

Математические модели аномалий
и методы их обнаружения
в данных высокой размерности

А. В. Артёмов

ФТиАД ФКН ВШЭ,
Вероятностные модели и прикладная статистика
в финансовой математике, весна 2018

Содержание

- 1 Аномалии, их характеристизация и свойства
- 2 Обзор математических моделей аномалий
- 3 Обнаружение аномалий без учителя в данных высокой размерности
 - Задачи снижения размерности и линейная модель РСА
 - Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
 - Стохастические модели «нормальности». Модель смеси гауссиан
- 4 Резюме лекции

Содержание

- 1 Аномалии, их характеристизация и свойства
- 2 Обзор математических моделей аномалий
- 3 Обнаружение аномалий без учителя в данных высокой размерности
 - Задачи снижения размерности и линейная модель PCA
 - Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
 - Стохастические модели «нормальности». Модель смеси гауссиан
- 4 Резюме лекции

Что такое аномальные данные?

- ▶ Нет строгого определения термину «**аномалии**»
- ▶ Аномалии: тестовые данные, отличные от обучающей выборки
- ▶ Используются и другие термины: *выбросы, новинки, шумы, отклонения, исключения*
- ▶ “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”
[D. M. Hawkins, 1980]

Причины появления аномальных данных

- ▶ Данные, полученные другим механизмом или от иного класса объектов, чем основная выборка
 - ▶ мошеннические vs. органические транзакции
 - ▶ заболевшие vs. здоровые испытуемые
- ▶ Естественные вариации генерирующего распределения
 - ▶ хвосты нормального распределения
- ▶ Ошибки и сбои измерений, сбора данных и т.п.
- ▶ В задачах классификации:
 - ▶ недостаточный объем выборки
 - ▶ недостаток значимых признаков
 - ▶ подавляющий объем примеров фонового класса

Где встречаются аномальные данные?

- ▶ Сложные (многокомпонентные, нелинейные) системы с большими объемами данных
 - ▶ Медицинская диагностика [Clifton и др., 2011; Quinn и др., 2007]
 - ▶ Сложные промышленные системы [Tarassenko и др., 2009]
 - ▶ Внедрения в системы безопасности [Jyothsna и др., 2011; Patcha и др., 2007]
 - ▶ Аномальные полеты в авиации [Matthews и др., 2013]
- ▶ Получить представление о взаимосвязях между компонентами системы крайне сложно
- ▶ Существует большое количество аномальных режимов (возможно, неизвестных заранее) \implies классификация неэффективна

Выявление предпосылок к летным происшествиям

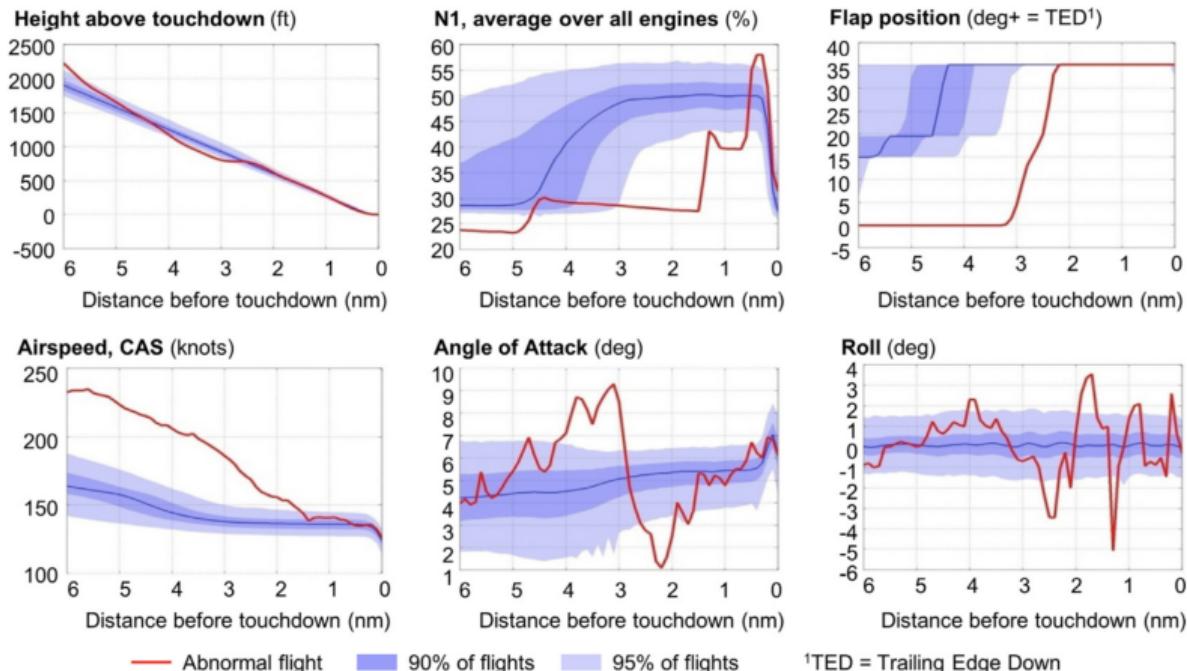


Рис. 1: Различные характеристики полетов, выявленные по большой выборке данных, и характеристики **аномального полета**, показанные красным цветом.

Ключевые трудности в обнаружении аномалий

- ▶ Трудно определить репрезентативную «нормальную» область данных
- ▶ Граница, отделяющая «нормальное» и «аномальное» поведение, часто неточна
- ▶ Нет единого определения понятию «выброса» или «аномалии»
- ▶ Доступность размеченных данных для обучения и валидации подходов ограничена
- ▶ Шум данных может быть неотличим от аномального поведения
- ▶ В некоторых приложениях (мошенничество) аномалии адаптивно изменяются во времени
- ▶ Нормальное поведение изменяется во времени

Содержание

1 Аномалии, их характеристизация и свойства

2 Обзор математических моделей аномалий

3 Обнаружение аномалий без учителя в данных высокой размерности

- Задачи снижения размерности и линейная модель PCA
- Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
- Стохастические модели «нормальности». Модель смеси гауссиан

4 Резюме лекции

Алгоритм действий при решении задачи обнаружения аномалий

На практике для построения алгоритма детектирования аномальных данных требуется:

1. Построить модель нормального состояния системы.
2. Определить, как измерить отклонение от нормального состояния.
3. Построить вычислительный алгоритм, маркирующий объекты, которые отклоняются от нормальных условий.

Простейший пример: правило «трех сигм» в теории вероятностей

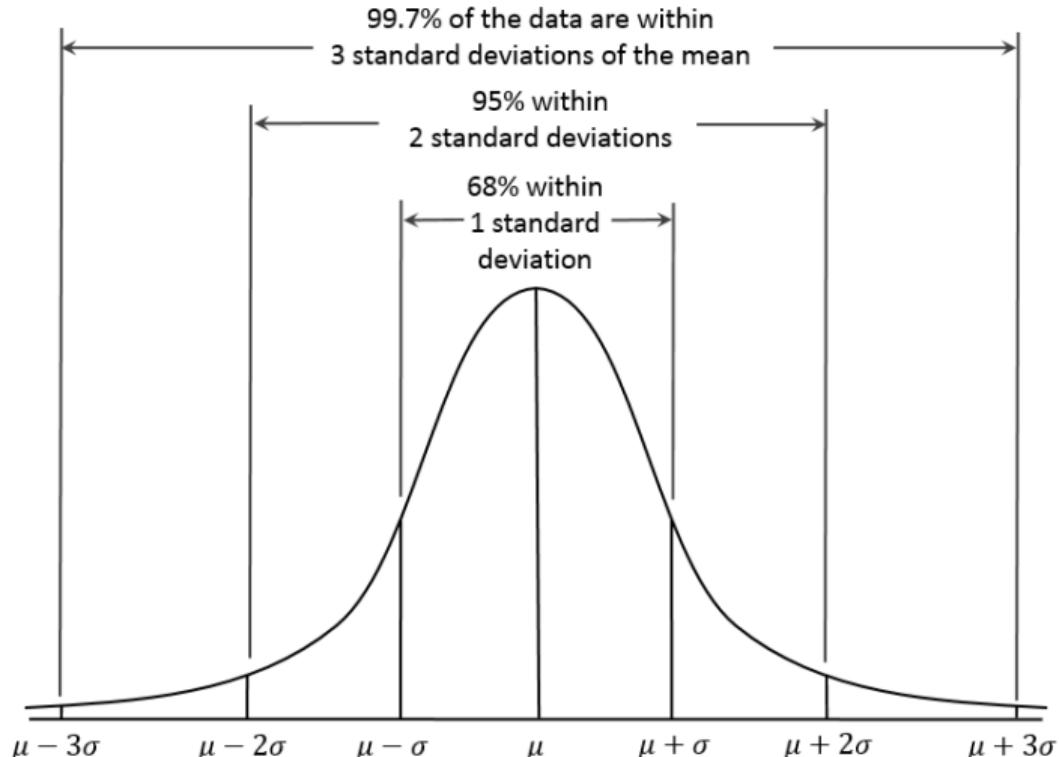


Рис. 2: Источник: Wikipedia

Правило трех σ

Имеем выборку x_1, \dots, x_n

1. Построить модель нормального состояния системы.
 - ▶ Нормальное состояние определяется средним значением $\mu = \frac{1}{n} \sum_{i=1}^n x_i$
2. Определить, как измерить отклонение от нормального состояния.
 - ▶ Отклонение измеряется с помощью расстояния до среднего значения, согласно среднеквадратичному отклонению и дисперсии $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$
3. Построить вычислительный алгоритм, маркирующий объекты, которые отклоняются от нормальных условий.
 - ▶ Все элементы, находящиеся более чем на $3 \cdot \sigma$ помечаются как аномалии $\frac{x-\mu}{\sigma} > 3$

Приложение: Хадлум против Хадлума (1949) [Barnett, 1978]

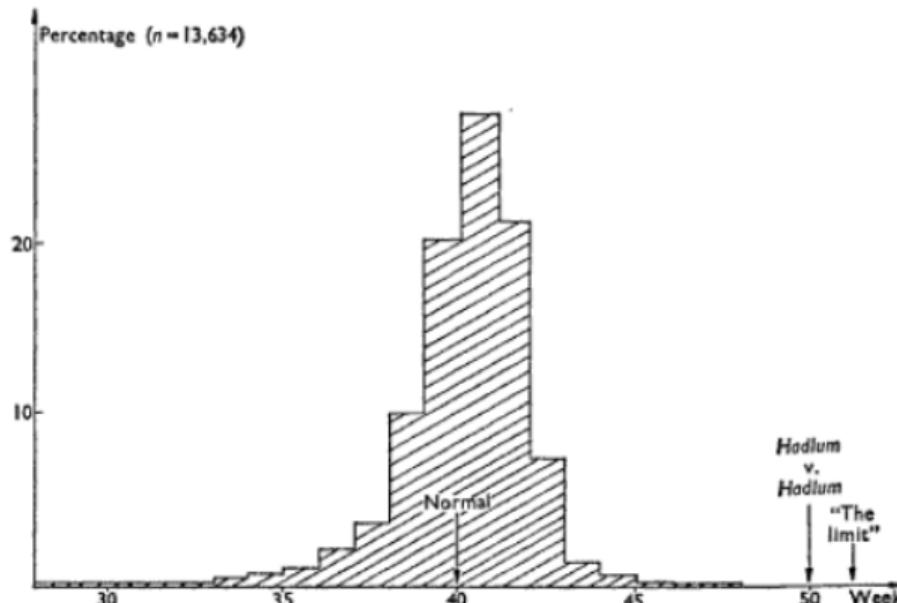
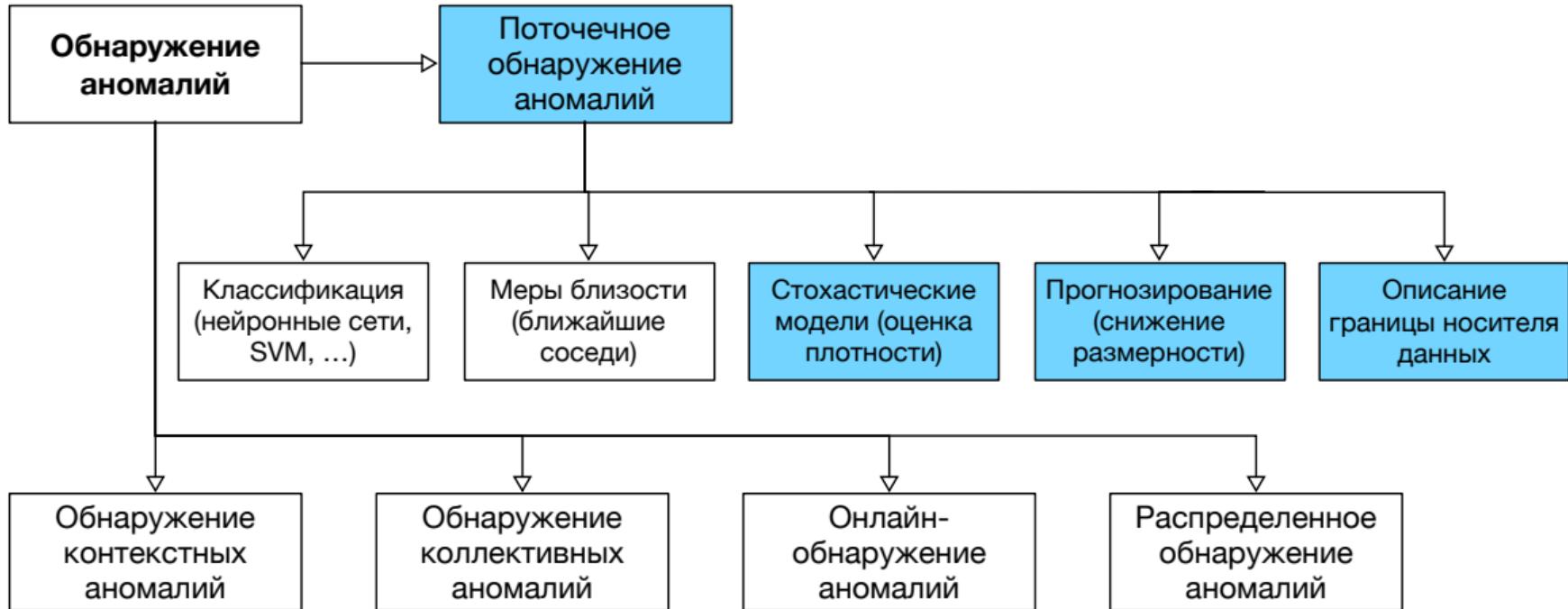


FIG. 1. Distribution of human gestation periods.

