Федеральное государственное бюджетное образовательное учреждение высшего образования

«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА и ГОСУДАРСТВЕННОЙ СЛУЖБЫ при Президенте Российской Федерации»

ИНСТИТУТ ЭКОНОМИКИ, МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

НАПРАВЛЕНИЕ 38.03.01 ЭКОНОМИКА

Группа ОБ-7281-20 Кафедра микроэкономики

Допустить к защите

заведующий кафедрой микроэкономики

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Левин М.И.

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_ г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ОЦЕНКА КРИВОЙ ФИЛЛИПСА ДЛЯ РОССИЙСКОЙ ЭКОНОМИКИ

студент-бакалавр

Анфимов Александр Дмитриевич /\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_\_/

*(подпись) (дата)*

научный руководитель выпускной

квалификационной работы

к.э.н., Девятов Алексей Евгеньевич /\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_\_/

*(подпись) (дата)*

МОСКВА

2024 г.

**Оглавление**

[Введение 3](#_Toc220059700)

[Основная часть 5](#_Toc220059701)

[1. Обзор литературы 5](#_Toc220059702)

[1.1. Современное состояние области исследования непараметрических методов 5](#_Toc220059703)

[1.2. Текущие результаты из области оценивания непараметрической регрессии 8](#_Toc220059704)

[1.3. Перспективные походы оценивания непараметрической регрессии 10](#_Toc220059705)

[2. Постановка задачи, генерация и сбор данных 15](#_Toc220059706)

[2.1. Формальная постановка задачи оценки одномерной регрессии 15](#_Toc220059707)

[2.2. Общее описание датасетов 15](#_Toc220059708)

[3. Обзор используемых методов 21](#_Toc220059709)

[3.1. Ядерное сглаживание 21](#_Toc220059710)

[3.2. Информационный подход 28](#_Toc220059711)

[ПРИЛОЖЕНИЕ 2 30](#_Toc220059712)

# Введение

Непараметрическая регрессия – это метод моделирования зависимости между переменными, который не предполагает явно заданной формы функции (линейной, полиномиальной и т.д.). Основное преимущество такого подхода заключается в гибкости, то есть адаптации к структуре без жёстких допущений. Именно поэтому данный метод является важным и актуальным в современных реалиях. В эпоху больших данных, когда наблюдается рост сложности данных (нелинейность, выбросы и мультимодальность) он позволяет анализировать сложные и неизвестные зависимости, которые практически невозможно оценить с помощью стандартных параметрических методов. Особенно широкое применение данный метод находит в машинном обучении и таких прикладных областях как экономика и медицина. Более подробно актуальные направления рассмотрены в соответствующей главе исследования.

В качестве основных подходов к реализации данного метода можно выделить как классические непараметрические методы (ядерные регрессии, локально полиномиальные оценки и др.) так и байесовские (регрессия гауссовских процессов, байесовские аддитивные регрессионные деревья). Более подробно существующие методы, а также их преимущества и недостатки, разбираются в обзорной главе данной работы.

Стоит отметить пробел в исследованиях, на котором фокусируется данная работа, а именно отсутствие работ, которые прямо анализируют информационную эффективность существующих подходов к оцениванию непараметрической регрессии. В большинстве исследований акцент идёт на предсказательной способности или уровня подгонки модели, а также качестве извлекаемых знаний из выборки. Однако стандартные используемые метрики, такие как и , не отражают полноту извлекаемых данных.

Таким образом, данное исследование направлено на то, чтобы заполнить этот пробел, предлагая другой, теоретико-информационный подход, к оценке эффективности построения непараметрической регрессии. Цель данной работы заключается в сравнении существующих классических непараметрических и байесовских методов с точки зрения количества извлекаемой информации из данных и построении сравнительной шкалы на основе нескольких метрик. Практическая значимость заключается в том, что в работе также описаны рекомендации и особенности выбора методов в зависимости от типа задач и данных.

Структура работы представлена следующим образом. Первая часть работы посвящена обзору актуальных направлений применения и исследования непараметрической регрессии. Во второй части работы приводится обзор и описание существующих подходов к оценке непараметрической регрессии. В третьей части описываются модификации выбранных для сравнения моделей, а также информационные критерии. В четвёртой части описаны данные. В пятой представлены результаты оценки и сравнения. В шестой части приводится заключение исследования.

# Основная часть

## 1. Обзор литературы

### 1.1. Современное состояние области исследования непараметрических методов

Непараметрическая регрессия позволяет моделировать зависимость без предположений о форме зависимости. Говоря более глобально, непараметрические методы в целом не предполагают конкретной формы распределения – они опираются на данные для оценки сложных зависимостей. На текущий момент существует множество различных непараметрических подходов, некоторые методы из которых описаны в работах Анатольева С.[[1]](#footnote-1) и Matias S.[[2]](#footnote-2). Однако всё же можно выделить три наиболее популярных подхода: ядерные методы (kernel), метод анализа многообразия (manifold learning) и метод перебора (копульного моделирования). Разберем каждый из них немного более подробно.

#### Ядерное сглаживание

Ядерное сглаживание (Kernel Density Estimation, KDE) – это непараметрический метод оценки функции распределения вероятности некой случайной величины с помощью сглаживания на основе выборки. Другими словами, в каждой точке вычисляется сглаженная функция плотности с помощью весов, определяемых ядром и полосы пропускания (окна) по формуле, также называемой ядерной оценкой плотности Розенблата — Парзена[[3]](#footnote-3)[[4]](#footnote-4):

(1)

где:

* элементы выборки,
* ядерная функция,
* полоса пропускания (окно сглаживания).

Ядро — это непрерывная ограниченная симметричная вещественная функция K с единичным интегралом. Каждое ядро имеет свои преимущества и недостатки, однако более подробно особенности метода будут описаны в отдельной главе. В рамках данного подхода также присутствует несколько ответвлений, которые фокусируются на том или ином аспекте метода.

В первую очередь стоит отметить подход энтропийного оценивания. Данный метод имеет название ME-KDE (minimum entropy-based KDE) и хорошо описан в работе Jie Jiang и др.[[5]](#footnote-5). В основном он используется для сравнения моделей и выбора гиперпараметра h по принципу максимума энтропии, что снижает смещение оценки. То есть выбирается модель, максимизирующая функцию

(2)

Второй подход заключается в измерении ошибки между истинной плотностью и оценённой по нормы L1 и L2 соответственно:

* L1-ошибка: более устойчивая к выбросам, но не так хорошо улавливает локальные пики;
* L2-ошибка: чувствительная к выбросам, но лучше улавливает локальные пики.

На практике довольно часто в качестве ошибки используют mean integrated squared error (MISE) (3). Однако, так как в реальных задачах истинная плотность неизвестна, используется ассимптотическая оценка (AMISE), которая в свою очередь применяется через бутстрэппинг[[6]](#footnote-6), кросс-валидацию[[7]](#footnote-7) или эмпирическое правило Сильвермана[[8]](#footnote-8).

(3)

Третий подход к оцениванию – метод моментов. Как и другие методы, он используется для подбора гиперпараметров ядра, в частности полосы пропускания h, но основывается на согласовании теоретических и выборочных моментов исследуемой случайно величины. Другими словами, подбирается такое h, что моменты близки к выборочным. Примером использования этого метода может послужить работа Bradley W. и др.[[9]](#footnote-9).

Наконец, четвертый подход – метод инвариантов (кумулянтов) – заключается в том, чтобы использовать инварианты для анализа формы плотности, построения уточнённых оценок плотности через разложение в ряд Эджворта или Грама-Шарлье и, конечно, оценки полосы пропускания. Примером реального применения метода может послужить работа Yanhong Luo, Xu Wang и Shijie Yan[[10]](#footnote-10).

#### Метод анализ многообразия (manifold learning)

В данном подходе предполагается, что данные на самом деле лежат на скрытом многообразии меньшей размерности, вложенном в анализируемое высокоразмерное пространство. На основе этого предположения методы анализа многообразия понижают размерность данных и обнаруживают нелинейные зависимости, в то же время сохраняя геометрические свойства данных как локально, так и глобально. Некоторые методы данного подхода также связаны с ядерный трюком (kernel trick), когда связь переменных исследуется в другом пространстве с помощью скалярного произведения из этого пространства. Непараметрическим метод анализа многообразия считается так как не предполагает фиксированной параметрической формы многообразия. Другими словами, для анализа в рамках данного подхода не требуется знание уравнения многообразия – структура выводится из данных с помощью локальных статистик.

Методы в рамках данного подхода применяются в таких задачах как визуализация высокоразмерных данных, подготовка данных для машинного обучения, анализ топологии данных, улучшение глубокого обучения, обработка естественного языка (NLP). Таким образом, данный подход слабо применим в рамках данной работы, поэтому мы не будем останавливаться на нём более подробно.

#### Метод перебора (копульного моделирования)

Копула – это многомерная функция распределения, определённая на n-мерном единичном кубе , такая, что каждое её частное распределение равномерно. Она выступает инструментом для исследования взаимосвязи между переменными, разделяя структурную зависимость и маргинальное распределение. Сам метод перебора заключается в том, что эмпирическая копула приближается существующими параметрическим копулами (Гаусса, Стьюдента, Архимедовыми) путём перебора. Модели именно приближают копулу, так как в реальных данных связь обычно сложна и не описывается полностью стандартными копулами. Несмотря на использование параметрических копул, данных подход является непараметрическим, так как никаких априорных предположений о форме зависимости нет, и мы просто подбираем эмпирически наиболее подходящее приближение. Отдельно стоит отметить, что так как истинная копула неизвестна для оценки ошибки используются методы, схожие с теми, которые обсуждались при оценке плотности в KDE, например оценка на основе MISE. Аналогично, методам многообразия данных подход сложно применить в рамках построения непраметрической регрессии, поэтому он не будет рассматриваться более подробно.

### 1.2. Текущие результаты из области оценивания непараметрической регрессии

Непараметрическая регрессия играет важную роль в современной статистике и машинном обучении. Она часто применяется в задачах, где требуется анализировать сложные, малоизученные и нелинейные данные, так как не предполагает явно заданной функциональной формы зависимости.

В последние годы наблюдается рост вычислительных мощностей и развитие больших данных. Особенно сильно это заметно в таких областях, как машинное обучение, искусственный интеллект, медицина, экономика и финансы, поэтому непараметрическая регрессия часто применяется для решения задач из этих областей, о чём и пойдёт речь в этой главе.

В качестве примера работы в области медицины можно выделить статью Альберта Канона, Гианлуки Байо и Иоанна Манолопполоу[[11]](#footnote-11). В ней авторы рассматривают современные подходы к оценке гетерогенных и индивидуальных эффектов лечения (CATE/ITE) с использованием методов непараметрической регрессии. Такой подход позволяет убрать или уменьшить необходимость в традиционных рандомизированных экспериментах, что особенно актуально в сферах, где они невозможны. В статье используется два основных непараметрических подхода: Bayesian Additive Regression Trees (BART) , а именно Bayesian Causal Forest (BCF) и мета-алгоритмы (S-, T-, X-, R-Learners). Эти методы позволили исследователям без жёстких предположений о функциональной зависимости оценить сложную связь между лечением и ковариатами. Авторы отмечают, что гибкий подход был критичным для минимизации систематических ошибок в оценках причинных эффектов и подчерчивают преимущество такого подхода перед параметрическими методами в контексте высокой размерности данных, несбалансированных групп и смешивающих факторов. Результативность методов демонстрировалась на двух датасете. Сначала исследователи провели анализ влияния школьных обеденных программ на здоровье детей, а потом показали высокую эффективность методов на синтетических данных. Однако авторы указывают на ряд важных ключевых допущений, соблюдение которых необходимо для эффективного применения предложенных методов, а именно отсутствие скрытых смешивателей и перекрытие групп.

