Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ Г. М. Цибульский
«____» _____ 2017 г.

# МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа

09.04.02 Информационные системы и технологии

09.04.02.01 Информационно-управляющие системы

| | | |
|---|---|---|
| Руководитель | проф., д-р.физ.-мат. наук | Б. С. Добронец |
| Студент | КИ15-02-1/1М 031513681 | А. А. Чевер |
| Рецензент | канд. физ.-мат. наук, ст. науч. сотр. | А. Н. Рогалев |
| Нормоконтролер | доцент | М. А. Аникьева |

Красноярск 2017

Продолжение титульного листа магистерской диссертации по теме «Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа».

Нормоконтролер М. А. Аникьева

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой
___ Г. М. Цибульский
« ___» _____ 2017 г.

**ЗАДАНИЕ**
**НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ**
**в форме магистерской диссертации**

Студенту Чевер Александре Александровне.

Группа КИ15-02-1/1М. Направление 09.04.02 Информационные системы и технологии.

Тема магистерской диссертации «Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа».

Утверждена приказом по университету №_____от_____.

Руководитель магистерской диссертации Б. С. Добронец, профессор, доктор физико-математических наук, заведующий кафедрой систем искусственного интеллекта ИКИТ СФУ.

Исходные данные для магистерской диссертации: методические указания научного руководителя, статьи, книги, научные журналы, монографии по теме исследования.

Перечень разделов ВКР: введение, анализ предметной области, численный вероятностный анализ, ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ, ЗАКЛЮЧЕНИЕ, СПИСОК ИСПОЛЬЗОВАННЫХ СОКРАЩЕНИЙ, СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ, ПРИЛОЖЕНИЯ.

Перечень графического материала: плакаты презентации, выполненная в Microsoft Office PowerPoint 2010.

Руководитель ВКР _____ Б. С. Добронец

Задание принял к исполнению _____ А. А. Чевер

« __ » _____ 2017 г.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ Г. М. Цибульский
«____» _____ 2017 г.

# ГРАФИК
# НАПИСАНИЯ И ОФОРМЛЕНИЯ
# ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
## в форме магистерской диссертации

Студент: Чевер Александра Александровна.

Группа: КИ15–02–1/1М. Направление: 09.04.02.01 Информационно-управляющие системы.

Тема выпускной квалификационной работы: разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа.

График выполнения выпускной квалификационной работы (ВКР) приведён в таблице 1.

Таблица 1 – График выполнения этапов ВКР

| Наименование / содержание этапа | Срок выполнения | Примечания |
|---|---|---|
| Анализ предметной области, подбор литературы | До 14 февраля 2016 | |
| Составление плана работы над ВКР | До 14 марта 2016 | |
| Разработка и предоставление на проверку первой главы | До 30 мая 2016 | |
| Разработка и предоставление на проверку второй главы | До 30 сентября 2016 | |
| Работа над экспериментальной частью исследования | До 29 декабря 2016 | |
| Разработка и предоставление на проверку третьей главы | До 31 января 2017 | |
| Доработка ВКР в соответствии с полученными замечаниями | До 8 мая 2017 | |
| Разработка тезисов доклада и подготовка презентации для защиты | До 1 июня 2017 | |
| Согласование с руководителем тезисов доклада и презентации | До 15 июня 2017 | |
| Прохождение нормоконтроля | До 14 июня 2017 | |
| Ознакомление с отзывом и рецензией | До 16 июня 2017 | |
| Завершение ВКР к защите с учётом отзыва и рецензии | До 19 июня 2017 | |

# РЕФЕРАТ

Выпускная квалификационная работа по теме «Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа» содержит 56 страниц текстового документа, 26 иллюстраций, 1 таблицу, 30 использованных источника, 2 приложения.

Объект исследования — эмпирические данные.

ЭМПИРИЧЕСКИЕ ДАННЫЕ, ЧИСЛЕННЫЙ ВЕРОЯТНОСТНЫЙ АНАЛИЗ, СЛУЧАЙНАЯ ВЕЛИЧИНА, ПЛОТНОСТЬ ВЕРОЯТНОСТИ, МЕТОДЫ ВОССТАНОВЛЕНИЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ, ГИСТОГРАММА, ОСРЕДНЕНИЕ СМЕЩЁННЫХ ГИСТОГРАММ, МЕТОДЫ АППРОКСИМАЦИИ, СГЛАЖИВАЮЩИЙ СПЛАЙН.

Цель исследования — повышение эффективности и качества оценки состояния технических систем, изделий и исследований в условиях неопределённости с использованием обработки эмпирических данных методом восстановления плотности вероятности.

С этой целью в результате выполнения работы был проведён анализ данной области, выявлена актуальность темы исследования, проведены многочисленные исследования методов, позволяющих извлекать максимум информации из эмпирических данных в условиях неопределённости.

На основании результатов, полученных в ходе исследования, был предложен и реализован метод восстановления плотности вероятности — метод осреднения смещённых гистограмм. Преимущества данного метода в сравнении с уже существующими методами — простота реализации, наглядность, точность результатов. Данный метод применим в областях, где приходится сталкиваться с обработкой эмпирических данных (медицина, космическая отрасль, гидроэнергетика и др.).

СОДЕРЖАНИЕ

# ВВЕДЕНИЕ

Тема магистерской диссертации «Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа».

Анализ изучаемой области показал, что тема обработки эмпирических данных с получением максимально надёжных и достоверных результатов в настоящее время является актуальной и имеет применения в следующих областях: системы ответственного назначения, производство и эксплуатация дорогостоящих и высоконадежных технических изделий, медицинские исследования, повышение качества оценки состояния технических систем и др.

Основная проблема исследования — обработка эмпирических данных в условиях неопределённости. В качестве объекта исследования в работе выступают эмпирические данные. Доказано, что наиболее исчерпывающая информация об объекте исследования в условиях неопределённости на основе полученных эмпирических данных получается путём восстановления плотности вероятности неизвестной случайной величины. Следовательно, в роли предмета исследования выступают методы восстановления плотности вероятности. В настоящее время существует достаточное количество методов исследования, каждый из которых обладает как положительными, так и отрицательными сторонами: большие вычислительные затраты, потеря точности и т. д.

Главной целью исследования является повышение эффективности и качества оценки состояния технических систем, изделий и исследований в условиях неопределённости с использованием метода восстановления плотности вероятности.

Для достижения поставленной цели в ходе исследования необходимо решить следующие основные задачи:

1) анализ предметной области и выявление проблем исследования;

2) анализ существующих методов восстановления плотности вероятности случайных величин;

3) разработка метода восстановления плотности вероятности с учётом опыта исследователей;

4) анализ существующих методов сглаживания и применение оптимальных к полученным результатам;

5) выводы о проделанной работе, научно значимости и целесообразности дальнейших исследований.

С учётом опыта исследований и разработок в данной области в качестве решения поставленной проблемы предлагается один из методов восстановления плотности вероятности — метод осреднения смещённых гистограмм. Преимущества метода в сравнении с уже существующими методами заключаются в быстроте реализации, минимальных вычислительных затратах, графическом представлении обрабатываемых данных, хорошем качестве приближения данных.

Текст диссертации состоит из введения; трёх глав, содержащих теоретические сведения и практическую часть; заключения, списка сокращений, списка использованных источников, двух приложений.

В ходе экспериментального исследования был реализован метод осреднения смещённых гистограмм. В тексте диссертации представлен один из примеров и полученные результатов.

## 1 Анализ предметной области

Тема магистерской диссертации «Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа».

В современной практике часто приходится обрабатывать результаты многократных измерений наблюдаемой величины. В качестве примера можно взять следующие области: службы контроля качества предприятий; производство и эксплуатация дорогостоящих и высоконадежных технических изделий. Аналогичные примеры можно отыскать в медицине, биологии, гидроэнергетике, ракетостроении и т. д. [6]. Также одним из примеров могут стать системы ответственного назначения. Технические системы ответственного назначения — это системы, изготавливаемые в небольшом числе экземпляров, эксплуатирующиеся в особых условиях и реализующие экстремальные технологии. Обычно относятся к числу уникальных систем.

В этих и других областях приходится обрабатывать выходные сигналы нескольких датчиков, измеряющих одну и ту же величину (температура, давление, изменение состояние объекта или навигационные параметры и т. д). Алгоритм обработки полученных измерительных данных напрямую зависит от объёма исходной информации, её качества и наличия помех.

В процессе работы с такого рода системами необходимо решать большое число задач, связанных с управлением надёжностью, оценкой сопротивляемости технических устройств, прогнозированием отказов технических изделий и многие другие.

Оценка показателей надежности осуществляется на основе информации, полученной в результате специально организованных испытаний. Наиболее объективной информацией для определения характеристик надежности объектов является эксплуатационная информация, т. к. такого рода информация отражает всю специфику функционирования объектов, все воздействующие на них факторы и другие особенности эксплуатации. Прогнозирование состояния

и надежности необходимо при планировании эксплуатации таких систем полностью зависит от результатов обработки данных. Следовательно, необходим метод, позволяющий получать исчерпывающую информацию о необходимом объекте исследования и давать максимально надёжное описание.

## 1.1 Проблема исследования

Однако не каждую задачу можно решить в ходе эксплуатации или путём непрерывного контроля состояния системы, поскольку в течение определённых промежутков времени эксплуатации техническое обслуживание становится невозможным (летательных аппаратов — во время полёта, надводных и подводных судов — во время рейса, радиоэлектронной аппаратуры — в течение сеанса связи или слежения за целью и т. д.).

Основные трудности при решении задачи прогнозирования связаны с тем, что прогноз приходится осуществлять для каждого объекта индивидуально, при различных объемах исходной информации и в присутствии помех (ошибок контроля), статистические свойства которых достоверно не известны. Решение проблем обработки и получение в дальнейшем адекватной информации имеет важнейшее значение при оценке проведенных исследований и дальнейшем планировании [22].

Основные трудности, с которыми приходится сталкиваться при обработке эмпирических данных:

1) решения принимаются в условиях неопределённости;

2) выбор оптимального метода обработки информации, позволяющий максимально извлекать полезную информацию из любого объёма данных;

3) в случае отказа систем, действующих на основе обработанных данных, возникает высокий уровень риска.

## 1.2 Объект и предмет исследования

В настоящее время особую актуальность имеют вопросы обеспечения надёжного функционирования объектов абсолютно различных областей. Именно поэтому возникает необходимость разработки специальных методов обработки статистической информации, чтобы в дальнейшем определять характеристики надёжности этих объектов.

Решение задач, описанных ранее, характеризуются высоким уровнем неопределённости. Поэтому поиск методов и оптимальных подходов к построению решений в условиях неопределенности является важной и значимой практической задачей.

Задача оценивания неизвестных параметров по некоторому числу наблюдений является одной из основных в параметрической статистике. За время развития теории вероятностей и математической статистики создана теория оценивания, решающая эту проблему.

При обработке некоторых видов информации используют параметрические методы (t-критерий Стьюдента, хи-квадрат Пирсона и т. д.). Однако использование параметрических методов связано с имеющимися предположениями о виде закона распределения наблюдаемых случайных величин. В некоторых случаях тяжело, а иногда и вовсе невозможно определить закон распределения. Кроме того, параметрические формулы дают приближённое значение для истинных плотностей вероятности. В некоторых случаях погрешности таких приближений оказываются ввиду своих вычислительных затрат являются достаточно высокими.

Следовательно, необходимы методы, не уступающие параметрическим методам в эффективности, а в лучшем случае, были бы малочувствительны к нарушению предположений, лежащих в основе параметрической модели. Поэтому для анализа такой статистической информации принято использовать непараметрические методы. непараметрические методы с более трудоёмкие с вычислительной точки зрения, чем параметрические, а порой и очень сложные.

Это сдерживало их применение раньше. Однако после появления компьютеров положение изменилось и теперь во всех наиболее распространенных пакетах прикладных программ реализованы и непараметрические процедуры.

В настоящее время разработка непараметрических методов обработки информации является актуальным и важным вопросом. К непараметрическим методам построения плотности вероятности относятся гистограммные, ядерные, частотные методы и др.

Одной важной характеристикой, описывающей поведение случайной величины, является её плотность вероятности $f(x)$. Зная плотность вероятности случайной величины, можно определить вероятность отказа, интенсивность отказа, среднее время между отказами систем и др. При этом важно оценить, какое количество информации содержится в выборке заданного объёма и какое количество информации необходимо для получения результата с заданной точностью и достоверностью.

Объектом исследования выступят эмпирические данные, требующие обработки. Предметом исследования станут методы восстановления плотности вероятности.

### 1.3 Цели и задачи исследования

Основная проблема исследования — обработка эмпирических данных в условиях неопределённости. Решение проблемы позволит повысить эффективность и качество оценки состояния технических систем, изделий и исследований в условиях неопределённости с использованием метода восстановления плотности вероятности. В данном исследовании в роли метода восстановления плотности вероятности будет метод осреднения смещённых гистограмм.

Для достижения поставленной цели в ходе исследования необходимо решить следующие задачи:

1) анализ предметной области;

2) выявление проблем исследования;

3) определение актуальности исследования;

4) анализ существующих методов восстановления плотности вероятности случайных величин;

5) разработка метода восстановления плотности вероятности с учётом опыта исследователей;

6) анализ существующих методов сглаживания;

7) оптимизация параметров и алгоритмов для дальнейшего использования;

8) выводы о проделанной работе, научно значимости и целесообразности дальнейших исследований.

## 2 Численный вероятностный анализ

В ходе изучения различного рода систем приходится сталкиваться с некоторыми трудностями, одной из которых является недостаточное количество информации об объекте и его процессах. В результате чего необходимо решать задачи обработки информацией в целях дальнейших исследований в условиях неопределённости. На качество принимаемых решений будет влиять полученный уровень знаний, который будет определяться различными подходами и методами. Кроме того, особое значение имеют методы и способы обработки этой информации, отличающиеся между собой сложностью реализации. Поэтому возникает ещё одна задача разработки метода, который позволит не только увеличить достоверность информации, но и понизить уровень сложности реализации самого метода, чтобы не только сэкономить временные, трудовые ресурсы, но и исключить возможность ошибки за счёт реализации сложного алгоритма обработки. Благодаря численному вероятностному анализу решение может быть найдено.

Численный вероятностный анализ — раздел вычислительной математики, занимающийся решением задач со случайными входными данными.

