

Метод главных компонент

Баранов Н. М.
19.Б13-ПУ

Проклятие размерности



Ричард Эрнест Бёллман (26 августа 1920, Нью-Йорк, США — 19 марта 1984, Лос-Анджелес, США) — американский математик, один из ведущих специалистов в области математики и вычислительной техники. Ввел термин «проклятие размерности» в 1961 году.

Пример проклятия размерности

Рассмотрим единичный интервал $[0,1]$. 100 равномерно разбросанных точек будет достаточно, чтобы покрыть этот интервал с частотой не менее 0,01.

Теперь рассмотрим 10-мерный куб. Для достижения той же степени покрытия потребуется уже 10^{20} точек. То есть, по сравнению с одномерным пространством, требуется в 10^{18} раз больше точек.

Поэтому, например, использование переборных алгоритмов становится неэффективным при возрастании размерности системы.

Сферы ВОЗНИКНОВЕНИЯ

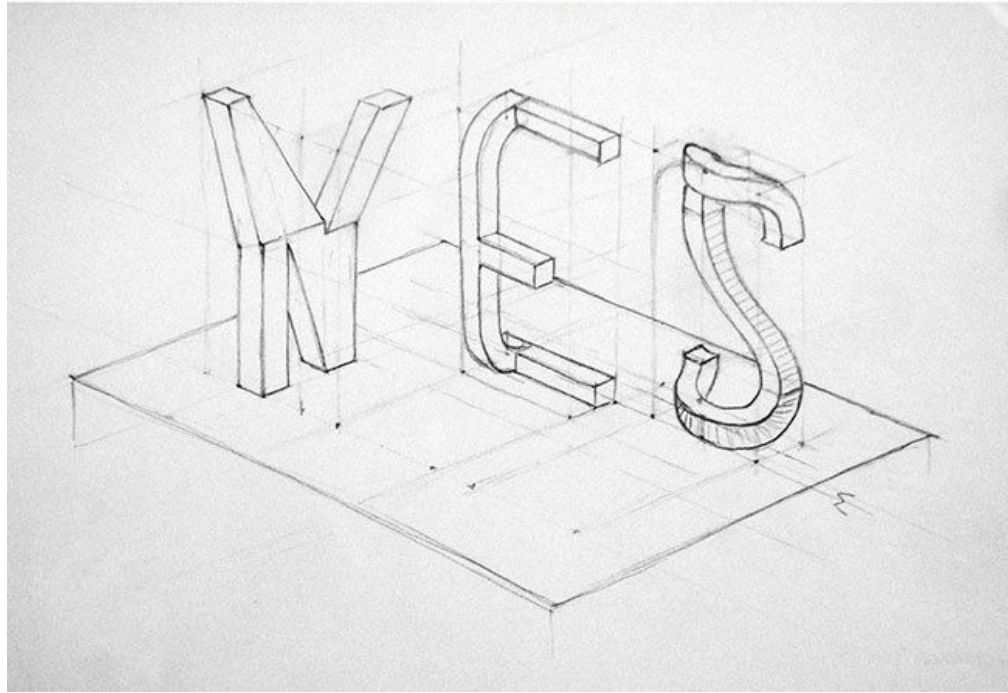
- Машинное обучение
- Задачи распознавания
- Задачи оптимизации
- Комбинаторная геометрия
- Работа со сложными системами

Трудности при работе со СЛОЖНЫМИ системами

- Трудоемкость вычислений
- Необходимость хранения огромного количества данных
- Увеличение доли шумов