В области экономики непараметрическая регрессия также является относительно популярным методов, в основном, потому что изучаемые зависимости неизвестны и довольно сложны. Хорошим примером этому может послужить работа Мингминг Джанга и др.[[12]](#footnote-12). В статье исследуется влияние инвестиций в возобновляемую энергетику на выбросы в Китае. Зависимость исследуется с помощью непараметрической аддитивной регрессионной модели STIRPAT, позволяющей учитывать как линейные, так и нелинейные взаимосвязи. Таким образом, исследователи могут учесть как известную информацию о функциональной зависимости через линейные коэффициенты (влияние таких стандартных факторов, как ВВП, уровень индустриализации и тд.), так и неизвестную изучаемую часть (в частности, влияние инвестиций) с помощью непараметрической регрессии. Проведённый анализ показал, что инвестиции оказываю сложное влияние на выбросы: сначала увеличивают их из-за инфраструктурных затрат и низкой эффективности, потом уменьшают благодаря замещению ископаемого топлива, однако снова начинают увеличивать выбросы при дальнейшем росте из-за экологических проблем инфраструктуры и компенсации углём. Также были обнаружены нелинейные зависимости для экономического роста (монотонный рост выбросов) и урбанизации (обратная U-кривая). В заключение авторы работы подчёркивают важность применения непараметрической регрессии в контексте исследования скрытых паттернов в экономике, которые сложно измерить параметрическими методами, а также предлагают практические рекомендации к повышению эффективности “зелёных” инвестиций в Китае.

Другим примером эффективного применения непараметрической регрессии в экономике может послужить работа Тингтинг Ченга и др.[[13]](#footnote-13). В ней авторы прогнозируют доходность акций с помощью непараметрической регрессии и других моделей. Для этого в исследовании была разработана многопериодная модель, которая учитывает локально стационарные предикторы (например, соотношении цены и доходности, дивидендную доходность) и оценивает регрессию с помощью ядерных методов сглаживания. По результатам исследования непараметрическая регрессия и её модификации (МА1 и МА2) показали качество лучшее по сравнению с линейными моделями, историческим средним уровнем и методами машинного обучения. Особенно сильно видно превосходство модели в долгосрочной перспективе (3–12 месяцев) и во времена кризисов и рыночной нестабильности (например, кризис 2009 года). Авторы указывают на то, что модель улавливает сложные паттерны, включая нелинейные связи между предикторами и доходностью, а также адаптируется к изменяющимся условиям через временную локализацию. Более того, инвестиционная стратегия построенная на модифицированной непараметрической регрессии МА2 показала высокую доходность (до 16% годовых) при низком риске. Таким образом, исследователи делают вывод, что непараметрическая регрессия является мощным и актуальным инструментом для инвестиционных стратегий и анализа рынка акций за счёт гибкости и устойчивости к структурным сдвигам в данных, так как финансовая область весьма волатильна.

Основным выводом из данных работ является факт того, что непараметрическая регрессия на сегодняшний день является важным и мощным инструментом во многих прикладных задачах, требующих гибкости и адаптивности модели.

### 1.3. Перспективные походы оценивания непараметрической регрессии

Большинство современных методов построения непараметрической регрессии связаны с машинным обучением, поэтому множество исследований прямо или косвенно посвящены именно этой теме. Однако машинное обучение и нейросети применяются в работах по-разному.

Одним из ключевых направлений исследования являются проблемы построения непараметрической регрессии в высокоразмерных пространствах. Здесь можно выделить работу Елены Брадик и др.[[14]](#footnote-14), в которой авторы изучают современные методы в описанных выше условиях, фокусируясь на корневой n-согласованности оценок линейных функционалов (например, регрессионного наклона или среднего эффекта). В статье авторы модифицируют подходы, использующие двойное (debiased) машинное обучение и разреженные аппроксимации. Таким образом, предложенный подход позволил им оценить параметры при отсутствии знаний о структуре данных. В качестве результатов исследования авторы выделяют установление условий разреженности и гладкости , при которых достижима скорость , и построение устойчивых к гетероскедастичности оценок через регуляризованные представители Рисса. В завершение авторы подчёркивают преимущество подобного подхода по сравнению с такими методами как debiased Lasso и series estimation, а именно высокая эффективность в условиях слабых предположениях о данных. Таким образом, данная работа обобщает непараметрические идеи и показывает, как с помощью машинного обучения можно улучшить качество непараметрической регрессии в высокоразмерных пространствах, расширяя применимость метода в экономике и других сферах.

Другой подход применяется в работе Юпинг Сонга и др.[[15]](#footnote-15). В своей статье авторы фокусируются на интеграции современных методов глубокого машинного обучения с классической ядерной непараметрической регрессией. Качество моделей изучается в контексте прогнозирования высокочастотных финансовых данных, которые сложно предсказывать из-за нелинейности, высокой волатильности, долговременной памяти и частых скачков. Авторы подмечают, что обычные подходы плохо справляются с этой задачей: традиционные эконометрические модели (ARIMA, GARCH) ограничены при работе с подобными данными из-за предположений о линейности и стационарности; машинное обучение (SVM, нейросети) справляются с нелинейностью, однако не улавливают долговременную память; глубокие нейросети хорошо прогнозируют, но не интерпретируемы. В работе предлагается гибридная модель NR-LSTM, объединяющая интерпретируемость непараметрической регрессии со способностью глубокой нейросети улавливать долговременные эффекты и краткосрочные аномалии. В статье непараметрическая регрессия используется для моделирования тренда, в то время как глубокая нейросеть корректирует остатки, повышая общую точность прогноза. Полученная модель эмпирически проверялась на разнообразных финансовых данных (CSI 300, FTSE 100, S&P 500) и показала значительное улучшение по сравнению с другими подходами по метрике средней абсолютной ошибки (MAE), демонстрируя устойчивость и универсальность. Таким образом, данная работа подчеркивает перспективность гибридных методов, так как они позволяют использовать сильные стороны как традиционных статистических подходов, так и современных алгоритмов ИИ для решения задач в условиях больших данных и неопределённости.

Очень важной в данном контексте является статья Minshuo Chen и др.[[16]](#footnote-16). В ней авторы проводят теоретический анализ построения непараметрической регрессии на низкоразмерных многообразиях с помощью глубоких нейросетей с функцией активации ReLU. В точной постановке задачи исследуется восстановление функции , заданной на мерном римановом многообразии , вложенным в пространство , по зашумлённым данным , где – независимый субгауссовский шум. В работе авторы доказывают, что ReLU-сети могут эффективно аппроксимировать любую гладкую или гёльдерову функцию на многообразии с нужной погрешностью. Однако самое главное, что полученная скорость сходимости для такого метода построения непараметрической регрессии зависит от внутренней размерности d, то есть сложности самой зависимости, а не от размерности внешнего пространства D. В заключение, авторы продемонстрировали на примере задачи обработки картинок из датасета ImageNet, что классическими подходами практически невозможно получить высокую точность из-за проклятья размерности, в том время как предложенный метод вполне осуществим. Работа развивает идеи исследований Schmidt-Hieber (2019) и Nakada & Imaizumi (2020) об адаптации нейросетей к свойствам данных и демонстрирует преимущество подходов на основе глубоких нейросетей перед классическими подходами в задачах с высокоразмерными, но структурированными данными, где моделируемые зависимости гораздо проще пространства, в котором они находятся.

Несмотря на популярность нейросететевого подхода также остаются модели, фокусирующиеся на модификации классических подходов. Особенно актуальной в контексте построения непараметрической регрессии является статья Taha Hussein Ali и др.[[17]](#footnote-17). В своей работе авторы представляют новый метод оценки параметра ширины окна (полосы пропускания) в непараметрической регрессии Надарая-Ватсона. В то время как классические подходы предлагают использовать фиксированнную или переменную (локально оптимальную) ширину окна, авторы предлагают улучшенный подход, основанный на универсальном пороговом уровне, вычисленном с помощью вейвлета Добеши. Для оценки шума используется медиана абсолютных отклонений (MAD), а порог определяется как , где – количество вейвлет коэффициентов. Предложенный подход авторы сравнивали с методами, которые основаны на геометрическом среднем, арифметическом среднем, размахе и медиане, на синтетических данных (размеры выборок – 32, 64 и 128) и реальных (извержения гейзера Old Faithful). По результатам оценки улучшенный подход показал наименьшую MSE по сравнению с классическими методами при всех объемах выборки и значениях начальной полосы пропускания на всех датасетах. Таким образом, данный метод является очень перспективным, особенно на маленьких выборках – он адаптивно настраивает ширину окна в регрессии Надарая-Ватсона, тем самым улучшая прогнозную способность модели. Более того, авторы подчеркивают, что использование других вейвлетов и пороговых стратегий (например, кросс-валидации) может ещё улучшить метод под конкретную задачу.

Другим примером современного исследования классических методов является работа Xin Bing, Xin He, Chao Wang[[18]](#footnote-18). В ней авторы рассматривают ядерную гребневую непараметрическую регрессию (KRR) и доказывают её эффективность в условиях, когда исходные данные зашумлены. Эта задача особенно актуально в текущем ландшафте высокоразмерных и зашумлённых данных, когда в моделях используются не сами признаки, а скрытые факторы из них, например полученные методом опорных векторов. В статье авторы предполагают, что данные X связаны со скрытыми факторами через функцию g (например, ). Вместо истинных факторов используется их оценка , из данных , полученная с помощью ядерной регрессии. Далее авторы исследуют ошибку предсказания при таком подходе. В результате в работе доказывается теоретическая гарантия точности предложенного метода и показывается, что он достигает оптимальной скорости сходимости даже при наличии ошибок в данных. Таким образом, данное исследование подтверждает устойчивость KRR к высокой размерности и шуму в данных, тем самым показывая её преимущество перед другими методами, которые сталкиваются с проклятьем размерности и ограничениями на наблюдаемые признаки.

В заключение главы хочется отметить, что тема данного диплома не рассматривалась в полной мере в исследованиях за последние годы. Таким образом, это ещё раз подчеркивает научную ценность и актуальность исследования этой области.

## Постановка задачи, генерация и сбор данных

### Формальная постановка задачи оценки одномерной регрессии

Цель задачи – восстановить неизвестную функциональную зависимость между одной независимой переменной (регрессор) и зависимой переменной на основе конечной выборки наблюдений, чтобы либо минимизировать ошибку предсказаний для новых значений и/или интерпретировать структуру связи и .

Пусть имеется обучающая выборка данных , где – наблюдения регрессора, наблюдения зависимой переменной, размер выборки. Также предполагается следующая структурная модель:

где истинная функция регрессии (неизвестна, предполагается полиномиальная зависимость), случайная ошибка с существующими конечными моментами первого и второго порядков, такие что .

Требуется найти оценку функции регрессии , минимизирующую заданную функцию потерь на обучающей выборке и обобщающую на новые . Функция потерь зависит от выбора метода и подхода в рамках него и будет описана в соответствующих разделах обзоров выбранных методов.

### Общее описание датасетов

Для оценки качества методов, а также их сравнения в данной работе используется ряд синтетических наборов данных, каждый из которых содержит наблюдений. Для получения достоверных оценок качества методы оцениваются несколько раз на одном типе датасета, и на каждой итерации оценивания датасет нужного типа генерируется заново. Реплицируемость результатов контролируется с помощью seed – они указаны в соответствующих разделах практической части. Генерация происходит в несколько этапов:

1. Выбирается 30 точек, равномерно расположенных на отрезке от 0 до 1;
2. Для каждой точки рассчитывается значение истинной функции регрессии на основе одной из моделей генерации. Модели генерации представлены в разделе 2.2.3;
3. К полученному значению прибавляется сгенерированная ошибка, если это ещё не сделано в модели. Генерация ошибок обсуждается в разделе 2.2.2.

Исключением является модель генерации, основанная на смеси нормальных распределений (см. раздел 2.2.3).

