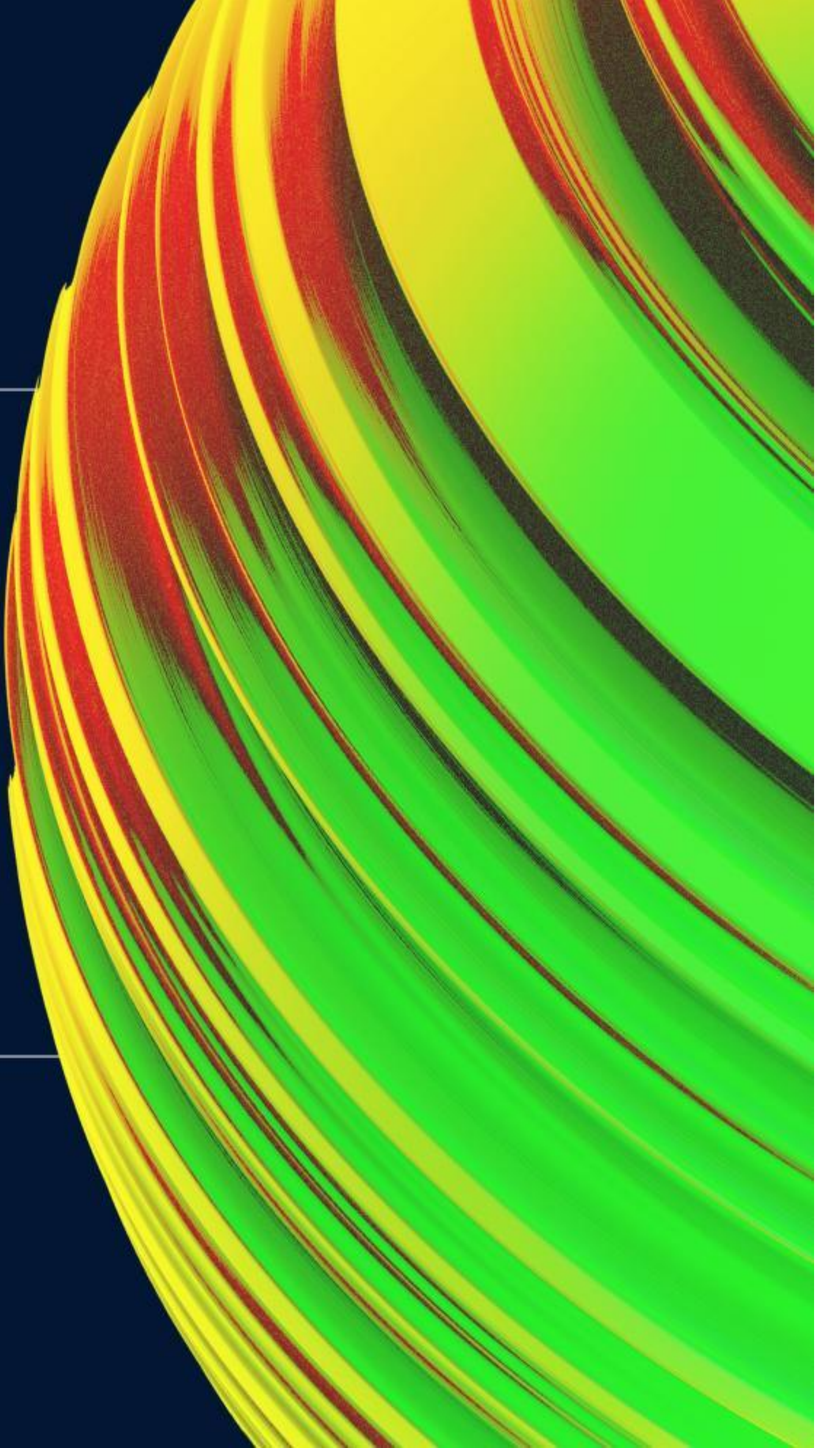
УНИВЕРСИТЕТ ИННОПОЛИС

Снижение размерности. PCA

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- @SparrowMaria



Проклятье размерности. Простой пример

Представьте, что вы находитесь в небольшой комнате и вы пытаетесь найти своего друга. В комнате всего один стол и стул, поэтому вам легко будет найти друга, потому что у вас мало места для поиска.

