Санкт-Петербургский государственный университет
Прикладная математика, программирование и искусственный интеллект

Отчет по учебной практике 2 (научно-исследовательской работе) (семестр 3)

Сингулярное разложение и анализ главных компонент.

Выполнила:
Барабашева Анастасия Дмитриевна,
группа 22.Б04-мм
Научный руководитель:
Кандидат физико-математических наук,
доцент
Голяндина Нина Эдуардовна.
Кафедра статистического моделирования

Работа выполнена на хорошем уровне
и может быть зачтена с оценкой А.

Санкт-Петербург

2023

I. Введение

В ходе работы я познакомилась с сингулярным разложением и с основами анализа главных компонент. Я выполнила упражнения, связанные с сингулярным разложением, также провела анализ данных при помощи анализа главных компонент, математической основой которого является сингулярное разложение.

II. Основная часть

1. Сингулярное разложение

Сингулярным разложением (SVD) матрицы называем равенство

$X = \sum_{i=1}^{p} \sqrt{\lambda_i} U_i V_i^T$.

Где $\sqrt{\lambda_i}$ - сингулярные числа матрицы X, векторы U_i и V_i - левые и правые сингулярные векторы матрицы X. Набор $(\sqrt{\lambda_i}, U_i, V_i)$ называется i-той собственной тройкой матрицы X.

Здесь U_i - ортонормальные линейно независимые собственные векторы матрицы XX^T, а λ_i собственные векторы этой матрицы.

$V_i = \frac{1}{\sqrt{\lambda_i}} X^T U_i$

Основные свойства.
Пусть есть разложение матрицы $X = \sum_{i=1}^{L} c_i P_i Q_i^T$, где $P_1, ..., P_L, Q_1, ..., Q_L$ - некоторые ортонормированные системы в R_L, R_K, $c_1 >= ..... >= c_L >= 0$.
Тогда:
1) $P_1, ..., P_L$ - собственные векторы матрицы XX^T, P_i соответствует собственному числу λ_i
2) $c_i^2 = \lambda_i \quad (i = 1, \ldots, d)$

3) $Q_i = \frac{X^T P_i}{\sqrt{\lambda_i}} \quad (i = 1, \ldots, d)$

Вектор
$Z_i = (c_1(U_i), \ldots, c_K(U_i))^T = X^T U_i$
называем вектором i-х главных компонент.

Перейдем к выполнению упражнений, связанных с сингулярным разложением.

Упражнение 1:

Как, не делая сингулярного разложения (не считая собственных векторов и пр.), легко ответить на вопрос, является ли разложение сингулярным (имеется в виду разложение в сумму матриц ранга 1)
Задание:
Является ли разложение матрицы
$Y = (1, 1)^T (1, 1, 1) + (-1, 1)^T (1, -1, 1)$
сингулярным?
А это (матрица другая)
$Y = (1, 1)^T (1, 1, 1) + (-1, 1)^T (2, -1, -1)$?
Если разложение сингулярное, выпишите сингулярные тройки, упорядочив их по λ_i.

Решение:
Если есть некоторое разложение матрицы $X = \sum_{i=1}^{L} c_i P_i Q_i^T$, то из свойств достаточно будет проверить ортонормированность P_i и Q_i

У матрицы 1 векторы (1, 1, 1) и (1, -1, 1) не ортогональны, так как их скалярное произведение не равно 0, значит, разложение не сингулярное.

У матрицы 2 векторы (1, 1, 1) и (2, -1, -1) и векторы (1, 1) и (-1, 1) являются ортогональными, так как их скалярное произведение равно нулю. Значит разложение является сингулярным,

(не)нормировка не мешает. Перенормируем векторы и запишем сингулярные тройки, упорядочив по вкладу, первая тройка с наибольшим вкладом.

