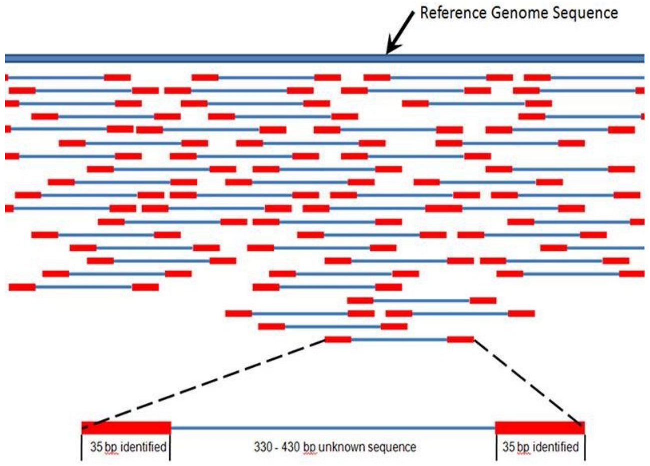
Понижение размерности

Биоинформатика

- Задачи анализа генома человека
- Признаки: характеристики генов (более 20.000)
- Маленькие выборки (расшифровка генома — сложная и дорогостоящая процедура)
- Признаков существенно больше, чем объектов!



Категориальные признаки

- Пример: предсказать, понравится ли пользователю фильм
- Объект: пара «пользователь-фильм»
- Признаки: ID пользователя, ID фильма, ID жанра, ID режиссёра, ID главных актёров, ID композитора, …
- Как много фильмов снято за всю историю?
- IMDB: >330 тысяч
- После бинарного кодирования получим миллионы признаков

Анализ текстов

- Пример: предсказание популярности фильма по тексту его сценария
- Признаки: количество вхождений каждого слова из словаря
- Сколько слов в словаре?
- Сотни тысяч признаков
- Если учитывать n-граммы, то десятки миллионов признаков

Анализ данных ЭЭГ

- Энцефалограф: 64 датчика, частота сигнала 256 Гц
- Объект: результаты измерений для одного пациента
- За 5 секунд измерений: $64 * 256 * 5 = 81\,920$ признаков

UCI Machine Learning Repository

About Citation Policy Donate a Data Set Contact

Search

Repository Web

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

View ALL Data Sets

Browse Through: **557** Data Sets

Table View List View

Default Task
- Classification (417)
- Regression (129)
- Clustering (112)
- Other (56)

Attribute Type
- Categorical (38)
- Numerical (375)
- Mixed (55)

Data Type
- Multivariate (434)
- Univariate (27)
- Sequential (55)
- Time-Series (112)
- Text (63)
- Domain-Theory (23)
- Other (21)

Area
- Life Sciences (131)
- Physical Sciences (55)
- CS / Engineering (205)
- Social Sciences (31)
- Business (40)
- Game (10)
- Other (80)

 # Attributes
- Less than 10 (141)
- 10 to 100 (252)
- Greater than 100 (99)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
URL Reputation	Multivariate, Time-Series	Classification	Integer, Real	2396130	3231961	2009
KASANDR	Multivariate	Causal-Discovery	Integer	17764280	2158859	2017
Gas sensor arrays in open sampling settings	Multivariate, Time-Series	Classification	Real	18000	1950000	2013
YouTube Multiview Video Games Dataset	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013
Twin gas sensor arrays	Multivariate, Time-Series, Domain-Theory	Classification, Regression	Real	640	480000	2016
Deepfakes: Medical Image Tamper Detection	Multivariate	Classification	Real	20000	200000	2020
Gas sensor array exposed to turbulent gas mixtures	Multivariate, Time-Series	Classification, Regression	Real	180	150000	2014
ElectricityLoadDiagrams20112014	Time-Series	Regression, Clustering	Real	370	140256	2015

https://archive.ics.uci.edu/ml/index.html

https://archive.ics.uci.edu/ml/datasets.php

Задача понижения размерности

- Дано: матрица «объекты-признаки» $\boldsymbol{X}$ размера $\ell \times \boldsymbol{D}$
- Найти: новую матрицу «объекты-признаки» $\boldsymbol{Z}$ размера $\ell \times \boldsymbol{d}$
- $d < D$

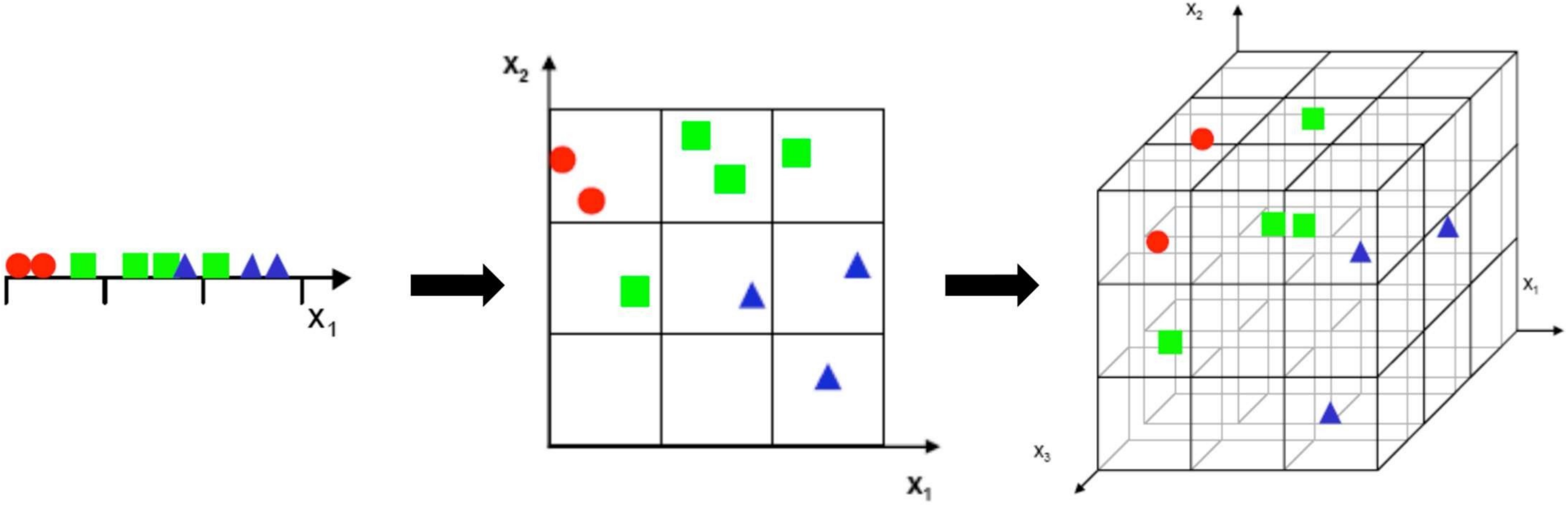
Но зачем?

