Санкт-Петербургский Политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Курсовая работа на тему

«Исследование возможности различения двух выборок с помощью метода главных компонент»

по дисциплине
«Стохастические модели и анализ данных»

Выполнил:
Чернова В.С.
Группа:
3640102/90201

Проверил:
к.ф.-м.н., доцент
Баженов А.Н.

Санкт-Петербург
2020 г

ОГЛАВЛЕНИЕ

ПОСТАНОВКА ЗАДАЧИ

1) Исследовать возможность различать к какой группе относится проба с помощью метода главных компонент;

2) Найти главные компоненты предоставленных выборок.

ТЕОРЕТИЧЕСКОЕ ВВЕДЕНИЕ

Метод главных компонент – это технология многомерного статистического анализа, используемая для сокращения размерности пространства признаков с минимальной потерей полезной информации. Предложен К. Пирсоном в 1901 г., а затем детально разработан американским экономистом и статистиком Г. Хоттелингом.

С математической точки зрения метод главных компонент представляет собой ортогональное линейное преобразование, которое отображает данные из исходного пространство признаков в новое пространство меньшей размерности.

При этом первая ось новой системы координат строится таким образом, чтобы дисперсия данных вдоль неё была бы максимальна. Вторая ось строится ортогонально первой так, чтобы дисперсия данных вдоль неё, была бы максимальной их оставшихся возможных и т.д. Первая ось называется первой главной компонентой, вторая - второй и т.д.

На рис. 1 показано снижение размерности исходного 2-мерного пространства (X_1, X_2) с помощью метода главных компонент до 1-мерного. Первая главная компонента PC_1 ориентирована вдоль направления наибольшей вытянутости эллипсоида рассеяния точек объектов исходного набора данных в пространстве признаков, т.е. с ней связана наибольшая дисперсия.

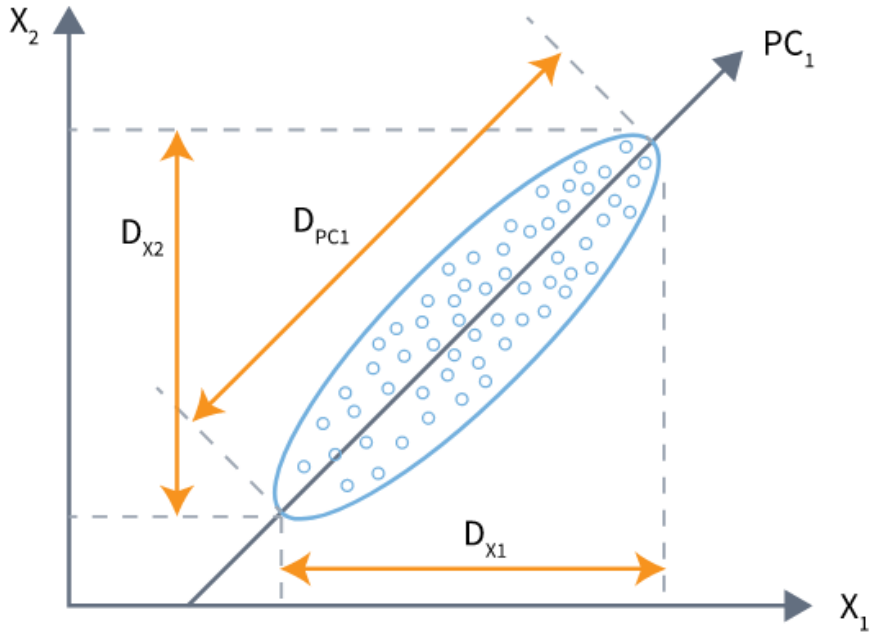

Рис. 1. Снижение размерности 2-мерного пространства до 1-мерного

Задача метода главных компонент заключается в том, чтобы построить новое пространство признаков меньшей размерности, дисперсия между осями которой будет перераспределена так, чтобы максимизировать дисперсию по каждой из них. Для этого выполняется последовательность следующих действий:

- Вычисляется общая дисперсия исходного пространства признаков;
- Вычисляются собственные векторы и собственные значения ковариационной матрицы, определяющие направления главных компонент и величину связанной с ними дисперсии;
- Производится снижение размерности. Диагональные элементы ковариационной матрицы показывают дисперсию по исходной системе координат, а её собственные значения - по новой. Тогда разделив дисперсию, связанную с каждой главной компонентой на сумму дисперсий по всем компонентам, получаем долю дисперсии, связанную с каждой компонентой. После этого отбрасывается столько главных компонент, чтобы доля оставшихся составляла 80-90%.