Изучение объектов и систем с использованием численного вероятностного анализа — особо важный и полезный инструмент в условиях неопределённости и риска. Предметом численного вероятностного анализа является решение различных задач со стохастическими неопределенностями в данных с использованием численных операций над плотностями вероятностей случайных величин и функций со случайными аргументами. Для этого предлагается разнообразный инструментарий, включающий такие понятия, как гистограммная арифметика, вероятностные, естественные и гистограммные расширения, гистограммы второго порядка.

Численный вероятностный анализ представляет собой непараметрический подход и может успешно применяться для вероятностного

описания систем в рамках визуально-интерактивного моделирования, повышая тем самым качество исследования систем.

С целью снижения уровня информационной неопределенности и получения дополнительной информации о распределении параметров в условиях информационной недостаточности предлагается использовать гистограммный подход. Для решения таких задач можно также использовать интервальные гистограммы и гистограммы второго порядка. В тех случаях, когда нет возможности получить точную функцию распределения случайной величины задают оценки плотности распределения. Подробнее о гистограммном подходе см. в пункте 2.2.1.

## 2.1 Функция плотности вероятности

Как говорилось ранее, в большинстве случаев решения задач принимаются в условиях неопределённости.

Случайная величина, одно из понятий теории вероятности — это величина, которая принимает определённое значение, и это значение заранее неизвестно, оно определяется в ходе эксперимента. Случайная величина $X$ задаётся на интервале $(a, b)$, где $X \in (a, b)$. Если границы неопределённые и значения случайной величины занимают некоторый промежуток, то такие величины называются непрерывными случайными.

Одной из основных идей в статистике является понятие функции плотности вероятности $f(x)$. Функция плотности вероятности — важная характеристика случайной величины. Поэтому проблема оценки данной функции важна. Восстановление плотности вероятности позволяет получить надежное описание системы с получением максимально достоверных результатов.

Пусть имеется непрерывная случайная величина $X$ с функцией распределения $F(x)$, непрерывная и дифференцируемая. Необходимо

вычислить вероятность $P$ попадания этой случайной величины в заданный интервал $(x, x + \Delta x)$:

$$P(x < X < x + \Delta x) = F(x + \Delta x) - F(x) \quad (1)$$

Функция плотности вероятности $f(x)$ непрерывной случайной величины — производная функции распределения $F(x)$ — характеризует плотность, с которой распределяются значения случайной величины в определенной точке [27]:

$$f(x) = F'(x) \quad (2)$$

Ещё одной важной характеристикой случайной величины является математическое ожидание $M[X]$, среднее значение случайной величины:

$$M[X] = \sum_{i=1}^{n} x_i p_i \quad (3)$$

При восстановлении данных по выборке из генеральной совокупности большинство исследователей данной области считают плотность вероятности исчерпывающей характеристикой для любого закона распределения вероятности. Восстановление плотности вероятности позволяет получить максимальное и надёжное описание параметров системы. Проблема восстановления плотности вероятности актуальна для широкого спектра прикладных наук.

### 2.2 Методы восстановления плотности вероятности

Задача восстановления плотности вероятности по выборке является основной проблемой математической статистики [4]. В большинстве случаев точный вид закона распределения рассматриваемой генеральной совокупности

неизвестен, поэтому приходится восстанавливать плотность вероятности по имеющейся выборке. Задача восстановления плотности вероятности имеет практическое применение во многих областях: наука, техника, статистика, медицина, биология и т. д.

Существует достаточное количество непараметрических методов и алгоритмов восстановления плотности вероятности. Все они в равной степени обладают как определёнными преимуществами, так и недостатками. Работоспособность методов при анализе экспериментальных данных можно подтвердить результатами данных, полученных исследователями в своих трудах (см. список использованных источников). Некоторые методы восстановления плотности вероятности приведены ниже:

1) метод гистограмм;
2) метод ядерных оценок;
3) метод интегральной оценки плотности вероятности (даёт хорошие результаты при работе с малыми объёмами данных);
4) метод стохастической регуляризации (используется при больших объёмах данных);
5) метод оценки максимума правдоподобия;
6) метод рекуррентных ядерных оценок;
7) метод интерполяции сплайнами;
8) проекционный метод;
9) метод структурной минимизации риска (используется при работе с малыми объёмами данных);
10) метод корневой оценки;
11) метод Карандеева-Эйсымонта (даёт точные значения при небольших выборках) и др.

Подобное разнообразие методов сформировалось в результате того, что задача восстановления плотности вероятности является актуальной для очень широкого спектра прикладных наук. Однако при определённых параметрах все эти методы могут быть идентичными друг другу.

### 2.2.1 Метод гистограмм

Восстановления плотности вероятности методом гистограмм является наиболее старой и традиционной непараметрической оценки плотности вероятности из-за простоты в реализации. При правильном подходе данный метод по достоверности полученных данных может быть эффективен и сравним с другими методами.

Термин «гистограмма» был введён статистиком Карлом Пирсоном [25] и состоит из слов «столб» и «нечто записанное». Но гистограммы использовались ещё до того, как получили своё название. Они полезны, даже если отсоединить их от графического представления и рассматривать как чисто математические объекты, сохраняющие приближенное распределение данных.

Гистограмма — это геометрическое изображение эмпирической функции плотности вероятности некоторой случайной величины, построенное по выборке. Высота каждого столбца указывает на частоту появления значений параметров в выбранном диапазоне, а количество столбцов – на число выбранных диапазонов. Особо важное достоинство гистограмм заключается в том, что они позволяют наглядно представить изменения параметров объекта и зрительно их оценить [18].

Пусть $x_0$ и $x_N$ — наименьшее и наибольшее значения выборки. Величина $R = (x_0 - x_N)$ — размах выборки. Гистограмма — это кусочно-постоянная функция, определяющаяся сеткой $\{x_i | i = 0, \dots, N\}$, откладывающейся на горизонтальной оси, и на каждом $[x_i, x_{i+1}]$ отрезке принимающая постоянное значение $p_i$. Если все интервалы были равными, то высота каждого получившегося прямоугольника пропорциональна числу элементов выборки, попадающих в соответствующий интервал:

$$p_i = \frac{v_i}{N(x_i - x_{i-1})}, \tag{4}$$

где $v_i$ — число точек, попадающих в интервал $[x_i, x_{i+1})$;

$x \in [x_{i-1}, x_i)$;

$h$ — ширина шага.

На рисунке 1 показан один из примеров построения гистограммы.

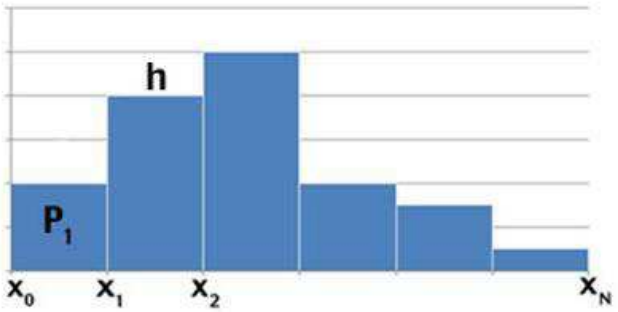


Рисунок 1 — Построение гистограммы

Вероятности событий, относящихся к наблюдаемой случайной величине, можно выразить через её функцию распределения $F$. Следовательно, зная приближение к функции $F$, можно получить приближенные значения для любых вероятностей.

### 2.2.2 Частотный полигон

Полигон частот — один из способов графического представления плотности вероятности случайной величины. Это ломаная, соединяющая точки $(x_i, n_i)$, где $i = 1,2, \dots, m$.

При построении полигона на горизонтальной оси (ось абсцисс) откладывают значения вариант $x_i$, а на вертикальной оси — соответствующие частоты $n_i$. Точки соединяются прямыми отрезками, в результате чего и получается полигон частот (рисунок 2).

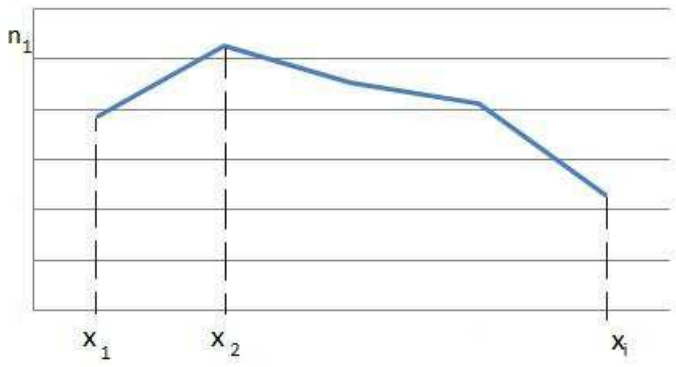


Рисунок 2 — Построение полигона частот

Полигон частот используется для представления непрерывного и дискретного распределения. Если распределение непрерывное и график эмпирического распределения описывается плавной зависимостью, то полигон частот является лучшим способом графического представления, чем гистограмма.

Кроме того, существует полигон накопленных частот. Полигон накопленных частот получается в результате соединения отрезков прямых точек, координаты которых соответствуют верхним интервалам группировкам и накопленным частотам.

### 2.2.3 Метод ядерных оценок

Впервые ядерная оценка была получена учеными Розенблаттом и Парзеном. Процедура оценки плотности вероятности в одной точке выглядит следующим образом:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-T_i}{h}\right), \quad (5)$$

где $T$ — выборка, полученная в ходе наблюдения за объектом;

$K$ — симметричное ядро;

$h$ — диапазон, параметр сглаживания, влияющий на точность оценок;

$n$ — размер выборки.

В качестве ядра используется Гауссово ядро:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad (6)$$

Таким образом, плотность вероятности в точке $X$ вычисляется как сумма значений ядра для величин, определяемых разностями между значением $X$ и значениями последовательности. Точки $X$ при этом, в которых вычисляется плотность, могут не совпадать со значениями самой последовательности.

### 2.3 Метод осреднённых смещённых гистограмм Дэвида Скотта

В одной из своих работ [29] Дэвид Скотт предложил метод осреднённых смещённых гистограмм, позволяющий учитывать все имеющиеся пустоты между столбцами частотного полигона, сохраняя при этом вычислительные преимущества оценки плотности распределения. В ходе эксперимента параметры сглаживания оставались неизменными, но при этом менялись начальные точки построения столбцов. Чтобы решить проблему сглаживания, Скотт предложил метод осреднения нескольких смещённых гистограмм. По словам учёного, данный метод достаточно хорош по сравнению с остальными для оценки плотности распределения. Рассмотрим алгоритм построения.

Пусть имеется m гистограмм $\widehat{f_1}$, $\widehat{f_2}$, … , $\widehat{f_m}$ с шириной каждого столбца h и с начальным столбцом в точке $t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \dots, \frac{(m-1)h}{m}$. Исходная осреднённая смещённая гистограмма определяется как $\hat{f}(\cdot) = \hat{f}_{ASH}(\cdot) = \frac{1}{m}\sum_{i=1}^{m} \widehat{f_i}(\cdot)$. На каждом из интервалов шириной $\delta = \frac{h}{m}$ является кусочно-линейной. В качестве примера в работе Скотта рассмотрен следующий.

Имеется набор данных, изображённый рисунке 4 ($m = 1$), ширина каждого столбца (шаг) $h = 12{,}5$ дюймов. В ходе эксперимента происходит смещение гистограмм относительно оси x на значение $\delta = \frac{h}{m}$. Максимальное

число смещения гистограмм $m = 32$. Как видно из рисунка 3, полученные данные получились гладкими.

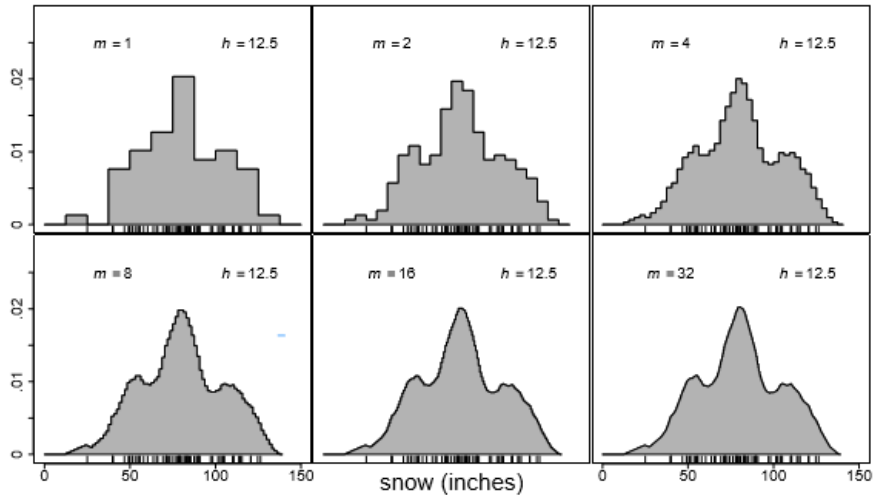


Рисунок 3 — Смещение гистограмм $m = 32$

В случае работы с многомерными данными формула исходной осреднённой смещённой гистограммы приобретает следующий вид:

$$\hat{f}(\cdot,\cdot) = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \widehat{f_{\iota,J}}(\cdot,\cdot) \qquad (7)$$

Начальный столбик двумерной гистограммы смещается на $\widehat{f_{\iota,J}}$ и вычисляется как точка $(x, y) = ((i-1)\delta_i, (j-1)\delta_2$. Пример использования метода осреднения гистограмм в случае многомерных данных представлен на рисунке 4.

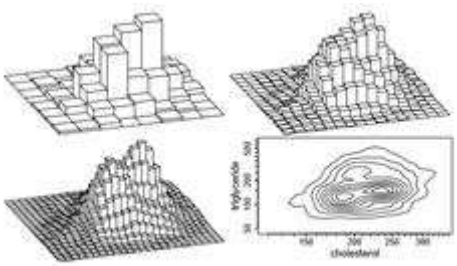

Рисунок 4 — Метод осреднения гистограмм в случае многомерных данных

На интервалах $[k\delta,(k+1)\delta]$, где $\delta=\frac{h}{m}$, осредненная смещённая гистограмма является кусочно-постоянной. Удобно ссылаться на этот интервал как на столбик $B_K$. Число столбцов для гистограммы может быть получено путем добавления m соседних столбцов $\{v_k\}$.