- Проклятие размерности
- Шумовые признаки
- Переобучение
- Интерпретируемость модели
- Скорость работы модели
- Визуализация данных

Проклятие размерности

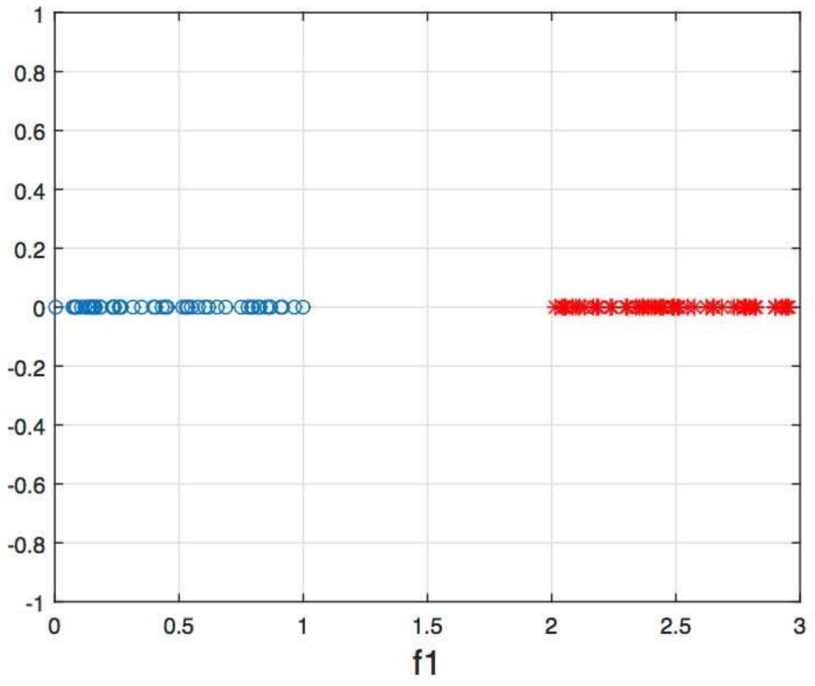
- Задача: классификация пончиков на вкусные и невкусные
- 100 объектов
- Цвет: 10 вариантов
- Цвет + размер: $10 * 4 = 40$ вариантов
- Цвет + размер + форма: $10 * 4 * 4 = 160$ вариантов