Кроме того, для оценки качества работы моделей на реальных данных используется известный датасет, содержащий извержения гейзера Старый Служака. Датасет представлен в виде фрейма данных с 272 наблюдениями по времени извержения в минутах и времени ожидания следующего извержения в минутах. Он часто используется в исследованиях, посвящённых оценке непараметрической регрессии[[19]](#footnote-19). На рис. 1 видно, что в среднем есть положительная зависимость между переменными. Кроме того, видно, что извержения можно условно разделить на два типа – короткие с маленьким временем ожидания и длинные с большим временем ожидания.

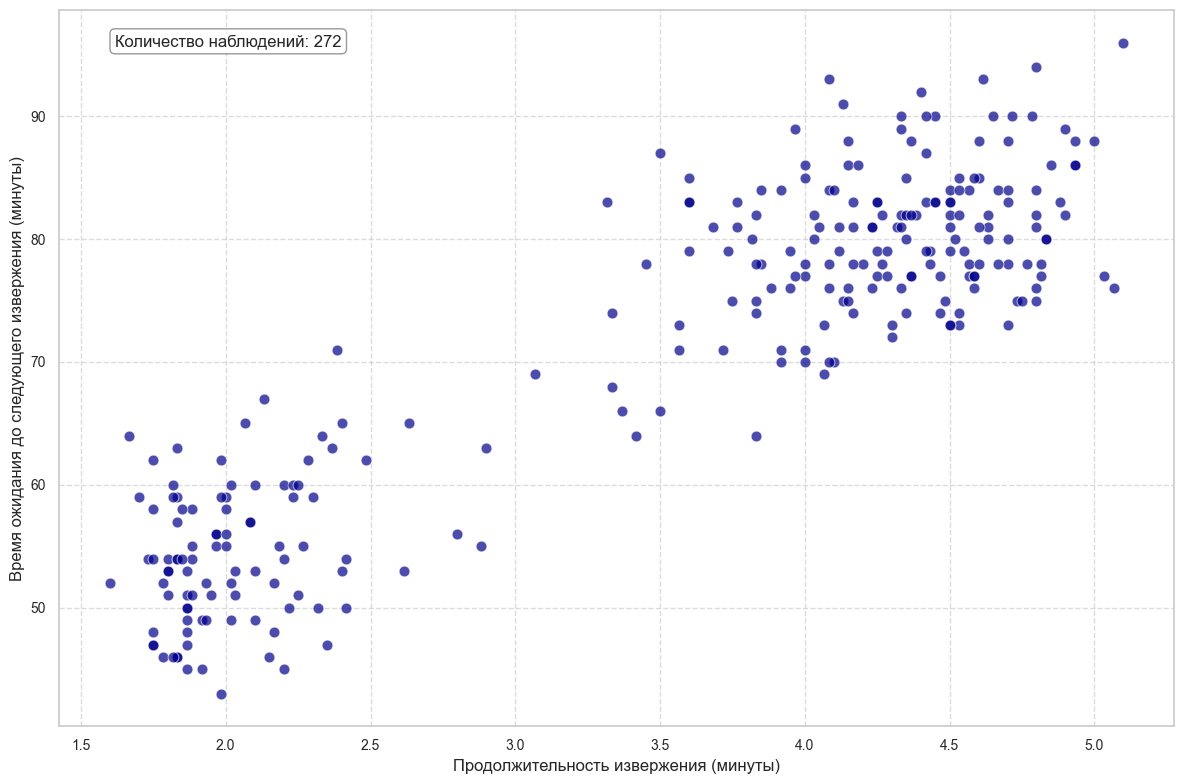


Рис. 1. Зависимость между продолжительностью извержения и временем ожидания, Гейзер "Старый Служака" (Озеро Йеллоустон). Источник: работа автора.

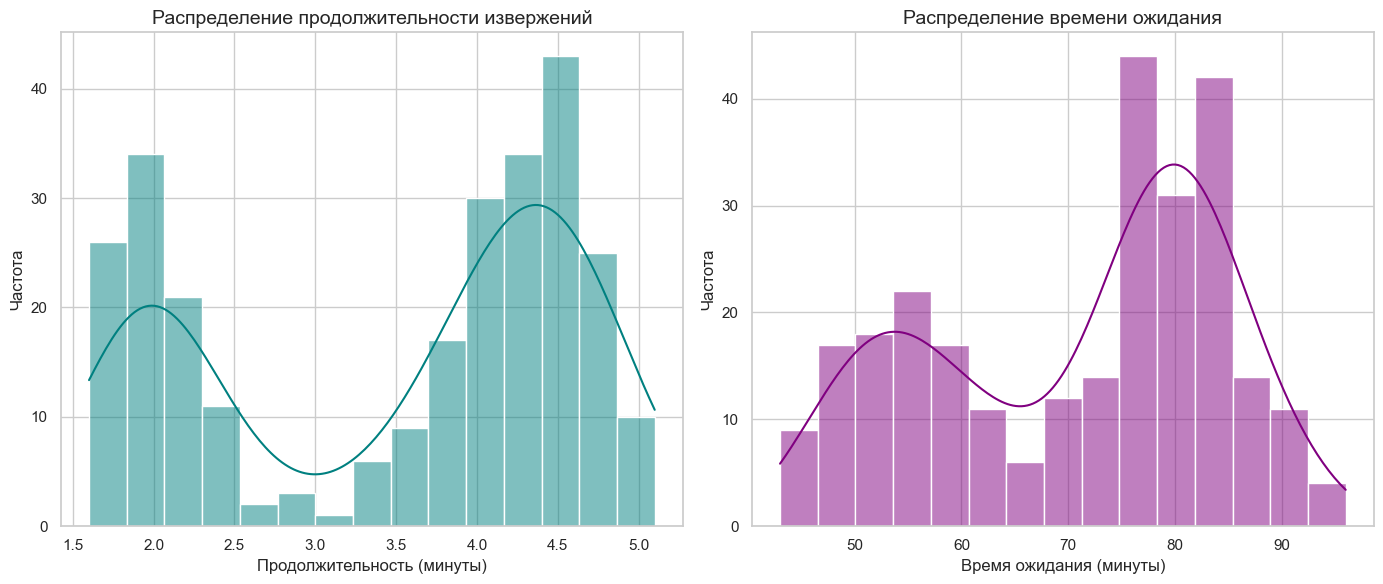


Рис. 2. Распределение данных по извержению гейзера. Источник: работа автора.

#### Обоснование размера выборки

Столь маленькое количество наблюдений () в наборе обусловлено рядом причин. Во-первых, в рамках данного исследования мы фокусируемся на оценке методов в сложных условиях, когда данных мало и ошибки распределены по разным законам. Именно в таких условиях используемые подходы имеют теоретическое преимущество по сравнению с классическими методами вроде МНК, которые показывают более низкую эффективность из-за нарушения предположений о нормальности ошибок. В частности, предлагаемый в данной работе метод энтропийного оценивания с жесткими балансовыми ограничениями изначально разработан для работы с малыми выборками и асимметричными/тяжелохвостыми распределениями ошибок, что делает его робастным в экстремальных сценариях. Таким образом, можно не только протестировать его в подходящих условиях, но и смоделировать реальные задачи, где сбор больших объёмов данных затруднён, например, исследование редких экономических событий.

Во-вторых, выбран как компромисс вычислительной сложностью и статистической осуществимостью. С одной стороны, этот размер достаточен для применения базовых методов кросс-валидации и оценки дисперсии коэффициентов[[20]](#footnote-20). С другой стороны, это граница применимости ряда важных асимптотических приближений, таких как ЦПТ[[21]](#footnote-21). Такой подход позволяет нам оценить модели в “неудобных” для классических методов условиях, о чем упоминалось выше.

В-третьих, основной метод, предлагаемый в данной статье, а именно энтропийное оценивание с жесткими балансовыми ограничениями, вычислительно требователен, так как требует решения системы уравнений размерности (см. главу ВСТАВИТЬ ГЛАВУ). Поэтому с точки зрения временных затрат использование больших наборов данных было бы непрактичным.

#### Генерация ошибок

Для каждой точки независимо генерируется ошибка. Ошибки берутся из разных распределений, но параметры подбираются так, чтобы матожидание ошибки было нулевым, а дисперсия единичной, т.е. . Таким образом, для 30 точек в каждом датасете генерируется соответственно 30 ошибок из разных распределений. Ошибки берутся в случайном порядке. Используются следующие распределения:

1. **Группа 1: Классические симметричные распределения**
   * Нормальное — базовое распределение для сравнения;
   * Логистическое — похоже на нормальное, но с более тяжелыми хвостами;
   * Распределение Лапласа — остроконечное с экспоненциальными хвостами;
   * Гиперболический секанс — интересная форма с более легкими хвостами, чем у нормального;
   * Обобщенное нормальное (beta=0.5) — очень остроконечное распределение.
2. **Группа 2: Распределения с тяжелыми хвостами**
   * Распределение Стьюдента (df=3) — очень тяжелые хвосты;
   * Распределение Стьюдента (df=30) — почти нормальное, но с чуть более тяжелыми хвостами;
   * Двойное Вейбулла (c=1.5) — асимметричное с тяжелыми хвостами;
   * Двойное Вейбулла (c=3.5) — почти симметричное с легкими хвостами;
   * Устойчивое (alpha=1.8) — распределение с бесконечной дисперсией при alpha≤2, но для alpha=1.8 можно настроить конечную дисперсию.
3. **Группа 3: Распределения экстремальных значений**
   * Гумбель для максимумов — распределение максимумов;
   * Гумбель для минимумов — распределение минимумов;
   * Обобщенное экстремальное (c=0.5) — гибкое распределение для экстремумов;
   * Обобщенное экстремальное (c=-0.5) — другая форма для экстремумов;
   * Распределение Мояла — асимметричное распределение для энергетических измерений.
4. **Группа 4: Скошенные (асимметричные) распределения**
   * Скошенное нормальное (a=5) — сильная правая асимметрия;
   * Скошенное нормальное (a=-5) — сильная левая асимметрия;
   * Нецентральное t (df=10, nc=3) — асимметричное распределение из статистики;
   * Экспоненциально модифицированное нормальное (K=1) — правая асимметрия;
   * Экспоненциально модифицированное нормальное (K=2) — более сильная правая асимметрия.
5. **Группа 5: Специальные распределения**
   * Распределение Джонсона SU (a=2, b=1) — гибкое семейство распределений;
   * Распределение Джонсона SU (a=0.5, b=2) — другая конфигурация для разнообразия;
   * Обобщенное логистическое (c=0.5) — похоже на логистическое, но более гибкое;
   * Обобщенное логистическое (c=2.0) — другая форма обобщенного логистического;
   * Распределение Тьюки (lambda=0.5) — гибкое распределение, переходящее между различными формами;
   * Распределение Тьюки (lambda=-0.5) — другая конфигурация для асимметрии.
6. **Группа 6: Менее известные, но интересные распределения**
   * R-распределение (c=0.5) — семейство распределений с разной степенью эксцесса;
   * R-распределение (c=2.0) — другая конфигурация с более легкими хвостами;
   * Степенное нормальное (c=0.3) — преобразованное нормальное распределение;
   * Степенное нормальное (c=0.8) — другая степень преобразования.

Все распределения соответствуют следующим требованиям:

* Определены на всей числовой оси ;
* Параметры подобраны так, что ;
* Достаточно сильно различаются друг от друга;
* Реализованы в одной библиотеке и контролируются одним seed (для удобства и реплицируемости).

Более подробно описание используемых распределений и их параметров представлено в Приложении 1.

#### Описание моделей генерации

* + - 1. **Полиномиальная регрессия**

Для генерации зависимости в данных используются полиномы до 6 степени включительно, т.е.

где степень полинома.

Фактический процесс генерации для каждого коэффициента выглядит следующим образом:

1. Генерируется нормальная величина ;
2. Она умножается на фактор масштаба и фактор смещения ;
3. Полученное число случайно умножается либо на 1, либо на -1.

