Федеральное государственное бюджетное образовательное учреждение высшего образования

«Новосибирский государственный технический университет»

УДК 519.23

На правах рукописи

БОБОЕВ ШАРАФ АСРОРОВИЧ

ПОСТРОЕНИЕ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ С ИСПОЛЬЗОВАНИЕМ КВАДРАТИЧНОЙ ФУНКЦИИ ПОТЕРЬ В МЕТОДЕ ОПОРНЫХ ВЕКТОРОВ (LS–SVM)

Диссертация на соискание ученой степени
кандидата технических наук по специальности
1.2.7 – Теоретические основы информатики

Научный руководитель:
доктор технических наук, профессор
Попов Александр Александрович

Душанбе – 2025

ОГЛАВЛЕНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ 5

ВВЕДЕНИЕ 6

ГЛАВА 1. РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПО МЕТОДУ LS–SVM 15

1.1 Исторический обзор 15

1.2 Основные понятия и определения 17

1.3 Ядра и спрямляющие пространства 21

1.4 LS–SVM регрессия 22

1.5 Критерии выбора модели оптимальной сложности 24

1.5.1 Критерий LOO CV 25

1.5.2 Критерий K-FOLD CV 25

1.6 Подбор метапараметров алгоритма LS–SVM 27

1.7 Исследования 29

1.8 Выводы 37

ГЛАВА 2. РОБАСТНОЕ РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПО МЕТОДУ LS–SVM 39

2.1 Основные понятия и определения 39

2.2 Метод M–оценивания 40

2.3 Метод псевдонаблюдений на основе функций потерь Хьюбера 41

2.4 Взвешенный метод LS–SVM на основе весовой функции Сайкенса 43

2.5 Взвешенный метод на основе весовой функции потерь Хьюбера 44

2.6 Робастные решения на основе функций потерь Эндрюса и биквадратной Тьюки 45

2.7 Робастные критерии выбора оптимальной модели 46

2.7.1 Критерий RLOO–P 47
2.7.2 Критерий RLOO 47
2.8 Исследования 48
2.9 Выводы 64
ГЛАВА 3. РАЗРЕЖЕННОЕ РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПО МЕТОДУ LS–SVM 65
3.1 Основные понятия и определения 65
3.2 Разреженное решение 66
3.3 Оптимальные планы. D–оптимальный план 67
3.4 Разбиение выборки с использованием D–оптимального планирования эксперимента 69
3.5 Разбиение выборки с использованием внешних критериев оценки качества моделей 72
3.5.1 Внешние критерии оценки качества моделей 72
3.5.2 Вариант включения 73
3.5.3 Вариант исключения 74
3.5.4 Вариант замены 74
3.5.5 Вариант Add/Del 75
3.5.6 Вариант Del/Add 75
3.6 Исследования 76
3.7 Выводы 91
ГЛАВА 4. ПРИМЕНЕНИЕ МЕТОДА LS–SVM ДЛЯ РЕШЕНИЯ ПРАКТИЧЕСКИХ ЗАДАЧ 92
4.1 Выборка LIDAR 92
4.2 Выборка MOTORCYCLE 96

4.3 Изучение процесса комплексообразования переходных металлов с производными тиомочевины в водных и водно-органических растворах 98
4.4 Выводы 111
ГЛАВА 5. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ПОСТРОЕНИЯ LS–SVM РЕГРЕССИИ 112
ЗАКЛЮЧЕНИЕ 124
СПИСОК ЛИТЕРАТУРЫ 126
ПРИЛОЖЕНИЕ А 143
ПРИЛОЖЕНИЕ Б 158
ПРИЛОЖЕНИЕ В 160

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

SVM	–	Support Vector Machines (Метод опорных векторов)
SVR	–	Support Vector Regression (Регрессия опорных векторов)
LS–SVM	–	Least Squares Support Vector Machines (Метод опорных векторов с квадратичной функцией потерь)
FLSA-SVM	–	Forward Least Squares Approximation Support Vector Machines
WLS-SVM	–	Weighted Least Squares Support Vector Machines
QP	–	Quadratic Programming (квадратичное программирование)
RBF	–	Radial Basis Function (радиально-базисные функции)
CV	–	Cross Validation (перекрестная проверка)
LOO	–	Leave One Out (исключение по одному)
LOO CV	–	Leave One Out Cross Validation (перекрестная проверка по отдельным объектам)
K-FOLD CV	–	K-FOLD Cross Validation
MSE	–	Mean Square Error (среднеквадратическая ошибка)
REG	–	Regularity (регулярность)
STAB	–	Stability (стабильность)
RLOO	–	Robust Leave One Out
RLOO-P	–	Robust Leave One Out - Pseudo
СЛАУ	–	Система линейных алгебраических уравнений
LU-разложение	–	LowerUpper-разложение
LIDAR	–	Light Detection And Ranging
$\hat{Y}(R)$	–	робастное решение
$\hat{Y}(S)$	–	разреженное решение
МНК	–	Метод наименьших квадратов
IFOST	–	International Forum Of Strategic Technology
АПЭП	–	Актуальные проблемы электронного приборостроения
НГТУ	–	Новосибирский государственных технический университет
СибГУТИ	–	Сибирский государственный университет телекоммуникаций и информатики
РФ	–	Российская Федерация
РТ	–	Республика Таджикистан
ВАК	–	Высшая аттестационная комиссия
Минобрнауки	–	Министерство образования и науки

ВВЕДЕНИЕ

Актуальность темы исследования. В настоящее время машинное обучение и анализ данных становятся все более важными в различных областях, таких как промышленность, финансы, медицина и многие другие. В этом контексте построение эффективных моделей регрессии, способных предсказывать значение непрерывной целевой переменной на основе набора признаков, является ключевой задачей. Развитие данного направления связано с возможностями современной вычислительной техники, оснащенной пакетами прикладных программ для машинной обработки статистической информации. Благодаря таким передовым вычислительным технологиям и программам для обработки данных, на сегодняшний день анализ информации стал быстрым и простым. Раньше аналитики тратили массу времени на этот процесс, часто сталкиваясь с ограничениями. Однако с развитием компьютерных технологий они получили большую свободу выбора объектов исследования.

Одной из областей, обладающих такой свободой, является непараметрическая оценка регрессии, также называемая сглаживающим методом моделирования. Цель непараметрических методов заключается в снижении ограничений на функциональную форму оцениваемых объектов. Эти методы становятся все более популярными и востребованными в прикладных исследованиях и применяются в тех случаях, когда параметрические модели не подходят для решения поставленной задачи. Непараметрические методы обладают более универсальной структурой, имеют широкий спектр применения и способны работать в условиях высокой неопределенности априорной информации. Кроме того, они эффективны для анализа большого объема данных при малом количестве переменных. К основным методам построения гибких моделей относятся ядерные методы, сглаживание сплайнами, методы ближайших соседей и нейронные сети.

Одним из таких методов, относящихся к классу ядерных методов, является метод опорных векторов с квадратичной функцией потерь.

Метод опорных векторов (SVM) является широко используемым инструментом моделирования зависимостей в условиях структурной неопределенности. Известная его модификация на основе квадратичной функции потерь (LS–SVM) позволяет получать решения в явном виде. При этом сохраняется возможность получать как гладкие, так и достаточно сложные зависимости за счет соответствующего выбора всех параметров алгоритма (вида ядерных функций, их числа и параметров и т.д.). Однако имеются определенные сложности использования данной технологии. Можно здесь выделить два основных отрицательных момента:

1. решения получаются неразреженными;

2. квадратичная функция потерь не обеспечивает в общем случае получение робастных решений.

Степень изученности научной темы. Основными отличиями метода LS–SVM от метода SVM являются:

- Классический SVM решает задачу оптимизации с ограничениями, используя методы квадратичного программирования (QP).
- LS–SVM заменяет запас устойчивости (slack variables) на квадратичную функцию потерь и использует метод наименьших квадратов (Least Squares).
- В результате получается система линейных уравнений, которую можно решить эффективными численными методами, такими как метод Гаусса, метод Гаусса-Жордана или LU-разложением.

Преимуществом метода LS–SVM заключается в:

- более быстрой оптимизации – за счет решения линейных уравнений, а не квадратичного программирования.
- упрощенной реализации – методы решения линейных уравнений проще в реализации, чем методы оптимизации QP.

- эффективности при применении на больших данных – LS-SVM легче масштабируется.
- гибкости – применим как к задачам классификации, так и к регрессии.

Полномасштабное исследование этих проблем позволило бы усовершенствовать данный подход, сделав его полноценным инструментом в руках исследователей.

Связь исследования с программами и научной тематикой. Метод LS–SVM может применятся в различных областях, включая:

1. **Машинное обучение и анализ данных** в котором решаются задачи: классификации изображений и текста, распознавание речи и анализ временных рядов.
2. **Биоинформатика и медицина** для диагностики заболеваний на основе медицинских данных, анализ ДНК и белков и обнаружение аномалий в медицинских снимках.
3. **Финансовая аналитика** для прогнозирования биржевых котировок, оценки кредитного риска и обнаружение мошеннических транзакций.
4. **Промышленность и инженерия** для контроля качества продукции, предсказание отказов оборудования и управление сложными системами.
5. **Робототехника и автономные системы** для обучения интеллектуальных агентов, распознавание объектов и планирование движений.

Общая характеристика исследования. Диссертационная работа посвящена разработке и исследованию методов получения робастных и разреженных регрессионных моделей на базе метода опорных векторов с использованием квадратичной функции потерь (LS–SVM), ориентированных на применение в условиях неопределенности, нестабильности и зашумленности исходных данных.

Цель исследования. Целями диссертационной работы являются разработка и исследование способов получения применительно к LS–SVM:

- робастных решений с использованием методов М–оценивания и взвешивания;
- робастных критериев подбора метапараметров алгоритма и выбора оптимальных устойчивых моделей с оценкой их качества;
- разреженных решений с использованием методов планирования эксперимента для априорного разбиения выборки на части, а также дальнейшего уточнения разбиения выборки с использованием внешних критериев оценки качества моделей.

Задачи исследования. Поставленная цель диссертационной работы предопределила необходимость решения следующих основных задач:

- рассмотреть теоретические взгляды отечественных и зарубежных ученых на способы построения робастных и регрессионных моделей;
- проанализировать современное подходы построения робастных и разреженных регрессионных моделей;
- разработать новые алгоритмы построения робастных регрессионных моделей;
- разработать новые алгоритмы для построения разреженных регрессионных моделей;
- разработать новые способы разбиения выборки на обучающую и тестовую части для получения разреженных решений.

Объект исследования. Объектом исследования диссертационной работы являются методы построения регрессионных зависимостей в условиях неопределенности и нестабильности, ориентированные на получение робастных и разреженных решений.

Методы исследования. Для достижения целей и решения поставленных задач использовались методы математической статистики, теории вероятностей, математического программирования, вычислительной математики и статистического моделирования.

Предмет исследования. Разработка новых алгоритмов получения робастных и разреженных решений в регрессионном моделировании с использованием метода LS–SVM.

Теоретическая и методологическая основы исследования. Теоретической основой исследования послужили труды отечественных и зарубежных ученых по проблемам восстановления регрессионных зависимостей методом LS–SVM, разбиения выборки на части, получения робастных и разреженных решений.

Исследование было методологически обосновано путем использования таких методов, как аппарат теории вероятностей, математической статистики, вычислительной математики, математического программирования и статистического моделирования.

Научная новизна работы состоит в следующем:

- предложены новые робастные варианты критерия скользящего контроля (RLOO-P, RLOO);
- предложены новые способы получения робастных регрессионных моделей на базе метода LS–SVM с использованием метода псевдонаблюдений и взвешенного метода на основе функций потерь Хьюбера;
- предложен адаптивный вариант функции потерь Хьюбера для получения псевдонаблюдений и весовой функции потерь;
- предложены новые способы разбиения выборки на части с использованием методов планирования эксперимента для получения разреженных регрессионных моделей на базе метода LS–SVM;
- предложены новые способы (алгоритмы) разбиения выборки на части с использованием критериев оценки качества моделей для получения разреженных регрессионных моделей на базе метода LS–SVM.

Положения, выносимые на защиту:

- разработанные алгоритмы получения робастных решений с использованием методов псевдонаблюдений и взвешивания;
- предложенные робастные варианты критерия скользящего контроля и предложенный адаптивный вариант функции потерь Хьюбера;
- разработанные алгоритмы разбиения выборки на части для получения разреженного решения.

Теоретическая и практическая значимость исследования. Разработанные алгоритмы получения робастных и разреженных регрессионных моделей на основе метода LS–SVM реализованы в зарегистрированной программе для ЭВМ – «Получение робастных и разреженных решений методов LS SVM “Robast_Sparse_LS-SVM”» (свидетельство о государственной регистрации программы для ЭВМ № 2018619675 (2018 г.). М.: Федеральный институт промышленной собственности).

Результаты диссертационных исследований и разработанное программное обеспечение используются в учебном процессе, научных исследованиях и решении практических задач.

Степень достоверности результатов диссертации и обоснованность научных положений, выводов и рекомендаций подтверждены апробацией результатов диссертационной работы на научно-практических конференциях и обеспечиваются корректным применением математического аппарата и методов многомерного статистического анализа. Полученные решения при использовании предложенных методов и алгоритмов согласуются с результатами применения известных подходов на тестовых примерах и задачах.

Соответствие диссертации паспорту научной специальности

Содержание диссертации соответствует пунктам 5 – «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование

методов и алгоритмов анализа текста, устной речи и изображений», 14 – «Разработка теоретических основ создания программных систем для новых информационных технологий», 17 – «Разработка методов обеспечения обработки информации и обеспечения помехоустойчивости систем обработки данных с целью разработки новых вычислительных систем» и 18 – «Исследование и разработка моделей и алгоритмов анализа данных различной природы: текстов, устной речи и изображений с использованием регрессионного анализа, методов машинного обучения и анализа закономерностей; разработка инструментов для извлечения знаний из неструктурированной информации и моделирования эмпирического опыта» паспорта специальности 1.2.7 – Теоретические основы информатики.

Личный вклад автора заключается в:

- проведении исследований, обосновывающих основные положения, выносимые на защиту;
- программной реализации алгоритмов, описанных в диссертационной работе.

Апробация и внедрение. Основные положения и отдельные результаты диссертационной работы докладывались: на XI международном форуме по стратегическим технологиям IFOST–2016 (Новосибирск, НГТУ, 2016); на XIII международной научно-технической конференции «Актуальные проблемы электронного приборостроения (АПЭП–2016)» (Новосибирск, НГТУ, 2016); на Российской научно-технической конференции «Обработка информации и математическое моделирование» (Новосибирск, 2016); на Российской научно-технической конференции «Обработка информации и математическое моделирование» (Новосибирск, 2017); на XIV Международной научно-технической конференции «Актуальные проблемы электронного приборостроения (АПЭП–2018)» (Новосибирск, НГТУ, 2018); на Российской научно-технической конференции «Обработка информации и математическое моделирование» (Новосибирск, 2018), на научно-практической конференции «XI Ломоносовские чтения», посвященной 30-летию Государственной

независимости Республики Таджикистан (Душанбе–2021); на Международной научно-практической конференции «XIII Ломоносовские чтения», посвященной 115-летию академика Бободжона Гафурова (Душанбе–2023).

Публикации по теме диссертации. По результатам выполненных в работе исследований опубликованы 21 печатных работ, в том числе 7 статей в рецензируемых изданиях, рекомендованных ВАК при Минобрнауки РФ и ВАК при Президенте РТ, 3 статьи в изданиях, индексируемых в наукометрических системах «Scopus» и «Web of Science», 10 статей в прочих изданиях и 1 свидетельство №2018619675 о государственной регистрации программы для ЭВМ в Российской Федерации.

Структура и объем диссертации. Диссертационная работа изложена на 160 страницах и состоит из: введения; 5 разделов; заключения и списка литературы. Список литературы содержит 126 наименований. Работа иллюстрирована 53 рисунками, 32 таблицами и содержит приложения А, Б и В.

Краткое содержание работы. В первой главе диссертационной работы рассмотрены подходы к построению регрессионных моделей на основе метода LS–SVM, описаны основные концепции регрессионного анализа, ядерные функции, используемые в методе LS–SVM, представлен обзор критериев для оценки качества получаемых моделей, приведены алгоритмы построения регрессионной модели и подбора метапараметров метода LS–SVM.

Во второй и третьей главах приведены разработанные автором алгоритмы и критерии оценки качества моделей.

Во второй главе рассмотрены основные подходы к построению робастных регрессионных моделей. В качестве таких подходов рассматривались метод М–оценивания, в основе которого лежит метод псевдонаблюдений, и метод взвешивания. Для построения робастных регрессионных моделей методами псевдонаблюдений и взвешивания использовались обычная и адаптивная (предложена автором) функции потерь Хьюбера. Предложены робастные варианты критерия скользящего контроля,

при помощи которых были подобраны метапараметры алгоритма LS–SVM и оценено качество полученных результирующих робастных моделей. Также использованы функции потерь Эндрюса и биквадратной Тьюки, и сравнены эффективности и недостатки этих функций потерь с функциями потерь Хьюбера.

В третьей главе рассмотрены основные способы разбиения выборки на обучающую и тестовую части с использованием D–оптимального плана и критериев оценки качества моделей. Проведены исследования по подбору метапараметров алгоритма LS–SVM с использованием внешних критериев оценки качества моделей. Проведен сравнительный анализ эффективности использования определенных критерий для разбиения выборки на части и оценки качества полученных разреженных моделей.

В четвертой главе рассмотрены способы применения метода LS–SVM для решения практических задач. В качестве объекта исследования были использованы известные выборки LIDAR, которая использует отражение света, излучаемого лазером, для обнаружения химических компонентов в атмосфере, и Motorcycle, которая содержит результаты аварийных испытаний, проведенных с использованием манекенов, установленных на мотоциклах. Кроме того, рассмотрен способ применения метода LS–SVM для определения комплексообразования равновесной концентрации химических элементов.

В пятой главе приведен разработанный автором программный продукт с описанием функциональности и скринами основных окон, который реализован на основе алгоритмов и подходов, предложенных в предыдущих главах. Также приведены виды функции генерирования шумов с соответствующими формулами.

ГЛАВА 1. РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПО МЕТОДУ LS–SVM

В данной главе приведены основные понятия, описан алгоритм опорных векторов с квадратичной функцией потерь, дан обзор современного состояния данного направления и рассмотрены существующие проблемы подбора значений метапараметров алгоритма и подходы к их решению.

1.1 Исторический обзор

Рассмотрим один из методов, который принадлежит к классу ядерных методов – метод опорных векторов с квадратичной функцией потерь (LS–SVM), представляющий модификацию алгоритма опорных векторов (SVM).

Впервые метод опорных векторов был разработан В. Вапником [1–5] в 1995 году и использовался в задачах классификации [6]. Основная идея метода SVM заключается в построении гиперплоскости, который оптимально разделяет классы в пространстве признаков с максимальным зазором между ними. Это обеспечивает высокую обобщающую способность модели и делает метод SVM одним из самых эффективных для решения задач классификации, особенно при работе с данными небольшого объема. В 1996 году этот метод был обобщен В. Вапником, Н. Дракером, К. Берджесом, Л. Кауфманом и А. Смолой для оценивания действительных функций [7–14], который стал основой появления нового направления и данный метод был назван регрессией опорных векторов (SVR). SVR использует аналогичный принцип построения оптимальной гиперплоскости, но с учетом отклонений предсказанных значений от истинных в пределах заданного допускного интервала. Метод SVR оказался востребованным в решении задач прогнозирования и аппроксимации сложных зависимостей. В последствии метод SVM успешно применялся как для выполнения регрессионного моделирования, так и для решения задач классификации и диагностирования.

Позднее, Дж. Сайкенсом было предложено расширение SVM с использованием квадратичной функции потерь [15], которое и получило название LS–SVM. Данный метод позволяет упростить вычисления за счет решения СЛАУ вместо задачи квадратичного программирования. Также большой вклад в развитие теоретических основ SVM внесли Н. Кристианини, Дж. Шов-Тейлор, М. Айзерман, Б. Шелкопф и К. Берджес, К. Кортес и др.

Метод LS–SVM эффективно уменьшает алгоритмическую сложность, однако, в отличие от оригинального метода опорных векторов, решения, получаемые с его использованием, как правило, являются полностью плотными. В связи с этим появился ряда модификаций, направленных на достижение разреженности решений LS–SVM. Дж. Сайкенс и др. в своей работе [16] предложили сократить количество обучающих образцов с минимальным значением множителя Лагранжа. Б. де Крюи и Т. де Фрисс исключали образцы, имеющие наименьшую ошибку аппроксимации [17]. С. Цзэн и С. Чэнь представили алгоритм сокращения, который вызывает наименьшее изменение двойственной целевой функции [18]. В 2013 году С.-Л. Ся и др. был опубликован алгоритм FLSA-SVM (The Forward Least Squares Approximation SVMs), который использует число опорных векторов в качестве параметра регуляризации [19]. Данный алгоритм позволяет повысить разреженность модели и сделать её более устойчивой к шуму в данных.

Использование функции потерь в виде суммы квадратов ошибок в LS–SVM–решении привело к получению менее устойчивых оценок при наличии выбросов в данных, что является недостатком. Для обычного SVM изначально основанного на использовании функции потерь Вапника вопросы получения робастных решений рассматривались в работах [20–22]. Для того, чтобы воспользоваться наборами данных, содержащие выбросы, Дж. Сайкенс и др. предложили взвешенную LS–SVM модель (WLS–SVM) [23].

Настройка ряда внутренних параметров является одним из важных этапов построения регрессионной модели с применением метода LS–SVM [24–26]. Использование произвольных значений этих параметров в алгоритме

LS–SVM может привести к существенному изменению качества получаемых решений. Выбор оптимальных значений этих параметров критически важен, поскольку их неудачная настройка может привести к значительному ухудшению качества модели и её способности к обобщению. Современные подходы к настройке параметров включают методы перекрестной проверки, байесовскую оптимизацию и алгоритмы на основе эволюционных вычислений.

1.2 Основные понятия и определения

Пусть $x \in X \subset \mathbb{R}^d$ определяет вещественный входной вектор, а $y \in Y \subset \mathbb{R}$ – вещественную выходную переменную с совместным распределением F_{XY}. В регрессионном анализе интерес представляет нахождение измеримой функции $f : X \to Y$, такой, что $f(x)$ – «хорошая аппроксимация y». Вследствие принятия решения по выбору искомой зависимости, любые отклонения наблюдаемых величин от предсказанных означают некоторые *потери* в точности предсказаний, например, из-за случайного шума (ошибок).

В связи с этим введем в рассмотрение *функцию потерь (функция штрафа, ошибки, риска, loss function, error function)*, зависящую от выходных переменных модели и объекта:

$$L(f(x), y).$$

К настоящему времени накоплен существенный положительный опыт применения метода штрафных функций для решения ряда практических задач оптимизации. Цель его использования – количественная характеристика потерь, связанных с недостижением абсолютно точного решения задачи идентификации. В качестве штрафной функции мы можем использовать L_2 – функцию риска или среднеквадратичную ошибку f :

$$R(f) = E\left[L\left(f(x), y\right)\right] = E\left[\left(f(x) - y\right)^2\right],$$

которую необходимо минимизировать. Существует следующая причина для рассмотрения L_2 – функции риска: упрощение математического решения всей задачи, т.е. попытка свести к минимуму L_2 и риск естественным образом приводит к оценкам, которые могут быть быстро вычислены. Далее интерес представляет измеримая функция $m^* : X \to Y$ такая, что

$$m^*(x) = \arg\min_{f:\mathbb{R}^d \to \mathbb{R}} E\left[\left(f(x) - y\right)^2\right].$$

Такая функция может быть получена в явном виде следующим образом: пусть

$$m(x) = E\left[Y \mid X = x\right]$$

есть условное математическое ожидание, также известное как регрессионная функция.

Таким образом, когда наилучшее приближение построено с использованием среднеквадратической ошибки, лучшей оценкой Y в любой точке $X = x$ является условное среднее. Действительно, для произвольной $f : \mathbb{R}^d \to \mathbb{R}$ имеем

$$E\left[\left(f(X) - Y\right)^2\right] = E\left[\left(f(X) - m(X) + m(X) - Y\right)^2\right] = \\ = E\left[\left(f(X) - m(X)\right)^2\right] + E\left[\left(m(X) - Y\right)^2\right],$$

где мы используем:

$$E\left[(f(X) - m(X))(m(X) - Y)\right] = E\left[E\left[f(X) - m(X)\right](m(X) - Y) \mid X\right] = \\ = E\left[(f(X) - m(X))E\left[(m(X) - Y) \mid X\right]\right] = \\ = E\left[(f(X) - m(X))(m(X) - m(X))\right] = 0.$$

Таким образом:

$$E\left[(f(X)-Y)^2\right]=\int_{\mathbb{R}^d}(f(x)-m(X))^2dF(x)+E\left[(m(X)-Y)^2\right],$$

где первое слагаемое всегда неотрицательно и равно нулю при $f(x)=m(x)$. Поэтому $m^*(x)=m(x),$ т. е. оптимальная аппроксимация *Y* функцией от *X* задается $m(X)$.

В большинстве случаев распределение F_{XY} неизвестно. Поэтому с помощью $m(X)$ невозможно оценить *Y*. Но зачастую существует возможность получать данные в соответствии с распределением F_{XY}, а также оценить регрессионную функцию по этим данным. В задаче оценивания регрессионной функции используются данные $D_n=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ с целью получения оценки $m_n: X\to Y$ регрессионной функции. В общем случае, оценка не будет равняться регрессионной функции.

Напомним, что основной целью было найти функцию f при условии, что L_2 риск $E\left[\left(f(x)-y\right)^2\right]$ будет достаточно мал. Его минимальное значение равняется $E\left[\left(m(x)-y\right)^2\right]$, и достигает за счет регрессионной функции m. Можно показать, учитывая данные $D_n=\{(x_1,y_1),\ldots,(x_n,y_n)\}$, что L_2 риск $E\left[\left(\hat{m}_n(x)-y\right)^2\right]$ оценки $\hat{m}_n$ близок к оптимальному значению только в том случае, если L_2 ошибка

$$\int_{\mathbb{R}^d}\left(\hat{m}_n(x)-m(x)\right)^2dF(x)$$

близка к нулю. Поэтому будем использовать L_2 ошибку для определения качества оценки.

Однако, с учетом эмпирических данных $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, минимизация функционала эмпирического L_2 риска, определенного как

$$R_{emp}(f) = \frac{1}{n}\sum_{k=1}^{n}\left(f(x_k) - y_k\right)^2, \qquad (1.1)$$

ведет к бесконечному множеству решений: любая функция f_n, проходящая через обучающие точки из множества D_n, является решением. Для того чтобы получить приемлемые результаты для конечного n, необходимо «ограничить» решение (1.1) малым набором функций. Таким образом, предварительно выбираем подходящий класс функций F, вслед за этим – функцию $f : X \to Y$, где $f \in F_n$, которая минимизирует функционал эмпирического L_2 риска. Т.е. определим оценку $\hat{m}_n$ как

$$\frac{1}{n}\sum_{k=1}^{n}(\hat{m}_n(x_k) - y_k)^2 = \min_{f \in F_n}\frac{1}{n}\sum_{k=1}^{n}(f(x_k) - y_k)^2,$$

где $\hat{m}_n \in F_n$.

Напомним, что эмпирические данные $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ можно представить в виде

$$y_k = m(x_k) + e_k,\ k = 1, \ldots, n.$$

Здесь предполагается, что ошибка e в модели наблюдения имеет нулевое математическое ожидание и постоянную дисперсию σ^2, то есть, $E[e_k \mid X = x_k] = 0$ и $E[e_k^2] = \sigma^2 < \infty$, и что $\{e_k\}$ - некоррелированные величины.

1.3 Ядра и спрямляющие пространства

Пусть $x \in X \subset \mathbb{R}^d$ определяет действительный входной вектор, а $y \in Y \subset \mathbb{R}$ действительную переменную отклика, и пусть $\Psi \in \mathbb{R}^{nf}$ определяет пространство функций высокой размерности. Ключевая составляющая метода опорных векторов заключается в следующем: он отображает входной вектор $x \in X \subset \mathbb{R}^d$ в пространство признаков высокой размерности Ψ через некоторое нелинейное отображение $\varphi : X \to \Psi$. В этом пространстве мы рассмотрим класс линейных функций

$$F_\Psi = \left\{ f : f(x) = \omega^T \varphi(x) + b : \varphi : X \to \Psi, \omega \in \mathbb{R}^{nf}, b \in \mathbb{R} \right\}. \quad (1.2)$$

Однако, даже если линейная функция в пространстве признаков (1.2) хорошо обобщает данные и теоретически может быть найдена, остается проблема работы с пространством функций высокой размерности. Важно отметить, что для построения линейной функции (1.2) в пространстве признаков Ψ отсутствует необходимость рассмотрения пространства признаков в явном виде. Достаточно заменить скалярное произведение в пространстве признаков $\varphi(x_k)^T \varphi(x_l)$ на соответствующую ядерную функцию $K(x_k, x_l)$.

Предположим, что $\omega = \sum_{k=1}^{n} \beta_k \varphi(x_k)$ и класс линейных функций в пространстве признаков (1.2) имеет следующее эквивалентное представление во входном пространстве X:

$$F_X = \left\{ f : f(x) = \sum_{k=1}^{n} \beta_k K(x, x_k) + b : b \in \mathbb{R}, \beta_k \in \mathbb{R} \right\},$$

где x_k – векторы и $K(x, x_k)$ – ядерная функция.

Существует ряд ядерных функций, которые могут быть использованы в моделях метода опорных векторов, таких как: линейные, полиномиальные и радиальные базисные функции (RBF) [27].

В данной работе будут рассмотрены следующие ядра:

- линейное: $K(x,z) = x^T z + c$;
- полиномиальное: $K(x,z) = (ax^T z + c)^d$;
- RBF (гауссово): $K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$.

1.4 LS–SVM регрессия

Рассмотрим задачу восстановления зависимости по зашумленным данным. Дана обучающая выборка

$$D_n = \left\{\left(x_k, y_k\right) : x_k \in X, y_k \in Y; k = 1, \ldots, n\right\}$$

объема n наблюдений с неизвестным распределением F_{XY} вида

$$y_k = m\left(x_k\right) + e_k, k = 1, \ldots, n, \tag{1.3}$$

где $e_k \in R$ будем считать независимо и одинаково распределенной ошибкой с $E\left[e_k \mid X = x_k\right] = 0$ и $Var\left[e_k\right] = \sigma^2 < \infty$, $m\left(x\right) \in F_\Psi$ – неизвестная действительная гладкая функция и $E\left[y_k \mid x = x_k\right] = m\left(x_k\right)$, Ψ – пространство функций высокой размерности. Нашей целью является поиск параметров ω и b исходного пространства, которые минимизируют эмпирический функционал риска

$$R_{emp}\left(\omega, b\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\left(\omega^T \varphi\left(x_k\right) + b\right) - y_k\right)^2.$$

Задачу оптимизации нахождения вектора ω и смещения $b \in R$ можно свести к решению следующей задачи оптимизации [28]

$$\min_{\omega,b,e} J(\omega,e) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{n} e_k^2, \qquad (1.4)$$

в предположении, что $y_k = \omega^T\varphi(x_k) + b + e_k$, $k = 1,\ldots,n$ и где γ – параметр регуляризации.

Стоит заметить, что относительная значимость слагаемых функции затрат J определяется положительной действительной константой γ [29].

Решение задачи (1.4) обычно проводят в двойственном пространстве с использованием функционала Лагранжа

$$L(\omega,b,e,\alpha) = J(\omega,e) - \sum_{k=1}^{n}\alpha_k\left(\omega^T\varphi(x_k) + b + e_k - y_k\right),$$

с лагранжевыми множителями $\alpha_k \in R$ (опорными значениями).

Условия оптимальности задаются следующим образом:

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \to \omega = \sum_{k=1}^{n}\alpha_k\varphi(x_k); & \dfrac{\partial L}{\partial e_k} = 0 \quad \to \alpha_k = \gamma e_k,\ k = 1,\ldots,n; \\ \dfrac{\partial L}{\partial b} = 0 \to \sum_{k=1}^{n}\alpha_k = 0; & \dfrac{\partial L}{\partial \alpha_k} = 0 \to \omega^T\varphi(x_k) + b + e_k = y_k,\ k = 1,\ldots,n. \end{cases}$$

После исключения ω и e, получаем решение

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & \Omega + \dfrac{1}{\gamma} I_n \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \qquad (1.5)$$

где $y = (y_1,\ldots,y_n)^T$, $1_n = (1,\ldots,1)^T$, $\hat{\alpha} = (\hat{\alpha}_1,\ldots,\hat{\alpha}_n)^T$ и $\Omega_{kl} = \varphi(x_k)^T\varphi(x_l)$ для $k,l = 1,\ldots,n$. Результирующая LS–SVM модель имеет вид

$$y(x)=\sum_{k=1}^{n}\hat{\alpha}_k K(x,x_k)+\hat{b}, \qquad (1.6)$$

где $K(x,x_k)$ – ядро скалярного произведения,

$$\hat{b}=\frac{1_n^T\left(\Omega+\frac{1}{\gamma}I_n\right)^{-1}y}{1_n^T\left(\Omega+\frac{1}{\gamma}I_n\right)^{-1}1_n},\quad \hat{\alpha}=\left(\Omega+\frac{1}{\gamma}I_n\right)^{-1}\left(y-1_n\hat{b}\right). \qquad (1.7)$$

В случае выборок большого размера для получения оценок всех параметров вместо обращения матриц в (1.7) решают систему уравнений (1.5). Точность получаемого решения (1.6) во многом определяется настройкой внутренних параметров алгоритма LS–SVM, к числу которых относят параметр регуляризации, и параметров ядерных функций. Настройка этих параметров идет, как правило, с использованием внешних критериев качества моделей [28, 30–45].

1.5 Критерии выбора модели оптимальной сложности

В алгоритме LS–SVM, как и в других алгоритмах построения регрессионной модели, на точность итоговой модели оказывает влияние ряд задаваемых параметров. К таким параметрам можно отнести тип ядерной функции (каждая из них имеет свой собственный набор параметров) и коэффициент регуляризации γ. При необходимости все перечисленные значения можно задавать вручную, если эти параметры выбраны правильно, иначе может быть получена модель, которая не лучшим образом описывает исходную выборку. Однако, ручная настройка параметров обычно является весьма трудоемкой процедурой. Использование автоматизированного метода подбора параметров позволяет значительно упростить этот процесс, повысить качество работы алгоритма и сэкономить время. Задача автоматического подбора параметров заключается в последовательном сравнении

промежуточных результатов, полученных при фиксированных значениях параметров, и выборе наилучшего. Для автоматизации подбора параметров обычно используются критерии оценки качества моделей. К таким критериям относятся критерии: скользящего контроля, регулярности, стабильности, согласованности.

1.5.1 Критерий LOO CV

Контроль по отдельным объектам (leave-one-out cross-validation, LOO CV) является частным случаем полного скользящего контроля. Преимущество критерия LOO CV заключается в том, что каждый объект ровно один раз участвует в контроле, а длина обучающих подвыборок лишь на единицу меньше длины полной выборки. Недостатком критерия LOO CV является большая ресурсоёмкость, поскольку задачу обучения приходится решать много раз, в соответствии с объёмом исходной выборки, что сопряжено со значительными вычислительными затратами. Критерий LOO CV вычисляется по формуле:

$$\mathrm{LOO} = \sum_{i=1}^{n} \left(y_i - y_i\left(x_i\right) \right)^2,$$

где $y_i\left(x_i\right)$ – оценка параметров по полной выборке с исключенным i – ым наблюдением [46–50].

1.5.2 Критерий K-FOLD CV

Критерий регулярности K-FOLD CV состоит в том, что исходная выборка разбивается некоторое количество раз на выборки: обучающую и контрольную, объемом в K наблюдений с усреднением результатов [51, 52]. Значение данного критерия можно вычислить по формуле:

$$\text{K- FOLD} = \frac{1}{k} \sum_{i=1}^{k} \Delta_i^2\left(A_i, B_i\right).$$

Здесь значение Δ_i^2 вычисляется по следующей формуле:

$$\Delta_i^2\left(A_i, B_i\right) = \frac{1}{k}\sum_{i=1}^{k}\left(y_{test}\left(x_i\right) - y_{test}\left(x_i\right)\right)^2.$$

Кроме вышеупомянутых критериев также можно воспользоваться удобным критерием контроля точности получаемых моделей, который является среднеквадратичной ошибкой предсказания MSE, его значение вычисляется по формуле:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(u_i - y_i\right)^2,$$

где $u_i, y_i, i = 1,\ldots,n$ соответственно незашумленное и оцененное по модели значение отклика.

Алгоритм получения регрессии методом LS–SVM

1. На основе обучающей выборки $\left\{x_k, y_k\right\}_{k=1}^{N}$ производится подбор оптимального значения параметра γ и параметров выбранной ядерной функции для линейной системы (1.5) с использованием критериев оценки качества моделей (в нашем случае критерии LOO CV и K-FOLD CV).
2. При использовании выбранных оптимальных параметров вычисляются значения α_k и b по (1.7) путем решения СЛАУ (1.5).
3. По полученным значениям α_k и b строится результирующая модель

$$y_n\left(x\right) = \sum_{k=1}^{n}\alpha_k K\left(x, x_k\right) + \hat{b} \quad (1.6).$$

1.6 Подбор метапараметров алгоритма LS–SVM

Важную роль при использовании какого-либо алгоритма играет его настройка, особенно правильный выбор и настройка его параметров. В методе опорных векторов с квадратичной функцией потерь (LS–SVM) используется ряд метапараметров, которые оказывают влияние на точность итоговой модели. К таким параметрам относятся параметры выбранной ядерной функции (каждая из которых обладает собственным набором параметров) и коэффициент регуляризации γ. Например, для полиномиальной ядерной функции $K(x,z)=\left(ax^Tz+c\right)^d$ основным параметром является параметр d, а для RBF-ядра (радиально-базисные функции) $K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$ основным параметром является σ^2 [28, 30]. Значения перечисленных параметров при необходимости могут быть заданы вручную, при условии, что выбор осуществлен правильно, иначе будет получена модель, не лучшим образом аппроксимирующая исходную выборку. Ручная настройка, однако, обычно является весьма трудоёмкой процедурой. Использование автоматизированного способа подбора параметров позволяет значительно упростить этот процесс и повысить качество работы алгоритма. Порядок, в котором происходит настройка параметров, также оказывает влияние на качество работы алгоритма. Соответственно для подбора значений метапараметров и определения того факта, что полученная результирующая модель является самой оптимальной, необходимо воспользоваться критериями оценки качества модели.

Результаты работы алгоритма LS–SVM зависят от настройки таких параметров, как коэффициент регуляризации γ и параметров ядра, а также от их возможных комбинаций.

Будем считать, что численные параметры могут принимать конечное количество заранее выбранных значений. Такой подход потенциально

ограничивает точность настройки параметров алгоритма, но при этом значительно уменьшается время работы при выполнении вычислительных операций.

Как было отмечено выше, для полиномиальной и RBF ядерных функций только один из параметров будет задействован в настройке. Его значение определяется как численный, а значения всех остальных параметров фиксируются и остаются постоянными до окончания работы алгоритма. Вид ядерной функции будет жестко задан и не участвует в подборе. Для коэффициента регуляризации γ определяются границы поиска оптимального значения.

Таким образом, задача автоматической настройки определяется как выбор набора параметров, дающих наилучшее качество результатов настраиваемого алгоритма.

Эффективность и точность выбора параметров алгоритма оцениваются определенными критериями предназначенные для оценки качества моделей.

Алгоритм подбора метапараметров

1. Выбирается тип ядра K_i и соответствующий настраиваемый параметр, при условии его наличия: линейный (настраиваемый параметр отсутствует), RBF (σ^2), полиномиальный (степень полинома – d).
2. Для настраиваемых параметров предоставляется множество допустимых значений. В случае полиномиального ядра, степень полинома (d) варьируется (в нашем случае) на промежутке $[2.5;9.5]$ с шагом 0.1, а при использовании RBF-ядра, значение σ^2 варьируется на промежутке $[10^{-2};10^2]$, где шаг изменения степени равен 0.1.
3. Для всех параметров задаются начальные значения.

4. В заданных границах по результатам критерия оценки качества модели находится оптимальное значение параметра γ с использованием метода золотого сечения.
5. Для подобранного на шаге 4 значения γ ищется оптимальное значение настраиваемого параметра ядра, если таковой присутствует, иначе подбор считается завершенным.
6. Если полученные на шагах 4 и 5 значения γ и настраиваемого параметра совпадают со значениями на предыдущей итерации, подбор считается завершенным, иначе для функции ядра задается параметр с шага 5 и осуществляется переход на шаг 4.

В случае обнаружения зацикливания алгоритма следует завершить подбор и считать оптимальными подобранную на текущей итерации пару значений.

Метод подбора значений метапараметров

В качестве метода подбора значений настраиваемого параметра можно использовать любой из следующих перечисленных методов: метод золотого сечения, метод половинного деления, метод градиентного спуска, метод Ньютона и т.д. В данной работе для подбора параметров был использован метод золотого сечения [31, 32].

1.7 Исследования

Целью исследований являлось выяснение возможности выбора внутренних параметров алгоритма, а именно параметра регуляризации γ и параметров ядерных функций, опираясь на значения критериев LOO CV и K-FOLD CV. Выбор данных критериев обусловлен их способностью оценивать обобщающую способность модели с минимизацией риска переобучения и обеспечивая надежные прогнозные характеристики модели.

Для проведения исследований использовалась следующая тестовая функция: $y = 7 / (e^{(x+0.75)^2}) + 3x$, заданная на отрезке $[-1; 1]$, что позволяло

анализировать поведение модели на ограниченной области определения. Основной задачей являлось изучение влияния различных значений параметра регуляризации и параметров ядра на точность предсказаний модели. При проведении экспериментов использовались следующие фиксированные значения для параметра регуляризации γ: 1, 5, 10, 20, 50, 100, 150, 200, 250, 500. Данные значения выбирались с учетом их распространенного применения в литературе и практических задачах, что позволяло более точно определить их влияние на обобщающую способность модели. Подбор оптимального решения осуществлялся по параметру масштаба RBF–ядра, который варьировался от 10^{-1} до 10^{1} с шагом 0.1. Масштаб ядра определяет степень сглаживания модели и влияет на способность алгоритма улавливать сложные зависимости в данных. Недостаточно большое значение масштаба может привести к переобучению, в то время как слишком большое значение, наоборот, вызовет недообучение и ухудшение качества аппроксимации.

Ниже в таблице **1.1** приведены результаты выбора оптимального значения масштаба σ^2 для гауссового ядра при фиксированном значении параметра регуляризации $\gamma = 50$ [30]. В столбцах с названиями MSE, LOO CV и K-FOLD CV приведены значения: среднеквадратичной ошибки, критерия LOO CV и критерия K-FOLD CV соответственно для каждого значения параметра масштаба RBF–ядра.

Таблица 1.1 – Значения MSE, LOO CV и K-FOLD при 20% уровне шума

σ^2	MSE	LOO CV	K-FOLD CV
0,1	0,095616	0,000538	2,396715
0,125893	0,082117	0,000526	2,328368
0,158489	0,065770	0,000519	2,284321
0,199526	0,051358	0,000517	2,264950
0,251189	0,040352	0,000515	2,244617
0,316228	0,036063	0,000514	2,220304
0,398107	0,034364	0,000508	2,187641
0,501187	0,031540	0,000501	2,159995

σ^2	MSE	LOO CV	K-FOLD CV
0,630957	0,028479	0,000495	2,143648
0,794328	**0,027049**	0,000490	2,132019
1	0,027307	0,000485	2,117301
1,258925	0,029243	0,000480	2,098099
1,584893	0,034299	0,000476	2,076935
1,995262	0,046697	**0,000475**	**2,059767**
2,511886	0,072529	0,000479	2,060920
3,162278	0,110966	0,000488	2,085587
3,981072	0,158537	0,000499	2,135367
5,011872	0,221113	0,000515	2,231796
6,309573	0,320730	0,000545	2,428731
7,943282	0,494121	0,000602	2,804796
10	0,766768	0,000698	3,398345

Критерий K-FOLD CV вычислялся по схеме: усредненное по двадцати испытаниям значение ошибки прогноза в точке случайно выбираемой тестовой части выборки объемом в 10 наблюдений. Из таблицы **1.1** видно, что использование критериев качества как LOO CV, так и K-FOLD CV позволяет выбрать параметр масштаба гауссового ядра, близкий к тому, который выбирается на основе среднеквадратичной ошибки.

Качество восстановленной зависимости при выбранном $\sigma^2 = 10^{0.3}$ иллюстрируется на рисунке **1.1** [30].

Далее в таблицах **1.2**–**1.5** приведены результаты проведенных экспериментов для полиномиального и RBF-ядра со значениями параметра ядра (столбцы с названиями d и σ^2), при которых значения критерия скользящего контроля и регулярности были минимальными и при этом полученные результирующие модели были оптимальными (рисунки **1.2**–**1.5**). В столбцах с названием γ указаны значения параметра регуляризации. В экспериментах объем выборки равнялся 100, уровень шума выбирался как 10% от мощности исходной выборки. Был выбран следующий набор

возможных значений параметра регуляризации для тестирования: [0.001,0.01,0.1, 1, 10, 100, 1000, 10000,100000].

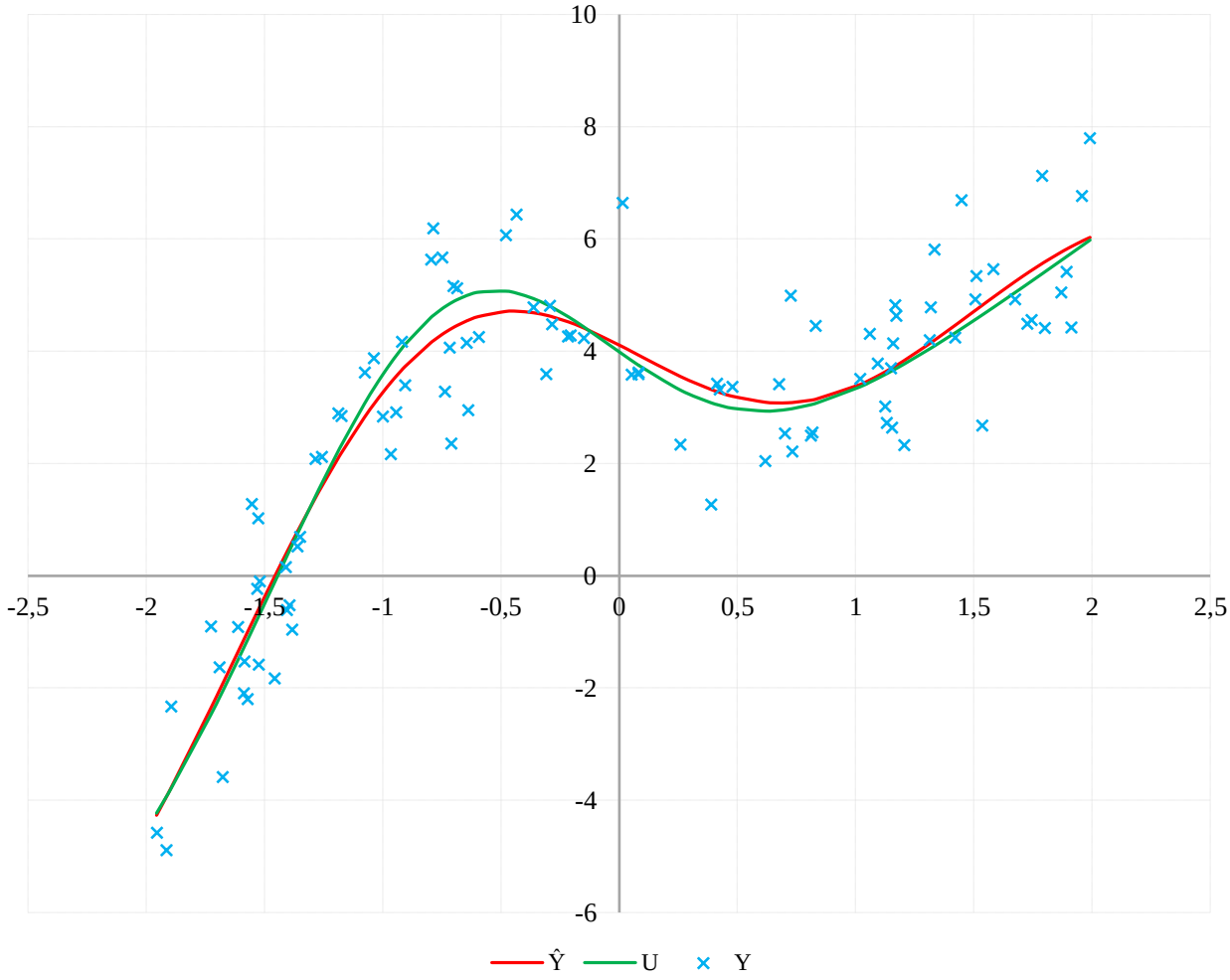


Рисунок 1.1. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по построенной модели

Таблица 1.2 – Значения LOO CV и MSE для полиномиального ядра (при 10% уровне шума)

γ	d	MSE	LOO CV
0,001	2,5	5,81809880000	6,28715948100
0,01	2,5	5,35059207900	5,78925412200
0,1	2,5	3,43872154100	3,67469889100
1	3,2	1,61994217200	1,52894318100
10	2,5	0,82907601400	0,17891527400
100	2,5	0,27279700400	0,00273542400
1000	2,5	0,18328192300	0,00002876390
10000	2,5	0,15213275200	0,00000028358
100000	2,5	0,03963974900	0,00000000238

Таблица 1.3 – Значения LOO CV и MSE для RBF-ядра (при 10% уровне шума)

γ	σ^2	MSE	LOO CV
0,001	0,25119	5,69174286800	6,16039195300
0,01	0,25119	4,40504277000	4,84097803900
0,1	0,25119	1,17524824900	1,48245869400
1	0,39811	0,11256190100	0,18072492900
10	1	0,02427337800	0,00510740500
100	1,25893	0,01788542000	0,00006054890
1000	1,99526	0,02035758700	0,00000062074
10000	2,51189	0,02026132500	0,00000000624
100000	3,98107	0,02173390300	0,00000000006

Таблица 1.4 – Значения K-FOLD CV и MSE для полиномиального ядра (при 10% уровне шума)

γ	d	MSE	K-FOLD CV
0,001	2,5	5,81809880015	0,64398972503
0,01	2,5	5,35059207907	0,38387222957
0,1	2,5	3,43872154100	0,58165398900
1	6,1	1,63164189200	0,11189477400
10	4,4	0,43421147600	0,11267343600
100	9,5	0,44928567400	0,06141019400
1000	9,5	0,14134665300	0,03141929100
10000	7,3	0,03196114000	0,04239044600
100000	3,9	0,02963846500	0,06690219800

Таблица 1.5 – Значения K-FOLD CV и MSE для RBF-ядра (при 10% уровне шума)

γ	σ^2	MSE	K-FOLD CV
0,001	0,631	5,70644645200	0,42874856100
0,01	0,1995	4,40287842600	0,28804060900
0,1	0,1995	1,16178188900	0,18082372000
1	0,631	0,14080184000	0,05394704300
10	2,5119	0,13692751700	0,01834252900
100	1,9953	0,03322894200	0,09253119200
1000	1,5849	0,01893449500	0,05024016600
10000	0,0631	0,12076364500	0,03564891300
100000	19,953	0,19636161800	0,01982553300

Стоит отметить, что среди использованных ядерных функций больше всего минимальные значения MSE были получены на основе RBF-ядра. Далее,

на рисунках приведены результаты полученных моделей на основе полиномиального и RBF-ядра.