Способ решения

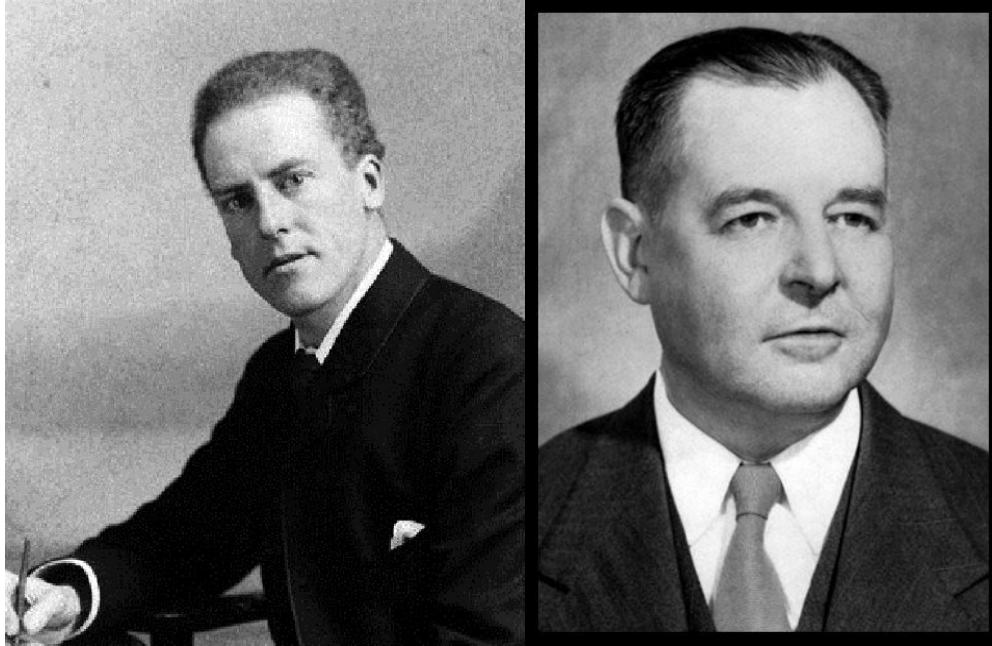
Основная идея при решении проблемы — понизить размерность пространства, а именно спроецировать данные на подпространство меньшей размерности. На этой идее и основан метод главных компонент.



Пример проецирования данных на подпространство меньшей размерности

История появления

Карл Пирсон (27 марта 1857, Лондон – 27 апреля 1936, Лондон) – английский математик, статистик, биолог и философ; основатель математической статистики, один из основоположников биометрики. Предложил идею метода главных компонент в 1901. В русскоязычных источниках его иногда называют Чарлз Пирсон.

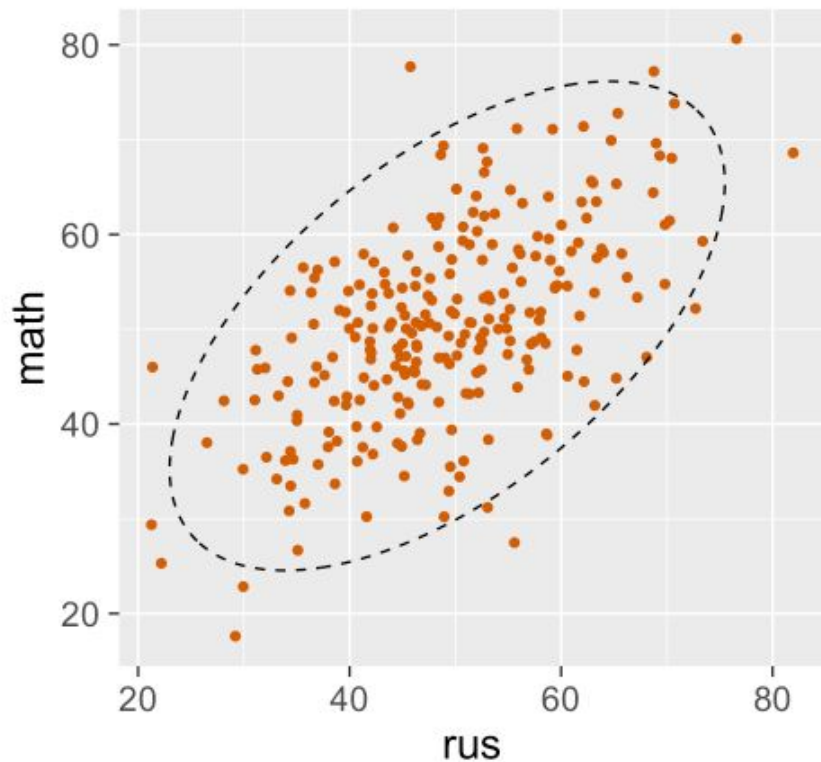


Гарольд Хотеллинг (29 сентября 1895, Фулда, Миннесота – 26 декабря 1973, Чапел-Хилл, Северная Каролина) – американский экономист и статистик. Детально разработал метод главных компонент, предложенный Карлом Пирсоном.