Так как масштаб ошибок не меняется, то факторы и по сути определяют влияние ошибок. В рамках данного исследования они контролируемо подбираются по метрике относительного уровня шума:

где стандартное отклонение ошибок, стандартное отклонение истинных точек регрессии.

Эта метрика широко используется при генерации данных для непараметрической регрессии в научной литературе[[22]](#footnote-22) [[23]](#footnote-23) [[24]](#footnote-24). В качестве умеренного уровня обычно[[25]](#footnote-25) выбирается из диапазона от 0.2 до 0.4. Относительный уровень шума взаимно-однозначно конвертируется в исходя из определений по формуле:

Для сравнения качества оценок методов в разных условиях генерируется выборки со следующими:

* шум почти не мешает восстановлению структуры;
* баланс между структурой и шумом;
* шум существенно затрудняет восстановление

Таким образом общий процесс генерации для выбранного диапазона и степени выглядит следующим образом:

1. Генерируются ошибки как описано в главе 2.2.2.;
2. На вход идёт заранее определённое ;
3. Выбираются факторы и , описанные выше. Для того, чтобы форма полиномов дополнительно различалась, факторы генерируются случайно как

где случайная величина, равномерно распределённая на отрезке . Случайность контролируется seed;

1. Наконец, чтобы достичь требуемого относительного уровня шума, значения функции дополнительно корректируются по формуле:

где изначально сгенерированные точки, выборочное стандартное отклонение сгенерированных ошибок, выбранный для данной генерации уровень относительного шума, выборочное стандартное отклонение . Суть преобразования заключается в домножении всех коэффициентов полинома на нужное число, чтобы в точности достичь требуемой дисперсии.

Всего генерируется 300 датасетов каждого вида, т.е. для каждой комбинации степени и уровня шума генерируется 300 выборок.

* + - 1. **Двумерная смесь нормальных распределений**

Также исследуется модель двухкомпонентной смеси двумерных нормальных величин. В этом случае двумерный вектор имеет распределение смеси:

где вес первого компонента, вектор средних для компонента k, ковариационная матрица для компонента k. В данной работе взяты следующие параметры компонент для генерации: , , , .

В Приложении 2 представлен вывод, показывающий, что условное матожидание можно представить в виде регрессии (4). Таким образом, данная модель генерации подходит для оценки качества и сравнения моделей в рамках поставленной задачи. Кроме того, ошибки в данном случае будут удовлетворять нашим требованиям – они не будут одинаково распределены и будут иметь при фиксированном сложную зависимую структуру внутри одного наблюдения.

Для этой модели генерация отличается и, кроме того, дополнительно разделена на два случая:

1. Детерминировано генерируется 24 точки из первого распределения и 6 из второго, после чего они случайно перемешиваются. Этот подход используется, чтобы оценить базовую погрешность метода, изолировав её от шума в данных.
2. Сначала случайно согласно бернуллиевскому распределению с выбирается компонента, после чего из соответствующего распределения генерируется точка. В этом случае имитируются более реалистичные условия и исследуется робастность метода к вариациям в данных.

#### Критерии оценки качества

В данной главе приведены выбранные в рамках исследования метрики оценки качества полученных регрессий, а также описан общий процесс их оценки.

* + 1. **Для синтетических датасетов**

Особенностью синтетических данных является то, что известна истинная функция регрессии. Это позволяет довольно точно оценить фактическое качество оценки регрессии. Таким образом, главной метрикой в данном случае выбрана интегральная среднеквадратичная ошибка – стандартный выбор для количественного оценки близости оценённой функции к истинной :

где для полиномиальных регрессий и для смесей. Выбранный интервал для исследуемых смесей покрывает приблизительно 95% вероятностной массы смеси распределений, так как

где функция Лапласа.

Кроме того, используются две дополнительные метрики:

Интегральная средняя абсолютная ошибка дополняет , так как она более робастна к экстремальным ошибкам в выборках и показывает типичное отклонение в интерпретируемом виде. Максимальная абсолютная ошибка показывает, насколько велики ошибки в критических зонах, где метод терпит неудачу.

Интегральные ошибки вычисляются на фиксированном интервале методом адаптивной квадратуры (алгоритм QUADPACK[[26]](#footnote-26)), реализованным в SciPy (SciPy[[27]](#footnote-27) quad) вычисляется методом ограниченной оптимизации (scipy.optimize.minimize\_scalar с методом 'bounded'). Все метрики считаются с абсолютной точностью с абсолютной точностью . Она пренебрежимо мала по сравнению со статистической вариацией оценок метрик между итерациями (см. практическую часть). Такой подход обеспечивает воспроизводимость оценок и исключает смещение из-за выбора сетки. Однако результат расчёта дополнительно сравнивается с дискретными аналогами метрик, рассчитанных на сетке из 1000 точек, для контроля возможных артефактов интегрирования.

* + 1. **Для реальных данных**

Для набора данных по извержениям гейзера Старый Служака, так как истинная функция неизвестна, используются метрики аналогичные прошлым, но подходящие под реальные данные:

Для получения более точной оценки качества метрики вычисляются усреднением по k-fold кросс-валидации[[28]](#footnote-28). Данные разбиваются на k частей – на k-1 обучается модель, а на k-ом фолде замеряется качество. После тестирования на k фолдах метрика усредняется. В силу вычислительных ограничений энтропийного метода для реального датасета в данной работе используется k=5.

## Обзор используемых методов

В данной главе подробно описаны используемые методы оценки регрессии.

### Ядерное сглаживание

#### Обзор метода

Данный подход уже рассматривался в главе, посвящённой непараметрическим подходам, поэтому этой главе мы остановимся лишь на ключевых аспектах применения данного метода в контексте оценки непараметрической регрессии. Впервые данный метод оценки функции плотности случайной величины был предложен в работе М. Розенблата[[29]](#footnote-29) и доработан в исследовании Е. Парзена[[30]](#footnote-30), однако в дальнейшем был многократно модифицирован. В задаче непараметрической регрессии полученная оценка плотности используется для вычисления последовательности ядерных весов, которые были предложены в работах Надарая[[31]](#footnote-31) и Ватсона[[32]](#footnote-32):

(4)

где ширина окна зависит от объёма выборки. В тех же работах предлагается использовать данную весовую схему для оценки условного математического ожидания , то есть непараметрической регрессии по формуле, называемой оценкой Надарая-Ватсона:

(5)

Как уже упоминалось ранее, основной задачей в рамках данного подхода является подбор полосы пропускания h. Чаще всего в качестве критерия оптимальности выбирают ожидаемую функцию потерь L2, а именно MISE (3). Однако так как истинная плотность не известна на практике используются автоматические, основывающиеся на данных методы подбора h. Наиболее эффективными в этой сфере считаются подходы, использующие подключаемые выборочные функции[[33]](#footnote-33) [[34]](#footnote-34) или кросс-валидацию[[35]](#footnote-35) [[36]](#footnote-36) [[37]](#footnote-37). Также можно выделить бутстраповские методы выбора ширины окна, например, как в работе Faraway, Jhun[[38]](#footnote-38), однако их недостатком является случайность целевой функции, которая приводит к вычислительным сложностям. Наконец, существуют эвристические правила подбора, в частности самое популярное – если используется гауссово ядро, и истинная плотность предполагается нормальной, то используется эмпирическое правило Сильвермана[[39]](#footnote-39) для оценки h по формуле:

, (6)

где размер выборки, выборочное среднеквадратичное отклонение. На практике с осторожностью данное правило применяют даже когда истинная плотность ненормальная, например, в качестве первоначального приближения h.

#### Функции ядра

Функция ядра определяет веса наблюдений, их релевантность для оценки плотности в конкретной точке, поэтому несомненно является важным гиперпараметром метода. Ядро должно удовлетворять следующим условиям[[40]](#footnote-40):

1. (неотрицательность);
2. (нормировка);
3. (симметричность и центрированность);
4. (конечная дисперсия).

Существует несколько основных ядер, каждое из которых имеет свои особенности. В данном разделе мы рассмотрим только ядра, наиболее часто применяемые на практике[[41]](#footnote-41).

* + - 1. **Гауссово ядро[[42]](#footnote-42)**

(7)

*Достоинства:* бесконечно дифференцируемая функция, даёт аналогичные свойства оценке плотности. Как следствие предыдущего пункта, не имеет резких границ, то есть при оценке используется информация из всех наблюдений.

*Недостатки:* на больших выборках вычислительно дорогая из-за расчёта экспоненты, теоретически неоптимальна по критерию MISE[[43]](#footnote-43) и склонна к лишнему сглаживанию функции плотности из-за использования всех наблюдений.

Стоит отметить, что несмотря на недостатки гауссово ядро на практике работает почти так же хорошо, как оптимальные ядра, и поэтому широко используется.

* + - 1. **Ядро Епанечникова[[44]](#footnote-44)**

(8)

*Достоинства:* теоретически оптимально по критерию MISE среди всех ядер с конечной схемой взвешивания и благодаря конечному взвешиванию быстрее вычисляется.

*Недостатки:* имеет резкие, недифференцируемые границы, из-за чего производная получается негладкой. На практике данное ядро даёт результат, визуально близкий к применению гауссова ядра.

* + - 1. **Равномерное (прямоугольное) ядро**

(9)

*Достоинства:* вычислительно самое простое ядро.

*Недостатки:* негладкая оценка плотности, плохая сходимость по критерию MISE, очень чувствительна к выбору ширины окна.

Несмотря на недостатки, данный метод иногда применяется в качестве бейзлайна или основы для других методов.

* + - 1. **Треугольное ядро**

(10)

Данное ядро имеет преимущества и недостатки аналогичные равномерному, за исключением немного более гладкой функции.

* + - 1. **Квартичное ядро (Биквадратное)**

(11)

Модификация ядра Епанечникова – вычислительно более сложное и уже не оптимально по MISE, но дифференцируемо один раз.

* + - 1. **Триквадратное ядро**

(12)

Следующая модификация ядра Епанечникова, дифференцируемая уже два раза. Широко используется в методах оценки локальной регрессии.

В заключение обзора стоит отметить, что на практике выбор ядерных функций влияет на точность оценки значительно меньше, чем корректный подбор оптимальной ширины окна.

#### Преимущества

Как и у в случае других непараметрических методов, отсутствие предположений о функции регрессии позволяет адаптироваться у сложным, нелинейным зависимостям. Среди всех представленных данный метод является одним из самых простых и интуитивно понятных – чем ближе наблюдение, тем она важнее для данной точки, и тем больше её вес.

В исследованиях Надарая и Ватсона, упоминавшихся выше, также доказывается, что данная оценка сходится по вероятности к истинному математическому ожиданию при стандартных условиях гладкости. Она асимптотически нормальна, что позволяет строить доверительные интервалы. Кроме того, при использовании ограниченных ядер, например ядра Епанечникова, оценка становится устойчивой к выбросам и аномалиям вне окна.

#### Ограничения метода

Из схемы взвешивания следует, что при ядерном сглаживании возникает систематическое смещение на границах данных[[45]](#footnote-45), так как с какой-то стороны становиться слишком мало наблюдений, либо они не имеются вовсе. Например, если на краю регрессия возрастает, то оценка будет занижать фактическое значение. Для решения этой проблемы иногда применяется локально полиномиальная регрессия, в самом простом случае локально линейная.