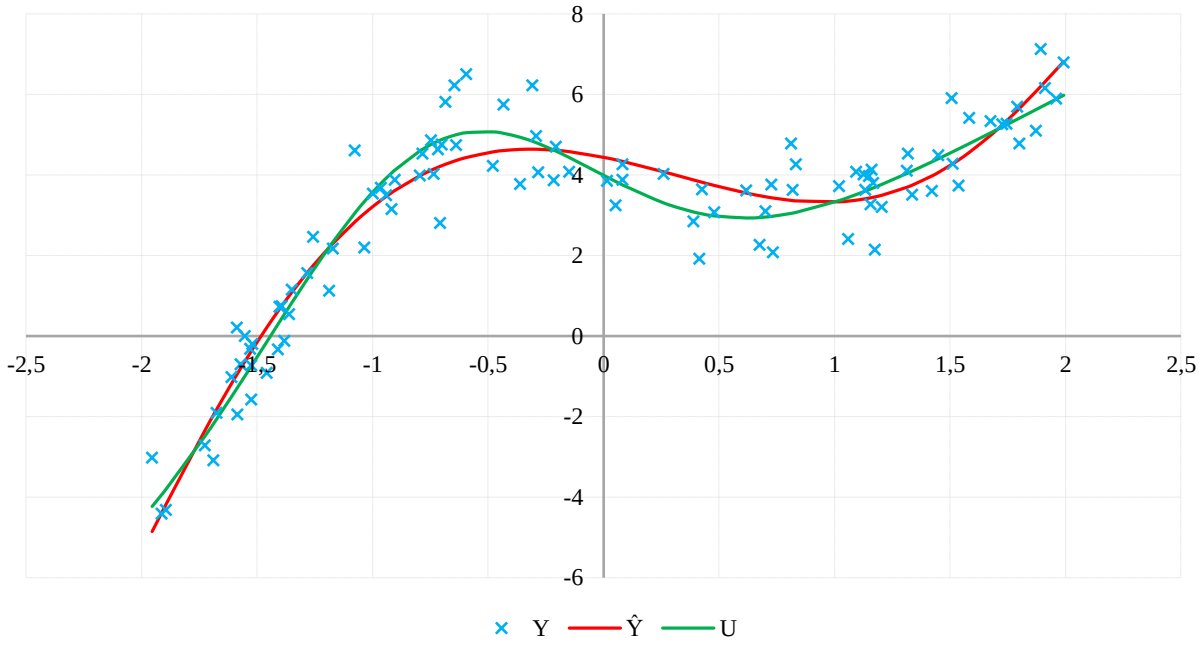


Рисунок 1.2. Результаты полученных оптимальных моделей с использованием полиномиального ядра и критерия LOO CV при γ=1000 (U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по построенной модели)

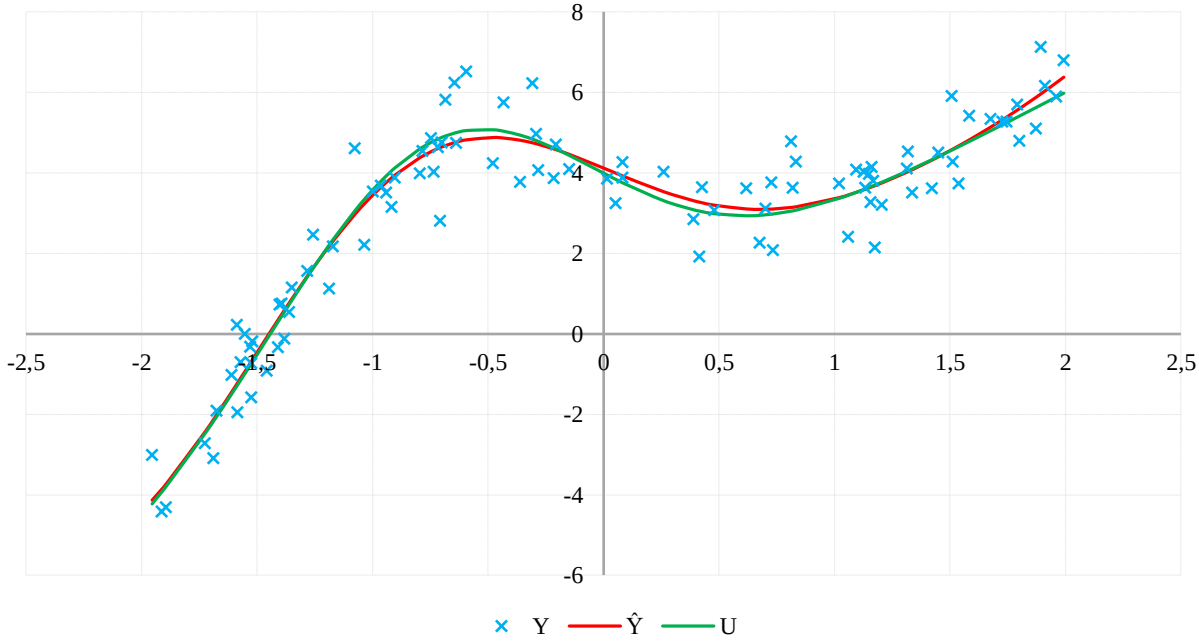


Рисунок 1.3. Результаты полученных оптимальных моделей с использованием RBF-ядра и критерия LOO CV при γ=1000 (U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по построенной модели)

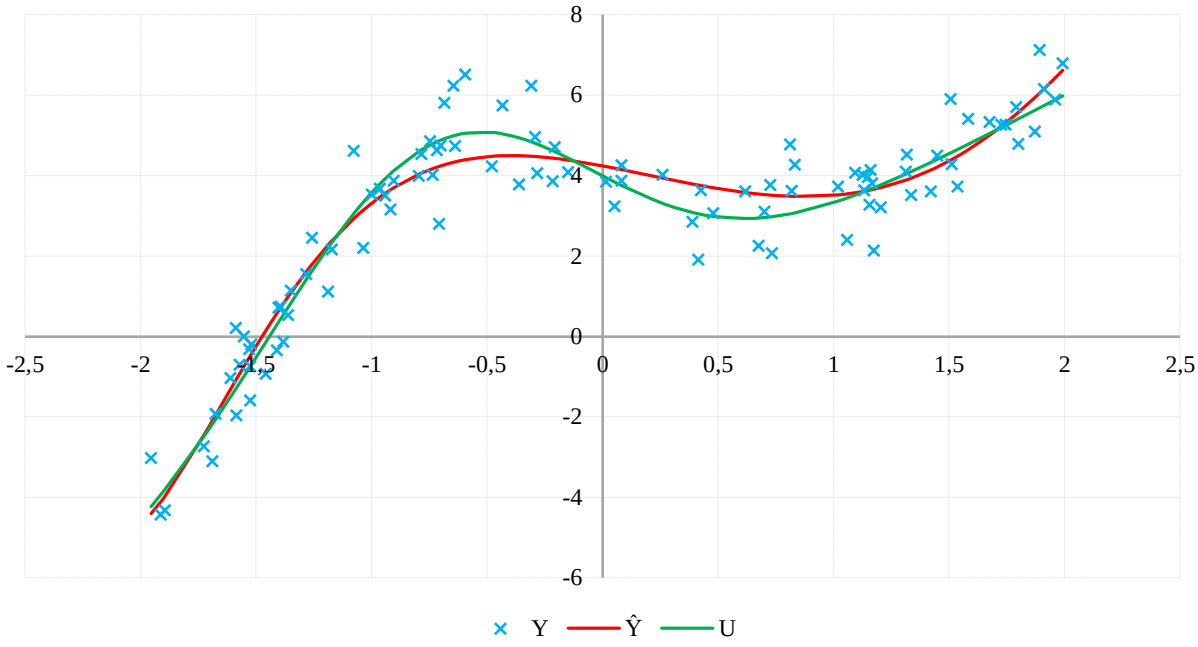


Рисунок 1.4. Результаты полученных оптимальных моделей с использованием полиномиального ядра и критерия K-FOLD CV при γ=1000 (U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по построенной модели)

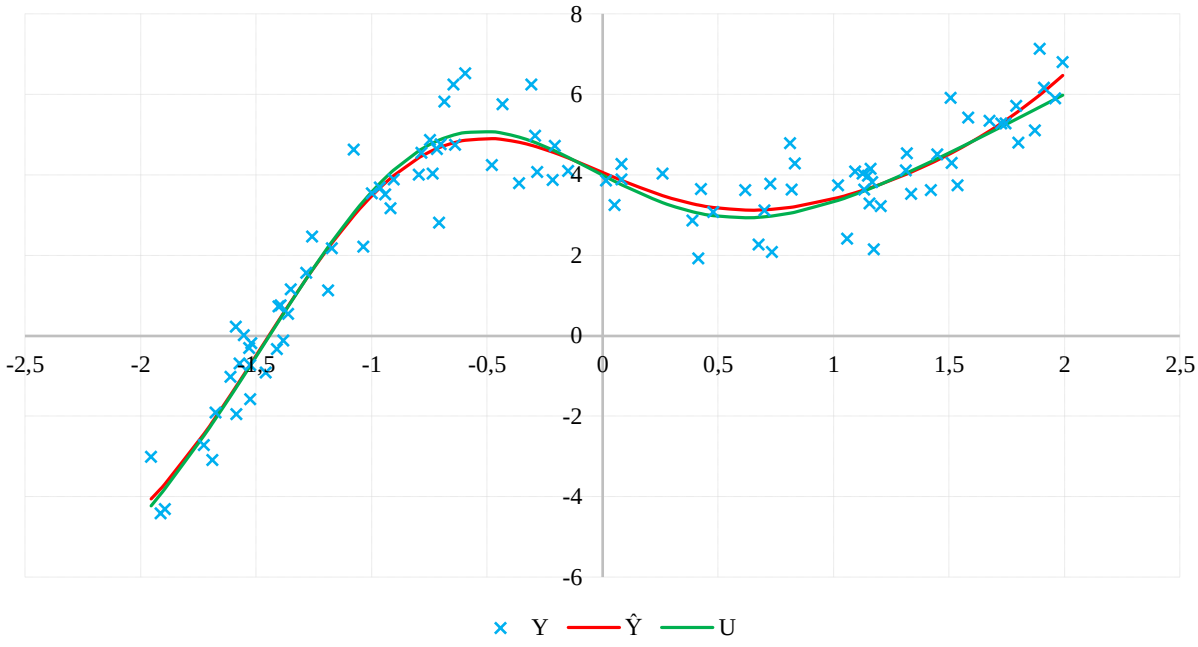


Рисунок 1.5. Результаты полученных оптимальных моделей с использованием RBF-ядра и критерия K-FOLD CV при γ=1000 (U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по построенной модели)

Из приведенных выше графиков видно, что при использовании RBF-ядра полученные модели являются наиболее оптимальными в отличие от полученных моделей на основе полиномиального ядра. Кроме того, чем больше значение коэффициента регуляризации, тем точнее может получиться результирующая модель и при этом подбирается наибольшее значение параметра ядра. Таким образом, заметно, что влияние на получаемые решения оказывает крепкая взаимосвязь всех параметров алгоритма между собой.

В рамках проведенных исследований была установлена эффективность использования критерия K-FOLD CV для подбора метапараметров алгоритма LS–SVM, так как важную роль в точности получаемых решений играет правильный выбор этих параметров. Кроме того, метапараметры алгоритма LS–SVM также можно подобрать, опираясь на критерий LOO CV. При правильном использовании перечисленных критериев можно получить модели, которые обладают хорошей обобщающей способностью.

На рисунках **1.6** и **1.7** приведены средние значения MSE полученные на основе критериев LOO CV и K-FOLD CV при использовании полиномиального и RBF – ядер для выборки с уровнями шума 5% и 10%.

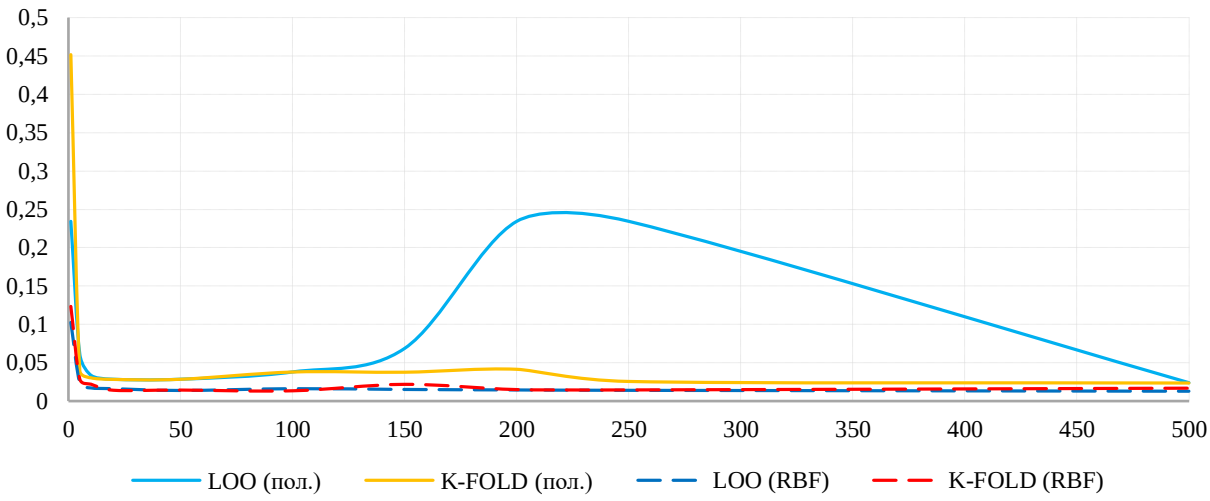


Рисунок 1.6. График средних значений MSE при использовании критериев LOO CV и K-FOLD CV полученные на основе полиномиального и RBF ядра с уровнем шума 5%

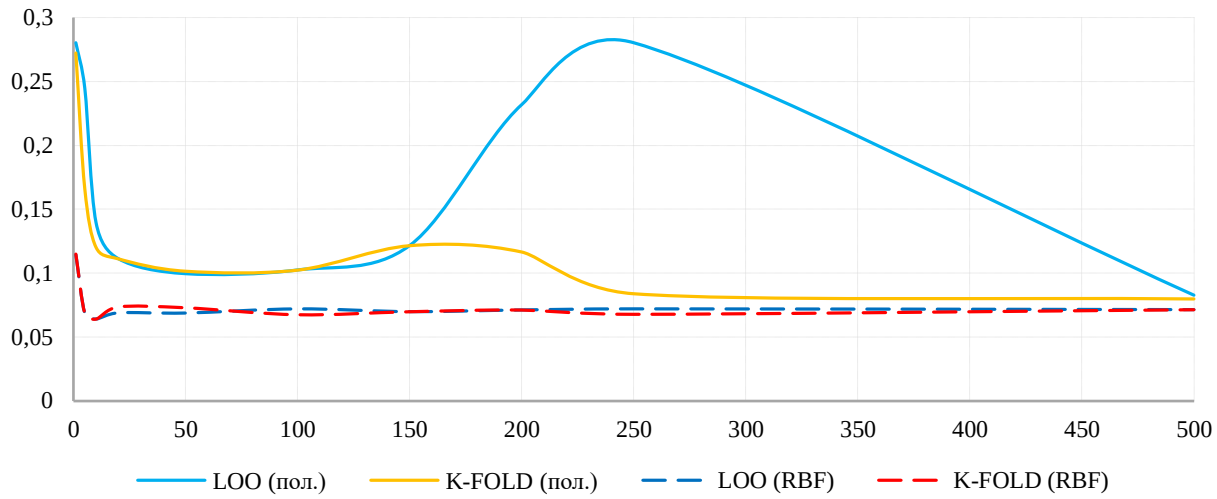


Рисунок 1.7. График средних значений MSE при использовании критериев LOO CV и K-FOLD CV полученные на основе полиномиального и RBF ядра с уровнем шума 10%

Анализ графиков показывает, что наилучшие результаты достигаются при использовании RBF-ядра. При этом следует отметить, что применение критерия K-FOLD CV позволяет получить наиболее оптимальные регрессионные модели. Этот факт подтверждает высокую эффективность данного критерия при выборе параметров модели и оценке её обобщающей способности. Таким образом, сочетание RBF-ядра и критерия K-FOLD-CV даёт наилучшие результаты, обеспечивая точность и стабильность модели даже при наличии шума в данных.

1.8 Выводы

В данной главе диссертационной работы рассмотрены способы построения регрессионных моделей на основе метода LS–SVM, описаны основные концепции регрессионного анализа, ядерные функции, используемые в методе LS–SVM. Представлен обзор критериев для оценки качества получаемых моделей, приведены алгоритмы построения регрессионной модели и подбора метапараметров метода LS–SVM. Так как настройка внутренних параметров алгоритма LS–SVM является важным этапом построения регрессии, а использование произвольных значений таких параметров могут существенно влиять на качество получаемых моделей,

критерии оценки качества получаемых моделей позволяют подбирать значения параметров алгоритма так, чтобы в итоге получить оптимальную результирующую модель. Проведены вычислительные эксперименты для модельной задачи, сравнены критерии оценки качества моделей с подбором метапараметров алгоритма, из которых очевидна актуальность решения следующих задач:

- построение регрессионных моделей на основе метода LS–SVM;
- использование критериев оценки качества моделей для подбора метапараметров метода LS–SVM.

ГЛАВА 2. РОБАСТНОЕ РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПО МЕТОДУ LS–SVM

В данной главе рассматриваются способы построения робастных регрессионных моделей с использованием методов М–оценивания и взвешивания, робастные варианты критерия скользящего контроля, адаптивный вариант функции потерь Хьюбера, а также функции потерь Эндрюса и Тьюки.

2.1 Основные понятия и определения

На протяжении последних десятилетий росло понимание того факта, что некоторые наиболее распространенные статистические процедуры весьма чувствительны к довольно малым отклонениям от предположений. Поэтому появились иные процедуры – «робастные». Робастность означает нечувствительность к малым отклонениям от предположений [53].

Оценки параметров называются робастными, если вне зависимости от присутствия больших выбросов они не меняют своих значений [54].

Если в наборе данных отсутствуют большие ошибки, то робастные оценки будут менее эффективными, но зато наиболее надежными [55, 56].

Широко известным методом получения робастных оценок параметров регрессионных моделей является метод М–оценивания [53]. Известны различные варианты реализации алгоритмов вычисления М–оценок [57], такие, например, как метод модифицированных весов и метод псевдонаблюдений. Данные методы позволяют повысить устойчивость модели к выбросам, минимизируя их влияние на итоговые результаты регрессионного анализа.

2.2 Метод М–оценивания

Метод опорных векторов с квадратичной функцией потерь (LS–SVM) является модификацией стандартного метода опорных векторов (SVM), в которой:

- Вместо решения задачи оптимизации с ограничениями применяется квадратичная функция потерь;
- Задача классификации или регрессии сводится к решению системы линейных алгебраических уравнений, что позволяет упростить вычисления.

Однако LS–SVM является чувствительным к выбросам, так как квадратичная функция потерь сильно штрафует большие отклонения. Поэтому для решения проблемы влияния больших выбросов можно применять метод М-оценок.

Метод М-оценок вводится в LS–SVM для повышения устойчивости к большим выбросам за счет использования робастной функции потерь $\rho(x)$ вместо стандартной квадратичной функции. Данный подход позволяет снижать влияние аномальных данных, улучшая обобщающую способность модели [53].

В случае стандартного LS–SVM решается задача минимизации (1.4). Соответственно при применении метода М-оценок в алгоритме LS–SVM задача минимизации примет следующий вид:

$$\min_{\omega,b,e} J(\omega,e) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{n}\rho(e_k).$$

Рассмотрим пример. Допустим, что мы получили значения y_i и остатки $r_i = y_i - y_i$.

Пусть s_i – некоторая оценка стандартной ошибки наблюдений y_i (или стандартной ошибки остатков r_i).

Метрически винсоризуем наблюдения y_i, заменяя их псевдонаблюдениями y_i^*:

$$y_i^* = \begin{cases} y_i, & |r_i| \le cs_i \\ y_i - cs_i, & r_i < -cs_i \\ y_i + cs_i, & r_i > cs_i \end{cases}.$$

Константа c регулирует степень робастности, её значения хорошо выбирать из промежутка от 1 до 2, например, чаще всего $c = 1.5$.

Далее, по псевдонаблюдениям y_i^* вычисляются новые значения y_i подгонки (и новые s_i). Действия повторяются до достижения сходимости.

Таким способом можно получить М-оценки робастной регрессии [55, 58–61].

2.3 Метод псевдонаблюдений на основе функций потерь Хьюбера

На первом этапе по методу LS–SVM с настройкой параметра регуляризации γ и параметров выбранного ядра (в нашем случае σ для RBF-ядра и d для полиномиального ядра) получаем предсказанные значения y_i, $i = 1, 2, ..., n$ [30], которые будем называть обычным решением.

Далее вычисляем остатки по формуле $r_i = y_i - y_i$ и, используя функцию потерь Хьюбера:

$$\rho\left(\frac{r_i}{s}\right) = \begin{cases} \frac{1}{2}\left|\frac{r_i}{s}\right|^2, & \left|\frac{r_i}{s}\right| \le c \\ c\left|\frac{r_i}{s}\right| - \frac{c^2}{2}, & \left|\frac{r_i}{s}\right| > c \end{cases}, \; i = 1, 2, ..., n$$

определяем псевдонаблюдения:

$$y_i^* = \begin{cases} y_i, & |r_i| \le c \\ y_i - r_i^*, & r_i < -cs \\ y_i + r_i^*, & r_i > cs \end{cases}, i = 1,2,...,n,$$

где s – робастная оценка стандартного отклонения, посчитанная, например, через медиану: $s = \underset{i}{med}|r_i| / 0.67449$ и скорректированные остатки вычисляются как квадратный корень из значения функции потерь Хьюбера для линейной зоны: $r_i^* = \sqrt{\left(2cs|r_i| - c^2 s^2\right)}$, где параметр c может меняться в пределах от 0.5 до 5.

В адаптированном варианте функции потерь Хьюбера:

$$\rho\left(\frac{r_i}{s}\right) = \begin{cases} \frac{1}{2}\left|\frac{r_i}{s}\right|^2, & \left|\frac{r_i}{s}\right| \le c \\ \tau\left(c\left|\frac{r_i}{s}\right| - \frac{c^2}{2}\right) + (1-\tau)\frac{c^2}{2}, & \left|\frac{r_i}{s}\right| > c \end{cases}, i = 1,...,n,$$

скорректированные остатки вычисляются по формуле: $r_i^* = \sqrt{\tau(2cs|r_i| - c^2 s^2) + (1-\tau)c^2 s^2}$, где параметр τ может меняться в пределах от 0 до 1.

Далее с использованием псевдонаблюдений составляем СЛАУ вида:

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & \Omega + \frac{1}{\gamma} I_n \end{bmatrix} \cdot \begin{bmatrix} \hat{b} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ y^* \end{bmatrix}. \quad (2.1)$$

Решая СЛАУ (2.1) получаем оценки $\alpha, \hat{b}$ и, используя их, построим робастную модель.

По полученной модели

$$y(x)=\sum_{k=1}^{n}\hat{\alpha}_k K(x,x_k)+\hat{b}$$

вновь вычисляем остатки и псевдонаблюдения. Эти операции повторно выполняем до тех пор, пока не выполняется условие: $\max_i \left|\frac{y_i^{(k)}-y_i^{(k-1)}}{y_i^{(k-1)}}\right| < 0.00001.$

2.4 Взвешенный метод LS–SVM на основе весовой функции Сайкенса

Сайкенс (J.A.K. Suykens) в одной из своих работ предложил для построения робастного решения использовать взвешивание наблюдений [62]. В этом случае поиск параметров ω^* и b^* в модели наблюдения $y_k=\omega^{*T}\varphi(x_k)+b^*+e_k^*,\ k=1,\ldots,n$ сводится к решению следующей задачи:

$$\min_{\omega,b,e} J(\omega^*,e^*)=\frac{1}{2}\omega^{*T}\omega^*+\frac{1}{2}\gamma\sum_{k=1}^{n}v_k e_k^{*2}. \qquad (2.2)$$

Решение задачи (2.2) обычно проводят в двойственном пространстве с использованием функционала Лагранжа

$$L(\omega^*,b^*,e^*,\alpha^*)=J(\omega^*,e^*)-\sum_{k=1}^{n}\alpha_k^*\left(\omega^{*T}\varphi(x_k)+b^*+e_k^*-y_k\right),$$

с лагранжевыми множителями $\alpha_k^*\in R$.

После исключения ω^* и e^* из условий оптимальности, получаем систему уравнений:

$$\begin{bmatrix}0 & 1_n^T\\ 1_n & \Omega+V_\gamma\end{bmatrix}\begin{bmatrix}b^*\\ \alpha^*\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix}, \qquad (2.3)$$

где диагональные элементы матрицы V_γ равны $V_\gamma = diag\left\{\frac{1}{\gamma v_1}, \frac{1}{\gamma v_2}, ..., \frac{1}{\gamma v_n}\right\}$ и значения v_k определяются с помощью весовой функции Сайкенса:

$$v_k = \begin{cases} 1, & если \left|\frac{r_i}{s}\right| \le c_1 \\ \frac{c_2 - \left|r_i / s\right|}{c_2 - c_1}, & если\, c_1 \le \left|\frac{r_i}{s}\right| \le c_2 \\ 10^{-4}, & иначе \end{cases} \quad [28, 62].$$

Решив систему (2.3) получим робастное решение по методу LS–SVM на основе весовой функции Сайкенса, которое имеет следующий вид:

$$y(x) = \sum_{k=1}^{n} \hat{\alpha}_k^* K(x, x_k) + \hat{b}^* .$$

2.5 Взвешенный метод на основе весовой функции потерь Хьюбера

Робастные решения на основе функции потерь Хьюбера можно получать с использованием взвешенного метода на основе алгоритма LS–SVM, предложенного в работах Сайкенса [28, 62]. Как и в предыдущем алгоритме, на первом этапе получаем обычное решение по LS–SVM. Для функции потерь Хьюбера определяем весовую функцию:

$$v_i = \begin{cases} 1; & |r_i| \le c \\ \frac{c}{|r_i / s|}; & |r_i| > c \end{cases},$$

где S – робастная оценка стандартного отклонения, посчитанная, например, через медиану: $s = \underset{i}{med}\,|r_i| / 0.67449$ [63–65].

Далее, составляем матрицу V_γ, элементы главной диагонали которой равны: $V_\gamma = diag\left\{\frac{1}{\gamma v_1};\frac{1}{\gamma v_2};...;\frac{1}{\gamma v_n}\right\}$ и на итерациях решаем СЛАУ:

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & \Omega + V_\gamma \end{bmatrix} \cdot \begin{bmatrix} \hat{b} \\ \hat{\alpha}_n \end{bmatrix} = \begin{bmatrix} 0 \\ y_n \end{bmatrix}. \quad (2.4)$$

Весовая функция на основе адаптивной функции потерь Хьюбера имеет следующий вид:

$$v_i = \begin{cases} 1; & |r_i| \le c \\ \dfrac{\tau c}{|r_i / s|}; & |r_i| > c \end{cases}.$$

По полученной на новой итерации модели

$$y(x) = \sum_{k=1}^{n} \hat{\alpha}_k K(x, x_k) + \hat{b}$$

вновь вычисляем остатки и скорректированные веса. Итерации выполняем до тех пор, пока не выполнятся условия останова, описанные в предыдущем пункте. Полученная модель с использованием взвешенного метода тоже является робастной.

Надо отметить, что во всех алгоритмах при получении обычных решений по методу LS–SVM для настройки внутренних параметров алгоритма можно использовать критерий скользящего контроля LOO CV [22].

2.6 Робастные решения на основе функций потерь Эндрюса и биквадратной Тьюки

Робастные решения с применением функции потерь Эндрюса и биквадратной функции потерь Тьюки строятся аналогично как в случаях с функции потерь Хьюбера. Подробно такие подходы рассматривались в работах [65, 66].

Функция потерь Эндрюса и ее весовой вид описывается следующим образом:

$$\rho\left(\frac{r_i}{s}\right)=\begin{cases} c^2\left(1-\cos\dfrac{\frac{r_i}{s}}{c}\right), & \left|\dfrac{r_i}{s}\right|<\pi c \\ 2c^2, & \left|\dfrac{r_i}{s}\right|\geq \pi c \end{cases}, \; v(z)=\begin{cases} \dfrac{1}{z}\sin\dfrac{z}{c}, & |z|<\pi c \\ 0, & |z|\geq \pi c \end{cases}.$$

В случае использования функции потерь Эндрюса для получения псевдонаблюдений скорректированные остатки вычисляются как квадратный корень для линейной зоны: $r_i^*=\sqrt{2c^2}$, где параметр c может меняться в переделах от 1.1 до 2.5.

Робастное решение на основе перечисленных функций строятся аналогично как в случае с функциями потерь Хьюбера по алгоритмам, приведенным в пунктах 2.3 и 2.5.

Аналогичные способы можно использовать и для биквадратичной функции потерь Тьюки, которые имеют следующий вид:

$$\rho\left(\frac{r_i}{s}\right)=\begin{cases} \dfrac{\left(\frac{r_i}{s}\right)^6}{6c^4}-\dfrac{\left(\frac{r_i}{s}\right)^4}{2c^2}+\dfrac{\left(\frac{r_i}{s}\right)^2}{2}, & \left|\dfrac{r_i}{s}\right|<c \\ \dfrac{c^2}{6}, & \left|\dfrac{r_i}{s}\right|\geq c \end{cases}, \; v_i=\begin{cases} \left(1-\left(\dfrac{r_i/s}{c}\right)^2\right)^2, & \text{если } \left|\dfrac{r_i}{s}\right|<c \\ 0, & \text{если } \left|\dfrac{r_i}{s}\right|\geq c \end{cases}.$$

Скорректированные остатки для составления псевдонаблюдений с использованием биквадратной функции потерь Тьюки для линейной зоны вычисляются: $r_i^*=\sqrt{c^2/6}$, где параметр c может меняться в переделах от 0.5 до 5.

2.7 Робастные критерии выбора оптимальной модели

Для оценки качества полученных робастных решений и подбора метапараметров алгоритма LS–SVM использование стандартных критериев не

дают возможность более точно оценить качество полученных робастных моделей, так как при использовании таких критериев на их значение влияют большие выбросы, имеющиеся в исходной выборке. Поэтому рассматриваем робастные критерии, которые были получены на основе критерия скользящего контроля LOO CV с использованием метода М-оценки [68, 69].

2.7.1 Критерий RLOO–P

Значение критерия LOO CV существенно зависит от имеющихся в выборке выбросов. Для уменьшения влияния выбросов при вычислении LOO CV предлагается использовать псевдонаблюдения, которые формируются на последнем шаге итерационного процесса построения М–оценок параметров. Обозначим данный критерий как RLOO–P и вычислим его по формуле

$$RLOO\text{-}P = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_r^{*}(x_i)\right)^2,$$

где $y_r^{*}(x_i)$ – прогнозное значение в точке x_i, построенное на основе робастного решения, которое было получено по методу М–оценивания, по выборке в которой отсутствовало i –е наблюдение. При этом в выборке, по которой проводилось оценивание всех n моделей, наблюдения за откликом были заменены на псевдонаблюдения.

2.7.2 Критерий RLOO

В общем случае для уменьшения влияния выбросов при вычислении критерия скользящего контроля можно использовать не квадратичную метрику, а какую–либо функцию потерь, которая применяется для построения робастных решений. Получаемый в этом случае критерий обозначим как RLOO и вычислим по следующей формуле:

$$\mathrm{RLOO} = \frac{1}{n}\sum_{i=1}^{n}\rho\left(y_i - y_r(x_i)\right).$$

В случае адаптивной функции потерь Хьюбера $\rho(r_i)$ имеет вид:

$$\rho(r_i)=\begin{cases} r_i^2 s, & \left|\frac{r_i}{s}\right| \le c \\ \tau\left(2cs|r_i|-c^2s^2\right)+(1-\tau)c^2s^2, & \left|\frac{r_i}{s}\right| > c \end{cases}.$$

2.8 Исследования

Целью исследований являлась оценка возможности получения робастных решений по методу LS–SVM с использованием методов псевдонаблюдений и взвешивания, полученные на основе обычной и адаптивной функций потерь Хьюбера, Эндрюса и биквадратной Тьюки и разработка робастных вариантов критерия скользящего контроля LOO CV. Для проведения исследования использовалась следующая тестовая функция: $y = 7/(e^{(x+0.75)^2}) + 3x$. В качестве ядер использовались полиномиальное и RBF-ядро. Уровень помехи выбирался как 5% и 10% от мощности исходной выборки и уровень засорения выбирался как 5%, 10%, 15% и 20%. Засоряющее распределение помехи имело нормальное распределение с дисперсией равной трехкратному значению дисперсии основного распределения помехи. Рассматривались варианты симметричного и несимметричного засорения. Для получения асимметричного засорения в засоряющем распределении использовалось смещение, равное 5. Количество наблюдений выбиралось равным 50. При проведении экспериментов использовались следующие фиксированные значения для параметра регуляризации γ: 1, 5, 10, 50, 100.

Ниже в таблицах **2.1** и **2.2** приведены усредненные по 10 реализациям результаты использования адаптивной функции потерь Хьюбера в LS–SVM с RBF-ядром, при зафиксированном значении $\tau = 0$ для различных значений параметров γ (параметр регуляризации), c (параметр функции потерь), σ (параметр RBF-ядра) для одного из вариантов засорения. В качестве значений

параметра σ RBF-ядра использовались $10^{-1}, 10^{-0.9}, 10^{-0.8}, ..., 10^{0.8}, 10^{0.9}, 10^{1}$ [70, 71].

Таблица 2.1 – Значения *c* и MSE, при 5% уровне шума и 20% уровне засорения

Уровень шума	Уровень засорения	γ	c	σ	MSE
5%	20%	1	7	0,398107	0,344667
		5	3,2	0,794328	0,094257
		10	1,9	1	0,057238
		50	1,4	1	0,025469
		100	1,2	1	0,020957

Таблица 2.2 – Значения c1, c2 и MSE, при 5% уровне шума и 20% уровне засорения для весовой функции Сайкенса

Уровень шума	Уровень засорения	γ	c_1	c_2	σ	MSE
5%	20%	1	4	4,5	0,398107	0,577013
		5	0,6	3,3	0,794328	0,145060
		10	4,7	4,8	1	0,054527
		50	0,2	5	1	0,038916
		100	0,2	4,5	1	0,038061

Полученные результаты показывают, что при использовании адаптивной функции потерь Хьюбера вместо функции потерь, предложенный Сайкенсом, для одних и тех же значений метапараметров алгоритма LS–SVM значения среднеквадратичной ошибки (MSE) становятся намного меньше. Это доказывает эффективность предложенного варианта адаптивной функции потерь Хьюбера для построения робастных регрессионных моделей в условиях присутствия больших выбросов в исходной выборке.

Качество восстановленных зависимостей, с использованием адаптивной функции Хьюбера и весовой функции Сайкенса при коэффициенте регуляризации $\gamma = 100$, иллюстрируются на рисунках **2.1** и **2.2** [71–73].

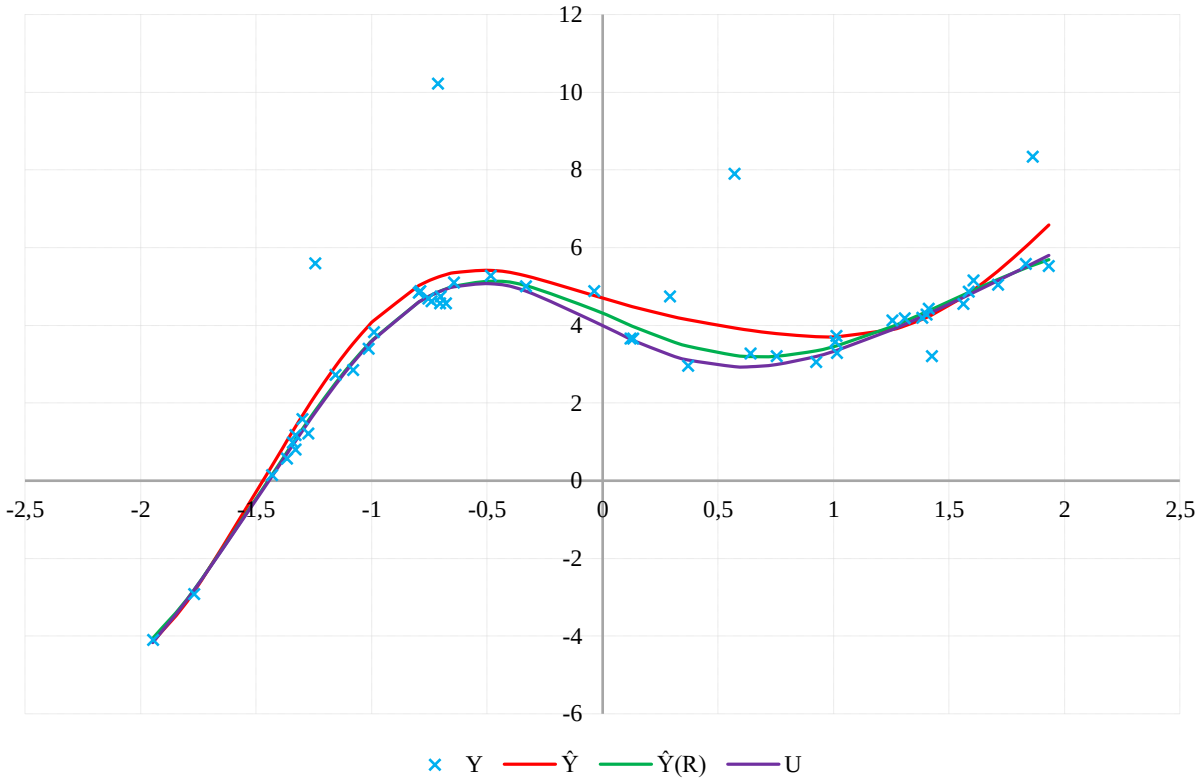


Рисунок 2.1. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(R) – робастное решение

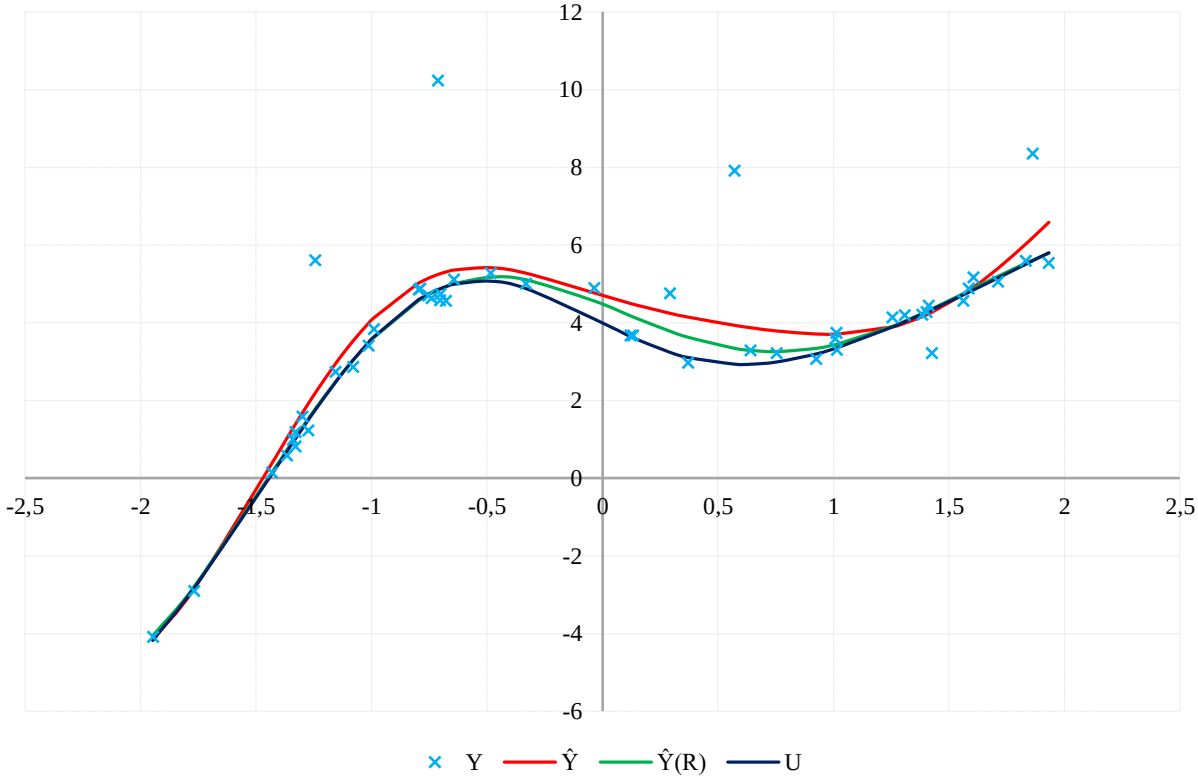


Рисунок 2.2. Графики зависимостей с использованием весовой функции Сайкенса: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(R) – робастное решение

Более подробные результаты проведенных вычислительных экспериментов с использованием обычной и адаптивной функции потерь Хьюбера в методах псевдонаблюдений и взвешиваний на основе алгоритма LS–SVM с несимметричным засорением приведены в приложении А в таблицах А.1–А.8.

Ниже в таблицах **2.3**–**2.6** приведены усредненные по 10 реализациям значения MSE при использовании обычной и адаптивной функции потерь Хьюбера в LS–SVM с RBF–ядром при зафиксированном значении $c = 1.345$ для различных значений параметра γ. Для заданного значения параметра регуляризации γ осуществлялась настройка параметра масштаба RBF–ядра по минимуму используемого критерия скользящего контроля. При сравнении полученных результатов можно видеть, что робастные варианты критерия скользящего контроля RLOO и RLOO–P позволяют выбирать внутренние параметры алгоритма LS–SVM, при которых получаемое решение имеет значительно меньшее смещение, чем при использовании обычного варианта LOO CV. При построении критерия RLOO возможно использовать известные робастные функции потерь, а не только рассмотренные в данной работе функции потерь Хьюбера.

В приведенных в таблицах **2.3**–**2.6** результатов можно увидеть, что наиболее хорошие результаты выдают решения полученные при использовании критерия RLOO. При этом критерий RLOO-P тоже показывает неплохие результаты по сравнению с обычной LOO CV.

Таблица 2.3 – Среднее значение критерия MSE для обычной (при *τ=1*) и адаптивной (при *τ=0*) функции потерь Хьюбера с 5% уровне шума и 10% уровне засорения (симметричное засорение)

Простая функция потерь Хьюбера			Адаптивная функция потерь Хьюбера		
γ	Критерий	MSE	γ	Критерий	MSE
1	LOO	0.56866	1	LOO	0.94254
	RLOO–P	0.44156		RLOO–P	0.84693
	RLOO	0.44152		RLOO	0.87027
5	LOO	1.21362	5	LOO	1.20214
	RLOO–P	0.06895		RLOO–P	0.05172
	RLOO	0.06921		RLOO	0.05324
10	LOO	1.42386	10	LOO	1.41556
	RLOO–P	0.05311		RLOO–P	0.03368
	RLOO	0.05266		RLOO	0.03505
50	LOO	0.41411	50	LOO	0.40468
	RLOO–P	0.04792		RLOO–P	0.02375
	RLOO	0.04372		RLOO	0.02337
100	LOO	0.35753	100	LOO	0.35484
	RLOO–P	0.07067		RLOO–P	0.06378
	RLOO	0.04695		RLOO	0.02971

Таблица 2.4 – Среднее значение критерия MSE для обычной (при *τ=1*) и адаптивной (при *τ=0*) функции потерь Хьюбера с 10% уровне шума и 10% уровне засорения (симметричное засорение)

Простая функция потерь Хьюбера			Адаптивная функция потерь Хьюбера		
γ	Критерий	MSE	γ	Критерий	MSE
1	LOO	0.80979	1	LOO	1.13295
	RLOO–P	0.52576		RLOO–P	0.86498
	RLOO	0.52227		RLOO	0.88751
5	LOO	1.59712	5	LOO	1.59090
	RLOO–P	0.25886		RLOO–P	0.20846
	RLOO	0.25742		RLOO	0.20804
10	LOO	1.42884	10	LOO	1.40354
	RLOO–P	0.27168		RLOO–P	0.17945
	RLOO	0.27026		RLOO	0.17800
50	LOO	0.57159	50	LOO	0.52826
	RLOO–P	0.33926		RLOO–P	0.15609
	RLOO	0.33767		RLOO	0.14289
100	LOO	0.43154	100	LOO	0.23591
	RLOO–P	0.35244		RLOO–P	0.14755
	RLOO	0.35038		RLOO	0.14274

Таблица 2.5 – Среднее значение критерия MSE для обычной (при *τ=1*) и адаптивной (при *τ=0*) функции потерь Хьюбера с 5% уровне шума и 10% уровне засорения (несимметричное засорение)

Простая функция потерь Хьюбера			Адаптивная функция потерь Хьюбера		
γ	Критерий	MSE	γ	Критерий	MSE
1	LOO	1.56640	1	LOO	1.81917
	RLOO–P	0.66466		RLOO–P	1.00344
	RLOO	0.65808		RLOO	1.00151
5	LOO	1.38660	5	LOO	1.29340
	RLOO–P	0.37796		RLOO–P	0.23374
	RLOO	0.36746		RLOO	0.18752
10	LOO	1.42670	10	LOO	1.33846
	RLOO–P	0.38186		RLOO–P	0.20068
	RLOO	0.37496		RLOO	0.15183
50	LOO	0.78629	50	LOO	0.65641
	RLOO–P	0.38275		RLOO–P	0.19350
	RLOO	0.38035		RLOO	0.13173
100	LOO	0.61613	100	LOO	0.47726
	RLOO–P	0.38933		RLOO–P	0.19762
	RLOO	0.38482		RLOO	0.13487

Таблица 2.6 – Среднее значение критерия MSE для обычной (при *τ=1*) и адаптивной (при *τ=0*) функции потерь Хьюбера с 10% уровне шума и 10% уровне засорения (несимметричное засорение)

Простая функция потерь Хьюбера			Адаптивная функция потерь Хьюбера		
γ	Критерий	MSE	γ	Критерий	MSE
1	LOO	1.42685	1	LOO	1.62151
	RLOO–P	0.66406		RLOO–P	0.92719
	RLOO	0.66209		RLOO	0.92734
5	LOO	1.54316	5	LOO	1.37960
	RLOO–P	0.39691		RLOO–P	0.16207
	RLOO	0.38711		RLOO	0.16422
10	LOO	1.52367	10	LOO	1.32574
	RLOO–P	0.39082		RLOO–P	0.11778
	RLOO	0.38083		RLOO	0.11373
50	LOO	0.65953	50	LOO	0.44159
	RLOO–P	0.41297		RLOO–P	0.11466
	RLOO	0.40500		RLOO	0.12432
100	LOO	0.51622	100	LOO	0.30157
	RLOO–P	0.40757		RLOO–P	0.11971
	RLOO	0.40464		RLOO	0.13584

В таблицах 2.7–2.10 приведены усредненные значения критериев MSE и RLOO при использовании обычной и адаптивной функции потерь Хьюбера, функции потерь Эндрюса и биквадратной Тьюки в LS–SVM с RBF–ядром при различных значений параметра регуляризации γ.

Анализ полученных результатов показывает, что наибольшую эффективность показывают обычная и адаптивная функции потерь Хьюбера. Они обеспечивают устойчивые результаты в различных случаях, что делает их предпочтительными для широкого спектра задач.

Функция потерь Эндрюса показывает хорошие результаты при небольших значениях коэффициента регуляризации в сочетании с методом взвешивания. Однако ее эффективность остается ниже по сравнению с другими функциями потерь, что ограничивает ее применение в некоторых случаях.

Биквадратичная функция потерь Тьюки показывает наилучшую эффективность при низком уровне засорения выборки в случае использования метода псевдонаблюдений. Однако при применении метода взвешивания ее результаты оказываются хуже по сравнению с другими функциями потерь, что свидетельствует о ее чувствительности к способу обработки данных.

Таблица 2.7 – Среднее значение критериев MSE и RLOO при 5% уровне шума и 5% уровне засорения, полученные методом псевдонаблюдений (несимметричное засорение)

γ	Обычная функция потерь Хьюбера		Адаптивная функция потерь Хьюбера		Функция потерь Эндрюса		Биквадратная функция потерь Тьюки	
	MSE	RLOO	MSE	RLOO	MSE	RLOO	MSE	RLOO
1	0,01248	0,57906	0,01337	0,57927	0,00784	0,58043	0,01248	0,57906
5	0,00276	0,58610	0,00511	0,57927	0,00443	0,60106	0,00212	0,58931
10	0,00185	0,58838	0,00407	0,58197	0,01528	0,60681	0,00147	0,58987
50	0,00114	0,58556	0,00261	0,58246	0,03919	0,62236	0,00083	0,59366
100	0,00160	0,58515	0,00219	0,58324	0,03918	0,62261	0,00053	0,59333

Таблица 2.8 – Среднее значение критериев MSE и RLOO при 5% уровне шума и 5% уровне засорения, полученные методом взвешивания (несимметричное засорение)

γ	Обычная функция потерь Хьюбера		Адаптивная функция потерь Хьюбера		Функция потерь Эндрюса		Биквадратная функция потерь Тьюки	
	MSE	RLOO	MSE	RLOO	MSE	RLOO	MSE	RLOO
1	0,06504	0,59211	0,06504	0,59211	0,04110	0,55249	0,16816	0,63808
5	0,03633	0,55722	0,02850	0,56878	0,04847	0,54608	0,11926	0,63451
10	0,03656	0,55412	0,02604	0,57262	0,05174	0,54261	0,11664	0,63260
50	0,04047	0,54770	0,03182	0,55841	0,04637	0,55176	0,11340	0,62956
100	0,04261	0,54467	0,03034	0,55973	0,04348	0,55272	0,11236	0,62890

Таблица 2.9 – Среднее значение критериев MSE и RLOO при 5% уровне шума и 10% уровне засорения, полученные методом псевдонаблюдений (несимметричное засорение)

γ	Обычная функция потерь Хьюбера		Адаптивная функция потерь Хьюбера		Функция потерь Эндрюса		Биквадратная функция потерь Тьюки	
	MSE	RLOO	MSE	RLOO	MSE	RLOO	MSE	RLOO
1	0,01894	0,24348	0,01327	0,24540	0,08249	0,26575	0,01896	0,24343
5	0,01022	0,23183	0,00463	0,23822	0,08640	0,25811	0,01320	0,23140
10	0,01037	0,22969	0,00414	0,23675	0,08188	0,25755	0,01454	0,22895
50	0,01280	0,22641	0,00484	0,23407	0,07272	0,25696	0,01776	0,22598
100	0,01348	0,22517	0,00504	0,23337	0,06556	0,25720	0,01874	0,22474

Таблица 2.10 – Среднее значение критериев MSE и RLOO при 5% уровне шума и 10% уровне засорения, полученные методом взвешивания (несимметричное засорение)

γ	Обычная функция потерь Хьюбера		Адаптивная функция потерь Хьюбера		Функция потерь Эндрюса		Биквадратная функция потерь Тьюки	
	MSE	RLOO	MSE	RLOO	MSE	RLOO	MSE	RLOO
1	0,06088	0,27448	0,06088	0,27448	0,02273	0,22597	0,09979	0,29117
5	0,01715	0,22904	0,01460	0,23301	0,02473	0,22523	0,05188	0,27742
10	0,01983	0,22622	0,01258	0,23298	0,02428	0,22642	0,04883	0,27507
50	0,02365	0,22313	0,01959	0,25156	0,00742	0,24366	0,04388	0,27061
100	0,02510	0,22161	0,01881	0,25085	0,00746	0,24357	0,04286	0,26963

На основе приведенных результатов можно сделать следующие общие выводы:

1. Обычная и адаптивная функции потерь Хьюбера являются наиболее универсальными и устойчивыми.
2. Функция потерь Эндрюса может быть эффективной при определенных параметрах, но уступает другим функциям.
3. Биквадратичная функция потерь Тьюки показывает высокую эффективность в специфических условиях, но ее применение требует учета метода обработки данных.

На рисунках **2.3**–**2.10** приведены результаты построения робастных решений с применением методов псевдонаблюдений и взвешивания в сочетании с различными функциями потерь. Робастные модели построены с использованием выборки в котором присутствуют шум с уровнем 5% и засорение с уровнями 5% и 10%. На данных графиках можно наблюдать влияние различных функций потерь на устойчивость и точность оценок в условиях наличия выбросов. Особое внимание уделяется сравнению эффективности методов при изменении уровня засорения выборки. Представленные результаты позволяют оценить, какие комбинации метода и функции потерь обеспечивают наилучшую робастность и минимальные искажения оценок в зависимости от характера данных.