Рис. 3: Мистер Хадлум обжаловал отказ в удовлетворении своей петиции о разводе: ребенок миссис Хадлум родился 349 дней спустя того, как мистер Хадлум уехал на войну в Европу. Среднее время беременности у женщин составляет 280 дней.

Таксономия подходов к обнаружению аномалий



Обнаружение аномалий на основе вероятностных моделей

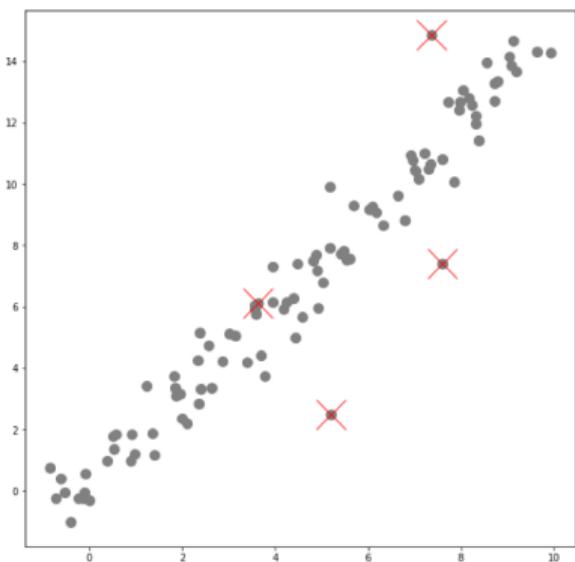


Рис. 4: Исходные
данные в \mathbb{R}^2

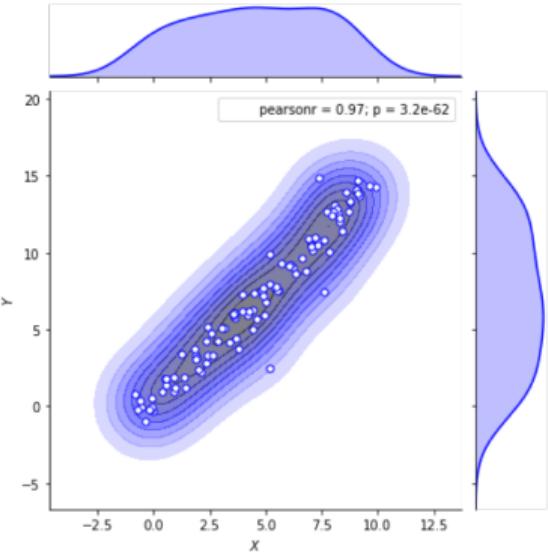


Рис. 5: Оценка
двумерной
плотности

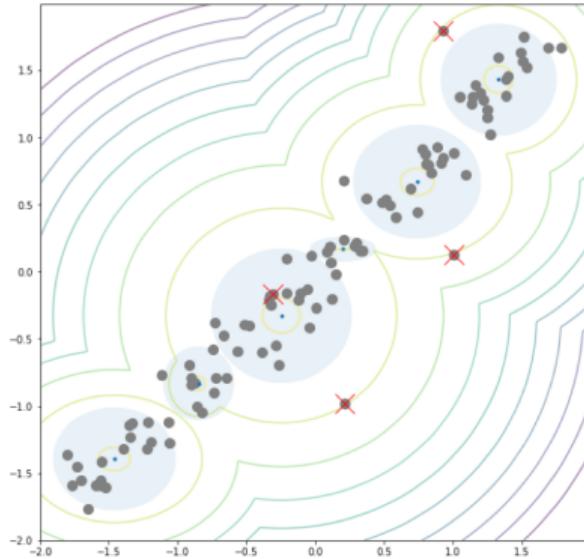


Рис. 6:
Кластеризация
смесью гауссиан

Серые точки ● — наблюдения, красные крестики ✕ — аномалии

Обнаружение аномалий на основе вероятностных моделей

- ▶ **Предположение:** выборка порождена некоторым вероятностным распределением, которое можно оценить
- ▶ Оценка генеративной плотности вероятностного распределения данных
- ▶ Урезание распределения по порогу для определения границ нормальности
- ▶ Простейший случай: статистические тесты на обнаружение выбросов
- ▶ Широко используется гипотеза нормальности данных
- ▶ Параметрические методы: смеси гауссиан (Gaussian mixture models, GMM) [Chandola и др., 2009; Markou и др., 2003; Miljković, 2010]
- ▶ Непараметрические методы: ядерная оценка плотности (kernel density estimate, KDE) [Chandola и др., 2009; Duda и др., 2012; Markou и др., 2003]

Обнаружение аномалий на основе мер близости

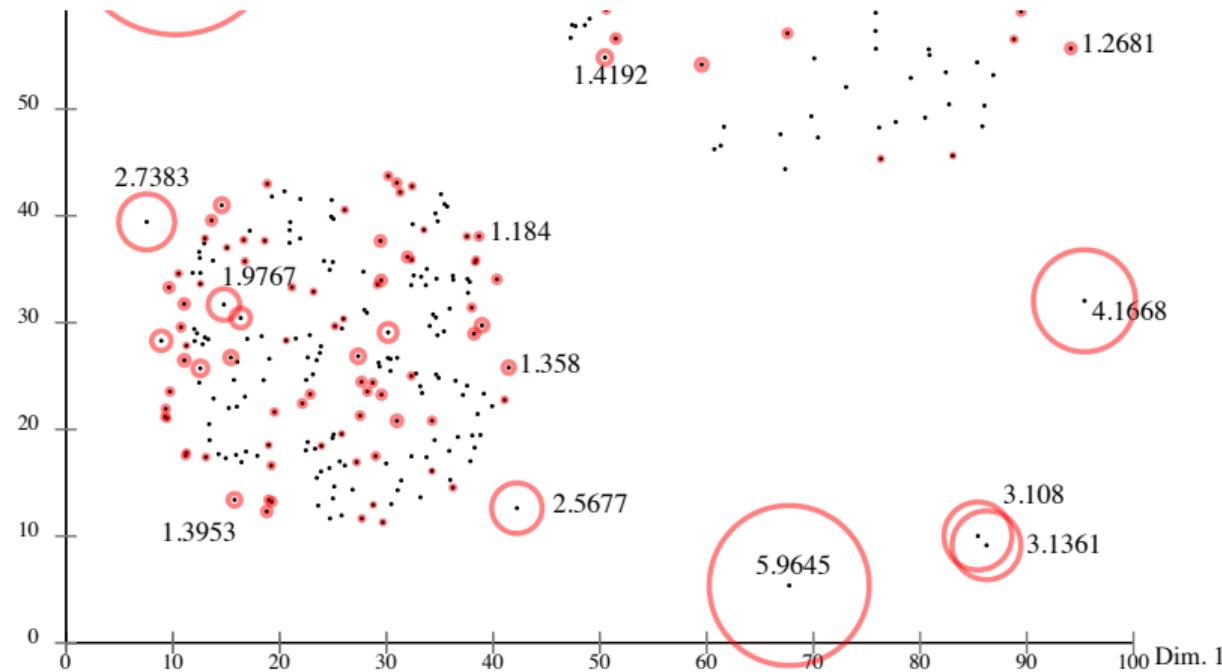


Рис. 7: Большие кружки соответствуют высокой вероятности аномальности, а маленькие – невысокой.

Обнаружение аномалий на основе мер близости

- ▶ **Предположение:** существует метрика, адекватно отражающая близость между наблюдениями (*нормальные данные окружены какими-то другими данными, а аномальные – одиноки*)
- ▶ Евклидова метрика, расстояние Махalanобиса, ...
- ▶ Плохое качество в случае многомерных данных
- ▶ Кластеризация: k -ближайших (k nearest neighbours, k NN) [A. Srivastava, 2006; Ashok N Srivastava и др., 2005]
- ▶ Локальная плотность данных: local outlier factor, LOF [Breunig и др., 2000]

Обнаружение аномалий на основе прогнозирования

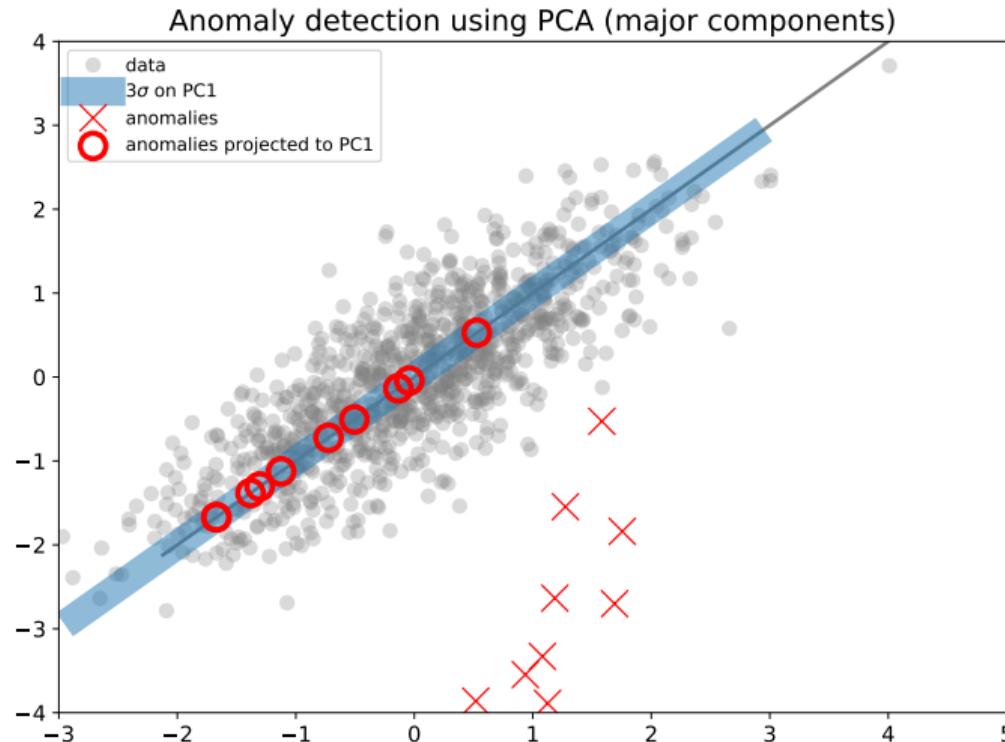


Рис. 8: Серые точки ● — наблюдения, красные крестики ✕ — аномалии

Обнаружение аномалий на основе прогнозирования

- ▶ **Предположение 1:** объекты тестовой выборки можно прогнозировать, высокая ошибка прогноза – признак аномальности
 - ▶ Сжатие данных репликативными нейросетями [Augusteijn и др., 2002; S. Hawkins и др., 2002; Williams и др., 2002]
- ▶ **Или предположение 2:** вложение данных в низкоразмерное подпространство позволит отличить нормальные и аномальные наблюдения
 - ▶ Данные «живут» на линейном подпространстве исходного пространства (предположения PCA) [Dutta и др., 2007; Hoffmann, 2007; Schölkopf, Smola и др., 1998; Shyu и др., 2003]

Обнаружение аномалий на основе описания границы носителя данных

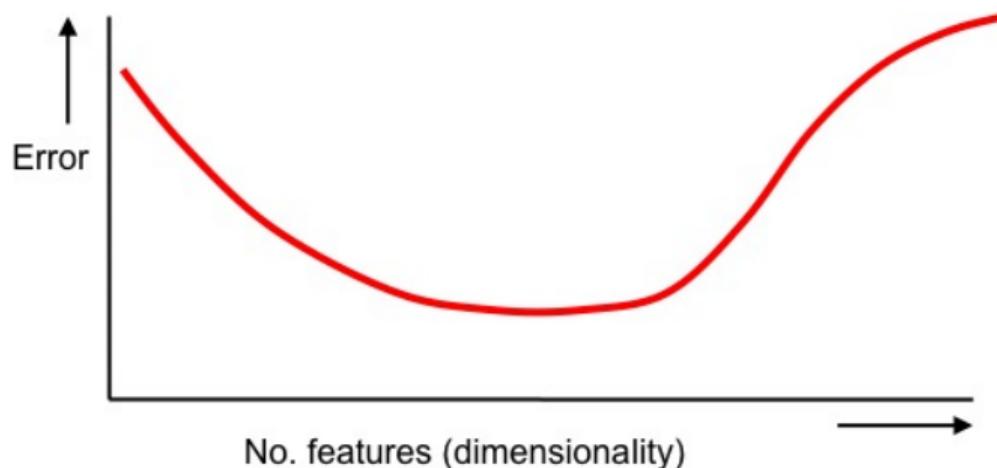
- ▶ Описание **границы носителя нормальных данных**
- ▶ Методы, нечувствительные к плотности нормального класса (описывают только границу)
 - ▶ Одноклассовый SVM определяет положение границы аномальной области используя только наблюдения, близкие к границе [Manevitz и др., 2001; Schölkopf, Williamson и др., 2000; Tax и др., 1999]