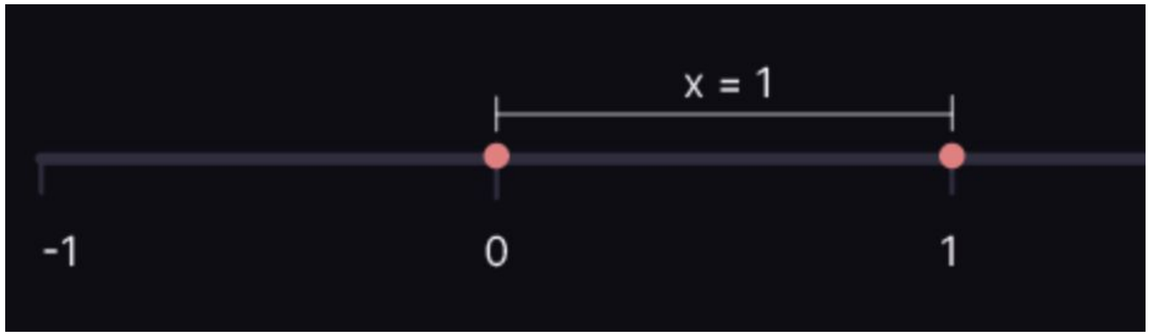
Совершенно другая ситуация, когда вы находитесь на огромном футбольном стадионе, и ваш друг тоже где-то там. Стадион в тысячу раз больше, чем комната, и вам гораздо сложнее найти друга, потому что у вас гораздо больше пространства для поиска.

Это и есть суть "проклятия размерности" — когда у вас много признаков (или измерений), поиск нужной информации становится очень трудным.

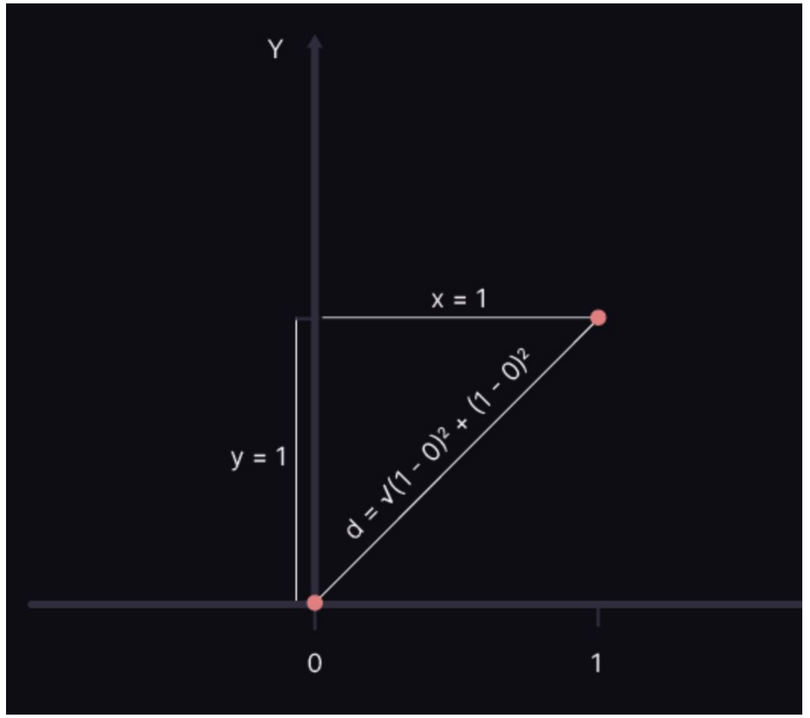
Проклятье размерности

Термин «**проклятие размерности**» в 1961 году ввел американский математик Ричард Беллман.