Проклятие размерности



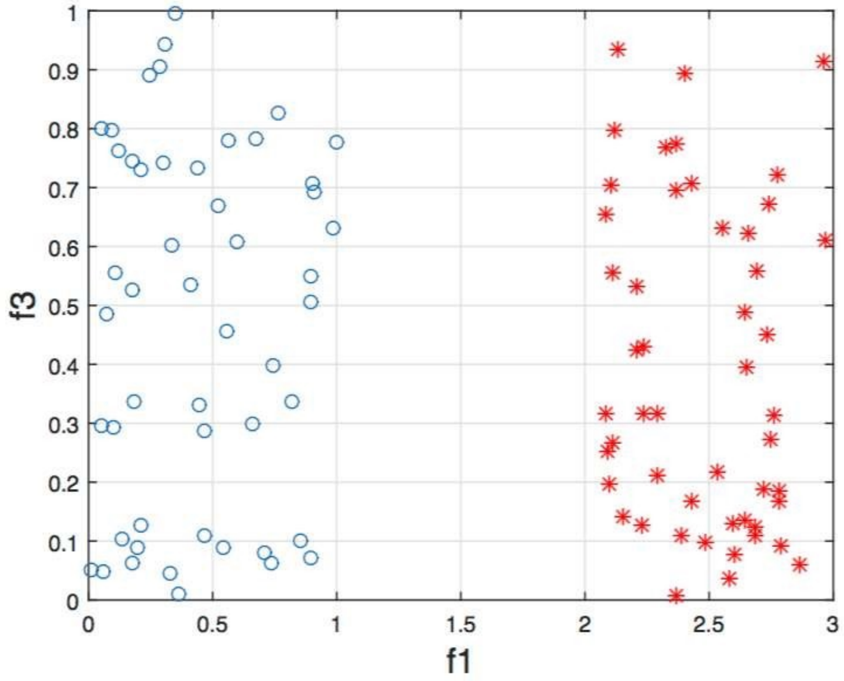
Плохие признаки

Информативный признак



Плохие признаки

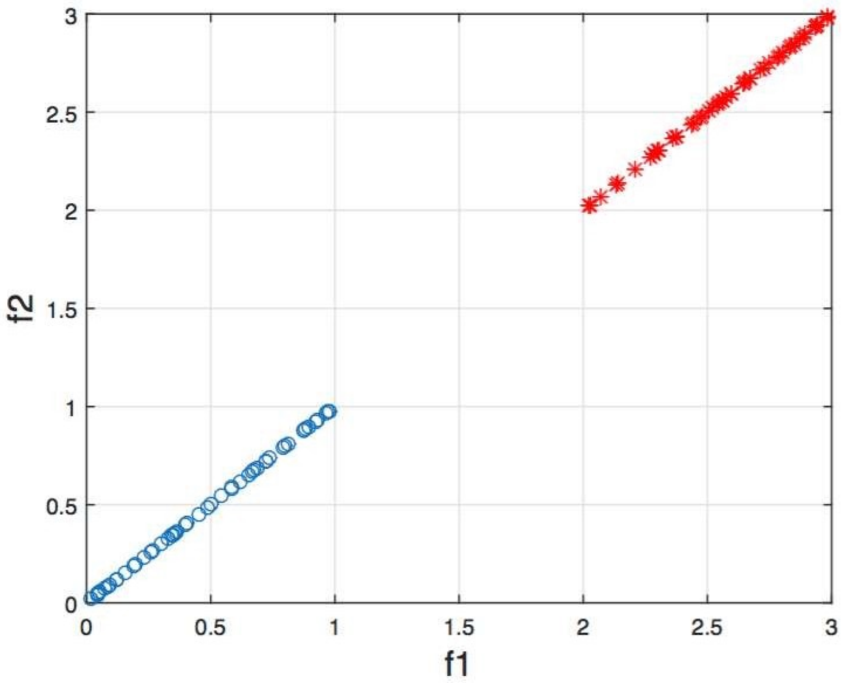
f3 — шумовой признак



Плохие признаки

Коррелирующие признаки

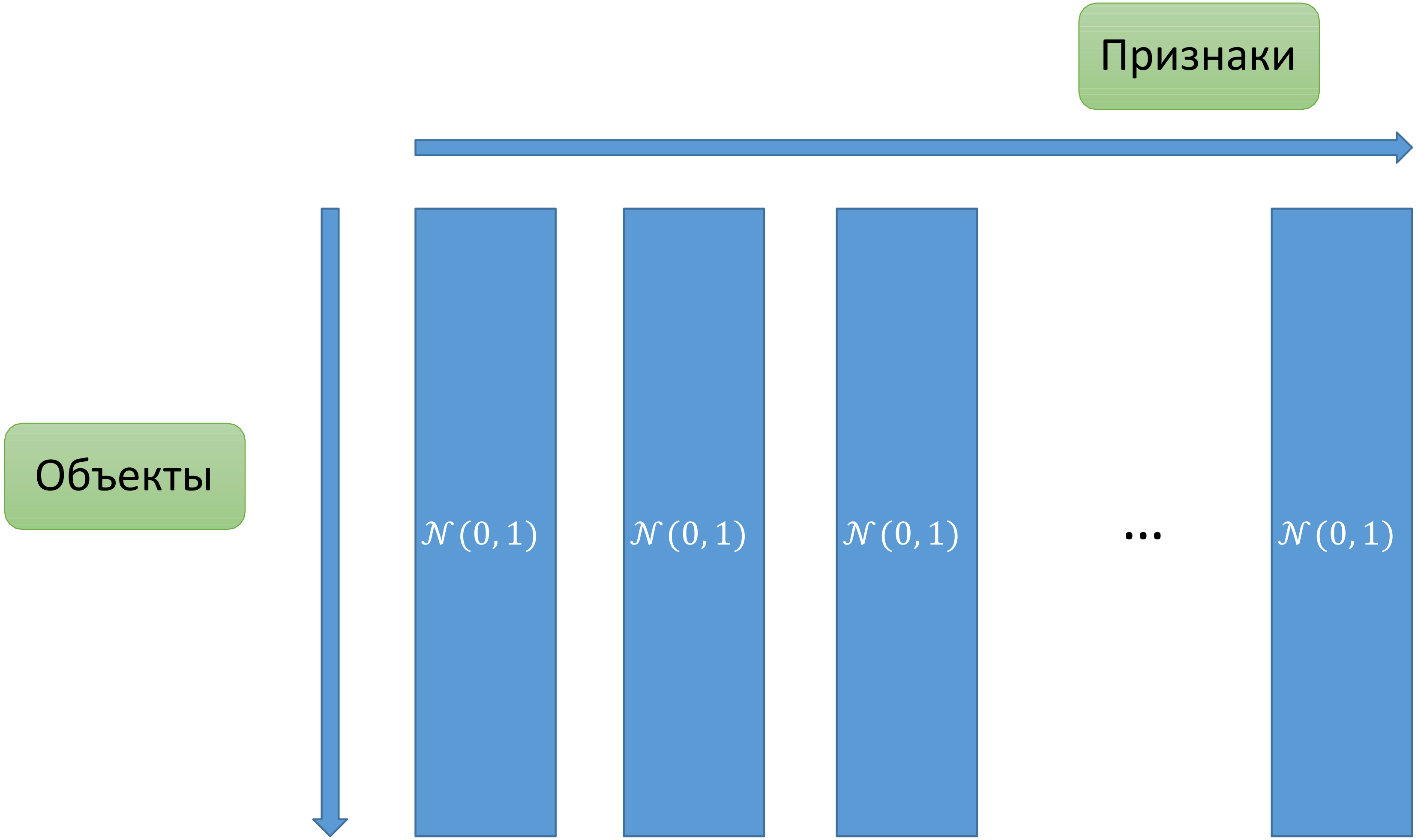
f2 — избыточный признак



Шумовые признаки

- Признаки, которые никак не связаны с целевой переменной
- Но по обучающей выборке это не всегда можно понять

Шумовые признаки



Шумовые признаки

- Генерируем случайные признаки
- Если их много, то некоторые будут хорошо коррелировать с ответами

y	x_1	x_2	x_3	x_4
-1	1.11	**-0.5**	0.42	0.33
-1	1.22	**-0.46**	-1.98	-0.55
1	-1.56	**0.04**	0.39	-1.67
1	-0.48	**1.32**	0.88	-0.27

Ускорение моделей

- Чем больше признаков, тем дольше обучаются модели
- Чем дольше обучаются модели, тем меньше экспериментов удаётся провести
- Чем сложнее модели, тем дольше они вычисляют прогнозы
- Могут быть жёсткие ограничения на скорость
- Пример: рекомендательные системы

Методы понижения размерности

- Отбор признаков (feature selection)
  - Выбрать d самых важных признаков
- **Извлечение признаков (feature extraction)**
  - Найти d новых признаков, выражающихся через исходные

МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS, PCA)

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Цель: *хотим придумать новые признаки, каким-то образом выражающиеся через старые, причем новых признаков хочется получить меньше, чем старых.*

Сегодня будем рассматривать только случай, когда новые признаки **линейно** выражаются через старые.

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Постановка задачи:

- $x_1, \dots, x_n$ - исходные числовые признаки
- $z_1, \dots, z_d$ – новые числовые признаки, $d \leq n$

Хотим:

1. чтобы новые числовые признаки z_j линейно выражались через исходные признаки x_i

2. чтобы при переходе к новым признакам было потеряно наименьшее количество исходной информации

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

1. чтобы новые числовые признаки z_j линейно выражались через исходные признаки x_i