Содержание

- 1 Аномалии, их характеристизация и свойства
- 2 Обзор математических моделей аномалий
- 3 Обнаружение аномалий без учителя в данных высокой размерности
 - Задачи снижения размерности и линейная модель РСА
 - Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
 - Стохастические модели «нормальности». Модель смеси гауссиан
- 4 Резюме лекции

«Проклятие размерности» в задачах классификации

- ▶ Не всегда доступны размеченные данные, содержащие аномальные события
- ▶ Малая размерность влечёт недостаточное описание для классификации
- ▶ Чрезмерно большая — влечёт разреженность пространства объектов и переобучение классификатора
- ▶ **Проклятие размерности:** задачи классификации плохо решаются в данных высокой размерности



Проклятие размерности

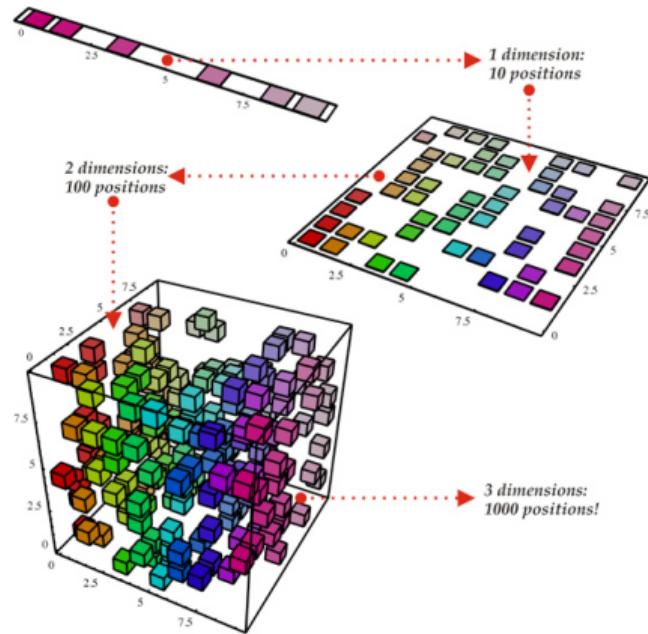


Рис. 9: Большая размерность требует большее число объектов для равномерного «покрытия» части пространства.

Проклятие размерности

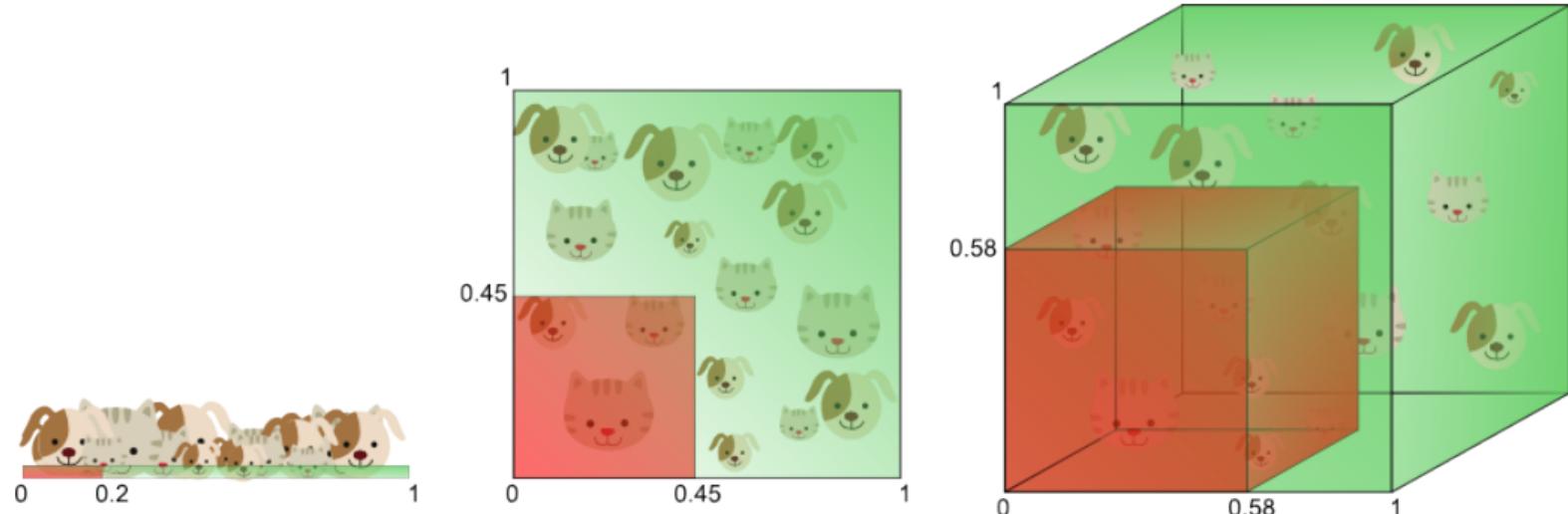


Рис. 10: Чтобы занять 20% объема пространства объектов, нужно захватить всё большую часть оси.

Цели снижения размерности

- ▶ **Проблема мультиколлинеарности:** при решении задачи регрессии методом наименьших квадратов

$$\beta = (X^T X)^{-1} X^T$$

$\det(X^T X) \approx 0 \Rightarrow$ неустойчивая оценка параметров моделей и их дисперсий

- ▶ **Упрощение вычислений:** меньше размерность \Rightarrow проще вычисления, требуется меньше памяти для хранения выборки.

Представление по выборке многообразия данных

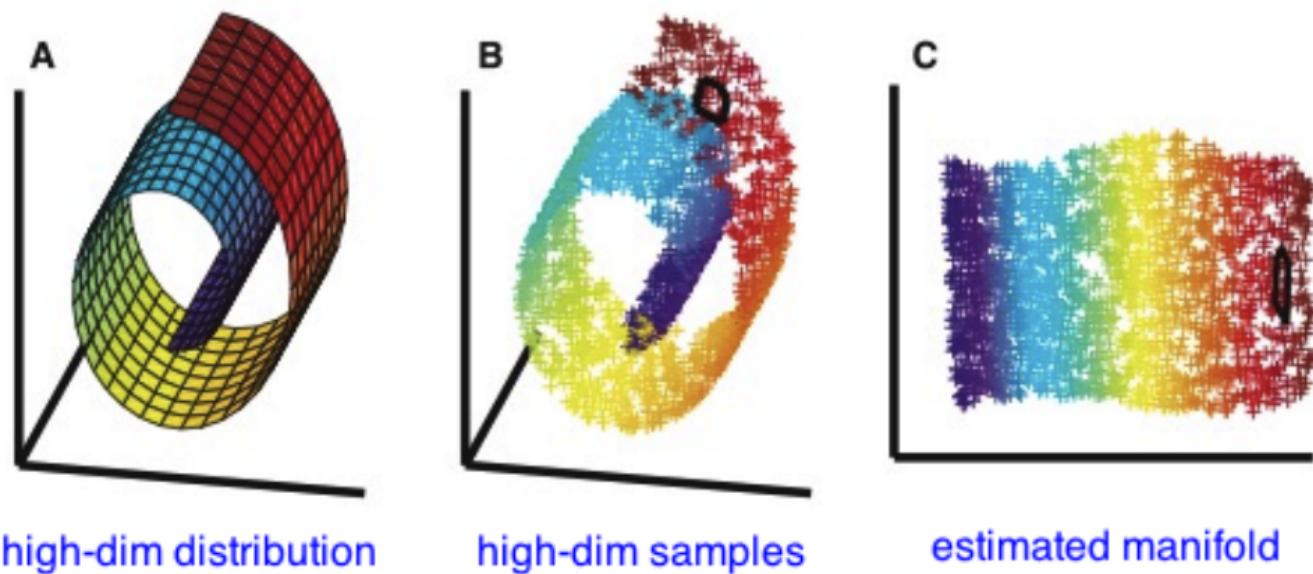


Рис. 11: Цель: необходимо обнаружить некоторое многообразие, лежащее в пространстве высокой размерности.

Представление по выборке многообразия данных

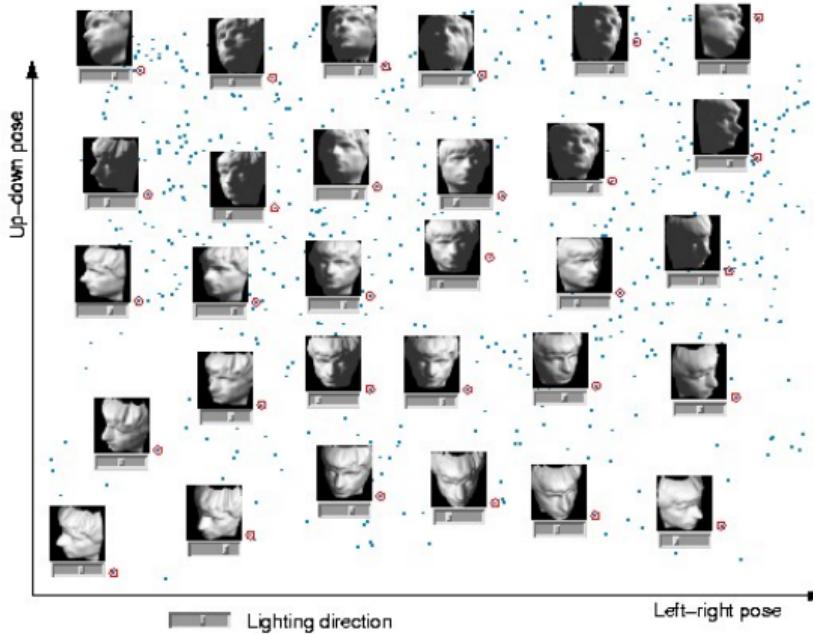


Рис. 12: В пространстве всех возможных фотографий лежит пространство низкой размерности — фото лица со всех возможных сторон. Мы можем описать изображение, задав положение камеры на единичной сфере. Это можно сделать с помощью двух углов.

Представление по выборке многообразия данных

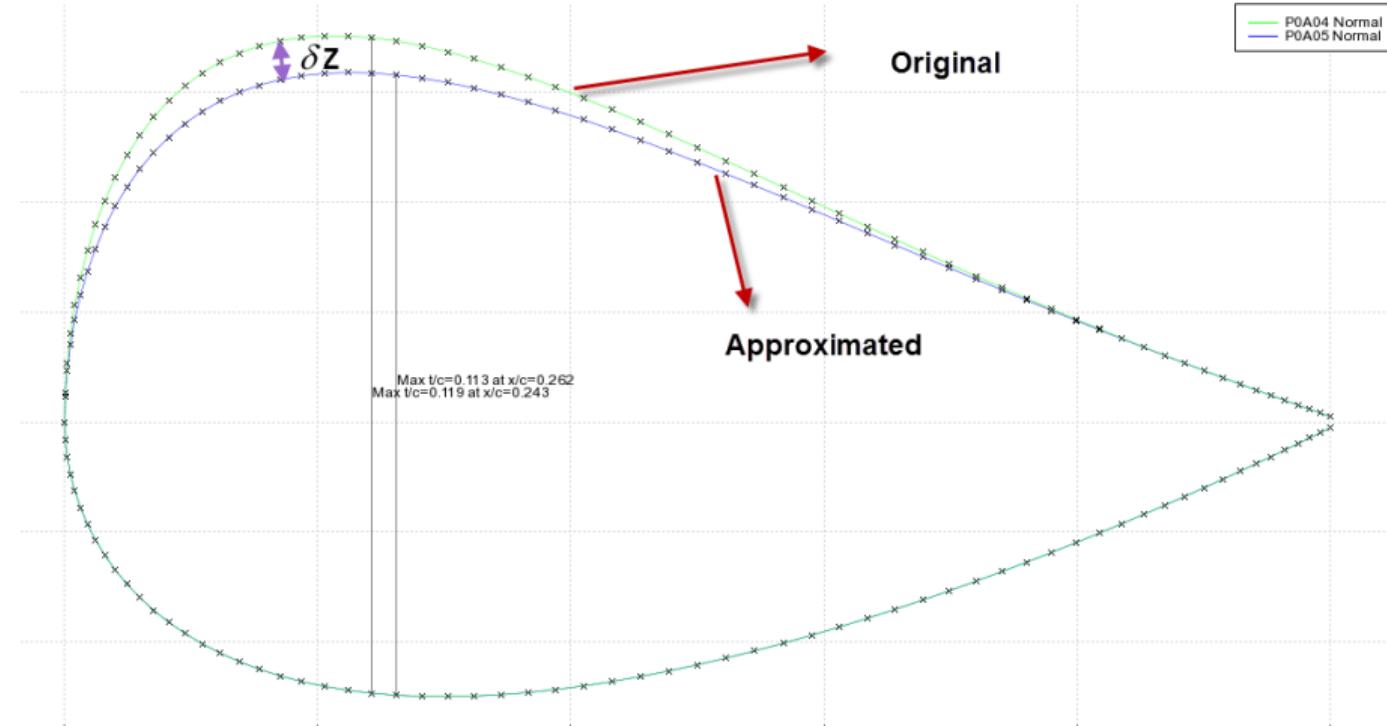


Рис. 13: Множество сечений (профилей) некоего крыла, изначально описанного 59 координатами, может быть описано шестью координатами.