Предположим, у нас есть две точки на прямой, 0 и 1. Эти две точки находятся на расстоянии друг от друга =1 Теперь мы вводим вторую ось Y – второе измерение. Положение точек определяется теперь списком из двух чисел – (0,0) и (1,1). Расстояние между точками теперь подсчитывается с помощью Евклидова расстояния и оно равно 1.44. В трехмерном пространстве будет 1.73

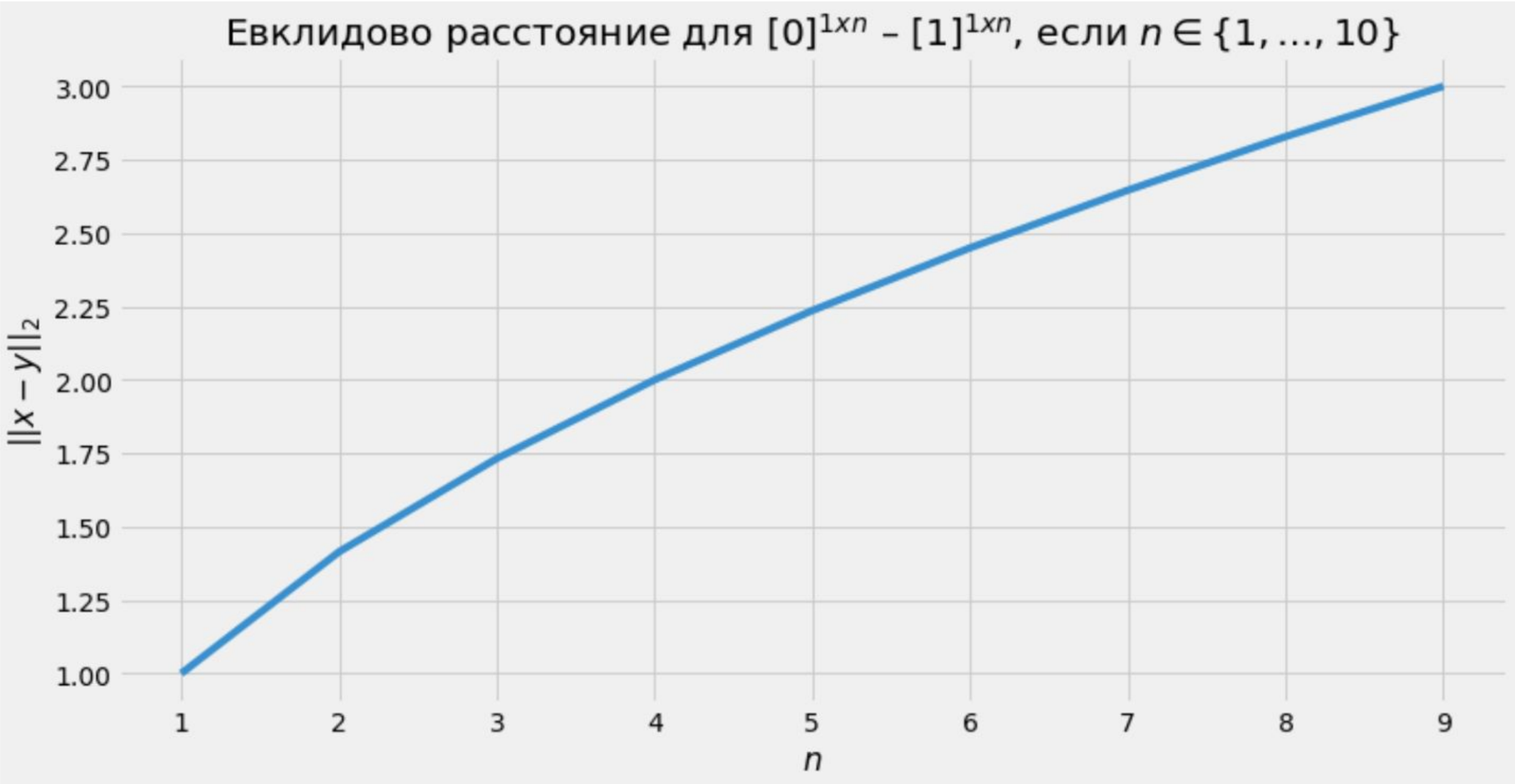


можно почитать - https://www.helenkapatsa.ru/prokliatiie-razmiernostiei



Проклятье размерности

Чем больше размерность, тем больше расстояние между точками (0,..0) и (1,..,1)



Как снизить размерность?

