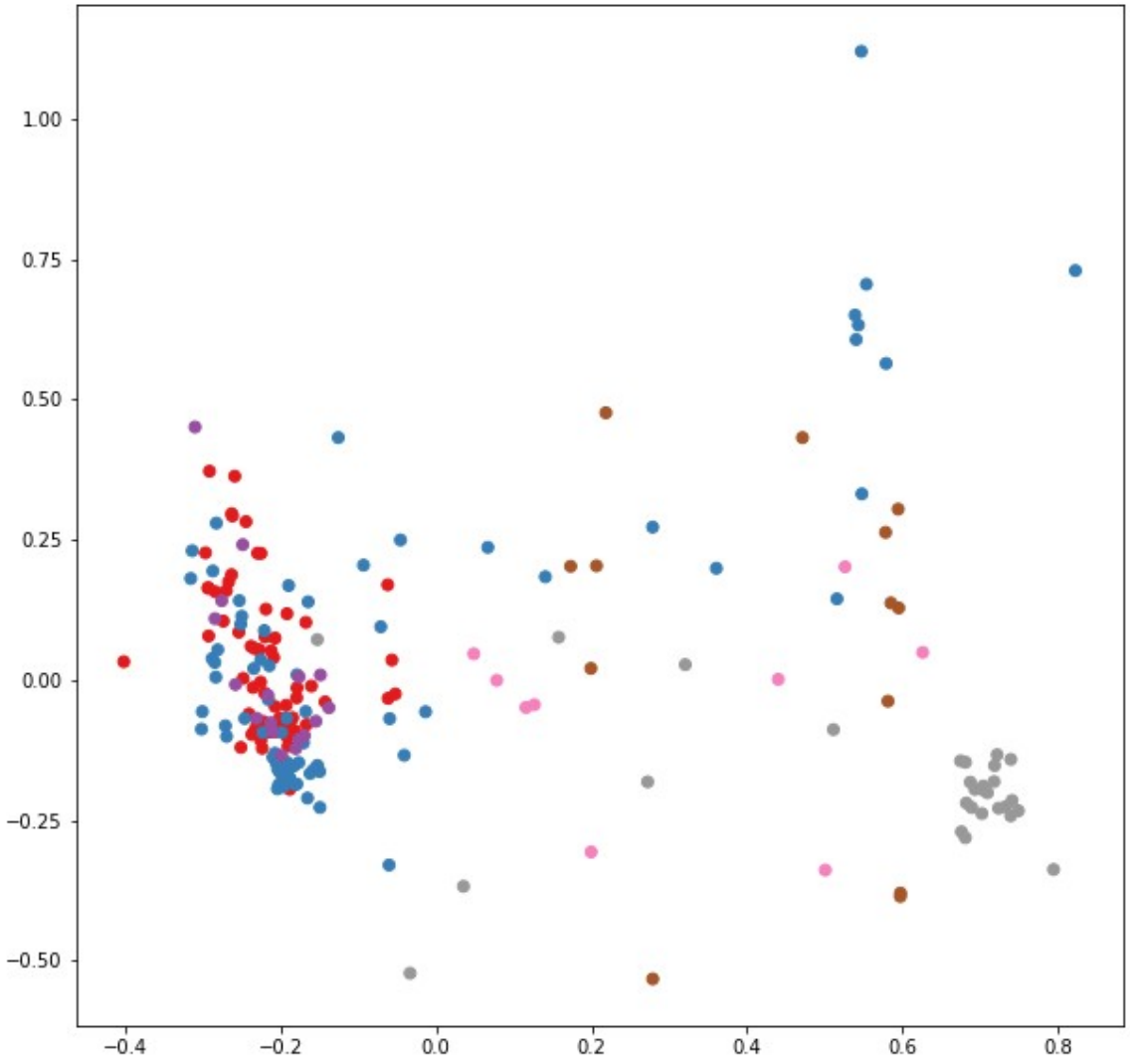
   - auto (default)
   - full

```
Объясненная дисперсия в %:  [0.45429569 0.17990097]
Собственные числа:          [5.1049308  3.21245688]
```