Рассмотрим оценку осреднённой смещённой гистограммы с начальным столбцом $B_0$. Высота гистограммы в точке $B_0$ — это осредненная высота $m$ смещённых гистограмм, каждая шириной $h=m\delta$ со столбцом $B_0$: $\frac{v_{1-m}+\cdots+v_0}{nh},\frac{v_{2-m}+\cdots+v_0+v_1}{nh},\ldots,\frac{v_0+\cdots+v_{m-1}}{nh}$. Следовательно, общее выражение для осреднённой смещённой гистограммы выглядит следующим образом:

$$\hat{f}(x;y)=\frac{1}{m}\sum_{i=1-m}^{m-1}\frac{(m-|i|)v_{k+i}}{nh}=\frac{1}{nh}\sum_{i=1-m}^{m-1}(1-\frac{|i|}{m})v_{k+i}, \qquad (8)$$

где $x\in B_k$.

Весовая функция общего числа столбцов, рассчитанного по уравнению (6), принимает форму равнобедренного треугольника с основанием $(-1;1)$. Другие формы могут рассматриваться как формы равномерных весовых функций или сглаживаемой (дифференцируемой) формы. Общая формула для осреднённой смещённой гистограммы использует произвольные весовые функции $w_m(i)$ и определяется как

$$\hat{f}(x;y)=\frac{1}{nh}\sum_{|i|<m}w_m(i)v_{k+i} \qquad (9)$$

Для того чтобы $\int\hat{f}(x;m)dx=1$, суммирование весовых функций должно происходить по $m$. Простой способ определения общей весовой функции:

$$w_m(i)=m\cdot\frac{K(\frac{i}{m})}{\sum_{j=1-m}^{m-1}K(\frac{j}{m})}, \qquad (10)$$

где $K$ – непрерывная функция, определяемая на интервале $(-1; 1)$.

Функцию $K$ часто выбирают как функцию плотности вероятности, которая называется двойной весовой функцией ядра или ядрами 4-го порядка.

Вычислительный алгоритм в общем виде для осреднённой смещённой гистограммы выглядит следующим образом. Необходимо построить равномерно распределенную сетку шириной δ на интервале $(a, b)$ и вычислить соответствующее начало отсчёта $\{v_k, k = 1, \ldots, nbin\}$ для $N$ точек. Как правило, $\delta \ll h$, и $nbin$ относится к числу столбцов шириной $\delta$. Это вычисление выполняется по алгоритму BIN1 (вычисление первого столбца), приведенному на рисунке 5.

**BIN1**$(x, n, a, b, nbin)$**Algorithm**: (* Bin univariate data *)

$\delta = (b - a)/nbin$
for $k = 1, nbin$ $\{v_k = 0\}$
for $i = 1, n\{$
$k = (x_i - a)/\delta + 1$ (* integer part *)
if $(k \in [1, nbin])$ $v_k = v_k + 1\}$
return $(\{v_k\})$

Рисунок 5 – Алгоритм BIN1

Далее вычислим вектор весовых коэффициентов $\{w_m(i)\}$, как и в уравнении (10). Следующим этапом необходимо произвести оценку одномерной осреднённой гистограммы $\{f_k, k = 1, \ldots, nbin\}$ по интервалам $nbin$, которые могут быть вычислены эффективно путем изменения порядка операций, указанных в формуле (9). Вместо того чтобы производить вычисления оценок осреднённых смещённых гистограмм индивидуально для каждого столбца, захватывая $(2m - 1)$ соседних столбцов, за один раз производится подсчёт с весовыми функциями $(2m - 1)$, применимых к соседним оценкам осреднённых смещённых гистограмм.

Данная модификация позволяет избежать многократного взвешивания пустых столбцов. Алгоритм предполагает, что существуют по крайней мере $(m - 1)$ пустых столбцов с каждого конца. Отметим, что объём исследования

определяется числом $m$ и количеством непустых столбцов. Алгоритм является достаточно эффективным даже при N $>10^6$, в этом случае большую часть работы включает в себя составление таблицы в несколько сотен столбцов.

Для большого объёма данных дополнительно необходимо использовать визуальное сглаживание, даже когда сглаживание параметров незначительно (рисунок 6).

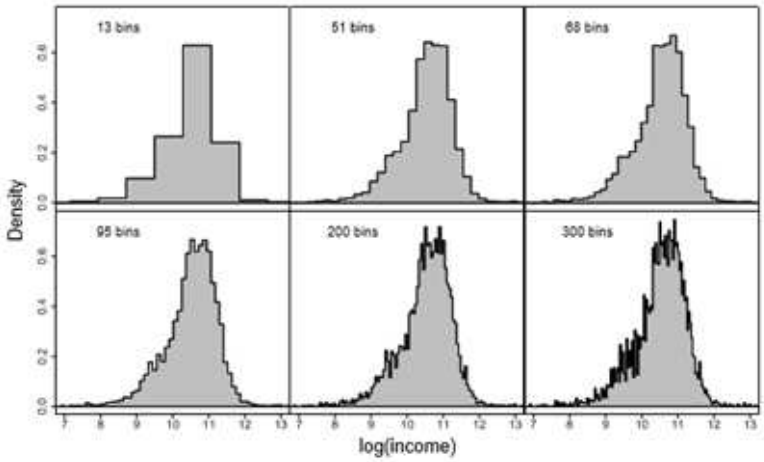


Рисунок 6 — 6 гистограмм из численного примера

На практике ширина столбца $\delta$ обычно фиксируется первыми выбранными в диапазоне от 50 до 500 столбцами от всего размера выборки (значение увеличивается на 5-10 % для того, чтобы включить некоторые пустые столбцы с обеих сторон). Значение $h$ $(m\delta)$ — это только значения параметра сглаживания $h$, которые являются целыми и кратными $\delta$ и могут быть рассмотрены, хотя с них легко снять ограничения. С другой стороны, если ширина шага $h$ известна, тогда $\delta$ может быть вычислена как $\frac{h}{5}$ и $\frac{h}{10}$.

Однако такой случай является редким. Многие данные большого объёма предварительно группируются; то есть, исходные данные не записываются, а записывается только число столбцов. Если ширина этих столбцов называется $\delta$ и $h^*$ и оказывается близка к $\delta$, то никакого дополнительного сглаживания не применяется, начиная с $m = 1$, где это единственный вариант. Тщательный анализ поможет избежать такого плохого исхода.

Например, используя сглаживание, можно выбрать $\delta$ достаточно малым или $N$ достаточно большим, таким, что $\delta < h_{OS}/25$ или $\delta < h_{OS}/50$. Для того чтобы оценить дисперсию данных, применяя правило ширины столбцов для сглаживания частотного полигона, необходимо проводить исследования.

AMISE для исходных осреднённых смещённых гистограмм (рисунок 7).

| Estimator | $N(\bar{x}, s^2)$ | ASH | FP-ASH | FP | Histogram |
|---|---|---|---|---|---|
| Sample Size | 100 | 436 | 436 | 546 | 2,297 |

**Theorem** 5.1: *For the naive ASH with the isosceles triangle kernel,*

$$\text{AMISE} = \frac{2}{3nh}\left(1+\frac{1}{2m^2}\right) + \frac{h^2}{12m^2}R(f') + \frac{h^4}{144}\left(1-\frac{2}{m^2}+\frac{3}{5m^4}\right)R(f''). \quad (5.8)$$

Рисунок 7 – Эквивалентные примеры размеров выборок

Первое выражение в AMISE даёт ошибку из-за интегрированной дисперсии. ИКС или смещённая часть AMISE связывается с $R(f_0)$ и $R(f_{00})$, которые были найдены в ИКС гистограммы и полигоне частот.

Легко проверить, что первые 2 условия в результате соответствуют обычным гистограммам теоремы, где $m = 1$. С другой стороны, при $m \to \infty$, вторая гистограмма — как исчезающая и смещающая — аналогична частоте полигона. Как правило, для $m \geq 10$ среднее слагаемое ничтожно мало по сравнению с последним слагаемым, которое может быть принято равным $h^4/144$. Сравнивая уравнения, видим, что слагаемые идентичны, а ИКС для осреднённой смещённой гистограммы составляет $41\,\%$ для частотного полигона. Оптимальная ширина столбца для простой осреднённой смещённой гистограммы при $m \to \infty$:

$$h^*{}_{m=\infty} = \left[\frac{24}{nR(f'')}\right]^{1/5} \quad (11)$$

Размеры выборки на рисунке 8 обобщают эффективность осреднённой смещённой гистограммы и других оценок с имеющимися нормальными данными. Осреднённая смещённая гистограмма требует 80 % выборки, необходимой частотному полигону для достижения той же средней интегральной ошибки. Фактически, этот показатель на 80 % справедлив для любой плотности распределения.

В некоторых сложных ситуациях, например, когда есть небольшая выборка с неточной плотностью распределения, обычная гистограмма может составить конкуренцию осреднённой смещённой гистограмме. Но асимптотическая эффективность гистограммы будет нулевая относительно осреднённой смещённой или частотного полигона из-за разной скорости сходимости. Конечно, усовершенствование нового метода по отношению к частотному полигону не столь грандиозно, как усовершенствование частотного полигона по отношению к гистограмме, поскольку уменьшается время на принятие решения.

Выражение для нахождения асимптотической погрешности $L2$ частотного полигона или линейный интерполяционной исходной осреднённой смещённой гистограммы гораздо проще в реализации, чем исходная осреднённая смещённая гистограмма (рисунок 8).

**Theorem** 5.2: *For the frequency polygon interpolant of the naive ASH,*

$$\text{AMISE} = \frac{2}{3nh} + \frac{h^4}{144}\left(1 + \frac{1}{m^2} + \frac{9}{20m^4}\right) R(f''). \qquad (5.9)$$

Рисунок 8 — Нахождения асимптотической погрешности L2

Гистограмма как смещённое слагаемое включает в себя $R(f_0)$ и обращается в $0$. Кроме того, в зависимости от остальных условий, выбор $m$ значительно снижается. Как правило, $m \geq 3$ является достаточным для

достижения 20 % улучшения AMISE над частотным полигоном, а не при $m \geq 10$ как рекомендуется для самой осреднённой смещённой гистограммы.

Многомерный частотный полигон, изученный Скоттом (1985), использует треугольную сетку, но линейные сочетания результатов Хьорта (1986) немного лучше в реализации. Обозначим индексы по $f$ как частные производные (рисунок 9).

**Theorem** 5.3: *The AMISE of the multivariate linear blend of the naive ASH equals*

$$\frac{2^d}{3^d n\, h_1 \cdots h_d} + \frac{1}{720}\sum_{i=1}^{d} \delta_i^4 R(f_{ii}) + \frac{1}{144}\int_{\Re^d}\left[\sum_{i=1}^{d} h_i^2\left(1+\frac{1}{2m_i^2}\right) f_{ii}\right]^2 . \quad (5.10)$$

Рисунок 9 – Многомерный частотный полигон

За исключением особых обстоятельств, в замкнутом виде выражения для оптимальных параметров сглаживания недоступны. Они должны быть получены путем решения системы нелинейных уравнений. Если δi ≈ 0, тогда:

$$h * i \; = \; O(n - 1/(4 + d)) \qquad (12)$$

$$AMISE \; * = \; O(n - 4/(4 + d)) \qquad (13)$$

Выражения сопоставимы с результатами для многомерного частотного полигона в уравнении. Пока скорость сходимости та же, многомерный частотный полигон будет уступать методу осреднённых смещённых гистограмм.

Параметры в одномерной осреднённой смещённой гистограмме становятся векторами в алгоритме с двумерными ОСГ. Алгоритмы BIN1 и ASH легко распространяются на случаи, когда $d = 3$ и $d = 4$ при увеличении размерностей векторов и матриц. Для размеров больших, чем 4, вообще невозможно, чтобы существовало соответствие массивам достаточных

измерений непосредственно в памяти компьютера. В таких случаях, алгоритм ОСГ может быть модифицирован так, чтобы могли вычисляться только двух или трёхмерные части многомерных ОСГ.

### 2.4 Модифицированный метод осреднения смещённых гистограмм

Одним из самых распространенных методов восстановления плотности вероятности является метод гистограмм. Способ построения гистограмм описан в пункте 2.4. Несмотря на свою простоту, гистограммы охватывают весь диапазон представления плотностей вероятности. Процедура построения гистограмм достаточно проста: строится сетка $\omega = \{x_i | i = 1,2, \dots n\}$, вертикальные прямоугольники с площадью $\frac{m_i}{N}$, где высота прямоугольника $P_i = \frac{m_i}{N(x_i - x_i - 1)}$.

Основанием для использования гистограммы $ph(x)$ в качестве оценки неизвестной плотности вероятностей $f(x)$ является кусочно-интегральная сходимость $ph(x)$ к $f(x)$, которая вытекает из того, что относительная частота $\frac{m_i}{N}$ события $X \in (x_i - 1, x_i]$ сходится к его вероятности $f_i$:

$$\frac{m_i}{N} \to \int_{x_{i-1}}^{x_i} f(x)dx \tag{14}$$

На качество гистограмм влияют несколько факторов: объем выборки $N$ (объём исходной информации, рекомендуемый не менее 5), величина интервалов группировки, точность измерений. Кроме того, сложность состоит в том, что влияние этих факторов зависит от случайного распределения плотности вероятности, заранее неизвестной. На качество гистограммы будет влиять и точность измерений.

В то же время существует метод ядерных оценок, который, по сравнению с методом гистограмм, даёт наиболее точные результаты. При этом к получившейся функции плотности восстановления необязательно применять

метод сглаживания полученных результатов. Однако метод ядерных оценок сложен в своей реализации. Ниже представлена формула процедуры оценки функции в одной точке

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right), \quad (15)$$

где K – ядро;

$x$ – последовательность, длиной $n$;

$h$ – шаг.

Метод осреднения смещенных гистограмм показал хорошие результаты, сравнимые с методом ядерных оценок.

В работе Дэвида Скотта [29] был предложен метод, который по качеству получаемых результатов не уступает методу ядерных оценок. Это метод осреднения смещенных гистограмм. Изучив все предложенные Скоттом идеи и модернизировав его метод, представляем метод, суть которого описана ниже. Оригинальность метода заключается в том, что в нашем случае мы используем лишь середины столбцов гистограмм (точка $z$) и производим сглаживание получившихся точек. Осреднение смещенных гистограмм — это непараметрическая оценка плотности вероятности от ряда гистограмм. Вместо того чтобы выбирать оптимальные пары $(h, x_0)$, в [29] предложен метод усреднения смещенных гистограмм.

Необходимо построить $m$ гистограмм, сдвинуть каждую гистограмму на шаг $h$, с постоянным началом координат для каждой гистограммы в точке $t_0$. При всём при этом, вместо того, чтобы оставлять все смещенные гистограммы, мы по-прежнему сдвигаем эти гистограммы, оставляя при этом лишь середины отрезков $(z)$.