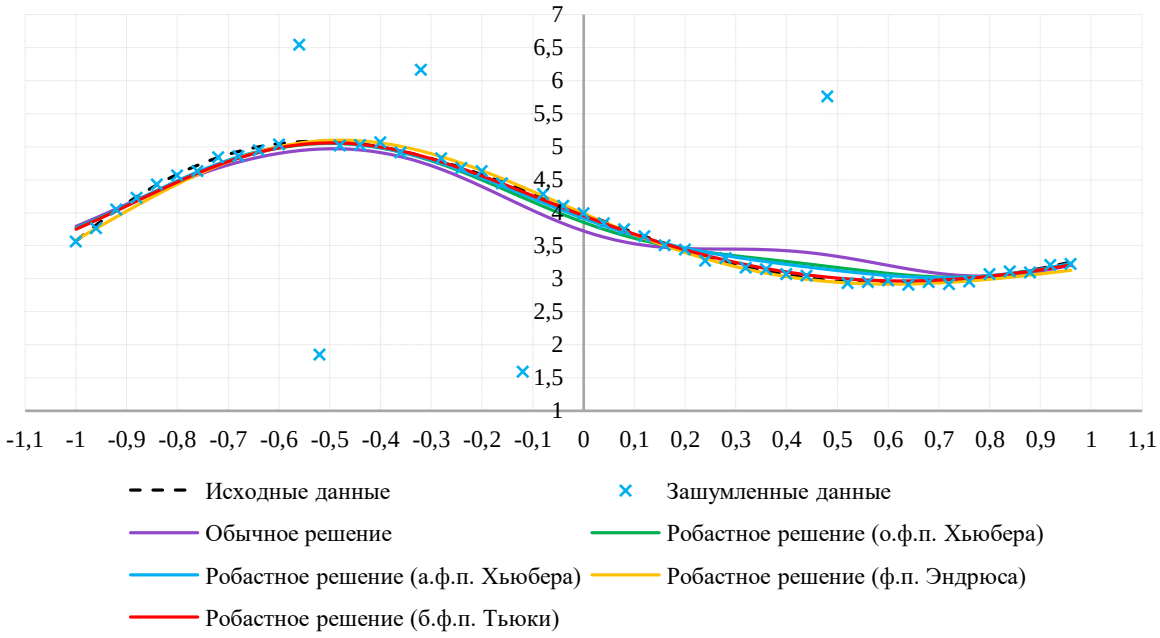


Рисунок 2.3. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 5$

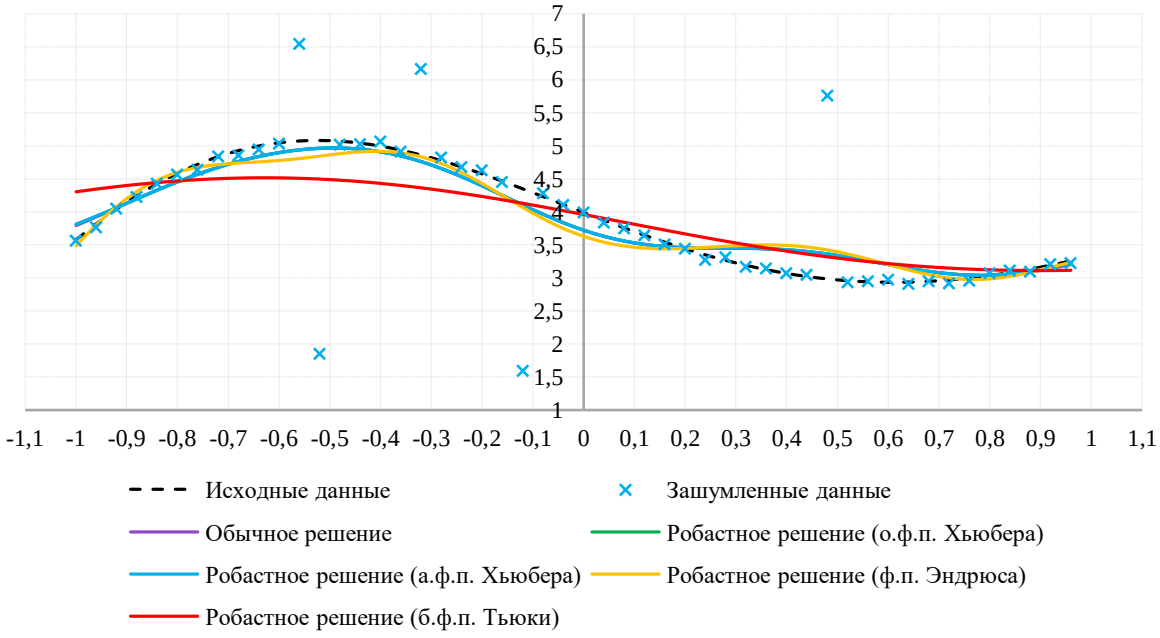


Рисунок 2.4. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 5$

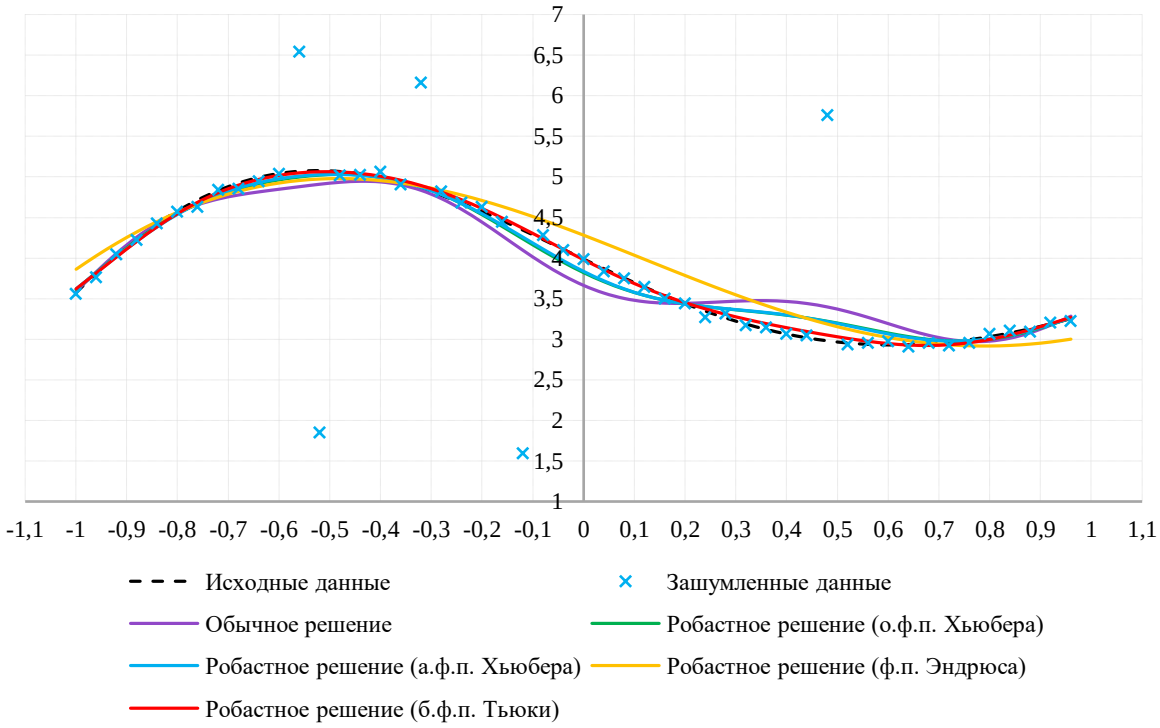


Рисунок 2.5. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 50$

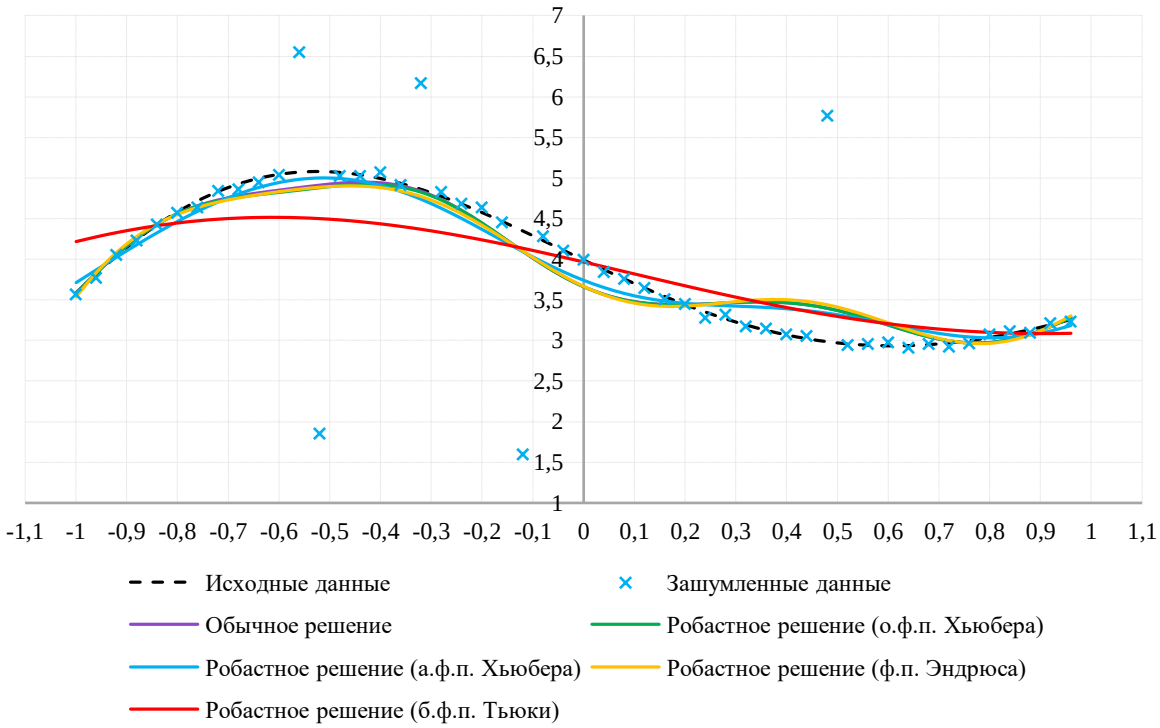


Рисунок 2.6. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 50$

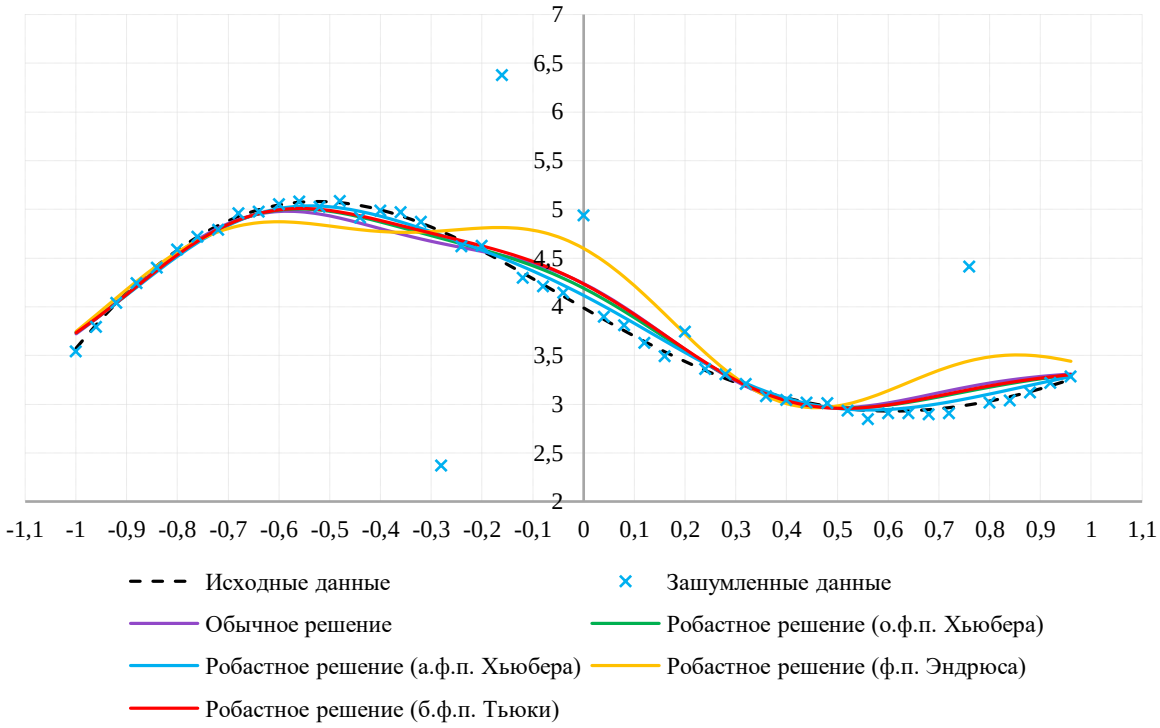


Рисунок 2.7. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 5$

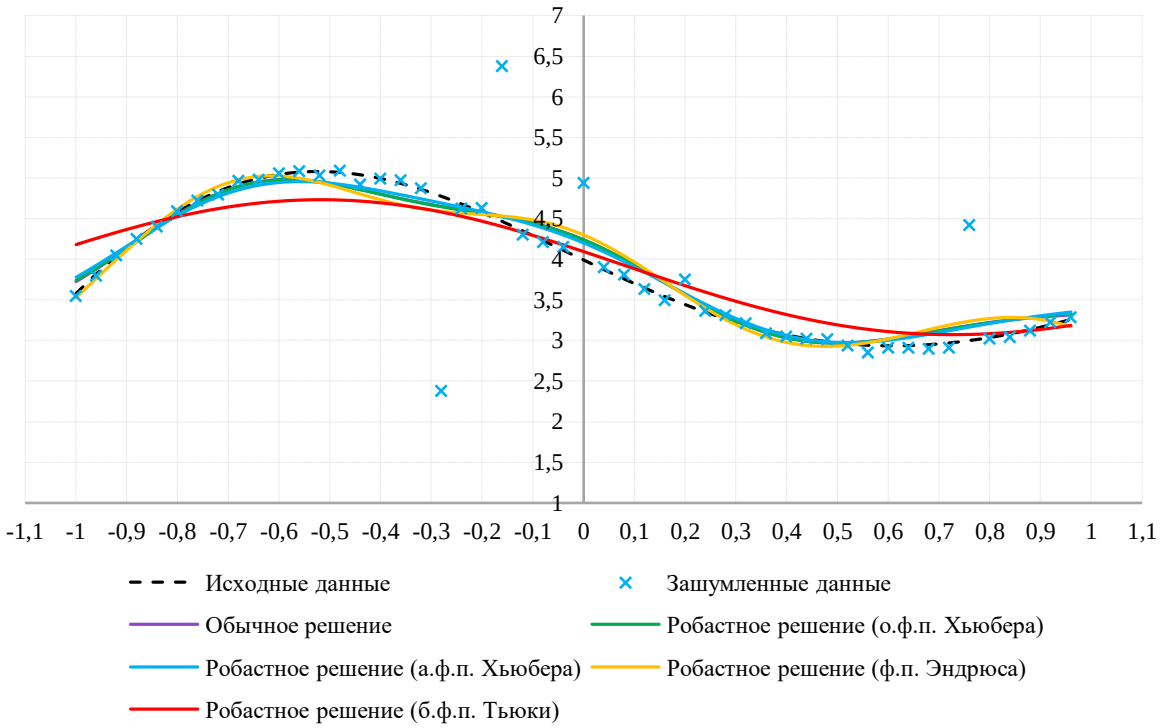


Рисунок 2.8. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 5$

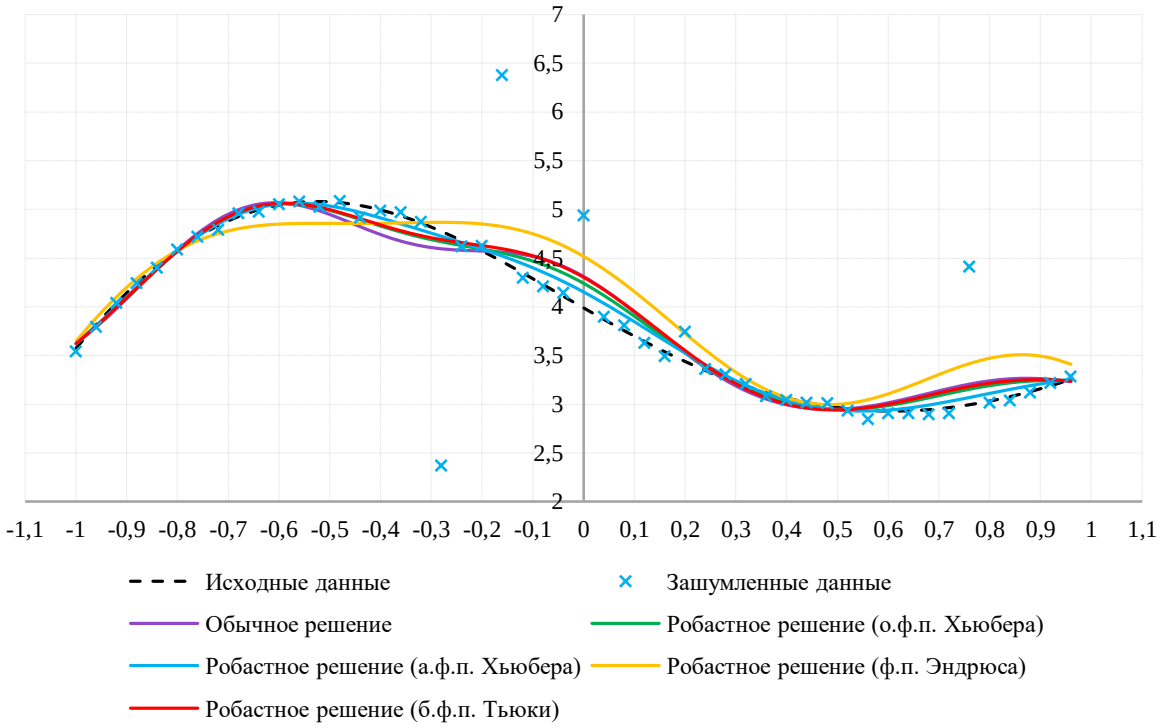


Рисунок 2.9. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 50$

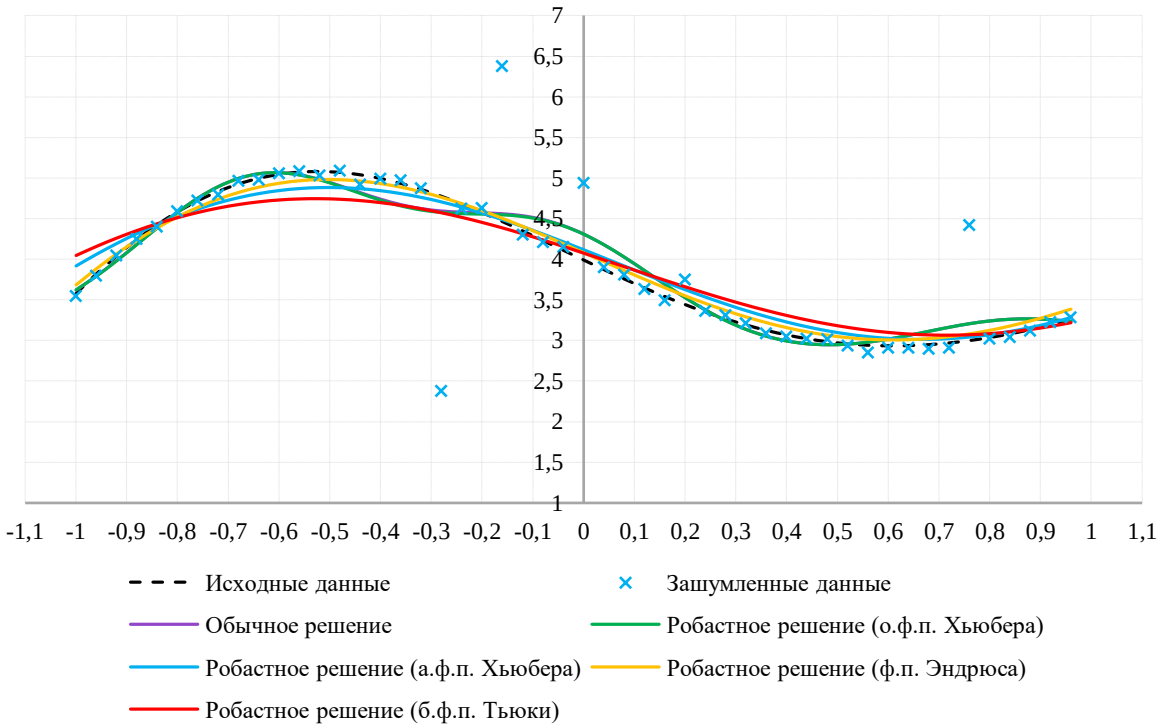


Рисунок 2.10. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 50$

Более подробные результаты приведены в виде графиков в приложении А.

Из графиков видно, что результаты, полученные с использованием обычной и адаптивной функции потерь Хьюбера являются наиболее устойчивыми. Результаты, полученные с использованием функции потерь Эндрюса являются менее эффективными, а результаты полученные на основе биквадратичной функции потерь Тьюки являются эффективными в случаях меньшего уровня засорения. Данные выводы являются подтверждением анализа результатов приведенные в таблицах 2.7–2.10.

На рисунках **2.11**–**2.14** приведены средние значения MSE в случаях присутствия симметричного и несимметричного засорения выборки, полученные с использованием критерия скользящего контроля LOO CV и ее предложенных автором робастных вариантов RLOO и RLOO-P.

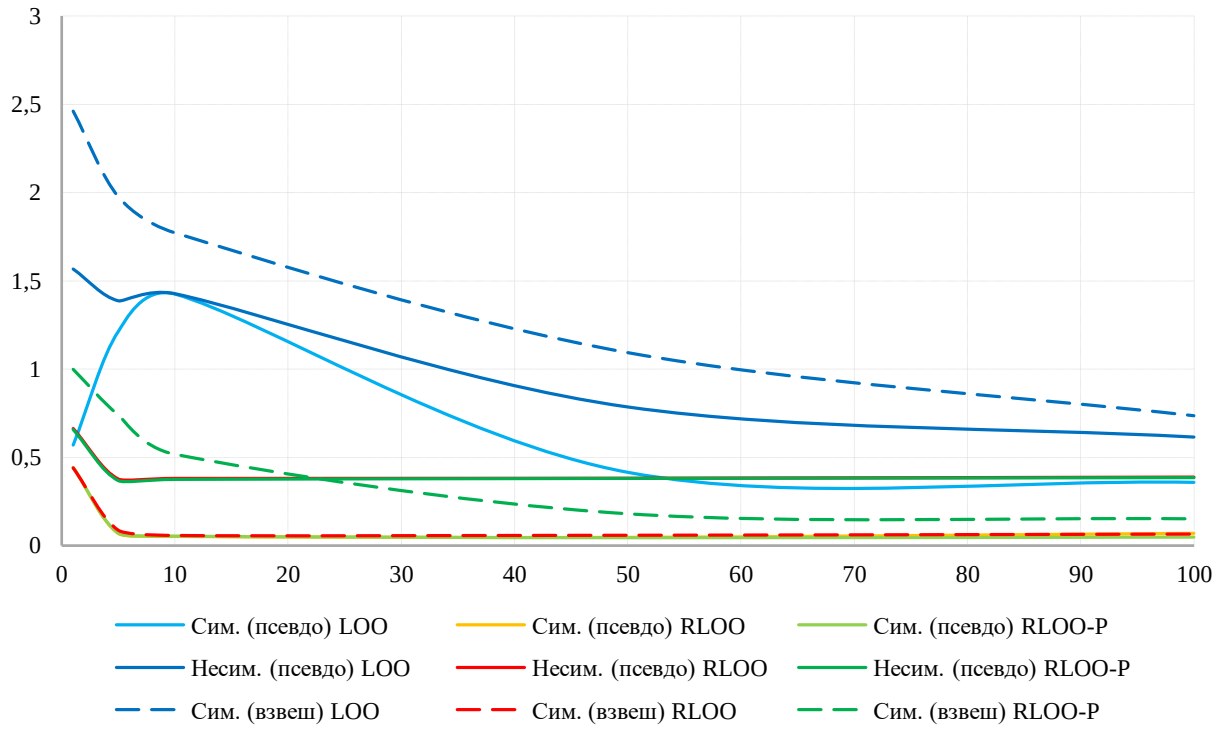


Рисунок 2.11. График средних значений MSE полученные с использованием обычной функции потерь Хьюбера при 5% уровне шума, 10% уровне засорения

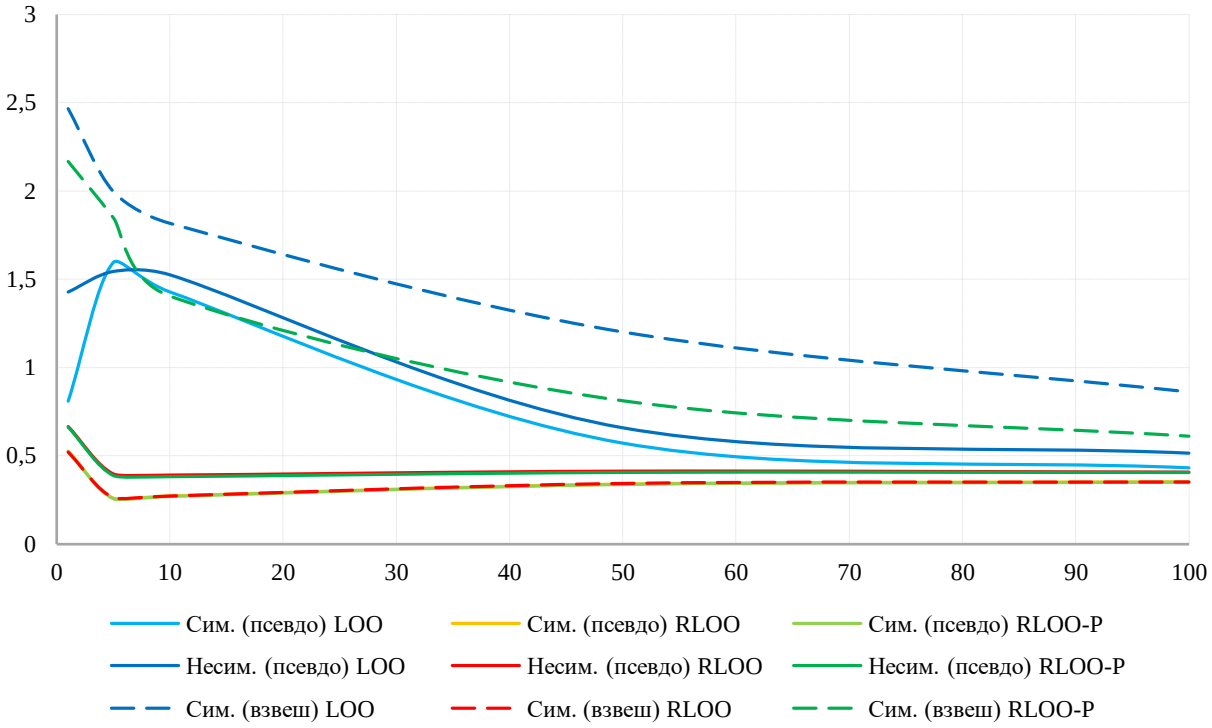


Рисунок 2.12. График средних значений MSE полученные с использованием обычной функции потерь Хьюбера при 10% уровне шума, 10% уровне засорения

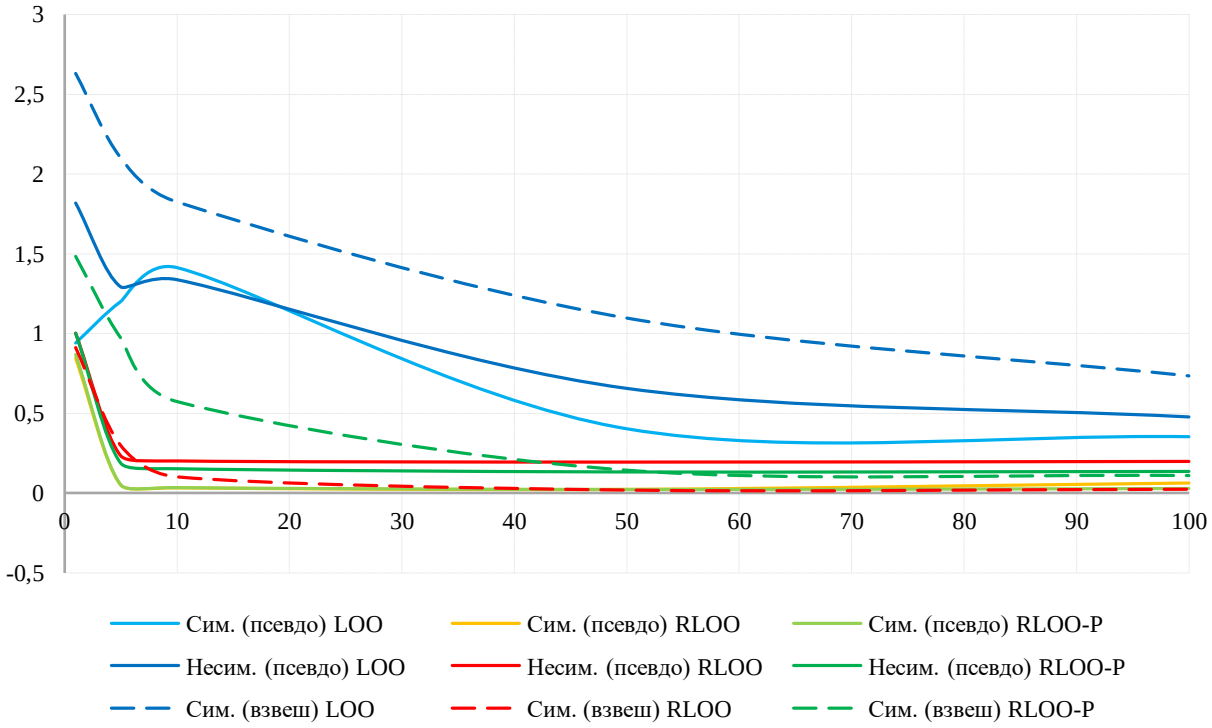


Рисунок 2.13. График средних значений MSE полученные с использованием адаптивной функции потерь Хьюбера при 5% уровне шума, 10% уровне засорения

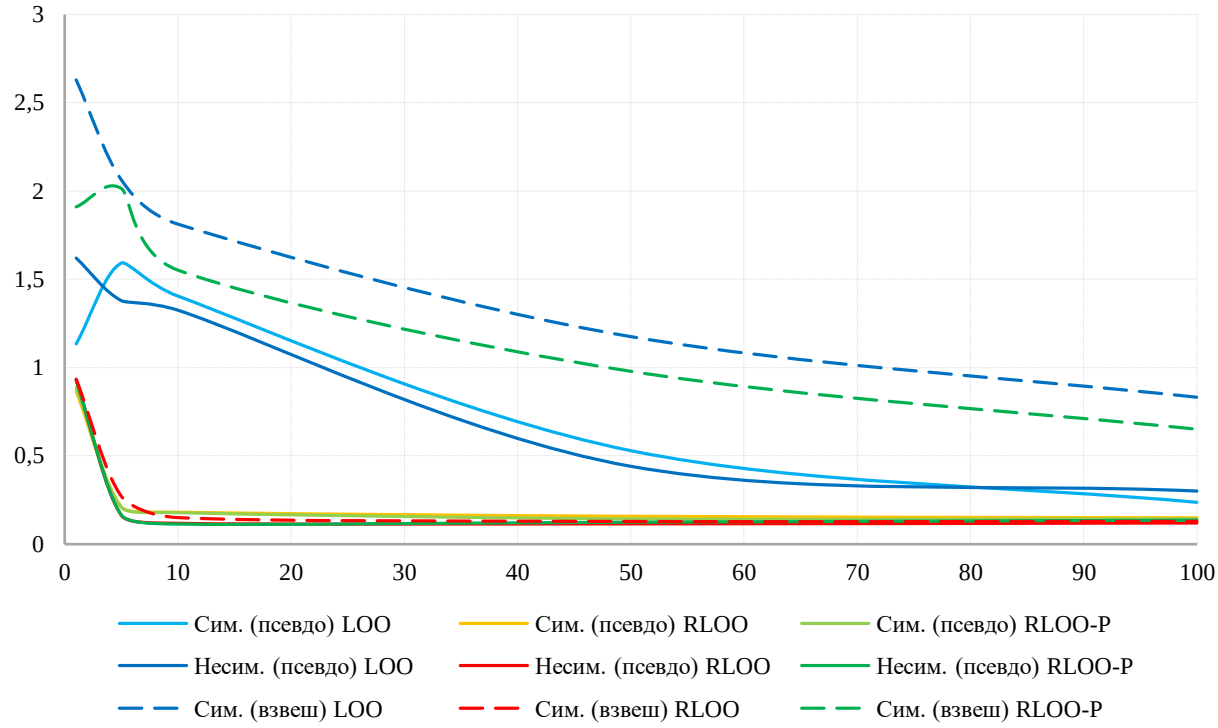


Рисунок 2.14. График средних значений MSE полученные с использованием адаптивной функции потерь Хьюбера при 10% уровне шума, 10% уровне засорения

Анализ графиков показывает, что наиболее эффективным критерием для оценки качества моделей является робастный вариант критерия скользящего контроля – RLOO. Данный критерий обеспечивает более стабильные и точные оценки по сравнению с классическим LOO-CV, особенно в присутствии выбросов и шумов в данных.

При этом критерий RLOO-P также демонстрирует хорошие результаты в отдельных случаях, превосходя традиционный LOO-CV за счет учета устойчивости к выбросам.

Кроме того, следует отметить, что метод псевдонаблюдений является эффективным инструментом для построения робастных регрессионных моделей, особенно в сочетании с функциями потерь Хьюбера. Использование данного подхода позволяет минимизировать влияние выбросов и улучшить качество аппроксимации, что делает его предпочтительным выбором для анализа данных с высоким уровнем шума.

2.9 Выводы

В данной главе рассмотрены основные подходы к построению робастных регрессионных моделей. В качестве таких подходов рассматривались метод М–оценивания, в основе которого лежат методы псевдонаблюдений, и метод взвешивания. Для построения робастных регрессионных моделей методами псевдонаблюдений и взвешивания использовались обычная и адаптивная (предложена автором) функции потерь Хьюбера. Также предложены робастные варианты критерия скользящего контроля при помощи которых были подобраны метапараметры алгоритма LS–SVM и оценены качество полученных результирующих робастных моделей.

На основании результатов проведенных исследований можно сделать следующие выводы:

1. Предложены подходы построения робастных регрессионных моделей с использованием методов псевдонаблюдений и взвешивания с использованием функций потерь Хьюбера на основе алгоритма LS–SVM.

2. Проведено исследование эффективности обычной и адаптивной вариантов функций потерь Хьюбера и весовой функции потерь Сайкенса в LS–SVM. В результате установлено, что в большинство случаев, при наличии больших выбросов в исходной выборке данных предложенный адаптивный вариант функции потерь Хьюбера дает результаты с меньшим смещением, чем другие функции потерь.

3. Предложены робастные варианты критерия скользящего контроля LOO CV. Установлена эффективность использования робастного варианта критерия скользящего контроля RLOO, который существенно увеличивает качество робастных моделей.

4. Проведены сравнительные анализы использования функций потерь Эндрюса и биквадратной функции потерь Тьюки с функциями потерь Хьюбера и установлены эффективность использования перечисленных функций потерь.

ГЛАВА 3. РАЗРЕЖЕННОЕ РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПО МЕТОДУ LS–SVM

В данной главе представлена теоретическая основа построения разреженных регрессионных моделей на основе алгоритма LS–SVM. Предложены два новых способа разбиения выборки на части: с использованием D–оптимального планирования и с использованием критериев оценки качества моделей при построения разреженных решений. Приведены алгоритмы разбиения выборки на обучающую и тестовую части.

3.1 Основные понятия и определения

Важнейшей характеристикой тех или иных методов построения регрессионных зависимостей является их способность осуществлять сжатие информации. В случае классических линейных параметрических моделей это достигается тем, что используются экономичные по числу регрессоров модели. В случае непараметрических методов построения регрессии, к коим относится и метод SVM, ситуация иная. В общем случае не удается существенно сжать информацию, поскольку в модельном описании используется значительная часть обучающей выборки. Это приводит к необходимости хранить в описании модели всю использованную для ее построения выборку. В тоже время при использовании классического SVM имеется возможность построения разреженных решений. Разреженное решение характеризуется тем, что в аддитивном его разложении по ядерным функциям задействуются не все точки выборки. Это достигается за счет использования функции потерь ε-нечувствительности Вапника. Именно те точки, которые попадают в зону ε-нечувствительности, и не участвуют в построении решения.

При использовании метода SVM с квадратичной функцией потерь для получения разреженных решений приходится прибегать к специальным приемам. Например, для их получения можно воспользоваться подходом,

предложенным в работе [62]. Основная идея при этом состоит в отбрасывании точек выборки, для которых параметры в аддитивном разложении решения по ядерным функциям имеют малые по модулю значения. Другим подходом в получении разреженных решений может быть разбиение имеющейся выборки на обучающую и тестовую части. При этом обучающая часть и составит выборку, по которой будут вычисляться решения. По отношению к полной выборке эти решения будут разреженными [74, 75]. Важно здесь и то, что получаемая тестовая выборка может быть использована для вычисления на ней критериев качества решений, по которым можно вести настройку внутренних параметров алгоритма. Эти критерии, связанные с точностью прогноза на тестовой выборке, относятся к классу внешних критериев и широко используются для выбора линейных параметрических моделей оптимальной сложности [76–87]. Разбиение выборки на обучающую и тестовую части можно проводить с использованием методов оптимального планирования эксперимента [88–96]. Данный способ получения разреженного решения предполагает использование обучающей выборки, свойства которой обеспечивают получение на ней решения, обеспечивающего наименьшую дисперсию предсказания на тестовой части выборки. Можно воспользоваться и другим способом, формируя обучающую выборку через минимизацию критериев оценки качества моделей [97–100].

3.2 Разреженное решение

Предположим, что выборка наблюдений $\left(x_k, y_k\right);\ k = 1, \ldots, n$ разбита на две части A и B. Соответственно n_A – объем обучающей выборки, а n_B – объем тестовой выборки. Тогда, для получения разреженного решения переписываем (1.5) в следующем виде:

$$\begin{bmatrix} 0 & 1_{n_A}^T \\ 1_{n_A} & \Omega_A + \frac{1}{\gamma} I_{n_A} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{\alpha}_A \end{bmatrix} = \begin{bmatrix} 0 \\ y_A \end{bmatrix}, \quad (3.1)$$

где $y_A = \left(y_1, \ldots, y_{n_A}\right)^T$ – точки из обучающей выборки, $1_{n_A} = (1, \ldots, 1)^T$, $\hat{\alpha}_A = \left(\hat{\alpha}_1, \ldots, \hat{\alpha}_{n_A}\right)$ и $\Omega_{kl} = \varphi\left(x_{Ak}\right)^T \varphi\left(x_{Al}\right)$ для $k, l = 1, \ldots, n_A$. Результирующая разреженная LS–SVM модель имеет вид:

$$y(x) = \sum_{k=1}^{n_A} \hat{\alpha}_{Ak} K\left(x, x_k\right) + \hat{b}_A,$$

где значение α_A и $\hat{b}_A$ вычисляются с помощью:

$$\hat{b}_A = \frac{1_{n_A}^T \left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-1} y_A}{1_{n_A}^T \left(\Omega A + \frac{1}{\gamma} I_{n_A}\right)^{-1} 1_{n_A}}, \quad \hat{\alpha}_A = \left(\Omega_A + \frac{1}{\gamma} I_{n_A}\right)^{-1} \left(y_A - 1_{n_A} \hat{b}\right).$$

3.3 Оптимальные планы. D–оптимальный план

Планирование оптимальных экспериментов основано на критериях оптимальности планов, однако выбор подходящего критерия представляет собой сложную задачу, не всегда поддающуюся формализации. Обычно критерий должен учитывать, как затраты на построение модели (время, материальные и финансовые ресурсы), так и потери, связанные с недостаточной точностью идентификации модели исследуемой системы.

Тем не менее, оценить затраты на построение модели в приемлемой количественной форме достаточно сложно. Поэтому на практике чаще всего используются критерии оценки «доброкачественности» модели, имеющие статистический смысл.

Однако даже такое упрощение критериев не всегда приводят к однозначным результатам. Статистические показатели модели нельзя охарактеризовать с использованием какой-то одной величины, поскольку их много. Поэтому логично требовать, чтобы выбранный план был «лучше» других в том смысле, что он имеет «минимальную» дисперсионную матрицу. Понятие «минимальной» дисперсионной матрицы должно быть характеризовано определенным функционалом от матрицы (например, определителем, следом, максимальным собственным числом и т.д.).

Определение. ***Планом*** эксперимента называется совокупность величин вида

$$\xi_N = \begin{Bmatrix} x_1, x_2, ..., x_n \\ r_1, r_2, ..., r_n \end{Bmatrix},$$

где $\sum_{i=1}^{n} r_i = N$, x_i – точка, в которой проводится r_i наблюдений, N – общее число наблюдений. Совокупность точек $x_1, x_2, ..., x_n$ называется ***спектром*** плана ξ_N.

Под ***оптимальным планированием эксперимента*** будем понимать выбор плана эксперимента в соответствии с теми или иными критериями оптимальности.

Определение. План ε^* называется D***–оптимальным,*** если

$$\varepsilon^* = Arg \max_{\varepsilon} \left| M(\varepsilon) \right|$$

или

$$\varepsilon^* = Arg \min_{\varepsilon} \left| D(\varepsilon) \right|$$ [92–94, 101–105].

3.4 Разбиение выборки с использованием D–оптимального планирования эксперимента

При рассмотрении точности оценивания модели (1.3) основное внимание будем уделять точности оценивания параметров α, убирая из рассмотрения параметр b через центрирование отклика по схеме $y^* = y - \hat{b}$, как это сделано в (1.7).

Обозначим оценки параметров α, полученные на обучающей выборке, как:

$$\hat{\alpha}_A = \left(\Omega_A + \frac{1}{\gamma} I_{n_A} \right)^{-1} \left(y_B^* \right),$$

где $\Omega_A = K\left(x_i, x_j \right),\ i, j = 1, \ldots, n_A$.

Для удобства различения точек обучающей и тестовой выборок будем обозначать координаты точек обучающей выборки через x, а координаты точек тестовой выборки через z. С учетом этого, элементы ядерной матрицы Φ_B для вычисления прогноза в точках тестовой выборки будем обозначать как:

$$(\Phi_B)_{ij} = K(z_i, x_j), i = 1, \ldots, n_B,\ j = 1, \ldots, n_A.$$

Прогнозные значения по модели, полученной на выборке A, рассчитываются как:

$$y_B = \Phi_B \hat{\alpha}_A + \hat{b}_A.$$

Ковариационная матрица ошибок прогноза на выборку B имеет вид:

$$\operatorname{cov}(y_B) = (\sigma^2 + \operatorname{cov}(\hat{b}_A)) \Phi_B (\Omega_A + \frac{1}{\gamma} I_{n_A})^{-2} \Phi_B^T + \operatorname{cov}(\hat{b}_A),$$

где $\mathrm{cov}(\hat{b}_A) = \sigma^2 \dfrac{1_{n_A}^T \left(\Omega + \dfrac{1}{\gamma} I_{n_A} \right)^{-2} 1_{n_A}}{\left[1_{n_A}^T \left(\Omega + \dfrac{1}{\gamma} I_{n_A} \right)^{-1} 1_{n_A} \right]^2}$.

Средняя дисперсия прогноза вычисляется как:

$$\bar{\sigma}^2(y_B) = (\sigma^2 + \mathrm{cov}(\hat{b}_A)) tr(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-2} \Phi_B^T \Phi_B / n_B + \mathrm{cov}(\hat{b}_A).$$

Опираясь на определение, минимизировать среднюю дисперсию $\bar{\sigma}^2(y_B)$ будем опосредовано через минимизацию определителя дисперсионной матрицы оценок параметров α. В нашем случае эта дисперсионная матрица имеет вид:

$$\mathrm{cov}(\hat{\alpha}_A) = (\sigma^2 + \mathrm{cov}(\hat{b}_A))(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-2}. \qquad (3.2)$$

Поведение $\mathrm{cov}(\hat{b}_A)$ можно рассмотреть на следующем примере. В качестве ядерной функции возьмем Гауссово (RBF ядро). В матрице Ω все диагональные элементы $k(x_i, x_i) = 1.$ Недиагональные элементы неотрицательны. Сумма всех элементов матрицы $\left(\Omega + \dfrac{1}{\gamma} I_{n_A} \right)^{-1}$ достигает минимального значения в том случае, когда ее недиагональные элементы близки к нулю. Это возможно, когда параметр масштаба гауссовой ядерной функции выбран достаточно большим, или, когда точки выборки расположены на достаточно большом расстоянии друг от друга. Будем считать, что это так, тогда

$$\mathrm{cov}(\hat{b}_A) \cong \sigma^2 \frac{n_A(\gamma/(1+\gamma))^2}{n_A^2(\gamma/(1+\gamma))^2} = \frac{\sigma^2}{n_A}.$$

Именно такой дисперсией обладает оценка параметра b в виде среднего $\tilde{b} = 1^T y / n_A$. Учитывая, что

$$\left|(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-2}\right| = \left|(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-1}\right| * \left|(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-1}\right|$$

и то, что матрица $(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-1}$ положительно определена, будем рассматривать минимизацию определителя $\left|(\Omega_A + \frac{1}{\gamma} I_{n_A})^{-1}\right|$ или, что намного проще – максимизацию определителя $\left|(\Omega_A + \frac{1}{\gamma} I_{n_A})\right|$. Для определителя положительно определенной матрицы известно свойство, что он меньше либо равен произведению диагональных ее элементов. Поскольку все диагональные элементы матрицы $\Omega_A + \frac{1}{\gamma} I_{n_A}$ равны $(\gamma + 1)/\gamma$, можно заключить, что максимум определителя достигается при диагональной матрице Ω_A. При этом $\mathrm{cov}(\hat{b}_A) = \frac{\sigma^2}{n_A}$. Таким образом в целях упрощения задачи минимизации определителя матрицы ковариации (3.2) будем решать задачу максимизации определителя $\left|(\Omega_A + \frac{1}{\gamma} I_{n_A})\right|$. Тем самым мы будем строить дискретный D–оптимальный план объёмом в n_A наблюдений, используя все точки имеющейся выборки.

В нашем случае для построения дискретного D–оптимального плана удобно воспользоваться хорошо себя зарекомендовавшими последовательными алгоритмами [93, 94].

Обозначим через G_s матрицу размером $s \times s$ для обучающей выборки объемом в s наблюдений и состоящую из элементов

$$(G_s)_{ij} = K(x_i, x_j) + \frac{1}{\gamma} I_s, i, j = 1, \ldots, s.$$

Тогда на шаге $s+1$ матрица G_{s+1} будет иметь вид:

$$G_{s+1} = \begin{pmatrix} G_s & F(x_{s+1}) \\ F^T(x_{s+1}) & K(x_{s+1}, x_{s+1}) + \frac{1}{\gamma} \end{pmatrix},$$

где $F^T(x_{s+1}) = (K(x_1, x_{s+1}), K(x_2, x_{s+1}), \ldots, K(x_s, x_{s+1}))$.

Определитель окаймленной матрицы легко вычисляется:

$$|G_{s+1}| = |G_s| * \Delta(x_{s+1}),$$

где $\Delta(x_{s+1}) = [K(x_{s+1}, x_{s+1}) + \frac{1}{\gamma} - F^T(x_{s+1}) G_s^{-1} F(x_{s+1})]$.

Таким образом, очередная точка, включаемая в обучающую выборку, отыскивается по следующей схеме:

$$x_{s+1} = Arg \max_x \Delta(x),$$

где аргумент x принимает значения координат точек исходной выборки, еще не включенных в обучающую часть [90, 106, 107].

3.5 Разбиение выборки с использованием внешних критериев оценки качества моделей

3.5.1 Внешние критерии оценки качества моделей

Рассматриваем другие пути разбиения выборки на части с использованием внешних критериев качества моделей [90, 91]. К таким критериям относятся критерии:

- регулярности:

$$\Delta^2 = (y_B - \hat{y}_B)^T (y_B - \hat{y}_B) / n_B,$$

где $y_B = \Phi_B \hat{\alpha}_A + \hat{b}_A$ – прогнозные значения по модели, оцененной на обучающей выборке, Φ_B – матрица элементов ядерной функции по тестовой выборке;

- стабильности:

$$S^2 = (y - \hat{y}_A)^T (y - \hat{y}_A) + (y - \hat{y}_B)^T (y - \hat{y}_B),$$

где $y_A = \Phi \hat{\alpha}_A + \hat{b}_A$ – прогнозные значения на часть A выборки по модели, оцененной на обучающей выборке, Φ – матрица элементов ядерной функции по всей выборки;

- согласованности (вариативности):

$$V^2 = (\hat{y}_A - \hat{y})^T (\hat{y} - \hat{y}_B).$$

Для разбиения выборки на обучающую и тестовую части с использованием внешних критериев качества моделей Δ_K (регулярности, стабильности, согласованности и др.) приведем несколько алгоритмов. С помощью этих алгоритмов разбиение выборки на части выполняется путем включения, исключения и заменой точек в обучающей части [108–113].

3.5.2 Вариант включения

1. Выполняем предварительное оценивание параметра той или иной ядерной функции с использованием критерия скользящего контроля (LOO).

2. Последовательно, по схеме наращиваем объем обучающей выборки до заданного объема. Каждая новая точка, включаемая в обучающую часть, выбирается по минимуму выбранного критерия Δ_K. Периодически или после каждого включения новой точки производим подстройку параметров ядерной функции. Новое значение параметра ядерной функции выбираем в том случае, если при нем удается получить лучшее значение выбранного критерия Δ_K.

3.5.3 Вариант исключения

1. Выполняем предварительное оценивание параметра той или иной ядерной функции с использованием критерия LOO.

2. Последовательно, по схеме наращиваем объем тестовой выборки, т.е. исключаем точки из обучающей выборки. Каждая новая точка, включаемая в B, выбирается по минимуму выбранного критерия Δ_K. Точки, не включенные в B, образуют обучающую выборку. Периодически или после каждого включения новой точки производим подстройку параметров ядерной функции. Новое значение параметра ядерной функции выбираем в том случае, если при нем удается получить лучшее значение выбранного критерия Δ_K.

3.5.4 Вариант замены

1. Выполняем предварительное оценивание параметра той или иной ядерной функции с использованием критерия LOO.

2. Выполняем D–оптимальное разбиение исходной выборки на части A и B с числом наблюдений $n_A + n_B = n$. Выполняем поиск оптимального значения параметров ядерных функций по выбранному критерию Δ_K.

3. Алгоритм последовательной замены по уточнению состава обучающей выборки A. Точки из выборки A поочередно заменяем точками выборки B. После каждой замены вычисляем значение выбранного критерия, который будем обозначать как $\Delta_K \pm$, а значение критерия до замены обозначим как Δ_K. Если окажется, что $\Delta_K \pm < \Delta_K$, то данную замену оставляем в силе и присваиваем Δ_K значение, равное $\Delta_K \pm$. Если удачных замен нет, то останов. После шага 3 состав точек в обучающей выборке может измениться. Можно попытаться улучшить получаемое разреженное решение, вводя дополнительный шаг.

4. Выполняем поиск оптимального значения параметров ядерных функций по выбранному критерию Δ_K с использованием обучающей выборки, полученной на шаге 3.

3.5.5 Вариант Add/Del

1. Выполняем предварительное оценивание параметра используемой ядерной функции с использованием критерия LOO.

2. Выполняем последовательную схему наращивания объема обучающей выборки до заданного объема. При этом при включении в обучающую выборку более 2-х точек попытаемся заменить какую-то из ранее включенных на точку из тестовой выборки. Перебираем все ранее включенные точки. Эту позицию будем называть Del. Если удается найти хорошую замену, то ее оставляем в силе и переходим к очередному шагу добавления новой точки в обучающую выборку. Эту позицию будем называть Add. Добавление и удаление точек контролируем по выбранному критерию Δ_K. Также, периодически или после каждого включения новой точки производим подстройку параметров ядерной функции. Новое значение параметров ядерной функции выбираем в том случае, если удается получить при нем лучшее значение выбранного критерия Δ_K.

3.5.6 Вариант Del/Add

1. Выполняем предварительное оценивание параметра используемой ядерной функции с использованием критерия LOO.

2. Выполняем последовательную схему наращивания объема тестовой выборки до заданного объема. При этом при включении в тестовую выборку более 2-х точек попытаемся заменить какую-то из ранее включенных на точку из обучающей выборки. Перебираем все ранее включенные точки. Эту позицию будем называть Add. Если удается найти хорошую замену, то ее оставляем в силе и переходим к очередному шагу добавления новой точки в тестовую выборку. Эту позицию будем называть Del. Добавление и удаление

точек контролируем по выбранному критерию Δ_K. Также, периодически или после каждого включения новой точки производим подстройку параметров ядерной функции. Новое значение параметров ядерной функции выбираем в том случае, если удается получить при нем лучшее значение выбранного критерия Δ_K.

Алгоритм получения разреженной LS–SVM регрессии

1. На основе обучающей выборки $\left\{x_k, y_k\right\}_{k=1}^{N}$ производится подбор оптимального значения параметра γ и параметров выбранной ядерной функции для линейной системе (1.5).
2. При использовании выбранных оптимальных параметров вычисляются значения α_k и b по (1.7).
3. Производится разбиение выборки на обучающую и тестовую части по выбранному способу (*D*–оптимальным планированием или критерием оценки качества моделей). При необходимости выполняется оптимизация разбиения по алгоритмам рассмотренные в пункте 3.5.
4. Решается СЛАУ (3.1), с получением разреженной модели вида: $y(x) = \sum_{k=1}^{n_A} \hat{\alpha}_{Ak} K(x, x_k) + \hat{b}_A$.

3.6 Исследования

Целью исследований являлось сравнение получаемых разреженных решений с разбиением выборки на тестовую и обучающую части с помощью вышеперечисленных алгоритмов.

