Курсовая работа

«Применение метода главных компонент для классификации выборок»

по дисциплине

«Стохастические модели и анализ данных»

Выполнила: Кацман Н.И.

группа: 3640102/90201

Проверил: к.ф.-м.н., доцент

Баженов А.Н.

Оглавление

[Постановка задачи 3](#_Toc67966506)

[Теоретическая часть 3](#_Toc67966507)

[Метод главых компонент (МГК) 3](#_Toc67966508)

[Выбор числа главных компонент 3](#_Toc67966509)

[Подготовка данных 4](#_Toc67966510)

[Практическая часть 4](#_Toc67966511)

[Импорт данных 4](#_Toc67966512)

[Приведение таблицы к удобному для импорта виду 4](#_Toc67966513)

[Исключение столбцов, которые не потребуются 5](#_Toc67966514)

[Исключение строк с пустыми значениями 5](#_Toc67966515)

[Импорт 5](#_Toc67966516)

[Программная предобработка данных 5](#_Toc67966517)

[Нормализация по объёму, в котором посчитаны содержания газов 5](#_Toc67966518)

[Центрирование и нормирование данных 5](#_Toc67966519)

[Применение метода главных компонент 6](#_Toc67966520)

[Коэффициенты корреляции Спирамена 7](#_Toc67966521)

[Заключение 8](#_Toc67966522)

[Список литературы 8](#_Toc67966523)

# Постановка задачи

Имеется exel файл с данными геолого-биологичеких проб.

В ходе работы требуется:

1. Считать данные из этого файла
2. Провести предварительную обработку данных (так чтобы для них имело смысл использовать МГК)
3. Применить МГК к подготовленным данным
4. Проанализировать результаты работы МГК

# Теоретическая часть

## Метод главых компонент (МГК)

*Метод Главных Компонент* (англ. Principal Components Analysis, PCA) — один из основных способов уменьшить [размерность](http://www.machinelearning.ru/wiki/index.php?title=%D0%A0%D0%B0%D0%B7%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D1%8C&action=edit" \o "Размерность) данных, потеряв наименьшее количество [информации](http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D1%8F&action=edit).

В общем, многомерном случае, процесс выделения главных компонент происходит так:

1. Ищется центр облака данных, и туда переносится новое начало координат –  это нулевая главная компонента (PC0)
2. Выбирается направление максимального изменения данных – это первая главная компонента (PC1)
3. Если данные описаны не полностью (шум велик), то выбирается еще одно направление (PC2) – перпендикулярное к первому, так чтобы описать оставшееся изменение в данных и т.д.

В результате, мы переходим от большого количества переменных к новому представлению, размерность которого значительно меньше. Часто удается упростить данные на порядки: от 1000 переменных перейти всего к двум. При этом ничего не выбрасывается – все переменные учитываются. В то же время несущественная для сути дела часть данных отделяется, превращается в шум. Найденные главные компоненты и дают нам искомые скрытые переменные, управляющие устройством данных.{\displaystyle x}

## Выбор числа главных компонент

Обычно считается, что нужно извлекать столько главных компонент, сколько можем содержательно проинтерпретировать. Но, конечно, существуют и более формальные критерии.

Так, например, по правилу Кайзера нужно извлекать столько главных компонент, сколько имеется компонент с дисперсией (собственным числом) более 1.

Другое эмпирическое правило гласит, что нужно извлекать столько главных компонент, сколько смогут объяснить хотя бы 70%-80% дисперсии наших исходных данных.

Еще существует правило Кеттела, которое имеет логику, схожую с методом согнутого колена (локтя) в кластерном анализе. Нужно построить график, по горизонтальной оси которого отложены номера главных компонент, а по вертикальной – их дисперсии (собственные значения ковариационной матрицы) и остановиться нужной на той главной компоненте, которая находится на “изгибе” или до него.

## Подготовка данных

Во многих случаях, перед применением PCA, исходные данные нужно предварительно подготовить: отцентрировать и/или отнормировать. Эти преобразования проводятся по столбцам – переменным.

*Центрирование* – это вычитание из каждого столбца **x***j* среднего (по столбцу) значения

.

Центрирование необходимо потому, что оригинальная PCA модель ([2](https://rcs.chemometrics.ru/old/Tutorials/pca.htm" \l "Eq2)) не содержит свободного члена.

Второе простейшее преобразование данных – это *нормирование*. Это преобразование выравнивает вклад разных переменных в PCA модель. При этом преобразовании каждый столбец **x***j* делится на свое стандартное отклонение.