INNOPOLIS UNIVERSITY

- удалять признаки, которые слабо коррелируют с целевой переменной
- выбрасывать признаки по одному и проверять качество модели на тестовой выборке
- перебирать случайные подмножества признаков в поисках лучших наборов
- использовать регуляризацию L1
- использовать метод главных компонент (principal component analysis, PCA)

Метод главных компонент

Метод главных компонент (Principal Component Analysis или же PCA) — алгоритм **обучения без учителя**, используемый для понижения размерности и выявления наиболее информативных признаков в данных.

Суть метода заключается в предположении о **линейной зависимости данных и их проекции на подпространство ортогональных векторов**, в которых дисперсия будет максимальной.

Метод главных компонент

PCA находит новые признаки, которые представляют собой комбинации старых и при этом несут максимум информации.

Эти новые признаки называются главными компонентами.

Простой пример: вы нашли главные маршруты в большом городе, которые помогают быстро добраться до любого места.

Методы понижения размерности

1-ая главная компонента должна иметь наибольшую изменчивость, то есть должна быть наибольшая выборочная дисперсия

Биология	Математика
4	5
4	2
4	5
4	4
4	3
4	4
3	3
5	3

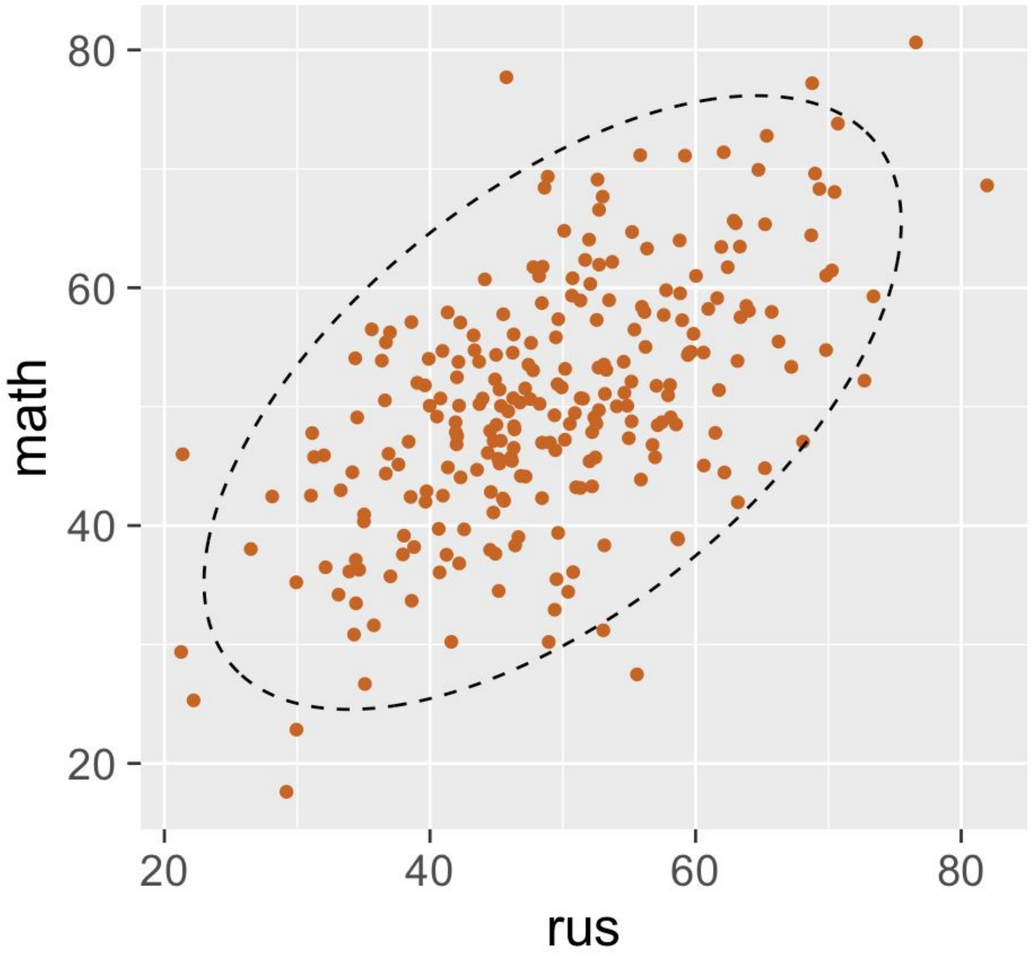
То есть первая главная компонента будет - оценка по математике