Не менее важной проблемой является сильная чувствительность оценки к выбору полосы пропускания. Слишком высокие значения приводят к недосглаживанию, то есть переобучению, в то время как слишком низкие, наоборот, приводят к сверхсглаживанию и слишком гладкой форме кривой. Hardle в своей книге[[46]](#footnote-46) разложил MSE на квадрат смещения и дисперсию и доказал, что оптимальная ширина окна должна балансировать их. В теории для каждой точки существует оптимальная ширина окна[[47]](#footnote-47), однако проблема заключается в том, что она зависит от функции плотности и её производных, то есть того, что требуется оценить. Кроме того, даже при наличии такой возможности требовалось бы непараметрически оценить все эти параметры для каждой точки, что с вычислительной точки зрения слишком затратно. Именно поэтому на практике чаще всего используется глобально оптимальная ширина окна, несмотря на её недостатки. Однако стоит отметить, что есть исследования посвящённые области адаптивных ядерных оценок[[48]](#footnote-48) [[49]](#footnote-49), позволяющим ширине окна меняться в точке. Их известным недостатком является введение ложного шума в оценку плотности.

Как и многие другие методы основанные на схемах взвешивания данный метод сильно страдает от проклятья размерности. В частности, Scott в своей книге[[50]](#footnote-50) показывает, что в многомерном случае скорость сходимости выражается как . Таким образом, при увеличении размерности для точной оценки требуется экспоненциально больше данных.

#### Используемый подход

В данной работе используется метод ядерного сглаживания с классическим гауссовым ядром (7). В качестве baseline оценки оптимальной ширины окна используется эмпирическое правило Сильвермана для гауссова ядра (6). Для более точной оценки используется метод LOOCV (Leave-One-Out Cross-Validation), описанный, например, в статье Анатольева С.[[51]](#footnote-51). В данном подходе критерием минимизации является функция кросс-валидации, в нашем случае основанная на средневадратичной ошибке:

где это оценка Надарайа-Уотсона в точке , использующая все наблюдения кроме i-го:

Таким образом, оптимальная ширина в данном методе находится из задачи минимизации .

У предложенного выше метода есть серьёзный недостаток. LOOCV оптимален асимптотически, но при малых его оценка MSE имеет большую дисперсию[[52]](#footnote-52), чем у k-fold CV с . Так как используются все данные для обучения (все точки кроме одной), то метод слишком подстраивается к данным, что приводит к недосглаживанию и увеличивает дисперсию предсказаний на новых данных.

Поэтому в качестве третьего метода для сравнения используется k-fold кросс-валидация[[53]](#footnote-53). Суть метода заключается в том, что данные разбиваются поровну на k частей, потом итеративно модель оценивается на всех частях, кроме одной, на которой замеряется критерий – в нашем случае MSE. Итоговой функцией кросс-валидации для минимизации является суммарное MSE по всем фолдам. Для маленьких датасетов оптимальное[[54]](#footnote-54) по дисперсии ошибки выбора 5 или 10. В силу небольшого размера синтетического датасета () в данной работе выбирается равным 5, так как при выборе k = 10, на этап тестирования придётся всего 3 наблюдения. Таким образом, проблемой 5-fold CV в данном случае будет зашумленность тестирования, так как точек будет всего 6.

Оба метода кросс валидации сталкиваются с ограничениями на таком маленьком датасете, поэтому сложно заранее определить какой из них окажется лучше, поэтому в данном исследовании применяются оба варианта.

### Информационный подход

#### Обзор метода

Данный подход является комбинацией метода энтропийно-робастного оценивания с «мягкой» рандомизацией и сравнения моделей с помощью информационного критерия BIC.

* + - 1. **Изначальный подход**

Используемый метод энтропийно-робастного оценивания, основан на изначальном подходе с «жёсткой» рандомизацией, который впервые был предложен в работе A. Golan[[55]](#footnote-55) и продолжал применяться и развиваться в том числе и в отечественных исследованиях[[56]](#footnote-56) [[57]](#footnote-57) [[58]](#footnote-58). Далее представлена краткая формулировка метода в рамках нашей задачи. Пусть оцениваемая зависимость представляет собой модель с параметрами , где количество параметров, т.е.

Параметры и ошибки предполагаются непрерывными случайными величинами, которые определены на некоторых интервалах и . Целью метода является восстановление по данным оптимальных функций плотности распределения (ФПР) случайных величин и . Это достигается путём максимизации функционала информационной энтропии:

при выполнении условий нормировки на искомые ФПР:

и ограничений балансировки на выход модели:

Как показано в работе Дубнова Ю. А. и Булычева А. В.[[59]](#footnote-59) данная задача является вариационной задачей Ляпуновского типа и имеет аналитическое решение, параметризованное множителями Лагранжа:

где множители Лагранжа вычисляются из системы уравнений, полученной подстановкой полученных плотностей (19) и (20) в балансовые ограничения (18). Отсюда следует основное ограничение метода – необходимость решать систему уравнений размерности . При большом количестве точек это становится очень затратной с вычислительной точки зрения задачей в связи с проклятьем размерности. Кроме того, на практике также возникают проблемы, связанные с наличием множества локальных оптимумов в многомерном пространстве. Именно поэтому в рамках данного исследования используется подход с «мягкой» рандомизацией, который за счёт регуляризации позволяет найти приближенную функцию распределения с меньшим количеством вычислительных затрат.

* + - 1. **Модифицированный метод**

Метод в первые был предложен[[60]](#footnote-60) и исследован[[61]](#footnote-61) в работах Ю. Дубнова и А. Булычёва. В данной работе исследуются возможности и качество оценки предложенного метода в контексте оценки регрессии.

* + - * 1. **Теоретический вывод метода**

Для начала кратко изложим суть приближённого метода. Постановка задачи в виде (15) на некоторых интервалах и сохраняется, однако вместо точной оптимизации используется приближённое балансирование предсказаний и истинных данных через гельдеровы нормы[[62]](#footnote-62):

где расстояние между моделью и данными, норма вектора шума .

Далее вводится математическое ожидание норм как «эмпирический риск»[[63]](#footnote-63), который позволяет учесть статистику распределения параметров и ошибок :

Для оптимизации строится синтетический функционал , объединяющий энтропийную меру неопределённости (16), «эмпирический риск» (23) и среднюю норму ошибок (24):

Задача сводится к максимизации (25) при условиях нормировки (17). Таким образом, для решения больше не требуется выполнения балансовых ограничений, а следовательно, и решения системы из уравнений. Для учёта ограничений вводится функционал Лагранжа:

где и – штрафные параметры. Условия оптимальности выводятся через производные Гато[[64]](#footnote-64). Так как , то их вариацию можно представить в виде:

где . Тогда производные Гато и условия оптимальности для параметров имеют следующий вид:

Для шума условия имеют аналогичный вид:

Обращение в ноль подынтегральных выражений является необходимым и достаточным условием, чтобы уравнения (29) и (30) выполнялись при любых функциях и . Таким образом, решение задачи в общем виде имеет следующий вид:

Таким образом, оптимальные функции плотности вероятности имеют экспоненциальную форму, т.е. вероятность параметров уменьшается экспоненциально с ростом ошибки и аналогично для ошибок . Однако точная структура распределения определяется самими функциями эмпирического риска и нормы ошибок . Авторы работы также пишут «поскольку величина эмпирического риска зависит не только от параметров модели, но и от имеющихся измерений, то функцию можно трактовать как аналог функции правдоподобия при имеющейся выборке наблюдений. Следовательно, в задачах оценивания при отсутствии информации о виде распределения ошибок и настоящей функции правдоподобия технология «мягкого» энтропийного оценивания позволяет восстановить ее альтернативу.»[[65]](#footnote-65). Таким образом, по полученной плотности можно получить точечную оценку параметров путём нахождения точки максимума плотности и это будет аналог оценки максимума правдоподобия в контексте энтропийно-оптимальных функций. Эта задача эквивалентна задаче минимизации «эмпирического риска» (21):

Также существует второй способ оценки путём нахождения математического ожидания параметров по плотности , однако в данной работе применяется именно первый метод.

* + - * 1. **Применение для оценки регрессии**

В рамках данной работы для оценки регрессии используются полиномиальные регрессии (5) со степенями от 1 до 6, которые в свою очередь оценивается предложенным энтропийным методом.

В качестве нормы в данном исследовании берётся комбинация и норм:

где компонента отбрасывает нерелевантные параметры, зануляя их, а компонента контролирует масштаб коэффициентов, стабилизируя оценки. Для выбора оптимальной степени полинома, а также и можно использовать как эмпирические методы, так и теоретические критерии. В данной работе для этой задачи применяется метод .632+ бутстреп[[66]](#footnote-66) (см. главу 3.2.4). Общее решение (31) представляется в следующем виде:

В заключение стоит отметить, что в случае , т.е. использовании в расчётах только регуляризации оценка как точка максимума энтропийно-оптимальной функции плотности распределения совпадает с оценкой по МНК. Однако в остальных случаях экспоненциальные распределения имеют более сложную форму и приводят к другим оценкам.

#### Преимущества метода

Основное теоретическое преимущество энтропийного метода заключается в его фокусе на первичной информации, а именно функции плотности распределения параметров. Любая интерпретация функции распределения — это сжатие информации из выборки с потерями. Энтропийное оценивание позволяет отыскать именно функцию распределения для параметров. В дальнейшем можно проводить операции с функцией распределения, но именно она представляет из себя первичную информацию: вычисление энтропийных статистик, в частности, моментов информационного содержания, которые проявляют устойчивость для отдельных видов распределений. Например, для нормального дисперсия информационного разброса равна 0.5.

Другим серьёзным преимуществом предложенного метода является отсутствие каких-либо ограничений на ошибки и параметры, кроме отрезка определения. Более того, при выборе оценки как точки максимума энтропийно-оптимальной функции плотности распределения в общем случае не требуется даже отрезок определения, так как его можно выбрать произвольным. Энтропийный подход позволяет получить обоснованные оценки плотности параметров и ошибок в сложных случаях, когда предпосылки классических подходов нарушаются. В частности, можно использовать плотность в качестве аналога функции правдоподобия, когда не представляется возможным построить её обычными методами. Наконец, сами оценки параметров можно выбирать как максимум плотности, математическое ожидание, медиану или другими методами, основанными на плотности распределения параметров.

Преимуществом конкретно метода «мягкой» рандомизации по сравнению с базовой версией является отсутствие необходимости вычислять множители Лагранжа в системе балансовых уравнений размерности . Это позволяет кратно уменьшить вычислительные затраты метода, так как задача сводится к оптимизации с размерностью равной количеству параметров. Благодаря этому метод освобождается от проклятья размерности и независимо от количества точек считается достаточно быстро при выборе классических норм[[67]](#footnote-67). Более того, при выборе уже упомянутого выше подхода оценки максимума плотности пропадает необходимость считать нормировочный интеграл, и задача сводится к минимизации выбранной нормы по параметрам, что значительно ускоряет расчёты, особенно на больших датасетах, но эффект сильно заметен даже на исследуемых датасетах из 30 наблюдений.

Эффективность метода «мягкой» рандомизации в контексте оценки линейной регрессии была экспериментально исследована в работе[[68]](#footnote-68) Дубнова Ю. А. и Булычева А. В.. Результаты указывают на то, что метод является более эффективным по сравнению с аналогами при ассиметричном распределении ошибок модели или если они распределены не нормально, а также при оценке на малых выборках.

#### Ограничения метода

Несмотря на значительное ускорение вычислений энтропийного метода с «мягкой» рандомизацией, существуют системные ограничения, связанные с использованием приближённого подхода.

Во-первых, происходит потеря информации из-за отклонения от приницпа максимума энтропии и асимптотического характера гарантий. В методе с «жёсткой» рандомизацией при выполнении балансовых ограничений полученные в результате максимизации энтропии оценки теоретически являются наилучшими в условиях неопределённости[[69]](#footnote-69) [[70]](#footnote-70) [[71]](#footnote-71). Предложенный приближённый метод, исходя из вывода, асимптотически сохраняет это свойство, однако за счёт потери части информации.

Во-вторых, выбор нормы для «эмпирического риска» определяет итоговые оценки, тем самым косвенно сужая возможные решения. Таким образом, форма «эмпирического риска» в некотором роде параметризует оценки, превращая метод в параметризированный компромисс между данными и субъективным выбором метрики отклонения.