Визуализация

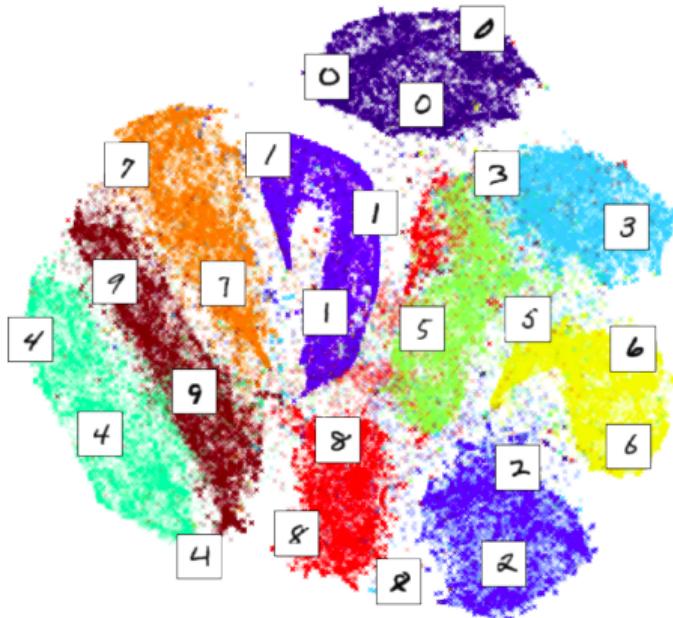


Рис. 14: Для человека удобным представлением данных является двухмерное (трёхмерное) представление. В данном случае размерность данных MNIST(изображения рукописных цифр 28x28) снижена до 2.

Постановка задачи снижения размерности

- ▶ Пусть объект $O \in S$ описывается вектором $X(O) \in \mathbb{R}^p$.
- ▶ Всё множество объектов описано набором $\mathbb{X} = \{X(O), O \in S\}$.
- ▶ Пусть описание доступно только для объектов $\{O_1, \dots, O_n\}$:

$$\mathbf{X}_n = \{X_i = X(O_i), i = 1, 2, \dots, n\} = \{X_1, X_2, \dots, X_n\}$$

- ▶ **Задача:** необходимо построить правило краткого описания объектов $y(X(O)) \in \mathbb{R}^q, q < p$ без «значимой» потери точности описания

Виды задачи

- ▶ **Embedding Problem (E-problem)**
 - ▶ Отображение должно быть определено только на обучающей выборке
- ▶ **Extended Embedding Problem (EE-problem)**
 - ▶ Отображение должно быть определено на любом детальном описании объекта
- ▶ **Full Dimension Reduction problem (Full DR-problem)**
 - ▶ Нужно также найти некоторое обратное преобразование из краткого описания в детальное

Содержание

- 1 Аномалии, их характеристизация и свойства
- 2 Обзор математических моделей аномалий
- 3 Обнаружение аномалий без учителя в данных высокой размерности
 - Задачи снижения размерности и линейная модель РСА
 - Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
 - Стохастические модели «нормальности». Модель смеси гауссиан
- 4 Резюме лекции

Метод главных компонент

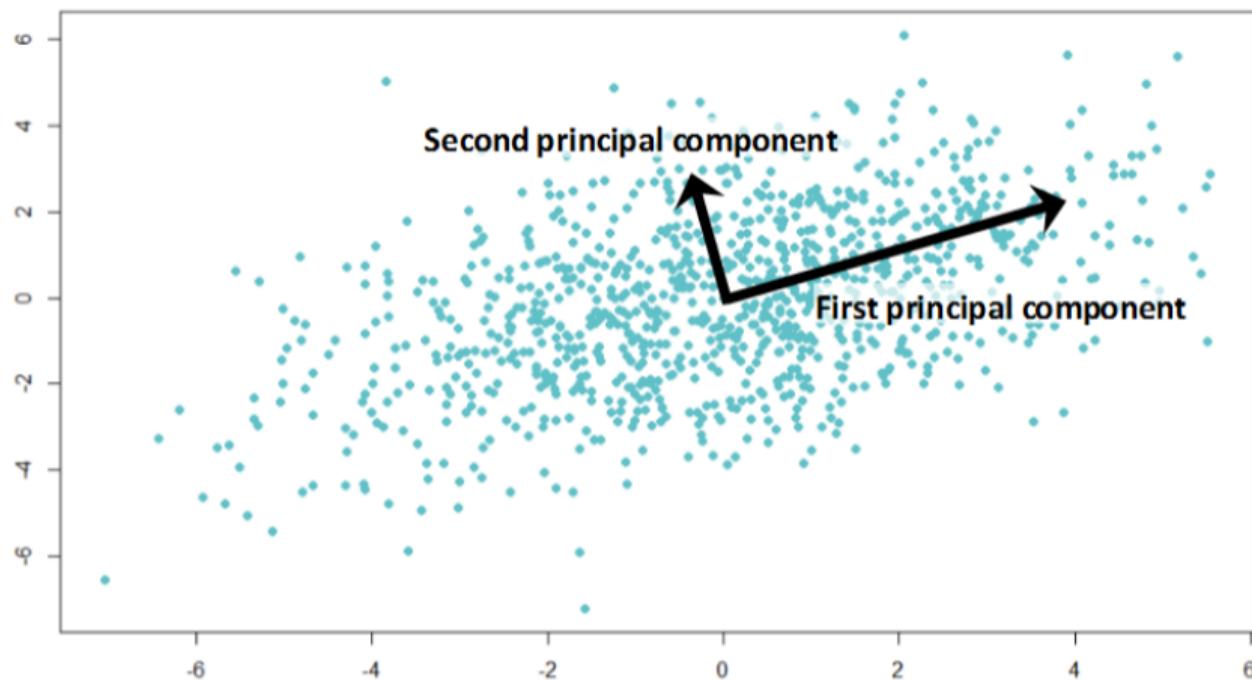
- ▶ Доступно описание n объектов размерности p : X_1, X_2, \dots, X_n
- ▶ Мы хотим построить отображение $y : \mathbb{R}^p \rightarrow \mathbb{R}^q, q < p$
- ▶ Идея: приблизим исходное пространство линейной оболочкой из q компонент p -мерного ортогонального базиса и p -мерного смещения:

$$x \approx \mu + \mathbf{V}_q y(x), \quad \mathbf{V}_q \in \mathbb{R}^{p \times q}, \quad \mathbf{V}_q^T \mathbf{V}_q = \mathbf{I}_p,$$

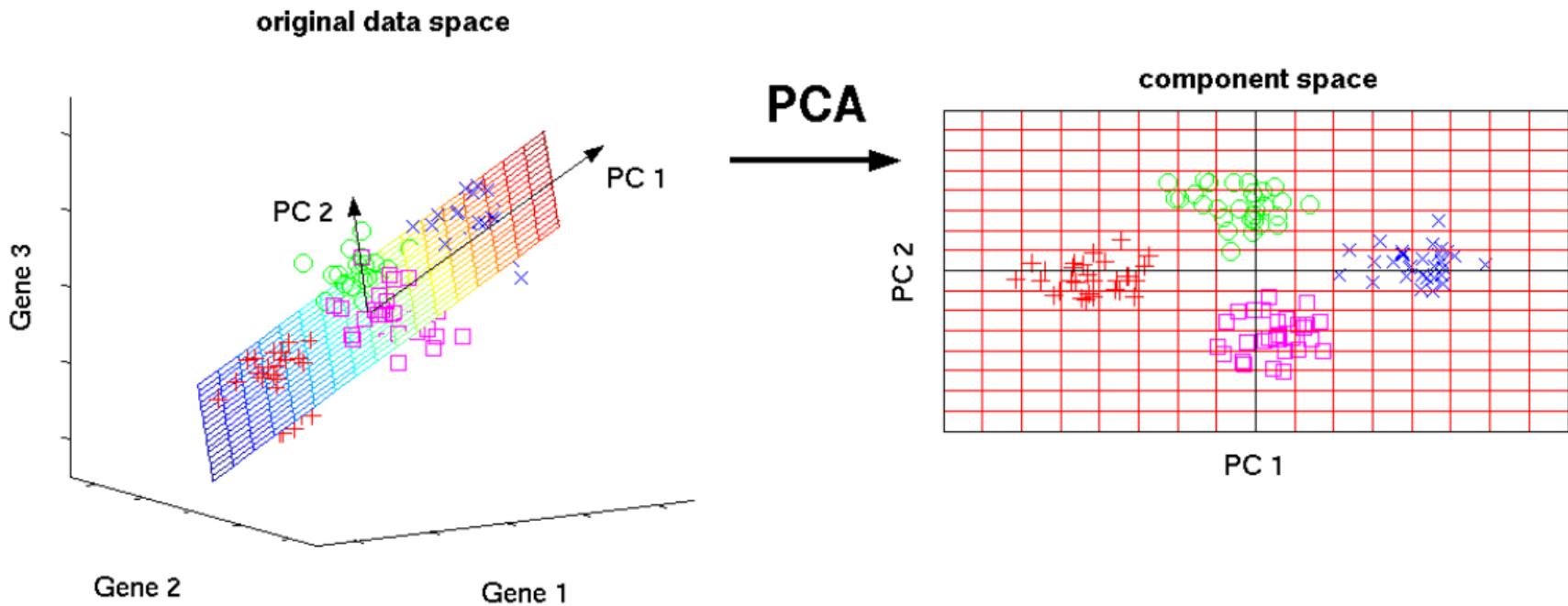
минимизируя квадрат ошибки восстановления:

$$\min_{\mu, \{y_i\}_{i=1}^n, \mathbf{V}_q} \sum_{i=1}^n \|X_i - \mu - \mathbf{V}_q y_i\|^2$$

Метод главных компонент



Метод главных компонент



Метод главных компонент

- Имеем функцию штрафа:

$$\min_{\mu, \{y_i\}_{i=1}^n, \mathbf{V}_q} \sum_{i=1}^n \|X_i - \mu - \mathbf{V}_q y_i\|^2$$

- Продифференцировав, можно найти:

$$\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad \{y_i = \mathbf{V}_q^T X_i\}_{i=1}^n.$$

- Тогда задача сводится к

$$\min_{\mathbf{V}_q} \sum_{i=1}^n \|(X_i - \bar{X}_n) - \mathbf{V}_q \mathbf{V}_q^T (X_i - \bar{X}_n)\|^2,$$

где $\mathbf{V}_q \mathbf{V}_q^T$ -матрица, производящая последовательное сжатие и разжатие центрированной выборки.

Метод главных компонент

- ▶ Решение задачи на минимум

$$\min_{\mathbf{V}_q} \sum_{i=1}^n \left\| (X_i - \bar{X}_n) - \mathbf{V}_q \mathbf{V}_q^T (X_i - \bar{X}_n) \right\|^2,$$

может быть найдено через SVD-разложение центрированной матрицы исходных данных:

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T, \quad \mathbf{X} = [(X_1 - \bar{X}_n), (X_2 - \bar{X}_n), \dots, (X_n - \bar{X}_n)]^T$$

- ▶ При этом

- ▶ \mathbf{V}_q состоит из первых q столбцов матрицы \mathbf{V}
- ▶ Столбцы матрицы \mathbf{V}_q называют **главными компонентами** матрицы \mathbf{X}
- ▶ Матрица \mathbf{X} есть матрица, строками которой являются векторы $X_i, i = 1, \dots, n$

Отбор признаков в методе главных компонент

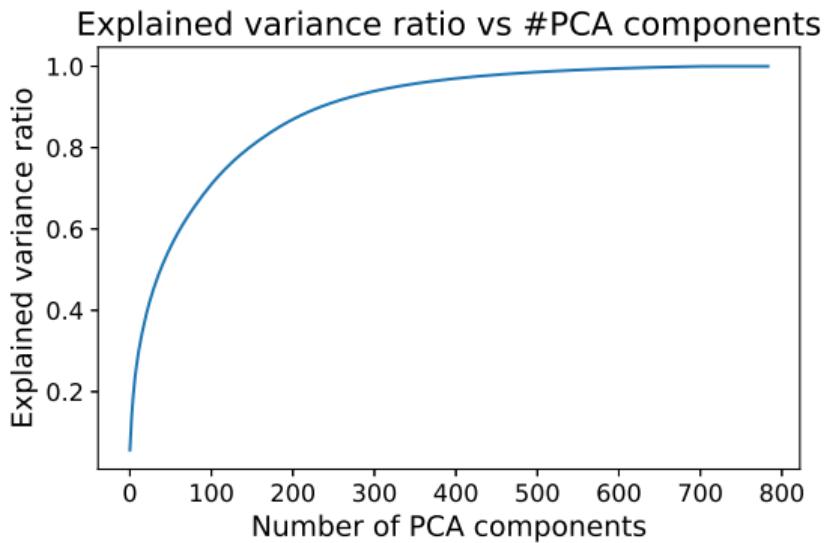
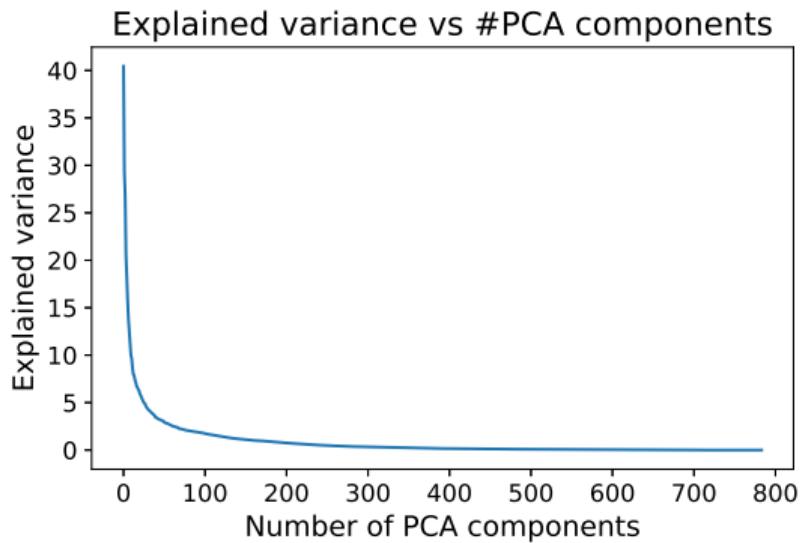


Рис. 15: Из графика можно видеть, что переменные «выбираются» по принципу убывания вдоль них дисперсии(«если дисперсия большая, то от этой переменной многое зависит, компонента значима для описания данных»).