# Практическая часть

Для выполнения данный работы был выбран язык **R.**

Для выполнения данного курсового проекта был предоставлен файл с данными геолого-биологических проб. В этом файле четыре листа данных, оформленных разным способом и внушающих более/менее положительное впечатление. Лучше всего оформлен првый лист, так что его и будеи использовать (так же в нём и существенно больше данных)

## Импорт данных

Данные для исследования представлены в формате .xlsx, что не совсем удобно, так что начнём с экспорта их в .csv формат. Возможность такого экспорта встроена в Microsoft Excel, ей и воспользуемся.

### Приведение таблицы к удобному для импорта виду

Чтобы импортировать имеющуюся таблицу в среду R, нужно немного модифицировать её. Необходимыми являются следующие действия:

* Доименовать столбцы. Последний столбец в исходных данных не имеет полноценного названия. И лучше для удобства из трёх строк на названия оставить только одну.
* Изменить формат хранения дробных чисел. Необходимо чтобы разделителем была точка, а не запятая, так как запятая – это разделитель колонок в формате .csv

### Исключение столбцов, которые не потребуются

Так как нам уже оказалось нужно форматировать таблицу до её импорта, то заодно удалим и колонки, которые нам не нужны. В работе планируется использовать в качестве признаков, только доли содержания разных газов в пробе, так что почти все колонки, кроме отвечающих за это содержание, удалим. Почти все, так как следует оставить ещё колонку с объёмами проб газа, в которых эти содержания были вычислены (эта колонка пригодится в следующей части работы).

### Исключение строк с пустыми значениями

Ещё одним действием, которое проделаем до импорта будет удаление из таблицы всех строк, в которых есть пропуск значений.

### Импорт

После проведения всех вышеупомянутых действий можем импортировать наш .csv файл в Среду разработки R.

## Программная предобработка данных

### Нормализация по объёму, в котором посчитаны содержания газов

Посмотрев на таблицу, можем заключить, что количества газов в пробах - это содержание этих газов в каком-то объёме пробы, причём объёмы эти для разных проб различны. Так что было бы хорошо пересчитать содержания газов в пробах так, чтобы это были содержания газов в одном и том же объёме пробы.

Выберем объём газа записанный для самой первой пробы () – будем для всех проб пересчитывать данные, так чтобы они соответствовали именно .  
Пробежимся по всем пробам.   
Для i-той пробы коэффициент преобразования . Пересчитаем для i-той пробы содержание каждого газа в ней, домножив его на .

Теперь столбик объемов содержит одинаковые значения для всех пробы, так что его можно удалить.

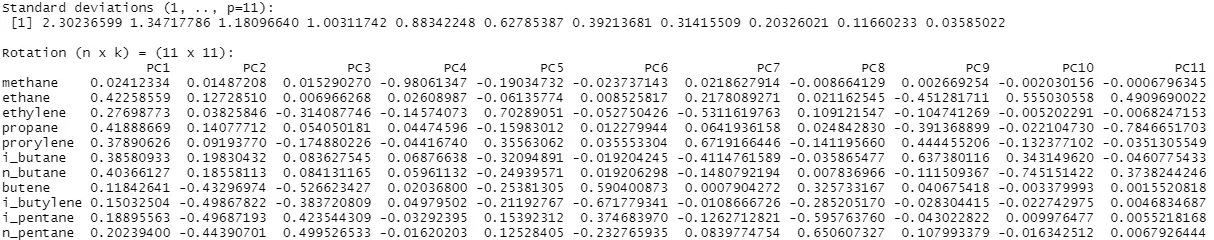
### Центрирование и нормирование данных

Оно отдельно не производится, так как во встроенной функции языка R есть возможность задать параметрами то, что данные надо центрировать и нормировать.

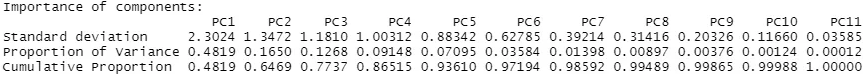
### Применение метода главных компонент

Используем встроенную функцию языка R ***prcomp*** для нахождеия главных компонент и дополнительной информации по ним.