Основными ограничениями метода главных компонент являются:

- невозможность смысловой интерпретации компонент, поскольку они "вбирают" в себя дисперсию от нескольких исходных переменных;
- метод может работать только с непрерывными данными.

ХОД РАБОТЫ

1. Описание программы

Программа была написана на языке программирования Python 3.7.

Входными данными являются три таблицы формата .xlsx, в которых записаны данные о концентрации различных веществ в некоторой среде. Каждая таблица содержит несколько проб из какого-то определенного места. С помощью метода главных компонент программа понижает размерность данных до двух таким образом, что по ним в последствии можно определить из какого места была взята та или иная проба.

2. Предобработка данных

Перед тем как начать понижать размерность данных с помощью метода главных компонент, необходимо произвести предобработку данных, а именно:

- заполнить пропущенные данные;
- сгладить выбросы;
- произвести стандартизацию (масштабирование) данных.

На место пропусков было решено ставить средние значения по столбцам.

Для сглаживания выбросов была проведена так называемая винсоризация: если значение фактора больше (меньше) квантиля заданного уровня, то этому значению присваивается значение выбранного граничного квантиля.

Масштабирование данных производилось следующим образом: для каждого значения было произведено вычитания среднего по столбцу, а затем деление на стандартное отклонение по столбцу.

3. Применение метода главных компонент

Сначала было определено распределение объясненной дисперсии по главным компонентам. Соответствующая гистограмма представлена на рис. 2. Из рис. 2 видно, что наиболее значимой компонентой является компонента под номером 0, что соответствует метану. Второй по значимости является этан. Следовательно, эти две компоненты мы и будем использовать для понижения размерности данных.