Эквивалентные постановки метода главных компонент

1. Аппроксимировать данные линейными многообразиями меньшей размерности
2. Найти подпространства меньшей размерности, в ортогональной проекции на которые разброс данных максимален
3. Найти подпространства меньшей размерности, в ортогональной проекции на которые среднеквадратичное расстояние между точками максимально
4. Для данной многомерной случайной величины построить такое ортогональное преобразование координат, что в результате корреляции между отдельными координатами обратятся в ноль.

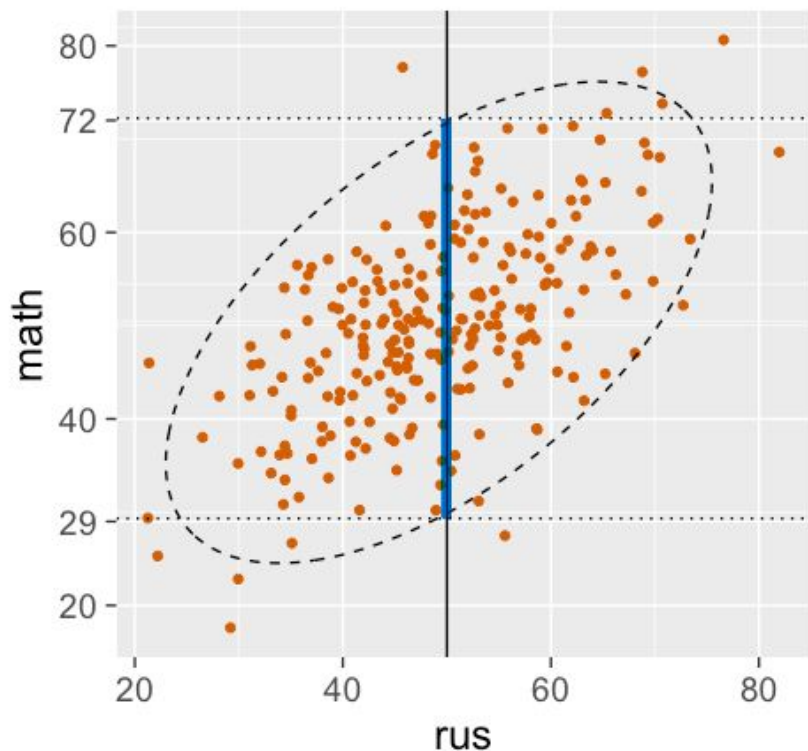
Пример “Школьные оценки”



Пусть у нас имеются результаты теста для школьников по двум предметам — например, по русскому языку и математике. Тогда мы можем построить по этим результатам график.

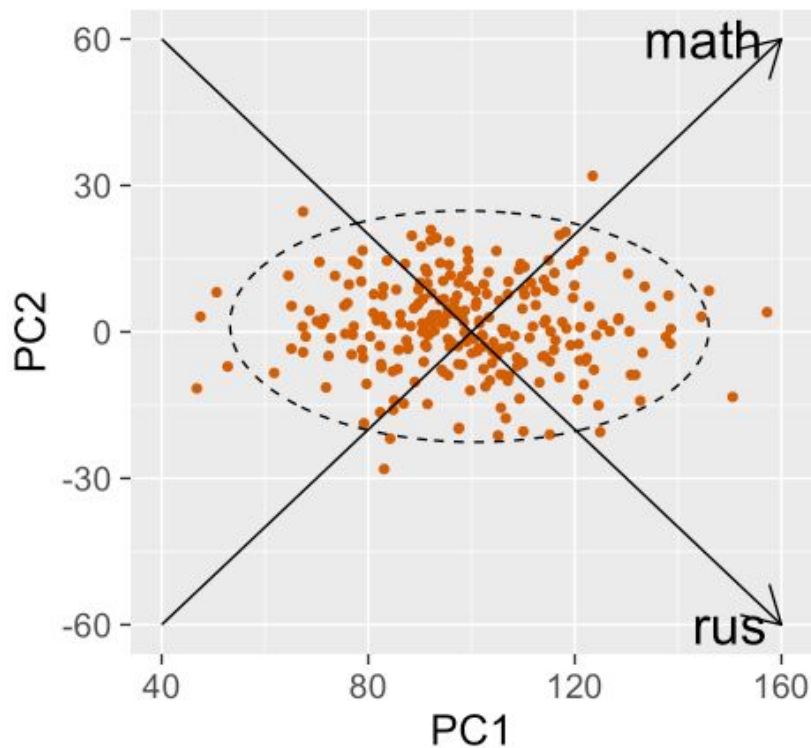
Предположим, что нам надо уменьшить размерность — вместо двух чисел на каждого школьника хранить только одно число.