Для проведения исследования использовалась тестовая функция: $m(x) = 7 / e^{(x+0.75)^2} + 3x$, заданная на отрезке $[-1;1]$. В качестве ядерной функции использовалось RBF ядро. В качестве помехи использовались

нормально распределенные величины. Уровень помехи (дисперсия случайной величины) выбирался от 5% до 25% от мощности незашумленного сигнала. Количество наблюдений выбиралось равным 10, 20, 30 и 50. При проведении вычислительных экспериментов параметр регуляризации принимал фиксированное значение равное 10. Подбор лучшего решения осуществлялось по параметру масштаба RBF ядра, который варьировался от 10^{-5} до 10^{0} с шагом 0.1.

Ниже в таблицах **3.1**–**3.5** приведены усредненные по 600 реализациям шума значения среднеквадратичной ошибки, рассчитанной по полученным решениям на основе того или иного алгоритма извлечения обучающей выборки. В строках "Без разбиения" указываются значения MSE для неразреженных решений, полученных на полных выборках. Настройка параметров ядерных функций в этом случае проводилась по критерию LOO. Условия экспериментов по столбцам различались тем, что использовалось различное количество точек в тестовой части в % от объема полной выборки [106, 108].

Таблица 3.1 – Значение MSE при 5% уровне шума

Объем выборки	Вариант разбиения	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
N=10	без разбиения	0,0159	0,0159	0,0159	0,0159	0,0159	0,0159	0,0159	0,0159	0,0159	0,0159
	D-опт. план	0,0159	0,0159	0,0228	0,0228	0,0051	0,0051	0,0051	0,0051	0,0051	0,0051
	замена	0,0159	0,0159	0,3142	0,3142	0,0051	0,0051	0,0046	0,0046	0,0046	0,0046
	исключение	0,2031	0,2031	0,3142	0,3142	0,3142	0,3142	0,0051	0,0051	0,0051	0,0051
	включение	0,0159	0,0159	0,0228	0,0228	0,0051	0,0051	0,0046	0,0046	0,0051	0,0051
N=20	без разбиения	0,0040	0,0040	0,0040	0,0040	0,0040	0,0040	0,0040	0,0040	0,0040	0,0040
	D-опт. план	0,0079	0,2050	0,2381	0,2381	0,2381	0,0025	0,0029	0,0029	0,0025	0,0029
	замена	0,0079	0,0029	0,2381	0,2381	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029
	исключение	0,2050	0,0029	0,2381	0,0029	0,0029	0,0029	0,0029	0,0029	0,0025	0,0029
	включение	0,0079	0,2050	0,2381	0,2381	0,2381	0,0025	0,0029	0,0029	0,0025	0,0029
N=30	без разбиения	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029
	D-опт. план	0,0163	0,2225	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027
	замена	0,2225	0,2225	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027
	исключение	0,2225	0,2225	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027
	включение	0,0163	0,2225	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027	0,0027
N=50	без разбиения	0,0011	0,0011	0,0011	0,0011	0,0011	0,0011	0,0011	0,0011	0,0011	0,0011
	D-опт. план	0,1973	0,1973	0,0022	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026
	замена	0,1973	0,1973	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026
	исключение	0,1973	0,1973	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026
	включение	0,1973	0,1973	0,0022	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026

Таблица 3.2 – Значение MSE при 10% уровне шума

Объем выборки	Вариант разбиения	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
N=10	без разбиения	0,0226	0,0226	0,0226	0,0226	0,0226	0,0226	0,0226	0,0226	0,0226	0,0226
	D-опт. план	0,0532	0,0532	0,0532	0,0532	0,3147	0,3147	0,0123	0,0123	0,0123	0,0123
	замена	0,0532	0,0532	0,3147	0,3147	0,0123	0,0123	0,0118	0,0118	0,0118	0,0118
	исключение	0,2505	0,2505	0,0118	0,0118	0,0123	0,0123	0,0123	0,0123	0,0123	0,0123
	включение	0,0532	0,0532	0,0532	0,0532	0,3147	0,3147	0,0123	0,0123	0,0123	0,0123
N=20	без разбиения	0,0051	0,0051	0,0051	0,0051	0,0051	0,0051	0,0051	0,0051	0,0051	0,0051
	D-опт. план	0,0051	0,0043	0,2381	0,2381	0,0043	0,0039	0,0043	0,0043	0,0039	0,0043
	замена	0,0143	0,0043	0,2381	0,2381	0,0043	0,0043	0,0043	0,0043	0,0043	0,0043
	исключение	0,1763	0,2381	0,2381	0,2381	0,0043	0,0039	0,0043	0,0043	0,0039	0,0043
	включение	0,0051	0,0043	0,2381	0,2381	0,0043	0,0043	0,0043	0,0043	0,0039	0,0043
N=30	без разбиения	0,0045	0,0045	0,0045	0,0045	0,0045	0,0045	0,0045	0,0045	0,0045	0,0045
	D-опт. план	0,0100	0,2218	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0040
	замена	0,0044	0,2218	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044
	исключение	0,0044	0,2218	0,0044	0,0044	0,0044	0,0044	0,0047	0,0047	0,0044	0,0040
	включение	0,0100	0,2218	0,0044	0,0044	0,0044	0,0044	0,0044	0,0047	0,0044	0,0040
N=50	без разбиения	0,0017	0,0017	0,0017	0,0017	0,0017	0,0017	0,0017	0,0017	0,0017	0,0017
	D-опт. план	0,1988	0,1988	0,0062	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067
	замена	0,1988	0,1988	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067
	исключение	0,1988	0,1988	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067
	включение	0,1988	0,1988	0,0062	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067	0,0067

Таблица 3.3 – Значение MSE при 15% уровне шума

Объем выборки	Вариант разбиения	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
N=10	без разбиения	0,0304	0,0304	0,0304	0,0304	0,0304	0,0304	0,0304	0,0304	0,0304	0,0304
	D-опт. план	0,0614	0,0614	0,0614	0,0614	0,3112	0,3112	0,0118	0,0118	0,0118	0,0118
	замена	0,0614	0,0614	0,3112	0,3112	0,3112	0,3112	0,0118	0,0118	0,0112	0,0112
	исключение	0,2050	0,2050	0,3112	0,3112	0,3112	0,3112	0,3112	0,3112	0,0118	0,0118
	включение	0,0614	0,0614	0,0614	0,0614	0,3112	0,3112	0,0118	0,0118	0,0118	0,0118
N=20	без разбиения	0,0102	0,0102	0,0102	0,0102	0,0102	0,0102	0,0102	0,0102	0,0102	0,0102
	D-опт. план	0,0091	0,2417	0,2417	0,2417	0,2417	0,0173	0,0173	0,0173	0,0173	0,0184
	замена	0,0091	0,2417	0,2417	0,2417	0,0173	0,0173	0,0173	0,0173	0,0184	0,0184
	исключение	0,2417	0,2417	0,2417	0,2417	0,0173	0,0184	0,0184	0,0184	0,0184	0,0184
	включение	0,0091	0,2417	0,2417	0,2417	0,2417	0,0173	0,0173	0,0173	0,0173	0,0184
N=30	без разбиения	0,0059	0,0059	0,0059	0,0059	0,0059	0,0059	0,0059	0,0059	0,0059	0,0059
	D-опт. план	0,2225	0,2225	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089
	замена	0,2225	0,2225	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089
	исключение	0,2225	0,2225	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089
	включение	0,2225	0,2225	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089	0,0089
N=50	без разбиения	0,0024	0,0024	0,0024	0,0024	0,0024	0,0024	0,0024	0,0024	0,0024	0,0024
	D-опт. план	0,1976	0,1976	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0055
	замена	0,1976	0,1976	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061
	исключение	0,1976	0,1976	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0055
	включение	0,1976	0,1976	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0061	0,0055

Таблица 3.4 – Значение MSE при 20% уровне шума

Объем выборки	Вариант разбиения	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
N=10	без разбиения	0,0367	0,0367	0,0367	0,0367	0,0367	0,0367	0,0367	0,0367	0,0367	0,0367
	D-опт. план	0,0685	0,0685	0,0685	0,0685	0,3290	0,3290	0,0250	0,0250	0,0250	0,0250
	замена	0,0685	0,0685	0,3290	0,3290	0,3290	0,3290	0,0250	0,0250	0,0250	0,0250
	исключение	0,2844	0,2844	0,0238	0,0238	0,0250	0,0250	0,0250	0,0250	0,0250	0,0250
	включение	0,0685	0,0685	0,0685	0,0685	0,3290	0,3290	0,0250	0,0250	0,0250	0,0250
N=20	без разбиения	0,0079	0,0079	0,0079	0,0079	0,0079	0,0079	0,0079	0,0079	0,0079	0,0079
	D-опт. план	0,0138	0,0195	0,2375	0,0213	0,0213	0,0213	0,0220	0,0220	0,0220	0,0220
	замена	0,0277	0,2375	0,2375	0,2375	0,0213	0,0213	0,0220	0,0220	0,0213	0,0220
	исключение	0,0277	0,2375	0,2375	0,0213	0,0213	0,0213	0,0220	0,0220	0,0220	0,0220
	включение	0,0138	0,0195	0,2375	0,0213	0,0213	0,0213	0,0220	0,0220	0,0220	0,0220
N=30	без разбиения	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021
	D-опт. план	0,0486	0,2237	0,0135	0,0135	0,0145	0,0145	0,0145	0,0145	0,0145	0,0145
	замена	0,2237	0,2237	0,0135	0,0135	0,0145	0,0135	0,0145	0,0145	0,0145	0,0145
	исключение	0,2237	0,2237	0,0135	0,0135	0,0145	0,0145	0,0145	0,0145	0,0145	0,0145
	включение	0,0486	0,2237	0,0135	0,0135	0,0145	0,0145	0,0145	0,0145	0,0145	0,0145
N=50	без разбиения	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021	0,0021
	D-опт. план	0,1982	0,1982	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139
	замена	0,1982	0,1982	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139
	исключение	0,1982	0,1982	0,1982	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139
	включение	0,1982	0,1982	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139	0,0139

Таблица 3.5 – Значение MSE при 25% уровне шума

Объем выборки	Вариант разбиения	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
N=10	без разбиения	0,0221	0,0221	0,0221	0,0221	0,0221	0,0221	0,0221	0,0221	0,0221	0,0221
	D-опт. план	0,0289	0,0289	0,0300	0,0300	0,1594	0,1594	0,3039	0,3039	0,0300	0,0300
	замена	0,0255	0,0255	0,0300	0,0300	0,3039	0,3039	0,3039	0,3039	0,0300	0,0300
	исключение	0,2466	0,2466	0,0300	0,0300	0,3039	0,3039	0,3039	0,3039	0,0300	0,0300
	включение	0,0289	0,0289	0,0300	0,0300	0,1594	0,1594	0,3039	0,3039	0,0300	0,0300
N=20	без разбиения	0,0133	0,0133	0,0133	0,0133	0,0133	0,0133	0,0133	0,0133	0,0133	0,0133
	D-опт. план	0,0207	0,2404	0,2404	0,0262	0,0268	0,0268	0,0262	0,0262	0,0262	0,0262
	замена	0,0207	0,2404	0,2404	0,0262	0,0262	0,0262	0,0262	0,0262	0,0268	0,0262
	исключение	0,2029	0,2404	0,2404	0,2404	0,0268	0,0262	0,0262	0,0262	0,0262	0,0262
	включение	0,0207	0,2404	0,2404	0,0262	0,0268	0,0268	0,0262	0,0262	0,0268	0,0262
N=30	без разбиения	0,0033	0,0033	0,0033	0,0033	0,0033	0,0033	0,0033	0,0033	0,0033	0,0033
	D-опт. план	0,2241	0,2241	0,0261	0,0271	0,0271	0,0261	0,0271	0,0261	0,0271	0,0271
	замена	0,2241	0,2241	0,0261	0,0271	0,0261	0,0261	0,0271	0,0261	0,0261	0,0271
	исключение	0,2241	0,2241	0,0261	0,2241	0,2241	0,0261	0,0271	0,0261	0,0261	0,0271
	включение	0,2241	0,2241	0,0261	0,0271	0,0271	0,0261	0,0271	0,0261	0,0271	0,0271
N=50	без разбиения	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044	0,0044
	D-опт. план	0,1978	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190
	замена	0,1978	0,1978	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190
	исключение	0,1978	0,0190	0,0190	0,0190	0,0200	0,0200	0,0200	0,0200	0,0200	0,0190
	включение	0,1978	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190	0,0190

На рисунках **3.1**–**3.10** приводится результаты полученных разреженных моделей при использовании D–оптимального плана и других вариантов разбиения выборки. Из рисунков видно, что чем большее количество точек попадают в тестовой части выборки, тем хуже становится результат. В случаях, когда в тестовой части попадут до 25% точек из общего количества, результаты получаются наиболее оптимальными. Исходя из этих фактов, можно делать выводы, что для получения разреженных решений по методу LS–SVM, при разбиении выборки на обучающую и тестовую частей, в обучающей части можно оставить до 75% точек из выборки для подбора и настройки параметров алгоритма, а остальные точки добавить в тестовой части.

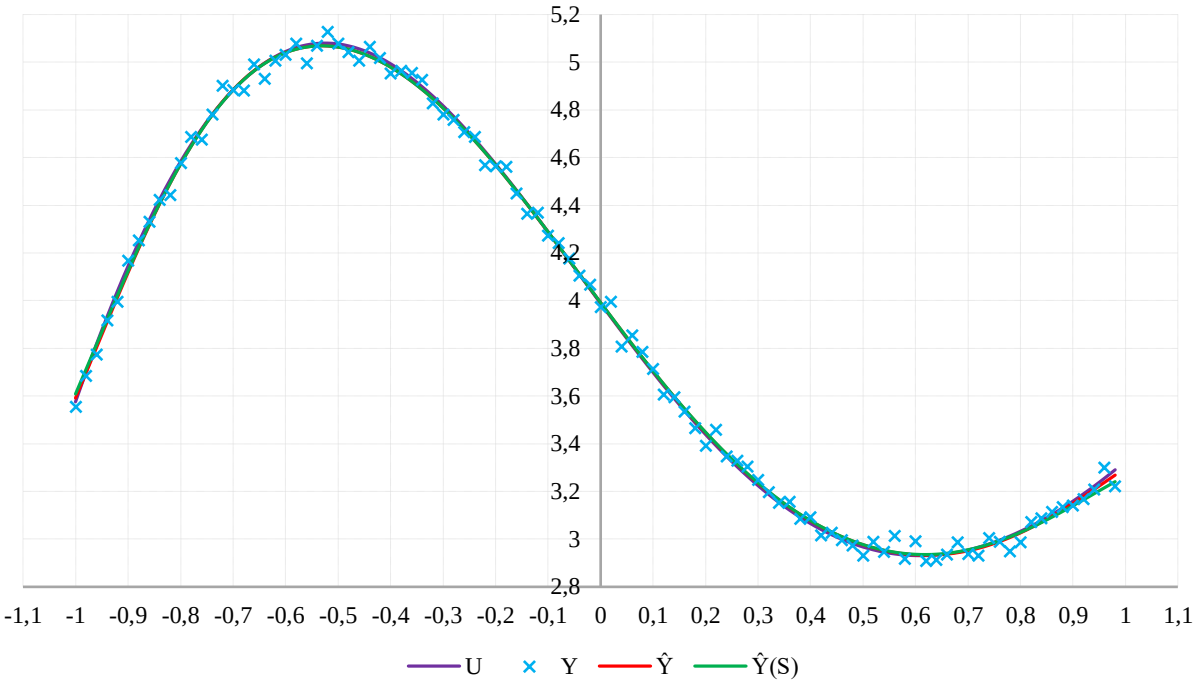


Рисунок 3.1. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 5% количество точек в тестовой части

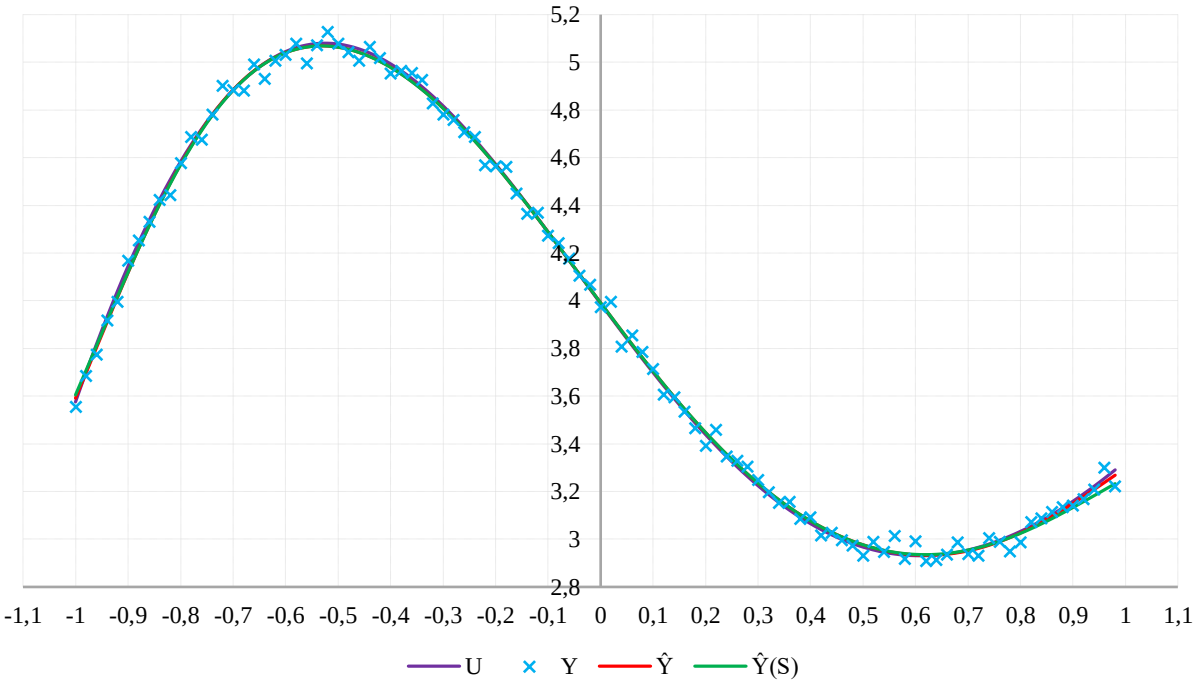


Рисунок 3.2. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 10% количество точек в тестовой части

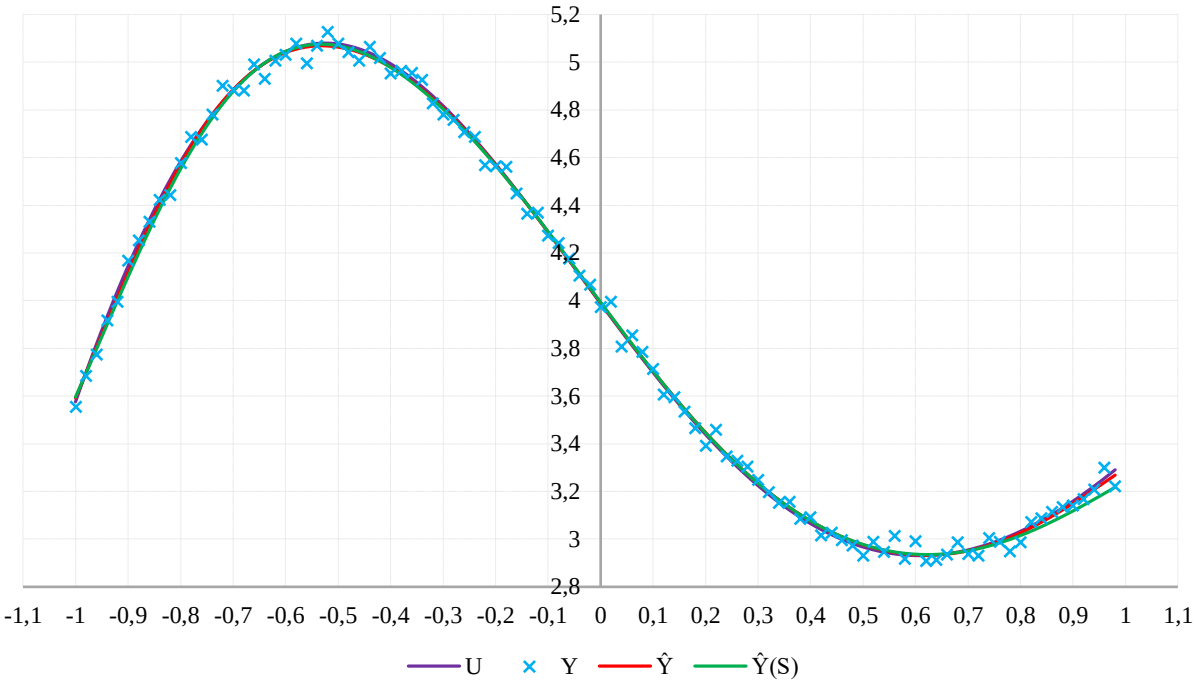


Рисунок 3.3. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 15% количество точек в тестовой части

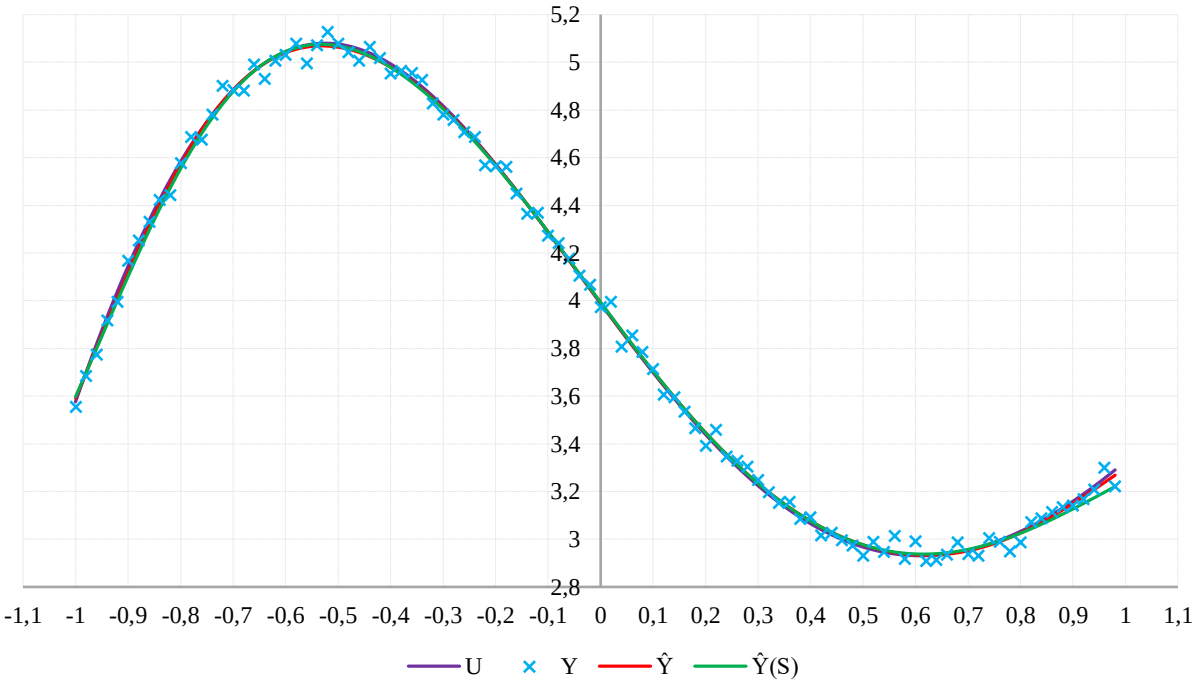


Рисунок 3.4. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 20% количество точек в тестовой части

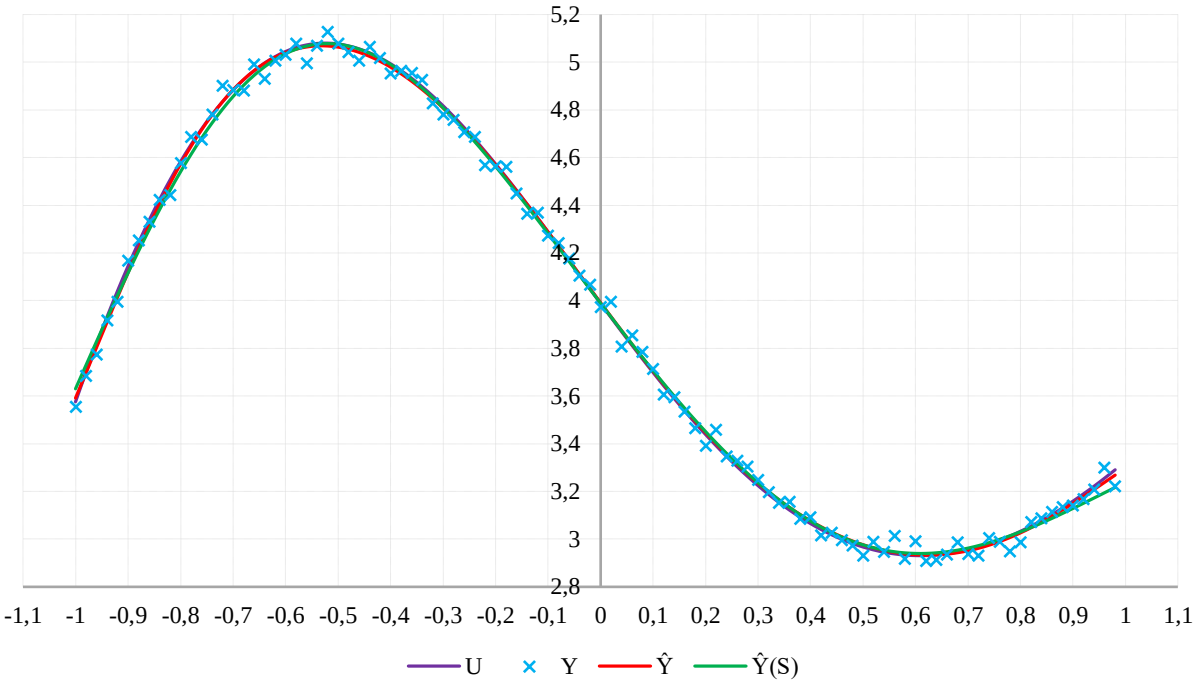


Рисунок 3.5. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 25% количество точек в тестовой части

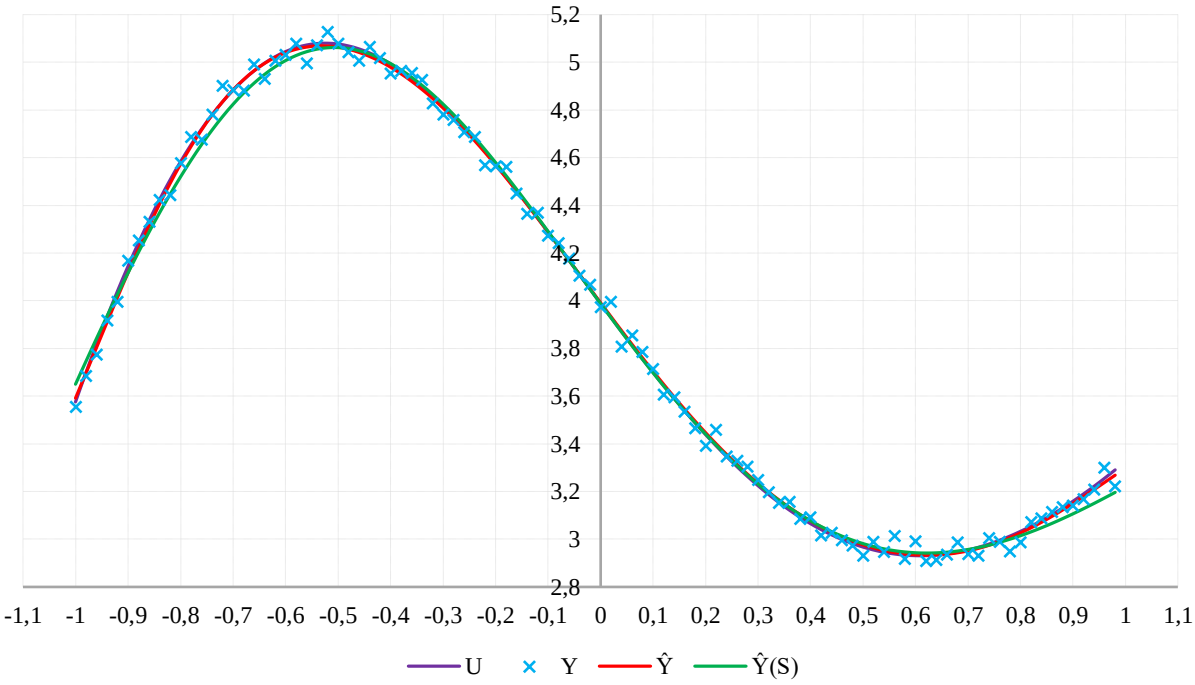


Рисунок 3.6. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 30% количество точек в тестовой части

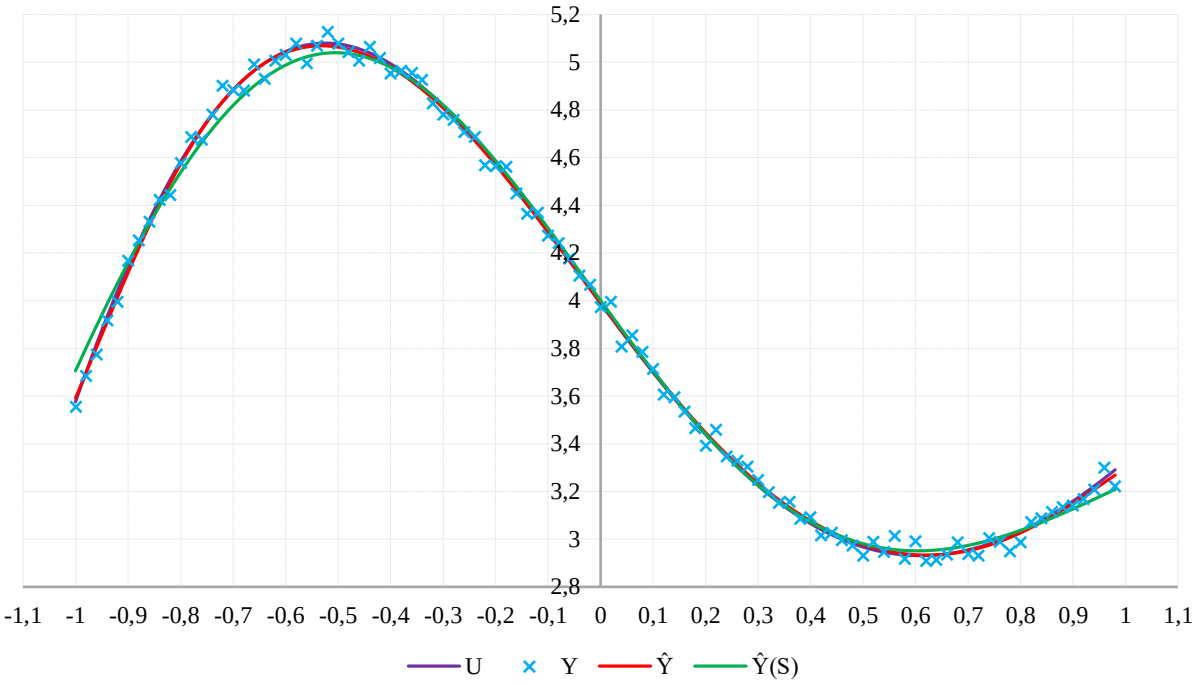


Рисунок 3.7. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 35% количество точек в тестовой части

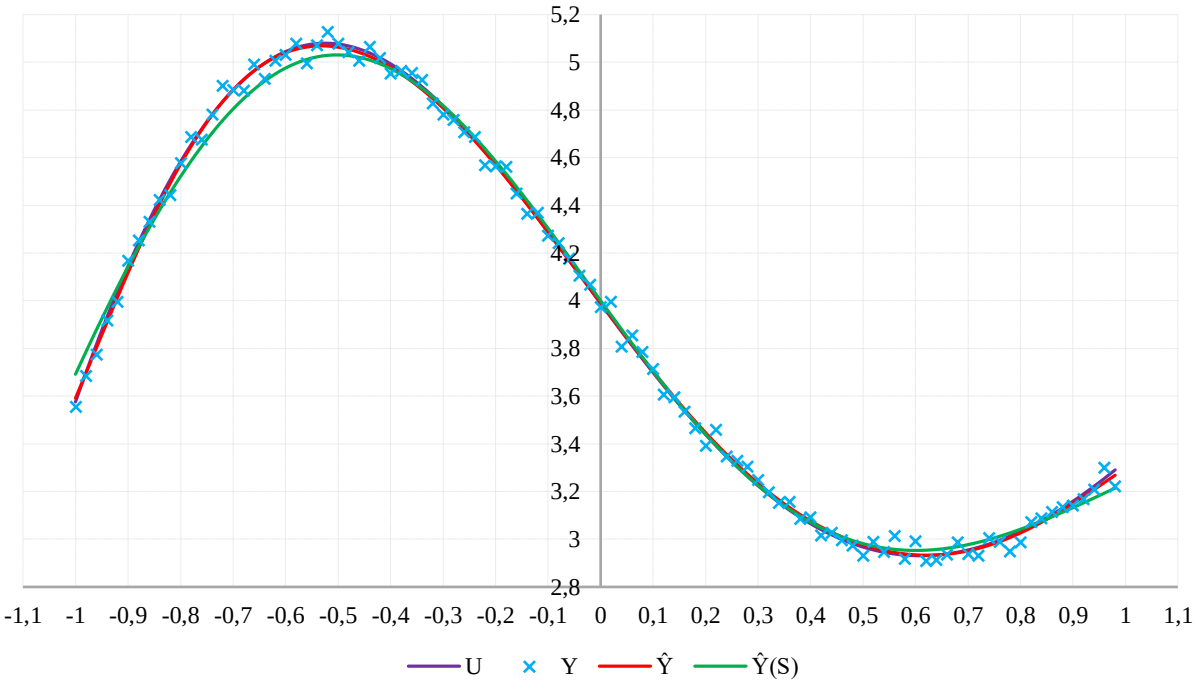


Рисунок 3.8. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 40% количество точек в тестовой части

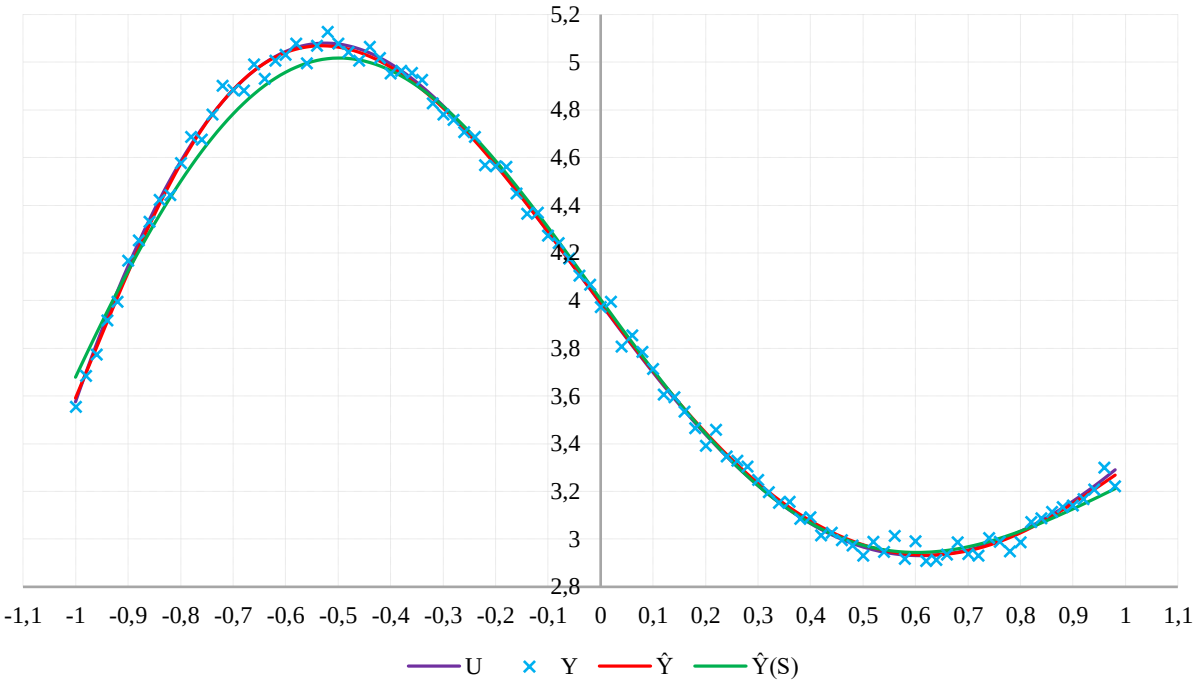


Рисунок 3.9. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 45% количество точек в тестовой части

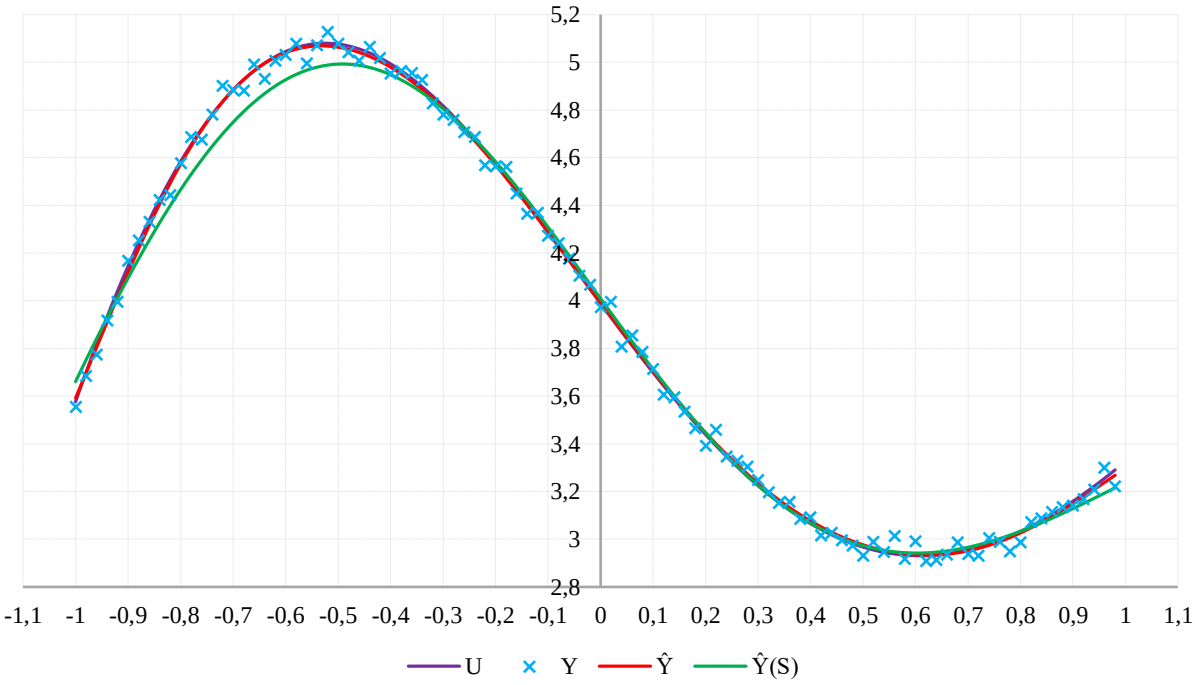


Рисунок 3.10. Графики зависимостей: U – незашумленный отклик, Y – зашумленный отклик, Ŷ – решение по обычной LS–SVM модели, Ŷ(S) – разреженное решение при 50% количество точек в тестовой части

Ниже на рисунках **3.11** и **3.12** приведены графики значений MSE при использовании D–оптимального плана и других вариантов разбиения выборки с использованием критерия согласованности [108].

Можно сравнивать между собой неразреженное решение и наполовину разреженное при 50% тестовой части. Видно, что получаемые разреженные решения при D–оптимальном разбиении выборки лишь немногим проигрывают неразреженным по величине MSE. При этом если использовать вариант разбиения на основе критерия согласованности, то улучшения качества решения с позиции MSE чаще всего не наблюдается. Это позволяет говорить о том, что для получения разреженного решения можно использовать обучающую выборку, полученную с использованием D–оптимального разбиения ее на части.

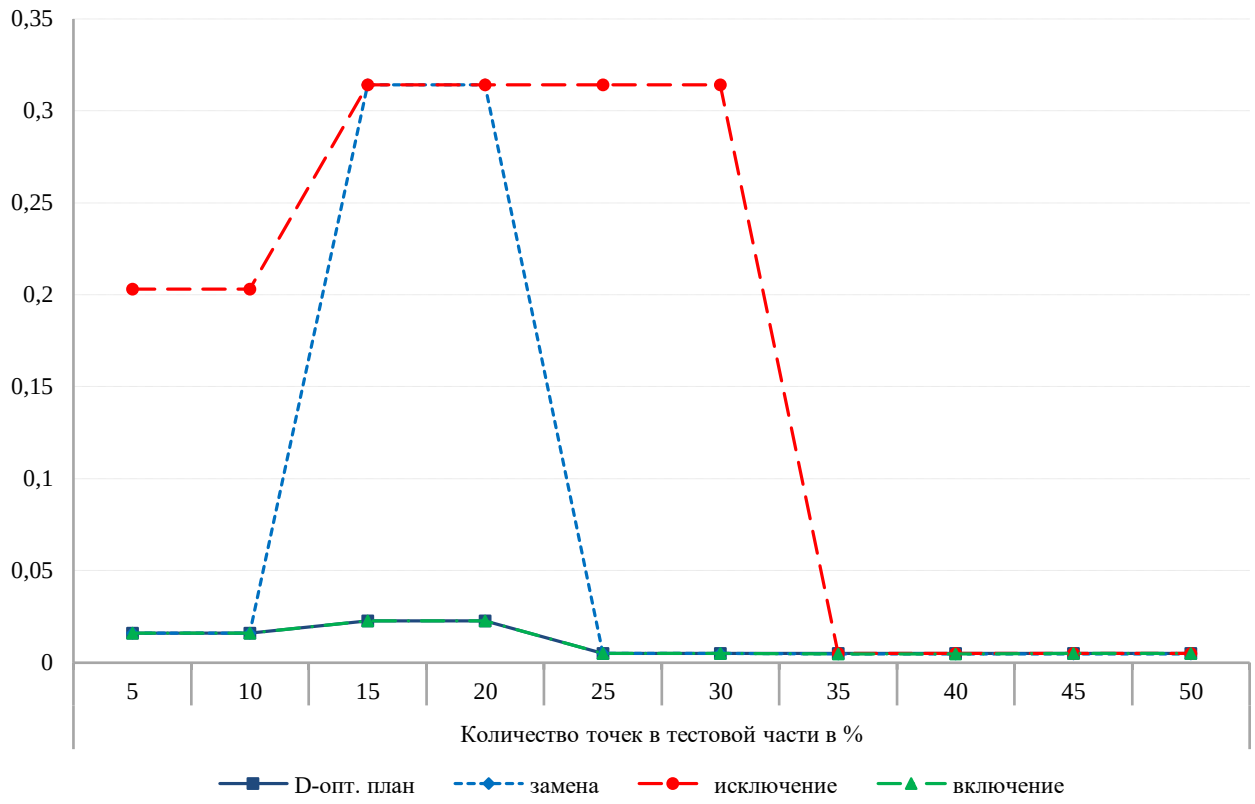


Рисунок 3.11. График значений MSE для выборки объема 10 при 5% уровне шума

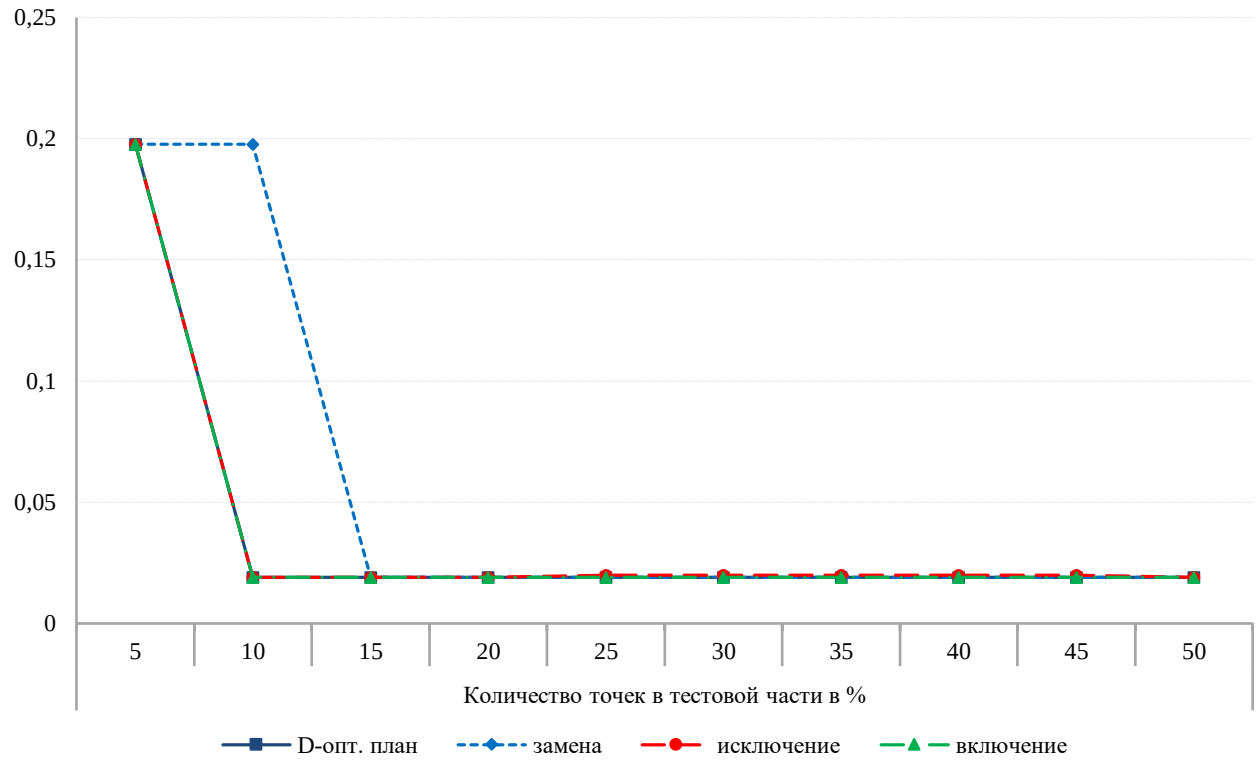


Рисунок 3.12. График значений MSE для выборки объема 50 при 25% уровне шума

В таблицах **3.6** и **3.7** приведены усредненные по 600 реализациям шума значения среднеквадратичной ошибки, рассчитанной по полученным решениям, выбранным с помощью того или иного внешнего критерия. Внешние критерии оценки качества моделей использовались для подбора метапараметров алгоритма LS–SVM и разбиении выборки на части. В таблицах в строках, озаглавленных как CV, REG, STAB, представлены соответственно средние значения MSE, полученные при использовании критерия скользящего контроля, критерия регулярности и критерия стабильности. Условия экспериментов по столбцам различались тем, что использовалось различное количество точек в тестовой части в % от объема полной выборки [107, 112].

Анализ таблиц **3.6**, **3.7** показывает, что эффективность критерия перекрестной проверки выше эффективности критериев регулярности и стабильности в условиях повышенного шума и использования тестовых

выборок малого относительного объема. Эффективность использования критерия стабильности, как правило, выше, чем у критерия регулярности.

Таблица 3.6 – Среднее значение MSE при 5% уровне шума

Объем выборки	Критерий	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
N=10	CV	0,0152	0,0152	0,0152	0,0152	0,0152	0,0152	0,0152	0,0152	0,0152	0,0152
	REG	0,0113	0,0113	0,0289	0,0289	0,0279	0,0279	0,0144	0,0144	0,0068	0,0068
	STAB	0,0034	0,0034	0,0049	0,0049	0,0049	0,0049	0,0049	0,0049	0,0053	0,0053
N=20	CV	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055
	REG	0,0051	0,0050	0,0041	0,0030	0,0031	0,0015	0,0016	0,0017	0,0019	0,0019
	STAB	0,0015	0,0015	0,0015	0,0015	0,0015	0,0015	0,0015	0,0015	0,0016	0,0016
N=30	CV	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029	0,0029
	REG	0,0037	0,0035	0,0016	0,0010	0,0011	0,0011	0,0012	0,0012	0,0015	0,0015
	STAB	0,0009	0,0009	0,0009	0,0009	0,0009	0,0009	0,0009	0,0009	0,0010	0,0009
N=50	CV	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013
	REG	0,0021	0,0013	0,0008	0,0008	0,0008	0,0007	0,0007	0,0006	0,0007	0,0007
	STAB	0,0007	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005

Таблица 3.7 – Среднее значение MSE для выборки объема 20 наблюдений при различных уровнях шума

Уровень шума	Критерий	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
5%	CV	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055	0,0055
	REG	0,0051	0,0050	0,0041	0,0030	0,0031	0,0015	0,0016	0,0017	0,0019	0,0019
	STAB	0,0015	0,0015	0,0015	0,0015	0,0015	0,0015	0,0015	0,0015	0,0016	0,0016
10%	CV	0,0062	0,0062	0,0062	0,0062	0,0062	0,0062	0,0062	0,0062	0,0062	0,0062
	REG	0,0077	0,0064	0,0057	0,0053	0,0053	0,0035	0,0032	0,0032	0,0032	0,0032
	STAB	0,0034	0,0032	0,0030	0,0029	0,0029	0,0028	0,0027	0,0027	0,0027	0,0027
15%	CV	0,0074	0,0074	0,0074	0,0074	0,0074	0,0074	0,0074	0,0074	0,0074	0,0074
	REG	0,0111	0,0094	0,0084	0,0088	0,0081	0,0060	0,0060	0,0059	0,0056	0,0056
	STAB	0,0071	0,0067	0,0061	0,0059	0,0057	0,0056	0,0051	0,0050	0,0049	0,0049
20%	CV	0,0091	0,0091	0,0091	0,0091	0,0091	0,0091	0,0091	0,0091	0,0091	0,0091
	REG	0,0135	0,0128	0,0119	0,0123	0,0114	0,0096	0,0091	0,0090	0,0083	0,0083
	STAB	0,0127	0,0118	0,0108	0,0102	0,0099	0,0095	0,0085	0,0084	0,0079	0,0078

На рисунке **3.13** представлены достигнутые средние значения MSE при использовании трех внешних критериев в зависимости от величины тестовой выборки для случая 15% шума и выборки объемом в 30 наблюдений. Средние значения MSE, достигнутые при использовании критерия скользящего

контроля показаны горизонтальной прямой. Видно, что выигрыш от использования критериев регулярности и стабильности, как правило, может быть достигнут при относительно большой тестовой части выборки. На рисунке **3.14** представлены достигнутые средние значения MSE при использовании критерия стабильности для выборки в 30 наблюдений при вариации объема тестовой части и изменении уровня шума от 5 до 20%. Видим, что при увеличении уровня шума минимум критерия сдвигается вправо. Это говорит о том, что в условиях использования сильно зашумленных выборок целесообразно по возможности использовать критерий стабильности с относительно большой тестовой частью. Аналогичные результаты для критерия регулярности представлены на рисунке **3.15** [107].