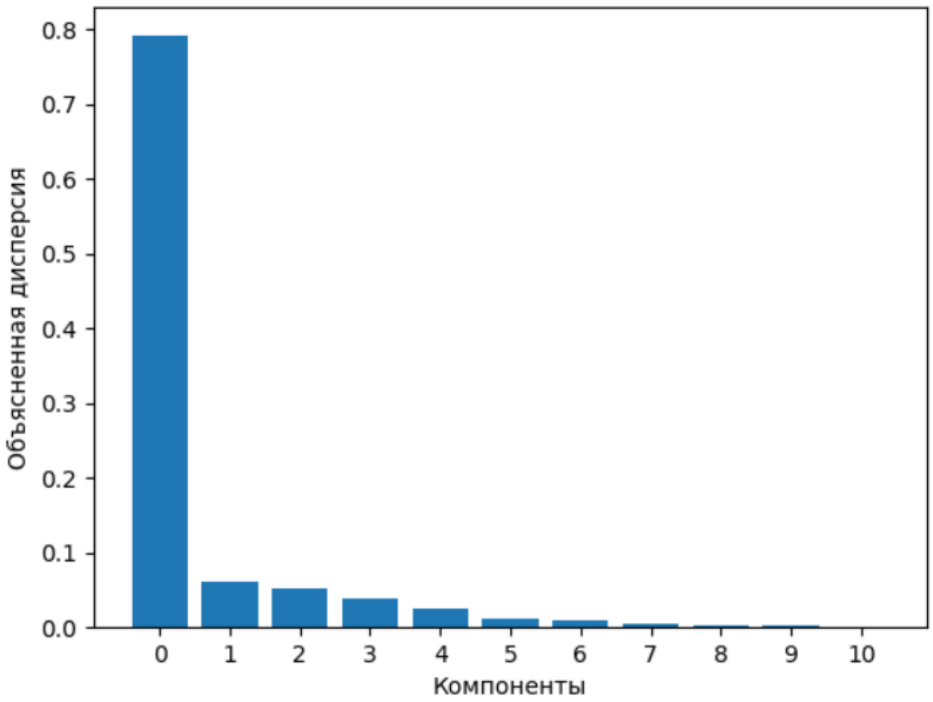

Рис. 2. Распределение объясненной дисперсии по главным компонентам

На рис. 3. Показан результат сжатия данных до двух компонент. Первая компонента PC1 расположена по оси X, а вторая PC2 – по оси Y. Из рисунка видно, что данные образуют три четко разделимых кластера, то есть двух компонент достаточно, чтобы определить к какой местности относятся взятые пробы. Из этого же рисунка можно сделать вывод, что одной самой значимой компоненты было бы недостаточно для разделения данных, так как при проекции точек на ось PC1 или PC2 кластеры накладывались бы друг на друга.

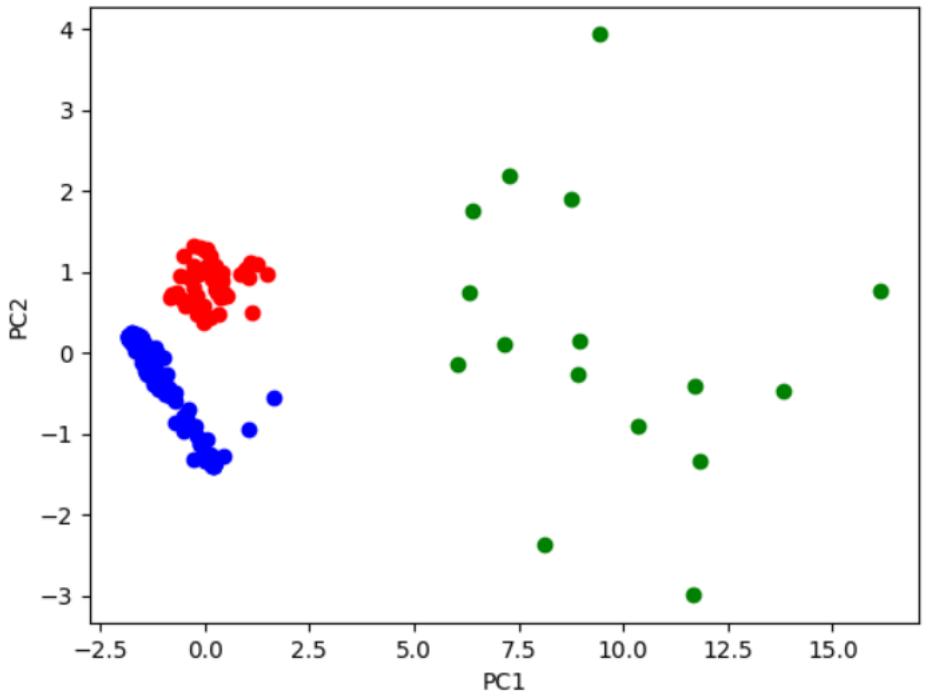

Рис. 3. Разделение выборок по двум компонентам

Далее удалим самую значимую компоненту, которая соответствует метану. Возьмем следующие по значимости после нее две компоненты, а именно этан и этилен. Результат сжатия данных до этих двух компонент представлен на рис. 4.