Смысл метода заключается в том, что начальная точка $x_0$ сетки выбирается равномерно из некоторого отрезка $[a_0, a_1]$ $m$ число раз. Соответственно для каждой сетки $\omega^k$ строится своя гистограмма $H_k$, $k = 1,\ldots,m$.

Гистограмма $H_k$ характеризуется своей сеткой $\omega^k = \{x_i^k = x_0^k + ih \mid i = 0,...,n\}$ и принимает на каждом отрезке $[x_{i-1}^k, x_i^k]$ постоянное значение $p_i^k$. В качестве приближения для функции плотности вероятности $f(x)$ строится функция $\hat{f}(x) = 1/m \sum_{k=1}^{m} H_k(x).$

Предполагаем, что носитель неизвестной плотности вероятности $[a,b]$. Таким образом, $x_0 = a$, $x_n = b$ оставляем неизменными. Выберем параметр $\delta \in (0,2h)$ и положим $x_1 = \delta$, далее сетку строим равномерной $x_i = \delta + (i-1)h$ $i = 1,2,...,n-1$. Выбирая $m$ значений $\delta \in (0,2h)$, получаем $m$ гистограмм. Усредним полученные гистограммы простым осреднением. В дальнейшем предлагается использовать для усреднения различные процедуры сглаживания.

В качестве численного примера рассмотрено восстановление функции плотности вероятности $f(x)$ суммы четырех равномерных на [0,1] случайных величин. На рисунке 10 приведен результат работы алгоритма. На рисунке 11 представлены сглаженные значения $\hat{f}_m$.

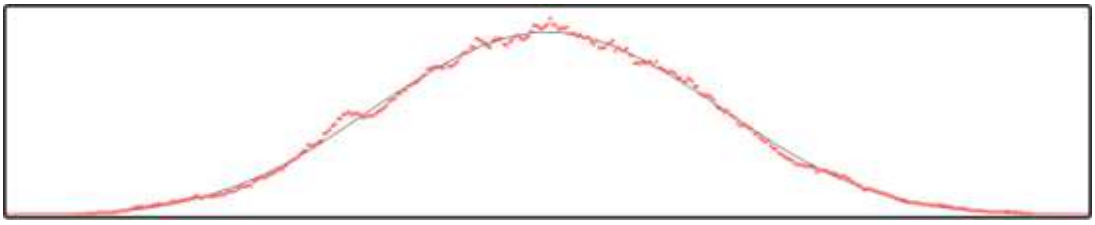

Рисунок 10 — Множество Z

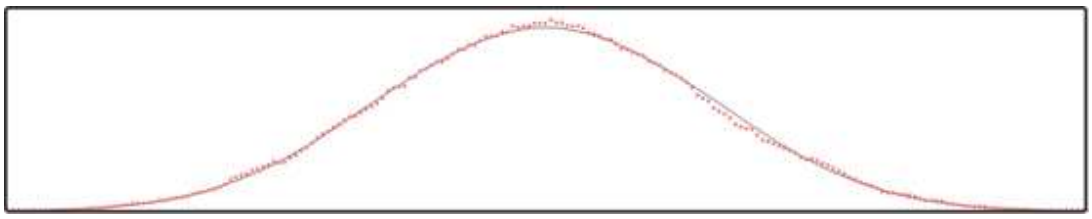

Рисунок 11 — Сглаженные значения $\hat{f}_m$

При этом $h = 0{,}1$, $n = 40$, $m = 20$ и $f(x)$ — сплошная линия, точками показано множество $Z$. Подробнее о ходе эксперимента см. в главе 3.

## 2.5 Использование пакета Pascal для реализации алгоритма

Среди огромного множества языков программирования для реализации алгоритма осреднения смещённых гистограмм, описанного в разделе 2.6, был выбран язык *PascalABC*.

*PascalABC* — это язык программирования, который включает в себе классический Паскаль, элементы языка программирования *Delphi* и собственные дополнительные расширения, что является несомненным плюсом.

На официальном сайте разработчика *(http://pascalabc.net)* имеется свободный доступ к пакету программы. Кроме того, имеется текстовая документация и все необходимые дополнительные компоненты.

Для первоначальной графической реализации алгоритма использовались возможности модуля *GraphABC* (*uses GraphABC*) – графической библиотеки языка, предназначенной для рисования объектов в графическом окне (*window*). Этот модуль содержит множество процедур и функций, предназначенных для работы с графическим экраном, а также константы и переменные, которые также могут потребоваться при работе с графикой.

Модуль обладает набором параметров и свойств пера, кисти и шрифта. В ходе рисование использовались следующие процедуры: *Pen.Color, putpixel, lineto, setpencolor* и другие.

Реализация метода осреднения смещённых гистограмм в среде *PascalABC* приведена в пункте 3.1.

## 2.6 Возможности пакета Matlab для аппроксимации

Среда *MATLAB* — это пакет прикладных программ, реализованный на языке С и включающий в себя интерпретатор команд, графическую систему, пакеты расширений. Работа в среде осуществляется через командное окно *Command Window,* открывающееся при запуске программы *matlab.exe*. Среда

позволяет выполнять математические вычисления, анализ данных, численные методы, программировать, обладает средствами графической визуализации и анимационной графикой (построение графиков функции, трёхмерная графика и т.д.).

Среда *MATLAB* достаточно гибкая и способна адаптироваться к различным задачам пользователей самых разных категорий. Кроме того, она может быть интегрирована с другими системами, что может привести к решению ряда более сложных задач. Весь необходимый стандартный набор компонент и документация по работе в среде имеется в свободном доступе в сети интернет.

На официальном сайте разработчика *www.mathworks.com* можно получить как полную версию программы, так и пробную. В нашем случае использовалась версия *MATLAB R2010a*.

Существует множество специальных пакетов инструментов (дополнительных приложений), разработанных для среды *MATLAB* и посвящённых решению более узких проблем: обработка сигналов и изображений, создание блок-схем, решение экономических задач и другие. Приложение *cftool* — одно из них. Оно позволяет сглаживать импортированные данные: работать с несколькими наборами данных, строить для них различные правила исключения, параметрические и непараметрические модели и создавать собственные модели с определёнными параметрами. В разделе 3.3 будут подробно рассмотрены и использованы на практике возможности данной среды разработки.

## 3 Экспериментальная часть

После проведение теоретических исследований, сбора необходимых знаний и опыта учёных и исследователей в данной области следует приступить к практической реализации алгоритма, описанного в пункте 2.5.

### 3.1 Использование возможностей среды PascalABC

Поставленная практическая задача – восстановить плотность вероятности по имеющимся данным. Воспользуемся средой разработки *PascalABC*. Листинг программы представлен в Приложении 1. Тестирование осуществлялось на данных большого объёма ($N = 10^5; 10^4; 10^3$). Значения случайных величин равномерно распределялись на интервале $[1, 4]$. На рисунке 12 представлен фрагмент кода программы.

```
begin
r:=0;
for i:= 1 to 4 do
r := r +  Random;
r4 := r;
end;
```

Рисунок 12 — Распределение случайных величин,
код программы

Функция плотности вероятности в случае 4 равномерно распределённых величин определялась формулой (16).

В ходе эксперимента менялся объём данных $N$, значения ширины шага $h$. Подробную информацию об оптимальных параметрах см. в пункте 3.4.

$$p(x) = \begin{cases} \frac{1}{6}x^3, \text{если } 0 \le x \le 1; \\ -\frac{1}{2}x^3 + 2x^2 - 2x + \frac{2}{3}, \text{если } 1 \le \text{ч} \le 2; \\ \frac{1}{2}x^3 - 4x^2 + 10x - \frac{22}{3}, \text{если } 2 \le x \le 3; \\ -\frac{1}{6}x^3 + 2x^2 - 8x + \frac{32}{3}, \text{если } 3 \le x \le 4. \end{cases} \quad (16)$$

Обязательно необходимо задать точную функцию плотности вероятности и графически изобразить её. Это позволит сравнивать результаты, полученные экспериментальным путём с истинным значением функции (рисунок 13).

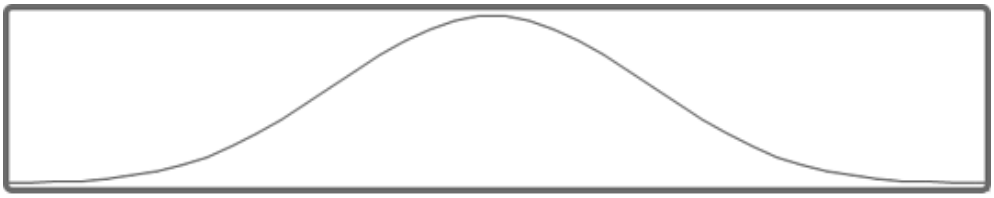

Рисунок 13 — Точная функция плотности вероятности

При построении гистограмм необходимо учитывать две важные величины — ширину шага $h$ и начальную точку отсчёта $x_0$. Иногда приходится сталкиваться с проблемой выбора шага $h$ и начальной точки $x_0$. Необходимо выбрать такие пары $(h, x_0)$, чтобы построенная гистограмма давала наилучшее приближение в некоторой норме, чаще всего в норме $L_2$.

Начальная точка $x_0$ каждой смещённой гистограммы задавалась следующим образом:

$$x_0 = -0.5 * h + k * \frac{h}{10}, \quad (17)$$

где $h$ – величина шага;

$k$ – порядковый номер гистограмм.

На данном этапе работы величина шага задавалась как $h = \frac{4}{n}, n = 40$. Результат построения точек 40 смещенных гистограмм представлен на рисунке 14.

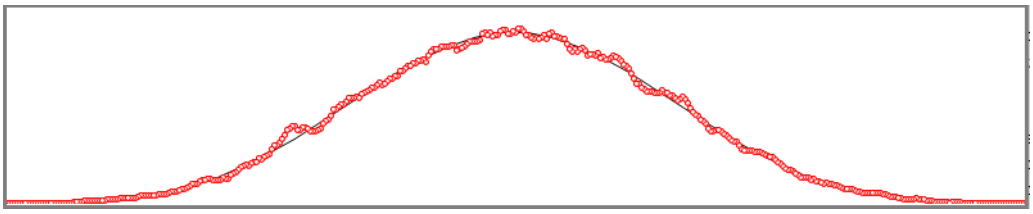

Рисунок 14 — Результат смещения 40 гистограмм с шагом $h = \frac{4}{n}$

При каждом запуске алгоритма генерируется новый набор случайных чисел. Следовательно, результат зависит от полученной случайной выборки. Он может быть как лучше, так и хуже. В качестве примера на рисунке 15 показаны результаты работы алгоритма с теми же параметрами, но уже при следующем запуске алгоритма. В данном случае отклонение от точной функции плотности вероятности увеличилось.

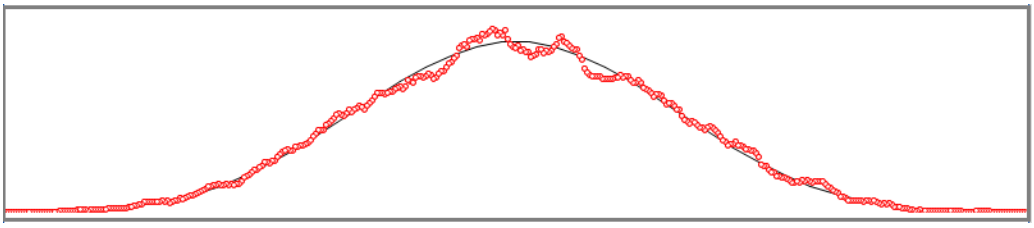

Рисунок 15 – Результат смещения 40 гистограмм с шагом $h = \frac{4}{n}$ (второй запуск)

Изменив шаг $h = 0{,}2$ видим, что алгоритм даёт более точную оценку (рисунок 16). Выводы об оптимальных параметрах см. в пункте 3.4.

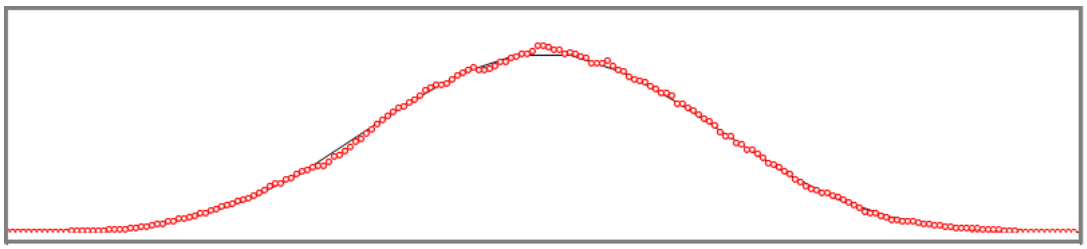

Рисунок 16 — Результат смещения 40 гистограмм с шагом $h = 0{,}2$

Для дальнейшей работы с алгоритмом и обработки полученных данных воспользуемся средой разработки *MATLAB*.

### 3.2 Использование возможностей среды MATLAB для аппроксимации

Чтобы приступить к сглаживанию полученных ранее данных (см. пункт 3.1), необходимо импортировать данные *save.txt* в среду *MATLAB*. Данные из файла считываются (значения $X$, экспериментальные значения $Y$, точные значения функции $Z$) в редакторе *Editor* среды следующим образом. Пример показан на рисунке 17.

```
clear all; //очищаем рабочую область
fileID = fopen ('save_exp.txt'); //открываем файл
formatSpec = '%f %f\n'; //формат данных
sizeA = [3 Inf]; //размер данных
B = fscanf(fileID,formatSpec,sizeA); //считывание данных
x = B(1,:);
y = B(2,:);
z = B(3,:);
```

Рисунок 17 — Пример работы в редакторе *Editor*

Сохраняем файл с командами в формате *read.m*. В дальнейшем этот файл не будет изменяться, а будет лишь вызываться при необходимости (в случае изменения данных в файле *save.txt*).

В ходе эксперимента было проведено свыше 20 испытаний при каждом изменении параметров. Результаты всех испытаний в большей или меньшей степени имеют отклонения от точной функции распределения. Выводы обо всех проделанных испытаниях будут приведены в последнем из разделов этой главы. В качестве наглядного примера рассмотрен следующий. На рисунке 18 представлены экспериментальные данные и точная функция распределения при $n = 10\,000, h = 0{,}1$.