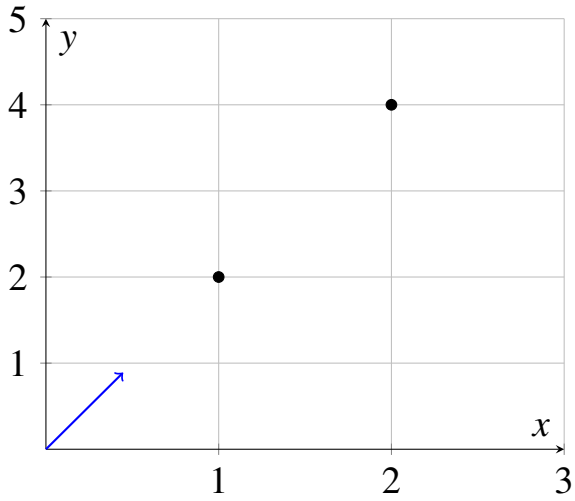
$(2\sqrt{3}, (-1/\sqrt{2}, 1/\sqrt{2}), (2/\sqrt{6}, -1/\sqrt{6}, -1/\sqrt{6}))$;
$(\sqrt{6}, (1/\sqrt{2}, 1/\sqrt{2}), (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}))$

Упражнение 2.

Условие:
Пусть матрица X имеет размерность 2 на 4. Это можно представить как четыре двумерных вектора, т.е., их можно рисовать как точки на плоскости. Нарисуйте точки так, чтобы ранг r матрицы X был 1 и 2. Нарисуйте вектора $U_i, i = 1, \ldots, r$, которым (примерно, навскидку) соответствуют ваши рисунки. Проверьте вычислениями в R.

Решение:
Если ранг равен 1, то векторы лежат на одной прямой, главный вектор единственен и тоже лежит на этой прямой. Пример такой матрицы:

$$X = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 2 & 4 & 2 & 4 \end{bmatrix}$$



$U = (\sqrt{0,2}; 2\sqrt{0,2})$

Проверим вычислениями в R:

```
mat <- matrix(c(1, 2, 1, 2, 2, 4, 2, 4), nrow = 2)

svd_result <- svd(mat)

U <- svd_result$u

print(U)
```

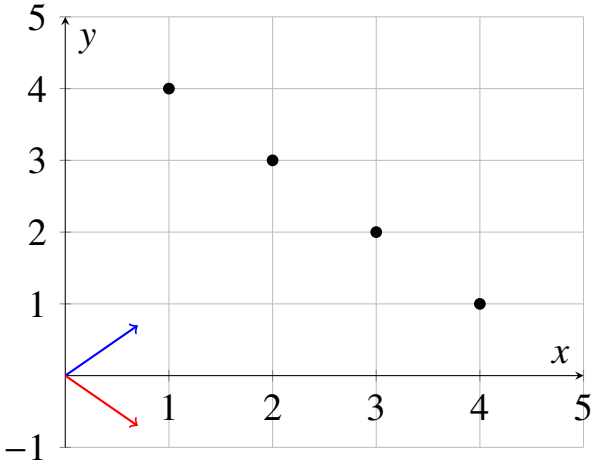
Вывод:
-0.4472136 -0.8944272
-0.8944272 0.4472136

Полученные векторы совпали с верными с точностью до знака.

Если ранг равен 2 (максимальный) - векторы расположены как угодно на плоскости. Тогда U_1 поместим визуально "посередине"векторов и чуть-чуть в сторону более длинных (т.к. лин. пространство, по-

рождаемое этим вектором лучше всего приближает набор векторов). U_2 будет ортогонален U_1 (из свойств неважное в какую сторону). Пример матрицы:

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$



$U_1 = (\sqrt{0,5}, \sqrt{0,5})$
$U_2 = (\sqrt{0,5}, -\sqrt{0,5})$

Проверим вычислениями в R:

```
mat <- matrix(c(1, 2, 3, 4, 4, 3, 2, 1), nrow = 2)

svd_result <- svd(mat)

U <- svd_result$u

print(U)
```

Вывод:
-0.7071068 -0.7071068
-0.7071068 0.7071068

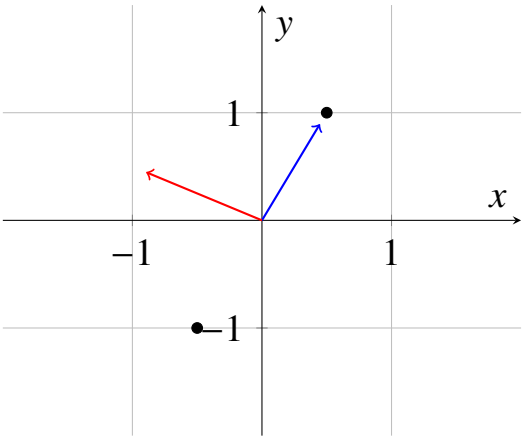
Полученные векторы совпали с верными с точностью до знака.