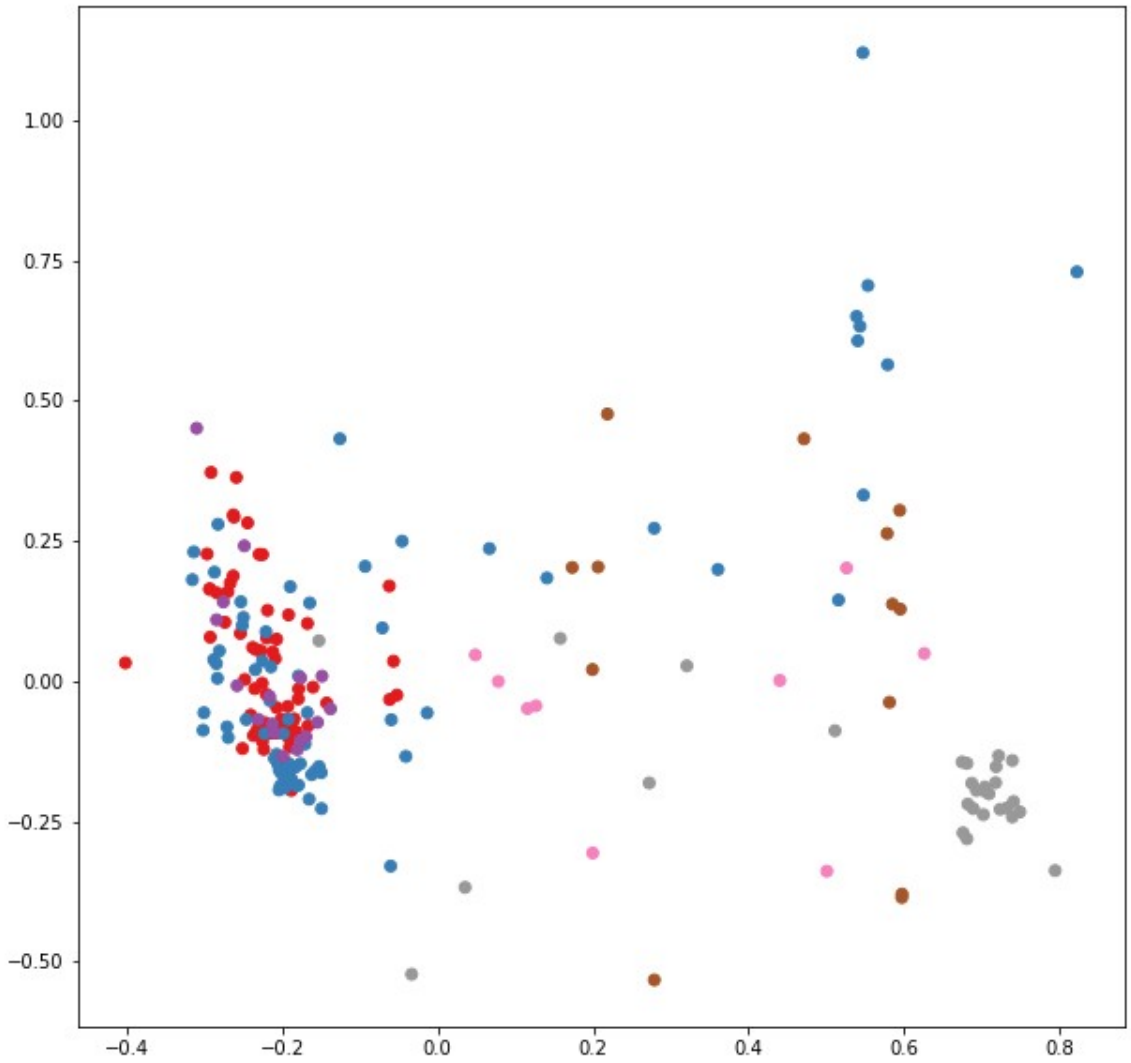
- arpack

```
Объясненная дисперсия в %:  [0.45429569 0.17990097]
Собственные числа:          [5.1049308  3.21245688]
```