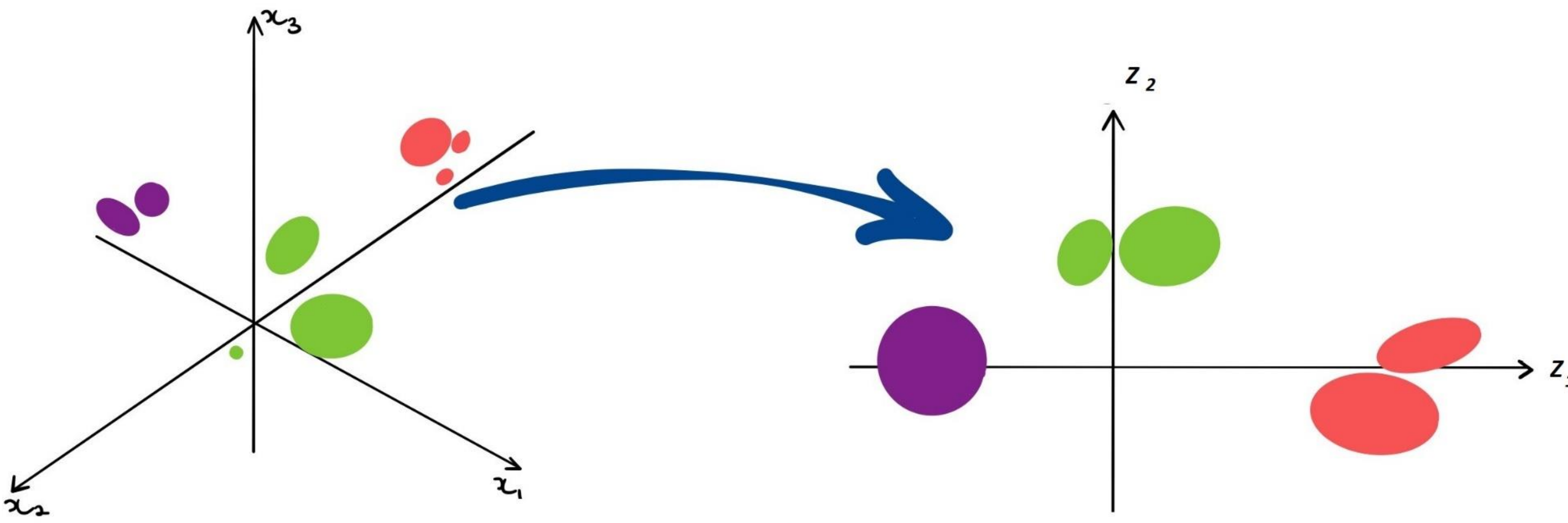
$$\begin{cases} z_1 = u_{11}x_1 + \cdots + u_{1n}x_n \\ z_2 = u_{21}x_1 + \cdots + u_{2n}x_n \\ \qquad\qquad \dots \\ z_d = u_{d1}x_1 + \cdots + u_{dn}x_n \end{cases}$$

Геометрическая интерпретация: новые признаки z_i – это проекции исходных признаков x_i на некоторые векторы (компоненты) u.

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

1. чтобы новые числовые признаки z_j линейно выражались через исходные признаки x_i

Геометрически это означает, что мы проецируем пространство признаков размерности n на некоторое линейное подпространство размерности d:



ПОЯСНЕНИЕ: ПРОЕКЦИЯ

- Проекция вектора x на вектор (компоненту) u_i:

$$(x, u_i)$$

- Проекция выборки X на компоненту u_i:

$$Xu_i$$

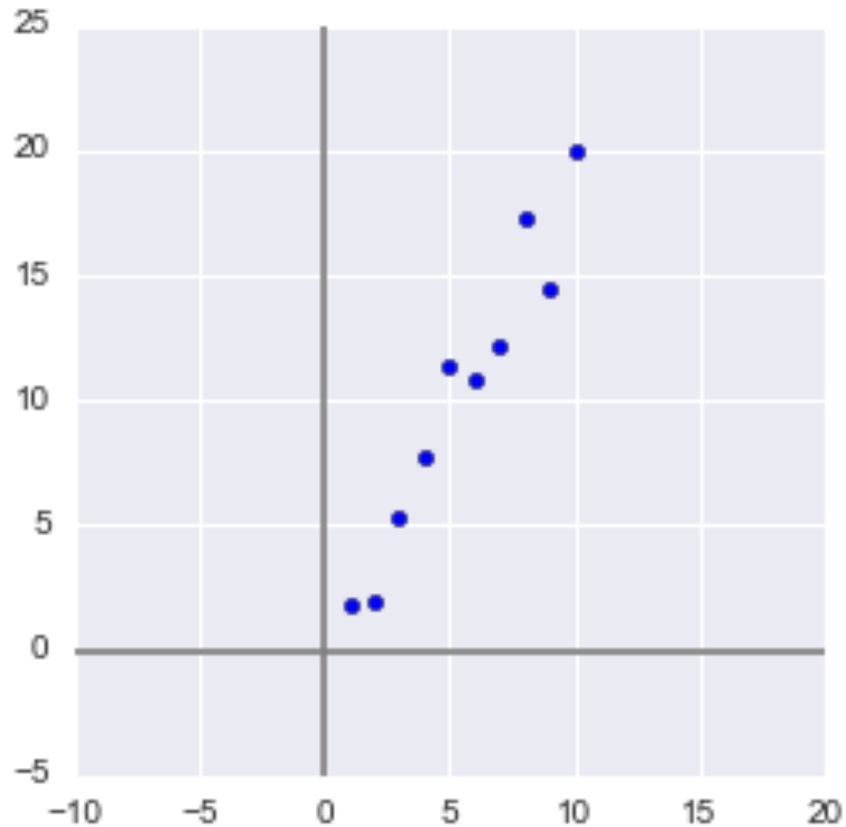
МЕТОД ГЛАВНЫХ КОМПОНЕНТ

2. чтобы при переходе к новым признакам было потеряно наименьшее количество исходной информации.

Дисперсия выборки, посчитанная в новых признаках, показывает, как много информации нам удалось сохранить после понижения размерности, поэтому дисперсия в новых признаках должна быть максимальной.