Пример “Школьные оценки”



Мы можем отбросить одну из переменных и оставить другую. Например, можно записывать в аттестат только оценку по русскому языку, а оценку по математике игнорировать. Но в таком случае мы потеряем слишком много информации.

Пример “Школьные оценки”



Нам необходимо выбрать такую систему координат, в которой мы сможем избавиться от зависимостей между переменными. Именно благодаря этому новая система координат будет «экономнее» старой и мы можем выделить в ней переменную PC1, содержащую большую часть информации.

ОСНОВНЫЕ ПОНЯТИЯ

- Линейное многообразие

$$M = \{v + x \mid x \in L\}$$

- Линейная комбинация
- Ортонормированная система
- Ортогональное преобразование
- Ковариационная матрица

$$\text{Cov}(X_i, X_j) = E[(X_i - E(X_i)) * (X_j - E(X_j))]$$

Аппроксимация данных линейными многообразиями

Дано: $x_1, x_2, \dots, x_m \in R^n, k = 1, 2, \dots, n - 1$

Найти: $L_k \subset R^n : \sum_{i=1}^m \text{dist}^2(x_i, L_k) \rightarrow \min$

$\forall L_k = \{a_0 + \beta_1 a_1 + \dots + \beta_k a_k, \beta_i \in R\}$, где параметры β_i пробегают вещественную прямую $R, a_0 \in R^n, \{a_1, a_2, \dots, a_k\} \subset R^n$ – ортонормированный набор векторов.

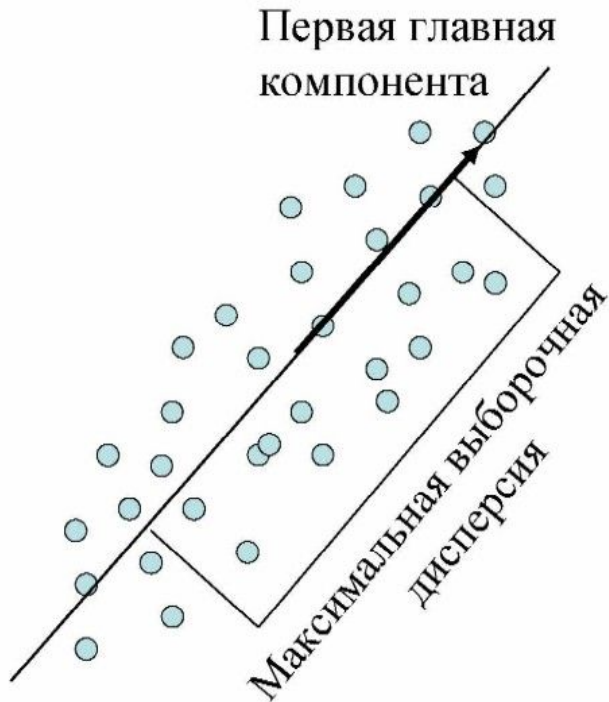
Аппроксимация данных линейными многообразиями

$$\sum_{i=1}^m \text{dist}^2(x_i, L_k) = \left\| x_i - a_0 - \sum_{j=1}^k a_j (a_j x_j - a_0) \right\|^2$$