Результаты сингулярного разложения на данных MNIST

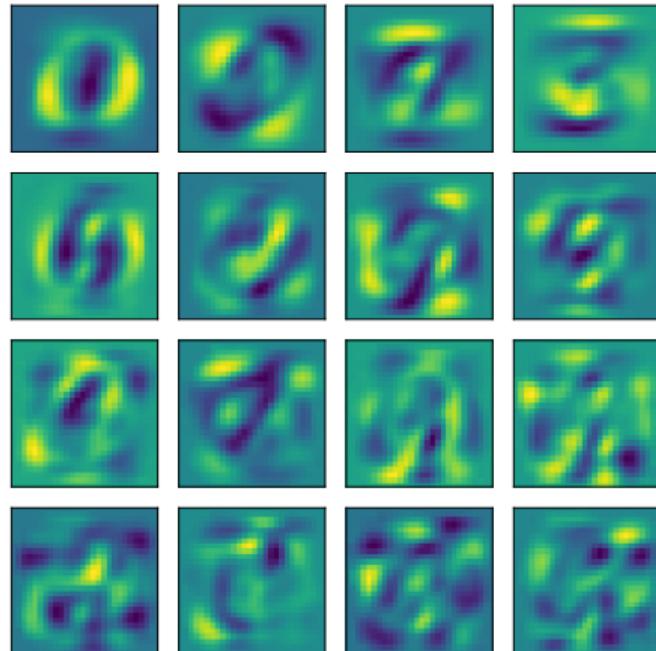


Рис. 16: Первые 16 столбцов
матрицы \mathbf{V}_q для всех изображений

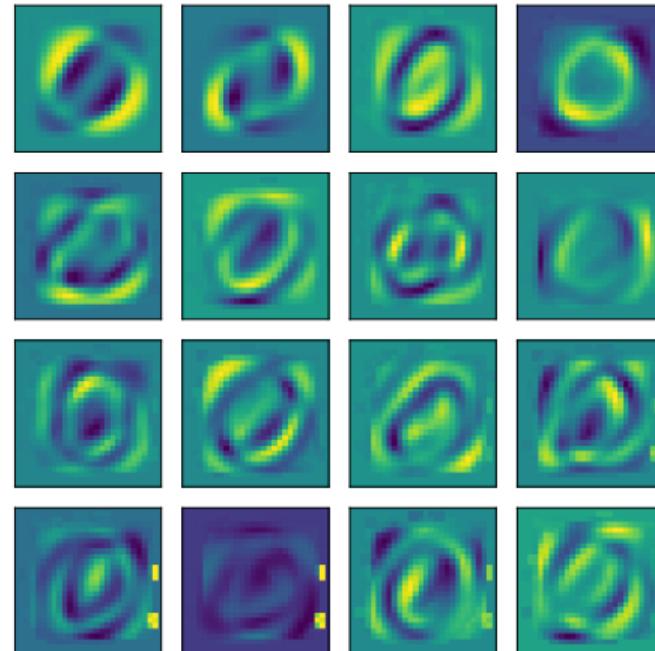


Рис. 17: Первые 16 столбцов
матрицы \mathbf{V}_q для изображений нуля

Преимущества метода

- ▶ Легко считается, высокая производительность
- ▶ Относительно интерпретируем
- ▶ Позволяет проводить и прямое, и обратное преобразование

Обнаружение аномалий на основе PCA

- ▶ Рассмотрим проекцию некоторого тестового наблюдения \mathbf{z} на линейную оболочку $L(\mathbf{v}_1, \dots, \mathbf{v}_q)$ главных компонент $\mathbf{v}_1, \dots, \mathbf{v}_q$ (столбцов матрицы \mathbf{V}_q):

$$\mathbf{y} = \mathbf{P}_L \mathbf{z} = (\mathbf{z}, \mathbf{v}_1) \mathbf{v}_1 + \dots + (\mathbf{z}, \mathbf{v}_q) \mathbf{v}_q$$

(берется стандартизованная версия $(z_k - \hat{\mu}_k)/\hat{\sigma}_{kk}$ координат \mathbf{z})

- ▶ Величина

$$\sum_{i=1}^q \frac{y_i^2}{\lambda_i^2} = \frac{y_1^2}{\lambda_1^2} + \dots + \frac{y_q^2}{\lambda_q^2}, \quad q \leq p,$$

эквивалентна расстоянию Махalanобиса от \mathbf{z} до выборочного среднего $\hat{\mu}$

- ▶ **Мера аномальности:** отличие аномального экземпляра от нормальных, выражаемое расстоянием в пространстве главных компонент

Правило обнаружения аномалий (1)

- ▶ Правило 1: считать наблюдение

аномалией, если

$$\sum_{i=1}^q \frac{y_i^2}{\lambda_i^2} > \chi_q^2(\alpha)$$

- ▶ Интуиция: $\forall y_i = \lambda_i^2$, поэтому, если $y_i \sim \mathcal{N}(0, 1)$, то $\sum_{i=1}^q \frac{y_i^2}{\lambda_i^2} \sim \chi_q^2$
- ▶ Классификация больших отклонений в направлениях главных компонент

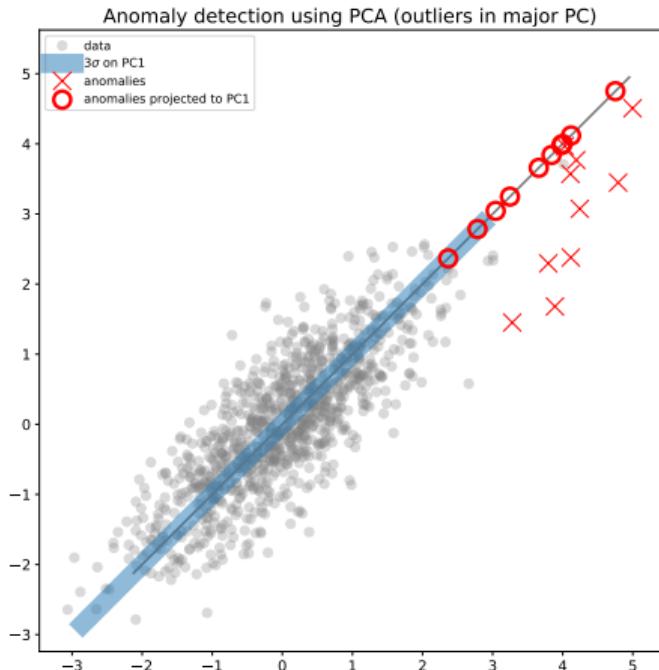


Рис. 18: Обнаружение сильных отклонений в первой главной компоненте данных

Правило обнаружения аномалий (2)

- ▶ Правило 2: считать наблюдение аномалией, если

$$\sum_{i=p-r+1}^p \frac{y_i^2}{\lambda_i^2} > \chi_r^2(\alpha)$$

- ▶ Интуиция: $\mathbb{V}y_i = \lambda_i^2$, поэтому, если $y_i \sim \mathcal{N}(0, 1)$, то $\sum_{i=1}^q \frac{y_i^2}{\lambda_i^2} \sim \chi_q^2$
- ▶ Классификация больших отклонений в остатках
- ▶ Соответствует высокой погрешности приближения исходного вектора \mathbf{z} вектором $P_L \mathbf{z}$ из линейной оболочки первых q главных компонент

$$||\mathbf{z} - P_L \mathbf{z}||^2 \text{ is large}$$

Правило обнаружения аномалий (2)

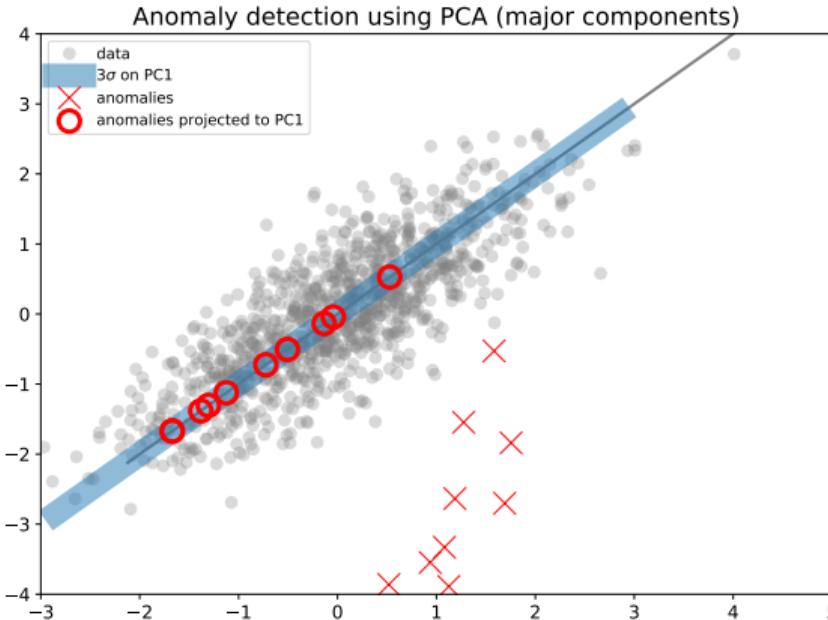


Рис. 19: Данные больше не являются аномальными в проекции на главные компоненты с большой дисперсией

А. В. Артёмов

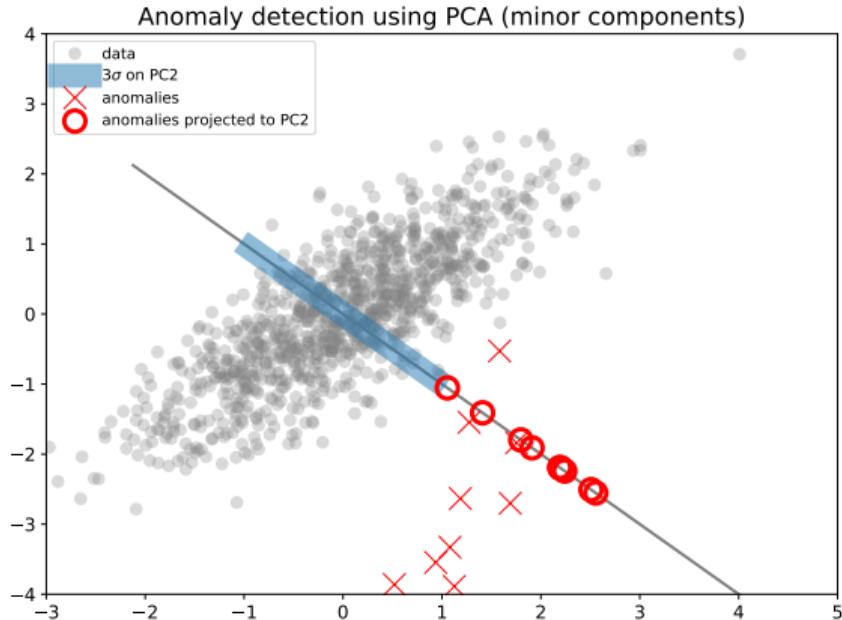


Рис. 20: Для определения аномальности требуется рассматривать ортогональное дополнение

Методы обнаружения аномалий

Свойства подхода к обнаружению аномалий на основе РСА

- ▶ **Преимущества:**

- ▶ Свободен от ограничительных гипотез о распределении вероятностей данных (нормальности и т. п.)
- ▶ В результате приводит к простым линейным правилам классификации, эффективным при высоких размерностях входа

- ▶ **Недостатки:**

- ▶ Предположение о линейности многообразия, содержащего нормальные точки данных

Содержание

- 1 Аномалии, их характеристизация и свойства
- 2 Обзор математических моделей аномалий
- 3 Обнаружение аномалий без учителя в данных высокой размерности
 - Задачи снижения размерности и линейная модель PCA
 - Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
 - Стохастические модели «нормальности». Модель смеси гауссиан
- 4 Резюме лекции

Одноклассовая машина опорных векторов

- ▶ Переформулирование задачи поиска аномалий в задачу одноклассовой классификации:
 - ▶ Отделение выборки от начала координат на максимальный отступ
 - ▶ Поиск небольшой области, в которой лежит большая часть данных, и классификация всех точек в этой области как принадлежащих одному классу
 - ▶ Параметры: ожидаемое число аномалий, сглаживание разделяющей гиперплоскости
- ▶ Отделение областей пространства, содержащих данные, от пустых областей

Сфера минимального объема и алгоритм SVDD

Пусть x_1, \dots, x_l – точки в некотором нормированном пространстве.