2-ая главная компонента - оценка по биологии

лекция https://www.youtube.com/watch?v=NKmwnILrHD8

Метод главных компонент на простом примере

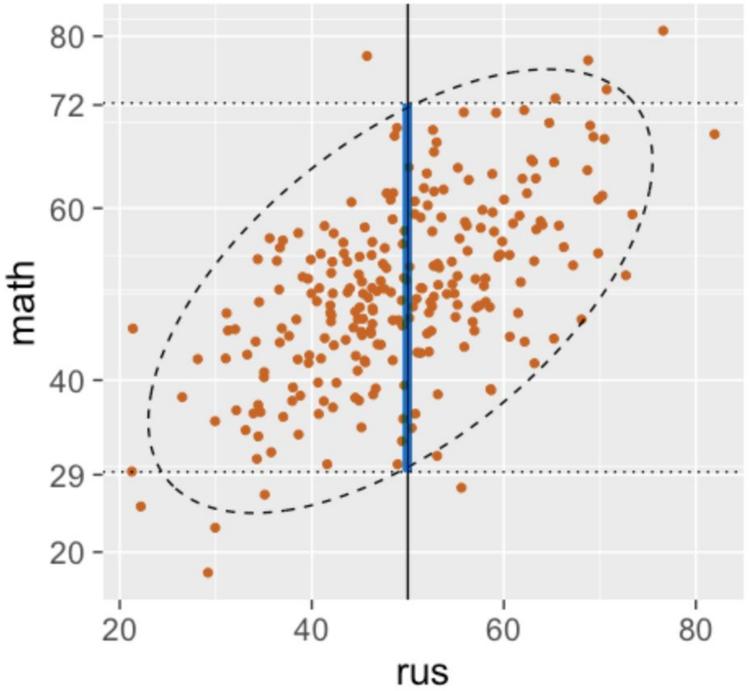
Рис. 1



Хотим уменьшить размерность и студентов оценивать по 1 признаку, например, по оценке по русскому

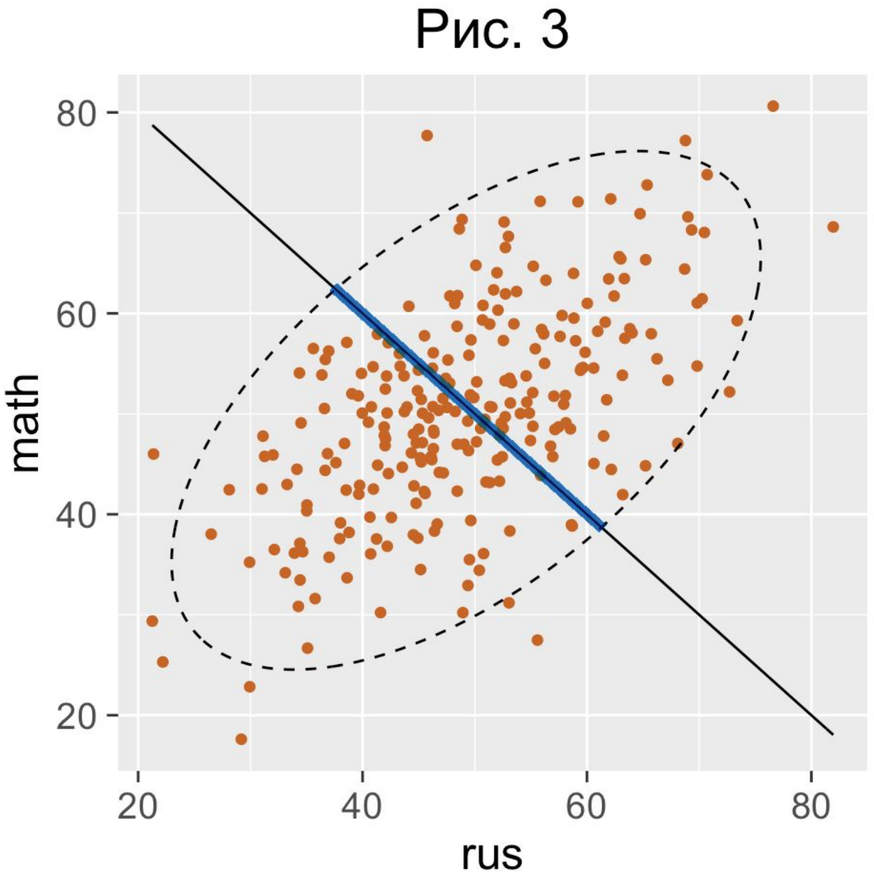
Таким образом мы много теряем, зная оценку по русскому, мы ничего не знаем о матем способностях