- randomized

```
Объясненная дисперсия в %:  [0.45429569 0.17990097]
Собственные числа:          [5.1049308  3.21245688]
```



Метод главных компонент при различных параметрах svd_solver дает одинаковый результат.

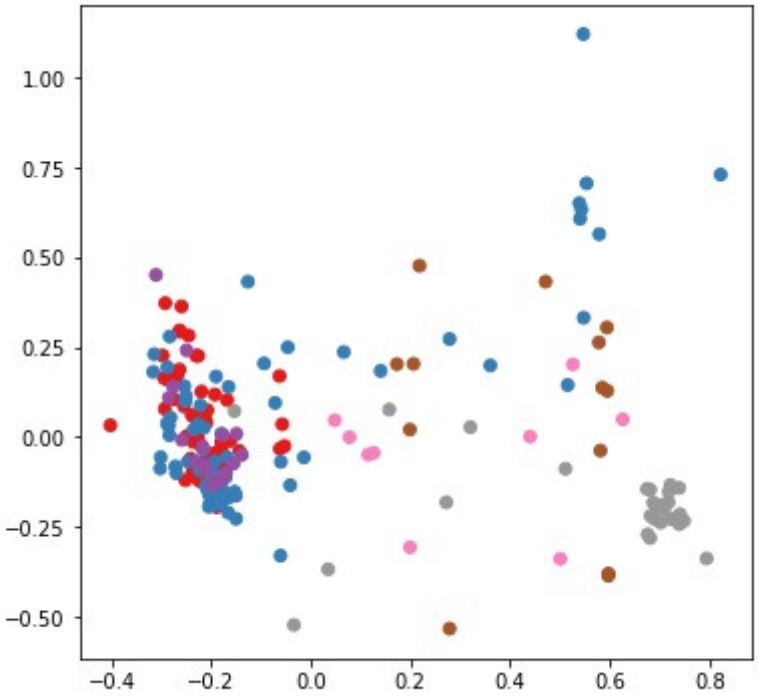
Модификация метода главных компонент

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

   Параметр *kernel* может принимать следующие значения:

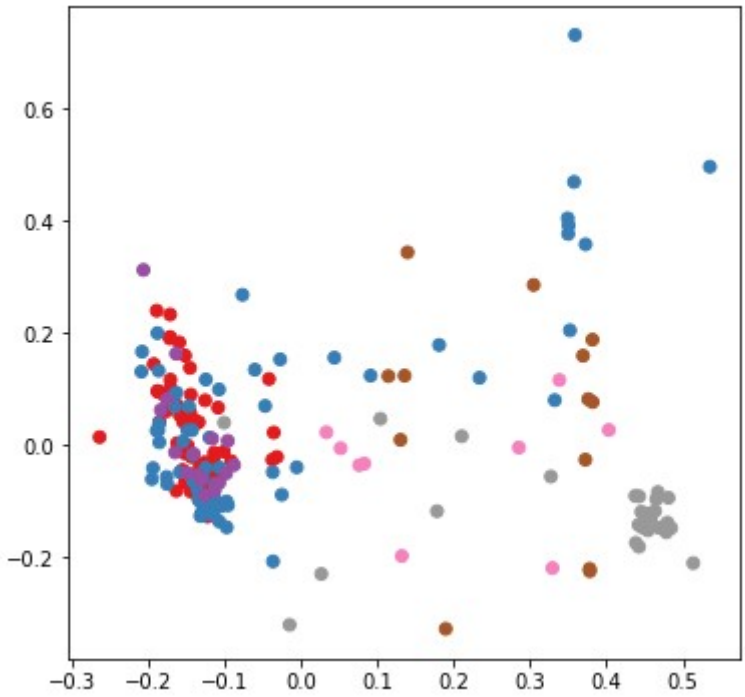
   - linear (default)