Оптимизационная задача для отыскания сферы минимального объема, содержащей все точки:

$$\begin{aligned} R^2 &\rightarrow \min_{R,a} \\ \|x_i - a\|^2 &\leq R^2 \end{aligned}$$

Сфера минимального объема и алгоритм SVDD

Отобразим данные в пространство более высокой размерности с помощью $\phi(\cdot)$:

$$\begin{aligned} R^2 &\rightarrow \min_{R,a} \\ \|\textcolor{red}{x}_i - a\|^2 &\leq R^2 \end{aligned}$$

Сфера минимального объема и алгоритм SVDD

Отобразим данные в пространство более высокой размерности с помощью $\phi(\cdot)$:

$$\begin{aligned} R^2 &\rightarrow \min_{R,a} \\ \|\phi(x_i) - a\|^2 &\leq R^2 \end{aligned}$$

Сфера минимального объема и алгоритм SVDD

Если точка находится слишком далеко от центра, она может быть аномальной и может негативно повлиять на решение.

Позволим некоторым точкам быть вне сферы и рассмотрим релаксацию изначальной задачи.

$$R^2 + C \sum_{i=1}^l \max(0, \|\phi(x_i) - a\|^2 - R^2) \rightarrow \min_{R,a}$$

$\max(0, \|\phi(x_i) - a\|^2 - R^2)$ равно нулю для точек внутри сферы

Сфера минимального объема

Перепишем задачу с помощью вспомогательных переменных ξ_i

$$R^2 + \sum_{i=1}^l \max(0, \|\phi(x_i) - a\|^2 - R^2) \rightarrow \min_{R,a}$$

Определим $\xi_i = \max(0, \|\phi(x_i) - a\| - R^2)$

Сфера минимального объема

Перепишем задачу с помощью вспомогательных переменных

$$\begin{aligned} R^2 + \sum_{i=1}^l \xi_i &\rightarrow \min_{R,a} \\ \|\phi(x_i) - a\|^2 &\leq R^2 + \xi_i \\ \xi_i &\geq 0 \end{aligned}$$

Сфера минимального объема

Последний шаг: сделаем задачу выпуклой

$$\begin{aligned} R^2 + \sum_{i=1}^l \xi_i &\rightarrow \min_{R,a} \\ \|\phi(x_i) - a\|^2 &\leq R^2 + \xi_i \\ \xi_i &\geq 0 \end{aligned}$$

Сфера минимального объема

Последний шаг: сделаем задачу выпуклой

$$\begin{aligned} \textcolor{red}{R} + \sum_{i=1}^l \xi_i &\rightarrow \min_{R,a} \\ \|\phi(x_i) - a\|^2 &\leq \textcolor{red}{R} + \xi_i \\ \xi_i &\geq 0 \\ \textcolor{red}{R} &\geq 0 \end{aligned}$$

Постановка оптимизационной задачи

Вход:

- ▶ Точки $X_1, \dots, X_l \subset \mathbb{R}^m$
- ▶ Отображение $\phi : \mathbb{R}^m \rightarrow \mathbb{H}_\phi$

Хотим:

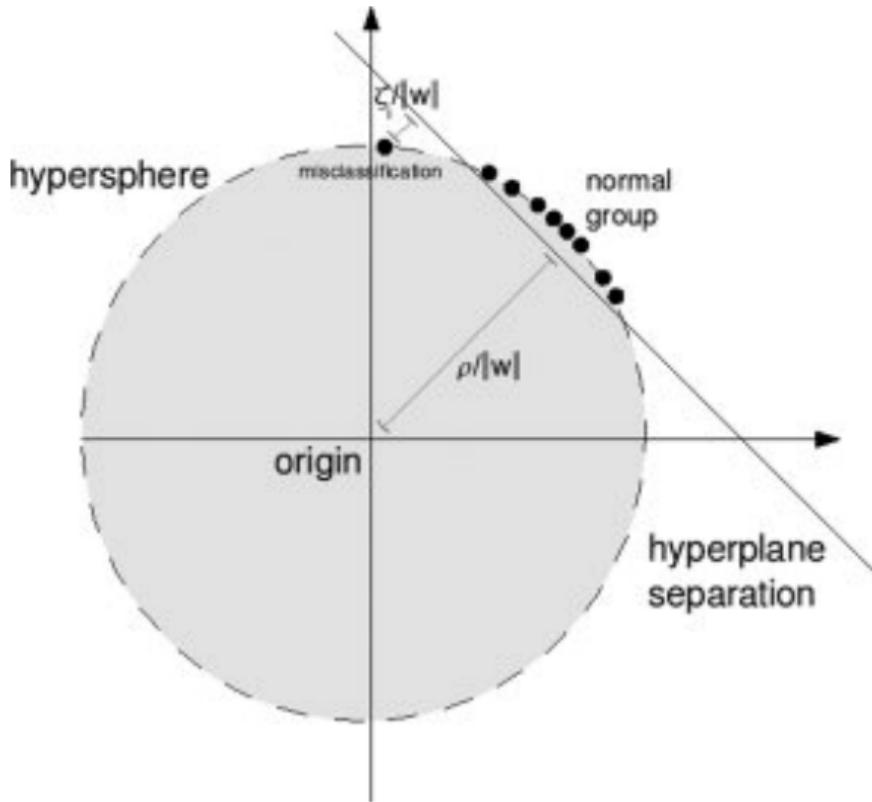
- ▶ Отделить точки от начала координат \mathbb{H}_ϕ

Оптимизационная задача:

$$\frac{\nu l}{2} \|w\|^2 - \rho \nu l + \sum_{i=1}^l \xi_i \rightarrow \min_{w, \rho, \xi}$$

$$(w \cdot \phi(X_i)) \geq \rho - \xi_i$$

$$\xi_i \geq 0$$



Оптимизационная задача машины опорных векторов

- ▶ Исходная задача эквивалентна дуальной:

$$Q = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \min_{\{\alpha_i\}}$$

при условиях:

$$0 \leq \alpha_i \leq \frac{1}{l\nu}, \quad \sum_i \alpha_i = 1, \rho \geq 0, \quad \nu \in [0; 1],$$

где α_i — множители Лагранжа, ν — верхняя граница доли ошибок на \mathbf{X}^ℓ

- ▶ Когда оптимизационная задача решена, найдутся минимум $l\nu$ точек с ненулевыми множителями Лагранжа (опорных векторов)
- ▶ Алгоритм вычисляет решающую функцию:

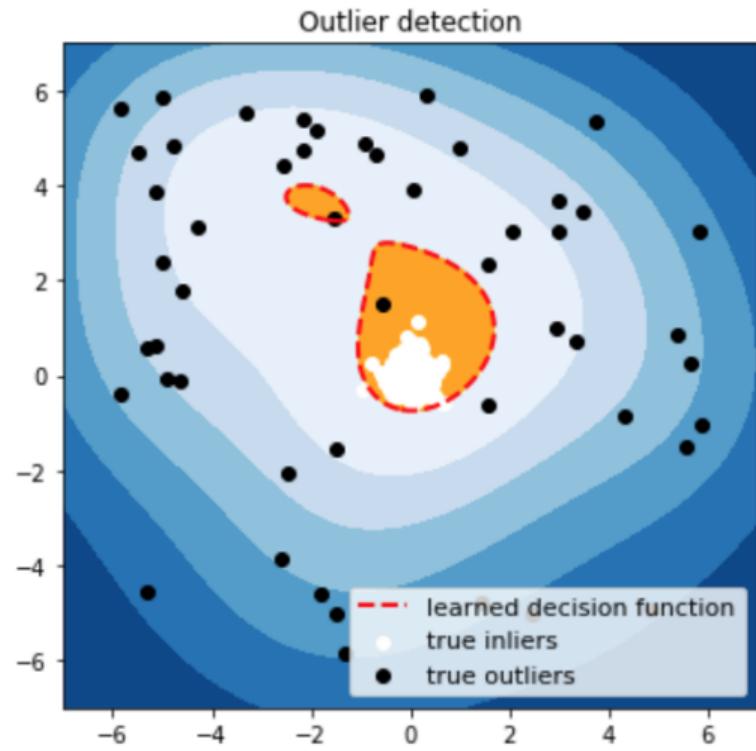
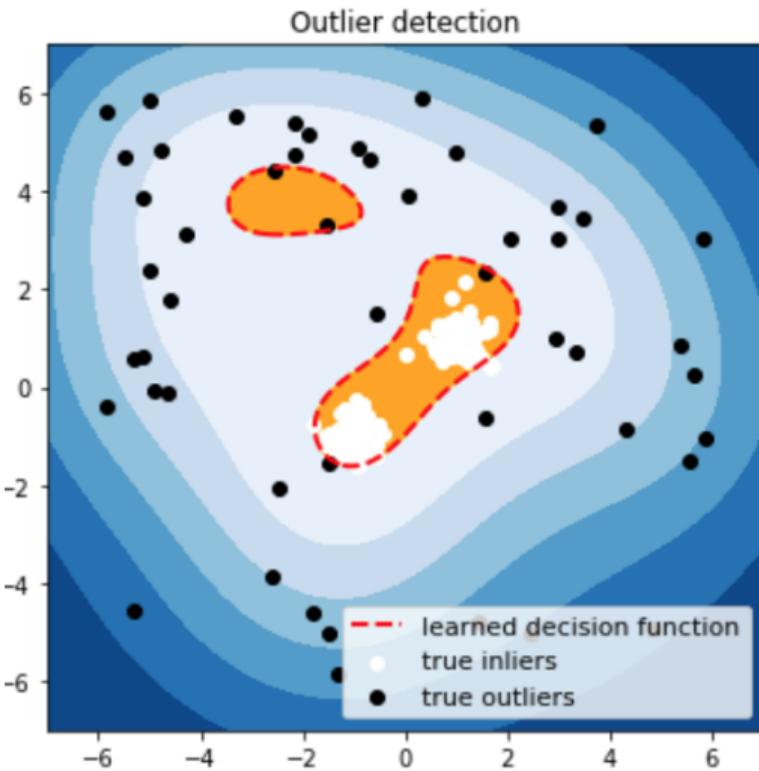
$$f(\mathbf{x}) = \text{sign}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho \right)$$

Примеры ядер

В зависимости от задачи используются разные типы ядер:

- ▶ $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{x}_2$ — линейное ядро
- ▶ $k(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1^T \mathbf{x}_2 + c)^d$ — полиномиальное ядро
- ▶ $k(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\|\mathbf{x}_1 - \mathbf{x}_2\|^2/\sigma^2)$ — радиальное базисное ядро
- ▶ $k(\mathbf{x}_1, \mathbf{x}_2) = \frac{LCS(\mathbf{x}_1, \mathbf{x}_2)}{\sqrt{l_{\mathbf{x}_1} l_{\mathbf{x}_2}}}$: $LCS(\mathbf{x}_1, \mathbf{x}_2)$ — наибольшая общая подпоследовательность (строк) \mathbf{x}_1 и \mathbf{x}_2 .

Пример использования: синтетическая выборка



Пример: выявление аномальных начертаний MNIST

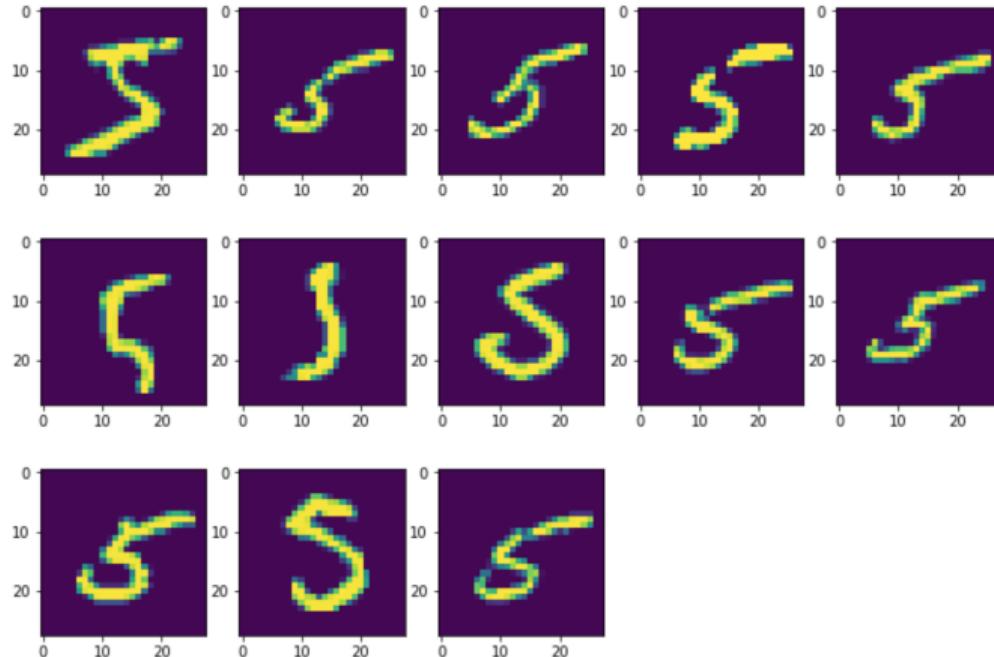


Рис. 22: Визуализация найденных аномальных изображений фиксированной цифры (13 аномалий из 200 примеров)