#### Используемый подход

В рамках данного исследования используется описанный выше энтропийный подход с «мягкой» рандомизацией и оценка как точка максимума энтропийно-оптимальной функции плотности распределения . Максимум ищется с помощью метода minimize библиотеки Python scipy[[72]](#footnote-72) методом Пауэлла[[73]](#footnote-73), так как он работает с негладкими функциями. В качестве «эмпирического риска» используется комбинация норм и , как описано в главе 3.2.1.2.2. Коэффициенты и подбираются по сетке соотношений: 1:0, 0:1, 1:1, 3:4, 4:3, 1:2, 2:1, 1:4, 4:1.

Оптимальная комбинация степени и соотношения , как описывалось выше, подбирается с помощью метода .632+ бутстрап[[74]](#footnote-74). Кратко опишем алгоритм подбора для каждой комбинации.

Для каждой модели генерируется бутстреп-выборок размера с возвращением и для каждой выборки определяются выброшенные точки , которые не попали в . Количество выборок на практике рекомендуется брать от 200 до 500. В силу большого количества датасетов для тестирования в данной работе берётся равным 200.

Для каждой бутстреп-итерации обучается модель на , вычисляется ошибка на обучении и выброшенных точках . Считается пессимистичная ошибка модели, которая делает предсказания независимо от данных:

где квадрат ошибки. по сути является ошибкой при случайном сопоставлении.

Ошибки усредняются по выборкам:

Вычисляется относительное переобучение:

где числитель – это абсолютное переобучение, а знаменатель – это максимально возможное переобучение. На основе относительно переобучения вычисляется адаптивный вес

и итоговая ошибка

При сильном переобучении вес ошибки на обучающих точках уменьшается, что предотвращает выбор излишне сложных моделей.

Таким образом, алгоритм применяется ко всем комбинациям степени и соотношения , после чего выбирается вариант с наименьшей ошибкой .

Данный метод был выбран, так как модели оцениваются на очень маленьких выборках . В этом случае методы кросс-валидации имеют большую дисперсию, чем предложенный метод, так как валидация происходит на очень маленьком количестве точек, а бутстреп использует перекрывающиеся выборки, что уменьшает дисперсию по сравнению с кросс-валидацией. Кроме того, полиномы оцениваются очень быстро энтропийным методом, поэтому в данном случае применение бутстрепа является оправданным. В работе Эфрона[[75]](#footnote-75) также доказывается ассимптотическая несмещённость метода.

## Практическая часть

В данной главе подробно описаны реализации предложенных в главе 3 методов, а также представлены результаты генерации и оценивания.

### Генерация датасетов

#### Полиномиальные регрессии

По описанному в главе 2.2.3.1. процессу было сгенерировано 5400 датасетов (6 степеней, 3 диапазона относительного уровня ошибок, по 30 точек в каждом). Для контроля случайности были использованы seed от 1 до 5400 соответственно, что гарантирует реплицируемость генерации. На рис. 2, 3 и 4 приведены примеры этих датасетов при разных уровнях шума и степени.

Изображение выглядит как диаграмма, График, текст, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 2. Примеры датасетов при низком относительном уровне шума. Источник: работа автора.

Изображение выглядит как текст, диаграмма, График, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 3. Примеры датасетов при умеренном относительном уровне шума. Источник: работа автора.

Изображение выглядит как текст, диаграмма, График, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 4. Примеры датасетов при высоком относительном уровне шума. Источник: работа автора.

При низком уровне шума линия регрессии визуально хорошо различима, шум почти не искажает данные, что позволяет тестировать способность методов различать локальные особенности. Эти датасеты моделируют выскокачественные измерения в лабораторных экспериментах.

При высоком уровне шума, наоборот, линию регрессии становится сложно точно определить. Этот вариант моделируют сложные условия, такие как сенсорные данные в реальном времени. На данном типе датасетов проверятся робастность методов к случаным возмущениям.

Средний уровень шума представляет собой баланс этих краевых случаев и является стандартом в исследовании методов оценки регресии[[76]](#footnote-76).

В итоговых датасетах достигается достаточно высокое разнообразие параметров генерации и точность относительного уровня шума. На рис. 5 представлена диаграмма рассеяния, которая демонстрирует равномерное покрытие как фактора масштаба, так и фактора смещения для всех 540 датасетах. Цветовая кодировка по степени указывает на отсутствие систематического смещения в этой части генерации. Кроме того, боксплоты отклонений фактического относительного уровня шума от целевых значений указывает на корректность подгонки коэффициентов и машинную точность генерации. Ожидаемо, отклонение больше при генерации с более высоким шумом. При этом максимальное отклонение составляет порядка . Таким образом, влияние шума на оценку методов является полностью контролируемым.

Изображение выглядит как снимок экрана, Красочность, текст, шаблон

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 5. Распределение параметров генерации полиномов. Источник: работа автора.

Изображение выглядит как диаграмма, текст, снимок экрана, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 6. Точность достижения целевого уровня шума. Источник: работа автора.

Далее на рис. 7-12 ниже представлены распределения коэффициентов разных степеней полиномов при умеренном уровне шума. За вычетом некоторых выбросов для всех степеней наблюдается приблизительно нормальное распределение с нулевым средним и конечной дисперсией, что соответствует предположениям о структуре регрессионной зависимости. Отсутствие вырожденных распределений указывает на корректность генерации, исключая влияние артефактов на оценку качества методов.

Изображение выглядит как текст, диаграмма, линия, График

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 7. Распределение коэффициентов полинома 1-й степени (300 датасетов, умеренный шум). Источник: работа автора.

Изображение выглядит как диаграмма, График, линия, снимок экрана

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 8. Распределение коэффициентов полинома 2-й степени (30 датасетов, умеренный шум). Источник: работа автора.

Изображение выглядит как линия, диаграмма, снимок экрана, График

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 9. Распределение коэффициентов полинома 3-й степени (30 датасетов, умеренный шум). Источник: работа автора.

Изображение выглядит как диаграмма, текст, План, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 10. Распределение коэффициентов полинома 4-й степени (30 датасетов, умеренный шум). Источник: работа автора.

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 11. Распределение коэффициентов полинома 5-й степени (30 датасетов, умеренный шум). Источник: работа автора.

Изображение выглядит как текст, диаграмма, План, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 12. Распределение коэффициентов полинома 6-й степени (30 датасетов, умеренный шум). Источник: работа автора.

Таким образом, предложенных подход корректно генерирует синтетические данные для оценки полиномов, позволяя качественно протестировать и оценить предложенные методы и ответить на ключевые исследовательские вопросы дипломной работы.

#### Двумерная смесь нормальных распределений

По описанному в главе 2.2.3.2. методу было сгенерировано 300 датасетов детерминированной смеси и 30 датасетов стохастической смеси. Примеры генераций представлены на рис. 13. Детерминированный подход позволяет точно контролировать соотношение точек из различных компонент, что позволяет изолировать влияние случайных отклонений на оценку качества методов и сосредоточится на базовой погрешности. Стохастический подход более реалистично моделирует распределение, что добавляет шум, связанный с разным количеством точек, но за счёт этого создаёт условия для проверки предложенных методов на робастность к случайным отклонениям в структуре данных.

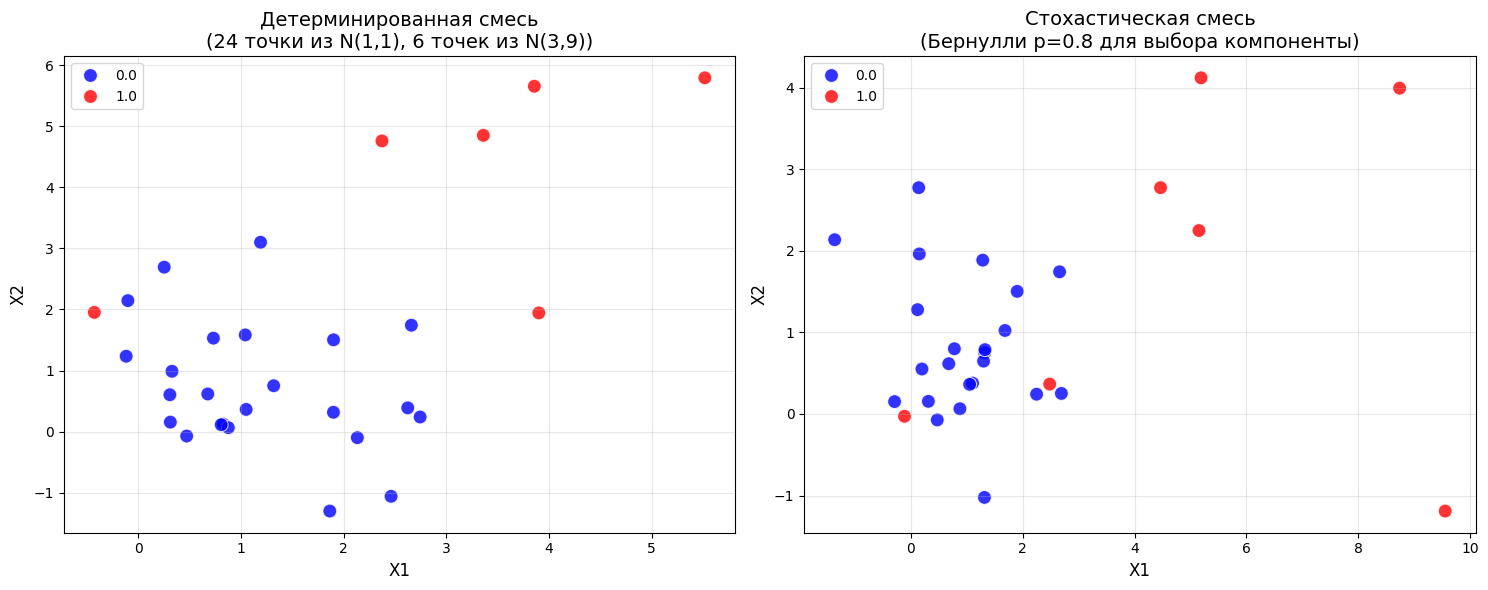


Рис. 13. Примеры генераций смеси распределений. Источник: работа автора.

Как указано в главе 2.3.1 для оценки качества модели по метрикам IMSE и IMAE был выбран отрезок от -1.5 до 6.5, так как он содержит 95% вероятностой массы распределения. Эмпирический анализ сгенерированных датасетов подтверждает теоретическую корректность такого выбора (см. рис. 14): в детерминированной смеси в диапазон попадает точек, а в стохастической . Кроме того, формы распределения компонент по x практически совпадают. Статистические характеристики также практически не отличаются (см. рис. 15). Таким образом, несмотря на различия в подходах, конечная модель синтетических данных остаётся в достаточной степени одинаковой.

Изображение выглядит как текст, диаграмма, График, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 14. Анализ распределения точек смеси по оси x в генерациях. Источник: работа автора.

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 15. Полная статистическая характеристика сгенерированных смесей. Источник: работа автора.

Однако важно отметить, что несмотря на одинаковость в среднем, следующую из теории, в отдельных отклонениях различия, за счёт которых тестируется робастность, наблюдаются. На рис. 16. видно количество точек из второго компонента отличатся от датасета к датасету и приблизительно следует биномиальному распределению .

Изображение выглядит как текст, диаграмма, График, линия

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 16. Распределение количества точек из компонента 2 (стохастическая генерация, 300 датасетов). Источник: работа автора.

Таким образом, анализ даёт представление о статистических свойствах сгенерированных данных и демонстрирует, что предложенная генерация корректна в рамках данного исследования.

### Оценка моделей на синтетических данных

#### Полиномиальные данные

* + - 1. **Оценка ядерной регрессии**

По результатам оценки были получены оценки метрик, представленные в таблицах 1-6.