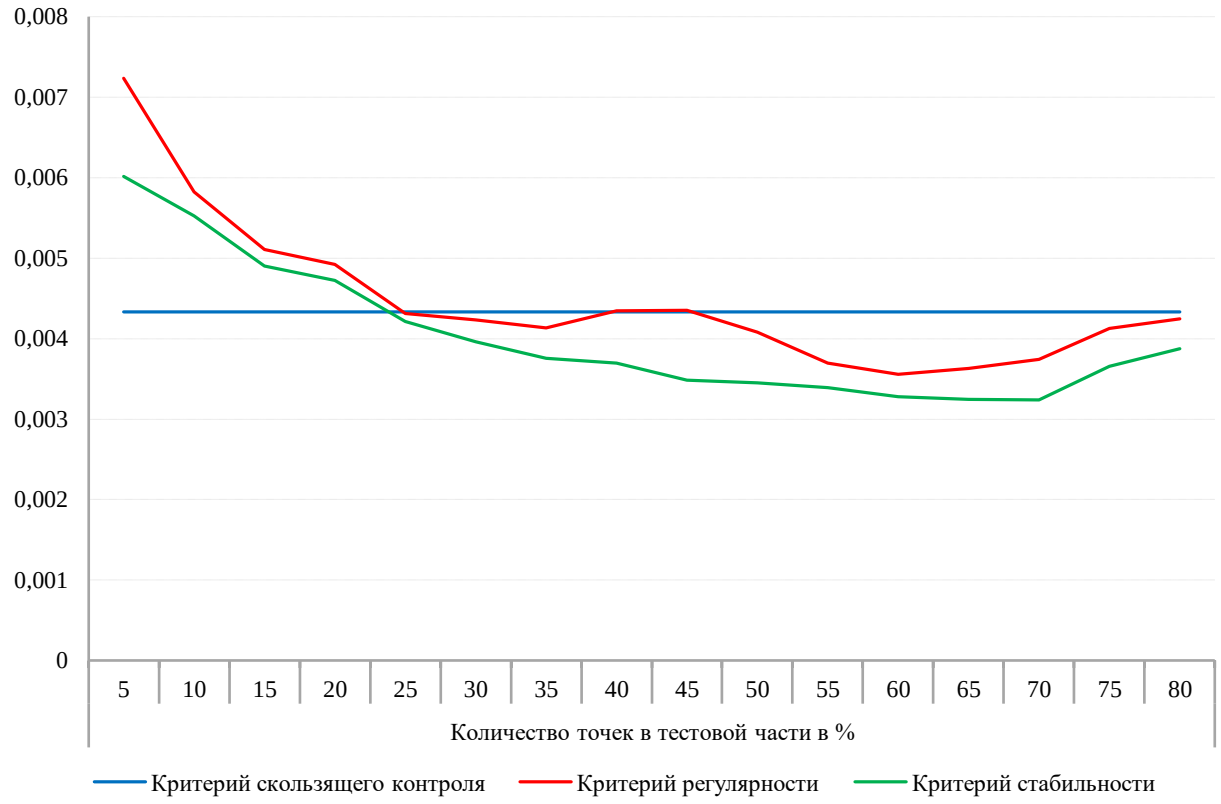


Рисунок 3.13. График средних значений MSE для выборки объема 30 при 15% уровне шума

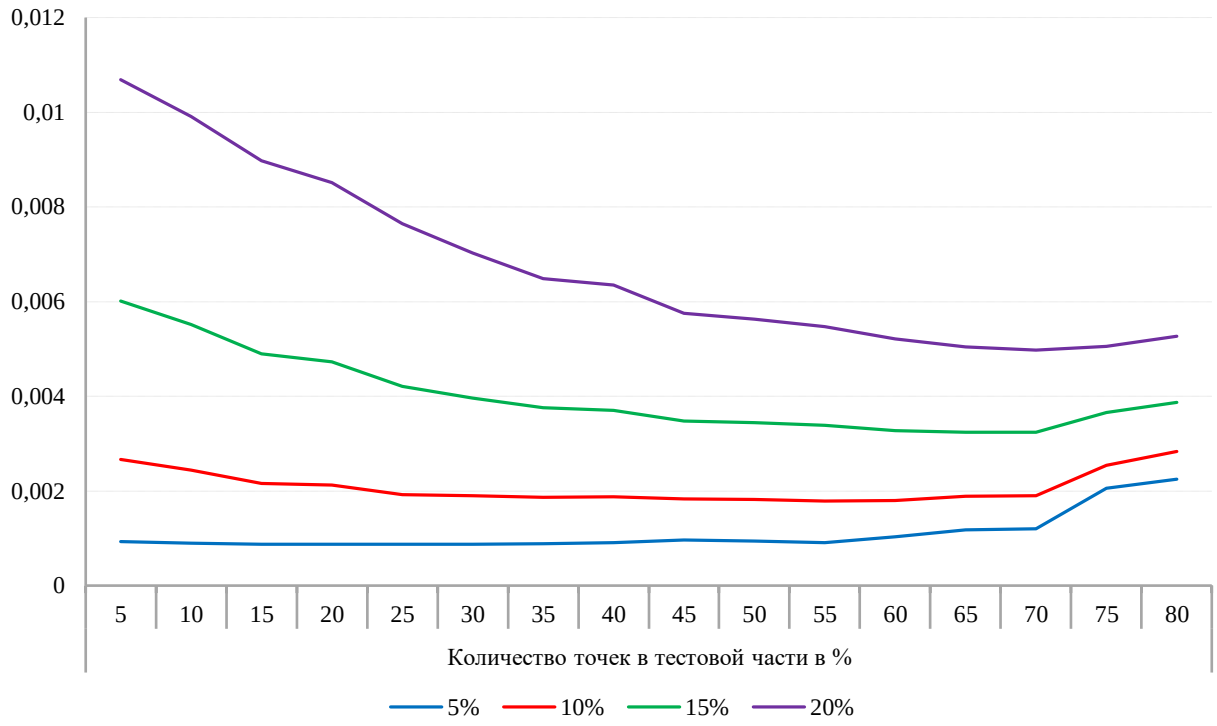


Рисунок 3.14. График средних значений MSE при использовании критерия стабильности для выборки объема 30 с изменением уровня шума от 5% до 20%

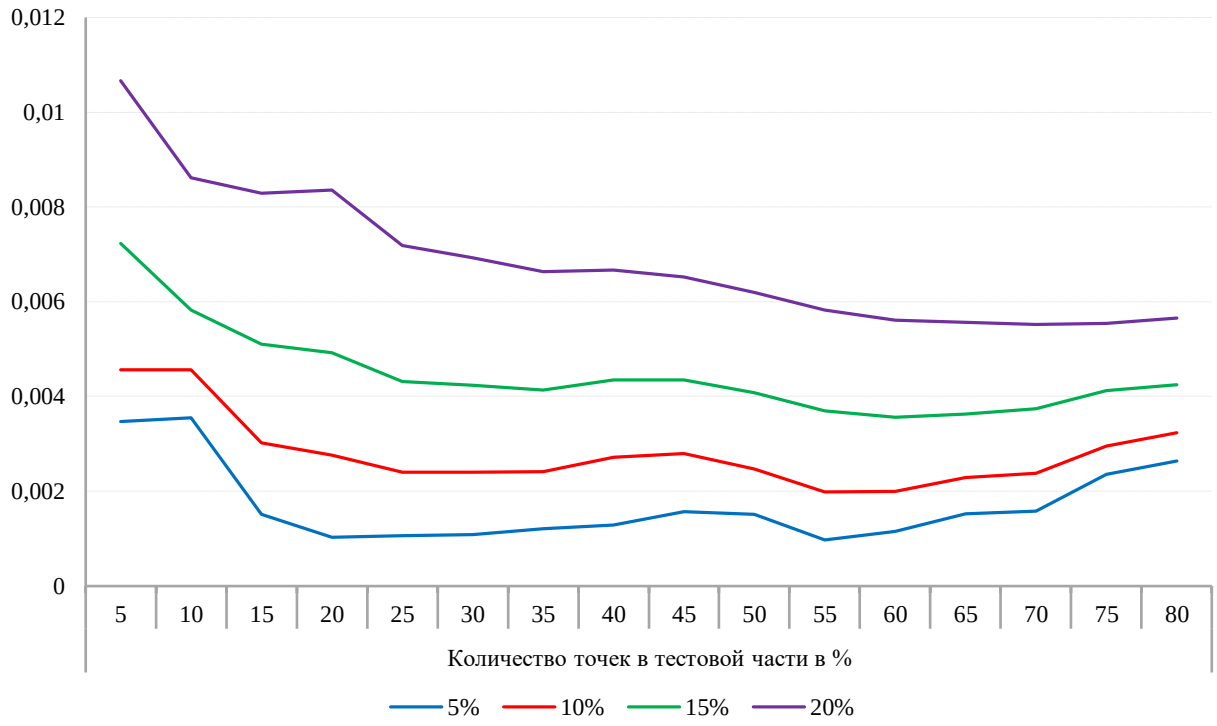


Рисунок 3.15. График средних значений MSE при использовании критерия регулярности для выборки объема 30 с изменением уровня шума от 5% до 20%

3.7 Выводы

В данной главе рассмотрены основные способы разбиения выборки на обучающую и тестовую части с использованием D–оптимального плана и критериев оценки качества моделей. Проведены исследования по подбору метапараметров алгоритма LS–SVM с использованием внешних критериев оценки качества моделей.

Основные полученные результаты в данной главе можно сформулировать следующим образом:

1. Предложен алгоритм построения разреженных LS–SVM решений с использованием обучающей и тестовой части выборки.
2. Разработан алгоритм разбиения выборки на обучающую и тестовую части D–оптимальным планированием эксперимента.
3. Разработаны несколько алгоритмов разбиения выборки на обучающую и тестовую части с использованием критериев оценки качества моделей. Эти алгоритмы можно использовать в качестве дополнительной оптимизации состава точек в обучающей и тестовой частях.
4. Для подбора метапараметров алгоритма LS–SVM использованы внешние критерии оценки качества моделей, такие как критерии перекрестной проверки, регулярности и стабильности.

ГЛАВА 4. ПРИМЕНЕНИЕ МЕТОДА LS–SVM ДЛЯ РЕШЕНИЯ ПРАКТИЧЕСКИХ ЗАДАЧ

В данной главе рассматривается применение алгоритма LS–SVM для построения регрессионных моделей на основе известных выборок LIDAR, Motorcycle и данных по определению общих констант устойчивости комплексов серебра.

4.1 Выборка LIDAR

Технология LIDAR (Light Detection And Ranging), основанная на измерении отраженного света от объектов с помощью лазера, применяется для обнаружения химических компонентов в атмосфере. Эта технология успешно доказала свою эффективность как средство для мониторинга распределения загрязняющих веществ в атмосфере [114, 115].

Типичные данные, используемые в технологии LIDAR, приведены на рисунке **4.1**. Горизонтальная переменная *range* определяет расстояние, пройденное отраженным светом до момента его возвращения к своему источнику. Переменная *log ratio*, являющаяся вертикальной, представляет собой логарифм коэффициента частоты резонанса (Гц) рассматриваемого компонента, в данном случае ртути. Другой источник имеет частоту, отличную от данной резонансной частоты. Более подробное описание выборки приводится в [114].

Данная выборка характеризуется нелинейностью и явным непостоянством дисперсии (гетероскедастичностью), которая значительно увеличивается вдоль горизонтальной оси. Интерес представляет оценка зависимости *logration* от *range*, а также оценка производной этой зависимости.

Отметим, что данная зависимость не подчиняется ни линейной, ни квадратичной форме. Применение полиномов не всегда эффективно для описания имеющихся данных. В случаях, когда структура исходной модели

неизвестна и существенно нелинейна, более целесообразно применять непараметрические и полу-параметрические подходы.

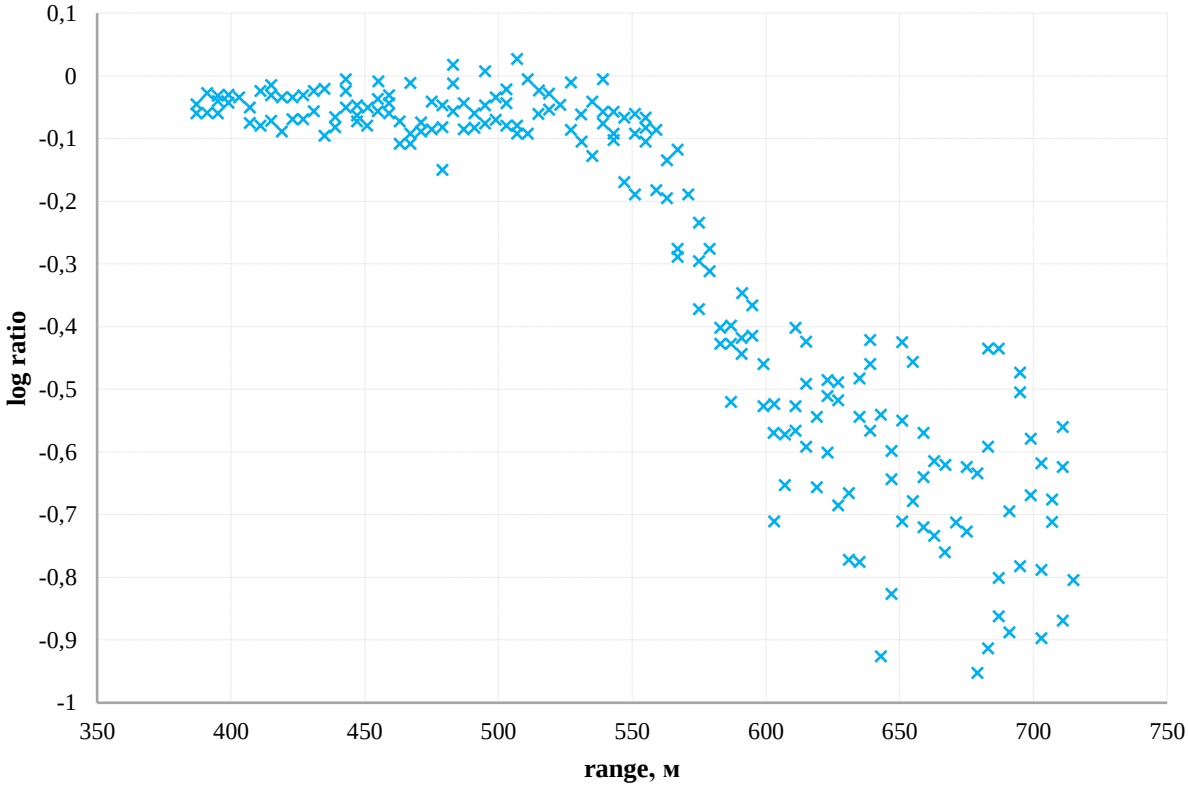


Рисунок 4.1. Данные выборки LIDAR

Таким образом, при проведении вычислительных экспериментов для метода LS–SVM в качестве ядерной функции использовалось гауссово ядро. В качестве параметра регуляризации использовалось фиксированное значение, равное 10. Параметры ядерной функции выбирались по значениям различных критериев оценки качества моделей. Целью вычислительного эксперимента являлось сравнение полученных обычных, робастных и разреженных решений [118, 119].

Полученные результаты критерия детерминации R^2 для обычной, робастной и разреженной модели LS–SVM приведены в таблицах **4.2**–**4.3**.

Поскольку в данных наблюдается эффект гетероскедастичности, то вполне можно применение робастных процедур для нивелирования эффекта. Это особенно важно, если проводить разреживание решения. Падение качества разреженных решений можно наблюдать по результатам таблицы **4.3**.

Если же использовать робастные процедуры (таблица **4.1**), то видим, что на разреженных решениях нет падения коэффициента детерминации.

Таблица 4.1 – Значение R^2 при робастной модели LS–SVM

Метод и вид функции потерь	Критерий	Количество точек в тестовой части в %									
		5	10	15	20	25	30	35	40	45	50
псевдонаблюдения, простая функция потерь	RLOO-P	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
	RLOO	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
	REG	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
взвешивание, простая функция потерь	RLOO-P	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
	RLOO	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
	REG	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
псевдонаблюдения, адаптивная функция потерь	RLOO-P	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
	RLOO	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
	REG	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
взвешивание, адаптивная функция потерь	RLOO-P	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910
	RLOO	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910
	REG	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910	0,910

Таблица 4.2 – Значение R^2 при обычной модели LS–SVM

Критерий	Количество точек в тестовой части в %									
	5	10	15	20	25	30	35	40	45	50
LOO	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915
K-FOLD	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915	0,915

Таблица 4.3 – Значение R^2 при разреженной модели LS–SVM

Вариант разбиения	Количество точек в тестовой части в %									
	5	10	15	20	25	30	35	40	45	50
D-опт. план	0,915	0,914	0,915	0,912	0,908	0,900	0,899	0,896	0,904	0,902
замена	0,915	0,915	0,914	0,914	0,914	0,914	0,915	0,915	0,913	0,911
исключение	0,915	0,914	0,915	0,912	0,908	0,900	0,899	0,896	0,904	0,902
включение	0,915	0,680	0,358	0,781	0,908	0,900	0,899	0,896	0,904	0,902
Add/Del	0,910	0,914	0,915	0,912	0,908	0,900	0,899	0,896	0,904	0,902
Del/Add	0,915	0,914	0,915	0,912	0,908	0,900	0,899	0,896	0,904	0,902

Построенные LS–SVM-регрессии для данной выборки представлены на рисунке **4.2**.

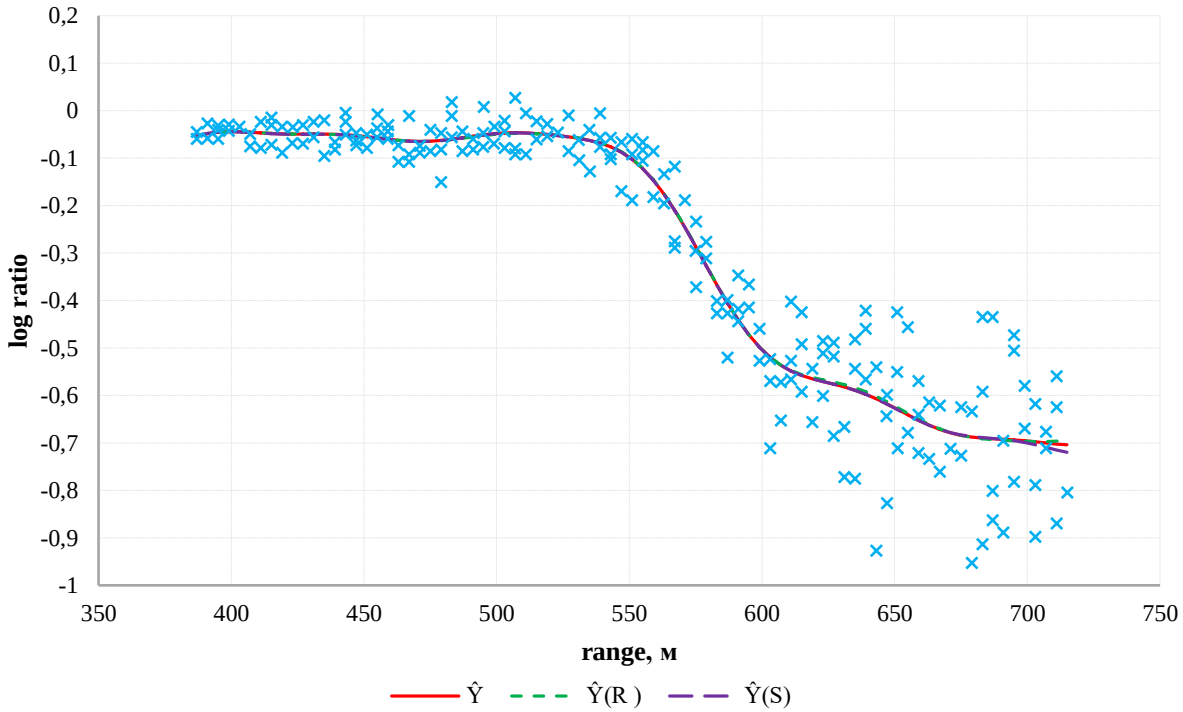


Рисунок 4.2. Обычная, робастная и разреженная LS–SVM-регрессии с использованием гауссова ядра для выборки LIDAR (Ŷ – обычное решение, Ŷ(R) – робастное решение, Ŷ(S) – разреженное решение)

По результатам проведенных вычислительных экспериментов можно увидеть, что в некоторых случаях разреженные решения проигрывают обычным и робастным решениям по значению R^2. Также робастный вариант решения с использованием взвешенного метода на основе адаптивного вида функции потерь Хьюбера предусматривает другие способы получения робастного решения. Видно, что получаемые разреженные решения при D–оптимальном разбиении выборки лишь немногим уступают не разреженным по величине R^2. При этом если использовать вариант разбиения на основе критерия регулярности, то улучшения качества решения с позиции R^2 чаще всего не наблюдается. Это позволяет говорить о том, что для получения разреженного решения можно использовать обучающую выборку, полученную с использованием D–оптимального разбиения ее на части.

Вместе с тем включение дополнительной оптимизации состава точек (алгоритм «Замена») позволяет в ряде случаев получать лучшие решения.

4.2 Выборка MOTORCYCLE

Motorcycle Accident Dataset представляет собой классический пример данных, часто упоминаемых в современной литературе для оценки качества регрессионных моделей [116, 117]. В результате испытаний шлемов на ударную прочность с использованием манекенов, установленных на мотоциклах, были получены эти данные. Ускорение головы манекена записывалось в течение определенного времени после столкновения со стеной. Эксперименты проводились для того, чтобы определить эффективность защитных шлемов. Данные представляют собой измерения ускорения головы мотоциклиста, зависимого от времени (рисунок **4.3**). Можно выделить как минимум две или даже три области, в которых отклик демонстрирует различное поведение с точки зрения структуры зависимости и уровня шума. Подробное описание выборки приводится в [116].

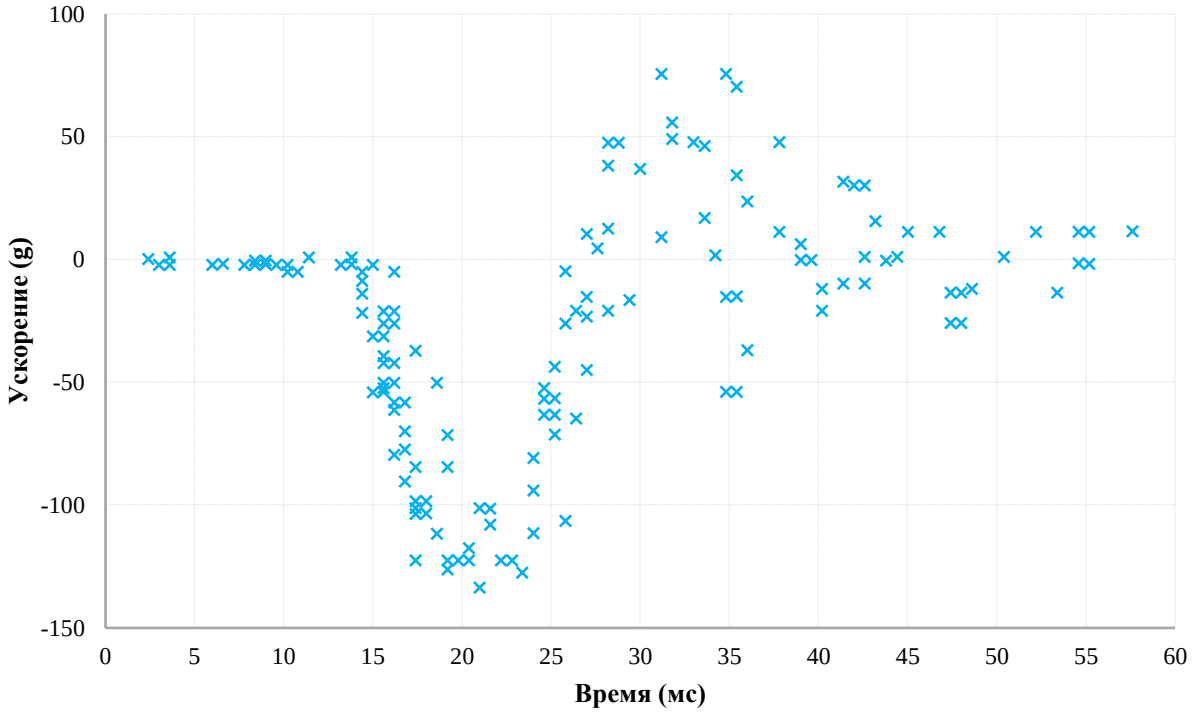


Рисунок 4.3. Данные выборки Motorcycle

Полученные результаты критерия детерминации R^2 для обычной, робастной и разреженной модели LS–SVM приведены в таблицах **4.4**–**4.6** [118, 119].

Таблица 4.4 – Значение R^2 при обычной модели LS–SVM

Критерий	Количество точек в тестовой части в %									
	5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
LOO	0,805	0,805	0,805	0,805	0,805	0,805	0,805	0,805	0,805	0,805
K-FOLD	0,805	0,805	0,805	0,805	0,805	0,805	0,805	0,805	0,805	0,805

Таблица 4.5 – Значение R^2 при робастной модели LS–SVM

Метод и вид функции потерь	Критерий	Количество точек в тестовой части в %									
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
псевдонаблюдения, простая функция потерь	RLOO-P	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804
	RLOO	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804
	REG	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804	0,804
взвешивание, простая функция потерь	RLOO-P	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796
	RLOO	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796
	REG	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796	0,796
псевдонаблюдения, адаптивная функция потерь	RLOO-P	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803
	RLOO	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803
	REG	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803	0,803
взвешивание, адаптивная функция потерь	RLOO-P	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681
	RLOO	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681
	REG	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681	0,681

Таблица 4.6 – Значение R^2 при разреженной модели LS–SVM

Вариант разбиения	Количество точек в тестовой части в %									
	5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
D-опт. план	0,803	0,804	0,803	0,790	0,785	0,784	0,776	0,768	0,764	0,757
замена	0,803	0,803	0,797	0,801	0,801	0,803	0,801	0,801	0,788	0,783
исключение	0,803	0,804	0,803	0,790	0,785	0,784	0,776	0,768	0,764	0,757
включение	0,803	0,804	0,803	0,790	0,785	0,784	0,776	0,768	0,764	0,757
Add/Del	0,803	0,804	0,803	0,790	0,785	0,784	0,776	0,768	0,764	0,757
Del/Add	0,803	0,804	0,803	0,790	0,785	0,784	0,776	0,768	0,764	0,757

Построенные LS–SVM-регрессии для данной выборки представлены на рисунке **4.4**.

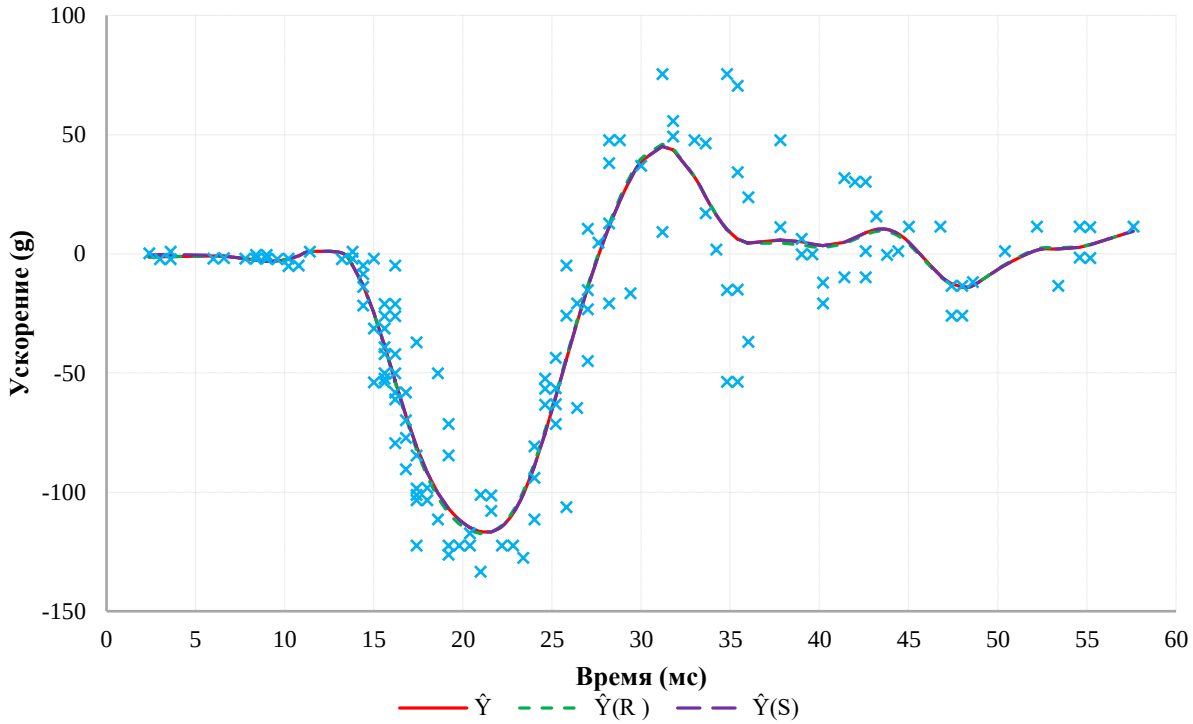


Рисунок 4.4. Обычная, робастная и разреженная LS–SVM-регрессии для выборки Motorcycle (Ŷ – обычное решение, Ŷ(R) – робастное решение, Ŷ(S) – разреженное решение)

По полученным результатам можно прийти к аналогичным выводам, приведенным в анализе выборки LIDAR.

4.3 Изучение процесса комплексообразования переходных металлов с производными тиомочевины в водных и водно-органических растворах

Практическая значимость водно-органических растворителей обусловлена их способностью изменять физико-химические свойства, что важно для создания условий, способствующих протеканию различных процессов, включая реакции комплексообразования.

Метод температурного коэффициента является один из наиболее распространенным методом, который используется для оценки

термодинамических функций процесса образования комплексных частиц. Именно с использованием этого метода были рассчитаны термодинамические функции процесса комплесообразования [120–123].

Известно, что на устойчивость комплексных соединений влияет не только природа раствора, но и другие факторы, связанные с природой органического лиганда. Большое практическое значение представляют координационные соединения ионов различных металлов с серосодержащими лигандами. К числу таких лигандов относятся и производные тиомочевины. Тиомочевина и её производные широко используются в аналитической химии и в настоящее время являются основой многих лекарственных и биологически активных веществ. Известно, что по определённому уравнению $2R = SR - S - S - R$ тиомочевина и некоторые её производные [120, 121] окисляются до соответствующих дисульфидов. Эта система широко используется для изучения процесса комплексообразования тиомочевины с переходными металлами. Введение этильных радикалов в молекулу тиомочевины может повлиять на её восстановительную способность, которая количественно характеризуется значением электродного потенциала системы.

Цель работы заключалось в исследовании процесса комплексообразования Ag(I) с тиокарбогидразидом (ТКЗ) в водно-спиртовых растворах, содержащих 25, 50 и 75 обьемн. % метанола и этанола при температурах 321, 325 и 335 К.

Равновесную концентрацию иона серебра можем определять по уравнению:

$$\lg\left[Ag^{+}\right] = \lg C_{Ag^{+}} - \frac{\Delta E}{1.985 \cdot 10^{-4} \cdot T},$$

где $\left[Ag^{+}\right]$ – равновесная концентрация ионов серебра в каждой точке титрования, $\Delta E = E_1 - E_2$, E_1 – начальный потенциал системы, E_2 –

потенциал системы в каждой точке титрования, C_{Ag^+} – концентрация серебра в каждой точке титрования с учетом разбавления.

Равновесную концентрацию тиокарбогидразида можем определять по уравнению:

$$\mathrm{L} = C_L - n\left(C_{Ag^+} - [Ag]\right),$$

где C_L – концентрация тиокарбогидразида в каждой точке титрования, n – число молекул тиокарбогидразида, присоединенных серебром (I), C_{Ag^+} – концентрация серебра (1) в каждой точке титрования, $\left[Ag^+\right]$ – равновесная концентрация ионов серебра в каждой точке титрования.

Исследование проводились в водно-спиртовых растворах при температуре 298 К и ионной силе 0,1 моль/л, создаваемой $NaNO_3$. Начальная концентрация $AgNO_3$ составляла $1\cdot10^{-4}$ моль/л, а начальная концентрация тиокарбогидразида – $1\cdot10^{-2}$ моль/л. Для потенциометрического титрования использовалось ячейка с переносом. Индикаторным электродом служила серебряная пластинка, а электродом сравнения – хлорсеребряный электрод. Потенциал системы при потенциометрическом титровании измерялось с помощью pHMeterpH-150МП. Равновесное значение потенциала на индикаторном электроде устанавливалось в течение 10-15 минут.

В таблице **4.7** в качестве примера приведены данные по определению равновесной концентрации иона серебра(I), равновесной концентрации тиокарбогидразида и функции Ладена в растворе, содержащем 25 обьемн. % метанола при 298 К по данным потенциометрического титрования.

Для определения количества частиц, образующихся при действии серебра в водно-спиртовых растворах был использован метод Яцимирского [120]. На рисунке **4.5** представлена зависимость ΔE от –lg[L] для тиокарбогидразидных комплексов серебра(I) в водно-метанольных растворах при содержании спирта равном 25 (1), 50 (2) и 75 (3) обьемн. % при

температуре 321, 325 и 335 К. Угол наклона зависимостей ΔЕ от –lg[L] при избытке тиокарбогидразида в растворе при вышеуказанных соотношениях (рисунок **4.5**). равняется 0,180 В*л*моль-1(1), 0,182 В*л*моль-1(2) и 0,183 В*л*моль-1(3), что свидетельствует о присоединении трех молекул тиокарбогидразида к иону серебра(I).

Таблица 4.7 – Данные по определению равновесной концентрации иона серебра(I), тиокарбогидразида и функции Ледена (F) в растворе, содержащем 25 обьемн. % метанола при 298 К по данным потенциометрического титрования. С ТКЗ =1·10^{-2}моль/л; CAg+=1·10^{-4} моль/л

ΔЕ прак, В	**ΔЕ теор, В**	**СТКЗ·10^{-3} моль/л**	**[Ag]·10^{-12} моль/л**	**[L]·10^{-3} моль/л**	**lgF0**
1	2	3	4	5	6
0,212	0,193	4	25380,75	0,108636	7,559
0,219	0,217	5,2	19178,5	0,221293	7,372
0,234	0,232	6,3	10573,87	0,332216	7,454
0,242	0,243	7,4	7715,83	0,441468	7,468
0,248	0,253	8,4	6181,93	0,549075	7,469
0,263	0,265	10	3343,75	0,707473	7,626
0,274	0,275	11,6	2180,39	0,862368	7,726
0,282	0,283	1,3,1	1561,19	1,013873	7,801
0,289	0,29	14,6	1175,99	1,162097	7,864
0,295	0,297	16	935,55	1,307146	7,913
0,301	0,305	17,9	718,99	1,495777	7,968
0,305	0,312	19,7	597,41	1,679168	7,999
0,311	0,318	21,5	473,83	1,857536	8,055
0,319	0,324	23,7	347,92	2,073738	8,142
0,329	0,33	25,8	231,84	2,282782	8,276
0,337	0,336	27,8	163,83	2,485017	8,39
0,342	0,341	29,8	131,67	2,68077	8,452
0,351	0,349	33,5	90,58	3,054038	8,558
0,356	0,356	37	71,47	3,404819	8,614
0,364	0,368	43,4	49,76	4,046591	8,696
0,371	0,377	49,1	36,42	4,619355	8,774
0,376	0,384	54,3	28,12	5,133674	8,841
0,39	0,393	61,1	15,15	5,813744	9,055
0,397	0,4	67	10,94	6,40354	9,155
0,405	0,407	7,37	7,33	7,078049	9,285
0,41	0,411	7,81	5,62	7,516092	9,374
0,414	0,416	8,32	4,5	8,027402	9,442
0,417	0,421	8,87	3,65	8,572549	9,504
0,422	0,424	9,31	2,79	9,018182	9,6
0,424	0,428	9,71	2,4	9,41831	9,646

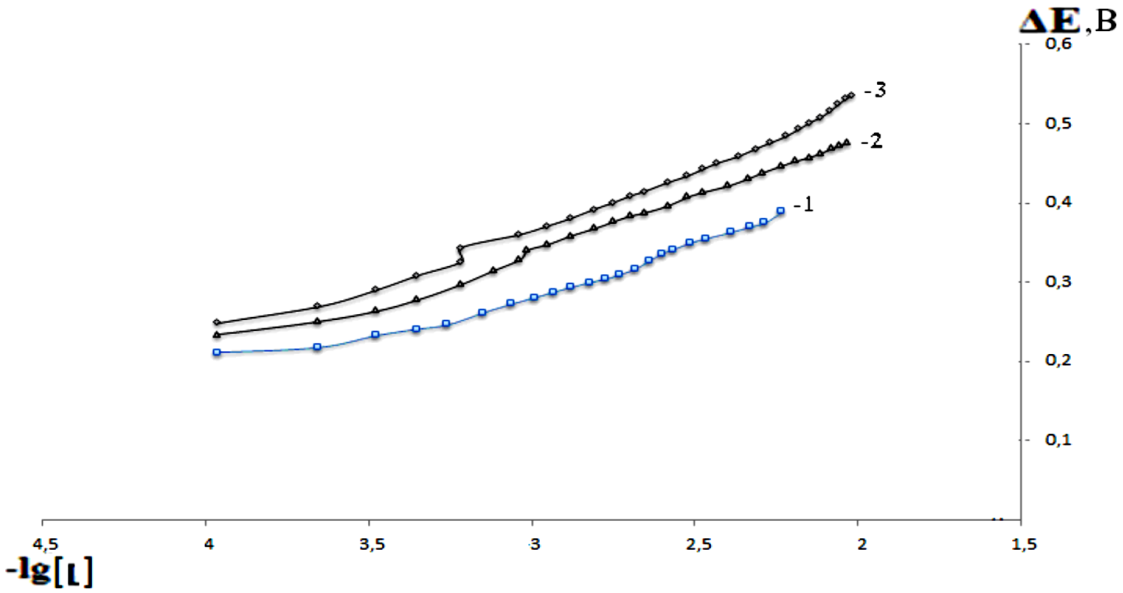


Рисунок 4.5. Зависимость ΔE от –lg[L] для тиокарбогидразидных комплексов серебра (I) в водно-метанольных растворах, содержащих 1-25, 2-50 и 3-75 обьемн. % спирта при температуре 298 К

Проведенные исследования показали, что комплексообразование серебра(I) с тиокарбогидразидом в водно-этанольных растворах, содержащих 1-25, 2-50 и 3-75 объем. % спирта при 298 К, близко по характеру комплексообразования, которое протекает в водно-метанольных растворах. Проведенные исследования показали, что состав смешанного раствора не влияет на количество комплексных частиц, образующихся в системе Ag(1)-ТКЗ-Р; где Р–смешанный раствор [121–123].

Для определения общих констант устойчивости комплексов серебра (I) с тиокарбогидразидом, по данным потенциометрического титрования, использовали нелинейный метод наименьших квадратов (МНК), основные положения которого изложены в работе [124] и метод LS–SVM [125].

Переходим к определению констант устойчивости образования комплексов серебра методом LS–SVM. Для этой цели воспользуемся алгоритмами, приведенными в работах [58, 68–71, 106–112, 126].

Во время проведения химических экспериментов в качестве ядерной функции метода LS–SVM использовалось полиномиальное ядро. Значение

параметра регуляризации было установлено фиксированным, и оно равно 1000. В качестве данных использовались выборки объемом 30, 31, 32, 33, 34 точки.

Далее, в таблицах **4.8**–**4.10** приведены полученные результаты значения критерия детерминации R^2 по методам МНК и LS–SVM.

При анализе полученных результатов можно увидеть, что в большинстве случаев результаты метода LS–SVM превосходят результаты МНК. Это позволяет говорить о том, что для определения общих констант устойчивости комплексов переходных металлов целесообразно использовать метод LS–SVM.

Для определения общих констант устойчивости комплексов переходных металлов использовался метод LS–SVM. Результаты проведенных вычислительных экспериментов были сравнены с результатами МНК, показаны преимущества использования метода LS–SVM. По полученным результатам можно рекомендовать использовать метод LS–SVM для решения задач, связанных с комплексообразованием переходных металлов с производными тиомочевины в водных и водно-органических растворах [125].

Проведенные исследования продемонстрировали хорошие возможности использования робастной и разреженной вариантов построения регрессии методом LS–SVM для решения прикладных задач. Исходя из вариантов исходных выборок робастные и разреженные решения, полученные методом LS–SVM являются достаточно хорошими (гладкими). Особенно робастный вариант решения, полученный алгоритмом LS–SVM с псевдонаблюдениями на основе обычной и адаптивной функции потерь Хьюбера с использованием робастного варианта критерия скользящего контроля RLOO-P, является достаточно неплохим по сравнению с другими решениями. Разреженный вариант алгоритма LS–SVM с D–оптимальным разбиением выборки на основе критерия регулярности тоже, в большинство случаев, является наиболее качественным.

По результатам проведенных вычислительных экспериментов можно сделать выводы о том, что эффективность использования робастной и разреженной вариантов алгоритма LS–SVM вполне пригодны для решения прикладных задач и построения регрессии на их основе.

Таблица 4.8 – Значение R^2 при температурном режиме 335 К

Тип модели	Критерий	Вариант разбиения	Количество точек в тестовой части в %									
			5	10	15	20	25	30	35	40	45	50
МНК		без разбиения	0,562	0,562	0,562	0,562	0,562	0,562	0,562	0,562	0,562	0,562
обычный LS–SVM	LOO	без разбиения	0,828	0,828	0,828	0,828	0,828	0,828	0,828	0,828	0,828	0,828
	K-FOLD	без разбиения	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
робастный LS–SVM	RLOO-P	без разбиения, псевдонаблюдения, простая функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991
	REG		0,999	0,999	0,999	0,991	0,999	0,999	0,999	0,999	0,999	0,994
	RLOO-P	без разбиения, взвешивание, простая функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992
	REG		0,999	0,999	0,999	0,994	0,999	0,999	0,999	0,999	0,999	0,993
	RLOO-P	без разбиения, псевдонаблюдения, адаптивная функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
	REG		0,999	0,999	0,999	0,987	0,999	0,999	0,999	0,999	0,999	0,994
	RLOO-P	без разбиения, взвешивание, адаптивная функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992
	REG		0,987	0,993	0,993	0,998	0,989	0,993	0,995	0,995	0,998	0,999
разреженный LS–SVM	регулярности	D-опт. план	0,999	0,999	0,999	0,995	0,999	0,998	0,989	0,994	0,981	0,982
		замена	0,999	0,999	0,999	0,999	0,999	0,999	0,989	0,999	0,999	0,999
		исключение	0,999	0,999	0,999	0,995	0,999	0,998	0,989	0,994	0,981	0,982
		включение	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
		Add/Del	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,982
		Del/Add	0,999	0,999	0,999	0,995	0,999	0,998	0,989	0,994	0,981	0,982
	стабильности	D-опт. план	0,998	0,992	0,975	0,967	0,951	0,973	0,928	0,911	0,774	0,498
		замена	0,999	0,999	0,983	0,987	0,983	0,980	0,989	0,989	0,701	0,277
		исключение	0,998	0,992	0,975	0,967	0,951	0,973	0,928	0,911	0,774	0,498
		включение	0,999	0,992	0,975	0,967	0,999	0,999	0,928	0,911	0,774	0,498
		Add/Del	0,999	0,992	0,975	0,967	0,999	0,973	0,928	0,911	0,774	0,498
		Del/Add	0,998	0,992	0,975	0,967	0,951	0,973	0,928	0,911	0,774	0,498
	согласованности	D-опт. план	0,999	0,999	0,985	0,921	0,549	0,383	0,984	0,919	0,859	0,992
		замена	0,680	0,991	0,986	0,999	0,914	0,998	0,967	0,234	0,168	0,675
		исключение	0,999	0,988	0,999	0,565	0,999	0,383	0,984	0,919	0,859	0,992
		включение	0,999	0,999	0,985	0,921	0,549	0,383	0,984	0,991	0,985	0,992
		Add/Del	0,999	0,999	0,985	0,921	0,549	0,383	0,984	0,919	0,859	0,992
		Del/Add	0,999	0,988	0,999	0,565	0,741	0,994	0,984	0,919	0,859	0,992

Таблица 4.9 – Значение R^2 при температурном режиме 325 К

Тип модели	Критерий	Вариант разбиения	Количество точек в тестовой части в %									
			5	10	15	20	25	30	35	40	45	50
МНК		без разбиения	0,935	0,935	0,935	0,935	0,935	0,935	0,935	0,935	0,935	0,935
обычный LS–SVM	LOO	без разбиения	0,828	0,828	0,828	0,828	0,828	0,828	0,828	0,828	0,828	0,828
	K-FOLD	без разбиения	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
робастный LS–SVM	RLOO-P	без разбиения, псевдонаблюдения, простая функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991
	REG		0,999	0,999	0,999	0,991	0,999	0,999	0,999	0,999	0,999	0,994
	RLOO-P	без разбиения, взвешивание, простая функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992
	REG		0,999	0,999	0,999	0,994	0,999	0,999	0,999	0,999	0,999	0,993
	RLOO-P	без разбиения, псевдонаблюдения, адаптивная функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
	REG		0,999	0,999	0,999	0,987	0,999	0,999	0,999	0,999	0,999	0,994
	RLOO-P	без разбиения, взвешивание, адаптивная функция потерь	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
	RLOO		0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992	0,992
	REG		0,987	0,993	0,993	0,998	0,989	0,993	0,995	0,995	0,998	0,999
разреженный LS–SVM	регулярности	D-опт. план	0,999	0,999	0,999	0,995	0,999	0,998	0,989	0,994	0,981	0,982
		замена	0,999	0,999	0,999	0,999	0,999	0,999	0,989	0,999	0,999	0,999
		исключение	0,999	0,999	0,999	0,995	0,999	0,998	0,989	0,994	0,981	0,982
		включение	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
		Add/Del	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,982
		Del/Add	0,999	0,999	0,999	0,995	0,999	0,998	0,989	0,994	0,981	0,982
	стабильности	D-опт. план	0,998	0,992	0,975	0,967	0,951	0,973	0,928	0,911	0,774	0,498
		замена	0,999	0,999	0,983	0,987	0,983	0,980	0,989	0,989	0,701	0,277
		исключение	0,998	0,992	0,975	0,967	0,951	0,973	0,928	0,911	0,774	0,498
		включение	0,999	0,992	0,975	0,967	0,999	0,999	0,928	0,911	0,774	0,498
		Add/Del	0,999	0,992	0,975	0,967	0,999	0,973	0,928	0,911	0,774	0,498
		Del/Add	0,998	0,992	0,975	0,967	0,951	0,973	0,928	0,911	0,774	0,498
	согласованности	D-опт. план	0,999	0,999	0,985	0,921	0,549	0,383	0,984	0,919	0,859	0,992
		замена	0,680	0,991	0,986	0,999	0,914	0,998	0,967	0,234	0,168	0,675
		исключение	0,999	0,988	0,999	0,565	0,999	0,383	0,984	0,919	0,859	0,992
		включение	0,999	0,999	0,985	0,921	0,549	0,383	0,984	0,991	0,985	0,992
		Add/Del	0,999	0,999	0,985	0,921	0,549	0,383	0,984	0,919	0,859	0,992
		Del/Add	0,999	0,988	0,999	0,565	0,741	0,994	0,984	0,919	0,859	0,992

Таблица 4.10 – Значение R^2 при температурном режиме 321 К

Тип модели	Критерий	Вариант разбиения	Количество точек в тестовой части в %									
			5	10	15	20	25	30	35	40	45	50
МНК		без разбиения	0,661	0,661	0,661	0,661	0,661	0,661	0,661	0,661	0,661	0,661
обычный LS–SVM	LOO	без разбиения	0,315	0,315	0,315	0,315	0,315	0,315	0,315	0,315	0,315	0,315
	K-FOLD	без разбиения	0,943	0,943	0,943	0,924	0,941	0,941	0,941	0,941	0,943	0,924
робастный LS–SVM	RLOO-P	без разбиения, псевдонаблюдения, простая функция потерь	0,943	0,943	0,943	0,943	0,943	0,943	0,943	0,943	0,943	0,943
	RLOO		0,849	0,849	0,849	0,849	0,849	0,849	0,849	0,849	0,849	0,849
	REG		0,939	0,756	0,933	0,756	0,928	0,756	0,933	0,879	0,943	0,881
	RLOO-P	без разбиения, взвешивание, простая функция потерь	0,926	0,926	0,926	0,926	0,926	0,926	0,926	0,926	0,926	0,926
	RLOO		0,765	0,765	0,765	0,765	0,765	0,765	0,765	0,765	0,765	0,765
	REG		0,796	0,865	0,796	0,865	0,862	0,865	0,796	0,742	0,887	0,763
	RLOO-P	без разбиения, псевдонаблюдения, адаптивная функция потерь	0,942	0,942	0,942	0,942	0,942	0,942	0,942	0,942	0,942	0,942
	RLOO		0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941
	REG		0,934	0,890	0,934	0,890	0,920	0,890	0,934	0,881	0,940	0,882
	RLOO-P	без разбиения, взвешивание, адаптивная функция потерь	0,928	0,928	0,928	0,928	0,928	0,928	0,928	0,928	0,928	0,928
	RLOO		0,790	0,790	0,790	0,790	0,790	0,790	0,790	0,790	0,790	0,790
	REG		0,502	0,911	0,643	0,911	0,818	0,911	0,643	0,640	0,848	0,694
разреженный LS–SVM	регулярности	D-опт. план	0,937	0,877	0,934	0,881	0,926	0,884	0,931	0,882	0,922	0,731
		замена	0,943	0,930	0,941	0,937	0,925	0,925	0,879	0,930	0,939	0,885
		исключение	0,943	0,877	0,934	0,881	0,926	0,884	0,931	0,882	0,922	0,731
		включение	0,943	0,935	0,935	0,931	0,925	0,925	0,885	0,925	0,844	0,883
		Add/Del	0,943	0,926	0,926	0,925	0,926	0,934	0,936	0,882	0,922	0,731
		Del/Add	0,943	0,877	0,934	0,881	0,926	0,884	0,931	0,882	0,922	0,731
	стабильности	D-опт. план	0,943	0,853	0,916	0,902	0,248	0,929	0,326	0,811	0,757	0,685
		замена	0,934	0,935	0,933	0,931	0,920	0,935	0,932	0,891	0,757	0,685
		исключение	0,934	0,853	0,916	0,902	0,248	0,929	0,326	0,811	0,757	0,685
		включение	0,942	0,853	0,938	0,938	0,248	0,929	0,838	0,830	0,757	0,685
		Add/Del	0,934	0,853	0,936	0,938	0,937	0,936	0,838	0,830	0,757	0,685
		Del/Add	0,934	0,853	0,916	0,902	0,248	0,929	0,326	0,811	0,757	0,685
	согласованности	D-опт. план	0,902	0,874	0,898	0,870	0,566	0,960	0,211	0,194	0,327	0,458
		замена	0,231	0,887	0,892	0,942	0,856	0,529	0,273	0,354	0,241	0,182
		исключение	0,231	0,893	0,898	0,852	0,915	0,273	0,280	0,887	0,895	0,193
		включение	0,902	0,874	0,898	0,870	0,566	0,960	0,211	0,194	0,327	0,458
		Add/Del	0,902	0,874	0,898	0,870	0,566	0,960	0,211	0,194	0,327	0,458
		Del/Add	0,231	0,893	0,898	0,852	0,915	0,823	0,873	0,194	0,327	0,458

Из результатов приведенных таблиц можно увидеть, что метод LS–SVM в большинство случаев выигрывает метод МНК. Самые плохие решения получились при обычной LS–SVM модели с использованием критерия скользящего контроля для подбора метапараметров. Также в разреженных

решениях при большем количестве точек в тестовой части путем разбиения выборки с использованием критерия стабильности и согласованности проигрывают решения, полученные таким же способом, но с использованием критерия регулярности.

Результаты построенных регрессионных моделей на основе выборок по определению равновесной концентрации ионов серебра приведены для различных температурных режимов приведены в рисунках **4.6**–**4.11**. Обозначения, используемые на графиках, являются: Y – исходные данные, $\hat{Y}^*$ – решение по МНК, $\hat{Y}(O)$ – обычное решение, $\hat{Y}(R)$ – робастное решение и $\hat{Y}(S)$ – разреженное решение.