Пример: выявление аномальных начертаний MNIST

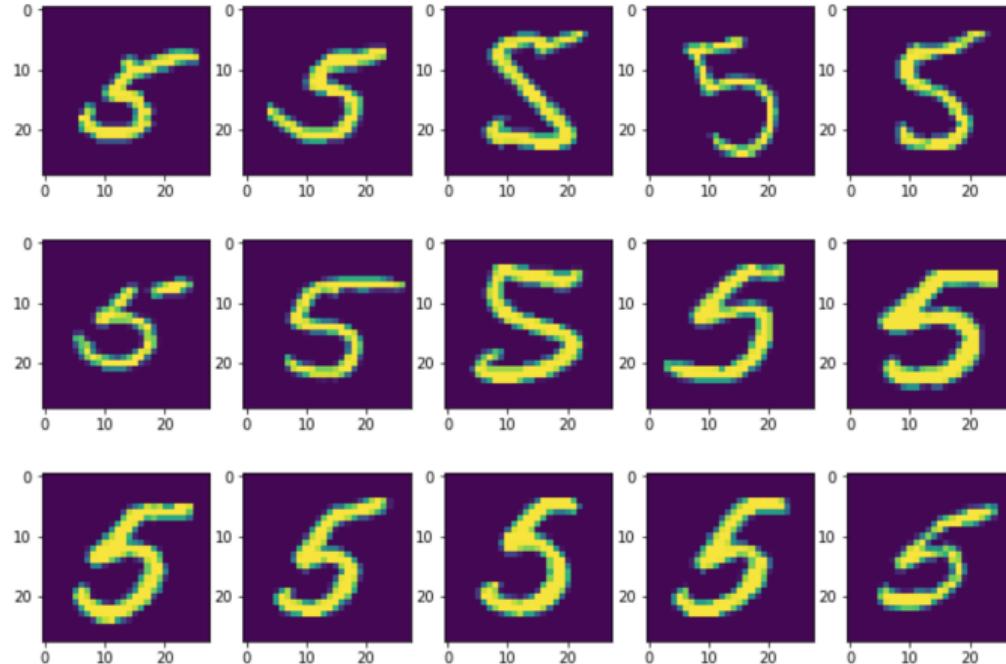


Рис. 23: Визуализация нормальных изображений фиксированной цифры (15 нормальных примеров из 200 цифр)

Пример: выявление аномальных начертаний MNIST (связь с t-SNE)

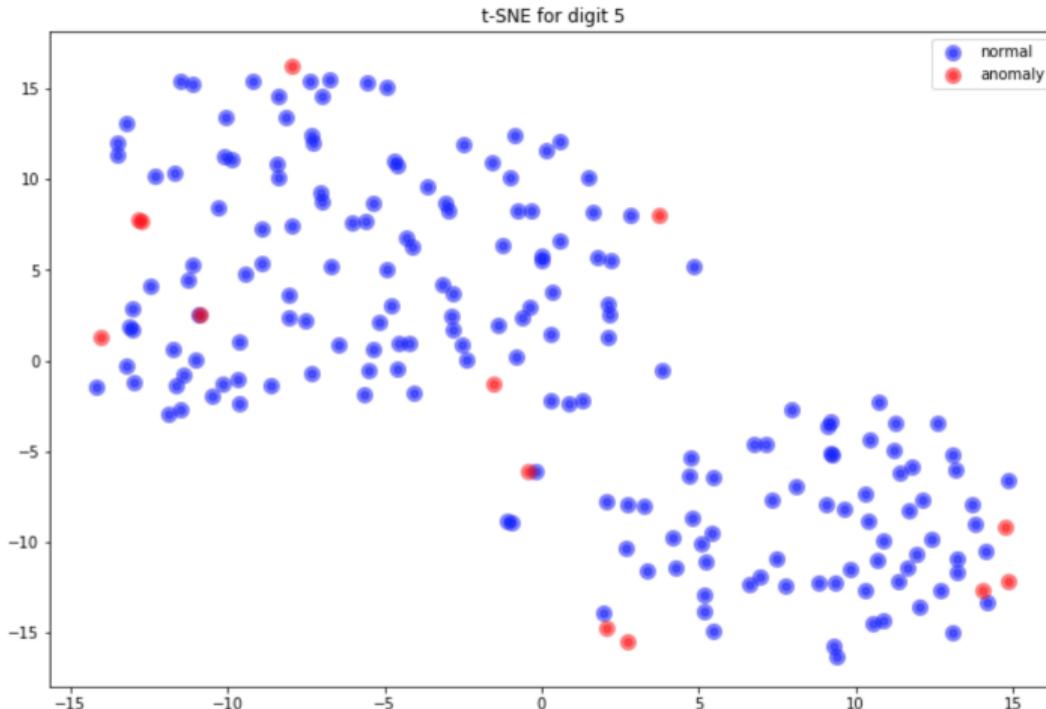


Рис. 24: Найденные аномалии на проекции mnist, построенной с помощью t-SNE. Видно, что найденные аномалии находятся на границах кластера цифры 5.

Свойства подхода к обнаружению аномалий на основе One-class SVM

- ▶ **Преимущества:**

- ▶ Применим в случае данных высокой размерности, когда другие подходы (оценка плотности) не справляются
- ▶ Гибкость выбора ядра
- ▶ Существуют онлайн-варианты, легко инкорпорировать привилегированную информацию [Burnaev и др., 2016]

- ▶ **Недостатки:**

- ▶ Проблема выбора модели: ширина ядра, параметр ν
- ▶ Отсутствуют оценки связи обобщающей способности и типа ядра

Содержание

- 1 Аномалии, их характеристизация и свойства
- 2 Обзор математических моделей аномалий
- 3 Обнаружение аномалий без учителя в данных высокой размерности
 - Задачи снижения размерности и линейная модель PCA
 - Нелинейная ядерная модель «нормальности» и алгоритм one-class SVM
 - Стохастические модели «нормальности». Модель смеси гауссиан
- 4 Резюме лекции

Стохастические модели «нормальности»

- ▶ Точки выборки моделируются с помощью вероятностного распределения \Rightarrow новые точки помечаются аномальными согласно оценке этого распределения
- ▶ **Преимущества:**
 - ▶ Использование существующих статистических подходов для моделирования различных типов распределений
- ▶ **Недостатки:**
 - ▶ Трудность описания нормального режима в данных высокой размерности
 - ▶ Реальные данные часто не соответствуют модельным параметрическим предположениям о них
- ▶ Параметрические подходы: моделирование на основе фиксированного класса распределений
- ▶ Непараметрические подходы: моделирование на основе непараметрической оценки плотности

Стохастические модели «нормальности»

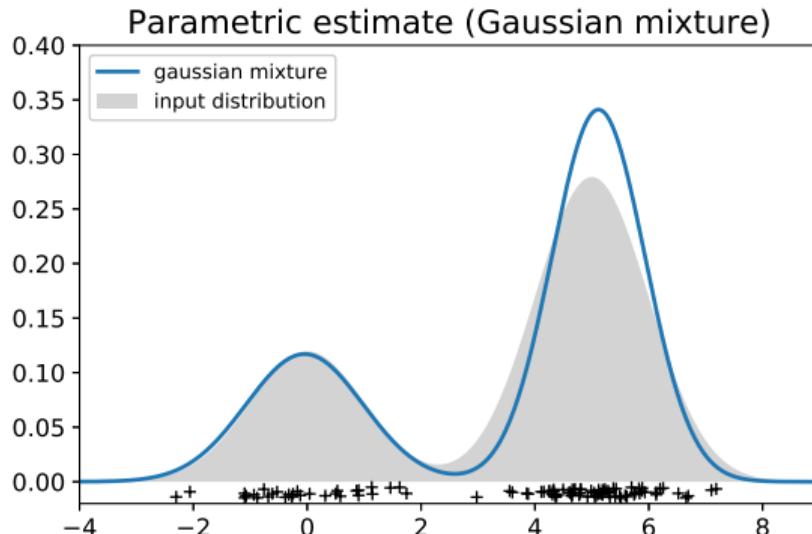
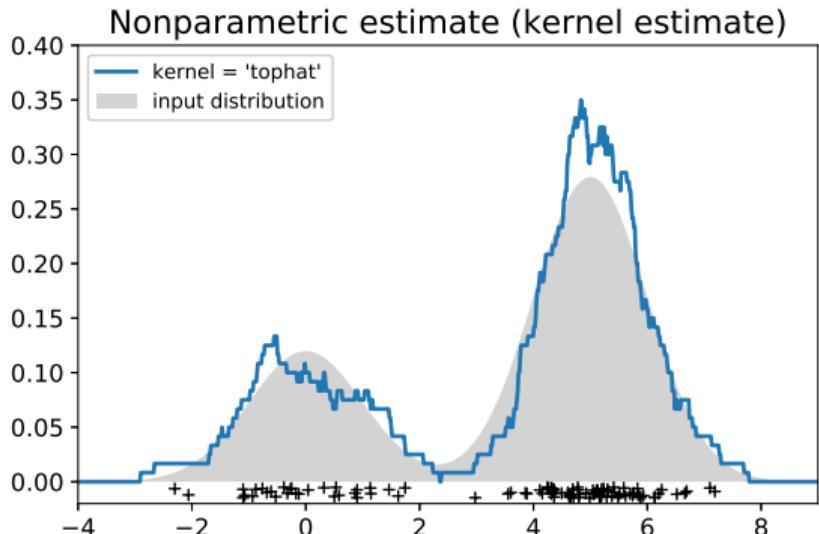


Рис. 25: Непараметрическая оценка плотности (слева) использует ядерное сглаживание для вычисления вероятности в каждой точке. Параметрическая модель (справа) оценивает только параметры распределения, которыми параметризована плотность.

Модель гауссовых смесей (Gaussian mixture model, GMM)

- ▶ Взвешенная смесь из K гауссовых распределений — распределение в многомерном пространстве \mathbb{R}^M объектов выборки с плотностью:

$$p(\mathbf{x}|\theta) = \sum_{i=1}^K \omega_i \rho(\mathbf{x}|\mu_i, \Sigma_i)$$

с параметрами $\theta = \{(\omega_i, \mu_i, \Sigma_i)\}_{i=1}^K$:

- ▶ $\rho(\mathbf{x}|\mu, \Sigma)$ — плотность гауссового распределения в точке $\mathbf{x} \in \mathbb{R}^M$,
- ▶ ω_i — вес кластера в общей смеси,
- ▶ μ_i, Σ_i — параметры гауссового распределения кластера i

Модель гауссовских смесей (Gaussian mixture model, GMM)

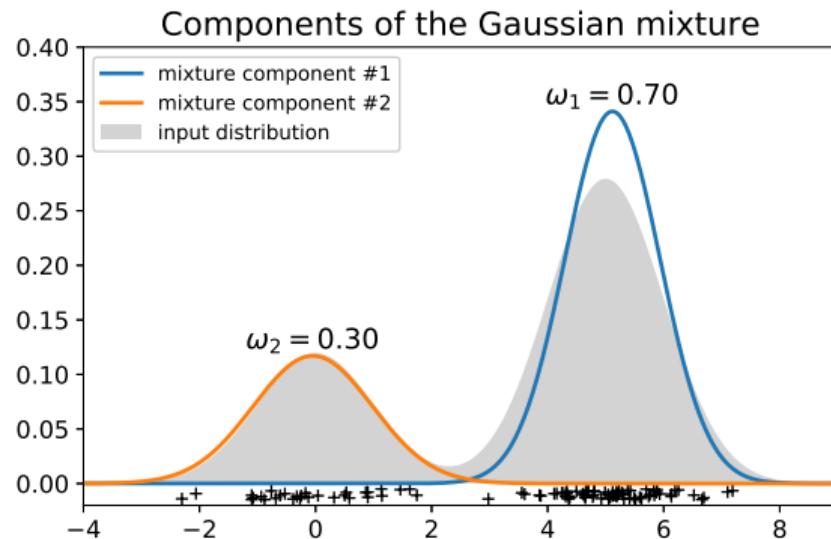
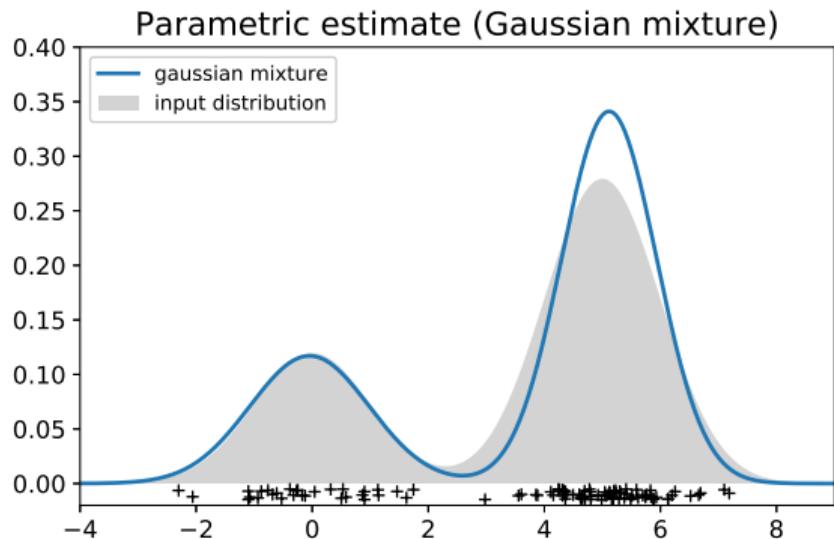


Рис. 26: Распределение вероятностей, которое представляет единую стохастическую модель данных (слева), на самом деле является взвешенной смесью нескольких распределений (справа).