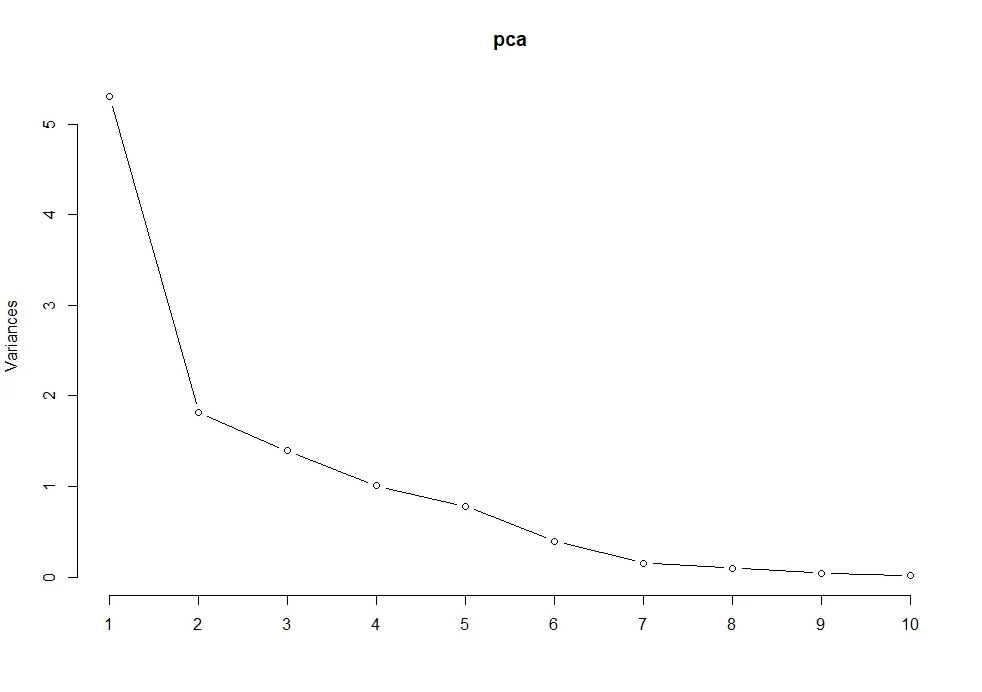
Ниже представлены стандартные отклонения главных компонент и матрица перехода от изначальных признаков к новым (т.е. главным компонентам)