Теперь посмотрим что будет с примерами, если данные центрировать по строкам, т.е. из каждой строки вычесть среднее арифметическое.

Матрица 1. Центрированная матрица:

$$X = \begin{bmatrix} -0,5 & 0,5 & -0,5 & 0,5 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

$U_1 = (\sqrt{0.2}, 2\sqrt{0}.2)$
$U_2 = (-2\sqrt{0.2}, \sqrt{0.2})$

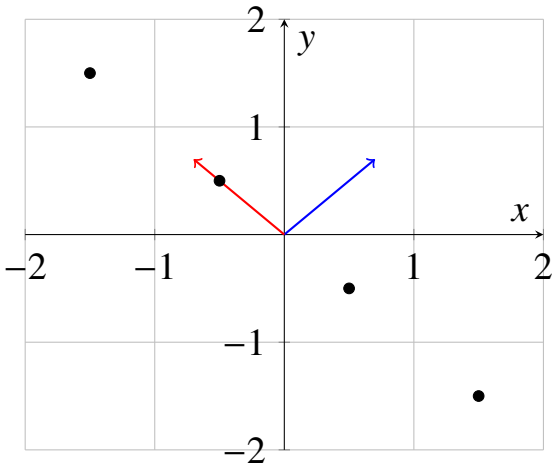


Была произведена проверка на языке R - результаты совпали. Ранг матрицы остался единичным.

Матрица 2. Центрированная матрица:

$$X = \begin{bmatrix} -1,5 & -0,5 & 0,5 & 1,5 \\ 1,5 & 0,5 & -0,5 & -1,5 \end{bmatrix}$$

$U_1 = (\sqrt{0,5}, \sqrt{0,5})$
$U_2 = (-\sqrt{0,5}, \sqrt{0,5})$



Была произведена проверка на языке R - результаты совпали. Ранг матрицы стал единичным.

Вывод: при центрировании матрицы по строкам собственные векторы не меняются, при этом ранг либо остается таким же (если минимальный), либо уменьшается, при этом работать с матрицей становится удобней.

2. Анализ главных компонент

Суть анализа главных компонент: имеются данные, по столбцам в которых находятся некоторые признаки. С помощью сингулярного разложения находим новые признаки (главные компоненты), которые раскладываются в линейную комбинацию изначальных. По новыым признакам анализировать данные удобней. Напомню, что вектор
$Z_i = (c_1(U_i), \ldots, c_K(U_i))^T = X^T U_i$
называем вектором i-х главных компонент.
Работа проводилась с данными, в которых строки — это школьники, которые поступают в школу ФТШ. Признаки — баллы, полученные за решение задач по математике и по физике, всего их 11. Столбец res — прошел или нет мальчик во второй тур. Отрезок данных для примера:

N	$M1$	$M2$	$M3$	$M4$	$P1$	$P2$	$P3$	$P4$	$P5$	$P6$	$P7$	RES
1	0	0	0	0	2	1	0	0	0	2	0	0
2	0	0	0	0	3	1	0	0	0	0	0	0
3	0	0	0	6	3	1	0	3	0	0	0	0
4	0	1	2	0	3	0	1	3	6	5	0	1
5	5	2	2	0	0	3	1	3	2	5	0	1
6	8	0	0	0	3	2	0	3	0	5	0	1
7	0	2	4	3	3	1	1	3	3	5	4	1
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	3	3	0	3	0	5	0	1
10	0	0	4	0	3	1	3	0	3	5	0	0

В ходе работы был проведен анализ главных компонент, получены коэффициенты линейной комбинации и по ним сделаны выводы о интерпретации новых признаков.
Ниже приведен код на языке R.

```
data <- read.table("exboy.txt", header = TRUE)

res_column <- data$RES

data <- data[, 2:12]

data_scaled <- scale(data)

svd_result <- svd(data_scaled)

loadings <- svd_result$v

print(loadings)

new_features <- data_scaled %*% loadings

data_with_res <- cbind(data, RES = res_column)
```

```
plot(new_features[,1], new_features[,2],
col = ifelse(data_with_res$RES == 1, "red", "blue"),
pch = 16, main = "PCA: Novye priznaki", xlab = "PC1",
ylab = "PC2")
```