Рис. 2



http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html

Метод главных компонент на простом примере



Оценивать ученика по 1-ой оценке плохо, давайте будем оценивать по $PC_1 = rus + math$

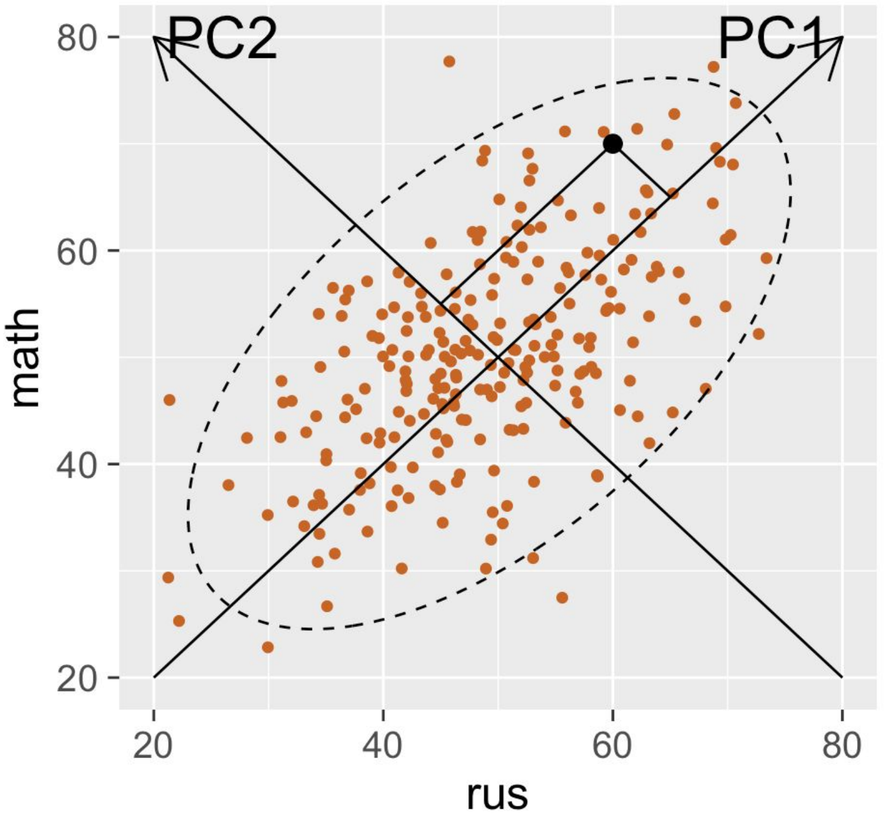
Измерять степень нашего незнания длиной того самого отрезка, на котором может оказаться школьник, то мы видим, что она уменьшилась: новый отрезок короче старого, потому что сейчас мы пересекаем эллипс «поперек», а раньше пересекали «наискосок».