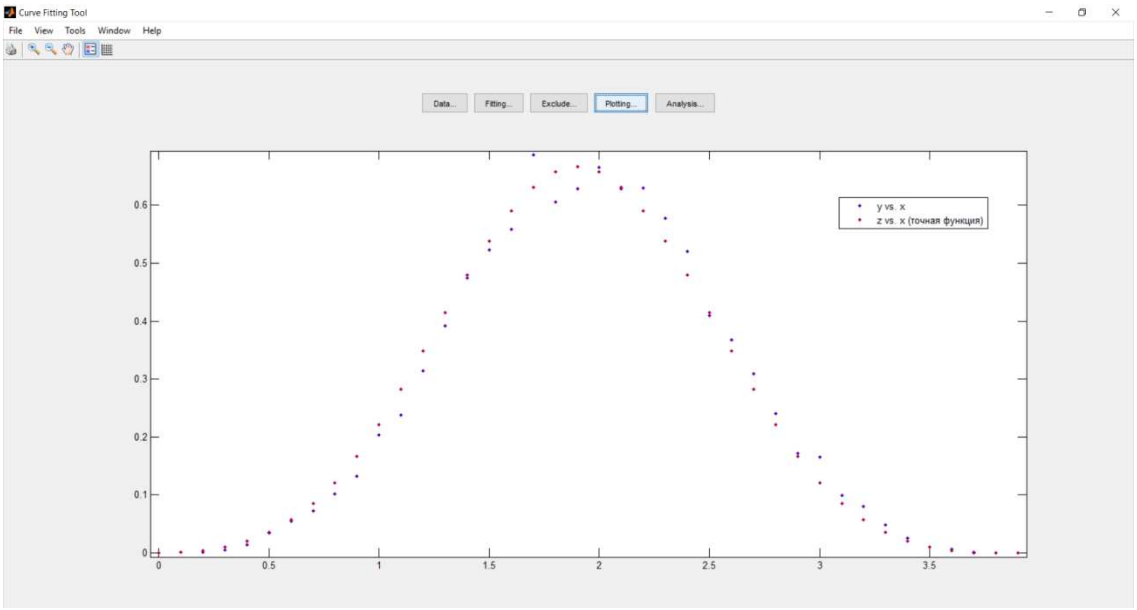


Рисунок 18 — Экспериментальные данные и точная функция распределения ($n = 10\,000, h = 0{,}1$)

После построения случайных точек, становятся актуальными проблемы сглаживания и фильтрации данных. Необходимо построить для исходных экспериментальных точек $y(x_i)$ зависимость $f(\mathrm{x})$, которая должна приближать эти точки. При этом нужно учитывать, что данные $(x_i, y_i)$ получены с некоторой погрешностью, выражающей шумовую компоненту измерений. А

функция $f(\mathrm{x})$ с помощью того или иного алгоритма сглаживания уменьшает погрешность, присутствующую в данных $(x_i, y_i)$.

Сглаживание путём построения регрессии данных — это частный случай фильтрации. Регрессия — наиболее простой и эффективный способ сглаживания. Но регрессия часто уничтожает информативную составляющую данных, оставляя лишь заданную исследователем зависимость. Она эффективна в тех случаях, когда известен закон распределения данных $(x_i, y_i)$.

Для сокращения временных ресурсов при исследовании подходящих методов сглаживания полученных данных, использовалось приложение *cftool* пакета *MATLAB*. Приложение вызывается одноимённой командой из рабочего окна *Command Window*. Приложение позволяет работать с импортированными данными, создавать правила-исключения, выбирать модели сглаживания или создавать собственные, подбирать параметры, границы, начальные приближения, анализировать графики и т. д.

Предположим, что полученные ранее данные зашумлены, следовательно, имеет смысл произвести их сглаживание. Однако стоит учитывать тот факт, что сглаживание обычно используется для получения информации о возможном выборе типа параметрической модели.

Для осуществления сглаживания берутся несколько подряд идущих точек. Сглаживание уничтожает стандартное предположение регрессионного анализа о том, что распределение ошибки в исходных данных подчиняется нормальному закону распределения. Если построена достаточно хорошая модель, то остатки (разность значений данных и приближения) также должны подчиняться нормальному закону. Поэтому сглаживание следует использовать как инструмент для получения первоначального предположения о возможной параметрической модели в случае зашумленных данных, а строить модель следует для не сглаженных исходных данных.

Для осуществления сглаживания воспользуемся уже имеющимися возможностями приложения *cftool* и вкладкой *smooth*. Осуществляется подбор параметров подходящей параметрической модели, входящей в набор

стандартных параметрических моделей *Curve Fitting Toolbox (Moving Average, Lowess (linear fit), Loess (quadratic fit), Savitzky-Golay, Robust Lowess (linear fit), Robust Loess (quadratic fit)).* Данные модели применяются в случае, если данные не сильно зашумлены, т.е. не требуется создавать правила исключения или использовать процедуры сглаживания для получения представления о возможной параметрической модели. В ходе исследования были рассмотрены все методы апплоксимации. Далее разберём и рассмотрим подробнее несколько методов, дающих наилучшие результаты.

### 3.2.1 Метод скользящего среднего

Метод скользящего среднего — один из методов экстраполяции (метод научного исследования, основанного на распространении прошлых и настоящих закономерностей и информации об объекте). Значения функции в каждой точке определяются как среднее значение исходной функции за предыдущий период.

В строке *Span* задаём число соседних точек, по которым будет осуществляться сглаживание (для взвешенной локальной регрессии можно задавать меньшее единицы число, оно воспринимается как процент от общего числа точек, используемых для сглаживания).

В методе скользящего среднего исходные данные $y_i$ сглаживаются по следующему правилу:

$$ys_i = \frac{1}{2N+1}\sum_{k=-N}^{N} y(i+k), \tag{18}$$

где $(2N + 1)$ — число точек, выбираемых для сглаживания.

Слева и справа от текущей точки выбирается по $N$ точек (число точек, участвующих в сглаживании, должно быть нечётным). Данные, расположенные в точках, близких к границам отрезка, не сглаживаются, т.к. не хватает точек

справа или слева от текущей, в которой в данный момент производится сглаживание.

Результат работы данного метода сглаживания применительно к имеющимся данным представлен на рисунке 19. Метод показал достаточно неплохие результаты, однако в верхних точках произошёл некий «срез» графика, за счёт чего произошла потеря части данных.

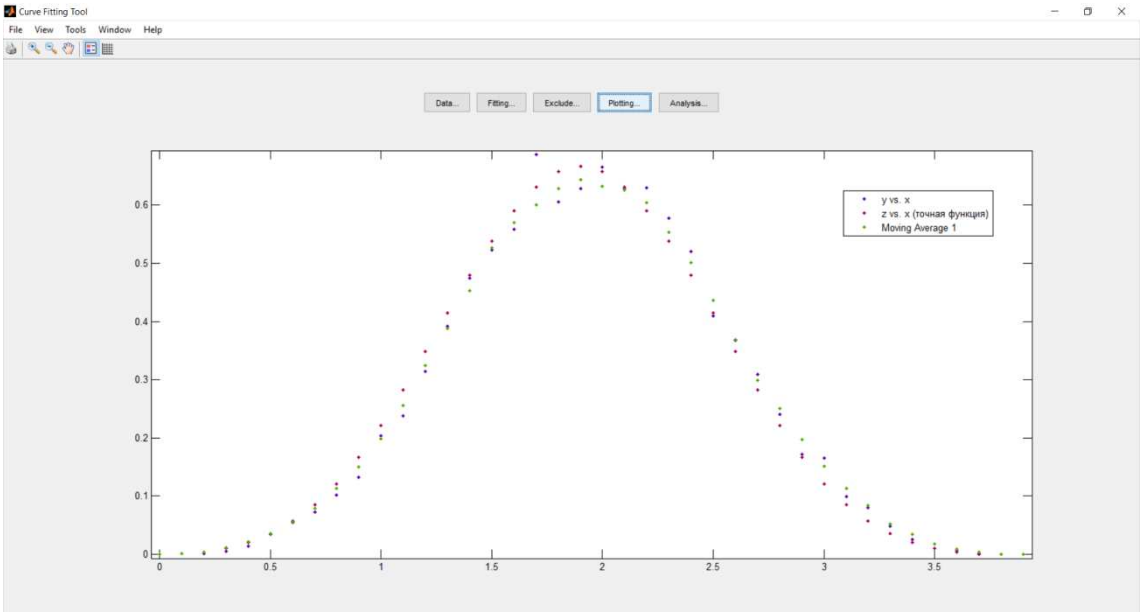


Рисунок 19 — Метод скользящего среднего

### 3.2.2 Фильтр Савицкого-Голея

Фильтр Савицкого-Голея — сглаживающий фильтр, аппроксимирующий отдельные входные данные по критерию минимума квадратической ошибки.

Число соседних данных (задаваемое в строке ввода *Span*) приложения *cftool* для сглаживания в точке должно быть нечётное, а степень полинома (задаваемая в строке ввода *Degree*) меньше, чем число соседних данных.

Этот способ фильтрации хорошо подходит для зашумленных сигналов, в которых при сглаживании требуется сохранить высокие частоты. При повышении степени полинома, применяемого в локальной регрессии, лучше воспроизводятся узкие и высокие пики.

Результаты работы аппроксимации фильтром Савицкого-Голея представлены на рисунке 20, где (*yvs.x)* — это экспериментальные данные; (*zvs.x* ) — точная функция распределения.

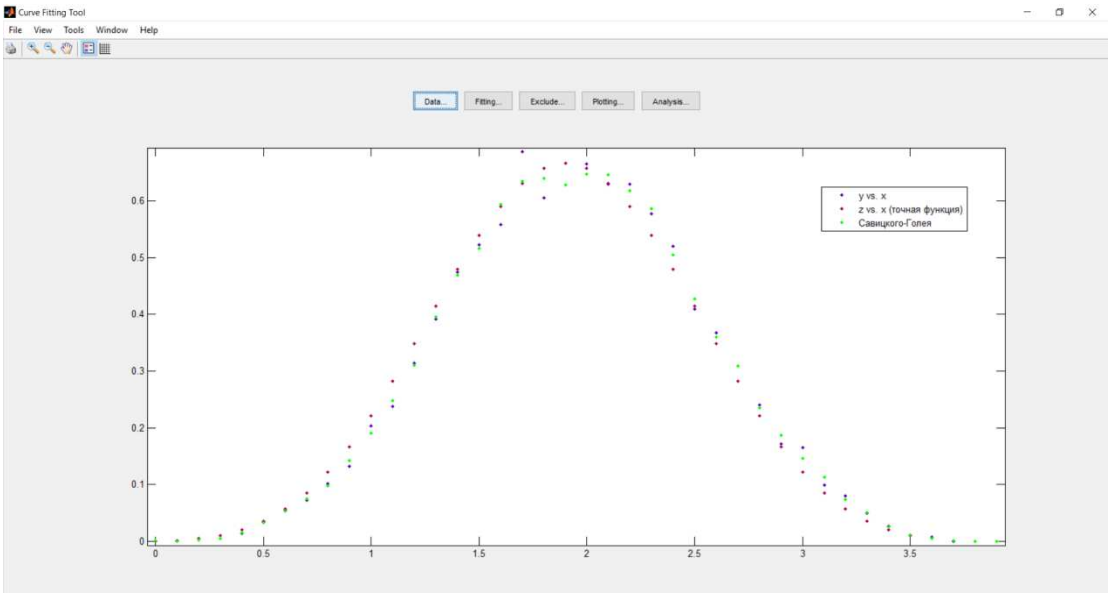


Рисунок 20 — Применение фильтра Савицкого-Голея

Кроме вышеперечисленных и рассмотренных методов сглаживания, приложение *cftool* пакета *MATLAB* позволяет работать со следующими методами:

1) экспоненциальные функции;
2) ряды Фурье и степенные ряды;
3) гауссианы;
4) полиномы высоких степеней (до девятой степени);
5) функции Вейбула;
6) сглаживающие сплайны;
7) рациональные функции;
8) суммы синусов и др.

Рассмотрим эти методы применительно к нашей задаче. Работа со всеми методами, задание нужных параметров, условий, выбор данных, получение результатов — всё это осуществляется в рабочем окне *Fitting* (рисунок 21).

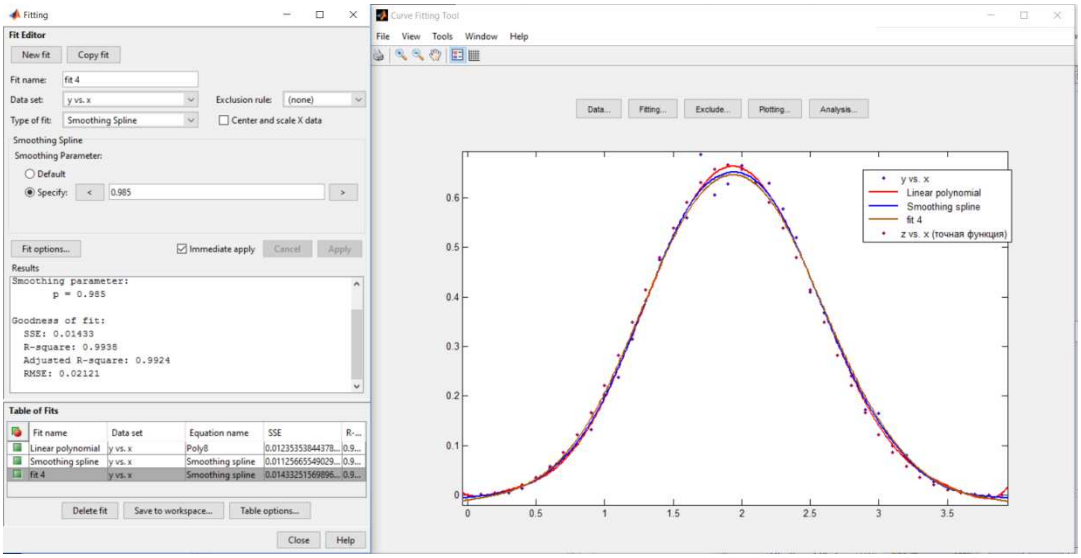


Рисунок 21 — Рабочее окно Fitting

Эти стандартные регрессионные модели включают в себя оптимизированные параметры решателя и начальные условия для повышения качества приближения. Кроме того, здесь можно задавать уравнения *(Custom Equation)* для построения собственных регрессионных моделей.

Для стандартных и пользовательских параметрических моделей предлагаются по умолчанию (там, где это нужно, в зависимости от типа параметрической модели) следующие границы параметров и начальные приближения, представленные в таблице 1.