Шаги EM-алгоритма

- ▶ Вход: многомерная выборка $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$
- ▶ Инициализация параметров кластеров θ алгоритмом k средних
- ▶ Применение EM-алгоритма:
 - ▶ Е-шаг: расчет вероятностей принадлежности точек к кластеру i :

$$P(i|\mathbf{x}, \theta) = \frac{\omega_i \rho(\mathbf{x}|\mu_i, \Sigma_i)}{\sum_{j=1}^K \omega_j \rho(\mathbf{x}|\mu_j, \Sigma_j)}$$

- ▶ М-шаг: пересчет оценок параметров гауссовских распределений $\theta = (\omega_i, \mu_i, \Sigma_i)$ выборочными оценками
- ▶ Выход: оценка параметров гауссовой смеси θ
- ▶ Ускорение вычислений: $\Sigma_i = \text{diag}(\lambda_1, \dots, \lambda_m)$
- ▶ (Reynolds (2008) показал, что такое ограничение не ухудшает качества детектора при увеличении числа кластеров)

Пример работы ЕМ-алгоритма для построения ГММ

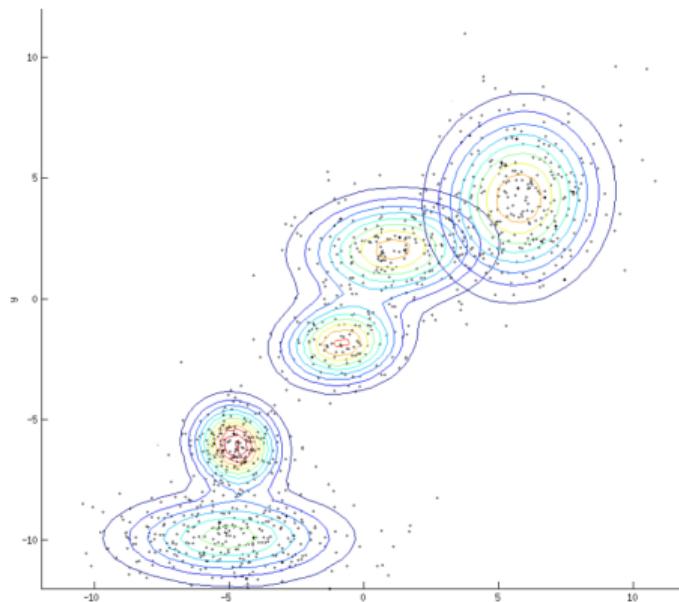


Рис. 27: ЕМ-алгоритм оптимизирует функционал сходится к оценке максимального правдоподобия θ параметров смеси, если выборка взята из взвешенной смеси гауссовских распределений.

Пример: аномальные начертания MNIST

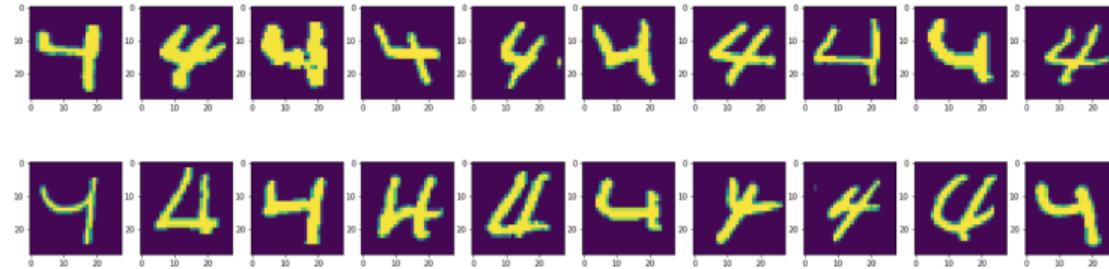


Рис. 28: $t \in \{1 \dots 28\}$ (построчная кластеризация), изображения с наименьшей сглаженной оценкой правдоподобия (1000 примеров, топ-20)

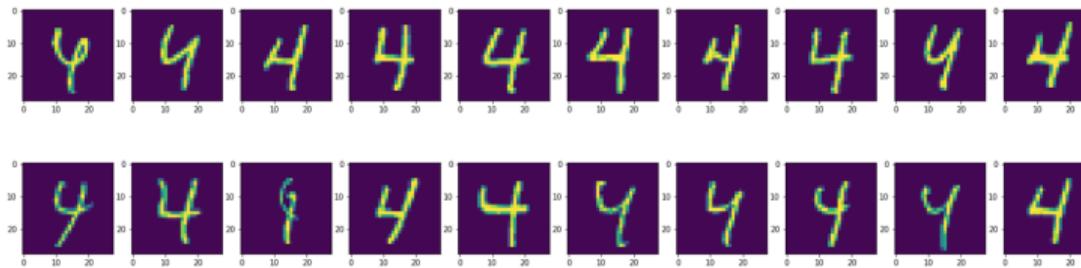


Рис. 29: Топ-20 примеров с максимальным правдоподобием

Пример: аномальные начертания MNIST (связь с t-SNE)

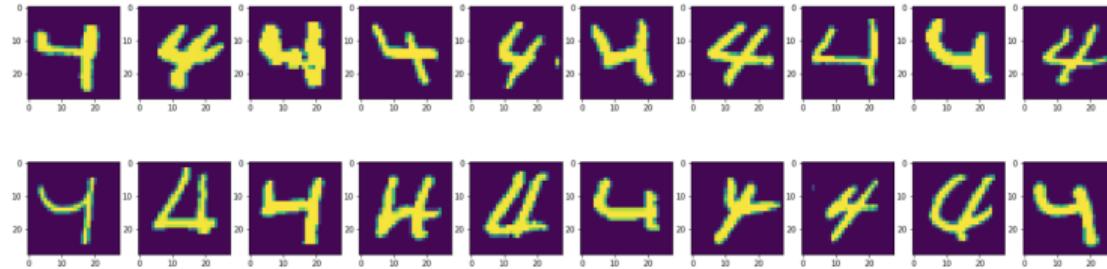


Рис. 30: Топ-20 аномальных начертаний

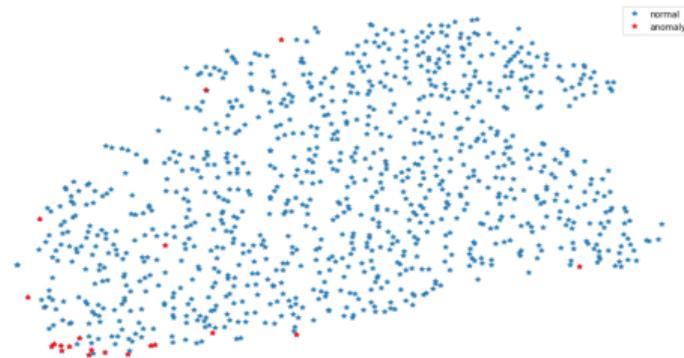


Рис. 31: t-SNE над 28x28 мерными векторами изображений

Пример: MNIST — аномальные строки

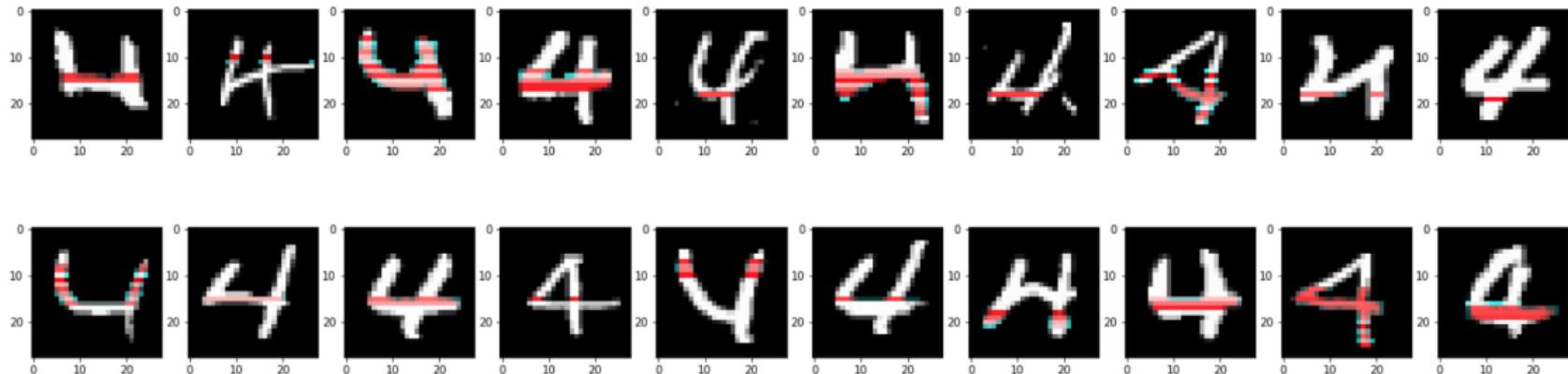


Рис. 32: Тепловая карта аномальности для топ-20 изображений, содержащих самые аномальные строки (с наименьшим правдоподобием)

Обнаружение аномалий на основе стохастических моделей

- ▶ **Преимущества:**
 - ▶ Использование существующих статистических подходов для моделирования различных типов распределений
- ▶ **Недостатки:**
 - ▶ Трудность описания нормального режима в данных высокой размерности
 - ▶ Реальные данные часто не соответствуют модельным параметрическим предположениям о них (гауссовость, факторизация компонент)

Резюме лекции и выводы

- ▶ Аномалии: тестовые данные, отличные от обучающей выборки
- ▶ Доступно множество методов обнаружения аномалий (хотя не для всех легко доступны программные реализации)
- ▶ Простой детектор: «энергия» остатков в методе PCA
- ▶ Вероятностное описание: смеси гауссиан (GMM)
- ▶ Построение нелинейной разделяющей гиперплоскости методом одноклассовой машины опорных векторов (one-class SVM)

Литература |

- Augusteijn, MF и BA Folkert (2002). «Neural network classification and novelty detection». B: *International Journal of Remote Sensing* 23.14, с. 2891—2902.
- Barnett, Vic (1978). «The study of outliers: purpose and model». B: *Applied Statistics*, с. 242—250.
- Breunig, Markus M и др. (2000). «LOF: identifying density-based local outliers». B: *ACM sigmod record*. Т. 29. 2. ACM, с. 93—104.
- Burnaev, Evgeny и Dmitry Smolyakov (2016). «One-class SVM with privileged information and its application to malware detection». B: *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*. IEEE, с. 273—280.
- Chandola, Varun, Arindam Banerjee и Vipin Kumar (2009). «Anomaly detection: A survey». B: *ACM computing surveys (CSUR)* 41.3, с. 15.
- Clifton, Lei и др. (2011). «Identification of patient deterioration in vital-sign data using one-class support vector machines». B: *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*. IEEE, с. 125—131.

Литература II

- Duda, Richard O, Peter E Hart и David G Stork (2012). *Pattern classification*. John Wiley & Sons.
- Dutta, Haimonti и др. (2007). «Distributed top-k outlier detection from astronomy catalogs using the demac system». В: *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM, с. 473—478.
- Hawkins, Douglas M (1980). *Identification of outliers*. Т. 11. Springer.
- Hawkins, Simon и др. (2002). «Outlier detection using replicator neural networks». В: *DaWaK*. Т. 2454. Springer, с. 170—180.
- Hoffmann, Heiko (2007). «Kernel PCA for novelty detection». В: *Pattern Recognition* 40.3, с. 863—874.
- Jyothsna, V, VV Rama Prasad и K Munivara Prasad (2011). «A review of anomaly based intrusion detection systems». В: *International Journal of Computer Applications* 28.7, с. 26—35.
- Manevitz, Larry M и Malik Yousef (2001). «One-class SVMs for document classification». В: *Journal of Machine Learning Research* 2.Dec, с. 139—154.

Литература III

- Markou, Markos и Sameer Singh (2003). «Novelty detection: a review—part 1: statistical approaches». B: *Signal processing* 83.12, c. 2481—2497.
- Matthews, Bryan и др. (2013). «Discovering anomalous aviation safety events using scalable data mining algorithms». B: *Journal of Aerospace Information Systems* 10.10, c. 467—475.
- Miljković, Dubravko (2010). «Review of novelty detection methods». B: *Mipro, 2010 proceedings of the 33rd international convention*. IEEE, c. 593—598.
- Patcha, Animesh и Jung-Min Park (2007). «An overview of anomaly detection techniques: Existing solutions and latest technological trends». B: *Computer networks* 51.12, c. 3448—3470.
- Quinn, John A и Christopher KI Williams (2007). «Known unknowns: Novelty detection in condition monitoring». B: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, c. 1—6.
- Schölkopf, Bernhard, Alexander Smola и Klaus-Robert Müller (1998). «Nonlinear component analysis as a kernel eigenvalue problem». B: *Neural computation* 10.5, c. 1299—1319.

Литература IV

- Schölkopf, Bernhard, Robert C Williamson и др. (2000). «Support vector method for novelty detection». B: *Advances in neural information processing systems*, c. 582—588.
- Shyu, Mei-Ling и др. (2003). *A novel anomaly detection scheme based on principal component classifier*. Tex. отч. MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL и COMPUTER ENGINEERING.
- Srivastava, AN (2006). «Enabling the discovery of recurring anomalies in aerospace problem reports using high-dimensional clustering techniques». B: *Aerospace Conference, 2006 IEEE*. IEEE, 17—pp.
- Srivastava, Ashok N и Brett Zane-Ulman (2005). «Discovering recurring anomalies in text reports regarding complex space systems». B: *Aerospace conference, 2005 IEEE*. IEEE, c. 3853—3862.
- Tarassenko, Lionel и др. (2009). «Novelty detection». B: *Encyclopedia of Structural Health Monitoring*.
- Tax, David MJ и Robert PW Duin (1999). «Support vector domain description». B: *Pattern recognition letters* 20.11, c. 1191—1199.

Литература V

Williams, Graham и др. (2002). «A comparative study of RNN for outlier detection in data mining». В: *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, с. 709—712.