Таблица 1. Результаты оценки ядерной регрессии на полиномиальных данных (1 степень). Источник: Работа автора.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Шум** | **Метод** |  |  |  |  |  |  |
| low | silverman | 0,160 | 0,016 | 0,279 | 0,007 | 0,538 | 0,035 |
| low | 5-fold cv | 0,146 | 0,013 | 0,264 | 0,009 | 0,521 | 0,024 |
| low | loocv | 0,187 | 0,054 | 0,265 | 0,010 | 0,597 | 0,086 |
| moderate | silverman | 0,176 | 0,029 | 0,277 | 0,009 | 0,532 | 0,032 |
| moderate | 5-fold cv | 0,166 | 0,034 | 0,258 | 0,012 | 0,516 | 0,031 |
| moderate | loocv | 0,168 | 0,035 | 0,264 | 0,012 | 0,517 | 0,029 |
| high | silverman | 0,178 | 0,019 | 0,282 | 0,008 | 0,534 | 0,036 |
| high | 5-fold cv | 0,157 | 0,015 | 0,270 | 0,009 | 0,533 | 0,024 |
| high | loocv | 0,151 | 0,014 | 0,268 | 0,009 | 0,514 | 0,021 |

Таблица 2. Результаты оценки ядерной регрессии на полиномиальных данных (2 степень). Источник: Работа автора.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Шум** | **Метод** |  |  |  |  |  |  |
| low | silverman | 0,421 | 0,029 | 0,478 | 0,017 | 0,788 | 0,031 |
| low | 5-fold cv | 0,868 | 0,477 | 0,406 | 0,015 | 0,696 | 0,029 |
| low | loocv | 0,781 | 0,469 | 0,398 | 0,014 | 0,693 | 0,028 |
| moderate | silverman | 0,266 | 0,020 | 0,383 | 0,013 | 0,647 | 0,024 |
| moderate | 5-fold cv | 0,378 | 0,079 | 0,375 | 0,012 | 0,669 | 0,029 |
| moderate | loocv | 0,394 | 0,085 | 0,370 | 0,012 | 0,673 | 0,040 |
| high | silverman | 0,250 | 0,026 | 0,356 | 0,012 | 0,617 | 0,030 |
| high | 5-fold cv | 0,268 | 0,033 | 0,351 | 0,011 | 0,618 | 0,034 |
| high | loocv | 0,295 | 0,044 | 0,354 | 0,012 | 0,667 | 0,047 |

Таблица 3. Результаты оценки ядерной регрессии на полиномиальных данных (3 степень). Источник: Работа автора.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Шум** | **Метод** |  |  |  |  |  |  |
| low | silverman | 0,442 | 0,043 | 0,469 | 0,017 | 0,814 | 0,035 |
| low | 5-fold cv | 0,428 | 0,098 | 0,406 | 0,011 | 0,729 | 0,031 |
| low | loocv | 0,335 | 0,032 | 0,396 | 0,010 | 0,716 | 0,029 |
| moderate | silverman | 0,324 | 0,038 | 0,396 | 0,013 | 0,673 | 0,028 |
| moderate | 5-fold cv | 0,405 | 0,070 | 0,394 | 0,013 | 0,723 | 0,041 |
| moderate | loocv | 0,417 | 0,059 | 0,399 | 0,013 | 0,756 | 0,042 |
| high | silverman | 0,194 | 0,019 | 0,319 | 0,010 | 0,577 | 0,031 |
| high | 5-fold cv | 0,224 | 0,022 | 0,331 | 0,009 | 0,560 | 0,022 |
| high | loocv | 0,358 | 0,115 | 0,343 | 0,012 | 0,609 | 0,033 |

Таблица 4. Результаты оценки ядерной регрессии на полиномиальных данных (4 степень). Источник: Работа автора.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Шум** | **Метод** |  |  |  |  |  |  |
| low | silverman | 0,653 | 0,078 | 0,541 | 0,021 | 0,946 | 0,042 |
| low | 5-fold cv | 0,410 | 0,052 | 0,421 | 0,012 | 0,744 | 0,032 |
| low | loocv | 0,396 | 0,050 | 0,416 | 0,011 | 0,765 | 0,039 |
| moderate | silverman | 0,390 | 0,035 | 0,429 | 0,016 | 0,747 | 0,031 |
| moderate | 5-fold cv | 0,370 | 0,041 | 0,397 | 0,012 | 0,780 | 0,055 |
| moderate | loocv | 0,391 | 0,057 | 0,394 | 0,012 | 0,737 | 0,038 |
| high | silverman | 0,295 | 0,041 | 0,369 | 0,013 | 0,648 | 0,035 |
| high | 5-fold cv | 0,376 | 0,088 | 0,378 | 0,012 | 0,686 | 0,028 |
| high | loocv | 0,384 | 0,092 | 0,377 | 0,012 | 0,686 | 0,031 |

Таблица 5. Результаты оценки ядерной регрессии на полиномиальных данных (5 степень). Источник: Работа автора.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Шум** | **Метод** |  |  |  |  |  |  |
| low | silverman | 0,786 | 0,070 | 0,606 | 0,021 | 1,120 | 0,043 |
| low | 5-fold cv | 0,826 | 0,368 | 0,454 | 0,017 | 0,904 | 0,061 |
| low | loocv | 1,090 | 0,472 | 0,469 | 0,019 | 0,903 | 0,047 |
| moderate | silverman | 0,413 | 0,031 | 0,454 | 0,014 | 0,824 | 0,032 |
| moderate | 5-fold cv | 0,411 | 0,053 | 0,410 | 0,011 | 0,819 | 0,055 |
| moderate | loocv | 0,541 | 0,164 | 0,410 | 0,015 | 0,845 | 0,060 |
| high | silverman | 0,303 | 0,027 | 0,380 | 0,011 | 0,653 | 0,030 |
| high | 5-fold cv | 0,401 | 0,073 | 0,388 | 0,011 | 0,645 | 0,022 |
| high | loocv | 1,232 | 0,878 | 0,400 | 0,018 | 0,671 | 0,028 |

Таблица 6. Результаты оценки ядерной регрессии на полиномиальных данных (6 степень). Источник: Работа автора.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Шум** | **Метод** |  |  |  |  |  |  |
| low | silverman | 0,972 | 0,073 | 0,660 | 0,022 | 1,281 | 0,051 |
| low | 5-fold cv | 0,606 | 0,086 | 0,469 | 0,012 | 0,913 | 0,043 |
| low | loocv | 0,580 | 0,086 | 0,461 | 0,012 | 0,850 | 0,033 |
| moderate | silverman | 0,591 | 0,079 | 0,499 | 0,018 | 0,940 | 0,043 |
| moderate | 5-fold cv | 0,465 | 0,122 | 0,414 | 0,013 | 0,804 | 0,072 |
| moderate | loocv | 0,460 | 0,120 | 0,410 | 0,013 | 0,827 | 0,076 |
| high | silverman | 0,320 | 0,035 | 0,396 | 0,012 | 0,683 | 0,026 |
| high | 5-fold cv | 0,286 | 0,020 | 0,379 | 0,010 | 0,722 | 0,035 |
| high | loocv | 0,337 | 0,041 | 0,383 | 0,011 | 0,793 | 0,055 |

# ПРИЛОЖЕНИЕ 2

**Вывод формулы регрессии для двумерной смеси двумерных нормальных распределений**

**Вывод регрессии в общем виде:**

Пусть двумерный вектор имеет распределение смеси:

где вес первого компонента, вектор средних для компонента k, ковариационная матрица для компонента k.

Для смеси условное распределение также является смесью:

где апостериорная вероятность принадлежности к первому компоненту при условии согласно теореме Байеса, маргинальная плотность .

Согласно свойствам нормального распределения, условное распределение для каждого компонента k нормально и имеет следующие параметры:

По свойству условного ожидания для смеси:

Определим случайную ошибку как:

Тогда по построению:

Таким образом, мы получаем искомое разложение:

**Вывод формулы истинной регрессии при известных параметрах:**

В данной работе взяты следующие параметры компонент для генерации: , , , .

Поскольку ковариации в обоих компонентах равны нулю, то в силу нормальности распределений и независимы внутри каждого компонента. Тогда:

* Для компонента 1:
* Для компонента 1:

По формуле Байеса вычислим апостериорные вероятности принадлежности к компонентам при условии :

где маргинальная плотность X в компоненте .

Подставляем указанные выше параметры смеси:

где

Условное математическое ожидание:

Подставляем :

Возвращаем A и B:

Эта формула и является истинным уравнением регрессии в данном случае.