Таблица 1 — Границы параметров и начальные приближения

| Тип модели | Начальное приближение | Ограничения на параметры |
|---|---|---|
| Линейная модель, определенная пользователем | Не требуется | Нет |
| Нелинейная модель, определенная пользователем | Случайное для каждого параметра из интервала [0,1] | Нет |

Продолжение таблицы 1

| Тип модели | Начальное приближение | Ограничения на параметры |
|---|---|---|
| Отрезки ряда Фурье $a_0 + \sum_{k=1}^{n} a_k \cos(nwx) + b_x \sin(nwx)$, $n \leq 8$ | Вычисляется по начальным данным по эвристическому алгоритму | Нет |
| Гауссова $\sum_{k=1}^{n} a_k e^{-(\frac{x-b_k}{c_k})^2}$, n≤ 8 | Вычисляется по начальным данным по эвристическому алгоритму | |
| Полиномиальная $\sum_{k=1}^{n+1} a_k x^{n+1-k}$, $1 \leq n \leq 9$ | Не требуется | Нет |
| Степенные $ax^b$, $ax^b + c$ | Вычисляется по начальным данным по эвристическому алгоритму | Нет |
| Дробно-рациональная $\frac{\sum_{k=1}^{n+1} p_k x^{n+1-k}}{x^m + \sum_{k=1}^{m} q_k x^{n-k}}$ | Случайное для каждого параметра из интервала [0,1] | Нет |
| Сумма синусов $\sum_{k=1}^{n} a_k \sin(b_k x + c_k)$, $n \leq 8$ | Вычисляется по начальным данным по эвристическому алгоритму | $b_k > 0$ |
| Вейбула $abx^{b-1}e^{-ax^b}$ | Случайное для каждого параметра из интервала [0,1] | $a > 0, b > 0$ |

В ходе многочисленных экспериментальных исследований, сравнения полученных результатов как визуальных, так и численных, было выявлено преимущество полиномиального метода и метода сглаживания сплайнами. Рассмотрим данные методы и полученные результаты подробнее.

### 3.2.3 Полиномиальный подход

Линейная регрессия — простой и наиболее часто используемый вид регрессии. Приближение данных $(x_i, y_i)$ осуществляется линейной функцией $y(x) = b + ax$. Коэффициенты $a$ и $b$ вычисляются из условия минимизации суммы квадратов ошибок $|b + a * (x_i - y_i)|^2$. В общем случае полиномиальная регрессия означает приближение данных $(x_i, y_i)$ полиномом -й степени:

$$A(x) = a + b * x + c * x^2 + \cdots + h * x^k \quad (19)$$

На практике обычно применяют $k < 5$. Для построения регрессии полиномом -й степени необходимо, чтобы было минимум $(k + 1)$ точек данных.

В рабочей среде уже присутствуют следующие полиномиальные подходы: линейный, квадратичный, кубический, полиномы до 9 степени. Визуально полином 8 степени даёт хорошее приближение (рисунок 22).

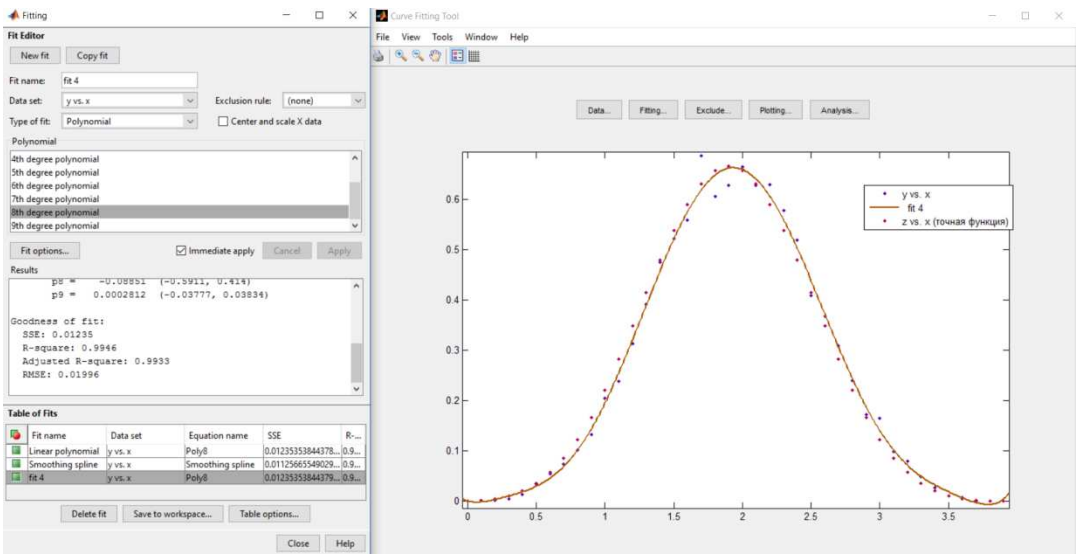


Рисунок 22 — Сглаживание функции полиномом 8-й степени

Особое внимание обращаем на критерий пригодности приближения $SSE$. Критерий вычисляется по формуле:

$$SSE = \sum_{k=1}^{n} w_k (y_k - \hat{y}_k)^2, \quad (14)$$

где $w_k$ — веса (если не заданы, то считаются равными нулю);

$y_k$ — данные в $x_k$;

$k$ — значения параметрической модели в $x^k$.

В приведённом примере $SSE = 0{,}01235$. Близость $SSE$ к нулю говорит о хорошем качестве приближения данных параметрической моделью.

### 3.2.4 Сглаживающий сплайн

Хотя сглаживающий сплайн и относят к непараметрическим моделям, он содержит задаваемый пользователем параметр $p$. Сглаживающий сплайн определяется как сплайн, который минимизирует следующий функционал, зависящий так же и от некоторого параметра $p$:

$$I(s; p) = p \sum_{k=1}^{n} w_k (y_k - s(x_k))^2 + (1 - p) \int_{x_1}^{x_n} \left(\frac{d^2 s(x)}{dx^2}\right)^2 dx, \quad (20)$$

где $(x_k, y_k)_k = 1,2, \dots, n$ — приближаемые данные;

$w_k$ — веса данных (если они не были заданы, то принимаются равными единице);

$p$ — сглаживающий параметр, изменяющийся от 0 до 1, который определяет кривизну получающегося сплайна.

Если задавать значения сглаживающего параметра близкие к нулю, то сглаживающий сплайн будет похож на прямую, приближающую данные в смысле наименьших квадратов, поскольку основным в минимизируемом функционале станет второе слагаемое $(1 - p) \int_{x_1}^{x_n} \left(\frac{d^2 s(x)}{dx^2}\right)^2 dx$, которое как раз и отвечает за гладкость. Его минимизация соответствует построению сплайна с наименьшим значением второй производной (ноль, для полинома первого порядка). Если значение сглаживающего параметра близко к единице, то

основным в минимизируемом функционале станет первое слагаемое $p\sum_{k=1}^{n} w_k(y_k - s(x_k))^2$, которое отвечает за прохождение сплайна через заданные точки. При $p = 1$ сглаживающий сплайн превращается в обыкновенный кубический сплайн. На практике при применении сглаживающего сплайна часто значение сглаживающего параметра выбирают $p \approx \frac{1}{1+\frac{h^3}{6}}$, где $h$ — среднее расстояние между точками, в которых определены приближаемые данные.

В нашем случае сглаживающий параметр $p = 0{,}994$, $SSE = 0{,}01125$. Как показали полученные экспериментальные данные, метод сглаживания сплайнами даёт наилучшие результаты как графически, так и согласно критерию пригодности (рисунок 23).

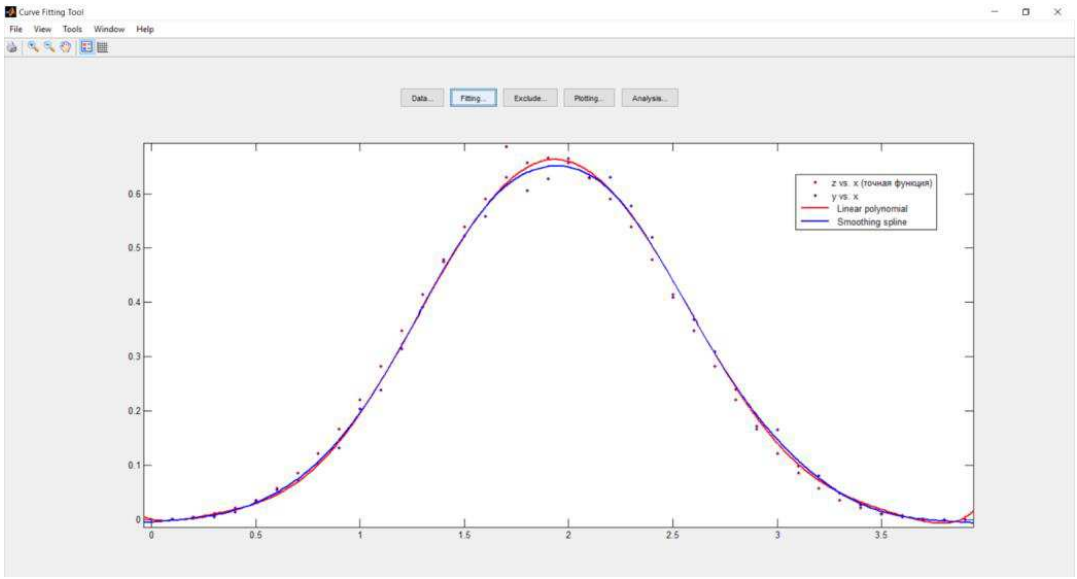


Рисунок 23 — Сглаживающий сплайн

Проведём последнее испытание, изменив размер исходной информации $N = 1\,000$. Все остальные параметры останутся неизменными. Для сглаживания будем использовать метод сплайнов (рисунок 24).

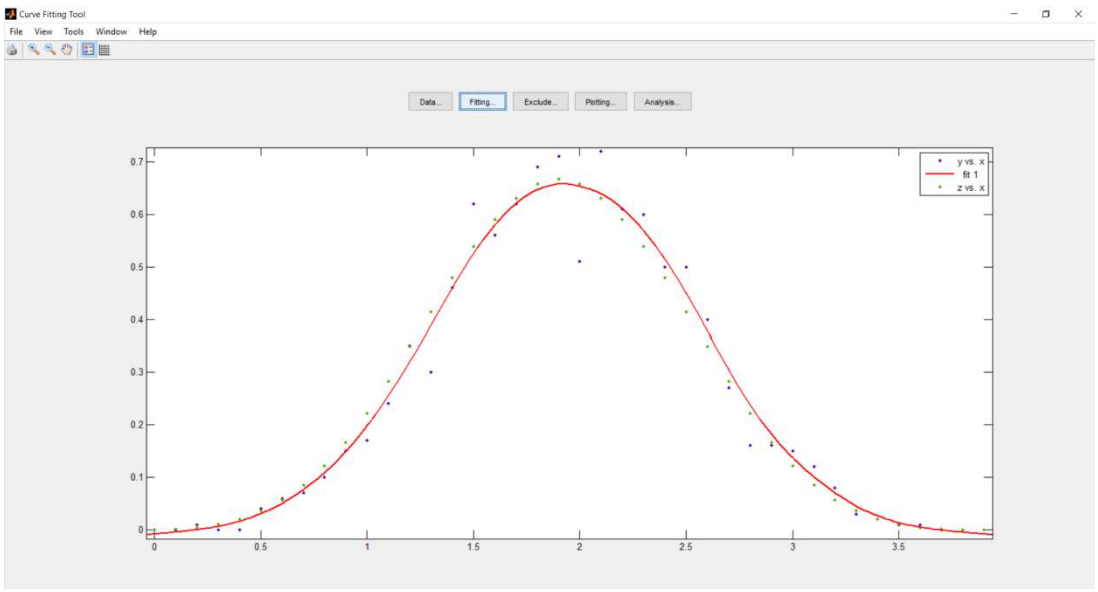


Рисунок 24 – Восстановление плотности вероятности при $N = 1\,000$

Таким образом, использование подхода основанного на методе сдвинутых гистограмм и сглаживания сплайнами позволяет восстанавливать функцию плотности вероятности с точностью не хуже ядерных оценок.

### 3.3 Оптимальные параметры для адекватной работы метода

В ходе проведения экспериментов на модельных данных проводились исследования зависимости качества восстановленной плотности вероятности от параметров входных данных. В качестве исходной информации использовалась информация, полученная в пункте 3.1.

Качество восстановленной плотности вероятности напрямую зависит от выбора величины параметра $h$. Разница между ними как раз в величине шага. Чем меньше шаг $h$, тем больше дисперсия оценки плотности распределения будет стремиться к бесконечности. И наоборот, чем больше шаг $h$, тем больше вероятность ошибки результатов. Его значение оказывает большое влияние на вид оценок плотностей распределения и на их точность. В нашем случае $h = 0{,}1$ оптимальное значение параметра. В качестве примера продублируем результаты экспериментов (рисунок 25 и рисунок 26).

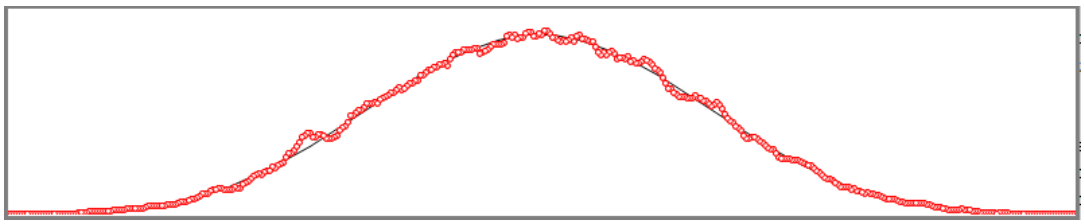

Рисунок 25 — Результат смещения 40 гистограмм с шагом $h = 0{,}1$ и $N = 10\,000$

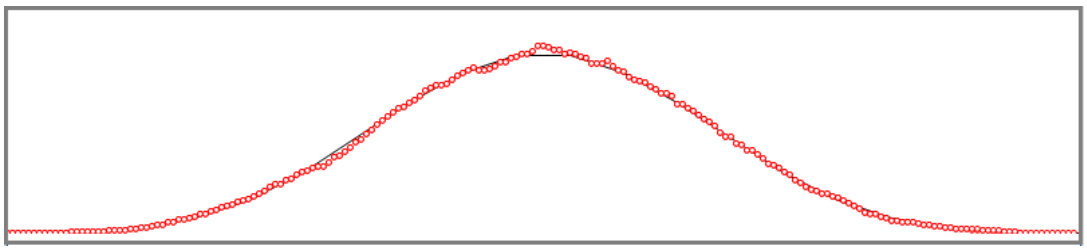

Рисунок 26 — Результат смещения 40 гистограмм с шагом $h = 0{,}2$ и $N = 10\,000$

Возникшую в ходе исследования проблему повышения качества восстановления плотности вероятности можно повысить за счёт метода, предложенного в работе [9]. Для повышения точности восстановления функции плотности вероятности $z$ авторы используют комбинацию оценок с параметрами $h$ и $2h$.