Решение задачи аппроксимации для $k = 1, 2, \dots, n - 1$ дается набором вложенных линейных многообразий $L_0 \subset L_1 \subset \dots \subset L_{n-1}$, $L_k = \{a_0 + \beta_1 a_1 + \dots + \beta_k a_k, \beta_i \in R\}$. Эти линейные многообразия определяются ортонормированным набором векторов $\{a_1, \dots, a_{n-1}\}$ – векторами главных компонент и вектором a_0 . Вектор a_0 ищется как решение задачи минимизации для L_0 :

$$a_0 = \arg \min \left(\sum_{i=1}^m \text{dist}^2(x_i, L_k) \right)$$

Аппроксимация данных линейными многообразиями



Ищем такой вектор, при котором максимизировалась бы дисперсия проекции нашей выборки на него.

Диагонализация ковариационной матрицы

$$C = [c_{ij}], c_{ij} = \frac{1}{m-1} \sum_{l=1}^m (x_{li} - \bar{X}_i)(x_{lj} - \bar{X}_j)$$

Векторы главных компонент для задач о наилучшей аппроксимации и о поиске ортогональных проекций с наибольшим рассеянием – это ортонормированный набор $\{a_1, \dots, a_n\}$ собственных векторов эмпирической ковариационной матрицы C , расположенных в порядке убывания собственных значений λ : $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$. Данные векторы служат оценкой для собственных векторов ковариационной матрицы $\text{cov}(X, X)$. В базисе из собственных векторов ковариационной матрицы она, естественно, диагональна, и в этом базисе коэффициент ковариации между различными координатами равен нулю.

Диагонализация ковариационной матрицы

Если спектр ковариационной матрицы вырожден, то выбирают произвольный ортонормированный базис собственных векторов. Он существует всегда, а собственные числа ковариационной матрицы всегда вещественны и неотрицательны

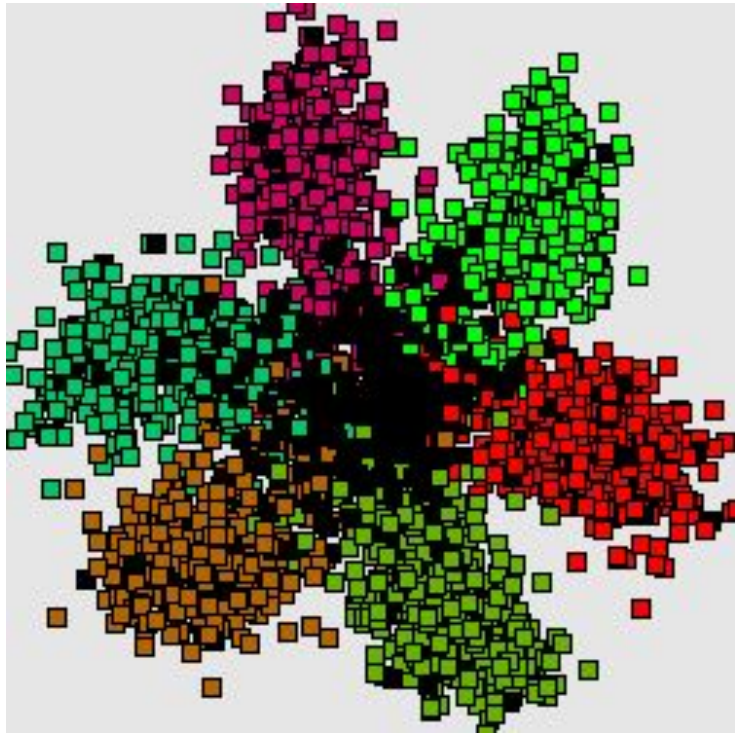
Подавление шума на изображениях



Примеры

- Биоинформатика
- Хемометрика
- Индексация видео
- Общественные науки

Применение для визуализации



Проекция ДНК-блуждания на первые 2 главные компоненты для генома бактерии «*Streptomyces coelicolor*».

Спасибо за внимание!

A dark blue, solid-colored shape that starts as a thin line at the bottom left and expands diagonally upwards and to the right, filling the bottom right portion of the slide.