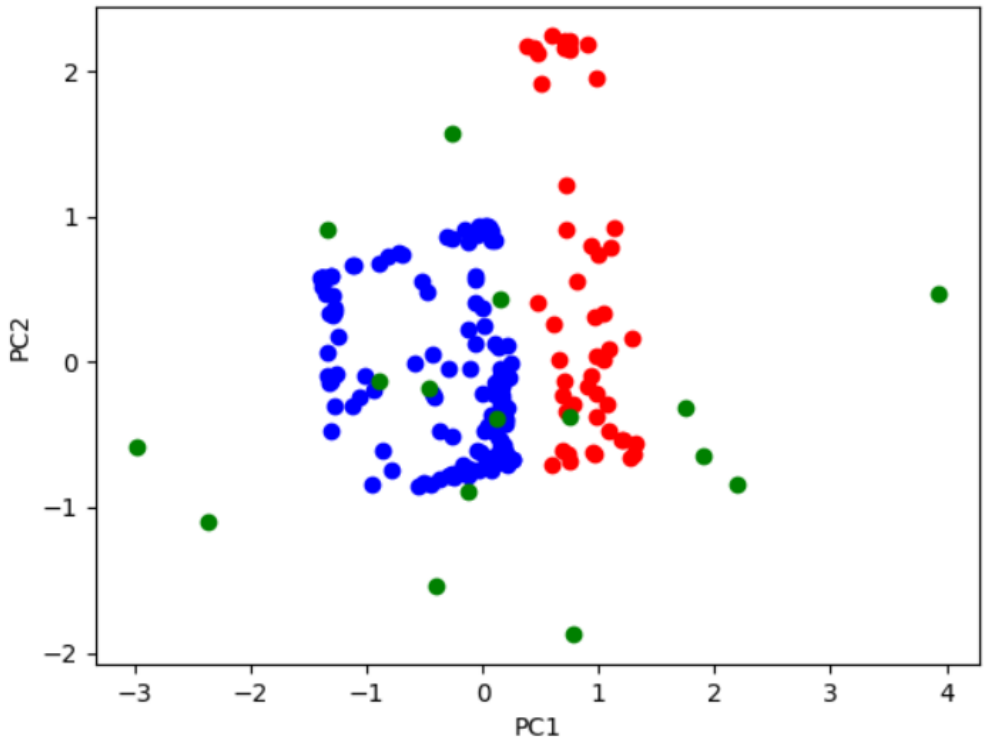

Рис. 4. Разделение выборок по компонентам «этан» и «этилен»

Из рис. 4 видно, что первые две выборки (синяя и красная) все еще хорошо разделимы, но вот третью выборку (зеленую) от них отличить уже практически не удастся.

Попробуем вместо двух компонент использовать три. То есть, к уже выбранным ранее компонентам прибавим третью по значимости, а именно пропан. Построим соответствующий 3D график (рис. 5). По рисунку видно, что точки третьей (зеленой) выборки расположены как-бы вокруг точек из первой и второй выборок. Стоит уточнить, что ни одна из зеленых точек не расположена внутри синей и красной области, в этом можно убедиться посчитав их, на картинке их ровно 16, столько и есть проб в третьей выборке. Однако, среди зеленых точек можно заметить несколько синих точек, поэтому даже используя три компоненты точно разделить данные выборки не получится, но точность все же будет лучше, чем при разделении с помощью двух компонент.