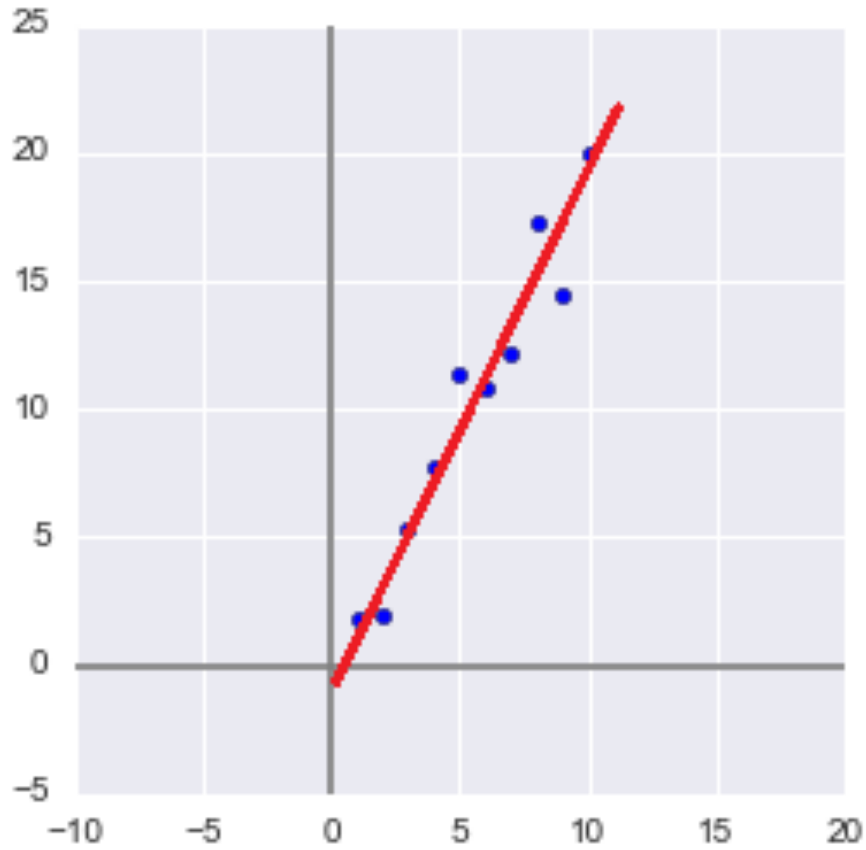
ПРИМЕР

Хотим спроецировать двумерные данные X на одномерный вектор u так, чтобы дисперсия проекции Xu была максимальной:



ПРИМЕР

Хотим спроецировать двумерные данные X на одномерный вектор u так, чтобы дисперсия проекции Xu была максимальной:



ПОСТАНОВКА ЗАДАЧИ

Будем искать такие компоненты $u_1, u_2, \dots, u_d$, что:

1) Они ортогональны, т.е. $(u_i, u_j) = 0$

2) Они нормированы, т.е. $||u_i|| = 1$

3) дисперсия проекции выборки на них максимальна:

$$D(Xu_i) \to \max_{u_i} \quad , \qquad i = 1, \dots, d$$

ВАЖНОЕ ДЕЙСТВИЕ

Центрируем исходные данные, то есть вычтем из каждого признака его среднее значение.

ДИСПЕРСИЯ ПРОЕКЦИИ

- Мы уже выяснили, что проекция выборки X на компоненту u_i:

$$Xu_i$$

- Тогда проекция выборки на первые d компонент, задаваемых столбцами матрицы U_d:

$$XU_d$$

ДИСПЕРСИЯ ПРОЕКЦИИ

- Мы уже выяснили, что проекция выборки X на компоненту u_i:

$$Xu_i$$

- Тогда проекция выборки на первые d компонент, задаваемых столбцами матрицы U_d:

$$XU_d$$

- Тогда дисперсия проекции – это след ковариационной матрицы:

$$tr\left((XU_d)^T(XU_d)\right) = \sum_{i=1}^{d} \left||Xu_i|\right|^2 \to \max_u$$

ПЕРВЫЙ ШАГ

- Будем искать первую компоненту, u_1:

$$\begin{cases} \|Xu_1\|^2 \to \max\limits_{u_1} \\ \|u_1\|^2 = 1 \end{cases}$$

ПЕРВЫЙ ШАГ

- Будем искать первую компоненту, u_1:

$$\begin{cases} ||Xu_1||^2 \to \max\limits_{u_1} \\ ||u_1||^2 = 1 \end{cases}$$

Решение:

- Запишем лагранжиан

$$L(u_1, \lambda) = ||Xu_1||^2 + \lambda(||u_1||^2 - 1)$$

ПЕРВЫЙ ШАГ

- Будем искать первую компоненту, u_1:

$$\begin{cases} ||Xu_1||^2 \to \max\limits_{u_1} \\ ||u_1||^2 = 1 \end{cases}$$

Решение:

- Запишем лагранжиан

$$L(u_1, \lambda) = ||Xu_1||^2 + \lambda(||u_1||^2 - 1)$$

- $\frac{\partial L}{\partial u_1} = ?$

ПЕРВЫЙ ШАГ

- Будем искать первую компоненту, u_1:

$$\begin{cases} \left|\left|Xu_1\right|\right|^2 \to \max\limits_{u_1} \\ \left|\left|u_1\right|\right|^2 = 1 \end{cases}$$

Решение:

- Запишем лагранжиан

$$L(u_1, \lambda) = \left|\left|Xu_1\right|\right|^2 + \lambda(\left|\left|u_1\right|\right|^2 - 1)$$

- $\frac{\partial L}{\partial u_1} = 2X^T X u_1 + 2\lambda u_1 = 0 \Rightarrow X^T X u_1 = -\lambda u_1$ - собств.в-р.

ПЕРВЫЙ ШАГ

- Будем искать первую компоненту, u_1:

$$\begin{cases} \left|\left|Xu_1\right|\right|^2 \to \max\limits_{u_1} \\ \left|\left|u_1\right|\right|^2 = 1 \end{cases}$$

Решение:

- Запишем лагранжиан

$$L(u_1, \lambda) = \left|\left|Xu_1\right|\right|^2 + \lambda(\left|\left|u_1\right|\right|^2 - 1)$$

- $\frac{\partial L}{\partial u_1} = 2X^T X u_1 + 2\lambda u_1 = 0 \Rightarrow X^T X u_1 = -\lambda u_1$ - собств.в-р.
- $\left|\left|Xu_1\right|\right|^2 = u_1^T X^T X u_1 = \lambda u_1^T u_1 = \lambda \to \max_{u_1}$ - max собств. значение.

ПЕРВЫЙ ШАГ

- Будем искать первую компоненту, u_1:

$$\begin{cases} ||Xu_1||^2 \to \max\limits_{u_1} \\ ||u_1||^2 = 1 \end{cases}$$

Ответ:

u_1 - собственный вектор матрицы ковариаций $X^T X$ с максимальным собственным значением.

ПРОЕКЦИИ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

- Пусть X – матрица объект-признак для исходных признаков.
- Метод главных компонент делает проекцию исходных объектов на гиперплоскость некоторой размерности d.

Теорема. Базисные векторы этой гиперплоскости – это собственные векторы матрицы $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ (матрица ковариаций), соответствующие d её наибольшим собственным значениям.