Поэтому сообщать наше число PC1 лучше, чем сообщать только одну из оценок (если, конечно, мы не знаем заранее, что получателю этой информации какая-то из двух оценок важнее другой).

http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html

Метод главных компонент на простом примере

Рис. 4



Метод главных компонент — это история про введение новой, более экономной системы координат, в которой описывать наши данные проще. Вот как эта система координат будет устроена в нашем примере с оценками

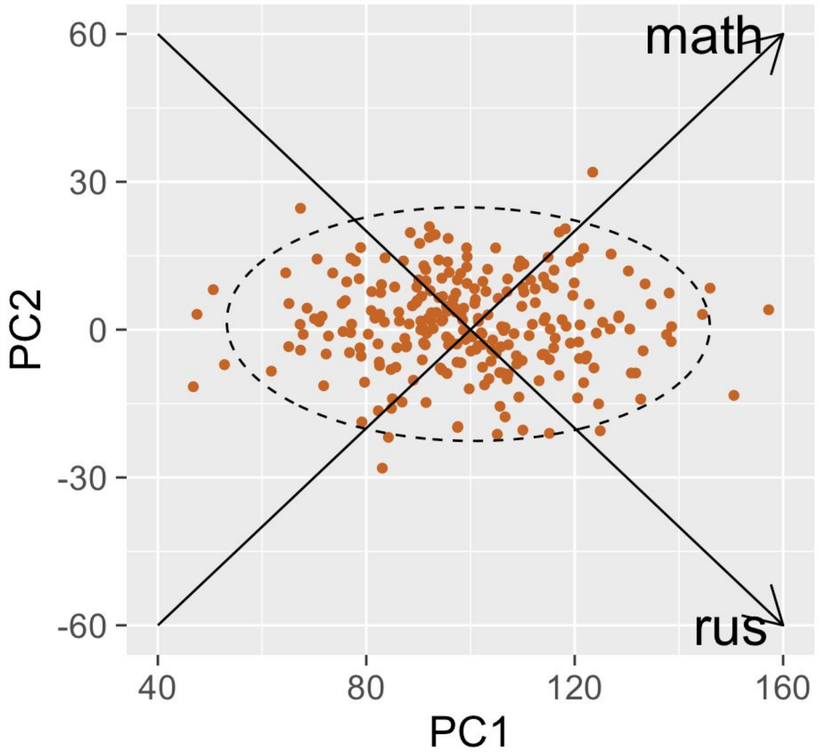
В качестве первой координаты точки мы возмём PC1

, то есть сумму её старых координат, а в качестве второй координаты (обозначим её через PC2) возьмём разность её старых координат:

http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html

Метод главных компонент на простом примере

Рис. 5



Теперь перейдем к координама PC1 и PC2

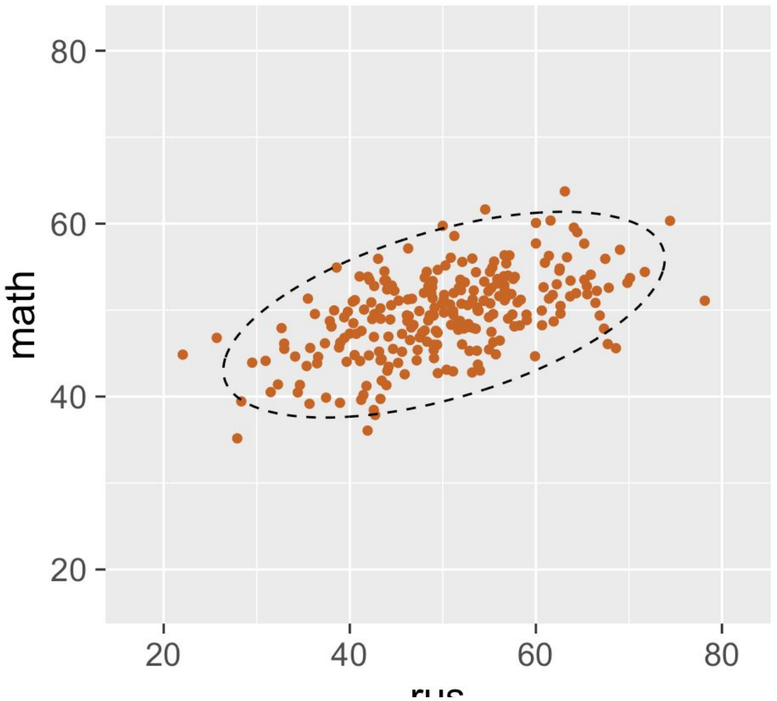
Заметим, что на новой картинке PC1 и PC2 имеют нулевую корреляцию: раньше мы знали, что школьник, хорошо успевающий по русскому, скорее всего имеет неплохую оценку и по математике, а сейчас знание PC1 ничего не говорит нам о том, велик или мал PC2. Геометрически это соответствует тому, что эллипс теперь не имеет никакого ярко выраженного наклона.