Сначала загружаем данные, затем сохраняем столбец RES и удаляем его из данных. Затем стандартизируем данные, находим главные компоненты и коэффициенты в разложении новых признаков по старым, рассчитываем новые признаки. Затем добавляем столбец RES обратно в данные и визуализируем первую и вторую компоненты, выделяя цветом поступивших и непоступивших.

Коэффициенты линейной комбинации при разложении новых признаков по старым можно найти в приложении - признаки получены с помощью кода, приведенного выше. Помимо этого в приложении есть график, на котором красные обозначены как успешно поступившие, синим - не поступившие.
По коэф-там можно посмотреть, какие начальные признаки оказывают наибольшее влияние на каждую из компонент.
Первую можно интерпретировать как общий успех решения задач, у нее все положительные веса.
У второй:
Положительные веса: M1, M2, M3, M4
Отрицательные веса: P1, P2, P3, P4, P5, P6, P7
Все результаты по математике вносят положительный вклад, по физике - отрицательный, ученики с высокими баллами по математике вносят большой положительный вклад, а ученики с высокими баллами по физике - отрицательный. Интерпретируем ее как разницу между способностями по математике и по физике.

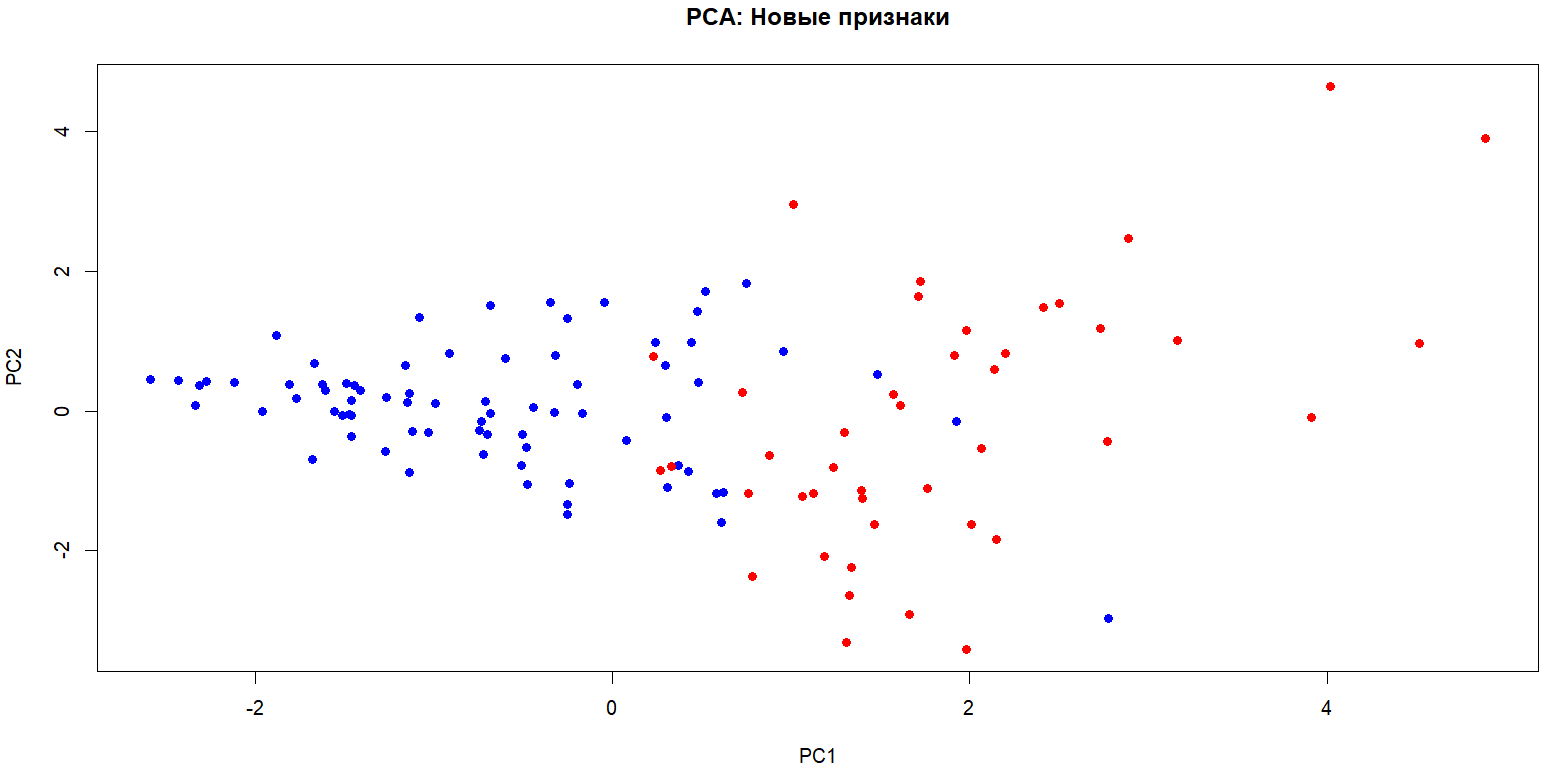
III. Заключение

В ходе выполнения работы я укрепила знания в области линейной алгебры, познакомилась с сингулярным разложением и анализом главных компонент, изучила теоретические основы методов. Выполнила несколько упражнений на свойства и оптимальные свойства сингулярного разложения, проверив правильность их выполнения на языке R, внутри которых исследовала влияние централизации на собственные векторы. В процессе выполнения работы я проанализировала данные при помощи анализа главных компонент, получив новые признаки и проинтепретировав их, построила график по полученным признакам.