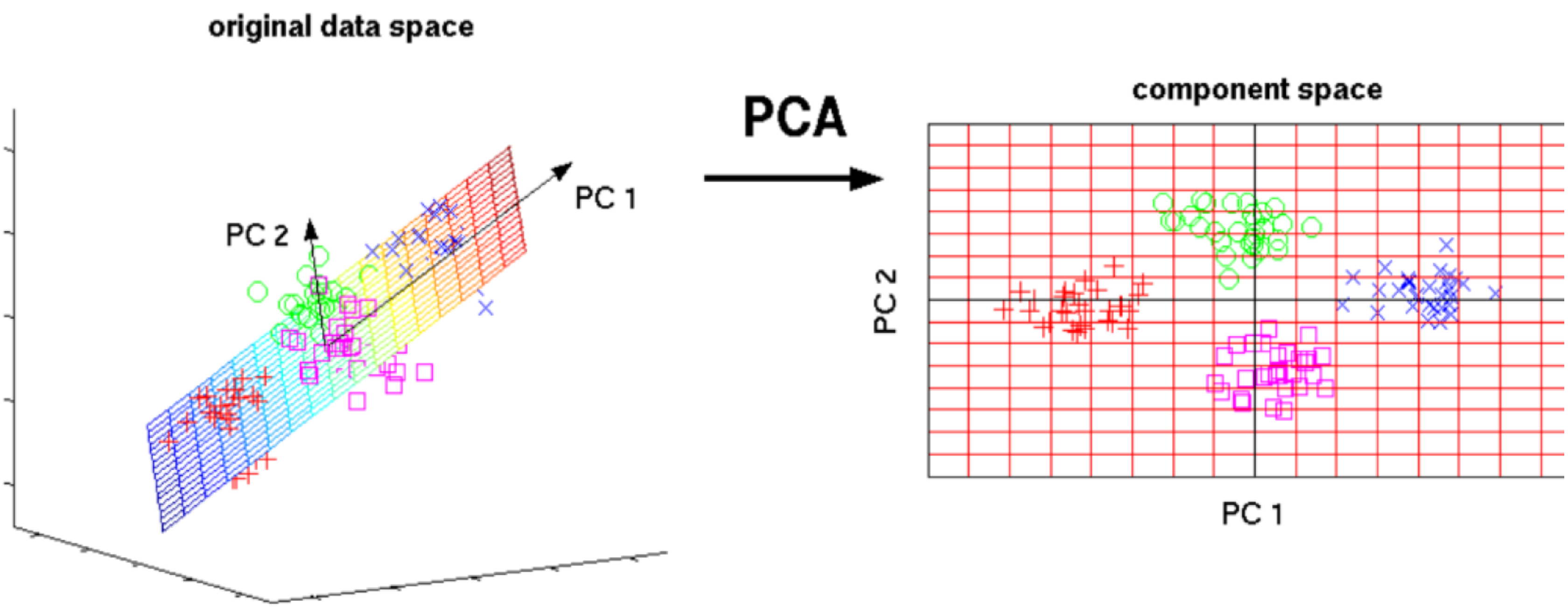
КОНСТРУКТИВНОЕ ПОСТРОЕНИЕ БАЗИСА В PCA

- Находим вектор $u_1 = argmax_u\big(D(Xu)\big)$ и нормируем его: $u_1 \to \frac{u_1}{||u_1||}$
- Находим вектор $u_2 = argmax_u\big(D(Xu)\big)$ такой, что $(u_1, u_2) = 0$ и нормируем его: $u_2 \to \frac{u_2}{||u_2||}$
- Находим вектор $u_3 = argmax_u\big(D(Xu)\big)$ такой, что $(u_1, u_3) = (u_2, u_3) = 0$ и нормируем его: $u_3 \to \frac{u_3}{||u_3||}$.

И т.д.

Получаем ортонормированный базис $\{u_1, u_2, \dots, u_d\}$.

ПРОЕКЦИЯ НА ГИПЕРПЛОСКОСТЬ



ПРИМЕНЕНИЕ МЕТОДА

- Когда главные компоненты найдены, можно проецировать на них и новые данные:
$$Z' = X' U_d.$$

ДОЛЯ ОБЪЯСНЕННОЙ ДИСПЕРСИИ

- Упорядочим собственные значения матрицы $X^T X$ по убыванию: $\lambda_1 \geq \lambda_2 \geq \cdots > \lambda_n \geq 0$.

- Доля дисперсии, объяснённой j-й компонентой (explained variance ratio):

$$\boldsymbol{\delta_j = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_n}}$$

- Доля дисперсии, объясняемой первыми k компонентами:

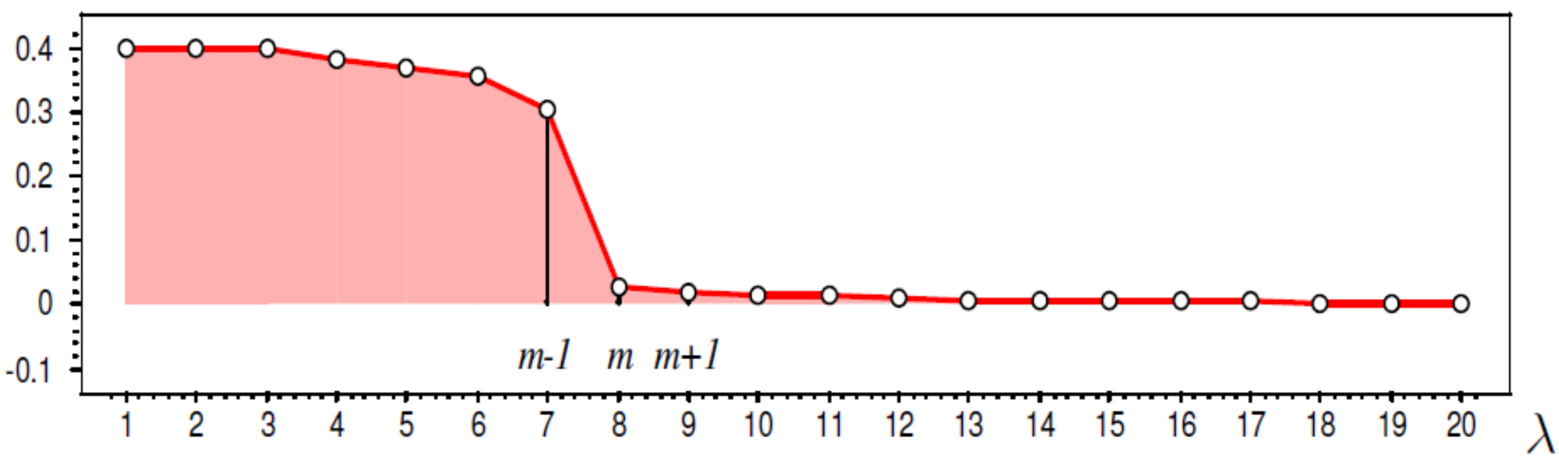
$$\boldsymbol{\delta = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_n}}$$