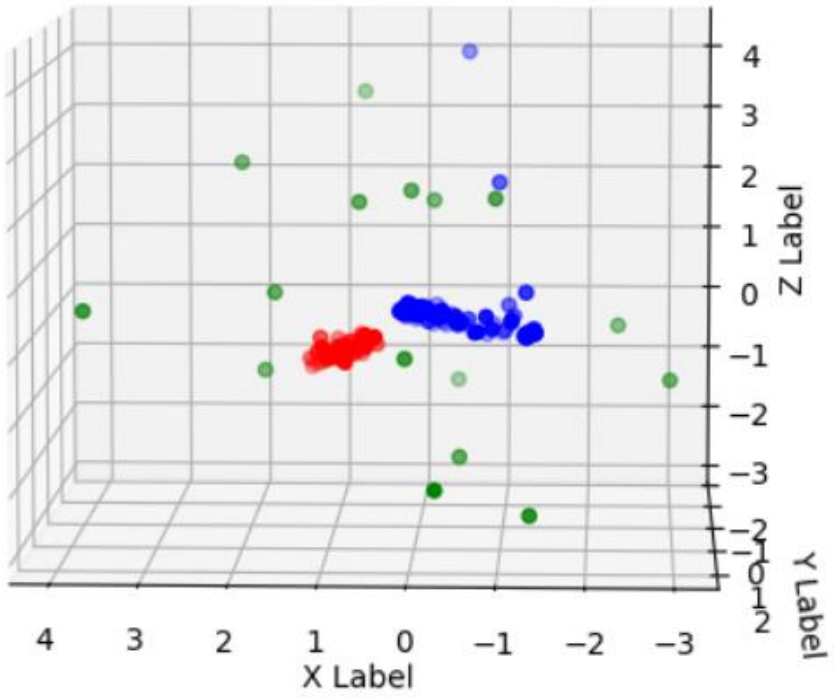

Рис. 5. Разделение выборок по трем компонентам

4. Главные компоненты выборок

Рассмотрим главные компоненты для каждой выборки отдельно. Исследуем какие вещества вносят наибольший вклад в дисперсию данных и, таким образом, вносят решающий вклад в различение двух выборок. Таблицы отсортированы по убыванию вклада в PC1.

Таблицы 1-2. Вклады веществ в PC1 для проб из выборок 1-2

	Пробы из выборки 1	
№	Вещество	Вклад в PC1
1	метан	0.999
2	этан	0.032
3	пропан	0.025
4	n-бутан	0.009
5	n-пентан	0.003
6	пропилен	0.002
7	i-бутан	0.001
8	этилен	0.001
9	i-бутилен	0.0001
10	i-пентан	0.00004
11	бутен-1	0.00003

	Пробы из выборки 2	
№	Вещество	Вклад в PC1
1	метан	0.999
2	этан	0.023
3	n-пентан	0.0058
4	пропан	0.0024
5	n-бутан	0.0023
6	этилен	0.0021
7	i-бутан	0.0017
8	i-пентан	-0.0017
9	i-бутилен	0.00024
10	бутен-1	-0.00013
11	пропилен	0.000006

Таблицы 3-4. Вклады веществ в PC1 для проб из выборки 3
и для объединенных выборок

	Пробы из выборки 3	
№	Вещество	Вклад в PC1
1	метан	0.984
2	этан	0.163
3	пропан	0.061
4	пропилен	0.023
5	этилен	0.016
6	n-бутан	0.010
7	i-бутилен	0.0063
8	i-пентан	0.0048
9	i-бутан	0.0038
10	бутен-1	0.0037
11	n-пентан	0.0029

	Объединенные выборки 1-3	
№	Вещество	Вклад в PC1
1	пропилен	0.330
2	этилен	0.325
3	этан	0.324
4	i-бутан	0.315
5	пропан	0.313
6	n-бутан	0.311
7	i-пентан	0.305
8	бутен-1	0.300
9	i-бутилен	0.272
10	n-пентан	0.271
11	метан	0.230

Из таблиц 1-3 видно, что наиболее значимым признаком для всех выборок является метан, а вторым по значимости – этан. Остальные же признаки вносят совсем малый вклад в дисперсию данных.

Из таблицы 4 можно сделать вывод, что в случае объединенных выборок все признаки являются примерно одинаковыми по значимости.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы был изучен и реализован метод главных компонент. С помощью данного метода была исследована возможность различать к какой группе относится проба. Кроме того, были исследованы главные компоненты каждой выборки в отдельности.

Был сделан вывод, что для первых двух выборок двух компонент будет достаточно для возможности их различия. Однако, при добавлении третьей выборки для возможности их различия необходимо ввести третью компоненту.

Для рассмотренных выборок было определено, что наиболее значимым признаком для всех выборок является метан, а вторым по значимости – этан. Остальные же признаки вносят совсем малый вклад в дисперсию данных. А в случае объединенных выборок все признаки являются примерно одинаковыми по значимости.

Код написанной программы представлен по следующей ссылке: https://github.com/nika2506/stochastic_labs/tree/main/Coursework.

СПИСОК ЛИТЕРАТУРЫ

1. Метод главных компонент [Электронный ресурс], URL: https://wiki.loginom.ru/articles/principal-component-analysis.html. Дата обращения: 20.12.2020

2. Winsorizing [Электронный ресурс], URL: https://en.wikipedia.org/wiki/Winsorizing. Дата обращения: 20.12.2020

3. SciPy Tutorial [Электронный ресурс], URL: https://docs.scipy.org/doc/scipy/reference/index.html. Дата обращения: 20.12.2020