     Диаграмма рассеяния для первых 2 компонент:

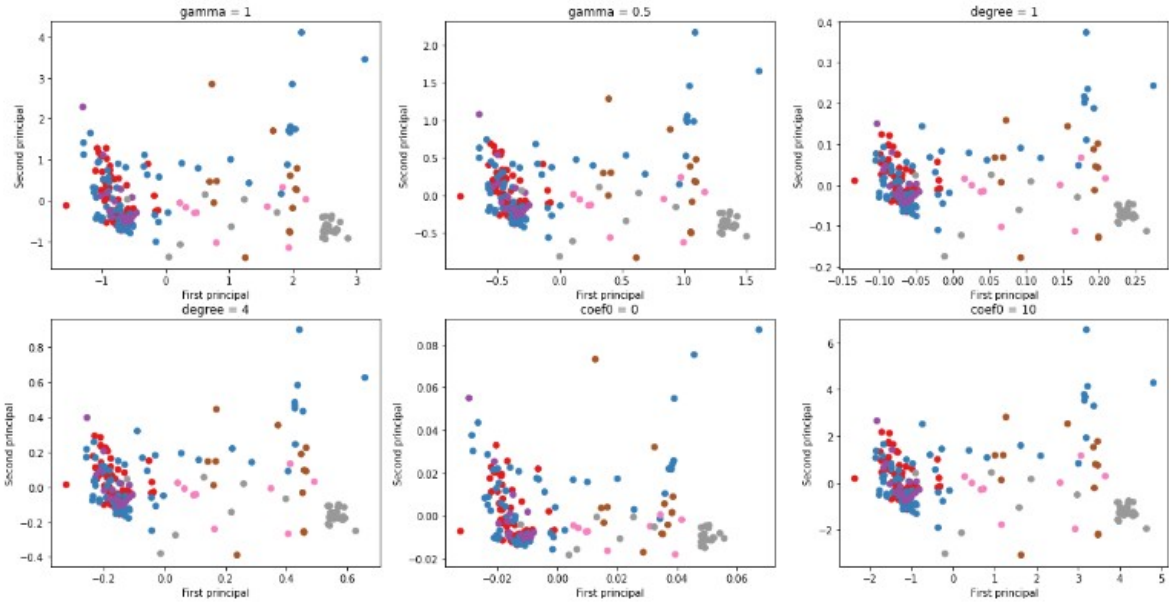


- poly

  Диаграмма рассеяния для первых 2 компонент при значениях по умолчанию:



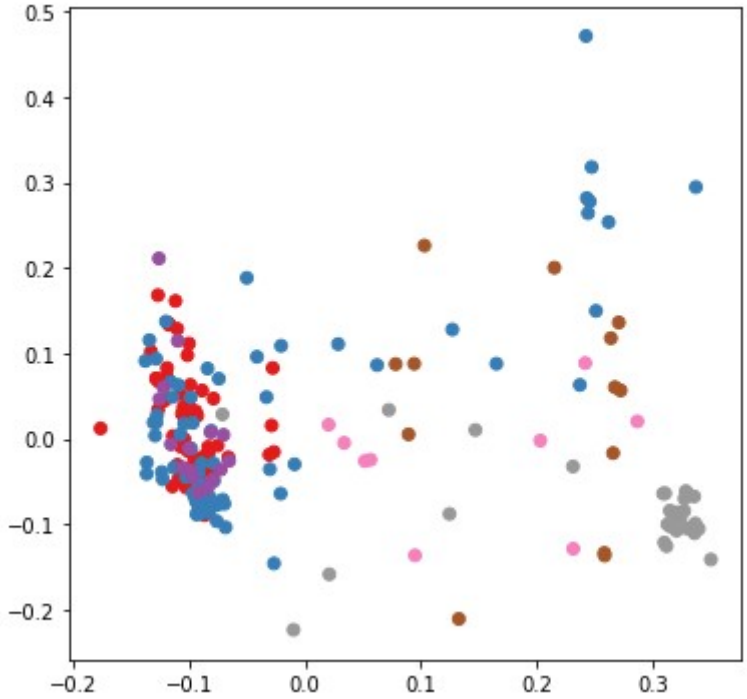
  Диаграммы рассеяния для первых 2 компонент при различных параметрах ядра:



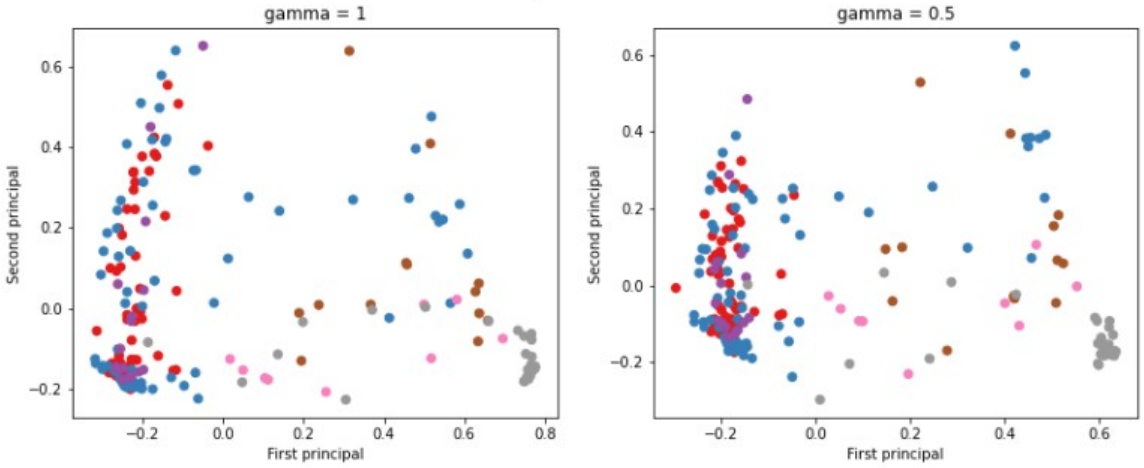
	1	2	explained, %
default	10.918196	4.319377	62.425064
gamma = 1	344.004959	139.713954	58.030996
gamma = 0.5	93.580987	37.372218	59.766201
degree = 1	2.895591	1.146653	63.419666
degree = 4	16.329092	6.467189	61.822522
coef0 = 0	0.131363	0.058304	53.776446
coef0 = 10	889.810298	352.277001	63.321358

- rbf

  Диаграмма рассеяния для первых 2 компонент:



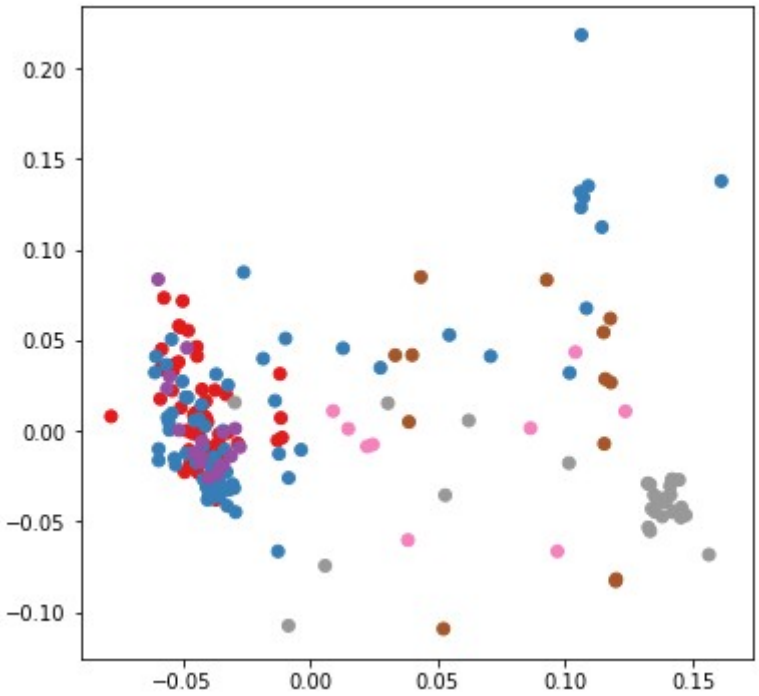
  Диаграммы рассеяния для первых 2 компонент при различных параметрах ядра:



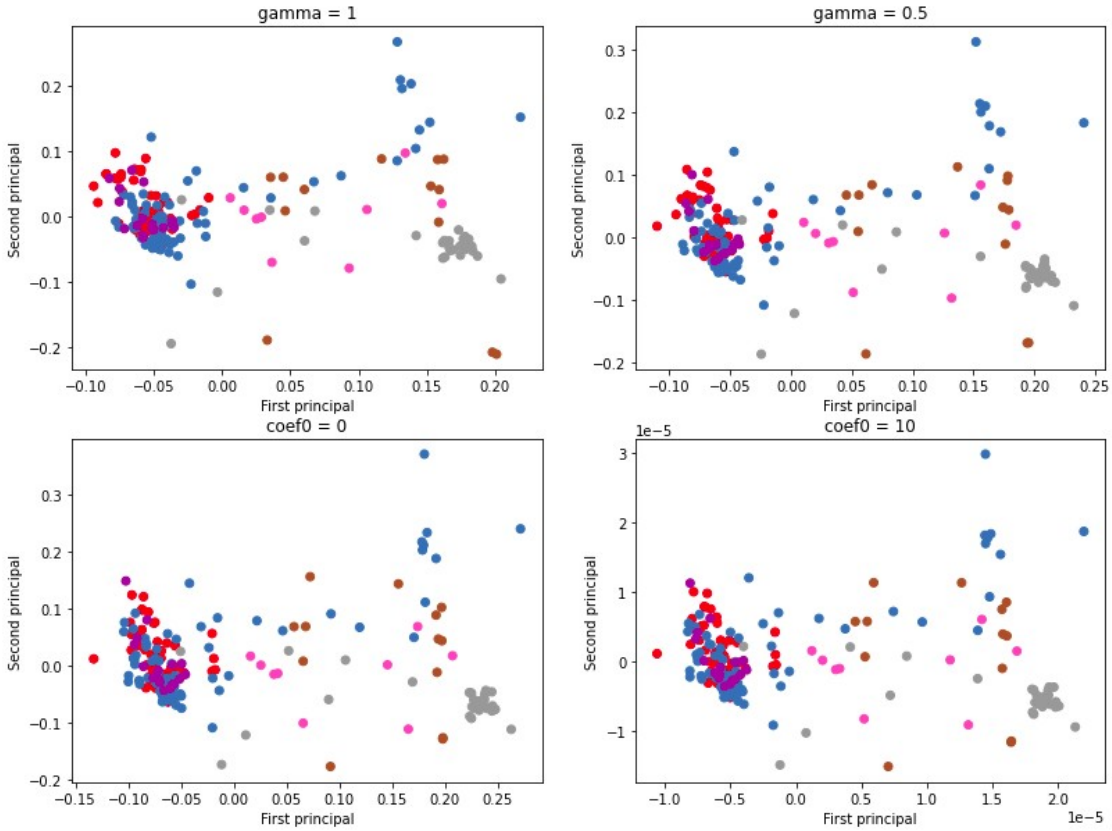
	1	2	explained, %
default	5.351453	2.018054	62.786000
gamma = 1	28.217500	9.985555	60.213668
gamma = 0.5	18.713311	6.434521	60.889491