ВЫБОР ЧИСЛА ГЛАВНЫХ КОМПОНЕНТ

- Эффективная размерность выборки **–** это наименьшее целое m, при котором *доля необъясненной дисперсии*

$$E_m = \frac{||ZU^T - X||^2}{\left||X|\right|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\sum_{i=1}^{n} \lambda_i} \le \varepsilon$$

Критерий крутого склона:

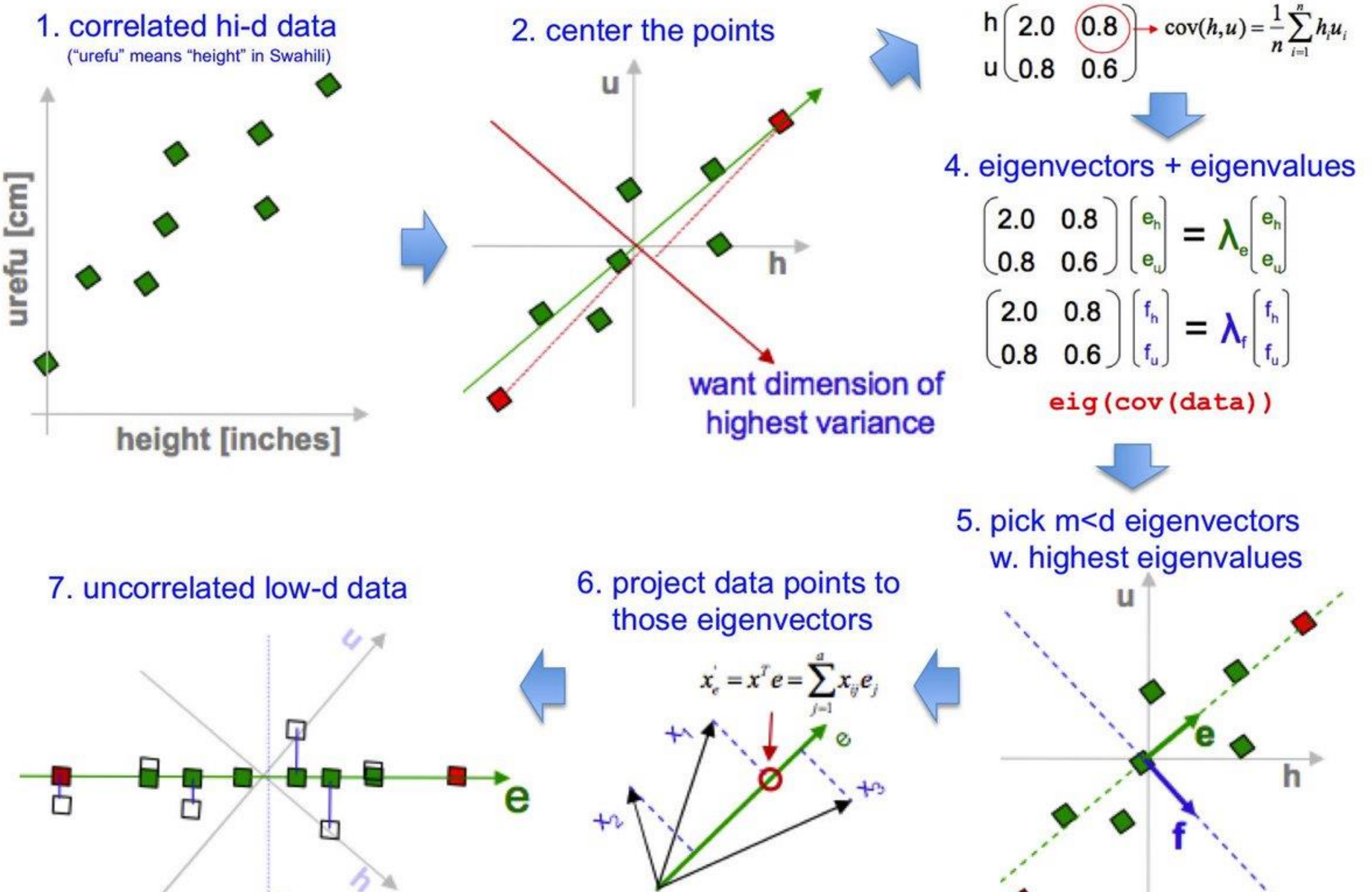


ПРИМЕР: FACES DATASET

FACES DATASET (ГЛАВНЫЕ КОМПОНЕНТЫ)

ВОССТАНОВЛЕННОЕ ИЗОБРАЖЕНИЕ

PCA in a nutshell



СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ МАТРИЦЫ (SINGULAR VALUE DECOMPOSITION, SVD)

Теорема. Матрицу $A \in \mathbb{R}^{m \times n}$ можно представить в виде

$$\boldsymbol{A = U\Sigma V^T},$$

- где $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ - ортогональные матрицы,
- $\Sigma \in \mathbb{R}^{m \times n}$ - диагональная матрица с ненулевыми элементами $\sigma_i = \sqrt{\lambda_i}$, где λ_i - собственные значения матрицы $A^T A$.

СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ МАТРИЦЫ (SVD)

Теорема. Матрицу $A \in \mathbb{R}^{m \times n}$ можно представить в виде

$$\boldsymbol{A = U\Sigma V^T},$$

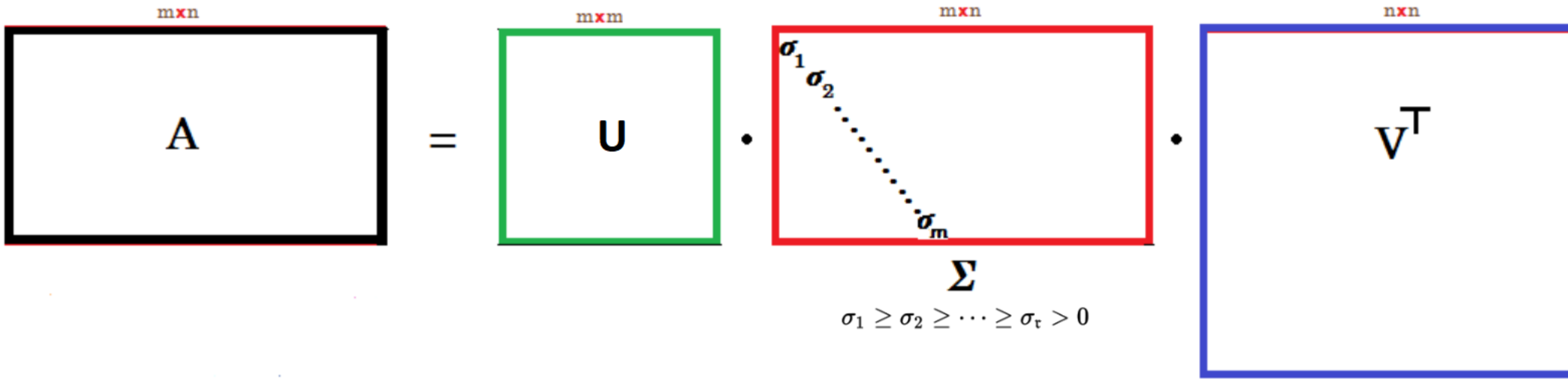
- где $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ - ортогональные матрицы,
- $\Sigma \in \mathbb{R}^{m \times n}$ - диагональная матрица с ненулевыми элементами $\sigma_i = \sqrt{\lambda_i}$, где λ_i - собственные значения матрицы $A^T A$.

При этом

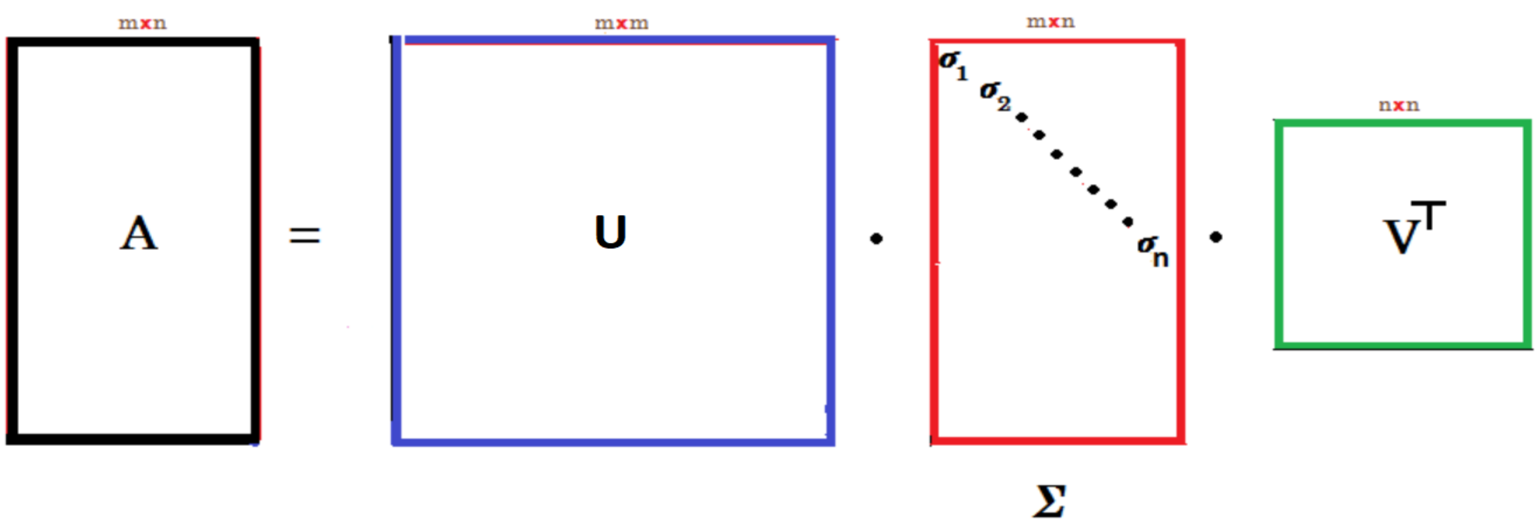
- Столбцы матрицы U являются собственными векторами матрицы AA^T
- Столбцы матрицы V являются собственными векторами матрицы $A^T A$.

SINGULAR VALUE DECOMPOSITION

- При $m \leq n$:



- При $m > n$:



СВЯЗЬ SVD И PCA

Пусть X – матрица объект-признак, для которой мы хотим снизить размерность и $X = U\Sigma V^T$ её SVD-разложение.

Тогда:

- Столбцы матрицы $\boldsymbol{V}$ – это собственные векторы матрицы $X^T X$, т.е. векторы $v_1, \dots, v_n$ – главные компоненты.
- Столбцы матрицы $\boldsymbol{U\Sigma}$ – это новые признаки, то есть, проекции исходных признаков на главные компоненты $Z = Xv$

($X = U\Sigma\mathrm{V}^{\mathrm{T}} \Leftrightarrow U\Sigma = XV$).

- Сингулярные числа матрицы Σ – это корни из собственных чисел матрицы $X^T X$.

СВЯЗЬ SVD И PCA

- Столбцы матрицы V – это собственные векторы матрицы $X^T X$, т.е. векторы $v_1, \dots, v_n$ — главные компоненты.
- Столбцы матрицы $U\Sigma$ – это новые признаки $z = Xv$

($X = U\Sigma \mathrm{V}^{\mathrm{T}} \Leftrightarrow U\Sigma = XV$).

- Сингулярные числа матрицы Σ – это корни из собственных чисел матрицы $X^T X$.

Для снижения размерности берем первые k столбцов матрицы U и верхний $k \times k$-квадрат матрицы Σ, тогда матрица $U_k \Sigma_k$ содержит k новых признаков, соответствующих первым k главным компонентам.

ЧТО ЛУЧШЕ: PCA ИЛИ SVD?

- Существуют вычислительные трудности с нахождением собственных значений, в этом недостаток PCA.
- Существует итерационный алгоритм для нахождения SVD (без нахождения собственных значений) http://www.machinelearning.ru/wiki/index.php?title=Простой_итерационный_алгоритм_сингулярного_разложения.

Поэтому вычислительно эффективнее использовать SVD при прочих равных.