IV. Список литературы

1. Н.Э.Голяндина «Метод «Гусеница»-SSA : анализ временных рядов : Учеб. пособие.» - СПб., 2004 - 76 с. (Приложение А: сингулярное разложение матриц стр.56)

V. Приложение


PCA: Новые признаки
PC2
PC1
4
2
0
-2
-2
0
2
4

```
> print(loadings)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
M1	0.1024850	0.029501291	-0.608169682	0.4701964	-0.39726943	0.22596972	-0.05215319	-0.38052498	-0.13537721	-0.14897408	0.02732625
M2	0.2920734	0.483101607	-0.200037009	0.1876333	0.08454086	-0.01578747	0.34074782	0.31541117	0.06630398	0.52863451	0.31503102
M3	0.3309481	0.402665199	-0.210166664	-0.1806907	0.31631606	0.14290274	0.09979127	0.21156786	-0.13360698	-0.55638191	-0.38588466
M4	0.2392464	0.402205075	0.139187442	-0.2519013	-0.56286609	-0.21325456	-0.52724825	0.04925168	-0.06724594	0.14695090	-0.17127772
P1	0.3963334	-0.068894534	0.219965729	-0.0375327	0.06907231	-0.26913098	0.38255844	-0.55972298	-0.35615377	0.21870990	-0.27723089
P2	0.2970503	-0.450130296	0.072611649	0.3002552	-0.19290328	0.19716896	0.04915009	0.46737353	0.08741199	0.19391011	-0.52109038
P3	0.2824015	-0.365791984	-0.301415151	-0.2764955	0.27829798	0.09302892	-0.39155386	0.12652666	-0.49991534	0.19702427	0.27383074
P4	0.3841025	-0.128589568	0.383503480	0.1554806	-0.27998632	-0.06265185	0.15982066	0.18064574	-0.14793599	-0.47015673	0.53033508
P5	0.3562116	-0.001317689	0.174050134	-0.2686729	0.02640441	0.65834088	-0.06216255	-0.32844613	0.44978232	0.08860852	0.12093982
P6	0.3318585	-0.034386218	-0.001461324	0.4060333	0.40563282	-0.41785891	-0.42910459	-0.13811902	0.41798887	-0.07959506	0.02050890
P7	0.1667429	-0.288005878	-0.452429765	-0.4673026	-0.23472499	-0.38843822	0.27857362	0.03473335	0.41424431	-0.08780079	0.04569380