- sigmoid

  Диаграмма рассеяния для первых 2 компонент:



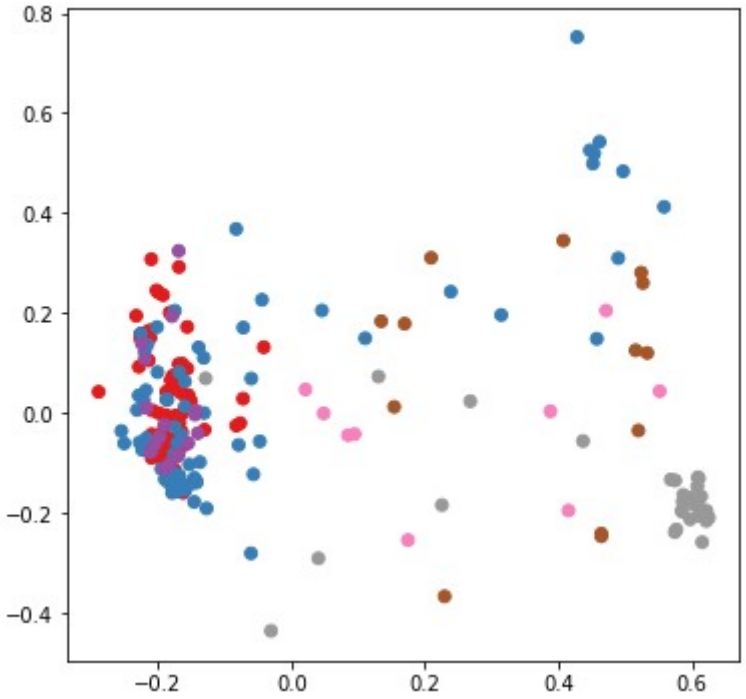
  Диаграммы рассеяния для первых 2 компонент при различных параметрах ядра:



	1	2	explained, %
default	1.006181e+00	3.998375e-01	63.905427
gamma = 1	1.652027e+00	7.699351e-01	65.616211
gamma = 0.5	2.202103e+00	9.114499e-01	65.291688
coef0 = 0	2.852693e+00	1.129250e+00	63.514218
coef0 = 10	1.876208e-08	7.467428e-09	64.004929

- cosine

  Диаграмма рассеяния для первых 2 компонент:



- precomputed

Ядерные функции для каждого параметра:

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.

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

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

2. Определить, при каких параметрах KernelPCA работает также как PCA.

   PCA работает аналогично KernelPCA с линейным ядром (kernel=linear).

3. Аналогично исследовать SparsePCA

   SparsePCA производит анализ разреженных компонент, что позволяет наиболее оптимальным обарзом восстановить данные.

   Диаграмма рассеяния для первых 2 компонент для различных *alpha*, *method = lars:*