1. Anatolyev, Stanislav. (2009). Nonparametric regression (in Russian). Quantile. 37-52. [↑](#footnote-ref-1)
2. Salibian-Barrera, Matias. (2022). Robust nonparametric regression: review and practical considerations. 10.48550/arXiv.2211.08376. [↑](#footnote-ref-2)
3. Rosenblatt, Murray. “Remarks on Some Nonparametric Estimates of a Density Function.” Annals of Mathematical Statistics 27 (1956): 832-837. [↑](#footnote-ref-3)
4. Parzen, Emanuel (1962) On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33 (3) 1065-1076 doi:10.1214/aoms/1177704472 [↑](#footnote-ref-4)
5. Jie Jiang, Yu-Lin He, De-Xin Dai, Joshua Zhexue Huang, A new kernel density estimator based on the minimum entropy of data set, Information Sciences, Volume 491, 2019, Pages 223-231, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.04.010. [↑](#footnote-ref-5)
6. Miecznikowski, J. C., Wang, D., & Hutson, A. (2010). Bootstrap MISE Estimators to Obtain Bandwidth for Kernel Density Estimation. Communications in Statistics - Simulation and Computation, 39(7), 1455–1469. https://doi.org/10.1080/03610918.2010.500108 [↑](#footnote-ref-6)
7. Mammen, E., Martínez Miranda, M. D., Nielsen, J. P., & Sperlich, S. (2011). Do-Validation for Kernel Density Estimation. Journal of the American Statistical Association, 106(494), 651–660. https://doi.org/10.1198/jasa.2011.tm08687 [↑](#footnote-ref-7)
8. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 45. ISBN 978-0-412-24620-3. [↑](#footnote-ref-8)
9. Wakefield, B., Lin, YX., Sarathy, R. et al. Moment-based density estimation of confidential micro-data: a computational statistics approach. Stat Comput 33, 35 (2023). https://doi.org/10.1007/s11222-022-10203-1 [↑](#footnote-ref-9)
10. Yanhong Luo, Xu Wang, Shijie Yan, Risk assessment of photovoltaic distribution network based on adaptive kernel density estimation and cumulant method, Energy Reports, Volume 8, Supplement 13, 2022, Pages 1152-1159, ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2022.08.156. [↑](#footnote-ref-10)
11. Alberto Caron, Gianluca Baio, Ioanna Manolopoulou, Estimating Individual Treatment Effects using Non-Parametric Regression Models: a Review, Journal of the Royal Statistical Society Series A: Statistics in Society, Volume 185, Issue 3, July 2022, Pages 1115–1149, https://doi.org/10.1111/rssa.12824 [↑](#footnote-ref-11)
12. Mingming Zhang, Zikun Yang, Liyun Liu, Dequn Zhou, Impact of renewable energy investment on carbon emissions in China - An empirical study using a nonparametric additive regression model, Science of The Total Environment, Volume 785, 2021, 147109, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2021.147109. [↑](#footnote-ref-12)
13. Cheng, Tingting and Gao, Jiti and Linton, Oliver B. and Yan, Yayi, Nonparametric Predictive Regression for Stock Return Prediction 1 (February 05, 2025). Available at SSRN: https://ssrn.com/abstract=5126279 or http://dx.doi.org/10.2139/ssrn.5126279 [↑](#footnote-ref-13)
14. Bradic, Jelena & Chernozhukov, Victor & Newey, Whitney & Zhu, Yinchu. (2019). Minimax Semiparametric Learning With Approximate Sparsity. 10.48550/arXiv.1912.12213. [↑](#footnote-ref-14)
15. Yuping Song, Chunchun Cai, Dexiang Ma, Chen Li, Modelling and forecasting high-frequency data with jumps based on a hybrid nonparametric regression and LSTM model, Expert Systems with Applications, Volume 237, Part A, 2024, 121527, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.121527. [↑](#footnote-ref-15)
16. Chen, Minshuo & Jiang, Haoming & Liao, Wenjing & Zhao, Tuo. (2022). Nonparametric regression on low-dimensional manifolds using deep ReLU networks: function approximation and statistical recovery. Information and Inference: A Journal of the IMA. 11. 10.1093/imaiai/iaac001. [↑](#footnote-ref-16)
17. Taha Hussein Ali, Heyam Abd Al-Majeed Hayawi & Delshad Shaker Ismael Botani (2021): Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2021.1884719 [↑](#footnote-ref-17)
18. Xin Bing, Xin He, Chao Wang, 2025, Kernel Ridge Regression with Predicted Feature Inputs and Applications to Factor-Based Nonparametric Regression, https://doi.org/10.48550/arXiv.2505.20022 [↑](#footnote-ref-18)
19. Taha Hussein Ali, Heyam Abd Al-Majeed Hayawi & Delshad Shaker Ismael Botani (2021): Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2021.1884719 [↑](#footnote-ref-19)
20. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation //Journal of the American statistical association. – 1983. – Т. 78. – №. 382. – С. 316-331. [↑](#footnote-ref-20)
21. van der Vaart, A. W. (1998). Asymptotic Statistics: U-Statistics. https://doi.org/10.1017/CBO9780511802256.013 [↑](#footnote-ref-21)
22. Loader C. Local Regression and Likelihood / C. Loader. — New York : Springer, 1999. — 308 p. — ISBN 0-387-98775-6. [↑](#footnote-ref-22)
23. Ruppert D. Semiparametric Regression / D. Ruppert, M. P. Wand, R. J. Carroll. — Cambridge : Cambridge University Press, 2003. — 386 p. — ISBN 0-521-78012-8. [↑](#footnote-ref-23)
24. Härdle W. Applied Nonparametric Regression / W. Härdle. — Cambridge : Cambridge University Press, 1990. — 334 p. — ISBN 0-521-36985-8 [↑](#footnote-ref-24)
25. Fan J. Local Polynomial Modelling and Its Applications / J. Fan, I. Gijbels. — London : Chapman & Hall, 1996. — 396 p. — ISBN 0-412-98321-4. [↑](#footnote-ref-25)
26. Piessens R. и др. Quadpack. : Springer Berlin Heidelberg, 1983. [↑](#footnote-ref-26)
27. Virtanen, P. et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, pp.261–272. [↑](#footnote-ref-27)
28. Analyzing Microarray Gene Expression Data. Wiley Series in Probability and Statistics. Wiley. 2004. doi:10.1002/047172842X. ISBN 978-0-471-22616-1. [↑](#footnote-ref-28)
29. Rosenblatt M. Remarks on Some Nonparametric Estimates of a Density Function // The Annals of Mathematical Statistics. — 1956. — Т. 27, вып. 3. — doi:10.1214/aoms/1177728190. [↑](#footnote-ref-29)
30. Parzen, Emanuel (1962) On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33 (3) 1065-1076 doi:10.1214/aoms/1177704472 [↑](#footnote-ref-30)
31. Nadaraya, E. A. On Estimating Regression // Theory of Probability and its Applications: journal. — 1964. — Vol. 9, no. 1. — P. 141—142. — doi:10.1137/1109020 [↑](#footnote-ref-31)
32. Watson, G. S. Smooth regression analysis // Sankhyā: The Indian Journal of Statistics, Series A. — 1964. — Т. 26, № 4. — С. 359—372. — JSTOR 25049340 [↑](#footnote-ref-32)
33. Botev Z.I., Grotowski J.F., Kroese D.P. Kernel density estimation via diffusion // Annals of Statistics. — 2010. — Т. 38, вып. 5. — doi:10.1214/10-AOS799. — arXiv:1011.2602. [↑](#footnote-ref-33)
34. Sheather S.J., Jones M.C. A reliable data-based bandwidth selection method for kernel density estimation // Journal of the Royal Statistical Society, Series B. — 1991. — Т. 53, вып. 3. — JSTOR 2345597. [↑](#footnote-ref-34)
35. Rudemo M. Empirical choice of histograms and kernel density estimators // Scandinavian Journal of Statistics. — 1982. — Т. 9, вып. 2. — JSTOR 4615859. [↑](#footnote-ref-35)
36. Bowman A.W. An alternative method of cross-validation for the smoothing of density estimates // Biometrika. — 1984. — Т. 71, вып. 2. — doi:10.1093/biomet/71.2.353. [↑](#footnote-ref-36)
37. Hall P., Marron J.S., Park B.U. Smoothed cross-validation // Probability Theory and Related Fields. — 1992. — Т. 92. — С. 1–20. — doi:10.1007/BF01205233. [↑](#footnote-ref-37)
38. Faraway, J., M. Jhun Bootstrap choice of bandwidth for density estimation. Journal of the American Statistical Association 85. — 1990. — С. 1119–1122. [↑](#footnote-ref-38)
39. Silverman B.W. Density Estimation for Statistics and Data Analysis. — London: Chapman & Hall/CRC, 1986. — ISBN 0-412-24620-1. [↑](#footnote-ref-39)
40. Watson, G. S. Smooth regression analysis // Sankhyā: The Indian Journal of Statistics, Series A. — 1964. — Т. 26, № 4. — С. 359—372. — JSTOR 25049340 [↑](#footnote-ref-40)
41. Wand, Matt P. and M. Chris Jones. “Kernel Smoothing.” (1995). [↑](#footnote-ref-41)
42. Silverman B.W. Density Estimation for Statistics and Data Analysis. — London: Chapman & Hall/CRC, 1986. — ISBN 0-412-24620-1. [↑](#footnote-ref-42)
43. Wand, Matt P. and M. Chris Jones. “Kernel Smoothing.” (1995). [↑](#footnote-ref-43)
44. Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153–158 [↑](#footnote-ref-44)
45. Fan, J. (1996). Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66 (1st ed.). Routledge. https://doi.org/10.1201/9780203748725 [↑](#footnote-ref-45)
46. Applied Nonparametric Regression W. Härdle Cambridge University Press, 1990. Econometric Theory. 8. 413-419. 10.1017/S0266466600013025. [↑](#footnote-ref-46)
47. Anatolyev, Stanislav. (2009). Nonparametric regression (in Russian). Quantile. 37-52. [↑](#footnote-ref-47)
48. Abramson, I.S. On bandwidth variation in kernel estimates – a square root law. Annals of Statistics 10. — 1982. — С. 1217–1223. [↑](#footnote-ref-48)
49. Breiman, L., W. Meisel, E. Purcell Variable kernel estimates of multivariate densities. Technometrics 19. — 1977. — С. 135–144. [↑](#footnote-ref-49)
50. Scott, D.W. (2012). Multivariate Density Estimation and Visualization. [↑](#footnote-ref-50)
51. Anatolyev, Stanislav. (2009). Nonparametric regression (in Russian). Quantile. 37-52. [↑](#footnote-ref-51)
52. Bengio Y., Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation //Journal of machine learning research. – 2004. – Т. 5. – №. Sep. – С. 1089-1105. [↑](#footnote-ref-52)
53. Analyzing Microarray Gene Expression Data. Wiley Series in Probability and Statistics. Wiley. 2004. doi:10.1002/047172842X. ISBN 978-0-471-22616-1. [↑](#footnote-ref-53)
54. Arlot, S. and Celisse, A. (2010) A Survey of Cross-Validation Procedures for Model Selection. Statistics Surveys, 4, 40-79. https://doi.org/10.1214/09-SS054 [↑](#footnote-ref-54)
55. Amos Golan, George G. Judge, Douglas Miller. Maximum Entropy Econometrics: Robust Estimation with Limited

    Data. – John Wiley and Sons Ltd. Chichester, U.K., 1996. [↑](#footnote-ref-55)
56. Yu. S. Popkov , Yu. A. Dubnov. Entropy-robust randomized forecasting under small sets of retrospective data // Automation and Remote Control. 2016, Volume 77, Issue 5, pp 839-854. [↑](#footnote-ref-56)
57. Popkov, Y.S.; Dubnov, Y.A.; Popkov, A.Y. New Method of Randomized Forecasting Using Entropy-Robust Estimation: Application to the World Population Prediction. // Mathematics, 2016, Vol. 4, Iss.1, p.1-16. [↑](#footnote-ref-57)
58. Дубнов Ю. А., Булычев А. В. Приближенное оценивание с помощью ускоренного метода наибольшей энтропии. Часть 1. Постановка задачи и реализация для задачи регрессии //Информационные технологии и вычислительные системы. – 2022. – №. 4. – С. 69-80. [↑](#footnote-ref-58)
59. Там же [↑](#footnote-ref-59)
60. Там же. [↑](#footnote-ref-60)
61. Дубнов Ю. А., Булычев А. В. Приближенное оценивание с помощью ускоренного метода наибольшей энтропии. Часть 2. Исследование свойств оценок //Информационные технологии и вычислительные системы. – 2023. – №. 1. – С. 71-81. [↑](#footnote-ref-61)
62. Воеводин В.В., Кузнецов Ю.А. Матрицы и вычисления. М., Наука, 1984. [↑](#footnote-ref-62)
63. Цыпкин Я.З. Основы теории обучающихся систем. М., Наука, 1970. [↑](#footnote-ref-63)
64. Kaashoek M.A., Seatzu S., van der Mee C. Recent Advances in Operator Theory and its Application. 2006, Springer, p.478. [↑](#footnote-ref-64)
65. Дубнов Ю. А., Булычев А. В. Приближенное оценивание с помощью ускоренного метода наибольшей энтропии. Часть 1. Постановка задачи и реализация для задачи регрессии //Информационные технологии и вычислительные системы. – 2022. – №. 4. – С. 69-80. [↑](#footnote-ref-65)
66. Efron B., Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method //Journal of the American Statistical Association. – 1997. – Т. 92. – №. 438. – С. 548-560. [↑](#footnote-ref-66)
67. Дубнов Ю. А., Булычев А. В. Приближенное оценивание с помощью ускоренного метода наибольшей энтропии. Часть 2. Исследование свойств оценок //Информационные технологии и вычислительные системы. – 2023. – №. 1. – С. 71-81. [↑](#footnote-ref-67)
68. Там же. [↑](#footnote-ref-68)
69. Jaynes E.T. Information Theory and Statistical Mechanics // Physics Review Notes. 1957. V. 106. P. 620–630. [↑](#footnote-ref-69)
70. R.D. Levin, M. Tribus. The maximum entropy formalism.MIT Press, 1979. [↑](#footnote-ref-70)
71. Jaynes E.T. Probability Theory. The logic and science. Cambrige Univ. Press, 2003. [↑](#footnote-ref-71)
72. Virtanen P. SciPy 1.0: fundamental algorithms for scientific computing in Python / P. Virtanen, R. Gommers, T. E. Oliphant [et al.] // Nature Methods. — 2020. — Vol. 17, № 3. — P. 261–272. — DOI: 10.1038/s41592-019-0686-2. [↑](#footnote-ref-72)
73. M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives (англ.) // The Computer Journal. — 1964. — 1 January (vol. 7). — P. 155–162. [↑](#footnote-ref-73)
74. Efron B., Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method //Journal of the American Statistical Association. – 1997. – Т. 92. – №. 438. – С. 548-560. [↑](#footnote-ref-74)
75. Там же. [↑](#footnote-ref-75)
76. Härdle W. Applied Nonparametric Regression / W. Härdle. — Cambridge : Cambridge University Press, 1990. — 334 p. — ISBN 0-521-36985-8 [↑](#footnote-ref-76)