В новой системы координат мы избавились от зависимостей между переменными.

http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html

Метод главных компонент на простом примере

Рис. 6



Ситуация меняется, когда оценки неравноправны

Здесь эллипс повёрнут не на 45 градусов, а на меньший угол — его длинная ось лишь немножко отклоняется от горизонтали. Если бы мы ввели такие же координаты PC1 и PC2, как и раньше, они бы не были оптимальными. Раньше координаты rus и math были равноправными, а теперь они явно неравноправны: оценка по русскому содержит больше информации о школьнике, чем оценка по математике. Это связано с тем фактом, что разброс оценок по математике в этом случае гораздо меньше разброса оценок по горизонтали. Мы хотим выбрать ось PC1 таким образом, чтобы разброс значений по этой оси был максимально возможным (чтобы она содержала максимум информации), то есть вдоль длинной оси эллипса (см. конец предыдущего параграфа).

http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html

Метод главных компонент

PCA используется для уменьшения размерности данных, сохраняя как можно больше вариации в данных.

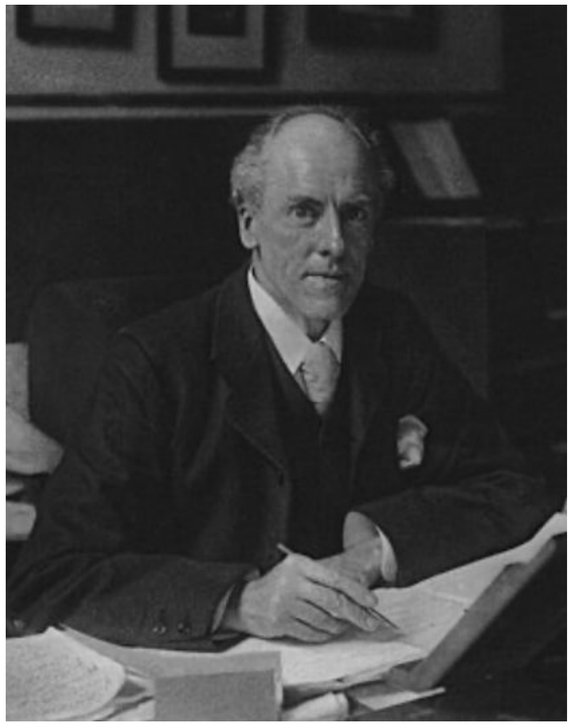
Это достигается путем проекции данных на новые ортогональные оси (главные компоненты), которые направлены в направлении наибольшей дисперсии данных.

PCA находит собственные значения и собственные векторы ковариационной матрицы данных. Главные компоненты — это собственные векторы ковариационной матрицы, отсортированные по величине собственных значений.

Метод главных компонент

- Уменьшение размерности
- Борьба с мультиколлинеарностью
- Ускорение вычислений

Методы понижения размерности



PCA - один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.

Изобретён Карлом Пирсоном в 1901 году.
Применяется во многих областях, в том числе в эконометрике, биоинформатике, обработке изображений, для сжатия данных, в общественных науках

УНИВЕРСИТЕТ
ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