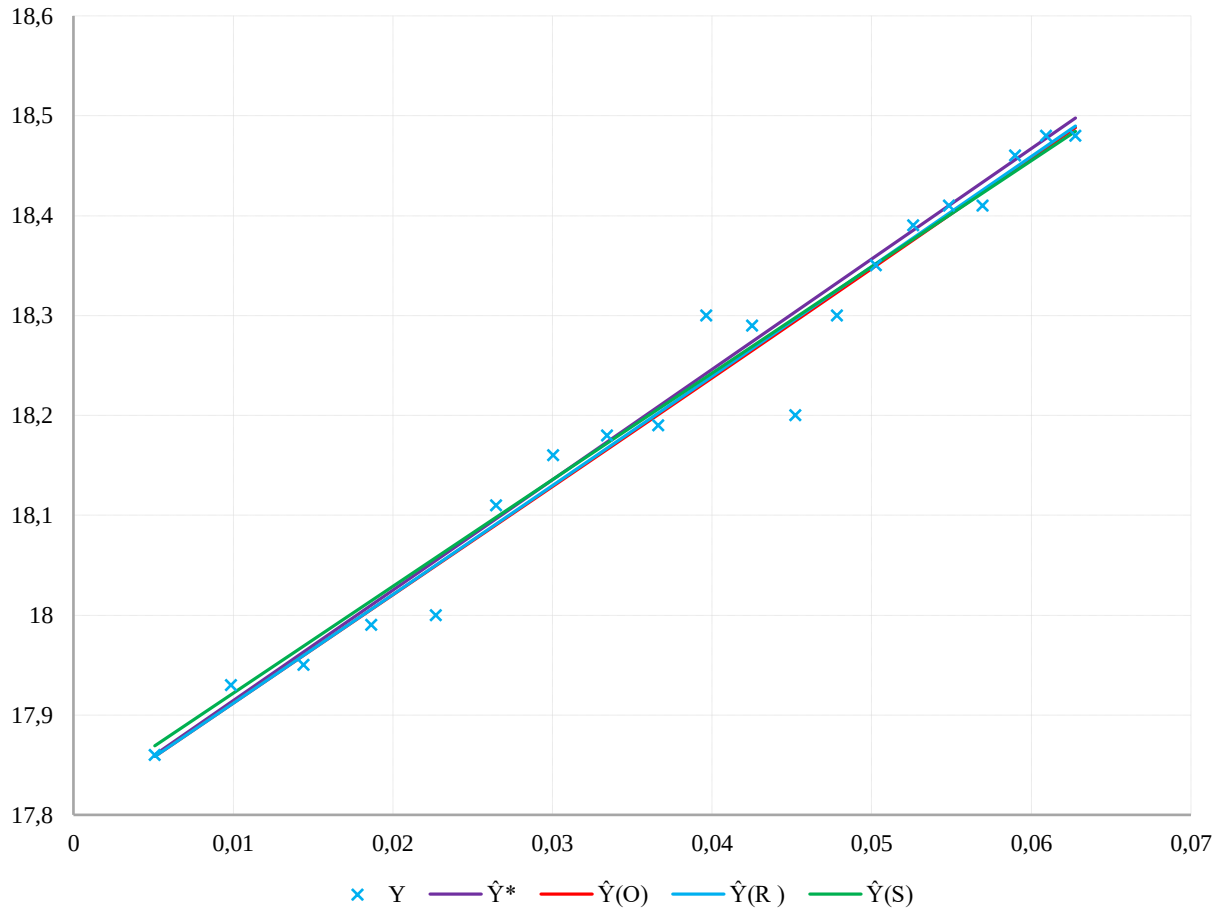


Рисунок 4.6. Зависимости ΔE от –lg[L] при температуре 321 К и 2-50 объем. % спирта

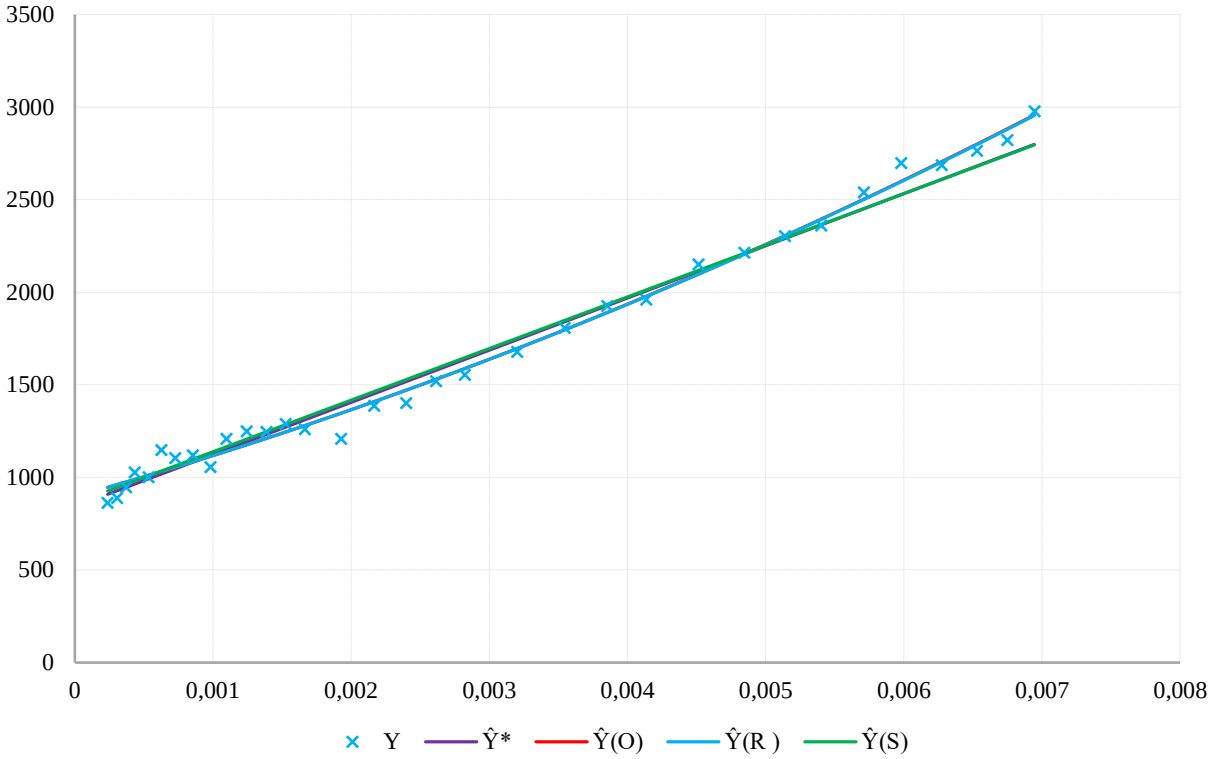


Рисунок 4.7. Зависимости ΔE от –lg[L] при температуре 325 К и 2-50 объем. % спирта

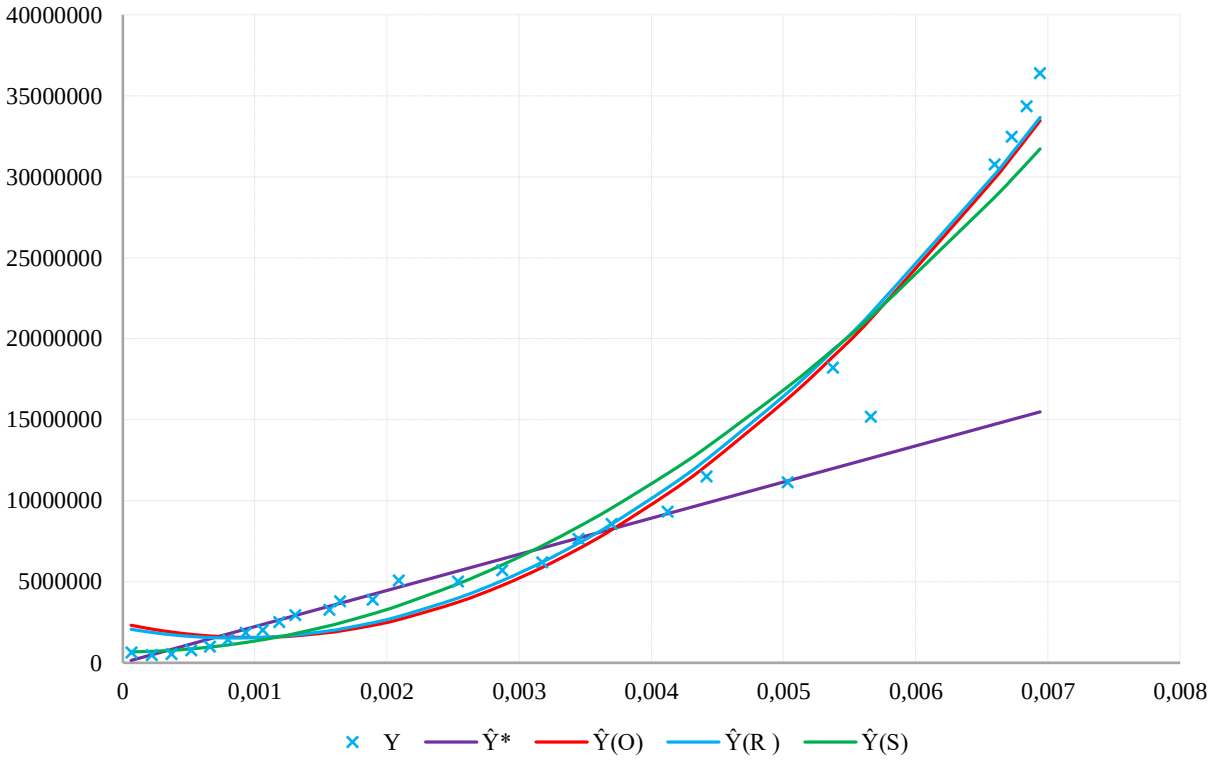


Рисунок 4.8. Зависимости ΔE от –lg[L] при температуре 335 К и 2-50 объем. % спирта

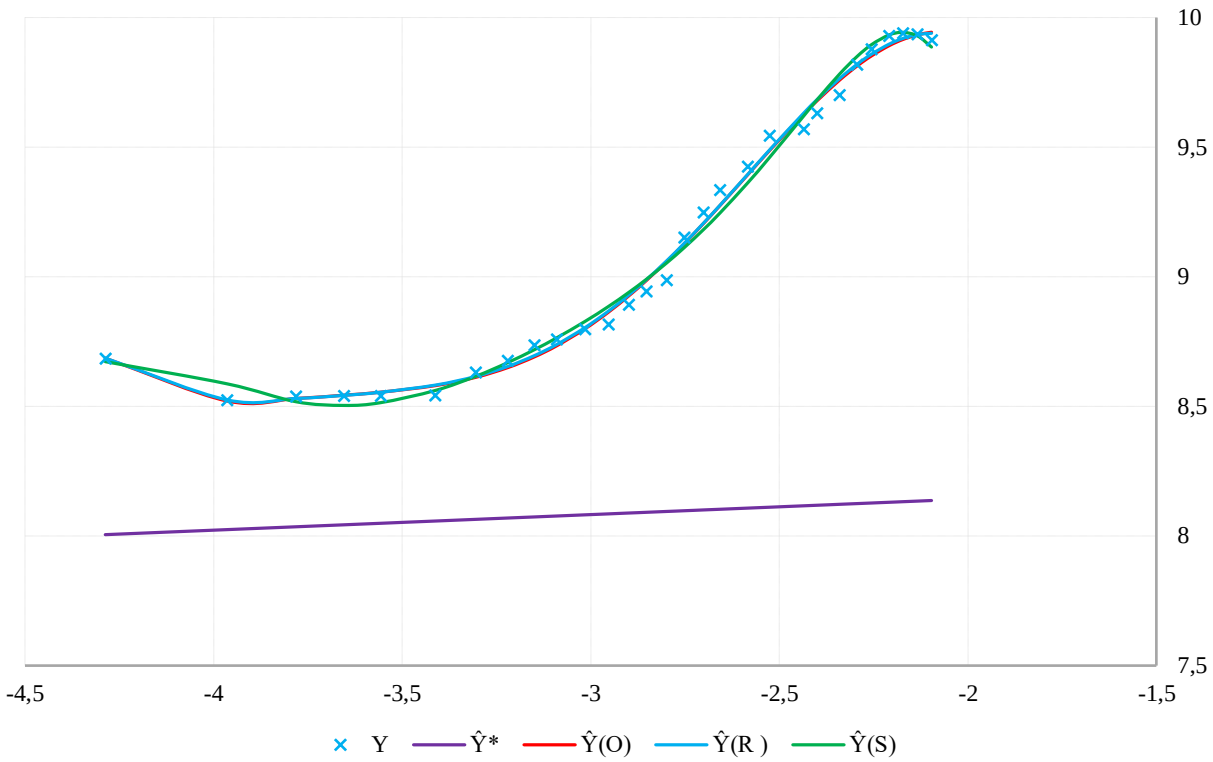


Рисунок 4.9. Зависимости ΔE от –lg[L] при температуре 321 К и 3-75 объем. % спирта

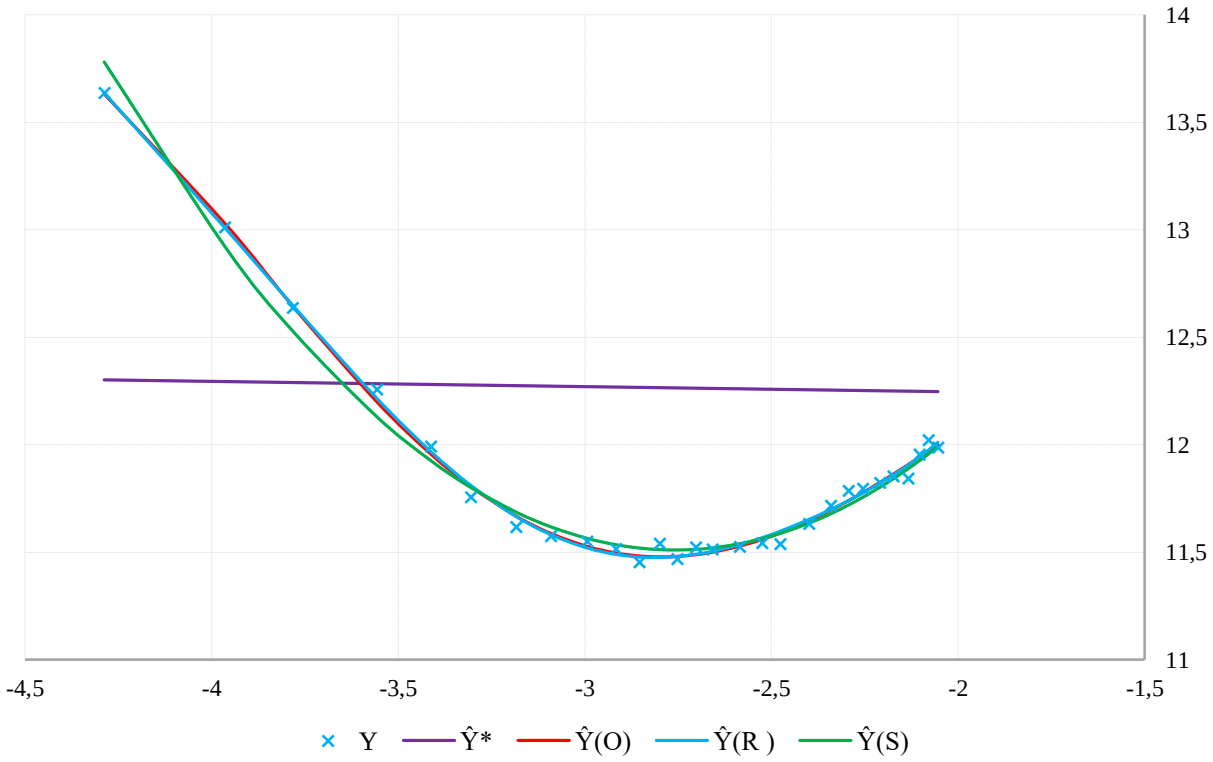


Рисунок 4.10. Зависимости ΔE от –lg[L] при температуре 325 К и 3-75 объем. % спирта

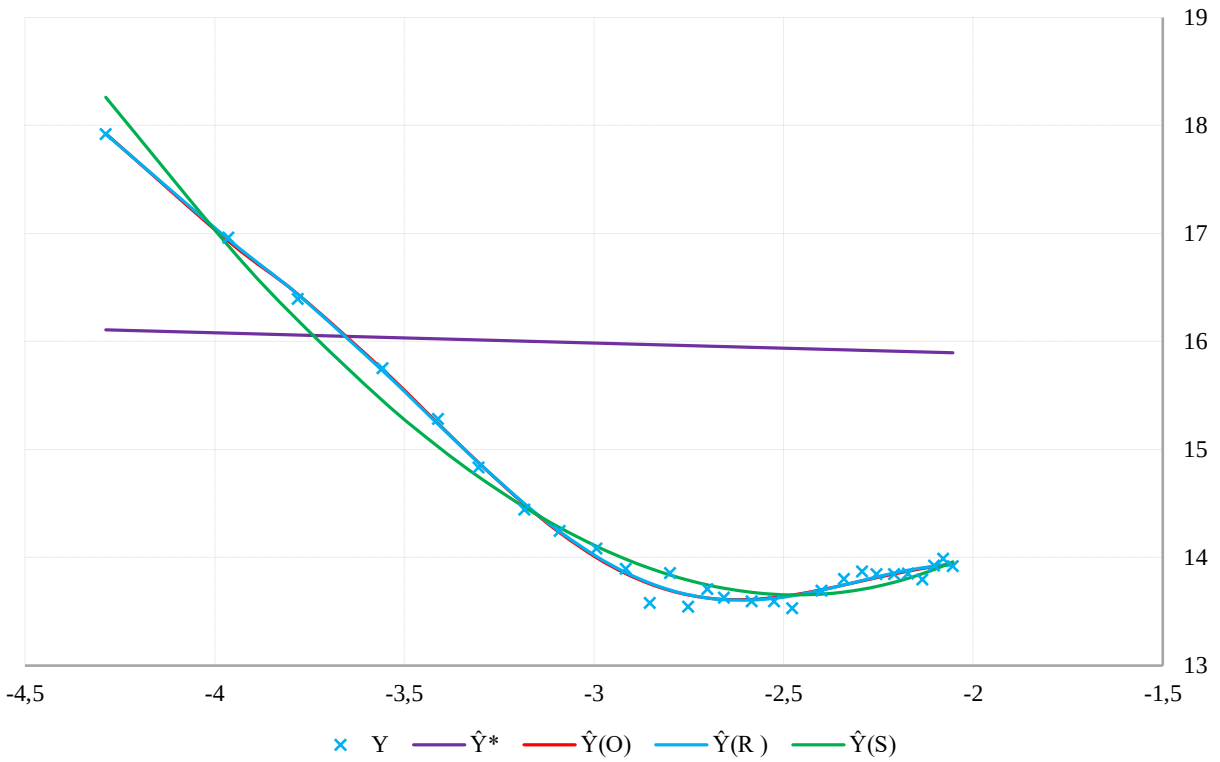


Рисунок 4.11. Зависимости ΔE от –lg[L] при температуре 335 К и 3-75 объем. % спирта

Анализ приведённых графиков показывает, что наиболее точно описывают данные как обычные, так и робастные модели. Они демонстрируют высокую точность аппроксимации и устойчивость к выбросам, что делает их предпочтительными в большинстве случаев.

Разреженные модели, хотя и обладают определёнными преимуществами, в некоторых ситуациях плохо описывают данные. Однако в целом они также показывают приемлемые результаты, особенно в условиях, когда требуется снизить сложность модели или улучшить её интерпретируемость.

Метод наименьших квадратов (МНК) в большинстве случаев продемонстрировал неудовлетворительные результаты. Он плохо справляется с аппроксимацией данных, особенно в ситуациях, когда точки не располагаются вдоль линейной зависимости. В таких случаях МНК даёт значительные отклонения и теряет точность, что ограничивает его применение в задачах, где данные содержат выбросы или имеют сложную структуру.

4.4 Выводы

В данной главе рассмотрены способы применения метода LS–SVM для решения практических задач. В качестве объекта исследования были использованы известные выборки: LIDAR – которая использует отражение света, излучаемого лазером, для обнаружения химических компонентов в атмосфере, и Motorcycle – которая содержит результаты проведенных испытаний шлемов на ударную прочность с использованием манекенов, установленных на мотоциклах. Кроме того, рассмотрен способ применения метода LS–SVM для определения комплексообразования равновесной концентрации химических элементов.

Основные полученные результаты:

1. Проведены исследования с использованием известных выборок LIDAR и Motorcycle и определены эффективность использования алгоритма LS–SVM для построения регрессии с использованием данных выборок.

2. Проведены исследования с использованием данных по образованию комплексов переходных металлов. Сравнены полученные на основе алгоритма LS–SVM решения с результатами, полученными алгоритмом МНК.

3. Определены степени гладкости полученных робастных и разреженных моделей.

ГЛАВА 5. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ПОСТРОЕНИЯ LS–SVM РЕГРЕССИИ

В рамках диссертационного исследования автор принимал участие в разработке программной системы «ПОСТРОЕНИЕ РОБАСТНЫХ И РАЗРЕЖЕННЫХ РЕШЕНИЙ МЕТОДОМ LS SVM “Robast_Sparse_LS-SVM”» для построения регрессии методом LS–SVM. Программа разработана в языке программирования Delphi 7 в среде Borland с использованием библиотек генерации случайных чисел по нормальному закону и визуальных нестандартных компонент, предназначенных для работы с вещественными числами. Автор разработал интерфейс программы и запрограммировал все разработанные методы построения робастных и разреженных регрессионных моделей, приведенные в главах 2 и 3.

Актуальность программного обеспечения

На практике часто приходится решать задачи построения регрессии непараметрическими методами. Существуют множество систем для решения таких задач, но пользователь не всегда имеет возможность воспользоваться такими программными продуктами из-за высокой стоимости и закрытости исходного кода таких продуктов. Поэтому пользователю при использовании таких продуктов приходится ограничиваться набором уже реализованных методов. Для использования всей функциональности и возможности алгоритмов становится востребованным программный продукт, который расширяет возможность и функционал известных пакетов программ.

Назначение программного продукта

Разработанный программный продукт направлен на построение регрессии для данных (выборок) с присутствием шума (помехи) и/или больших выбросов (засорений). Программный продукт предоставляет возможность построить робастные регрессионные модели с использованием псевдонаблюдений и весовой функции и разреженные регрессионные модели путем разбиения выборки на части *D*–оптимальным планированием и/или

использованием различных критериев оценки качества моделей. Имеется возможность оценить качество полученных моделей различными значениями критериев, таких как: значение MSE, значение RSS, значение R и значение R^2. Программный продукт могут использовать различные вузы, научные коллективы и научные сотрудники различных научных организаций.

Интерфейс и возможности программного продукта

Разработанный программный продукт предоставляет следующие возможности пользователю:

1. Импорт/экспорт данных в форматах xls, xlsx, txt, автоматически генерировать данные через саму программу или ввести данные вручную;
2. Генерировать шум (помеха) и засорение (большие выбросы) (рис. **5.1**, **5.2**);
3. Выбрать ядерную функцию и различные критерии для подбора параметров алгоритма (рис. **5.3**);
4. Выбрать тип разбиения выборки и/или ее оптимизации (рис. **5.3**, **5.5**);
5. Выбрать тип регрессионной модели (обычная, робастная, разреженная) (рис. **5.3**);
6. Выбрать нужные значения для вывода;
7. Визуальное отображение результатов вычисления в виде таблиц и графиков (рис. **5.4**, **5.6**).

Интерфейс программы состоит из следующих вкладок:

- Данные;
- Шум (помехи);
- Настройки;
- Результаты;
- Результаты разбиения выборки;
- Графики.

Далее, на рисунках **5.1**–**5.6** приведены вкладки интерфейса программы.

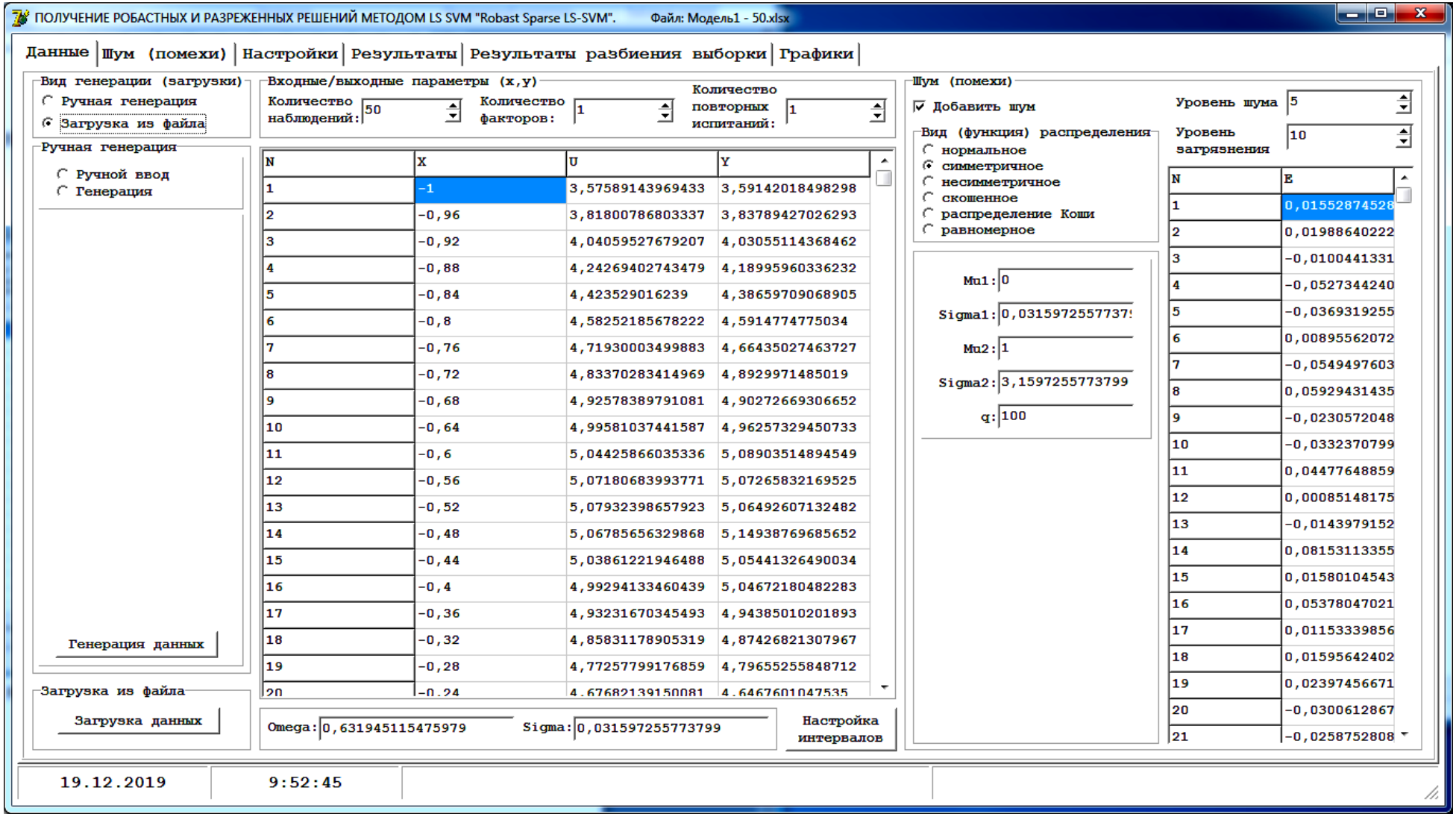


Рисунок 5.1. Вкладка «Данные»

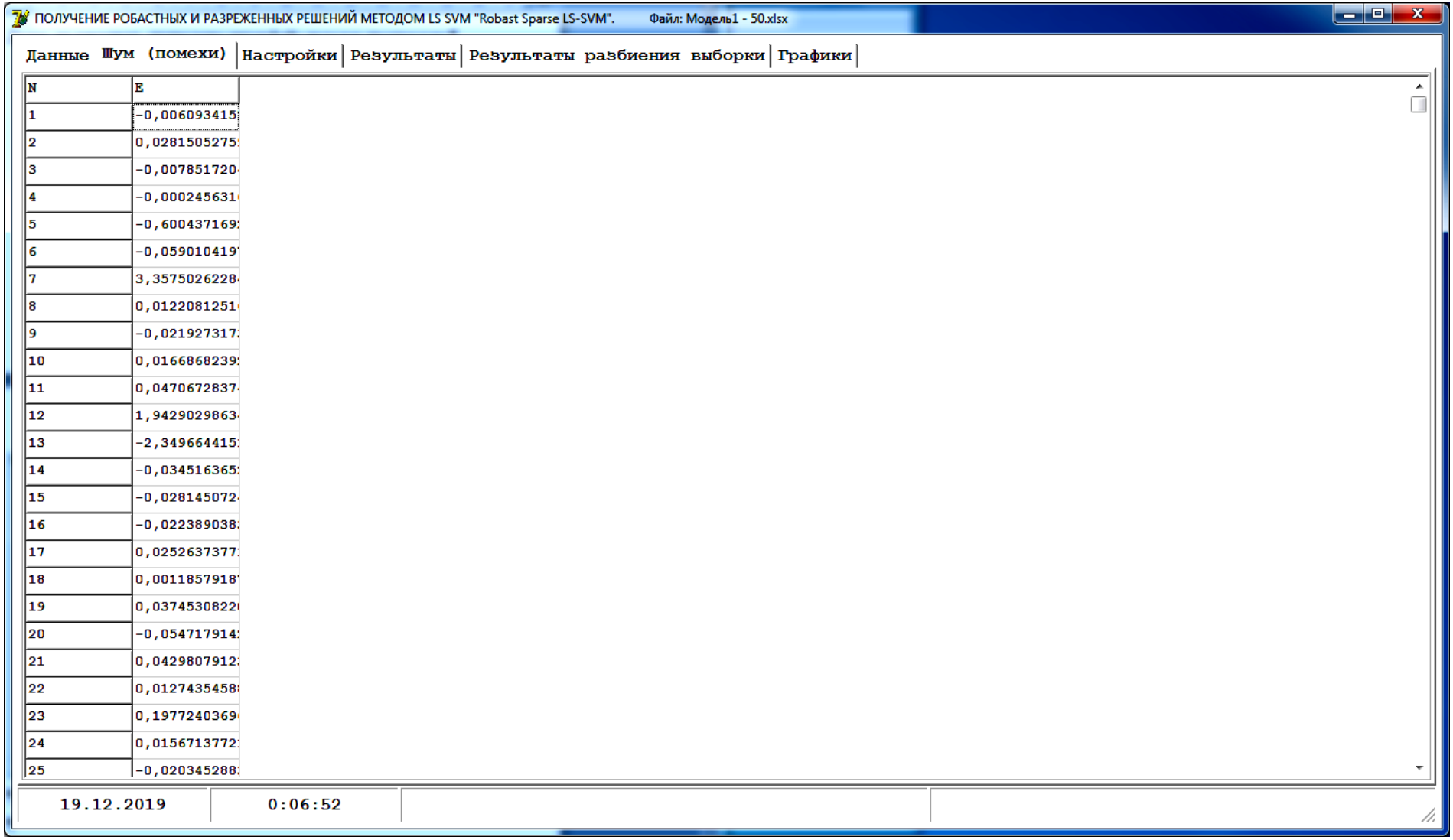


Рисунок 5.2. Вкладка «Шум (помехи)»

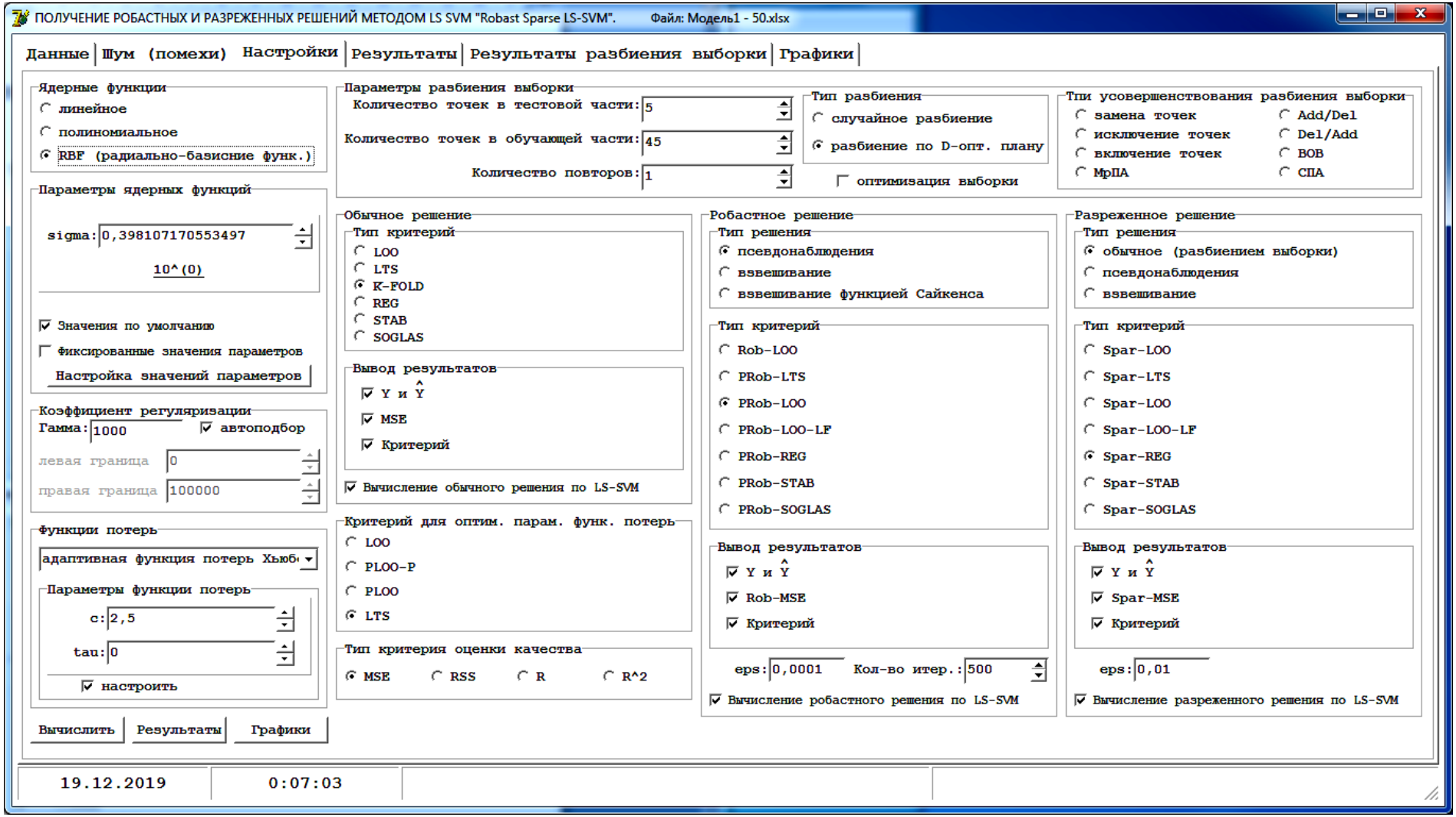


Рисунок 5.3. Вкладка «Настройки»

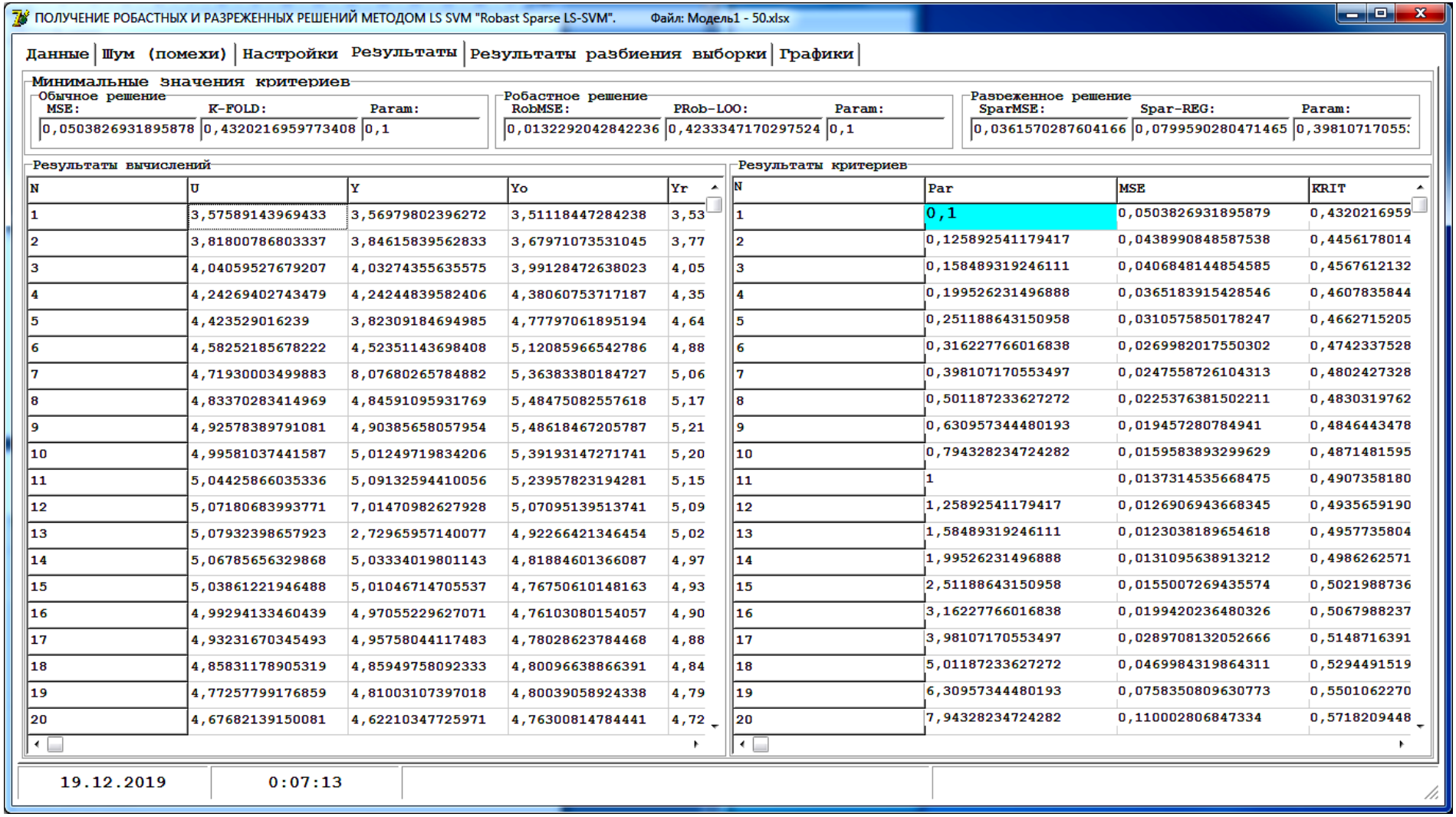
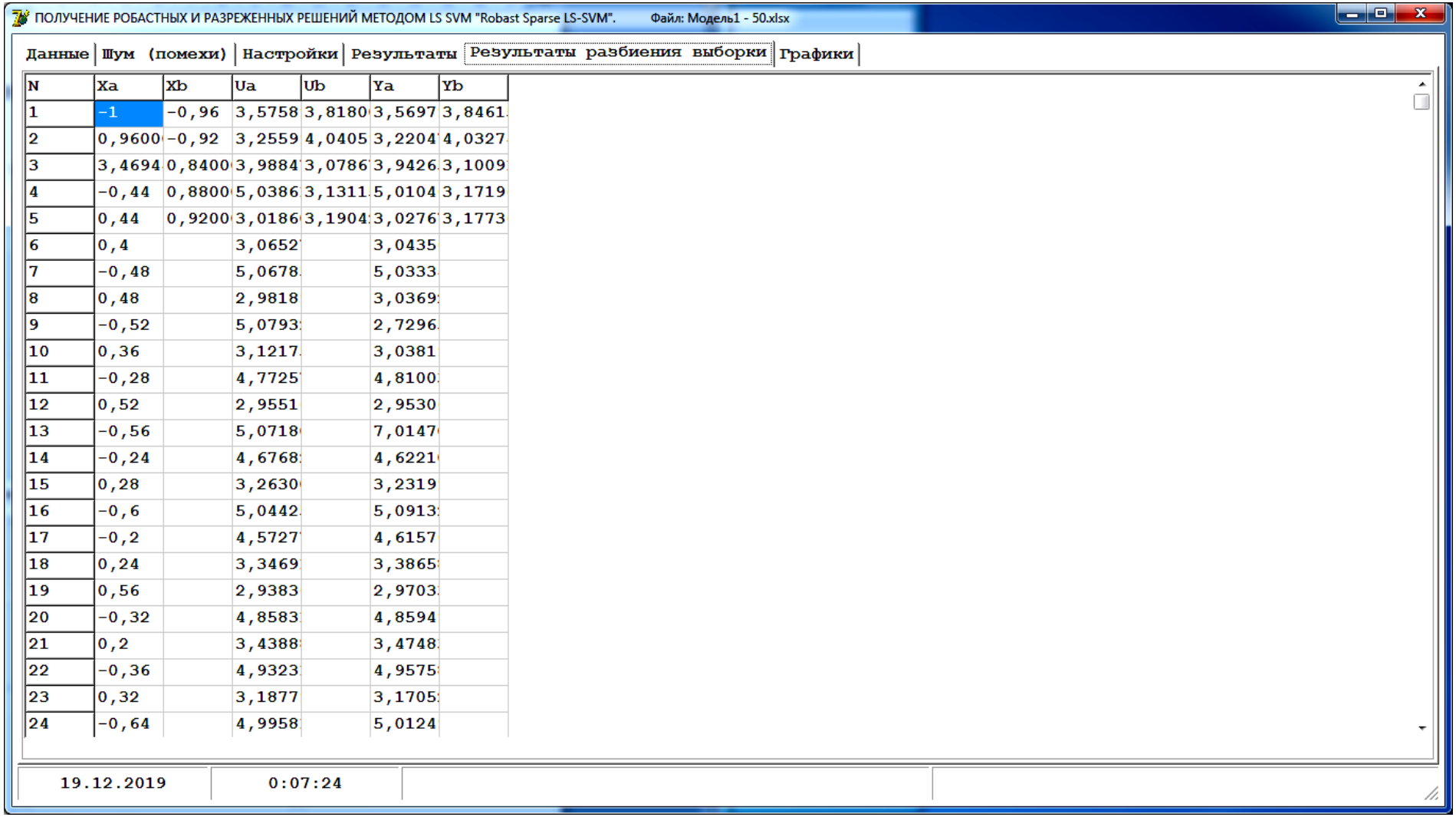


Рисунок 5.4. Вкладка «Результаты»



N	Xa	Xb	Ua	Ub	Ya	Yb
1	-1	-0,96	3,5758	3,8180	3,5697	3,8461
2	0,9600	-0,92	3,2559	4,0405	3,2204	4,0327
3	3,4694	0,8400	3,9884	3,0786	3,9426	3,1009
4	-0,44	0,8800	5,0386	3,1311	5,0104	3,1719
5	0,44	0,9200	3,0186	3,1904	3,0276	3,1773
6	0,4		3,0652		3,0435	
7	-0,48		5,0678		5,0333	
8	0,48		2,9818		3,0369	
9	-0,52		5,0793		2,7296	
10	0,36		3,1217		3,0381	
11	-0,28		4,7725		4,8100	
12	0,52		2,9551		2,9530	
13	-0,56		5,0718		7,0147	
14	-0,24		4,6768		4,6221	
15	0,28		3,2630		3,2319	
16	-0,6		5,0442		5,0913	
17	-0,2		4,5727		4,6157	
18	0,24		3,3469		3,3865	
19	0,56		2,9383		2,9703	
20	-0,32		4,8583		4,8594	
21	0,2		3,4388		3,4748	
22	-0,36		4,9323		4,9575	
23	0,32		3,1877		3,1705	
24	-0,64		4,9958		5,0124	

Рисунок 5.5. Вкладка «Результаты разбиения выборки»

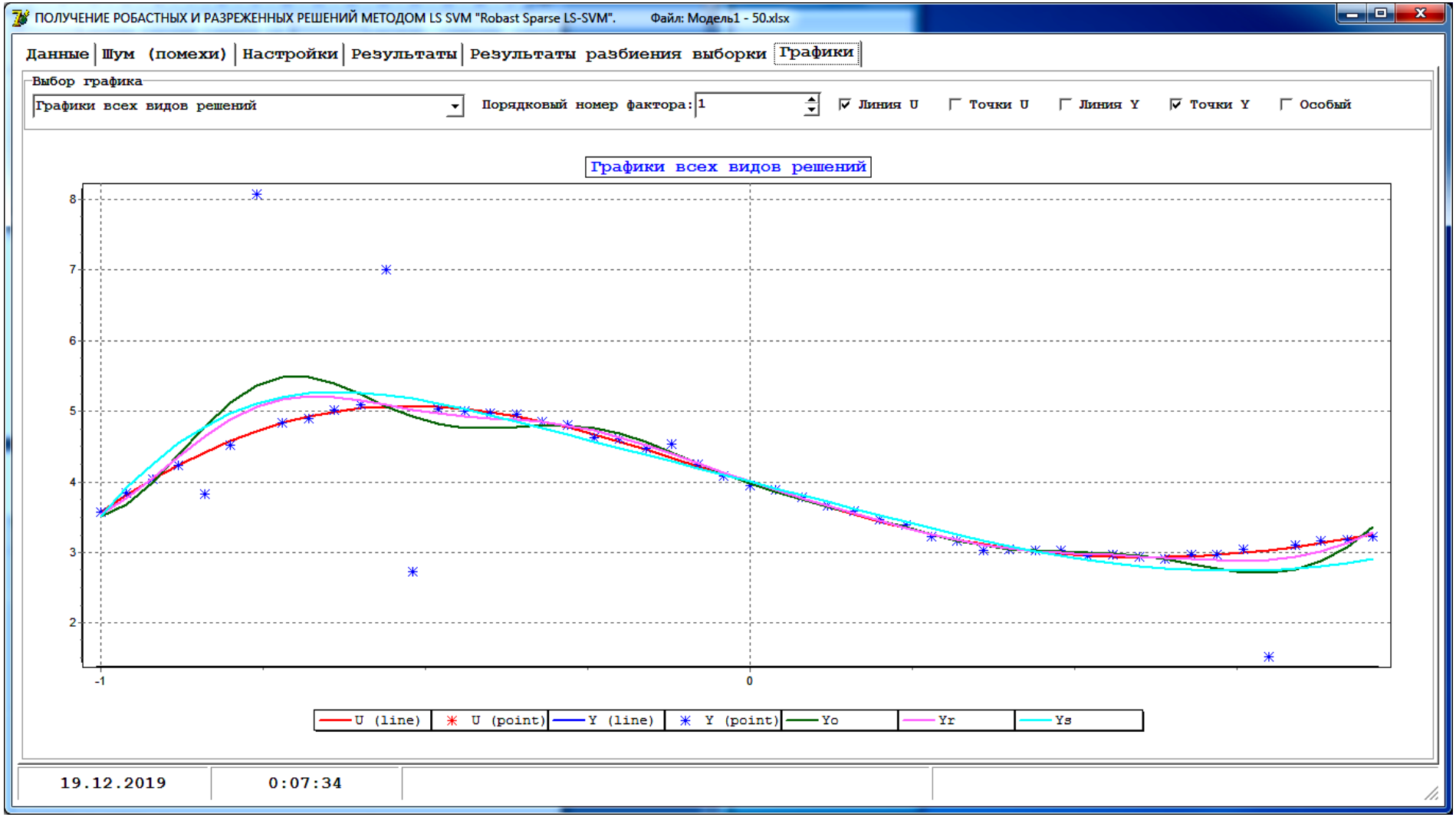
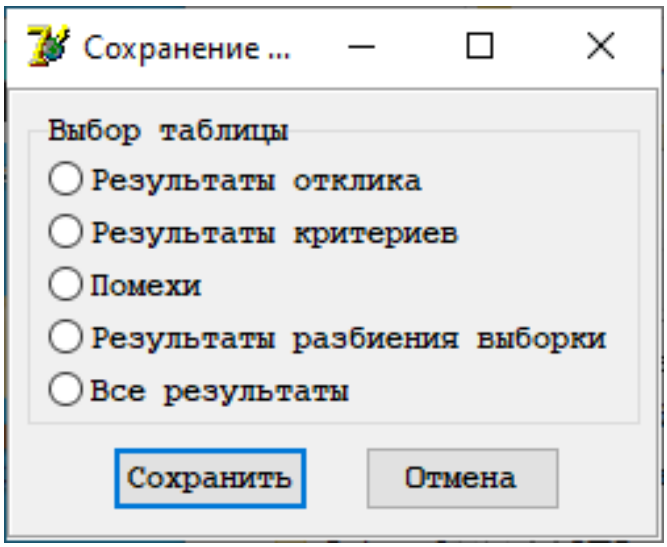


Рисунок 5.6. Вкладка «Графики»

Кроме приведенных вкладок в программе имеется функция сохранения результатов в котором по желанию и/или необходимости можно выбрать какие результаты необходимо сохранить. Данную функциональность приведем в следующем рисунке:



Как видно из рисунка имеется возможность сохранения результатов отклика, результатов критериев, результатов генерированному шума (помехи), результатов разбиения выборки на обучающую и тестовую части и сохранения всех результатов.

Работа с программой

Для построения регрессии методом LS–SVM приходится решать СЛАУ. Для выполнения данной операции в программе реализовано решение СЛАУ методами Гаусса-Жордана и LU–разложения. Также, для определения обратной матрицы могут использоваться эти же методы.

После выполнения операции вычисления на вкладке «Результаты» автоматически выводятся результаты полученной модели и значения выбранных критериев с пошаговым изменением значения основного параметра ядра (для полиномиального и RBF-ядра) в табличном виде, а также оптимальные значения выбранных критериев и параметра ядра в отдельных полях.

На вкладке «Графики» пользователь может выбрать вид графика, который хочет посмотреть.

В программе можно генерировать следующие типы зашумления данных:

- нормальное;
- несимметричное;
- скошенное;
- распределение Коши;
- равномерное.

Также можно генерировать следующие типы засорения данных (добавление больших выбросов):

- симметричное;
- несимметричное;
- скошенное.

Функции генерации шума и засорения и подбора оптимальных параметров алгоритма

Как было перечислено выше, в программе имеется возможность генерировать шум для зашумления данных. Генерация каждого из типов шумов выполняются соответствующими функциями. Для каждого из типов шумов приведем функции, используемые в программе по которым генерируются эти шумы.

Для моделирования помехи e_i, $i = 1, 2, \ldots, n$ необходимо использовать какой-либо доступный датчик псевдослучайных нормально распределенных величин. Моделируемая помеха должна иметь нулевое математическое ожидание и дисперсию σ^2. Дисперсию σ^2 помехи e целесообразно выбирать в виде некоторой доли ρ от мощности ω^2 сигнала $u = \eta\left(\underline{x}, \theta\right)$. Мощность сигнала определим

$$\omega = \frac{\left(u - \bar{u}\right)^T \left(u - \bar{u}\right)}{n - 1},$$

где u – вектор истинных значений отклика, $\bar{u}$ – вектор, все элементы которого есть среднее значение сигнала по выборке. Долю ρ можно брать в пределах 5…15 %. Тогда значение дисперсии определяется по формуле $\sigma^2 = \frac{\omega \cdot \rho}{100}$.

Шум по **нормальному** типу генерируется по формуле:

$$e = N\left(\mu, \sigma^2\right),$$

где N – функция распределения по нормальному закону, μ – математическое ожидание входного сигнала.

Шум по **несимметричному** типу генерируется по формуле:

$$e = N_1\left(\mu, \sigma^2\right) + N_2\left(\mu_1, q \cdot \sigma^2\right),$$

где q – коэффициент, определяющий масштабирование шума.

Шум по **скошенному** типу генерируется по формуле:

$$e = N\left(\mu, \sigma^2\right) + \left(\left(\sigma^2 \cdot N(0,1) + \sqrt{1 - \left(\frac{\delta}{\sqrt{1+\delta^2}}\right)^2} \cdot N(0,1)\right) \cdot q \cdot \sigma^2 + \mu\right),$$

где δ – параметр асимметрии (скоса) распределения.

Шум по **распределению Коши** генерируется по формуле:

$$e = N\left(\mu, \sigma^2\right) + \left(\mu + \tan\left(\pi \cdot \left(Random - 0.5\right)\right)\right) \cdot \sigma^2,$$

где $Random$ – равномерно распределенное случайное число в диапазоне $[0;1]$.

Шум по **равномерному** типу генерируется по формуле:

$$e = N\left(\mu, \sigma^2\right) + \left(\mu + Random \cdot \left(\sigma^2 - \mu\right)\right).$$

Выводы

Разработанный программный продукт содержит реализацию методов и подходов построения LS–SVM регрессии, рассмотренных в 1–3 главах данной диссертационной работы. С его помощью можно получить обычную, робастную или разреженную регрессионную модель для различных выборок данных. В программе существуют функциональность по настройке различных параметров

алгоритма LS–SVM, таких как коэффициент регуляризации, внутренние параметры ядерных функций, способы разбиения исходной выборки на части и выбор внешних критериев для оценки качества результирующих моделей.

Аналог разработанного программного продукта используется в научно-исследовательском институте Таджикского национального университета для построения моделей, которые используются для определения образования комплексов переходных металлов с производными тиомочевины в водно и водно-органических растворах.

ЗАКЛЮЧЕНИЕ

В диссертационной работе в соответствии с поставленными задачами исследования были получены следующие основные результаты:

1. Рассмотрены критерии оценки качества моделей LOO CV и K-FOLD CV для подбора метапараметров алгоритма LS–SVM. Были показаны эффективность использования данных критериев по проведенным вычислительным экспериментам.

2. Предложен адаптивный вариант функции потерь Хьюбера для получения псевдонаблюдений и взвешивания наблюдений.

3. Проведен сравнение результатов робастных моделей полученных на основе методов псевдонаблюдений и взвешивания с использованием функций потерь Эндрюса и биквадратной Тьюки и установлены степень их эффективности.

4. Предложены робастные варианты критерия скользящего контроля: RLOO-P и RLOO, которые служат для подбора метапараметров алгоритма LS–SVM и оценки качества получаемых робастных моделей.

5. На основе LS–SVM предложен метод построения робастных решений с использованием псевдонаблюдений и функции весов на основе обычной и адаптивной функций потерь Хьюбера. Проведенные исследования продемонстрировали хорошие возможности использования адаптивной функции потерь Хьюбера, предложенной автором, для получения устойчивых решений с малым смещением в условиях засорения наблюдений.