Кроме того, важную роль играет объём полученной информации, определяющей достоверность последующих результатов. Чем больше объём выборки, тем выше точность восстановления плотности вероятности. В нашей работе в качестве оптимального объёма выборки рассматривалась $N = 10\,000$.

В качестве наиболее эффективного метода сглаживания были выявлены и доказаны преимущества сглаживания методом со сглаживающим параметром $p = 0{,}994$.

При правильном подходе и использовании оптимальных параметров модернизированный метод осреднения смещённых гистограмм можно рассматривать и для данных меньшего объёма. Так, например, в одном из экспериментов была рассмотрена выборка $N = 1\,000$. (рисунок 24). При использовании разработанных ранее методов и подходов результаты получились сравнимы с методом ядерных оценок.

Однако не стоит забывать и про погрешности, возникающие в ходе экспериментов. В работе [9] предлагается один из методов повышения точности оценки функции плотности вероятности. Данный подход основан на применении правила Рунге для вычисления второй производной оценки функции плотности вероятности.

Оценки второй производной позволяют оценить математическое ожидание в $L_2$ норме, чтобы определить погрешность функции плотности вероятности. Знание этих оценок даёт возможность определить оптимальное значение параметра $h$.

# ЗАКЛЮЧЕНИЕ

Тема научно-исследовательской работы являлась «Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа». В ходе практического и теоретического исследования был реализован метод осреднения смещённых гистограмм, позволяющий извлекать максимум полезной информацию из определённого набора эмпирических данных. Данный результат позволит повысить эффективность и качество оценки состояния технических систем, изделий и исследований в условиях неопределённости.

Преимуществом данного реализованного метода являются быстрота реализации, минимальные вычислительные затраты, графическое представление обрабатываемых данных, хорошее качество приближения данных в сравнении с другими существующими методами оценки.

На определённых этапах исследования были решены следующие основные задачи:

1. В ходе теоретического исследования проведён анализ данной предметной области. Оказалось, что тема восстановления плотности вероятности весьма актуальна, поскольку имеет место быть во многих сферах деятельности.

2. Основной проблемой, с которой пришлось столкнуть, стала проблема отсутствия единого универсального метода для обработки данных в условиях неопределённости (точнее — восстановление плотности вероятности), дающего эффективные результаты.

3. Следующим этапом стало изучение и исследование уже существующих методов восстановления плотности вероятности. Каждый из рассмотренных методов имел как плюсы, так и минусы: точность получаемых результатов, сложность реализации метода, наличие необходимых исходных данных – при исследовании учитывались все эти факторы и не только.

4. Оценив достоинства и недостатки всех методов, в качестве дальнейшей работы для восстановления плотности вероятности был выбран метод осреднения смещённых гистограмм [29].

5. Изучив данный метод и модернизировав его, в ходе проведения экспериментов было выявлено, что использование середин гистограмм дают оценку такую же, как и метод ядерных оценок. Однако метод осреднённых смещённых гистограмм проще в реализации, следовательно, может исключать возникновение погрешностей.

6. Далее исследовались методы сглаживания полученных результатов. При проведении численного эксперимента оказалось, что наилучшие результаты при использовании данного метода восстановления плотности вероятности даёт метод сглаживающих сплайнов. Таким образом, результатом работы стало определение оптимальных методов восстановления плотности вероятности, методов сглаживания полученных данных и выявлены оптимальные параметры при работе алгоритма.

Метод, полученный в результате выполнения научно исследовательской работы, имеет место быть во многих областях, где приходится сталкиваться с обработкой эмпирических данных. В дальнейшем проводя исследования в данной области можно решить проблему обработки данных небольшого объёма в условиях неопределённости и повысить точность получаемых в ходе экспериментов результатов.

По результатам исследования было опубликовано 4 статьи:

1. Подходы к обработке экспериментальных данных в условиях ограниченной информации // Журнал «Научные исследования и разработки молодых учёных», 2015г. Стр. 136 – 140.

2. Методы восстановления плотности вероятности в условиях ограниченных объёмов данных // Международная научная конференция «Молодежь и наука: проспект Свободный». 2016. Стр. 127 – 129.

3. Модернизированный метод смещенных гистограмм для восстановления плотности вероятности // Журнал «Решетнёвские чтения», 2016 г. Стр. 237 – 238.

4. Построение функции плотности вероятности на основе сглаживания смещенных гистограмм // Международная научная конференция «Молодежь и наука: проспект Свободный», 2017.

# СПИСОК СОКРАЩЕНИЙ

BIN1 — наименование алгоритма вычисления первого столбца гистограммы;

AMISE (от англ. «asymptotic mean integrated squared error») — асимптотическая средняя интегральная ошибка;

ASH — average shifted histogram;

ОСГ — осреднённая смещённая гистограмма;

SSE (от англ. «sum of squares due to error») — сумма квадратов ошибок;

ISB (от англ. «integrated squared bias») — интегрированное квадратное смещение, ИКС.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1) Акимов С. С., Оценка методов восстановления законов распределения вероятности и обоснование предпочтения на основании некоторых свойств эмпирического массива данных // Актуальные вопросы современной науки, вып.№ 31 / 2014. С.139.

2) Антонов А. В., Зюляева Н. Г., Чепурко В. А., Исследование метода ядерной оценки плотности распределения // Надёжность. – Москва: Издательский дом «Технология», 2005. С. 3.

3) Бардасов С. А., Гистограммы. Критерии оптимальности // Монография. Тюмень, 2014. – 95 с.

4) Вапник В. Н., Восстановление зависимостей по эмпирическим данным // Москва, 1979 г. – 447 с.

5) Водолазская И. В., Введение в систему Matlab // Астрахань, 2004. – 48 с.

6) Вовк А. А., Основы общей теории статистики. – Москва: Маршрут, 2006. – 240 с.

7) Воронцов К. В., Статистические алгоритмы классификации // 2008. – 39 с.

8) Глаголев М. В., Сабреков А. Ф., О восстановлении плотности вероятности методом гистограмм в почвоведении и экологии // 1998. С.48–57.

9) Деврой Л., Дьёрфи Л., Непараметрическое оценивание плотности. L1-подход. // Пер. с англ. – Москва: Мир. 1988. 408с.

10) Добронец Б. С., Попова О.А., Численный вероятностный анализ для исследования систем в условиях неопределенности // Красноярск, СФУ, 2014. – 169 с.

11) Мартынов Н. Н., Иванов А. А., Matlab: вычисления, визуализация, программирование // Москва, 2000. – 337 с.

12) Попова О. А., Гистограммы второго порядка для численного моделирования в задачах с информационной неопределённостью // Известия

Южного федерального университета. Технические науки. 2014. № 6 (155). С. 6-14.

13) Попова О. А., Технология извлечения и визуализации знаний на основе численного вероятностного анализа неопределенных данных // Информатизация и связь. 2013. № 2. С. 63–66.

14) Поршнев С. В., Копосов А. С., Использование аппроксимации Розенблатта-Парзена для восстановления функции распределения непрерывной случайной величины с ограниченным одномодальным законом распределения // Политематический сетевой электронный научный журнал, выпуск № 92, 2013.

15) Рубан А. И., Кузнецов А. В., Методы обработки экспериментальных данных // Красноярск, 2008.

16) Тарасенко Ф. П. Непараметрическая статистика // Томск, 1976. – 293 с.

17) Ужга-Ребров О. И., Управление неопределенностями // 2004. 293 с.

18) Фастовец Н. О., Попов М. А., «Математическая статистика примеры, задачи и типовые задания // Москва 2012.

19) Фастовец Н. О., Попов М. А., «Математическая статистика примеры, задачи и типовые задания» Тюмень: Изд-во «ТюмГНГУ», 2007. – 202 с.

20) Хардле В., Прикладная непараметрическая регрессия: пер. С англ. М., Мир, 1993. – 349 с.

21) Dobronets B. S., Popova O. A., Improving the accuracy of the probability density function estimation // Journal of Siberian Federal University. Mathematics & Physics, 2016.

22) Dobronets B., Popova O., Numerical probabilistic approach for data nonparametric analysis // Applied methods of statistical analysis. nonparametric approach proceedings of the international workshop. 2015. С. 376–384.

23) Dobronets B., Popova O., Numerical Probabilistic Approach for optimization Problems. Scientific Computing, Computer Arithmetic, and Validated

Numerics. Lecture Notes in Computer Science 9553. Springer International Publishing, Cham, 43 – 53 (2016).

24) Parsen E., On estimation of a probability density function and mode //AMS. 1962. 33. Pp. 1065–1076.

25) Pearson K., Contributions to the Mathematical Theory of Evolution, II: skew variation in homogeneous material // Philosophical Transactions of the Royal Society of London (A). 1895. Vol. 186. Pp. 343–414.

26) Popova O., Information approach to a posteriori error estimates of numerical modeling. Informatization and Communication. 2, 29–32 (2016).

27) Popova O. A., Optimization problems with random data // Журнал Сибирского федерального университета. Серия: Математика и физика. 2013. Т. 6. № 4. С. 506–515

28) Rosenblatt M., Remarks on some nonparametric estimates of a density functions // AMS. 1956. 27 Pp. 832–837

29) Scott R. W., Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons. 2015. 381 p.

30) Uglev V. A., Popova O. A., Dobronets B. S., The accuracy calculation control of reliability indices for equipment responsible appointment // 2015 International Siberian Conference on Control and Communications, SIBCON 2015 – Proceedings 2015.

**Электронные ресурсы**

1) Сайт разработчика пакета прикладных программ Matlab [Электронный ресурс] – Режим доступа: http://www.mathworks.com, свободный.

2) Справочник по работе в среде MATLAB MathWorks [Электронный ресурс]. Режим доступа: http://www.mathworks.com, свободный.

## ПРИЛОЖЕНИЕ А

### Листинг программы алгоритма

```
program AverShift_points;
uses GraphABC;
Const n=40;
     nr=10000;
type his =  array[0..n+1] of real;

Var fout : text;
   hk,ts, tscor, tspr2, sfpr2 :real;
   i,ns,nmat : Longint;

function r4:real;
Var r, t, r1: real;
   i,j : integer;

begin
r:=0;
for i:= 1 to 4 do
r := r +  Random;
r4 := r;
//setpixel(i,round(r),clblack);
end;

function pfr4(x:real):real;
Var r, t, x2,x3 : real;
   i,j : integer;
begin
```

```
x2:=x*x;
x3:=x2*x;
//плотность вероятность
if x < 0 then r:=0;
if (x >=0) and (x < 1)  then r:= x3/6;
if (x >=1) and (x < 2)  then r:= -x3/2 +x2*2 - 2*x +2/3;
if (x >=2) and (x < 3)  then r:= x3/2  -4*x2 + 10*x -22/3;
if (x >=3) and (x < 4)  then r:= -x3/6 +x2*2 - 8*x +32/3;
if x >= 4 then r:=0;
pfr4:=r;
end;

function pfr4_pr2(x:real):real;
Var r : real;
  // вторая производная
begin
if x < 0 then r:=0;
if (x >=0) and (x < 1)  then r:= x;
if (x >=1) and (x < 2)  then r:= 4-3*x;
if (x >=2) and (x < 3)  then r:= 3*x-8;
if (x >=3) and (x < 4)  then r:= 4-x;
if x >= 4 then r:=0;
pfr4_pr2:=r;
end;

Procedure tt;
Var i,k,j, kpt, n1,n2 :Longint;
    t, x, r,r1,r2,h, x0,fpr2, s, scor, spr2 : real;
    x1, x2: his;
    rd : array[1..nr] of real;
```

```
   t1:integer;

begin
h:=4/n;
kpt:=0;
for i:=1 to n+1 do
begin
x1[i]:=0;
end;
x0:=0;
for i:= 1 to nr do
rd[i]:= r4;
for i:= 1 to nr do
begin
j:=Round( Int( rd[i]/h ) )+1;
x1[j]:=x1[j]+1;
end;
for i:=1 to n+1 do
begin
x1[i]:=x1[i]/(h*nr);
end;
for i:=1 to n do
begin
t:= (i-1/2)*h;
//writeln(t : 6:3, x1[i]:8:4, pfr4(t):8:4);
end;

  window.title :='Гистограмма';
  window.SetSize(1000,200);
  Pen.Color:=clblack;
```

```
 // Line(0,0,0,200);//ось y
  //Line (0,200,800,200);//ось x
  var tt1,xx1:integer;
  //отрисовываем точную линию
  putpixel(0,0,clblack);//начальная точка, от которой начинаем рисовать линию точную
  for i:=1 to n do
  begin
  t:= (i-1/2)*h;
  tt1:=round(200*t);
  xx1:=round(200*pfr4(t));
  lineto(tt1,xx1);
  //line(0,200,800,200)
  end;
setpencolor(clred);
for k:=1 to 20 do
begin
h:=0.20;
x0:=-0.5*h+k*h/10;
for i:=1 to n+1 do
begin
x1[i]:=0;
end;
for i:= 1 to nr do
begin
j:=Round( Int( (rd[i]-x0)/h ) )+1;
x1[j]:=x1[j]+1;
end;
for i:=1 to n+1 do
begin
```

```
x1[i]:=x1[i]/(h*nr);
end;
for i:=1 to n do
begin
t:= (i-1/2)*h+x0;
tt1:=round(200*t);
xx1:=round(200*x1[i]);
//putpixel(tt1,xx1,clred);
circle(tt1,xx1,2);
end;
end;
end;
begin
Randomize;
Assign(fout,'ASPoint_007.tex'); rewrite(fout);
tt;
close(fout);
end.
```

# ПРИЛОЖЕНИЕ Б

## Плакаты презентации


Федеральное государственное автономное образовательное учреждение
Сибирский Федеральный Университет
Кафедра систем искусственного интеллекта

**Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа**

Магистр группы КИ15-02-1/1М А. А. Чевер
Научный руководитель профессор,
доктор физ.-мат. наук Б. С. Добронец

Рисунок 1 — Слайд 1

**Актуальность** 2

**Повышение качества состояния систем в следующих областях:**

- медицина,
- ядерная энергетика,
- космические исследования,
- гидроэнергетика и др.