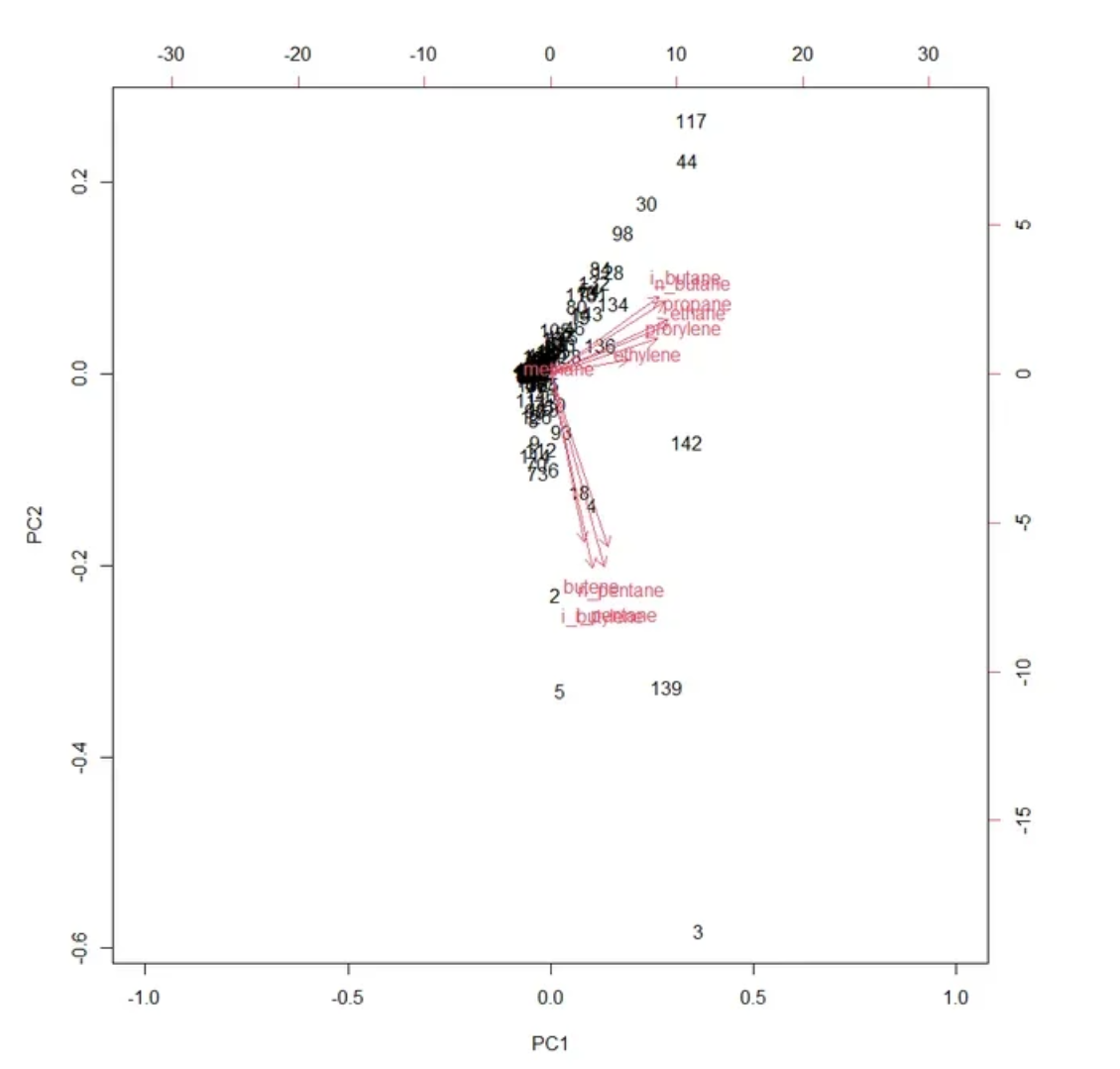


Так же полезны будут и данные о долях дисперсии исходных данных, которую объясняет каждая главная компонента, и кумулятивная доля объясненной дисперсии – дисперсия исходных данных, которая объясняется текущей и предыдущими главными компонентами. Эти данные придставлены на ниже



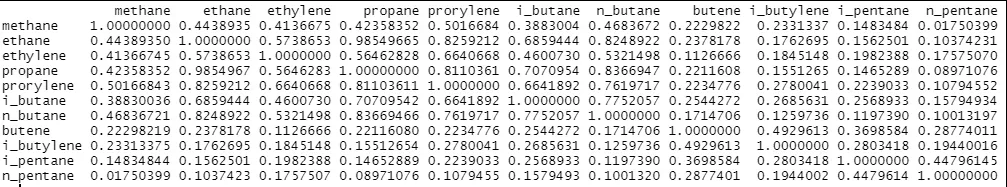
Интересным будет ещё взглянуть на изображение дисперсий главных компонент в виде графика от их номера. В теоретической части упоминалось правило Коттела выбора существенных главных компонент. В нём речь идёт как раз о таком графике. Собственно, его можно увидеть ниже.



На следующем графике представлены номаера проб и проекции «осей» первоначальных признаков на плоскость первых двух главных компонент. На этом графике проекции «осей» первоначальных признаков изображены красными векторами и подписаны.

### Коэффициенты корреляции Спирамена

Ради интереса и дополнительной информации так же вычислим коэффициенты корреляции Спирмена. По ним можно будет предположить о взаимосвязанности некоторых признаков, а если, например, мы знаем, что признаки не связаны, то они вероятнее всего не окажутся одновременно ни в одной из первых главных компонент.

Ниже представлены вычесленные корреляции  


# Заключение

На основе полученных результатов и графиков можно сделать следующие выводы:

1. Полученный набор данных не слишком хорошо поддаётся сокращению размерности. Можно сократить её до двух, но это приведёт к существенной потере инфомации. Скорее всего, достаточно неплохо информацию из исходных данных сохранит сокращение размерности до четырёх. В таком случае будет покрыто около 86% дисперсии исходных данных.  
   Недостаточность первых двух главных компонент в качестве единственных новых признаков ярко показывает последний график. Можно заметить, что разброс данных на нём не сильно удачен: большинство проб кучкуется рядом с точкой (0, 0), тогда как если бы рассматривались все изначальные признаки, вероятно они бы основательно отличались.
2. Можно заметить связь между корреляциями Спирмена данных и полученными главными векторами. Например, для этана в таблице корреляций явно выделяются высокие значения корреляции с пропаном, пропиленом и n-бутаном. И если мы посмотрим на последний график, то увидим, что как раз этан, пропан, припилен и n-бутан вносят наибольший вклад во вторую главную компоненту.
3. Похоже, что правило Коттела не всегда стоит использовать для выбора количества главных компонент: построенный график довольно явно заявляет, что можно, следуя правилу Коттела, ограничиться превыми двумя главными компонентами, но выше уже сделан вывод, о том что эти две компоненты не учмтывают существенное количество информации.

Файл кода начальные и обработанные данные и этот отчёт можно посмотреть [здесь](https://github.com/Hadegda/Stochastic-labs/tree/main/PCA).

# Список литературы

1. Метод главных компонент [Электронный ресурс]   
   url: <https://rcs.chemometrics.ru/old/Tutorials/pca.htm>

Дата обращения: 29.03.21

1. Метод главных компонент [Электронный ресурс]   
   url: https://rpubs.com/AllaT/pca-intro

Дата обращения: 29.03.21