6. Для получения разреженного решения на основе метода LS–SVM предложены различные способы разбиения выборки на обучающую и тестовую части с использованием D–оптимального разбиения и внешних критериев оценки качества моделей. Для каждого из способов разбиения выборки предложены последовательные алгоритмы.

7. Рассмотрены решения прикладных задач с использованием известных выборок и данные по определению образования комплексов переходных металлов с производными теомочевины в водно и водно-органических растворах. По

результатам проведенных вычислительных экспериментов определилась эффективность использования разработанных автором алгоритмов построения робастной и разреженной LS–SVM регрессии для решения прикладных задач.

8. Разработано программное обеспечение для построения робастных и разреженных LS–SVM регрессий. Программное обеспечение имеет государственную регистрацию №2018619675, предусмотренную для программ ЭВМ.

Научные результаты диссертационной работы и разработанное программное обеспечение внедрены в научно-исследовательском институте Таджикского государственного университета, а также нашли практическое применение в учебном процессе на факультете прикладной математики и информатики ФГБОУ ВО «Новосибирский государственный технический университет», что подтверждается соответствующими актами о внедрении.

Разработанные методы могут использоваться при проведении научных исследований с целью восстановления зависимостей в прикладных выборках данных.

СПИСОК ЛИТЕРАТУРЫ

1. Cristianini N. An introduction to support Vector Machines: and other kernel-based learning methods / N. Cristianini, J. Shawe-Taylor – Cambridge: Cambridge University Press, 2000. – 189 p. – ISBN: 0-521-78019-5.

2. Györfi L. A Distribution-Free Theory of Nonparametric Regression / L. Györfi M. Kohler, A. Krzyzak, H. Walk, N.Y.: Springer-Verlag, 2002. – 656 p. – ISBN: 0-387-95441-4.

3. Schölkopf B. Advances in Kernel Methods – Support Vector Learning / B. Schölkopf, C.J.C.Burges, A.J. Smola, Eds., Cambridge, MA: MIT Press, – 386 p.– ISBN-10: 0262194163.

4. Schölkop B. Nonlinear component analysis as a kernel eigenvalue problem / B. Schölkop, A.J. Smola, K.R. Müller // Neural Computation, 1998. – vol. 10, no. 5, pp. 1299–1319.

5. Vapnik V. Statistical Learning Theory / V. Vapnik, NY: John Wiley, 1998. – 768 p. – ISBN: 978-0-471-03003-4.

6. Boser B.A, Guyon I., Vapnik V.N. A training algorithm for optimal margin classifiers // Proc. of the 5th Annual ACM Workshop on Computational Learning Theory (COLT). – Pittsburgh: ACM Press, 1992. – pp. 144–152.

7. Support vector regression machines / H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, Eds. // In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems, – Cambridge: MA, MIT Press, 1997. – vol. 9, – pp. 155–161.

8. Vapnik V. Pattern recognition using generalized portrait method / V. Vapnik, A. Lerner, Automation and Remote Control, 1963. – vol. 24, – no. 6, – pp. 774 – 780.

9. Vapnik V. A note on one class of perceptrons / V. Vapnik, A. Chervonenkis, Automation and Remote Control, 1964. – vol. 25, – no. 1, – pp. 821 – 837.

10. Айзерман М. А. Метод потенциальных функций в теории обучения машин / М. А. Айзерман, Браверман Э. М., Розоноэр Л. И – М.: Наука, 1970. – 384 с.

11. Вапник В. Н. Теория распознавания образов. Статистические проблемы обучения / В. Н. Вапник, А. Я. Червоненкис, – М.: Наука, 1974. – 416 с.

12. Вапник В. Н. Восстановление зависимостей по эмпирическим данным / В. Н. Вапник, – М.: Наука, 1979. – 448 с.

13. Smola A. J. A Tutorial on Support Vector Regression / A. J. Smola, B. Schölkopf, // NeuroCOLT, Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK, 1998.

14. Cortes C. Support vector networks / C. Cortes and V. Vapnik // Machine Learning ML: 1995. – vol. 20, – no. 3, – pp. 273 – 297.

15. Suykens, J.A.K. Least squares support vector machine classifiers / J.A.K. Suykens, J. Vandewalle // Neural processing letters. – 1999. – vol. 9, Iss. 3. – pp. 293–300.

16. Weighted least squares support vector machines: robastness and sparse approximation / J.A.K. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle. Neurocomputing. – 2002. – vol. 48, – pp. 85–105.

17. De Kruif B.J. Pruning error minimization in least squares support vector machines / B.J. De Kruif T.J.A. De Vries, // IEEE Transactions on Neural Networks. – 2003. – vol. 14, – no. 3, – pp. 696–702.

18. Zeng X. SMO-based pruning methods for sparse least squares support vector machines / X. Zeng, X.W. Chen // IEEE Transactions on Neural Networks, – 2005. – vol. 16, – no. 6, – pp. 1541–1546.

19. Xia X.L. A novel sparse least squares support vector machine/ X.L. Xia, K. Li, G. Irwin // Mathematical Problems in Engineering – 2013. vol. 2013, – Article ID 602341, – 10 p.

20. Попов А.А. Построение регрессионных зависимостей с использованием алгоритма опорных векторов с адаптивными функциями потерь / А.А. Попов, А.С. Саутин // Научный вестник НГТУ. – 2011. – № 1 (42). – С. 17–26.

21. Popov A.A. Adaptive Huber loss function in support vector regression / A.A. Popov, A.S. Sautin // IFOST 2009. Proceedings of 2009 international forum on strategic

technologies, Vietnam, Ho Chi Minh City, Vietnam, 21–23 Oct. 2009. – Ho Chi Minh City, 2009. – Sess. 2. – pp. 114–118.

22. Попов А.А. Использование робастных функций потерь в алгоритме опорных векторов при решении задачи построения регрессии / А.А. Попов, А.С. Саутин // Научный вестник НГТУ. – 2009. – № 4 (37). – С. 45–56.

23. Espinoza M. LS-SVM Regression with Autocorrelated Errors / M. Espinoza, J. Suykens, B. De Moor // Proc. of the 14th IF AC Symposium on System Identification (SYSID). 2006. - Vol. 15. - P. 582-587.

24. Cherkassky V., Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression / V. Cherkassky, Y. Ma // Neural Networks. – 2004. – no. 17. – pp. 113–126.

25. Попов А.А. Определение параметров алгоритма опорных векторов при решении задачи построения регрессии / А.А. Попов, А.С. Саутин // Сборник научных трудов НГТУ. – 2008. – № 2(52). – С. 35–40.

26. Popov A.A. Selection of support vector machines parameters for regression using nested grids / A.A. Popov, A.S. Sautin // The Third International Forum on Strategic Technology. Novosibirsk, – 2008. – pp. 329–331.

27. Kernel Functions for Machine Learning Applications [Электронный ресурс]. – Режим доступа: http://crsouza.blogspot.ru/2010/03/kernel-functions-for-machine-learning.html.

28. Suykens J.A.K. Least Square Support Vector Machines / Johan A.K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, Joos Vandewalle. World Scientific, New Jersey-London-Singapore-Hong Kong, – 2002. – 290 p.

29. Brabanter J.D. LS-SVM Regression Modelling and its Applications / J.D. Brabanter. – Leuven (Heverlee): Katholieke Universiteit Leuven, – 2004. – 243 c.

30. Попов А.А. Построение регрессионных зависимостей с использованием квадратичной функции потерь в методе опорных векторов / А.А. Попов, Ш.А. Бобоев // Сборник научных трудов Новосибирского государственного технического университета. – 2015. – № 3 (81). – С. 69–78.

31. Бобоев Ш.А. Критерии подбора метапараметров алгоритма LS–SVM / Ш.А. Бобоев // Вестник филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. Серия естественных наук. – 2023. – Т. 1, № 2 (31). – С. 5–15.

32. Бобоев Ш.А. Подбор коэффициента регуляризации в методе LS–SVM / Ш.А. Бобоев // Материалы международной научно-практической конференции «XIII Ломоносовские чтения», посвященной 115-летию академика Бободжона Гафурова (28-29 апреля 2023 года). – Душанбе. – 2023. – С. 27–32.

33. Vapnik V.N. Estimation of Dependences Based on Empirical Data / V.N. Vapnik. – New York: Springer Verlag, 1982. – 399 pp.

34. Vapnik V.N. The Nature of Statistical Learning Theory / V.N. Vapnik. – New York: Springer Verlag, 1995. – 188 pp.

35. C.M. Huang, Y.J. Lee. Model selection for support vector machines via uniform design // Computational Statistics & Data Analysis. 2007. No. 52. P. 335-346.

36. Попов А. А. Планирование эксперимента в задачах структурного моделирования с использованием критерия скользящего прогноза / А. А. Попов // Заводская лаборатория. – 1996. – № 10. – С. 42-44.

37. J.A.K. Suykens. Regularization, Optimization, Kernels, and Support Vector Machines / J.A.K. Suykens, M. Signoretto, A. Argyriou (Eds.). – Chapman & Hall/CRC, Machine Learning & Pattern Recognition Series, Boca Raton US. – 2014. – 525 p.

38. Gavin C. Cawley. Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters / Gavin C. Cawley, Nicola L. C. Talbot. // Journal of Machine Learning Research. – 2007. – v. 8. – pp. 841–861.

39. Wentao Mao. Leave-one-out cross-validation-based model selection for multi-input multi-output support vector machine / Wentao Mao, Xiaoxia M, Yanbin Zheng, Guirong Yan // Neural Computing and Application. – 2014. – №2 (24). – pp. 441–451.

40. Pablo Rivas-Perea. A nonlinear least squares quasi-Newton strategy for LP-SVR hyper-parameters selection / Pablo Rivas-Perea, Juan Cota-Ruiz, Jose-Gerardo Rosiles. // International Journal of Machine Learning and Cybernetics. – 2014. – №4 (5). – pp. 579–597.

41. Amit Kumar Gupta. Optimisation of turning parameters by integrating genetic algorithm with support vector regression and artificial neural networks / Amit Kumar Gupta, Sharath Chandra Guntuku, Raghuram Karthik Desu, Aditya Balu // The International Journal of Advanced Manufacturing Technology. – 2015. – №1-4 (77). – pp. 331–339.

42. Гультяева Т.А. Методы статистического обучения в задачах регрессии и классификации: монография / Т. А. Гультяева, А. А. Попов, А. С. Саутин. – Новосибирск: Изд-во НГТУ, 2016. – 322 с.

43. B. Scholkopf, A.J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA. – 2002. – 644 p.

44. Jin C., Jin S.-W. Software reliability prediction model based on support vector regression with improved estimation of distribution algorithms // Applied Soft Computing. – 2014. – no. 15, pp. 113–120.

45. Utkin L.V. A framework for imprecise robust one-class classification models // International Journal of Machine Learning and Cybernetics. – 2014. no. 5(3), pp. 379–393.

46. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов / К. В Воронцов // Математические вопросы кибернетики. – М.: Физматлит, 2004. – №13. – С. 5 - 36.

47. M. Stone. Cross-validatory choice and assessment of statistical predictions. // Journal of the Royal Statistical Society. – 1974. – No. 36(1). – pp. 111–147.

48. Wahba G. A survey of some smoothing problems and the method of generalized cross-validation for solving them / G. Wahba // Application of Statistics. – 1977. – pp. 507-523.

49. Wahba G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV / G. Wahba // Advances in Kernel Methods – Support Vector Learning. – Cambridge: MIT Press. – 1999. – pp. 69-88.

50. Скользящий контроль [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=CV#.D0.9A.D0.BE.D0.BD.D1.82.D1.80.D0.BE.D0.BB.D1.8C_.D0.BF.D1.80.D0.B8_.D0.BD.D0.B0.D1.80.D0.B0.D1.81

.D1.82.D0.B0.D1.8E.D1.89.D0.B5.D0.B9_.D0.B4.D0.BB.D0.B8.D0.BD.D0.B5_.D0.BE.D0.B1.D1.83.D1.87.D0.B5.D0.BD.D0.B8.D1.8F.

51. Zhao Y. Fast Leave-one-out Evaluation and Improvement on Inference for LS-SVMs / Y. Zhao, C. K. Kwoh // 17th International Conference on Pattern Recognition (ICPR'04). – Cambridge UK: 2004. – vol. 3. – pp. 494-497.

52. Гладкова А.В. Выбор настраиваемых параметров алгоритма опорных векторов с квадратичной функцией потерь / А. В. Гладкова, А. А. Попов // Обработка информации и математическое моделирование: материалы Рос. науч.-техн. конф. [Новосибирск, 24–25 апр. 2015 г.]. – Новосибирск: СибГУТИ, 2015. – С. 62–66.

53. Huber P.J. Robust Statistics / P.J. Huber, New York: Wiley, 1981. – 304 p.

54. Robust Statistics: The Approach Based on Influence Functions / F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel. New York: Wiley Series in Probability and Statistics, – 1986. – 536 p – ISBN: 978-0-471-73577-9.

55. Айвазян С.А. и др. Прикладная статистика: Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин; под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1985. – 487с.

56. Айвазян С.А. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика. – 1989. – 606 с.

57. A. Smola, Learning with Kernels, Ph. D. Thesis. Published by: GMD-First, Birlinghoven, – 1999.

58. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов / А.А. Большаков, Р.Н. Каримов – М.: Горячая линия-Телеком, 2007. – 522 с.

59. Дрейпер Н. Прикладной регрессионный анализ, 3-е изд. / Н. Дрейпер, Г. Смит. – М.: Издательский дом «Вильямс», 2007. – 912 с.

60. Хьюбер Дж.П. Робастность в статистике. – М.: Мир, 1984. – 304 с.

61. Попов А.А. Математические методы планирования эксперимента: учеб.-метод. пособие / А.А. Попов, Д.В. Лисицин. – Новосибирск: Изд-во НГТУ, 2000. – 28 с.

62. Suykens J.A.K. Sparse approximation using least-squares support vector machines / J.A.K. Suykens, L Lukas,J. Vandewalle // Neurocomputing, 2000, – vol. 48, – pp. 85–105.

63. David H. A. Early sample measures of variability / H. A. David // StatisticalScience, – 1998. – vol. 13, no. 4, – pp. 368 – 377.

64. Rousseeuw P. J. Robust Regression and Outlier Detection / P. J. Rousseeuw, A. M. Leroy. Wiley-Interscience, N. Y. (Series in Applied Probability and Statistics): 1987. – 329 p. – ISBN: 0-471-85233-3.

65. Jiyan H. Robust location algorithm based on weighted least squares support vector machine (WLS-SVM) for nonline-of-sight environments/ H. Jiyan, G. Guan,W. Qun // International Journal of the Physical Sciences, – 23 Oct., 2011. – Vol. 6, no. 25, – pp. 5897 – 5905.

66. Бобоев Ш.А. Построение робастных регрессионных моделей по методу LS–SVM с использованием функции потерь Эндрюса / Ш.А. Бобоев // Вестник филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. Серия естественных наук. – 2024. – Т. 1, № 3 (41). – С. 30–36.

67. Бобоев Ш.А. Использование биквадратной функции потерь Тьюки для построения робастных регрессионных моделей по методу LS–SVM / Ш.А. Бобоев // Материалы международной научно-практической конференции «XIV Ломоносовские чтения», «Роль филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе в развитии науки и образования» (22–23 ноября 2024 г.). – Душанбе. – 2024. – Часть II. Естественные науки. – С. 12–19.

68. Popov A.A. The use of robust criteria for the choice of regression model by LS-VM method / A.A. Popov, Sh.A. Boboev // В сборнике: Труды XIII международной научно-технической конференции Актуальные проблемы электронного приборостроения (АПЭП-2016). В 12 томах. – Т. 1, Ч. 2. – 2016. – С. 313–316.

69. Попов А.А. Использование робастных критериев для выбора регрессионной модели по методу LS–SVM / А.А. Попов, Ш.А. Бобоев // В сборнике: Труды XIII международной научно-технической конференции Актуальные проблемы электронного приборостроения (АПЭП-2016). В 12 томах. – Т. 8. – 2016. – С. 145–148.

70. Popov A.A. The construction of the robust regression models with the LS–SVM method using a nonquadratic loss function / A.A. Popov, Sh.A. Boboev // В сборнике: Proceedings of IFOST-2016 11th International Forum on Strategic Technology IFOST-2016. – 2016. – С. 394–396.

71. Попов А.А. Построение робастных регрессионных моделей по методу LS–SVM с использованием функций потерь Хьюбера и взвешивания наблюдений / А.А. Попов, Ш.А. Бобоев // В сборнике: Обработка информации и математическое моделирование материалы Российской научно-технической конференции. –2016. – С. 118–124.

72. Холкин В.В. Построение робастных решений при восстановлении зависимостей по методу опорных векторов с квадратичной функцией потерь / В.В. Холкин, А.А. Попов // Сборник научных трудов конференции «Наука. Технологии. Инновации», в 9 частях. – 2016. – Часть 2. – С. 198–200.

73. Попов А.А. Построение робастных и разреженных решений по методу опорных векторов с функцией потерь Йохана Сайкинса / А.А. Попов, В.В. Холкин // Материалы Российской научно-технической конференции «Обработка информации и математическое моделирование». – 2018. – С. 117-122.

74. Suykens J.A.K. Sparse Least Squares Support Vector Machine Classifiers / J.A.K. Suykens, L. Lukas, J. Vandewalle. // In: ESANN’2000 European Symposium on Artificial Neural Networks. – 2000. – pp. 37-42.

75. Suykens J.A.K. Sparse Approximation Using Least Squares Support Vector Machines / J.A.K. Suykens, L. Lukas, J. Vandewalle // In: IEEE International Symposium on Circuits and Systems ISCAS’2000. – 2000. – Vol. 2. – pp. 757-760.

76. Перельман И. И. Методология выбора структуры модели при идентификации объектов управления / И. И. Перельман // Автомат. и телемеханика. – 1983. – № 11. – С. 5–29.

77. Романов В.Л. Выбор наилучшей линейной регрессии: сравнение формальных критериев / В. Л. Романов // Зав. лаборатория – 1990. – №1. – С. 90 – 95.

78. Себер Дж. Линейный регрессионный анализ / Дж. Себер. – М.: Мир, 1980. – 456 с.

79. Степашко В.С. Методы и критерии решения задач структурной идентификации / В.С. Степашко, Ю.Л. Кочерга // Автоматика. – 1985. – № 5. – С. 29–37.

80. Кочерга Ю.Л. J–оптимальная редукция структуры модели в схеме Гаусса–Маркова / Ю.Л. Кочерга // Автоматика. – 1988. – № 4. – С. 34–38.

81. Сарычев А.П. Усредненный критерий регулярности метода группового учета аргументов в задаче поиска наилучшей регрессии / А.П. Сарычев // Автоматика. – 1990. – № 5. – С. 28–33.

82. Степашко В.С. Асимптотические свойства внешних критериев выбора моделей / В.С. Степашко // Автоматика. – 1988. – № 6. – С. 75–82.

83. Степашко В.С. Потенциальная помехоустойчивость моделирования по комбинаторному алгоритму МГУА без использования информации о помехах // Автоматика. 1983. № 3. С. 18–28.

84. Степашко В.С. Селективные свойства критерия непротиворечивости моделей / В.С. Степашко // Автоматика. – 1986. – № 2. – С. 40–49.

85. Попов А.А. Использование повторных выборок в критериях селекции моделей / А.А. Попов // Планирование эксперимента, идентификация, анализ и оптимизация многофакторных систем. Новосибирский электротехнический институт. Новосибирск, – 1990. – С. 82–88.

86. Лисицин Д.В., Попов А.А. Исследование критериев селекции многооткликовых регрессионных моделей / Д.В. Лисицин, А.А. Попов // Сборник научных трудов НГТУ. – 1996. – Вып. 2. – С. 19–28.

87. Лисицин Д.В., Попов А.А. Конструирование критериев селекции многомерных регрессионных моделей / Д.В. Лисицин, А.А. Попов // Сборник научных трудов НГТУ. – 1996. – Вып. 1. – С. 13–20.

88. Попов А.А. Планирование эксперимента в задачах разбиения выборки в МГУА / А.А. Попов // Сборник научных трудов НГТУ. – 1995. – Вып. 2. – С. 35–40.

89. Попов А.А. Разбиение выборки для внешних критериев селекции моделей с использованием методов планирования эксперимента / А.А. Попов // Заводская лаборатория. – 1997. – № 1. – С. 49–53.

90. Попов А.А. Получение тестовой выборки в методе LS–SVM с использованием оптимального планирования эксперимента / А.А. Попов, Ш.А. Бобоев // Научный вестник Новосибирского государственного технического университета. – 2016. – № 4 (65). – С. 80–99. DOI: 10.17212/1814-1196-2016-4-80-99.

91. Ванюкевич О.Н., Попов А.А. Критерии выбора модели при построении размытой регрессионной зависимости // Сборник научных трудов НГТУ. 2004. № 4 (38). С. 15–20.

92. Попов А.А. Оптимальное планирование эксперимента в задачах структурной и параметрической идентификации моделей многофакторных систем: монография. / А.А. Попов, Новосибирск: Изд-во НГТУ, 2013. – 296 с.

93. Попов А.А. Последовательные схемы построения оптимальных планов эксперимента / А.А. Попов // Сборник научных трудов НГТУ. – 1995. – Вып. 1. – С. 39–44.

94. Попов А.А. Последовательные схемы синтеза оптимальных планов эксперимента / А.А. Попов // Доклады академии наук высшей школы России. – 2008. – № 1 (10). – С. 45–55.

95. Юрачковский Ю.П. Применение канонической формы внешних критериев для исследования их свойств / Ю.П. Юрачковский, А.Н. Грошков // Автоматика. – 1979. – №3. – С. 85–89.

96. Федоров В.В. Активные регрессионные эксперименты / В.В. Федоров // Математические методы планирования эксперимента. – Новосибирск: Наука. – 1981. – С. 19 – 73.

97. Попов А.А. Методы планирования эксперимента в задачах синтеза моделей оптимальной сложности / А.А. Попов // Машинные методы планирования эксперимента и оптимизации многофакторных систем / Новосиб. электротехн. ин–т. – Новосибирск, 1987. – С. 54–58.

98. Лисицин Д.В. Выбор структуры для многомерной динамической системы / Д.В. Лисицин, А.А. Попов // Сб. науч. тр. НГТУ. – Новосибирск, 1997. – Вып. 1(6). – С. –33-40.

99. Лисицин Д.В. Исследование критериев селекции многомерных моделей при наличии разнотипных факторов / Д.В. Лисицин, А.А. Попов // Труды третьей межд. научн.–техн. конф. "Актуальные проблемы электронного приборостроения" АПЭП –96. – Новосибирск, 1996, т. 6, Ч. 1. – С. 54–58.

100. Лисицин Д.В. Структурная оптимизация многомерных регрессионных моделей / Д.В. Лисицин, А.А. Попов // Второй Сибирский Конгресс по Прикладной и Индустриальной Математике, Новосибирск, 1996.: Тез. докл. – Новосибирск, 1996. – С. 179.

101. Демиденко Е.З. Линейная и нелинейная регрессии / Е.З. Демиденко. – М.: Финансы и статистика, 1981. – 304 с.

102. Дрейпер Н. Прикладной регрессионный анализ: В 2-х кн. / Н. Дрейпер, Г. Смит. – М.: Финансы и статистика, 1986. – 366 с.

103. Ермаков С.М., Жиглявский А.А. Математическая теория оптимального эксперимента / С.М. Ермаков, А.А. Жиглявский. – М.: Наука, 1987. – 320 с.

104. Попов А.А. Конструирование линейных регрессионных моделей с разнотипными переменными / А.А. Попов. – Новосибирск: Изд-во НГТУ, 1999.

105. Федоров В.В. Теория оптимального эксперимента / В.В. Федоров. – М.: Наука, 1971. – 312 с.

106. Попов А.А. Получение разреженных решений методом LS–SVM через построение обучающей выборки / А.А. Попов, Ш.А. Бобоев // Вестник

Таджикского национального университета. Серия естественных наук. – 2017. – № 1-5. – С. 183–191.

107. Суходолов А.П. Настройка параметров ядерных функций в методе LS–SVM с использованием внешних критериев качества моделей / А.П. Суходолов, А.А. Попов, Ш.А. Бобоев // Доклады Академии наук высшей школы Российской Федерации. – 2017. – № 3 (36). – С. 88–104.

108. Попов А.А. Получение разреженных решений методом LS–SVM через построение выборки с помощью методов оптимального планирования и внешних критериев качества моделей / А.А. Попов, Ш.А. Бобоев // Вестник Иркутского государственного технического университета. – 2018. – Т. 22. – № 1 (132). – С. 100–117.

109. Попов А. А. Построение разреженных решений при использовании алгоритма опорных векторов в задаче восстановления зависимости / А. А. Попов, А. С. Саутин // Научный вестник НГТУ. – 2010. – № 2(39). – С. 31–42.

110. Popov A.A. Comparison of sparse solutions obtained by splitting the sample into parts based on external quality criteria of models in the LS–SVM method / A.A. Popov, Sh.A. Boboev // В сборнике: Труды XIV международной научно-технической конференции Актуальные проблемы электронного приборостроения (АПЭП-2018). В 8 томах. – Т. 1, Ч. 4. – 2018. – С. 236–240.

111. Попов А.А. Сравнение разреженных решений, получаемых разбиением выборки на части на основе внешних критериев качества моделей в методе LS–SVM / А.А. Попов, Ш.А. Бобоев // В сборнике: Обработка информации и математическое моделирование Материалы Российской научно-технической конференции. – 2018. – С. 102–109.

112. Попов А.А. Использование методов оптимального планирования эксперимента для разбиения выборки на части и настройка параметров ядерных функций в методе LS–SVM на основе внешних критериев качества моделей / А.А. Попов, Ш.А. Бобоев // В сборнике: Обработка информации и математическое моделирование Материалы Российской научно-технической конференции. – 2017. – С. 135–142.

113. Бобоев Ш.А. Способы построения разреженных регрессионных моделей по методу LS–SVM / Ш.А. Бобоев // Материалы научно-практической конференции «XI Ломоносовские чтения», посвященной 30-летию Государственной независимости Республики Таджикистан (29-30 апреля 2021 года). – Душанбе. – 2021. – С. 9–10.

114. Sigrist M. Air Monitoring by Spectroscopic Techniques / M. Sigrist. N.Y.: Wiley, 1994. – 560 p.

115. SemiPar: Semiparametic Regression: [Электронный ресурс]. URL: https://cran.r-project.org/web/packages/SemiPar/index.html.

116. Silverman B.W. Some aspects of the spline smoothing approach to nonparametric regression curve fitting / B.W. Silverman // Journal of the Royal Statistical Society. – 1985. – vol. 47, no. 1. – pp. 1–52.

117. Mcycle R dataset - Simulated Motorcycle Accident Data: [Электронный ресурс]. URL: https://www.kaggle.com/datasets/nirmalsankalana/mcycle?resource=download.

118. Smola A. Regression estimation with support vector learning machines: master's thesis / A. Smola – Technische Universität München. – München, 1996. – 78 p.

119. Бобоев Ш.А. Применение метода LS–SVM для анализа выборок LIDAR и Motorcycle / Ш.А. Бобоев // Сборник научных трудов НГТУ. – 2019. – № 1 (94). – С. 85–99. DOI: 10.17212/2307-6879-2019-1-85-99.

120. Новаковский М.С. Лабораторные работы по химии комплексных соединений / М.С. Новаковский – 2-е изд., перераб. и доп. – Харьков: Издательство Харьковского университета, 1972. – 232 с.

121. Мудинов Х.Г. Комплексообразование серебра с 1,2,4-триазолом в водно-спиртовых растворах / Х.Г. Мудинов, С.М. Сафармамадов // Вестник Таджикского национального университета. – 2015. – № 1/6 (91). – С. 98–102.

122. Сангов М.М. Комплексообразование Ag (I) с тиокарбогидразидом в интервале 288–328К / М.М. Сангов, С.М. Сафармамадов // Вестник Таджикского национального университета. – 2015. – № 1/6 (91). – С. 74–79.

123. Сангов М.М. Комплексообразование Ag (I) с тиокарбогидразидом в водно-спиртовых растворах / М.М. Сангов, С.М. Сафармамадов // Вестник Таджикского национального университета. – 2016. – № 1/3 (200). – С. 179–183.

124. Капустин Е.И. Решение некоторых классов математических задач в программе Excel [Электронный ресурс]. – URL: http://old.exponenta.ru/educat/systemat/kapustin/014.asp (дата обращения: 06.06.2019).

125. Бобоев Ш.А. Использование метода LS–SVM для изучения процесса комплексообразования переходных металлов с производными тиомочевины в водных и водно-органических растворах / Ш.А. Бобоев // Сборник научных трудов НГТУ. – 2019. – № 1 (94). – С. 71–84. DOI: 10.17212/2307-6879-2019-1-71-84.

126. Попов А.А. Получение разреженных решений с использованием D–оптимального разбиения исходной выборки на обучающую и тестовую части и критерия регулярности / А.А. Попов, Ш.А. Бобоев // Вестник кибернетики. – 2018. – № 3 (31). – С. 162–168.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи, опубликованные в научных журналах, рекомендованных ВАК при Минобрнауки РФ и ВАК при Президенте РТ

1. **Бобоев Ш.А.** Получение тестовой выборки в методе LS–SVM с использованием оптимального планирования эксперимента / А.А. Попов, Ш.А. Бобоев // Научный вестник Новосибирского государственного технического университета. – 2016. – № 4 (65). – С. 80-99.
2. **Бобоев Ш.А.** Настройка параметров ядерных функций в методе LS–SVM с использованием внешних критериев качества моделей / А.П. Суходолов, А.А. Попов, Ш.А. Бобоев // Доклады Академии наук высшей школы Российской Федерации. – 2017. – № 3 (36). – С. 88-104.
3. **Бобоев Ш.А.** Получение разреженных решений методом LS–SVM через построение обучающей выборки / А.А. Попов, Ш.А. Бобоев // Вестник Таджикского национального университета. Серия естественных наук. – 2017. – № 1-5. – С. 183-191.
4. **Бобоев Ш.А.** Получение разреженных решений методом LS–SVM через построение выборки с помощью методов оптимального планирования и внешних

критериев качества моделей / А.А. Попов, Ш.А. Бобоев // Вестник Иркутского государственного технического университета. – 2018. – Т. 22. – № 1 (132). – С. 100-117.

5. **Бобоев Ш.А.** Получение разреженных решений с использованием D–оптимального разбиения исходной выборки на обучающую и тестовую части и критерия регулярности / А.А. Попов, Ш.А. Бобоев // Вестник кибернетики. – 2018. – № 3 (31). – С. 162–168.

6. **Бобоев Ш.А.** Критерии подбора метапараметров алгоритма LS–SVM / Ш.А. Бобоев // Вестник филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. Серия естественных наук. – 2023. – Т. 1. – № 2 (31). – С. 5-15.

7. **Бобоев Ш.А.** Построение робастных регрессионных моделей по методу LS–SVM с использованием функции потерь Эндрюса / Ш.А. Бобоев // Вестник филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. Серия естественных наук. – 2024. – Т. 1. – № 3 (41). – С. 30–36.

Публикации в изданиях, индексируемых Web of Science и Scopus

8. **Boboev Sh.A.** The construction of the robust regression models with the LS–SVM method using a nonquadratic loss function / A.A. Popov, Sh.A. Boboev // В сборнике: Proceedings of IFOST-2016 11th International Forum on Strategic Technology IFOST-2016. – 2016. – С. 394-396.

9. **Boboev Sh.A.** The use of robust criteria for the choice of regression model by LS-VM method / A.A. Popov, Sh.A. Boboev // В сборнике: Труды XIV международной научно-технической конференции Актуальные проблемы электронного приборостроения Proceedings: in 12 volumes. – 2016. – Vol. 1. P. 2. – Pp. 313-316.

10. **Boboev Sh.A.** Comparison of sparse solutions obtained by splitting the sample into parts based on external quality criteria of models in the LS–SVM method / A.A. Popov, Sh.A. Boboev // В сборнике: Труды XIII международной научно-технической конференции Актуальные проблемы электронного приборостроения Proceedings: in 8 volumes. – 2018. – Vol. 1. P. 4. – Pp. 236-240.

Публикации в других изданиях

11. **Бобоев Ш.А.** Построение регрессионных зависимостей с использованием квадратичной функции потерь в методе опорных векторов / А.А. Попов, Ш.А. Бобоев // Сборник научных трудов Новосибирского государственного технического университета. – 2015. – № 3 (81). – С. 69-78.

12. **Бобоев Ш.А.** Построение робастных регрессионных моделей по методу LS–SVM с использованием функций потерь Хьюбера и взвешивания наблюдений – / А.А. Попов, Ш.А. Бобоев // В сборнике: Обработка информации и

математическое моделирование материалы Российской научно-технической конференции. – 2016. – С. 118-124.

13. **Бобоев Ш.А.** Использование робастных критериев для выбора регрессионной модели по методу LS–SVM / А.А. Попов, Ш.А. Бобоев // В сборнике Актуальные проблемы электронного приборостроения (АПЭП – 2016), труды XIII международной научно-технической конференции: в 12 томах. – Новосибирск. – 2016. – Том 8. – С. 145-148.

14. **Бобоев Ш.А.** Использование методов оптимального планирования эксперимента для разбиения выборки на части и настройка параметров ядерных функций в методе LS–SVM на основе внешних критериев качества моделей / А.А. Попов, Ш.А. Бобоев // В сборнике: Обработка информации и математическое моделирование Материалы Российской научно-технической конференции. – 2017. – С. 135-142.

15. **Бобоев Ш.А.** Сравнение разреженных решений, получаемых разбиением выборки на части на основе внешних критериев качества моделей в методе LS–SVM / А.А. Попов, Ш.А. Бобоев // В сборнике Обработка информации и математическое моделирование, материалы Российской научно-технической конференции. – Новосибирск. – 2018. – С. 102-109.

16. **Бобоев Ш.А.** Использование метода LS–SVM для изучения процесса комплексообразования переходных металлов с производными теомочевины в водных и водно-органических растворах / Ш.А. Бобоев // Сборник научных трудов Новосибирского государственного технического университета. – 2019. – № 1 (94). – С. 71-84.

17. **Бобоев Ш.А.** Применение метода LS–SVM для анализа выборок LIDAR и MOTORCYCLE / Ш.А. Бобоев // Сборник научных трудов Новосибирского государственного технического университета. – 2019. – № 1 (94). – С. 85-99.

18. **Бобоев Ш.А.** Способы построения разреженных регрессионных моделей по методу LS–SVM / Ш.А. Бобоев // Материалы международной научно-практической конференции «XI Ломоносовские чтения», посвященной 30-летию Государственной независимости Республики Таджикистан (29-30 апреля 2021 года). – Душанбе. – 2021. – С. 9-10.

19. **Бобоев Ш.А.** Подбор коэффициента регуляризации в методе LS–SVM / Ш.А. Бобоев // Материалы международной научно-практической конференции «XIII Ломоносовские чтения», посвященной 115-летию академика Бободжона Гафурова (28-29 апреля 2023 года). – Душанбе. – 2023. – Ч. 3. – С. 27-32.

20. **Бобоев Ш.А.** Использование биквадратной функции потерь Тьюки для построения робастных регрессионных моделей по методу LS–SVM / Ш.А. Бобоев // Материалы международной научно-практической конференции «XIV Ломоносовские чтения», «Роль филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе в развитии науки и

образования» (22–23 ноября 2024 г.). – Душанбе. – 2024. – Часть II. Естественные науки. – С. 12-19.

Свидетельства о Государственной регистрации программы для ЭВМ

21. **Бобоев Ш.А.**, Попов А.А. Свидетельство №2018619675 о государственной регистрации программы на ЭВМ Программа «ПОЛУЧЕНИЕ РОБАСТНЫХ И РАЗРЕЖЕННЫХ РЕШЕНИЙ МЕТОДОМ LS–SVM "Robast_Sparse_LS–SVM"». Дата регистрации: 09.08.2018.

ПРИЛОЖЕНИЕ А

Таблицы значения MSE при использовании простого и адаптивного вариантов функции потерь Хьюбера для получения робастных решений

Таблица А.1. Значения *c* и MSE, при 5% уровне шума, при использовании обычной функции потерь Хьюбера в методе псевдонаблюдений (несимметричное засорение)

Уровень шума	**Уровень засорения**	**γ**	**c**	**σ**	**MSE**
5%	5%	1	9	0,398107	0,235250
		5	4,3	0,794328	0,065455
		10	2,4	1,000000	0,048811
		50	1,9	1	0,037158
		100	1,9	1	0,038941
5%	10%	1	8	0,398107	0,260005
		5	4	0,794328	0,062145
		10	2,4	1,000000	0,035534
		50	0,5	1	0,027667
		100	0,5	1	0,031760
5%	15%	1	14	0,398107	0,190563
		5	5	0,794328	0,037210
		10	2	1,000000	0,021405
		50	0,5	1	0,026432
		100	0,5	1	0,033889
5%	20%	1	4,3	0,398107	0,380655
		5	1,7	0,794328	0,135097
		10	1,1	1,000000	0,102778
		50	0,5	1	0,071496
		100	0,5	1	0,074302

Таблица А.2. Значения c и MSE, при 10% уровне шума, при использовании обычной функции потерь Хьюбера в методе псевдонаблюдений (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
10%	5%	1	7	0,398107	0,322583
		5	2,3	0,630957	0,124637
		10	1,5	0,794328	0,096524
		50	0,5	1	0,072369
		100	0,5	1	0,069525
10%	10%	1	4,5	0,398107	0,319364
		5	1,6	0,630957	0,074075
		10	1,1	0,794328	0,045200
		50	0,5	1	0,027963
		100	0,5	1	0,030195
10%	15%	1	1,8	0,398107	0,423099
		5	0,5	0,630957	0,164099
		10	0,6	0,794328	0,131422
		50	0,5	1	0,123752
		100	0,5	1	0,130369
10%	20%	1	2,4	0,398107	0,281896
		5	0,7	0,630957	0,041415
		10	0,5	0,794328	0,021457
		50	0,5	1	0,040908
		100	0,5	1	0,057931

Таблица А.3. Значения c и MSE, при 5% уровне шума, при использовании адаптивной функции потерь Хьюбера в методе псевдонаблюдений (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
5%	5%	1	9	0,398107	0,235250
		5	4,3	0,794328	0,062662
		10	2,4	1,000000	0,031040
		50	1,9	1	0,013700
		100	1,9	1	0,014085
5%	10%	1	8	0,398107	0,260005
		5	4	0,794328	0,061700
		10	2,4	1,000000	0,027324
		50	1,1	1	0,010004
		100	0,9	1	0,007811
5%	15%	1	14	0,398107	0,190563
		5	6	0,794328	0,035626
		10	4,5	1,000000	0,023762
		50	0,8	1	0,011677
		100	0,5	1	0,010888
5%	20%	1	7	0,398107	0,344667
		5	3,2	0,794328	0,094257
		10	1,9	1,000000	0,057238
		50	1,4	1	0,025469
		100	1,2	1	0,020957

Таблица А.4. Значения c и MSE, при 10% уровне шума, при использовании адаптивной функции потерь Хьюбера в методе псевдонаблюдений (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
10%	5%	1	9,3	0,398107	0,320095
		5	3,6	0,630957	0,114603
		10	3	0,794328	0,087465
		50	1,5	1	0,061764
		100	1,4	1	0,056479
10%	10%	1	5,6	0,398107	0,334688
		5	2,6	0,630957	0,070495
		10	1,7	0,794328	0,039896
		50	0,9	1	0,023365
		100	1	1	0,023508
10%	15%	1	3,5	0,398107	0,410236
		5	2	0,630957	0,085645
		10	1,4	0,794328	0,051979
		50	1,2	1	0,039874
		100	1,1	1	0,040020
10%	20%	1	5	0,398107	0,253204
		5	2,5	0,630957	0,033777
		10	1,7	0,794328	0,013886
		50	1,2	1	0,004655
		100	1,1	1	0,005535

Таблица А.5. Значения c и MSE, при 5% уровне шума, при использовании обычной функции потерь Хьюбера в методе взвешивания (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
5%	5%	1	9	0,398107	0,235250
		5	3,8	0,794328	0,058096
		10	2,3	1,000000	0,030925
		50	0,7	1	0,009446
		100	0,5	1	0,007946
5%	10%	1	8	0,398107	0,260005
		5	3,2	0,794328	0,058899
		10	2,2	1,000000	0,027134
		50	1,1	1	0,010001
		100	0,9	1	0,007808
5%	15%	1	13,5	0,398107	0,190563
		5	5,5	0,794328	0,035344
		10	3,5	1,000000	0,019037
		50	0,8	1	0,011682
		100	0,6	1	0,011450
5%	20%	1	6	0,398107	0,348757
		5	3,2	0,794328	0,094128
		10	1,9	1,000000	0,057317
		50	1,4	1	0,025451
		100	1,2	1	0,020899

Таблица А.6. Значения c и MSE, при 10% уровне шума, при использовании обычной функции потерь Хьюбера в методе взвешивания (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
10%	5%	1	9	0,398107	0,320106
		5	6	0,630957	0,117393
		10	3	0,794328	0,087465
		50	1,5	1	0,061498
		100	1,4	1	0,056302
10%	10%	1	6	0,398107	0,338705
		5	2,6	0,630957	0,070556
		10	1,7	0,794328	0,039901
		50	0,7	1	0,023172
		100	0,5	1	0,023006
10%	15%	1	3,5	0,398107	0,410229
		5	2	0,630957	0,085553
		10	1,4	0,794328	0,051962
		50	1,2	1	0,039839
		100	1,1	1	0,039929
10%	20%	1	3,5	0,398107	0,410229
		5	2	0,630957	0,085553
		10	1,4	0,794328	0,051962
		50	1,2	1	0,039839
		100	1,1	1	0,039929

Таблица А.7. Значения c и MSE, при 5% уровне шума, при использовании адаптивной функции потерь Хьюбера в методе взвешивания (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
5%	5%	1	9	0,398107	0,235250
		5	4,6	0,794328	0,048758
		10	4,2	1,000000	0,025408
		50	1,4	1	0,006107
		100	1,7	1	0,004548
5%	10%	1	8	0,398107	0,260005
		5	4,3	0,794328	0,056707
		10	3,4	1,000000	0,032218
		50	3,6	1	0,008674
		100	3,6	1	0,006816
5%	15%	1	13,5	0,398107	0,190563
		5	7,5	0,794328	0,036668
		10	9	1,000000	0,030389
		50	2,5	1	0,010750
		100	1,9	1	0,007287
5%	20%	1	6	0,398107	0,297404
		5	9	0,794328	0,079160
		10	4	1,000000	0,058763
		50	1,9	1	0,010521
		100	2	1	0,008374

Таблица А.8. Значения c и MSE, при 10% уровне шума, при использовании адаптивной функции потерь Хьюбера в методе взвешивания (несимметричное засорение)

Уровень шума	Уровень засорения	γ	c	σ	MSE
10%	5%	1	14	0,398107	0,326742
		5	5	0,630957	0,114983
		10	6	0,794328	0,086804
		50	1,8	1	0,061602
		100	2,3	1	0,027879
10%	10%	1	10	0,398107	0,411757
		5	2,5	0,630957	0,058059
		10	2,3	0,794328	0,027093
		50	2,2	1	0,014628
		100	2,2	1	0,014852
10%	15%	1	3,9	0,398107	0,269401
		5	3,5	0,630957	0,057514
		10	3,4	0,794328	0,034222
		50	1,7	1	0,040732
		100	1,8	1	0,040300
10%	20%	1	4,5	0,398107	0,308330
		5	3,5	0,630957	0,057514
		10	3,4	0,794328	0,034222
		50	1,7	1	0,040732
		100	1,8	1	0,040300