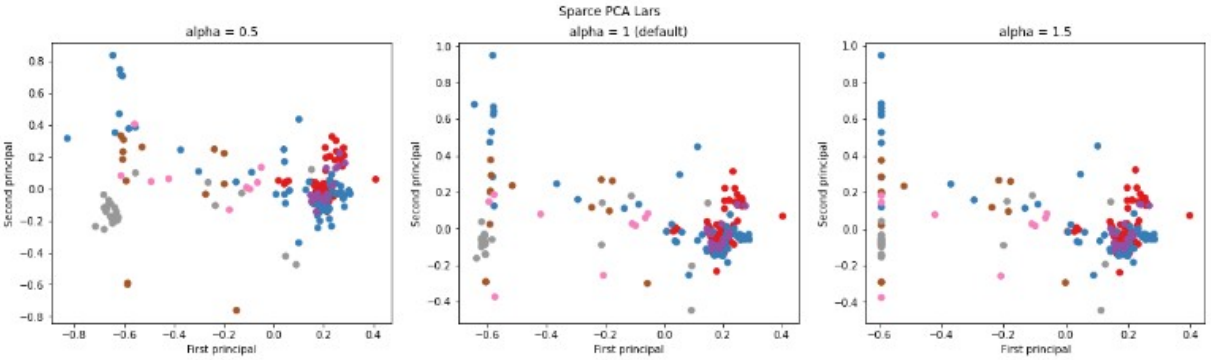
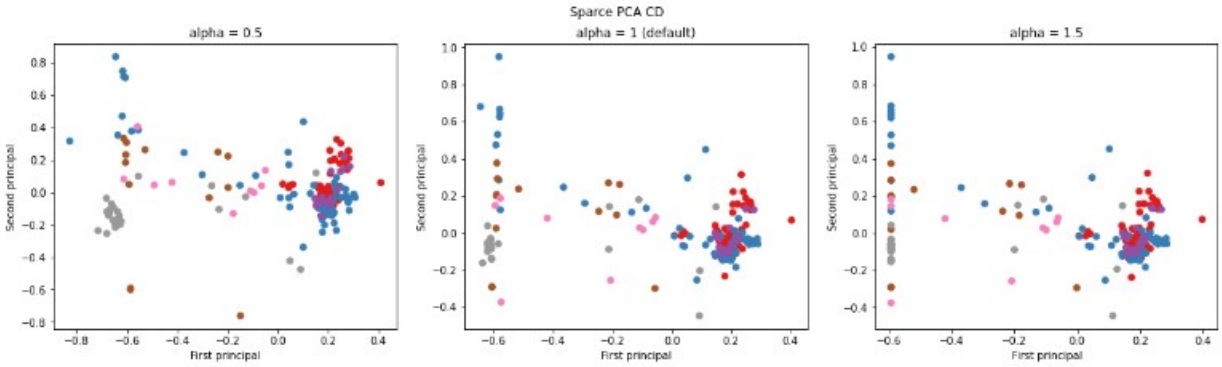


Диаграмма рассеяния для первых 2 компонент для различных *alpha*, *method = cd:*

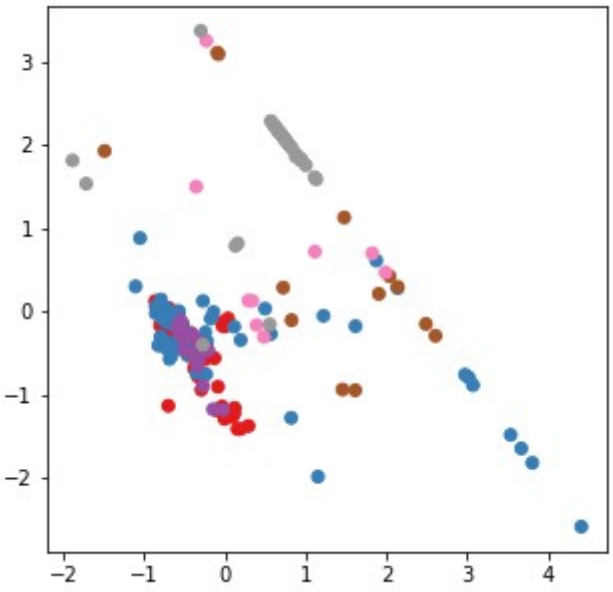


4. Проанализировать и обосновать полученные результаты

   Диаграммы показывают, что для данного набора данных методы *lars* и *cd* не различимы. Значение *alpha* определяет как сильно будут разреженны компоненты.

Факторный анализ

1. Провести понижении размерности используя факторный анализ FactorAnalysis.



2. Сравнить 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. Объяснить 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.

Выводы

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.