**Решения принимаются в условиях неопределённости.**

**В случае отказа – высокий уровень риска.**

Рисунок 2 — Слайд 2

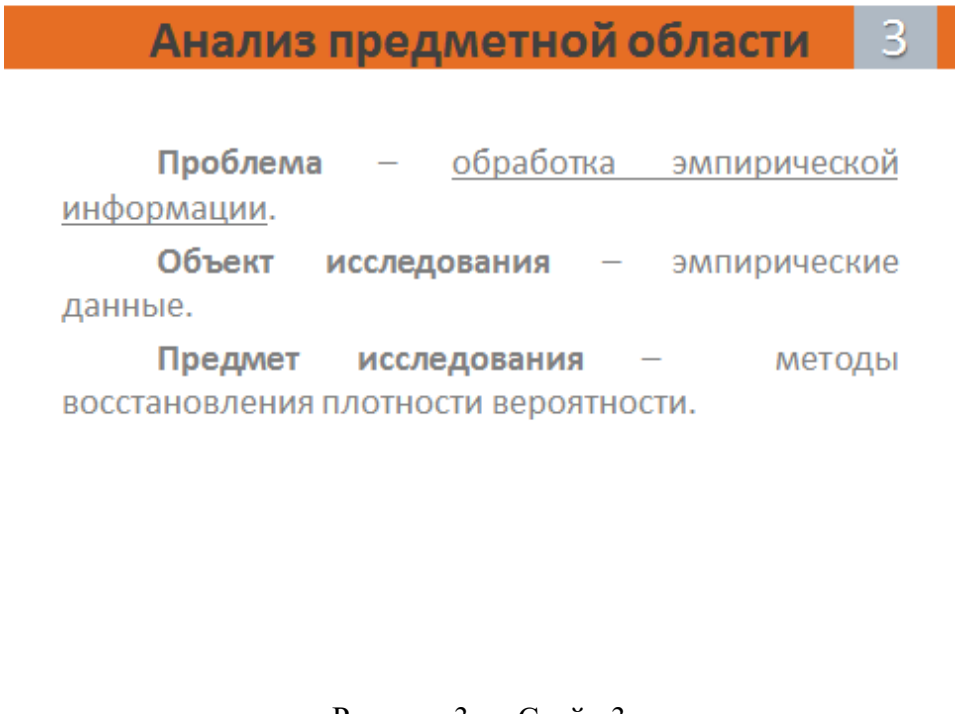


Рисунок 3 — Слайд 3

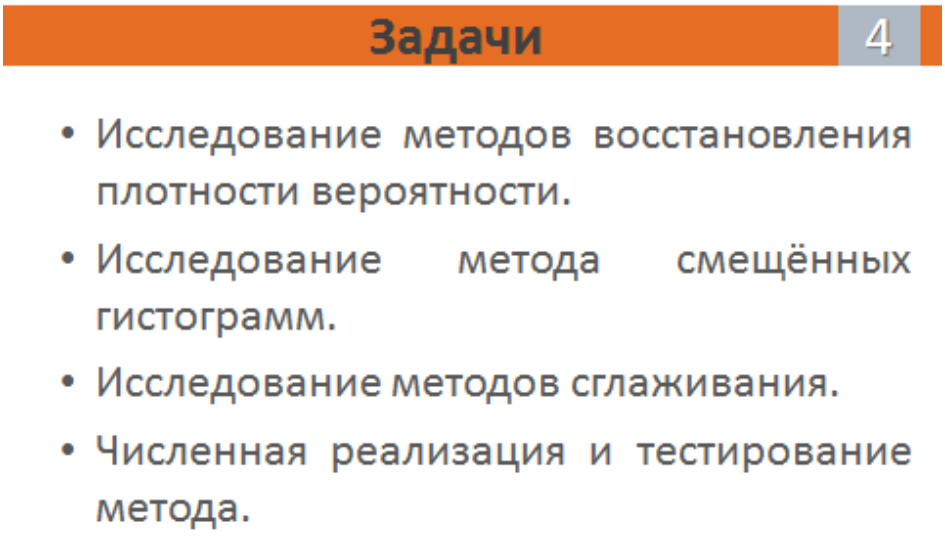


Рисунок 4 — Слайд 4

## Функция плотности вероятности 5

Случайная величина $X \xleftarrow{\text{характеристика}}$ функция плотности вероятности $f(x) \xleftarrow{\text{описание } X}$ восстановление плотности вероятности.

**Функция плотности вероятности** $f(x)$ – плотность, с которой распределяются значения случайной величины в определенной точке.

**Методы восстановления плотности вероятности:**

- *гистограммный подход;*
- метод ядерных оценок;
- частотный полигон.

Рисунок 5 — Слайд 5

## Частотный полигон 6

**Частотный полигон** – способов графического представления плотности вероятности случайной величины. Это ломаная, соединяющая точки $(x_i, n_i)$, где $i = 1,2,\dots,m$; $x_i$, – значения вариант; $n_i$ – соответствующие частоты.

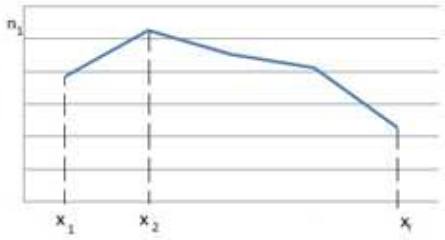


Рисунок 1 – Построение полигона частот

Рисунок 6 — Слайд 6

## Метод ядерных оценок 7

Процедура оценки плотности вероятности в одной точке:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right),\ h \to 0, nh \to \infty,$$

где $x$ – элемент выборки размера $n$, полученная в ходе наблюдения за объектом; $K$ – симметричное ядро; $h$ – диапазон, параметр сглаживания, влияющий на точность оценок.

Использовано Гауссова ядро $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$

Рисунок 7 — Слайд 7

## Гистограммный подход 8

**Гистограмма** – это кусочно-постоянная функция, определяющаяся сеткой $\{x_i | i = 0, \dots, N\}$, откладывающейся на горизонтальной оси, и на каждом $[x_i, x_{i+1}]$ отрезке принимающая постоянное значение $p_i$.

$$p_i = \frac{v_i}{N(x_i - x_{i-1})}, x \in [x_{i-1}, x_i),$$

где $v_i$ – число точек, попадающих в интервал $[x_i, x_{i+1})$; $h$ – ширина шага.

$\frac{m_i}{N} \to \int_{x_{i-1}}^{x_i} f(x)dx,$

где $\frac{m_i}{N}$ - относительная частота события X ∈ $(x_i{-}1, x_i]$.

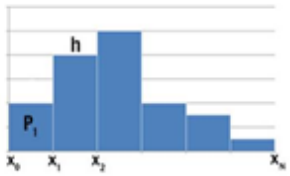


Рисунок 2 – Построение гистограммы

Рисунок 8 — Слайд 8

## Метод смещения гистограмм 9

**Монография:** Scott David Multivariate density estimation: theory, practice, and visualization.

**Суть метода:** пусть у нас есть m гистограмм $\widehat{f_1}, \widehat{f_2}, \dots, \widehat{f_m}$ с шириной каждого столбца h и с начальным столбцом в точке $t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \dots, \frac{(m-1)h}{m}$. Исходная осреднённая смещённая гистограмма определяется как $\hat{f}(\cdot) = \hat{f}_{ASH}(\cdot) = \frac{1}{m}\sum_{i=1}^{m}\widehat{f}_i(\cdot)$.

На каждом из интервалов шириной δ=h/m является кусочно-линейной.

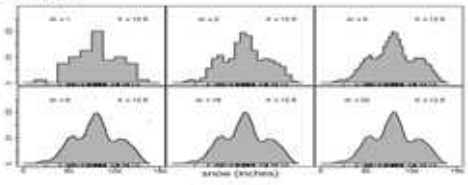

Рисунок 3 – Смещение гистограмм

Рисунок 9 — Слайд 9

## Программная реализация 10

Значения случайных величин равномерно распределялись на интервале [1,4]. Функция плотности вероятности в случае 4 равномерно распределённых величин определялась формулой:

$$p(x) = \begin{cases} \frac{1}{6}x^3, \text{если } 0 \le x \le 1; \\ -\frac{1}{2}x^3 + 2x^2 - 2x + \frac{2}{3}, \text{если } 1 \le x \le 2; \\ \frac{1}{2}x^3 - 4x^2 + 10x - \frac{22}{3}, \text{если } 2 \le x \le 3; \\ -\frac{1}{6}x^3 + 2x^2 - 8x + \frac{32}{3}, \text{если } 3 \le x \le 4. \end{cases}$$

Рисунок 10 — Слайд 10

## Программная реализация 11

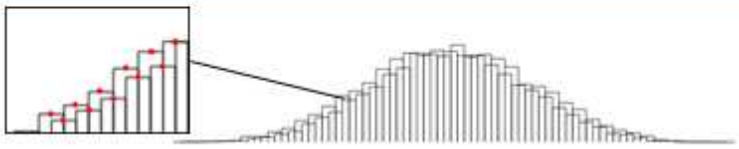

Рисунок 4 – Результат смещения двух гистограмм

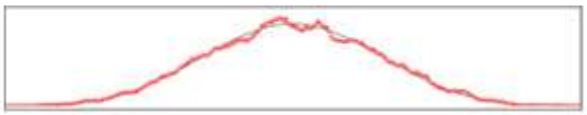

Рисунок 5 – Результат смещения 40 гистограмм с шагом $h = 0{,}1$ (2)

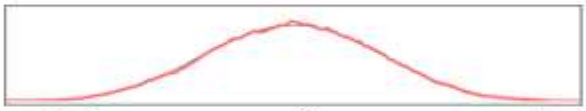

Рисунок 6 – Результат смещения 40 гистограмм с шагом $h = 0{,}2$

Рисунок 11 — Слайд 11

## Метод осреднения смещённых гистограмм 12

Середины $z$ столбцов $h$ гистограмм дают наилучшее приближение.

Рисунок 7 – Результат смещения гистограмм (20 гистограмм)

Рисунок 8 – Метод ядерных оценок (20 гистограмм)

Рисунок 12 — Слайд 12

## Сглаживание данных 13

- Метод скользящего среднего;
- Фильтр Савицкого-Голея;
- Метод полиномов;
- Сглаживание Вейбула;
- Сглаживающий сплайн и др.

Рисунок 13 — Слайд 13

## Сглаживающий сплайн 14

Сглаживающий сплайн - сплайн, который минимизирует следующий функционал, зависящий от некоторого параметра $p$:

$$I(s;p) = p\sum_{k=1}^{n} w_k(y_k - s(x_k))^2 + (1-p)\int_{x_1}^{x_n}\left(\frac{d^2s(x)}{dx^2}\right)^2 dx,$$

где $(x_k, y_k)_k = 1,2,\dots,n$ – приближаемые данные; $w_k$ – веса данных; $p$ - сглаживающий параметр, изменяющийся от 0 до 1.

Рисунок 14 — Слайд 14

## Программная реализация 15

Параметры: $n = 10\,000$; $h = 0{,}1$; $nr = 40$

$p = 0{,}994$; $SSE = 0{,}01125$ (сумма квадратов ошибок).

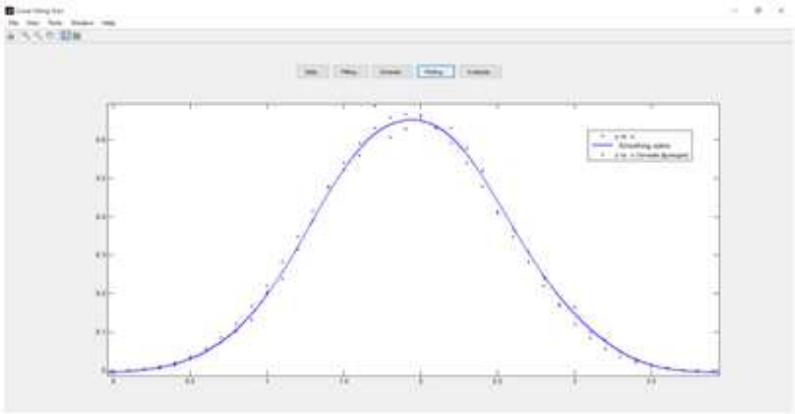

Рисунок 8 – Сглаживание восстановленной функции плотности вероятности

Рисунок 15 — Слайд 15

## Выводы 16

- Разработка и реализация метода осреднения смещённых гистограмм с оптимальными параметрами.
- Применения сглаживающего сплайна.
- Метод сопоставим с ядерными оценками.
- Повышение точности оценки.
- Дальнейшее использование комплекса методов на практике.

Рисунок 16 — Слайд 16

## Публикации 17

- Подходы к обработке экспериментальных данных в условиях ограниченной информации // Журнал «Научные исследования и разработки молодых учёных», 2015г. Стр. 136-140.
- Методы восстановления плотности вероятности в условиях ограниченных объёмов данных // Международная научная конференция «Молодежь и наука: проспект Свободный». 2016. Стр. 127 – 129.
- Модернизированный метод смещенных гистограмм для восстановления плотности вероятности // Журнал «Решетнёвские чтения», 2016 г. Стр. 237-238.
- Построение функции плотности вероятности на основе сглаживания смещенных гистограмм // Международная научная конференция «Молодежь и наука: проспект Свободный», 2017.

Рисунок 17 — Слайд 17

СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
SIBERIAN FEDERAL UNIVERSITY

**СПАСИБО ЗА ВНИМАНИЕ!**

Рисунок 18 — Слайд 18

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ Г. М. Цибульский
«___» _____ 2017 г.

# МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Разработка моделей и алгоритмов обработки эмпирических данных на основе численного вероятностного анализа

09.04.02 Информационные системы и технологии

09.04.02.01 Информационно-управляющие системы

| | | | |
|---|---|---|---|
| Руководитель | | проф., д-р.физ.-мат. наук | Б. С. Добронец |
| Студент | | КИ15-02-1/1М 031513681 | А. А. Чевер |
| Рецензент | | канд. физ.-мат. наук, ст. науч. сотр. | А. Н. Рогалев |
| Нормоконтролер | | доцент | М. А. Аникьева |

Красноярск 2017