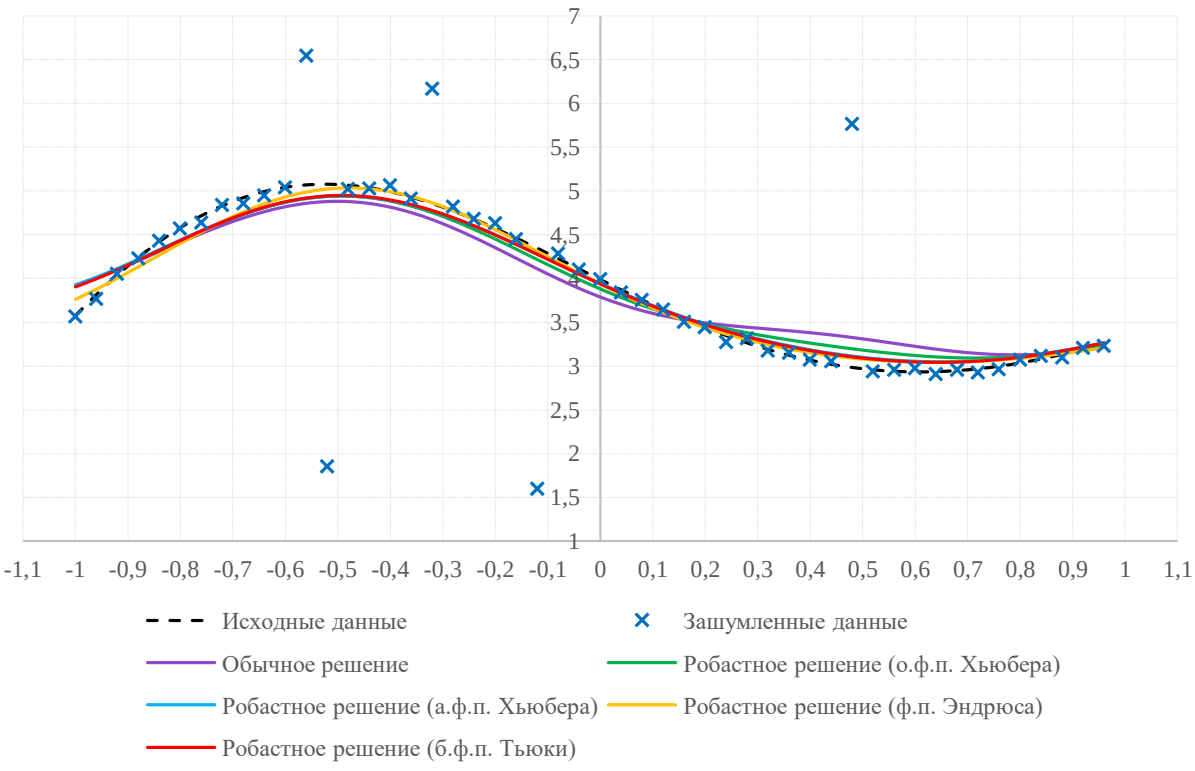


Рисунок А.1. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 1$

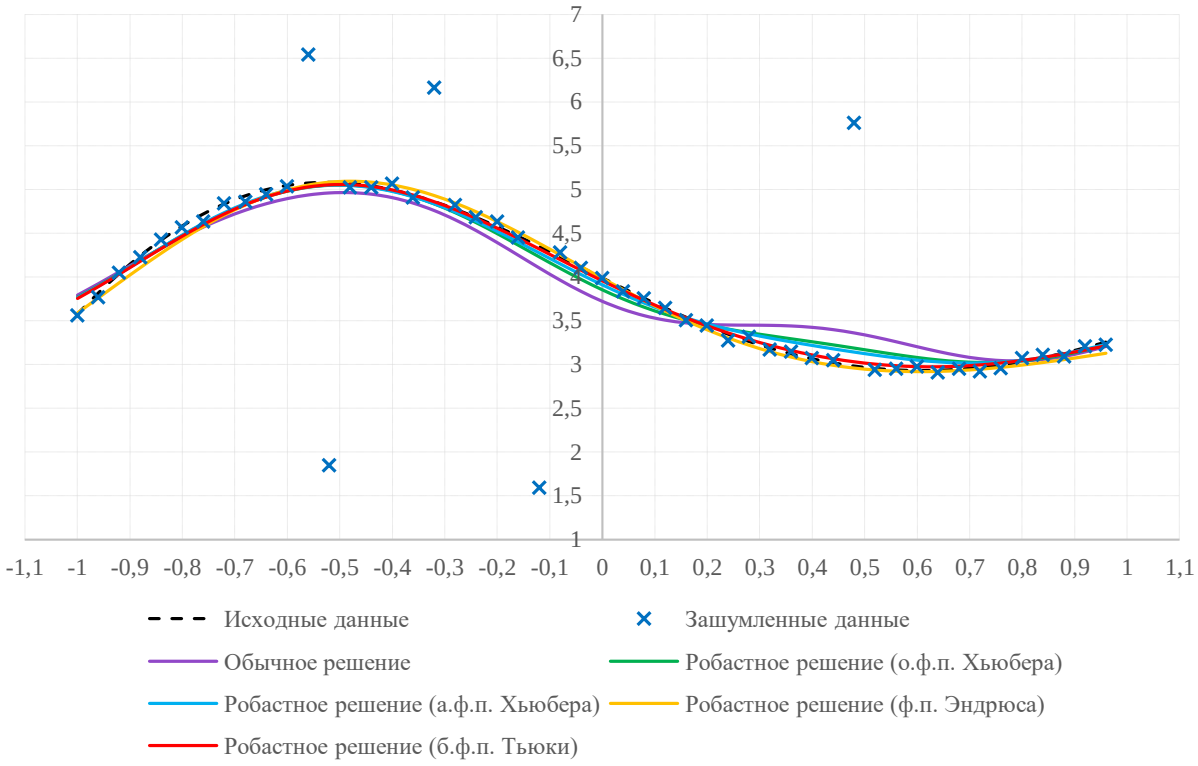


Рисунок А.2. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 5$

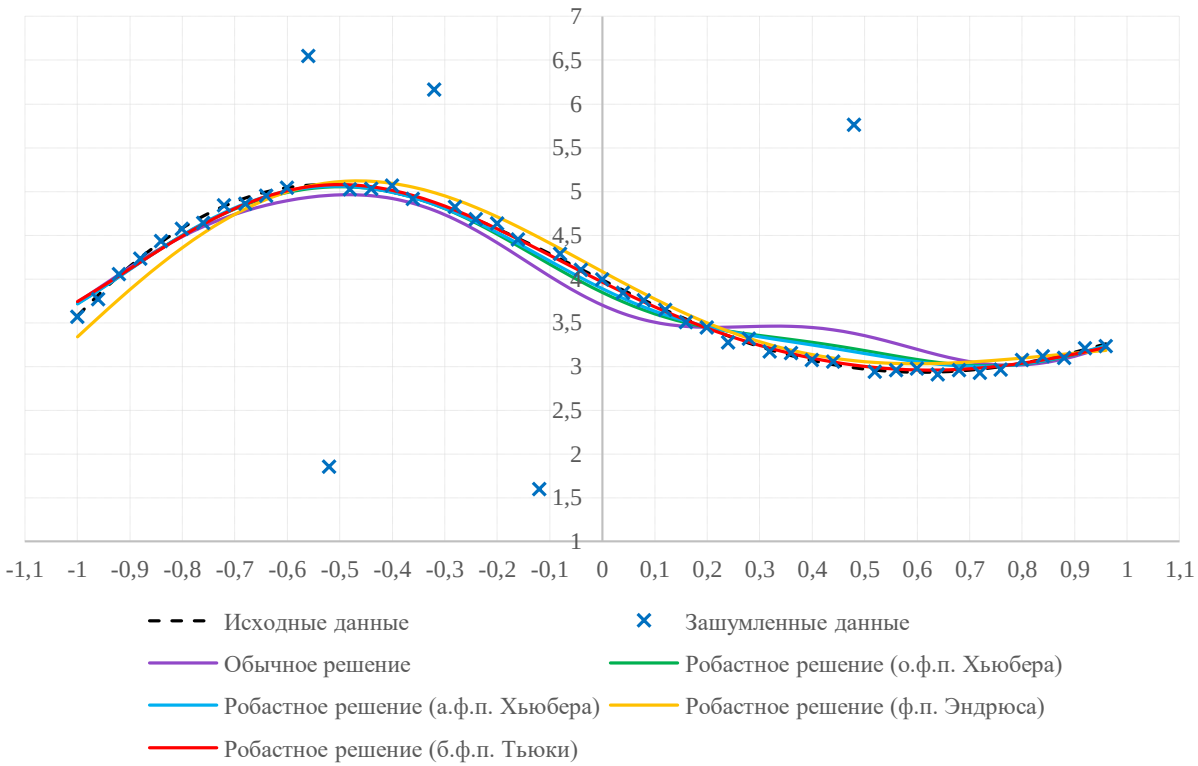


Рисунок А.3. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 10$

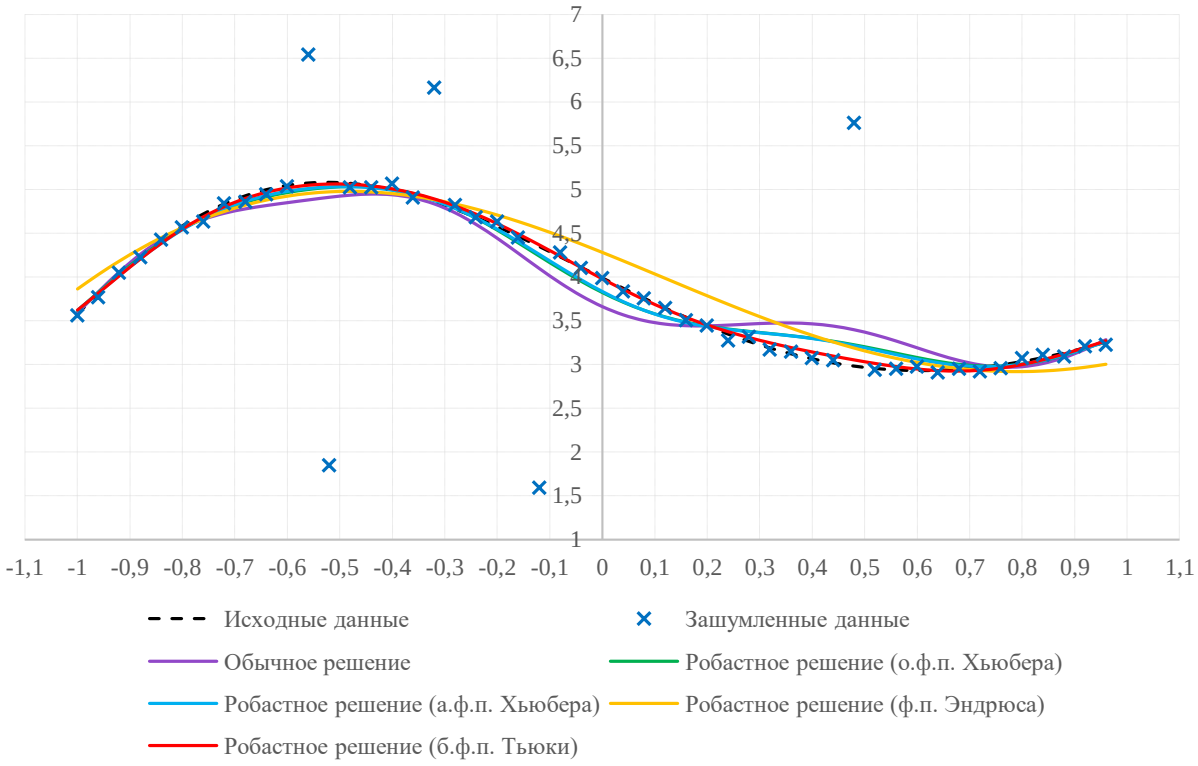


Рисунок А.4. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 50$

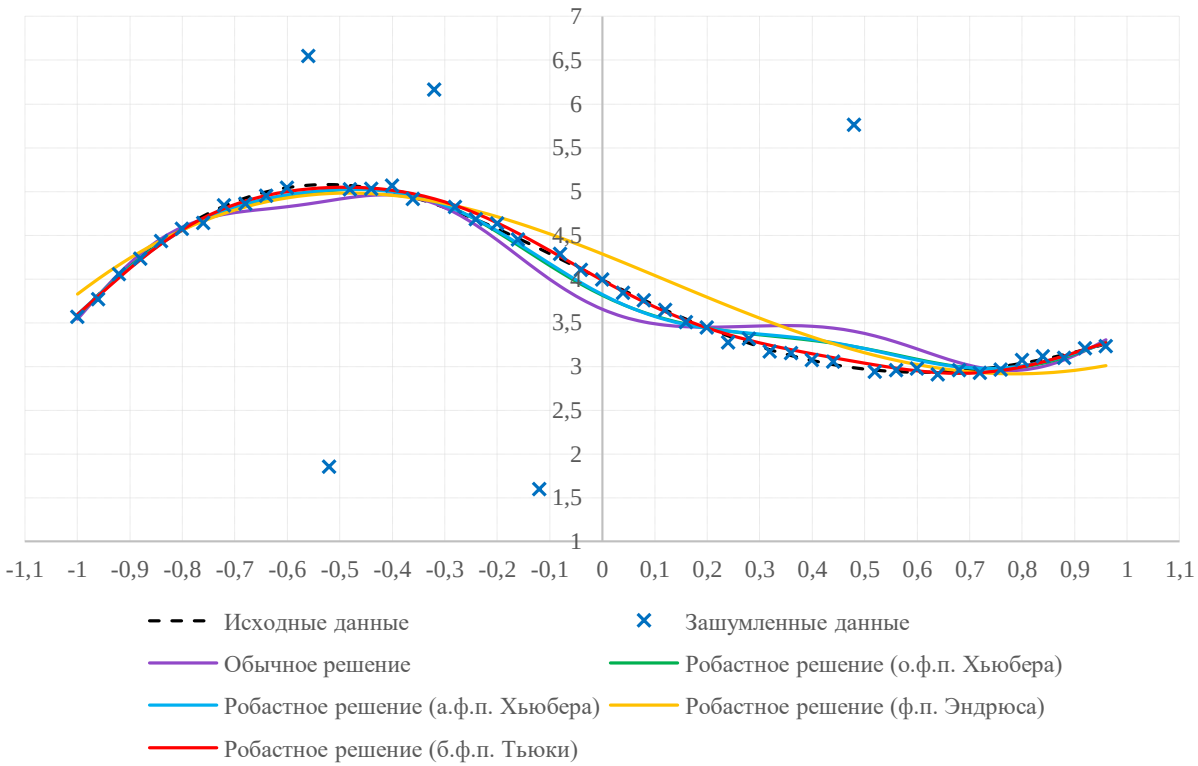


Рисунок А.5. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 100$

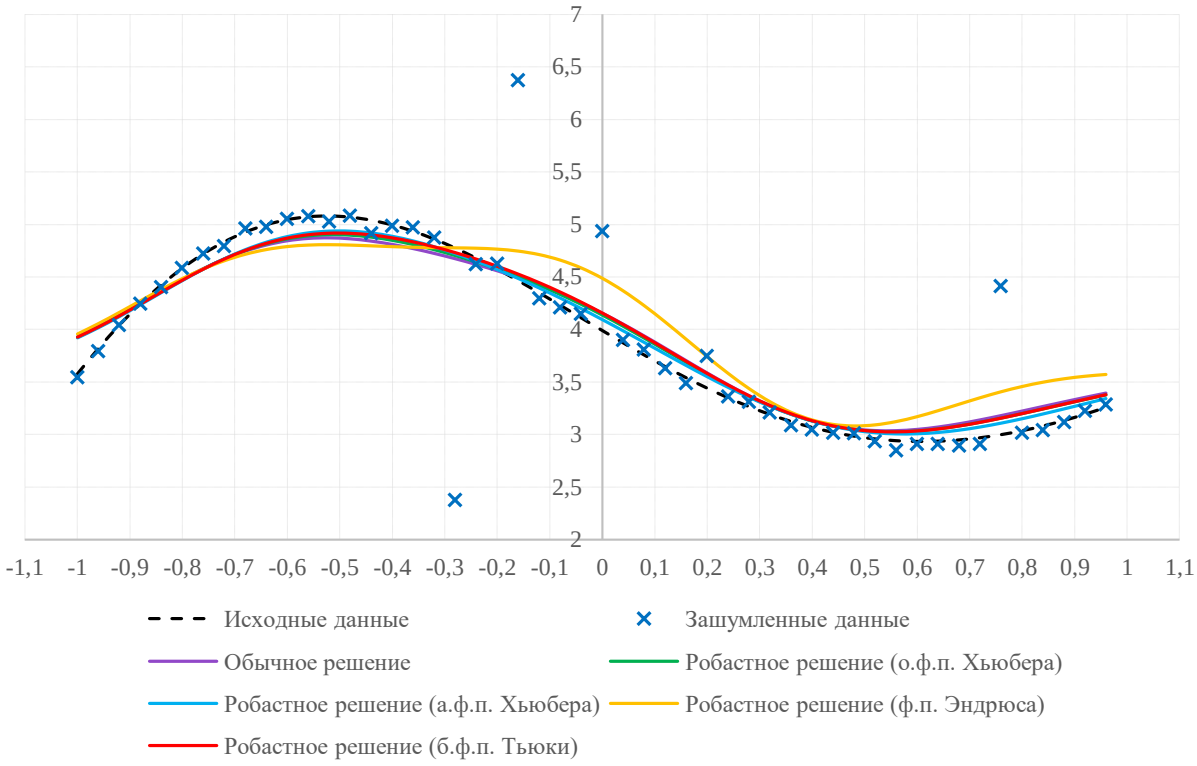


Рисунок А.6. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 1$

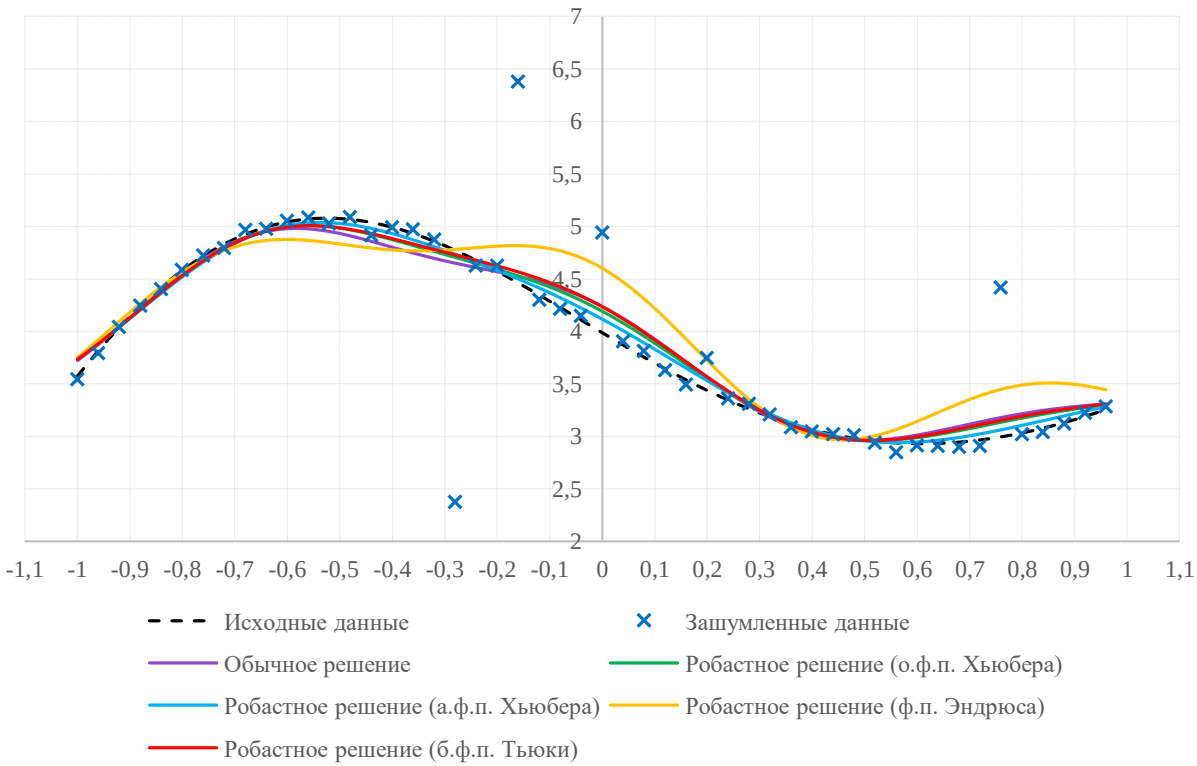


Рисунок А.7. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 5$

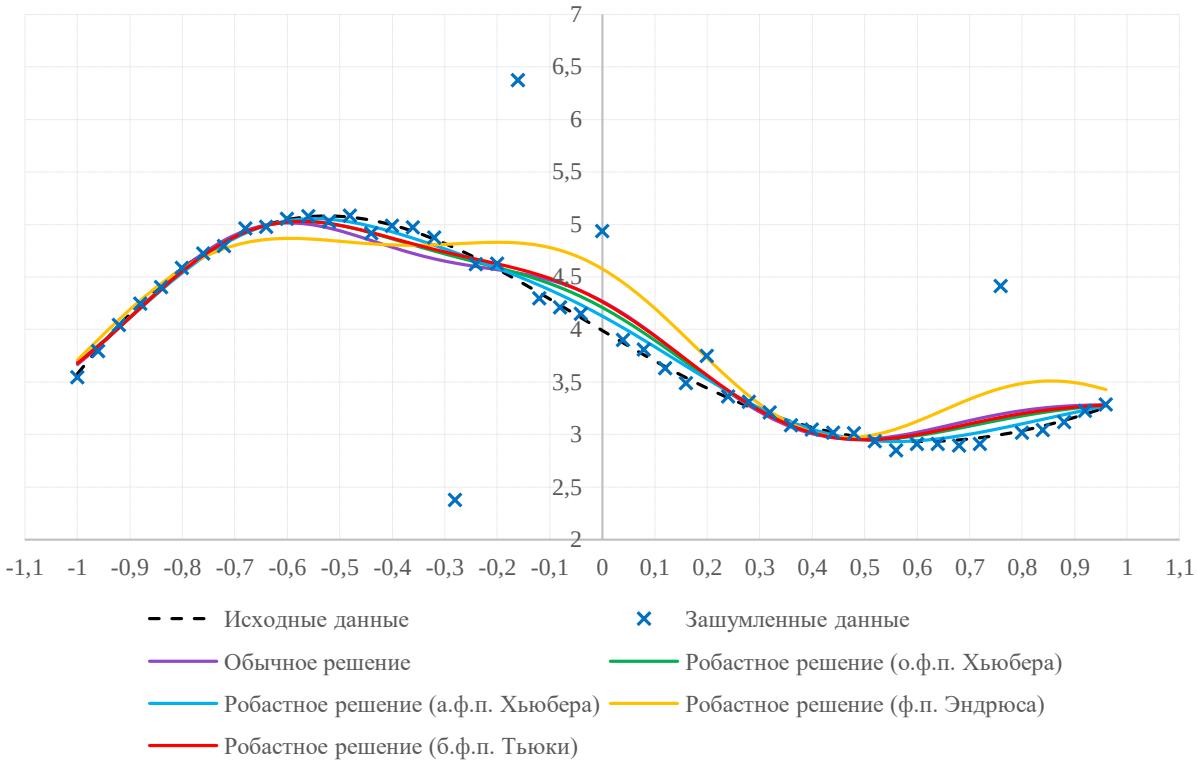


Рисунок А.8. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 10$

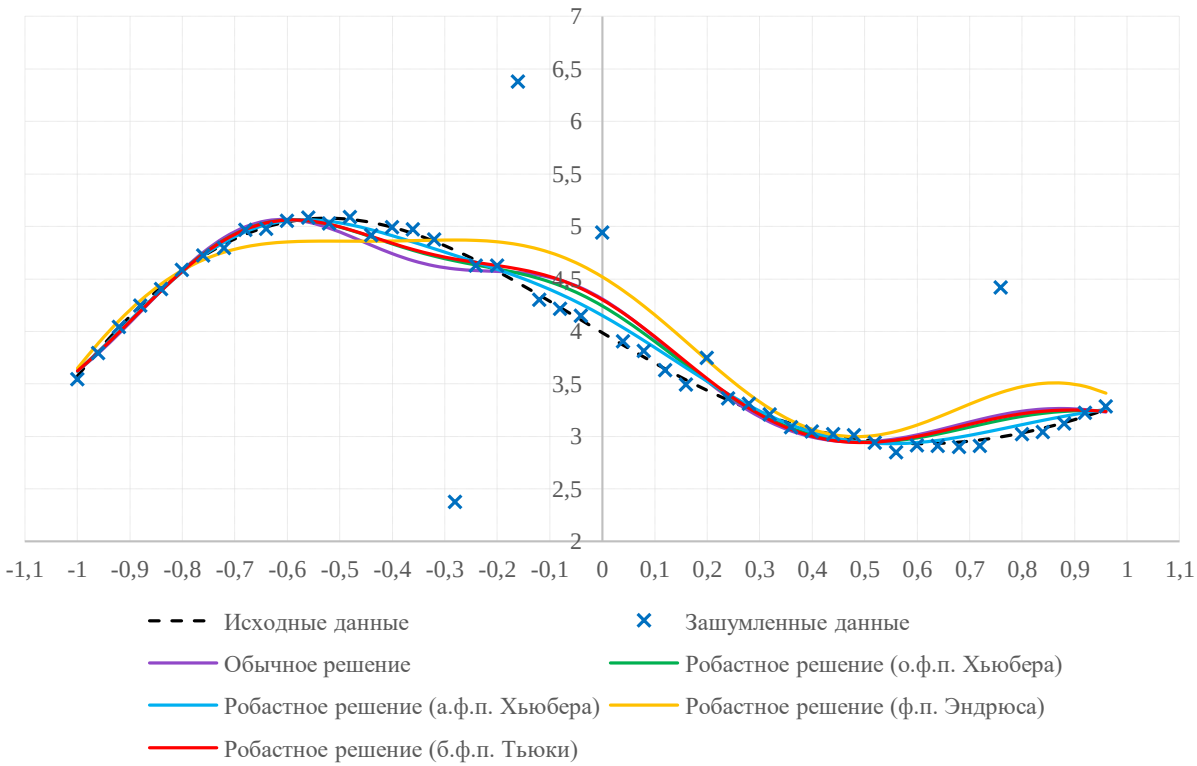


Рисунок А.9. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 50$

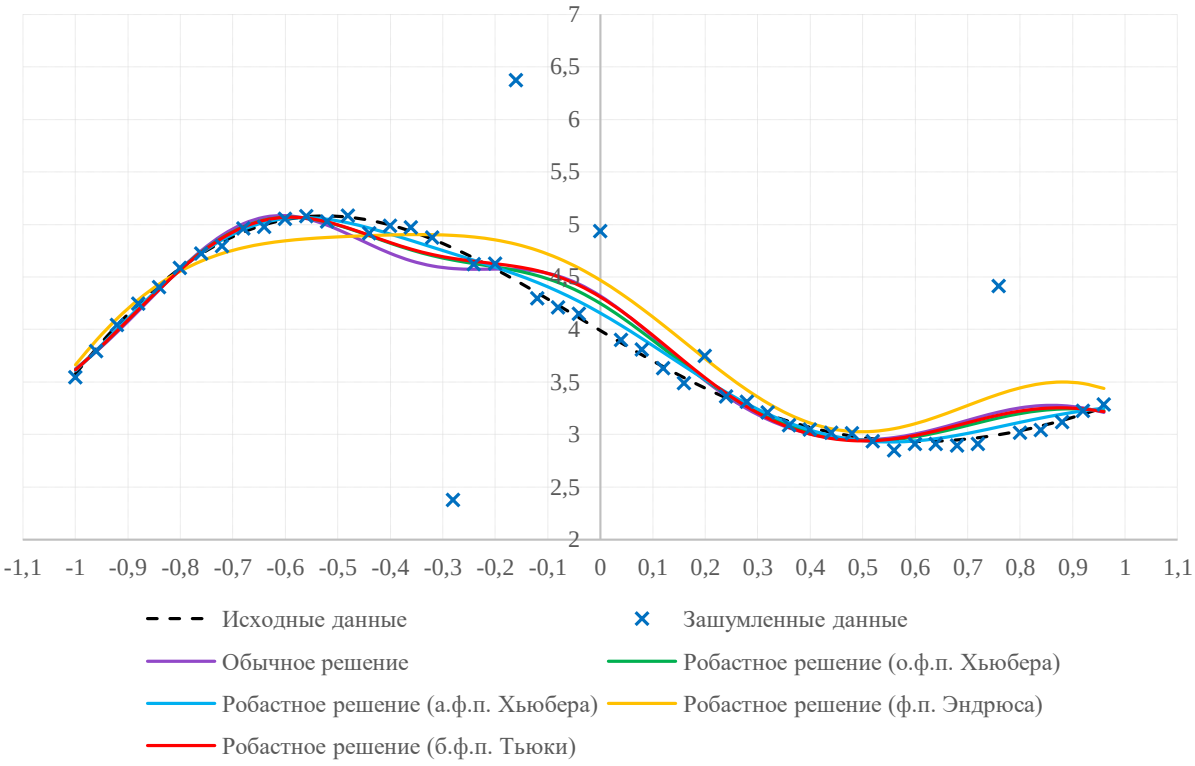


Рисунок А.10. Графики робастных моделей полученные методом псевдонаблюдений при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 100$

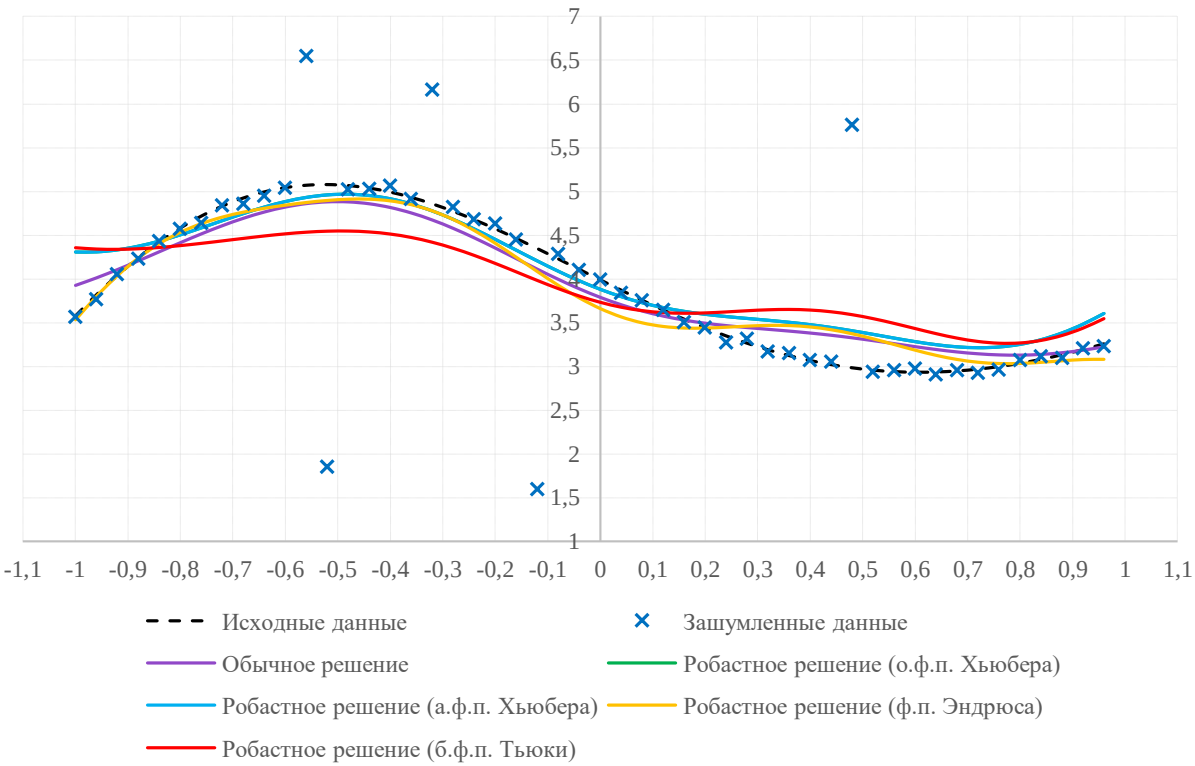


Рисунок А.11. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 1$

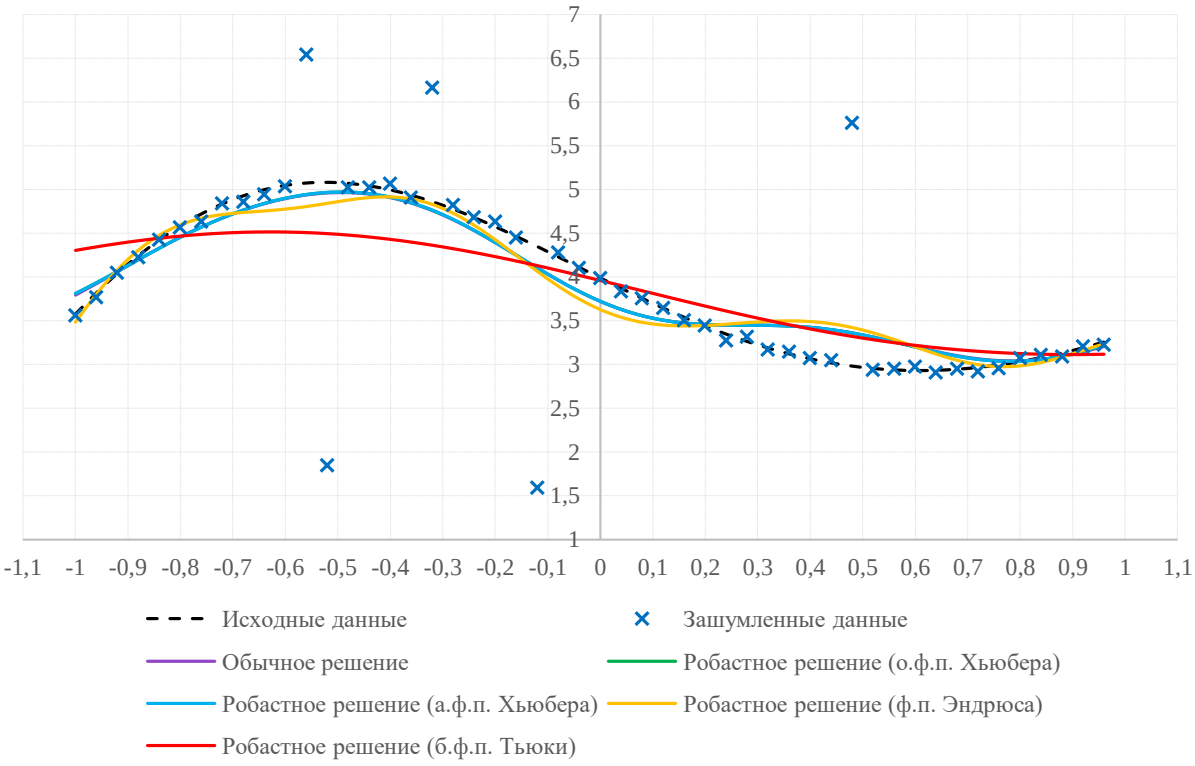


Рисунок А.12. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 5$

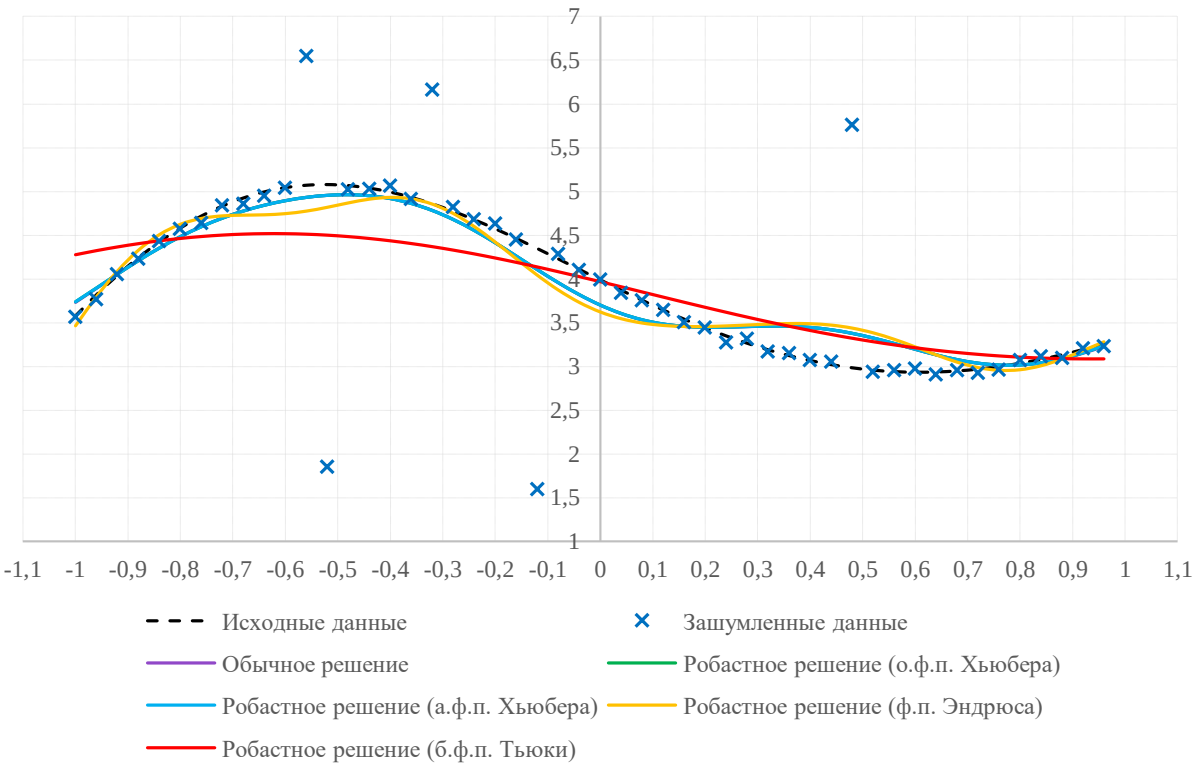


Рисунок А.13. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 10$

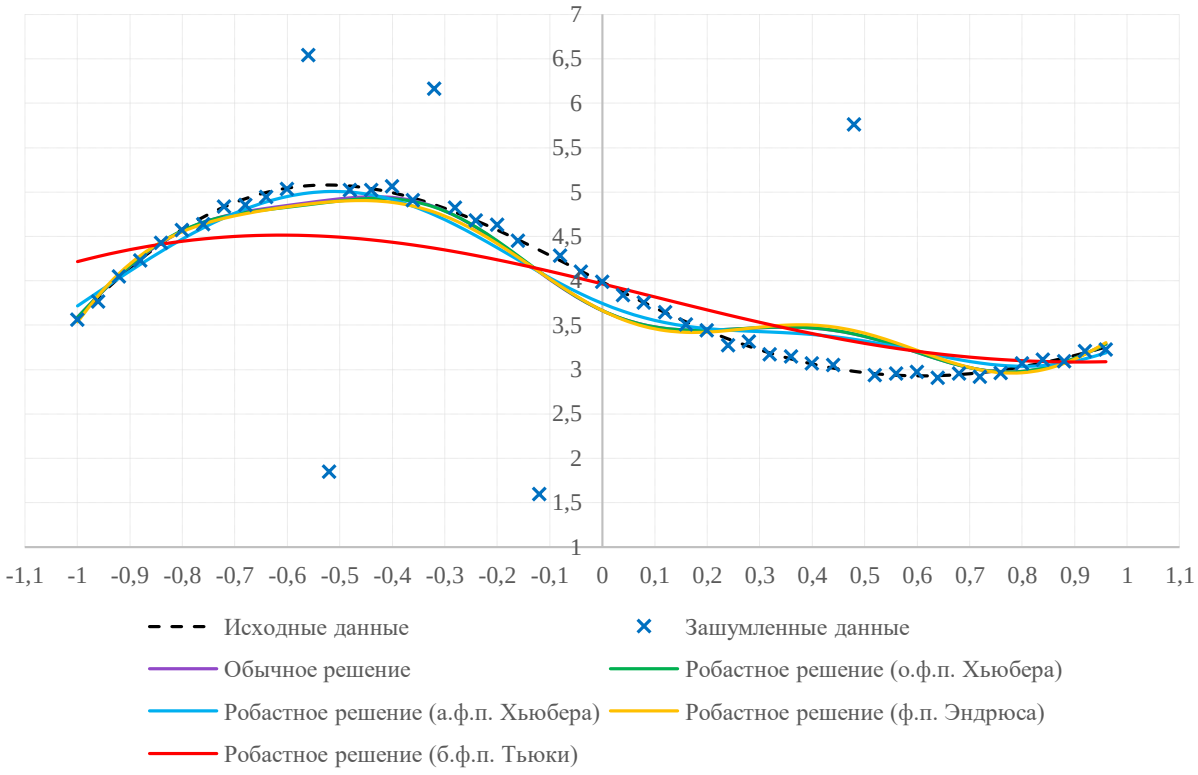


Рисунок А.14. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 5% уровне засорения и коэффициентом регуляризации $\gamma = 50$

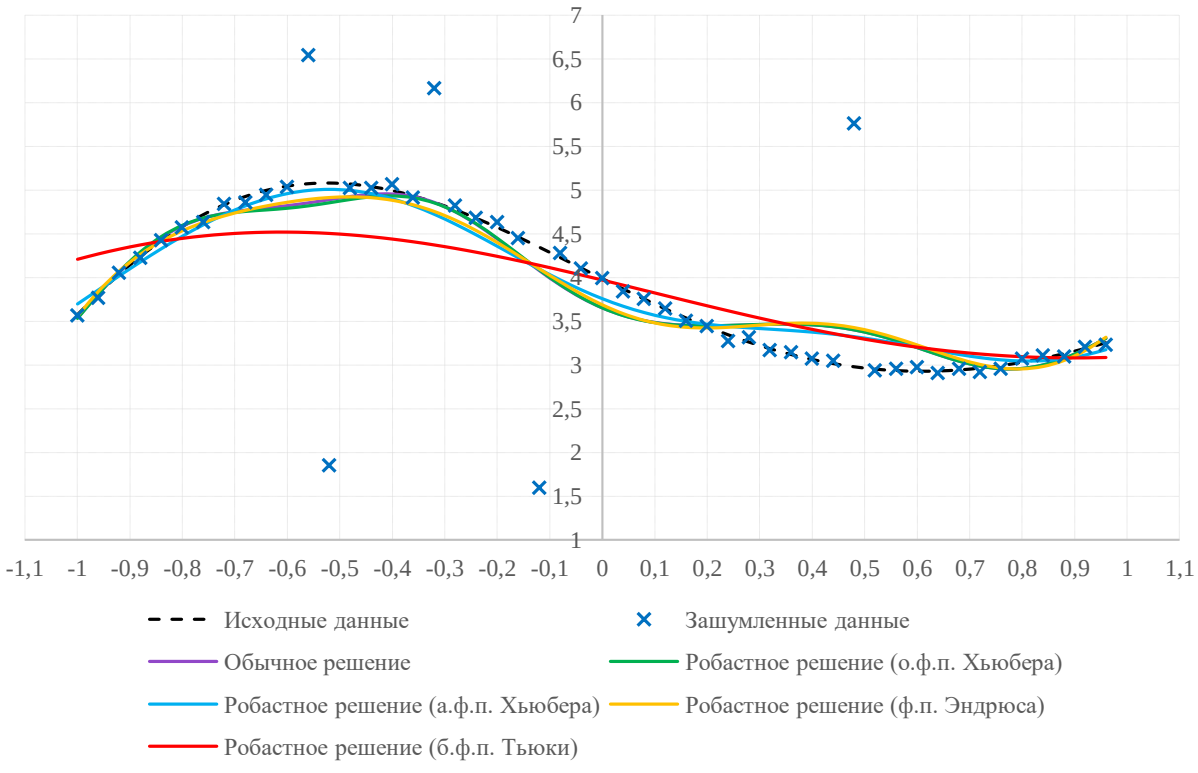


Рисунок А.15. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 100$

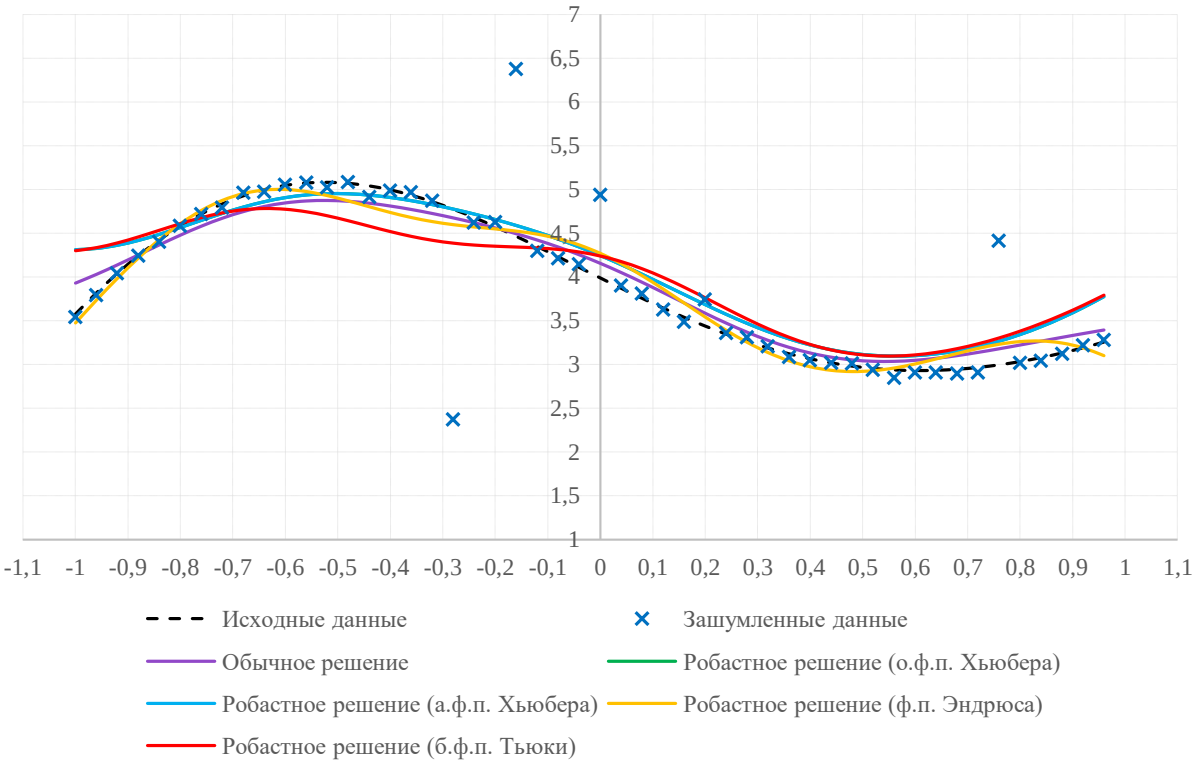


Рисунок А.16. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 1$

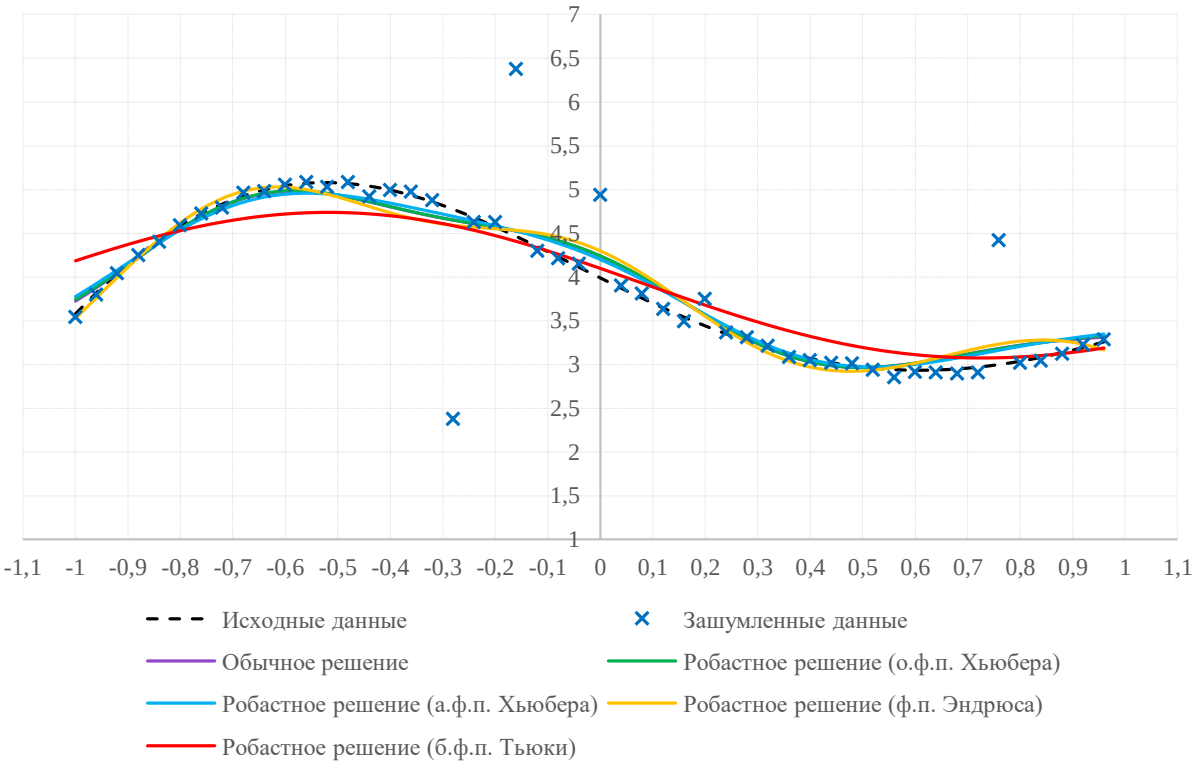


Рисунок А.17. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 5$

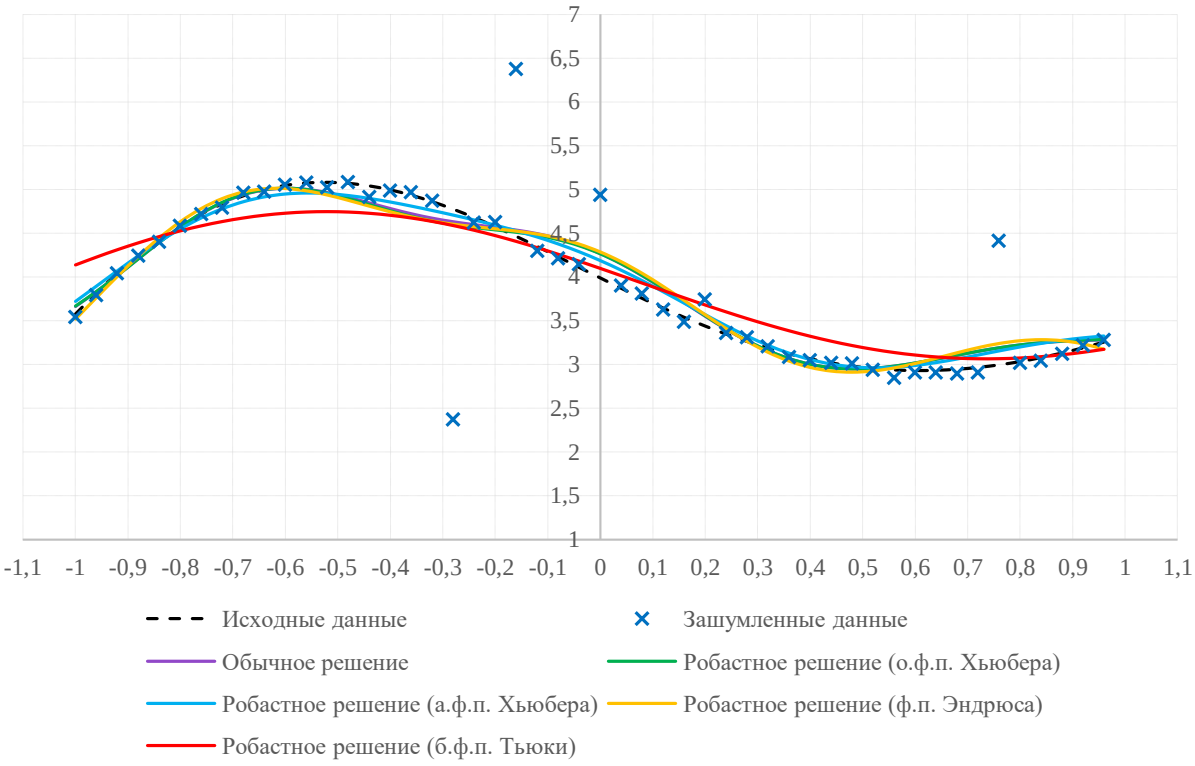


Рисунок А.18. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 10$

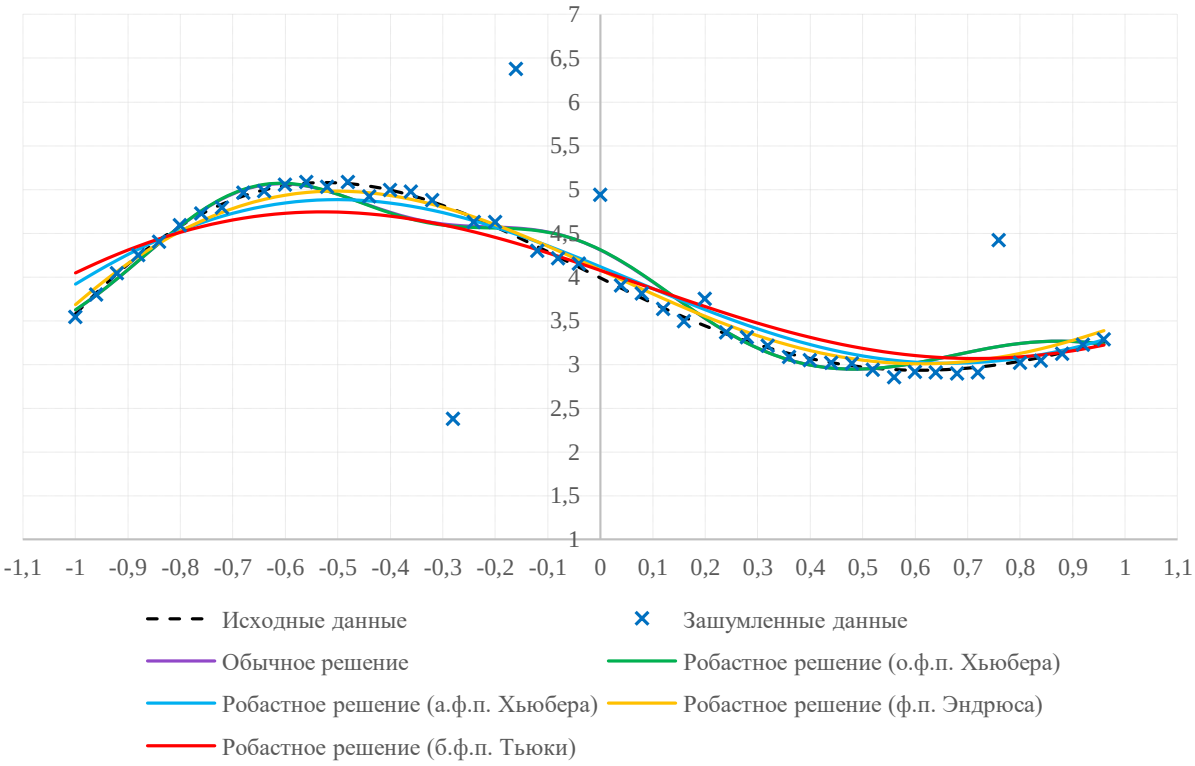


Рисунок А.19. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 50$

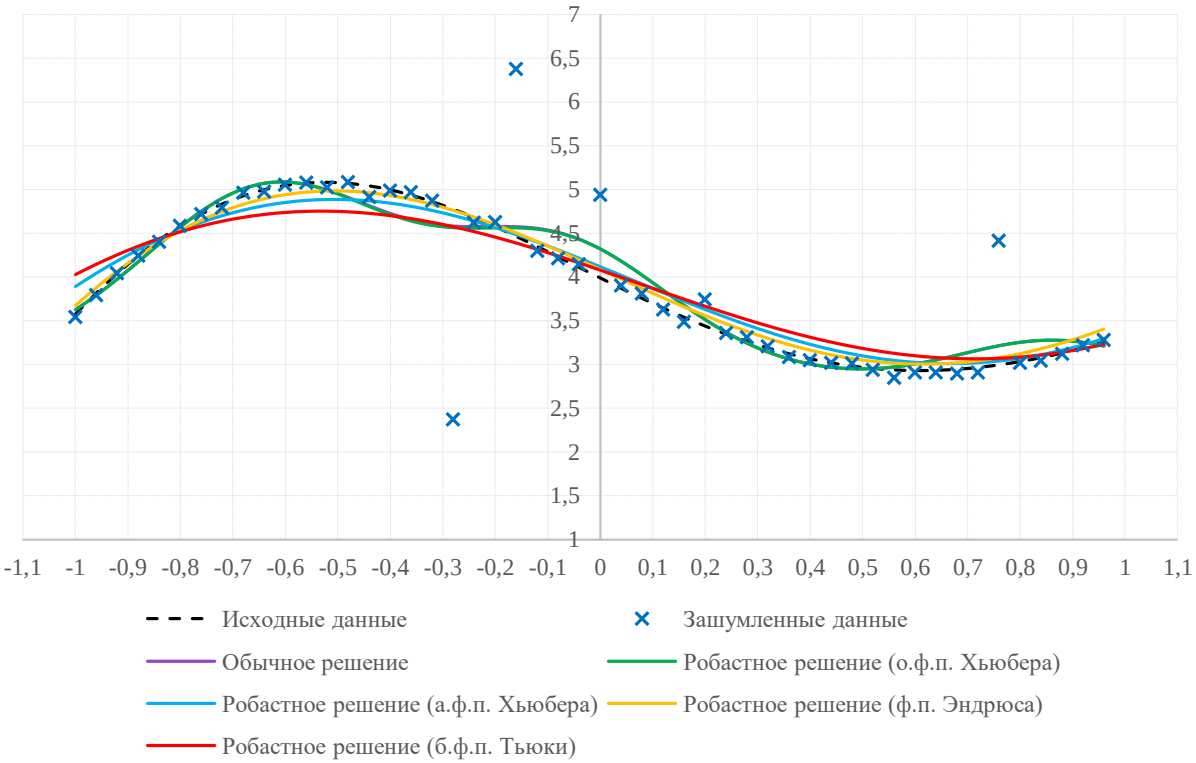


Рисунок А.20. Графики робастных моделей полученные методом взвешивания при 5% уровне шума, 10% уровне засорения и коэффициентом регуляризации $\gamma = 100$

ПРИЛОЖЕНИЕ Б

Акты о внедрении результатов диссертационной работы

УТВЕРЖДАЮ:
Проректор по учебной работе
д.т.н., доцент
_____ С.В. Брованов
«05» июня 2018 г.

АКТ

о внедрении в учебный процесс Новосибирского государственного технического университета результатов диссертационной работы Бобоева Ш.А.

Настоящим актом подтверждается, внедрение результатов диссертационной работы Ш.А. Бобоева «Построение регрессионных зависимостей с использованием квадратичной функции потерь в методе опорных векторов (LS SVM)» в учебный процесс факультета прикладной математики и информатики Новосибирского государственного технического университета.

Разработанные в диссертационной работе алгоритмы дают возможность получить робастные решения с использованием методов М-оценивания и разреженные решения путем разбиения выборки на части с использованием методов планирования эксперимента и внешних критериев оценки качества моделей. Основные положения и результаты диссертации включены в дисциплины «Статистические методы анализа данных», «Основы теории машинного обучения» и «Методы планирования эксперимента». Материалы диссертационной работы успешно используются при написании бакалаврских и магистерских диссертаций, а также в исследованиях аспирантов.

Заведующий кафедрой ТПИ
д.т.н., доцент _____ В.М. Чубич

ҶУМҲУРИИ ТОҶИКИСТОН
ДОНИШГОҲИ МИЛЛИИ ТОҶИКИСТОН

РЕСПУБЛИКА ТАДЖИКИСТАН
ТАДЖИКСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ

734025, ш. Душанбе, хиёбони Рӯдакӣ, 17 734025, г. Душанбе, проспект Рудаки, 17
тел.: (+992-37) 221-77-11, факс: (+992-37) 221-48-84
e-mail: tgnu@mail.tj, tnu.int.re@gmail.com

аз «____» _____ соли 201__ сод.№ _____
от «8» 01 2019 года исх.№ 12/ре

АКТ

о внедрении результатов диссертационной работы Бобоева Шарафа Асроровича «Построение регрессионных зависимостей с использованием квадратичной функции потерь в методе опорных векторов (LS SVM)»

Настоящим актом подтверждается, внедрение результатов диссертационной работы аспиранта кафедры ТПИ НГТУ Ш.А. Бобоева «Построение регрессионных зависимостей с использованием квадратичной функции потерь в методе опорных векторов (LS SVM)» для изучения процесса комплексообразования переходных металлов с производными теомочевины, 1, 2, 4 – триозола в водных и водно-органических растворах в Научно-исследовательском институте ТНУ.

Проректор по научной работе,
д.х.н., профессор С.М. Сафармамадов

Директор НИИ ТНУ
к.фарм.н., доцент Н.Б. Саидов

ПРИЛОЖЕНИЕ В

Свидетельство о государственной регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2018619675

Программа «ПОЛУЧЕНИЕ РОБАСТНЫХ И РАЗРЕЖЕННЫХ РЕШЕНИЙ МЕТОДОМ LS SVM "Robast_Sparse_LS-SVM"»

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Новосибирский государственный технический университет» (RU)***

Авторы: ***Бобоев Шараф Асрорович (TJ), Попов Александр Александрович (RU)***

Заявка № **2018617162**
Дата поступления **09 июля 2018 г.**
Дата государственной регистрации в Реестре программ для ЭВМ ***09 августа 2018 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев